BAYESIAN VARIABLE SELECTION MODEL AND DIFFERENTIAL CO-EXPRESSION NETWORK ANALYSIS FOR MULTI-OMICS DATA INTEGRATION

by

Li Zhu

MS, Biostatistics, Duke University, 2014

BS in Traditional Chinese Pharmacology, Zhejiang University, China, 2011

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Li Zhu

It was defended on

April 9th 2019

and approved by

George C. Tseng, ScD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Robert Krafty, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Lu Tang, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Wei Chen, PhD, Associate Professor, Department of Pediatrics, School of Medicine, University of Pittsburgh

Daniel E. Weeks, PhD, Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: George C. Tseng, ScD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

BAYESIAN VARIABLE SELECTION MODEL AND DIFFERENTIAL CO-EXPRESSION NETWORK ANALYSIS FOR MULTI-OMICS DATA INTEGRATION

Li Zhu, PhD

University of Pittsburgh, 2019

ABSTRACT

Due to the large accumulation of omics data sets in public repositories, innumerable studies have been designed to analyze omics data for various purposes. However, the analysis of a single data set may only provide limited information or suffer from small sample size and lack of reproducibility, thus data integration is gaining more and more attention nowadays. This dissertation focuses on developing methods for variable selection in regression (Chapter 2) and clustering (Chapter 3) for multi-omics data integration, and identification of differential co-expression networks (Chapter 4) in the transcriptomics meta-analysis setting.

In Chapter 2, we propose a Bayesian indicator variable selection model to incorporate multi-layer overlapping group structure (MOG) in the regression setting, motivated by the structure commonly encountered in multi-omics applications, in which a biological pathway contains tens to hundreds of genes and a gene can be mapped to multiple experimentally measured features (such as its mRNA expression, copy number variation and methylation levels at possibly multiple sites). We evaluated the model in simulations and two breast cancer examples, and demonstrated that this approach not only enhances prediction accuracy but also improves variable selection and model interpretation that lead to deeper biological insight into disease. In Chapter 3, we extended MOG to Gaussian mixture models for clustering, aiming to identify disease subtypes and detect subtype-predictive omics features.

In Chapter 4, we present a meta-analytic framework for detecting differential co-expression networks (MetaDCN). Differential co-expression (DC) analysis, different from conventional differential expression (DE) analysis, helps detect alterations of gene-gene correlations in case/control comparison, which is likely to be missed in DE analysis.

Public health significance:

Methods proposed in Chapter 2 - 3 not only can predict disease outcome or identify disease subtypes, but also determine relevant biomarkers, which can potentially facilitate the design of a test assay to monitor disease progression, predict disease subtypes, and guide treatment decisions. The method developed in Chapter 4 provides a novel framework for identifying differentially co-expressed genes to help us better understand how gene-gene interactions are altered in disease and to provide potential new molecular targets for drug development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my advisor Dr. George Tseng, who has introduced me to interesting research questions, provided insightful advices and inspired me to think deeper and harder whenever I was struggling. Those weekly meetings over the past five years have been proved to be one of my most valuable and enjoyable learning experiences at Pitt. George himself is an excellent example of a good statistician by always being humble, patient and curious. Being a member of his outstanding group is a great honor and an experience that I will benefit from for a lifetime.

I would like to express my gratitude to my committee members. Dr. Wei Chen introduced me to the exciting field of single cell RNAseq analysis and provided me tremendous guidance and help. Dr. Daniel Weeks is one of the most thorough reviewers, helping me realize the importance of details. I am also thankful to Dr. Robert Krafty and Dr. Lu Tang for all the helpful advices in both research and career.

I am also grateful to have Dr. Steffi Oesterreich and Dr. Adrian Lee at Magee Womens Research Institute (MWRI) as my mentors over the past four years. They generously provided me the opportunities to participate into their exciting projects and encouraged me to join their lab meeting, journal club and conferences, in which I obtained a lot of knowledge about breast cancer research.

I would like to thank my supportive lab mates and friends. Many thanks to Zhiguang (Caleb) Huo, Tianzhou (Charles) Ma and Shuchang (Silvia) Liu for their detailed guidance during my first few years in the lab. I also would like to thank Kevin Levine, Nilgun Tasdemir, Tian Du and Sayali Onkar from Dr. Oesterreich's lab for their support and expertise in breast cancer research.

I also want to show my deepest gratitude to my parents Aizhen Qian and Faming Zhu for letting me pursue my dream for so long and so far away from home. Finally, I would like to thank my husband Wen Sun for his unconditional love and support, for always believing in me and encouraging me to be a better me.

1.0 INTRODUCTION

In this chapter, background knowledge for this dissertation will be introduced. Section 1.1 presents an overview of the multi-omics data, followed by a review of the commonly used statistical models for various objectives in Section 1.2. In Section 1.3, we will review multi-omics data integration and two of its major directions: horizontal and vertical integration. Finally, an overview of the dissertation will be introduced in Section 1.4.

1.1 OVERVIEW OF MULTI-OMICS DATA

Omics data often refer to the data sets with the names ending with "-omics". They are measurements of an organism's genetic materials (genomics), epigenetic modifications (epigenomics), RNA transcripts (transcriptomics), and proteins (proteomics), which are the materials involved in the central dogma of biology: DNA $\rightarrow$ RNA $\rightarrow$ protein. Multi-omics data simply denote a collection of different types of omics data sets. Due to the rapid advance in high-throughput technologies, large volumes of multi-omics data have been accumulated in the past two decades. Innumerable studies have been conducted to analyze those data sets, resulting in a significantly enhanced understanding of biological processes. In this section, we will briefly review some omics data types which are relevant to this dissertation.

1.1.1 Genomics data

A genomics data set is a set of genetic materials, which, in humans, contains approximately 3.2×10^9 base pairs, distributed in 23 pairs of chromosomes. The average proportion of nucleotide differences between a randomly chosen pair of humans is estimated to be between 1/1500 to 1/1000 (Jorde and Wooding, 2004).

Genetic variations can be attributed to single base-pair substitution, insertion or deletion, structural variation, etc. Single base-pair substitution indicates the change of a single base in the nucleotide sequence of the genome. If the single nucleotide variant occurs in at least 1% of the population, it is called single nucleotide polymorphism (SNP). Genome-wide association study (GWAS) is commonly conducted to assess the association between a phenotype and each SNP. On the other hand, insertion and deletion indicate one or more base pairs are inserted or deleted in the nucleotide sequence.

Structural variation is generally defined as variation affecting more than one thousand base pairs, commonly referred as copy number variation (CNV). It can be caused by deletion or duplication of large regions of DNA. Numerous studies have shown that CNV is associated with disease phenotype (McCarroll and Altshuler, 2007).

1.1.2 Epigenomics data

An epigenomics data set is a set of epigenetic modifications, including DNA methylation, histone modification and chromatin structure change, which all play an important role in gene regulation, in addition to genotype. DNA methylation, as the mostly studied epigenomic alteration, indicates the addition of a methyl group to the DNA cytosine, resulting in a 5-methylcytosine. This process usually happens at a CpG site, which are the regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases along its 5' $\rightarrow$ 3' direction. DNA methylation is shown to be associated with several process, including genomic imprinting, silencing of repetitive DNA, and X-chromosome inactivation (Schübeler, 2015).

To quantify methylation levels, the β value is defined as the percentage of methylated events, which is between 0 and 1. It is also often transferred to the M value as $\log_2(\beta/(1-\beta))$.

1.1.3 Transcriptomics data

A transcriptomics data set is a set of RNA molecules, which includes messenger RNA (mRNA), ribosome RNA, transfer RNA, and other non-coding RNA, such as microRNA (miRNA). mRNA, as one of the mostly studied transcriptomics data, conveys the most genetic information from DNA transcription, and is directly related to protein translation. Unlike genomics data which are mostly same in all cells, mRNA represents the expression level of all genes inside a particular cell or tissue at one particular time. It is extensively studied due to fact that mRNA is highly indicative of function and current biological condition.

Currently, microarray and RNA-sequencing are the two mostly commonly used platforms to quantify mRNA levels. However, due to the lower background noise and high dynamic range, RNA-sequencing is gradually taking over microarray platform.

miRNA, which is a small non-coding RNA, is also gaining more attention nowadays. It functions by base-pairing with complementary mRNA, so that mRNA will be silenced (Bartel, 2009). Dysregulation of miRNA has been shown to be associated with several diseases (Jiang et al., 2008).

1.2 RELATED TOPICS OF STATISTICAL LEARNING IN OMICS DATA

Due to the large accumulation of omics data sets in public repositories, innumerable studies have been designed to analyze omics data for various purposes, including marker discovery, clinical outcome prediction, disease subtype identification, etc. In this section, we will review some commonly used statistical methods in omics data application.

1.2.1 Regression analysis

In statistics, a regression model is usually used to assess the relationship between a dependent variable and several independent variables. In biological sciences, it can be used to predict clinical outcomes, such as survival and disease subtypes, and simultaneously

identify associated biomarkers. Denoting the clinical outcome of interest as y_i $(i = 1, \cdots, n)$ for n subjects, and x_{ij} as the jth independent variable (covariate) in ith subject, the linear regression model can be expressed as

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i,$$

where ϵ_i is the error term, which usually is assumed to be independently drawn from $\mathcal{N}(0, \sigma^2)$. In multi-omics data analysis, the total number of covariates p is usually much larger than the number of subjects n. Therefore, feature selection is necessary for parameter identification, avoiding over-fitting, and better interpretation.

One of the frequently used feature selection methods is the regularization approach, which essentially imposes a constraint on the coefficients $\beta = (\beta_1, \cdots, \beta_p)^T$, so that the objective function becomes

$$\hat{\beta} = arg\min_{\beta}(\|y - X\beta\|^2 + \lambda\|\beta\|_q),$$

where $\lambda\|\beta\|_q$ is the penalty term, λ is a tuning parameter, and β_0 is omitted assuming data are centered. Several regularization methods have been proposed, with different selections of q. AIC/BIC corresponds to $q = 0$; Lasso (Tibshirani, 1996) corresponds to $q = 1$ (a.k.a. L_1 norm); Ridge regression uses the square of L_2 $(q = 2)$ norm as the penalty; Elastic net uses a combination of L_1 and L_2 norms (Zou and Hastie, 2005). In some cases, features are naturally structured into groups. Here, we denote β_g as the coefficients of features belonging to group g $(1 \leq g \leq G)$. To select or drop an entire group, Yuan and Lin (2006) proposed the group lasso (GL) using $\lambda \sum_{g=1}^{G} ||\beta_g||^2$ as the penalty. Later, to allow sparsity inside selected groups, Simon et al. (2013) proposed the sparse group lasso (SGL) with both L_1 norm penalty and group lasso penalty.

In a Bayesian framework, variable selection can be viewed as identifying nonzero variables in the posterior distribution. Park and Casella (2008) proposed a full Bayesian lasso model assuming Laplace distribution as the prior for coefficients, which provides more shrinkage than the normal distribution (see Figure 1A). Kyung et al. (2010) further derived the Bayesian version of the group lasso and the elastic net. Mitchell and Beauchamp (1988) proposed another type of prior called the "spike and slab" prior, which is a mixture

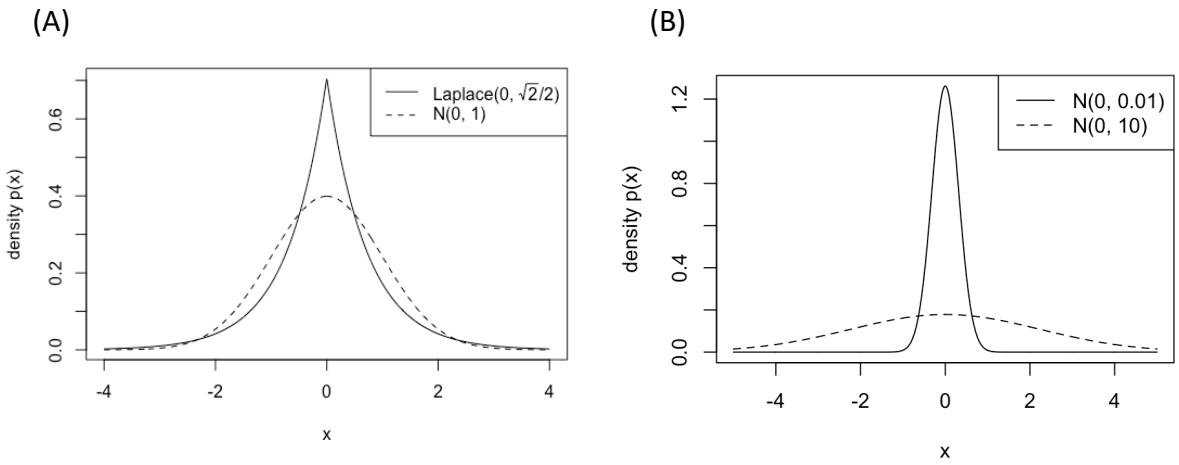


Figure 1: Demonstrating plots of two commonly used Bayesian variable selection priors. (A) shows a normal distribution and a Laplace distribution with the same mean and variance. (B) shows the two components of a spike and slab prior — a "spike" denoted by a normal distribution centered around zero and a "flat" denoted by a flat normal distribution.

of a point mass at zero (or a distribution centered around zero with small variance) and a diffuse uniform or large variance distribution (Figure 1B, see also George and McCulloch (1993) and Kuo and Mallick (1998)). Hernández-Lobato et al. (2013) generalized the spike-and-slab prior for group feature selection. Xu et al. (2015), Zhang et al. (2014a), and Chen et al. (2016) extended the spike-and-slab to achieve sparsity both at the group level and within groups.

However, in multi-omics data integration, the feature structure can be more complicated than the group structure. For instance, a biological pathway contains tens to hundreds of genes and a gene can be mapped to multiple different levels of measurements (such as mRNA expression, copy number variation and methylation levels of possibly multiple sites) See Figure 2 for an example. In Chapter 2, we will propose a Bayesian indicator variable selection model to incorporate multi-layer overlapping group structure.

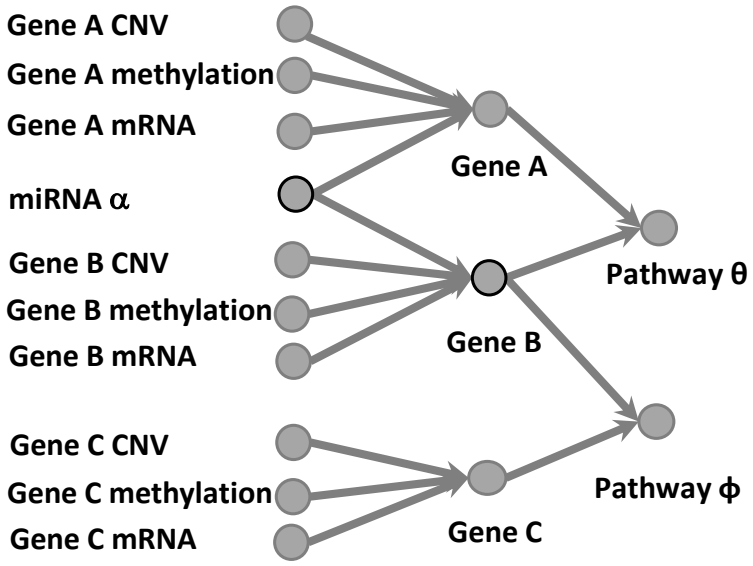


Figure 2: An example of a multi-layer overlapping group structure in a multi-omics dataset. Multi-omics features (mRNA expression, copy number variation (CNV), and DNA methylation) are mapped to genes, and genes are grouped into pathways. Some multi-omics features may belong to multiple gene groups. For example, miRNA α regulates both gene A and gene B. A gene may also belong to multiple pathways due to its multiple functions, such as gene B.

1.2.2 Clustering

Increasing evidence suggests that complex disease is usually not a single disease, instead it encompasses many subtypes. Identification of disease subtypes using clustering of multi-omics data is raising wide attention, because different subtypes are often related to different molecular mechanisms and require tailored treatment. For instance, Sørlie et al. (2003) utilized the transcriptomic profiles to group a breast tumor to one of the five subtypes of Luminal A, Luminal B, Normal-like, Her2-enriched, and Basal-like, which were demonstrated to have distinct molecular characteristics and clinical outcomes.

Classical clustering methods can be grouped into two main categories: distance-based methods, and model-based methods. K-means is one of the most commonly used distance-based clustering methods. Denoting x_{ij} as the observed feature j $(1 \leq j \leq p)$ in sample i $(1 \leq i \leq n)$, K-means aims to minimize the within-cluster sum of squares (WCSS):

$$\min_{C} \sum_{j=1}^{p} WCSS_j(C) = \min_{C} \sum_{j=1}^{p} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{s,t \in C_k} d_{st,j}, \tag{1.1}$$

where $C = (C_1, C_2, \ldots, C_K)$ is a partition corresponding to the clustering result, n_k is the number of samples in cluster k, and $d_{st,j} = (x_{sj} - x_{tj})^2$ is the Euclidean distance of gene j between sample s and t. Realizing that the total sum of squares ($TSS_j = \frac{1}{n}\sum_{s,t} d_{st,j}$), which is the sum of the $WCSS_j$ and the between-cluster sum of squares ($BCSS_j$), is a constant irrelevant to clustering, an equivalent objective function to Equation 1.1 is to maximize $BCSS_j$ as

$$\max_{C} \sum_{j=1}^{p} BCSS_j(C) = \max_{C} \sum_{j=1}^{p} \left(\frac{1}{n} \sum_{s,t} d_{st,j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{s,t \in C_k} d_{st,j} \right). \tag{1.2}$$

In high-dimensional data analysis, including all p features for clustering is likely to produce undesired clustering result due to too much noise, therefore further extensions were proposed to allow feature selection. Witten and Tibshirani (2010) proposed the sparse K-means (SPKM) algorithm by introducing a weight for each feature and a L_1 norm penalty of the weights to the K-means objective function in Equation 1.2:

$$\min_{C,w} - \sum_{j=1}^{p} w_j BCSS_j(C) + \lambda ||w||_1, \quad \text{subject to } ||w||_2 \leq 1, w_j \geq 0,$$

where λ is a tuning parameter. Huo and Tseng (2017) further proposed the integrative sparse K-means (ISKM) method, adding overlapping group lasso penalty to the Sparse K-means objective function to incorporate the group structure and allow groups to overlap:

$$\min_{C,w} -\sum_{j=1}^{p} w_j \frac{BCSS_j(C)}{TSS_j} + \lambda\alpha||w||_1 + \lambda(1-\alpha)\Omega(w),$$

$$\text{subject to } ||w||_2 \leq 1, w_j \geq 0,$$

where $\alpha \in [0,1]$ controls the balance between individual feature penalty ($||w||_1$) and group feature penalty ($\Omega(w)$). Define $\mathcal{J}_g$ as the set of index of features which belong to group g ($1 \leq g \leq G_0$), of which the cardinality $|\mathcal{J}_g|$ is the number of features that belong to group g, and $h(j) = \sum_{1\leq g\leq G_0} \mathbb{I}(j \in \mathcal{J}_g)$ as the number of groups containing feature j, the overlapping group lasso penalty term is defined as

$$\Omega(w) = \sum_{1\leq g\leq G} v_g ||m_g \bigcirc w||_2$$

where $v_g = \sqrt{\sum_{j\in\mathcal{J}_g} 1/h(j)}$ is the weight for group g, $m_g = (m_{g1}, \ldots, m_{gp})$ is the design vector with $m_{gj} = \mathbb{I}(j \in \mathcal{J}_g)/\sqrt{h(j)}$, and $\bigcirc$ is the Hadamard product.

Among model-based clustering methods, the Gaussian mixture model is frequently used, with the likelihood specified as

$$f(x_i) = \sum_{k=1}^{K} \pi_k \phi(\mu_k, \Sigma_k), \tag{1.3}$$

$$\phi(\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|} \exp\left(-\frac{1}{2}(x_i - \mu_k)\Sigma_k^{-1}(x_i - \mu_k)\right), \tag{1.4}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the observed feature vector in sample i, π_k is the probability of belonging to cluster k, $\mu_k = (\mu_{1,k}, \mu_{2,k}, \ldots, \mu_{p,k})$ is the mean vector in cluster k, and Σ_k is a $p \times p$ covariance matrix in cluster k. Similarly, to allow feature selection in high-dimensional data analysis, Pan and Shen (2007) introduced a L_1 penalty $h_\lambda = \lambda \sum_{j=1}^{p} \sum_{k=1}^{K} |\mu_{j,k}|$ into Equation 1.3 to encourage $\mu_{j,k}$ to be same across clusters. Xie et al. (2008) further proposed an extension adding the group lasso penalty.

In Chapter 3, similar to the regression analysis discussed in Section 1.2.1, we will introduce a Bayesian indicator variable selection prior into the Gaussian mixture model aiming to incorporate complex feature structures in multi-omics data integration.

1.2.3 Network analysis

Network analysis is often be used to understand the interactions between components in a biological system. The basic elements in a network are nodes and edges. Taking gene expression data as an example, each gene can be a node, and the edge between each pair of genes can denote the gene-gene correlation. Several attributes were defined to describe the properties and characteristics of each network. For example, the size of a network can be described by the number of nodes N and the number of edges E. For a simple undirected graph, which has at most one edge between any pairs of nodes and does not have a node with edge pointing to itself, the maximum number of edges is $E_{max} = \binom{N}{2}$. Then, the density of a network is defined as $D = \frac{E}{E_{max}}$.

There are various types of networks, including binary network, weighted network, directed network and undirected network. A simple way to generate a binary network is using correlation. For instance, let w_{ij} denote the correlation between nodes x_i and x_j, $e_{ij} = 1$ denote there exists an edge between them, and $e_{ij} = 0$ denotes there does not exist an edge between them. We will then define $e_{ij} = 1$, if $|w_{ij}| \geq \lambda$; and define $e_{ij} = 0$, otherwise. This correlation-based binary network can be used for differential co-expression analysis.

Differential co-expression (DC) refers to the change in gene-gene correlations between two conditions (e.g., cases and controls). Changes in gene-gene correlation may occur in the absence of differential expression, meaning that a gene may undergo radical changes in regulatory patterns that would be undetected by traditional differential expression (DE) analyses. Therefore, DC analysis can provide complementary information to standard differential expression (DE) analyses. Differential co-expression in two conditions could shed light on novel biological mechanisms. For example, a group of genes may be regulated by a common transcription factor or epigenetic modification, which is active in one condition but disrupted in the other.

In the literature, Lai et al. (2004) proposed an expected conditional F-statistics to identify differential co-expressed gene pairs, while Amar et al. (2013) and Bhattacharyya and Bandyopadhyay (2013) developed methods for direct identification of DC gene

modules. Choi and Kendziorski (2009) detected differential co-expression using predefined gene sets such as Gene Ontology (GO) categories. Although this approach incorporates prior biological information, it lacks the ability to detect novel DC modules. Another class of methods detected differential modules with genes highly co-expressed in one reference condition but with little or no correlation in the other condition. These types of methods rely on applying clustering methods to one reference condition, causing case-control asymmetry in the analysis (Watson, 2006; Ihmels et al., 2005). To circumvent this problem, Zhang and Horvath (2005) identified co-expressed modules in the entire (cases and controls combined) cohort through clustering and then evaluated their differential co-expression across conditions. Similarly, Tesson et al. (2010) extended this framework to detect differential co-expression modules by introducing the correlation changes between conditions into the dissimilarity matrix for clustering (DiffCoEx).

All methods described above for DC network detection focused on single transcriptomic study analysis. In Chapter 4, we will propose a meta-analytic framework for detecting differentially co-expressed networks (MetaDCN).

1.3 OMICS DATA INTEGRATION

Single data set, either single omics type or single cohort, may only provide limited information or suffer from small sample size and lack of reproducibility. Due to large accumulation of omics data sets, integrating multiple data sets serves as a better alternative to gain power and provide more robust conclusions.

There are mainly two types of integration in multi-omics applications: 1) "horizontal integration", which denotes the integration of multiple cohorts of subjects, who are measured for the same type of omics data (such as gene expression); 2) "vertical integration", which means an integration of multiple kinds of omics data types for the same set of subjects. Below, we will explain several methods developed in those two directions.

1.3.1 Horizontal integration

Many statistical methods aiming for "horizontal integration" have been proposed for different objectives. Tseng et al. (2012) reviewed 333 papers about "horizontal integration" of microarray data for purposes including differential expression analysis, pathway analysis, prediction, network analysis, etc. More recently, additional methods have been developed for clustering (Huo et al., 2016), dimension reduction (Kim et al., 2017), prediction using pairs of features (Kim et al., 2016), and differential expression analysis (Ma et al., 2017).

The meta-analysis framework for differential co-expression network (MetaDCN) developed in Chapter 4 belongs to this category of integration.

1.3.2 Vertical integration

Public data depositories, such as TCGA, have different omics data types collected for each sample. This has made the vertical integration possible.

In the regression setting, (Wang et al., 2012) proposed an integrative Bayesian analysis of genomics data (iBAG) framework, which identified important genes that were associated with clinical outcome when integrating gene expression and methylation data in TCGA. Fang et al. (2018) further extended iBAG to allow missing data.

Vertical integration is also common in clustering. Lock and Dunson (2013) proposed a method to cluster multi-omics data which allows both common and omic-type specific patterns. Shen et al. (2009) developed method iCluster, using a latent variable to cluster samples integrating multi-omics data. Huo and Tseng (2017) extended the sparse K-means framework by adding a penalty term similar to overlap group lasso to incorporate the group structure in multi-omics data.

The Bayesian indicator variable selection models developed in Chapter 2 and 3 implement vertical integration.

1.4 OVERVIEW OF THE THESIS

This dissertation includes five chapters. Chapter 1 gives an overview of the multi-omics data, relevant statistical methods and data integration, which serve as background information and motivation for methods developed in Chapters 2, 3 and 4.

In Chapter 2, we introduce a Bayesian indicator variable selection model to incorporate multi-layer overlapping group structure (MOG) in the regression setting, motivated by the structure commonly encountered in multi-omics applications, in which a biological pathway contains tens to hundreds of genes and a gene can be mapped to multiple experimentally measured features (such as its mRNA expression, copy number variation and methylation levels of possibly multiple sites). The contents of this chapter is currently under minor revision in the journal *The Annals of Applied Statistics*.

In Chapter 3, we extend the Bayesian indicator variable selection prior to the Gaussian mixture model for clustering, incorporating single- and multi-layer overlapping groups. To avoid determining of the number of clusters, which is often difficult, we will further extend the finite mixture model to Dirichlet process mixtures (DPM), allowing for more flexibility.

In Chapter 4, we will present a meta-analytic framework for detecting differential co-expression networks (MetaDCN), which can identify alterations of gene-gene correlations in case/control comparison that is likely to be missed in DE analysis. The contents of this chapter were published in the journal *Bioinformatics* (Zhu et al., 2016).

Chapter 5 contains the discussion and future work.

2.0 BAYESIAN INDICATOR VARIABLE SELECTION TO INCORPORATE HIERARCHICAL OVERLAPPING GROUP STRUCTURE IN MULTI-OMICS APPLICATIONS

The contents of this Chapter is currently under minor revision in the journal *The Annals of Applied Statistics*. An earlier version received one of the International Biometric Society Eastern North American Region's (ENAR) Distinguished Student Paper Awards.

2.1 INTRODUCTION

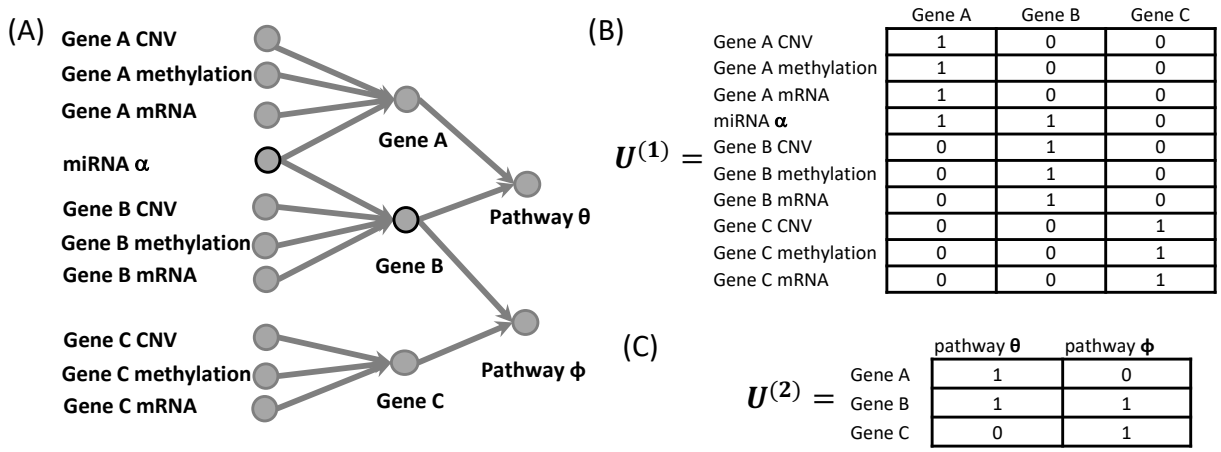
Variable selection is a pervasive problem in statistical applications, intended to search for the best model by eliminating unnecessary features. It gains increasing attention particularly in high dimensional data analysis, where the number of features often greatly exceeds the number of samples. For high-dimensional regression problems, it is commonly believed that only a small set of features have non-trivial effect on the outcome, while most other features have little or no effect. In the literature, the penalized regression method — lasso (Tibshirani, 1996) uses an L_1 norm penalty to achieve variable selection, however it tends to randomly select one out of a set of highly correlated variables while ignoring the others. Zou and Hastie (2005) proposed the elastic net method with a combination of L_1 and L_2 norm penalties to overcome this problem. When prior information of grouped variables is available and variable selection by groups is desired, Yuan and Lin (2006) proposed the group lasso penalty so that variables inside the same group are selected or dropped together. In order to further allow sparsity within selected groups, Simon et al. (2013) proposed the sparse group lasso with both a L_1 norm penalty and a group lasso penalty. In the counterpart of Bayesian

framework, variable selection can be viewed as identifying nonzero variables (or elimination of variables very close to zero) in the posterior distribution. Tibshirani (1996) pointed out that the lasso estimator is equivalent to the posterior median of a Gaussian model using the double exponential (Laplace) prior for each variable. Inspired by the hierarchical structure of Laplace prior, Park and Casella (2008) proposed a full Bayesian lasso model and Kyung et al. (2010) further derived the Bayesian version for the group lasso and the elastic net. Mitchell and Beauchamp (1988) proposed another popular type of prior called the "spike and slab" prior, which is a mixture of a point mass at zero (or a distribution centered around zero with small variance) and a diffuse uniform or large variance distribution (see also George and McCulloch (1993) and Kuo and Mallick (1998)). Hernández-Lobato et al. (2013) generalized the spike-and-slab prior for group feature selection and implemented the expectation propagation algorithm. Xu et al. (2015) and Zhang et al. (2014a) extended the spike-and-slab to achieve sparsity both at the group level and within groups. Under mild conditions, the posterior median estimator for a normal mean sample with the spike-and-slab prior is a soft-thresholding estimator with desired selection consistency and asymptotic normality properties (Johnstone and Silverman, 2004; Xu et al., 2015). Chen et al. (2016) developed a similar Bayesian model by introducing separate binary selection indicators for each group and each feature inside each group, which can also lead to sparsity at the group level and within groups.

All aforementioned methods allow only non-overlapping and single layer group structures. In this chapter, we consider a motivating example that requires incorporation of a hierarchical overlapping group structure. Suppose SNP array, methylation array, miRNA array and RNA-seq are performed on n tumor tissues to obtain genome-wide copy number variation (CNV), methylation, miRNA and mRNA expression measurements. Integration of such multi-level omics data has become prevalent in the research of many diseases and brought new statistical challenges (see Richardson et al. (2016) for review). Denote p as the total umber of variables in the union of all CNV, methylation sites, miRNA and mRNA expression features. The input data $X = \{x_{ij}\}$ is a $n \times p$ matrix, where n is the number of samples. Figure 3A shows an example of hierarchical overlapping group structure with two layers of groups. In the first layer of groups, four features belong to the gene A group: mRNA, CNV and methylation

probe of gene A, and miRNA α that targets gene A (knowledge known a priori from miRNA target database). Similarly, group gene B contains three multi-omics features, and miRNA α also targets this gene. The group structure of gene A and B, at the first layer, is an example of overlapping group structure as they are both targeted by miRNA α. In the second layer, pathway θ contains these two genes, and another pathway ϕ contains gene B and C. As a result, pathways θ and ϕ represent overlapping groups at the second layer as they both contain gene B. To formally represent such structure, we introduce two membership matrices for this example in Figure 3B and 3C. $U^{(1)}$ is a matrix with row dimension equal to the number of multi-omics features (i.e. $p = 10$), and column dimension equal to the number of genes (i.e. $m_1 = 3$, m_1 is the total gene number). $U^{(1)}_{jk} = 1$ denotes multi-omics feature j $(1 \leq j \leq p)$ belonging to gene k $(1 \leq k \leq m_1)$, otherwise $U^{(1)}_{jk} = 0$. Furthermore, we also introduce $U^{(2)}$ matrix with row dimension equal to the number of genes (i.e. $m_1 = 3$), and column dimension equal to the number of pathways (i.e. $m_2 = 2$, m_2 is the number of pathways). Again, $U^{(2)}_{kl} = 1$ denotes that gene k $(1 \leq k \leq m_1)$ belongs to pathway l $(1 \leq l \leq m_2)$, otherwise $U^{(2)}_{kl} = 0$. In this chapter, we consider a multi-omics linear regression setting $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where dependent variable $Y = \{y_i\}_{1 \leq i \leq n}$ is the clinical outcome, $X = \{x_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq p}$ represents measurements of multi-omics features, and feature number p is usually much larger than sample size n. Since $p >> n$, variable selection that properly incorporates prior biological knowledge is crucial. In our situation, the group structure of "multi-omics feature $\Rightarrow$ gene $\Rightarrow$ pathway" demonstrates a hierarchical overlapping group structure that brings challenges for variable selection in the regression setting.

A similar but simplified version of this structure was studied by Zhao et al. (2009) and Jenatton et al. (2011), in which, features are structured to form a tree, but the groups defined by nodes at the same depth are not allowed to overlap. They designed a specific group penalty, so that a child node group will only be selected when its parent node is selected. For general overlapping group structure, Jacob et al. (2009) proposed the concept of latent feature decomposition, which led to the solution support as the union of groups. Similarly, in the Bayesian framework, Zhang et al. (2014b) decomposed the marginal regression coefficient of a feature shared by multiple groups to be the sum of partial effects contributed by each group. With the hierarchical overlapping group structure, the target



	Gene A	Gene B	Gene C
Gene A CNV	1	0	0
Gene A methylation	1	0	0
Gene A mRNA	1	0	0
miRNA α	1	1	0
Gene B CNV	0	1	0
Gene B methylation	0	1	0
Gene B mRNA	0	1	0
Gene C CNV	0	0	1
Gene C methylation	0	0	1
Gene C mRNA	0	0	1

	pathway θ	pathway ϕ
Gene A	1	0
Gene B	1	1
Gene C	0	1

Figure 3: Motivating example of a multi-layer overlapping group structure in multi-omics dataset with membership matrices.

(A) The same example of a multi-layer overlapping group structure as shown in Figure 2. (B) $U^{(1)}$ membership matrix denotes if a multi-omics feature belongs to a certain gene. (C) $U^{(2)}$ membership matrix denotes if a gene belongs to a certain pathway.

function of penalized regression approaches generally becomes intractable to optimize. A Bayesian hierarchical model provides a natural alternative for incorporating the hierarchical overlapping group structure. We propose a multi-layer indicator variable selection model extended from Kuo and Mallick (1998) where three levels of binary indicators illustrate whether the corresponding multi-omics features, genes or pathways are selected. For overlapping groups, we adopt from Zhang et al. (2014a) the additive effect assumption for each overlapping group. We will show that incorporation of the hierarchical overlapping group structure enhances prediction accuracy and improves both feature selection and model interpretation.

The paper is organized as follows. In Section 2, we review the indicator variable selection model, and propose a Bayesian indicator variable selection model with single-layer and hierarchical (multi-layer) overlapping group structures. We describe the detailed MCMC algorithms for each model and extend the models to binary and survival outcomes. In Section 3, we illustrate the capabilities and limitations of existing methods compared to

our proposed model. In Section 4, four simulations are demonstrated to compare the performance of the proposed model and other existing methods. We further apply the model to data from two real examples in breast cancer, using multi-omics features to predict estrogen receptor (ER) status and histological subtype (invasive lobular carcinoma (ILC) versus invasive ductal carcinoma (IDC)) in Section 5. Section 6 contains final conclusion and discussion.

2.2 METHODS

2.2.1 Review of indicator variable selection model

Consider a linear regression setting, in which $y = (y_1, ..., y_n)^T$ denotes the outcomes for n samples, and X denotes an $n \times p$ covariate matrix for p variables. Assume data are centered and thus the intercept can be omitted. Under linear regression assumptions, $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, and $i = 1, \ldots, n$.

Bayesian indicator variable selection model was first proposed in Kuo and Mallick (1998). It embeds binary indicators into regression model to incorporate all 2^p candidate models. Denoting the binary indicator as γ_j, the indicator variable selection model is

$$y_i = \sum_{j=1}^{p} \beta_j \gamma_j x_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$
$$\beta = (\beta_1, \ldots, \beta_p)^T \sim N(0, D_0), \quad \gamma_j \sim Bern(\pi).$$

If $D_0 = s^2 I_{p\times p}$ is a diagonal matrix, where $I_{p\times p}$ is an identity matrix with dimension $p \times p$, and we define $\beta_j^* = \beta_j \gamma_j$, the indicator prior is equivalent to a spike-and-slab prior:

$$\beta_j^* \sim (1 - \pi)\delta_0(\cdot) + \pi N(0, s^2),$$

where $\delta_a(\cdot)$ is a Dirac delta function putting all mass at a.

This method is free of tuning and can be easily extended to more complicated modeling, such as a model with interactions. However, if the prior is too vague, mixing can be poor, as the sampled values of β_j may only rarely be in the region with high posterior support

(O'Hara et al., 2009). Other alternatives have been proposed (George and McCulloch, 1993), but most of them require additional tuning parameters.

2.2.2 SOG: Bayesian indicator variable selection with Single-layer Overlapping Groups

Motivated by the indicator variable selection model, we propose a Bayesian indicator variable selection model with Multi-layer hierarchical Overlapping Groups (MOG). We first introduce a simple version with only Single-layer Overlapping Groups (SOG).

Under the same linear regression setting in Section 2.2.1, we assume p variables (level-0 variables) belonging to m_1 possibly overlapping groups (level-1 groups). For instance, p experimentally measured features belong to m_1 genes. We define a $p \times m_1$ matrix $U^{(1)}$ to denote the group membership of level-0 variables, with $U^{(1)}_{j,k} = 1$ denoting that level-0 variable j belongs to level-1 group k, and $U^{(1)}_{j,k} = 0$ otherwise. We propose the following model:

$$Y_i \sim N\left(\sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2\right), \quad \beta_j = \sum_{k=1}^{m_1} U^{(1)}_{jk}\beta_{jk} \tag{2.1}$$

$$(\beta_{jk}|U^{(1)}_{jk} = 1) = \gamma^{(1)}_k \gamma^{(0)}_{jk} b_{jk}, \quad (\beta_{jk}|U^{(1)}_{jk} = 0) \sim \delta_0(\cdot), \tag{2.2}$$

$$\gamma^{(1)}_k \sim Bern(\pi^{(1)}), \quad \gamma^{(0)}_{jk} \sim Bern(\pi^{(0)}_k/R_j), \quad b_{jk} \sim N(0, s^2), \quad p(\sigma^2) \propto 1/\sigma^2, \tag{2.3}$$

where $R_j = \sum_{k=1}^{m_1} U^{(1)}_{jk}$ is the number of level-1 groups which includes level-0 variable j. The reason for scaling by R_j in the prior of $\gamma^{(0)}_{jk}$ is to ensure the same selection probability and variance of the marginal effect $\beta_j = \sum_{k=1}^{m_1} U^{(1)}_{jk}\beta_{jk}$ in the prior distribution. The justification is outlined below in Remark (2). In SOG, $\gamma^{(1)}_k$ can be interpreted as the selection indicator for level-1 group k; if $\gamma^{(1)}_k = 1$, $\gamma^{(0)}_{jk}$ can be interpreted as the selection indicator for level-0 variable j belonging to the level-1 group k; $\beta_{jk} \neq 0$ if and only if $\gamma^{(1)}_k = 1$ and $\gamma^{(0)}_{jk} = 1$. A singleton will be treated as a group with itself as its only member.

Markov chain Monte Carlo (MCMC) is used for model fitting. When groups do not overlap, all the full conditional distributions are available for Gibbs sampling; otherwise, Metropolis-Hastings is used to update $\pi^{(0)}_k$. See Appendix A.1.1.

Remarks:

(1) $U^{(1)}$ is a sparse matrix, most of whose entries are 0's and a few are 1's. $\sum_{k=1}^{m_1} U_{jk}^{(1)}$ is the number of level-1 groups that level-0 variable j belongs to. If $\sum_{k=1}^{m_1} U_{jk}^{(1)} > 1$, level-0 variable j belongs to multiple groups. $\sum_{j=1}^{p} U_{jk}^{(1)}$ is the number of level-0 variables that belong to level-1 group k. If $\sum_{j=1}^{p} U_{jk}^{(1)} U_{jk'}^{(1)} \geq 1$, level-1 groups k and k' overlap. β is also a $p \times m_1$ sparse matrix, with $\beta_{jk} \neq 0$ only when $U_{jk}^{(1)} = 1$.

(2) Assuming $\pi_k^{(0)} = \pi^{(0)}$ for all $1 \leq k \leq m_1$, prior of $Pr(\beta_j \neq 0) = 1 - \prod_{k=1}^{m_1} Pr(\beta_{jk} \neq 0)^{U_{jk}^{(1)}} = 1 - (1 - \pi^{(1)}\pi^{(0)}/R_j)^{R_j} \approx 1 - (1 - \pi^{(1)}\frac{\pi^{(0)}}{R_j} R_j)$ (if ignoring higher order terms) $= \pi^{(1)}\pi^{(0)}$, which is free of R_j; Meanwhile, prior for $Var(\beta_j) = E(\beta_j^2) = E\left(\sum_{k=1}^{m_1} U_{jk}^{(1)} \beta_{jk}\right)^2 = R_j\left(\pi^{(1)}\frac{\pi^{(0)}}{R_j} s^2\right) = \pi^{(1)}\pi^{(0)} s^2$, which is also free of R_j.

(3) In the case of duplicated variables such as $\beta_1 = \beta_{11} + \beta_{12}$ (feature 1 is shared by group 1 and 2), partial effects (β_{11},β_{12}) are not separately identifiable in the classical frequentist sense, since different parameter values can correspond to the same likelihood through the equal sum. This may seem to violate another definition of identifiability in the Bayesian framework, which we will refer to as "unidentifiable by likelihood "(Gelfand and Sahu, 1999). However, with an informative prior, or if the separate parameters share information from other parameters (e.g. β_{11} shares information with other parameters in group 1 and β_{12} shares information with other parameters in group 2 in our case), identifiability is not an issue, although slow convergence or unstable MCMC can be a problem (Eberly and Carlin, 2000). Nevertheless, the marginal parameter β_1 is our main interest of inference and is always identifiable by likelihood no matter in a frequentist or a Bayesian framework.

(4) For binary indicators $\gamma_k^{(1)}$ and $\gamma_{jk}^{(0)}$, there are three situations potentially not identifiable by likelihood (suppose two features belong to group k): (1) $\gamma_k^{(1)} = 0$, and $\gamma_{1k}^{(0)} = \gamma_{2k}^{(0)} = 0$; (2) $\gamma_k^{(1)} = 1$, and $\gamma_{1k}^{(0)} = \gamma_{2k}^{(0)} = 0$; and (3) $\gamma_k^{(1)} = 0$, and $\gamma_{1k}^{(0)} = 1$ or $\gamma_{2k}^{(0)} = 1$. Chen et al. (2016) used a conditional prior to avoid situation (3), so that whenever $\gamma_k^{(1)}$ is zero, $\gamma_{1k}^{(0)}$ and $\gamma_{2k}^{(0)}$ have to be zero. This conditional prior can be adopted into our model easily, but it still cannot distinguish situation (1) and (2) by avoiding a "false group" with all zero features in situation (2). Stingo et al. (2011) imposed three additional constraints for interpretability and identifiability. When $\gamma_k^{(1)} = 0$, they forced $\gamma_{1k}^{(0)} = \gamma_{1k}^{(0)} = 0$; and if

$\gamma_k^{(1)} = 1$, they eliminated the possibility of having $\gamma_{jk}^{(0)} = 0$ for all $j = 1, \cdots, p$. However, this constrained prior makes the Gibbs sampling infeasible. Thus, they have to adopt the Metropolis-Hastings algorithm, which can be inefficient when multi-layers of groups are introduced and feature dimension becomes large. Therefore, we decided not to add constraints in our prior, at a price that individual indicators $\gamma_k^{(1)}$ and $\gamma_{jk}^{(0)}$ may not be interpretable occasionally. Instead, they are used to impose group level and variable level sparsity. Variable selection eventually is determined at level-0 variable level by $\eta_{jk} = \gamma_k^{(1)}\gamma_{jk}^{(0)}$. Higher level group selection will be defined through group impact score (i.e. pathway impact score, PIS; see Section 2.5) to provide interpretation of selection at different layers of groups.

(5) s^2 controls the magnitude of the effect size. Here, for simplicity, we assume all b_{jk} are from the same distribution with common s^2. However, when dealing with multi-omics data in all our applications, we let s^2 be platform specific. In other words, methylation, CNV and gene expression can have different levels of variability.

(6) We assign hyper-priors: $\pi^{(1)} \sim Beta(a_1, b_1)$, $\pi_k^{(0)} \sim Beta(a_0, b_0)$, and $s^2 \sim Inverse - Gamma(a_s, b_s)$. If prior information is not available, we set $a_1 = b_1 = a_0 = b_0 = 1$, and $p(s^2) \propto 1/s^2$ (i.e. $a_s = b_s \approx 0$) as a non-informative prior. When the group size varies, borrowing information across groups will stabilize the estimate of $\pi_k^{(0)}$ for groups with small size. We consider two possible ways of information sharing: one is to assume that genes can be categorized into clusters, each with cluster-specific sparsity prior (Lock and Dunson, 2017), and the other is to use a common informative prior to stabilize the estimates. Since the former option is similar to the design of level-2 group sparsity, which will be proposed later, in this situation, we choose the second option and propose an empirical Bayes approach to estimate a_0 and b_0: (1) We first apply lasso regression (Tibshirani, 1996), ignoring any group structure; (2) Then, group specific sparsity $\hat{\pi}_k^{(0)}$ is estimated by counting number of non-zero coefficients inside each group k; (3) Finally, by moment matching, hyper-parameters are estimated as $\hat{a}_0 = \left(\frac{1-\hat{E}}{\hat{V}} - \frac{1}{\hat{E}}\right)\hat{E}^2$ and $\hat{b}_0 = \left(\frac{1-\hat{E}}{\hat{V}} - \frac{1}{\hat{E}}\right)(1-\hat{E})\hat{E}$, where $\hat{E}$ and $\hat{V}$ are the sample expectation and variance of $\hat{\pi}_k^{(0)}$ $(k = 1, \cdots, m_1)$. A simulation was conducted to evaluate the performance of borrowing information using the proposed empirical Bayes approach

(simulation V in Appendix A.3). When a large number of groups with a reasonable number of variables inside each group exist, borrowing information can better estimate $\pi_k^{(0)}$. When the number of groups or the number of variables in each group is small, this approach may produce inaccurate estimate of $\pi_k^{(0)}$, because a_0 and b_0 cannot be correct inferred. Due to the pros and cons of borrowing information to help estimate group specific sparsity, we allow users to choose the new empirical Bayes approach or the original non-informative approach in our R package. Users can decide which approach to use by evaluating performance in cross-validation. For all simulations and applications in this chapter, we will apply the original non-informative prior.

(7) The prior specified in Equation 2.1 does not imply each β_j is independent of others. In fact, the correlation structure is captured by groups, so that features belonging the same group share the common group selection indicator.

2.2.3 MOG: Bayesian indicator variable selection with Multi-layer hierarchical Overlapping Groups

In the presence of multi-layer (say s layers) hierarchical overlapping groups, we define $U^{(1)}$, $\ldots$, $U^{(s)}$, each with dimension $p \times m_1$, $m_1 \times m_2$, $\ldots$, $m_{s-1} \times m_s$ respectively, to specify the group structures. The multi-level omics data example in the introduction (Figure 3A) corresponds to a structure with $s = 2$. Below, we use $s = 2$ to illustrate the motivating example, but the model can be extended to $s > 2$. The proposed model for two-layer overlapping groups is

$$
\begin{gathered}
Y_i \sim N\left(\sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2} x_{ij}\beta_{jkl}, \sigma^2\right), \\
(\beta_{jkl}|U_{jk}^{(1)}U_{kl}^{(2)} = 1) = \gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}b_{jkl}, \quad (\beta_{jkl}|U_{jk}^{(1)}U_{kl}^{(2)} = 0) \sim \delta_0(\cdot), \\
\gamma_l^{(2)} \sim Bern(\pi^{(2)}), \quad \gamma_{kl}^{(1)} \sim Bern(\pi_l^{(1)}/D_k), \quad \gamma_{jkl}^{(0)} \sim Bern(\pi_{kl}^{(0)}/R_j), \\
b_{jkl} \sim N(0, s^2), \quad p(\sigma^2) \propto 1/\sigma^2,
\end{gathered}
$$

where $D_k = \sum_{l=1}^{m_2} U_{kl}^{(2)}$ is the number of level-2 groups which share level-1 group k. Similar to R_j in SOG, D_k and R_j here are used to ensure the same selection probability and variance for the marginal effect β_j in the prior distribution. In MOG, $\gamma_l^{(2)}$ can be interpreted as the selection indicator for level-2 group l; if $\gamma_l^{(2)} = 1$, $\gamma_{kl}^{(1)}$ can be interpreted as the selection indicator for level-1 group k belonging to level-2 group l; if $\gamma_l^{(2)}\gamma_{kl}^{(1)} = 1$, $\gamma_{jkl}^{(0)}$ can be interpreted as the selection indicator for level-0 variable j belonging to level-1 group k and level-2 group l; $\beta_{jkl} \neq 0$ if and only if $\gamma_l^{(2)} = 1$, $\gamma_{jk}^{(1)} = 1$, and $\gamma_{jkl}^{(0)} = 1$.

When prior information is not available, we assign non-informative hyper-priors similar to SOG: $\pi^{(2)} \sim Beta(1,1)$, $\pi_l^{(1)} \sim Beta(1,1)$, $\pi_{kl}^{(0)} \sim Beta(1,1)$, and $s^2 \propto 1/s^2$. MCMC sampling are described in Appendix A.1.2.

Asymptotic properties of SOG and MOG under orthogonal design are provided in Appendix A.2. Briefly, we show that the posterior median estimator of β_{jkl} is a soft-thresholding estimator with selection consistency and asymptotic normality, when the design matrix is orthogonal, p is fixed and $n \to \infty$. Although the conditions generally do not hold in multi-omics applications, it provides some insights to the proposed method.

2.2.4 Extension to binary and survival outcomes

For a binary outcome, we adopt the data augmentation from Albert and Chib (1993) introducing latent variable Z_i $(i = 1, ..., n)$ to replace Y_i in the regression:

$$Y_i = \begin{cases} 1, & \text{if } Z_i \geq 0 \\ 0, & \text{otherwise} \end{cases}, \quad Z_i = \beta_0 + x_i^T\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0,1),$$

where β_0 is the intercept, for which a non-informative prior $N(0, 100)$ is given.

For a survival outcome, we apply similar data augmentation (Tanner and Wong, 1987) for accelerated failure time (AFT) model, introducing a latent variable Z_i for time to event t_i and censor indicator δ_i ($\delta_i = 1$ indicating event happened):

$$\begin{cases} \log(t_i) = Z_i, & \text{if } \delta_i = 1 \\ \log(t_i) < Z_i, & \text{if } \delta_i = 0 \end{cases}, \quad Z_i = \beta_0 + x_i^T\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

2.3 RELATED METHODS

2.3.1 Capabilities and limitations of existing methods

Many methods have been proposed for variable selection with or without group structures. Here, we illustrate the major capabilities and limitations of several related methods comparing to SOG/MOG. Table 1 tabulates the key features and comparison of all methods.

- Penalized regression
  - Lasso (Tibshirani, 1996): One of the most popular variable selection methods using L_1 penalty at individual variable level and without any group structure.
  - Group lasso (GL) (Yuan and Lin, 2006): Variable selection is performed on the selection of groups using L_2 penalty. But, since there is no sparsity on individual variables, variables in each group are either entirely selected or entirely dropped.
  - Sparse group lasso (SGL) (Simon et al., 2013): The penalty term combines L_2 penalty on group level and L_1 penalty on individual variable level to achieve both group selection and sparsity inside a selected group. However, it is only applicable to single-layer group structure.
  - Tree structured group lasso (TGL) (Zhao et al., 2009; Liu et al., 2009): It is designed for hierarchical tree structured variables and it can lead to a sparsity pattern in which a group defined by child node is always selected after its parent node. However, groups defined by nodes at the same depth are not allowed to overlap and thus TGL is not applicable to overlapping group structure.
- Bayesian methods
  - BSGS (Bayesian sparse group lasso, Chen et al. (2016)):

$$
\begin{aligned}
&\gamma_k^{(1)} \sim Bern(1-\pi_k^{(1)}) \\
&\gamma_{jk}^{(0)}|\gamma_k^{(1)} \sim (1-\gamma_k^{(1)})\delta_0 + \gamma_k^{(1)} Bern(1-\pi_{jk}^{(0)}) \\
&\beta_{jk}|\gamma_k^{(1)}, \gamma_{jk}^{(0)} \sim (1-\gamma_k^{(1)}\gamma_{jk}^{(0)})\delta_0 + \gamma_k^{(1)}\gamma_{jk}^{(0)} N(0, s_{jk}^2),
\end{aligned}
$$

where $\pi_k^{(1)}$ and $\pi_{jk}^{(0)}$ are set to be 1/2, and s^2 is selected by cross-validation. This Bayesian hierarchical model is similar to SOG except that it pre-determines some hyper-parameters, assumes common group sparsity, and assumes conditional priors on binary indicators (see Section 2.2.2 Remark (4)). It does not allow a multi-layer hierarchical group structure.

– BSGS-SS (Bayesian sparse group selection with spike and slab prior, Xu et al. (2015)):

$$\begin{aligned}
&\beta_k = V_k^{1/2} b_k, \quad V_k^{1/2} = diag\left(\tau_{k1}, \ldots, \tau_{km_k}\right), \\
&b_k \sim (1-\pi_0) N_{m_k}(0, I_{m_k}) + \pi_0 \delta_0 \\
&\tau_{kj} \sim (1-\pi_1) N^+(0, s^2) + \pi_1 \delta_0
\end{aligned}$$

where β_k is the vector of coefficients corresponding to the features in level-1 group k, m_k is the number of level-0 features belonged to group k, V_k controls the magnitude of elements of β_k, $N^+(0, s^2)$ is a truncated normal distribution above zero, with mean as zero and variance s^2, and $N_{m_k}(0, I_{m_k})$ is the m_k dimensional normal distribution with mean as 0 and covariance as the identity matrix. Compared to SOG or BSGS, this Bayesian hierarchical model constructs binary indicators differently. But it still assumes common group sparsity and does not allow a multi-layer hierarchical group structure.

– HSVS (hierarchical structured variable selection, Zhang et al. (2014a)):

$$\begin{aligned}
&\beta_k | \gamma_k, \sigma^2, \tau_k^2 \sim (1-\gamma_k^{(1)}) \delta_0 + \gamma_k N(0, \sigma^2 D_{\tau_k}) \\
&D_{\tau_k} = diag(\tau_{k1}^2, \ldots, \tau_{km_k}^2), \\
&\gamma_k | \pi \sim Bern(\pi), \quad \tau_{kj}^2 | \lambda_k \sim Exp(\lambda_k^2/2),
\end{aligned}$$

where β_k and m_k are defined the same as in BSGS-SS, and $Exp(\lambda_k^2/2)$ is the exponential distribution with the rate parameter $\lambda_k^2/2$. This is another Bayesian indicator variable selection model similar to SOG. The method applies Laplace prior and does not generate exact zero estimates in MCMC. Sparsity is achieved by truncation at an arbitrary threshold. It does not allow a multi-layer hierarchical group structure.

Table 1: Compare MOG/SOG to some existing methods

Method	Feature selection	Exact zero in feature selection	Group selection	Exact zero in group selection	Varying sparsity inside groups	overlapping groups	multi-layer groups	Reference
MOG	✓	✓	✓	✓	✓	✓	✓	
SOG	✓	✓	✓	✓	✓	✓	✗	
BSGS	✓	✓	✓	✓	✗	◇	✗	Chen et al. (2016)
BSGS-SS	✓	✓	✓	✓	✗	◇	✗	Xu et al. (2015)
HSVS	✓	✗	✓	✓	✓	◇	✗	Zhang et al. (2014a)
TGL	✓	✓	✓	✓	-	✗	✓	Zhao et al. (2009)
SGL	✓	✓	✓	✓	-	✗	✗	Simon et al. (2013)
GL	✗	✗	✓	✓	-	✗	✗	Yuan and Lin (2006)
Lasso	✓	✓	✗	✗	-	✗	✗	Tibshirani (1996)

✓indicates it can be achieved; ✗indicates it cannot be achieved; ◇ indicates it cannot be achieved by the original method, but it can be achieved by an extended version in this chapter; - indicates it is not applicable.

2.3.2 Implementation and evaluation to compare with other existing models

We compared our model to three existing Bayesian models BSGS (Chen et al. (2016)), BSGS-SS (Xu et al., 2015) and HSVS (Zhang et al., 2014a), all of which can perform variable selection at the group level and within groups. Since BSGS requires all hyper-parameters to be pre-determined, we set them to the software default if available. The choice of τ^2 in BSGS, which serves the same purpose as s^2 in SOG, is a sensitive hyper-parameter without default value, and the details will be discussed in each simulation. When overlapping groups existed, we assumed that the marginal coefficient can be decomposed into the sum of partial coefficient as in Equation 2.1 when implementing all Bayesian models. When dealing with the binary outcome, we applied the same data augmentation in Section 2.2.4 to BSGS-SS and HSVS. The built-in function for binary outcome in BSGS package reported a fatal error, so we excluded it from our comparison. We also compared our model to lasso (Tibshirani, 1996), group lasso (GL) (Yuan and Lin, 2006), sparse group lasso (SGL) (Simon et al., 2013), and tree structured group lasso (TGL) (Zhao et al., 2009). Since TGL reduces to SGL when only a single-layer of groups exist, and it does not allow groups of the same level to overlap, we only evaluated its performance in simulation IV in which tree structured variables were simulated. The mixing weight α in SGL was set to be 0.95 by software default, thus more similar to lasso. The performance was evaluated by accuracy of both variable selection and prediction. In all the simulations and applications, data were randomly split into five folds, with four folds as training sets and one fold as the testing set.

In terms of variable selection performance, when the true β is known in simulation, the performance of variable selection relies on a tuning parameter or a cutoff. To eliminate the influence of arbitrary cutoffs in different methods, we derived sensitivity and specificity of variable selection under different cutoffs and calculated the area under the receiver operating curves (AUC) for a fair comparison. MOG (SOG), BSGS, and BSGS-SS can obtain exact zero estimates inside groups in each MCMC iteration, so level-0 variables were sorted according to posterior mean of the selection probability, which was calculated as $\hat{Pr}(\beta_j \neq 0|y,x) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\beta_j^{(b)} \neq 0)$, where $\beta_j^{(b)}$ is the b-th iteration of totally B converged MCMC samples. HSVS uses Laplace prior within groups and cannot obtain exact zeros

inside a group if the group is selected, even though the estimates are shrunk towards zero. As a result, we sorted the features based on $\max(p_{pos}, 1 - p_{pos})$, where p_{pos} is the posterior mean of $P(\beta_j > 0|y)$. For lasso, GL, SGL, and TGL, we applied multiple tuning parameters that detected different numbers of variables and formed the basis of ROC curve. Default tuning parameter sequences were used in lasso, GL, and SGL, whereas TGL calculated the max tuning parameter automatically and we selected a sequence of ratios as 0.01 to 1, with an increment of 0.01. For MOG (SOG), BSGS, and BSGS-SS, which produces coefficients as exact zeroes, we also controlled Bayesian false discovery rate (BFDR, Newton et al. (2004)) at the nominal level of 10% to compare their true FDRs (the number of false positives/the number of claimed positives) and false omission rates (FOR, the number of false negatives/number of claimed negatives).

To evaluate model prediction accuracy, the coefficient estimates need to be calculated. All Bayesian methods (MOG, SOG, BSGS, BSGS-SS, and HSVS) used posterior median estimator, whereas penalized regression methods (Lasso, GL, SGL, and TGL) used tuning parameters selected by 10-fold cross-validation. Note that, for features shared by more than one group, we only summarized the performance for the marginal effect β_j instead of partial coefficients. For continuous outcomes, we compared prediction mean squared error (MSE) in the testing set, i.e. $MSE = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} (x_{test,i}^T \hat{\beta} - y_{test,i})^2$, where n_{te} is the sample size in the testing set and $y_{test,i}$ is the i-th observation in the testing set. If the outcomes are binary, we sorted the samples in the testing set based on the predicted probability and calculated the prediction AUC.

R was used to implement all methods, except that TGL was implemented in Matlab. Gibbs sampler of all Bayesian models used 3,000 MCMC iterations (2,000 as burn-in) in simulations, and 20,000 iterations (10,000 as burn-in) in applications. BSGS, by default, uses Monte Carlo standard error (MCSE) for convergence diagnosis, and it only updates one group at each iteration. To make comparison fair but also save time, we applied 30,000 iterations (20,000 as burn-in) simulations with 10 groups in simulation I and II, and 200,000 iterations (100,000 as burn-in) in simulation III with 100 groups. In the end, we only included simulations which achieved MCSE below 0.1. When groups overlaped, SOG/MOG used the Metropolis-Hastings algorithm keeping 5000 iterations from stationary distribution, which

was monitored by Gekewe diagnosis (Geweke et al., 1991). We applied R packages MBSGS, glmnet, grplasso, SGL, and Matlab package SLEP (Liu et al., 2009), for BSGS-SS, lasso, GL, SGL, and TGL, respectively. R packages/functions for BSGS and HSVS were provided by the original authors. We provided all data and programming code in github (see Discussion section) to reproduce all results in this chapter.

2.4 SIMULATIONS

2.4.1 Simulation I: Single-layer non-overlapping groups

We first simulated data with single-layer non-overlapping groups to evaluate the performance of SOG. We set $n = 125$, $p = 200$, $m_1 = 10$, and $U^{(1)}$ with block diagonal structure as below:

$$U^{(1)} = \begin{bmatrix} 1_{20} & 0_{20} & \dots & 0_{20} \\ 0_{20} & 1_{20} & \dots & 0_{20} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{20} & 0_{20} & \dots & 1_{20} \end{bmatrix},$$

where 1_m (0_m) denotes an $m \times 1$ column matrix with all values equal to 1(0). In this setting, all 10 level-1 groups are disjoint, each having 20 level-0 variables. To model the within level-1 group correlation to be 0.5, for each level-1 group k ($k = 1, \dots, m_1$), we drew $z_k^{(1)}$ independently from $N(0,1)$, and sampled $x_{ij} = \left(z_k^{(1)} + e_{ij}\right)/\sqrt{2}$, where $e_{ij} \sim N(0,1)$, $1 \le i \le n$, and $1 \le j \le p$. The total number of effective β_{jk}'s with corresponding $U_{jk}^{(1)} = 1$ was 200. We set 50 out of those 200 β's to be non-zero, generated from $N(0,5)$. Other β's were set to be 0. We set the sparsity to vary among level-1 groups: group 1 had all 20 β's as non-zero; group 2 and 3 had 10 out of 20 β's as non-zero; group 4 and 5 had 5 out of 20 β's as non-zero. All other groups had all β's as zero. The outcomes were generated as $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_i \sim N(0,1)$.

We repeated the simulation 100 times, each evaluated by 5-fold cross validation. We compared the variable selection and prediction performance of SOG with BSGS, BSGS-SS, HSVS, lasso, GL, and SGL in Table 2, with the evaluation criteria described in Section 2.3.2.

We applied two values of τ^2 in BSGS, one being the truth $\tau^2 = 5$, and the other being $\tau^2 = 1$ to evaluate the sensitivity.

From Table 2, we can see that SOG had the best variable selection performance and prediction accuracy, with the highest AUC and the smallest MSE. For BSGS, even with the large amount of MCMC iterations, the number of valid simulations (MCSE < 0.1) left were limited, with 54 and 11 simulations left for $\tau^2 = 5$ and 1 respectively. Among the valid simulations, BSGS with the correct setting ($\tau^2 = 5$) had the similar feature selection AUC with SOG, but MSE was slightly larger. This is probably because BSGS estimates β as the average of its non-zero MCMC samples, which may be biased as it ignores the zeros; in addition, it assumes the same sparsity inside each group. BSGS-SS also had similar feature selection AUC and slightly higher prediction MSE, possibly also due to the same assumption of equal within group sparsity. HSVS had larger MSE and smaller AUC, which was likely because the Laplace prior failed to provide exact zero estimates. Lasso and GL both had poor performance as expected, since lasso does not consider group structure and GL does not consider sparsity within selected groups. SGL improved feature selection AUC over GL as expected, but it implicitly assumes equal proportion of true non-zero β's in each group. For SOG, BSGS, and BSGS-SS, the posterior distribution of feature selection can allow control of BFDR. Under nominal level of BFDR at 10%, the true FDR given the simulation truth are shown in Table 2. BSGS-SS was anti-conservative with 45% true FDR, while SOG and BSGS ($\tau^2 = 5$) properly controlled true FDR at 8% and 6%, respectively. In addition to smaller true FDR, SOG had only slightly higher FOR than BSGS-SS and lower than BSGS, showing its better feature selection performance.

2.4.2 Simulation II: Single-layer overlapping groups

We next simulated data with single-layer overlapping groups to evaluate the performance of SOG with BSGS, BSGS-SS and HSVS. The setting was exactly the same as simulation I in Section 2.4.1, except now $U^{(1)}_{1,1} = U^{(1)}_{1,2} = 1$ and $U^{(1)}_{41,3} = U^{(1)}_{41,4} = 1$ (see Figure 4A). In other words, we set level-0 variable 1 to belong to both level-1 group 1 and 2; level-0 variable 41 to belong to both group 3 and 4. To maintain the within group correlation at 0.5, for variables

Table 2: Variable selection and prediction performance from 100 repeats in simulation I-IV (mean(SE)).

	Model	Feature selection			Prediction
		Cutoff-free	Control nominal BFDR=0.1		
		AUC	True FDR	True FOR	MSE
Simulation I single-layer non-overlapping	SOG	0.99 (0.00)	0.08 (0.00)	0.03 (0.00)	3.15 (0.13)
	BSGS ($\tau^2 = 5$)	0.97 (0.00)	0.06 (0.01)	0.06 (0.00)	7.39 (0.98)
	BSGS ($\tau^2 = 1$)	0.95 (0.01)	0.15 (0.04)	0.07 (0.00)	7.92 (0.8)
	BSGS-SS	0.97 (0.00)	0.45 (0.02)	0.01 (0.00)	7.07 (2.50)
	HSVS	0.96 (0.00)	–	–	6.68 (0.31)
	Lasso	0.78 (0.00)	–	–	28.3 (1.44)
	GL	0.51 (0.00)	–	–	193.75 (11.31)
	SGL	0.74 (0.00)	–	–	41.64 (1.82)
Simulation II single-layer overlapping	SOG	0.98 (0.00)	0.11 (0.01)	0.04 (0.01)	5.27 (0.93)
	BSGS ($\tau^2 = 5$)	0.96 (0.01)	0.07 (0.02)	0.06 (0.00)	10.06 (2.11)
	BSGS ($\tau^2 = 1$)	0.87 (0.02)	0.26 (0.06)	0.07 (0.00)	23.66 (3.10)
	BSGS-SS	0.97 (0.00)	0.44 (0.01)	0.01 (0.00)	5.57 (1.04)
	HSVS	0.97 (0.00)	–	–	5.93 (0.29)
Simulation III U=0.2 two-layer overlapping	MOG	0.99 (0.00)	0.03 (0.00)	0.04 (0.00)	0.75 (0.03)
	SOG	0.97 (0.02)	0.02 (0.02)	0.11 (0.05)	3.92 (1.29)
	BSGS	0.86 (0.00)	0.03 (0.01)	0.25 (0.00)	29.64 (1.05)
	BSGS-SS	0.92 (0.00)	0.02 (0.00)	0.22 (0.01)	8.91 (0.33)
	HSVS	0.82 (0.01)	–	–	11.85 (0.46)
	Lasso	0.74 (0.00)	–	–	8.96 (0.25)
	GL	0.75 (0.00)	–	–	5.64 (0.17)
	SGL	0.74 (0.00)	–	–	8.52 (0.24)
Simulation III U=0.5 two-layer overlapping	MOG	1.00 (0.00)	0.1 (0.00)	0.00 (0.00)	0.54 (0.01)
	SOG	1.00 (0.00)	0.1 (0.01)	0.00 (0.00)	2.09 (0.09)
	BSGS	0.94 (0.01)	0.08 (0.02)	0.04 (0.00)	17.01 (1.36)
	BSGS-SS	0.96 (0.00)	0.07 (0.01)	0.11 (0.01)	28.99 (1.86)
	HSVS	0.98 (0.01)	–	–	4.57 (0.81)
	Lasso	0.77 (0.00)	–	–	42.15 (1.30)
	GL	0.81 (0.00)	–	–	20.51 (0.69)
	SGL	0.75 (0.00)	–	–	43.10 (1.24)
Simulation IV U=0.2 two-layer non-overlapping	MOG	1.00 (0.00)	0.03 (0.00)	0.04 (0.00)	0.76 (0.028)
	TGL	0.86 (0.04)	–	–	5.47 (1.37)
	Lasso	0.74 (0.00)	–	–	9.21 (0.19)
	GL	0.77 (0.00)	–	–	6.00 (0.20)
	SGL	0.74 (0.00)	–	–	8.34 (0.18)
Simulation IV U=0.5 two-layer non-overlapping	MOG	1.00 (0.00)	0.10 (0.00)	0.00 (0.00)	0.55 (0.015)
	TGL	0.88 (0.04)	–	–	18.47 (6.05)
	Lasso	0.77 (0.00)	–	–	42.26 (1.05)
	GL	0.80 (0.00)	–	–	22.14 (0.90)
	SGL	0.76 (0.00)	–	–	42.69 (0.97)

shared by more than one group, such as $x_{i,1}$, we first generated "pseudo" variables such as $x_{i,11}$ $(k = 1)$ and $x_{i,12}$ $(k = 2)$ as described in Section 2.4.1, and then set $x_{i,1}$ as the average of $x_{i,11}$ and $x_{i,12}$. β_{jk}'s and outcome y_i were generated the same way as in Section 2.4.1. We applied SOG, BSGS, BSGS-SS, and HSVS using the same duplication approach. Table 2 shows the evaluation results using 100 simulated data sets.

From the results, SOG continued to have the best variable selection and prediction performance. In fact, the results were very similar to simulation I. Even though we introduced overlapping feature coefficients (e.g. β_{11} and β_{12}) which were unidentifiable by likelihood, we were still able to estimate the marginal effects (e.g. $\beta_1 = \beta_{11} + \beta_{12}$), which were identifiable.

2.4.3 Simulation III: Two-layer overlapping groups

In this simulation, we simulated two-layers of overlapping groups to evaluate the performance of MOG. We set $n = 200$, $p = 300$, $m_1 = 100$, $m_2 = 10$, $U^{(1)}$ and $U^{(2)}$ with structures in Figure 4B and 4C. $U^{(1)}$ had a block diagonal structure, i.e. every three features belonged to one level-1 group; $U^{(2)}$ had a block diagonal structure in the most parts except $U^{(2)}_{1,1} = U^{(2)}_{1,2} = 1$ and $U^{(2)}_{21,3} = U^{(2)}_{21,4} = 1$, i.e. level-1 group 1 belonged to level-2 group 1 and 2; level-1 group 21 belonged to level-2 group 3 and 4.

In this setting, we only had overlapping level-2 groups while level-1 groups were disjoint. As a result, we could still compare MOG to SOG, BSGS, BSGS-SS, HSVS, GL, and SGL, as they only use level-1 group structure and ignore level-2 group structure. We used a similar approach to model the within group correlation. For each level-1 group k, we drew $z^{(1)}_k \sim N(0, 0.3)$; for each level-2 group l, we drew $z^{(2)}_l \sim N(0, 0.2)$; then we set $x_{ij} = z^{(1)}_k + z^{(2)}_l + e_{ij}$, where $e_{ij} \sim N(0, 0.5)$. In this way, $Var(X_{ij}) = 1$. For variables belonging to the same level-1 group, the correlation was 0.5; for variables belonging to the same level-2 group but different level-1 groups, the correlation was 0.2. Variables shared by more than one group were generated the same way as in simulation II. Outcomes were also generated as $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$.

We set 5 out of 10 level-2 groups to contain relevant features. Inside these 5 level-2 groups, we set 4 out of 10 level-1 groups to have strong signals (all three features in each

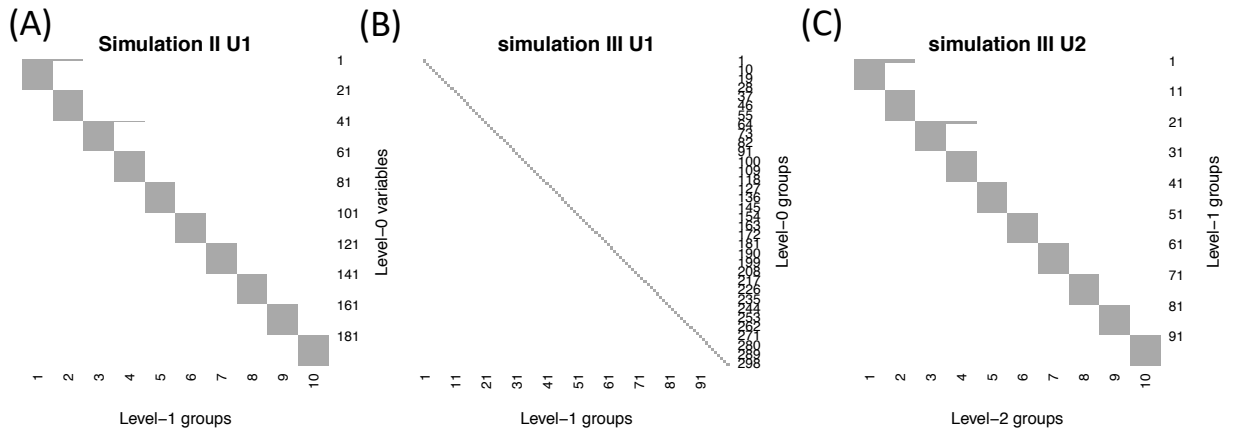


Figure 4: $U^{(1)}$ matrix in simulation II, $U^{(1)}$ and $U^{(2)}$ in simulation III.

Grey denotes 1, which means variable/group in the row belongs to the group in the column, while white denotes 0.

level-1 group have coefficients $\beta \sim Unif(2U, 3U)$; U will vary); the other 2 of 10 level-1 groups to have medium signals (all three variables in each level-1 group have coefficients $\beta \sim Unif(U, 2U)$). The remaining 4 level-1 groups had all 3 features with zero coefficients. We set U to be 0.2 and 0.5. In this setting, we did not have the true τ^2 to set in BSGS. Instead, we tested $\tau^2 = 1, 2, \ldots, 5$, performing 3-fold cross-validation in the training set, and then selected the one with the smallest MSE.

Table 2 shows the comparison results for 100 simulated data sets. MOG had the best performance in both variable selection AUC and prediction MSE, especially when U was small. When U=0.2, SOG had better performance than other models and MOG further improved SOG. The result showed the benefit of incorporating level-2 grouping structure. When the signal was weak, BSGS had a severe convergence issue, even with 200,000 MCMC iterations, which also impaired its feature selection and prediction performance. BSGS-SS had smaller FDR but higher FOR than SOG. This is because BSGS-SS assumes same sparsity inside groups. In the presence of groups with weak signals, it missed some weak features. At U=0.5, all four Bayesian models obtained similar good performance in feature selection, because all of them can perform variable selection both at group and within group level, and the sparsity inside level-1 groups with relevant features did not exist (i.e. all β's

were non-zero). But for prediction MSE, MOG still clearly outperformed other Bayesian models. Lasso, GL and SGL had poorer selection and prediction performance even when U was large. GL performed better than Lasso and SGL, because sparsity was not designed inside level-1 groups in this simulation.

2.4.4 Simulation IV: Two-layer non-overlapping groups

This simulation was designed to compare the performance of MOG and TGL since TGL does not allow groups at the same level to overlap as in simulation III. The only difference from the setting of simulation III was that level-2 groups did not overlap, so it was a straightforward extension of tree structure which included multiple trees. The implementation of TGL is described in Section 2.3.2, and results are shown in Table 2.

Compared to lasso, GL and SGL, TGL had better variable selection and prediction performance as expected. However, MOG still outperformed TGL in tree structure setting, regarding both variable selection and prediction. The improved performance is possibly because penalized regression methods are optimization-based and cannot incorporate complex structure and information flow as efficiently and naturally as Bayesian hierarchical models.

2.5 APPLICATIONS

2.5.1 Predict ER+ versus ER- breast cancer

We applied MOG to n=727 (560 ER+, 167 ER-) breast cancer patients retrieved from The Cancer Genome Atlas (TCGA). Each sample had mRNA expression, methylation, and copy number variations (CNV) available. This application aimed to predict estrogen receptor (ER) status and identify associated pathways, genes, and multi-level omics features simultaneously. We first filtered out genes with mRNA expression mean and variance below the median and constructed one summary methylation value for each gene by averaging the beta-values within 50 kb of the gene starting position. Beta-values were

later transformed to M-values to better fit model assumptions. After this filtering step, 14,976 features were left, of which 5,125 were from mRNA expression data, 4,816 were from CNV data, and 5,035 were from methylation data. We then downloaded KEGG pathways from MsigDB http://software.broadinstitute.org/gsea/msigdb/index.jsp.

Since BSGS-SS and HSVS are computationally intensive, we want to further filter genes and pathways. We first tested each mRNA expression feature for equal expression levels in ER+ and ER- groups, since all the genes were mapped to mRNA expression features. Then, we only kept the pathways containing 40-50 genes and having 80% of the genes with mRNA expression different in two groups (t-test $p-val < 0.05$), and filtered out features mapped to the genes that were not included in those selected pathways. A total of $p = 824$ multi-level omics features (level-0 variables) belonging to $m_1 = 276$ genes (level-1 groups) in $m_2 = 8$ pathways (level-2 groups) were left for analysis. Among 824 features, 276 were from gene expression data, 274 were from CNV data and the remaining 274 were from methylation data.

For another more realistic setting, we relaxed our filtering criteria. We kept pathways containing 20-50 genes and filtered out the features mapped to the genes which were not included in the selected pathways. This setting was used to compare the performance of MOG, SOG, lasso, GL, and SGL. In this way, $p = 11785$ multi-level omics features (1,316 from mRNA expression data, 1,292 from CNV data and 1,302 from methylation data) belonging to $m_1 = 1316$ genes (level-1 groups) in $m_2 = 123$ pathways (level-2 groups) were left for analysis.

Obviously, the "ER signaling pathway" should predict the ER status well. It was covered in both 8 and 123 pathways selected and could serve as an internal control. We applied SOG, BSGS-SS, HSVS, GL, and SGL, by using genes as group structure and ignoring level-2 pathway groups; we also applied lasso ignoring all group structures. Lasso, GL, and SGL used 10-fold cross-validation in the training set to select tuning parameters. Performance was evaluated using 5-fold cross-validation by keeping original case/control ratio in all folds. Each time, four folds of the ER+ and ER- samples were left for training, and one fold was left for testing. To avoid local optimal trapping and save time, when applying MOG and SOG, we used estimates from lasso as initial values. It took BSGS-SS and HSVS 1.4 and

19.7 hours to complete eight pathways example respectively, much longer than that of MOG (0.1 hours). These two models, especially HSVS, became inapplicable in larger data set such as those with 123 pathways.

To prioritize variable and group selection, we defined a feature impact score FIS_j in MOG as the posterior average of the selection probability of feature j, i.e. $FIS_j = AVE(\sum_{k=1}^{m_1}\sum_{l=1}^{m_2}\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}U_{jk}^{(1)}U_{kl}^{(2)})$, where $AVE(\cdot)$ was the average over all MCMC iterations after burn-in. The pathway impact score PIS_l was then defined as the average of the selection probability of all level-0 variables included in pathway l, i.e. $PIS_l = AVE(\sum_{j=1}^{p}\sum_{k=1}^{m_1}\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}U_{jk}^{(1)}U_{kl}^{(2)})$. In SOG, FIS and PIS were defined similarly, $FIS_j = AVE(\sum_{k=1}^{m_1}\gamma_k^{(1)}\gamma_{jk}^{(0)}U_{jk}^{(1)})$ and $PIS_l = AVE(\sum_{j=1}^{p}\sum_{k=1}^{m_1}\gamma_k^{(1)}\gamma_{jk}^{(0)}U_{jk}^{(1)}U_{kl}^{(2)})$. Setting $\gamma_k^{(1)} = 1$, denoting $\gamma_{jk}^{(0)} = 1$ if $\beta_{jk} \neq 0$, and denoting $\gamma_{jk}^{(0)} = 0$ otherwise, the definitions of FIS and PIS for BSGS-SS and HSVS are the same as SOG. We ranked the pathways and variables based on their impact scores averaged over 5-fold cross validations in Table 3 and 4. Top 20 selected multi-omics features by MOG are also listed in Table 12. Penalized regression models including lasso, GL, and SGL, cannot readily prioritize variables and pathways. Instead, we performed pathway enrichment analysis applying Fisher's exact test to features selected at least once in 5-fold cross-validation to prioritize the top pathways.

It is well known that the mRNA expression of ESR1 is predictive of ER status, defined by the immunohistochemistry (IHC) assay of estrogen receptor (ER). In both settings with 8 and 123 pathways, MOG detected the ER signaling pathway as the top selected pathway with the highest PIS, and ESR1-mRNA, ESR1-methyl and ESR1-CNV were among the top selected features. To obtain a better sense of the feature selection, we plotted the number of selected features ranked by FIS (x-axis) versus the number of selected features belonging to the ER signaling pathway (y-axis) in Figure 5. For lasso, GL, SGL, and TGL, for which FIS was not available, we used the feature selection results with the first fold data left out, as leaving different folds out gave similar results. Most of the top features selected by MOG, belonged to the ER signaling pathway (e.g. 92 out of top 100 in Figure 5A). Nonetheless other models had much fewer features in ER signaling (e.g. SOG had 27 out of top 100 in Figure 5A).

To compare the prediction performance, we calculated ER prediction AUC for samples in the testing set. For Bayesian models, we performed two predictions: (1) plugging posterior median estimates of β into $\hat{Pr}(Y_i = 0) = \Phi(X\hat{\beta}^{Med})$ to obtain AUC_1; (2) using model averaging by calculating posterior mean of $\Phi(X\hat{\beta})$ to generate AUC_2. For lasso, GL, and SGL, we selected tuning parameter from 10-fold cross-validation and plugged in $\hat{\beta}$. Having strong predictive genes such as ESR1, all models generated high AUCs in the testing set as expected. Comparing two AUCs, AUC_2 was slightly higher than AUC_1 in general for the Bayesian models, consistent with the common belief that averaging over all models from MCMC provides better predictive ability than using a single plug-in estimate. MOG using model averaging predictor generated the highest prediction AUC although the differences were not statistically significant given the almost perfect prediction for all models, which is probably because there exists other features correlated with features mapped to ESR1.

2.5.2 Predict invasive lobular carcinoma (ILC) versus invasive ductal carcinoma (IDC)

We next applied MOG to predict histological subtypes (ILC/IDC) for 669 patients (496 IDCs, 173 ILCs) in the same TCGA data set. Invasive lobular carcinoma (ILC) constituting 10% of all invasive breast cancer cases, is the second most frequently diagnosed subtype, following invasive ductal carcinoma (IDC, 80%) (Ciriello et al., 2015). We chose the same feature set preprocessed in Section 2.5.1, including 11,785 multi-level omics features (1,316 from mRNA expression data, 1,292 from CNV data and 1,302 from methylation data) belonging to $m_1 = 1316$ genes (level-1 groups) in $m_2 = 123$ KEGG pathways (level-2 groups), to compare the performance of MOG, SOG, lasso, GL, and SGL. Variable selection and prediction performances are summarized in Table 5, and top 20 multi-omics features selected by MOG are listed in Table 13. Similar to ER status, there exists a well-known strong predictor CDH1 mRNA expression, as the loss of CDH1 is the hallmark of ILC (Ciriello et al., 2015). Thus all models had good prediction $AUCs$. Since ILC is a less-studied subtype in breast cancer research, there is no annotated pathway specifically for this histologic subtype. The pathways identified by MOG provides proof-of-principle, as the top identified pathway

termed "Endometrial Cancer" not only includes E-cadherin (CDH1), but also contains PI3K and Akt, two kinases that are activated as a result of loss of CDH1 (Ciriello et al., 2015; Teo et al., 2018). And finally, there are a couple of genes such as APC, TCF7/TCF7L (Ravindranath and Cadigan, 2016) and LEF1 (Santiago et al., 2017) that all belong to the Wnt signaling pathway, highlighting a unique role for this pathway as we (Sikora et al., 2016) and others (Turashvili et al., 2007; van Hengel et al., 1999) have previously shown. Another top pathway identified is related to "Amoebiasis", and it includes many genes known to play diverse roles in movement and motility of cells, such as serpins, laminins, and extracellular movement, which we hypothesize is likely related to the different behavior of ILC cells, as a result of loss CDH1, and decreased cell-cell attachment, a phenotype that we have recently described in great detail (Tasdemir et al., 2018).

2.6 CONCLUSION AND DISCUSSION

In modern small-n-large-p applications, effective variable selection has become an increasingly important component in statistical methodologies. models that incorporate prior structural knowledge of variables (e.g. group lasso and fused lasso) can improve variable selection, prediction accuracy and model interpretation. In this chapter, we consider a hierarchical overlapping group structure that is commonly seen in the "multi-level omics features $\Rightarrow$ genes $\Rightarrow$ pathways" scenario in genomic applications. Our proposed Bayesian indicator variable selection model has several innovations and advantages for the targeted problem. Firstly, Bayesian hierarchical model and indicator variable selection model allow for natural incorporation of hierarchical group structure with fast MCMC sampling. Secondly, we explicitly model group-specific proportions of non-zero β values (i.e. $\pi_k^{(0)}$) for different sparsity levels in different selected groups. Thirdly, our Bayesian approach allows for a simple duplication technique to incorporate overlapping groups. Fourthly, the proposed model can be extended to more than two layers of overlapping group structure. The result gives clear interpretation of which features, genes and pathways contributing to the prediction. Finally, the posterior distribution from MCMC samples provides easy post hoc inferences,

such as characterization of variability and BFDR control of feature selection. Using four simulation settings, we demonstrated superior performance of the proposed method in terms of variable selection and prediction accuracy. In the applications to breast cancer patients, we also showed better performance of the proposed models in variable selection and model interpretation.

Our proposed model has several limitations to be improved in the future. First, as noted in the paper, the MCMC mixing rate in the indicator model can be unstable, leading to slow convergence. Although our current simulation and application can be implemented adequately, we expect worse performance when p increases or the data signal becomes weaker. A modification to spike-and-slab prior with a small-variance Gaussian spike might alleviate the computing difficulty. Secondly, in SOG/MOG, feature sparsity varies by gene groups. To better allow for heterogeneity among multi-omics platforms, a more sophisticated sparsity modeling may be needed to allow for different levels of sparsity in different platforms. Specifically, taking MOG as an example, we can design feature sparsity prior through a probit function: $\gamma_{jkl}^{(0)} \sim Bern(\Phi(\mu_{kl}^{(0)}/R_j + \mu_m))$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution, $\mu_{kl}^{(0)}$ is the feature selection strength of gene k in pathway l, and μ_m is the feature selection strength of multi-omics platform m. Since this implementation may bring computational challenges and significantly slow down computing time, we did not implement it in this chapter to allow for practical omics applications, but we consider this as a future extension. As large data sets with complex prior information structure continue to accumulate in data science, we expect to encounter the hierarchical overlapping group structure more often in the future and the proposed method can better incorporate prior information to improve statistical learning performance.

An R package "MOG" calling C++ using RcppEigen (Bates and Eddelbuettel, 2013) and the code to generate simulation and application results are available at github https://github.com/lizhu06/MOG. The computing time for MOG to predict ER+/ER- with 123 pathways is 2.33 hours, and computing time to predict ILC/IDC is 2.26 hours with 16 CPU cores, 1.4 GHz and 128 GB RAM.

Table 3: 5-fold cross-validation AUC in breast cancer ER+/- application.

A. 8 pathways

Bayesian model	AUC_1 (SE)	AUC_2 (SE)
MOG	0.943 (0.008)	0.949 (0.010)
SOG	0.945 (0.009)	0.948 (0.010)
BSGS-SS	0.947 (0.009)	0.948 (0.013)
HSVS	0.942 (0.012)	0.944 (0.012)
Penalized regression	AUC (SE)	
Lasso	0.945 (0.008)	
GL	0.943 (0.011)	
SGL	0.764 (0.108)	
TGL	0.946 (0.010)	

B. 123 pathways

Bayesian	AUC_1 (SE)	AUC_2(SE)
MOG	0.940 (0.013)	0.944 (0.011)
SOG	0.943 (0.009)	0.944 (0.011)
Penalized regression	AUC (SE)	
Lasso	0.946 (0.006)	
GL	0.942 (0.011)	
SGL	0.681 (0.111)	
TGL	0.944 (0.009)	

AUC_1: Plug-in $\hat{\beta}$ (posterior median); AUC_2: Model averaging

Table 4: Top pathways and features selected in breast cancer ER+/- application. Results are from 5-fold cross-validation.

A. 8 pathways

Bayesian model	Top pathway by PIS	PIS	Top 3 selected features by FIS
MOG	ER signaling	0.109	ESR1-mRNA, ESR1-methyl, ESR1-CNV
SOG	ER signaling	0.053	ESR1-mRNA, ESR1-methyl, NME3-mRNA
BSGS-SS	ER signaling	0.020	ESR1-mRNA, NME3-mRNA, ADCY9-mRNA
HSVS	ER signaling	0.027	ESR1-mRNA, ESR1-CNV, ESR1-methyl
Penalized regression	Top pathway by Fisher's exact test	Fisher's exact test p-val	–
Lasso	Calcium signaling	0.179	–
GL	ER signaling	0.999	–
SGL	ER signaling	0.152	–
TGL	AMPK signaling	0.204	

B. 123 pathways

Bayesian	Top pathway	PIS	Top 3 selected features
MOG	ER signaling	0.044	ESR1-mRNA, ESR1-methyl, ESR1-CNV
SOG	Prolactin signaling	0.031	ESR1-mRNA, MARCKS-mRNA, ESR1-methyl
Penalized regression	Top pathway by Fisher's exact test	Fisher's exact test p-val	–
Lasso	Dilated cardiomyopathy	0.0004	–
GL	RNA transport	0.528	–
SGL	Prolactin signaling	0.021	–
TGL	Adrenergic signaling	0.007	–

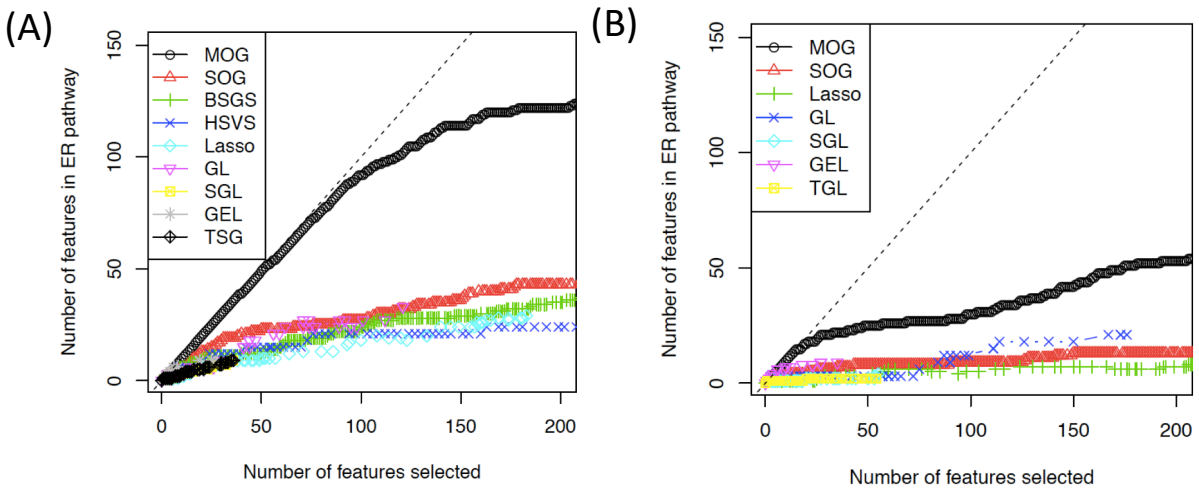


Figure 5: The number of top selected features versus the number of selected features belonging to the ER signaling pathway in breast cancer ER+/- application.

Application using (A) 8 pathways and (B) 123 pathways.

Table 5: Feature selection and prediction results in breast cancer ILC/IDC application. Results are from 5-fold cross-validation

Bayesian model [c]	Top pathway from PIS	PIS	Top 3 selected features by FIS	AUC_1 [a] (SD)	AUC_2 [b] (SD)
MOG	Endometrial cancer	0.064	CDH1-mRNA, LAMA3-mRNA, CDH1-methyl	0.941 (0.008)	0.949 (0.014)
SOG	Viral myocarditis	0.044	CDH1-mRNA, MAP3K1-mRNA, SHROOM1-mRNA	0.911 (0.010)	0.950 (0.012)
Penalized regression	Top pathway by Fisher's exact test	Fisher's exact test p-val	–	*AUC* (SD)	
Lasso	Thyroid hormone synthesis	0.008	–	0.956 (0.009)	
GL	Notch signaling	0.593	–	0.953 (0.011)	
SGL	Endometrial cancer	0.017	–	0.901 (0.011)	
TGL	AMPK signaling	0.005	–	0.955 (0.010)	

[a] Plug-in $\hat{\beta}$ (posterior median)
[b] Model averaging
[c] Computation not affordable with 123 pathways in BSGS-SS and HSVS models

3.0 BAYESIAN CLUSTERING WITH INDICATOR VARIABLE SELECTION MODEL TO INCORPORATE MULTI-LAYER OVERLAPPING GROUP STRUCTURE IN MULTI-OMICS APPLICATIONS

3.1 INTRODUCTION

Increasing evidence has suggested that complex disease, such as cancer, is not a single disease, instead it encompasses several different subtypes. Breast cancer, as one representative example, can be categorized into different subtypes based on histopathological type, tumor grade, tumor stage or molecular markers such as the estrogen receptor. Identifying disease subtypes using clustering methods has received wide attention, because different subtypes are often related to different molecular mechanisms and require tailored treatment. For instance, Sørlie et al. (2003) utilized the transcriptomic profiles to group breast tumors to one of the five subtypes of Luminal A, Luminal B, Normal-like, Her2-enriched, and Basal-like, which were demonstrated to have distinct molecular characteristics and clinical outcomes. Nowadays, with the rapid development of next-generation sequencing, large amounts of omics data sets are accumulated in public repositories. The Cancer Genome Atlas (TCGA), as an example, includes genomics, epigenomics, transcriptomics, and proteomics data for more than 11,000 patients spanning 33 cancer types. This kind of database provides unprecedented opportunities to better characterize subtypes but also raises several challenges: (1) How can we integrate multiple omics data types of the same set of patients, which may provide more comprehensive clustering and increase statistical power? (2) How to effectively select relevant features from the huge feature set with such a complex structure? (3) How to correctly determine the number of clusters that fully capture all unique subtypes but are also biologically interpretable?

Classical clustering methods can be grouped into two main categories: distance-based methods (e.g. hierarchical clustering and K-means) and model-based methods (e.g. Gaussian mixture models). In high dimensional data analysis, several extensions have been made in both categories to incorporate feature selection. Witten and Tibshirani (2010) proposed the sparse K-means (SPKM) algorithm by introducing a weight for each feature and a L_1 norm of the weights as the penalty term to the K-means objective function, which maximizes between-cluster variation. Huo and Tseng (2017) proposed a further extension — the integrative sparse K-means (ISKM) method, adding a penalty term similar to the overlapping group lasso penalty to the SPKM objective function to incorporate the overlapping group structure. On the other hand, Gaussian mixture model is widely used for model-based clustering. To accommodate feature selection, Pan and Shen (2007) introduced a L_1 norm penalty of the mean parameter in the Gaussian mixture model to encourage features to have the same mean across clusters. Xie et al. (2008) further extended it by adding the group lasso like penalty to the mean parameter to incorporate the group structure. Section 1.2.2 introduces the details of each method.

In this chapter, we extend the Bayesian indicator variable selection prior proposed in Chapter 2 to a Gaussian mixture model for the purpose of clustering. This extension, to our best knowledge, is the first Bayesian clustering method allowing overlapping group structures among features. More importantly, it allows the incorporation of multi-layor overlapping group structures commonly encountered in multi-omics data. To avoid determining the number of clusters, which is often difficult, we further extend the finite mixture model to Dirichlet process mixtures (DPM), allowing more flexibility.

This chapter is organized as follows. In Section 3.2, we propose a Bayesian clustering model with the indicator variable selection prior for single- and multi-layer overlapping group structure, and a further extension to DPM. In Section 3.3, three simulations are demonstrated to compare the performance of proposed models with other existing methods. In Section 3.4, we apply our models to two real data sets with single- or multi-layer overlapping groups. Finally, Section 3.5 contains the conclusion and discussion.

3.2 METHODS

3.2.1 Bayesian Clustering with indicator variable selection for Single-layer Overlapping Groups (SOGC)

Denote the observed data matrix as $X_{n\times p}$, where n is the number of samples, and p is the number of features; and denote $z_{n\times 1} \in \{1, \cdots, K\}$ as the unobserved cluster labels, where K is the total number of clusters. We assume p features (level-0 variables) belong to m_1 possibly overlapping groups (level-1 groups). For instance, p experimentally measured features are mapped to m_1 genes, or p genes are grouped into m_1 pathways. We define a $p \times m_1$ matrix $U^{(1)}$ to denote the group membership of level-0 variables, with $U^{(1)}_{jg} = 1$ denoting that level-0 variable j belongs to level-1 group g, and $U^{(1)}_{jg} = 0$ otherwise. With data centered to zero and scaled to have variance as one, we assume most features are irrelevant to clustering (i.e. subtype-irrelevant features) and have a mean of zero in all clusters, while only a few of them are subtype-relevant features. We propose a Bayesian clustering model for single-layer overlapping groups (SOGC) as

$$
\begin{gathered}
X_{ij}|z_j = k \sim \mathcal{N}(\mu_{j,k}, \sigma_j^2), \quad \mu_{j,k} = \sum_{g=1}^{m_1} U^{(1)}_{jg}\mu_{jg,k}, \\
(\mu_{jg,k}|U^{(1)}_{jg} = 0) \sim \delta_0(\cdot), \quad (\mu_{jg,k}|U^{(1)}_{jg} = 1) = \gamma^{(1)}_{g,k}\gamma^{(0)}_{jg,k}\mu_{jg,k}, \\
\gamma^{(1)}_{g,k} \sim Bern(\pi^{(1)}_k), \quad \gamma^{(0)}_{jg,k} \sim Bern(\pi^{(0)}_{g,k}/R_j), \quad \mu_{jg,k} \sim \mathcal{N}(0, s^2), \\
p(\sigma_j^2) \propto 1/\sigma_j^2, \quad p(s^2) \propto 1/s^2,
\end{gathered}
$$

where $\delta_a(\cdot)$ is a Dirac delta function putting all mass at a, $R_j = \sum_{g=1}^{m_1} U^{(1)}_{jg}$ is the number of level-1 groups containing level-0 feature j, which is to used to ensure that the prior selection probability for all features are the same. $\gamma^{(1)}_{g,k}$ can be interpreted as the selection indicator for level-1 group g in cluster k; if $\gamma^{(1)}_{g,k} = 1$, $\gamma^{(0)}_{jg,k}$ can be interpreted as the selection indicator for level-0 variable j belonging to level-1 group g in cluster k; $\mu_{jg,k} \neq 0$ if and only if $\gamma^{(1)}_{g,k} = 1$ and $\gamma^{(0)}_{jg,k} = 1$. If feature j belongs to multiple level-1 groups, we assume the same decomposition as in Chapter 2 that $\mu_{j,k}$ can be decomposed into the sum of partial means $\mu_{jg,k}$. A singleton will be treated as a group with itself as its only member.

3.2.2 Bayesian clustering with indicator variable selection for Multi-layer Overlapping Groups (MOGC)

In the presence of multi-layer (say s layers) overlapping groups, we define $U^{(1)}$, ..., $U^{(s)}$, each with dimension $p \times m_1$, $m_1 \times m_2$, ..., $m_{s-1} \times m_s$ respectively, to specify the group structures. We propose a Bayesian clustering model with indicator variable selection prior for multi-layer overlapping groups (MOGC) as

$$
\begin{aligned}
X_{ij}|z_i = k \sim \mathcal{N}(\mu_{j,k}, \sigma_j^2), \quad \mu_{j,k} = \sum_{g=1}^{m_1}\sum_{l=1}^{m_2} U_{jg}^{(1)} U_{gl}^{(2)} \mu_{jgl,k}, \\
(\mu_{jgl,k}|U_{jg}^{(1)}U_{gl}^{(2)} = 0) \sim \delta_0(\cdot), \quad (\mu_{jgl,k}|U_{jg}^{(1)}U_{gl}^{(2)} = 1) = \gamma_{l,k}^{(2)}\gamma_{gl,k}^{(1)}\gamma_{jgl,k}^{(0)} b_{jgl,k}, \\
\gamma_{l,k}^{(2)} \sim Bern(\pi_k^{(2)}), \quad \gamma_{gl,k}^{(1)} \sim Bern(\pi_{l,k}^{(1)}/D_g), \quad \gamma_{jgl,k}^{(0)} \sim Bern(\pi_{gl,k}^{(0)}/R_j), \\
b_{jgl,k} \sim \mathcal{N}(0, s^2), \quad p(\sigma_j^2) \propto 1/\sigma_j^2, \quad p(s^2) \propto 1/s^2
\end{aligned}
$$

where $\gamma_{l,k}^{(2)}$ is now the selection indicator for level-2 group l in cluster k; if $\gamma_{l,k}^{(2)} = 1$, $\gamma_{gl,k}^{(1)}$ can be interpreted as the selection indicator for level-1 group g belonging to level-2 group l in cluster k; similarly, if $\gamma_{l,k}^{(2)}\gamma_{gl,k}^{(1)} = 1$, $\gamma_{jgl,k}^{(0)}$ is the selection indicator for level-0 variable j belonging to level-1 group g and level-2 group l in cluster k; $\mu_{jgl,k} \neq 0$ if and only if $\gamma_{l,k}^{(2)} = 1$, $\gamma_{gl,k}^{(1)} = 1$ and $\gamma_{jgl,k}^{(0)} = 1$.

3.2.3 Dirichlet process mixture model (SOGC$_{dp}$ and MOGC$_{dp}$)

Determine the number of clusters K is often challenging, especially when the prior knowledge is limited (Tibshirani et al., 2001). To avoid arbitrarily determining K and allow for more flexibility, we further extend SOGC and MOGC to Dirichlet process mixture model (DPM). Denote all the cluster-specific parameters as ϕ_k and the sample realization as $\theta_i = \phi_{z_i}$, the formal definition of SOGC and MOGC Dirichlet process mixture models ($SOGC_{dp}$ and $MOGC_{dp}$) can be expressed as

$$
\begin{aligned}
Y_i|\theta_i &\sim F(\theta_i), \\
\theta_i|G &\sim G, \\
G &\sim DP(G_0, \alpha),
\end{aligned}
$$

where $F(\theta_i)$ is the multi-variate normal likelihood defined in SOGC and MOGC, G is a random probability measure that is from a Dirichlet process (DP) with a base distribution G_0 and a concentration parameter α. Here, the base distribution G_0 is the joint prior of all cluster-specific parameters ϕ_k. Another constructive representation of DP is the stick-breaking representation (Sethuraman, 1994):

$$\theta_i \sim G = \sum_{k=1}^{\infty} p_k \delta_{\phi_k},$$
$$p_1 = v_1, \quad p_k = (1-v_1)(1-v_2)\dots(1-v_{k-1})v_k, \quad \text{for } k = 2, \cdots, \quad v_k \sim Beta(1, \alpha).$$

We utilize the blocked Gibbs sampler algorithm proposed in Ishwaran and James (2002), approximating a DPM by a truncation of finite mixtures with an upper bound K_{max}:

$$G = \sum_{k=1}^{Kmax} p_k \delta_{\phi_k},$$
$$p_1 = v_1, \quad p_k = (1-v_1)(1-v_2)\dots(1-v_{k-1})v_k, \quad \text{for } k = 2, \dots, K_{max}-1,$$
$$v_k \sim Beta(1, \alpha), \quad v_{K_{max}} = 1.$$

3.2.4 Implementation and evaluation to compare with other existing models

We compared SOGC, SOGC$_{dp}$, MOGC and MOGC$_{dp}$ with two existing methods – SPKM and ISKM. The tuning parameters in SPKM and ISKM were both selected using default setting in the original package. For ISKM, we implemented both $\alpha = 0.5$ (equal weight on the lasso and the overlapping group lasso penalties, see details in Section 1.2.2) and $\alpha = 0.01$ (99% weight on the overlapping group lasso penalty). In all simulations, we set $K_{max} = 10$ in SOGC$_{dp}$ and MOGC$_{dp}$, but reduced K_{max} to 5 in the real data application for an easier interpretation of the clusters identified. Random initial values were used in SPKM, and the cluster labels obtained from SPKM were used as the initial values for all other methods.

In the end, adjusted Rand index (ARI) were used to quantify the clustering performance, and area under the receiver operating characteristic curve (AUC) were used to measure the feature selection performance. For SPKM and ISKM, all the features were prioritized based on the feature weights. For SOGC, SOGC$_{dp}$, MOGC, and MOGC$_{dp}$, features were ordered by the posterior selection probability $Pr(\eta_j = 1|X)$, where $\eta_j = 1$ if $\mu_{j,k} = 1$ for any $k \in \{1, \dots, K\}$.

3.3 SIMULATIONS

3.3.1 Simulation I: Single-layer non-overlapping groups

We first simulated data with single-layer non-overlapping groups to evaluate the performance of SOGC and SOGC$_{dp}$. We generated three clusters, each with 40 subjects, and 240 features grouped to eight level-1 groups. $U^{(1)}$ had block diagonal structure so that all eight level-1 groups were disjoint, each with 30 level-0 features. Among all level-1 groups, only two groups were subtype-relevant, which means they contained subtype-relevant features. Each subtype-relevant group contained 50% of all level-0 features as strong subtype-relevant features with cluster-specific means $\mu_{j,k}$ as E, 0 and $-E$ in cluster 1, 2 and 3, respectively. 30% of the features were weak subtype-relevant features with cluster-specific means as Et, 0 and $-Et$, where $t \in [0, 1]$. The remaining 20% of the features were subtype-irrelevant. All the features were generated as $x_{i,j}|z_i = k \sim \mathcal{N}(\mu_{j,k}, 1)$.

We repeated 50 simulations and compared the performance of SOGC, SOGC$_{dp}$, SPKM, ISKM with $\alpha = 0.5$ (50% weight on overlapping group lasso penalty) and ISKM with $\alpha = 0.01$ (99% weight on overlapping group lasso penalty). All the methods, except for SOGC$_{dp}$, assumed the underlying number of clusters $K = 3$ was known. We set $K_{max} = 10$ for $SOGCC_{dp}$. See more details about the implementation and evaluation in Section 3.2.4. Clustering and feature selection results from 50 repeats are shown in Figure 6 (A).

As shown in Figure 6 (A), when the effect size was small, SOGC and SOGC$_{dp}$ outperformed SPKM and ISKM in both clustering and feature selection. Since 80% of the features inside the subtype-relevant groups were subtype-relevant, ISKM with larger weight ($\alpha = 0.01$) on group penalty outperformed that with smaller weight ($\alpha = 0.5$). In fact, we suspect that the poor performance of ISKM with $\alpha = 0.5$ might be due to local optimal trapping.

3.3.2 Simulation II: Single-layer overlapping groups

We next simulated data with single-layer overlapping groups. The setting was exactly the same as in simulation I in Section 3.3.1, except that $U^{(1)}_{1,1} = U^{(1)}_{1,2} = 1$. In other words, we set level-0 feature 1 to belong to both level-1 groups 1 and 2.

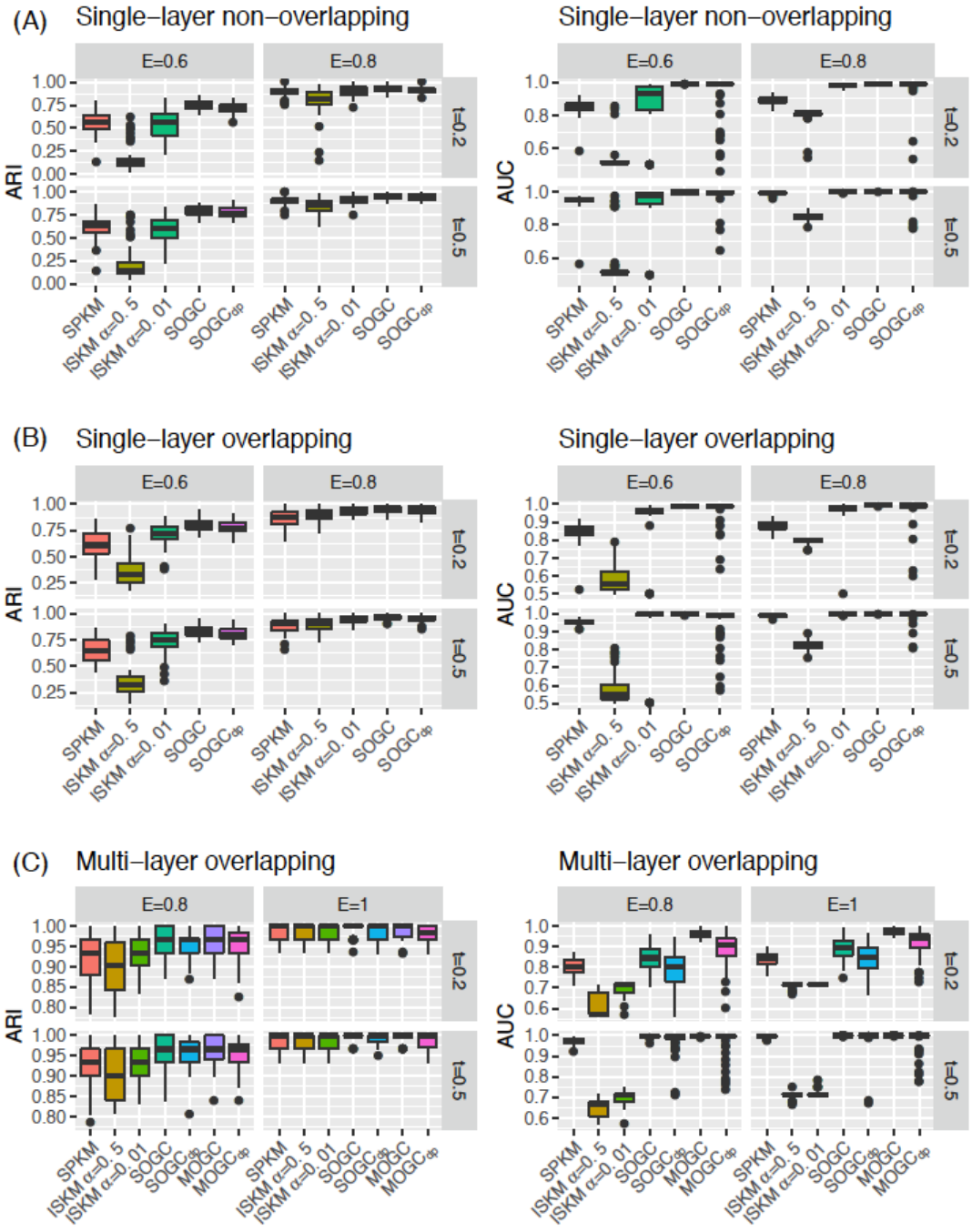


Figure 6: Results from 50 replicates

(A) simulation I, (B) simulation II and (C) simulation III. E is the effect size of strong subtype-relevant features; $E * t$ is the effect size of weak subtype-relevant features.

We also repeated 50 simulations for the same comparisons with the results showed in Figure 6 (B) . Similar to simulation I, SOGC and SOGC$_{dp}$ were demonstrated to have the best performance in both clustering and feature selection, despite the groups overlapping.

3.3.3 Simulation III: Two-layer overlapping groups

In this simulation, we generated 90 subjects evenly distributed in three clusters and 120 features with two-layer overlapping group structure. 120 features were firstly grouped into 40 non-overlapping level-1 groups, each of which contained 3 features. 40 level-1 groups were further assigned to 4 level-2 groups, each of which had 10 level-1 groups, except that level-2 group 2 had 11 level-1 groups. Besides, level-2 groups 1 and 2 shared one level-1 group. Group membership matrices is shown in Figure 7.

Among four level-2 groups, two of them were subtype-relevant, including four level-1 groups that were strongly predictive of subtypes, four level-1 groups that were weakly predictive of subtypes, and two level-1 groups that were not predictive of subtypes. For each of the level-1 groups that were strongly predictive of subtypes, all three level-0 features included had the cluster-specific means as E, 0 and $-E$ for cluster 1, 2 and 3, respectively. For each of the level-1 groups that were weakly predictive of subtypes, all the three level-0 features included had the corresponding cluster-specific means as Et, 0 and $-Et$, where $t \in [0, 1]$. For features shared by more than one group, we first generated partial mean $\mu_{j,k}$, and then defined the marginal mean as $\mu_{j,k} = \sum_{g=1}^{m_1} \sum_{l=1}^{m_2} U_{jg}^{(1)} U_{gl}^{(2)} \mu_{jgl,k}$. Given the cluster-specific means, features were generated the same as in simulation I.

We compared the performances of SPKM, ISKM ($\alpha = 0.01$ and 0.5), SOGC, SOGC$_{dp}$, MOGC and MOGC$_{dp}$. Only MOGC and MOGC$_{dp}$ utilized the level-2 group structure. All the methods, except for SOGC$_{dp}$ and MOGC$_{dp}$, assumed the underlying number of clusters $K = 3$ was known. We set $K_{max} = 10$ in SOGC$_{dp}$ and MOGC$_{dp}$. See more details in Section 3.2.4.

Figure 6 (C) shows the clustering and feature selection results from 50 replicates. All the methods had similar good clustering performance, but when the effect size was

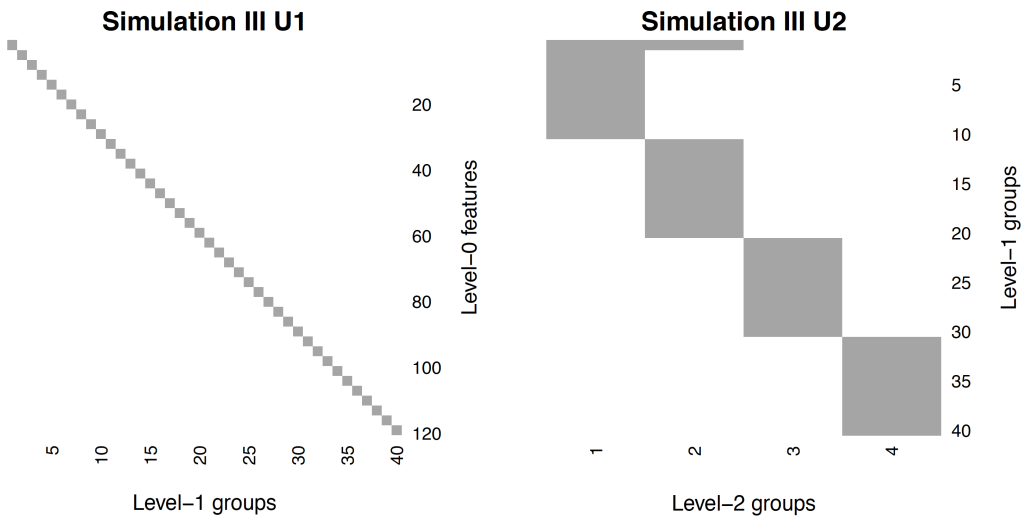


Figure 7: Membership matrices in simulation III

relatively small, SOGC, SOGC$_{dp}$, MOGC and MOGC$_{dp}$ outperformed SPKM and ISKM. The advantages of incorporating level-2 groups were more obviously demonstrated by the better feature selection performance of MOGC and MOGC$_{dp}$.

3.4 APPLICATIONS

3.4.1 Leukemia transcriptomic datasets using pathway database as prior knowledge

In the first application, we evaluated the performance of SOGC and SOGC$_{dp}$ given a single-layer overlapping group structure in a real data set. We adopted a same data set used in Huo and Tseng (2017) — a transcriptomcis data set of acute myeloid leukemia (AML) patients (Verhaak et al., 2009) with pathway databases to define the level-1 groups. We only kept 89 samples in one of the three fusion gene subtypes: inv(16) (inversions at chromosome 16), t(15;17) (translocations between chromosome 15 and 17) and t(8,21) (translocations between chromosome 8 and 21), which were previously shown to have distinct survival outcomes and

treatment responses. These three subtypes were also treated as the true cluster labels to evaluate clustering performance.

The original data downloaded from NCBI GEO website contained 54,613 probes. For multiple probes that were mapped to the same gene symbol, we selected the probe with the least interquartile range (IQR). After this step, we ended up with 20,154 unique genes and 89 samples. We then considered three pathway databases: Biocarta (217 pathways), KEGG (186 pathways) and Reactome (674 pathways), all of which were downloaded from MSigDB http://software.broadinstitute.org/gsea/msigdb. Since KEGG and Reactome contained more and larger pathways than Biocarta, we performed different filtering steps for each of the pathway databases to achieve similar number of features for input. For the Bocarta pathway database, we first filtered out genes which had an average gene expression below 30% or were not included in any of the Biocarta pathways. Then, we only kept pathways with 15-100 genes, and removed genes that were not included in the remaining pathways. For the KEGG and Reactome pathway databases, we filtered out genes with an average gene expression below 50% or which were not included in the pathway database, and only kept pathways with 50-100 genes. The final data dimensions are listed in Table 6.

Comparing the ARIs in Table 6, all the methods had similar clustering performance, except for SOGC$_{dp}$. This is expected given that SOGC$_{dp}$ did not assume $K = 3$ was known. To compare the feature selection performance, we selected the top 500 features for the Biocarta data base and top 1,000 features for the KEGG and Rectome databases of each method (see the details about feature sorting in Section 3.2.4) and performed Fisher's exact test for pathway enrichment analysis. SOGC and $SOGC_{dp}$ had more pathways significantly enriched with $p < 0.05$ as shown in Table 6. To better compare the significance of enriched pathways, we plotted $\log_{10}$(p) from Fisher's exact test in Figure 8. SOGC and SOGC$_{dp}$ clearly had more pathways with very small p-values.

3.4.2 Integrating TCGA Breast cancer mRNA, CNV and methylation

We next applied our models to a multi-omics data set of breast cancer patient for a setting of multi-layer overlapping group structure. Gene expression data (IlluminaHiSeq RNAseqV2,

20,531 genes), copy number variation (CNV) data (BI gistic2, 24,776 genes) and methylation data (Methylation450, 485,577 probes) of 770 breast cancer patients were downloaded from TCGA NIH official website. We followed the same pre-processing steps as in Huo and Tseng (2017), firstly removing features with any missing values, then transforming the gene expression FPKM to $log_2(\cdot + 1)$ and converting methylation β value to the M value ($M = \log_2(\beta/(1-\beta))$. For methylation data, if multiple probes were matched to the same gene symbol, we selected one probe with the largest average correlation with other probes matched to the same gene symbol. After pre-processing, there were 20,147 genes with expression data, 24,776 genes with CNV data, and 20,531 genes with methylation data.

To reduce the data dimension, we filtered out 70% genes with lower average expression, then another 70% genes with lower variance, and finally the genes that were not included in the KEGG pathways. In the end, there were 2507 features (846 genes with expression data, 825 genes with CNV data, and 836 genes with methylation data) mapped to 846 genes and 285 pathways.

To evaluate clustering performance, we calculated the ARI of the cluster labels identified by each method with the labels of well established intrinsic subtypes — Luminal/Normal (Luminal A and B, and Normal-like), Basal-like, and Her2-enriched (Sørlie et al., 2003). We then compared the feature selection performance by selecting the top 1000 features of each method for a pathway enrichment analysis.

As shown in Table 7, all the methods except for SOGC_{dp} and MOGC_{dp} had similar clustering performance. However, incorporating the pathway as the level-2 groups, MOGC and MOGC_{dp} identified much more significantly enriched pathways with $p < 0.05$ than other methods. SOGC_{dp}, although did not consider pathway structures, also led to more enriched pathways, which we suspected was due to the fact that it identified more clusters.

3.5 CONCLUSION AND DISCUSSION

Structured feature selection is extensively studied in the regression setting, nevertheless its extension in the clustering setting is less explored. This chapter extends the indicator variable

selection prior proposed in Chapter 2 into Gaussian mixture models to incorporate single- and multi-layer overlapping group structures for clustering. In simulations, both SOGC (SOGC_{dp}) and MOGC (MOGC_{dp}) can better select features that were weakly predictive of subtypes by borrowing the information from groups, and further improve clustering. In the applications, although incorporating extra group information did not improve the clustering performance, it better selected the subtype-predictive groups.

Table 6: Results of clustering the transcriptomics data set of leukemia patients integrating pathway database

	Pathway database	Biocarta	KEGG	Reactome
Input dimension	No. genes	798	1352	1977
	No. pathways	121	41	82
ARI	SPKM	0.87	0.89	0.86
	ISKM ($\alpha = 0.5$)	0.86	0.86	0.46
	ISKM ($\alpha = 0.01$)	0.86	0.86	0.86
	SOGC	0.86	0.86	0.86
	SOGC_{dp}	0.75	0.78	0.77
Number of pathways enriched ($p < 0.05$)	SPKM	14	8	13
	ISKM ($\alpha = 0.5$)	15	11	4
	ISKM ($\alpha = 0.01$)	7	5	16
	SOGC	88	31	25
	SOGC_{dp}	69	33	29

ARI were calculated comparing clustering labels with three fusion gene subtypes. Enriched pathways were defined as the pathways with Fisher's exact test $p < 0.05$

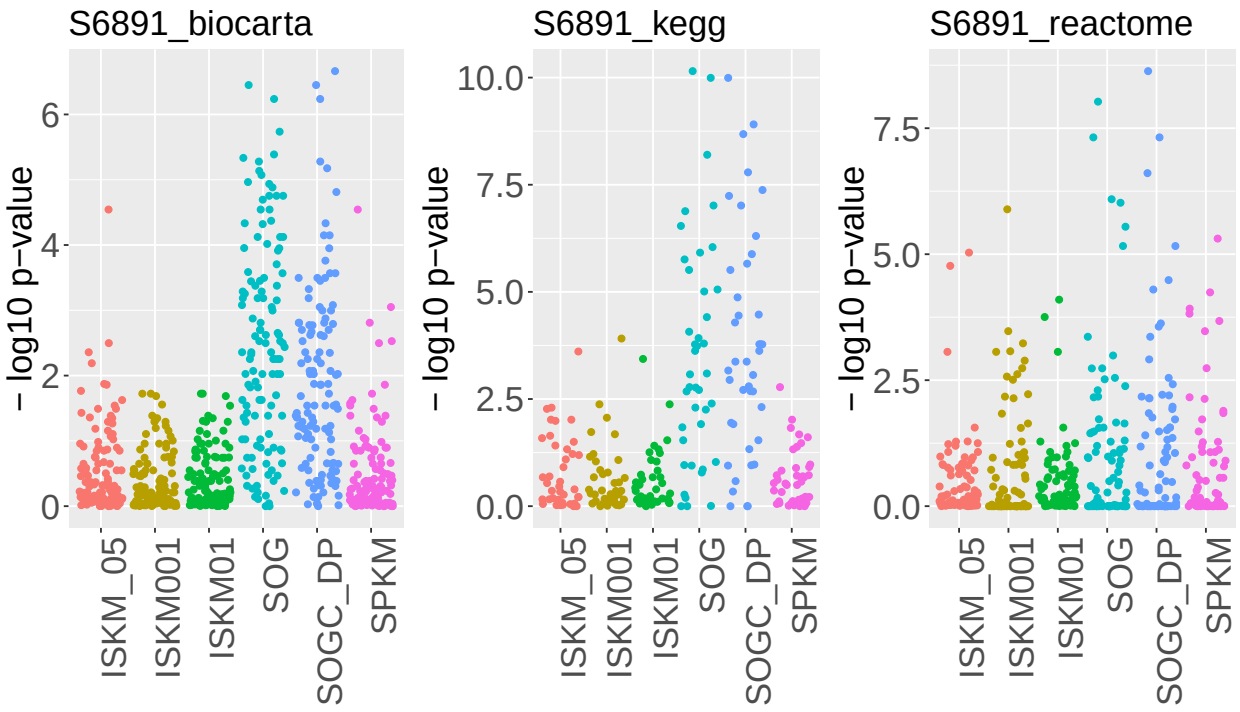


Figure 8: Pathway enrichment p-values from clustering the transcriptomics data of leukemia patients.

Fisher's exact test was performed on the top 500 (Biocarta) or 1,000 (KEGG and Reactome) features from each method.

Table 7: Clustering and feature selection results of TCGA Breast cancer application

Method	ARI	No. pathways $p < 0.05$
SPKM	0.34	174
ISKM ($\alpha = 0.5$)	0.34	168
ISKM ($\alpha = 0.01$)	0.34	165
SOGC	0.33	193
SOGC_{dp}	0.16	203
MOGC	0.33	203
MOGC_{dp}	0.16	204

4.0 METADCN: META-ANALYSIS FRAMEWORK FOR DIFFERENTIAL CO-EXPRESSION NETWORK DETECTION WITH AN APPLICATION IN BREAST CANCER

The contents of this Chapter is published in the journal *Bioinformatics* with three joint first authors (Zhu et al., 2016). Ying Ding initialized the project, Cho-Yi Chen revised the algorithm and developed the first version of the visualization tool MetaDCNExplorer. Li Zhu further revised the algorithm, performed all the simulations and applications, and drafted the manuscript.

4.1 INTRODUCTION

Differential co-expression (DC) refers to the change in gene-gene correlations between two conditions (e.g., cases and controls). Changes in gene-gene correlation may occur in the absence of differential expression, meaning that a gene may undergo radical changes in regulatory patterns that would be undetected by traditional differential expression (DE) analyses (see Figure 9A). A specific phenotype could be contributed to by differential co-expression without altering the expression levels of genes. This phenomenon has been found in aging (Southworth et al., 2009) as well as in other biological conditions (Gaiteri et al., 2014). Disease-associated alterations in the regulatory systems that create co-expression changes may be revealed through comparing gene-gene correlations that are computed separately in control and disease populations. Therefore, DC analysis can provide complementary information to standard differential expression (DE) analyses. Differential co-expression in two conditions could shed light on novel biological mechanism. For

example, a group of genes may be regulated by a common transcription factor or epigenetic modification, which is active in one condition but disrupted in the other.

In the literature, Lai et al. (2004) proposed an expected conditional F-statistics to identify differential co-expressed gene pairs, while Amar et al. (2013) and Bhattacharyya and Bandyopadhyay (2013) developed methods for direct identification of DC gene modules. Choi and Kendziorski (2009) detected differential co-expression using predefined gene sets such as Gene Ontology (GO) categories. Although this approach incorporates prior biological information, it lacks the ability to detect novel DC modules. Another class of methods detected differential modules with genes highly co-expressed in one reference condition but with little or no correlation in the other condition. These types of methods rely on applying clustering methods to one reference condition, therefore the clusters are more related to one condition, causing case-control asymmetry in the analysis (Watson, 2006; Ihmels et al., 2005). To circumvent this problem, Zhang and Horvath (2005) identified co-expressed modules in the entire (cases and controls combined) cohort through clustering and then evaluated their differential co-expression across conditions. Similarly, Tesson et al. (2010) extended this framework to detect differential co-expression modules by introducing the correlation changes between conditions into a dissimilarity matrix for clustering (DiffCoEx).

All methods described above for DC network detection focused on single transcriptomic study analysis. A differential correlation relationship could arise from meaningful biological sources as well as uncorrected technical biases (see Figure 1 in Gaiteri et al. (2014)). Any mechanism that synchronously regulates transcription of multiple genes, unwanted batch effects, or mixture of tissues could potentially contribute to co-expression relationships. Therefore, instead of looking for DC networks between two conditions in a single study, differential co-expression may be confirmed across multiple datasets via meta-analyses to increase detection power and stability. DC networks that are significant in one dataset may become more convincing if the DC patterns are preserved across multiple datasets. DC between conditions can be assessed by different choices of measures; for example, differential modules with a predominant measure such as density (Li et al., 2011) or other sophisticated network measures (Kugler et al., 2011; Langfelder et al., 2011).

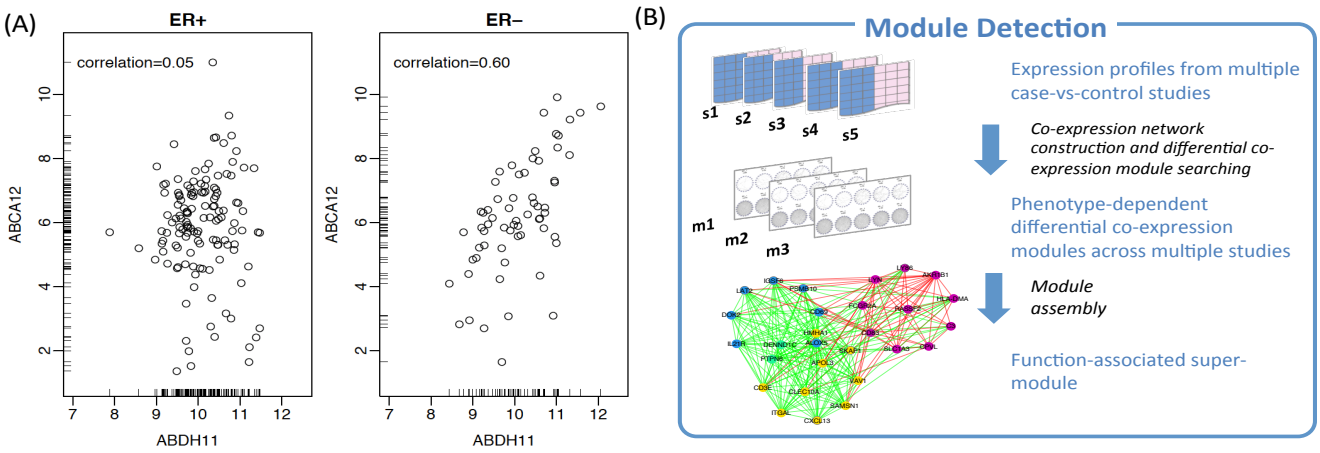


Figure 9: MetaDCN example and pipeline

(A) An example of differential co-expression between ER+ and ER- breast cancers. Each dot represents one sample. Strong co-expression between *ABCA12* and *ABHD11* can be observed in ER- tumors (right) but not in ER+ tumors (left). Samples are from GSE7390. (B) Diagram of procedures for basic module detection by energy function optimization and supermodule assembly via pathway enrichment criterion.

So far, few studies attempted to detect DC networks across multiple studies. Mehan et al. (2009) proposed a simulated-annealing-based method to detect DC modules of which the network density changes were associated with phenotype. However, their method embedded pathway enrichment in the optimization of objective function; that is, the optimization phase heavily depended on the prior knowledge and also the output module sizes from the method were generally small. In this chapter, we have developed a new meta-analytic framework, namely MetaDCN, to search for initial DC modules without prior information. Our method included additional network properties in the energy function to detect biologically meaningful "basic DC modules" and the false discovery rate (FDR) was controlled by permutation analysis. We then further combined basic DC modules that share common pathway annotation into more interpretable DC supermodules. We evaluated the method on simulated data and breast cancer studies to search for differential co-expression network (DCN) between ER+ vs. ER- and invasive lobular carcinoma (ILC) vs. invasive ductal carcinoma (IDC). The identified DCNs were further validated in independent breast cancer studies. The result identified pathways such as ER-mediated immune functions and extracellular matrix heterogeneity between ILC and IDC that help elucidate the underlying disease mechanisms.

4.2 METHODS

MetaDCN combines multiple case-control transcriptomic studies to detect disease-associated modules such that genes in the modules are highly correlated in control samples but the correlations are disrupted in cases or vice versa. An energy function is introduced to detect modules of DC networks across studies. Since direct optimization for large modules is computationally challenging and unstable, we first aim for detecting a sequence of small "basic DC modules" of sizes between 3 and 30. Basic DC modules are then combined into DC supermodules via a module assembly algorithm based on pathway enrichment information (see Figure 9B). Such pathway-centric assembly improves functional annotation of detected supermodules that can advance disease understanding and guide further hypothesis generation.

4.2.1 Basic DC module detection

The algorithm to detect basic DC modules is outlined below. Details of energy function, optimization procedure and false discovery rate control are described.

Energy function

Consider N transcriptomic studies, each containing case and control samples. Gene co-expression networks are first constructed among cases and among controls for each of the N studies, thus generating $2N$ co-expression networks. In this chapter, we demonstrate our method based on unweighted networks but the method can be extended to weighted networks. To build unweighted networks and normalize them across different studies, we first calculated pair-wise gene-gene Spearman's correlations for robust comparisons. In contrast to Pearson correlation, Spearman's correlation can capture both linear and non-linear association. Considering the large number of possible edges and computation complexity, we then select the correlation cut-off for edge connections so that only the top 0.4% of possible connections in each network were kept (Lee et al., 2003). Assuming the total number of genes is p, the number of all possible connections is $\binom{p}{2}$, therefore this cutoff is not very stringent. This procedure provides robustness because different studies usually have different sample sizes and are conducted using different experimental platforms, which could result in distinct correlation distributions. Our proposed algorithm is developed for the more popular unweighted network but it can be modified for weighted network if desired.

We propose to minimize the following energy function (target function) for detection of gene modules with differential co-expression:

$$E_{\text{tot}} = w_1 E_{\text{diff\_mean}} + w_2 E_{\text{size}} + w_3 E_{\text{diff\_var}}$$

The proposed target function comprises the following three components: I) $E_{\text{diff\_mean}}$ for mean network density difference between two phenotypes across N studies, II) E_{size} for size of module, and III) $E_{\text{diff\_var}}$ for the consistency of the density difference between the two phenotypes across N studies. Gene modules minimizing E_{tot} have consistently large correlation differences between cases and controls across multiple studies, as well as reasonable large size. The search direction is bidirectional, meaning that we will identify

modules with significantly higher connections in case networks than in control networks and then repeat reversely.

Each component in the target function is described by an exponential decay function. The first component is defined as $E_{\text{diff\_mean}} = \exp\left\{-\alpha_1\left(\frac{\sum_{i=1}^{N}(\delta_{i,\text{cases}}-\delta_{i,\text{controls}})}{N}\right)\right\}$, where $\delta_{i,\text{cases}}$ and $\delta_{i,\text{controls}}$ are the densities of case network and control network respectively in study i. The exponential decay function favors larger mean density differences between cases and controls and is the major target of our algorithm. The second component $E_{\text{size}} = \exp\{-\alpha_2(|x|/\gamma)\}$ (where x denotes the genes in the module, $|x|$ is the module size) is related to the size of the modules which favors larger modules and penalizes smaller modules. We restricted the module size no larger than 30 due to large searching space. We set $\gamma = 30$ to rescale the ratio ranging from 0 to 1 to make the three decay parameters (α_1, α_2 and α_3) comparable in later parameter selection. Without E_{size}, dimers or triplets with density 1 or 0 could easily dominate the output by random chance and increase false positives. The third component $E_{\text{diff\_var}} = \exp\left\{-\alpha_3\left(1-\sqrt{\frac{\sum_i(\delta_{i,\text{cases}}-\delta_{i,\text{controls}})^2}{N}}\right)\right\}$ quantifies the variance of the paired difference of network densities between cases and controls across studies to favor consistent differential co-expression among studies. In all three components, the parameters (α_1, α_2 and α_3) control the decay rate in the exponential function. In our implementation, we set $\alpha_2 = 10$ for E_{size} and $\alpha_1 = \alpha_3 = 5$ for $E_{\text{diff\_mean}}$ and $E_{\text{diff\_var}}$. The higher α_2 was used for E_{size} to avoid extremely small modules as previously mentioned.

To tune the parameters w_1, w_2, and w_3 in the target function, we first constrain the sum of the three parameters to be 1000, i.e. $w_1 + w_2 + w_3 = 1000$. We assigned equal importance to $E_{\text{diff\_mean}}$ and $E_{\text{diff\_var}}$ by setting $w_1 = w_3$, and searched for optimal w_2 from 100 to 700 with 100 increments that could output the largest number of basic DC modules under FDR 0.3 (see below for detection of basic DC modules and FDR control).

Optimization by simulated annealing

Due to the non-convex nature of E_{tot}, we applied simulated annealing, a stochastic algorithm for non-convex optimization (Kirkpatrick et al., 1983). In each Monte Carlo (MC) step with simulated annealing, a new state is proposed and denoted as X_{new}, which is either adding a node (gene) from trial set to selected set or removing a node (gene) from selected set to trial set. At the beginning, the trial set is determined as the set

of genes that have at least one edge connected to the seed module genes (initial selected set) in any of the N case co-expression networks. If the resulting energy is smaller, the state is accepted. If not, the state is accepted with an acceptance probability as $P_{\text{acc}} = \min\left(1, \frac{\pi(x_{\text{new}})p(x_{\text{new}\to\text{old}})}{\pi(x_{\text{old}})p(x_{\text{old}\to\text{new}})}\right)$, where P_{acc} is the acceptance probability and $p(x_{\text{old}\to\text{new}})$ is the transition probability from old state to the new state. If a genes is added from trial set to selected set, $p(x_{\text{old}\to\text{new}})/p(x_{\text{new}\to\text{old}}) = |\text{trial set}|/|\text{selected set}|$; if a genes is removed from selected set to trial set, $p(x_{\text{old}\to\text{new}})/p(x_{\text{new}\to\text{old}}) = |\text{selected set}|/|\text{trial set}|$, where $|x|$ denotes the size of set x. $\pi(x_{\text{new}})$ is the Boltzmann distribution of the energy function to be minimized: $\pi(x_{\text{new}}) = \exp-\{-E_{\text{tot}}(x_{\text{new}})/T\}$, where T is a temperature parameter. When temperature is high, new trial moves will be accepted easily and thus more freely jump out of the local minimum. When temperature gets lower, it tends to converge to a local minimum. We apply the temperature schedule $T_{(k+1)} = 0.95 \cdot T_k$ and stop the annealing run if the acceptance ratio is smaller than 2%, where the acceptance ratio is calculated as the ratio of steps accepted in every 400 MC steps. Due to large searching space, we bounded the module size between 3 and 30. If current module size is 3, only addition of new node is allowed for a new state while if module size is 30, only node removal is allowed.

Although simulated annealing helps improve local minimum trapping, a good starting point, which is called seed module here, is critical for optimization in high dimensional space. Instead of randomly selecting a subset of genes from the genome to be the seed modules, an edge-study matrix of Spearman's correlations was constructed where rows represent all possible edges and columns represent all studies in two conditions of size $2N$ (Walley et al., 2012). For each edge on the rows, a simple paired t-test is applied to the Spearman's correlations to assess candidate differential co-expression edges (require paired t-test p-value < 0.1 and absolute mean difference of Spearman's correlation > 0.1). Based on these candidate differential co-expression edges, an initial network is constructed and multiple (denoted as K) connected graphs in the network are identified. If the size of a connected graph is larger than 30, we randomly sample 10 genes from it as the initial seed module for optimization starting points; if the size is smaller than 3, we discard it. Otherwise, the optimization starts from the connected graphs as the seed module directly. In our evaluation for simulations and application, such an algorithm to generate seed modules has performed

well. But it is possible to apply other community detection algorithms for this purpose (Fortunato, 2010).

Although simulated annealing helps improve local optimum problem, optimization instability still exists. We will repeat the optimization by starting from K initial seed modules and repeat R times of simulated annealing repeats. For two repeats with Jaccard index greater than 0.8, we will select the one with smaller E_{tot}. This will generate $\sum_{k=1}^{K} R_k$ basic differential co-expression (DC) modules for supermodule assembly, where R_k is the number of basic modules from the kth seed modules with pairwise Jaccard index smaller than 0.8.

Control of false discovery rate

To avoid detection of spurious modules by chance, the false discovery rate is assessed for detected $\sum_{k=1}^{K} R_k$ basic DC modules as described below. Denote by E_{kj} the optimized energy value for detected basic DC module u_{kj} from the k-th seed module and j-th simulated annealing repeat, where $1 \leq k \leq K$ and $1 \leq j \leq R_k$. We first permute the case-control class labels for samples in each study and then reconstruct the case and control co-expression network as described previously. Simulated annealing optimization is similarly applied to detect $\sum_{k=1}^{K} R'_k$ "null" basic DC modules, where R'_k is the number of basic modules dected from permuted network with pairwise Jaccard index smaller than 0.8. Suppose the permutation is repeated for B times and the resulting energy values are denoted as $E_{k,j}^{(b)}$ where $1 \leq b \leq B$, $1 \leq k \leq K^{(b)}$, $1 \leq j \leq R_k^{'(b)}$. Under null hypothesis, the resulting case and control co-expression networks from permutation have no difference and $E_{k,j}^{(b)}$ will form a null distribution to assess p-values of E_{kj}. The p-values of basic DC modules u_{kj} are estimated as $p(u_{kj}) = \frac{\sum_{b=1}^{B}\sum_{k=1}^{K^{(b)}}\sum_{j=1}^{R_k^{'(b)}} I\{E_{kj}^{(b)} \leq E_{kj}\}+1}{\sum_{b=1}^{B}\sum_{k=1}^{K^{(b)}} R_k^{'(b)}+1}$. A pseudo count of 1 is added to both the denominator and the numerator to avoid zero p-values (Phipson and Smyth, 2010). FDR is controlled by Benjamini-Hochberg correction to account for multiple comparisons.

4.2.2 Supermodule assembly, summarization and visualization

DC supermodule assembly

Since the current approach limits the size of the basic DC modules between 3 and 30, small modules often do not yield significant pathway enrichment annotation to inspire further

hypothesis generation. Therefore, in order to obtain larger DC modules, we proposed to use statistical significance of pathway enrichment to guide module assembly. Firstly, we applied pathway enrichment analysis using Fisher's exact test on detected basic DC modules (here we choose FDR $\leq$ 0.3) against 2,379 pathways downloaded from MSigDB http://www.broadinstitute.org/gsea/msigdb/, which contained Biocarta, KEGG, Reactome and Gene Ontology databases (excluding large pathways with more than 250 genes). For each given pathway, we applied Fisher's meta-analysis method to combine p-values across basic DC modules and selected the top 150 pathways with the most significant meta-analyzed p-values. The restriction is not necessary, but will reduce the computation cost, without changing the results much. For each of the 150 candidate pathways, we searched among combinations of up to three basic DC modules (including both over-connected and under-connected DC modules of case-control comparison) and identify the assembled supermodule such that its pathway enrichment p-value is minimized. The reason to search for up to three basic DC modules is to ensure reasonable supermodule size (3-90), and meanwhile reduce computation burden. Take the immune response pathway in Figure 11C as an example, the pathway enrichment p-values for modules H9, L1 and L2 (H stands for modules with higher density in ER+ patients; L stands for modules with lower density in ER+ patients) are 0.018, 7×10^{-4} and 0.02 with module sizes 10, 10 and 11, respectively. The supermodule combining these three basic DC modules contains 28 genes with Fisher's exact test p-value = 1.33×10^{-6}. Assembly of multiple basic DC modules can yield larger supermodules with more genes involved in a specific pathway, which provides better biological interpretation and hypothesis generation. Additionally, if the assembled supermodule contains both over-connected and under-connected basic DC modules (see red and green edges in Figure 11C), it may suggest an interesting alternative activation mechanism in the pathway related to disease development.

Summarization and visualization of DC supermodules

Visualization of basic DC modules across N studies can be easily done by displaying the $2N$ co-expression networks as shown in Figure 10(A-B). For DC supermodules, however, smarter design of visualization is needed. Figure 11(C-D) shows our proposed visualization

display for DC supermodules. On the left plot, three basic DC modules (gene nodes displayed by red, blue and yellow) are combined to form the DC supermodule. The edge widths between any pair of gene nodes i and j are controlled proportionally by a standardized score Z_{ij} to represent the degree of differential co-expression. Denote by $u_{ij}^{(s)}$ and $v_{ij}^{(s)}$ the Spearman's correlation between gene i and j in study s in case and control samples, respectively. Let $d_{ij}^{s} = u_{ij}^{(s)} - v_{ij}^{(s)}$ and let $\bar{d}_{ij}$ be the mean of paired correlation differences of all studies, σ_{ij} be the standard deviation of paired correlation differences and σ_0 be the fudge parameter estimated by the median of all standard deviations σ_{ij}'s. The edge widths are proportionally to the standardized score $Z_{ij} = \frac{\bar{d}_{ij}}{(\sigma_{ij}+\sigma_0)}$. The fudge parameter σ_0 is used to avoid accidentally large Z_{ij} due to small σ_{ij} (Tusher et al., 2001). As a result, the DC supermodule can be represented as a weighted undirected network. P-values of the Z scores were calculated by permuting case and control subjects in each study and randomly subsampling the same number of genes to calculate null permutated Z scores and comparing with them. Only edges with significant p-values passing a certain p-value threshold are displayed in the network plot.

We further developed a Cytoscape plug-in application, called "MetaDCNExplorer", which utilizes the power of the Cytoscape Java API in visualizing complex networks and integrating topology with attributes. The interface allows users to load the input supermodule attributes and generate interactive network visualization with additional context annotations. Firstly, the user selects a DC supermodule of interest to visualize from the list of biological pathways ranked by the significance of enrichment. The attributes of that supermodule will be loaded. The absolute values of the standardized Z scores for each gene pairs will be interpreted as edge widths, and the initial network view will be generated using edge-weighted force directed layout algorithm provided from prefuse toolkit (see Appendix B.1 for more details). In the network view, the edge width represents the edge weight. Nodes with their neighbors connected by high-weight edges will be automatically clustered together, so that the modular organization will be revealed. The edge color represents the direction of differential co-expression interpreted from edge Z scores, in which positive values (red color) indicate over-connected edges and negative values represent under-connected edges in case-control comparison. Node color represents the original basic DC modules where the gene

belongs, and the genes annotated under the selected biological pathway are highlighted with outer black circles. To account for the fact that different diseases might have different range of differential co-expression signals, we thus introduced additional factors that control the repelling and attracting force between and within the modules. These factors, together with the edge p-value cut-off threshold, are adjustable in a control panel so that users can update the network view in real time. In summary, this application is designed to reveal the modular organization of DC supermodules and to suggest alternatively activated sub-pathways that allow biologists to further explore and generate biological hypotheses on potential disease mechanisms.

4.2.3 Data sets

In this chapter, we applied MetaDCN to two breast cancer applications. In the first application, DC supermodules are detected for ER+ versus ER- comparison in five training studies and validated in three independent testing studies (see Table 8). The second application examines invasive lobular carcinoma (ILC) and invasive ductal carcinoma (IDC) comparison in two training studies (with both ILC and IDC samples) and partially validated in two testing studies, where only ILC subjects are available (see Table 9). Details of data description and data preprocessing are available in Appendix B.2.

4.3 RESULTS

4.3.1 Simulation

We first applied MetaDCN to a simulated dataset including 5 studies. Each study contained case and control groups, with number of subjects in each group drawn from $Poisson(50)$. We generated 1000 artificial genes named 1 to 1000 and a subset of them belonged to 5 gene modules (non-overlapping), each of which contains the number of genes $g_m \sim Poisson(20)$ $(1 \leq m \leq 5)$. Let $x_c^{(s)(m)}$ denote the vector of expression intensities of the g_m genes in the m-th module in group c in study s. We generated $x_c^{(s)(m)} \sim N(0, \Sigma_c^{(s)(m)})$, where $\Sigma_c^{(s)(m)} \sim$

Table 8: Description of breast cancer datasets for comparing ER+ vs. ER-

	Data sets	Study index	Sample size (ER+ vs. ER-)	Platform
Training	TCGA	S1	406(319 vs. 87)	RNA-Seq
	GSE7390	S2	198(134 vs. 64)	Affymetrix HG-U133A
	GSE2034	S3	286(209 vs. 77)	Affymetrix HG-U133A
	METABRIC	S4	1981(1512 vs. 469)	Illumina
	GSE4922	S5	245(211 vs. 34)	Affymetrix HG-U133A
Testing	GSE23720	S6	197(131 vs. 66)	Affymetrix HG-U133 Plus 2.0
	GSE58215	S7	270(218 vs. 52)	Agilent-028004
	GSE22220	S8	216(134 vs. 82)	Illumina humanRef8

Table 9: Description of breast cancer datasets for comparing ILC vs. IDC

	Data sets	Sample size (ILC vs. IDC)	Platform
Training	TCGA	470 (159 vs. 311)	RNA-Seq
	METABRIC	598 (65 vs. 533)	Illumina
Testing	Sotiriou	147 ILCs	Affymetrix HG-U133 Plus 2.0
	RATHER	111 ILCs	Agilent custom-designed platform

Inverse-Wishart$(60, (1-\rho_c^m)I + \rho_c^m J)$, $I_{g_m \times g_m}$ is the identity matrix, $J_{g_m \times g_m}$ is a matrix with all entries as 1, $c = 1$ for controls, $c = 2$ for cases, $s = 1, 2, \ldots, 5$, and $m = 1, 2, \ldots, 5$. We set different $(\rho_{c=1}^m, \rho_{c=2}^m)$ pairs for five modules to include both strong and weak signals. They were set to be (0.3, 0.1), (0.1, 0.3), (0.5, 0.1), (0.1, 0.5), and (0.7, 0.1) for 5 modules respectively. Therefore, the first and second modules have smaller signals, while the fifth module has the strongest signal. For genes outside the modules, the expressions were i.i.d. drawn from $N(0, 1)$.

With this simulated dataset, we constructed the edge-study matrix based on Spearman's correlation. The module search was performed using simulated annealing algorithm with a maximum of 500 iterations. $R = 3$ trials with different initial seed modules were repeated, and p-value was calculated using $B = 10$ permutations. In the end, the best module among 3 repeats was selected based on optimal p-value and energy. For simplicity, here we only evaluate performance of basic DC modules without module assembly. If the Jaccard index (ratio of the intersection set divided by the union set) of the identified basic DC module to the underlying truth is greater than 0.5, we denote this searching as a successful hit.

We generated 50 datasets and compared the performance of MetaDCN with an existing method DiffCoEx (Tesson et al., 2010). The implementation of DiffCoEx used the R code directly from the original paper with the default setting. The soft threshold in DiffCoEx, as the most sensitive tuning parameter, was chosen based on scale free topology fit (Zhang and Horvath, 2005). The hierarchical tree for clustering in DiffCoEx was cut using dynamicTreeCut R package (Langfelder et al., 2008).

Table 10 shows the lower and upper quartile of number of detected modules and the percentage of successful hits for each of the five modules under different FDR cut-offs for permutation test in the 50 repeated simulations. The result shows that DiffCoEx tends to detect many false positives while still missing the underlying true DC networks.

4.3.2 Breast cancer studies (ER+ vs. ER-)

We next applied our method to identify differentially co-expressed modules between networks from ER+ patients and networks from ER- patients. Estrogen receptor, indicating the

Table 10: MetaDCN simulation results

Method	FDR	Upper and lower quartile	M1 (%)	M2 (%)	M3 (%)	M4 (%)	M5 (%)
MetaDCN	0.1	(3, 5)	56	58	96	96	100
	0.2	(4, 5)	72	74	100	100	100
	0.3	(5, 5)	78	82	100	100	100
DiffCoEx	–	(3, 39)	8	8	30	26	37

Percentage of successful hits (Jaccard index>0.5) in simulation study (50 repeats). Upper and lower quantile indicates the upper and lower quartile of the number of detected modules in 50 repeats.

cancer cell response to hormone estrogen, is an important marker in breast cancer cases for treatment selection. Detecting differential co-expression network between ER+ and ER- patients can help us better understand the difference of disease mechanism, thus designing specific therapies for ER+/ER- patients. In the analysis of training data, five pairs of gene co-expression networks were constructed for ER+ patients and ER- patients across the five studies. Edge-study matrices were calculated and connected components were obtained as initial seed modules for simulated annealing algorithm. FDR was calculated for each of the modules with $B = 10$ permutations. The best weights were selected based on the results from first 3 repeats with different initial modules. With the optimal weights and $R = 10$ repeats, at FDR ≤ 0.3, 12 basic DC modules were detected as over-connected in ER+ networks while another 12 basic DC modules were detected as over-connected in ER- networks. Two example modules, one densely connected in ER+ networks and one densely connected in ER- networks, are illustrated in Figure 10. Both modules achieved FDR 0.02.

We tested varying number of repeats (R) in each seed module and the results of pathway-centric assembly are quite consistent. We identified 20 supermodules engaged in 40 pathways, sharing at least 3 overlapping genes with the enriched pathway. The top pathways associated with the assembled modules were listed in Table 11(A). Among the list of summarized

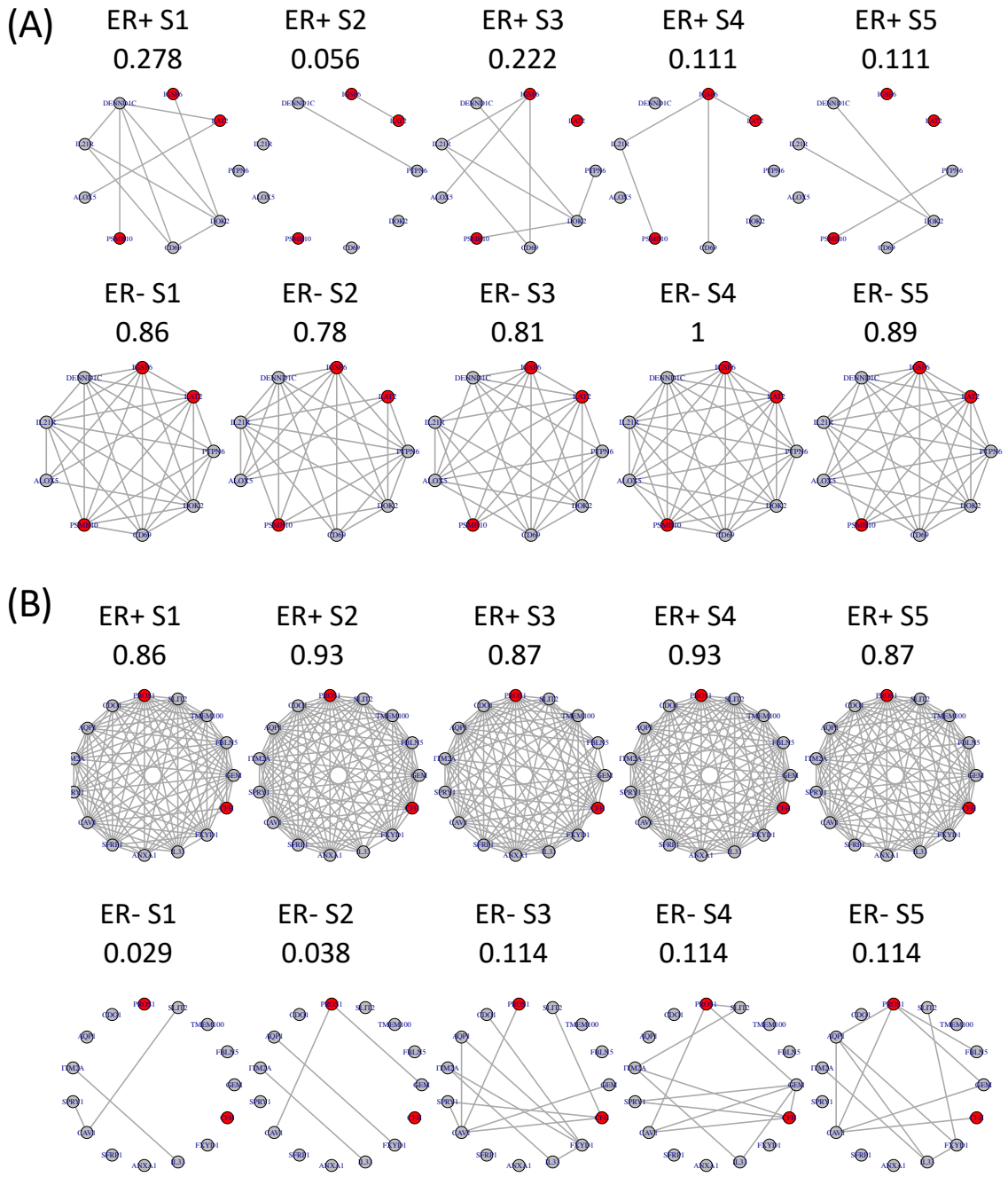


Figure 10: Examples of basic modules detected in ER+ vs. ER- comparison of five studies

(A) Example module (L2) more densely connected in ER- group with red nodes indicate genes belonging to the immune response pathway. (B) Example module (H7) more densely connected in ER+ group with red nodes indicate genes belonging to the complement cascade pathway. Nodes represent genes and links between them represents co-expression relationship. Each column corresponds to one independent study. The number above each module indicates the module density.

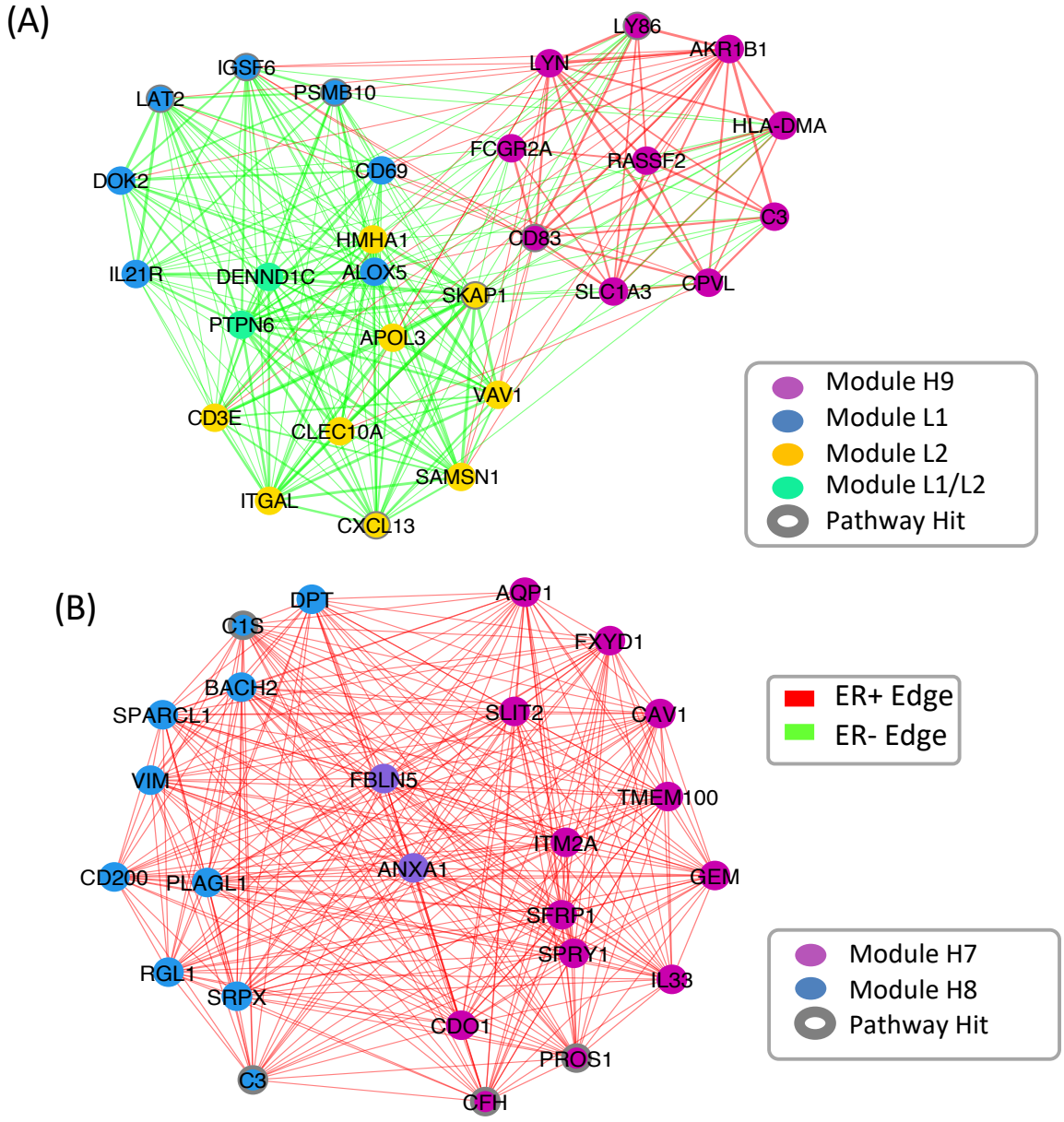


Figure 11: Examples of supermodules ensembled in ER+ vs. ER- comparison of five studies

(A) Visualization of immune response pathway supermodule. (B) Visualization of complement cascades pathway supermodules. The edge color represents the direction of differential gene co-expression, in which positive values (red color) represent ER-positive-favored co-expression and negative values (green color) show ER-negative-favored co-expression. Node color represents its origin of sub-modules, and the genes annotated in the immune response pathway are highlighted with dark circles. Edge width represents edge weight (Z score of differential co-expression).

DC supermodules, "complement cascade pathway" was with highest significance followed by "immune response pathway". Figure 11 showed the network view for these two DC supermodules.

In the literature, studies have shown that estrogen receptors can regulate innate immune cells (Kovats, 2015). Cunningham and Gilkeson (2011) found ERs have prominent effects on immune function in both the innate and adaptive immune responses. ERα expression is associated with outcome in patients with autoimmune diseases such as lupus. Possible alternative activations of immune and complement pathway between ER+ and ER- breast cancer patients have also been revealed in several research studies. Teschendorff et al. (2007) found that heterogeneity in clinical outcomes of ER- breast cancer patients is related to the complement and immune pathway, while this association is not observed in ER+ patients.

We next validated those two supermodules in leave-one-out cross-validation (LOOCV). Each time, we left one study out as testing set and used the remaining four studies as training set to perform module searching and module ensemble. In each LOOCV, 2 or 3 basic DC modules were merged into a DC supermodule in each pathway. We calculated network averaged densities in each basic DC module in the four training studies (on the left of dashed line) as well as the testing study (on the right of dashed line) in Figure 12(A) and Figure 13(A). Similarly, box-plots of Spearman correlation distributions are plotted in Figure 12(B) and Figure 13(B). The result consistently shows good validation of the finding.

Finally, we used the top two pathways and the DC supermodules obtained from five training studies and tested in the three independent validation studies. Same set of genes was used for constructing co-expression network. If genes were not available in a study with different platform, the overlapped gene set was used. Following Figures 12 (A-B) and Figure 13(A-B) for LOOCV, we plotted the average network densities and box-plots of Spearman correlation distribution in Figure 12(C-D) and Figure 13(C-D) for the basic DC modules of the supermodules enriched in those two pathways. The result provides consistent validation of the differential co-expression pattern of gene modules enriched in these pathways.

As a comparison, we also applied DiffCoEx (Tesson et al., 2010) to our datasets. Since DiffCoEx is only applicable to a single study, we applied it to the largest study METABRIC using the same procedure as described in the simulation section and evaluated the validation

Table 11: Top pathway-centric supermodules in ER+ vs ER- comparison of five studies

(A) Pathway name (ER+ vs. ER-)	Pathway size	Module size	# pathway genes	q-value	p-value	Module
Reactome complement cascade	32	25	4	2.14E-05	1.93E-07	H7,H8
GO immune response	235	28	7	5.63E-05	1.33E-06	H9,L1,L2
Reactome regulation of complement cascade	14	25	3	5.63E-05	2.47E-06	H7,H8
GO organ morphogenesis	144	35	6	5.63E-05	2.80E-06	H3,H5,L9
Biocarta tcytotoxic pathway	14	23	3	5.63E-05	2.85E-06	H3,L5
(B) Pathway name (ILC vs. IDC)	Pathway size	Module size	# pathway genes	q-value	p-value	Module
GO protease inhibitor activity	41	27	3	0.003	6.13E-05	L2,L4,L8
GO proteinaceous extracellular matrix	98	15	3	0.003	0.00085	L5,L7
GO extracellular matrix	100	15	3	0.003	0.00085	L5,L7

Top pathway-centric supermodules with at least 3 pathway overlapping genes (with 10 repeats with different initial modules). Module starts with H indicates it is more densely connected in ER+ or ILC network, while module starts with L indicates it is more densely connected in ER- or IDC network.

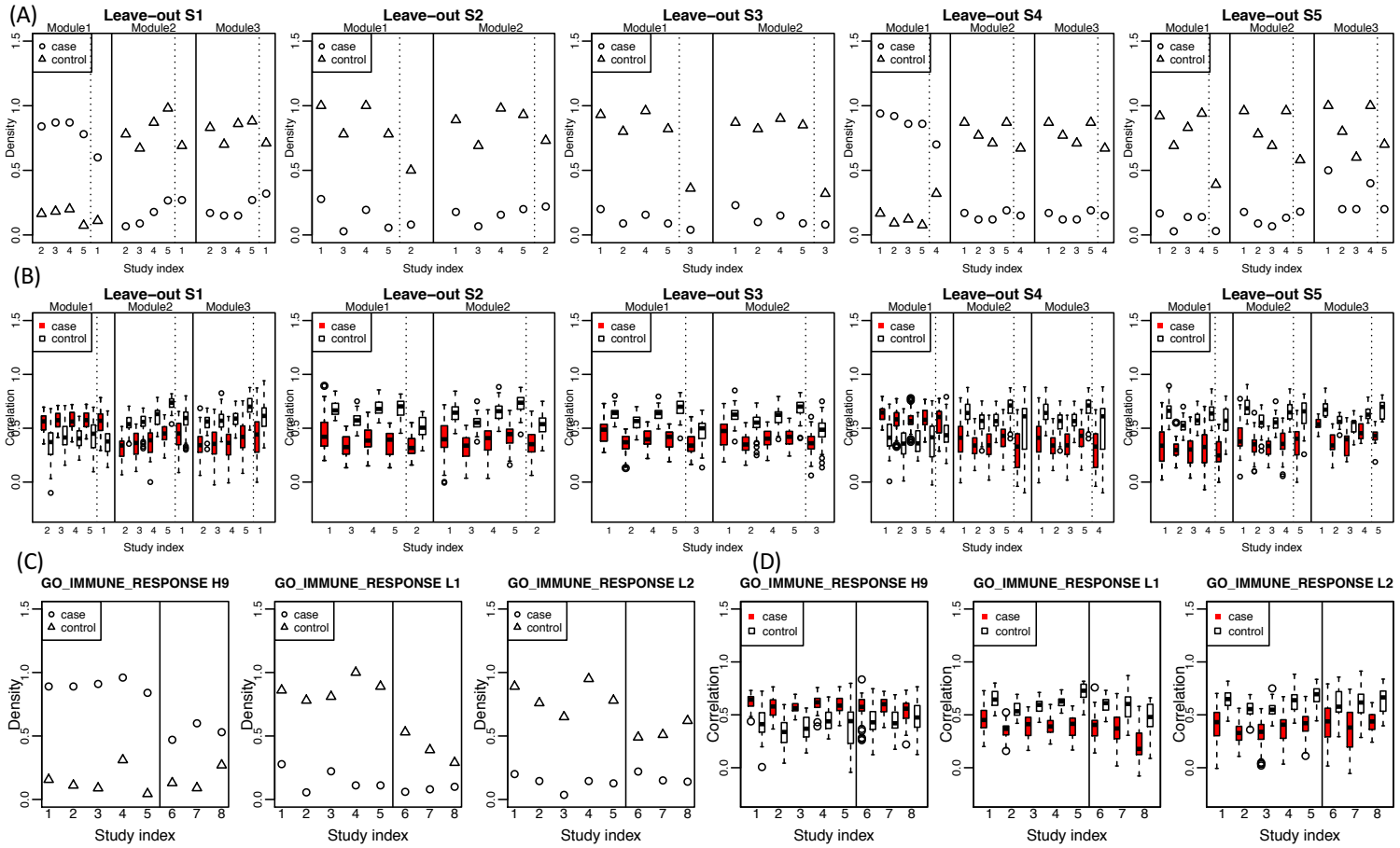


Figure 12: Validation of immune response pathway supermodules from ER+ vs ER- comparison

(A) Densities and (B) correlations of the basic modules assembled into immune response pathway supermodules in leave-one-out cross-validation. Solid lines separate modules, and dashed lines separate training set and testing set. (C) Module density and (D) correlations of genes in the basic modules enriched in immune response supermodule in independent validation studies. Solid lines separate training sets and testing sets.

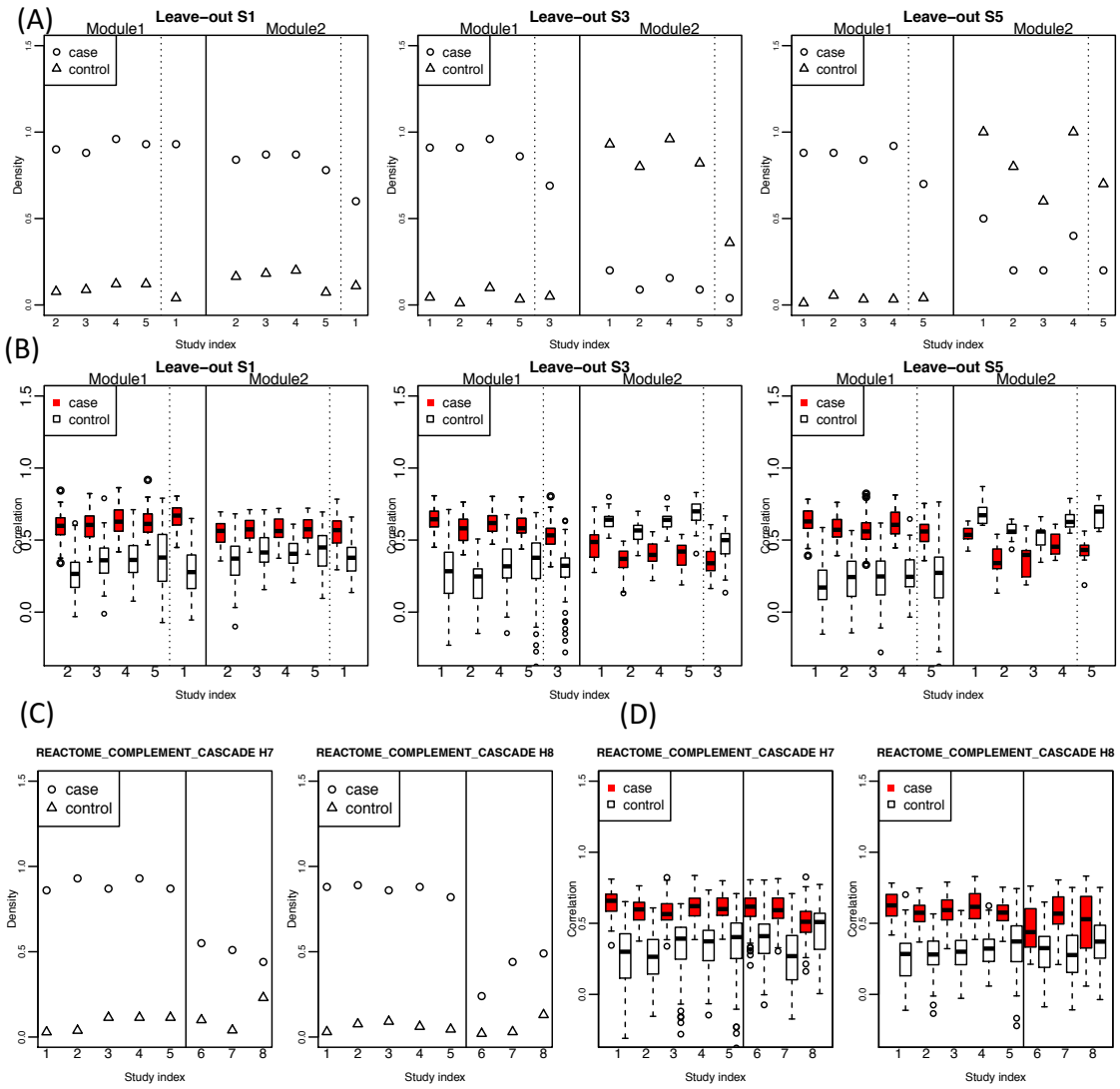


Figure 13: Validation of complement cascade pathway supermodules from ER+ vs ER- comparison

(A) Densities and (B) correlations of the basic modules assembled into complement cascade pathway supermodules in leave-one-out cross-validation. Leaving out study 2 and 4 do not give supermodules significant enriched in complement cascade pathway. Solid lines separate modules, and dashed lines separate training set and testing set. (C) Module density and (D) correlations of genes in the basic modules enriched in complement cascade supermodule in independent validation studies. Solid lines separate training sets and testing sets.

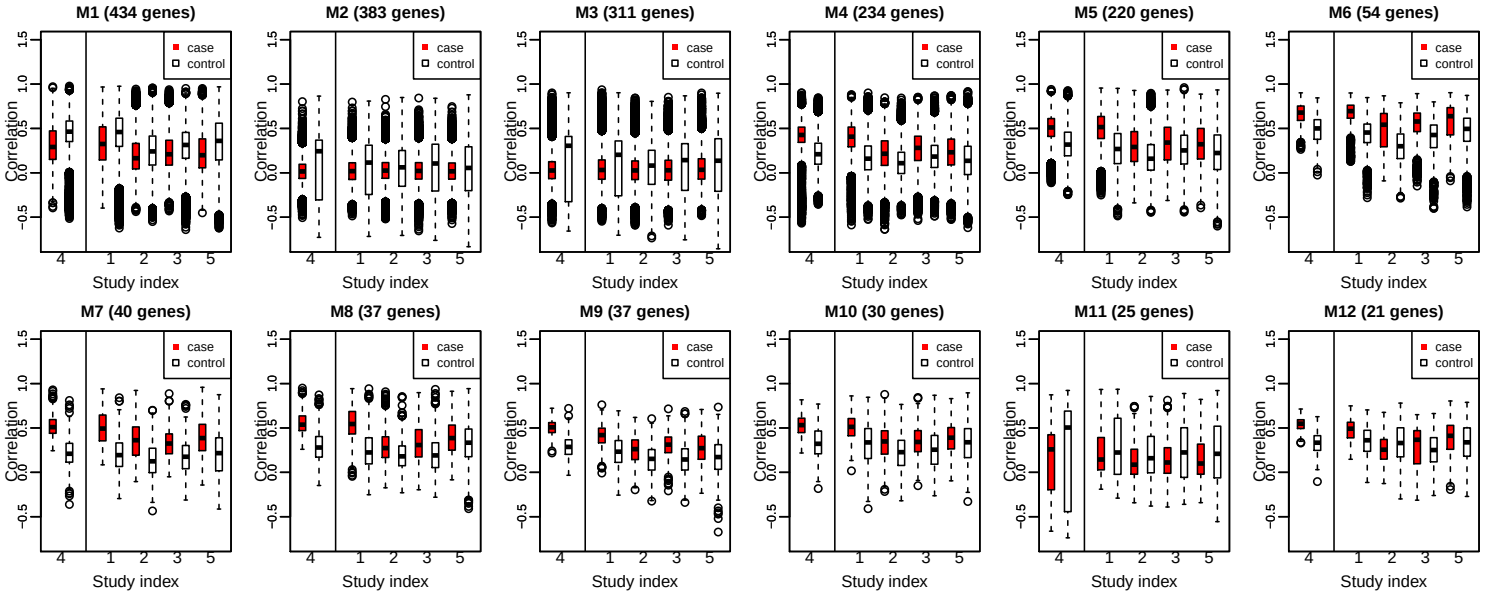


Figure 14: Result from applying DiffCoEx to METABRIC

Apply DiffCoEx (Tesson et al., 2010) on METABRIC: (A) Module density and (B) gene-gene pairwise correlation distributions of 12 modules detected. Solid lines separate training sets (S4) and testing sets (S1, S2, S3, S5)

in other studies. By selecting soft threshold based on free topology fit (Zhang and Horvath, 2005) and cutting hierarchical tree using dynamicTreeCut R package (Langfelder et al., 2008), 12 modules were detected in METABRIC using DiffCoEx. The gene-gene pairwise correlation distributions were calculated for METABRIC as well as the other four studies and the boxplots are shown in Figure 14. Most detected modules only showed moderate degree of validation.

4.3.3 Breast cancer studies (ILC vs. IDC)

We finally applied our method to search for DCN between two breast cancer histological subtypes: ILC (invasive lobular carcinoma) and IDC (invasive ductal carcinoma). IDC and ILC are the two most common subtypes of breast cancers, representing 60-75% and 5-15% of all breast cancer cases, respectively (Guiu et al., 2014). Several studies have

shown that they are two biological distinct diseases by comparing their genomic profiles, but the biological process driving for different subtypes are still largely unknown (Michaut et al., 2016a). Identifying differential co-expression network between ILC and IDC can potentially unveil different biological mechanisms and provide targets for precise treatment for ILC. Using similar parameter settings, with the optimal weights and $R = 10$ repeats, at FDR$\leq$ 0.3, 11 basic DC modules were detected as over-connected in IDC, and no modules were detected as over-connected in ILC. Pathway-enrichment-guided module assembly was performed for varying number of repeats with different initial seed modules. The results were quite consistent. We identified 4 supermodules engaged in 5 pathways, sharing at least 3 overlapping genes with enriched pathway. The top pathways associated with the assembled modules from 10 repeats were listed in Table 11 (B). Figure 15 (A-B) shows the visualizations of two supermodules enriched in protease inhibitor activity and proteinaceous extracellular matrix pathways. We also validated the densities and correlations of the basic modules ensembles in those two pathways in the validation sets (see Figure 15 (C-D)).

In the literature, alteration of extracellular matrix in tumor stroma has been shown relevant to metastatic potential (Oskarsson, 2013). Previous imaging analysis has further demonstrated different evolution of fibrillary collagen changes in ILC versus IDC throughout tumor progression (Burke et al., 2013).

4.4 CONCLUSION

In this study, we proposed a method, MetaDCN, to detect consensus differential co-expression (DC) networks across multiple studies with respect to certain phenotype of interest (e.g. case versus control or ER+ versus ER-). The method optimizes a target function to detect biologically meaningful DC modules. Since global optimization is computationally infeasible and unstable, we developed a simulated annealing algorithm to detect small (size 3 to 30) basic DC modules and assessed their false discovery rate. Through a pathway-guided module assembly algorithm, basic DC modules passing FDR threshold were merged into DC supermodules that were enriched in certain pathways to

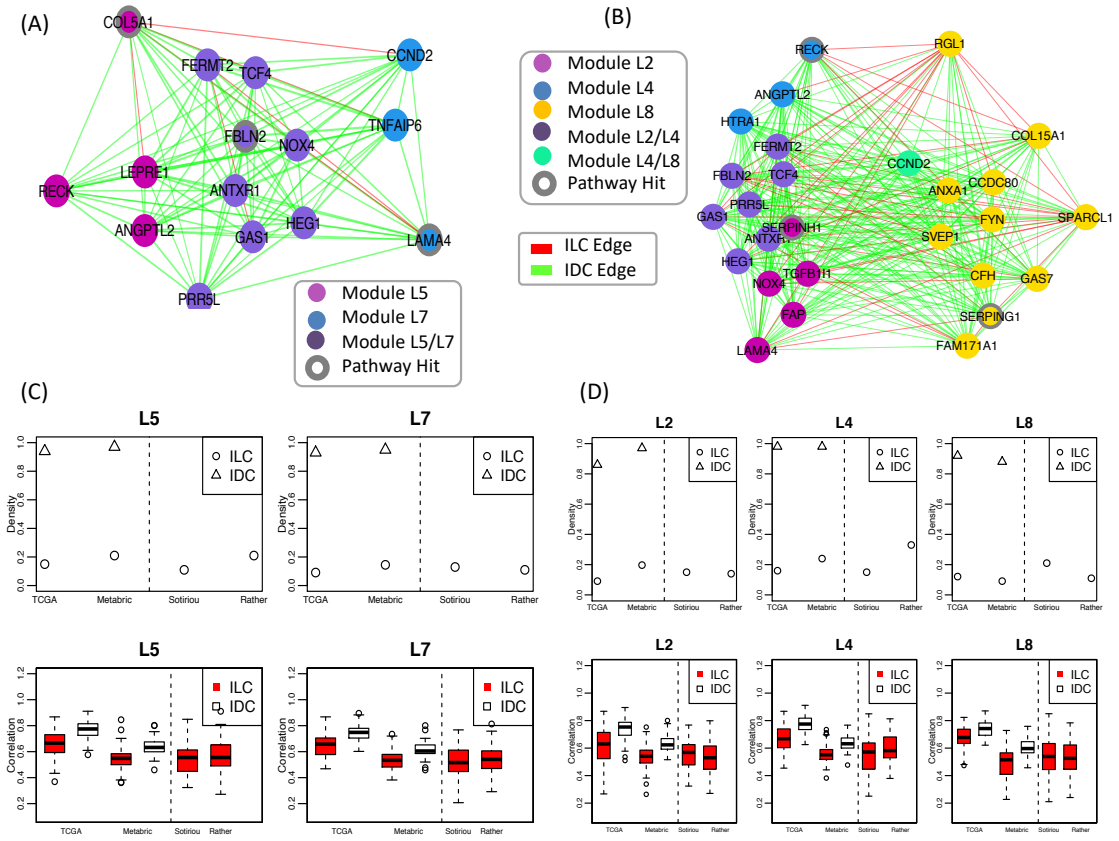


Figure 15: Supermodules from ILC vs IDC comparison and validation

(A) Visualization of proteinaceous extracellular matrix pathway supermodule. (B) Visualization of protease inhibitor activity pathway supermodules. The edge color represents the direction of differential gene co-expression, in which the positive value implies ILC-favored co-expression and the negative value implies IDC-favored co-expression. Node color represents its origin of sub-modules, and the genes annotated in the immune response pathway are highlighted with dark circles. Edge width represents edge weight (Z score of differential co-expression). (C) Module densities and (D) gene-gene pairwise correlations of the basic modules enriched in those two pathways in TCGA, METABRIC, Sotiriou and RATHER. Dashed lines separate training and testing sets.

allow biological interpretation and hypothesis generation. The module assembly approach also allowed over- and under-connected basic DC modules to be simultaneously merged in a DC supermodule, representing possible alternative sub-pathway activation under different phenotypic conditions. Simulations and two real applications in breast cancer studies (ER+ vs. ER- and ILC vs. IDC) demonstrated superior performance of MetaDCN to elucidate novel disease-related differential co-expression mechanisms. DC supermodules identified by training breast cancer studies were further validated in independent studies. A Cytoscape plug-in software, MetaDCNExplorer, was developed to visualize and interactively explore the identified DC networks.

Given limited sample size and potentially biased patient cohort or experimental platform in a single transcriptomic study, detection of DC modules from one study is deemed unstable and often difficult to validate. With the rapid accumulation of transcriptomic studies in the public domain, a meta-analytic approach to combine multiple transcriptomic studies is promising to identify biological meaningful and verifiable DC modules. MetaDCN meets the urgent need for this purpose and is expected to elucidate novel mechanisms in many disease investigations.

5.0 DISCUSSION AND FUTURE WORK

5.1 DISCUSSION

With the increasingly accumulated multi-omics data sets, data integration becomes increasingly important. However, integration of multiple data sets, especially with high dimensions, also poses statistical challenges. This dissertation is focused on two interesting questions in multi-omics data integration: structured feature selection in prediction and clustering models, and differential co-expression network analysis.

The Bayesian indicator variable selection prior proposed in Chapter 2 and 3 can incorporate the "multi-level omics features $\rightarrow$ genes $\rightarrow$ pathway" structure in multi-omics data set. We implemented such a prior into both a linear prediction model (Chapter 2) and the clustering setting (Chapter 3), utilizing the flexibility of Bayesian hierarchical models. We also believe such models need not to be restricted to omics-data integration, but can also be applied to other fields which have multi-layer overlapping group structure.

Differential gene-gene correlation has long been recognized, however, DC analysis is still less studied compared to DE analysis. Given limited sample size and potentially biased patient cohort or experimental platform in a single transcriptomic study, detection of DC modules from one study is deemed unstable and often difficult to validate. With the rapid accumulation of transcriptomic studies in the public domain, the meta-analysis framework proposed in Chapter 4 can be used to detect consensus differential co-expression (DC) networks across multiple studies, which are biological meaningful and verifiable.

5.2 FUTURE WORK

In Chapter 2, we developed a Bayesian indicator variable selection prior for the regression setting and demonstrated its superiority over other methods in simulations and applications. We also proved its oracle property (i.e. feature selection consistency and optimal convergence rate of $\sqrt{n}$) under orthogonal design when groups do not overlap (Appendix A.2). It is well known that lasso and group lasso estimators do not have the oracle properties, and other alternatives (Zou, 2006; Wang and Leng, 2008) were proposed, which may be worth comparing with.

In Chapter 2-3, we considered the vertical integration of different types of mics data of the same cohort of samples. A straightforward extension is the two-way integration, i.e. integrate multi-omics data of multiple cohorts of samples. We believe such extension will not only enjoy the rich information provided by multi-omics data, but also gain statistical power due to increased sample size.

The meta-analytic framework we developed in Chapter 4 utilizes the marginal correlation to construct co-expression network. We are also interested in extending the framework to other metrics. For example, instead of using marginal correlation, the partial correlation can better reveal the correlation between any two genes after correcting for other genes. We expect such modification will not need a lot of changes in the framework but may face new challenges such as sparsity.

APPENDIX A

APPENDIX FOR BAYESIAN INDICATOR VARIABLE SELECTION TO INCORPORATE HIERARCHICAL OVERLAPPING GROUP STRUCTURE IN MULTI-OMICS APPLICATIONS

A.1 MCMC SAMPLING

A.1.1 MCMC sampling of SOG

When groups do not overlap, for the SOG model constructed in Section 2.2.2, following full conditionals can be used for Gibbs sampling:

$$Pr(\gamma_k^{(1)} = 1|-)$$
$$= \left(1 + \frac{(1-\pi^{(1)})}{1-\pi^{(1)}} \exp\left\{\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2\beta_{jk}^2(U_{jk}^{(1)})^2 - 2\sum_{i=1}^{n} y_{i,k}\sum_{j=1}^{p} x_{ij}\beta_{jk}U_{jk}^{(1)}\right)\right\}\right)^{-1}$$
$$Pr(\gamma_{jk}^{(0)} = 1|U_{jk}^{(1)} = 1, -)$$
$$= \left(1 + \frac{(1-\pi_k^{(0)}/R_j)}{\pi_k^{(0)}/R_j} \exp\left\{\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_{ij}^2\beta_{jk}^2(U_{jk}^{(1)})^2 - 2\sum_{i=1}^{n} y_{i,(jk)}x_{ij}\beta_{jk}U_{jk}^{(1)}\right)\right\}\right)^{-1},$$
$$(b_{jk}|\gamma_k^{(1)}\gamma_{jk}^{(0)} = 0, U_{jk}^{(1)} = 1, -) \sim N(0, s^2),$$
$$(b_{jk}|\gamma_k^{(1)}\gamma_{jk}^{(0)} = 1, U_{jk}^{(1)} = 1, -) \sim N\left(\mu_b = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_{ij}y_{i,(jk)})\sigma_b^2, \sigma_b^2 = \frac{1}{\frac{\sum_{i=1}^{n} x_{ij}^2}{\sigma^2} + \frac{1}{s^2}}\right),$$
$$\sigma^2|- \sim \text{Inverse-Gamma}\left(n/2, 1/2\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\sum_{k=1}^{m_1} x_{ij}\beta_{jk}U_{jk}^{(1)})^2\right),$$
$$s^2|- \sim \text{Inverse-Gamma}\left(\sum_{j=1}^{p}\sum_{k=1}^{m_1} U_{jk}^{(1)}/2, 1/2\sum_{j=1}^{p}\sum_{k=1}^{m_1} b_{jk}^2\right),$$
$$\pi^{(1)}|- \sim Beta\left(\sum_{k=1}^{m_1}\gamma_k^{(1)} + 1, m_1 - \sum_{k=1}^{m_1}\gamma_k^{(1)} + 1\right),$$

$$\pi_k^{(0)}|- \sim Beta\left(\sum_{j=1}^{p}\gamma_{jk}^{(0)}+1, \sum_{j=1}^{P}U_{jk}^{(1)}-\sum_{j=1}^{p}\gamma_{jk}^{(0)}+1\right),$$

where "$-$" stands for all other variables, $y_{i,k} = y_i - \sum_{j=1}^{p}\sum_{k'\neq k} x_{ij}\beta_{jk'}U_{jk'}^{(1)}$, and $y_{i,(jk)} = y_i - \sum_{j'\neq j}\sum_{k'=1}^{m_1} x_{ij}\beta_{j'k'}U_{j'k'}^{(1)} - \sum_{k'\neq k} x_{ij}\beta_{jk'}U_{jk'}^{(1)}$.

When groups overlap, $\pi_k^{(0)}$ is drawn using Metropolis-Hasting. We choose the proposal distribution as $\pi_k^{(0),new} \sim Beta(\nu\pi_k^{(0),old}, \nu(1-\pi_k^{(0),old}))$. We set $\nu = 10$, but ν can be tuned with larger n indicating more concentrated around $\pi_k^{(0),old}$. Then, $\pi_k^{(0)}$ is accepted with probability $\min\left(1, \frac{P(\pi_k^{(0),new})\prod_{j=1}^{p}P(\gamma_{jk}^{(0)}|\pi_k^{(0),new},-)P(\pi_k^{(0),old}|\pi_k^{(0),new})}{P(\pi_k^{(0),old})\prod_{j=1}^{p}P(\gamma_{jk}^{(0)}|\pi_k^{(0),old},-)P(\pi_k^{(0),new}|\pi_k^{(0),old})}\right)$.

A.1.2 MCMC sampling of MOG

When groups at same layer do not overlap, MOG model constructed in Section 2.2.3 has following full conditionals for Gibbs sampling:

$$Pr(\gamma_l^{(2)}=1|-)$$
$$= \left(1+\frac{(1-\pi^{(2)})}{\pi^{(2)}}\exp\left\{\left(\sum_{i,j,k}x_{ij}^2\beta_{jkl}^2(U_{jk}^{(1)}U_{kl}^{(2)})^2 - 2\sum_{i=1}^{n}y_{i,l}\sum_{j,k}x_{ij}\beta_{jkl}U_{jk}^{(1)}U_{kl}^{(2)}\right)/\sigma^2\right\}\right)^{-1},$$
$$Pr(\gamma_{kl}^{(1)}=1|U_{kl}^{(2)}=1,-) =$$
$$\left(1+\frac{(1-\pi_l^{(1)}/D_k)}{\pi_l^{(1)}/D_k}\exp\left\{\left(\sum_{i,j}x_{ij}^2\beta_{jkl}^2(U_{jk}^{(1)}U_{kl}^{(2)})^2 - 2\sum_{i=1}^{n}y_{i,(kl)}\sum_{j=1}^{p}x_{ij}\beta_{jkl}U_{jk}^{(1)}U_{kl}^{(2)}\right)/\sigma^2\right\}\right)^{-1},$$
$$Pr(\gamma_{jkl}^{(0)}=1|U_{jk}^{(1)}U_{kl}^{(2)}=1,-) =$$
$$\left(1+\frac{(1-\pi_{kl}^{(0)}/R_j)}{\pi_{kl}^{(0)}/R_j}\exp\left\{\left(\sum_{i=1}^{n}x_{ij}^2\beta_{jkl}^2(U_{jk}^{(1)}U_{kl}^{(2)})^2 - 2\sum_{i=1}^{n}y_{i,jkl}x_{ij}\beta_{jkl}U_{jk}^{(1)}U_{kl}^{(2)}\right)/\sigma^2\right\}\right)^{-1},$$
$$(b_{jkl}|\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}=0, U_{jk}^{(1)}U_{kl}^{(2)}=1,-) \sim N(0,s^2),$$
$$(b_{jkl}|\gamma_l^{(2)}\gamma_{kl}^{(1)}\gamma_{jkl}^{(0)}=1,-) \sim N\left(\mu_b=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_{ij}y_{i,jkl})\sigma_b^2, \sigma_b^2=\frac{1}{\frac{\sum_{i=1}^{n}x_{ij}^2}{\sigma^2}+\frac{1}{s^2}}\right),$$
$$\sigma^2|- \sim \text{IG}\left(n/2, 1/2\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2}x_{ij}\beta_{jkl}U_{jk}^{(1)}U_{kl}^{(2)})^2\right),$$
$$s^2|- \sim \text{Inverse-Gamma}\left(\sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2}U_{jk}^{(1)}U_{kl}^{(2)}/2, 1/2\sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l=1}^{m_2}b_{jkl}^2\right),$$
$$\pi^{(2)}|- \sim Beta\left(\sum_{l=1}^{m_2}\gamma_l^{(2)}+1, M_1-\sum_{l=1}^{m_2}\gamma_l^{(2)}+1\right),$$

$$\pi_l^{(1)}|- \sim Beta\left(\sum_{k=1}^{m_1}\gamma_{kl}^{(1)}+1, \sum_{k=1}^{m_2}U_{kl}^{(2)}-\sum_{k=1}^{m_1}\gamma_{kl}^{(1)}+1\right),$$
$$\pi_{kl}^{(0)}|U_{kl}^{(2)}=1,- \sim Beta\left(\sum_{j=1}^{p}\gamma_{jkl}^{(0)}+1, \sum_{j=1}^{p}U_{jk}^{(1)}U_{kl}^{(2)}-\sum_{k=1}^{p}\gamma_{jkl}^{(0)}+1\right)$$

where $y_{i,l} = y_i - \sum_{j=1}^{p}\sum_{k=1}^{m_1}\sum_{l'\neq l} x_{ij}\beta_{jkl'}U_{jk}^{(1)}U_{kl'}^{(2)}$, $y_{i,kl} = y_i - \sum_{j=1}^{p}\sum_{k'=1}^{m_1}\sum_{l'\neq l} x_{ij}\beta_{jk'l'}U_{jk'}^{(1)}U_{k'l'}^{(2)}$-$\sum_{j=1}^{p}\sum_{k'\neq k}\sum_{l'=1}^{m_2} x_{ij}\beta_{jk'l'}U_{jk'}^{(1)}U_{k'l'}^{(2)}$, and $y_{i,jkl} = y_i - \sum_{j'\neq j}^{p}\sum_{k'=1}^{m_1}\sum_{l'=1}^{m_2} x_{ij'}\beta_{j'k'l'}U_{j'k'}^{(1)}U_{k'l'}^{(2)} - \sum_{k'\neq k}\sum_{l'=1}^{m_2} x_{ij}\beta_{jk'l'}U_{jk'}^{(1)}U_{k'l'}^{(2)} - \sum_{l'\neq l} x_{ij}\beta_{jkl'}U_{jk}^{(1)}U_{kl'}^{(2)}$.

Similar to SOG, when groups at the same layer overlap, $\pi_{kl}^{(0)}$ and $\pi_l^{(1)}$ are drawn using Metropolis-Hasting. Same proposal distributions are used: $\pi_{kl}^{(0),new} \sim Beta(\nu\pi_{kl}^{(0),old}, \nu(1-\pi_{kl}^{(0),old}))$ and $\pi_l^{(1),new} \sim Beta(\nu\pi_l^{(1),old}, \nu(1-\pi_l^{(1),old}))$, with $\nu = 10$. Then, $\pi_{kl}^{(0),new}$ is accepted with probability $min\left(1, \frac{P(\pi_{kl}^{(0),new})\prod_{j=1}^{p}P(\gamma_{jkl}^{(0)}|\pi_{kl}^{(0),new},-)P(\pi_{kl}^{(0),old}|\pi_{kl}^{(0),new})}{P(\pi_k^{(0),old})\prod_{j=1}^{p}P(\gamma_{jkl}^{(0)}|\pi_{kl}^{(0),old},-)p(\pi_{kl}^{(0),new}|\pi_{kl}^{(0),old})}\right)$, and $\pi_l^{(1),new}$ is accepted with probability $min\left(1, \frac{P(\pi_l^{(1),new})\prod_{k=1}^{m_1}P(\gamma_{kl}^{(1)}|\pi_l^{(1),new},-)P(\pi_l^{(1),old}|\pi_l^{(1),new})}{P(\pi_l^{(1),old})\prod_{k=1}^{m_1}P(\gamma_{kl}^{(1)}|\pi_{kl}^{(1),old},-)p(\pi_{kl}^{(1),new}|\pi_{kl}^{(1),old})}\right)$.

A.2 PROOFS OF ASYMPTOTIC PROPERTIES OF POSTERIOR MEDIAN ESTIMATOR

Lemma 1. Assuming orthogonal design (i.e. $X^TX = nI$) and all levels of groups are disjoint, given $\pi^{(0)}$, $\pi^{(1)}$, $\pi^{(2)}$, s^2, σ^2 are known, the posterior median estimator $\hat{\beta}_{jkl}^{Med}$, i.e. the median of $P(\beta_{jkl}|y,X)$, in MOG is a soft-thresholding estimator.

Proof: The marginal prior of β_{jkl} in MOG is a "spike-and-slab":

$$\beta_{jkl}|U_{jk}^{(1)}U_{kl}^{(2)}=1 \sim (1-\pi^{(2)}\pi_l^{(1)}\pi_{kl}^{(0)})\delta_0(\beta_{jkl}) + \pi^{(2)}\pi_l^{(1)}\pi_{kl}^{(0)}N(0,s^2).$$

Here, we are going to omit $U_{jk}^{(1)}U_{kl}^{(2)} = 1$, i.e. assuming feature j belongs to level-1 group k and level-2 group l. To simplify the notation, we define $\pi^* \triangleq \pi^{(2)}\pi_l^{(1)}\pi_{kl}^{(0)}$, $y_{i,jkl}$ as the residual subtracting out the contribution of all other features $y_{i,jkl} \triangleq y_i -$

$\sum_{j'\neq j}^{p}\sum_{k'=1}^{m_1}\sum_{l'=1}^{m_2} x_{ij'}\beta_{j'k'l'}U^{(1)}_{j'k'}U^{(2)}_{k'l'} - \sum_{k'\neq k}\sum_{l'=1}^{m_2} x_{ij}\beta_{jk'l'}U^{(1)}_{jk'}U^{(2)}_{k'l'} - \sum_{l'\neq l} x_{ij}\beta_{jkl'}U^{(1)}_{jk}U^{(2)}_{kl'}$, and $\beta_{(-jkl)}$ as the vector β excluding β_{jkl}. We also use $\sum$ to denote $\sum_{i=1}^{n}$, unless specified otherwise.

1. Derive the posterior distribution $P(\beta_{jkl}|y,X)$.

$$
\begin{aligned}
&P(\beta_{jkl}|y,X,\beta_{(-jkl)}) \propto P(\beta_{jkl}|\beta_{-jkl})P(Y|X,\beta_{jkl},\beta_{(-jkl)})\\
&= \left(\pi^{(2)}\pi^{(1)}_{l}\pi^{(0)}_{kl}\frac{1}{\sqrt{2\pi s^2}}\exp\left(-\frac{\beta^2_{jkl}}{2s^2}\right) + (1-\pi^{(2)}\pi^{(1)}_{l}\pi^{(0)}_{kl})\delta_0(\beta_{jkl})\right)\\
&\quad\times\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\sum\left(y_i - \sum_{j'=1}^{p}\sum_{k'=1}^{m_1}\sum_{l'=1}^{m_2}x_{ij'}\beta_{j'k'l'}\right)^2/2\sigma^2\right)\\
&\propto \pi^*\frac{1}{\sqrt{2\pi s^2}}\exp\left(-\frac{\beta^2_{jkl}}{2s^2}\right)\exp\left(-\sum\left(y_{i,jkl}-x_{ij}\beta_{jkl}\right)^2/2\sigma^2\right)\\
&+(1-\pi^*)\exp\left(-\frac{\sum y^2_{i,jkl}}{2\sigma^2}\right)\delta_0(\beta_{jkl})\\
&= \pi^*\frac{1}{\sqrt{2\pi s^2}}\exp\left(-\left(\frac{\beta^2_{jkl}}{2s^2}+\frac{\sum y^2_{i,jkl}-2\beta_{jkl}\sum y_{i,jkl}x_{ij}+\beta^2_{jkl}\sum x^2_{ij}}{2\sigma^2}\right)\right)\\
&+(1-\pi^*)\exp\left(-\frac{\sum y^2_{i,jkl}}{2\sigma^2}\right)\delta_0(\beta_{jkl})\\
&= \pi^*\frac{1}{\sqrt{2\pi s^2}}\exp\left(-\left(\frac{1}{2s^2}+\frac{\sum x^2_{ij}}{2\sigma^2}\right)\beta^2_{jkl}+\left(\frac{\sum y_{i,jkl}x_{ij}}{2\sigma^2}\right)2\beta_{jkl}-\frac{\sum y^2_{i,jkl}}{2\sigma^2}\right)\\
&+(1-\pi^*)\exp\left(-\frac{\sum y^2_{i,jkl}}{2\sigma^2}\right)\delta_0(\beta_{jkl})\\
&\propto \pi^*\frac{1}{\sqrt{2\pi s^2}}\exp\left(-\frac{1}{2}\left(\frac{1}{s^2}+\frac{\sum x^2_{ij}}{\sigma^2}\right)\left(\beta^2_{ijk}-\left(\frac{\sum y_{i,jkl}x_{ij}}{2\sigma^2}\right)/\left(\frac{1}{2s^2}+\frac{\sum x^2_{ij}}{2\sigma^2}\right)2\beta_{jkl}\right)\right)\\
&+(1-\pi^*)\delta_0(\beta_{jkl}),
\end{aligned}
$$

Since $X^TX = nI$, i.e. $x^T_{\cdot j}x_{\cdot j} = n$ and $x^T_{\cdot j}x_{\cdot j'} = 0$, each $\hat{\beta}^{LS}_{jkl}$ $(j = 1,\ldots,p; k = 1,\ldots,m_1; l = 1,\ldots,m_2)$ is independent from each other. In fact, $\hat{\beta}^{LS}_{jkl} = \sum x_{ij}y_i/n$, and hence, $\sum y_{i,jkl}x_{ij} = \sum\left(y_i - \sum_{j'\neq j}x_{ij'}\beta_{j'}\right)x_{ij} = \sum y_i x_{ij} = n\hat{\beta}^{LS}_{jkl}$, which is independent from $\beta_{(-jkl)}$. Therefore,

$$
\begin{aligned}
&P(\beta_{jkl}|y,X,\beta_{(-jkl)}) = P(\beta_{jkl}|y,X)\\
&\propto \pi^*\frac{1}{\sqrt{s^2}}\sqrt{\frac{s^2\sigma^2}{\sigma^2+ns^2}}\frac{1}{\sqrt{2\pi\left(\frac{s^2\sigma^2}{\sigma^2+ns^2}\right)}}\exp\left(-\frac{\left(\beta_{jkl}-\frac{ns^2\hat{\beta}^{LS}_{jkl}}{\sigma^2+ns^2}\right)^2}{2\left(\frac{s^2\sigma^2}{\sigma^2+ns^2}\right)}\right)\exp\left(\frac{n^2s^2(\hat{\beta}^{LS}_{jkl})^2}{2\sigma^4+2\sigma^2ns^2}\right)\\
&+(1-\pi^*)\delta_0(\beta_{jkl})
\end{aligned}
$$

$$\propto \pi^* \sqrt{\frac{\sigma^2}{\sigma^2+ns^2}} \phi\left(\beta_{jkl}; \frac{ns^2\hat{\beta}_{jkl}^{LS}}{\sigma^2+ns^2}, \frac{s^2\sigma^2}{\sigma^2+ns^2}\right) \exp\left(\frac{n^2s^2\left(\hat{\beta}_{jkl}^{LS}\right)^2}{2\sigma^4+2\sigma^2ns^2}\right) + (1-\pi^*)\delta_0(\beta_{jkl}),$$

where $\phi(x;\mu,\sigma^2)$ indicates the normal density of x with mean μ and variance σ^2. We further normalize the distribution to make it proper,

$$\begin{aligned} r_{jkl} &\triangleq Pr(\beta_{jkl} \neq 0|y,X) = \frac{\pi^* \sqrt{\frac{\sigma^2}{\sigma^2+ns^2}} \exp\left(\frac{n^2s^2\left(\hat{\beta}_{jkl}^{LS}\right)^2}{2\sigma^4+2\sigma^2ns^2}\right)}{1-\pi^*+\pi^* \sqrt{\frac{\sigma^2}{\sigma^2+ns^2}} \exp\left(\frac{n^2s^2\left(\hat{\beta}_{jkl}^{LS}\right)^2}{2\sigma^4+2\sigma^2ns^2}\right)} \\ &= \frac{\pi^*}{\pi^*+(1-\pi^*)\left(\frac{\sigma^2}{\sigma^2+ns^2}\right)^{-1/2} \exp\left(-\frac{n^2s^2(\hat{\beta}_{jkl}^{LS})^2}{2\sigma^4+2n\sigma^2s^2}\right)}. \end{aligned}$$

To simplify the notation, we define $B_n \triangleq \frac{\sigma^2}{\sigma^2+ns^2}$ and $\tau^2 \triangleq \frac{s^2}{\sigma^2}$, then

$$r_{jkl} = \frac{\pi^*}{\pi^*+(1-\pi^*)(1+n\tau^2)^{1/2} \exp\left(-\frac{(1-B_n)}{2\sigma^2} n(\hat{\beta}_{jkl}^{LS})^2\right)}.$$

Also, the posterior distribution of β_{jkl} is:

$$\beta_{jkl}|y,X \sim (1-r_{jkl})\delta_0(\beta_{jkl}) + r_{jkl} N\left((1-B_n)\hat{\beta}_{jkl}^{LS}, \sigma^2(1-B_n)/n\right).$$

2. Derive the soft-thresholding estimator.

We define $\hat{\beta}_{jkl}^{Med} \triangleq Med(\beta|y,X)$ (i.e. median of $P(\beta|y,X)$) as the posterior median estimator. The posterior distribution of β_{jkl} is a mixture of point mass at zero and a normal distribution with mean and variance denoted as $M \triangleq (1-B_n)\hat{\beta}_{jkl}^{LS}$ and $Var \triangleq \sigma^2(1-B_n)/n$.

(1) Since $Pr(\beta_{jkl}=0|y,X) = 1-r_{ijk}$, if $r_{ijk} < 1/2$, $\hat{\beta}_{jkl}^{Med} = 0$.

(2) Then, we consider $r_{ijk} \geq 1/2$.
If $\hat{\beta}_{jkl}^{LS} \geq 0$, $(1-B_n)\hat{\beta}_{jkl}^{LS} \geq 0$. Then,

$$\Phi\left(-\frac{\hat{\beta}_{jkl}^{Med}-M}{\sqrt{Var}}\right) = \frac{1}{2r_{jkl}} \Rightarrow \hat{\beta}_{jkl}^{Med} = (1-B_n)\hat{\beta}_{jkl}^{LS} - \frac{\sigma}{\sqrt{n}}\sqrt{1-B_n}\Phi^{-1}\left(\frac{1}{2r_{jkl}}\right)$$

If $\hat{\beta}_{jkl}^{LS} < 0$, $(1-B_n)\hat{\beta}_{jkl}^{LS} < 0$. Then,

$$\Phi\left(\frac{\hat{\beta}_{jkl}^{Med}-M}{\sqrt{Var}}\right) = \frac{1}{2r_{jkl}} \Rightarrow \hat{\beta}_{jkl}^{Med} = (1-B_n)\hat{\beta}_{jkl}^{LS} + \frac{\sigma}{\sqrt{n}}\sqrt{1-B_n}\Phi^{-1}\left(\frac{1}{2r_{jkl}}\right)$$

Combine all together,

$$\hat{\beta}_{jkl}^{Med} = sgn(\hat{\beta}_{jkl}^{LS}) \left((1 - B_n)|\hat{\beta}_{jkl}^{LS}| - \frac{\sigma}{\sqrt{n}} \sqrt{1 - B_n} \Phi^{-1} \left(\frac{1}{2 \max(r_{jkl}, 1/2)} \right) \right)_+ ,$$

where sgn is the sign function, Φ is cumulative distribution function (CDF) of standard normal distribution, and $(x)_+$ takes the value of x if $x > 0$, and zero otherwise. ■

Theorem 1. Define an index vector of true non-zero features $\mathcal{A} = (I(\beta_{jkl}^0 \neq 0), j = 1, .., p; k = 1, .., m_1; l = 1, ..., m_2)$ and $\mathcal{A}_n$ for index vector from posterior median estimator. Under the assumption that all levels of groups are disjoint, $X^T X = nI$, $\sqrt{n} s^2/\sigma^2 \to \infty$ and $\log(s^2/\sigma^2)/n \to 0$ as $n \to \infty$, we have

$$\begin{aligned} &\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1 \qquad &\text{(Selection consistency)}, \\ &\sqrt{n}(\hat{\beta}_{\mathcal{A}}^{Med} - \beta_{\mathcal{A}}^0) \to N(0, \sigma^2 I) \qquad &\text{(Asymptotic normality)}. \end{aligned}$$

Proof:

(1) Prove selection consistency. We consider $\beta_{jkl}^0 = 0$ and $\beta_{jkl}^0 \neq 0$ separately.

(i) When $\beta_{jkl}^0 = 0$, from the proof of Lemma,

$$Pr(\hat{\beta}_{jkl}^{Med} = 0 | y, X) = Pr \left(\frac{\sqrt{1 - B_n} \sqrt{n} |\hat{\beta}_{jkl}^{LS}|}{\sigma \Phi^{-1} \left(\frac{1}{2 \max(1/2, r_{jkl})} \right)} < 1 \right).$$

Since $n\tau^2 \to \infty$, $B_n = 1/(1 + n\tau^2) \to 0$; Also, $X^T X = nI$, $\frac{\sqrt{n}|\hat{\beta}_{jkl}^{LS} - \beta_{jkl}^0|}{\sigma} \xrightarrow{d} z \sim N(0, 1)$, hence, $\sqrt{n}|\hat{\beta}_{jkl}^{LS}| = O_p(1)$, and $r_{jkl} \xrightarrow{p} 0$. Therefore, $Pr(\hat{\beta}_{jkl}^{Med} = 0 | y, X) \to 1$.

(ii) When $\beta_{jkl}^0 \neq 0$,

$$Pr(\hat{\beta}_{jkl}^{Med} \neq 0 | y, X) = Pr \left(\frac{\sigma \Phi^{-1} \left(\frac{1}{2 \max(1/2, r_{jkl})} \right)}{\sqrt{1 - B_n} \sqrt{n} |\hat{\beta}_{jkl}^{LS}|} < 1 \right).$$

Now, by definition, $r_{jkl} = \frac{\pi^*}{\pi^* + (1 - \pi^*) \frac{(1+n\tau^2)^{1/2}}{\exp\left(-\frac{(1-B_n)}{2\sigma^2} n (\hat{\beta}_{jkl}^{LS})^2\right)}}$. By L'Hospital's rule, $\lim_{n\to\infty} \frac{(1+n\tau^2)^{1/2}}{\exp\left(-\frac{(1-B_n)}{2\sigma^2} n\right)} = \lim_{n\to\infty} \frac{1/2\tau^2 (1+n\tau^2)^{-1/2}}{\exp\left(-\frac{1-B_n}{2\sigma^2} n\right)\left(-\frac{1-B_n}{2\sigma^2}\right)} = 0$, since $\tau^2/\exp(n) \to 0$ and $B_n \to 0$. Also, $\hat{\beta}_{jkl}^{LS} \xrightarrow{p} \beta_{jkl}^0 \neq 0$. By continuous mapping theorem, $r_{jkl} \xrightarrow{p} 1$ and $Pr(\hat{\beta}_{jkl}^{Med} \neq 0 | y, X) \to 1$.

This concludes the selection consistency of each coefficients $\hat{\beta}^{Med}_{jkl}$. Under orthogonal design, each $\hat{\beta}^{LS}_{jkl}$ is independent, and so is $\hat{\beta}^{Med}_{jkl}$. Therefore, we conclude the selection consistency for all coefficients.

(2) Prove the asymptotic normality of $\beta^{Med}_{\mathcal{A}}$.

When $\beta^0_{jkl} \neq 0$,

$$\begin{aligned}
&\left|\sqrt{n}(\hat{\beta}^{Med}_{jkl} - \hat{\beta}^{LS}_{jkl})\right| \\
&= \left|\sqrt{n}B_n \left|\hat{\beta}^{LS}_{jkl}\right| + \sigma\sqrt{1-B_n}\Phi^{-1}\left(\frac{1}{2\max(\frac{1}{2}, r_{jkl})}\right)\right| I(\hat{\beta}^{Med}_{jkl} \neq 0) \\
&+ \sqrt{n}\left|\hat{\beta}^{LS}_{jkl}\right| I(\hat{\beta}^{Med}_{jkl} = 0) \xrightarrow{\text{p}} 0,
\end{aligned}$$

since $\sqrt{n}B_n \to 0$, $\hat{\beta}^{LS}_{jkl} \xrightarrow{p} \beta^0_{jkl} \neq 0$, $B_n \to 0$, $r_{ikl} \xrightarrow{p} 1$, $I(\hat{\beta}^{Med}_{jkl} \neq 0) \xrightarrow{p} 1$, and $\sqrt{n}I(\hat{\beta}^{Med}_{jkl} = 0) \xrightarrow{p} 0$.

Again, under orthogonal design, $\sqrt{n}(\hat{\beta}^{Med}_{\mathcal{A}} - \hat{\beta}^{LS}_{\mathcal{A}}) \xrightarrow{p} 0$. Since $\sqrt{n}(\hat{\beta}^{LS}_{\mathcal{A}} - \beta^0_{\mathcal{A}}) \xrightarrow{\text{d}} N(0, \sigma^2 I)$, by Slutsky theorem, $\sqrt{n}(\hat{\beta}^{Med}_{\mathcal{A}} - \beta^0_{\mathcal{A}}) \xrightarrow{\text{d}} N(0, \sigma^2 I)$. ■

Theorem 2. With the same assumptions in Theorem 1, but groups allowed to be overlapped, define the marginal coefficient as β_{mar} with each $\beta_{mar,j} = \sum_{k=1}^{m_1}\sum_{l=1}^{m_2}\beta_{jkl}U^{(1)}_{jk}U^{(2)}_{kl}$, and similarly define $\hat{\beta}^{Med}_{mar,j}$ as the posterior median estimator, and $\mathcal{A}_{mar} = (I(\beta^0_{mar,j} \neq 0), j = 1, .., p; k = 1, .., m_1; l = 1, ..., m_2)$ and $\mathcal{A}_{mar,n} = (I((\hat{\beta}^{Med}_{mar,j})^0 \neq 0), j = 1, .., p; k = 1, .., m_1; l = 1, ..., m_2)$, then

$$\lim_{n\to\infty} P(\mathcal{A}_{mar,n} = \mathcal{A}_{mar}) = 1 \qquad \text{(Selection consistency)}.$$

Proof: For simplicity, we assume level-0 feature j is shared by two level-1 groups k and k', each of which belongs to a level-2 group l and l', then the marginal coefficient $\beta_{mar,j} = \beta_{jkl} + \beta_{jk'l'}$ ($k' \neq k$ and $l' \neq l$). Even though each individual partial effect is not identifiable, the marginal effect $\beta_{mar,j}$ is identifiable.

Different from the disjoint group setting, the marginal prior for overlapping feature $\beta_{mar,j}$ is "one-spike-and-two-slabs":

$$\begin{aligned}
\beta_{mar,j} \sim\ & \pi_A\pi_B N(0, 2s^2) + (\pi_A(1-\pi_B) + (1-\pi_A)\pi_B)\, N(0, s^2) \\
&+ (1-\pi_A)(1-\pi_B)\delta_0(\beta_{mar,j}),
\end{aligned}$$

where $\pi_A \triangleq \pi^{(2)}\frac{\pi^{(1)}_l}{D_k}\frac{\pi^{(0)}_{kl}}{R_j}$ and $\pi_B \triangleq \pi^{(2)}\frac{\pi^{(1)}_{l'}}{D'_k}\frac{\pi^{(0)}_{k'l'}}{R_j}$.

(1) Derive the posterior distribution.
Similar to the proof of Lemma,

$$
\begin{aligned}
&P(\beta_{mar,j}|y,X) \propto P(\beta_{mar,j})P(Y|X,\beta_{mar,j}) \\
&\propto \left(\pi_A\pi_B\phi(\beta_{mar,j};0,2s^2) + (\pi_A(1-\pi_B)+(1-\pi_A)\pi_B)\,\phi(\beta_{mar,j};0,s^2)\right. \\
&\quad \left.+(1-\pi_A)(1-\pi_B)\delta_0(\beta_{mar,j})\right) \times \exp\left(-\sum\left(y_i - \sum_{j'} x_{ij'}\beta_{mar,j'})^2\right) / \left(2\sigma^2\right)\right) \\
&\propto \pi_A\pi_B\sqrt{\frac{\sigma^2}{\sigma^2+2ns^2}}\exp\left(\frac{n^2(\hat{\beta}^{LS}_{mar,j})^2 2s^2}{2\sigma^4+4\sigma^2 ns^2}\right)\phi\left(\beta_{mar,j};\frac{2n\hat{\beta}^{LS}_{mar,j}s^2}{\sigma^2+2ns^2},\frac{2s^2\sigma^2}{\sigma^2+2ns^2}\right) \\
&\quad + (\pi_A(1-\pi_B)+(1-\pi_A)\pi_B)\sqrt{\frac{\sigma^2}{\sigma^2+ns^2}}\exp\left(\frac{n^2(\hat{\beta}^{LS}_{mar,j})^2 s^2}{2\sigma^4+2\sigma^2 ns^2}\right) \\
&\quad \times\phi\left(\beta_{mar,j};\frac{n\hat{\beta}^{LS}_{mar,j}s^2}{\sigma^2+ns^2},\frac{s^2\sigma^2}{\sigma^2+ns^2}\right) + (1-\pi_A)(1-\pi_B)\delta_0(\beta_{mar,j})
\end{aligned}
$$

To simplify the notation, we further define $B_n^* = 1/(1+2ns^2/\sigma^2)$. We can easily notice that $B_n^* \to 0$ and $\sqrt{n}B_n^* \to 0$, similar to B_n. Again, we need further normalize the distribution, and the marginal posterior distribution is a mixture of a point mass at 0 and two normal distributions:

$$
\begin{aligned}
\beta_{mar,j}|y,X \sim\ & r_A N\left((1-B_n^*)\hat{\beta}^{LS}_{mar,j},\sigma^2(1-B_n^*)/n\right) \\
&+ r_B N\left((1-B_n)\hat{\beta}^{LS}_{mar,j},\sigma^2(1-B_n)/n\right) \\
&+ r_C\delta_0(\beta_{mar,j}),
\end{aligned}
$$

where r_A, r_B and r_C are the normalized posterior weights for the three distributions with following forms:

$$
\begin{aligned}
r_A &= C_{cons}\pi_A\pi_B(1+2n\tau^2)^{-1/2}\exp\left(\frac{(1-B_n^*)}{2\sigma^2}n(\hat{\beta}^{LS}_{mar,j})^2\right), \\
r_B &= C_{cons}(\pi_A(1-\pi_B)+(1-\pi_A)\pi_B)(1+n\tau^2)^{-1/2}\exp\left(\frac{(1-B_n)}{2\sigma^2}n(\hat{\beta}^{LS}_{mar,j})^2\right), \\
r_C &= C_{cons}(1-\pi_A)(1-\pi_B),
\end{aligned}
$$

where C_{cons} is the normalizing constant to ensure $r_A + r_B + r_C = 1$.
More explicitly,

$$
\begin{aligned}
r_C &= P(\beta_{mar,j}=0|y,X) \\
&= (1-\pi_A)(1-\pi_B)/\left(\pi_A\pi_B(1+2n\tau^2)^{-1/2}\exp\left(\frac{(1-B_n^*)}{2\sigma^2}n(\hat{\beta}^{LS}_{mar,j})^2\right)\right.
\end{aligned}
$$

$$+ (\pi_A(1-\pi_B) + (1-\pi_A)\pi_B)(1+n\tau^2)^{-1/2} \exp\left(\frac{(1-B_n)}{2\sigma^2} n(\hat{\beta}^{LS}_{mar,j})^2\right)$$
$$+ (1-\pi_A)(1-\pi_B)).$$

Following the similar proof in Theorem 1, one can easily prove

- when $\beta^0_{mar,j} = 0$, $\sqrt{n}\hat{\beta}^{LS}_{mar,j} = O_p(1)$, $r_C \xrightarrow{p} 1$.
- when $\beta^0_{mar,j} \neq 0$, $\hat{\beta}^{LS}_{mar,j} \xrightarrow{p} \beta^0_{mar,j}$, $r_C \xrightarrow{p} 0$.

(2) Prove the selection consistency.

The CDF of $\beta_{mar,j}$ is

$$\begin{aligned}
F(\beta_{mar,j}|y,X) =& r_A\Phi\left(\frac{\beta_{mar,j} - (1-B^*_n)\hat{\beta}^{LS}_{mar,j}}{\sqrt{\sigma^2(1-B^*_n)/n}}\right) \\
&+ r_B\Phi\left(\frac{\beta_{mar,j} - (1-B_n)\hat{\beta}^{LS}_{mar,j}}{\sqrt{\sigma^2(1-B_n)/n}}\right) \\
&+ r_C\mathbb{I}(\beta_{mar,j} \geq 0) \\
=& r_A\Phi\left(\frac{\sqrt{n}\beta_{mar,j} - \sqrt{n}(1-B^*_n)\hat{\beta}^{LS}_{mar,j}}{\sqrt{\sigma^2(1-B^*_n)}}\right) \\
&+ r_B\Phi\left(\frac{\sqrt{n}\beta_{mar,j} - \sqrt{n}(1-B_n)\hat{\beta}^{LS}_{mar,j}}{\sqrt{\sigma^2(1-B_n)}}\right) \\
&+ r_C\mathbb{I}(\beta_{mar,j} \geq 0).
\end{aligned}$$

We further define $V^*_n \triangleq \frac{\sqrt{n}\beta_{mar,j} - \sqrt{n}(1-B^*_n)\hat{\beta}^{LS}_{mar,j}}{\sqrt{\sigma^2(1-B^*_n)}}$ and $V_n \triangleq \frac{\sqrt{n}\beta_{mar,j} - \sqrt{n}(1-B_n)\hat{\beta}^{LS}_{mar,j}}{\sqrt{\sigma^2(1-B_n)}}$, for notation simplicity.

(i) First consider $\beta^0_{mar,j} = 0$.

When $\beta_{mar,j} = 0$, since $r_C \xrightarrow{p} 1$, i.e. $\forall\ \delta_1 > 0$, $\epsilon_1 > 0$, $\exists\ N_1$, s.t. $Pr(r_C > 1-\epsilon_1) > 1-\delta_1$, whenever $n > N_1$. Also, as $r_A\Phi(V^*_n) + r_B\Phi(V_n) \geq 0$,

$$Pr\left(F(\beta_{mar,j} = 0|y,X) > 1-\epsilon_1\right) = Pr\left(r_A\Phi(V^*_n) + r_B\Phi(V_n) + r_C > 1-\epsilon_1\right)$$
$$\geq Pr(r_C > 1-\epsilon_1) > 1-\delta_1, \text{ whenever } n > N_1.$$

In other words, $F(\beta_{mar,j} = 0|y,X) \xrightarrow{p} 1$.

For any $\beta_{mar,j} < 0$, since $r_C \xrightarrow{p} 1$, $r_A + r_B = 1 - r_C \xrightarrow{p} 0$ by Slutsky's theorem. Also, $0 \leq \Phi^{-1}(V_n^*), \Phi^{-1}(V_n) \leq 1$, hence, $\forall\ \delta_2 > 0$, $\epsilon_2 > 0$, $\exists\ N_2$, s.t.

$$\begin{aligned} &Pr\left(F(\beta_{mar,j}|y,X) < \epsilon_2\right) = Pr\left(r_A\Phi(V_n^*) + r_B\Phi(V_n) < \epsilon_2\right) \\ &\geq Pr(r_A + r_B < \epsilon_2) > 1 - \delta_2, \text{ whenever } n > N_2. \end{aligned}$$

In other words, $F(\beta_{mar,j}|y,X) \xrightarrow{p} 0$, for any $\beta_{mar,j} < 0$.

Combine them together, if we set $\epsilon_1 = \epsilon_2 = 1/3$, for any $\delta_1 > 0$, $\delta_2 > 0$, there exists $N = \max(N_1, N_2)$, s.t.

$$\begin{aligned} &Pr\left(Pr(\beta_{mar,j} \leq 0|y,X) \geq 1/2 \text{ and } Pr(\beta_{mar,j} \geq 0|y,X) \geq 1/2\right) \\ &\geq Pr\left(F(\beta_{mar,j} = 0|y,X) > 1 - \epsilon_1 \text{ and } F(\beta_{mar,j} < 0|y,X) < \epsilon_2\right) \\ &= 1 - Pr\left(F(\beta_{mar,j} = 0|y,X) \leq 1 - \epsilon_1 \text{ or } Pr(\beta_{mar,j} < 0|y,X) \geq \epsilon_2\right) \\ &> 1 - \delta_1 - \delta_2, \text{ whenever } n > N. \end{aligned}$$

By the definition of median, $Pr\left(\hat{\beta}_{mar,j}^{Med} = 0\right) \to 1$.

(ii) Next, consider $\beta_{mar,j}^0 > 0$.

When $\beta_{mar,j} = 0$, we have $\Phi(V_n^*) \xrightarrow{p} 0$, $\Phi(V_n) \xrightarrow{p} 0$. Since $\hat{\beta}_{mar,j}^{LS} \xrightarrow{p} \beta_{mar,j}^0 > 0$, $r_C \xrightarrow{p} 0$, $0 \leq r_A, r_B \leq 1$, we have $F(\beta_{mar,j} = 0|y,X) = r_A\Phi(V_n^*) + r_B\Phi(V_n) + r_C \xrightarrow{p} 0$, which means $Pr\left(Pr\left(\beta_{mar,j} \leq 0|y,X\right) < 1/2\right) \to 1$. Therefore, by definition, $Pr\left(\hat{\beta}_{mar,j}^{Med} \neq 0\right) \to 1$.

(iii) Lastly, consider $\beta_{mar,j}^0 < 0$.

When $\beta_{mar,j} = -1/\sqrt{n} < 0$, since $\Phi(V_n^*) \xrightarrow{p} 1$, $\Phi(V_n) \xrightarrow{p} 1$, $r_C \xrightarrow{p} 0$, $F\left(\beta_{mar,j} = -1/\sqrt{n}|y,X\right) = r_A\Phi(V_n^*) + r_B\Phi(V_n) \xrightarrow{p} 1$. So, $Pr\left(Pr\left(\beta_{mar,j} \geq 0|y,X\right) < 1/2\right) \to 1$. So, by definition, $Pr\left(\hat{\beta}_{mar,j}^{Med} \neq 0\right) \to 1$.

Combine (i)-(iii), we conclude the selection consistency.

A.3 SIMULATION V: BORROWING INFORMATION ACROSS GROUPS IN SOG

In this simulation, we generated data in SOG setting to evaluate the benefit of borrowing information across groups. We simulated m_{1s} groups with non-zero β's and 10 groups with all β's as zero ($m_1 = m_{1s}+10$). Half of the m_{1s} groups had q features per group (large group); the other half only had $q/4$ features per group (small group); and the other 10 groups had $q/2$ features per group. We did not generate overlapping groups for simplicity, which means $U^{(1)}$ was a block diagonal matrix as below

$$U^{(1)} = \begin{bmatrix} 1_q & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1_q & & & & & & \\ & & & 1_{q/4} & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1_{q/4} & & & \\ & & & & & & 1_{q/2} & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1_{q/2} \end{bmatrix}.$$

Same as simulation 1 in Section 2.4, we modeled the within level-1 group correlation to be 0.5. For m_{1s} groups with non-zero β's, we simulated $\pi_k^{(0)} \sim Beta(10, 10)$, with $\gamma_{jk}^{(0)} \sim Bern(\pi_k^{(0)})$ and $b_{jk} \sim N(0, 5)$. Then, outcomes were generated as $y_i = x_i^T\beta + \epsilon_i$ where $\epsilon_i \sim N(0, 1)$. We varied m_{1s} = 20, 30, 40, q = 20, 30, 40, and set sample size as n = 1000 to compare the performance of SOG with (EB) or without (NonEB) information sharing using empirical Bayes method described in section 2.2. Figure 16 shows the $MSE(\pi^0) = \frac{1}{m_1}\sum_{j=1}^{p}(\hat{\pi}_k^{(0)} - \pi_k^{(0)})^2)$, $MSE(\beta) = \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j^0)^2$, and feature selection AUC.

When large number of groups and reasonable number of variables existed, borrowing information could better estimate $\pi_k^{(0)}$, but the gain in overall model fitting and variable selection was small, possibly because likelihood still had dominant contribution. When the number of groups were small and number of variables inside each group was limited,

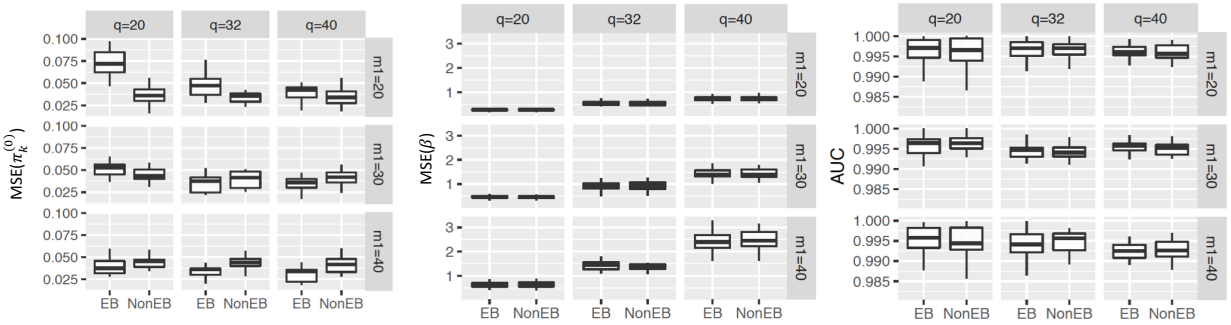


Figure 16: Simulation V results

q determines the number of variables inside each group; m_1 is the total number of groups; EB denotes borrowing information across groups using empirical Bayes method; NonEB denotes not borrowing information. Results are from 100 simulation.

borrowing information (EB) did not help better estimate $\pi_k^{(0)}$, but instead made the estimates unstable, which may be due to limited information to share with.

A.4 TOP 20 MULTI-OMICS FEATURES SELECTED BY MOG IN APPLICATIONS

Table 12: Top 20 multi-omics features selected by MOG in Application ER+ vs. ER- with 123 pathways

Gene name	Multi-omics type	Average FIS
ESR1	mRNA	1
ESR1	methylation	0.59
ESR1	CNV	0.52
SHC2	mRNA	0.20
ADCY9	mRNA	0.20
FKBP4	methylation	0.17
ADCY9	CNV	0.16
SHC2	CNV	0.15
FKBP4	CNV	0.15
FKBP5	CNV	0.15
FKBP5	methylation	0.14
FKBP4	mRNA	0.14
SHC2	methylation	0.14
ADCY9	methylation	0.14
CREB3L4	mRNA	0.13
PEX12	mRNA	0.11
ABAT	mRNA	0.10
MMP2	mRNA	0.09
CREB3L4	CNV	0.09
PXMP2	mRNA	0.09

Multi-omics features are sorted by FIS averaged over 5-fold cross-validation.

Table 13: Top 20 multi-omics features selected by MOG in Application ILC vs IDC with 123 pathways

Gene name	Multi-omics type	Average FIS
CDH1	mRNA	1
LAMA3	mRNA	0.80
CDH1	methylation	0.65
CDH1	CNV	0.64
LAMA3	methylation	0.55
LAMA3	CNV	0.52
DGKD	mRNA	0.43
MAP3K1	mRNA	0.33
APC	CNV	0.33
DGKD	methylation	0.30
DGKD	CNV	0.28
ALDH1B1	mRNA	0.28
APC	methylation	0.28
DTX3	methylation	0.22
PCLO	mRNA	0.22
MAP3K1	CNV	0.22
APC	mRNA	0.21
AKR1B1	mRNA	0.21
AKR1B1	methylation	0.21
MAP3K1	methylation	0.20

Multi-omics features are sorted by FIS averaged over 5-fold cross-validation.

APPENDIX B

APPENDIX FOR METADCN: META-ANALYSIS FRAMEWORK FOR DIFFERENTIAL CO-EXPRESSION NETWORK DETECTION WITH AN APPLICATION IN BREAST CANCER

B.1 METADCNEXPLORER ALGORITHM

MetaDCNExplorer is a Cytoscape application (App) for visualization of differential co-expression networks (DCNs). MetaDNCExplorer utilizes the power of Cytoscape Java API to visualize complex networks. The graphical user interface (GUI) allows users to load input network files and any node/edge attribute tables associated with the networks. Users can manage the imported networks and all aesthetic elements via control panel. MetaDCNExplorer was designed to generate visualization for differential co-expression networks in which nodes represent genes and edges represent co-expression relationships. Each edge should be associated with following two attributes: 1) the directional effect size (e.g., Z-score) and 2) the statistical significance (P-value) of differential co-expression. A gene can belong to one or many modules in a network. Node attributes should specify the module membership of the gene. All above-mentioned attributes, along with the modular network they attached to, are necessary for MetaDCNExplorer, and can be automatically generated from the analysis pipeline of MetaDCN R package. MetaNetworkExplorer was developed in Java programming language and built on OSGi (Open Service Gateway Initiative) Java framework. The implement was based on Cytoscape archetype cyaction-app version 3.0.0 and was built as Bundle App that can be dynamically loaded by Cytoscape main program. By

default the prefuse force-directed layout was used to visualize the modular structure hidden in the input network. This layout is based on the force simulation algorithm implemented as part of the prefuse toolkit (Heer et al., 2005), integrated in the Cytoscape main program. The algorithm positions nodes based on a physics simulation of interacting forces that consist of node repelling force, edge spring force, and air drag forces. The absolute effect size of differential co-expression (i.e., Z-score) reflects the spring length in the simulation. Inter-module repelling factor and intra-module attracting factor is provided for tuning. User can also select either linear force or exponential force. The estimated running time of this layout algorithm on a network with N nodes and E edges will be the grater of O(N log N) and O(E).

B.2 DATA DESCRIPTION AND PREPROCESSING

Eight breast cancer datasets (five training sets and three testing sets) were used for comparing ER+ and ER- patients, including six GEO datasets, The Cancer Genome Atlas (TCGA) breast cancer dataset, and Molecular Taxonomy of Breast Cancer International Consortium dataset (METABRIC) (see Table 8). The TCGA breast cancer dataset was downloaded from the Cancer Genome Atlas (TCGA) website http://tcga-data.nci.nih.gov/tcga in October 2012. Level 3 RNA-Seq data were extracted from the Illumina HiSeq 2000 platform. We selected the TCGA breast cancer dataset that contained expression data of n=406 tumor samples. The METABRIC gene expression and clinical data were retrieved from Synapse https://www.synapse.org/#!Synapse:syn2133309 where we obtained 1981 samples (Curtis et al., 2012). In all studies, microarrays were scanned and summarized by manufacturers' defaults. For the six studies from GEO, data from Affymetrix arrays were processed by robust multi-array (RMA) method and data from Illumina arrays by manufacturer's BeadArray software for probe analysis. Oligonucleotide probes (or probesets) were matched to gene symbols using hgu133plus2.db and illuminaHumanv4.db Bioconductor packages. If multiple probes matched to the same gene, the probe with the largest inter-quartile range (IQR) was used. After matching all the genes across the eight studies, we further filtered away genes

with average standard deviation smaller than 0.2 across all studies, which left 10,636 genes for the following analysis.

Four breast cancer datasets (2 training sets and 2 testing sets) were used for comparing invasive lobular carcinoma (ILC) and invasive ductal carcinoma (IDC) (see Table 9). We included ILC and IDC of Lumina A subtypes from METABRIC and TCGA datasets to gain better homogeneity in patients. TCGA transcript per million (TPM) data were achieved from GSE62944 (Rahman et al., 2015). PAM50 (Parker et al., 2009) subtypes of TCGA patients were called by applying genefu R package (Haibe-Kains et al., 2012), using an ER balanced subsamples for median centering (Curtis et al., 2012). We also included ILCs in a dataset from Sotiriou Lab (Metzger-Filho et al., 2013) and a dataset from Rational Therapy for Breast Cancer (RATHER) consortium (Michaut et al., 2016b) excluding overlapping patients in METABRIC, for validation. The pre-processing step is similar to the previous section. After matching genes and filtering out all the genes with average gene expression or average standard deviation smaller than 50% across two studies, 4552 genes left for following analysis.

All these studies were approved by the University of Pittsburgh Institutional Review Board (IRB PRO16020311).

BIBLIOGRAPHY

Albert, J. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679.

Amar, D., Safer, H., and Shamir, R. (2013). Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS computational biology*, 9(3):1553–7358.

Bartel, D. P. (2009). Micrornas: target recognition and regulatory functions. *Cell*, 136(2):215–233.

Bates, D. and Eddelbuettel, D. (2013). Fast and elegant numerical linear algebra using the RcppEigen package. *Journal of Statistical Software*, 52(5):1–24.

Bhattacharyya, M. and Bandyopadhyay, S. (2013). Studying the differential co-expression of microRNAs reveals significant role of white matter in early Alzheimer's progression. *Molecular bioSystems*, 9(3):457–66.

Burke, K., Tang, P., and Brown, E. (2013). Second harmonic generation reveals matrix alterations during breast tumor progression. *Journal of biomedical optics*, 18(3):31106.

Chen, R.-B., Chu, C.-H., Yuan, S., and Wu, Y. N. (2016). Bayesian sparse group selection. *Journal of Computational and Graphical Statistics*, 25(3):665–683.

Choi, Y. and Kendziorski, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics*, 25(21):2780–2786.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519.

Cunningham, M. and Gilkeson, G. (2011). Estrogen receptors in immunity and autoimmunity. *Clinical Reviews in Allergy and Immunology*, 40(1):66–73.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G.,

Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C., and Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–52.

Eberly, L. E. and Carlin, B. P. (2000). Identifiability and convergence issues for markov chain monte carlo fitting of spatial models. *Statistics in Medicine*, 19(17-18):2279–2294.

Fang, Z., Ma, T., Tang, G., Zhu, L., Yan, Q., Wang, T., Celedón, J. C., Chen, W., and Tseng, G. C. (2018). Bayesian integrative model for multi-omics data with missingness. *Bioinformatics*, 34(22):3801–3808.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486:75–174.

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

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Guiu, S., Wolfer, A., Jacot, W., Fumoleau, P., Romieu, G., Bonnetain, F., and Fiche, M. (2014). Invasive lobular breast cancer and its variants: How special are they for systemic therapy decisions? *Critical reviews in oncology/hematology*, 92(3):235–257.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., and Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, 104(4):311–325.

Heer, J., Card, S. K., and Landay, J. A. (2005). Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430. ACM.

Hernández-Lobato, D., Hernández-Lobato, J. M., Dupont, P., et al. (2013). Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14(1):1891–1945.

Huo, Z., Ding, Y., Liu, S., Oesterreich, S., and Tseng, G. (2016). Meta-analytic framework for sparse k-means to identify disease subtypes in multiple transcriptomic studies. *Journal of the American Statistical Association*, 111(513):27–42.

Huo, Z. and Tseng, G. (2017). Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *The Annals of Applied Statistics*, 11(2):1011.

Ihmels, J., Bergmann, S., Berman, J., and Barkai, N. (2005). Comparative Gene Expression Analysis by a Differential Clustering Approach: Application to the Candida albicans Transcription Program. *PLoS Genetics*, 1(3):e39.

Ishwaran, H. and James, L. F. (2002). Approximate dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical statistics*, 11(3):508–532.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2011). Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12(Jul):2297–2334.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2008). mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic acids research*, 37(suppl_1):D98–D104.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *Annals of Statistics*, pages 1594–1649.

Jorde, L. B. and Wooding, S. P. (2004). Genetic variation, classification and 'race'. *Nature genetics*, 36(11s):S28.

Kim, S., Kang, D., Huo, Z., Park, Y., and Tseng, G. C. (2017). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics*, 34(8):1321–1328.

Kim, S., Lin, C.-W., and Tseng, G. C. (2016). Metaktsp: a meta-analytic top scoring pair method for robust cross-study validation of omics prediction analysis. *Bioinformatics*, 32(13):1966–1973.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by Simulated Annealing. *Science*, 220(4598):671–680.

Kovats, S. (2015). Estrogen receptors regulate innate immune cells and signaling pathways. *Cellular immunology*, 294(2):63–69.

Kugler, K. G., Mueller, L. A. J., Graber, A., and Dehmer, M. (2011). Integrative network biology: graph prototyping for co-expression cancer networks. *PloS one*, 6(7):e22843.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Lai, Y., Wu, B., Chen, L., and Zhao, H. (2004). A statistical method for identifying differential gene–gene co-expression patterns. *Bioinformatics*, 20(17):3146–3155.

Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is My Network Module Preserved and Reproducible? *PLoS Computational Biology*, 7(1):e1001057.

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2003). Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome Research*, 14:1085–1094.

Li, W., Liu, C.-C., Zhang, T., Li, H., Waterman, M. S., and Zhou, X. J. (2011). Integrative Analysis of Many Weighted Co-Expression Networks Using Tensor Computation. *PLoS Computational Biology*, 7(6):e1001106.

Liu, J., Ji, S., Ye, J., et al. (2009). Slep: Sparse learning with efficient projections. *Arizona State University*, 6(491):7.

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

Lock, E. F. and Dunson, D. B. (2017). Bayesian genome-and epigenome-wide association studies with gene level dependence. *Biometrics*, 73(3):1018–1028.

Ma, T., Liang, F., and Tseng, G. C. (2017). Biomarker detection and categorization in ribonucleic acid sequencing meta-analysis using bayesian hierarchical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(4):847–867.

McCarroll, S. A. and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature genetics*, 39:S37.

Mehan, M. R., Nunez-Iglesias, J., Kalakrishnan, M., Waterman, M. S., and Zhou, X. J. (2009). An Integrative Network Approach to Map the Transcriptome to the Phenome. *Journal of computational biology*, 16(8):232–245.

Metzger-Filho, O., Michiels, S., Bertucci, F., Catteau, A., Salgado, R., Galant, C., Fumagalli, D., Singhal, S. K., Desmedt, C., Ignatiadis, M., Haussy, S., Finetti, P., Birnbaum, D., Saini, K. S., Berlière, M., Veys, I., de Azambuja, E., Bozovic, I., Peyro-Saint-Paul, H., Larsimont, D., Piccart, M., and Sotiriou, C. (2013). Genomic grade adds prognostic value in invasive lobular carcinoma. *Annals of Oncology*, 24(2):377–384.

Michaut, M., Chin, S.-F., Majewski, I., Severson, T. M., Bismeijer, T., de Koning, L., Peeters, J. K., Schouten, P. C., Rueda, O. M., Bosma, A. J., et al. (2016a). Integration of

genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Scientific Reports*, 6.

Michaut, M., Chin, S.-F., Majewski, I., Severson, T. M., Bismeijer, T., de Koning, L., Peeters, J. K., Schouten, P. C., Rueda, O. M., Bosma, A. J., Tarrant, F., Fan, Y., He, B., Xue, Z., Mittempergher, L., Kluin, R. J., Heijmans, J., Snel, M., Pereira, B., Schlicker, A., Provenzano, E., Ali, H. R., Gaber, A., O'Hurley, G., Lehn, S., Muris, J. J., Wesseling, J., Kay, E., Sammut, S. J., Bardwell, H. A., Barbet, A. S., Bard, F., Lecerf, C., O'Connor, D. P., Vis, D. J., Benes, C. H., McDermott, U., Garnett, M. J., Simon, I. M., Jirström, K., Dubois, T., Linn, S. C., Gallagher, W. M., Wessels, L. F., Caldas, C., and Bernards, R. (2016b). Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Scientific Reports*, 6(November 2015):18517.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117.

Oskarsson, T. (2013). Extracellular matrix components in breast cancer progression and metastasis. *Breast (Edinburgh, Scotland)*, 22 Suppl 2(2013):S66–72.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(May):1145–1164.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Parker, J. S., Mullins, M., Cheung, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenhush, J. F., Stijleman, I. J., Palazzo, J., Matron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

Phipson, B. and Smyth, G. K. (2010). Permutation P-values Should Never be Zero. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article39.

Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H., and Piccolo, S. R. (2015). Alternative preprocessing of RNA-Sequencing data in the Cancer Genome Atlas leads to improved analysis results. *Bioinformatics*, 31(22):3666–3672.

Ravindranath, A. and Cadigan, K. (2016). The role of the c-clamp in wnt-related colorectal cancers. *Cancers*, 8(8):74.

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and Its Application*, 3:181–209.

Santiago, L., Daniels, G., Wang, D., Deng, F.-M., and Lee, P. (2017). Wnt signaling pathway protein lef1 in cancer, as a biomarker for prognosis and a target for treatment. *American journal of cancer research*, 7(6):1389.

Schübeler, D. (2015). Function and information content of dna methylation. *Nature*, 517(7534):321.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

Sikora, M. J., Jacobsen, B. M., Levine, K., Chen, J., Davidson, N. E., Lee, A. V., Alexander, C. M., and Oesterreich, S. (2016). Wnt4 mediates estrogen receptor signaling and endocrine resistance in invasive lobular carcinoma cell lines. *Breast Cancer Research*, 18(1):92.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the national academy of sciences*, 100(14):8418–8423.

Southworth, L. K., Owen, A. B., and Kim, S. K. (2009). Aging Mice Show a Decreasing Correlation of Gene Expression within Genetic Modules. *PLoS Genetics*, 5(12):e1000776.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). Incorporating biological information into linear models: A bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*, 5(3).

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tasdemir, N., Bossart, E. A., Li, Z., Zhu, L., Sikora, M. J., Levine, K. M., Jacobsen, B. M., Tseng, G. C., Davidson, N. E., and Oesterreich, S. (2018). Comprehensive phenotypic characterization of human invasive lobular carcinoma cell lines in 2d and 3d cultures. *Cancer research*, pages canres–1416.

Teo, K., Gómez-Cuadrado, L., Tenhagen, M., Byron, A., Rätze, M., van Amersfoort, M., Renes, J., Strengman, E., Mandoli, A., Singh, A. A., et al. (2018). E-cadherin loss induces targetable autocrine activation of growth factor signalling in lobular breast cancer. *Scientific reports*, 8(1):15454.

Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., and Caldas, C. (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome biology*, 8(8):R157.

Tesson, B. M., Breitling, R., and Jansen, R. C. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC bioinformatics*, 11:497.

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

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9):3785–3799.

Turashvili, G., Bouchal, J., Baumforth, K., Wei, W., Dziechciarkova, M., Ehrmann, J., Klein, J., Fridman, E., Skarda, J., Srovnal, J., et al. (2007). Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC cancer*, 7(1):55.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–21.

van Hengel, J., Vanhoenacker, P., Staes, K., and Van Roy, F. (1999). Nuclear localization of the p120ctn armadillo-like catenin is counteracted by a nuclear export signal and by e-cadherin expression. *Proceedings of the National Academy of Sciences*, 96(14):7980–7985.

Verhaak, R. G., Wouters, B. J., Erpelinck, C. A., Abbas, S., Beverloo, H. B., Lugthart, S., Löwenberg, B., Delwel, R., and Valk, P. J. (2009). Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*, 94(1):131–134.

Walley, A., Jacobson, P., Falchi, M., Bottolo, L., Andersson, J., Petretto, E., Bonnefond, a., Vaillant, E., Lecoeur, C., Vatin, V., Jernas, M., Balding, D., Petteni, M., Park, Y., Aitman, T., Richardson, S., Sjostrom, L., Carlsson, L., and Froguel, P. (2012). Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue. *International Journal of Obesity*, 36(1):137–147.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational statistics & data analysis*, 52(12):5277–5286.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2012). ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159.

Watson, M. (2006). CoXpress: differential co-expression in gene expression data. *BMC bioinformatics*, 7:509.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726.

Xie, B., Pan, W., and Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, 64(3):921–930.

Xu, X., Ghosh, M., et al. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, B. and Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., and Do, K.-A. (2014a). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(4):595–620.

Zhang, L., Morris, J. S., Zhang, J., Orlowski, R. Z., and Baladandayuthapani, V. (2014b). Bayesian joint selection of genes and pathways: Applications in multiple myeloma genomics. *Cancer informatics*, 13(Suppl 2):113.

Zhao, P., Rocha, G., Yu, B., et al. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497.

Zhu, L., Ding, Y., Chen, C.-Y., Wang, L., Huo, Z., Kim, S., Sotiriou, C., Oesterreich, S., and Tseng, G. C. (2016). Metadcn: meta-analysis framework for differential co-expression network detection with an application in breast cancer. *Bioinformatics*, 33(8):1121–1129.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.