Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Qingpo Cai	Date

Statistical Methods for Biomedical Network Data

By

Qingpo Cai

Doctor of Philosophy

Biostatistics

Jian Kang, Ph.D.
Advisor

Tianwei Yu, Ph.D.
Advisor

Jessica Alvarez, Ph.D.
Committee Member

Zhaohui Qin, Ph.D.
Committee Member

Hao Wu, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for Biomedical Network Data

By

Qingpo Cai
MS, Emory University, 2017
BS, University of Science and Technology of China, 2013

Advisors: Jian Kang, Ph.D., Tianwei Yu, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2018

Abstract

Statistical Methods for Biomedical Network Data

By

Qingpo Cai

There are tens of thousands of units in a biological system. Network representations have been used to describe interactions between these units. Studying biological networks is a key to understand complex biological activities. In this dissertation, we develop statistical methods for analyzing biological network data, aiming to find subnetwork or network marker strongly associated with the clinical outcome of interest.

Selecting informative nodes over large-scale networks becomes increasingly important in many research areas. Most existing methods focus on the local network structure and incur heavy computational costs for the large-scale problem. In the **first project**, we propose a novel prior model for Bayesian network marker selection in the generalized linear model (GLM) framework: the Thresholded Graph Laplacian Gaussian (TGLG) prior, which adopts the graph Laplacian matrix to characterize the conditional dependence between neighboring markers accounting for the global network structure. Under mild conditions, we show the proposed model enjoys the posterior consistency with a diverging number of edges and nodes in the network. We also develop a Metropolis-adjusted Langevin algorithm (MALA) for efficient posterior computation, which is scalable to large-scale networks. We illustrate the superiorities of the proposed method compared with existing alternatives via extensive simulation studies and an analysis of the breast cancer gene expression dataset in the Cancer Genome Atlas (TCGA).

Untargeted metabolomics using high-resolution liquid chromatography - mass spectrometry (LC-MS) is becoming one of the major areas of high-throughput biology. Functional analysis, i.e. analyzing the data based on metabolic pathways or the genome-scale metabolic network, is critical in feature selection and interpretation of metabolomics data. One of the main challenges in the functional analyses is the lack of the feature identity in the LC-MS data itself. By matching mass-to-charge ratio (m/z) values of the features to theoretical values derived from known metabolites, some features can be matched to one or more known metabolites. When multiple matching occurs, in most cases only one of the matchings can be true. At the same time, some known metabolites are missing in the measurements. Current network/pathway analysis methods ignore the uncertainty in metabolite identification and the missing observations, which could lead to errors in the selection of significant subnetworks/pathways. In the **second project**, we propose a flexible network feature selection framework that combines metabolomics data with the genome-scale metabolic network. The method adopts a sequential feature screening procedure and

machine learning-based criteria to select important sub-networks and identify the optimal feature matching simultaneously. Simulation studies show that the proposed method has a much higher sensitivity than the commonly used maximal matching approach. For demonstration, we apply the method on a cohort of healthy subjects to detect subnetworks associated with the Body Mass Index (BMI). The method identifies several subnetworks that are supported by the current literature, as well as detect some subnetworks with plausible new functional implications.

Mediation analysis is a modelling framework to study the relationship between the independent variable (exposure) and the dependent variable (outcome) via including the mediator variable. Traditionally, mediation analysis is developed under regression and causal inference framework, which focuses on measuring or testing the mediation effect. Alternative to existing mediation analysis framework, we propose a new mediation analysis framework focusing on predictive modeling in the **third project**. We propose new definitions for predictive exposure, predictive mediator and predictive network mediator. An estimation procedure is proposed to identify predictive exposure and predictive mediator and simulation studies are conducted to illustrate the performance of the proposed estimation procedure. Two greedy algorithms are proposed to identify network mediator for single and multiple exposure variable and are applied on a dataset from Emory-Georgia Tech Predictive Health Initiative Cohort of the Center for Health Discovery and Well Being.

Statistical Methods for Biomedical Network Data

By

Qingpo Cai
MS, Emory University, 2017
BS, University of Science and Technology of China, 2013

Advisor: Jian Kang, Ph.D., Tianwei Yu, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2018

Acknowledgement

First, I would like to express my deepest gratitude to my advisors, Dr. Jian Kang and Dr. Tianwei Yu, for their enormous help during my graduate study at Emory University. I really appreciate their insightful guidance, generous support and unending encouragement throughout these years. They have always been patient and encouraging whenever I encountered difficulties during my research. I feel so lucky to have them as my advisors. Without their guidance and constant feedback, this PhD would not have been achievable.

I would also like to thank all my committee members, Dr. Jessica Alvarez, Dr. Zhaohui Qin and Dr. Hao Wu, for generously offering their precious time during the whole process. Their valuable suggestions and comments have significantly improved this dissertation work.

I also appreciate the excellent academic support provided by the Department of Biostatistics and Bioinformatics at Emory University. I am grateful to all the faculty members for offering a comprehensive collection of lectures which gets my graduate career started on the right foot. I would also like to thank all the staff members, Bob, Mary, Melissa, for their great service to the department. My sincere thanks also goes to my Emory colleagues and friends, for accompanying me throughout this wonderful Ph.D journey, and for all the memories we have had in the last five years.

Last but not least, I would like to thank my family for their unconditional love and support throughout my whole life.

Contents

List of Figures

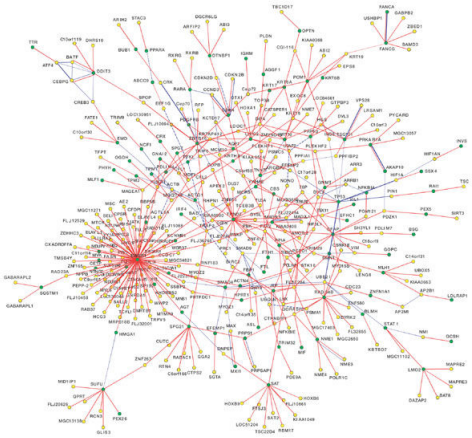
List of Tables

Chapter 1

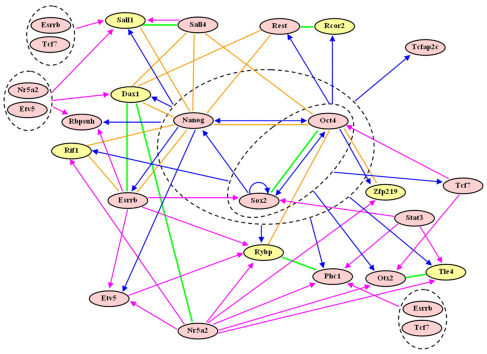
Introduction

1.1 Overview

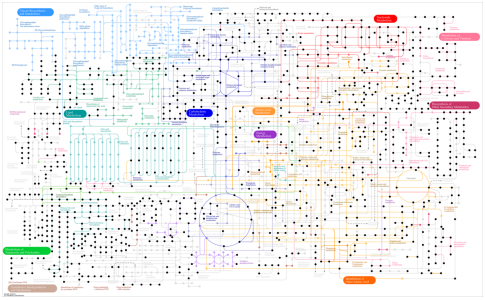
There are tens of thousands of units in a biological system. Network representations have been used to describe interactions between these units. There are various kinds of network in biomedical studies, i.e. protein-protein interaction network (Figure 1.1) (a)), gene regulatory network (Figure 1.1) (b)), Metabolomic network (Figure 1.1) (c)), functional brain network (Figure 1.1) (d)). Studying biological networks is a key to understand complex biological activities (Barabási et al., 2011; Chan and Loscalzo, 2012a; Yu et al., 2013). In this dissertation, we develop statistical methods for analyzing biological network data, aiming to find subnetwork or network marker strongly associated with the clinical outcome of interest.



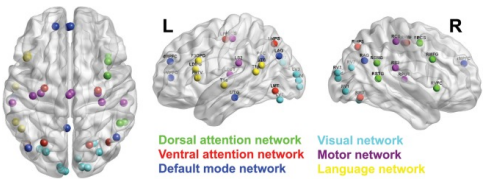
(a) Protein-protein interaction network



(b) Gene regulatory network



(c) Metabolomic network



(d) Functional brain network

Figure 1.1: Examples of different kinds of biomedical networks. (a) Protein-protein interaction network (Rual et al., 2005); (b) Gene regulatory network (Zhou et al., 2007); (c) Metabolomic network (Kyoto Encyclopedia of Genes and Genomes, KEGG); (d) Functional brain network (Schäfer et al., 2014).

1.2 Variable selection methods for genomic network data

In biomedical research, complex biological systems are often modeled or represented as biological networks. High-throughput technology such as next generation sequencing, mass spectrometry and medical imaging has generated massive datasets related to those biological networks. For example, in omics studies, a biological network may represent the interactions or dependences among a large set of genes/proteins/metabolites; and the expression data are a number of observations at each node of the network. In neuroimaging studies, a biological network may refer to the functional connectivity among many brain regions or voxels; and the neural activity can be measured at each node of the network. In many biomedical studies, one important research question is to select informative nodes from tens of thousands of candidate nodes that are strongly associated with the disease risk or other clinical outcomes. We refer to these informative nodes as network markers and the selection procedure as network marker selection. One promising solution is to perform network marker selection under regression framework where the response variable is the clinical outcome and predictors are nodes in the network. The classical variable selection in the regression model can be considered as a special case of the network marker selection, where the variable refers to the nodes in the network that has no edges.

For variable selection in regression models, many regularization methods have been investigated for various penalty terms, including the least absolute shrinkage and selection operator or the L_1 penalty (Tibshirani, 1996; Zou, 2006, LASSO), elastic-net or the L_1 plus L_2 penalty (Zou and Hastie, 2005), the Smoothly Clipped Absolute Deviation penalty (Fan and Li, 2001, SCAD), the minimax concave penalty (Zhang, 2010, MCP) and so on. Several network constrained regularized regression approaches have been developed to improve the selection accuracy and increase the prediction power.

One pioneering work is the graph-constrained estimation (Li and Li, 2008, Grace), which adopts the normalized graph Laplacian matrix to incorporate the network dependent structure between connected nodes. As an extension of Grace, the adaptive Grace (Li and Li, 2010, aGrace) makes constraints on the absolute values of weighted coefficients between connected nodes. Alternatively, an L_γ norm group penalty (Pan et al., 2010) and a fused LASSO type penalty (Luo et al., 2012) have been proposed to penalize the difference of absolute values of coefficients between neighboring nodes. Instead of imposing constraints on coefficients between neighboring nodes, an L_0 loss to penalize their selection indicators (Kim et al., 2013) has been proposed, leading to a non-convex optimization problem for parameter estimation, which can be solved by approximating the non-continuous L_0 loss using the truncated lasso penalty (TLP).

In addition to the frequentist approaches, Bayesian variable selection methods have received much attention recently with many successful applications. The Bayesian methods are natural to incorporate the prior knowledge and make posterior inference on uncertainty of variable selection. A variety of prior models have been studied, such as the spike and slab prior (George and McCulloch, 1993), the LASSO prior (Park and Casella, 2008), the Horseshoe prior (Polson and Scott, 2012), the non-local prior (Johnson and Rossell, 2012) , the Dirichlet Laplace prior (Bhattacharya et al., 2015) and more. To incorporate the known network information, Stingo et al. (2011) proposed Markov Random Field to capture network dependence and to joint select pathways and genes and Chekouo et al. (2016) adopted a similar approach for imaging genetics analysis. Zhou and Zheng (2013) proposed rGrace, a Bayesian random graph-constrained model to combine network information with empirical evidence for pathway analysis. A partial least squares (PLS) g-prior was developed in Peng et al. (2013) to incorporate prior knowledge on gene-gene interactions or functional relationship for identifying genes and pathways. Chang et al. (2016) proposed a Bayesian shrinkage prior which smoothed shrinkage parameters of connected nodes to a similar

degree for structural variable selection. Another commonly used Bayesian structural variable selection method is the Ising model, which has been adopted as a prior model for latent selection indicators that lay on an undirected graph characterizing the local network structure. They are especially successful for variable selection over the grid network motivated by some applications, for example, the motif finding problem (Li and Zhang, 2010) and the imaging data analysis (Goldsmith et al., 2014; Li et al., 2015). However, it is very challenging for fully Bayesian inference on the Ising model over the large-scale network due to at least two reasons: 1) The posterior inference is quite sensitive to the hyperparameter specifications in the Ising priors based on empirical Bayes estimates or subjective prior elicitation in some applications. However, fully Bayesian inference on those parameters is difficult due to the intractable normalizing constant in the model. 2) Most posterior computation algorithms, such as the single-site Gibbs sampler and the Swendsen-Wang algorithm, incur heavy computational costs for updating the massive binary indicators over large-scale networks with complex structures. In addition, Dobra (2009); Kundu et al. (2015); Liu et al. (2014) and Peterson et al. (2016) also proposed Bayesian variable selection approaches for predictors with unknown network structure.

1.3 Metabolomic network data

Metabolomics is the comprehensive analysis of metabolites, i.e. low molecular weight components, in a biological system (Issaq et al., 2009). In recent years, metabolomics has become one of the major areas of interest in high-throughput biology (Zhou et al., 2012; Johnson et al., 2015). There are two general categories of metabolomics: targeted and untargeted. While targeted metabolomics seeks to accurately quantify a limited number of metabolites, untargeted metabolomics seeks to profile the entire metabolome in an unbiased manner (Jones et al., 2012; Sumner et al., 2007). It

helps to discover biomarkers, unravel disease etiology, evaluate systematic response to drugs, and detect environmental chemicals in humans (Nicholson et al., 2008; Zhou et al., 2012).

Untargeted metabolomics is largely made possible by the advances of high-resolution mass spectrometry platforms, which generate highly accurate mass-to-charge ratio (m/z) measurements, greatly facilitating metabolite identification (Patti et al., 2012). Complex preprocessing routines are necessary to ensure high-quality peak detection, quantification, and alignment across profiles (Zhou et al., 2012; Want and Masson, 2011). After alignment, an aligned peak across the LC-MS profiles is referred to as a feature. In the downstream data analysis, a major aspect is functional analysis, i.e. finding pathways or subnetworks that are associated with the clinical outcome (Johnson et al., 2015).

Unlike other omics technologies, metabolomic profiling by LC-MS does not directly provide a critical piece of information - the molecular identities of the features. A single metabolite can produce one or more ion species, due to the presence of various adduct ions, multiple charge states, and isotopic peaks (Kind and Fiehn, 2010). One way to tackle this issue is to first reduce the data by grouping and annotating features derived from the same metabolite (Silva et al., 2014; Kuhl et al., 2012). However this can be difficult for metabolites that exist in low abundance in the biological sample. Hence a common practice in biomarker selection and functional analysis is to match features individually to known metabolites based on mass-to-charge ratio (m/z) (Li et al., 2013a). Retention time can also be used to improve the matching when such information is available on known metabolites, and it can help to determine whether two features are likely to be derived from the same metabolite.

Often a feature can be matched to more than one known metabolite, and a metabolite can be matched to multiple features. This is due to several reasons including (1) some metabolites have the same molecular weight, (2) some features have extremely

close m/z values, and (3) various adduct ions and isotopic peaks are possible. At the same time, due to the sensitivity limits of the technology, some metabolites are not detected in the data. Figure 1.2 displays a sub-network with potentially matched features, which is derived from real data. Several features have been matched to multiple metabolites, and some metabolites have been matched to multiple features.

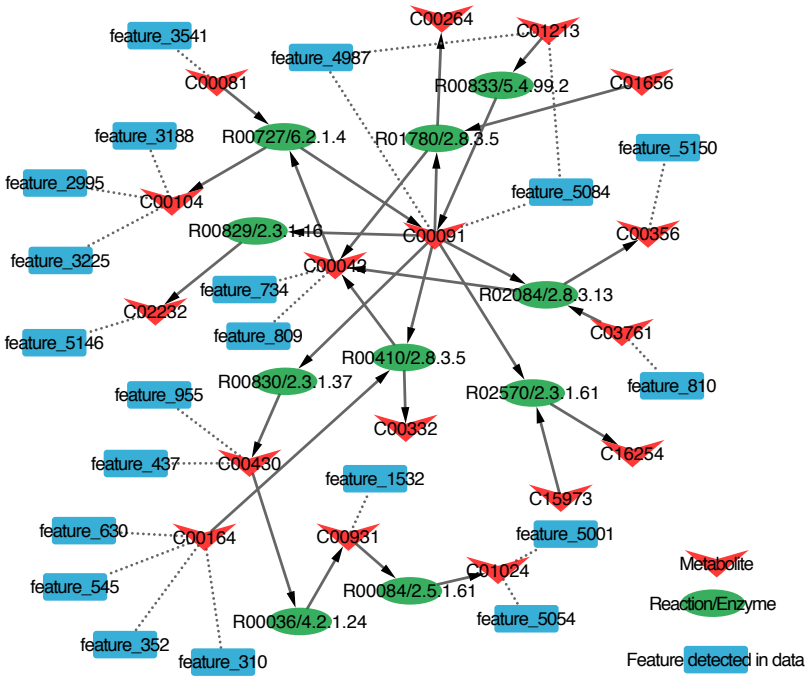


Figure 1.2: An example sub-network presented as a bi-partite network, with potentially matched LC-MS features linked to the metabolites (dotted lines). The cyan nodes represent metabolites, the yellow nodes represent reaction, and dark blue nodes represent LC-MS data features. Four adduct ions are considered: $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, and $[M+NH4]^+$, and the m/z tolerance is 10 ppm.

So far, metabolic pathway analyses are conducted without addressing the matching uncertainty problem (Xia and Wishart, 2010; Kessler et al., 2013; Aggio et al., 2010; Li et al., 2013a; Barupal et al., 2012), i.e. one feature can be matched to more than one metabolite, with only one of the matching being true. In addition, analyzing the biological network directly, without dissecting the overall network artificially into pathways, has been shown to be a very promising approach in other areas of omics

(Chan and Loscalzo, 2012b). Currently for metabolic network analysis, there are few dedicated network analysis methods available (Li et al., 2013a). Although methods can be borrowed from the gene expression field, such methods are not designed to take into account the the matching uncertainty issue.

1.4 Mediation analysis

Mediation analysis is a modelling framework to study the relationship between the independent variable (exposure) and the dependent variable (outcome) via including a third variable, which is defined as the mediator variable. In mediation analysis framework, besides a direct effect between the exposure and the outcome, it's assumed that the exposure has an effect on the mediator, which in turn has an effect on the outcome. Figure 1.3 depicts the typical setting in mediation analysis.

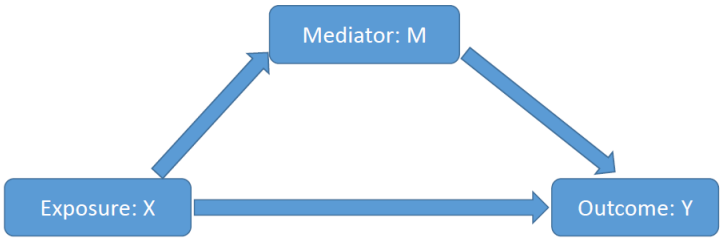


Figure 1.3: A typical setting in mediation analysis

An early approach for mediation analysis was proposed in Baron and Kenny (1986), which is a still quite commonly used approach in behavior sciences and psychological studies (VanderWeele and Vansteelandt, 2009). In Baron and Kenny (1986), three linear regression models need to be estimated for testing mediation effect. See Equation (1)-(3) for details. Equation (1) is regressing the outcome on the exposure. Equation (2) is regressing the mediator on the exposure. Equation (3) is regressing the outcome on both the mediator and the exposure. Several conditions need to be satisfied to establish mediation effect. First, β_{11} in Equation (1) is significant, which

requires the exposure must have an effect on the outcome. Second, β_{21} in Equation (2) is significant, which requires the exposure must have an effect on the mediator. Third, β_{32} in Equation (3) is significant and β_{31} in Equation (3) is smaller than β_{11} in Equation (1), which requires that the mediator must have an effect on the outcome and that part of the effect of the exposure on the outcome is mediated through the mediator. Perfect mediation occurs if β_{31} in Equation (3) is not significant. Within this regression mediation analysis framework, Sobel's test (Sobel, 1982) is a widely used approach to test mediation effect.

$$Y = \beta_{10} + \beta_{11}X + \varepsilon_1 \tag{1.1}$$

$$M = \beta_{20} + \beta_{21}X + \varepsilon_2 \tag{1.2}$$

$$Y = \beta_{30} + \beta_{31}X + \beta_{32}M + \varepsilon_2 \tag{1.3}$$

The above mentioned regression based mediation analysis has been widely utilized in social science and psychology studies, which usually does not imply causal relationship between the exposure and the outcome. Recent advances in mediation analysis has adopted the potential outcome or counterfactual outcome framework in causal inference (Rubin, 1978; Robins and Greenland, 1992; Pearl, 2001) and has been widely used for researchers in biostatistics, epidemiology, causal inference field (VanderWeele, 2016). Consider an exposure X, a mediator M and an outcome Y (Figure 1.3). Variables like $Y(x)$ is defined as the 'potential outcome' or 'counterfactual outcome' (Robins and Greenland, 1992; Pearl, 2001), which denotes the outcome value if exposure X were set to x. Similarly, $M(x), Y(x, m), Y(x, M(x'))$ denote the mediator value if exposure X were set to x, the outcome value if exposure X were set to x and

mediator were set to m, the outcome value if exposure X were set to x and mediator were set to $M(x')$. Given this potential outcome framework, researchers have proposed the following definitions for direct effect, indirect effect and total effect.

- Direct effect: $E[Y(x, M(x)) - Y(x^*, M(x))]$
- Indirect effect: $E[Y(x^*, M(x)) - Y(x^*, M(x^*))]$
- Total effect: $E[Y(x, M(x)) - Y(x^*, M(x^*))] =$
  $\underbrace{E[Y(x, M(x)) - Y(x^*, M(x))]}_{direct} + \underbrace{E[Y(x^*, M(x)) - Y(x^*, M(x^*))]}_{indirect}$

Here direct effect measures the effect of the exposure on the outcome while controlling the mediator and indirect effect measures the effect of the mediator on the outcome while controlling for the exposure. Total effect is the sum of direct effect and indirect effect. This causal inference framework for mediation analysis with only one single mediator has been applied in biostatistics (Albert and Nelson, 2011; Zheng and van der Laan, 2012), epidemiology (Albert, 2012), social sciences (Imai and Yamamoto, 2013) and so on. Built upon this framework for single mediator analysis, researchers have proposed various approaches to extend the causal inference framework that allows multiple mediators by including the interaction effect between the exposure and the mediator and the interaction between the mediators (VanderWeele and Vansteelandt, 2009; Imai et al., 2010; Tchetgen and Shpitser, 2012; van der Laan and Petersen, 2008; VanderWeele and Vansteelandt, 2014; Daniel et al., 2015; Huang et al., 2014).

1.5 Outline

In this dissertation, we propose several novel statistical methods to analyze biomedical network data. In chapter 2, we introduce the thresholded graph Laplacian Gaussian (TGLG) prior and propose a model for variable selection with incorporating network

structure under the GLM framework. In chapter 3, we propose a framework for subnetwork selection and addressing the multiple matching issue with application to metabolomics data. In chapter 4, we propose a new mediation analysis framework focusing on predictive modeling. We end this dissertation with a brief discussion and future work in chapter 5.

Chapter 2

Bayesian network marker selection via the thresholded graph Laplacian Gaussian prior

2.1 Introduction

In this chapter, we propose a new prior model: the thresholded graph Laplacian Gaussian (TGLG) prior, to perform network marker selection over the large-scale network by thresholding a latent continuous variable that is attached to each node. The joint distribution of all the latent variables is a multivariate Gaussian distribution with mean zero and covariance matrix constructed by the normalized graph Laplacian matrix to model the selection dependence over the network. The effect size of each node is modeled through an independent Gaussian distribution.

Threshold priors have been proposed for Bayesian modeling of sparsity in various applications. Motivated by the analysis of financial time series data, Nakajima and West (2013a) and Nakajima and West (2013b) proposed a latent threshold approach to imposing dynamic sparsity in the simultaneous autoregressive models (SAR). Nakajima et al. (2017) further extended this type of models for the analysis of EEG data. To analyze neuroimaging data, Shi and Kang (2015) proposed a hard-thresholded Gaussian process prior for image-on-scalar regression; and Kang et al. (2018) introduced a soft-thresholded Gaussian process for scalar-on-image regression. To construct the directed graphs in genomics applications, Ni et al. (2017) adopted a hard threshold Gaussian prior in a structural equation model. However, all the existing threshold prior models do not incorporate the useful network structural information, and thus are not directly applicable to the network marker selection problem.

In this work, we propose to build the threshold priors using the graph Laplacian matrix, which has been used to capture the structure dependence between neighboring nodes (Li and Li, 2008; Zhe et al., 2013; Li and Li, 2010). Most of those frequents methods directly specify the graph Laplacian matrix from the existing biological network. Liu et al. (2014) has proposed a Bayesian regularization graph Laplacian (BRGL) approach which utilizes the graph Laplacian matrix to specify *a priori* precision matrix of regression coefficients. However, BRGL is fundamentally

different from our method in that it is one type of continuous shrinkage priors for regression coefficients which have quite different prior supports compared with our TGLG priors. BRGL were developed only for linear regression and its computational cost can be extremely heavy for large-scale networks. In addition, there is lack of theoretical justifications for BRGL for large-scale networks with a diverging number of edges and nodes.

Our method is a compelling Bayesian approach to network marker selection over large-scale networks. The TGLG prior has at least four markable features: 1) Fully Bayesian inference for large-scale networks is feasible in that the TGLG prior does not involve any intractable normalizing constants. 2) Posterior computation can be more efficient, since the TGLG-based inference avoids updating the latent binary selection indicators and instead updates the latent continuous variables, to which many existing approximation techniques can be potentially applied. 3) The graph Laplacian matrix (Chung, 1997; Li and Li, 2008; Zhe et al., 2013) based prior can incorporate the topological structure of the network which has been adopted in genomics. 4) The TGLG prior enjoys the large support for Bayesian network marker selection over large-scale networks, leading to posterior consistency of model inference with a diverging number of nodes and edges under the generalized linear model (GLM) framework.

The remainder of this chapter is organized as follows. In Section 2.2, we introduce the TGLG prior and propose our model for network marker selection under the GLM framework. In Section 2.3, we study the theoretical properties for the TGLG prior and show the posterior consistency of model inference. In Section 2.4, we discuss the hyper prior specifications and the efficient posterior computation algorithm. We illustrate the performance of our approach via simulation studies and an application on the breast cancer gene expression dataset from The Cancer Genome Atlas (TCGA) in Section 2.5. We conclude our paper with a brief discussion in Section 2.6.

2.2 The model

Suppose the observed dataset includes a network with p_n nodes, one response variable and q confounding variables. For each node, we have n observations. For observation i, $i = 1, \ldots, n$, let y_i be the response variable, $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip_n})^{\mathrm{T}}$ be the vector of nodes and $\mathbf{z}_i = (z_{i1}, \cdots, z_{iq})^{\mathrm{T}}$ be the vector of confounding variables. Denote by $D_n = \{\mathbf{z}_i, \mathbf{x}_i, y_i\}_{i=1}^n$ the dataset. We write the number of nodes as p_n to emphasize on the diverging number of nodes in our asymptotical theory. Drop subscript i to have generic notation for a response variable y, a vector of nodes $\mathbf{x}$ and a vector of confounders $\mathbf{z}$. Generalized linear model (GLM) is a flexible regression model to relate a response variable to a vector of nodes and confounding variables. The GLM density function for $(y, \mathbf{x}, \mathbf{z})$ with one natural parameter is:

$$f^*(y, h^*) = \exp\{a(h^*)y + b(h^*) + c(y)\}, \tag{2.1}$$

where $h^* = \mathbf{z}^{\mathrm{T}}\boldsymbol{\omega}^* + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*$ is the linear parameter in the model, $\boldsymbol{\omega}^*$ and $\boldsymbol{\beta}^*$ are true coefficients that generate data, $a(h)$ and $b(h)$ are continuous differentiable functions. The true mean function is

$$\mu^* = \mathrm{E}(y \mid \mathbf{z}, \mathbf{x}) = -b'(h^*)/a'(h^*) \equiv g^{-1}(\mathbf{z}^{\mathrm{T}}\boldsymbol{\omega}^* + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*),$$

where $g^{-1}(\cdot)$ is an inverse link function, which can be chosen according to the specific type of the response variable. For example, one can choose the identity link for the continuous response and the logit link for the binary response.

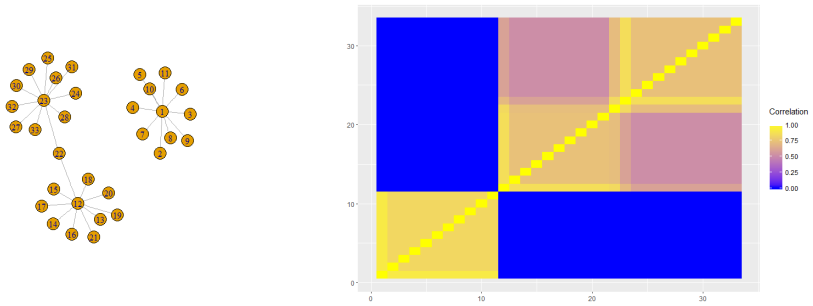
In (2.1), coefficient vector $\boldsymbol{\omega}$ is a nuisance parameter to adjust for the confounder effects, for which we assign a Gaussian prior with mean zero and independent covariance, i.e. $\boldsymbol{\omega} \sim \mathrm{N}(0, \sigma_\omega^2 \mathbf{I}_q)$ for $\sigma_\omega^2 > 0$. Here $\mathbf{I}_d$ represents an identity matrix of dimension d for any $d > 0$. Coefficient vector $\boldsymbol{\beta}$ represents the effects of nodes on the

response variable. Here we perform network marker selection by imposing sparsity on $\boldsymbol{\beta}$. To achieve this goal, we develop a new prior model for $\boldsymbol{\beta}$: the thresholded graph Laplacian Gaussian (TGLG) prior. Suppose the observed network can be represented by a graph G, with each vertex corresponding to one node in the network. Let $j \sim k$ indicate there exists an edge between vertices j and k in G. Let d_j represent the degree of vertex j, i.e., the number of nodes that are connected to vertex j in G. Denote by $\mathbf{L} = (L_{jk})$ a $p_n \times p_n$ normalized graph Laplacian matrix, where $L_{jk} = 1$ if $j = k$ and $\deg(v_j) \neq 0$, $L_{jk} = -1/\sqrt{d_j d_k}$ if $j \sim k$, and $L_{jk} = 0$ otherwise. For any $d > 0$, denote by $\mathbf{0}_d$ an all zero vector of dimension d. For any $\lambda, \varepsilon, \sigma_\alpha^2, \sigma_\gamma^2 > 0$, we consider an element-wise decomposition of $\boldsymbol{\beta}$ for the prior specifications:

$$\boldsymbol{\beta} = \boldsymbol{\alpha} \circ \mathbf{t}_\lambda(\boldsymbol{\gamma}), \quad \boldsymbol{\gamma} \sim \mathrm{N}\{\mathbf{0}_{p_n}, \sigma_\gamma^2(\mathbf{L} + \varepsilon \mathbf{I}_{p_n})^{-1}\}, \quad \boldsymbol{\alpha} \sim \mathrm{N}(\mathbf{0}_{p_n}, \sigma_\alpha^2 \mathbf{I}_{p_n}). \tag{2.2}$$

Here $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p_n})^{\mathrm{T}}$ represents the effect size of nodes. The operator "$\circ$" is the element-wise product. The vector thresholding function is $\mathbf{t}_\lambda(\boldsymbol{\gamma}) = \{I(|\gamma_1| > \lambda), \ldots, I(|\gamma_{p_n}| > \lambda)\}^{\mathrm{T}}$, where $I(\mathcal{A})$ is the event indicator with $I(\mathcal{A}) = 1$ if $\mathcal{A}$ occurs and $I(\mathcal{A}) = 0$ otherwise. The latent continuous vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_n})^{\mathrm{T}}$ controls the sparsity over graph G. We refer to (2.2) as the TGLG prior for $\boldsymbol{\beta}$, denoted as $\boldsymbol{\beta} \sim \mathrm{TGLG}(\lambda, \varepsilon, \sigma_\gamma^2, \sigma_\alpha^2)$. The TGLG prior implies that for any two nodes j and k, γ_j and γ_k are conditionally dependent given others if and only if $j \sim k$ over the graph G. In this case, their absolute values are more likely to be smaller or larger than a threshold value λ together. This further implies that nodes j and k are more likely to be selected as network marker or not selected together if $j \sim k$. Figure 2.1 shows an example of a graph and the corresponding correlation matrix of $\boldsymbol{\gamma}$ for $\varepsilon = 10^{-2}$, where the γ's of connected vertices are highly correlated.

There are four hyperparameters in the TGLG prior model. The threshold λ controls *a priori* the sparsity. When $\lambda \to 0$, all the nodes tend to be selected. When



(a) network (b) correlation

Figure 2.1: An example of the graph and the corresponding correlation matrix of $\boldsymbol{\gamma}$ that was constructed from the inverse graph Laplacian matrix

$\lambda \to \infty$, none of them will be selected. The parameter ε determines the impact of the network on the sparsity. When $\varepsilon \to \infty$, γ's of connected vertices tend to be independent while they tend to be perfectly correlated when $\varepsilon \to 0$. The two variance parameters σ_γ^2 and σ_α^2 control the prior variability of the latent vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ respectively.

Now we discuss how to specify the hyperparameters. For variance terms σ_γ^2 and σ_α^2, we use the conjugate prior model by assigning the Inverse-Gamma distribution $\text{IG}(a_\gamma, b_\gamma)$ and $\text{IG}(a_\alpha, b_\alpha)$ respectively. We fix σ_ω^2 as a large value. We assign the uniform prior to the threshold parameter λ, i.e. $\lambda \sim \text{Unif}(0, \lambda_u)$ with upper bound $\lambda_u > 0$. We choose a wide range by set $\lambda_u = 10$ in the following content. For parameter ε, we can either assign an log-normal prior ($\log\varepsilon \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$) or set as a fixed small value.

2.3 Theoretical Properties

In this section, we examine the theoretical properties of TGLG prior based network marker selection under the GLM framework. In particular, we establish the posterior consistency with a diverging number of nodes in the large-scale networks.

Let $\xi \subset \{1, 2, \cdots, p_n\}$ denote the set of selected node indices, i.e. $I(|\gamma_j| > \lambda) = 1$,

if $j \in \xi$, $I(|\gamma_j| > \lambda) = 0$, otherwise. The number of nodes in ξ is denoted as $|\xi|$. For a model $\xi = (i_1, \cdots, i_{|\xi|})$, denote by $\boldsymbol{\beta}_\xi = (\beta_{i_1}, \cdots, \beta_{i_{|\xi|}})^{\mathrm{T}}$ the coefficient of interest, respectively. Let $\pi(\xi, d\boldsymbol{\beta}_\xi, d\boldsymbol{\omega})$ represent the joint prior probability measure for model ξ, parameters $\boldsymbol{\beta}_\xi$ and confounding coefficients $\boldsymbol{\omega}$. Their joint posterior probability measure conditional on dataset D_n is:

$$\pi(\xi, d\boldsymbol{\beta}_\xi, d\boldsymbol{\omega} \mid D_n) = \frac{\prod_{i=1}^n f(y_i, h_i)\pi(\xi, d\boldsymbol{\beta}_\xi, d\boldsymbol{\omega})}{\sum_{\xi'} \int_{\boldsymbol{\beta}_{\xi'}} \prod_{i=1}^n f(y_i, h_i)\pi(\xi', d\boldsymbol{\beta}'_\xi, d\boldsymbol{\omega})},$$

where $h_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\omega} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. We examine the asymptotic properties of posterior density function regarding to the Hellinger distance (Jiang, 2007; Song and Liang, 2015) under appropriate regularity conditions. The Hellinger distance $d(f_1, f_2)$ between two density functions $f_1(x, y)$ and $f_2(x, y)$ is defined as

$$d(f_1, f_2) = \left[\int\int \{f_1^{1/2}(x, y) - f_2^{1/2}(x, y)\} dx dy\right]^{1/2}.$$

We list all the regularity conditions in the Appendix. We show that the TGLG prior enjoys the following properties:

Theorem 1. *(Large Support for Network Marker Selection) Assume a sequence $\epsilon_n \in (0, 1]$ with $n\epsilon_n^2 \to \infty$ and a sequence of nonempty models ξ_n. Assume conditions (C1)–(C3) and (C7) hold. Given σ_α^2 and σ_γ^2, for any sufficiently small $\eta > 0$, there exists N_η such that for all $n > N_\eta$, we have*

$$\pi(\xi = \xi_n) \geq e^{-n\epsilon_n^2/128} \text{ and} \tag{2.3}$$

$$\pi(\boldsymbol{\beta}_\xi \in B(\xi_n, \eta) \mid \xi = \xi_n) \geq e^{-n\epsilon_n^2/128} \text{ with } B(\xi_n, \eta) = \{\beta_j^* \pm \eta\epsilon_n^2/|\xi_n|\}_{j \in \xi_n}. \tag{2.4}$$

There exists $C_n > 0$, such that for all sufficiently large n and for any $j \in \xi_n$:

$$\pi(|\beta_j| > C_n \mid \xi_n) \leq e^{-n\epsilon_n^2/4}. \tag{2.5}$$

This theorem shows that the TGLG prior has a large support for the network marker selection. Particularly, (2.3) states that the TGLG prior can select the true network marker with a positive prior probability bounded away from zero, (2.4) ensures that the prior probability of the coefficients falling within an arbitrarily small neighborhood of the true coefficients with probability bounded away from zero, and (2.5) indicates a sufficiently small tail probability of the TGLG prior.

Theorem 2. *(Posterior Consistency for Network Marker Selection) For the GLM with bounded covariates, i.e.* $|x_j| \leq M$ *for all* $j = 1, \cdots, p_n$ *and* $|z_k| \leq M$ *for all* $k = 1, \cdots, q$*, suppose the true node regression coefficients satisfy*

$$\lim_{n\to\infty} \sum_{j=1}^{p_n} |\beta_j^*| < \infty.$$

Let $\epsilon_n \in (0, 1]$ *be a sequence such that* $n\epsilon_n^2 \to \infty$*. Assume conditions (C1)–(C7) hold. Then we have the following results:*

(i) $\lim_{n\to\infty} \mathrm{P}\{\pi[d(f, f^*) \leq \epsilon_n | D_n] \geq 1 - 2e^{-n\epsilon_n^2/64}\} = 1.$

(ii) For all sufficiently large n: $\mathrm{P}\{\pi[d(f, f^*) > \epsilon_n | D_n] \geq 2e^{-n\epsilon_n^2/64}\} \leq 2e^{-n\epsilon_n^2/64}.$

(iii) For all sufficiently large n: $\mathrm{E}\{\pi[d(f, f^*) > \epsilon_n | D_n]\} \leq 4e^{-n\epsilon_n^2/32}$*, where* $d(f, f^*)$ *is the Hellinger distance between the true density* f^* *and the density function* f *simulated from posterior. Probability measure* P *and expectation* E *are both with respect to data* D_n*.*

This theorem shows that as sample size n goes to infinity, density f sampled from the posterior converges to true density f^* with regarding to Hellinger distance. Please refer to the Appendix for proofs for both Theorems.

2.4 Posterior Computation

For the nuisance parameter $\boldsymbol{\omega}$, it can be estimated by its posterior expectation $E(\boldsymbol{\omega})$. Our interest is to estimate the regression coefficients for nodes $\boldsymbol{\beta}$. According to the model specification, it is straightforward to see that $\beta_j = 0$ for $\{j : |\gamma_j| \leq \lambda\}$. For $\{j : |\gamma_j| > \lambda\}$, we estimate β_j by posterior expectation $E(\beta_j||\gamma_j| > \lambda)$.

We adopt an efficient Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Rosenthal, 1998) for posterior computation. We introduce a smooth approximation for the thresholding function:

$$I(|\gamma_j| > \lambda) \simeq \frac{1}{2}\left\{1 + \frac{2}{\pi}\arctan(\frac{\gamma_j^2 - \lambda^2}{\varepsilon_0})\right\} \quad \text{for} \quad \varepsilon_0 \to 0,$$

leading to the analytically tractable first derivative:

$$\frac{\partial \beta_j}{\partial \gamma_j} = \alpha_j \frac{2\gamma_j/\varepsilon_0}{\pi(1 + (\gamma_j^2 - \lambda^2)^2/\varepsilon_0^2)}.$$

We choose $\varepsilon_0 = 10^{-8}$ in the simulation studies and real data application in this article.

Denote by $f(y_i \mid \boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \lambda)$ the likelihood function for all the parameters of interests for observation i. Let $\phi(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\phi_+(x \mid \mu, \mu_l, \mu_u, \sigma^2)$ denote the density of a truncated normal distribution $N_+(\mu, \mu_l, \mu_u, \sigma^2)$ density with mean μ, variance σ^2 and interval $[\mu_l, \mu_u]$. Let $V_\omega = \sigma_\omega^2 I_q$ be the variance of the prior distribution for $\boldsymbol{\omega}$. Let $\Lambda_\gamma = (\mathbf{L} + \varepsilon \mathbf{I}_{p_n})^{-1}$. The key steps in our posterior computation algorithm include:

- Update $\boldsymbol{\omega}$ (Random Walk): Given current $\boldsymbol{\omega}$, Draw $\boldsymbol{\omega}^{new} \sim \mathrm{N}(\boldsymbol{\omega}, \tau_\omega^2 I_q)$. Set $\boldsymbol{\omega} \longleftarrow \boldsymbol{\omega}^{new}$ with probability $\min\left\{1, \frac{\phi(\boldsymbol{\omega}^{new}|0,V_\omega)\prod_i f(y_i|\boldsymbol{\omega}^{new},\bullet)}{\phi(\boldsymbol{\omega}|0,V_\omega)\prod_i f(y_i|\boldsymbol{\omega},\bullet)}\right\}$.

- Update $\boldsymbol{\gamma}$ (MALA): Given current $\boldsymbol{\gamma}$, draw $\boldsymbol{\gamma}^{new} \sim \mathrm{N}\{\mu(\boldsymbol{\gamma}), \tau_\gamma^2 I_p\}$, where $\mu(\boldsymbol{\gamma}) = \boldsymbol{\gamma} + \frac{\tau_\gamma^2}{2}(\frac{\partial \log f}{\partial \boldsymbol{\gamma}} - \frac{1}{2}\sigma_\gamma^2 \Lambda_\gamma \boldsymbol{\gamma})$ with $\frac{\partial \log f}{\partial \gamma_j} = \sum_{i=1}^n (a'(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\omega} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + b'(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\omega} +$

$\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}))x_{ij}\frac{\partial\beta_j}{\partial\gamma_j}$. Set $\boldsymbol{\gamma} \longleftarrow \boldsymbol{\gamma}^{new}$ with probability min $\left\{1, \frac{\phi(\boldsymbol{\gamma}|\mu(\boldsymbol{\gamma}^{new}),\tau_\gamma^2 I_p)\phi(\boldsymbol{\gamma}^{new}|0,\sigma_\gamma^2\Lambda_\gamma)\prod_i f(y_i|\boldsymbol{\gamma}^{new},\bullet)}{\phi(\boldsymbol{\gamma}^{new}|\mu(\boldsymbol{\gamma}),\tau_\gamma^2 I_p)\phi(\boldsymbol{\gamma}|0,\sigma_\gamma^2\Lambda_\gamma)\prod_i f(y_i|\boldsymbol{\gamma},\bullet)}\right\}$.

- Update ξ: Given $\boldsymbol{\gamma}$ and λ, update $\xi = \{j : \gamma_j > \lambda\}$.

- Update $\boldsymbol{\alpha}$ (MALA): For $j \notin \xi$, sample $\alpha_j \sim \mathrm{N}(0, \sigma_\alpha^2)$. Draw $\boldsymbol{\alpha}_\xi^{new} \sim \mathrm{N}\left\{\mu(\boldsymbol{\alpha}_\xi), \tau_\alpha^2 I_{|\xi|}\right\}$, where $\mu(\boldsymbol{\alpha}_\xi) = \boldsymbol{\alpha}_\xi + \frac{\tau_\alpha^2}{2}(\frac{\partial \log f}{\partial \boldsymbol{\alpha}_\xi} - \frac{1}{2}\Sigma_\xi\boldsymbol{\alpha}_\xi)$ with $\frac{\partial \log f}{\partial \alpha_j} = \sum_{i=1}^n (a'(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\omega} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + b'(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\omega} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}))x_{ij}$ for $j \in \xi$ and $\Sigma_\xi = \sigma_\alpha^2 I_{|\xi|}$. Update $\boldsymbol{\alpha}_\xi \longleftarrow \boldsymbol{\alpha}_\xi^{new}$ with probability

  $\min\left\{1, \frac{\phi(\boldsymbol{\alpha}_\xi|\mu(\boldsymbol{\alpha}_\xi^{new}),\tau_\alpha^2 I_{|\xi|})\phi(\boldsymbol{\alpha}_\xi^{new}|0,\Sigma_\xi)\prod_i f(y_i|\boldsymbol{\alpha}_\xi^{new},\bullet)}{\phi(\boldsymbol{\alpha}_\xi^{new}|\mu(\boldsymbol{\alpha}_\xi),\tau_\alpha^2 I_{|\xi|})\phi(\boldsymbol{\alpha}_\xi|0,\Sigma_\xi)\prod_i f(y_i|\boldsymbol{\alpha}_\xi,\bullet)}\right\}$.

- Update σ_γ^2: Draw $\sigma_\gamma^2 \sim \mathrm{IG}(a^\gamma, b^\gamma)$ where $a^\gamma = a_\gamma + \frac{p}{2}$ and $b^\gamma = b_\gamma + \frac{\boldsymbol{\gamma}^{\mathrm{T}}\Lambda_\gamma^{-1}\boldsymbol{\gamma}}{2}$.

- Update σ_α^2: Draw $\sigma_\alpha^2 \sim \mathrm{IG}(a^\alpha, b^\alpha)$ where $a^\alpha = a_\alpha + \frac{p}{2}$ and $b^\alpha = b_\alpha + \frac{\sum_j \alpha_j^2}{2}$.

- Update ε (Random Walk, optional) Draw $\varepsilon^{new} \sim N(\varepsilon, \tau_\varepsilon^2)$. Update $\varepsilon \longleftarrow \varepsilon^{new}$ with probability $\min\left\{1, \frac{|\mathbf{L}+\varepsilon^{new}\mathbf{I}_{p_n}|^{\frac{1}{2}}\frac{1}{\varepsilon^{new}}\exp(-\frac{\varepsilon^{new}\boldsymbol{\gamma}^T\boldsymbol{\gamma}}{2\sigma_\gamma^2} - \frac{(log\varepsilon^{new}-\mu_\varepsilon)^2}{2\sigma_\varepsilon^2})}{|\mathbf{L}+\varepsilon\mathbf{I}_{p_n}|^{\frac{1}{2}}\frac{1}{\varepsilon}\exp(-\frac{\varepsilon\boldsymbol{\gamma}^T\boldsymbol{\gamma}}{2\sigma_\gamma^2} - \frac{(log\varepsilon-\mu_\varepsilon)^2}{2\sigma_\varepsilon^2})}\right\}$.

- Update λ: Given λ, draw $\lambda^{new} \sim N_+(\lambda, \lambda_l, \lambda_u, \sigma_l^2)$. Set $\lambda \longleftarrow \lambda^{new}$ with probability min $\left\{1, \frac{\phi_+(\lambda|\lambda^{new},\lambda_l,\lambda_u,\sigma_l^2)\prod_i f(y_i|\lambda^{new},\bullet)}{\phi_+(\lambda^{new}|\lambda,\lambda_l,\lambda_u,\sigma_l^2)\prod_i f(y_i|\lambda,\bullet)}\right\}$.

The proposal variances τ_γ^2, τ_α^2 and τ_ω^2 are all adaptively chosen by tuning acceptance rates to 30% for random walk and 50% for MALA (Roberts and Rosenthal, 1998).

Denote by $\boldsymbol{\gamma}^{(i)}, \boldsymbol{\alpha}^{(i)}, \lambda^{(i)} (i = 1, \cdots, N)$ the MCMC samples obtained after burn-in. We estimate the posterior inclusion probability for node $j (j = 1, \cdots, p_n)$ by

$$\widehat{\Pr}(\beta_j \neq 0 \mid D_n) = \frac{1}{N}\sum_{i=1}^{N} I\{|\gamma_j^{(i)}| > \lambda^{(i)}\}.$$

According to Barbieri et al. (2004), we select the informative nodes with at least 50% inclusion probability, denote by $\widehat{M} = \{j : \widehat{\Pr}(\beta_j \neq 0 \mid D_n) > 0.5\}$ the indices of all

the informative nodes. To estimate regression coefficients of informative nodes, we choose the estimated conditional expectation of β_j given $\beta_j \neq 0$ by

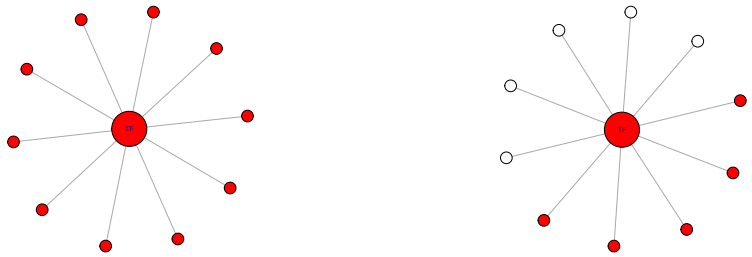
$$\widehat{\mathrm{E}}\{\beta_j \mid \beta_j \neq 0, D_n\} = \frac{\sum_{i=1}^{N} \alpha_j^{(i)} I(|\gamma_j^{(i)}| > \lambda^{(i)})}{\sum_{i=1}^{N} I\{|\gamma_j^{(i)}| > \lambda^{(i)}\}}, \text{for } j \in \widehat{M}.$$

2.5 Numerical Studies

2.5.1 Simulation studies

We conducte simulation studies to evaluate performance of the proposed methods compared with existing methods for many different scenarios.

To generate the networks and observations on nodes, we follow the simulation settings in Li and Li (2008), Zhe et al. (2013) and Kim et al. (2013). See Figure 2.2 for network structure. We simulate gene networks consisting of m subnetworks where each subnetwork contains one transcription factor (TF) gene and ten target genes that are connected to the TF gene. The TF gene expression levels, denoted by X_{TF}, is generated from the standard normal distribution. Given the X_{TF}, the target gene expression data, denoted by X_{tg}, are independently sampled from a normal distribution with mean $0.5X_{\mathrm{TF}}$ and variance 0.75. This implies that the marginal correlation between X_{TF} and X_{tg} is 0.5. Two types of the true network markers are considered in Simulation 1. Type 1 network markers include one TF gene and its connected targets; see Figure 2.2(a); Type 2 network markers have one TF gene along with only part of its connected target genes, see Figure 2.2(b). In Type 1 network, the model assumptions are satisfied since if one TF gene has signal, all its connected neighbors also have signal. In Type 2 network, when one TF gene has signal, only part of its connected neighbors have signal. For Simulation 1, we consider two numbers of subnetworks: $m = 3$ and $m = 10$. The coefficients are random generated from Unif$(1, 3)$ and then randomly assigned as positive or negative.



(a) Simulation 1 Type 1 (b) Simulation 1 Type 2

Figure 2.2: Network settings in simulation studies Case 1. Red means network markers in the true subnetwork. (a) and (b) shows the structure of a subnetwork.

To generate the response variable given the true network markers, we consider binary and continuous cases, where the continuous response variable are generated from linear regression, i.e. $y \sim \mathrm{N}(X\boldsymbol{\beta}, \sum_i \beta_i^2/3)$; and the binary response are generated from logistic regression, i.e. $\Pr(y = 1) = 1/\{1 + \exp(-X\boldsymbol{\beta})\}$.

We generate 50 datasets for each scenario. For linear regression, each dataset contains 100 training samples and 100 test samples; for logistic regression, each dataset contains 200 training samples and 200 test samples.

We compare the proposed TGLG approach with the following existing methods: Lasso (Tibshirani, 1996), Elastic-net (Zou and Hastie, 2005), Grace (Li and Li, 2008), aGrace (Li and Li, 2010), L_∞ and aL_∞ (Luo et al., 2012), TTLP and LTLP (Kim et al., 2013), BRGL (Liu et al., 2014) and Ising model (Goldsmith et al., 2014; Li et al., 2015). For the hyper priors in the TGLG approach, we assign weakly informative priors: $\sigma_\gamma^2 \sim \mathrm{IG}(0.01, 0.01), \sigma_\alpha^2 \sim \mathrm{IG}(0.01, 0.01)$. For all the regularized approaches, we adopt 3-fold cross validations to choose tuning parameters. For the Ising prior model, we specify the priors as

$$p(\boldsymbol{\gamma}) = \phi(a, b) \exp\left[a \sum_i \gamma_i + \sum_i \{\sum_{j \in N_i} bI(\gamma_i = \gamma_j)\}\right]$$

and $\beta_i|\gamma_i = 1 \sim \mathrm{N}(0, \sigma_\beta^2)$, where N_i denotes the neighbor nodes set of node i. For hyper prior specifications in Ising model, we fix $a = -2$ and choose b from $2, 5, 7$ and 10 based on model performance. We implement a single-site Gibbs sampler for Ising model. For the parameter $\varepsilon(\boldsymbol{\gamma} \sim \mathrm{N}\{\mathbf{0}_{p_n}, \sigma_\gamma^2(\mathbf{L} + \varepsilon\mathbf{I}_{p_n})^{-1}\})$ in TGLG, we consider to fix $\varepsilon = 10^{-5}$ and to update ε with a prior $log\varepsilon \sim N(-5, 9)$. To check the influence of including the known network information, we also consider an i.i.d prior for selection variable as $\boldsymbol{\gamma} \sim \mathrm{N}\{\mathbf{0}_p, \sigma_\gamma^2 I_p)$, which means network information is not included. For the posterior computation of the Ising model and TGLG, we ran $30,000$ MCMC iterations with the first $20,000$ as burn-in. For BRGL by Liu et al. (2014), the network markers are selected when the posterior probability $\mathrm{P}(|\beta_j| > \sqrt{\mathrm{Var}(\beta_j)}|D_n)$ exceeds 0.5. For different methods, we compare true positives, false positives and AUC for true network markers recovery, prediction mean squared error (PMSE) for linear regression and classification error (CE) for logistic regression regarding to outcome. We report the mean and standard error over 50 datasets for each metric we choose to compare in the result table.

Table 2.1 summarizes the results for linear regression under different settings. As we can see, in most cases, the TGLG approach with incorporating network structure has smallest PMSE, smallest number of false positives with a comparable amount of true positives compared with existing regularized approaches and Bayesian methods BRGL and Ising prior model. For the Ising model, we only report the results in the case of $b = 7$ since it has an overall best performance among all choices of b values. In fact, the performance of the Ising model varies greatly for different choices of values for b and it may perform vary bad with an inappropriate value of b. Table 2.3 shows the mean computation time over 50 datasets for Ising model and TGLG. It shows that our method is much more computationally efficient than Ising model, especially for the large-scale networks. As for the three cases of adopting TGLG approaches, TGLG with updating ε has the best overall performance regarding to PMSE and false

positives. TGLG with fixed $\varepsilon = 10^{-5}$ tends to have a larger false positive than TGLG updating ε and TGLG with i.i.d prior since selection variable for connected nodes are highly dependent when fixed $\varepsilon = 10^{-5}$. However, TGLG with fixed $\varepsilon = 10^{-5}$ still has a smaller PMSE than TGLG with i.i.d prior. Compared with TGLG with i.i.d prior, TGLG with updating ε has a smaller FP and PMSE in most cases. These facts show that incorporating network structure can improve model prediction performance in linear regression.

Table 2.2 summarizes the results for the logistic regression under different simulation settings. Here the TGLG is only compared with Lasso, Elastic-net and the Ising model. The Ising model has smaller number of false positives in Simulation 1 Type 1 setting than TGLG under all three settings. However, the Ising model has a larger prediction error and a smaller number of true positives. In all other scenarios, TGLG outperforms the Ising model. Table 2.3 demonstrates the TGLG approach is much more computational efficient than the Ising model in Logistic regression. In addition, TGLG with fixed $\varepsilon = 10^{-5}$ and TGLG with updating ε have a smaller number of false positives and classification error than TGLG with i.i.d prior in most cases, which indicates that including network structure could improve model performance in logistic regression.

2.5.2 scalefree network

To further evaluate the performance of our proposed method, we conduct another more complicated simulation study. In this simulation study, we generate 50 datasets with sample size 200 for scalefree network with dimension 1,000. In this simulation, we consider different network structure and different coefficients for each dataset. For each dataset, we randomly generate a scalefree network with dimension 1,000. Covariates X are generated from a multivariate Gaussian distribution $X \sim \mathrm{N}(0, 0.3^D)$, where D is the shortest path distance matrix between nodes in the generated scalefree

Table 2.1: Simulation results for linear regression. PMSE: prediction mean squared error. TP: true positives, FP: false positives. Numbers of true network markers in Type 1 and Type 2 are 22 and 12, respectively.

Method	PMSE	TP	FP	AUC	PMSE	TP	FP	AUC
	Simulation 1 Type 1 $p = 33$				Simulation 1 Type 1 $p = 110$			
Lasso	52.3(1.6)	20.6(0.2)	7.3(0.3)	0.861(0.007)	71.6(1.9)	17.2(0.3)	19.6(1.2)	0.847(0.005)
Elastic-net	50.9(1.4)	21.8(0.1)	10.4(0.2)	0.847(0.007)	73.7(1.8)	19.6(0.3)	46.6(2.9)	0.871(0.004)
Grace	56.8(1.5)	21.6(0.1)	10.1(0.2)	0.864(0.007)	87.5(2.0)	17.9(0.4)	37.5(2.5)	0.897(0.004)
aGrace	53.7(1.5)	22.0(0.0)	10.7(0.1)	0.875(0.007)	76.4(2.1)	20.6(0.3)	65.9(3.6)	0.899(0.005)
L_∞	51.4(1.5)	21.8(0.1)	8.9(0.4)	0.970(0.006)	66.5(1.7)	21.5(0.2)	22.7(1.5)	0.973(0.005)
aL_∞	54.2(1.3)	21.8(0.1)	8.2(0.6)	0.669(0.034)	63.5(1.5)	21.5(0.2)	19.6(1.4)	0.946(0.010)
TTLP	54.3(1.6)	21.9(0.0)	10.1(0.4)	0.834(0.019)	72.6(2.0)	20.9(0.4)	44.2(4.6)	0.920(0.004)
LTLP	51.3(1.2)	22.0(0.0)	8.8(0.6)	0.933(0.005)	67.1(1.7)	21.5(0.2)	57.6(2.7)	0.897(0.009)
BRGL	51.0(1.3)	19.5(0.2)	4.1(0.3)	0.883(0.008)	79.7(1.8)	17.9(0.2)	22.1(0.9)	0.867(0.006)
Ising(b=7)	54.9(3.0)	19.7(0.7)	2.9(0.7)	0.925(0.017)	94.9(5.9)	15.1(0.9)	33.9(2.4)	0.786(0.023)
TGLG ($\boldsymbol{\gamma} \sim \mathrm{N}\{\mathbf{0}_{p_n}, \sigma_\gamma^2 I_p)$)	50.1(1.3)	21.9(0.1)	10.7(0.2)	0.863(0.010)	81.4(2.1)	14.8(0.5)	22.6(2.6)	0.779(0.009)
TGLG ($\varepsilon = 10^{-5}$)	45.2(1.2)	22.0(0.0)	2.2(0.6)	0.912(0.032)	63.9(2.8)	19.7(0.4)	17.8(2.9)	0.899(0.016)
TGLG ($log\varepsilon \sim N(-5, 9)$)	46.0(1.3)	21.9(0.1)	1.7(0.5)	0.968(0.016)	74.1(2.4)	17.1(0.5)	19.3(2.7)	0.847(0.013)
	Simulation 1 Type 2 $p = 33$				Simulation 1 Type 2 $p = 110$			
Lasso	23.1(0.6)	11.7(0.1)	11.8(0.6)	0.904(0.007)	30.6(0.8)	9.5(0.2)	19.1(1.1)	0.874(0.007)
Elastic-net	23.4(0.6)	11.8(0.1)	15.4(0.6)	0.809(0.006)	31.4(0.9)	10.6(0.2)	34.0(2.1)	0.842(0.006)
Grace	25.8(0.6)	11.4(0.1)	14.7(0.6)	0.813(0.005)	35.2(0.8)	9.1(0.2)	25.8(1.9)	0.855(0.005)
aGrace	25.9(0.7)	12.0(0.0)	20.3(0.3)	0.868(0.006)	32.8(0.8)	11.6(0.1)	73.0(3.5)	0.895(0.007)
L_∞	23.8(0.6)	11.9(0.1)	17.2(0.6)	0.812(0.005)	30.3(0.7)	11.3(0.2)	28.9(1.9)	0.928(0.005)
aL_∞	26.1(0.7)	11.9(0.1)	16.9(0.6)	0.643(0.018)	30.6(0.6)	11.3(0.2)	27.1(1.7)	0.893(0.009)
TTLP	25.9(0.8)	12.0(0.0)	20.0(0.5)	0.801(0.008)	32.2(0.8)	11.6(0.2)	64.3(5.2)	0.923(0.004)
LTLP	24.7(0.7)	12.0(0.0)	20.4(0.4)	0.825(0.008)	30.6(0.7)	11.7(0.2)	75.1(3.6)	0.864(0.006)
BRGL	23.7(0.6)	11.4(0.1)	7.3(0.4)	0.938(0.007)	37.7(0.9)	9.9(0.1)	23.8(1.1)	0.876(0.008)
Ising(b=7)	27.8(1.5)	9.9(0.5)	11.6(0.8)	0.855(0.024)	45.8(2.6)	7.6(0.6)	44.5(2.0)	0.709(0.032)
TGLG ($\boldsymbol{\gamma} \sim \mathrm{N}\{\mathbf{0}_{p_n}, \sigma_\gamma^2 I_p)$)	23.7(0.6)	10.8(0.2)	8.0(0.9)	0.918(0.006)	33.9(0.9)	7.2(0.3)	7.6(1.5)	0.829(0.011)
TTGLG ($\varepsilon = 10^{-5}$)	22.8(0.6)	11.4(0.1)	10.2(0.7)	0.901(0.015)	28.7(1.1)	10.5(0.3)	14.2(2.1)	0.922(0.012)
TGLG ($log\varepsilon \sim N(-5, 9)$)	22.3(0.6)	11.6(0.1)	8.9(0.6)	0.930(0.008)	28.8(0.9)	8.8(0.3)	6.4(1.1)	0.908(0.011)

Table 2.2: Simulation results for logistic regression with sample size is 200. CE: classification error, number of wrong prediction classification. TP: true positive, FP: False Postive. Number of true network markers in setup 1 and setup 2 are 22 and 12, respectively.

Method	CE	TP	FP	AUC	CE	TP	FP	AUC
	Simulation 1 Type 1 $p = 33$				Simulation 1 Type 1 $p = 110$			
Lasso	20.8(0.7)	21.2(0.1)	6.9(0.4)	0.915(0.006)	30.8(1.1)	19.1(0.4)	25.1(1.7)	0.907(0.004)
Elastic-net	21.0(0.8)	21.4(0.1)	8.4(0.4)	0.920(0.005)	32.6(0.8)	19.9(0.2)	29.4(2.1)	0.916(0.004)
Ising(b=5)	39.2(3.0)	15.2(1.2)	0.0(0.0)	0.937(0.011)	47.6(4.1)	13.5(1.1)	10.2(2.9)	0.826(0.031)
TGLG ($\boldsymbol{\gamma} \sim \mathrm{N}\{\mathbf{0}_{p_n}, \sigma_\gamma^2 I_p)$)	19.2(0.6)	21.9(0.1)	10.0(0.2)	0.877(0.011)	30.5(0.9)	17.1(0.3)	16.0(1.4)	0.851(0.008)
TGLG ($\varepsilon = 10^{-5}$)	19.4(0.7)	21.8(0.1)	8.0(0.5)	0.858(0.021)	30.8(1.1)	17.6(0.4)	13.0(1.1)	0.870(0.007)
TGLG ($log\varepsilon \sim N(-5, 9)$)	18.7(0.7)	21.8(0.1)	7.5(0.5)	0.875(0.018)	30.4(1.0)	17.3(0.3)	13.4(1.1)	0.858(0.008)
	Simulation 1 Type 2 $p = 33$				Simulation 1 Type 2 $p = 110$			
Lasso	25.2(0.9)	11.7(0.1)	10.1(0.7)	0.934(0.005)	32.7(1.0)	10.6(0.2)	22.7(2.2)	0.941(0.005)
Elastic-net	26.1(0.8)	11.9(0.0)	13.2(0.7)	0.876(0.004)	36.6(1.2)	10.5(0.3)	25.9(2.5)	0.915(0.004)
Ising(b=5)	27.4(1.4)	9.5(0.4)	7.2(0.4)	0.899(0.016)	37.7(2.8)	7.4(0.5)	9.0(1.7)	0.820(0.025)
TGLG ($\boldsymbol{\gamma} \sim \mathrm{N}\{\mathbf{0}_{p_n}, \sigma_\gamma^2 I_p)$)	22.6(0.8)	11.4(0.1)	4.8(0.6)	0.961(0.007)	29.4(1.2)	9.7(0.3)	6.9(0.9)	0.897(0.0012)
TGLG ($\varepsilon = 10^{-5}$)	23.2(0.8)	11.5(0.1)	6.3(0.6)	0.941(0.010)	29.3(1.0)	9.9(0.3)	6.7(0.6)	0.903(0.010)
TGLG ($log\varepsilon \sim N(-5, 9)$)	22.1(0.8)	11.6(0.1)	5.8(0.7)	0.959(0.005)	28.6(1.0)	10.1(0.2)	6.2(0.8)	0.921(0.009)

Table 2.3: Average computing time with standard deviation in seconds for Ising model and TGLG based network marker selection. All the calculations are executed on a desktop computer with 3.40 GHz i7 CPU and 16 GB memory

	Linear regression		Logistic regression	
	Ising	**TGLG**	Ising	**TGLG**
Simulation 1 Type 1 $p = 33$	140.1(0.5)	21.5(0.2)	230.1(7.6)	26.7(0.3)
Simulation 1 Type 2 $p = 33$	140.1(0.5)	21.0(0.3)	229.9(7.6)	26.4(0.2)
Simulation 1 Type 1 $p = 110$	1191.4(7.1)	31.7(0.2)	1210.1(10.1)	37.7(1.0)
Simulation 1 Type 2 $p = 110$	1153.4(8.5)	30.6(0.1)	1203.6(8.4)	36.5(0.9)

network. Coefficients $\boldsymbol{\beta}$ are randomly generated from Unif$(1, 3)$ and are randomly assigned as positive or negative for each dataset. Response variable Y is generated using $Y \sim \mathrm{N}(X\boldsymbol{\beta}, \sum \beta_i^2/3)$ for linear regression and $\Pr(Y = 1) = 1/\{1 + \exp(-X\boldsymbol{\beta})\}$ for logistic regression. As for the true signal nodes settings, we consider two cases: 1, random select 10 connected nodes; 2, random select 10 disconnected nodes. In this way, we could see the robustness for our proposed method when the model assumption are violated in case 2 true signal nodes setting. Another concern for TGLG is that TGLG prior construction depends on the correctly specified network structure. So in this simulation, we also adopt TGLG with a misspecified network structure, where 20% of the nodes are randomly permuted.

Table 2.4 summarizes the results for scale free network under different simulation settings. As we can see, when true signal nodes are connected, TGLG with updating ε has the overall best performance regarding to PMSE or CE, and number of FP. When true signal nodes are disconnected, TGLG with updating ε still has the best performance in linear regression, but is slightly worse than TGLG with i.i.d prior in logistic regression. This fact indicates that our model is robust to true signal settings. In both true signal nodes settings, TGLG with misspecified network performs worse than TGLG with correctly misspecified network, but still better than Lasso and Elastic-net. This fact indicates that prior network specification has an impact on the performance of TGLG and TGLG is robust to network specification to some extent.

2.5.3 Application to breast cancer data from the Cancer Genome Atlas

In the real data application, we use the High-quality INTeractomes (HINT) database for the biological network (Das and Yu, 2012). We apply our method to the TCGA breast cancer (BRCA) RNA-seq gene expression dataset with 762 subjects and $10,792$ genes in the network. The response variable we consider here is ER status - whether

Table 2.4: Simulation results for scale free network. True TP is 10. Sample size is 200 and dimension is 1,000.

Method	PMSE	TP	FP	AUC	CE	TP	FP	AUC
	Linear regression				Logistic regression			
true signal nodes are connected								
Lasso	21.7(0.6)	9.5(0.1)	54.4(3.8)	0.982(0.004)	43.3(1.6)	8.4(0.2)	29.6(3.4)	0.954(0.007)
Elastic-net	23.2(0.7)	9.6(0.1)	69.0(3.9)	0.975(0.005)	57.9(2.4)	7.7(0.2)	22.4(3.2)	0.961(0.006)
TGLG ($\boldsymbol{\gamma} \sim \mathrm{N}\{\mathbf{0}_{p_n}, \sigma_\gamma^2 I_p)$)	21.7(0.8)	9.1(0.1)	13.5(1.9)	0.950(0.007)	37.2(1.3)	7.7(0.2)	8.9(0.9)	0.892(0.011)
TGLG ($\varepsilon = 10^{-5}$)	21.8(0.9)	9.3(0.1)	14.6(1.5)	0.968(0.006)	35.2(1.3)	8.0(0.2)	7.8(0.9)	0.902(0.011)
TGLG ($log\varepsilon \sim N(-5, 9)$)	20.7(0.7)	9.1(0.1)	10.1(1.5)	0.957(0.006)	35.4(1.4)	7.9(0.3)	8.3(1.0)	0.893(0.011)
TGLG ($log\varepsilon \sim N(-5, 9)$) misspecified	21.2(0.8)	9.1(0.1)	11.3(1.5)	0.952(0.007)	37.1(1.3)	7.8(0.2)	9.3(1.1)	0.892(0.012)
True signal nodes are disconnected								
Lasso	20.8(0.6)	9.8(0.1)	55.0(3.7)	0.989(0.003)	43.4(1.2)	8.9(0.2)	26.8(3.0)	0.979(0.004)
Elastic-net	22.2(0.7)	9.8(0.1)	68.6(3.9)	0.988(0.003)	55.7(1.9)	8.4(0.2)	27.3(4.0)	0.981(0.003)
TGLG ($\boldsymbol{\gamma} \sim \mathrm{N}\{\mathbf{0}_{p_n}, \sigma_\gamma^2 I_p)$)	21.4(0.9)	9.4(0.1)	13.4(2.0)	0.974(0.006)	35.4(1.3)	8.6(0.2)	7.9(0.8)	0.931(0.009)
TGLG ($\varepsilon = 10^{-5}$)	21.7(0.8)	9.4(0.1)	16.7(1.9)	0.971(0.006)	35.5(1.4)	8.4(0.2)	7.8(0.9)	0.922(0.010)
TGLG ($log\varepsilon \sim N(-5, 9)$)	20.6(0.8)	9.6(0.1)	11.6(2.1)	0.980(0.004)	36.9(1.5)	8.5(0.2)	9.4(1.1)	0.925(0.009)
TGLG ($log\varepsilon \sim N(-5, 9)$) misspecified	21.3(0.9)	9.4(0.1)	11.4(1.7)	0.969(0.005)	35.3(1.2)	8.5(0.2)	8.4(0.9)	0.928(0.008)

the cancer cells grow in response to the estrogen. The ER status is a molecular characteristic of the cancer which has important implications in prognosis. The purpose here is not focused on prediction. Rather we intend to find genes and functional modules that are associated with ER status, through which biological mechanisms differentiating the two subgroups of cancer can be further elucidated.

We code ER-positive as 1 and ER-negative as 0. We remove subjects with unknown ER status. In total, there are 707 subjects with 544 ER-positive and 163 ER-negative. We remove 348 gene nodes with low count number, which leaves us with 10,444 nodes. To apply our methods, we first standardize the gene nodes and then apply a logistic regression model for network marker selection. For prior settings, we use $\sigma_\gamma^2 \sim \mathrm{IG}(0.01, 0.01), \sigma_\alpha^2 \sim \mathrm{IG}(0.01, 0.01)$ and $\sigma_\omega^2 = 50$. We fix λ at different grid values and choose $\lambda = 0.004$ by maximizing the likelihood values. The MCMC algorithms runs 100,000 iterations with first 90,000 as burn-in and thin by 10. We run the chain with 3 different initial values and the Gelman-Rubin diagnostic statistic is [1.07,1.15], which shows convergence of the chain.

A total of 470 genes are selected as network marker by our approach. To facilitate data interpretation, we conduct the community detection on the network containing the selected network markers and their one-step neighbors (Clauset et al., 2004). There is a total of eight modules that contain 10 or more selected genes. The modules

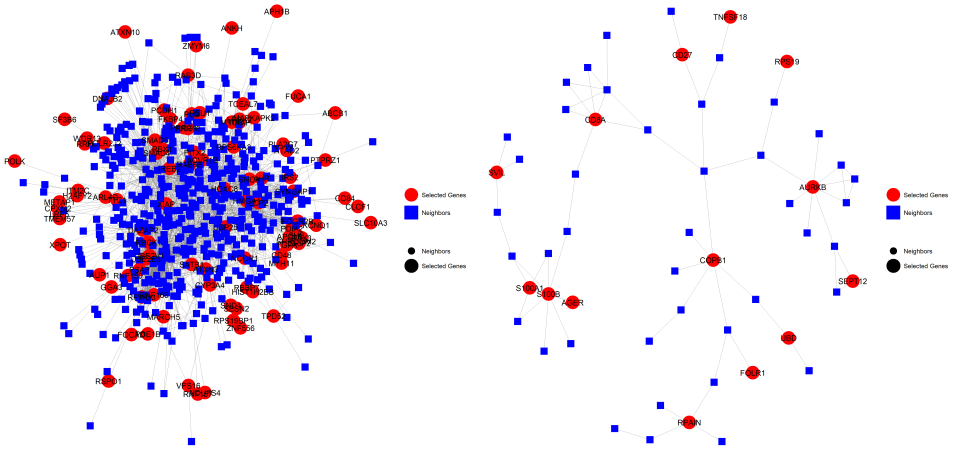
and their over-represented biological process are identified using the 'GOstats' package (Falcon and Gentleman, 2007)

Figure 2.3 shows two example network modules. The first example (Figure 2.3(a)) contains 95 selected gene network markers, including 14 that are connected with other network markers. The top 5 biological processes associated with these 95 genes are listed in Table 2.5. The most significant biological process that is over-represented by the selected genes in this module is regulation of cellular response to stress (p=0.00016), with 14 of the selected genes involved in this biological process. Besides the general connection between stress response and breast cancer, ER status has some specific interplay with various stress response processes. For example, breast cancer cells up-regulate hypoxia-inducible factors, which cause higher risk of metastasis (Gilkes and Semenza, 2013). Hypoxia inducible factors can influence the expression of estrogen receptor (Wolff et al., 2017). In addition, estrogen changes the DNA damage response by regulating proteins including ATM, ATR, CHK1, BRCA1, and p53 (Caldon, 2014). Thus it is expected that DNA damage response is closely related to ER status.

Five other genes in this module are involved in the pathway of regulation of anion transport, which include the famous mTOR gene, which is implicated in multiple cancers (Le Rhun et al., 2017). The PI3K/AKT/mTOR pathway is an anticancer target in ER+ breast cancer (Ciruelos Gil, 2014). The other four genes, ABCB1 (Jin and Song, 2017), SNCA (Li et al., 2018), IRS2 (Yin et al., 2017) and NCOR1 (Lopez et al., 2016) are all involved in some other types of cancer.

In ER- breast cancer cells, the lack of ER signaling triggers the epigenetic silencing of downstream targets (Leu et al., 2004), which explains the significance of the biological process "negative regulation of gene silencing". Many genes in the "cardiac muscle cell development" processes are also part of the growth factor receptor pathway, which has a close interplay with estrogen signaling (Osborne et al., 2005). Four

of the genes fall into the process "regulation of B cell proliferation". Among them, AHR has been identified as a potential tumor suppressor (Formosa et al., 2017). ERα is recruited in AhR signaling (Matthews and Gustafsson, 2006). IRS2 responds to interleukin 4 treatment, and its polymorphism is associated with colorectal cancer risk (Yin et al., 2017). CLCF1 signal tranduction was found to play a critical role in the growth of malignant plasma cells (Burger et al., 2003). It appears that these genes are found due to their functionality in signal transduction, rather than specific functions in B cell proliferation.



(a) (b)

Figure 2.3: Two example modules of selected genes.

The second example is a much smaller module including 14 selected genes. Six of the 14 genes are involved in both hemopoiesis and immune system development (Table 2.5). They are all signal transducers. Among them, AGER is a member of the immunoglobulin superfamily of cell surface receptors, which also acts as a tumor suppressor (Wu et al., 2018). CD27 is a tumor necrosis factor (TNF) receptor. Treatment with the estrogen E2 modulates the expression of CD27 in the bone marrow and spleen cells (Stubelius et al., 2014). TNFSF18 is a cytokine that belongs to the tumor necrosis factor (TNF) ligand family. Although its relation with estrogen and breast cancer is unclear, its receptor GITR shows increased expression in tumor-

Table 2.5: Selected goterm results for the two selected modules shown in Figure 2.3. The upper part is the Goterm results for Figure 2.3(a) and the lower part is the Goterm results for Figure 2.3(b).

GOBPID	Pvalue	Term
GO:0080135	0.0001618	regulation of cellular response to stress
GO:0044070	0.000381	regulation of anion transport
GO:0060969	0.0004409	negative regulation of gene silencing
GO:0055013	0.000757	cardiac muscle cell development
GO:0030888	0.0009629	regulation of B cell proliferation
GOBPID	Pvalue	Term
GO:0030097	0.00006398	hemopoiesis
GO:1902533	0.0003036	positive regulation of intracellular signal transduction
GO:0002250	0.0004063	adaptive immune response
GO:0032467	0.0004452	positive regulation of cytokinesis
GO:0070229	0.0005767	negative regulation of lymphocyte apoptotic process

positive lymph nodes from advanced breast cancer patients (Krausz et al., 2012), and is targeted by some anti-cancer immunotherapy (Schaer et al., 2012). UBD is a ubiquitin-like protein, which promotes tumor proliferation by stabilizing the translation elongation factor eEF1A1 (Liu et al., 2016).

Interestingly, three of the other top biological processes are also immune processes. In normal immune cells, estrogen receptors regulate innate immune signaling pathways (Kovats, 2015). In addition, some of the selected genes in these pathways have been found to associate with cancer. Examples include AURKB, which belongs to the family of serine/threonine kinases, and contributes to chemo-resistance and poor prognosis in breast cancer (Zhang et al., 2015), and SVIL, which mediates the suppression of p53 protein and enhances cell survival (Fang and Luna, 2013).

Overall, genes selected by TGLG are easy to interpret. Many known links exist between these genes and ER status, or breast cancer in general. Still many of the selected genes are not reported so far to be linked to ER status or breast cancer. Our results indicate they may play important roles.

2.6 Discussion

In summary, we propose a new prior model: TGLG prior for Bayesian network marker selection over large-scale networks. We show the proposed prior model enjoys large prior support for network marker selection over large-scale networks, leading to the posterior consistency. We also develop an efficient Metropolis-adjusted Langevin algorithm (MALA) for posterior computation. The simulation studies show that our method performs better than existing regularized regression approaches with regard to the selection and prediction accuracy. Also, the analysis of TCGA breast cancer data indicates that our method can provide biologically meaningful results.

Chapter 3

Network Marker Selection for Untargeted LC-MS Metabolomics Data

3.1 Introduction

In this chapter, we propose a unified framework for network feature selection from the metabolic network along with optimal matching detection. First, we adopt the ego-network concept for easy delineation of subnetworks from a large-scale biological network (Yang et al., 2014); Next, we develop a sequential optimizing procedure which conducts feature selection of subnetworks based on their predictive power of the clinical outcome, and detects the optimal matching between features and metabolites. To the best of our knowledge, we are the first to address the matching uncertainty issue in metabolomic network analysis. Our proposed framework provides a very flexible sequential optimization procedure that can incorporate various machine learning algorithms to identify the most important subnetworks while finding optimal matching, including the Naive Bayes method which is in concept close to the common enrichment-based methods.

In actual application the user can choose what adduct ions and isotope peaks should be allowed. There is clearly a trade-off. The more adduct ions and isotope peaks allowed, the more potential matching between features and metabolites. However at the same time, more false matchings are included in the computation, because there are features derived from pure noise in untargeted metabolomics data. In this study we choose to use a conservative approach, allowing only four common adduct ions and the most abundant isotopes: $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, and $[M+NH4]^+$. We evaluate the performance of our proposed method using simulation studies, and illustrate the proposed framework on a metabolome-wide association study (MWAS) of body mass index (BMI) in a healthy cohort.

3.2 Method

3.2.1 The setup of the problem

Suppose the dataset contains n samples with p features. For $i = 1, \cdots, n$ and $j = 1, \cdots, p$, denoted by x_{ij} observation i for feature j and by y_i an outcome variable, which could be continuous or categorical. Write $x_j = (x_{1j}, \cdots, x_{nj})^T$, $X = (x_1, \cdots, x_p)$ and $y = (y_1, \cdots, y_n)^T$. A network of q metabolites is also given. Let $D = \{D_{kl}\}_{q \times q}$ denote the distance matrix of the metabolic network, where $D_{kl} \in \{1, 2, \cdots\}$ indicates the distance between metabolites k, l. The distance here is defined as the shortest path between two nodes in a graph. In the rest of paper we will not distinguish metabolite and node. Given the allowed adduct ions and m/z difference tolerance level, let m_k denote the number of features that could possibly match to metabolite k in the network and $f_k = \{f_{k1}, \cdots, f_{km_k}\}$ be the collection of those features, where $f_{kh} \in \{1, \cdots, p\}, h \in \{1, \cdots, m_k\}$. The uniqueness of this problem lies in the fact that $f_k \cap f_l \neq \varnothing$, for some $k \neq l$. This indicates that one feature may be matched to multiple metabolites.

Let t_j denote the number of possible metabolites that could match with feature j in the network and $u_j = \{u_{j1}, \cdots, u_{jt_j}\}$ be the collection of those metabolites, where $u_{jg} \in \{1, \cdots, q\}$, for $g = 1, \cdots, t_j$. In addition to the metabolomics data, suppose r demographic covariates are collected, denoted by $z_i = (z_{i1}, \cdots, z_{ir})^T$. Write $Z = (z_1, \cdots, z_n)^T$. The goal of this paper is to develop a framework to simultaneously select important network markers and identify the optimal matching of features to the metabolites on the network, while adjusting for demographic covariates.

3.2.2 Metabolic ego networks

In this study, we adopt the ego-network approach to delineate the sub-network structure, which is a well-defined notation in social network studies (Borgatti et al., 2009),

Table 3.1: Notation Definition

Notation	Definition
Z	demographic covariates
$\mathcal{D}$	largest ego-radius
v_k^s	nodes with distance $\leq s$ to node k
F_k^s	feature set of nodes v_k^s
e_k^s	predictive error for F_k^s and Z
E_k	Ego-network for node k
T	number of iterations for optimal matching
F_k	feature sets in E_k
$\sharp\{F_k\}$	number of unique features in F_k
N_k	node sets in E_k
e_k	predictive error of E_k
M	predefined largest predictive error
$\tilde{F}_k$	selected feature from $\{F_k, Z\}$
Q	set for ego node id

and previously applied in the genomics setting (Yang et al., 2014). An ego-network consists of a centroid node, referred as *ego-node*, and its neighborhood defined as a set of nodes within certain distance to the *ego-node* over the network. We refer to this distance as the *ego-radius*. An ego-network can be grown by increasing the corresponding *ego-radius* and including more nodes. Let $\mathcal{D}$ be the upper bound of the *ego-radius* for all the possible ego-networks in the network. We fix $\mathcal{D} = 2$ in the following content. Given an *ego-radius*, we can obtain all the nodes and potentially matched features of the ego network. Furthermore, we can evaluate the performance of the ego network based on a criterion, i.e., capability of the matched features to predict the clinical outcome in cross-validation, based on which we can rank all the ego-networks. Our framework is general such that any machine learning or statistical predictive model can be used, as long as they are capable of variable selection.

Table 3.1 provides a summary of all the notations and their definitions used in the general workflow (Figure 3.1). Specifically, let v_k^s be a set of metabolites with distance smaller than or equal to s to metabolite k. Denote by F_k^s the matched feature set

of metabolites v_k^s and by e_k^s the predictive error of ego-network for metabolite k with *ego-radius* s, which is calculated using some machine learning algorithms based on outcome y and covariates $\{Z, F_k^s\}$. The left box of Figure 3.1 is the proposed algorithm to determine the ego-network for each node in the network without considering the multiple matching issue.

3.2.3 Optimal matching

Next we address the multiple matching issue for network feature selection. The proposed method uses a sequential feature screening procedure to select important sub-networks and identifies the optimal matching. Let E_k denote the ego-network for node k and e_k is the predictive error of E_k. In addition, F_k and N_k are the set of features and nodes of E_k. $\tilde{F}_k$ is the set of feature selected from $\{F_k, Z\}$ of optimal matching, besides demographic variables. Let T ($T \leq q$) be the number of top ego-networks that we select and conduct the optimal matching. T is pre-specified and fixed in the algorithm. The algorithm for ego-network selection and optimal matching is described in Figure 3.1.

Our main idea for developing this algorithm lies in that for features that can match to multiple metabolites, the true matching more likely corresponds to the one where it yields the lowest predictive error, together with neighboring metabolites. In each iteration, we first select the ego-network with the lowest predictive error and conduct a statistical feature selection procedure within the ego network. Based on the feature selection results, we assign each selected feature to the ego-network, and most likely to a specific metabolite when no two metabolites share molecular weight in the ego network. This can bring changes to the matching between features and metabolites in some other ego-networks. Then we keep the selected ego-network fixed and re-fit predictive models for all other ego-networks affected by the change of matching. Repeat the procedures for ego-network selection and feature/matching selection until

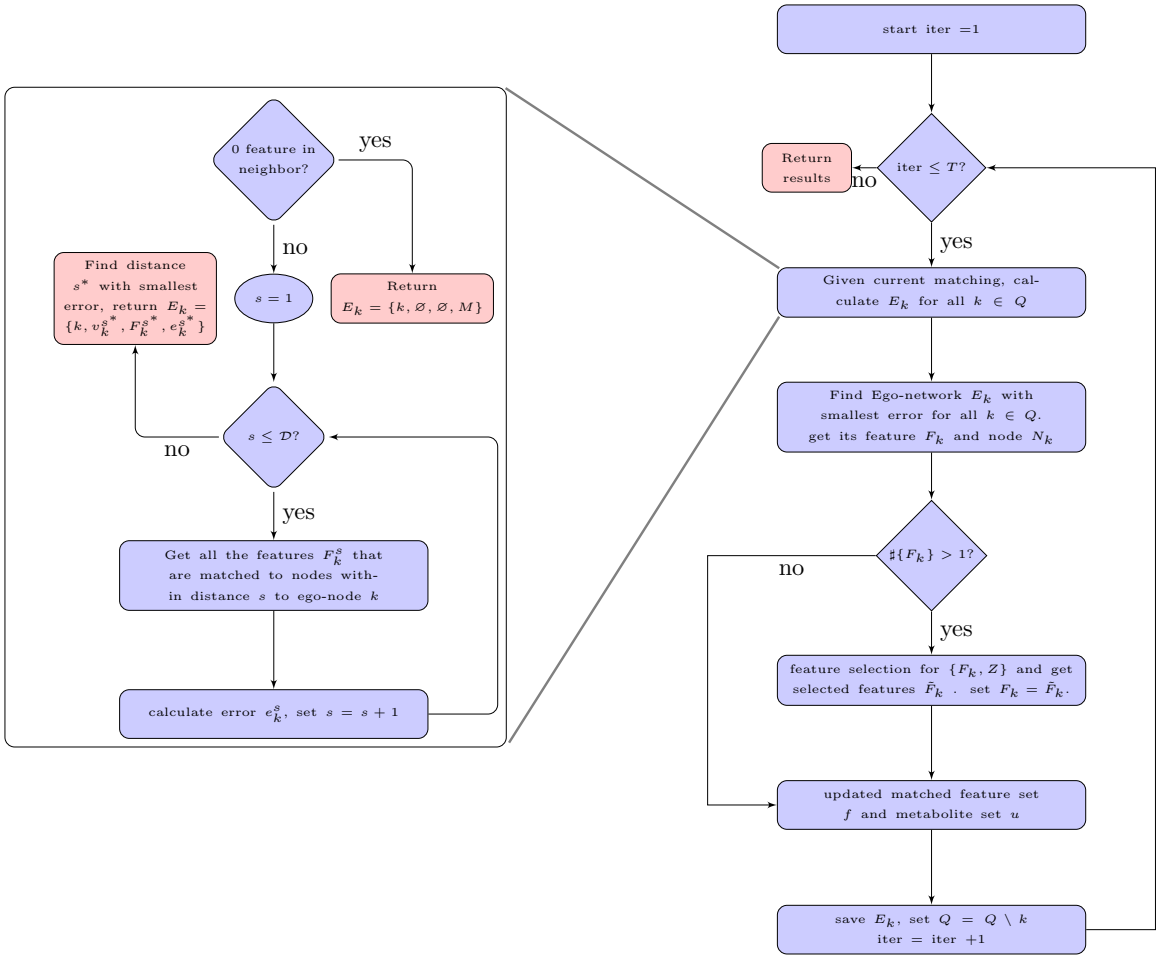


Figure 3.1: The general workflow of the method.

enumerating all ego-networks or a predefined number of iterations is met.

In many applications, it is desirable to consider demographic variables, e.g. age, gender, ethnicity *etc*. Our framework easily accommodates this need by forcing all the demographic variables to be used in predictive model fitting, regardless of the feature selection results. In rare occasions, features with the same m/z value but different retention time are matched to the same metabolite. However, only one matching can be true. In this case, the matching with the lowest predictive error is retained.

3.3 Simulation

Simulation studies are conducted to evaluate the performance of our proposed method. Here we directly adopt the KEGG human metabolic network (Kanehisa et al., 2016), as well as a real metabolomics dataset. The KEGG network was downloaded and extracted using R packages KEGGREST (Tenenbaum, 2016) and igraph (Csardi and Nepusz, 2006). For the real data, demographic covariates are omitted in simulation. In total, 1074 features are matched to 944 unique metabolites in the network, with another 1306 metabolites not matched by any feature. Of all these features, 685 have only one matched metabolite and 389 have been matched to multiple metabolites. For the purpose of simulation, a random matching for these 389 features with multiple matched nodes is set as the true matching.

Given some features have zero readings from some samples, which can be caused by either true non-presence or non-detection due to low signal strength, we use features with less than 20% 0's to generate the response variable and use features with less than 50% 0's to calculate the predictive error of an ego-network. For each simulation, we randomly choose a sub-network as the ground truth ego-network and randomly sample more than two features with less than 20% 0's from the selected sub-network

to generate response variable using a logistic regression model, i.e. for $i = 1, \ldots, n$,

$$\Pr(y_i = 1) = \frac{1}{1 + \exp(-x_i^T \beta)},$$

where the response variable $y_i \in \{0, 1\}$. The sample size is 499, and the number of true predictors is between 2 and 11.

We generate 100 datasets in this simulation study. The mean and median number of features in the true ego-network are 3.61 and 3, respectively. In this simulation, we only make changes to the top 20 ego-networks ranked by classification accuracy ($T = 20$). Four methods are considered here for comparisons: logistic regression (LR), naive Bayes classifier (NBC), random forest (RF) (Breiman, 2001) and support vector machine (SVM) (Cortes and Vapnik, 1995). We calculate the predictive error using a 5-fold cross validation. Recursive Feature Elimination (RFE) (Kuhn, 2008) procedure implemented in `R` package "caret" is adopted for feature selection. For comparison, we also conduct the simulation without considering the multiple matching issue - every ego-network uses all the features possibly matched to the ego-network, which we refer to as "maximum matching" in the following discussion. Note that all currently existing methods implicitly use the maximum matching method as they do not consider the multiple matching issue. Also, the naive Bayes (NB) approach paired with the maximum match is in essence similar to the predominant enrichment-based analysis, such as Mummichog etc (Li et al., 2013a).

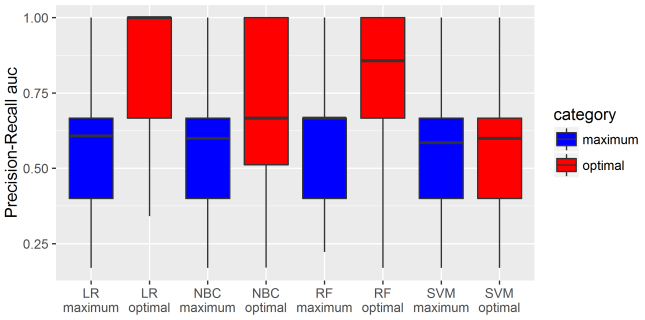
Figure 3.2(a) presents a summary of the network marker selection accuracy for our proposed optimal matching method and the maximum matching method. We first compare the sensitivity and specificity of selecting the correct features from all features. Because most features are in the negative class, we use the precision-recall curve to summarize the results from each simulated dataset, and compute the Area Under Curve (AUC). We then use boxplots to compare the AUC values of the

methods (Figure 3.2a). It is clear that the optimal matching produced better AUC than maximum matching. Among the four prediction methods, logistic regression performed the best, which is no surprise given the data is simulated from a logistic regression model. Among the more flexible machine learning methods, Random Forest achieved the best performance.

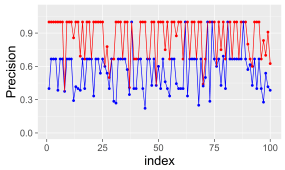
We then focus on the true ego networks that are used to generate the y value, and compare the selection accuracy of the features in the true ego network. Figure 3.2(b-e) display the precision (true positive divided by all selected) of the features, where the optimal matching approach produces a better precision (red) compared to the maximum matching approach (blue) in most cases.

Figure 3.2(f) compares the recovery of the predictive features in the true ego network. Four categories for the selection results are considered here: features in selected ego-network are exactly the same as all the true features (Same); features in selected ego-network contain all the true features and some false positive features (Large); features in selected ego-network are a subset of the true features (Small); features in selected ego-network and true features partially overlap (Mixed). Computational time for different methods are also compared. Computation time is the average CPU time in seconds per simulation across 100 simulations. All the simulations are executed on a desktop computer with 3.40 GHz i7 CPU and 16 GB memory.

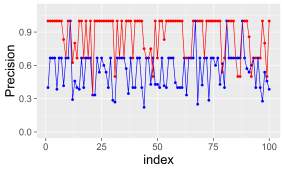
In addition to the frequency in all categories, the size ratios are also calculated (Figure 3.2(g)) except for the "Same" category. In a "Large" or "Small" selected ego-network, the size ratio is defined as the ratio of the number of selected features over the number of true features. For a "Mixed" selected ego-network, the size ratio is defined as the ratio of the number of the shared features over the number of true features. Of note, a "Large" selected ego-network with a smaller size ratio has a more accurate selection. On the other hand, a "Small" selected ego-network or a "Mixed" selected ego-network with a larger size ratio indicates a better selection result. Overall, the



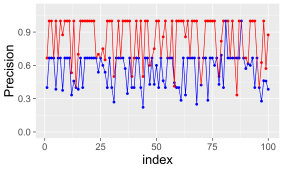
(a) Feature level precision-recall curve AUC



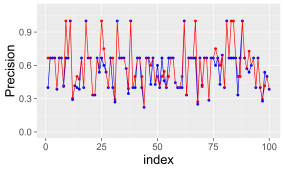
(b) Logistic Regression



(c) Naive Bayes Classifier



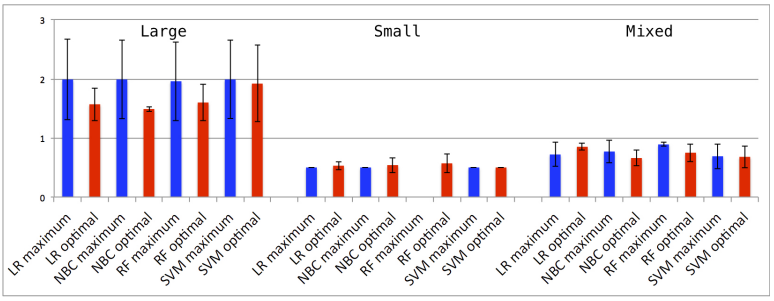
(d) Random Forest



(e) Support Vector Machine

	Mixed	Small	Large	Same	Time (seconds)
LR maximum	0.07 (0.02,0.14)	0.04 (0.00,0.11)	0.87 (0.82,0.94)	0.02 (0.00,0.09)	44.1 (6.4)
LR optimal	0.05 (0.00,0.16)	0.04 (0.00,0.15)	0.34 (0.25,0.45)	0.57(0.48,0.68)	25.7 (1.9)
NBC maximum	0.07 (0.02,0.13)	0.02 (0.00,0.08)	0.89 (0.84,0.95)	0.02 (0.00,0.08)	467.8 (62.0)
NBC optimal	0.17 (0.07,0.27)	0.22(0.12,0.32)	0.20 (0.10,0.30)	0.41(0.31,0.51)	418.7 (21.5)
RF maximum	0.03 (0.00,0.07)	0.00 (0.00,0.04)	0.95 (0.92,0.99)	0.02 (0.00,0.06)	339.2 (62.4)
RF optimal	0.07 (0.00,0.17)	0.05 (0.00,0.15)	0.39 (0.29,0.49)	0.49(0.39,0.59)	216.2 (8.3)
SVM maximum	0.06 (0.01,0.13)	0.05 (0.00,0.12)	0.87 (0.82,0.94)	0.02 (0.00,0.09)	717.3 (171.9)
SVM optimal	0.13 (0.06,0.22)	0.05 (0.00,0.14)	0.75 (0.68,0.84)	0.07(0.00,0.16)	185.0 (7.2)

(f) Recovery of the predictive features in the true ego network



(g) Selection ratio of the true ego network (closer to 1 is better)

Figure 3.2: A comparison of sensitivities between the proposed optimal matching and the maximum matching method. (a) Boxplots of area under the curve (AUC) of feature-level precision-recall (PR) curve. (b) to (e) Comparison of precision in the true ego network for individual simulated datasets. Red: optimal matching; blue: maximum matching. (f) Feature selection accuracy for the true ego network. Ratios of the four categories are followed by 95% confidence interval. Computing time (last column) is followed by standard deviation. (g) Selection ratios of the methods.

optimal matching approach (Figure 3.2(b), orange bars) produces better size ratios.

3.4 Application

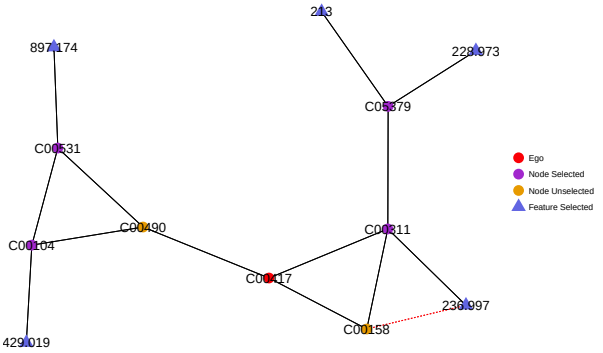
3.4.1 Dataset

We test our network marker selection framework in the Emory-Georgia Tech Predictive Health Initiative Cohort of the Center for Health Discovery and Well Being. This is an ongoing, cohort of generally healthy university employees, ages 18 and older, recruited between January 2008 and February 2013 (http: //predictivehealth.emory.edu) (Brigham, 2010). All participants are free of any acute illness, uncontrolled or unstable chronic disease, hospitalizations within the year prior to study entry, substance or drug abuse within the past year, or active malignant neoplasm or history of malignancy other than basal cell skin cancer within the previous 5 years. Subjects undergo an extensive medical and metabolic assessment annually. The study is approved by the Emory Institutional Review Board, and all participants provide informed consent prior to any testing. For this study, only subjects with available high-resolution plasma metabolomics data are assessed ($N = 371$). For metabolic network, we use the KEGG human metabolic network (Kanehisa et al., 2016), and removed all nodes with degrees of 20 or higher. Such highly connected nodes are involved in too many reactions for their concentration level to be informative. In addition, the subnetwork surrounding such a node may be too diverse to carry a clear biological theme. From a network analysis point of view, the presence of such nodes makes the distance between most node pairs very small, making it difficult to select meaningful subnetworks. We conducted a systematic study of network characteristics versus the cutoff value, and determine 20 is a good cutoff value.

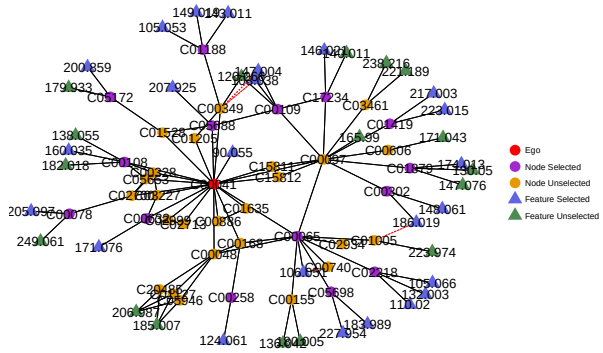
3.4.2 Results

We choose to use BMI as the outcome variable to assess our proposed framework because of the vast literature, including metabolomics studies Ho et al. (2016); Boulet et al. (2015); Newgard et al. (2009); Bogl et al. (2016); Moore et al. (2014), linking BMI, obesity, and adiposity to major metabolic pathways. This would allow us to evaluate the biological plausibility of our models. We compute the classification accuracy based on the four methods: logistic regression (LR) , naive Bayes classifier (NBC), random forest (RF), and support vector machine (SVM). All methods result in some degree of biological plausibility with regard to ego-network links to BMI. However, the random forest method, in additional to being one of the best-performing methods in the simulation study (above), provide the most consistent ego-networks in terms of the resultant ego-nodes and selected metabolites fitting within a specific metabolic pathway or common unifying metabolite. Given the nature of ego-networks, some of the selected ego-networks are partially overlapping, as their ego nodes are neighbors in the KEGG network. Some metabolic pathways are represented by several of the selected ego-networks. We select the top 30 ego-networks generated using the random forest method, excluding those supported by a single feature. They are metabolically connected to several pathways or specific metabolites, all of which have been biologically linked to BMI.

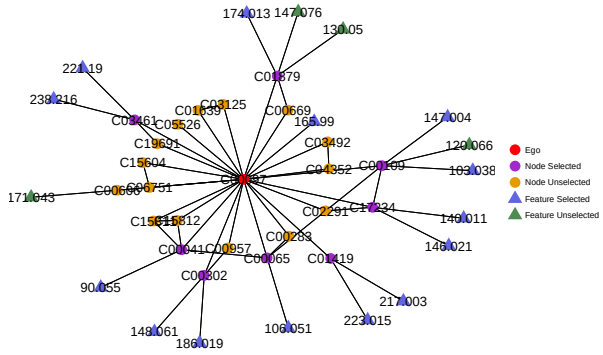
Several of the ego-networks are related to the tricarboxylic acid cycle (TCA) cycle. The TCA cycle is an essential mitochondrial component of the metabolism of carbohydrates, fats, and proteins for the production of energy. Impaired mitochondrial activity has long been implicated in the development in obesity and its metabolic sequelae given the role of the mitochondria in energy expenditure and lipid storage and mobilization (Christe et al., 2013). It is, therefore, expected that BMI would be linked to such a key pathway, in addition to several metabolites that function as substrates for the TCA cycle (pyruvate, lactate, alanine, cysteine, glutamate, phenylalanine, and



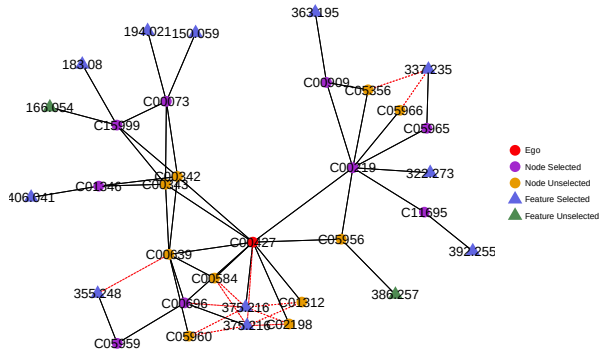
(a) C00417 (cis-Aconitate)



(b) C00041 (Alanine)



(c) C00097 (Cysteine)



(d) C00427 (Prostaglandin H2)

Figure 3.3: Some example ego-networks selected by Random Forest. Red dotted line means the matching between feature and node is eliminated by our algorithm.

tryptophan). Specific intermediates of the TCA cycle identified as either ego-nodes or metabolites within the ego-networks include citrate, oxalosuccinate, and cis-aconitate (Figure 3.3(a))(Akram, 2014; Bender DA, 2015). Obesity has been shown to impair the rate-limiting enzyme of the TCA cycle, citrate synthase, which catalyzes the production of citrate (Christe et al., 2013).

Our data are also supported by a recent metabolomics study showing a relationship between BMI and several intermediates of the TCA cycle, including cis-aconitate (Figure 3.3(a)), as well as lactate and several amino acids and their intermediates, including alanine (Figure 3.3(b)), tryptophan (Figure 3.3(b), non-ego node), cysteine (Figure 3.3(c)), and phenylalanine (a non-ego node among top 30 ego networks) Ho et al. (2016). Interestingly, the tryptophan intermediates within our ego-networks, anthranilate, 3-hydroxyanthranilate, and 3-hydroxy-L-kynurenine are consistent with studies indicating that obesity induces the increase of indoleamine 2, 3-dioxygenase through a pro-inflammatory pathway (Wolowczuk et al., 2012; Favennec et al., 2015). Additional metabolomics and amino acid studies confirm relationships between BMI or other indicators of adiposity and the circulating amino acids related to our selected ego-networks (Boulet et al., 2015; Newgard et al., 2009; Bogl et al., 2016; Moore et al., 2014; Felig et al., 1969).

Intermediates in the metabolism of sulfur-containing amino acids, cysteine and glutathione, are prominent among our BMI-associated ego-networks (Figure 3.3(c)). Cysteine has been implicated in the promotion of obesity through various epidemiological and experimental studies (Elshorbagy et al., 2012). Glutathione, the major intracellular antioxidant, is decreased in circulation in obesity (Di Renzo et al., 2010), consistent with the oxidative environment associated with excess adiposity (Marseglia et al., 2015). Hydrogen sulfide is identified as an ego-node in our study. This metabolite has been shown to suppress oxidative stress by promoting the transport of cysteine towards glutathione production (Kimura et al., 2009), and circulating hydrogen sul-

fide is inversely associated with obesity (Whiteman et al., 2010).

Prostaglandin H2 is an ego-node that is associated with BMI (Figure 3.3(d)). Prostaglandins are eicosanoids derived from arachidonic acid that play important roles in pro-inflammatory responses (Ricciotti and FitzGerald, 2011). Prostaglandins, in various biological sources, correlate positively with BMI and/or obesity (Martinez et al., 1999; Morris et al., 2013; Sinaiko et al., 2005; Subbaramaiah et al., 2012). Metabolites within this ego-network included arachidonate and its other derivative lipid mediators, thromboxane A2 and leukotriene A4, as well as related intermediates. Arachidonic acid, thromboxanes, and leukotrienes also correlate with adiposity (Back et al., 2014; Giouleka et al., 2011; Kaplon-Cieslicka et al., 2014; Savva et al., 2004). The link between BMI and the arachidonic acid pathway may reflect dietary differences in polyunsaturated fatty acid intakes (Ailhaud et al., 2008).

3.5 Discussion and Conclusion

Functional analysis, including network analysis and pathway analysis, is important for data interpretation and feature selection in metabolomics data. For untargeted metabolomics, the issue of multiple matching has existed for a long time and has been overlooked, which can lead to erroneous results. In this paper, we propose a flexible sequential optimizing procedure that can incorporate various machine learning algorithms to address this multiple matching issue in metabolomics data, along with identifying sub-networks which are highly relevant to the clinical outcome. The method ranks ego networks. The number of top ego-networks to study is a user-defined parameter. In practice, one can also choose the parameter based on predictive accuracy, i.e. stopping the program when the prediction accuracy is smaller than a threshold.

Simulation studies show that our method greatly improve the selection accuracy

compared with the existing maximum matching approach. Application to a real dataset also proves that our method can detect important sub-networks associated with the outcome variable. The same idea can be easily adapted to pathway analysis, where pre-determined pathways, rather than ego-networks, are used. We note that the method is based on matching of m/z values to theoretical values of known metabolites, which can only indicate, but not confirm the identities of features. Experimental approaches, such as chemical spike-in and LC-MS/MS, should be used to confirm the identities of features found to be relevant.

Chapter 4

A new framework for predictive network mediator analysis

4.1 Introduction

In this chapter, our work is motivated by a dataset from Emory-Georgia Tech Predictive Health Initiative Cohort of the Center for Health Discovery and Well Being. The dataset contains demographic variables, nutrition variables, metabolomics variables and BMI value for 179 subjects. Network structure between the metabolomics variables is also known. Here we consider nutrition variables as the exposure, metabolomics variables as the mediator and binary overweight status based on BMI as the outcome. Our goal is to find the combination of metabolomic variable and single or multiple nutrition variable with incorporating the dependence between the exposure and the mediator and the dependence between mediators that has high predictive performance about the outcome of interest. However, the goal of traditional mediation analysis is to measure or test the mediation effect. Thus, we could not directly apply the existing mediation analysis framework to achieve our goal. So we proposed a new framework for predictive mediation analysis with incorporating the dependence structure in this paper.

The remainder of the paper is organized as follows. In section 4.2, we introduce the definitions of predictive exposure, predictive mediator and predictive mediator network. An estimation procedure is also introduced in section 4.2 to identify the above definitions. In section 4.3, we conduct simulation studies to illustrate the performance of the proposed estimation procedure. We apply the proposed predictive mediation analysis framework on the motivated dataset in Section 4.4. We conclude our paper with a brief summary and discussion on future work in Section 4.5.

4.2 A predictive mediation analysis framework

Suppose there are m exposure variables $X_i, i = 1, \cdots, m$ and v mediator variable $M_j, j = 1, \cdots, v$ and a binary outcome variable Y. The network structure between the

mediator variables is also known. Our goal is to build a mediation analysis framework focusing on predictive modeling. Thus, we need to choose a evaluation metric first for evaluating predictive performance. The receiver operating characteristic (ROC) curve is a plot to show the diagnostic ability of a binary classifier and has been widely used in fields such as machine learning, biomedical, psychology (Hanley and McNeil, 1982; Hand and Till, 2001; Lasko et al., 2005; Krzanowski and Hand, 2009; Gonçalves et al., 2014) and so on. The area under the curve (AUC) is a commonly used summary measure of the diagnostic ability of the ROC curve (de Carvalho et al., 2013; Bamber, 1975). Thus, we adopt the AUC as the evaluation metric in this paper.

4.2.1 The area under the curve (AUC)

We show the definition for the area under the curve (AUC) in this section. Let D denote a dichotomous variable which takes value 1 for positive subjects and value 0 for negative subjects. Let X and B denote the diagnostic test values for positive subjects and negative subjects. F_1 and F_0 are the cumulative distribution functions (cdf) for X and B respectively, i.e. $X \sim F_1, B \sim F_0$. Given an cut-off value c, a subject has a positive test result if its diagnostic test value is larger than c and negative otherwise. The ROC curve is defined as the plot of the true positive rate against the false positive rate, where true positive rate (TPR) is defined as as the probability that a positive subject has a positive test result and false positive rate (FPR) is defined as the probability of a negative subject has a positive test result. Given a cut-off value c, we have that $TPR(c) = 1 - F_1(c)$ and $FPR(c) = 1 - F_0(c)$. Thus, ROC curve is the plot of $\{(FPR(c), TPR(c)), -\infty < c < \infty\}$ or $\{(t, ROC(t)), 0 \leq t \leq 1\}$ where $ROC(t) = 1 - F_0(F_1^{-1}(1-t))$ (Gonçalves et al., 2014). The AUC is defined as:

$$AUC = \int_0^1 ROC(u)du.$$

AUC denotes the probability that, in a randomly selected pair of positive and negative subjects, the diagnostic test value for positive subject is higher, i.e. $P(X > B)$ (Gonçalves et al., 2014). AUC has been shown to have a strong connection with the popular nonparametric Mann−Whitney test (Bamber, 1975; Faraggi and Reiser, 2002). There are several methods for empirical estimation of AUC. See Gonçalves et al. (2014) for details.

4.2.2 Predictive mediation analysis framework

We propose several new definitions for building a predictive mediation analysis framework in this section. We use $AUC(X_i)$ and $AUC(X_i, M_j)$ to denote the AUC for classifier that only uses feature X_i and uses features X_i, M_j respectively. In the definitions proposed for predictive mediation analysis, we also need to state whether the evaluation metric AUC has predictive power or not. We consider this as a subjective statement to say whether a certain AUC value can be considered as predictive or not. Thus we set up a predefined threshold value T as an user input here, i.e. $T = 0.6$. An AUC value is considered as predictive if it's larger than threshold value T.

4.2.2.1 Predictive exposure

An exposure variable X_i is a predictive exposure about the outcome Y if $AUC(X_i) > T$. This definition can be used as a screening procedure to find predictive exposures.

4.2.2.2 Predictive mediator

Predictive mediator is defined for a given exposure. M_j is a predictive mediator of exposure X_i if the following conditions hold:

- X_i, M_j are significantly correlated
- AUC $(X_i, M_j) > T$

- AUC $(X_i, M_j) >$ AUC (X_i)
- AUC $(X_i, M_j) >$ AUC (M_j)

Note here X_i does not need to be a predictive exposure. The first condition requires that X_i and M_j are significantly correlated. If they are uncorrelated, there is nothing to mediate. The other three conditions require that the AUC of the combination of exposure and mediator should be larger than the AUC of exposure or mediator alone.

4.2.2.3 Building network

Our goal is to find the combination of metabolomic variable and single or multiple nutrition variable with incorporating the dependence structure that has high predictive performance about the outcome of interest. In this paper, we use network to denote the dependence structure. There are two kinds of dependence structure here. One is the dependence between the exposure and the mediator, which is established using the definition of predictive mediator. The other is the dependence between mediators, which is given by the network denote the functional link between mediators. We define that a mediation link between an exposure X_i and a mediator M_j exists if M_j is a predictive mediator of X_i. Using this definition for mediation link, we can build a network between all exposures and their predictive mediators. Then we can combine the given the functional network between mediators and the network built on mediation link between exposure and mediator to a combined network G. Given network G, we can adopt machine learning algorithms to find subnetwork combined with exposures and mediators that are not only highly predictive about the outcome, but also easy to interpret from a biological perspective.

4.2.2.4 Predictive network mediator for single exposure

Given the network G defined in Section 2.2.3, we can now define the predictive network mediator for a single exposure. This definition is motivated by the fact that sometimes

researchers might have an exposure of particular interest and they would like to see which mediators combined with this exposure of interest are highly predictive about the outcome. Given an exposure X and a set of mediator nodes $M = \{M_{i_1}, \cdots, M_{i_k}\}$ from the combined network G, M is predictive network mediator of exposure X if the following three conditions hold:

- at least one mediator in M is a predictive mediator of X
- remove one mediator from M will decrease $AUC(X, M)$
- add one mediator to M will not increase $AUC(X, M)$

The first condition requires that exposure X is connected to at least one mediator in M. The other two conditions requires that predictive performance of exposure X and mediator set M is optimal.

4.2.3 Estimation procedure and algorithm

In section 2.2, we proposed definitions for predictive exposure, predictive mediator and predictive network mediator. In this section, we propose estimation procedure and algorithm for those above mentioned definitions.

4.2.3.1 Estimation for predictive exposure

Predictive exposure is defined as an exposure variable X_i that satisfies $AUC(X_i) > T$. This AUC is for evaluating predictive performance and should be evaluated on test data. In this paper, we adopt cross-validation to estimate AUC value. We propose an estimation procedure to identify predictive exposure using a repeated cross-validation. We use the mean of all the repeated cross-validation AUC as the final estimation for AUC. See below for details of the estimation procedure for predictive exposure:

- Given a repeated number of R, for $r = 1, \cdots R$, calculate $\widehat{AUC}_r(X_i)$

- Check $\sum_r[\widehat{AUC}_r(X_i) > T] > H$

We can declare an exposure as a predictive exposure about the outcome if $\sum_r[\widehat{AUC}_r(X_i) > T] > H$, where threshold value is also an user input value. A higher value for H results in a low TPR and FPR and a lower value for H results in a high TPR and FPR. Estimation for $AUC(X_i)$ is $\widehat{AUC}(X_i) = \frac{1}{R}\sum_r \widehat{AUC}_r(X_i)$.

4.2.3.2 Estimation for predictive mediator

Assume a given exposure X_i and a given mediator M_j. First we need to check whether X_i and M_j are significantly correlated. There are several correlation metrics that are commonly used in statistics to measure the dependence, i.e. Pearson correlation, Kendall rank correlation, Brownian distance correlation (Székely et al., 2009). Among all those choices, Brownian distance correlation can measure the dependence between two random vectors with arbitrary dimensions. Thus, we adopt the Brownian distance correlation in the following content to measure the dependence between exposure and mediator. A statistical hypothesis testing procedure for is provided in Székely et al. (2007); Székely and Rizzo (2013) to test the significance of the Brownian distance correlation.

Given X_i and M_j are significantly correlated, we propose an estimation procedure for predictive mediator similar to the estimation procedure given in predictive exposure. See below for details:

- Given a repeated number of R, for $r = 1, \cdots R$, calculate $\widehat{AUC}_r(X_i), \widehat{AUC}_r(M_j), \widehat{AUC}_r(X_i, M_j)$
- Check $\sum_r[\widehat{AUC}_r(X_i, M_j) > T] > H$
- Check $\sum_r[\widehat{AUC}_r(X_i, M_j) > \widehat{AUC}_r(X_i)] > H$
- Check $\sum_r[\widehat{AUC}_r(X_i, M_j) > \widehat{AUC}_r(M_j)] > H$

If all three inequalities hold, M_j is a predictive mediator of X_i.

4.2.3.3 Greedy algorithms for predictive network mediator

In this section, we first propose a greedy algorithms (Algorithm 1) for finding predictive network mediator for a single exposure of interest. However, in some cases, researchers might doesn't have a particular exposure of interest and only want to find the combination of mediator and exposure from network G that are highly predictive about the outcome. To address this need, we propose another greedy algorithm (Algorithm 2). Algorithm 2 can be used as a screening procedure to find subnetwork of exposure and mediator from network G that are highly predictive about the outcome.

Algorithm 1 Greedy algorithm for single exposure

Input: exposure X_i, network G
Step 1: select $M_j = argmax_{M_j}\widehat{AUC}(X_i, M_j)$, where M_j is unvisited mediator neighbor nodes of X_i.
If M_j exists
    mark M_j as visited and denote $C = \{X_i, M_j\}$
Else
    Stop. Procedure is finished
Step 2: find all unvisited mediator neighbor nodes Ne for C
For each $Ne_i \in Ne$
    adopt estimation procedure for $AUC(C, Ne_i) > AUC(C)$
    calculate $\widehat{AUC}(C, Ne_i)$
If there exists Ne_i that survives the estimation procedure
    Set $C = \{C, Ne_i\}$ where $Ne_i = argmax_{Ne_i}\widehat{AUC}(C, Ne_i)$ for $Ne_i \in Ne$.
    Mark Ne_i as visited. Go to **Step 2**
Else
    Store result C. Go to **Step 1**

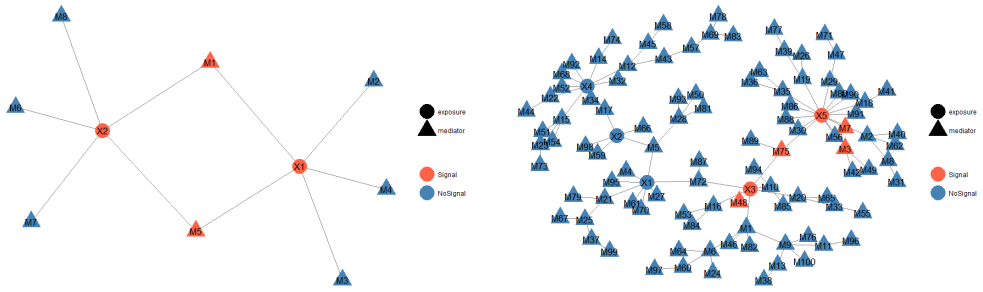
4.3 Simulation

Simulation studies are conducted to evaluate the performance of the proposed estimation procedure for identifying predictive exposure and predictive mediator. We consider two network settings for generating value for exposure and mediator, a simple network and a complex scalefree network. See Figure 4.3 for details. Exposure and

Algorithm 2 Greedy algorithm for multiple exposure

Input: network G
Step 1: select unvisited pair $(X_i, M_j) = argmax_{X_i,M_j}\widehat{AUC}(X_i, M_j)$ among all unvisited mediators
If M_j exists
    mark M_j as visited and denote $C = \{X_i, M_j\}$
Else
    Stop. Procedure is finished
Step 2: find all unvisited neighbor nodes Ne for C
For each $Ne_i \in Ne$
    adopt estimation procedure for $AUC(C, Ne_i) > AUC(C)$
    calculate $\widehat{AUC}(C, Ne_i)$
If there exists Ne_i that survives the estimation procedure
    Set $C = \{C, Ne_i\}$ where $Ne_i = argmax_{Ne_i}\widehat{AUC}(C, Ne_i)$ for $Ne_i \in Ne$.
    Mark Ne_i as visited if Ne_i is a mediator. Go to **Step 2**
Else
    Store result C. Go to **Step 1**

mediator values are generated using multivariate Gaussian distribution with mean 0 and covariance matrix $\Sigma = 0.3^D$, where D is the distance matrix between nodes in the graph. Here in generating data, a link between two nodes means there exists functional linking between connected nodes.



(a) case1 (b) scalefree

Figure 4.1: Network structure for exposure and mediator that are used to generating data. Triangle denotes mediator and circle denotes exposure. Nodes with orange color are used to generate Y.

We use a logistic regression model to generate the outcome Y. We also consider both linear and nonlinear relation between $logit(Y)$ and (X, M). We also multiply a coefficient value 'a' to adjust for signal to noise ratio (SNR). A larger value for 'a'

corresponds to a larger SNR. To be specific, in linear case, Y is generated as follows:

- Case 1: $\text{logit}(P(y=1)) = a * [X_1 - X_2 + M_1 - M_5]$.
- Scalefree: $\text{logit}(P(y=1)) = a * [X_3 - X_5 + M_3 - M_7 + M_{48} - M_{75}]$

In nonlinear case, Y is generated using:

- Case 1: $\text{logit}(P(y=1)) = a * \{X_1 * [1 + \sin(M_1) + \cos(M_5)] - X_2 * [1 + \sin(M_1) + \cos(M_5)]\}$
- Scalefree: $\text{logit}(P(y=1)) = a * \{X_3 * [1 + \sin(M_{48}) - \cos(M_{75})] - X_5 * [1 - \sin(M_3) + \cos(M_7) - \sin(M_{75})]\}$

Coefficient value 'a' is set as $1, 2, 5, 20$ to denote an increasing SNR. Sample size 100 and 500 are both considered in simulation studies. Brownian distance correlation is adopted here to measure the dependence between exposure and mediator. In the proposed estimation procedure, repeated number R is set as $R = 20$ and 5-fold cross validation is adopted to calculate AUC. Two threshold values are set as $T = 0.6$ and $H = 14$. Logistic regression (LR)and random forest (Breiman, 2001, RF) are the classifiers used to calculate AUC.

To summarize the simulation results, we report the true positive rate (TPR) and the false positive rate (FPR), which is the fraction of true positives that are diagnosed as positive and the fraction of true negatives that are diagnosed as positive respectively. To report simulation results, we are interested in whether this estimation procedure identifies predictive exposure, predictive mediator and predictive mediator to its exposure correctly. Thus, we report TPR and FPR for exposure, mediator and pair of exposure and mediator (edge in result table).

Table 1 and Table 2 summarizes the simulation results for linear case for both case 1 network and scalefree network. When looking at each row, the results show that TPR increases as SNR increases. Increasing sample size results in both a higher

TPR and a higher FPR. In the simple case 1 network setting, LR has a higher TPR on exposure, mediator and edge than RF in most SNR settings when sample size is 100. When sample size increases to 500, TPR on exposure, mediator and edge of LR is 1 and TPR on mediator and edge is close or equal to 1 for RF. When sample size is 500, LR still performs better than RF regarding to have a higher TPR on exposure, mediator and edge in most settings. Also, FRP on mediator and edge is much higher of RF than that of LR, especially for high SNR settings. In the more complex scalefree network setting, the performance of LR and RF on TPR is not as good as the simple network setting. When sample size is 500, TPR on exposure, mediator and edge of LR is 1 and TPR on mediator and edge is close or equal to 1 for RF. Overall, in linear case setting, LR performs better than RF regarding to TPR and FPR on exposure, mediator and edge. The results in Table 1 and Table 2 also show that the proposed estimation procedure is pretty accurate in linear case setting.

Table 3 and Table 4 summarizes the simulation results for nonlinear case for both case 1 network and scalefree network. The effect of SNR and sample size on TPR and FPR shows a similar pattern to that in linear case. In the simple case 1 network setting, TPR on exposure, mediator and edge are both very high for LR and RF. However, RF still has a much higher FPR on mediator and edge than LR for both sample size. In the scalefree network setting, TPR on mediator and edge is pretty high for RF, which is much better than that for LR. Overall, RF performs better than RF regarding to TPR on exposure, mediator and edge in most cases.

4.4 Real data application

This paper is motivated by a dataset from Emory-Georgia Tech Predictive Health Initiative Cohort of the Center for Health Discovery and Well Being. There are totally 179 subjects in this dataset. For each subject, the dataset contains demographic

Table 4.1: Simulation results linear case for case 1 network

		a	1	2	5	20
				Size 100		
LR	exposure	TPR	0.6(0.348)	0.8(0.299)	1(0)	0.925(0.183)
	mediator	TPR	0.9(0.205)	0.975(0.112)	0.925(0.183)	1(0)
		FPR	0.1(0.137)	0.05(0.078)	0.133(0.139)	0.108(0.112)
	edge	TPR	0.5(0.181)	0.688(0.179)	0.762(0.263)	0.825(0.143)
		FPR	0.1(0.137)	0.05(0.078)	0.133(0.139)	0.125(0.142)
RF	exposure	TPR	0.175(0.294)	0.4(0.348)	0.5(0.397)	0.5(0.281)
	mediator	TPR	0.675(0.294)	0.925(0.183)	0.9(0.205)	0.975(0.112)
		FPR	0.175(0.175)	0.2(0.149)	0.225(0.146)	0.225(0.249)
	edge	TPR	0.412(0.247)	0.675(0.164)	0.662(0.284)	0.725(0.197)
		FPR	0.175(0.175)	0.208(0.161)	0.233(0.157)	0.242(0.289)
				Size 500		
LR	exposure	TPR	1(0)	1(0)	1(0)	1(0)
	mediator	TPR	1(0)	1(0)	1(0)	1(0)
		FPR	0.208(0.186)	0.225(0.182)	0.167(0.108)	0.083(0.115)
	edge	TPR	1(0)	1(0)	1(0)	1(0)
		FPR	0.208(0.186)	0.25(0.199)	0.167(0.108)	0.083(0.115)
RF	exposure	TPR	0.2(0.251)	0.4(0.262)	0.55(0.32)	0.75(0.344)
	mediator	TPR	1(0)	1(0)	1(0)	1(0)
		FPR	0.217(0.196)	0.492(0.268)	0.6(0.198)	0.758(0.245)
	edge	TPR	0.812(0.179)	0.975(0.077)	0.988(0.056)	1(0)
		FPR	0.217(0.196)	0.533(0.323)	0.658(0.245)	0.767(0.232)

Table 4.2: Simulation results linear case for scalefree network

		a	1	2	5	20
				Size 100		
LR	exposure	TPR	0.6(0.308)	0.7(0.299)	0.7(0.377)	0.85(0.235)
		FPR	0.017(0.075)	0.017(0.075)	0.017(0.075)	0(0)
	mediator	TPR	0.588(0.233)	0.712(0.219)	0.75(0.256)	0.725(0.213)
		FPR	0.03(0.017)	0.036(0.026)	0.038(0.021)	0.023(0.016)
	edge	TPR	0.52(0.219)	0.66(0.185)	0.62(0.267)	0.65(0.193)
		FPR	0.031(0.017)	0.04(0.03)	0.041(0.024)	0.026(0.016)
RF	exposure	TPR	0.175(0.294)	0.225(0.302)	0.2(0.251)	0.35(0.286)
		FPR	0.033(0.103)	0.067(0.137)	0(0)	0.033(0.149)
	mediator	TPR	0.425(0.282)	0.65(0.274)	0.638(0.222)	0.612(0.236)
		FPR	0.061(0.052)	0.057(0.042)	0.076(0.052)	0.078(0.051)
	edge	TPR	0.34(0.252)	0.55(0.25)	0.53(0.227)	0.52(0.238)
		FPR	0.065(0.053)	0.062(0.045)	0.08(0.051)	0.082(0.053)
				Size 500		
LR	exposure	TPR	0.875(0.222)	0.975(0.112)	0.975(0.112)	1(0)
		FPR	0(0)	0(0)	0(0)	0(0)
	mediator	TPR	0.975(0.077)	1(0)	1(0)	1(0)
		FPR	0.069(0.026)	0.071(0.029)	0.073(0.027)	0.073(0.032)
	edge	TPR	0.98(0.062)	1(0)	1(0)	1(0)
		FPR	0.072(0.028)	0.074(0.028)	0.077(0.025)	0.079(0.033)
RF	exposure	TPR	0.125(0.222)	0.3(0.299)	0.45(0.359)	0.45(0.224)
		FPR	0(0)	0(0)	0(0)	0(0)
	mediator	TPR	0.888(0.151)	0.975(0.077)	0.988(0.056)	0.962(0.092)
		FPR	0.107(0.074)	0.164(0.09)	0.216(0.071)	0.201(0.068)
	edge	TPR	0.81(0.165)	0.97(0.073)	0.97(0.073)	0.95(0.089)
		FPR	0.112(0.078)	0.172(0.09)	0.225(0.074)	0.207(0.069)

Table 4.3: Simulation results nonlinear case for case 1 network

		a	1	2	5	20
				Size 100		
LR	exposure	TPR	0.95(0.154)	1(0)	1(0)	1(0)
	mediator	TPR	0.575(0.373)	0.625(0.358)	0.725(0.255)	0.85(0.235)
		FPR	0.1(0.126)	0.075(0.101)	0.142(0.124)	0.117(0.122)
	edge	TPR	0.3(0.192)	0.375(0.236)	0.462(0.247)	0.575(0.245)
		FPR	0.1(0.126)	0.075(0.101)	0.142(0.124)	0.133(0.139)
RF	exposure	TPR	0.625(0.393)	0.8(0.34)	0.95(0.154)	0.95(0.154)
	mediator	TPR	0.5(0.397)	0.7(0.34)	0.775(0.302)	0.875(0.275)
		FPR	0.267(0.244)	0.325(0.239)	0.383(0.217)	0.4(0.183)
	edge	TPR	0.375(0.358)	0.512(0.309)	0.55(0.288)	0.638(0.25)
		FPR	0.283(0.271)	0.333(0.259)	0.408(0.226)	0.45(0.265)
				Size 500		
LR	exposure	TPR	1(0)	1(0)	1(0)	1(0)
	mediator	TPR	1(0)	1(0)	1(0)	1(0)
		FPR	0.167(0.132)	0.192(0.135)	0.175(0.157)	0.192(0.156)
	edg	TPR	0.8(0.208)	0.975(0.077)	0.975(0.077)	1(0)
		FPR	0.167(0.132)	0.208(0.161)	0.217(0.196)	0.242(0.232)
RF	exposure	TPR	0.55(0.359)	0.95(0.154)	1(0)	1(0)
	mediator	TPR	0.975(0.112)	1(0)	1(0)	1(0)
		FPR	0.717(0.254)	0.833(0.195)	0.85(0.194)	0.808(0.218)
	edge	TPR	0.888(0.172)	0.9(0.17)	0.975(0.077)	1(0)
		FPR	0.8(0.34)	0.933(0.267)	0.967(0.268)	0.925(0.273)

Table 4.4: Simulation results nonlinear case for scalefree network

		a	1	2	5	20
				Size 100		
LR	exposure	TPR	0.5(0.162)	0.575(0.183)	0.55(0.154)	0.575(0.183)
		FPR	0(0)	0(0)	0.033(0.103)	0.017(0.075)
	mediator	TPR	0.138(0.172)	0.162(0.147)	0.125(0.128)	0.062(0.111)
		FPR	0.015(0.013)	0.022(0.021)	0.018(0.011)	0.019(0.016)
	edge	TPR	0.11(0.137)	0.13(0.117)	0.1(0.103)	0.05(0.089)
		FPR	0.014(0.013)	0.021(0.02)	0.018(0.011)	0.019(0.016)
RF	exposure	TPR	0.35(0.235)	0.5(0.162)	0.55(0.154)	0.625(0.222)
		FPR	0.017(0.075)	0.017(0.075)	0.033(0.149)	0.033(0.149)
	mediator	TPR	0.388(0.319)	0.512(0.25)	0.5(0.181)	0.575(0.245)
		FPR	0.078(0.051)	0.106(0.064)	0.096(0.045)	0.111(0.053)
	edge	TPR	0.33(0.27)	0.4(0.195)	0.43(0.163)	0.48(0.199)
		FPR	0.077(0.052)	0.106(0.065)	0.099(0.05)	0.112(0.054)
				Size 500		
LR	exposure	TPR	0.5(0)	0.525(0.112)	0.525(0.112)	0.6(0.205)
		FPR	0(0)	0(0)	0(0)	0(0)
	mediator	TPR	0.2(0.208)	0.275(0.18)	0.275(0.228)	0.388(0.172)
		FPR	0.029(0.018)	0.034(0.02)	0.028(0.016)	0.036(0.031)
	edge	TPR	0.13(0.117)	0.2(0.13)	0.21(0.165)	0.23(0.073)
		FPR	0.029(0.017)	0.035(0.021)	0.03(0.017)	0.04(0.029)
RF	exposure	TPR	0.5(0)	0.5(0)	0.5(0)	0.5(0)
		FPR	0(0)	0(0)	0(0)	0(0)
	mediator	TPR	0.75(0.199)	0.825(0.143)	0.888(0.128)	0.938(0.111)
		FPR	0.197(0.043)	0.209(0.047)	0.212(0.042)	0.234(0.034)
	edge	TPR	0.57(0.149)	0.66(0.114)	0.7(0.103)	0.79(0.165)
		FPR	0.192(0.042)	0.206(0.045)	0.209(0.042)	0.232(0.032)

variables such as age, gender, nutrition variables, metabolomics variables and BMI value. There are 142 nutrition variables that are transformed from a food intake questionnaire for each subject. There are $2,321$ metabolomics feature that could be matched to metabolites for each subject. A network between the metabolites is also known. To analyze the data, we consider nutrition variables as the exposure, metabolomics variables as the mediator and binary overweight status based on BMI as the outcome (BMI¿24.9). Then we apply the two greedy algorithms on this dataset to find combination of metabolomics features and single or multiple nutrition variable that are highly predictive about overweight status. To facilitate the interpretation for the selected metabolomics features in each subnetwork, we conduct a Mummichog analysis (Li et al., 2013b) to find significant pathways for those metabolites.

Figure 4.2 shows the network mediator for single nutrition variable 'cholesterol'. One significant pathway from Mummichog analysis is Bile acid biosynthesis, which has been reported to be related to human obesity in Haeusler et al. (2016); Tomkin and Owens (2016); Ma and Patti (2014); Haeusler et al. (2016). Another significant pathway from Mummichog analysis is Arginine and Proline Metabolism, which also has been report to have an impact on BMI in the literature (Wu et al., 2009; Martin-Lorenzo et al., 2015).

Figure 4.3 shows a selected subnetwork combination with multiple nutrition variable and multiple metabolomics variable that are highly predictive about the overweight status. One significant pathway from Mummichog analysis is Vitamin D3 (cholecalciferol) metabolism, which has been reported to BMI in Araghi et al. (2015); Bikle (2014); Cipriani et al. (2014).

Overall, we find easily find biological interpretation for the pathways identified in predictive network mediator for single and multiple nutrition variable.

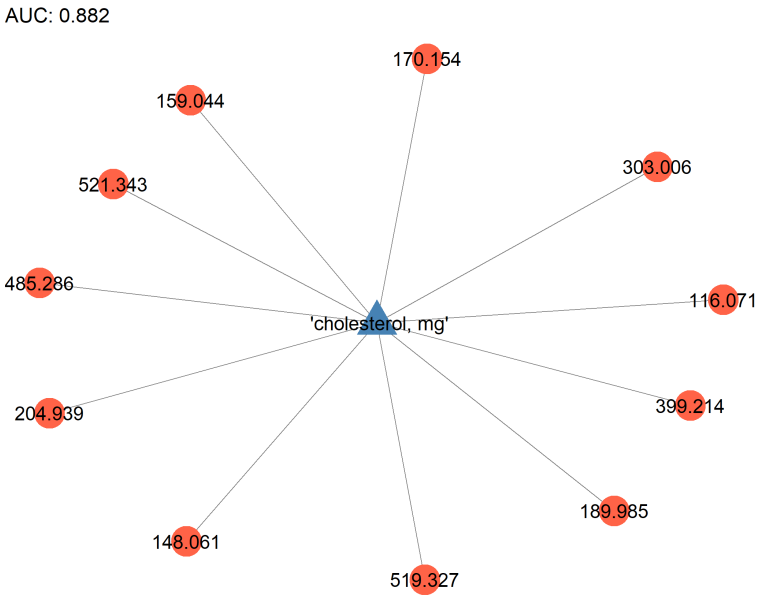


Figure 4.2: Predictive network mediator for single nutrition variable. Triangle denote nutrition variable and circle denote metabolites. The value on each circle is the mz value for metabolite.

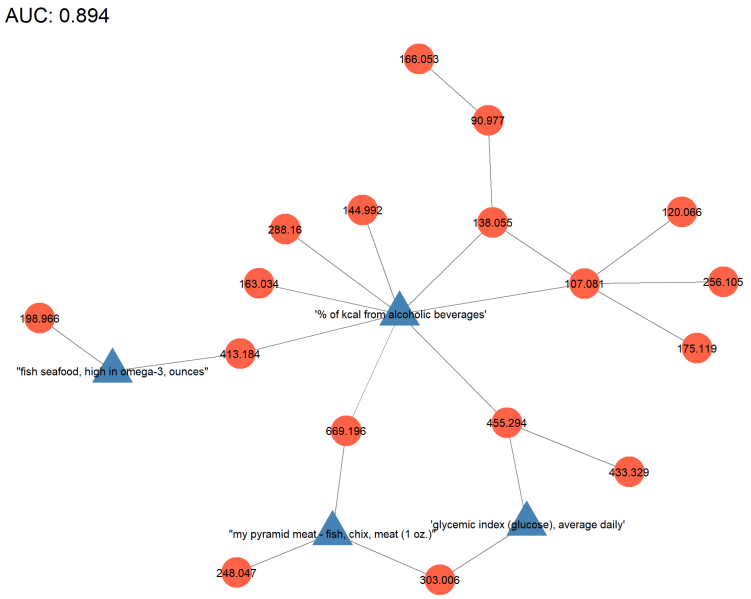


Figure 4.3: Predictive network mediator for multiple nutrition variables. Triangle denote nutrition variable and circle denote metabolites. The value on each circle is the mz value for metabolite.

4.5 Discussion

In this paper, we propose a new mediation analysis framework focusing on predictive modeling. We propose new definitions for predictive exposure, predictive mediator and predictive network mediator. Estimation procedure is also proposed to identify predictive exposure and predictive mediator. Following the definition of predictive mediator, we propose to build a network that combines the mediation dependence between mediator and exposure and the functional dependence between mediators. Then we propose two greedy algorithms that can incorporate various machine learning algorithms to find subnetwork of exposures and mediators from this combined network that are highly predictive about the outcome. Simulation studies show that the estimation procedure can identify predictive exposure and predictive mediator accurately. Application to a real dataset also proves that our proposed greedy algorithm can detect subnetwork that are not only highly predictive about the outcome, but also have meaningful biological interpretation.

Chapter 5

Future work

In this dissertation, we propose several new statistical methods to analyze biomedical network data. In the first project, we propose a novel prior, the thresholded graph Laplacian Gaussian (TGLG) prior, to perform network marker selection over the large-scale network under the GLM framework. In the second project, we propose a unified framework for network feature selection from the metabolic network along with optimal matching detection. In the third project, we propose a new mediation analysis framework for biomedical network data focusing on predictive modeling. In the end, we identify several directions for future work.

For the first project, first we can apply the TGLG prior for network marker selection under other modeling framework such as the survival model and the generalized mixed effects model. Second, the current posterior computation can be further improved by utilizing the parallel computing techniques within each iteration of the MCMC algorithm, for updating the massive latent variables simultaneously. Third, another promising direction is to use the integrated nested laplace approximations (INLA) for Bayesian approximating computation taking advantages of the TGLG prior involving high-dimensional Gaussian latent variables.

For the second project, metabolomics data is very complex, which brings lots of interesting and difficult statistical issues. As we mentioned above, many features have lots of 0's, either because the metabolite is truly non-present in the samples, or because the low peaks cannot be differentiated form noise using current technology. Imputing these missing data in features could improve the power for statistical inference in analyzing metabolomics data. This could be seen as a possible extension of our paper. Another possible extension is to develop a systematic Bayesian modeling framework for feature selection over the network and while addressing the multiple matching issue.

Appendix A

Appendix for Chapter 2

A.0.1 Regularity conditions

First, we introduction the following notations. We define a pre-specified upper bound $\bar{r}_n$ for model size: $|\xi| \leq \bar{r}_n$. For two sequences a_n and b_n, let $a_n = o(b_n)$ denote $\lim_{n\to\infty} a_n/b_n = 0$. Denote by $a \vee b$ the max number between a and b. Define $\triangle(r_n) = \inf_{\xi:|\xi|=r_n} \sum_{j:j\notin\xi} |\beta_j^*|$ and $D(R) = 1 + R \times \sup_{|h|\leq R} |a'(h)| \times \sup_{|h|\leq R} |g^{-1}(h)|$.

We consider the following conditions: (C1) $\bar{r}_n \log(1/\epsilon_n^2) = o(n\epsilon_n^2)$; (C2) $\bar{r}_n \log p_n = o(n\epsilon_n^2)$; (C3) $r_n = o(p_n)$; (C4) $\triangle(r_n) = o(\epsilon_n^2)$; (C5) $(\bar{r}_n + q)\log D((\bar{r}_n + q)(n\epsilon_n^2(\sigma_\alpha^2 \vee \sigma_\omega^2)/2)^{1/2}) = o(n\epsilon_n^2)$; (C6) $1 \leq r_n \leq \bar{r}_n < p_n$; and (C7) $\inf\{\sigma_j^2\}_{j=1}^{r_n} > 0$, where the sequence of $\sigma_1^2, \ldots, \sigma_{r_n}^2$ are defined as follows: Let $\Lambda_\gamma = (\mathbf{L} + \varepsilon \mathbf{I}_p)^{-1}$. After a permutation for rows and columns, Λ_γ can be decomposed as $\tilde{\Lambda}_\gamma = \begin{pmatrix} \tilde{\Lambda}_{11} & \tilde{\Lambda}_{12} \\ \tilde{\Lambda}_{21} & \tilde{\Lambda}_{22} \end{pmatrix}$, where $\tilde{\Lambda}_{11}$ is a numerical value and $\tilde{\Lambda}_{21}$ and $\tilde{\Lambda}_{22}$ are the corresponding $p_n - 1$ vector and $(p_n - 1) \times (p_n - 1)$ submatrix of $\tilde{\Lambda}_\gamma$. Set $\sigma_1^2 = \tilde{\Lambda}_\gamma^R = \tilde{\Lambda}_{11} - \tilde{\Lambda}_{12}\tilde{\Lambda}_{22}^{-1}\tilde{\Lambda}_{21}$. Consider the same procedure for $\tilde{\Lambda}_{22}$ and we can get $\sigma_2^2 = \tilde{\Lambda}_{22}^R$. Repeat the above procedure, then a sequence $\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_{r_n}^2\}$ can be obtained.

Condition (C7) is not a strong condition and we conduct an empirical study to show the correctness of condition (C7). We generate scale free network and random

network with different edge probability using R package igraph. The number of nodes we consider ranges from 500 to 5,000. We set $r_n = \sqrt{p_n}$ and fix $\varepsilon = 10^{-5}$. Figure A.1 shows the value of $\inf\{\sigma_j^2\}_{j=1}^{r_n} > 0$ defined in condition (C7) with different number of nodes. As we can see, all infimum values are bounded away from 0.

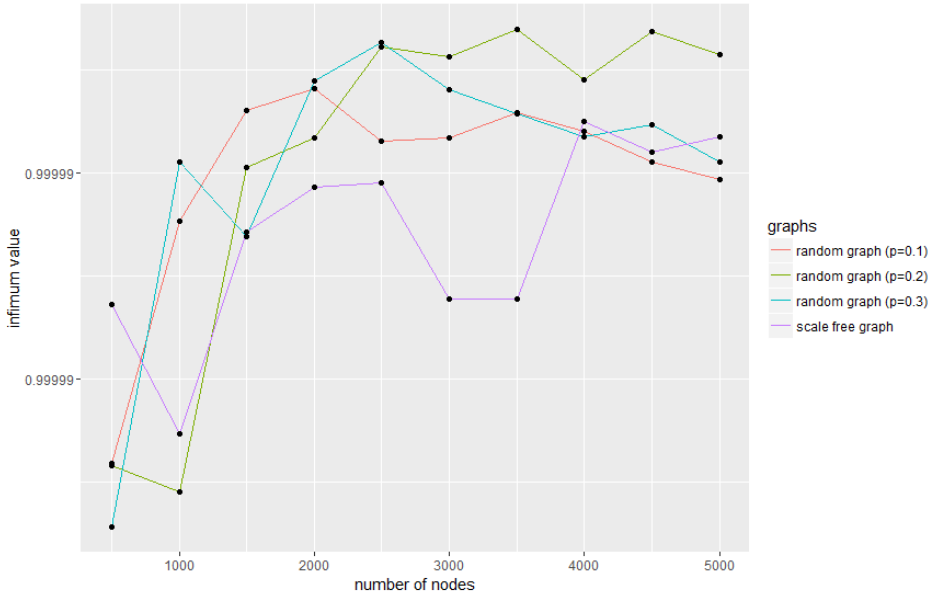


Figure A.1: An illustration example for condition (C7)

A.0.2 Lemmas

Given Theorem 1, the proof of Theorem 2 directly follows proof of Theorem 4 in Jiang (2007) and Theorem 1 in Song and Liang (2015). To prove Thoerem 1, we need to first introduce the following two lemmas.

Lemma 1. *For a one dimension Gaussian random* $Y \sim \mathrm{N}(0, \sigma^2)$ *with* $\sigma^2 > 0$*, denote* $P_{\sigma^2}(\mathcal{C}) = P(Y \in \mathcal{C})$ *for set* $\mathcal{C} \in \mathcal{R}$*. We have (i)* $P_{\sigma^2}(\mathcal{C}) \leq P_{\sigma^2}(\mathcal{C} - z)$ *for* $\mathcal{C} = \{Y \mid |Y| > \lambda\}$ *and* $z \in \mathcal{R}$*; (ii)* $\sigma_1^2 \leq \sigma_2^2 \Rightarrow P_{\sigma_1^2}(\mathcal{C}) \leq P_{\sigma_2^2}(\mathcal{C})$*.*

Proof. (i) First, without loss of generality, we assume $z > 0$. Let $\phi(y)$ denote the

density for $Y \sim \mathrm{N}(0, \sigma^2)$. Then we have

$$\begin{aligned}\mathrm{P}_{\sigma^2}(\mathcal{C}) - \mathrm{P}_{\sigma^2}(\mathcal{C} - z) &= \int_{\lambda}^{\infty} \phi(y)dy + \int_{-\infty}^{-\lambda} \phi(y)dy - \int_{\lambda - z}^{\infty} \phi(y)dy - \int_{-\infty}^{-\lambda - z} \phi(y)dy \\ &= \int_{-\lambda - z}^{-\lambda} \phi(y)dy - \int_{\lambda - z}^{\lambda} \phi(y)dy = \int_{\lambda}^{\lambda + z} \phi(y)dy - \int_{\lambda - z}^{\lambda} \phi(y)dy \leq 0.\end{aligned}$$

The inequality holds since $\phi(y)$ is symmetric around 0 and $\phi(y)$ is smaller in $[\lambda, \lambda + z]$ than $[\lambda - z, \lambda]$ since $\lambda > 0$.

(ii) Assume $Y_1 \sim \mathrm{N}(0, \sigma_1^2)$ and $Y_2 \sim \mathrm{N}(0, \sigma_2^2)$. We can have $Y_2 \overset{d}{=} Y_1 + Z$ with $Z \sim \mathrm{N}(0, \sigma_2^2 - \sigma_1^2)$ and Y_1, Z are independent.

$$\mathrm{P}_{\sigma_2^2}(\mathcal{C}) = \mathrm{E}\{P_{\sigma_1^2}[Y_1 \in \mathcal{C} - Z \mid Z]\} \geq \mathrm{E}\{P_{\sigma_1^2}[Y_1 \in \mathcal{C}]\} = P_{\sigma_1^2}(\mathcal{C}).$$

□

Lemma 2. *Suppose a $p-$dimension multivariate Gaussian variable $Y \sim \mathrm{N}(\mathbf{0}_p, \Sigma)$. We partition Y as*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Y_1 is of dimension one and Y_2 is of dimension $p - 1$. $\sigma_{11}, \sigma_{12}, \sigma_{21}$ and Σ_{22} are the corresponding value, vector and sub-matrix from Σ. For $\mathcal{C}_1 = \{y \mid |y| > \lambda\}$ and $\mathcal{C}_2 \in \mathcal{R}^{p-1}$, We have:

$$\mathrm{P}(Y_1 \in \mathcal{C}_1, Y_2 \in \mathcal{C}_2) \geq \mathrm{P}_{\sigma_{11.2}}(\mathcal{C}_1)\mathrm{P}_{\Sigma_{22}}(\mathcal{C}_2),$$

where $\sigma_{11.2} = \sigma_{11} - \sigma_{12}\Sigma_{22}^{-1}\sigma_{21}$.

Proof. Note $V = Y_1 - \sigma_{12}\Sigma_{22}^{-1}Y_2 \sim \mathrm{N}(0, \sigma_{11.2})$ and $Y_2 \sim \mathrm{N}(0, \Sigma_{22})$. V, Y_2 are indepen-

dent.

$$\begin{aligned} \mathrm{P}(Y_1 \in \mathcal{C}_1, Y_2 \in \mathcal{C}_2) &= \mathrm{E}\{\mathrm{P}[V \in \mathcal{C}_1 - \sigma_{12}\Sigma_{22}^{-1}Y_2, Y_2 \in \mathcal{C}_2|Y_2]\} \\ &= \mathrm{E}\{\mathrm{P}[V \in \mathcal{C}_1 - \sigma_{12}\Sigma_{22}^{-1}Y_2|Y_2]I_{\mathcal{C}_2}(Y_2)\} \\ &\geq \mathrm{E}\{\mathrm{P}[V \in \mathcal{C}_1|Y_2]I_{\mathcal{C}_2}(Y_2)\} = \mathrm{P}_{\sigma_{11.2}}(\mathcal{C}_1)\mathrm{P}_{\Sigma_{22}}(\mathcal{C}_2). \end{aligned}$$

□

A.0.3 Proof for Thoerem 1

Now we are in a good position to prove Theorem 1.

Proof. We have $\boldsymbol{\gamma} \sim \mathrm{N}(0, \sigma_\gamma^2\Lambda_\gamma)$. Set $\boldsymbol{\gamma}_{\xi_n} = \{\gamma_j, j \in \xi_n\}$ as a vector of length $|\xi_n|$ and $\boldsymbol{\gamma}_{-\xi_n} = \{\gamma_j, j \notin \xi_n\}$. Write the thresholding parameter as λ_n to denote that λ_n might increase as n increases. Let $\mathcal{C} = \mathcal{C}_1 \bigcup \mathcal{C}_2 = \{\bigcup_{j\in\xi_n}\{|\gamma_j| > \lambda_n\}\} \bigcup \{\bigcup_{j\notin\xi_n}\{|\gamma_j| \leq \lambda_n\}\}$.

$$\pi(\xi = \xi_n) = \pi(\boldsymbol{\gamma}_{\xi_n} \in \mathcal{C}_1, \boldsymbol{\gamma}_{-\xi_n} \in \mathcal{C}_2).$$

Denote $\xi_n = (i_1, \cdots, i_{|\xi_n|})$. Set $\mathcal{C}_{i_1} = \{|\gamma_{i_1}| > \lambda_n\}, \mathcal{C}_{-i_1} = \mathcal{C}\backslash\mathcal{C}_1$. Let $\sigma_{i_1} = \sigma_{11} - \sigma_{12}\Sigma_{22}^{-1}\sigma_{21}$ for

$$\begin{pmatrix} \gamma_{i_1} \\ \boldsymbol{\gamma}_{-i_1} \end{pmatrix} \sim \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

According to Lemma 2, we have :

$$\begin{aligned} \pi(\xi = \xi_n) &= \pi(\boldsymbol{\gamma}_{\xi_n} \in \mathcal{C}_1, \boldsymbol{\gamma}_{-\xi_n} \in \mathcal{C}_2) = \pi(\gamma_{i_1} \in \mathcal{C}_{i_1}, \boldsymbol{\gamma}_{-i_1} \in \mathcal{C}_{-i_1}) \\ &\geq \pi_{\sigma_{i_1}}(\gamma_{i_1} \in \mathcal{C}_{i_1})\pi_{\Sigma_{22}}(\boldsymbol{\gamma}_{-i_1} \in \mathcal{C}_{-i_1}). \end{aligned}$$

Similarly, we can apply the same procedure to $\pi_{\Sigma_{22}}(\boldsymbol{\gamma}_{-i_1} \in \mathcal{C}_{-i_1})$ until we have gone

through all the elements in ξ_n. Then we have

$$\pi(\xi = \xi_n) \geq \{\Pi_{j\in\xi_n}\pi_{\sigma_j}(|\gamma_j| > \lambda_n)\} \times \pi_{\Sigma_{-\xi_n}}(\boldsymbol{\gamma}_{-\xi_n} \in \mathcal{C}_2),$$

where $\Sigma_{-\xi_n} = \sigma_\gamma^2\Lambda_\gamma(i,j)_{i,j\notin\xi_n}$. By (C7), we have $\sigma_1^2 = \inf_{j\in\xi_n}\sigma_j > 0$. According to Lemma 1, $\pi_{\sigma_j}(|\gamma_j| > \lambda_n) \geq \pi_{\sigma_1^2}(|\gamma_j| > \lambda_n)$, for all $j \in \xi_n$.

According to Anderson (1955), for a p dimension multivariate Gaussian random variable $Y \sim N_p(0,\Sigma)$, we have $\Sigma_1 \leq \Sigma_2 \Rightarrow P_{\Sigma_1}(\mathcal{C}) \geq P_{\Sigma_2}(\mathcal{C})$ for every centrally symmetric convex set $\mathcal{C}$. In our case, $\mathcal{C}_2$ is a centrally symmetric convex set. According to Chung (1997), the eigenvalue for a p dimension Graph Laplacian matrix L is : $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p$. So the maximum eigenvalue for $\Sigma_{-\xi_n}$ is smaller than σ_γ^2/ϵ since the maximum eigenvalue for Λ_γ is smaller than $1/\epsilon$. Then we could have $\pi_{\Sigma_{-\xi_n}}(\boldsymbol{\gamma}_{-\xi_n} \in \mathcal{C}_2) \geq \pi_{\sigma_\gamma^2/\epsilon I_{p_n-|\xi_n|}}(\boldsymbol{\gamma}_{-\xi_n} \in \mathcal{C}_2) = (\pi(|Y| \leq \lambda_n))^{p_n-|\xi_n|}$ where $Y \sim \mathrm{N}(0,\sigma_\gamma^2/\epsilon)$. Note that $\pi(\xi = \xi_n) \geq (\pi_{\sigma_1^2}(|\gamma_j| > \lambda_n))^{|\xi_n|}(\pi_{\sigma_\gamma^2/\epsilon}(|\gamma_j| \leq \lambda_n))^{p_n-|\xi_n|}$. Take λ_n such that $\min\{\pi_{\sigma_1^2}(|\gamma_j| > \lambda_n), \pi_{\sigma_\gamma^2/\epsilon}(|\gamma_j| > \lambda_n)\} = |\xi_n|/p_n$. Then we have $-\log\pi(\xi = \xi_n) \leq -|\xi_n|\log(|\xi_n|/p_n) - (p_n - |\xi_n|)\log(1 - |\xi_n|/p_n) \leq |\xi_n|\log p_n + |\xi_n| = o(n\epsilon_n^2)$, since $|\xi_n| = o(p_n)$ and $|\xi_n|\log p_n \leq \bar{r}_n \log p_n = o(n\epsilon_n^2)$.

Given ξ_n and σ_α^2, we have $\alpha_j \overset{\text{i.i.d}}{\sim} \mathrm{N}(0,\sigma_\alpha^2)$ and $\beta_j = \alpha_j$ for $j \in \xi_n$. For $j \in \xi_n$, denote ϕ the infimum of the density for $\mathrm{N}(0,\sigma_\alpha^2)$ for all $\{\beta_j^* \pm \eta\epsilon_n^2/|\xi_n|\}_{j\in\xi_n}$. So $-\log\pi(\boldsymbol{\beta}_\xi \in B(\xi_n,\eta)|\xi = \xi_n) \leq -|\xi_n|\log(2\phi\eta\epsilon_n^2/|\xi_n|) \leq |\xi_n|\log|\xi_n| + C|\xi_n| + |\xi_n|\log(1/\epsilon_n^2) = o(n\epsilon_n^2)$ where C is some constant, since $|\xi_n|\log(1/\epsilon_n^2) \leq \bar{r}_n\log(1/\epsilon_n^2) = o(n\epsilon_n^2)$ and $|\xi_n|\log|\xi_n| \leq \bar{r}_n\log p_n = o(n\epsilon_n^2)$.

By Mill's ratio, for all $j \in \xi_n$, we have $\pi(|\beta_j| > C_n) \leq 2e^{-C_n^2/(2\sigma_\alpha^2)}/\sqrt{2\pi C_n^2/\sigma_\alpha^2}$ with $\beta_j \sim \mathrm{N}(0,\sigma_\alpha^2)$. Choose $C_n = \sqrt{n\epsilon_n^2\sigma_\alpha^2/2}$. Then we have $\pi(|\beta_j| > C_n) \leq e^{-n\epsilon_n^2/4}$ for large enough n. This completes the proof for Theorem 1. □

A.0.4 Proof for Thoerem 2

Next, we show the proof for Theorem 2.

Assume $\mathcal{P}_n$ is a sequence of sets of probability densities and ϵ_n is a sequence of positive number. Denote $N(\epsilon_n, \mathcal{P}_n)$ as the minimal number of Hellinger balls of radius ϵ_n that are needed to cover $\mathcal{P}_n$. Denote $d_0(f, f^*) = \int f^* ln(f^*/f)$ as the Kullback-Leibler divergence between two densities f and f^* and define $d_t(f, f^*) = t^{-1}(\int f^*(f^*/f)^t - 1)$ for any $t > 0$. It's easy to see that d_t decrease to d_0 as t decrease to 0. Define $\hat{\pi}(\epsilon) = \pi[d(p, p^*) \geq \epsilon | D_n]$, for any $\epsilon > 0$.

Following from Theorem 6 of Jiang (2005) and Proposition 1 of Jiang (2007), we have the Following **Lemma 3**.

Lemma 3 Assume there is a sequence $\epsilon_n \in (0, 1]$ such that $n\epsilon_n^2 \to \infty$. If for all large enough n, the following conditions hold:

(a) $\ln \mathrm{N}(\epsilon_n/4, \mathcal{P}_n) \leq n\epsilon_n^2/16$;

(b) $\pi(\mathcal{P}_n^c) \leq e^{-n\epsilon_n^2/8}$;

(c) $\pi[p : d_t(p, p^*) \leq \epsilon_n^2/64] \geq e^{-n\epsilon_n^2/64}$ for some $t > 0$.

Then under (a), (b), (c), we have:

(i) $P\{\hat{\pi}(\epsilon_n) \geq 2e^{-n\epsilon_n^2 \min\{1/32, t/64\}}\} \leq 2e^{-n\epsilon_n^2 \min\{1/32, t/64\}}$;

(ii)$E\hat{\pi}(\epsilon_n) \leq 4e^{-n\epsilon_n^2 \min\{1/16, t/32\}}$

Next, we prove **Theorem 2** by checking that our model settings satisfy conditions (a), (b), (c) in **Lemma 3**.

Proof. **Check condition (c)**:

Consider the case for $t = 1$. The GLM density is $f(y, h) = \exp\{a(h)y + b(h) + c(y)\}$. Let $p^* = f(y, h^*)$ with $h^* = \mathbf{x}^T\boldsymbol{\beta}^* + \mathbf{z}^T\boldsymbol{\omega}^*$ and $p_{\xi_n} = f(y, h_{\xi_n})$ with $h_{\xi_n} = \mathbf{z}^T\boldsymbol{\omega} + \mathbf{x}_{\xi_n}^T\boldsymbol{\beta}_{\xi_n} = \mathbf{z}^T\boldsymbol{\omega} + \sum_{j \in \xi_n} x_j\beta_j$. As shown in the proof of Theorem 4 of Jiang (2007), when h^* and h_{ξ_n} are close enough, $d_1(p_{\xi_n}, p^*)$ can be put in the form as $d_1(p_{\xi_n}, p^*) = E_x g(\tilde{h})(h^* - h_{\xi_n})$, where g is a continuous derivative function in a neighborhood of

h^* and $\tilde{h}$ is an intermediate point between h^* and h_{ξ_n}.

Denote r_n as the model size for ξ_n ($|\xi_n| = r_n$) and $Q = \{1, \cdots, q\}$. We have $|\tilde{h} - h^*| \leq |h^* - h_{\xi_n}| \leq |\sum_{k\in Q} z_k(\omega_k - \omega_k^*)| + |\sum_{j\in\xi_n} x_j(\beta_j - \beta_j^*)| + |\sum_{j\notin\xi_n} x_j\beta_j^*| \leq qM\max_{k\in Q}\{|\omega_k - \omega_k^*|\} + r_nM\max_{j\in\xi_n}\{|\beta_j - \beta_j^*|\} + M\Delta(r_n)$. If there exists small enough δ_1 and δ_2 such that $\beta_j \in (\beta_j^* \pm \delta_1)$ for all $j \in \xi_n$ and $\omega_k \in (\omega_k^* \pm \delta_2)$ for all $k \in Q$, we could have that $|\tilde{h} - h^*| \leq |h^* - h_{\xi_n}| \leq M\Delta(r_n) + Mr_n\delta_1 + Mq\delta_2$. Here we have $\Delta(r_n) = o(\epsilon_n^2)$ by condition (C4).

For sufficiently small δ_1, δ_2, $|g(\tilde{h})|$ is bounded since $|\tilde{h}| \leq |h^*| + |\tilde{h} - h^*| \leq B_0 + M\Delta(r_n) + Mr_n\delta_1 + Mq\delta_2$ is bounded, where $B_0 = \lim_{n\to\infty}\sum_{j=1}^{p_n}|\beta_j^*|$. Then we could have $d_1(p_{\xi_n}, p^*) \leq C(M\Delta(r_n) + Mr_n\delta_1 + Mq\delta_2)$ for some constant C and small enough δ_1, δ_2. Take $\delta_1 = \eta_1\epsilon_n^2/(Mr_n)$ and $\delta_2 = \eta_2\epsilon_n^2/(Mq)$ for small enough η_1, η_2. Then we could have $d_1(p_{\xi_n}, p^*) \leq \epsilon_n^2/64$ for large enough n and small enough η_1, η_2. So we can conclude that $\{\xi = \xi_n, \beta_j \in (\beta_j^* \pm \eta_1\epsilon_n^2/(Mr_n)), j \in \xi_n, \omega_k \in (\omega_k^* \pm \eta_2\epsilon_n^2/(Mq)), k \in Q\} \subset \{p : d_1(p_{\xi_n}, p^*) \leq \epsilon_n^2/64\}$.

As shown in the proof of Theorem 1, we have $-\ln\pi(\xi = \xi_n) = o(n\epsilon_n^2)$ and $-\ln\pi(\beta_\xi \in \{\beta_j^* \pm \eta_1\epsilon_n^2/(M|\xi_n|)\}_{j\in\xi_n}|\xi = \xi_n) = o(n\epsilon_n^2)$. Similarly, it's easy to show $-\ln\pi(\boldsymbol{\omega} \in \{\omega_k^* \pm \eta_2\epsilon_n^2/(Mq)\}_{k\in Q}) = o(n\epsilon_n^2)$. Then we can have $-\ln\pi\{p : d_1(p_{\xi_n}, p^*) \leq \epsilon_n^2/64\} = o(n\epsilon_n^2)$.

Check condition (a): Let $\mathcal{P}_n = \{f(y; \xi, \beta_\xi, \boldsymbol{\omega}) : |\xi| \leq \bar{r}_n, |\beta_j|_{j\in\xi} \leq C_n, |\omega_k|_{k\in Q} \leq C_n\}$, for some $C_n > 0$. For each model ξ in $\mathcal{P}_n$, there are $|\xi| + q$ nonzero elements, with each element bounded by $[-C_n, C_n]$. It takes at most $[C_n/\delta + 1]^{|\xi|+q}$ balls with radius $\delta(\delta > 0)$ to cover the parameter space of model ξ. For each model ξ with size $|\xi| = r$, there are at most p_n^r models for $r = 0, 1, \cdots, \bar{r}_n$. Then we could have that the number of radius-δ balls $N(\delta)$ needed to cover the parameter space in $\mathcal{P}_n$ is at most $\sum_{r=0}^{\bar{r}_n} p_n^r[C_n/\delta + 1]^{r+q}$, which is bounded by $(\bar{r}_n + 1)p_n^{\bar{r}_n}[C_n/\delta + 1]^{\bar{r}_n+q}$. This means that for any density in $\mathcal{P}_n$ that can be represented by a set of regression parameters $\{\omega_k^u\}_1^q, \{u_j\}_1^{p_n}$, it must fall in these $N(\delta)$ balls with center $\{\omega_k^v\}_{k=1}^q, \{v_j\}_{j=1}^{p_n}$,

i.e. $\{\omega_k^v \pm \delta\}_{k=1}^q, \{v_j \pm \delta\}_{j=1}^{p_n}$, where u_j and v_j are zero for the same model ξ and $|\xi| \leq \bar{r}_n$.

Consider the corresponding GLM densities $f_u = \exp\{a(h_u)y + b(h_u) + c(y)\}$ and $f_v = \exp\{a(h_v)y + b(h_v) + c(y)\}$ with $h_u = \sum_{k=1}^q z_k\omega_k^u + \sum_{j=1}^{p_n} x_j u_j$ and $h_v = \sum_{k=1}^q z_k\omega_k^v + \sum_{j=1}^{p_n} x_j v_j$. It's easy to show that the Hellinger distance between f_u and f_v is smaller than the square root KL divergence, $d(f_u, f_v) \leq \sqrt{d_0(f_u, f_v)}$. The KL divergence is $d_0(f_u, f_v) = E_{\mathbf{z},\mathbf{x}} \int f_v(\ln f_v - \ln f_u)\nu_y(dy)$. Integrate out y and apply a Taylor expansion. We can show that $d_0(f_u, f_v) \leq E_{\mathbf{z},\mathbf{x}}(a'(\tilde{h})\psi(h_v) + b'(\tilde{h}))(h_v - h_u)$, where $\psi = -b'/a$ and $\tilde{h}$ is an intermediate point between h_v and h_u. By definition, we could have $h_u, h_v, \tilde{h}$ are all bounded by $(\bar{r}_n + q)C_n$. Note that $|h_u - h_v| = |\sum_{k\in Q} z_k(\omega_k^u - \omega_k^v)| + |\sum_{j\in\xi} x_j(u_j - v_j)| \leq M(\bar{r}_n + q)\delta$ since $|u_j - v_j| \leq \delta$ and $|\omega_k^u - \omega_k^v| \leq \delta$. Therefore,

$$d_0(f_u, f_v) \leq 2 \sup_{|h|\leq(\bar{r}_n+q)C_n} |a'(h)| \times \sup_{|h|\leq(\bar{r}_n+q)C_n} |\psi(h)| M(\bar{r}_n + q)\delta$$

and

$$d(f_u, f_v) \leq \{2 \sup_{|h|\leq(\bar{r}_n+q)C_n} |a'(h)| \times \sup_{|h|\leq(\bar{r}_n+q)C_n} |\psi(h)| M(\bar{r}_n + q)\delta\}^{1/2}$$

Let $\delta = \epsilon_n^2/\{32 \sup_{|h|\leq(\bar{r}_n+q)C_n} |a'(h)| \times \sup_{|h|\leq(\bar{r}_n+q)C_n} |\psi(h)| M(\bar{r}_n + q)\}$ and we have $d(f_u, f_v) \leq \epsilon_n/4$. So we have the Hotelling covering number:

$$\begin{aligned}
\ln N(\epsilon_n/4, \mathcal{P}_n) &\leq \ln N(\delta) \\
&\leq \ln(\bar{r}_n + 1) + \bar{r}_n \ln p_n + \bar{r}_n \ln(32\epsilon_n^{-2} \sup_{|h|\leq\bar{r}_n C_n} |a'(h)| \times \sup_{|h|\leq\bar{r}_n C_n} |\psi(h)| + 1) \\
&\leq \ln(\bar{r}_n + 1) + \bar{r}_n \ln p_n + (\bar{r}_n + q)\ln D((\bar{r}_n + q)C_n) + \bar{r}_n \ln 32
\end{aligned}$$

Since $\bar{r}_n \ln p_n = o(n\epsilon_n^2)$ and $(\bar{r}_n+q)\ln D((\bar{r}_n+q)C_n) = o(n\epsilon_n^2)$ with $C_n = \sqrt{n\epsilon_n^2(\sigma_\alpha^2 \vee \sigma_\omega^2)/2}$, we have $\ln N(\epsilon_n/4, \mathcal{P}_n) = o(n\epsilon_n^2)$

Check condition (b): $\mathcal{P}_n = \{f(y; \xi, \beta_\xi) : |\xi| \leq \bar{r}_n, |\beta_j|_{j\in\xi} \leq C_n, |\omega_k|_{k\in Q} \leq C_n\}$, for some $C_n > 0$. By the definition of $\bar{r}_n$, we have $\pi(|\xi| > \bar{r}_n) = 0$. So we could

get $\pi(\mathcal{P}_n^c) \leq \max_{\xi:|\xi|\leq \bar{r}_n} \pi(\xi)\pi(\cup_{j\in\xi}[|\beta_j| > C_n] \mid \xi) + \pi(\cup_{k\in Q}[|\omega_k| > C_n])$. By Mill's ratio, we have $\pi(|\beta_j| > C_n) \leq 2e^{-C_n^2/(2\sigma_\alpha^2)}/\sqrt{2\pi C_n^2/\sigma_\alpha^2}$ for $\beta_j \sim N(0, \sigma_\alpha^2)$ and $j \in \xi$. Similarly, $\pi(|\omega_k| > C_n) \leq 2e^{-C_n^2/(2\sigma_\omega^2)}/\sqrt{2\pi C_n^2/\sigma_\omega^2}$ for $\omega_k \sim N(0, \sigma_\omega^2)$ and $k \in Q$. Choose $C_n = \sqrt{n\epsilon_n^2(\sigma_\alpha^2 \vee \sigma_\omega^2)/2}$. Then we have $\pi(|\beta_j| > C_n) \leq e^{-n\epsilon_n^2/4}$ and $\pi(|\omega_k| > C_n) \leq e^{-n\epsilon_n^2/4}$ for large enough n. So $\pi(\mathcal{P}_n^c) \leq (2 + \bar{r}_n + q)e^{-n\epsilon_n^2/4} \leq e^{-n\epsilon_n^2/8}$ for large enough n, since $\ln(2 + \bar{r}_n + q) \leq \bar{r}_n \ln p_n = o(n\epsilon_n^2)$.

This completes the proof for Theorem 2. □

Bibliography

Aggio, R. B. M., Ruggiero, K. and Villas-Bôas, S. G. (2010), 'Pathway activity profiling (papi): from the metabolite profile to the metabolic pathway activity', Bioinformatics **26**(23), 2969–76.

Ailhaud, G., Guesnet, P. and Cunnane, S. C. (2008), 'An emerging risk factor for obesity: does disequilibrium of polyunsaturated fatty acid metabolism contribute to excessive adipose tissue development?', Br J Nutr **100**(3), 461–70.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/18307824*

Akram, M. (2014), 'Citric acid cycle and role of its intermediates in metabolism', Cell biochemistry and biophysics **68**(3), 475–478.

Albert, J. M. (2012), 'Mediation analysis for nonlinear models with confounding', Epidemiology (Cambridge, Mass.) **23**(6), 879.

Albert, J. M. and Nelson, S. (2011), 'Generalized causal mediation analysis', Biometrics **67**(3), 1028–1038.

Anderson, T. W. (1955), 'The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities', Proceedings of the American Mathematical Society **6**(2), 170–176.

Araghi, S. O., van Dijk, S., Ham, A., Brouwer-Brolsma, E., Enneman, A., Sohl, E., Swart, K., Van der Zwaluw, N., van Wijngaarden, J., Dhonukshe-Rutten, R. et al.

(2015), 'Bmi and body fat mass is inversely associated with vitamin d levels in older individuals', The journal of nutrition, health & aging **19**(10), 980–985.

Back, M., Avignon, A., Stanke-Labesque, F., Boegner, C., Attalin, V., Leprieur, E. and Sultan, A. (2014), 'Leukotriene production is increased in abdominal obesity', PLoS One **9**(12), e104593.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/25437865*

Bamber, D. (1975), 'The area above the ordinal dominance graph and the area below the receiver operating characteristic graph', Journal of mathematical psychology **12**(4), 387–415.

Barabási, A.-L., Gulbahce, N. and Loscalzo, J. (2011), 'Network medicine: a network-based approach to human disease', Nature Reviews Genetics **12**(1), 56–68.

Barbieri, M. M., Berger, J. O. et al. (2004), 'Optimal predictive model selection', The annals of statistics **32**(3), 870–897.

Baron, R. M. and Kenny, D. A. (1986), 'The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.', Journal of personality and social psychology **51**(6), 1173.

Barupal, D. K., Haldiya, P. K., Wohlgemuth, G., Kind, T., Kothari, S. L., Pinkerton, K. E. and Fiehn, O. (2012), 'Metamapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity', BMC Bioinformatics **13**, 99.

Bender DA, Mayes PA. The Citric Acid Cycle: The Central Pathway of Carbohydrate, L. . A. A. M. (2015), Harper's illustrated biochemistry, 30e. Eds. Victor W. Rodwell et al., New York: McGraw-Hill.

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015), 'Dirichlet–laplace priors for optimal shrinkage', Journal of the American Statistical Association **110**(512), 1479–1490.

Bikle, D. D. (2014), 'Vitamin d metabolism, mechanism of action, and clinical applications', Chemistry & biology **21**(3), 319–329.

Bogl, L. H., Kaye, S. M., Ramo, J. T., Kangas, A. J., Soininen, P., Hakkarainen, A., Lundbom, J., Lundbom, N., Ortega-Alonso, A., Rissanen, A., Ala-Korpela, M., Kaprio, J. and Pietilainen, K. H. (2016), 'Abdominal obesity and circulating metabolites: A twin study approach', Metabolism **65**(3), 111–21.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/26892522*

Borgatti, S. P., Mehra, A., Brass, D. J. and Labianca, G. (2009), 'Network analysis in the social sciences', Science **323**(5916), 892–895.

Boulet, M. M., Chevrier, G., Grenier-Larouche, T., Pelletier, M., Nadeau, M., Scarpa, J., Prehn, C., Marette, A., Adamski, J. and Tchernof, A. (2015), 'Alterations of plasma metabolite profiles related to adipose tissue distribution and cardiometabolic risk', Am J Physiol Endocrinol Metab **309**(8), E736–46.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/26306599*

Breiman, L. (2001), 'Random forests', Machine learning **45**(1), 5–32.

Brigham, K. L. (2010), 'Predictive health: the imminent revolution in health care', J Am Geriatr Soc **58 Suppl 2**, S298–302.

Burger, R., Bakker, F., Guenther, A., Baum, W., Schmidt-Arras, D., Hideshima, T., Tai, Y.-T., Shringarpure, R., Catley, L., Senaldi, G., Gramatzki, M. and Anderson, K. C. (2003), 'Functional significance of novel neurotrophin-1/b cell-stimulating factor-3 (cardiotrophin-like cytokine) for human myeloma cell growth and survival', Br J Haematol **123**(5), 869–78.

Caldon, C. E. (2014), 'Estrogen signaling and the dna damage response in hormone dependent breast cancers', Front Oncol **4**, 106.

Chan, S. Y. and Loscalzo, J. (2012a), 'The emerging paradigm of network medicine in the study of human disease', Circulation research **111**(3), 359–374.

Chan, S. Y. and Loscalzo, J. (2012b), 'The emerging paradigm of network medicine in the study of human disease', Circ Res **111**(3), 359–74.

Chang, C., Kundu, S. and Long, Q. (2016), 'Scalable bayesian variable selection for structured high-dimensional data', arXiv preprint arXiv:1604.07264 .

Chekouo, T., Stingo, F. C., Guindani, M., Do, K.-A. et al. (2016), 'A bayesian predictive model for imaging genetics with application to schizophrenia', The Annals of Applied Statistics **10**(3), 1547–1571.

Christe, M., Hirzel, E., Lindinger, A., Kern, B., von Flüe, M., Peterli, R., Peters, T., Eberle, A. N. and Lindinger, P. W. (2013), 'Obesity affects mitochondrial citrate synthase in human omental adipose tissue', ISRN obesity **2013**.

Chung, F. R. (1997), Spectral graph theory, Vol. 92, American Mathematical Soc.

Cipriani, C., Pepe, J., Piemonte, S., Colangelo, L., Cilli, M. and Minisola, S. (2014), 'Vitamin d and its relationship with obesity and muscle', International journal of endocrinology **2014**.

Ciruelos Gil, E. M. (2014), 'Targeting the pi3k/akt/mtor pathway in estrogen receptor-positive breast cancer', Cancer Treat Rev **40**(7), 862–71.

Clauset, A., Newman, M. E. and Moore, C. (2004), 'Finding community structure in very large networks', Physical review E **70**(6), 066111.

Cortes, C. and Vapnik, V. (1995), 'Support-vector networks', Machine learning **20**(3), 273–297.

Csardi, G. and Nepusz, T. (2006), 'The igraph software package for complex network research', InterJournal **Complex Systems**, 1695.
URL: *http://igraph.org*

Daniel, R., De Stavola, B., Cousens, S. and Vansteelandt, S. (2015), 'Causal mediation analysis with multiple mediators', Biometrics **71**(1), 1–14.

Das, J. and Yu, H. (2012), 'Hint: High-quality protein interactomes and their applications in understanding human disease', BMC Syst Biol **6**, 92.

de Carvalho, V. I., Jara, A., Hanson, T. E., de Carvalho, M. et al. (2013), 'Bayesian nonparametric roc regression modeling', Bayesian Analysis **8**(3), 623–646.

Di Renzo, L., Galvano, F., Orlandi, C., Bianchi, A., Di Giacomo, C., La Fauci, L., Acquaviva, R. and De Lorenzo, A. (2010), 'Oxidative stress in normal-weight obese syndrome', Obesity **18**(11), 2125–2130.
URL: *http://dx.doi.org/10.1038/oby.2010.50*

Dobra, A. (2009), 'Variable selection and dependency networks for genomewide data', Biostatistics **10**(4), 621–639.

Elshorbagy, A. K., Kozich, V., Smith, A. D. and Refsum, H. (2012), 'Cysteine and obesity: consistency of the evidence across epidemiologic, animal and cellular studies', Curr Opin Clin Nutr Metab Care **15**(1), 49–57.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/22108094*

Falcon, S. and Gentleman, R. (2007), 'Using gostats to test gene lists for go term association', Bioinformatics **23**(2), 257–8.

Fan, J. and Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', Journal of the American statistical Association **96**(456), 1348–1360.

Fang, Z. and Luna, E. J. (2013), 'Supervillin-mediated suppression of p53 protein enhances cell survival', J Biol Chem **288**(11), 7918–29.

Faraggi, D. and Reiser, B. (2002), 'Estimation of the area under the roc curve', Statistics in medicine **21**(20), 3093–3106.

Favennec, M., Hennart, B., Caiazzo, R., Leloire, A., Yengo, L., Verbanck, M., Arredouani, A., Marre, M., Pigeyre, M., Bessede, A., Guillemin, G. J., Chinetti, G., Staels, B., Pattou, F., Balkau, B., Allorge, D., Froguel, P. and Poulain-Godefroy, O. (2015), 'The kynurenine pathway is activated in human obesity and shifted toward kynurenine monooxygenase activation', Obesity (Silver Spring) **23**(10), 2066–74.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/26347385*

Felig, P., Marliss, E. and Cahill, G. F., J. (1969), 'Plasma amino acid levels and insulin secretion in obesity', N Engl J Med **281**(15), 811–6.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/5809519*

Formosa, R., Borg, J. and Vassallo, J. (2017), 'Aryl hydrocarbon receptor (ahr) is a potential tumour suppressor in pituitary adenomas', Endocr Relat Cancer **24**(8), 445–457.

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

Gilkes, D. M. and Semenza, G. L. (2013), 'Role of hypoxia-inducible factors in breast cancer metastasis', Future Oncol **9**(11), 1623–36.

Giouleka, P., Papatheodorou, G., Lyberopoulos, P., Karakatsani, A., Alchanatis, M., Roussos, C., Papiris, S. and Loukides, S. (2011), 'Body mass index is associated with leukotriene inflammation in asthmatics', Eur J Clin Invest **41**(1), 30–8.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/20825465*

Goldsmith, J., Huang, L. and Crainiceanu, C. M. (2014), 'Smooth scalar-on-image regression via spatial bayesian variable selection', Journal of Computational and Graphical Statistics **23**(1), 46–64.

Gonçalves, L., Subtil, A., Oliveira, M. R. and Bermudez, P. (2014), 'Roc curve estimation: An overview', REVSTAT–Statistical Journal **12**(1), 1–20.

Haeusler, R. A., Camastra, S., Nannipieri, M., Astiarraga, B., Castro-Perez, J., Xie, D., Wang, L., Chakravarthy, M. and Ferrannini, E. (2016), 'Increased bile acid synthesis and impaired bile acid transport in human obesity', The Journal of Clinical Endocrinology & Metabolism **101**(5), 1935–1944.

Hand, D. J. and Till, R. J. (2001), 'A simple generalisation of the area under the roc curve for multiple class classification problems', Machine learning **45**(2), 171–186.

Hanley, J. A. and McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (roc) curve.', Radiology **143**(1), 29–36.

Ho, J. E., Larson, M. G., Ghorbani, A., Cheng, S., Chen, M. H., Keyes, M., Rhee, E. P., Clish, C. B., Vasan, R. S., Gerszten, R. E. and Wang, T. J. (2016), 'Metabolomic profiles of body mass index in the framingham heart study reveal distinct cardiometabolic phenotypes', PLoS One **11**(2), e0148361.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/26863521*

Huang, Y.-T., VanderWeele, T. J. and Lin, X. (2014), 'Joint analysis of snp and gene expression data in genetic association studies of complex diseases', The annals of applied statistics **8**(1), 352.

Imai, K., Keele, L. and Tingley, D. (2010), 'A general approach to causal mediation analysis.', Psychological methods **15**(4), 309.

Imai, K. and Yamamoto, T. (2013), 'Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments', Political Analysis **21**(2), 141–171.

Issaq, H. J., Van, Q. N., Waybright, T. J., Muschik, G. M. and Veenstra, T. D. (2009), 'Analytical and statistical approaches to metabolomics research', journal of separation science **32**(13), 2183–2199.

Jiang, W. (2005), Bayesian variable selection for high dimensional generalized linear models, Technical report, Technical Report 05-02, Dept. Statistics, Northwestern Univ. Available at newton. stats. northwestern. edu/˜ jiang/tr/glmone2. tr. pdf.

Jiang, W. (2007), 'Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities', The Annals of Statistics **35**(4), 1487–1511.

Jin, S.-S. and Song, W.-J. (2017), 'Association between mdr1 c3435t polymorphism and colorectal cancer risk: A meta-analysis', Medicine (Baltimore) **96**(51), e9428.

Johnson, C. H., Ivanisevic, J., Benton, H. P. and Siuzdak, G. (2015), 'Bioinformatics: the next frontier of metabolomics', Anal Chem **87**(1), 147–56.

Johnson, V. E. and Rossell, D. (2012), 'Bayesian model selection in high-dimensional settings', Journal of the American Statistical Association **107**(498), 649–660.

Jones, D. P., Park, Y. and Ziegler, T. R. (2012), 'Nutritional metabolomics: progress in addressing complexity in diet and health', Annual review of nutrition **32**, 183.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016), 'Kegg as a reference resource for gene and protein annotation', Nucleic Acids Res **44**(D1), D457–62.

Kang, J., Reich, B. J. and Staicu, A.-M. (2018), 'Scalar-on-image regression via the soft-thresholded gaussian process', Biometrika **105**(1), 165–184.

Kaplon-Cieslicka, A., Postula, M., Rosiak, M., Peller, M., Kondracka, A., Serafin, A., Trzepla, E., Opolski, G. and Filipiak, K. J. (2014), 'Younger age, higher body mass index and lower adiponectin concentration predict higher serum thromboxane b2 level in aspirin-treated patients with type 2 diabetes: an observational study', Cardiovasc Diabetol **13**, 112.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/25123549*

Kessler, N., Neuweger, H., Bonte, A., Langenkämper, G., Niehaus, K., Nattkemper, T. W. and Goesmann, A. (2013), 'Meltdb 2.0-advances of the metabolomics software system', Bioinformatics **29**(19), 2452–9.

Kim, S., Pan, W. and Shen, X. (2013), 'Network-based penalized regression with application to genomic data', Biometrics **69**(3), 582–593.

Kimura, Y., Goto, Y.-I. and Kimura, H. (2009), 'Hydrogen sulfide increases glutathione production and suppresses oxidative stress in mitochondria', Antioxidants & Redox Signaling **12**(1), 1–13.
URL: *http://dx.doi.org/10.1089/ars.2008.2282*

Kind, T. and Fiehn, O. (2010), 'Advances in structure elucidation of small molecules using mass spectrometry', Bioanal Rev **2**(1-4), 23–60.

Kovats, S. (2015), 'Estrogen receptors regulate innate immune cells and signaling pathways', Cell Immunol **294**(2), 63–9.

Krausz, L. T., Fischer-Fodor, E., Major, Z. Z. and Fetica, B. (2012), 'Gitr-expressing regulatory t-cell subsets are increased in tumor-positive lymph nodes from advanced breast cancer patients as compared to tumor-negative lymph nodes', Int J Immunopathol Pharmacol **25**(1), 59–66.

Krzanowski, W. J. and Hand, D. J. (2009), ROC curves for continuous data, CRC Press.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. and Neumann, S. (2012), 'Camera: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets', Anal Chem **84**(1), 283–9.

Kuhn, M. (2008), 'Caret package', journal of Statistical Software **28**(5).

Kundu, S., Shin, M., Cheng, Y., Manyam, G., Mallick, B. K. and Baladandayuthapani, V. (2015), 'Bayesian variable selection with structure learning: Applications in integrative genomics', arXiv preprint arXiv:1508.02803 .

Lasko, T. A., Bhagwat, J. G., Zou, K. H. and Ohno-Machado, L. (2005), 'The use of receiver operating characteristic curves in biomedical informatics', Journal of biomedical informatics **38**(5), 404–415.

Le Rhun, E., Bertrand, N., Dumont, A., Tresch, E., Le Deley, M.-C., Mailliez, A., Preusser, M., Weller, M., Revillion, F. and Bonneterre, J. (2017), 'Identification of single nucleotide polymorphisms of the pi3k-akt-mtor pathway as a risk factor of central nervous system metastasis in metastatic breast cancer', Eur J Cancer **87**, 189–198.

Leu, Y.-W., Yan, P. S., Fan, M., Jin, V. X., Liu, J. C., Curran, E. M., Welshons, W. V., Wei, S. H., Davuluri, R. V., Plass, C., Nephew, K. P. and Huang, T. H.-M. (2004), 'Loss of estrogen receptor signaling triggers epigenetic silencing of downstream targets in breast cancer', Cancer Res **64**(22), 8184–92.

Li, C. and Li, H. (2008), 'Network-constrained regularization and variable selection for analysis of genomic data', Bioinformatics **24**(9), 1175–1182.

Li, C. and Li, H. (2010), 'Variable selection and regression analysis for graph-structured covariates with an application to genomics', The annals of applied statistics **4**(3), 1498.

Li, F. and Zhang, N. R. (2010), 'Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics', Journal of the American Statistical Association **105**(491), 1202–1214.
URL: *http://dx.doi.org/10.1198/jasa.2010.tm08177*

Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., Coan, J. A. et al. (2015), 'Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression', The Annals of Applied Statistics **9**(2), 687–713.

Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., Jones, D. P. and Pulendran, B. (2013a), 'Predicting network activity from high throughput metabolomics', PLoS Comput Biol **9**(7), e1003123.

Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., Jones, D. P. and Pulendran, B. (2013b), 'Predicting network activity from high throughput metabolomics', PLoS computational biology **9**(7), e1003123.

Li, Y.-X., Yu, Z.-W., Jiang, T., Shao, L.-W., Liu, Y., Li, N., Wu, Y.-F., Zheng, C., Wu, X.-Y., Zhang, M., Zheng, D.-F., Qi, X.-L., Ding, M., Zhang, J. and Chang, Q. (2018), 'Snca, a novel biomarker for group 4 medulloblastomas, can inhibit tumor invasion and induce apoptosis', Cancer Sci **109**(4), 1263–1275.

Liu, F., Chakraborty, S., Li, F., Liu, Y., Lozano, A. C. et al. (2014), 'Bayesian regularization via graph laplacian', Bayesian Analysis **9**(2), 449–474.

Liu, X., Chen, L., Ge, J., Yan, C., Huang, Z., Hu, J., Wen, C., Li, M., Huang, D., Qiu, Y., Hao, H., Yuan, R., Lei, J., Yu, X. and Shao, J. (2016), 'The ubiquitin-

like protein fat10 stabilizes eef1a1 expression to promote tumor proliferation in a complex manner', Cancer Res **76**(16), 4897–907.

Lopez, S. M., Agoulnik, A. I., Zhang, M., Peterson, L. E., Suarez, E., Gandarillas, G. A., Frolov, A., Li, R., Rajapakshe, K., Coarfa, C., Ittmann, M. M., Weigel, N. L. and Agoulnik, I. U. (2016), 'Nuclear receptor corepressor 1 expression and output declines with prostate cancer progression', Clin Cancer Res **22**(15), 3937–49.

Luo, C., Pan, W. and Shen, X. (2012), 'A two-step penalized regression method with networked predictors', Statistics in biosciences **4**(1), 27–46.

Ma, H. and Patti, M. E. (2014), 'Bile acids, obesity, and the metabolic syndrome', Best practice & research Clinical gastroenterology **28**(4), 573–583.

Marseglia, L., Manti, S., D'Angelo, G., Nicotera, A., Parisi, E., Di Rosa, G., Gitto, E. and Arrigo, T. (2015), 'Oxidative stress in obesity: a critical component in human diseases', Int J Mol Sci **16**(1), 378–400.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/25548896*

Martin-Lorenzo, M., Zubiri, I., Maroto, A. S., Gonzalez-Calero, L., Posada-Ayala, M., de la Cuesta, F., Mourino-Alvarez, L., Lopez-Almodovar, L. F., Calvo-Bonacho, E., Ruilope, L. M. et al. (2015), 'Klk1 and zg16b proteins and arginine–proline metabolism identified as novel targets to monitor atherosclerosis, acute coronary syndrome and recovery', Metabolomics **11**(5), 2.

Martinez, M. E., Heddens, D., Earnest, D. L., Bogert, C. L., Roe, D., Einspahr, J., Marshall, J. R. and Alberts, D. S. (1999), 'Physical activity, body mass index, and prostaglandin e2 levels in rectal mucosa', J Natl Cancer Inst **91**(11), 950–3.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/10359547*

Matthews, J. and Gustafsson, J.-A. (2006), 'Estrogen receptor and aryl hydrocarbon receptor signaling pathways', Nucl Recept Signal **4**, e016.

Moore, S. C., Matthews, C. E., Sampson, J. N., Stolzenberg-Solomon, R. Z., Zheng, W., Cai, Q., Tan, Y. T., Chow, W. H., Ji, B. T., Liu, D. K., Xiao, Q., Boca, S. M., Leitzmann, M. F., Yang, G., Xiang, Y. B., Sinha, R., Shu, X. O. and Cross, A. J. (2014), 'Human metabolic correlates of body mass index', Metabolomics **10**(2), 259–269.

URL: *http://www.ncbi.nlm.nih.gov/pubmed/25254000*

Morris, P. G., Zhou, X. K., Milne, G. L., Goldstein, D., Hawks, L. C., Dang, C. T., Modi, S., Fornier, M. N., Hudis, C. A. and Dannenberg, A. J. (2013), 'Increased levels of urinary pge-m, a biomarker of inflammation, occur in association with obesity, aging, and lung metastases in patients with breast cancer', Cancer Prev Res (Phila) **6**(5), 428–36.

URL: *http://www.ncbi.nlm.nih.gov/pubmed/23531446*

Nakajima, J. and West, M. (2013a), 'Bayesian analysis of latent threshold dynamic models', Journal of Business & Economic Statistics **31**(2), 151–164.

Nakajima, J. and West, M. (2013b), 'Bayesian dynamic factor models: Latent threshold approach', Journal of Financial Econometrics **11**, 116–153.

Nakajima, J., West, M. et al. (2017), 'Dynamics & sparsity in latent threshold factor models: A study in multivariate eeg signal processing', Brazilian Journal of Probability and Statistics **31**(4), 701–731.

Newgard, C. B., An, J., Bain, J. R., Muehlbauer, M. J., Stevens, R. D., Lien, L. F., Haqq, A. M., Shah, S. H., Arlotto, M., Slentz, C. A., Rochon, J., Gallup, D., Ilkayeva, O., Wenner, B. R., Yancy, W. S., J., Eisenson, H., Musante, G., Surwit, R. S., Millington, D. S., Butler, M. D. and Svetkey, L. P. (2009), 'A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans

and contributes to insulin resistance', Cell Metab **9**(4), 311–26.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/19356713*

Ni, Y., Stingo, F. C. and Baladandayuthapani, V. (2017), 'Bayesian graphical regression', Journal of the American Statistical Association (just-accepted).

Nicholson, J. K., Holmes, E. and Elliott, P. (2008), 'The metabolome-wide association study: a new look at human disease risk factors', J. Proteome Res **7**(9), 3637–3638.

Osborne, C. K., Shou, J., Massarweh, S. and Schiff, R. (2005), 'Crosstalk between estrogen receptor and growth factor receptor pathways as a cause for endocrine therapy resistance in breast cancer', Clin Cancer Res **11**(2 Pt 2), 865s–70s.

Pan, W., Xie, B. and Shen, X. (2010), 'Incorporating predictor network in penalized regression with application to microarray data', Biometrics **66**(2), 474–484.

Park, T. and Casella, G. (2008), 'The bayesian lasso', Journal of the American Statistical Association **103**(482), 681–686.

Patti, G. J., Yanes, O. and Siuzdak, G. (2012), 'Innovation: Metabolomics: the apogee of the omics trilogy', Nature reviews Molecular cell biology **13**(4), 263–269.

Pearl, J. (2001), Direct and indirect effects, in 'Proceedings of the seventeenth conference on uncertainty in artificial intelligence', Morgan Kaufmann Publishers Inc., pp. 411–420.

Peng, B., Zhu, D., Ander, B. P., Zhang, X., Xue, F., Sharp, F. R. and Yang, X. (2013), 'An integrative framework for bayesian variable selection with informative priors for identifying genes and pathways', PloS one **8**(7), e67672.

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

Polson, N. G. and Scott, J. G. (2012), 'Local shrinkage rules, lévy processes and regularized regression', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **74**(2), 287–311.

Ricciotti, E. and FitzGerald, G. A. (2011), 'Prostaglandins and inflammation', Arterioscler Thromb Vasc Biol **31**(5), 986–1000.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/21508345*

Roberts, G. O. and Rosenthal, J. S. (1998), 'Optimal scaling of discrete approximations to langevin diffusions', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **60**(1), 255–268.

Robins, J. M. and Greenland, S. (1992), 'Identifiability and exchangeability for direct and indirect effects', Epidemiology pp. 143–155.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N. et al. (2005), 'Towards a proteome-scale map of the human protein–protein interaction network', Nature **437**(7062), 1173.

Rubin, D. B. (1978), 'Bayesian inference for causal effects: The role of randomization', The Annals of statistics pp. 34–58.

Savva, S. C., Chadjigeorgiou, C., Hatzis, C., Kyriakakis, M., Tsimbinos, G., Tornaritis, M. and Kafatos, A. (2004), 'Association of adipose tissue arachidonic acid content with bmi and overweight status in children from cyprus and crete', Br J Nutr **91**(4), 643–9.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/15035692*

Schaer, D. A., Murphy, J. T. and Wolchok, J. D. (2012), 'Modulation of gitr for cancer immunotherapy', Curr Opin Immunol **24**(2), 217–24.

Schäfer, C. B., Morgan, B. R., Ye, A. X., Taylor, M. J. and Doesburg, S. M. (2014), 'Oscillations, networks, and their development: Meg connectivity changes with age', Human brain mapping **35**(10), 5249–5261.

Shi, R. and Kang, J. (2015), 'Thresholded multiscale gaussian processes with application to bayesian feature selection for massive neuroimaging data', arXiv preprint arXiv:1504.06074 .

Silva, R. R., Jourdan, F., Salvanha, D. M., Letisse, F., Jamin, E. L., Guidetti-Gonzalez, S., Labate, C. A. and Vêncio, R. Z. N. (2014), 'Probmetab: an r package for bayesian probabilistic annotation of lc-ms-based metabolomics', Bioinformatics **30**(9), 1336–7.

Sinaiko, A. R., Steinberger, J., Moran, A., Prineas, R. J., Vessby, B., Basu, S., Tracy, R. and Jacobs, D. R., J. (2005), 'Relation of body mass index and insulin resistance to cardiovascular risk factors, inflammatory factors, and oxidative stress during adolescence', Circulation **111**(15), 1985–91.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/15837953*

Sobel, M. E. (1982), 'Asymptotic confidence intervals for indirect effects in structural equation models', Sociological methodology **13**, 290–312.

Song, Q. and Liang, F. (2015), 'A split-and-merge bayesian variable selection approach for ultrahigh dimensional regression', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **77**(5), 947–972.

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).

Stubelius, A., Erlandsson, M. C., Islander, U. and Carlsten, H. (2014), 'Immunomodulation by the estrogen metabolite 2-methoxyestradiol', Clin Immunol **153**(1), 40–8.

Subbaramaiah, K., Morris, P. G., Zhou, X. K., Morrow, M., Du, B., Giri, D., Kopelovich, L., Hudis, C. A. and Dannenberg, A. J. (2012), 'Increased levels of cox-2 and prostaglandin e2 contribute to elevated aromatase expression in inflamed breast tissue of obese women', Cancer Discov **2**(4), 356–65.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/22576212*

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J. and Viant, M. R. (2007), 'Proposed minimum reporting standards for chemical analysis chemical analysis working group (cawg) metabolomics standards initiative (msi)', Metabolomics **3**(3), 211–221.

Székely, G. J. and Rizzo, M. L. (2013), 'The distance correlation t-test of independence in high dimension', Journal of Multivariate Analysis **117**, 193–213.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), 'Measuring and testing dependence by correlation of distances', The annals of statistics pp. 2769–2794.

Székely, G. J., Rizzo, M. L. et al. (2009), 'Brownian distance covariance', The annals of applied statistics **3**(4), 1236–1265.

Tchetgen, E. J. T. and Shpitser, I. (2012), 'Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis', Annals of Statistics **40**(3), 1816.

Tenenbaum, D. (2016), 'Keggrest: Client-side rest access to kegg', R package version **1**(1).

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

Tomkin, G. H. and Owens, D. (2016), 'Obesity diabetes and the role of bile acids in metabolism', Journal of translational internal medicine **4**(2), 73–80.

van der Laan, M. J. and Petersen, M. L. (2008), 'Direct effect models', The International Journal of Biostatistics **4**(1).

VanderWeele, T. J. (2016), 'Mediation analysis: a practitioner's guide', Annual review of public health **37**, 17–32.

VanderWeele, T. J. and Vansteelandt, S. (2009), 'Conceptual issues concerning mediation, interventions and composition', Statistics and its Interface **2**(4), 457–468.

VanderWeele, T. and Vansteelandt, S. (2014), 'Mediation analysis with multiple mediators', Epidemiologic methods **2**(1), 95–115.

Want, E. and Masson, P. (2011), Processing and analysis of gc/lc-ms-based metabolomics data, in 'Metabolic Profiling', Springer, pp. 277–298.

Whiteman, M., Gooding, K. M., Whatmore, J. L., Ball, C. I., Mawson, D., Skinner, K., Tooke, J. E. and Shore, A. C. (2010), 'Adiposity is a major determinant of plasma levels of the novel vasodilator hydrogen sulphide', Diabetologia **53**(8), 1722–6.

URL: *http://www.ncbi.nlm.nih.gov/pubmed/20414636*

Wolff, M., Kosyna, F. K., Dunst, J., Jelkmann, W. and Depping, R. (2017), 'Impact of hypoxia inducible factors on estrogen receptor expression in breast cancer cells', Arch Biochem Biophys **613**, 23–30.

Wolowczuk, I., Hennart, B., Leloire, A., Bessede, A., Soichot, M., Taront, S., Caiazzo, R., Raverdy, V., Pigeyre, M., Consortium, A., Guillemin, G. J., Allorge, D., Pattou, F., Froguel, P. and Poulain-Godefroy, O. (2012), 'Tryptophan metabolism activation by indoleamine 2,3-dioxygenase in adipose tissue of obese women: an

attempt to maintain immune homeostasis and vascular tone', Am J Physiol Regul Integr Comp Physiol **303**(2), R135–43.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/22592557*

Wu, G., Bazer, F. W., Davis, T. A., Kim, S. W., Li, P., Rhoads, J. M., Satterfield, M. C., Smith, S. B., Spencer, T. E. and Yin, Y. (2009), 'Arginine metabolism and nutrition in growth, health and disease', Amino acids **37**(1), 153–168.

Wu, S., Mao, L., Li, Y., Yin, Y., Yuan, W., Chen, Y., Ren, W., Lu, X., Li, Y., Chen, L., Chen, B., Xu, W., Tian, T., Lu, Y., Jiang, L., Zhuang, X., Chu, M. and Wu, J. (2018), 'Rage may act as a tumour suppressor to regulate lung cancer development', Gene **651**, 86–93.

Xia, J. and Wishart, D. S. (2010), 'Metpa: a web-based metabolomics tool for pathway analysis and visualization', Bioinformatics **26**(18), 2342–4.

Yang, R., Bai, Y., Qin, Z. and Yu, T. (2014), 'Egonet: identification of human disease ego-network modules', BMC genomics **15**(1), 314.

Yin, J., Zhang, Z., Zheng, H. and Xu, L. (2017), 'Irs-2 rs1805097 polymorphism is associated with the decreased risk of colorectal cancer', Oncotarget **8**(15), 25107–25114.

Yu, D., Kim, M., Xiao, G. and Hwang, T. H. (2013), 'Review of biological network data and its applications', Genomics & informatics **11**(4), 200–210.

Zhang, C.-H. (2010), 'Nearly unbiased variable selection under minimax concave penalty', The Annals of statistics pp. 894–942.

Zhang, Y., Jiang, C., Li, H., Lv, F., Li, X., Qian, X., Fu, L., Xu, B. and Guo, X. (2015), 'Elevated aurora b expression contributes to chemoresistance and poor prognosis in breast cancer', Int J Clin Exp Pathol **8**(1), 751–7.

Zhe, S., Naqvi, S. A., Yang, Y. and Qi, Y. (2013), 'Joint network and node selection for pathway-based genomic data analysis', Bioinformatics **29**(16), 1987–1996.

Zheng, W. and van der Laan, M. J. (2012), 'Targeted maximum likelihood estimation of natural direct effects', The international journal of biostatistics **8**(1), 1–40.

Zhou, B., Xiao, J. F., Tuli, L. and Ressom, H. W. (2012), 'Lc-ms-based metabolomics', Molecular BioSystems **8**(2), 470–481.

Zhou, H. and Zheng, T. (2013), 'Bayesian hierarchical graph-structured model for pathway analysis using gene expression data', Statistical applications in genetics and molecular biology **12**(3), 393–412.

Zhou, Q., Chipperfield, H., Melton, D. A. and Wong, W. H. (2007), 'A gene regulatory network in mouse embryonic stem cells', Proceedings of the National Academy of Sciences **104**(42), 16438–16443.

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.