Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Hesen Peng Date

High-dimensional Universal Dependence Discovery

By

Hesen Peng
Doctor of Philosophy

Biostatistics and Bioinformatics

Tianwei Yu, Ph.D.
Adviser

Vicki Hertzberg, Ph.D.
Committee Member

Zhaohui Qin, Ph.D.
Committee Member

Glen Satten, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

Date

High-dimensional Universal Dependence Discovery

By

Hesen Peng
B.S., Fudan University, 2008

Adviser: Tianwei Yu, Ph.D.

An Abstract of
A dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biostatistics
2012

Abstract

High-dimensional Universal Dependence Discovery

By Hesen Peng

The emergence of high-throughput data in biological science and computer networks has generated novel challenges for statistical methods. Nonlinear relationships and multivariate interactions are abundant. The sheer volume of high-throughput data has limited the application for traditional case-by-case analysis methods, whose model assumptions, like linearity, are often not supported in high-throughput scenarios.

To meet these challenges, we developed Mira score, a novel probabilistic association statistic that accounts for high-dimensional universal dependence. Mira score is defined as a function of observation graph, and thus circumvents the curse of dimensionality in high-dimensional data. The superior statistical property enjoyed by Mira score has led to our development of an efficient network reverse-engineering procedure for multivariate dependence. As an example, the procedure has been applied to celiac disease and lung cancer pathway interaction analysis, and has achieved interesting findings.

Further more, in the supervised-machine learning scenario, we proposed SeMira procedure, an efficient variable selection procedure that accounts for high-dimensional universal dependence. The SeMira procedure is capable of identifying universal probabilistic association between multivariate response variables and high-dimensional predictors. The highly desirable statistical property of the SeMira procedure is discussed and numerical study is conducted using both simulated and real genetic pathway data.

High-dimensional Universal Dependence Discovery

By Hesen Peng

B.S., Fudan University, 2008

Adviser: Tianwei Yu, Ph.D.

A dissertation submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biostatistics

2012

To Guangguang

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 High-dimensional nonlinear associations

The emergence of high-throughput data in biology and computer science has motivated the proposal of *universal dependence* statistic. *Universal dependence* is defined to be capable of accounting for probabilistic association of arbitrary type between arbitrary number of variables. For example, association study in social network analysis aims at identifying the factors that affects user behaviors in social network, whose interactions involve the complicated interaction of multiple random variables and may go well beyond linear functions (Boyd and Ellison [2008]). Functional MRI analysis aims at locating interrelated brain regions connected to disease symptoms. The response for each brain region is represented using multivariate time series through fMRI observation. And major signals have been reported to follow nonlinear relationships with excitation input (Toyoda et al. [2008], Johnston et al. [2008], Gautama et al. [2003]). Genetic pathway analysis aims at discovering higher-level gene expression interactions between pathways defined by existing biological findings. And gene expression dependency between pathways may feature the co-regulation of multiple

genes (Li [2002], Li et al. [2004]) and nonlinear relationships (Hasty et al. [2001], Ritchie et al. [2003]).

A plethora of statistical methods have been proposed to meet the challenge of high-dimensional nonlinear dependence, and might encompass the numerous fields of research (see Hastie et al. [2009] for a comprehensive survey). *Dimension reduction* extracts summary statistics through a linear combination of multiple variables (Zou et al. [2006], Tamayo et al. [2007], Mashal et al. [2005]), but may discard relevant information by ignoring secondary components. *Mutual information* by Margolin et al. [2006] has been widely applied in the discovery of nonlinear dependence, but suffer from curse of dimensionality when the nonlinear dependence between more than two variables are considered. Liquid association is an innovative method developed to study the interaction involving three and more gene expressions (Li et al. [2004], Li [2002]) of specific interaction types. Brownian covariate is able to account for universal types of dependencies (Szkely and Rizzo [2009]), but lacks the expandability into efficient variable selection procedure.

Great advancement in variable selection strategy has been made in the recent decade. The Lasso family and related penalized regression methods (Tibshirani [1996], Zou [2006], Fan and Lv [2008], Fan and Li [1999], Hastie and Efron [2007]) are capable of efficiently selecting variables of linear association in high-dimensional scenario. Slice inverse regression is capable of conducting variable selection that accounts for functional nonlinear association between response variable and linear combination of predictors (Li [1991], Ferre [1998]). Numerous pairwise mutual-information-based heuristic approach have been proposed to account for nonlinear association (Peng et al. [2005], Durand et al. [2007], May et al. [2008]). However, no method upon our literature survey has the capability of conducting variable selection for universal dependence in high-dimensional data.

1.2 Road map towards universal dependence discovery

Before proceeding, we would like to give a more mathematically explicit definition of the problem. Given random vectors $(X_1, \ldots, X_p)$ and $(Y_1, \ldots, Y_q)$, we are interested in making statistical inferences between these two random vectors given n pairs of independent samples. In this dissertation, we endeavor to develop a suit of statistical methods that is capable to

1. identify probabilistic dependence of *arbitrary type*,
2. detect probabilistic dependence of *arbitrary dimension*.

In this dissertation, we have focused on the development of universal dependence discovery methods for *continuous variables*. Universal dependence discovery and inference methods involving categorical variables has been briefly investigated during our research and will be presented only in future essays. Our work can be outlines in two stages.

In Chapter 2, we will present our work on Mira score. Mira score is a universal dependence association statistic that is capable to *identify* continuous probabilistic dependence of arbitrary type involving arbitrary number of variables. The Mira score is defined as 1-nearest neighbor edge sum of the observation graph. This graph-based definition will be capable of circumventing the curse of dimensionality, and account for multivariate universal associations. Mira score enjoys asymptotic normality in large samples. We proposed Gaussian plug-in permutation test of association test the existence of universal association.

In Chapter 3, we extended out work on Mira score and proposed SeMira procedure for *variable selection*. SeMira procedure is capable to conduct variable selection on

high-dimensional predictors having universal association with multivariate responses. SeMira has computational complexity of $O(n^2p)$, which means that the computational burden of variable selection while accounting for universal dependence only increase linearly with the number of predictors.

On the application side, the Mira score permutation test and SeMira procedure have been applied to biological data. We developed network reverse engineering algorithm based on Mira permutation test in Chapter 2 to investigate changes in pathway interactions in disease state compared with normal state. The algorithm has been applied to lung cancer data set (NCBI data set GDS2771) and celiac disease data set (NCBI data set GDS3646) to reveal differentially interacting pathways in disease state. Our findings have been evaluated as consistent with biological mechanism through literature survey. In Chapter 3, we applied the SeMira procedure to identify genetic pathways that are associated with the white blood cell counts in primary acute lymphoblastic leukemia (ALL) study with methotrexate treatment. The majority of identified pathways are related to the protein creation process.

Chapter 2

High-dimensional Universal Dependence Statistic

2.1 Introduction

Given random vectors $(X_1, \ldots, X_p)$ and $(Y_1, \ldots, Y_q)$, we are interested in testing the existence of probabilistic dependence between these two vectors given n pairs of independent samples. In this chapter we propose the Mira score, a universal dependence statistic that is capable of

1. identifying probabilistic dependence of *arbitrary type*,
2. detecting probabilistic dependence of *arbitrary dimension*.

Universal dependence statistic has been motivated by the emergence of high-throughput data in biology and computer science. Association studies in social network analysis aims at identifying the factors that affects user behaviors in social network, whose interaction may go well beyond linear functions (Boyd and Ellison [2008]). Functional MRI analysis aims at locating interrelated brain regions connected to disease

symptoms, where major signals have been reported to follow nonlinear relationships with excitation input (Toyoda et al. [2008], Johnston et al. [2008], Gautama et al. [2003]). Pathway analysis aims at discovering higher-level gene expression interactions between pathways defined by existing biological findings. And gene expression dependency between pathways may feature the co-regulation of multiple genes (Li [2002], Li et al. [2004]) and nonlinear relationships (Hasty et al. [2001], Ritchie et al. [2003]).

A plethora of statistical methods have been proposed to meet the challenge of high-dimensional nonlinear dependence (see Hastie et al. [2009] for a comprehensive survey). Dimension reduction extracts summary statistics through a linear combination of multiple variables (Zou et al. [2006], Tamayo et al. [2007], Mashal et al. [2005]), but may discard relevant information by ignoring secondary components. Mutual information by Margolin et al. [2006] has been widely applied in the discovery of nonlinear dependence, but suffer from curse of dimensionality when the nonlinear dependence between more than two variables are considered. Liquid association is an innovative method developed to study the interaction involving three and more gene expressions (Li et al. [2004], Li [2002]) of specific interaction types. Brownian covariate is able to account for universal types of dependencies (Szkely and Rizzo [2009]), but lacks the expandability into efficient variable selection procedure.

We endeavor to address the problem of high dimensional universal probabilistic dependence discovery from a graph theory approach. This computationally efficient approach circumvents the necessity of joint probability density estimation and does not suffer from the curse of dimensionality. Thus we name the statistic as "Mira" score, short for the self-referential phrase *Mira Is Really Adaptive*.

2.2 Theory

Denote n-pairs of random samples as $\{x_{i1}, \ldots, x_{ip}, y_{i1}, \ldots, y_{iq}\}$, with $i = 1, \ldots, n$ where x_{ij}, y_{ik} are the observation for X_j and Y_k in the i-th pair of observation, respectively. In the $\mathcal{R}^{p+q}$ space, we define a complete graph with n vertices where the coordinates for the i-th vertex equals to $(x_{i1}, \ldots, y_{iq})$. We define sample distance matrix $D = (d_{ij})$ with inter-observation distance $d_{ij} = \sum_{l=1}^{p} |x_{il} - x_{jl}| + \sum_{k=1}^{q} |y_{ik} - y_{jk}|$ for $i, j = 1, \ldots, n$. Then Mira score is defined as the 1-nearest neighbor edge sum for the observation graph. More specifically,

$$S = \sum_{i=1}^{n} d_{(i)} \tag{2.1}$$

where $d_{(i)} = \min_{j \neq i} d_{ij}$ is the nearest neighbor edge length for the i-th observation. To put each variable on a comparable scale, we assume that samples for each variable have been standardized using the normal Gaussian score transformation with mean 0 and unit variance. The n observations of the i-th variable are compared to obtain the ranks $r_{i1}, \ldots, r_{in}$. And then each x_{ij} of the vector is replaced by $\Phi^{-1}(r_{ij}/(n+1))$, where $\Phi(\cdot)$ is the cumulative normal distribution (Yu et al. [2011]). Here d_{ij} is defined using Manhattan distance for the ease of mathematical deduction of our working paper, in which an efficient variable selection procedure for universal dependency is proposed. We would like to point out that the definition of d_{ij} may be substituted with Euclidean or any other reasonable distance metrics.

Given $(X_1, \ldots, X_p)$ and $(Y_1, \ldots, Y_q)$ of fixed marginal distribution, samples from joint distribution with inter-group probabilistic dependence may expect a smaller 1-nearest neighbor edge sum than samples from the independent case. This inspiration for Mira score comes from observing graphical distributions in 2-dimension cases. For illustration consider a bivariate normal case where (X, Y) follow bivariate normal

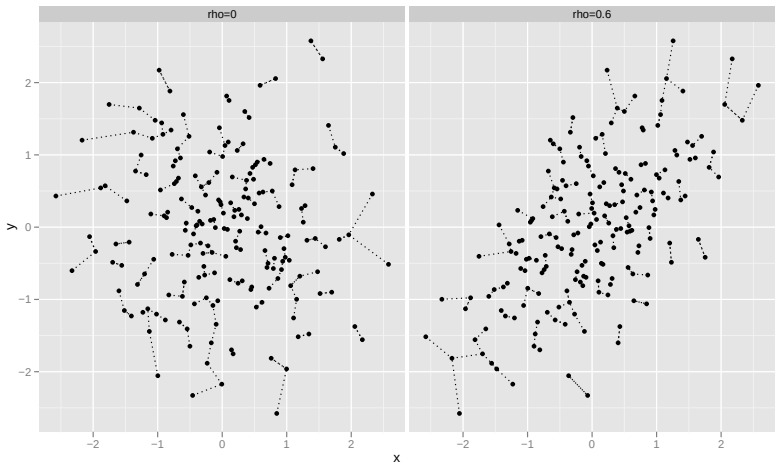

Figure 2.1: Scatter plot of 400 random samples drawn from (X, Y) with $\rho = 0$ (left, Mira score $S = 42.57$) and $\rho = 0.7$ (right, Mira score $S = 37.83$). Dotted line connects two observation points if one is the nearest neighbor of another.

distribution of mean 0 and variance 1, with correlation coefficient ρ. Figure 2.1 is the scatter plot for 400 random samples drawn from (X, Y) with $\rho = 0$ or $\rho = 0.7$. Samples for each variable were Gaussian standardized to have 0 mean and unit variance. Both cases have identical marginal distribution. The Mira score, calculated as the total edge length of the nearest neighbor connection, is observed to be smaller for the $\rho = 0.7$ case than its independent counterpart.

Quantities similar to Mira score were first utilized to assess goodness of fit in literature dating back to 1980s (Bickel and Breiman [1983], Penrose and Yukich [2001]). Nearest neighbor based entropy estimate was also exploited the observational spatial property to describe multivariate dependencies (Mnatsakanov et al. [2008], Leonenko et al. [2008], Beirlant et al. [1997]). However, no efficient procedure has been proposed for the discovery of high-dimensional dependency based on these valuable findings upon our literature survey.

Mira score enjoys asymptotic normality following Bickel and Breiman [1983]. That is, regardless of the joint distribution of $(X_1, \ldots, Y_q)$ or the norm used to define the distance matrix D, if $(x_{i1}, \ldots, y_{iq})_{i=1}^{n}$ is a sample from the joint distribution whose

dependence is characterized by the Mira score S, then if we calculate the Mira score, we will have

$$\frac{1}{\sqrt{n}}(S - E(S)) \to N(0, \sigma^2) \quad \text{as } n \to \infty \tag{2.2}$$

where $E(S)$ is the expectation of Mira score S, and σ is the asymptotic standard deviation. This property leads to the proposal of Gaussian plug-in permutation test in the next section, and dramatically reduces the computational burden of simulating the Mira score distribution under the null hypothesis of no dependence between $(X_1, \ldots, X_p)$ and $(Y_1, \ldots, Y_q)$.

2.2.1 Permutation test of association

We propose a Gaussian plug-in permutation test of probabilistic dependence between $(X_1, \ldots, X_p)$ and $(Y_1, \ldots, Y_q)$ give n pairs of independent samples $\{(x_{i1}, \ldots, y_{iq})\}_{i=1}^n$. Test statistic is generated as follows:

1. Calculate sample Mira score S_0. Set $r = 1$.
2. Permute the observation indeces for observations from Y, and generate new permuted sample $\{(x_{i1}, \ldots, x_{ip}, y_{(i)1}, \ldots, y_{(i)q})\}_{i=1}^n$. Calculate new Mira score S_r based on permuted sample.
3. Set $r = r + 1$
4. Repeat the above two steps R times.
5. Calculate $\hat{\mu} = \sum S_r/R$, $\hat{\sigma}^2 = \sum(S_r - \bar{S})^2/(R-1)$.
6. Test p-value equals to $\Phi_{\hat{\mu},\hat{\sigma}}(S_0)$, where $\Phi_{\hat{\mu},\hat{\sigma}}(\cdot)$ is the cumulative density function for normal distribution with mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$.

By default the number of permutations R is set to 100. The advantageous asymptotic normality of Mira score guarantees the performance of the test even with a small number of permutations.

2.3 Numerical Study

In this section we conduct numerical study to investigate the power of the Gaussian score permutation test. Three major aspect of comparison are considered. First, we investigated the power trade-off in bivariate cases compared with linear regression test of significance. Then we conducted simulation study in high-dimensional scenario and power comparison was make with competing methods. Finally, we studied the specificity of Mira score permutation test by evaluating the false positive rate under numerous null hypothesis settings.

2.3.1 Comparison with linear regression test of significance

In order to identify probabilistic dependence of arbitrary type, Mira score has made no distribution and functional assumptions. Thus the Mira score is expected to have weaker power in the perfectly linear case compared with linear regression test of significance. The study described is aimed at elucidating the trade-off between power and universal identifiability for the Mira score permutation test.

For simulation we considered the following hybrid bi-variate normal setting. Consider (X, Y) following bivariate normal distribution with standard normal marginal distribution and correlation coefficient ρ.

With probability θ, we assign $Z = -Y$, and let $Z = Y$ otherwise. We observe n samples from (X, Z) and set $n = 100$ in our simulation study. We conducted linear

regression significance test and Mira test for each combination of (θ, ρ) respectively with 100 replications. Here θ ranged from 0 to 0.5 with steps of 0.05. ρ ranged from 0 to 1 with steps of 0.05. Finally, the power are compared using the median p-value at each combination of (θ, ρ).

This simulation scenario is designed to compare the power of Mira score permutation test against linear regression test of significance on bivariate association conditions shifting with θ from perfectly linear correlation to mixed nonlinear association. X and Z are perfectly linearly correlated hen $\theta = 0$. In contrast, X and Z form a cross-shaped association with theoretical correlation coefficient 0 on the R^2 plane when $\theta = 0.5$. In addition, ranging ρ given fixed θ will compare the power between Mira permutation test and linear regression test of significance under different noise levels.

Simulated power for each combination of (θ, ρ) in the simulation with significance level 0.05 were plotted in Figure 2.2. In our study, Mira permutation test retained the power of identify probabilistic dependence in scenarios with high levels of nonlinearity (θ ranging from 0.4 to 0.5) and strong signals, while linear association based significance test fails to produce satisfactory results. However, Mira score permutation test do sacrifice the identification power in perfectly linear conditions (θ smaller than 0.2, ρ ranging from 0.2 to 0.5).

2.3.2 High-dimensional Power comparison

A simulation study was conducted to evaluate the performance of the Mira score permutation test. Power comparison was made with Brownian covariate (dCov) by (Szkely and Rizzo [2009]). Two major scenarios were considered:

Log dependent: X is p-variate standard normal. $Y_i = \log(X_i^2)$ for $i = 1, \ldots, p$. And n independent samples from (X, Y) are observed. Mira permutation test and

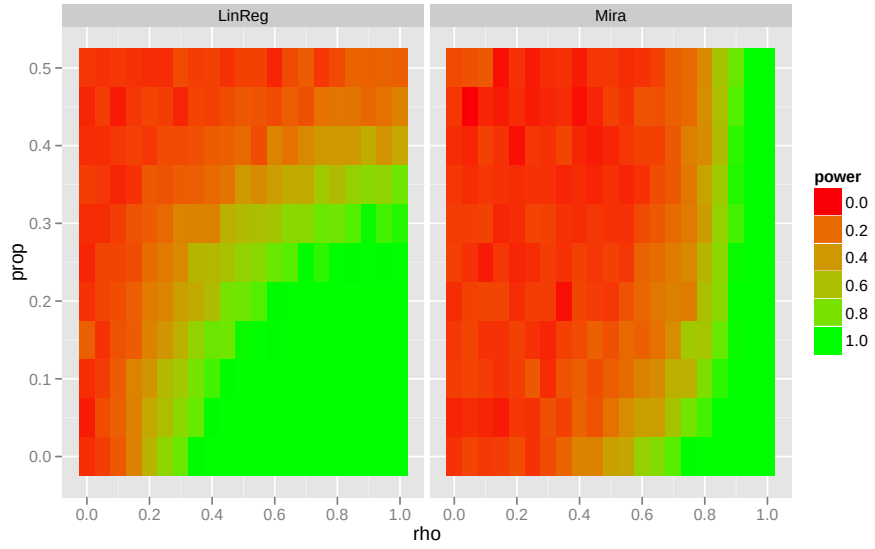

Figure 2.2: Power comparison for each combination of θ (prop) and ρ (rho) using Mira permutation test and linear regression test of significance.

Brownian covariate test were conducted with $p = 5$ on sample size $n = 20$ to 150 with step of 5. For each n, 10000 replications were made.

Normal Product: X and V are both p-variate standard normal random variables. $Y_i = X_i V_i$ for $i = 1, \ldots, p$. And n independent samples from (X, Y) are observed. Mira score permutation test and Brownian covariate bootstrap test was conducted with $p = 5$ on sample size $n = 20$ to 150 with step of 5. 10000 replications were made for each n.

Power comparison for the simulation study was made between Mira score and Brownian covariate (Figure 2.3). For both scenarios, the Mira score permutation test enjoyed higher power compared with Brownian covariate. Besides, Mira permutation test enjoys power higher than 0.90 even with small sample size ($n = 60$).

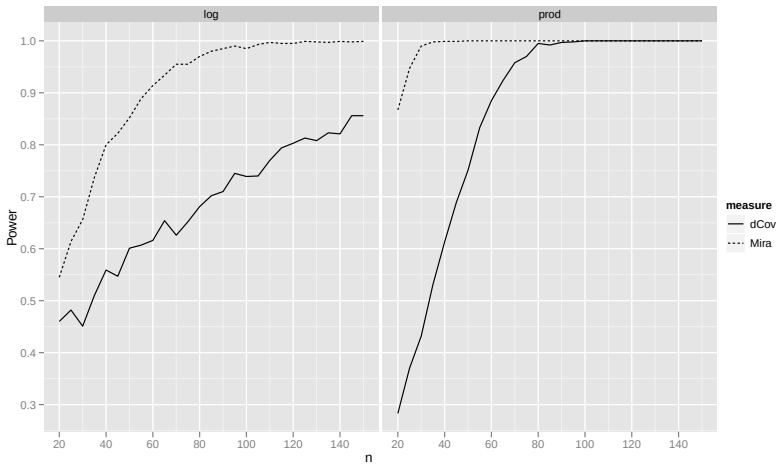

Figure 2.3: Power comparison of Mira permutation test and Brownian covariate test in log-dependent (left) and normal product (right) scenarios.

2.3.3 Specificity study under null hypothesis

A study of false positive rate using the Mira score permutation test has been highly recommended throughout the study. In an ideal statistical test, the false positive rate is expected to be close or equal to the confidence level α. The investigation may reveal the goodness of normal approximation used in generating null distribution for the test statistic.

The simulation was constructed based on testing the association between independent p-vector $(X_i, \ldots, X_p)$ and $(Y_i, \ldots, Y_p)$, both of which follow multivariate normal distribution with 0 mean vector and standard normal marginal distribution. Two covariate settings were considered. In the first setting, $\{X_i\}$ and $\{Y_i\}$ are mutually independent. In the second setting, both random vectors feature exchangeable covariate matrix with 0.2 correlation coefficient. In the simulation, p ranged for each setting from 5 to 20 with steps of 5. The sample size n ranged from 100 to 500 with steps of 100. Each combination of covariate matrix, n and p has been repeated 1000 times and the false positive rate at 0.05 significance level was evaluated.

Figure 2.4 shows the simulated false positive rate under each simulation setting at

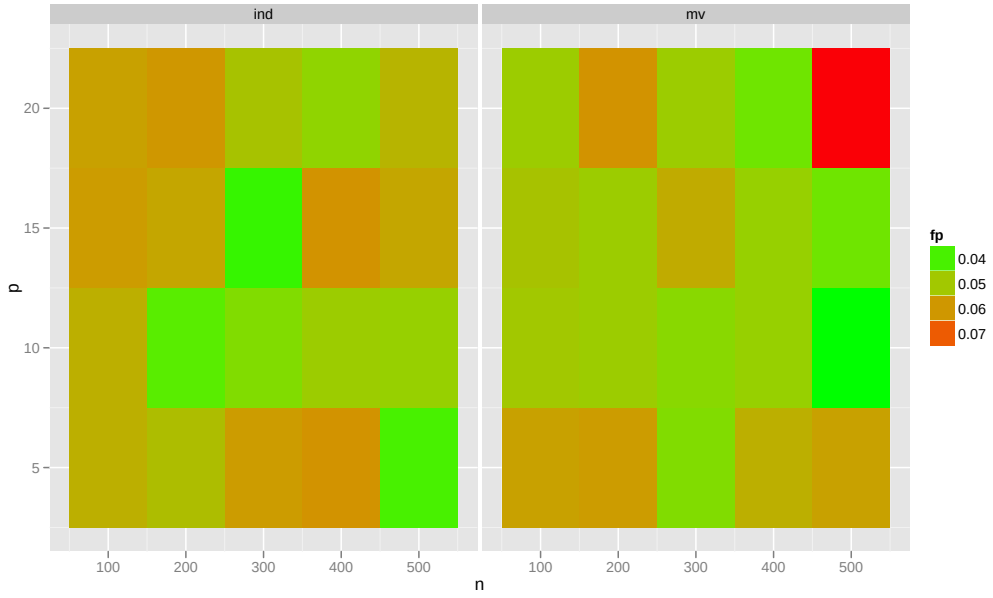

Figure 2.4: False positive rate in each scenario with identity covariate (left) and exchangeable covariate (right) matrix.

0.05 significance level. The false positive rates ranged between 0.04 and 0.07. And no trend with respect to sample size n or variable size p has been observed. Moreover, even with the presence of probablistic dependence within each random vectors, the false positive rates in the exchangeable covariate case are still around 0.05.

2.4 Differential pathway interaction network discovery

The change of pathway interactions under different cell status is of crucial interest in biomedical study. For example, certain interactions between pathways may be amplified or suppressed in disease state compared with healthy states. These changes in interaction may facilitate the discovery of cell regulatory mechanism.

We applied a Mira-score-based network reverse engineering procedure for pathway interaction to celiac disease data (NCBI data set GDS3646) and lung cancer data

(NCBI data set GDS2771), in which we are specifically interested in identifying pathway interactions that are amplified or suppressed in the disease state.

The celiac disease data consists of gene expression levels of untouched primary leukocytes from 132 unrelated celiac disease individuals and 22185 probes(Heap et al. [2009]). Of the 132 individuals, 110 have sustained celiac disease, and 22 are healthy control individuals. Illumina HumanWGv2 annotation data (Dunning et al.) was used to group probe reads into 214 KEGG pathway groups, covering 5201 genes (23.4%) of the data set. The lung cancer data consists of gene expression levels of large airway epithelial cells from cigarette smokers without cancer, with cancer, and with suspect lung cancer (Spira et al. [2007]). The probe reads were grouped into 214 KEGG pathways using Affymetrix Human Genome U133A database (Carlson et al.).

The amplified/suppressed pathway interactions were identified using the following procedure:

1. For each pair of non-overlapping pathways i and j, a Mira permutation test of association was applied for the disease group and control group, respectively. Denote the test p-value for disease group as p_{ij}^D, and p_{ij}^C for control group.

2. Rank $\{p_{ij}^D\}$ and $\{p_{ij}^C\}$ in ascending and obtain $\{r_{ij}^D\}$ and $\{r_{ij}^C\}$, respectively.

3. Calculate between state rank differences $d_{ij} = r_{ij}^D - r_{ij}^C$.

4. Pathway pairs with rank change d_{ij} smaller than the 1% quantile in $\{d_{ij}\}$ are identified as amplified in association at the disease state. Pairs with d_{ij} greater than 99% quantile are identified as suppressed in association in disease state.

The identified pathways interactions are then checked for their biological meanings and discussed in the result.

2.4.1 Celiac Disease Pathway Interaction

The newly developed method was used to analyze the Gene Expression Omnibus (GEO) data set GDS 3646. GDS3646 record is an expression analysis of untouched primary leukocytes from unrelated celiac disease individuals (Heap et al. [2009]). In the study, the gene expression in untouched primary leucocytes from individuals with celiac disease were compared with an EBV-transformed HapMap B cell line data. Celiac disease , a multifactorial disorder with complex genetics, is an enteropathy caused by autoimmune response against wheat gluten, the protein component of the cereals wheat, rye and barley in genetically susceptible individuals (Alaedini and Green [2005]). Patients with celiac disease have a wide spectrum of gastrointestinal and extraintestinal manifestations, characterized by intestinal malabsorption and atrophy of intestinal villi (Malandrino et al. [2008], Rubio-Tapia and Murray [2010]). Celiac patients experience altered carbohydrate, lipid, pipetide/protein, metabolism levels (Townley [1973], Pumarino et al. [1985], Vuoristo et al. [1993]). Untreated celiac patients oxidize more carbohydrates as energy substrate compared to treated subject (Malandrino et al. [2008]).

The identified pathways are dominantly related to nutrition absorption and metabolism (Figure 2.5). Other pathways potentially linked with celiac disease were also identified. For example, the 04062 chemokine signaling pathway appears 6 times on the list. Chemokines are small peptides that provide directional cues for the cell trafficking and thus are vital for protective inflammatory immune response that responses to requires the recruitment of leukocytes to the site of inflammation upon foreign insult. Celiac disease is known to be an inflammation disease caused by dietary gluten. In genetically predisposed people, gliadin peptides (derivatives of gluten) provokes immune response, which leads to the production of pro-inflammatory cytokines and subsequent damage to, and increased permeability of the intestinal epithelium (De

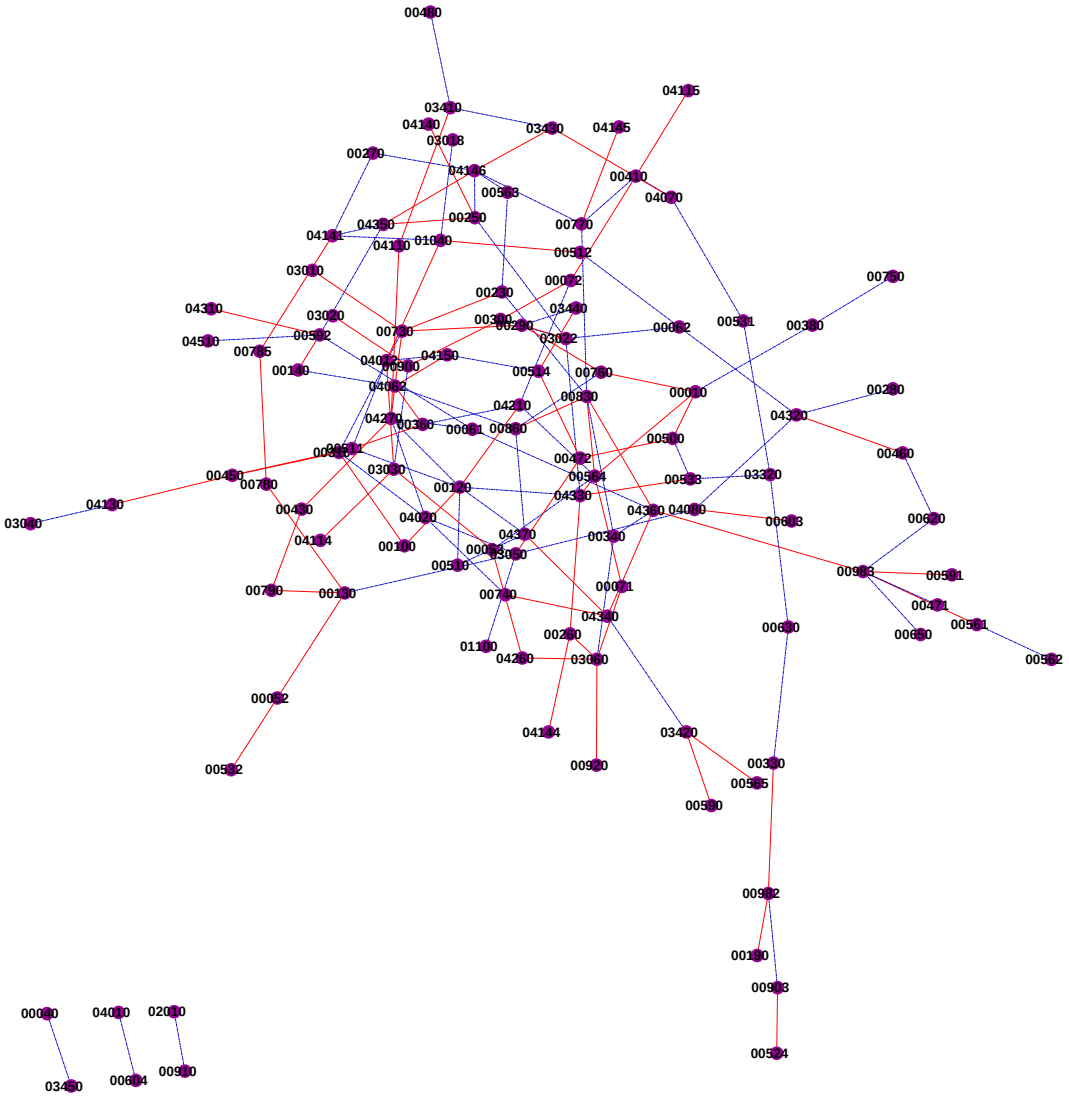

Figure 2.5: Network interaction for celiac disease pathways. Red edge indicates that the interaction between connected pathways are amplified in disease individuals. Grey edge indicates the interaction suppressed in disease individuals.

Carolis et al. [2004], Meresse et al. [2009], Bianchi [2010]).

Another example is the 02010 ABC transporters pathway. The results show that the interaction between ABC transporter pathway and nitrogen metabolism pathway is suppressed in the celiac disease patients. The ATP-binding cassette (ABC) transporters are protein families that couple ATP hydrolysis to activate transport of a wide variety of substrates such as ions, sugars, lipids, sterols, peptides, proteins, and drugs (Teodori et al. [2006], Linton [2007]). ABC transporters have been confirmed to be related to celiac disease. It has been reported that a close association exists between polymorphism of TAP1 and TAP2 (ABC transporter gene) and disease susceptibility among southern european populations (Tighe et al. [1994]). The products of TAP1 and TAP2 are ABC transporters, which are believed to transport antigenic peptides from the cytoplasm into the endoplasmic reticulum. It was reported that nitrogen balance was modulated in celiac patients (Caughey et al. [1955]). In addition, both nitrate/nitrite are transported by ATP-binding cassette (ABC) transporters (Maeda and Omata [2009]).

In addition, the correlation between 04370 VEGF signaling pathway and several pathways is found to be modulated in celiac patients, including 04340 hedgehog signaling, 510 N-Glycan biosynthesis, 00860 porphyrin and chlorophyll metabolism, 00120 primary bile acid biosynthesis. Vascular endothelial growth factor (VEGF) family and its receptor systems have been demonstrated to be the fundamental regulator in the cell signaling of angiogenesis. Angiogenesis is an essential biological process involved in the progression of a variety of major diseases such as cancer, diabetes and inflammation (Shibuya [2001]). It was reported that small-bowel mucosal vascular network was altered in untreated coeliac disease. The study found that on a gluten-containing diet the mucosal vasculature in the small intestine of untreated coeliac disease patients was altered in overall organization as well as in the number and maturity of the vessels when compared to healthy subjects. In patients on a gluten-free diet, the

vasculature normalized parallel to mucosal recovery (Myrsky et al. [2009]). Angiogenesis is reported to be related to hedgehog signaling (Koyama et al. [2007], Chen et al. [2011]), bile acid (Soma et al. [2006]), glycan biosynthesis (Banerjee [2007]), porphyrin (Aviezer et al. [2000], Lee et al. [2011]). 04210 apoptosis pathway, the programmed cell death, also frequently appears on the list. Much evidence showed increased small intestinal apoptosis in celiac disease (Moss et al. [1996]). Some other study demonstrated that enterocyte apoptosis induced by activated intraepithelial lymphocytes is increased in celiac disease (Giovannini et al. [2000]).

2.4.2 Lung Cancer Pathway Interaction

The newly developed method is also tested on GDS2771 data set, which is the microarray data of large airway epithelial cells from cigarette smokers without cancer, with cancer, and with suspect lung cancer. Many studies demonstrated the correlation of altered metabolism with lung cancer, including basal metabolism (Tokovoi and Matytsin [1967], Kurgan [1970, 1969]), carbohydrate metabolism (Heber et al. [1982], Giovacchini et al. [2009]), amino acid metabolism (Heber et al. [1982], Gabazza et al. [1995], Koukourakis et al. [2007]), lipid metabolism (Dessi et al. [1992]), and xenobiotic metabolism (Kiyohara et al. [2002]).

The result show that many correlations between metabolism related pathways are regulated (Figure 2.6). Take the TCA pathway as an example, the citrate cycle (TCA cycle, Krebs cycle) is an important aerobic pathway for the final steps of the oxidation of carbohydrates and fatty acids. Modulation of TCA cycle enzymes have been demonstrated in lung cancer. Decreased activities of TCA cycle key enzymes were observed in lung cancer bearing animals (Senthilnathan et al. [2006]). Some specific pathways that have been demonstrated to relate to lung cancer are also caught on the list: It is showed that the correlation between pathway 00072, the synthesis

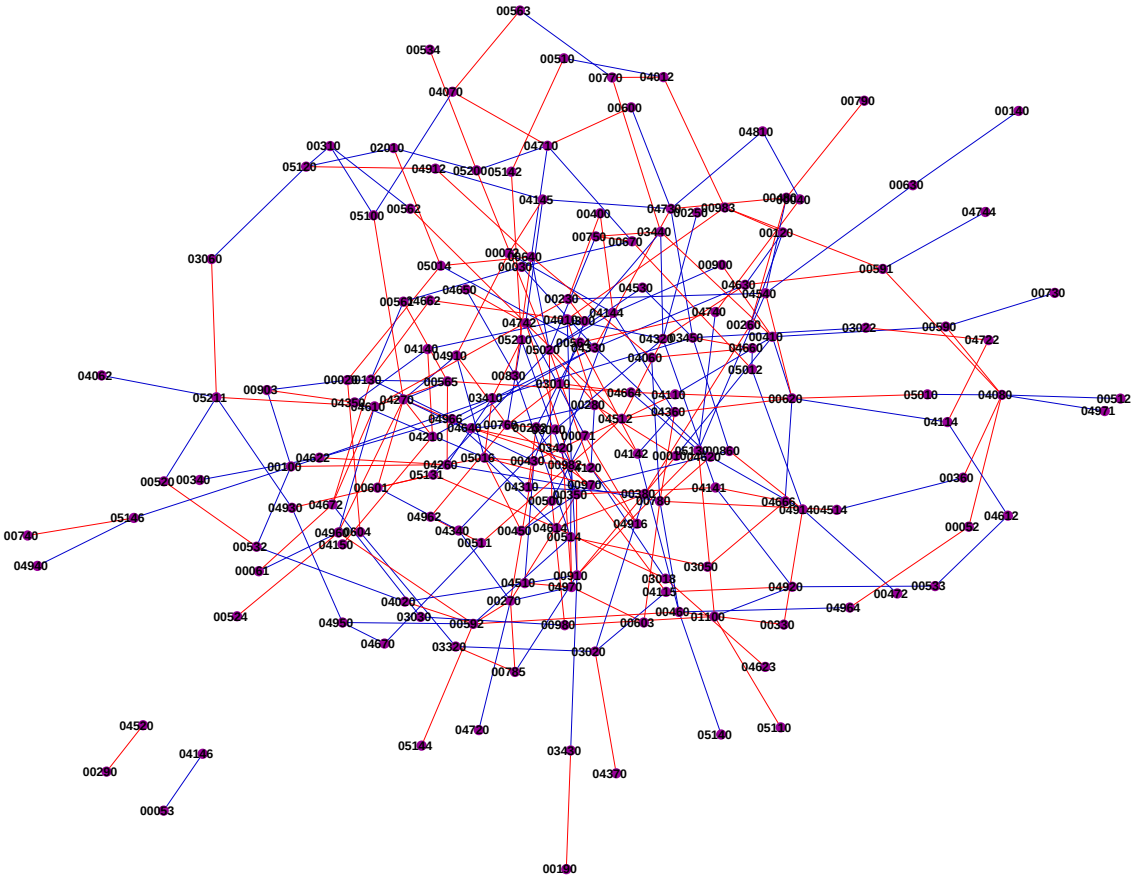

Figure 2.6: Network interaction for lung cancer pathways. Red edge indicates that the interaction between connected pathways are amplified in disease individuals. Grey edge indicates the interaction suppressed in disease individuals.

and degradation of ketone bodies pathway and 04145, the phagosome pathway is amplified in lung cancer patients. Phagocytosis is the cellular process of engulfing solid particles by the cell membrane to form an internal phagosome, which is a central mechanism in both immune and apotosis responses. There is a broad accepted view that bronchial neoplasms or its products suppress phagocytic functions of alveolar macrophages (Sulowicz [1983]).

The alveolar macrophage is believed to be of central importance in the immune response against infection and tumour. It has been reported that there are type-specific alterations in phagocytosis ability of alveolar macrophage in lung cancer patients, which may result in an inability to stimulate anti-tumour immunity and subsequently cause observed differences between lung cancer subgroups. Altered blood monocyte (BM) phagocytosis ability was also observed in patients with lung cancer (Hosker and Corris [1991], Pouniotis et al. [2006]). More importantly, some studies also proved that ketone bodies affect the phagocytic activity of macrophages and leukocytes (Klucinski et al. [1988]).

Another interesting example is that the correlation between 00232, the pathway of caffeine metabolism, and 00760, the nicotinate and nicotinamide metabolism, amplified in smokers with lung cancer. First of all, both caffeine and nicotine metabolism are generally believed to related to the risk of lung cancer. Cigarette smoking is a clear risk factor for lung cancer. Even though nicotine, one of the major ingredients and the causative agent for addiction of cigarette smoking, is generally believed not a carcinogen by itself. However, several studies have shown that nicotine can induce cell proliferation and angiogenesis (Puliyappadamba et al. [2010]). Nicotine metabolism by cytochrome P450 2A6 (CYP2A6) varies across ethnicity and race, which is indicated to be related to smoking behavior and lung cancer risk (Derby et al. [2008], Murray et al. [2009]). The same as smoking, the consumption of coffee is a very old and popular habit. Coffee contains catechins and flavonoids, which exhibit anticar-

cinogenic properties. Conversely, caffeine may elevate cancer risk through a variety of mechanisms (Martinet and Debry [1992], Baker et al. [2005]). Caffeine, an environmentally prominent phosphodiesterase, has been proved to selectively stimulate the growth of pulmonary adenocarcinoma and small airway epithelial cells (Al-Wadei et al. [2006]). Not only are both nicotine and caffeine related to lung cancer, but also many evidences suggested that the metabolism of caffeine and nicotine are closely correlated. Caffeine is mainly metabolized by cytochrome P450 1A2 (CYP1A2). Actually caffeine metabolism has been used as an in vivo marker of CYP1A2 activity, which has been clearly demonstrated to be induced by cigarette smoking(Zevin and Benowitz [1999]). The difference of caffeine intake and plasma concentrations among smokers and nonsmokers was reported (de Leon et al. [2003]). The results from 69 US samples showed that smokers had significantly higher caffeine intake than nonsmokers and the ratio of concentration/dose of caffeine was approximately four-fold higher in nonsmokers than in smokers (de Leon et al. [2003]). In animal studies, nicotine have been proved to induce the activity of several enzymes, including CYP1A2 (Carrillo and Benitez [1996]). It explains why nonsmokers have high plasma caffeinine concentration after intake of the same dose of caffeine compared to smokers. Some other research articles reported that the combined NAT2/CYP1A2 status was related to lung adenocarcinoma (Murray et al. [2009]).

2.5 Discussion

In this chapter we have proposed the Mira score, a novel association statistic for universal probabilistic dependencies. The Mira framework is very versatile and open to future possibilities of development. The definition of the Mira score is a special

case of functions defined on observation graphs

$$S = \sum_{i,j} d_{ij} w_{ij}$$

where $w_{ij} = \mathbf{1}(d_{ij} \leq d_{ik}, \forall k \neq j, k \neq i)$. Future improvement is possible through the development of different forms of w_{ij}'s as functions of distance matrix D. For example, we may define $W = (w_{ij})_{n \times n}$ as k-nearest neighbor connectivity matrix to incorporate more information around each observation point. Asymptotic normality is a straight forward result from existing literature (Bickel and Breiman [1983]). And gaussian plug-in permutation test for universal dependence can still be applied.

Bayesian inference can further expand the application of Mira score into the incorporation of information from multiple sources. For example, biological researchers may boost the power of association test for new experiments using the Mira scores of pre-existing experiments as prior. The prior distribution can be approximated using normal distribution with mean $\bar{S}$ and standard deviation $\hat{\sigma}$ both calculated from historical Mira scores.

The most exciting extension of the Mira score is the efficient variable selection procedure for supervised machine learning scenario proposed in the next chapter. Consider the situation where we have response variable $Y = (Y_1, \ldots, Y_q)$ and predictors $X = (X_1, \ldots, X_p)$. We are interested selecting a subset of variables from X that are most highly associated with Y based on n independent samples from (Y, X). In our working paper, we have proposed a very efficient procedure for universal dependency variable selection in high-dimensional data analysis scenario. Simulation study has shown very satisfactory statistical property. And the procedure has generated findings that are consistent with biological results in pathway analysis study.

Chapter 3

High-dimensional Universal Dependence Variable Selection

3.1 Introduction

Given random vectors $Y = (Y_1, \ldots, Y_q)$ and $X = (X_1, \ldots, X_p)$, we are interested in identifying a subset of $\{X_i\}$ that are probabilistically associated with $(Y_1, \ldots, Y_q)$. In this chapter we introduce SeMira, a variable selection procedure for high-dimensional universal association that is capable of accounting for probabilistic dependence of *arbitrary dimension* and *arbitrary forms of association*.

The proposal of a high-dimensional universal dependence variable selection procedure is motivated by the emergence of high-throughput data in biology and computer science. For example, in genetic pathway analysis, biologists are interested in discovering a small set of genes associated with the expression level of a pathway, which usually involves the interrelated expression of multiple genes. In functional magnetic resonance imaging (fMRI) studies, physicians are curious about the brain regions that are associated with the signal of a specific brain region, which may consists

of multivariate magnetic resonance signals from each brain regional. These practical challenges involve abundant nonlinear associations and high dimensional interactions, which has been the holy grail of a plethora of statistical methods in the recent decade (see Hastie et al. [2009] for a comprehensive review).

On the theory side, multiple nonlinear association measures have been developed. Dimension reduction extracts summary statistics through a linear combination of multiple variables (Zou et al. [2006], Tamayo et al. [2007], Mashal et al. [2005]), but may discard relevant information by ignoring secondary components. Mutual information by Margolin et al. [2006] has been widely applied in the discovery of nonlinear dependencies, but suffer from curse of dimensionality when the nonlinear dependency between more than two variables are considered. Liquid association is an innovative method developed to study the interaction involving three and more gene expressions (Li et al. [2004], Li [2002]) of specific interaction types. The Brownian covariate method is able to account for universal types of dependencies (Szkely and Rizzo [2009]), but lacks the expandability into efficient variable selection procedure.

Great advancement in variable selection strategy has been made in the recent decade. The Lasso family and related penalized regression methods (Tibshirani [1996], Zou [2006], Fan and Lv [2008], Fan and Li [1999], Hastie and Efron [2007]) are capable of efficiently selecting variables of linear association in high-dimensional scenario. Slice inverse regression is capable of conducting variable selection that accounts for functional nonlinear association between response variable and linear combination of predictors (Li [1991], Ferre [1998]). Numerous pairwise mutual-information-based heuristic approach have been proposed to account for nonlinear association (Peng et al. [2005], Durand et al. [2007], May et al. [2008]). However, no method upon our literature survey has the capability of conducting variable selection for universal dependence in high-dimensional data.

In our original proposal in the last chapter, we developed the Mira score, a universal dependence statistic that is capable to discover probabilistic association of any type involving arbitrary number of variables. The Mira score is defined as

$$S = \sum_{i=1}^{n} d_{(i)} \tag{3.1}$$

where $d_{(i)} = \min_{j \neq i} d_{ij}$ is the nearest neighbor edge length for the i-th observation. The Mira score has desirable statistical properties. Its extension to network reverse-engineering has been capable of finding meaningful biological patterns for subsequent research.

The application of the Mira score alone as a universal dependence statistic is limited in the high-throughput data analysis scenario given the huge number of parameters and the exponentially growing combination of predictors. Based on the Mira score, we propose SeMira procedure for high-dimensional nonlinear dependency variable selection.

The content of this chapter is structured as follow: in Section 3.2, we will introduce the SeMira procedure and discuss its mathematical relevance with the Mira score. In Section 3.3, we will compare the SeMira procedure with existing variable selection methods and evaluate their performance in numerous scenarios. The SeMira procedure is applied to genetic pathway interaction discovery in clinical gene expression data set in Section 3.4. And we point out directions for future research in Section 3.5.

3.2 SeMira procedure for variable selection

In this section we develop the SeMira procedure, an efficient variable selection procedure based on Mira score for universal types of probabilistic dependence.

Denote observation as $\{y_{i1}, \ldots, y_{iq}, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$, where $(y_{i1}, \ldots, y_{iq})$ are the multi/univariate response variable, and $(x_{i1}, \ldots, x_{ip})$ are the predictors. We define p-vector $v = (v_1, \ldots, v_p)$, with $v_i \in [0, 1]$ as the presence indicator for the variable X_i. Here $v_i = 1$ indicates the inclusion of the X_i in the model and 0 otherwise. We redefine the inter-sample distance as $d_{ij} = \sum_{l=1}^{p} |x_{il} - x_{jl}| v_l + \sum_{k=1}^{q} |y_{ik} - y_{jk}|$. Then we select a subset of the variables $\{X_i\}$ as probabilistically relevant to response variable $(Y_1, \ldots, Y_q)$ with

$$\begin{aligned} \hat{v} &= \arg\min_{v} \sum_{i,j} d_{ij} w_{ij} \\ &\quad \text{s.t. } |v|_1 \geq s \end{aligned} \tag{3.2}$$

where $s > 0$ is a tuning parameter and $W = (w_{ij})_{n \times n}$ is the 1-nearest neighbor connectivity matrix for the distance matrix $D = (d'_{ij})_{n \times n}$ defined using

$$d'_{ij} = \frac{s}{p} \sum_{l=1}^{p} |x_{il} - x_{jl}| + \sum_{k=1}^{q} |y_{ik} - y_{jk}| \tag{3.3}$$

Given tuning parameter s, the computation of solution to SeMira procedure 3.2 is straight forward:

1. Calculate distance matrix D and corresponding 1-nearest neighbor connectivity matrix W.
2. Calculate $u_l = \sum_{ij} |x_{il} - x_{jl}| w_{ij}$ for $l = 1, \ldots, p$.
3. Set $v_l = 1$ if u_l has the ascending rank in $\{u_l\}$ no larger than s. $v_l = 0$ otherwise.

It should be noted that the procedure above has computationally complexity of $O(n^2 p)$, which is very suitable for large p, moderate n problems in high-dimensional biological study.

This algorithm offers unprecedented computational efficiency. The daunting task of variable selection problem for probabilistic dependence with universal type and arbitrary dimension is reduced to three simple steps of matrix manipulation, which can be further accelerated using GPU computing or Map Reduce distributed computing facilities.

3.2.1 Geometry of Minimum Mira score Estimate

The proposal of SeMira procedure is based on Minimum Mira score Estimate, a computationally challenging estimator that we will discuss in detail in this section. Minimum Mira score estimator is worth mentioning here for its connection with Mira score and potential expansion into other forms of universal variable selection procedures. Under regularity conditions (Appendix .1), the result of SeMira procedure is identical to Minimum Mira score estimator.

Continuing the notations in last section, we define the Minimum Mira score Estimator as

$$\begin{aligned} \hat{v} &= \arg\min_{v} \sum_{i,j} d_{ij} w_{ij} \\ \text{s.t.} \quad & |v|_1 \geq s \quad \text{and } v_i \in [0,1] \\ & w_{ij} = \mathbf{1}\{d_{ij} \leq d_{ik}, \forall k \neq i\} \\ & d_{ij} = \sum_{l=1}^{p} v_l |x_{il} - x_{jl}| + \sum_{k=1}^{q} |y_{ik} - y_{jk}| \end{aligned} \tag{3.4}$$

where $s > 0$ is a tuning parameter for the estimate. A short comparison with SeMira procedure defined in Equation 3.2 can show that in SeMira, 1-nearest neighbor connectivity matrix remained static throughout the calculation. However, in Minimum Mira score Estimate, the connectivity matrix is updated with any updated v.

The design of Minimum Mira-score Estimate was inspired by Lasso (Tibshirani [1996]),

and the intuition is straight forward: for any given s, variables with higher dependence with response variable Y tend to generate a smaller sample Mira score. Thus we can select the relevant variables by minimizing the Mira score $\sum_i d_{ij} w_{ij}$ with respect to v.

Estimator 3.4 can be reduced to an optimization problem which requires gradient search in $\mathcal{R}^{p+n(n-1)}$ parameter space. High-throughput data problems with large n and p will pose tremendous computational challenge to the Minimum Mira score Estimator. However, we have proved that under regularity condition defined in Appendix .1, the solution to Minimum Mira score Estimator is identical to the SeMira procedure result defined in Equation 3.2. The intuition of this is straight forward: when the signal from the group of predictors are strong enough, the 1-nearest neighbor connectivity matrix we obtain using distance matrix defined by Equation 3.3 becomes identical to what we get with only the selected predictors defined in Equation 3.4.

3.2.2 Parameter tuning

Parameter tuning on s for SeMira procedure is challenging due to the fact that we do not have an explicit model for outcome prediction. Thus there is no traditional method to follow based on predictive error. Thus we resorted to tune s based on the penalty term $\sum d_{ij} w_{ij}$.

The tuning process is defined below:

1. Given s, and it's corresponding 1-nearest neighbor connectivity matrix W obtained using Equation 3.3, we calculate the contribution of the k-th parameter using $C_k = \sum_{ij} d_{ij}^k w_{ij}$, where d_{ij}^k is the distance between observation i and j based on variable X_k. That is, $d_{ij}^k = |x_{ik} - x_{jk}|$.

2. We ascend sort $\{C_k\}_{k=1}^p$ and calculate the increase of penalty contribution from

each predictor as $E_k = C_{k+1} - C_k$ for $k = 1 \dots k-1$.

3. Then we identify the largest increase in penalty term $\arg\max E_k$ and calculate its absolute difference with s, denoted as $A_s = |\arg\max_k E_k - s|$.

4. We iterate the above steps for all s and use the s corresponding to the smallest A_s as the optimum s.

The intuition for the procedure is based on the fact that if any given variable is *not* associated with the response variable, then its contribution to the penalty term should be larger compared with the variables that are associated with Y. Thus if we assume s as the optimum tuning parameter, then we may simply expect a jump of penalty contribution on the $s+1$-th smallest element in $\{C_k\}$.

3.3 Numerical study

In this section we conduct simulation studies to investigate the statistical property of SeMira procedure. The SeMira procedure does not make functional assumptions between response variable and predictor set. And the performance is evaluated using false discovery rate and false negative rate. As has been discussed in Section 3.2.2, the choice of tuning parameter s is heuristic at current stage. Thus two major scenarios are considered. First, we evaluated the performance of SeMira in Section 3.3.1 by using the tuning paramter selection procedure. Second, in Section 3.3.2 we evaluated the SeMira performance assuming that we have the right choice of s, and the heuristic tuning procedure was not used in the process.

3.3.1 SeMira procedure performance

Simulation study was conducted to investigate the power of variable selection of SeMira procedure in different scenarios, ranging from multivariate linear to more complicated nonlinear interactions with multivariate response variables. For each scenario, n independent random samples were generated from (Y, X) where X is a p-variate random vector following *i.i.d.* standard normal distribution, and Y is a q-vector response variable. In the simulation study, n ranged from 100 to 1000 with steps of 100. p ranged from 50 to 100 with steps of 5. q ranged from 3 to 6. Three major settings were considered:

Multivariate Additive Normal model is simulated to investigate the performance of SeMira in linear scenario. For the i-th sample, $Y_{ij} = X_{ij} + e_{ij}$ for $j = 1 \dots q$. Here $\{e_{ij}\}$ is *i.i.d.* standard normal independent of X.

Multivariate variance dependent model is simulated to investigate the performance of SeMira when the predictors only have functional effect on the variance of response variables. For the i-th sample, Y_{ij} follows normal distribution with mean 0 and variance $|X_{ij}|$ for $j = 1 \dots q$.

Multivariate triple interaction model is simulated to investigate the performance of SeMira in situations where variables are marginally independent but jointly dependent. For the i-th sample, $Y_{ij} = |e_{ij}|\text{sign}(X_{ij}X_{i(j+q)})$ for $i = 1 \dots n$, $j = 1 \dots q$. Here $\{e_{ij}\}$ are *i.i.d.* standard normal independent of $\{X_i\}$.

Simulation for each scenario and the combination of (n, p, q) is repeated 1000 times. Simulation results are evaluated using median false discovery rate and false negative rate for each setting with corresponding combination of (n, p, q).

False discovery rates presented in Figure 3.1 demonstrate the trade-off of specificity between different scenarios. The SeMira procedure has apparently good performance

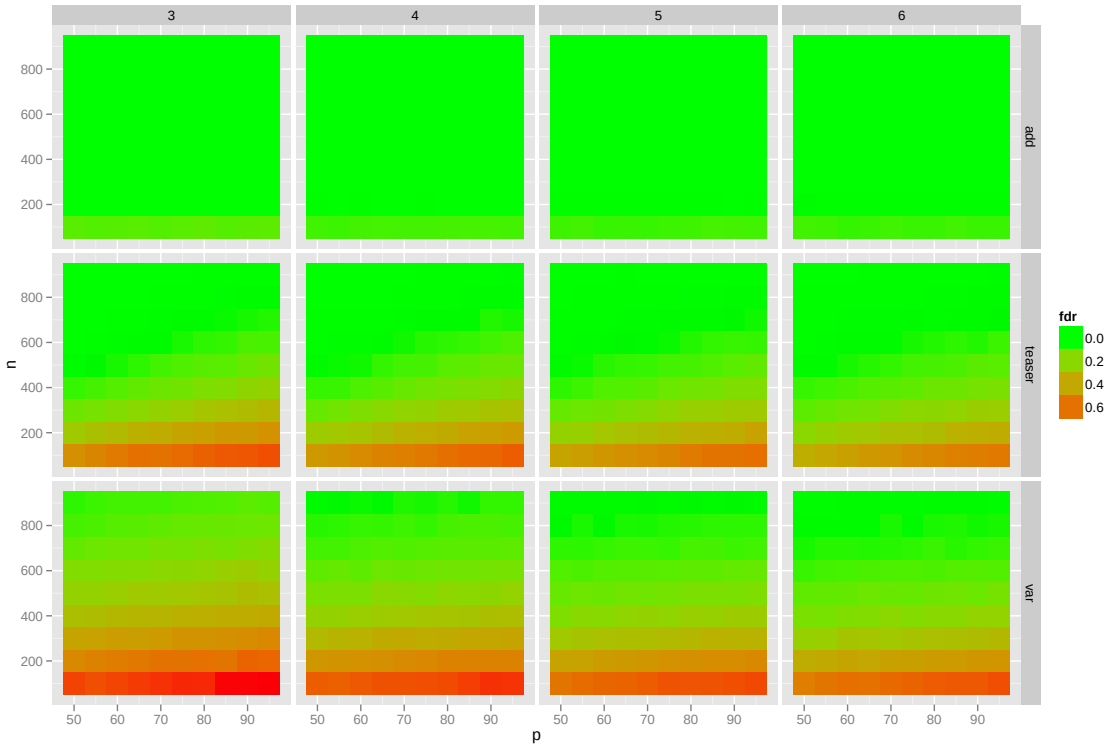

Figure 3.1: False discovery rate (FDR) for multivariate additive normal (add), variance dependent (var), and triple interaction (teaser) settings from the variable selection simulation study under different combinations of sample size (n), predictor size (p), and response variable size (q).

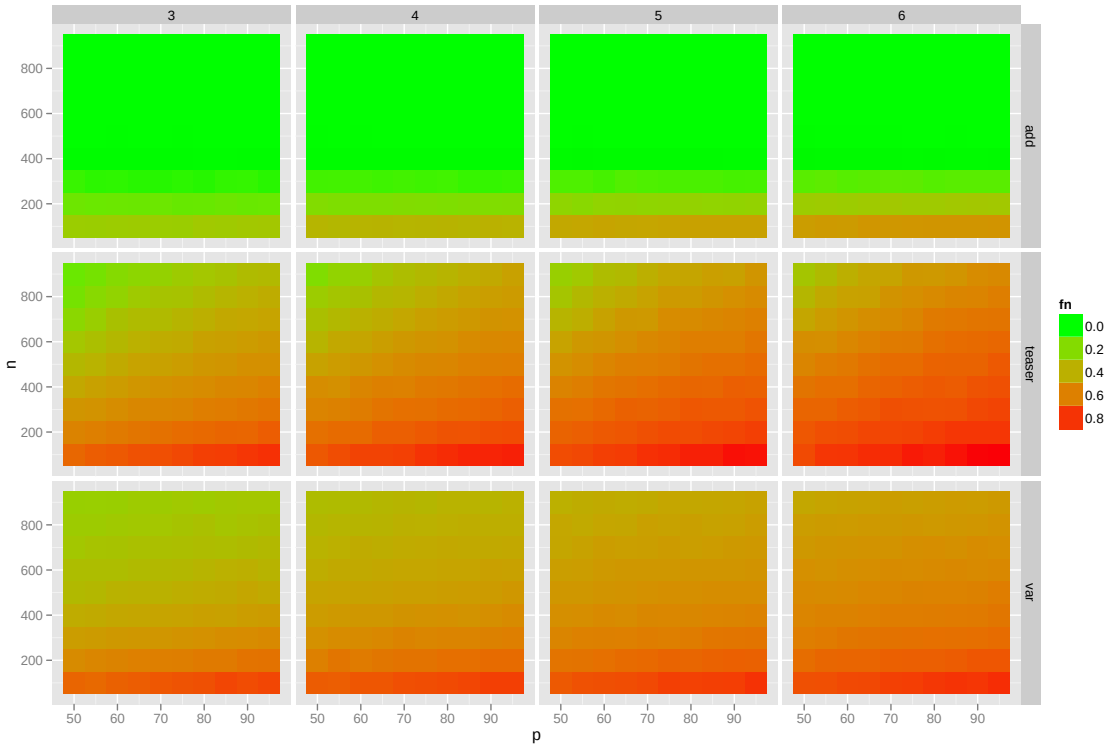

Figure 3.2: False negative rate (FN) for multivariate additive normal (add), variance dependent (var), and triple interaction (teaser) settings from the variable selection simulation study under different combinations of sample size (n), predictor size (p), and response variable size (q).

given large sample size ($n > 300$) across all three settings ($fdr < 0.1$). While it performs well with small sample size in linear additive setting, it suffers high false discovery rate with at variance dependent and multivariate triple dependent setting when sample size n is smaller than 200. In addition, we only observe a slight increase in FDR with growing number of unrelated predictors p.

False negative rate presented in Figure 3.2 demonstrate the trade-off of sensitivity between different scenarios. The SeMira procedure has small false negative rate ($fn < 0.1$) give relatively large number of samples ($n > 200$) in multivariate additive setting. However, the procedure suffered large false negative rate in variance dependent and triple interaction settings. We consider this a result of incomplete use of information from the observation graph that has originated from the definition of

Mira score. Recall that the Mira score is defined as the sum of 1-nearest neighbor edge length, which is only utilizing the information of each observation and its smallest neighborhood. Thus the contribution of variables related to Y might be buried under noise given small sample size and complicated dependence.

3.3.2 SeMira performance with known s

In this section we conducted numerical study to evaluate the statistical performance of SeMira procedure with known s, the number of variables involves with the interaction with response variables. The simulation settings were identical to Section 3.3.1, except that the tuning parameter s is set to equal to the true number of predictors associated with the response variable. On special consideration for situations with known s is that the inclusion of a variable unassociated with response in the underlying model will result in the exclusion of an associated variable. Thus the false discovery rate equals to false negative rate in this case. And the SeMira procedure is evaluated only using false discovery rate.

Simulation result is shown in Figure 3.3. Compared with simulation results using the heuristic tuning procedure in Figure 3.1, false discovery rate has increased in variance dependence, and triple interaction case. Meanwhile, the false negative rate has decreased with known s compared with Figure 3.2. Thus we can see that the heuristic parameter tuning procedure has been too conservative in selecting the associated variables.

Meanwhile, we can observe that even with known s, the number of variables associated with the response, we still need sufficiently large number of observations ($n > 800$) in order to achieve good selection result (FDR ¡ 0.10). This suggests that sample information may not have been fully utilized for the identification of relevant variables, as we will discuss in detail in Section 3.5.

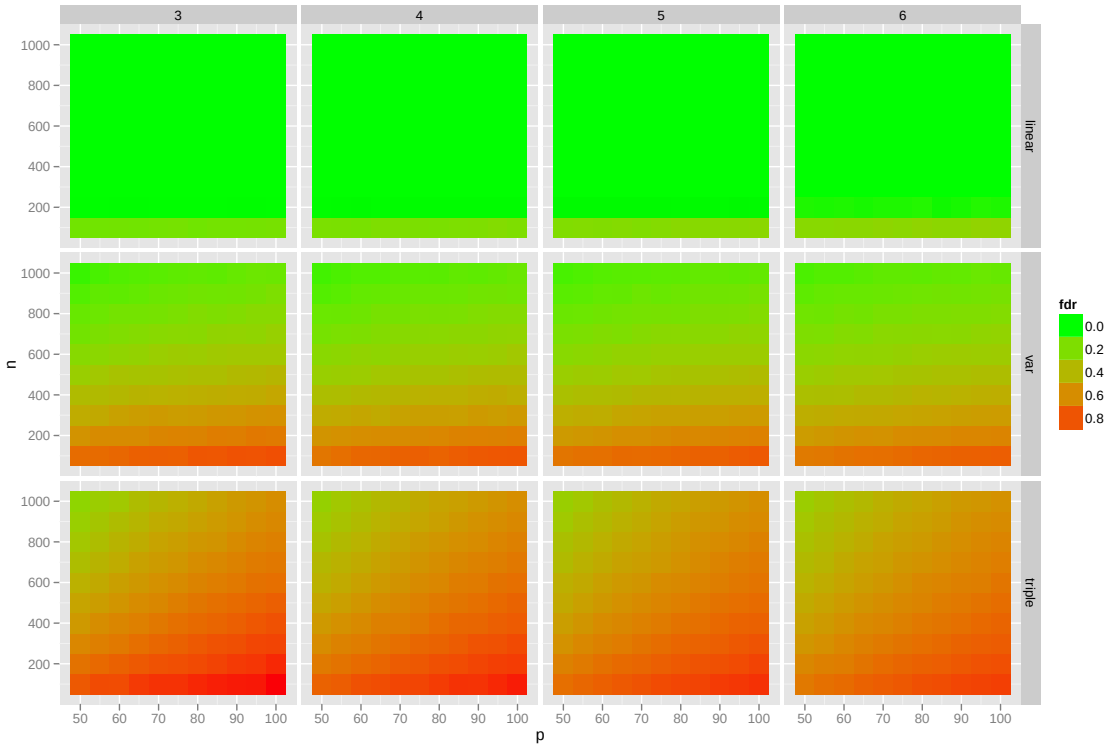

Figure 3.3: Mean false discovery rate (FDR) with different combinations of (n, p, q) and settings using SeMira procedure with known s.

3.4 Clinical outcome pathway interaction analysis

We applied the SeMira procedure to identify genetic pathways that are related to clinical response in the primary acute lymphoblastic leukemia (ALL) study with methotrexate (MTX) treatment using GEO data set GS10255 by Sorich et al. [2008]. The predictors in the data includes 12704 gene expression in primary acute lymphoblastic leukemia cells from 161 children diagnosed with childhood acute lymphoblastic leukemia. Log-transformed white blood cell counts at the beginning of MTX treatment and 3 days after the treatment are identified as clinical outcomes. Samples of each variable from the data was preprocessed with standard normal inverse-Gaussian transformation.

To elucidate the pathways that are associated with the white blood cell outcome, we selected top 1% genes that are associated with the outcome variables using SeMira procedure. Hypergeometric test for over-concentration using KEGG pathway terms with HGU133a annotation was used to identify the pathways enriched in the selected gene set. And 7 pathways were identified as significantly over-concentrated with p-value smaller than 0.01 (Table 3.4).

The set of enriched pathways are primarily focused on DNA transcription(Spliceosome pathway), protein degradation related to cell cycles. This has coincided with the findings that methotrexatec an inhibit the synthesis of DNA, RNA, thymidylates, and proteins.

3.5 Discussion

The SeMira procedure we proposed in this chapter is designed to give efficient solution to variable selection problems concerning probabilistic dependence of arbitrary

KEGGID	p-value	Term
03040	5.25×10^{-6}	Spliceosome
04120	7.27×10^{-5}	Ubiquitin mediated proteolysis
03050	1.55×10^{-3}	Proteasome
04141	1.80×10^{-3}	Protein processing in endoplasmic reticulum
05110	2.96×10^{-3}	Vibrio cholerae infection
00020	4.95×10^{-3}	Citrate cycle (TCA cycle)
00970	5.99×10^{-3}	Aminoacyl-tRNA biosynthesis

Table 3.1: Pathways identified as associated with white blood cell counts at the beginning of the MTX treatment, and 3 days after the treatment. Cutoff p-value 0.01.

dimension involving arbitrary types of associations. The procedure is defined based on Mira score, a very nascent universal dependence measure. We would like to point out that, as the definition of Mira score evolves, the corresponding SeMira procedure may also improve with its versatile forms. For example, we may define sum of K-nearest neighbor edge length in the observation graph as a new measure of probabilistic dependence. And the corresponding variable selection procedure is expected to be:

$$\begin{aligned} \hat{v} &= \arg\min_{v} \sum_{i,j} d_{ij} w_{ij} \\ &\quad \text{s.t. } |v|_1 \geq s \end{aligned} \tag{3.5}$$

where $w_{ij} = 1$ only when j-th observation is one of the k-nearest neighbors of i-th observation, and $w_{ij} = 0$ otherwise. Though the above procedure has not been formally studied, it is expected to have even better statistical property thanks to its ability to utilize more local information around each observation.

The computational complexity for the procedure is $O(n^2 p)$. We would like to point out that, as though the method is capable of handling large p, moderate n problems, the method can be further modified to adapt to even greater number of observations and get reduced to $O(np)$. The intuition is comes from the fact that a random m

subset of n observations from interrelated variables tend to be *closer* to each other than subset of samples from independent variables. Then instead of calculating the complete distance matrix with computational burden of $O(n^2)$, we can simply fix m, the subset size, and calculate 1-nearest neighbor distance for each variable in its specific random subset. Thus the computational complexity of this problem can be reduced to $O(np)$, which is very suitable for problems with large n and p.

Chapter 4

Quantification and Deconvolution Of Asymmetric LC-MS Peaks Using The Bi-Gaussian Mixture Model And Statistical Model Selection

4.1 Introduction

Liquid chromatography-mass spectrometry (LC-MS) is one of the major techniques in metabolomics (Issaq et al. [2009], Dettmer et al. [2007], Dunn [2008], Griffin and Kauppinen [2007]), as well as a key component in MS-based proteomics (Chen and Pramanik [2009], Ahmed [2009]). The pre-processing of LC-MS data involves a complex workflow including noise reduction, peak identification and quantification, retention time correction, peak alignment and weak signal recovery (Katajamaa and Oresic

[2007], Smith et al. [2006]). We have previously reported the apLCMS package which carries out the entire workflow with new algorithms specifically designed for LC-MS data with high mass resolution (Yu et al. [2009]). High-resolution mass spectrometry, such as Fourier transform mass spectrometry (FT-MS), allows the separation of m/z values at or below 10 ppm level (Ahmed [2008]), resulting in good separation between metabolites. The high resolution facilitates the use of empirical peak shape models to accurately quantify peaks, which is critical in biomarker studies where the relative quantities of metabolites are compared across samples. Currently, LC-MS peaks are quantified either by summation of ion count, or using symmetric peak shape models, such as the Gaussian function (Katajamaa and Oresic [2007], Smith et al. [2006], Yu et al. [2009]). Both methods have serious drawbacks. The method of ion count summation results in biased quantification when the ion trace has missing intensities, which often occurs in high-resolution LC-FTMS data. The Gaussian peak model can result in bias in peak location estimation and peak quantification when the peaks are asymmetric. Hence asymmetric peak models are necessary for the accurate quantification and identification of metabolites. In addition, some metabolites may share m/z and partially overlap in retention time, which necessitates the development of deconvolution procedures. A large number of empirical peak shape models have been developed for asymmetric peaks in chromatography, most of which were summarized in Di Marco and Bombi [2001]. For a few of the models, advanced deconvo- lution procedures are available (Felinger [1994], Johansson et al. [1993], Papai and Pap [2002], Youn et al. [1992], TorresLapasio et al. [1997b], Caballero et al. [2002]). Examples include the non-linear deconvolution based on Powells method (Powell [1965]) for the polynomial-modified Gaussian (PMG) model (TorresLapasio et al. [1997b,a]) regression-based methods for the parabolic-Lorentzian modified Gaussian (PLMG) model (Caballero et al. [2002]), and various deconvolution methods for the exponentially modified Gaussian (EMG) model (Felinger [1998], Johansson et al. [1993]).

The estimating procedures for asymmetric peak models in chromatographic data generally assume low noise level. In LC-MS data, the noise level is magnitudes higher, and the intensity observations are obtained at much fewer time points. Thus a simple, robust model that can be fitted using a limited number of intensity observations is necessary. The bi-Gaussian peak model (Figure 1a) has been described in the context of chromatography (Ahmed [2008], Felinger [1998]).

Empirical and theoretical results have shown that the bi-Gaussian model is well suited for asymmetric peaks (Buys and De Clerk [1972], Felinger [1998]). With four parameters and a simple functional form thats amenable to maximum likelihood estimation, the bi-Gaussian model is suitable for LC-MS data. A parameter estimation method for the bi-Gaussian model has been developed in the openMS environment (Sturm et al. [2008]). The method relies on the observed maximum intensity for the determination of the peak summit location, which could lead to inaccurate estimates when the signal-to-noise ratio is low. Currently no deconvolution method is available for the bi-Gaussian mixture model.

In this paper, we first develop a new algorithm to fit the bi-Gaussian function to noisy ion traces. Simulation study is then conducted to compare the performance of proposed procedure with competing methods. All the algorithms described here have been implemented to improve the apLCMS package for high-resolution LC-MS data analysis (Katajamaa and Oresic [2007]).

4.2 Methods

4.2.1 The bi-Gaussian peak model

The model involves four parameters – the location of the peak summit α, the standard deviation of the half Gaussian function to the left of the summit σ_1, the standard deviation of the half Gaussian function to the right of the summit σ_2, and the scaling factor δ. The intensity as a function of retention time is modeled by:

$$g(t) = \frac{\delta}{\sqrt{2\pi}} \exp\{-\frac{(t-\alpha)^2}{2}[\frac{1(t \leq \alpha)}{\sigma_1^2} + \frac{1(t > \alpha)}{\sigma_2^2}]\} \tag{4.1}$$

The areas of the two regions to the left/right of the peak summit are $\delta\sigma_1/2$ and $\delta\sigma_2/2$, respectively.

4.2.2 Likelihood-based estimation method

Assuming we have observation $\{(t_i, x_i)\}_{i=1}^n$, where t_i is the retention time for i-th observation, x_i is the corresponding intensity (weight). Continuing notations in Section 4.2.1, and scaling out δ with $\sigma_1 + \sigma_2$, the likelihood function is defined as

$$L(\sigma_1, \sigma_2, \alpha) \propto \frac{1}{(\sigma_1 + \sigma_2)^{\sum_{x_i}}} \exp\{-\sum x_i (t_i - \alpha)^2 [\frac{1(t_i \leq \alpha)}{\sigma_1^2} + \frac{1(t_i > \alpha)}{\sigma_2^2}]\}$$

Then the problem is identifying $(\sigma_1, \sigma_2, \alpha)$ is converted to the estimation based on the above likelihood function. The profile likelihood estimation procedure below is then used to generate estimation in an iterative manner:

1. Set $i = 0$. Generate initial estimation with $\hat{\alpha}^{(0)} = t_{(0)}$, where the corresponding $x_{(0)}$ has the greatest intensity among observations.

2. Given $\hat{\alpha}^{(i)}$, we define $u = \sum x_t(t_i - \hat{\alpha}^{(i)})^2 1(t_i \leq \alpha^{(0)})$, $v = \sum x_t(t_i - \hat{\alpha}^{(i)})^2 1(t_i > \alpha^{(i)})$, and $s = \sum x_i$. Then maximizing $L(\sigma_1, \sigma_2, \hat{\alpha}^{(i)}$, the estimation for bi-gaussian spans are calculated as:

$$\begin{aligned} \hat{\sigma}_1^{(i+1)} &= \sqrt{\frac{u}{s}(1 + \sqrt[3]{v/u})} \\ \hat{\sigma}_2^{(i+1)} &= \sqrt{\frac{v}{s}(1 + \sqrt[3]{u/v})} \end{aligned}$$

3. Given $(\hat{\alpha}^{(i)}, \hat{\sigma}_i^{(i+1)}, \hat{\sigma}_2^{(i+1)})$, we define $w = \frac{1}{\hat{\sigma}_1^{(i+1)}} \sum x_i 1(t_i \leq \hat{\alpha}^{(i)}) + \frac{1}{\hat{\sigma}_2^{(i+1)}} \sum x_i 1(t_i > \hat{\alpha}^{(i)})$, and $r = \frac{1}{\hat{\sigma}_1^{(i+1)}} \sum x_i t_i 1(t_i \leq \hat{\alpha}^{(i)}) + \frac{1}{\hat{\sigma}_2^{(i+1)}} \sum x_i t_i 1(t_i > \hat{\alpha}^{(i)})$. the estimation for α is updated as

$$\hat{\alpha}^{(i+1)} = r/w$$

4. Increase $i + 1$ and repeat the previous steps until the estimation converges.

Finally, Since log-likelihood function given α is concave, we and get consistent estimate for σ_1 and σ_2 if we have correct starting point of α.

4.2.3 Choosing the number of components of the mixture by statistical model selection

In the previous subsection, the kernel smoother is employed to obtain an initial estimate of the number of components and the parameters. When the data is noisy, changing the window size of the kernel smoother could result in different numbers of components of the mixture. To find the best model to explain the data, we utilize statistical model selection based on the Bayesian information criterion (BIC) (Schwarz [1978]). BIC is one of the most popular criteria for the selection among a set of parametric models with different number of parameters. It penalizes the number of free

parameters. The model with lower BIC value is preferred. First, a reasonable range of the window-size parameter is determined based on biological/chemical considerations about potential peak width. It can be quite lenient to cover a wide range of potential values. Several window size values spanning the range are selected. Starting from each of the window-size value, we compute the kernel smoother, and run the EM-like algorithm described in the previous sub-section. The corresponding BIC value is computed by:

$$N \log \left(\frac{1}{N} \sum_i (x_i - \sum_j \hat{z}_{ij})^2\right) + 4J \log N$$

where N is the total number of time points with observed intensities, and J is the number of bi-Gaussian components in the model. The model with the lowest BIC value is selected. In the setting of LC-MS data, this is a heuristic criterion, because the data we observe are not random samples, and the Gaussian error assumption of BIC may not be satisfied. We justify the usage of the criterion by extensive simulations.

4.3 Numerical Simulation

To assess the performance of the proposed method, extensive simulations were conducted. The bi-Gaussian mixture model with BIC model selection was compared with two other methods - the Gaussian mixture model (Yu et al. [2009]) with BIC model selection, and the peak quantification based on kernel smoother and signal summation. The data were generated from a 3-component bi- Gaussian mixture model, with different levels of peak asymmetry, noise and peak overlap. Given the parameters, the data from each component are generated from the bi-Gaussian functions:

$$g_j(t) = \frac{\delta_j}{\sqrt{2\pi}} \exp\{-\frac{(t-\alpha_j)^2}{2}[\frac{1(t \leq \alpha_j)}{\sigma_{j1}^2} + \frac{1(t > \alpha_j)}{\sigma_{j2}^2}]\}$$

After summing the intensities from the components, multiplicative noise was added to the data. In addition, a portion of the values were turned into zero to mimic the behavior of real high-resolution LC-MS data:

$$\begin{aligned} x_i &= \sum g_j(t_i) \exp(\epsilon_i)\mu_i \\ \epsilon_i &\sim N(0, \eta) \\ \mu_i &\sim \text{Binom}(\theta) \end{aligned}$$

The parameter η is the standard deviation of the noise added at the log-scale. Three levels of η were used in the simulations $(0.2, 0.4, 0.6)$. At the high noise level of $\eta = 0.6$, 50% of the intensity values were changed by 1.5 fold or more, and 25% were changed by two fold or more. The parameter θ controls the percentage of values turned into zero using random samples from the binomial distribution. Three levels of were used $(0, 0.25, 0.5)$. The value of θ directly corresponds to the proportion of intensities turned into zero. In addition, various levels of peak asymmetry and overlap were considered. In total 864 parameter combinations were tested. At each parameter setting, the simulation was performed 100 times.

4.4 Results

First, we compared the rate of successfully selecting the correct number of components between the bi-Gaussian mixture model and the Gaussian mixture model (Figure 4.1). The method of kernel smoother combined with signal summation was not compared because no BIC model selection could be performed using this method, which is a shortcoming in itself. In summarizing the results, the level of peak overlap is defined by the ratio r between the lowest point of the valley between two peaks and the lower of the peak summits, before noise is intro- duced. Because two valleys exist between

the three simu- lated peaks, the larger r value is taken for each simulation setting. For the purpose of plotting, we roughly divide the amount of overlap into four categories: little overlap ($r < 0.2$), moderate overlap ($0.2r < 0.5$), strong overlap ($0.5r < 0.75$), and severe overlap ($r0.75$). The level of overlapping is color-coded. The point size corresponds to the three levels of noise added to the data ($\eta = 0.2, 0.4, 0.6$). The fill of the point represents the proportion of missing values ($0\%, 25\%\text{and}50\%$). When the peaks were symmetric (Figure 4.1, upper-left panel), the Gaussian mixture model showed a slight advantage when the overlapping was strong (red and magenta points). When the peaks were asymmetric (Figure 4.1, upper-right and lower-left panels), the bi-Gaussian mixture model showed a clear advantage. When the peak overlapping was not strong (blue and green points), the success rate of the bi-Gaussian mixture model was mostly higher than 90%, even when thenoise level was high. When there was strong peak overlapping and the noise level was high (larger sized red and magenta points), the rate of successfully selecting the correct number of components was reduced for both the bi-Gaussian mixture model and the Gaussian mixture model.

Secondly, we compared the percentage error in peak area quantification between the three methods, when all three methods were able to identify the correct number of components (not necessarily the best BIC value). Compared to the Gaussian mixture model, the biGaussian mixture model yielded much smaller errors when the peaks were asymmetric (Figure 4.2, upper-right and lower-left panels). Compared to the method of kernel smoother combined with signal summation, the bi-Gaussian mixture model showed a clear advantage when some of the intensity values were missing (filled points) (Figure 4.3). When the peak overlapping was not strong (blue and green points), the error of the bi- Gaussian mixture model was mostly under 15%. The bi-Gaussian mixture model also clearly out-performed the other two methods in those aspects.

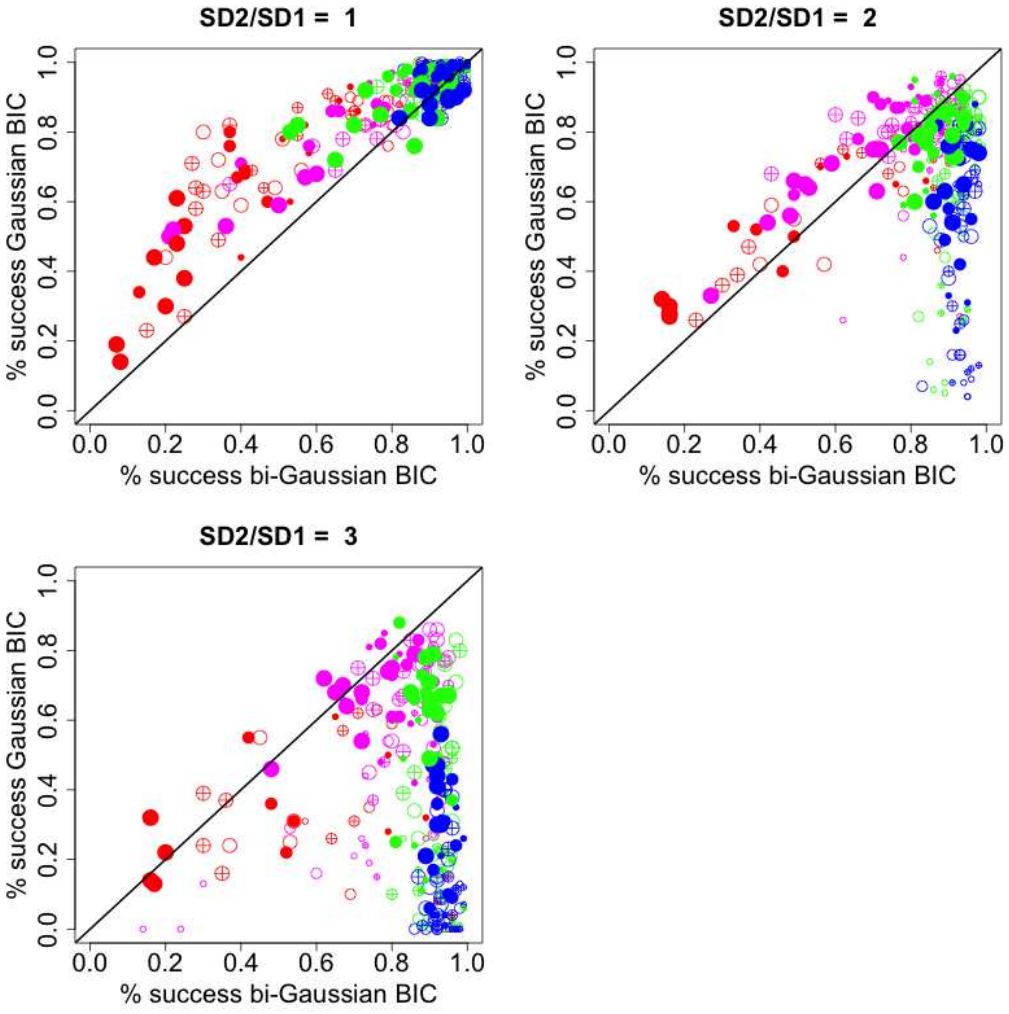

Figure 4.1: Comparison of the rate of successfully selecting the correct number of components between the bi-Gaussian mixture model and the Gaussian mixture model. Each sub-plot corresponds to a different degree of asymmetry, as shown in the titles of the sub-plots (ratios between the right- and left- standard deviations). Each dot represents a simulated situation. The values were obtained by averaging the results from 100 simulations. The color represents the level of overlaps between the simulated peaks. The size of the dot represents the amount of noise added to the data. The fill of the dot represents the percentage of values missing in the ion trace.

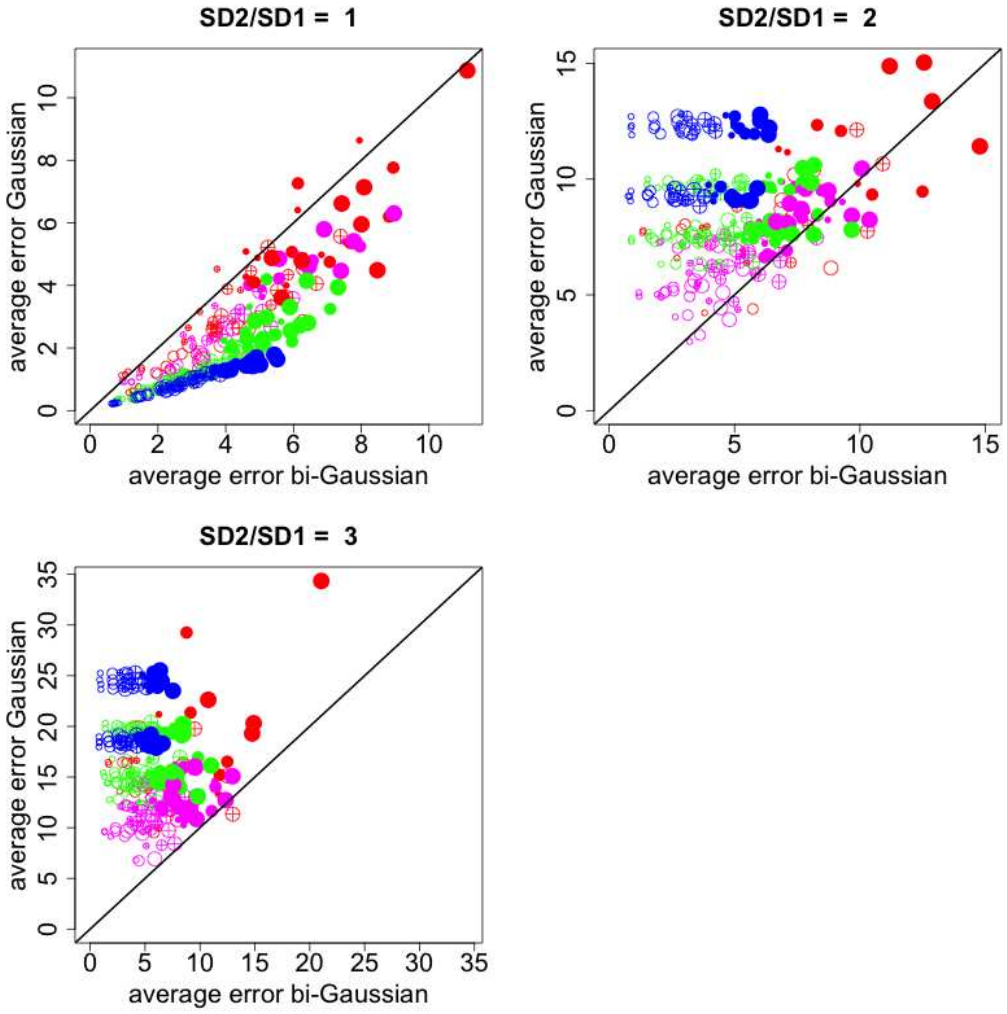

Figure 4.2: Comparison of the accuracy in peak size quantification between the bi-Gaussian mixture model and the Gaussian mixture model. Each sub-plot corresponds to a different degree of asymmetry, as shown in the titles of the sub-plots (ratios between the right- and left- standard deviations). Each dot represents a simulated situation. The values were obtained by averaging the results from 100 simulations. The color represents the level of overlaps between the simulated peaks. The size of the dot represents the amount of noise added to the data. The fill of the dot represents the percentage of values missing in the ion trace.

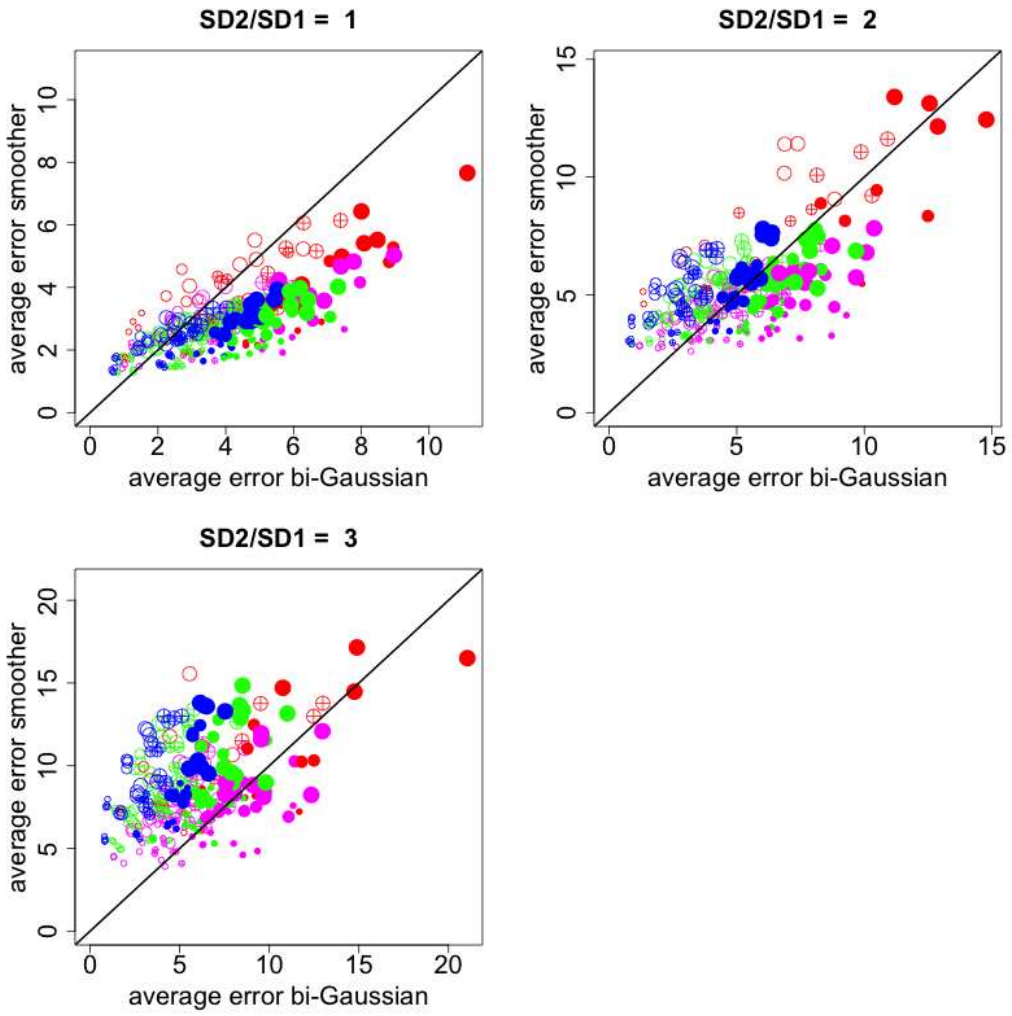

Figure 4.3: Comparison of the accuracy in peak size quantification between the bi-Gaussian mixture model and the method of kernel smoother combined with signal summation. Each sub-plot corresponds to a different degree of asymmetry, as shown in the titles of the sub-plots (ratios between the right- and left- standard deviations). Each dot represents a simulated situation. The values were obtained by averaging the results from 100 simulations. The color represents the level of overlaps between the simulated peaks. The size of the dot represents the amount of noise added to the data. The fill of the dot represents the percentage of values missing in the ion trace.

4.5 Discussion

In this manuscript, we presented a method to fit the bi-Gaussian curve to noisy LC-MS ion traces, as well as an EM-like algorithm paired with BIC model selection for the deconvolution of partially overlapping peaks. Currently, the methods were implemented in the apLCMS package for the pre-processing of high-resolution LC-MS data. The same modeling procedure can be adapted easily into other pipelines for the quantification of both metabolites and peptides.

Compared to the Gaussian peak shape model, which has been used in some model-based data processing pipelines (Smith et al. [2006], Yu et al. [2009]), the bi-Gaussian model provides extra flexibility to fit asymmetric peaks, while suffering little disadvantage when the true peak shape is symmetric. Compared to the method of kernel smoother combined with signal summation, fitting a bi-Gaussian mixture model disentangles partially overlapping peaks, and copes with the issue of missing intensities in high-resolution LC-FTMS data much better. The bi-Gaussian model is among many asymmetric peak models in chromatographic peak modeling. A large number of other models could potentially be used for the processing of LC-MS data (Di Marco and Bombi [2001]). Advanced deconvolution methods already exist for a few of the models (Felinger [1998], Johansson et al. [1993], Papai and Pap [2002], Youn et al. [1992], TorresLapasio et al. [1997b], Caballero et al. [2002], TorresLapasio et al. [1997a]). However, modifications to the existing estimation procedures may be necessary to suit the characteristics of LC-MS data, i.e. sparser data points and much higher noise.

In related study, the parameter estimation for a single peak is done by numerically solving an equation that involves the zero and second moments of the truncated distribution functions. This is an alternative route compared to the maximum likelihood method proposed in this chapter. We its performance with the moment-based

method in simulations. The likelihood-based algorithm was slower in computation due to its iterative nature, and it did not achieve better estimation accuracy over the moment-based method. And we would like to recommend using the moment-based method introduced by Yu and Peng [2010]. Under the settings of our simulations, five window size values were used for the initiation of the model selection process. With both methods programmed in R, using a single core of a 2.26 GHz Xeon CPU, the median CPU time for solving the three-component mixture was 0.15 second for the moment-based method, and 0.33 second for the likelihood-based method.

Chapter 5

Summary

In the dissertation we have pinpointed the necessity of universal dependence discovery amongst the emergency of high-throughput data, and developed novel statistical methods to utilize information that might include probabilistic association for continuous variables involving *arbitrary number* of variables and of *arbitrary types* of interaction.

Our work make two major contributions to the discovery of universal probabilistic dependence. First of all, we proposed Mira score, a universal association statistic that is capable of identifying probabilistic dependence of arbitrary type involving arbitrary number of variables. The Mira score permutation test enjoys superior power compared with existing method (Brownian co variate), and has been applied to the discovery of genetic network interaction in clinical pathways.

Second, we proposed SeMira procedure for variable selection based on probabilistic association of arbitrary dimension and of arbitrary type. The SeMira procedure is the *first* procedure capable of conducting the aforementioned task. Besides, the SeMira procedure allows variable selection for multivariate response variables without dimension reduction, a valuable feature to fully preserve sample information in

variable selection process.

The proposal of Mira score and SeMira procedure has opened up a brand new area of statistical research in universal dependence discovery and statistical inference. We would like to point out that, in addition to the prospective development mentioned in Section 2.5 and Section 3.5, endless opportunities for expansion awaits for the Mira-based methods.

One very desirable feature is the the capacity to conduct prediction of based on predictor $x_0 = (x_{01}, \ldots, x_{0p})$ with the constructed Mira based model. In our preliminary research, we proposed a predictor aimed at minimizing the prediction Mira score through

$$\hat{y}_0 = \arg\min_y \sum_{i=0}^{i} d_{(0)}$$

where $d_{(0)}$ is the distance of between predicted observation $(\hat{y}_0, x_0)$ and its nearest neighbor. This is identical as taking the 1-nearest neighbor estimate based on predictors. And we find, through numerical study, that the performance of this method is poor in high-dimensional scenario. The poor performance comes partly from the "hubness" of high-dimensional data. That is a few observation points have high-probability of becoming the k-nearest neighbor of other observations in high-dimensional space (Radovanović et al. [2010]). And our numerical study has found that in problems involving 1000 predictors and 100 samples, the majority of predictions of $\hat{y}_0$ are concentrated on the y_i of one single observation even if the predictor x_0 may scatter around the predictor space. We would like to admit that, although we boast that Mira score and SeMira procedure has circumvented the curse of dimensionality in previous chapter, the curse of dimensionality is actually taking effect on the Mira-based methods in prediction scenario.

On the other hand, endless possibilities exists for the development of even more powerful methods that better overcome the curse of dimensionality. We would like to

point out that the Mira-based method does not make any functional assumptions about random interactions. However, through our research, we have found that in very complex probabilistic interactions, local functional interaction may still be utilized in the solution of the problem.

Appendices

.1 Regularity condition for SeMira procedure

Continuing the notation in Section 3.2, we denote the 1-nearest neighbor connectivity matrix using distance matrix $D = (d_{ij})_{n\times n}$ with $d_{ij} = \frac{s}{p}\sum_{l=1}^{p}|x_{il} - x_{jl}| + |y_i - y_j|$ as $\hat{W} = (\hat{w}_{ij})_{n\times n}$. Denote the vectorized $n(n-1)$-vector for W as W^*, and matrix $E_{p\times n(n-1)}$ where each row of E is the vectorized distance matrix for the i-th variable.

In this section we prove that under regularity condition, the set of variable what we select using SeMira procedure is consistent with the set of variables we select using Minimum Mira score Estimator. This consistency can be intuitively explained by evaluating in the $\mathcal{R}^{p+n(n-1)}$ space the gradient of objective function $f(v, W) = \sum_{i,j} d_{ij}w_{ij}$ at O', the projection of origin in the hyperplane $C = \{(v, W) : |v|_1 = s, \sum_{i\neq j} w_{ij} = 1, \forall j\}$. The projection of $\frac{\partial f}{\partial(v,W)}|_{O'}$ on the hyperplane C is

$$\begin{aligned}\frac{\partial f}{\partial v}|_{O'} &= 0\\ \frac{\partial f}{\partial W}|_{O'} &= \frac{s}{p}\mathbf{1}_p E + F\end{aligned}$$

which suggests that for gradient search starting at O' on hyperplane C, one should first follow the direction of 1-nearest neighbor connectivity matrix defined by Equation 3.3, and then update on v. This direction of search is identical as the direction to solution defined as the Minimum Mira score estimate. Or more mathematically,

Definition 1 (Regularity condition)**.** *Given $p \geq s > 0$, regularity condition is satisfied when*

$$v^T E(W^* - W) \leq \frac{s}{p}\mathbf{1}_p E(W^* - W)$$

for vectorized 1-connectivity matrix W, and any $v = (v_1, \ldots, v_p)$ such that $|v|_1 = s$ and $v_i \in [0, 1]\forall i$. □

To prove the claim in Section 3.2.1 that the SeMira procedure result defined in Equation 3.2 is identical to Minimum Mira score Estimate defined in 3.4 under regularity condition, we use the prove by contrary procedure.

Assume that there exists a p-vector v and vectorized 1-connectivity matrix W such that the corresponding objective function value at (v, W) is smaller than that of MME (v^*, W^*), that is:

$$v^T EW + F^T W < v^{*T} EW^* + F^T W^*$$

where F is the vectorized distance matrix for response observations $\{y_i\}_{i=1}^n$. Making use of the fact that

$$\begin{aligned} \frac{s}{p}(\mathbf{1}_p^T E + F^T)W^* &\leq \frac{s}{p}(\mathbf{1}_p^T E + F^T)W \\ v^{*T} EW^* &\leq v^T EW \end{aligned}$$

we have

$$v^T E(W^* - W) > \frac{s}{p}\mathbf{1}_p E(W^* - W)$$

which contradicts with the aforementioned regularity condition.

Bibliography

FE Ahmed. Utility of mass spectrometry for proteome analysis: part i. conceptual and experimental approaches. *Expert Rev Proteomics*, 5(6):841–864, 2008. doi: 10.1586/14789450.5.6.841.

FE Ahmed. Utility of mass spectrometry for proteome analysis: part ii. ion-activation methods, statistics, bioinformatics and annotation. *Expert Rev Proteomics*, 6(2): 171–197, 2009. doi: 10.1586/epr.09.4.

H. A. Al-Wadei, T. Takahashi, and H. M. Schuller. Caffeine stimulates the proliferation of human lung adenocarcinoma cells and small airway epithelial cells via activation of pka, creb and erk1/2. *Oncol Rep*, 15(2):431–5, Feb 2006.

A. Alaedini and P. H. Green. Narrative review: celiac disease: understanding a complex autoimmune disorder. *Ann Intern Med*, 142(4):289–98, Feb 15 2005.

D. Aviezer, S. Cotton, M. David, A. Segev, N. Khaselev, N. Galili, Z. Gross, and A. Yayon. Porphyrin analogues as novel antagonists of fibroblast growth factor and vascular endothelial growth factor receptor binding that inhibit endothelial cell proliferation, tumor progression, and metastasis. *Cancer Res*, 60(11):2973–80, Jun 1 2000.

J. A. Baker, S. E. McCann, M. E. Reid, S. Nowell, G. P. Beehler, and K. B. Moysich.

Associations between black tea and coffee consumption and risk of lung cancer among current and former smokers. *Nutr Cancer*, 52(1):15–21, 2005.

D. K. Banerjee. Requirement of protein kinase type i for camp-mediated up-regulation of lipid-linked oligosaccharide for asparagine-linked protein glycosylation. *Cell Mol Biol (Noisy-le-grand)*, 53(3):55–63, 2007.

J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Meulen. Nonparametric entropy estimation: An overview. *International Journal of the Mathematical Statistics Sciences*, 6:17–39, 1997. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.87.5281`.

M. L. Bianchi. Inflammatory bowel diseases, celiac disease, and bone. *Arch Biochem Biophys*, 503(1):54–65, Nov 1 2010.

Peter J. Bickel and Leo Breiman. Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *The Annals of Probability*, 1983.

Danah M. Boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2008. ISSN 1083-6101. doi: 10.1111/j.1083-6101.2007.00393.x. URL `http://dx.doi.org/10.1111/j.1083-6101.2007.00393.x`.

TS Buys and K De Clerk. Bi-gaussian fitting of skewed peaks. *Analytical Chemistry*, 44(7):1273–1275, 1972. doi: 10.1021/ac60315a005.

RD Caballero, MC Garcia-Alvarez-Coque, and JJ Baeza-Baeza. Parabolic-lorentzian modified gaussian model for describing and deconvolving chromatographic peaks. *Journal of Chromatography A*, 954(1-2):59–76, 2002. doi: 10.1016/S0021-9673(02)00194-2.

Marc Carlson, Seth Falcon, Herve Pages, and Nianhua Li. *hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a)*. R package version 2.4.5.

J. A. Carrillo and J. Benitez. Cyp1a2 activity, gender and smoking, as variables influencing the toxicity of caffeine. *Br J Clin Pharmacol*, 41(6):605–8, Jun 1996.

R. H. Caughey, Crory Ww Mc, and R. Kaye. Nitrogen balances in a patient with fibrocystic disease of the pancreas and a patient with the celiac syndrome and the effect of aureomycin. *Pediatrics*, 16(2):174–83, Aug 1955.

G Chen and BN Pramanik. Application of lc/ms to proteomics studies: current status and future prospects. *Drug Discov Today*, 14(9-10):465–471, 2009. doi: 10.1016/j.drudis.2009.02.007.

W. Chen, T. Tang, J. Eastham-Anderson, D. Dunlap, B. Alicke, M. Nannini, S. Gould, R. Yauch, Z. Modrusan, K. J. Dupree, W. C. Darbonne, G. Plowman, F. J. de Sauvage, and C. A. Callahan. Canonical hedgehog signaling augments tumor angiogenesis by induction of vegf-a in stromal perivascular cells. *Proc Natl Acad Sci U S A*, 108(23):9589–94, Jun 7 2011.

S. De Carolis, A. Botta, G. Fatigante, S. Garofalo, S. Ferrazzani, A. Gasbarrini, and A. Caruso. Celiac disease and inflammatory bowel disease in pregnancy. *Lupus*, 13 (9):653–8, 2004.

J. de Leon, F. J. Diaz, T. Rogers, D. Browne, L. Dinsmore, O. H. Ghosheh, L. P. Dwoskin, and P. A. Crooks. A pilot study of plasma caffeine concentrations in a us sample of smoker and nonsmoker volunteers. *Prog Neuropsychopharmacol Biol Psychiatry*, 27(1):165–71, Feb 2003.

K. S. Derby, K. Cuthrell, C. Caberto, S. G. Carmella, A. A. Franke, S. S. Hecht, S. E. Murphy, and L. Le Marchand. Nicotine metabolism in three ethnic/racial

groups with different risks of lung cancer. *Cancer Epidemiol Biomarkers Prev*, 17 (12):3526–35, Dec 2008.

S. Dessi, B. Batetta, D. Pulisci, O. Spano, R. Cherchi, G. Lanfranco, L. Tessitore, P. Costelli, F. M. Baccino, C. Anchisi, and et al. Altered pattern of lipid metabolism in patients with lung cancer. *Oncology*, 49(6):436–41, 1992.

K Dettmer, PA Aronov, and BD Hammock. Mass spectrometry-based metabolomics. *Mass Spectrom Rev*, 26(1):51–78, 2007. doi: 10.1002/mas.20108.

VB Di Marco and GG Bombi. Mathematical functions for the representation of chromatographic peaks. *J Chromatogr A*, 931(1-2):1–30, 2001. doi: 10.1016/S0021-9673(01)01136-0.

WB Dunn. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Phys Biol*, 5(1):11001, 2008. doi: 10.1088/1478-3975/5/1/011001.

Mark Dunning, Matt Ritchie, and Nuno Barbosa-Morais. *illuminaHumanv2.db: Illumina HumanWGv2 annotation data (chip illuminaHumanv2)*. R package version 1.8.0.

A. Durand, O. Devos, C. Ruckebusch, and J.P. Huvenne. Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cottonviscose textiles. *Analytica Chimica Acta*, 595(12):72 – 79, 2007. ISSN 0003-2670. doi: 10.1016/j.aca.2007.03.024. URL `http://www.sciencedirect.com/science/article/pii/S0003267007005387`. ¡ce:title¿Papers presented at the 10th International Conference on Chemometrics in Analytical Chemistry¡/ce:title¿ ¡ce:subtitle¿CAC 2006¡/ce:subtitle¿.

Jianqing Fan and Runze Li. Variable Selection via Penalized Likelihood. *Journal of American Statistical Association*, 96:1348–1360, 1999. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.4256.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, November 2008. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00674.x. URL http://dx.doi.org/10.1111/j.1467-9868.2008.00674.x.

A Felinger. Deconvolution of overlapping skewed peaks. *Analytical Chemistry*, 66 (19):3066–3072, 1994. doi: 10.1021/ac00091a013.

A Felinger. Data analysis and signal processing in chromatography. 1998.

Louis Ferre. Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93(441):pp. 132–140, 1998. ISSN 01621459. URL http://www.jstor.org/stable/2669610.

E. C. Gabazza, O. Taguchi, M. Yoshida, T. Yamakami, H. Kobayashi, H. Ibata, and T. Shima. Neutrophil activation and collagen metabolism in lung cancer. *Clin Chim Acta*, 236(1):101–8, Apr 30 1995.

T. Gautama, D.P. Mandic, and M.M. Van Hulle. Signal nonlinearity in fmri: a comparison between bold and mion. *Medical Imaging, IEEE Transactions on*, 22 (5):636 –644, may 2003. ISSN 0278-0062. doi: 10.1109/TMI.2003.812248.

G. Giovacchini, M. Picchio, S. Schipani, C. Landoni, L. Gianolli, V. Bettinardi, N. Di Muzio, M. C. Gilardi, F. Fazio, and C. Messa. Changes in glucose metabolism during and after radiotherapy in non-small cell lung cancer. *Tumori*, 95(2):177–84, Mar-Apr 2009.

C. Giovannini, M. Sanchez, E. Straface, B. Scazzocchio, M. Silano, and M. De Vincenzi. Induction of apoptosis in caco-2 cells by wheat gliadin peptides. *Toxicology*, 145(1):63–71, Apr 7 2000.

JL Griffin and RA Kauppinen. A metabolomics perspective of human brain tumours. *Febs J*, 274(5):1132–1139, 2007. doi: 10.1111/j.1742-4658.2007.05676.x.

Trevor Hastie and Brad Efron. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2007. URL `http://www-stat.stanford.edu/~hastie/Papers/#LARS`. R package version 0.9-7.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, 2nd ed. 2009. corr. 3rd printing 5th printing. edition, September 2009. ISBN 0387848576. URL `http://www-stat.stanford.edu/\~{}tibs/ElemStatLearn/main.html`.

J. Hasty, D. Mcmillen, F. Isaacs, and J. J. Collins. Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genetics*, 2(4): 268–279, April 2001. ISSN 1471-0056. doi: http://dx.doi.org/10.1038/35066056. URL `http://dx.doi.org/10.1038/35066056`.

Graham A. Heap, Gosia Trynka, Ritsert C. Jansen, Marcel Bruinenberg, Morris A. Swertz, Lotte C. Dinesen, Karen A. Hunt, Cisca Wijmenga, David A. Vanheel, and Lude Franke. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC medical genomics*, 2, 2009. ISSN 1755-8794. doi: 10.1186/1755-8794-2-1. URL `http://dx.doi.org/10.1186/1755-8794-2-1`.

D. Heber, R. T. Chlebowski, D. E. Ishibashi, J. N. Herrold, and J. B. Block. Abnormalities in glucose and protein metabolism in noncachectic lung cancer patients. *Cancer Res*, 42(11):4815–9, Nov 1982.

H. S. Hosker and P. A. Corris. Alveolar macrophage and blood monocyte function in lung cancer. *Cancer Detect Prev*, 15(2):103–6, 1991.

HJ Issaq, QN Van, TJ Waybright, GM Muschik, and TD Veenstra. Analytical and statistical approaches to metabolomics research. *J Sep Sci*, 32(13):2183–2199, 2009. doi: 10.1002/jssc.200900152.

M Johansson, M Berglund, and DC Baxter. Improving accuracy in the quantitation of overlapping, asymmetric, chromatographic peaks by deconvolution - theory and application to coupled gas-chromatography atomic-absorption spectrometry. *Spectrochim Acta B*, 48(11):1393–1409, 1993. doi: 10.1016/0584-8547(93)80127-G.

Leigh A. Johnston, Eugene Duff, Iven Mareels, and Gary F. Egan. Nonlinear estimation of the BOLD signal. *NeuroImage*, 40(2):504–514, April 2008. doi: 10.1016/j.neuroimage.2007.11.024. URL `http://dx.doi.org/10.1016/j.neuroimage.2007.11.024`.

M Katajamaa and M Oresic. Data processing for mass spectrometry-based metabolomics. *J Chromatogr A*, 1158(1-2):318–328, 2007. doi: 10.1016/j.chroma.2007.04.021.

C. Kiyohara, T. Shirakawa, and J. M. Hopkin. Genetic polymorphism of enzymes involved in xenobiotic metabolism and the risk of lung cancer. *Environ Health Prev Med*, 7(2):47–59, may 2002.

W. Klucinski, A. Degorski, E. Miernik-Degorska, S. Targowski, and A. Winnicka. Effect of ketone bodies on the phagocytic activity of bovine milk macrophages and polymorphonuclear leukocytes. *Zentralbl Veterinarmed A*, 35(8):632–9, Sep 1988.

M. I. Koukourakis, A. Giatromanolaki, G. Bougioukas, and E. Sivridis. Lung cancer: a comparative study of metabolism related protein expression in cancer cells and tumor associated stroma. *Cancer Biol Ther*, 6(9):1476–9, Sep 2007.

E. Koyama, B. Young, M. Nagayama, Y. Shibukawa, M. Enomoto-Iwamoto, M. Iwamoto, Y. Maeda, B. Lanske, B. Song, R. Serra, and M. Pacifici. Conditional kif3a ablation causes abnormal hedgehog signaling topography, growth plate dysfunction, and excessive bone and cartilage formation during mouse skeletogenesis. *Development*, 134(11):2159–69, Jun 2007.

J. Kurgan. [basal metabolism in patients with lung cancer and the effect of corticosteroid therapy]. *Gruzlica*, 37(4):297–302, Apr 1969.

J. Kurgan. [basal metabolism in patients with lung cancer and its modification by corticotherapy]. *Z Erkr Atmungsorgane Folia Bronchol*, 132(2):181–6, 1970.

J. M. Lee, W. H. Lee, H. Y. Kay, E. S. Kim, A. Moon, and S. G. Kim. Hemin, an iron-binding porphyrin, inhibits hif-1alpha induction through its binding with heat shock protein 90. *Int J Cancer*, Mar 16 2011.

Nikolai Leonenko, Luc Pronzato, and Vippal Savani. Estimation of entropies and divergences via nearest neighbors. In *Tatra Mt. Math. Publ. ProbaStat 2006*, volume 39, pages 265–273, Smolenice Slovakia, 2008. URL `http://hal.archives-ouvertes.fr/hal-00322783/en/`. 94A15, 62G20.

K. C. Li. Genome-wide coexpression dynamics: theory and application. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):16875–16880, December 2002. ISSN 0027-8424. doi: 10.1073/pnas.252466999. URL `http://dx.doi.org/10.1073/pnas.252466999`.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):pp. 316–327, 1991. ISSN 01621459. URL `http://www.jstor.org/stable/2290563`.

Ker-Chau Li, Ching-Ti Liu, Wei Sun, Shinsheng Yuan, and Tianwei Yu. A system for enhancing genome-wide coexpression dynamics study. *Proceedings of the Na-*

tional Academy of Sciences of the United States of America, 101(44):15561–15566, 2004. doi: 10.1073/pnas.0402962101. URL http://www.pnas.org/content/101/44/15561.abstract.

K. J. Linton. Structure and function of abc transporters. *Physiology (Bethesda)*, 22: 122–30, Apr 2007.

S. Maeda and T. Omata. Nitrite transport activity of the abc-type cyanate transporter of the cyanobacterium synechococcus elongatus. *J Bacteriol*, 191(10):3265–72, may 2009.

N. Malandrino, E. Capristo, S. Farnetti, L. Leggio, L. Abenavoli, G. Addolorato, and G. Gasbarrini. Metabolic and nutritional features in adult celiac patients. *Dig Dis*, 26(2):128–33, 2008.

Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7 Suppl 1, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-S1-S7. URL http://dx.doi.org/10.1186/1471-2105-7-S1-S7.

Y. Martinet and G. Debry. [effects of coffee on the respiratory system]. *Rev Mal Respir*, 9(6):587–92, 1992.

N. Mashal, M. Faust, and T. Hendler. The role of the right hemisphere in processing nonsalient metaphorical meanings: Application of principal components analysis to fmri data. *Neuropsychologia*, 43(14):2084 – 2100, 2005. ISSN 0028-3932. doi: DOI: 10.1016/j.neuropsychologia.2005.03.019. URL http://www.sciencedirect.com/science/article/B6T0D-4G0YTS1-2/2/f9f75ab71e48750ade0243fbebb6e125.

Robert J. May, Holger R. Maier, Graeme C. Dandy, and T.M.K. Gayani Fernando. Non-linear variable selection for artificial neural networks using par-

tial mutual information. *Environmental Modelling and Software*, 23(1011):1312 – 1326, 2008. ISSN 1364-8152. doi: 10.1016/j.envsoft.2008.03.007. URL http://www.sciencedirect.com/science/article/pii/S1364815208000467.

B. Meresse, J. Ripoche, M. Heyman, and N. Cerf-Bensussan. Celiac disease: from oral tolerance to intestinal inflammation, autoimmunity and lymphomagenesis. *Mucosal Immunol*, 2(1):8–23, Jan 2009.

R. Mnatsakanov, N. Misra, Sh. Li, and E. Harner. k_n-nearest neighbor estimators of entropy. *Mathematical Methods of Statistics*, 17:261–277, 2008. ISSN 1066-5307. URL http://dx.doi.org/10.3103/S106653070803006X. 10.3103/S106653070803006X.

S. F. Moss, L. Attia, J. V. Scholes, J. R. Walters, and P. R. Holt. Increased small intestinal apoptosis in coeliac disease. *Gut*, 39(6):811–7, Dec 1996.

R. P. Murray, J. E. Connett, and L. M. Zapawa. Does nicotine replacement therapy cause cancer? evidence from the lung health study. *Nicotine Tob Res*, 11(9):1076–82, Sep 2009.

E. Myrsky, M. Syrjanen, I. R. Korponay-Szabo, M. Maki, K. Kaukinen, and K. Lindfors. Altered small-bowel mucosal vascular network in untreated coeliac disease. *Scand J Gastroenterol*, 44(2):162–7, 2009.

Z Papai and TL Pap. Determination of chromatographic peak parameters by nonlinear curve fitting using statistical moments. *Analyst*, 127(4):494–498, 2002. doi: 10.1039/b111304f.

Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226 –1238, aug. 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.159.

Mathew D. Penrose and J. E. Yukich. Central limit theorems for some graphs in computational geometry. *The Annals of Applied Probability*, 11(4):pp. 1005–1041, 2001. ISSN 10505164. URL http://www.jstor.org/stable/2699907.

D. S. Pouniotis, M. Plebanski, V. Apostolopoulos, and C. F. McDonald. Alveolar macrophage function is altered in patients with lung cancer. *Clin Exp Immunol*, 143(2):363–72, Feb 2006.

MJD Powell. A method for minimizing a sum of squares of non-linear functions without calculating derivatives. *Comput J*, 7(4):303–307, 1965.

Vineshkumar Puliyappadamba, Vino Cheriyan, Arun Kumar Thulasidasan, Smitha Bava, Balachandran Vinod, Priya Prabhu, Ranji Varghese, Arathy Bevin, Shalini Venugopal, and Ruby Anto. Nicotine-induced survival signaling in lung cancer cells is dependent on their p53 status while its down-regulation by curcumin is independent. *Molecular Cancer*, 9(1):220, 2010. ISSN 1476-4598. doi: 10.1186/1476-4598-9-220. URL http://www.molecular-cancer.com/content/9/1/220.

H. Pumarino, C. Campino, R. Palma, H. Michelsen, and G. Generini. [mineral metabolism and secondary hyperparathyroidism in celiac disease]. *Rev Med Chil*, 113(11):1065–71, Nov 1985.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, 11:2487–2531, December 2010. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1756006.1953015.

Marylyn Ritchie, Bill White, Joel Parker, Lance Hahn, and Jason Moore. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioin-*

formatics, 4(1):28, 2003. ISSN 1471-2105. doi: 10.1186/1471-2105-4-28. URL http://www.biomedcentral.com/1471-2105/4/28.

A. Rubio-Tapia and J. A. Murray. Celiac disease. *Curr Opin Gastroenterol*, 26(2): 116–22, Mar 2010.

G Schwarz. Estimating dimension of a model. *Ann Stat*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136.

P. Senthilnathan, R. Padmavathi, V. Magesh, and D. Sakthisekaran. Modulation of tca cycle enzymes and electron transport chain systems in experimental lung cancer. *Life Sci*, 78(9):1010–4, Jan 25 2006.

M. Shibuya. Structure and function of vegf/vegf-receptor system involved in angiogenesis. *Cell Struct Funct*, 26(1):25–35, Feb 2001.

CA Smith, EJ Want, G O'Maille, R Abagyan, and G Siuzdak. Xcms: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*, 78(3):779–787, 2006. doi: 10.1021/ac051437y.

T. Soma, J. Kaganoi, A. Kawabe, K. Kondo, S. Tsunoda, M. Imamura, and Y. Shimada. Chenodeoxycholic acid stimulates the progression of human esophageal cancer cells: A possible mechanism of angiogenesis in patients with esophageal cancer. *Int J Cancer*, 119(4):771–82, Aug 15 2006.

Michael J Sorich, Nicolas Pottier, Deqing Pei, Wenjian Yang, Leo Kager, Gabriele Stocco, Cheng Cheng, John C Panetta, Ching-Hon Pui, Mary V Relling, Meyling H Cheok, and William E Evans. In vivo response to methotrexate forecasts outcome of acute lymphoblastic leukemia and has a distinct gene expression profile. *PLoS Med*, 5(4):e83, 04 2008. doi: 10.1371/journal.pmed.0050083. URL http://dx.doi.org/10.1371\%2Fjournal.pmed.0050083.

Avrum Spira, Jennifer E. Beane, Vishal Shah, Katrina Steiling, Gang Liu, Frank Schembri, Sean Gilman, Yves-Martine Dumas, Paul Calner, Paola Sebastiani, Sriram Sridhar, John Beamis, Carla Lamb, Timothy Anderson, Norman Gerry, Joseph Keane, Marc E. Lenburg, and Jerome S. Brody. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, 13(3):361–366, March 2007. ISSN 1078-8956. doi: 10.1038/nm1556. URL `http://dx.doi.org/10.1038/nm1556`.

M Sturm, A Bertsch, C Gropl, A Hildebrandt, R Hussong, E Lange, N Pfeifer, O Schulz-Trieglaff, A Zerck, and K Reinert. Openms - an open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9:163, 2008. doi: 10.1186/1471-2105-9-163.

W. Sulowicz. Phagocytosis and peroxidase activity in neutrophils from peripheral blood of patients with malignant tumours of lung, stomach and large intestine. *Folia Haematol Int Mag Klin Morphol Blutforsch*, 110(1):48–54, 1983.

Gbor J. Szkely and Maria L. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 2009.

Pablo Tamayo, Daniel Scanfeld, Benjamin L. Ebert, Michael A. Gillette, Charles W. M. Roberts, and Jill P. Mesirov. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proceedings of the National Academy of Sciences*, 104(14):5959–5964, 2007. doi: 10.1073/pnas.0701068104. URL `http://www.pnas.org/content/104/14/5959.abstract`.

E. Teodori, S. Dei, C. Martelli, S. Scapecchi, and F. Gualtieri. The functions and structure of abc transporters: implications for the design of new inhibitors of pgp and mrp1 to control multidrug resistance (mdr). *Curr Drug Targets*, 7(7):893–909, Jul 2006.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. doi: 10.2307/2346178. URL http://dx.doi.org/10.2307/2346178. the very original paper of Lasso that started the mess.

M. R. Tighe, M. A. Hall, E. Cardi, A. Ashkenazi, J. S. Lanchbury, and P. J. Ciclitira. Associations between alleles of the major histocompatibility complex-encoded abc transporter gene tap2, hla class ii alleles, and celiac disease susceptibility. *Hum Immunol*, 39(1):9–16, Jan 1994.

V. A. Tokovoi and A. N. Matytsin. [basal metabolism in patients with lung cancer]. *Vopr Onkol*, 13(8):74–7, 1967.

JR TorresLapasio, JJ BaezaBaeza, and MC GarciaAlvarezCoque. A model for the description, simulation, and deconvolution of skewed chromatographic peaks. *Analytical Chemistry*, 69(18):3822–3831, 1997a. doi: 10.1021/ac970223g.

JR TorresLapasio, MC GarciaAlvarezCoque, and JJ BaezaBaeza. Global treatment of chromatographic data with michrom. *Anal Chim Acta*, 348(1-3):187–196, 1997b. doi: 10.1016/S0003-2670(97)00066-4.

R. R. Townley. Celiac disease–an inborn error of metabolism. *Am J Dig Dis*, 18(9): 797–800, Sep 1973.

Hiroshi Toyoda, Kenichi Kashikura, Tomohisa Okada, Satoru Nakashita, Manabu Honda, Yoshiharu Yonekura, Hideo Kawaguchi, Atsushi Maki, and Norihiro Sadato. Source of nonlinearity of the BOLD response revealed by simultaneous fMRI and NIRS. *NeuroImage*, 39(3):997–1013, February 2008. doi: 10.1016/j.neuroimage.2007.09.053. URL http://dx.doi.org/10.1016/j.neuroimage.2007.09.053.

M. Vuoristo, Y. A. Kesaniemi, H. Gylling, and T. A. Miettinen. Metabolism of cholesterol and apolipoprotein b in celiac disease. *Metabolism*, 42(11):1386–91, Nov 1993.

DY Youn, SJ Yun, and KH Jung. Improved algorithm for resolution of overlapped asymmetric chromatographic peaks. *J Chromatogr*, 591(1-2):19–29, 1992. doi: 10.1016/0021-9673(92)80219-K.

T Yu, Y Park, JM Johnson, and DP Jones. aplcms–adaptive processing of high-resolution lc/ms data. *Bioinformatics*, 25(15):1930–1936, 2009. doi: 10.1093/bioinformatics/btp291.

Tianwei Yu and Hesen Peng. Quantification and deconvolution of asymmetric lc-ms peaks using the bi-gaussian mixture model and statistical model selection. *BMC Bioinformatics*, 11(1):559, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-559. URL `http://www.biomedcentral.com/1471-2105/11/559`.

Tianwei Yu, Hesen Peng, and Wei Sun. Incorporating nonlinear relationships in microarray missing value imputation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8:723–731, 2011. ISSN 1545-5963. doi: http://doi.ieeecomputersociety.org/10.1109/TCBB.2010.73.

S. Zevin and N. L. Benowitz. Drug interactions with tobacco smoking. an update. *Clin Pharmacokinet*, 36(6):425–38, Jun 1999.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735. URL `http://pubs.amstat.org/doi/abs/10.1198/016214506000000735`.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006. doi:

10.1198/106186006X113430. URL http://pubs.amstat.org/doi/abs/10.1198/106186006X113430.