Summer 2022

Statistical Methods for Analyzing Multi-omics Data: Dependence Structure and Missing Values

Wenda Zhang

Statistical Methods for Analyzing Multi-omics Data: Dependence Structure and Missing Values

by

Wenda Zhang

Bachelor of Science
Beijing University of Technology, 2012

Master of Science
Beijing WUZI University, 2016

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2022

Accepted by:

Yen-Yi Ho, Major Professor

Lianming Wang, Committee Member

Karl Gregory, Committee Member

Daping Fan, Committee Member

Tracey L. Weldon, Vice Provost and Dean of the Graduate School

Dedication

This dissertation is dedicated to

MY WIFE

Siqi Yang

MY MOTHER & MY FATHER

Liping Wang & Wushun Zhang

All my families, friends, and people who support me

Acknowledgments

First of all, I want to thank my advisor Dr. Yen-Yi Ho from the bottom of my heart. She opened the door to academic research for me and took me into the hall of statistics. She is a knowledgeable mentor, and a great person. Her enthusiasm for statistics, and attitude towards research deeply influenced me. Dr. Ho provided me with many valuable and pertinent suggestions, which enabled me to overcome difficulties and challenges in the study and research work of my doctoral career. I have benefited a lot from this.

Secondly, I want to thank my committee members Dr. Lianming Wang, Dr. Karl Gregory, and Dr. Daping Fan. They offered many invaluable and useful suggestions for my doctoral dissertation.

Thirdly, I want to thank my wife, my best friend, and my soulmate, Siqi Yang. She always stands behind me unconditionally to support me, encourage me, understand me and love me. During my doctoral career, I have been frustrated and self-doubting in the face of setbacks. It was her love and support that lifted me up again and again and made it possible for me to get through those hard times with courage and strength. She is my warrior, my spiritual support, and my everything. So I would like to give her my deepest gratitude.

Last but not the least, I would like to thank my parents, my families, my friends and people who helped me and supported me.

Abstract

The advancements in high-throughput technologies have made it possible to generate a huge number of "omics" data, including genomics, proteomics, transcriptomics, epigenomics, metabolomics, and microbiomics. Combining multiple data sources and performing joint analyses with all available information and the phenotypic outcome can reflect various aspects in complex biological systems, such as revealing regulation processes, discovering novel associations between biological entities, and identifying relevant biomarkers for certain diseases or phenotypic outcomes. This dissertation focuses on developing statistical models for analyzing multi-omics data. It is comprised of three topics: (1) integrative analysis for multi-omics data with missing observations in intermediate variables; (2) modeling the dynamic gene co-expression (DC) in a genome-wide search space with the implementation of variable selection techniques in a Bayesian framework; and (3) mixed-effect variable selection model for identifying DC using scRNA-seq count data and DC-based strategy in subject subgroup classification.

In Chapter 2, we propose a novel integrative multi-omics analytical framework based on p-value weight adjustment in order to incorporate observations with a large proportion of missing values in the analysis. The occurrence of missing values is an inevitable issue in multi-omics data because some measurements, such as mRNA gene expression levels, often require invasive tissue sampling from patients. To incorporate the incomplete information from measurements with missing records, we split the data into a complete set with full information and an incomplete set with missing measurements, and introduce mechanisms to derive weights and weight-adjusted

p-values from the two sets. Through simulation analyses, we demonstrate that the proposed framework achieves considerable statistical power gains compared to a complete case analysis or multiple imputation approaches. In experimental data analysis, the implementation of our proposed framework is illustrated by a joint analysis of DNA methylation, mRNA, and the phenotypic outcome in a study of preterm infant birth weight.

Chapter 3 proposes two models to apply the genome-wide search for identifying dynamic gene co-expression from genomic datasets. In a biological system, genetic interactions are tightly regulated and are often highly dynamic. The interactions can change flexibly under various internal cellular signals or external stimuli. Previous studies have developed statistical methods to examine these dynamic changes in genetic interactions. However, a common challenge encountered in the existing approaches is the computational intensiveness due to the massive number of gene combinations needed to be considered in a typical genomic dataset, yet often a much smaller proportion of gene interactions exhibit dynamic co-expression changes.

To solve this problem, we propose variable selection methods in Bayesian frameworks with spike-and-slab priors. The proposed algorithms focus on subsets of promising gene combinations in the search space. We also adopt a Bayesian false discovery control procedure for testing the significance of dynamic gene co-expression changes. A series of simulation studies are then conducted to present a comparison between our proposed approaches to the existing exhaustive search heuristics. We also demonstrate the implementation of our proposed approaches in experimental data analysis to study gene co-expression changes associated with colorectal cancer recurrence-free survival.

In Chapter 4, we develop subject-specific methods which combines mixed-effect model and spike-and-slab variable selection method in identifying DC for scRNA-seq read count data. In a typical scRNA-seq dataset, gene expression profiles are usually

collected from multiple subjects such as different patients or tissues. Considering subject-specific characterization can increase the accuracy of identifying DC and the sparsity of DC signals in scRNA-seq datasets, we propose a mixed-effect variable selection model for identifying subject-specific DC gene pairs while incorporating both subject-specific random effects, across-specific fixed effects, zero-inflation and over-dispersion in scRNA-seq datasets. We also propose a DC-based strategy to classify subjects into subgroups by using subject-specific DC as inputs. Through simulation study, we show that our proposed ME-SPSL model outperforms the mixed-effect model without using variable selection technique and the existing method without considering subject-specific random effects. We also demonstrate the implementation of our proposed method in a melanoma scRNA-seq dataset to estimate subject-specific DC and use the DC gene pairs as the biomarkers to classify the melanoma samples into immunotheropy-resistant and non-resistant groups.

Table of Contents

Dedication iii

Acknowledgments iv

Abstract v

List of Tables xi

List of Figures xiii

Chapter 1 Introduction 1

1.1 Multi-omics Integrative Analysis 1

1.2 Missing Values and Data Imputation Algorithms 1

1.3 Dynamic Gene Co-expression 2

1.4 Bayesian variable selection: Spike-and-slab priors 4

1.5 Structure of the dissertation 7

Chapter 2 Multi-omics integrative analysis for incomplete data 9

2.1 Introduction 9

2.2 Datasets and Databases 11

2.3 Models 11

2.4 Simulation 15

2.5 Experimental Data Analysis 23

2.6 Discussion 27

Chapter 3 Identification of dynamic gene co-expression via Bayesian variable selection 32

3.1 Introduction 32

3.2 Datasets and Databases 33

3.3 Model 34

3.4 Simulation 40

3.5 Experimental Data Analysis 43

3.6 Discussion 47

Chapter 4 Subject-specific Model for Identifying Dynamic Gene Co-expression as Biomarkers for Classification using Bayesian Variable Selection for Single-Cell Count Data 50

4.1 Introduction 50

4.2 Materials and Methods 53

4.3 Simulation 56

4.4 Experimental Data Analysis 63

4.5 Discussion 68

Bibliography 70

Appendix A Likelihood and Posterior Distributions for SPSL and C-SPSL 78

Appendix B Adaptive Metropolis-Hastings Sampling Algorithm . 82

Appendix C Empirical distributions of $\widehat{\tau}_1$ in CRC Dataset 86

Appendix D Convergence Diagnostics for SPSL and C-SPSL . . . 87

Appendix E Determination of the Recurrence-Free Survival Time 93

List of Tables

Table 2.1 Table for FWER after weighted Bonferroni method and FDR after weighted BH method when all associations were set equal to 0. 20

Table 2.2 Type I error control for FWER using weighted Bonferroni method and FDR using q-value method when $\gamma_{MG} = \gamma_{GY} = 0$ under MAR. 23

Table 2.3 Top table for 30 CpG sites associated with birth weight scores of preterm infants with the smallest weighted p-values derived from the proposed omnibus method. Chromosomes, CpG sites, UCSC RefSeq gene names, weighted p-values and q-values are reported. The results were adjusted for paternal age, maternal age, paternal BMI, maternal BMI, maternal smoking status before pregnancy, and the gender of infants. 28

Table 3.1 Comparison of ES, SPSL and C-SPSL model based on 100 simulation iterations in scenario I (sparsity = 70%). The true values of $\tau_{\mathbf{1}}$ are set to be $(0, 0, 0, 0, 0, 0, 0, 1, 1, 1)^T$ and the true values $\tau_{\mathbf{0}}$ are set to 0. The false discovery rate (FDR) and false negative rate (FNR) are reported. 42

Table 3.2 Comparison of SPSL and C-SPSL model based on 100 simulation iterations in scenario II with with 1,225 gene pairs. The true values of $\tau_{\mathbf{1}} = 0$ except for 10 gene pairs $\tau_{\mathbf{1}} = 1$ (sparsity=99.2%). The true values of $\tau_{\mathbf{0}}$ are set to be all 0. The false discovery rate (FDR) and false negative rate (FNR) are reported. . 43

Table 3.3 The top 30 significant gene pairs with the largest posterior mean $|\widehat{\tau}_1|$ associated with recurrence-free survival time. The 95% posterior credible intervals (95% CI) for each parameter are provided in the parentheses. 45

Table 4.1 Gelman-Rubin convergence diagnostics results for a randomly picked sample among 200 simulations. Point estimates of the potential scale reduction factor (Point est.) and their upper confidence limits (Upper C.I.) are provided. 60

Table 4.2 Coverage probability (CP) and the length of 95% credible interval (CI length), mean square errors (MSE), mean bias errors (MBE) are used as metrics to evaluate the performance of estimating τ_1 ranging from 0 to 0.5 for ME-SPSL, ZENCO-ME and ZENCO models based on 200 simulations. 62

Table 4.3 The top 30 significant gene pairs with the largest posterior mean $|\widehat{\tau}_1|$ associated with OE score. The 95% posterior credible intervals of $|\widehat{\tau}_1|$ (95% CI) are provided. 66

Table D.1 Gelman-Rubin diagnostics results for SPSL model with 5 genes (10 pairs) and sample size of 200. 88

Table D.2 Gelman-Rubin diagnostics results for C-SPSL model with 5 genes (10 pairs) and sample size of 200. 88

Table D.3 Gelman-Rubin diagnostics results of $\tau_{1,j}$ for the top 30 gene pairs with the largest $|\widehat{\tau}_{1,j}|$. 91

List of Figures

Figure 2.1 The schematic diagram for partitions of the complete set and incomplete set in datasets with missing intermediate mRNA measurements. Dimensions, sample sizes, the complete set, and the incomplete set are shown in the plot. 12

Figure 2.2 Power comparisons of omnibus method, MICE, KNN imputation, IG and linear model for various missing rates and γ_{MG}. The value of γ_{GY} was set equal to 0.1, 0.2 and 0.5. The standard deviations of both γ_{MG} and γ_{GY} were set equal to 1. 21

Figure 2.3 Power comparisons of omnibus method, general weight scheme, reverse weight scheme, IG and linear model for various γ_{MG} and γ_{GY} with different missing rates. 22

Figure 2.4 Power curves for omnibus method and competing methods in high-dimensional case. The number of factors $k = 5$ was determined by the 10-fold cross-validation. The gene-phenotype association γ_{GY} was set to be 0.1, 0.2 and 0.5. 22

Figure 2.5 Power comparisons of omnibus method, MICE, KNN imputation, IG and linear model for various missing rates and γ_{MG} under MAR. The value of γ_{GY} was set equal to 0.1, 0.2 and 0.5. The standard deviations of both γ_{MG} and γ_{GY} were set equal to 1. 24

Figure 2.6 Power comparisons of omnibus method, general weighting scheme, reverse weighting scheme, IG and linear model for various γ_{MG} and γ_{GY} with different missing rates under MAR. The standard deviations of γ_{MG} and γ_{GY} were both set equal to 1. 25

Figure 2.7 Power curves for omnibus method and competing methods in high-dimensional case under MAR. The number of factors $k = 5$ was determined by the 10-fold cross-validation. The value of γ_{GY} was set to be 0.1, 0.2 and 0.5. 25

Figure 2.8 QQ-plots for p-values after SVA and genomic inflation factor adjustment for the proposed omnibus method. 26

Figure 2.9 Plot for percentage of variance explained by the principal components (PC) over the number of PC. The first 10 principal components explain 97.9% of the variance of gene expression data. 27

Figure 3.1 Conditional density plots for the bimodal inverse gamma distributions with $\alpha_1 = 5$ and $\alpha_2 = 50$. Plot (a) and (b) present the different shapes of the densities with $w = 0.2$ and $w = 0.95$ respectively while fixing $v_0 = 0.005$. Plot (c) and (d) are generated with fixing $v_0 = 0.08$ and varied $w = 0.2$ and $w = 0.95$ respectively. 37

Figure 3.2 Profile plots for the top two gene pairs with the largest $|\widehat{\tau}_1|$ associated with recurrence-free survival time. (a) presents the profile plots between *C18orf12* and *MEP1B* with $\widehat{\tau}_1 = 2.98$; (b) presents the profile plots between *SCEL* and *TGIF1* with $\widehat{\tau}_1 = -2.93$. Gene expression levels were standardized to have mean 0 and variance 1 in both plots. 46

Figure 3.3 Network plot for gene pairs with $|\widehat{\tau}_{1,j}| \geq 2.5$ (33 genes with 19 pairs). Nodes represent genes and the thickness of the link edges indicate the values of $|\widehat{\tau}_{1,j}|$. Dotted lines indicate gene pairs with negative $\widehat{\tau}_{1,j}$ while solid lines denote positive $\widehat{\tau}_{1,j}$ values. 48

Figure 4.1 Power curves for τ_1 for ME-SPSL, ZENCO-ME, and ZENCO model based on 200 simulations and fixed σ_1^2. 59

Figure 4.2 Power curves for τ_1 for ME-SPSL, ZENCO-ME and ZENCO model based on 200 simulations and random σ_1^2. 61

Figure 4.3 ROC curves for subgroup classification. Black lines represent the average ROC curve for ME-SPSL method and red dashed lines represent the ZENCO-ME method. 63

Figure 4.4 Density plots of the posterior mean ($\widehat{\tau}_1$) for the groups including or excluding zero from the 95% credible intervals (95% CI) of $\widehat{\tau}_1$. The value of ϵ denotes the minimum of $|\widehat{\tau}_1|$ in the significant group that 95% CI excludes 0. 65

Figure 4.5 Average ROC curves for validation sets of ICI-resistant classification. The average AUC score of DC-based method (black solid line) is 0.65 and the average AUC score of the expression-based method (red dashed line) is 0.61. 67

Figure C.1 Empirical density plot of $\widehat{\tau}_1$ for all $949,280$ gene-pair combinations. The proportion of $|\widehat{\tau}_1|$ being less than 0.5 is 0.977. 86

Figure D.1 Trace plots for SPSL model with sample size of 200. 89

Figure D.2 Trace plots for C-SPSL model with sample size of 200. Notice c is a global parameter for all j. 90

Figure D.3 Trace plots of $\tau_{1,j}$ for gene pairs with the largest ten $|\widehat{\tau}_{1,j}|$ in the colorectal cancer dataset. 92

Chapter 1

Introduction

1.1 Multi-omics Integrative Analysis

Advancements in high-throughput technologies have enabled the generation of large-scale multi-omics data from multiple sources. Increasingly, multi-omics data such as DNA sequences, copy number variations, methylation, miRNA, and gene expression are collected from the same individuals in biomedical studies. The benefits of combining multiple data sources and performing joint analyses with all available genomic information and the phenotypic outcome are multifold. First, different data types could reflect various aspects of the underlying biological system (Song et al., 2020; Kristensen et al., 2014). Second, if multiple data sources all pinpoint the same gene or pathway, then it is less likely to be a false positive. Third, combining data from various sources can lead to a better statistical performance in detecting signals among the noise.

1.2 Missing Values and Data Imputation Algorithms

In integrative multi-omics data analysis, mRNA gene expression often serves as the intermediate variable in many underlying etiological mechanisms. Due to the fact that mRNA measurements often require invasive tissue sampling from participants, it is common to obtain a large portion of the data with missing values in mRNA gene expression measurements as shown in Figure 2.1. A straightforward approach for handling missing values is to implement a complete case analysis by removing

observations with incomplete information (Guillermo et al., 2021; Ramaswami et al., 2020; de Silva and Perera, 2017). Another solution could be to apply imputation methods (Lin et al., 2016; Rubin, 2004; Van Buuren and Groothuis-Oudshoorn, 2011; Troyanskaya et al., 2001; Shah et al., 2014). Multiple imputation (Rubin, 2004) is a widely used solution to the missing value problem. Van Buuren and Groothuis-Oudshoorn (2011) developed a useful tool, multivariate imputation by chained equations (MICE), for implementing multiple imputation to iteratively generate missing values from conditional distributions on the basis of the observed data while considering the relationships between variables.

Although imputation methods offer practical alternatives for handling missing values, in the situation where there is a high rate of missing values, imputation approaches might not perform well (Yang et al., 2014; Yu et al., 2020). Multiple imputation algorithms such as MICE can quickly become computationally intensive as the number of variables with missing values increases (Ratolojanahary et al., 2019). Furthermore, imputation methods mainly use information from single omics data rather than considering the connections among multi-omics data, which can lead to biases in the final imputation (Lin et al., 2016).

1.3 Dynamic Gene Co-expression

High-throughput technologies have generated a wealth of gene expression data and provided exciting opportunities to study genetic regulations and interactions in biological systems. Genetic interactions, often tightly regulated and highly dynamic in a complex regulatory network (Green et al., 2009; Sambandan et al., 2006; Qiu et al., 2011; Yu, 2018), can be activated or switched off in response to changes in cellular signals or environmental conditions. Changes in cellular states and modulatory patterns can result in changes in correlations between genes (Li, 2002; Lai et al., 2004).

In the literature (Li, 2002; Gunderson and Ho, 2014; Kinzy et al., 2019; Yang and Ho, 2021), the dynamic change in gene correlation was referred to as dynamic gene co-expression (DC). It was defined as the conditional correlation between the expression levels of two genes X_1 and X_2 given a modulating factor Z, $\rho(X_1, X_2 \mid Z = z) = E(X_1X_2 \mid Z = z)$, when X_1 and X_2 have zero means and unit standard deviations.

Previous studies have established several statistical methods for quantifying DC (Li, 2002; Lai et al., 2004; Ho et al., 2011; Kinzy et al., 2019). Li (2002) proposed a direct approach named liquid association (LA) to measure the dynamic pattern of co-expression between two genes modulated by the expression level of a third gene. He defined a LA score to describe the expected changing rate of co-expression towards the modulatory variable Z as follows, $LA(X_1, X_2 \mid Z) = E(g^{'}(Z))$, where $g(Z) = \rho(X_1, X_2 \mid Z)$. By assuming the gene expression levels to be standard normal, the conditional correlation can be written as the expected value of the product of gene expression levels, $\rho(X_1, X_2 \mid Z) = E(X_1X_2 \mid Z)$. According to the Stein's lemma, if Z follows the standard normal distribution, the magnitude of DC can be quantified based on the three-product moment estimator: $\frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2}z_i$ (n is sample size).

However, the assumption that the modulating factor (Z) follows a standard normal distribution might be violated when considering to other types of modulators, such as patients survival time or cancer metastatic potency. To relax the normality assumption, Ho et al. (2011) extended Li's work and proposed a conditional normal model (CNM) to quantify DC. In relevant pathways or whole-genome datasets, this approach can be implemented by searching through all possible gene pair combinations (Gunderson and Ho, 2014; Kinzy et al., 2019). However, when the number of genes under consideration increases, this exhaustive search (ES) strategy quickly become computationally intractable. In a typical genomic dataset, often a much smaller proportion of gene interactions exhibit dynamic co-expression changes. This

issue motivated us to focus on the promising gene set combinations during the search process.

1.4 Bayesian variable selection: Spike-and-slab priors

Li's work (Li, 2002) established a concise method for quantifying DC using gene expression modulators. However, the normality assumption cannot be fulfilled when it comes to other types of modulating factors, such as RNA-seq data or cancer metastatic potency. Thus developing methods with more flexibility on the distribution of modulating factors can be beneficial in the discovery of underlying regulatory mechanisms.

On the other hand, DC signals between gene pairs with respect to a modulating factor in a genomic dataset can be highly sparse. A large proportion of gene-pair combinations only exhibit weak DC signals. The original LA and similar approaches generally implement a genome-scale scan over all possible gene-pair combinations (Li, 2002; Li et al., 2004; Lai et al., 2004; Ho et al., 2011; Kinzy et al., 2019; Yang and Ho, 2021). As the dimensionality of genome-wide gene-pair combinations approaches billions ($10^8 \sim 10^{10}$), this scanning strategy quickly becomes computationally intractable. Applying screening measures (Gunderson and Ho, 2014; Yu, 2018) was a frequently used remedy for the issue of dimensionality. However, even the screening techniques can effectively reduce the search space to a smaller subset containing a larger proportion of gene pairs with significant DC signals, sparsity remains an issue in the subset for DC estimation.

Motivated by these two issues, we proposed a novel DC model with the implementation of the variable selection technique in a Bayesian framework. First, we assumed a linear relationship between $\rho(X_1, X_2|Z)$ and Z with a link function of Fisher's Z-transformation as suggested by Ho et al., (2011), Kinzy et al., (2019) and Yang et al., (2021) (Ho et al., 2011; Kinzy et al., 2019; Yang and Ho, 2021). The modulating

factor Z is the predictor without the normality assumption so that a broader range of factors can be considered in the proposed model. Second, the variable selection technique was used for systematically studying associations within a sparse matrix by shrinking insignificant signals towards zero and yielding a parsimonious model. Third, by constructing a multivariate variable selection model, we could consider the dependencies among the gene-pair combinations and improve the effectiveness of DC estimation.

It is noted that solutions for variable selection are available in both frequentist and Bayesian frameworks. Adopting a variable selection method in a Bayesian perspective offers several advantages (Van Erp et al., 2019). First, it is available to incorporate the prior knowledge in the estimation process and derive the posterior distribution of the parameters of interest. Second, Bayesian approaches estimate the shrinking parameter that controls the shrinkage effect using a prior mechanism. The Markov Chain Monte Carlo (MCMC) sampling technique provides greater computational flexibility for parameter estimation. Third, the Bayesian framework facilitates a more natural and intuitive interpretation of parameters. For instance, a shrinkage parameter that is tightly concentrated at 0 can be interpreted as that the corresponding coefficient is dropped out.

We incorporate the spike-and-slab priors in the Bayesian variable selection framework to identify sparse signals among a large set of possible gene combinations. Compared to ES, the searching process can be drastically improved by identifying smaller subsets of promising gene combinations with significant nonzero effects while shrinking other small coefficients to zero (George and McCulloch, 1993).

The spike-and-slab prior, first proposed by Mitchell and Beauchamp (1988), is a scaling mixture of two distributions, one of which concentrates at 0 and the other of which disperses. George and McCulloch (1993) defined the discrete spike-and-slab prior, which is a frequently used prior in the Bayesian variable selection framework.

Let $\beta = \{\beta_j : j = 1, ..., q\}$ be a vector of parameters of interest in a Bayesian model. The general form of the discrete spike-and-slab priors is represented by a normal mixture,

$$\beta_j \mid s_j \overset{ind}{\sim} s_j N(0, v^2) + (1 - s_j)N(0, cv^2),$$
$$s_j \overset{ind}{\sim} Bernoulli(p_j),$$

where v^2 is a small value and c is a large value to generate a large variance (e.g. $v^2 = 0.01$ and c=10,000 so that $cv^2 = 100$). The variable s_j determines whether the "spike" part with a smaller variance or the "slab" part with a larger variance is activated. The probability of the selection is given by $P(s_j = 1) = 1 - P(s_j = 0) = p_j$. When p_j is close to 0, the parameter β_j will shrink to 0 with high probability.

The point mass measure at zero was widely used on the "spike" part of the discrete spike-and-slab priors (Lee et al., 2017); however, the computation was challenging due to the discrete probability measures (Shin and Liu, 2021). To address this problem, Ishwaran et al. (2005) developed continuous spike-and-slab priors by introducing continuous bimodal priors with the following hierarchical structure:

$$\beta_j \mid \gamma_j, v_j^2 \overset{ind}{\sim} N(0, \gamma_j v_j^2),$$
$$v_j^2 \mid \alpha_1, \alpha_2 \overset{i.i.d.}{\sim} \mathcal{IG}(\alpha_1, \alpha_2),$$
$$\gamma_j \overset{ind}{\sim} w_j \mathcal{I}_1 + (1 - w_j)\mathcal{I}_{v_0},$$
$$w_j \overset{i.i.d.}{\sim} U(0, 1),$$

where α_1 and α_2 are the shape and scale parameters of the inverse gamma distribution, and $\mathcal{I}_{v_0}$ refers to the point mass measure concentrating at a small value v_0 (e.g. 0.005).

The hypervariance of the normal prior is the product of two parameters, γ_j and v_j^2. The scaling parameter γ_j controls the shrinkage effect while v_j^2 describes the variance (Polson and Scott, 2010). With different γ_j and v_j^2 values, various bimodal inverse gamma distributions can be constructed, which is either centered near 0 or spreading out. On the basis of the continuous spike-and-slab prior (Ishwaran et al.,

2005), we developed our variable selection method to systematically study DC with sparse signals.

1.5 Structure of the dissertation

The rest of the dissertation is organized as follows. In Chapter 2, integrative multi-omics analysis models are proposed for incorporating incomplete information in intermediate variables using weighted p-value adjustment approaches. We describe the proposed weighted p-value mechanisms in Section 2.3. In Section 2.4, a series of simulation studies are presented in both low-dimensional cases and high-dimensional cases to demonstrate the advantages of our proposed approach compared to imputation algorithms in various aspects. To illustrate the implementation of the proposed approaches in real cases, we apply them to jointly analyze DNA methylation, gene expression, and phenotypic outcome in a preterm infant birth weights study in Section 2.5. Finally, the conclusion and future studies are discussed in Section 2.6.

In Chapter 3, we propose two models to systematically identify dynamic gene-coexpression in whole-genome datasets using Bayeisan variable selection techniques with spike-and-slab priors. Section 3.1 reviews the main issues and the related literature. Section 3.2 provides an overview of the dataset. Section 3.3 introduces the proposed approaches and describes the estimation processes. Section 3.4 presents two simulation studies in both low- and high-dimesional search spaces with different degrees of sparsity, which illustrates the advantages of implementing the proposed models compared to the existing exhaustive search strategy. Section 3.5 demonstrates the implementation of our proposed approaches with the experimental data analysis on a colorectal cancer (CRC) dataset. In Section 3.6, we provide some discussions and suggestions for future research on the related topics.

In Chapter 4, we proposed a mixed-effect variable selection model to study the subject-specific DCs for zero-inflated and over-dispersed scRNA-seq count data to

account for both the random effect of each subject and the fixed effect across all subjects on DC. A DC-based classification strategy is proposed to implement subgroup classification using identified subject-specific DC gene pairs as the biomarkers. Section 4.1 provides an overview of the background and the idea of our proposed method. Section 4.2 elaborates the proposed methods and introduced the melanoma dataset used in this chapter. Section 4.3 presents three simulation studies to illustrate the performance of the proposed methods in subject-specific DC identification and subgroup classification. In Section 4.4, we illustrate the implementation of our proposed methods through an experimental data analysis using the melanoma dataset to estimate the subject-specific DC and implement the immunotheropy-resistant sample classification. At the end, Section 4.5 provides discussion for the proposed methods and the future work in the related area.

Chapter 2

Multi-omics integrative analysis for incomplete data

2.1 Introduction

The advancements in high-throughput technologies provide exciting opportunities to obtain multi-omics data from the same individual in a biomedical study. Joint analyses of data from multiple sources offers many benefits. However, the occurrence of missing values is an inevitable issue in multi-omics data because some measurements, such as mRNA gene expression levels often require invasive tissue sampling from patients. Common approaches for addressing missing measurements include analyses based on observations with complete data or multiple imputation methods.

To solve this issue, we propose a novel integrative analytical framework using weighted p-value adjustment approaches to incorporate both the complete and incomplete (with missing mRNA gene expression measurements) observations in multi-omics analyses. The weighted p-value adjustment approaches were proposed in the context of multiple hypothesis testing to incorporate external information or prior knowledge while maintaining the type I error rate (Roeder and Wasserman, 2009). Several weighting procedures have been proposed in the literature, such as weighted Bonferroni method for family-wise error rate (FWER) control (Roeder and Wasserman, 2009; Li et al., 2013), weighted Benjamini-Hochberg (BH) method (Genovese et al., 2006; Habiger, 2017) and q-value method (Storey and Tibshirani, 2003; Storey et al., 2004) for false discovery rate (FDR) control, and grouped FDR methods (Igna-

tiadis and Huber, 2021; Roquain and Van De Wiel, 2009). To ensure the independence between p-values and the derived weights (Roeder and Wasserman, 2009), the sample splitting strategy (Rubin et al., 2006; Roeder et al., 2007) provides a useful tool that uses a subset of the data to generate weights and the remaining data to compute p-values.

In our proposed approaches, we split the samples into a complete set with full information and an incomplete set with missing mRNA gene expression measurements. Two weighted p-value mechanisms (general and reverse weighting schemes) are proposed. Compared to integrative procedures that utilize Markov chain Monte Carlo such as iBAG (Wang et al., 2013), Bayesian integrative model (Fridley et al., 2012), multi-dataset integration (Kirk et al., 2012), and Bayesian consensus clustering (Lock and Dunson, 2013), our proposed approach is fast and computationally simple for a whole-genome study. Computational efficiency is particularly critical for integrating multi-omics data since the interactions between multiple data types grow exponentially with the number of variables considered in the study.

This chapter is structured as follows. The experiemental datasets used in this chapter is described in Section 2.2. Next, the details of proposed weighted p-value mechanisms are presented in Section 2.3. In Section 2.4, we implement three simulation studies in both low- and high-dimensional cases to demonstrate the advantages of our proposed approach compared to the existing imputation algorithms and complete-case studies. To illustrate the use of our proposed approaches, we apply the proposed models to jointly analyze DNA methylation, gene expression, and phenotypic outcome in a preterm infant birth weights study in Section 2.5. Finally, we conclude this chapter and discuss the future areas of research in Section 2.6.

2.2 Datasets and Databases

The dataset utilized in this chapter is from a genetic association study for preterm infants by Kashima et al. (2021) and can be accessed at Gene Expression Omnibus (GEO) with accession number GSE110828. This study contains 157 observations with DNA methylation and phenotypic outcome information. However, mRNA gene expression measurements were collected for only 55 observations (65% missing). DNA methylation levels were measured using the Ilumina HumanMethylation450 BeadChip for 410,735 cytosine-phosphate-guanine (CpG) sites and reported after quantile normalization and background correction. The mRNA gene expression levels of 46,789 transcripts were profiled using the SurePrint G3 Human GE microarray 8×60K version 3.0 (Agilent Technologies). Transcriptional activities were analyzed using GeneSpring14.5 to perform probe-filtering and quantile-normalization to report the gene expression signal levels.

2.3 Models

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ be the vector of phenotypic outcome with n representing the total number of subjects, $\boldsymbol{X}$ be the matrix of clinical covariates, and $\boldsymbol{M} = (\boldsymbol{M}_1, ..., \boldsymbol{M}_q)$ be the matrix of DNA methylation levels of q CpG sites, where $\boldsymbol{M}_j = (M_{1j}, ..., M_{nj})^T$, $j = 1, ..., q$, is the vector of methylation levels for the j-th CpG site. Let $\boldsymbol{G} = (\boldsymbol{G}_1, ..., \boldsymbol{G}_d)$ be the matrix of standardized mRNA gene expression data (mean = 0 and standard deviation = 1) of d genes and $\boldsymbol{G}_l = (G_{1l}, ..., G_{n_1 l})^T$ be the vector of expression levels for the l-th gene ($l = 1, ..., d$) with n_1 representing the number of subjects of gene expression data ($n_1 \leq n$). All subjects can be split into two subsets: a complete set ($Z^{(1)} = (\boldsymbol{M}^{(1)}, \boldsymbol{Y}^{(1)}, \boldsymbol{X}^{(1)}, \boldsymbol{G})$) with n_1 subjects where mRNA expression data can be observed and an incomplete set ($Z^{(2)} = (\boldsymbol{M}^{(2)}, \boldsymbol{Y}^{(2)}, \boldsymbol{X}^{(2)})$) with n_2 subjects where the mRNA gene expression data are completely missing. The total

number of subjects is $n = n_1 + n_2$. Figure 2.1 provides a schematic diagram for this data structure.

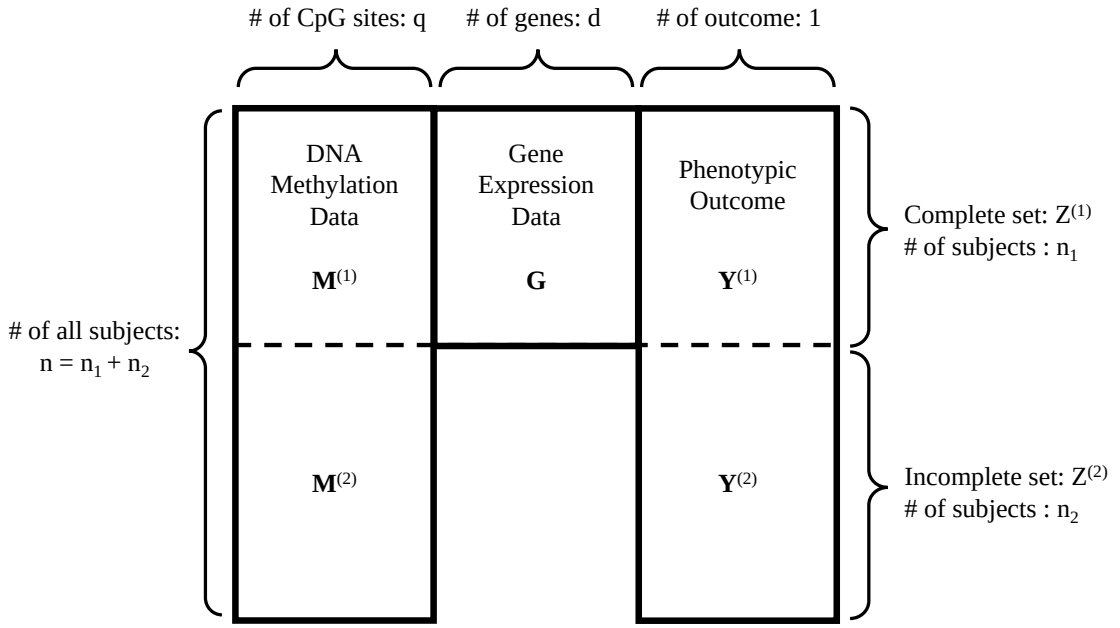


Figure 2.1. The schematic diagram for partitions of the complete set and incomplete set in datasets with missing intermediate mRNA measurements. Dimensions, sample sizes, the complete set, and the incomplete set are shown in the plot.

2.3.1 General weight

In the complete set, $Z^{(1)}$, we implement the integrative analytical framework (IG) suggested by Zhao et al. (2014) to integrate the DNA methylation, mRNA gene expression data, and the phenotypic outcome to derive the p-value (p_j^{IG}) for testing the association between the j-th DNA methylation measurement ($j = 1, ..., q$) and the phenotypic outcome. Briefly, p_j^{IG} is calculated via two linear models formulated as follows:

$$E(Y_i^{(1)}|\boldsymbol{G}_i, \boldsymbol{X}_i^{(1)}) = \alpha_0 + \boldsymbol{G}_i^T\boldsymbol{\alpha}_{\boldsymbol{G}} + (\boldsymbol{X}_i^{(1)})^T\boldsymbol{\alpha}_{\boldsymbol{X}} \tag{2.1}$$

$$\boldsymbol{G}_i^T\boldsymbol{\alpha}_{\boldsymbol{G}} = \beta_{0j} + \beta_{M_j}M_{ij}^{(1)} + (\boldsymbol{X}_i^{(1)})^T\boldsymbol{\beta}_{\boldsymbol{X}} + u_{ij}, \tag{2.2}$$

where α_0 and β_{0j} are intercepts; $\boldsymbol{\alpha_G}$ and $\boldsymbol{\alpha_X}$ are coefficients describing the associations between mRNA gene expression, clinical covariates, and the outcome. The parameter of interest, β_{M_j}, measures the association between the j-th DNA methylation level and the phenotypic outcome via the regulation of mRNAs; and $u_{ij} \sim \mathcal{N}(0, \sigma_u^2)$ is the error term ($i = 1, ..., n_1$) with variance σ_u^2. Let $\widehat{\boldsymbol{\alpha_G}}$ and $\widehat{\beta_{M_j}}$ be the estimates of α_G and $\beta_{M_j} = 0$. In practice, $\widehat{\beta_{M_j}}$ can be estimated via Equation (2) using $\widehat{\boldsymbol{\alpha_G}}$ derived from Equation (1). Under the null hypothesis of no association between M_j and Y ($\beta_{M_j} = 0$), the p-value (p_j^{IG}) can be calculated based on $\widehat{\beta_{M_j}}$ and var($\widehat{\beta_{M_j}}$).

In the incomplete set, $Z^{(2)}$, we implement the linear model as follows:

$$Y_i^{(2)} = \gamma_{0j} + \gamma_{M_j} M_{ij}^{(2)} + (\boldsymbol{X}_i^{(2)})^T \boldsymbol{\gamma_X} + \epsilon_{ij}, \tag{2.3}$$

where $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$, $i = 1, ..., n_2$ and $j = 1, ..., q$, is the error term with variance σ_ϵ^2; γ_{0j} is the intercept and γ_{M_j} represents the association between j-th methylation data measurement and the phenotypic outcome; and $\boldsymbol{\gamma_X}$ is the vector of coefficients for the covariates. Let $\widehat{\gamma_{M_j}}$ be the estimate of γ_{M_j}. The p-value (p_j^{LM}) can be derived based on $\widehat{\gamma_{M_j}}$ and var($\widehat{\gamma_{M_j}}$) under the null hypothesis $\gamma_{M_j} = 0$.

In the general weighting scheme, the p-value derived from the incomplete set (p_j^{LM}) is used to generate the weight. According to Li et al. (2013), we choose $w_{G_j} = \sqrt{-log_{10}(p_j^{LM})}$ when $p_j^{LM} < 0.05$, and set $w_{G_j} = 1$ otherwise, as the functional form to incorporate the information of the association between M_j and Y into the weights. Since smaller p-values are associated with null hypotheses that are more likely false, the proposed weights are anticipated to be positively correlated with optimal weights and perform well (Habiger, 2017). To control the type I error, the general weights are then divided by the average weight $w_{G_j}^* = \frac{w_{G_j}}{\overline{w_G}}$ ($\overline{w_G} = \frac{1}{q}\sum_{j=1}^{q} w_{G_j}$) to ensure $\overline{w_G^*} = 1$ (Genovese et al., 2006; Wasserman and Roeder, 2006). Finally, the adjusted p-values for general weighting scheme can be calculated as $p_{1j} = \frac{p_j^{IG}}{w_{G_j}^*}$.

2.3.2 Reverse weight

The general weighting scheme is more effective when the missing rate of gene expression data is low. When the missing rate is high (i.e. $> 50\%$), we propose a reverse weighting scheme to increase the power of identifying significant CpG sites. This approach to deriving weights by a reverse weighting scheme is similar to the general weighting scheme but uses a complete set ($Z^{(1)}$) to obtain weights while deriving p-values using the incomplete set ($Z^{(2)}$).

In the incomplete set, $Z^{(2)}$, the p-value (p_j^{LM}) is derived based on $\widehat{\gamma_{M_j}}$ and var($\widehat{\gamma_{M_j}}$) from the linear model as described in Equation (2.3). Then, the weights are calculated in terms of p_j^{IG} obtained from the complete set ($Z^{(1)}$) by implementing the IG model. The reverse weight is set to be $w_{R_j} = \sqrt{-log_{10}(p_j^{IG})}$ when $p_j^{IG} < 0.05$ and $w_{R_j} = 1$ otherwise. Then the weights are adjusted by the average value as follows, $w_{R_j}^* = \frac{w_{R_j}}{\overline{w_R}}$ ($\overline{w_R} = \frac{1}{q}\sum_{j=1}^{q} w_{R_j}$) to ensure $\overline{w_R^*} = 1$ (Genovese et al., 2006; Wasserman and Roeder, 2006). Finally, we derive the p-value adjusted by the corresponding reverse weight, $p_{2j} = \frac{p_j^{LM}}{w_{R_j}^*}$.

The null hypotheses are $\beta_{M_j} = 0$ and $\gamma_{M_j} = 0$ corresponding to p_j^{IG} and p_j^{LM}, respectively, in the general and reverse weighting scheme. According to Zhao et al. (2014) when Equation (2.1) and (2) hold, we can plug Equation (2) into Equation (2.1), hence the null hypothesis $\beta_{M_j} = 0$ is equivalent to $\gamma_{M_j} = 0$. Next, we present an omnibus approach to combine the weight adjusted p-values (p_{1j}, p_{2j}) from the two weighting schemes.

2.3.3 Omnibus Method

For this study, we also consider an omnibus approach, the Aggregated Cauchy Association Test (ACAT) (Liu et al., 2019; Liu and Xie, 2020), to combine the adjusted p-values from the general weighting scheme and the reverse weighting scheme. The ACAT calculates the test statistic via a weighted sum of Cauchy transformations of

the component p-values and assumes the test statistic to follow a Cauchy distribution under the null hypothesis. The test statistic for the j-th CpG site ($j = 1, ..., q$) is given by,

$$T_j^{ACAT} = (1 - \lambda) \times tan\{(0.5 - p_{1j})\pi\} + \lambda \times tan\{(0.5 - p_{2j})\pi\}, \tag{2.4}$$

where two component models (p_{1j} for general weighting scheme and p_{2j} for the reverse weighting scheme) are combined in the ACAT and λ is determined by the missing rate. The general scheme is more powerful in a low missing rate dataset, while the reverse scheme becomes more effective when the missing rate is greater than 50%. To maintain a desirable performance under various missing rates, we chose the λ in terms of the missing rates of gene expression data ($0 \leq \lambda \leq 1$). Thus, we could emphasize the general weighting scheme in studies with low missing rate and prioritize the reverse weighting scheme in high-missing-rate studies.

2.4 Simulation

We conducted simulation studies to compare the performance of the proposed weighting approaches to the IG method (Zhao et al., 2014), the popular MICE imputation (Van Buuren and Groothuis-Oudshoorn, 2011) and the K-nearest neighbor (KNN) imputation method (Batista et al., 2002) under three scenarios. In this section, we use the notation γ_{MG} to describe the DNA-gene association between DNA methylation (M) and gene expression (G), and γ_{GY} to denote the gene-phenotype association between gene expression (G) and the phenotypic outcome (Y). Since there were 157 observed subjects in the experimental data set, we generated $n = 150$ samples in all scenarios and studied the power of models averaged over 1,000 simulations.

The following steps describe the data generation procedures for Scenarios I and II with low-dimensional gene expression data. For the i-th subject, we first generated data for $q = 5$ DNA methylation loci ($\boldsymbol{M}_i$) and $r = 2$ clinical covariates ($\boldsymbol{X}_i$) indepen-

dently from standard normal distributions. A single CpG site (M_{i1}) was selected to be the true underlying methylated CpG site associated with the phenotypic outcome through the regulation of the mRNA gene expression of three modulating genes.

In the second step, we considered $d = 8$ genes with expression levels ($\boldsymbol{G}_i$), of which three genes were simulated based on the underlying CpG site (M_{i1}) via the linear model,

$$\boldsymbol{G}_i = \boldsymbol{\gamma}_{0G} + M_{i1}\boldsymbol{\gamma}_{MG} + \boldsymbol{X}_i^T\boldsymbol{\gamma}_{XG} + \boldsymbol{\epsilon}_{i1}, \tag{2.5}$$

where $\boldsymbol{\gamma}_{MG}$ is the vector of DNA-gene association; and $\boldsymbol{\epsilon}_{i1} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{I}$ is an identity matrix. The values of the intercept ($\boldsymbol{\gamma}_{0G}$) were determined based on the mean expression levels of randomly selected genes from the experimental dataset. The values of the elements in the coefficient vectors ($\boldsymbol{\gamma}_{XG}$) were all set equal to 0.5. The five other independent genes served as unrelated signals and were generated independently from the normal distribution, $\mathcal{N}(\mu_{0G}, 1)$, where μ_{0G} was also determined by the mean expression level of a randomly picked gene from the experimental dataset.

In the third step, we simulated the phenotypic outcome (Y_i) based on the mRNA expression levels of the three modulating genes, according to the second linear model,

$$Y_i = \gamma_{0Y} + \boldsymbol{G}_i^T\boldsymbol{\gamma}_{GY} + \boldsymbol{X}_i^T\boldsymbol{\gamma}_{XY} + \epsilon_{i2}, \tag{2.6}$$

where $\epsilon_{i2} \sim \mathcal{N}(0, 1)$ is the error term with the variance of the phenotypic outcome being set equal to 1. Here the intercept (γ_{0Y}) was set equal to the mean birth weight score of the preterm infants in the experimental dataset, and the associations between the clinical covariates and the phenotypic outcome ($\boldsymbol{\gamma}_{XY}$) were all set equal to 0.5. For simplicity, we also assumed the same values for all the elements in the vector $\boldsymbol{\gamma}_{GY}$. After generating $n = 150$ subjects, which is close to the sample size of the experimental dataset, multiple records of gene expression levels were removed completely at random.

2.4.1 Scenario I

After obtaining the weight-adjusted p-values via the proposed weighting schemes, we considered two multiple testing adjustment procedures, one is the weighted Bonferroni method (Bland and Altman, 1995) for maintaining FWER and the other is the weighted BH method (Benjamini and Hochberg, 1995) for FDR control. We set both γ_{MG} and γ_{GY} equal to 0 and reported the results for both the FWER and the FDR at the nominal level of 0.05 (Table 2.1). The FWER was calculated as the proportion of times that at least one significant CpG site was observed among all CpG sites. The FDR was calculated as the ratio of falsely detected CpG sites after the BH procedure.

2.4.2 Scenario II

In this scenario, we set the missing rate of gene expression data to 20% (low), 50% (medium), and 70% (high) to assess the power of the proposed methods for identifying the underlying CpG site (M_1) with different amounts of missing gene expression data. To comprehensively present the performance of the methods under various combinations of DNA-gene and gene-phenotype associations, γ_{MG} was set equal to 0, 0.1, 0.2, and 0.5, and γ_{GY} was set equal to 0.1, 0.2, and 0.5. The empirical power of detecting M_1 over 1,000 simulations was reported.

2.4.3 Scenario III

In this scenario, we considered 1000 mRNA expression measurements and assumed that the underlying DNA methylation (M_{i1}) was associated with the phenotypic outcome for the i-th subject (Y_i) through the regulation of $k = 5$ genetic pathways ($\boldsymbol{f}_i$). The associations between gene expression levels ($\boldsymbol{G}_i$) and the pathway activities ($\boldsymbol{f}_i$) could be estimated using a factor model,

$$\boldsymbol{G}_i = \boldsymbol{B}\boldsymbol{f}_i + \boldsymbol{U}_i, \tag{2.7}$$

where $\boldsymbol{f}_i \in \mathcal{R}^k$ $(k < p)$ is the vector of latent factors with $cov(\boldsymbol{f}_i) = \boldsymbol{I}_k$, $\boldsymbol{U}_i \in \mathcal{R}^p$ is the error term, and $B \in \mathcal{R}^{p \times k}$ is the loading matrix describing the gene-factor associations.

Data simulation was performed in a series of steps. To avoid further confusion, we still used γ_{MG} and γ_{GY} to denote the DNA-gene association and gene-phenotype association. For i-th subject, five factors $(\boldsymbol{f}_i)$ were first simulated based on the underlying CpG site (M_{i1}) from the equation,

$$\boldsymbol{f}_i = M_{i1}\boldsymbol{\gamma}_{MG} + \boldsymbol{X}_i^T\boldsymbol{\gamma}_{XG} + \boldsymbol{\epsilon}_{i1}, \tag{2.8}$$

where $\boldsymbol{\gamma}_{MG}$ is the vector of DNA-gene associations, $\boldsymbol{\epsilon}_{i1} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{I})$ were normal, and the the covariate coefficients $\boldsymbol{\gamma}_{XG}$ were set equal to 0.5. Second, Y_i was generated based on the latent factors from the equation,

$$Y_i = \gamma_{0Y} + \boldsymbol{f}_i^T\boldsymbol{\gamma}_{GY} + \boldsymbol{X}_i^T\boldsymbol{\gamma}_{XY} + \epsilon_{i2}, \tag{2.9}$$

where $\epsilon_{i2} \sim \mathcal{N}(0, 1)$ is the error term, $\boldsymbol{\gamma}_{GY}$ is the vector the gene-phenotype associations. $\boldsymbol{\gamma}_{XY}$ and γ_{0Y} were set to the same values as the low-dimensional cases.

In the last step, the gene expression data $(\boldsymbol{G}_i)$ were generated based on the factors $(\boldsymbol{f}_i)$, the error term $(\boldsymbol{U}_i \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{I}))$, and the loading matrix $(\boldsymbol{B})$ following Baek et al. (2020). We formed $\boldsymbol{B} = \frac{1}{\sqrt{n}}\boldsymbol{L}^T\boldsymbol{E}$ where $\boldsymbol{L} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{E}$ is an $n \times k$ orthogonal matrix formulated by the eigenvectors corresponding to the k largest eigenvalues of $\boldsymbol{L}\boldsymbol{L}^T$. The optimal k can be estimated by minimizing the cross-validated mean squared error (MSE), as suggested by Owen and Perry (2009).

After deriving the latent gene factors via factor analysis (Baek et al., 2020), p_j^{IG} was derived according to Equation (2.1) and Equation (2) by using $\boldsymbol{f}_i$ instead of $\boldsymbol{G}_i$ to implement the proposed methods. Next, we reported the empirical power of the underlying CpG site with methylation for γ_{MG} ranging from 0 to 0.5. The gene-phenotype association γ_{GY} was set equal to 0.1, 0.2, and 0.5, and the missing

rate was 70%. Notably, in the high-dimensional case, multiple imputation becomes computationally infeasible.

2.4.4 Competing Methods

Four competing methods were considered in the simulation studies: (1) complete case studies using only the complete set with the integrative analytical framework (Zhao et al., 2014), (2) KNN imputation algorithm (Batista et al., 2002) to estimate the missing values using the mean of the nearest values of k-th closest subjects, (3) multivariate imputation via chained equations (MICE) method to estimate the missing values by combining results derived from multiple imputed datasets, and (4) linear model on all subjects of $\boldsymbol{M}$ and $\boldsymbol{Y}$. In (2) and (3), the IG model was implemented to identify the underlying CpG site after imputing the missing values in the gene expression data.

We implemented a 10 nearest-neighbor imputation method with the *impute* package (Hastie T and G, 2021) and applied the MICE algorithm with the *mice* package (van Buuren et al., 2021) in R. The maximum number of iterations was set equal to 5 in MICE and 5 datasets were generated for pooling results. Due to the intractable computational time in the high-dimensional case, MICE was not implemented in Scenario III.

2.4.5 Simulation Results

Table 2.1 reports the FWER and the FDR of testing the CpG sites with $\gamma_{MG} = \gamma_{GY} = 0$ (no association with the outcome $\boldsymbol{Y}$). The results show that our proposed methods and the existing method maintained FWER and FDR at the nominal 0.05 level.

Figure 2.2 presents the average power of the proposed omnibus method and the competing methods to compare the performance for identifying the underlying CpG

Table 2.1. Table for FWER after weighted Bonferroni method and FDR after weighted BH method when all associations were set equal to 0.

		IG	MICE	KNN Impute	General Weight	Reverse Weight	Omnibus Method	Linear Model
20%	FWER	0.037	0.010	0.045	0.037	0.048	0.047	0.053
	FDR	0.037	0.010	0.052	0.037	0.049	0.053	0.057
50%	FWER	0.063	0.013	0.061	0.061	0.060	0.065	0.053
	FDR	0.062	0.013	0.061	0.062	0.059	0.077	0.057
70%	FWER	0.028	0.026	0.044	0.027	0.045	0.048	0.053
	FDR	0.031	0.027	0.050	0.030	0.040	0.054	0.057

site. In datasets with a high missing rate ($\geq 50\%$), the proposed omnibus method is more powerful than the IG model with a complete case approach and the imputation algorithms. For example, when the missing rate is 70% and $\gamma_{MG} = \gamma_{GY} = 0.2$, the proposed omnibus method achieves the highest power, which is 10.5% higher than the IG model and 52.3% higher than MICE.

Figure 2.3 presents the performance using the general weighting scheme, the reverse weighting scheme, and the omnibus method. Based on our simulation results, the general weighting scheme performes better in datasets with a low missing rate while the reverse weighting scheme performs better with a high missing rate ($> 50\%$) as well as a large γ_{GY} (e.g. $\gamma_{GY} = 0.5$). We set λ in the ACAT test statistic in terms of the missing rate as described in Section 2.3.3. Our results show that the proposed omnibus method demonstrates a more competitive performance compared to the other methods.

The performance of our proposed method in a high-dimensional case is illustrated in Figure 2.4. In this scenario, the missing rate was set equal to 70% to mimic that of the experimental dataset used in this chapter. As shown in the power plots, the omnibus method is more powerful than the IG in all cases except when γ_{MG} and γ_{GY} are extremely weak. The proposed method outperforms the other competing methods when the gene-phenotype signal is moderate $\gamma_{GY} = 0.2$ but is comparable to the linear model when $\gamma_{GY} = 0.5$.

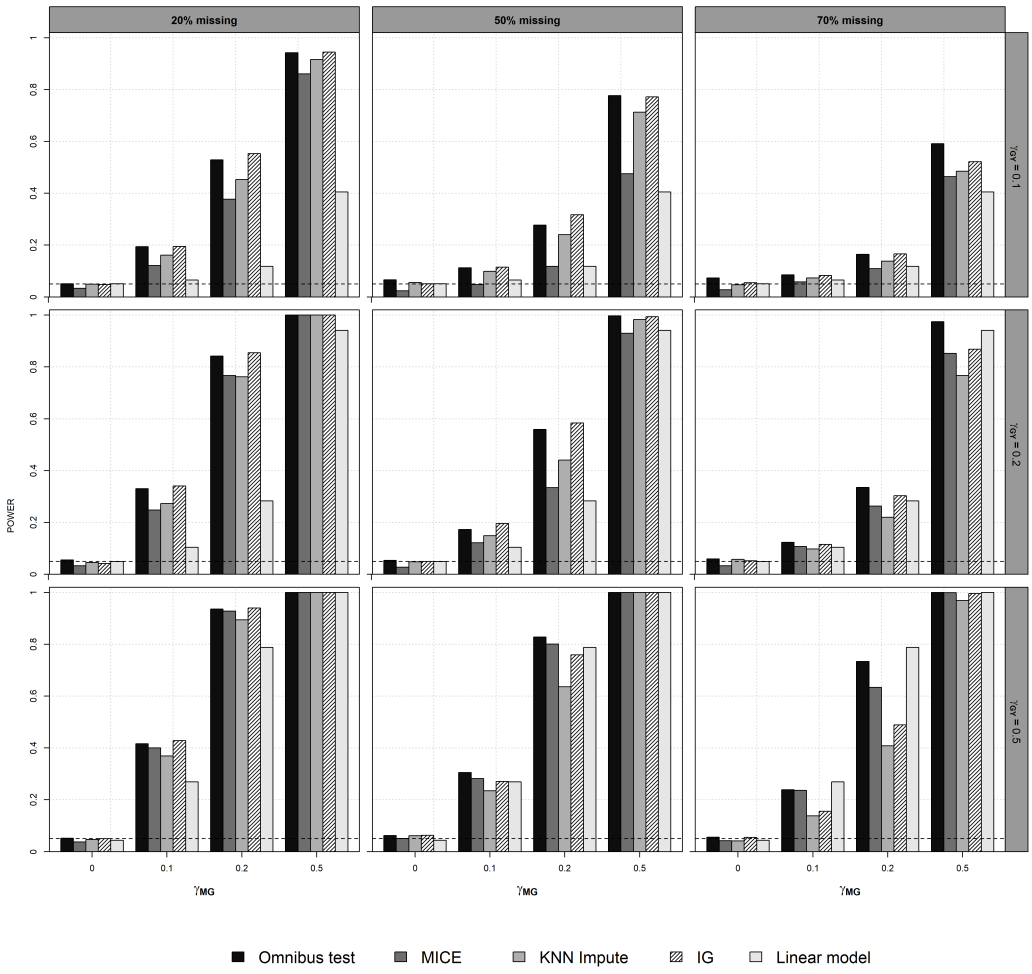


Figure 2.2. Power comparisons of omnibus method, MICE, KNN imputation, IG and linear model for various missing rates and γ_{MG}. The value of γ_{GY} was set equal to 0.1, 0.2 and 0.5. The standard deviations of both γ_{MG} and γ_{GY} were set equal to 1.

In consideration of the case where the missingness is explained by observed variables, we further conducted simulation studies for a missing at random case. Table 2.2 reports the FWER and FDR of testing the CpG sites with $\gamma_{MG} = \gamma_{GY} = 0$ after the weighted bonferroni method and q-value method as described in Section 2.4.1. The results show that all the methods maintain both the FWER and the FDR at the nominal 0.05 level.

Figure 2.5 and Figure 2.6 present the power comparisons of the proposed omnibus

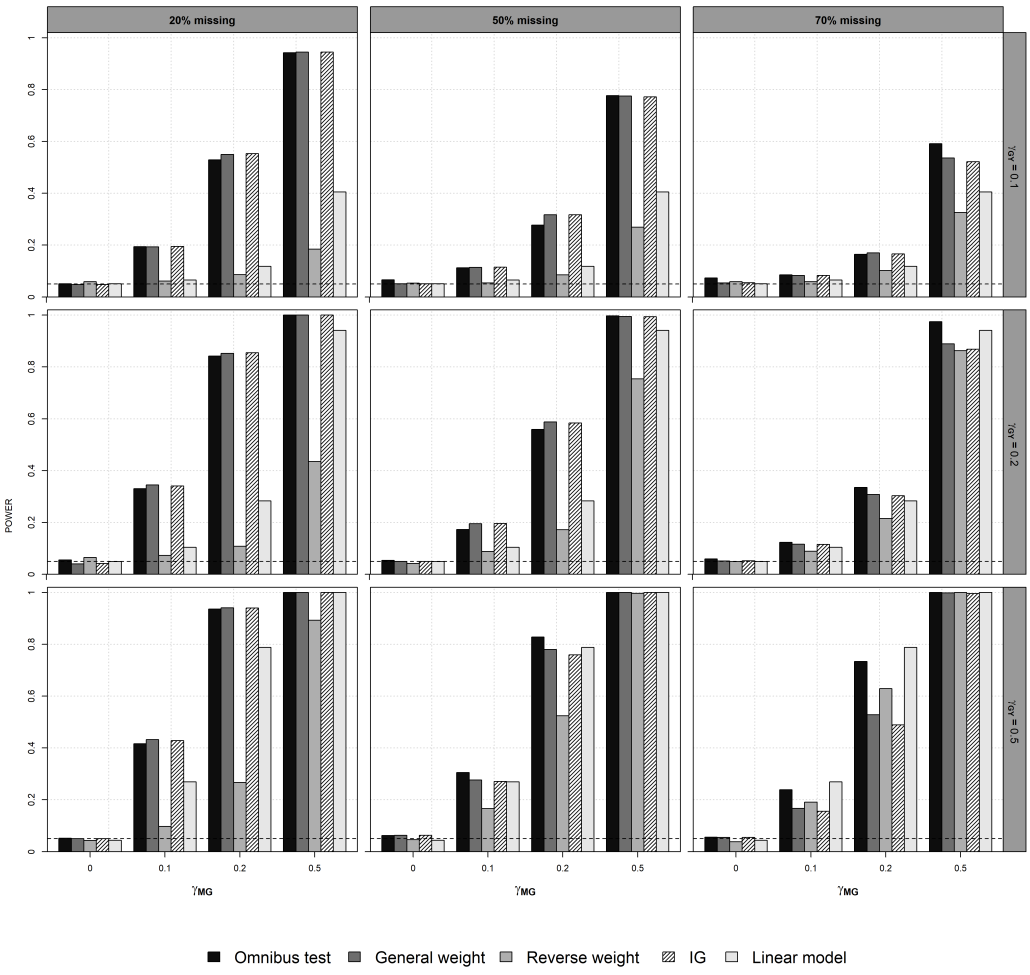


Figure 2.3. Power comparisons of omnibus method, general weight scheme, reverse weight scheme, IG and linear model for various γ_{MG} and γ_{GY} with different missing rates.

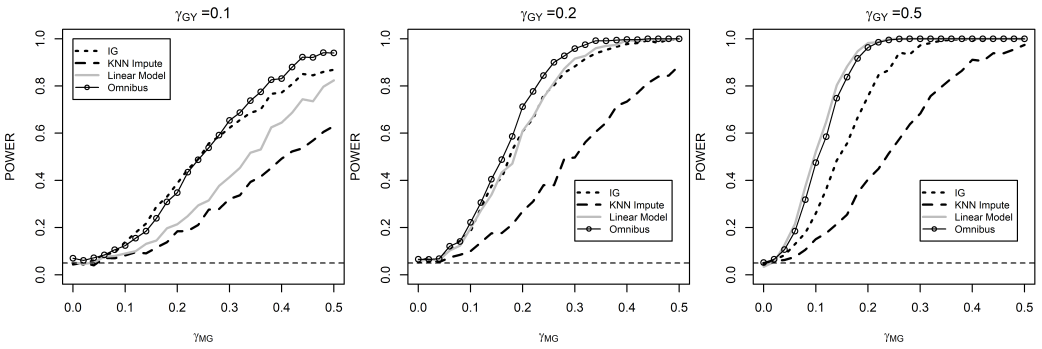


Figure 2.4. Power curves for omnibus method and competing methods in high-dimensional case. The number of factors $k = 5$ was determined by the 10-fold cross-validation. The gene-phenotype association γ_{GY} was set to be 0.1, 0.2 and 0.5.

Table 2.2. Type I error control for FWER using weighted Bonferroni method and FDR using q-value method when $\gamma_{MG} = \gamma_{GY} = 0$ under MAR.

		IG	MICE	KNN Impute	General Weight	Reverse Weight	Omnibus Method	Linear Model
20%	FWER	0.047	0.007	0.051	0.046	0.058	0.054	0.055
	FDR	0.044	0.007	0.058	0.044	0.055	0.064	0.058
50%	FWER	0.051	0.020	0.047	0.051	0.037	0.042	0.050
	FDR	0.053	0.021	0.051	0.054	0.041	0.051	0.052
70%	FWER	0.048	0.022	0.058	0.048	0.055	0.068	0.056
	FDR	0.051	0.024	0.060	0.051	0.052	0.081	0.055

method to existing methods (MICE, KNN, IG, and LM) and the component methods (general and the reverse weighting schemes) under MAR for identifying the underlying CpG site. The results are similar to the MCAR case in Section 2.4.2, which demonstrates that our proposed methods outperform the other methods.

Figure 2.7 presents the performance of our proposed method in the high-dimensional case. As shown in the power plots, the omnibus method with the circled line is more powerful than the IG in all cases except when the signals of γ_{MG} and γ_{GY} are weak.

2.5 Experimental Data Analysis

We implemented our proposed omnibus weighting approach using the preterm infant data described in section 2.2. The infants birth weight scores were used as the phenotypic outcome (Y). The weight scores were calculated by the normal quantile of the birth weights for each gestational age in the entire population of newborn infants so that they are normally distributed, as described in the Japanese reference data (Kashima et al., 2021; Agha et al., 2016; Oken et al., 2003). According to Kashima et al. (2021), the methylation levels ($\boldsymbol{M}$) were measured by β values ranging from 0 (completely unmethylated) to 1 (completely methylated) to indicate the intensity of methylation on each CpG site. Both the methylation levels and the birth weight scores were scaled to have a mean of 0 and a variance of 1.

The clinical covariates considered in this analysis included paternal age, maternal

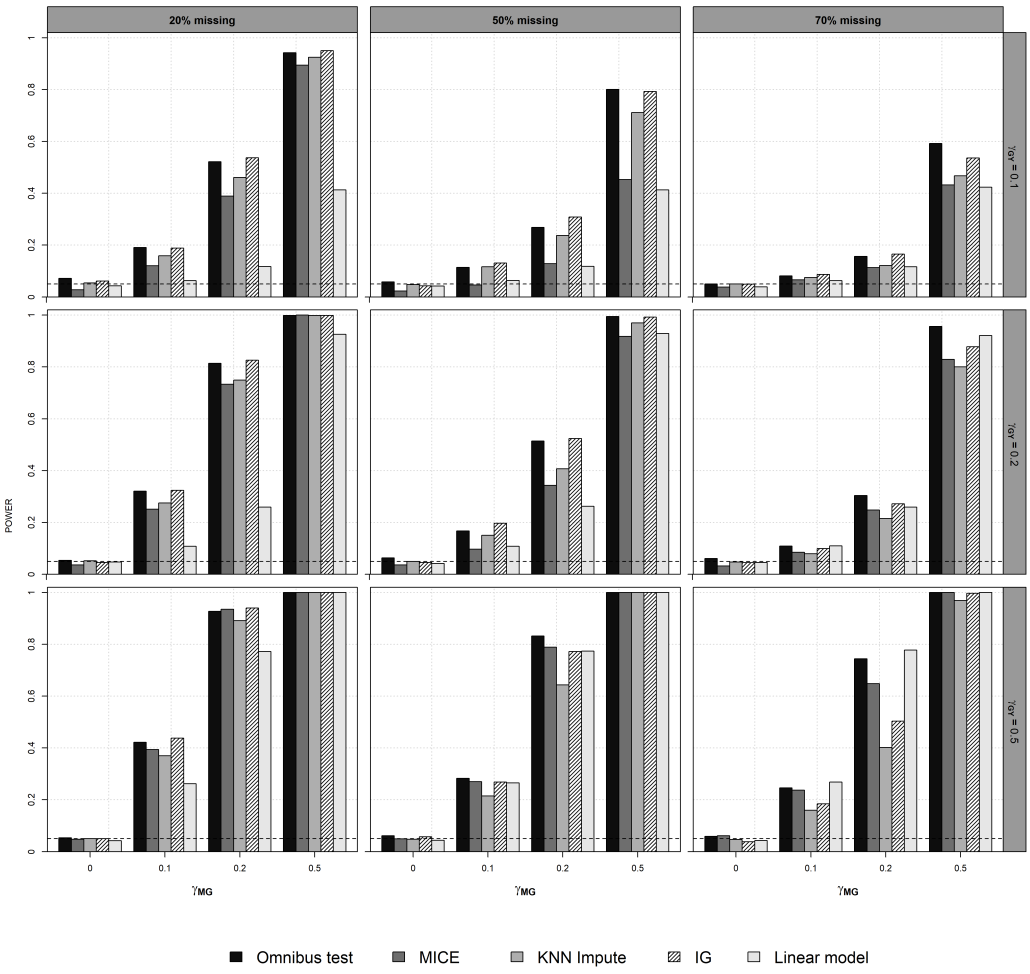


Figure 2.5. Power comparisons of omnibus method, MICE, KNN imputation, IG and linear model for various missing rates and γ_{MG} under MAR. The value of γ_{GY} was set equal to 0.1, 0.2 and 0.5. The standard deviations of both γ_{MG} and γ_{GY} were set equal to 1.

age, paternal body mass index (BMI), maternal BMI, maternal smoking status before pregnancy, and the gender of the infants. To correct for population stratification, we implemented surrogate variable analysis (SVA) (Leek and Storey, 2007) to account for the unobserved effect from the clinical covariates. The genomic inflation factor (van Iterson et al., 2017) was used after SVA for adjusting the inflated p-values due to population stratification. Figure 2.8 presents the QQ-plot for omnibus method and suggests proper type I error control.

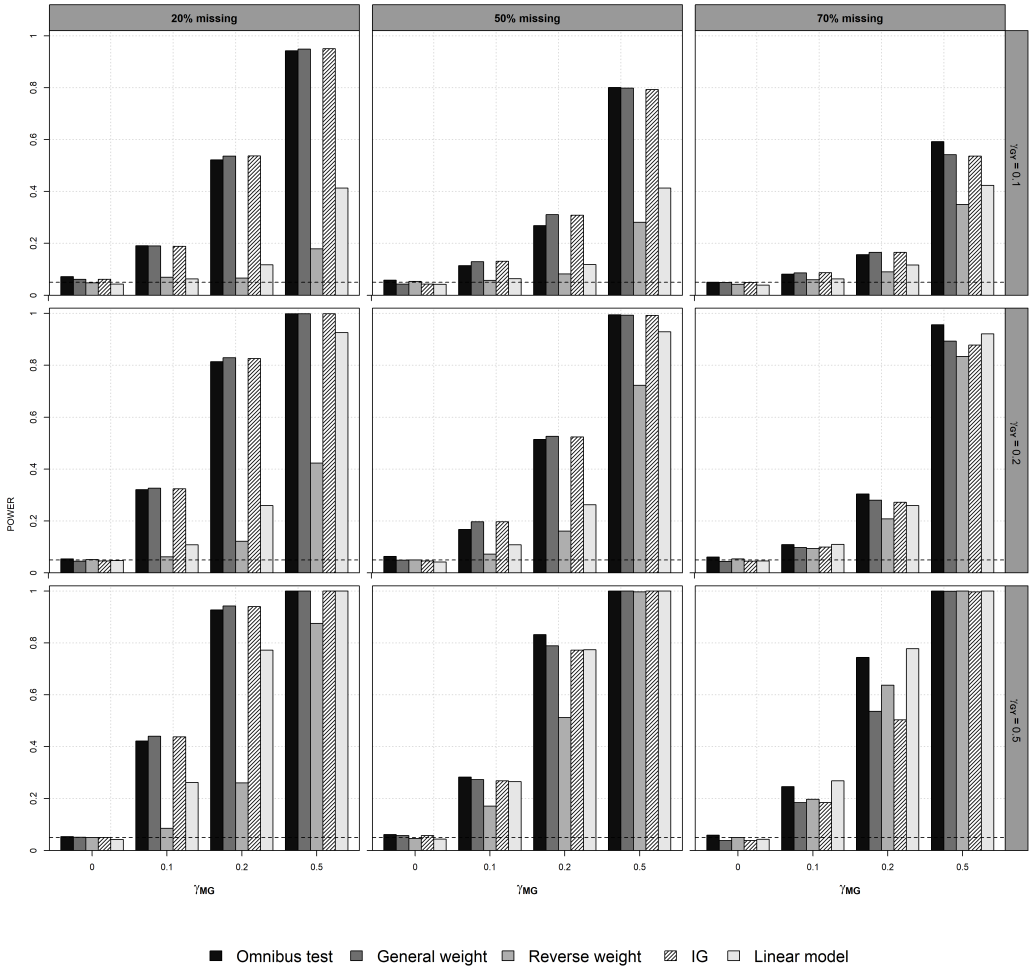


Figure 2.6. Power comparisons of omnibus method, general weighting scheme, reverse weighting scheme, IG and linear model for various γ_{MG} and γ_{GY} with different missing rates under MAR. The standard deviations of γ_{MG} and γ_{GY} were both set equal to 1.

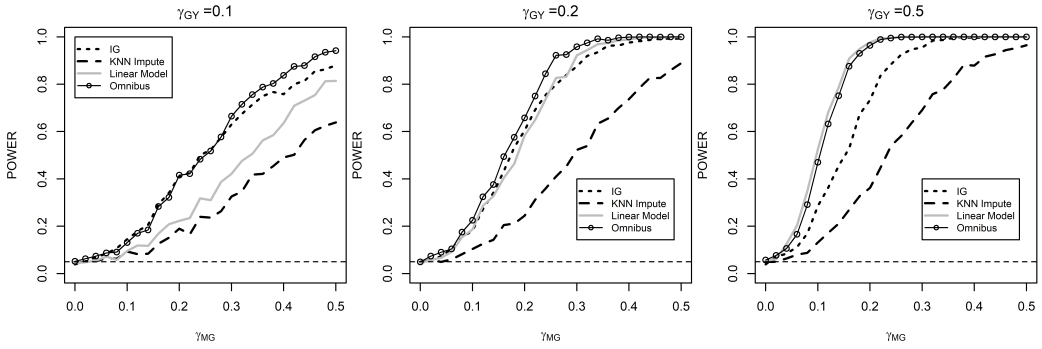


Figure 2.7. Power curves for omnibus method and competing methods in high-dimensional case under MAR. The number of factors $k = 5$ was determined by the 10-fold cross-validation. The value of γ_{GY} was set to be 0.1, 0.2 and 0.5.

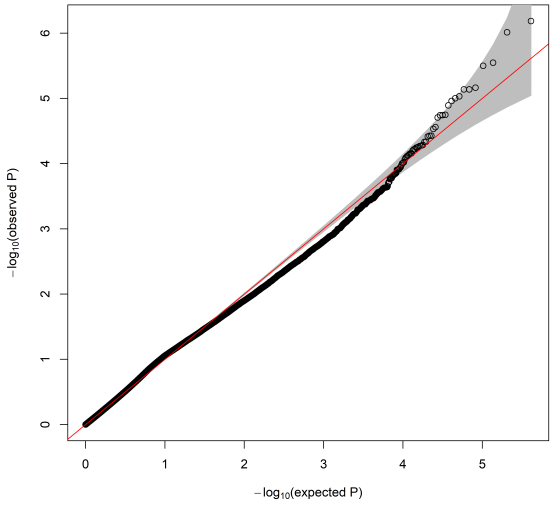


Figure 2.8. QQ-plots for p-values after SVA and genomic inflation factor adjustment for the proposed omnibus method.

As discussed in section 2.2, data was collected from 157 participants. However, mRNA gene expression measurements were only available for 55 participants. In the complete data set, after standardizing the mRNA expression measurements, we implemented the factor analysis as described in section 2.4.3 for 46,789 mRNA expression measurements. The optimal number of factors ($k = 7$) was determined by minimizing the Wold-style 10-fold cross-validated MSE (Owen and Perry, 2009).

For 410,735 CpG sites, we implemented our proposed weighting schemes one CpG site at a time to derive the weighted p-values for identifying the association with infant birth weights. The weighted p-values for all CpG sites with the corresponding chromosomes are presented in Figure 4. The number of significant CpG sites are provided in Web Appendix C. As presented in Web Table 2, none of the CpG sites were identified as significant in either q-value method or the weighted Bonferroni method after implementing our proposed omnibus method.

The top 30 CpG sites are listed in Table 2.3 with the corresponding reference sequence (RefSeq) gene symbols. The CpG sites are listed in the ascending order of p-values derived from the omnibus method after SVA and the genomic inflation

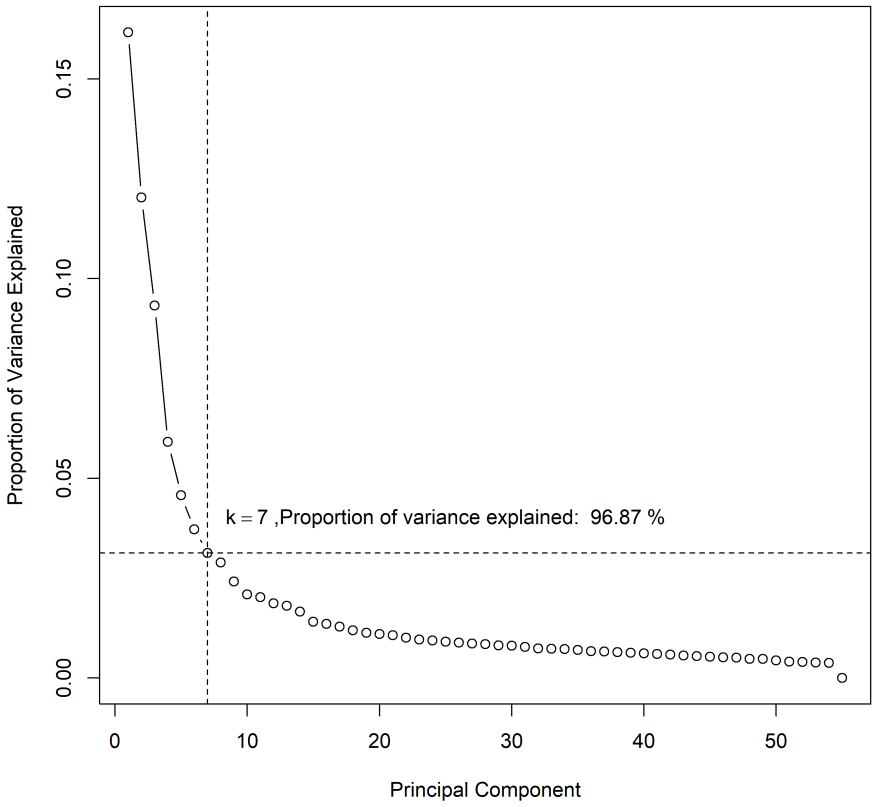


Figure 2.9. Plot for percentage of variance explained by the principal components (PC) over the number of PC. The first 10 principal components explain 97.9% of the variance of gene expression data.

factor adjustment. As shown in Table 2.3, *PIM1* is the gene that the second most significant CpG site is located on. *PIM1* was found to be expressed in B-lymphoid and myeloid cells and the methylation of cg25325512 (*PIM1*) mediates the effect of maternal smoking on the birth weight of infants (Witt et al., 2018).

2.6 Discussion

In this chapter, we propose a novel framework to implement integrative analysis for multi-omics data where the intermediate variables, such as mRNA gene expression measurements, are completely missing for a large proportion of subjects. Existing

Table 2.3. Top table for 30 CpG sites associated with birth weight scores of preterm infants with the smallest weighted p-values derived from the proposed omnibus method. Chromosomes, CpG sites, UCSC RefSeq gene names, weighted p-values and q-values are reported. The results were adjusted for paternal age, maternal age, paternal BMI, maternal BMI, maternal smoking status before pregnancy, and the gender of infants.

Rank	Chromosome	CpG site	Gene name	Weighted p-value	q-value
1	2	cg10855531		6.52e-07	0.199
2	6	cg25325512	PIM1	9.70e-07	0.199
3	2	cg20219891		2.84e-06	0.324
4	2	cg13872898		3.16e-06	0.324
5	3	cg25773386	GPR62	6.85e-06	0.428
6	2	cg00637745		7.29e-06	0.428
7	2	cg14311362		7.30e-06	0.428
8	2	cg07133097		9.30e-06	0.45
9	3	cg08120226	MUC4	9.96e-06	0.45
10	2	cg17870997		1.10e-05	0.45
11	11	cg08162803	TCIRG1	1.28e-05	0.479
12	7	cg18158419	FIS1	1.78e-05	0.533
13	18	cg15871086		1.81e-05	0.533
14	11	cg13746740		1.82e-05	0.533
15	5	cg25368647	MXD3	1.97e-05	0.538
16	7	cg06557644	NOD1	2.76e-05	0.708
17	6	cg16199747		2.94e-05	0.711
18	1	cg23305408	COL16A1	3.74e-05	0.79
19	2	cg17979068	C2orf48	3.75e-05	0.79
20	12	cg21052873	NCOR2	3.85e-05	0.79
21	14	cg07159958	ATL1	4.61e-05	0.829
22	2	cg02100410	LOC151534	4.66e-05	0.829
23	5	cg05992340	PDE4D	5.17e-05	0.829
24	6	cg13892322	LY6G5C	5.28e-05	0.829
25	11	cg10469774		5.36e-05	0.829
26	1	cg12723026	KCNC4	5.50e-05	0.829
27	19	cg24449014	MEGF8	5.61e-05	0.829
28	6	cg21457401	PNLDC1	5.69e-05	0.829
29	3	cg13268590	MAGI1	5.99e-05	0.829
30	2	cg02640173	GLI2	6.05e-05	0.829

multi-omics integrative studies require removing missing records or applying data imputation techniques to prepare a complete dataset for the analysis. However, when the missing rate is high, especially higher than 50%, the power of complete case analyses and imputation methods decreases drastically due to the reduction in sample size. Our proposed framework utilizes a p-value weighted adjustment and hence incorporates information from both complete and incomplete observations in the data.

The advantages of implementing the proposed framework in the multi-omics integrative analysis are multifold. First, by incorporating the information from incomplete observations, our proposed approaches boost the power of multi-omics integrative analyses compared to the existing methods. Second, our proposed approaches perform well even in situations with a large missing proportion of intermediate variables. Third, the two component weighting schemes combined in the omnibus test can provide flexibility in the implementation of multi-omics integrative analyses with missing rates ranging from 0 to 1. Furthermore, our simulation analyses showed that the proposed method maintains proper FWER control with the weighted bonferroni method and FDR control with the q-value method. According to Storey et al. (Storey and Tibshirani, 2003; Storey et al., 2004), the utilization of q-value method can also maintain FDR even with weak dependence structures between CpG sites. In addition, the proposed method can be implemented in a computationally efficient manner because the weights can be easily computed.

In our proposed method, the two component weighting schemes perform differently in datasets with various missing rates. According to section 2.4.5, the general weighting scheme achieves greater power in cases where the missing rates are lower than 50% while the reverse weighting scheme achieves better performance when the missing rates are high. A possible strategy to utilize the advantages of both methods is to set a dynamic contribution constant in terms of the missing rate in the omnibus test to up-weight the general scheme in situations where the missing rate is low and

emphasize the reverse scheme when the missing rate is high. The simulation studies show that the proposed omnibus method demonstrates more competitive results in both low and high missing rate situations. However, our proposed method can only be applied to datasets with missing records in the intermediate variables of the multi-omics integrative analysis. When phenotypic outcomes or independent variables are also missing, methods such as data imputations could be considered.

We implemented the proposed method in a birth weight study of preterm infants and identified the CpG sites with DNA methylation that are associated with birth weights of preterm infants via the regulation of gene expression. In practice, our analytical framework can be directly applied to any continuous independent variables (such as DNA methylation levels), or discrete variables (such as SNP genotypes or DNA mutation status). Since the measurements were of intensity values generated via microarray experiments, we assumed that the intermediate variables followed normal distributions. However, if the intermediate measurements are generated by high-throughput RNA sequencing (RNAseq), preprocessing procedures such as normalization by the sequencing depth and log transformation of the data, as described in the limma (Cloonan et al., 2008) or voom (Law et al., 2014), could be applied for ensuring normality before implementing the weighting schemes.

As discussed in sections 2.3 and 2.5, the implementation of SVD-based dimension reduction techniques allows us to apply our integrative framework to datasets with high-dimensional intermediate variables. Other variable selection approaches such as LASSO could also be used to reduce the dimension of mRNA gene expression in the analysis. Another future research area is the Cox model for survival outcomes in the integrative framework. Applying the Cox regression in our proposed integrative framework would require the implementation of estimating equation theory and to derive the asymptotic distribution of the estimates (Zhao et al., 2014). Therefore, further work is needed to develop multi-omics integration frameworks for survival

outcomes.

Chapter 3

Identification of dynamic gene co-expression via Bayesian variable selection

3.1 Introduction

A wealth of gene expression data generated by high-throughput techniques provide exciting opportunities for studying genetic interactions systemically. Genetic interactions in a biological system are tightly regulated and are often highly dynamic. The interactions can change flexibly under various internal cellular signals or external stimuli. Previous studies have developed statistical methods to examine these dynamic changes in genetic interactions. However, a common challenge encountered in the existing approaches is the computational intensiveness due to the massive amount of gene combinations needed to be considered in a typical genomic dataset, and yet often a much smaller proportion of gene interactions exhibit dynamic co-expression changes.

To solve this problem, we propose variable selection methods in Bayesian frameworks with spike-and-slab priors. The proposed algorithms focus on subsets of promising gene combinations in the search space. We also adopt a Bayesian false discovery control procedure for testing the significance of dynamic gene co-expression changes. Simulation studies are conducted comparing our proposed approaches to existing exhaustive search heuristics. We demonstrate the implementation of our proposed approaches in experimental data analysis to study gene co-expression changes associated with colorectal cancer recurrence-free survival.

In this chapter, we set out to identify dynamic gene co-expression (DC) patterns that are associated with cancer recurrence risk. Previous studies of cancer recurrence risk mainly focused on one-dimensional gene expression levels (Mallmann et al., 2010; Pan et al., 2019; Shi et al., 2017). Depicting dynamic joint gene co-expression patterns would offer new insights into the mechanisms of cancer recurrence and provide new biomarkers for treatment selection and prognosis prediction.

This chapter is structured as follows. Section 3.2 provides an overview of the dataset and Section 3.3 introduces the proposed models as well as the estimation approaches of the proposed spike-and-slab model. Section 3.4 presents two simulation studies with different degrees of sparsity in the search space. Section 3.5 demonstrates the implementation of our proposed approaches with the experimental data analysis on a colorectal cancer (CRC) dataset. In section 3.6, we provide discussions and suggestions for future research on these topics.

3.2 Datasets and Databases

The CRC dataset used in this chapter is based on the study by Smith et al. (2010) available in Gene Expression Omnibus with accession number GSE17536. This study included data from 177 CRC patients, and gene expression levels for 54,675 transcripts were measured using Affymetrix Human Genome U133 Plus 2.0 (hgu133plus2) platform. For identifying DC gene pairs associated with recurrence-free survival, 20,186 genes with unique gene symbols were selected. We excluded the participants who dropped out at the beginning of the study and use the remaining 145 records in the experimental data analysis. The cancer recurrence-free survival time was defined as the length of time that CRC patients live without cancer recurrence after the surgery.

3.3 Model

Let $\mathbf{X}_i = (X_{i,1}, ..., X_{i,q})^T$ be the vector of q gene expression levels of the i-th participant, $i = 1, ..., n$, with each following a standard normal distribution. The vector of the pair-wise products of $\mathbf{X}_i$ is denoted as $\mathbf{Y}_i = (Y_{i,1}, ..., Y_{i,m})^T = (X_{i,1} \cdot X_{i,2}, X_{i,1} \cdot X_{i,3}, ..., X_{i,q-1} \cdot X_{i,q})^T$. In j-th pair of gene combinations, $j = 1, ..., m$, the pair of genes used to calculate the product can be denoted as X_{i,j_1} and X_{i,j_2}, $1 \leq j_1 < j_2 \leq q$. For a set of q genes, the total number of gene-pair combinations is $m = \binom{q}{2}$. In practice, q could be around 20,000 and $m \approx 2 \times 10^8$ as demonstrated in the experimental data analysis.

Let z_i be the modulating factor, which represents the recurrence-free survival time for the i-th participant. As discussed in section 3.1, DC can be demonstrated using the conditional expectation of the product of two gene expression levels given the modulating factor if the two genes are standard normal (Li, 2002), that is, $\rho(X_{i,j_1}, X_{i,j_2} \mid z_i) = E(X_{i,j_1} X_{i,j_2} \mid z_i)$. Based on the normality assumption for $\mathbf{X}_i$, the product vector of the expression levels $\mathbf{Y}_i$ can be used as the measurement of the gene co-expression to formulate the model of DC.

Nadarajah and Pogány (2016) derived the probability density function for the product of two standard normal gene expression levels, $Y_{i,j} \mid \rho_{i,j} \sim \mathcal{NP}(\rho_{i,j})$, with $\rho_{i,j}$ presenting the correlation coefficient between j-th pair of genes for i-th participant. The distribution has the following formulation,

$$f(Y_{i,j} = y_{i,j} \mid \rho_{i,j}) = \frac{1}{\pi\sqrt{1-\rho_{i,j}^2}} e^{\frac{\rho_{i,j} y_{i,j}}{1-\rho_{i,j}^2}} K_0\left(\frac{|y_{i,j}|}{1-\rho_{i,j}^2}\right), \tag{3.1}$$

with $K_0(\cdot)$ being the second class zero-order modified Bessel function, and $E(Y_{i,j}) = \rho_{i,j}$. Next, we consider a general form of DC between the modulating factor and the co-expression of the gene pairs (Ho et al., 2011; Kinzy et al., 2019; Yang and Ho, 2021),

$$g(\rho_{i,j}) = g(\rho(X_{i,j_1}, X_{i,j_2} \mid z_i)) = \tau_{0,j} + \tau_{1,j} \times z_i, \tag{3.2}$$

where $g(\cdot) = \frac{1}{2}\log\left(\frac{1+(\cdot)}{1-(\cdot)}\right)$ is the Fisher's Z-transformation to ensure that $\rho_{i,j}$ falls within (-1,1). Considering the inverse function of the Fisher's Z-transformation, $g^{-1}(\cdot) = \frac{e^{2(\cdot)}-1}{e^{2(\cdot)}+1}$, DC between j-th pair of genes has the following form, $\rho(X_{i,j_1}, X_{i,j_2} \mid z_i) = E(X_{i,j_1}X_{i,j_2} \mid z_i) = E(Y_{i,j} \mid z_i) = g^{-1}(\tau_{0,j} + \tau_{1,j}z_i)$. Our goal is to estimate $\tau_{1,j}$ to quantify DC with respect to the modulating factor z_i. Thus the DC model can be formulated as follows,

$$Y_{i,j} \mid \tau_{0,j}, \tau_{1,j}, z_i \overset{ind}{\sim} \mathcal{NP}\{g^{-1}(\tau_{0,j} + \tau_{1,j}z_i)\}, \tag{3.3}$$

In Equation (3.3), $\tau_{1,j}$ is the main parameter of interest and represents the strength of association between $\mathbf{z}$ and the correlation between j-th pair of genes. A positive $\tau_{1,j}$ suggests that the increasing positive correlation is associated with longer recurrence-free survival time. When $\tau_{1,j} = 0$, the correlation of gene pair j is not associated with the recurrence-free survival time. The intercept $\tau_{0,j}$ is the Fisher's Z-transformed correlation coefficient between j-th pair of genes when the patient's colorectal cancer recurrence-free survival time is 0.

To develop the variable selection models in a Bayesian framework, we need to specify the prior distributions for model parameters. The prior for $\tau_{0,j}$ can be set to $\mathcal{N}(0,1)$. Theoretically, the approximate variance of Fisher's Z-transformation is $\frac{1}{n-3}$, with n presenting the sample size (Fisher et al., 1921). The $\mathcal{N}(0,1)$ prior for $\tau_{0,j}$ yields sufficiently large variance and performs well in our simulation analyses as described in Section 3.4. The prior distributions for $\tau_{1,j}$ are the cores in the structure of the variable selection models proposed in this chapter, which will be discussed next.

3.3.1 SPSL model

To effectively identify the dynamic gene co-expression within sparse DC signals, we consider the continuous spike-and-slab prior (Ishwaran et al., 2005) for $\tau_{1,j}$ as follows,

$$\begin{aligned} \tau_{1,j} \mid \gamma_j, v_j^2 &\overset{ind}{\sim} N(0, \gamma_j v_j^2), \\ v_j^2 \mid \alpha_1, \alpha_2 &\overset{i.i.d.}{\sim} \mathcal{IG}(\alpha_1, \alpha_2), \\ \gamma_j \mid w_j &\overset{ind}{\sim} w_j \mathcal{I}_1 + (1 - w_j)\mathcal{I}_{v_0}, \\ w_j &\overset{i.i.d.}{\sim} \mathcal{U}(0, 1). \end{aligned} \tag{3.4}$$

In this model framework, the variance of the prior distribution for $\tau_{1,j}$ is the product of two parameters γ_j and v_j^2. The prior of the γ_j is a mixture of two point mass measures, and v_j^2 has an inverse gamma prior with shape and scale hyperparameters $\alpha_1 = 5$ and $\alpha_2 = 50$ (mean is 12.5). Figure 3.1 illustrates the shapes of the conditional densities for the bimodal inverse gamma distributions ($\gamma_j v_j^2$) with different values for w_j and v_0. The extent of shrinkage of $\tau_{1,j}$ toward 0 is controlled by the selection variable w_j. A small value of w_j leads to more shrinkage. We set $v_0 = 0.005$ in our analyses following Ishwaran et al. (2005).

3.3.2 C-SPSL model

Consider the fact that genetic interactions can be modulated by same factors, the shrinking of different gene-pair combinations can not be entirely independent (Lee et al., 2017). Therefore, we further develop a correlated spike-and-slab model (C-SPSL) to incorporate the dependencies among selection parameters $\mathbf{w}$. Instead of assigning independent uniform priors for $\mathbf{w}$, we take the idea of assuming a global prior for the probabilities of shrinking to derive dependent posterior distributions of

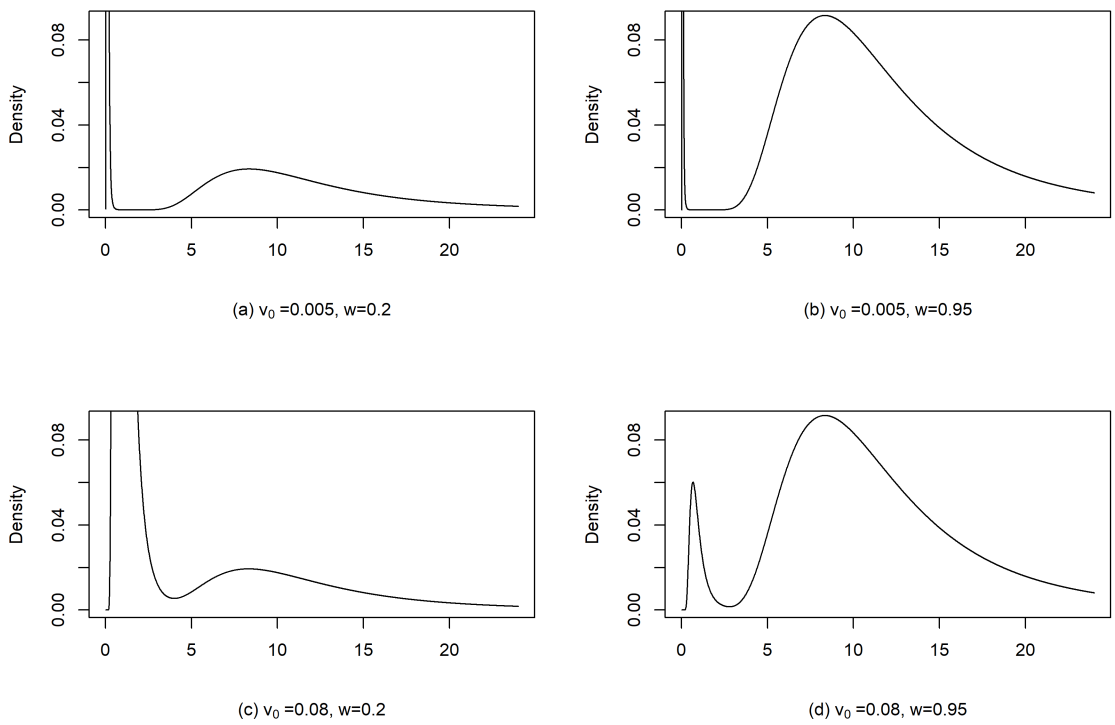


Figure 3.1. Conditional density plots for the bimodal inverse gamma distributions with $\alpha_1 = 5$ and $\alpha_2 = 50$. Plot (a) and (b) present the different shapes of the densities with $w = 0.2$ and $w = 0.95$ respectively while fixing $v_0 = 0.005$. Plot (c) and (d) are generated with fixing $v_0 = 0.08$ and varied $w = 0.2$ and $w = 0.95$ respectively.

w. The structure of C-SPSL model is described as follows:

$$\begin{aligned} \tau_{1,j} \mid \gamma_j, v_j^2 &\overset{ind}{\sim} N(0, \gamma_j v_j^2) \\ v_j^2 \mid \alpha_1, \alpha_2 &\overset{i.i.d.}{\sim} \mathcal{IG}(\alpha_1, \alpha_2), \\ \gamma_j \mid w_j &\overset{ind}{\sim} w_j \mathcal{I}_1 + (1 - w_j)\mathcal{I}_{v_0} \\ w_j \mid c &\overset{i.i.d.}{\sim} Beta(c, 1 - c), \\ c &\sim \mathcal{U}(0, 1), \end{aligned} \tag{3.5}$$

In Equation (3.5), **w** is set with priors of the beta distributions with a global hyperparameter c instead of independent standard uniform priors. The global hyperparameter c allows the marginal distributions of **w** to be correlated. The priors for $\tau_{0,j}$, v_j^2, γ_j and the settings of hyperparameters v_0, α_1 and α_2 are the same as in the SPSL model.

3.3.3 Estimation

Let θ denote the vector of all the parameters. The likelihood function of θ is given by

$$\mathcal{L}(\theta \mid \mathbf{Y}, \mathbf{z}) = \prod_{i=1}^{n} \prod_{j=1}^{m} f(Y_{i,j} \mid z_i, \theta), \tag{3.6}$$

where $f(\cdot)$ is the density function of $Y_{i,j}$ given in Equation (3.3). The posterior joint distribution for θ with prior $\pi(\theta)$ can be described as

$$\pi(\theta \mid \mathbf{Y}, \mathbf{z}) \propto \prod_{i=1}^{n} \prod_{j=1}^{m} f(Y_{i,j} \mid z_i, \theta) \times \pi(\theta). \tag{3.7}$$

The main difference between the SPSL model and the C-SPSL model is the prior distribution of w_j. The full conditional distributions involved in the posterior computation are derived and sketched in Web Appendix A.

In order to develop an efficient Markov Chain Monte Carlo (MCMC) algorithm, we implement the adaptive Metropolis algorithm (Haario et al., 2001). Different multivariate normal distributions are used as the proposal distributions in our algorithm to maintain an acceptance rate close to 20-30%. The details of the adaptive algorithm and operation process are provided in Web Appendix B.

3.3.4 Hypothesis Testing

To declare significant $\tau_{1,j}$, we implement the following hypothesis testing procedure. Because the distributions of $\tau_{1,j}$ are continuous and hence $P(\tau_{1,j} = 0) = 0$, a small value $\epsilon > 0$ is used to define the local null and alternative hypotheses. For j-th pair of genes, the local hypotheses are defined as, $H_{0j} : |\tau_{1,j}| < \epsilon$ vs. $H_{aj} : |\tau_{1,j}| \geq \epsilon$, $j = 1, ..., m$. In practice, The value of ϵ can be determined based on expert knowledge as illustrated in our experimental data analysis (section 3.5). With the MCMC samples, we can calculate the posterior probability of the local null hypothesis,

$$\mathrm{P}(H_{0j} \mid Data) = \frac{1}{h} \sum_{l=1}^{h} 1_{(|\tilde{\tau}_{1,j}^{(l)}| < \epsilon)}, \tag{3.8}$$

where h is the number of iterations after the burn-in period and $\tilde{\tau}_{1,j}^{(l)}$ is the sample of $\tau_{1,j}$ at the l-th iteration after the burn-in period. The local alternative hypothesis is defined as $\mathrm{P}(H_{aj} \mid Data) = 1 - \mathrm{P}(H_{0j} \mid Data)$.

We propose to use Bayes factor to assess the degree of evidence in favor of the alternative hypothesis. Unlike using the posterior probability of the alternative hypothesis, using Bayes factor will balance out the effect of the prior by taking into account the prior probabilities of the hypotheses. Specifically, the Bayes factor of the alternative hypothesis for the j-th local hypothesis test is defined as follows,

$$BF_j = \frac{\frac{P(H_{aj}|Data)}{P(H_{0j}|Data)}}{\frac{P(H_{aj})}{P(H_{0j})}},$$

where $P(H_{0j})$ and $P(H_{aj})$ represent the prior probabilities of local null and alternative hypotheses respectively. We reject the null hypothesis H_{0j} and claim the j-th gene pair is significant if BF_j is greater than some critical value r. In the Bayesian literature, taking $r = 3$ and 10 represents substantial and strong evidence of favoring the alternative hypothesis, respectively.

Considering the large number of local hypotheses tests in our study, multiplicity has to be addressed to avoid an inflated false positive rate (FDR). For this purpose, a Bayesian multiple testing procedure proposed by Müller et al. (2004) and Wang and Dunson (2010) is applied to control the expected false discovery rate. As suggested by Wang and Dunson (2010), the expected false discovery rate ($\overline{\mathrm{FDR}}$) is defined as a function of r in the following form,

$$\overline{\mathrm{FDR}}(r) = \frac{\sum_j 1_{(BF_j \geq r)}(1 - \delta_j)}{\sum_j 1_{(BF_j \geq r)} + e},$$

where $\delta_j = \mathrm{P}(H_{aj} \mid Data)$ is the posterior probability of the local alternative hypothesis H_{aj} and e is a small positive value (e.g. 0.0001) that prevents the denominator from being zero. By controlling the $\overline{\mathrm{FDR}}$ at the pre-specified significance level α, we can determine the optimal critical value: $r^{opt} = \min_r\{r \in [0, +\infty) : \overline{\mathrm{FDR}}(r) \leq \alpha\}$ and reject H_{0j} if $BF_j > r^{opt}$ for each j.

3.4 Simulation

In this section, we assessed the performance of the proposed DC variable selection models in a Bayesian framework and compared them with a competing approach via implementing simulation studies. The simulated data sets were generated under two scenarios with different degrees of sparsity as well as different dimensions of search space.

3.4.1 Settings

In the first simulation scenario, we compared our proposed models (SPSL and C-SPSL) to the conditional normal model (Ho et al., 2011; Kinzy et al., 2019) in a Bayesian framework. The competing method assumed each gene pair to be a bivariate normal distribution while allowing the modulating factors to have an arbitrary distribution. Comparing to our proposed models, the conditional normal model used an exhaustive search strategy to implement a genome-scale scan for all possible gene-pair combinations without handling the sparsity. To present the difference between our proposed approaches and the competing method, we named the competing method as exhaustive search method (ES).

We set a small number of gene-pair combinations and moderately sparse DC signals in this scenario with 5 genes ($q = 5$), resulting in 10 genes pairs ($m = 10$). The vector of main parameters $\tau_{\mathbf{1}}$ was set to be $(0, 0, 0, 0, 0, 0, 0, 1, 1, 1)^T$, where 70% of the values were zero so that the sparsity in this scenario was 70%.

In the second scenario, we considered a larger number of gene-pair combinations to evaluate the performance of our proposed models in large search space with highly sparse DC signals. In this simulation, 50 genes ($q = 50$) and hence 1,225 gene-pair combinations ($m = 1,225$) were generated. In the vector of τ_1, only 10 gene pairs were set with $\tau_{1,j} = 1$ and all other pairs were 0 so that the sparsity indicated by the proportion of zero values is 99.18%. As discussed in section 3.1, however, the

ES method quickly became computational intractable as the number of gene-pair combinations increased. In this scenario, the ES method was inapplicable due to the astronomical computational time under the same number of iterations and chains of MCMC as the proposed methods. We therefore only presented the results of the proposed models in this case. The true values of the intercepts $\tau_{\mathbf{0}} = (\tau_{0,1}, ..., \tau_{0,m})^T$ were all set equal to zero in both scenarios.

The simulation data were generated in the same procedures in both scenarios. For observation i, we first simulated the covariate z_i from $\mathcal{U}(0,1)$ which represented the recurrence-free survival time. The correlation $\rho_{i,j}$ between the j-th pair of gene X_{i,j_1} and X_{i,j_2}, $1 \leq j_1 < j_2 \leq q$, was then calculated using Equation (3.2). Next, the covariance matrix Σ_i for all genes can be constructed as follows, $\Sigma_i = cov(\mathbf{X}_i)$, where $cov(X_{i,j_1}, X_{i,j_2}) = 1$ if $j_1 = j_2$, and $cov(X_{i,j_1}, X_{i,j_2}) = cov(X_{i,j_2}, X_{i,j_1}) = \rho_{i,j}$ if $j_1 \neq j_2$. After deriving the covariance matrix, the gene expression vector $\mathbf{X}_i$ was generated from a multivariate normal distribution, $\mathbf{X}_i \sim N_q(\mathbf{0}, \Sigma_i)$. It was noted that the marginal distributions of $\mathbf{X}_i$ were all standard normal distributions due to unit diagonal elements in Σ_i. In the last step, we obtained $\mathbf{Y}_i$ by taking pairwise products of the gene expression vector $\mathbf{X}_i$.

For hypothesis testing, the value of ϵ was set equal to 0.2 in the first scenario (moderately sparse case), and 0.5 in the second scenario (highly sparse case). Since the true statuses of the hypotheses were known in the simulation, the empirical false discovery rate (FDR), false negative rate (FNR) can be calculated based on the threshold of the Bayes factor described in Section 3.3.4:

$$\text{FDR} = \frac{\sum_j 1_{(BF_j \geq r^{opt})} \times 1_{(\tau^0_{1,j}=0)}}{\sum_j 1_{(BF_j \geq r^{opt})} + e}, \tag{3.9}$$

$$\text{FNR} = \frac{\sum_j 1_{(BF_j < r^{opt})} \times 1_{(\tau^0_{1,j}\neq 0)}}{\sum_j 1_{(BF_j < r^{opt})} + e}, \tag{3.10}$$

where r^{opt} is the optimal criterion that is determined by controlling the $\overline{\text{FDR}}$ and $\tau^0_{1,j}$ is the true value of dynamic gene co-expression for the j-th pair of gene combinations.

3.4.2 Results

The adaptive MCMC procedures were implemented to estimate the parameters in SPSL and C-SPSL models. We set the number of iterations to be 20,000 with a burn-in period of 10,000 iterations. To evaluate the convergence of MCMC sampling, we implemented the Gelman-Rubin convergence diagnostics (Gelman et al., 1992) on three chains by comparing the between-chains and within-chains variances. As suggested by Brooks and Gelman (1998), a criterion of the diagnostic statistics being less than 1.2 for all parameters indicated convergence for MCMC samples. As presented in diagnostic results (Web Appendix D), the Gelman-Rubin statistics were close to 1 for all $\tau_{1,j}$ and the tracing plots are mixed well. The results indicate good convergence for both SPSL and C-SPSL. All simulations were implemented on the Intel(R) Xeon(R) CPU @ 2.40GHz processors.

Table 3.1 presents FDR and FNR for our proposed approaches and the ES method with different sample sizes, 200, 500, and 1,000. The proposed approaches have lower FDR than ES, while the FNR of the proposed models is slightly higher when the sample sizes are small. As the sample size increases to 1,000, the FNR of both SPSL and C-SPSL decreases to 0 while still maintaining a smaller FDR than ES.

Table 3.1. Comparison of ES, SPSL and C-SPSL model based on 100 simulation iterations in scenario I (sparsity = 70%). The true values of $\tau_{\mathbf{1}}$ are set to be $(0,0,0,0,0,0,0,1,1,1)^T$ and the true values $\tau_{\mathbf{0}}$ are set to 0. The false discovery rate (FDR) and false negative rate (FNR) are reported.

Sample size	Method	FDR	FNR
$n = 200$	ES	0.1790	0.0067
	SPSL	0.0200	0.0652
	C-SPSL	0.0108	0.0903
$n = 500$	ES	0.0530	0.0000
	SPSL	0.0150	0.0012
	C-SPSL	0.0100	0.0012
$n = 1,000$	ES	0.0175	0.0000
	SPSL	0.0025	0.0000
	C-SPSL	0.0025	0.0000

To further compare the efficiency of SPSL and C-SPSL, we performed a second simulation study with 50 genes and 1,225 gene-pair combinations with highly sparse signals. It was impractical to implement ES in this scenario due to the computational intensiveness, and only the two proposed approaches were implemented. In Table 3.2, the C-SPSL are preferable as FDR and FNR are both much smaller.

Table 3.2. Comparison of SPSL and C-SPSL model based on 100 simulation iterations in scenario II with with 1,225 gene pairs. The true values of $\tau_{\mathbf{1}} = 0$ except for 10 gene pairs $\tau_{\mathbf{1}} = 1$ (sparsity=99.2%). The true values of $\tau_{\mathbf{0}}$ are set to be all 0. The false discovery rate (FDR) and false negative rate (FNR) are reported.

	FDR	FNR
SPSL	0.0306	0.0055
C-SPSL	0.0197	0.0042

3.5 Experimental Data Analysis

In this section, we implemented the proposed C-SPSL model to identify gene pairs with significant non-zero $\tau_{1,j}$ (DC gene pairs) associated with the recurrence-free survival time. In the CRC dataset considered in this study, a total number of 2.04×10^8 gene-pair combinations needed to be considered with the 20,186 unique gene expression measurements. A screening measure $\zeta_j = Cor(|X_{j_1}|, |X_{j_2}|) - |Cor(X_{j_1}, X_{j_2})|$ was applied to identify 949,280 ($\approx 10^6$) gene pairs that potentially exhibit DC patterns (Yu, 2018), where j_1 and j_2 represent the two genes for j-th pair. After the screening step, we chose to implement C-SPSL on all 949,280 gene pairs because it performed slightly better than SPSL in the highly sparse situation considered in this experimental data analysis. To improve the computational efficiency, we used computer clusters to separate all selected gene pairs into 1,000 groups with each including 950 pairs and apply the C-SPSL model for each group parallelly.

We used individual patients' recurrence-free survival time ($\mathbf{z}$) to indicate their ability to survive without CRC recurrence or metastasis after the surgery. If the

recurrence-free survival time of i-th observation was right-censored, the predicted value $E(T_i \mid T_i > t_{0_i})$ was used, with T_i representing the CRC recurrence-free survival time and $t_{0,i}$ representing the censoring time for observation i. The calculation of $E(T_i \mid T_i > t_{0,i})$ was described in Web Appendix E. Finally, the recurrence-free survival time estimates were re-scaled by their maximum value in the CRC dataset so that all z were within $(0, 1]$.

The parameters in the C-SPSL model qwre estimated using the adaptive MCMC algorithm with 20,000 iterations and 10,000 burn-in. As described in Section 3.3.4, The hypothesis testing procedure was implemented to identify gene pairs with significant DC signals. We set $\epsilon = 0.5$ in the local hypotheses as presented in Equation (3.8). By controlling the $\overline{\text{FDR}}$ at 0.05 level, 2,570 gene pairs were identified to be statistically significant. The DC signals in the CRC dataset can be considered as highly sparse ($\frac{2{,}570}{949{,}280} \approx 0.3\%$). The empirical distribution of $\widehat{\tau}_1$ (the posterior mean of τ_1) based on all $949,280$ gene-pair combinations was presented in Web Figure 1. It was observed that the proportion of $|\widehat{\tau}_1|$ being greater than 0.5 is 0.023.

The top 30 gene pairs with the largest posterior mean $|\widehat{\tau}_1|$ from the CRC dataset are listed in Table 3.3. The trace plots and the Gelman-Rubin diagnostic results for the top $|\widehat{\tau}_{1,j}|$ are provided in Web Appendix D.

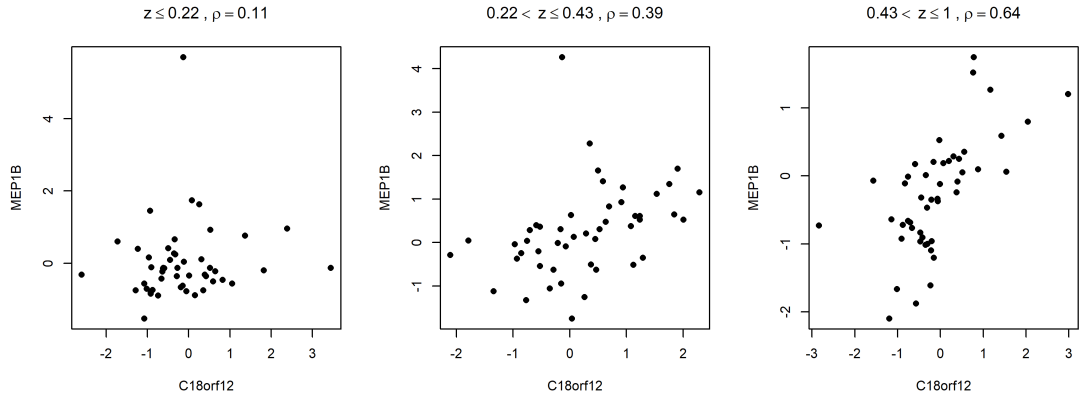
To visualize the DC patterns, the profile plots of the top two gene pairs are presented in Figure 3.2 with the recurrence-free survival time (z) binned into 3-quantiles. Among all the gene pairs, *MEP1B* and *C18orf12* have the largest $\widehat{\tau}_1 = 2.981$ associated with the recurrence-free survival time. The meprin β (*MEP1B*) was found to be positively correlated with the colorectal cancer survival prognosis (Peters and Becker-Pauly, 2019). Although the function of *C18orf12*-encoded protein has not been fully characterized, our analysis reports the positive τ_1 estimate between *MEP1B* and *C18orf12* with the colorectal cancer recurrence-free survival time. Our results suggest patients with a larger correlation between *MEP1B* and *C18orf12* gene

Table 3.3. The top 30 significant gene pairs with the largest posterior mean $|\widehat{\tau}_1|$ associated with recurrence-free survival time. The 95% posterior credible intervals (95% CI) for each parameter are provided in the parentheses.

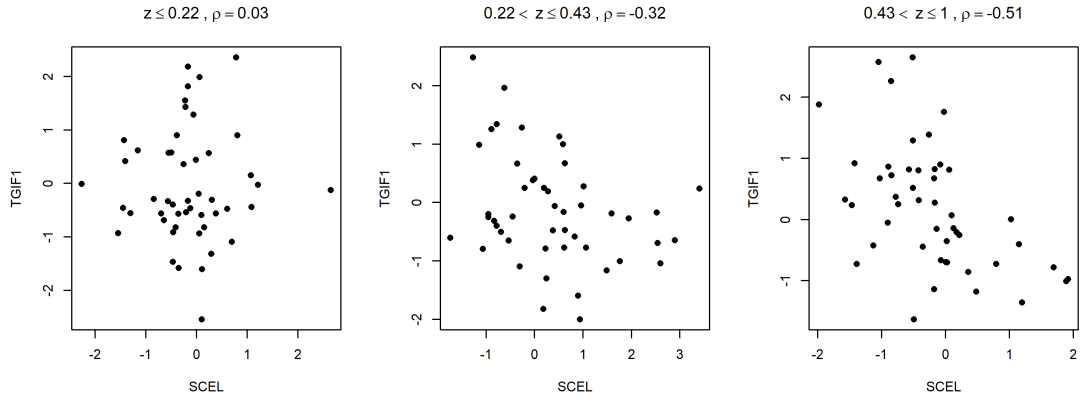
j	Gene I	Gene II	$\widehat{\tau}_{1,j}$ (95% CI)	$\widehat{\tau}_{0,j}$ (95% CI)
1	C18orf12	MEP1B	2.981 (1.722, 4.235)	-0.197 (-0.682, 0.221)
2	SCEL	TGIF1	-2.933 (-4.226,-1.493)	0.614 (0.082, 1.123)
3	POLR3A	MEP1B	2.927 (1.712, 4.150)	0.131 (-0.324, 0.510)
4	MEP1B	MIR4500HG	2.890 (1.543, 4.314)	-0.161 (-0.636, 0.281)
5	PRR16	PLAU	2.822 (1.426, 4.122)	0.089 (-0.307, 0.508)
6	LOC642696	DDX25	2.812 (1.374, 4.073)	-0.193 (-0.587, 0.236)
7	METTL6	MEP1B	2.807 (1.408, 4.164)	-0.035 (-0.531, 0.460)
8	ZMYM2	AMELX	-2.732 (-3.736,-1.689)	0.814 (0.404, 1.231)
9	PIK3CA	PIP	-2.730 (-3.509,-1.826)	0.874 (0.561, 1.145)
10	KIF19	LOC105379426	-2.679 (-3.596,-1.711)	0.905 (0.525, 1.324)
11	KCNS1	TEX48	2.666 (1.187, 5.143)	0.337 (-1.066, 0.850)
12	LINC01013	SMYD1	2.635 (1.423, 3.771)	-0.369 (-0.826, 0.056)
13	LOC100129175	ZNF787	2.626 (1.388, 3.757)	-0.947 (-1.385,-0.466)
14	PIK3CA	SCGB1D1	-2.615 (-3.495,-1.618)	0.884 (0.548, 1.197)
15	FMN2	SH2B1	-2.564 (-3.285,-1.667)	0.965 (0.633, 1.261)
16	SCGB1D2	SNHG10	2.552 (1.526, 3.590)	-0.902 (-1.317,-0.497)
17	C16orf74	TBC1D21	2.539 (1.644, 3.365)	-0.832 (-1.182,-0.440)
18	LOC389199	MYH7B	-2.535 (-3.340,-1.737)	1.150 (0.778, 1.512)
19	MYH7B	UCP1	-2.516 (-3.375,-1.545)	1.056 (0.667, 1.411)
20	PLEK	AQP9	2.498 (1.212, 3.745)	0.169 (-0.261, 0.558)
21	FILIP1L	GDPD1	2.464 (0.749, 3.758)	0.001 (-0.443, 0.538)
22	LOC101928760	TMEM151B	2.463 (1.443, 3.594)	-0.821 (-1.237,-0.450)
23	MYH7B	ADARB2	-2.440 (-3.243,-1.684)	1.139 (0.805, 1.464)
24	LOC101928553	SPINT3	2.421 (1.267, 3.596)	0.060 (-0.332, 0.437)
25	KCNIP2	MLANA	2.419 (1.211, 3.383)	-0.879 (-1.302,-0.374)
26	MEP1B	NIPAL3	2.419 (1.027, 3.763)	-0.041 (-0.566, 0.467)
27	C11orf42	SNHG10	2.418 (0.995, 3.391)	-0.898 (-1.293,-0.386)
28	FAM218A	DIPK1B	2.379 (1.244, 3.331)	-0.725 (-1.131,-0.262)
29	LOC101928760	CLECL1	2.361 (1.361, 3.349)	-0.786 (-1.159,-0.415)
30	LINC00889	MEP1B	2.359 (1.069, 3.745)	0.157 (-0.259, 0.550)

expression are associated with longer CRC recurrence-free survival time.

Additionally, *SCEL* and *TGIF1* have the largest negative $\widehat{\tau}_1 = -2.933$ associated with the recurrence-free survival time. *TGIF1* was found to be positively correlated with the proliferation and migration of the CRC cancer cells (Wang et al., 2017) while the knockdown of *SCEL* was discovered to promote the migration and metastasis of CRC (Chou et al., 2016). Based on our analysis results, the negative correlation between *SCEL* and *TGIF1* expression is associated with longer CRC recurrence-free survival time.



(a) Profile plots between *C18orf12* and *MEP1B* with $\widehat{\tau}_1 = 2.98$



(b) Profile plots between *SCEL* and *TGIF1* with $\widehat{\tau}_1 = -2.93$

Figure 3.2. Profile plots for the top two gene pairs with the largest $|\widehat{\tau}_1|$ associated with recurrence-free survival time. (a) presents the profile plots between *C18orf12* and *MEP1B* with $\widehat{\tau}_1 = 2.98$; (b) presents the profile plots between *SCEL* and *TGIF1* with $\widehat{\tau}_1 = -2.93$. Gene expression levels were standardized to have mean 0 and variance 1 in both plots.

In Figure 3.3, the top 19 gene pairs (33 genes) with $|\widehat{\tau}_{1,j}|$ (≥ 2.5) are shown as nodes in this network plot. The thickness of the edges represents the values of $|\widehat{\tau}_{1,j}|$. Genes linked with dash lines in the network represent negative $\widehat{\tau}_{1,j}$ values while edges with solid lines denote positive $\widehat{\tau}_{1,j}$ values. The genes linked together can be seen as a co-expression gene group where the correlations among genes are closely related to the recurrence-free survival time. For example, *MEP1B*, *C18orf12*, *METTL6* and *POLR3A* are a such group where *MEP1B* acts as a "hub" that shows dynamic gene co-expression associated with colorectal cancer recurrence-free survival time with all other genes in this group.

3.6 Discussion

In the past two decades, several useful statistical tools (Li, 2002; Lai et al., 2004; Ho et al., 2011; Gunderson and Ho, 2014; Kinzy et al., 2019) have been developed for identifying dynamic co-expression in single combination of genes. When considering a small number of genes, one can perform the analysis for all possible gene combinations one after another. This search strategy is known as the exhaustive search (ES) algorithm. In a typical genome-wide dataset, the ES algorithm can quickly become computational intensive as the number of considered genes increases. This chapter aims to develop new DC models to supplement the statistical tools for high dimensional search space with highly sparse DC signals where the existing ES algorithm is intractable.

In this chapter, we propose the SPSL model and the C-SPSL model with the implementation of variable selection technique in a Bayesian framework. The proposed approaches possess several advantages. First, in a sparse data matrix, the proposed methods can shrink gene pairs with weak or zero DC signals via variable selection procedure so that the algorithms can reduce the false discovery error rate. Second, the proposed method has the capacity to incorporate the dependence structure be-

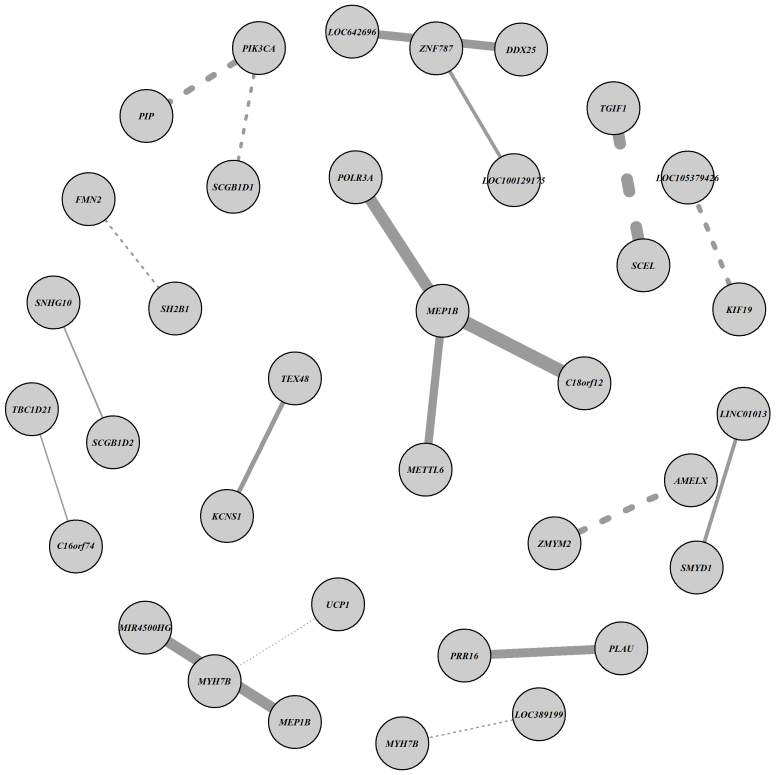


Figure 3.3. Network plot for gene pairs with $|\widehat{\tau}_{1,j}| \geq 2.5$ (33 genes with 19 pairs). Nodes represent genes and the thickness of the link edges indicate the values of $|\widehat{\tau}_{1,j}|$. Dotted lines indicate gene pairs with negative $\widehat{\tau}_{1,j}$ while solid lines denote positive $\widehat{\tau}_{1,j}$ values.

tween the parameters of interest $\tau_{\mathbf{1}}$ to improve the performance of estimates. By setting a global prior for the selection variable as presented in Equation (3.5), the C-SPSL model achieves lower FDR and FNR comparing to SPSL, in the situation with highly sparse signals and a large search space as seen in our simulation scenario II. Third, the proposed methods do not assume the modulating factors to be normally distributed gene expression data, providing great flexibility in the discovery of modulatory mechanisms in the biological systems. Fourth, the shrinkage nature allows the proposed method to focus on the small subset of gene-pair combinations that exhibit DC signals so that they can alleviate the computational intensiveness during the estimation process.

To control the extent of parameter shrinkage and dispersion, the values of the hyperparameters in the prior distributions can be modified accordingly. For example, as demonstrated in Figure 3.1, a larger value of v_0 can be set for less shrinkage in the situation with less sparsity.

In addition, we assumed Fisher's Z-transformation of the gene interactions to be linear with the cancer recurrence risk as presented in Equation (3.2). In practice, the association patterns could be more flexible. Using various transformations and non-linear associations can be easily adapted to the proposed modeling framework.

With the increasing availability of data generated by next-generation RNA sequencing (RNA-seq), it is of practical significance to discuss how to apply our proposed models to RNA-Seq read count data. While the models were designed for the gene normally distributed expression data, our proposed models are still applicable to count data after proper pre-processing. One can use the log normalized RNA-Seq read count data through adaptive technology such as Voom (Law et al., 2014) to ensure the transformed data correspond to a normal distribution. Moreover, developing a negative binomial model for systematically identifying DC is also a feasible option for modeling RNA-Seq read count data for future research.

Chapter 4

Subject-specific Model for Identifying Dynamic Gene Co-expression as Biomarkers for Classification using Bayesian Variable Selection for Single-Cell Count Data

4.1 Introduction

Advancements of next-generation sequencing technologies have enabled the generation of genetic data in individual cells and provided valuable insights into functions and behaviors of cells at various stages and cell types. In recent years, single-cell studies have been increasingly carried out to unveil the potential discoveries in genetic regulation mechanisms and interactions (Hwang et al., 2018). In a complex biological system, however, gene interaction in a genetic circuit are often highly dynamic and tightly coordinated by other modulating factors (Green et al., 2009; Qiu et al., 2011; Sambandan et al., 2006). In response to various external signals or changes in cellular states, interactions between genes can be activated or switched off. The dynamic change between gene interactions under various biological conditions is named dynamic gene coexpression (DC). Studying the dynamic changes using single-cell RNA sequencing (scRNA-seq) data can uncover the complex underlying genetic regulatory mechanisms in a high-resolution perspective.

Statistical methods for the identification of DC have been established and validated in past studies (Li, 2002; Li et al., 2004; Lai et al., 2004; Ho et al., 2011;

Gunderson and Ho, 2014; Yu, 2018; Kinzy et al., 2019). Li (2002) proposed a direct approach named liquid association to measure the DC between a pair of genes conditioning on a third gene using a three-product moment statistic. This method formulated the foundation of DC estimation and provided a new perspective on studying gene interactions and the underlying genetic regulatory mechanisms. Ho et al. (2011) extended the work and proposed a conditional normal model to consider interdependencies of mean, variance as well as the correlation in a gene triplet. To accelerate the process of identification of DC in high dimensional datasets, Gunderson and Ho (2014) developed a useful screening approach to reduce the number of gene triplets under consideration hence significantly reducing the computational time.

These methods mainly focused on continuous transcriptional data and assumed the gene expression levels to be Gaussian. However, scRNA-seq read count data are usually non-negative integers with a proportion of raw counts to be zero (Ma et al., 2020; Yang and Ho, 2021). The implementation of the aforementioned approaches require transforming count data into continuous measures and hence is unable to account for zero-inflation and over-dispersion in scRNA-seq data. To handle the over-dispersion and count-based marginal distributions for scRNA-seq data, Ma et al. (2020, 2022) established frameworks of using Poisson-Gamma mixture models to generate negative binomial distributions for count data and implementing copula methods to estimate DC in gene pairs. Yang and Ho (2021) proposed zero-inflated negative binomial dynamic correlation model (ZENCO) to incorporation zero-inflation in DC estimation with a Bernoulli-negative-binomial mixture model. These work provided ground-breaking opportunities to identify dynamic patterns of gene interactions for scRNA-seq count data across all cells.

In a typical scRNA-seq dataset, gene expression profiles are usually collected from multiple subjects such as different patients or tissues. The multi-subject scRNA-seq data are often characterized by individual difference and sparsity in gene co-

expressions (He et al., 2021). Considering subject-specific characterization can increase the accuracy of identifying DC gene pairs and provide great opportunity in using individual DC gene pairs as biomarkers in studying clinical outcomes. Motivated by this, we propose a mixed-effect model for scRNA-seq DC identification while incorporating subject-specific random effects and zero-inflation in scRNA-seq datasets.

Since the DC signal is often sparse among gene pairs, we also consider a Bayesian spike-and-slab variable selection approach in the mixed-effect framework to identify DC gene-pairs. The proposed mixed-effect spike-and-slab (ME-SPSL) model can focus on subsets of individuals with significant random effects while shrinking others to zero (George and McCulloch, 1993). Furthermore, we set out to use individual DC gene-pairs as biomarkers to classify subjects into different subgroups. In terms of the melanoma datasets (Jerby-Arnon et al., 2018) used in the experimental analysis in this article, the cells sampled from different patients were characterized by immunotheropy-resistant and non-resistant subgroups. The proposed methods can offer new insights into gene-pair biomarkers in clinical studies.

This article is organized in the following structure. Section 4.2 elaborates the proposed mixed-effect models for DC estimation using spike-and-slab variable selection method (ME-SPSL) and provides the description of the dataset used in this paper. Section 4.3 presents three simulation studies to compare the performance of the proposed model with the competing methods in identifying DC gene pairs and using the DC measurement as the biomarkers for subject classification. Section 4.4 describes an experimental data analysis using the proposed ME-SPSL model in a melanoma dataset for DC estimation and immunotheropy-resistant sample classification. At the end, Section 4.5 concludes the article and discusses the future work in the related area.

4.2 Materials and Methods

4.2.1 Datasets and Databases

The dataset used in this paper were obtained from an immunotherapy study for melanoma and can be accessed from Gene Expression Omnibus (GEO) with accession number GSE115978. This study aimed to identify cells with scRNA-seq profiles that were associated with immune checkpoint inhibitor (ICI) resistance. The dataset contains 7,186 cells from 33 melanoma tumors, among which 15 samples were ICI-resistant. The scRNA-seq data were collected using 10x Genomics Chromium platform (Jerby-Arnon et al., 2018) and the gene profiles were mapped to the UCSC hg19 human transcriptome using STAR (Dobin et al., 2013). The cells were identified based on the expression levels and inferred copy-number variation (CNV) profiles (Jerby-Arnon et al., 2018). After cells identification and gene profiles mapping, 10,483 genes and 6,173 melanoma cells were available in the experimental dataset.

4.2.2 Methods

Suppose scRNA-seq read count data are collected from K subjects and n_k cells per subject. Let Y_{ijk} represent the scRNA-seq count data for i-th gene in j-th cell from k-th subject, where $i = 1, 2$, $j = 1, ..., n_k$, and $k = 1, ..., K$. As suggested by Yang and Ho (2021); Ma et al. (2020), a negative binomial mixture can be used to model count data while account for zero-inflation. Considering the subject-specific random effects, the negative binomial mixture is formulated as follows:

$$Y_{ijk} \mid \gamma_{ijk}, \mu_{ik}, \phi_i \sim \gamma_{ijk}\mathcal{I}_0 + (1 - \gamma_{ijk})NB(\mu_{ik}, \phi_i), \tag{4.1}$$

where $\mathcal{I}_0$ is a point mass function at 0; and $\gamma_{ijk} \sim Bernoulli(p_{ik})$ is a binary dropout indicator with cell-invariant dropout probability of p_{ik}. As $\gamma_{ijk} = 1$, the expression level of gene $Y_{ijk} = 0$ and vice versa. The negative binomial component is assumed to have mean (μ_{ik}) for all cells and measurement of dispersion (ϕ_i) for all cells and all

subjects. To account for the subject-specific random effects on the mean, we assume $\mu_{ik} \sim N(\mu, \sigma_\mu^2)$, where μ is the fixed effect on the mean of the negative binomial distribution and σ_μ^2 is the variance. We set $\mu \sim N(0, 10^3)$ and $\sigma_\mu^2 \sim \mathcal{IG}(1, 0.001)$.

The dropout probability p_{ik} can be modeled as a logistic function of the subject-specific mean ($p_{ik} = \frac{e^{b_0+b_1\mu_{ik}}}{1+e^{b_0+b_1\mu_{ik}}}$) and estimated from the real dataset. Based on the melanoma dataset as described in Section 4.2.1, the parameters b_0 and b_1 can be estimated using a logistic regression.

Our goal of this research is to model the dynamic gene co-expression between Y_{1jk} and Y_{2jk} modulated by x_{jk}. Following Ma et al. (2020), we model the correlation between Y_{1jk} and Y_{2jk} by introducing latent bivariate normal variables and constructing the negative binomial distribution in Equation (4.1) using a poisson-gamma mixture. Let $\mathbf{z}_{jk} = (z_{1jk}, z_{2jk})^T$ be latent bivariate normal variables such that

$$\mathbf{z}_{jk} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{jk} \\ \rho_{jk} & 1 \end{bmatrix} \right), \tag{4.2}$$

where ρ_{jk} is the correlation coefficient between Y_{1jk} and Y_{2jk} in j-th cell and k-th subject; and $z_{ijk} \sim N(0, 1)$, marginally. Let $Y_{ijk} \mid u_{ijk}, \mu_{ik} \sim Poisson(\mu_{ik}u_{ijk})$ and $u_{ijk} \mid \phi_i \sim Gamma(1/\phi_i, 1/\phi_i)$. By integrating out u_{ijk} from the posterior distribution of $Y_{ijk}, u_{ijk} \mid \mu_{ik}, \phi_i$, the negative binomial distribution can be derived: $Y_{ijk} \mid \mu_{ik}, \phi_i \sim NB(\mu_{ik}, \phi_i)$ with $E(Y_{ijk}) = \mu_{ik}$ and $Var(Y_{ijk}) = \mu_{ik}(1 + \phi_i\mu_{ik})$. Furthermore, let $F(\cdot)$ be the cumulative distribution function of gamma distribution with same shape and scale parameter $1/\phi_i$. Therefore, the distribution of Y_{ijk} can be written as:

$$Y_{ijk} \mid \gamma_{ijk}, \mu_{ik}, \phi_i, z_{ijk} \sim \gamma_{ijk}\mathcal{I}_0 + (1 - \gamma_{ik})Poisson(F_{\alpha_i}^{-1}[\Phi(z_{ijk})]\mu_{ik}). \tag{4.3}$$

When $\gamma_{ijk} = 0$, Y_{ijk} follows a negative binomial distribution.

We define the dynamic gene co-expression between Y_{1jk} and Y_{2jk} changing with

the modulating factor x_{jk} as follows:

$$\log\left(\frac{1+\rho_{jk}}{1-\rho_{jk}}\right) = \tau_{0k} + \tau_{1k} \times x_{jk}, \tag{4.4}$$

where $\log\left(\frac{1+(\cdot)}{1-(\cdot)}\right)$ is the Fisher's Z-transformation to ensure that ρ_{jk} falls within (-1,1) and hence $\rho_{jk} = \frac{e^{(\tau_{0k}+\tau_{1k}\times x_{jk})}-1}{e^{(\tau_{0k}+\tau_{1k}\times x_{jk})}+1}$; τ_{0k} and τ_{1k} represent the measurement of baseline co-expression (BC) and the dynamic gene co-expression (DC). By assuming $\tau_{0k} \sim N(\tau_0, \sigma_0^2)$ and $\tau_{1k} \sim N(\tau_1, \sigma_1^2)$, where $\sigma_0^2 > 0$ and $\sigma_1^2 > 0$ are variances, the subject-specific random effect on the dynamic gene co-expression can be modeled. Here τ_0 and τ_1 are fixed effect of BC and DC across subjects and τ_1 is the parameter of interest.

Considering the subject-specific random effects on DC can be sparse, we propose a spike-and-slab random effect model (ME-SPSL) to apply variable selection to DC random effects. As suggested by George and McCulloch (1993); Ishwaran et al. (2005), we implement a spike-and-slab prior on τ_{1k} and formulate the proposed structure as follows,

$$\begin{aligned}
\mathbf{z}_{jk} \mid \tau_{0k}, \tau_{1k}, x_{jk} &\sim N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \frac{e^{(\tau_{0k}+\tau_{1k}\times x_{jk})}-1}{e^{(\tau_{0k}+\tau_{1k}\times x_{jk})}+1}\\ \frac{e^{(\tau_{0k}+\tau_{1k}\times x_{jk})}-1}{e^{(\tau_{0k}+\tau_{1k}\times x_{jk})}+1} & 1\end{bmatrix}\right)\\
\tau_{0k} \mid \tau_0, \sigma_0^2 &\sim N(\tau_0, \sigma_0^2)\\
\tau_{1k} \mid w_k &\sim w_k N(\tau_1, \mathscr{I}) + (1-w_k)\mathcal{I}_{\tau_1}\\
w_k \mid s_k &\sim Bernoulli(s_k)\\
s_k &\sim U(0,1).
\end{aligned} \tag{4.5}$$

In this framework, τ_{1k}'s prior is assumed to be a Gaussian mixture. The "slab" part is a normal distribution $N(\tau_1, \mathscr{I})$ with a mean of fixed effect τ_1 and hypervariance $\mathscr{I}$. The "spike" part is a point mass function centered at τ_1 to ensure the expectation of τ_{1k}'s prior distribution is τ_1. The shrinkage indicator w_k is a binary variable with 1 indicating normal estimation and 0 referring to τ_{1k}'s shrinking towards τ_1. In the proposed framework, we set non-informative conjugate priors to τ_0, τ_1 and $\mathscr{I}$ to

ensure that the parameters can be well estimated: $\tau_0 \sim N(0, 10^3)$, $\tau_1 \sim N(0, 10^3)$ and $\mathscr{I} \sim \mathcal{IG}(0.01, 0.01)$.

4.2.3 Estimation

Let θ represent the vector of all the parameters and latent variables including τ_{0k}'s, τ_{1k}'s, τ_0, τ_1, σ_0^2, μ_{1k}'s, μ_{2k}'s, μ, σ_μ^2, ϕ_1, ϕ_2, γ_{1k}'s, γ_{2k}'s, z_{1jk}'s, z_{2jk}'s, $\mathscr{I}$, w_k's and s_k's. The likelihood function of θ is given by

$$\mathcal{L}(\theta \mid \mathbf{Y}, \mathbf{x}) = \prod_{k=1}^{K} \prod_{j=1}^{n_k} f(Y_{1jk}, Y_{2jk} \mid x_{jk}, \theta),$$

where $f(\cdot)$ is the joint density function of $(Y_{1jk}, Y_{2jk})^T$. The posterior joint distribution for θ with prior joint distribution $\pi(\theta)$ can be written as

$$\pi(\theta \mid \mathbf{Y}, \mathbf{z}) \propto \prod_{k=1}^{K} \prod_{j=1}^{n_k} f(Y_{1jk}, Y_{2jk} \mid x_{jk}, \theta) \times \pi(\theta).$$

4.3 Simulation

To assess the performance of our proposed ME-SPSL model, three simulation studies were conducted to compare the proposed model with two other competing approaches, zero-inflated negative binomial dynamic correlation model Yang and Ho (2021) with random effects (ZENCO-ME) and ZENCO without random effects.

4.3.1 Scenario I

In the first simulation scenario, we evaluated the performance of our proposed ME-SPSL model and the competing methods in identifying DC from a single pair of zero-inflated scRNA-seq count data. We simulated $K = 30$ subjects, each with $n_k = 50$ cells. The simulated data contain a single pair of count data $\mathbf{Y}_{jk} = (Y_{1jk}, Y_{2jk})^T$ and a continuous modulating factor x_{jk} in j-th cell from k-th subject, where $j = 1, ..., n_k$ and $k = 1, ..., K$. The data were generated in the following procedure.

First, we sampled subject-specific measurements of DC. For k-th subject, τ_{0k} was sampled from $N(\tau_0, \sigma_0^2)$ and τ_{1k} was sampled from $N(\tau_1, \sigma_{1k}^2)$, where τ_1 is the fixed effect on DC across all subjects. The variance of random effects σ_1^2 was set equal to 10^{-4} and 10^{-2}. We used the same simulation procedure to simulate $n_k = 50$ cells for each subject. A set of $\{x_{jk}\}_{j=1}^{n_k}$ from $N(0,1)$ was generated as the modulating factor.

Next, the correlation coefficient was calculated, $\rho_{jk} = \frac{e^{(\tau_{0k}+\tau_{1k}x_{jk})}-1}{e^{(\tau_{0k}+\tau_{1k}x_{jk})}+1}$. Then the latent variables $\mathbf{z}_{jk} = (z_{1jk}, z_{2jk})^T$ were generated such that $\mathbf{z}_{jk} \sim N_2\Big(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho_{jk}\\ \rho_{jk} & 1\end{bmatrix}\Big)$. The count data $\mathbf{Y}_{jk}$ were simulated using $\mathbf{z}_{jk}$ as discussed in Equation (4.3). The dropout indicators γ_{1jk} and γ_{2jk} were generated from Bernoulli distributions, where the dropout probabilities p_{1k} and p_{2k} were estimated from the real dataset as discussed in Section 4.2.2.

In this scenario, we evaluated the performance of the proposed model to identify the measurement of the fixed effect of dynamic gene co-expression (τ_1) while considering the subject-specific random effects in a single pair of genes. Power analysis was used to compare the model with two other competing approaches: ZENCO-ME and ZENCO. For testing the existence of dynamic co-expression changes, our hypotheses were set up as:

$$H_0 : \tau_1 = 0 \text{ versus } H_1 : \tau_1 \neq 0.$$

We set $b_0 = -0.698$, $b_1 = -0.076$ which were estimated using the real dataset. The other parameters were set as $\mu_1 = \mu_2 = 30$, $\sigma_\mu = 1$, $\phi_1 = \phi_2 = 2$, $\sigma_0^2 = 10^{-4}$. The true value of τ_1 was set ranging from 0 to 0.5 while τ_0 was set equal to 0. The statistical power can be calculated as the percentage of the posterior 95% credible intervals of τ_1 not including zero in the 200 simulations.

4.3.2 Scenario II

An additional simulation was conducted to evaluate the performance of the proposed model and competing methods when the variances of τ_{1k} are random. The variance of DC for the k-th subject σ_1^2 was chosen at randomly from the set of $(10^{-4}, 10^{-2}, 1)$, resulting in some τ_{1k}'s widely spread and some highly concentrating around τ_1. While maintaining all parameters the same as in Section 4.3.1, we implemented power analysis to evaluate the performance of identifying τ_1.

4.3.3 Scenario III

In this scenario, we evaluated the performance of our proposed ME-SPSL model compared to ZENCO-ME in subjects classification using subject-specific DC as the biomarkers. Following the data generation procedure in Section 4.3.1, we simulated 4 pairs of count data $\{\mathbf{Y}_{jk}^{(l)}\}_{l=1}^{3}$ based on the same x_{jk} and a set of $\tau_1 = (\tau_1^{(1)}, \tau_1^{(2)}, \tau_1^{(3)}, \tau_1^{(4)}) = (0, 0.2, 0.3, 0.5)$ in j-th cell from k-th subject. For l-th pair, we generated $\tau_{1k}^{(l)} \sim N(\tau_1^{(l)}, \sigma_{1k}^2)$ with corresponding $\tau_1^{(l)}$, $l = 1, 2, 3, 4$, while σ_{1k}^2 was randomly chosen from the set of $(10^{-4}, 10^{-2}, 1)$ as described in Section 4.3.1. Other parameters including the number of cells and subjects were all set the same as in Section 4.3.1. Next, the binary class for k-th subject was generated in terms of a Bernoulli distribution with probability p_k^*, which was calculated based on the logistic function $p_k^* = \frac{e^{\beta_0+\beta\tau_{1k}}}{1+e^{\beta_0+\beta\tau_{1k}}}$. The parameters β were set equal to (1, 1, 1, -1) with the intercept β_0 being equal to 0 so that the number of classes were not imbalanced.

In ME-SPSL and ZENCO-ME models, τ_{1k} were estimated first from all combinations of $\mathbf{Y}_{jk}$ and then used as biomarkers to classify subjects into two groups using logistic regression. Pairs of genes were selected based on the significance of τ_1. Receiver operating characteristics (ROC) curve and area under curve (AUC) score were used to evaluate the performance in classification.

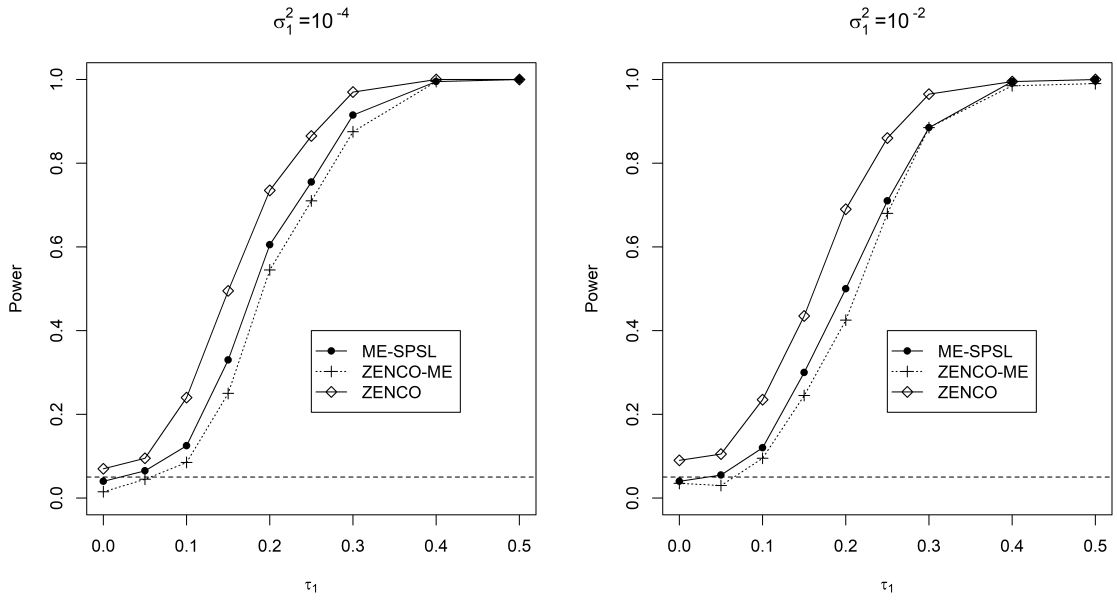


Figure 4.1. Power curves for τ_1 for ME-SPSL, ZENCO-ME, and ZENCO model based on 200 simulations and fixed σ_1^2.

4.3.4 Results

All methods were implemented through automated Markov Chain Monte Carlo (MCMC) compiler JAGS (Plummer et al., 2003) with 10,000 iterations, 2,000 burn-ins, and 3 chains. The results of Gelman-Rubin convergence diagnostics (Gelman et al., 1992) are provided in Table 4.1. Gelman-Rubin convergence diagnostics statistics were all less than 1.2, which indicates good convergence (Brooks and Gelman, 1998). The calculation was implemented on computing clusters nodes each with the Intel(R) Xeon(R) CPU @ 2.40GHz processor.

Figure 4.1 presents the power curves for scenario I (Section 4.3.1) with fixed σ_1^2. The ME-SPSL is more powerful than ZENCO-ME when σ_1^2 is small. As shown in Figure 4.2, ZENCO fails to control the type I error in all cases. Our proposed ME-SPSL can successfully maintain type I error and achieve greater power than ZENCO-ME.

Figure 4.2 presents the power plots for the proposed ME-SPSL model and the

Table 4.1. Gelman-Rubin convergence diagnostics results for a randomly picked sample among 200 simulations. Point estimates of the potential scale reduction factor (Point est.) and their upper confidence limits (Upper C.I.) are provided.

	Point est.	Upper C.I.		Point est.	Upper C.I.
σ_0^2	1.01	1.01	σ_1^2	1.00	1.01
τ_0	1.00	1.00	τ_1	1.00	1.00
$\tau_{0,1}$	1.00	1.00	$\tau_{1,1}$	1.00	1.00
$\tau_{0,2}$	1.00	1.00	$\tau_{1,2}$	1.00	1.01
$\tau_{0,3}$	1.00	1.01	$\tau_{1,3}$	1.00	1.01
$\tau_{0,4}$	1.00	1.00	$\tau_{1,4}$	1.00	1.00
$\tau_{0,5}$	1.00	1.01	$\tau_{1,5}$	1.00	1.00
$\tau_{0,6}$	1.00	1.00	$\tau_{1,6}$	1.00	1.00
$\tau_{0,7}$	1.00	1.00	$\tau_{1,7}$	1.01	1.03
$\tau_{0,8}$	1.00	1.00	$\tau_{1,8}$	1.00	1.00
$\tau_{0,9}$	1.00	1.00	$\tau_{1,9}$	1.00	1.00
$\tau_{0,10}$	1.00	1.00	$\tau_{1,10}$	1.01	1.02
$\tau_{0,11}$	1.00	1.00	$\tau_{1,11}$	1.00	1.01
$\tau_{0,12}$	1.00	1.00	$\tau_{1,12}$	1.00	1.01
$\tau_{0,13}$	1.00	1.00	$\tau_{1,13}$	1.01	1.02
$\tau_{0,14}$	1.00	1.00	$\tau_{1,14}$	1.00	1.00
$\tau_{0,15}$	1.00	1.00	$\tau_{1,15}$	1.00	1.00
$\tau_{0,16}$	1.00	1.00	$\tau_{1,16}$	1.00	1.00
$\tau_{0,17}$	1.00	1.00	$\tau_{1,17}$	1.00	1.00
$\tau_{0,18}$	1.00	1.00	$\tau_{1,18}$	1.00	1.00
$\tau_{0,19}$	1.00	1.00	$\tau_{1,19}$	1.00	1.01
$\tau_{0,20}$	1.00	1.00	$\tau_{1,20}$	1.00	1.00
$\tau_{0,21}$	1.00	1.00	$\tau_{1,21}$	1.01	1.01
$\tau_{0,22}$	1.00	1.01	$\tau_{1,22}$	1.00	1.01
$\tau_{0,23}$	1.00	1.00	$\tau_{1,23}$	1.00	1.01
$\tau_{0,24}$	1.00	1.00	$\tau_{1,24}$	1.00	1.00
$\tau_{0,25}$	1.00	1.01	$\tau_{1,25}$	1.00	1.01
$\tau_{0,26}$	1.00	1.00	$\tau_{1,26}$	1.00	1.00
$\tau_{0,27}$	1.00	1.01	$\tau_{1,27}$	1.00	1.02
$\tau_{0,28}$	1.00	1.01	$\tau_{1,28}$	1.00	1.01
$\tau_{0,29}$	1.00	1.01	$\tau_{1,29}$	1.01	1.01
$\tau_{0,30}$	1.00	1.00	$\tau_{1,30}$	1.00	1.01

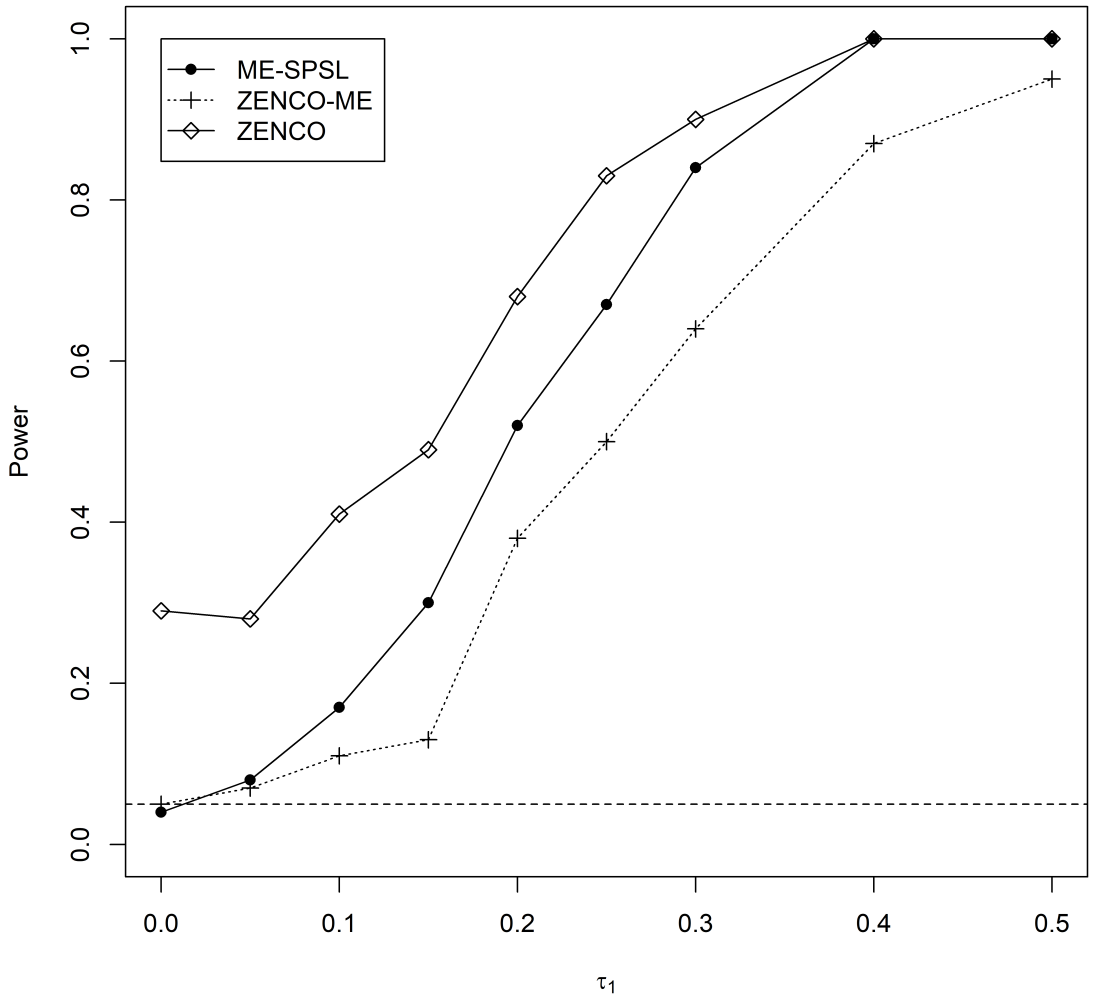


Figure 4.2. Power curves for τ_1 for ME-SPSL, ZENCO-ME and ZENCO model based on 200 simulations and random σ_1^2.

competing methods: ZENCO-ME and ZENCO models over 200 simulations for scenario II as described in Section 4.3.2. As shown in Figure 4.2, ME-SPSL outperforms ZENCO-ME while ZENCO fails to control the type I error. Since ZENCO estimates τ_1 without considering the subject-specific random effects, it tends to overestimate τ_1 when the values of τ_1 are near 0. The implementation of variable selection can effectively identify the fixed effect τ_1 while shrinking the small random effects to 0 to increase the power of identification of τ_1.

Table 4.2 lists the coverage probability of the 95% credible interval (CP), the length of the 95% credible interval (CI length), mean square errors (MSE) and mean bias errors (MBE). The MSE and MBE evaluate the deviations between the posterior mean of τ_1 from the MCMC samples and the true values. The MSE and MBE are calculated as $MSE = \frac{1}{n_s}(\hat{\tau}_1 - \tau_1)^2$ and $MBE = \frac{1}{n_s}(\hat{\tau}_1 - \tau_1)$, where n_s is the number of simulations; $\hat{\tau}_1$ is the posterior mean of τ_1 estimates from the MCMC samples. As presented in Table 4.2, CPs of ME-SPSL and ZENCO-ME are all around 0.95 and ME-SPSL has narrower credible intervals than ZENCO-ME.

Table 4.2. Coverage probability (CP) and the length of 95% credible interval (CI length), mean square errors (MSE), mean bias errors (MBE) are used as metrics to evaluate the performance of estimating τ_1 ranging from 0 to 0.5 for ME-SPSL, ZENCO-ME and ZENCO models based on 200 simulations.

		τ_1								
Model	Metric	0	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
ME-SPSL	CP	0.93	0.95	0.97	0.92	0.96	0.95	0.94	0.97	0.97
	CI length	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
	MSE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	MBE	0.00	0.00	0.00	-0.01	-0.01	-0.00	-0.01	-0.01	0.00
ZENCO-ME	CP	0.90	0.93	0.98	0.94	0.95	0.96	0.94	0.97	0.94
	CI length	0.49	0.48	0.48	0.49	0.48	0.49	0.49	0.48	0.49
	MSE	0.02	0.02	0.01	0.02	0.01	0.02	0.01	0.01	0.02
	MBE	0.00	0.01	0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.01
ZENCO	CP	0.72	0.77	0.80	0.76	0.78	0.78	0.76	0.81	0.78
	CI length	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
	MSE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	MBE	0.00	0.00	0.00	-0.02	-0.02	-0.02	-0.03	-0.03	-0.03

Figure 4.3 presents the results of subgroup classification for scenario III as described in Section 4.3.3, which compares the classification results between ME-SPSL and ZENCO-ME using subject-specific DC estimates as the biomarkers. Samples were randomly separated into a traning set (20 samples) and a validation set (10 samples). The average ROC curves for validation sets over the 200 simulations are presented in bold. The ME-SPSL outperforms the ZENCO-ME since the average AUC score of both the training sets and the validation sets of ME-SPSL (0.81/0.72)

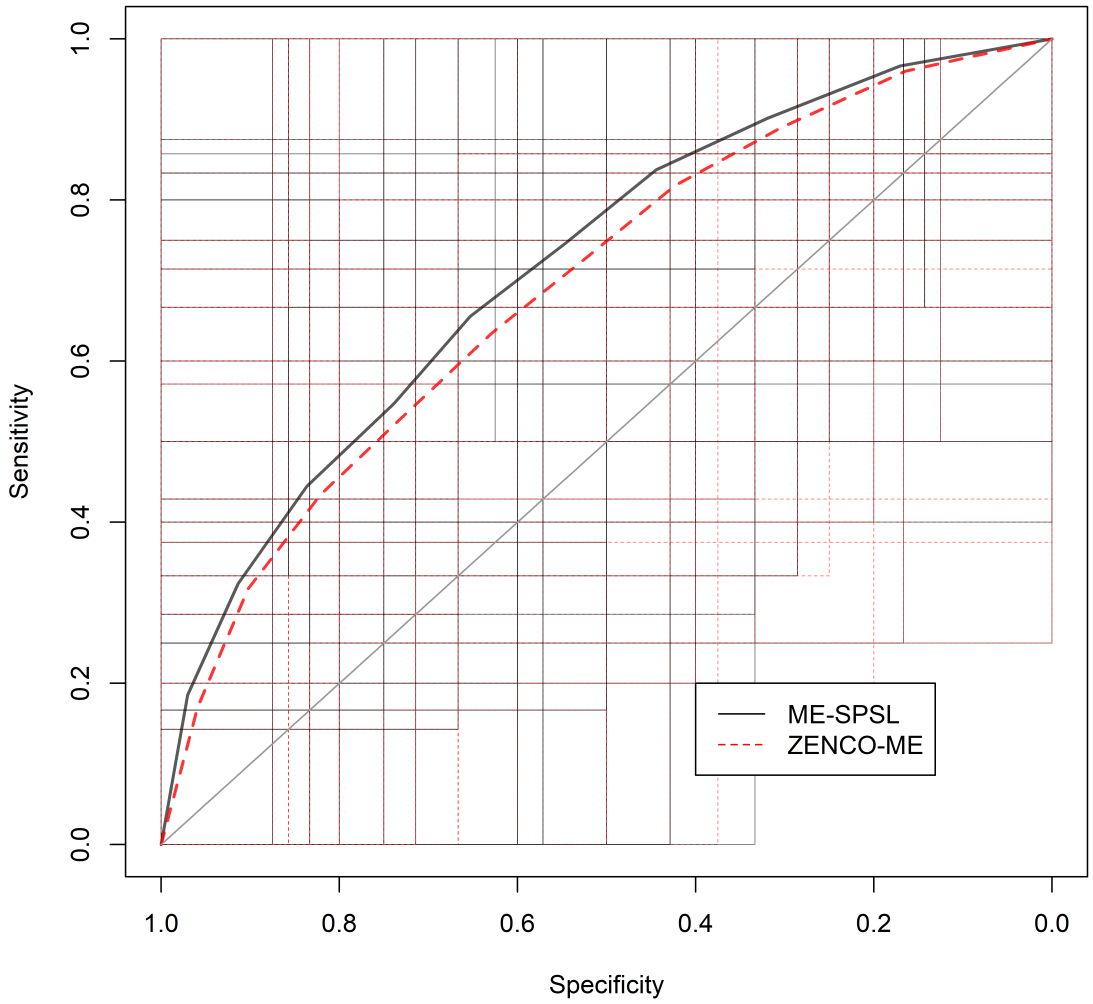


Figure 4.3. ROC curves for subgroup classification. Black lines represent the average ROC curve for ME-SPSL method and red dashed lines represent the ZENCO-ME method.

is higher than ZENCO-ME (0.78/0.70).

4.4 Experimental Data Analysis

In this section, we demonstrated the implementation of our proposed ME-SPSL model and DC-based classification strategy with an experimental analysis using the melanoma dataset as discussed in Section 4.2.1. This dataset contains 10,483 genes,

6,173 melanoma cells from 32 different melanoma samples. We considered genes in the KEGG melanoma pathway with Entry ID of hsa05218 (Kanehisa and Goto, 2000). After screening out genes with more than 80% zero counts, 32 genes remained, resulting in 496 gene-pair combinations.

In this study, our goal was to identify gene pairs with significant non-zero τ_1 associated with the overall expression (OE) score of the gene sets and use the subject-specific τ_{1k}'s, where k refers to the k-th melanoma sample ($k = 1, ..., 32$), in the corresponding significant gene pairs as the biomarkers to classify the melanoma samples into ICI-resistant group and non-resistant group. The cell-based OE score was used as the modulating factor x, which is the measurement for the average expression of genes across all cells and the subjects in a single-cell genomic dataset (Jerby-Arnon et al., 2018). After standardizing x with the mean and the standard deviation, the proposed ME-SPSL was implemented on the pairwise gene combinations to estimate τ_1 and τ_{1k}'s.

We implemented the ME-SPSL model using MCMC compiler JAGS with 20,000 iterations, 10,000 burn-ins and 3 chains. As suggested by Wang and Dunson (2010), the Bayesian multiple hypothesis testing procedure was applied to control the false discovery rate (FDR) of identifying DC gene pairs. For the j-th gene pair, the local hypothesis was defined as, $H_0^{(j)} : |\tau_1^{(j)}| < \epsilon$ vs. $H_a^{(j)} : |\tau_1^{(j)}| \geq \epsilon$. The threshold ϵ was determined based on the distributions of $\widehat{\tau}_1$ (the posterior mean of τ_1). As shown in Figure 4.4 which presents the comparison between the distributions of $\widehat{\tau}_1$ that the 95% credible intervals of τ_1 (95% CIs) from the MCMC posterior samples include or exclude zero, $\epsilon = 0.16$ is the minimum $\widehat{\tau}_1$ that the 95% CIs exclude zero. By controlling the FDR at 0.05 level, 140 gene pairs were identified as significant.

Table 4.3 lists the top 30 gene pairs with the largest $|\widehat{\tau}_1|$ within the 140 significant gene pairs. Gene *IGF1R* is included in all the top three pairs. *IGF1R* is a well-known growth factor I receptor that regulates cells' growth and apoptotic events and was

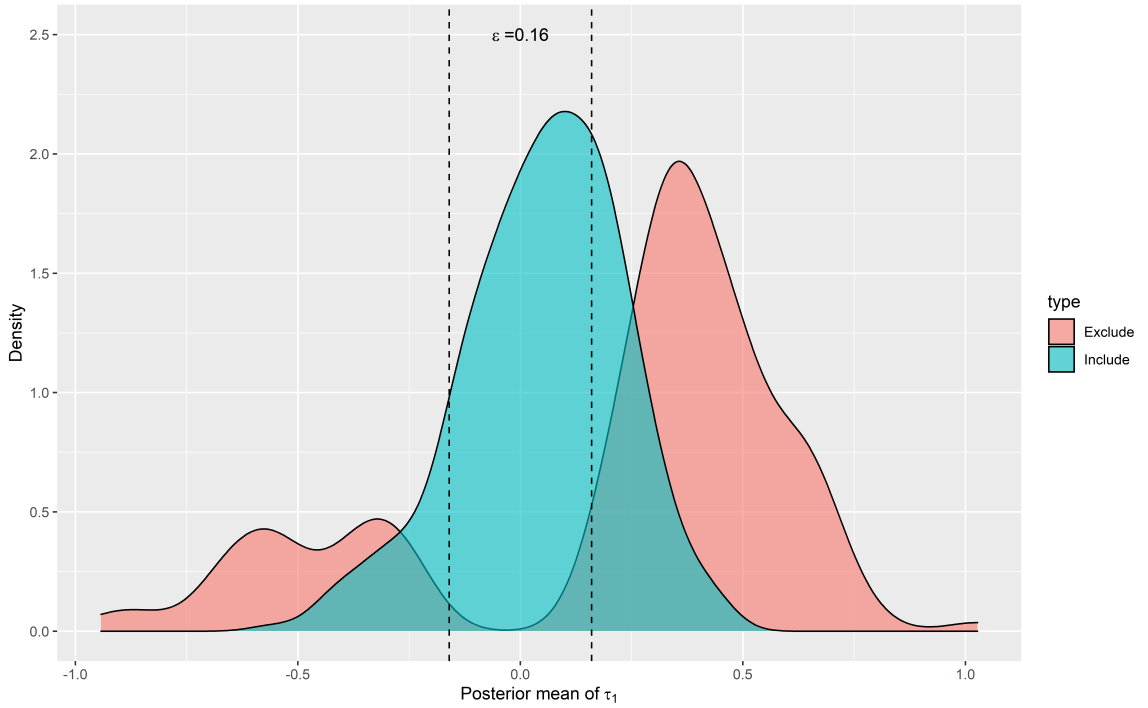


Figure 4.4. Density plots of the posterior mean ($\widehat{\tau}_1$) for the groups including or excluding zero from the 95% credible intervals (95% CI) of $\widehat{\tau}_1$. The value of ϵ denotes the minimum of $|\widehat{\tau}_1|$ in the significant group that 95% CI excludes 0.

found to play an important role in the treatment of melanoma (Karasic et al., 2010).

We then use the subject-specific τ_{1k}'s in the significant gene-pairs as the biomarkers (DC-based strategy) and in logistic regression to classify the melanoma samples into ICI-resistant and non-resistant groups. An expression-based strategy was considered as the benchmark method, which considers the subject-specific mean expression of genes as the biomarkers, to demonstrate the performance of DC-based strategy in classification. We also implemented principle component analysis to reduce the dimension of the biomarkers to avoid singularity issues in logistic regression procedure. The optimal number of principal components was determined as 6 for both strategies based on wold-style cross validation (Owen and Perry, 2009). Ten-fold Repeated random sub-sampling validation (Monte Carlo cross-validation) (Dubitzky et al., 2007) was used to split the dataset into ten training (20 samples) and validation sets (12 samples). The predictive accuracy was assessed using average ROC curve and AUC

Table 4.3. The top 30 significant gene pairs with the largest posterior mean $|\widehat{\tau}_1|$ associated with OE score. The 95% posterior credible intervals of $|\widehat{\tau}_1|$ (95% CI) are provided.

Rank	Gene I	Gene II	$\widehat{\tau}_1$	95% CI
1	IGF1R	BAK1	1. 027	(0.721, 1.325)
2	AKT2	IGF1R	-0. 942	(-1.528,-0.319)
3	PIK3CD	IGF1R	-0. 882	(-1.202,-0.456)
4	MET	CDH1	-0. 828	(-1.147,-0.551)
5	AKT3	ARAF	0. 782	(0.381, 1.125)
6	MAPK1	DDB2	0. 732	(0.300, 1.168)
7	AKT3	BAK1	0. 699	(0.365, 1.031)
8	AKT2	FGFR1	-0. 694	(-1.030,-0.420)
9	RAF1	CDK4	0. 694	(0.460, 0.912)
10	PIK3R1	GADD45B	0. 691	(0.389, 1.032)
11	AKT2	IGF1	0. 679	(0.252, 0.969)
12	IGF1R	DDB2	0. 678	(0.128, 1.137)
13	CDK4	BAK1	0. 675	(0.406, 0.912)
14	PIK3CD	CDK4	-0. 675	(-0.887,-0.428)
15	AKT2	MITF	-0. 670	(-1.067,-0.333)
16	PIK3CD	AKT3	-0. 664	(-0.972,-0.314)
17	E2F3	MITF	0. 654	(0.180, 1.161)
18	AKT2	GADD45B	0. 650	(0.189, 1.016)
19	IGF1R	MDM2	-0. 647	(-1.135,-0.241)
20	RAF1	E2F3	0. 641	(0.289, 1.094)
21	BAX	RAF1	0. 640	(0.337, 1.016)
22	CDK4	DDB2	0. 635	(0.314, 0.961)
23	CDK4	E2F3	0. 633	(0.373, 0.968)
24	PIK3R1	POLK	0. 632	(0.282, 0.988)
25	FGFR1	GADD45B	0. 631	(0.288, 1.016)
26	PIK3R1	MITF	-0. 624	(-0.909,-0.340)
27	AKT2	RB1	0. 622	(0.319, 0.918)
28	MET	E2F3	0. 612	(0.206, 0.963)
29	IGF1	MET	0. 609	(0.008, 1.078)
30	MET	GADD45B	0..601	(0.266, 0.894)

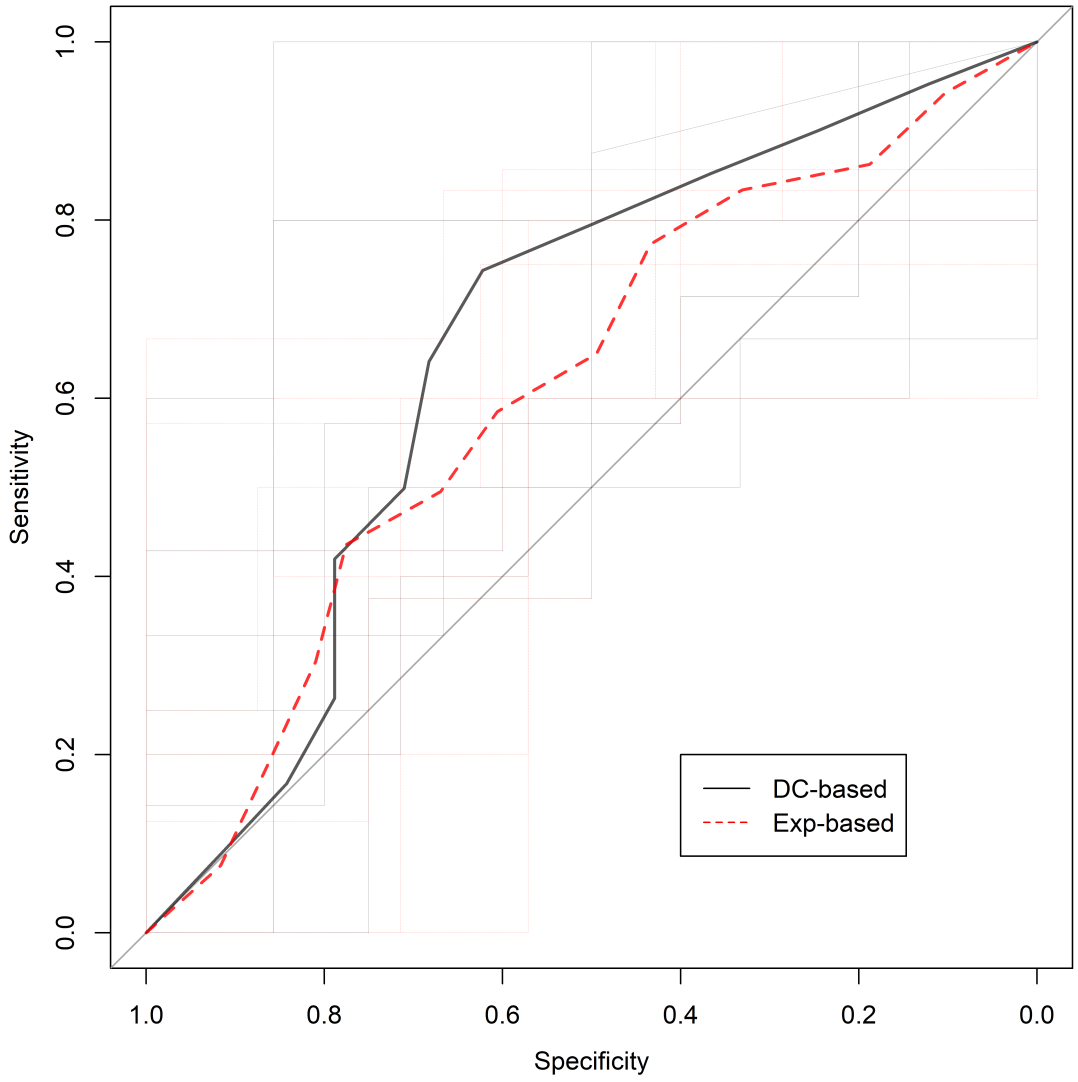


Figure 4.5. Average ROC curves for validation sets of ICI-resistant classification. The average AUC score of DC-based method (black solid line) is 0.65 and the average AUC score of the expression-based method (red dashed line) is 0.61.

score through the multiple validation sets.

Figure 4.5 presents the average ROC curves for the ten validation sets over the 10-fold Monte Carlo cross-validation. The average AUC scores of the training/validation sets of the DC-based strategy are 0.81 and 0.65 and that of the expression-based strategy are 0.80 and 0.61. The DC-based strategy provides a preferable tool for clinical outcome prediction and classification compared to the expression-based strategy.

4.5 Discussion

In this paper, we proposed a mixed-effect model to study the subject-specific DCs for zero-inflated and over-dispersed scRNA-seq count data to account for both the random effect of each subject and the fixed effect across all subjects on DC. Considering the sparsity of subject-specific DC signals, we included a spike-and-slab method as the variable selection technique in the mixed-effect model to exploit the sparsity to achieve better performance. Furthermore, we proposed to use subject-specific DC gene pairs as the biomarkers in subgroup classification. In the proposed DC-based classification strategy, gene-pair biomarkers are identified first in terms of significant fixed-effect of DC. The subject-specific DCs are then estimated and used as the inputs of the logistic regression to implement subgroup classification.

The advantages of the proposed method in studying DC in scRNA-seq datasets are multifold. First, compared with the existing method, the proposed method can estimate both the random effects and the fixed effects while account for the zero-inflation and over-dispersion in count data. The proposed method can avoid overestimating the DC when random effects exist and hence perfectly control the type I error. Second, the ME-SPSL method can exploit the sparsity of DC signals to boost the statistical power in identifying DC, especially when subject-specific random effects have different variances. Third, the proposed subject-specific model provides an exciting opportunity to implement DC-based classification strategy to use DC gene pairs as biomarkers in clinical subgroups classification. Compared to the expression-based strategy which directly uses the subject-specific mean expression of each gene as the biomarkers in classification, the DC-based strategy can achieve better performance. Finally, The DC-based classification strategy provides an innovative perspective in disease-related biomarkers identification and clinical outcomes predictive studies.

We applied our proposed method to a melanoma scRNA-seq dataset. The overall expression for each cell in the entire dataset was used as the modulating factor.

The OE score is continuous and normally distributed. In practice, our proposed method can be used to model DC associated with data from other distributions including count data. In the experimental analysis, an exhaustive search strategy was used to scan over all gene-pair combinations to estimate DC and required screening procedure to reduce the computational intensiveness. Furthermore, we used the DC gene pairs as the biomarkers to classify the melanoma samples into ICI-resistant and non-resistant groups. Through Monte-Carlo cross-validation, we showed that the DC-based strategy outperformed expression-based strategy in subgroup classification.

In this paper, we focused on identifying DC in a single gene pair. A possible future work is to incorporate the dependence structure among DCs of different gene pairs to extend the proposed model to higher dimensions. In addition, our proposed model suggests a negative binomial distribution to model the count data. The implementation of copula methods can incorporate more feasible distributions in DC estimation. Moreover, in subgroup classification, considering other classification methods including machine learning classification techniques, deep learning models is a feasible area of future work.

Bibliography

Agha, G., Hajj, H., Rifas-Shiman, S. L., Just, A. C., Hivert, M.-F., Burris, H. H., Lin, X., Litonjua, A. A., Oken, E., DeMeo, D. L., et al. (2016). Birth weight-for-gestational age is associated with DNA methylation at birth and in childhood. *Clinical Epigenetics*, 8(1):1–12.

Baek, S., Ho, Y.-Y., and Ma, Y. (2020). Using sufficient direction factor model to analyze latent activities associated with breast cancer survival. *Biometrics*, 76(4):1340–1350.

Batista, G. E., Monard, M. C., et al. (2002). A study of k-nearest neighbour as an imputation method. *His*, 87(251-260):48.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *BMJ*, 310(6973):170.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Chou, C.-K., Fan, C.-C., Lin, P.-S., Liao, P.-Y., Tung, J.-C., Hsieh, C.-H., Hung, M.-C., Chen, C.-H., and Chang, W.-C. (2016). Sciellin mediates mesenchymal-to-epithelial transition in colorectal cancer hepatic metastasis. *Oncotarget*, 7(18):25742.

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5(7):613–619.

de Silva, H. M. and Perera, A. S. (2017). Evolutionary k-nearest neighbor imputation algorithm for gene expression data. *ICTer*, 10(1).

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Dubitzky, W., Granzow, M., and Berrar, D. P. (2007). *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media.

Fisher, R. A. et al. (1921). 014: On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

Fridley, B. L., Lund, S., Jenkins, G. D., and Wang, L. (2012). A Bayesian integrative genomic model for pathway analysis of complex traits. *Genetic Epidemiology*, 36(4):352–359.

Gelman, A., Rubin, D. B., et al. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical science*, 7(4):457–472.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

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

Green, L. M., Wagner, K. J., Campbell, H. A., Addison, K., and Roberts, S. G. (2009). Dynamic interaction between WT1 and BASP1 in transcriptional regulation during differentiation. *Nucleic Acids Research*, 37(2):431–440.

Guillermo, R., Elena, V., Martin, K., and Chris, W. (2021). Rápidopgs: A rapid polygenic score calculator for summary gwas data without a test dataset. *Bioinformatics*.

Gunderson, T. and Ho, Y.-Y. (2014). An efficient algorithm to explore liquid association on a genome-wide scale. *BMC Bioinformatics*, 15(1):371.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Habiger, J. D. (2017). Adaptive false discovery rate control for heterogeneous data. *Statistica Sinica*, pages 1731–1756.

Hastie T, Tibshirani R, N. B. and G, C. (2021). *impute: impute: Imputation for microarray data*. `https://bioconductor.org/packages/impute/` (accessed December 2nd).

He, L., Davila-Velderrain, J., Sumida, T. S., Hafler, D. A., Kellis, M., and Kulminski, A. M. (2021). Nebula is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Communications biology*, 4(1):1–17.

Ho, Y.-Y., Parmigiani, G., Louis, T. A., and Cope, L. M. (2011). Modeling liquid association. *Biometrics*, 67(1):133–141.

Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14.

Ignatiadis, N. and Huber, W. (2021). Covariate powered cross-weighted multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):720–751.

Ishwaran, H., Rao, J. S., et al. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Jerby-Arnon, L., Shah, P., Cuoco, M. S., Rodman, C., Su, M.-J., Melms, J. C., Leeson, R., Kanodia, A., Mei, S., Lin, J.-R., et al. (2018). A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell*, 175(4):984–997.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Karasic, T. B., Hei, T. K., and Ivanov, V. N. (2010). Disruption of IGF-1R signaling increases TRAIL-induced apoptosis: a new potential therapy for the treatment of melanoma. *Experimental cell research*, 316(12):1994–2007.

Kashima, K., Kawai, T., Nishimura, R., Shiwa, Y., Urayama, K. Y., Kamura, H., Takeda, K., Aoto, S., Ito, A., Matsubara, K., et al. (2021). Identification of epigenetic memory candidates associated with gestational age at birth through analysis of methylome and transcriptional data. *Scientific Reports*, 11(1):1–16.

Kinzy, T. G., Starr, T. K., Tseng, G. C., and Ho, Y.-Y. (2019). Meta-analytic framework for modeling genetic coexpression dynamics. *Statistical Applications in Genetics and Molecular Biology*, 18(1).

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297.

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313.

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

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

Lee, K. H., Tadesse, M. G., Baccarelli, A. A., Schwartz, J., and Coull, B. A. (2017). Multivariate Bayesian variable selection exploiting dependence structure among outcomes: Application to air pollution effects on DNA methylation. *Biometrics*, 73(1):232–241.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161.

Li, K.-C. (2002). Genome-wide coexpression dynamics: Theory and application. *Proceedings of the National Academy of Sciences*, 99(26):16875–16880.

Li, K.-C., Liu, C.-T., Sun, W., Yuan, S., and Yu, T. (2004). A system for enhancing genome-wide coexpression dynamics study. *Proceedings of the National Academy of Sciences*, 101(44):15561–15566.

Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., Lin, X., and Liang, L. (2013). Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in Genetics*, 4:103.

Lin, D., Zhang, J., Li, J., Xu, C., Deng, H.-W., and Wang, Y.-P. (2016). An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics*, 17(1):1–12.

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3):410–421.

Liu, Y. and Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402.

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

Ma, Z., Davis, S. W., and Ho, Y.-Y. (2022). Flexible copula model for integrating correlated multi-omics data from single-cell experiments. *Biometrics*.

Ma, Z., Hanson, T. E., and Ho, Y.-Y. (2020). Flexible bivariate correlated count data regression. *Statistics in Medicine*, 39(25):3476–3490.

Mallmann, M. R., Staratschek-Jox, A., Rudlowski, C., Braun, M., Gaarz, A., Wolfgarten, M., Kuhn, W., and Schultze, J. L. (2010). Prediction and prognosis: Impact of gene expression profiling in personalized treatment of breast cancer patients. *EPMA Journal*, 1(3):421–437.

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

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays. *Journal of the American Statistical Association*, 99(468):990–1001.

Nadarajah, S. and Pogány, T. K. (2016). On the distribution of the product of correlated normal random variables. *Comptes Rendus Mathematique*, 354(2):201–204.

Oken, E., Kleinman, K. P., Rich-Edwards, J., and Gillman, M. W. (2003). A nearly continuous measure of birth weight for gestational age using a United States national reference. *BMC Pediatrics*, 3(1):1–10.

Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the SVD and the non-negative matrix factorization. *The Annals of Applied Statistics*, 3(2):564–594.

Pan, F., Chen, T., Sun, X., Li, K., Jiang, X., Försti, A., Zhu, Y., and Lai, M. (2019). Prognosis prediction of colorectal cancer using gene expression profiles. *Frontiers in Oncology*, 9:252.

Peters, F. and Becker-Pauly, C. (2019). Role of meprin metalloproteases in metastasis and tumor microenvironment. *Cancer and Metastasis Reviews*, 38(3):347–356.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9(501-538):105.

Qiu, Y., Stavreva, D. A., Luo, Y., Indrawan, A., Chang, M., and Hager, G. L. (2011). Dynamic interaction of HDAC1 with a glucocorticoid receptor-regulated gene is modulated by the activity state of the promoter. *Journal of Biological Chemistry*, 286(9):7641–7647.

Ramaswami, G., Won, H., Gandal, M. J., Haney, J., Wang, J. C., Wong, C. C., Sun, W., Prabhakar, S., Mill, J., and Geschwind, D. H. (2020). Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism. *Nature Communications*, 11(1):1–14.

Ratolojanahary, R., Ngouna, R. H., Medjaher, K., Junca-Bourié, J., Dauriac, F., and Sebilo, M. (2019). Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. *Expert Systems with Applications*, 131:299–307.

Roeder, K., Devlin, B., and Wasserman, L. (2007). Improving power in genome-wide association studies: Weights tip the scale. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 31(7):741–747.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 24(4):398.

Roquain, E. and Van De Wiel, M. A. (2009). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics*, 3:678–711.

Rubin, D., Dudoit, S., and Van der Laan, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology*, 5(1).

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Sambandan, D., Yamamoto, A., Fanara, J.-J., Mackay, T. F., and Anholt, R. R. (2006). Dynamic genetic interactions determine odor-guided behavior in drosophila melanogaster. *Genetics*, 174(3):1349–1363.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology*, 179(6):764–774.

Shi, H., Zhang, L., Qu, Y., Hou, L., Wang, L., and Zheng, M. (2017). Prognostic genes of breast cancer revealed by gene co-expression network analysis. *Oncology Letters*, 14(4):4535–4542.

Shin, M. and Liu, J. S. (2021). Neuronized Priors for Bayesian Sparse Linear Regression. *Journal of the American Statistical Association*, pages 1–43.

Smith, J. J., Deane, N. G., Wu, F., Merchant, N. B., Zhang, B., Jiang, A., Lu, P., Johnson, J. C., Schmidt, C., Bailey, C. E., et al. (2010). Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*, 138(3):958–968.

Song, M., Greenbaum, J., Luttrell IV, J., Zhou, W., Wu, C., Shen, H., Gong, P., Zhang, C., and Deng, H.-W. (2020). A review of integrative imputation for multi-omics datasets. *Frontiers in Genetics*, 11.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

The survival package (2021). The Comprehensive R Archive Network. `https://cran.r-project.org/web/packages/survival/`. (accessed May 13, 2021).

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(1):1–67.

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., and Jolani, S. (2021). *package 'mice'*. `https://github.com/amices/mice` (accessed December 2nd).

Van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50.

van Iterson, M., van Zwet, E. W., and Heijmans, B. T. (2017). Controlling bias and inflation in epigenome-and transcriptome-wide association studies using the empirical null distribution. *Genome biology*, 18(1):1–13.

Wang, J.-L., Qi, Z., Li, Y.-H., Zhao, H.-M., Chen, Y.-G., and Fu, W. (2017). TGFβ induced factor homeobox 1 promotes colorectal cancer development through activating Wnt/β-catenin signaling. *Oncotarget*, 8(41):70214.

Wang, L. and Dunson, D. B. (2010). Semiparametric Bayes multiple testing: Applications to tumor data. *Biometrics*, 66(2):493–501.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2013). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159.

Wasserman, L. and Roeder, K. (2006). Weighted hypothesis testing. *arXiv preprint math/0604172*.

Witt, S. H., Frank, J., Gilles, M., Lang, M., Treutlein, J., Streit, F., Wolf, I. A., Peus, V., Scharnholz, B., Send, T. S., et al. (2018). Impact on birth weight of maternal smoking throughout pregnancy mediated by dna methylation. *Bmc Genomics*, 19(1):1–10.

Yang, Y., Wang, Q., Chen, Q., Liao, R., Zhang, X., Yang, H., Zheng, Y., Zhang, Z., and Pan, Y. (2014). A new genotype imputation method with tolerance to high missing rate and rare variants. *PloS One*, 9(6):e101025.

Yang, Z. and Ho, Y.-Y. (2021). Modeling dynamic correlation in zero-inflated bivariate count data with applications to single-cell RNA sequencing data. *Biometrics*.

Yu, L., Zhou, R., Chen, R., and Lai, K. K. (2020). Missing data preprocessing in credit classification: One-hot encoding or imputation? *Emerging Markets Finance and Trade*, pages 1–11.

Yu, T. (2018). A new dynamic correlation algorithm reveals novel functional aspects in single cell and bulk RNA-seq data. *PLoS Computational Biology*, 14(8):e1006391.

Zhao, S. D., Cai, T. T., and Li, H. (2014). More powerful genetic association testing via a new statistical framework for integrative genomics. *Biometrics*, 70(4):881–890.

Appendix A

Likelihood and Posterior Distributions for SPSL and C-SPSL

In this section, we provide the likelihood function and the full conditional distributions of the parameters involved in the parameter estimation process of the proposed spike-and-slab (SPSL) model and the correlated spike-and-slab (C-SPSL) model as described in Section 3.3. The conditional distribution of the response variable $Y_{i,j}|z_i, \tau_{0,j}, \tau_{1,j}$ given the covariate z_i and parameters of interest $\tau_{0,j}$ and $\tau_{1,j}$ for i-th observation $(i = 1, ..., n)$ and j-th gene pair $(j = 1, ..., m)$ is a product distribution of two standard normal distributions with correlation coefficient $g^{-1}(\tau_{0,j} + \tau_{1,j} z_i)$ as described in Section 3.3, where $g^{-1}(\cdot) = \frac{e^{2(\cdot)}-1}{e^{2(\cdot)}+1}$ is the inverse function of Fisher's Z-transformation. Based on the independence assumption, the likelihood function is given as follows,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\tau_0}, \boldsymbol{\tau_1}|\boldsymbol{Y}, \boldsymbol{z}) = \prod_{i=1}^{n}\prod_{j=1}^{m} &\frac{1}{\pi\sqrt{1-\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}^2}} e^{\frac{\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}y_{i,j}}{1-\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}^2}} \\ &K_0\left[\frac{|y_{i,j}|}{1-\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}^2}\right],\end{aligned} \tag{A.1}$$

where $K_0(\cdot)$ is the second class zero-order modified Bessel function.

Since the prior distributions of $\boldsymbol{\tau_0}$ are assumed to be independent standard normal distributions, the full conditional distribution of $\boldsymbol{\tau_0}|\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{\tau_1}$ can be derived as follows,

$$\begin{aligned}\pi(\boldsymbol{\tau_0}|\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{\tau_1}) \propto &\mathcal{L}(\boldsymbol{\tau_0}, \boldsymbol{\tau_1}|\boldsymbol{Y}, \boldsymbol{z}) \times \pi(\boldsymbol{\tau_0}) \\ \propto &\prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\pi\sqrt{1-\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}^2}} e^{\frac{\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}y_{i,j}}{1-\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}^2}} \\ &K_0\left[\frac{|y_{i,j}|}{1-\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}^2}\right] \times e^{-\frac{\tau_{0,j}^2}{2}}.\end{aligned} \tag{A.2}$$

To find the full conditional distribution of $\boldsymbol{\tau_1}$, we introduce a latent variable $\phi_j = \gamma_j \cdot v_j^2$ for j-th pair of genes as an auxiliary variable to simplify the calculation of the posterior density of $\boldsymbol{\tau_1}$. Based on the independence assumption and the likelihood function derived above, the full conditional distribution of $\boldsymbol{\tau_1}|\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{\tau_0}, \boldsymbol{\phi}$ is provided as follows,

$$\begin{aligned}\pi(\boldsymbol{\tau_1}|\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{\tau_0}, \boldsymbol{\phi}) \propto & \mathcal{L}(\boldsymbol{\tau_0}, \boldsymbol{\tau_1}|\boldsymbol{Y}, \boldsymbol{z}) \times \pi(\boldsymbol{\tau_1}|\boldsymbol{\phi}) \\ \propto & \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\pi\sqrt{1-\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}^2}} e^{\frac{\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}y_{i,j}}{1-\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}^2}} \\ & K_0\left[\frac{|y_{i,j}|}{1-\{g^{-1}(\tau_{0,j}+\tau_{1,j}z_i)\}^2}\right] \times \phi_j^{-\frac{1}{2}} e^{-\frac{\tau_{1,j}^2}{2\phi_j}}.\end{aligned} \tag{A.3}$$

To find the full conditional distribution of $\boldsymbol{\phi}|\boldsymbol{\tau_1}, \boldsymbol{w}$, we need to derive the conditional distribution of $\phi_j|w_j$ first. Since $\gamma_j|w_j$ is a discrete probability measure concentrated on either 1 or v_0 with probabilities $P(\gamma_j = 1|w_j) = w_j$ and $P(\gamma_j = v_0|w_j) = 1 - w_j$, the conditional distribution of $\phi_j|\gamma_j, w_j$ can also have two measures. Recall that ϕ_j is the product of γ_j and v_j^2, which is assumed to be an inverse gamma distribution with shape and scale parameters α_1 and α_2. The conditional distribution of $\phi_j|w_j, \gamma_j = 1$ is $\mathcal{IG}(\alpha_1, \alpha_2)$ while the conditional distribution of $\phi_j|w_j, \gamma_j = v_0$ is an inverse gamma distribution with a smaller scale parameter $\mathcal{IG}(\alpha_1, v_0\alpha_2)$.

Then we marginalize γ_j to derive the conditional distribution of $\phi_j|w_j$,

$$\begin{aligned}\pi(\phi_j|w_j) &= \pi(\phi_j|w_j, \gamma_j = 1) \times P(\gamma_j = 1|w_j) + \pi(\phi_j|w_j, \gamma_j = v_0) \times P(\gamma_j = v_0|w_j) \\ &= \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}\phi_j^{-\alpha_1-1}e^{-\frac{\alpha_2}{\phi_j}} \times w_j + \frac{(v_0\alpha_2)^{\alpha_1}}{\Gamma(\alpha_1)}\phi_j^{-\alpha_1-1}e^{-\frac{v_0\alpha_2}{\phi_j}} \times (1-w_j),\end{aligned} \tag{A.4}$$

which follows a mixture of inverse gamma distributions and can be denoted as $w_j\mathcal{IG}(\alpha_1, \alpha_2) + (1 - w_j)\mathcal{IG}(\alpha_1, v_0\alpha_2)$. Therefore, the full conditional distribution of $\boldsymbol{\phi}|\boldsymbol{\tau_1}, \boldsymbol{w}$ can be derived as follows,

$$
\begin{aligned}
\pi(\boldsymbol{\phi}|\boldsymbol{\tau_1},\boldsymbol{w}) =& \frac{\pi(\boldsymbol{\tau_1}|\boldsymbol{\phi})\times\pi(\boldsymbol{\phi}|\boldsymbol{w})}{\int \pi(\boldsymbol{\tau_1}|\boldsymbol{\phi})\times\pi(\boldsymbol{\phi}|\boldsymbol{w})d\boldsymbol{\phi}} \\
=& \prod_{j=1}^{m} \frac{w_j\phi_j^{-(\alpha_1+\frac{1}{2})-1}e^{-\frac{\tau_{1,j}^2+2\alpha_2}{2\phi_j}} + (1-w_j)\phi_j^{-(\alpha_1+\frac{1}{2})-1}e^{-\frac{\tau_{1,j}^2+2v_0\alpha_2}{2\phi_j}}v_0^{\alpha_1}}{\Gamma(\alpha_1+\frac{1}{2})\left\{w_j\left(\frac{\tau_{1,j}^2}{2}+\alpha_2\right)^{-\alpha_1-\frac{1}{2}} + (1-w_j)v_0^{\alpha_1}\left(\frac{\tau_{1,j}^2}{2}+v_0\alpha_2\right)^{-\alpha_1-\frac{1}{2}}\right\}} \\
=& \prod_{j=1}^{m}\left\{\mathcal{C}_j\times\frac{\left(\frac{\tau_{1,j}^2}{2}+\alpha_2\right)^{\alpha_1+\frac{1}{2}}}{\Gamma(\alpha_1+\frac{1}{2})}\phi_j^{-(\alpha_1+\frac{1}{2})-1}e^{-\frac{\tau_{1,j}^2+2\alpha_2}{2\phi_j}} + \right. \\
& \left.(1-\mathcal{C}_j)\times\frac{\left(\frac{\tau_{1,j}^2}{2}+v_0\alpha_2\right)^{\alpha_1+\frac{1}{2}}}{\Gamma(\alpha_1+\frac{1}{2})}\phi_j^{-(\alpha_1+\frac{1}{2})-1}e^{-\frac{\tau_{1,j}^2+2v_0\alpha_2}{2\phi_j}}\right\},
\end{aligned}
\tag{A.5}
$$

where $\mathcal{C}_j = \dfrac{w_j\left(\frac{\tau_{1,j}^2}{2}+\alpha_2\right)^{-\alpha_1-\frac{1}{2}}}{w_j\left(\frac{\tau_{1,j}^2}{2}+\alpha_2\right)^{-\alpha_1-\frac{1}{2}} + (1-w_j)v_0^{\alpha_1}\left(\frac{\tau_{1,j}^2}{2}+v_0\alpha_2\right)^{-\alpha_1-\frac{1}{2}}}$ is a constant for ϕ_j. Therefore, for j-th gene pair, $\phi_j|\tau_{1,j},w_j$ also follows a mixture of inverse gamma distributions which can be denoted as $\mathcal{C}_j\times\mathcal{IG}(\alpha_1+\frac{1}{2},\frac{\tau_{1,j}^2}{2}+\alpha_2)+(1-\mathcal{C}_j)\times\mathcal{IG}(\alpha_1+\frac{1}{2},\frac{\tau_{1,j}^2}{2}+v_0\alpha_2)$.

As discussed in the Section 3.3, the difference between the SPSL model and the C-SPSL model is that the SPSL model assumes independent uniform prior distributions for $\boldsymbol{w}$ while the C-SPSL model uses a beta-uniform prior structure to incorporate the dependence structure in the model. Therefore, the full conditional distribution of $\boldsymbol{w}|\boldsymbol{\phi}$ in the SPSL model can be derived as follows,

$$
\begin{aligned}
\pi(\boldsymbol{w}|\boldsymbol{\phi}) =& \frac{\pi(\boldsymbol{\phi}|\boldsymbol{w})\times\pi(\boldsymbol{w})}{\int \pi(\boldsymbol{\phi}|\boldsymbol{w})\times\pi(\boldsymbol{w})d\boldsymbol{\phi}} \\
=& \prod_{j=1}^{m}\frac{\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}\phi_j^{-\alpha_1-1}e^{-\frac{\alpha_2}{\phi_j}}\times w_j + \frac{(v_0\alpha_2)^{\alpha_1}}{\Gamma(\alpha_1)}\phi_j^{-\alpha_1-1}e^{-\frac{v_0\alpha_2}{\phi_j}}\times(1-w_j)}{\frac{\Gamma(2)\Gamma(1)}{\Gamma(3)}\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}\phi_j^{-\alpha_1-1}e^{-\frac{\alpha_2}{\phi_j}} + \frac{\Gamma(2)\Gamma(1)}{\Gamma(3)}\frac{(v_0\alpha_2)^{\alpha_1}}{\Gamma(\alpha_1)}\phi_j^{-\alpha_1-1}e^{-\frac{v_0\alpha_2}{\phi_j}}} \\
=& \prod_{j=1}^{m}\left\{\mathcal{D}_j\times\frac{\Gamma(3)}{\Gamma(2)\Gamma(1)}w_j + (1-\mathcal{D}_j)\times\frac{\Gamma(3)}{\Gamma(2)\Gamma(1)}(1-w_j)\right\},
\end{aligned}
\tag{A.6}
$$

where $\mathcal{D}_j = \dfrac{e^{-\frac{\alpha_2}{\phi_j}}}{e^{-\frac{\alpha_2}{\phi_j}}+e^{-\frac{v_0\alpha_2}{\phi_j}}}$ is a constant for w_j. The conditional distribution of $w_j|\phi_j$ is a mixture of beta distribution which can be denoted as $\mathcal{D}_j\times\mathcal{B}(2,1)+(1-\mathcal{D}_j)\times\mathcal{B}(1,2)$.

In the C-SPSL model, we assume the conditional distribution of $w_j|c$ for j-th pair of genes to be a beta distribution with global hyperparameters c and $1-c$. The full conditional distribution of $\boldsymbol{w}|\boldsymbol{\phi}$ in the C-SPSL model is given by,

$$\begin{aligned}
\pi(\boldsymbol{w}|\boldsymbol{\phi}, c) =& \frac{\pi(\boldsymbol{\phi}|\boldsymbol{w}) \times \pi(\boldsymbol{w}|c)}{\int \pi(\boldsymbol{\phi}|\boldsymbol{w}) \times \pi(\boldsymbol{w}|c) d\boldsymbol{\phi}} \\
=& \prod_{j=1}^{m} \frac{\left\{ \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \phi_j^{-\alpha_1-1} e^{-\frac{\alpha_2}{\phi_j}} \times w_j + \frac{(v_0\alpha_2)^{\alpha_1}}{\Gamma(\alpha_1)} \phi_j^{-\alpha_1-1} e^{-\frac{v_0\alpha_2}{\phi_j}} \times (1-w_j) \right\} \times w_j^{c-1}(1-w_j)^{-c}}{\frac{\Gamma(1+c)\Gamma(1-c)}{\Gamma(2)} \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \phi_j^{-\alpha_1-1} e^{-\frac{\alpha_2}{\phi_j}} + \frac{\Gamma(c)\Gamma(2-c)}{\Gamma(2)} \frac{(v_0\alpha_2)^{\alpha_1}}{\Gamma(\alpha_1)} \phi_j^{-\alpha_1-1} e^{-\frac{v_0\alpha_2}{\phi_j}}} \\
=& \prod_{j=1}^{m} \left\{ \mathcal{E}_j \times \frac{\Gamma(2)}{\Gamma(1+c)\Gamma(1-c)} w_j^{(1+c)-1}(1-w_j)^{(1-c)-1} \right. \\
& \left. + (1-\mathcal{E}_j) \times \frac{\Gamma(2)}{\Gamma(c)\Gamma(2-c)} w_j^{c-1}(1-w_j)^{(2-c)-1} \right\},
\end{aligned} \tag{A.7}$$

where $\mathcal{E}_j = \frac{\Gamma(1+c)\Gamma(1-c)e^{-\frac{\alpha_2}{\phi_j}}}{\Gamma(1+c)\Gamma(1-c)e^{-\frac{\alpha_2}{\phi_j}} + \Gamma(c)\Gamma(2-c)e^{-\frac{v_0\alpha_2}{\phi_j}}}$ is a constant for w_j. Therefore, the conditional distribution of $w_j|\phi_j, c$ can be denoted as a mixture of beta distributions, $\mathcal{E}_j \times \mathcal{B}(1+c, 1-c) + (1-\mathcal{E}_j) \times \mathcal{B}(c, 2-c)$.

By assuming a uniform distribution to c, the conditional distribution of $c|\boldsymbol{w}$ in the C-SPSL model can be derived as follows,

$$\begin{aligned}
\pi(c|\boldsymbol{w}) \propto& \pi(\boldsymbol{w}|c) \times \pi(c) \\
\propto& \prod_{j=1}^{m} \frac{1}{\Gamma(c)\Gamma(1-c)} w_j^{c-1}(1-w_j)^{-c}.
\end{aligned} \tag{A.8}$$

With the likelihood function and all the conditional distributions for the hyperparameters derived above, we implemented the adaptive Metropolis-Hastings sampling scheme and obtained the posterior samples of the parameters of interest. More details are discussed in the next section.

Appendix B

Adaptive Metropolis-Hastings Sampling Algorithm

As discussed in Section 3.3, we developed an efficient Markov Chain Monte Carlo (MCMC) algorithm using the adaptive Metropolis (AM) algorithm (Haario et al., 2001). Algorithm 1 and 2 provide the procedures of the adaptive MCMC implemented in SPSL and C-SPSL, respectively. In the AM algorithm, the proposal distribution is a multivariate normal distribution centered on the current state, and the covariance matrix is calculated from a fixed number of previous states. Due to the ranges of support for various parameters, appropriate transformations are considered with samples from the multivariate proposal distribution. Because of the symmetry of the normal proposal distributions suggested in the AM algorithm, the acceptance-rejection rate is equivalent to the posterior ratios.

As presented in Algorithm 1 and 2, the sampling process is cut into two stages. In the initial stage, the proposal distributions are independent normal distributions with the step size $\lambda = 0.05$. After the cutting point t_h, multivariate normal distribution with covariance matrix calculated from the history states are used as the proposal distributions. The covariance matrix is multiplied by a different step size $S_d = 10$ in this stage to give an acceptance rate close to 20-30%. A small value ξ (e.g. 0.00001) is added to the diagonal elements of the covariance matrix to avoid singularity in the variance-covariance structure. We set the history tracing size as $h = 200$ when calculating the covariance matrix to reduce the computation cost ($h \leq t_h$).

In the adaptation process (when $t_h < i < t_N$ and t_N is the total number of iteration), the proposal distributions are separated into four groups in the SPSL model and five groups in the C-SPSL model. Within each group, the covariance matrix of the multivariate normal distribution is calculated. The groups are segregated in terms of the objects of the parameters. For example, in the SPSL model, the four groups are $\boldsymbol{\tau}_0$, $\boldsymbol{\tau}_1$, $\boldsymbol{\phi}$, and $\boldsymbol{w}$ while in the C-SPSL model, the five groups are $\boldsymbol{\tau}_0$, $\boldsymbol{\tau}_1$, $\boldsymbol{\phi}$, $\boldsymbol{w}$, and c. The dimension of parameters in each group is denoted as $d^{'}$.

Algorithm 1: The adaptive Metropolis algorithm implemented in SPSL

Initialize the parameters $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\tau_0^{(0)}}, \boldsymbol{\tau_1^{(0)}}, \boldsymbol{\phi^{(0)}}, \boldsymbol{w^{(0)}})$, $\boldsymbol{\theta} \in \mathbb{R}^d$;
for *i=1* ***to*** t_N **do**
  if $i \leq t_h$ **then**
    Derive $\boldsymbol{q}^*$ from the proposal distribution $\mathcal{N}_d(\boldsymbol{\theta}^{(i-1)}, \lambda \boldsymbol{I}_d)$ with transformations in terms of $\boldsymbol{\theta}^{(i-1)}$;
    for $j = 1$ ***to*** d **do**
      Calculate the acceptance-rejection rate :
      $A(q_j^*, \theta_j^{(i-1)}) = min\left\{1, \frac{\pi(q_j^*|\cdot)}{\pi(\theta_j^{(i-1)}|\cdot)}\right\}$;
      if $A(q_j^*, \theta_j^{(i-1)}) \geq \mathcal{U}(0,1)$ *or* $A(q_j^*, \theta_j^{(i-1)}) = 1$ **then**
        $\theta_j^{(i)} = q_j^*$;
      else
        $\theta_j^{(i)} = \theta_j^{(i-1)}$;
      end
    end
  end
  if $i > t_h$ **then**
    while $\boldsymbol{\theta}^{'}$ *in each group of* $(\boldsymbol{\tau_0}, \boldsymbol{\tau_1}, \boldsymbol{\phi}, \boldsymbol{w})$, $\boldsymbol{\theta}^{'} \in \mathbb{R}^{d^{'}}$ **do**
      Derive $\boldsymbol{q}^*$ from the proposal distribution $\mathcal{N}_{d'}(\boldsymbol{\theta}^{'(i-1)}, S_d cov(\boldsymbol{\theta}^{'(i-h)}, ..., \boldsymbol{\theta}^{'(i-1)}) + \xi \boldsymbol{I}_{d'})$ with transformations in terms of $\boldsymbol{\theta}^{'(i-1)}$;
      for $j = 1$ ***to*** $d^{'}$ **do**
        Calculate the acceptance-rejection rate:
        $A(q_j^*, \theta_j^{'(i-1)}) = min\left\{1, \frac{\pi(q_j^*|\cdot)}{\pi(\theta_j^{'(i-1)}|\cdot)}\right\}$;
        if $A(q_j^*, \theta_j^{'(i-1)}) \geq \mathcal{U}(0,1)$ *or* $A(q_j^*, \theta_j^{'(i-1)}) = 1$ **then**
          $\theta_j^{'(i)} = q_j^*$;
        else
          $\theta_j^{'(i)} = \theta_j^{'(i-1)}$;
        end
      end
    end
  end
end

Algorithm 2: The adaptive Metropolis algorithm implemented in C-SPSL

Initialize the parameters $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\tau_0^{(0)}}, \boldsymbol{\tau_1^{(0)}}, \boldsymbol{\phi^{(0)}}, \boldsymbol{w^{(0)}}, c)$, $\boldsymbol{\theta} \in \mathbb{R}^d$;
for *i=1* ***to*** t_N **do**
  if $i \leq t_h$ **then**
    Derive $\boldsymbol{q}^*$ from the proposal distribution $\mathcal{N}_d(\boldsymbol{\theta}^{(i-1)}, \lambda \boldsymbol{I}_d)$ with transformations in terms of $\boldsymbol{\theta}^{(i-1)}$;
    for $j = 1$ ***to*** d **do**
      Calculate the acceptance-rejection rate :
      $A(q_j^*, \theta_j^{(i-1)}) = min\left\{1, \frac{\pi(q_j^*|\cdot)}{\pi(\theta_j^{(i-1)}|\cdot)}\right\}$;
      if $A(q_j^*, \theta_j^{(i-1)}) \geq \mathcal{U}(0,1)$ *or* $A(q_j^*, \theta_j^{(i-1)}) = 1$ **then**
        $\theta_j^{(i)} = q_j^*$;
      else
        $\theta_j^{(i)} = \theta_j^{(i-1)}$;
      end
    end
  end
  if $i > t_h$ **then**
    while $\boldsymbol{\theta}^{'}$ *in each group of* $(\boldsymbol{\tau_0}, \boldsymbol{\tau_1}, \boldsymbol{\phi}, \boldsymbol{w}, c)$, $\boldsymbol{\theta}^{'} \in \mathbb{R}^{d^{'}}$ **do**
      Derive $\boldsymbol{q}^*$ from the proposal distribution $\mathcal{N}_{d'}(\boldsymbol{\theta}^{'(i-1)}, S_d cov(\boldsymbol{\theta}^{'(i-h)}, ..., \boldsymbol{\theta}^{'(i-1)}) + \xi \boldsymbol{I}_{d'})$ with transformations in terms of $\boldsymbol{\theta}^{'(i-1)}$;
      for $j = 1$ ***to*** $d^{'}$ **do**
        Calculate the acceptance-rejection rate:
        $A(q_j^*, \theta_j^{'(i-1)}) = min\left\{1, \frac{\pi(q_j^*|\cdot)}{\pi(\theta_j^{'(i-1)}|\cdot)}\right\}$;
        if $A(q_j^*, \theta_j^{'(i-1)}) \geq \mathcal{U}(0,1)$ *or* $A(q_j^*, \theta_j^{'(i-1)}) = 1$ **then**
          $\theta_j^{'(i)} = q_j^*$;
        else
          $\theta_j^{'(i)} = \theta_j^{'(i-1)}$;
        end
      end
    end
  end
end

Appendix C

Empirical distributions of $\widehat{\tau}_1$ in CRC Dataset

Let $\widehat{\tau}_1$ be the posterior mean of τ_1 based on all $949,280$ gene pairs after the screening step as described in the Section 3.5. The empirical density plot of $\widehat{\tau}_1$ is presented in Figure C.1. As observed in Figure C.1, the proportion of $|\widehat{\tau}_1|$ being greater than 0.5 is 0.023. It implies that the DC signals in the CRC dataset are highly sparse.

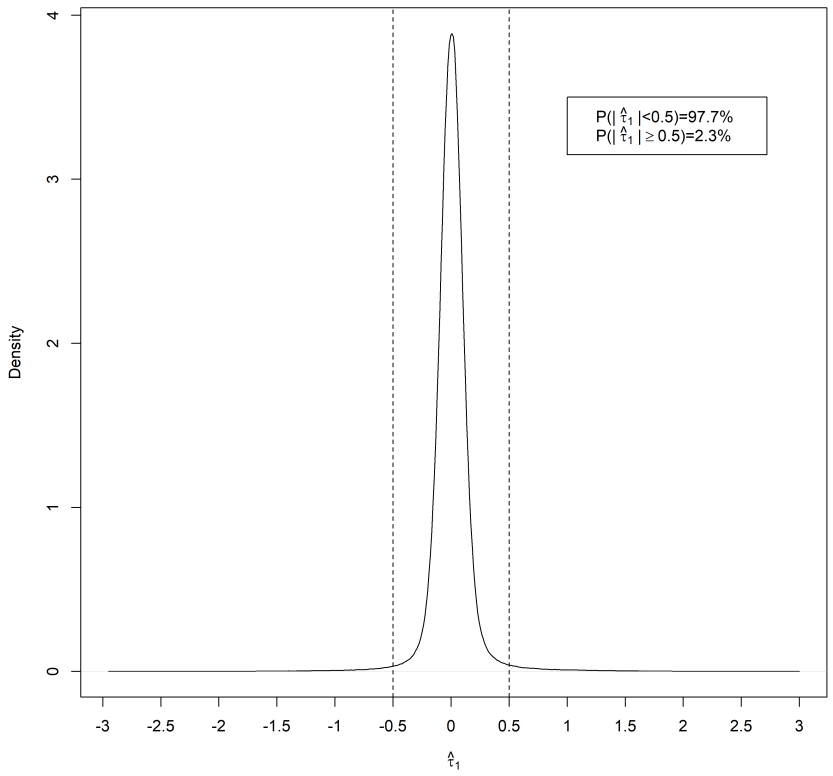


Figure C.1. Empirical density plot of $\widehat{\tau}_1$ for all $949,280$ gene-pair combinations. The proportion of $|\widehat{\tau}_1|$ being less than 0.5 is 0.977.

Appendix D

Convergence Diagnostics for SPSL and C-SPSL

To examine MCMC convergence in simulations, Gelman-Rubin diagnostics (Gelman et al., 1992) are used to provide numerical convergence summaries for both the SPSL model and the C-SPSL model in simulation scenario I. For each simulated dataset in the 100 simulation iterations, 3 MCMC chains were implemented for deriving the Gelman-Rubin diagnostics summaries. In this section, we pick one dataset and present the results of both SPSL and C-SPSL for all parameters in Table D.1 and Table D.2. As shown in the tables, the Gelman-Rubin statistics are all close to 1, which indicates good convergence for both SPSL and C-SPSL. In addition, all the other simulated datasets also have good convergence in MCMC chains.

Moreover, to visualize the convergence for the two models, trace plots generated from three mixed MCMC chains for all hyperparameters $(\tau_{0,j}, \tau_{1,j}, \phi_j, w_j, c)$ in both models for the same simulatied dataset are presented in Figure D.1 and Figure D.2. We choose the first gene pair ($\tau_{1,1} = 0$) and the tenth ($\tau_{1,10} = 1$) gene pair to show the convergence results for both zero and non-zero dynamic co-expression (DC) gene pairs. As presented in the Figures, MCMC chains for all parameters mixed well for both SPSL and C-SPSL.

The results of Gelman-Rubin diagnostics for the top 30 gene pairs with the largest $|\widehat{\tau}_{1,j}|$ based on Table 3.3 in Section 3.5 from the experimental data analysis are presented in Table D.3. All Gelman-Rubin statistics are close to 1. We also provide the trace plots for the $\tau_{1,j}$ of the top ten gene pairs in Figure D.3. The results of Gelman-Rubin diagnostics and the trace plots both indicate that the MCMC chains

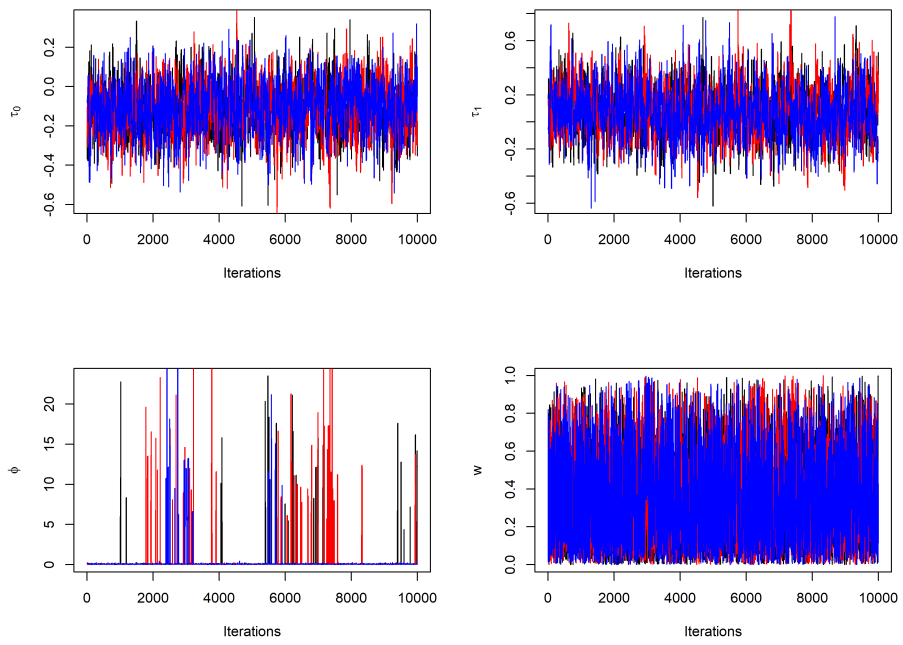
mixed well and converged for all the top gene pairs.

Table D.1. Gelman-Rubin diagnostics results for SPSL model with 5 genes (10 pairs) and sample size of 200.

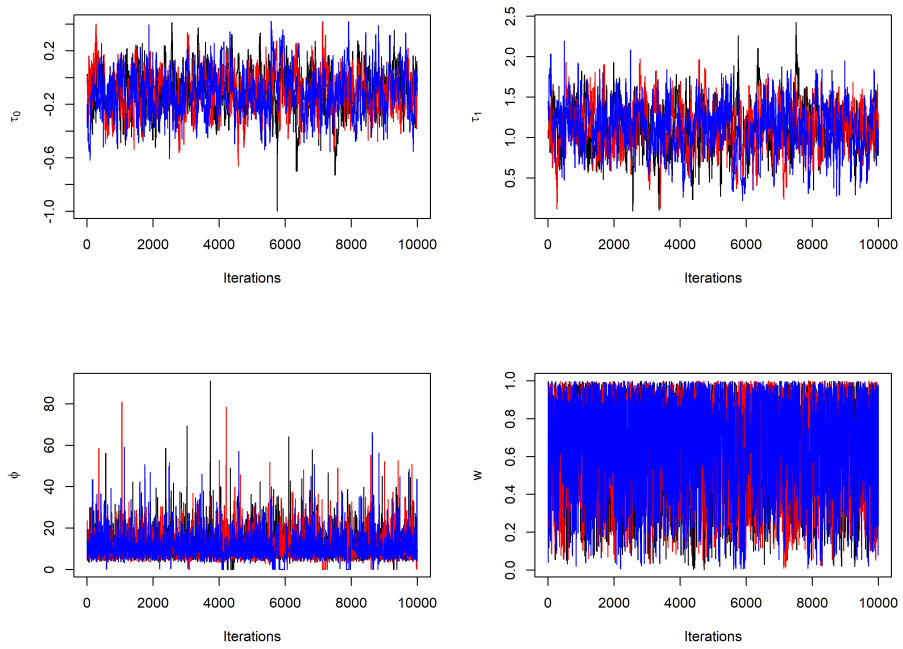
	$\tau_{0,j}$		$\tau_{1,j}$		ϕ_j		w_j	
j	Point est.	Upper C.I.	Point est.	Upper C.I.	Point est.	Upper C.I.	Point est.	Upper C.I.
1	1.00	1.01	1.00	1.00	1.05	1.08	1.00	1.00
2	1.00	1.00	1.00	1.00	1.03	1.04	1.00	1.01
3	1.01	1.02	1.01	1.02	1.00	1.02	1.00	1.00
4	1.01	1.02	1.01	1.03	1.03	1.07	1.00	1.00
5	1.00	1.01	1.00	1.01	1.08	1.18	1.00	1.02
6	1.00	1.00	1.00	1.00	1.08	1.17	1.01	1.02
7	1.01	1.03	1.01	1.03	1.02	1.04	1.00	1.00
8	1.01	1.02	1.01	1.02	1.00	1.00	1.00	1.00
9	1.00	1.01	1.01	1.02	1.04	1.09	1.00	1.01
10	1.00	1.01	1.00	1.01	1.01	1.01	1.00	1.00

Table D.2. Gelman-Rubin diagnostics results for C-SPSL model with 5 genes (10 pairs) and sample size of 200.

	$\tau_{0,j}$		$\tau_{1,j}$		ϕ_j		w_j	
j	Point est.	Upper C.I.	Point est.	Upper C.I.	Point est.	Upper C.I.	Point est.	Upper C.I.
1	1.00	1.01	1.00	1.00	1.01	1.02	1.01	1.02
2	1.00	1.00	1.00	1.00	1.01	1.03	1.01	1.02
3	1.01	1.02	1.01	1.02	1.05	1.11	1.01	1.03
4	1.00	1.01	1.00	1.00	1.12	1.21	1.00	1.01
5	1.01	1.02	1.01	1.03	1.02	1.04	1.00	1.01
6	1.00	1.01	1.00	1.00	1.01	1.02	1.00	1.00
7	1.00	1.01	1.00	1.00	1.06	1.08	1.00	1.00
8	1.01	1.03	1.01	1.03	1.01	1.03	1.01	1.03
9	1.04	1.12	1.05	1.16	1.09	1.23	1.02	1.06
10	1.00	1.01	1.00	1.01	1.06	1.08	1.01	1.03

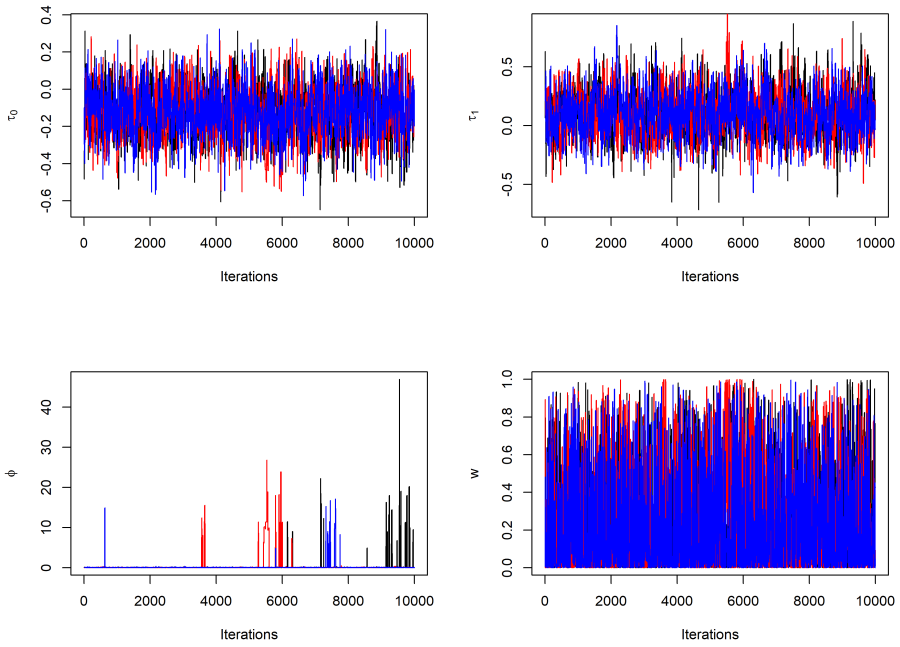


(a) Trace plots for $\tau_{0,10}$, $\tau_{1,10}$, ϕ_{10}, w_{10} when $\tau_{1,10} = 0$.

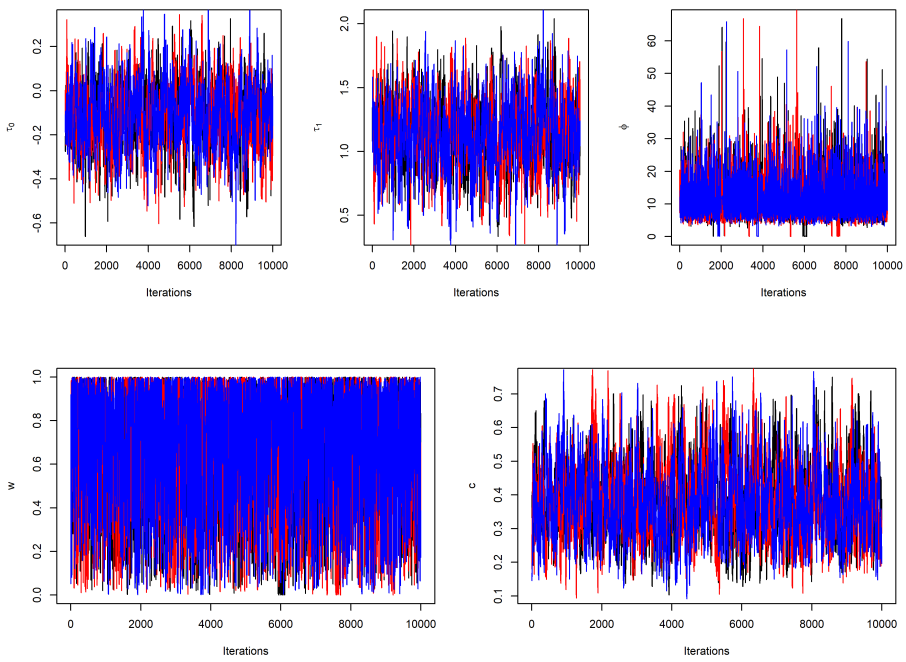


(b) Trace plots for $\tau_{0,10}$, $\tau_{1,10}$, ϕ_{10}, w_{10} when $\tau_{1,10} = 1$.

Figure D.1. Trace plots for SPSL model with sample size of 200.



(a) Trace plots for $\tau_{0,10}$, $\tau_{1,10}$, ϕ_{10}, w_{10} when $\tau_{1,10} = 0$.



(b) Trace plots for $\tau_{0,10}$, $\tau_{1,10}$, ϕ_{10}, w_{10} when $\tau_{1,10} = 1$.

Figure D.2. Trace plots for C-SPSL model with sample size of 200. Notice c is a global parameter for all j.

Table D.3. Gelman-Rubin diagnostics results of $\tau_{1,j}$ for the top 30 gene pairs with the largest $|\widehat{\tau}_{1,j}|$.

j	Gene I	Gene II	Point est.	Upper C.I.
1	C18orf12	MEP1B	1.14	1.40
2	SCEL	TGIF1	1.04	1.11
3	POLR3A	MEP1B	1.00	1.01
4	MEP1B	MIR4500HG	1.01	1.02
5	PRR16	PLAU	1.01	1.02
6	LOC642696	DDX25	1.01	1.03
7	METTL6	MEP1B	1.01	1.01
8	ZMYM2	AMELX	1.27	1.92
9	PIK3CA	PIP	1.00	1.02
10	KIF19	LOC105379426	1.00	1.01
11	KCNS1	TEX48	1.02	1.05
12	LINC01013	SMYD1	1.01	1.01
13	LOC100129175	ZNF787	1.00	1.00
14	PIK3CA	SCGB1D1	1.03	1.05
15	FMN2	SH2B1	1.00	1.00
16	SCGB1D2	SNHG10	1.00	1.01
17	C16orf74	TBC1D21	1.01	1.02
18	LOC389199	MYH7B	1.01	1.02
19	MYH7B	UCP1	1.00	1.01
20	PLEK	AQP9	1.01	1.02
21	FILIP1L	GDPD1	1.00	1.01
22	LOC101928760	TMEM151B	1.00	1.01
23	MYH7B	ADARB2	1.00	1.01
24	LOC101928553	SPINT3	1.01	1.02
25	KCNIP2	MLANA	1.00	1.01
26	MEP1B	NIPAL3	1.01	1.01
27	C11orf42	SNHG10	1.07	1.16
28	FAM218A	DIPK1B	1.00	1.00
29	LOC101928760	CLECL1	1.01	1.02
30	LINC00889	MEP1B	1.00	1.00

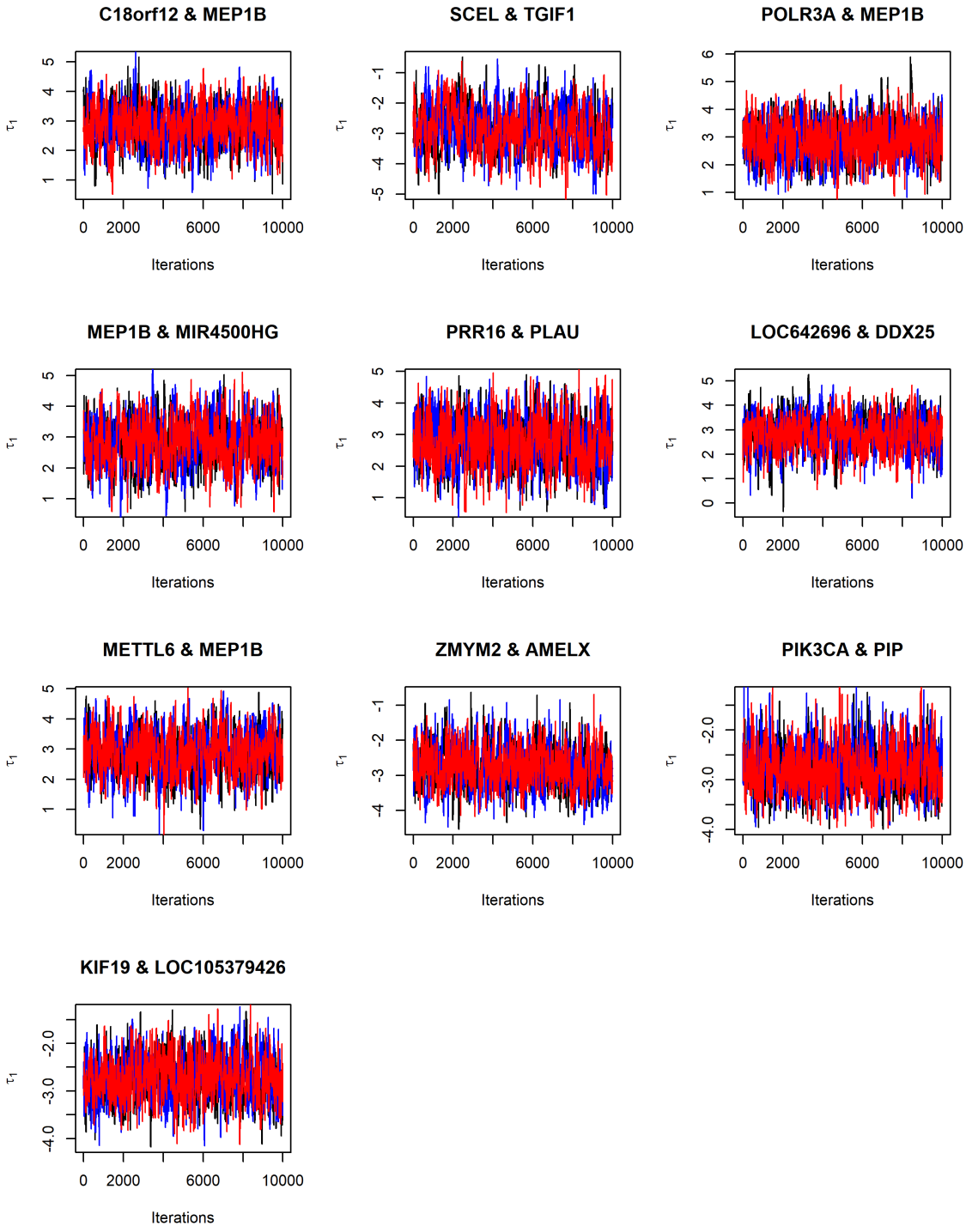


Figure D.3. Trace plots of $\tau_{1,j}$ for gene pairs with the largest ten $|\widehat{\tau}_{1,j}|$ in the colorectal cancer dataset.

Appendix E

Determination of the Recurrence-Free Survival Time

In order to implement the proposed model in the experimental data analysis to identify gene pairs with significant non-zero $\tau_{1,j}$ (DC gene pairs) associated with the recurrence-free survival time, we use individual patients' recurrence-free survival time ($\boldsymbol{z}$) to indicate their ability to survive without CRC recurrence or metastasis after the surgery. Let T_i be the cancer recurrence-free survival time for subject i. If subject i experience the CRC recurrence or metastasis after the surgery, then T_i is exactly observed and will be used for z_i. However, if subject i does not experience the CRC recurrence or metastasis after the surgery, T_i is right-censored at the last examination (or follow-up) time $t_{0,i}$. In this case, we will use the predictive survival time $E(T_i|T_i > t_{0,i})$ for z_i. In the following we describe how we obtain an estimate of $E(T_i|T_i > t_{0,i})$ for right-censored T_i.

Define $S(t|\nu) = P(T > t|\nu)$ as the marginal probability of T greater than t for a subject with covariates ν. One can obtain the estimated survival function based on the observed data using standard statistical packages, such as the survival package in R (The survival package, 2021). The conditional expectation is calculated as,

$$
\begin{aligned}
E(T_i|T_i > t_{0,i}, \boldsymbol{\nu}_i) &= \int_{t_{0,i}}^{\infty} t \cdot \frac{f(t|\boldsymbol{\nu}_i)}{P(T_i > t_{0,i}|\boldsymbol{\nu}_i)} dt \\
&= -\frac{1}{S(t_{0,i}|\boldsymbol{\nu}_i)} \int_{t_{0,i}}^{\infty} t dS(t|\boldsymbol{\nu}_i) \\
&= \frac{1}{S(t_{0,i}|\boldsymbol{\nu}_i)} \left\{ t_{0,i} S(t_{0,i}|\boldsymbol{\nu}_i) + \int_{t_{0,i}}^{\infty} S(t|\boldsymbol{\nu}_i) dt \right\} \\
&= t_{0,i} + \frac{1}{S(t_{0,i}|\boldsymbol{\nu}_i)} \int_{t_{0,i}}^{\infty} S(t|\boldsymbol{\nu}_i) dt.
\end{aligned}
\tag{E.1}
$$

One can estimate the above quantity by plugging in the estimated survival function $\widehat{S}(t|\nu)$ for $S(t|\nu)$. To calculate the integral, we propose to use importance sampling as follows. Given a set of samples $\omega_1, ..., \omega_l$ sampled from the proposal distribution with pdf $g(\cdot)$, the estimator for $\int_{t_{0,i}}^{\infty} \widehat{S}(t|\boldsymbol{\nu}_i) dt$ is given by the following equation,

$$
\begin{aligned}
\int_{t_{0,i}}^{\infty} \widehat{S}(t|\boldsymbol{\nu}_i) dt &= \int_{t_{0,i}}^{\infty} \frac{\widehat{S}(t|\boldsymbol{\nu}_i)}{g(t)} \cdot g(t) dt \\
&= E_g(\frac{\widehat{S}(t|\boldsymbol{\nu}_i)}{g(t)}) \\
&\simeq \frac{1}{l} \sum_{k=1}^{l} \frac{\widehat{S}(\omega_k|\boldsymbol{\nu}_i)}{g(\omega_k)},
\end{aligned}
\tag{E.2}
$$

where $\widehat{S}_i(t|\boldsymbol{\nu}_i) = e^{-\widehat{\Lambda}_0(t) e^{\boldsymbol{\nu}_i^T \widehat{\boldsymbol{\beta}}}}$ is the estimator of the survival function, $\widehat{\Lambda}_0(t)$ is the estimated cumulative baseline hazard function at time t and $\widehat{\boldsymbol{\beta}}$ is the estimated coefficient vector. The covariates $\boldsymbol{\nu}_i$ are specified as patients age, gender, and stage of colorectal cancer before surgery. Here we take the shifted exponential distribution $g(t) = e^{-(t-t_{0,i})} \mathcal{I}(t > t_{0,i})$ as the proposal distribution in the sampling process. The number of samples of the proposal distributions $l = 1,000$ ensures the accuracy of the approximation.

Algorithm 3 below summarizes how we obtain the predictive survival $E(T_i|T_i > t_{0,i})$ for right-censored observation i. The survival time for right-censored records are approximated using the predicted survival.

After calculating the expected survival times $E(T_i|T_i > t_{0,i})$ for right-censored observations, we can derive the covariate vector $\boldsymbol{z}$ as described above. At the end,

Algorithm 3: Calculate the expected survival times for right-censored observations

1. Build Cox model with covariates $\boldsymbol{\nu}$ (for all observations).
2. Sample $\omega_1, ...\omega_{1000}$ from $g(t) = e^{-(t-t_{0,i})}\mathcal{I}(t > t_{0,i})$ for the i-th observation.
3. Estimate $\widehat{S}(\cdot|\boldsymbol{\nu}_i)$ and calculate $\frac{\widehat{S}(\omega_k|\boldsymbol{\nu}_i)}{g(\omega_k)}$, $k = 1, ..., l$.
4. Approximate the predictive survival $E(T_i|T_i > t_{0,i}, \boldsymbol{\nu}_i)$ using $t_{0,i} + \frac{1}{\widehat{S}(t_{0,i}|\boldsymbol{\nu}_i)} \times \frac{1}{l}\sum_{k=1}^{l} \frac{\widehat{S}(\omega_k|\boldsymbol{\nu}_i)}{g(\omega_k)}$.

the recurrence-free survival time estimates are re-scaled by the maximum value so that each z_i is within $(0, 1]$.