5-2016

INTEGRATION OF MULTI-PLATFORM HIGH-DIMENSIONAL OMIC DATA

Xuebei An

INTEGRATION OF MULTI-PLATFORM HIGH-DIMENSIONAL OMIC DATA

by

Xuebei An, M.S.

APPROVED:

_____
Kim-Anh Do, Ph.D.
Advisory Professor

_____
Jianhua Hu, Ph.D.

_____
Veerabhadran Baladandayuthapani, Ph.D.

_____
Bradley M. Broom, Ph.D.

_____
Guillermina (Gigi) Lozano, Ph.D.

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

INTEGRATION OF MULTI-PLATFORM HIGH-DIMENSIONAL OMIC DATA

A DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

M.D. Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment of Requirements

for the Degree of

Doctor of Philosophy

by

Xuebei An

Houston, Texas, USA

May 2016

To my parents

who

always support me pursuing my dream.

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my advisors Dr. Kim-Anh Do and Dr. Jianhua Hu for their inspiring and constructive advice to my projects. I sincerely appreciate their countless hours of reflecting, reading, and encouraging during the preparation of this dissertation. I would also like to thank the members of my doctoral committee for their helpful comments and suggestions. Finally, I would thank my beloved family and friends for supporting me and encouraging me as always. Their willingness and excitement made the completion of this long journey an enjoyable experience.

INTEGRATION OF MULTI-PLATFORM HIGH-DIMENSIONAL OMIC DATA

Xuebei An, M.S.

Advisory Professor: Kim-Anh Do, Ph.D.

The development of high-throughput biotechnologies have made data accessible from different platforms, including RNA sequencing, copy number variation, DNA methylation, protein lysate arrays, etc. The high-dimensional *omic* data derived from different technological platforms have been extensively used to facilitate comprehensive understanding of disease mechanisms and to determine personalized health treatments. Although vital to the progress of clinical research, the high dimensional multi-platform data impose new challenges for data analysis. Numerous studies have been proposed to integrate multi-platform *omic* data; however, few have efficiently and simultaneously addressed the problems that arise from high dimensionality and complex correlations.

In my dissertation, I propose a statistical framework of shared informative factor model (*SIFORM*) that can jointly analyze multi-platform *omic* data and explore their associations with a disease phenotype. The common disease- associated sample characteristics across different data types can be captured through the shared structure space, while the corresponding weights of genetic variables directly index the strengths of their association with the phenotype. I compare the performance of the proposed method with several popular regularized regression methods and canoni-

cal correlation analysis (*CCA*)-based methods through extensive simulation studies and two lung adenocarcinoma applications. The two lung adenocarcinoma applications jointly explore the associations of mRNA expression and protein expression with smoking status and survival using The Cancer Genome Atlas (TCGA) datasets. The simulation studies demonstrate the superior performance of *SIFORM* in terms of biomarker detection accuracy. In lung cancer applications, *SIFORM* identifies many biomarkers that belong to key pathways for lung tumorigenesis. It also discovers potential prognostic biomarkers for lung cancer patients survival and some biomarkers that reveal different tumorigenesis mechanisms between light smokers and heavy smokers.

To improve the prediction accuracy and interpretability of the proposed model, I extend it to *PSIFORM* by incorporating existing biological pathway information to current statistical framework. I adopt a network-based regularization to ensure that the neighboring genes in the same pathway tend to be selected (or eliminated) simultaneously. Through simulation studies and a TCGA kidney cancer application, I show that *PSIFORM* outperforms its competitors in both variable selection and prediction. The statistical framework of *PSIFORM* also has a great potential in incorporating the hierarchical order across the multi-platform *omic* measurements.

TABLE OF CONTENTS

Page

List of Figures

List of Tables

1. Introduction and Background

In this chapter, I review several methods commonly used for integrative *omic* data analysis and biological pathway knowledge incorporation.

1.1 Introduction

The success of the Human Genome Project allows for genome-wide studies of various biological activities in humans and other organisms. The extensive development of high-throughput biotechnologies have made data accessible from different platforms, including RNA sequencing, copy number variation, DNA methylation, and protein lysate arrays. These high-dimensional *omic* data derived from different technological platforms have been extensively used to facilitate comprehensive understanding of disease mechanisms (e.g., the genetic profiles that are associated with tumor pathogenesis, progression, and prognosis) and to determine personalized health treatments, especially for cancer patients [1, 2].

Although vital to the progress of biomedical research, these multi-platform data impose new challenges for data analysis. In addition to high dimensionality and complex correlations within and across platforms, different types of *omic* data likely have different scales and distributions [3]. It is also recognized that the incorporation of existing biological information (e.g., biological pathways) into the analysis of multi-platform *omic* data can lead to more accurate prediction and improved interpretability of the results [4, 5]. Therefore, effective analytical methods are desirable to extract and integrate useful information from the emerging data platforms with proper utilization of the existing biological knowledge.

Numerous integrative analyses of multi-platform data have been proposed to study the interplay within and between different levels of biological data[6, 7, 8]. Some of the analyses mainly depend on biological knowledge and experimental results[9, 10, 11], while others are more statistical methodology [12, 13, 14]. This dissertation focuses on the latter. A large number of statistical methods have been discussed for different applications and purposes. For example, pairwise correlation of genetic variables (e.g., weighted gene correlation network analysis[15]) is used to infer molecular network interactions. Network analysis (e.g., differential network analysis in genomics[16]) identifies active or aberrant subsets in a certain biological system by using molecular network interactions based on graphical models. Bayesian network approaches[17, 18] identify biomarkers that are associated with clinical outcomes by incorporating another type of biomarkers collected from the same samples through prior distributions. Penalized likelihood analysis (e.g., *Lasso*[19]) handles high-dimensional multi-omic data by regularizing the coefficients. The commonly used regularized regression methods are briefly discussed in Section 1.2. Canonical correlation analysis (CCA) determines the linear relationship between two sets of biological variables on the same set of samples. This line of work includes sparse *CCA*, supervised *CCA*, and multiple *CCA* [35, 36, 39], which are briefly introduced in Section 1.3. However, few of these methods have efficiently and simultaneously addressed the problems that arise from high dimensionality and complex correlations. In this dissertation, I propose a regularized regression-based framework of the shared informative factor model (*SIFORM*) to detect predictive biomarkers from high-throughput multi-platform data for a response variable of interest, which is typically a disease-associated phenotype.

Methods for the incorporation of biological knowledge have also been discussed in the statistical literature [4, 5, 20, 21, 22, 23, 24, 25]. This dissertation focuses on incorporating existing pathway information into statistical models. Biologically related genes are known to form groups called pathways. The genes on the same

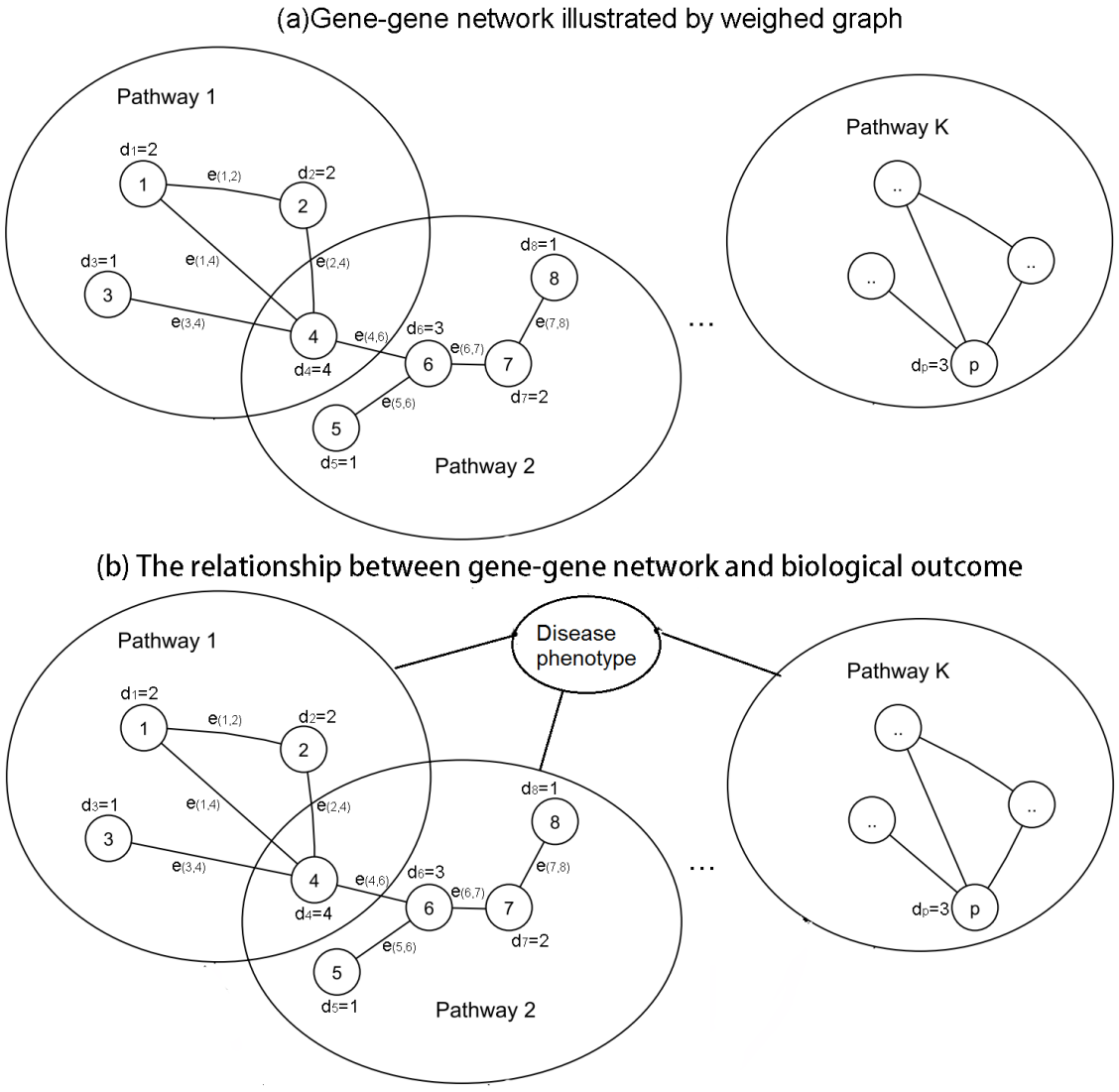
pathway have similar functions and may be associated with a disease phenotype in a similar manner. In addition, the genes within and between pathways interact with each other in different ways [20]. Many databases have been developed to store the existing biological pathway knowledge attained from biomedical research. Some examples include protein-protein interaction (PPI) networks that are available from the Biomolecular Interaction Network Database (BIND) [26], and biological pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [27]. Hence, it is both desirable and feasible to incorporate pathway information to improve the prediction accuracy and interpretability of statistical methods.

Most of the pathway incorporation methods use graphical models to characterize the network structure of genetic variables. A gene-gene network can be represented as a weighted graph $G = (V, E, W)$, where V is the set of vertices that correspond to the p genetic variables, $E = \{u \sim v\}$ is the set of edges indicating that the variables u and v are linked within the network through edge $e = (u \sim v)$, and W is the weights of the edges. In W, $\omega(u, v)$ denotes the weight of edge $e = (u \sim v)$, which represents the probability that vertices u and v are connected [21]. In applications, $\omega(u, v)$ is chosen as $\sqrt{d_j}$ or simply 1, where d_j is the degree of node j, that is, the number of edges connected to j [21, 23]. The idea of weighted graphs is demonstrated in Figure 1.1(a).

Current graphical model-based statistical methods for incorporating pathway information mainly include Bayesian models [4, 20, 24, 25] and penalized regression approaches [5, 21, 22, 23]. In Bayesian settings, some researchers use a Markov random field prior that captures the gene-gene interaction network to select genetic variables correlated with outcomes [24, 25]. Others combine penalized regression methods with Bayesian approaches. For example, Rockova proposed a Bayesian *Lasso* method by using pathway information to pose within-group sparsity rather than as a source of strict regularization constraints [20]. The penalized regression models post con-

straints on the coefficients by making assumptions about the similarity of genes in the same pathway, as Figure 1.1(b) shows. This line of work is discussed in Section 1.4 in detail. In this dissertation, a novel penalized regression approach is proposed to incorporate prior pathway knowledge into the shared informative factor model.

Figure 1.1.: Graph-based biological pathway structure



1.2 Regularized regression methods

In this section, I present a brief review of commonly used variable selection methods, including the L_1-penalty-based *Lasso* [19], its improved version, adaptive *Lasso* (*adaLasso*) [28], a combination of L_1 and L_2, the elastic net [31], and smoothly clipped absolute deviation (*SCAD*) [32, 33].

Suppose that we have n samples, each of which has p predictor variables and one response variable. This can be expressed as $(\boldsymbol{x_i}, y_i), i = 1, 2, \cdots, n$, where $\boldsymbol{x_i}$= $(x_{i1}, \cdots, x_{ip})^T$ are the predictor variables and y_i is the response. It is usually assumed that the conditional mean of Y given $\boldsymbol{X}$ depends on the linear predictor $\boldsymbol{\beta^T X}$ with $\boldsymbol{\beta}$ =$(\beta_1, \cdots, \beta_p)^T$. When $p \gg n$, conventional methods fail to handle the parameter estimation robustly and efficiently due to the singularity of the design matrix. Therefore, penalized likelihood methods are proposed to cope with the problems that arise from high dimensionality. In the above settings, the generalized form of the objective function is

$$l(\beta) - \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{1.1}$$

where $l(\beta)$ is the log-likelihood function, and $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ [40].

The variable selection and regression coefficient estimation can be done simultaneously by maximizing the penalized likelihood function (1.1). When $p_\lambda(|\beta|) = \lambda|\beta|^q$, the regression is called penalized L_q-regression. This generalized form includes some popular variable selection methods, such as penalized L_1-regression (*Lasso*) and penalized L_2-regression(ridge regression)[19, 30].

1.2.1 Lasso

When q=1, the *Lasso* estimate is defined as follows

$$\hat{\beta}^{Lasso} = \underset{\beta}{argmax} \left\{ l(\beta) - \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{1.2}$$

There is no closed-form solution for *Lasso* because the solution is nonlinear in Y, but there several algorithms have been proposed for maximizing the penalized likelihood function numerically. Osborne et al. converted the maximization problem into a quadratic programming problem and solved it in an iterative way [42]. Efron et al. proposed the least angle regression (LARS) algorithm, which efficiently produces the entire path of *Lasso* estimates for all values of λ [43].

The tuning parameter λ controls the amount of regularization. As $\lambda \to 0$, we obtain the least squares solutions; as $\lambda \to \infty$, all the coefficients are approximately estimated as zeros. When λ is sufficiently but reasonably large, *Lasso* will cause some of the coefficients to be exactly zero, leading to a sparse interpretable model with excellent prediction accuracy. However, *Lasso* has limited capacity for selecting a consistent model, and the selected model tends to give biased estimates for the large coefficients and may have many false positives [19, 40, 41].

In 2001, Fan and Li proposed oracle properties as criteria for good penalty functions. [32] The oracle properties include

1. *Unbiasedness*: The resulting estimator is nearly unbiased especially when the true unknown parameter is large;

2. *Sparsity*: The resulting estimator automatically shrinks the sufficiently small estimated coefficients to zero to accomplish variable selection and to reduce model complexity; and

3. *Continuity*: The resulting estimator is continuous in the data to reduce instability in model prediction.

Given the inconsistency and biasedness of *Lasso*, the oracle properties do not hold [28, 32].

1.2.2 Adaptive Lasso

Adaptive *Lasso* (*adaLasso*), a weighted version of the *Lasso*, was proposed by Zou, where the adaptive weights ω_j's are adopted in the L_1 regularized regression to penalize different coefficients [28]. The *adaLasso* estimates are defined as

$$\hat{\beta}^{adaLasso} = \underset{\beta}{argmax} \left\{ l(\beta) - \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \right\}. \tag{1.3}$$

The adaptive weight vector $\boldsymbol{\omega}$ is usually chosen as $|\hat{\beta}|^{-\gamma}$, where $\gamma > 0$ and $\hat{\beta}$ is a root-n consistent estimator of β_0. Typically, ordinary least squares (OLS) estimates are used when no collinearity is assumed, and L_2 regression estimates are used when correlation exists among predictors. Similar to *Lasso*, the *adaLasso* estimates can be obtained by the LARS algorithm [28, 29, 43].

The *adaLasso* yields consistent estimates of the parameters while retaining the attractive property of the *Lasso*. Furthermore, it has been shown to enjoy the oracle properties [28].

However, *adaLasso* may not be able to deal with correlated predictors efficiently for moderate sample sizes [29], and has one more parameter (γ) to tune compared to *Lasso*.

1.2.3 Elastic Net

Zou and Hastie (2005) introduced the elastic net, a linear combination of L_1 and L_2 penalties, as equation (1.4) shows [31].

$$\hat{\beta}^{EN} = \underset{\beta}{argmax} \left\{ l(\beta) - (1-\alpha) \sum_{j=1}^{p} |\beta_j| - \alpha \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{1.4}$$

where the function $(1-\alpha)|\beta| + \alpha\beta^2$ is called the elastic net penalty for $\alpha \in [0, 1)$. When α=0, the elastic net becomes *Lasso* regression; when α=1, it reduces to simple ridge regression [31].

Maximizing equation (1.4) is equivalent to a lasso-type optimization problem, so elastic net estimates can also be obtained by LARS. Moreover, a LARS-based algorithm called LARS-EN was proposed by Zou to solve the elastic net more efficiently [31]. The L_2 penalty in the elastic net encourages the simultaneous selection or elimination of strongly correlated variables for the model, which is called a "grouping effect". However, the elastic net does not reveal the underlying group structure in its solution and does not possess oracle properties [44].

1.2.4 Smoothly Clipped Absolute Deviation (SCAD)

In L_q-regression, the penalty with $q = 1$ does not satisfy the condition of unbiasedness; the penalty with $q > 1$ does not satisfy the sparsity condition; and the penalty with $0 \leq q < 1$ does not satisfy the continuity condition. For these reasons, Fan and Li (2001) introduced the smoothly clipped absolute deviation (SCAD), which has the following form[40]:

$$\hat{\beta}^{SCAD} = \underset{\beta}{argmax} \left\{ l(\beta) - \sum_{j=1}^{p} [\lambda^2 - (|\beta_j| - \lambda)^2 I(|\beta_j| < \lambda)] \right\}. \tag{1.5}$$

The *SCAD* penalty is $P_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$, and its first-order derivative is $P'_\lambda(|\beta|) = \lambda I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{a-1} I(|\beta| > \lambda)$, where a is suggested to be 3.7. The SCAD penalty satisfies the aforementioned oracle properties [32].

Since the *SCAD* penalty is nonconcave and nondifferentiable, it cannot be solved by convex optimization. Several algorithms are proposed to optimize nonconcave penalized likelihood functions, including local quadratic approximation (LQA) and local linear approximation (LLA)[32, 45].

The LQA algorithm proposed by Fan and Li [32] locally approximates the objective function by a quadratic function. Given an initial value $\beta^{(k)}$, the penalty function $P_\lambda(|\beta|)$ can be locally approximated by a quadratic function, as shown by (1.6).

$$P_\lambda(|\beta|) \approx P_\lambda(|\beta^{(k)}|) + \frac{1}{2}\frac{P'_\lambda(|\beta^{(k)}|)}{|\beta^{(k)}|}[\beta^2 - \beta^{(k)2}]. \tag{1.6}$$

With this approximation, maximizing the penalized likelihood can be converted to a least squares problem with a closed-form solution. However, none of the elements of $\beta^{(k)}$ shrink exactly to zero. Instead, the very small coefficients need to be set to 0 manually [32, 40, 45].

To solve this problem, Zou and Li [45] proposed a better approximation, LLA, which can be represented in the following form:

$$P_\lambda(|\beta|) \approx P_\lambda(|\beta^{(k)}|) + P'_\lambda(|\beta^{(k)}|)(|\beta| - |\beta^{(k)}|). \tag{1.7}$$

The LLA estimators at each iteration naturally adopt a sparse representation. Therefore, Zou and Li advocate the one-step LLA estimator for the final estimates. The one-step estimator has been proven to be efficient as the fully iterative estimator and enjoys the oracle properties, provided that the initial values and tuning parameters are appropriately chosen. Furthermore, it can dramatically reduce the computational cost [45].

The commonly used penalized regression methods satisfactorily handle the $p \gg n$ problem. However, these variable selection approaches simply aggregate genetic variables derived from different platforms as covariates in regression models. By doing so, these methods lack the ability to consider the discrepancy in different platforms. They also fail to detect the correlation structure among the variables [12, 46].To address these problems, I apply the regularization method to the proposed shared informative factor model to integrate multi-platform data and select the genetic variables that

are associated with disease phenotype simultaneously. Specifically, *SCAD* is adopted for its good theoretical properties and promising empirical results.

1.3 Canonical correlation analysis (CCA)-based methods

This section consists of a brief review of several commonly used CCA-based methods: sparse *CCA*, supervised CCA (*sCCA*), multiple CCA (*mCCA*), and collaborative regression (*CollRe*) [34, 35, 36, 39].

Assuming n samples have two standardized datasets $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ of dimensions $n \times p_1$ and $n \times p_2$, CCA seeks the linear combinations (or canonical variates) of the variables in $\boldsymbol{X}_1$ and the variables in $\boldsymbol{X}_2$ that are maximally correlated with each other. This problem is equivalent to determining $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ by maximizing the *CCA criterion* [35]

$$\underset{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}{max}\ \boldsymbol{\omega}_1^T X_1^T X_2 \boldsymbol{\omega}_2 \text{ subject to } \boldsymbol{\omega}_1^T \boldsymbol{X}_1^T \boldsymbol{X}_1 \boldsymbol{\omega}_1 = \boldsymbol{\omega}_2^T \boldsymbol{X}_2^T \boldsymbol{X}_2 \boldsymbol{\omega}_2 = 1, \tag{1.8}$$

where $\boldsymbol{\omega}_1 \in \mathbb{R}^{p_1}$ and $\boldsymbol{\omega}_2 \in \mathbb{R}^{p_2}$ are canonical vectors (or weights). The optimal canonical vectors identify the genetic variables in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ that maximize the correlation between linear combinations $\boldsymbol{X}_1\boldsymbol{\omega}_1$ and $\boldsymbol{X}_2\boldsymbol{\omega}_2$. There are closed-form solutions for optimal canonical vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$.

CCA limits the probability of committing type I errors because it allows for simultaneous comparisons among the variables, which avoids inflated p-values from multiple testing [37]. However, *CCA* may fail when the relationship between the canonical variates is nonlinear. In addition, since this method is based on the correlation coefficient, it is very sensitive to outliers [38].

1.3.1 Sparse CCA

In genetic studies, the number of features of genomic data usually greatly exceeds the number of samples. *CCA* cannot be directly applied to solve this high-dimensional

problem. Therefore, sparse *CCA* has been proposed [35, 36]. In sparse *CCA*, penalty functions $P_{1,\lambda_1}(\cdot)$ and $P_{2,\lambda_2}(\cdot)$ are incorporated to determine the sparse canonical vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ by maximizing the *sparse CCA criterion*:

$$\underset{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}{max}\ \boldsymbol{\omega}_1^T X_1^T X_2 \boldsymbol{\omega}_2 \text{ subject to } ||\boldsymbol{\omega}_1||^2 \leqslant 1, ||\boldsymbol{\omega}_2||^2 \leqslant 1, P_{1,\lambda_1}(\boldsymbol{\omega}_1) \leqslant c_1, P_{2,\lambda_2}(\boldsymbol{\omega}_2) \leqslant c_2. \tag{1.9}$$

Typically, convex penalty functions $P_{1,\lambda 1}(\cdot)$ and $P_{2,\lambda 2}(\cdot)$ (e.g., *Lasso*, fused *Lasso*) are used to render (1.9) to a convex problem. The convex problem can be solved by an iterative algorithm, which maximizes the objective function (1.9) by fixing $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ alternately in each step. The tuning parameters λ_1 and λ_2 are determined by a permutation-based algorithm [35].

1.3.2 Sparse multiple CCA (mCCA)

The availability of multi-platform biological data makes it necessary to investigate the association among multiple datasets. Sparse *CCA* can be easily extended to sparse multiple CCA (*mCCA*) to accommodate the complex relationship among multiple datasets.

Suppose we have n samples with K standardized data sets denoted by $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_K$ with $p_i$variables in each dataset. Similar to (1.9), sparse *mCCA* can be demonstrated by [35]:

$$\max_{\boldsymbol{\omega}_1,\cdots\boldsymbol{\omega}_K} \sum_{i<j} \boldsymbol{\omega}_i^T \boldsymbol{X}_i^T \boldsymbol{X}_j \boldsymbol{\omega}_j \text{ subject to } ||\boldsymbol{\omega}_i||^2 \leqslant 1, P_i(\boldsymbol{\omega}_i) \leqslant C_i, \forall i, \tag{1.10}$$

where $\boldsymbol{\omega}_i \in \mathbb{R}^{p_i}$ and P_i's are convex penalty functions. The optimal canonical vectors $\boldsymbol{\omega}_i$ can be obtained by a similar iterative procedure. In each iteration step, $\boldsymbol{\omega}_i$ is obtained by fixing $\boldsymbol{\omega}_j$s for all $j \neq i$. The algorithm repeats until a convergence criterion is reached [35].

1.3.3 Sparse supervised CCA (sCCA)

The above methods focus only on the association among genetic variables, without making use of the disease phenotype or clinical outcome (e.g., smoking status, survival). Sparse supervised *CCA* (*sCCA*) is thus proposed to study the association of genetic variables with the clinical outcome.

Suppose that a quantitative clinical outcome $\boldsymbol{y} \in \mathbb{R}^n$ is available in addition to genetic matrices $\boldsymbol{X}_i$'s, then sparse *sCCA* is defined to seek the linear combination of the variables in $\boldsymbol{X}_i$'s that are highly correlated with each other and associated with the outcome $\boldsymbol{y}$ as follows:

$$\max_{\boldsymbol{\omega}_1,\cdots,\boldsymbol{\omega}_K} \sum_{i<j} \boldsymbol{\omega}_i^T X_i^T X_j \boldsymbol{\omega}_j \text{ subject to } ||\boldsymbol{\omega}_i||^2 \leqslant 1, P_i(\boldsymbol{\omega}_i) \leqslant C_i, \boldsymbol{\omega}_{il} = 0 \ \forall l \notin \boldsymbol{Q}_i, \quad (1.11)$$

where $\boldsymbol{Q}_i$ is the set of features in $\boldsymbol{X}_i$ that are correlated with $\boldsymbol{y}$. The optimal canonical vectors of sparse *sCCA* can be obtained by a similar iterative algorithm, as described in previous sections.

1.3.4 Collaborative regression (CollRe)

A novel form of supervised CCA, collaborative regression (*CollRe*), is proposed to study the relationship between the response variable and genetic variables from different assays [39].

It is assumed that we have n samples, each of which has two genetic matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ of dimensions $n \times p_x$ and $n \times p_z$ and a response vector $\boldsymbol{y}$. The coefficients $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_z$, which contain the weights of the features for the outcome, can be obtained by minimizing the objective function [39]:

$$J(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z) = \frac{b_{xy}}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_x||^2 + \frac{b_{zy}}{2}||\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta}_z||^2 + \frac{b_{xz}}{2}||\boldsymbol{X}\boldsymbol{\theta}_x - \boldsymbol{Z}\boldsymbol{\theta}_z||^2, \quad (1.12)$$

where b_{xy}, b_{yz},and b_{xz} are parameters that control the relative importance of the three terms in the objective. For simplicity, b_{xy}, b_{yz},and b_{xz} can be set as 1. The key idea

of *CollRe* is to force the two genetic matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ to make similar contributions to prediction by penalizing $||\boldsymbol{X}\boldsymbol{\theta}_x - \boldsymbol{Z}\boldsymbol{\theta}_z||$, the difference between the fitted linear predictors $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_x||$ and $||\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta}_z||$.

It is also easy to extend *CollRe* to account for sparse genetic matrices by defining a penalized version of *CollRe* (*pCollRe*). The *pCollRe* coefficients can be obtained by minimizing the following objective [39]:

$$F(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z) = \frac{b_{xy}}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}_x||^2+\frac{b_{zy}}{2}||\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\theta}_z||^2+\frac{b_{xz}}{2}||\boldsymbol{X}\boldsymbol{\theta}_x-\boldsymbol{Z}\boldsymbol{\theta}_z||^2+P^x_{\lambda_1}(\boldsymbol{\theta}_x)+P^z_{\lambda_2}(\boldsymbol{\theta}_z), \tag{1.13}$$

where $P^x_{\lambda_1}(\boldsymbol{\theta}_x)$ and $P^z_{\lambda_2}(\boldsymbol{\theta}_z)$ are convex penalty functions (e.g., *Lasso*, ridge, and fused *Lasso*). *pCollRe* is more efficient than sparse *CCA*, since the tuning parameters λ_1 and λ_2 can be determined by the efficient LARS algorithm rather than the permutation-based algorithm. However, this method is not well suited for prediction [39].

CCA provides an efficient way of performing an integrative analysis for high-dimensional genetic variables from two platforms. The extensions of *CCA* allow for the integration of more than two data sets and the incorporation of an outcome into the analysis. However, the series of *CCA*-based methods have the inherent drawback of not being able to capture the nonlinear relationships between canonical variates.

1.4 Statistical methods for biological knowledge incorporation

A penalized regression-based method is used for incorporating known pathway information due to its easy application to *SIFORM*. In this section, I review the following specific penalized regression-based methods that incorporate pathway information: the graph constrained estimation (*Grace*) method [21], an adaptive version of graph constrained estimation (*aGrace*) [22], a generalized class of the network-

constrained penalty proposed by Pan et al. (grouped L_γ-norm penalty) [5], and a less stringent version of Pan's proposed penalty function ($TTLP_I$ and $LTLP_I$) [23].

1.4.1 Graph constrained estimation (Grace)

Recall the graphical models introduced in Section 1.1. In the graphical model, genetic variables are indicated by a set of vertices, and two directly connected vertices are represented as $j \sim j'$. The graph constrained estimation (*Grace*) method proposed by Li and Li (2008) assumes that the genes connected within a subnetwork have similar functions and therefore smoothed regression coefficients. This idea can be illustrated by a penalty function with two terms, as equation (1.14) illustrates. The first term is a L_1-penalty for variable selection, and the second term is used to smooth the weighted coefficients over the network [21].

$$P^{Grace}_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j\sim j'} (\frac{\beta_j}{\sqrt{d_j}} - \frac{\beta_{j'}}{\sqrt{d_{j'}}})^2, \tag{1.14}$$

where d_j is the degree (the number of edges) of node j. The degrees scale the coefficients to allow the highly connected genes, the hub genes, to have larger coefficients so that small changes in the expressions of such genes can lead to large changes in the response [21].

Maximizing the penalized regression with penalty function (1.14) is equivalent to solving a *Lasso*-type optimization problem and thus enjoys the computational advantage of *Lasso* [21].

1.4.2 Adaptive graph constrained estimation (aGrace)

In some applications, two connected genes might be negatively correlated with the phenotypes and therefore have opposite signs in their regression coefficients. However, the above *Grace* method fails to smooth the genes that are linked within the subnetwork and which have different regression coefficients signs. To account for the

sign differences, Li and Li (2010) [22] proposed an adaptive version of *Grace* method (*aGrace*). In this method, $|\beta_j|/\sqrt{d_j} = |\beta_{j'}|/\sqrt{d_{j'}}$ for $j \sim j'$ is encouraged in the penalty function, as equation (1.15) shows.

$$P_{\lambda_1,\lambda_2}^{aGrace}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j\sim j'} (\frac{sign(\tilde{\beta}_j)\beta_j}{\sqrt{d_j}} - \frac{sign(\tilde{\beta}_{j'})\beta_{j'}}{\sqrt{d_{j'}}})^2, \tag{1.15}$$

where $\tilde{\beta}_j$ is an initial estimate. Typically, the OLS estimate is employed for $p < n$, and the elastic net estimate is employed for $p \geqslant n$. The main idea is to use $sign(\tilde{\beta}_j)$ to estimate $sign(\beta_j)$, which, however, may not work well for high-dimensional data: since we do not even know which β_j's are 0 for variable selection, it is more difficult to estimate their signs [22, 23].

The coordinate descent algorithm is used to optimize the penalized likelihood function that uses (1.15) as a penalty function [47].

1.4.3 Grouped L_γ-norm penalty

In addition to the incapability of capturing the connected genes with opposite regression coefficient signs, the *Grace* approach also fails to accomplish grouped variable selection. To address this problem, Pan et al. (2010) [5] proposed a grouped L_γ-norm penalty to automatically realize grouped variable selection and exploit grouping effects. Given a pre-specified $\gamma > 1$, this class of penalties can be illustrated by equation (1.10):

$$P_{\lambda,\gamma}(\boldsymbol{\beta}) = \lambda \sum_{j\sim j'} [(\frac{|\beta_j|}{\omega_j})^\gamma + (\frac{|\beta_{j'}|}{\omega_{j'}})^\gamma]^{\frac{1}{\gamma}}, \tag{1.16}$$

where ω_j is a user-specified weight to realize different types of shrinkages. Usually, three choices are considered for ω_j: (i) $\omega_j = d_j^{\frac{\gamma+1}{2}}$; (ii)$\omega_j = d_j$; (iii) $\omega_j = d_j^\gamma$, which lead to three different types of smoothness on the parameters [5].

Pan et al. proposed a slightly modified generalized boosted Lasso (GBL) algorithm to maximize the penalized likelihood function [48]. The GBL algorithm yields

an approximate solution path $\beta(\hat{\lambda})$ through a coordinate-wise search and repeated calculations of the objective function. Among the set of β's at a finite number of tuning parameter values, the final coefficient estimates $\beta_{\lambda_{(k_0)}}$ are obtained at tuning parameter $\lambda_{(k_0)}$, which minimizes the prediction mean squared error [5].

The empirical results of this method demonstrate better performance in variable selection than that of *Grace*, but its parameter estimates may be severely biased [23].

1.4.4 Truncated lasso penalty (TLP)-based penalty

All the above methods assume the smoothness of weighted $\boldsymbol{\beta}$ or $|\boldsymbol{\beta}|$ over the network. This assumption may be too stringent in some applications, because the genes connected within the same subnetwork may be correlated with the outcome with different effect sizes. Hence, Kim et al. proposed a new network-based penalty without the assumption of smoothness on regression coefficients. Instead, the proposed penalty only assumes that two connected genes are more likely to participate (or not participate) together in the same biological process than two distant genes, as equation (1.17) shows [23].

$$P(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} I(|\beta_j| \neq 0) + \lambda_2 \sum_{j \sim j'} |I(\frac{|\beta_j|}{\omega_j} \neq 0) - I(\frac{|\beta_{j'}|}{\omega_{j'}} \neq 0)|, \tag{1.17}$$

where the first penalty term is for variable selection and the second one encourages the simultaneous selection (or elimination) of connected nodes in a subnetwork. Since the indicator function $I(\cdot)$ is not continuous, Kim et al. adopted a truncated *Lasso* penalty (TLP), $J_\tau(|z|) = min(\frac{|z|}{\tau}, 1)$, which approximates $I(|z| \neq 0)$ as $\tau \to 0^+$. The tuning parameter τ determines the degree of approximation [23, 49]. The TLP is applied to (1.17) as a computational surrogate of $I(|z| \neq 0)$ and therefore leads

to the following penalty function $TTLP_I$, with a TLP for variable selection and a TLP-based penalty for grouping the genetic variables [23]:

$$P^{TTLP}_{\tau,\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} J_\tau(|\beta_j|) + \lambda_2 \sum_{j\sim j'} |J_\tau(\frac{|\beta_j|}{\omega_j}) - J_\tau(\frac{|\beta_{j'}|}{\omega_{j'}})|. \tag{1.18}$$

For computational efficiency, Kim et al. proposed another penalty function $LTLP_I$ (1.19), which uses *Lasso* rather than TLP for variable selection [23]:

$$P^{LTLP}_{\tau,\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j\sim j'} |J_\tau(\frac{|\beta_j|}{\omega_j}) - J_\tau(\frac{|\beta_{j'}|}{\omega_{j'}})|. \tag{1.19}$$

Since the TLP function is non-convex, difference convex (DC) programming is used to decompose the non-convex function into a difference of two convex functions, which then can be solved by convex optimization [23, 50].

The existing pathway knowledge can be structured by the network-based penalty functions introduced in Section 1.4 and then can be easily applied to the framework of *SIFORM*. Specifically, I adopt the idea of TLP for its flexibility and its better approximation to reality. I propose a penalty function that uses $SCAD$ for variable selection and TLP for simultaneous selection of connected variables. The proposed penalty with its computation issues and theoretical properties is discussed in Chapter 3.

2. Shared Informative Factor Models for Integration of Multi-platform Bioinformatic Data (SIFORM)

In this chapter, I propose a statistical framework of shared informative factor models (*SIFORM*) to integrate multi-platform *omic* data and explore their association with a disease phenotype. The common disease-associated sample characteristics across different data types can be captured through the shared structure space, while the corresponding weights of genetic variables directly index the strengths of their association with the phenotype.

Extensive simulation studies demonstrate the superior performance of *SIFORM* in terms of biomarker detection accuracy compared to the performance of several popular regularized regression methods and *CCA*-based methods. I also applied *SIFORM* to two lung adenocarcinoma datasets from The Cancer Genome Atlas (TCGA) database to jointly explore the associations of mRNA expression and protein expression with smoking status and survival rate. Both studies identified many biomarkers that belong to key pathways for lung tumorigenesis, some of which are known to show differential expression across smoking levels or between long-term survivors and short-term survivors. *SIFORM* also discovered potential biomarkers that reveal different mechanisms of lung tumorigenesis between light smokers and heavy smokers and prognostic biomarkers for survival.

2.1 Motivation

High-dimensional *omic* data derived from different platforms have been extensively used to facilitate comprehensive understanding of disease mechanisms [1, 2].

Numerous studies have integrated multi-platform *omic* data [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]; however, few have efficiently and simultaneously addressed the problems that arise from high dimensionality and complex correlations.

For instance, the penalized regression-based variable selection methods introduced in Section 1.2 [19, 28, 31, 32] satisfactorily address the extremely high dimensionality of the covariate space, which is much larger than the number of samples. However, these methods aggregate genetic variables from different platforms as covariates in regression models without accounting for their correlation. The *CCA*-based methods introduced in Section 1.3 [35, 36, 39] efficiently integrate genetic variables from multiple platforms but do not achieve satisfactory predictive performance and fail to handle the nonlinear relationship among multi-platform data.

To address this problem, I propose a statistical framework, *SIFORM*, to integrate multiple types of *omic* data, and investigate the association of the integrated data with a disease-associated response variable. In contrast to the conventional factor models, I incorporate the disease phenotype information into the factor space to detect genetic variables that interact with the response variable. I also assume a common structured factor across multiple data types for the purpose of detecting different levels of the important disease-associated genetic variables. The proposed framework lays a foundation for incorporating existing biological knowledge into statistical models to improve the biological relevance of results.

2.2 Methodology

2.2.1 A framework of shared informative factor models

In this section, I describe a new model framework to explore associations between a disease phenotype and high-throughput genetic data generated from multiple (≥ 2) platforms. Herein, I focus on two-platform data for the purpose of demonstration.

Let the $n \times p_1$ matrix $\mathbf{X_1}$ and $n \times p_2$ matrix $\mathbf{X_2}$ denote two data matrices containing the intensity measurements of p_1 genetic variables obtained from platform 1 and those of p_2 genetic variables obtained from platform 2, respectively. Correspondingly, we let the length-n vector $\mathbf{y}$ denote the phenotypes of n subjects (e.g., smoking status, cancer subtype). The system is built upon the generalized linear models, considering exponential family of distributions for genetic variables in $\mathbf{X_1}$ and $\mathbf{X_2}$ (e.g., continous measurements, count data). The transformed mean functions of the expression intensities of genetic variables can be expressed via canonical link functions g_1 and g_2, respectively corresponding to $\mathbf{X_1}$ and $\mathbf{X_2}$, as follows,

$$\begin{aligned} g_1\left\{E(\mathbf{X_1})\right\} &= \boldsymbol{\alpha}_1 + \boldsymbol{CA}, \\ g_2\left\{E(\mathbf{X_2})\right\} &= \boldsymbol{\alpha}_2 + \boldsymbol{CB}. \end{aligned} \tag{2.1}$$

For the demonstration, I focus on continuously measured genetic variables (e.g., gene expression, protein expression, DNA copy number), for which the identity link $g_1(\mu) = g_2(\mu) = \mu$ is used. I also assume the genetic variables in $\mathbf{X_1}$ and $\mathbf{X_2}$ follow normal distributions (posterior to transformation when appropriate). The $n \times 1$ parameter vectors $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ respectively correspond to baseline sample effects in two sets of genetic data. The jth columns of the $n \times p_1$ residual matrix $\boldsymbol{\varepsilon_1}$ of $\mathbf{X_1}$ and the $n \times p_2$ residual matrix $\boldsymbol{\varepsilon_2}$ of $\mathbf{X_2}$ follow mean-0 normal distributions with the respective variances σ^2_{1j} and σ^2_{2j}. Our interest is drawn to the multiplicative terms $\boldsymbol{CA}$ and $\boldsymbol{CB}$, which capture the associations between the phenotype and genetic variables. The proposed framework has two main distinctions from the conventional factor models that have similar forms. First, the $n \times 1$ parameter vector $\boldsymbol{C}$ is structured to deliver the phenotype information. In the case of a K-categorical phenotype (e.g., K levels of smoking status), $\boldsymbol{C} = \{c_1, ..., c_1, ..., c_l, ..., c_l, ..., c_K, ..., c_K\}^T$, with c_l indicating the common effect that the subjects in the l^{th} phenotype contribute to the associations between the phenotypes and genetic variables. Second, the same $\boldsymbol{C}$ is employed in these two models to represent the common intrinsic sample characteristics in the two

sets of genetic data. The length-p_1 vector $\boldsymbol{A}$ and length-p_2 vector $\boldsymbol{B}$ contain the weights that the genetic variables contribute to the structured sample characteristics shared by the two sets of genetic data. The elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ are called association scores. This new system is called shared informative factor models (*SIFORM*). For model identifiability, the constraint $\sum_{i=1}^{n} c_i^2 = 1$ is imposed on $\boldsymbol{C}$. I also follow the established practice of standardizing the intensity values of each genetic variable prior to the downstream analysis.

2.2.2 Parameter Estimation

In the above-described framework, the log-likelihood of the two sets of observed genetic data is

$$\begin{aligned} l = \quad & \sum_{j=1}^{p_1}\sum_{i=1}^{n}\left\{-\frac{(x_{1ij}-\alpha_{1i}-c_l\cdot a_j)^2}{2\sigma_{1j}^2}-\frac{1}{2}log(2\pi\sigma_{1j}^2)\right\} \\ & +\sum_{j=1}^{p_2}\sum_{i=1}^{n}\left\{-\frac{(x_{2ij}-\alpha_{2i}-c_l\cdot b_j)^2}{2\sigma_{2j}^2}-\frac{1}{2}log(2\pi\sigma_{2j}^2)\right\}. \end{aligned} \tag{2.2}$$

To identify important genetic variables associated with the phenotype of interest, I also impose a sparsity regularization on the elements of the parameter vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ in model estimation. I adopt the *SCAD* penalization method [32], which uses symmetric penalty functions that are non-concave on $(0,\infty)$ to simultaneously select the variables and estimate the coefficients. *SCAD* has been shown to have good theoretical properties and to offer promising empirical results.

The parameter estimates are obtained via minimizing the following objective function, which is simply the negative penalized log-likelihood

$$S = -l + \sum_{j=1}^{p_1} P_{\delta_1}(|a_j|) + \sum_{j=1}^{p_2} P_{\delta_2}(|b_j|), \tag{2.3}$$

where $P_\delta(.) = \delta^2 - (.-\delta)^2 I(. < \delta)$, and its first-order derivative is $P'_\delta(.) = \delta I(. \leqslant \delta) + \frac{(a\delta - .)_+}{a-1} I(. > \delta)$. I take the value of $a = 3.7$ as suggested by [32]. Herein, δ_1 and

δ_2 are the penalty tuning parameters that respectively correspond to the two sets of genetic data.

I adopt the local linear approximation of [45] for the sparsity penalty term and find a convex function to implement the efficient majorization-minimization (MM) algorithm [51]. The MM algorithm, which stands for the minorize-maximize or majorize-minimize algorithm, is a class of algorithms for finding a maximizer or minimizer of non-differentiable penalty functions in an iterative way. Specifically, the minimization problem consists of two steps at the k^{th} iteration: majorization and minimization. In the majorization step, a differentiable function $g(\theta|\theta^{(k)})$ is created as a surrogate for the non-differentiable objective function $f(\theta)$. The function $g(\theta|\theta^{(k)}) \geqslant f(\theta)$ is said to majorize function $f(\theta)$ at $\theta^{(k)}$, and has the following properties. In the minimization step, the differentiable majorization function $g(\theta|\theta^{(k)})$ is minimized rather than the actual objective function $f(\theta)$. If $\theta^{(k+1)})$ denotes the minimizer of $g(\theta|\theta^{(k)})$, then it can be shown that $f(\theta^{(k+1)}) \geqslant f(\theta^{(k)})$. This is called the descent property of the MM algorithm, which guarantees the numerical stability of the algorithm [51].

Using $P_{\delta_1}(|a_j|)$ as an example, the local linear approximation can be expressed as

$$P_{\delta_1}(|a_j|) \approx P_{\delta_1}(|a_j^{(k)}|) + P'_{\delta_1}(|a_j^{(k)}|)(|a_j| - |a_j^{(k)}|), \tag{2.4}$$

where $a_j^{(k)}$ is the value of a_j estimated at step k. Then the majorization function of the local linear approximated penalty is

$$G_{\delta_1}(|a_j|) = P'_{\delta_1}(|a_j^{(k)}|)\frac{a_j^2 + a_j^{(k)2}}{2|a_j^{(k)}|} + P_{\delta_1}(|a_j^{(k)}|) - P'_{\delta_1}(|a_j^{(k)}|)(|a_j^{(k)}|). \tag{2.5}$$

The majorization function $G_{\delta_2}(|b_j|)$ for $P_{\delta_2}(|b_j|)$ can be obtained in similar way.

Thus, the parameter estimates are obtained by iteratively minimizing the following objective function,

$$S^* = -l + \sum_{j=1}^{p_1} G_{\delta_1}(|a_j|) + \sum_{j=1}^{p_2} G_{\delta_2}(|b_j|).$$

At the $(k+1)^{th}$ step, the closed-form penalized log-likelihood estimators for the elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ can be obtained as

$$\hat{a_j}^{(k+1)} = \frac{|a_j^{(k)}| \sum_{i=1}^{n} c_i(x_{1ij} - \alpha_{1i})}{|a_j^{(k)}| + \sigma_{1j}^2 P'_{\delta_1}(|a_j^{(k)}|)} \tag{2.6}$$

and

$$\hat{b_j}^{(k+1)} = \frac{|b_j^{(k)}| \sum_{i=1}^{n} c_i(x_{2ij} - \alpha_{2i})}{|b_j^{(k)}| + \sigma_{2j}^2 P'_{\delta_2}(|b_j^{(k)}|)}. \tag{2.7}$$

The closed-form solutions of all the other parameter estimates are obtained via taking the partial derivative of log-likelihood function (2.2), as the following equations show:

$$\hat{\sigma}_{1j}^2 = \frac{\sum_{i=1}^{n}(x_{1ij} - \hat{\alpha}_{1j} - \hat{c}_i \cdot \hat{a}_j)^2}{n}; \tag{2.8}$$

$$\hat{\sigma}_{2j}^2 = \frac{\sum_{i=1}^{n}(x_{2ij} - \hat{\alpha}_{2j} - \hat{c}_i \cdot \hat{b}_j)^2}{n}; \tag{2.9}$$

$$\hat{\alpha}_{1i}^2 = \frac{\sum_{j=1}^{p_1} x_{1ij} - \hat{c}_i \sum_{j=1}^{p_1} \hat{a}_j}{p_1}; \tag{2.10}$$

$$\hat{\alpha}_{2i}^2 = \frac{\sum_{j=1}^{p_2} x_{2ij} - \hat{c}_i \sum_{j=1}^{p_2} \hat{b}_j}{p_2}; \tag{2.11}$$

$$\hat{c}_l = \frac{\sum_{j=1}^{p_1} \sum_{i=n_{l-1}+1}^{n_l} \frac{\hat{a}_j}{\hat{\sigma}_{1j}^2}(x_{1ij} - \hat{\alpha}_{1i}) + \sum_{j=1}^{p_2} \sum_{i=n_{l-1}+1}^{n_l} \frac{\hat{b}_j}{\hat{\sigma}_{2j}^2}(x_{2ij} - \hat{\alpha}_{2i})}{(n_l - n_{l-1})(\sum_{j=1}^{p_1} \frac{\hat{a}_j^2}{\hat{\sigma}_{1j}^2} + \sum_{j=1}^{p_2} \frac{\hat{b}_j^2}{\hat{\sigma}_{2j}^2})}, \tag{2.12}$$

where $l = 1, \cdots K$ indicates the smoking status l.

Algorithm 1 demonstrates the details of the iterative parameter estimation procedure. The convergence threshold ε is set at 10^{-5}. Extensive simulations show that the estimation seems to stabilize within 300 iterations. Therefore, the iterative procedure stops when either convergence is reached or 300 iterations are completed, whichever occurs first.

Algorithm 1 Iterative parameter estimation procedure

1. Input: X_1, X_2, y

2. Initialization:

2.1. $\alpha_1^{(0)} \leftarrow 0, \alpha_2^{(0)} \leftarrow 0$

2.2. $\sigma_{1j}^{(0)} \leftarrow var(X_{1(\cdot j)}), \sigma_{2j}^{(0)} \leftarrow var(X_{2(\cdot j)})$

2.3. Assign 1,$\cdots$,K to c_i according to the K-categorical phenotype y, then scale C to have $\sum_{i=1}^{n} c_i^2 = 1$. Use the scaled C as the initial value, $C^{(0)}$.

2.4. Use least square estimators for $A^{(0)}, B^{(0)}$: $a_j^{(0)} \leftarrow \sum_{i=1}^{n} c_i x_{1ij}, b_j^{(0)} \leftarrow \sum_{i=1}^{n} c_i x_{2ij}$

3. Do-while loop:

do{

Update 3.1-3.3 using closed-form estimators

3.1. $\alpha_{1,2}^{(k+1)} \leftarrow (X_1, X_2, \sigma_{1,2}^{2(k)}, C^{(k)}, A^{(k)}, B^{(k)})$

3.2. $\sigma_{1,2}^{2(k+1)} \leftarrow (X_1, X_2, \alpha_{1,2}^{(k+1)}, C^{(k)}, A^{(k)}, B^{(k)})$

3.3. $C^{(k+1)} \leftarrow (X_1, X_2, \alpha_{1,2}^{(k+1)}, \sigma_{1,2}^{2(k+1)}, A^{(k)}, B^{(k)})$

Update 3.4 using formula (2.6)

3.4. $A^{(k+1)} \leftarrow (X_1, C^{(k+1)}, \alpha_1^{(k+1)}, \sigma_1^{2(k+1)})$

Update 3.5 using formula (2.7)

3.5. $B^{(k+1)} \leftarrow (X_2, C^{(k+1)}, \alpha_2^{(k+1)}, \sigma_2^{2(k+1)})$

3.6. For each estimate, compute the difference between current and previous iteration steps:

$\Delta_1 = |\alpha_1^{(k+1)} - \alpha_1^{(k)}|, \Delta_2 = |\alpha_2^{(k+1)} - \alpha_2^{(k)}|, \Delta_3 = |\sigma_1^{(k+1)} - \sigma_1^{(k)}|, \Delta_4 = |\sigma_2^{(k+1)} - \sigma_2^{(k)}|, \Delta_5 = |A^{(k+1)} - A^{(k)}|, \Delta_6 = |B^{(k+1)} - B^{(k)}|, \Delta_7 = |C^{(k+1)} - C^{(k)}|$

} while$(\Delta_1 > \varepsilon || \Delta_2 > \varepsilon || \Delta_3 > \varepsilon || \Delta_4 > \varepsilon || \Delta_5 > \varepsilon || \Delta_6 > \varepsilon || \Delta_7 > \varepsilon)$

4. Output:$\hat{A} \leftarrow A^{(k+1)}, \hat{B} \leftarrow B^{(k+1)}, \hat{C} \leftarrow C^{(k+1)}$

2.2.3 Tuning Parameter Selection

The tuning parameters $\boldsymbol{\lambda} = (\delta_1, \delta_2)$ are important for calibrating the goodness-of-fit for the data and model sparsity. I adopt the Bayesian information criterion (BIC) [52], which is widely used in high-dimensional studies [53, 54], and defined as

$$BIC = -2l + log(n(p_1 + p_2)) \cdot df(\lambda),$$

where $df(\lambda)$ is proportional to the summation of the number of nonzeros in $\hat{\boldsymbol{A}}$ and the number of nonzeros in $\hat{\boldsymbol{B}}$.

The pair of (δ_1, δ_2) that achieves the smallest BIC value is obtained by using a two-dimensional grid search over a pre-determined space. A large value of $\boldsymbol{\lambda}$ may lead to an under-fitted model while a smaller $\boldsymbol{\lambda}$ leads to an over-fitted model. I use the range between 0 and 10 for each tuning parameter.

The BIC curve is expected to be convex over the tuning parameters. When the BIC curve is convex, it is very easy to find its minimum. Sometimes the BIC curve is L-shaped. When it is L-shaped, the point of maximum curvature is used as the minimum, which is called the knee of the curve. To find the knee, I locate the point on the curve that is furthest from a line fitted to the entire curve. This criterion considers all data points on the curve at the same time.

2.3 Simulation

2.3.1 Data description

Extensive simulations are conducted to assess the performance of the proposed *SIFORM* and compare its performance to those of three popular regularized regression methods: *SCAD*, *Lasso*, adaptive Lasso (*adaLasso*) and *CCA*-based methods: *mCCA* and *pCollRe*. These methods can be implemented directly using the respective R

packages *ncvreg*, *glmnet*, *parcor*, and *PMA*. The tuning parameters are selected by the 5-fold cross-validation procedure.

Seven simulation scenarios are used to study the various data generation models, inter-genetic marker dependence structures, and residual variances. In scenarios 1-3 and 5 - 6, the phenotype variable $\mathbf{Y}$ is categorized into four groups and two high-dimensional genetic profiling matrices, $\mathbf{X}_1$ and $\mathbf{X}_2$. In the three penalized regression methods, all the genetic variables in $\mathbf{X}_1$ and $\mathbf{X}_2$ are treated as the covariates to predict the phenotype. Note that multinomial logistic regression models are used for the response variable of smoking status in four categories. For *SCAD*, the union of all the nonzero coefficients identified in three separate logistic regression models is taken, treating the category of non-smoker as the reference group. In contrast, *SIFORM* conveniently uses $\boldsymbol{A}$ and $\boldsymbol{B}$ to identify the important biomarkers despite the different disease phenotypes. The performance of *SIFORM* is also investigated when data come from three platforms in scenarios 4 and 7.

In brief, scenarios 1-4 generate genetic data $\mathbf{X}_1$ and $\mathbf{X}_2$ from model (2.1), and scenarios 5-7 simulate discrete phenotype data from multinomial logistic regression models. The matrices $\mathbf{X}_k (k = 1, 2, 3)$ in scenarios 5-7, and $\boldsymbol{\varepsilon_k}$ (k=1, 2, 3) in scenarios 1-4 are simulated from multivariate mean-0 normal distributions. The corresponding variances are set to be 1 for all the multivariate normally-distributed variables in scenarios 1-4. In scenario 6, they are sampled from a lung cancer data set obtained from TCGA. In terms of the data dependence structure, the genetic variables in $\mathbf{X}_1$ and $\mathbf{X}_2$ are mutually independent in scenario 1; the genetic variables with nonzero coefficients are correlated with $\rho = 0.8$, and $14\% - 32\%$ of the remaining genetic variables are weakly correlated with $\rho = 0.2$ in scenarios 2-5. Throughout all the scenarios, sparse true biomarkers that are associated with the phenotype are assumed. Additional details for the five simulation scenarios follow.

- Scenario 1: I equally assign 120 samples to the four categories of the phenotype $\mathbf{y}$, which gives the corresponding $\boldsymbol{C} = (0.033, \cdots, 0.033, 0.067, \cdots, 0.067, 0.100, \cdots, 0.100, 0.133, \cdots, 0.133)$. Each sample has intensity measurements collected over 100 genetic variables in data set 1 ($\mathbf{X}_1$) and over 100 genetic variables in data set 2 ($\mathbf{X}_2$). In each data set, only the first 5 genetic variables are associated with the phenotype, with the sparse coefficient vectors $\boldsymbol{A}_{1\times 100} = (5, 7, 11, 15, 18, 0, \cdots, 0)$ and $\boldsymbol{B}_{1\times 100} = (4, 8, 10, 14, 20, 0, \cdots, 0)$. All the genetic variables are assumed to be mutually independent. The baselines $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ are set at 0, and the intensity measurements in $\mathbf{X}_1$ and $\mathbf{X}_2$ are simulated from model (2.1).

- Scenario 2: This scenario is the same as scenario 1, except that we assume a blockwise compound symmetry correlation structure among the genetic variables to mimic real studies.

- Scenario 3: In this higher dimensional case, 160 samples are equally assigned to four categories, and there are 1,000 variables in each of the two genetic data sets. Among the genetic variables, only the first 50 have nonzero coefficients in each data set. The data are generated in the same way as in scenario 2.

- Scenario 4: This scenario is the same as scenario 2, except that each sample has 100 additional genetic variables in data set 3 $\mathbf{X}_3$, with sparse coefficient vectors $\mathbf{M}_{1\times 100} = (3, 6, 9, 12, 17, 0, \cdots, 0)$.

- Scenario 5: I use a multinomial logistic regression model to generate the phenotype. I consider a total of 200 samples and 100 genetic variables, where the first 5 variables are true biomarkers, in each of $\mathbf{X_1}$ and $\mathbf{X_2}$. Letting the sparse coefficient vectors $\mathbf{A_{(1)1\times 100}} = (6, 14, 14, 24, 20, 0, \cdots, 0)$, $\mathbf{B_{(1)1\times 100}} = (4, 12, 10, 18, 18, 0, \cdots, 0)$, $\mathbf{A_{(2)1\times 100}} = (3, 14, 11, 24, 24, 0, \cdots, 0)$ and $\mathbf{B_{(2)1\times 100}} = (2, 10, 13, 24, 23, 0, \cdots, 0)$, $\mathbf{A_{(3)1\times 100}} = (2, 14, 15, 20, 25, 0, \cdots, 0)$ and $\mathbf{B_{(3)1\times 100}}$

$= (2, 10, 11, 21, 23, 0, \cdots, 0)$, we have the logit transformed predictor $\eta_{il} = log(\frac{P(Y_i=l)}{P(Y_i=4)}) = A_{(l)}X_{1i} + B_{(l)}X_{2i},\ l = 1, 2, 3$. The probability of $Y_i = l$ is $\mathbf{p_{il}} = \mathbf{e}^{\eta_{il}}/(\mathbf{1} + \sum_{\mathbf{k=1}}^{\mathbf{3}} \mathbf{e}^{\eta_{ik}})$. Accordingly, I sample Y_i from the multinomial distribution $MN(1, p_{i1}, p_{i2}, p_{i3}, p_{i4})$. The genetic intensity values follow multivariate mean-0 normal distributions. The generated data are fairly balanced. For example, a simulated data set contains 66, 40, 43, and 51 samples in the four respective categories.

- Scenario 6: I investigate a higher dimensional setting for the data generated under a multinomial logistic regression model, which is similar to scenario 4. I consider 230 samples, with $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively containing 3500 and 150 genetic variables. This dimensionality is comparable to that of the first TCGA lung cancer data set I used to illustrate the real application of the proposed method. Among the genetic variables, only the first 30 in $\mathbf{X}_1$ and the first 10 in $\mathbf{X}_2$ are the true predictors of the phenotype. As an example, a simulated data set produces 59, 47, 49, and 75 samples in the four respective phenotype categories.

- Scenario 7: I also investigate a 3-platform setting for the data generated under a multinomial logistic regression model. In this scenario, the same $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{A_{(1,2,3)1\times 100}}$ and $\mathbf{B_{(1,2,3)1\times 100}}$ are used as in scenario 5. The only difference is that each sample has 100 additional genetic variables in $\mathbf{X_3}$, with the sparse coefficient vectors $\mathbf{M_{(1)1\times 100}} = (3, 9, 12, 15, 22, 0, \cdots, 0)$, $\mathbf{M_{(2)1\times 100}} = (4, 8, 12, 20, 23, 0, \cdots, 0)$ and $\mathbf{M_{(3)1\times 100}} = (3, 10, 15, 19, 24, 0, \cdots, 0)$. Y_i is generated is the same way as described in scenario 5. As an example, a simulated data set contains 29, 36, 47, and 88 samples in the four respective categories.

2.3.2 Results

One hundred simulations are run for each scenario. I report the average value and standard error of the true positive rate (TPR), true negative rate (TNR), and false discovery rate (FDR) among the top genetic variables detected that have the largest values of $|\boldsymbol{A}|$ or $|\boldsymbol{B}|$ for *SIFORM*, and the largest absolute values of the regression coefficients for the other methods. For scenario 3 with 100 true biomarkers, I focus on the top 10 biomarkers detected by different methods. For all the other scenarios, I focus on the top 5 biomarkers detected.

I report the variable selection results in Table 2.1. The first four scenarios correspond to the generation of data from the shared informative factor model (2.1), considering independent and dependent genetic variables, different levels of dimensionality, and different numbers of platforms. The simulations demonstrate the consistently superior performance of the *SIFORM* method compared to those of the other five methods in terms of biomarker detection accuracy across the scenarios. The TPRs of *SIFORM* are strikingly higher than those of all the other methods in each scenario, and *SIFORM* maintains the smallest FDRs and comparable TNRs across all scenarios. Among the remaining methods, *pCollRe* generally has the best performance under every criteria, especially in the high-dimensional scenario 3. This is because both L_1- and L_2-norm penalties are used in *pCollRe* to smooth the estimates [39]. *mCCA* has an overall large variance, and its performance seems easily affected by dimension. Among the penalized regression methods, *SCAD* detects the highest number of true biomarkers while sacrificing false positives, and thus yields a high FDR. The remaining two penalized regression approaches, *Lasso* and *adaLasso*, detect fewer true biomarkers than the other two methods, and *adaLasso* produces a lower FDR than *Lasso* and *SCAD*.

Table 2.1: Comparison of TPRs, TNRs, and FDRs between *SIFORM* and five other methods under seven simulation scenarios.

Scenario	Method	TPR	TNR	FDR
Scenario 1	SIFORM	0.983(0.038)	0.985(0.008)	0.000(0.000)
	SCAD	0.676(0.129)	0.921(0.030)	0.342(0.319)
	Lasso	0.466(0.133)	0.987(0.013)	0.170(0.159)
	adaLasso	0.571(0.128)	0.991(0.011)	0.052(0.093)
	pCollRe	0.677(0.104)	1.000(0.000)	0.014(0.051)
	mCCA	0.603(0.243)	0.991(0.029)	0.048(0.090)
Scenario 2	SIFORM	0.989(0.031)	0.987(0.008)	0.000(0.000)
	SCAD	0.266(0.178)	0.964(0.036)	0.350(0.320)
	Lasso	0.323(0.094)	0.958(0.024)	0.436(0.178)
	adaLasso	0.111(0.057)	0.995(0.001)	0.192(0.039)
	pCollRe	0.540(0.091)	1.000(0.000)	0.068(0.099)
	mCCA	0.688(0.277)	0.984(0.037)	0.022(0.142)
Scenario 3	SIFORM	0.959(0.015)	0.994(0.002)	0.000(0.000)
	SCAD	0.195(0.036)	0.988(0.005)	0.441(0.222)
	Lasso	0.093(0.020)	0.998(0.002)	0.136(0.121)
	adaLasso	0.075(0.012)	0.999(0.001)	0.050(0.085)
	pCollRe	0.542(0.031)	0.996(0.002)	0.001(0.010)
	mCCA	0.068(0.014)	1.000(0.000)	0.000(0.000)
Scenario 4	SIFORM	0.950(0.090)	1.000(0.000)	0.000(0.000)
	SCAD	0.364(0.102)	0.989(0.160)	0.335(0.299)
	Lasso	0.287(0.112)	0.995(0.242)	0.212(0.144)
	adaLasso	0.390(0.096)	1.000(0.147)	0.069(0.099)
	pCollRe	0.430(0.072)	1.000(0.000)	0.086(0.115)

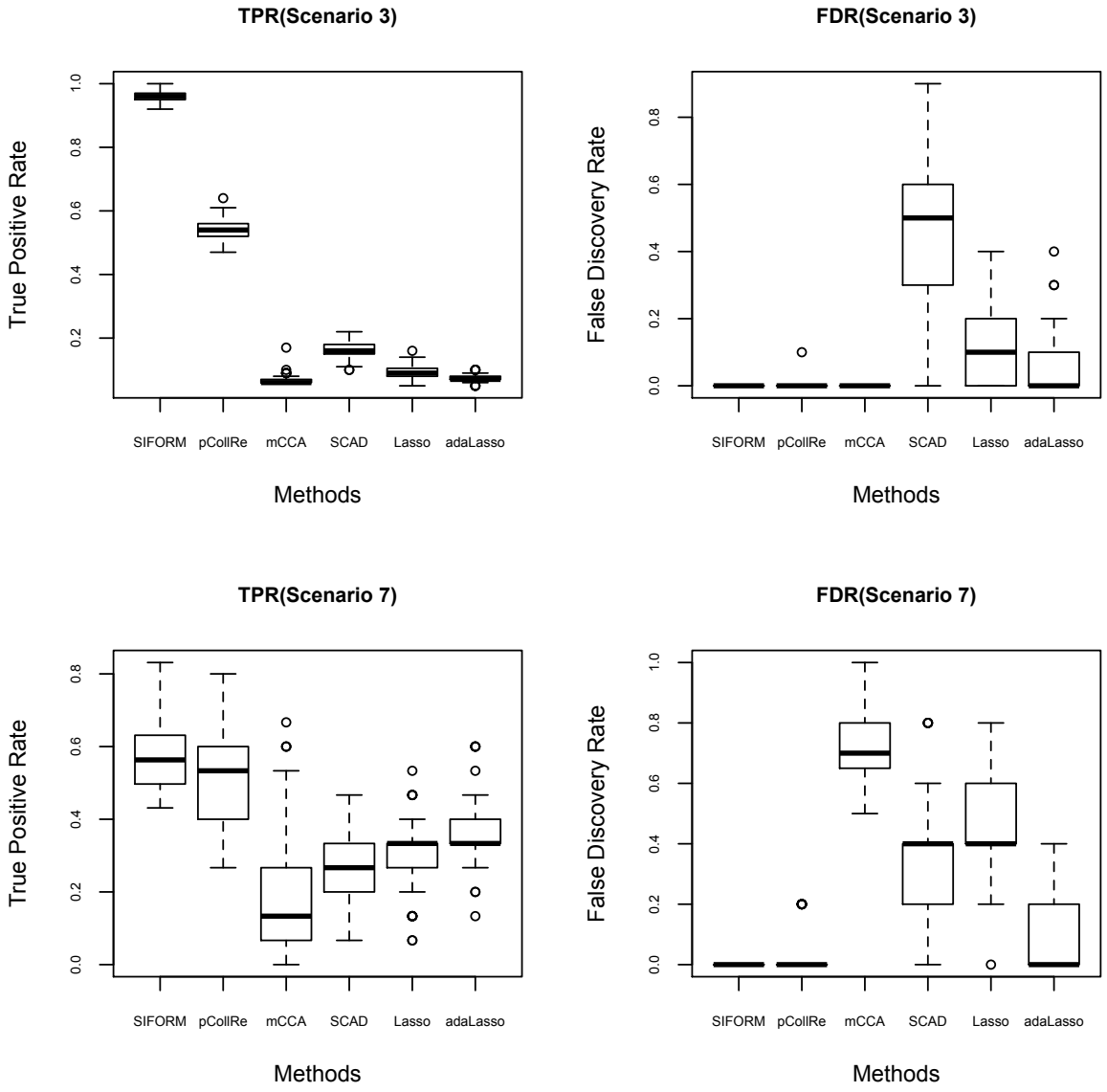
	mCCA	0.316(0.161)	0.999(0.005)	0.529(0.069)
Scenario 5	SIFORM	0.791(0.123)	0.987(0.007)	0.012(0.048)
	SCAD	0.754(0.105)	0.925(0.024)	0.334(0.203)
	Lasso	0.642(0.136)	0.903(0.028)	0.170(0.140)
	adaLasso	0.675(0.130)	0.984(0.020)	0.016(0.055)
	pCollRe	0.627(0.112)	0.994(0.000)	0.024(0.071)
	mCCA	0.398(0.286)	0.970(0.045)	0.376(0.216)
Scenario 6	SIFORM	0.392(0.067)	0.999(0.005)	0.000(0.000)
	SCAD	0.246(0.062)	0.988(0.004)	0.598(0.208)
	Lasso	0.132(0.044)	0.990(0.003)	0.390(0.174)
	adaLasso	0.106(0.033)	0.999(0.001)	0.110(0.140)
	pCollRe	0.498(0.066)	0.996(0.001)	0.014(0.051)
	mCCA	0.206(0.220)	0.973(0.044)	0.740(0.280)
Scenario 7	SIFORM	0.622(0.094)	1.000(0.000)	0.000(0.000)
	SCAD	0.262(0.094)	0.967(0.014)	0.346(0.170)
	Lasso	0.315(0.101)	0.929(0.019)	0.458(0.174)
	adaLasso	0.413(0.090)	0.999(0.007)	0.056(0.095)
	pCollRe	0.520(0.109)	1.000(0.000)	0.018(0.058)
	mCCA	0.270(0.200)	0.966(0.043)	0.726(0.119)

I also consider generating data from logistic regression models under scenarios 5, 6, and 7, which represent different dimensionalities and numbers of platforms of genetic data. While the differences between *SIFORM* and the other three methods are not as large as in the first three scenarios, the advantage of using *SIFORM* in terms of selecting true biomarkers is still clear. In most scenarios, *SIFORM* detects the most true biomarkers with the smallest FDR. In contrast, *SCAD*, *Lasso*, and *mCCA* yield high FDR, implying that large numbers of trivial genetic variables are selected

into the models. The results are still solid when genetic variables come from three different platforms. *pCollRe* discovers more true positives than *SIFORM* in scenario 6. However, *pCollRe* may not be well-suited for prediction based on the evaluation of the predictive performance of *SIFORM* and *pCollRe* by cross-validation [39]. For simplicity, for each of the 100 data sets in scenario 6, I focus on n samples in the two most extreme categories: y=1 and y=4. Specifically, leave-one-out cross-validation is used. Among the n samples, $n-1$ are used as the training set to build a classifier, which is then used to predict the phenotype of the remaining sample. This procedure is repeated n times to obtain the overall misclassification rate for each data set. To make a fair comparison, I focus on the predictive performance of the top 5 biomarkers with the largest absolute estimated coefficient values as identified by each method. The means (and standard deviations) of the misclassification rates of *SIFORM* and *pCollRe* are respectively 0.283(0.085) and 0.409(0.090). These results indicate that *SIFORM* significantly outperforms *pCollRe* in terms of prediction.

Figure 2.1 depicts the boxplots of the number of TPRs and FDRs obtained by all six methods over 100 simulations under scenarios 3 and 7.

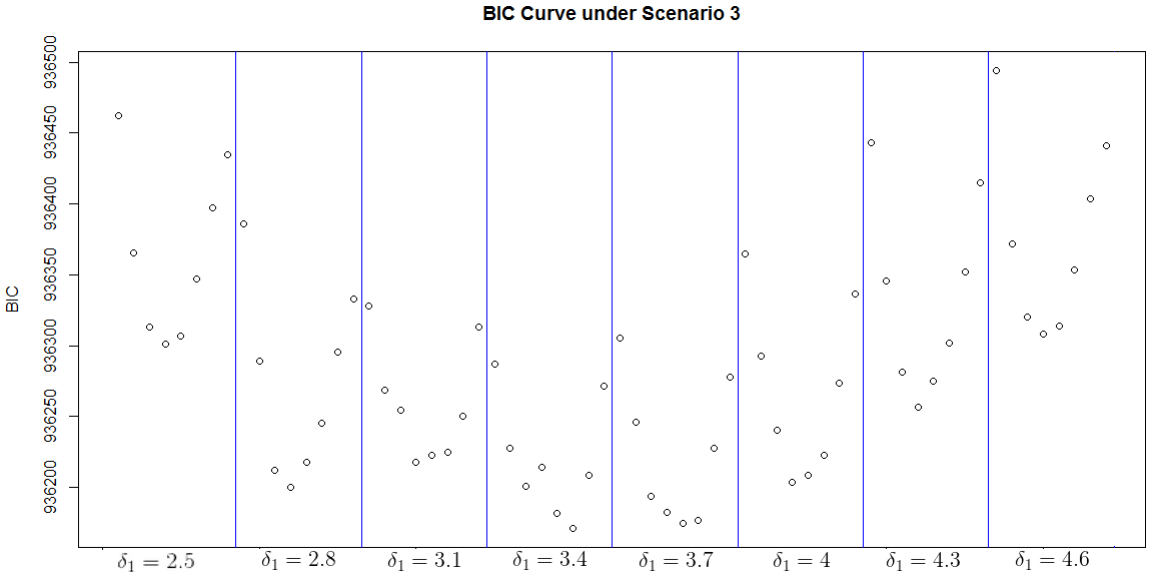
Figure 2.1.: Comparison of number of TPRs and FDRs across 6 methods under scenarios 3 and 7 (*SIFORM*)



I also assess the performance of the BIC in the proposed framework. Using scenario 3 as an example, Figure 2.2 shows the BIC values along the tuning parameters δ_1 and δ_2 in a simulated dataset. Reading from left to right in Figure 2.2, the subpanels correspond to δ_1 values increasing from 2.5 to 4.6. In each subpanel, the BIC values

are plotted against the values of δ_2 in increments of 0.3. The good behavior of the BIC method is indicated by clear convex curves, with $(\delta_1 = 3.4, \delta_2 = 4)$ chosen as the optimal tuning parameter values for this dataset. In a common scenario, it can be told that the results are very similar across different simulated datasets, and thus, I use the tuning parameter values obtained from a single dataset for computational consideration.

Figure 2.2.: BIC curve in two-dimensional grid search under scenario 3



2.4 Lung adenocarcinoma applications

Lung cancer is the leading cause of cancer deaths in the United States [55]. Among the major histological types of lung cancer, adenocarcinoma is the most common form in non-smokers [56]. I investigate the applicability of *SIFORM* to the TCGA lung adenocarcinoma data set available through the data portal hosted by the National Cancer Institute (http://cancergenome.nih.gov/). In particular, two disease phenotypes, smoking status and survival, are studied to demonstrate the application of

SIFORM and explore new molecular biomarkers that may facilitate the comprehensive understanding of lung cancer.

2.4.1 The association of biomarkers with smoking status

Cigarette smoking is the most important risk factor for lung cancer; however, approximately 25% of lung cancer cases are not attributable to tobacco use. Striking differences in the epidemiological, clinical and molecular characteristics of lung cancer have been demonstrated when the cancer arises in a non-smoker versus a smoker [56, 57, 58, 59]. This suggests that lung cancers are likely caused by separate biological mechanisms in smokers versus non-smokers.

Substantial studies have identified the genes and pathways that contribute to lung tumorigenesis in smokers (e.g., *EGFR*, *KRAS* and *TP53*) [56, 60], and these discoveries have led to the development of targeted therapies (e.g., *EGFR* tyrosine kinase inhibitor erlotinib; anti-VEGF antibody bevacizumab) [61]. However, most of the studies have analyzed genetic profiles in a single assay, and the etiology of lung tumors that arise in non-smokers remains unclear [56]. Therefore, it is worthwhile to integrate multi-platform bioinformatic data to discover predictive biomarkers that are associated with the smoking status of patients diagnosed with lung adenocarcinoma.

Data description

The TCGA lung adenocarcinoma data set includes samples from 225 patients, for whom the expression intensities of 20,531 genes and 160 proteins were measured using the respective platforms of Illumina RNA sequencing and reverse-phase protein array (RPPA) technology. The variable of smoking status has four categories. Among the 225 patients represented in the samples, there were 49 with the status of "current

smoker", 86 with "current reformed smoker for <or=15 years", 58 with "current reformed smoker for >15 years", and 32 with "lifelong non-smoker".

The gene expression data were normalized using the RNA-seq by expectation-maximization approach [62] and logarithm-transformed prior to downstream analysis. The protein concentration data were also normalized by subtracting the median, both column-wise and row-wise [63].

A subsample study

The six methods that I investigated aim to simultaneously analyze genetic variables to detect predictive biomarkers for a given phenotype. I am further interested in investigating the capabilities of the methods to detect biomarkers that are likely marginally associated with the phenotype, using individual marker tests as the reference.

I subsampled 300 genes and 100 proteins from all the genetic variables. Because the truth is unknown in a real data study, I identified 30 genes via gene shaving [64] or with the smallest p-values obtained from a univariate F-test; these 30 genes were treated as the true biomarkers. The remaining 270 "null" genes are genes that have the largest p-values obtained from the univariate F-test. Among the 100 proteins, 10 were treated as truly important as determined by gene shaving and univariate test results. The other 90 "null" proteins are proteins with the largest p-values based on the univariate F-test.

I report the results of all six methods in Table 2.2, which lists the TPR, TNR, and FDR values. Compared to the other methods, *SIFORM* has superior performance in terms of accurate biomarker detection. *pCollRe* has the second highest true positive rate among the remaining methods. *Lasso* performs the worst in both true and false biomarker detection, while *SCAD*, *adaLasso*, and *mCCA* have similar performances,

with TPRs and TNRs that fall in between those of the other two methods. These results are consistent with the simulation results demonstrated in the previous section.

Table 2.2: Comparison of six methods in a subsample lung study with smoking status as outcome.

Method	TPR	TNR	FDR
SIFORM	0.950	1.000	0.000
SCAD	0.350	0.994	0.200
Lasso	0.100	0.986	0.400
adaLasso	0.275	1.000	0.000
pCollRe	0.775	0.994	0.000
mCCA	0.500	0.986	0.000

Full data analysis

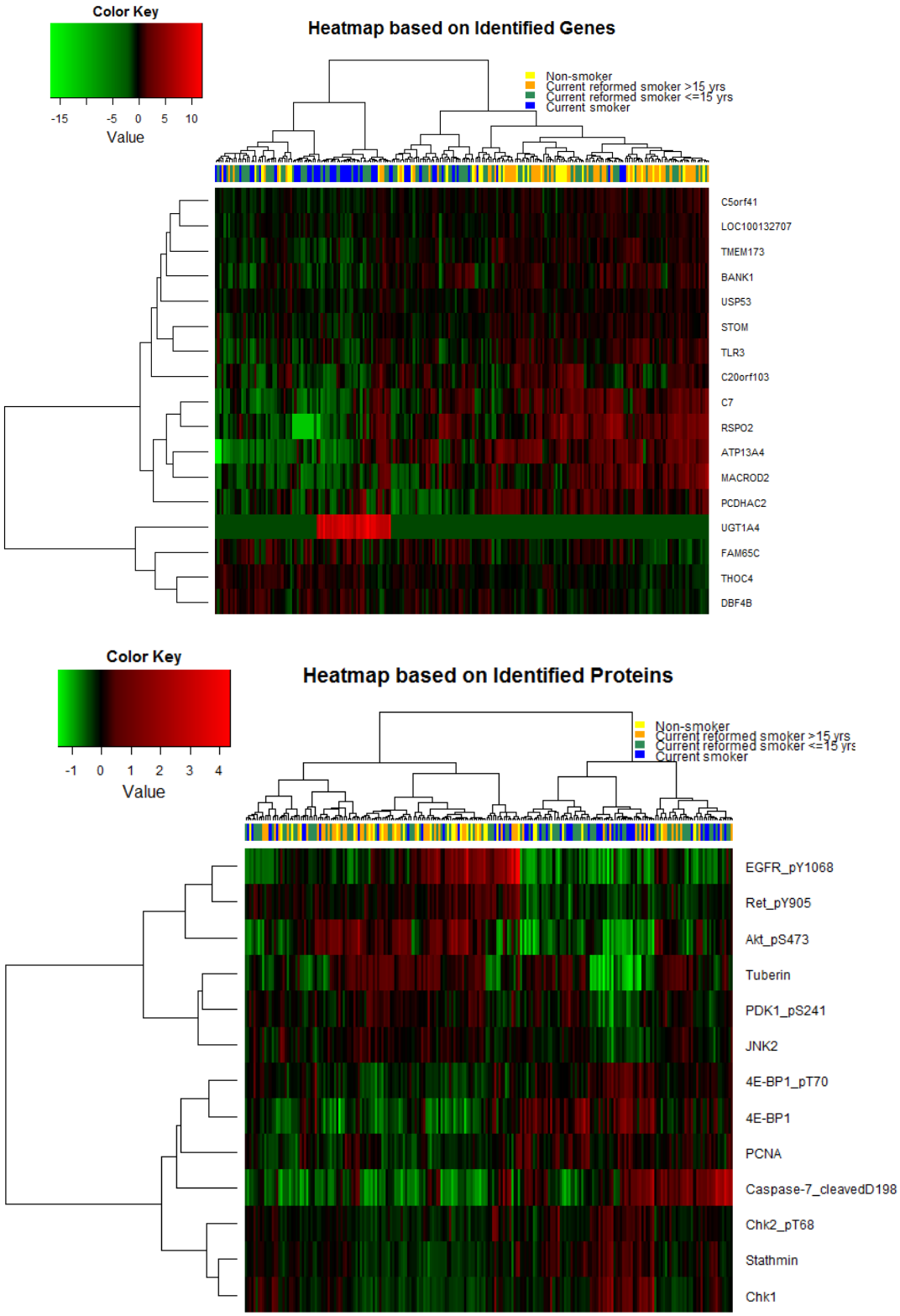
The following steps were implemented to filter out trivial genes in the RNA-seq data. First, I removed 5% of the genes that had extremely small coefficient of variation values, where the coefficient of variation is defined as the ratio of the standard deviation to the absolute value of the mean of the gene expression intensities. Second, I removed genes for which the difference between the top 90% quantile and the bottom 10% quantile was no larger than 0.8. Then, I filtered out the genes with p-value ≥ 0.03 based on a univariate analysis of the variance F-test. This procedure retained 3,707 genes for further analysis.

I implemented all six methods for the integrative analysis of the genomic and proteomic data. To determine the tuning parameter values in *SIFORM*, I performed a grid search over $[4, 14]$ for δ_1 and $[1, 10]$ for δ_2. I obtained $(\delta_1, \delta_2) = (5.2, 4)$ as the local optimal values. *SIFORM* identified 17 genes that had nonzero coefficients in

$\boldsymbol{A}$: ATP13A4, BANK1, C20orf103, C5orf41, C7, DBF4B, FAM65C, LOC100132707, MACROD2, PCDHAC2, RSPO2, STOM, THOC4, TLR3, TMEM173, UGT1A4, and USP53. *SIFORM* also detected 13 proteins based on $\boldsymbol{B}$: 4E-BP1, 4E-BP1_pT70, Akt_pS473, caspase-7_cleavedD198, Chk1, Chk2_pT68, EGFR_pY1068, JNK2, PCNA, PDK1_pS241, Ret_pY905, stathmin, and tuberin.

I performed hierarchical clustering of the samples based on the Pearson correlation distance and applied Ward's linkage method to the selected genes and proteins. The clustering heatmaps using the 17 genes and 13 proteins, respectively, are displayed in Figure 2.3. In both panels, it can be observed that most "non-smokers" (yellow) and "current reformed smokers for >15 years" (orange) are clustered together. In addition, most of the 49 "current smokers" (blue) are mixed with some "current reformed smokers for <=15 years" (green). However, the data for some of the 86 recently reformed smokers (green) behave differently from the rest of the group; these samples enlarge the overall difference between the green and blue groups. The observations can be somewhat reflected from the estimate of $\boldsymbol{C} = (-0.064, -0.074, 0.015, 0.104)$. The close values (c_1, c_2) between "non-smokers" and "current reformed smokers for >15 years" indicate similarity between the major genetic profiling characteristics of these two categories of smoking status. The relatively large difference between (c_1, c_2), c_3, and c_4 manifests possible genetic profiling variations between (roughly) non-smokers, recent smokers, and current smokers. In particular, the most different genetic profiles are those of non-smokers compared to current smokers, which agrees with the intuitive expectation.

Figure 2.3.: Sample clustering based on genes and proteins selected by *SIFORM*



It can also be seen from Figure 2.3 that the most identified biomarkers show clear differential expression between light smokers and heavy smokers. For instance, the genes contained in the first 13 rows of the heatmap in the upper panel have lower expression levels in heavy smokers but higher expression levels in light smokers, and the bottom four genes show the opposite expression pattern. As an example, gene *UGT1A4* (the 14^{th} gene in the upper panel of Figure 2.3) appears to be up-regulated in the heavy smokers but down-regulated in the light smokers. Extensive evidence shows that *UGT1A4* can be induced by cigarette constituents [65, 66].

The comparisons among the six methods in terms of variable selection are summarized as follows.

- *SCAD*, *Lasso*, *adaLasso*, *pCollRe*, and *mCCA* identified 60, 22, 33, 27, and 12 genes, respectively. *SCAD* detected more genes than the other methods; however, it likely yielded the largest number of false positives, as indicated by the simulation studies. No common gene was identified across the six methods. This implies an ongoing challenge for genetic studies of the association between smoking and lung cancer, which is also reflected in the very limited medical literature in this area of research. *SCAD*, *Lasso*, and *mCCA*, all of which are known to produce high false positives, identified 5 genes in common. *AdaLasso* and *pCollRe*, which are known to have appealing empirical and theoretical properties [28, 39], identified very few genes that overlap with those identified by the other comparative methods but substantially overlap with those identified by *SIFORM*. (There are 13 genes identified by both *SIFORM* and *adaLasso*; 10 genes identified by both *SIFORM* and *pCollRe*.)

- In terms of proteomic profiling, *SCAD*, *Lasso*, *adaLasso*, and *mCCA* identified only 3, 0, 1, and 1, proteins, respectively. In contrast, the proposed method, *SIFORM*, detected 13 proteins and *pCollRe* detected 10 proteins. *SIFORM* and

pCollRe identified 9 common proteins, which makes the results of *SIFORM* more convincing due to the good empirical results of *pCollRe* in simulation studies.

Next, the biological relevance of the genes and proteins detected by *SIFORM* is studied. Conducting pathway analysis with a web-based tool, Pathway Commons (http://www.pathwaycommons.org/pcviz/), I discovered that 13 genes are likely to interact with each other through functional pathways and networks. This is displayed in Figures 2.4, where these genes are highlighted in green. The important miRNAs associated with a common pathway are shown in red. The 8 detected proteins appear to be in several common pathways, as highlighted in green in Figure 2.5. Several of them are known to be associated with lung cancer.

EGFR, which was detected in the study, has been found to play a key role in lung tumorigenesis. Approximately 10% of non-small-cell lung carcinoma (NSCLC) cases in the U.S. and 35% in East Asia have tumor-associated *EGFR* mutations. *EGFR* mutations are more often found in tumors from non-smokers with adenocarcinoma histology (http://www.mycancergenome.org/content/disease/lung-cancer/egfr/)[68], which supports the biological relevance of our finding.

Activated *EGFR* can regulate the activity of another detected gene, *PCNA*, and trigger its important downstream signaling pathway AKT/PI3K [69]. *PCNA*, an index of tumor cell proliferation, has high expression in poorly differentiated lung adenocarcinomas [70, 71]. Another detected biomarker, *AKT1*, is an isoform of *AKT* [72]. The signal transduction pathway AKT/PI3K is involved in the regulation of cell proliferation, survival, differentiation, adhesion, motility and invasion. Aberrations of this pathway have been implicated in lung cancer development and progression [72].

Figure 2.4.: The network structure of genes identified by *SIFORM*

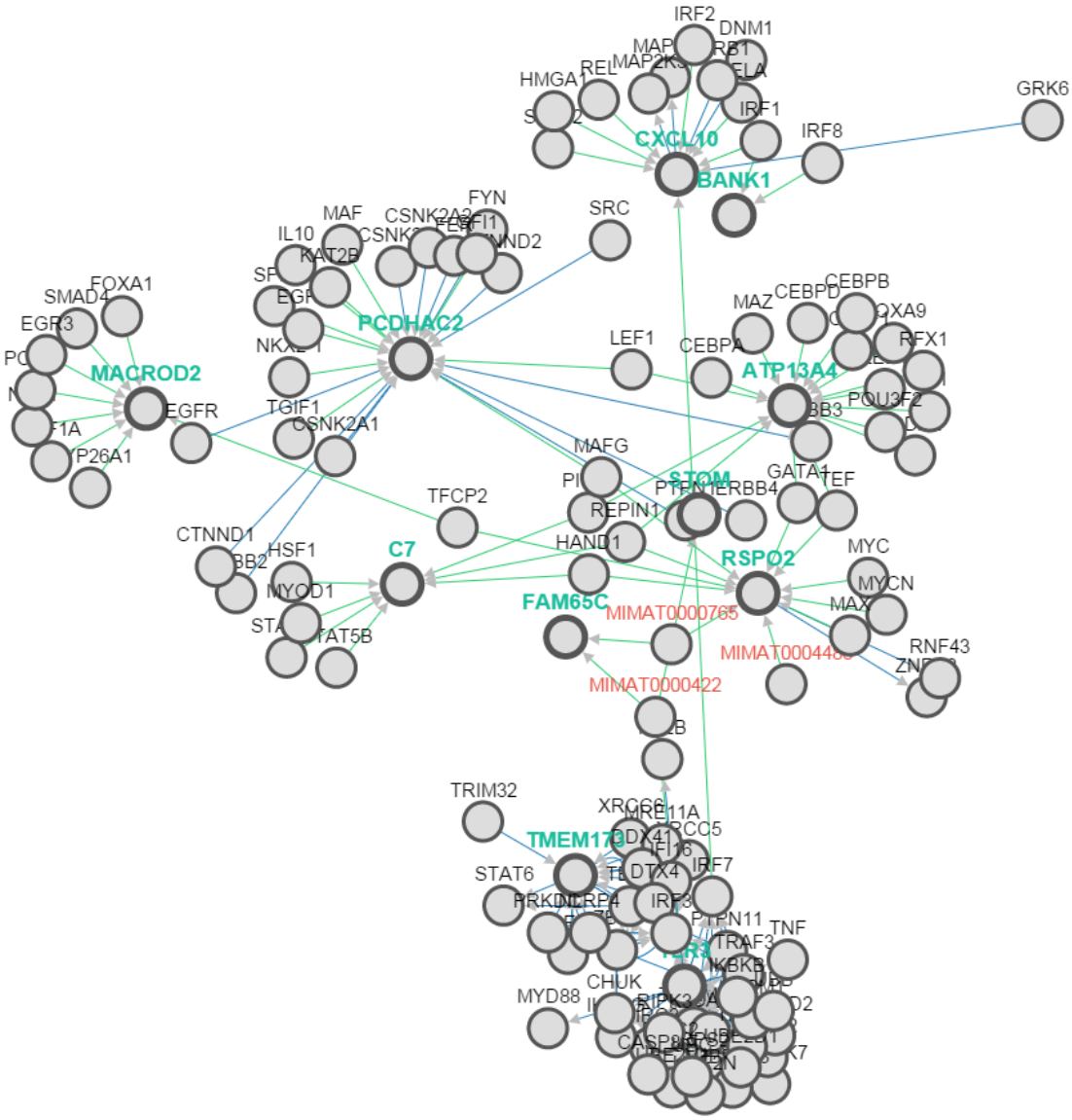
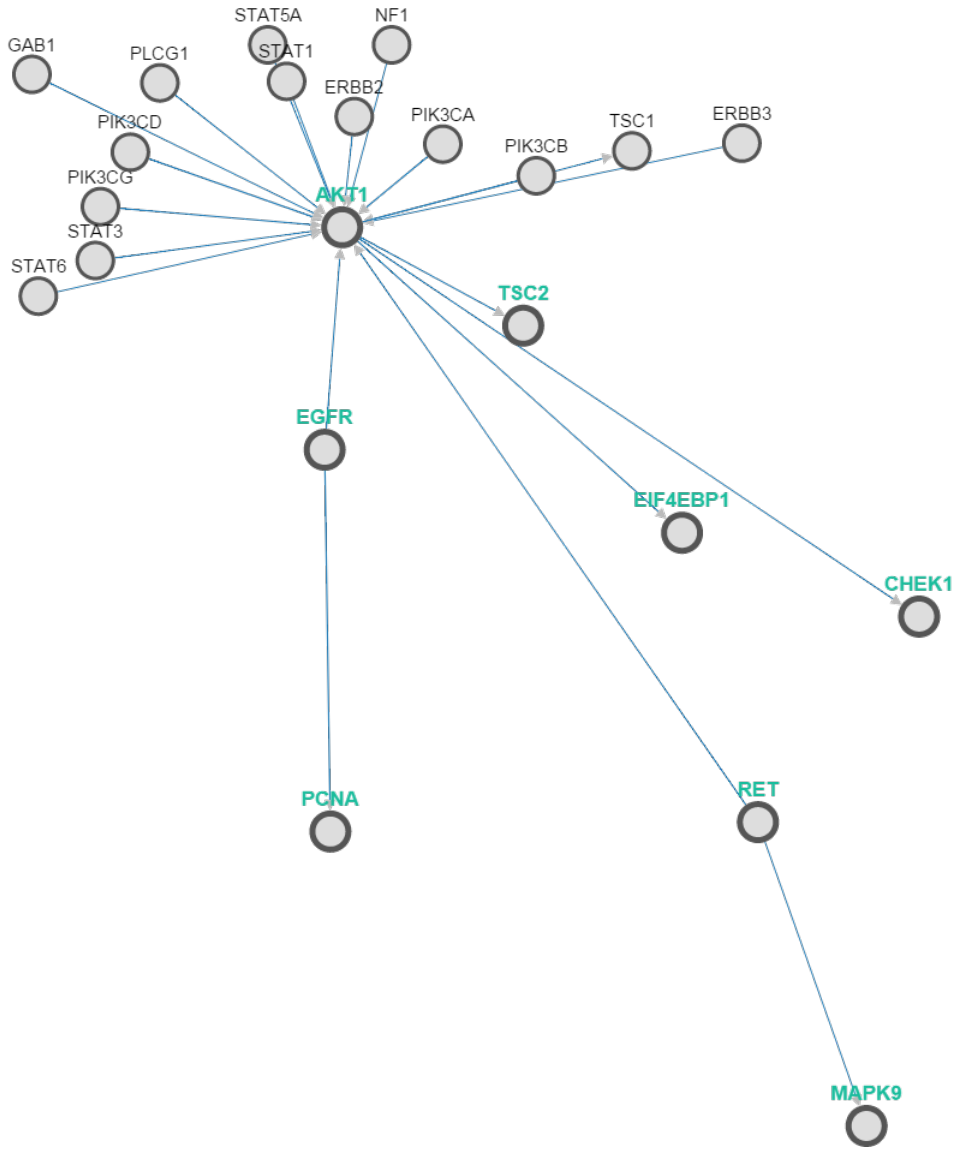


Figure 2.5.: The network structure of proteins identified by *SIFORM*



Most of the proteins selected by *SIFORM* are associated with the tumorigenesis of NSCLC through the AKT pathway, as shown by Figure 2.5. First, the activated AKT pathway regulates tuberin (*TSC2*), which inhibits the *mTOR* nutrient signaling input through the tuberous sclerosis complex. *mTOR* has been correlated with NSCLC tumor progression [69, 72]. Second, the activated *AKT* phosphorylates *CHEK1*, an integral component of the DNA damage response. The overexpression of *CHEK1* is

associated with poor tumor differentiation and significantly worse patient survival in NSCLC [73]. In addition, activated *AKT* induces the phosphorylation and inactivation of *EIF4EBP1*, and the increased phosphorylation of *EIF4EBP1* is found to be associated with the progression of several types of cancer, including lung adenocarcinoma [74, 75, 76]. In addition to *EGFR*, we identified another activator of the AKT/PI3K pathways—*RET*. The alteration of *RET* has key roles in cell growth, differentiation, and survival, and has been associated with NSCLC [72]. In addition to *PI3K/AKT*, *RET* signaling activates the *MAPK* family, including *MAPK9 (JNK2)* [77]. Some studies have found that *MAPK9* is frequently activated in NSCLC [78].

In summary, *EGFR* and its downstream *PI3K/AKT* pathway are among the most important molecular therapeutic targets for NSCLC [79, 80]. Existing studies suggest differences in *EFGR* mutations between smokers and non-smokers, and thus implicate the *AKT/PI3K* pathway and its downstream targets as playing nontrivial roles that differ in patients who develop lung cancer and have a history of smoking versus those who have never smoked. In contrast, other methods have identified only one protein, collagen VI, which is an extracellular matrix protein. Although collagen VI has been correlated with tumor progression, its role in NSCLC is rarely discussed [81, 82].

Cross-validation: I evaluated the predictive performance of each method used in simulation. For simplicity, I focused on the 81 patients whose habits placed them in the two most extreme categories related to smoking: 49 who were current smokers and 32 who were lifelong non-smokers. Specifically, I used leave-one-out cross-validation to predict the smoking status of each patient. This procedure was repeated 81 times to obtain the overall misclassification rate. To make a fair comparison, I focused on the predictive performance of the top 5 biomarkers with the largest absolute coefficient estimates. The misclassification rates of *SIFORM*, *SCAD*, *Lasso*, *adaLasso*,*mCCA* and *pCollRe* are respectively 0.407, 0.469, 0.593, 0.556, 0.605 and 0.593. These results indicate that *SIFORM* has the best prediction accuracy. The overall high mis-

classification rates across the six methods might arise from the limited sample sizes and the recognized difficulty of differentiating the genetic mechanisms of lung cancer attributable to smoking versus not smoking.

2.4.2 The association of biomarkers with survival

The 5-year survival rate of advanced lung cancer is only approximately 15%, and conventional treatments (e.g., chemotherapy, radiotherapy) have limited impact on improving survival for patients with advanced NSCLC. Mutations in genetic biomarkers result in aberrant function and uninterrupted signaling, which may lead to pharmacological inhibition for conventional treatments. Therefore, the identification of these biomarkers plays a key role in developing targeted cancer therapy and forming the basis of personalized medicine [83]. The integrative analysis of multi-platform bioinformatic data is particularly useful for precise understanding of the underlying relationship between genetic mutations and survival in patients with lung adenocarcinoma.

Data description

I used the same gene expression and protein expression data from the same set of patients that are used in Section 2.4.1. In this study, survival, instead of smoking status, is used as the outcome. Two hundred nineteen patients with valid survival information are divided into two groups (long-term survival (LTS) and short-term survival (STS)) based on their survival duration. By using an extreme discordant phenotype design [84], patients whose survivals rank in the top 25% (55 patients, surviving >896 days) are defined as LTS, and those in the bottom 65% (29 patients, surviving <895 days) are defined as STS. Finally, 84 patients remain for further analysis.

A subsample study

As in Section 2.4.1, I subsampled 300 genes and 100 proteins from all the genetic variables. Among them, 30 genes and 10 proteins were selected via gene shaving [64] or the univariate F-test. They were treated as true biomarkers. The remaining 270 "null" genes and 90 "null" proteins were biomarkers that have the largest p-values obtained from the univariate F-test.

I report the results of the implementation of all six methods in Table 2.3. Compared to the other methods, *SIFORM* has superior performance in terms of accurate biomarker detection. Again, *pCollRe* performs the second best in both true and false biomarker detection; *Lasso* has the smallest TNR among the other methods; *SCAD*, *adaLasso*, and *mCCA* have similar and moderate performances. These results are consistent with the simulation results.

Table 2.3: Comparison of six methods in a subsample lung study with discretized survival as outcome.

Method	TPR	TNR	FDR
SIFORM	0.950	1.000	0.000
SCAD	0.325	0.992	0.000
Lasso	0.325	0.958	0.000
adaLasso	0.250	1.000	0.000
pCollRe	0.800	0.992	0.000
mCCA	0.475	0.989	0.000

Full data analysis

Using steps that are similar to those described in Section 2.4.1, I filtered out the trivial genes in the RNA-seq data. First, I removed 10% of the genes that had

extremely small coefficient of variation values, where the coefficient of variation is defined as the ratio of the standard deviation to the absolute value of the mean of the gene expression intensities. Second, I removed genes for which the difference between the top 90% quantile and the bottom 10% quantile was no larger than 0.8. Then, I filtered out the genes with p-value ≥ 0.05 based on a univariate analysis of the variance F-test. This procedure retained 2,529 genes for further analysis.

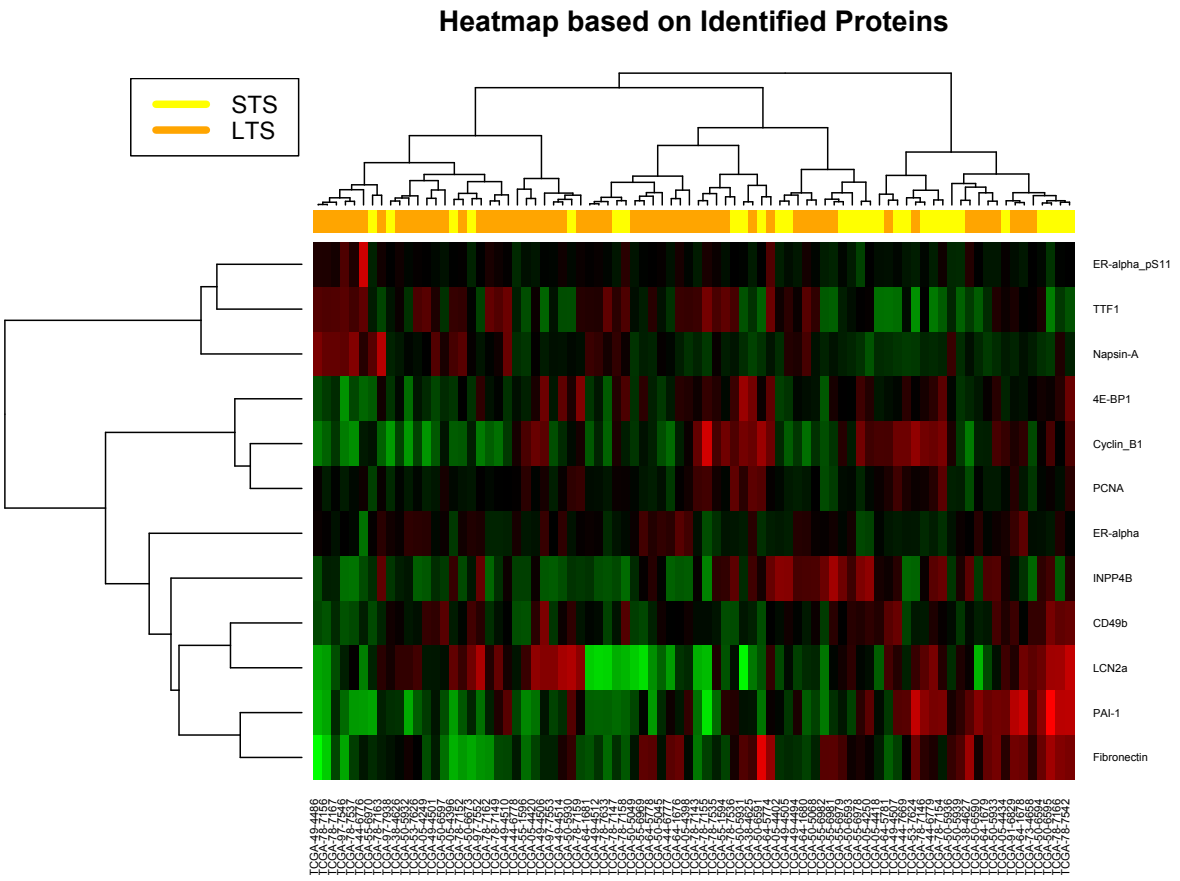
I implemented all six methods for the integrative analysis of the genomic and proteomic data. To determine the tuning parameter values in *SIFORM*, I performed a grid search over $[1, 10]$ for δ_1 and $[0, 4]$ for δ_2. The parameters $(\delta_1, \delta_2) = (4.7, 1.9)$ are obtained as the local optimal values. *SIFORM* identified 12 proteins that had nonzero coefficients in $\boldsymbol{B}$: 4E-BP1, CD49b, Cyclin_B1, ER-alpha, ER-alpha_pS118, Fibronectin, INPP4B, LCN2a, Napsin-A, PAI-1, PCNA, and TTF1. No genes were identified as prognostic biomarkers for survival outcome in this data set.

The other methods discovered more genes (24 by *SCAD*, 40 by *Lasso*, 17 by *adaLasso*, 215 by *mCCA*, and 29 by *pCollRe*) but fewer proteins (1 by *SCAD*, 1 by *Lasso*, 1 by *adaLasso*, 1 by *mCCA*, and 4 by *pCollRe*) compared to *SIFORM*. There is no common gene that was identified across all 5 comparative methods. There are 8 genes that were identified by 4 methods simultaneously: CLEC4G, EPGN, LOC392196, LOC728643, NRL, RPS6KL1, TFAP2A, and TMEM9B. However, none of them is known to be correlated with lung cancer survival times. In terms of proteomic profiling, estrogen receptor (ER)-alpha is the only protein that was discovered by 4 methods. The prognostic values of ER expressions in lung cancer have been extensively discussed. ER proteins, including ER-alpha, have been found to be associated with a poorer prognosis among NSCLC patients [85, 86]. ER-alpha was also discovered by *SIFORM*.

As in the previous section, I performed hierarchical clustering of the samples based on the Pearson correlation distance and applied Ward's linkage method to the selected

proteins. From Figure 2.6, it can be observed that most "STS" (yellow) and "LTS" (orange) are clustered together. It can also be told from this figure that most of the identified proteins show clear differential expression between "LTS" and "STS".

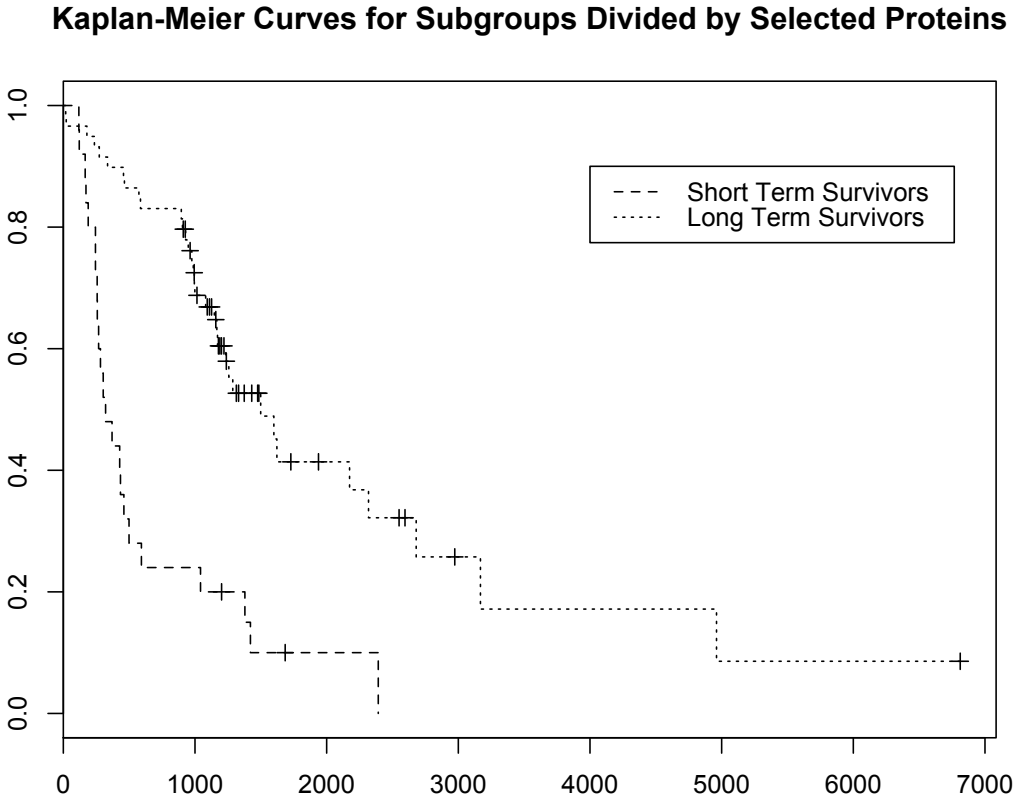
Figure 2.6.: Sample clustering based on proteins selected by *SIFORM*



I also used Kaplan-Meier curves to further investigate whether the selected proteins can differentiate patients with different survival durations. Specifically, I fit logistic regression based on discretized survival and the 12 selected proteins, and then calculated survival scores for each patient using the estimated coefficients. Each patient was then reclassified into two subgroups ("long-term survivors" vs. "short-term survivors") based on the calculated scores. Kaplan-Meier curves are plotted for the two subgroups in Figure 2.7. It is observed that the subgroups divided by selected

proteins have apparently different survival curves, which implies that the 12 proteins selected by *SIFORM* can be good prognostic biomarkers for lung adenocarcinoma patients.

Figure 2.7.: Kaplan-Meier curves for subgroups divided by *SIFORM*-selected proteins



Most of the proteins identified by *SIFORM* are found to be biologically relevant to lung cancer prognosis. 4E-BP1 is an eIF4E binding protein, and eIF4E is known to be related to reduced survival in a variety of cancers, including lung adenocarcinoma. High 4E-BP1 expression is thus found to be correlated with worse overall survival of lung cancer patients [76, 87, 88]. Cyclin B1 plays a key role in the G2-M phase transition of the cell cycle, and elevated cyclin B1 expression levels have been found to

indicate a poor prognosis in NSCLC [89, 90, 91]. Fibronectin, which plays an important role in cell adhesion, migration, growth and differentiation by mediating cellular interactions with the extracellular matrix, stimulates NSCLC cell growth and survival through activation of Akt/mTOR/p70S6K and inactivation of LKB1/AMPK signal pathways [93, 94]. Napsin-A and TTF1 have been extensively studied as correlated prognostic factors for survival among lung cancer patients. Patients with high expression levels of TTF-1 and Napsin-A have better survivals than those with low levels of expression of these factors [92]. Finally, high expression levels of PAI-1 and PCNA proteins indicate a shorter survival for patients diagnosed with lung adenocarcinoma [95, 96].

Again, the predictive performance of each method was evaluated in this study by leave-one-out cross-validation. The misclassification rates of *SIFORM*, *SCAD*, *Lasso*, *adaLasso*, *mCCA* and *pCollRe* are respectively 0.322, 0.369, 0.631, 0.560, 0.417 and 0.607. *SIFORM* still outperformed all the other methods in terms of prediction accuracy.

2.5 Inference on Regularized Regression Estimates

In Sections 2.2-2.4, I proposed a statistical framework, *SIFORM*, to integrate multi-platform data, and developed an iterative procedure for regularized parameter estimation. The superior performance of *SIFORM* in terms of biomarker detection has been demonstrated by extensive simulations and lung cancer applications. In this section, the assessment of statistical inference for the regularized estimates is discussed.

It remains difficult to develop well-performing confidence intervals (CIs) and hypothesis testing procedures for regularized estimates in a high-dimensional setting. To date, only limited work in the mainstream statistical literature has addressed this problem. Potscher et al. proposed a coverage probability theory of the confidence

intervals for *adaLasso*-type estimators [97, 98], which was shown to be infeasible when the true parameter is of similar magnitude to $n^{-\frac{1}{2}}$. Lockhart and Tibshirani introduced a statistic called the *covariance test statistic*, which is only applicable to test the significance of the selected variables in the *Lasso* model [99]. Instead, Minnier et al. proposed a perturbation method that can approximate the distribution of regularized estimates for a general class of models. This perturbation method is based on a resampling procedure [100].

I adopted this resampling-based perturbation procedure to derive CIs for biomarker effects under our framework because of its generality to different classes of penalty functions and its robustness to a misspecified working model [100].

2.5.1 A perturbation method for inference

Recall the negative likelihood function (2.2) given in Section 2.2.2, which can be re-expressed as

$$\begin{aligned} l \quad &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{p_1} [-\frac{(x_{1ij} - \alpha_{1i} - c_l \cdot a_j)^2}{2\sigma_{1j}^2} - \frac{1}{2} log(2\pi\sigma_{1j}^2)] \right. \\ &\quad \left. + \sum_{j=1}^{p_2} [-\frac{(x_{2ij} - \alpha_{2i} - c_l \cdot b_j)^2}{2\sigma_{2j}^2} - \frac{1}{2} log(2\pi\sigma_{2j}^2)] \right\} \\ &= \sum_{i=1}^{n} L(\theta; D_i) = n\widetilde{L}(\boldsymbol{\theta}), \end{aligned} \tag{2.13}$$

where $\boldsymbol{D} = (\mathbf{y}, \mathbf{X_1}, \mathbf{X_1})^{\mathbf{T}}$ and $\boldsymbol{\theta} = (\boldsymbol{C}, \boldsymbol{A}^T, \boldsymbol{B}^T, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\sigma}_1^{2T}, \boldsymbol{\sigma}_2^{2T})^T$. By using a one-step estimator with the local linear approximation [45], minimizing the negative penalized log-likelihood function (2.3) is equivalent to minimizing

$$\widehat{L}(\theta) = \widetilde{L}(\theta) + \sum_{j=1}^{p_1} p'_{\delta_{1n}j}(|\widetilde{a}_j|)|a_j| + \sum_{j=1}^{p_2} p'_{\delta_{2n}j}(|\widetilde{b}_j|)|b_j|. \tag{2.14}$$

The regularized estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ is a minimizer of the regularized objective function (2.14).

The resampling-based perturbation method proposed by Minnier et al. uses the distribution of perturbed estimators $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}_0)$ to approximate the distribution of penalized estimators $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, which lays a foundation for assessing statistical inference for the regularized estimates [100]. The first step of this method is to simulate a set of independent and identically distributed (i.i.d.) positive random variables $\mathbf{S} = \{G_i; i = 1, \cdots, n\}$, where G_i has mean and variance equal to 1. Then, by plugging G_i into the initial objective function (2.13), we obtain the perturbed objective function

$$\widetilde{L}^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{L}(\theta; D_i)G_i.$$

Similarly, the perturbed version of the regularized objective function is defined by

$$\hat{L}^*(\boldsymbol{\theta}) = \widetilde{L}^*(\boldsymbol{\theta}) + \sum_{j=1}^{p_1} p'_{\delta_{1nj}}(|\widetilde{a}_j|)|a_j| + \sum_{j=1}^{p_2} p'_{\delta_{2nj}}(|\widetilde{b}_j|)|b_j|, \tag{2.15}$$

and the perturbed regularized estimator $\hat{\boldsymbol{\theta}}^*$ can be obtained in the following way:

$$\hat{\boldsymbol{\theta}}^* = \underset{\theta}{\operatorname{argmin}}\ \hat{L}^*(\boldsymbol{\theta}). \tag{2.16}$$

Let $\Delta = \{j : \theta_{0j} \neq 0\}$ and $\Delta^C = \{j : \theta_{0j} = 0\}$, the resampling-based perturbation method can be justified by proving the following three statements:

(i) $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^*_\Delta - \boldsymbol{\theta}_{0\Delta})$ converges in distribution to $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), N(0, \mathbb{A}_{11}^{-1}\mathbb{B}_{11}\mathbb{A}_{11}^{-1})$, where $\mathbb{A}_{11}$ and $\mathbb{B}_{11}$ are the respective q*q submatrices corresponding to Δ.

(ii) The perturbed estimator has consistency in variable selection, that is, $P^*(\hat{\boldsymbol{\theta}}^*_{\Delta^C} = 0) \to 1$, where P^* is the probability measure generated by both $\boldsymbol{D}$ and $\mathbf{S}$.

(iii) The distribution of $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^*_\Delta - \hat{\boldsymbol{\theta}}_\Delta)|\boldsymbol{D}$ approximates the unconditional distribution of $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_\Delta - \boldsymbol{\theta}_{0\Delta})$.

Suppose a large number, say, M, of random samples S is generated, the perturbed estimator $\hat{\theta}^*_m$ can be obtained for each sample $m = 1, \cdots, M$. Then, the above asymptotic properties (i)(ii)(iii) allow us to approximate the theoretical distribution

of $\hat{\boldsymbol{\theta}}$ by empirical distribution $\left\{\hat{\theta}^*_m, m = 1, \cdots, M\right\}$. Statistical inference and the significance test for $\hat{\boldsymbol{\theta}}$ can then be carried out on the basis of $\hat{\theta}^*_m$.

The confidence intervals of $\hat{\boldsymbol{\theta}}$ can be constructed from $\hat{\boldsymbol{\theta}}^*$ in both parametric and non-parametric ways. The parametric confidence interval, normal confidence interval (CI^N), is constructed on the assumption that $\hat{\boldsymbol{\theta}}$ is normally distributed. Then, the (1-α)100% confidence interval CI^N_j for θ_{0j} can be approximated by $(\hat{\theta}_j - Z_\alpha \cdot \hat{\sigma}^*_j, \hat{\theta}_j + Z_\alpha \cdot \hat{\sigma}^*_j)$, where the standard deviation $\hat{\sigma}^*_j = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}^*_{mj} - \hat{\theta}_j)^2}$. To construct the non-parametric confidence interval, quantile confidence interval (CI^Q), I simply take the 2.5^{th} and 97.5^{th} quantiles of $\hat{\theta}^*_{mj}$ as the upper and lower bounds of CI^Q_j.

I demonstrate the resampling-based perturbation method for statistical inference through a simulation study and a lung cancer application.

Simulation study

To validate the perturbation method in finite samples, I compare the perturbed CI^Ns and CI^Qs with empirical CIs using simulated data sets in simulation scenario 1. Recall that we have 120 samples in scenario 1, each of which has 100 genetic variables from platform 1($\mathbf{X}_1$) and 100 genetic variables from platform 2($\mathbf{X}_2$). The matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are generated by

$$\mathbf{X_1} = \boldsymbol{\alpha_1} + \boldsymbol{CA} + \boldsymbol{\varepsilon_1},$$

$$\mathbf{X_2} = \boldsymbol{\alpha_2} + \boldsymbol{CB} + \boldsymbol{\varepsilon_2},$$

where the true coefficient vectors $\boldsymbol{A}_{1\times 100} = (5, 7, 11, 15, 18, 0, \cdots, 0)$ and $\boldsymbol{B}_{1\times 100} = (4, 8, 10, 14, 20, 0, \cdots, 0)$; the four categories of the phenotype vector $\mathbf{y}$ is delivered by $\boldsymbol{C} = (0.033, \cdots, 0.033, 0.067, \cdots, 0.067, 0.100, \cdots, 0.100, 0.133, \cdots, 0.133)$; the baselines $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ are set as 0; all $\epsilon_{\cdot j}$'s are simulated from distribution N(0,1).

To construct 95% empirical CIs, T (T=500) data sets are generated from the simulation scenario 1 settings as described above, where $\epsilon_{1(\cdot j)}$ and $\epsilon_{2(\cdot j)}$ are simulated

from N(0,1) with different seeds for each data set. Then, the 95% empirical CIs for $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ are calculated from the 2.5^{th} and 97.5^{th} percentiles of the distribution $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$, $t = 1, \cdots, T$.

To construct perturbed CI^Ns and CI^Qs, I generate M (M=500) perturbed samples for each data set t among the T data sets. Each of the perturbed samples has a random variable set $\mathbf{S}_m = (G_1, \cdots, G_{120})$ generated from an exponential distribution with rate=1. For m=1,...,M, I plug in $\mathbf{S}_m$ and obtain the following closed-form solutions for perturbed estimators $\hat{\boldsymbol{\theta}}^*_m$:

$$\hat{\sigma}^2_{m1j} = \frac{\sum_{i=1}^{n} G_i(x_{1ij} - \hat{\alpha}_{1j} - \hat{c}_i \cdot \hat{a}_j)^2}{n};$$

$$\hat{\sigma}^2_{m2j} = \frac{\sum_{i=1}^{n} G_i(x_{2ij} - \hat{\alpha}_{2j} - \hat{c}_i \cdot \hat{b}_j)^2}{n};$$

$$\hat{\alpha}^2_{m1i} = \frac{\sum_{j=1}^{p_1} x_{1ij} - \hat{c}_i \sum_{j=1}^{p_1} \hat{a}_j}{p_1};$$

$$\hat{\alpha}^2_{m2i} = \frac{\sum_{j=1}^{p_2} x_{2ij} - \hat{c}_i \sum_{j=1}^{p_2} \hat{b}_j}{p_2};$$

$$\hat{c}_{ml} = \frac{\sum_{j=1}^{p_1} \frac{\hat{a}_j}{\hat{\sigma}^2_{1j}} \sum_{i=n_{l-1}+1}^{n_l} G_i(x_{1ij} - \hat{\alpha}_{1i}) + \sum_{j=1}^{p_2} \frac{\hat{b}_j}{\hat{\sigma}^2_{2j}} \sum_{i=n_{l-1}+1}^{n_l} G_i(x_{2ij} - \hat{\alpha}_{2i})}{\sum_{i=n_{l-1}+1}^{n_l} G_i(\sum_{j=1}^{p_1} \frac{\hat{a}_j^2}{\hat{\sigma}^2_{1j}} + \sum_{j=1}^{p_2} \frac{\hat{b}_j^2}{\hat{\sigma}^2_{2j}})};$$

$$\hat{a_{mj}}^{(k+1)} = \frac{|a_j^{(k)}| \sum_{i=1}^{n} G_i c_i (x_{1ij} - \alpha_{1i})}{|a_j^{(k)}| \sum_{i=1}^{n} G_i c_i^2 + \sigma^2_{1j} P'_{\delta_1}(|a_j^{(k)}|)};$$

$$\hat{b_{mj}}^{(k+1)} = \frac{|b_j^{(k)}| \sum_{i=1}^{n} G_i c_i (x_{2ij} - \alpha_{2i})}{|b_j^{(k)}| \sum_{i=1}^{n} G_i c_i^2 + \sigma^2_{2j} P'_{\delta_2}(|b_j^{(k)}|)}.$$

The perturbed 95% CI^Ns for $\hat{\boldsymbol{A}}^*$ and $\hat{\boldsymbol{B}}^*$ are calculated from $(\hat{a}_j - 1.96 \cdot \hat{\sigma}^*_{Aj}, \hat{a}_j + 1.96 \cdot \hat{\sigma}^*_{Aj})$ and $(\hat{b}_j - 1.96 \cdot \hat{\sigma}^*_{Bj}, \hat{b}_j + 1.96 \cdot \hat{\sigma}^*_{Bj})$, where the standard errors $\hat{\sigma}^*_{Aj}$ and $\hat{\sigma}^*_{Bj}$ come from $\boldsymbol{A}_m$ and $\boldsymbol{B}_m$. The perturbed 95% CI^Qs for $\hat{\boldsymbol{A}}^*$ and $\hat{\boldsymbol{B}}^*$ are calculated from the 2.5^{th} and 97.5^{th} percentiles of the distribution $\boldsymbol{A}_m$ and $\boldsymbol{B}_m$.

I report the coverage probabilities (CPs) and the widths of the 95% confidence intervals CI^N and CI^Q for the nonzero coefficients in $\boldsymbol{A}^*$ and $\boldsymbol{B}^*$ in Table 2.4. The widths of the 95% empirical CIs for the nonzero coefficients in $\boldsymbol{A}$ and $\boldsymbol{B}$ are also listed for comparison.

Table 2.4: Coverage probabilities and widths of perturbed CIs (CI^N and CI^Q), compared with the widths of empirical CIs(CI^E) based on 500 simulation data sets

Coeff	CI^N			CI^Q			CI^E
	CP(β_0)	CP($\hat{\beta}$)	Width [$\mu(\sigma)$]	CP(β_0)	CP($\hat{\beta}$)	Width [$\mu(\sigma)$]	Width
a_1	0.350	1	4.865(1.133)	0.452	1	4.594(1.240)	4.598
a_2	0.768	1	6.628(1.121)	0.794	1	5.560(1.031)	6.277
a_3	0.962	1	4.856(1.081)	0.946	1	4.229(1.052)	4.543
a_4	0.948	1	3.891(0.344)	0.948	1	3.866(0.354)	4.054
a_5	0.950	1	3.896(0.352)	0.948	1	3.874(0.366)	3.908
b_1	0.252	1	3.766(1.210)	0.324	1	3.342(1.322)	3.373
b_2	0.852	1	6.558(1.062)	0.846	1	6.432(1.042)	6.599
b_3	0.942	1	5.401(1.230)	0.942	1	4.998(1.206)	5.339
b_4	0.948	1	3.932(0.375)	0.950	1	3.911(0.405)	4.353
b_5	0.944	1	3.869(0.351)	0.942	1	3.842(0.356)	3.695

Overall, the widths of CI^N, CI^Q, and CI^E are very comparable. The widths of CI^N are closer to those of CI^E compared to CI^Q, which tends to have a slightly narrower confidence interval. The CP($\hat{\beta}$)s of both CI^N and CI^Q equal 1 for all coefficients. Both types of confidence intervals cover $\hat{\beta}$ with 100% probability, which supports the expectation that the distribution of $\hat{\beta}^*$ is a good approximation of the distribution of $\hat{\beta}$. The CP(β_0)s of CI^N and CI^Q for each coefficient are very close. As expected, the CP(β_0)s of CI^N and CI^Q are very close to 0.95 when the true coefficients

are large enough, although they are apparently smaller than 0.95 for small coefficients due to the thresholding rule of the $SCAD$ penalty. The thresholding rule of $SCAD$ makes small coefficients be set at zero and moderate coefficients be shrunk towards zero, and keeps large coefficients as they are, as the following equation shows. [32]

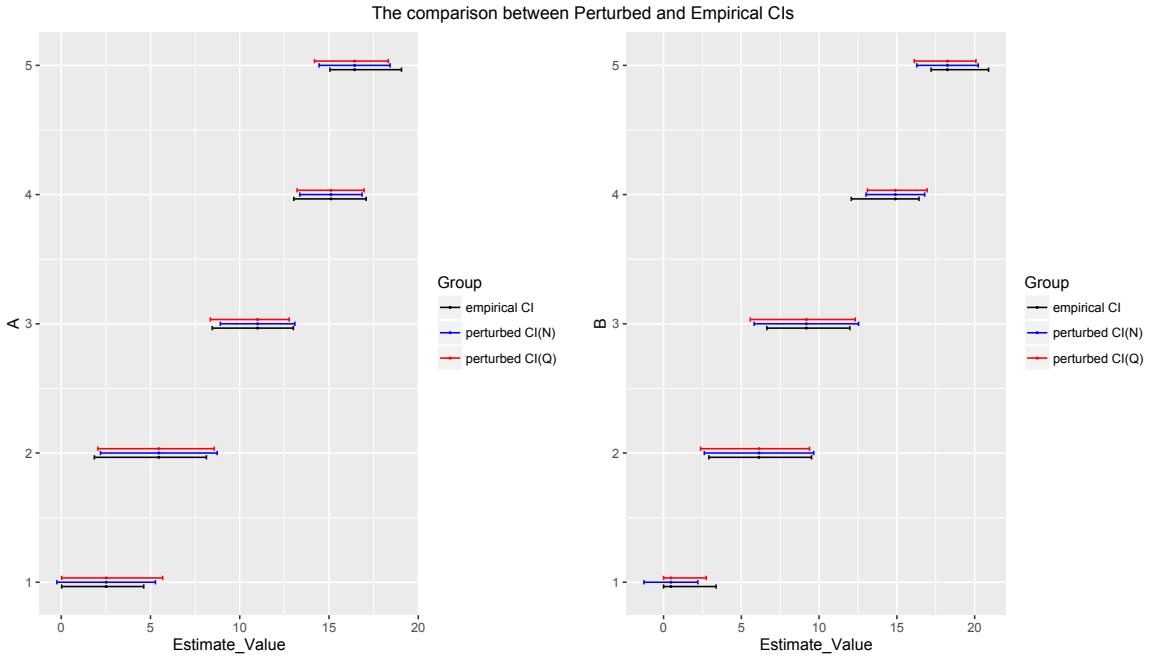
$$\hat{\beta} = \begin{cases} sgn(\beta)(|\beta| - \delta)_+ & \text{, when } |\beta| \leqslant 2\delta \\ \{(a-1)\beta - sgn(\beta)a\delta\}/(a-1) & \text{, when } 2\delta < |\beta| \leqslant a\delta \\ \beta & \text{, when } \beta > a\delta, \end{cases}$$

where a = 3.7, as suggested by Fan et al. [32]. In this simulation scenario, $\delta_1 = \delta_2 = 2.8$, and are determined by BIC. Therefore, any coefficients $|\beta| \leqslant 2.8$ are set to 0, the coefficients $2.8 < |\beta| \leqslant 5.6$ are underestimated by 2.8, the coefficients $5.6 < |\beta| \leqslant 10.36$ have some shrinkage, and the coefficients $|\beta| > 10.36$ have unbiased estimators. This explains the low CP(β_0)s of all CIs for coefficients a_1 , a_2, b_1, and b_2.

The 95% CI^N and CI^Q of all the 190 true zero coefficients in $\boldsymbol{A}$ and $\boldsymbol{B}$ have 100% coverage probabilities (both CP(β_0) and CP($\hat{\beta}$)). The simulation results show that the perturbed CIs have high probability of covering the true coefficient values and comparable widths of confidence intervals compared to the empirical CIs.

I illustrate the difference between the perturbed and empirical CIs by showing CI^N, CI^Q, and the empirical CI of the nonzero coefficients in Figure 2.8. For demonstration, I only show the CI^N and CI^Q from one (out of the 500) data sets. It can be told from the figure that both CI^N and CI^Q are very comparable to the empirical CIs. They have similar lengths and coverage regions. Slight shifts from the empirical CIs are observed for a_5, b_4, and b_5, which may be due to the difference between data set 1 and the whole 500 data sets.

Figure 2.8.: Comparison of 95% perturbed CI^Ns and CI^Qs with empirical 95% CIs for nonzero elements in A and B (simulation)

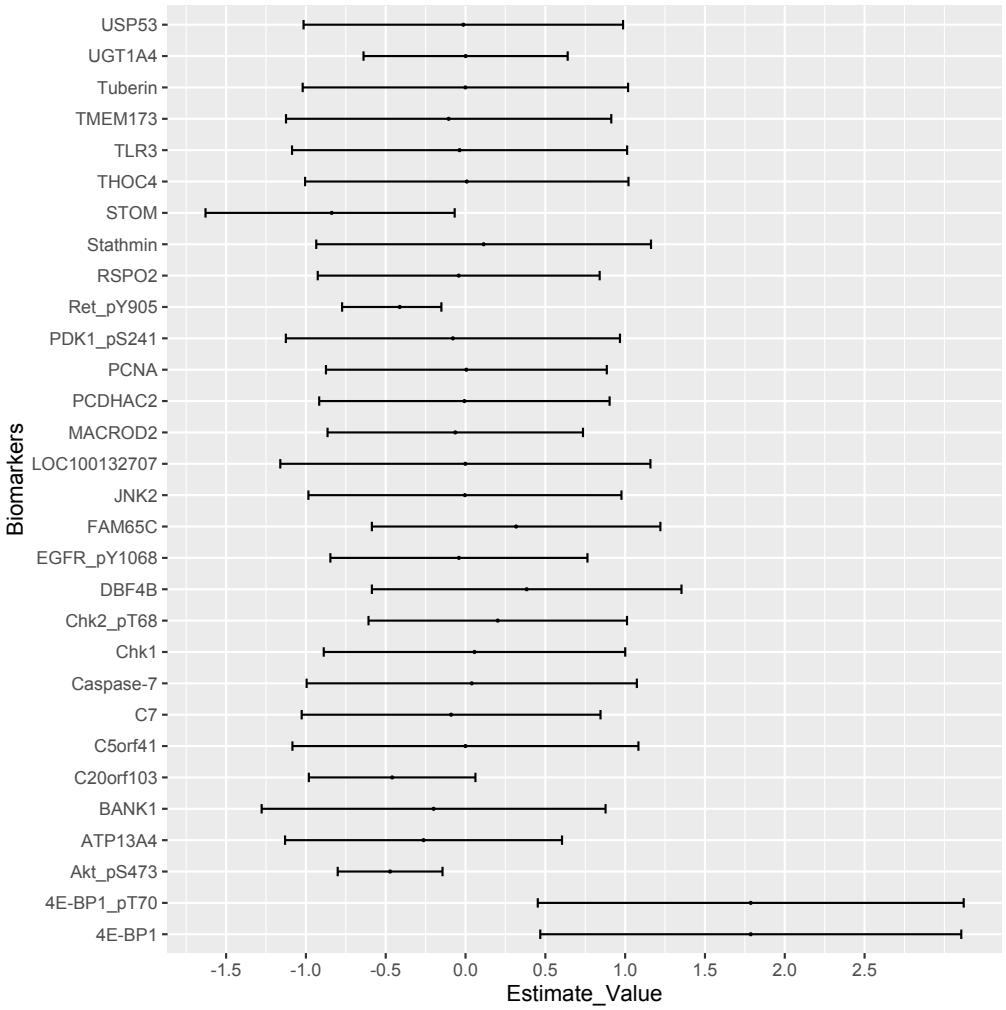


Lung adenocarcinoma application

I used the lung adenocarcinoma data described in Section 2.4.1 to further demonstrate the resampling-based perturbation method. Recall that we have 225 samples in the lung cancer study, each of which has 3,707 genes, 160 proteins, and 1 four-categorical smoking status. The association between the integrated genetic variables and smoking status was investigated. *SIFORM* identified 17 genes that may be associated with smoking status: ATP13A4, BANK1, C20orf103, C5orf41, C7, DBF4B, FAM65C, LOC100132707, MACROD2, PCDHAC2, RSPO2, STOM, THOC4, TLR3, TMEM173, UGT1A4, and USP53. *SIFORM* also detected 13 proteins: 4E-BP1, 4E-BP1_pT70, Akt_pS473, caspase-7_cleavedD198, Chk1, Chk2_pT68, EGFR_pY1068, JNK2, PCNA, PDK1_pS241, Ret_pY905, stathmin, and tuberin.

I generated 500 random variable sets $S_m = (G_1, \cdots, G_{225})$, $m = 1, \cdots, 500$, where G_i is simulated from the exponential distribution exp(1). I calculated 95% perturbed CI^N's based on the 500 perturbed samples because of the better approximation to the empirical Cis. The perturbed CIs of the 17 identified genes and 13 identified proteins are displayed in Figure 2.9. Only 5 biomarkers do not cover 0, including 4 proteins: 4E-BP1, 4E-BP1_pT70, Akt_pS473, and Ret_pY905 and 1 gene: STOM. This may indicate that many of the identified biomarkers are false positives.

Figure 2.9.: 95% perturbed CI^Ns for selected biomarkers (lung cancer application)



2.5.2 Justification for the perturbation method

This section justifies the resampling-based perturbation method. First, to ensure the validity of the justification, we required the following regularity conditions, as in the paper by Minnier et al. [100]:

C1: $P\{L(\boldsymbol{\theta};\boldsymbol{D})\}$ has a unique minimum at $\boldsymbol{\theta}_0$ and a continuous secondary derivative with a positive definite $\mathbb{A} = \partial^2 P\{l(\boldsymbol{\theta};\boldsymbol{D})\}/\partial\boldsymbol{\theta}\boldsymbol{\theta}^T|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} > 0$, where P is the probability measure generated by data $\boldsymbol{D}$.

C2: The class of functions indexed by $\boldsymbol{\theta}, \{L(\boldsymbol{\theta};\boldsymbol{D})|\boldsymbol{\theta}\in\Omega\}$, is Glivenko-Cantelli [101], where Ω is the compact parameter space containing $\boldsymbol{\theta}_0$.

C3: There exists a quasi-derivative function $U(\boldsymbol{\theta};\boldsymbol{D})$ for $L(\boldsymbol{\theta};\boldsymbol{D})$ such that for any positive sequence $\delta_n \to 0$, we have the following denotations.

(a). $P\left\{U^{\otimes 2}(\boldsymbol{\theta}_0;\boldsymbol{D})\right\} = \mathbb{B}$, a positive definite matrix.

(b). $P\{L(\boldsymbol{\theta};D) - L(\boldsymbol{\theta}_0;\boldsymbol{D}) - U(\boldsymbol{\theta}_0;\boldsymbol{D})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\} = \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbb{A}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + o(||\boldsymbol{\theta}-\boldsymbol{\theta}_0||^2)$, where $||\boldsymbol{\theta}-\boldsymbol{\theta}_0|| \leqslant \delta_n$.

(c). $P_n\{L(\boldsymbol{\theta}_1;\boldsymbol{D}) - L(\boldsymbol{\theta}_2;\boldsymbol{D}) - U(\boldsymbol{\theta}_2;\boldsymbol{D})(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\} = \frac{1}{2}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^T\mathbb{A}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2) + o(||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||^2 + n^{-1/2}||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||)$, almost surely, uniformly in $||\boldsymbol{\theta}_1-\boldsymbol{\theta}_0|| \leqslant \delta_n, ||\boldsymbol{\theta}_2-\boldsymbol{\theta}_0|| \leqslant \delta_n$, where P_n denotes the empirical measure: $\forall f \in \{L(\boldsymbol{\theta};\boldsymbol{D})\}, P_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$.

To prove (i): $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^*_\Delta - \boldsymbol{\theta}_{0\Delta})$ converges in distribution to $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, there are the following three steps.

(1) First, we show that $\widetilde{\boldsymbol{\theta}}^* \xrightarrow{P} \boldsymbol{\theta}_0$.

$$
\begin{aligned}
&|\widetilde{L}^*(\boldsymbol{\theta}) - P\{L(\boldsymbol{\theta};\boldsymbol{D})\}| \\
&\leqslant |\widetilde{L}^*(\boldsymbol{\theta}) - \widetilde{L}(\boldsymbol{\theta})| + |\widetilde{L}(\boldsymbol{\theta}) - P\left\{\widetilde{L}(\boldsymbol{\theta};\boldsymbol{D})\right\}| \\
&= |\frac{1}{n}\sum_{i=1}^n L(\theta;D_i)G_i - \frac{1}{n}\sum_{i=1}^n L(\theta;D_i)| + |\frac{1}{n}\sum_{i=1}^n \widetilde{L}(\theta;D_i) - P\left\{\widetilde{L}(\boldsymbol{\theta};\boldsymbol{D})\right\}| \\
&= |P_n^*\{L(\boldsymbol{\theta};\boldsymbol{D})\} - P_n\{L(\boldsymbol{\theta};\boldsymbol{D})\}| + |P_n\left\{\widetilde{L}(\boldsymbol{\theta};\boldsymbol{D})\right\} - P\left\{\widetilde{L}(\boldsymbol{\theta};\boldsymbol{D})\right\}|
\end{aligned}
$$

By Corollary 10.14 in Kosorok et al. [101], $\underset{\theta\in\Omega}{sup}|P_n^*\{L(\boldsymbol{\theta};\boldsymbol{D})\}-P_n\{L(\boldsymbol{\theta};\boldsymbol{D})\}|\to 0$ as $n\to\infty$. In addition, since $\{L(\boldsymbol{\theta};\boldsymbol{D})|\boldsymbol{\theta}\in\Omega\}$ is Glivenko-Cantelli, $\underset{\theta\in\Omega}{sup}|P_n\left\{\widetilde{L}(\boldsymbol{\theta};\boldsymbol{D})\right\}-P\left\{\widetilde{L}(\boldsymbol{\theta};\boldsymbol{D})\right\}|\to 0$ as $n\to\infty$ [101]. Thus, $|\widetilde{L}^*(\boldsymbol{\theta})-P\{L(\boldsymbol{\theta};\boldsymbol{D})\}|$ uniformly converges to 0.

Then, by Condition C1 and Theorem 2.1 in Newey et al. [102], $\widetilde{\boldsymbol{\theta}}^*\xrightarrow{P}\boldsymbol{\theta}_0$.

(2) In the second step, we show that $\hat{\boldsymbol{\theta}}^*\xrightarrow{P}\boldsymbol{\theta}_0$.

Since it is proved that $\widetilde{\boldsymbol{\theta}}^*\xrightarrow{P}\boldsymbol{\theta}_0$, $|\widetilde{a}_j^*|\xrightarrow{P}|a_{0j}|$ and $|\widetilde{b}_j^*|\xrightarrow{P}|b_{0j}|$.

For the SCAD penalty, $p'_{\delta_{1n}^*j}(|\widetilde{a}_j^*|)=\delta_{1n}I(|\widetilde{a}_j^*|\leqslant\delta_{1n})+(a\delta_{1n}-|\widetilde{a}_j^*|)_+I(|\widetilde{a}_j^*|>\delta_{1n})/(a-1)$. When $a_{0j}\neq 0$, we have $\delta_{1n}\to 0$ and $|\widetilde{a}_j^*|\xrightarrow{P}|a_{0j}|$; thus, $I(|\widetilde{a}_j^*|\leqslant\delta_{1n})\xrightarrow{P}0$ and $(a\delta_{1n}-|\widetilde{a}_j^*|)_+\xrightarrow{P}0$. When $a_{0j}=0$, we have $\delta_{1n}\to 0$ and $(a\delta_{1n}-|\widetilde{a}_j^*|)_+\xrightarrow{P}0$. Therefore, $p'_{\delta_{1n}^*j}(|\widetilde{a}_j^*|)\xrightarrow{P}0$.

Since $\boldsymbol{\theta}$ lies in a compact space, $\sum_{j=1}^{p_1}p'_{\delta_{1n}^*j}(|\widetilde{a}_j^*|)|a_j|\leqslant\tau\sum_{j=1}^{p_1}p'_{\delta_{1n*}^*j}(|\widetilde{a}_j^*|)\leqslant||\boldsymbol{A}||\boldsymbol{B}_n$, where $\tau=max\{|a_j|\},\boldsymbol{B}_n=o_P(1)$, and $p'_{\delta_{1n}^*j}(|\widetilde{a}_j^*|)\xrightarrow{P}0$ for each j, by Lemma 2.9 in Newey et al. [102], $\underset{\theta\in\Omega}{sup}|\sum_{j=1}^{p_1}p'_{\delta_{1n}^*j}(|\widetilde{a}_j^*|)|a_j||\xrightarrow{P}0$.

Similarly, $\underset{\theta\in\Omega}{sup}|\sum_{j=1}^{p_2}p'_{\delta_{2n}^*j}(|\widetilde{b}_j^*|)|b_j||\xrightarrow{P}0$.

$$
\begin{aligned}
&|\hat{L}^*(\boldsymbol{\theta})-P\{L(\boldsymbol{\theta};\boldsymbol{D})\}|\\
&=|\widetilde{L}^*(\boldsymbol{\theta})+\sum_{j=1}^{p_1}p'_{\delta_{1n}^*j}(|\widetilde{a}_j^*|)|a_j|+\sum_{j=1}^{p_2}p'_{\delta_{2n}^*j}(|\widetilde{b}_j^*|)|b_j|-P\{L(\boldsymbol{\theta};\boldsymbol{D})\}|\\
&\leqslant|\widetilde{L}^*(\boldsymbol{\theta})-P\{L(\boldsymbol{\theta};\boldsymbol{D})\}|+|\sum_{j=1}^{p_1}p'_{\delta_{1n}^*j}(|\widetilde{a}_j^*|)|a_j||+|\sum_{j=1}^{p_2}p'_{\delta_{2n}^*j}(|\widetilde{b}_j^*|)|b_j||
\end{aligned}
$$

Given that $\underset{\theta\in\Omega}{sup}|\sum_{j=1}^{p_1}p'_{\delta_{1n}^*j}(|\widetilde{a}_j^*|)|a_j||\xrightarrow{P}0$, $\underset{\theta\in\Omega}{sup}|\sum_{j=1}^{p_2}p'_{\delta_{2n}^*j}(|\widetilde{b}_j^*|)|b_j||\xrightarrow{P}0$, and $|\widetilde{L}^*(\boldsymbol{\theta})-P\{L(\boldsymbol{\theta};\boldsymbol{D})\}|$ uniformly converges to 0, $|\hat{L}^*(\boldsymbol{\theta})-P\{L(\boldsymbol{\theta};\boldsymbol{D})\}|$ uniformly converges to 0. Then, with an argument similar to that of step (1), we have $\hat{\boldsymbol{\theta}}^*\xrightarrow{P}\boldsymbol{\theta}_0$.

(3) Finally, we show that $\hat{\boldsymbol{\theta}}^*$ converges to $\boldsymbol{\theta}_0$ with the rate of $n^{-\frac{1}{2}}$: $||\hat{\boldsymbol{\theta}}^*-\boldsymbol{\theta}_0||=O_{P^*}(n^{-\frac{1}{2}})$. It is sufficient to show that for any $\epsilon>0$, there exists $C>0$ such that

$$
P^*\left\{\underset{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||\geqslant Cn^{-\frac{1}{2}}}{inf}\hat{L}^*(\boldsymbol{\theta})>\hat{L}^*(\boldsymbol{\theta}_0)\right\}>1-\varepsilon \tag{2.17}
$$

Consider $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}$. By condition C3(b), we have

$$\begin{aligned}
&\frac{P_n\left\{L(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}) - L(\boldsymbol{\theta}_0) - n^{-\frac{1}{2}}U(\boldsymbol{\theta}_0;\boldsymbol{D})^T\mathbf{u}\right\} - \frac{1}{2}n^{-1}\mathbf{u}^T\mathbb{A}\mathbf{u}}{||n^{-\frac{1}{2}}\mathbf{u}||} \\
&= \frac{\frac{1}{2}(n^{-\frac{1}{2}}\mathbf{u})^T\mathbb{A}n^{-\frac{1}{2}}\mathbf{u} + o(||n^{-\frac{1}{2}}\mathbf{u}||^2) - \frac{1}{2}n^{-1}\mathbf{u}^T\mathbb{A}\mathbf{u}}{||n^{-\frac{1}{2}}\mathbf{u}||} \\
&= \frac{o(||n^{-\frac{1}{2}}\mathbf{u}||^2)}{||n^{-\frac{1}{2}}\mathbf{u}||} = o_P(1)
\end{aligned}$$

uniformly in $\mathbf{u}$. By the multiplier central limit theorem (Theorem 10.1) in Kosorok et al. [101],

$$\frac{P_n^*\left\{L(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u})\boldsymbol{G} - L(\boldsymbol{\theta}_0)\boldsymbol{G} - n^{-\frac{1}{2}}U(\boldsymbol{\theta}_0;\boldsymbol{D})^T\mathbf{u}\boldsymbol{G}\right\} - \frac{1}{2}n^{-1}\mathbf{u}^T\mathbb{A}\mathbf{u}}{||n^{-\frac{1}{2}}\mathbf{u}||} = o_{P^*}(1)$$

uniformly in $\mathbf{u}$. Since $P_n^*\left\{L(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u})\boldsymbol{G} - L(\boldsymbol{\theta}_0)\boldsymbol{G} - n^{-\frac{1}{2}}U(\boldsymbol{\theta}_0;\boldsymbol{D})^T\mathbf{u}\boldsymbol{G}\right\} = \frac{1}{n}\sum_{i=1}^n\left\{L(\boldsymbol{\theta} + n^{-\frac{1}{2}}\mathbf{u})G_i - L(\boldsymbol{\theta})G_i - n^{-\frac{1}{2}}U(\boldsymbol{\theta}_0;\boldsymbol{D})^T\mathbf{u}G_i\right\} = \widetilde{L}^*(\boldsymbol{\theta}_0+n^{-\frac{1}{2}}\mathbf{u}) - \widetilde{L}^*(\boldsymbol{\theta}_0) - n^{-\frac{1}{2}}P_n\left\{U(\boldsymbol{\theta}_0;\boldsymbol{D})\boldsymbol{G}\right\}\mathbf{u}$, we have

$$\widetilde{L}^*(\boldsymbol{\theta}_0+n^{-\frac{1}{2}}\mathbf{u}) - \widetilde{L}^*(\boldsymbol{\theta}_0) = n^{-\frac{1}{2}}P_n\left\{U(\boldsymbol{\theta}_0;\boldsymbol{D})\boldsymbol{G}\right\}\mathbf{u} + \frac{1}{2}n^{-1}\mathbf{u}^T\mathbb{A}\mathbf{u} + o_{P^*}(n^{-1}||\mathbf{u}||). \quad (2.18)$$

From (2.18), it can be shown that

$$\begin{aligned}
&n\left\{\hat{L}^*(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}) - \hat{L}^*(\boldsymbol{\theta}_0)\right\} \\
&= n\left\{\widetilde{L}^*(\boldsymbol{\theta}_0 + n^{-\frac{1}{2}}\mathbf{u}) + \sum_{j=1}^{p_1}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|a_{0j} + n^{-\frac{1}{2}}u_j| + \sum_{j=1}^{p_2}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|b_{0j} + n^{-\frac{1}{2}}u_j|\right. \\
&\left.-\widetilde{L}^*(\boldsymbol{\theta}_0) - \sum_{j=1}^{p_1}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|a_{0j}| - \sum_{j=1}^{p_2}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|b_{0j}|\right\} \\
&= n^{\frac{1}{2}}(P_n\left\{U(\boldsymbol{\theta}_0;\boldsymbol{D})\boldsymbol{G}\right\} - P\left\{U(\boldsymbol{\theta}_0;\boldsymbol{D})\boldsymbol{G}\right\})\mathbf{u} + n\sum_{j=1}^{p_1}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)(|a_{0j} + n^{-\frac{1}{2}}u_j| - |a_{0j}|) \\
&+n\sum_{j=1}^{p_2}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)(|b_{0j} + n^{-\frac{1}{2}}u_j| - |b_{0j}|) + \frac{1}{2}\mathbf{u}^T\mathbb{A}\mathbf{u} + o_{P^*}(||\mathbf{u}||) \\
&= G_n\left\{U(\boldsymbol{\theta}_0;\boldsymbol{D})\boldsymbol{G}\right\}\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbb{A}\mathbf{u} + n\sum_{j=1}^{p_1}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)(|a_{0j} + n^{-\frac{1}{2}}u_j| - |a_{0j}|) \\
&+n\sum_{j=1}^{p_2}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)(|b_{0j} + n^{-\frac{1}{2}}u_j| - |b_{0j}|) + o_{P^*}(||\mathbf{u}||^2 + ||\mathbf{u}||), \quad (2.19)
\end{aligned}$$

where $G_n = n^{\frac{1}{2}}(P_n - P)$. When $\mathbf{u} = C > 0$, (2.19) apparently deviates from 0, then (2.17) holds.

With steps (1), (2), and (3), $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0)$ converges in distribution to $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

Then, we prove (ii): $P^*(\hat{\boldsymbol{\theta}}^*_{\Delta^C} = 0) \to 1$.

It suffices to show that for any constant C and given $\widetilde{\boldsymbol{\theta}}_\Delta$ such that $||\widetilde{\boldsymbol{\theta}}_\Delta - \boldsymbol{\theta}_{0\Delta}|| = O_{P^*}(n^{-\frac{1}{2}})$

$$P^*\left\{\underset{||\boldsymbol{\theta}_{\Delta^C}||\leqslant Cn^{-\frac{1}{2}}}{\operatorname{argmin}} \hat{L}^*[(\widetilde{\boldsymbol{\theta}}_\Delta^T, \boldsymbol{\theta}_{\Delta^C}^T)^T] = 0\right\} \to 1 \tag{2.20}$$

Let $\widetilde{\mathbf{u}}_\Delta$ and $\mathbf{u}_{\Delta^C}$ denote $n^{\frac{1}{2}}(\widetilde{\boldsymbol{\theta}}_\Delta - \boldsymbol{\theta}_{0\Delta})$ and $n^{\frac{1}{2}}\boldsymbol{\theta}_{\Delta^C}$, respectively. By (2.19), we have

$$\begin{aligned}
&n[\hat{L}^*\left\{(\boldsymbol{\theta}_{0\Delta}^T + n^{-\frac{1}{2}}\widetilde{\mathbf{u}}_\Delta^T, n^{-\frac{1}{2}}\widetilde{\mathbf{u}}_{\Delta^C}^T)^T\right\} - \hat{L}^*\left\{(\boldsymbol{\theta}_{0\Delta}^T + n^{-\frac{1}{2}}\widetilde{\mathbf{u}}_\Delta^T, 0^T)^T\right\}] \\
&= [G_n\left\{U(\boldsymbol{\theta}_0; \boldsymbol{D})_{\Delta^C}^T \boldsymbol{G}\right\} + \widetilde{\mathbf{u}}_\Delta^T \mathbb{A}_{12}]\mathbf{u}_{\Delta^C} + \frac{1}{2}\mathbf{u}_{\Delta^C}^T \mathbb{A}_{22}\mathbf{u}_{\Delta^C} + \\
&n\left\{\sum_{j\in\Delta^C} p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|n^{-\frac{1}{2}}u_j| + \sum_{j\in\Delta} p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)(|a_{0j} + n^{-\frac{1}{2}}\cdot 0| - |a_{0j}|)\right\} \\
&+n\left\{\sum_{j\in\Delta^C} p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|n^{-\frac{1}{2}}u_j| + \sum_{j\in\Delta} p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)(|b_{0j} + n^{-\frac{1}{2}}\cdot 0| - |b_{0j}|)\right\} \\
&+o_{P^*}(||\mathbf{u}_{\Delta^C}||^2 + ||\mathbf{u}_{\Delta^C}||) \\
&= [G_n\left\{U(\boldsymbol{\theta}_0; \boldsymbol{D})_{\Delta^C}^T \boldsymbol{G}\right\} + \widetilde{\mathbf{u}}_\Delta^T \mathbb{A}_{12}]\mathbf{u}_{\Delta^C} + \frac{1}{2}\mathbf{u}_{\Delta^C}^T \mathbb{A}_{22}\mathbf{u}_{\Delta^C} + n^{\frac{1}{2}}\sum_{j\in\Delta^C} p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|u_j| + \\
&n^{\frac{1}{2}}\sum_{j\in\Delta^C} p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|u_j| + o_{P^*}(||\mathbf{u}_{\Delta^C}||^2 + ||\mathbf{u}_{\Delta^C}||) \\
&= \sum_{j\in\Delta^C}\left\{n^{\frac{1}{2}}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|u_j| + n^{\frac{1}{2}}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|u_j|\right\} + R_n(\mathbf{u}_{\Delta^C}),
\end{aligned} \tag{2.21}$$

where $sup||\mathbf{u}_{\Delta^C}|| \leqslant CR_n(\mathbf{u}_{\Delta^C})/(||\mathbf{u}_{\Delta^C}||^2 + ||\mathbf{u}_{\Delta^C}||) = o_{P^*}(1)$, from which we have $sup||\mathbf{u}_{\Delta^C}|| \leqslant o_{P^*}(1)$ and $R_n(\mathbf{u}_{\Delta^C}) \leqslant C(||\mathbf{u}_{\Delta^C}||^2 + ||\mathbf{u}_{\Delta^C}||)$. Thus, there exists $C_0 > 0$, which lets $P^*\left\{R_n(\mathbf{u}_{\Delta^C}) \leqslant C_0\sum_{j\in\Delta^C}|u_j|\right\} \geqslant 1 - \varepsilon$ for $||\mathbf{u}_{\Delta^C}|| < C$.

In addition, it is shown in Zou et al. [45] that $n^{\frac{1}{2}}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|) \xrightarrow{P} \infty$ and $n^{\frac{1}{2}}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|) \xrightarrow{P} \infty$for $j \in \Delta^C$. Thus, for any $\epsilon > 0$, there exists $C_1 > C_0 > 0$ such that

$$P^*\left\{\sum_{j\in\Delta^C}[n^{\frac{1}{2}}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|u_j| + n^{\frac{1}{2}}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|u_j|] \geqslant C_1\sum_{j\in\Delta^C}|u_j|\right\} \geqslant 1-\varepsilon.$$

Therefore,

$$P^*\left\{\sum_{j\in\Delta^C}[n^{\frac{1}{2}}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|u_j| + n^{\frac{1}{2}}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|u_j|] - R_n(u_{\Delta^C}) \geqslant (C_1-C_0)\sum_{j\in\Delta^C}|u_j|\right\} \geqslant 1-\varepsilon \geqslant 0,$$

which implies that $n[\hat{L}^*\left\{(\widetilde{\boldsymbol{\theta}}^T_\Delta, n^{-\frac{1}{2}}u^T_{\Delta^C})^T\right\} - \hat{L}^*\left\{(\widetilde{\boldsymbol{\theta}}^T_\Delta, 0^T)^T\right\}] \geqslant 0$, then (2.20) holds.

Finally, we prove (iii): The distribution of $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^*_\Delta - \hat{\boldsymbol{\theta}}_\Delta)|\boldsymbol{D}$ approximates to the unconditional distribution of $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_\Delta - \boldsymbol{\theta}_{0\Delta})$.

Since $P^*(\hat{\boldsymbol{\theta}}^*_{\Delta^C} = 0) \to 1$, we have $\widehat{\boldsymbol{\theta}}^*_\Delta = \text{argmin}\widehat{L}^*_\Delta(\boldsymbol{\theta}_\Delta) = \text{argmin}\widehat{L}^*\left\{(\boldsymbol{\theta}^T_\Delta, 0^T)^T\right\}$.

Denote $\hat{\mathbf{u}}^{(n)}_\Delta = \text{argmin}\widehat{L}^*_\Delta(\boldsymbol{\theta}_{0\Delta} + n^{-\frac{1}{2}}\mathbf{u}_\Delta)$, where

$$\begin{aligned}
&\widehat{L}^*_\Delta(\boldsymbol{\theta}_{0\Delta} + n^{-\frac{1}{2}}u_\Delta)\\
&= \widetilde{L}^*_\Delta(\boldsymbol{\theta}_{0\Delta} + n^{-\frac{1}{2}}\mathbf{u}_\Delta) + \sum_{j\in\Delta}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|a_{0j} + n^{-\frac{1}{2}}u_j| + \sum_{j\in\Delta}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|b_{0j} + n^{-\frac{1}{2}}u_j|\\
&= \widetilde{L}^*\left\{(\boldsymbol{\theta}^T_{0\Delta} + n^{-\frac{1}{2}}\mathbf{u}^T_\Delta, \mathbf{0}^T)^T\right\} + \sum_{j\in\Delta}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|a_{0j} + n^{-\frac{1}{2}}u_j| + \sum_{j\in\Delta}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|b_{0j} + n^{-\frac{1}{2}}u_j|\\
&= \frac{1}{n}\sum_{i=1}^n\left\{L[(\boldsymbol{\theta}^T_{0\Delta} + n^{-\frac{1}{2}}\mathbf{u}^T_\Delta, \mathbf{0}^T)^T; D_i]G_i\right\} + \sum_{j\in\Delta}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|a_{0j} + n^{-\frac{1}{2}}u_j|\\
&+ \sum_{j\in\Delta}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|b_{0j} + n^{-\frac{1}{2}}u_j|\\
&= P_n\left\{L[(\boldsymbol{\theta}^T_{0\Delta} + n^{-\frac{1}{2}}\mathbf{u}^T_\Delta, \mathbf{0}^T)^T; D_i]G_i\right\} + \sum_{j\in\Delta}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)|a_{0j} + n^{-\frac{1}{2}}u_j|\\
&+ \sum_{j\in\Delta}p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)|b_{0j} + n^{-\frac{1}{2}}u_j| \qquad (2.22)
\end{aligned}$$

Note that $\hat{\mathbf{u}}^{(n)}_\Delta$ is also a minimizer of $V^*_n(\mathbf{u}_\Delta) \equiv \widehat{L}^*_\Delta(\boldsymbol{\theta}_{0\Delta} + n^{-\frac{1}{2}}\mathbf{u}_\Delta) - L^*(\boldsymbol{\theta}_0)$, since $L^*(\boldsymbol{\theta}_0)$ is constant.

Again, from (2.19),

$$Vn^*(\mathbf{u}_\Delta) = n^{\frac{1}{2}}\mathbf{u}_\Delta^T P_n\left\{U_\Delta(\boldsymbol{\theta}_0; D)\boldsymbol{G}\right\} + \frac{1}{2}\mathbf{u}_\Delta^T\mathbb{A}_{11}\mathbf{u}_\Delta + n\sum_{j\in\Delta} p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)(|a_{0j} + n^{-\frac{1}{2}}u_j| - |a_{0j}|)$$

$$+n\sum_{j\in\Delta} p'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)(|b_{0j} + n^{-\frac{1}{2}}u_j| - |b_{0j}|) + o_{P^*}(||\mathbf{u}_\Delta|| + ||\mathbf{u}_\Delta||^2). \quad (2.23)$$

When $j \in \Delta$, $a_{0j} \neq 0$, $n^{\frac{1}{2}}(|a_{0j}+n^{-\frac{1}{2}}u_j|-|a_{0j}|) \xrightarrow{P} u_j sgn(a_{0j})$. Also, it is proved by Zou et al.[45] that $n^{\frac{1}{2}}p'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|) \xrightarrow{P} 0$ for the *SCAD* penalty. By Slutsky's theorem, we have $np'_{\delta^*_{1n}j}(|\widetilde{a}^*_j|)(|a_{0j}+n^{\frac{1}{2}}u_j|-|a_{0j}|) = o_{P^*}(1)$. Using the same arguments, $np'_{\delta^*_{2n}j}(|\widetilde{b}^*_j|)(|b_{0j}+ n^{\frac{1}{2}}u_j| - |b_{0j}|) = o_{P^*}(1)$ can be proved. Therefore,

$$Vn^*(\mathbf{u}_\Delta) = \mathbf{u}_\Delta^T\boldsymbol{G}_n\left\{U_\Delta(\boldsymbol{\theta}_0; \boldsymbol{D})\boldsymbol{G}\right\} + \frac{1}{2}\mathbf{u}_\Delta^T\mathbb{A}_{11}\mathbf{u}_\Delta + o_{P^*}(1 + ||\mathbf{u}_\Delta|| + ||\mathbf{u}_\Delta||^2) \quad (2.24)$$

Thus, $\hat{\mathbf{u}}^{(n)}_\Delta = -\mathbb{A}_{11}^{-1}G_n\left\{U_\Delta(\boldsymbol{\theta}_0; \boldsymbol{D})\boldsymbol{G}\right\}+o_{P^*}(1)$. Since $G_n\left\{U_\Delta(\boldsymbol{\theta}_0; \boldsymbol{D})\boldsymbol{G}\right\} \xrightarrow{d} N(0, \mathbb{B}_{11})$, $\hat{\mathbf{u}}^{(n)}_\Delta \xrightarrow{d} N(0, \mathbb{A}_{11}^{-1}\mathbb{B}_{11}\mathbb{A}_{11}^{-1})$. That is, $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^*_\Delta - \boldsymbol{\theta}_{0\Delta}) \xrightarrow{d} N(0, \mathbb{A}_{11}^{-1}\mathbb{B}_{11}\mathbb{A}_{11}^{-1})$.

Using similar arguments, we can obtain $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_\Delta - \boldsymbol{\theta}_{0\Delta}) = -\mathbb{A}_{11}^{-1}G_n\left\{U_\Delta(\boldsymbol{\theta}_0; \boldsymbol{D})\right\} + o_{P^*}(1)$. Therefore, $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^*_\Delta - \hat{\boldsymbol{\theta}}_\Delta) = -\mathbb{A}_{11}^{-1}G_n\left\{U_\Delta(\boldsymbol{\theta}_0; \boldsymbol{D})(\boldsymbol{G}-1)\right\} + o_{P^*}(1)$. Since $-\mathbb{A}_{11}^{-1}G_n\left\{U_\Delta(\boldsymbol{\theta}_0; \boldsymbol{D})(\boldsymbol{G}-1)\right\}|\boldsymbol{D} \xrightarrow{d} N(0, \mathbb{A}_{11}^{-1}\hat{\mathbb{B}}_{11}\mathbb{A}_{11}^{-1})$ and $\hat{\mathbb{B}}_{11} \xrightarrow{P} \mathbb{B}_{11}$, $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^*_\Delta - \hat{\boldsymbol{\theta}}_\Delta)|\boldsymbol{D}$ and $n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_\Delta - \boldsymbol{\theta}_{0\Delta})$ converge in distribution to the same limit.

2.6 Discussion

I have proposed a generalized statistical framework *SIFORM* to jointly model high-throughput *omic* data produced by multiple platforms and to discover the associations between these genetic variables and a disease-associated phenotype. The new method conveniently produces direct rankings of genetic variables in terms of the strength of association with the response variable. Extensive simulation studies demonstrated the superior performance of *SIFORM* in terms of biomarker detection accuracy, regardless of inter-variable correlations, compared to the performances of other penalized variable selection methods. Biological meaningfulness of the proposed method is supported by two TCGA lung adenocarcinoma studies that investigated

the association of the integrated mRNA expression and RPPA protein concentration data with smoking status and discretized survival. In particular, most of the identified proteins either belong to known key pathways for NSCLC tumorigenesis or are prognostic biomarkers in NSCLC. I also discovered some proteins that can potentially predict survival for NSCLC patients or which are associated with the different ways in which NSCLC develops in smokers versus non-smokers. The statistical significance of the regularized estimates is also assessed by a resampling-based method.

One possibility for future research is to focus on allowing for mandatory variables (e.g., biomarkers that are known to be correlated with the phenotype) in the proposed system. In a different direction, information on biological pathways and networks can be intuitively incorporated into the proposed framework to further improve the scientific validity of biomarker detection. Chapter 3 discusses the methodology development, simulations, and real applications for the incorporation of pathway information in the analyses.

3. Biological Pathway Information Incorporated in a Structured Model (PSIFORM)

In this chapter, I focus on incorporating molecular pathway information into the proposed statistical framework, *SIFORM*. The extended version of the structured model is called pathway information incorporated in a structured model (*PSIFORM*). The pathway structure among genetic variables is characterized by a graphical model, and the network-based penalty is used to deliver the pathway information in *PSIFORM*.

This chapter is organized as follows. Section 3.1 provides the motivation of the project. In section 3.2, I apply a non-convex network-based penalty to our framework to characterize the correlation structures between the genetic variables on biological pathways. The parameter estimation procedure is also discussed in this section. Section 3.3 presents extensive simulation studies to compare the proposed method with other network-based penalized regression methods. Section 3.4 contains a kidney cancer application, through which I show that *PSIFORM* can jointly explore the associations of mRNA expression and protein expression with discretized survival while utilizing existing pathway information from public databases. Section 3.5 concludes with a summary and a discussion about future research directions.

3.1 Motivation

Genes regulate disease progression and impact phenotypes through functional groups called pathways. The genes from the same pathway usually have similar functions and are highly correlated. Incorporating pathway information into the integrative analysis of *omic* data can lead to more accurate and improved interpretability

of the results [4, 5]. Many public and commercial databases have been developed to collect the existing biological knowledge (e.g., protein-protein interaction networks, biological pathways). Such databases enable us to utilize the existing pathway information [26, 27].

Numerous methods have been proposed to incorporate the biological pathway structures into statistical analyses. Most methods use graphical models to describe the complex structure of genetic variables between and within pathways. Specifically, the graphical model-based methods can be classified into two major categories: the Bayesian approach [4, 20, 24, 25] and the penalized regression approach [5, 21, 22, 23]. I focus on the penalized regression approach because it can be easily fitted into the proposed framework.

Section 1.4 provides a brief introduction to several popular penalized regression models for pathway incorporation. Among them, *Grace*, *aGrace*, and grouped L_γ-norm assume that the genes from the same pathway must have smoothed-regression coefficients [5, 21, 22]. This assumption may be too stringent in real biological processes. To address this problem, Kim et al. proposed a new set of network-based penalty functions, $TTLP_I$ and $LTLP_I$, which only assume that the genes from the same pathway are more likely to participate (or not participate) together in the same biological process [23]. However, this flexible method does not take multi-platform data into consideration.

Therefore, I modify the network-based penalty of Kim et al. [23] and apply it to *SIFORM*. By doing so, I model both the common structured factor across multiple data types and the network structure within the data types in *PSIFORM*. The biological relevance of the results and prediction accuracy can be further improved in *PSIFORM*.

3.2 Methodology

In this section, I follow the same framework set-up as described in Section 2.2.1. Recall that the $n \times 1$ vector $\mathbf{y}$ denotes the phenotypes of n subjects, each of whom has had continuously measured genetic variables collected from two different platforms. The $n \times p_1$ matrix $\mathbf{X_1}$ and $n \times p_2$ matrix $\mathbf{X_2}$ respectively contain the intensity measurements of p_1 genetic variables obtained from platform 1 and the intensity measurements of p_2 genetic variables obtained from platform 2. The generalized statistical framework *PSIFORM* is composed of

$$\begin{aligned} g\left\{E(\mathbf{X_1})\right\} &= \boldsymbol{\alpha}_1 + \boldsymbol{CA} + \boldsymbol{\varepsilon}_1, \\ g\left\{E(\mathbf{X_2})\right\} &= \boldsymbol{\alpha}_2 + \boldsymbol{CB} + \boldsymbol{\varepsilon}_2. \end{aligned} \tag{3.1}$$

Similar to Section 2.2.1, for demonstration, I focus on continuously measured genetic variables and use identity link $g(\mu) = \mu$. The $n \times 1$ parameter vectors $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ respectively correspond to baseline sample effects in two sets of genetic data. The jth columns of the $n \times p_1$ residual matrix $\boldsymbol{\varepsilon_1}$ and the $n \times p_2$ residual matrix $\boldsymbol{\varepsilon_2}$ follow mean-0 normal distributions with respective variances σ_{1j}^2 and σ_{2j}^2. The $n \times 1$ parameter vector $\boldsymbol{C} = \{c_1, ..., c_1, ..., c_l, ..., c_l, ..., c_K, ..., c_K\}^T$ is structured to deliver the K-categorical phenotype information and is employed in both equations to represent the common intrinsic sample characteristics in the two sets of genetic data. Constraint $\sum_{i=1}^{n} c_i^2 = 1$ is imposed for model identifiability. The length-p_1 vector $\boldsymbol{A}$ and length-p_2 vector $\boldsymbol{B}$ contain the weights that the genetic variables contribute to the structured sample characteristics shared by the two sets of genetic data.

Parameter Estimation

To identify the important genetic variables associated with the phenotype of interest and incorporate pathway information at the same time, I impose a network-based regularization on the elements of the parameter vectors $\boldsymbol{A}$ and $\boldsymbol{B}$. First, I adopt

SCAD, the same penalization method used in Section 2.2.2, for variable selection. The *SCAD* estimators can be explicitly given by

$$\hat{\beta} = \begin{cases} sgn(\beta)(|\beta| - \delta)_+ & \text{, when } |\beta| \leqslant 2\delta \\ \{(a-1)\beta - sgn(\beta)a\delta\}/(a-1) & \text{, when } 2\delta < |\beta| \leqslant a\delta \\ \beta & \text{, when } \beta > a\delta, \end{cases} \tag{3.2}$$

where a = 3.7, as suggested by Fan et al. [32]. Equation (3.2) shows that the *SCAD* penalization method shrinks any coefficients $< \delta$ to 0 [32].

Second, for incorporating pathway information, it is assumed that two neighboring genes in the same subnetwork tend to be selected (or eliminated) simultaneously [23]. I adopt the penalty $\sum_{i\sim j}\left|I(\frac{|\beta_i|}{\omega_i} \neq 0) - I(\frac{|\beta_j|}{\omega_j} \neq 0)\right|$ to model the network structure of the genetic variables [23, 49]. However, the non-continuous indicator function $I(\cdot)$ is not computationally tractable. To solve this problem, I use a truncated *Lasso* penalty (*TLP*) as a computational surrogate of $I(|\beta| \neq 0)$, where *TLP* $J_\tau(|\beta|) = min(\frac{|\beta|}{\tau}, 1)$ approaches $I(|\beta| \neq 0)$ as $\tau \to 0^+$. The parameter τ determines the degree of approximation [23, 49]. That is, any genetic variables with coefficients $< \tau$ will be eliminated from the model. To make the thresholds of *SCAD* and *TLP* in accordance with each other, I let δ and τ be the same.

Thus, given tuning parameters $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, the network-based penalty for incorporating pathway information has the following form:

$$\begin{aligned} P(\boldsymbol{A}, \boldsymbol{B}) &= P^{SCAD}_{\lambda_1}(\boldsymbol{A}) + P^{TLP}_{\lambda_2,\lambda_1}(\boldsymbol{A}) + P^{SCAD}_{\lambda_3}(\boldsymbol{B}) + P^{TLP}_{\lambda_4,\lambda_3}(\boldsymbol{B}) \\ &= \lambda_1 \sum_{j=1}^{p_1} \left\{\lambda_1^2 - (|a_j| - \lambda_1)^2 I(|a_j| < \lambda_1)\right\} \\ &+ \lambda_2 \sum_{j\sim j'} \left| J_{\lambda_1}(\frac{|a_j|}{\omega_j}) - J_{\lambda_1}(\frac{|a_{j'}|}{\omega_{j'}}) \right| \\ &+ \lambda_3 \sum_{j=1}^{p_2} \left\{\lambda_3^2 - (|b_j| - \lambda_3)^2 I(|b_j| < \lambda_3)\right\} \\ &+ \lambda_4 \sum_{j\sim j'} \left| J_{\lambda_3}(\frac{|b_j|}{\omega_j}) - J_{\lambda_3}(\frac{|b_{j'}|}{\omega_{j'}}) \right|. \end{aligned} \tag{3.3}$$

The log-likelihood of the two sets of observed genetic data, l, takes the same form as equation (2.2); therefore, the parameter estimates are obtained via minimizing the following negative penalized log-likelihood,

$$S = -l + P_{\lambda_1}^{SCAD}(A) + P_{\lambda_2,\lambda_1}^{TLP}(A) + P_{\lambda_3}^{SCAD}(B) + P_{\lambda_4,\lambda_3}^{TLP}(B). \tag{3.4}$$

However, minimizing the negative penalized likelihood function is computationally challenging, because the penalty functions are non-concave and non-differentiable. The following transformations are applied to make the penalty functions differentiable.

First, for SCAD penalties $P_{\lambda_1}(A)$ and $P_{\lambda_3}(B)$, I adopt the local linear approximation as in Section 2.2.2 [45]. The linear approximation to $P_{\lambda_1}(|a_j|)$ has the same form as equation (2.4):

$$P_{\lambda_1}(|a_j|) \approx P_{\lambda_1}(|a_j^{(k)}|) + P'_{\lambda_1}(|a_j^{(k)}|)(|a_j| - |a_j^{(k)}|), \tag{3.5}$$

where $a_j^{(k)}$ is the value of a_j estimated at step k.

Second, the non-convex TLP can be decomposed into a difference of two convex functions by DC programming [50]. Specifically, two tricks are used here: (1) The non-convex function $J_\tau(|z|)$ can be decomposed into the difference between two convex functions: $J_\tau(|z|) = \frac{1}{\tau}(|z| - max(|z| - \tau, 0))$; and (2) $|f_1 - f_2|$ can be decomposed as $|f_1 - f_2| = 2max(f_1, f_2) - (f_1 + f_2)$, where f_1 and f_2 are convex functions [23].

For example, after applying these two DC decompositions to $P_{\lambda_2,\lambda_1}(|a_j|)$, I have

$$P_{\lambda_2,\lambda_1}(|a_j|) = \frac{\lambda_2}{\lambda_1}\sum_{j\sim j'}[2max(u_{j,j'}, v_{j,j'}) - (u_{j,j'} + v_{j,j'})], \tag{3.6}$$

where $u_{j,j'} = \frac{|a_j|}{\omega_j} + max(\frac{|a_{j'}|}{\omega_{j'}} - \lambda_1, 0)$ and $v_{j,j'} = \frac{|a_{j'}|}{\omega_{j'}} + max(\frac{|a_j|}{\omega_j} - \lambda_1, 0)$.

After replacing non-differentiable penalties with the above transformations, the penalty function (3.3) can be rewritten as the difference between two convex functions P_1 and P_2, which can be minimized iteratively by the DC algorithm:

$$P(A, B) = P_1 - P_2. \tag{3.7}$$

In this equation,

$$
\begin{aligned}
P_1 = & \sum_{j=1}^{p_1} [P_{\lambda_1}(|a_j^{(k)}|) - P'_{\lambda_1}(|a_j^{(k)}|)|a_j^{(k)}| + P'_{\lambda_1}(|a_j^{(k)}|)|a_j|] \\
& + \sum_{j=1}^{p_2} [P_{\lambda_3}(|b_j^{(k)}|) - P'_{\lambda_3}(|b_j^{(k)}|)|b_j^{(k)}| + P'_{\lambda_3}(|b_j^{(k)}|)|b_j|] \\
& + \frac{\lambda_2}{\lambda_1} \sum_{j \sim j'} 2max(u_{j,j'}, v_{j,j'}) + \frac{\lambda_4}{\lambda_3} \sum_{j \sim j'} 2max(m_{j,j'}, n_{j,j'})
\end{aligned} \tag{3.8}
$$

and

$$
P_2 = \frac{\lambda_2}{\lambda_1} \sum_{j \sim j'} (u_{j,j'} + v_{j,j'}) + \frac{\lambda_4}{\lambda_3} \sum_{j \sim j'} (m_{j,j'} + n_{j,j'}), \tag{3.9}
$$

where $m_{j,j'} = \frac{|b_j|}{\omega_j} + max(\frac{|b_{j'}|}{\omega_{j'}} - \lambda_3, 0)$ and $n_{j,j'} = \frac{|b_{j'}|}{\omega_{j'}} + max(\frac{|b_j|}{\omega_j} - \lambda_3, 0)$. By linearizing P_2 at a current estimate $\hat{A}$ and $\hat{B}$ and ignoring terms independent of A and B, a convex approximation of S can be obtained at the $(k+1)^{th}$ step:

$$
\begin{aligned}
S^{(k+1)} = & \sum_{j=1}^{p_1}\sum_{i=1}^{n}[\frac{(x_{1ij}-\alpha_{1i}-c_i a_j)^2}{2\sigma_{1j}^2}+\frac{1}{2}log(2\pi\sigma_{1j}^2)] \\
& + \sum_{j=1}^{p_2}\sum_{i=1}^{n}[\frac{(x_{2ij}-\alpha_{2i}-c_i b_j)^2}{2\sigma_{2j}^2}+\frac{1}{2}log(2\pi\sigma_{2j}^2)] \\
& + \sum_{j=1}^{p_1}[P_{\lambda_1}(|a_j^{(k)}|)-P'_{\lambda_1}(|a_j^{(k)}|)|a_j^{(k)}|)+P'_{\lambda_1}(|a_j^{(k)}|)|a_j|] \\
& + \sum_{j=1}^{p_2}[P_{\lambda_3}(|b_j^{(k)}|)-P'_{\lambda_3}(|b_j^{(k)}|)|b_j^{(k)}|)+P'_{\lambda_3}(|b_j^{(k)}|)|b_j|] \\
& + \frac{\lambda_2}{\lambda_1}\sum_{j\sim j'}2max(u_{j,j'},v_{j,j'})+\frac{\lambda_4}{\lambda_3}\sum_{j\sim j'}2max(m_{j,j'},n_{j,j'}) \\
& - \Bigg\{\frac{\lambda_2}{\lambda_1}\sum_{j\sim j'}[\frac{a_j sign(\hat{a_j}^{(k)})}{\omega_j}(1+I(\frac{|\hat{a_j}^{(k)}|}{\omega_j}>\lambda_1)) \\
& + \frac{a_{j'} sign(\hat{a_{j'}}^{(k)})}{\omega_{j'}}(1+I(\frac{|\hat{a_{j'}}^{(k)}|}{\omega_{j'}}>\lambda_1))] \\
& + \frac{\lambda_4}{\lambda_3}\sum_{j\sim j'}[\frac{b_j sign(\hat{b_j}^{(k)})}{\omega_j}(1+I(\frac{|\hat{b_j}^{(k)}|}{\omega_j}>\lambda_3)) \\
& + \frac{b_{j'} sign(\hat{b_{j'}}^{(k)})}{\omega_{j'}}(1+I(\frac{|\hat{b_{j'}}^{(k)}|}{\omega_{j'}}>\lambda_3))]\Bigg\}
\end{aligned}
\tag{3.10}
$$

Since $S^{(k+1)}$ is convex, I use Matlab package CVX to minimize it in each iteration step [103].

The closed-form solutions of all the other parameter estimates can be obtained from equations (2.8)-(2.12); therefore, the details are omitted here. In practice, I obtain the estimates via the iterative parameter estimation procedure as described in algorithm 1 in Chapter 2, and the same stopping rule is adopted here. Similarly, the optimal tuning parameters $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are determined by BIC by using a four-dimensional grid search over a pre-determined space [52].

3.3 Simulation

3.3.1 Data description

I conduct extensive simulations to assess the performance of *PSIFORM* and compare it to those of seven regularized regression methods: *Lasso*, *adaLasso*, *SCAD*, *Grace*, *aGrace*, $TTLP_I$, and $LTLP_I$. *Lasso*, *adaLasso*, and *SCAD* can be implemented directly using the respective R packages *ncvreg*, *glmnet*, and *parcor*; the other network-based penalized regression methods are implemented in Matlab. The tuning parameters of all these comparative methods are selected by a 5-fold cross-validation procedure.

As in Section 2.3.1, I assign a phenotype variable $\mathbf{Y}$ categorized into four groups and two high-dimensional genetic profiling matrices, $\mathbf{X}_1$ and $\mathbf{X}_2$, for each sample. In the seven comparative penalized regression methods, all the genetic variables in $\mathbf{X}_1$ and $\mathbf{X}_2$ are treated as the covariates to predict the phenotype. I use seven simulation scenarios to study the various data generation models, inter-genetic marker dependence structures, and residual variances.

In all scenarios, I use set-ups similar to those in three published studies [5, 21, 23] to incorporate the dependence structure among genetic variables. I simulate 120 samples in scenarios 1, 2, 3, 5, and 6, and 60 samples in scenarios 4 and 7. In scenarios 2, 3, 5, and 6, each sample has 110 genetic variables in $\mathbf{X}_1$ and 110 in $\mathbf{X}_2$. The 110 genetic variables consist of 10 independent subnetworks, each including one transcription factor (TF) and 10 target genes (TGs); each TF is connected to each of its 10 TGs with ρ=0.8. The first two networks (that is, the first 22 genetic variables) are correlated with the phenotype variable $\mathbf{Y}$, and the TFs have larger coefficients than the TGs. Scenario 1 is set up in a similar way. The only difference is that the genetic variables in $\mathbf{X}_1$ and $\mathbf{X}_2$ in scenario 1 are mutually independent. Scenarios 4 and 7 consider a sparser situation. In scenarios 4 and 7, the subnetwork

and correlation structures remain the same, but each sample has 660 genetic variables (60 subnetworks accordingly) in $\mathbf{X}_1$ and 110 genetic variables in $\mathbf{X}_2$. The first two subnetworks in $\mathbf{X}_1$ and the first subnetwork in $\mathbf{X}_2$ are correlated with the phenotype variable $\mathbf{Y}$.

In scenarios 1-4, the relationship between $\mathbf{Y}$ and genetic data $\mathbf{X_1}$ and $\mathbf{X_2}$ are generated from model (3.1). In scenarios 5-7, the discrete phenotype $\mathbf{Y}$ is generated from multinomial logistic regression models. Scenarios 1-4 model the inter-genetic dependence structure in residual matrices $\boldsymbol{\varepsilon_1}$ and $\boldsymbol{\varepsilon_2}$. In scenario 1, σ_{1j}^2 and σ_{2j}^2 are independently simulated from $N(0,1)$. In scenarios 2-4, the residuals of the TFs are distributed as $N(0,1)$; conditional on TF, the residuals of the TGs are distributed as $N(0.5\sigma_{TF}^2, 0.75)$, and any two of the residuals are independent with each other. In contrast, scenarios 5-7 model the data dependence structure in $\mathbf{X}_1$ and $\mathbf{X}_2$. The TF genes in $\mathbf{X}_1$ and $\mathbf{X}_2$ are distributed as N (0, 1). Conditional on the TF expression level X_{TF}, the TG expression levels X_{TG}'s are distributed as N ($0.5X_{TF}$, 0.75). Similarly, any two X_{TG}'s are conditionally independent given X_{TF}. Additional details for the seven simulation scenarios are described below.

- Scenario 1: In this scenario, 120 samples are equally assigned to the four categories of the phenotype $\mathbf{y}$, which gives the corresponding $\boldsymbol{C} = (0.033, \cdots, 0.033, 0.067, \cdots, 0.067, 0.100, \cdots, 0.100, 0.133, \cdots, 0.133)$. Each of the 120 samples has 110 genetic variables in data set 1 ($\mathbf{X}_1$) and 110 genetic variables in data set 2 ($\mathbf{X}_2$). In each data set, the first 22 genetic variables are associated with the phenotype, and the weighted magnitudes of the nonzero coefficients in the same subnetwork are close to each other: $\boldsymbol{A}_{1\times 110} = (20, \frac{20}{\sqrt{10}}, \cdots, \frac{20}{\sqrt{10}}, -15, \frac{-15}{\sqrt{10}}, \cdots, \frac{-15}{\sqrt{10}}, 0, \cdots, 0)$ and $\boldsymbol{B}_{1\times 110} = (24, \frac{24}{\sqrt{10}}, \cdots, \frac{24}{\sqrt{10}}, -18, \frac{-18}{\sqrt{10}}, \cdots, \frac{-18}{\sqrt{10}}, 0, \cdots, 0)$. All the genetic variables are mutually independent. The baselines $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ are set to be 0, and the intensity measurements in $\mathbf{X}_1$ and $\mathbf{X}_2$ are simulated from model (3.1).

- Scenario 2: This scenario is the same as scenario 1, except that I assume a block-wise compound symmetric correlation structure among the genetic variables to mimic real biological data.

- Scenario 3: This scenario is the same as scenario 2, except that the nonzero coefficients in the same subnetwork are randomly drawn from a uniform distribution. Specifically, the generated nonzero regression coefficient vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ are: $a_1 = 15, a_2, \cdots, a_6 \sim Unif(5, 15), a_7, \cdots, a_{11} \sim Unif(-15, -5), a_{12} = 10, a_{13}, \cdots, a_{17} \sim Unif(4, 10), a_{18}, \cdots, a_{22} \sim Unif(-10, -4)$ and $b_1 = 16, b_2, \cdots, b_6 \sim Unif(6, 16), b_7, \cdots, b_{11} \sim Unif(-16, -6), b_{12} = 12, b_{17}, \cdots, b_{22} \sim Unif(4, 12), b_{18}, \cdots, b_{22} \sim Unif(-12, -4)$. This scenario is used to determine whether *PSI-FORM* can outperform the methods that have a stringent assumption on the smoothness of the coefficients (e.g., *Grace*, *aGrace*).

- Scenario 4: A sparser setting for the data generated from the model is investigated in this scenario. I consider 60 samples that are equally assigned to the four categories of the phenotype $\mathbf{y}$, with $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively containing 660 and 110 genetic variables. Among the genetic variables, only the first 22 in $\mathbf{X}_1$ and the first 11 in $\mathbf{X}_2$ are the true predictors of the phenotype. The nonzero coefficients in coefficient vectors $\mathbf{A_{(1)}}, \mathbf{A_{(2)}}, \mathbf{A_{(3)}}, \mathbf{B_{(1)}}, \mathbf{B_{(2)}}$, and $\mathbf{B_{(3)}}$ are randomly drawn from $Unif(-25, -5) \cup Unif(5, 25)$.

- Scenario 5: In this scenario, I use a multinomial logistic regression model to generate the phenotype $\mathbf{y}$. I consider a total of 120 samples and 110 genetic variables, where the first 22 variables are true biomarkers, in each of $\mathbf{X_1}$ and $\mathbf{X_2}$. The genetic intensity variables in $\mathbf{X_1}$ and $\mathbf{X_2}$ follow multivariate mean-0 normal distributions. By setting the sparse coefficient vectors $\mathbf{A_{(1)1\times 110}} = (20, \frac{20}{\sqrt{10}}, \cdots, \frac{20}{\sqrt{10}}, -15, \frac{-15}{\sqrt{10}}, \cdots, \frac{-15}{\sqrt{10}}, 0, \cdots, 0), \mathbf{B_{(1)1\times 110}} = (24, \frac{24}{\sqrt{10}}, \cdots, \frac{24}{\sqrt{10}}, -18, \frac{-18}{\sqrt{10}}, \cdots, \frac{-18}{\sqrt{10}}, 0, \cdots, 0), \mathbf{A_{(2)1\times 110}} = (18, \frac{18}{\sqrt{10}}, \cdots, \frac{18}{\sqrt{10}}, -9, \frac{-9}{\sqrt{10}}, \cdots, \frac{-9}{\sqrt{10}}, 0, \cdots, 0),$

$\mathbf{B_{(2)1\times 110}} = (20, \frac{20}{\sqrt{10}}, \cdots, \frac{20}{\sqrt{10}}, -15, \frac{-15}{\sqrt{10}}, \cdots, \frac{-15}{\sqrt{10}}, 0, \cdots, 0), \mathbf{A_{(3)1\times 110}} = (10, \frac{10}{\sqrt{10}}, \cdots, \frac{10}{\sqrt{10}}, -3, \frac{-3}{\sqrt{10}}, \cdots, \frac{-3}{\sqrt{10}}, 0, \cdots, 0), \mathbf{B_{(3)1\times 110}} = (11, \frac{11}{\sqrt{10}}, \cdots, \frac{11}{\sqrt{10}}, -23, \frac{-23}{\sqrt{10}}, \cdots, \frac{-23}{\sqrt{10}}, 0, \cdots, 0)$, we have the logit transformed predictor $\eta_{il} = log(\frac{P(Y_i=l)}{P(Y_i=4)}) = A_{(l)}X_{1i} + B_{(l)}X_{2i}, l = 1, 2, 3$. The probability of $Y_i = l$ is $\mathbf{p_{il}} = \mathbf{e^{\eta_{il}}/(1 + \sum_{k=1}^{3} e^{\eta_{ik}})}$. Accordingly, I sample Y_i from the multinomial distribution $MN(1, p_{i1}, p_{i2}, p_{i3}, p_{i4})$. One of the simulated data sets contains 71, 30, 47, and 92 samples in the four respective categories.

- Scenario 6: This scenario is the same as scenario 5, except that the nonzero coefficients in coefficient vectors $\mathbf{A_{(1)}}, \mathbf{A_{(2)}}, \mathbf{A_{(3)}}, \mathbf{B_{(1)}}, \mathbf{B_{(2)}}$, and $\mathbf{B_{(3)}}$ are randomly drawn from $Unif(-25, -5) \cup Unif(5, 25)$. As an example, a simulated data set produces 34, 39, 27, and 20 samples in the four respective phenotype categories.

- Scenario 7: This scenario investigates the sparser settings for data generated under a multinomial logistic regression model by simulating 60 samples, with $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively containing 660 and 110 genetic variables. Among the genetic variables, only the first 22 in $\mathbf{X}_1$ and the first 11 in $\mathbf{X}_2$ are the true predictors of the phenotype. Similar to scenario 5, the nonzero coefficients in coefficient vectors $\mathbf{A_{(1)}}, \mathbf{A_{(2)}}, \mathbf{A_{(3)}}, \mathbf{B_{(1)}}, \mathbf{B_{(2)}}$, and $\mathbf{B_{(3)}}$ are randomly drawn from $Unif(-25, -5) \cup Unif(5, 25)$. A simulated data set produces 17, 15, 15, and 13 samples in the four respective phenotype categories.

3.3.2 Results

Similar to the scenarios described in Section 2.3.2, each scenario has 100 simulations for stability. The mean and standard error of the true positive rate (TPR), true negative rate (TNR), and false discovery rate (FDR) among the top genetic variables detected that have the largest values of $|A|$ or $|B|$ for our method, and the largest

absolute values of the regression coefficients for the other methods, are reported in Table 3.1. To make a fair comparison, we focus on the top 10 biomarkers detected by the different methods in each scenario.

Table 3.1: Comparison of TPRs, TNRs, and FDRs between *PSIFORM* and the other seven methods under seven simulation scenarios

Scenario	Method	TPR	TNR	FDR
Scenario 1	PSIFORM	1.000(0.002)	0.998(0.006)	0.000(0.000)
	Lasso	0.186(0.061)	0.987(0.012)	0.234(0.130)
	adaLasso	0.204(0.064)	0.994(0.008)	0.165(0.085)
	SCAD	0.105(0.067)	0.971(0.024)	0.390(0.250)
	Grace	0.356(0.053)	0.986(0.008)	0.100(0.100)
	aGrace	0.3542(0.053)	0.985(0.008)	0.100(0.100)
	$TTLP_I$	0.369(0.092)	0.999(0.002)	0.054(0.098)
	$LTLP_I$	0.435(0.155)	0.996(0.005)	0.008(0.031)
Scenario 2	PSIFORM	1.000(0.001)	0.993(0.019)	0.000(0.000)
	Lasso	0.204(0.068)	0.981(0.016)	0.262(0.139)
	adaLasso	0.212(0.062)	0.993(0.008)	0.170(0.089)
	SCAD	0.105(0.072)	0.969(0.026)	0.396(0.269)
	Grace	0.368(0.052)	0.986(0.010)	0.100(0.100)
	aGrace	0.358(0.046)	0.986(0.011)	0.100(0.052)
	$TTLP_I$	0.362(0.136)	0.999(0.002)	0.042(0.103)
	$LTLP_I$	0.557(0.232)	0.994(0.009)	0.042(0.143)
Scenario 3	PSIFORM	1.000(0.001)	0.992(0.024)	0.000(0.000)
	Lasso	0.273(0.061)	0.977(0.099)	0.161(0.110)
	adaLasso	0.308(0.058)	0.988(0.100)	0.103(0.022)
	SCAD	0.110(0.077)	0.964(0.100)	0.305(0.241)

	Grace	0.287(0.106)	0.992(0.006)	0.075(0.188)
	aGrace	0.297(0.141)	0.993(0.006)	0.046(0.148)
	$TTLP_I$	0.391(0.050)	0.999(0.003)	0.018(0.041)
	$LTLP_I$	0.472(0.049)	0.997(0.007)	0.022(0.042)
Scenario 4	PSIFORM	1.000(0.001)	0.989(0.037)	0.000(0.000)
	Lasso	0.200(0.053)	0.999(0.001)	0.101(0.093)
	adaLasso	0.219(0.038)	1.000(0.000)	0.081(0.042)
	SCAD	0.178(0.070)	0.994(0.004)	0.380(0.194)
	Grace	0.266(0.091)	0.999(0.001)	0.095(0.049)
	aGrace	0.289(0.149)	1.000(0.001)	0.056(0.037)
	$TTLP_I$	0.441(0.056)	1.000(0.056)	0.000(0.000)
	$LTLP_I$	0.474(0.058)	1.000(0.000)	0.000(0.000)
Scenario 5	PSIFORM	0.763(0.097)	0.964(0.034)	0.000(0.000)
	Lasso	0.240(0.045)	0.945(0.025)	0.393(0.084)
	adaLasso	0.178(0.072)	0.991(0.012)	0.185(0.105)
	SCAD	0.147(0.055)	0.981(0.009)	0.275(0.131)
	Grace	0.666(0.086)	0.997(0.003)	0.091(0.145)
	aGrace	0.597(0.102)	0.997(0.005)	0.079(0.130)
	$TTLP_I$	0.729(0.103)	1.000(0.000)	0.000(0.000)
	$LTLP_I$	0.777(0.111)	1.000(0.000)	0.000(0.000)
Scenario 6	PSIFORM	0.739(0.175)	0.995(0.057)	0.085(0.227)
	Lasso	0.250(0.066)	0.967(0.018)	0.267(0.129)
	adaLasso	0.108(0.074)	0.988(0.020)	0.154(0.161)
	SCAD	0.251(0.052)	0.974(0.011)	0.225(0.138)
	Grace	0.546(0.091)	0.989(0.010)	0.181(0.300)
	aGrace	0.561(0.089)	0.990(0.009)	0.181(0.300)
	$TTLP_I$	0.702(0.091)	0.981(0.020)	0.130(0.216)

	$LTLP_I$	0.720(0.107)	0.976(0.020)	0.089(0.107)
Scenario 7	PSIFORM	0.636(0.145)	0.976(0.052)	0.085(0.027)
	Lasso	0.108(0.074)	0.988(0.020)	0.154(0.161)
	adaLasso	0.251(0.052)	0.974(0.011)	0.235(0.138)
	SCAD	0.250(0.066)	0.967(0.018)	0.267(0.129)
	Grace	0.415(0.063)	0.988(0.006)	0.206(0.107)
	aGrace	0.410(0.063)	0.988(0.006)	0.206(0.185)
	$TTLP_I$	0.524(0.072)	0.980(0.022)	0.270(0.156)
	$LTLP_I$	0.545(0.099)	0.975(0.022)	0.213(0.235)

The first four scenarios correspond to the generation of data from model (3.1). Across the first four scenarios, *PSIFORM* has the largest TPRs, smallest FDRs, and comparable TNRs compared to the other methods. The simulation results demonstrate the consistently superior performance of *PSIFORM* compared to all the other methods in terms of biomarker detection accuracy when data are generated from a model, regardless of the independence or dependence of the genetic variable structure, different levels of dimensionality or sparsity, and various coefficient set-ups.

In scenarios 5-7, the performances of the different methods are compared when data are generated from a multinomial logistic regression model. Different coefficient set-ups and different dimensionalities of simulation data in the multinomial logistic regression model are explored. Although the differences between *PSIFORM* and the other methods are not as large as in the first four scenarios, the advantage of using the proposed method in terms of selecting true biomarkers is still clear: across the three scenarios, *PSIFORM* always has the smallest FDRs and generally has the largest TPRs. $TTLP_I$ and $LTLP_I$ outperform *Grace* and *aGrace* in terms of TPRs.

Across all scenarios, the network-based methods ($PSIFORM/LTLP_I/TTLP_I$/Grace /aGrace) obviously perform better than the penalized regression methods (SCAD/Lasso/adaLasso)

in terms of variable selection accuracy. When the true regression coefficients of neighboring genes are randomly generated, *PSIFORM*, $LTLP_I$, and $TTLP_I$ have higher TPRs and TNRs than *Grace* and *aGrace*, and *aGrace* outperforms *Grace* under almost all criteria, as expected. Moreover, $LTLP_I$ has overall slightly larger TPRs but lower TNRs compared to $TTLP_I$. In sparse models (scenarios 4 and 7), *PSIFORM* shows substantial advantages over the other methods in variable selection under all criteria. The performance of *PSIFORM* is the most robust when a sparsity structure exists among the variables.

In summary, *PSIFORM* has the best biomarker detection capability, especially when the neighboring true predictors do not share the same coefficients or when the model is sparse.

3.4 Kidney cancer application

In this section, I investigated the applicability of *PSIFORM* to the TCGA kidney renal clear cell carcinoma (KIRC) data set, which is available through the data portal hosted by the National Cancer Institute (http://cancergenome.nih.gov/). The objective of this study is to integrate genomic and proteomic data to discover predictive biomarkers that are associated with the 5-year survival rates of patients diagnosed with clear cell renal cell carcinoma (ccRCC), and incorporate pathway information to improve the accuracy of the results.

Renal cell carcinoma (RCC) is the most common and lethal type of kidney cancer. It is classified into four major histological cell types: clear cell (cc), papillary (10-15%), chromophobe (5%), and collecting duct (1%). Among them, ccRCC is the most common type of RCC, which accounts for 75-80% of total RCC cases [105].

Survival is affected by the stage and grade of cancer and the clinical characteristics of the patients. In advanced ccRCC, the 5-year survival rate is only approximately 5%-15%. Such tumors are usually resistant to chemotherapy and radiation therapy

[106]. Therefore, it would be very valuable to identify prognostic biomarkers and develop targeted treatments to supplement traditional therapies for ccRCC. Some biomarkers have been well recognized for ccRCC prognosis and pathogenesis, such as VHL, VEGFR, mTOR, HGF/c-MET, and Wnt/β-catenin signaling pathways. A number of targeted therapies have also been developed by blocking these critical signaling pathways [105]. The major ones include VEGF-targeted agents (sunitinib, sorafenib and bevacizumab) and mTOR inhibitors (everolimus and temsirolimus) [107, 108].

However, despite the advances in targeted therapies, these agents play limited roles in the eradication of RCC because most tumors develop acquired drug resistance (e.g., 30% of tumors show no response to sunitinib) [105]. There is a compelling need to understand the tumorigenesis and progression mechanisms more comprehensively and identify new prognostic biomarkers to develop novel treatments for RCC. Therefore, it is worthwhile to incorporate biological pathway information into the integrative analysis of mRNA and protein data to more accurately discover prognostic biomarkers for patients with RCC.

3.4.1 Data description

The TCGA KIRC data set includes 451 patients samples, from which 20,531 genes and 212 proteins were collected using the respective platforms of Illumina HiSeq2000 RNA sequencing and reverse-phase protein array (RPPA) technology. Among the 451 patients, 301 are censored and 150 are dead. I partitioned the patients into two groups on the basis of their survival times. According to an extreme discordant phenotype design [84], I took the top 20% (91 patients, surviving $>$ 1862 days) as long-term survivors (LTSs) and the bottom 45% (90 patients, surviving $<$ 885 days) as short-term survivors (STSs). By doing so, 181 patient samples remained for further analysis. The gene expression data were normalized using the RNA-

seq by expectation-maximization approach [62] and were logarithm-transformed prior to downstream analysis. The protein concentration data were also normalized by subtracting the median, both column-wise and row-wise [63].

I implemented the following steps to filter out trivial genes in the RNA-seq data. First, I removed genes with >40% missing data and used the k-nearest neighbor algorithm with k=10 to impute genes with <40% missing data by using the R package *impute* [104]. Second, I removed 25% of the genes that had extremely small coefficient of variation values, where the coefficient of variation is defined as the ratio of the standard deviation to the absolute value of the mean of the gene expression intensities. Then, I removed genes for which the difference between the top 80% quantile and the bottom 20% quantile was no larger than 0.75. Finally, I filtered out the genes with p-value >0.03 based on a univariate t-test. This procedure retained 3,361 genes for further analysis.

I downloaded pathway information from the KEGG database [27]. After mapping the Entrez Gene ID to the Pathway ID, among the 3,361 genes, we identified 294 genes that are associated with 32 KEGG pathways.

3.4.2 Results

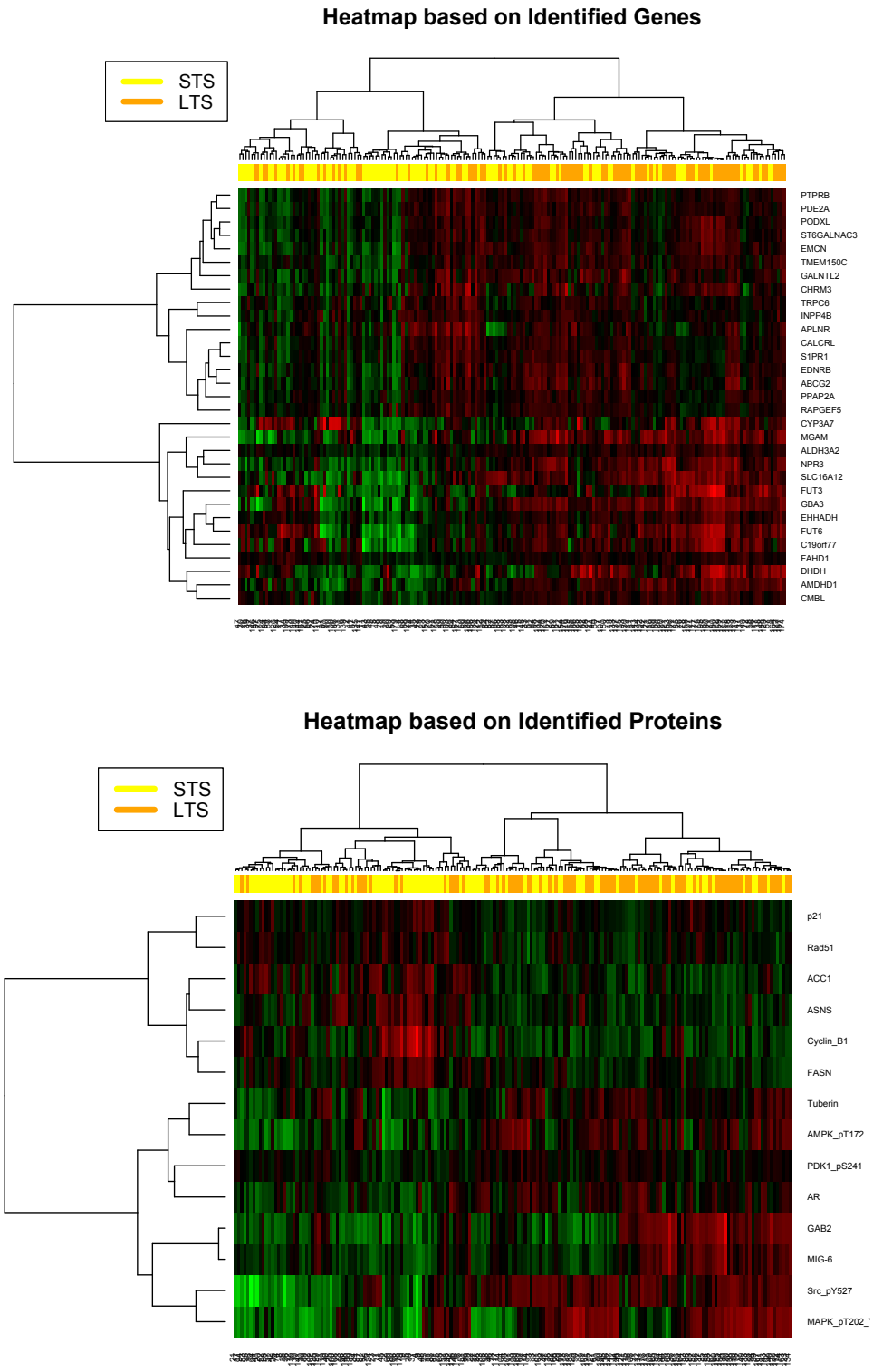
I implemented all the aforementioned network-based penalized regression methods for the integrative analysis of the genomic and proteomic data. To determine the tuning parameter values in the proposed method, I performed a grid search for tuning parameters λ_1, λ_2, and λ_3. Since pathway information for protein data is not available, $P_{\lambda_4}^{TLP}$ is not used and there is no need to tune λ_4. Specifically, the gird search was conducted over [4, 10] for λ_1, [0.05, 5] for λ_2, and [2, 6] for λ_3. I obtained $(\lambda_1, \lambda_2, \lambda_3)$ = (6, 2.9, 4.2) as the local optimal values.

PSIFORM identified 31 genes that had nonzero coefficients in $\boldsymbol{A}$. Among them, 15 genes belong to a metabolic pathway: ALDH3A2, DHDH, EHHADH, MGAM,

CYP3A7, PPAP2A, GBA3, CMBL, FUT6, INPP4B, FUT3, ST6GALNAC3, AMDHD1, FAHD1, GALNTL2; 5 genes are on the neuroactive ligand-receptor interaction pathway: CALCRL, CHRM3, S1PR1, APLNR, EDNRB, where EDNRB also belongs to the cGMP-PKG signaling pathway; 2 genes are on the cGMP-PKG signaling pathway: TRPC6, PDE2A; and 9 independent genes were identified: ABCG2, C19orf77, EMCN, NPR3, PODXL, PTPRB, RAPGEF5, SLC16A12, TMEM150C. Fourteen proteins were also identified by *PSIFORM*: ACC1, AMPK_pT172, AR, Cyclin_B1, GAB2, MAPK_pT202_Y204, MIG-6, PDK1_pS241, Rad51, Src_pY527, Tuberin, ASNS, FASN, and p21.

I also performed hierarchical clustering of the samples based on the Pearson correlation distance and applied Ward's linkage method to the selected genes and proteins. The clustering heatmaps based on the 31 identified genes and 15 proteins are displayed in Figure 3.1. In both panels, it can be observed that most "STS" (yellow) and "LTS" (orange) can be differentiated, as the different estimates of the two group characteristics in C=(-0.0747, 0.0739) reveal. It can also be seen from Figure 3.1 that the genes in the same pathway have similar gene expression patterns. For instance, in the first heatmap, most of the genes in the metabolic pathway, CYP3A7, MGAM, ALDH3A2, FUT3, GBA3, EHHADH, FUT6, FAHD1, DHDH, AMDHD1, and CMBL, cluster together at the bottom. These genes behave similarly as a functional group: they have higher expression levels in "LTS" but lower expression levels in "STS".

Figure 3.1.: Sample clustering based on genes and proteins selected by *PSIFORM*



ccRCC is fundamentally a metabolic disorder, and metabolic pathways have been found to be mostly deregulated in ccRCC by vast cancer profiling studies [109, 110, 111]. Many of the metabolic genes identified by *PSIFORM* are known to be correlated with ccRCC. For instance, CYP3A is confirmed to have an important role in the detoxification of chemotherapeutic agents and is a potential predictive biomarker for patients with metastatic renal cell carcinoma treated with sunitinib [114]; FUT3 and FUT6 genes are up-regulated in RCC [112]; and GALNTL2 is involved in the metastatic spread of ccRCC [113, 115]. Furthermore, it is also well established that the gene mutations in particular metabolic pathways correlate with the prognosis in ccRCC. Some studies have revealed that many genes in metabolic pathways display reduced expression in patients with more advanced ccRCC progression, which is consistent with our findings [111].

The neuroactive ligand-receptor interaction pathway and the cGMP-PKG signaling pathway are also found to be involved in the regulation of RCC [116, 117, 118]. Specifically, some of the genes identified in these two pathways are known to be correlated with the progression or prognosis of RCC. For example, the overexpression of S1PR1 is associated with increased survival times for RCC patients [119]. Similarly, the higher expression of EDNRB indicates a significantly longer survival time among patients with ccRCC, and has been identified as an independent prognostic marker for ccRCC [120].

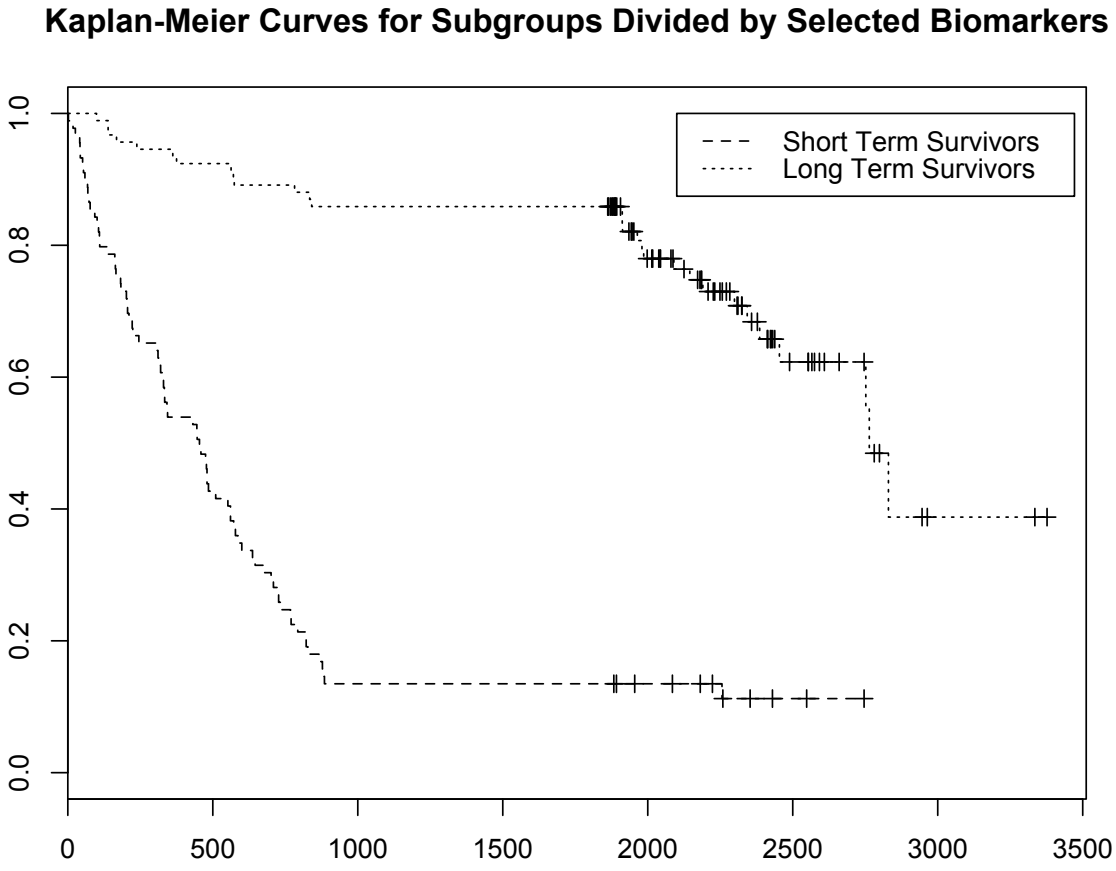
Some of the 9 independent genes are also identified as potential prognostic biomarkers for RCC [121, 122, 123]. For example, NPR3 has been found to be significantly associated with the survival times of RCC patients [121]. PODXL regulates cell-to-cell adhesion, and has been recognized as an independent predictor for survival times among patients with RCC in multiple studies [122, 123].

In terms of protein profiling, most of the *PSIFORM*-selected proteins are known critical potential markers for survival among patients with ccRCC. AMPK is a well-

established prognostic biomarker for survival among patients with RCC. The activation of AMPK inhibits the growth and survival of renal cell carcinoma cells by negatively regulating the mTOR signaling pathway. The down-regulation of AMPK is associated with poor response to cancer treatments and shorter survival times among patients with RCC [124, 125, 126]. The up-regulation of FASN, the fatty acid synthesis gene, is correlated with worse survival times among patients with RCC [125]. Similarly, the overexpression of cyclin B1 is associated with poor prognosis in patients with RCC [126, 127]. MAPK cascades are key signaling pathways involved in the regulation of cell proliferation, survival and differentiation, and the suppression of MAPK signaling pathways is known to inhibit RCC tumor growth [128]. The down-regulation of PDK1 may be associated with aggressive disease progression in RCC [129]. Finally, p21 is a valuable prognostic factor for ccRCC. In metastatic ccRCC, a higher level of p21 expression is associated with a shorter survival time [130]. All these scientific conclusions support our findings.

Kaplan-Meier curves are used to further investigate whether the 31 genes and 15 proteins selected by *PSIFORM* can differentiate patients with different survivals. I fitted a logistic regression model with discretized survival as the outcome and the 46 selected genetic variables as covariates. A survival score for each patient was calculated using the estimated coefficients. Based on the calculated scores, each patient was then reclassified into "long-term survivors" vs. "short-term survivors". Figure 3.2 shows the Kaplan-Meier curves of the two subgroups. It is observed that the selected biomarkers clearly differentiate "long-term survivors" from "short-term survivors" and thus can be potential prognostic biomarkers for patients diagnosed with ccRCC.

Figure 3.2.: Kaplan-Meier curves for subgroups divided by *PSIFORM*-selected biomarkers



The comparisons of the other four network-based penalized regression methods are summarized as follows.

- $LTLP_I$, $TTLP_I$, *Grace*, and *aGrace* identified 35, 31, 43, and 43 genes, respectively. $LTLP_I$ and $TTLP_I$ yielded very similar results; *Grace* and *aGrace* selected almost the same set of genes. Nine genes were identified by all these four methods. They are BST1, C21orf121, DHDH, GJD4, KALRN, NPR3, PID1, TMEM150C, and UBE2QL1. Among them, DHDH, NPR3, and TMEM150C

were also identified by *PSIFORM*. BST1, KALRN, DHDH are on metabolic pathways, but none of the remaining genes belong to any group with the same biological function. Out of the 9 genes, NPR3 and UBE2QL1 have been found to be associated with duration of survival among patients with RCC [131].

- In terms of proteomic profiling, $LTLP_I$, $TTLP_I$, *Grace*, and *aGrace* identified 11, 5, 10, and 10 proteins, respectively. Five proteins were identified by all four methods: GAB2, Ku80, PDK1_ps241, G6PD, and GYS. GAB2 and PDK1_ps241 were also identified by *PSIFORM*. PDK1 has been recognized as being associated with renal tumor progression. The elevated expression level of G6PD has also been observed in RCC patients [132].

None of the four methods satisfactorily captured the complex pathway structure from the high-dimensional multi-platform data. Furthermore, only a limited number of genes identified by the four methods belong to key pathways for tumorigenesis or prognosis of RCC. In contrast, *PSIFORM* identified many gene pathways and proteins that are known to be associated with the prognosis of patients with RCC, which supports the capability of *PSIFORM* to identify biomarkers that are potentially predictive of survival among patients with RCC.

Cross-validation: Finally, I compared the predictive performance of each method using leave-one-out cross-validation. The prescreened KIRC data set consists of 181 patients. I fitted the *PSIFORM* model based on mRNA and protein expression data from 180 patients, and then used the selected covariates to predict the discretized survival times for the remaining patient. This procedure was repeated 181 times to obtain the overall misclassification rate. To make a fair comparison, I focused on the predictive performance of the top 10 biomarkers with the largest absolute coefficient estimates. The misclassification rates of *PSIFORM*, $LTLP_I$, $TTLP_I$, *Grace*,and *aGrace* are respectively 0.127, 0.233, 0.242, 0.291, and 0.291. *PSIFORM* demonstrated the best prediction accuracy in this data set. It is noteworthy that the mis-

classification rates in this project are much lower than those in the *SIFORM* project. The incorporation of biological pathway information may contribute to the improved predictive performance.

3.5 Summary

In Chapter 2, I propose a generalized statistical framework *SIFORM* to jointly model multi-platform high-dimensional omic data and to discover the associations between these genetic variables and a disease-associated phenotype. In this chapter, I extend *SIFORM* to *PSIFORM*, which allows for incorporating pathway information in multi-platform data integration to improve the accuracy and interpretability of the results. The network structures of genes in biological pathways are characterized by a graphical model, and a network-based penalty is used to incorporate the graphic structure into the proposed statistical framework. Given biological pathway information, *PSIFORM* is shown to be a powerful tool in both biomarker detection and prediction through extensive simulations and a kidney cancer application. *PSIFORM* is able to detect the biomarkers that are associated with the disease phenotype of interest in different settings of dimensionality and coefficients.

PSIFORM also has the potential to accommodate a hierarchical order among genetic variables across different platforms by adding additional constraints P(A,B) to the coefficients of the variables from different platforms. For example, for any gene i and protein j encoded by gene i, I can add the *TLP* penalty function $P(A, B) = p(a_i, b_j) = |J_\tau(|a_i|) - J_\tau(|b_j|)|$ to ensure that the genes and their encoded proteins are selected into or eliminated from the model simultaneously. By modifying the penalized likelihood function, *PSIFORM* can easily incorporate the hierarchical order across the multi-platform *omic* measurements.

Currently, I use DC programming to convert a non-convex minimization problem into an iterative convex problem, and use the Matlab package CVX to solve the convex

problem. However, that package is not designed for high-dimensional data; hence, the implementation speed can be severely affected when the dimension of the data increases [103]. Therefore, a possible direction of future research is to develop more efficient numerical approaches to solve the optimization problem in a high-dimensional space.

Bibliography

[1] Brower,V. (2011)Epigenetics: Unravelling the cancer code. *Nature*, 471, S12-13.

[2] Chin,L., Andersen,J.N., Futreal,P. A. (2011) Cancer genomics: from discovery science to personalized medicine. *Nature Med.*17, 297-303.

[3] Zhang, S., Liu, CC., Li, W., Shen, H., Laird, PW., Zhou, XJ. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucl. Acids Res.*, 40(19), 9379-9391.

[4] Stingo,F., Chen,Y., Tadesse, M., Vannucci,M. (2011). Incorporating Biological Information into Linear Models: A Bayesian Approach to the Selection of Pathways and Genes. *The Annals of Applied Statistics*, 5: 1202-1214.

[5] Pan,W., Xie,B., Shen,X. (2010) Incorporating predictor network in penalized regression with application to microarray data, *Biometrics*, 66(2):474-84.

[6] Lingjarde,OC., Russnes,HG., Vollan,HM., Frigessi,A., Borresen-Dale AL. (2014) Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer*, 14, 299-313.

[7] Chari,R., Coe,BP., Vucic,EA., Lockwood,WW., Lam,WL. (2010) An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer, *BMC Syst. Biol.*, 4, 67. doi: 10.1186/1752-0509-4-67.

[8] Rhodes,D.R. and Chinnaiyan, A.M. (2005) Integrative analysis of the cancer transcriptome, *Nat. Genet.*, 37, S31-S37.

[9] Chin, K. DeVries,S., Fridlyand,J., Spellman,PT., Roydasgupta,R., Kuo,WL., Lapuk,A., Neve,RM., Qian,Z., Ryder,T., Chen,F., Feiler,H., Tokuyasu,T., Kingsley,C., Dairkee,S., Meng,Z., Chew,K., Pinkel,D., Jain,A., Ljung,BM., Esserman,L., Albertson,DG., Waldman,FM., Gray,JW. (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies, *Cancer Cell*, 10, 529-541.

[10] Segal,E., Friedman,N., Kaminski,N., Regev,A., Koller,D. (2005) From signatures to models: understanding cancer using microarrays,*Nature Genet.*, 37, S38-S45.

[11] Ovaska,K., Laakso,M., Haapa-Paananen,S., Louhimo,R., Chen,P., Aittomaki,V., Valo,E., Nunez-Fontarnau,J., Rantanen,V., Karinen,S., Nousiainen,K., Lahesmaa-Korpinen,AM., Miettinen,M., Saarinen,L., Kohonen,P., Wu,J., Westermarck,J., Hautaniemi,S. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme, *Genome Med.*, 2, 65.

[12] Bovelstad,H.M., Nygard,S., Storvold,HL., Aldrin,M., Borgan,O., Frigessi,A., Lingjarde,OC. (2007) Predicting survival from microarray data - a comparative study. *Bioinformatics*, 23: 2080-2087.

[13] Nowak,G., Hastie,T., Pollack,JR., Tibshirani,R. (2011) A fused lasso latent feature model for analyzing multi- sample aCGH data. *Biostatistics*,12, 776-791.

[14] Imoto,S., Higuchi,T., Goto,T., Tashiro,K., Kuhara,S., Miyano, S. (2004) Combining microarrays and biological knowledge for estimating gene networks via bayesian networks.*J. Bioinform. Comput. Biol.*, 2, 77-98.

[15] Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559. doi: 10.1186/1471-2105-9-559.

[16] Ha,MJ., Baladandayuthapani,V., Do,KA. (2015) DINGO: differential network analysis in genomics.*Bioinformatics*, Jul 6.

[17] Ni, Y., Stingo,FC., Baladandayuthapani,V. (2014) Integrative Bayesian network analysis of genomic data, *Cancer Inform.*, 13(Suppl 2), 39 -48.

[18] Chekouo,T., Stingo,F.C. and Doecke,J.D. (2015) miRNA-target gene regulatory networks: A Bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics*, 71(2), 428-438.

[19] Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, 58(1), 267-288.

[20] Rockova,V., Lesaffre,E. (2014) Incorporating Grouping Information in Bayesian Variable Selection with Applications in Genomics, *Bayesian Analysis*, 1: 221-258

[21] Li, C., Li H. (2008) Network-constrained regularization and variable selection for analysis of genomic data, *Bioinformatics*, 24(9), 1175-82.

[22] Li, C., Li H. (2010) Variable Selection and Regression Analysis for Graph-Structured Covariates with an Application to Genomics, *Ann Appl Stat.*,4(3):1498-1516.

[23] Kim,S., Pan,W., Shen, X. (2013) Network-Based Penalized Regression With Application to Genomic Data, *Biometrics*, 69:582-593.

[24] Li,F., Zhang,N. (2010) Bayesian Variable selection in structured high-dimensional covariate space with application in genomics, *Journal of the American Statistical Association*, 105:1202-1214.

[25] Stingo,F., Vannucci, M. (2011). Variable Selection for Discriminant Analysis with Markov Random Field Priors for the Analysis of Microarray Data, *Bioinformatics*, 27(4): 495-501.

[26] Alfarano,C., Andrade,CE., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E., Buzadzija,K., Cavero,R., D'Abreo,C., Donaldson,I., Dorairajoo,D., Dumontier,MJ., Dumontier,MR., Earles,V., Farrall,R., Feldman,H., Garderman,E., Gong,Y., Gonzaga,R., Grytsan,V., Gryz,E., Gu,V., Haldorsen,E., Halupa,A., Haw,R., Hrvojic,A., Hurrell,L., Isserlin,R., Jack,F., Juma,F., Khan,A., Kon,T., Konopinsky,S., Le,V., Lee,E., Ling,S., Magidin,M., Moniakis,J., Montojo,J., Moore,S., Muskat,B., Ng,I., Paraiso,JP., Parker,B., Pintilie,G., Pirone,R., Salama,JJ., Sgro,S., Shan,T., Shu,Y., Siew,J., Skinner,D., Snyder,K., Stasiuk,R., Strumpf,D., Tuekam,B., Tao,S., Wang,Z., White,M., Willis,R., Wolting,C., Wong,S., Wrong,A., Xin,C., Yao,R., Yates,B., Zhang,S., Zheng, K., Pawson,T., Ouellette,BFF., Hogue,CWV. (2005) The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic Acids Res.*, January 1 (33)

[27] Kanehisa,M., Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res*, Jan 1;28(1):27-30.

[28] Zou,H. (2006) The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, 101(476), 1418-1429.

[29] Savin, I., Winker, P. (2013), Lasso-type and heuristic strategies in model selection and forecasting, *Journal of Economics and Statistics*, Vol. 233, No. 4, pp. 526-549

[30] Hoerl,AE., Kennard,RW. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, Vol. 12, No. 1, pp. 55-67

[31] Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, 67(2), 301–320.

[32] Fan,J. and Li,R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456), 1348-1360.

[33] Fan,J. and Peng,H. (2004) Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.*, 32(3), 928-961.

[34] Hotelling, H. (1936) Relations between two sets of variates.*Biometrika*. Vol. 28, No. 3/4, 321-377

[35] Witten,DM. and Tibshirani,RJ. (2009) Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology*, Volume 8, Issue 1, Article 28

[36] Witten,DM, Tibshirani, R., Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3): 515-534.

[37] Thompson, B. (1991) A Primer on the Logic and Use of Canonical Correlation Analysis. *Measurement and Evaluation in Counseling and Development*, 24(2): p80-93

[38] Mandal, A. and Cichocki,A. (2013) Non-Linear Canonical Correlation Analysis Using Alpha-Beta Divergence. *Entropy*, 15, 2788-2804

[39] Gross, SM and Tibshirani,RJ. (2014) Collaborative Regression, *Biostatistics*,16(2).

[40] Fan,J., Lv,J. (2010) A selective overview of variable selection in high dimensional feature space, *Stat. Sin.*, 20(1), 101-148.

[41] Hastie,T., Tibshirani,R., Friedman,J. (2001) The Elements of Statistical Learning, *Springer New York Inc.* , New York, NY, USA . p61-79

[42] Osborne, M., Presnell, B. and Turlach, B. (2000) A new approach to variable selection in least squares problems, *IMA Journal of Numerical Analysis*, 20, 389-403.

[43] Efron,B., Hastie,T., Johnstone,I., Tibshirani,R. (2004) Least Angle Regression, *The Annals of Statistics*, Vol. 32, No. 2, 407- 499

[44] Xie,J., Zeng,L. (2010) Group Variable Selection Methods and Their Applications in Analysis of Genomic Data, *Frontiers in Computational and Systems Biology*,Volume 15 of the series Computational Biology, pp 231-248

[45] Zou,H., Li,R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.*, 36(4), 1509-1533.

[46] Mankoo, PK., Shen, R., Schultz, N., Levine,DA., Sander,C.(2011) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles, *PLoS ONE*, Volume 6, Issue 11

[47] Friedman,J., Hastie,T., Hoefling,H., Tibshirani,R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.*, 1, 302-332.

[48] Zhao,P., Yu,B. (2004). Boosted Lasso. Technical Report, Department of Statistics, UC-Berkeley.

[49] Shen,X., Pan,W., and Zhu,Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107, 223-232.

[50] Horst,R., Thoai,N.V. (1999) Dc programming: An overview, *Journal of Optimization Theory and Application*, 93(1):1-43

[51] Hunter,D.R., Li,R. (2005) Variable selection using MM algorithm. *Ann. Stat.*, 33(4), 1617-1642.

[52] Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, 6(2), 461-464.

[53] Fan,Y.(2013) Tuning parameter selection in high dimensional penalized likelihood, *J. R. Stat. Soc. B*, 75(3), 531-552.

[54] Zou,H., Hastie,T., Tibshirani,R. (2007) On the "degrees of freedom" of the lasso. *Ann. Stat.*, 35(5), 2173-2192.

[55] Sanchez-Cespedes,M.S., Ahrendt,SA., Piantadosi,S., Rosell,R., Monzo,M., Wu,L., Westra,WH., Yang,SC., Jen,J., Sidransky,D. (2001) Chromosomal alterations in lung adenocarcinoma from smokers and nonsmokers. *Cancer Res.*, 61, 1309-1313.

[56] Sun,S., Schiller,J.H., Gazdar,A.F. (2007) Lung cancer in never smokers - a different disease. *Nat. Rev. Cancer*, 7, 778-790.

[57] Samet,J., Avila-Tang,E., Boffetta,P., Hannan,LM., Olivo-Marston,S., Thun,MJ., Rudin,CM.(2009) Lung cancer in never smokers: clinical epidemiology and environmental risk factors. *Clin. Cancer Res.*, 15(18), 5626-5645.

[58] Yano,T., Miura,N., Takenaka,T., Haro,A., Okazaki,H., Ohba,T., Kouso,H., Kometani,T., Shoji,F., Maehara,Y. (2008) Never-smoking nonsmall cell lung cancer as a separate entity — the clinico-pathologic features and survival. *Cancer*, 113, 1012-1018.

[59] Yano,T., Haro,A., Shikada,Y., Maruyama,R., Maehara,Y.(2011) Non-small cell lung cancer in never smokers as a representative "non-smoking-associated lung cancer": epidemiology and clinical features. *Int. J. Clin. Oncol.*, 16(4), 287-293.

[60] Ding,L. Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216), 1069-1075.

[61] Serke M. (2007) Lung cancer: targeted therapy, *Pneumologie*, 61, 162-170.

[62] Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.

[63] Li,J., Lu,Y., Akbani,R., Ju,Z., Roebuck,PL., Liu,W., Yang,YJ., Broom,BM., Verhaak,RGW., Kane,DW., Wakefield,C., Weinstein,JN., Mills,GB., Liang,H. (2013) TCPA: a resource for cancer functional proteomics data. *Nat. Methods*, 10, 1046-1047.

[64] Hastie,T., Tibshirani,R. and Eisen,M.B. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, 1(2), research0003.1-0003.21.

[65] Collier, A.C., Tingle,MD., Paxton,JW., Mitchell,MD., Keelan,JA. (2002) Metabolizing enzyme localization and activities in the first trimester human placenta: The effect of maternal and gestational age, smoking and alcohol consumption. *Hum. Reprod.*, 17, 2564-2572.

[66] Bock,K.W., Schrenk,D., Forster,A., Griese,EU., Morike,K., Brockmeier,D., Eichelbaum,M.(1994) The influence of environmental and genetic factors on CYP2D6, CYP1A2 and UDP-glucuronosyltransferases in man using sparteine, caffeine, and paracetamol as probes. *Pharmacogenetics*, 4: 209-218.

[67] Cerami,E.G., Gross,BE., Demir,E., Rodchenkov,I., Babur,O., Anwar,N., Schultz,N., Bader,GD., Sander,C. (2010) Pathway Commons, a web resource for biological pathway data. *Nucl. Acids Res.*, 30, D685-D690.

[68] Lovly,C., Horn,L. and Pao,W. (2014) EGFR mutations in non-small cell lung cancer (NSCLC). *My Cancer Genome*, http://www.mycancergenome.org/content/disease/lung-cancer/egfr/ (Updated June 18).

[69] Sarris,E.G., Saif,MW., Syrigos,KN. (2012) The biological role of PI3K pathway in lung cancer. *Pharmaceuticals (Basel)*, 5 (2012), 1236-1264.

[70] Brand,T.M., Iida M, Li C, Wheeler DL. (2011) The nuclear epidermal growth factor receptor signaling network and its role in cancer. *Discov. Med.*,12(66), 419-432.

[71] Ma,J., Fan,K., Zhang,Y., Song,D., Ma,J.(2008) Clinicopathological significance of E-cadherin and PCNA expression in human non-small cell lung cancer. *Chin. J. Clin. Oncol.*, 5, 87-92.

[72] Fumarola,C., Bonelli,MA., Petronini,PG., Alfieri,RR. (2014) Targeting PI3K/AKT/mTOR pathway in non small cell lung cancer, *Biochem. Pharmacol.*, 3(90), 197-207.

[73] Grabauskiene,S., Bergeron,EJ., Chen,G., Thomas,DG., Giordano,TJ., Beer,DG., Morgan,MA., Reddy,RM. (2014) Checkpoint kinase 1 protein expression indicates sensitization to therapy by checkpoint kinase 1 inhibition in non-small cell lung cancer. *J. Surg. Res.*, 187(1), 6-13.

[74] Gingras,A.C., Kennedy,SG., O'Leary,MA., Sonenberg,N., Hay,N. (1998) 4E-BP1, a repressor of mRNA translation, is phosphorylated and inactivated by the Akt(PKB) signaling pathway. *Genes & Dev.*, 12, 502-513.

[75] Sekia,N., Takasu T, Sawada S, Nakata M, Nishimura R, Segawa Y, Shibakuki R, Hanafusa T, Eguchi K. (2010) Prognostic significance of expression of eukaryotic initiation factor 4E and 4E binding protein 1 in patients with pathological stage I invasive lung adenocarcinoma. *Lung Cancer*, 70(3), 329-334.

[76] Dumstorf,C.A., Konicek BW, McNulty AM, Parsons SH, Furic L, Sonenberg N, Graff JR.(2010) Modulation of 4E-BP1 function as a critical determinant of enzastaurin-induced apoptosis. *Mol. Cancer Ther.*, 9, 3158-3163.

[77] Giunti,S., Antonelli,A., Amorosi,A., Santarpia,L. (2013) Cellular signaling pathway alterations and potential targeted therapies for medullary thyroid carcinoma. *Int. J. Endocrinol.*, 2013, 803171. doi: 10.1155/2013/803171.

[78] Nitta,R.T., Del Vecchio C.A., Chu A.H., Mitra S.S., Godwin A.K., Wong A.J. (2011) The role of the c-Jun N-terminal kinase 2-alpha-isoform in non-small cell lung carcinoma tumorigenesis. *Oncogene*, 30(2), 234-244.

[79] Cooper,W.A., Lam DC, O'Toole SA, Minna JD.(2013) Molecular biology of lung cancer, *J. Thorac. Dis.*, 5(Suppl 5), S479–S490.

[80] Ren,J.H., He WS, Yan GL, Jin M, Yang KY, Wu G. (2012) EGFR mutations in non-small-cell lung cancer among smokers and non-smokers: a meta-analysis. *Environ. Mol. Mutagen.*, 53(1), 78-82.

[81] Chen,P., Cescon,M. and Bonaldo,P. (2013) Collagen VI in cancer and its biological mechanisms. *Trends Mol. Med.*, 19(7), 410-417.

[82] Voiles,L. Lewis DE, Han L, Lupov IP, Lin TL, Robertson MJ, Petrache I, Chang HC. (2014) Overexpression of type VI collagen in neoplastic lung tissues. *Oncol. Rep.*, 32(5), 1897-1904.

[83] Kuykendall,A., Chiappori,A. (2014) Advanced EGFR Mutation-Positive Non-Small-Cell Lung Cancer: Case Report, Literature Review, and Treatment Recommendations. *Cancer Control*, 21(1): p.67-73

[84] Nebert,D.W. (2000) Extreme discordant phenotype methodology: an intui- tive approach to clinical pharmacogenetics, *Eur. J. Pharmacol.*, 410,107-120.

[85] Kawai, H. Ishii, A., Washiya, K., Konno, T., Kon, H., Yamaya, C., Ono, I., Minamiya, Y., Ogawa, J.(2005) Estrogen receptor alpha and beta are prognostic factors in non-small cell lung cancer. *Clin Cancer Res.*, 11(14): 5084-9

[86] Olivo-Marston,S., Mechanic, LE., Mollerup, S., Bowman, ED., Remaley, AT., Forman, MR., Skaug, V., Zheng, YL., Haugen, A., Harris, CC. (2010) Serum

estrogen and tumor-positive estrogen receptor-alpha are strong prognostic classifiers of non-small-cell lung cancer survival in both men and women. *Carcinogenesis*, 31(10),1778-86.

[87] Lee,HW. Lee, EH., Lee, JH., Kim, JE., Kim, SH., Kim, TG., Hwang, SW., Kang, KW.(2015) Prognostic significance of phosphorylated 4E-binding protein 1 in non-small cell lung cancer. *Int J Clin Exp Pathol.* , 8(4): 3955-3962.

[88] Lv, T. Wang, Q., Cromie, M., Liu, H., Tang, S., Song, Y., Gao, W. (2015) Twist1-mediated 4E-BP1 regulation through mTOR in non-small cell lung cancer. *Oncotarget.*, 6(32): 33006-18.

[89] Cooper, WA. Kohonen-Corish, MR., McCaughan, B., Kenned, C., Sutherland, RL., Lee, CS.(2009) Expression and prognostic significance of cyclin B1 and cyclin A in non-small cell lung cancer. *Histopathology*, 55(1): 28-36

[90] Yoshida, T., Tanaka, S., Mogi, A., Shitara, Y., Kuwano, H. (2004) The clinical significance of Cyclin B1 and Wee1 expression in non-small-cell lung cancer. *Ann Oncol.*, 15(2): 252-6

[91] Arinaga,M., Noguchi, T., Takeno, S., Chujo, M., Miura, T., Kimura, Y., Uchida, Y. (2003) Clinical implication of cyclin B1 in non-small cell lung cancer. *Oncol Rep.*, 10(5): 1381-6

[92] Ma,Y., Fan, M., Dai, L., Kang, X., Liu, Y., Sun, Y., Yan, W., Liang, Z., Xiong, H., Chen, K. (2015) The expression of TTF-1 and Napsin A in early-stage lung adenocarcinoma correlates with the results of surgical treatment. *Tumour Biol.*, 36(10): 8085-92

[93] Pankov,R., Yamada,KM. (2002) Fibronectin at a glance. *Journal of Cell Science*, 115: 3861-3863

[94] Han, S., Khuri, FR., Roman, J.(2006) Fibronectin stimulates non-small cell lung carcinoma cell growth through activation of Akt/mammalian target of rapamycin/S6 kinase and inactivation of LKB1/AMP-activated protein kinase signal pathways. *Cancer Res.* , 66(1): 315-23

[95] Di Bernardo,MC. Matakidou, A., Eisen, T., Houlston, RS.(2009) Plasminogen activator inhibitor variants PAI-1 A15T and PAI-2 S413C influence lung cancer prognosis. *Lung Cancer.*, 65(2): 237-41

[96] Robert, C., Bolon, I., Gazzeri, S., Veyrenc, S., Brambilla, C., Brambilla, E. (1999) Expression of Plasminogen Activator Inhibitors 1 and 2 in Lung Cancer and Their Role in Tumor Progression. *Clin Cancer Res.*, 5: 2094

[97] Potscher, BM. and Schneider, U. (2009) On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference*,139: 2775-2790.

[98] Potscher, BM. and Schneider, U. (2010) Confidence Sets Based on Penalized Maximum Likelihood Estimators in Gaussian Regression. *Electronic Journal of Statistics*, 4:334-360.

[99] Lockhart,R., Taylor,J., Tibshirani, RJ., Tibshirani,R. (2014) A SIGNIFICANCE TEST FOR THE LASSO. *Ann. Statist.*, Volume 42, Number 2, 413-468.

[100] Minnier,J., Tian,L, Cai,T. (2011) A Perturbation Method for Inference on Regularized Regression Estimates. *J Am Stat Assoc.*, 106(496): 1371-1382

[101] Kosorok, M. (2008) Introduction to empirical processes and semiparametric inference. *New York: Springer Verlag.*

[102] Newey, W, McFadden, D. (1994) Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4: 2111-2245.

[103] Grant, M. and Boyd, S. (2015). CVX: Matlab software for disciplined convex programming, version 2.1 http://cvxr. com/cvx

[104] Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein, D., Altman, RB. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, Vol. 17 no. 6:520-525

[105] Banumathy,G., Cairns, P. (2010) Signaling pathways in renal cell carcinoma. *Cancer Biol Ther.* ,Oct 1;10(7): 658- 664

[106] Li,C., Shu,F., Lei,B., Lv,D., Zhang,S., Mao, X. (2015) Expression of PGAM1 in renal clear cell carcinoma and its clinical significance. *Int J Clin Exp Pathol.*8(8): 9410-9415.

[107] Powles,T., Chowdhury,S., Jones,R., Mantle,M., Nathan,P., Bex,A., Lim, L., Hutson,T. (2011) Sunitinib and other targeted therapies for renal cell carcinoma. *British Journal of Cancer*, 104, 741-745.

[108] Fishman,MN. (2013) Targeted Therapy of Kidney Cancer: Keeping the Art Around the Algorithms. *Cancer Control.* Jul;20(3):222-32

[109] Zaravinos,A., Pieri, M., Mourmouras, N., Anastasiadou, N., Zouvani, I., Delakas, D., Deltas, C. (2014) Altered metabolic pathways in clear cell renal cell carcinoma: A meta-analysis and validation study focused on the deregulated genes and their associated networks. *Oncoscience.*1(2): 117-131.

[110] White, NM., Newsted, DW., Masui, O., Romaschin, AD., Siu, KW., Yousef, GM. (2014) Identification and validation of dysregulated metabolic pathways in metastatic renal cell carcinoma. *Tumour Biol.* Mar; 35(3):1833-46.

[111] Hakimi,A., Reznik,E., Lee,C., Creighton,C., Brannon,AR., Luna,A., Aksoy,BA., Liu,E., Shen,R., Lee,W., Chen,Y., Stirdivant,SM., Russo,P.,

Chen,YB., Tickoo,S., Reuter,V., Cheng,E., Sander,C., Hsieh,J. (2016) An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma. *Cancee Cell* Volume 29, Issue 1, Pages 104-116

[112] Liou, LS., Shi, T., Duan, ZH., Sadhukhan, P., Der, SD., Novick, AA., Hissong, J., Skacel, M., Almasan, A., DiDonato, JA. (2004) Microarray gene expression profiling and analysis in renal cell carcinoma. *BMC Urol.*, Jun 22;4:9.

[113] Khella,HW., White,NM., Faragalla,H., Gabril,M., Boazak,M., Dorian,D., Khalil,B., Antonios,H., Bao,T., Pasic,MD., Honey,RJ., Stewart,R., Pace,KT., Bjarnason,GA., Jewett,MA., Yousef,GM. (2012) Exploring the role of miRNAs in renal cell carcinoma progression and metastasis through bioinformatic and experimental analyses. *Tumor Biol.*, 33:131-140

[114] Diekstra,MH., Swen,JJ., Boven,E., Castellano,D., Gelderblom,H., Mathijssen,RH., Rodrguez-Antona,C., Garca-Donas,J., Rini,BI., Guchelaar,HJ. (2015) CYP3A5 and ABCB1 Polymorphisms as Predictors for Sunitinib Outcome in Metastatic Renal Cell Carcinoma. *European Urology*, Volume 68 Issue 4, Pages 621-629

[115] Wuttig,D., Baier,B., Fuessel,S., Meinhardt,M., Herr,A., Hoefling,C., Toma,M., Grimm,MO., Meye,A., Rolle,A., Wirth,MP. (2009) Gene signatures of pulmonary metastases of renal cell carcinoma reflect the disease-free interval and the number of metastases per patient. *Int. J. Cancer*: 125, 474-482

[116] Liu,X., Wang,J., Sun,G. (2015) Identification of Key Genes and Pathways in Renal Cell Carcinoma Through Expression Profiling Data. *Kidney Blood Press Res*; 40: 288-297

[117] Yang,W., Yoshigoe,K., Qin,X., Liu,J., Yang,J., Niemierko,A., Deng,Y., Liu,Y., Dunker,A., Chen,Z., Wang,L., Xu,D., Arabnia,H., Tong,W., Yang,M. (2014)

Identification of genes and pathways involved in kidney renal clear cell carcinoma. *BMC Bioinformatics*, 15(Suppl 17): S2

[118] Ren,Y., Zheng,J., Yao,X., Weng,G., Wu,L. (2014) Essential role of the cGMP/PKG signaling pathway in regulating the proliferation and survival of human renal carcinoma cells. *Int J Mol Med.* Nov;34(5):1430-8.

[119] Wozniak,MB., Le Calvez-Kelm,F., Abedi-Ardekani,B., Byrnes,G., Durand,G., Carreira,C., Michelon,J., Janout,V., Holcatova,I., Foretova,L., Brisuda,A., Lesueur,F., McKay,J., Brennan,P., Scelo,G. (2013) Integrative Genome-Wide Gene Expression Profiling of Clear Cell Renal Cell Carcinoma in Czech Republic and in the United States. *PLoS ONE 8(3)*: e57886.

[120] Wuttig,D., Zastrow,S., Fussel,S., Toma,MI., Meinhardt,M., Kalman,K., Junker,K., Sanjmyatav,J., Boll,K., Hackermuller,J., Rolle,A., Grimm,MO., Wirth,MP. (2012) CD31, EDNRB and TSPAN7 are promising prognostic markers in clear-cell renal cell carcinoma revealed by genome-wide expression analyses of primary tumors and metastases.*Int J Cancer.* Sep 1;131(5): E693-704.

[121] Liu,Q., Zhao,S., Su,P., Yu,S. (2013) Gene and isoform expression signatures associated with tumor stage in kidney renal clear cell carcinoma, *BMC Systems Biology*, 7(Suppl 5): S7

[122] Saukkonen,K., Hagstrm,J., Mustonen,H., Juuti,A., Nordling,S., Fermr,C., Nilsson,O., Seppnen,H., Haglund,C. (2015) Podocalyxin Is a Marker of Poor Prognosis in Pancreatic Ductal Adenocarcinoma. *PLoS ONE*, 10(6): e0129012.

[123] Hsu,YH., Lin,WL., Hou,YT., Pu,YS., Shun,CT., Chen,CL., Wu,YY., Chen,JY., Chen,TH., Jou,TS. (2010) Podocalyxin EBP50 ezrin molecular complex enhances the metastatic potential of renal cell carcinoma through recruiting Rac1 guanine nucleotide exchange factor ARHGEF7.*Am J Pathol*, 176: 3050-3061

[124] Woodard,J., Joshi,S., Viollet,B., Hay,N., Platanias,LC., (2010) AMPK as a therapeutic target in renal cell carcinoma. *Cancer Biol Ther.*, Dec 1;10(11):1168-77.

[125] The Cancer Genome Atlas Research Network (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499: 43-49

[126] Tsavachidou-Fenner,D., Tannir,N., Tamboli,P., Liu,W., Petillo,D., Teh,B., Mills,GB., Jonasch,E. (2010) Gene and protein expression markers of response to combined antiangiogenic and epidermal growth factor targeted therapy in renal cell carcinoma. *Ann Oncol*, 21: 1599-1606.

[127] Ikuerowo SO, Kuczyk MA, Mengel M, van der Heyde E, Shittu OB, Vaske B, Jonas U, Machtens S, Serth J (2006) Alteration of subcellular and cellular expression patterns of cyclin B1 in renal cell carcinoma is significantly related to clinical progression and survival of patients. *Int J Cancer*, 119: 867-874

[128] Huang,D., Ding,Y., Luo,WM., Bender,S., Qian,CN., Kort,E., Zhang,ZF., VandenBeldt,K., Duesbery,NS., Resau,JH., Teh, BT. (2008) Inhibition of MAPK kinase signaling pathways suppressed renal cell carcinoma growth and angiogenesis in vivo. *Cancer Res.*, Jan 1;68(1): 81-8.

[129] Baumunk,D., Reichelt,U., Hildebrandt,J., Krause,H., Ebbing,J., Cash,H., Miller,K., Schostak,M., Weikert,S. (2013) Expression parameters of the metabolic pathway genes pyruvate dehydrogenase kinase-1 (PDK-1) and DJ-1/PARK7 in renal cell carcinoma (RCC). *World J Urol.*, Oct;31(5): 1191-6.

[130] Weiss,RH., Borowsky,AD., Seligson,D., Lin,PY., Dillard-Telm,L., Belldegrun,AS., Figlin,RA., Pantuck,AD. (2007) p21 is a prognostic marker for renal cell carcinoma: implications for novel therapeutic approaches. *J Urol.*, Jan; 177(1): 63-8.

[131] Wake,NC., Ricketts,CJ., Morris,MR., Prigmore,E., Gribble,SM., Skytte,AB., Brown,M., Clarke,N., Banks,RE., Hodgson,S., Turnell,AS., Maher,ER., Woodward,ER. (2013) UBE2QL1 is Disrupted by a Constitutional Translocation Associated with Renal Tumor Predisposition and is a Novel Candidate Renal Tumor Suppressor Gene.*Hum Mutat.*, 34(12):1650-61.

[132] Langbein,S., Frederiks,WM., zur Hausen,A., Popa,J., Lehmann,J., Weiss,C., Alken,P., Coy,JF. (2008) Metastasis is promoted by a bioenergetic switch: new targets for progressive renal cell cancer. *Int J Cancer.* , 122(11): 2422-2428.

VITA

Xuebei An was born in Shanghai, China in December, 1986. She received the degree of Bachelor of Engineering with a major in Computer Science from Shanghai University in July, 2009. After that, she obtained Master of Science degree with a major in Biostatistics from Case Western Reserve University in 2012. In August of 2012, she entered the The University of Texas Graduate School of Biomedical Sciences at Houston.