Bayesian Non-Linear Methods for Survival Analysis and Structural Equation Models

---

A Thesis presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Zhenyu Wang

Dr. Sounak Chakraborty, Co-Adviser

Dr. (Tony) Jianguo Sun, Co-Adviser

JULY 2014

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

Bayesian Non-Linear Methods for Survival Analysis
and Structural Equation Models

presented by Zhenyu Wang,
a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Sounak Chakraborty

Dr. (Tony) Jianguo Sun

Dr. Chong (Zhuoqiong) He

Dr. Lori Thombs

Dr. Bimal Ray

ACKNOWLEDGMENTS

I would love to express my appreciation to those who have guided and supported me throughout the research process and provided assistance for my venture.

I would first like to thank my co-adviser, Dr. Sounak Chakraborty and co-adviser, Dr. Tony Sun for their exceptional guidance, thoughtful supervision, and tireless encouragement. Without their support and persistent help this dissertation would not have been possible.

I would also like to thank Dr. Lori Thombs who supported and leaded me for working as a research assistant at the Social Science Statistics Center and participate in my committee.

I would like to show my gratitude to the rest of my committee members, Dr. Chong (Zhuoqiong) He and Dr. Bimal Ray for their time, encouraging words, thoughtful criticism, and attention during busy semesters.

I would like to thank Dr. Larry Ries, who helped and supported me for teaching as a graduate instructor. I would also like to thank all the professors, who throughout my educational career have supported and encouraged me to believe in my abilities. They have directed me through various situations, allowing me to reach this accomplishment. I would also like to show my gratitude to my colleagues for the stimulating discussions, for teamwork and encouragement, and for all the fun we have had in the last five years.

Finally, I would like to thank my family, especially my mother, Ailing Li, who is always supporting me and encouraging me with her best wishes.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

High dimensional data are more common nowadays, because the collection of such data becomes larger and more complex due to the technology advance of the computer science, biology, etc. The analysis of high dimensional data is different from traditional data analysis, and variable selection for high dimensional data becomes very challenging. Structural equation modeling (SEM) analyzes the relationship between manifest variables and latent variables. The structural equation focuses on analyzing the relationship between latent variables. New proposed methods of these topics are discussed in the dissertation.

In the first chapter, we review the basic concept of survival analysis, SEM, and current method of variable selection in those two scenarios. We also introduce the available software package for current methods and relevant data set.

In the second chapter, we develop a Bayesian kernel machine model with incorporating existing information on pathways and gene networks in the analysis of DNA microarray data. Each pathway is modeled nonparametrically using reproducing kernel Hilbert space. The pathways and the genes are selected via assigning mixture priors on the pathway indicator variable and the gene indicator variable. This approach helped us in flexible modeling of the pathway effects, which can capture both linear and non-linear effect. Moreover, the model can also pinpoint the important pathways and the important active genes within each pathway. We have also developed an efficient Markov Chain Monte Carlo (MCMC) algorithm to fit our model. We used simulations and a real data analysis, [van 't Veer et al., 2002] breast cancer microarray data, to illustrate the proposed method.

In the third chapter, we extend the idea of semiparametric structural equation model where the nonlinear functional relationships are approximated using basis expansions [Guo et al., 2012]. Many basis expansion methods, including cubic splines, are known to induce correlations. In this chapter we compare standard Lasso, Fused Lasso and Elastic Net to account for correlations in both the covariate and basis expansions. To illustrate the usefulness of the proposed methods, a simulation study and a real data study have been performed. The semiparametric structural equation models based on Bayesian fused Lasso and Bayesian elastic-net outperform the Bayesian Lasso model.

In the fourth chapter, we apply Bayesian Graph Laplacian Model, developed by [Liu et al., 2014] and generalized the graph Laplacian allowing both positively and negatively correlated variable, to analyze gene expression data from Michigan prostate cancer study [Dhanasekaran et al., 2001]. We find out the underlie gene network and interaction related to prostate cancer and discuss the possible extensions for Bayesian Graph Laplacian Model, including analyzing multiple pathways simultaneously and pathways selection, right censored data as response variable and binomial or multinomial data as response variable.

Chapter 1

Introduction

Survival analysis focus on analyzing time to events such as death, disease occurrence, and malfunction in mechanical system. The event can be referred to as the failure.

Suppose we analyze the data consisting of time to the occurrence of certain type of cancer. It is possible that some of the patients have no occurrence at the end of the study. As a result, the exact failure times of such patients are unknown, but they are only unknown to be greater than certain amount of time. This feature is referred to as censoring in survival analysis.

The gene expression microarray data contain the information for thousands of genes. Oncologist have been trying to identify genes related to different cancers. Using microarray data and survival time for the patients to identify important genes presents a challenge in data analysis. This chapter contains a literature review on the analysis of right-censored survival data and the variable selection in high dimensional situation.

In psychology, latent variables represent the variables which cannot be measured

directly. Structural Equation Modeling (SEM) analyzes the relationship between latent variables and manifest variables. In this chapter, we present a literature review on SEM.

We review general concept of survival analysis and survival models under frequentist perspective in section 1.1; in section 1.2, we review parametric and semiparametric Bayesian model analyzing survival data; section 1.3 covers the current methods on variable selection in high dimensional data and discusses the difficulties and challenges of the current methods; we introduce the concept of genetic pathway in section 1.4; section 1.5 contains the introduction for Structural equation modeling; we discuss the motivation and outline of this thesis in section 1.6; and section 1.7 lists the available software package and relevant data set.

1.1 Survival Analysis Under Frequentist Perspective

Let T be a non-negative continuous random variable denoting the failure time of a subject. The probability of a subject surviving beyond a specific time t is given by the survival function, defined as

$$S(t) = P(T > t), \tag{1.1}$$

where $S(t)$ is a monotonically decreasing, right-continuous, $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$ function.

The hazard function, $\lambda(t)$, is the instantaneous rate at which failures occur given

the condition that subjects survive at the time t or later. It is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{1.2}$$

The probability density function of T is $f(t) = -dS(t)/dt$, where $t \in [0, +\infty)$, therefore (1.2) follows that

$$\lambda(t) = f(t)/S(t) = -d \log S(t)/dt. \tag{1.3}$$

Given $S(0) = 1$, by integrating t from both sides of (1.3) we get

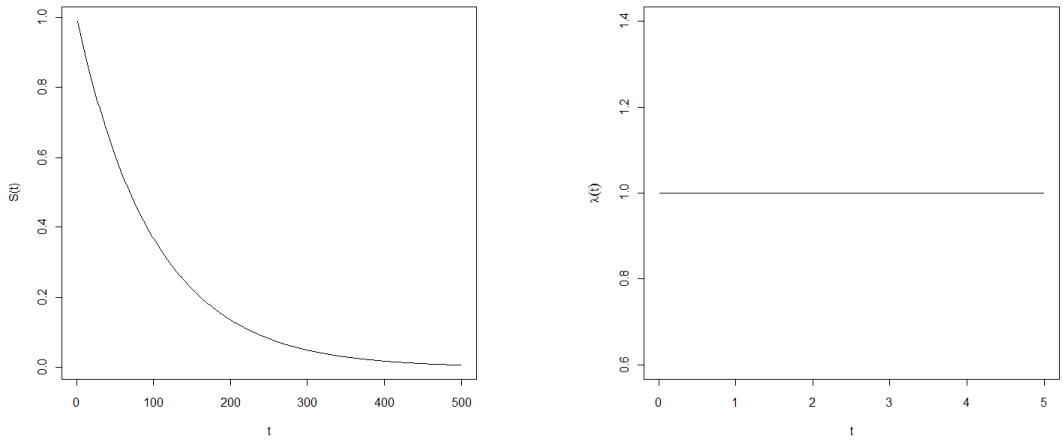
$$S(t) = \exp\{-\int_0^t \lambda(s)ds\} = \exp\{-\Lambda(t)\}, \tag{1.4}$$

where$\Lambda(t)$ is called cumulative hazard function and $\Lambda(t) = \int_0^t \lambda(s)ds$. Taking derivative with respect to t in (1.4), we obtain

$$s(t) = \lambda(t) \exp\{-\Lambda(t)\}. \tag{1.5}$$

We use log logistic probability density function to illustrate the survival function and responding hazard function in figure (1.1).

One special feature of the survival data is known as censoring. Because of the time limit, cost concern or incidence related to experimental subjects, the investigators terminate the research before all subjects realize their event of interest or some of the subjects leave the research before research ends. As a result, survival times from some of the subjects are longer than some certain values. It is called right censored when the survival time of a subject exceeds certain censoring time, C_r, and left censored



(a) Survival function (b) Hazard function

Figure 1.1: An example of hazard function and survival function

when the survival time is only know to be less than a censoring time, C_l. Interval censoring occurs when the precise survival time is unknown, but it is within a known interval, (C_l,C_r).

1.1.1 Proportional Hazards Regression

Proportional hazards Regression model is one of the regression models for survival data. According to [Cox, 1972], for a subject with covariate vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$, the hazard rate at time t can be expressed as:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{1.6}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$ is a $p \times 1$ vector of regression parameters corresponding to $\boldsymbol{x}$, and $\lambda_0(\cdot)$ is an arbitrary unknown baseline hazard function. With fixed covariate, the ratio of hazards between each subject is constant over time.

The survival function (1.4) corresponding to (1.6) is

$$S(t|\boldsymbol{x}) = \exp\{-\exp(\boldsymbol{x}'\boldsymbol{\beta}) \int_0^t \lambda_0(\mu)d\mu\}, \tag{1.7}$$

and the density function of T corresponding to (1.6) is

$$f(t|\boldsymbol{x}) = \lambda_0(t)exp(\boldsymbol{x}'\boldsymbol{\beta}) \exp\{-\exp(\boldsymbol{x}'\boldsymbol{\beta}) \int_0^t \lambda_0(\mu)d\mu\}. \tag{1.8}$$

There are two important extension of the proportional hazards regression model: (i) stratified Cox model and (ii) time-dependent covariate model. In stratified Cox model, if $\lambda_0(\cdot)$ is arbitrary and there are J strata in the population, the hazard function for jth stratum is

$$\lambda_j(t|\boldsymbol{x}) = \lambda_{0j}(t) \exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{1.9}$$

for $j = 1, \cdots, J$, where $\lambda_{0j}(t)$ is the corresponding baseline hazard function for the jth stratum.

When the covariates are time-dependent. Cox model can be easily extended to time-variant covariates:

$$\lambda(t|\boldsymbol{x}(t)) = \lambda_0(t)exp(\boldsymbol{x}(t)'\boldsymbol{\beta}). \tag{1.10}$$

When $n > p$ and only one subject fails at each time, maximizing the partial

likelihood is used to find the estimate of $\boldsymbol{\beta}$ [Cox, 1975],

$$L(\boldsymbol{\beta}) = \prod_{k \in \mathcal{D}} \frac{\exp(\boldsymbol{x}_k' \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}_k} \exp(\boldsymbol{x}_l' \boldsymbol{\beta})}, \tag{1.11}$$

where $\mathcal{D}$ is the set of indicators of failure times and $\mathcal{R}_k$ is the set of indicator of subjects at risk right before t_k. If there are ties, we use the approximation [Breslow and Crowley, 1974] or [Efron, 1977] to the partial likelihood (1.11).

1.1.2 Accelerated Failure Time Models

In the hazard function (1.6), the multiplicative effect of the covariate has a clear meaning, but because of unknown baseline hazard function $\lambda_0(\cdot)$, there is no direction relationship between covariate $\boldsymbol{x}$ and the survival time T. Suppose a linear model $Y = \boldsymbol{x}'\beta + \theta$, where $Y = \log(T)$ and ϵ is an error variable with some density function. The model can be written as $T = \exp(\boldsymbol{x}'\beta)V$, where $V = \exp(\theta)$ has hazard function $\lambda_0(v)$. Then the hazard function for T with covariates $\boldsymbol{x}$ can be written as

$$\lambda(t|\boldsymbol{x}) = \exp(-\boldsymbol{x}'\beta)\lambda_0\{t \exp(-\boldsymbol{x}'\beta)\}. \tag{1.12}$$

In (1.12), it is obvious that the effect of covariates in the model is multiplicative on t. When $\boldsymbol{x} = \mathbf{0}$, there is a baseline hazard function $\lambda_0(t)$; when $\boldsymbol{x} \neq \mathbf{0}$, the covariates of each subject affects the hazard rate along with t. The role of covariate is to accelerate (or decelerate) the time to failure. The corresponding survivor function

is

$$\begin{aligned} S(t|\boldsymbol{x}) &= \exp\{-\int_0^t \exp(-\boldsymbol{x}'\beta)\lambda_0(\mu e^{-\boldsymbol{x}'\beta})d\mu\} \\ &= \exp\{-\Lambda_0(te^{-\boldsymbol{x}'\beta})\}, \end{aligned}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(\mu)d\mu$.

The extensions of the model include stratifying the model and incorporating time-dependent covariates.

1.2 Survival Analysis under the Bayesian Perspective

1.2.1 Parametric Models

Parametric modeling is straightforward, and many Bayesian analyses in practice are based on a parametric model. In this section, we cover the Weibull model, one of the most widely used parametric survival model.

Suppose we have survival times $\boldsymbol{t} = (t_1, t_2, \cdots, t_n)'$, each independent and identically following Weibull distribution, $\mathcal{W}(\alpha, \lambda)$, as

$$f(t_i|\alpha, \lambda) = \alpha t_i^{\alpha-1} \exp(\lambda - \exp(\lambda)t_i^{\alpha}), \tag{1.13}$$

where $i = 1, \cdots, n$. The corresponding survival function is $S(t_i|\alpha, \lambda) = \exp(-\exp(\lambda)t_i^{\alpha})$. The censoring indicator is given as $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_n)$ with $\delta_i = 0$ when t_i is right censored time and $\delta_i = 1$. When t_i is a known survival time, we can write the likelihood

function of (α, λ) as

$$\begin{aligned} L(\alpha, \lambda | n, \boldsymbol{t}, \boldsymbol{\delta}) &= \prod_{i=1}^{n} f(t_i|\alpha, \lambda)^{\delta_i} S(t_i|\alpha, \lambda)^{1-\delta_i} \\ &= \alpha^{\sum_{i=1}^{n} \delta_i} \exp\{\lambda \sum_{i=1}^{n} \delta_i + \sum_{i=1}^{n} (\delta_i(\alpha - 1)\log(t_i) - \exp(\lambda) t_i^{\alpha})\}. \end{aligned} \tag{1.14}$$

To form a Weibull regression model, let $\lambda_i = \boldsymbol{x}_i' \boldsymbol{\beta}$, where the covariate $\boldsymbol{x}_i$ is a $p \times 1$ vector and corresponding regression coefficient parameter $\boldsymbol{\beta}$ is also a $p \times 1$ vector.

Let $N_p(\boldsymbol{\mu}_0, \Sigma_0)$ to be the normal prior for $\boldsymbol{\beta}$ and $\mathcal{G}(\alpha_0, \kappa_0)$ to be the gamma prior for α, we have the joint posterior as

$$\begin{aligned} \pi(\boldsymbol{\beta}, \alpha | n, \boldsymbol{t}, \boldsymbol{d}elta) \propto \alpha^{\sum_{i=1}^{n} \delta_i + \alpha_0 - 1} \exp\{\sum_{i=1}^{n} (\delta_i \boldsymbol{x}_i' \boldsymbol{\beta} + \delta_i(\alpha - 1)\log(t_i) - t_i^{\alpha} \exp(\boldsymbol{x}_i' \boldsymbol{\beta}) \\ - \kappa_0 \alpha - \frac{(\boldsymbol{\beta} - \boldsymbol{\mu}_0)\Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)}{2}\} \end{aligned} \tag{1.15}$$

The posterior distribution of $\boldsymbol{\beta}$ does not have a closed form, so numerical integration or MCMC methods are used to estimate the posterior distribution of $\boldsymbol{\beta}$.

1.2.2 Semiparametric Models

In this section, we consider Bayesian semiparametric approach for the accelerated failure time model and Bayesian Cox proportional model.

Let data set without censoring, $\boldsymbol{t} = (t_1, t_2, \cdots, t_n)'$ be independently and identically distributed, X is $n \times p$ matrix of covariates with i_{th} row $\boldsymbol{x}_i'$ representing a vector of covariates for subject i, and $\boldsymbol{\beta}$ is the corresponding coefficient of the covariates.

From section 1.1.2, the probability model is,

$$t_i = \exp(-\boldsymbol{x}_i'\boldsymbol{\beta})\nu_i, \tag{1.16}$$

where $\nu_i = \exp(\theta_i)$. A mixture of Dirichlet processes (MDP) is used as a prior for θ_i by [Kuo and Mallick, 1997]. Assume ν_i are independently and identically distributed with density:

$$f(\nu_i|G) = \int f(\nu_i|\psi_i)G(d\psi_i), \tag{1.17}$$

where unknown G is given by a Dirichlet-process prior with known parameters and $f(\nu_i|\psi_i)$ is a kernel density with kernel parameter ψ_i. With (1.2.2) and (1.17), the likelihood function of Y can be written as,

$$f(Y|\boldsymbol{\beta}, G) = \prod_{i=1}^{n} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \int f(y_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})|\psi_i)G(d\psi_i). \tag{1.18}$$

Let the prior of $\boldsymbol{\beta} = \pi(\boldsymbol{\beta})$, the posterior of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\psi, Y \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) f(y_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})|\psi_i). \tag{1.19}$$

[Kuo and Mallick, 1997] has more details. [Ghosh and Ghosal, 2005] prove the posterior consistency of semiparametric AFT models with censored data. For the censored data, data augmentation is used. Let $\delta_i = I\{t_i \leq c_i\}$ be the censoring indicator and $W = (w_1, \cdots, w_n)'$, where $w_i = log(t_i)$, be the augmented data, we have

$$\begin{cases} w_i = \log(t_i^*) & \text{if} \quad \delta_i = 1 \\ w_i > \log(t_i^*) & \text{if} \quad \delta_i = 0. \end{cases} \tag{1.20}$$

[Sha et al., 2006] assumes the θ in are iid $N(0, \sigma^2)$, as a result the T's are log-normally distributed. The augmented data follow normal distribution, $W|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, with $\boldsymbol{I}_{n\times n}$ the identity matrix.

The priors for this model are following,

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \sigma^2\Sigma_0) \tag{1.21}$$

$$\sigma^2 \sim IG(v_0/2, v_0\sigma_0^2/2). \tag{1.22}$$

[Sha et al., 2006] is interested in variable selection rather than estimation of β's. The mixture priors for variable selection is

$$\beta_j|\gamma_j, \sigma^2 \sim (1-\gamma_j)I(0) + \gamma_j N(0, \sigma^2\tau_j), \tag{1.23}$$

where τ_j is the jth diagonal element of σ_0. $\pi(\gamma_j)$ is the prior for γ_j following independent Bernoulli distribution. $\gamma_j = 1$ indicates j-th variable is selected in the model.

After integrating out $\boldsymbol{\beta}$ and σ^2, marginal likelihood of the augmented data is a multivariate t-distribution,

$$W|\boldsymbol{X}_{(\gamma)} \sim \mathcal{T}_{v_0}[\boldsymbol{X}\boldsymbol{\beta}_0, \sigma_0(\boldsymbol{I} + \boldsymbol{X}\Sigma_0\boldsymbol{X}')]. \tag{1.24}$$

Posterior for the γ is,

$$p(\gamma|\boldsymbol{X}, W) \propto \prod_{j=1}^{n} p(\gamma_j)p(W|\boldsymbol{X}_{(\gamma)}). \tag{1.25}$$

[Sha et al., 2006] also discuss the model with log-t prior for β's.

The other model is one of the most convenient and popular models for semiparametric survival analysis, the Cox proportional hazards model. Instead of assuming multiplicative effect on the t, cox model assumes a multiplicative effect on the hazard functions. To construct this model, we first consider a finite partition of time, $0 < s_1 < s_2 < \cdots < s_K$, with $s_K > t_i$ for all subjects from $i = 1, 2, \cdots, n$. Thus, we form K intervals, and the k intervals is $I_k \in (s_{k-1}, s_k]$. To form a piecewise constant hazard model, let the baseline hazard $\lambda_0(t) = \lambda_k$ for $t \in I_k$ and $\boldsymbol{t} = (t_1, t_2, \cdots, t_n)$. The maximum likelihood function is

$$p(\boldsymbol{t}|\boldsymbol{\beta}, \boldsymbol{\lambda}, X, \nu, \delta) = \prod_{i=1}^{n}\prod_{k=1}^{K}(\lambda_j \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))^{\delta_{ik}\nu_i} \exp\{-\delta_{ik}[\lambda_j(t_i - s_{k-1}) + \sum_{g=1}^{k-1}\lambda_g(s_g - s_{g-1})] \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\}, \tag{1.26}$$

where $\boldsymbol{\nu} = (\nu_1, \nu_2, \cdots, \nu_n)'$ with $\nu_i = 1$ if the i^{th} subject has an exact failure time and 0 right censored time, $\delta_{ik} = 1$ if the i^{th} subject failed or is censored in the interval I_k.

Prior of the baseline hazard $\boldsymbol{\lambda}$ follows independent gamma distribution and prior of $\boldsymbol{\beta}$ follows independent normal distribution,

$$\pi(\lambda_k) \sim \mathcal{G}(\alpha_{0k}, \lambda_{0k}) \tag{1.27}$$

$$\pi(\beta_j) \sim N(0, \sigma_0^2), \tag{1.28}$$

where α_{0k}, λ_{0k} and σ_0^2 are known. And the joint posterior distribution of $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ is,

$$p(\boldsymbol{\lambda}, \boldsymbol{\beta}|\boldsymbol{t}, X, \nu, \delta) \propto p(\boldsymbol{t}|\boldsymbol{\beta}, \boldsymbol{\lambda}, X, \nu, \delta)\pi(\boldsymbol{\lambda})\pi(\boldsymbol{\beta}), \tag{1.29}$$

[Sinha et al., 1999] considers discrete hazard model allows for time-dependent regression coefficients.

1.3 Variable Selection in Gene Expression Data and Related Difficulties

DNA microarrays measure the expression levels of large numbers of genes simultaneously. These measurements, gene expression profiling, can identify between cells that are actively dividing, or show how the cells react to a treatment. One of the objectives to analyze gene expression is to link the survival time to certain subset of genes, pathway or both. The identified genes/pathways in the subset can be used either to inform biologists to do more research on the related subset or to build a statistical model to predict the survival time of new patients.

In statistics analysis, one difficulty is the incomplete data due to censoring. The other difficulty is the number of genes(p) is usually much larger than the number of experimental subjects(n), because of the nature of gene expression data. Therefore transitional regression analysis is not useful in this scenario. Supervised principal components [Bair et al., 2006] which reduces the dimension of the predictor can be used to solve this difficulty.

Even though $p \gg n$, the number of genes related to survival time is usually very small comparing to p. To encourage the sparsity of the coefficients and model selection in the same time, [Tibshirani, 1996] introduced the least absolute shrinkage and selection operator (Lasso) penalty based on the L_1-norm. The lasso method makes some coefficient exactly equal to 0 and hence the genes related to survival time

can be identified. [Tibshirani, 1997] extend the Lasso method to Cox proportional model. [Gui and Li, 2005] applies least-angle regression (LARS) method to Cox model. LARS-COX procedure reduce the computational difficulty of Lasso Method based on the L_1-norm in the Cox Model. However, all the methods above can only select at most n genes. If the n is relatively very small, this limitation would cause a problem. Bayesian framework can handle this limitation.

Bayesian framework can solve the limitation we mentioned above. Variable selection under the Bayesian framework traditionally has been done by the SSVS(Stochastic Search Variable Selection) procedure [George and McCulloch, 2005]. The predictors that have higher posterior probability can form a promising subset in this procedure. [E. et al., 2003] and [Tibshirani et al., 2005b] extend the SSVS procedure to discrete response models. [Tanner and Wong, 1987] introduce data augmentation by calculating the posterior distribution of missing data. This approach is widely used to impute the censored data in survival analysis. [Sha et al., 2006] consider accelerated failure time (AFT) models to select important genes with augmented survival data assuming the survival time follows log-normal or log-t distribution. Bayesian gene selection applies in non-linear binary and multiclass problems by [Chakraborty et al., 2007] and [Chakraborty, 2009]

1.3.1 Supervised Principal Components Regression

[Bair et al., 2006] propsed supervised principal components regression (SPC) by adapting ideas of dimension reduction and penalized regression. The idea of SPC is to compute univariate standard regression coefficients for each predictor and only keep the predictors whose absolute value of univariate coefficient exceeds a threshold

θ. The remained predictors form a reduced matrix. Then we compute the first (or first few) principal component of the reduce matrix and use them to predict the response variable. SPC has a consistent estimation for regression coefficient parameters as n and $p \to \infty$, but the usual principal components regression does not.

1.3.2 Cox Univariate Shrinkage Method

[Tibshirani, 2009] proposed Cox univariate shrinkage (CUS) estimator, which finds estimate using a set of simple one-dimensional maximization with the Lasso penalty under the assumption that the features are independent. Under this assumption, the partial likelihood (1.11) can be written as following:

$$L(\boldsymbol{\beta}) \propto \prod_{j-1}^{p} \prod_{k \in \mathcal{D}} \frac{\exp(x_{kj}\beta_j)}{\sum_{m \in \mathcal{R}_k} \exp(x_{mj}\beta_j)} \tag{1.30}$$

The log partial likelihood is

$$l(\boldsymbol{\beta}) \propto \sum_{j=1}^{p} \sum_{k=1}^{K} (x_{kj}\beta_j - \log \sum_{m \in \mathcal{R}_k} \exp(x_{mj}\beta_j)) \tag{1.31}$$

where K represent the total number of different failure times. The proposed CUS estimator is as the maximizer of the penalized partial log-likelihood,

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} g_j(\beta_j) - \lambda \sum |\beta_j| \tag{1.32}$$

where $g_j(\beta) \equiv \sum_{k=1}^{K} (x_{kj}\beta_j - \log \sum_{m \in \mathcal{R}_k} \exp(x_{mj}\beta_j))$ and $\lambda \geq 0$ is the tuning parameter. The problem can be solved for a range of λ values and it is a set of one-dimensional

maximizations, because each function $g_j(\beta_j) - \lambda \sum |\beta_j|$ in (1.32) can be maximized separately.

1.3.3 Iterative Bayesian Model Average

To Analyze survival data with microarray predictors, [Annest et al., 2009] developed the iterative Bayesian Model Average (BMA) algorithm. In this algorithm, the partial log likelihood of each genomic variable is calculated and the top 25 geneomic variables with a largest log likelihood value are chosen in the initial model. After applying iterative BMA algorithm, the 25 genes which have low posterior probabilities, generally the threshold is 1%, would be removed from the initial model. Suppose we have k number of genes removed from the initial model, then k genes with highest log likelihood value next to the initial 25 genes would be selected in the model. The process continues until all the genes have been considered. The traditional BMA algorithm includes the leaps and bounds algorithm and it is not efficient when the number of predictors is greater than 30, so only 25 genes are considered at each iteration. As a result, iterative BMA cannot selected more than 25 genes in our case, more generally, more than the size of the BMA window (maximum 30).

1.3.4 Bayesian Variable Selection in AFT Model

Based on variable selection in regression and multinomial probit models [Sha et al., 2004], [Sha et al., 2006] extended this Bayesian variable selection approach to accelerated failure time (AFT) models. The censored survival times are imputed using a data augmentation approach proposed by [Tanner and Wong, 1987] with log-normal

or log-t distributional assumptions. The full conditional of a censored case follows a univariate truncated t-distribution and it can be updated using Gibbs sampling. The regression coefficients are assumed to arise from a scale mixture of a point mass at 0 and a normal density [George and McCulloch, 2005] by adding a latent vector, γ, with Bernoulli distribution to the prior of coefficients. The joint posterior distribution of γ or the marginal posterior distributions of its elements can be used to make the variable selection. We discuss the model on detail in section(1.2.2).

There are some major limitations for microarray data analysis when only one gene is considered individually, because cellular processes often affect sets of genes instead of one, and the biological mechanisms are more related to moderate changes in several genes than dramatic change in a single gene [Mootha et al., 2003]. [Liu et al., 2007] consider a semiparametric regression model with covariates and a genetic pathway. The covariates are modeled parametrically and the genes in the pathway are modeled using least-squares kernel machines (LSKMS). The overall effect of the pathway can be tested in the semiparametric model.

[Stingo et al., 2011] considers the selection of pathways and genes simultaneously with biological information, which includes the membership of genes in pathways and the relationships between genes, Kyoto Encyclopedia of Genes and Genomes (KEGG)[Kanehisa and Goto, 2000].

[Stingo et al., 2011] apply PLS regression of Y on a subset of selected genes and pathways,

$$Y = \mathbf{1}\alpha + \sum_{k=1}^{K_\theta} T_{k(\gamma)}\boldsymbol{\beta}_{k(\gamma)} + \epsilon, \tag{1.33}$$

where θ is the indicator of selected pathways, $K_\theta = \sum_{k=1}^{K} \theta_k$ is the number of selected pathways, γ is the indicator of selected genes, $\gamma\, T_{k(\gamma)}$ is the first latent PLS component

from microarray data of selected pathway k and corresponding selected genes, and $\epsilon \sim N(0, \sigma^2)$. The model can be also written as,

$$Y|\boldsymbol{X}, \alpha, \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{1}\alpha + \sum_{k=1}^{K_\theta} T_{k(\gamma)}\boldsymbol{\beta}_{k(\gamma)}, \sigma^2\boldsymbol{I}). \tag{1.34}$$

Priors for regression parameters are,

$$\beta_k|\theta_k, \sigma^2 \sim \theta_k N(\beta_0, h\sigma^2) + (1-\theta_k)\delta_0(\beta_k) \tag{1.35}$$

$$\alpha|\sigma^2 \sim N(\alpha_0, h_0\sigma^2) \tag{1.36}$$

$$\sigma^2 \sim IG(v_0/2, v_0\sigma_0^2/2), \tag{1.37}$$

where α_0, β_0, h_0, h, v_0 and σ_0^2 are known.

Priors for pathway and gene selection indicator are,

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mu, \eta) \propto \prod_{k=1}^{K} \psi_k^{\theta_k}(1-\psi_k)^{1-\theta_k} \exp(\mu\mathbf{1}'\boldsymbol{\gamma} + \eta\boldsymbol{\gamma}'\boldsymbol{R}\boldsymbol{\gamma}), \tag{1.38}$$

where ψ_k, μ, and η are known. $\boldsymbol{R}$ is the gene relationship matrix.

By multiplying the prior of α, $\boldsymbol{\beta}$ and σ^2 to (1.34) and then integrating out them, we get a multivariate t-distribution,

$$f(Y|\boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \sim \mathcal{T}_{v_0}(\alpha_0\mathbf{1} + T_{\theta,\gamma}\beta_0, \sigma_0^2(\boldsymbol{I} = h_0\mathbf{1}\mathbf{1}' + T_{\theta,\gamma}\Sigma_0\boldsymbol{T}'_{\theta,\gamma})). \tag{1.39}$$

And the joint posterior distribution of the pathway and gene selection indicators is

$$f(\boldsymbol{\theta}, \boldsymbol{\gamma}, \eta|\boldsymbol{T}, Y) \propto f(Y|\boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\gamma})p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mu, \eta). \tag{1.40}$$

Pathways and genes are selected by three moves: adding/removing a pathway and a gene; adding/removing a gene; adding/removing a pathway.

1.4 Genetic pathways

A genetic pathway figure(1.2) is the set of interactions occurring between a group of genes. The interactions together execute certain biological function(s). As we mention before, biological mechanisms are more related to moderate changes in several genes than dramatic change in a single gene. Studying pathway makes us better understanding biological mechanisms. It is possible that more than one pathway related to a certain disease, and finding those related pathways will help us learn more about disease process figure(1.3).

1.5 Structural Equation Modeling

Structural Equation Modeling can be used where the data set contain manifest(observed) and latent(unobserved) variables. Manifest variables can be measured directly, while latent variables cannot.

1.5.1 Introduction

Exploratory factor analysis (EFA) is a basic SEM, and it is defined as follow:

$$\boldsymbol{y} = \boldsymbol{\Lambda\omega} + \boldsymbol{\epsilon} \tag{1.41}$$

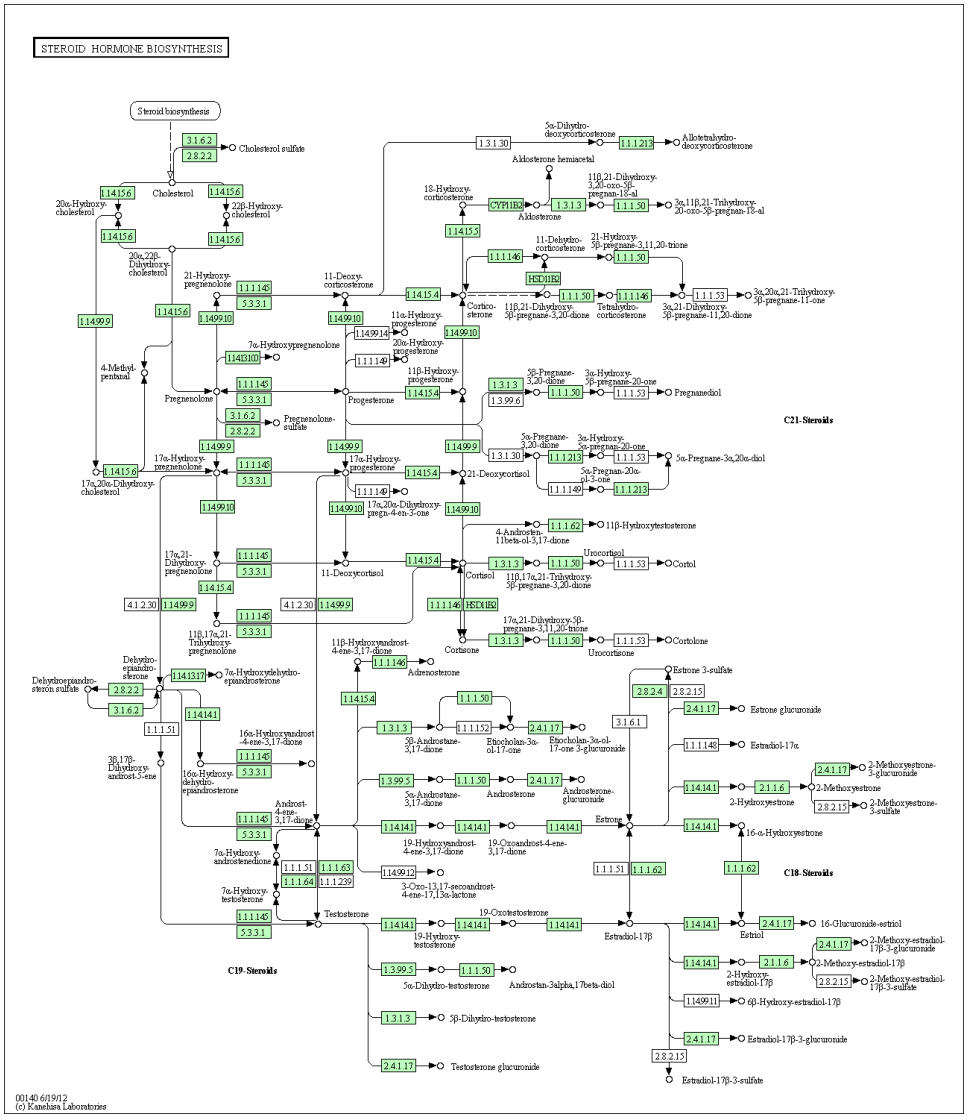


Figure 1.2: Steroid hormone biosynthesis pathway

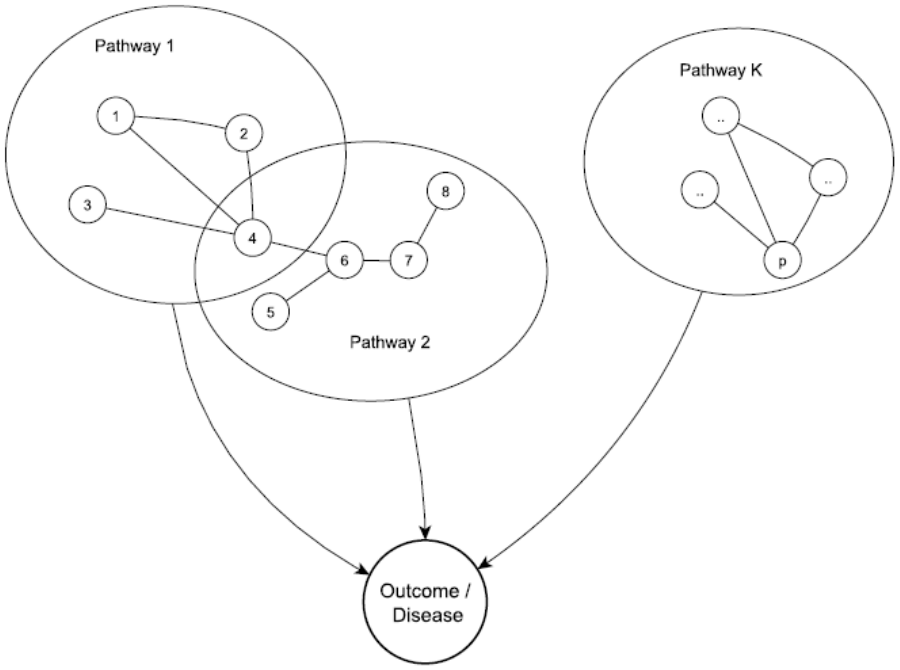


Figure 1.3: Schematic representation of the relationship among genes, pathways and diseases [Stingo et al., 2011]

where $\mathbf{\Lambda}$ is a $p \times q$ unknown parameter matrix (factor loadings), $\boldsymbol{\omega}$ is a $q \times 1$ vector of latent variables, $\boldsymbol{\epsilon}$ is a $p \times 1$ vector of measurement errors. $\boldsymbol{\omega}$ and $\boldsymbol{\epsilon}$ are independent. $\boldsymbol{\omega}$ follows a $N[\mathbf{0}, \boldsymbol{I}]$ distribution and $\boldsymbol{\epsilon}$ follows normal distribution as $N[\mathbf{0}, \mathbf{\Psi}_\epsilon]$, where $\mathbf{\Psi}_\epsilon$ is a diagonal matrix. The observable response variables $\boldsymbol{y}$ follows a $N[\mathbf{0}, \mathbf{\Sigma}]$, where $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}_\epsilon$.

The latent variables are correlated with each other, which means $\boldsymbol{\epsilon}$ follows a $N[\mathbf{0}, \mathbf{\Phi}]$ and $\mathbf{\Phi}$ is a positive definite covariance, so that the previous model becomes the confirmatory factor analysis (CFA) model, which is a natural extension of the EFA model. And $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^T + \mathbf{\Psi}_\epsilon$

(1.41) is refereed to be measurement equation, which represent the relationship between Manifest variables and latent variable. In a general structural equation model, the relationship among latent variables is also considered. If SEMs assume linear relations among latent variables [Jőreskog, 1973], the full structural model is defined as follows:

$$\boldsymbol{\eta} = \mathbf{\Pi}\boldsymbol{\eta} + \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{1.42}$$

where $\boldsymbol{\eta}$ is a $q_1 \times 1$ vector of endogenous latent variables and $\boldsymbol{\xi}$ is a $q_2 \times 1$ vector of exogenous latent variables, $\mathbf{\Pi}$ is a $q_1 \times q_1$ unknown matrix of regression coefficients relating the latent endogenous variables to each other and $\mathbf{\Gamma}$ is a $q_2 \times q_2$ unknown matrix of regression coefficients relating the exogenous latent variables to the endogenous latent variables. In this case, $\boldsymbol{\omega}$ can be defined as $\boldsymbol{\omega} = (\boldsymbol{\eta}^T, \boldsymbol{\xi}^T)^T$, so the measurement equation for the general structure equation model is still (1.41).

1.5.2 Bayesian Estimation

To illustrate the Bayesian method, let us consider CFA model. Suppose there are n observations and $i = 1, \cdots, n$, so (1.41) becomes:

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i. \tag{1.43}$$

Priors

Let $\boldsymbol{\Lambda}_k$ be the kth column of $\boldsymbol{\Lambda}$ and kth diagonal elements of $\boldsymbol{\Psi}$ be $\psi_{\epsilon k}$, conjugate priors for $\boldsymbol{\Lambda}_k$ and $\psi_{\epsilon k}$ are,

$$\begin{aligned} \boldsymbol{\Lambda}_k|\psi_{\epsilon k} &\sim N(\boldsymbol{\Lambda}_{0k}, \psi_{\epsilon k}\boldsymbol{H}_{0yk}) && (1.44) \\ \psi_{\epsilon k}^{-1} &\sim Gamma(\alpha_{0\epsilon k}, \beta_{0\epsilon k}), && (1.45) \end{aligned}$$

where $\alpha_{0\epsilon k}$, $\beta_{0\epsilon k}$, $\boldsymbol{\Lambda}_{0k}$ and positive definite matrix $\boldsymbol{H}_{0yk}$ are hyperparameters.

For $\boldsymbol{\Phi}$, a conjuage prior is a q dimensional Inverted Wishart distribution:

$$\boldsymbol{\Phi} \sim IW_q(\boldsymbol{R}_0, \rho_0), \tag{1.46}$$

where positive definite matrix $\boldsymbol{R}_0$ and ρ are hyperparameters.

Full conditional distribution

The parameters of interest are $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon, \boldsymbol{\Phi})^T$. Let $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_n)$ and the conditional distribution of $\boldsymbol{\Omega}$ is:

$$p(\boldsymbol{\Omega}|\boldsymbol{Y}, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\boldsymbol{\theta}) p(\boldsymbol{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta}), \tag{1.47}$$

where $\boldsymbol{\omega}_i \sim N(\mathbf{0}, \boldsymbol{\Phi})$ and $\boldsymbol{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\theta} \sim N(\boldsymbol{\Lambda}\boldsymbol{\omega}_i, \boldsymbol{\Psi}_\epsilon)$, so

$$\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{\theta} \sim N((\boldsymbol{\Phi}^{-1} + \boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\boldsymbol{y}_i, (\boldsymbol{\Phi}^{-1} + \boldsymbol{\Lambda}^T\boldsymbol{\Psi}_\epsilon^{-1}\boldsymbol{\Lambda})^{-1}). \tag{1.48}$$

For $\boldsymbol{\theta}$, the conditional distributions are:

$$\psi_{\epsilon k}^{-1}|\boldsymbol{Y}, \boldsymbol{\Omega} \sim Gamma(n/2 + \alpha_{0\epsilon k}, \beta_{\epsilon k}) \tag{1.49}$$

$$\boldsymbol{\Lambda}_k|\boldsymbol{Y}, \boldsymbol{\Omega}, \psi_{\epsilon k}^{-1} \sim N(\boldsymbol{a}_k, \psi\epsilon k\boldsymbol{A}_k) \tag{1.50}$$

$$\boldsymbol{\Phi}|\boldsymbol{Y}, \boldsymbol{\Omega} \sim IW_q(\boldsymbol{\Omega}\boldsymbol{\Omega}^T + \boldsymbol{R}_0^{-1}, n + \rho_0), \tag{1.51}$$

where $\boldsymbol{A}_k = (\boldsymbol{H}_{0yk}^{-1} + \boldsymbol{\Omega}\boldsymbol{\Omega}^T)^{-1}$ and $\boldsymbol{a}_k = \boldsymbol{A}_k(\boldsymbol{H}_{0yk}^{-1}\boldsymbol{\Lambda}_{0k} + \boldsymbol{\Omega}\boldsymbol{Y}_k)$

The Gibbs sampler can be used to generate the posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$.

1.6 Motivation and Outline of the Study

Much attention has been given recently to the development of methods that utilize the large quantity of genetic information. Most of the proposed methods look at the entire set of genes and their impact on a disease. Recently a new philosophy emerged which considers the genetic pathways, which contain sets of genes, combined effect

on a disease. Under the new philosophy the goal is to identify the significant genetic pathways and the corresponding influential genes in regards to different diseases.

In Chapter 2, a Bayesian kernel machine model which incorporates existing information on pathways and gene networks in the analysis of DNA microarray data is developed. Each pathway is modeled nonparametrically using a reproducing kernel Hilbert space. Mixture priors on the pathway indicator variable and the gene indicator variable are assigned. This approach can be used to model both linear and non-linear pathway effects and can pinpoint the important pathways along with the active genes within each pathway. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed to fit our model. A simulation study and a real data analysis, using, [van 't Veer et al., 2002] breast cancer microarray data, are used to illustrate the proposed method.

In Chapter 3, we focus on Structural equation modeling. Structural equation models are a well-developed statistical tool for dealing with multivariate data that contain latent variables. Recently much attention has been given to developing structural equation models that account for nonlinear relationships between the endogenous latent variable and the covariates and endogenous latent variables. [Guo et al., 2012] developed a semiparametric structural equation model where the nonlinear functional relationships were approximated using basis expansions. Many basis expansion methods, including cubic splines, are known to induce correlations. In this chapter, we compare standard Lasso, Fused Lasso and Elastic Net to account for correlations in both the covariate and basis expansions. To illustrate the usefulness of the proposed method a simulation study has been performed. Results indicate that the Elastic Net is most efficient at approximating the nonlinear relationships between the endogenous

latent variable and the covariates and endogenous latent variables.

1.7 Software and Data Sets

1.7.1 Software

The methodologies that we mentioned on 1.3, Supervised Principal Components Regression, Cox Univariate Shrinkage Method and Iterative Bayesian Model Average, R packages are available. They are superpc, uniCox and iterativeBAMsurv respectively. Matlab codes for BVSME-Surv and bvssurv are available *http://www.stat.rice.edu/ marina/software.html*.

- superpc- The package superpc uses the functions, superpc.train and superpc.predict to predict a quantitative regression or survival outcome using supervised principal components method. The accuracy of the estimation can be set by n.threshold option, which decides the number of the thresholds to consider.

- uniCox- The package uniCox uses Univariate Shrinkage to fit a high dimensional Cox model. The estimation accuracy and computation time are decided by the option nlam, the number of λ values to consider.

- iterativeBAMsurv- The package iterativeBAMsurv use the function iterativeBAMsurv.train to implement iterative BMA for variable selection on microarray data and survival analysis.

- BVSME-Surv- The Matlab program BVSME-Surv use function bvsme_aft to implement Bayesian variable selection method in AFT model using Metropolis

search for the micorarray data related to survival time.

- bvssurv The Matlab program bvssurv is used to implement Bayesian variable selection method to choose important pathways and genes simultaneously by incorporating information of the relationship of pathway and genes in the analysis of DNA microarray data.

- Bayesian Lasso for Semiparametric Structural Equation Models Illustrative Code The C++ program of the Illustrative Code use a simulation study to implement Bayesian Lasso method with a Markov Chain Monte Carlo (MCMC) algorithm to SEM.

1.7.2 Data Sets

- **NCI breast cancer data** [van 't Veer et al., 2002] - This data set has 295 consecutive patients with primary breast cancer. 151 had lymph-node-negaitve disease, and 144 had lymph-node-positive disease. It also includes 24481 gene-expression signatures.

- **Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2010** - This survey is conducted by the University of Michigan's Institute for Social Research. This data set has total 12999 observations and some 1400 variables.

Chapter 2

Bayesian Kernel Based Modeling and Selection of Genetic Pathways and Genes for Cancer

2.1 Introduction

DNA microarray data have been used as an approach to cancer classification previous knowledge of those classes [Golub et al., 1999]. A lot of statistical methods have been develop to identify important genes related certain diseases, prognosis etc. However, gene selection may not be enough for more completed disease, especially in cancer. Cancer is result of deregulation of one or more signaling pathways which are caused by one or several set(s) of gene mutation[Sherr, 1996],[Hanahan and Weinberg, 2000]. Some types of cancers are more complicated, for instance, breast cancer. It is possible that the genes or pathways which mutate to cause breast cancer are mostly different between two breast cancer patients. The difference of genes or pathways mutation

may be related to the patients' cancer recurrence possibilities and times. Our goal is to find out those important genes and pathways that might be related to the recurrence possibilities and tumor free time. In this chapter, we construct a semiparametric Bayesian model which enable us to select important pathways and individual important genes from the pathway by mixed priors through Bayesian variable selection scheme.

We extend the idea of AFT models for survival data in the situation where there are much more variables than observations. The model that we propose consider both genes and pathways and combines information of pathway relationships and gene networks in DNA microarray data analysis. The pathway and gene mapping information are obtained from Kyoto Encyclopedia of Genes and Genomes(KEGG). The gene networks information is used not only to define Markov random field prior [Stingo et al., 2011] but also to structure the Markov chain Monte Carlo (MCMC) moves. The interactions among the genes in one pathway are very complex and the function form of the overall pathway effects is not clearly understood, so that we adopt a reproducing kernel Hilbert spaces (RKHS) [Aronszajn, 1950] approach, therefore we model the pathway effects nonparametrically. A big advantage of our approach is that both linear and non-linear pathway effect can be used in the model, so that the model is more flexible. Moreover, the model can perform different important genes and pathways selection criteria by choosing different kernel functions.

Section 2 of this chapter introduced our semiparametric Bayesian model. Step by step MCMC algorithm is introduced in Section 3. Section 4 provides simulation studies and compare the performance of our proposed model against the method proposed by [Stingo et al., 2011]. The application of the model for a real data set is

discussed in Section 5. Finally, some discussion and concluding remarks are made in section 6.

2.2 Bayesian Kernel Based Model

Accelerated failure time (AFT) models assumes multiplicative effect of the covariates on the survival time, and the general AFT model is as,

$$\log(\boldsymbol{T}_i) = \alpha + \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \cdots, n \tag{2.1}$$

where $\boldsymbol{T}_i$ is the survival time, α is the intercept, p-vector $\boldsymbol{x}_i$ is covariates, p-vector $\boldsymbol{\beta}$ is regression parameters corresponding to the covariates, and ϵ_i's are the error term which independent and identically distributed random variables with a common distribution.

Suppose a data set consists of n subjects. For the subject i, we have the survival time t_i. Let c_i be the censoring time independent of t_i. Let $\delta_i = I\{t_i \leq c_i\}$ to be censored indicator function and $t_i^* = \min(t_i, c_i)$. We impute the censored data by using the [Tanner and Wong, 1987] data augmentation approach. Let $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)'$, and y_i is the augmented data as,

$$\begin{cases} y_i = \log(t_i^*) & \text{if} \quad \delta_i = 1 \\ y_i > \log(t_i^*) & \text{if} \quad \delta_i = 0 \end{cases} \tag{2.2}$$

The covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ in (2.1) is extended as $\boldsymbol{f}_i = (f_1(\boldsymbol{X}_i^1), f_2(\boldsymbol{X}_i^2), \cdots, f_J(\boldsymbol{X}_i^J))$, $j = 1, 2, \cdots, J$. $f_j(\boldsymbol{X}_i^j)$ is the overall effect of the pathway j with gene set $\boldsymbol{X}_i^j$ and J

is the total number of the pathway. Assuming the error term is *iid* following standard normal distribution, the model (2.1) becomes:

$$Y_i = \alpha + f_1(\boldsymbol{X}_i^1) + \cdots + f_J(\boldsymbol{X}_i^J) + e_i, \quad e_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{2.3}$$

We adopt RKHS approach to model function $f_j(\boldsymbol{X}_i^j)$. A Hilbert space is a vector space H with an inner product $\langle g_1, g_2 \rangle$ and the norm $\|g_1\| = \langle g_1, g_1 \rangle^{1/2}$. An RKHS H is a Hilbert space of "smooth" functions defined by kernel. In an RHKS, there is a function $K : T \times T \to \mathcal{R}$ with the properties:

- $K(\cdot, \boldsymbol{x}) \in H$ and
- for any $g \in H$ and $\boldsymbol{x} \in T$, $\langle K(\cdot, \boldsymbol{x}), g(\cdot) \rangle = g(\boldsymbol{x})$

Following the representation theorem [Kimeldorf and Wahba, 1971], we have

$$f_j(\boldsymbol{X}_i^j) = \sum_{l=1}^{n} \beta_l^j K(\boldsymbol{X}_i^j, \boldsymbol{X}_l^j | \theta_j) \tag{2.4}$$

where $K(\boldsymbol{X}_i^j, \boldsymbol{X}_l^j \mid \theta_j)$ is the Kernel, $\theta_j > 0$, is the Kernel parameter and β_l^j is the Kernel weight. In our research, we choose the Kernel as,

$$K(\boldsymbol{X}_i^j, \boldsymbol{X}_l^j \mid \theta_j) = \exp(-\frac{\|\boldsymbol{X}_i^j - \boldsymbol{X}_l^j\|}{\theta}) \tag{2.5}$$

Our model will become:

$$\boldsymbol{Y} = \alpha \mathbf{1}_n + \mathcal{K}\boldsymbol{B} + \boldsymbol{e} \qquad \boldsymbol{e} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$$

where $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)'$; $\mathcal{K}$ is n by $n \times J$ matrix,

$$\mathcal{K} = \begin{pmatrix} K(\boldsymbol{X}_1^1, \boldsymbol{X}_1^1|\theta_1) & K(\boldsymbol{X}_1^1, \boldsymbol{X}_2^1|\theta_1) & \cdots & K(\boldsymbol{X}_1^1, \boldsymbol{X}_n^1|\theta_1) & \cdots & K(\boldsymbol{X}_1^J, \boldsymbol{X}_n^J|\theta_J) \\ K(\boldsymbol{X}_2^1, \boldsymbol{X}_1^1|\theta_1) & K(\boldsymbol{X}_2^1, \boldsymbol{X}_2^1|\theta_1) & \cdots & K(\boldsymbol{X}_2^1, \boldsymbol{X}_n^1|\theta_1) & \cdots & K(\boldsymbol{X}_2^J, \boldsymbol{X}_n^J|\theta_J) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ K(\boldsymbol{X}_n^1, \boldsymbol{X}_1^1|\theta_1) & K(\boldsymbol{X}_n^1, \boldsymbol{X}_2^1|\theta_1) & \cdots & K(\boldsymbol{X}_n^1, \boldsymbol{X}_n^1|\theta_1) & \cdots & K(\boldsymbol{X}_n^J, \boldsymbol{X}_n^J|\theta_J) \end{pmatrix} \tag{2.6}$$

; and B is $n \times J$ vector,

$$\boldsymbol{B}^T = (\beta_1^1, \beta_2^1, \cdots, \beta_n^1, \beta_1^2, \beta_2^2, \cdots, \beta_n^2, \cdots \beta_1^J, \beta_2^J, \cdots, \beta_n^J) \tag{2.7}$$

In order to use the information from KEGG, two matrices, $\boldsymbol{S}$ and $\boldsymbol{R}$, are constructed [Stingo et al., 2011]. $\boldsymbol{S}$ is a $J \times p$ matrix representing the relationship between genes and pathways. If gene k belongs to pathway j $s_{jk} = 1$, otherwise $s_{jk} = 0$, where $k = 1, \cdots, p$. The construction of the matrix R is different from [Li and Zhang, 2010]. Matrix $\boldsymbol{R}$ indicates the relationship between genes. We consider two types of gene relationships in our model. The first one is the genes whose coded proteins combine and form a protein compound. If gene k_1 and k_2 are in this case, $r_{k_1,k_2} = 1$. The other relationship between genes is that proteins coded by those genes signal each other. In this case, $r_{k_1,k_2} = q$, where $q > 0$. And $r_{k_1,k_2} = 0$ other wise.

2.2.1 Priors for Regression Parameters

Suppose the t_i's follow a log-normal distribution, the augmented data y_i's in (2.2) are normally distributed as:

$$\boldsymbol{Y}|\mathcal{K}, \alpha, \boldsymbol{B}, \sigma^2 \sim \mathcal{N}(\alpha\mathbf{1} + \mathcal{K}\boldsymbol{B}, \sigma^2\boldsymbol{I}_n) \tag{2.8}$$

where $\boldsymbol{I}$ is the identity matrix.

The conjugate priors for the model are given by

$$\begin{aligned} &\alpha|\sigma^2 \sim \mathcal{N}(\alpha_0, a_\alpha) \\ &\boldsymbol{B}|\sigma^2 \sim \mathcal{N}(\boldsymbol{B}_0, a_B\sigma^2\boldsymbol{I}_{n\times J}) \\ &\sigma^2 \sim \mathcal{IG}(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned} \tag{2.9}$$

where the hyperparameters are α_0, a_α, $\boldsymbol{B}_0$, a_B, ν_0 and σ_0. We choose vague priors on α and $\boldsymbol{B}$: $\alpha_0 = 0$ and large a_α; $\boldsymbol{B}_0 = \mathbf{0}$ and small a_B. Small value of ν_0 gives weakly informative prior for σ^2. Without losing generality and letting $\alpha_0 = 0$ and $\boldsymbol{B}_0 = \mathbf{0}$. The priors become:

$$\begin{aligned} &\alpha|\sigma^2 \sim \mathcal{N}(0, a_\alpha) \\ &\boldsymbol{B}|\sigma^2 \sim \mathcal{N}(\mathbf{0}, a_B\sigma^2\boldsymbol{I}) \\ &\sigma^2 \sim \mathcal{IG}(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned} \tag{2.10}$$

2.2.2 Marginal Likelihood of the Augmented data

To simplify the calculation, we integrate out α, $\boldsymbol{B}$ and σ^2.

- First, we integrate out $\boldsymbol{B}$

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{X},\mathcal{K},\alpha,\sigma^2) &= \int p(\boldsymbol{Y},\boldsymbol{B}|\boldsymbol{X},\mathcal{K},\alpha,\sigma^2)d\boldsymbol{B} \\ &= \int p(\boldsymbol{Y}|\boldsymbol{X},\mathcal{K},\boldsymbol{B},\alpha,\sigma^2)p(\boldsymbol{B}|\sigma^2)d\boldsymbol{B} \\ &\propto \frac{1}{(\sigma^2)^{n/2}}\exp[-\frac{(\boldsymbol{Y}-\mathcal{K}\boldsymbol{B})^T(\boldsymbol{Y}-\mathcal{K}\boldsymbol{B})}{2\sigma^2}] \\ &\quad \times\frac{1}{(a_B\sigma^2)^{Jn/2}}\exp[-\frac{\boldsymbol{B}^T\boldsymbol{B}}{2a_B\sigma^2}]d\boldsymbol{B} \\ &\propto \frac{1}{(\sigma^2)^{n/2}}\exp(\frac{\boldsymbol{Y}^T(I+a_\beta\mathcal{K}\mathcal{K}^T)^{-1}\boldsymbol{Y}}{2\sigma^2}) \end{aligned} \tag{2.11}$$

- We integrate out α

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{X},\mathcal{K},\sigma^2) &= \int p(\boldsymbol{Y},\alpha|\boldsymbol{X},\mathcal{K},\sigma^2)d\alpha \\ &= \int p(\boldsymbol{Y}|\boldsymbol{X},\mathcal{K}\alpha,\sigma^2)p(\alpha|\sigma^2)d\alpha \\ &\propto \int \frac{1}{(\sigma^2)^{n/2}}\exp(\frac{\boldsymbol{Y}^T(I+a_{beta}\mathcal{K}\mathcal{K}^T)^{-1}\boldsymbol{Y}}{2\sigma^2})\frac{1}{(a_\alpha\sigma^2)^{1/2}}\exp(-\frac{\alpha^2}{2a_\alpha\sigma^2})d\alpha \\ &\propto \frac{1}{(\sigma^2)^{n/2}}\exp(-\frac{\boldsymbol{Y}^T(I_n+a_\alpha\mathbf{1}^T\mathbf{1}+a_\beta\mathcal{K}\mathcal{K}^T)^{-1}\boldsymbol{Y}^T}{2\sigma^2}) \end{aligned} \tag{2.12}$$

- Finally, we integrate out σ^2

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{X},\mathcal{K}) &= \int p(\boldsymbol{Y},\sigma^2|\boldsymbol{X},\mathcal{K})d\sigma^2 \\ &= \int p(\boldsymbol{Y}|\boldsymbol{X},\mathcal{K},\sigma^2)p(\sigma^2)d\sigma^2 \\ &\propto \int (\sigma^2)^{-\nu_0/2-1}\exp(-\frac{\nu_0\sigma_0^2/2}{\sigma^2})\frac{1}{(\sigma^2)^{n/2}} \\ &\quad \exp(-\frac{\boldsymbol{Y}^T(I_n+a_\alpha\mathbf{1}^T\mathbf{1}+a_\beta\mathcal{K}\mathcal{K}^T)^{-1}\boldsymbol{Y}^T}{2\sigma^2})d\sigma^2 \\ &\propto (\nu_0\sigma_0^2+\boldsymbol{Y}^T(I_n+a_\alpha\mathbf{1}^T\mathbf{1}+a_\beta\mathcal{K}\mathcal{K}^T)^{-1}\boldsymbol{Y}^T)^{-\nu_0/2-n/2} \end{aligned} \tag{2.13}$$

2.2.3 Mixture Priors for Variable selection

The regression coefficient $\boldsymbol{B}$ in (2.7) can be written as:

$$\boldsymbol{B}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \cdots, \boldsymbol{\beta}_J^T) \tag{2.14}$$

$\boldsymbol{\beta}_j$ measures the effect of pathway j, but not all pathways are related to the dependent variable. In order to identify the important pathways, we use Bayesian methods for variable selection by applying a latent J-vector $\boldsymbol{\phi}$ with binary entries. [Chipman et al., 2001] review a vast amount of literature on Bayesian variable selection methodologies. [George and McCulloch, 2005] assumed the regression coefficients arise from a scale mixture of a point mass at 0 and a normal distribution, and our model follows this assumption, so we have:

$$\boldsymbol{\beta}^j|\phi_j,\sigma^2 \sim \phi_j\mathcal{N}(\mathbf{0},a_\beta\sigma^2\boldsymbol{I}_n)+(1-\phi_j)\mathcal{I}(0) \tag{2.15}$$

where ϕ_j's are independent Bernoulli random variables. When the pathway j is not related to the dependent variable Y, the coefficient related to the pathway j, $\boldsymbol{\beta}_j$ are all $\mathbf{0}$'s.

2.2.4 Priors for Pathway and Gene Selection indicators

From the previous subsection, we know ϕ_j is the pathway selection indicator for jth pathway. Let $\boldsymbol{\phi}$ be the pathway selection indicator, and $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_J)$.

$$\begin{cases} \phi_j = 1 & \text{when pathway is } j \text{ selected in the model} \\ \phi_j = 0 & \text{otherwise} \end{cases} \tag{2.16}$$

The priors for the pathway selection is:

$$p(\boldsymbol{\phi}|\omega_j) = \prod_{j=1}^{J} \omega_j^{\phi_j}(1-\omega_j)^{1-\phi_j} \tag{2.17}$$

where ω_j is a constant, which represents the *priori* probability of pathway j in the model.

Let the gene selection indicator, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_p)^T$. The prior distribution should be able to consider both the pathway membership of each gene and the biological relationships between genes, which we use matrix R to indicate. We model these relations using a Markov random field(MRF)[Li and Zhang, 2010]. Different from [Li and Zhang, 2010], we consider two types of gene relationships. A MRF is a set of random variables with a Markov property described by an undirected graph. If two genes are not related, they are considered to be conditionally independent given

all other genes [Besag, 1974]. The following probability indicates the connections on the MRF.

$$p(\gamma_k|\eta, \gamma_l, l \in N_k) = \frac{\exp(\gamma_k F(\gamma_k))}{1 + \exp(\gamma_k F(\gamma_k))} \tag{2.18}$$

where $F(\gamma_k) = (\mu + \eta \sum_{l \in N_k} \gamma_k)$ and N_k is the set of genes in the same protein compound of gene k and genes that receive/send signal from/to gene k in the MRF given that the pathway of those genes are in the model. The global distribution on the MRF is as:

$$p(\boldsymbol{\gamma}|\boldsymbol{\phi}) \propto \exp(\mu \mathbf{1}_p^T \boldsymbol{\gamma} + \eta \boldsymbol{\gamma}^T \boldsymbol{R} \boldsymbol{\gamma}) \tag{2.19}$$

where matrix $\boldsymbol{R}$ are introduced in the beginning of this section, μ is the parameter that relates to the sparsity of the model, and η controls the prior probability of gene selection depending on how many of its related genes are selected, so η sets the smoothness of the distribution of $\boldsymbol{\gamma}$ over the undirected graph. If a protein from one gene is isolated, then its prior distribution (2.18) becomes a Bernoulli distribution, $p = \exp(\mu)/[1 + \exp(\mu)]$. On the other hand, the higher values of η is, the more likely a gene is selected if many of its related genes are alright in the model.

Following [Stingo et al., 2011], three restrictions are needed to make sure both interpretability and identifiablility of the model.

- Empty pathways, which means a pathways is selected but none of its member genes are in the model.
- Orphan genes, which means a gene is selected in the model but none of pathways having this gene is in the model.
- Different selected pathways have the same subset of genes selected in the model.

Given these three restrictions, some of the combination of $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ are not in the model. The joint prior probability for $(\boldsymbol{\phi}, \boldsymbol{\gamma})$ is as:

$$p(\boldsymbol{\phi}, \boldsymbol{\gamma}) \propto \begin{cases} \prod_{j=1}^{J} \omega_j^{\phi_j} (1-\omega_j)^{1-\phi_j} \exp(\mu \mathbf{1}_p^T \boldsymbol{\gamma} + \eta \boldsymbol{\gamma}^T \boldsymbol{R} \boldsymbol{\gamma}) & \text{for valid configurations} \\ 0 & \text{for invalid configurations} \end{cases} \tag{2.20}$$

2.2.5 Prior for the Kernel Parameters

θ_j is the Kernel parameters for the pathway j. We use uniform distribution prior for θ_j as:

$$p(\theta_j) = \frac{1}{d_{\theta_j} - c_{\theta_j}} \tag{2.21}$$

We only consider the θ's that the corresponding pathways are in the model, so the overall distribution of Kernel Parameters is:

$$p(\boldsymbol{\theta}) = \begin{cases} \prod_{j \in \mathcal{J}} \frac{1}{d_{\theta_j} - c_{\theta_j}} & \mathcal{J} \text{ is the subset of pathways that are in the model} \\ 0 & \text{otherwise} \end{cases} \tag{2.22}$$

2.2.6 Marginal Posterior Probabilities

After integrating out $(\alpha, \boldsymbol{\beta}, \sigma^2)$ in 2.2.2, our model has following parameters $(\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta})$. given all the priors, (2.13) can be written as:

$$f(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{K}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \propto (\nu_0\sigma_0^2 + \boldsymbol{Y}^T(I_n + a_\alpha \mathbf{1}^T\mathbf{1} + a_\beta \mathcal{K}\mathcal{K}^T)^{-1}\boldsymbol{Y}^T)^{-\nu_0/2-n/2} \tag{2.23}$$

This is a multivariate t-distribution

$$\boldsymbol{Y}|\boldsymbol{X}, \mathcal{K}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta} \sim \mathcal{T}_{\nu_0}[\mathbf{0}, \sigma_0(I_n + a_\alpha \mathbf{1}^T\mathbf{1} + a_\beta \mathcal{K}\mathcal{K}^T)] \tag{2.24}$$

with truncation given by (2.2). When y_i is censored with $\delta_i = 0$, it follows a univariate truncated t-distribution and can be updated by Gibbs sampling.

The posterior probability distribution of the pathway and gene selection indicators is as:

$$f(\boldsymbol{\phi}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}, \boldsymbol{\theta}) \propto f(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{K}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \cdot p(\boldsymbol{\phi}, \boldsymbol{\gamma}) \tag{2.25}$$

Similarly, the posterior probability distribution of kernel parameter is:

$$f(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}, \boldsymbol{\phi}, \boldsymbol{\gamma}) \propto f(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{K}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) \tag{2.26}$$

2.3 MCMC algorithm and Posterior Inference

There are two MCMC steps in the model:

- sampling pathway and gene selection indicators from $p(\boldsymbol{\phi}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}, \boldsymbol{\theta})$
- sampling kernel parameters from $p(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}, \boldsymbol{\phi}, \boldsymbol{\gamma})$

2.3.1 MCMC algorithm for Pathway and Gene Selection Indicators

Metropolis-Hastings algorithm is used when updating the pathway and gene selection indicators parameter $(\boldsymbol{\phi}, \boldsymbol{\gamma})$. The MCMC moves follows [Stingo et al., 2011]. In order to select pathways and genes simultaneously and make sure the selections follow the these restriction we mentioned before, they choose one of the following moves randomly in each iteration:

1. Change the indicator of gene and pathway in the same time

- Add a pathway and a gene Randomly select a pathway from the subset of pathways that are not in the model and neither is their member genes, so $\phi_j^o = 0$ and $p_j^o = 0$, where p_j^o represent the number of genes that are included in the model in pathway j. And then randomly select one gene k from the pathway j $(\gamma_{j,k}^o = 0)$. Include both pathway and gene in the model, so that $\phi_j^p = 1$ and $p_j^p = 1$. The proposed (ϕ_j^p, p_j^p) is accepted with probability:

$$\min\{1, \frac{f(\boldsymbol{\phi}^p, \boldsymbol{\gamma}^p | \boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})}{f(\boldsymbol{\phi}^o, \boldsymbol{\gamma}^o | \boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})} \cdot \frac{\omega_j \cdot \sum_{j=1}^{J} I\{\phi_j^o = 0, p_j^o = 0\}}{\sum_{j=1}^{J} I\{\phi_j^p = 1, p_j^p = 1, \text{C1}\}}\} \tag{2.27}$$

  where C1 is clarified below.

- Remove a pathway and a gene Find a subset of the pathways that are in the model and only one of their member genes are selected in the model. Randomly select one of them, so $\phi_j^o = 1$ and $p_j^o = 1$. In addition, the removal of gene does not create identical subset of genes in different pathways, and this restriction is C1. Remove both pathway and gene from the model, so that $\phi_j^p = 0$ and

$p_j^p = 0$. The proposed (ϕ_j^p, p_j^p) is accepted with probability:

$$\min\{1, \frac{f(\boldsymbol{\phi}^p, \boldsymbol{\gamma}^p|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})}{f(\boldsymbol{\phi}^o, \boldsymbol{\gamma}^o|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})} \cdot \frac{\sum_{j=1}^J I\{\phi_j^o = 1, p_j^o = 1, \text{C1}\}}{\omega_j \cdot \sum_{j=1}^J I\{\phi_j^p = 0, p_j^p = 0\}}\} \tag{2.28}$$

2. Only change the indicator of gene in an included pathway

- Add a gene in an included pathway Find the subset($\mathcal{J}$) of pathways that have some of their member genes, but not all, in the model. And randomly select one of them j, so $\phi_j = 1$ and $p_j^o < p_j$, where p_j represent the total number of genes in pathway j. And then randomly select one gene k from non-included genes from the pathway $j(\gamma_{j,k}^o = 0)$. Include the gene in the model, so that $\gamma_{j,k}^p = 1$. The proposed move is accepted with probability:

$$\min\{1, \frac{f(\boldsymbol{\phi}^o, \boldsymbol{\gamma}^p|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})}{f(\boldsymbol{\phi}^o, \boldsymbol{\gamma}^o|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})} \cdot \frac{\sum_{j=1}^J I\{\phi_j^o = 1, p_j^o < p_j\} \cdot \sum_{j\in\mathcal{J}} \frac{1}{p_j^p(\text{C2}\gamma,\ \text{CI2}\gamma)}}{\sum_{j=1}^J I\{\phi_j^p = 1, p_j^p > 1, \text{C2}\theta,\ \text{CI2}\theta\} \sum_{j\in\mathcal{J}} \frac{1}{p_j - p_j^o}}\} \tag{2.29}$$

where C2γ, CI2γ, C2θ and CI2θ are clarified below.

- Remove a gene from an included pathway Find the subset($\mathcal{J}$) of pathways that have more than one of their member genes in the model. And randomly select one of them j, so $\phi_j = 1$ and $p_j^o > 1$. Moreover, at least one of the included genes in the pathway j may not be the sole gene in other included pathways, so that removal of the gene do not create a empty pathway. The these restrictions are corresponding to C2θ, CI2θ in (2.29) and (2.30). After the pathway is selected, the subset of the included member genes is the genes that

are not solely representative in other included pathways. And this restrictions are corresponding to C2γ, CI2γ. A gene k is randomly selected from the subset to be removed. C2θ and C2γ ensure the restrictions of the combination of pathways and genes. The proposed move is accepted with probability:

$$\min\{1, \frac{f(\boldsymbol{\phi}^o, \boldsymbol{\gamma}^p|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})}{f(\boldsymbol{\phi}^o, \boldsymbol{\gamma}^o|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})} \cdot \frac{\sum_{j=1}^{J} I\{\phi_j^o = 1, p_j^o > 1, \text{C2}\theta, \text{CI2}\theta\} \sum_{j\in\mathcal{J}} \frac{1}{p_j - p_j^p}}{\sum_{j=1}^{J} I\{\phi_j^p = 1, p_j^p < p_j\} \cdot \sum_{j\in\mathcal{J}} \frac{1}{p_j^o(\text{C2}\gamma, \text{CI2}\gamma)}}\} \tag{2.30}$$

3. Only change the indicator of the pathways not the genes

- Add a pathway Find the subset of the non-included pathways that have some of the member genes included in the model, and randomly select a pathway j ($\phi_j^o = 0$ and $p_j^o \geq 0$). Change the status of the pathway j, $\phi_j^p = 1$. In addition, avoid to select a pathway whose included genes are exactly the same as the include genes from an already included pathway. This is responding to CI3 below. Include the pathway j in the model, $p_j^p = 1$. The proposed move is accepted with probability:

$$\min\{1, \frac{f(\boldsymbol{\phi}^p, \boldsymbol{\gamma}^o|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})}{f(\boldsymbol{\phi}^o, \boldsymbol{\gamma}^o|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})} \cdot \frac{\sum_{j=1}^{J} I\{\phi_j^o = 1, p_j^o \geq 1, \text{CI3}\}}{\sum_{j=1}^{J} I\{\phi_j^p = 1, p_j^p \geq 1, \text{C3}\}}\} \tag{2.31}$$

  where C3 is clarified below.

- Remove an included pathway Find the subset of the included pathways that all of the included member genes are associated with other included pathways and randomly select one of them $j(\phi_j^o = 1$ and C3). This guarantees that no

orphan gene in the model. Remove the pathway j from the model, $p_j^p = 0$. The proposed move is accepted with probability:

$$\min\{1, \frac{f(\boldsymbol{\phi}^p, \boldsymbol{\gamma}^o|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})}{f(\boldsymbol{\phi}^o, \boldsymbol{\gamma}^o|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})} \cdot \frac{\sum_{j=1}^{J} I\{\phi_j^o = 1, p_j^o \geq 1, \text{C3}\}}{\sum_{j=1}^{J} I\{\phi_j^p = 1, p_j^p \geq 1, \text{CI3}\}}\} \tag{2.32}$$

2.3.2 MCMC algorithm for Kernel Parameters

From (2.22), $\boldsymbol{\theta}$ have flat priors. Metropolis algorithm is used when updating $\boldsymbol{\theta}$. For each included pathway, the proposed θ_j^p is accepted with probability

$$\min\{1, \frac{f(\theta_j^p|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}, \boldsymbol{\phi}, \boldsymbol{\gamma})}{f(\theta_j^o|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}, \boldsymbol{\phi}, \boldsymbol{\gamma})}\} \tag{2.33}$$

2.3.3 Posterior Inference

In each iteration, the MCMC algorithm produce one model with included pathways and gene, indicated by $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ respectively. And the whole procedure produce a list of models. To estimate the marginal posterior probability for pathway j, $p(\phi_j|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K})$, we can count the number of the pathway j appeared in the included pathways for every iteration after certain burnin point and then divide the count by the total iteration after the burnin point. If the posterior probabilities of the pathways pass some threshold, they are identified as the important pathways. Similarly, posterior probabilities of gene k in the important pathway j, $p(\gamma_k|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}, I\{\phi_j s_j k = 1\})$ can be calculated. The other way to find out important genes is to count the number of gene k appear in the model ignoring whether it is in the important pathways. The posterior probabilities of gene k can be calculate as $p(\gamma_k|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}\})$.

2.4 Simulation Study

In this section, we evaluate the proposed model performance using simulated data and comparing the result with [Stingo et al., 2011].

To simulate the data, we randomly chose 50 pathways from 243 pathways, and 1656 genes responding to these 50 pathways. The relationship between the pathways and genes in the simulation data are based on the gene-pathway relations, $\boldsymbol{S}$, and the gene relations, $\boldsymbol{R}$. We randomly chose 4 important pathways from the 50 pathways. For each of the 4 important pathways, one important gene is selected randomly and also the genes near it to code the same protein compound. Moreover, we selected the genes that send or receive signals from the important protein compound. Then we have 4 pathways and 27 important genes: 5 important genes out of 14 genes from the first pathway, 6 out of 11 from the second, 8 out of 43 form the third and 8 out of 38 from the fourth. To simulate the data like these is based on the fact the one gene mutation usually will not form cancer, but several related genes mutation might cause cancer. We also add 2 fake important genes and 2 corresponding fake important pathways in the model to check if the proposed model can avoid those 2 fake important pathways and genes. In reality, it is possible that patients have some mutation genes not related to cancer.

To generate important genes' microarray data, we first pick a gene which only sends, but not receives signal among the important genes in a pathway and let X_{00} represent the value of that gene. Let X_0 be either 2 or -2, and draw X_{00} from $X_{00} \sim U(X_0 - .5, X_0 + .5)$. For the other genes that are in the same protein compound, select one of them and denote it by $X_0 1$ and $X_{01} \sim \mathcal{N}(\rho X_{00}, .5)$, where ρ a multiplier and here we let $\rho = .95$. For the kth genes in the same protein compound, its value

is generated by $X_{0j} \sim \mathcal{N}(\rho X_{0(j-1)}, .5)$. In the neighborhood protein compound, the gene that receives the signal from X_{00} is denoted by X_{10} and $X_{10} \sim \mathcal{N}(\rho X_{00}, .5)$ and we can generate other genes value by $X_{1j} \sim \mathcal{N}(\rho X_{1(j-1)}, .5)$. If we have other neighborhood protein compound, their values are generated in similar format. For the fake important gene, we only need to draw X_{00} from $X_{00} \sim U(X_0 - .5, X_0 + .5)$. The rest of unimportant genes are simulated from a stand normal distribution. We assume non-linear relationship between the response variable and the important genes variable. The response variables are generated from:

- Let the effect of the first important pathway to be $F1$ and X_k^1 to be the kth important genes in the first important pathway.

$$F1 = \cos(X_1^1) - 1.5(X_2^1)^2 + \exp(-X_3^1)X_4^1 - .8\sin(X_5^1)\cos(X_3^1) + 2X_1^1 * X_5^1 \quad (2.34)$$

- Let the effect of the second important pathway to be $F2$ and X_k^2 to be the kth important genes in the first important pathway.

$$F2 = \cos(X_1^2) - 1.5(X_2^2)^2 + \exp(-X_3^2)X_4^2 - .8\sin(X_5^2)\cos(X_3^2) + 2X_1^2 * X_5^2 + .9X_6^2 \quad (2.35)$$

- Let the effect of the third important pathway to be $F3$ and X_k^3 to be the kth important genes in the first important pathway.

$$\begin{aligned} F3 \;=\; & \cos(X_1^3) - 1.5(X_2^3)^2 + \exp(-X_3^3)X_4^3 - .8\sin(X_5^3)\cos(X_3^3) \quad (2.36) \\ & + 2X_1^3 * X_5^3 + .9X_6^3\sin(X_7^3) - .8\cos(X_6^3)X_7^3 + 2X_8^4 \end{aligned}$$

- Let the effect of the forth important pathway to be $F4$, and it can be generated similar to the third pathway.

- The response variables y would be

$$y = F1 + F2 + F3 + F4 + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1) \tag{2.37}$$

We choose weak information priors for our prior parameters. On 2.2.1, we already set $\alpha_0 = 0$ and $\boldsymbol{B}_0 = \mathbf{0}$. Moreover, we set $a_\alpha = 100$ and $a_\beta = 0.1$. The variance of the inverse gamma distribution exists when the shape parameter greater than 2 and we set the shape parameter, $\nu_0/2 = 3$, which is the smallest integer that greater than 2. The scale parameter, $\nu_0\sigma_0^2/2 = 0.6$, which form a weakly informative prior. μ decides the sparsity of the model. In simulation our goal is to select 2 to 6 important pathways for each iteration, but different simulation data set require slightly different μ to achieve this goal. We let the model change the μ value based on average pathways in the model in very 2000 iterations. The change of value of μ will increase the instability of the pathway selection, but once the suitable μ is selected, it will unlikely change. We choose $\eta = 0.08$, which controls the prior probability of gene selection depending on how many its related genes are selected. We set $\omega = .1$ as the prior for the probability of important pathway.

We have two different simulation scenarios: first one is that the important genes are only in the important pathways; the other one is that the important genes are not only in the important pathways, but the subset of them are in other pathways. The detail of the scenarios setting is as:

- Simulation 1 We make sure the 27 important genes only appear in the four

	TP	FP	TN	FN
True	4	0	46	0
Est.	3.24	0.26	45.74	0.76
Est. by [Stingo et al., 2011]	2.46	10.78	35.22	1.54

Table 2.1: Comparison between our model and [Stingo et al., 2011] model in simulation 1

important pathways. The sample size $n = 100$. We replicate the simulation 50 times, generating 50 different data sets. $\mu = -2.8$ is the initial value.

- Simulation 2 Similar to simulation, except we add subset of the important genes to 3 different pathways and $\mu = -5$ is the initial value.

We compare to result with [Stingo et al., 2011] model in the Simulation 1, table(2.1). On average, in our method, we choose 3.42 out of 4 important pathways and that is almost 1 more than [Stingo et al., 2011] model. Our model have 0.26 false pathway selection, which is much less than theirs(10.78). In 50 different simulations, our model selected none of the 2 fake important pathways. For each selected important pathways, at least one of the important genes are selected and none of the non-important genes are selected. Our model does not select any of the fake important pathways during the 50 different simulations; while [Stingo et al., 2011] model selects one of the fake important pathways frequently. This is a very good advantage of our model, because model does not select those mutated genes, when mutation happens during cell reproduction process but not related to the metastasis.

Table(2.2) shows the result of simulation 2. Our model selected 3.02 out of 4 important pathways on average, which is slightly lower than simulation 1, while the false selection becomes 1.02, which is greater than the simulation 1. Most of the false

	TP	FP	TN	FN
True	4	0	46	0
Est.	2.92	1.04	44.96	1.08
Est. by [Stingo et al., 2011]	1.78	11.12	34.88	2.22

Table 2.2: Comparison between our model and [Stingo et al., 2011] model in simulation 2

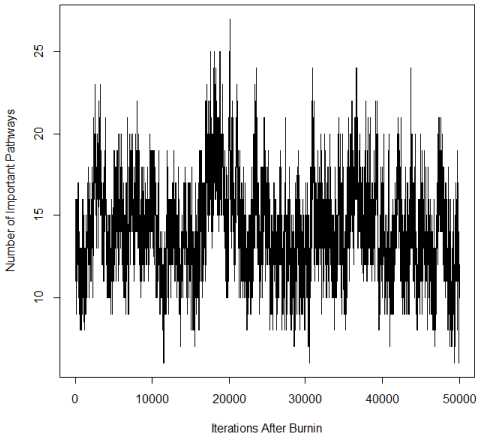
selections are related to non-important pathways with important genes. [Stingo et al., 2011] model also performs worse than the simulation 1.

2.5 Application

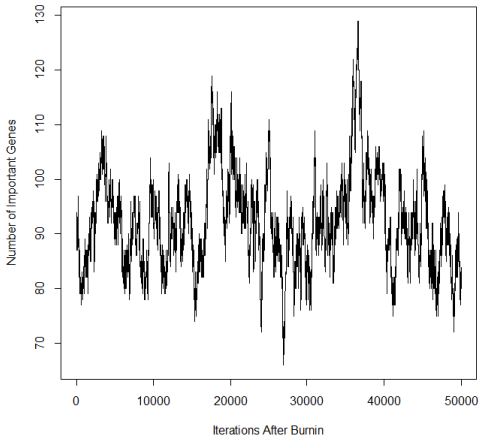
We used the [van 't Veer et al., 2002] breast cancer microarray data. In the data set there are 337 patients with 24481 microarray probes. We focus on 54 lymph-node-negative stage 0 patients. 21 of them developed distant metastasis and the rest of them are censored. our goal is to identify the pathways and genes related to breast cancer distant metastasis.

Follow [Troyanskaya et al., 2001] imputation method, we applied a 10-nearest neighbor algorithm to impute missing gene expressions data. The pathway and gene mapping information are obtained from Kyoto Encyclopedia of Genes and Genomes(KEGG). The gene identifiers are Entrez Gene in KEGG and GenBank accession in breast cancer microarray data. We convert the GenBank accession to Entrez Gene by using Gene ID converter *http://idconverter.bioinfo.cnio.es/*. We then were able to match the gene expression data with the pathway and gene mapping information. A total of 243 pathways and 4363 genes were included in this study.

We set $\omega = .1$ as the prior for the probability of important pathway, and $\mu = -1.7$



(a) The number of selected pathways after burnin



(b) The number of selected genes after burnin

Figure 2.1: The trace plots for the number of selected pathways and selected genes

and $\eta = 0.04$ for the gene selection. We choose weak information priors for our prior parameters. We set $\alpha_0 = 0$, $\boldsymbol{B}_0 = \mathbf{0}$ section(2.2.1), and $a_\alpha = 1000$, $a_\beta = 0.1$. Similar to the simulation study, the variance of the inverse gamma distribution exists when the shape parameter greater than 2 and we set the shape parameter, $\nu_0/2 = 3$, which is the smallest integer that greater than 2. The scale parameter, $\nu_0\sigma_0^2/2 = 0.6$, which form a weakly informative prior.

We used 100,000 iterations with burnin 50,000 iterations. Figure(2.1) shows the number of selected pathways and genes after burnin. The number selection pathways for each iteration mostly is between 10 and 20; while the number selection genes is always between 80 and 110.

Figure(2.2) shows that the 9 selected pathways which have highest posterior probabilities. We chose 0.4 as threshold, because there is a gap between 0.3 and 0.4 in the figure. Figure(2.3) shows the posterior probabilities of selected genes. We also

chose 0.4 as the threshold, and we got 94 important genes Table2.3. From literature search, we noticed that at least one of the important genes in each selected pathways are related to breast cancer or breast cancer metastasis.

Let us look at some of the important pathways:

- Steroid Hormone Biosynthesis pathway Figure(2.4). Steroid hormones belong to the group of chemical compounds known as steroids. These compounds are biologically synthesized by several organs of the human as well as other animals and they perform essential functions to maintain homeostasis. These functions include control of metabolism, inflammation, immune functions, salt and water balance, development of sexual characteristics, and the ability to cope with illness and injury. Such functional activities of steroid hormones require a strict balance of their synthesis to assure appropriate host response. Any abnormal changes in the biosynthetic pathway for steroid hormones can lead to imbalance of the hormonal level in the body. A consequence of such an event will be the abnormality in cellular function and abnormal growth. Postmenopausal women have altered level of steroid hormones and are more susceptible to develop breast cancer. Studies have shown that higher blood levels of testosterone may increase the risk of breast cancer in postmenopausal women. Furthermore, some evidence suggests that higher blood levels of testosterone may also increase breast cancer risk in premenopausal women. Estrogen plays a critical role in hormone-receptor-positive breast cancer growth. These findings underscore the importance of the association of several enzymes and other bioactive molecules in the steroid hormone biosynthetic pathway for analyzing the possibility of the involvement of some of these molecules in breast cancer growth and spread.

Important Pathways	Important Genes(Entrez)						Total number of Genes
Steroid Hormone Biosynthesis	10720	10941	1109	1543	1545	1551	
	1576	1577	1584	1586	1588	3283	24
	3284	3290	3291	3293	3294	6715	
	7364	7365	7366	7367	8630	8644	
Tyrosine Metabolism	124	125	126	127	128	130	13
	131	218	220	221	222	2954	
	4128						
Glutathione Metabolism	2678	2877	2878	2879	2882	2937	
	2938	2939	2940	2941	2946	2947	24
	2948	2949	2950	2952	2953	2954	
	373156	4257	4258	4259	51060	9446	
Arachidonic Acid Metabolism	1558	1562	1571	1579	2678	2877	11
	2878	2882	8644	873	874		
Folate Biosynthesis	10170	10720	10941	124	125	126	
	127	128	130	131	1543	1544	
	1548	1551	1553	1558	1562	1576	
	1577	1579	1592	216	50700	53630	37
	54884	56603	5959	6121	7364	7365	
	7366	7367	8228	8608	8854	9227	
	9249						
Retinol Metabolism	10720	10941	7364	7365	7366	7367	6
Porphyrin & Chlorophyll	10720	10941	1109	124	125	126	
Metabolism	127	128	130	131	1543	1544	
	1545	1548	1551	1553	1558	1562	
	1565	1571	1572	1576	1577	2052	
	218	220	221	222	22977	27294	57
	2938	2939	2940	2941	2946	2947	
	2948	2949	2950	2952	2953	2954	
	3290	3291	373156	4257	4258	4259	
	7364	7365	7366	7367	8574	8644	
	873	874	9446				
Drug Metabolism -	10720	10941	124	125	126	127	
Cytochrome P450	128	130	131	1544	1548	1551	
	1553	1558	1562	1565	1571	1576	
	1577	218	220	221	222	2326	
	2327	2328	2329	2330	2938	2939	50
	2940	2941	2946	2947	2948	2949	
	2950	2952	2953	2954	373156	4128	
	4257	4258	4259	7364	7365	7366	
	7367	9446					
Drug Metabolism -	10720	10941	1548	1551	1553	1576	11
Other Enzymes	1577	7364	7365	7366	7367		

Table 2.3: Summation of Important Pathways and Genes

- Tyrosine Metabolism Figure(2.5). Tyrosine is an amino acid that plays an essential role in the metabolism. Tyrosine metabolism is crucial component to breast cancer and other cancer as well.

- Glutathione Metabolism Figure(2.6). Glutathione plays important roles in antioxidant defense, nutrient metabolism, and regulation of cellular events including gene expression, DNA and protein synthesis, cell proliferation and apoptosis, signal transduction, cytokine production and immune response, and protein glutathionylation.Physiological relevance of glutathione makes the pathway and components of glutathione metabolism a vital part of cancer-associated studies.

- Arachidonic Acid Metabolism Figure(2.7). Arachidonic acid and certain other polyunsaturated fatty acids may be transformed into prostaglandins (PG) by the enzyme prostaglandin endoperoxide synthase (PES). Level of prostaglandin E2 is elevated in malignant human breast tissue. An increase of inflammatory component may have a direct consequence in the development of inflammatory breast cancer, a deadliest form of breast cancer.

Additionally, , the boxes represent the protein compound coded by genes. The red number above the boxes are the important genes known to be related to breast cancer that our model found. The number is Entrez. In the lower center of the graph, Entrez gene 1588 is called CYP19A1. CYP19A1 gene codes for an enzyme called aromatase. One of the critical functions of this aromatase is to convert testosterone to estradiol, a form of estrogen [Meinhardt and Mullis, 2002]. Increased activity of aromatase has been linked to breast cancer due to abundance of estradiol accumulation in the cancer cells [Chen, 1998]. Estradiol increases tumor growth in breast cancer by promoting

cell proliferation [L. et al., 2013]. In cancer cells it binds to estrogen receptor which subsequently activates a group of hormone-responsive genes that promote DNA synthesis and cell proliferation [DeMayo et al., 2002]. Therefore over-active CYP19A1 gene in breast cancer cells has a poor prognosis for patients due to over-expression of aromatase-driven estradiol synthesis. Consequently, specific aromatase inhibitors have been found to be useful in the treatment of breast cancer [M. et al., 1999].

Some of selected pathways and genes are known to be related to breast cancer and breast cancer metastasis. Some of genes have no information according to the literature search. From the biology research point of view, the biologists can use the information from the posterior probability to prioritize the pathways and genes for future research.

2.6 Discussion

In this chapter, we have proposed a Bayesian variable selection model with prior biological information from pathways and genes relationship. We have considered AFT to model the relationship between covariates and augmented failure time. We have adopted an RKHS-based method to nonparametrically model the pathways effect and have built into the model a variable selection mechanism that selects genes and pathways simultaneously. Simulation studies and breast cancer microarray data have been used to illustrate the proposed method.

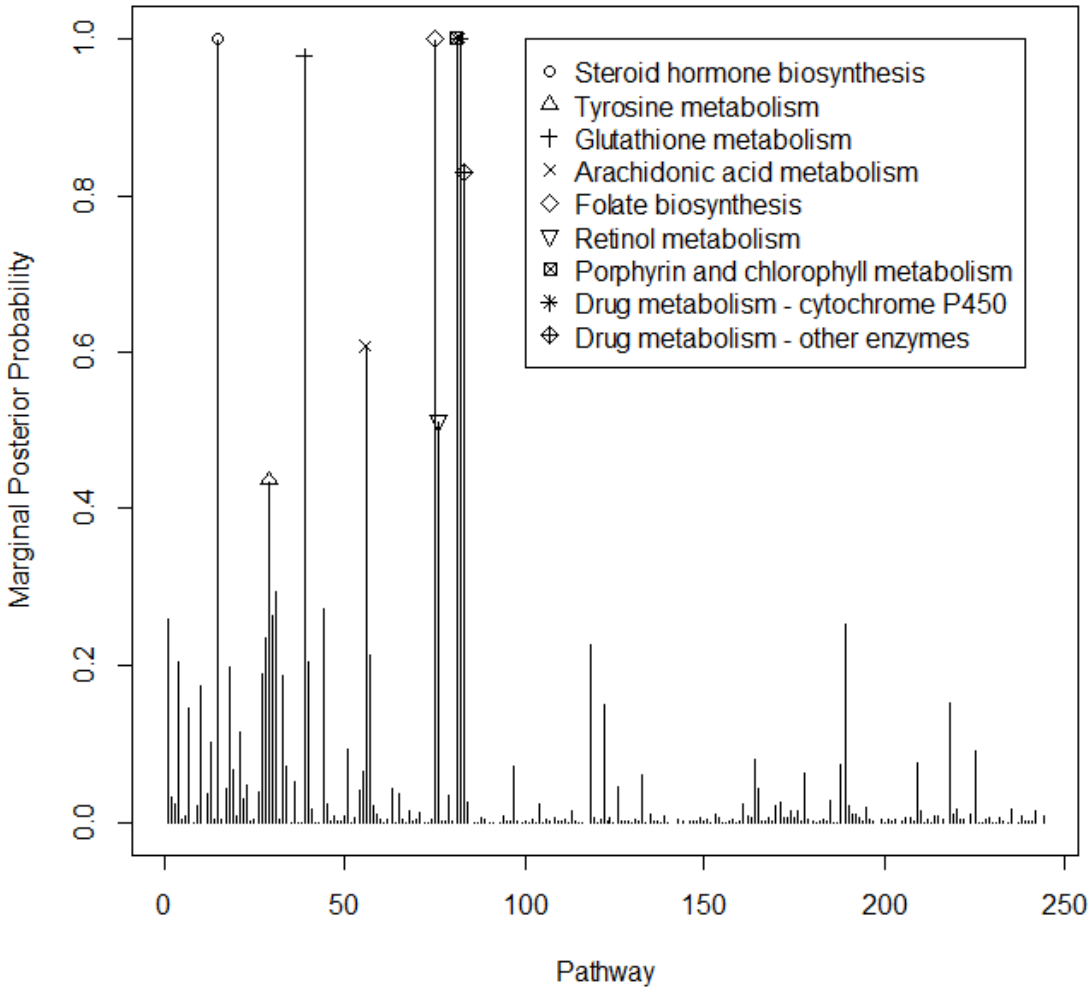


Figure 2.2: Marginal posterior probabilities for pathway selection, $p(\phi_j|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}) > 0.4$

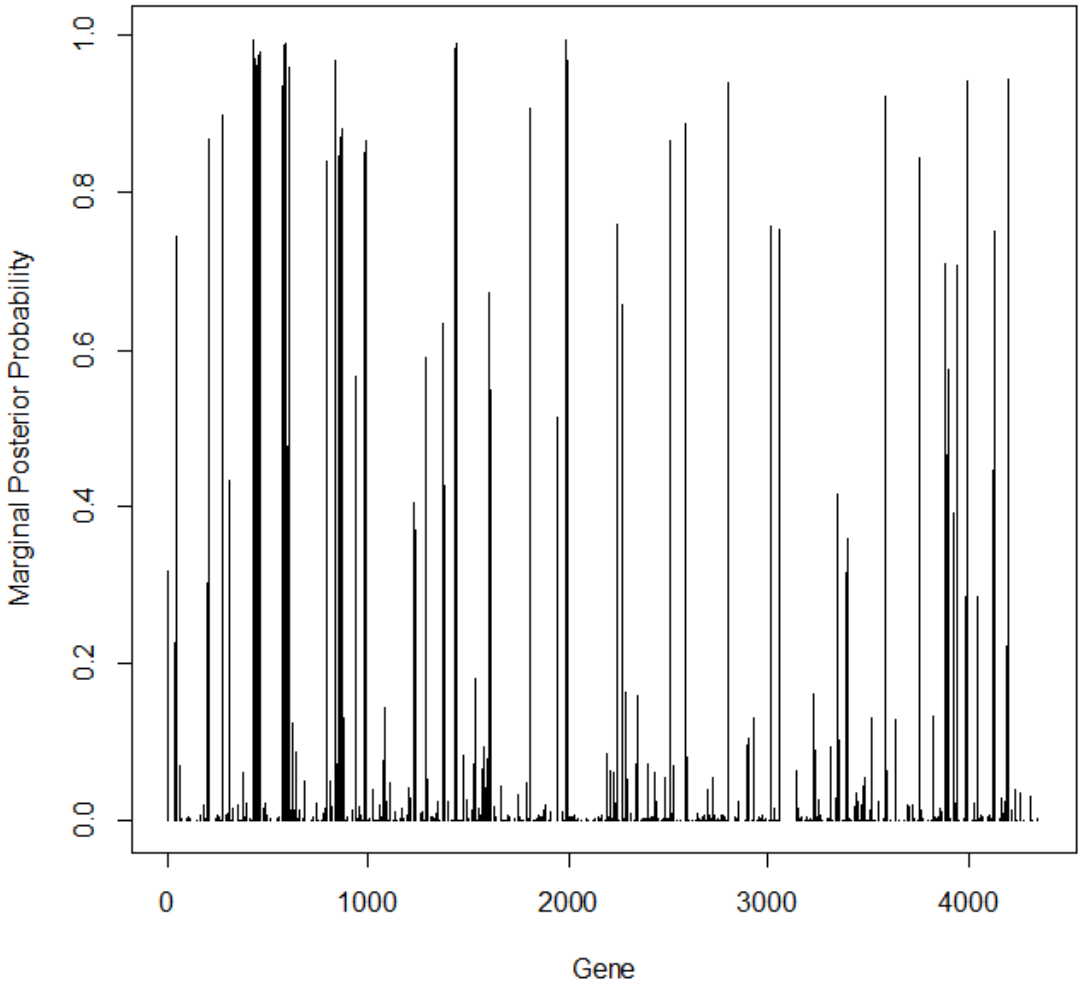


Figure 2.3: Marginal posterior probabilities for gene selection, $p(\gamma_j|\boldsymbol{Y}, \boldsymbol{X}, \mathcal{K}) > 0.4$

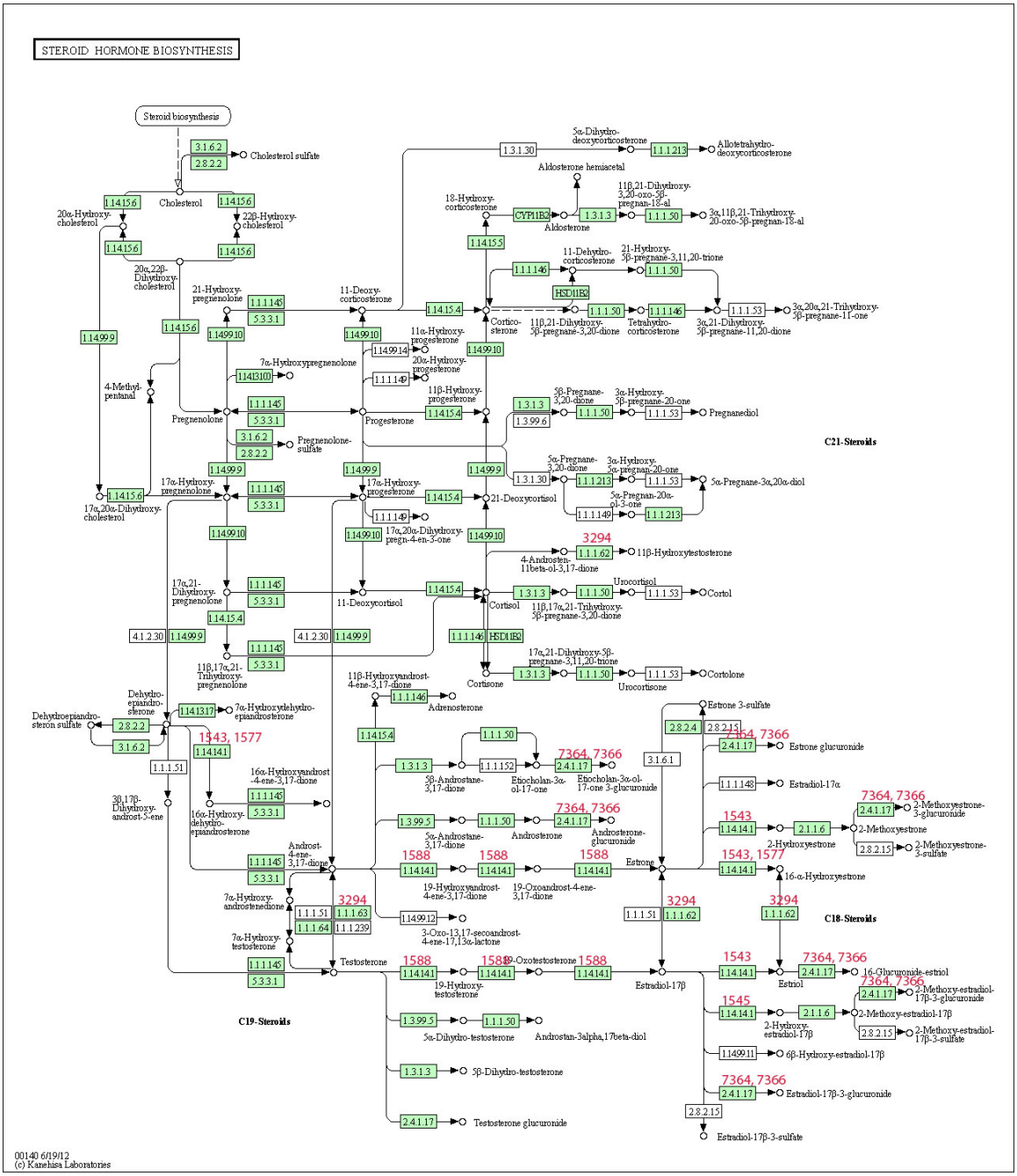


Figure 2.4: Steroid hormone biosynthesis pathway with important genes that related to breast cancer

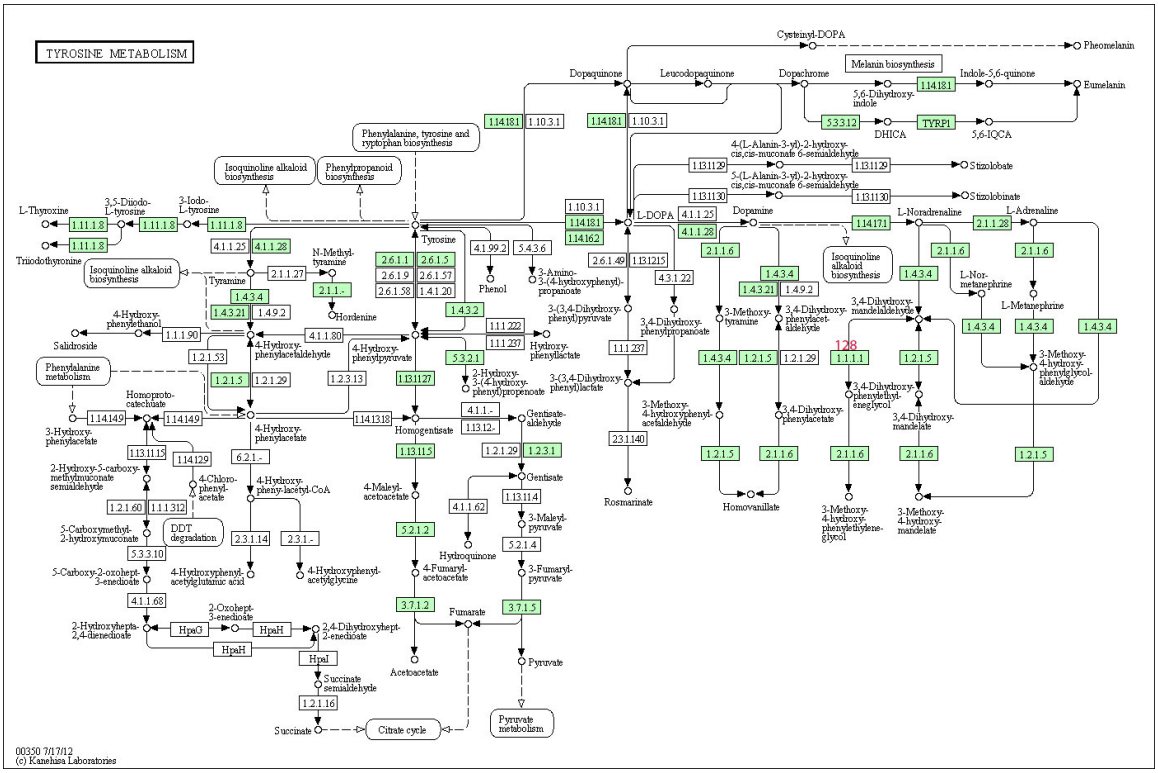


Figure 2.5: Tyrosine Metabolism pathway with important genes that related to breast cancer

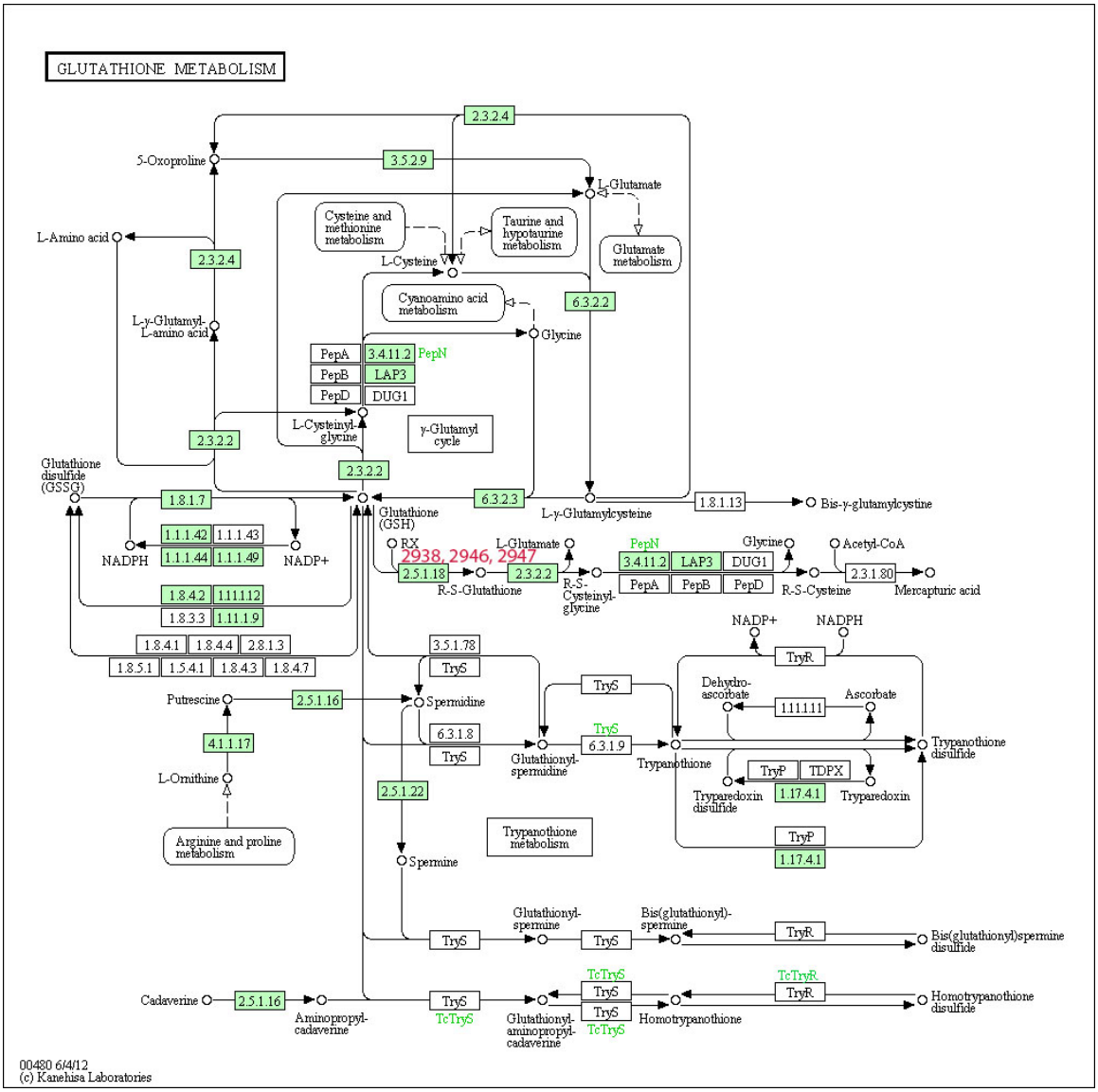


Figure 2.6: Glutathione Metabolism pathway with important genes that related to breast cancer

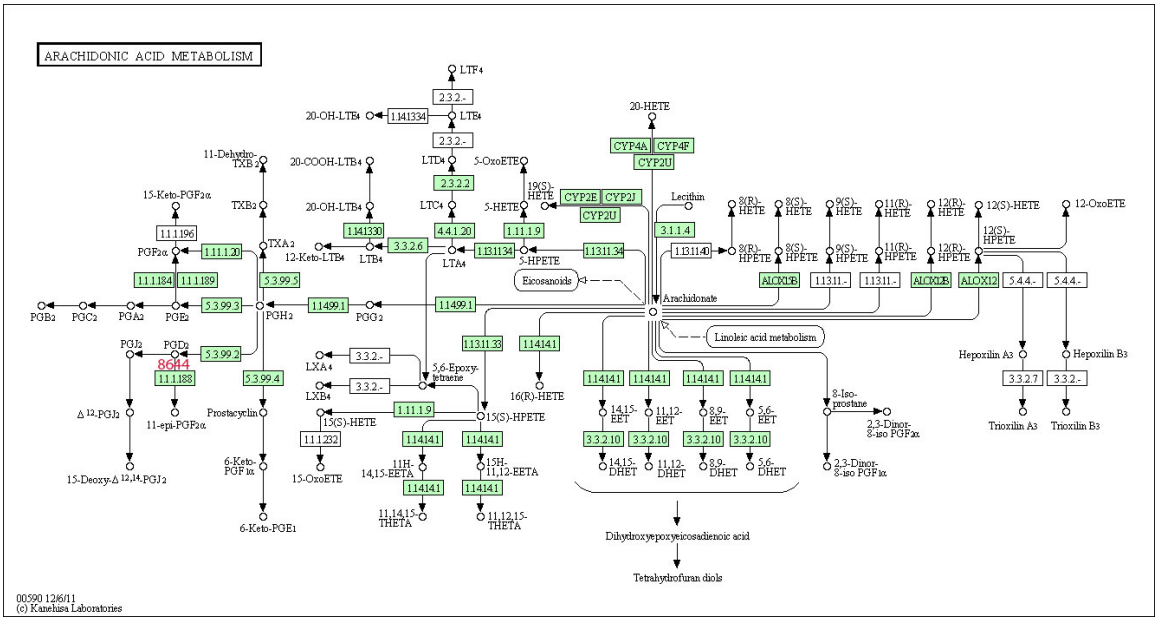


Figure 2.7: Arachidonic acid metabolism pathway with important genes that related to breast cancer

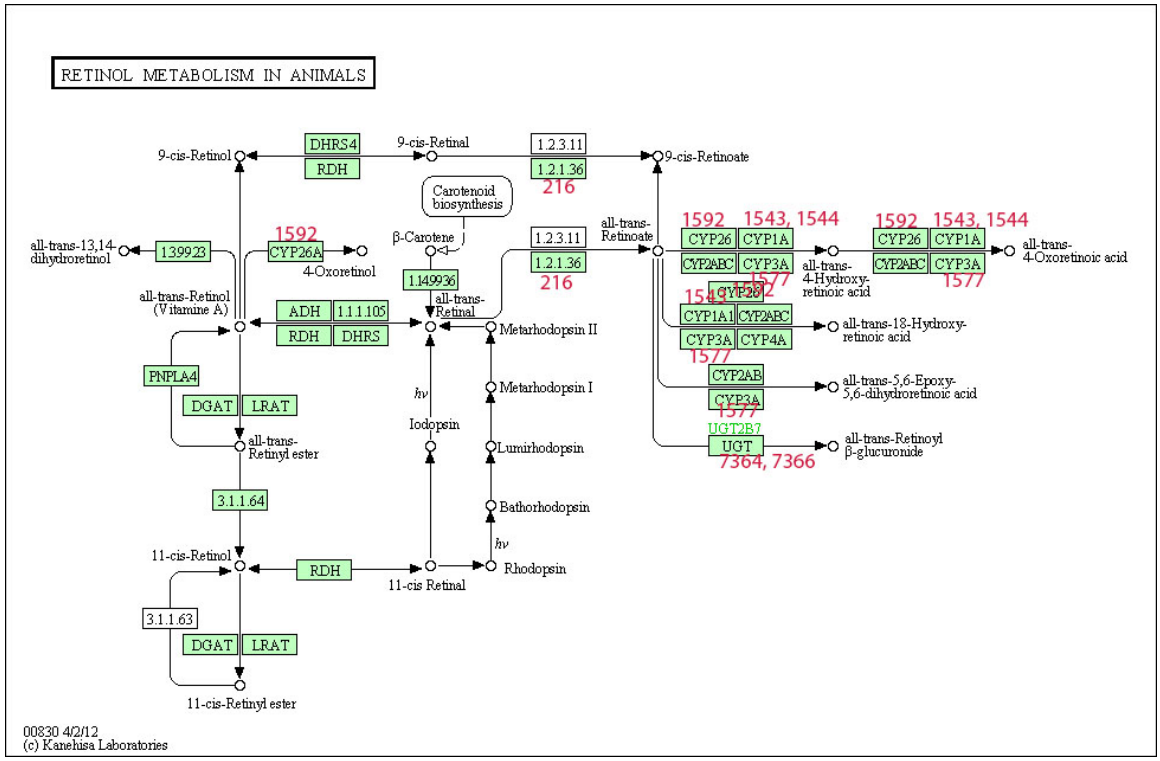


Figure 2.8: Retinol metabolism pathway with important genes that related to breast cancer

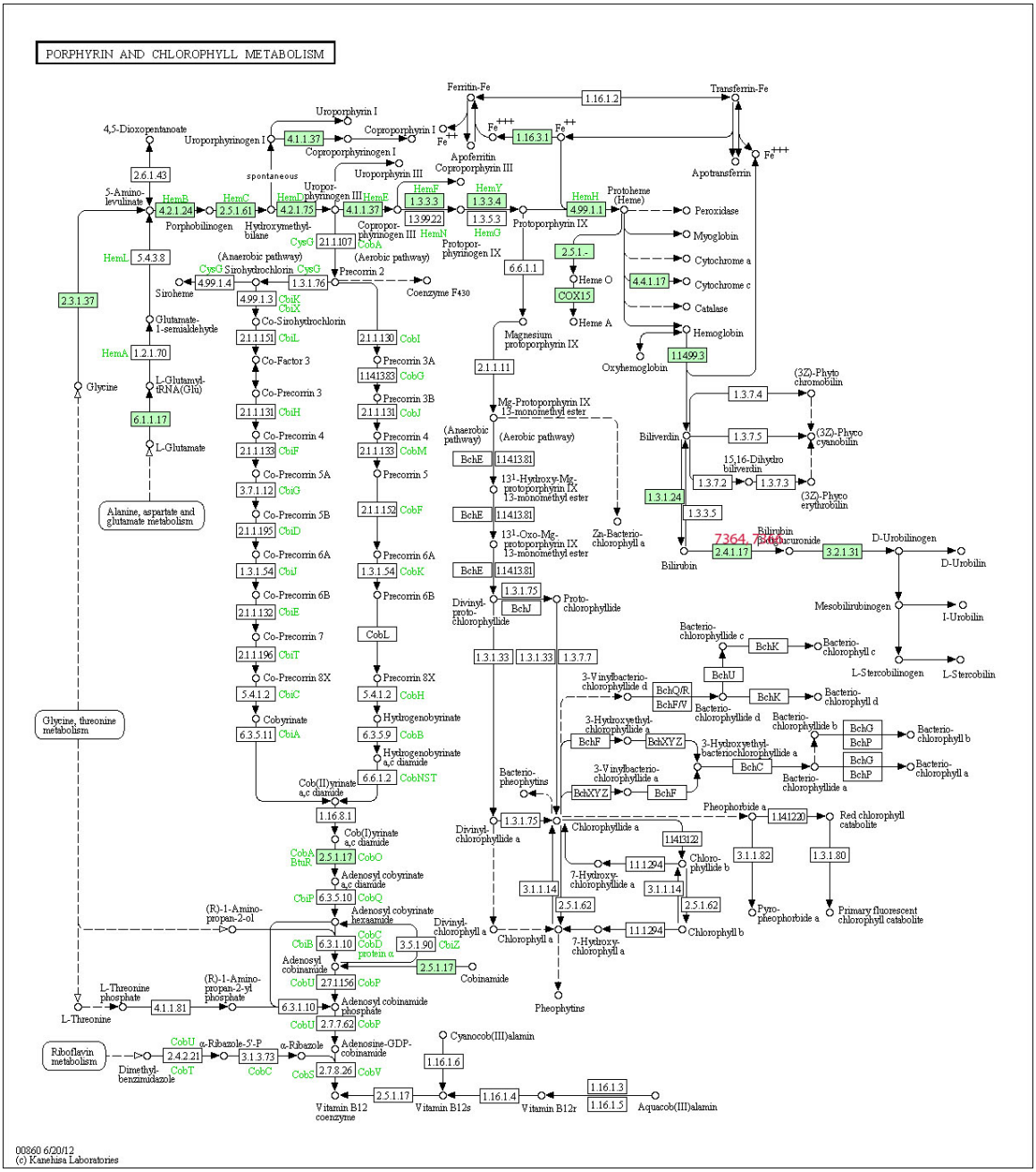


Figure 2.9: Porphyrin and chlorophyll metabolism pathway with important genes that related to breast cancer

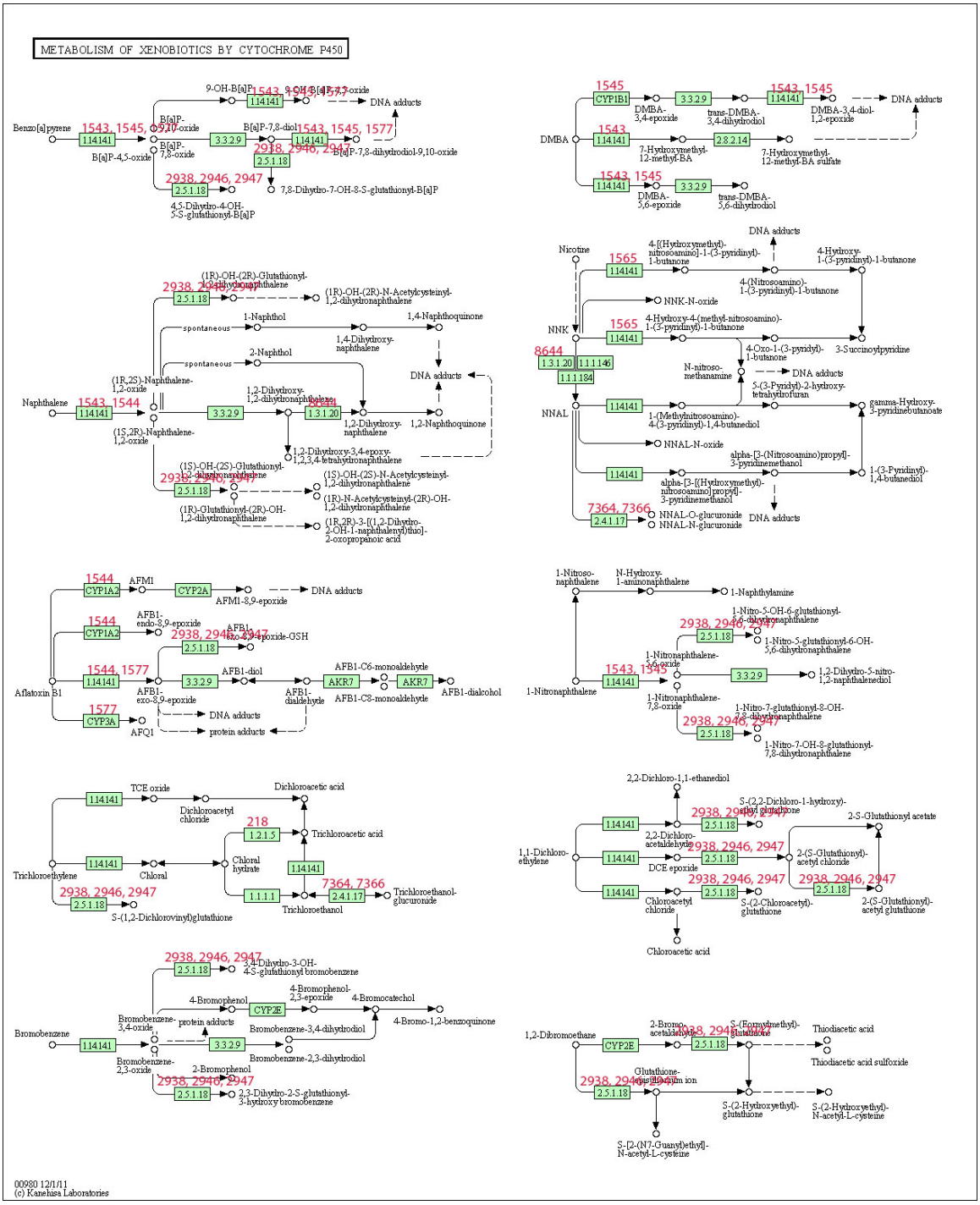


Figure 2.10: Metabolism of xenobiotics by cytochrome P450 pathway with important genes that related to breast cancer

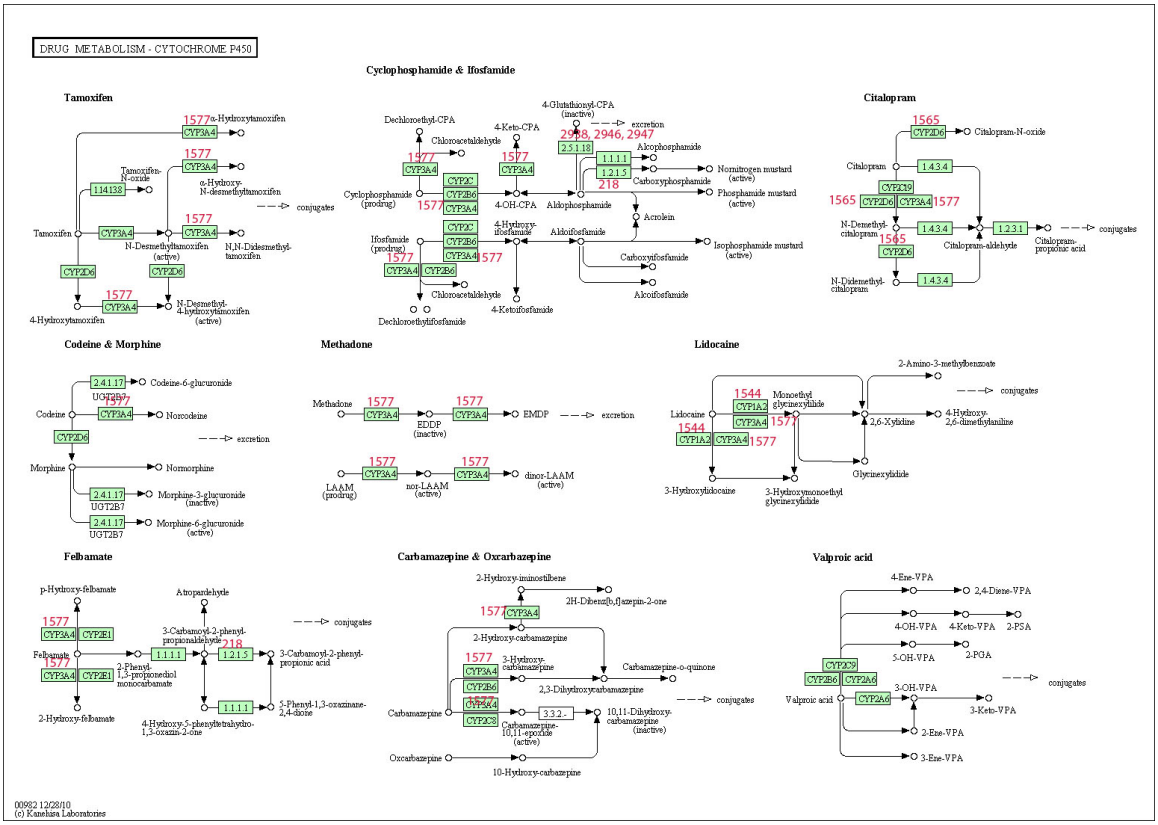


Figure 2.11: Drug metabolism - cytochrome P450 pathway with important genes that related to breast cancer

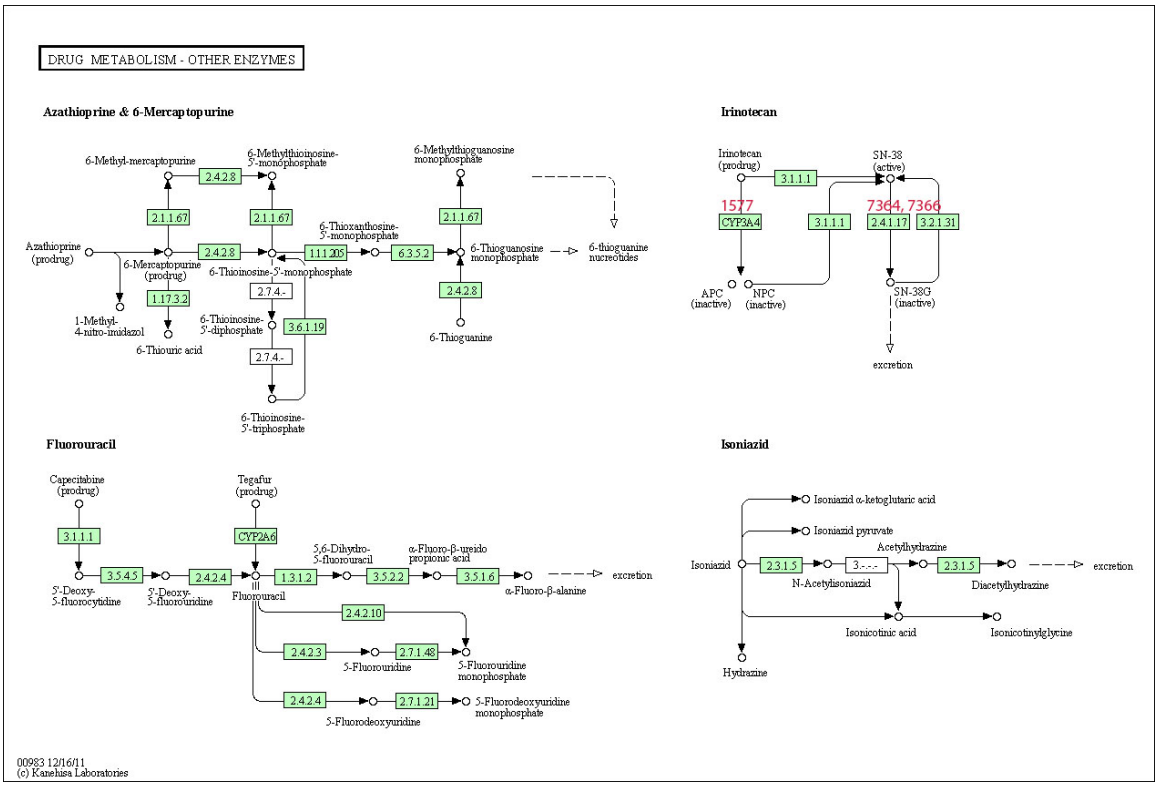


Figure 2.12: Drug metabolism - other enzymes pathway with important genes that related to breast cancer

Chapter 3

Bayesian Elastic-Net and Fused Lasso for Semiparametric Structural Equation Models

3.1 Introduction

Structural equation models (SEMs) are a well developed statistical tool that are useful for datasets with latent variables, which are not observed directly, but are estimated from observed variables. SEMs consist of two parts, a measurement equation and a structural equation. The measurement equation investigates the relationship between the unobservable latent variables and observed manifest variables; whereas the structural equation measures the relationship between the endogenous latent variables and exogenous latent variables, and the covariates. The primary research interest is typically the structural equation. SEMs are commonly used in Psychology, Biology, etc., where latent variables are common. For example, see [Martens, 2005], [Lee and Zhu,

2000], [Liu et al., 2008], etc.

Traditionally, SEMs assume linear relationships among latent variables in the structural equation. [Kenny and Judd, 1984] introduced a nonlinear SEM (NSEM) that extended this methodology to include relationships such as interaction and quadratic terms. [Lee, 2007] generalized NSEM to include a broader set of nonlinear relationships. However, misspecification of the parametric form at the latent level, whether the model is linear or nonlinear, can result in very poor estimation. Recently, some semiparametric approaches have been developed. [Bauer, 2005], [Fahrmeir and Raach, 2007], [Guo et al., 2012], etc used basis expansions to approximate the nonlinear structural relationships using semiparametric SEM (SSEM). To achieve simultaneous estimation and model selection [Guo et al., 2012] applied the Bayesian Lasso method to the SSEM. The Bayesian Lasso performs well in SSEM, however, it ignores correlation of the features which leads to inefficient parameter estimation and model selection.

This is concerning when cubic splines are used, because they tend to be highly correlated since each column is a transformed version of the same variables [Keele, 2008]. This chapter accesses this correlation by considering fused Lasso and elastic net. The fused lasso has been shown to be a good method for multiple linear regression when the features have a natural order, specifically when there is side by side correlation [Tibshirani et al., 2005a]. [Zou and Hastie, 2005], show that elastic net often outperforms regular Lasso in both real world data set and simulation studies, and they still have a similar sparse representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together.

3.2 Model

3.2.1 Semiparametric Structural Equation Models

Semiparametric structural equation models consist of two parts, a measurement equation and a structural equation. For a random sample of n independent subjects, the measurement equation defines the relationship between the observed $p \times 1$ vector of manifest variables $\boldsymbol{y}_i$ and the unobserved $q \times 1$ vector of latent variables $\boldsymbol{w}_i$ as follows:

$$\boldsymbol{y}_i = \boldsymbol{A}\boldsymbol{c}_i + \boldsymbol{\Lambda}\boldsymbol{w}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, ..., n$$

where $\boldsymbol{c}_i$ is an $r \times 1$ vector of known functions of the $s \times 1$ vector of fixed covariates $\boldsymbol{x}_i$, $\boldsymbol{A}$ and $\boldsymbol{\Lambda}$ are unknown parameter matrices, $\boldsymbol{\epsilon}_i$ is a $p \times 1$ vector of measurement errors.

The latent variable $\boldsymbol{w}_i$ is written in two parts, a $q_1 \times 1$ vector of endogenous latent variables $\boldsymbol{\eta}_i$ and a $q_2 \times 1$ vector of exogenous latent variables $\boldsymbol{\xi}_i$, i.e. $\boldsymbol{w}_i = (\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$. Then the structural equation, which defines the relationship between the exogenous and endogenous latent variables, is

$$\boldsymbol{\eta}_i = \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{F}(\boldsymbol{x}_i, \boldsymbol{\xi}_i) + \boldsymbol{\zeta}_i, \quad i = 1, 2, ..., n, \tag{3.1}$$

where $\boldsymbol{\zeta}_i$ is a vector of residuals and $\boldsymbol{F}(\boldsymbol{x}_i, \boldsymbol{\xi}_i)$ is a vector of unknown functions of the covariates $\boldsymbol{x}_i$ and exogenous latent variables $\boldsymbol{\xi}_i$.

For this model, we require the following assumptions:

- $\boldsymbol{\epsilon}_i$ are independently distributed as $N(\boldsymbol{0}, \boldsymbol{\Psi}_\epsilon)$ with $\boldsymbol{\Psi}_\epsilon = diag(\psi_{\epsilon 1}, \psi_{\epsilon 2}, ..., \psi_{\epsilon p})$.

- $\boldsymbol{w}_i$ and $\boldsymbol{\epsilon}_i$ are independent, and $\boldsymbol{w}_i$ are independently distributed.
- $\boldsymbol{\zeta}_i$ follows $N(\mathbf{0}, \boldsymbol{\Psi}_\zeta)$ with $\boldsymbol{\Psi}_\zeta = diag(\psi_{\zeta 1}, \psi_{\zeta 2}, ..., \psi_{\zeta q_1})$.
- $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are independently distributed, and $\boldsymbol{\xi}_i$ follows $N(\mathbf{0}, \boldsymbol{\Phi})$
- $\Pi_0 = I - \Pi$ is nonsingular and $|\Pi_0|$ is independent of the elements of Π.

Theoretically, $\boldsymbol{F}(\boldsymbol{x}_i, \boldsymbol{\xi}_i)$ can be any linear or nonlinear function of of $\boldsymbol{x}_i$ and $\boldsymbol{\xi}_i$ with or without interaction terms like $\xi_{i1}\xi_{i2}$. In this project, we consider a nonparametric structural equation similar to [Guo et al., 2012] and we approximate the nonparametric function $\boldsymbol{F}(\boldsymbol{x}_i, \boldsymbol{\xi}_i)$ using basis expansions. The structural equation 3.1, in general case, can be represented as

$$\boldsymbol{\eta}_i = \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{B}\boldsymbol{H}(\boldsymbol{x}_i, \boldsymbol{\xi}_i) + \boldsymbol{\zeta}_i,$$

where $\boldsymbol{H}(\boldsymbol{x}_i, \boldsymbol{\xi}_i)$ is an $N_H \times 1$ vector of basis functions, and $\boldsymbol{B}_{q_i \times N_H}$ is the coefficient parameter matrix associated with $\boldsymbol{H}(\boldsymbol{x}_i, \boldsymbol{\xi}_i)$.

To illustrate the structural equation, consider a simple example with $\Pi = 0$, one covariate, one endogenous and two exogenous latent variables. Any function $\boldsymbol{F}(\boldsymbol{x}_i, \boldsymbol{\xi}_i)$ can be decomposed into two parts, functions with only one variable as f_1, f_2 and f_3, which could be constant, and functions with interactions as f_{12}, f_{13} and f_{23}, which must be functions of both two parameters, i.e.,

$$\begin{aligned}
\eta_i &= F(x_i, \xi_{i1}, \xi_{i2}) + \zeta_i \\
&= f_1(x_i) + f_2(\xi_{i1}) + f_3(\xi_{i2}) + f_{12}(x_i, \xi_{i1}) + f_{13}(x_i, \xi_{i2}) \\
&\quad + f_{23}(\xi_{i1}, \xi_{i2}) + f_{123}(x_i, \xi_{i1}, \xi_{i2}) + \zeta_i,
\end{aligned}$$

It indicates that for modeling f_1, f_2 and f_3, a linear basis expansion can be used, such as piecewise polynomials, natural cubic splines, etc. In such cases,

$$f_j(.) = \sum_{m_j=1}^{M_j} \beta_{jm_j} h_{jm_j}(.), \quad j = 1, 2, 3$$

where $\{h_{jm_j}(.), m_j = 1, ..., M_j\}$ are basis functions. For modeling f_{12}, f_{13} and f_{23}, tensor product basis expansion can be used as follows:

$$f_{kl}(.,.) = \sum_{m_k=1}^{M_k} \sum_{m_l=1}^{M_l} \beta_{m_k m_l}^{(kl)} h_{km_k}(.) h_{lm_l}(.), \quad k, l = 1, 2, 3.$$

3.2.2 Bayesian Fused Lasso in the Semiparametric SEM

The unknown parameters in the measurement equation are $\boldsymbol{\Lambda}_y = (\boldsymbol{A}, \boldsymbol{\Lambda})$ and $\boldsymbol{\Psi}_{\boldsymbol{\epsilon}}$, in structural equation, the unknown parameters are $\boldsymbol{\Lambda}_w = (\boldsymbol{\Pi}, \boldsymbol{B})$, $\boldsymbol{\Psi}_{\boldsymbol{\zeta}}$ and $\boldsymbol{\Phi}$. Some elements of $\boldsymbol{\Lambda}_y$ must be fixed for identifiability purposes.

For the measurement equation, an index matrix $\boldsymbol{M} = (m_{kj})_{p\times(r+q)}$ is created as follows [Lee and Zhu, 2000],

$$m_{kj} = \begin{cases} 1 & \text{if} \quad \lambda_{ykj} \text{ is unknown} \\ 0 & \text{otherwise} \end{cases}$$

where λ_{ykj} is the kj-th element of $\boldsymbol{\Lambda}_y$. If there is an unknown parameter in k-th row of $\boldsymbol{\Lambda}_y$ for $k = 1, ..., p$, this means that $r_{yk} = \sum_{j=1}^{r+q} m_{kj} > 0$. We denote $\boldsymbol{\Lambda}_{yk}^*$ as the

$r_{yk} \times 1$ vector of unknown parameters and specified a conjugate prior for $\{\boldsymbol{\Lambda}^*_{yk}, \psi_{\epsilon k}\}$,

$$\boldsymbol{\Lambda}^*_{yk} | \psi_{\epsilon k} \sim N_{r_{yk}}(\mu^*_{0yk}, \psi_{\epsilon k} \boldsymbol{H}^*_{0yk}) \tag{3.2}$$

$$\psi^{-1}_{\epsilon k} \sim Gamma(\alpha_{0\epsilon k}, \beta_{0\epsilon k}) \tag{3.3}$$

where μ^*_{0yk}, $\boldsymbol{H}^*_{0yk}$, $\alpha_{0\epsilon k}$ and $\beta_{0\epsilon k}$ are hyperparameters.

For the structural equation, let Λ_{wh}be the h-th row of $\boldsymbol{\Lambda}_w$ where $h = 1, ..., q_1$. As mentioned earlier, we assigned Bayesian fused Lasso priors for each Λ_{wh} and assigned the inverse-Wishart prior for $\boldsymbol{\Phi}$.

$$\boldsymbol{\Lambda}_{wh} | \psi_{\zeta h}, \boldsymbol{\tau}_{\Lambda_{wh}}, \boldsymbol{\upsilon}_{\Lambda_{wh}} \sim N(0, \psi_{\zeta h} \boldsymbol{D}_{\Lambda_{wh}}),$$

$$\psi^{-1}_{\zeta h} \sim Gamma(\alpha_{0\zeta h}, \beta_{0\zeta h}),$$

$$\pi(\boldsymbol{\tau}^2_{\Lambda_{wh}}) \propto \prod_{j=1}^{q_1} \frac{\lambda^2_{\Pi_h}}{2} e^{-\lambda^2_{\Pi_h} \tau^2_{\Pi_h j}/2} \prod_{j=1}^{N_X} \frac{\lambda^2_{B_{1h}}}{2} e^{-\lambda^2_{B_{1h}} \tau^2_{B_{1j}j}/2} \prod_{j=1}^{N_T} \frac{\lambda^2_{B_{2h}}}{2} e^{-\lambda^2_{B_{2h}} \tau^2_{B_{2h}j}/2},$$

$$\pi(\boldsymbol{\upsilon}^2_{\Lambda_{wh}}) \propto \prod_{j=1}^{N_T} \frac{\mu^2_{B_{2h}}}{2} e^{-\mu^2_{B_{2h}} \upsilon^2_{B_{2h}j}/2},$$

$$\boldsymbol{\Phi} \sim IW(\boldsymbol{R}_0, \rho_0),$$

where N_h is the number of non-constant spline basis functions, and $N_h = N_x + N_T$, where N_x is the number of basis functions related x's, and N_T is the number of basis functions related to exogenous latent variables. $\boldsymbol{B}_h = (\boldsymbol{B}^T_{1h}, \boldsymbol{B}^T_{2h})^T$, where $\boldsymbol{B}_{1h}$ are the coefficients corresponding to the x's and $\boldsymbol{B}_{2h}$ are the coefficients corresponding to the exogenous latent variables. $\boldsymbol{\tau}_{\Lambda_{wh}}$ and $\boldsymbol{\upsilon}_{\Lambda_{wh}}$ are mutually independent, and the covariance matrix $\boldsymbol{D}^{-1}_{\Lambda_{wh}}$ is a diagonal tridiagonal mixed matrix.

$\boldsymbol{D}^{-1}_{\Lambda_{wh}} = diag(\boldsymbol{D}^{11}_{q_1\times q_1}, \boldsymbol{D}^{22}_{N_X\times N_X}, \boldsymbol{D}^{33}_{N_T\times N_T})$, where $\boldsymbol{D}^{11}_{q_1\times q_1}$ is a diagonal matrix with

$$\text{main diagonal} = \Big\{\frac{1}{\tau^2_{\Pi_{hj}}}, j = 1, ..., q_1\Big\}$$

$\boldsymbol{D}^{22}_{N_X\times N_X}$ is also a diagonal matrix with

$$\text{main diagonal} = \Big\{\frac{1}{\tau^2_{B_{1hj}}}, j = 1, ..., N_X\Big\}$$

$\boldsymbol{D}^{33}_{N_T\times N_T}$ is a tridiagonal matrix with

$$\text{main diagonal} = \Big\{\frac{1}{\tau^2_{B_{2hj}}} + \frac{1}{\upsilon^2_{B_{2hj-1}}} + \frac{1}{\upsilon^2_{B_{2hj}}}, j = 1, ..., N_T\Big\}$$

$$\text{off diagonals} = \Big\{ -\frac{1}{\upsilon^2_{B_{2hj}}}, j = 1, ..., N_T - 1\Big\}$$

All the λ's are tuning parameters with gamma priors.

The extended Bayesian Fused Lasso prior has additional parameters, however, with the priors specified as above, it is straightforward to derive the full conditional distribution[Kyung et al., 2010]. As a result we can use MCMC methods to generate samples from the joint posterior distribution of parameters.

The model can be easily extended to the case where X's has side by side correlation. We only need to change $\boldsymbol{D}^{22}_{N_X\times N_X}$ to tridiagonal matrix with

$$\text{main diagonal} = \Big\{\frac{1}{\tau^2_{B_{1hj}}} + \frac{1}{\upsilon^2_{B_{1hj-1}}} + \frac{1}{\upsilon^2_{B_{1hj}}}, j = 1, ..., N_X\Big\}$$

$$\text{off diagonals} = \Big\{ -\frac{1}{\upsilon^2_{B_{1hj}}}, j = 1, ..., N_X - 1\Big\}$$

$$\pi(\boldsymbol{\upsilon}_{\Lambda_{wh}}) \propto \prod_{j=1}^{N_X} \frac{\mu^2_{B_{1h}}}{2} e^{-\mu^2_{B_{1h}} \upsilon^2_{B_{1h}j}/2} \prod_{j=1}^{N_T} \frac{\mu^2_{B_{2h}}}{2} e^{-\mu^2_{B_{2h}} \upsilon^2_{B_{2h}j}/2},$$

It is easy to derive the full conditional distribution and use MCMC methods to generate samples from the joint posterior distribution of parameters for this scenario as well.

3.2.3 Bayesian Elastic Net in the Semiparametric SEM

The measurement equation is exactly the same as in 3.2.2, however, for the structural equation, the priors become:

$$\boldsymbol{\Lambda}_{wh} | \psi_{\zeta h}, \boldsymbol{\tau}_{\Lambda_{wh}}, \boldsymbol{\upsilon}_{\Lambda_{wh}} \sim N(0, \psi_{\zeta h} \boldsymbol{D}_{\Lambda_{wh}}),$$

$$\psi^{-1}_{\zeta h} \sim Gamma(\alpha_{0\zeta h}, \beta_{0\zeta h}),$$

$$\pi(\boldsymbol{\tau}_{\Lambda_{wh}}) \propto \prod_{j=1}^{q_1} \frac{\lambda^2_{\Pi_h}}{2} e^{-\lambda^2_{\Pi_h} \tau^2_{\Pi_h j}/2} \prod_{k=1}^{N_G} \prod_{j=1}^{N_k} \frac{\lambda^2_{1B_h k}}{2} e^{-\lambda^2_{1B_h k} \tau^2_{B_h kj}/2}$$

$$\boldsymbol{\Phi} \sim IW(\boldsymbol{R}_0, \rho_0),$$

where $\boldsymbol{X}$ is reordered. Strongly correlated covariates are grouped together, so we have N_G blocks of $\boldsymbol{X}$'s, including one block for independent $\boldsymbol{X}$'s if any exists. And $k = 1, \cdots, N_G$. For block, k, N_k is the total number of members in the block. $\boldsymbol{D}_{\Lambda_{wh}}$ is a diagonal matrix with diagonal elements. If $\boldsymbol{X}$'s in the corresponding block k are correlated, the diagonal elements are $(\tau^{-2}_{B_h kj} + \lambda_{2B_h k})^{-1}$; if X's in the corresponding block k are independent, the diagonal elements are $\tau^2_{2B_h kj}$, in other words $\lambda_{2B_h k} = 0$. And similar to the Bayesian fused lasso, all the λ's have gamma priors. It is still straightforward to derive the full conditional distribution [Li and Lin, 2010], and

use MCMC methods to generate samples from the joint posterior distribution of parameters.

3.3 Posterior Distribution in the Semiparametric SEM

3.3.1 Posterior Distribution in the measurement equation

Using the conjugate prior for $\boldsymbol{\Lambda}_{yk}^*$ and $\psi_{\epsilon k}$ from 3.2 and 3.3, we can easily get the posterior distributions as:

$$\boldsymbol{\Lambda}_{yk}^*|rest \sim N_{r_{yk}}(\boldsymbol{H}_{yk}(\boldsymbol{H}_{0yk}^{*-1}\boldsymbol{\mu}_{0yk}^* + \boldsymbol{G}_{yk}\boldsymbol{y}_k^*), \psi_{\epsilon k}(\boldsymbol{H}_{0yk}^{*-1} + \boldsymbol{G}_{yk}\boldsymbol{G}_{yk}^T)^{-1}) \tag{3.4}$$

$$\psi_{\epsilon k}^{-1}|rest \sim Gamma(\alpha_{0\epsilon k} + n/2, \beta_{0\epsilon k} + \frac{1}{2}(\boldsymbol{y}_k^{*T}\boldsymbol{y}_k^* + \boldsymbol{\mu}_{0yk}^{*T}\boldsymbol{H}_{0yk}^{*-1}\boldsymbol{\mu}_{0yk} - \boldsymbol{\mu}_{yk}^T\boldsymbol{H}_{yk}^{-1}bm\mu_{yk})) \tag{3.5}$$

where $\boldsymbol{G}_y = (\boldsymbol{C}^T, \boldsymbol{\Omega}^T)^T$, $\boldsymbol{C} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_n\}$ and $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_n\}$.

3.3.2 Posterior Distribution in the Structure Equation of Fused Lasso

Let $\boldsymbol{G}_\omega = (\boldsymbol{g}_{\omega 1}, \cdots, \boldsymbol{g}_{\omega n})$, where $\boldsymbol{g}_{\omega i} = (\boldsymbol{\eta}_i^T, \boldsymbol{H}(x_i, \boldsymbol{\xi}_i)^T)^T$. Full conditionals in the structure equation for the h-th row of $\boldsymbol{\Lambda}_\omega$ is:

$$\boldsymbol{\Lambda}_{wh}|\boldsymbol{\Omega}, \psi_{\zeta h}, \boldsymbol{\tau}_{\Lambda_{wh}}, \boldsymbol{\upsilon}_{\Lambda_{wh}} \sim N_{q_1+N_H}((\boldsymbol{G}_\omega^T\boldsymbol{G}_\omega + \boldsymbol{D}_{\Lambda_{wh}}^{-1})^{-1}\boldsymbol{G}_\omega^T(\boldsymbol{\eta}_h - \beta_{0h}\boldsymbol{1}_n), \psi_{\zeta h}(\boldsymbol{G}^T\boldsymbol{G}_\omega + \boldsymbol{D}_{\Lambda_{wh}}^{-1})^{-1}), \tag{3.6}$$

where $\boldsymbol{\Lambda}_{\omega h} = (\boldsymbol{\Pi}_h^T, \boldsymbol{B}_h^T)^T$. N_h is the number of non-constant spline basis functions,

and $N_h = N_x + N_T$, where N_x is the number of basis functions related x's, and N_T is the number of basis functions related to exogenous latent variables.

Let $\boldsymbol{B}_h = (\boldsymbol{B}_{1h}^T, \boldsymbol{B}_{2h}^T)^T$, where $\boldsymbol{B}_{1h}$ are the coefficients corresponding to the x's and $\boldsymbol{B}_{2h}$ are the coefficients corresponding to the exogenous latent variables. Note that $\boldsymbol{\tau}_{\Lambda_{\omega h}} = (\tau_{\Pi_{h1}^2}, \cdots, \tau_{\Pi_{hq_1}^2}, \tau_{\Pi_{B_h1}^2}, \cdots, \tau_{\Pi_{B_hN_H}^2})^T$, and the full conditional distribution for $\boldsymbol{\tau}_{\Lambda_{\omega h}}$ are:

$$1/\tau_{\Pi_h j}^2 | \boldsymbol{\Pi}_h, \psi_{\zeta h} \sim IN(\sqrt{\frac{\lambda_{\Pi_h}^2 \psi_{\zeta h}}{\Pi_{hj}^2}}, \lambda_{\Pi_h}^2)$$

$$1/\tau_{B_{1hj}}^2 | \boldsymbol{B}_{1h}, \psi_{\zeta h} \sim IN(\sqrt{\frac{\lambda_{B_{1h}}^2 \psi_{\zeta h}}{(B_{1hj})^2}}, \lambda_{B_{1h}}^2)$$

$$1/\tau_{B_{2hj}}^2 | \boldsymbol{B}_{2h}, \psi_{\zeta h} \sim IN(\sqrt{\frac{\lambda_{B_{2h}}^2 \psi_{\zeta h}}{(B_{2h}j)^2}}, \lambda_{B_{2h}}^2)$$

$$1/\upsilon_{B_{2hj}}^2 | \boldsymbol{B}_{2h}, \psi_{\zeta h} \sim IN(\sqrt{\frac{\lambda_4^2 \psi_{\zeta h}}{(B_{2h(j+1)} - B_{2h(j)})^2}}, \lambda_4^2)$$

for $j = 1, \cdots, NT - 1$.

The full conditional of $\psi_{\zeta h}$ is:

$$\psi_{\zeta h} | \boldsymbol{\Lambda}_{wh}, \boldsymbol{G}_\omega \sim IG(\alpha_{0\zeta h} + \frac{n + q_1 + N_H + 1}{2}, \beta_{1\zeta h})$$

where $\beta_{1\zeta h} = \beta_{0\zeta h} + \frac{1}{2}[(\boldsymbol{\eta}_h - \beta_{0h}\mathbf{1}_n - \boldsymbol{G}_\omega^T \Lambda_{\omega h})^T (\boldsymbol{\eta}_h - \beta_{0h}\mathbf{1}_n - \boldsymbol{G}_\omega^T \Lambda_{\omega h}) + \boldsymbol{\Lambda}_{\omega h}^T \boldsymbol{D}_{\omega h}^{-1} \boldsymbol{\Lambda}_{\omega h}]$

Let the prior of λ's to be Gamma distribution and the full conditional distributions of them is:

$$\lambda_{\Pi_h}^2 | \boldsymbol{\tau}_{\Pi_h} \sim Gamma(q_1 + r_{0\omega}, \sum_{j=1}^{q_1} \tau_{\Pi_h j}^2/2 + \delta_{0\Pi})$$

$$\lambda^2_{B_{1h}}|\boldsymbol{\tau}_{B_{1h}} \sim Gamma(N_X + r_{0B_1}, \sum_{j=1}^{N_X} \tau^2_{B_{1h}j}/2 + \delta_{0B_1})$$

$$\lambda^2_{B_{2h}}|\boldsymbol{\tau}_{B_{2h}} \sim Gamma(N_T + r_{0B_2}, \sum_{j=1}^{N_T} \tau^2_{B_{2h}j}/2 + \delta_{0B_2})$$

$$\lambda^2_4|\boldsymbol{\upsilon}_{B_{2h}} \sim Gamma(N_T + r_{0B_{22}} - 1, \sum_{j=1}^{N_T-1} \upsilon^2_{B_{2h}j}/2 + \delta_{0B_{22}})$$

3.3.3 Posterior Distribution in the Structure Equation of Elastic Net

Full conditionals in the structure equation for the h-th row of $\boldsymbol{\Lambda}_\omega$ is:

$$\boldsymbol{\Lambda}_{wh}|\boldsymbol{\Omega}, \psi_{\zeta h}, \boldsymbol{\tau}_{\Lambda_{wh}}, \sim N_{q_1+N_H}((\boldsymbol{G}^T_\omega \boldsymbol{G}_\omega + \boldsymbol{D}^{-1}_{\Lambda_{wh}})^{-1}\boldsymbol{G}^T_\omega(\boldsymbol{\eta}_h - \beta_{0h}\mathbf{1}_n), \psi_{\zeta h}(\boldsymbol{G}^T\boldsymbol{G}_\omega + \boldsymbol{D}^{-1}_{\Lambda_{wh}})^{-1}),$$

$$1/\tau_{B_hkj}|\boldsymbol{\Lambda}_{\omega h}, \psi_{\zeta h} \sim IG(\sqrt{\frac{\lambda^2_{1\Lambda_h k}\psi_{\zeta h}}{\Lambda^2_{\omega hkj}}}, \lambda^2_{1\Lambda_h k})$$

for $j = 1, \cdots, N_k$, where $\lambda_{1\Lambda_h k} = \lambda_{1\Pi_h k}$, when $\boldsymbol{\Lambda}_{whk}$ are the coefficients of the endogenous latent variables; and $\lambda_{1\Lambda_h k} = \lambda_{1B_h k}$, when $\boldsymbol{\Lambda}_{whk}$ are the coefficients of the exogenous latent variables.

The full conditional of $\psi_{\zeta h}$ is:

$$\psi_{\zeta h}|\boldsymbol{\Lambda}_{wh}, \boldsymbol{G}_\omega \sim IG(\alpha_{0\zeta h} + \frac{n + q_1 + N_H + 1}{2}, \beta_{1\zeta h})$$

where $\beta_{1\zeta h} = \beta_{0\zeta h} + \frac{1}{2}[(\boldsymbol{\eta}_h - \beta_{0h}\mathbf{1}_n - \boldsymbol{G}^T_\omega\Lambda_{\omega h})^T(\boldsymbol{\eta}_h - \beta_{0h}\mathbf{1}_n - \boldsymbol{G}^T_\omega\Lambda_{\omega h}) + \boldsymbol{\Lambda}^T_{\omega h}\boldsymbol{D}^{-1}_{\omega h}\boldsymbol{\Lambda}_{\omega h}]$

Let the prior of λ's to be Gamma distribution and the full conditional distributions

of them is:

$$\lambda^2_{\Pi_h}|\boldsymbol{\tau}_{\Pi_h} \sim Gamma(q_1 + r_{0\Pi}, \sum_{j=1}^{q_1} \tau^2_{\Pi_h j}/2 + \delta_{0\Pi})$$

$$\lambda^2_{1B_{hk}}|\boldsymbol{\tau}_{\Lambda_{hk}} \sim Gamma(N_k + r_{1B_{hk}}, \sum_{j=1}^{N_k} \tau^2_{1B_h kj}/2 + \delta_{1B_{hk}})$$

$$\lambda^2_{2B_{hk}}|\boldsymbol{B} \sim Gamma(N_k + r_{2B_{hk}}, \frac{1}{2\psi_{\zeta h}} \sum_{j=1}^{N_k} \Lambda^2_{\omega hkj} + \delta_{2B_{hk}})$$

where $\boldsymbol{\Lambda}_{\omega hk}$ represent the Λ's belong to the group k.

3.3.4 MCMC Algorithm

The parameters from the measurement equation are denoted as $\theta_1^T = \{\Lambda_y, \Psi_\epsilon\}$, while the parameters from the structure equation are denoted as $\theta_2^T = \{\Lambda_\omega, \Psi_\xi, \Phi\}$. Let the parameter of interest be $\theta = (\theta_1^T, \theta_2^T)^T$.

Here are the variables we use in MCMC Algorithm:

- $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$, and $\boldsymbol{y}_i$ is $p \times 1$ vector of manifest variables.
- $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, and $\boldsymbol{x}_i$ is $s \times 1$ vector of fixed covariates.
- $\boldsymbol{C} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_n\}$, and $\boldsymbol{c}_i$ is $r \times 1$ vector of known function of $\boldsymbol{x}_i$.
- $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_n\}$, and $\boldsymbol{\omega}_i$ is $q \times 1$ vector of latent variables.

where $i = 1, \cdots, n$

$\boldsymbol{\Omega}$ are unobservable latent variables, we can generate it from the full conditional distribution $p(\boldsymbol{\Omega}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{\theta})$. Because the latent variables are independent among the subjects, we can write the full conditional distribution as $p(\boldsymbol{\Omega}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{\theta}) =$

$\prod_{i=1}^{n} p(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{c}_i, \boldsymbol{\theta})$. Let $g_{yi} = (\boldsymbol{c}_i^T, \boldsymbol{\omega}_i^T)^T$. The full conditional distribution of $\boldsymbol{\omega}_i$ is:

$$\begin{aligned} p(\boldsymbol{\omega}_i|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{c}_i, \boldsymbol{\theta}) &\propto p(\boldsymbol{y}_i|\boldsymbol{c}_i, \boldsymbol{\omega}_i, \theta_1) p(\boldsymbol{\eta}_i|\boldsymbol{x}_i, \boldsymbol{\xi}_i, \theta_2) p(\boldsymbol{\xi}_i|\theta_2) \\ &\propto \exp\{-\frac{1}{2}(\boldsymbol{y}_i - \Lambda_y g_{yi})^T \boldsymbol{\Psi}_\epsilon^{-1} (\boldsymbol{y}_i - \Lambda_y g_{yi}) - \frac{1}{2}\boldsymbol{\xi}_i^T \boldsymbol{\Phi}_i^{-1} \boldsymbol{\Phi}_i \\ &\quad -\frac{1}{2}(\boldsymbol{\eta}_i - \beta_0 - \Lambda_\omega g_{\omega i})^T \boldsymbol{\Psi}_\zeta^{-1} (\boldsymbol{\eta}_i - \beta_0 - \Lambda_\omega g_{\omega i})\} \end{aligned} \tag{3.7}$$

$\boldsymbol{\omega}_i$ can be sampled using Metropolis Hastings (MH) algorithm with a proposal distribution $q(\boldsymbol{\omega}_i^*|\sigma_\omega^2) \sim N(\boldsymbol{\omega}_i^{(j)}, \sigma_\omega^2 \Sigma_\omega)$, where $\boldsymbol{\omega}_i^*$ is the proposed new value and $\omega_i^{(j)}$ is the value from previous step (jth step). From [Guo et al., 2012],

$$\Sigma_\omega^{-1} = \Lambda^T \boldsymbol{\Psi}^{-1} \Lambda + \begin{pmatrix} \Pi_0^T \boldsymbol{\Psi}_\zeta^{-1} \Pi_0 & -\Pi_0^T \boldsymbol{\Psi}_\zeta^{-1} \boldsymbol{B} \Delta_H \\ -\Delta_H^T \boldsymbol{B}^T \boldsymbol{\Psi}_\zeta^{-1} \Pi_0 & \boldsymbol{\Phi}^{-1} + \Delta_H^T \boldsymbol{B}^T \boldsymbol{\Psi}_\zeta^{-1} \boldsymbol{B} \Delta_H \end{pmatrix} \tag{3.8}$$

where $\Delta_H = \partial \boldsymbol{H}(\boldsymbol{x}_i, \boldsymbol{\xi}_i)/\partial \boldsymbol{\xi}_i^T|_{\boldsymbol{\xi}_i = \boldsymbol{0}}$. The proposed $\boldsymbol{\omega}_i^*$ can be accepted with the probability $\min\{1, \frac{p(\boldsymbol{\omega}_i^*|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{c}_i, \boldsymbol{\theta})}{p(\boldsymbol{\omega}_i^{(j)}|\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{c}_i, \boldsymbol{\theta})}\}$. $\boldsymbol{\Omega}$ can be sampled using Gibbs sampler.

For θ_1, sample $\boldsymbol{\Lambda}_{yk}^*|rest$ and $\psi_{\epsilon k}|rest$ from 3.4 and 3.5 respectively.

For θ_2, the posterior distribution of the parameters are different between Bayesian fused Lasso and Bayesian Elastic Net. We can sample the unknown parameters from the posterior distribution we get on section 3.3.2 and section 3.3.3.

3.4 Simulation Study

To illustrate the fused Lasso and elastic net we have considered the case where the covariates have correlations. Under this framework it is of interest to compare among

the Bayesian Lasso, the Bayesian fused Lasso and Bayesian elastic net.

3.4.1 Simulation 1

We follow the simulation setup on [Guo et al., 2012], setting $n = 500$, $p = 9$, $q_1 = 1$, $q_2 = 2$ and $\boldsymbol{A} = \text{diag}(0^*, 0^*, 0^*, \mu_4, \ldots, \mu_9)$, $\boldsymbol{c_i} = (1, \ldots, 1)^T$,

$$\boldsymbol{\Lambda}^T = \begin{bmatrix} 1.0^* & \lambda_{21} & \lambda_{31} & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 1.0^* & \lambda_{52} & \lambda_{62} & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 1.0^* & \lambda_{83} & \lambda_{93} \end{bmatrix},$$

where $\mu_4 = \ldots = \mu_9 = \lambda_{21} = \ldots = \lambda_{93} = \zeta = .36$ and $\{\phi_{11}, \phi_{12}, \phi_{22}\} = \{1, .25, 1\}$. The function, $f(\xi_{i1}, \xi_{i2}) = f_1(\xi_{i1}) + f_2(\xi_{i2}) + f_{12}(\xi_{i1}, \xi_{i2})$, where $f_1(\xi_{i1}) = \sin(\xi i1) - \xi_{i1}$, $f_1(\xi_{i1}) = \exp(\xi_{i2})/2.5 - 3.0$ and $f_{12}(\xi_{i1}, \xi_{i2}) = 0$, has been used to define the underlying relationship between the endogenous and exogenous latent variables. Also, this function is considered unknown and will be approximated using natural cubic splines, i.e.,

$$\begin{aligned} f_j(\xi_{ij}) &\approx \beta_{j2}\xi_{ij} \sum_{m=1}^{K-2} \beta_{j,m+2} \left(d_m(\xi_{ij}) - d_{K-1}(\xi_{ij})\right) \\ f_{12}(\xi_{i1}, \xi_{i2}) &\approx \beta_{12}^{(12)}\xi_{i1}\xi_{i2} + \sum_{m_1=1}^{K-2} \xi_{i2} \left(d_{m_1}(\xi_{i1}) - d_{K-1}(\xi_{i1})\right) \\ &\quad + \sum_{m_2=1}^{K-2} \xi_{i1} \left(d_{m_2}(\xi_{i2}) - d_{K-1}(\xi_{i2})\right) \\ &\quad + \sum_{m_1=1}^{K-2} \sum_{m_2=1}^{K-2} \left(d_{m_1}(\xi_{i1}) - d_{K-1}(\xi_{i1})\right) \left(d_{m_2}(\xi_{i2}) - d_{K-1}(\xi_{i2})\right), \end{aligned}$$

with $d_k(\xi_{ij}) = \left[(\xi_{ij} - \kappa_k)_+ - (\xi_{ij} - \kappa_k)_+\right] / (\kappa_K - \kappa_k)$ where K is the number of knots and $(\kappa_k, k = 1, \ldots, K)$ are the location of the knots. The knot locations are selected using a truncated power series basis developed in [Hastie et al., 2009]. In general cubic splines will be correlated, thus the use of the fused Lasso is appropriate.

We consider $s = 35$ with true parameter values

$$b_l = \begin{cases} 0.5 \text{ if } l \in \{1, 2, 3\} \\ -0.7 \text{ if } l \in \{4, 5\} \\ 0.85 \text{ if } l \in \{6, \cdots, 15\} \\ 0.7 \text{ if } l = 32 \\ 0.5 \text{ if } l = 33 \\ 0 \text{ otherwise} \end{cases}.$$

To induce correlation of the covariates $x_1, \ldots, x_{31}, x_{34}, x_{35}$ are simulated from a multivariate standard normal distribution where $\text{corr}(x_i, x_j) = .5^{|i-j|}$, $i \neq j \in (6, \ldots, 15)$, $\text{corr}(x_i, x_j) = .7$, $i - j = 1, i \in (1, 2, 3)$, $\text{corr}(x_i, x_j) = .9$, $i \neq j \in (4, 5)$ and all other correlations equal to 0. The covariate of $x_{32} \sim 2\text{Binomial}(1, .5)$ and $x_{33} \sim N(-0.5, 1)$.

Table 3.1 summarizes the parameter estimates from the 50 simulations using the fused Lasso, elastic net and standard Lasso. The b_i parameters which relate the covariates to the endogenous latent variable are slightly closer to the true value when the fused Lasso is used, however for most of the parameters it is only a slight improvement. The covariates with $\text{corr}(x_i, x_j) = .7$, $i \neq j \in (1, 2, 3)$ have the most marked improvement when the fused Lasso is used instead of the standard Lasso or elastic

net. All models are efficient at shrinking the insignificant parameters to 0.

There is a fairly significant difference in the spline estimates between the standard Lasso and the other two models. For the spline parameters that are not equal to zero it is not possible to determine which of the models is better in terms of estimation. However, in many of these cases the standard deviations of standard Lasso model are significantly higher; while fused lasso and elastic net are similar to each other. For the spline parameters that are equal to zero both fused Lasso and elastic net models shrink the estimates nearer to zero than standard Lasso and many have significantly lower standard deviations. Moreover, elastic net is slightly better than fused Lasso.

To measure the models efficiency at predicting the endogenous latent variable using the covariates and exogenous latent variables, we consider three measures of RMSE.

- $\text{RMSE}(\hat{f}) = \sqrt{\sum_{i=1}^{n} \left(\hat{f}(\xi_{i1}, \xi_{i2}) - f(\xi_{i1}, \xi_{i2})\right)^2 / n}$ is a measure of the models ability to approximate the nonlinear relationship between the endogenous and exogenous latent variables,
- $\text{RMSE}(\hat{B}) = \sqrt{\sum_{i=1}^{n} \left(\boldsymbol{X}\hat{\boldsymbol{B}} - \boldsymbol{X}\boldsymbol{B}\right)^2 / n}$ is a measure of the models ability to relate the covariates to the endogenous latent variables and
- $\text{RMSE} = \sqrt{\sum_{i=1}^{n} \left(\left(\boldsymbol{X}\hat{\boldsymbol{B}} + \hat{f}(\xi_{i1}, \xi_{i2})\right) - (\boldsymbol{X}\boldsymbol{B} + f(\xi_{i1}, \xi_{i2}))\right)^2 / n}$ is a measure of the models overall ability to predict the endogenous latent variable.

The most significant improvement in the fused Lasso and elastic net appears to be in the $\text{RMSE}(\hat{f})$ which suggests that it is much better at defining the relationship between the endogenous and exogenous latent variables. And $\text{RMSE}(\hat{f})$ of elastic net is slightly lower than fused Lasso's. A possible reason there was little impact from

on the covariate parameters is that it is very difficult to simulate complex correlation structures. If more covariance structures are examined we believe the difference could be significant.

3.4.2 Simulation 2

In order to compare the difference defining the relationship between the endogenous and exogenous latent variables among these three model. We randomly choose one of the simulation study and let the coefficient of the covariate to be zeros and plot the surface of $f(\xi_{i1}, \xi_{i2})$. Figure(3.1) shows the true relationship between exogenous latent variables and endogenous latent variable based on function $\eta = F(x, \boldsymbol{\xi})$; figure (3.2) shows the relationship between them based on the simulation data, and some of the surface does not have data. figure(3.3, 3.4 and 3.5) show the estimated surface via original Lasso, Fused Lasso and Elastic Net. In figure(3.3), Lasso perform badly when η_1 and η_2 both greater than 0. From figure(3.2), there are no data when both η_1 and η_2 are greater than 2.5. Fused Lasso and Elastic perform similarly. In this simulation, Fused Lasso might perform a little better, when both η_1 and η_2 are less than 0.

3.5 Application

We apply Lasso and Elastic net models to analyze Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey). There are three exogenous latent variables of interests, cigarette morbidity, marijuana morbidity and behavior risk index; one endogenous latent variable, alcohol morbidity. We want to analyze

		Fused		Elastic Net		Standard	
Para.	True	Est.	STD	Est.	STD	Est.	STD
b_1	0.5	0.4604	0.0885	0.4417	0.1241	0.4524	0.1767
b_2	0.5	0.5596	0.1230	0.5768	0.1739	0.5685	0.2484
b_3	0.5	0.4512	0.0909	0.4374	0.1280	0.4469	0.1798
b_4	-0.7	-0.6817	0.0808	-0.6800	0.0894	-0.6793	0.0949
b_5	-0.7	-0.7120	0.0809	-0.7099	0.0908	-0.7143	0.0955
b_6	0.85	0.8430	0.0429	0.8460	0.0485	0.8460	0.0481
b_7	0.85	0.8466	0.0537	0.8480	0.0612	0.8481	0.0608
b_8	0.85	0.8465	0.0519	0.8405	0.0598	0.8409	0.0586
b_9	0.85	0.8408	0.0542	0.8448	0.0634	0.8455	0.0643
b_{10}	0.85	0.8579	0.0555	0.8543	0.0623	0.8523	0.0638
b_{11}	0.85	0.8470	0.0513	0.8440	0.0574	0.8439	0.0562
b_{12}	0.85	0.8454	0.0516	0.8433	0.0537	0.8428	0.0535
b_{13}	0.85	0.8430	0.0510	0.8499	0.0549	0.8495	0.0547
b_{14}	0.85	0.8499	0.0502	0.8404	0.0568	0.8415	0.0566
b_{15}	0.85	0.8513	0.0488	0.8582	0.0507	0.8581	0.0490
b_{16}	0	0.0043	0.0382	0.0014	0.0387	0.0026	0.0425
b_{17}	0	0.0049	0.0380	0.0051	0.0347	0.0063	0.0388
b_{18}	0	0.0050	0.0432	0.0036	0.0418	0.0036	0.0454
b_{19}	0	-0.0003	0.0382	-0.0003	0.0362	-0.0001	0.0389
b_{20}	0	0.0048	0.0397	0.0004	0.0400	0.0006	0.0440
b_{21}	0	-0.0006	0.0393	-0.0034	0.0367	-0.0036	0.0400
b_{22}	0	-0.0030	0.0377	-0.0017	0.0399	-0.0025	0.0427
b_{23}	0	-0.0024	0.0429	-0.0021	0.0417	-0.0014	0.0452
b_{24}	0	-0.0028	0.0413	0.0022	0.0353	0.0030	0.0380
b_{25}	0	0.0027	0.0388	-0.0042	0.0374	-0.0048	0.0399
b_{26}	0	-0.0030	0.0367	-0.0031	0.0368	-0.0017	0.0422
b_{27}	0	0.0039	0.0388	0.0026	0.0337	0.0022	0.0376
b_{28}	0	-0.0015	0.0385	0.0024	0.0368	0.0018	0.0403
b_{29}	0	0.0061	0.0408	0.0052	0.0377	0.0052	0.0402
b_{30}	0	0.0008	0.0360	-0.0029	0.0367	-0.0033	0.0401
b_{31}	0	0.0036	0.0361	-0.0009	0.0337	-0.0016	0.0364
b_{32}	0.7	0.6908	0.0452	0.6870	0.0442	0.6948	0.0443
b_{33}	-0.5	-0.4932	0.0407	-0.4932	0.0412	-0.5001	0.0409
b_{34}	0	-0.0064	0.0368	-0.0028	0.0363	-0.0029	0.0406
b_{35}	0	0.0055	0.0384	0.0047	0.0384	0.0052	0.0421
β_0	-	-2.1529	0.0917	-2.1641	0.0940	-2.2231	0.1641
β_{12}	-	-0.1947	0.0772	-0.2165	0.0788	-0.2744	0.1721
β_{13}	-	-0.0607	0.0233	0.0117	0.0403	0.3379	0.4499
β_{14}	-	-0.0380	0.0283	-0.0308	0.0267	-0.0199	0.0465
β_{15}	-	-0.0213	0.0322	-0.0793	0.0456	-0.3637	0.3492
β_{22}	-	0.0805	0.0404	0.0781	0.0611	0.1753	0.1352
β_{23}	-	0.1528	0.0324	0.1163	0.0354	-0.0002	0.1777
β_{24}	-	0.1985	0.0556	0.1766	0.0478	0.1118	0.0704
β_{25}	-	0.1866	0.0650	0.2782	0.1173	0.3291	0.2340
$\beta_{22}^{(12)}$	0	0.0411	0.0632	0.0208	0.0551	0.1011	0.1753
$\beta_{23}^{(12)}$	0	-0.0320	0.0634	-0.0256	0.0553	-0.1101	0.1630
$\beta_{24}^{(12)}$	0	-0.0446	0.0808	-0.0198	0.0618	-0.0912	0.1752
$\beta_{25}^{(12)}$	0	-0.0369	0.0732	-0.0148	0.0688	-0.0600	0.2716
$\beta_{32}^{(12)}$	0	-0.0207	0.0331	-0.0093	0.0234	-0.0823	0.1394
$\beta_{33}^{(12)}$	0	0.0197	0.0190	0.0190	0.0273	0.0406	0.2214
$\beta_{34}^{(12)}$	0	0.0415	0.0304	0.0279	0.0344	0.1178	0.1546
$\beta_{35}^{(12)}$	0	0.0444	0.0329	0.0378	0.0431	0.2527	0.2123
$\beta_{42}^{(12)}$	0	0.0009	0.0318	-0.0110	0.0263	-0.0676	0.1108
$\beta_{43}^{(12)}$	0	0.0205	0.0329	0.0155	0.0319	-0.0275	0.1377
$\beta_{44}^{(12)}$	0	0.0362	0.0439	0.0246	0.0390	0.0510	0.1011
$\beta_{45}^{(12)}$	0	0.0382	0.0472	0.0352	0.0480	0.1807	0.2341
$\beta_{52}^{(12)}$	0	-0.0029	0.0461	-0.0118	0.0310	-0.0572	0.1702
$\beta_{53}^{(12)}$	0	0.0190	0.0569	0.0116	0.0381	-0.1153	0.1802
$\beta_{54}^{(12)}$	0	0.0469	0.0848	0.0215	0.0451	-0.0199	0.1947
$\beta_{55}^{(12)}$	0	0.0760	0.1133	0.0323	0.0538	0.1275	0.4199
RMSE($\hat{f}$)	0	0.6676	0.3864	0.6493	0.3778	1.5127	1.5713
RMSE($\hat{B}$)	0	0.2323	0.0299	0.2350	0.0291	0.2436	0.0278
RMSE	0	0.6966	0.3738	0.6764	0.3693	1.5311	1.5627

Table 3.1: Simulation Result for Fused Lasso, Elastic Net and Standard Lasso

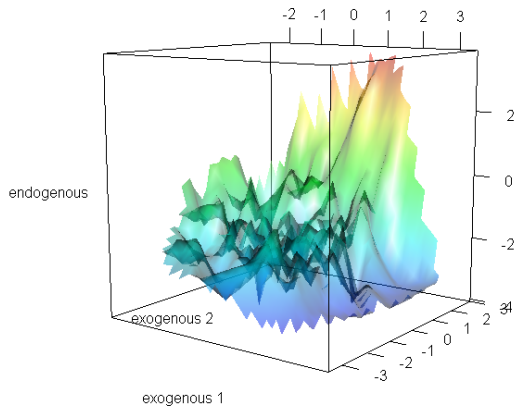


Figure 3.1: True surface for $\eta = F(x, \boldsymbol{\xi})$

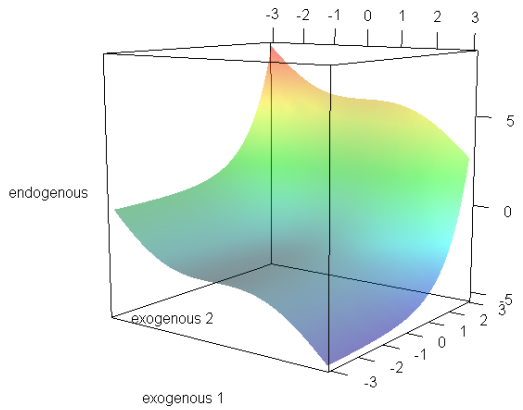


Figure 3.2: True surface for simulated data

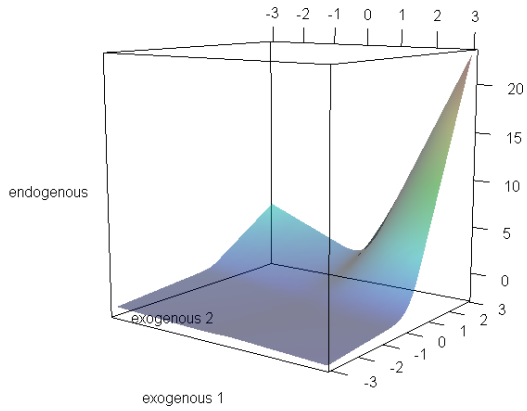


Figure 3.3: Estimated surface via Lasso

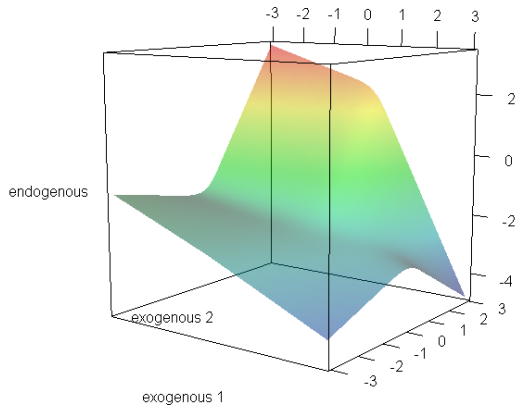


Figure 3.4: Estimated surface via Fused Lasso

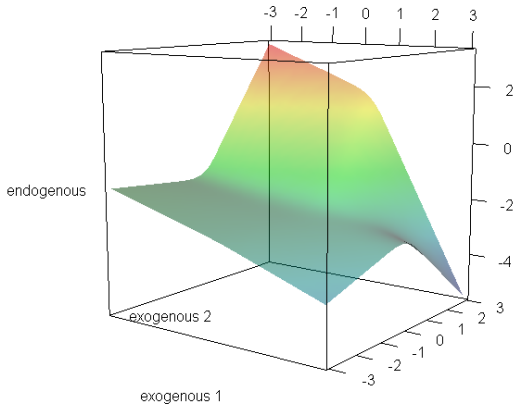


Figure 3.5: Estimated surface via Elastic Net

how cigarette morbidity, marijuana morbidity and behavior risk index affect alcohol morbidity. We used the subset from the Monitoring the Future data: 1878 students who had drinking experience.

The endogenous latent variable, alcohol morbidity, is measured by following items:

- The occasions that students had alcoholic beverages to drink, more than just a few sips in their lifetime.
- The occasions that students had alcoholic beverages to drink, more than just a few sips last year.
- The occasions that students had alcoholic beverages to drink, more than just a few sips last month.
- The number of times that the students had five or more drinks in a row in the last two weeks.

The first exogenous latent variable, cigarette morbidity, is measure by following items:

- The occasions that students smoked cigarettes in their lifetime.
- The occasions have students smoked cigarettes during the past 30 days.

The second exogenous latent variable, marijuana morbidity, is measure by following items:

- The occasions that students smoked marijuana in their lifetime.
- The occasions that students smoked marijuana last year.

- The occasions that students smoked marijuana last month.

The third exogenous latent variable, behavior risk index, is measure by following items:

- During the last four weeks, the number of whole days of school students have missed because they skipped.
- During the last four weeks, the number of whole days of school students have missed because other reasons.
- During a typical week, the number of evenings students go out for fun and recreation.
- On the average, how often students go out with a date.
- During an average week, how much students usually drive.

As a result, there are totally 14 manifest variables. The $\mathbf{\Lambda}$ in the measurement equation is:

$$\mathbf{\Lambda}^T = \begin{pmatrix} 1 & \lambda_{21} & \lambda_{31} & \lambda_{41} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_{62} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{83} & \lambda_{93} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{11,4} & \lambda_{12,4} & \lambda_{13,4} & \lambda_{14,4} \end{pmatrix} \tag{3.9}$$

Let $\boldsymbol{A} = \text{diag}(0, \cdots, 0, \mu_5, \cdots, \mu_{14})$ and $\boldsymbol{c}_i = (1, \cdots, 1)^T$. In addition, we have five covariates, which are gender, geographic area, living with siblings, father education level and mother education level. Let $\boldsymbol{x}_i = (x_{1i}, \cdots, x_{5i})$ To study the interaction

between the exogenous latent variables and endogenous latent variable, we proposed following structure equation model:

$$\eta_i = \boldsymbol{x}_i \boldsymbol{b}^T + f_1(\xi_{1i}) + f_2(\xi_{2i}) + f_3(\xi_{3i}) + f_{12}(\xi_{1i}, \xi_{2i}) + f_{13}(\xi_{1i}, \xi_{3i}) + f_{23}(\xi_{2i}, \xi_{3i}) \quad (3.10)$$

where $\boldsymbol{b} = (b_1, \cdots, b_5)$. Similar to simulation study, natural cubic splines are used in function $f(\cdot)$ with 5 knots. MCMC chains of 20,000 iterations are generated and the burnin is 10,000. We use both Bayesian fused lasso and Bayesian elastic net to solve the problem, and compare the result with Bayesian Lasso. Table 3.2 shows the estimates from measurement equation. The estimates are very similar among three methods.

Para.	Fused LASSO Est.	Elastic Net Est.	LASSO Est.
$\lambda_{2,1}$	0.8477	0.8358	0.836
$\lambda_{3,1}$	0.5202	0.5067	0.5069
$\lambda_{4,1}$	0.4098	0.3982	0.3983
$\lambda_{6,2}$	1.0505	0.9976	1.047
$\lambda_{8,3}$	1.2825	1.2866	1.2965
$\lambda_{9,3}$	1.2088	1.2127	1.2219
$\lambda_{11,4}$	0.7485	0.5973	0.6634
$\lambda_{12,4}$	0.3364	0.4089	0.3993
$\lambda_{13,4}$	0.1659	0.1558	0.1583
$\lambda_{14,4}$	0.109	0.1978	0.1796
μ_5	3.133	3.1667	3.1099
μ_6	1.9161	1.9526	1.8925
μ_7	5.3888	5.4149	5.4285
μ_8	4.4709	4.5048	4.5217
μ_9	2.8753	2.9067	2.9222
μ_10	2.0841	2.1565	2.1203
μ_11	1.7787	1.8275	1.8065
μ_12	3.9871	4.0141	3.999
μ_13	3.1242	3.1356	3.1303
μ_14	3.5832	3.5943	3.5879

Para.: parameter

Est. posterior estimates

Table 3.2: Non-Spline Parameter Estimation

The structure equation results for fused Lasso and Elastic Net are on Table 3.3 and Table 3.4 respectively. Some of the β's from Elastic Net are not converged.

Comparing to the result form Bayesian Lasso Table 3.5. All the β's from Bayesian Lasso are not converged. Fused lasso performed best in this application, all the β's from Fused Lasso converge, and result shows that there is interaction between marijuana morbidity and behavior risk index. The main effect of cigarette morbidity is also significant. The graphs of the two-way interaction of these three exogenous latent variables shows their relation with endogenous latent variable. Figure 3.6 shows there is not obviously interaction between cigarette morbidity and marijuana morbidity, but both main effects are significant. When cigarette morbidity or/and marijuana morbidity increase, alcohol morbidity increases. Figure 3.7 shows similar pattern with cigarette morbidity and behavior risk index. Figure 3.8 shows the interaction between marijuana morbidity and behavior risk index. When behavior risk index is in the higher level, as marijuana morbidity increases, alcohol morbidity increases faster.

Para.	Est.	Para.	Est.	Para.	Est.	Para.	Est.
b_1	0.1216*	$\beta_1^{(12)}$	-0.0553	$\beta_1^{(13)}$	-0.0156	$\beta_1^{(23)}$	-0.1149^*
b_2	0.0243	$\beta_2^{(12)}$	-0.008	$\beta_2^{(13)}$	0.0073	$\beta_2^{(23)}$	-0.0056
b_3	0.0835	$\beta_3^{(12)}$	0.0081	$\beta_3^{(13)}$	0.0179	$\beta_3^{(23)}$	0.0246
b_4	0.0845	$\beta_4^{(12)}$	0.0152	$\beta_4^{(13)}$	0.0248	$\beta_4^{(23)}$	0.0327
b_5	-0.0322	$\beta_5^{(12)}$	0.0258	$\beta_5^{(13)}$	0.033	$\beta_5^{(23)}$	0.0286
β_0	5.8534*	$\beta_6^{(12)}$	0.0063	$\beta_6^{(13)}$	0.007	$\beta_6^{(23)}$	0.0165
β_{12}	0.2729*	$\beta_7^{(12)}$	-0.0045	$\beta_7^{(13)}$	-0.0041	$\beta_7^{(23)}$	0.009
β_{13}	0.0303	$\beta_8^{(12)}$	-0.0044	$\beta_8^{(13)}$	-0.0029	$\beta_8^{(23)}$	0.0068
β_{14}	-0.0037	$\beta_9^{(12)}$	0.0166	$\beta_9^{(13)}$	0.0163	$\beta_9^{(23)}$	0.0113
β_{15}	0.0067	$\beta_{10}^{(12)}$	0.0007	$\beta_{10}^{(13)}$	-0.006	$\beta_{10}^{(23)}$	0.0002
β_{22}	0.1142*	$\beta_{11}^{(12)}$	-0.0086	$\beta_{11}^{(13)}$	-0.0139	$\beta_{11}^{(23)}$	-0.0047
β_{23}	0.0654*	$\beta_{12}^{(12)}$	-0.0065	$\beta_{12}^{(13)}$	-0.012	$\beta_{12}^{(23)}$	-0.0066
β_{24}	0.0382	$\beta_{13}^{(12)}$	0.0176	$\beta_{13}^{(13)}$	0.0048	$\beta_{13}^{(23)}$	-0.002
β_{25}	0.0441	$\beta_{14}^{(12)}$	-0.0025	$\beta_{14}^{(13)}$	-0.0273	$\beta_{14}^{(23)}$	-0.0168
β_{32}	0.1745*	$\beta_{15}^{(12)}$	-0.0151	$\beta_{15}^{(13)}$	-0.0529	$\beta_{15}^{(23)}$	-0.0252
β_{33}	0.0986*	$\beta_{16}^{(12)}$	-0.0217	$\beta_{16}^{(13)}$	-0.0824	$\beta_{16}^{(23)}$	-0.0314
β_{34}	0.0417						
β_{35}	-0.0039						

Para.: parameter

Est. posterior estimates

* marked values indicates 90% of the distribution is greater than 0 or less than 0.

Table 3.3: Spline Parameter Estimation using Bayesian Fused Lasso

Para.	Est.	Para.	Est.	Para.	Est.	Para.	Est.
b_1	0.0169	$\beta_1^{(12)}$	-0.0026	$\beta_1^{(13)}$	-0.1193	$\beta_1^{(23)}$	0.0023
b_2	0.0018	$\beta_2^{(12)}$	-0.0054	$\beta_2^{(13)}$	-0.4759	$\beta_2^{(23)}$	0.0031
b_3	0.0035	$\beta_3^{(12)}$	-0.005	$\beta_3^{(13)}$	0.462	$\beta_3^{(23)}$	0.0039
b_4	0.004	$\beta_4^{(12)}$	-0.0045	$\beta_4^{(13)}$	2.5223§	$\beta_4^{(23)}$	0.0046
b_5	-0.0042	$\beta_5^{(12)}$	0.0022	$\beta_5^{(13)}$	-0.3639	$\beta_5^{(23)}$	-0.0008
β_0	5.9121	$\beta_6^{(12)}$	0.0012	$\beta_6^{(13)}$	-0.4635	$\beta_6^{(23)}$	0.0015
β_{12}	-0.0291	$\beta_7^{(12)}$	0.0013	$\beta_7^{(13)}$	-0.6698	$\beta_7^{(23)}$	0.0018
β_{13}	-0.0016	$\beta_8^{(12)}$	0.0011	$\beta_8^{(13)}$	-1.2979§	$\beta_8^{(23)}$	0.0027
β_{14}	0.0043	$\beta_9^{(12)}$	0.003	$\beta_9^{(13)}$	-0.1049	$\beta_9^{(23)}$	-0.0011
β_{15}	0.0156	$\beta_{10}^{(12)}$	0.0021	$\beta_{10}^{(13)}$	-0.1743	$\beta_{10}^{(23)}$	0.0021
β_{22}	-0.0002	$\beta_{11}^{(12)}$	0.0022	$\beta_{11}^{(13)}$	-0.5128	$\beta_{11}^{(23)}$	0.0021
β_{23}	-0.0026	$\beta_{12}^{(12)}$	0.0022	$\beta_{12}^{(13)}$	-1.5011§	$\beta_{12}^{(23)}$	0.003
β_{24}	-0.0025	$\beta_{13}^{(12)}$	0.0044	$\beta_{13}^{(13)}$	1.0263§	$\beta_{13}^{(23)}$	-0.0011
β_{25}	-0.0026	$\beta_{14}^{(12)}$	0.0045	$\beta_{14}^{(13)}$	1.3305§	$\beta_{14}^{(23)}$	0.0017
β_{32}	2.4961§	$\beta_{15}^{(12)}$	0.0042	$\beta_{15}^{(13)}$	0.4733	$\beta_{15}^{(23)}$	0.0024
β_{33}	18.3222§	$\beta_{16}^{(12)}$	0.0041	$\beta_{16}^{(13)}$	-0.53	$\beta_{16}^{(23)}$	0.0031
β_{34}	-35.6403§						
β_{35}	16.0324§						

Para.: parameter

Est. posterior estimates

§ indicates the estimates are not converged..

Table 3.4: Spline Parameter Estimation using Bayesian Elastic Net

Para.	Est.	Para.	Est.	Para.	Est.	Para.	Est.
b_1	0.0159	$\beta_1^{(12)}$	$-2.7382^{\S}$	$\beta_1^{(13)}$	$1.0451^{\S}$	$\beta_1^{(23)}$	$0.2214^{\S}$
b_2	0.0014	$\beta_2^{(12)}$	$8.284^{\S}$	$\beta_2^{(13)}$	$22.0392^{\S}$	$\beta_2^{(23)}$	$0.3895^{\S}$
b_3	0.0008	$\beta_3^{(12)}$	$-3.644^{\S}$	$\beta_3^{(13)}$	$-3.7627^{\S}$	$\beta_3^{(23)}$	$4.5003^{\S}$
b_4	0.0074	$\beta_4^{(12)}$	$-1.4991^{\S}$	$\beta_4^{(13)}$	$-15.6498^{\S}$	$\beta_4^{(23)}$	$-2.7686^{\S}$
b_5	-0.0038	$\beta_5^{(12)}$	$9.7528^{\S}$	$\beta_5^{(13)}$	$-7.5459^{\S}$	$\beta_5^{(23)}$	$1.2014^{\S}$
β_0	5.9608	$\beta_6^{(12)}$	$-5.0116^{\S}$	$\beta_6^{(13)}$	$-11.514^{\S}$	$\beta_6^{(23)}$	$-7.5515^{\S}$
β_{12}	$-1.3933^{\S}$	$\beta_7^{(12)}$	$-3.7479^{\S}$	$\beta_7^{(13)}$	$-10.617^{\S}$	$\beta_7^{(23)}$	$-18.5468^{\S}$
β_{13}	$-9.2749^{\S}$	$\beta_8^{(12)}$	$-1.6634^{\S}$	$\beta_8^{(13)}$	$4.8088^{\S}$	$\beta_8^{(23)}$	$-6.5395^{\S}$
β_{14}	$8.9574^{\S}$	$\beta_9^{(12)}$	$-7.53^{\S}$	$\beta_9^{(13)}$	$2.7655^{\S}$	$\beta_9^{(23)}$	$-4.1878^{\S}$
β_{15}	$0.6507^{\S}$	$\beta_{10}^{(12)}$	$-0.5455^{\S}$	$\beta_{10}^{(13)}$	$0.9769^{\S}$	$\beta_{10}^{(23)}$	$3.9873^{\S}$
β_{22}	$-1.7371^{\S}$	$\beta_{11}^{(12)}$	$3.0895^{\S}$	$\beta_{11}^{(13)}$	$0.6814^{\S}$	$\beta_{11}^{(23)}$	$4.1692^{\S}$
β_{23}	$3.1955^{\S}$	$\beta_{12}^{(12)}$	$5.5655^{\S}$	$\beta_{12}^{(13)}$	$18.2107^{\S}$	$\beta_{12}^{(23)}$	$34.5997^{\S}$
β_{24}	$-1.5914^{\S}$	$\beta_{13}^{(12)}$	$0.4208^{\S}$	$\beta_{13}^{(13)}$	$3.3887^{\S}$	$\beta_{13}^{(23)}$	$2.5759^{\S}$
β_{25}	$-0.5296^{\S}$	$\beta_{14}^{(12)}$	$-0.4141^{\S}$	$\beta_{14}^{(13)}$	$-1.0071^{\S}$	$\beta_{14}^{(23)}$	$8.6044^{\S}$
β_{32}	$4.3805^{\S}$	$\beta_{15}^{(12)}$	$1.5452^{\S}$	$\beta_{15}^{(13)}$	$-2.2612^{\S}$	$\beta_{15}^{(23)}$	$-0.4574^{\S}$
β_{33}	$3.9027^{\S}$	$\beta_{16}^{(12)}$	$-1.8298^{\S}$	$\beta_{16}^{(13)}$	$-2.0676^{\S}$	$\beta_{16}^{(23)}$	$-21.5666^{\S}$
β_{34}	$-5.7997^{\S}$						
β_{35}	$1.9083^{\S}$						

Para.: parameter

Est. posterior estimates

§ indicates the estimates are not converged.

Table 3.5: Spline Parameter Estimation using Bayesian LASSO

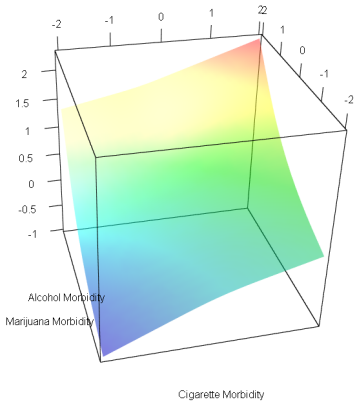


Figure 3.6: Estimated surface for cigarette morbidity and marijuana morbidity

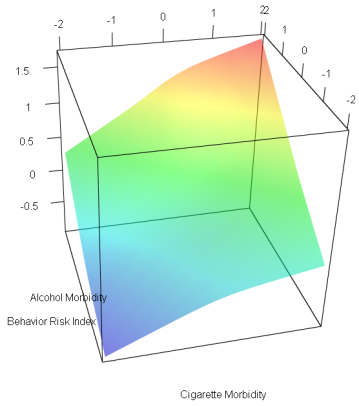


Figure 3.7: Estimated surface for cigarette morbidity and behavior risk index

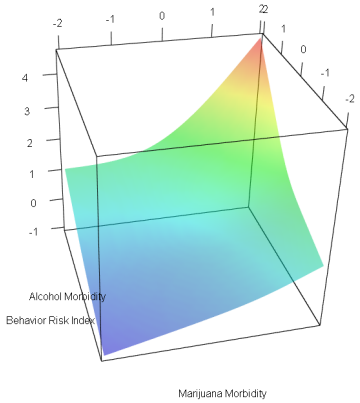


Figure 3.8: Estimated surface for marijuana morbidity and behavior risk index

3.6 Discussion

We adapt Bayesian fused Lasso and Bayesian elastic net for using in semiparametric structural equation models. Basis expansions are used to approximate the nonparametric relationships between the endogenous latent variables and the exogenous latent variables and covariates. When cubic splines are used as the basis expansion, it is beneficial to use the fused Lasso or the elastic net to estimate the parameters since cubic splines are correlated in general. In the simulation study, both fused Lasso and elastic net reduce the standard deviations of the spline parameters and shrink the estimates of the spline parameters closer to zero when the true values of those parameters are equal to zero. More importantly, RMSE($\hat{f}$) of fused Lasso and elastic net is about half of RMSE($\hat{f}$) of standard Lasso.

There are benefits to use the fused Lasso to estimate the coefficients of the covariates, however, it is difficult to generate realistic correlation structures. The usefulness of this method will depend greatly on the type of correlation. In our simulation study, the fused Lasso has a remarkable improvement over the standard Lasso for the tridiagonal structure with correlation equal to 0.70. However, it is difficult to simulate tridiagonal structures since we often get negative eigenvalues. We believe that if a natural order are present in a real data set the fused Lasso would lead to much better results.

In the application, we treated the ordinal valuables as continuous. All of these three methods have similar estimates for the measurement equations. However, the Bayesian Lasso and Bayesian Elastic Net are not converged when estimating the structure equation. Bayesian fused Lasso is converged and show the interaction between behavior risk index and marijuana morbidity.

The proposed model includes two way interaction of the exogenous latent variables, and it is straightforward to extend to three way interaction, when the problem has at least three exogenous latent variables. However, that will increase a great amount of the number of coefficients needed to estimate, depending on the number of knots. In our study, the options of the psychology survey are mostly ordinal data. In some cases, the options might be dichotomous and that would violate the continuous assumption of the manifest variable. Further research is needed to extend the manifest variable to binary and nominal response. Also it is worthwhile to extend it to other basis expansion methods.

Chapter 4

Discovering Gene Network and Interactions using Bayesian Graph Laplacian Model

4.1 Introduction

In chapter 2, we select important pathways and genes with the help of pathways information and genes relationship. The matrix R in chapter 2 indicates what genes are related, but does not specify how strong they are related. Also, there are other limitation: it is possible that different diseases cause genes interact differently within the same pathway; biology technology upgrades frequently and new findings in genetic research happen all the time, which means the genes relationship matrix R might be renewed every few years. It will be a great advantage in statistical analysis if we can remove the independence assumption, a priori between variables or a completely known dependence structure, *i.e.*, matrix R, when analyze the data. [Liu et al., 2014]

propose a Bayesian method that models the dependence structure through a graph Laplacian matrix. The main methods to find out similarity between data points and spectral clustering are Graph Laplacian matrices. We believe this method can be used to show the underlie dependent structure among the genes, which can be used as a potential guideline for further biological study about the interaction among genes.

4.2 Graph Laplacian Matrix

Spectral clustering algorithms concentrate on finding good clusters in statistical learning and data mining. One of the main tools of spectral clustering algorithms are graph Laplacian or the laplace matrices of graphs. Followed [von Luxburg, 2007], let similarity graph $G = (V, E)$ represent a set of n data points, and v_i represents a vertex, $i = 1, \cdots, n$ and E is a set of edge. Let $s_{ij} \geq 0$ be the measure of similarity between two vertices v_i and v_j, and they are connected by an edge $s_{ij} > 0$. Define a weighted adjacency matrix $\boldsymbol{W} = (w_{ji})_{i,j=1,\cdots,n}$ and $w_{ij} = w_{ji} \geq 0$. $d_i = \sum_{j=1}^{n} w_{ij}$ is the degree of a vertex v_i. Let $\boldsymbol{D} = diag(d_1, \cdots, d_n)$ and the graph Laplacian of G is $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$. The spectral clustering algorithms works effectively, but there are several limitations. In regression analysis, it clusters the independent variables rather than their coefficients; secondly, it is under an assumption that there is available information about the weighted adjacency matrix, but that is not necessary the case when we analyze gene relationship; lastly, the restriction of $w_{ij} \geq 0$ is not realistic by assuming positive partial correlations between all pairs of variables. [Liu et al., 2014] overcome these difficulties by extending the graph Laplacian model.

4.3 Graph Laplacian Model

Consider a linear regression:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1}$$

where dependent variable $\boldsymbol{Y}$ is a $n \times 1$ vector, independent variables $\boldsymbol{X}$ is $n \times p$ matrix, corresponding $\boldsymbol{\beta}$ is a $p \times 1$ vector and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$.

4.3.1 Prior Distribution

The prior distribution for $\boldsymbol{\beta}$ is:

$$\boldsymbol{\beta} = N(\mathbf{0}, \frac{\sigma^2}{r}\boldsymbol{\Lambda}^{-1}), \tag{4.2}$$

where $r \geq 0$ and $\boldsymbol{\Lambda}$ is the graph Laplacian matrix:

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 + \lambda_{11} + \sum_{j \neq 1} |\lambda_{1j}| & \lambda_{12} & \cdots & \lambda_{1p} \\ \lambda_{21} & 1 + \lambda_{22} + \sum_{j \neq 2} |\lambda_{2j}| & \cdots & \lambda_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \cdots & \cdots & 1 + \lambda_{pp} + \sum_{j \neq p} |\lambda_{pj}| \end{pmatrix} \tag{4.3}$$

where $\lambda_{ij} = \lambda_{ji}$ and $\lambda_{ii} > 0$

The prior for λ's is as follows:

$$\pi(\boldsymbol{\lambda}) \propto C_{a,b} |\boldsymbol{\Lambda}|^{-1/2} \prod_{i=1}^{p} \lambda_{ii}^{-3/2} \exp(-\frac{a^2}{2\lambda_{ii}}) I(\lambda_{ii} > 0) \prod_{j<i} |\lambda_{ij}|^{-3/2} \exp(-\frac{b^2}{2|\lambda_{ij}|}) \tag{4.4}$$

where $\boldsymbol{\lambda}$ is the collection of all λ's in $\boldsymbol{\Lambda}$ and $C_{a,b}$ is the normalizing constant.

The prior for σ^2 is $\pi(\sigma^2) \propto 1/\sigma^2$

4.3.2 Posterior Distribution

The likelihood function from (4.1) is:

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{X}, \boldsymbol{y}) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\} \tag{4.5}$$

After multiplying the priors of σ^2, $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, the joint posterior distribution is:

$$\begin{aligned} &\pi(\sigma^2, \boldsymbol{\beta}, \boldsymbol{\lambda} | \boldsymbol{X}, \boldsymbol{y}) \propto \\ &\sigma^{-(n+p+2)} \{\prod_i \lambda_{ii}^{-3/2} \prod_{j<i} |\lambda_{ij}|^{-3/2}\} \exp\{-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\} \\ &\times \exp\{-\frac{r}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{\Lambda}\boldsymbol{\beta} - \frac{a^2}{2}\sum_i \lambda_{ii}^{-1} - \frac{b^2}{2}\sum_{j<i} |\lambda_{ij}|^{-1}\} \end{aligned} \tag{4.6}$$

The full conditional posterior distribution for $\boldsymbol{\beta}$ is followed a normal distribution as:

$$\boldsymbol{\beta} | \sigma^2, \boldsymbol{\lambda}, \boldsymbol{X}, \boldsymbol{Y} \sim N_p((\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{\Lambda})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{\Lambda})^{-1}) \tag{4.7}$$

Integrating out $\boldsymbol{\beta}$ from (4.6), they get the posterior distribution of σ^2:

$$\sigma^2 | \boldsymbol{\lambda}, \boldsymbol{X}, \boldsymbol{y} \sim Inv - Gamma(n/2, \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{\Lambda})^{-1}\boldsymbol{X}')\boldsymbol{y}/2) \tag{4.8}$$

The conditional posterior distribution for $\boldsymbol{\lambda}$ does not have a closed form, but it

can be obtained as:

$$\pi(\boldsymbol{\lambda}|\boldsymbol{\beta},\sigma^2,\boldsymbol{X},\boldsymbol{y}) \propto \prod_i \lambda_{ii}^{-3/2} \prod_{j<i} |\lambda_{ij}|^{-3/2} \exp\{-\frac{r}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{\Lambda}\boldsymbol{\beta} - \frac{a^2}{2}\sum_i \lambda_{ii}^{-1} - \frac{b^2}{2}\sum_{i<j}|\lambda_{ij}|^{-1}\} \tag{4.9}$$

4.3.3 MCMC

In order to sample from (4.9), the parameter space is augmented. Let $\eta_{ij} = |\lambda_{ij}|$ and $c_{ij} = sign(\lambda_{ij})$. And c_{ij} can be either +1 or -1 here. Let p_{ij} be the probability that $c_{ij} = +1$, and c_{ij} follows Bernoulli distribution:

$$\pi(c_{ij}|\boldsymbol{\beta},\sigma^2,\boldsymbol{X},\boldsymbol{y}) = p_{ij} \tag{4.10}$$

where $p_{ij} = [1 + \exp\{-rb(|\beta_i - \beta_j| + |\beta_i + \beta_j|)/2\sigma\}]^{-1}$

$\boldsymbol{\eta}$ can be divided into 2 cases, η_{ii} and η_{ij}.

$$\eta_{ii}|\boldsymbol{\beta},\sigma^2,\boldsymbol{X},\boldsymbol{y} \sim Inv - N(a\sigma|\sqrt{r}\beta_i|^{-1}, a^2) \tag{4.11}$$

and

$$\eta_{ij}|\boldsymbol{\beta},\sigma^2,\boldsymbol{X},\boldsymbol{y} \sim Inv - N(b\sigma|\sqrt{r}(\beta_i + c_{ij}\beta_j)|^{-1}, b^2) \tag{4.12}$$

The Gibbs sampler is developed as follow:

- Update σ^2 from (4.6).
- Update $\boldsymbol{\beta}$ from (4.7).
- Update $\boldsymbol{c}$ from (4.10).

- Update $\boldsymbol{\eta}$ from (4.11) and (4.12). Set $\lambda_{ii} = \eta_{ii}$ and $\lambda_{ij} = c_{ij}\eta_{ij}$

4.3.4 Choice for Hyperparameters

Conditioned on $\boldsymbol{c}$ and $\boldsymbol{\beta}$, the hyperparameters r, a, and b are:

$$r|a, b, \boldsymbol{c}, \boldsymbol{\beta} \sim Gamma(\frac{p}{2} + h_r, \frac{\sum_i \beta_i^2}{2\sigma^2} + \frac{a\sum_i |\beta_i|}{2\sigma} + \frac{b\sum_{i<j} |\beta_i + c_{ij}\beta_j|}{2\sigma}) \tag{4.13}$$

$$a|r, b, \boldsymbol{c}, \boldsymbol{\beta} \sim \exp(g_a + \frac{r\sum_i |\beta_i|}{2\sigma}) \tag{4.14}$$

$$b|r, a, \boldsymbol{c}, \boldsymbol{\beta} \sim \exp(g_a + \frac{r\sum_{i<j} |\beta_i + c_{ij}\beta_j|}{2\sigma}) \tag{4.15}$$

In order to get a relatively flat prior, g_a, h_b and g_b should be small. In each iteration, these hyperparameters are updated by drawing samples from their full conditional distributions.

4.4 Software

RCPP package, BVSG.cpp, is available to preform Bayesian Graph Laplacian Model. g_a, h_b and g_b should be set to small values. The function *BVSGR* can be used to find out the posterior of β's and the correlation matrix between them. The function *myheatmap* can be used to show the correlation matrix graphically. Beside heat maps, it also includes the dependence structure among the β's.

4.5 Application

The data we use are from the Michigan prostate cancer study [Dhanasekaran et al., 2001]. In order to screen prostate cancer, Prostate Specific antigen (PSA) is used as a biomarker. [Dhanasekaran et al., 2001] shows that some of the genetic pathways relative to non-cancerous tissue seemed to be impaired in the prostate cancer, and [Tang et al., 2013] indicates 16 KEGG pathways might be related to prostate cancer. Due to the complexity of the gene interaction within a pathway, we use the Bayesian Graph Laplacian Model to model the pathway effect.

There are 101 patients with 7103 gene microarray expression in the data set. We select 77 patients who had preoperative prostate-specific antigen (PSA) information and 368 genes from 16 KEGG pathways related to prostate cancer in our study. We applied [Liu et al., 2014] method to the prostate cancer data. For each pathway, preoperative PSA is the response variable and the genes are independent variables.

We used 20,000 iterations with burnin 10,000 iterations. If the absolution values of the coefficients are greater than 2 or the correlation between two genes are great than 0.2 or less than -0.2, we consider those genes are important. Table 4.1 and 4.2 summarized those important genes. Based on literature research, * marked genes are related to cancer and § marked genes are related to prostate cancer. For example, one of the critical determinants for the development and progression of human prostate cancers is the androgen receptor (AR). In prostate cancer cells, AR-mediated gene expression is suppressed by inhibition of PI3K activity, and [Zhu et al., 2008] shows PIK3R1 (hsa:5295) is one of the primary genes they are interested in after the treatment with PI3K inhibitors. Figure 4.1 to 4.32 show the heat map and dependent structure among the genes in these 16 pathways. Graph Laplacian shows us what

genes might be related and how strong their relations are. For instance, Figure 4.5 shows the heat map of mTOR signaling pathway. The tiles off the diagonal show the correlation between two genes. The darker the color is, the stronger the correlation is. If the absolute value of the correlation between two genes are more than 0.2, the two genes are assumed to be related. Figure 4.6 shows the dependent structure among the genes. The number indicates the order of the genes in mTOR signaling pathway. For example, 4 denotes hsa:9706 and 12 denotes hsa:7248. Those two genes are connected with a blue line, which indicates that they are related. The thicker the blue line is, the stronger the two genes are related.

The information of how strong the genes are related is not available in the KEGG data, let alone the interaction between genes might be different among different diseases. By using the graph Laplacian model, we can find out underlie relationship between the genes in response to prostate cancer or other diseases.

Important Pathways	Important Genes(Entrez) $\|\beta\| > 2$						Important Genes correlation great than 0.2			
MAPK signaling pathway	3845	9261	1386	5062*	10746§	9479§	5320	6654*	776	3925
	5530						7157*	4217§	10746§	4609*
ErbB signaling pathway	6198*	5062*	6777	3725*	5291*	5894*	3725*	5062*	5295§	5894*
	673	3845								
mTOR signaling pathway	51719*	673	9706	6194*	6198*	7422*	51719*	6198*	6194*	9706
	7248	6195*	6197§	5291*	5295§	5296	6199*	3091§	5291*	7422*
							7249	7248	6195*	6197§
							5290*	5295§	3479	
Wnt signaling pathway	5530	8322*	8324	5881*	595	3725*	5332*	3725*	4316	4609*
	1488	6885*	5515	5516	1454		8454*	6424		
Axon guidance	2773	7852*	5881*	4775	5530	5532	5881*	5058*	64221	1969*
	5533	3983*	2534	5747*	998*	9475	1948	5530	9475	4690
	5058*	5062*	387§	3688	4690	64221	1949*	4233*	6387*	2932
	1969*	1947*	2048*	2050	23365§	3845				
	5362	6405*	10500	9901	6387*					
Focal adhesion	5728*	5747*	1292	3909	7058*	7060	10398	1281	5290*	3480§
	858	7422*	5062*				5295§	7414	2335	7058*
							3479	3725*		
Long-term potentiation	5530	6195*	5566	10411	5894*	673	5578§	4659	5894*	5908
	3845	5502					1387	5530	5906	5330*
							5332*			
Neurotrophin signaling pathway	5291*	3845	673	4217§	3667	9261	5295§	4217§	3725*	397
	998									
Insulin signaling pathway	7248	51763	5792*	5565	5573	31	2194	5567	5257	8835*
	6194*	2308	3845	6464	6198*	7249	6198*	31	673	5290*
	673	5106*	10891	5291*	2932	5584				
	5590	10211								
Pathways in cancer	3908	3909	5728*	8322*	8324	3815	3688	999*	3685*	3480§
	2261	7175	7184	2932	4193*	6772	5290*	5295§	2335	3815
	673	5979	2353	54583	4436	1488	3725*	2950§	5602*	7175
	2950§	5915	3728	5371	8554§	9063	3479	6772	3091§	4436
	5925*	329	9618	7422*	1436	2247	7042	5743*	329	

* marked indicates the gene related to cancer

§ marked indicates the gene related to prostate cancer

Table 4.1: Summary of the important genes in 16 pathways (1)

Important Pathways	Important Genes(Entrez) $\lvert\beta\rvert > 2$						Important Genes correlation great than 0.2			
Colorectal cancer	8322*	8324	5291*	4436	3845		5295§	3725*		
Endometrial cancer	3845	2932	842	6934	83439	5728*	6654*	5295§		
	4609*	5291*	5295§	5894*	673					
Glioma	1956*	5156	6464	5894*	673	5290*	3479	6464	6654*	1956*
	5291*	207	5728*	1019*	1021	5925	5154	5291*	3480§	5295§
	1869§	5335	3845				1021	5335	3845	
Prostate cancer	5728*	2308	3845	5925*	2932	7184*	1956*	3480§	3479	
	5291*	673								
Chronic myeloid leukemia	7042	3066*	1488	6777	4792	6464	3066*	5291*	5290*	5295§
	5925*	5291*	5296	5894*	673	3845	3845			
Non-small cell lung cancer	842	5291*	5296	1019*	1021	1869§	5578§	842	5291*	1021
	3845	1956*	6789*	2064*	5925*	5915	5295§	6654*	207	595
	369	5894*	673				3845	1956*	6655	

* marked indicates the gene related to cancer

§ marked indicates the gene related to prostate cancer

Table 4.2: Summary of the important genes in 16 pathways (2)

4.6 Discussion

In this Chapter, we apply the Bayesian Graph Laplacian Model to analyze the gene network and interaction in response to prostate cancer. The table 4.1 and 4.2 shows the model findings. According to current research literature, we notice that the important genes picked up by interaction relationship are more likely related to prostate cancer or cancer than those picked up by larger absolute value of β's. This might be because cancer is caused by the interaction of a set of genes rather than individual genes. In genetic study, the gene relationship is not available sometimes, and often incomplete. It is also possible that genes relate differently in response to different dis-

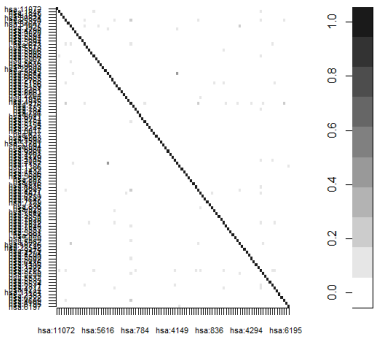


Figure 4.1: Heat Map of MAPK signaling pathway

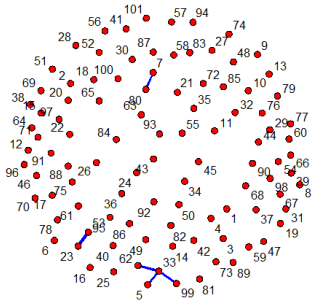


Figure 4.2: Dependence Structure among Genes of MAPK signaling pathway

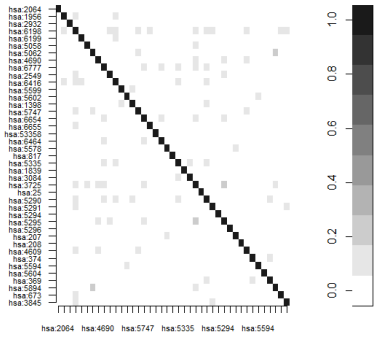


Figure 4.3: Heat Map of ErbB signaling pathway

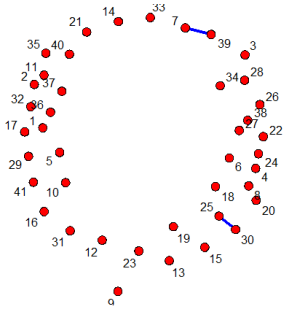


Figure 4.4: Dependence Structure among Genes of ErbB signaling pathway

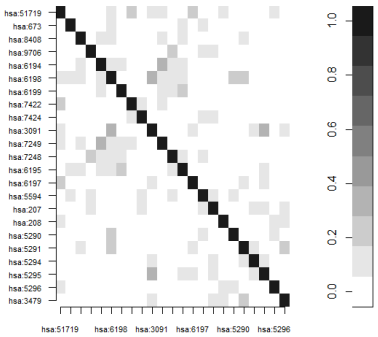


Figure 4.5: Heat Map of mTOR signaling pathway

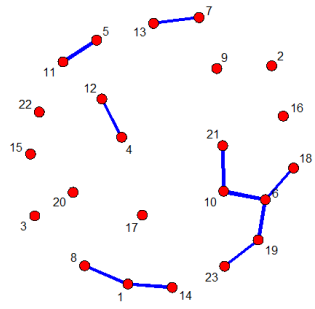


Figure 4.6: Dependence Structure among Genes of mTOR signaling pathway

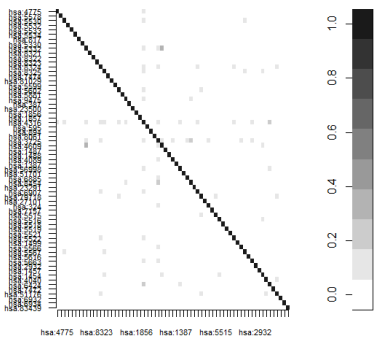


Figure 4.7: Heat Map of Wnt signaling pathway

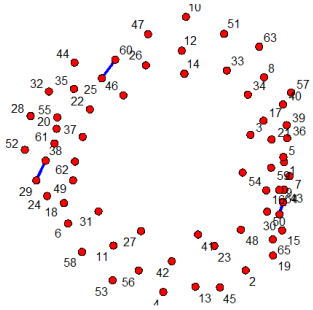


Figure 4.8: Dependence Structure among Genes of Wnt signaling pathway

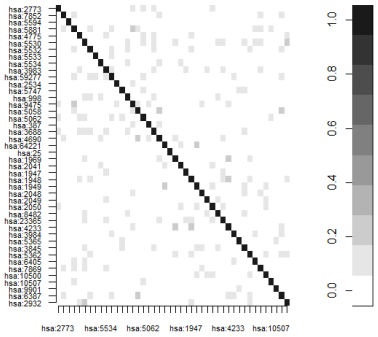


Figure 4.9: Heat Map of Axon guidance

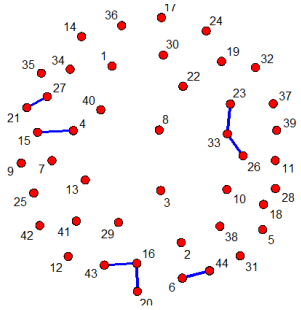


Figure 4.10: Dependence Structure among Genes of Axon guidance

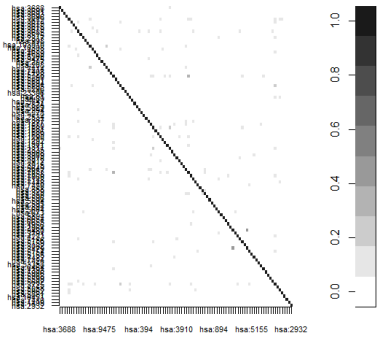


Figure 4.11: Heat Map of Focal adhesion

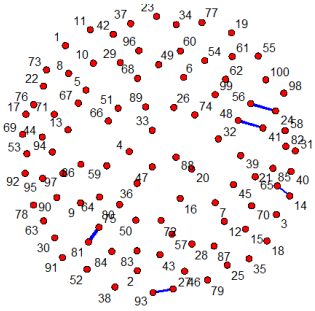


Figure 4.12: Dependence Structure among Genes of Focal adhesion

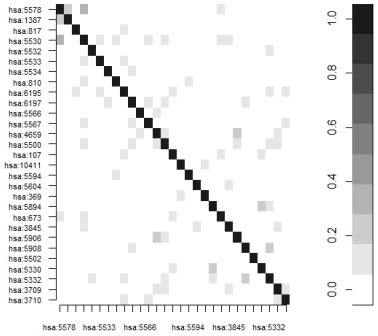


Figure 4.13: Heat Map of Long-term potentiation

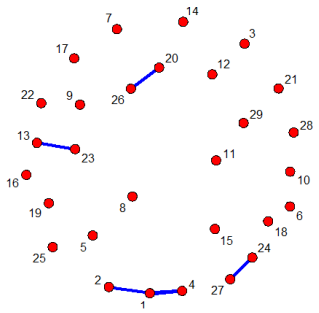


Figure 4.14: Dependence Structure among Genes of Long-term potentiation

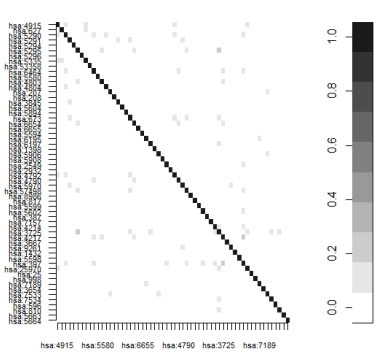


Figure 4.15: Heat Map of Neurotrophin signaling pathway

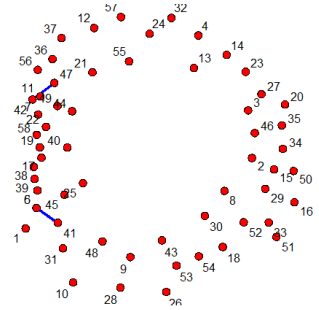


Figure 4.16: Dependence Structure among Genes of Neurotrophin signaling pathway

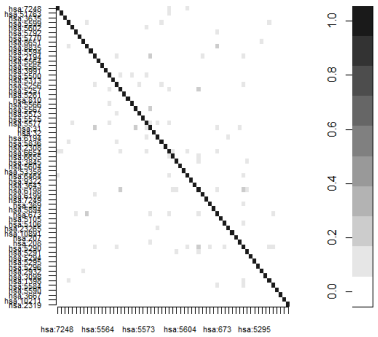


Figure 4.17: Heat Map of Insulin signaling pathway

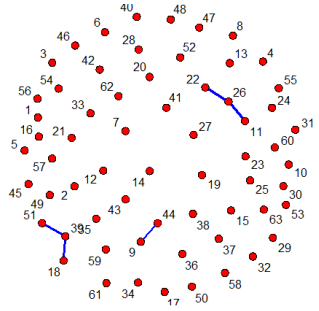


Figure 4.18: Dependence Structure among Genes of Insulin signaling pathway

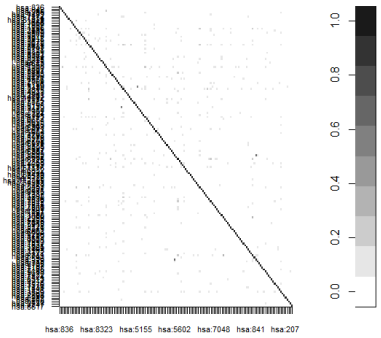


Figure 4.19: Heat Map of Pathways in cancer

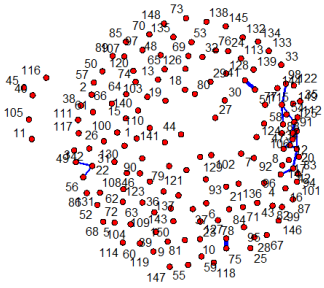


Figure 4.20: Dependence Structure among Genes of Pathways in cancer

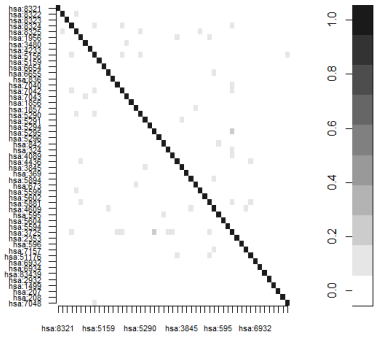


Figure 4.21: Heat Map of Colorectal cancer

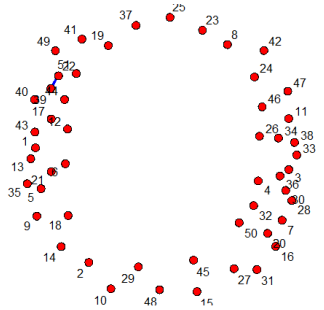


Figure 4.22: Dependence Structure among Genes of Colorectal cancer

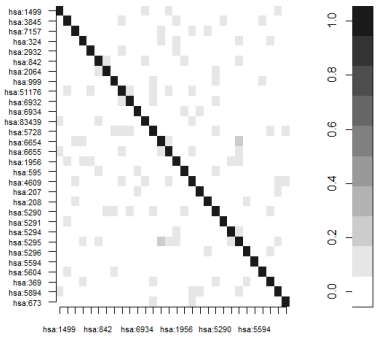


Figure 4.23: Heat Map of Endometrial cancer

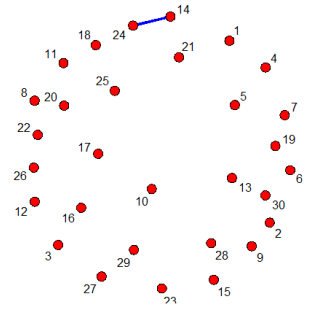


Figure 4.24: Dependence Structure among Genes of Endometrial cancer

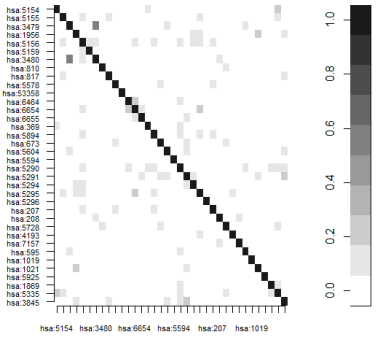


Figure 4.25: Heat Map of Glioma

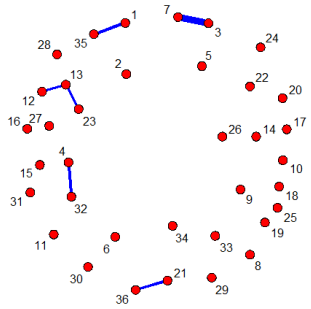


Figure 4.26: Dependence Structure among Genes of Glioma

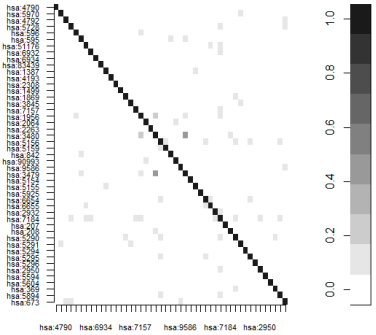


Figure 4.27: Heat Map of Prostate cancer

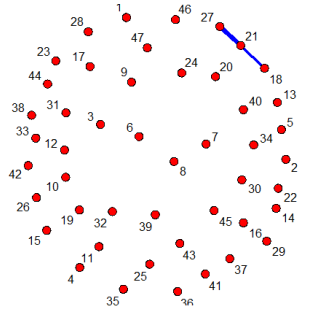


Figure 4.28: Dependence Structure among Genes of Prostate cancer

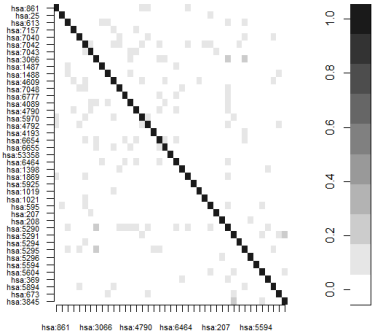


Figure 4.29: Heat Map of Chronic myeloid leukemia

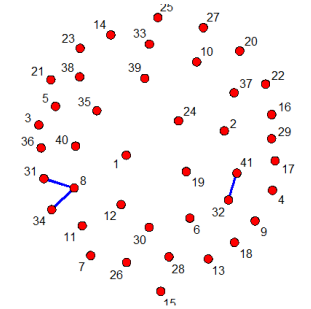


Figure 4.30: Dependence Structure among Genes of Chronic myeloid leukemia

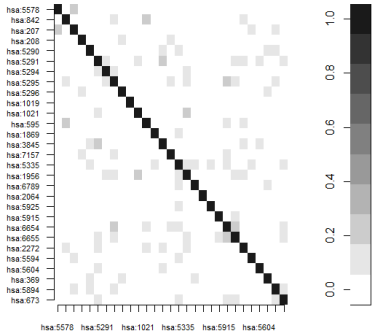


Figure 4.31: Heat Map of Non-small cell lung cancer

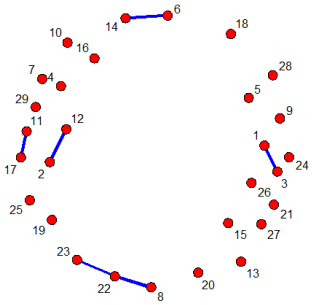


Figure 4.32: Dependence Structure among Genes of Non-small cell lung cancer

eases. One of the advantages of Bayesian Graph Laplacian Model is that a completely known dependence structure is not needed.

[Liu et al., 2014] compares the performance of Bayesian Graph Laplacian Model with that of Lasso, EN, OSCAR, Bayesian Lasso and Bayesian Elastic Net in 5 different scenario simulation studies. Bayesian Graph Laplacian Model preform best in four scenarios and second to the best in one scenario.

[Liu et al., 2014] proposed method can be used for one pathway each time, with the information that we already know the potential important pathways. Some of the diseases, especially cancers, might be caused by the multiplied pathways interacting with each other. It is worthwhile to extend the method to pin point important pathways among all potential disease related pathways. We could possibly find out the interaction between pathways through the interaction among genes, because one disease related gene can be in multiple different pathways.

Chapter 5

Future Study

5.1 Multiple Pathways Simultaneous Analysis and Pathways Selections

When we analyze the gene expression data in Chapter 4, one of the limitations of Bayesian Graph Laplacian Model is that we can only analyze each pathway individually. However, in disease research, especially in cancer, it is important to consider the interaction between pathways, as some of the diseases are the result of several pathways interactions. According to the current biology research, there are more than 200 pathways information available ,but most of the pathways are unrelated to the diseases, so it is important for us to select possible disease related pathways. We would like to extend the Graph Laplacian method with pathway selection.

First consider a linear regression function,

$$y_i = \boldsymbol{x}_i^{(1)T}\boldsymbol{\beta}_1 + \boldsymbol{x}_i^{(2)T}\boldsymbol{\beta}_2 + \cdots + \boldsymbol{x}_i^{(L)T}\boldsymbol{\beta}_L + e_i, \tag{5.1}$$

where $e_i \sim N(0, \sigma^2)$, y_i is continuous variables for ith observation, $i = 1, \cdots, n$, $\boldsymbol{x}_i^{(l)}$ is the microarray expression data for lth pathway, and $l = 1, \cdots, L$. In this section y_i is continuous variable , similar to preoperative PSA. $\boldsymbol{x}_i^{(l)T}\boldsymbol{\beta}_l$ is the lth pathway effect and we assume additive effect on the response variable.

Similar to chapter 2, we assume the regression coefficients arise from a scale mixture of a point mass of 0 and a normal distribution, so we have:

$$\boldsymbol{\beta}_l|\phi_l = \phi_l N_{p_l}(\mathbf{0}, \frac{\sigma^2}{r_l}\boldsymbol{\Lambda}_l^{-1}) + (1-\phi_l)I(0), \tag{5.2}$$

where ϕ_l is the pathway selection indicator for lth pathway,

$$\begin{cases} \phi_l = 1 & \text{when pathway is } l \text{ selected in the model} \\ \phi_l = 0 & \text{otherwise} \end{cases} \tag{5.3}$$

and it follows Bernoulli distribution:

$$\phi_l = \omega_l^{\phi_l}(1-\omega_l)^{\phi_l} \tag{5.4}$$

And $\sigma_l^2 \sim IG(a_l, b_l)$, a, b, $r_l \geq 0$ and ω_l are hyperparameters. Moreover, $\boldsymbol{\Lambda}_l^{-1}$ is the inverse of the graph Laplacian matrix, and it is as follows:

$$\boldsymbol{\Lambda}_l = \begin{pmatrix} 1+\lambda_{11}^{(l)}+\sum_{j\neq 1}|\lambda_{1j}^{(l)}| & \lambda_{12}^{(l)} & \cdots & \lambda_{1p}^{(l)} \\ \lambda_{21}^{(l)} & 1+\lambda_{22}^{(l)}+\sum_{j\neq 2}|\lambda_{2j}^{(l)}| & \cdots & \lambda_{2p}^{(l)} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1}^{(l)} & \cdots & \cdots & 1+\lambda_{pp}^{(l)}+\sum_{j\neq p}|\lambda_{pj}^{(l)}| \end{pmatrix} \tag{5.5}$$

where $\lambda_{ij}^{(l)} = \lambda_{ji}^{(l)}$ and $\lambda_{ii}^{(l)} > 0$ The prior for $\lambda^{(l)}$ is:

$$\pi(\boldsymbol{\lambda}^{(l)}) \propto C_{a,b}|\boldsymbol{\Lambda}_l|^{-1/2} \prod_{i=1}^{p} (\lambda_{ii}^{(l)})^{-3/2} \exp(-\frac{a^2}{2\lambda_{ii}^{(l)}}) I(\lambda_{ii}^{(l)} > 0) \prod_{j<i}^{(l)} |\lambda_{ij}|^{-3/2} \exp(-\frac{b^2}{2|\lambda_{ij}^{(l)}|}) \tag{5.6}$$

The prior distribution for β's will shrink the β's values close to 0 when those β's and corresponding pathways are not important. In such way, we can select pathways.

5.2 Survival Time as Response Variable

In disease related research, it is common to use survival time as the response variable. We would like to extend the Graph Laplacian method with right censored response. We replace the original regression model with Accelerated Failure Time (AFT) models. AFT models assume multiplicative effect of the pathway effect on the survival time:

$$\log(t_i) = \boldsymbol{x}_i^{(1)T}\boldsymbol{\beta}_1 + \boldsymbol{x}_i^{(2)T}\boldsymbol{\beta}_2 + \cdots + \boldsymbol{x}_i^{(L)T}\boldsymbol{\beta}_L + e_i, \tag{5.7}$$

where we have the survival time t_i for subject i. Let c_i be the censoring time independent of t_i. Let $\delta_i = I\{t_i \leq c_i\}$ to be censored indicator function and $t_i^* = \min(t_i, c_i)$. We impute the censored data by using the [Tanner and Wong, 1987] data augmentation approach. Let $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)'$, and y_i is the augmented data as,

$$\begin{cases} Y_i = \log(t_i^*) & \text{if} \quad \delta_i = 1 \\ Y_i > \log(t_i^*) & \text{if} \quad \delta_i = 0 \end{cases} \tag{5.8}$$

Assuming the error term is *iid* following standard normal distribution, the model (5.7) becomes:

$$Y_i = \boldsymbol{x}_i^{(1)T}\boldsymbol{\beta}_1 + \boldsymbol{x}_i^{(2)T}\boldsymbol{\beta}_2 + \cdots + \boldsymbol{x}_i^{(L)T}\boldsymbol{\beta}_L + e_i, \quad e_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{5.9}$$

5.3 Binary Response Variable

In medical research, sometimes, researchers are interested in classifying or discriminating different diseases or cancers. In this case, the response variables are dichotomous or categorical. So we would like to extend the Graph Laplacian method with binary response and multinomial response to analyze those research questions.

Suppose y_i is binary random variables, where $i = 1, \cdots, n$, and $y_i = 1$ or $y_i = 0$. We use the data augmentation method introduced by [Tanner and Wong, 1987]. Let $z_1, \cdots, z_n$ be n latent variables, and z_i are independent $N(\boldsymbol{X}_i^T\boldsymbol{B}, 1)$, where $\boldsymbol{X}_i = c((\boldsymbol{x}_i^{(1)})^T, \cdots, (\boldsymbol{x}_i^{(L)})^T)^T$ and $\boldsymbol{B} = c(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_L)$. Let

$$\begin{cases} y_i = 1 & \text{if} \quad z_i > 0 \\ y_i = 0 & \text{if} \quad z_1 \leq 0 \end{cases} \tag{5.10}$$

As a result, y_i follows Bernoulli distribution with $p_i = P(y_1 = 1) = \Phi(\boldsymbol{X}_i^T\boldsymbol{B})$. The prior of $\boldsymbol{\beta}_l$ is the same as (5.2).

For the multinomial response, if the responses are ordinal variables, we suppose y_i takes K ordered categories. Let $p_{ik} = P(y_i = k)$ and cumulative probabilities $\eta_{ik} = \sum_{k=1}^{K} p_{ik}$, where $k = 1, \cdots, K-1$. Following [McCullagh, 1980], we have $\eta_{ik} = \Phi(\gamma_k - \boldsymbol{X}_i^T\boldsymbol{B})$. Similarly, if z_i are independent $N(\boldsymbol{X}_i^T\boldsymbol{B}, 1)$, let $y_i = k$, when

we have $\gamma_{k-1} < z_i \leq \gamma_k$.

Lastly, if the responses are nominal variables. [Aitchison and Bennett, 1970] applied Gibbs sampling approach to multinomial probit model. In our specific model setting, let $\boldsymbol{z}_i = z_{i1}, \cdots, z_{iK}$, where $i = 1, \cdots, n$ and $K > 2$, we have:

$$z_{ik} = \boldsymbol{X}_{ik}^T \boldsymbol{B}_k + e_{ik}, \tag{5.11}$$

where $\boldsymbol{e}_i = (e_{i1}, \cdots, e_{iK})$. Pathway selections are different in different diseases, so disease k has its own set of genes $\boldsymbol{X}_{ik}$ and corresponding $\boldsymbol{\beta}_k$. Disease k is observed if $z_{ik} > z_{im}$ for all $k \neq m$.

We believe with these three extensions, Bayesian Graph Laplacian Model can be used in most of biological or medical research data.

Bibliography

J. Aitchison and J. A. Bennett. Polychotomous quantal response by maximum indicant. *Biometrika*, 57:253–262, 1970.

A. Annest, R. E. Bumgarner, A. E. Raftery, and K. Y. Yeung. Iterative bayesian model averaging: a method for the application of survival analysis to high- dimensional microarray data. *BMC Bioinformatics*, 10:72, 2009.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101:119–137, 2006.

D. J. Bauer. A semiparametric approach to modeling nonlinear relations among latent variables. *Structural Equation Modeling*, 12:513–535, 2005.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36:192236, 1974.

N. Breslow and J. Crowley. A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics*, 2:437–453, 1974.

S. Chakraborty. Simultaneouscancerclassification and geneselection with bayesian nearest neighbor method: An integrated approach. *Computational Statistics and Data Analysis*, 53:1462–1474, 2009.

S. Chakraborty, B. K. Mallick, D. Ghosh, M. Ghosh, and E. Dougherty. Gene expression-based glioma classification using hierarchical bayesian vector machines. *Sankhy*, 69:514–547, 2007.

S. Chen. Aromatase and breast cancer. *Frontiers in Bioscience*, 3:922–933, 1998.

H. Chipman, George E. I., McCulloch R. E., M. Clyde, Foster D. P., and R. A. Stine. The practical implementation of bayesian model selection. *IMS Lecture Notes-Monograph Series*, 38, 2001.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

F. J. DeMayo, B. Zhao, N. Takamoto, and S. Y. Tsai. Mechanisms of action of estrogen and progesterone. *Annals of the New York Academy of Sciences*, 955: 48–59, 2002.

S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412:822–826, 2001.

Lee K. E., N. Sha, E. R. Dougherty, M Vannucci, and B. K. Mallick. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 2003.

B. Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72:557–565, 1977.

L. Fahrmeir and A. Raach. A bayesian semiparametric latent variable model for mixed responses. *Psychometrika*, 72:327–346, 2007.

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 2005.

S. K. Ghosh and S. Ghosal. Semiparametric accelerated failure time models for censored data. *Bayesian Statistics and Its Applications*, pages 213–229, 2005.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531537, 1999.

J Gui and H Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21:3001–3008, 2005.

R. Guo, H. Zhu, S. Chow, and J. G. Ibrahim. Bayesian lasso for semiparametric structural euqation models. *Biometrics*, 68:567–577, 2012.

D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100:5770, 2000.

T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York, New York, fifth edition, 2009.

K. G. Jőreskog. *A general method for estimating a linear structural equation system.* In Structural Equation Models in the Social Sciences: A. S. Goldberger and O. D. Duncan(eds). 85-112, New York: Seminar Press, 1973.

M. Kanehisa and S. Goto. Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:2730, 2000.

L. Keele. *Semiparametric Regression for the Social Sciences.* Wiley, Chichester UK, 2008.

D. A. Kenny and C. M. Judd. Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96:201–210, 1984.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian splinefunctions. *Journal of Mathematical Analysis and Applications*, 33:8295, 1971.

L. Kuo and B. Mallick. Bayesian semiparametric inference for the accelerated failuretime model. *Canadian Journal of Statistics*, 25:457–472, 1997.

M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5:369–412, 2010.

Felice D. L., L. El-Shennawy, S. Zhao, D. L. Lantvit, Q. Shen, T. G. Unterman, S. M. Swanson, and J. Frasor. Growth hormone potentiates 17β-estradiol-dependent breast cancer cell proliferation independently of igf-i receptor signaling. *Endocrinology*, 154:3219–3227, 2013.

S. Lee and H. Zhu. Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, 53:209232, 2000.

S.Y. Lee. *Stuctural Equation Modeling: A Bayesian Approach*. Wiley, Chichester, England, 2007.

F. Li and N. R. Zhang. Bayesian variable selection in structured high-dimensional covariate space with application in genomics. *Journal of the American Statistical Association*, 105:1202–1214, 2010.

Q. Li and N. Lin. The bayesian elastic net. *Bayesian Analysis*, 5:151–170, 2010.

B. Liu, A. Fuente, and I. Hoeschele. Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, 178:1763–1776, 2008.

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

F. Liu, S. Chakraborty, F. Li, Y. Liu, and A. C. Lozano. Bayesian regulation via graph laplacian. *Bayesian Analysis*, 9:449–474, 2014.

Gutman M., S. Couillard, F. Labrie, B. Candas, and C. Labrie. Effects of the antiestrogen em-800 (sch 57050) and cyclophosphamide alone and in combination on growth of human zr-75-1 breast cancer xenografts in nude mice. *Cancer Research*, 59:5176–5180, 1999.

M. P. Martens. The use of structural equation modeling in counseling psychology research. *The Counseling Psychologist*, 33:269–298, 2005.

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B*, 42:109–142, 1980.

U. Meinhardt and P. E. Mullis. The aromatase cytochrome p-450 and its clinical impact. *Hormone Research*, 57:145–152, 2002.

V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature*, 34:267–273, 2003.

N. Sha, M. Vannucci, M. G. Tadesse, P. J. Brown, I. Dragoni, N. Davies, T. C. Roberts, A. Contestabile, M. Salmon, C. Buckley, and F. Falciani. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, 60:812819, 2004.

N. Sha, M. G. Tadesse, and M. Vannucci. Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, 22:2262–2268, 2006.

C. J. Sherr. Cancer cell cycles. *Science*, 274:1672–1677, 1996.

D. Sinha, M. Chen, and S. K. Ghosh. Bayesian analysis and predictive model diagnostics for interval-censored survival data. *Annals of Statistics*, 55:585–590, 1999.

F. C. Stingo, Y. A. Chen, M. G. Tadesse, and M. Vannucci. Incorporating biological information into linear models: A bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*, 5:1978–2002, 2011.

Y. Tang, W. Yan, J. Chen, C. Luo, A. Kaipia, and B. Shen. Identification of novel

microrna regulatory pathways associated with heterogeneous prostate cancer. *BMC Systems Biology*, 7, 2013.

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

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16:385–395, 1997.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, 67:91–108, 2005a.

R. Tibshirani, M. Saunders, S. Rosset, J Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:91–108, 2005b.

R. J. Tibshirani. Univariate shrinkage in the cox model for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, 8:Article 21, 2009.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17:520–525, 2001.

L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y.D. He, A. A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber,

R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17: 395–416, 2007.

Q. Zhu, H. Youn, J. Tang, O. Tawfik, K. Dennis, P. F. Terranova, J. Du, P. Raynal, J. B. Thrasher, and B. Li. Phosphoinositide 3-oh kinase p85 and p110 are essential for androgen receptor transactivation and tumor progression in prostate cancers. *Oncogene*, 27:45694579, 2008.

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

VITA

Zhenyu Wang was born in Guangzhou, China. He got his bachelor degree in Business at Sichuan University and He received his master degree and PhD in Statistics at the University of Missouri. He will serve as a Postdoctoral Associate in the Department of Biostatistics and Bioinformatics at Duke University in August 2014.