ABSTRACT

ZHU, HONGJIE. Pharmacometabolomics Data Analysis and Nonlinear Sufficient Dimension Reduction for Genome-Scale Studies. (Under the direction of Drs. Zhaobang Zeng and Lexin Li.)

This thesis is devoted to the large area of omics data analysis. Its contribution has two aspects: a practical guideline for pharmacometabolomics data analysis and a methodological development of a framework of nonlinear sufficient dimension reduction methods for biological group selection.

Pharmacometabolomics is a new omics discipline that aims to understand the mechanisms of action of drugs and variations in the drug response phenotype. We propose an integrated pipeline for in-depth mining of pharmacometabolomics data. Not only do we cover a variety of routine topics, such as data cleaning, detection of drug exposure signatures, predictive biomarkers, drug response signatures, and validation of results, but also we attempt to probe into the hidden mechanism of drug response variation through deciphering the relationships among biomarkers and pathway analysis. Also discussed are means to connect various types of omics studies for the same cohort of subjects. We then demonstrate two pharmacometabolomics studies: one reveals a role for gut-derived factors in modulating statin efficacy; the other maps a pathway that is implicated in mechanism of variation in response to an antidepressant selective serotonin reuptake inhibitor (SSRI), which then help define a gene implicated in response to SSRI.

To meet the high-dimensionality challenge from omics studies while considering the widely existing regulatory and other complex interactive relationships among units in biological systems, we propose a framework of kernel-based nonlinear supervised dimension reduction methods. It aims at capturing nonlinear effect of a group of variables on

a phenotype of interest using a small number of summary variables. Its computational advantage can handle a very large number of input variables, even if it far exceeds the number of observations. We then present two applications of the proposed approach to omics data analysis. The first application is to summarize biological pathways for the purpose of identifying those related to a clinical phenotype of interest. Simulations demonstrates the theoretical advantages of our approaches over existing solutions to pathway selection. An analysis of a glioblastoma microarray data using our approach finds four pathways related to patients' time to death that have evidence of support from the biological literature. The second application is to aggregate SNP information within a SNP-set, e.g., a gene or pathway, for genetic association testing of complex trait. Since kernel functions play a critical role in the nonlinear sufficient dimension reduction approach while commonly used kernels for continuous variables may not perform well for discrete genetic data, we propose a class of new kernels specifically designed for genotype data based on Markov chain theory. Our new summary statistic based on nonlinear dimension reduction coupled with the new kernels shows superior performance over existing methods over various disease models simulated from the sequence data of real genes.

Finally, the thesis concludes with a discussion of future research work.

Pharmacometabolomics Data Analysis and Nonlinear Sufficient Dimension Reduction for Genome-Scale Studies

by
Hongjie Zhu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics and Statistics

Raleigh, North Carolina

2011

APPROVED BY:

Rima Kaddurah-Daouk	David Bird
Hua Zhou	Lexin Li Co-chair of Advisory Committee

Zhaobang Zeng
Co-chair of Advisory Committee

DEDICATION

To my beloved parents, grandparents, and Man Luo.

BIOGRAPHY

Hongjie Zhu was born in Shandong province, China. He received his Bachelor of Science in Bioinformatics from Zhejiang University (Hangzhou, Zhejiang, China) in 2007. During his undergraduate course, he also received training in mathematics and computer science from Chu Kechen Honors College of Zhejiang University. He entered the bioinformatics Ph.D. program in 2007 at North Carolina State University (NCSU) and pursued a Ph.D. co-major in bioinformatics and statistics. Under the direction of Drs. Zhaobang Zeng and Lexin Li, he focused on computational systems biology during his study at NCSU, which includes omics data analysis and integration, network modeling, and related statistical methodologies, such as model selection and dimension reduction. He is also a member of Pharmacometabolomics Research Network, and involved in several studies on interpersonal variations in therapeutic response.

ACKNOWLEDGEMENTS

I am deeply indebted to Dr. Zhaobang Zeng for his scientific guidance, his faith and generous support at the critical moments of my life. I feel amazingly fortunate to have such a good advisor who allows me to explore different scientific projects and related fields, and at the same time make great efforts to keep me on the right track. My utmost gratitude also goes to Dr. Lexin Li, who opens the doors of many research areas for me, and provides very helpful suggestions for the problems I encountered during my research with great patience. I will never forget the enormous sincerity, encouragement and direction from Dr. Rima Kaddurah-Daouk. Her incredible enthusiasm to science has been and will always be great inspiration for me to push forward in my own research. Dr. Hua Zhou is one of the smartest people I have ever seen in my life. I am enlightened from every discussion we have had. I am also deeply impressed by his generosity in sharing his valuable research ideas. Dr. David Bird tries to help his students in every possible way. Not only do I clearly remember his illumination in our scientific discussions, but also his warm-hearted explanations of American customs to me.

I would also like to express my sincere appreciation to Drs. Richard M. Weinshilboum, Wayne Matson, Yuan Ji, Stephen Boyle, and all my collaborators from the Pharmacometabolomics Research Network, for their high quality data, thoughtful insight into biomedical science, and priceless contributions to our projects.

I am very grateful for Drs. Eric Stone, Steffen Heber, Alison Motsinger, and all the other faculties, post-docs and my fellow graduate students in the Bioinformatics Research Center (BRC), for all the warm help and hearty scientific discussions they have kindly offered. I sincerely appreciate staffs at BRC for their energetic and zealous assistance to me in all aspects during my study course.

Lastly, I cannot acknowledge enough for my family and beloved girlfriend Man Luo. Without their continuous support and great encouragement, I cannot reach where I am standing now.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Introduction to omics studies and data analysis challenges

Biological research is advancing into an *omics* era: the target of more and more research effort is moving from individual genes, proteins and metabolites to genome-wide or system-wide spectrums of these biological units. Two major forces have driven the emergence of omics studies. First, there is a need. Decades of intensive biological studies based on the so-called *reductionism*, the idea that biological systems should be investigated from the lowest level (Brigandt and Love, 2008), have made unquestionable achievements. At the same time, reductionism enables people to be aware that the function of an integrated biological system is largely based on mutual interactions among those basic biological units that form a wide spread and complex network. Therefore, it is unrealistic to predict or reveal the integrated mechanism of a behavior of the system by merely looking at isolated components (Ahn *et al.*, 2006; Strange, 2005). Second, there is a way. Technological breakthroughs catalyze the evolution of biological research

approaches, and make it possible to generate rich datasets within an acceptable amount of time. There is no doubt that, without the invention of high-throughput sequencing, cDNA microarray, or instruments that can separate and quantitate the small and large molecules of interest, genomics, transcriptomics, proteomics, or metabolomics would be nothing more than armchair strategies.

Omics studies share several common features. First, unlike a traditional reductionism approach that is typically driven by a specific hypothesis, omics studies are usually regarded as "discovery tools" (Ahn *et al.*, 2006; Strange, 2005). Second, omics studies deal with large datasets (Joyce and Palsson, 2006). Hundreds to even tens of thousands of biomarkers are examined simultaneously (Clarke *et al.*, 2008). Third, the basic units under study are well connected in a global and hidden network. These unique features of omics studies not only provide us a wealth of opportunities to derive fundamental information about whole biological systems, but also raise as many methodological challenges on how to mine the vast amount of data efficiently and thoroughly. For the rest of this section, we will illustrate these features by a brief survey of major omics disciplines, with emphasis on the common scientific questions raised and the types of data generated. Each of these omics studies focuses on one particular molecular biology layer. Common computational challenges shared by omics studies will be summarized at the end of this section.

1.1.1 Genomics

Genomics is the broad study of individual organism's whole DNA set. The field includes all the intensive efforts to determine the entire DNA sequence of organisms and reveal the hereditary information encoded in DNA. A significant portion of this field aims at deciphering the genetic basis of a phenotype of interest, e.g., disease susceptibility, based on

genetic variations in DNA sequence. There are different types of genetic variants. Examples are restriction fragment length polymorphism (RFLP), microsatellites, copy number variation (CNV) and single nucleotide polymorphisms (SNP) (Strachan and Read, 2010). Among all the genetic variants, SNP is one of the most widely used genetic biomarkers. SNPs exist almost all over the genome. The international HapMap project (International HapMap Consortium, 2005) intended to provide a library of human SNPs. It focused only on common SNPs (minor allele frequency (MAF) $\geq 1\%$), and identified several million well-defined SNPs. To find genetic factors involved in a particular phenotype, one can select a set of tag SNPs (SNPs that can capture most of the haplotypes in a region of linkage disequilibrium), and determine their genotypes for a cohort of subjects with heterogeneous phenotype. From this, a genome-wide association study (GWAS) can be performed to examine SNPs' association with the phenotype at a global scale. A GWAS typically involves several hundreds to thousands of subjects, up to millions of SNPs, and a continuous or categorical phenotype of interest.

Due in part to the limitation of sequencing techniques, conventional GWA studies typically evaluated the common disease common variant (CDCV) hypothesis. This hypothesis posits that common, interacting disease alleles underlie most common diseases (Wang *et al.*, 2005), and each has only small or medium effect. Studies that followed the CDCV hypothesis, however, enjoyed only limited successes over the last decade: only a limited amount of the heritable components of complex disease have been identified (Easton and Eeles, 2005; Frazer *et al.*, 2009; Lettre and Rioux, 2008). As a result of the exponential decrease in the cost and the increase in speed of whole genome sequencing, the availability of next generation sequencing (NGS) data makes possible a fundamental transformation within this field. NGS data covers much denser genetic markers, including many rare variants, which allows us to evaluate an distinct hypothesis: the common

disease rare variant (CDRV) hypothesis. This hypothesis presumes that rare variants play a critical role in the heritable part of common diseases, and each possesses medium to large effect.

1.1.2 Transcriptomics and proteomics

According to the central dogma of molecular biology, information in the genome is turned into mature products through two major steps: the information in DNA is first transcribed into RNA, and then translated into proteins. Each step is modulated by a complicated and integrated system of mechanisms. These regulations are also temporally and spatially specific. Therefore, although most of the modulators of these mechanisms are essentially gene products and encoded in the genome (Strachan and Read, 2010), the mechanism of variation in many complex phenotypes cannot be perfectly recovered by merely using DNA information. Transcriptomics and proteomics focus on end products after each step of genetic information transmission, and therefore are able to (partially) capture the regulatory information.

First, transcription of genes is regulated by interactions between several transcription factors and RNA polymerase II(the enzyme that categorizes the transcription event) (Latchman, 2005), and finally generates mRNA. The abundance of mRNA corresponding to one particular gene reflects gene activity at the specific time and location. Transcriptomics studies the transcriptome, the expression profiles of genes in one or a population of cells by measuring the abundance of mRNA transcripts on a genome-wide basis. cDNA microarray is an important transcriptomics technique. It is able to capture a snapshot of the expression state of thousands of genes at the time the target sample is collected. In brief, this technique hybridizes the target sample with an array of probes on a chip, each of which is a specific DNA sequence representing one gene. A typical microarray

dataset is a matrix with each row representing one sample and each column representing one probe of a gene. The number of probes can reach several thousands or even tens of thousands, while the sample size rarely exceeds one hundred. Expression of genes can be highly correlated, especially for those that act together in a pathway (Qiu *et al.*, 2005). One primary goal of microarray studies (Gohlmann and Talloen, 2009) is to compare gene expression profiles among two or more classes of samples, e.g., different molecular subtypes of human cancers (Rhodes and Chinnaiyan, 2005), so as to determine which genes are differentially expressed. Another goal is to decipher mutual regulatory relationships among genes (Filkov, 2005) by their correlated expression profile, which may involve network reconstruction.

The abundance of mRNA is not perfectly proportional to the abundance of corresponding proteins. After mRNA is formed, its translation rate is regulated by another system of mechanisms (Latchman, 2005). Moreover, alternative splicing to mRNA can happen in eukaryotes after transcription so that multiple proteins can be produced from one gene. Therefore, a transcriptome can only be regarded as a precursor for the proteome, the entire complement of proteins produced by a cell or a population of cells at a specific stage of its life circle. Strictly speaking, proteomics is the study of all aspects of the proteome: their structures, cellular localizations, interactions and their expression profile. Here, we will emphasize the last one. Mass spectrometry (MS) combined with chromatography is a popular technique (Patterson and Aebersold , 2003) to identify and quantify the cellular levels of genome-wide proteins. Unlike cDNA microarray studies, raw data from MS platforms has to go through several preprocessing steps in order to be transformed from complex spectrograms into protein identities and their relative abundance, and meanwhile account for various causes of variation (Wagner *et al.*, 2003). A final dataset after data preprocessing, however, is very similar to a cDNA microarray

dataset in that there are typically tens of rows representing subjects and thousands of columns representing proteins. For a proteomics study of protein expression profile, we are usually interested in identifying proteins that change in response to treatments or discriminate classes of subjects. Therefore, many statistical methods adapted to transcriptomics studies also apply here.

1.1.3 Metabolomics

An omics study that is even closer to the phenotype is metabolomics, which studies the entire repertoire of metabolites contained in cells and/or tissues (Kaddurah-Daouk *et al.*, 2008, 2009; Kaddurah-Daouk and Krishnan, 2009). The identities and concentrations of all the metabolites represent a complex interplay between the gene products and environmental factors. Therefore, metabolomics provides a functional readout of the entire cellular state (Joyce and Palsson, 2006). On the other hand, compared with the biological units studied by transcriptomics and proteomics, metabolites are highly diverse and dynamic in their range of concentrations, which represents a major challenge to this discipline.

Several techniques have been used to separate and quantitate hundreds of metabolites simultaneously in body fluids, such as urine and plasma, many of which are shared with proteomics. Examples are gas chromatography in conjunction with mass spectroscopy (GC/MS), liquid chromatography with mass spectroscopy (LC/MS), and nuclear magnetic resonance spectroscopy (NMR) (Nicholson *et al.*, 2007). A typical metabolomics dataset is also very similar to a gene expression dataset, except that the gene probes here become metabolites. Compared to DNA sequencing and microarray, metabolomics techniques are not mature yet: the set of detected metabolites can not cover the global metabolic network very well, and the identities of some detected signals cannot be fully

determined.

A typical metabolomics study involves the collection of concentrations of hundreds of metabolites in an attempt to gain an overall understanding of metabolism and/or metabolic dynamics associated with some conditions of interest, including disease (Kaddurah-Daouk *et al.*, 2009; Kaddurah-Daouk and Krishnan, 2009) and drug exposure (Holmes *et al.*, 2008; Kaddurah-Daouk *et al.*, 2008). Despite those challenges stated above, metabolomics still enjoys a number of successful stories in detecting biomarkers for diseases status (Bogdanov *et al.*, 2008; Brindle *et al.*, 2002; Han *et al.*, 2001; Kaddurah-Daouk *et al.*, 2007, 2009; Kaddurah-Daouk and Krishnan, 2009) and therapeutic response (Ji *et al.*, 2011; Kaddurah-Daouk *et al.*, 2008).

1.1.4 Methodological challenges for Computational Biology and Statistics

The challenges in omics data analysis not only stem from the statistical properties of high-dimensional data, which are also faced by many other data-rich disciplines (Joyce and Palsson, 2006), but also reflect some inherent and unique biological properties of omics data. This section discusses these challenges and their implications.

The "curse of dimensionality"

It can be easily seen from the above examples of omics studies that the number of biomarkers, or in statistical jargon, the dimensionality of the input space can easily reach several thousands, more than enough to create annoying mathematical phenomena that prohibit the application of a majority of traditional statistical models. In 1961, Richard E. Bellman first coined the term "curse of dimensionality" when he referred to

the phenomena that the amount of data to sustain a given spatial density of data points increases exponentially with the dimension of the input space. Here we borrow this term to represent a series of problems caused by the high dimensionality in omics data analysis.

Consider a general problem of identifying biomarkers associated with some phenotype of interest. If marginal tests are used, then we will first miss all the interactions among the candidates. Besides, the statistical significance of marginal tests has to be adjusted for multiple testing, which may dramatically reduce the power of the study. Suppose in a transcriptomics study we want to identify disease biomarkers among 1,000 genes using 50 patients and 50 controls. Here, a Bonfferoni correction will result in a p-value criteria of 5×10^{-5}. This means there is only a 40% chance we can detect a true biomarker with a large Cohen's d effect size of 0.8 (Cohen, 1988), and less than a 5% chance to detect a medium effect size of 0.5. Performance of many classical multivariate methods is also severely impacted by the high ratio of number of variables to sample size. First, any statistical method based on spatial density of samples in the input space, such as neighborhood-based classification approaches and nonparametric local smoothing approaches (Scott, 1992), will be adversely influenced. Second, parameter estimation of many other multivariate models, e.g., the widely used ordinary least square (OLS) for the multiple linear regression models, even become not applicable, when the sample size is smaller than the number of variables (known as small-n-large-p problems).

Nonlinearity and multi-modality

These two features have to do with the underlying model that generates the phenotype. Despite that a linear additive function is probably the simplest way to model joint effect of multiple biomarkers, this is not the optimal way in the face of many biological mechanisms. Interaction represents a large class of nonlinear effects and widely exists

in biological systems. A good example is the genetic epistasis, or the the phenomenon where the effects of one gene are modified by one or several other genes, which is an important component of the genetic architecture of many biological phenotypes (Roth *et al.*, 2009). The inherent network structure within omics data, which will be further discussed later, can be regarded as a representation of interactions at the global level. Considering the complex interaction mechanisms in biological pathways and networks, naively adding interaction terms to the original linear additive functions is usually not a satisfactory solution. Besides, adding interaction terms will further increase the already high dimensionality of the model and make it even harder to solve.

Multi-modality is another confounding factor that adds to the complexity of biological data analysis. Several different but inter-related biological processes can control one phenotype of interest. Multi-modality refers to the cases where one biological unit can participate in multiple processes, and its true effect can be obscured if its multiple roles are not properly recognized and characterized. In statistics, this means the response variable (phenotype) can be generated from two or more mathematical functions, which share explanatory variables (biomarkers). The shared explanatory variables may possess distinct effects in different functions, which may be hard to be captured if treated as one unknown but constant value. Clarke *et al.* (2008) provided rich biological evidences for several sources of multi-modality.

Correlation and network

Nontrivial correlations, usually found in omics datasets, can come from a variety of sources. For instance, SNPs that are physically close to one another may exhibit high correlation (linkage disequilibrium); expression of genes under the regulation of the same transcript factor can be highly correlated; metabolites in the same compound class (e.g.,

bile acids) can be highly correlated. On one hand, the highly correlated nature of omics datasets is challenging for data analysis: it may cause a so-called "multicollinearity" problem, which can lead to unstable estimation for regression models. When univariate tests are controlled for the multiple comparison issue, the high correlation among variables may obscure the effective number of tests conducted and inflate the false negative rate. On the other hand, the correlation structure within omics datasets can potentially provide evidence on some in-depth biological mechanisms, which is related to the following important feature of omics data.

There are inherent network structures behind omics datasets. Examples are gene regulatory networks behind transcriptomics data, and metabolic pathways behind metabolomics data. In a gene regulatory network, two genes are linked if the product of one gene affects the transcription of the other gene. In a metabolic network, metabolic reaction products are linked to their precursors. Such network information has been accumulated by decades of intensive biological research, and are accessible from many public databases (Kanehisa and Goto, 2000; Karp *et al.*, 2005; Matthews *et al.*, 2008). This information can (and should) be used to facilitate statistical analysis. For example, if a variable is found related to the phenotype, then it is reasonable to assume that its neighbors in the network are also likely to be involved in the mechanism. It also becomes more interesting to identify phenotype-relevant functional modules or pathways in the network.

1.2 Penalized regression models and variations

Many statistical methods have been developed to meet (some of) the challenges mentioned above. In this section we review a family of penalized regression models. They

work for small n large p problems, and some are robust for multicollinearity. These models also serve as a basis of statistical methods for biological network inference and reconstruction, which are reviewed in sections 1.2.2 and 1.2.3. These methods are also embedded in a comprehensive omics data analysis pipeline proposed in Chapter 2.

1.2.1 Penalized linear regression models

Consider a classical linear regression model where the phenotype Y is predicted by p variables, $X_1, X_2, \ldots, X_p$,

$$Y = \mu + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_p X_p + \varepsilon.$$

In order to estimate the coefficient $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, OLS minimizes the squared residual error

$$f(\mu, \boldsymbol{\theta}) = (\mathbf{Y} - \mu - \mathbf{X}\boldsymbol{\theta})^\intercal (\mathbf{Y} - \mu - \mathbf{X}\boldsymbol{\theta}), \tag{1.1}$$

where $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^\intercal$ is a vector of n phenotype observations and $\mathbf{X} = (x_1, x_2, \ldots, x_n)^\intercal$ is a $n \times p$ design matrix. The OLS estimates often have low bias but large variance; when $n < p$, the solution to (1.1) is not unique. To increase numerical stability, a trade-off between bias and variance can be adopted by imposing a penalty term to (1.1):

$$\tilde{f}(\mu, \boldsymbol{\theta}) = (\mathbf{Y} - \mu - \mathbf{X}\boldsymbol{\theta})^\intercal (\mathbf{Y} - \mu - \mathbf{X}\boldsymbol{\theta}) + g(\boldsymbol{\lambda}, \boldsymbol{\theta}),$$

where $\boldsymbol{\lambda}$ represents penalty coefficients, and $g(\boldsymbol{\lambda}, \boldsymbol{\theta})$ is the penalty term. Given a fixed $\boldsymbol{\lambda}$, estimation of coefficients $\boldsymbol{\theta}$ is a balance between minimizing the squared residual error

and the penalty. Different penalty terms grant different model properties. The bridge regression (Frank and Friedman, 1993) adopts $\lambda\|\boldsymbol{\theta}\|_k = \lambda\sum_{j=1}^{p}\boldsymbol{\theta}_j^k$, $k \geq 0$, as its penalty term, which has several well-known special cases. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) criteria for subset selection adopt $\lambda\|\boldsymbol{\theta}\|_0$ with different λ. The ridge regression, which can date back to 1960s, uses $\lambda\|\boldsymbol{\theta}\|_2$ as its penalty term and is more robust to multicollinearity among predictors compared to OLS. The so-called "lasso" penalty (Tibshirani, 1996) is $\lambda\|\boldsymbol{\theta}\|_1$. A favorable property of lasso is that it is able to shrink coefficients of trivial variables to 0 and therefore remove them from the model. After lasso has been invented, a number of its variants have also been proposed. Examples are 1) elastic net (Zou and Hastie, 2005), which uses a mixture of $\lambda_1\|\boldsymbol{\theta}\|_1$ and $\lambda_2\|\boldsymbol{\theta}\|_2$ as a penalty and achieves both the ability of variable selection and some robustness to multicollinearity; 2) group lasso (Yuan and Lin, 2006), which selects a pre-defined group of variables at a time; 3) fused lasso (Tibshirani *et al.*, 2005), which assumes variables can be ordered in a meaningful way and penalizes both the coefficients and their successive differences.

1.2.2 Network-constrained penalized regression models

Recently, massive amount of biological network information have been accumulated by intensive biological and biomedical research, and stored in several publicly accessible biological function/pathway databases, such as Genome Ontology (GO) (The Gene Ontology Consortium, 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), Reactome (Matthews *et al.*, 2008), and BioCyc (Karp *et al.*, 2005). This section will review regression models that perform variable selection on a known network. In other words, pathway and network structure is incorporated into statistical models to facilitate variable selection. A key assumption made by these methods is that variables

directly connected in the network is more likely to behave similarly. This family of methods can be summarized as follows: introduce a network-constrained penalty term to a classical regression model.

For example, in the model proposed by Li and Li (2008), imposing their network-constrained penalty term to lasso introduced in the section above gives the following objective function:

$$f(\mu, \boldsymbol{\theta}) = (\mathbf{Y} - \mu - \mathbf{X}\boldsymbol{\theta})^{\intercal}(\mathbf{Y} - \mu - \mathbf{X}\boldsymbol{\theta}) + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \sum_{i \sim j} (\frac{\theta_i}{\sqrt{d_i}} - \frac{\theta_j}{\sqrt{d_j}})^2 W_{ij} \quad (1.2)$$

where λ_1 and λ_2 are two non-negative penalty coefficients, $\sum_{i \sim j}$ denotes a sum over all pairs of neighboring variables in the network, d_i is the degree of the i-th variable, and W stores weights of network edges. The penalty of this model contains two parts. The first penalty term is a lasso penalty, which maintains model sparseness. The second penalty term encourages smoothness of coefficients on the network: minimizing (1.2) will shrink the weighted difference between coefficients of neighboring variables towards 0. Furthermore, variables with higher degree or more neighbors tend to have larger coefficients. Again, the biological motivation of this penalty is that if we expect the biological variables that are linked on the network to have similar functions, then they should have similar regression coefficients. On the other hand, the model does not force these coefficients to be exactly the same so as to avoid the trap of self-fulfilling prophesy. Computationally, the authors show that the model can be easily reformulated to a standard lasso-type problem and therefore solved by algorithms for lasso.

Following the same logic, Pan *et al.* (2010) proposed another penalty term, which is a sum of L_k-norm of the two coefficients for each pair of neighboring nodes in the network. Kim, Sohn and Xing (2009) proposed graph-guided fused lasso models in order

to incorporate network structure among multiple response variables. Their penalty term is a sum of absolute difference between each two regression coefficients corresponding to the same predictor for two neighboring response variables in the network.

1.2.3 Construction of relevance network

The previous section discussed several ways to incorporate information about network structure into statistical models. However, as has been mentioned in Section 1.1.1, many biological interactions are very "dynamic", which may change with spatial, temporal and/or other micro or macro conditions. For this reason, people are also interested in inferring relationships among those dynamic biological units, such as transcripts and metabolites, under a specific circumstance. This section reviews a series of statistical methods for biological relevance network reconstruction.

In brief, the purpose of network construction is to infer mutual-dependence relationship among a set of variables using a certain number of observations on these variables. A network can be always classified as a directed network or undirected network, which can be easily distinguished by whether its edges are directed or not directed. Bayesian network is a framework for directed network inference, while Markov random field (MRF) is a class of graphical models that are widely used to infer undirected networks. In this section, we will mainly focus on MRF.

Strictly, an undirected network G is considered a MRF with respect to the joint probability distribution of its nodes or variables $P(X = x)$, if and only if graph separation in G implies conditional independence: if two nodes X_i and X_j are separated in G after removing a set of nodes Z from G, then X_i and X_j must be conditionally independent given Z. A straightforward deduction from this definition is that an edge between X_i and X_j is absent, if and only if X_i and X_j are independent conditional on all the rest

variables in the network.

One popular assumption imposed on MAF is the joint Gaussian distribution of all variables in the network or $X \sim N(\mu, \Sigma)$. MAF with joint Gaussian assumption is called Gaussian Markov random field (GMRF). Denote Σ^{-1} as Ω. One of the reasons why the Gaussian assumption is imposed is the following famous observation: given $X \sim N(\mu, \Sigma)$, X_i and X_j are independent conditional on all the rest variables on the network, if and only if $\Omega_{ij} = 0$ (Edwards, 2000). Ω is also called the precision matrix or concentration matrix. Therefore, network construction is equivalent to the identification of non-zero elements in Ω, which is essentially a model selection problem.

When the sample size n is relatively small compared to the number of variables p, traditional approaches based on the maximum likelihood estimation for the multivariate normal distribution of X can give very unstable estimation of Ω; they become inapplicable when $n < p$. Many methods have been suggested to face the challenge. A large branch of these is based on penalized regression models we reviewed in Section 1.2.1. Meinshausen and Buhlmann (2006) were among the earliest of those who proposed such a method. They simply fit one lasso model for each variable using all the other variables as predictors. Ω_{ij} is determined as non-zero, if and only if, both the estimated coefficients for variables X_i and X_j are non-zero in the other model.

Later, researchers developed methodologies that fit just one penalized model, all of which can be formulated as minimizing the following penalized loss function

$$\Gamma(\Theta, \mathbf{X}) = L(\Theta, \mathbf{X}) + J(\Theta) \tag{1.3}$$

where Θ is the overall set of parameters to be estimated, $L(\Theta, \mathbf{X})$ represents the original loss function of certain form, and $J(\Theta)$ stands for a penalty term. In the remainder

of this section, we assume observations are properly standardized by sample mean and standard deviation.

Yuan and Lin (2007b) proposed to directly penalize the off-diagonal elements of matrix Ω, resulting in the following loss function:

$$\{-\ln|\Omega| + \text{trace}(\Omega S)\} + \lambda \sum_{i \neq j} |\Omega_{ij}|, \tag{1.4}$$

where S represents sample covariance matrix, and the quantity in $\{\ \}$ is the original negative log likelihood of the GMRF. Only positive definite matrices are considered candidates for Ω. They also proposed a non-negative garrote estimator for Ω based on a relatively reliable preliminary estimator $\tilde{\Omega}$. The nonnegative-type panelized loss function is

$$\{-\ln|\Omega| + \text{trace}(\Omega S)\} + \lambda \sum_{i \neq j} D_{ij}, \tag{1.5}$$

subject to $D_{ij} > 0$ for all $i \neq j$, where $\Omega_{ij} = D_{ij}\tilde{\Omega}_{ij}$, and the minimization is performed on D instead of Ω. The shrinkage parameter can be selected based on model BIC score. Asymptotically, this nonnegative-type estimator enjoys a property that, as $n \to \infty$ and p is fixed (though $p \to \infty$ would be more realistic), it selects the correct network with probability tending to one and at the same time gives $\sqrt{n}$ consistent estimator for the non-zero elements of Ω. Computationally, since the objective function is non-linear with respect to elements of Ω, Yuan and Lin (2007b) adopted interior point optimization methods to solve the problem. Friedman *et al.* (2008) later proposed an efficient algorithm, glasso, to implement this method, which is much more efficient to handle high dimensional data. However, the computation complexity is still $O(p^3)$, which is less efficient

than the following method when $n < p$.

The sparse partial correlation estimation (SPACE) proposed by Peng *et al.* (2009) is driven by the following lemma: for GMRF, X_i can be expressed as the regression model $X_i = \sum_{j \neq i} \beta_{ij} X_j + \varepsilon_i$ such that ε_i is uncorrelated with all the variables except X_i, if and only if $\beta_{ij} = -\Omega_{ij}/\Omega_{ii} = \rho_{ij|rest}\sqrt{\Omega_{jj}/\Omega_{ii}}$, where $\rho_{ij|rest}$ is the partial correlation between X_i and X_j conditional on all the rest variables. Hence, the proposed penalized loss function is

$$\sum_{i=1}^{p} w_i \|\mathbf{X}_i - \sum_{j \neq i} \beta_{ij} \mathbf{X}_j\|^2 + \lambda \sum_{1 \leq i < j \leq p} |\rho_{ij|rest}|, \tag{1.6}$$

where w_i is a weight for variable X_i or the i-th regression model, and $\mathbf{X}_i$ is the i-th column of $\mathbf{X}$ or a vector of the n observations for variable X_i. The assignment of weights grants much flexibility to this model, and simply switching the settings of weights can generate a number of variants. In theory, there are several advantages of this model over the penalized likelihood method proposed by Yuan and Lin (2007b): 1) the weights in (1.6) can be utilized to incorporate prior knowledge of network structure; 2) the lasso-type penalization is used, but the model is flexible to incorporate other types of penalization. On the other hand, the penalized likelihood method proposed by Yuan and Lin (2007b) can not do so, because their computation algorithm is specifically designed for the lasso-type penalty; 3) its computation complexity is $min(O(np^2), O(p^3))$, and is therefore more efficient when $n < p$; 4) this method can be shown to identify consistently the correct network structure when both n and p go to infinity. One disadvantage is that this method cannot guarantee positive definiteness of Ω. However, the authors argued that, based on their comprehensive simulation studies, nonpositive definiteness only occurs when the resulting model is huge and much larger than that of the true model, which rarely occurs

in practice.

1.3 Dimension reduction methods

Briefly speaking, dimension reduction methods aim to reduce the number of variables in a scientific problem. According to this definition, a variety of variable selection methods fall into this category. However, in this section, we will focus on another type of methods, that seek linear combinations of the original variables to capture some essential information of interest. Due to this key feature, these methods serve as good "summary" tools for aggregating information from a group of variables. These methods can be classified further into several categories. In this section, we will review several classical dimension reduction methods by their categories: 1) principle component analysis (PCA) representing unsupervised dimension reduction methods; 2) ordinary least square (OLS) and partial least square (PLS, Wold, 1996) representing supervised dimension reduction methods; 3) sliced inverse regression (SIR, Li, 1991) representing sufficient dimension reduction (SDR) methods, which is a subclass of supervised dimension reduction, but aims to capture all the information in the original predictors about the response variable. In the later chapters, the readers will see some other sufficient dimension reduction methods, and their extensions, to better capture and aggregate information of interest. Performances of dimension reduction methods introduced in this section and the following chapters will be compared by simulation studies in Chapters 3 and 4.

1.3.1 Unsupervised dimension reduction

PCA is a well-known representation of unsupervised dimension reduction methods, which only perform dimension reduction to predictors, but do not take into account any response

information. PCA seeks linear combinations that explain the most variance in the original predictors. Specifically, the k-th principle component has the highest variance among all possible linear combinations orthogonal to the previous components:

$$a_k = \arg \max_{\substack{a \in R^{p\times 1} \\ a^\top a=1 \\ a^\top a_j=0, j=1,\ldots,k-1}} \text{var}\{a^\top X\}.$$

It can be shown that the PCA directions are essentially leading eigenvectors of the covariance matrix of the original predictors. The corresponding eigenvalues are proportional to the variability that each direction can explain, and, therefore, serve as a measure of importance.

Limitations of PCA are obvious. First, since it is an unsupervised method, the PCA components are not necessarily related to the response variable at all. Second, the PCA estimates are restricted to be linear combinations. This vastly simplified the problem, but at the same time limits its ability to handle cases where the underlying major components are nonlinear functions of the original predictors.

1.3.2 Supervised dimension reduction

The well known OLS can be regarded as one of the simplest supervised dimension reduction methods. The purpose of OLS is to identify a linear combination of the original predictors with the maximum correlation with a continuous response variable Y

$$a_{OLS} = \arg \max_{\substack{a \in R^{p\times 1} \\ a^\top a=1}} \text{corr}^2\{Y, a^\top X\}.$$

a_{OLS} is derived by minimizing the sum of squared differences between the observed responses and the responses predicted by the linear approximation, $a^{\intercal}X$, with respect to a.

OLS is widely used in linear regression models, but it is not applicable to problems where the sample size is less than the number of variables, which prevents its application to many problems emerged from omics studies. The reason is the OLS solution depends on the inverse of $\mathbf{X}^{\intercal}\mathbf{X}$, which does not exist if $\mathbf{X}^{\intercal}\mathbf{X}$ is not of full rank. The dependence of OLS on this inversion also results in its vulnerability to high correlation among predictors.

PLS (Wood, 1966) serves as an alternative to OLS for the poorly or ill-conditioned problems. It draws away the estimation from the OLS solution toward directions in the predictor space of larger sample spread: (Frank and Friedman, 1993; Garthwaite, 1994)

$$a_k = \arg \max_{\substack{a \in R^{p\times 1} \\ a^{\intercal}a=1 \\ a^{\intercal}a_j=0, j=1,\ldots,k-1}} \text{var}\{a^{\intercal}X\}\text{corr}^2\{Y, a^{\intercal}X\}. \tag{1.7}$$

The PLS solution is obtained in an iterative manner. In the initial step, Y is marginally regressed on each predictor X_i, and the first PLS component T_1 is a sum of products of all predictors with their marginal regression coefficients. Regressing Y on T_1 then obtain a residual of Y, and a residual is also obtained for each predictor X_i by regressing X_i on T_1. Subsequently, in the m-th step, the m-th PLS component T_m is obtained in a similar way except that Y and X are replaced with their residuals from last step. In this way, PLS obtains a series of mutually uncorrelated components.

OLS and PLS, among many other supervised dimension reduction methods, aim to capture information within the original predictors that is related to the response variable. However, none of these can guarantee that there is no missing information. Below, we introduce sufficient dimension reduction, a sub-class of supervised dimension reduction,

which aims to capture all the information that the predictors contain about the response variable.

1.3.3 Sufficient dimension reduction

Before explaining sufficient dimension reduction, we will first introduce several relevant concepts. First, a dimension reduction subspace (DRS) is span(B), where B is a $p \times d$ matrix that satisfies $Y \perp X|B^{\intercal}X$. Obviously, given a pair of Y and X, there can be infinite number of DRS'. A central subspace (CS) $S_{Y|X}$ is the intersection of all DRS', and itself is a DRS. By definition, it is clear that a CS is a minimum DRS; therefore, it achieves the maximum dimension reduction. It can be shown that such a CS uniquely exists under mild conditions (Cook, 1998). Sufficient dimension reduction aims to estimate CS assuming its existence.

SIR is one of the earliest sufficient dimension reduction methods. Suppose a true model is

$$Y = f(\beta_1^{\intercal}X, \beta_2^{\intercal}X, ..., \beta_d^{\intercal}X, \varepsilon), \tag{1.8}$$

then $S_{Y|X} = \text{span}(\beta_1, \beta_2, ..., \beta_d)$. Given Z is the standardized version of X, $Z = \Sigma^{-1/2}(X - E(X))$, then $S_{Y|Z} = \text{span}(\nu_1, \nu_2, ..., \nu_d)$, where $\nu_i = \Sigma^{1/2}\beta_i$. The idea of SIR to estimate CS stems from the following proposition: under the linearity condition (that is $E(Z|\gamma^{\intercal}Z$ is linear in columns of $\gamma^{\intercal}Z$, where γ denotes of a basis of $S_{Y|Z}$), $E(Z|Y) \in S_{Y|Z}$. To capture the high variance directions of $E(Z|Y)$, we can apply PCA to span($E(Z|Y)$), which results in the following eigen decomposition problem

$$\text{Cov}\big\{E(Z|Y)\big\}\theta_j = \lambda_j\theta_j, \text{ j=1, ..., d.} \tag{1.9}$$

The leading eigenvectors of $\mathrm{Cov}\big\{E(Z|Y)\big\}$ span a subspace of $S_{Y|Z}$. The word *sliced* in the name of SIR comes from the manner by which it estimates $\mathrm{Cov}\big\{E(Z|Y)\big\}$. It first slices the range of Y into h non-overlapping intervals denoted by $\tilde{Y}$. Within each interval, the expectation of Z is estimated by

$$\hat{E}(Z|\tilde{Y}=s)=\frac{1}{n_s}\sum_{Y_i\in s} z_i, \tag{1.10}$$

where n_s is the number of samples within interval s. $\mathrm{Cov}\big\{E(Z|Y)\big\}$ can then be estimated by

$$\widehat{\mathrm{Cov}}\big\{E(Z|Y)\big\}=\sum_{s=1}^{h}\frac{n_s}{n}\hat{E}(Z|\tilde{Y}=s)\hat{E}(Z|\tilde{Y}=s)^{\intercal}. \tag{1.11}$$

It can be also easily shown that solving (1.9) is equivalent to solve

$$\mathrm{Cov}\big\{E(X|Y)\big\}\eta_j=\lambda_j\Sigma_X\eta_j,\ \text{j=1, ..., d.} \tag{1.12}$$

The leading eigenvectors from this generalized eigen decomposition problem span a subspace of $S_{Y|X}$. Tests have been designed to determine the dimension of CS or how many eigenvectors should be chosen (Cook, 1998). The number of slices h is another tuning parameter; however, usually it does not significantly influence the estimation as long as $h>d$.

From (1.8), it is noted that SIR makes no assumptions on models of any form. This is actually a key advantage enjoyed by many SDR methods. However, SIR itself is still unable to handle the $n<p$ problems, as the inversion of the $p\times p$ matrix Σ_x is inevitable. In Chapter 3, we will extend SIR and some other dimension reduction methods by incorporating kernel functions, which addresses this issue and in addition

grant them some other favorable properties. These properties make these novel methods more appropriate for solving problems that emerge from omics studies.

1.4 Thesis contributions and organization

1.4.1 In-Depth mining of pharmacometabolomics data

In Section 1.1, we have surveyed several omics disciplines, and outlined major challenges for omics data analysis. Much effort in the field of statistics has been made to meet these challenges, some of which were reviewed in Section 1.2. Based on these current developments, and on our experience with metabolomics studies, we propose a pipeline for in-depth mining of data from pharmacometabolomics studies of interpersonal differences in therapeutic response, which is organized by scientific questions that are frequently asked in these studies. The readers will also note that novel statistical techniques are integrated into the pipeline to better answer those questions.

1.4.2 Nonlinear sufficient dimension reduction for biological pathway selection

The high dimensionality of omics data suggests that dimension reduction techniques can find enormous potential applications in this large area. However, one fundamental criticism to dimension reduction techniques lies in the interpretability of results. For instance, in gene expression studies, the linear combinations of the original genes produced by dimension reductions are named "super genes" (Antoniadis *et al.*, 2003). However, people find it difficult to appreciate the biological meaning behind this new terminology. In this thesis, we emphasize applications of dimension reduction in summarizing and aggregating

information in groups of variables with sound biological interpretation.

In Section 1.3, we have reviewed several dimension reduction methods. Meanwhile, we can also observe that there are several limitations with the reviewed approaches, which may impede their wide applications in biological omics data analysis. One common issue with all of these methods is that they can only obtain linear combinations, and are, therefore, deficient in capturing any nonlinear effects, which, however, has been mentioned as one of the major challenges from biological data analysis. This limitation motivates the theoretical research in Chapter 3, where we propose a framework of nonlinear extensions to the existing sufficient dimension reduction approaches by incorporating kernel functions. This upgrade also immunizes them against the small-n-large-p problem. These novel approaches are then used to solve two scientific problems in omics data analysis.

The first one is the general pathway selection problem. It is often believed that most biological systems operate not based on just isolated elements, but instead on their mutual and complex interactions and regulations, which form biological pathways. Therefore, a keen interest in analyzing omics data is to identify pathways that are related to or influence a phenotype of interest. Based on nonlinear sufficient dimension reduction approaches, we propose a two-step procedure for the purpose of pathway selection. It permits flexible and complex predictor interactions within the pathways, as well as nonlinear pathway effects on the response variable, without imposing any strong parametric model assumptions. Its computational advantage can effectively handle a very large number of variables, even if it far exceeds the number of observations.

1.4.3 Nonlinear sufficient dimension reduction for association testing of complex traits

Chapter 4 applies the proposed nonlinear dimension reduction approaches to GWAS. In Section 1.1.1, we mentioned that the CDRV hypothesis is attracting greater attention. However, the extremely small number of samples carrying the rare variants becomes an overwhelming difficulty of screening in-depth sequencing data for phenotype-related rare variants. A widely accepted general strategy is to aggregate genetic effect of multiple markers within a genome region, which becomes another natural application of the approaches proposed in Chapter 3. However, unlike Chapter 3 which emphasizes the overall methodological framework of nonlinear sufficient dimension reduction, we want to stress in this chapter one key ingredient of these approaches: the kernel functions. In our context, a kernel function should capture the genetic similarity between two individuals. The popular Gaussian kernel used in Chapter 3 may work well for continuous predictors but perform poorly on categorical predictors such as SNPs. Therein, we propose a unified way to generate kernels for such data based on Markov chain theory. Simulation studies demonstrate their advantages.

Finally, in Chapter 5, we make general conclusions and discuss possible future research.

Chapter 2

In-Depth Mining of Pharmacometabolomics Data: A Question-Oriented Data Analysis Procedure

2.1 Introduction

The classical biochemical studies often focus on single metabolites or single metabolic reactions; therefore, lack a systematic view. Recent technological advances have made it possible to move biochemistry to the global level, and give rise to a new omics field - metabolomics. The analytical platforms, including but not limited to gas chromatography(GC) or liquid chromatography (LC) in conjunction with mass spectroscopy (MS), and nuclear magnetic resonance (NMR) spectroscopy, are able to simultaneously identify and quantify hundreds to thousands of compounds present in a biological tissue or body

fluid, which offers us a unique opportunity to extensively study the global metabolic network.

Pharmacometabolomics is the application of metabolomics in the field of pharmacology with primary interests in understanding drugs' mechanisms of action (MOA) and variations in the drug response phenotypes (Kaddurah-Daouk *et al.*, 2008, 2007, 2009). In contrast to most traditional biochemical studies, which can be usually regarded as hypothesis-testing procedures, pharmacometabolomics is typically used as an exploratory or hypothesis-generating tool. This feature, which is shared among most omics studies, indicates that one can (and should) mine a pharmacometabolomics dataset in several different ways with different specific aims. One can test the impact of a drug treatment to patients' metabotype and futher infer the relationship among the change of metabotype components. This analysis can potentially facilitate the understanding of drugs' MOA. In another aspect, for the same dataset, one can also detect biomarkers that are associated with drug response variation, use them to predict drug response phenotype(s), decipher their relationships, identify drug response-related pathways, and connect them with genetic factors. This information can serve as the basis for the generation and testing of mechanistic hypotheses regarding the underlying basis for individual variation in drug response. In this chapter, we propose a practical pipeline for pharmacometabolomics data analysis, where both routine statistical methods and recently developed statistical techniques are integrated to fill the gap of current analysis procedure to answer the above scientific questions. The pipeline is organized by research questions of interest, instead of data analysis methods, in order to guide a comprehensive and in-depth mining of a pharmacometabolomics dataset.

The remainder of this chapter is organized as follows: Section 2.2 describes the data analysis pipeline in detail, which is further divided by major research questions; Sec-

tion 2.3 demonstrates examples of real data analysis according to the procedure; the last section draws conclusions and discusses limitations and possible extensions to this pipeline.

2.2 The pipeline

2.2.1 Experimental design and data structure

In this chapter, we focus on a commonly used design in the metabololomics studies of drug actions and response variations. In these studies, body fluid samples (e.g., plasma or urine) are collected before and after drug treatment for each enrolled subject. The samples are typically processed through one or several metabolomic analytical platforms. The original output from these platforms is a complex spectrogram for each sample, based on which the identity and relative concentration of a series of metabolites are determined. The whole spectrum of identified metabolites and their relative concentrations can be regarded as a snapshot of the global metabolic state for a subject at a specific time point. Note that the method to transform the raw output to the absolute or relative concentration of metabolites is itself a key data pre-treatment step. It is usually performed by data-generating centers and beyond the scope of this chapter. A typical dataset supplied to the subsequent data analysis procedure includes tens to hundreds of samples, two or several corresponding to one subject, and the relative concentration of hundreds to thousands of metabolites within each sample. Also recorded in the dataset are typically the measurement of one or several drug response and/or side effect phenotypes and some other clinical variables, e.g., age, gender, race, etc.

2.2.2 Data cleaning, quality control and pre-data analysis

Before performing data analysis, one needs to be aware that several data quality issues, including outliers, missing data, and skewed variable distributions, may severely impact many of the following numerical methods. Therefore, we begin our procedure with cleaning the raw data for a proper data quality assurance and quality control (QA/QC).

Determination of outliers

A search for subject outliers and possible miss-annotations can be accomplished with principal components analysis (PCA), where a number of metabolites are reduced to a small number of principle components, and samples are plotted in two or three-dimensional spaces for the first few components. Individual samples that lie far away from the majority are considered potential outliers. When outliers are detected, one needs to determine whether the outliers are due to miss-annotations that need to be corrected, or unusual medical conditions of the subjects, which may constitute the ground for exclusion of the subjects from further analysis.

Missing data

For missing data, it is critical to first clarify the *missingness mechanism*. In statistics, there are several different kinds of missingness mechanisms (Gelman and Hill, 2006): missingness completely at random, missingness at random, missingness that depends on unobserved predictors, and missingness that depends on the missing value itself. Missingness completely at random indicates that the probability of missingness is the same for all subjects, while missingness at random represents cases where the probability of missingness is not the same for all subjects, but can be obtained or estimated based only

on available data. In practice, the first three mechanisms are usually hard to distinguish from each other. Therefore, we typically assume missing completely at random, unless there is clear evidence to reject the assumption. One special case is that the sample metabolite concentration is below the detection level of the platform, which is an example of the fourth mechanism. In this case, it is reasonable to substitute a missing data point by a low quantity value, such as a value that reflects half of the detection limit for the platform.

There are several naïve but frequently used methods to handle randomly missing data: 1) remove subjects with missing data; 2) remove variables with missing data; 3) replace each missing data point with the mean of observed values of that variable. Methods 1) and 2) are usually performed when a metabolite or a subject has a significant proportion, say $\geq 10\%$ of data-points missing. In some other circumstances, especially when the missing data distributes throughout the dataset, methods 1) and 2) will severely reduce the sample size and therefore the study power. Method 3) is one of the simplest ways of implementing a statistical procedure called imputation. However, method 3) can distort distribution for the variable toward a lower standard deviation, and pull correlations between this variable and other variables toward 0.

We recommend another way of imputation, which predicts missing points using all the metabolites without missing data. A variety of statistics and machine learning methods can be used to build the predictive model, and many of them are able to provide satisfactory performance, which is probably due to the high correlation among metabolites. In practice, we use penalized regressions for imputation. This family of methods fit linear regression models, which differ from the widely used ordinary least square (OLS) in that they introduce an additional penalty term to the square error loss function of OLS for the seek of higher prediction accuracy. A more detailed description will be given jointly with

other predictive models in Section 2.2.4. A high quality imputation (predictive $R^2 > 0.7$) can be reasonably expected for most metabolites.

Distribution of metabolites

We perform a number of analysis to check the distributions of metabolites. In general, metabolites do not follow the well-known normal distribution. Instead, most of them are heavily right skewed (figure 2.1). The influence of right skewness of data distribution on the subsequent data analysis depends on what types of analysis will be used. Nonparametric methods, which mainly use the rank of data points instead of their original values, are robust to the skewed distribution. However, to the knowledge of the author, statistical questions that can be addressed by such kind of methods are quite limited. Parametric models that make use of original values is much more versatile, but may suffer from the distribution issue and produce unstable estimation. Therefore, log transformation is usually performed before the following data analysis.

After going through the pre-processing steps, we are ready to analysis the data in several different ways so as to address different research questions.

2.2.3 Drug exposure signatures

The metabolome represents a metabolic state for an individual at a particular time point, which is jointly regulated by interactions between genetic effects and environment influences (Kaddurah-Daouk and Krishnan, 2009). An environmental stimulus, e.g., a drug treatment, may disrupt metabolism and result in changes that are long lasting and that can be captured as metabolic drug exposure signatures. Analysis of these signatures and their mutual relationships can potentially provide information with regard to drug mechanisms and lead to better understanding of both drug efficacy and side effects.

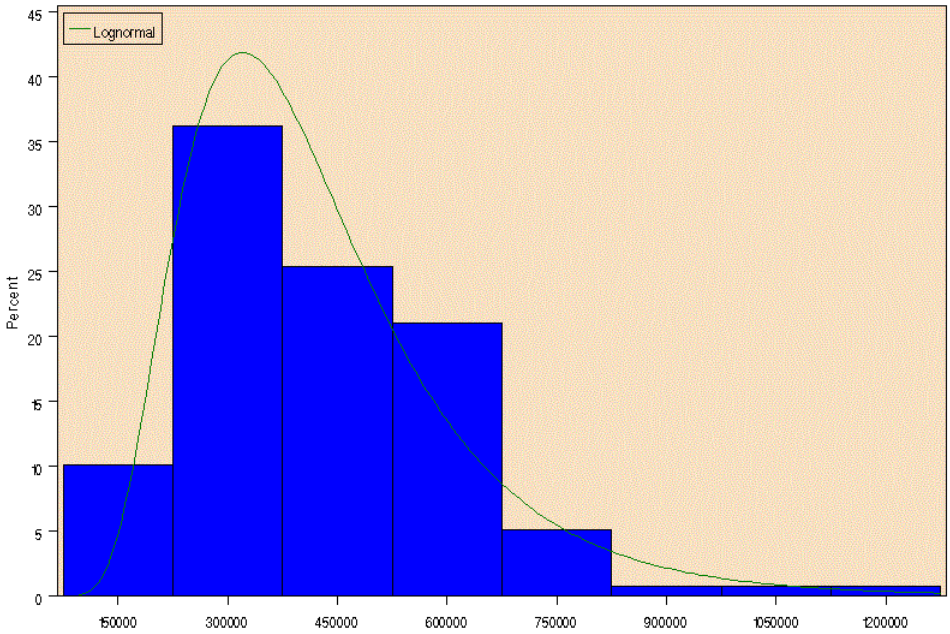

Figure 2.1: A typical distribution of metabolite and a fit log-normal distribution. It is clearly seen that the distribution of this metabolite is right-skewed.

Detection of drug exposure signatures

Change of metabolite concentration can be tested by a parametric paired t-test or a non-parametric Wilcoxon signed rank test. It is always informative to perform such analysis in a systemic way. The information can also be used to compare with other more sophisticated analysis results. Notice that one should control simultaneous multiple testing inherent in the need to conduct a test for each metabolite. Several statistical techniques have been developed for this purpose. A significant portion of them, e.g., Bonferroni correction, control for the family-wise error rate, which is the probability of at least one false positives existing among all the hypotheses. However, it can be argued that controlling the family-wise error rate is unnecessarily stringent in the context of metabolomics studies, because for those exploratory studies, falsely declaring several metabolites as

being significant will not be a serious problem if the majority of significant metabolites are correctly identified, while adopting a very stringent criteria will dramatically decrease the study power. Therefore, another class of methods are proposed based on the concept of the false discovery rate, which is the expected portion of false positives among all significant hypotheses. For instance, Storey and Tibshirani (2003) proposed to estimate a q-value for each individual metabolite, which is the expected portion of false positives incurred when this metabolite and all the other metabolites with stronger evidence (e.g., smaller p-value) are declared significant. This approach is frequently used in our studies.

Deciphering the relationship among drug exposure signatures

The correlation structure among the change of metabolite concentration could potentially point to altered metabolic pathway activates. For each pair of metabolites, one can calculate the parametric Pearson's correlation coefficient or the nonparametric Spearman's correlation coefficient. Based on pairwise correlations, signatures can be further clustered according to some clustering algorithm. The modulated modularity clustering (MMC, Stone and Ayroles, 2009) is frequently used in our studies. This algorithm is built on pairwise correlations among metabolites. It seeks clustering of metabolites that maximizes a modularity objective function, which encourages strong correlations within clusters while penalizes correlations between clusters. The algorithm searches through different numbers of clusters and ways of assigning metabolites to clusters, and determines the optimal combination. This feature makes MMC different from many other hierarchical or partitioned clustering approaches, which depend on either a pre-specified number of clusters or an external criteria to choose the optimal number of clusters (Stone and Ayroles, 2009). Figure 2.2 illustrates one example of a correlation heatmap of drug exposure signatures which are ordered using the resulting clustering generated by MMC.

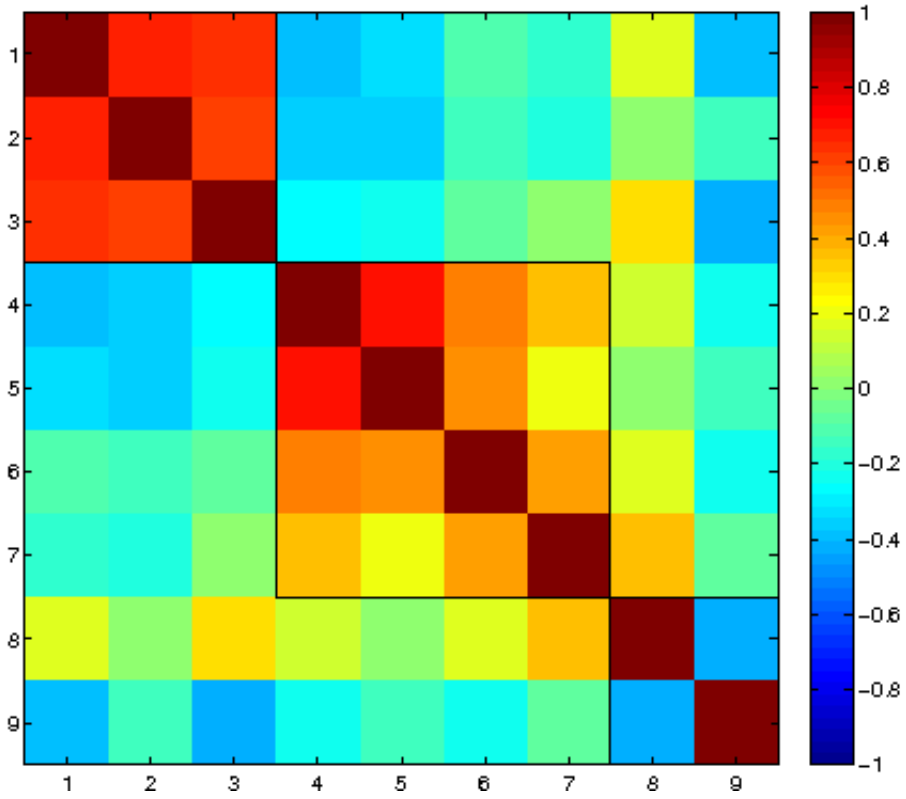

Figure 2.2: An example clustering of drug exposure signatures generated by MMC (Stone and Ayroles, 2009). Strength and direction of correlations are color coded.

Identification of differences in drug exposure signatures between sub-populations

Drug exposure signatures can be quite different among sub-populations of patients, which are defined by one or more clinical variables, e.g., male and females, or African Americans and Caucasians. One subpopulation can be extraordinarily sensitive to a specific therapeutic treatment, while another is not. Differences in drug exposure signatures between two sub-populations can be tested by a parametric two sample t-test, or a nonparametric Wilcoxon rank sum test. If there are more than two sub-populations under study, a parametric one-way analysis of variance (ANOVA) or a nonparametric Kruskal-Wallis ANOVA can be used.

Clusters of drug exposure signatures can be also obtained for each of the sub-population using the same approaches mentioned in the previous section. Comparing resulting clustering of sub-populations may shed light on their differences in metabolic response pathways to therapeutic treatment.

Detection of drug exposure signatures of other types

Drug exposure signatures are not limited to the change of absolute concentration of individual metabolites. On a metabolic pathway, the product-to-precursor ratios reflect enzyme activities (Yao *et al.*, 2010a,b). Therefore, change of such ratios after therapeutic treatment may reflect a change of enzyme activities. From increased product-to-precursor ratios, a possible interpretation is that the reaction may be up-regulated by the treatment. Since such ratios can be obtained for every subject, most of the analyses described above are also applicable here.

Correlation among metabolites can be another indicator of metabolic pathway ac-

tivity. Therefore, correlation comparison before and after drug treatment is another possible way of inspecting drug effect to patients' metabolism. Since correlations under comparison are derived from the same cohort of subjects, their non-independence must be properly adjusted. The Pearson-Filon test (Raghunathan *et al.*, 1996), based on Fisher's Z transformation, can be used for this purpose. Note that, since only one correlation can be obtained per pair of variables per cohort of subjects, there is a lack of straightforward methods to study the inter-relationship among the change of correlations, or to compare such inter-relationships between sub-populations of patients.

2.2.4 Baseline metabotype related to drug response phenotype

Detection of drug response predictive biomarkers

One of the most important objectives of pharmacomatabolomics studies is to determine how the baseline metabotype can be used to predict drug response phenotype. A number of statistical studies (e.g., Fan and Lv (2008)) published recently demonstrated the importance and the effectiveness of marginal tests, especially in an initial analysis stage when the number of candidate metabolites is large. Therefore, for this specific aim, one can begin with a univariate analysis as an exploratory step. When good and poor drug responders have been defined, a parametric two sample t-test, or a nonparametric Wilcoxon rank sum test, can be used to test significant differences on metabolites between drug response groups. A parametric Pearson's correlation or a nonparametric Spearman's correlation can be used to associate metabolites at baseline with a continuous drug response phenotype. Again, both p-value and q-value (Storey and Tibshirani, 2003) should be reported as measures of test significance. Non-linear marginal effect of metabolites can be inspected through the scatter plots of each metabolite versus response. Smoothing

techniques can be used to model a non-linear relationship (Hastie *et al.*, 2009).

Similar to drug exposure signatures, drug response predictive biomarkers can be different among sub-populations defined by one or more clinical variables. Incidentally, we observed cases where correlations between metabolites and drug response phenotype are in the opposite direction between two sub-populations. The various signals can negate the other if the sub-populations are merged and their difference is not properly treated. The difference among sub-populations in effect of metabolites to drug response can be determined either from prior knowledge or from data analysis. Specifically, a regression model with interaction terms can be built:

$$Y_i = \mu + \beta^m \times M_i + \sum_{j=1}^{p'} \beta_j^c \times C_{ij} + \sum_{j=1}^{p'} \beta_j^{mc} M_i \times C_{ij} + \varepsilon_i,$$

where p' is the number of clinical variables of interest, M represents the p-dimensional predictor of metabolites, M_i is the i-th metabolite, and C's represent clinical variables. A significant interaction term indicates a significant difference in metabolite effect among sub-populations defined by the corresponding clinical variable.

We also find it very informative to present the pairwise correlation structure among metabolites and between metabolites and drug response variables in a heat-map. In such a heat-map, the order of metabolites can be arranged either according to their pathway relationship and distance, or according to some cluster algorithm (e.g., Stone and Ayroles (2009), the same clustering algorithm used in Section 2.2.3) that groups highly correlated metabolites. In the former case, one can visualize whether metabolites in the same pathway tends to be more correlated in the study sample and are related to treatment response similarly. In the latter case, one can identify modular structure of metabolites that are highly correlated with each other and with the response phenotypes.

It is very illuminating to examine what metabolites are in those significant modulars and why they are in those modulars. In Section 2.3, we provide an example to illustrate the advantage of this type of data analysis and presentation approach.

Ratios between metabolites have been discussed in Section 2.2.3 as another type of drug exposure signatures. Here, we emphasize that these ratios can also be potential drug response predictive biomarkers. These ratios at baseline reflect patients' metabolic activities and regulation before drug treatment, which may influence their response to the drug. The analysis strategies discussed above are also meaningful when applied to ratios between products and precursors in a metabolic pathway.

We have also mentioned that differences in correlations among metabolites in different physiological states may indicate differences in metabolic regulation. Therefore, in order to further identify differences in metabolic regulation that are associated with variations in drug response, the strength of pairwise correlations among metabolites can be associated with variation in drug response. If good and poor responders have been defined, pairwise correlations among metabolites can be obtained for each drug response group. Then A two-sample Z-test after hyperbolic tangent transformation can be used to test difference in correlations that are obtained from the two independent cohorts of subjects. If a continuous drug response phenotype is used, the strength of pairwise correlations among metabolites can be associated with it using the liquid association (LA) theory (Li, 2002). Instead of regarding a correlation between two variables (e.g., two metabolites) as a constant value, LA treats it as a variable, which may change consistently with some other variables (e.g., reaction activities under regulation). This analysis can potentially reveal what is unique to the metabolism of good responders to the drug.

Prediction of Drug response

After performing marginal effect analysis, one can proceed to use the entire repertoire of baseline metabolites to predict drug response phenotype. A number of statistical analysis approaches are available for this purpose. One can build a linear regression model using metabolites as predictors and drug response phenotype as the response variable:

$$Y = \mu + \beta_1 M_1 + \beta_2 M_2 + \ldots + \beta_p M_p + \varepsilon.$$

Note that the prevalent OLS for the estimation of regression coefficients will become an ill-conditioned problem when the number of metabolites p is larger than the sample size n, which is frequently observed in metabolomics studies. PLS (Wood, 1966) serves as an alternative to OLS for the poorly or ill-conditioned problems. It is well known that OLS seeks a linear combination of M with the maximum correlation with Y. PLS draws away the estimation from the OLS solution toward directions in the predictor space of larger sample spread (Frank and Friedman, 1993; Garthwaite, 1994):

$$a_k = \arg \max_{\substack{a \in R^{p\times 1} \\ a^\top a = 1 \\ a^\top a_j = 0, j = 1, \ldots, k-1}} \operatorname{var}\{a^\top M\} \operatorname{corr}^2\{Y, a^\top M\}. \tag{2.1}$$

In the estimation procedure, coefficient in the so-called loading vectors $\{a_1, a_2, \ldots, a_d\}$ is obtained through individual univariate regressions. Therefore, the large p small n problem often diminishes. PLS is original designed for continuous response, but it can be also adapted to classification. Boulesteix (2004) has showed a connection between PLS for classification and Fisher's linear discriminant analysis. Alternatively, one can build *off-the-shelf* classification models on the latent PLS variables $\{a_1^\top M, a_2^\top M, \ldots, a_d^\top M\}$, e.g., Nguyen and Rocke (2002a) and Nguyen and Rocke (2002b). Penalized regression is

another alternative solution. This approach imposes an additional penalty term to the square error loss function of OLS:

$$f(\boldsymbol{\beta}) = (\mathbf{Y} - \mu - \mathbf{M}\boldsymbol{\beta})^{\intercal}(\mathbf{Y} - \mu - \mathbf{M}\boldsymbol{\beta}) + g(\boldsymbol{\lambda}, \boldsymbol{\beta}),$$

where $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^{\intercal}$ is a vector of n phenotype observations and $\mathbf{M}$ is a $n \times p$ design matrix for metabolites, $\boldsymbol{\lambda}$ represents penalty coefficienst, and $g(\boldsymbol{\lambda}, \boldsymbol{\beta})$ is the penalty term. Given a fixed $\boldsymbol{\lambda}$, estimation of coefficients $\boldsymbol{\beta}$ is a balance between minimizing the squared residual error and the penalty. Different penalty terms grant further different model properties. The bridge regression (Frank and Friedman, 1993) adopts $\lambda\|\boldsymbol{\beta}\|_k = \lambda \sum_{j=1}^{p} \boldsymbol{\beta}_j^k$, $k \geq 0$, as its penalty term, which has several well-known special cases. The ridge regression uses $\lambda\|\boldsymbol{\beta}\|_2$ as its penalty term and is more robust to multicollinearity among predictors compared to OLS. The so-called "lasso" penalty (Tibshirani, 1996) is $\lambda\|\boldsymbol{\beta}\|_1$. A favorable property of lasso is that it is able to shrink coefficients of trivial variables to 0 and therefore remove them from the model. Elastic net (Zou and Hastie, 2005), which uses a mixture of $\lambda_1\|\boldsymbol{\beta}\|_1$ and $\lambda_2\|\boldsymbol{\beta}\|_2$ as a penalty and achieves both the ability of variable selection and some robustness to multicollinearity. Imposing penalties to generalized linear models (GLM) can deal with both continuous and categorical response variables.

In contrast to linear regression models, random forest (Breiman, 2001) is a non-parametric method, which does not rely on assumptions about the form of relationships between predictors and the response variable. Random forest builds a number of decision trees (regression trees for continuous continuous response and classification trees for binary response). Each decision tree is constructed based on a bootstrap sample from the overall training subjects, and the rest subjects are used to evaluate the prediction error of this tree. At each internal node, the method randomly selects a subset of predic-

tors and determine the best split using only these predictors. The overall prediction is given by the the average response (regression) or majority vote (classification) from all individually trained trees.

Identification of drug response-related pathways

The above data analysis methods can be used to identify metabolites that are most significantly correlated with drug response phenotypes individually or jointly. Chemical reactions among metabolites compose metabolic pathways. Since a pathway has some functional interpretations, it is often the case that we are also interested in identifying metabolic pathways that are significantly associated with treatment response. Observed metabolites can be partitioned into different groups according to their pathway relationships retrieved from a pathway database, e.g., KEGG (Kanehisa and Goto, 2000). This group identification can potentially point to a functional basis for the drug response, and can also enhance the statistical power of biological feature identification. For this purpose, we have developed the following analysis steps to aid us to identify responsible metabolite pathways for pharmacometabolomics studies.

1. First perform statistical analysis to detect metabolites that are most significantly associated with the treatment response.

2. Map the selected metabolites onto the pathway database along with other unselected known metabolites. All pathways are considered.

3. Group the mapped metabolites in identifiable pathway groups based on their relationships in the pathways.

4. Perform a multiple regression analysis between the treatment response and metabolites in those pathways that contain marginally significant metabolites, and do

variable selection within each identified pathway.

Besides treating each pathway independently, we have also developed a new and more systematical pathway selection strategy. The new method is a two-step procedure. In the first step, a nonlinear dimension reduction method is employed for multiple metabolites in each identified pathway. This step essentially condenses the (linear and non linear) complex information from a number of metabolites into one or two summary features for each pathway. In the second step, we build a generalized linear model (GLM) (McCullagh and Nelder, 1989) or a generalized additive model (GAM) (Hastie and Tibshirani, 1990) for all the pathways based on the new condensed pathway summary variables to assess the relative importance of pathways jointly. Chapter 3 describes this method in detail.

Aside from the grouping structure, one can further make use of the network structure within and among metabolic pathways, which can be also obtained from pathway databases. For instance, methods have been proposed to incorporate network structure into regression framework to explain a phenotype of interest, which can potentially identify important subnetworks or functional modules of the network but do not rely on pre-defined pathways. This family of methods are penalized regression models, and can be summarized as follows: introduce a network-constrained penalty term to a classical regression model. A key assumption made by these methods is that variables directly connected in the network tend to behave similarly. One representative example in this class is the network-constrained regularization and variable selection method proposed by Li and Li (2008), which imposes a network-constrained penalty term to lasso introduced in Section 2.2.4 above, and gives the following objective function:

$$f(\boldsymbol{\theta}) = (\mathbf{Y} - \mu - \mathbf{M}\boldsymbol{\theta})^{\intercal}(\mathbf{Y} - \mu - \mathbf{M}\boldsymbol{\theta}) + \lambda_1\|\boldsymbol{\theta}\|_1 + \lambda_2 \sum_{i\sim j} (\frac{\theta_i}{\sqrt{d_i}} - \frac{\theta_j}{\sqrt{d_j}})^2 W_{ij} \qquad (2.2)$$

where λ_1 and λ_2 are two non-negative penalty coefficients, $\sum_{i\sim j}$ denotes a sum over all pairs of neighboring metabolites in the network, d_i is the degree of the i-th metabolite, and W stores weights of network edges. The penalty of this model contains two parts. The lasso penalty maintains model sparseness. The second penalty term encourages smoothness of coefficients on the network: minimizing (2.2) will shrink the weighted difference between coefficients of neighboring metabolites towards 0. In this way, a metabolite with large effect will make it easier for its neighbors to be also selected. Following the same logic, Pan *et al.* (2010) proposed another penalty term, which is a sum of L_k-norm of the two coefficients for each pair of neighboring metabolites in the network.

Use of baseline metabotypes to explain differential drug response among sub-populations

Occasionally, it is observed that the drug response is significantly different among sub-populations defined by one or more clinical variables. In other words, these clinical variables are significantly associated with the drug response; therefore, it is interesting to determine whether the differential drug response between sub-populations can be explained by their differential baseline metabotype. A statistical procedure called mediation analysis can be used for this purpose. For a response variable Y, a predictor X, and a mediator T, the classic Baron and Kenny approach (Baron and Kenny, 1986) provide four criteria to establish mediation effect of T between X and Y:

1. There is a significant relation of X to Y;
2. There is a significant relation of X to T;
3. T is significantly related to Y after adjusting for X;

4. There is a significant decrease (mediation effect) in the coefficient relating X to Y after adjusting for T.

Imai *et al.* (2010) proposed two methods that can test the mediation effects for linear and nonlinear relationships, with parametric and nonparametric models, with continuous and discrete mediators, and with various types of outcome variables. One method is applied only to parametric models, whose result can be viewed as an approximation to the Bayesian posterior distribution of the mediation effect. The other is based on bootstrap and can be applied to semi- and nonparametric models. The first method is more computationally efficient, but, in practice, we usually find the second method takes a reasonable amount of time to finish sufficient number of re-sampling steps. Note that one should be cautious with interpreting results from this analysis. A significant mediation effect cannot assert causality relationship.

2.2.5 Identification of drug induced changes in metabotypes that associate with drug response

The relationship between the drug induced change in metabotypes and response could provide excellent insight into the mechanisms of drug effect. Therefore, another typical scientific question of interest is how a change in metabotype is related to drug response phenotype. Specifically, one can ask similar questions for change of metabolites as we have asked for baseline metabolites:

1. How are changes of individual metabolites correlated with response phenotype?

2. Are these correlations different among sub-populations defined by any clinical variables?

3. Are the ratios and/or correlation patterns among changes of metabolites different between good and poor responders?

4. How well can change of metabotype classify good and poor responders?

5. How are the changes in metabolic pathways related to the drug response phenotype?

6. Can the changes of metabolites mediate the significant association between the drug response and some clinical variables?

All of the questions can be addressed using the same strategies and techniques discussed in Section 2.2.4.

2.2.6 Connecting metabolomics with other omics

Many large scale biological and biomedical studies contain omics data on multiple layers of molecular biology that have been jointly generated for the same cohort of samples. Metabolomics data has been collected along with genomics, transcriptomics, and/or proteomics data (Lindon *et al.*, 2007). In this section, we will briefly discuss methods that can be used to connect the metabolomic biomarkers identified from sections 2.2.4 and 2.2.5 with other omics biomarkers in deeper molecular biology layer, which can potentially facilitate mapping genes and enzymes implicated in mechanisms of variation in response to treatment.

Again, one can always start with a univariate analysis. For instance, Kruskal-Wallis ANOVA can be used to test association between each pair of metabolomic biomarkers and discrete genetic markers; Spearman or Pearson correlations can be tested between continuous biomarkers.

Some well-known biological hypotheses and/or observations can guide us to design multivariate models, which can potentially improve the statistical power and, at the same time, reduce the false discovery rate. For instance, considering the metabolism of specific metabolomic biomarkers may be governed by more than one genetic factors, one can build one multiple regression model for each metabolomic biomarker using all genetic markers as potential predictors and perform variable selection. Furthermore, a genetic variant may have pleiotropic effects on multiple metabolomic biomarkers and metabolomic biomarkers are very likely to be highly correlated. In order to incorporate this information, one can adopt the following two-step analysis procedure. First, a relevance network among metabolomic biomarkers can be built using either some simple strategies, e.g., giving a threshold for pairwise correlations among metabolomic biomarkers, or the more sophisticated Marcov network construction methods (Edwards, 2000; Meinshausen and Buhlmann, 2006; Peng *et al.*, 2009; Yuan and Lin, 2007b). Second, a graph-guided fused lasso models proposed by Kim, Sohn and Xing (2009) can incorporate this network information into a multivariate regression model using all the metabolomic biomarkers as response variables and the genetic biomarkers as explanatory variables. This model belongs to the family of penalized regression. It penalizes the absolute difference between each pair of regression coefficients corresponding to the same predictor for two neighboring response variables in the network. This approach will allow us to identify genetic biomarkers that jointly influence subgroups of highly correlated metabolomic biomarkers, which can potentially represent a compound class or a functional module on the metabolic network. Finally, these multivariate methods can be also used to connect transcriptomic and proteomic biomarkers with metabolomic biomarkers.

2.2.7 Validation of Results

It is important to validate the results after exploratory analysis, especially for multivariate models with the number of predictors exceeding the sample size, which is a typical characteristic of omics studies. Cross validation is frequently used for this purpose. In order to validate a prediction model for drug response phenotype, the prediction accuracy based on a separate test set should be reported, with its sensitivity and specificity, especially for an unbalanced study. Model selection is sometimes desired for multivariate models, either using rigorous statistical procedure or heuristic criteria. A common, but inconspicuous, mistake is the involvement of the test set in any part of the model building procedure, especially when the model selection appears independent from subsequent models to be evaluated. For example, before a standard PLSDA model is constructed, a pre-selection of predictors (metabolites) may be performed based on results from univariate test for each metabolite. It is not uncommon for some researchers to use the whole dataset to perform the univariate tests. In this case, the original training set for each cross-validation step should be split further into a training set and an internal test set for model selection. The selected model should then be evaluated by an independent external test set.

2.3 Example studies

In this section, we will demonstrate examples of pharmacometabolomics studies, in which we have been involved, and where methods described in the above pipeline identified important biological evidence for drug response mechanisms and linked different omics studies.

2.3.1 Pharmacometabolomics of statins in the treatment of hyperlipidemia

Statins are HMG-CoA reductase inhibitors that lower lipids and are widely used to treat cardiovascular disease (CAD). The primary clinical rationale for the use of statins is to reduce the level of low-density lipoprotein cholesterol (LDLC). However, efficacy of statins can vary greatly among individuals, and the biochemical basis for the variation remains poorly understood. Therefore, identification of metabolic biomarkers that are related to drug response would be useful for understanding the drug response mechanisms, and for targeting the therapy to population who will respond to it.

In this study, pre-treatment plasma samples of 100 subjects were analyzed using a targeted metabolomics approach focusing on 26 metabolites in a sterol metabolism pathway. The pathway can be divided into three parts: cholesterol synthesis, dietary sterol absorption, and bile acid formation (Figure 2.3). The drug response of patients was measured by the change in LDLC.

We tested the correlation between metabolites at baseline and drug response (Figures 2.3 and 2.4). Correlations between each pair of metabolites at baseline were also obtained. A correlation heatmap (Figure 2.4) was constructed, where metabolites were listed in the order of the pathway map for sterol metabolism.

The correlation matrix (Figure 2.4) showed a clear block pattern, which is roughly consistent with the pathway structure. All of the metabolites within the cholesterol biosynthetic pathway, except cholestanol (CSTN), were positively correlated with each other but not strongly correlated with drug response. The dietary sterols, sitosterol (SITO), stigmasterol (STIG), and campesterol (CAMP), were highly positively correlated with each other, and were negatively correlated with most cholesterol biosynthesis

metabolites and with bile acids. COPR did not correlate with the other dietary sterols and was negatively correlated with drug response to statins, while SITO, STIG and CAMP were slightly positively correlated with drug response. Most bile acids were positively correlated with each other, and a set of primary and secondary bile acids, including taurocholic acid (TCA), glycocholic acid (GCA), taurochenodeoxycholic acid (TCDCA), glycochenodeoxycholic acid (GCDCA), and glycoursodeoxycholic acid (GUDCA), were positively and significantly correlated with change in LDLC. Additionally, we observed that several bile acids showed significant associations with plasma concentration of simvastatin and one SNP in a gene whose product transports simvastatin from plasma into the liver (results not shown). Since most of these bile acids are either direct products, or closely related to products of intestinal microbiome, the overall study suggests a role for gut-derived factors in modulating statin efficacy.

2.3.2 Pharmacometabolomics-guided pharmacogenomics of selective serotonin reuptake inhibitors(SSRIs) in the treatment of depression

In this section, we exemplify the part of pathway analysis in our pipeline with a pharmacometabolomics-guided pharmacogenomics study of interpersonal difference in SSRI response (Ji *et al.*, 2011).

SSRIs are an important class of drugs used to treat major depressive disorder (MDD), a common psychiatric disease. However, fewer than half of depressed patients have a sustained response to therapy with these agents. Therefore, there is an urgent need for validated biomarkers that predict treatment responses prior to therapeutic treatment, and reveal drug response mechanisms. Many pharmacogenetic studies of antidepressant

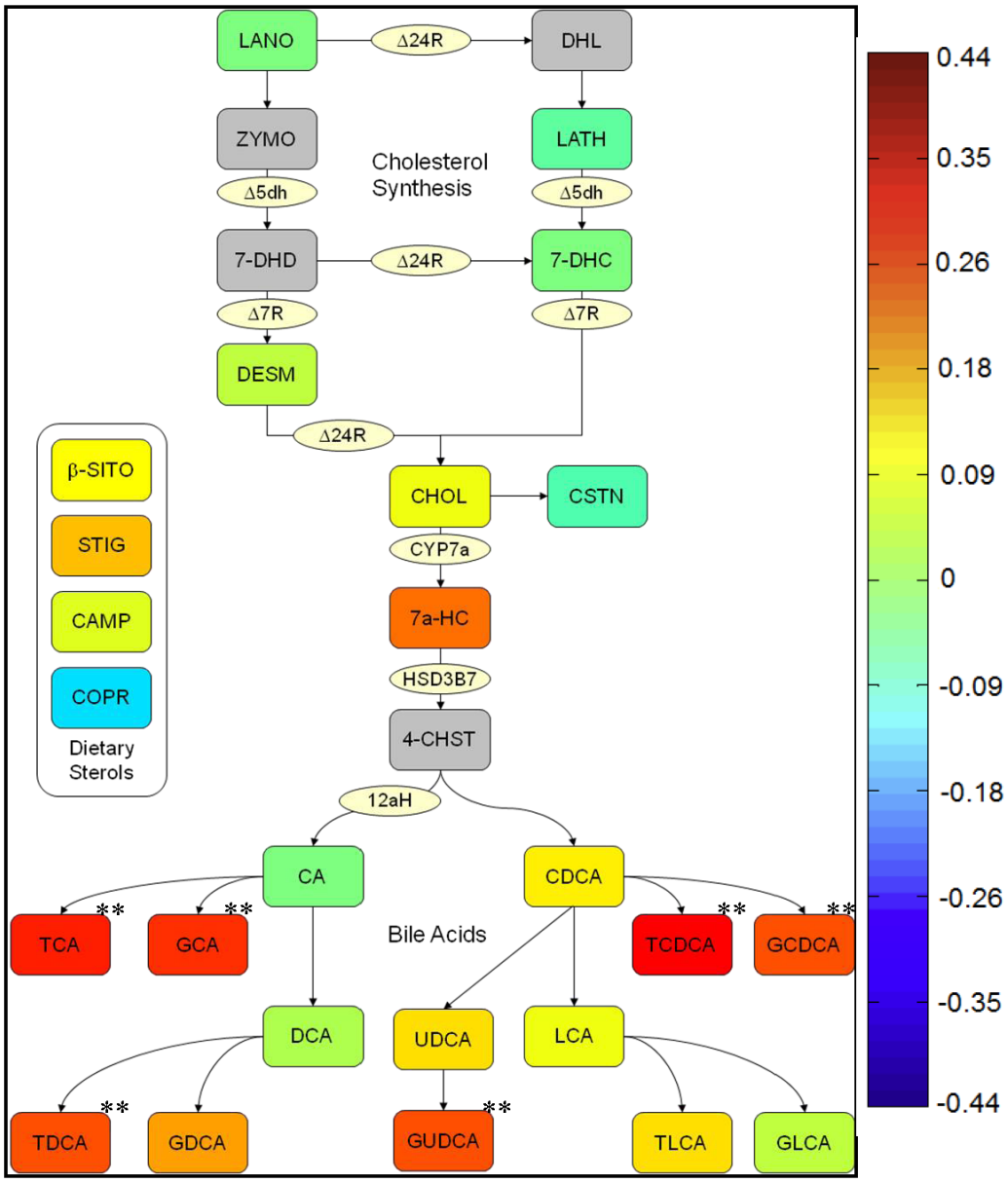

Figure 2.3: Sterol metabolism pathway showing the association of metabolites at baseline with change of LDLC by statin treatment. The color scheme corresponds to correlation strength as shown by the color bar. Red: Better response, more reduction or less increase of the metabolite. Blue: Better response, less reduction or more increase of the metabolite.

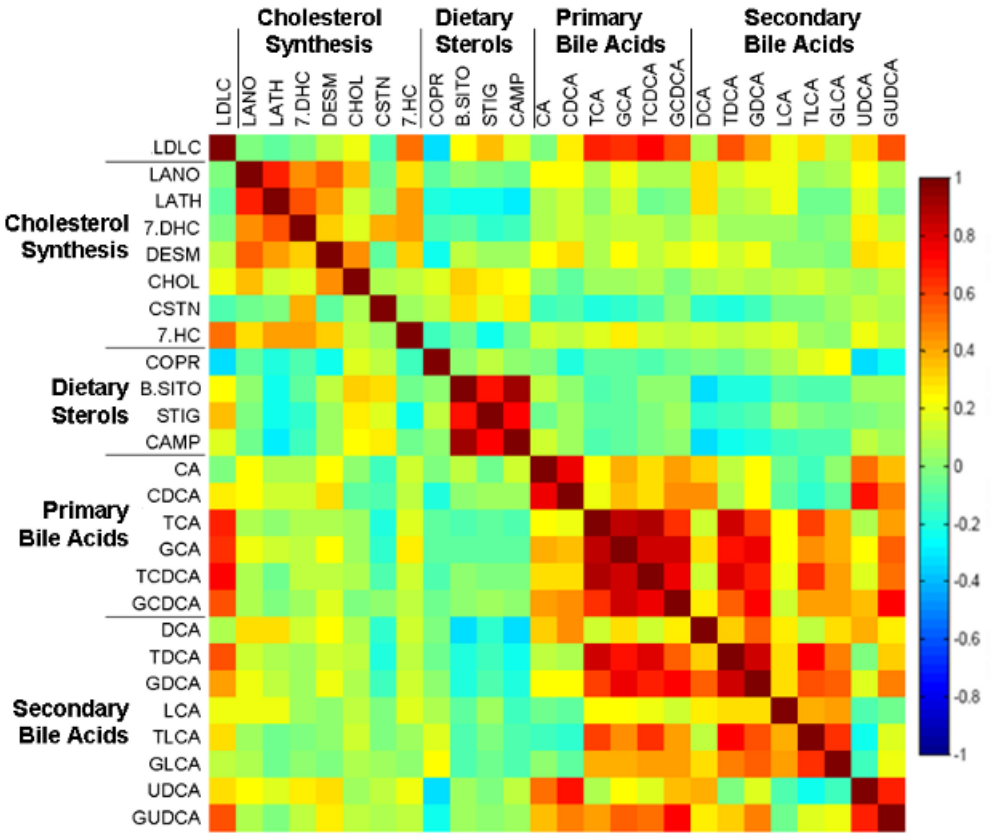

Figure 2.4: Correlation heatmap of metabolites at baseline and change of LDLC. Correlations to change of LDLC are given in the first row and column, and are rescaled (divided by the largest absolute value of them) to be clearer in the map. The color scheme corresponds to correlation strength, as shown by the color bar. In the first row and column: Red: Better response, more reduction or less increase of the metabolite; Blue: Better response, less reduction or more increase of the metabolite. In the rest rows and columns: Red: positive correlation between two metabolites; Blue: negative correlation between two metabolites.

drugs, particularly SSRIs, have been performed. Those studies have generally focused on polymorphisms in several candidate genes, including those encoding the serotonin transporter; a variety of serotonin receptors; enzymes involved in serotonin biosynthesis; and drug metabolizing enzymes (Kato and Serretti , 2008; Mrazek *et al.*, 2009; Serretti and Artioli, 2004). However, these candidate gene-based studies, and even some recently published genome-wide association studies (GWAS), have failed to provide reliable biomarkers for SSRI treatment outcome (Hamilton *et al.*, 2010; Holsboer *et al.*, 2009; Uher *et al.*, 2010). We proposed to first identify metabolomic biomarkers which then guide the identification of genetic biomarkers.

Specifically, plasma samples were obtained prior to treatment from 20 MDD Escitalopram (an antidepressant of the SSRI class) *remitters* and 20 *non-remitters*. The samples were assayed by a MS-based metabolomic platform, and 251 metabolites were quantified and composed a metabotype. The drug response was evaluated by the percentage change in patients' Quick Inventory of Depressive Symptomatology (QIDS) score after the drug treatment. QIDS is a highly validated scoring system that is used to evaluate the severity of depression. To associate patients' baseline metabotype with their drug response, we first performed a univariate correlation analysis among metabolites at the baseline and the drug response. However, we are aware of the fact that the significant marginal associations do not necessarily point to specific metabolic pathways. We then performed the pathway-specific regression analysis outlined in the first part of Section 2.2.4 to identify significant pathways. Along with several other pathways, the nitrogen metabolism pathway scored as the most significant one from this pathway analysis. Among all the six metabolites that we can map from our dataset to the nitrogen metabolism pathway, five are jointly significantly associated with the drug response (Table 2.1). The regression model has an adjusted $R^2 = 0.55$. The results reflect some antagonistic correlations

among the metabolites in the nitrogen pathway with respect to the drug response, which may explains why most simple (or marginal) regression coefficients of the metabolites to the drug response are not as highly significant as the partial regression coefficients. This also demonstrates the importance of our pathway-specific analysis approach in the identification of responsible pathway.

Table 2.1: Simple and partial regression coefficients of nitrogen metabolites on the Citalopram/Ecitalopram response.

Compound	Simple regression (p-value)	Partial regression (p-value)
Glycine	2.71×10^{-5} (0.0054)	3.68×10^{-5} (<0.0001)
Glutamic acid	2.44×10^{-4} (0.22)	4.54×10^{-4} (0.0088)
Aspartic acid	-2.29×10^{-3} (0.069)	-2.41×10^{-3} (0.029)
Asparagine	-9.46×10^{-5} (0.70)	-6.04×10^{-4} (0.0057)
Hydrozylamine	-1.33×10^{-4} (0.0022)	-1.02×10^{-4} (0.0035)

Among the five metabolites in table 2.1, glycine is the most significant one. Besides, glycine has been known as an inhibitory neurotransmitter as well as a key metabolite in the Folate Cycle. Therefore, we focused on glycine as a candidate metabolite. We then tested the hypothesis that DNA sequence variation in genes encoding glycine metabolism enzymes might be associated with the drug response. The biosynthesis and metabolic degradation of glycine involved several enzymes. Glycine is synthesized from serine in a reaction catalyzed by two serine hydroxymethyltransferase (SHMT) isoforms (McNeil *et al.*, 1994; Turner *et al.*, 1992), and it is degraded by a multi-enzyme "glycine cleavage system". This glycine cleavage system includes aminomethyltransferase (AMT), dihydrolipoamide dehydrogenase (DLD), glycine cleavage system protein H (GCSH), and

glycine dehydrogenase (GLDC) (Ichinohe *et al.*, 2004; Kikuchi *et al.*, 2008). We therefore selected 135 tag SNPs for these genes for an association analysis using DNA from 512 subjects enrolled in the Mayo Clinic-NIH Pharmacogenetics Research Network Citalopram/Escitalopram Pharmacogenomic study, the same study population from which the subjects for metabolomic profiling had been selected. Among the 135 SNPs analyzed, the top 4 SNPs ($p < 0.05$) that were associated with the remission status were all within the GLDC gene on chromosome 9, and the rs10975641 SNP was most significantly associated with remission status (p = 0.008). Rs10975641 was also associated with the continuous percentage change of QIDS score (p = 0.006). This finding was then validated using DNA samples from the large NIMH-sponsored Sequenced Treatment Alternatives to Relieve Depression study. In summary, our pathway analysis identified glycine as a potential candidate compound, which led to the identification of a common SNP that was associated with SSRI treatment outcome in two large SSRI pharmacogenomic studies. This study also provides proof of principle for a pharmacometabolomics-informed pharmacogenomic approach that might be applied in future biomarker discovery studies.

2.4 Discussion

In this chapter, we proposed a comprehensive data analysis pipeline for pharmacometabolomics studies. We summarized several common specific aims of metabolomics studies of individual variation in therapeutic response, and provided a variety of numerical approaches and procedures that address these aims from different angles. Most of the approaches reviewed in this chapter have been implemented in different packages of R, a free software environment for statistical computing. MMC is implemented in MATLAB, and the authors also provide a web server for MMC (http://mmc.gnets.ncsu.edu/).

Though we attempted to be comprehensive in this chapter, the pipeline only covers one typical study design: all subjects are patients that take one drug, and for each subject, two samples were collected before and after drug treatment. There exist many other study designs in the field of metabolomics. For instance, we have been involved in studies of diseases where healthy controls are also recruited, and studies of drug response where patients are treated either by the drug or placebo. Additionally, when data on more than two types of drugs or treatments for the same symptom are collected, one additional specific aim is to learn similarities and differences in their mechanisms. Finally, we have also observed longitudinal metabolomics data collected at tens of time points to investigate the dynamics of metabolic system (Jansen *et al.*, 2004). Analysis strategies should also be developed for these study designs with new research objectives.

In another aspect, despite that most of the adopted approaches in our pipeline have been implemented under one or even several statistical or mathematical programming environments, there is still a lack of a freely accessible and easy-to-use software that connects them in one pathway. Considering the steady blooming of metabolomics, it is worthwhile to turn our pipeline into such a product.

Chapter 3

Nonlinear Sufficient Dimension Reduction for Biological Pathway Selection

3.1 Introduction

In biomedical research, high-throughput technologies are generating massive amounts of high-dimensional data. A representative example is cDNA microarray, which simultaneously measures expressions of thousands or tens of thousands of genes in a single experiment. These fast developing techniques have created a research area called the "omics" studies, which aim at examining the whole spectrum of basic biological units such as genes, proteins and metabolites. Examples include genomics that studies the whole genome, i.e., the sum of all genes of an individual organism, proteomics that studies proteome, i.e., the entire complement of proteins produced in an organism, specific type of tissues or other biological system, and metabolomics that studies metabolome,

i.e., the entire repertoire of metabolites present in cells and/or tissues, which represents the final product of certain cellular processes. An important question in such studies is to identify biological units, i.e., genes, proteins or metabolites, that are related to and influence various clinically relevant phenotypes, for instance, survival outcome from cancer treatment, or risk of developing cancers. Given the very high dimensionality and often the limited number of sample units of such omics data, new challenges arise for conventional statistical analysis of the data. As a consequence, it has motivated developments of many new statistical methods; see Bickel *et al.* (2009) for an excellent review and the references therein.

Most existing approaches treat genes, proteins or metabolites individually, and as such ignore any prior biological knowledge that characterizes inter-relationship among those biological units. It is commonly believed that most biological systems operate not based on just a single unit, but instead on their mutual interactions and regulations. Attesting to this belief, recent studies have suggested that clinical outcomes of complex diseases are often associated with multiple genes rather than a single one. This has led researchers to form and define clusters of genes, proteins and metabolites, referred to as "pathways", that are believed to function in a coordinated and interactive fashion. Years of intensive biomedical research has accumulated an immense wealth of pathway knowledge, which is primarily available through some well-known pathway databases, for instance, Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), Reactome (Matthews *et al.*, 2008), and BioCyc (Karp *et al.*, 2005). Intuitively, employment of such biological pathway knowledge would greatly facilitate our statistical analysis, because it offers a means to reduce the potentially enormous modeling space by focusing on biologically meaningful interactions among genes, proteins or metabolites in related pathways. In addition, the outcomes are expected to be more readily interpretable

biologically when the pathway information is taken into account.

In the light of this observation, there have been some recent developments of pathway-based statistical analysis that search for pathways that are associated with disease phenotypes. Early efforts have focused on gene set enrichment analysis, i.e., to identify pathways that are over-represented by differentially expressed genes (Subramanian *et al.*, 2005), which provides a useful and informative tool (Tian *et al.*, 2005). However, such analysis treats each pathway separately without accounting for possible interactions between pathways. Wei and Li (2007) proposed a nonparametric pathway based regression (NPR) model that considers multiple pathways simultaneously and allows complex interactions among genes within the pathways. To fit their NPR model, they employed a gradient descent boosting algorithm and used classification and regression tree as the base learner. However, the algorithm only produces a ranking based on relative importance of the pathways rather than an explicit selection of relevant pathways. Luan and Li (2008) considered a group additive regression (GAR) model along with a group gradient descent boosting algorithm to identify pathways related to clinical outcomes. But again, their procedure only gives a pathway ranking. Besides, the model only permits linear combinations of genes within the pathways since their choice of the base learner is a linear model. More recently, Ma and Kosorok (2009) used principal components analysis (PCA) as the first step to combine genes in the same pathways into a small number of summary features. They then built statistical models associating the induced summary features with the phenotypes. It is well known, however, that PCA is an unsupervised dimension reduction technique that does not take into account the response information. For this reason, as we will show later, PCA could yield inferior results.

In this chapter, we propose a method of nonlinear dimension reduction followed by a generalized additive model (GAM) or a generalized linear model (GLM) with L_1 regular-

ization for the purpose of selecting informative pathways that are related to the phenotype of interest. In the first step, we couple a reproducing kernel map with a family of sufficient dimension reduction estimators to produce summary features of constituent units for each pathway. This step turns the high-dimensional data into its low-dimensional projections. Due to flexibility of kernels and their computational advantages, the new dimension reduction method allows complex interactive relations among biological units such as genes, and can also handle a very large number of units, even if it far exceeds the number of observations. In addition, the proposed method does not impose any strong parametric model assumptions in the phase of dimension reduction, and as such it grants full flexibility in subsequent model formulation and selection. Our method is motivated by kernel sliced inverse regression first proposed by Wu (2008) and Wu *et al.* (2008), but we extend their work to a whole family of dimension reduction methods. We also derive a generalized cross validation criterion for the regularized estimation. In the second step, we build a generalized additive model or a generalized linear model based on the induced summary features from all the pathways. We produce both a ranking of pathways by employing an L_1 regularization, as well as an explicit pathway selection by proposing a pseudo pathway selection strategy. Compared with Ma and Kosorok (2009), the main difference lies in the step of dimension reduction. Our method extends PCA in the sense that it takes into account the response information during the reduction, and it permits complex nonlinear associations of genes when producing the summary features. Compared with Wei and Li (2007) and Luan and Li (2008), our method produces both pathway ranking and selection. Our simulations studies confirm the advantages of our proposal over those existing solutions. An analysis of a glioblastoma microarray data identifies four informative pathways that have good evidences in the biology literature to support possible associations with the disease.

3.2 Nonlinear dimension reduction

3.2.1 Linear dimension reduction

For a regression of a response Y given a p-dimensional predictor X, sufficient dimension reduction (SDR) seeks a minimum number of linear combinations, $\eta_1^\intercal X, \ldots, \eta_d^\intercal X$, such that

$$Y \perp\!\!\!\perp X \mid (\eta_1^\intercal X, \ldots, \eta_d^\intercal X). \tag{3.1}$$

The space spanned by $\eta = (\eta_1, \ldots, \eta_d)$, called the central subspace and denoted as $\mathcal{S}_{Y|X}$, uniquely exists under minor conditions (Cook, 1996). Given (3.1), we can replace the originally p-dimensional X with now d-dimensional $\eta^\intercal X$. In practice, d is often much smaller than p, and thus dimension reduction is achieved. For ease of reference, we call $(\eta_1^\intercal X, \ldots, \eta_d^\intercal X)$ the sufficient predictors, which will serve as the induced summary features in subsequent modeling.

There have been many methods proposed to estimate $\mathcal{S}_{Y|X}$, most of which can be formulated as a generalized eigen decomposition problem. Specifically, an estimate of a basis of $\mathcal{S}_{Y|X}$ can be obtained by the first d eigenvectors η_j's that correspond to the nonzero eigenvalues λ_j's in a descending order from the decomposition,

$$\Omega_x \eta_j = \lambda_j \Sigma_x \eta_j, \quad j = 1, \ldots, d, \tag{3.2}$$

where $\Sigma_x = \mathrm{Cov}(X)$, and Ω_x is a method-specific $p \times p$ semi-positive definite matrix. Some representative estimators include sliced inverse regression(SIR) (Li, 1991), where $\Omega_x = \mathrm{Cov}\{E(X|Y) - E(X)\}$; sliced average variance estimation (SAVE) (Cook and Weisberg,

1991), where $\Omega_x = E[\{\Sigma_x - \mathrm{Cov}(X - E(X)|Y)\}^2]$; and directional regression(DIR) (Li and Wang, 2007), where $\Omega_x = 2E[\{\Sigma_x - \mathrm{Cov}(X - E(X)|Y)\}^2] + 2\mathrm{Cov}^2\{E(X|Y) - E(X)\} + 2E[\{E(X|Y) - E(X)\}^\tau\{E(X|Y) - E(X)\}]\mathrm{Cov}\{E(X|Y) - E(X)\}$. All those methods involve the inverse moments $E(X|Y)$ and $\mathrm{Cov}(X|Y)$. To estimate those quantities, we often first partition the sample space of Y into H non-overlapping intervals, or say slices, and then obtain the sample average or sample covariance within each slice. It is also interesting to note that, most of those SDR methods impose no parametric assumption on $Y|X$. Instead they require the marginal distribution of X to satisfy that $E(X|\eta^\tau X)$ is linear in $\eta^\tau X$. This is often viewed as a mild condition, since it holds when X is elliptically symmetric and is approximately true when p goes to infinity. In this chapter we assume the condition holds approximately since we are dealing with a very large p.

SDR methods following (3.1) and (3.2) yield *linear* dimension reduction, because the reduction admits the form of linear combinations of X. This could have some limitations. Consider an illustrative example, where $X = (X_1, \ldots, X_6)$ and

$$Y = X_1 + X_2X_3 + X_4^2 + X_5X_6 + \varepsilon, \tag{3.3}$$

with an independent error ε. Then $\mathcal{S}_{Y|X} = \mathbb{R}^6$, with no reduction in dimension possible. Another limitation associated with the SDR methods in (3.2) is that one needs to invert a $p \times p$ covariance matrix Σ_x. When the number of predictors p exceeds the sample size n, one can not invert the sample estimator of Σ_x. To address this $n < p$ issue, there have been proposals employing the ridge regression idea (Li and Yin, 2008) or the partial least squares idea (Cook *et al.*, 2007; Li *et al.*, 2007). However, the proposed remedies are very computationally intensive when p becomes very large. Next we consider a *nonlinear* dimension reduction strategy to address those limitations.

3.2.2 Nonlinear dimension reduction using kernel methods

In the usual kernel-based methods, a set of features are chosen that define a space $\mathcal{F}$, where it is hoped relevant structure will be revealed. The data $(x_1, \ldots, x_n)$ in the input space $\mathcal{X}$ are then mapped to the feature space $\mathcal{F}$ using a mapping $\phi : \mathcal{X} \to \mathcal{F}$, and classification, regression, or clustering is performed in $\mathcal{F}$ using traditional methods. If $\mathcal{F}$ is chosen to be an inner product space and if one defines the kernel function k, with the associated Gram matrix $K \in \mathbb{R}^{n \times n}$ as $K_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, then any algorithm whose operations can be expressed in terms of inner products in the input space can be generalized to an algorithm which operates in the feature space by substituting a kernel function for the inner product. Using the kernel k instead of an inner product in the input space corresponds to mapping the data into a high-dimensional inner product space $\mathcal{F}$ by a usually nonlinear mapping ϕ, and taking inner products there, i.e., $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. Common choices of kernels include the polynomial kernel $k(x, x') = (1 + x^\top x')^r$ for some positive integer r, and the Gaussian kernel $k(x, x') = \exp\{-\|x - x'\|^2/(2\sigma^2)\}$.

In our context of nonlinear dimension reduction, the basic idea is to carry out a linear dimension reduction in the space of $\phi(X)$, which in effect results in a nonlinear dimension reduction in the original input space of X. The well-known kernel trick turns the primal problem that depends on the dimension of $\phi(X)$, which is high or even infinite, to a dual problem that only depends on the sample size. As such, the method works for any p regardless of n.

More specifically, in analogy to linear reduction in (3.1), nonlinear dimension reduction seeks

$$Y \perp\!\!\!\perp X \mid \big(\langle \beta_1, \phi(X) \rangle, \ldots, \langle \beta_{\tilde{d}}, \phi(X) \rangle\big). \tag{3.4}$$

Comparing with (3.1), the linear combinations $(\eta_1^\intercal X, \ldots, \eta_d^\intercal X)$ are replaced by $\tilde{d}$ inner products $(\langle \beta_1, \phi(X)\rangle, \ldots, \langle \beta_{\tilde{d}}, \phi(X)\rangle)$, and Y depends on X only through those inner products. We again refer them as the sufficient predictors, and assume $\tilde{d} \leq \min(n, p)$.

In terms of estimation, conceptually, one can estimate β's in a way analogous to (3.2), i.e., through the eigen decomposition

$$\Omega_\phi \beta_j = \rho_j \Sigma_\phi \beta_j, \quad j = 1, \ldots, \tilde{d}, \tag{3.5}$$

where $\Sigma_\phi = \mathrm{Cov}\{\phi(X)\}$, and Ω_ϕ is defined similarly as Ω_x except we replace X with $\phi(X)$. See Wu *et al.* (2008) for a theoretical justification of (3.5) for sliced inverse regression, while similar arguments apply to other SDR estimators. Given $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, estimation of β_j's are obtained by replacing Ω_ϕ and Σ_ϕ in (3.4) with their corresponding sample counterparts $\hat{\Omega}_\phi$ and $\hat{\Sigma}_\phi$, i.e.,

$$\hat{\Omega}_\phi \beta_j = \rho_j \hat{\Sigma}_\phi \beta_j, \quad j = 1, \ldots, \tilde{d}, \tag{3.6}$$

On the other hand, depending on the choice of the kernel function k, the dimension of the induced mapping $\phi(X)$ can be very high, sometimes even infinite. As such, a direct decomposition through (3.6) is not feasible computationally.

The problem can be solved by noting that, the target of nonlinear dimension reduction estimation are the inner products $\langle \beta_j, \phi(X)\rangle$, rather than β_j themselves. These inner products can be obtained by solving a dual problem to (3.6),

$$\tilde{K} J \tilde{K} \alpha_j = \rho_j \tilde{K}^2 \alpha_j, \quad j = 1, \ldots, \tilde{d}, \tag{3.7}$$

where $\tilde{K} \in \mathbb{R}^{n \times n}$ is the centered Gram matrix with the (i, j)-th element $\tilde{K}_{ij} = K_{ij} -$

$K_{i\cdot} - K_{\cdot j} + K_{\cdot\cdot}$, $K_{\cdot\cdot}$ is the mean of K and $K_{i\cdot}$ and $K_{\cdot j}$ are the means of the i-th row and j-th column of K. J is a method-specific $n \times n$ matrix that takes the form,

$$\begin{aligned}
J_{\mathsf{SIR}} &= \sum_{h=1}^{H} n_h^{-1} b_h b_h^{\intercal}, \\
J_{\mathsf{SAVE}} &= \sum_{h=1}^{H} B_h \tilde{K} B_h, \quad B_h = n^{-1} n_h^{1/2} I_n - n^{-2} n_h^{1/2} \mathbf{1}_n \mathbf{1}_n^{\intercal} - n_h^{-1/2} \mathrm{diag}(b_h) + n_h^{-3/2} b_h b_h^{\intercal}, \\
J_{\mathsf{DIR}} &= 2J_{\mathsf{SAVE}} + 2n^{-1} J_{\mathsf{SIR}} \tilde{K} J_{\mathsf{SIR}} + 2n^{-1} \left(\sum_{h=1}^{H} n_h^{-1} b_h^{\intercal} \tilde{K} b_h \right) J_{\mathsf{SIR}},
\end{aligned}$$

where $\mathbf{1}_n$ is an $n \times 1$ vector with all elements equal to 1, b_h is an $n \times 1$ vector with its i-th element equal to 1 if y_i belongs to the h-th slice and 0 otherwise, n_h is the number of observations in the h-th slice, and I_n is an n-dimensional identity matrix. Then for a new observation $x \in \mathcal{X}$,

$$\langle \beta_j, \phi(x) \rangle = \alpha_j^{\intercal} [\tilde{k}(x_1, x), \ldots, \tilde{k}(x_n, x)]^{\intercal} \tag{3.8}$$

where $\tilde{k}(x_i, x) = k(x_i, x) - n^{-1} \sum_{l=1}^{n} k(x_l, x), i = 1, \ldots, n$. So the inner product $\langle \beta_j, \phi(x) \rangle$ can be obtained from the kernel k and α_j's from (3.7). An outline of derivation of the above kernel version of SDR estimators is given in the Appendix.

Furthermore, to induce numerical stability to the eigen decomposition (3.7), we follow Wu *et al.* (2008) to introduce a ridge regularization,

$$\tilde{K} J \tilde{K} \alpha_j = \rho_j (\tilde{K}^2 + n^2 \tau I_n) \alpha_j, \quad j = 1, \ldots, \tilde{d}, \tag{3.9}$$

where $\tau \geq 0$ is a ridge parameter. We will discuss tuning of τ in the next section.

Due to (3.7), the proposed nonlinear dimension reduction only involves decomposition

of an $n \times n$ matrix, so it can handle $n < p$. Its flexible reduction form beyond the linear combination is also expected to facilitate dimension reduction. As a simple illustration, we re-consider the example (3.3). If one employs a quadratic kernel, then only one linear combination in the mapped feature space is needed to summarize all regression information, and thus substantial reduction is achieved. Finally, if a linear kernel is employed, the proposed method reduces to the usual linear dimension reduction.

3.2.3 Tuning

There are two tuning parameters in the aforementioned nonlinear dimension reduction procedure: one is the number of slices H, and the other is the ridge parameter τ.

We first consider a variant of dimension reduction estimator of Zhu *et al.* (2010), which allows one to avoid the tuning parameter H. The basic idea is to always dichotomize the sample space of Y by the observed y_i, $i = 1, \ldots, n$, and then substitute the average of Ω_x for all such binary partitions in (3.2). In our context of nonlinear dimension reduction, the solution continues to admit the eigen decomposition of (3.9) except that the J matrix now takes the form

$$J_{\mathsf{APME}} = n^{-1} \sum_{i=1}^{n} J_i, \tag{3.10}$$

where J_i is the J matrix derived by performing nonlinear dimension reduction for the i-th dichotomization. Also note that (3.10) is applicable for all the aforementioned kernel SDR estimators.

Next we develop a generalized cross validation (GCV) criterion to choose the ridge parameter τ. Consider the eigen decomposition $n^{-1}\tilde{K}J\tilde{K} = \sum_{j=1}^{m} u_j v_j v_j^{\intercal}$, where $u_1 \geq \ldots \geq u_p$ are the eigenvalues in descending order, $v_1, \ldots, v_p$ are the correspond-

ing eigenvectors, and m is the number of nonzero eigenvalues. Then following Li and Yin (2008), the solution α_j from the eigen decomposition (3.9) can be equivalently obtained by minimizing the objective function,

$$L(A,C) = \sum_{j=1}^{m} u_j \|v_j - n^{-1}\tilde{K}^2 A c_j\|^2 + n\tau \mathrm{vec}(AC)^\intercal \{U \otimes (n^{-1}\tilde{K}^2)\}\mathrm{vec}(AC),$$

where $A \in \mathbb{R}^{n\times \tilde{d}}$ and $C = (c_1, \ldots, c_m) \in \mathbb{R}^{\tilde{d}\times m}$, $U = \mathrm{diag}(u_1, \ldots, u_m)$, $\otimes$ stands for the kronecker product, and $\mathrm{vec}(\cdot)$ is a matrix operator that stacks all columns of a matrix into a vector. Letting $(\hat{A}, \hat{C}) = \arg\min L(A,C)$, then $\mathrm{span}(\hat{A}) = \mathrm{span}(\alpha_1, \ldots, \alpha_{\tilde{d}})$. Note that minimizing $L(A,C)$ is a least squares type problem. As such we can derive a GCV criterion for choosing τ, following Li and Yin (2008),

$$\frac{\|(I_{nm} - Q)(U^{1/2} \otimes I_n)\mathrm{vec}(V)\|^2}{nm\{1 - \mathrm{trace}(Q)/(nm)\}^2}, \tag{3.11}$$

where $V = (v_1, \ldots, v_m) \in \mathbb{R}^{n\times m}$,

$$Q = \left\{ U^{1/2}\hat{A}(\tau)^\intercal \left(\hat{A}(\tau) U \hat{A}(\tau)^\intercal\right)^{-1} \hat{A}(\tau) U^{1/2} \right\} \otimes \left\{ \tilde{K}(\tilde{K}^2 + n^2\tau I_n)^{-1}\tilde{K} \right\},$$

and $\hat{A}(\tau)$ denotes the minimizer of $L(A,C)$ for a given τ. We choose τ such that (3.11) attains its minimum.

3.3 Pathway ranking and selection

3.3.1 Groupwise dimension reduction

Biological units such as genes, proteins and metabolites have inherent pathway structures, and units in the same pathway often have coordinated functions in affecting phenotype activities. In this chapter, we focus on gene regulatory pathway, while the methodology applies to other biological pathways as well. We will adopt a simple view of gene pathways by treating each pathway as a static gene cluster. This view has been adopted in many gene pathway studies, e.g., Luan and Li (2008); Ma and Kosorok (2009); Pan and Zhao (2008); Shi and Ma (2008) and Wei and Li (2007).

Specifically, our goal is to identify pathways that are relevant to phenotype activities and to model their effects. For that purpose, we first construct gene pathways using information retrieved from public databases; in our study, we use KEGG (`http://www.genome.ad.jp/kegg`). We then divide genes into groups by the pathways, and conduct nonlinear dimension reduction of the phenotype given all the genes in that pathway. Li (2009) noted that some additional conditions are needed to ensure such groupwise dimension reduction to preserve full information. However, the sensitivity analysis in Li (2009) also suggested that, with a reasonable sample size, this groupwise reduction strategy often works satisfactorily. For simplicity, we adopt this strategy in our analysis. For each pathway, we extract one or a few sufficient predictors $\langle \beta_j, \phi(x) \rangle$'s as summary features. Then we fit a generalized linear model or a generalized additive model to connect the phenotype with all the pathways under study. We emphasize that dimension reduction is an important intermediate step in this procedure. It brings down the dimensionality of the data to a much lower and manageable scale, and it grants full flexibility in subsequent model building and selection. Without dimension reduction, it

would have been far more difficult, and sometimes even intractable, to apply the conventional methods like GLM and GAM. We also point out that our solution is similar in spirit to Ma and Kosorok (2009), who summarized pathways using the first few principal components of genes. Unlike the PCA-based method, however, our approach takes into account the response information directly in the process of dimension reduction, whereas PCA does not. Moreover, our method permits flexible interactive relations among genes, while PCA is often restricted to main effects only, or at most to second-order interactions, due to its computational feasibility.

In practice we often select a small number of summary features for each pathway. In our simulation studies in Section 3.4, we choose only the leading sufficient predictor. It is not our intention to suggest that one summary feature would always be sufficient for all data analysis. However, based on our limited experiences, we do find that it is often the case that a small number of sufficient predictors are good enough, especially for a nonlinear dimension reduction method. Moreover, a small number of summary features are also easier for the interpretation purpose. A similar view can be found in the schematic model of Chatterjee et al. (2006, Figure 1) in the genetics studies.

3.3.2 Model-based pathway selection after dimension reduction

Suppose there are G pathways. We denote the summary features from the g-th pathway as $Z_{g1}, \ldots, Z_{gd_g}$, for $g = 1, \ldots, G$. We then fit a generalized linear model (GLM) (McCullagh and Nelder, 1989) or a generalized additive model (GAM) (Hastie and Tibshirani, 1990) of Y on Z's,

$$\begin{aligned} g\{E(Y|Z)\} &= \theta_0 + \theta_{11}Z_{11} + \ldots + \theta_{1d_1}Z_{1d_1} + \ldots + \theta_{G1}Z_{G1} + \ldots + \theta_{Gd_G}Z_{Gd_G}, \\ g\{E(Y|Z)\} &= \theta_0 + f_1(Z_{11}, \ldots, Z_{1d_1}; \theta_1) + \ldots + f_G(Z_{G1}, \ldots, Z_{Gd_G}; \theta_G), \end{aligned}$$

where $g(\cdot)$ is a known link function, and f_g's are unspecified smooth functions.

For the purpose of both pathway ranking and selection, we employ L_1 regularization in GLM and GAM. For GLM, we adopt the group Lasso penalty of Yuan and Lin (2006), since the presence or absence of a pathway in the model depends on the entire group of parameters $(\theta_{g1}, \ldots, \theta_{gd_g})$, so they should be shrunk to zero simultaneously. For GAM, we adopt the two-step procedure of nonnegative garrote (Breiman, 1995; Yuan and Lin, 2007a). That is, we first obtain the GAM estimates $f_g(Z_{g1}, \ldots, Z_{gd_1}; \hat{\theta}_g)$ of f_g, $g = 1, \ldots, G$, with no penalty. We then introduce a shrinkage coefficient w_g for each f_g, and penalize on those shrinkage coefficients w's. As an example, for the Gaussian data, we consider the minimization of

$$\left\{Y - w_1 f_1(Z_{11}, \ldots, Z_{1d_1}; \hat{\theta}_1) - \ldots - w_G f_G(Z_{G1}, \ldots, Z_{Gd_G}; \hat{\theta}_G)\right\}^2$$

over $(w_1, \ldots, w_G)$ subject to the constraints that $\sum_{g=1}^{G} w_g < \zeta$ for some nonnegative shrinkage parameter ζ, and $w_g \geq 0, g = 1, \ldots, G$. A decreasing penalty parameter ζ would shrink some w_g's to be exactly zero, which in effect screening out those irrelevant pathways. Computationally, the minimization can be solved by calling readily available standard Lasso algorithm such as LARS (Efron *et al.*, 2004), and is straightforward and fast.

We note that the entire solution path can be obtained for the above procedure. As such, the order of those pathways entering the model provides a natural way to rank the pathways. Moreover, an explicit pathway selection is possible, provided that an appropriate criterion is available to tune ζ. However, our numerical experiences have found that the criterion like Bayesian information criterion that is often recommended in the penalized estimation literature does not work satisfactorily in our context. For this rea-

son, we propose here a heuristic pseudo pathway selection strategy. The idea is similar in spirit to, though not the same as, Wu *et al.* (2007), who achieved variable selection in a conventional linear model setup by introducing pseudo variables into the data. Our strategy calls to add a group of pseudo variables that are generated from a standard normal distribution and uncorrelated with the existing pathways. We treat these pseudo variables as if they were from a single pathway, and apply the nonlinear dimension reduction technique to obtain summary features of this pseudo pathway. We then amend the pseudo pathway to the original G pathways and conduct pathway ranking. We repeat this procedure for B times (say, $B = 100$). For each of those G pathways in the original data, we record the frequency of times that pathway shows up earlier than the known pseudo pathway on the solution path. We declare a pathway relevant to the response if its associated frequency is above a pre-specified thresholding rate r (say, $r = 0.9$).

3.4 Simulations

3.4.1 Simulation setup

We generate totally G pathways, each of which contains p_G variables. Each variable follows a uniform distribution between 0 and 1, and variables within the pathway admit a compound symmetric correlation structure with correlation 0.2. The phenotype Y is related to four pathways, $P_1, \ldots, P_4$, through the function

$$Y = 9P_1 + 1.5\exp(3P_2) + 75(P_3 - 0.5)^2 + 3.75\sin\{2\pi(P_4 - 1)/3\} + \varepsilon,$$

where ε is a standard normal error independent of all pathways. The coefficients in front of each pathway function are to make their effects comparable in magnitude. The four

relevant pathways are composed of individual variables as,

$$\begin{aligned} P_1 &= X_{11} - X_{12}, \\ P_2 &= X_{21} + X_{22} + X_{23} - X_{21} \times X_{22} - X_{22} \times X_{23} - X_{21} \times X_{23}, \\ P_3 &= X_{31}, \\ P_4 &= X_{41} + X_{42} + X_{43} + X_{44} + X_{45}, \end{aligned}$$

where X_{gj} denotes the j-th gene in the g-th pathway. So the first, third and fourth pathways depend on their individual variables in a linear fashion, whereas the second pathway involves both linear and interaction terms. Meanwhile, the first pathway affects the response through a linear function whereas the other three through nonlinear functions. The sample size is set as $n = 100$. We consider both a relatively small number of pathways, $G = 10$, which is encountered in some Metabolomics studies, and a large number of pathways, $G = 50$, which is closer to gene pathway analysis. For each pathway, we first set a small number of variables, $p_G = 5$, which is mainly because some methods we are to compare can only work when $n > p$. Later, we will also consider the $p_G = 150$ case. In the real data analysis, we will further complement our numerical study with a case where the number of predictors in groups is large, and some are larger than the sample size.

3.4.2 Pathway ranking

We compare our method of nonlinear dimension reduction followed by the regularized generalized linear model or generalized additive model, with group additive regression (GAR) of Luan and Li (2008), and nonparametric pathway-based regression (NPR) of Wei and Li (2007). It is noted that both those methods produce only a rank of pathways

according to their relevance to the response. For this reason, we repeat the data replications 100 times and report the number of times that each of those four truly relevant pathways is ranked among the top four pathways as our evaluation criterion. We also compare some variants at the steps of dimension reduction and post reduction model selection. Specifically, at the dimension reduction step, we compare principal components analysis (PCA) (Ma and Kosorok, 2009), partial least square (PLS), the conventional SIR and SAVE, and regularized sliced inverse regression using a Gaussian kernel (KSIR). We choose the regularized kernel version of SIR only, due to the simplicity and the popularity of SIR in the dimension reduction literature. At the modeling step, we compare the generalized linear model (GLM), which reduces to a linear model (LM) with a normally distributed response in our setup, and the generalized additive model (GAM) when $G = 10$. We use LM only when $G = 50$ due to the limited sample size ($n = 100$).

Left part of table 3.1 reports the results of pathway ranking. Comparing five dimension reduction methods first under $G = 10$, we see that PCA followed by GAM works fine for pathways P_2, P_3 and P_4 but fails for pathway P_1. This is not surprising though, since by design the leading principal component direction should be close to $(1, 1, 1, 1, 1)^\intercal$, which is perpendicular to the actual direction $(1, -1, 0, 0, 0)^\intercal$. This simple example reveals a key drawback of PCA, which does dimension reduction in an unsupervised fashion that does not take into account the response information. In our example, the response is affected by the first pathway that is aligned along the direction $(1, -1, 0, 0, 0)^\intercal$, whereas PCA totally ignores that information. PLS does take into account the response information, but it still fails for P_2 and P_3. Similar behavior is observed for SIR. This is partly due to that, pathway P_2 consists of both main effects and interaction terms of the individual variables, and for this reason, one linear combination can not capture all regression information. For pathway P_3, although it only consists of a single variable,

Table 3.1: Pathway ranking and pathway selection for simulation studies. Under "pathway ranking" are numbers of times of the 4 truly relevant pathways being ranked as the top 4 most important ones among all pathways out of 100 data replications. Under "pathway selection" are numbers of times of each pathway being selected by the pseudo pathway selection strategy. Column P_{rest} reports the average times selected times of all the rest of irrelevant pathways.

Method		Pathway Ranking				Pathway Selection				
Reduction	Model	P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4	P_{rest}
PCA	GAM	16	94	74	100	9	89	61	98	12
	LM	41	35	37	100	16	21	22	99	16
PLS	GAM	98	48	52	99	98	39	46	96	12
	LM	100	32	25	99	97	20	13	92	14
SIR	GAM	99	56	48	97	98	40	40	93	11
	LM	100	33	25	97	98	20	14	92	14
SAVE	GAM	51	89	99	44	33	82	99	33	11
	LM	65	37	42	66	52	23	20	48	17
KSIR	GAM	95	90	92	95	98	90	94	95	12
	LM	94	90	92	94	97	90	92	95	12
GAR (Luan and Li, 2008)		78	42	40	0	–	–	–	–	–
NPR (Wei and Li, 2007)		87	37	91	84	–	–	–	–	–

its effect in the response model is dominated by a quadratic trend to which PLS and SIR are insensitive. SAVE performs unsatisfactorily for this example, mainly due to the limited sample size. Finally, KSIR achieves the best performance across all pathways, thanks to its flexibility to accommodate various structures. Moreover, it is interesting to note that, after KSIR, GAM and LM yield very similar results, indicating that KSIR has captured most nonlinear effects in the model. This observation is further reinforced by Figure 3.1, which shows the GAM fit for each pathway based on the summary feature produced by KSIR. It is seen from the plot that, although some functions show a little nonlinear trend, overall the functions are all close to being linear, which partly explains why LM works about the same as GAM in this example.

Next we compare our method with GAR of Luan and Li (2008) and NPR of Wei and Li (2007). It is clearly seen that GAR performs poorly for pathways P_2, P_3 and P_4. For P_2, this is because the base learner in GAR only models the main effects, whereas P_2 involves the interaction terms; and for P_3 and P_4, GAR is not designed to handle nonlinear effects, whereas P_3 and P_4 both have a dominating nonlinear association with the response. NPR is designed to tackle nonlinear effects, which explains why it works well with P_3 and P_4. On the other hand, its performance is less satisfactory for P_2, mainly due to the limited maximum depth of regression tree (which is set as 2) as its base learner. By slightly increasing the maximum depths, NPR is seen (results not shown here) to identify P_2 more frequently, but at the cost of decreasing frequency of finding the other three pathways.

Simulation results of 50 pathways exhibit similar qualitative patterns (results not shown), although for pathway ranking, sensitivity of most dimension reduction based methods is seen slightly decreased, the degradation of GAR and NPR is even more significant. Overall, our proposed method achieves the best performance, and is seen to

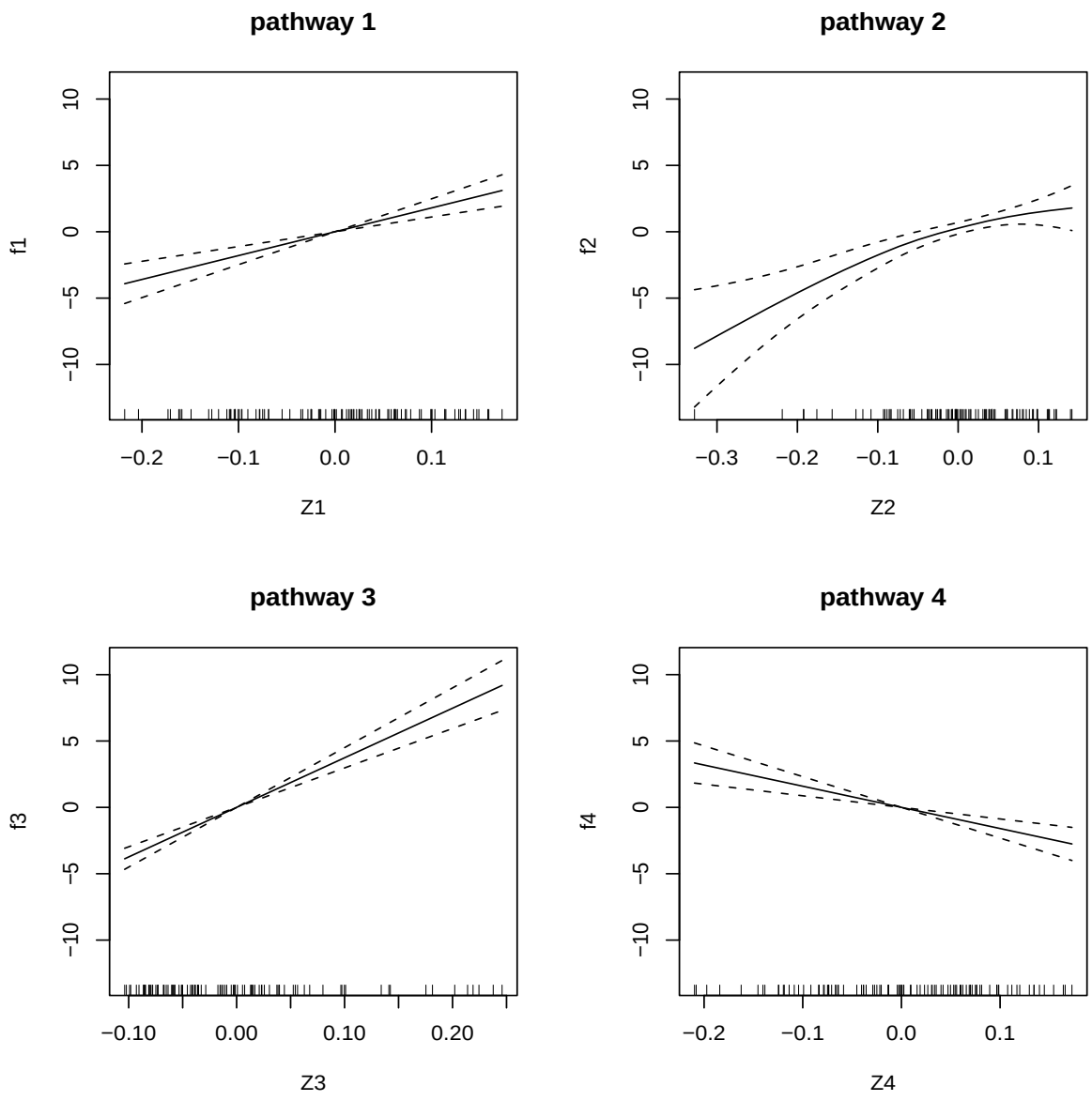

Figure 3.1: Component smooth functions of the fitted generalized additive model based on the kSIR summary features of the four relevant pathways.Upper and lower dashed lines are 2 standard errors above and below the estimate of the smooth function (solid line).

be capable of handling both complicated variable structures within a pathway as well as nonlinear pathway effects on the response.

We also consider a case where the number of variables in pathway is larger than the sample size. Specifically, we generate 10 pathways, each of which consists of 150 variables. Only the first pathway P_1 is relevant to the response through the model $Y = \exp(P_1) + \varepsilon$, and $P_1 = X_1 + \ldots + X_6 - X_7 - \ldots - X_{10} + X_1^i + X_2^i - X_3^i - X_4^i$, where the last four terms are interactions that are randomly selected from all possible two-way interactions of the first ten predictors without replacement. For this small-n-large-p setup, the usual SIR, SAVE and GAR are not applicable. The number of times that P_1 is ranked as the top pathways is recorded for 100 data replications. With the LM fitting, the results for PCA, PLS, and KSIR are 73, 0, and 85, respectively, whereas with the GAM fitting, the numbers are 47, 0, and 81. The number for NPR is 72. Particularly, PLS fails due to its sensitivity to the sparsity of true model. Overall, our proposed method, again, achieves the best performance in this setup.

3.4.3 Pathway selection

Next we examine the performance of our method in terms of pathway selection using the pseudo pathway strategy proposed in Section 3.3.2. Specifically, to match our simulation example, we generate five pseudo variables for a pseudo pathway, obtain the leading sufficient predictor by a dimension reduction method, and then amend it with the original G pathways to produce the solution path. We repeat this procedure for $B = 100$ times, and declare one of those G pathways being selected if its frequency of showing up earlier on the solution path than the pseudo pathway out of $B = 100$ times is above the pre-specified threshold value $r = 0.9$. Right part of table 3.1 reports the number of times that each of the truly relevant pathways being selected out of 100 data replications, plus the

average number of all the rest of irrelevant pathways. We first observe that the proposed method works pretty well in the sense that it achieves a high rate of identifying the truly relevant pathways whereas maintaining a low rate of selecting those irrelevant pathways, and the results are similar for $G = 10$ and $G = 50$ (results not shown). In addition, it echoes our observations in Section 3.4.2 that KSIR ourperforms all other dimension reduction solutions. Moreover, GAM and LM perform similarly if KSIR is employed in the reduction step.

3.5 A real data analysis

We analyze a microarray gene expression data of Horvath *et al.* (2006) that studies glioblastoma. Glioblastoma is the most common primary malignant brain tumor and one of the most lethal one among all cancers. A dataset of gene expressions of 120 patients were collected and normalized. Among those patients, only 9 were alive at the end of the study. To avoid complication of the response censoring, we focus on those $n = 111$ patients who died, and use their (uncensored) time to death as the response variable. All the genes were mapped to the 33 regulatory pathways recorded in the KEGG database. Of the total 1668 nodes of the 33 pathways, 1498 were found in our dataset. The list of those 33 pathways are given in Table 3.2 (column 1), along with the number of genes in that pathway from the KEGG database (column 2) and the number found in our dataset (column 3). Note that for some pathways, the number of genes are larger than the sample size. The goal of this analysis is then to identify pathways that are strongly associated with the patient's survival time from the brain cancer.

We analyze this data using kernel sliced inverse regression for dimension reduction, followed by fitting a generalized additive model or a linear model with L_1 regularization.

We report the results for both pathway ranking, based on the order of appearance along the entire solution path, and pathway selection, based on the pseudo pathway strategy. Table 3.2 reports the number of times out of 100 pseudo pathway generations that each original pathway appears before the known pseudo one on the solution path (column 4 for the results based on GAM, and column 5 for LM). The table also reports the ranking of all the pathways (column 6 for GAM, and column 7 for LM). It is seen that four pathways stand out clearly: complement and coagulation cascades, mitogen-activated protein kinase (MAPK) signaling pathway, cytokine-cytokine receptor interaction, and neuroactive ligand-receptor interaction. The four pathways are ranked as top four by both methods, and the number of times they appear ahead of a pseudo pathway are all above 95. Besides, the results based on GAM and LM agree well.

There seem to exist recurring biological evidences to support our findings of those four important pathways. For the complement and coagulation cascades pathway, It was reported in Liu *et al.* (2008) that the glioma cell invasiveness depends on proteases of the coagulation and complement cascades. In addition, suppression of the tissue factor-dependent coagulation cascade is found to be a contributing factor for the development of intratumoral hemorrhage ire glioblastoma (Takeshima *et al.*, 2000). Coagulation cascade activation was also found to be related to glioma cell proliferation (Ogiichi *et al.*, 2000). The MAPK signaling pathway is involved in various cellular functions, including cell proliferation, differentiation and migration. It has been linked to being responsible for the malignant phenotype, including increased proliferation, defects in apoptosis, ability to induce neovascularization, and invasiveness (Liu *et al.*, 2008), which are important prerequisites for the infiltrative and destructive growth patterns of malignant gliomas. Moreover, Mawrin *et al.* (2003) and Pelloski *et al.* (2006) reported the prognostic relevance of MAPK expression in glioblastoma multiforme. It is worth to mention that,

when Li and Li (2008) analyzed the same dataset using a network-constrained regularization and variable selection method, they identified a well connected sub-network of genes, while genes from the MAPK signaling pathway constitute the majority of that sub-network. The goal of Li and Li (2008) was not to select individual pathways, but to enhance gene selection by using the network information. But both studies point to the same pathway of interest, i.e., the MAPK signaling pathway. Cytokine-cytokine receptor interactions and Neuroactive ligand-receptor interactions are also highly likely to be associated with glioblastoma. Cytokines are important regulators and mobilizers of cells involved in cell growth, differentiation and death, angiogenesis, and development and repair processes aimed at the restoration of homeostasis. Cytokine-cytokine receptor interaction is the way Cytokines induces responses to an activating signal. Neuroactive ligand-receptor interactions also involve signaling molecules and interactions that play critical roles in a variety of important cellular processes, such as apoptosis, cytolysis, and cell proliferation.

3.6 Discussion

The focus of this chapter is to propose a two-step statistical method for analyzing high-throughput biological data while utilizing pathway information. An important intermediate step is nonlinear dimension reduction, which turns high-dimensional data into a much lower-dimensional and manageable summary features, and permits flexible predictor interactions within the groups as well as flexible pathway effects on the phenotype response. Although there is no guarantee that nonlinear dimension reduction would always be superior than linear reduction, its flexible reduction form may facilitate both dimension reduction itself, as partly illustrated by example (3.3), and modeling after

Table 3.2: Pathway ranking and pathway selection for the analysis of the glioblastoma microarray data. The top four pathways are marked in bold face.

	No. of genes		Selection		Ranking	
Pathway name	database	dataset	GAM	LM	GAM	LM
MAPK signaling pathway	269	249	**97**	**100**	**2**	**3**
Calcium signaling pathway	154	142	15	5	12	19
Cytokine-cytokine receptor interaction	142	126	**99**	**100**	**4**	**4**
Phosphatidylinositol signaling system	70	65	7	8	29	13
Neuroactive ligand-receptor interaction	133	116	**100**	**100**	**1**	**1**
Cell cycle	48	44	17	55	13	5
Ubiquitin mediated proteolysis	21	19	7	5	19	22
Apoptosis	75	72	44	0	14	24
Wnt signaling pathway	143	124	35	0	17	33
TGF-beta signaling pathway	66	61	1	0	32	31
Axon guidance	111	100	5	5	25	17
Focal adhesion	121	115	50	7	5	14
ECM-receptor interaction	53	50	6	2	24	25
Cell adhesion molecules	91	79	28	53	9	6
Adherens junction	70	68	10	5	15	23
Tight junction	105	91	12	1	10	27
Gap junction	92	88	3	0	30	29
Complement and coagulation cascades	53	51	**95**	**100**	**3**	**2**
Toll-like receptor signaling pathway	73	71	25	21	8	9
Jak-STAT signaling pathway	98	91	28	0	7	30
Natural killer cell mediated cytotoxicity	121	109	4	7	27	15
Circadian rhythm	7	6	10	22	16	8
Regulation of actin cytoskeleton	169	154	9	5	31	20
Insulin signaling pathway	134	129	5	9	23	12
Adipocytokine signaling pathway	64	60	1	5	33	21
Type II diabetes mellitus	42	40	8	0	18	32
Type I diabetes mellitus	5	5	4	1	28	28
Alzheimer's disease	12	11	37	35	6	7
Prion diseases	9	7	12	16	11	10
Unknown	9	9	7	6	21	16
Unknown	22	21	19	5	20	18
Unknown	11	11	9	2	22	26
Unknown	10	8	4	14	26	11

dimension reduction, as demonstrated by our simulation studies. Moreover, depending on the choice of kernels, the kernel based nonlinear dimension reduction includes the usual linear reduction as a special case. The new method comes with a price of a more complicated but tractable implementation, and the computer code (in R) is available upon request.

There are a number of possible avenues for future extensions. First, we have been focusing on ranking and selecting the entire pathway as a whole, while it could also be of interest to select individual genes within the pathway that are relevant to the phenotype. Both Wei and Li (2007) and Luan and Li (2008) can target individual genes. To accomplish this task, it requires an effective variable selection technique in the kernel setting, which, to our knowledge, has only limited success (Zhang, 2006). Secondly, the pathway information is now only used to divide the genes into groups, while how to incorporate further pathway information, e.g., the known network structure of a given pathway, is important. Finally, kernel selection remains a critical yet open question for all kernel based methods. All these problems are warranted for future research.

Chapter 4

Nonlinear Sufficient Dimension Reduction for Association Testing of Complex Traits

4.1 Introduction

Genome-wide association studies (GWAS) based on single nucleotide polymorphism (SNP) chips have enjoyed varying degrees of success in identifying the genes associated with some complex diseases or traits (Easton and Eeles, 2005; Frazer *et al.*, 2009; Lettre and Rioux, 2008). Now it has been widely recognized that common variants explain only a small fraction of the population variation of most disease traits (Frazer *et al.*, 2009). The common disease rare variants (CDRV) hypothesis states that the complex disease may be attributed to multiple rare variants with relatively high risks and has attracted much attention recently. Deep resequencing is emerging as a new and potent means for mapping complex trait genes (Hodges *et al.*, 2007; Turner *et al.*, 2009). Se-

quence data characterizes all variants within each gene and allow direct search for the functional variants with biological effects. Multiple deleterious or protective rare variants have been identified for low plasma levels of high-density lipoprotein cholesterol (Cohen *et al.*, 2004), hypertension (Ji *et al.*, 2008), and type-I diabetes (Nejentsev *et al.*, 2009).

Resequencing delivers orders of magnitude more variants than SNP chips and many of them are rare variants with minor allele frequency (MAF) less than 1%. The traditional single marker test for common variants is underpowered to detect causal variants with low frequency. A particular rare disease predisposing allele may be present in only a handful of patients. Hence, statistical tests that capture only marginal effects are doomed to low power. This suggests focusing on disease gene discovery rather than disease variant discovery. A successful strategy should effectively merge information in SNP variants by genes or pathways. Li and Leal (2008) proposed a group-wise test exploiting both multivariate and collapsing strategies that possess higher power than a simple multivariate test or simple collapsing. Madsen and Browning (2009) extended the method by incorporating weights (dependent on allele frequency) into the group-wise statistics and approximating p-values by permutations within each group. Both methods consider rare variants with minor allele frequencies (MAF) falling below a pre-specified threshold and exclude more common variants from analysis. It would be counterproductive to discard the common variants because in reality there is no way to know in advance which SNPs are informative. Thus, statistical methods that can analyze both rare and common variants simultaneously are preferable. The pooling strategy of Price *et al.* (2010) circumvents the issue of arbitrarily chosen frequency threshold by calculating a group-wise statistics under a variety of thresholds. Higher power is achieved at the cost of an increased computational burden. These methods have certain drawbacks: 1) environmental predictors are excluded from analysis even though they may contribute significantly to

an association; 2) gene-by-gene and gene-by-environment interactions remain unresolved; 3) more importantly, existing methods are sensitive to the classification of variants. If all types of variants (deleterious, protective, or neutral) coexist, then the various signals can cancel one another and potentially compromise statistical power. Liu and Leal (2010) proposed a SNP-set genotype-based statistic for rare variants and declared that the common variants and environment factors can be modeled together with the rare variant statistic in a logistic regression model. However, interactions between rare and common variants cannot be explicitly modeled in this way, and the method handles only dichotomous traits.

The field of dimension reduction (DR) appeals for this high-dimensional problem. It is based on the belief that high dimensional data can be effectively summarized on a low dimensional space, and it can reduce the dimensionality of the predictor vector prior to subsequent modeling and analysis efforts. For example, Chen *et al.* (2010) apply principle component analysis (PCA) to combine SNP information within pathways and generate the so-called eigen-SNPs, which are used in followed association testing. PCA has at least two limitations. First, PCA aggregates SNP information regardless of the trait information. Because the ultimate goal is mapping the genes associated to the traits being studied, it is advantageous to aggregate SNP information under the guidance of trait information (disease status or quantitative traits). In statistical jargon, we want to perform *supervised* dimension reduction. Second, the eigen-SNPs, or principle components, are *linear* combinations of the SNPs. As a consequence, association tests based on eigen-SNPs are doomed for low power when there are interactions between SNPs.

Motivated by these drawbacks, in this chapter we use the kernel-based nonlinear dimension reduction methods proposed in last chapter for association test based on SNP

or sequence data. Unlike Chapter 3 which emphasizes the overall methodological framework, we want to stress in this chapter one key ingredient of this approach: the kernel functions, which essentially measure similarity between subjects. The commonly used kernels, such as Gaussian, polynomial, spline, sigmoid, etc, work successfully with continuous attributes, but may perform poorly for discrete genetic data. We propose to use Markov chain theory to generate kernels that effectively capture the similarity between individual genotypes. Comparison with currently available collapsing methods shows superior power of our approach over a variety of scenarios.

The remainder of this chapter is organized as follows. Section 4.2 describes our method for obtaining summary statistics for a SNP-set and the kernel generation mechanism. Various methods for grouping SNP information are compared on a simulation study in Section 4.3. Finally, Section 4.4 makes conclusions and discusses potentially helpful extensions.

4.2 Methods

Suppose n study individuals are genotyped at p SNPs, denoted by $X_1, \ldots, X_p$, and the trait Y can be either binary (case control study) or quantitative. Potential covariates (sex, age, smoke, etc) are denoted by $C_1, \ldots, C_t$. The goal is to test the association of the trait and all markers jointly with adjustment of covariates. In sequence studies, the number of markers p is potentially large and can outnumber the number of observations n. Our method first performs dimension reduction on the markers and then tests association based on the summary statistics.

4.2.1 Calculation of Summary Statistics

The classical principal component analysis (PCA) has been applied to several genetic and genomic problems. For instance, PCA is employed to adjust for population stratification (Price *et al.*, 2006) in GWA studies. PCA has also been applied to produce a set of eigen-genes or super-genes for association mapping (Chen *et al.*, 2010). Given p SNPs, PCA seeks linear combinations of these SNPs that have maximal variances. Typically PCA is solved by an eigen-decomposition of SNP covariance matrix. Then the eigenvectors with leading eigenvalues give the coefficients of linear combinations being sought. The linearly transformed SNPs form the eigen-genes used in follow-up analysis. PCA is simple to use, significantly reduces the dimensionality, and is effective for handling collinearity. Despite its widespread applications, there has been a common criticism of principle components based regression. The principle components are computed purely based on the SNP information and the trait value information is not taken into account. Since the data reduction is achieved without regard to the response, there is no reason that the response should not be associated with the least important principle components. Consider a simple example. Suppose two SNPs S_1 and S_2 are in linkage disequilibrium (LD). Then the covariance of the linear combination $S_1 + S_2$ is larger than that of $S_1 - S_2$ and thus the eigen-gene found by PCA will be closer to the direction $S_1 + S_2$ than to $S_1 - S_2$. If in truth these two SNPs have opposite effects – one deleterious the other protective –then the trait depends on the SNPs through $S_1 - S_2$ and the eigen-gene would contain no signal for association.

Intuitively it is natural to incorporate the trait information during the phase of dimension reduction. In Chapter 3, we have proposed a framework of nonlinear sufficient dimension reduction approaches that can be applied for this purpose. For the sake of

completeness of this chapter, we will brief review these approaches. Consider a regression of a response Y given a p-dimensional predictor X. The basic idea of nonlinear extension to linear dimension reduction is to employ a function $\phi(\cdot)$, with an associated gram matrix K, to map X to $\phi(X)$; one then carries out a linear dimension reduction in the space of $\phi(X)$, which in effect results in a nonlinear dimension reduction in the original predictor space $\mathcal{X}$. The well-known kernel trick turns the primal problem that depends on the dimension of $\phi(X)$, which is high or even infinite, to a dual problem that only depends on the sample size. Consequently, the method works for small-n-large-p problems.

Specifically, nonlinear dimension reduction seeks the inner products $(\langle\beta_1, \phi(X)\rangle, \ldots, \langle\beta_{\tilde{d}}, \phi(X)\rangle)$, such that

$$Y \perp\!\!\!\perp X \mid \big(\langle\beta_1, \phi(X)\rangle, \ldots, \langle\beta_{\tilde{d}}, \phi(X)\rangle\big). \tag{4.1}$$

That is Y depends on X only through those inner products. We refer them as the *nonlinear sufficient predictors*, and assume the number $\tilde{d}$ of inner products $\leq \min(n, p)$.

In terms of estimation, these inner products can be obtained by solving the following generalized eigen decomposition problem:

$$\tilde{K} J \tilde{K} \alpha_j = \rho_j \tilde{K}^2 \alpha_j, \quad j = 1, \ldots, \tilde{d}, \tag{4.2}$$

where $\tilde{K} \in \mathbb{R}^{n\times n}$ is the centered Gram matrix, J is a method-specific $n \times n$ matrix. Then for a new observation $x \in \mathcal{X}$, $\langle\beta_j, \phi(x)\rangle = \alpha_j^{\intercal}[\tilde{k}(x_1, x), \ldots, \tilde{k}(x_n, x)]^{\intercal}$, where $\tilde{k}(x_i, x) = k(x_i, x) - n^{-1}\sum_{l=1}^{n} k(x_l, x),\ i = 1, \ldots, n$. So the inner product $\langle\beta_j, \phi(x)\rangle$ can be obtained from the kernel k and α_j's from (4.2). Due to (4.2), the proposed nonlinear sufficient

dimension reduction approaches only involve decomposition of an $n \times n$ matrix, so it can handle $n < p$. Its flexible reduction form beyond the linear combination is also expected to facilitate dimension reduction. Consider the illustrative example used in Chapter 3, where $X = (X_1, \ldots, X_6)$ and $Y = X_1 + X_2X_3 + X_4^2 + X_5X_6 + \varepsilon$, with an independent error ε, where no linear reduction is possible. However, if one employs a quadratic kernel, then only one linear combination in the mapped feature space is needed to summarize all regression information, and thus substantial reduction is achieved. Finally, if a linear kernel is employed, the proposed method reduces to the usual linear dimension reduction.

4.2.2 Choice of Kernels

In previous discussion we have omitted the specific form of kernel. The choice of kernel turns out to be critical for nonlinear dimension reduction. Ideally the kernel should capture the genetic similarity between two individuals. The most popular Gaussian kernel works well for continuous predictors but may perform poorly on categorical predictors such as SNPs. Some specialized kernels have been crafted for SNP data. The identity-by-sharing (IBS) kernel (Wessel and Schork, 2006) calculates the distance between two genotypes x_i and x_j coded with numbers of minor alleles as

$$K(x_i, x_j) = \frac{\sum_{s=1}^{p} 2I(x_{is} = x_{js}) + I(|x_{is} - x_{js}| = 1)}{2p}.$$

The more general weighted IBS kernel (Kwee *et al.*, 2008; Wu *et al.*, 2010) takes the form

$$K(x_i, x_j) = \frac{\sum_{s=1}^{p} w_s\big(2I(x_{is} = x_{js}) + I(|x_{is} - x_{js}| = 1)\big)}{2\sum_{s}^{p} w_s}$$

which offers flexibility of incorporating variant specific weights into kernel. Kwee *et al.* (2008) and Wu *et al.* (2010) use $w_s = 1/\sqrt{f_s}$ where f_s is the MAF of the variant, up-weighting the importance of rare variants. Critical to the successes of various kernel-based methods, designing optimal kernels that capture genetic similarity becomes ever pressing (Schaid, 2010a,b). Here, we propose a class of new kernels based on Markov chain theory. We first briefly review some basics about Markov chains.

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space equipped with a σ-finite measure μ. Suppose we are given a Markov chain on state space $\mathcal{X}$ described by its transition density $K(x, x')$ with respect to $\mu(dx')$. Suppose further that the chain has stationary measure $\pi(dx) = \pi(x)\mu(dx)$. $K^l(x, \cdot)$ denotes the density of the chain started at state x after l steps. Let $l^2(\pi) = \{f : \mathcal{X} \to \mathbb{R} : \int_{\mathcal{X}} f^2(x)\pi(x)\mu(dx) < \infty\}$ denote the Hilbert space equipped with inner product

$$\langle f, g \rangle_{l^2(\pi)} = \mathbf{E}_\pi[f(X)g(X)] = \int_{\mathcal{X}} f(x)g(x)\pi(x)\mu(dx).$$

The Markov chain K operates on $l^2(\pi)$ by

$$Kf(x) = \int_{\mathcal{X}} K(x, y)f(y)\mu(dy).$$

K is called reversible when

$$\pi(x)K(x, x') = \pi(x')K(x', x) \text{ for all } x, x' \in \mathcal{X},$$

or equivalently, when the operator K is self-adjoint

$$\langle Kf, g \rangle_{l^2(\pi)} = \langle f, Kg \rangle_{l^2(\pi)}.$$

Suppose that $l^2(\pi)$ admits an orthonormal basis of eigenfunctions $\{\phi_n\}_{n\geq 0}$ with $\phi_0 \equiv 1$ such that

$$K\phi_n(x) = \beta_n \phi_n(x), \quad n \geq 0,$$

where the eigenvalues $\{\beta_n\}_{n\geq 0}$ satisfy $\beta_0 = 1$, $|\beta_n| \leq 1$, and $\sum_{n=0}^{\infty} \beta_n^2 < \infty$. Then K is a Hilbert-Schmidt operator and

$$K^l(x, x') = \sum_{n=0}^{\infty} \beta_n^l \phi_n(x)\phi_n(x')\pi(x').$$

The series on the right hand side converges in the $l^2(\pi \times \pi)$ sense. This shows that all such Markov kernels are Mercel kernels.

Wright-Fisher Kernel

For genetic data, we focus on the discrete state space $\mathcal{X} = \mathbb{N}_N^p = \{x \in \mathbb{N}^p : \sum_{j=1}^p x_j = N\}$, and propose new kernels based on Markov chains on $\mathcal{X}$. For example, the Wright-Fisher process (Ewens, 2004) in population genetics gives a natural kernel. First we model a genotype x as a Dirichlet-Multinomial random variable with parameter $\alpha = (\alpha_1, \ldots, \alpha_p)$, which is estimated from genotypes of all the individuals. The transition kernel of Wright-Fisher process between two genotypes x_i and x_j is

$$P(x_i, x_j) = \binom{N}{x_j} \prod_{s=1}^{p} \pi(x_{is})^{x_{js}}$$

where $\pi(x_{is}) = \frac{x_{is}+\alpha_s}{N+|\alpha|}$ and $|\alpha| = \sum_{s=1}^p \alpha_s$. The stationary distribution of P is not known explicitly but can be approximated by a Dirichlet-Multinomial distribution with parameter α. Note that P is an irreversible Markov kernel, but either the additive or

multiplicative symmetrization gives a symmetric kernel: $K = \frac{P+P^T}{2}$ and $K = PP^T$. It can be shown that both are positive definite and we call this the Wright-Fisher (WF) kernel. The simulation studies in Section 4.3 shows the promise of the WF kernel combined with nonlinear dimension reduction method proposed in Chapter 3.

4.2.3 SNP-set selection

SNPs in GWA studies can be grouped by genes or pathways. Our goal is to identify groups of SNPs that are relevant to a trait and to model their effects. Similar to Chapter 3, we propose a two-step procedure for this purpose based on nonlinear dimension reduction methods coupled with a specific kernel. We first conduct nonlinear dimension reduction of the phenotype given all the SNPs in a SNP-set, for which we extract one or a few sufficient predictors $\langle \beta_j, \phi(x) \rangle$'s as summary features. Then we conduct a permutation test to evaluate significance of the SNP-set using these summary features. Specifically, we first fit a generalized linear model (GLM) to connect the phenotype with all the summary features obtained from the SNP-set. A test statistics, say a F-value, can be provided by GLM for the significance of all the summary features. We then acquire a null distribution for this test statistics by randomly permuting phenotypes among all individuals for B times (say, $B = 10,000$). Each time we obtain the same test statistics for the gene-set with respect to permuted phenotype following the same procedure as above. Mapping the test statistics obtained from the real phenotypes to its null distribution gives an empirical p-value for the SNP-set.

4.3 Simulation studies

4.3.1 Simulation setup

We have performed some simple simulation studies to illustrate the promise of the information aggregation method using the nonlinear dimension reduction and the additively symmetrized Wright-Fisher (WF) kernel discussed above. We investigate the empirical type I error and power of association test based on summary statistics from different information aggregation methods. Simulation studies are designed based on the sequencing data of the genes *TG* and *TIA1* for 697 individuals compiled from the 1,000 Genomes Project (The 1000 Genomes Project Consortium, 2010) by the Genetic Analysis Workshop 17 (GAW17). *TG* encodes a protein called thyroglobulin (Genetics Home Reference, 2011). Mutations in this gene has been found related to congenital hypothyroidism and autoimmune disorders (Genetics Home Reference, 2011). In the dataset, *TG* contains 146 SNPs, among which 33 are common variants (MAF$>$ 1%). Figure 4.1 plots LD structure of *TG*. High LD is only observed between a few pairs of SNPs. Figure 4.2 draws a histogram of MAFs for the 146 SNPs in *TG* compared to a histogram for all the SNPs in the whole dataset. The two distributions agree well except that *TG* contains slightly higher percentage of common variants. This gene is chosen because we need a certain number of common variants to simulate situations where there are multiple informative common variants with additive or epistasis effects. Gene *TIA1* encodes a poly(A)-binding protein, which is associated with apoptosis (GeneCards, 2011). In the dataset, it contains only 12 rare variants with low pairwise LD. This gene is used to simulate situations where rare variants are causal to the trait.

For each simulation scenario, a total of 1,000 replications are simulated with sample size $n = 697$. In each replication, we first simulate a quantitative trait under the null

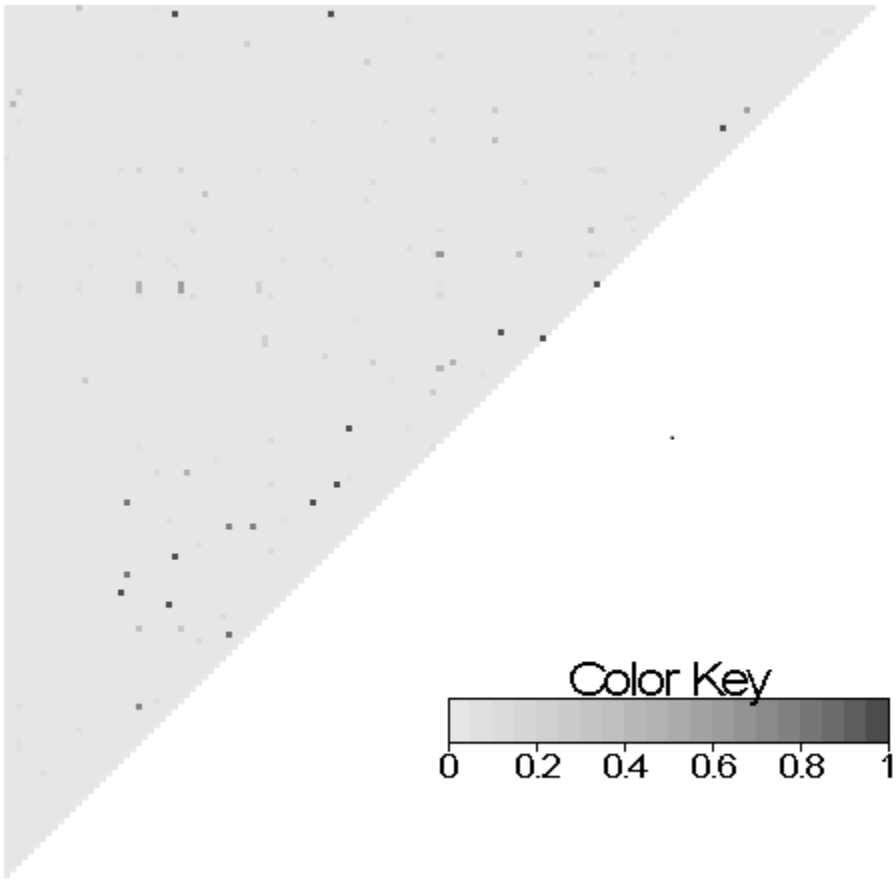

Figure 4.1: LD structure of the 146 SNPs in *TG*.

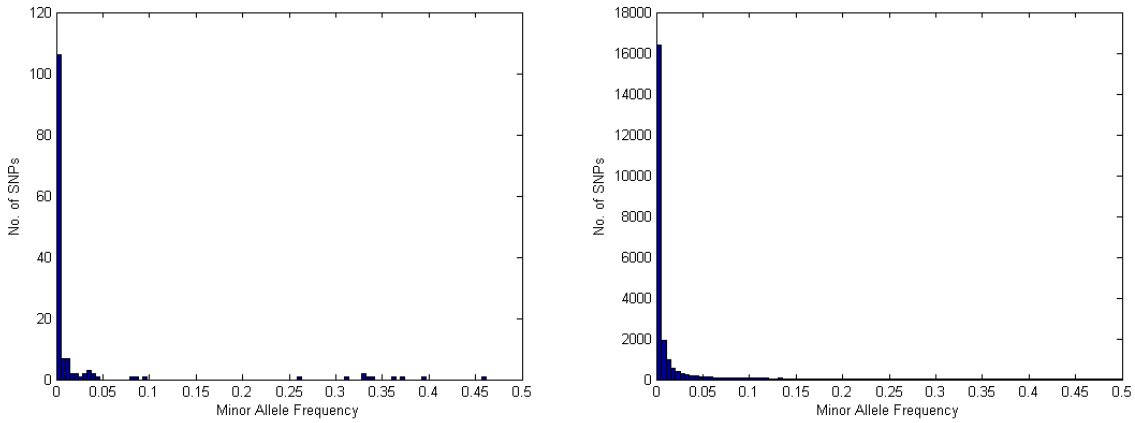

Figure 4.2: Histograms of MAFs for SNPs in TG (left) and the whole GAW17 dataset (right).

model $Q_0 = \epsilon$, where ϵ is a standard normal noise. The top 50% of the distribution of Q_0 are then declared affected, by which we define a binary disease status. It can be easily seen that under the null model, the quantitative trait and disease status do not depend on genotypes.

Then under the alternative model, a quantitative trait is generated according to

$$Q_1 = f(X_{(1)}, ..., X_{(p)}) + Q_0, \tag{4.3}$$

where X's are SNPs in descending order according to their minor allele frequencies, and f is the true genetic effect model, which differs among simulation studies (table 4.1). Various scenarios are simulated, encompassing rare variants, presence of both deleterious and protective variants, interaction effects, and nonlinear interaction effects. A binary disease status based on Q_1 is defined in a similar way as Q_0. For reference purpose, if in (4.3) $f = \sum_{j=1}^{p} \beta_j X_{(j)}$ is a linear additive model, then conditional on the fact that this is an half-affected-half-control sample and assuming existence of causal SNPs in the

gene does not heavily influence the general cutoff to Q_1 (which is the case for most of our simulation studies shown below), there is an approximate correspondence between coefficient β_j and odds ratio of the SNP $X_{(j)}$: Odds Ratio $\approx (1 - \Phi(-\beta_j; 0, 1))/\Phi(-\beta_j; 0, 1)$, where $\Phi(x; \mu, \sigma)$ represent cumulative normal distribution function of mean μ and standard deviation σ. Therefore, a coefficient β from 0.4 to 1.2 corresponds to an odds ratio from 2 to 7. On the contrary, a coefficient β from -0.4 to -1.2 reduces odds ratio to 1/2 to 1/7.

4.3.2 Simulation results

We compared our method of dimension reduction followed by logistic regression model (a special case of GLM for binary trait) with kernel-based adaptive cluster (KBAC) of Liu and Leal (2010), and variable threshold (VT) test of Price *et al.* (2010). For KBAC, if a SNP-set contains only rare variants (e.g., *TIA1*), we adopt a permutation test based on the KBAC statistic; otherwise, following Liu and Leal (2010), a logistic regression model will be used to incorporate the common variants and a variable for the kernel weight of KBAC. Standard permutation procedure is then applied to evaluate the significance. VT test can be used regardless of the existence of common variants. For our approach, we compare some variants at the steps of dimension reduction, namely principal components analysis (PCA), sliced inverse regression(SIR), and kernel sliced inverse regression (kSIR) with various kernel functions (Gaussian, IBS, and WF). For each dimension reduction-based method, the leading summary variable is used to represent the SNP-set. All the methods under comparison are permutation-based. We used 10,000 permutations for every simulation replication. For each method under comparison, an empirical power is then obtained by counting how many of empirical p-values over 1,000 replications are less than the nominal significance level $\alpha = 0.05$ under the alternative model using

the designated causal SNPs. An empirical type I error is evaluated in a similar fashion except that the trait is generated under the null model. We designed various situations to evaluate methods. Results are summarized in Table 4.1.

Empirical power Comparison

Simulations 1-4 represent main effect models. When the main effect model are only composed of deleterious rare variants (simulation 1), KBAC works the best and supervised dimension reduction-based methods outperform VT. Note that few people have totally more than one minor alleles over these three SNPs, whose MAFs are respectively 0.0057, 0.0050, and 0.0029. The existence of protective variants destroy the performance of both KBAC and VT (simulation 2), but performance of dimension reduction-based methods is not significantly affected. In simulations 3 and 4, informative SNPs are all common variants with very low LD between each other. Their MAFs all fall into the range of $0.32 \sim 0.40$. When all SNPs are deleterious (simulation 3), VT works the best, and WF outperforms IBS and KBAC. However, when two of the informative alleles are protective (simulation 4), the performance of VT drops dramatically, while the performance of dimension reduction-based methods does not. KBAC maintains similar performance because common variants are essentially tested by a logistic regression model. Note that under these linear genetic models, performance of IBS kernel SIR is still better than or comparable with linear SIR, while performance of Gaussian kernel SIR is not. It is also true for WF kernel under main effect genetic models only composed of common variants. This reflects the superiority of these kernels for sequence data.

Simulation 5 represents a model with both main and epistasis effects. Both of the two informative SNPs are common variants with very low LD. In this case, WF works better than IBS, which then outperforms KBAC, VT and other dimension reduction-

based methods. Simulations 6-9 are pure epistasis model among three common SNPs with low LD between each other. Again WF outperforms the rest. Simulations 10 and 11 represent epistasis between common and rare variants. It is clear that IBS kernel outperform KBAC and VT, however the performance of WF drops dramatically. Note that whenever there are protective variants, KBAC performs poorly.

When the phenotype is the quantitative trait in (4.3), similar pattern is observed (results not shown), except that KBAC is no longer applicable.

In general, the KBAC statistic and VT test are extremely vulnerable to protective variants, and do not capture epistasis effect as well as kernel dimension reduction-based methods. Among the different kernel methods, WF works the best for cases where common variants dominate the genetic effect, which, however, is less sensitive to rare variants. IBS is slightly worse than WF in common variant models, yet significant better in rare variant models. Finally, they beat PCA in all simulated situations.

Empirical type I error

The last two rows of Table 4.1 show that whichever gene is used, the empirical type I errors of all the methods are relatively close to the nominal significance level.

Table 4.1: Results of simulations studies. Each row represents one simulation study mimicking a specific type of true genetic effect. Under "Genetic Effect" are the true genetic effects that generate the true quantitative trait. The numbers under the names of different methods are their empirical power in different studies.

			Current Methods		Linear DR		Kernel SIR		
Index	Gene	Genetic Effect	KBAC	VT	PCA	SIR	Gauss	IBS	WF
1	*TIA1*	$1.2\big(X_{\{1\}} + X_{\{4\}} + X_{\{7\}}\big)$	**0.74**	0.50	0.13	0.69	0.60	0.62	0.57
2	*TIA1*	$1.2\big(- X_{\{1\}} + X_{\{4\}} + X_{\{7\}}\big)$	0.24	0.20	0.19	**0.67**	0.57	0.61	0.60
3	*TG*	$0.2\big(X_{\{2\}} + X_{\{3\}} + X_{\{4\}} + X_{\{7\}}\big)$	0.29	**0.92**	0.13	0.63	0.43	0.68	0.86
4	*TG*	$0.2\big(X_{\{2\}} - X_{\{3\}} - X_{\{4\}} + X_{\{7\}}\big)$	0.31	0.05	0.40	0.73	0.52	0.77	**0.97**
5	*TG*	$\big(X_{\{6\}} + X_{\{7\}} + X_{\{6\}} \times X_{\{7\}}\big)/8$	0.19	0.17	0.27	0.30	0.23	0.34	**0.56**
6	*TG*	$\big(X_{\{3\}} \times X_{\{4\}} + X_{\{3\}} \times X_{\{7\}} + X_{\{4\}} \times X_{\{7\}}\big)/6$	0.36	0.68	0.15	0.61	0.56	0.65	**0.83**
7	*TG*	$\big(X_{\{3\}} \times X_{\{4\}} + X_{\{3\}} \times X_{\{7\}} - X_{\{4\}} \times X_{\{7\}}\big)/6$	0.13	0.20	0.06	0.19	0.25	0.21	**0.39**
8	*TG*	$exp\Big(\big(X_{\{3\}} \times X_{\{4\}} + X_{\{3\}} \times X_{\{7\}} + X_{\{4\}} \times X_{\{7\}}\big)/8\Big)$	0.39	0.68	0.16	0.66	0.65	0.70	**0.86**
9	*TG*	$exp\Big(\big(X_{\{3\}} \times X_{\{4\}} + X_{\{3\}} \times X_{\{7\}} - X_{\{4\}} \times X_{\{7\}}\big)/6\Big)$	0.17	0.34	0.06	0.29	0.37	0.33	**0.55**
10	*TG*	$X_{\{2\}} \times \sum_{i=36\sim40} X_{\{i\}} + X_{\{3\}} \times \sum_{i=41\sim45} X_{\{i\}}$	0.00	0.45	0.06	0.44	0.43	**0.58**	0.13
11	*TG*	$X_{\{2\}} \times \sum_{i=36\sim40} X_{\{i\}} - X_{\{3\}} \times \sum_{i=41\sim45} X_{\{i\}}$	0.11	0.04	0.28	0.45	0.41	**0.59**	0.16
Type I Error	*TIA1*	Null model	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Type I Error	*TG*	Null model	0.05	0.05	0.05	0.06	0.05	0.06	0.05

4.4 Discussion

Ideally, an association test should be able to handle:

1. both rare and common variants, with somewhat greater stress on the former,
2. high dimensionality of genomic dataset, which typically far exceeds the sample size,
3. quantitative traits as well as disease dichotomies,
4. additive, recessive, and dominant models of gene action,
5. non-genetic predictors,
6. gene-gene and gene-environment interactions.

All current methods are compromises and fail to deliver across the board judged by these criteria. To fill the gaps, we use the nonlinear supervised dimension reduction methods proposed in Chapter 3 to aggregate SNP information within a gene or pathway, which increases the likelihood of detecting multiple causal rare variants as a group. It also turns the high-dimensional data into its low-dimensional projections, and allows complex interactive relationships among genetic variants. Gene-gene interactions can be taken into account by considering a pathway of genes as a SNP-set. Non-genetic predictors and/or environmental effects can be incorporated by the following GLM model. The flexibility of both dimension reduction approaches and GLM enables our approach to handle both quantitative and categorical traits.

An important development in this chapter is the proposal of new kernels for kernel-based dimension reduction approaches using Markov chain theory. Simulation studies have demonstrated the advantage of one of them in aggregating genetic information over

some existing and commonly used kernels for some scenarios. However, it is still suboptimal for genetic effects involving rare variants, and should be further tuned for better performance. Other Markov chain kernels that can be generated in a similar way include Ehrenfest kernels, hypergeometric kernels, and Dirichlet-Multinomial kernels (Khare and Zhou, 2009; Zhou and Lange, 2009), which should be also extensively explored. Additionally, our grouping strategy has not taken advantage of any available information regarding the relationships among SNPs within a group, e.g., their genetic distances, which, however, can potential facilitate the design of biologically meaningful kernels. Finally, generating new kernels based on Markov theory can potentially benefit many other kernel methods, such as kernel PCA, support vector machines, and nonparametric and semiparametric regressions, which is another interesting research topic.

Chapter 5

Summary and future research

5.1 Summary

This thesis is devoted to the large area of omics data analysis. Different types of omics data share a great deal of similarity with each other in several aspects. Most omics data studies are often challenged with the problem of dealing with a limited sample size and a high-dimensional feature space. Several unique characteristics of biological data, such as nonlinearity, multicollinearity, as well as the inherent correlation and network structure contribute further to the problem complexity. In terms of study objectives, many omics studies aim to identify biomarkers for a phenotype of interest in hope of deciphering the underlining mechanisms.

People have developed various statistical and machine learning approaches to meet the methodological challenges mentioned above. However, as there is still a lack of practical guidelines for metabolomics data analysis which incorporates current numerical strategies, we proposed in Chapter 2 an integrated pipeline for in-depth mining of data from pharmacometabolomics studies of interpersonal variations in therapeutic response.

Not only did we cover a variety of routine topics, such as data cleaning, detection of drug exposure signatures, predictive biomarkers, drug response signatures, and validation of results, but also we attempted to probe into the hidden mechanism of drug response variation through deciphering the relationships among biomarkers and pathway analysis. Also discussed are means to connect various types of omics studies for the same cohort of subjects. Although this pipeline is primarily aimed at pharmacometabolomics studies of drug response, it adapts easily to metabolomics studies focused on other phenotypes, e.g., diseases status. This pipeline for metabolomics can be also adopted by other omics studies, given their similarities in data structure and problems of interest.

To meet the high-dimensionality challenge from omics studies while considering their unique biological characteristics, in Chapter 3, we proposed a framework of nonlinear supervised dimension reduction methods. For a large number of predictors and a phenotype under study, our approach is able to capture flexible and complex interaction effects of predictors on the phenotype using a small number of summary features. Therefore, it effectively turn the high-dimensional data of predictors into its most informative low-dimensional projection. The computational advantage of our approach can also effectively handle a very large number of variables, even if it far exceeds the number of observations. Additionally, the proposed methods do not impose any strong parametric model assumptions, and as such they grant full flexibility in subsequent model formulation and selection based on the summary features they generate.

We then demonstrated two important applications of our nonlinear sufficient dimension reduction approach to omics data analysis.

First, it is commonly believed that most biological systems operate not based on a simple gathering of isolated biological units, but instead on their complex mutual interactions and regulations forming biological pathways. Therefore, in Chapter 3, we

proposed a two-step procedure for the purpose of identifying biological pathways that are related to or influence a clinical phenotype of interest. In the first step, the novel nonlinear dimension reduction approach is applied for variables in each pathway. This step essentially condenses the (linear and non linear) complex information of predictors with respect to the phenotype into a small number of summary features for each pathway. In the second step, we build a generalized additive model or a generalized linear model for all the pathways based on the new condensed pathway summary variables to assess the relative importance of pathways jointly. We can produce both a ranking of pathways by employing an L_1 penalty, as well as an explicit pathway selection by proposing a pseudo pathway selection strategy. Simulations demonstrated the theoretical advantages of our approaches over the existing solutions to pathway selection. An analysis of a glioblastoma microarray data using our approach found four pathways related to patients' time to death that have evidence of support from the biological literature.

Second, association tests based on next-generation sequencing data are often underpowered due to presence of rare variants and large amount of neutral or protective variants. A successful strategy is to aggregate genetic information within meaningful SNP-sets, e.g., genes or pathways, and test association on SNP-sets. Many existing methods for group-wise tests require specific assumptions about the direction of individual SNP effects and perform poorly in the presence of interactions, which motivates the application of nonlinear dimension reduction approach to fill this gap in Chapter 4. Accompanying this new joint association test strategy, we also proposed a class of new kernels specifically designed for genotype data, which can boost the power of various kernel-based methods, including, but not limited to, nonlinear dimension reduction and semi-parametric regressions. Our new summary statistic based on nonlinear dimension reduction coupled with the new kernels shows superior performance over existing methods

over various disease models simulated from the sequence data of real genes.

5.2 Future research

5.2.1 Bayesian pathway and gene selection that incorporates network information

In the last section of Chapter 3, we discussed several possible extensions to our dimension reduction-based pathway selection approach. It is desirable to select further the variables within pathways and incorporate network structure among variables. It seems less straightforward to achieve both goals with the classical dimension reduction approaches, but the Bayesian sufficient dimension reduction techniques proposed by Reich *et al.* (2011) offers an alternative avenue toward this aim.

For simplicity, let us consider the case where only one summary feature is to be obtained from each pathway. The idea is to first specify a weighted normal mixture model for the conditional distribution $f(Y|\eta_{(1)}^{\mathsf{T}}X_{(1)},...,\eta_{(G)}^{\mathsf{T}}X_{(G)})$

$$f(Y|\eta_{(1)}^{\mathsf{T}}X_{(1)},...,\eta_{(G)}^{\mathsf{T}}X_{(G)}) = \frac{1}{G}\sum_{g=1}^{G}\sum_{k=1}^{M} P_k(\eta_{(g)}^{\mathsf{T}}X_{(g)})N(\mu_{k(g)},\sigma_Y^2), \tag{5.1}$$

where G is the total number of pathways, $X_{(g)}$ represents variables in the g-th pathway, $\eta_{(g)}$ denotes their coefficients to be estimated. The full range of $\eta_{(g)}^{\mathsf{T}}X_{(g)}$ is divided into M intervals, and we provide a different normal distribution $N(Y|\mu_{k(g)},\sigma_Y^2)$ as a feature distribution for each interval. Denote $\phi_i < ... < \phi_{M+1}$ as the cutpoints that generate the

M intervals, $P_k(\eta_{(g)}^{\intercal} X_{(g)})$ is defined as

$$P_k(\eta_{(g)}^{\intercal} X_{(g)}) = \Phi(\frac{\phi_{k+1} - \eta_{(g)}^{\intercal} X_{(g)}}{\sigma_z}) - \Phi(\frac{\phi_k - \eta_{(g)}^{\intercal} X_{(g)}}{\sigma_z}). \tag{5.2}$$

where Φ represents a cumulative standard normal distribution. In this way, the conditional distribution of Y gives more weight to the feature distributions of intervals that are closer to $\eta_{(g)}^{\intercal} X_{(g)}$. The set of $\mu_{k(g)}$'s for the g-th pathway shape the effect of $\eta_{(g)}^{\intercal} X_{(g)}$ for its full range. In this way, the mixture modeling structure in (5.1) is capable of capturing complex nonlinear effect of pathways.

In order to perform pathway selection, we can use a normal distribution $N(0, \sigma^2_{\mu(g)})$ as the prior distribution of $\mu_{k(g)}$'s, and then adopt stochastic search variable selection(SSVS) via the two component mixture prior (George and McCulloch, 1993)

$$\sigma^2_{\mu(g)} = \pi_{(g)} + e(1 - \pi_{(g)}), \tag{5.3}$$

where $\pi_{(g)} \sim Bern(\bar{\pi})$ with $\bar{\pi}$ being the prior pathway inclusion probability, and $0 < e < 1$ is a small constant. $\pi_{(g)}$ is a binary parameter representing whether pathway g is selected; if it is 1, then the g-th pathway is included in the model, and if it is 0, the prior variance of $\mu_{k(g)}$ is close to 0. Therefore, the g-th pathway is effectively removed from the model.

The network structure can be incorporated during variable selection within pathways. For variable selection within pathways, we assume $\eta_{l(g)}$, the coefficient corresponding to the l-th variable in the g-th pathway, is normal with mean 0 and variance $\sigma^2_{\eta l(g)}$. Then we adopt SSVS again for $\eta_{l(g)}$. That is we use a normal distribution $N(0, \sigma^2_{\eta l(g)})$ as the prior distribution of $\eta_{l(g)}$, and define $\sigma_{\eta l(g)}$ via

$$\sigma_{\eta l(g)} = \pi_{\eta l(g)} + e'(1 - \pi_{\eta l(g)}), \tag{5.4}$$

where $\pi_{l(g)} \sim Bern(\bar{\pi}')$ with $\bar{\pi}$ is the prior variable inclusion probability, and $0 < e' < 1$ is a small constant. In this variable selection procedure, there are two alternative methods to incorporate the network information according to Tai *et al.* (2009): Gaussian Markov random field (GMRF) or Binary Markov Random Field (BMRF). Let $w_{l(g)} = Pr(\pi_{l(g)} = 1)$, and $\theta_{l(g)} = \log(w_{l(g)}/(1-w_{l(g)}))$. A GMRF models the distribution of $\theta_{l(g)}$ conditional on all the rest θ's by

$$\theta_{l(g)}|\theta_{(-l(g))} \sim N(\frac{1}{m_{l(g)}} \sum_{j \in \Delta_{l(g)}} \theta_j, \frac{\tau^2}{m_{l(g)}}), \tag{5.5}$$

where $\theta_{(-i)} = \{\theta_j : j \neq i\}$, $\Delta_{l(g)}$ is the set of indices of direct neighbors of the l-th variable in the g-th pathway, τ is a variance parameter and $m_{l(g)} = |\Delta_{l(g)}|$ is the number of direct neighbors of the l-th variable in the g-th pathway. A BMRF directly regresses $\theta_{l(g)}$ on $T_{l(g)}$:

$$\theta_{l(g)} = \beta_0 + \beta_{l(g)} T_{l(g)}, \tag{5.6}$$

where $T_{l(g)} = (m_{1,l(g)} - m_{0,l(g)})/(m_{1,l(g)} + m_{0,l(g)})$, $m_{1,l(g)} = \sum_{j \in \Delta_{l(g)}} I(\pi_j = 1)$ is the number of selected neighbors of the l-th variable in the g-th pathway, and $m_{0,l(g)} = \sum_{j \in \Delta_{l(g)}} I(\pi_j = 0)$ is the number of unselected neighbors. Both of the two models account for spatial smoothness among the prior probabilities of the variables being selected.

We can also account for pathway cross-talk in the network by specifying GMRF or BMRF priors for $\pi_{(g)}$. The rest prior specification can be adopted from Reich *et al.* (2011) and Tai *et al.* (2009). It is still not clear how to incorporate kernel functions for Bayesian nonlinear dimension reduction, which demands further investigation.

5.2.2 Improving the power of statistical tests for omics data by incorporating additional information from omics databases

A great amount of omics data has been generated over the last two decades from a number of different experiments. Each contains tens or hundreds of subjects, which is usually much smaller than the number of variables, and, therefore, suffers from a lack of statistical power and inflated variance of estimation. Therein, a natural question of interest is whether we can make use of the previously accumulated data to enhance the power of new studies and stabilize their estimation. Consider the common variable selection problem: given an omics dataset, we want to identify biomarkers from a set of candidates that are associated with some phenotype of interest. In addition to the new dataset, we also have a database containing data from a number of previous experiments for the same set of variables. One intuitive approach is to use dimension reduction techniques to summarize the database, and to then integrate this prior information with the new data through Bayesian models.

To illustrate our idea, we take a typical pharmacometabolomics study as an example. Assume that plasma samples of a cohort of patients have been analyzed by a GC/MS platform, which quantified 500 metabolites. These patients suffer from the same disease, and are treated with the same drug; however, only half of these patients have been fully remitted, while the other half have not. The objective is to identify biomarkers among the 500 metabolites that are significantly different between good and poor responders to the specific drug.

All we need from the new experiment data is a statistical test result for each variable. Typically, it can be summarized by a test statistic and a degree of significance, e.g., a

p-value. A Z-score can then be calculated from the test statistic and the p-value. In our example, a two-sample t-test can be applied to every metabolite, which provides a t-statistic t and a p-value P for each metabolite. Then a Z-score can be calculated by

$$Z = \begin{cases} -\Phi^{-1}(1 - P/2) & t < 0 \\ \Phi^{-1}(1 - P/2) & t \geq 0, \end{cases}$$

where again Φ represents a cumulative standard normal distribution.

In addition to the new dataset, we have a metabolomics database, which collects data from a number of old experiments also studying interpersonal variation in drug response with the same analytical platform. In order to summarize information from the database, we use un-supervised dimension reduction techniques, e.g., PCA, to build database *profile* variables from the database. At this stage, metabolites are temporarily treated as individual observations, while samples under various experimental conditions in the database are regarded as explanatory variables. The first D PCA components are chosen as database profile variables according to the amount of overall variance they can explain. The rationale for this proposal is that the experiments in the database are not independent, but instead some of these experiments share very similar objectives and/or study designs. If the database profile variables are able to capture the vast majority of variance in the metabolic data under all different kinds of experimental conditions, then at least some of them should be also able to explain metabolites' behavior in the new experiment as well.

Next, we propose a Bayesian logistic normal mixture model, which utilizes the database profile variables to define prior probabilities of metabolites' being informative. In this way, database information is incorporated to enhance the statistical test results

based on the new dataset. For our example, a three-component logistic normal mixture model for the i-th metabolite is

$$f(Z_i) = \pi_{i,1} f_1(Z_i) + \pi_{i,2} f_2(Z_i) + \pi_{i,3} f_3(Z_i). \tag{5.7}$$

where Z_i is the Z score of the i-th metabolite from marginal test using new experimental data. The three terms in (5.7) represent the three categories of the i-th metabolite respectively: no significant difference between good and poor responders, lower in good responders, and high in good responders. Let T_i=1, 2, or 3 denotes the true category of the i-th metabolite. Then our ultimate goal is to obtain the posterior distributions of T_i's. In (5.7), $f_1 = N(0, \sigma_1^2)$, $f_2 = N(\mu_2, \sigma_2^2)$, and $f_3 = N(\mu_3, \sigma_3^2)$ are conditional distribution functions of Z_i given T_i; μ_2 is forced to be smaller than 0, while it is the opposite for μ_3. Both of them are estimated from the data. $\pi_{i,1} = Pr(T_i = 1)$, $\pi_{i,2} = Pr(T_i = 2)$ and $\pi_{i,3} = Pr(T_i = 3)$ are prior probabilities for T_i. They are modeled by linear functions of database profile variables through the following logit models:

$$\begin{aligned} log(\frac{\pi_{i,2}}{\pi_{i,1}}) &= \sum_{j=0}^{D} \beta_{j,1} \mathbf{D}_{i,j} \\ log(\frac{\pi_{i,3}}{\pi_{i,1}}) &= \sum_{j=0}^{D} \beta_{j,2} \mathbf{D}_{i,j}, \end{aligned} \tag{5.8}$$

where $\mathbf{D}$ is the design matrix of database profile variables with the first column being $\mathbf{1}_p$, and β's are corresponding coefficients to be estimated. Note that we do not pre-determine the influences of profile variables to the prior probability, but instead they are estimated from the new data. The posterior distribution of β's will determine whether their corresponding profile variables have significant effects to the prior probabilities or not. However, we do assume that metabolites that share similar values for the significant

profile variables tend to have similar behavior in the new experiment. To assure the total probabilities sum to one and each estimated probability lies in [0, 1], the following standardization is performed:

$$\begin{aligned}
\pi_{i,1} &= 1/\big(1+exp(\sum_{j=0}^{D}\beta_{j,1}\mathbf{D}_{i,j})+exp(\sum_{j=0}^{D}\beta_{j,2}\mathbf{D}_{i,j})\big) \\
\pi_{i,2} &= exp(\sum_{j=0}^{D}\beta_{j,1}\mathbf{D}_{i,j})/\big(1+exp(\sum_{j=0}^{D}\beta_{j,1}\mathbf{D}_{i,j})+exp(\sum_{j=0}^{D}\beta_{j,2}\mathbf{D}_{i,j})\big) \qquad (5.9) \\
\pi_{i,3} &= exp(\sum_{j=0}^{D}\beta_{j,2}\mathbf{D}_{i,j})/\big(1+exp(\sum_{j=0}^{D}\beta_{j,1}\mathbf{D}_{i,j})+exp(\sum_{j=0}^{D}\beta_{j,2}\mathbf{D}_{i,j})\big)
\end{aligned}$$

Intuitively, given the model specification above for each metabolite, the final decision is made by both its significance from marginal test based on new experiment data, which contributes to the conditional distributions, and its behavior in a number of old experiments, which contribution to the prior distributions. Another interesting observation is that we actually turn the large number of variables from an unfavorable condition into a favorable one, as more variables in the study mean more observations for (5.8) and (5.9), which results in more reliable estimation of the true effects of profile variables.

The model should be fit in a Bayesian framework. Following prior specifications in Wei and Pan (2008), we can give non-informative prior distributions to mean and variance parameters. Particularly, $\mu_2 \sim N(0, 10^6)I(a, 0)$, $a < 0$, $\mu_3 \sim N(0, 10^6)I(0, b)$, $b > 0$, and $\sigma_t^2 \sim \Gamma^{-1}(0.1, 0.1)$ for t=1, 2, and 3, where Γ^{-1} represents inverse gamma distribution. Vague priors are also given to intercepts and coefficients in the logic model: $\beta_{j,k} \sim N(0, 10^6)$ for $j = 0 \sim D$ and $k = 1,$ or 2.

Some simple simulation studies demonstrate the advantage of incorporating the database information, given our assumption is true. We first simulated 10,000 metabo-

lites falling into two alternative categories: 1) no significant difference between good and poor responders or 2) higher in good responders. The true category of the i-th metabolite is randomly drawn from $Categorical([P_{i,1}, P_{i,2}])$. The probabilities $P_{i,1}$ and $P_{i,2}$ are calculated from

$$\begin{aligned} P_{i,1} &= 1/\big(1 + exp(-3 + \sum_{j=1}^{10} \mathbf{D}_{i,j} + e)\big) \\ P_{i,2} &= 1 - P_{i,1}, \end{aligned}$$

where $\mathbf{D}$ is a design matrix of ten independent variables sampled from standard normal distribution, mimicking summarized profile variables from a database, and e is a standard normal random error, which is independent from profile variables and adds to the uncertainty of the true metabolite category given the profile variables. In this settings, about 2,000 metabolites fall in the second category. We simulated 40 good responders and 40 poor responders. Metabolites that are not different between the two classes of subjects are randomly sampled from a standard normal distribution. The rest metabolites in poor responders are sampled from a standard normal distribution, while in good responders they are sampled from $N(0.5, 1)$.

We compare our approach with a standard normal mixture model based on standard t-test (ST), and a standard normal mixture model based on moderated t-test (SMT) (Smyth, 2004) implemented in the LIMMA package. Both of them only use the new experimental data. Accordingly, we build logit normal mixture model based on the two types of t-tests, denoted as LT and LMT respectively. All the four models provide posterior probabilities for metabolite categories. By varying cutoff criteria for these posterior probabilities, we are able to draw receiver operating characteristic (ROC) curves for each method and compare their area-under-the-curve (AUC) scores. Figure 5.1 shows

the results. It can be seen clearly that LT and LMT outperform the other two.

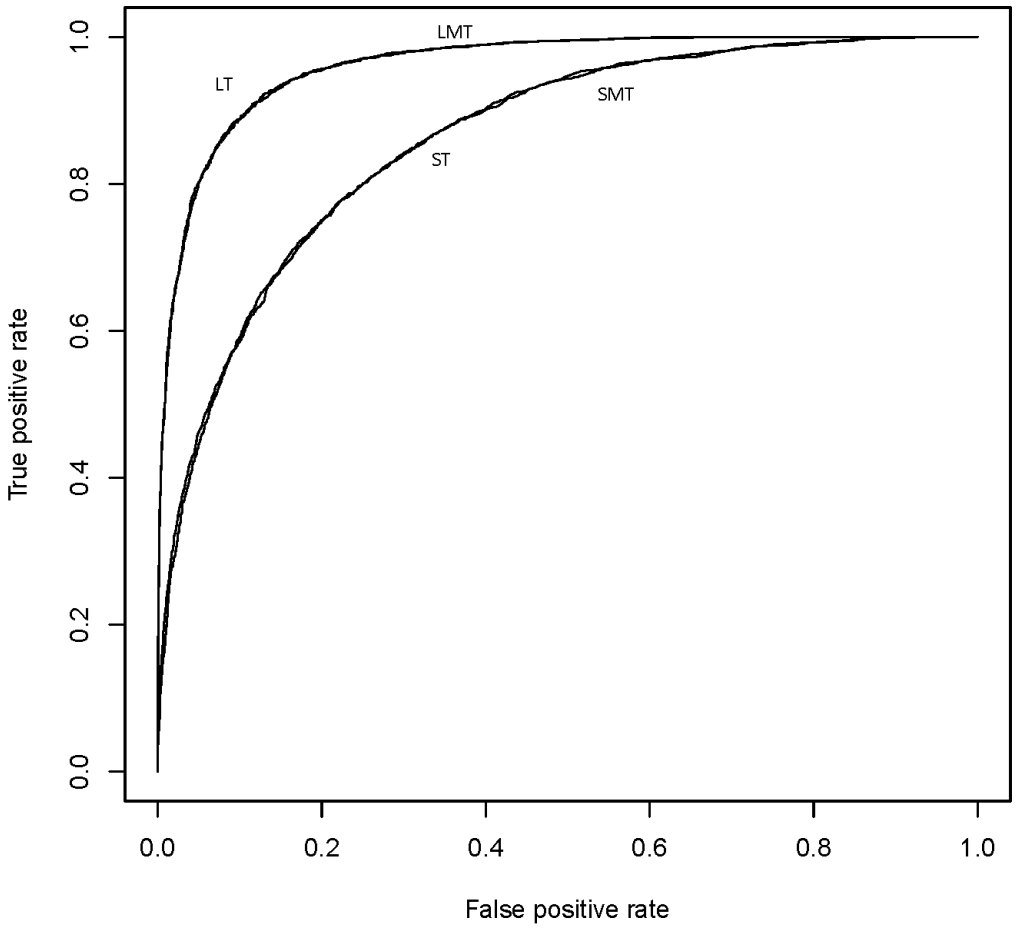

Figure 5.1: ROC curves of ST, SMT, LST and LMT models in the first simulation study. AUC scores of the 4 models are respectively: ST: 0.86; SMT: 0.863; LT: 0.962; LMT: 0.962.

In a second simulation study, we examine the case where only two of the ten database

profile variables are truly informative. Specifically, $P_{i,1}$ and $P_{i,2}$ are now calculated from

$$\begin{aligned} P_{i,1} &= 1/\big(1 + exp(-2 + \sum_{j=1}^{2} X_{i,j} + e)\big) \\ P_{i,2} &= 1 - P_{i,1}. \end{aligned}$$

There are still about 2,000 metabolites in the second category. Other settings remains the same. Results are given in Figure 5.2. It can be seen that even if only a small portion of the database profile variables are informative, LT and LMT are still able to beat ST and SMT, although their performances are significantly decreased compared to the first simulation study. The second simulation study indicates that LT and LMT are not very sensitive to non-informative database profile variables in the modeling.

This approach should be further evaluated using real data, and there are a number of possible avenues for future exploration. First, many other dimension reduction methods, such as principle coordinate analysis, multi-dimensional scaling, self-organized map as well as kernel PCA, can be used in place of PCA and compared. Second, instead of performing dimension reduction on the whole database, we can perform dimension reduction for every old experiment. Bayesian variable selection methods can be then applied to the profile variables representing different experimental conditions, so that the prior distributions are only determined by information retrieved from relevant old experiments.

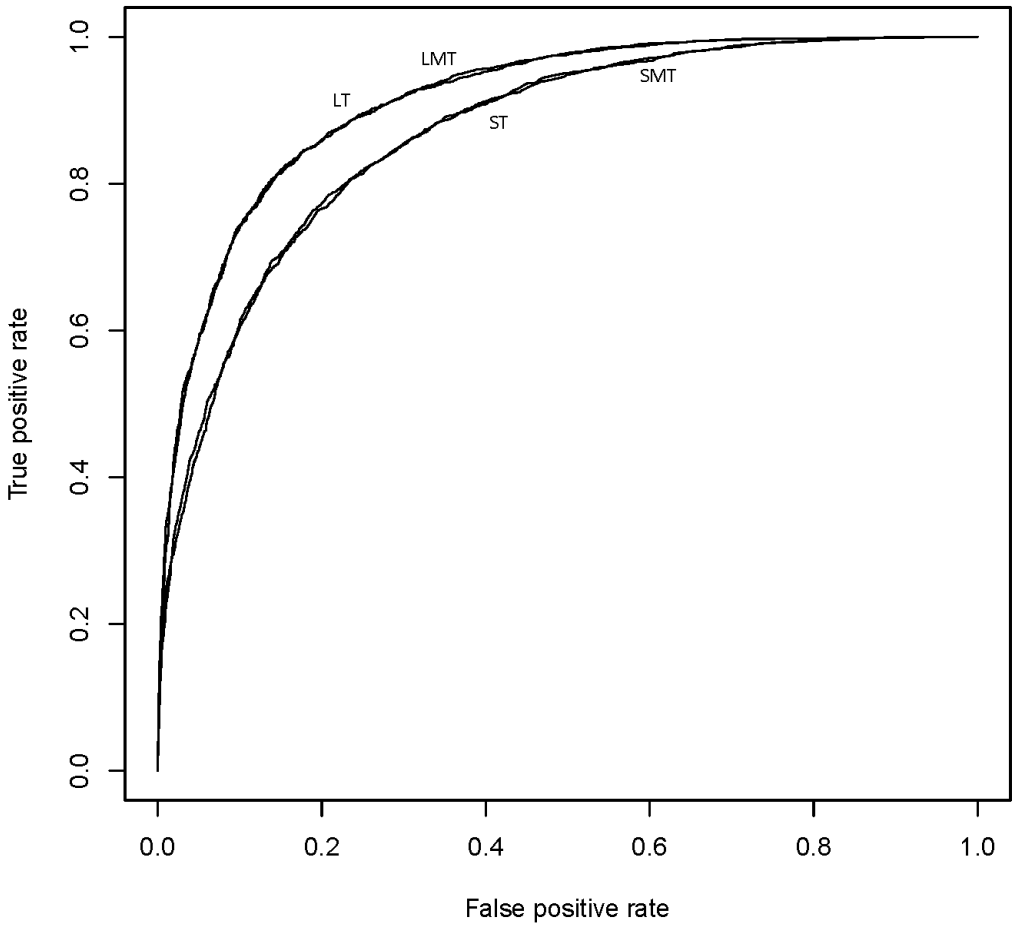

Figure 5.2: ROC curves of ST, SMT, LST and LMT models in the second simulation study. AUC scores of the 4 models are: ST: 0.867; SMT: 0.869; LT: 0.913; LMT: 0.913.

REFERENCES

Ahn, A.C., Tewari, M., Poon, C.S. and Phillips, R.S. (2006). The limits of reductionism in medicine: Could systems biology offer an alternative? *Plos Medicine*, **3**, 709-713.

Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data, *Bioinformatics*, **19**, 563-570.

Ballard, D.H., Cho, J. and Zhao, H.Y. (2010) Comparisons of Multi-Marker Association Methods to Detect Association Between a Candidate Region and Disease. *Genetic Epidemiology*, **34**, 201-212.

Baron R.M., and Kenny D.A.(1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Personal. Soc. Psychol.*, **51**, 1173-1182.

Bellman, R.E.(1961). *Adaptive Control Processes*. Princeton: Princeton University Press.

Bickel, P., Brown, J., Huang, H., and Li, Q. (2009). An overview of recent developments in genomics and the statistical methods that bear on them. Technical report, University of California, Berkeley.

Bogdanov, M., Matson, W.R., Wang, L., Matson, T., Saunders-Pullman, R., Bressman, S.S. and Flint Beal, M. (2008) Metabolomic profiling to develop blood biomarkers for Parkinson's disease. *Brain*, **131**, 389-396.

Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 33.

Breiman, L.(1995).Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, **37**, 373-384.

Breiman, L. (2001). "Random Forests". *Machine Learning*, **45**, 5-32.

Brindle, J.T., Antti, H., Holmes, E., Tranter, G., Nicholson, J.K., Bethell, H.W.L., Clarke, S., Schofield, P.M., McKilligin, E., Mosedale, D.E. et al.. (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using H-1-NMR-based metabonomics. *Nature Medicine*, **8**, 1439-1444.

Brigandt, I., and Love, A. (2008). Reductionism in Biology. *The Stanford Encyclopedia of Philosophy*, Fall 2008 Edition, Edward N. Zalta (ed.), URL = ¡http://plato.stanford.edu/archives/fall2008/entries/reduction-biology/¿.

Camacho, D., de la Fuente, A., Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics*, **1**, 53-63.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *American Journal of Human Genetics*, **79**, 1002-1016.

Chen, L.S., Hutter, C.M., Potter, J.D., Liu, Y., Prentice, R.L., Peters, U. and Hsu, L. (2010). Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data. *Am J Hum Genet*, **86**, 860-871.

Chuang, C. and Cox, C. (1985). Pseudo Maximum-Likelihood Estimation for the Dirichlet-Multinomial Distribution. *Commun Stat-Theor M*, **14**, 2293-2311.

Clarke, R., Ressom, H.W., Wang, A.T., Xuan, J.H., Liu, M.C., Gehan, E.A., and Wang,

Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer*, **8**, 37-49.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2ed ed., Lawrence Erlbaum Associates.

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004) Multiple rare Alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305, 869-872.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983-992.

Cook, R.D. (1998). *Rgression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York.

Cook, R. D., Li, B., and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika*, **94**, 569-584.

Cook, R.D., and Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, **86**, 328-332.

Cook, R.D. and Yin, X.R. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Aust Nz J Stat*, **43**, 147-177.

Davidson, E., and Levin, M. (2005). Gene regulatory networks. *Proc. Natl. Acad. Sci.*, **102**, 4935.

Easton, D.F., and Eeles, R.A. (2008). Genome-wide association studies in cancer. *Human Molecular Genetics*, **17**, 109-115.

Edwards, D.M. (2000). Introduction to Graphical Modelling. New York: Springer.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407-451.

Ewens,W.J. (2004). *Mathematical Population Genetics, I. Theoretical Introduction*, 2nd ed. NewYork: Springer.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, **70**, 949-911.

Filkov, V. (2005). Identifying Gene Regulatory Networks from Gene Expression Data. *Handbook of Computational Molecular Biology*. Chapman and Hall/CRC Press.

Frank, I.E. and Friedman, J.H. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-135.

Frazer, K.A., Murray, S.S., Schork, N.J. and Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. *Nat Rev Genet*, **10**, 241-251.

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

Garthwaite, P.H. (1994). An Interpretation of Partial Least-Squares. *Journal of the American Statistical Association*, **89**, 122-127.

Gelman, A., and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

GeneCards (2011). Retrieved Jun 24, 2011, from `http://www.genecards.org/cgi-bin/carddisp.pl?gene=TIA1`.

Genetics Home Reference (2011). Retrieved Jun 24, 2011, from `http://ghr.nlm.nih.gov/gene/TG`.

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

Gohlmann, H., and Talloen, W. (2009). *Gene Expression Studies Using Affymetrix Microarrays*. Chapman and Hall/CRC press.

Golub, G.H., Heath, M. and Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21, 215-223.

Hamilton, S. P., Garriock, H. A., Kraft, J. B., Shyn, S. I., Peters, E. J., Yokoyama, J. S., Jenkins, G. D., Reinalda, M. S., Slager, S. L. and Mcgrath, P. J. (2010). A Genomewide Association Study of Citalopram Response in Major Depressive Disorder. *Biological Psychiatry*, *67*, 133-138.

Han, X.L., Holtzman, D.M. and McKeel, D.W. (2001) Plasmalogen deficiency in early Alzheimer's disease subjects and in animal models: molecular characterization using electrospray ionization mass spectrometry. *Journal of Neurochemistry*, **77**, 1168-1180.

Hastie, T., Rosset, S., Tibshirani, R., Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5, 1391-1415.

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hastie,T.,Tibshirani, R.,and Friedman, J. (2009).*Element of statistical learning*, 2nd ed. New York: Springer-Verlag.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J. et al. (2007). Genome-wide in situ exon capture for selective resequencing. *Nat Genet*, **39**, 1522-1527.

Holmes, E., Wilson, I.D. and Nicholson, J.K. (2008). Metabolic phenotyping in health and disease. *Cell*, **134**, 714-717.

Holsboer, F., Ising, M., Lucae, S., Binder, E. B., Bettecken, T., Uhr, M., Ripke, S., Kohli, M. A., Hennings, J. M., Horstmann, S., et al. (2009). A Genomewide Association Study Points to Multiple Loci That Predict Antidepressant Drug Treatment Outcome in Depression. *Archives of General Psychiatry*, **66**, 966-975.

Horvath, S., Zhang, B., Carlson, M., Lu, K.V., Zhu, S., Felciano, R.M., Laurance, M.F., Zhao, W., Shu, Q., Lee, Y. *et al.* (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a novel molecular target. *Proceedings of National Academy of Sciences*, **103**, 17402-17407.

Ichinohe, A., Kure, S., Mikawa, S., Ueki, T., Kojima, K., Fujiwara, K., Iinuma, K., Matsubara, Y. And Sato, K. (2004). Glycine cleavage system in neurogenic regions. *European Journal of Neuroscience*, **19**, 2365-2370.

Imai, K., Keele, L. and Tingley, D. (2010). A General Approach to Causal Mediation Analysis. *Psychol Methods*,**15**, 309-334.

International HapMap Consortium(2005). A haplotype map of the human genome. *Nature*, **437**, 1299-1320.

Jansen, J.J., Hoefsloot, H.C.J., Boelens, H.F.M., van der Greef, J. and Smilde, A.K. (2004) Analysis of longitudinal metabolomics data. *Bioinformatics*, **20**, 2438-2446.

Ji, Y., Hebbring, S., Zhu, H., Jenkins, G.D., Biernacka, J., Snyder, K., Drews, M., Fiehn, O., Zeng, Z., Schaid, D. et al.. (2011). Glycine and a Glycine Dehydrogenase (GLDC) SNP as Citalopram/Escitalopram Response Biomarkers in Depression: Pharmacometabolomics-Informed Pharmacogenomics. *Clinical Pharmacology and Therapeutics*, **89**, 97-104.

Ji, W., Fool, J.N., O'Roak, B.J., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., State, M.W., Levy, D. and Lifton, R.P. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*, **40**, 592-599.

Joyce, A.R. and Palsson, B.O. (2006). The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Bio, 7, 198-210.

Kaddurah-Daouk, R., Kristal B.S., and Weinshilboum R.M. (2008). Metabolomics: A Global Biochemical Approach to Drug Response and Disease. *Annu Rev Pharmacol Toxicol*, **48**, 653-683.

Kaddurah-Daouk, R., McEvoy, J., Baillie, R.A., Lee, D., Yao, J.K., Doraiswamy, P.M. and Krishnan, K.R.R. (2007) Metabolomic mapping of atypical antipsychotic effects in schizophrenia. *Mol Psychiatr*, **12**, 934-945.

Kaddurah-Daouk, R., Soares, J.C., Quinones, M.P.(2009). Metabolomics: A Global Biochemical Approach to the Discovery of Biomarkers for Psychiatric Disorders. Springer Science 2009.

Kaddurah-Daouk, R., and Krishnan, K.R.(2009). Metabolomics: a global biochemical approach to the study of central nervous system diseases. *Neuropsychopharmacology.* **34**, 173-186.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**, 27-30.

Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L. , Kaipa, P., Ahren, D.,Tsoka, S., Darzentas, N.,Kunin, V., and Lopez-Bigas, N. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, **19**, 6083-89.

Kato, M., and Serretti, A. (2008). Review and meta-analysis of antidepressant pharmacogenetic findings in major depressive disorder. *Mol Psychiatry*, **15**, 473-500.

Khare, K., and Zhou, H. (2009). Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. *Ann. Appl. Probab*, **19**, 737C777.

Kikuchi, G., Motokawa, Y., Yoshida, T. and Hiraga, K. (2008). Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycinemia. *Proc Jpn Acad Ser B Phys Biol Sci.*, **84**, 246-263.

Kim, S., Sohn, K. A. and Xing, E. P. (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*. **25**, 204-212.

Kristal, B.S., Kaddurah-Daouk, R., Beal, M.F., and Matson, W.R. (2007). Metabolomics: concept and potential neuroscience application. In Handbook of Neurochemistry and Molecular Neurobiology: Brain Energetics. Integration of Molecular and Cellular Processes, pp. 889-912.

Kristal, B.S., Shurubor, Y.I., Kaddurah-Daouk, R., and Matson, W.R.(2007). High-performance liquid chromatography separations coupled with coulometric electrode array detectors: a unique approach to metabolomics. *Methods Mol Biol*, **358**, 159-174.

Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. and Sunyaev, S.R. (2009) Power of deep, all-exon resequencing for discovery of human trait genes. *P Natl Acad Sci USA*, **106**, 3871-3876.

Kwee, L.C., Liu, D., Lin, X., Ghosh, D. and Epstein, M.P. (2008). A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet*, **82**, 386-397.

Latchman, David S. (2005). *Gene regulation: a eukaryotic perspective*. Psychology Press.

Lettre, G. and Rioux, J.D. (2008) Autoimmune diseases: insights from genome-wide association studies. *Hum. Mol. Genet.*, **17**, ddn246+.

Li, B. and Leal, S.M.M. (2008) Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *The American Journal of Human Genetics*, **83**, 311-321.

Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316-327.

Li, K.C. (2002) Genome-wide coexpression dynamics: Theory and application. *P Natl Acad Sci USA*, **99**, 16875-16880.

Li, B., and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**, 997-1008.

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

Li, L. (2009). Exploiting predictor domain information in sufficient dimension reduction. *Computational Statistics and Data Analysis*, **53**, 2665-2672.

Li, L., Cook, R.D., and Tsai, C.L.(2007). Partial inverse regression. *Biometrika*, **94**, 615-625.

Li, L., and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, bf 64, 124-131.

Lindon, J.C., Holmes, E., and Nicholson, J.K. (2007). Global systems biology through integration of "omics" results. *The Handbook of Metabonomics and Metabolomics*. Elsevier Science.

Liu, D. and Leal, S.M. (2010). A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *Plos Genetics*, **6**, e1001156.

Liu, Z., Gartenhaus, R.B., Tao, M., and Jiang, F. (2008). Gene and pathway identification with Lp penalized Bayesian logistic regression. *BMC Bioinformatics*, **9**:412.

Luan, Y., and Li, H. (2008). Group additive regression models for analysis of genomic data. *Biostatistics*, **9**, 100-113.

Ma, S., and Kosorok, M.R. (2009). Identification of differential gene pathways with principal component analysis. *Bioinformatics*, **25**, 882-889.

Madsen, B.E. and Browning, S.R. (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *Plos Genetics*, **5**, e1000384.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., De Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., D'Eustachio, P.(2008).

Reactome knowledgebase of biological pathways and processes. *Nucleic Acids Research*, **37**, 619-622.

Mawrin, C., Diete, S., Treuheit, T., Kropf, S., Vorwerk, C.K., Boltze, C., Kirches, E., Firsching, R., and Dietzmann, K. (2003). Prognostic relevance of MAPK expression in glioblastoma multiforme. *International Journal of Oncology*, **33**, 641-648.

Mccullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

McNeil, J.B., McIntosh, E.M., Taylor, B.V., Zhang, F.R., Tang, S. and Bognar, A.L. (1994). Cloning and molecular characterization of three genes, including two genes encoding serine hydroxymethyltransferases, whose inactivation is required to render yeast auxotrophic for glycine. *J Biol Chem.*, **269**, 9155-9165.

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

Mrazek, D.A., Rush, A.J., Biernacka, J.M., O'Kane, D.J., Cunningham, J.M., Wieben, E.D., Schaid, D.J., Drews, M.S., Courson, VL, Snyder, K.A., Black, J.L., Weinshilboum, R.M. (2009). SLC6A4 variation and citalopram response. *Am J Med Genet B Neuropsychiatr Genet*, 2009 **150**, 341-51.

Muller-Linow, M., Weckwerth, W., Hutt, M.T. (2007). Consistency analysis of metabolic correlation networks. *BMC Syst Biol*, 1:44.

Nejentsev, S., Walker, N., Riches, D., Egholm, M. and Todd, J.A. (2009) Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. *Science*, **324**, 387-389.

Nguyen, D. and Rocke, D. (2002a). Muti-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**, 1216-1226.

Nguyen, D. and Rocke, D. (2002b). Classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39-50.

Nicholson, J.K., Holmes, E., and Lindon, J.C. (2007). Metabonomics and metabolomics techniques and their applications in mammalian systems. *The Handbook of Metabonomics and Metabolomics*. Elsevier Science.

Ogiichi, T., Hirashima, Y., Nakamura, S., Endo, S., Kurimoto, M., and Takaku, A. (2000). Tissue factor and cancer procoagulant expressed by glioma cells participate in their thrombin-mediated proliferation. *Journal of Neuro-Oncology*, **46**, 1-9.

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

Pang, H., and Zhao, H. (2008). Building pathway clusters from random forests classification using class votes. *BMC Bioinformatics*, **9**, 87.

Patterson, S.D., and Aebersold, R. H. (2003). Proteomics: the first decade and beyond. *N*ature Genet., **33**, 311-323.

Pelloski, C.E., Lin, E., Zhang, L., Yung, W.K.A., Colman, H., Liu, J., Woo, S.Y., Heimberger, A.B., Suki, D., Prados, M., Chang, S., Barker, F.G., Fuller, G.N., and Aldape, K.D. (2006). Prognostic associations of activated mitogen-activated protein kinase and akt Pathways in glioblastoma. *Clinical Cancer Research*, **12**, 3935-3941.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009) Partial correlation estimation by joint sparse regression models, *Journal of American Statistical Association*, **104**, 735-746.

Price, A.L., Kryukov, G.V., de Bakker, P.I.W., Purcell, S.M., Staples, J., Wei, L.J. and Sunyaev, S.R. (2010). Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am J Hum Genet*, **86**, 832-838.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, **38**, 904-909.

Qiu, X., Brooks, A.I., Klebanov, L., and Yakovlev, N. (2005). The effects of normalization on the correlation structure of microarray data. *BMC Bioinform*,6:120.

Raghunathan, T.E., Rosenthal, R., and Rubin, D.B. 1996. Comparing correlated but nonoverlapping correlations. *Psychological Methods*, **1**, 178-183.

Reich, B.J., Bondell, H.D., and Li, L. (2011). Sufficient dimension reduction via Bayesian mixture modeling. *Biometrics*, In press.

Rhodes, D.R., and Chinnaiyan, A.M. (2005). Integrative analysis of the cancer transcriptome. *Nature Genet*, **37**, 31C37.

Roth, F., Lipshitz, H., and Andrews, B. (2009). Q and A: Epistasis. *J. Biol.*, *8*:35

Schaid, D.J. (2010) Genomic similarity and kernel methods I: methods for genomic information. *Hum Hered*, **70**, 132-140.

Schaid, D.J. (2010) Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered, 70, 109-131.

Scholkopf, B., and Smola, A. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press.

Scott, D.W. (1992). *Multivariate Density Estimation*. John Wiley; Hoboken.

Shi, M., and Ma, S. (2008). Identifying subset of genes that have influential impacts on cancer progression: a new approach to analyze cancer microarray data.*Funct. Integr. Genomics*, **8**, 361-373.

Serretti, A., and Artioli, P. (2004). The pharmacogenomics of selective serotonin reuptake inhibitors. *Pharmacogenomics J.*, **4**, 233-244.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, No. 1, Article 3.

Steuer, R., Kurths, J., Fiehn, O., Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, **19**, 1019-1026.

Stone E.A., Ayroles J.F.(2009). Modulated Modularity Clustering as an Exploratory Tool for Functional Genomic Inference. *PLoS Genet.*, **5**, e1000479.

Storey, Jd, and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440-9445.

Strange, K. (2005). The end of "naive reductionism": rise of systems biology or renaissance of physiology? *Am J Physiol-Cell Ph*, **288**, 968-974.

Strachan., T., and Read, A. (2010). *Human molecular genetics*, 4th ed. Garland Science.

Subramanian, A, Tamayo, P, Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005).

Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.*Proceedings of the National Academy of Sciences*, **102**, 15545-15550.

Tai, F., Pan, W., Shen, X. (2009). Bayesian Variable Selection in Regression with Networked Predictors. Technical report, University of Minnesota, Minneapolis.

Takeshima, H., Nishi, T., Kuratsu, J., Kamikubo, Y., Kochi, M., and Ushio, Y. (2000). Suppression of the tissue factor-dependent coagulation cascade: A contributing factor for the development of intratumoral hemorrhage ire glioblastoma. *Internaltional Journal of Molecular Medicine*, **6**, 271-276.

Tian, L, Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S., and Park, P.J.(2005) Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, **102**, 13544-13549.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *J Roy Stat Soc B*, **67**, 91-108.

The 1000 Genomes Project Consortium . A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25-9.

Trygg, J., and Lundstedt, T. (2007). Chemometrics Techniques for Metabonomics. *The Handbook of Metabonomics and Metabolomics*. Elsevier Science.

Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. and Shendure, J. (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods*, **6**, 315-316.

Turner, S.R., Ireland, R., Morgan, C., and Rawsthorne, S. (1992). Identification and localization of multiple forms of serine hydroxymethyltransferase in pea (Pisum sativum) and characterization of a cDNA encoding a mitochondrial isoform. *J Biol Chem.*, **267**, 13528-13534.

Uher, R., Perroud, N., Ng, M. Y. M., Hauser, J., Henigsberg, N., Maier, W., Mors, O., Placentino, A., Rietschel, M., Souery, D., et.al. (2010). Genome-Wide Pharmacogenetics of Antidepressant Response in the GENDEP Project. *American Journal of Psychiatry*, **167**, 555-564.

Vapnik, V. (1996). The Nature of Statistical Learning. Springer-Verlag.

Wagner, M., Naik, D., and Pothen, A. (2003). Protocols for disease classification from mass spectrometry data. *Proteomics*, **3**, 1692-1698.

Wang W.Y., Barratt B.J., Clayton D.G., and Todd J.A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, **6**, 109-118.

Wei, P. and Pan, W. (2008). Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, **24**,404-411.

Wei, Z., and Li, H. (2007). Nonparametric pathways-based regression models for analysis of genomic data. *Biostatistics*, **8**, 265-284.

Wessel, J. and Schork, N.J. (2006) Generalized Genomic Distance-Based Regression

Methodology for Multilocus Association Analysis. *The American Journal of Human Genetics*, **79**, 792 - 806.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, New York: Academic Press.

Wu, H.M. (2008). Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics*, **17**, 590-610

Wu, M., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J. and Lin, X.H. (2010). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *Am J Hum Genet*, **86**, 929-942.

Wu, Q., Liang, F., and Mukherjee, S. (2008). Regularized sliced inverse regression for kernel models. Technical report, Duke University.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007). Controlling Variable Selection by the Addition of Pseudovariables. *Journal of the American Statistical Association*, **477**, 235-243.

Yao, J.K., Dougherty, G.G., Jr., Reddy, R.D., Keshavan, M.S., Montrose, D.M., Matson, W.R., McEvoy, J. and Kaddurah-Daouk, R. (2010) Homeostatic imbalance of purine catabolism in first-episode neuroleptic-naive patients with schizophrenia. *PLoS One*, **5**, e9508.

Yao, J.K., Dougherty, G.G., Jr., Reddy, R.D., Keshavan, M.S., Montrose, D.M., Matson, W.R., Rozen, S., Krishnan, R.R., McEvoy, J. and Kaddurah-Daouk, R. (2010). Altered interactions of tryptophan metabolites in first-episode neuroleptic-naive patients with schizophrenia. *Mol Psychiatry*, **15**, 938-953.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, **68**, 49-67.

Yuan, M., and Lin, Y. (2007).On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, **69**,143-161.

Yuan, M. and Lin, Y. (2007) Model selection and estimation in the Gaussian graphical models, Biometrika, 94, 19-35.

Zhang, H.H. (2006). Variable selection for support vector machines via smoothing spline ANOVA. *Statistica Sinica*, **16**, 659-674.

Zhou, H., and Lange, K. (2009). Composition Markov chains of multinomial type. *Adv. in Appl. Probab.*, **41**, 270-291.

Zou, H., and Hastie, T. (2006). Regularization and variable selection via the elastic net. *J.R. Statist. Soc. B*, **67**, 301-320.

Zhu, L.P., Wang, T., Zhu, L.X., and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, **97**, 295-304.

Zhu, Z. and Liu, Y. (2009). Estimating spatial covariance using penalized likelihood with weighted L1 penalty. *Journal of Nonparametric Statistics*, **21**, 925-942.

APPENDIX

Appendix A

Derivation of the Kernel Sufficient Dimension Reduction Estimators

We outline the derivation of the kernel sufficient dimension reduction estimators given in Section 3.2.2. Theoretical arguments will be carried out in terms of the matrix form for simplicity, assuming the dimension p_ϕ of the kernel map ϕ is finite. When ϕ's dimension is infinite, the notion of matrix needs to be replaced by an operator. A rigorous treatment can be obtained in analogous to Wu *et al.* (2008), and the details are omitted here.

Denote Φ as an $n \times p_\phi$ matrix with its i-th row as $\phi(x_i)$, $i = 1, \ldots, n$. Our goal here is to establish the relation between (3.6) and (3.7), and the key is the equation that $K = \Phi\Phi^\intercal$. Define $L_n = I_n - n^{-1}\mathrm{l}_n\mathrm{l}_n^\intercal$, where l is a vector of ones with length n. Right multiplying L_n to a matrix is to perform a centering by the columns of that matrix. Besides, the centered gram matrix $\tilde{K} = L_n K L_n$. We first note that $\hat{\Sigma}_\phi = n^{-1}(\Phi^\intercal L_n)(\Phi^\intercal L_n)^\intercal$, and that β_j in (3.6) and α_j in (3.7) are connected by the equation $\beta_j = \Phi^\intercal L_n \alpha_j$. Then (3.6)

can be written as

$$n\,\hat{\Omega}_\phi \Phi^\intercal L_n \alpha_j = \rho_j \Phi^\intercal L_n L_n \Phi\Phi^\intercal L_n \alpha_j = \rho_j \Phi^\intercal L_n \tilde{K} \alpha_j.$$

Multiplying $L_n\Phi$ on both sides yields

$$n\,L_n\Phi\hat{\Omega}_\phi \Phi^\intercal L_n \alpha_j = \rho_j \tilde{K}^2 \alpha_j.$$

Next we derive $nL_n\Phi\hat{\Omega}_\phi\Phi^\intercal L_n$ on the left hand side in terms of $\tilde{K}J\tilde{K}$ for different SDR estimators. Following the usual protocol, we partition the response Y into H non-overlapping intervals.

For sliced inverse regression (SIR),

$$\begin{aligned}
\hat{\Omega}_{\phi\mathsf{SIR}} &= \sum_{h=1}^{H} \frac{n_h}{n} \left\{\hat{E}(\Phi|Y=h) - E(\Phi)\right\} \left\{\hat{E}(\Phi|Y=h) - E(\Phi)\right\}^\intercal \\
&= \frac{1}{n}\Phi^\intercal L_n \left(\sum_{h=1}^{H} \frac{1}{n_h} b_h b_h^\intercal\right) L_n \Phi = \frac{1}{n}\Phi^\intercal L_n J_{\mathsf{SIR}} L_n \Phi,
\end{aligned}$$

where b_h and J_{SIR} are as defined in Section 3.2.2. Then $nL_n\Phi\hat{\Omega}_{\phi\mathsf{SIR}}\Phi^\intercal L_n = \tilde{K}J_{\mathsf{SIR}}\tilde{K}$.

For sliced average variance estimation (SAVE),

$$\begin{aligned}
\hat{\Omega}_{\phi\mathsf{SAVE}} &= \sum_{h=1}^{H} \frac{n_h}{n} \left\{\widehat{\mathrm{Cov}}(\Phi) - \widehat{\mathrm{Cov}}(\Phi|Y=h)\right\}^2 \\
&= \sum_{h=1}^{H} \frac{n_h}{n} \Phi^\intercal \left(\frac{1}{n}L_n - \frac{1}{n_h}P_h^\intercal L_{n_h} P_h\right) \Phi\Phi^\intercal \left(\frac{1}{n}L_n - \frac{1}{n_h}P_h^\intercal L_{n_h} P_h\right) \Phi
\end{aligned}$$

where P_h is an $n_h \times n$ matrix that consists of all the non-zero rows of the diagonal matrix $\mathrm{diag}(b_h)$, and L_{n_h} is defined in the same way as L_n. We can further rewrite

$\sqrt{n_h}\left(\frac{1}{n}L_n - \frac{1}{n_h}P_h^\mathsf{T} L_{n_h} P_h\right) = n^{-1}n_h^{1/2}I_n - n^{-2}n_h^{1/2}\mathbf{1}_n\mathbf{1}_n^\mathsf{T} - n_h^{-1/2}\mathrm{diag}(b_h) + n_h^{-3/2}b_h b_h^\mathsf{T} \equiv B_h$ as defined in Section 3.2.2, and we note that $L_n B_h L_n = B_h$. Then $nL_n\Phi\hat{\Omega}_{\phi\mathsf{SAVE}}\Phi^\mathsf{T}L_n = \tilde{K}\left(\sum_{h=1}^H B_h\tilde{K}B_h\right)\tilde{K} = \tilde{K}J_{\mathsf{SAVE}}\tilde{K}$.

For directional regression (DIR),

$$\begin{aligned}
\hat{\Omega}_{\phi\mathsf{DIR}} &= 2\hat{\Omega}_{\phi\mathsf{SAVE}} + 2\hat{\Omega}_{\phi\mathsf{SIR}}{}^2 + \\
&\qquad 2\sum_{h=1}^H \frac{n_h}{n}\left\{\hat{E}(\Phi|Y=h) - E(\Phi)\right\}^\mathsf{T}\left\{\hat{E}(\Phi|Y=h) - E(\Phi)\right\}\ \hat{\Omega}_{\phi\mathsf{SIR}} \\
&= 2\hat{\Omega}_{\phi\mathsf{SAVE}} + 2\hat{\Omega}_{\phi\mathsf{SIR}}{}^2 + \frac{2}{n}\sum_{h=1}^H \frac{1}{n_h}\left(b_h^\mathsf{T}L_n\Phi\right)\left(\Phi^\mathsf{T}L_n b_h\right)\ \hat{\Omega}_{\phi\mathsf{SIR}} \\
&= 2\hat{\Omega}_{\phi\mathsf{SAVE}} + 2\hat{\Omega}_{\phi\mathsf{SIR}}{}^2 + \frac{2}{n}\sum_{h=1}^H \frac{1}{n_h}b_h^\mathsf{T}\tilde{K}b_h\ \hat{\Omega}_{\phi\mathsf{SIR}}
\end{aligned}$$

Then $nL_n\Phi\hat{\Omega}_{\phi\mathsf{DIR}}\Phi^\mathsf{T}L_n = 2\tilde{K}J_{\mathsf{SAVE}} + 2n^{-1}J_{\mathsf{SIR}}\tilde{K}J_{\mathsf{SIR}} + 2n^{-1}\left(\sum_{h=1}^H n_h^{-1}b_h^\mathsf{T}\tilde{K}b_h\right)J_{\mathsf{SIR}} = \tilde{K}J_{\mathsf{DIR}}\tilde{K}$.