Statistical Methods for Genetic Pathway-Based Data Analysis

Lulu Cheng

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Inyoung Kim, Committee Chair
Eric P. Smith
Leanna L. House
Yili Hong

October 28, 2013
Blacksburg, Virginia

KEYWORDS: Adaptive GLASSO; Gaussian Random Process; Gene Expression Data; GLASSO; Marginal Likelihood; Multi-Level Gaussian Graphical Model; Pathway-Based Analysis; Unknown Link Estimation; Zero Inflated Poisson.

Statistical Methods for Genetic Pathway Based Data Analysis

Lulu Cheng

(ABSTRACT)

The wide application of the genomic microarray technology triggers a tremendous need in the development of the high dimensional genetic data analysis. Many statistical methods for the microarray data analysis consider one gene at a time, but they may miss subtle changes at the single gene level. This limitation may be overcome by considering a set of genes simultaneously where the gene sets are derived from the prior biological knowledge and are called "pathways". We have made contributions on two specific research topics related to the high dimensional genetic pathway data. One is to propose a semiparametric model for identifying pathways related to the zero inflated clinical outcomes; the other is to propose a multilevel Gaussian graphical model for exploring both pathway and gene level network structures.

For the first problem, we develop a semiparametric model via a Bayesian hierarchical framework. We model the pathway effect nonparametrically into a zero inflated Poisson hierarchical regression model with unknown link function. The nonparametric pathway effect is estimated via the kernel machine and the unknown link function is estimated by transforming a mixture of beta cumulative density functions. Our approach provides flexible semiparametric settings to describe the complicated association between gene microarray expressions and the clinical outcomes. The Metropolis-within-Gibbs sampling algorithm and Bayes factor are used to make the statistical inferences. Our simulation results support that the semiparametric approach is more accurate and flexible than the zero inflated Poisson regression with the canonical link function, this is especially true when the number of genes is large. The usefulness of our approaches is demonstrated through its applications to a canine gene expression data set (Enerson et al., 2006). Our approaches can also be applied to other settings where a large number of highly correlated predictors are present.

Unlike the first problem, the second one is to take into account that pathways are not independent of each other because of shared genes and interactions among pathways. Multi-pathway analysis has been a challenging problem because of the complex dependence structure among pathways. By considering the dependency among pathways as well as genes within each pathway, we propose a multi-level Gaussian graphical model (MGGM): one level is for pathway network and the second one is for gene network. We develop a multilevel L1 penalized likelihood approach to achieve the sparseness on both levels. We also provide an iterative weighted graphical LASSO algorithm(Guo et al., 2011) for MGGM. Some asymptotic properties of the estimator are also illustrated. Our simulation results support the advantages of our approach; our method estimates the network more accurate on the pathway level, and sparser on the gene level. We also demonstrate usefulness of our approach using the canine genes-pathways data set.

To my Dad and Mom, who showed me the way,
To Jue, who accompanied and will accompany with me all through the way,
To Enzo, who made the way full of fun.

Acknowledgments

I would like to express my sincere gratitude and appreciation to my advisor Dr. Inyoung Kim for her support, guidance, encouragement and trust throughout my graduate studies. She has given me the oppotunity to work on challenging and interesting statistical modeling research projects, and guided me through reading, thinking, presenting and writing to grow as a statistician.

It is also my honor to be advised by a group of prominent and brilliant committee members, Dr. Eric P. Smith, Dr. Leanna L. House, and Dr. Yili Hong. All of whom I have my utmost respect for.

I would like to extend my deepest thanks to all group members, who offered me invaluable advice and support whenever I need a practice presentation, encounter a bottleneck in my research, or seek a general discussion on our research, work and life. I cherish the hardworking environment in which we could learn from each other.

I have also had the privilege to talk and discussion with a number of talented individual in Department of Statistics at Virginia Tech. Thank you all the professors for your inspiring courses and guidance. Thank you all my friends for your generous collaboration and discussion on statistical issues that widened my vision. I also would like to thank all staff of Department of Statistics for the selfless support to the students.

Lastly, I would like to thank my beloved family, for their great and never-ending support. None of this would be possible without the love of my family.

Contents

List of Figures

List of Tables

Chapter 1

Background and Outline

The modern science of genetics could trace back to Greogor Mendel's famous experiments on hybridization of the pea plants, published in 1866. He conceived the idea of the heredity unit, which he called it "factors" at that time and was later widely known as "genes". It has laid the foundation for the explosion of genetic research thereafter. People believed the variation of phenotypes of organisms were originated from the alternative forms of genotypes, and the inheritance principals derived from his work were extensively applied to study genetic diseases or phenomena in bacteria, plants, animals and human beings under the developed statistical framework of population genetics.

Later in the middle of last century, biologists found that DNA was the portion of chromosomes that carried that inheritance unit–"gene", and along with the discovery of the double helical structure of DNA in 1953, the genetic research made its transition to the molecular level, so called "molecular biology". It attempted to explain the phenomena of life starting from the macromolecular properties that generate them. Within the scope of molecular genetics, it was of main interests to sequence both nucleic acids (such as DNAs), and proteins, and learn the relationship between the two forms were of main

interests. This relationship was also related to the term "gene expression", which was defined as the process that converts information encoded in a gene into functional products of the cell, such as proteins. There are a tremendous amount of application through understanding sequencing of the DNA and the mechanism of gene expression, including genotyping of specific viruses to direct appropriate treatment; identification of oncogenes and mutations linked to different forms of cancer; the design of medication and more accurate prediction of their effects; advancement in forensic applied sciences; biofuels and other energy applications; agriculture, livestock breeding, bioprocessing; risk assessment; bioarcheology, anthropology, and evolution.

In the last decades of the 20th century, gene expression technologies evolved so rapidly that the regulation of gene expression could be controlled and manipulated through genetic engineering; the Human Genome Project was completed in 2003 to provide a large and growing library of organisms with already sequenced genomes; and the genomic microarray technology was developed to allow quick and inexpensive measurements of gene expression level for thousands of genes simultaneously(Dziuda, 2010). Our research is based on this high-throughput microarray gene expression data, where the measurement is the preprocessed data by standardization of gene expression level refers to the number of transcripts from DNA to RNA.

Many statistical methods for the microarray data analysis consider one gene at a time, and they may miss subtle changes at the single gene level. This limitation may be overcome by considering a set of genes simultaneously where the gene sets are derived from prior biological knowledge. This dissertation focus on two major research problems related to high dimensional genetic pathway-based data. In Chapter 2, we will introduce a flexible semiparametric approach for evaluating the pathway effects on the zero inflated clinic outcomes. And in Chapter 3, we will propose a multilevel Gaussian graphical

model for pathway and genes.

1.1 A Semiparametric Model for Pathway Analysis with Zero Inflated Outcomes

Some work has been done in the regression setting to study the effects of clinical covariates and expression levels of genes in a pathway for continuous and binary clinical outcomes but not zero inflated outcomes which often arise in practice. Hence, in Chapter 2, we propose a semiparametric regression approach for identifying pathways related to zero inflated clinical outcomes. Our approach is developed by using a Bayesian hierarchical framework. We model the nonparametric pathway effect into a zero inflated Poisson hierarchical regression model with unknown link function. The nonparametric pathway effect is estimated via the kernel machine technique, while the unknown link function is estimated by transforming a mixture of beta cumulative density functions. Our approach provides flexible semiparametric settings to describe the complicated association between genes microarray expressions and the clinical outcomes. The Metropolis-within-Gibbs sampling algorithm and Bayes factor are adopted to make the statistical inference. Our simulation results support that the semiparametric approach is more accurate and flexible than the zero inflated Poisson regression with the canonical link function, this is especially so when the number of genes is large. The usefulness of our approaches is demonstrated through its applications to the Canine data set from Enerson et al. (2006). Our approaches can also be applied to other settings where a large number of highly correlated predictors are present.

The chapter related to this first problem is organized as follows. In Section 2.1, the

background of pathway data analysis is introduced. In Section 2.2, two semiparametric regression model for zero inflated outcomes are proposed. The first model takes two independent nonparametric functions for the mixing proportion and the mean of Poisson regression, respectively, while the second one assumes the only one nonparametric function and is a special case of the first model. The Gaussian kernels are used for constructing the covariance matrix of Gaussian random process, which is used to estimate the nonparametric pathway effect. In Section 2.3, the Bayesian approach is described. The Metropolis-within-Gibbs sampling algorithm is explained in Section 2.3.1 and the Bayesian inference based on the Bayes factor is proposed in Section 2.3.2, respectively. Two approaches, one based on the Monte Carlo and the other one based on Laplace approximation combining with Monte Carlo, are discussed in this section, too. In Section 2.4, we report the simulation results to study the performance of our approach and compare it with a zero inflated Poisson regression with canonical link functions. We compare Bayes factors of the two approaches. In Section 2.5, we apply our approaches to our motivated example, pathway study of the dog lesion. We then report the significant gene pathways using Bayes factors, and visually check the superiority of our approach comparing with Poisson regression and canonical link model.

1.2 Multilevel Gaussian Graphical Model

Gaussian graphical models have become a popular tool to represent networks among variables such as genes. It uses the conditional correlations from the joint distribution to describe the dependencies between gene pairs, and equivalently it employs the precision matrix of the genes. Because of the sparsity nature of the gene networks and small sample sizes in almost all high dimensional genetic data, regularization approaches attracted

much attention in aim at obtaining the shrinkage estimates of the precision matrix. Methods such as LASSO (Friedman et al., 2008), adaptive LASSO(Zou, 2006; Zou and Li, 2008; Fan and Li, 2001) and SCAD(Fan and Li, 2001) were introduced to Gaussian graphical models in recent studies. However these existing methods have been focused on the Gaussian graphical model among genes, that is, they are only applicable to single Gaussian graphical model. It is known that pathways are not independent of each other because of shared genes and interactions among pathways. Developing multipathway analysis has been a challenging problem because of the complex dependence structure among pathways. Hence, in Section 3, by considering the dependency among pathways as well as genes within each pathway, we propose a multi-level Gaussian graphical model.

The chapter related to this second problem is organized as follows. In Section 3.1, we review recent studies of Gaussian graphical models with application to gene networks, including the challenges, issues we would like to address, and how current methods handle the sparsity estimation of the precision matrix. In Section 3.2, we describe our multilevel Gaussian graphical model which helps identify association among both genes and pathways. The penalized estimation approach for this model will be described in Section 3.3, while some asymptotic properties will be depicted in Section 3.4. In Section 3.5, we show our simulation settings for comparing methods, and report analysis for the canine data results in Section 3.6. And the Conclusion and discussion are covered in Section 3.7.

Chapter 2

Bayesian Semiparametric Regression for Pathway-Based Analysis with Zero Inflated Clinical Outcomes

2.1 Introduction

The microarray technique allows measuring the expression level of a large number of genes, and given an efficient statistical analysis, the bio-molecular information could become even more essential than the traditional clinical factors. Hence, high-throughput gene expression has become one of the most important tools for functional genomic studies. However, most of the methods involve single gene-based analysis. These methods analyze genes marginally and do not model the dependencies among them. One way to address the limitation of the single gene-based analysis is to analyze gene sets derived from prior biological knowledge to uncover patterns among the genes within a set.

A number of methods and programs have been developed to consider gene groupings based on Gene Ontology (GO) (Harris et al., 2004). These methods have been successful in detecting subtle changes in expression levels, which could be missed using single gene-based analysis (Mootha et al., 2004; Hosack et al., 2003; Rajagopalan and Agarwal, 2005). These sets of genes, called pathways, are predefined and ranking the ones relevant to a particular phenotype or clinical outcome can help researchers focus on a few sets of genes. The primary advantage of pathway-based analysis is that it can detect subtle changes in gene expression levels, which may not be possible with the single gene-based analysis.

A number of methods have been proposed to identify pathways relevant to a particular disease. Goeman et al. (2004) proposed a global test to identify pathways to distinguish between two binary groups. Their model was based on the generalized linear random effects model with a logit link. Gene Set Enrichment Analysis (GSEA), proposed by Subramanian et al. (2005), examined the overall strength of top signals in a given pathway. While the global test and GSEA mainly focused on the detection of differentially expressed pathways associated with binary outcomes, a Random Forests approach proposed by Pang et al. (2006) is applicable to both continuous and binary outcomes. Liu et al. (2007) proposed a semiparametric logistic regression model for pathway-based analysis. Stingo et al. (2011) proposed a Bayesian approach and incorporated genes in a pathway into a parametric mean function. All of these approaches specify the canonical link function such as the logit, log, or identity link functions. Furthermore, these existing approaches are not applicable when events are rare or when zero-inflated outcomes exist.

Our study is motivated by both this need for an approach suitable for zero-inflated outcomes, as well as a data set from Enerson et al. (2006). This data set contains microarray gene expression data measured in 14 dogs with lesions and 15 without. The goal of

this study is to identify pathways to distinguish between dogs with and without lesions in their liver. There was a total of 441 pathways and 6,592 genes. The Canine dataset was generated from investigative toxicology studies designed to identify the molecular pathogenesis of drug-induced vascular injury in coronary arteries of dogs treated with adenosine receipt eragonist CI-947. The Canine genes were mapped to human orthologs for pathway analysis. The human orthologs for dogs were generated by matching gene sequences using BLASTx (Enerson et al., 2006). For each dog, the number of lesions was counted, having a range between 0 and 21. Among the 29 dogs, more than half of the dogs had zero count outcomes, which would result in an overdispersion if the Poisson or Negative binomial regression was applied. Since this clinical outcome is zero-inflated, there are a limited number of methods which can be used to handle this situation given the fact that most existing methods focus on binary or continuous outcomes. This has motivated us to develop statistical methods under a regression setting to study the effects of clinical covariates and expression levels of genes in a pathway on zero-inflated clinical outcomes. However, since the relationship between zero-inflated clinical outcomes and expression levels of genes in a pathway are complex, it is not possible to describe it using a parametric model. Thus, we consider the development of a statistical method to cover a broad class of nonparametric settings for not only expression levels of genes in a pathway, but also the link function. We know that if the link function is specified as the canonical link and the parametric predictive function form is used for an unknown relationship, the model may easily turn out to be misspecified, and overdispersion problems may arise, influencing the statistical inferences (Liu et al., 2007; Hilbe, 2009, 2011). Therefore, in this chapter, we develop a nonparametric setting which can include the traditional model with the canonical link and linear predictive function as a special case, and also cover the noncanonical link and nonlinear model in general.

We propose a semiparametric regression model for evaluating pathway effects on zero-inflated clinical outcomes. In our model, the pathway effect is nonparametrically modeled into a zero-inflated Poisson regression (ZIP). We do not specify link functions for the mixing proportion or the mean of the Poisson regression. That is, the mixing proportion is modeled using a nonparametric function of gene expression without a logit link and the mean function of Poisson regression is also modeled using nonparametric function without a log link.

The nonparametric function of the gene expressions is estimated by connecting a kernel machine with the Gaussian random process to construct the dependencies among genes in a pathway. By using the Gaussian random process for pathway effect, we can obtain a similarity matrix of observations and also address issues of high dimensional space of multivariate covariates and interactions among them.

The unknown link function is estimated by transforming it to a unit interval so that we can estimate this transformation as a mixture of strictly increasing beta cumulative distribution functions using the method described by Mallick and Gelfand (1994). We give an illustration of the mixture of the cumulative beta distribution function to mode thel link function in Figure 2.1. The link functions used in practice, such as canonical link functions, can be obtained using the class of mixtures of the cumulative beta distribution functions. For example, the dashed curves represent five mixtures of cumulative beta distribution functions. The logit link, which is the solid curve with circles, is covered by the mixture of the five cumulative beta distribution functions with equal weights $(0.2, 0.2, 0.2, 0.2, 0.2)$. The probit link function, which is the dotted curve with triangles, is very close to one of the five beta cumulative distribution functions, Beta density $\mathrm{Beta}(3, 3)$ with parameters 3 and 3, which can be obtained by using mixing weights $(0, 0, 1, 0, 0)$. The 5th power of the logit link, which is represented by the dashed curve with plus symbols, is

covered by a beta cumulative distribution, based on a Beta density with parameters 5 and 1, denoted by $\mathrm{Beta}(5, 1)$, and can also be obtained by using mixing weights $(0, 0, 0, 0, 1)$. As we can see from Figure 1, by varying weights in the mixture of cumulative beta distribution functions, we can achieve a broader class of link functions than the conventional method. To obtain a desired monotone link function, we could change the number of mixtures and estimate the mixing weights.

Our approach is developed using a Bayesian hierarchical framework because it not only provides a intuitive way to estimate the complex statistical model but also easily incorporates prior knowledge. The Metropolis-within-Gibbs strategy is adopted for sampling from the posterior distribution. We make a Bayesian inference based on the Bayes factor to identify significant pathways. Since our approach does not require strong model specifications, it is more flexible and robust than other existing approaches.

This chapter is organized as follows. In Section 2.2, two semiparametric regression models for zero-inflated outcomes are proposed. The first model takes two independent nonparametric functions for the mixing proportion and the mean of Poisson regression, respectively, while the second one assumes only one nonparametric function and is a special case of the first model. The Gaussian kernels are used for constructing the covariance matrix of the Gaussian random process, which is used to estimate the nonparametric pathway effect. In Section 2.3, the Bayesian approach is described; specifically, the Metropolis-within-Gibbs sampling algorithm is explained in Section 2.3.4 and Bayesian inference based on the Bayes factor is described in Section 2.3.5. In Section 2.4, we report simulation results to study the performance of our approach by comparing it with a ZIP with canonical link functions and evaluating its testing of pathway effects in terms of Type I errors and powers. In Section 2.5, we apply our approach to our primary example, the pathway study of dog lesions. We then report the significant gene pathways using the

Figure 2.1: The dashed curves represent the five mixtures of the cumulative beta distribution functions. The logit link, which is the solid curve with circles, is covered by the mixture of the cumulative beta distribution functions with equal weights $(0.2, 0.2, 0.2, 0.2, 0.2)$. The probit link function, which is the dotted curve with triangles, is very close to one of the five beta cumulative distribution functions, Beta density $Beta(3, 3)$ with parameters 3 and 3, which can be obtained by using mixing weights $(0, 0, 1, 0, 0)$. The 5th power of the logit link, which is represented by the dashed curve with plus symbols, is covered by the beta cumulative distribution, Beta density $Beta(5, 1)$ with parameters 5 and 1, which can be obtained by using mixing weights $(0, 0, 0, 0, 1)$.

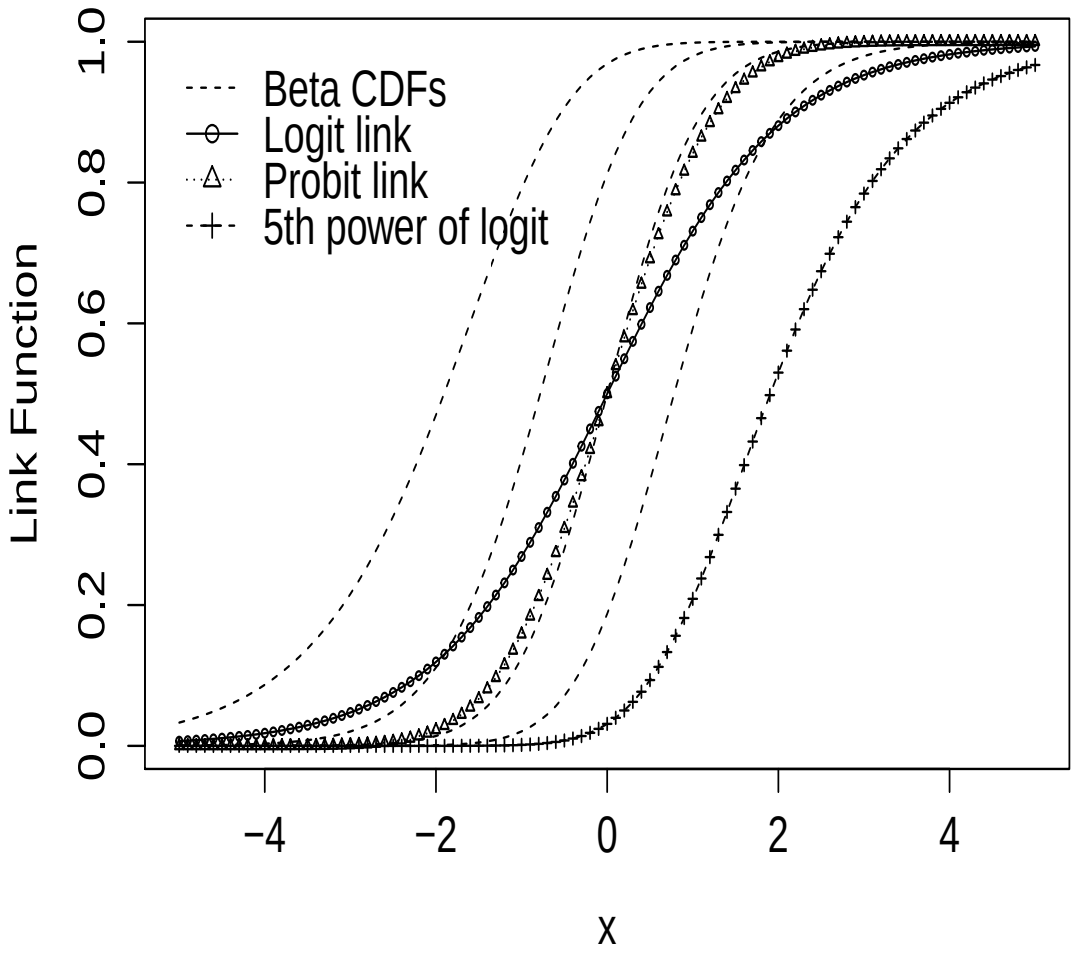


Illustration of Mixture of Cumulative Beta Distribution Function as Link Functions

Bayes factor.

2.2 Semiparametric regression for Zero-Inflated Outcomes

In this chapter, we propose two semiparametric regression models for zero-inflated outcomes. The first model involves two nonparametric functions, for the mean of the Poisson counts and the probability of dogs being healthy, respectively. The second model is a special case of the first model, in which the probability of dogs being healthy is a function of the mean number of dog lesions, and thus only one nonparametric function is needed.

2.2.1 Semiparametric model with two nonparametric functions

Let Y_i be a clinical outcome measured on the ith subject for $i = 1, 2, \ldots, n$. The outcome Y_i can be a nonnegative integer value $k \geq 0$. Let X_{ij} denote the gene expression value of the jth gene in a pathway for the ith subject, $j = 1, 2, \cdots, p$ (in gene expression studies, $p >> n$). Let $\mathbf{x}_i$ denote the ith subject's gene expression vector and $\mathbf{X}$ represent an $n \times p$ matrix with elements X_{ij}. Our semiparametric regression model for zero-inflated outcomes can be written as

$$P(Y_i = k|\mathbf{x}_i) = \begin{cases} \pi_i\{\gamma_1(\mathbf{x}_i)\} + [1 - \pi_i\{\gamma_1(\mathbf{x}_i)\}] \exp[-\lambda_i\{\gamma_2(\mathbf{x}_i)\}] & \text{if } k = 0; \\ [1 - \pi_i\{\gamma_1(\mathbf{x}_i)\}] \frac{[\lambda_i\{\gamma_2(\mathbf{x}_i)\}]^k \exp[-\lambda_i\{\gamma_2(\mathbf{x}_i)\}]}{k!} & \text{if } k > 0, \end{cases} \tag{2.1}$$

where $\gamma_1(\mathbf{x}_i)$ and $\gamma_2(\mathbf{x}_i)$ are unknown nonlinear smooth functions of $\mathbf{x}_i$, $\pi_i\{\gamma_1(\mathbf{x}_i)\}$ is the probability of being healthy, and $1 - \pi_i\{\gamma_1(\mathbf{x}_i)\}$ is the probability that the subject is unhealthy. With the above parameterization, if a subject is in the unhealthy state, it still has a chance to have an outcome of zero. Thus, Poisson regression is a special case of the

model when $\pi_i\{\gamma_1(\mathbf{x}_i)\} = 0$. The nonparametric mean function of the Poisson distribution is represented by $\lambda_i\{\gamma_2(\mathbf{x}_i)\}$. We model $\pi_i\{\gamma_1(\mathbf{x}_i)\}$ and $\lambda_i\{\gamma_2(\mathbf{x}_i)\}$ as $\pi_i(\mathbf{x}_i) = g_1^{-1}\{\gamma_1(\mathbf{x}_i)\}$ and $\lambda_i(\mathbf{x}_i) = g_2^{-1}\{\gamma_2(\mathbf{x}_i)\}$, where $g_1(\cdot)$ and $g_2(\cdot)$ are the unspecified link functions for $\pi_i(\cdot)$ and $\lambda_i(\cdot)$ respectively.

First we explain how to estimate the nonparametric functions, $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$. Because of the high dimensional space of multivariate covariates X and interaction among $\mathbf{x}$'s, we estimate $\gamma_1(\mathbf{x}_i)$ and $\gamma_2(\mathbf{x}_i)$ by connecting the kernel machine with the Gaussian random process, that is $\gamma_l(\cdot)$, $l = 1, 2$, follows the Gaussian Process with mean 0 and covariance $\text{cov}\{\gamma_l(\mathbf{x}_i), \gamma_l(\mathbf{x}_j)\} = \tau_l^{-1}K(\mathbf{x}_i, \mathbf{x}_j)$, where τ_l^{-1} is an unknown positive parameter and $K(\cdot, \cdot)$ is the kernel function. Therefore, we can rewrite $\boldsymbol{\gamma}_l(\mathbf{X}) = \{\gamma_l(\mathbf{x}_1), \gamma_l(\mathbf{x}_2), \cdots, \gamma_l(\mathbf{x}_n)\}^T \sim MN\{0, \tau_l^{-1}\mathbf{K}(X)\}$ and $\mathbf{K}(X)$ is an $n \times n$ matrix whose ijth element is $K(\mathbf{x}_i, \mathbf{x}_j)$. For the kernel function $K(\cdot, \cdot)$, we consider two kernels: one is the Gaussian kernel and the other is the independent kernel.

- Gaussian kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\rho})$, where ρ is a scale parameter; the "similarity" in this kernel is measured using the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Note that the smaller the distance, the larger similarity between the two observations. We fix ρ as $2p$.

- Independent kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = 1$ if $i = j$ and 0 otherwise; the "similarity" in this kernel is measured under the assumption that this set of genes are independent of each other.

By comparing these two kernels using Bayes factor, we will make Bayesian inference to identify significant pathways related to zero-inflated clinical outcomes. We further discuss this Bayesian inference based on the Bayes factor in Section 2.3.

In addition to the Gaussian kernel, other kernels can be used, such as linear, polyno-

mial, or spline kernels in practice. Since pathway gene expression data are high dimensional and have complicated interactions among genes in a pathway, we use the Gaussian kernel. This is because the Gaussian kernel can be seen as an infinite order of polynomial kernel with all orders of interactions. In addition, in situations where $p < n$, the variance-covariance matrix based on a polynomial kernel can not be positive definite and may lead to computational problems.

The following assumption is required for a Gaussian kernel K:

$$Pr\left(l_1 < \exp[-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\rho}] < 1 - l_2\right) = \eta > 0, \tag{2.2}$$

where $0 < l_1, l_2 < 0.5$ are the two values such that the off-diagonal entries of K are different from 0 or 1, and η is sufficiently large. The appropriate choice of ρ can make η large. When $\rho \to \infty$, K approaches the matrix of 1's, and when $\rho \to 0$, K approaches the identity matrix. Both cases cause an identifiability issue. We can estimate ρ by considering it as an unknown parameter, or we can fix it at a value such that η is optimized. It turns out that the optimal choice is $\rho \approx \rho^* E(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, $1 \leq \rho* \leq 2$ (Fang Z., 2012). In addition, the value of $E(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ depends on a pathway. The number of genes in a pathway varies. Hence, we assume that $E(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ also depends on the number of genes p. In addition, Liu et al. (2007) showed that the choice of ρ is not sensitive on their testing procedure. Hence we fix ρ as $2p$.

Next, we describe how to estimate the link functions, $g_1(\cdot)$ and $g_2(\cdot)$. In the case of generalized linear models, one usually specifies the known link function using a canonical link, such as the logit link $g_1(p_i) = \log\{p_i/(1-p_i)\} = \gamma_1(\mathbf{x}_i)$ or the log link $g_2(\lambda_i) = \log(\lambda_i) = \gamma_2(\mathbf{x}_i)$. However, in our study we do not explicitly specify these link functions.

The link function g is a mapping from the space of Ω, the mean of Y, onto R^1, that is,

$g^{-1}(\cdot)$ is a mapping function from R^1 onto Ω. Let us define a strictly increasing and differentiable function $T_1(\cdot)$ that maps from Ω to a unit interval (0,1) with $J(\cdot) = T_1\{g^{-1}(\cdot)\}$, where J is a strictly increasing and differentiable distribution function. So modeling the function $g(\cdot)$ is equivalent to modeling an unknown distribution function. Let $g_0(\cdot)$ be a baseline link function for $g(\cdot)$, perhaps the canonical link, and let $J_0(\cdot) = T\{g_0^{-1}(\cdot)\}$ be the cumulative distribution function associated with $g_0(\cdot)$. Since discrete mixtures of Beta densities provide a continuous dense class of models for densities on (0,1) (Diaconis and Ylvisaker, 1985), a general member of model can be formed as

$$h(u) = \sum_{j=1}^{\infty} w_j \mathrm{Beta}(u|c_j, d_j),$$

where ω_j are the weights for the mixands satisfying $\sum_{j=1}^{\infty} \omega_j = 1$, and $\mathrm{Beta}(u|c_j, d_j)$ denotes the Beta density with parameters c_j and d_j. In practice,

$$h(u) \approx \sum_{j=1}^{r} w_j \mathrm{Beta}(u|c_j, d_j),$$

where r, the finite number of mixands, is adopted.

By specifying $T_1(\cdot)$ as the identity function, $\pi_i\{\gamma_1(\mathbf{x}_i)\}$ can be expressed as follows:

$$\begin{aligned} T_1[\pi_i\{\gamma_1(\mathbf{x}_i)\}] &= \pi_i\{\gamma_1(\mathbf{x}_i)\} \\ &= T_1[g_1^{-1}\{\gamma_1(\mathbf{x}_i)\}] = J_1\{\gamma_1(\mathbf{x}_i)\} \\ &= \sum_{j=1}^{r} \omega_{1j} IB[J_{10}\{\gamma_1(\mathbf{x}_i)\}, c_{1j}, d_{1j}], \end{aligned}$$

where $IB[\cdot,\cdot,\cdot]$ denotes the incomplete Beta function with parameters c_{1j} and d_{1j}, and $J_{10}(\cdot) = T_1\{g_{10}^{-1}(\cdot)\}$; $g_{10}(\cdot)$ is the baseline link function and can be chosen as the canonical link. For example, in our study, the logit link function for π_i can be used as the baseline

link function. Thus, $J_{10}\{\gamma_1(\mathbf{x}_i)\} = \exp(\gamma_1(\mathbf{x}_i))/\{1+\exp(\gamma_1(\mathbf{x}_i))\}$. Mallick and Gelfand (1994) showed that in practice, choosing r to be 3 or 4 is robust enough. Given r, the number of mixands, it is easier to assume that the component Beta densities are specified but the weights (ω_{1j}) are unknown. The set of c_1 and d_1 can be specified to provide a collection of Beta densities that blanket the range (0,1). Hence, the specification of $g_1(\cdot)$ is equivalent to the specification of ω.

For the unspecified link function of $\lambda_i\{\gamma_2(\mathbf{x_i})\}$, we similarly define $T_2(\cdot)$ as follows:

$$\begin{aligned} T_2[\lambda_i\{\gamma_2(\mathbf{x}_i)\}] &= \frac{\lambda_i\{\gamma_2(\mathbf{x}_i)\}}{1+\lambda_i\{\gamma_2(\mathbf{x}_i)\}} \\ &= T_2[g_2^{-1}\{\gamma_1(\mathbf{x}_i)\}] = J_2\{\gamma_2(\mathbf{x}_i)\} \\ &= \sum_{j=1}^{r} \omega_{2j} IB[J_{20}\{\gamma_2(\mathbf{x}_i)\}, c_{2j}, d_{2j}] \end{aligned}$$

where $J_{20}(\cdot) = T_2\{g_{20}^{-1}(\cdot)\}$. The baseline link g_{20} was chosen to be the logit link function, so $J_{20}\{\gamma_2(\mathbf{x}_i)\} = \exp(\gamma_2(\mathbf{x}_i))/\{1+\exp(\gamma_2(\mathbf{x}_i))\}$.

Our model can be interpreted as follows: since we are interested in overall pathway effect instead of single gene effect, we model these effects through the Gaussian process. The similarity is measured using Euclidean distance. If the samples have "similar" expression profiles for genes in a pathway, this pathway has similar effects on their clinical outcomes. If two pathways are close in terms of the Euclidean distance of their gene expressions, they have a high chance of having similar predictive functions and consequently similar clinical outcomes. For the link function, we use a mixture of cumulative beta distribution functions. Once the weights are estimated, the monotone link function will be determined by the data. We can then estimate the risk and odds ratio for the unhealthy state and also estimate the average number of counts.

2.2.2 Semiparametric model with one nonparametric function

We also construct a simpler model which has fewer parameters than model (1). We first consider the fact that λ and π depend on a common nonparametric function. Our semiparametric model is then

$$P(Y_i = k|\mathbf{x}_i) = \begin{cases} \pi_i\{\gamma(\mathbf{x}_i)\} + [1 - \pi_i\{\gamma(\mathbf{x}_i)\}] \exp[-\lambda_i\{-C \cdot \gamma(\mathbf{x}_i)\}] & \text{if } k = 0; \\ [1 - \pi_i\{\gamma(\mathbf{x}_i)\}] \frac{[\lambda_i\{-C \cdot \gamma(\mathbf{x}_i)\}]^k \exp[-\lambda_i\{-C \cdot \gamma(\mathbf{x}_i)\}]}{k!} & \text{if } k > 0. \end{cases} \tag{2.3}$$

which has only one nonparametric function, while model (1) has two.

This model results in a decrease of $(n+1)$ parameters by combining γ_2 and τ_2 in model (1) into one parameter C. This C requires the assumption of a proportional relationship between the predictors for λ and π. In our study, we assume C to be positive. The predictor $\gamma(\mathbf{x_i})$ and unknown link function are estimated using the same approach as in model (1). As for C, one would estimate its value by using a grid search and finally choosing the C that satisfies the largest marginal likelihood. In our study, we estimate it using a Bayesian framework. We choose a Uniform distribution, $\mathrm{Unif}[c_1, c_2]$, for the prior distribution of C, where $c_1 = 0.001$ and $c_2 = 100$.

2.3 Bayesian Approach

This section describes the Bayesian approach to estimate parameters of our semiparametric models for zero-inflated outcomes. Metropolis-within-Gibbs' algorithm is explained in Section 2.3.1-2.3.4 and the Bayesian inference based on the Bayes factor is described in Section 2.3.5.

2.3.1 The prior and full conditional distributions for model (1)

We assume that $\gamma_l(\mathbf{X}) \sim MN\{0, \tau_l^{-1}K(X)\}$, $l = 1, 2$. To complete the model specification, we further assume independent prior distributions: $\tau_l \sim Gamma(\alpha_l, \beta_l)$ and $\omega_l \sim Dir(\delta_l \mathbf{1}_r)$. Here by adopting the suggestion from Mallick and Gelfand (1994), we use $r = 3$ mixands.

Under this model specification, the likelihood is

$$L(\gamma_1, \gamma_2, \omega_1, \omega_2 | X_i, y_i) = \prod_{i=1}^{n} (\pi_i\{\gamma_1(\mathbf{x}_i)\} \cdot \boldsymbol{1}_{y_i=0} + [1 - \pi_i\{\gamma_1(\mathbf{x}_i)\}] \cdot \frac{\exp[-\lambda_i\{\gamma_2(\mathbf{x}_i)\}][\lambda_i\{\gamma_2(\mathbf{x}_i)\}]^{y_i}}{y_i!})$$

where

$$\begin{aligned} \pi_i\{\gamma_1(\mathbf{x}_i)\} &= T_1^{-1}[\sum_{j=1}^{3} \omega_{1j} IB\{J_{10}(\gamma_{1i}), c_{1j}, d_{1j}\}] \\ &= \sum_{j=1}^{3} \omega_{1j} IB\{J_{10}(\gamma_{1i}), c_{1j}, d_{1j}\} \end{aligned}$$

and

$$\begin{aligned} \lambda_i\{\gamma_2(\mathbf{x}_i)\} &= T_2^{-1}[\sum_{j=1}^{3} \omega_{2j} IB\{J_{20}(\gamma_{2i}), c_{2j}, d_{2j}\}] \\ &= \frac{\sum_{j=1}^{3} \omega_{2j} IB\{J_{20}(\gamma_{2i}), c_{2j}, d_{2j}\}}{1 - \sum_{j=1}^{3} \omega_{2j} IB\{J_{20}(\gamma_{2i}), c_{2j}, d_{2j}\}}. \end{aligned}$$

The full conditional distributions are

$$\begin{aligned} [\gamma_l | rest] &\propto L(\gamma_1, \gamma_2, \omega_1, \omega_2 | X_i, y_i) \times MN\{0, \tau_l^{-1}K(X)\} \times Gamma(\alpha_l, \beta_l), \\ [\tau_l | rest] &\propto MN\{0, \tau_l^{-1}K(X)\} \times Gamma(\alpha_l, \beta_l), \\ &\sim Gamma\{\alpha_l + \frac{n}{2}, \beta_l + \frac{1}{2}\gamma_l' K(X)\gamma_l\}, \\ [\omega_l | rest] &\propto L(\gamma_1, \gamma_2, \omega_1, \omega_2 | X_i, y_i) \times Dir(\delta_l \mathbf{1}_3). \end{aligned}$$

where $rest$ represents all other parameters.

Since there are no closed forms for the full conditional distribution for γ_l and ω_l, we adopt Metropolis-within-Gibbs algorithm to sample from the posterior distribution. We run M Metropolis-within-Gibbs and run m Metropolis sub-trajectories within each Metropolis-within-Gibbs iteration. We use the latter run as the Gibbs update. Each Metropolis chain has an acceptance rate of around 30%. Our Metropolis-within-Gibbs algorithm is summarized in the following steps:

- Step 1: Initialize $[\gamma_l^0, \tau_l^0, \omega_l^0]$, $l = 1, 2$.
- Step 2: Update parameters at the t^{th} iteration.
  - Step 2.1: Update γ_1^t;
    (i) Let $\gamma_1^t = \gamma_1^{t-1}$;
    (ii) Sample γ_1^* from a multivariate normal proposal distribution with mean γ_1^t and covariance $\tau_0^{-1} I$, where the scale parameter τ_0 is fixed and previously tuned through prior trials, to provide an acceptance rate of around 30%. Update $\gamma_1^t = \gamma_1^*$ with probability $p = \min(1, \frac{[\gamma_1^*|rest^{t-1}]}{[\gamma_1^t|rest^{t-1}]})$, where $rest^{t-1}$ represents all other parameters obtained at the $(t-1)$ iteration;
    (iii) Repeat (ii) m times and use the last updated sample as γ_1^t; in our case, the total number of iterations $m = 100$ is sufficient;
  - Step 2.2: Update $\tau_1^t \sim [\tau_1|\gamma_1^t, rest^{t-1}] = \text{Gamma}\{\alpha_1 + \frac{n}{2}, \beta_1 + \frac{1}{2}(\gamma_1^t)' K(X)(\gamma_1^t)\}$
  - Step 2.3: Update ω_1^t;
    (i) Let $\omega_1^t = \omega_1^{t-1}$;
    (ii) Sample ω_1^* from a Dirichlet proposal distribution with mean ω_1^t, and the weight parameter δ_0 is adjusted through prior trials to obtain an acceptance

rate of around 30%. Update $\omega_1^t = \omega_1^*$ with probability $p = \min(1, \frac{[\omega_1^*|\gamma_1^t,\tau_1^t,rest^{t-1}]}{[\omega_1^t|\gamma_1^t,\tau_1^t,rest^{t-1}]})$;

(iii) Repeat (ii) m times and use the last updated sample as ω_1^t; the total number of iterations is $m = 100$;

– Step 2.4: Update γ_2^t with Metropolis sub-trajectories within Gibbs iteration, similar to γ_1^t;

– Step 2.5: Update $\tau_2^t \sim [\tau_2|\gamma_2^t, rest^{t-1}] = \text{Gamma}\{\alpha_2 + \frac{n}{2}, \beta_2 + \frac{1}{2}(\gamma_2^t)'K(X)(\gamma_2^t)\}$;

– Step 2.6: Update ω_2^t with Metropolis sub-trajectories within Gibbs iteration, similar to ω_1^t;

- Step 3: Increase t until convergence; in our case, the total number of iterations $M = 10,000$ is sufficient.

After obtaining samples from the joint posterior distribution of γ_l, ω_l, and τ_l, $l = 1, 2$ using Gibbs updates, we can estimate these distributions by finding the modes of the distributions of each marginal distribution.

2.3.2 The prior and full conditional distributions for model (2)

Under this model (2) specification, the likelihood is

$$L(\gamma, \omega_1, \omega_2, C|X_i, y_i) = \prod_{i=1}^{n}(\pi_i\{\gamma(\mathbf{x}_i)\} \cdot \mathbf{1}_{y_i=0} + [1 - \pi_i\{\gamma(\mathbf{x}_i)\}] \cdot \frac{\exp[-\lambda_i\{C\gamma(\mathbf{x}_i)\}][\lambda_i\{C\gamma(\mathbf{x}_i)\}]^{y_i}}{y_i!})$$

where

$$\begin{aligned}\pi_i\{\gamma(\mathbf{x}_i)\} &= T_1^{-1}[\sum_{j=1}^{3}\omega_{1j}IB\{J_{10}(\gamma_{1i}), c_{1j}, d_{1j}\}] \\ &= \sum_{j=1}^{3}\omega_{1j}IB\{J_{10}(\gamma_{1i}), c_{1j}, d_{1j}\}\end{aligned}$$

and

$$\begin{aligned}\lambda_i\{\gamma(\mathbf{x}_i)\} &= T_2^{-1}[\sum_{j=1}^{3}\omega_{2j}IB\{J_{20}(\gamma_{2i}), c_{2j}, d_{2j}\}] \\ &= \frac{\sum_{j=1}^{3}\omega_{2j}IB\{J_{20}(\gamma_{2i}), c_{2j}, d_{2j}\}}{1-\sum_{j=1}^{3}\omega_{2j}IB\{J_{20}(\gamma_{2i}), c_{2j}, d_{2j}\}}.\end{aligned}$$

The full conditional distributions are

$$\begin{aligned}[\gamma|rest] &\propto L(\gamma, \omega_1, \omega_2, C|X_i, y_i) \times MN\{0, \tau^{-1}K(X)\} \times Gamma(\alpha, \beta), \\ [\tau|rest] &\propto MN\{0, \tau^{-1}K(X)\} \times Gamma(\alpha, \beta), \\ &\sim Gamma\{\alpha + \frac{n}{2}, \beta + \frac{1}{2}\gamma' K(X)\gamma\}, \\ [\omega_l|rest] &\propto L(\gamma, \omega_1, \omega_2|X_i, y_i) \times Dir(\delta_l \mathbf{1}_3) \\ [C|rest] &\propto L(\gamma, \omega_1, \omega_2|X_i, y_i) \times Unif[c_1, c_2].\end{aligned}$$

where $rest$ represents all other parameters.

We perform similar procedures as described in Section 3.1, except that we do not have Steps 2.4-2.5. The sampling of C is based on the Metropolis algorithm. Hence the samples from the reduced model can be drawn using Metropolis-within-Gibbs's algorithm.

2.3.3 Prior specification

In this study, our priors are specified as $\alpha = 0.001$ and $\beta = 0.001$ so that $E(\tau) = 1$, $\delta_l = 2$, $c_1 = 0.001$, and $c_2 = 100$. These priors are very close to noninformative priors. However, more informative priors can be implemented as follows: first, in the Gaussian process, the specification of $K(X)$ is important. With the Gaussian variance-covariance function, the scale parameter ρ will control the linearity of the predictive function. The closer to zero

it is, the noisier and local the random function is; the closer to infinity it is, the smoother and more global the random function is. Hence we can consider a prior distribution $\rho \sim Unif[L, U]$ with a fixed positive small value L and large value U with the assumption $U \sim O\{E(||x - x'||^2)\}$. For example, $L = 10^{-5}$ and $U = 10 \times E(||x - x'||^2)$. One can also use the polynomial kernel to quantify the similarity through the inner-product if strong belief in the inner product distances can better describe the similarity. However, we note that a polynomial kernel might create numerical problems. Firstly, the covariance matrix may not be positive definite; secondly, if the prior of the parameter τ, which controls the pathway effect, has a very large value, the variance-covariance matrix in the random function will shrink and no pathway effect can be modeled. Since the mean of $\tau = \alpha/\beta$ can represent the strength of the pathway effect and the variance of $\tau = \alpha/\beta^2$ can indicate how strong the prior belief is, one can incorporate the prior knowledge of pathway effect strength by specifying α and β. For pathways identified by a previous study, hyperpriors α and β can be selected to have large mean values, e.g, 0.7 and variance as 2$\times$mean, while for pathways which were not identified, we choose a small mean value, e.g, 0.3, and variance as 2$\times$mean. Such informative prior elicitation was used by Laud and Ibrahim (1995); for the prior of the weights ω and the base link which controls the shape of the link function, one can specify a concentrated Dirichlet prior distribution for ω's to have canonical link functions. If any prior study shows canonical link functions are adequate, once can specify concentrated Dirichlet prior distributions for ω's.

2.3.4 Bayesian inference

In this chapter, we describe the Bayesian inference to test (a) whether our approach can identify significant pathways related to zero-inflated clinical outcomes, (b) whether the zero-inflated Poisson distribution is necessary, compared to Poisson regression, and (c)

whether the unknown link function is advantageous over the canonical link function. All of these tests can be considered as model selection problems and can be performed using the Bayes factor. For each test, we describe the two models below:

- For testing (a), M_0 is a model under H_0 and M_1 is a model under H_1, where H_0: $\boldsymbol{\gamma}(X)$ does not depend on X and H_1: $\boldsymbol{\gamma}(X)$ is an unknown nonlinear smooth function of X. This means that in our study, the Gaussian random process under the null hypothesis is H_0: $\{\boldsymbol{\gamma}(X)$ has an identity covariance matrix as a function of $X\}$, i.e., $\mathbf{K}(X) = I$ and $\gamma_l(\mathbf{X}) \sim MN\{0, \tau_l^{-1} I\}$. We then compare it to the model M_1 with Gaussian kernel covariance matrix, $\mathbf{K}(X) = \left(e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2p}}\right)$. This means that in our study, the model M_1 under the alternative hypothesis H_1 is a semiparametric regression model, where $\gamma_l(\mathbf{X}) \sim MN\{0, \tau_l^{-1} K(X)\}$ and $\mathbf{K}(X) = \left(e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2p}}\right)$.

- For testing (b), M_0 is a semiparametric regression model assuming Y is following a Poisson distribution with a nonparametric mean function of X, and M_1 is a semiparametric regression model assuming Y is following a Zero-inflated Poisson distribution with a mixing proportion, where the mean of the Poisson distribution depends on a nonparametric mean function of X; the difference between M_0 and M_1 depends on whether $\pi_l\{\gamma_l(X)\} = 0$ for all $l = 1, 2, \ldots, n$.

- For testing (c), M_0 is a semiparametric regression model obtained using the canonical link function and M_1 is a semiparametric regression model obtained using an unknown link function. In model M_0, $g_1[\pi_i\{\gamma_1(X)\}] = \log \frac{\pi_i\{\gamma_1(X)\}}{1-\pi_i\{\gamma_1(X)\}}$ and $g_2[\lambda_i\{\gamma_2(X)\}] = \log[\lambda_i\{\gamma_2(X)\}]$.

To estimate the Bayes factor, we first considered Chib and Jeliazkov's approach (Chib and Jeliazkov, 2001). However, we found that this approach was very slow because it required sequential sampling from the full conditional distributions at fixed parameter

values. In addition, the result was unstable because it was calculated on a set of chosen values of parameters. Hence we estimated the marginal likelihood using a Laplace approximation, which we describe in the following Section 2.3.5.

2.3.5 Marginal likelihood estimation

We note that the marginal likelihood for model M_1 is an integral of the likelihood over the prior space

$$\begin{aligned}
&P(D|M_1) = \int P(D|\boldsymbol{\theta}_1, M_1)P(\boldsymbol{\theta}_1|M_1)d\boldsymbol{\theta}_1 \\
&= \int \cdots \int L(\gamma_1, \gamma_2, \omega_1, \omega_2|X_i, y_i)f(\gamma_1|\tau_1)f(\tau_1)f(\gamma_2|\tau_2)f(\tau_2)f(\omega_1)f(\omega_2)d\gamma_1 d\tau_1 d\gamma_2 d\tau_2 d\omega_1 d\omega_2 \\
&= \int_{(\gamma_1,\gamma_2,\omega_1,\omega_2)} L(\gamma_1, \gamma_2, \omega_1, \omega_2|X_i, y_i)\Big\{\int_{\tau_1} f(\gamma_1|\tau_1)f(\tau_1)d\tau_1\Big\}\Big\{\int_{\tau_2} f(\gamma_2|\tau_2)f(\tau_2)d\tau_2\Big\}f(\omega_1)f(\omega_2)d\gamma_1 d\gamma_2 d\omega_1 d\omega_2 \\
&= \int_{(\omega_1,\omega_2)} \Big\{\int_{(\gamma_1,\gamma_2)} L(\gamma_1, \gamma_2, \omega_1, \omega_2|X_i, y_i)f(\gamma_1)f(\gamma_2)d\gamma_1 d\gamma_2\Big\}f(\omega_1)f(\omega_2)d\omega_1 d\omega_2,
\end{aligned}$$

where $P(D|M_a)$ is the marginal likelihood for model M_a, $P(D|\boldsymbol{\theta}_a, M_a)$ is the density function of data D under model M_a given the model-specific parameter vector $\boldsymbol{\theta}_a$, and $P(\boldsymbol{\theta}_a|M_a)$ is the prior density of $\boldsymbol{\theta}_a$, $a = 0, 1$.

To calculate this marginal likelihood, we first obtain

$$\int_{\tau_l} f(\gamma_l|\tau_l)f(\tau_l)d\tau_l = f_l(\gamma_l) \propto \left(\beta_l + \frac{\gamma_l' K(X)^{-1}\gamma_l}{2}\right)^{-(\alpha_l + \frac{n}{2})}$$

where $K(X)$ is a matrix with the Gaussian kernel functions and (α_l, β_l) are the shape and rate hyper-parameters in the Gamma prior distribution of τ_l, $l = 1, 2$.

We then sample ω_l from Dirichlet distributions. For the tth sample ω_l^t, we can obtain

the Laplace approximation for

$$\begin{aligned} P(D|M_1)_t &= \int \cdots \int L(\gamma_1, \gamma_2, \omega_1^t, \omega_2^t | X_i, y_i) f_1(\gamma_1) f_2(\gamma_2) d\gamma_1 d\gamma_2 \\ &\approx h(\hat{\gamma}_1, \hat{\gamma}_2) \cdot |\Sigma|^{\frac{1}{2}}, \end{aligned}$$

where $h(\gamma_1, \gamma_2) = L(\gamma_1, \gamma_2, \omega_1^t, \omega_2^t | X_i, y_i) f_1(\gamma_1) f_2(\gamma_2)$, Σ is the inverse Hessian matrix of $\ln h(\gamma_1, \gamma_2)$ and $(\hat{\gamma}_1, \hat{\gamma}_2)$ is the maxima of $\ln h(\gamma_1, \gamma_2)$ which can be found by Newton Raphson method (See Section A in Supplementary Materials). Therefore, we finally obtain $P(D|M_1) = \sum_{t=1}^{T} P(D|M_1)_t / T$.

2.4 Simulation Study

We conducted simulations to study the advantage of our semiparametric regression model with unknown link functions. We compare three models: (1) ZIP with unknown link function and Gaussian process (ZIPUN), (2) ZIP with canonical link function and Gaussian process (ZIPCAN), and (3) Poisson with unknown link function and Gaussian process (PUN).

2.4.1 Comparison with ZIP with canonical link function

We study whether the zero-inflated Poisson distribution (ZIP) is more applicable than Poisson regression. That is, if excess zeros exist, the ZIP model should provide a larger marginal likelihood. We also study whether our semiparametric regression with unknown link function is more useful than the canonical link, that is, if the true link function is distinctly different from the canonical link function, our semiparametric approach would give a larger marginal likelihood. Our simulation setting are explained in detail as

follows:

- Generate the matrix X from the following two cases.
  - Case 1: The X matrix is generated from the Beta distribution $Beta(0.5, 0.5)$. We use this Beta distribution because it is a U-shaped distribution. Therefore it is easier to cluster X into two categories: unhealthy subjects with low values of X and healthy subjects with high values of X. Thus we can generate zero-inflated Poisson count outcome with excessive zeros.

    The number of genes p is set to be 5, 30, and 50 to simulate a range of the largest possible number of genes. The number of observations n is set to be 30 and 100 to represent small and large sample sizes, respectively.
  - Case 2: The X matrix is generated using the *simulator* within the "boost" R package (Dettling, 2004); this simulator allows us to retain the mean and correlation structure from real data. Pathway 119 (*Ion Channels and Their Functional Role in Vascular Endothelium*) is chosen for this case. Like our data, the number of observations n is set to 29, and 5, 15 and 50 genes are considered and generated.
- Generate the Gaussian random process for π_i and λ_i, $\gamma_1 \sim MN\{0, K(X)/2\}$ and $\gamma_2 \sim MN\{0, K(X)\}$, where $K(X) = \left(e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2p}}\right)$ and p is the number of genes generated.
- Set the true link function for λ_i as the log link function, $\lambda_i = \exp(\gamma_{2i})$.
- Set the true link function for π_i as the logit link with the power of tuning parameter a, $\pi_i = [\exp(\gamma_{1i})/\{1 + \exp(\gamma_{1i})\}]^a$, where a takes the values 0.1, 0.2, 0.5, 1, 2, 5, 10. It indicates how different the link function is from the canonical link function. As a

Table 2.1: Unknown link vs canonical link: the counts shows how many times the marginal likelihood for the ZIP unknown link approach (ZIPUN) is larger than that for the ZIP with canonical link approach (ZIPCAN) among the 100 simulations, for different true links, with p genes and sample size n.

$ZIPUN > ZIPCAN$	$n = 30$			$n = 100$		
	$p = 5$	$p = 15$	$p = 50$	$p = 5$	$p = 15$	$p = 50$
$a = 0.1$	67	94	100	64	84	100
$a = 0.2$	66	96	81	67	84	100
$a = 0.5$	57	94	77	50	97	95
$a = 1$	37	34	78	34	48	100
$a = 2$	32	49	90	31	39	100
$a = 5$	41	55	85	59	49	100
$a = 10$	54	57	84	50	56	100

gets closer to 1, the link becomes closer to the canonical link. When a is less than 1, the higher the chance that the response is 0, i.e. we have more zeros. When a is much larger than 1, π_i generally tends to be very small so that the real process is closer to a Poisson distribution without extra zeros.

- Generate the zero-inflated Poisson response variable in the following way: $y = 0$ if $u_i < \pi_i$, where $u_i \sim \text{Unif}[0, 1]$; otherwise $y \sim \text{Poisson}(\lambda_i)$.

For each combination of gene sizes p, sample size n, and level of tuning parameter a, we simulated 100 data sets and calculated the marginal likelihoods for each candidate models. We then count how many times the marginal likelihood of the ZIPUN approach is larger than that of the ZIPCAN approach among the 100 runs. These results are summarized in Tables 2.1 and 2.2.

We also calculate the number of times that the marginal likelihood of ZIPUN is larger than that for PUN, which is provided in Table 2.3. In Table 2.4, we summarize the median value of the marginal likelihoods for (i) ZIP with unknown link function (ZIPUN), (ii) ZIP

Table 2.2: Unknown link vs canonical link for a simulation withthe same data structure as pathway 119 generated from the "boost" package: the counts shows how many times the marginal likelihood for ZIP unknown link approach (ZIPUN) is larger than that for the ZIP with canonical link approach (ZIPCAN) among the 100 simulations, for different true links, p genes, and sample size $n = 29$, the same as our case study.

$ZIPUN > ZIPCAN$	$n = 29$		
	$p = 5$	$p = 15$	$p = 50$
$a = 0.1$	72	88	100
$a = 1$	22	46	35
$a = 10$	65	92	100

Table 2.3: ZIP vs Poisson: the counts shows how many times the marginal likelihood for ZIP unknown link approach (ZIPUN) is larger than that for the Poisson regression with unknown link approach (PUN) among the 100 runs, for different true links, p genes, and sample size n.

$ZIPUN > PUN$	$n = 30$			$n = 100$		
	$p = 5$	$p = 15$	$p = 50$	$p = 5$	$p = 15$	$p = 50$
$a = 0.1$	100	100	100	97	100	100
$a = 0.2$	90	100	100	92	98	100
$a = 0.5$	95	98	100	92	100	100
$a = 1$	76	100	100	96	100	100
$a = 2$	62	71	100	90	99	100
$a = 5$	56	70	99	73	94	100
$a = 10$	53	59	100	70	80	100

with canonical link (ZIPCAN), and (ii) Poisson regression with unknown link (PUN).

The simulation results suggest that when unknown the link function is quite different from the canonical link, the marginal likelihood of the former is likely to be larger than that of the latter, especially when the dimension of X is large (either p is large or both n and p are large). When a is less than 0, the probability of having a zero response is high, so the ZIP model indeed produces a larger marginal likelihood. In the case where a is large and the excess zeros are not as substantial, we compare the ZIPUN model with the PUN

Table 2.4: The median of the marginal likelihood obtained from the Laplace approximation for ZIP unknown link model, the ZIP with canonical link model, and Poisson regression model with unknown link, for each cell, with a combination of a true link, p genes and a sample size n.

Marginal likelihood	$n = 30$			$n = 100$		
	$p = 5$	$p = 15$	$p = 50$	$p = 5$	$p = 15$	$p = 50$
$a = 0.1$	0.0011	1.25e-6	2.76e-9	4.79e-13	1.35e-16	1.42e-8
	0.0008	1.01e-6	5.00e-10	3.56e-13	2.89e-17	3.15e-15
	0.0006	5.34e-8	3.05e-12	3.05e-12	1.35e-16	1.42e-8
$a = 0.2$	7.61e-14	1.98e-10	0.0004	1.89e-17	6.40e-10	1.38e-23
	9.62e-14	1.65e-10	4.93e-6	1.22e-17	2.37e-10	1.14e-25
	1.10e-15	1.08e-11	8.06e-6	5.55e-20	9.28e-12	3.87e-26
$a = 0.5$	2.66e-16	6.23e-19	4.49e-14	2.64e-11	3.74e-33	2.87e-34
	1.74e-16	8.82e-19	1.99e-14	3.10e-11	2.61e-33	7.65e-35
	1.64e-22	2.04e-18	8.46e-18	3.32e-13	7.93e-36	4.55e-38
$a = 1$	3.56e-12	4.07e-9	6.91e-18	1.59e-72	1.19e-50	6.67e-48
	3.89e-12	4.00e-9	7.78e-18	8.54e-73	4.18e-50	2.89e-48
	6.12e-12	7.89e-10	2.60e-17	4.79e-74	1.30e-56	9.18e-56

model. As we would expect, Poisson regression gives larger marginal likelihoods more often than ZIPUN. In fact, the values of the marginal likelihoods were actually very close to the Poisson model. When a is 1, where the true model has a canonical link function, the marginal likelihood of ZIPUN is close to the one obtained from the canonical link.

Overall, our simulation results suggest that when the true model is quite different from the classical specified form, which is often the case in practice, the zero-inflated nonparametric approach performs better. When the true model is close to the specified form, the nonparametric approach produces results similar to the classical models. Therefore we conclude that the ZIP model with unknown link approach is more flexible and gives better results in realistic situations.

2.4.2 Type I error using BF

We also conduct simulations to estimate type I error using our Bayesian approach. We estimate error by using the Bayes factor in favor of the alternative hypothesis H_1: $\{\boldsymbol{\gamma}(X)$ is an unknown nonlinear smooth function of X function$\}$, versus the null hypothesis H_0: $\{\boldsymbol{\gamma}(X)$ does not depend on $X\}$. Large values of BF are in favor of H_1; this means that the data indicates that H_1 is more strongly supported by the data than H_0. By following Jeffreys (1961) suggestion, we interpret the value of BF as not favoring if $BF \leq 1$, weakly favoring if $1 < BF \leq 3$, positively favoring if $3 < BF \leq 10$, and strong favor if $BF > 10$.

For the assessment of type I error, we consider the following cases 3-5 and simulate 1000 data sets with sample size $n = 30$, similar to our data, and three settings of the number of genes, $p = (5, 30, 100)$: $p = 5$ represents a relatively small number of genes compared to the sample size, $p = 30$ represents a comparable number of genes compared to the sample size, and $p = 100$ represents a relatively large number of genes compared to the sample size.

- Case 3: the probability of being healthy is $\pi = (0.1, 0.25, 0.5, 0.75, 0.9)$, and the mean of the number of lesions is $\lambda = (1, 5, 10)$; that is, both are independent of X. For each combination of π and λ, the type I error based on the relative frequency of the Bayes factor that falls above the cutoff values ($Pr(BF > BF_{cut})$) is shown in Table 2.5, where the three cutoff values of BF (BF_{cut}) used in this stimulation are 1, 3, and 10, which represent weak, positive and strong favors of H_1, respectively.

- Case 4: the same setting as case 3 except for that $\lambda = \exp(\gamma)$ where $\gamma \sim MN(0, I)$ and $\pi = (0.1, 0.25, 0.5, 0.75, 0.9)$; that is, both are independent of X. The type I error based on the relative frequency of the Bayes factor that falls above the cutoff

Table 2.5: Type I error for case 3: the mean of the number of lesions is $\lambda = (1, 5, 10)$, and the probability of being healthy is $\pi = (0.1, 0.25, 0.5, 0.75, 0.9)$, that is, both are independent of X. For each combination of π and λ, the type I error based on the relative frequency of the Bayes factor that falls above the cutoff, $Pr(BF > BF_{cut})$, is shown, where the three cutoff values of BF (BF_{cut}) used in this stimulation are 1, 3, and 10 which represent weak, positive and strong evidence favoring H_1, respectively.

			$p = 5$		$p = 30$			$p = 100$		
		BF_{cut} =1	3	10	1	3	10	1	3	10
$\lambda = 1$	$\pi = 0.1$	0.011	0.001	0	0.002	0.002	0	0	0	0
	$\pi = 0.25$	0.014	0.002	0	0	0	0	0	0	0
	$\pi = 0.5$	0.022	0	0	0.004	0	0	0	0	0
	$\pi = 0.75$	0.096	0	0	0.038	0	0	0.028	0	0
	$\pi = 0.9$	0.254	0	0	0.262	0	0	0.205	0	0
$\lambda = 5$	$\pi = 0.1$	0	0	0	0	0	0	0	0	0
	$\pi = 0.25$	0	0	0	0	0	0	0	0	0
	$\pi = 0.5$	0	0	0	0	0	0	0	0	0
	$\pi = 0.75$	0.026	0.008	0.002	0.014	0.006	0.002	0.012	0	0
	$\pi = 0.9$	0.159	0.004	0	0.117	0.008	0.002	0.122	0.006	0
$\lambda = 10$	$\pi = 0.1$	0	0	0	0	0	0	0	0	0
	$\pi = 0.25$	0	0	0	0	0	0	0	0	0
	$\pi = 0.5$	0	0	0	0	0	0	0	0	0
	$\pi = 0.75$	0.008	0.004	0.002	0.006	0	0	0.004	0.002	0
	$\pi = 0.9$	0.164	0.04	0.011	0.102	0.016	0.004	0.097	0.008	0

Table 2.6: Type I error for case 4: the mean of the number of lesions is $\lambda = \exp(\gamma)$, where $\gamma \sim MN(0, I)$, and $\pi = (0.1, 0.25, 0.5, 0.75, 0.9)$, that is, both are independent of X. The type I error based on the relative frequency of the Bayes factor that falls above the cutoff, $Pr(BF > BF_{cut})$, is shown. The three cutoff values of BF (BF_{cut}) used in this stimulation are 1, 3, and 10 which represent weak, positive and strong evidence favoring H_1, respectively.

	$p = 5$			$p = 30$			$p = 100$		
	BF_{cut} =1	3	10	1	3	10	1	3	10
$\pi = 0.1$	0.048	0.034	0.02	0.04	0.034	0.024	0.024	0.018	0.016
$\pi = 0.25$	0.072	0.052	0.034	0.03	0.022	0.01	0.046	0.032	0.024
$\pi = 0.5$	0.116	0.072	0.04	0.084	0.042	0.026	0.086	0.042	0.028
$\pi = 0.75$	0.172	0.042	0.024	0.15	0.042	0.02	0.136	0.048	0.022
$\pi = 0.9$	0.328	0.025	0.016	0.289	0.02	0.013	0.285	0.03	0.009

is shown in Table 2.6.

For case 3, the estimated type I errors in Table 2.5 are between 0 and 0.262 based on 1000 simulations. They vary depending on p, λ, π and BF_{cut}. When BF_{cut} is 10, the proportions of strong favor to H_1 are almost zero, which makes sense because the simulated data is independent of X. When BF_{cut} is 3, the proportions of positive favor to H_1 are around 0.05. When BF_{cut} is 1, the proportions of weak favor to H_1 are greater than 0.1. Note that the type I error increases as π increases, so when $\pi < 0.9$, the estimated type I error is within a reasonable range. However, when $\pi = 0.9$ and $BF_{cut} = 1$, the proportions of weak favor to H_1 are greater than 0.05. The results are similar for all λ. When BF_{cut} becomes larger, the estimated type I errors are mostly zero except for when $\lambda = 5$ and $\lambda = 10$. The estimated type I errors get smaller as p increases, although they are comparable for all p.

For case 4, the estimated type I errors in Table 2.6 are between 0.01 and 0.33. As in case 3, they also vary depending on p, π and the cutoff value of the Bayes factor (BF_{cut}). The estimated type I error is a little larger in case 4 than in case 3. When $\pi = 0.9$ and

$BF_{cut} = 1$, we obtain the largest type I error, 0.328. This is because there are too many zeros (90% of outcomes are zero). Overall, the results of case 4 are similar to those of case 3.

2.4.3 Power for testing pathway effect using BF

For the assessment of power, we consider the following case 5 and simulate 1000 data sets with $p = (5, 10)$, and $n = (30, 60)$.

- Case 5: This setting is the same as case 3 except for the mean of the number of counts λ is a nonlinear function of the X's. That is, $\lambda = \exp\{\gamma(X)\}$.
  - When $p = 5$, all X are generated from N(0,1) and $\gamma(X) = \cos(X_1) - 1.5X_2^2 + \exp(-X_3)X_4 - 0.8\sin(X_5)\cos(X_3) + 2X_1X_5$;
  - When $p = 10$, all X are generated from N(0,1) and $\gamma(X) = \cos(X_1) - 1.5X_2^2 + \exp(-X_3)X_4 - 0.8\sin(X_5)\cos(X_3) + 2X_1X_5 + 0.9X_6\sin(X_7) - 0.8\cos(X_6)X_7 + 2X_8\sin(X_9)\sin(X_{10}) - 1.5X_8^3 - X_8X_9 - 0.1\exp(X_{10})\cos(X_{10})$.

  A similar function form is used by Liu et al. (2007). The power based on the relative frequency of the Bayes factor falls above the cutoff $Pr(BF > BF_{cut})$ is shown in Table 2.7. The two cutoff values of BF (BF_{cut}) used in this stimulation are 1 and 3 which represent weak and positive favors of H_1. Note that we did not consider $BF_{cut} = 10$ because the estimated Type I error is too small.

The estimated powers in Table 2.7 vary depending on n, p, λ, π and the cutoff value of the Bayes factor, BF_{cut}. The estimated powers become smaller as BF_{cut} increases. The estimated powers are similar regardless of the sample size and the number of genes. Overall, our approach provides reasonable power when BF_{cut} is 1 or 3 and $\pi < 0.75$.

Table 2.7: Power for case 5: the mean of the count λ is a nonlinear function of X's. That is, $\lambda = \exp\{\gamma(X)\}$; when $p = 5$, all X are generated from N(0,1) and $\gamma(X) = \cos(X_1) - 1.5X_2^2 + \exp(-X_3)X_4 - 0.8\sin(X_5)\cos(X_3) + 2X_1X_5$, and when $p = 10$, all X are generated from N(0,1) and $\gamma(X) = \cos(X_1) - 1.5X_2^2 + \exp(-X_3)X_4 - 0.8\sin(X_5)\cos(X_3) + 2X_1X_5 + 0.9X_6\sin(X_7) - 0.8\cos(X_6)X_7 + 2X_8\sin(X_9)\sin(X_{10}) - 1.5X_8^3 - X_8X_9 - 0.1\exp(X_{10})\cos(X_{10})$. The power based on the relative frequency of the Bayes factor falls above the cutoff, $Pr(BF > BF_{cut})$, is shown. The two cutoff values of BF (BF_{cut}) used in this stimulation are 1 and 3, which represent weak and positive evidence favoring H_1, respectively.

			$p = 5$		$p = 10$
		BF_{cut} =1	3	1	3
$n = 30$	$\pi = 0.1$	0.768	0.59	0.728	0.66
	$\pi = 0.25$	0.859	0.718	0.831	0.817
	$\pi = 0.5$	0.91	0.846	0.831	0.831
	$\pi = 0.75$	0.833	0.695	0.771	0.634
$n = 60$	$\pi = 0.1$	0.829	0.685	0.864	0.849
	$\pi = 0.25$	0.699	0.563	0.825	0.537
	$\pi = 0.5$	0.72	0.587	0.78	0.699
	$\pi = 0.75$	0.814	0.694	0.811	0.786

2.4.4 Comparison with Zero-inflated negative binomial

We further conduct simulations to compare the performances of two models: (1) a zero-inflated Poisson regression with canonical link function and Gaussian process (ZIPCAN) and (2) a zero-inflated negative binomial regression with Gaussian process (ZINB). Under case 6, we estimate the power of testing for pathway effects and assess power in a simulation of 1000 data sets with $p = (5, 10)$, and $n = (30, 60, 100)$.

- Case 6: This setting is the same as case 5. We fit both ZIPCAN and ZINB models and compare their power.

For case 6, the estimated powers of ZIPCAN and ZINB are similar; they both vary depending on n, p, π and the cutoff value of the Bayes factor (BF_{cut}). The estimated powers decrease as BF_{cut} and π increase; that is, power becomes smaller as the number of zeros increases. The estimated powers are comparable regardless of the sample size and the number of genes. Overall, our approach provides reasonable power when BF_{cut} is 1 or 3 and $\pi < 0.75$; however, our approach is not applicable when $\pi > 0.75$.

2.5 Real Data Analysis

We applied our Bayesian semiparametric approach to the Canine data set from Enerson et al. (2006), described in Section 2.1. This data set is a microarray expression data set measured in 14 dogs with lesions and 15 without. There were a total of 441 pathways and 6,592 genes.

For our analysis, let Y be the number of lesions in the dogs and X be the $n \times p$ gene expression levels within each pathway, where the sample size n is 29 and the number of

Table 2.8: Power for case 6: the mean of the count λ is a nonlinear function of the X's. That is, $\lambda = \exp\{\gamma(X)\}$; when $p = 5$, all X are generated from N(0,1) and $\gamma(X) = \cos(X_1) - 1.5X_2^2 + \exp(-X_3)X_4 - 0.8\sin(X_5)\cos(X_3) + 2X_1X_5$, and when $p = 10$, all X are generated from N(0,1) and $\gamma(X) = \cos(X_1) - 1.5X_2^2 + \exp(-X_3)X_4 - 0.8\sin(X_5)\cos(X_3) + 2X_1X_5 + 0.9X_6\sin(X_7) - 0.8\cos(X_6)X_7 + 2X_8\sin(X_9)\sin(X_{10}) - 1.5X_8^3 - X_8X_9 - 0.1\exp(X_{10})\cos(X_{10})$. The outcome is generated from ZIP. The power based on the relative frequency of the Bayes factor that falls above the cutoff, $Pr(BF > BF_{cut})$, is shown. The two cutoff value of BF (BF_{cut}) used in this stimulation are 1 and 3 which represent weak and positive evidence favoring H_1, respectively.

Power				$p = 5$		$p = 10$
			BF_{cut} =1	3	1	3
$\pi = 0.1$	$n = 30$	ZIPCAN	0.843	0.699	0.695	0.601
		ZINB	0.818	0.636	0.889	0.566
	$n = 60$	ZIPCAN	0.811	0.744	0.795	0.768
		ZINB	0.814	0.711	0.95	0.91
	$n = 100$	ZIPCAN	0.927	0.854	0.816	0.816
		ZINB	0.925	0.786	0.969	0.938
$\pi = 0.25$	$n = 30$	ZIPCAN	0.74	0.521	0.761	0.642
		ZINB	0.707	0.314	0.87	0.48
	$n = 60$	ZIPCAN	0.806	0.709	0.754	0.729
		ZINB	0.792	0.529	0.93	0.9
	$n = 100$	ZIPCAN	0.8	0.7	0.891	0.86
		ZINB	0.794	0.479	0.98	0.97
$\pi = 0.5$	$n = 30$	ZIPCAN	0.686	0.372	0.75	0.5
		ZINB	0.698	0.219	0.778	0.294
	$n = 60$	ZIPCAN	0.817	0.659	0.766	0.68
		ZINB	0.748	0.466	0.9	0.6
	$n = 100$	ZIPCAN	0.834	0.737	0.882	0.869
		ZINB	0.848	0.511	0.94	0.89
$\pi = 0.75$	$n = 30$	ZIPCAN	0.508	0.344	0.407	0.116
		ZINB	0.643	0.256	0.37	0.10
	$n = 60$	ZIPCAN	0.694	0.414	0.691	0.441
		ZINB	0.643	0.307	0.77	0.31
	$n = 100$	ZIPCAN	0.79	0.619	0.849	0.745
		ZINB	0.756	0.427	0.9	0.62

genes p varies from 1 to 172 across the pathway considered. A total of 15 out of 29 dogs have zero count outcomes. Our goal is to identify pathways having a strong effect on the number of lesions in a dog. To identify significant pathways, we use our Bayesian inference based on the Bayes factor. Our approach took approximately 1-3 hours to run MH-within-Gibbs sampling for each pathway, depending on the size of the gene. Our code is written in R and is available upon request.

We first estimated the Bayes factor in favor of the semiparametric model with (1) two nonparametric functions versus (2) one nonparametric function. Since the Bayes factor was less than 5, we reported our analysis based on the semiparametric model (2) with one nonparametric function.

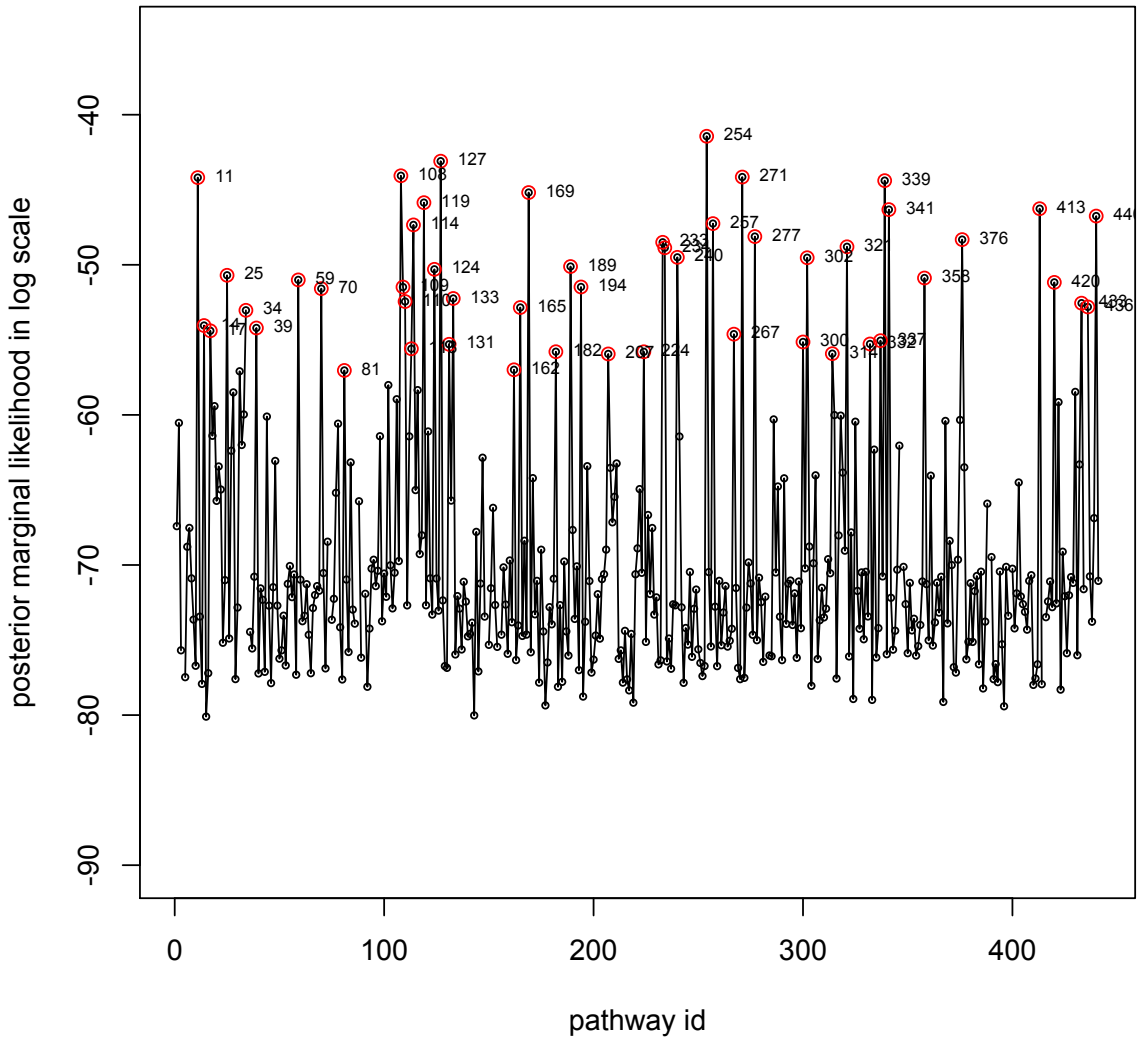
By comparing the Gaussian kernel with the independent kernel, we identify the top 50 pathways with a large Bayes factor, listed in Table 2.9. Marginal likelihood values for all 441 pathways are summarized in Figure 2.2.

We also compared the top 50 pathways identified by the global test (Goeman et al., 2004), Gene Set Enrichment Analysis (Subramanian et al., 2005), and the Random Forest approach (Pang et al., 2006). Since the global test and RF are applicable to either binary or continuous outcomes, we treat zero-inflated outcomes as continuous variables. GSEA is based on normalized Kolmogorov-Smirnov statistics and binary outcomes. Hence we used it for binary outcomes, that is, zero or nonzero count groups. We note that the proportion of overlap between the global and RF was 0.15. The proportion of overlap between our approach and the global test is 0.14. The proportion of overlap between our approach and random forest is 0.15. These results suggest that the proportions of overlap among different methods were small, meaning that each method detected different pathways.

Table 2.9: Top 50 significant pathways in terms of the Bayes factor in favor of Gaussian kernel on ZIP with unknown link over the Constant variance kernel.

Rank of Pathway	Pathway Id	Number of genes	Log marginal likelihood	Rank of Pathway	Pathway Id	Number of genes	Log marginal likelihood
1	254	29	-41.44	26	420	26	-51.17
2	127	49	-43.10	27	194	15	-51.48
3	108	152	-44.07	28	109	42	-51.48
4	271	29	-44.15	29	70	34	-51.59
5	11	47	-44.19	30	133	11	-52.24
6	339	25	-44.39	31	110	64	-52.45
7	169	20	-45.19	32	433	23	-52.56
8	119	40	-45.86	33	436	15	-52.81
9	413	26	-46.26	34	165	9	-52.85
10	341	14	-46.32	35	34	34	-53.03
11	440	59	-46.75	36	14	21	-54.05
12	257	23	-47.25	37	39	61	-54.21
13	114	40	-47.34	38	17	19	-54.40
14	277	20	-48.12	39	267	18	-54.62
15	376	27	-48.32	40	337	11	-55.05
16	233	21	-48.50	41	300	19	-55.14
17	321	17	-48.80	42	332	15	-55.28
18	234	28	-48.87	43	131	25	-55.30
19	240	27	-49.50	44	113	69	-55.59
20	302	32	-49.53	45	182	18	-55.79
21	189	23	-50.11	46	224	26	-55.79
22	124	26	-50.30	47	314	26	-55.92
23	25	34	-50.70	48	207	20	-55.94
24	358	21	-50.88	49	162	27	-56.99
25	59	69	-51.01	50	81	35	-57.04

Figure 2.2: Marginal likelihood values for all 441 pathways. The red circles represent top 50 pathways with the highest marginal likelihood.



We found that three pathways, being pathway 17 (*Androgen and estrogen metabolism*), pathway 39 (*Tryptophan metabolism*), and pathway 440 (*Leloir pathway of galactose metabolism*), were identified for all approaches. We note that pathway 4 (*Inhibition of Matrix Metalloproteinases*) overlapped with the random forest test and our approach.

We also found other pathways which were not identified using other existing approaches. The top 25 pathways and their names are listed in Table 2.10.

The pathways identified are clearly related to vascular injury, as they include pathways related to blood vessels and pathways that play a part in inflammation and tissue injury. It has been found that the MAPK signaling pathway (pathway 108) can be used as a target for anti-inflammatory therapy (Kaminska et al., 2005). Three sets of pathways are directly related to vascular injury: (1) the fatty acid metabolism pathway (pathway 11) is closely tied to vascular injury and disease (Semenkovich, 2004; Vecchione et al., 2006), (2) the signaling Pathway from G-Protein Families (pathway 271) contains receptor agonists and along with vascular injury causes vascular smooth muscle migration and proliferation (Ohta and Sitkovsky, 2001; Mallat and Lotersztajn, 2008), and (3) The pathway Actions of Nitric Oxide (pathway 339) in the Heart is particularly relevant as it has been noted in the literature that vascular injury results in a loss of endothelial nitric oxide (Ali, et al., 2008). The leloir pathway of the galactose metabolism pathway (pathway 440) contains the CCL2 gene that has been reported to affect gene expression in inflammatory vascular injury regulated by activation of NF-kappaB. Androgen and estrogen cell-specific interactions (pathway 17) were found to be in control of cellular proliferation in the vascular wall (Somjen et al., 1998). Gene RAC1 in the Cholera - Infection pathway has strong evidence in the literature to provide protection from diabetes-induced vascular injury (Vecchione et al. 2006). Moreover, Gene NFKB1 in the same pathway has been found to play an essential role in vascular healing (Vecchione et al., 2006). The deficiency

Table 2.10: Top 25 significant pathways and their names.

Rank of Pathway	Pathway Id	Pathway Name
1	254	fMLP induced chemokine gene expression in HMC-1 cells
2	127	Cholera - Infection
3	108	MAPK signaling pathway
4	271	Signaling Pathway from G-Protein Families
5	11	Fatty acid metabolism
6	339	Actions of Nitric Oxide in the Heart
7	169	ALK in cardiac myocytes
8	119	JAK-Stat Signaling Pathway
9	413	T Cell Receptor Signaling Pathway
10	341	Nitric Oxide Signaling Pathway
11	440	Leukocyte Adhesion(user defined)
12	257	Fc Epsilon Receptor I Signaling in Mast Cells
13	114	TGF-beta signaling pathway
14	277	Control of skeletal myogenesis by HDAC calcium/calmodulin-dependent kinase (CaMK)
15	376	Links between Pyk2 and Map Kinases
16	233	Erk and PI-3 Kinase Are Necessary for Collagen Binding in Corneal Epithelia
17	321	Role of MEF2D in T-cell Apoptosis
18	234	Phospholipids as signaling intermediaries
19	240	Regulation of eIF4e and p70 S6 Kinase
20	302	Integrin Signaling Pathway
21	189	BCR Signaling Pathway
22	124	Huntington's disease
23	25	Glycine, serine and threonine metabolism
24	358	Regulation of PGC-1a
25	59	Glycerolipid metabolism

of PRKCD, a gene in fMLP-induced chemokine gene expression in HMC-1 cells pathway (pathway 254), would accelerate lesions of an injured artery in mice (Bai, et al., 2010).

Furthermore, we calculated the posterior predictive probability $f(Y_{i,obs}|Y_{(i),obs})$, where $Y_{i,obs}$ and $Y_{(i),obs}$ represent the count of lesions for the ith dog and the counts of lesions for the other $(29-1)$ dogs, respectively, to show the agreement between the observations and the model (Pettit and Young, 1990). Denote all parameters in the model as a vector ν, We illustrate the calculation of $f(Y_{i,obs}|Y_{(i),obs})$ as below:

$$\begin{aligned} f(Y_{i,obs}|Y_{(i),obs}) &= \frac{f(Y_{i,obs}, Y_{(i),obs})}{f(Y_{(i),obs})} \\ &= \frac{1}{\int \frac{f(Y_{(i),obs},\nu)}{f(Y_{all,obs},\nu)} f(\nu|Y_{all,obs})d\nu} \\ &= \frac{1}{\int \frac{1}{f(Y_{i,obs}|\nu)} f(\nu|Y_{all,obs})d\nu}. \end{aligned}$$

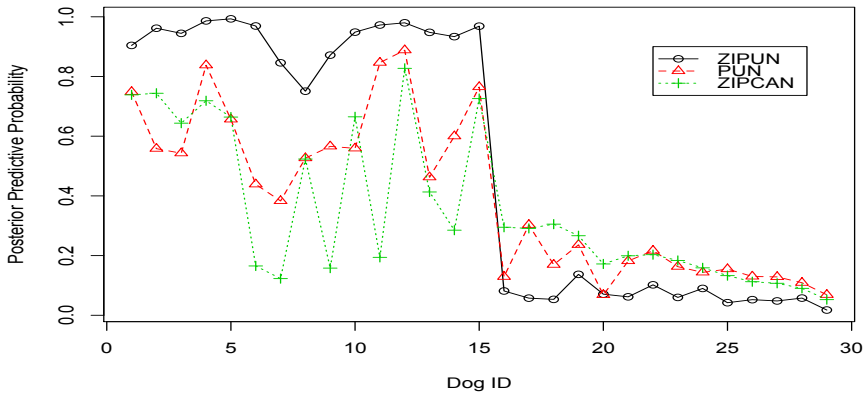
Therefore, with the posterior samples of ν, we can estimate the posterior predictive probability $f(Y_{i,obs}|Y_{(i),obs})$ as following:

$$\hat{f}(Y_{i,obs}|Y_{(i),obs}) = \frac{1}{\frac{1}{T}\sum_{t=1}^{T} \frac{1}{(Y_{i,obs}|\nu^t)}},$$

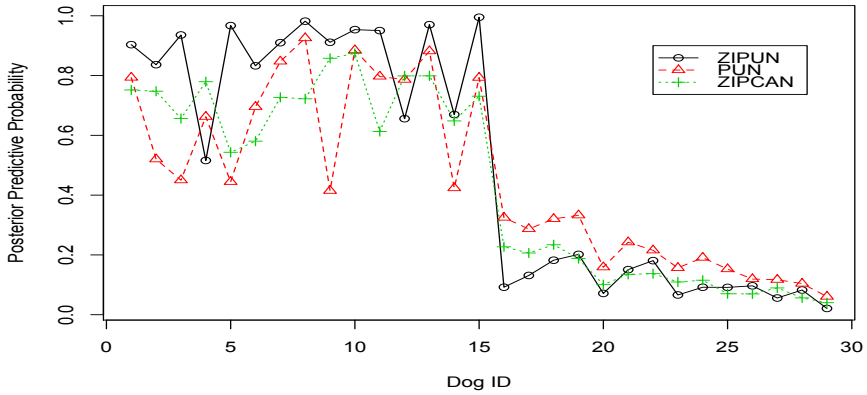
where T is the number of MCMC samples.

As an example, we randomly selected pathway 108 (*Map Kinase Inactivation of SMRT Corepressor*), pathway 119 (*Ion Channels and Their Functional Role in Vascular Endothelium*), and pathway 440 (*Leloir pathway of galactose metabolism*), which are ranked as 3, 8, and 11 out of 441 pathways, respectively, using our Bayesian inference. The posterior predictive probability $f(Y_{i,obs}|Y_{(i),obs})$ of pathways 108, 119, 440 is shown in Figure 2.3, suggesting that ZIP with unknown link successfully inflated the possibility of an observation with value 0 and also had a higher predictive density than ZIP with the canonical link.

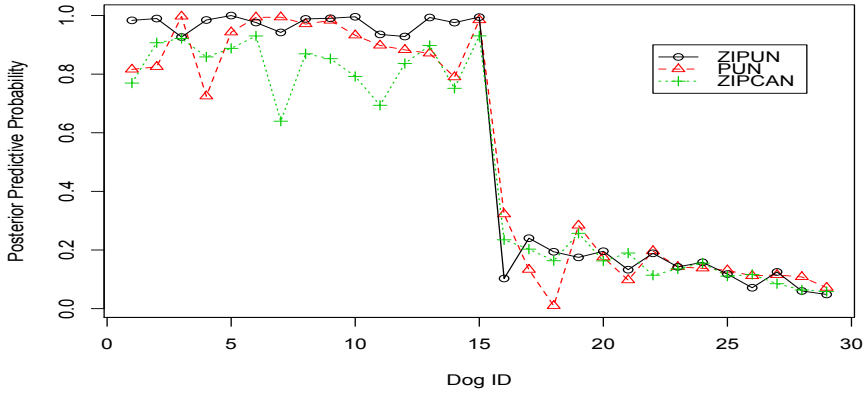
Figure 2.3: The conditional predictive ordinates of (1) ZIP with Unknown Link, (2) Poisson Regression with Unknown Link, and (3) ZIP with Canonical Link.



(a) Pathway 108



(b) Pathway 119



(c) Pathway 440

2.6 Conclusion and Discussion

In this chapter, we have proposed a semiparametric regression approach for pathway-based analysis with zero-inflated outcomes. Based on the best of our knowledge, there is no available approach for the zero-inflated outcome. Our approach is developed based on the Bayesian approach. Because of the high dimensional space of multivariate covariates, we estimate the nonparametric function of multivariate covariates by connecting a kernel machine with the Gaussian random process. The unknown link functions were estimated by transforming a mixture with an unknown link function. Our simulation results suggest that the Bayesian semiparametric approach has more accuracy and flexibility than a zero-inflated Poisson regression with canonical link function; this is especially true when the number of genes is large or the sample number and genes are both large. This result may be due to our nonparametric settings such as the Gaussian process and the unknown link which have the ability to describe the complicated association between gene expressions and clinical outcomes.

In many single gene-based analyses, many genes do not play an important role in differentiating the disease and non-disease groups. Similarly, in pathway analysis, we expect that many pathways are not involved. This mixture prior might be useful to incorporate such characteristics in the analysis. In order to incorporate prior knowledge into our Bayesian analysis, the prior π can be established from historical knowledge. Another hierarchical model we consider is under the same setting of the previous model (1) except for it uses the following mixture model:

$$\gamma(\mathbf{x}) \sim \pi\, GP\{0, \tau K(\mathbf{x}, \mathbf{x}')\} + (1-\pi)\, \delta_0(\mathrm{r}), \tag{2.4}$$

where $\delta_0(\mathrm{x})$ represents the point mass density at zero and $\pi \in [0, 1]$. With an additional

prior for π, e.g, uniform prior $\pi \sim \text{Unif}[0,1]$ or beta prior $\pi \sim B(a_\pi, b_\pi)$, the samples from the joint posterior distribution are drawn using the MH algorithm in a similar way to the method used in our previous model. However, we cannot calculate BF because point mass density at zero is an improper prior.

The model can be easily extended to situations with covariates Z in two ways. One is to assume an additive model that models $\gamma_1(\mathbf{x}_i) + z_i^T\beta_2$ and $\gamma_2(\mathbf{x}_i) + z_i^T\beta_1$; it implies that there is no interaction between $\mathbf{x}_i$ and z_i. The other way is to set the model matrix as $X' = (X, Z)$. When constructing the random predictive function, substitute $K(X)$ as $K(X')$ in order to model the pathway and covariate effect.

We note that we have analyzed each pathway separately in our analysis. It is known that pathways are not independent of each other because of interactions among them and shared genes. Extending our approach to multi-pathway analysis will be an interesting and challenging problem because of the complex dependence structure among pathways.

There are several other directions to be considered in future studies, such as multivariate analysis. Most methods for pathway analysis treat sets of genes as covariates to find which pathways are highly associated with clinical outcomes. The regression and classification models treat the gene expressions as the independent variables. However, as in some research, it is possible instead to treat the gene expressions as multivariate responses. One disadvantage of this approach is that if the outcome is count data, multiple comparison issues will arise, such as how many categories to classify the outcomes into, factoring in the need to balance the sample size in each category and the testing power. Our approach avoids this by treating the outcome as the response variable and simply specifying a discrete distribution for it. Both model directions lead to the association between the gene pathway changes and the distinct clinical outcomes. Comparison and connection between these two approaches are still open to future research. Hence

as a multivariate analysis, we can consider gene expression as the outcome and clinical outcomes as covariates. This is a worthwhile and challenging problem.

Another possible direction is to consider the measurement error of gene expression. This is fairly common, so it is important to develop methods to deal with this measurement error in pathway analysis. It may also be beneficial to consider whether measurement error from expression measurements would be more problematic when gene expression is a predictor rather than the outcome.

Chapter 3

Multilevel Gaussian Graphical Model for Gene and Pathway Networks

3.1 Introduction

Mathematically, a network can be thought as a collection of nodes that represent some physical units and edges that interconnect different nodes. We could use graph-theoretical notations to model mathematical networks. Let $G = (V, E, W)$, where $V = \{v_1, \ldots, v_p\}$ is the set of nodes, $E = (e_{i,j}), i, j = 1, \ldots, p$ is the set of edges, and $W = (W_{i,j}), i, j = 1, \ldots, p$ is the corresponding p by p adjacency matrix, with $W_{i,j} = 1$ when there is a link from i to j, and $W_{i,j} = 0$ when there is no link from i to j. Based on this conception, a gene network is a structure made up of genes called nodes, which are tied by one or more specific types of interdependency, such as co-functionality. Such ties can be viewed as edges in the graphical representation.

Gaussian graphical models (Dempster, 1972), also known as "covariance selection" or

"concentration graph" models, have recently become a popular tool to learn gene association networks. It assumes the nodes, i.e. gene expression data observed in our study, are randomly sampled observational or experimental data from a multivariate Gaussian distribution. That is, let $V = \{v_1, \ldots, v_p\}$ be the set of nodes (genes), and $X_1, \cdots, X_p$ denote the p genes, we assume that $(X_1, \cdots, X_p) \sim N(0, \Sigma)$ with positive definite variancecovariance matrix $\Sigma = (\sigma_{i,j})$ and precision matrix $\Omega = \Sigma^{-1} = (\omega_{i,j})$. Then, the Gaussian graphical model uses the precision matrix Ω as the adjacent matrix, i.e. $W_{i,j} = \delta(\omega_{i,j} \neq 0)$, and $\omega_{i,j} = 0 \Leftrightarrow e_{i,j} = 0$ means no association between the gene pair, and vice versa.

A related but completely different concept are the so-called gene "relevance networks", which are based on the covariance matrix Σ. The simple reason why Gaussian graphical models should be preferred over relevance networks for identification of gene networks is the off-diagonal elements of Ω are proportional to partial correlations, while the off-diagonal elements of Σ are proportional to marginal correlations. In the latter interactions are defined through standard correlation coefficients so that missing edges denote marginal independence only. The correlation coefficient is a weak criterion for measuring dependence, as marginally, i.e. directly and indirectly, more or less all genes will be correlated. This implies that zero marginal correlation is in fact a strong indicator for independence. On the other hand, partial correlation coefficients do provide a strong measure of dependence and, correspondingly, offer only a weak criterion of independence as most partial correlations coefficients usually vanish. And more often, with high dimension of genetic data, one would prefer concentrating the network size rather than trapping in a large amount of relevances resulted from relevance networks.

Application of Gaussian graphical models to high dimensional data is quite challenging because classical Gaussian graphical model theory is not valid in a small sample setting, but the number of genes p is usually much larger than the number of available sam-

ples n. Based on the Gaussian graphical model assumptions, the log-likelihood of the precision matrix is:

$$\log L(\Omega) = \log\{det(\Omega)\} - tr(S\Omega)$$

up to a constant. By taking the derivative of it, we can get

$$\frac{\partial \log L}{\partial \Omega} = \Omega^{-1} - S.$$

From the above, one can easily derive the maximum likelihood estimator for Ω is S^{-1}. However, such an estimator is expected to include many small values and therefore cannot be used directly to select edges of a concentrate graph. Also, when $n < p$, S is singular, the maximum likelihood estimator won't be unique.

There has been a number of studies work on estimating Ω. A popular way to estimate precision matrix for Gaussian graphical models with small sample modeling is to introduce penalty to the off diagonal elements in Ω, which is feasible in computing when $n < p$ and makes all off-diagonal element selection simutaneously. The sparsity of the obtained precision matrix would be able to take the nature of the genetic networks into account. Due to the small sample size in gene expression data, researchers usually take a penalized log likelihood approach and solve the following objective function,

$$\max_{\Omega} \big[\log\{det(\Omega)\} - tr(S\Omega) + \lambda P(\Omega) \big],$$

where λ is a non-negative penalty parameter and $P(\cdot)$ is a penalty function on the precision matrix elements. A popular penalty function is use the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996; Friedman et al., 2008; Yuan and Lin,

2007; Levina et al., 2008), which can be applied to shrink the off-diagonal elements in the precision matrix exactly to zeros. GLASSO algorithm (Friedman et al., 2008) based on a coordinate descent procedure is fast and can be adopted easily by many extension of LASSO. For example, to remedy the bias issue in LASSO, (Zou, 2006) proposed the adaptive LASSO penalty, by using the reciprocal of the absolute value of a consistent estimator raised to some power as the weight for each component. The solution can be obtained iteratively using weighted GLASSO. Another example is for joint estimation of multiple graphical models (Guo et al., 2011). They proposed a factor across data categories for each off diagonal elements to represent the homogeneity network structure and put LASSO penalty on both the elements and factors. Their solution could also be obtained from an iterative weighted GLASSO algorithm.

However, these recent studies only work on association among genes. That is, these methods can describe the association between single genes only. It is known that pathways are sets of genes which serve a particular cellular or physiological function. Hence pathways are not independent of each other because of shared genes and interactions among pathways. Multi-pathway analysis has been challenging problem because of the complex dependence structure among pathways. On the other hand, subtle connections between genes in two pathways may indicate strong connection between two pathways but can be ignored by individual gene network analysis. The main goal of our study is to develop a Gaussian graphical model for the gene and pathway network. Thus, by considering the dependency among pathways as well as genes winthin each pathway, we have proposed a multi-level Gaussian graphial model: one level is for pathway network structure and the second level is for gene network structure. We will propose a hierarchically structured graphical model for this in Section 3.2.

This chapter is organized as follows. In Section 3.2, we propose a multilevel Gaussian

graphical model for the gene and pathway network. Section 3.3 contains the penalized log likelihood approach and the development of the algorithm for the solution. In Section 3.4, we further illustrate asymptotic properties for the estimation. In Section 3.5 we compare our method with GLASSO method for individual gene network based on several criteria. We introduce a definition of pathway level connection degree. We also give a real data analysis in Section 3.6. Some conclusion and discussion is in Section 3.7.

3.2 Multilevel Gaussian Graphical Model for Gene and Pathway Networks

In this Section, we describe how to build a multilevel Gaussian graphical model for the gene and pathway networks. We'll firstly provide the precision matrix for the Gaussian graphical mode, then explain how to extract the pathway network information, and finally give and graphical illustration of the multilevel network model.

Suppose we have p genes, denoted by $X_1, \ldots, X_p$, the whole gene network can be represented by the precision matrix Ω,

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1p} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{p1} & \omega_{p2} & \cdots & \omega_{pp} \end{pmatrix}.$$

In this setting, if the off diagonal element $\omega_{ij} = 0$, it means the ith and jth genes are conditionally independent.

Furthermore, suppose these genes are in k predefined pathways, denoted by $P_1, \ldots, P_k$. Without loss of generality, we can re-denote the genes as: $X_{1,1}, \ldots X_{1,p_1}$, $X_{2,1}, \ldots, X_{2,p_2}, \ldots$, $X_{k,1}, \ldots, X_{k,p_k}$, where $p_1, p_2, \ldots, p_k$ are the number of genes in each pathway and $\sum_{i=1}^{k} p_i = p$. The conditional correlations among genes in k and k'th pathways can be rewrite as a p_k by $p_{k'}$ sub-block precision matrix $\Omega_{kk'}$,

$$\Omega_{kk'} = \begin{pmatrix} \omega_{11}^{(kk')} & \omega_{12}^{(kk')} & \cdots & \omega_{1p_{k'}}^{(kk')} \\ \omega_{21}^{(kk')} & \omega_{22}^{(kk')} & \cdots & \omega_{2p_{k'}}^{(kk')} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{p_k1}^{(kk')} & \omega_{p_k2}^{(kk')} & \cdots & \omega_{p_kp_{k'}}^{(kk')} \end{pmatrix}.$$

where $\omega_{ij}^{(kk')}$ is defined as follows. Let us reparameterize the precision matrix by introducing a pathway level factor $\theta_{kk'}$ for kth and k'th pathways. Then the partial correlation between jth gene in the kth pathway and $j'$the gene in the k'th pathway is $\omega_{j_k,j'_{k'}}^{(kk')} = \theta_{kk'}\gamma_{j_k,j'_{k'}}^{(kk')}$, where $\theta_{kk'} \geqslant 0$. The precision matrix becomes as follows;

$$\begin{aligned} \Omega &= \begin{pmatrix} \Omega_{11} & \Omega_{12} & \cdots & \Omega_{1K} \\ \Omega_{21} & \Omega_{22} & \cdots & \Omega_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_{K1} & \Omega_{K2} & \cdots & \Omega_{KK} \end{pmatrix} \\ &= \begin{pmatrix} \theta_{11}\Gamma_{11} & \theta_{12}\Gamma_{12} & \cdots & \theta_{1K}\Gamma_{1K} \\ \theta_{21}\Gamma_{21} & \theta_{22}\Gamma_{22} & \cdots & \theta_{2K}\Gamma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{K1}\Gamma_{K1} & \theta_{K2}\Gamma_{K2} & \cdots & \theta_{KK}\Gamma_{KK} \end{pmatrix} \end{aligned}$$

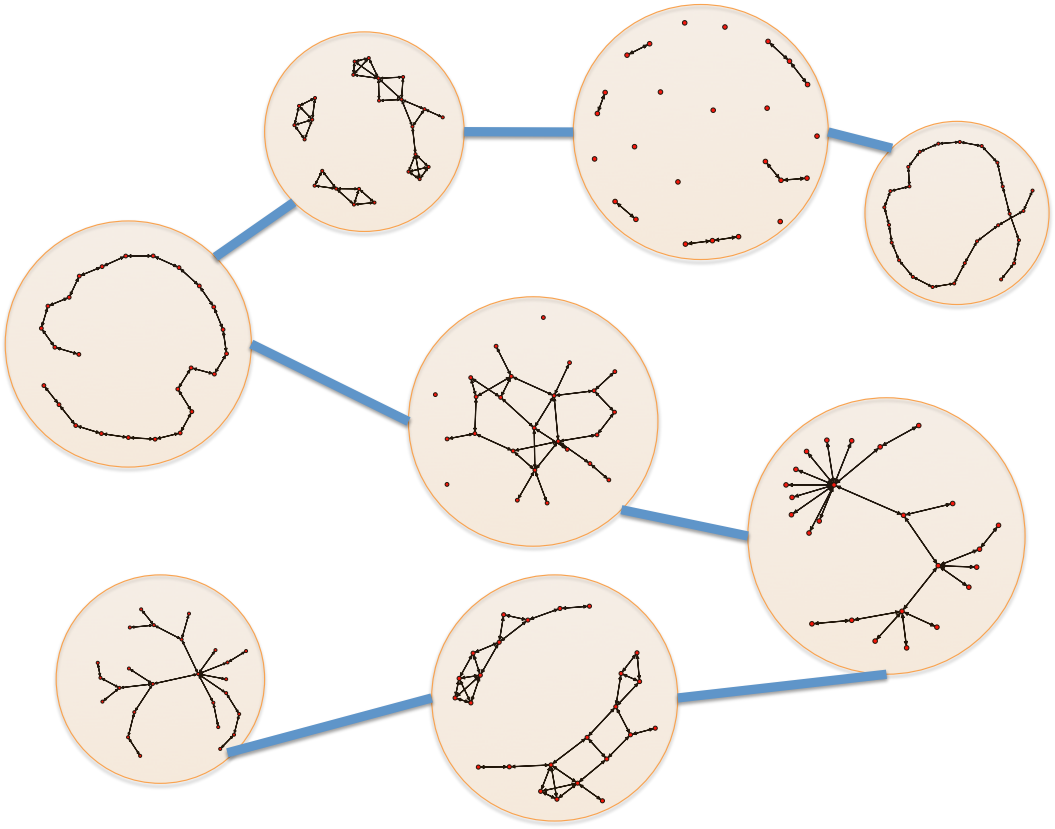
where

$$\Gamma_{kk'} = \begin{pmatrix} \gamma_{11}^{(kk')} & \gamma_{12}^{(kk')} & \cdots & \gamma_{1p_{k'}}^{(kk')} \\ \gamma_{21}^{(kk')} & \gamma_{22}^{(kk')} & \cdots & \gamma_{2p_{k'}}^{(kk')} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p_k1}^{(kk')} & \gamma_{p_k2}^{(kk')} & \cdots & \gamma_{p_kp_{k'}}^{(kk')} \end{pmatrix}$$

In our model, parameter $\theta_{kk'}$ indicates association between pathways and thus we call it "pathway level connection factor". Another parameter $\gamma_{j_k,j'_{k'}}^{(kk')}$ represents connections between individual genes, we call it "gene level connection factor". That is, if the off-diagonal factor $\theta_{kk'} = 0$, it means the k and k'th pathways are conditionally independent. Therefore, all the genes in the kth pathway has no connection to the genes in the other pathways. If $\theta_{kk'} \neq 0$, there is connection between these two pathways, to some extent. We further provide a definition of connection degree in Section 3.5 to measure how strong the connection is. This measurement is based on the individual gene level connection between genes in these two pathways, i.e. $\gamma_{j_k,j'_{k'}}^{(kk')}$. The diagnoal matrix Γ_{kk} represents the gene networks within kth pathway, so one should check the off diagonal elements of it for within pathway networks.

Figure 3.1 is an illustration of the proposed multi-level Gaussian graphical model. The large circles represent pathways, each dot in a circle represents a gene. The edges among circles represent dependencies among the pathways, and the lines beween dots represent dependencies among genes. With this graph, we could see which pathways are associated with each other, at the same time, we could see the gene networks within each pathway.

Figure 3.1: An illustration of the proposed multi-level Gaussian graphical model. The circles represent pathways and the points in the circles represent genes.



3.3 Estimation of Multi-level Gaussian Graphical Model

In this Chapter, we provide a penalized likelihood approach to estimate parameters and develop algorithm for the solution.

First we define some notation. Let A be a $p \times p$ matrix. We denote $det(A)$ as the determinant of A, $tr(A)$ as the trace of A, $\phi_{\max}(A)$ and $\phi_{\min}(A)$ as the maximum and minimum eigenvalues of A, respectively, and A^+ as the diagnoal matrix with the same diagnoal elements of A. We further define as follows: $A^- = A - A^+$, $||A||_F^2 = \sum_{i,j} a_{i,j}^2$,

$||A||^2 = \phi_{\max}(AA^T)$, and $|A|_1 = \sum_{i,j} |a_{i,j}|$.

3.3.1 The Penalized Approach

To estimate the multi-level Gaussian graphical model, and in order to obtain a sparse network at both gene and pathway level, we propose the following penalized log likelihood and denote it as **Q1**:

$$\min_{\{\Theta,\Gamma_{kk'},\Gamma_{kk}\}} \Big[\log L(\Omega) + \lambda P(\Omega) \Big] = \log\{det(\Omega)\} - tr(S\Omega) + \eta_1 |\Theta^-|_1 + \eta_2 \sum_{k \neq k'} |\Gamma_{kk'}|_1 + \eta_3 \sum_k |\Gamma^-_{kk}|_1$$

subject to

$$\Omega_{kk'} = \theta_{kk'}\Gamma_{kk'}, \theta_{kk'} > 0, 1 \leq k, k^{'} \leq K;$$
$$\theta_{kk'} = \theta_{k'k}, \Gamma_{kk'} = \Gamma^t_{k'k}, 1 \leq k \neq k^{'} \leq K;$$
$$\theta_k = 1, \Gamma_{kk} = \Omega_{kk}, 1 \leq k \leq K.$$

Note that there are three penalty functions applied to different components of Ω. The first one, $\eta_1|\Theta^-|_1$, is used to shrink the off diagonal pathway level connection factors. It will effectively remove really weak edges between pathways. The second one, $\eta_2 \sum_{k \neq k'} |\Gamma_{kk'}|_1$, controls sparsity of the network between genes in different pathways. That is, if some $\theta_{kk'}$ is not shrunk to zero, there would be connection between these two pathways. With this second penalty function, one can still obtain disconnection between genes in these two pathways. The third penalty, $\eta_3 \sum_k |\Gamma^-_{kk}|_1$, helps obtain a sparse gene network within each pathway. Although three penalty parameters provide more flexi-

bility, tuning three penalty parameters dramatically increase computing time. Thus we show that our problem can be simplified to the following version in Lemma 1

Lemma 1 Q1 is equivalent to the following **Q2**:

$$\min_{\{\Theta,\Gamma_{kk'},\Gamma_{kk}\}}\Big[\log L(\Omega)+\lambda P(\Omega)\Big]=\log\{det(\Omega)\}-tr(S\Omega) \\ +|\Theta^{-}|_1+\eta\sum_{k\neq k'}|\Gamma_{kk'}|_1+\eta_3\sum_{k}|\Gamma^{-}_{kk}|_1$$

subject to

$$\Omega_{kk'}=\theta_{kk'}\Gamma_{kk'},\theta_{kk'}>0,1\leq k,k^{'}\leq K;$$
$$\theta_{kk'}=\theta_{k'k},\Gamma_{kk'}=\Gamma^{t}_{k'k},1\leq k\neq k^{'}\leq K;$$
$$\theta_k=1,\Gamma_{kk}=\Omega_{kk},1\leq k\leq K.$$

To prove this, we need to show that if $(\hat{\Theta}^{**},\{\hat{\Gamma}^{**}_{kk'}\},\{\hat{\Gamma}^{**}_{kk}\})$ is a local minimizer of **Q1**, then there exist a local minimizer of **Q2**, denote as $(\hat{\Theta}^{*},\{\hat{\Gamma}^{*}_{kk'}\},\{\hat{\Gamma}^{*}_{kk}\})$, such that $\hat{\theta}^{**}_{kk'}\cdot\hat{\Gamma}^{**}_{kk'}=\hat{\theta}^{*}_{kk'}\cdot\hat{\Gamma}^{*}_{kk'}$, and vise versa. This proof is shown in Appendix A.

With this result, we have only two parameters η and η_3 instead of the three ones, which significantly reduces computing efforts.

3.3.2 The Algorithm

To make the optimization problem more achievable, **Q2** can be further reformulated. **Lemma 2** The **Q2** is equivalent to the following **Q3**:

$$\min_{\{\Theta,\Gamma_{kk'},\Gamma_{kk}\}}\Big[\log L(\Omega)+\lambda P(\Omega)\Big] = \log\{det(\Omega)\} - tr(S\Omega) + \lambda\sum_{k\neq k'}\sqrt{|\Omega_{kk'}|_1} + \eta_3\sum_{k}|\Gamma^{-}_{kk}|_1$$

where $\lambda = 2\sqrt{\eta}$.

To prove this, we need to show that if $\hat{\Omega}_{kk'}$ is a local minimizer of **Q3**, then there exists a local minimizer of **Q2**, denote as $(\hat{\Theta}, \{\hat{\Gamma}_{kk'}\}, \{\hat{\Gamma}_{kk}\})$, such that $\hat{\Omega}_{kk'} = \hat{\theta}_{kk'}\hat{\Gamma}_{kk'}$, and vise versa. We summarize the detail of the proof in Appendix B.

The square root in the term $\lambda\sum_{k\neq k'}\sqrt{|\Omega_{kk'}|_1}$ is like an umbrella penalty function which shaded every item under it. That is, it helps shrink all elements in $\Omega_{kk'}$ to zero at the same time. Therefore it plays a role of the pathway level connection factor and also gives less computation effort in practice.

The solution of **Q3** can be obtained though an interative approach based on the local linear approximation (Zou and Li, 2008). Let $\Omega^{(t)}_{kk'}$ be the estimation of the whole graph at the tth iteration, we can write the approximation as follows.

$$\sqrt{|\Omega_{kk'}|_1} \approx \frac{|\Omega_{kk'}|_1}{\sqrt{|\Omega^{(t)}_{kk'}|_1}}$$

Then at the $(t+1)$ iteration, the problem for the **Q3** can be written as **Q4**:

$$\min_{\{\Omega_{kk'},\Gamma_{kk}\}}\Big[\log L(\Omega)+\lambda P(\Omega)\Big] = \log\{det(\Omega)\} - tr(S\Omega) + \lambda\sum_{k\neq k'}\tau_{kk'}|\Omega_{kk'}|_1 + \eta_3\sum_{k}|\Gamma^{-}_{kk}|_1$$

where $\tau_{kk'} = 1/\sqrt{|\Omega^{(t)}_{kk'}|_1}$. The solution can be efficiently computed using the weighted GLASSO algorithm of Friedman et al. (2008). Since $\sqrt{|\Omega^{(t)}_{kk'}|_1}$ could be zero, we threshold it at 10^{-10} for computing stability.

In summary, the proposed algorithm at a candidate set of penalty parameters is following:

1 Initialize the precision matrix as $(\hat{\Sigma}+\nu I)^{-}1$ to guarantee a positive definite start.

2 Update $\Omega^{(t)}$ by **Q4** using weighted GLASSO;

3 Stop until convergence.

3.3.3 Model selection on the penalty parameter λ

For model selection on the penalty parameters, BIC criteria is often used. As the model is estimated through likelihood based approach, the BIC criteria is believed to be better than cross-validation. In the Gaussian graphical model, it is defined as

$$BIC(\lambda,\eta_3) = tr(S\hat{\Omega}) - \log|\hat{\Omega}| + \frac{\log(n)}{n}|\hat{\Omega}^{-}|_0.$$

where n is the sample size, S is the sample variance covariance matrix, $\hat{\Omega}$ is the estimated precision matrix and $|\hat{\Omega}^-|_0$ represents the number of nonzero elements in the off diagonal of $\hat{\Omega}$. The simplicity and effectiveness of the BIC have made it very attractive, and it is well known to lead to asymptotically consistent model selection in the setting of fixed number of variables p and growing sample size n. However, in a scenario where p grows moderately with n, it is observed to be usually too liberal to select a model with many spurious covariates (Bogdan et al., 2004; Broman and Speed, 2002; Siegmund, 2004).

Chen and Chen (2008) proposed an extended BIC for model selection with large model spaces. Drton and Foygel (2010) has shown that in chain graphical structures, this criteria performs better than either cross-validation or the ordinary BIC criteria, in terms of incurring a small loss in the positive selection rate but tightly controling the false discovery rate. The extended BIC in the graphical model is defined as

$$EBIC(\lambda, \eta_3) = tr(S\hat{\Omega}) - \log|\hat{\Omega}| + \frac{\log(n)}{n}|\hat{\Omega}^-|_0 + 4\xi\frac{\log(p)}{n}|\hat{\Omega}^-|_0.$$

where p is the number of nodes in the graphical model. There is a trade off between the positive selection rate and the false discovery rate based on the choice of the positive parameter ξ. From the simulations by Drton and Foygel (2010), $\xi = 0.5$ provides a good balance between the two evaluations.

While Gaussian graphical model is focused on the number of nodes in graph, our MGGM is more interested in off diagonal elements in the precision matrix. Hence we modify the extended BIC. We would like to use $\xi = 0.5$ in the extended BIC but with a modification on the degree of freedom part related to the number of parameters to be

estimated. It is denoted as EBIC$_m$ and is defined as

$$\begin{aligned} EBIC_m(\lambda, \eta_3) =& tr(S\hat{\Omega}) - \log(n)|\hat{\Omega}| + \frac{\log(n)}{n}|\hat{\Omega}^-|_0 \\ &+ 4\xi\frac{\log\{p(p-1)/2\}}{n}|\hat{\Omega}^-|_0. \end{aligned}$$

where $\log\{p(p-1)/2$ is the total number of off diagonal elements in the precision matrix. The extra degree of freedom results in stronger penalty of large graphs, which intuitively applicable to our scenario of sparsity at both gene and pathway level. We'll demonstrate its improvements of graph estimation at the pathway level as well in Section 3.5.

3.4 Asymptotic Properties

We consider two asymptotic properties. The first one is consistency. It implies that when the sample size n and model size p go to infinity and the tuning parameters goes to 0 at a certain rate, the estimated model goes to the true model at certain rate. The second one is sparsistency, which is more essential since one would prefer estimating zeros accurate in the covariance estimation.

We can show two asymptotic properties under two regularity conditions:

C1 There exist constants τ_1, τ_2 such that,

$$0 < \tau_1 < \phi_{\min}(\Omega_0) \leq \phi_{\max}(\Omega_0) < \tau_2 < \infty$$

where $\phi_{\min}$ and $\phi_{\max}$ indicate the minimal and maximal eigenvalues.

C2 There exist constant $\tau_3 > 0$ such that,

$$\min_{(j,j',k,k')\in T} |\omega_{0,jj'}^{kk'}| \geq \tau_3$$

where T is the nonzero set of $\{\omega_{0,jj'}^{kk'} : \omega_{0,jj'}^{kk'} \neq 0\}$.

The condition C1 regularize the eigen values for the true precision matrix that they should be all positive as the precision matrix requires positive definiteness, and also they should not be too large for well conditioned. The condition C2 implies the minimum element in the precision matrix has to be bounded away from 0.

With these conditions, the asymptotic properties on consistency and sparsity can be further addressed in Theorme 1 and 2.

Theorem 1 (Consistency)

Suppose that (i)-(iv) holds: (i) C1 and C2 hold; (ii) $(p+q)(\log p)/n = \circ(1)$*; and (iii)* $\Lambda_1\sqrt{(\log p)/n} \leq \lambda \leq \Lambda_2\sqrt{(1+p/q)(\log p)/n}$ *for some positive constants* Λ_1 *and* Λ_2*. Then, there exists a local minimizer* $\hat{\Omega}$*, such that,*

$$||\hat{\Omega} - \Omega_0||_F = O_P\left(\sqrt{\frac{(p+q)\log p}{n}}\right).$$

Details are given in the Appendix C. To prove Theorem 1, if we write $\Omega = \{\Omega_{kk'}\}$, $\Omega_0 = \{\Omega_{0,kk'}\}$, $\Delta = \{\Delta_{kk'}\}$, where$\Delta_{kk'} = \Omega_{kk'} - \Omega_{0,kk'}$, and let $G(\Delta) = Q_3(\Omega_0 + \Delta) - Q_3(\Delta)$, we need to show that if we take a closed bounded convex set $\mathcal{A}$ which contains 0, and G is strictly positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that G has a local minimum inside $\mathcal{A}$, since G is continuous and $G(0) = 0$.

Theorem 2 (Sparsity)

Suppose all conditions (i)-(iv) in Theorem 1 hold and (v) $||\hat{\Omega} - \Omega_0||_F = O_P(\eta_n)$ *hold, where* $\eta_n \to 0$ *and* $\{(\log p)/n\}^{1/2} + \eta_n^{1/2} = O(\lambda)$*. Then, with probability tending to 1, the local minimizer* $\hat{\Omega}$

satisfies $\hat{\omega}^{kk'}_{0,jj'} = 0$ *for all* $(j, j', k, k') \in T^c$*, where* T^c *is the gene and pathway pair index set where each element* (j, j', k, k') *in this set pointing to the pairs of gene* j *in pathway* k *and gene* j' *in pathway* k' *are independent.*

The proof of the Theorem 2 is given in Appendix D, where we show that $\forall(j, j', k, k') \in T^C$, the derivative $\partial Q_3/\partial\omega^{kk'}_{jj'}$ at $\hat{\omega}^{kk'}_{jj'}$ has the same sign as $\hat{\omega}^{kk'}_{jj'}$ with probability tending to 1. This is because suppose for some $(j, j', k, k') \in T^C$, the estimates $\hat{\omega}^{kk'}_{jj'} \neq 0$, without loss of generality, suppose $\hat{\omega}^{kk'}_{jj'} > 0$, then $\exists\xi > 0$ such that $\hat{\omega}^{kk'}_{jj'} - \xi > 0$. Since $\hat{\Omega}$ is a local minimizer, $\partial Q_3/\partial\omega^{kk'}_{jj'} < 0$ at $\hat{\omega}^{kk'}_{jj'} - \xi$ when ξ small enough, contracting the claim that $\partial Q_3/\partial\omega^{kk'}_{jj'}$ at $\hat{\omega}^{kk'}_{jj'}$ has the same sign as $\hat{\omega}^{kk'}_{jj'}$.

3.5 Simulation Study

In this section, we conduct simulation to compare the performance of our multi-level Gaussian graphical model with that of the regular graphical LASSO method.

3.5.1 Simulation Settings

The elements on the off diagonal is set to be $\exp(-\delta_g|i-j|)\exp(-\delta_p|p_i - p_j|)$, where δ_g and δ_p are set to ensure a well decayed positive definite matrix, i and j are the gene index and p_i and p_j are the pathway index. We set the gene disconnection rate $P(\gamma = 0) = 0.95$ and vary pathway disconnection rate $P(\theta = 0) = 0.9$,0.75,0.5. Two sample size are studied at $n = 50$ and $n = 200$. The number of pathways are also varied from $K = 5$ (10 pairs), 20 (45 pairs), to 50(190 pairs), with number of genes in each pathway varied from $p_k = 5$,10, to 50. Therefore, we have 72 ($= 4 \times 2 \times 3 \times 3$) combinations of simulation settings. For each combination, we generated 100 data sets.

We compare our approach with maximize likelihood method and regular graphical LASSO using 10 criteria. The first class of comparisons are the loss functions: entropy loss(EL), quadratic loss(QL), and Frobenius loss(FL) function. They are defined as

$$\begin{aligned}
\text{EL} &= tr(\Omega^{-1}\hat{\Omega}) - \log|\Omega^{-1}\hat{\Omega}| - n; \\
\text{QL} &= tr(\Omega^{-1}\hat{\Omega} - I)^2; \\
\text{FL} &= tr\{(\Omega^{-1} - \hat{\Omega} - I)^T(\Omega^{-1} - \hat{\Omega} - I)\}.
\end{aligned}$$

The second class of comparison is the sparsity that is the number of zeros estimated in the graphical model. The third class of comparison is the false selection rate at gene level: the false positive error(FP) and error rate(FPR), false negative error(FN) and error rate(FNR), as defined as follows:

$$\begin{aligned}
FP &= \text{\# non-zeros estimated if the trueth is zero}; \\
FPR &= \frac{FP}{\text{\# zeros in the true network}}; \\
FN &= \text{\# zeros estimated when the trueth is not zero}; \\
FNR &= \frac{FN}{\text{\# non-zeros in the true network}}.
\end{aligned}$$

Lastly, the fourth class of comparison is the assessment of pathway level selection accuracy using several measures. The first one we call it "pathway connection degree bias (PCDB)". To obtain it, we first define "pathway connection degree(PCD)": for the pathway k and k', there are $p_k \cdot p_{k'}$ possible edges, and the PCD is defined as the proportion of edges among these possible positions. When we consider all pairs of pathways, we can obtain a K by K PCD matrix, of which the diagonal elements are 1 and the off diagonal elements are the PCD's for each pathway pair. Next we take the PCD matrix for both

true and estimated gene network, and sum up the squared difference for the elements where the true PCD is positive, i.e. where the true pathway pairs are connected, one get the defined PCDB. Similarly, we define the "pathway disconnection degree bias (PDDB)". The difference is that it is based on "pathway disconnection degree(PDD)", and only sum up the squared PDD over the disconnected pathway pairs in the true precision matrix. Therefore we can formulate these measures as followings:

$$
\begin{aligned}
PCD_{kk'} &= P(\omega_{jj'}^{kk'} \neq 0) = \frac{\#\omega_{jj'}^{kk'} \neq 0}{p_k \cdot p_{k'}}; \\
PCDB &= \sum_{k \neq k'} (PCD_{kk'} - PCD_{0,kk'})^2 \mathbf{1}_{\{PCD_{0,kk'} > 0\}}; \\
PDD_{kk'} &= P(\omega_{jj'}^{kk'} = 0) = \frac{\#\omega_{jj'}^{kk'} = 0}{p_k \cdot p_{k'}}; \\
PDDB &= \sum_{k \neq k'} (PDD_{kk'} - PDD_{0,kk'})^2 \mathbf{1}_{\{PDD_{0,kk'} = 1\}}.
\end{aligned}
$$

3.5.2 Simulation Results

All of our simulations provided similar results, thus we summarize six common situations for demonstration:

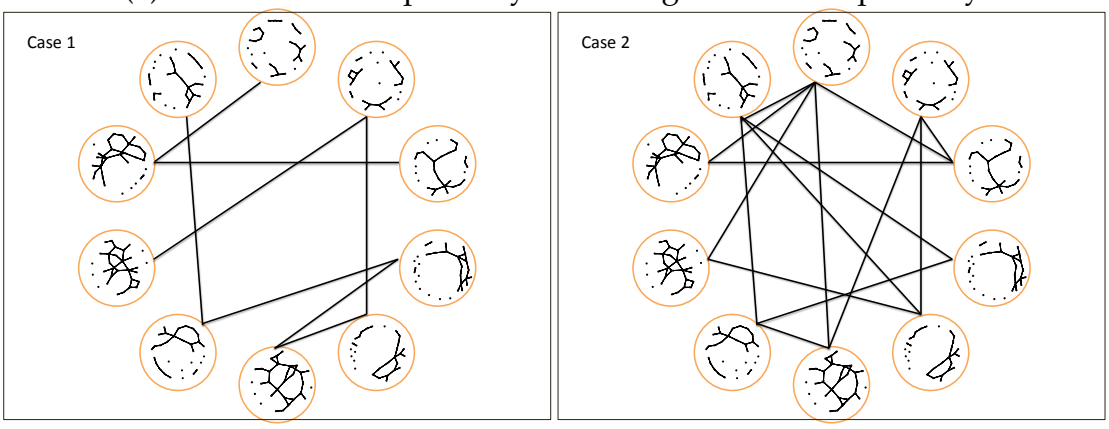
- Case 1: We generate 10 pathways ($K = 10$). Within each pathway, there are 30 genes ($p_k = 30$). The probability of the disconnection for each pair of genes $P_{\omega=0} = 0.9$, and the probability of the disconnection for each pair of pathway $P_{\theta=0} = 0.9$;
- Case 2: the same as case 1 except the pathway disconnection rate $P_{\theta=0} = 0.75$;
- Case 3: We generate 15 pathways ($K = 15$). Within each pathway, there are 30 genes ($p_k = 30$). The probability of the disconnection for each pair of genes $P_{\omega=0} = 0.9$, and the probability of the disconnection for each pair of pathway $P_{\theta=0} = 0.9$;

- Case 4: the same as case 3 except the pathway disconnection rate $P_{\theta=0} = 0.75$;
- Case 5: We generate 20 pathways ($K = 20$). Within each pathway, there are 5 genes ($p_k = 5$). The probability of the disconnection for each pair of genes $P_{\omega=0} = 0.9$, and the probability of the disconnection for each pair of pathway $P_{\theta=0} = 0.9$;
- Case 6: the same as case 5 except the pathway disconnection rate $P_{\theta=0} = 0.75$.

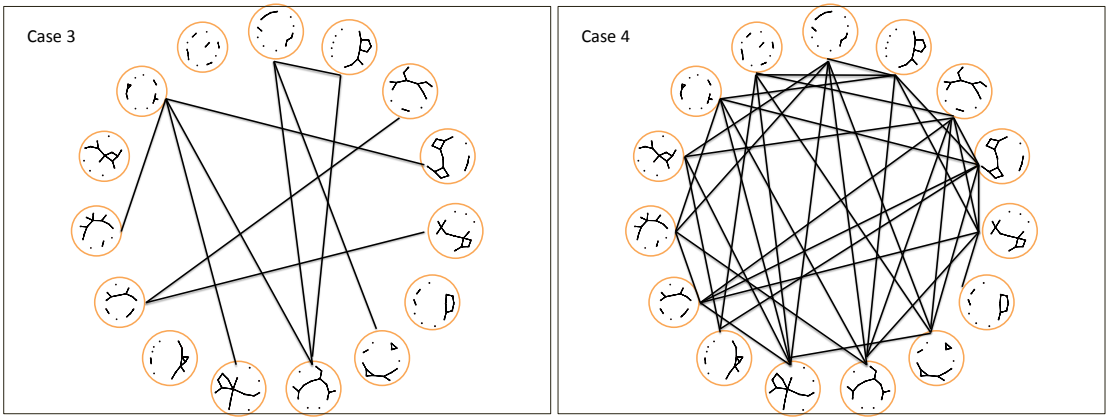
As illustrated by Figure 3.2, case 1 and 2 represent situations that we have small number of pathways but large number of genes in each pathway. Case 3 and 4 represent we have comparable number of genes within each pathway, relative to the number of pathways. And case 5 and 6 represent we have more pathways but small number of genes in each pathway. In case 1, 3, and 5, the probability of the connection for each pair of pathways $P_{\theta=0} = 0.9$, which means we have a sparser pathway level network than case 2, 4, and 6, where the pathway disconnection rate $P_{\theta=0} = 0.75$.

Figure 3.2: Simulated pathway and gene network for the simulation cases. The circles represent pathways and the points in the circles represent genes. Each row shows a different setting of the gene and pathway size. Left column shows sparser pathway networks with disconnection rate $P_{\theta=0} = 0.9$ while right column shows denser ones with $P_{\theta=0} = 0.75$.

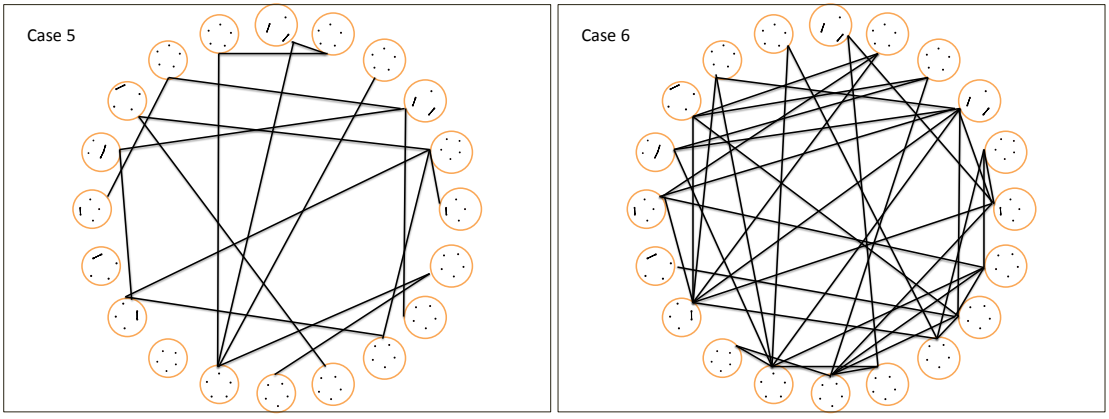
(a) Case 1 and 2: 10 pathways with 30 genes in each pathway.



(b) Case 3 and 4: 15 pathways with 15 genes in each pathway.



(c) Case 5 and 6: 20 pathways with 5 genes in each pathway.



In Section 3.3.3, we mentioned the preference of using the extended BIC to select the penalty parameters for the model. The advantage of the extended BIC has been demonstrated by Drton and Foygel (2010) in their simulation of one-level chain network structure. Comparing to the ordinary BIC, it will select a model with very few loss of positive selection rate but better control of false discovery rate. We would demonstrate that in our simulation settings, the extended BIC (with modified degree of freedom) also benefit the model selection by reduce the bias of pathway level connection or disconnection degree.

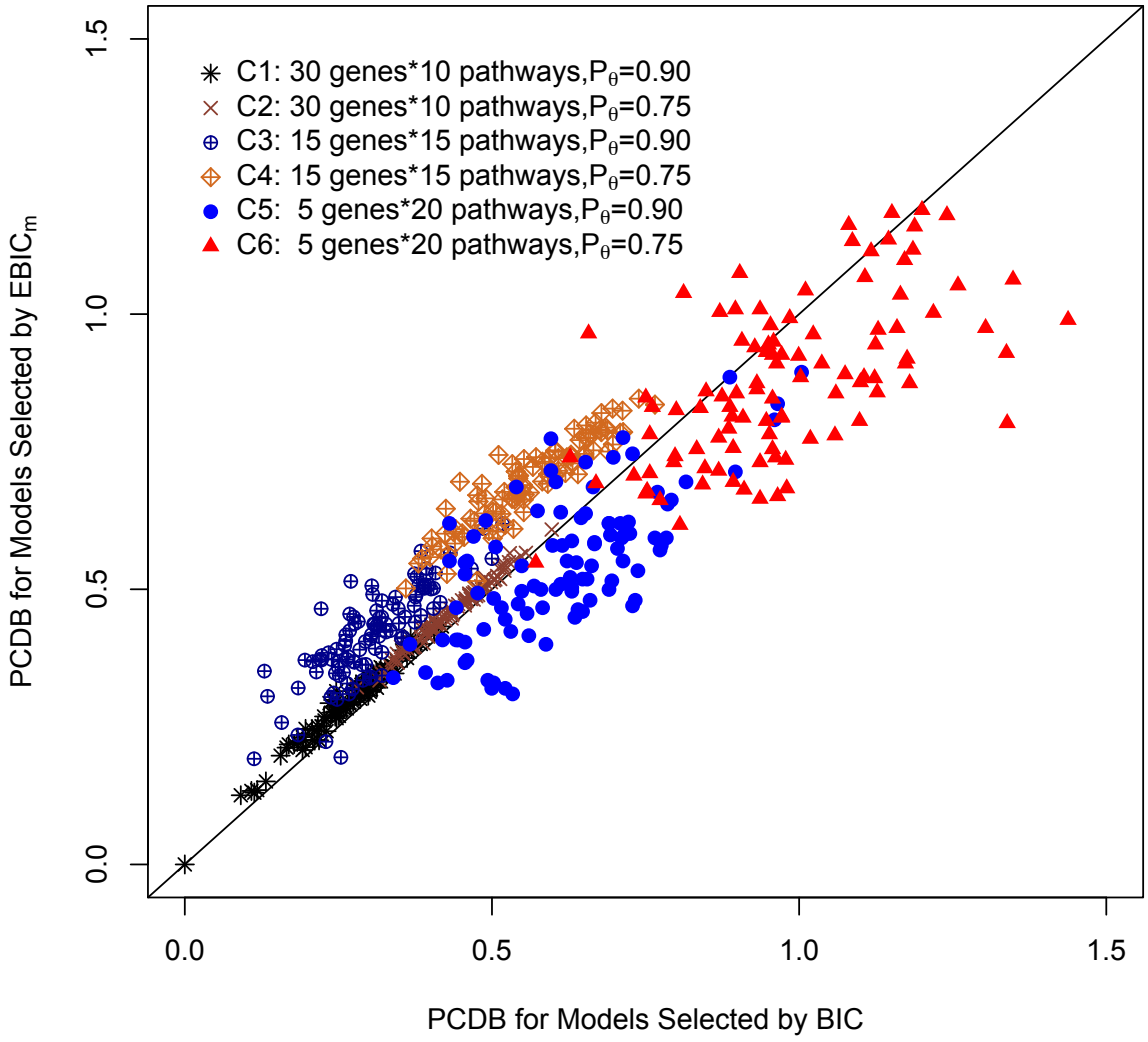
Figure 3.3 displays the scatter plot between the pathway connection degree bias (PCDB) obtained from the estimated models selected by EBIC$_m$, the modified extended BIC, versus those from the estimated models selected by the ordinary BIC. The smaller PCDB is, the better is. Each point in the figure represents the PCDB obtained from each case in a simulated data set. The diagonal line means the EBIC$_m$ and BIC will select a model producing the same PCDB, that is, they perform the same when estimating the overall connection strength for pathway pairs. If a point falls under the diagonal line, it means the EBIC$_m$-selected model results in a smaller bias when estimating the connection degree at the pathway level, and thus better than BIC. Similarly, if a point falls above the diagnoal line, it means EBIC$_m$ selects a model with larger pathway connection degree bias for this data set and thus worse than BIC.

We first compare PCDB among cases which have three different combinations of pathway and gene sizes. We could see that in case 1 and case 2 (10 pathways with 30 genes in each one), the PCDB from using BIC and EBIC$_m$ are the smallest and almost the same to each other. In case 3 and case 4 (15 pathways with 15 genes in each one), PCDB are generally larger than those in cases 1 and 2, and the comparison trend is parallelly shifted a little bit upward the diagonal line, which indicates a small but acceptable increase of PCDB incurs when using EBIC$_m$. In case 5 and case 6 (20 pathways with 5 genes in each

one), we could see that PCDB are larger than the other cases, and its variance from data set to data set are also the largest in all cases. There are more points under the diagonal line, indicating using EBIC$_m$ will be more likely to provide a smaller PCDB than using BIC. In general, the accuracy of EBIC$_m$ and BIC are similar for the inference of connection at the pathway level since they provide similar bias of pathway connection degree. Furthermore, we notice that when the pathway connection rate is higher, represented by points with cold tones (black, dark blue, and blue), the PCDB are smaller, which shows for sparser pathway network, the bias of pathway connection degree will be smaller.

Similarly, Figure 3.4 displays the scatter plot between pathway disconnection degree bias (PDDB) obtained from the estimated models selected by EBIC$_m$ and those from the estimated models selected by the ordinary BIC. The smaller PDDB is, the better is. Each point in the figure represents PDDB obtained from each case in a simulated data set. The diagonal line means the EBIC$_m$ and BIC will select a model producing the same PDDB, that is, they perform the same when estimating the overall disconnection degree for pathway pairs. If a point falls under the diagonal line, it means the EBIC$_m$-selected model results in a smaller bias when estimating the disconnection degree at the pathway level, and thus better than BIC. Similarly, if a point falls above the diagnoal line, it means EBIC$_m$ selects a model with larger pathway connection degree bias for this data set and thus worse than BIC.

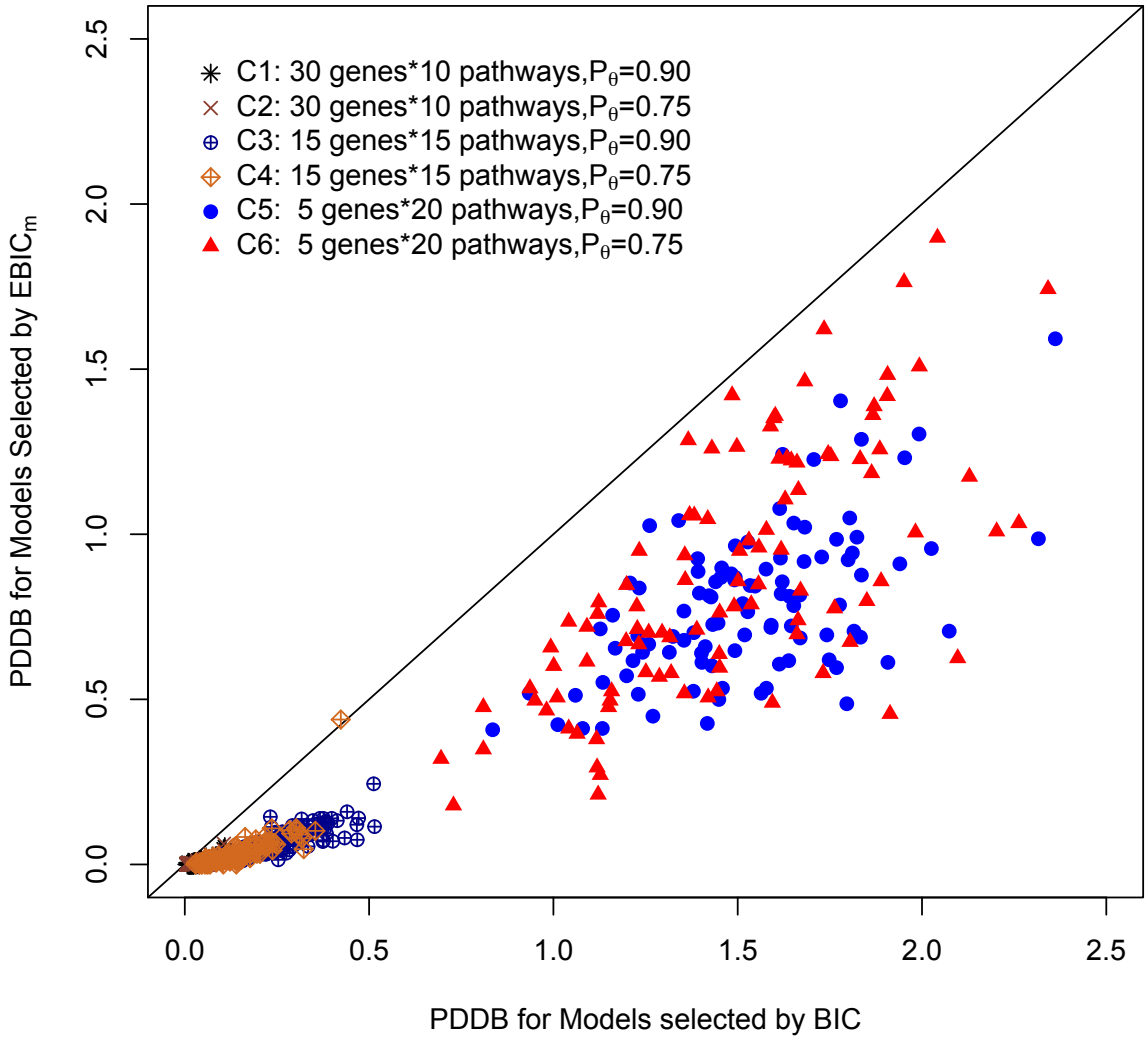
Figure 3.3: Pathway Connection Degree Bias(PCDB) comparison for model selection by the modified extended BIC ($EBIC_m$) and BIC criteria; C1-C6 represent Case 1-Case 6 which are considered in simulation study; Points are PCDBs obtained from different cases



We then compare PDDB among cases which have three different combination of pathway and gene sizes. We could see that in case 1 to case 4 (10 pathways with 30 genes in each one and 15 pathways with 15 genes in each one) the PDDB are smaller than those in

case 5 and 6, and the comparison trend turns underneath the diagonal line, with a rotation manner, not a parallel shifting, which indicates a significant increase of PDDB incurs when using BIC; in case 5 and case 6 (20 pathways with 5 genes in each one), PDDB are the largest and most scattering around among the pathway and gene size settings, with a similar comparison result to the cases 1-4, indicating using EBIC$_m$ provides significantly smaller PDDB than using BIC. In general, the accuracy of the selected model by EBIC$_m$ is much better than the one by BIC for the inference of disconnection at the pathway level since all the points are under the diagonal line. Furthermore, unlike PCDB, the PDDBs are similar to all cases no matter the pathway connection rate is 0.9 or 0.75. These results are shown in Figure 3.4. We can observe that the points are overlapped between two cases of different pathway connection rate under each setting of pathway and gene sizes.

Figure 3.4: Pathway Disconnection Degree Bias(PDDB) comparison for model selection by the modified extended BIC ($EBIC_m$) and BIC criteria; C1-C6 represent Case 1-Case 6 which are considered in simulation study; Points are PCDBs obtained from different Cases



The results of comparison on the maximum likelihood estimates(M.L.E.), graphical LASSO method(Glasso) and multilevel graphical LASSO(MGlasso) method based on the simulations are summarized in Tables 3.1-3.6. Table 3.1 and Table 3.2 show situations of

small number of pathways and large number of genes within each pathway. Table 3.3 and Table 3.4 show situations of the number of genes within pathways is comparable to the number of pathways. Table 3.5 and Table 3.6 show situations of larger number of pathways but less genes within pathways.

By checking all the cases, we can see that the maximum likelihood estimator almost always gives the smallest entropy loss because it doesn't shrink the parameters. Our method, the multilevel graphical LASSO, has the advantage of reducing the quadratic and Frobenius loss. The maximum likelihood estimator would never give a sparse network estimation, in terms of no zero appearing on the off diagonal in the estimated precision matrix, while graphcial LASSO and our method give similar sparsity, with our method a little bit sparser overall. GLASSO and our method have comparable false positive rate. But our method gives much better false negative rate. For the pathway level, our method gives comparable disconnection degree bias to the graphical LASSO, but has smaller bias for the connection degree. In general, our method is better for the gene level covariance selection.

Additionally, by comparing Table 3.1 and Table 3.2, which show the same setting of the pathway and gene size but with different sparsity on the pathway network, we can see when the pathway level network is sparser ($P_\theta = 0.9$), the more obvious advantage can be seen for our method in terms of the pathway level criteria PCDB. A similar story can be found by comparing Table 3.3 and Table 3.4, or Table 3.5 and Table 3.6.

By comparing Table 3.2, Table 3.4 and Table 3.6, which show situations of increasing number of pathways but reduced number of genes within pathways, we found that the pathway disconnection degree bias, PDDB, of our method performed better in the case where number of pathways are larger, because the it goes more smaller than those of graphical LASSO. A similar story can be found by comparing Table 3.1, Table 3.3 and

Table 3.1: $K = 10$, $p_k = 30$, $P_{\omega=0} = 0.9$, $P_{\theta=0} = 0.9$; M.L.E=Maximum likelihood estimator; Glasso=Graphical lasso; MGlasso=our multilevel Graphical lasso; EL=entropy loss; QL=quadratic loss; FL= Frobenius loss; $P_{\hat{\omega}=0}$ = relative frequency of estimated zeros; FP= false positive error; FN= false negative error; FPR= false positive error rate; FNR= false negative error ratel; PDDB= pathway disconnection degree bias; PCDB= pathway connection degree bias.

Method	EL	QL	FL	$P_{\hat{\omega}=0}$	FP
M.L.E	-INF	1.83e33	1.06e18	0	44787
Glasso	515.79	237.06	29.39	0.9997	9
MGlasso	221.60	117.22	22.16	0.9991	25
Method	FN	FPR	FNR	PDDB	PCDB
M.L.E	0	1.0000	0.00	9.27	1.9767
Glasso	59	0.0002	0.27	0.00	0.0205
MGlasso	50	0.0006	0.23	0.00	0.0005

Table 3.5.

3.6 Real Data Analysis

In this chapter we give a real data analysis example. The data set comes from Enerson et al. (2006). The goal of this study is to identify association among top ranked pathways and distinguish between lesion disease and non-lesion in dogs. There were 15 dogs with lesion and 14 without, with their microarray gene expression measurements. There were a total of 441 pathways and 6,592 genes. The Canine dataset is generated from investigative toxicology studies designed to identify the molecular pathogenesis of drug-induced vascular injury in coronary arteries of dogs treated with adenosine recept eragonist CI-947. Then, the Canine genes are mapped to human orthologs for pathway analysis. The human orthologs for dogs are generated by matching the genes sequence using BLASTx (Enerson et al., 2006). We picked the top 10 pathways ranked by the pathway evaluation of the dog lesions based on results from the random forest model (Pang et al., 2006).

Table 3.2: $K = 10$, $p_k = 30$, $P_{\omega=0} = 0.9$, $P_{\theta=0} = 0.75$; M.L.E=Maximum likelihood estimator; Glasso=Graphical lasso; MGlasso=our multilevel Graphical lasso; EL=entropy loss; QL=quadratic loss; FL= Frobenius loss; $P_{\hat{\omega}=0}$ = relative frequency of estimated zeros; FP= false positive error; FN= false negative error; FPR= false positive error rate; FNR= false negative error ratel; PDDB= pathway disconnection degree bias; PCDB= pathway connection degree bias.

Method	EL	QL	FL	$P_{\hat{\omega}=0}$	FP
M.L.E	-INF	4.35e33	2.60e18	0	43425
Glasso	INF	256.50	75.95	1.000	0
MGlasso	365.80	159.75	67.67	0.9979	52
Method	FN	FPR	FNR	PDDB	PCDB
M.L.E	0	1.0000	0.00	8.12	4.4296
Glasso	1425	0.0000	0.90	0.00	0.4707
MGlasso	1386	0.0012	0.88	0.00	0.4526

Table 3.3: $K = 15$, $p_k = 15$, $P_{\omega=0} = 0.9$, $P_{\theta=0} = 0.9$; M.L.E=Maximum likelihood estimator; Glasso=Graphical lasso; MGlasso=our multilevel Graphical lasso; EL=entropy loss; QL=quadratic loss; FL= Frobenius loss; $P_{\hat{\omega}=0}$ = relative frequency of estimated zeros; FP= false positive error; FN= false negative error; FPR= false positive error rate; FNR= false negative error ratel; PDDB= pathway disconnection degree bias; PCDB= pathway connection degree bias.

Method	EL	QL	FL	$P_{\hat{\omega}=0}$	FP
M.L.E	-INF	3.58e31	6.73e16	0	24793
Glasso	124.99	44.14	17.73	0.9939	83
MGlasso	118.79	42.07	17.52	0.9930	19
Method	FN	FPR	FNR	PDDB	PCDB
M.L.E	0	1.0000	0.00	13.71	4.1819
Glasso	338	0.0033	0.69	0.09	0.5200
MGlasso	368	0.0008	0.65	0.00	0.4529

Table 3.4: $K = 15$, $p_k = 15$, $P_{\omega=0} = 0.9$, $P_{\theta=0} = 0.75$; M.L.E=Maximum likelihood estimator; Glasso=Graphical lasso; MGlasso=our multilevel Graphical lasso; EL=entropy loss; QL=quadratic loss; FL= Frobenius loss; $P_{\hat{\omega}=0}$ = relative frequency of estimated zeros; FP= false positive error; FN= false negative error; FPR= false positive error rate; FNR= false negative error ratel; PDDB= pathway disconnection degree bias; PCDB= pathway connection degree bias.

Method	EL	QL	FL	$P_{\hat{\omega}=0}$	FP
M.L.E	-INF	4.64e31	1.16e17	0	24540
Glasso	122.31	54.71	20.38	0.9934	11
MGlasso	98.74	30.89	18.82	0.9941	4
Method	FN	FPR	FNR	PDDB	PCDB
M.L.E	0	1.0000	0.00	12.72	6.2167
Glasso	339	0.0004	0.82	0.009	0.7294
MGlasso	345	0.0002	0.81	0.000	0.7144

Table 3.5: $K = 20$, $p_k = 5$, $P_{\omega=0} = 0.9$, $P_{\theta=0} = 0.9$; M.L.E=Maximum likelihood estimator; Glasso=Graphical lasso; MGlasso=our multilevel Graphical lasso; EL=entropy loss; QL=quadratic loss; FL= Frobenius loss; $P_{\hat{\omega}=0}$ = relative frequency of estimated zeros; FP= false positive error; FN= false negative error; FPR= false positive error rate; FNR= false negative error ratel; PDDB= pathway disconnection degree bias; PCDB= pathway connection degree bias.

Method	EL	QL	FL	$P_{\hat{\omega}=0}$	FP
M.L.E	-22.98	127.51	50.60	0	4899
Glasso	-51.55	40.86	15.80	0.9977	5
Mglasso	-55.82	38.98	15.33	0.9901	24
Method	FN	FPR	FNR	PDDB	PCDB
M.L.E	0	1.0000	0.00	18.97	3.9642
Glasso	22	0.0010	0.44	0.00	0.6145
Mglasso	13	0.0049	0.25	0.00	0.5713

Table 3.6: $K = 20$, $p_k = 5$, $P_{\omega=0} = 0.9$, $P_{\theta=0} = 0.75$;M.L.E=Maximum likelihood estimator; Glasso=Graphical lasso; MGlasso=our multilevel Graphical lasso; EL=entropy loss; QL=quadratic loss; FL= Frobenius loss; $P_{\hat{\omega}=0}$ = relative frequency of estimated zeros; FP= false positive error; FN= false negative error; FPR= false positive error rate; FNR= false negative error ratel; PDDB= pathway disconnection degree bias; PCDB= pathway connection degree bias.

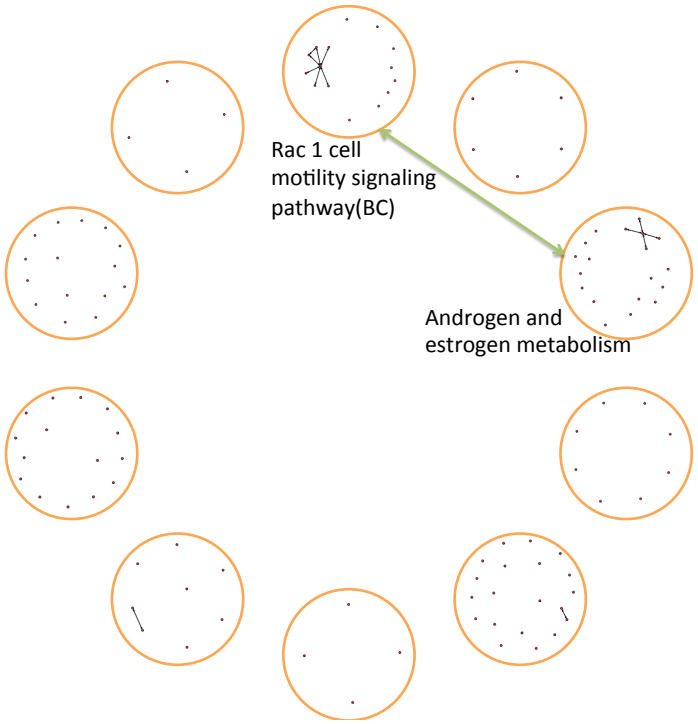
Method	EL	QL	FL	$P_{\hat{\omega}=0}$	FP
M.L.E	-28.47	120.33	47.35	0	4811
Glasso	-52.51	71.79	20.77	0.9985	3
Mglasso	-39.22	34.54	17.14	0.9947	15
Method	FN	FPR	FNR	PDDB	PCDB
M.L.E	0	1.0000	0.00	16.79	8.9248
Glasso	135	0.0006	0.71	0.05	1.0822
Mglasso	128	0.0031	0.67	0.00	1.0598

The gene and pathway network for the top 10 pathways for the two dog categories: healthy dog and unhealthy dog are displayed in Figure 3.5. We found one pair of pathways are connected in healthy dogs but the connection breaks in unhealthy dogs. This pair of pathways are "Androgen and estrogen metabolism" and "Rac 1 cell motility signaling pathway(BC)".

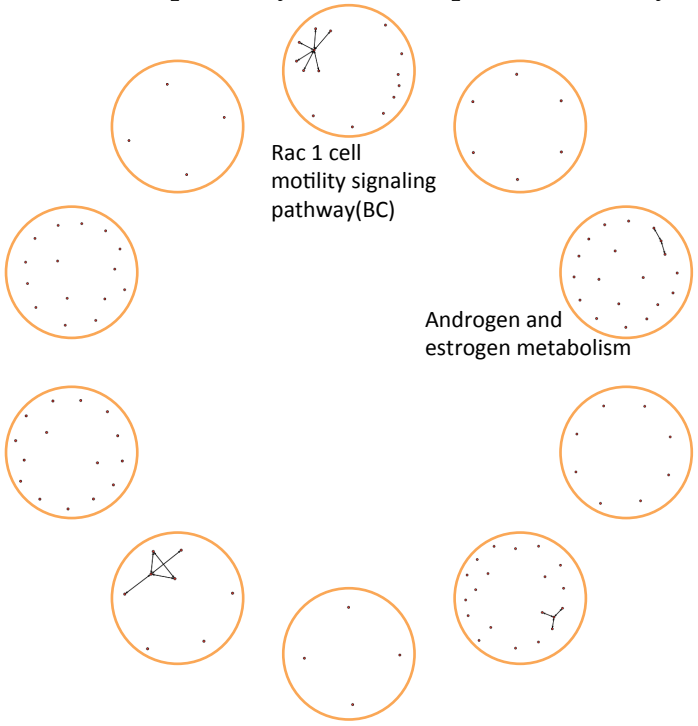
The pathway "Androgen and estrogen metabolism" describes the inactivation and catabolism of male (androgen) and female (estrogen) hormones. It is related to vascular diseases since the estrogen is demonstrated in many experiments to have vasodilator effects on endothelial cells, vascular smooth muscle and extracellular matrix(Mendelsohn, 2002; Smiley and Khalil, 2009). It is also reported that androgen and estrogen metabolism has effect on the inflammatory process, oxidative stress, and angiogenesis(Miller and Duckles, 2008). The association of estrogen with white adipose tissue mass with consequent changes in circulating lipid levels and inflammatory cytokines, pointing to another mechanism by which estrogen might benefit vascular health (Simpson et al., 2005). Massod et al. (2010) also learned the impact of the sex hormone metabolism on vascular

effects of hormone therapy in cardivascular disease. Rac-1 is a small G-protein in the Rho family that regulates cell motility in response to extracellular signals. Several changes in cytoskeletal structure and other aspects of cell structure are involved in cell motility. It was indicated that the characteristics of vascular smooth muscle cell phenotypes as they relate to cell migration (Louis and Zahradka, 2010). And the cell migration process is driven by the small GTPases from the Rho family, primarily by Rac and Cdc42 (Cosco et al., 1995; Olson et al., 1995; Pankova et al., 2010).

Figure 3.5: Network for pathways ranked top 10 using random forest classification (Up: healthy dogs; Down: unhealthy dogs).



(a) Network for pathways ranked top 10 for healthy dogs.



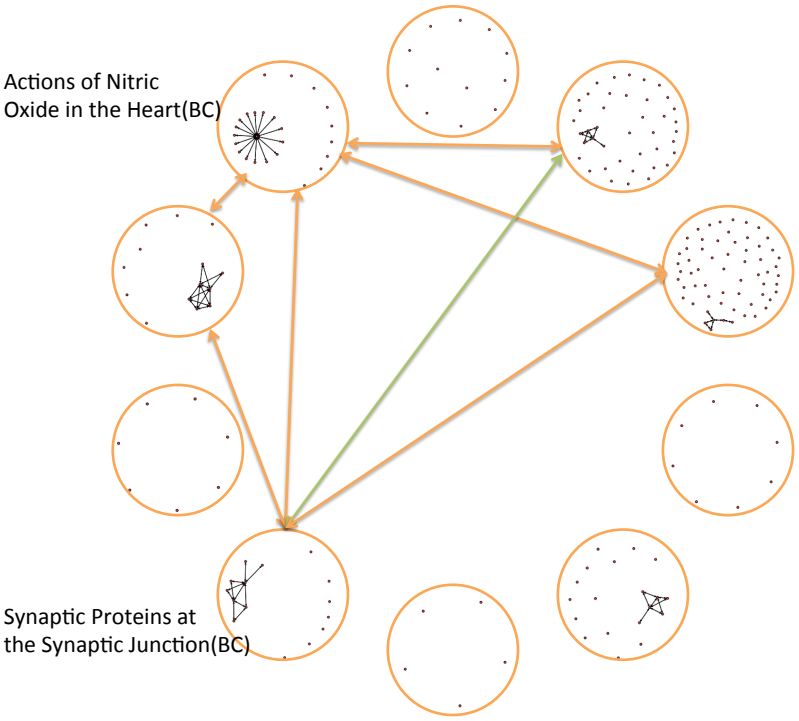
(b) Network for pathways ranked top 10 for unhealthy dogs.

There is some potential support for the interaction between these two pathways: It was examined that Rac1 GTPase gene-transcription and activity is down-regulated by an estdogen, 17 beta-estradiol, in vitro, vivo, and observed in human mononuclear cells, which may be an important molecular mechanism contributing to the cardiovascular effects of estrogens (Laufs et al., 2003; Li et al., 2011); and on the other hand, inhibition of the Rac1 decreases estrogen receptor levels (Rosenblatt et al., 2011).

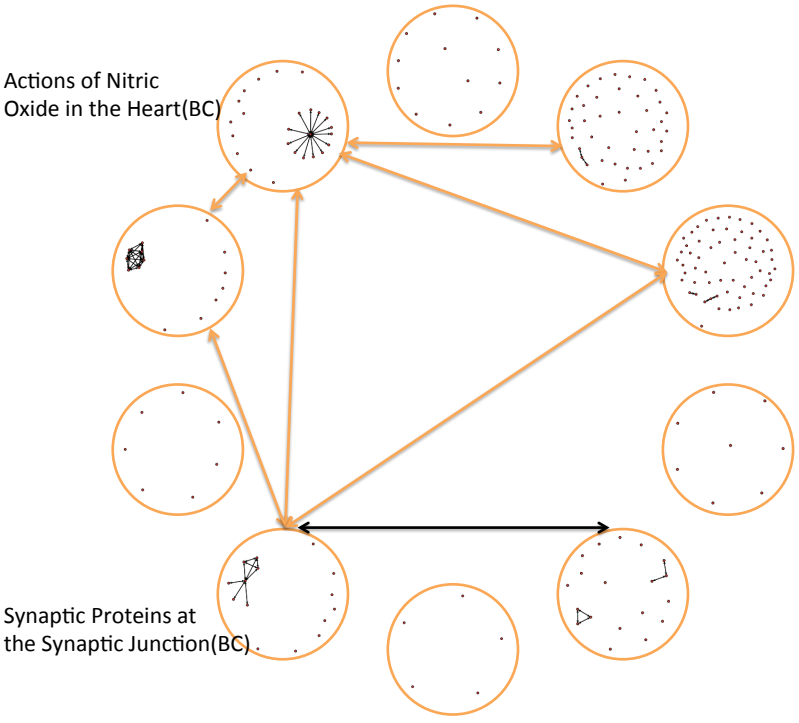
By examining the pathways ranked 11-20 for healthy and unhealthy dogs, as shown in Figure 3.6, we found two pathways that are important from another view: "Actions of Nitric Oxide in the Heart" and "Synaptic Proteins at the Synaptic Junctions(BC)". In the network, they are connected to many other top ranked pathways like a"hub".

The pathway"Actions of Nitric Oxide in the Heart" plays an important role in vasular diseases is easy to understand since nitric oxide is an important regulator and mediator of numerous processes in the nervous, immune, and cardiovascular systems, including vascular smooth muscle relaxation, resulting in arterial vasodilation and increasing blood flow. For the pathway "Synaptic Proteins at the Synaptic Junctions(BC)", research literatures related with vascular activity is limited. However, some researches has found neurexins and neuroligins, which constitute large and complex families of fundamental players in synaptic activity, are produced and processed by endothelial and vascular smooth muscle cells throughout the vasculature. Moreover, they are dynamically regulated during vessel remodeling and form endogenous complexes in large vessels as well as in the brain (Botto et al., 2009). Based on the large variety of biofunctions of nitric oxide, the extensive places of synapses in the animal or human body with their central station-styled transmition role, the active connections to other pathways could be expected.

Figure 3.6: Network for pathways ranked top 11-20 using random forest classification (Up: healthy dogs; Down: unhealthy dogs).



(a) Network for pathways ranked top 11-20 for healthy dogs.



(b) Network for pathways ranked top 11-20 for unhealthy dogs.

3.7 Conclusion and Discussion

In this chapter, we have proposed a multilevel Gaussian graphical model (MGGM), in which one level describes the networks for genes and the other for pathways. A penalized likelihood approach is developed to achieve the sparseness on both levels. We provided an iterative weighted graphical LASSO algorithm for MGGM. In addition to some common criteria people used to evaluate model estimation and selection, we developed two criteria: pathway connection degree bias and pathway disconnection degree bias, to evaluate the performance of a model for the data analysis at the pathway level. Our simulation results supported the advantages of our approach; our method estimated the network more accurate on the pathway level, and sparser on the gene level. We also demonstrated usefulness of our approach using a canine genes-pathways data set.

One advantage of our MGGM is that it is applicable to the case where some genes appear in different pathways. In practice, several pathways share certain genes. In MGGM, the dependency among genes is higly associated to the dependency among other genes and among pathways. The connections among pathways are affected by a set of genes. Hence, it is possible that the pathways, which have shared genes, may have weak connection, because other genes in these pathways may affect more than the shared ones. This can be seen from the penalty weight in the pathway connection. For this penalty, we add up the connection strength of all gene pairs between two pathways, then take reciprocal of the square root of it. Hence the penalty can still be large if gene pairs other than the shared ones are weakly connected. Therefore, the pathways connection will be shrunk to weak.

We note that our approach is based on a multilevel Gaussian graphical model for a single class. For example, we analyze the healthy and unhealthy dogs seperately. Devel-

oping our approach with multiple classes will be worthwhile for the future research.

Chapter 4

Summary and Future Work

Summary

In this dissertation, we have described two statistical topics with application to the high dimensional genetic pathway data analysis. Specifically, the analyses are all focused on pathway (i.e. multiple genes) level: one is for testing and ranking the pathway effect, and the other is for exploring the pathway dependencies.

We first discuss a zero inflated Poisson regression model for analyzing and ranking the pathway effects, under the Bayesian framework. The high dimensional problem is addressed through the Gaussian process. That is, instead of modeling the effect from each gene in the pathway and their interactions to fit a global model for all observations, we propose a random function from the Gaussian process, by treating multiple genes in a pathway as a whole and incorporating dependency among genes in a pathway into the variance covariance matrix of the random function. In this way, we can test the pathway effect by setting the null hypothesis as the variance covariance function is not related to

the pathway and applying Bayes factor to compare it with our semiparametric model using observed pathway data. The link function was modeled nonparametrically using a mixture of cumulative Beta density functions, which is a broader class covering the link functions than is typically used, and it is more flexible and data driven than the specified forms.

The second topic is to build a multilevel Gaussian graphical model. The multilevel means one level for pathway network and one for gene network within each pathway. The connection of genes is under a typical setting, i.e. by assuming the gene expressions are from a multivariate normal distribution, the off diagonal elements in the precision matrix represent the network connections. The dependency between each pathway pairs was modeled by extracting a non-negative factor for all pairs of genes from either pathway, and the interpretation of this factor is analogous to the off diagonal elements in the precision matrix, i.e. zero means two pathways are disconnected and vise versa. A weighted graphical LASSO algorithm is developed for parameter estimation. The asymptotic properties are also illustrated. In our simulation, we have found that for selection of the penalty parameters, rather than using BIC, the use of extended BIC produces lower bias at the pathway level in estimating the disconnection strength while maintaining almost the same accuracy in estimation the connection strength. Our simulation results support the advantage of the multilevel Gaussian graphical model; the maximum likelihood estimation never give us a sparse network with finite samples, and the ordinary graphical LASSO gives worse pathway level estimation accuracy and is less sparse at the gene level, although the loss functions and the gene level false positive and negative selection rates are comparable.

Future Work

The high dimensional genetic pathway-based data analysis is a open research area. There are still many remaining issues to address. Many interesting problems require further work.

For our first topic, a Bayesian semiparametric zero inflated Poisson regression model with unknown link function and Gaussian process to represent pathway effects on the clinical outcomes, we list the following directions for the future work:

- **Extension to measurement error model:** Gene expression level measured by the microarray technology may have variances and it is commonly observed the variance increases proportionally with the intensity (Wang et al., 2009). So, one direction of our future work could involve the measurement error model. Specifically, the Gaussian process will be built conditional on the unknown true gene expression level, while the observed gene expression level is centered around the true values with measurement error. By this means, the Gaussian process is a random function on a random field, i.e. it considers modeling the uncertainty of the variance covariance structure in the Gaussian process and thus increases the power of detecting the gene and pathway effects. In order to obtain a good estimate on the uncertainty and realize the benefit, a larger sample size or strong prior information may be desired.

- **Alternative method of modeling the unknown link function:** The support of beta cumulative density function is a unit interval. So for Poisson regression part in our model, if we choose the base link as log link, we need to transform it to the unit interval. Another possible choice also mentioned by Mallick and Gelfand (1994) is the mixture on gamma cumulative desity functions, which has support on R^+. If it works, there will be no need to transform the base link for the Poisson means.

- **Extension to multi-pathway analysis:** Our current approach for the pathway analysis are based on seperately analyzing single pathways. However, pathways share genes or have interactions, as we shown in our second topic. So constructing a similarity measure and a variance covariance structure in the Gaussian process to cover interactions of multiple pathways could be a challanging but exciting research area.

For our second topic, the Gaussian graphical model for both pathway and gene level network, we can further work on the following directions:

- **Application of other penalty functions:** Our current method is based on the LASSO penalty and derived to a adaptive LASSO type algorithm. The adaptive part, i.e. the weight on the L_1 penalty for a gene pair from different pathways has a form related to the L_1 norm of the total dependency at the pathway level. Another approach we could try is the L_2 norm, a grouped graphical LASSO type, of the total dependency at the pathway level. Danaher et al. (2013) has indicated this penalty method with ADMM algorithm (Boyd et al., 2010) has advantage in computational speed and convergency, and comaparable estimation quality for the joint estimation of gene level network for multiple trait categories. The study on this method for pathway network is of our future interest. Application and algorithm development of other penalty functions on pathway network, such as SCAD penalty which reduces the estimation bias for large signals, are also important to learn.

- **Development of criteria for evaluation of pathway level edge selection or estimation:** In our work, we have developed two criteria for the pathway level network estimation; one is the pathway connection degree bias (PCDB) and the other is the pathway disconnection degree(PDDB). They measured the bias of the estimated network for pathway connection or disconnection strength, comparing to the true net-

work. Thus they both serve for the simulations, analogize to the false positive and false negative rate. In the future research, we could develop the pathway level likelihood based criteria, for example, the pathway level BIC, or pathway level GCV, etc.

- **Extension to the joint estimation on multi-class situations:** We analyzed the healthy dogs and unhealthy dogs seperately on their gene and pathway networks. It is of interest if a joint estimation and in what situation it will benefit could be provided as Guo et al. (2011) and Boyd et al. (2010) learned in their single level network.

- **Alternative Bayesian approach:** The LASSO penalty is well known to be the same as implementing a double exponential prior on the penalized parameters. We could develop the Bayesian approach by learn how to introduce a prior at the pathway level as well. In our model, to avoid the identifiability problem, we constrain the pathway connection factor θ's as non-negative. So another possible path of the Bayesian approach is constrain this factor as an indicator of connection in $\{0, 1\}$, and then develop a stochastic search of variable selection (Geoge and McCulloch, 1993).

Acknowledgements

This study was supported by a grant from National Science Foundation (CNS-096480).

Bibliography

Ali, Z.A., Bursill, C.A., Douglas, G., McNeill, E., Papaspyridonos, M., Tatham, A.L., Bendall, J.K., Akhtar, A.M., Alp, N.J., Greaves, D.R., and Channon, K.M. (2008). CCR2-mediated anti-inflammatory effects of endothelial tetrahydrobiopterin inhibit vascular injury-induced accelerated atherosclerosis. *Circulation*, 118, S71-7

Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3, 205-238.

Bai, X., Margariti, A., Hu, Y., Sato, Y., Zeng, L., Ivetic, A., Habi, O., Mason, J.C., Wang, X., and Xu, Q. (2010). Protein kinase Cdelta deficiency accelerates neointimal lesions of mouse injured artery involving delayed reendothelialization and vasohibin-1 accumulation. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 30, 2467-74.

Bogdan, M., Doerge, R. and Ghosh, J.K. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, 167, 989-999.

Bottos, A., Destro, E., Rissone, A., Graziano, S., Cordara, G., Assenzio, B., Cera, M.R., Mascia, L., Bussolino, F., and Arese, M. (2009). The synaptic proteins neurexins and neuroligins are widely expressed in the vascular system and contribute to its functions. *Proceedings of the National Academy of Sciences*, 106, 49, 20782-87.

Broman, K.W. and Speed, T.P. (2002). A model selection approach for the identication of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society, Series B*, 64, 641-656.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends in Machine Learning*, 3, 1, 1-122.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95, 3, 759-771.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96, 270-281.

Coso, O.A., Chiariello, M., Yu, J.C., Teramoto, H., Crespo, P., Xu, N., Miki, T., and Gutkind, J.S. (1995). Rho, rac, and cdc42 GTPases regulate the assembly of multimolecular focal complexes associated with actin stress fibers, lamellipodia, and filopodia. *Cell*, 81, 7, 1137-1146.

Danaher, P., Wang, P., and Witten, D.M. (2013). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, online.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics*, 20, 18, 3583-3593.

Dempster, A.P. (1972). Covariance selection. *Biometrics*, 28, 1, 157-175.

Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion (with discussions). *Bayesian Statistics*, North-Holland, Amsterdam, 133-156.

Drton, M. and Foygel, R. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, 23, 2020-2028.

Dziuda, M.D. (2010) *Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data*, Hoboken, NJ: Wiley.

Enerson, B.E., Lin, A., Lu, B., Zhao, H., Lawton, M.P., and Floyd, E. (2006). Acute drug-induced vascular injury in beagle dogs: pathology and correlating genomic expression. *Toxicologic Pathology*, 34, 27-32.

Jeffreys, H. (1961). *The Theory of Probability*, Oxford, New York.

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. (2012). *Some advanced model selection topics for nonparametric/semiparametric models with high dimensional data*, Ph.D dissertation, Virginia Tech.

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

Gao, X., Iwai, M., Inaba, S., Tomono, Y., Kanno, H., Mogi, M., and Horiuchi, M. (2007). Attenuation of monocyte chemoattractant protein-1 expression via inhibition of nuclear factor-kappaB activity in inflammatory vascular injury. *American Journal of Hypertension*, 20, 1170-1175.

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

Goeman, J.J., van de Geer, S.A., de Kort, F., van Houwelingen, H.C., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and

Mesirovak, J.P. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20, 1, 93-99.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98, 1, 1-15.

Harris, M.A. et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32, D258-261.

Hilbe, J. M. (2009). *Logistic Regression Models*, Boca Raton, FL: Chapman & Hall/CRC.

Hilbe, J. M. (2011). *Negative Binomial Regression Extensions*, Cambridge University.

Hosack, A., Dennis, G., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4:R70.

Kaminska, B. (2005). MAPK signalling pathways as molecular targets for anti-inflammatory therapy–from molecular mechanisms to therapeutic benefits. *Biochimica et Biophysica Acta* , 1754, 253-262.

Kass, R.E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes modes). *Journal of the American Statistical Association*, 84, 717-726.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1-14.

Laud, P. and Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society Series B*, 57, 247-262.

Laufs, U., Adam, O., Strehlow, K., Wassmann, S., Konkol, C., Laufs, K., Schmidt, W., Bohm, M., and Nickenig, G. (2003). Down-regulation of rac-1 GTPase by estrogen. *The Journal of Biological Chemistry*, 278, 5956-5962.

Levina, E., Rothman, A. J., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Annals of Applied Statistics*, 2, 1, 245-263.

Li, Y., Wang, J., Santen, R., Kim, T., Park, H., Fan, P., and Yue, W. (2011). Estrogen stimulation of cell migration involves multiple signaling pathway interactions. *Endocrinology*, 151, 11, 5146-5156.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, 63, 4, 1079-1088.

Louis, F.S. and Zahradka, P. (2010). Vascular smooth muscle cell motility: From migration to invasion. *Experimental Experimental & Clinical Cardiology*, 15, 4, 175-185.

Mallat, A. and Lotersztajn, S. (2008). Endocannabinoids and liver disease. I. Endocannabinoids and their receptors in the liver. *American Journal of Physiology - Gastrointestinal and Liver Physiology*, 294, G9-G12

Mallick, B.K., and Gelfand, A.E. (1994). Generalized linear models with unknown link functions. *Biometrika*, 81, 2, 237-45.

Masood, D., Roach, C.E., Beauregard, G.K., and Khalil, A.R. (2010). Impact of sex hormone metabolism on the vascular effects of menopausal hormone therapy in cardiovascular disease. *Current Drug Metabolism*, 11, 8, 693-714.

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

Melaragno, M.G., Wuthrich, D.A., Poppa, V., Gill, D., Lindner, V., Berk, B.C., and Corson, M.A. (1998). Increased expression of Axl tyrosine kinase after vascular injury and regulation by G protein-coupled receptor agonists in rats. *Circulation Research*, 83, 697-704.

Mendelsohn, E.M. (2002) Genomic and nongenomic effects of estrogen in the vasculature.*The American Journal of Cardiology*, 90, 1, F3-F4.

Miller, M.V. and Duckles, P.S. (2008). Vascular actions of estrogens: functional implications. *Pharmacological Reviews*, 60, 2, 210-241.

Mootha, V. K., Handschin, C., Arlow, D., Xie, X., Pierre, J. S., Sihag, S., Yang, W., Altshuler, D., Puigserver, P., Patterson, N., Willy, P. J., Schulman, I. G., Heyman, R. A., Lander, E. S., and Spiegelman, B. M. (2004). Errα and Gabpa/b specify PGC-1α-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proceedings of the National Academy of Sciences*, 101, 6570-6575.

Ohta, A. and Sitkovsky M. (2001). Role of G-protein-coupled adenosine receptors in downregulation of inflammation and protection from tissue damage. *Nature*, 414, 6866, 916-920.

Olson, M.F., Ashworth, A., and Hall, A. (1995). An essential role for Rho, Rac, and Cdc42 GTPases in cell cycle progression through G1. *Science*, 269, 5228, 1270-1272.

Pankova, K., Rosel, D., Novotny, M., and Brabek, J. (2010). The molecular mechanisms of transition between mesenchymal and amoeboid invasiveness in tumor cells. *Cellular and Molecular Life Sciences*, 67, 1, 63-71.

Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E., and Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, 22, 2028-2036.

Pettit, L.I., and Young, K.D.S. (1990). Measuring the effect of observation on Bayes factors. *Biometrika*, 77, 455-466.

Rajagopalan, D., and Agarwal, P. (2005). Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, 21, 6, 788-93.

Rosenblatt, A.E., Garcia, M.I., Lyons, L., Xie, Y., Maiorino, C., Desire, L., Slingerland, J., and Burnstein, K.L. (2011). Inhibition of the Rho GTPase, Rac1, decreases estrogen receptor levels and is a novel therapeutic strategy in breast cancer. *Endocrine-Related Cancer*, 18, 2, 207-219.

Ruusalepp, A., Yan, Z.Q., Carlsen, H., Czibik, G., Hansson, G.K., Moskaug, J.Ø., Blomhoff, R., and Valen, G. (2006). Gene deletion of NF-kappaB p105 enhances neointima formation in a mouse model of carotid artery injury. *Cardiovascular Drugs and Therapy*, 20, 103-11.

Semenkovich, C.F. (2004). Fatty acid metabolism and vascular disease. *Trends in Cardiovascular Medicine*, 14, 72-6.

Siegmund, D. (2004). Model selection in irregular problems: Application to mapping quantitative trait loci. *Biometrika*, 91, 785-800.

Simpson, E.R., Misso, M., Hewitt, K.N., Hill, R.A., Boon, W.C., Jones, M.E., Kovacic, A., Zhou, J., and Clyne, C.D. (2005). Estrogen–the good, the bad, and the unexpected. *Endocrine Reviews*, 26, 3, 322-330.

Smiley, A.D. and Khalil, A.R. (2009). Estrogenic compounds, estrogen receptors and vascular cell signaling in the aging blood vessels. *Current Medicinal Chemistry*, 16, 15, 1863-87.

Somjen, D., Kohen, F., Jaffe, A., Amir-Zaltsman, Y., Knoll, E., and Stern, N. (1998). Effects of gonadal steroids and their antagonists on DNA synthesis in human vascular cells. *Hypertension*, 32, 39-45.

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. *Annals of Applied Statistics*, 5, 1978-2002.

Subramanian, A., Tamayo, P., Mootha, V.K. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 43, 15545-15550.

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

Vanhoutte, P.M. (2010). Regeneration of the endothelium in vascular injury. *Cardiovascular Drugs and Therapy*, 24, 299-303.

Vecchione, C., Aretini, A., Marino, G., Bettarini, U., Poulet, R., Maffei, A., Sbroggió, M., Pastore, L., Gentile, M.T., Notte, A., Iorio, L., Hirsch, E., Tarone, G., and Lembo, G. (2006). Selective Rac-1 inhibition protects from diabetes-induced vascular injury. *Circulation Research*, 98, 218-25.

Wang, Y., Ma, Y., and Carroll, R.J. (2009). Variance estimation in the analysis of microarray data. *Journal of the Royal Statistical Society, Series B*, 71, 2, 425-445.

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

Yuan, M. and Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 94, 1, 19-35.

Zhou, N., and Zhu, J. (2007). Group Variable Selection via a Hierarchical Lasso and Its Oracle Property. *Arxiv Preprint ArXiv*. 1006.2871.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 476, 1418-1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36, 4, 1509-1533.

Appendix A

Newton Raphson method

For the Laplace approximation, we calculate the maxima of $h = \log L(r_1, r_2|\omega_1, \omega_2)$ using Newton-Raphson method. In this section we sketch the first and second derivatives of h in order to use Newton-Raphson algorithm.

We first define some notations. Let r_{li} denote the ith observation's predictor in γ_l, $l = 1, 2$ and $i = 1, \ldots, n$. Since observations are conditionally independent, we ignore the index i for simplicity, i.e. r_l represents any element in the vector γ_l. Using the following notations,

$$\begin{aligned} u_l &= \frac{\exp(r_l)}{1+\exp(r_l)}, l = 1, 2, \\ \pi &= \sum_{l=1}^{3} \omega_{1j} IB(u_1, c_{1j}, d_{1j}), \\ \pi_2 &= \sum_{l=1}^{3} \omega_{2j} IB(u_2, c_{2j}, d_{2j}), \\ \lambda &= \frac{\pi_2}{1-\pi_2}, \end{aligned}$$

the first and second derivatives of π with respect to r_1 can be expressed as follows,

$$\begin{aligned}\frac{\partial \pi}{\partial r_1} &= \pi u_1(1-u_1),\\ \frac{\partial^2 \pi}{\partial r_1^2} &= \sum_{l=1}^{3} \omega_{1j} IB(u_1, c_{1j}, d_{1j})(c_{1j} - (c_{1j}+d_{1j})u_1).\end{aligned}$$

We then calculate these first and derivatives for all n observations and obtain $\partial \pi_{1i}/\partial r_{1i}$ and $\partial^2 \pi_{1i}/\partial r_{1i}^2$, $i = 1, \ldots, n$. Similar forms can be derived for $\partial \pi_2/\partial r_2$ and $\partial^2 \pi_2/\partial r_2^2$. Using these calculations, we can obtain the first and second derivatives of λ with respect to r_2,

$$\begin{aligned}\frac{\partial \lambda}{\partial r_2} &= \frac{1}{(1-\pi_2)^2}\frac{\partial \pi_2}{\partial r_2},\\ \frac{\partial^2 \lambda}{\partial r_2^2} &= -\frac{2}{(1-\pi_2)^3}(\frac{\partial \pi_2}{\partial r_2})^2 + \frac{1}{(1-\pi_2)^2}\frac{\partial^2 \pi_2}{\partial r_2^2}.\end{aligned}$$

Then we calculate these first and derivatives for all n observations and have $\partial \lambda_i/\partial r_{2i}$ and $\partial^2 \lambda_i/\partial r_{2i}^2$.

Further define some more notations,

$$\begin{aligned}\delta_1 &= \frac{1-\exp(-\lambda)}{\pi + (1-\pi)\exp(-\lambda)}\mathbf{1}_{y=0} - \frac{1}{1-\pi}\mathbf{1}_{y>0},\\ \delta_2 &= -\frac{(1-\pi)\exp(-\lambda)}{\pi + (1-\pi)\exp(-\lambda)}\mathbf{1}_{y=0} + (\frac{y}{\lambda} - 1)\mathbf{1}_{y>0},\\ \delta_3 &= -\delta_1^2,\\ \delta_4 &= \frac{\pi(1-\pi)\exp(-\lambda)}{\{\pi + (1-\pi)\exp(-\lambda)\}^2}\mathbf{1}_{y=0} - \frac{y}{\lambda^2}\mathbf{1}_{y>0},\\ \delta_5 &= \frac{\exp(-\lambda)}{\{\pi + (1-\pi)\exp(-\lambda)\}^2}\mathbf{1}_{y=0}.\end{aligned}$$

We also calculate $\delta_1, \ldots, \delta_5$ for all n observations and obtain $\delta_{1i}, \ldots, \delta_{5i}$.

Let $V(\cdot)$ be the vector composed of n elements in "()". Denote $B_l = b_l + \gamma_l' K^{-1} \gamma_l, l =$

$1, 2$. Then we can calculate the first derivatives of h with respect to γ_1 and γ_2 as follow,

$$\begin{aligned} g_1 = \frac{\partial h}{\partial \gamma_1} &= V(\delta_{11}\frac{\partial \pi_1}{\partial r_{11}}, \ldots, \delta_{1i}\frac{\partial \pi_i}{\partial r_{1i}}, \ldots, \delta_{1n}\frac{\partial \pi_n}{\partial r_{1n}}) + \frac{a_1 + N/2}{B_1}K^{-1}\gamma_1, \\ g_2 = \frac{\partial h}{\partial \gamma_2} &= V(\delta_{21}\frac{\partial \lambda_1}{\partial r_{21}}, \ldots, \delta_{2i}\frac{\partial \lambda_i}{\partial r_{2i}}, \ldots, \delta_{2n}\frac{\partial \lambda_n}{\partial r_{2n}}) + \frac{a_2 + N/2}{B_2}K^{-1}\gamma_2. \end{aligned}$$

By letting $D(\cdot)$ be the diagonal matrix with the n elements in "()" on the diagonal, we also obtain the second derivatives respect to γ_1 and γ_2,

$$\begin{aligned} H_{11} &= D\Big\{\delta_{11}\frac{\partial^2 \pi_1}{\partial r_{11}^2} + \delta_{31}(\frac{\partial \pi_1}{\partial r_{11}})^2, \ldots, \delta_{1n}\frac{\partial^2 \pi_n}{\partial r_{1n}^2} + \delta_{3n}(\frac{\partial \pi_n}{\partial r_{1n}})^2\Big\} + \frac{a_1 + N/2}{B_1^2}(B_1K^{-1} - K^{-1}\gamma_1\gamma_1'K^{-1}), \\ H_{12} &= H_{21} = D\Big(\delta_{51}\frac{\partial \pi_1}{\partial r_{11}}\frac{\partial \lambda_1}{\partial r_{21}}, \ldots, \delta_{5n}\frac{\partial \pi_n}{\partial r_{1n}}\frac{\partial \lambda_n}{\partial r_{2n}}\Big), \\ H_{22} &= D\Big\{\delta_{21}\frac{\partial^2 \lambda_1}{\partial r_{21}^2} + \delta_{41}(\frac{\partial \lambda_1}{\partial r_{21}})^2, \ldots, \delta_{2n}\frac{\partial^2 \lambda_n}{\partial r_{2n}^2} + \delta_{4n}(\frac{\partial \lambda_n}{\partial r_{2n}})^2\Big\} + \frac{a_2 + N/2}{B_2^2}(B_2K^{-1} - K^{-1}\gamma_2\gamma_2'K^{-1}). \end{aligned}$$

Using $g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$, $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{pmatrix}$, and initial values (γ_1^0, γ_2^0), the Newton Raphson method can be conducted iteratively in order to obtain the maxima $(\hat{\gamma}_1, \hat{\gamma}_2)$ of h. Finally we can obtain $h(\hat{\gamma}_1, \hat{\gamma}_2)$ and $\Sigma = H^{-1}(\hat{\gamma}_1, \hat{\gamma}_2)$ for the Laplace approximation of the integral in Section 2.3.5.

Appendix B

Proof of Asymptotic Result for MGGM

B.1 Proof of Lemma 1

We need to show that if $(\hat{\Theta}^{**}, \{\hat{\Gamma}^{**}_{kk'}\}, \{\hat{\Gamma}^{**}_{kk}\})$ is a local minimizer of **Q1**, then there exist a local minimizer of **Q2**, denote as $(\hat{\Theta}^{*}, \{\hat{\Gamma}^{*}_{kk'}\}, \{\hat{\Gamma}^{*}_{kk}\})$, such that $\hat{\theta}^{**}_{kk'} \cdot \hat{\Gamma}^{**}_{kk'} = \hat{\theta}^{*}_{kk'} \cdot \hat{\Gamma}^{*}_{kk'}$. And vise versa.

Let $\eta = \eta_1\eta_2$, we have:

$$Q_1(\Theta, \{\Gamma_{kk'}\}, \{\Gamma_{kk}\}, \eta_1, \eta_2) = Q_2(\eta_1\Theta, \{\frac{1}{\eta_1}\Gamma_{kk'}\}, \{\Gamma_{kk}\}, \eta).$$

Let $(\hat{\Theta}^{**}, \{\hat{\Gamma}^{**}_{kk'}\}, \{\hat{\Gamma}^{**}_{kk}\})$ be a local minimizer of **Q1**, then, $\exists\delta > 0, \forall(\Theta', \{\Gamma'_{kk'}\})$ satisfying $||\Theta' - \hat{\Theta}^{**}||_1 + \sum_{kk'} ||\Gamma'_{kk'} - \hat{\Gamma}^{**}_{kk'}||_1 < \delta$, we have:

$$Q_1(\Theta', \{\Gamma'_{kk'}\}, \{\Gamma_{kk}\}, \eta_1, \eta_2) \geq Q_1(\hat{\Theta}^{**}, \{\hat{\Gamma}^{**}_{kk'}\}, \{\Gamma_{kk}\}, \eta_1, \eta_2).$$

Choose δ' such that $\delta'/\{\min(\eta_1, 1/\eta_1)\} \leq \delta$,
$\forall(\Theta^{''}, \{\Gamma^{''}_{kk'}\})$ satisfying $||\Theta^{''} - \hat{\Theta}^*||_1 + \sum_{kk'} ||\Gamma^{''}_{kk'} - \hat{\Gamma}^*_{kk'}||_1 < \delta'$, we have:

$$\begin{aligned}
||\frac{\Theta^{''}}{\eta_1} - \hat{\Theta}^{**}|| + \sum_{kk'} ||\eta_1\Gamma^{''}_{kk'} - \hat{\Gamma}^{**}_{kk'}|| &\leq \frac{\eta_1||\frac{\Theta^{''}}{\eta_1} - \hat{\Theta}^{**}|| + (1/\eta_1)\sum_{kk'} ||\eta_1\Gamma^{''}_{kk'} - \hat{\Gamma}^{**}_{kk'}||}{\min(\eta_1, 1/\eta_1)} \\
&= \frac{||\Theta^{''} - \eta_1\hat{\Theta}^{**}|| + \sum_{kk'} ||\Gamma^{''}_{kk'} - (1/\eta_1)\hat{\Gamma}^{**}_{kk'}||}{\min(\eta_1, 1/\eta_1)} \\
&< \frac{\delta'}{\min(\eta_1, 1/\eta_1)} \\
&= \delta.
\end{aligned}$$

$$\therefore Q_1(\frac{\Theta^{''}}{\eta_1}, \{\eta_1\Gamma_{kk'}\}, \{\Gamma_{kk}\}, \eta_1, \eta_2) \geq Q_1(\hat{\Theta}^{**}, \{\hat{\Gamma}^{**}_{kk'}\}, \{\Gamma_{kk}\}, \eta_1, \eta_2),$$

$$\begin{aligned}
\therefore Q_2(\Theta^{''}, \{\Gamma_{kk'}\}, \{\Gamma_{kk}\}, \eta_1, \eta_2) &\geq Q_2(\eta_1\hat{\Theta}^{**}, \{\frac{1}{\eta_1}\hat{\Gamma}^{**}_{kk'}\}, \{\Gamma_{kk}\}, \eta_1, \eta_2) \\
&= Q_2(\hat{\Theta}^*, \{\hat{\Gamma}^*_{kk'}\}, \{\Gamma_{kk}\}, \eta_1, \eta_2).
\end{aligned}$$

Similarly, we can prove for the reverse side.

B.2 Proof of Lemma 2

We need to show that suppose $\hat{\Omega}_{kk'}$ is a local minimizer of **Q3**, then there exists a local minimizer of **Q2**, denote as $(\hat{\Theta}, \{\hat{\Gamma}_{kk'}\}, \{\hat{\Gamma}_{kk}\})$, such that $\hat{\Omega}_{kk'} = \hat{\theta}_{kk'}\hat{\Gamma}_{kk'}$.

Let $(\{\hat{\Omega}_{kk'}\}, \{\hat{\Gamma}^{-}_{kk'}\})$ is a local minimizer of **Q3**, and define

$$\hat{\theta}_{kk'} = \sqrt{\eta|\hat{\Omega}_{kk'}|_1},$$
$$|\hat{\Gamma}_{kk'}|_1 = \frac{|\hat{\Omega}_{kk'}|_1}{\sqrt{\eta|\hat{\Omega}_{kk'}|_1}},$$

$\exists \varepsilon > 0, \forall \{\Omega_{kk'}\}$ satisfying $\sum_{k \neq k'} |\Omega_{kk'} - \hat{\Omega}_{kk'}|_1 \leq \varepsilon$, we have:

$$Q_3(\{\Omega_{kk'}\}) \geq Q_3(\{\hat{\Omega}_{kk'}\}).$$

Let $\mathcal{F} = \{(\Theta, \{\Gamma_{kk'}\}) : |\Delta\Theta|_1 + \sum_{k \neq k'} |\Delta\Gamma_{kk'}|_1 \leq \varepsilon'\}$, where $\Delta\theta_{kk'} = \theta_{kk'} - \hat{\theta}_{kk'}$, $\Delta\Gamma_{kk'} = \Gamma_{kk'} - \hat{\Gamma}_{kk'}$, ε' is some constant satisfying $0 \leq \varepsilon' \leq (-U + \sqrt{U^2 + 4\varepsilon})/2$, and $U = \max(1, 4/\lambda)\max_{k \neq k'} \lambda\sqrt{|\Omega_{kk'}|_1}$.

Then $\forall(\Theta, \{\Gamma_{kk'}\}) \in \mathcal{F}$, we have:

$$
\begin{aligned}
|\Omega_{kk'} - \hat{\Omega}_{kk'}|_1 &= \sum_{j,j'} |\theta_{kk'}\gamma_{jj'}^{kk'} - \hat{\theta}_{kk'}\hat{\gamma}_{jj'}^{kk'}| \\
&= \sum_{j,j'} |(\hat{\theta}_{kk'} + \Delta\theta_{kk'})(\hat{\gamma}_{jj'}^{kk'} + \Delta\gamma_{jj'}^{kk'}) - \hat{\theta}_{kk'}\hat{\gamma}_{jj'}^{kk'}| \\
&\leq |\Delta\theta_{kk'}||\hat{\Gamma}_{kk'}|_1 + |\hat{\theta}_{kk'}||\Delta\Gamma_{kk'}|_1 + |\Delta\theta_{kk'}||\Delta\Gamma_{kk'}|_1. \\
\therefore \sum_{k\neq k'} |\Delta\Omega_{kk'}| &\leq (\sum_{k\neq k'} |\Delta\theta_{kk'}|)(\max_{k\neq k'} |\hat{\Gamma}_{kk'}|_1) \\
&\quad + (\max_{k\neq k'} |\hat{\theta}_{kk'}|_1)(\sum_{k\neq k'} |\Delta\Gamma_{kk'}|_1) \\
&\quad + (\sum_{k\neq k'} |\Delta\theta_{kk'}|)(\sum_{k\neq k'} |\Delta\Gamma_{kk'}|_1) \\
&\leq \max_{k\neq k'}(|\hat{\theta}_{kk'}|_1 + |\hat{\Gamma}_{kk'}|_1)(|\Delta\Theta|_1 + \sum_{k\neq k'} |\Delta\Gamma_{kk'}|_1) + (|\Delta\Theta|_1 + \sum_{k\neq k'} |\Delta\Gamma_{kk'}|_1)^2. \\
\because \max_{k\neq k'}(|\hat{\theta}_{kk'}|_1 + |\hat{\Gamma}_{kk'}|_1) &\leq \max(1, \frac{1}{\eta})\max_{k\neq k'}(|\hat{\theta}_{kk'}|_1 + \eta|\hat{\Gamma}_{kk'}|_1) \\
&= \max(1, \frac{4}{\lambda^2})\max_{k\neq k'}(\lambda\sqrt{|\hat{\Omega}_{kk'}|_1}) \\
&= U, \\
\therefore \sum_{k\neq k'} |\Delta\Omega_{kk'}| &\leq U\varepsilon' + \varepsilon'^2 \leq \varepsilon, \\
\therefore Q_2(\Theta, \{\Gamma_{kk'}\}) &= Q_3(\{\Omega_{kk'}\}) \\
&\geq Q_3(\{\hat{\Omega}_{kk'}\}) = Q_2(\hat{\Theta}, \{\hat{\Gamma}_{kk'}\}).
\end{aligned}
$$

B.3 The Proof of Theorem 1

We write $\Omega = \{\Omega_{kk'}\}$, $\Omega_0 = \{\Omega_{0,kk'}\}$, $\Delta = \{\Delta_{kk'}\}$, where$\Delta_{kk'} = \Omega_{kk'} - \Omega_{0,kk'}$, and let $G(\Delta) = Q_3(\Omega_0 + \Delta) - Q_3(\Delta)$. If we take a closed bounded convex set $\mathcal{A}$ which contains 0,

and show that G is strictly positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that G has a local minimum inside $\mathcal{A}$, since G is continuous and $G(0) = 0$. Specifically, we'll define $\mathcal{A} = \{\Delta : \sum_{kk'} ||\Delta_{kk'}||_F \leq Mr_n\}$, with boundary $\partial\mathcal{A} = \{\Delta : \sum_{kk'} ||\Delta_{kk'}||_F = Mr_n\}$, where M is a positive constant and $r_n = \sqrt{(p+q)(\log p)/n}$.

We can write $G(\Delta) = I_1 + I_2 + I_3 + I_4$, where

$$
\begin{aligned}
I_1 &= tr\{(S - \Sigma_0)\Delta\} \\
I_2 &= \tilde{\Delta}^T\{\int_0^1 (1-v)(\Omega_0 + v\Delta)^{-1} \bigotimes (\Omega_0 + v\Delta)^{-1} dv\}\tilde{\Delta} \\
I_3 &= \lambda \sum_{kk'} \sqrt{|\Delta_{T^C,kk'}|_1} + \lambda \sum_{kk'} \left(\sqrt{|\Omega_{T,kk'}|_1} - \sqrt{|\Omega_{0,T,kk'}|_1} \right) \\
&= I_{3,1} + I_{3,2} \\
I_4 &= \eta_3 \sum_{k=1}^{K} \left(|\Gamma^-_{kk}| - \Gamma^-_{0,kk} \right)
\end{aligned}
$$

For I_1, we have:

$$
I_1 \leq C_1 \sqrt{\frac{\log p}{n}} \sum_{kk'} |\Delta^-_{kk'}|_1 + C_2 \sqrt{\frac{p \log p}{n}} \sum_{kk'} ||\Delta^+_{kk'}||_F = I_{1,1} + I_{1,2}
$$

where

$$
\begin{aligned}
I_{1,1} &= C_1 \sqrt{\frac{\log p}{n}} \sum_{kk'} |\Delta^-_{T,kk'}|_1 + C_2 \sqrt{\frac{p \log p}{n}} \sum_{kk'} ||\Delta^+_{kk'}||_F \\
I_{1,2} &= C_1 \sqrt{\frac{\log p}{n}} \sum_{kk'} |\Delta^-_{T^C,kk'}|_1
\end{aligned}
$$

Since $|\Delta^-_{T,kk'}|_1 \leq \sqrt{q_k k'}||\Delta^-_{T,kk'}||_F$, we have:

$$\begin{aligned}|I_{1,1}| &\leq C_1\sqrt{\frac{q\log p}{n}}\sum_{kk'}||\Delta^-_{T,kk'}||_F + C_2\sqrt{\frac{p\log p}{n}}\sum_{kk'}||\Delta^+_{kk'}||_F \\ &\leq (C_1+C_2)\sqrt{\frac{(p+q)\log p}{n}}\sum_{kk'}||\Delta_{kk'}||_F = M(C_1+C_2)\frac{(p+q)\log p}{n}\end{aligned}$$

on the boundary $\partial\mathcal{A}$.

For r_n small enough, we have $I_{3,1} \geq \lambda\sum_{kk'}|\Delta^-_{T^C,kk'}|_1$, $I_{1,2}$ is dominated by the positive term $I_{3,1}$:

$$\begin{aligned}I_{3,1}+I_{1,2} &\geq \lambda\sum_{kk'}|\Delta^-_{T^C,kk'}|_1 - C_1\sqrt{\frac{\log p}{n}}\sum_{kk'}|\Delta^-_{T^C,kk'}|_1 \\ &\geq (\Lambda_1 - C1)\sqrt{\frac{\log p}{n}}\sum_{kk'}|\Delta^-_{T^C,kk'}|_1\end{aligned}$$

For I_2, we have:

$$I_2 \geq \frac{1}{4\tau_2^2}\sum_{kk'}||\Delta_{kk'}||_F^2 \geq \frac{M^2}{8\tau_2^2}\frac{(p+q)\log p}{n}$$

Finally, for $I_{3,2}$, we have:

$$\begin{aligned}|I_{3,2}| &\leq \lambda\sum_{kk'}|\left(\sqrt{|\Omega_{T,kk'}|_1} - \sqrt{|\Omega_{0,T,kk'}|_1}\right)| \\ &\leq \lambda\sum_{kk'}\frac{||\Omega_{T,kk'}|_1 - |\Omega_{0,T,kk'}|_1|}{\sqrt{|\Omega_{T,kk'}|_1}+\sqrt{|\Omega_{0,T,kk'}|_1}} \\ &\leq \frac{\lambda}{\sqrt{\tau_3}}\sum_{kk'}|\Omega_{kk'}-\Omega_{0,kk'}| \\ &\leq \frac{\lambda}{\sqrt{\tau_3}}\sqrt{q}\sum_{kk'}||\Delta_{kk'}||_F \\ &\leq \frac{M\Lambda_2}{\sqrt{\tau_3}}\frac{(p+q)(\log p)}{n}\end{aligned}$$

Since $I_2 > 0$, $I_{3,1} + I_{1,2} > 0$, we have:

$$\begin{aligned} G(\Delta) &\geq I_2 - I_{1,1} - I_{3,2} \\ &\geq \frac{M^2}{8\tau_2^2}\frac{(p+q)\log p}{n} - M(C_1+C_2)\frac{(p+q)\log p}{n} - \frac{M\Lambda_2}{\sqrt{\tau_3}}\frac{(p+q)(\log p)}{n} \\ &= M^2\frac{(p+q)(\log p)}{n}\left(\frac{1}{8\tau_2^2} - \frac{C_1+C_2+\lambda_2/\sqrt{\tau_3}}{M}\right) \end{aligned}$$

So, for M sufficiently large, $G(\Delta) > 0$ for any $\Delta \in \partial\mathcal{A}$.

B.4 The proof of Theorem 2

It's sufficient to show that $\forall(j, j', k, k') \in T^C$, the derivative $\partial Q_3/\partial\omega_{jj'}^{kk'}$ at $\hat{\omega}_{jj'}^{kk'}$ has the same sign as $\hat{\omega}_{jj'}^{kk'}$ with probability tending to 1. This is because suppose for some $(j, j', k, k') \in T^C$, the estimates $\hat{\omega}_{jj'}^{kk'} \neq 0$, without loss of generality, suppose $\hat{\omega}_{jj'}^{kk'} > 0$, then $\exists\xi > 0$ such that $\hat{\omega}_{jj'}^{kk'} - \xi > 0$. Since $\hat{\Omega}$ is a local minimizer, $\partial Q_3/\partial\omega_{jj'}^{kk'} < 0$ at $\hat{\omega}_{jj'}^{kk'} - \xi$ when ξ small enough, contracting the claim that $\partial Q_3/\partial\omega_{jj'}^{kk'}$ at $\hat{\omega}_{jj'}^{kk'}$ has the same sign as $\hat{\omega}_{jj'}^{kk'}$.

To show this, we get the derivative of the objective function as:

$$\frac{\partial Q_3}{\partial\omega_{jj'}^{kk'}} = 2\{W_1 + W_2 sgn(\omega_{jj'}^{kk'})\}$$

where $W_1 = \hat{\sigma}_{jj'}^{kk'} - \sigma_{jj'}^{kk'}$, and $W_2 = \lambda/\sqrt{|\Omega_{kk'}|_1}$. Refer to Lam and Fan (2009), one can show $W_1 = O[\{(\log p)/n\}^{1/2} + \eta_n^{1/2}]$. On the other hand, by Theorem 1, we have $|\Omega_{kk'} - \Omega_{0,kk'}|_1 \leq |\Omega_{kk'} - \Omega_{0,kk'}|_F = O(\eta_n) = \circ(1)$. Then for any $\varepsilon > 0$ and large enough n, we have $|\Omega_{kk'}| \leq |\Omega_{0,kk'}| + \varepsilon$. Then we have $|W_2| \geq \lambda/(1 + |\Omega_{0,kk'}|)$. By assumption, $\{(\log p)/n\}^{1/2} + \eta_n^{1/2} = O(\lambda)$,and thus the term W_2 dominate W_1 in the derivative of the

objective function. So we have:

$$sgn(\frac{\partial Q_3}{\partial \omega_{jj'}^{kk'}}|_{\omega_{jj'}^{kk'}=\hat{\omega}_{jj'}^{kk'}}) = sgn(\hat{\omega}_{jj'}^{kk'}).$$