2018

Classification of high-dimensional data based on multiple testing methods

Chong Ma
University of South Carolina

Classification of high-dimensional data based on multiple testing methods

by

Chong Ma

Bachelor of Science
Nankai University 2011

Master of Science
Bowling Green State University 2013

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2018

Accepted by:

David Hitchcock, Major Professor

Paramita Chakraborty, Major Professor

Yen-Yi Ho, Committee Member

John Grego, Committee Member

Steven L. Morgan, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

I would like to thank the people who stood me during my phd study at University of South Carolina. Without their support, there is no way for me to being accomplishing this great journey.

First of all, I would like to express my sincere gratitude to my advisors David B. Hitchcock, Paramita Chakrabory, and Yen-Yi Ho, and my other thesis committee-members: John Grego and Stephen L. Morgan, for their generous support and invaluable ideas.

My dissertation wound not have been possible without the support of David. I am very thankful to David for guiding me into research and consistently supporting me in any aspect. It is my honor to have David being a great adviser and a wonderful person to work with. He positively shares his insightful thoughts, patiently discusses ideas with me and always encourages me to create novel methods. He is very supportive and encourages me to explore applied statistics across interdisciplinary areas. I have learned a lot of things from David about functional data analysis which will benefit and influence my whole career.

I would like to express my great appreciation and extensive gratitude to the Paramita-John-James research group. I want to especially thank Paramita as an enthusiastic, passionate, and patient mentor in the projects on which I work. Paramita has generously provided me invaluable support in accomplishing my dissertation and carrying on academic career. John is a great person to nicely lead me into their research team and gives me tons of help in statistical computing. I am enormously thankful to James Lynch for sharing his deep and broad thinking with or without related to the project and lifting my eyes in different academic topics.

I am deeply indebted to Yen-Yi for being nicely mentoring me on the bioinformatic field. I would like to convey my sincere thanks and humble appreciation to Yen-Yi, for her time, patience and generous help. Yen-Yi is very enthusiastic and supportive for sharing her sparkling thoughts on genetics study, and inspires me exploring new knowledge and ideas in biology. Without her nice support, I could have spent more numerous time on mining the genetic and genomics study.

Thanks to my peers and other faculty and staff members in statistics department, they make my life easier and enjoyable. I would like to warmly thank my parents, my brothers and sisters-in-law who always stand me and encourage me to purse my dreams. Because of their endless support and love, it makes me fill with courage to overcome any challenges and finally achieve the success.

Abstract

Supervised and unsupervised classification are common topics in machine learning in both scientific and industrial fields, which usually involve three tasks: prediction, exploration, and explanation. False discovery rate (FDR) theory has a close connection to classical classification theory, which must be employed in a sophisticated way to achieve good performance in various contexts. The study aims to explore novel supervised classifiers and unsupervised classification approaches for functional data and high-dimensional data in genome study by using FDR, respectively. One work develops a novel classifier for functional data by casting the classification problem into a multiple testing task, which involves using statistical depth functions. The other two works essentially deal with p-values or tail-areas by using FDR in the large scale testing problem. One work proposes a novel algorithm to yield reproducible differential expression analysis for microarray and RNA-Seq data. The proposed algorithm combines the cross-validation type subsampling and false discovery rate, where the p-values obtained from the training data are used to fit a mixture of baseline and signal distributions by using the EM algorithm, which is in turn used to screen the significance for the p-values obtained from the testing data. Another work proposes a novel weighted p-value approach to explore the association between microRNAs and COPD emphysema severity by regulating the mRNA expressions, while integrating patient phenotype information. This proposed method can be applied to study the causality between miRNA and any particular disease, by exploring the precise role of miRNA in regulating genes.

Table of Contents

List of Tables

List of Figures

Chapter 1

Supervised Classification for Functional Data[1]

1.1 Introduction

In the past twenty years, functional data have been increasingly studied theoretically and functional data analysis has been applied in various fields such as physical science, genomes, forensic science, economics, and finance. Nowadays functional data analysis is becoming even more popular because the ubiquity of advanced data-gathering technology has made high-dimensional data common. With functional data, we observe a response function $Y(t)$ at an ordered set of measurement points $t_1, \ldots, t_n$ supported in a compact interval $\mathcal{I}$. Functional data may arise as temporal, spatial, electrical, or spectral measurements, among other applications. Two common goals of functional data analysis include (1) estimating the distribution of a functional random variable and (2) predicting the response related to the functional data. The popular methods for estimating the density of the functional random variable include nonparametric or distribution-free approaches based on estimators of the Nadaraya-Watson type. It is necessary to point out that the concept of a density for the functional random variable is difficult to define (Ferraty and Vieu, 2002, Hall and Heckman, 2002). (2) usually refers to functional data smoothing, functional principal component analysis and functional linear models, all of which have counterparts in multivariate analysis or generalized linear regression (Ramsay and Silverman, 2005).

Supervised classification for functional data has recently gained popularity in var-

[1] Chong Ma, David B. Hitchcock, Stephen L. Morgan. To be submitted to Journal of Applied Statistics

ious industrial fields such as speech recognition, differential analysis of gene expressions, disease diagnosis in public health, risk identification in finance, and so on. Penalized discriminant analysis (PDA) (Hastie et al., 1995), considered an early application of functional data analysis, cast the classification problem into a logistic regression framework via optimal scoring. PDA is a penalized version of linear discriminant analysis (LDA), where the within-class covariance in Mahalanobis distance in PDA is regularized by smoothing the discretized functional observations, in order to avoid the degeneracy of the inverse within-covariance. Recently, Ferraty and Vieu (2003) used a functional nonparametric approach for curve discrimination. Their approach is based on a kernel-type estimator of posterior probability with a tuning parameter bandwidth h. This approach also has a $k-$Nearest Neighbors (kNN) version that replaces the real-valued tuning parameter h with an integer parameter k (among a finite set). Recently, Llop et al. (2011) proposed a new nonparametric classification rule based on a $\sqrt{n}-$consistent nonparametric estimator for the marginal density function of an order-one stationary process. Ramsay and Silverman (2005) introduced how functional principal component analysis and canonical correlation analysis work in classification.

Recently, the notion of depth has become an important tool in classifying high-dimensional data, especially functional data. The concept of depth was originally developed for multivariate data, aiming to order them from center outwards, such that the more central observations have larger depths, and vice versa. Zuo and Serfling (2000a) summarized the general notions of statistical depth functions and proposed the key structural properties that statistical depth functions should satisfy (Liu, 1990). Recently, the concept of depth has been extended to functional data. Fraiman and Muniz (2001) proposed the integration data depth (ID), which is an integration of a univariate data depth analogous to Tukey's half-space depth on the supported compact domain $\mathcal{I}$. Later, López-Pintado and Romo (2009) proposed the band depth

(BD) and the generalized band depth (GBD). The GBD is closely related to ID, since MBD can also be considered as an integration of the univariate simplicial depth over the time points. Recently, Narisetty and Nair (2016) proposed extremal depth (ED) which is based on the "extreme outlyingness", as opposed to ID and GBD which consider more the centrality. Not surprisingly, ED is more resistant to functions that are outlying in small regions of the domain. Obviously, each depth function can be used to classify new curves by the essential structure properties of depth functions (Zuo and Serfling, 2000a) including maximality at center, monotonicity relative to deepest point and vanishing at infinity. Among the depth-based classification methods, the most straightforward classifier is the maximum depth rule (Ghosh and Chaudhuri, 2005) which assigns a new functional observation to the group within which it has the largest depth. Besides directly comparing depths in each group, distance-based rules appear in Cuevas et al. (2007) and López-Pintado and Romo (2006). The concepts of depth mainly focus on finding the most representative curve and detecting outlying curves. There are relatively fewer articles that study the distribution of depth for the purpose of classification.

In this article, we propose a novel method applying the multivariate functional depth for supervised classification of functional data. Instead of merely using the univariate functional observations, we propose to augment a univariate functional observation, creating a $(p+1)$-vector of functions by taking derivatives up to the $p-$th order. By taking into account the derivatives, we can use the multivariate functional depth to best capture the shapes, amplitude and phase variations of curves in various groups. As the functional depth is an extension of the multivariate depth, the multivariate functional depth is a combination of the functional depth and the multivariate depth. The simplicial band depth (López-Pintado et al., 2014) is essentially a Lebesgue measure of the region where a given multivariate function $\mathbf{x}(\mathbf{t})$ is contained in the simplicial region determined by $\mathbf{X_1}(\mathbf{t}), \ldots, \mathbf{X_{p+1}}(\mathbf{t})$. The simplicial region is a $(p+1)$-dimension "tunnel" which consists of a convex hull of $(p+1)$ vertices at each time point t. The multivariate functional halfspace depth (Claeskens et al., 2014) is an extension of the integration data depth (Fraiman and Muniz, 2001) which is a weighted average of Tukey's half-space over the time points on the domain $\mathcal{I}$. Hlubinka et al. (2015) proposed a modification of the integration data depth that takes into account the derivatives of smoothed functions. The modified integration data depth is also essentially a weighted average of integration data depths involving different order of derivatives, i.e., $ID(\mathbf{x};\mathbb{F}_{\mathbf{X}}) = \int D(\mathbf{x}(\mathbf{t});\mathbb{F}_{\mathbf{X}(\mathbf{t})})\,d\mathbf{t}$, where $\boldsymbol{x} = (x^{(i_1)},\ldots,x^{(i_l)})'$ is a vector of an observed curve and its derivatives, $\boldsymbol{X} = (X^{(i_1)},\ldots,X^{(i_l)})'$ is the vector of the corresponding functional random variable and its derivatives and $\mathcal{F}_{\mathbf{X}}$ is the cumulative distribution for $\mathbf{X}$, where the set of derivative orders is $\{i_1,\ldots,i_l\} \in \{0,\ldots,p\}$. A 2-fold cross-validation is used to select the optimal weights for the integrated data depth for achieving the minimum misclassification rate. Rather than considering a data-driven method, we propose a model-based classifier based on the depth functions. Our model-based classifier is constructed based on legitimate depth functions, and a smart choice of depth function can enhance the classification power of our proposed method. In addition, our work is also motivated by the DD-classifier (Li et al., 2012) which is quite sophisticated and which achieves an optimal polynomial curve separation in the depth-versus-depth plot (DD-plot). The DD-classifier is especially for multivariate data and it is of interest whether if the same theoretical result could also hold for functional data.

The paper is organized as follows. Section 2 briefly reviews some commonly used depth functions for multivariate data and functional data. The multivariate functional depths are introduced as well. We propose a novel classification method for functional data based on multivariate functional depth in Section 3. Section 4 briefly introduces some conventional supervised classification methods in both multivariate and functional context which will be compared later to our proposed method in

Section 5 via simulation studies and in Section 6 via real data applications. We give a conclusion in Section 7 and some acknowledgments in Section 8.

1.2 Background for Functional Depth

1.2.1 Notation

Consider a probability space $(\Omega, \mathcal{F}, P)$ where Ω is the space and $\mathcal{F}$ is an appropriate σ algebra on Ω and P is a probability measure. Let $\mathcal{I} \in \mathcal{B}(\mathbb{R})$ be a compact interval, a stochastic process is a mapping $X : (\mathcal{I}, \Omega) \rightarrow \mathbb{R}$ such that $X(t, \cdot)$ is measurable for every $t \in \mathcal{I}$. For notational convenience, denote by X this stochastic process, and assume it is differentiable up to p times. Denote by $C(\mathcal{I})^{p+1}$ the collection of continuous stochastic processes differentiable up to p times and then $\{X : X(t), t \in \mathcal{I}\} \in C(\mathcal{I})^{p+1}$. The bold capital letters (e.g. $\mathbf{X}$) are used to represent a vector of continuous functions $(X_0, X_1, \ldots, X_p)$ where X_i could be a certain derivative of X or some other transformation of X. The corresponding smaller letters x and $x(t)$ refer to the observed stochastic trajectory and its specific value at time t. And $\mathbf{x}$ and $\mathbf{x}(t)$ are the corresponding observed $p+1$-variate curves and $(p+1)-$variate point at time t. Without loss of generality, we set $\mathcal{I} = [0, 1]$.

1.2.2 Depth Functions

The depth function has been proposed for multivariate data for ordering the multivariate data from center outward such that the most central datum has the largest depth and the least central datum has the smallest depth. Zuo and Serfling (2000b) summarized statistical depth functions in terms of multivariate data and also established the desirable structural properties (Liu, 1990) that a legitimate statistical depth function should satisfy. For functional data, more specifically, univariate functional data, most depth functions are extensions of multivariate depth functions, with the caveat that some property that holds for multivariate data may not hold for functional data. Likewise, the multivariate functional depths are also related to functional or multivariate depth functions. In this article, we briefly introduce three multivariate functional depths: the multivariate functional halfspace depth (Claeskens et al., 2014), the multivariate functional h-Mode depth (Cuevas et al., 2007) and the multivariate functional simplicial band depth (López-Pintado et al., 2014).

Assume that $Y \in C(\mathcal{I})^{p+1}$ with respect to cdf F_Y, and then let $\mathbf{Y} = (Y^{(0)}, Y^{(1)}, \dots, Y^{(p)})'$ be a $(p+1)-$variate stochastic process with cdf $F_{\mathbf{Y}}$, where $Y^{(0)} = Y$ and $Y^{(i)}$ is the i^{th} derivative of Y. Consider an arbitrary $X \in C(\mathcal{I})^{p+1}$.

Definition 1.2.2.1. Multivariate Functional Halfspace Depth

Let $Z(t) = HD(\mathbf{X}(t); F_{\mathbf{Y}(t)}) = \inf_{\mathbf{u}\in\mathbb{R}^{p+1}, ||\mathbf{u}||=1} P(\mathbf{u}'\mathbf{Y}(t) \geq \mathbf{u}'\mathbf{X}(t)), \mathbf{X}(t) \in \mathbb{R}^{p+1}$. The population version of the multivariate functional halfspace depth for an arbitrary $\mathbf{X}$ with respect to $F_{\mathbf{Y}}$ is

$$D(\mathbf{X}; F_{\mathbf{Y}}) = \int_0^1 Z(t) \cdot w(t)\, dt, \tag{1.1}$$

where $w(t)$ is the weight function that may or may not depend on $F_{Y(t)}, t \in [0,1]$.

The multivariate functional halfspace depth (Claeskens et al., 2014) is a weighted average of the multivariate Tukey's halfspace depths over the time points, and the weight is actually a function of the time t that can be chosen to account for the local amplitude variability in order to reflect the functional nature of the data. It is essentially a sophisticated integrated data depth, since it is an integration of the weighted Tukey's halfspace depths over the domain $\mathcal{I} = [0,1]$. In this article, we set the weight function $w(t)$ uniform over the time domain $[0,1]$. The finite-sample version for the multivariate functional halfspace depth as in (1.1) based on $\mathbf{Y}_1, \dots, \mathbf{Y}_N \overset{i.i.d}{\sim} F_{\mathbf{Y}}$ is

$$D(\mathbf{X}; F_{\mathbf{Y},N}) = \int_0^1 Z_N(T)\, dt$$

where $Z_N(t) = HD(\mathbf{X}(t); F_{\mathbf{Y}(t),N}) = \frac{1}{N} \min_{\mathbf{u}\in\mathbb{R}^{p+1}, ||\mathbf{u}||=1} \#\{Y_n(t), n = 1, \dots, N : \mathbf{u}'\mathbf{Y}_n(t) \geq \mathbf{u}'\mathbf{X}(t)\}, \mathbf{X}(t) \in \mathbb{R}^{p+1}$

Definition 1.2.2.2. Multivariate Functional h-mode Depth

Let $m(\mathbf{X}, \mathbf{Y}) = \sqrt{||X - Y||^2 + ||X^{(1)} - Y^{(1)}||^2 + \cdots + ||X^{(p)} - Y^{(p)}||^2}$ be the metric for $(p+1)-$variate curves where $|| \cdot ||^2$ is the squared Euclidean L_2 norm such that $||X^{(0)} - Y^{(0)}||^2 = \int_0^1 (x(t) - y(t))^2 \, dt$. The population version of the multivariate functional h-mode depth for an arbitrary $\mathbf{X}$ with respect to $F_{\mathbf{Y}}$ is

$$D(\mathbf{X}; F_{\mathbf{Y}}) = E_{\mathbf{Y}}[K_h(m(\mathbf{X}, \mathbf{Y}))] \tag{1.2}$$

where $K_h(t)$ is a scaled asymmetric kernel such that $K_h(t) = \frac{1}{h} K(\frac{t}{h})$. Here K is a right-truncated normal probability density function since the metric m is non-negative and h is a tuning bandwidth parameter which takes a default value in the `depth.modep` function in the `fda.usc` R package.

The multivariate functional h-mode depth (Cuevas et al., 2007) for $\mathbf{X}$ with respect to $F_{\mathbf{Y}}$ measures how surrounded the $(p+1)-$variate set of curves $\mathbf{X}$ is in the $(p+1)$-variate stochastic process $F_{\mathbf{Y}}$. The finite-sample version for the multivariate functional h-Mode depth as in (1.2) based on $\mathbf{Y}_1, \ldots, \mathbf{Y}_N \overset{i.i.d}{\sim} F_{\mathbf{Y}}$ is

$$D(\mathbf{X}; F_{\mathbf{Y},N}) = \frac{1}{N} \sum_{i=1}^{N} K_h(m(\mathbf{X}, \mathbf{Y}_i))$$

As is well known, the multivariate functional halfspace depth is a type of integrated data depth (Fraiman and Muniz, 2001) which is related to the extreme depth function proposed by Narisetty and Nair (2016) and the multivariate simplicial depth function proposed by López-Pintado et al. (2014). It is certain that there are a lot of statistical depth concepts for functional data; however, in this article, we focus on the multivariate functional halfspace depth and the h-mode depth, since our method is built to perform with any legitimate statistical depth function. However, a good performance in the classification of functional data depends on a smart choice of the functional depth function.

1.3 Multivariate Functional Depth Classifier

In a multi-group supervised classification problem, assume we have K independent groups of functional data in the form of $(x_i, g_i), i = 1, \ldots, n$, and $g_i = k$ for some $k \in 1, \ldots, K$ on the compact domain $\mathcal{I}$ ($\mathcal{I} = [0, 1]$). Each group consists of $\{x_1^k, \ldots, x_{n_k}^k\}$ i.i.d. with each group member being a realization of an unknown stochastic process X_k with the cumulative distribution F_{X_k}, where $n_1 + \ldots + n_K = N, k = 1, \ldots, K$.

A statistical depth function measures how central a curve (or a vector of curves) is with respect to a group of curves (a curve vectors) in terms of an appropriate metric. Put it in another way, the depths of all groups of curves with respect to a certain target group reflect the similarity of these groups of curves to that target group. The larger the depth of a curve to a target group, the more similar the curve is to that target group. Given a certain target group, the depths of all groups of curves with respect to that target group can be assumed to follow a mixture model.

Step 1. Calculate depths of all groups of curves to each group $1, 2, \ldots, K$ by using an appropriate depth function, respectively. Denote depths of curve x_i in each group $1, 2, \ldots, K$ by $d_{i1}, d_{i2}, \ldots, d_{iK}$. For the sake of easy interpretation and model fitting, we calculate the ratio of depths by dividing d_{ik} by the summation of $d_{i1}, d_{i2}, \ldots, d_{iK}$ for each x_i, that is, $t_{ik} = \frac{d_{ik}}{d_{i1}+d_{i2}+\ldots+d_{iK}}, k = 1, 2, \ldots, K$. Therefore, $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{iK})'$ is K-dimensional probability vector and its components sum to one.

$$
\begin{array}{cc}
 & \overbrace{\begin{array}{cccc} 1 & 2 & \cdots & K \end{array}}^{\text{Groups}} \\
\begin{array}{c} g_1 = 1 \\ g_2 = 2 \\ g_3 = 3 \\ g_4 = 4 \\ \vdots \\ g_N = K \end{array} &
\begin{pmatrix}
d_{11} & d_{12} & \cdots & d_{1K} \\
d_{21} & d_{22} & \cdots & d_{2K} \\
d_{31} & d_{32} & \cdots & d_{3K} \\
d_{41} & d_{42} & \cdots & d_{4K} \\
\vdots & \vdots & \ddots & \vdots \\
d_{N1} & d_{N2} & \ldots & d_{NK}
\end{pmatrix}
\end{array}
\Rightarrow
\begin{array}{c}
\overbrace{\begin{array}{cccc} 1 & 2 & \cdots & K \end{array}}^{\text{Groups}} \\
\begin{pmatrix}
t_{11} & t_{12} & \ldots & t_{1K} \\
t_{21} & t_{22} & \ldots & t_{2K} \\
t_{31} & t_{32} & \ldots & t_{3K} \\
t_{41} & t_{42} & \ldots & t_{4K} \\
\vdots & \vdots & \ddots & \vdots \\
t_{N1} & t_{N2} & \ldots & t_{NK}
\end{pmatrix}
\end{array}
$$

Step 2. The k-th depth ratio t_k is assumed to be a K-mixture of logit-normal distributions, since the k-th depth ratios are calculated from the K groups of curves. Each component of the K-mixture of the k-th depth ratio t_k represents the similarity of a group of curves with respect to the k-th group. Thus, $\text{logit}(t_k)$ follows a K-mixture of Gaussian distributions such that

$$P(\text{logit}(t_k)) = \sum_{j=1}^{K} \pi_{kj}\phi\left(\text{logit}(t_k); \mu_{kj}, \sigma_{kj}\right)$$

where π_{kj} is the mixing proportion such that $\sum_{j=1}^{K} \pi_{kj} = 1$.

In fact, t_k is a univariate random variable in $[0, 1]$; therefore a mixture of logit-normal distributions is just a convenient model fitting method which might be not the best universally. A mixture of Beta distributions or a nonparametric method such as kernel density estimation could have better model fit in some cases. The reason for proposing the mixture of logit-normal distributions is convenient for proving that $\{T : T < t\}$ is a UMP test in Theorem 1.3.1.

Step 3. Given a new curve Z with unknown group label, we can calculate the multivariate functional depth of Z in each group and then obtain the depth ratio of

$D(Z;\mathbb{F}_{X_k})$ to the overall depth of $\mathbf{Z}$,

$$\begin{aligned} T(z) &= \left(\frac{D(z;\mathbb{F}_{X_1})}{D(z;\mathbb{F}_{X_1}) + \ldots + D(z;\mathbb{F}_{X_K})}, \ldots, \frac{D(z;\mathbb{F}_{X_K})}{D(z;\mathbb{F}_{X_1}) + \ldots + D(z;\mathbb{F}_{X_K})}\right) \\ &= (T_{X_1}(z), \ldots, T_{X_K}(z)) \end{aligned}$$

Intuitively, the larger the depth ratio of Z in a certain group, the more likelihood it belongs to the according group. We format the multi-group classification problem into a set of K hypothesis tests,

$$\mathrm{H}_0^k : Z \sim \mathbb{F}_{X_k} \text{ vs. } \mathrm{H}_a^k : Z \nsim \mathbb{F}_{X_k}$$

where $k = 1, \ldots, K$ and we assume Z must come from a certain group among the K groups. Thus, in each hypothesis test, either H_0^k or H_a^k must be correct. Recall that t_k is the observed depth ratio relative to group k. Under H_0^k, we take $T_{X_k}(Z)$ as the test statistic and let the observed tail area be $\Gamma(t_k) = \{T_{X_k}(Z) : T_{X_k}(Z) < t_k\}$, which is a Uniformly Most Powerful test shown by Theorem 1.3.1 under some assumptions. Because the true distribution of t_k depends on the distribution of K random functions and the depth function, it is barely possible to find the best fitted model. Though the mixture of logit-normal is an approximation to the true model that makes the Theorem 1.3.1 limited, the observed tail area $\{T : T < t\}$ intuitively makes sense in that the smaller the depth ratio of Z under H_0^k, the stronger evidence against H_0^k.

Storey et al. (2003) connected FDR to classical classification theory, in which he formulated multiple hypothesis testing as a classification problem by minimizing a weighted average of false discovery rate (FDR) and false nondiscovery rate (FNR). The “classification” in that work actually refers to unsupervised classification in which group labels for subjects are unknown. In this paper, we connect the FDR theory to supervised classification in which the group labels are known. We propose an M_1 score to measure the test’s accuracy for testing H_0^k by taking the harmonic mean of the negative predictive value (NPV) and false discovery rate (FDR). The M_1 score depends on the observed tail area $\Gamma(t_k)$. Thus, for the kth hypothesis test,

$$\mathrm{M}_1(t_k) = 2 \cdot \frac{1}{\frac{1}{\mathrm{NPV}(t_k)} + \frac{1}{\mathrm{FDR}(t_k)}} \tag{1.3}$$

where

$$\begin{aligned} \mathrm{NPV}(t_k) &= P(\mathrm{H}_0^k|\Gamma(t_k)^c) = \frac{\pi_{kk} P(T_{X_k}(Z) \geq t_k | Z \in \mathbb{F}_{X_k})}{\sum_{j=1}^{K} \pi_{kj} P(T_{X_k}(Z) \geq t_k | Z \in \mathbb{F}_{X_j})} \\ \mathrm{FDR}(t_k) &= P(\mathrm{H}_0^k|\Gamma(t_k)) = \frac{\pi_{kk} P(T_{X_k}(Z) < t_k | Z \in \mathbb{F}_{X_k})}{\sum_{j=1}^{K} \pi_{kj} P(T_{X_k}(Z) < t_k | Z \in \mathbb{F}_{X_j})} \end{aligned}$$

For each curve, we conduct K hypothesis tests on $\mathrm{H}_0^k : Z \sim \mathbb{F}_{X_k}$ versus $\mathrm{H}_0^k : Z \nsim \mathbb{F}_{X_k}$, $k = 1, 2, \ldots, K$. For easy interpretation, one can take the ratio of the measure of strength of $\mathrm{M}_1(t_k)$ to the sum of the M_1's, that is,

$$Q(Z \sim F_{X_k}|Z) = \frac{\mathrm{M}_1(t_k)}{\sum_{j=1}^{K} \mathrm{M}_1(t_j)}$$

subject to the constraint $\sum_{k=1}^{K} Q(Z \sim F_{X_k}|Z) = 1$. Therefore, the classifier for curve Z is determined by

$$\arg\max_k Q(Z \sim \mathbb{F}_{X_k}|Z) = \arg\max_k \mathrm{M}_1(t_k)$$

FDR is the global false discovery rate (Efron, 2007, Storey, 2007) which is a measure of the expected rate of false positives to all significant tests, taking into account all hypothesis tests. From a Bayesian perspective, FDR is a Bayesian sort of p-value. More specifically, $\mathrm{FDR}(t_k)$ is the posterior probability under H_0^k that a curve with a depth ratio as small or smaller than the observed depth ratio is truly from group k. A smaller $\mathrm{FDR}(t_k)$ gives stronger evidence for believing that it comes from H_1^k. On the other hand, a large $\mathrm{FDR}(t_k)$ in turn implies the curve is more likely to belong to H_0^k. Analogously, NPV is a measure of the expected rate of true positives to all nonsignificant tests, simultaneously considering all hypothesis tests. Specially, $\mathrm{NPV}(t_k)$ is a posterior probability under H_0^k that a curve with a depth ratio as large or larger than the observed depth ratio is truly from group k. Under H_0^k, either a large $\mathrm{FDR}(t_k)$ or large $\mathrm{NPV}(t_k)$ can be used as a measure of strength of that curve Z is truly from group k. When both of them are large simultaneously, it shows stronger evidence that curve Z belongs to group k. Conversely, when either of them is small, there is less evidence that curve Z belongs to group k.

We define a more general definition of M_1 score, denoted M_β where β is nonnegative. When $\beta = 0$, M_β equals $\mathrm{NPV}(t_k)$; when $\beta = 1$, $\mathrm{M}_\beta = \mathrm{M}_1$. In other words, one can decide how important the negative predictive value is relative to the false discovery rate on the measure of strength of evidence that a curve comes from H_0. For simplicity, we choose $\beta = 1$ in the following expression:

$$\mathrm{M}_\beta(t_k) = (1+\beta)\cdot\frac{\mathrm{NPV}(t_k)\mathrm{FDR}(t_k)}{\beta\mathrm{NPV}(t_k)+\mathrm{FDR}(t_k)} \tag{1.4}$$

In later sections, we will explore the effect of the choice of β on the classification results.

Theorem 1.3.1. *Suppose T is the test statistic for hypothesis test $\mathrm{H}_0 : Z \sim \mathbb{F}_X$ versus $\mathrm{H}_1 : Z \nsim \mathbb{F}_X$. Assume that $T|H \sim (1-H)\cdot\mathrm{F}_0 + H\cdot\mathrm{F}_1$ where $H = 0$ if Z is truly from H_0 and $H = 1$ if Z is truly from H_1. $H \sim \mathrm{Bernoulli}(1-\pi_0)$. Assume F_0 is a logit-normal distribution and F_1 is a mixture of logit normal distributions where the corresponding densities $f_0(t) =$ logit-norm$(t;\mu_0,\sigma)$ and $f_1(t) = \sum_{i=1}^{r}\gamma_i$logit-norm$(t;\mu_{1i},\sigma)$. Assuming that $\mu_0 > \max_i \mu_{1i}$, the uniformly most powerful test is $\{T : T < t\}$ at the size $\alpha = P(T < t|\mathrm{H}_0)$.*

Proof. Recall that by the Neyman-Pearson lemma we can have the set of observed tail areas $\mathcal{A}(\lambda)$ for $0 \le \lambda \le 1$ formed by

$$\mathcal{A}(\lambda) = \left\{ t : \frac{\pi_0 f_0(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)} \le \lambda \right\}$$

We will show that the set of observed tail areas $\mathcal{A}(\lambda)$ has the form $\{T : T < t\}$ in which t is related to λ. Note that $\mathcal{A}(\lambda)$ can be written as $\{t : \frac{f_1(t)}{f_0(t)} \ge \lambda\}$. Since

$f_0(t) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{t(1-t)}e^{-\frac{(\mathrm{logit}(t)-\mu_0)^2}{2\sigma^2}}$, $f_1(t) = \sum_{i=1}^{r}\gamma_i\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{t(1-t)}e^{-\frac{(\mathrm{logit}(t)-\mu_{1i})^2}{2\sigma^2}}$, then

$$\frac{f_1(t)}{f_0(t)} = \sum_{i=1}^{r}\gamma_i e^{\frac{\mu_0^2-\mu_{1i}^2}{2\sigma^2}} e^{\frac{\mu_{1i}-\mu_0}{2\sigma^2}\mathrm{logit}(t)} \geq r\cdot\min_i\gamma_i\cdot e^{\frac{\mu_0-\max\limits_i\mu_{1i}}{2\sigma^2}}\cdot e^{\frac{\min\limits_i\mu_{1i}-\mu_0}{2\sigma^2}\mathrm{logit}(t)} \geq \lambda$$

Note that $\mu_0 > \max\limits_i \mu_{1i} \geq \min\limits_i \mu_{1i}$, then $\min\limits_i \mu_{1i} - \mu_0 < 0$. And $\mathrm{logit}(t) = \log\left(\frac{t}{1-t}\right)$ is nondecreasing, so for any $\lambda \in (0,1)$, there exists a λ^* such that $\mathcal{A}(\lambda) = \{t : \frac{f_1(t)}{f_0(t)} \geq \lambda\} = \{t : t < \lambda^*\}$. Therefore, under the assumption, the observed tail area has the form $\{T : T < t\}$, and by the Neyman-Pearson Lemma, $\{T : T < t\}$ is a UMP test at the size $\alpha = P(T < t|\mathrm{H}_0)$. □

1.4 Classical Classification Methods and Depth-Based Classifiers

In this section, we relate our method to some conventional classification methods in the multivariate context and pure depth-based classifiers for functional data. In practice, functional data are discretized curves on a fine mesh that has infinite dimension theoretically but consists of many closely spaced measurement points. One approach to classification is to use functional principal component analysis (PCA) to reduce the infinite-dimensional curves to a finite-dimensional multivariate vectors, for the sake of applying conventional classification methods in the multivariate context. There are two approaches to ensuring smooth eigenfunctions during the implementation of the functional PCA. One is regularized principal component analysis based on the raw discretized curves, in which we find orthonormal eigenfunctions ξ_p, $p = 1, 2, 3, \ldots$ to maximize the penalized variance

$$\frac{\mathrm{var}(\int \xi_p(t)x_i(t)\,dt)}{||\xi_p||^2 + \lambda\int \xi_p''(t)^2\,dt}$$

subject to the constraints $\int \xi_p(t)\xi_q(t)\,dt + \int D^2\xi_p(t)D^2\xi_q(t)\,dt = 0$, for $p \neq q$. Here $D^2\xi(t) = \xi''(t)$. The other approach is principal component analysis on functional observations which have been smoothed via some appropriate spline smoothing technique. This is more convenient for conducting functional PCA, whose goal is to find orthonormal eigenfunctions ξ_p, $p = 1, 2, 3, \ldots$ to maximize the variance $\mathrm{var}(\int \xi_p(t) x_i(t)\, dt)$ subject to the constraints $\int \xi_p(t)\xi_q(t)\, dt = 0$ for $p \neq q$. The two functional PCA techniques perform similarly in terms of their impact on classification results. In our paper, we use the second one, i.e., PCA on smoothed functional observations, and assume that functional data are preprocessed properly. In other words, we assume curve registration and curve centering (subtract the mean curve from each curve) have been done before performing functional PCA.

The previously proposed distributional depth-based classifiers is built upon the original functional observations. For the sake of comparison to our method, we carry out two pure depth-based classifiers, that is, the Distance to the Trimmed Mean (DS) and the Trimmed Averaged Mean Distance (TAD) classifiers (López-Pintado and Romo, 2006), which is constructed on the generalized band depth. Both of the functional PCA and pure depth-based classification methods will be described in following.

1.4.1 Functional Principal Component Analysis (PCA) in Classification

Functional principal component analysis is used to reduce infinite-dimensional curves into conventional multivariate vectors composed of a set of appropriate finite-dimensional principal component scores that accounts for most of the variation among the curves. Assume that the functional observations are preprocessed by using an appropriate smoothing technique. In this paper, we apply B-spline smoothing to preprocess the discretized functional observations, in which the regularization parameter is determined by the generalized cross-validation method and the penalized term is the fourth derivative of the curves. We employ the functional PCA technique to convert functional observations into a number of principal component scores so that the components account for at least 90% of the total variation in the curves. Our study used the first four principal component scores, which usually dominate the overall variation of functional data in our study, though the number of principal component scores selected usually depends on the particular functional data set. Finally, we cast the classification problem for functional data as a classification task for the resulting multivariate data. We conduct functional principal component analysis as follows.

Step 1. Split the whole functional observations into a training set $\{(x^i_{\text{train}}, g^i_{\text{train}}), i = 1, 2, \ldots, n\}$ and a test set $\{(x^j_{\text{test}}, g^j_{\text{test}}), j = 1, 2, \ldots, m\}$. For all of the curves, subtract the curves from the mean curve $\bar{x}_{\text{train}}$ in the training set. Thus, the training set and test set in the functional PCA study yield $\{(\tilde{x}^i_{\text{train}}, g^i_{\text{train}}), i = 1, 2, \ldots, n\}$ and $\{(\tilde{x}^j_{\text{test}}, g^j_{\text{test}}), j = 1, 2, \ldots, m\}$ where $\tilde{x}^i_{\text{train}} = x^i_{\text{train}} - \bar{x}_{\text{train}}$ and $\tilde{x}^j_{\text{test}} = x^j_{\text{test}} - \bar{x}_{\text{train}}$.

Step 2. Based on the training set, calculate the first four principal component weight functions $\xi_1, \xi_2, \xi_3, \xi_4$ which are orthonormal. Convert the curves in the training and test sets into their corresponding multivariate vectors, which consist of a vector of four principal component scores. That is, $\tilde{x}$ is represented by $y = (\int \tilde{x}\xi_1, \int \tilde{x}\xi_2, \int \tilde{x}\xi_3, \int \tilde{x}\xi_4)'$. Here $\tilde{x}$ refers to an arbitrary curve in training and test sets and $\int \tilde{x}\xi_p = \int_{\mathcal{T}} \tilde{x}(t) \cdot \xi_p(t)\, dt$, $p = 1, 2, 3, 4$.

Step 3. Apply the conventional classification methods in the multivariate context including linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), generalized linear model (GLM), support vector machine (SVM), neural network (NNET), mixture discriminant analysis and flexible discriminant analysis (FDA) (Friedman et al., 2001), respectively. Based on the multivariate vectors of principal component scores, obtain the estimates of the group means and variance-covariance matrices from the training set, which are used to predict the group labels on the test set. These methods serve as standards of comparison to our method.

1.4.2 Depth-Based Classifiers

The pure depth-based classifiers used here are the Distance to the Trimmed Mean (DS) and the Trimmed Weighted Averaged Distance (TAD) (López-Pintado and Romo, 2006), which is built upon the generalized band depth defined as follows.

Definition 1.4.2.1. Generalized Band Depth

Let $Y_1, Y_2, \dots, Y_N \overset{i.i.d}{\sim} F_Y$ and $X \sim F_X$. Denote by $S^{(j)} = E[\lambda_r(A(X; Y_{i_1}, Y_{i_2} \dots, Y_{i_j}))]$ where $A(X; Y_{i_1}, Y_{i_2}, \dots, Y_{i_j}) = \{t \in \mathcal{I} : \min_{k=1,\dots,j} Y_{i_k}(t) \le X(t) \le \max_{k=1,\dots,j} Y_{i_k}(t)\}$ and $\lambda_r(A(X; Y_{i_1}, Y_{i_2}, \dots, Y_{i_j})) = \lambda(A(X; Y_{i_1}, Y_{i_2}, \dots, Y_{i_j}))/\lambda(\mathcal{I})$. The generalized band depth is defined by

$$D(X; F_Y) = \sum_{j=2}^{J} S^{(j)}(X; F_Y) \tag{1.5}$$

The sample version of $S^{(j)}(X; F_Y)$ is

$$S_N^{(j)}(x) = \binom{N}{j}^{-1} \sum_{1 \le i_1 < \dots < i_j \le N} \lambda\left(t \in \mathcal{I} : \min_{k=1,\dots,j} y_{i_k}(t) \le x(t) \le \max_{k=1,\dots,j} y_{i_k}(t)\right)$$

Note $\lambda\left(t \in \mathcal{I} : \min_{k=1,\dots,j} y_{i_k}(t) \le x(t) \le \max_{k=1,\dots,j} y_{i_k}(t)\right) = \int_{\mathcal{I}} I(y_{(1)}(t) \le x(t) \le y_{(j)}(t)\, dt$, is essentially the integrated univariate simplicial depth where $I(A)$ is the indicator function such that $I(A) = 1$ if event A is satisfied and zero otherwise. The sample version of the generalized band depth is a U-statistic and has some good properties such as consistency (Liu, 1990, Zuo and Serfling, 2000b). However, calculating the combinatorial sample statistics for each observed curve leads to extensive computations as complex as $O(n^J)$. For the convenience of computation, we use the default setting of $J = 2$.

The Distance to the Trimmed Mean (DS) classifier calculates the distance from the new functional observation to the trimmed means of each group, and then classifies it to the group which is closest to the new curve in terms of the calculated distance. The trimmed mean in each group is the average of a proportion of the deepest curves in that group. Analogously, the Trimmed Weighted Averaged Distance (TAD) classifier calculates the distance from a new functional observation to each group as a weighted average of distances to a proportion of the deepest curves in that group, where the weights are determined by those group members' depths in that group, and then classifies the new curve to the group which is closest to the new functional observation in terms of the weighted average distance.

1.5 Simulation Study

In this section we conduct two simulation studies to investigate the performance for our method by comparing it to classification based on different statistical depth methods. Because our proposed method is flexible for depth methods, for simplicity, we propose to use the multivariate functional Fraiman and Muniz (FM) depth function and the multivariate functional h-mode depth function. Simulation 1 involves two groups for classification and simulation 2 consists of three groups. For each simulation study, we randomly generate a data set from each of the main effect curve with Ornstein-Uhlenbeck process error. Moreover, the generated data are contaminated with some batch effects. In reality, often functional data are repeatedly measured using different equipment in different labs in forensic analysis, resulting in such batch effects the data. Similarly, mRNA gene expressions may be observed at different times or locations. In each simulation, we assign different prior probabilities on the group membership and investigate the robustness of our proposed method.

The simulation study shows that our method performs best by using the multivariate functional h-mode depth with respect to a bivariate functional observation composed of the raw curve and its first derivative, compared to the functional PCA and pure depth-based classification approaches proposed in section 4.

1.5.1 Simulation 1. Binary Classification

Model I. Without Batch Effects

Case I. Consider a binary group classification where the classes have unequal group sizes. From the Bayesian perspective, this is tantamount to the prior probabilities of group 1 and group 2 being different. Here, assume that the prior probabilities of x_1 and x_2 are $\boldsymbol{\pi} = (0.2, 0.8)$. The simulation is conducted as follows. Randomly generate 40 curves from X_1 and 160 curves from X_2. Conduct 100 cross-validations to investigate the performance of our proposed method. In each cross-validation, randomly select 30 and 120 curves from X_1 and X_2 accordingly, which constitute the training data set, and the testing data consist of the rest. We consider three scenarios for specifications of prior probabilities, which are $\boldsymbol{\pi_1} = (0.2, 0.8)$, $\boldsymbol{\pi_2} = (0.5, 0.5)$ and $\boldsymbol{\pi_3} = (0.8, 0.2)$, respectively.

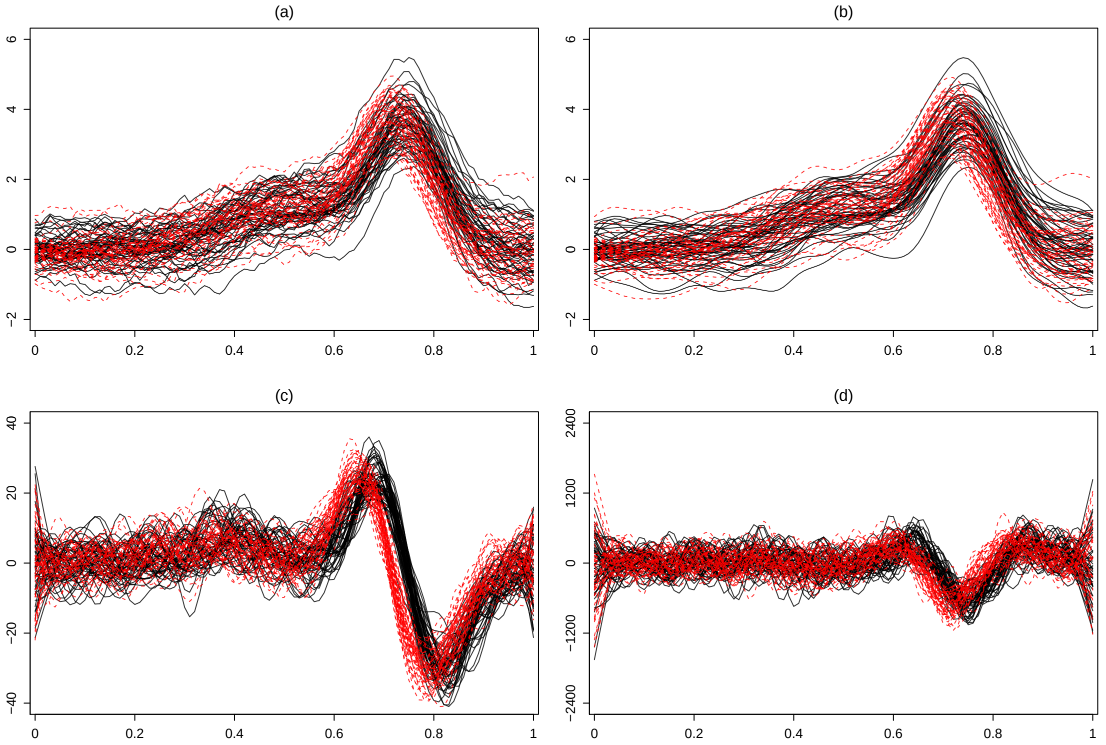


Figure 1.1: (a). 20 raw curves for two groups in Model I. (b). Smoothed curves. (c). First derivatives. (d). Second derivatives.

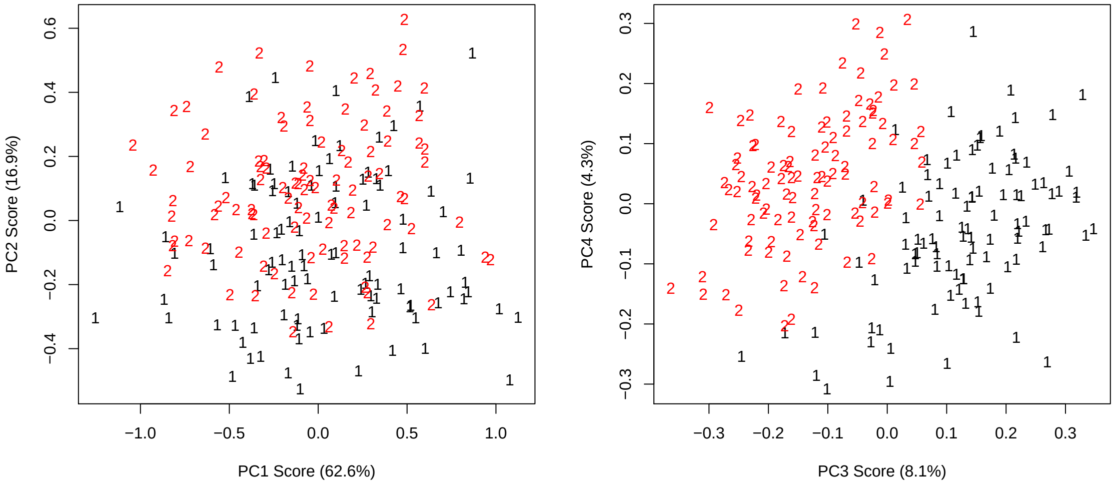


Figure 1.2: First four functional principal component scores of 20 curves for two groups in Model I. The first two scores account for about 80% of the variation of the five groups of curves, where group 1 and group 2 mask each other. However, the third and four principal component scores separate group 1 and group 2, though they explain only 12.5% of the total variation.

Case II. Consider a binary group classification where the classes have equal group sizes. Essentially, we assume that the two groups have equal prior probability. The simulation is realized as follows. Randomly generate 100 curves from x_1 and x_2 respectively. Conduct 100 cross-validations to investigate the performance of our proposed method, where each cross-validation randomly splits the whole data into training and testing data having equal sample sizes. More specifically, both the training and testing data have 50 curves from x_1 and 50 curves from x_2. Consider the same three scenarios for specifications of prior probabilities as in Case I. Our simulated curves follow the model:

$$x_{ij}(t) = \mu_i(t) + \varepsilon_j(t) \quad i = 1, 2; j = 1, \ldots, n_i \tag{1.6}$$

where

$$\mu_1(t) = 0.4\phi\left(\frac{t-0.52}{0.125}\right) + 0.6\phi\left(\frac{t-0.75}{0.224}\right) \tag{1.7a}$$

$$\mu_2(t) = 0.4\phi\left(\frac{t-0.35}{0.141}\right) + 0.6\phi\left(\frac{t-0.73}{0.1}\right) \tag{1.7b}$$

$$\varepsilon(t) = 10\int_0^t e^{-(t-s)}\,dW_s \tag{1.7c}$$

$\varepsilon(t)$ is the stationary Ornstein-Uhlenbeck process with mean $\alpha = 0$, decay-rate (growth-rate) $\beta = 1$ and noise variation $\sigma = 10$. W_s is the Wiener process with normally distributed increments.

Model II. With Batch Effects

Case I. Consider a binary group classification for a scenario of unequal group sizes, where each group is contaminated with different batch effects. This reflects that, in practice, data are obtained from different sources such as different hospitals or laboratories. Again, under this case, we assume that the prior probabilities of the two groups are different. The simulation process is the same as Case I in Model I, except that each group is randomly contaminated with one of three different batch effects with equal probability. The nature of the batch effects is explained below.

Case II. Consider a binary group classification for a scenario of equal group sizes, where each group is also contaminated with different batch effects. Under this case, we assume that the prior probabilities of the two groups are equal. The simulation process is the same as Case II in Model I, except that each group is randomly contaminated with one of three different batch effects, with equal probability.

The model for the data containing batch effects (which are denoted $\alpha(t)$) is:

$$x_{ij}(t) = \mu_i(t) + \alpha(t) + \varepsilon_j(t) \tag{1.8}$$

where $\mu_i(t)$ and $\varepsilon_j(t)$ are the same as in Model I Case I. $\alpha(t)$ is chosen at random from a set of possible batch effects:

$$\alpha(t) = \begin{cases} \sin(t+U_{11})\log(t+U_{12}) & w.p.\ 1/3 \\ -U_{21}t^2 + U_{22}t & w.p.\ 1/3 \\ \phi(\frac{t-U_{31}}{0.316}) + U_{32} & w.p.\ 1/3 \end{cases} \tag{1.9}$$

where $U_{21} \sim U(0.9, 1)$, $U_{22} \sim U(0.8, 0.9)$, $U_{11} \sim U(-0.02, 0.02)$, $U_{12} \sim U(0.01, 0.02)$,$U_{31} \sim U(0.475, 0.525)$ and $U_{32} \sim U(-0.3, -0.2)$. Here $U(a, b)$ represents the continuous uniform distribution between a and b.

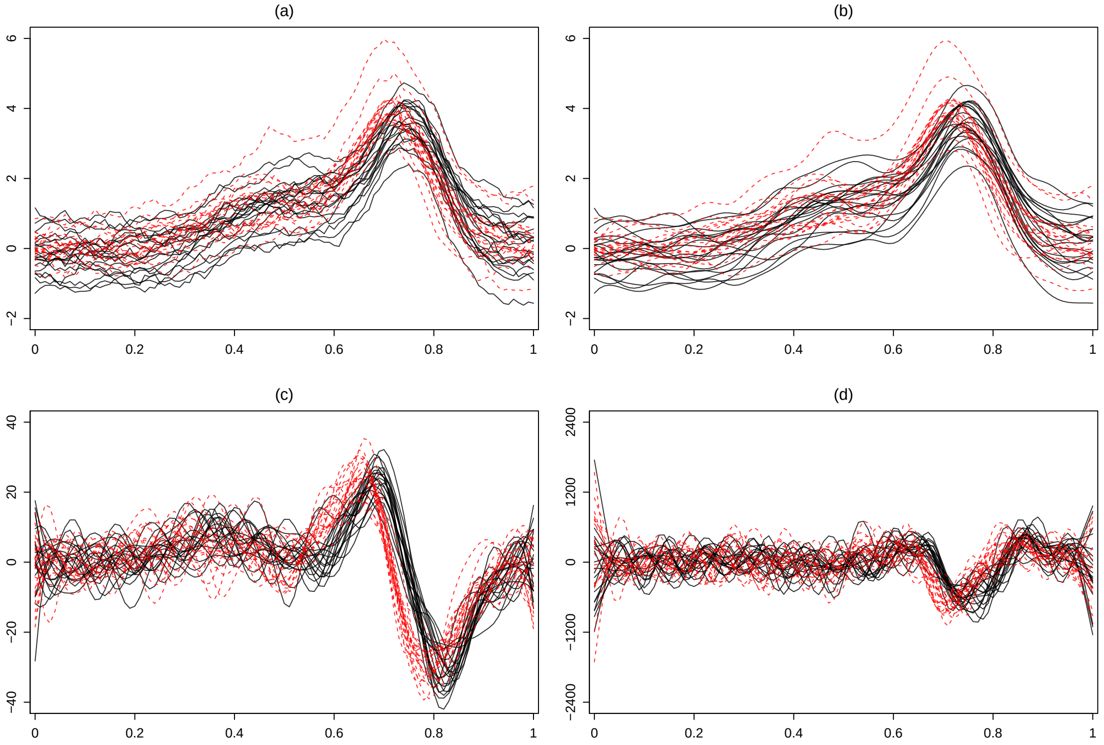


Figure 1.3: (a). 20 raw curves for two groups in Model II. (b). Smoothed curves. (c). First derivatives. (d). Second derivatives.

1.5.2 Simulation 2. Multi-Group Classification

Model III. Without Batch Effects

Case I. Consider a three-group classification where the classes have unequal group sizes. We assume that the prior probabilities of three groups are $\boldsymbol{\pi} = (0.1, 0.2, 0.7)$.

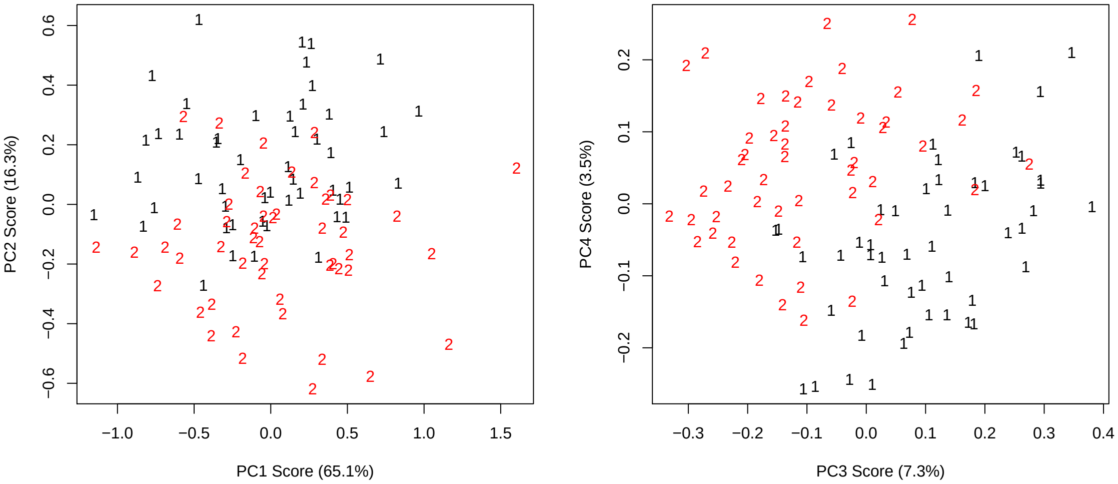


Figure 1.4: First four functional principal component scores of 20 curves for two groups in Model II. The first two scores account for about 80% of variation of the five groups of curves, where group 1 and group 2 mask each other. However, the third and four principal component scores separate the two groups, though they explain only 12.5% of the total variation.

The simulation is conducted as follows. Randomly generate 40, 80 and 280 curves from X_1, X_2 and X_3, respectively. Conduct 100 cross-validations to investigate the performance of our proposed method. In each cross-validation, randomly select 30, 60 and 210 curves from X_1, X_2 and X_3, constituting the training data, and the rest constitutes the testing data. In order to study the effect of the specification of prior probability for each group on classification performance, we propose three different specifications of prior probabilities, which are $\boldsymbol{\pi_1} = (0.1, 0.2, 0.7)$, $\boldsymbol{\pi_2} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and $\boldsymbol{\pi_3} = (0.2, 0.7, 0.1)$.

Case II. Consider a three-group classification with equal group sizes. Assume that the prior probabilities of three groups are $\boldsymbol{\pi} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. In the simulation, randomly generate 100 curves for each of three groups, that is, X_1, X_2 and X_3, respectively. Conduct 100 cross-validations to investigate the performance of our proposed method. In each cross-validation, randomly select 50 curves from X_1, X_2 and X_3

Table 1.1: The mean misclassification rate and the standard deviation (in parenthesis) (in percentage) for three different priors $\boldsymbol{\pi_1} = (0.2, 0.8)$, $\boldsymbol{\pi_2} = (0.5, 0.5)$ and $\boldsymbol{\pi_3} = (0.8, 0.2)$ using different data augmentations, where $\mathbf{x}_1 = c(x, x')$, $\mathbf{x}_2 = c(x, x', x^{(2)})$, $\mathbf{x}_3 = c(x, x', x^{(2)}, x^{(3)})$, and $\mathbf{x}_4 = c(x, x', x^{(2)}, x^{(3)}, x^{(4)})$, respectively. The true group prior probability is $\pi = (0.1, 0.2, 0.7)$. In case I, $\boldsymbol{\pi_1}$ is the correct prior guess but $\boldsymbol{\pi_2}$ and $\boldsymbol{\pi_3}$ are biased prior guesses; in case II, $\boldsymbol{\pi_2}$ is the correct prior guess but $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_3}$ are biased prior guesses. However, when using the raw and first derivative curves, the incorrect prior guess has little effect on the classification performance.

			FM depth			h-mode depth		
			π_1	π_2	π_3	π_1	π_2	π_3
Model I	Case I	$\mathbf{x}_1$	5.38(0.13)	5.03(0.12)	10.01(0.29)	**0.2**(**0.13**)	0.25(0.12)	0.52(0.29)
		$\mathbf{x}_2$	3.54(0.83)	**2.9**(**0.63**)	5.49(1.7)	1.39(0.83)	1.23(0.63)	2.55(1.7)
		$\mathbf{x}_3$	5.61(6)	3.41(4.18)	5.37(4.98)	24.15(6)	15.42(4.18)	21.97(4.98)
		$\mathbf{x}_4$	6.3(2.35)	3.89(2.86)	5.98(2.42)	46.31(2.35)	38.05(2.86)	46.45(2.42)
	Case II	$\mathbf{x}_1$	8.8(0.13)	5.89(0.1)	9.11(0.2)	0.28(0.13)	**0.26**(**0.1**)	0.41(0.2)
		$\mathbf{x}_2$	5.19(0.91)	**3.31**(**0.44**)	5.24(0.78)	1.96(0.91)	1.16(0.44)	1.86(0.78)
		$\mathbf{x}_3$	5.96(1.87)	3.88(2.13)	5.53(2.3)	41.23(1.87)	35.26(2.13)	41.64(2.3)
		$\mathbf{x}_4$	6.3(0.8)	4.05(1.78)	5.72(0.79)	49.78(0.8)	49.27(1.78)	49.89(0.79)
Model II	Case I	$\mathbf{x}_1$	4.84(0)	4.44(0)	8.74(0.67)	**0**(**0**)	0(0)	0.2(0.67)
		$\mathbf{x}_2$	2.78(0.98)	2.6(1.25)	4.9(1.96)	0.54(0.98)	0.92(1.25)	2.02(1.96)
		$\mathbf{x}_3$	3.46(3.9)	**2.58**(**5.72**)	4.22(9.01)	14.42(3.9)	13.18(5.72)	21.18(9.01)
		$\mathbf{x}_4$	4.48(2.12)	3.12(6.7)	4.58(8.11)	20.9(2.12)	31.9(6.7)	51.68(8.11)
	Case II	$\mathbf{x}_1$	14.03(0.66)	11.92(0.17)	15.58(0.38)	0.54(0.66)	**0.03**(**0.17**)	0.07(0.38)
		$\mathbf{x}_2$	8.63(2.08)	7.19(1.09)	9.91(2.76)	2.28(2.08)	0.87(1.09)	2.97(2.76)
		$\mathbf{x}_3$	8.49(6.01)	**6.35**(**5.52**)	9.7(7.47)	23.86(6.01)	19.21(5.52)	28.28(7.47)
		$\mathbf{x}_4$	9.21(3.88)	7.24(5.33)	11.12(3.22)	41.95(3.88)	39.09(5.33)	47.67(3.22)

constituting the training data and the rest constitutes the testing data. Analogous to Case I, we propose three scenarios for specifications of prior probabilities that are the same as $\boldsymbol{\pi_1}$, $\boldsymbol{\pi_2}$ and $\boldsymbol{\pi_3}$ in Case I for studying the robustness of our method to different priors. The simulated curves follow the model:

$$x_{ij}(t) = \mu_i(t) + \varepsilon_j(t), \quad i = 1, 2, 3; \; j = 1, \ldots, n_i, \tag{1.10}$$

where $\mu_1(t), \mu_2(t), \varepsilon(t)$ are the same as (1.7a), (1.7b), (1.7c) in Model I and $\mu_3(t) = 300t^6(1-t)^2$.

Model IV. With Batch Effects

Case I. Analogously to Case I in Model III, the simulation process is the same, except that the three groups of curves are randomly contaminated with three

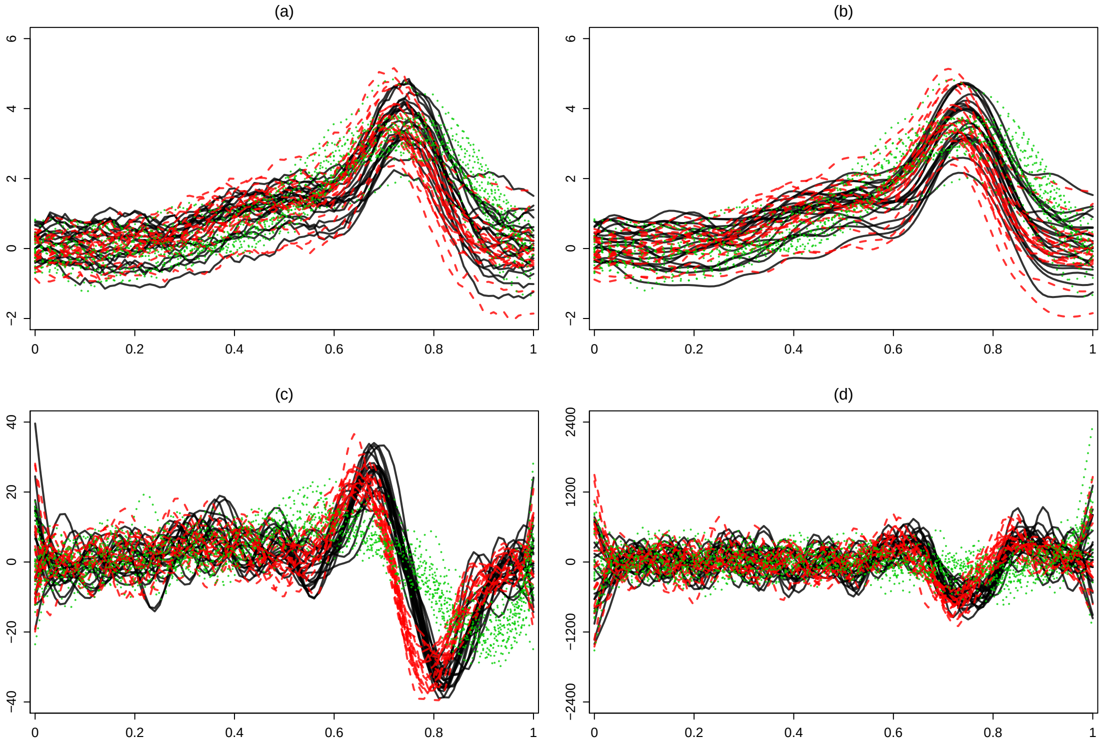


Figure 1.5: (a). 20 raw curves for three groups in Model III. (b). Smoothed curves. (c). First derivatives. (d). Second derivatives.

different batch effects α_1, α_2 and α_3, which are the same as Case I in Model II.

Case II. Analogously to Case II in Model III, the simulation process is the same, except that the three groups of curves are randomly contaminated with three different batch effects α_1, α_2 and α_3, which are the same as Case II in Model II. The simulated curves follow the model:

$$x_{ij}(t) = \mu_i(t) + \alpha_j(t) + \varepsilon_j(t), \quad i = 1, 2, 3;\ j = 1, \ldots, n_i, \tag{1.11}$$

where $\mu_1(t), \mu_2(t), \mu_3(t)$ and $\varepsilon(t)$ are the same as above Case I, and $\alpha(t)$ is (1.9) in Model II.

The enriched simulation study illustrates the performance of our method by using each of the multivariate functional Fraiman and Muniz (FM) depth and the multi-

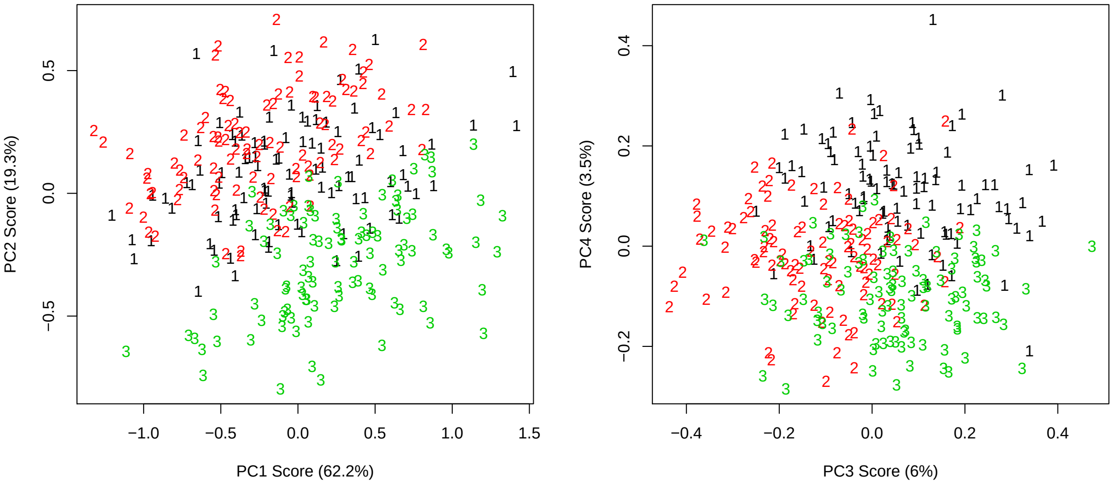


Figure 1.6: First four functional principal component scores of 20 curves for two groups in Model III. The first two scores accounts for about 81.5% of variation of the three groups of curves, where group 1 and group 2 mask somehow each other, but separate from group 3. However, the third and four principal component scores shows that group 1 and group 2 are quite distinguishable, though they explain about 9.5% of the total variation.

variate functional h-mode depth with respect to different curve vectors. It is shown that our method has competitive performance in classification when using the multivariate functional h-mode depth with respect to a bivariate functional observation vector consisting of the raw curve and its first derivative. Moreover, we compare our method (DB) to the conventional multivariate discriminant analysis that uses the functional principal component analysis to reduce the infinite-dimension functional observation to a finite-dimension vector of four principal component scores and the aforementioned pure depth-based methods. The conventional multivariate discriminant analysis methods used in our comparison includes linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), generalized linear model (GLM), support vector machine (SVM), neural network (NNET), mixture discriminant analysis and flexible discriminant analysis (FDA) (Friedman et al., 2001), respectively. The depth-based methods includes the Distance to trimmed mean (DS) and the Trimmed

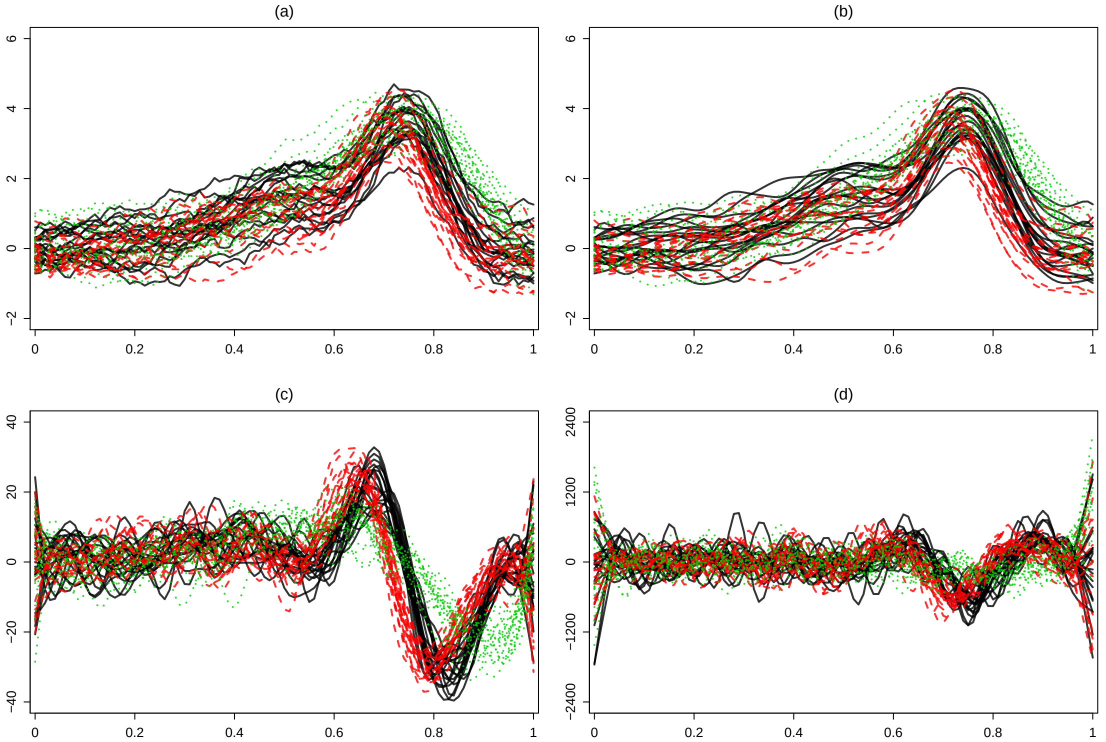


Figure 1.7: (a). 20 raw curves for three groups in Model IV. (b). Smoothed curves. (c). First derivatives. (d). Second derivatives.

Weighted Averaged Distance (TAD) based on the generalized band depth (López-Pintado and Romo, 2009). Table 1.3 displays that our method (DB) has lower misclassification than the other competitors in classification.

1.5.3 Sensitivity Analysis

We conducted a sensitivity analysis to understand how our proposed algorithm in Section 1.3 varies in terms of the classification performance based on using different smoothing parameters for preprocessing the functional data. We also investigated the effect of using various β, which quantifies the relative weight of NPV(t) to FDR(t) in the score function M_β. In the simulation subsections 1.5.1 and 1.5.2, we explored the proposed algorithm in four models, each of which involved using both equal and unequal prior probabilities to investigate the classification performance. In this sensitivity analysis study, we merely considered using the equal prior probability for

Table 1.2: The mean misclassification rate and the standard deviation (in parenthesis), in percentage, for three different priors $\pi_1 = (0.1, 0.2, 0.7)$, $\pi_2 = (1/3, 1/3, 1/3)$ and $\pi_3 = (0.2, 0.7, 0.1)$ using different data augmentations, where $\mathbf{x_1} = c(x, x')$, $\mathbf{x_2} = c(x, x', x^{(2)})$, $\mathbf{x_3} = c(x, x', x^{(2)}, x^{(3)})$, and $\mathbf{x_4} = c(x, x', x^{(2)}, x^{(3)}, x^{(4)})$, respectively. The true group prior probability is $\pi = (0.1, 0.2, 0.7)$. In case I, $\boldsymbol{\pi_1}$ is the correct prior guess but $\boldsymbol{\pi_2}$ and $\boldsymbol{\pi_3}$ are biased prior guesses; in case II, $\boldsymbol{\pi_1}$ is the correct prior guess but $\boldsymbol{\pi_2}$ and $\boldsymbol{\pi_3}$ are biased prior guesses. However, when using the raw and first derivative curves, the incorrect prior guess has little effect on the classification performance.

			FM depth			h-mode depth		
			π_1	π_2	π_3	π_1	π_2	π_3
Model III	Case I	$\mathbf{x}_1$	4(0.1)	3.94(0.15)	5.84(0.09)	**0.12**(**0.1**)	0.18(0.15)	0.12(0.09)
		$\mathbf{x}_2$	1.99(0.5)	1.89(0.48)	2.91(1.25)	1.33(0.5)	1.03(0.48)	3.72(1.25)
		$\mathbf{x}_3$	**2.11**(**2.52**)	2.12(2.58)	3.32(6.76)	12.92(2.52)	12.29(2.58)	20.64(6.76)
		$\mathbf{x}_4$	2.56(1.11)	2.56(2.68)	3.99(5.28)	28.77(1.11)	41.3(2.68)	68.77(5.28)
	Case II	$\mathbf{x}_1$	6.14(0.07)	5.08(0.08)	8.45(0.07)	**0.16**(**0.07**)	0.17(0.08)	0.18(0.07)
		$\mathbf{x}_2$	2.67(0.57)	**2.66**(**0.35**)	4.61(0.74)	1.49(0.57)	1.22(0.35)	1.63(0.74)
		$\mathbf{x}_3$	2.89(4.83)	2.83(2.79)	5.29(6.74)	20.16(4.83)	14.92(2.79)	25.98(6.74)
		$\mathbf{x}_4$	3.4(3.31)	3.26(2.91)	6.16(1.91)	59.55(3.31)	50.7(2.91)	64.15(1.91)
Model IV	Case I	$\mathbf{x}_1$	2.98(0)	2.29(0)	4.33(0.4)	**0**(**0**)	0(0)	0.2(0.4)
		$\mathbf{x}_2$	**0.81**(**1.36**)	1.01(1.29)	1.8(1.54)	2.57(1.36)	1.57(1.29)	3.05(1.54)
		$\mathbf{x}_3$	1.38(3.04)	1.25(3.8)	2.04(5.59)	15.16(3.04)	13.28(3.8)	24.32(5.59)
		$\mathbf{x}_4$	1.89(1.69)	1.74(4.59)	2.7(5.74)	29.31(1.69)	39.5(4.59)	65.95(5.74)
	Case II	$\mathbf{x}_1$	6.21(0.08)	5.54(0.09)	9.13(0.11)	**0.11**(**0.08**)	0.12(0.09)	0.14(0.11)
		$\mathbf{x}_2$	2.62(0.59)	**2.44**(**0.43**)	4.45(0.82)	1.15(0.59)	0.88(0.43)	1.55(0.82)
		$\mathbf{x}_3$	2.57(5.17)	2.6(2.96)	4.44(5.15)	22.26(5.17)	15.65(2.96)	23.83(5.15)
		$\mathbf{x}_4$	2.97(3.17)	3.04(2.66)	5(2.29)	61.48(3.17)	48.75(2.66)	62.92(2.29)

Table 1.3: Comparison of our method (DB) to the conventional multivariate discriminant analysis by using functional principal component analysis to reduce the infinite-dimension functional observation to a finite-dimension vector of four principal component scores and depth-based methods.

Scenario	Model	Group	Batch	DB	LDA	QDA	GLM	SVM	NNET	MDA	FDA	DS	TAD
Case I	Model I	2	No	0.2(0.13)	1.05(0.63)	1.12(0.61)	1.58(0.79)	1.53(0.69)	1.68(0.97)	1.11(0.62)	1.05(0.63)	12.23(7.19)	18.76(3.33)
	Model II	2	Yes	0(0)	0.42(1.15)	0.46(0.94)	0.6(1.46)	1.3(1.89)	0.6(1.62)	0.5(1.18)	0.42(1.15)	26.32(8.65)	21.98(6.3)
	Model III	3	No	0.12(0.1)	1.53(0.68)	1.73(0.74)	1.96(0.79)	3.4(0.82)	2.15(1.26)	1.59(0.72)	1.53(0.69)	18.03(3.65)	17.67(2.61)
	Model IV	3	Yes	0(0)	2.08(1.24)	2.16(1.17)	2.51(1.33)	3.67(1.66)	2.82(1.43)	1.96(1.14)	2.08(1.24)	21.55(5.77)	22.29(4.78)
Case II	Model I	2	No	0.26(0.1)	0.58(0.23)	0.66(0.26)	1.22(0.71)	1.07(0.42)	1.17(0.6)	0.62(0.29)	0.58(0.23)	12.31(6.84)	19.14(6)
	Model II	2	Yes	0.03(0.17)	0.27(0.55)	0.42(0.61)	1.37(1.23)	0.8(0.94)	1.08(0.9)	0.33(0.59)	0.27(0.55)	20.4(10.08)	25.71(8.84)
	Model III	3	No	0.17(0.08)	4.38(1.86)	4.49(1.78)	4.91(1.61)	4.86(1.69)	5.15(1.77)	4.59(1.81)	4.38(1.86)	19.43(4.3)	22.17(3.96)
	Model IV	3	Yes	0.12(0.09)	7.66(3.17)	7.8(3.3)	8.06(3.17)	8.39(3.37)	8.4(3.17)	7.79(3.27)	7.66(3.17)	21.19(4.16)	24.73(3.71)

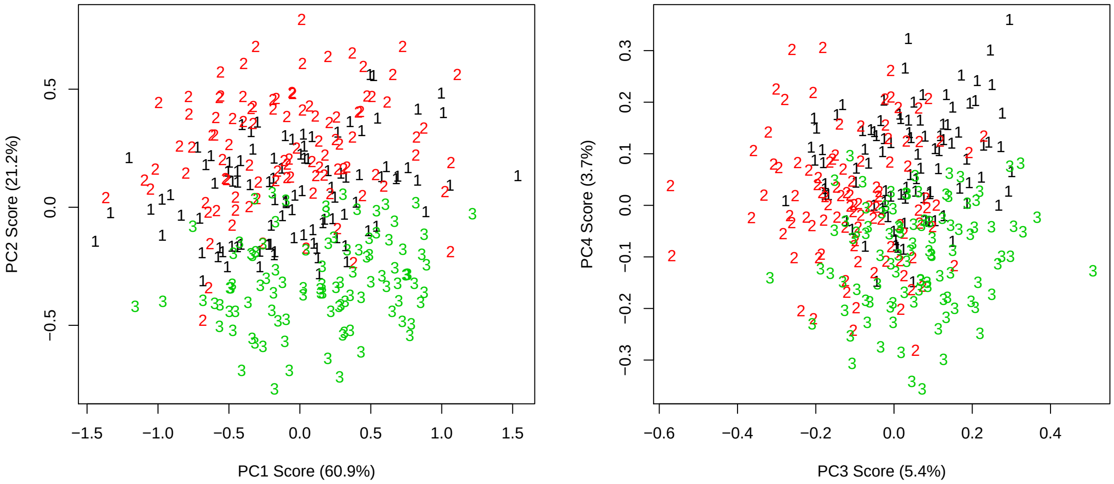


Figure 1.8: First four functional principal component scores of 20 curves for two groups in Model IV. The first two scores account for about 82.1% of the total variation in the three groups of curves, and the third and four principal component scores explain about 9.1% of the total variation. It is clear that the four principal scores can not make the groups as separable as in Model III since the batch effects add more noise within the groups.

the classification of each group in the algorithm. Model I and II illustrate the binary classification for functional data with and without batch effects, whereas Model III and IV involve multi-group classification for functional data with and without batch effects, respectively. The structure of the four models are displayed in Table 1.4.

Table 1.4: $\mu_i(t), \alpha(t), \alpha(t)$ are the same as in subsections 1.5.1 and 1.5.2.

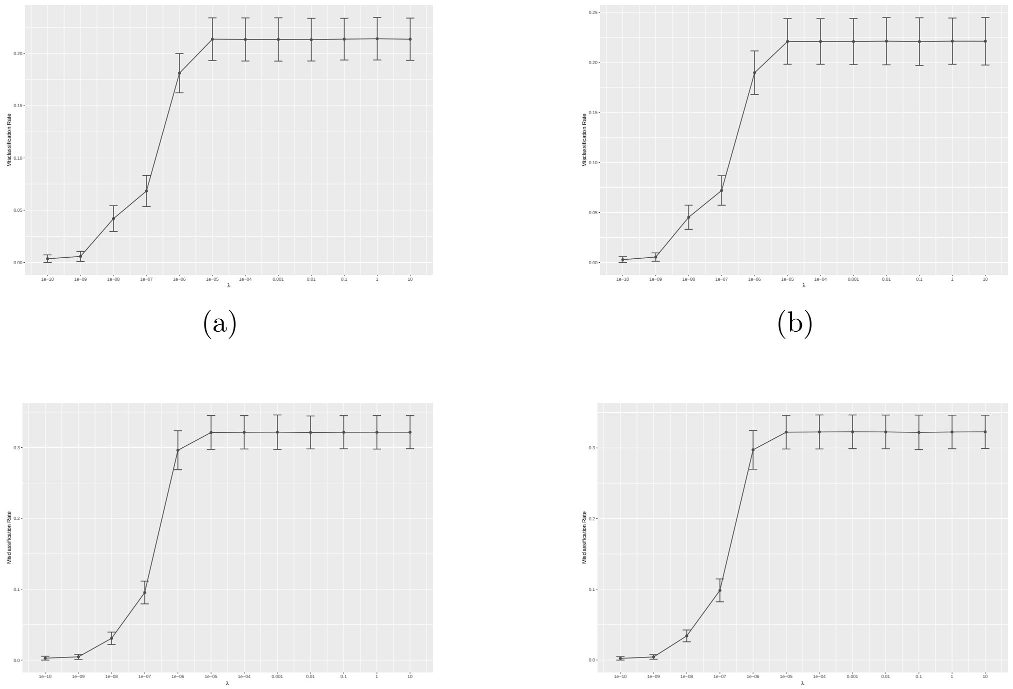
Scenario	Groups	Batch Effect	Model
Model I	Binary	No	$X_i(t) = \mu_i(t) + \varepsilon(t), i = 1, 2$
Model II	Binary	Yes	$X_i(t) = \mu_i(t) + \alpha(t) + \varepsilon(t), i = 1, 2$
Model III	Multi-Group	No	$X_i(t) = \mu_i(t) + \varepsilon(t), i = 1, 2, 3$
Model IV	Multi-Group	Yes	$X_i(t) = \mu_i(t) + \alpha(t) + \varepsilon(t), i = 1, 2, 3$

In each model, we repeated the sensitivity analysis 100 times, in which we simulated 100 functional data for each group based on the corresponding setting. For each simulated data, we conducted 100 cross-validations by using the proposed classification algorithm, where we randomly split the whole functional data into a training and testing data, both of which consisted of half of the functional data from each group. Moreover, for each cross-validation, instead of using the optimal smoothing parameter, we preprocessed all functional data using the B-spline smoothing method by varying the smoothing parameters from 10^{-10} through 10 by increasing one magnitude each time. Meanwhile, in order to explore how the different score function M_β influences the classification performance, we also changed β, the relative weight of $\mathrm{NPV}(t)$ to $\mathrm{FDR}(t)$, from 0 through 2.

The sensitivity study shows that there is no change on the classification performance when the value of β in the score function M_β varies from 0 through 2. This fact actually proves that M_β is, in a sense, a robust classification criterion in both binary and multi-group classification. Nonetheless, Figure 1.9 displays the mean misclassification rate (with one standard deviation) for the 100 repeated simulations at each smoothing parameter for each model. The mean misclassification rates become worse when oversmoothing occurs, which also causes the corresponding standard deviations to grow larger. This makes sense because the oversmoothing greatly impacts the original functional data so that it disguises the dissimilarity amongst the original groups of functional data. Therefore, the proposed classification algorithm cannot classify the groups of smoothed curves correctly. Overall, the sensitivity analysis study tells us that appropriate smoothing for the raw functional data matters on the classification performance by using our proposed algorithm. In practice, we recommend a small amount of presmoothing, to avoid the danger of oversmoothing the raw data.

1.6 Real Data Application

In this section, we apply our method to three real data sets and investigate its classification performance by comparing it to competing functional classification methods. Of the three real data cases, the first two are frequently analyzed benchmark data, in particular the famous Berkeley growth data (Ramsay and Silverman, 2005) and a set



(a)

(b)

(c)

(d)

Figure 1.9: Figure (a), (b), (c), (d) represent the mean misclassification rates with one standard deviation at various smoothing parameters from 10^{-10} through 10 for Model I, II, III, and IV, respectively.

of phoneme spectral data (Ferraty and Vieu, 2006). The third one is data from 12 groups of textile fibers in forensic casework described by Fuenffinger (2015). We compare our method (DB) to the competing classification methods proposed in Section 4, including linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) under the application of functional principal component analysis (PCA) and the Trimmed Weighted Averaged Distance (TAD) and Distance to the Trimmed Mean (DS) under the application of modified band depth. In order to obtain the best performance in classification for applying our method, we employ the multivariate functional h-mode depth by using a bivariate functional observation composed of the raw curve and the corresponding first derivative. For simplicity, we merely use the LDA and QDA in functional PCA approach, since other discriminant analysis methods like NNET and SVM perform similarly to them typically. Moreover, we

Table 1.5: The mean misclassification rate and standard deviation (in parenthesis) at various smoothing parameters λ from 10^{-10} through 10 for four models. Note that the values in the table are in percent.

	λ											
	1e-10	1e-09	1e-08	1e-07	1e-06	1e-05	1e-04	0.001	0.01	0.1	1	10
Model I	**0.35** (**0.37**)	0.58 (0.49)	4.19 (1.24)	6.84 (1.48)	18.11 (1.88)	21.36 (2.04)	21.33 (2.06)	21.33 (2.07)	21.31 (2.04)	21.36 (2)	21.4 (2.03)	21.35 (2.02)
Model II	**0.28** (**0.3**)	0.54 (0.42)	4.52 (1.21)	7.2 (1.47)	18.98 (2.18)	22.11 (2.28)	22.1 (2.28)	22.09 (2.3)	22.13 (2.36)	22.09 (2.39)	22.13 (2.31)	22.12 (2.38)
Model III	**0.28** (**0.27**)	0.46 (0.35)	3.08 (0.87)	9.54 (1.6)	29.62 (2.75)	32.14 (2.39)	32.16 (2.36)	32.18 (2.43)	32.14 (2.32)	32.17 (2.34)	32.17 (2.38)	32.17 (2.33)
Model IV	**0.23** (**0.24**)	0.43 (0.33)	3.41 (0.83)	9.85 (1.61)	29.74 (2.75)	32.23 (2.38)	32.26 (2.4)	32.28 (2.38)	32.26 (2.39)	32.2 (2.44)	32.25 (2.38)	32.27 (2.35)

also compare our method to a pure nonparametric (NP) method merely using the multivariate functional h-mode depth, which classifies a curve to the group in which it has the largest depth.

1.6.1 Berkeley Growth Data

The Berkeley growth data originally collected by Tuddenham and Snyder (1954) consists of the heights of 39 boys and 54 girls from age 1 to 18, measured intermittently. Of interest with these data is classifying a child's gender using that child's growth curve (i.e., the function of height over time). Based on our simulation results, a pre-processing smoothing technique can yield better results in terms of classification performance. So we first smooth all the curves by applying a B-spline smoother with the optimal penalty parameter chosen by generalized cross-validation, by regularizing the second derivatives. Then, we randomly select 20 centered smoothed curves for each group, i.e., boys and girls, and apply our method and its competitors to the rest of the curves. We calculate the misclassification rates of each classification method to judge their accuracy. We repeat this process 100 times, and Table 1.7 shows that our method performs competitively for this data set.

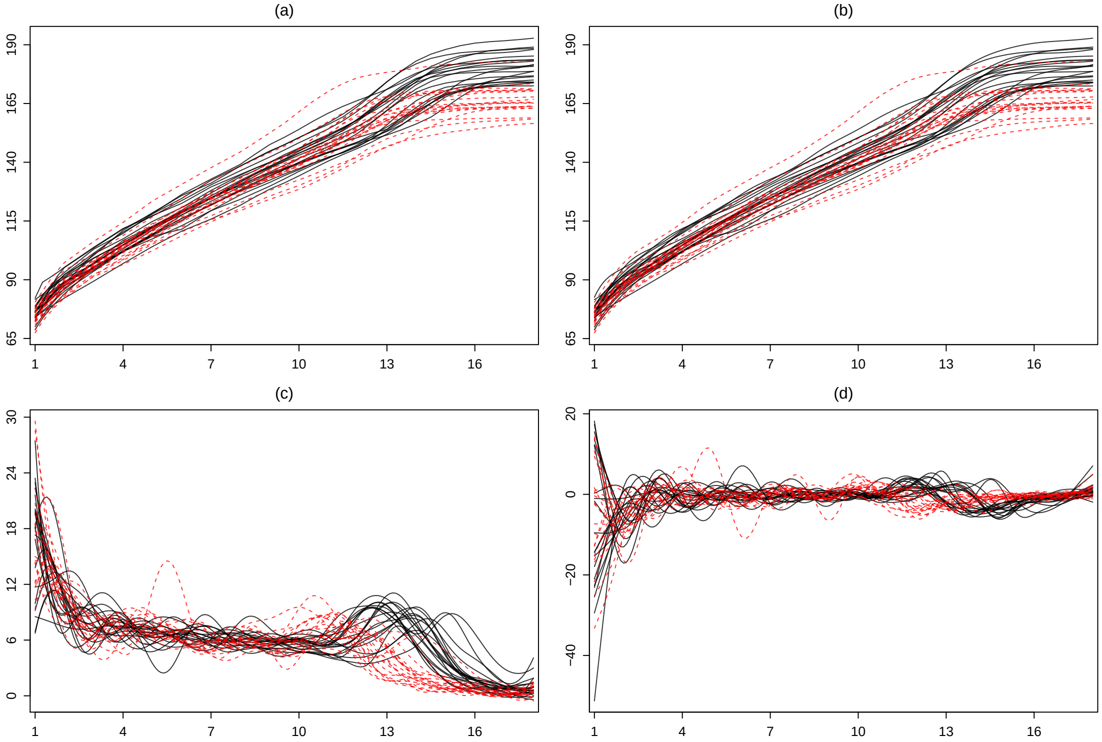


Figure 1.10: (a). 20 raw curves of growth height for girls and boys. (b). Smoothed curves. (c). First derivatives. (d). Second derivatives.

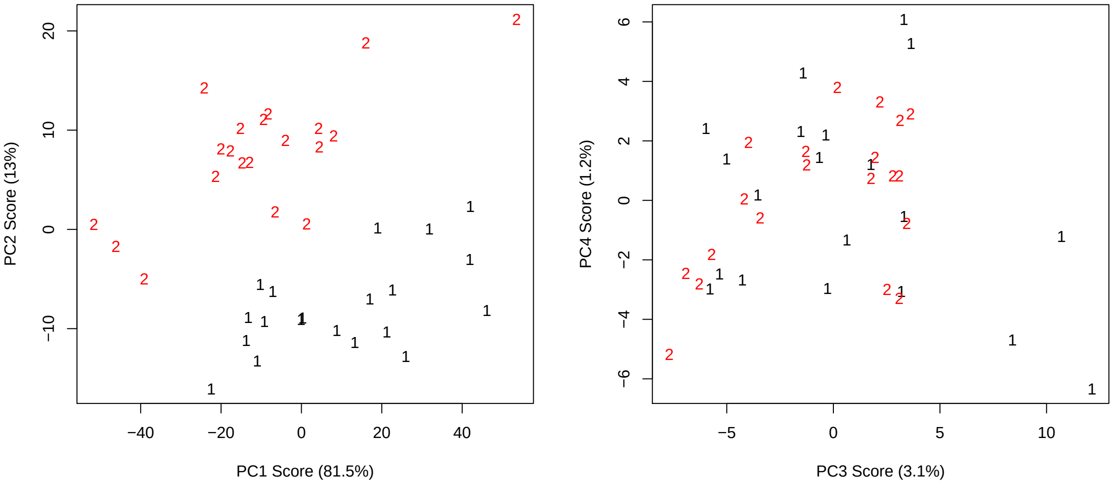


Figure 1.11: First four functional principal component scores of 20 curves of growth height for girls and boys. The first two scores account for about 94.5% of the variation of the five groups of curves, where group 1 and group 2 are clearly separate from each other.

1.6.2 Phoneme Spectral Data

The phoneme data consist of five groups of phoneme curves in log-periodogram at 150 frequency measurements (Ferraty and Vieu, 2006). Each group consists of 400 observed curves. The five phonemes are “aa”, “ao”, “iy”, “sh” and “dcl”, respectively. Similar to the pre-processing for the Berkeley growth data, we conduct B-spline smoothing on the phoneme curves by regularizing the second derivatives such that the first and second derivatives are smoothed. We randomly select 100 smoothed curves for each phoneme as the training set and use the rest as the testing set. We apply our method and the competitors on the testing data. We repeat the process 100 times and get the misclassification rate for each method. The result shows that our method is very competitive.

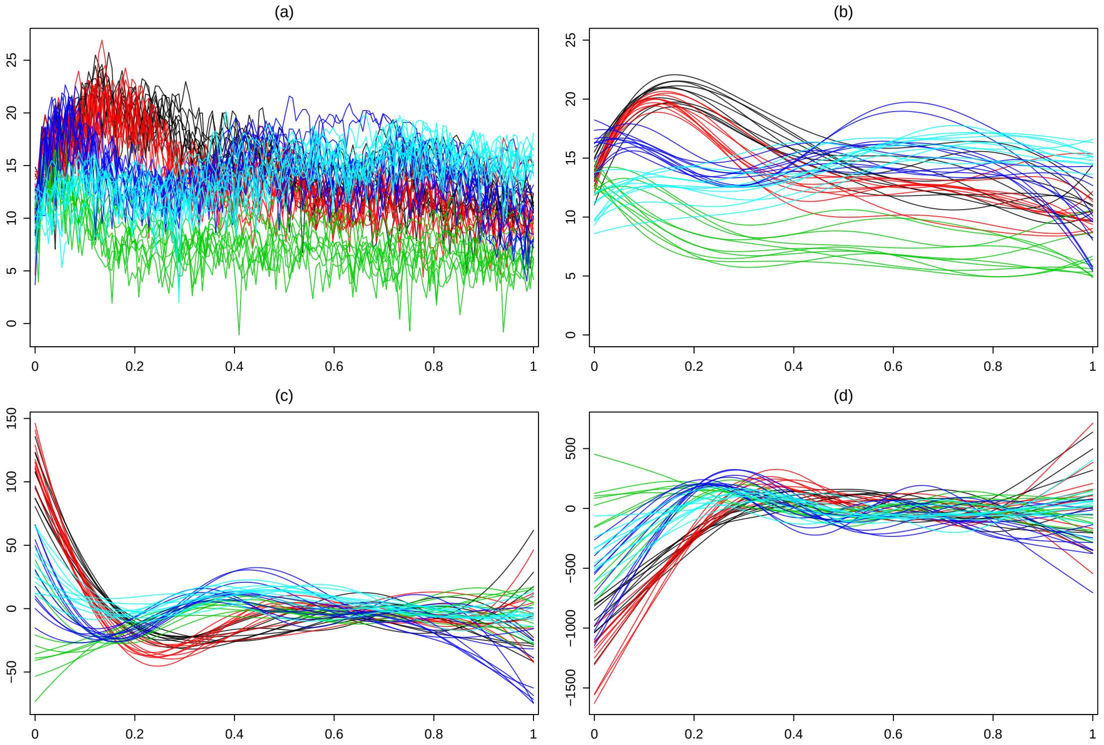


Figure 1.12: (a). 10 raw curves for each of phonemes “aa”, “ao”, “iy”, “sh” and “dcl”. (b). Smoothed curves. (c). First derivatives of the 50 curves. (d). Second derivatives of the 50 curves.

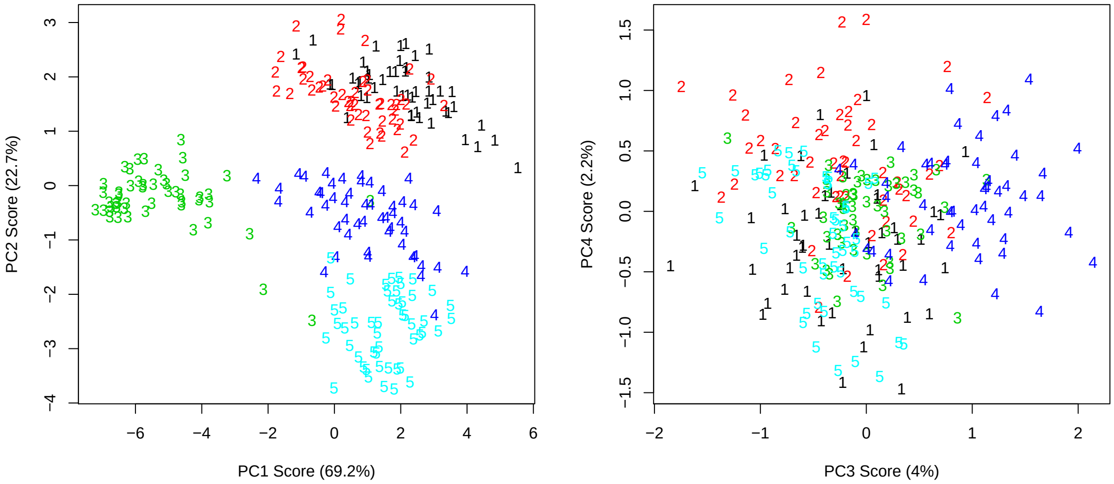


Figure 1.13: First four functional principal component scores for 50 curves of each five phonemes “aa”, “ao”, “iy”, “sh” and “dcl”. The first two scores account for about 92% variation of the five groups of curves, while group 1 and group 2 mask each other. Instead, though the third and fourth scores explain merely 6%, it might provide additional information to discriminate group 1 and group 2 somehow.

1.6.3 Forensic Data

The forensic data consist of 12 blue acrylic fibers, represented via UV-visible absorbance spectra (by Ultraviolet-visible microspectrophotometry) in a forensic study. The data are provided by Fuenffinger (2015) and Morgan (2014). The UV-visible absorbance spectra of 12 blue acrylic fiber types were examined 10 times each at 1175 spectral measurement points in the region 400-800 nm at five separate locations (including three academic research laboratories and two forensic laboratories). Our interest is to classify the fiber type given a new UV-visible absorbance curve. Because these UV-visible absorbance curves were measured in different locations, there are batch effects caused by the location. Analogously, we conduct the B-spline smoothing on these fiber absorbance curves by regularizing the second derivatives. Then, the smoothed curves are used to compare the classification performance between our method and the aforementioned competitors. We randomly select 30 curves from each group as the training data, and the rest become the testing data. We apply our method and the competitors to the training and testing data and we obtain the misclassification rate on the test data for each method. We repeat this process 100 times and the misclassification rate for each method shown in Table 1.7 represents that our method gives the best result.

Table 1.6: The confusion matrix of the predicted classes and true classes. The rows represent the true classes and the columns for the predicted classes. In each row, the cell values are the mean classification rate in percent in each possible class. It shows that our method has inferior performance on classifying the fiber 086 and fiber 112, which reflects the fact that these two groups of curves are highly similar each other.

	Predicted Class											
True Class	F086	F087	F088	F091	F092	F095	F098	F099	F112	F113	F114	F145
F086	**75.75**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	*24.25*	0.00	0.00	0.00
F087	0.00	**98.70**	1.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
F088	0.00	7.45	**92.50**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00
F091	0.05	0.30	0.05	**92.30**	0.05	0.00	0.00	2.50	0.00	4.55	0.20	0.00
F092	0.10	0.50	0.00	0.60	**97.60**	0.30	0.10	0.60	0.00	0.00	0.20	0.00
F095	0.05	0.15	0.15	4.55	0.00	**93.90**	0.05	0.95	0.00	0.05	0.15	0.00
F098	0.95	0.00	0.00	0.00	0.00	0.00	**98.60**	0.00	0.45	0.00	0.00	0.00
F099	0.00	0.60	0.00	0.00	0.00	0.00	0.05	**99.25**	0.00	0.00	0.10	0.00
F112	*37.80*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**62.20**	0.00	0.00	0.00
F113	0.00	0.65	0.25	0.60	0.00	0.00	0.00	0.00	0.00	**97.90**	0.60	0.00
F114	0.00	0.05	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**99.55**	0.00
F145	1.75	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.90	0.00	0.00	**95.35**

Table 1.7: Misclassification mean rate and standard deviation (in parenthesis) in percent for the three real data, that is, the Berkeley Growth Data (Growth), the Phoneme Data (Phoneme) and the Forensic Fiber Data (Forensic). Our method is quite competitive among the selected classification methods, and performs the best for the forensic fiber data.

	LDA	QDA	DS	TAD	DB	NP
Growth	**4.58(2.26)**	32.43(9.5)	16.74(8.4)	18.81(7.88)	*6.62(4.22)*	6.81(3.73)
Phoneme	**9.19(1.39)**	*10.42(1.57)*	14.04(1.53)	14.7(1.57)	10.88(1.8)	12.84(2.09)
Forensic	25.3(2.61)	14.68(2.1)	39.98(2.52)	40.89(3.19)	**8.03(1.78)**	*8.74(1.82)*

1.7 Conclusion

We have proposed a novel classification method for functional data by integrating statistical functional depth concepts and false discovery rate theory. In order to best

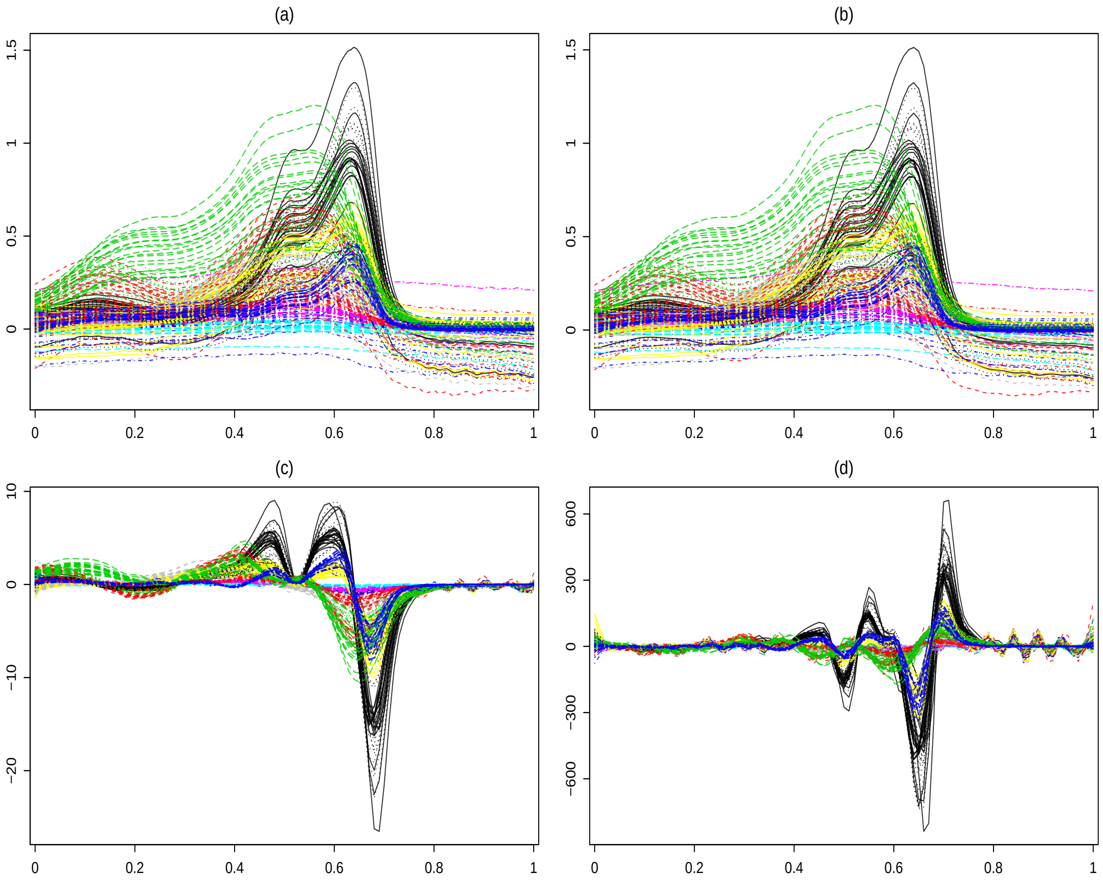


Figure 1.14: Forensic data: 12 blue acrylic fiber absorbance spectra plot. (a): Raw curves (b): Smoothed curves (c): First derivative (d): Second derivative

capture the shape, amplitude and phase variation between groups of curves, we augment the observed functional data by taking derivatives up to an appropriate order under some smoothing constraints. Based on the augmented multivariate functional data and a proper multivariate depth function, we propose a novel measure M_1 score to assess the posterior probability of a curve belonging to each group. Compared with some conventional classification methods on the multivariate data formed by the principal component scores, our method based on multivariate functional depth performs better in classification. The optimal dimension of our augmented multivariate functional observations depends on the choice of multivariate functional depth, and our simulation study shows that the multivariate functional halfspace depth is

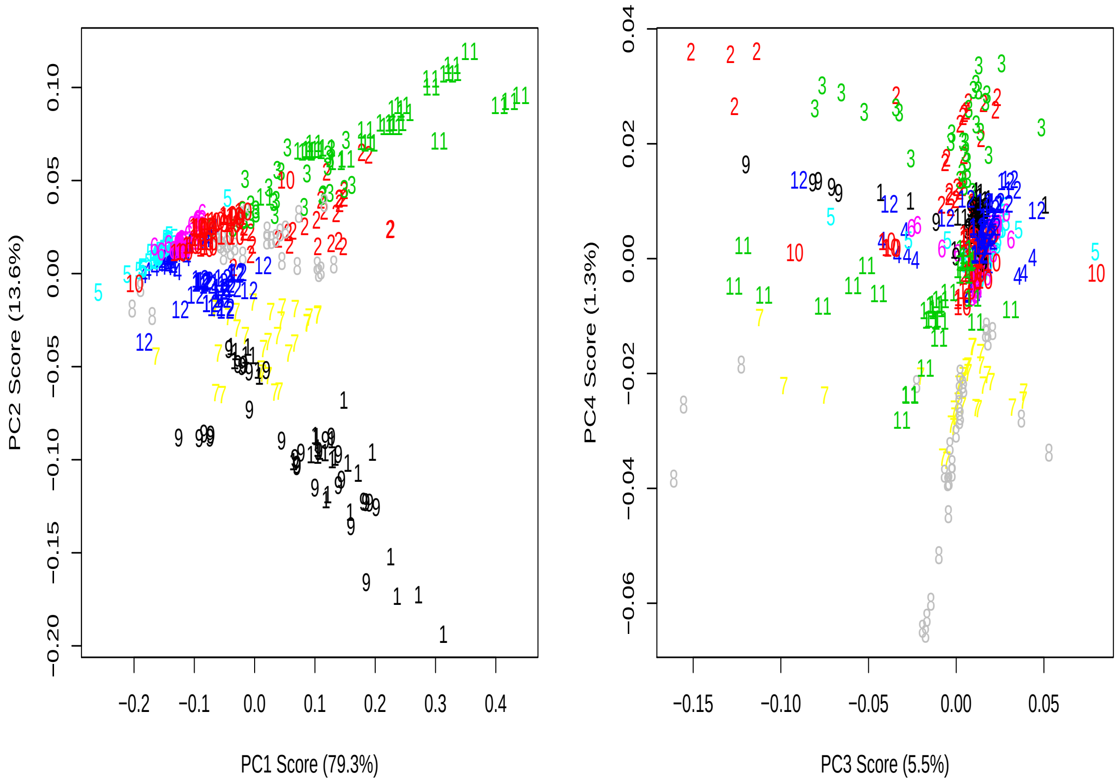


Figure 1.15: First four functional principal component scores for 20 curves of each of the 12 fiber groups. The first two scores account for about 92.9% of the variation of the 12 groups of curves, and the third and fourth scores explain merely 6.8%.

more robust to higher-order derivatives of functional observations than multivariate functional h-mode depth. Moreover, when using the multivariate functional halfspace depth, the optimal dimension of the multivariate functional observations is obtained by taking derivatives of curves up to the third order, while it is best to merely take the first derivative of curves when using multivariate functional h-mode depth. However, the best performance in classification using our method is based on the multivariate functional h-mode depth, by forming two-dimensional functional observations which include the raw curve and its corresponding first derivative.

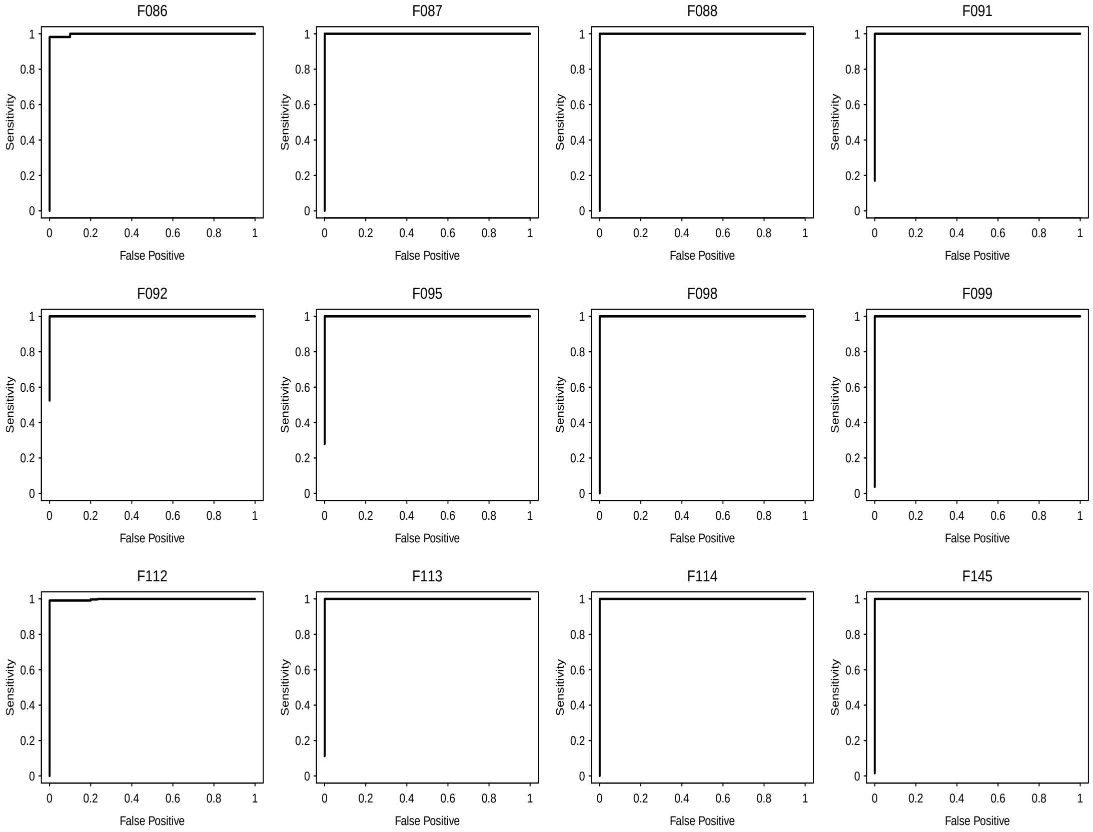


Figure 1.16: Power function for each of the 12 hypothesis tests.

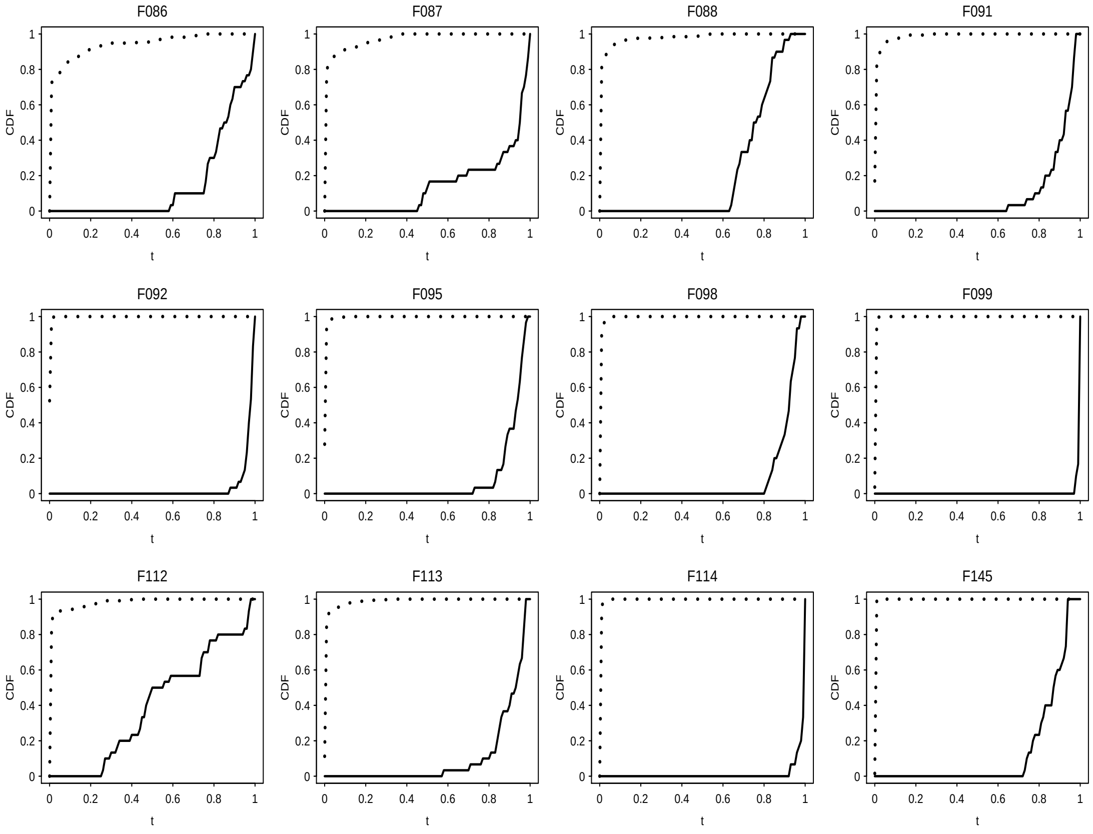


Figure 1.17: CDF's of depth ratio $T_{X_k}(Z)$ under H_0^k and H_a^k, $k = 1, 2, \ldots, 12$ respectively. The solid CDF is under the alternative hypothesis and the dotted CDF is under the null hypothesis for each k. The CDF's are estimated using the empirical CDF rather than using the mixture of logit-normal distributions.

Chapter 2

A Modified Mixture Model Approach to the Large Scale Multiple Testing Problem[1]

2.1 Introduction

In many inference problems, a large scale hypotheses are considered simultaneously. In such situations, traditional multiple testing methods can lead to numerous false discoveries and false non-discoveries that are not confirmed in later experiments. In this paper we discuss an approach that is useful in a simultaneous multiple hypothesis testing situation where the goal is to find the real "discoveries" or "significant" cases. Based on this approach, we also present a way to explore the inter-relationship between the hypotheses via a visual network pattern construction. The methodology developed here is based on a simple two-point mixture contamination model where one component corresponds to the baseline (background) information and the second to the sources which are the real discoveries (the contamination). The basic model for the density of the population under study is assumed to be:

$$f(x) = p_0 f_0(x) + p_1 f_1(x), \tag{2.1}$$

with $p_0 + p_1 = 1$. Here f_0 is the background density and f_1 is the contamination density or the density of the signal that one wants to find, while p_0 and p_1, respectively, denote the proportions of baseline and significant cases in the study.

[1] Paramita Chakraborty, Chong Ma, John Grego, James Lynch. Submitted to Statistics in Medicine

A notable work related to the model in (2.1) for large scale inference is the methodology described in Efron (2007, 2010). Efron's approach is to use the empirical Bayes (a mixture distribution where the mixing parameter is a latent variable) based on Robbins (1956), and Efron (2010, p. 14) notes that he and Morris originally "hijacked" Robbins' terminology for James-Stein estimation purposes (Efron and Morris, 1973). But his work (Efron, 2007, 2008, 2010) is along the lines of Robbins' original ideas in estimating the mixing distribution (the "empirical prior"). Efron also expends a good bit of effort in determining f_0 which he refers to as the "empirical null". Murlidharan (2010) subsumes this effort as well as that of choosing f_1 in a mixture model empirical Bayes method that is a specialization of Efron's Brown-Stein model. His method is based on mixing over an exponential family where f_0 and f_1 are submixtures of this mixture model.

We follow a similar setup and use the associated posterior probabilities for inference purposes. Our approach is to fit the mixture contamination model in (2.1) to the p-values or the left-tail areas (LTA's) from the test statistics associated with the array of hypotheses under study and use the fitted distributions with a tail adjustment for estimating the background and the contamination densities. This adjusted empirical fit can be used to approximate continuous or discrete data.

Based on Equation (2.1), the assignment functions are $\mathrm{A}_0(x) = p_0 f_0(x)/f(x)$ and $\mathrm{A}_1(x) = p_1 f_1(x)/f(x)$. These are empirical posterior probability densities of the background and the contamination classes for a given observation. The assignment function $\mathrm{A}_1(x)$ can be used to investigate the chance of an observation coming from the contamination class f_1, i.e. the observation is actually a significant case. The complementary assignment function $\mathrm{A}_0(x)$ is related to what is popularly considered as the local false discovery rate (local fdr or fdr) (Efron, 2010). In recent years, the false discovery rate has been presented as an effective tool to handle large scale multiple testing problems (Benjamini and Hochberg, 1995, Storey, 2002). If we label the background distribution f_0 as the null model and the signal/contamination distribution f_1 as the nonnull model, then it is easy to see that the assignment functions $\mathrm{A}_0(x)$ and $\mathrm{A}_1(x)$ are essentially the posterior probabilities $P(\text{null}|x)$ and $P(\text{nonnull}|x)$, respectively. In symbols, we refer to the null as H_0 and the non-null as H_1.

The local fdr can be used to calculate the tail-area false discovery rate $\mathrm{Fdr}(x) = P(\text{null}||X| > |x|)$ (assuming symmetric f for two-sided tests). The tail-area Fdr is a useful tool to screen for potential significant cases. Specifically, observations with small Fdr can be viewed as less likely to be a false discovery and thus can be considered as a significant case or a true discovery. Refer to Efron (2010) for detailed discussion of local and tail-area false discovery rates and their relationships with the FDR proposed by Benjamini and Hochberg (1995).

In many multiple testing situations, the entire data set is first used to fit a model, which is in turn used to detect significant cases using the entire data again. By doing this way, it could lead to over-fitting which may distort the real picture. In this article, we propose a modified approach that uses a mixture model for using the local fdr or the tail-area Fdr screening technique, which effectively deals with the over-fitting issue through the subsampling technique. The proposed methodology starts with randomly splitting the available data into two halves, where one part is used for model building and the other part is for anomaly detection. In addition, repeated sample splitting and resulting detection frequencies provide an informative look into the inter-relationship between the significant cases. We also present a power analysis to examine of the efficiency of the proposed method.

<u>Proposed Subsample-Splitting Analysis Methodology:</u>

(i) We first randomly split the subjects under study into two (equal) parts, *viz.* the training set and the verification set. The p-values or the LTA's from the test statistic associated with each of the hypotheses under study derived from the training set are named the training data and similar values derived from the verification set are named the verification data.

(ii) Then, a mixture contamination model $\hat{f}(x) = \hat{p}_0\hat{f}_0(x)+\hat{p}_1\hat{f}_1(x)$ is fitted by using only the training data, which is alternately adjusted to capture the baseline and the signal (empirical fit) appropriately.

(iii) The fitted model is used to derive the tail-area Fdr or the local fdr, based on the verification data. Given a predetermined cutoff value q, the cases with the tail-area Fdr (or the local fdr) less than q are identified as significant cases.

(iv) Repeat the stages (i), (ii), (iii) many times with different random splits of the training and the verification subsets. For each split/repetition, a set of significant cases is identified. The most frequently identified significant cases are considered as "potential discoveries".

(v) The screened cases detected together and their detection frequencies can be used to study the inter-relationships/dependencies between the significant cases. This frequency distribution is used to develop a network structure for the hypotheses that graphically describes these insights.

The subsampling approach not only circumvents the over-fitting in the mixture model, but also balances out other latent sources of variation in the data. The power and error probabilities associated with the union of rejection regions from all splits are calculated using the fitted mixture model and provide some objective understanding of the efficiency of this method. In addition, repeated sample splitting can be used to produce visualization tools such as frequency networks and parallel coordinate graphs, that provide useful summation of the data and are easy to understand. The screening of cases with high detection frequency can also be justified from the stability selection point of view (Meinshausen and Bühlmann, 2010).

The format of the paper is as follows: the theoretical background required for the proposed methodology and the associated power and precision probability calculation ideas are discussed in Section 2.2. The methodology is illustrated in Section 2.3 using a microarray data and a RNA-Sequencing data. A simulation study is also included in the same section along with related power analysis. Some discussion and concluding remarks are given in Section 2.4.

2.2 Identification of Significant Cases and Power Calculations

In this section, we discuss the theoretical model formulation and present the screening and power analysis tools. The random variables $X_1, \ldots, X_n$ under study are assumed to be i.i.d. with density (2.1). In this discussion X_i's can be the p-values or the LTA's from the test statistics. As noted earlier in Section 2.1, the local false discovery rate (fdr) (Efron, 2007) is essentially the same as the assignment function $\mathrm{A}_0(x)$. From the identity (2.15) (Efron, 2010), the relationship between the local fdr and the tail-area Fdr(B) for a given *tail-area* B is $\mathrm{Fdr}(\mathrm{B}) = E(\mathrm{fdr}(X)|X \in \mathrm{B})$.

2.2.1 Empirical Fit

Using the observed X_i's first fit a mixture of the Uniform distribution in [0,1] f_0^* and the Beta distribution f_1^* for the population density

$$\hat{f}(x) = p_0^* \cdot f_0^*(x) + p_1^* \cdot f_1^*(x)$$

Since our main interest is the identification of the most extreme cases, we adjust the signal (contamination) part as follows. Let $f_1^* = f_{01}^* + f_{11}^*$, where

$$f_{01}^*(x) = f_1^*(x) \cdot I\{f_1^*(x) < 1\} + 1 \cdot I\{f_1^*(x) > 1\} \tag{2.2}$$

$$f_{11}^*(x) = 0 \cdot I\{f_1^*(x) < 1\} + [f_1^*(x) - 1] \cdot I\{f_1^*(x) > 1\} \tag{2.3}$$

Therefore, f_{11}^* captures the more extreme part of the signal. Since f_{11}^* is not a density, it needs to be normalized as follows. Let $\int_{\mathbb{R}} f_{11}^*(x)dx = \mathrm{A}_{11}$ and define

$$\hat{f}_1(x) = \frac{1}{\mathrm{A}_{11}} f_{11}^*(x)$$

and $\hat{p}_1 = p_1^* \cdot \mathrm{A}_{11}$ and $\hat{p}_0 = 1 - \hat{p}_1$. Now let

$$\hat{f}_0(x) = \frac{p_0^*}{p_0} \cdot f_0^*(x) + \frac{p_1^*}{p_0} \cdot f_{01}^*(x).$$

Then the fitted model can be re-written as

$$\hat{f}(x) = \hat{p}_0 \hat{f}_0(x) + \hat{p}_1 \hat{f}_1(x). \tag{2.4}$$

Note that the fitted mixture model is unchanged; the terms have been rearranged so that $\hat{f}_0$ captures more of the middle part of the data while $\hat{f}_1$ captures the tail part. The rearrangement given in (2.4) is what we will refer to as the "empirical mixture model" and is related to Efron's "empirical null"; this representation of the fitted mixture model better captures the baseline and the signal distribution than the original Uniform/Beta mixture representation. To derive the estimates for expressions presented in next two subsections one has to just replace the subsequent terms in the assumed population density $f(x) = p_0 f_0(x) + p_1 f_1(x)$ by Equation (2.4). The cutoff point for the tail-adjustment in the equation (2.2) does not have to be equal to 1. Based on expected proportion of the signal present in a study, the cutoff point $c \in \mathbb{R}$ can be chosen subjectively. The assumption that the null data follows a Uniform distribution may not be a practical one specially in discrete cases (Murlidharan, 2010). But the tail adjustment part can compensate, at least in part, for the deviation from Uniform in the final adjusted form $\hat{f}_0$. One can also start with a Beta/Beta mixture at the first step if the null data is expected to deviate too much from the Uniform distribution and the rest of the adjustment and analysis steps will be exactly the same.

2.2.2 Screening Significant Cases Based on FDR and Sample Splitting

In case the p-values are used for the analysis, X_i's close to 0 are associated with the signal (contamination). If X_i's are the LTA's from the test statistic then X_i's close to 0 or 1 (or both) are associated with the signal/contamination depending on whether we are in a left-sided test, right sided-test (or two-sided test) situation. The advantage of using LTA's for the analysis is that the directions of deviation of the screened cases from the null can be easily identified. Thus the tail area B (left, right or two-sided) used for Fdr calculation will depend on the definition of X_i's and the direction of the hypotheses under study.

Let F be the cumulative distribution function corresponding to f. Let $F(\mathrm{B}) = \int_{\mathrm{B}} f(x)dx$, for any Borel set B. Similarly, write F_0 as the distribution function of the baseline distribution f_0. With this notation we derive the tail-area Fdr associated with a given tail-area B as follows:

$$\begin{aligned}\mathrm{Fdr}(\mathrm{B}) &= E(\mathrm{fdr}(X)|X \in \mathrm{B}) = \int_{\mathrm{B}} \frac{\mathrm{fdr}(x)dF(x)}{F(\mathrm{B})} \\ &= \frac{1}{F(\mathrm{B})} \int_{\mathrm{B}} \frac{p_0 f_0(x)}{f(x)} f(x)dx = \frac{p_0 F_0(\mathrm{B})}{F(\mathrm{B})}. \qquad (2.5)\end{aligned}$$

In practice, one can consider the tail-area $\mathrm{B}(x)$ associated with any value x under the support $\mathbb{S}$ of f. The tail-area false discovery rate $\mathrm{Fdr}\left(\mathrm{B}(x)\right)$ associated with x can be calculated using the equation (2.5).

If X_i's are p-values derived for each hypothesis under study, for any $x \in \mathbb{S}$, the appropriate tail area to use is $\mathrm{B}(x) = \{y \in \mathbb{S} : y < x\}$. When X_i's are LTA's from test statistics, one should use $\mathrm{B}(x) = \{y \in \mathbb{S} : y < x\}$ for a left sided test, $\mathrm{B}(x) = \{y \in \mathbb{S} : y > x\}$ for a right sided test. Whereas, in a two sided test situation with f symmetric around zero, the tail area simply is $\mathrm{B}(x) = \{y \in \mathbb{S} : |y| > |x|\}$. In general, for f which is not symmetric, the two-sided tail area can be derived using matching percentiles. Express any x as the p^{th} percentile of f, i.e., $\int_{-\infty}^{x} f(u)du = p$; then a complement x^* can be found such that $\int_{-\infty}^{x^*} f(u)du = 1 - p$. Now if $p < 0.5$ (x is smaller than the median) we choose

$$\mathrm{B}(x) = \{y \in \mathbb{S} : y < x\} \cup \{y \in \mathbb{S} : y > x^*\}. \qquad (2.6a)$$

On the other hand if $p > 0.5$ (x is larger than the median) we choose

$$\mathrm{B}(x) = \{y \in \mathbb{S} : y < x^*\} \cup \{y \in \mathbb{S} : y > x\}. \qquad (2.6b)$$

An observation x with $\mathrm{Fdr}\left(\mathrm{B}(x)\right)$ smaller than the predetermined critical value should be identified as significant. For the application part with tail-area Fdr screening, a training split is first used to fit the adjusted mixture model (2.4). Then for each data point x_i in the corresponding verification split, the appropriate tail area $\hat{\mathrm{B}}(x_i)$ is determined using the fitted model. Next $\hat{f}_0$, $\hat{p}_0$ and $\hat{f}$ from the training fit are used with the model (2.5) to derive the estimated observed tail-area Fdr, *viz.* $\widehat{\mathrm{Fdr}}(\hat{\mathrm{B}}(x_i))$. Any case with $\widehat{\mathrm{Fdr}}(\hat{\mathrm{B}}(x_i)) < q$ is screened as a potential discovery, where q is a pre-determined cutoff point.

For the screening with local fdr after fitting the adjusted mixture model with the training split, the fitted densities are used to calculate estimated local fdr $\widehat{\mathrm{fdr}}(x_i) = \frac{\hat{p}_0 \hat{f}_0(x_i)}{\hat{f}(x_i)}$ for each verification split data point. Cases with $\widehat{\mathrm{fdr}}(x_i)$ less than q (predetermined) are considered to be potential discoveries. In terms of the rejection region, for any fixed cutoff point q, depending on the tail-area Fdr or the local fdr screening, the theoretical rejection set from the k^{th} split is given by

$$R_k(q) := \{x \in \mathbb{S}_k : \widehat{\mathrm{Fdr}}\left(\mathrm{B}(x)\right) < q\} \tag{2.7a}$$

or

$$\tilde{R}_k(q) := \{x \in \mathbb{S}_k : \widehat{\mathrm{fdr}}\left(\mathrm{B}(x)\right) < q\}, \tag{2.7b}$$

where $\mathbb{S}_k$ is the support of $\hat{f}$ from the k^{th} sample split.

The above calculation is repeated a number of times. The potential significant cases can be identified from the combined rejection set $\bigcup_k R_k(q)$ or $\bigcup_k \tilde{R}_k(q)$. But to increase the precision, only the observations that have been detected repeatedly with high frequency across the $R_k(q)$'s or $\tilde{R}_k(q)$'s should be considered as potential true discoveries. The critical frequency of detection for an observation at screening can be set subjectively depending on what percentage of overall discoveries are expected for a given study.

2.2.3 Power and Error Probabilities Calculation

Efron (2007, 2010) uses a whole data fit on the z-values transformed from the p-values associated with the hypotheses under study and advocates the use of the local fdr for the screening of significant cases. In that analysis, for a given cutoff point q of the local fdr, the rejection region effectively is $\tilde{R}(q) = \{x \in \mathbb{R} : \mathrm{fdr}(x) < q\}$. The power diagnostic tools chosen in those discussions are the non-null average of the local fdr $E_{\mathrm{H}_1}(\mathrm{fdr})$ and the non-null cdf of the local fdr given by $G(q) = P_{\mathrm{H}_1}(\mathrm{fdr} < q) = P_{\mathrm{H}_1}\left(\tilde{R}(q)\right) = \int_{\tilde{R}(q)} f_1(x)dx$. Some empirical estimates of these functions were used in Efron (2007, 2010) for the power analysis.

For the sample splitting method proposed in this paper, where the model is fitted on the p-values or LTA's associated with test statistics, the rejection region from the k^{th} split can be obtained from (2.7). The combined rejection region from all splits can then be constructed as $R(q) = \bigcup_k R_k(q)$ or as $R(q) = \bigcup_k \tilde{R}_k(q)$, depending on the screening tool used. Considering the mixture model setup in (2.1), for a given rejection region $R(q)$ with a cutoff point q, the following probabilities can be used for the power analysis and a relative efficiency comparison:

$$\textbf{Power:}\ P_{\mathrm{H}_1}\left(R(q)\right) = \int\limits_{R(q)} f_1(x)dx \tag{2.8}$$

$$\textbf{Type I error:}\ P_{\mathrm{H}_0}\left(R(q)\right) = \int\limits_{R(q)} f_0(x)dx \tag{2.9}$$

$$\textbf{Type II error:}\ P_{\mathrm{H}_1}\left(R^c(q)\right) = \int\limits_{R^c(q)} f_1(x)dx \tag{2.10}$$

$$\textbf{Precision:}\ P\left(\mathrm{H}_1|R(q)\right) = \frac{p_1 \int_{R(q)} f_1(x)dx}{\int_{R(q)} f(x)dx} \tag{2.11}$$

Here, f_0, f_1 and f are the true densities that follow from the assumption that $X_1, \ldots, X_n$ are *i.i.d.* with p.d.f (2.1). These terms recalling, false negative rate, false positive rate, and precision are commonly used in machine learning (Powers, 2011).

The estimates of eqs. (2.8) to (2.11) for a given data set can be obtained from the following steps:

Rejection Region and Power Calculation Steps.

(i) Suppose sample splitting and subsequent screening were done N times following the steps in Section 2.2.2 and for a given q the rejection regions $R_k(q)$ or $\tilde{R}_k(q)$ (as in (2.7)) were obtained from each of the splits.

(ii) For a two sided test with LTA's the rejection region from the k^{th} split, using either the tail-area Fdr or the local fdr screening, can be written as:

$$R_k(q) = \{x \in \mathbb{R} : \mathrm{Fdr}\,(\mathrm{B}(x)) < q\} = \{x < x_k^*\} \cup \{x > x_k^{**}\}$$
$$\tilde{R}_k(q) = \{x \in \mathbb{R} : \mathrm{fdr}\,(\mathrm{B}(x)) < q\} = \{x < \tilde{x}_k^*\} \cup \{x > \tilde{x}_k^{**}\}.$$

Then writing

$$x^* = \max_k x_k^* \text{ and } \; x^{**} = \min_k x_k^{**},$$
$$\tilde{x}^* = \max_k \tilde{x}_k^* \text{ and } \; \tilde{x}^{**} = \min_k \tilde{x}_k^{**},$$

the combined rejection region can be expressed as:

$$R(q) = \bigcup_{k=1}^{N} R_k(q) = \{x < x^*\} \cup \{x > x^{**}\} \tag{2.12a}$$

or

$$R(q) = \bigcup_{k=1}^{N} \tilde{R}_k(q) = \{x < \tilde{x}^*\} \cup \{x > \tilde{x}^{**}\} \tag{2.12b}$$

depending on the choice of the screening tool. The equation (2.12) will include sets with one sided region only for p-value analysis or one sided tests with LTA's.

(iii) A mixture model $\tilde{f}(x) = \tilde{p}_0\tilde{f}_0(x) + \tilde{p}_1\tilde{f}_1(x)$ with tail adjustment, fitted to the entire data (without any data splitting) can be used for the estimates of the densities in eqs. (2.8) to (2.11).

(iv) Numerical integration and numerical root finding techniques can be used to estimate the probabilities in eqs. (2.8) to (2.11) and to find x_k^* and x_k^{**} or $\tilde{x}_k^*$ and $\tilde{x}_k^{**}$ from (2.7), where closed form solutions are not feasible.

The power and error probabilities in the steps above are associated with the final analysis method that combines all N splits and not with any single training split in particular. Therefore, for the estimation of eqs. (2.8) to (2.11) it is appropriate to use a mixture model fitted to the entire data for estimates of densities f_0, f_1 and f as suggested in step (iii) above. An individual fit from any single particular training split should not be used for the power analysis. Using the combined rejection region $R(q)$ for screening will increase the number of rejections compared to whole data based screening described in Efron (2007). This will naturally increase the power (2.8) of the proposed method but the payoff will be a loss of precision (2.11). Using only the cases in $R(q)$ with high detection frequencies as the potential discoveries will increase the precision of the method. An added benefit of expressing the error probabilities as a function of Fdr cutoff point q is that one can choose q where both the type I and the type II error probabilities are at a reasonable level. Alternatively, an appropriate q can also be chosen so that the proportion of correct classifications $\tilde{\mathrm{A}}(q) = p_1 F_1\left(R(q)\right) + p_0 F_0\left(R^c(q)\right)$ is at a desired level. $\tilde{\mathrm{A}}$ is also known as the "accuracy" function in machine learning.

2.3 Illustrative Examples

In this section we illustrate the proposed methodology with a microarray data set and a RNA-sequencing data set where the goal is to identify genes that are expressed at a significantly higher or lower level in the experiment group compared to the control group. Also, simulated data is used to present the power analysis and other related plots.

2.3.1 Microarray Data Analysis

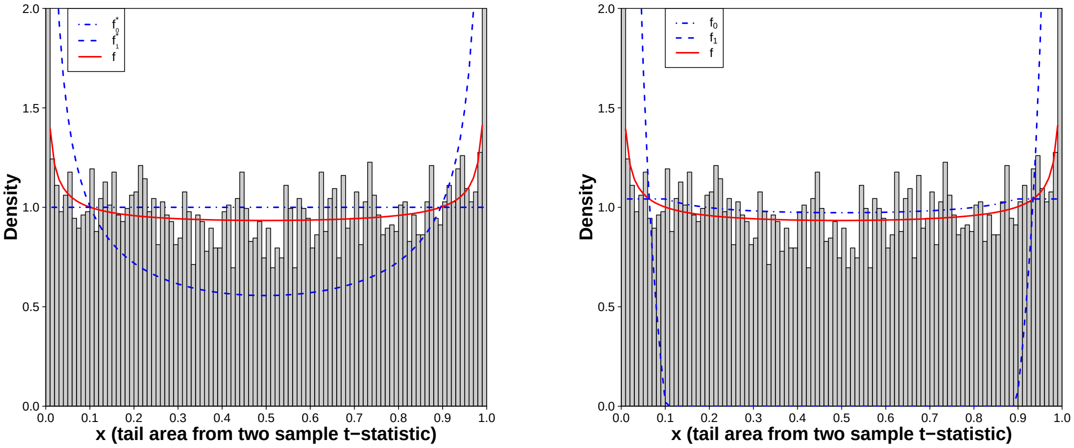
The data set used in this subsection is a prostrate cancer microarray data used in Efron (2010) from Singh et al. (2002), which is available in the R package `sda`, named after “singh2002”. The data consist of 102 microarray samples with expression levels for the same 6033 genes, where 52 samples are for prostate cancer patients and 50 for normal subjects. The analysis aims to detect genes that have significantly different expression levels between the cancer and the non-cancer group, and to explore the inter-relation between these genes. The genes that have significantly different expression levels between the cancer and the non-cancer group are supposed to be captured in the screening part, while a frequency network plot is generated to explore the inter-relation between these genes. We used LTA’s with tail-area Fdr screening for the analysis, although a local fdr screening also can be used following the steps described in Section 2.2.

To begin, the data was split into the training set and the verification set. The training split consisted of 26 randomly selected prostate cancer patients and 25 normal subjects. The remaining 26 patients and 25 non-cancer subjects formed the verification split. The training data was used to fit the contamination model (2.4). This is described next.

The two-sample t-statistic t_i, $i = 1, 2, \ldots 6033$, is calculated for each gene from the training group where it is assumed that t_i follows a central t-distribution with degrees of freedom 26+25-2=49. The LTA for each of these t_i’s is calculated as:

$$x_i = P(t < t_i), \; i = 1, 2, \ldots 6033. \tag{2.13}$$

Note that, x_i’s should be close to 0 or 1 for genes that are deemed significantly differentially expressed in the cancer and non-cancer groups. The histogram of the x_i’s for the 6033 genes in Figure 2.1 shows a bathtub shape that suggests that most of the gene expressions are uniformly distributed but with more than expected (under the uniform) close to 0 and 1. Prompted by the distinctive shape of the histogram, we first fitted a mixture of the uniform distribution in [0,1] Uniform$(0, 1)$ and a Beta distribution to the training data and then readjusted it as described in subsection 2.2.1.

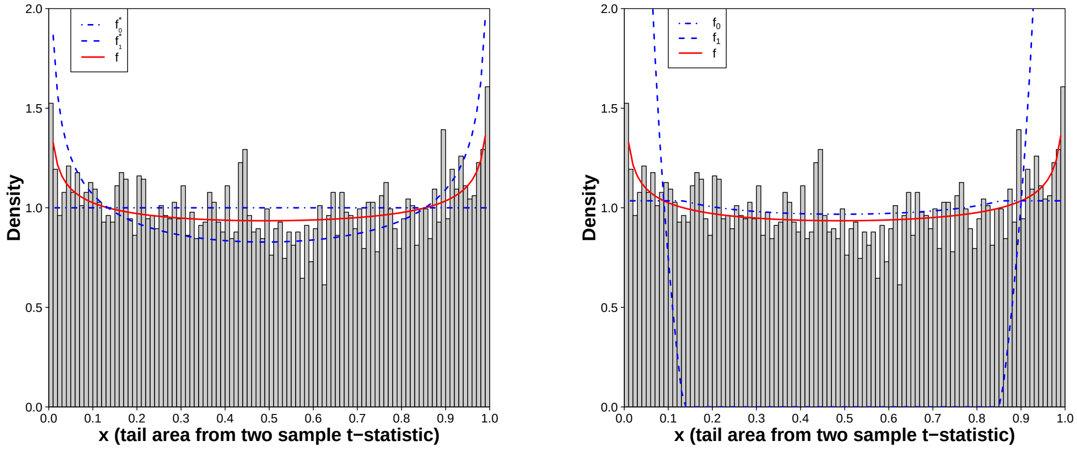


(a) Uniform-Beta mixture model.

(b) Adjusted Uniform-Beta mixture model.

Figure 2.1: The histogram of x_i (left-tail-area for the observed two sample t-statistic for genes $i = 1, 2, \ldots, 6033$) from the entire prostate cancer data set. Superimposed on Figure (a) are the fitted Uniform, Beta and the associated mixture distribution obtained from one particular training split as $\hat{f}(x) = 0.851 f_0^*(x) + 0.149 f_1^*(x)$ where f_0^* is the Uniform$(0, 1)$ p.d.f and f_1^* is the Beta$(0.417, 0.410)$ p.d.f. And on Figure (b) is the empirical null fit adjusted from the fitted Uniform-Beta mixture as in the equation (2.4) as $\hat{f}(x) = 0.96\hat{f}_0(x) + 0.04\hat{f}_1(x)$.

Next following (2.5), we compute the tail-area Fdr associated with each gene. Genes with tail-area Fdr less than 0.1 are declared as significantly different between cancer patients and non-cancer subjects. This procedure was repeated on 100 different sample splits. Out of these 100 repetitions, 66 verification groups identified at least one significant gene with associated tail-area Fdr less than 0.1. The other 34 verification groups failed to capture any significant gene. Few verification sets identified more than one significant gene. After 100 repetitions of this procedure, out of the total 6033 genes, 69 genes were identified to be significantly differentially expressed in the cancer patients compared to non-cancer subjects. These 69 significant



(a) Uniform-Beta mixture model.

(b) Adjusted Uniform-Beta mixture model.

Figure 2.2: The histogram of x_i (left-tail-area for the observed two sample t-statistic for genes $i = 1, 2, \ldots, 6033$) from a particular verification split consisting of half of the control and the treatment group respectively. Superimposed on (a) are the fitted Uniform-Beta and the associated mixture distribution obtained from the corresponding training split as $\hat{f}(x) = 0.622 f_0^*(x) + 0.378 f_1^*(x)$ where f_0^* is the Uniform$(0, 1)$ p.d.f and f_1^* is the Beta$(0.696, 0.736)$ p.d.f. Figure (b) is the empirical null fit adjusted from the fitted Uniform-Beta mixture distribution as in Equation (2.4) where $\hat{f}(x) = 0.966 \hat{f}_0(x) + 0.034 \hat{f}_1(x)$.

genes included some that are expressed at significantly higher levels among the cancer patients (resulting in large t-statistics and consequently x_i's close to 1) and some at significantly lower levels among the cancer patients (resulting in small t-statistics and consequently x_i's close to 0).

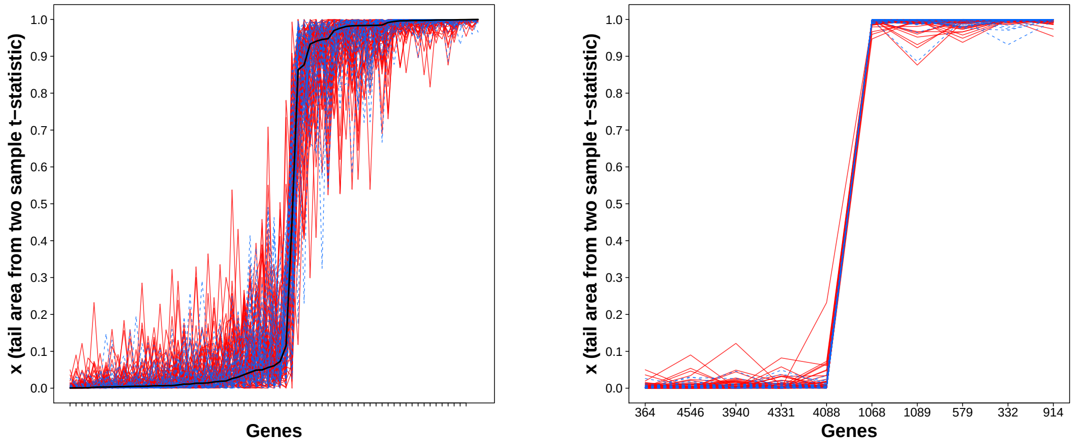
The parallel coordinates graph for 69 significant genes in Figures 2.3(a) and 2.3(b) show the variation among x_i's for these genes observed in 100 different verification splits. The plot reveals a consistent pattern across different splits. The 0.1 critical value for the tail-area Fdr was able to capture at least one significant gene in specific splits (66 splits) and missed the signal in other splits (34 splits). But the the patterns in the parallel coordinate plots indicate these genes are expressed consistently at higher or lower levels throughout all 100 splits. These patterns may suggest biological significance.

Table 2.1: 22 most significant genes from 100 sample splits of the prostate cancer data (with the detection frequency 2 or higher). The third column indicates the frequency of occurrence for the corresponding gene in the 100 splits. The columns med.x, avg.x and sd.x are the median, mean and standard deviation of tail area x (as in Equation (2.13)) for each gene computed from 100 randomly chosen verification data sets.

Gene	**freq**	**med.FDR**	**med**(x)	**avg**(x)	**sd**(x)
610	10	8.45E-02	1.00E+00	9.99E-01	3.93E-03
1720	9	8.43E-02	1.00E+00	9.98E-01	6.28E-03
914	7	7.93E-02	9.99E-01	9.97E-01	5.94E-03
4331	6	7.70E-02	1.22E-03	6.80E-03	1.32E-02
579	5	7.22E-02	9.99E-01	9.93E-01	1.15E-02
1068	4	8.05E-02	9.98E-01	9.96E-01	8.08E-03
4546	4	6.92E-02	1.04E-03	5.74E-03	1.27E-02
1089	3	8.88E-02	9.98E-01	9.91E-01	2.05E-02
364	3	4.17E-02	6.96E-04	3.43E-03	7.35E-03
4518	3	8.97E-02	9.96E-01	9.88E-01	2.38E-02
1130	2	7.55E-02	9.97E-01	9.89E-01	1.95E-02
1458	2	6.73E-02	3.67E-02	6.61E-02	7.41E-02
2856	2	8.84E-02	7.05E-03	2.43E-02	4.57E-02
2945	2	7.33E-02	6.62E-03	1.91E-02	3.30E-02
3017	2	7.19E-02	6.52E-03	2.09E-02	3.41E-02
332	2	4.94E-02	9.99E-01	9.97E-01	8.29E-03
3505	2	8.05E-02	6.90E-03	1.88E-02	2.74E-02
3647	2	7.57E-02	9.97E-01	9.91E-01	1.92E-02
3940	2	6.57E-02	1.10E-03	6.79E-03	1.52E-02
4000	2	6.06E-02	5.18E-03	1.94E-02	4.03E-02
4316	2	9.47E-02	3.07E-03	9.07E-03	1.50E-02
921	2	8.19E-02	4.27E-03	1.55E-02	2.84E-02

Table 2.2: The frequency of occurrence for pairs of significant genes in 100 verification data sets.

Gene-pairs	freq
(1068,914)	2
(914,1720)	2
(914,2945)	2



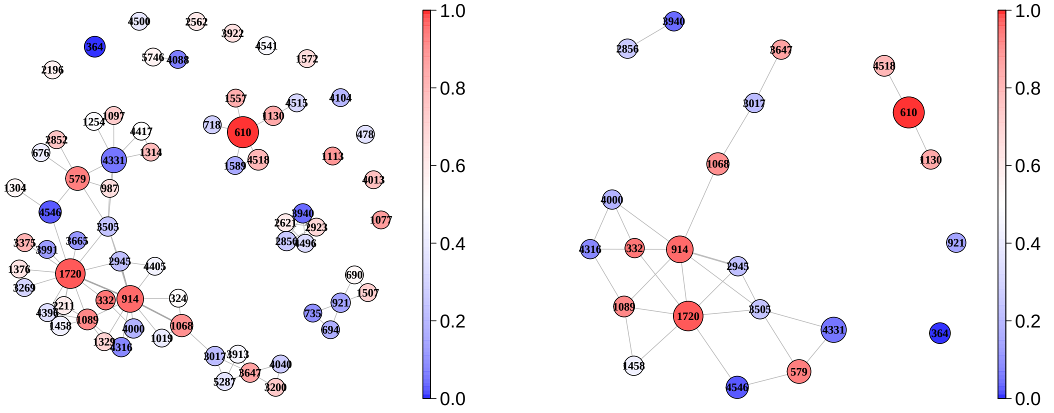
(a) Full parallel coordinate plot.

(b) Partial parallel coordinate plot.

Figure 2.3: Parallel coordinate plot for the detected significant genes using the tail-area Fdr cutoff value $q = 0.1$. Each tick on the horizontal axis represents a significant gene, and the vertical axis shows the left tail-area x from the two sample t-statistic obtained in each of the 100 validation samples. Figure (a) is a full profile for all of the detected significant genes and Figure (b) is the plot for the 10 most significantly differentially expressed genes.

Genes that are repeatedly detected as significant strongly confirm the difference between the patient and the control group. Table 2.1 shows the significant genes (with detection frequency at least 2) along with the number of times they were identified as significant through the 100 sample splits. Some sets of genes were identified as significant as a group more than once. With proper biological oversight and interpretation, these sets of genes may help in the understanding of network relationships. Table 2.2 shows genes identified as significant in groups with the number of times they were identified together (table shows pairs).

Figure 2.4(a) shows the frequently identified significant genes and the sets of genes with which they are simultaneously identified as significant across 100 sample splits and subsequent screenings. In this gene frequency network graph (F-network), the nodes and edges indicate the detected significant genes and the detection of two genes at the same time in a particular split. The node size indicates the frequency of



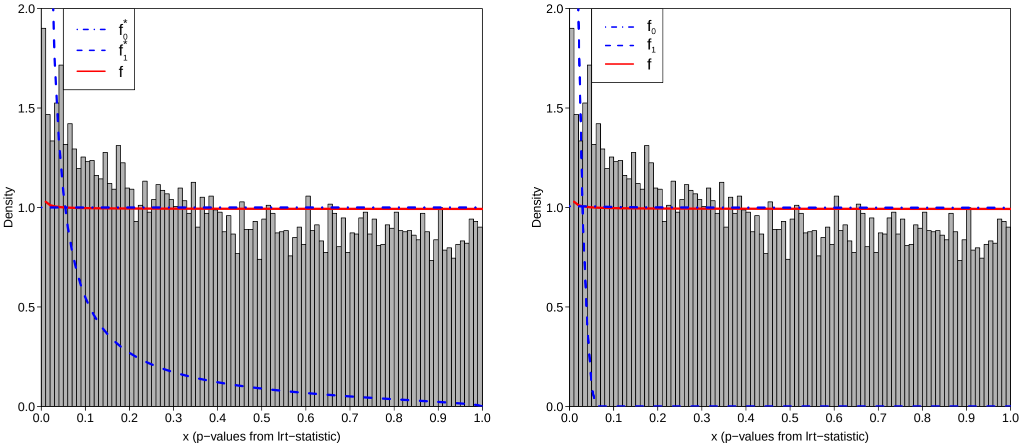
(a) Frequency network for significant genes. (b) Simplified frequency network for (a).

Figure 2.4: Figure 2.4(a) is the entire F-network of 69 significant genes detected at least once in the 100 cross-validation processes with tail-area Fdr ≤ 0.1. Figure (b) is the sparse F-network created from Figure (a) by deleting genes with detection frequency 1. The change of color from blue to red indicates the corresponding gene expression level changing from significantly lower to significantly higher in the cancer group compared to the control group.

detection for that gene and the edge width indicates the frequency of detection for the pair of significant genes at the same time. The node color represents the median tail area from the 100 verification data sets for that gene, for which the color turns from blue to red accordingly as the tail area increases from small (close to 0) to large (close to 1). That is, the red nodes represent genes expressed at significantly higher levels and the blue nodes represent genes expressed at significantly lower levels in the cancer patients' group compared to the controls. Figure 2.4(b) shows genes with at least three edges for a clearer picture into the network.

2.3.2 RNA-seq Data Analysis

The proposed methodology can be used to analyse data from more recent gene expression study mechanism like RNA-sequencing. In this subsection we present such a analysis and also illustrate how the method can be used with p-values. The data consist of RNA-Seq profiles of cell lines derived from lymphoblastoid cells from 69 different Yoruba individuals from Ibadan, Nigeria (Pickrell et al., 2010). The study group is an "opportunity" sample and the samples are likely to be genetically diverse. The aim of analysis is to investigate differentially expressed genes between males and females. The RNA count data are available in the R package `tweeDEseqCountData`. In the raw RNA count data, there are 38415 genes with or without defined annotations. To filter out the noninformative genes, we keep genes with both defined annotations and at least 1 count-per-million (cpm) in at least 20 samples. At last, there are $17,310$ genes remaining for the differential gene analysis.



(a) Unadjusted fit for the whole data. (b) Adjusted fit for the whole data.

Figure 2.5: The mixture model for the p-values using the whole data set. $f(x) = 0.993 f_0^*(x) + 0.007 f_1^*(x)$ where $f_0^*(x)$ is Uniform$(0,1)$ density, and $f_1^*(x)$ is the Beta$(\alpha = 0.064, \beta = 1.517)$ density. The adjusted mixture model is $f(x) = 0.994 f_0(x) + 0.006 f_1(x)$.

Of the 69 individuals, there are 40 females and 29 males. From the whole group 20 female and 15 male subjects were randomly selected to construct the training split, (the remaining subjects constructed the verification split). Assuming that the data follow a negative binomial distribution (Anders and Huber, 2010), the p-value for each gene was calculated using generalized likelihood ratio test comparing male and female subjects in the training split. The training data consist of these $17,310$

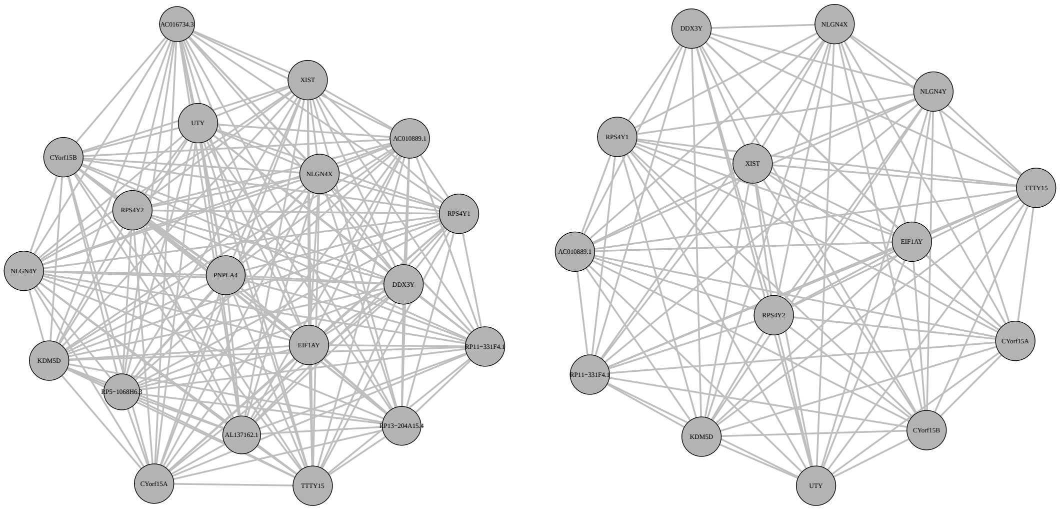
Table 2.3: 28 most frequently detected significant variables from 100 sample split data sets with the tail-area Fdr< 0.1 (detection frequencies 40 or higher). The third column indicates the frequency of occurrence for the corresponding variable (i) in the 100 cross validations. The columns med.x, avg.x and sd.x are the median, mean and standard deviation of the LTA's x for each variable computed from 100 randomly chosen verification data sets.

ID	**Symbol**	**Chrom**	**freq**	**med.FDR**	**med**(x)	**avg**(x)	**sd**(x)
ENSG00000229807	XIST	X	91	2.27E-19	8.92E-25	2.44E-23	7.24E-23
ENSG00000099749	CYorf15A	Y	91	1.70E-16	1.64E-21	1.00E-20	2.58E-20
ENSG00000131002	CYorf15B	Y	91	1.76E-14	1.92E-19	1.09E-18	2.13E-18
ENSG00000157828	RPS4Y2	Y	91	3.85E-14	2.32E-19	4.52E-18	1.49E-17
ENSG00000233864	TTTY15	Y	91	2.60E-13	1.31E-18	4.96E-17	2.03E-16
ENSG00000198692	EIF1AY	Y	91	1.39E-11	3.89E-16	7.25E-15	2.11E-14
ENSG00000165246	NLGN4Y	Y	91	7.37E-10	2.89E-14	3.86E-13	9.19E-13
ENSG00000183878	UTY	Y	91	1.97E-09	1.63E-13	1.87E-12	5.62E-12
ENSG00000243209	AC010889.1	Y	91	4.93E-09	6.55E-13	3.18E-12	1.01E-11
ENSG00000129824	RPS4Y1	Y	91	1.16E-08	3.81E-13	2.03E-11	6.14E-11
ENSG00000012817	KDM5D	Y	91	1.27E-08	1.41E-12	1.41E-11	4.08E-11
ENSG00000213318	RP11-331F4.1	16	91	2.12E-08	2.51E-12	2.82E-11	9.40E-11
ENSG00000067048	DDX3Y	Y	91	2.16E-07	1.56E-11	6.59E-10	2.70E-09
ENSG00000146938	NLGN4X	X	91	4.98E-06	7.24E-10	7.74E-09	1.86E-08
ENSG00000006757	PNPLA4	X	89	1.91E-04	9.08E-08	3.44E-07	8.81E-07
ENSG00000232928	RP13-204A15.4	X	88	2.41E-05	5.84E-09	1.84E-07	6.61E-07
ENSG00000214541	AL137162.1	20	84	4.20E-04	2.16E-07	1.48E-06	2.72E-06
ENSG00000226948	RP5-1068H6.3	20	76	9.72E-04	8.62E-07	2.95E-06	4.99E-06
ENSG00000229920	AC016734.3	2	72	7.49E-04	9.67E-07	1.10E-05	3.66E-05
ENSG00000242058	RP11-143J12.1	18	64	1.40E-03	2.47E-06	9.25E-06	1.58E-05
ENSG00000198034	RPS4X	X	64	1.88E-03	3.49E-06	9.70E-06	1.69E-05
ENSG00000244097	RP11-411G7.1	17	62	1.80E-03	3.95E-06	1.14E-05	2.14E-05
ENSG00000239490	RP11-863N1.1	18	61	9.45E-04	2.68E-06	3.21E-05	8.07E-05
ENSG00000239830	CTD-3116E22.2	19	58	2.64E-03	6.68E-06	2.39E-05	6.72E-05
ENSG00000214203	RP11-135F9.1	12	56	2.73E-03	5.29E-06	1.87E-05	3.00E-05
ENSG00000240371	RP11-624G17.1	11	56	2.82E-03	5.76E-06	1.66E-05	3.06E-05
ENSG00000130021	HDHD1	X	54	1.58E-03	4.20E-06	3.77E-05	9.29E-05
ENSG00000243663	RP11-21K20.1	12	48	2.10E-03	7.33E-06	2.88E-05	4.67E-05

p-values. Similar p-values from the verification split provided the verification data. A mixture of a Uniform and a Beta distribution was fitted to the training data p-values and was adjusted to get the empirical fit as described in Subsection 2.2.1. The training fit then was used to calculate the tail-area Fdr associated with the p-value in the verification data for each gene. Genes with Fdr less than 0.01 were detected as significantly differently expressed between male and female subjects. The process was repeated 100 times.

100 sample splits and subsequent screening detected a total of 83 significant genes

out of 17, 310. Table 2.3 present a partial summary of the detection results. Of these 83 detected significant genes, 56 appeared more than once in the 100 repetitions of the splitting and screening. Especially, the top significant gene XIST (X inactive specific transcript) is known to be expressed only in females, which works to suppress the other pair of X chromosome and then balance the population between females and males. And most frequently detected significant genes appear on the Y or X chromosomes, which is expected since they are differentially expressed between males and females.



(a) F-network for top 19 significant genes (b) F-network for top 14 significant genes

Figure 2.6: F-network for the most significant genes appearing more than 70 times in (a) and more than 90 times in part (b). The Fdr threshold is at 0.01.

Figure 2.6 shows the F-network plot for most significant genes. Since we are using p-values for the analysis, detected genes cannot be flagged as over-expressed or under-expressed (unlike LTA values), but can only be detected as significantly differently expressed in female subjects compared to male subjects. Therefore the color schemes of Figure 2.4 is absent in Figure 2.6. Each node represents a gene, and the gray link between nodes represents the pair of genes are simultaneously differentially expressed. In figure 2.6(a), most of the inner clustered genes comes from the top significant genes from the table 2.3, and the more outward the nodes in the frequency network, the lower the occurrence of those being differentially expressed.

2.3.3 A Simulation Study.

We used simulated data to study the relative efficiency of the proposed method. The data consisted of a treatment group of 50 subjects and a control group of 50 subjects independent from the treatment group. 1000 expressions were simulated for each of these 100 subjects. Out of the total 1000 variables, the first 30 were set to be the nonnull cases (expressions were simulated from distributions different for the treatment and control groups); while the last 970 variables were set to be the null cases (expressions were simulated from the same distribution for the treatment and control groups).

Further, to show that the F-network plots constructed using the detection frequencies can pick up an existing inter-relationship between screened cases, we added a correlation structure among the first 10 non-null variables. For a gene expression study if some genes are inter-related with each other they will work in tandem in any subject no matter whether from the control or from the treatment group, although their expression levels can be different between the two groups. Keeping that in mind, for the simulated data the same correlation structures were applied to both the treatment and the control group while keeping the non-null means different in the two groups.

We used the normal distribution for the simulation. The $N(6, 2)$ distribution was used for all 970 null variables for each subject in the treatment and the control group. The normal distribution parameters used to simulate the 30 non-null variables are described in Table 2.4. The 100 subjects were randomly split into a training and a verification set, each consisting of 25 subjects from the treatment group and 25 subjects from the control group. The data points were defined as $x_i = P(t < t_i)$,

Table 2.4: The simulation parameters for 30 non-null variables. Here 10 values for μ_{c_1} were generated from a $N(5,1)$ distribution, one for each of the variables 11 to 20 in the treatment group. Similarly, 10 values for μ_{c_2} were generated from a $N(7,1)$ distribution, one for each of the output variables 21 to 30 in the treatment group.

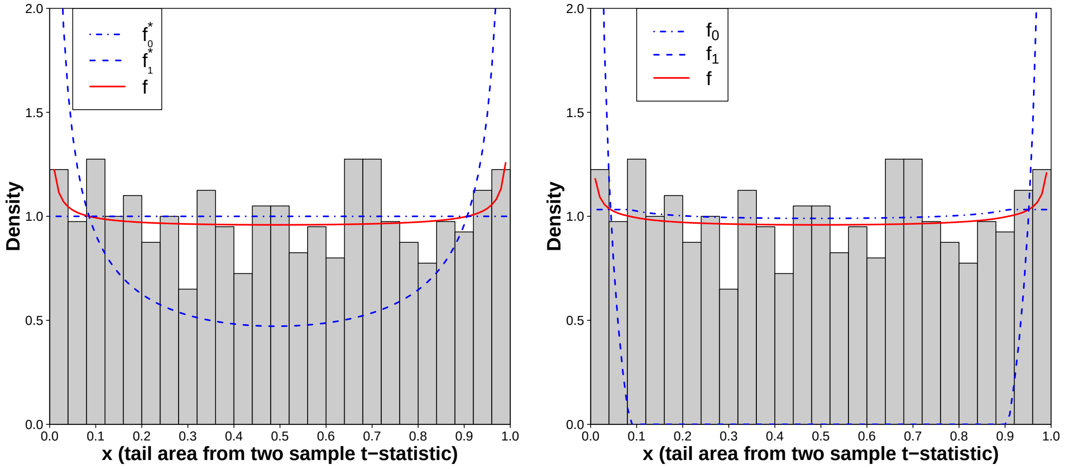
Output Variables	Mean		Variance (treatment and control)
	treatment	control	
$\{1,2\}$	$(7,7)'$	$(6,6)'$	$2 * \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$
$\{3,4\}$	$(5,5)'$	$(6,6)'$	$2 * \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$
$\{5,6,7\}$	$(7.5,7.5,7.5)'$	$(6,6,6)'$	$2 * \begin{bmatrix} 1 & 0.75 & 0.8 \\ 0.75 & 1 & 0.9 \\ 0.8 & 0.9 & 1 \end{bmatrix}$
$\{8,9,10\}$	$(4.5,4.5,4.5)'$	$(6,6,6)'$	$2 * \begin{bmatrix} 1 & -0.85 & -0.9 \\ -.85 & 1 & 0.61 \\ -0.9 & 0.61 & 1 \end{bmatrix}$
$\{11,12,\ldots,20\}$ (independent)	μ_{c_1}	6	2
$\{21,22,\ldots,30\}$ (independent)	μ_{c_2}	6	2

where t_i is the two-sample t-test statistic for each output variable from the 25 control and the 25 treatment samples in the training split. Then a mixture of a Uniform and a Beta distribution was fitted to the training split x_i's and was adjusted to better capture the background and the signal similar to (2.4) in Section 2.3.1. Figures 2.7(a) and 2.7(b) show a fit from one particular training split.

Here we used tail-area Fdr screening to construct the frequency table 2.5 and the F-network plots 2.9. The local fdr screening results are used for comparison purposes in Figures 2.10 and 2.11.

Table 2.5: 26 most frequently detected significant variables from 100 sample split data sets with the tail-area Fdr< 0.1 (detection frequencies 2 or higher). The third column indicates the frequency of occurrence for the corresponding variable (i) in the 100 cross validations. The columns med.x, avg.x and sd.x are the median, mean and standard deviation of the LTA's x for each variable computed from 100 randomly chosen verification data sets.

variable (i)	**freq**	**med.FDR**	**med**(x)	**avg**(x)	**sd**(x)
8	63	3.66E-02	6.05E-05	5.85E-04	1.53E-03
5	45	4.46E-02	1.00E+00	9.99E-01	1.46E-03
10	44	3.74E-02	4.03E-04	2.66E-03	8.62E-03
6	41	4.69E-02	1.00E+00	9.99E-01	2.16E-03
7	38	5.29E-02	1.00E+00	9.99E-01	3.03E-03
15	36	4.49E-02	5.11E-04	3.99E-03	1.05E-02
26	34	5.28E-02	1.00E+00	9.98E-01	3.62E-03
21	22	5.03E-02	9.99E-01	9.96E-01	1.45E-02
9	20	4.85E-02	2.13E-03	1.04E-02	1.89E-02
16	17	4.82E-02	2.33E-03	1.14E-02	2.41E-02
23	17	5.83E-02	9.98E-01	9.95E-01	1.29E-02
30	17	7.37E-02	9.98E-01	9.93E-01	1.57E-02
29	13	4.63E-02	9.98E-01	9.93E-01	1.86E-02
12	12	4.98E-02	6.48E-03	1.89E-02	3.58E-02
20	10	6.33E-02	5.18E-03	2.02E-02	3.75E-02
11	9	6.04E-02	3.11E-03	1.55E-02	3.76E-02
13	6	6.72E-02	1.67E-02	2.78E-02	3.88E-02
19	6	6.93E-02	9.58E-03	1.91E-02	2.50E-02
4	6	6.72E-02	1.16E-02	2.58E-02	4.41E-02
24	5	4.71E-02	9.97E-01	9.91E-01	1.48E-02
3	5	6.42E-02	8.61E-03	2.27E-02	3.69E-02
106	3	3.14E-02	2.36E-02	4.74E-02	6.81E-02
523	3	4.05E-02	2.28E-02	5.72E-02	1.05E-01
27	2	7.01E-02	9.89E-01	9.68E-01	5.34E-02
508	2	7.17E-02	4.12E-02	7.93E-02	1.00E-01



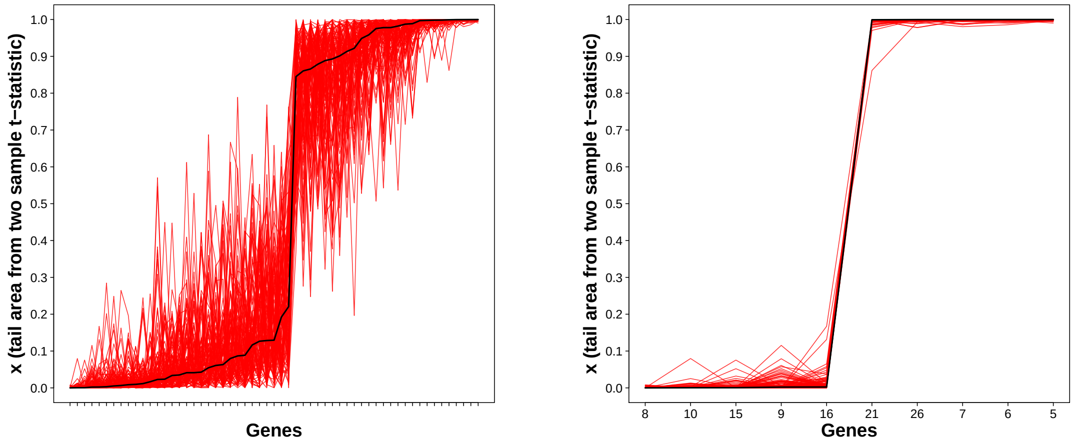
(a) Uniform-Beta mixture model. (b) Adjusted Uniform-Beta mixture model.

Figure 2.7: The histogram of x_i (left-tail-area for the observed two sample t-statistic for the simulated variables $i = 1, 2, \ldots, 1000$) from a verification data set consisting of half of the control and the treatment groups respectively. Superimposed on Figure (a) are the fitted Uniform, Beta and the associated mixture distribution obtained from the corresponding training split as $\hat{f}(x) = 0.923 f_0^*(x) + 0.077 f_1^*(x)$ where f_0^* is the Uniform$(0, 1)$ p.d.f and f_1^* is the Beta$(0.341, 0.319)$ p.d.f. Figure (b) shows the empirical null fit adjusted from the fitted Uniform-Beta mixture distribution to $\hat{f}(x) = 0.974\hat{f}_0(x) + 0.026\hat{f}_1(x)$ as in the equation (2.4).

The two-sample t for each variable in the verification split and its left-tail-area x_i formed the verification data. The training split fit (2.4) was used to obtain the tail-area Fdr associated with each verification data point x_i. The variables with tail-area Fdr less than 0.1 were detected as significant. The process was repeated 100 times. The top 26 most frequently detected significant variables are shown in Table 2.5.

Note that, although variables 1 to 30 out of the 1000 simulated variables were set as non-null, the groups of variables 5, 6, 7 and 8, 9, 10 deviated the most from the null. The mean vectors for variables 1, 2 and 3, 4 did not deviate enough from the null mean to produce significantly large or small t-statistic values. The analysis was done using the t-statistic tail-area and not the original normal distribution, thus naturally non-null variables 1, 2 were not captured in the screening process.

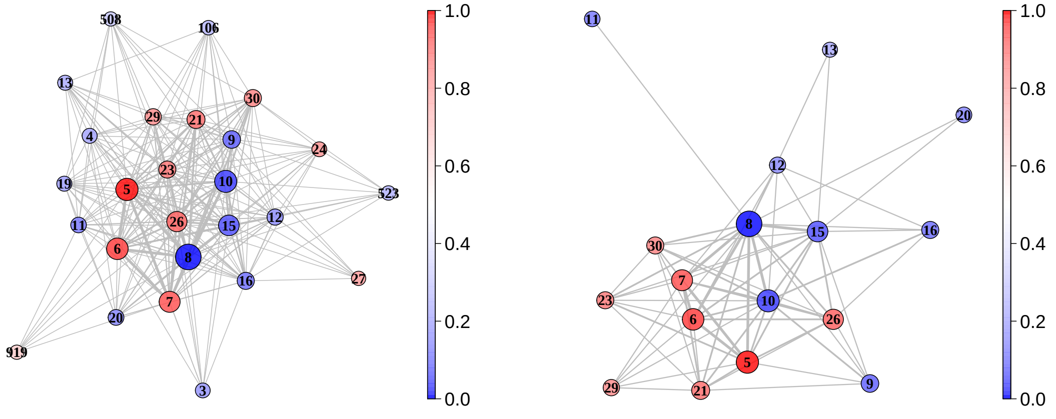
The F-network plots in Figures 2.9(a) and 2.9(b) show that among the detected



(a) Full parallel coordinate plot.

(b) Partial parallel coordinate plot.

Figure 2.8: Parallel coordinate plot for the detected significant variables with the tail-area Fdr less than 0.1. Each tick mark on the horizontal axis represents a significant variable, and the vertical axis shows the left tail-area x from the two sample t-statistic obtained from each of the 100 verification splits. Figure (a) is a full profile for all of the detected significant variables and Figure (b) is the plot for the 10 most significantly different variables.



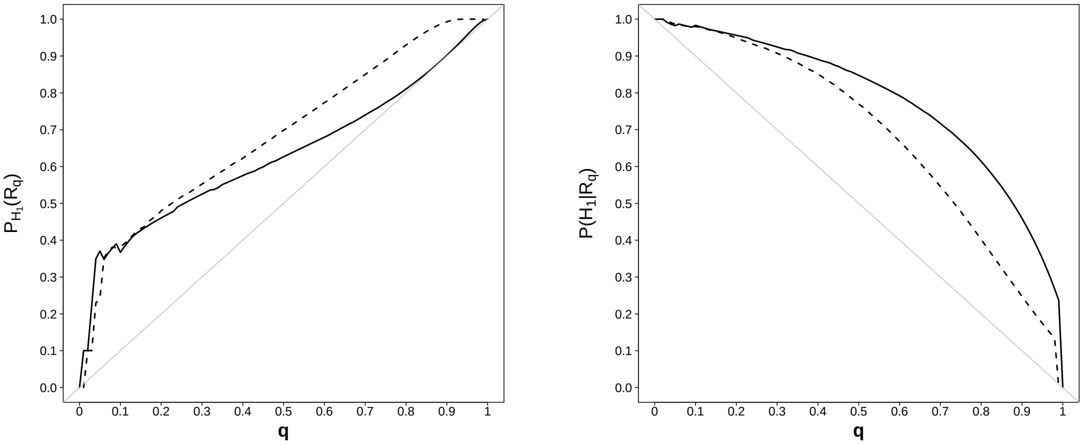
(a) F-network for top 26 significant variables.

(b) Simplified F-network for (a).

Figure 2.9: Figure (a) is the F-network plot of 26 significant variables appearing at least twice in the 100 sample splits by using the tail-area Fdr$\leq$ 0.1. Figure (b) is the simplified network of Figure (a) by deleting variables with less than 5 connected edges.

significant variables, the groups with strongest correlation structure, one with variables 5, 6 and 7 and another with variables 8, 10 (variable 9 had a smaller correlation coefficient 0.61 in the group), were captured successfully.

To compare the efficacy of the proposed method with the whole data fit/screening, we present the following power (recall) (2.8) and precision (2.11) comparison in Figure 2.10. Since Efron (2007) used the local fdr for screening, for comparison purposes with repeated sample splitting, the combined rejection region based on local fdr screening as in (2.12b) is used in Figure 2.10 along with the rejection region from the whole data fit and local fdr screening method.



(a) Power curve comparison.

(b) Precision curve comparison.

Figure 2.10: Figure (a) is the power curve comparison and Figure (b) is the precision curve comparison between the whole data fit/screening method and the proposed sample splitting method. Here q is the cutoff point of the local fdr. The solid line represents power or precision using the whole data fit/screening method and the dashed line represents the same using the sample splitting method.

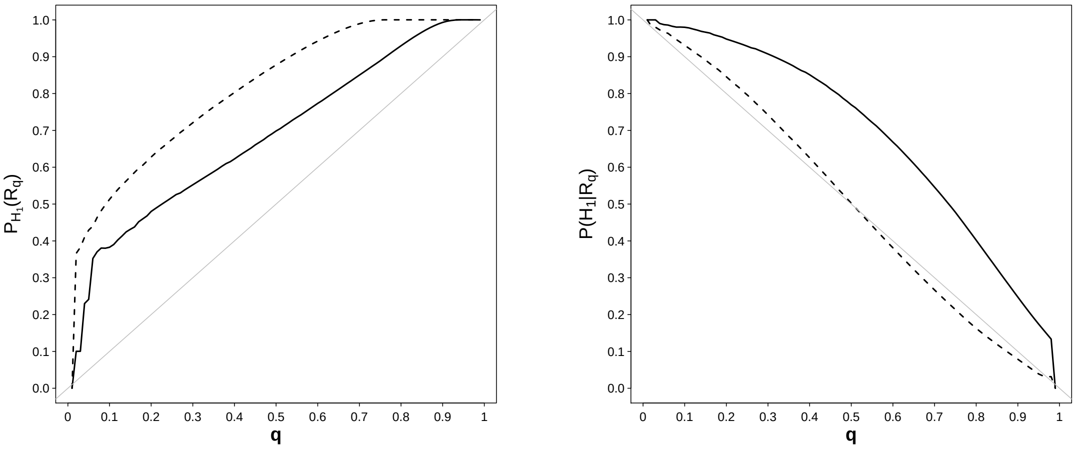
The combined rejection region from repeated sample splits results in a larger rejection region hence providing higher power as evident in Figure 2.10(a). But the inclusion of all detections increases the number of false discoveries and consequently decreases the precision as seen in Figure 2.10(b). However, we are proposing that only the variables with high detection frequencies should be screened as potential true discoveries and not the entire $R(q)$ as presented in Figure 2.10. Eliminating these variables should increase the precision. For example, in case of the tail-area Fdr screening 21 of 26 detections (about 81%) in Table 2.5 with detection frequencies of 2 or higher were true discoveries. If variables with detection frequency 5 or higher are considered, all 21 detections are true discoveries (100%) (recall that in the simulation, variables 1 to 30 were non-null and 31 to 1000 were null).

From Figure 2.10(b), note that even when the entire $R(q)$ was used with the proposed method as the rejection region, at relevant q values 0.1 to 0.3 (reasonable cutoff points for local fdr) the sample splitting technique shows a level of precision that is on par with the whole data fit/screening method. By adding a high frequency criterion to the combined rejection region, the precision is expected to improve, whereas with enough repeated sample splitting, the set of variables in $R(q)$ with high frequencies is unlikely to get much smaller compared to the rejection set produced by the existing methods. In other words, with sufficiently large repetition of sample splitting, the high frequency screening is expected to increase precision without significant loss of power.

Figure 2.11 shows the power and precision comparison between the tail-area Fdr screening and the local fdr screening used with the sample splitting technique. More precisely it compares the performances of the analyses between screening methods i.e,

- With the tail-area Fdr; when the theoretical rejection region is obtained using (2.12a).
- With the local fdr; when the theoretical rejection region is obtained using (2.12b).

The tail-area Fdr screening is expected to provide a larger rejection region at the same critical value q. Consequently the gain of power in Figure 2.11(a) is larger when



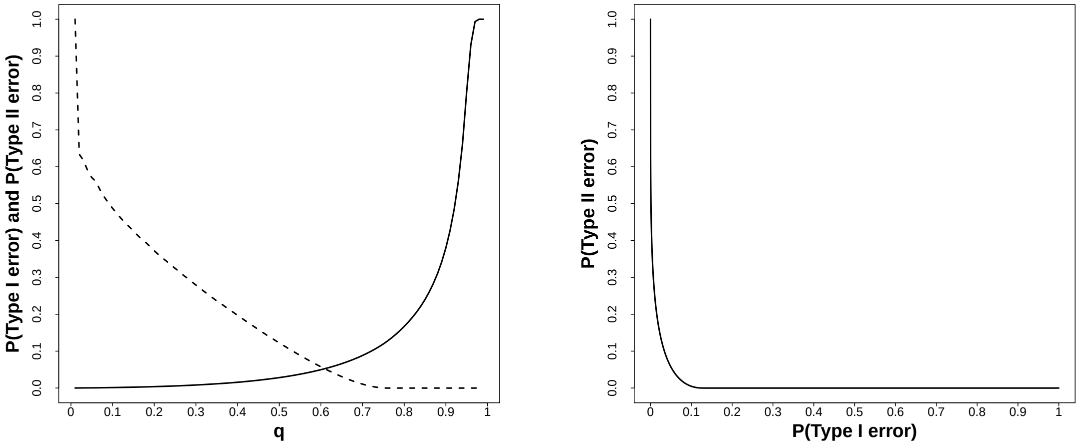
(a) Power curve comparison.

(b) Precision curve comparison.

Figure 2.11: Figure (a) is the power comparison between the tail-area Fdr and the local fdr screening used with sample splitting. Figure (b) is the precision comparison of the same. The dashed line shows power and precision when the tail-area Fdr was used for the screening. The solid line represents power and precision when the local fdr was used for the screening. The probabilities are calculated as a function of the local fdr or the tail-area Fdr cutoff point q, to obtain the corresponding combined rejection region $R(q)$ from 100 sample splits as Equations (2.12a) or (2.12b).

the tail-area Fdr is used as opposed to the local fdr screening. But subsequently the tail-area Fdr screening reduces the precision or increases the percentage of false discovery at the same q as seen in Figure 2.11(b). However, while applying the sample splitting we are using only half of the available information for model building and that is bound to cause some loss of power compared to when the full data is used for the model fitting. For that reason we favor the tail-area Fdr screening with the sample splitting method. Since as evident from the frequency of detection Table 2.5, precision can be greatly improved by considering only the high frequency cases for potential true discoveries.

Figure 2.12 shows type I and type II errors as functions of cutoff points q for the tail-area Fdr screening calculated from the simulated data following steps described in Section 2.2.3. This may help in the choice of appropriate cutoff point q for the main analysis.



(a) Type I and II errors.

(b) ROC curve for Type I and II error.

Figure 2.12: The solid line represents Type I error and the dashed line represents Type II error as in Equations (2.9) and (2.10). The error probabilities are calculated as a function of tail-area Fdr cutoff point q with the rejection region in the Equation (2.12a).

2.4 Discussion

Sample Splitting: In a multiple testing situation, if the entire available data is used for the fitting of a contamination model (null, non-null mixture), then using the same data for the non-null detection may cause a feedback loop. The sample splitting in the proposed method allows a part of the available information to be used for model building and the other part for screening significant cases, and hence avoids that drawback. Further, when the data is randomly split multiple times it produces a different (may not be disjoint) training set each time. When models are fitted based on these different training splits, it helps to neutralize the effect of sources of variation (noise) other than the one that is of interest in a study.

The use of only partial information for the model building part may lead to some loss of power for an individual training set, but repeated sample splitting and combining the resulting rejection regions overcomes that. However, the combined rejection region also accumulates false discoveries and reduces precision. Using cases with high detection frequencies for screening with enough repetition of splits can balance out the power and the precision.

F-Network Plots: The other benefit of the repeated sample splitting is the frequency network or F-network plots that are constructed based on the detection frequencies across the repeated splits. Here we want to emphasize that the F-network plots generated by using this method should not be used as a “proof” of group behavior among the cases. If two unrelated non-null cases deviate strongly from the null distribution, they are bound to be frequently detected as significant, no matter how the data is split. In that case these two unrelated non-null cases may appear in the detected sets repeatedly at the same time and consequently will show up in the F-network plot as a group. Therefore, the F-network plot is not intended to be used as a basis for a causal relation.

However, the simulation study shows that if some screened non-null cases are indeed correlated, that relation is captured in the group structure of the F-network plot. Thus we recommend that these plots to be used as an exploratory tool that precedes further investigation to establish possible causal relationships between cases that show high concurrent detection frequencies in this method. When a study includes thousands of cases, at least some starting point for an exploratory network analysis can be highly useful and cost-effective.

The relevance and effectiveness of the proposed method can be explained particularly well in microarray analysis where the main goal is the identification of groups of differentially expressed genes. These types of studies are commonly used to identify the genes that are associated with a specific biological behavior. Genes detected through the proposed methodology can be isolated for detailed follow-up functional studies. For example, a biologist may look into the few most frequently identified significant genes to distinguish the “regulators”. A systematic gene knockout experiment, conducted on the sets that appear in the F-network plot as a group, can reveal the effect of individual genes on the biological response of interest. This can provide a novel starting point for the elucidation of gene networks or hierarchical regulation patterns in a biological system. Thus, the proposed analysis can guide an exploratory biological study where, instead of experimental investigations of the effect of every gene, a small subset of significant ones can be selected for further experimentation to establish their individual or collective role in the biological response of interest.

Conditional Independence: An interesting question can be posed: “when does the empirical Bayes model (2.1) work?” It of course works when the observations are i.i.d according to (2.1). But in many cases (in particular, the microarray example considered here), it is not correct. It does work, though, in pooling observations where the observations are conditionally independent. For example, in a microarray, clusters of genes may be acting together but still conditionally independent. This is a typical argument used in multiple hypothesis testing cases (Karlin and Taylor, 1981).

Pooling Data: In situations where we have clusters of observations and within a cluster the observations are conditionally independent, pooling of the observations can result in the observations being i.i.d. from the pooled/mixed distribution model. Grego et al. (1990) were one of the first to suggest the use of mixed distribution methodology to analyze such data when the observations are exponentially distributed. Here we provide a justification of this type of analysis for more complicated situations such as the one considered in this paper.

To see this, consider k clusters, where there are n_i observations, $X_{i,1}, \ldots, X_{i,n_i}$ in cluster C_i. Consider the situation where the joint density of all the observations can be written as

$$g(I) \prod_{j=1}^{k} \left[\prod_{m=1}^{n_j} f\left(x_{j,m} \middle| \lambda_{j,m}\right) g_{j,m}\left(\lambda_{j,m} \middle| I_j\right) \right] \tag{2.14}$$

where $\boldsymbol{I} = (I_1, \ldots, I_k)$ is a vector of indicator variables indicating if the clusters are in the background state ($I_j = 0$) or in the signal state ($I_j = 1$). Note that, if $g(\boldsymbol{I}) = \prod_{j=1}^{k} g(I_j)$ (the indicator variables are independent), then the X's are independent.

For example, a cluster might be a biological network of genes where the indicator $I = 0$ denotes that the genes in the network are not being differentially expressed and any expressed genes are simply background while if $I = 1$ the network is being differentially expressed (signal). Typically, we do not know the clusters/networks and are simply pooling the data. Thus, from (2.14)

$$\{(X_{j,m}, \Lambda_{j,m})\} \text{ given } I \text{ are independent.} \tag{2.15}$$

Notice that this is almost an empirical Bayes or a mixture model formulation except that the distributions of the observations are not identically distributed.

However, notice that, given $(\boldsymbol{I}, \boldsymbol{\Lambda})$, the conditional distribution of the X's is given by $X_{j,m}|(\boldsymbol{I}, \boldsymbol{\Lambda}) \sim f(x_{j,m}|\Lambda_{j,m})$. Thus, if we pool the data, then, given $(\boldsymbol{I}, \boldsymbol{\Lambda})$, the resulting X's have marginal mixed density,

$$f(x) = \int f(x|\lambda)m(d\lambda) \tag{2.16}$$

with support $\boldsymbol{\Lambda}$ where the point masses are determined by $\{g_{j,m}\left(\Lambda_{j,m}|\, I_j\right)\}$. That is, we are observing X's for each gene from marginal density (2.16), where they can be considered conditionally independent in the pooled data. Thus, given $(\boldsymbol{I}, \boldsymbol{\Lambda})$, and (2.15) the form of (2.16) justifies the use of the mixture distribution/empirical Bayes that we developed.

2.5 Conclusion

In conclusion, we present a method that can be used for identifying significant cases when carrying out a large number of simultaneous tests. We propose a cross-validation type analysis where a part of the available information goes into the understanding of the underlying process or model fitting while the other part goes into screening for extreme cases. Random splitting and repeated screening provide a way to reduce the noise (other sources of variation) in the analysis and as a by-product we get an exploratory look into the network pattern for significant cases.

Chapter 3

Sparse Regulatory Network Between microRNA and mRNA By Using Weighted P-value Approach[1]

3.1 Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a progressive fatal lung disease with the potential for major complications and is often eventually fatal (Erb-Downward et al., 2011). COPD is characterized by chronically poor airflow caused by an inflammatory response in the lungs resulting in narrowing of the small airways and breakdown of lung tissue known as emphysema. It typically worsens over time and a major cause of disability, and it is the third leading cause of death in the U.S (Mannino and Buist, 2007). Although chronic exposure to smoking, pollutants, etc is known to be closely related to the onset of COPD, the precise mechanisms for the development of this disease have been not fully understood yet.

Recent studies showed that epigenetic alteration is associated with peripheral muscle dysfunction of COPD patients (Barreiro et al., 2005), and several microRNAs are associated with the development of this lung disease (Angulo et al., 2012, Hayashita et al., 2005). MicroRNAs are endogenous and small non-coding RNAs of approximately 21 to 25 nucleotide single-stranded RNAs. MicroRNAs regulate gene expression and it was initially thought that the alteration of microRNAs is as-

[1] Chong Ma, Yen-Yi Ho, Stephanie Christenson, Richard Nho. To be submitted to Bioinformatics

sociated with cancer development (Volinia et al., 2006). However, numerous studies demonstrated that microRNAs are also associated with various human diseases such as cardiovascular disease, idiopathic pulmonary fibrosis (IPF) (Corsten et al., 2010, King Jr et al., 2011), and several microRNAs are known to be deregulated in COPD patients. Furthermore, microRNA signature becomes altered based on the severity of COPD, suggesting that the alteration of microRNAs might be closely linked to the severity of COPD. Thus, defining the role of microRNAs in regulating their target mRNAs in non-severe to severe emphysema is highly imperative to understanding COPD pathogenesis and potentially aiding designing a molecular target for the treatment of COPD.

Christenson et al. (2013) profiled the association between microRNAs and emphysema severity by adjusting the fixed effects of different regions of the lung and random effect of subjects. Nonetheless, we are interested in studying the relationship between the alteration of microRNAs and emphysema severity by integrating gene information and patient demographic information together for gaining more power and removing potential confounding effects.

In this chapter, the goal of our study is to integrate microRNA, mRNA expressions, emphysema severity, and patient demographic information, for establishing a direct link between the alteration of microRNAs for mRNA regulation and emphysema severity in patients. We analyzed the association between mean linear intercept (Lm), a measure of alveolar destruction for lungs and microRNAs using a novel weighted p-value procedure. We obtained the weight-adjusted p-values for 397 miRNAs and further explored their association with genes altered in emphysema severity. Our results showed that 33 microRNAs are significantly associated with the changes in no emphysema to severe emphysema. Moreover, our approach enabled us to identify the potential regulation network between the significant microRNAs and their associated mRNAs in emphysema severity. We propose that our approach to find the direct relationship between severity of emphysema and mRNA changes by the alteration of microRNAs can be potentially applicable for understanding microRNA profiles in an individual patient and useful for the COPD patient-specific treatment in the future.

3.2 Methods

3.2.1 COPD Data

The data we used were obtained from Christenson et al. (2013), which consist of multiple specimens in different lung regions from 8 subjects. Six subjects had undergone lung transplantation for severe COPD and two subjects were donors without COPD. For each subject, paired samples were taken from different regions varying from apex to base in the lung. One sample was used to measure the emphysema severity by the mean linear intercept (Lm), and the adjacent sample was used to measure the 397 microRNA and $22,011$ mRNA expression levels. However, several samples from certain subjects were dropped for quality control reasons and in total, only 57 samples have the mRNA gene expression levels. Additional information on the subjects is also available including the COPD, age, gender, etc. The microRNA and mRNA expression profile datasets are available in Gene Expression Omnibus (GEO) for which the GEO accessions are GSE49881 and GSE27597, respectively.

Figure 3.1 illustrates the log emphysema severity (log(Lm)) from apex (slice 2) to the bottom (slice 13) in the lung for the 8 subjects in the data. Patients with COPD have higher emphysema severity than donors without COPD generally, although log(Lm) demonstrates subject-specific variations and variations in different lung regions, provided that it assumes a linear relationship between log(Lm) and the position of slices in the lung. Taking into account the missing values in various slices of the lung for subjects, we propose to fit the data by using the linear mixed model by dealing the measures of different slices in the lung as repeated measures.

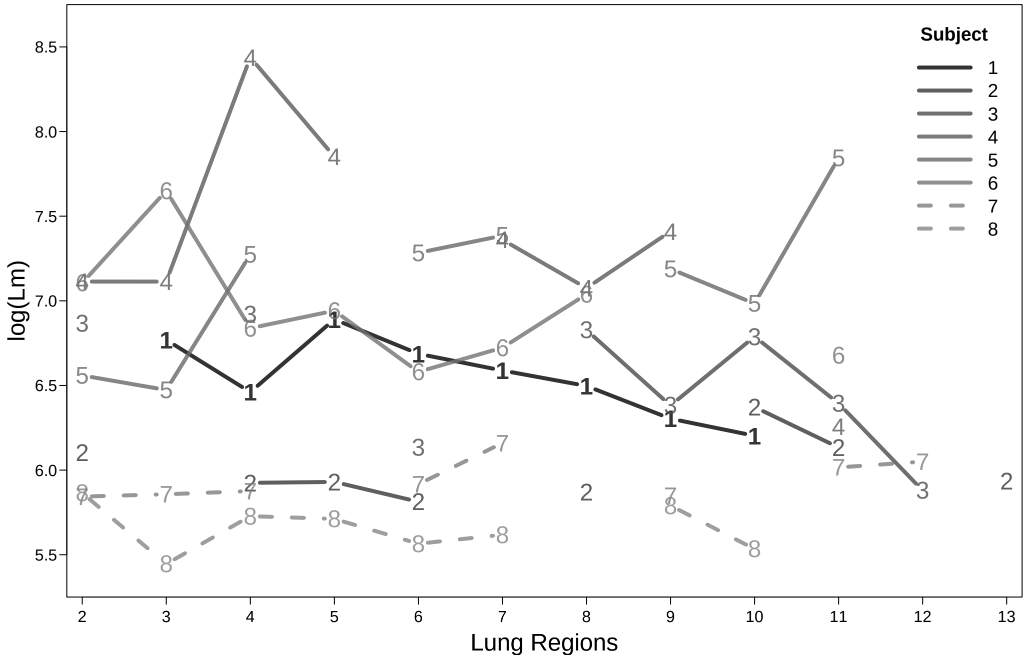


Figure 3.1: Spaghetti plot for log(Lm) with lung regions (slices) for each subject. Slice 2 is the apex in lung and slice 13 is the bottom in the lung. Subject 1 to 6 represent patients with COPD and subject 7 and 8 represent donors without COPD. The spaghetti plot shows a pattern of random intercept log(Lm) for subjects and the overall mean log(Lm) for COPD patients is higher than that for healthy donors.

3.2.2 Methodology

In this paper, we aim to study the association between miRNA and the emphysema severity (Lm) by integrating the mRNA gene information and other important covariate variables. In the analysis, our goal is to identify genetic connections depicted in Figure 3.2. Figure 3.2 indicates miRNAs associated with the emphysema severity by regulating mRNA gene expressions. Put in another way, if a miRNA is significantly associated with a mRNA (Link I) and that mRNA is also significantly associated with the emphysema severity, then it improves power to detect the association between the miRNA and the emphysema severity by carrying out the weighted p-value approach (Roeder and Wasserman, 2009). Meanwhile, it has a potential to unveil how microRNAs regulate COPD-associated gene expression network that underlies the emphysema severity.

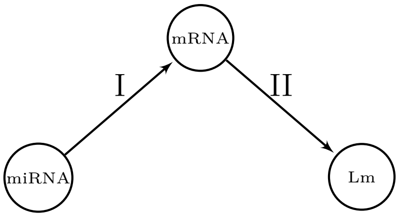


Figure 3.2: The assumed biological pathway amongst microRNA, mRNA and Lm. miRNA is short for microRNA.

Ho et al. (2014) proposed a novel weighted procedure motivated by Roeder et al. (2006) to integrate gene expressions in GWA analysis by gaining more power while controlling the familywise error rate (FWER) at the nominal level α. In this study, we integrate mRNA gene expression profiles and other important covariate variables for detecting significant COPD-associated miRNAs by implementing the weighted p-value procedure. Moreover, we obtain a regulation network between the significant miRNAs and a group of mRNAs (genes) selected by thresholding the weights at a certain quantile. In the COPD data set (Christenson et al., 2013), besides miRNA and mRNA expression levels, there are five important covariate variables available for the 8 subjects. For notation simplification, denote by $x_1 =$ COPD, $x_2 =$ packyears, $x_3 =$ age, $x_4 =$ sex, $x_5 =$ slices and lLm $= \log(\text{Lm})$. Taking the natural log transformation for Lm is beneficial for eliminating the effect of non-normality of the emphysema severity (Lm).

For the sake of model interpretability and avoiding overfitting, we select a best parsimonious model for fitting lLm on the covariate variables, before integrating the miRNA and mRNA expression levels into the selected model. The exploratory analysis implies the existence of nonlinear relationship between packyears, age and lLm. Therefore, we use truncated polynomial basis functions with degree 1 for packyears and age, where the knots are at age $= 61$ and packyears $= 25$ based on the exploratory study, respectively. At the end, the selected model with the minimum AIC is as follows,

$$\begin{aligned}\text{lLm} &= u + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_2 - 25)_+ + \beta_4 x_3 \\ &\quad + \beta_5 (x_3 - 61)_+ + \beta_6 x_4 + \varepsilon \\ &= u + \mathbf{x}'\boldsymbol{\beta} + \varepsilon \end{aligned} \tag{3.1}$$

where $\mathbf{x} = (1, x_1, x_2, (x_2 - 25)_+, x_3, (x_3 - 61)_+, x_4)'$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$ contains the corresponding fixed effects. Note that u is the random intercept indicating subject-specific effects illustrated in figure 3.1.

Now we apply the weighted p-value procedure in (Ho et al., 2014, Roeder et al., 2006) to integrate the miRNA and mRNA expression profiles into the parsimonious model (3.1), for discovering significant COPD-associated miRNAs. We illustrate the procedure in three steps as follows.

$$\text{lLm} = u + \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \tag{3.2}$$

$$\text{lLm} = u + \mathbf{x}'\boldsymbol{\beta} + \beta_{\text{miR}_j,\text{Lm}}\text{miRNA}_j + \varepsilon, \tag{3.3}$$

Step 1. Obtain p-values for 397 miRNAs, denoted by $p_j, j = 1, 2, \ldots, 397$, using likelihood ratio test by comparing the linear mixed effect model (3.2) and model (3.3).

$$\text{mRNA}_k = u + \mathbf{x}'\boldsymbol{\beta} + \beta_{\text{miR}_j,mR_k}\text{miRNA}_j + \beta_{\text{lLm}}\text{lLm} + \varepsilon \tag{3.4}$$

$$\text{lLm} = u + \mathbf{x}'\boldsymbol{\beta} + \beta_{\text{mR}_k,\text{lLm}}\text{mRNA}_k + \beta_{\text{miR}_j}\text{miRNA}_k + \varepsilon \tag{3.5}$$

Step 2. Calculate the weight matrix $\mathbf{W}$ in which each element w_{jk} represents to the extent that the jth miRNA is associated with the emphysema severity lLm by regulating the kth mRNA. The weight matrix $\mathbf{W}$ is by 397 × 22011, where each row represents a miRNA and each column represents a mRNA. The formula for w_{jk} is

$$w_{jk} = \underbrace{\left(\frac{\hat{\beta}_{\text{miR}_j,\text{mR}_k}}{SE(\hat{\beta}_{\text{miR}_j,\text{mR}_k})}\right)^2}_{w_{\text{miR}_j\text{mR}_k}} \times \underbrace{\left(\frac{\hat{\beta}_{\text{mR}_k,\text{lLm}}}{SE(\hat{\beta}_{\text{mR}_k,\text{lLm}})}\right)^2}_{w_{\text{mR}_k\text{lLm}}}$$

By Theorem 3.2.2, p_j is independent of $w_{\text{miR}_j,\text{mR}_k}$ and $w_{\text{mR}_k,\text{lLm}}$, respectively, thus p_j is independent of the weight w_{jk}. It also makes sense since we remove the effect of lLm in model (3.4) and the effect of miRNA$_k$ in model (3.5) respectively so that $w_{\text{miR}_j,\text{mR}_k}$ and $w_{\text{mR}_k,\text{lLm}}$ have no association with p_j. Because p_j is independent with w_{jk} for any k, then it makes p_j independent of the weight $w_{\text{miR}_j} = \max\limits_k w_{jk}$ which is proposed in step 3. The independence between p_j and w_{miR_j} is crucial to guarantee the success of the weight-adjusted p-value of approach for controlling the familywise error rate at level α.

Step 3. Calculate $w_{\text{miR}_j} = \max\limits_k w_{jk}$ and assign the average scaled weight $w^*_{\text{miR}_j} = \frac{w_{\text{miR}_j}}{c}$ to the jth miRNA, where $c = \mathbb{E}(\overline{w}_{\text{miR}})$ and $\overline{w}_{\text{miR}} = \frac{1}{m}\sum_{j=1}^{m} w_{\text{miR}_j}$ is the average of all w_{miR_j}. By Theorem 3.2.2, the rejection set $\mathcal{R} = \{j : \frac{p_j}{w^*_{\text{miR}_j}} < \frac{\alpha}{m}\}$ can control the familywise error rate at level of α, since $w^*_{\text{miR}} > 0$, $p_j \perp\!\!\!\perp w^*_{\text{miR}}$, and the average of all weights $\overline{w}^*_{\text{miR}}$ is 1. Since the distribution of $\overline{w}_{\text{miR}}$ might be complicated, we use the observed value to replace the expected value as an ad hoc approach. In fact, that the average of $w^*_{\text{miR}_j}$ is 1 is required to control familywise error at level α (Roeder and Wasserman, 2009).

Theorem 3.2.1. *Assume that (Y, X) is a random data matrix from $N(\mu, \mathbf{\Sigma})$ where $Y = (Y_1, Y_2, Y_3)$ and $X = (X_1, \ldots, X_p)$. Note that Y_i, X_j are $N \times 1$ vectors. Consider the following regression models*

$$Y_3 = \beta_{13}Y_1 + X\beta + \varepsilon$$

$$Y_2 = \beta_{12}Y_1 + \beta_3 Y_3 + X\beta + \varepsilon$$

$$Y_3 = \beta_{23}Y_2 + \beta_1 Y_1 + X\beta + \varepsilon$$

where $\varepsilon \sim N(\mathbf{0}, \mathbf{\Sigma})$. Let F_{13}, F_{12}, F_{23} denote the F-statistics for the significance of $\beta_{13}, \beta_{12}, \beta_{23}$ in the corresponding models. Then, F_{13} is independent of (F_{12}, F_{23}).

Theorem 3.2.2. *Let $\mathcal{H} = \{H_1, \ldots, H_m\}$ be a set of hypotheses, where $H_j = 0$ represents null and $H_j = 1$ for significance. Denote by $\mathcal{H}_0 = \{j : H_j = 0\}$ the* *set consisting of the true nulls. Suppose that $W_j > 0, j = 1, 2, \ldots, m$ are random variables following some known distributions, P_j is independent of W_j for all $j \in \mathcal{H}_0$, and c is a constant such that $c = \mathbb{E}_{\mathcal{H}}(\bar{W})$ where $\bar{W} = \frac{1}{m}\sum_{j=1}^{m} W_j$. Then the rejection set $\mathcal{R} = \{j : P_j < \frac{\alpha W_j}{mc}\}$ controls the familywise error rate at level α.*

The weight-adjusted p-value for the jth miRNA for demonstrating the statistical evidence of the association between the miRNA and Lm is $\frac{p_j}{w^*_{\text{miR}_j}}$. Here we use the maximum of the "crude" products of the two likelihood ratio test statistics, that is $w_{\text{miR}_j} = \max_k w_{\text{miR}_j,\text{mR}_k} \times w_{\text{mR}_k,\text{Lm}}$ as the weight to scale the "raw" p-value p_j for the significance of the jth miRNA in model (3.3). It makes sense to use a product because w_{jk} would become substantial while both of the two statistics $w_{\text{miR}_j,\text{mR}_k}$ and $w_{\text{mR}_k,\text{Lm}}$ are big enough simultaneously. If merely one of the two statistics is massive and the other is slight, then the product of the two statistics w_{jk} would be not so significant, which adjusts the raw p-value sensibly. In order to gain as much power as possible, we propose to use $\max_k w_{jk}$ as the weight.

There are definitely other ways to define the weight $w^*_{\text{miR}_j}$ by specific interests and purposes. In Table 3.1, we compare the way of taking the maximum of the "crude" products of $w_{\text{miR}_j,\text{mR}_k}$ and $w_{\text{mR}_k,\text{Lm}}$ across all mRNAs and the way of taking the average of them, that is $w_{\text{miR}_j} = \text{avg}_k w_{\text{miR}_j,\text{mR}_k} \times w_{\text{mR}_k,\text{Lm}}$.

In the linear mixed models (3.1), (3.2), (3.3), (3.4) and (3.5), $\mathbf{x}'\boldsymbol{\beta}$ and u have played the same roles in these according models, the fixed effects of COPD, pack-years, age and sex and the random intercept effect caused by lung regions (slices), respectively. By integrating the miRNA and mRNA expression profiles in these models, the weighted p-value procedure could gain more power in detecting the signals for miRNA associated with COPD by regulating mRNA and adjusting associated fixed and random effects, while controlling familywise error at the nominal level α (Ho et al., 2014).

Table 3.1: 33 miRNAs with weight-adjusted p-values (p-values) ≤ 0.05. p.value is calculated from the linear mixed effect models (3.2) and (3.3) by using likelihood ratio test. Weight1 and Weight2 are weights following the formula $w_{miR_j} = \max_k w_{jk}$ and $w_{miR_j} = \text{Avg}_k(w_{jk})$. Padj1 and Padj2 are adjusted p-values by scaling p.value by according weight.

miRNA	Estimate	p.value	Weight1	Padj1	Weight2	Padj2	mRNA
miR-133a	0.50	1.73e-04	1.37	1.26e-04	1.17	1.48e-04	PENK,GPRC6A,HIST2H2BE,PHC1
miR-122	0.40	8.51e-04	3.55	2.40e-04	4.20	2.03e-04	ENGASE,SLC25A45,IL2
miR-137	0.40	3.30e-03	0.54	6.11e-03	0.67	4.94e-03	NEFH,VSNL1,FLJ45244,FBXO10,NDUFA2
miR-96	0.40	5.07e-03	3.18	1.59e-03	0.86	5.89e-03	DSCR9,UBE2J2,TCP10L
miR-629	0.43	7.46e-03	0.55	1.35e-02	0.58	1.28e-02	MYADM,SERPINE1,C5orf17,VCAN
miR-582-5p	0.44	1.05e-02	0.67	1.57e-02	0.56	1.87e-02	LOC440173
miR-337-3p	0.34	1.06e-02	1.95	5.41e-03	1.84	5.73e-03	XKR8,GYPC,TBX19
miR-362-3p	-0.28	1.17e-02	0.98	1.20e-02	0.74	1.59e-02	FRS3,LCE2D,DOCK3,IFNA10
miR-939	0.55	1.25e-02	0.86	1.46e-02	1.16	1.08e-02	CLIC5,SOX4,OTUD6A,OR6A2
miR-374bS	0.36	1.49e-02	2.30	6.46e-03	0.73	2.03e-02	CRMP1,VSNL1,FLJ45244
miR-487b	-0.34	1.55e-02	1.27	1.22e-02	2.13	7.25e-03	HSBP1,GLIPR2,DBF4
miR-211	0.24	1.91e-02	1.77	1.08e-02	1.56	1.22e-02	XKR8,SNRPF,C5orf48
miR-19b-2S	0.33	2.13e-02	1.35	1.57e-02	0.48	4.40e-02	CD209,GYPC,XKR8
miR-181a-2S	-0.30	2.49e-02	0.82	3.04e-02	0.65	3.83e-02	TTYH1,IL23A,APOA2,MCART1
miR-518c	0.40	2.53e-02	0.66	3.80e-02	0.56	4.52e-02	LHB
miR-578	0.38	2.61e-02	1.84	1.42e-02	0.85	3.05e-02	KLHL12,ZDHHC22
miR-136	0.33	2.83e-02	0.54	5.23e-02	0.47	6.08e-02	C8orf86,MCART1,ZNF133,DIO3-OS
miR-520f	0.31	3.01e-02	0.46	6.62e-02	0.62	4.87e-02	NCRNA00161,MCART1
miR-194S	-0.35	3.05e-02	0.64	4.81e-02	0.90	3.38e-02	FLJ46111,VASH2,RERG
miR-924	0.23	3.77e-02	2.11	1.78e-02	0.64	5.87e-02	SSX1,TMCC2,DNAJB7
miR-299-5p	-0.30	3.88e-02	2.49	1.56e-02	2.33	1.67e-02	FCHSD1,IHH,CCDC28B,PVRIG
let-7b	-0.14	4.32e-02	1.78	2.43e-02	0.71	6.11e-02	ALDH1A2
let-7c	-0.14	4.34e-02	1.00	4.35e-02	0.84	5.15e-02	MLL,FBXO17
miR-130aS	0.27	4.42e-02	1.31	3.38e-02	0.91	4.83e-02	MCART1,DGKA,LY9,CD84
miR-593	0.23	4.73e-02	1.16	4.09e-02	1.11	4.28e-02	DIAPH3,PDPN,CEACAM6,RPUSD4,COX5B
miR-10b	-0.19	4.75e-02	1.15	4.11e-02	0.71	6.66e-02	FCHSD1,CD22,DOCK3,CHDH
miR-378	-0.19	4.82e-02	1.83	2.63e-02	0.74	6.50e-02	GPR119,PHC1,DKK3,CHST1
miR-505	0.32	4.95e-02	0.60	8.18e-02	0.53	9.26e-02	C8orf86,C3orf10
miR-128a	-0.31	5.76e-02	1.63	3.54e-02	0.79	7.33e-02	IRX6,FRS3,VSNL1,CYB5RL
miR-99bS	0.33	7.71e-02	1.75	4.41e-02	1.72	4.49e-02	TOR1AIP1,ATP8A2,GPX1
miR-518b	-0.24	8.41e-02	1.59	5.30e-02	3.08	2.73e-02	HNRNPM,CLIC5,ARHGEF10
miR-222S	0.13	9.40e-02	2.40	3.91e-02	3.43	2.74e-02	ZNF174,C19orf30,UCKL1,RCAN2
miR-106aS	0.21	9.82e-02	2.41	4.07e-02	1.63	6.02e-02	MAP2K7,APBA1,IRF2BP1,C5orf48,IMPDH1

3.3 Results

By conducting the weighted procedure in section 3.2.2, we obtain 33 miRNAs which are significantly associated with emphysema severity (Lm) by adjusting the covariates including sex, age and pack of cigarettes consumed each year, where the summary output is listed in Table 3.1. Table 3.1 consists of the miRNA name (miRNA), coefficient estimate for the corresponding miRNA in the linear mixed model (3.3) (Estimate) and its associated p-value (p.value), weights and weight-adjusted p-values and top 5 mRNAs which are most associated with the respective miRNA in terms of weight w_{jk}.

We propose two weighting procedures described in section 3.2.2. In Table 3.1, Weight1 and Padj1 are obtained by using the formula $w_{miR_j} = \max_k w_{jk}$, and Weight2 and Padj2 are obtained by using $w_{miR_j} = \text{Avg}_k(w_{jk})$, accordingly. The 33 miRNAs are selected as long as either of their p.value, Padj1, and Padj2 are less than 0.05. The last column (mRNA) in Table 3.1 is the top 5 mRNAs that are most associated with the corresponding miRNA in terms of the weight w_{jk}. Some of mRNAs do not have annotations though. About one-third of the 33 miRNAs have weights less than 1 and the left two-thirds have weights greater than 1. Roeder and Wasserman (2009) pointed out that power is increased when weight > 1 and decreased when weight < 1. miRNA-122, miRNA-96 and miRNA-229-5p have the top 3 largest weights. In particular, miRNA-128a, miRNA-9bS, miRNA-518b, miRNA-222S, and miRNA-106aS are originally not significant because their raw p-values are greater than 0.05. Because they all have relatively large weights more than 1, their adjusted p-values render them significant, while controlling the overall familywise error rate at level 0.05. The distribution of a section of weights for the 33 miRNAs is shown in figure 3.4.

3.3.1 miR-mRNA Sparse Network

Figure 3.3 illustrates a sparse miRNA-mRNA regulation network, where each node (miRNA) is connected to the top 5 most associated mRNAs in terms of weight w_{jk}. Interestingly, there appear several "cliques" of sub-networks among some miRNAs and mRNAs, in that some mRNAs could be regulated simultaneously by several miRNAs. This finding could be potentially beneficial to further study the pathogenesis of COPD.

miR-133a and miR-378 comprise a small clique which simultaneously regulates PHC1 on associating with emphysema severity. miR-378 gains more power by weight-

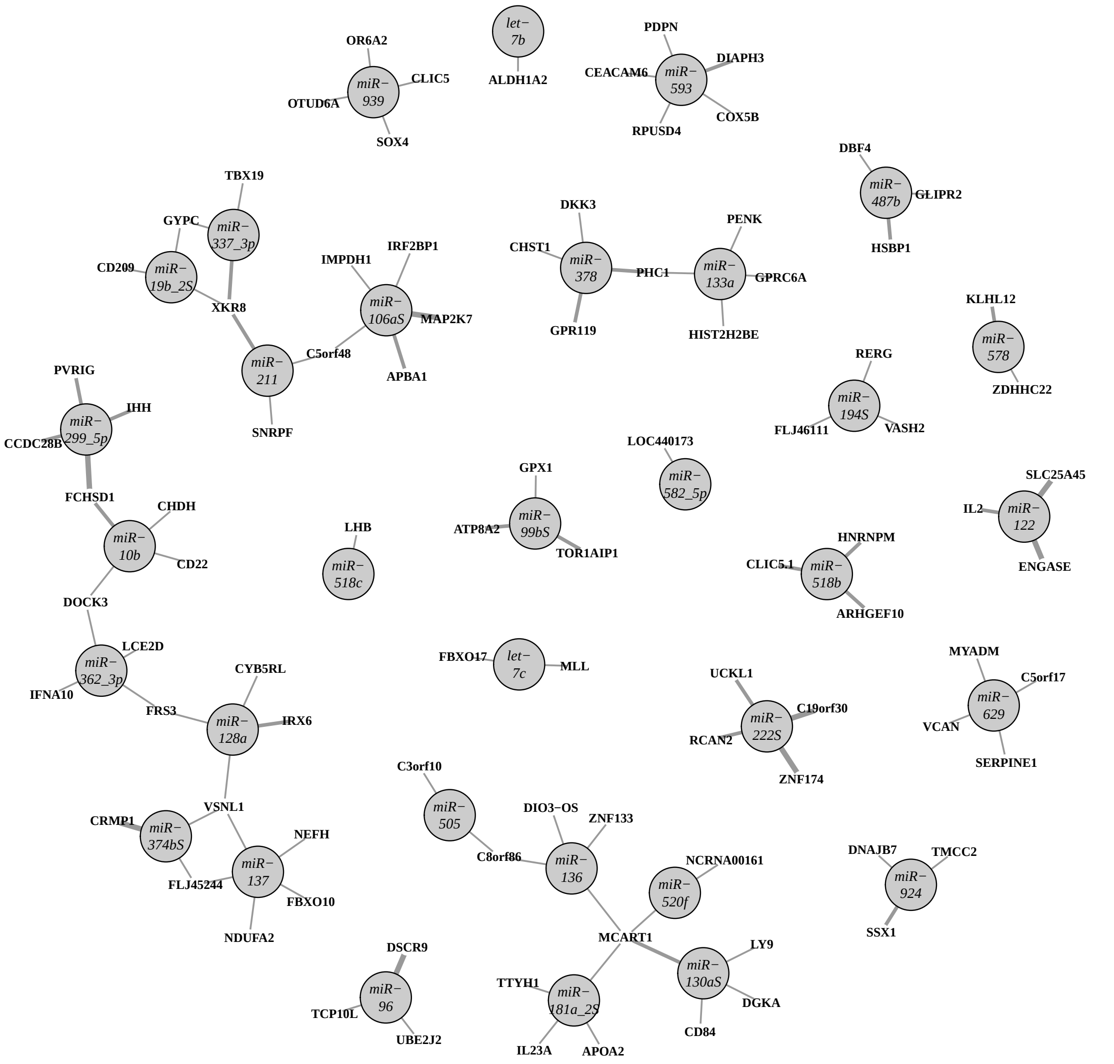


Figure 3.3: A sparse miRNA-mRNA regulation network in association with COPD. The nodes represent the 33 significant miRNAs, which are surrounded by their corresponding 5 most associated mRNAs in terms of weight w_{jk}. The width of a link indicates the strength of association between the miRNA and mRNA, that is, the weight w_{jk}. It shows that several "cliques" of sub-networks among some miRNAs and mRNAs, which are biologically beneficial to unveil the pathogenesis of COPD.

ing PHC1 though. PHC1 is a member of the Polycomb group of genes that can produce a component of a multimeric protein complex that contains EDR2 and the vertebrate Polycomb protein BMH1 (Pruitt et al., 2008). PHC1 is identified as one of the novel lung genes in the bronchial epithelium that is under-expressed for smokers. More interestingly, miR-378 has a negative relationship with emphysemas severity overall, indicating that miR-378 is a putative regulator for PHC1 in COPD.

It appears that miR-211, miR-19b-2S, miR-211 and miR-106aS comprise another clique which regulates genes GYPC, XKR8 and C5orf48 together. miR-211, miR19b-2s, and miR-337-3p are strongly associated with the severity of emphysema by targeting XKR8 and GYPC. XKR8 can promote phosphatidylserine exposure on the apoptotic cell surface, possibly by mediating phospholipid scrambling (Suzuki et al., 2013). XKR8 plays an important role in chronic lung inflammation by controlling apoptotic cell clearance (Grabiec and Hussell, 2016). Besides, miR-106aS is bolstered to be significant by the weighted procedure in which MAP2K7 contributes greatly. Qiu et al. (2017) finds that the p.Glu116Lys rare variant in human mitogen-activated protein kinase kinase 7 (MAP2K7) increases the risk of developing COPD, which could behave as a genetic biomarker for COPD in Chinese. It coincides with our study result that miR-106aS associates with emphysema severity by regulating MAP2K7.

A larger clique is constituted by miR-299-5p, miR-10b, miR-362-3p, miR-128a, miR-374bS and miR-137, regulating several genes such as FCHSD1. FCHSD1 is related to the Mammalian target of the rapamycin kinase (mTOR) pathway which could result in inducing dyskinesia in the treatment of Parkinson's disease. In the gene set enrichment analysis (GESA), Table 3.2 shows that the COPD-associated genes are related to Parkinson's disease. It might be interesting to further study whether FCHSD1 is somehow functional in the development of COPD as well. PVRIG and CCDC28B also have a large association with miRNA-299-5p. PVRIG is significantly down-regulated by human rhinovirus (HRV) infection and HRV alterations of pulmonary epithelial cells are associated with COPD exacerbation (Etemadi et al., 2017). CCDC28B is found to be associated with airway ciliary dysfunction in animal models (Ware et al., 2011). Moreover, CCDC28B is identified as a second site modifier of Bardet-Biedl syndrome (BBS) encoding a protein which affects ciliogenesis in cultured cells in zebra fish (Cardenas-Rodriguez et al., 2013). Though there is no substantial evidence on the relationship between BBS and COPD, it could be beneficial to investigate the functional insights on the cellular basis of CCDC28B effect in COPD patients.

miR-505, miR136, miR-520f, miR-130aS, and miR-181a-2S constitute another clique by mainly regulating MCART1 together. MCART1 is found on chromosome 9 (Pruitt et al., 2008) that is intronless and may be an evolving pseudogene. Because it is transcribed and it contains a full-length coding region, it is currently classified as a protein-coding locus but there is less study on the functioning of its encoded proteins.

In addition to the several cliques illustrated in figure 3.3, there are more isolated miRNAs regulating some genes on their own. miRNA-122 gains the largest power by the weighting procedure, in which IL2 is the third largest associated gene with miRNA-122. Rybka et al. (2016) identified IL2 as a putative inflammatory agent resulting in the depression symptom in COPD patients. ZNF174 plays a vital role in weighting miRNA-222S for gaining more power, and ZNF174 is found to be significantly expressed with sarcoidosis severity (Zhou et al., 2017) and also related to obliterative bronchiolitis in an animal model (Dong et al., 2015).

3.3.2 miR-mRNA Heatmap

Figure 3.4 illustrates a weight submatrix of miRNA by mRNA by 33 × 758, where rows represent the 33 significant miRNAs and columns represent the 758 mRNAs which are the combined top 5 mRNAs that are most associated with each of the overall 397 miRNAs in terms of the weight w_{jk}. Each pixel in Figure 3.4 represents the weight w_{jk} for the jth miRNA and kth mRNA, displayed by using the spectrum varying from white through shades of gray. A dark condensed ribbon implies that the weight w_{miR} used for adjusting the p-value is large. Like miRNA-122, miRNA-337-3p, miRNA-299-5p and miRNA-222S, they all have dark condensed weight ribbons in Figure 3.4 which coincide large weights in Table 3.1.

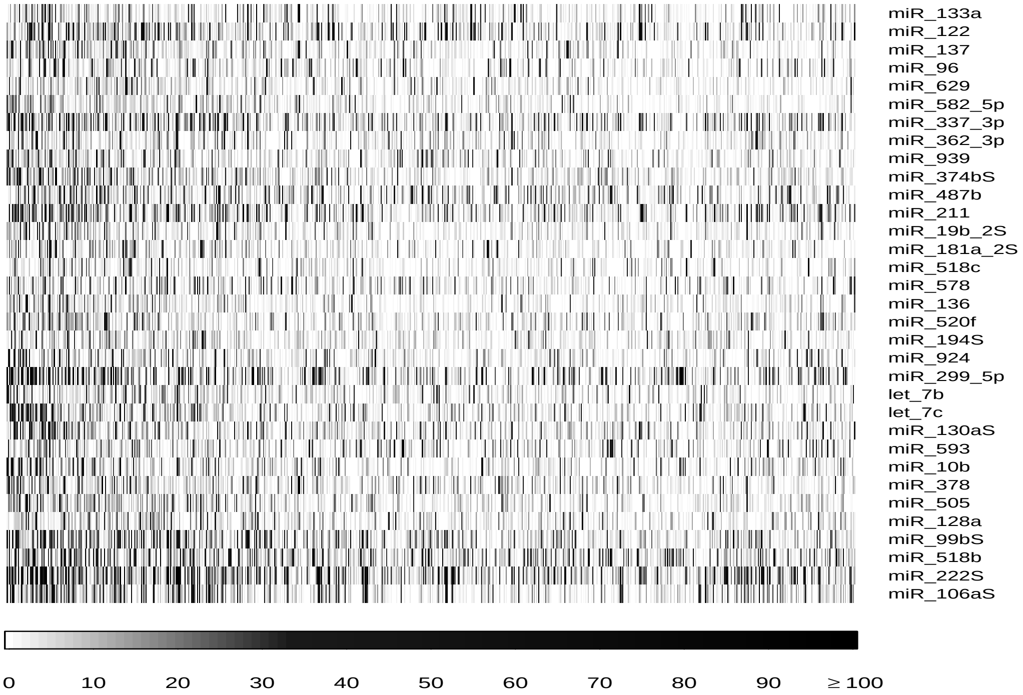


Figure 3.4: Weight matrix. Rows are 33 significant miRNAs by thresholding the weighted p-values (or p-values) ≤ 0.05. Columns are 758 mRNAs which are the combined top 5 mRNAs that are most associated with the 397 miRNAs. Each value in the matrix is calculated by the formula w_{jk} for the jth miRNA and kth mRNA, labeled by the spectrum from white to gray, from the smallest weight to the largest one.

3.3.3 miR-mRNA GSEA

For each gene expression, we access their strength of association with both miRNA and emphysema by summing their corresponding weight through all miRNA using the miRNA × mRNA weight matrix; we named this gene expression weighted score. We performed gene set enrichment analysis for the top 145 genes that are regulated by the 33 miRNAs using KEGG pathway categories. The results of enriched pathways are shown in Table 3.2 and Table 3.1.

Table 3.2: Enriched KEGG pathway using 145 top genes regulated by the 33 microRNAs in table 3.1.

ID	PathName	P-value	Odds	Expected
00983	Drug metabolism - other enzymes	0.022	9.588	0.232
05320	Autoimmune thyroid disease	0.023	9.346	0.238
05012	Parkinson's disease	0.025	5.265	0.633
05014	Amyotrophic lateral sclerosis (ALS)	0.033	7.617	0.288
04630	Jak-STAT signaling pathway	0.043	4.198	0.786

3.4 Discussion

COPD is still a major lung disease characterized by the obstruction of airway flow. Previous studies suggest that miRNAs are altered in the emphysematous lung of various severities but the direct connection between miRNAs and their target mRNAs in various severity degrees of emphysema has been not established yet. In this study, we addressed this question using a weighted p-value approach by integrating the miRNA and mRNA genotype information and other important covariate variables including age, sex, and pack of cigarettes consumed each year. The reason for taking into account those covariate variables is to remove the potential confounding factors which might affect the association between miRNAs and mRNAs. We obtained 33 significant microRNAs which are highly altered in non-severe to severe emphysematous tissues by adjusting effects of mRNAs and other covariates. The weighted procedure (Ho et al., 2014) used in this paper is more statistically powerful. Furthermore, we came up with a sparse miRNA-mRNA regulation network, which has exciting potential to unveil the pathogenesis of COPD.

Based on our results, it is thought that chromosome and mitochondrion homeostasis and apoptosis regulation may be important in the pathogenesis of COPD. Previous studies documented that miRNA let-7c expression was reduced in patients with COPD, and the target genes of let-7c were significantly enriched in the sputum of patients with severe COPD and considerably altered in severe emphysema (Takamizawa et al., 2004). Thus, current study further fortifies the engagement of several genes that are altered in severe COPD and suggests the involvement of potential miRNAs that target mRNA expressions based on the progression of emphysema. In conclusion, our study shows that several miRNAs are altered in severe emphysematous COPD. The confirmation of the level of changes in miRNA and mRNA profiles at the molecular levels from various degrees of emphysema severity will further aid in establishing the direct relationship between miRNA alteration and their target mRNA. MiRNA-based therapy has already been attempted to change the course of cancer and fibrosis. Therefore, obtaining the precise miRNA signature and their direct role in regulating mRNA can be potentially useful for patients via emphysema severity-specific treatment.

3.5 Conclusion

Although a prior study suggested that miRNAs participate in COPD development by changing mRNA expression, the relationship between miRNA alterations and the regulation of their target mRNAs in various degrees of emphysema severity is not yet established. To address this, we utilized a new methodology that permits us to establish 1) whether the progression of severe emphysema from no emphysema alters miRNA signature, 2) the relationship of altered miRNAs and their target mRNA changes in emphysema severity within an individual lung. We re-analyzed previous data with our new linear mixed model by integrating the miRNA and mRNA genotype information and patient demographic information, and we applied the weighted p-value procedure to gain more power while controlling the familywise error rate (FWER) at level α. This study permits us to further identify potential alterations of miRNA/mRNA profiles and whether there is a change of miRNA/mRNA as disease progression. We demonstrate that 33 miRNAs can regulate mRNA gene expressions to effect the emphysema severity. More importantly, several miRNAs appear being strongly significant such as miRNA-133a, miRNA-122, miRNA-137 and miRNA-96 before and after using the weighted procedure. The sparse miRNA-mRNA regulation-network could be substantially meaningful to further study the emphysema pathogenesis amongst the miRNAs and mRNAs.

APPENDIX

PROOF OF THEOREM 3.2.1. Under the normal distribution assumption, we have

$$F_{\beta_{13}} = (N-p-1)\frac{Y_1^T(P_{(X,Y_1)} - P_X)Y_3}{Y_3^T(I_N - P_{(X,Y_1)})Y_3}$$

$$F_{\beta_{12}} = (N-p-2)\frac{Y_2^T(P_{(X,Y_1,Y_3)} - P_{(X,Y_3)})Y_2}{Y_2^T(I_N - P_{(X,Y_1,Y_3)})Y_2}$$

$$F_{\beta_{23}} = (N-p-2)\frac{Y_3^T(P_{(X,Y_1,Y_2)} - P_{(X,Y_1)})Y_3}{Y_3^T(I_N - P_{(X,Y_1,Y_2)})Y_3}$$

Note that P_X is the projection matrix on X. Given (X, Y_1), the $F_{\beta_{13}} \sim F_{1,N-p-1}$ does not depend on (X, Y_1), hence $F_{\beta_{13}}$ is independent of (X, Y_1). Similarly, $F_{\beta_{12}}$ is independent of (X, Y_1, Y_3), and $F_{\beta_{23}}$ is independent of (X, Y_1, Y_2), respectively. Since $F_{\beta_{13}}$ is a function of (X, Y_1, Y_3), then $F_{\beta_{13}}$ is independent of $F_{\beta_{12}}$. Next, we prove $F_{\beta_{13}}$ is independent of $F_{\beta_{23}}$. Denote by

$$V_1 = Y_3^T(P_{(X,Y_1)} - P_X)Y_3$$

$$V_2 = Y_3^T(I_N - P_{(X,Y_1,Y_2)})Y_3$$

$$V_3 = Y_3^T(P_{(X,Y_1,Y_2)} - P_{(X,Y_1)})Y_3$$

where V_1, V_2, V_3 represent individual random variables. Because $P_{(X,Y_1)} - P_X$, $I_N - P_{(X,Y_1,Y_2)}$, and $P_{(X,Y_1,Y_2)} - P_{(X,Y_1)}$ are idempotent and orthogonal to each other, by Craig's theorem, V_1, V_2, V_3 are jointly independent. Under H_0, V_1, V_2, V_3 are from χ^2 distributions with different degrees of freedom so that $V_1 \sim \sigma^2\chi^2_1$, $V_2 \sim \sigma^2\chi^2_{N-p-2}$, and $V_3 \sim \sigma^2\chi^2_1$, respectively. For ease of notation, denote $d = N - p - 2$, $U = F_{\beta_{13}}$, $W = F_{\beta_{23}}$ and $Z = V_3$, accordingly. Since $I_N - P_{(X,Y_1)} = I_N - P_{(X,Y_1,Y_2)} + P_{(X,Y_1,Y_2)} - P_{(X,Y_1)}$, then we have the inverse transformation of (V_1, V_2, V_3) in terms of (U, W, Z) such that

$$\begin{cases} U = \frac{V_1}{V_2+V_3} \cdot (d+1) \\ W = \frac{V_3}{V_2} \cdot d \\ Z = V_3 \end{cases} \Rightarrow \begin{cases} V_1 = \frac{1}{d+1} UZ(\frac{d}{W}+1) \\ V_2 = \frac{Z}{W} \cdot d \\ V_3 = Z \end{cases}$$

And the Jacobian matrix is

$$J = \frac{\partial(V_1, V_2, V_3)}{\partial(U, W, Z)} = \begin{pmatrix} \frac{Z}{d+1}(\frac{d}{W}+1) & -\frac{d}{d+1}\frac{UZ}{W^2} & \frac{U}{d+1}(\frac{d}{W}+1) \\ 0 & -\frac{dZ}{W^2} & \frac{d}{W} \\ 0 & 0 & 1 \end{pmatrix}$$

Therefore, the joint distribution of (U, W, Z) is

$$\begin{aligned} f(u,w,z) &= f(v_1(u,w,z), v_2(u,w,z), v_3(u,w,z))|J| \\ &= f_1(v_1(u,w,z)) f_2(v_2(u,w,z)) f_3(v_3(u,w,z))|J| \\ &\propto \left(\frac{1}{d+1} uz(\frac{d}{w}+1)\right)^{-\frac{1}{2}} e^{-\frac{1}{2}\frac{1}{d+1}uz(\frac{d}{w}+1)} \times \\ &\quad \left(d\frac{z}{w}\right)^{\frac{d}{2}-1} e^{-\frac{1}{2}\frac{dz}{w}} \times z^{-\frac{1}{2}} e^{-\frac{z}{2}} \times \frac{d}{d+1}\frac{z^2}{w^2}(\frac{d}{w}+1) \\ &\propto u^{-\frac{1}{2}} w^{-(\frac{d}{2}+1)} \left(\frac{d}{w}+1\right)^{\frac{1}{2}} z^{\frac{d}{2}} e^{-\frac{z}{2}\left(\frac{u}{d+1}+1\right)\left(\frac{d}{w}+1\right)} \end{aligned}$$

Hence, the joint distribution of (U, W) can be obtained by integrating $f(u, w, z)$ over z, such that

$$\begin{aligned} f(u,w) &\propto u^{-\frac{1}{2}} \left(\frac{u}{d+1}+1\right)^{-\frac{d}{2}} w^{-(\frac{d}{2}+1)} (\frac{d}{w}+1)^{-\frac{d}{2}+1} \\ &\propto f_U(u) f_W(w) \end{aligned}$$

That is to say, the joint distribution of (U, W) can be written as a multiplication of two disjoint distributions of U and W. Therefore, (U, W) are independent and then we prove that $F_{\beta_{13}}$ is independent of $F_{\beta_{23}}$. Overall, F_{13} is independent of (F_{12}, F_{23}). □

PROOF OF THEOREM 3.2.2.

$$\begin{aligned}
P((\mathcal{R} \cap \mathcal{H}_0) > 0) &= P(P_j < \frac{\alpha W_j}{mc} \text{ for some } j \in \mathcal{H}_0) \\
&\leq \sum_{j \in \mathcal{H}_0} P(P_j < \frac{\alpha W_j}{cm}) \\
&\leq \sum_{j \in \mathcal{H}_0} E(I(P_j < \frac{\alpha W_j}{cm})) \\
&\leq \sum_{j \in \mathcal{H}_0} E\{E(I(P_j < \frac{\alpha W_j}{cm}|W_j))\} \\
&\leq \sum_{j \in \mathcal{H}_0} E\{P(P_j < \frac{\alpha W_j}{cm}|W_j)\} \\
&\leq \sum_{j \in \mathcal{H}_0} E(\frac{\alpha W_j}{cm}) \\
&\leq \frac{\alpha}{c} \sum_{j \in \mathcal{H}_0} \frac{E(W_j)}{m} \\
&\leq \alpha
\end{aligned}$$

□

Chapter 4

Discussion

In this dissertation, we explore novel supervised and unsupervised classification methods for functional data and high-dimensional data in genomics studies by employing false discovery rate theory. Supervised and unsupervised classification are common topics in scientific and industrial fields, which involve three tasks: prediction, exploration, and explanation. False discovery rate theory has a close connection to classical classification theory, which must be employed in a sophisticated way to achieve good performance in various contexts.

In Chapter 1, we develop a novel classifier for functional data, which casts the functional data classification problem as a multiple testing task, and the proposed classifier is based on statistical depth functions involving the application of false discovery rate and negative predictive value. Both the simulation studies and real benchmark data analysis illustrate that our proposed method is competitive with other classifiers, such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), and Neural Networks, etc., in the multivariate and functional contexts. Motivated by the success of studying false discovery rate in supervised classification, we present novel methods for applying false discovery rate in unsupervised classification for high-dimensional data. Chapter 2 and 3 essentially deal with the large scale testing problem in genomics studies by using false discovery rate in different perspectives.

In Chapter 2, we propose a novel algorithm to yield reproducible differential expression analysis for microarray and RNA-Seq data. In large scale testing problems, p-values are usually obtained by using the whole data, which are in turn used to conduct the significance screening for all of the hypotheses parametrically or nonparametrically. Our proposed algorithm combines the cross-validation type subsampling and false discovery rate, where the p-values obtained from the training data are used to fit a mixture of baseline and signal distributions, which is in turn used to screen the significance for the p-values obtained from the testing data. In this way, our proposed algorithm can not only overcome the overfitting issue but is also able to obtain reproducible significant detections for the large scale hypotheses. The simulation studies illustrate our proposed algorithm is more powerful and flexible than the general approach of applying the false discovery rate to the whole data once.

In Chapter 3, we propose a novel weighted p-value approach to explore the association between microRNAs and COPD emphysema severity by regulating the mRNA expressions, while integrating patient phenotype information. Our new approach also enables us to find a sparse regulatory network between the significant miRNAs and their most associated mRNAs. The simulation study shows that our method is more powerful than merely using the raw marginal p-values from multiple hypotheses, while controlling the familywise error rate or false discovery rate. Most importantly, under the normal distribution assumption, our proposed method can be applied to study the causality between miRNA and any particular disease, by exploring the precise role of miRNA in regulating genes.

Bibliography

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

Martín Angulo, Emilia Lecuona, and Jacob Iasha Sznajder. Role of micrornas in lung disease. *Archivos de Bronconeumología (English Edition)*, 48(9):325–330, 2012.

Esther Barreiro, Beatriz De La Puente, Joan Minguella, Josep M Corominas, Sergi Serrano, Sabah NA Hussain, and Joaquim Gea. Oxidative stress and respiratory muscle dysfunction in severe chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 171(10):1116–1124, 2005.

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

Magdalena Cardenas-Rodriguez, Daniel PS Osborn, Florencia Irigoín, Martín Graña, Héctor Romero, Philip L Beales, and Jose L Badano. Characterization of ccdc28b reveals its role in ciliogenesis and provides insight to understand its modifier effect on bardet–biedl syndrome. *Human genetics*, 132(1):91–105, 2013.

Stephanie A Christenson, Corry-Anke Brandsma, Joshua D Campbell, Darryl A Knight, Dmitri V Pechkovsky, James C Hogg, Wim Timens, Dirkje S Postma, Marc Lenburg, and Avrum Spira. mir-638 regulates gene expression networks associated with emphysematous lung destruction. *Genome medicine*, 5(12):114, 2013.

Gerda Claeskens, Mia Hubert, Leen Slaets, and Kaveh Vakili. Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 109(505):411–423, 2014.

Maarten F Corsten, Robert Dennert, Sylvia Jochems, Tatiana Kuznetsova, Yvan Devaux, Leon Hofstra, Daniel R Wagner, Jan Staessen, Stephane Heymans, and Blanche Schroen. Circulating microrna-208b and microrna-499 reflect myocardial damage in cardiovascular disease. *Circulation: Genomic and Precision Medicine*, pages CIRCGENETICS–110, 2010.

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496, 2007.

Ming Dong, Xin Wang, Hong-Lin Zhao, Xing-Long Chen, Jing-Hua Yuan, Jiu-Yi Guo, Ke-Qiu Li, and Guang Li. Integrated analysis of transcription factor, microrna and lncrna in an animal model of obliterative bronchiolitis. *International journal of clinical and experimental pathology*, 8(6):7050, 2015.

B. Efron. Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377, 2007.

B. Efron. Microarrays, empirical Bayes, and the two-groups model. *Statistical Science*, 23:1–22, 2008.

B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation*. Testing and Prediction. Cambridge University Press, Cambridge, UK, 2010.

B. Efron and C. Morris. Stein's estimation rule and its competitors–an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

John R Erb-Downward, Deborah L Thompson, Meilan K Han, Christine M Freeman, Lisa McCloskey, Lindsay A Schmidt, Vincent B Young, Galen B Toews, Jeffrey L Curtis, Baskaran Sundaram, et al. Analysis of the lung microbiome in the âĂIJhealthyâĂİ smoker and in copd. *PloS one*, 6(2):e16384, 2011.

Mohammad Reza Etemadi, King-Hwa Ling, Shahidee Zainal Abidin, Hui-Yee Chee, and Zamberi Sekawi. Gene expression patterns induced at different stages of rhinovirus infection in human alveolar epithelial cells. *PloS one*, 12(5):e0176947, 2017.

Frédéric Ferraty and Philippe Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17(4):545–564, 2002.

Frédéric Ferraty and Philippe Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics and Data Analysis*, 44(1):161–173, 2003.

Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media, 2006.

Ricardo Fraiman and Graciela Muniz. Trimmed means for functional data. *Test*, 10 (2):419–440, 2001.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics, New York, 2001.

Nathan C Fuenffinger. *Optical spectroscopy and chemometrics for discriminations of dyed textile fibers and magnetic audio tapes*. PhD thesis, University of South Carolina, 2015.

Anil K Ghosh and Probal Chaudhuri. On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 32(2):327–350, 2005.

Aleksander M Grabiec and Tracy Hussell. The role of airway macrophages in apoptotic cell clearance following acute and chronic lung inflammation. In *Seminars in immunopathology*, volume 38, pages 409–423. Springer, 2016.

J. Grego, H-L. Hsi, and J. Lynch. A strategy for analyzing mixed and pooled exponentials. *Applied Stochastic Models And Data Analysis.*, VI:59–70, 1990.

Peter Hall and Nancy E Heckman. Estimating and depicting the structure of a distribution of random functions. *Biometrika*, 89(1):145–158, 2002.

Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

Yoji Hayashita, Hirotaka Osada, Yoshio Tatematsu, Hideki Yamada, Kiyoshi Yanagisawa, Shuta Tomida, Yasushi Yatabe, Katsunobu Kawahara, Yoshitaka Sekido, and Takashi Takahashi. A polycistronic microrna cluster, mir-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer research*, 65(21):9628–9632, 2005.

Daniel Hlubinka, Irène Gijbels, Marek Omelka, and Stanislav Nagy. Integrated data depth for smooth functions and its application in supervised classification. *Computational Statistics*, 30(4):1011–1031, 2015.

Yen-Yi Ho, Emily C Baechler, Ward Ortmann, Timothy W Behrens, Robert R Graham, Tushar R Bhangale, and Wei Pan. Using gene expression to improve the power of genome-wide association analysis. *Human heredity*, 78(2):94–103, 2014.

Samuel Karlin and Howard E Taylor. *A second course in stochastic processes*. Elsevier, 1981.

Talmadge E King Jr, Annie Pardo, and Moisés Selman. Idiopathic pulmonary fibrosis. *The Lancet*, 378(9807):1949–1961, 2011.

Jun Li, Juan A Cuesta-Albertos, and Regina Y Liu. Dd-classifier: Nonparametric classification procedure based on dd-plot. *Journal of the American Statistical Association*, 107(498):737–753, 2012.

Regina Y. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

Pamela Llop, Liliana Forzani, and Ricardo Fraiman. On local times, density estimation and supervised classification from functional data. *Journal of Multivariate Analysis*, 102(1):73–86, 2011.

Sara López-Pintado and Juan Romo. Depth-based classification for functional data. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72: 103, 2006.

Sara López-Pintado and Juan Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009.

Sara López-Pintado, Ying Sun, Juan K Lin, and Marc G Genton. Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, 8(3):321–338, 2014.

David M Mannino and A Sonia Buist. Global burden of copd: risk factors, prevalence, and future trends. *The Lancet*, 370(9589):765–773, 2007.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Stephen L Morgan. Evaluation of statistical measures for fiber comparisons: Interlaboratory studies and forensic databases. 2014. [Online; accessed October-2014].

O. Murlidharan. An empirical Bayes mixture method for effect size and false discovery rate. *Annals of Applied Statistics*, 4(1):422–438, 2010.

Naveen N Narisetty and Vijayan N Nair. Extremal depth for functional data and applications. *Journal of the American Statistical Association*, 111(516):1705–1714, 2016.

Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768, 2010.

D. M. W Powers. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

Kim D Pruitt, Tatiana Tatusova, William Klimke, and Donna R Maglott. Ncbi reference sequences: current status, policy and new initiatives. *Nucleic acids research*, 37(suppl_1):D32–D36, 2008.

Fuman Qiu, Yinyan Li, Xiaoxiao Lu, Chenli Xie, Qingqing Nong, Di Wu, Jiansong Chen, Lei Yang, Yifeng Zhou, and Jiachun Lu. Rare variant of map2k7 is associated with increased risk of copd in southern and eastern chinese. *Respirology*, 22 (4):691–698, 2017.

Jim Ramsay and Bernard Silverman. *Functional Data Analysis*. Springer Science, New York, 2005.

H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 157–163. University of California Press, Berkeley, CA., 1956.

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

Kathryn Roeder, Silvi-Alin Bacanu, Larry Wasserman, and B Devlin. Using linkage genome scans to improve power of association in genome scans. *The American Journal of Human Genetics*, 78(2):243–252, 2006.

Joanna Rybka, S Mechiel Korte, Małgorzata Czajkowska-Malinowska, Małgorzata Wiese, Kornelia Kędziora-Kornatowska, and Józef Kędziora. The links between chronic obstructive pulmonary disease and comorbid depressive symptoms: role of il-2 and ifn-γ. *Clinical and experimental medicine*, 16(4):493–502, 2016.

D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, Lander, E., Loda, M., P. Kantoff, T. Golub, and R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell.*, 1: 203–209, 2002.

John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

John D Storey. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):347–368, 2007.

John D Storey et al. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

Jun Suzuki, Daniel P Denning, Eiichi Imanishi, H Robert Horvitz, and Shigekazu Nagata. Xk-related protein 8 and ced-8 promote phosphatidylserine exposure in apoptotic cells. *Science*, 341(6144):403–406, 2013.

Junichi Takamizawa, Hiroyuki Konishi, Kiyoshi Yanagisawa, Shuta Tomida, Hirotaka Osada, Hideki Endoh, Tomoko Harano, Yasushi Yatabe, Masato Nagino, Yuji Nimura, et al. Reduced expression of the let-7 micrornas in human lung cancers in association with shortened postoperative survival. *Cancer research*, 64(11): 3753–3756, 2004.

Read D Tuddenham and Margaret M Snyder. Physical growth of California boys and girls from birth to eighteen years. *Publications in child development. University of California, Berkeley*, 1(2):183, 1954.

Stefano Volinia, George A Calin, Chang-Gong Liu, Stefan Ambs, Amelia Cimmino, Fabio Petrocca, Rosa Visone, Marilena Iorio, Claudia Roldo, Manuela Ferracin, et al. A microrna expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National academy of Sciences of the United States of America*, 103(7):2257–2261, 2006.

Stephanie M Ware, Meral Gunay Aygun, and Friedhelm Hildebrandt. Spectrum of clinical diseases caused by disorders of primary cilia. *Proceedings of the American Thoracic Society*, 8(5):444–450, 2011.

Tong Zhou, Nancy Casanova, Nima Pouladi, Ting Wang, Yves Lussier, Kenneth S Knox, and Joe GN Garcia. Identification of jak-stat signaling involvement in sarcoidosis severity via a novel microrna-regulated peripheral blood mononuclear cell gene signature. *Scientific reports*, 7(1):4237, 2017.

Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of Statistics*, 28(2):461–482, 2000a.

Yijun Zuo and Robert Serfling. Structural properties and convergence results for contours of sample statistical depth functions. *Annals of Statistics*, 28(2):483–499, 2000b.