On Combination of Elliptically Distributed Biomarkers to Improve Diagnostic Accuracy

by Shiqi Dong

B.S. in Statistics, June 2016, Central China Normal University
M.S. in Statistics, May 2018, The George Washington University

A Dissertation submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

August 31, 2021

Dissertation directed by

Zhaohai Li
Professor of Statistics and Biostatistics

Aiyi Liu
Senior Investigator, Biostatistics & Bioinformatics Branch, NICHD/NIH

The Columbian College of Arts and Sciences of The George Washington University certifies that Shiqi Dong has passed the Final Examination for the degree of Doctor of Philosophy as of July 2021. This is the final and approved form of the dissertation.

On Combination of Elliptically Distributed Biomarkers to Improve Diagnostic Accuracy

Shiqi Dong

Dissertation Research Committee:

Zhaohai Li, Professor of Statistics and Biostatistics, Dissertation Co-Director

Aiyi Liu, Senior Investigator, Senior Investigator, Biostatistics & Bioinformatics Branch, NICHD/NIH, Dissertation Co-Director

Feifang Hu, Professor of Statistics, Committee Member

Lin Wang, Assistant Professor of Statistics, Committee Member

© Copyright 2021 by Shiqi Dong
All rights reserved

Acknowledgments

Time flies in a blink of an eye, and my studies at the George Washington University are about to come to an end. Through the past three years, I have received many helps from many people: my respected advisors, my dear classmates and my best friends. During the time when my graduation thesis is about to be completed and I am going to graduate, I would like to express my gratitude toward my advisors who have provided lots of guidance for my research, my family members who have supported my study and work, as well as friends who have helped me.

First of all, I would like to express my sincere thanks to my advisor Professor Zhaohai Li. In the past three years, his energy, sincerity and motivation have deeply inspired me. The knowledge learned in the books is indeed important, but when I worked with him, I learned more extracurricular things that help in life, social intercourse and job hunting. There are also helps provided by the co-advisor Dr. Aiyi Liu. When I faced difficulties, he always encouraged me carefully and patiently, and provided me with a lot of helpful suggestions. It is a great honor to work and study under his guidance.

I wish to thank Professor Feifang Hu and Professor Lin Wang for being the readers of my graduation committee. As the statistics PhD program director, Prof. Hu not only gave me helpful comments and suggestions for my dissertation, but also helped me a lot during my PhD studies. The interesting questions raised by Prof. Hu and Prof. Wang made me realize different angles of thinking about problems, which helped me expand my thinking.

I also want to appreciate Dr. Yuanzhang Li. I have accumulated a great amount of experience on how to apply statistical methods to actual data through working with him in the Walter Reed. Dr. Li provided me lots of advice and suggestions in real data analysis work. I appreciate that he could be a member in the committee. I also would like to express

my gratitude to the chair of the committee, Professor Qing Pan, for making the process organized. I had worked as the grader to the Prof. Pan and learned a great deal. In addition, I hope to thank another committee member, Professor Fang Jin, for her help and previous time.

I would also like to thank Prof. Tapan K. Nayak for supporting me to organize the data mining club. With his help, my team had the opportunity to participate in the data mining competition and won prizes. The experience in data mining club is one of the most memorable experiences during my studies at GWU.

And, my family. I cannot express my gratitude to my parents for their unconditional love in words. At the same time, I am very grateful to the classmates who have been with me. I will not forget the days of staying up late to review knowledge, helping and encouraging each other. And my lovely roommates, they saved me a lot of cooking time when I was busy studying and working.

Abstract of Dissertation

On Combination of Elliptically Distributed Biomarkers to Improve Diagnostic Accuracy

Diagnostic biomarkers play a critical role in biomedical research such as diagnosis and prediction of diseases, selection of treatments for patients, etc. It has been widely accepted by medical researchers that diagnosis based on one single biomarker may not provide sufficient accuracy. To improve the diagnostic performance, researchers proposed methods to combine multiple biomarkers and the corresponding measurements to help clinicians make better diagnostic judgment. Considerable research results have been derived based on the multivariate normal distribution, especially for the combination based on AUC values. In chapter 2, the properties of the ROC curve function under the multivariate normal distribution are discussed. We find that if the case and control group have different covariance matrices for the binormal model, there would always exist a segment of ROC curve which is lower than the chance line.

In addition, the normality is indeed a hard assumption in diagnostic medicine. It is common in practice that many biomarkers follow distributions that are far from normality, even after popular transformations. To make the combination method more useful in practice, we extend the combination of multiple biomarkers to a broader distribution family – elliptical distributions. It is well known that the likelihood ratio combination is the optimal combination but it's sensitive to the distribution assumption and complicated to calculate. Based on the assumption of elliptical distribution, we study the ROC curve function of likelihood ratio combination in chapter 4. The empirical estimation of elliptical likelihood ratio combination is proposed. Except for the optimal combination, we also extend the best linear combination AUC to the elliptical distribution and derive the linear combination ROC curve function of elliptically distributed biomarkers. We find that the best linear combination only depends on the Mahalanobis distance of elliptical distributions.

Many researchers focused on combinations to improve the resulting AUC value while the AUC still has drawbacks. Therefore, we research the combination of more indices besides AUC in order to improve the diagnosis performance. We develop the linear combination based on projected length (PLC) of ROC curves in the chapter 6. It is concluded that the best linear coefficient of PLC is the same as the combination coefficient of AUC under elliptical distribution assumption.

The proposed methods are applied to two case-control studies. The first example is predicting autism in children. Compared with the empirical best linear combination, the AUC value of elliptical likelihood ratio combination is increased by 7.69% in predicting autism, and the PLC value is improved by 16.78%. The AUC and PLC value of elliptical likelihood ratio combination is increased by 1.85% and 5.15% from normal likelihood ratio combination. Another application is the diagnosis of neural tube defects. We find that our method improve by 11.43% and 23.29% from normal likelihood ratio combined AUC value and empirical best linear combination, respectively.

Table of Contents

Acknowledgments iv

Abstract of Dissertation vi

List of Figures xi

List of Tables xii

1 Introduction 1

1.1 Diagnostic Accuracy of a Biomarker 2

1.2 Combination of Multiple Biomarkers 4

1.3 Literature Review of Combination of Biomarkers 5

1.3.1 Optimal Combination 6

1.3.2 Linear Combination 7

1.4 Motivation and Objectives 8

2 Properties of Normal ROC Curves 10

2.1 Slope of An ROC Curve 11

2.2 ROC Curve Intersection Point 12

3 Introduction of Elliptical Distribution 14

3.1 Definition and Density of Elliptical Distribution 14

3.2 Properties of Elliptical Distribution 16

3.2.1 Moments 17

3.2.2 Linear Transformation 17

3.2.3 Summation Stability 17

3.3 Estimation of Elliptical Distribution 18

3.3.1 Mean and Scalar Matrix Estimation 18

3.3.2 Density Generator Estimation 19

4 The Combination of Biomarkers based on Elliptical Distribution 21

4.1 Optimal Combination of Biomarkers with Elliptical Distribution 21

4.2 Simulation 25

4.2.1 Multivariate Normal Distribution 25

4.2.2 Multivariate T Distribution 27

4.2.3 Multivariate Laplace Distribution 28

4.3 Linear Combination of Biomarkers with Elliptical Distribution 29

5 Maximum Linear AUC Estimation under Elliptical Distribution 32

5.1 Empirical Estimation 32

5.2 Optimal Linear AUC of Elliptical Distribution 33

5.3 Numerical Results for Combination of Elliptical Distribution 36

5.3.1 AUC for Different Distributions 36

5.3.2 Simulation for the Same Scalar Matrix 40

5.3.3 Simulation for Different Scalar Matrices 45

6 Combination of Projected Length 50

6.1 Definition of Projected Length 51

6.2 Optimal Linear Combination of PLC Under Normal Distribution 52

6.3 Optimal Linear Combination of PLC Under Elliptical Distribution 54

6.4 Simulation 58

6.4.1 Multivariate Normal Distribution 59

6.4.2 Multivariate T Distribution 60

6.4.3 Multivariate Laplace Distribution 60

7 Application 62

7.1 Biomarkers for Autism/Autistic Disorder 62

7.2 Biomarkers for Neural Tube Defects (NTDs) 65

8 Discussion and Future Research **69**

8.1 Summary of Results 69

8.2 Future Research 70

8.2.1 Alternative Index - Youden's index 70

8.2.2 Comparing Two Groups of Biomarkers 71

Bibliography **76**

List of Figures

1.1 ROC Curve 3

2.1 ROC Curve of Normal Biomarkers with Different Variance Ratios 12

4.1 Multivariate Normal Distribution 26

4.2 Multivariate T Distribution 27

4.3 Multivariate Laplace Distribution 29

5.1 Multivariate T Distribution with Same Scalar Matrix 41

5.2 Multivariate Laplace Distribution with Same Scalar Matrix 43

5.3 Elliptical Distribution with Same Scalar Matrix 44

5.4 Multivariate T Distribution with Different Scalar Matrices 46

5.5 Multivariate Laplace Distribution with Different Scalar Matrices 47

5.6 Elliptical Distribution with Different Scalar Matrices 48

6.1 An Example of Drawbacks of AUC 50

6.2 Example: Projected Length of the ROC Curve 51

6.3 Projected Length of the Normal Distribution ROC Curve 59

6.4 Projected Length of the Multivariate T Distribution ROC Curve 60

6.5 Projected Length of the Multivariate Laplace Distribution ROC Curve . . . 61

7.1 Combinations of Autism Data and Corresponding AUC Values 64

7.2 Smoothed Combinations of Autism Data and Corresponding AUC and PLC Values 65

7.3 Combinations of NTDs Data and Corresponding AUC Values 66

7.4 Smoothed Combinations of NTDs Data and Corresponding AUC and PLC Values 67

8.1 An Example of Drawbacks of PLC 70

List of Tables

5.1 Combined AUC for Different Distributions with the Same Location Vector and Scalar Matrix 37

5.2 Combined AUC for Different Distributions with the Same Location Vector and Scalar Matrix 38

5.3 Combined AUC for Different Distributions with the Same Location Vector and Scalar Matrix 39

5.4 Combined AUC for Different Distributions with the Same Location Vector and Scalar Matrix 40

6.1 Combined PLC for Different Distributions and Mahalanobis Distances . . . 57

Chapter 1

Introduction

Diagnostic biomarkers are used for diagnosis, prediction of diseases and play a critical role in biomedical research such as clinical trails. In 2001, National Institutes of Health (NIH) Biomarkers Definitions Working Group defined that "biomarkers are characteristics that are objectively measured and evaluated as indicators of normal biological processes, pathogenic processes, or pharmacologic responses to therapeutic intervention" (Biomarkers Definitions Working Group, 2001). Some well-known biomarkers include glycated haemoglobin (HbA1c) for diagnosis of diabetes (World Health Organization, 2011), prostate-specific-antigen testing for cancer (Barry, 2001), to name a few. Since the breakout of the COVID-19 pandemic, scientists are rigorously discovering and developing biomarkers that can be used to screen for infection of the virus. More than forty articles have reported potential biomarkers in diagnosis of COVID-19 such as procalcitonin (PCT), C-reactive protein (CRP) and many others; see, e.g., Kermali et al. (2020); Zhang and Guo (2020). The importance of biomarkers in biomedical is plain to see.

However, most biomarkers' effectiveness in diagnosing diseases is not perfect. New biomarkers for the early detection and prevention of diseases are constantly being developed, and are compared with other biomarkers in terms of their accuracy in diagnosis of

diseases. But it has been widely accepted by medical researchers that diagnosis based on one single biomarker may not provide sufficient accuracy. To improve the diagnostic performance, researchers proposed lots of method to combine multiple biomarkers and the corresponding measurements to help clinicians make better diagnostic judgment.

In order to measure the ability of a biomarker to discriminate or predict disease, diagnostic accuracy indices are usually used. Most accuracy indices are calculated through receiver operating characteristic (ROC) curves, including AUC value, projected length and Youden's index.

1.1 Diagnostic Accuracy of a Biomarker

Consider using a biomarker M which is measured on a continuous scale to diagnose or predict a particular disease. The disease status of a subject is represented by a binary variable:

$$D = \begin{cases} 1 \text{ if the subject is diseased;} \\ 0 \text{ if the subject is not diseased.} \end{cases}$$

Assuming that subjects with higher biomarker levels are more likely to be diseased. If not, we could do some transformation. Corresponding to a cut point $c \in \mathbb{R}$, the sensitivity (also called the true positive rate) of this biomarker is the proportion that a diseased person is correctly identified, and the specificity (also called the true positive rate) is the proportion of a non-diseased person that is correctly identified, that is,

$$\begin{cases} s(c) = P(M > c|D = 1) \\ p(c) = P(M < c|D = 0) \end{cases} \tag{1.1.1}$$

where $s(c)$ is the sensitivity (TPR) and $p(c)$ is the specificity (TPR).

The ROC curve is plotted with the true positive rate (sensitivity) against the false

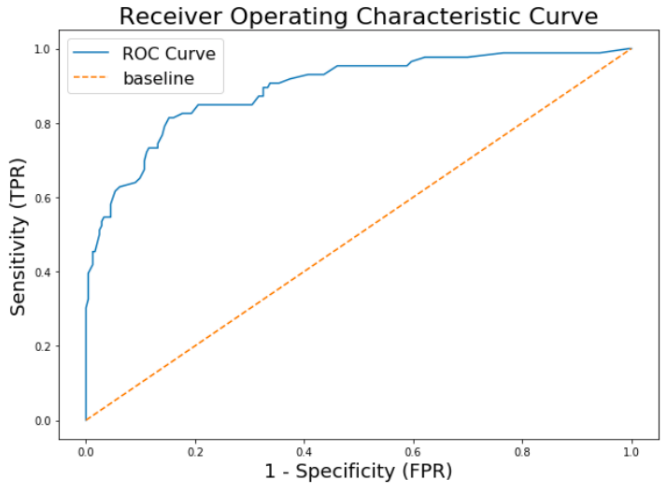

Figure 1.1: ROC Curve

positive rate (1-specificity) over all possible values of $c \in \mathbb{R}$. Higher ROC curves usually represent better diagnostic performance. Most measurement indices could be obtained from ROC curves. For example, the area under the ROC curve, denoted by AUC, is a summary measure of the diagnostic accuracy of a biomarker. A better biomarker should have a larger AUC value.

Krzanowski and Hand (2009) proposed the ROC curve function and corresponding estimation methods, both parametrically and non-parametrically. Suppose the level of one biomarker is denoted by X and Y with cumulative function F_X and F_Y for case and control group, respectively. The ROC curve is the function $s(t(c))$ plotted over all values of $c \in \mathbb{R}$ and it can be written as:

$$s(t) = 1 - F_x\left[F_y^{-1}(1-t)\right] \quad (0 \leq t \leq 1) \tag{1.1.2}$$

where $t(c)$ is the false positive rate i.e. $t(c) = 1 - p(c)$ and the AUC could be obtained by integration:

$$AUC = \int_0^1 s(t)dt \tag{1.1.3}$$

Krzanowski and Hand (2009) also performed the non-parametrical estimation of ROC

curves. Suppose there are n observations in case group denoted as $(X_1, X_2, \ldots, X_n)$ and m observations in control group denoted as $(Y_1, Y_2, \ldots, Y_m)$. Let n_c be the number of subjects in the case group and biomarker levels are greater than the cut point c, m_c be the number of subjects in the control group and biomarker levels are greater than the cut point c.

$$n_{(c)} = \sum_{i=1}^{n} I(X_i > c), \quad m_{(c)} = \sum_{j=1}^{m} I(Y_j > c)$$

where I is the indicator function. Then the empirical estimators of $s(c)$ and $p(c)$ in equation 1.1.1 is given by:

$$\begin{cases} \hat{s}(c) = \frac{n_{(c)}}{n} \\ \hat{p}(c) = \frac{m_{(c)}}{m} \end{cases} \tag{1.1.4}$$

The empirical ROC curve is obtained by plotting $\hat{s}(c)$ against $1 - \hat{p}(c)$ over all values of $c \in \mathbb{R}$.

1.2 Combination of Multiple Biomarkers

Suppose there are K biomarkers available for the diagnosis of a specific disease. We denote the levels of K biomarkers by $\mathbf{M} = (M_1, M_2, \ldots, M_K)$. Each biomarker is measured on a continuous scale. In order to improve the diagnostic performance, researchers are interested in combining these biomarkers.

Define a function $g : \mathbb{R}^K \to \mathbb{R}$ which maps the K-dimensional space onto one dimension. We call $g(\mathbf{M})$ is a combination of the K biomarkers. The sensitivity and specificity of the combination could be defined as:

$$\begin{cases} s_g(c) = P(g(\mathbf{M}) > c | D = 1); \\ p_g(c) = P(g(\mathbf{M}) < c | D = 0). \end{cases}$$

The ROC curve of the combination $g(\mathbf{M})$ is plotted with the $s_g(c)$ against the $1 - p_g(c)$ over all possible values of $c \in \mathbb{R}$. Many researchers are studying on deriving better combining functions not only in order to improve the diagnostic accuracy, but also make the methods more useful in real data analysis.

1.3 Literature Review of Combination of Biomarkers

Many researchers have studied the problem of combining biomarkers to improve diagnostic accuracy. Su and Liu (1993) proposed the optimal linear combinations that yield the largest AUC values, if the levels of the biomarkers in the control and case groups both follow multivariate normal distributions. They also proved that if case and control populations follow multivariate normal distributions with a same covariance matrix, then the linear combination is equivalent to the optimal combination. In 2000, Pepe and Thompson discussed an empirical estimation for the solutions of the optimal linear combinations. The empirical estimation is non-parametric and thus is robust against misspecified distributions. But, the empirical estimation is limited with the sample size. Liu et al. (2005) derived alternative linear combinations that have higher sensitivity over a range of high (or low) specificity. Furthermore, Pepe et al. (2006) pointed out that the likelihood ratio provides the globally optimal combination based on the popular Neyman-Pearson Lemma. But, working with the likelihood ratio to obtain the ROC curve and its area is very complicated. In addition, the likelihood ratio combination is extremely sensitive to the assumption on the joint distribution of the biomarkers' levels. Liu et al. (2011) proposed another non-parametric method, the min-max combination of biomarkers, to improve the diagnostic accuracy.

1.3.1 Optimal Combination

Pepe et al. (2006) proved that the likelihood ratio combination is the optimal combination that yields the largest sensitivity for every given specificity, that is, the higest ROC curve. Let the level of the biomarker be denoted by X and Y with density function f_X and f_Y for case and control group, respectively. The likelihood ratio combination is defined by

$$R(\mathbf{M}) = \frac{f_X(M_1, \ldots, M_K)}{f_Y(M_1, \ldots, M_K)}. \tag{1.3.1}$$

A subject will be diagnosed as diseased if $R(\mathbf{M})$ is larger than a cut point. $R(\mathbf{M})$ is the optimal combination. The proof is based on the Neyman-Pearson Lemma (Neyman and Pearson, 1993). A heuristic argument is given below.

Consider testing the null hypothesis $H_0 : \mathbf{M} \sim f_X$ against the alternative hypothesis $H_1 : \mathbf{M} \sim f_Y$ at a given level of significance. According to the Neyman-Pearson Lemma, $R(\mathbf{M})$ is the uniformly most powerful test (Neyman and Pearson, 1993). Now suppose a subject will be diagnosed as diseased if $R(\mathbf{M}) > c$ where c is a constant. Then the false positive rate is given by $P(R(\mathbf{M}) > c \mid D = 0)$, which is the type I error of the Neyman-Pearson test, and its true positive rate (sensitivity) is given by $P(R(\mathbf{M}) > c \mid D = 1)$, which is the power of the Neyman-Pearson test. It is concluded that, at any given false positive rate, the likelihood ratio has higher sensitivity than any other functions of $\mathbf{M}$ (Pepe et al., 2006).

In fact, given specificity, as long as the accuracy index is an increasing function of sensitivity, then the likelihood ratio would be the best combination of biomarkers. The likelihood ratio, though optimal, is sensitive to distributional assumptions which are often difficult to justify. In addition, The likelihood ratio combination is an exactly parametric method, it's optimal only if the (conditional) joint distribution function of the biomarkers given the disease status is correctly specified.

1.3.2 Linear Combination

An alternative method is the linear combination, which is more robust and easier to interpret and calculate. Let $l(\mathbf{M}) = \lambda^{\mathrm{T}}\mathbf{M}$, where the $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_K)^{\mathrm{T}}$ is the combination coefficient. A subject will be diagnosed as diseased if $l(\mathbf{M}) > c$ where c is a cut point value. If a linear combination ROC curve is higher than ROC curves of any other linear combinations, we call this linear combination is uniformly linear optimal.

McIntosh and Pepe (2002) pointed out that if the likelihood ratio $R(\mathbf{M})$ is a monotone function of $l(\mathbf{M})$, then the uniformly optimal linear combinations exist and it is $l(\mathbf{M})$. In this case, $l(\mathbf{M})$ is indeed the optimal combination among all combinations including nonlinear ones. However, uniform optimality is the a strict criteria since it requires the ROC curve be higher than any other curves at any specificity value. To solve this we could relax the comparison criteria and consider combinations that have special forms. In this respect, linear combinations that maximize the AUC becomes a most popular method.

According to Bamber (1975), the AUC of the ith biomarker can be represented by the probability that ranks a randomly chosen case instance higher than a randomly chosen control one, i.e.,

$$AUC = P(Y_i < X_i). \tag{1.3.2}$$

It could be calculated by the Wilcoxon rank-sum test statistic when we have observed data (Hanley and McNeil, 1982).

Given a linear combination coefficient $\lambda \in \mathbb{R}^K$, the AUC of the linear combination is given by

$$AUC(\lambda) = P\left(\lambda^{\mathrm{T}}Y < \lambda^{\mathrm{T}}X\right). \tag{1.3.3}$$

The best linear combination is the one that maximizes the AUC among all possible

values of λ. That is, the best $\hat{\lambda}$ should be:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \left(P\left(\lambda^{\mathrm{T}} Y < \lambda^{\mathrm{T}} X \right) \right). \tag{1.3.4}$$

In particular, if $X = (X_1, \ldots, X_k)$ follows a multivariate normal distribution with mean vector μ_X and covariance matrix Σ_x, $Y = (Y_1, \ldots, Y_k)$ follows a multivariate normal distribution with mean vector μ_Y and covariance matrix Σ_y, then the optimal linear combination coefficient is given by

$$\hat{\lambda} = \left(\Sigma_x + \Sigma_y \right)^{-1} \left(\mu_x - \mu_y \right), \tag{1.3.5}$$

and the corresponding AUC of optimal linear combination is:

$$AUC = \Phi\left(\left[\left(\mu_x - \mu_y \right)^{\mathrm{T}} \left(\Sigma_x + \Sigma_y \right)^{-1} \left(\mu_x - \mu_y \right) \right]^{\frac{1}{2}} \right) \tag{1.3.6}$$

where Φ is the cumulative function of standard normal distribution (Su and Liu, 1993). Usually we assume that $\mu_x \geq \mu_y$ since subjects with higher biomarker levels are more likely to be diseased. Su and Liu (1993) also proved that under the normality assumption, the best linear combination is equivalent to the likelihood ratio combination.

1.4 Motivation and Objectives

Considerable amount of research results have been derived based on normal distribution, especially for the combination of the AUC values. We will discuss some properties of the ROC curve function under the multivariate normal distribution. However, the normality is indeed a hard assumption in diagnostic medicine. To make the combination method more useful in practice, the main aim of my project is promoting the combination of multiple biomarkers to a broader distribution family – elliptical distribution. We will develop the combination of biomarkers based on the elliptical distribution.

It is well known that the likelihood ratio combination is the optimal combination but it's sensitive to the distribution assumption and complicated to calculate. Based on the assumption of elliptical distribution, we will study the ROC curve function of likelihood ratio combination. The empirical estimation of elliptical likelihood ratio combination is one of the most important parts of our researches. Except for the optimal combination, we also expect to extend the best linear combined AUC to elliptical distribution and derive the linear combined ROC curve of elliptically distributed biomarkers.

Many researchers focused on improving the combined AUC value while the AUC still has drawbacks. Therefore, another objective of our project is researching the combination of more indices besides AUC in order to improve the diagnosis performance. We will consider the combination of projected length of ROC curves as another measurement index.

Chapter 2

Properties of Normal ROC Curves

Consider a binormal model: a biomarker which follows normal distribution with mean μ_x and variance σ_x^2 in case group, denoted by X, and normal distribution with mean μ_y and variance σ_y^2 in control group, denoted by Y. Assume $\mu_x \geq \mu_y$. The sensitivity s could be expressed as a function of the false positive rate t:

$$s(t) = \Phi(a + b\Phi^{-1}(1-t)) \tag{2.0.1}$$

where $a = (\mu_x - \mu_y)/\sigma_x$ and $b = -\sigma_y/\sigma_x$ (Krzanowski and Hand, 2009). The equation 2.0.1 is the ROC curve function. Higher ROC curve means better diagnostic performance. In this chapter we will discuss some features of the ROC curve under normal distribution.

When there are multiple biomarkers that follow multivariate normal distribution, we just need to replace the μ_x, μ_y, σ_x^2 and σ_y^2 with the corresponding mean vector and covariance matrix under linear combination in the equation 2.0.1.

2.1 Slope of An ROC Curve

In this section, we study the trend of a normal ROC curve. We first examine whether the ROC curve would always be higher than the diagonal line, especially in the neighborhood of a boundary, so that the biomarker performs better than the chance line (flipping a coin). Let $s(t) = \Phi\left(\frac{\mu_x-\mu_y}{\sigma_x} - \frac{\sigma_y}{\sigma_x}\Phi^{-1}(1-t)\right)$, which is the ROC Curve function.

Since $\Phi(\Phi^{-1}(x)) = x$, take derivative with respect to x on both sides, we will have $\phi(\Phi^{-1}(x))(\Phi^{-1}(x))' = 1$, which implies $(\Phi^{-1}(x))' = \frac{1}{\phi(\Phi^{-1}(x))}$. It follows that

$$
\begin{aligned}
s'(t) &= \phi\left(\frac{\mu_x-\mu_y}{\sigma_x} - \frac{\sigma_y}{\sigma_x}\Phi^{-1}(1-t)\right)\frac{\sigma_y}{\sigma_x}(\Phi^{-1}(1-t))' \\
&= \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{\mu_x-\mu_y}{\sigma_x} - \frac{\sigma_y}{\sigma_x}\Phi^{-1}(1-t)\right)^2\right\}\frac{\sigma_y}{\sigma_x}\sqrt{2\pi}\exp\left\{\frac{1}{2}\left(\Phi^{-1}(1-t)\right)^2\right\} \\
&= \frac{\sigma_y}{\sigma_x}\exp\left\{\frac{1}{2}\left[\left(\Phi^{-1}(1-t)\right)^2 - \left(\frac{\mu_x-\mu_y}{\sigma_x} - \frac{\sigma_y}{\sigma_x}\Phi^{-1}(1-t)\right)^2\right]\right\} \\
&= \frac{\sigma_y}{\sigma_x}\exp\left\{\frac{1}{2}\left[\left(1-\frac{\sigma_y^2}{\sigma_x^2}\right)\left(\Phi^{-1}(1-t)\right)^2 + 2\frac{(\mu_x-\mu_y)\sigma_y}{\sigma_x^2}\Phi^{-1}(1-t) - \left(\frac{\mu_x-\mu_y}{\sigma_x}\right)^2\right]\right\} \\
&= \frac{\sigma_y}{\sigma_x}\exp\left\{\frac{1}{2}\left(1-\frac{\sigma_y^2}{\sigma_x^2}\right)\left[\Phi^{-1}(1-t)+C_1\right]^2 + C_2\right\},\ \sigma_x^2 \neq \sigma_y^2
\end{aligned}
$$

where $C_1 = \frac{(\mu_x-\mu_y)\sigma_y}{\sigma_x^2-\sigma_y^2}$ and $C_2 = -\frac{(\mu_x-\mu_y)^2}{2(\sigma_x^2-\sigma_y^2)}$.

If $\left(1-\frac{\sigma_y^2}{\sigma_x^2}\right) > 0$ (meaning $\sigma_x^2 > \sigma_y^2$), then $s'(t) \to +\infty$ as $t \to 0^+$ or $t \to 1^-$. The ROC curve increases sharply in the neighborhood of boundaries.

If $\left(1-\frac{\sigma_y^2}{\sigma_x^2}\right) < 0$ (meaning $\sigma_x^2 < \sigma_y^2$), then $s'(t) \to 0$ as $t \to 0^+$ or $t \to 1^-$. The ROC curve increases slowly in the neighborhood of boundaries.

If $\left(1-\frac{\sigma_y^2}{\sigma_x^2}\right) = 0$ (meaning $\sigma_x^2 = \sigma_y^2$), then $s'(t) = \exp\left\{\frac{(\mu_x-\mu_y)}{\sigma_x}\left(\Phi^{-1}(1-t) - \frac{(\mu_x-\mu_y)}{2\sigma_x}\right)\right\}$, $S'(t) \to +\infty$ as $t \to 0^+$ and $S'(t) \to 0$ as $t \to 1^-$. The ROC curve increases sharply in the beginning then it increases at a slow rate.

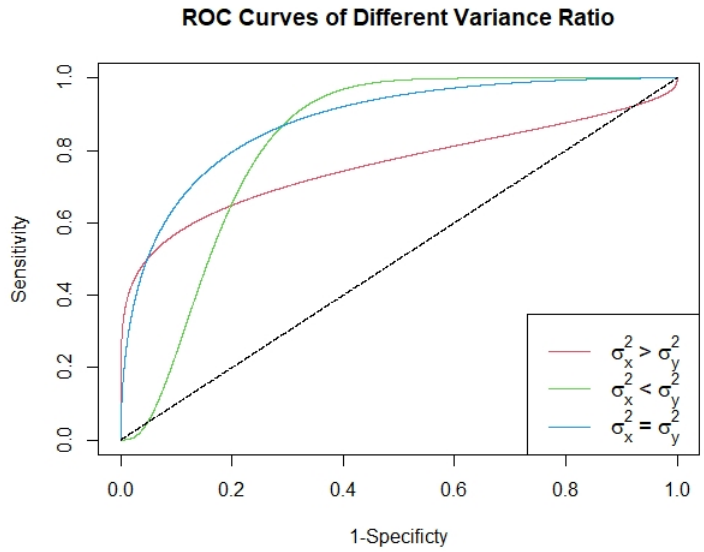

Figure 2.1: ROC Curve of Normal Biomarkers with Different Variance Ratios

Figure 2.1 shows curves under these three conditions. From the figure 2.1 we can see that when the biomarker in case and control group have unequal variances, the ROC curve would intersect with the diagonal line. Therefore, we present next the conditions under which the intersection points exist.

2.2 ROC Curve Intersection Point

Theorem 2.2.1. *Suppose a biomarker follows a normal distribution with mean* μ_x *and variance* σ_x^2 *in case group, denoted by* X*, and a normal distribution with mean* μ_y *and variance* σ_y^2 *in control group, denoted by* Y*.* $\mu_x \geq \mu_y$*. If* $\sigma_x \neq \sigma_y$*, then the ROC curve of this biomarker would always intersect with the diagonal line at the false positive rate* $t = \Phi\left(\frac{\mu_x - \mu_y}{\sigma_x - \sigma_y}\right)$.

Proof. To find the intersection of a normal ROC curve and the diagonal line, when $\sigma_x \neq \sigma_y$,

let $\Phi\left(\frac{\mu_x-\mu_y}{\sigma_x}-\frac{\sigma_y}{\sigma_x}\Phi^{-1}(1-t)\right)=t$, then we have:

$$\begin{aligned}\frac{\mu_x-\mu_y}{\sigma_x}-\frac{\sigma_y}{\sigma_x}\Phi^{-1}(1-t)&=\Phi^{-1}(t)\\ \Rightarrow \mu_x-\mu_y&=\sigma_y\Phi^{-1}(1-t)+\sigma_x\Phi^{-1}(t).\end{aligned}$$

Since $\Phi^{-1}(t)=-\Phi^{-1}(1-t)$, then:

$$\begin{aligned}\mu_x-\mu_y&=(\sigma_x-\sigma_y)\Phi^{-1}(t)\\ \Rightarrow t&=\Phi\left(\frac{\mu_x-\mu_y}{\sigma_x-\sigma_y}\right),\ \sigma_x\neq\sigma_y.\end{aligned}$$

Therefore, if $\sigma_x\neq\sigma_y$, then the ROC curve of a normally distributed biomarker intersects with the diagonal line. □

In conclusion, the ROC curve of a normally distributed biomarker would be higher than the diagonal line at any specificity point if and only if $\sigma_x=\sigma_y$. In other words, if the case and control group have different variances for the binormal model, there would always exist a segment of ROC curve which is lower than the chance line.

Chapter 3

Introduction of Elliptical Distribution

Although considerable amount of research results have been derived based on normal distribution, the normality is indeed a strict assumption. In clinical trails, sometimes the data is distributed with heavy tails, such as the multivariate t distribution, or light tails. In this chapter, we derive the combination of biomarkers based on elliptical distribution which is a broader distribution family. In this chapter, we will introduce the definition of elliptical distribution and some properties of elliptical distribution which we used in the chapter 4, 5 and 6.

3.1 Definition and Density of Elliptical Distribution

Definition 3.1.1. (Frahm, 2004) Let X be a d-dimensional random vector. X is said to be 'elliptically distributed', if and only if there exist a vector $\mu \in \mathbb{R}^d$, a positive semidefinite matrix $\Sigma \in \mathbb{R}^{d\times d}$, and a function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that the characteristic function $t \mapsto \varphi_{X-\mu}(t)$ of $X-\mu$ corresponds to $t \mapsto \phi\left(t^{\mathrm{T}}\Sigma t\right), t \in \mathbb{R}^d$.

An elliptically distributed random variable is denoted by $X \sim \mathcal{E}_d(\mu, \Sigma, \phi)$. Some examples of elliptical distributions include multivariate normal distribution, multivariate t-

distribution, symmetric multivariate stable distribution and so on. Following theorem introduces the density function of elliptical distribution.

Theorem 3.1.1. *(Fang et al., 1990) Let $X \sim \mathcal{E}_d(\mu, \Sigma, \phi)$ where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$, is positive definite. Then X can be represented stochastically by $X \overset{d}{=} \mu + \mathcal{R}\Lambda U^{(d)}$ with $\Lambda\Lambda^{\mathrm{T}} = \Sigma$, where $U^{(d)}$ is uniformly distributed on the unit hypersphere with $d-1$ topological dimensions, $\mathcal{R}$ is a non-negative random variable stochastically independent of $U^{(d)}$. Then the p.d.f of X is given by*

$$f_X(x) = \sqrt{\det(\Sigma^{-1})} \cdot g_{\mathcal{R}}\left((x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right), \quad x \neq \mu \tag{3.1.1}$$

where

$$g_{\mathcal{R}}(t) := \frac{\Gamma\left(\frac{d}{2}\right)}{2\pi^{d/2}} \cdot \sqrt{t}^{-(d-1)} \cdot f_{\mathcal{R}}(\sqrt{t}), \quad t > 0$$

and $f_{\mathcal{R}}$ is the p.d.f. of $\mathcal{R}$.

Or we can write:

$$f_X(x) = \left(\frac{1}{|\Sigma|}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{d}{2}\right)}{2\pi^{d/2}} \cdot \left((x-\mu_x)^{\mathrm{T}}\Sigma^{-1}(x-\mu_x)\right)^{-\frac{d-1}{2}} f_R\left(\sqrt{(x-\mu_x)^{\mathrm{T}}\Sigma^{-1}(x-\mu_x)}\right) \tag{3.1.2}$$

$\mathcal{R}$ is called the generating variate, $g_{\mathcal{R}}$ is the density generator and Σ is the covariance generating matrix or scalar matrix. Elliptical distribution family is broadly used in statistics studies such as robust statistics, generalized multivariate analysis and so on. Many studies of the normal distribution also have a good performance through elliptical distribution. Therefore, it is feasible to extend the linear combination from the normal distribution to the elliptical family. In this paper, we focus on the combination of biomarkers that follow an elliptical distribution.

From theorem 3.1.1 we can see that an elliptical distributed random variable is mainly determined by three items: the location vector μ, the scalar matrix Σ and density generator

$g_{\mathcal{R}}$. The family of density generators is divided into two classes: monotone and non-monotone.

Definition 3.1.2. (Chia and Nakano, 2008) We say that X is elliptical unimodal distributed if there exist a vector $\mu \in \mathbb{R}^d$, a positive semidefinite symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$ and a nonnegative random variable $\mathcal{R}$ with non-increasing density such that X can be represented stochastically by $X \stackrel{d}{=} \mu + \mathcal{R}\Lambda U^{(d)}$ with $\Lambda\Lambda^{\mathrm{T}} = \Sigma$. $U^{(d)}$ is uniformly distributed on the unit hypersphere with $d-1$ topological dimensions and independent of $\mathcal{R}$.

In other words, the elliptical vector X is unimodal if and only if it's density generator $g_{\mathcal{R}}$ is monotonically non-increasing. Most elliptical random vectors are unimodal, including multivariate normal distribution, multivariate T-distribution, multivariate Laplace distribution and many others (Bhattacharyya and Bickel, 2013). We will discuss the combination of biomarkers which are elliptical unimodal in chapter 4.1 and chapter 6.

3.2 Properties of Elliptical Distribution

The main reason that we choose elliptical distribution is because it has a variety of properties. It includes all multivariate symmetric distribution, and it's summation stable (which means the sum of independent elliptically distributed random vectors with the same covariance generating matrix is elliptical). It also has the property of 'infinite divisibility' (it can be expressed as the probability distribution of the sum of an arbitrary number of independent and identically distributed random variables) and self-decomposability. We will introduce some properties and estimation algorithms which are useful in the combination process.

3.2.1 Moments

From theorem 3.1.1, it's obvious that the covariance matrix of a d-dimensional elliptical random vector X is related to it's generating variate. But the scalar matrix generally does not equal to the covariance matrix.

Theorem 3.2.1. *(Frahm, 2004) Suppose X is a d-dimensional elliptical random vector which is stochastically represented by $X \stackrel{d}{=} \mu + \mathcal{R}\Lambda U^{(d)}$, then $E(X) = \mu$ and $\mathrm{Var}(X) = \frac{E(\mathcal{R}^2)}{d}\Sigma$ where $\Sigma = \Lambda\Lambda^{\mathrm{T}}$.*

The normal distribution is a special case whose covariance matrix is equivalent to the scalar matrix. However, we can always find an expression such that $\mathrm{Var}(X) = \Sigma$ by multiplying $\mathcal{R}$ with $\sqrt{\frac{d}{E(\mathcal{R}^2)}}$ (Hult and Lindskog, 2002).

3.2.2 Linear Transformation

Hult and Lindskog (2002) pointed out that any linear transformation of an elliptical distributed vector is still elliptical.

Theorem 3.2.2. *Suppose $X \sim \mathcal{E}_d(\mu, \Sigma, \phi)$ is a d-dimensional elliptical random vector which is stochastically represented by $X \stackrel{d}{=} \mu + \mathcal{R}\Lambda U^{(d)}$. Let $Y = a + AX$ where $a \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k\times d}$, then Y is a k-dimensional elliptical random vector with mean vector $a + A\mu$ and scalar matrix $A\Sigma A'$*

Note that the generating variate of a linear transformed random vector doesn't change which means Y has the 'same type' distribution of X.

3.2.3 Summation Stability

Theorem 3.2.3. *(Hult and Lindskog, 2002) Let X_1 and X_2 be two d-dimensional elliptically distributed random vectors with the same scalar matrix. Assume X_1 and X_2 could be stochastic represented by $X_1 \stackrel{d}{=} \mu_1 + \mathcal{R}_1\Lambda U_1^{(k)}$ and $X_2 \stackrel{d}{=} \mu_2 + \mathcal{R}_2\Lambda U_2^{(k)}$ respectively. Here*

$(\mathcal{R}_1, \mathcal{R}_2), U_1^{(k)}$ and $U_2^{(k)}$ are mutually independent but $\mathcal{R}_1$ and $\mathcal{R}_2$ may depend on each other. Then $X_1 + X_2$ is also elliptically distributed with location vector $\mu_1 + \mu_2$ and dispersion matrix $\Sigma = \Lambda\Lambda'$.

Theorem 3.2.3 states that the sum of a series of correlated elliptical vectors are still elliptical distribution if they have a same scalar matrix. Combining with the theorem 3.2.2, we can say that any linear combination of elliptical distributed vectors whose scalar matrices are equivalent are still elliptical.

3.3 Estimation of Elliptical Distribution

An elliptical distributed random variable is mainly determined by three parameters: the location vector μ, the scalar matrix Σ and density generator $g_{\mathcal{R}}$. In this section we will introduce the algorithms which could be used to estimate the distribution of elliptical vectors.

3.3.1 Mean and Scalar Matrix Estimation

Let $X \sim \mathcal{E}_d(\mu, \Sigma, \phi)$, in order to identify the location and dispersion of X, it's sufficient to find the consistent estimator of μ and Σ.

At first, assume that the location vector μ is known. Tyler (1983) proposed a fixed point solution of the maximum likelihood estimator of scalar matrix Σ. To get an unique solution of the maximum likelihood estimation, we could assume a constraint $\hat{\Sigma}_{11} = 1$. Alternative constraints are given by fixing the trace or the determinant of $\hat{\Sigma}$. Suppose there are n independent observations $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} \mathcal{E}_d(\mu, \Sigma, \phi)$ where μ is known. Following iterative algorithm is effective for finding the maximum likelihood estimator of scalar matrix Σ.

1. Choose a initial value $\hat{\Sigma}^{(0)} = I_d$ or by taking the sample covariance matrix;

2. For $i = 0,\ 1,\ 2,\ \ldots,\ N$, $\hat{\Sigma}^{(i+1)} = \frac{d}{n}\sum_{j=1}^{n} \frac{(X_j-\mu)(X_j-\mu)'}{(X_j-\mu)'\hat{\Sigma}^{(i)}(X_j-\mu)}$

3. Do the normalization $\hat{\Sigma} = \hat{\Sigma}^{(N)}/\hat{\Sigma}^{(N)}_{11}$

After a finite number of numerical iterations, the obtained estimator is always positive definite. Tyler (1987) proved that if n is large enough ($n > d(d-1)$), a positive definite fixed point almost surely exists and is unique. In practice, n is sufficient large in most cases.

When μ is unknown, we could use sample median $\hat{\mu}$ to estimate it and it's a consistent estimator (Tyler, 1987). Plug the sample median $\hat{\mu}$ into previous algorithm we can have $\hat{\Sigma}^{(i+1)} = \frac{d}{n}\sum_{j=1}^{n} \frac{(X_j-\hat{\mu})(X_j-\hat{\mu})'}{(X_j-\hat{\mu})'\hat{\Sigma}^{(i)}(X_j-\hat{\mu})}$.

3.3.2 Density Generator Estimation

In addition to the location and scalar matrix, another item we are concerned with is the density generator. A lot of research on estimation of the density generator has been conducted. One of the most popular method is the non-parametric maximum likelihood estimate (NPMLE) of density generator g_p for the elliptical unimodal distribution (Bhattacharyya and Bickel, 2013).

Suppose there are n independent observations $X_1,\ X_2,\ \ldots,\ X_n \overset{i.i.d}{\sim} \mathcal{E}_d(\mu, \Sigma, \phi)$ with density function $f_X(x) = \sqrt{\det(\Sigma^{-1})} \cdot g_{\mathcal{R}}\left((x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right)$, where $g_{\mathcal{R}}$ is the monotonically non-increasing density generator. For $i \in (1,\ 2,\ \ldots,\ n)$, let $Z_i = (X_i-\mu)^{\mathrm{T}}\Sigma^{-1}(X_i-\mu)$ which is the normalized observations. Then Z_i has the density function of $f_Z(z) = c_d z^{d/2-1} g_{\mathcal{R}}(z)$, where $c_d = \frac{\pi^{d/2}}{\Gamma(d/2)}$ and $z \in \mathbb{R}^+$.

Let $\left(Z_{(1)}, \ldots, Z_{(n)}\right)$ be the order statistics of $(Z_1,\ Z_2,\ \ldots,\ Z_n)$. Define

$$\hat{g}_i = \frac{d}{2c_d} \min_{s \le i-1} \max_{t \ge i} \frac{F_n\left(Z_{(t)}\right) - F_n\left(Z_{(s)}\right)}{Z_{(t)}^{d/2} - Z_{(s)}^{d/2}} \tag{3.3.1}$$

where F_n is the empirical cumulative distribution function (CDF) of the data $(Z_1, \ldots, Z_n)$.

Then the NPMLE $\hat{g}_n$ is given by

$$\hat{g}_n(z) = \hat{g}_n(z) = \begin{cases} \hat{g}_i, & \text{if } Z_{(i-1)} < z \leq Z_{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{3.3.2}$$

Bhattacharyya and Bickel (2013) provides more asymptotic properties of NPMLE $\hat{g}_n$. When the location vector and scalar matrix are unknown, we can use the estimation algorithm in section 3.3.1 to estimate μ and Σ. And Z_i is replaced by $\hat{Z}_i = (X_i - \hat{\mu})^{\mathrm{T}} \hat{\Sigma}^{-1} (X_i - \hat{\mu})$ in the density generator estimation process.

Chapter 4

The Combination of Biomarkers based on Elliptical Distribution

In this chapter, we propose the combination of elliptically distributed biomarkers. Considerable amount of research results have been derived based on the multivariate normal distribution. But the normality is indeed a strict assumption in diagnostic medicine. It is common in practice that many biomarkers follow distributions that are far from normality, even after popular transformations. Based on the properties of elliptical distribution in Chapter 3, we will derive the combination of biomarkers and the combined ROC curve function. Furthermore, we will show that our methods could significantly improve the diagnostic performance by simulations.

4.1 Optimal Combination of Biomarkers with Elliptical Distribution

If there are d biomarkers are denoted by $\mathbf{M} = (M_1, \ldots, M_d)^{\mathrm{T}}$. Let the level of the biomarkers be denoted by $X \sim \mathcal{E}_d(\mu_x, \Sigma_x; \phi_x)$ and $Y \sim \mathcal{E}_d(\mu_y, \Sigma_y; \phi_y)$, with density function

f_X and f_Y, for case and control group respectively. Pepe et al. (2006) showed that the likelihood ratio combination is the optimal combination. According to the theorem 3.1.1, X, Y could be stochastically represented by $X \overset{d}{=} \mu_x + \mathcal{R}_x \Lambda_x U_x^{(d)}$ and $Y \overset{d}{=} \mu_y + \mathcal{R}_y \Lambda_y U_y^{(d)}$ respectively, where $\Lambda_x \Lambda_x^{\mathrm{T}} = \Sigma_x$ and $\Lambda_y \Lambda_y^{\mathrm{T}} = \Sigma_y$. Then the likelihood ratio combination could be defined as:

$$
\begin{aligned}
R(\mathbf{M}) &= \frac{f_X\left(M_1, \ldots, M_d\right)}{f_Y\left(M_1, \ldots, M_d\right)} \\
&= \frac{\sqrt{\det\left(\Sigma_x^{-1}\right)} \cdot g_{\mathcal{R}_x}\left((\mathbf{M}-\mu_x)^{\mathrm{T}} \Sigma_x^{-1} (\mathbf{M}-\mu_x)\right)}{\sqrt{\det\left(\Sigma_y^{-1}\right)} \cdot g_{\mathcal{R}_y}\left((\mathbf{M}-\mu_y)^{\mathrm{T}} \Sigma_y^{-1} (\mathbf{M}-\mu_y)\right)} \\
&= \frac{\sqrt{\det\left(\Sigma_x^{-1}\right)} \left((\mathbf{M}-\mu_x)^{\mathrm{T}} \Sigma_x^{-1} (\mathbf{M}-\mu_x)\right)^{-\frac{d-1}{2}} f_{R_x}\left(\sqrt{(\mathbf{M}-\mu_x)^{\mathrm{T}} \Sigma_x^{-1} (\mathbf{M}-\mu_x)}\right)}{\sqrt{\det\left(\Sigma_y^{-1}\right)} \left((\mathbf{M}-\mu_y)^{\mathrm{T}} \Sigma_y^{-1} (\mathbf{M}-\mu_y)\right)^{-\frac{d-1}{2}} f_{R_y}\left(\sqrt{(\mathbf{M}-\mu_y)^{\mathrm{T}} \Sigma_y^{-1} (\mathbf{M}-\mu_y)}\right)}
\end{aligned}
\tag{4.1.1}
$$

A subject will be diagnosed as diseased if the likelihood ratio $R(\mathbf{M})$ is larger than a cut point $c \in \mathbb{R}$. Generally, obtaining likelihood ratio combination is complicated. But under elliptical distribution assumption, the likelihood ratio combination could be simplified to the ratio of density generators since $\sqrt{\frac{\det\left(\Sigma_x^{-1}\right)}{\det\left(\Sigma_y^{-1}\right)}}$ is a positive constant. Therefore, we get the following theorem:

Theorem 4.1.1. *Suppose there are d-dimensional biomarkers which are elliptically distributed denoted as $X \sim \mathcal{E}_d\left(\mu_x, \Sigma_x; \phi_x\right)$ and $Y \sim \mathcal{E}_d\left(\mu_y, \Sigma_y; \phi_y\right)$ in case and control group, respectively. X, Y could be stochastically represented by $X \overset{d}{=} \mu_x + \mathcal{R}_x \Lambda_x U_x^{(d)}$ and $Y \overset{d}{=} \mu_y + \mathcal{R}_y \Lambda_y U_y^{(d)}$ where $\Lambda_x \Lambda_x^{\mathrm{T}} = \Sigma_x$ and $\Lambda_y \Lambda_y^{\mathrm{T}} = \Sigma_y$. Then the likelihood ratio combination could be expressed as*

$$
R(\mathbf{M}) = \frac{g_{\mathcal{R}_x}\left((\mathbf{M}-\mu_x)^{\mathrm{T}} \Sigma_x^{-1} (\mathbf{M}-\mu_x)\right)}{g_{\mathcal{R}_y}\left((\mathbf{M}-\mu_y)^{\mathrm{T}} \Sigma_y^{-1} (\mathbf{M}-\mu_y)\right)} \tag{4.1.2}
$$

where $g_{\mathcal{R}_x}$ and $g_{\mathcal{R}_y}$ are density generators of X and Y.

In other words, the likelihood ratio combination could be simplified to the ratio of density generators of the elliptical distributions. After we have data, the density generators could be estimated by equation 3.3.1 and equation 3.3.2 if the biomarkers are elliptical unimodal. Next we will show examples about the likelihood ratio combination for some popular distributions.

Example 4.1.1. (Multivariate Normal Distribution) Suppose there are d biomarkers following multivariate normal distributions in case and control group: $X \sim MN_d(\mu_x, \Sigma_x)$, $Y \sim MN_d(\mu_y, \Sigma_y)$.

The density generator of d dimensional multivariate normal distribution is $g(t) = \frac{1}{(2\pi)^{d/2}} \cdot \exp\left(-\frac{t}{2}\right)$ (Frahm, 2004). Then the likelihood ratio combination is:

$$\begin{aligned} R(\mathbf{M}) &= \frac{g_{\mathcal{R}_x}\left((\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x)\right)}{g_{\mathcal{R}_y}\left((\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)\right)} \\ &= \exp\left\{-\tfrac{1}{2}\left((\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x) - (\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)\right)\right\} \\ &\propto -\left((\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x) - (\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)\right) \end{aligned}$$

A subject $\mathbf{M}$ will be diagnosed as diseased if $R(\mathbf{M}) > c$, which is equivalent to that value of $\left((\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x) - (\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)\right)$ less than a cut point value. In particular, when $\Sigma_x = \Sigma_y = \Sigma$, the likelihood ratio becomes:

$$\begin{aligned} R(\mathbf{M}) &= \exp\left\{(\mu_x-\mu_y)^{\mathrm{T}}\Sigma^{-1}\mathbf{M} - \tfrac{1}{2}\left(\mu_x^{\mathrm{T}}\Sigma^{-1}\mu_x - \mu_y^{\mathrm{T}}\Sigma^{-1}\mu_y\right)\right\} \\ &\propto (\mu_x-\mu_y)^{\mathrm{T}}\Sigma^{-1}\mathbf{M} \end{aligned}$$

A subject $\mathbf{M}$ will be diagnosed as diseased if $R(\mathbf{M}) > c$, which is equivalent to that $(\mu_x - \mu_y)^{\mathrm{T}}\Sigma^{-1}\mathbf{M}$ is larger than a cut point value. Under this condition, the likelihood ratio combination is a linear combination with linear coefficient $\lambda = \Sigma^{-1}(\mu_x - \mu_y)$ (Su and Liu, 1993).

Example 4.1.2. (Multivariate T Distribution) Suppose there are d biomarkers following

multivariate T distributions in case and control group: $X \sim t_d(\mu_x, \Sigma, v_x)$ and $Y \sim t_d(\mu_y, \Sigma, v_y)$.

From Frahm (2004), the density generator of multivariate t distribution $t_d(\mu, \Sigma, v)$ is $g_{\mathcal{R}}(t) = \frac{\Gamma\left(\frac{d+v}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \cdot \frac{1}{(v\pi)^{d/2}} \cdot \left(1 + \frac{t}{v}\right)^{-\frac{d+v}{2}}$. Then the likelihood ratio combination is:

$$\begin{aligned} R(\mathbf{M}) &= \frac{g_{\mathcal{R}_x}\left((\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x)\right)}{g_{\mathcal{R}_y}\left((\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)\right)} \\ &\propto \left(1 + \frac{(\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x)}{v_x}\right)^{-\frac{d+v_x}{2}} \Big/ \left(1 + \frac{(\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)}{v_y}\right)^{-\frac{d+v_y}{2}} \\ &\propto \frac{\left(v_x + (\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x)\right)^{-\frac{d+v_x}{2}}}{\left(v_y + (\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)\right)^{-\frac{d+v_y}{2}}} \end{aligned}$$

In particular, if $v_x = v_y = v$ i.e. the degree freedoms of case and control group are equal, the likelihood ratio combination is given by:

$$R(\mathbf{M}) = \left(\frac{v + (\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x)}{v + (\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)}\right)^{-\frac{d+v}{2}}$$

Under this condition, A subject $\mathbf{M}$ will be diagnosed as diseased if the value of $\frac{v+(\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x)}{v+(\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)}$ is larger than a cut point c.

Example 4.1.3. (Mixed Distribution) Suppose $X \sim MN_d(\mu_x, \Sigma_x)$ and $Y \sim t_d(\mu_y, \Sigma_y, v)$. Here the distributions in case and control group have different distribution "types".

Then the likelihood ratio combination is:

$$\begin{aligned} R(\mathbf{M}) &= \frac{g_{\mathcal{R}_x}\left((\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x)\right)}{g_{\mathcal{R}_y}\left((\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)\right)} \\ &\propto \exp\left(-\tfrac{1}{2}(\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x)\right)\left(1 + \frac{(\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)}{v}\right)^{\frac{d+v}{2}} \\ &\propto (d+v)\log\left(v + (\mathbf{M}-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(\mathbf{M}-\mu_y)\right) - (\mathbf{M}-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(\mathbf{M}-\mu_x) \end{aligned}$$

4.2 Simulation

Now we will show that the empirical ratio of density generators could be used to estimate the likelihood ratio combination by simulations. Consider the likelihood ratio combination of d biomarkers. Suppose there are n independent observations $(X_1,\ X_2,\ \ldots, X_n) \sim \mathcal{E}_d\left(\mu_x, \Sigma_x; \phi_x\right)$ in case group, and m independent observations $(Y_1,\ Y_2,\ \ldots, Y_m) \sim \mathcal{E}_d\left(\mu_y, \Sigma_y; \phi_y\right)$ in control group. The density generators in both groups are monotonically non-increasing which means these biomarkers are elliptical unimodal distributed. Following are main steps to get the estimation of likelihood ratio combination ROC curve.

1. Let $\hat{g}_x$ and $\hat{g}_y$ be the NPMLE estimators obtained by equation 3.3.1 and equation 3.3.2 of density generators $g_{\mathcal{R}_x}$ and $g_{\mathcal{R}_y}$.
2. For $i \in (1,\ 2,\ \ldots,\ n)$, let $\hat{R}_i(X) = \frac{\hat{g}_x\left((X_i-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(X_i-\mu_x)\right)}{\hat{g}_y\left((X_i-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(X_i-\mu_y)\right)}$. Similarly, let $\hat{R}_j(Y) = \frac{\hat{g}_x\left((Y_j-\mu_x)^{\mathrm{T}}\Sigma_x^{-1}(Y_j-\mu_x)\right)}{\hat{g}_y\left((Y_j-\mu_y)^{\mathrm{T}}\Sigma_y^{-1}(Y_j-\mu_y)\right)}$ for $j \in (1,\ 2,\ \ldots,\ m)$.
3. For a cut point value $c \in \mathbb{R}$, let $n_c = \sum_{i=1}^{n} I(\hat{R}_i(X) > c)$, $m_c = \sum_{j=1}^{m} I(\hat{R}_j(Y) > c)$, where I is the indicator function.
4. Let $\hat{s}(c) = \frac{n_{(t)}}{n}, \hat{p}(c) = \frac{m_{(t)}}{m}$. The estimated ROC curve of likelihood ration combination is obtained by plotting $\hat{s}(c)$ and $1 - \hat{p}(c)$ over all values of $c \in \mathbb{R}$.

4.2.1 Multivariate Normal Distribution

Firstly we begin with the easiest condition: multivariate normal distribution. Multivariate normal distribution is a special case in elliptical distribution family. The scalar matrix of a multivariate normal distribution random variable is equivalent to it's covariance matrix. In addition, when the biomarkers in case and control group have a same covariance matrix, the likelihood ratio combination is the optimal linear combination.

Suppose case group and control group follow the multivariate normal distribution with the expectation μ_x and μ_y, separately. Assume they share the same covariance ma-

trix Σ:

$$X \sim MN_d(\mu_x, \Sigma), \quad Y \sim MN_d(\mu_y, \Sigma)$$

$$\text{Let } \mu_x = \begin{pmatrix} 2 \\ 1.6 \\ 1.3 \end{pmatrix}, \mu_y = \begin{pmatrix} 1.4 \\ 1 \\ 1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.15 & 0.1 \\ 0.15 & 0.26 & 0.02 \\ 0.1 & 0.02 & 1 \end{pmatrix}.$$

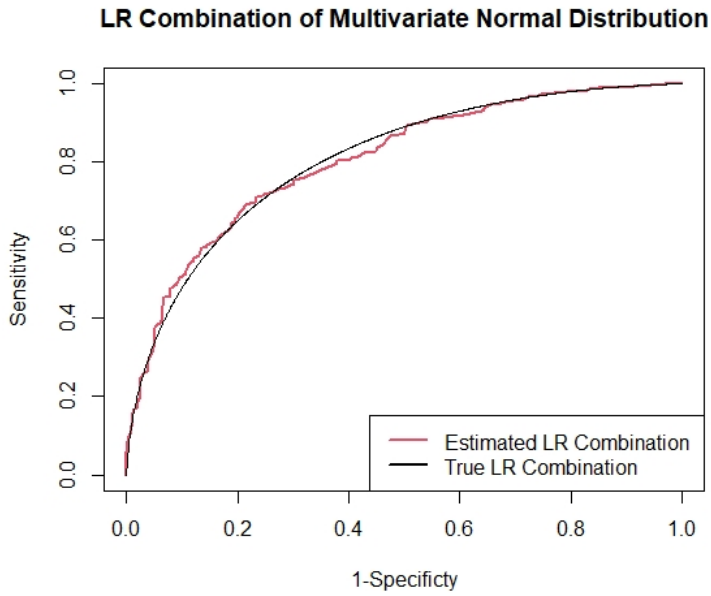

Figure 4.1: Multivariate Normal Distribution

Then the best linear combination coefficient is given by $\hat{\lambda}_{max}\Sigma^{-1}(\mu_x - \mu_y)$ which equals to $\begin{pmatrix} 0.2555 & 2.1425 & 0.2316 \end{pmatrix}^{\mathrm{T}}$ (Su and Liu, 1993). Let $\mu_x^* = \hat{\lambda}_{max}^{\mathrm{T}}\mu_x = 4.24$, $\mu_y^* = \hat{\lambda}_{max}^{\mathrm{T}}\mu_y = 2.7317$ and $\sigma^2 = \hat{\lambda}_{max}^{\mathrm{T}}\Sigma\hat{\lambda}_{max} = 1.5082$.

According to the equation 2.0.1, the likelihood ratio combined ROC curve is: $s(t) = \Phi(\left(\mu_x^* - \mu_y^*\right)/\sigma + \Phi^{-1}(1-t))$ which is shown as the black line in the Figure 4.1. The estimated likelihood ratio combined ROC curve which is obtained by the density generator estimation is shown as the red line in the Figure 4.1. We can see that the estimated curve is approximately around the true curve but with some noise.

4.2.2 Multivariate T Distribution

Multivariate t-distribution is a generalization to random vectors of the Student's t-distribution (Kotz and Nadarajah, 2004). Suppose the case and control group follow the multivariate-t distribution with the expectation μ_x and μ_y, separately. Assume they share the same covariance generating matrix Σ:

$$X \sim T_{v_x}(\mu_x, \Sigma), \quad Y \sim T_{v_y}(\mu_y, \Sigma)$$

Notice the covariance matrices V of case and control group are different when they have different degrees of freedom. To make sure that the covariance martix exists, the degree freedom is required to be larger than 2.

$$V_x = \frac{v_x}{v_x - 2}\Sigma, \quad V_y = \frac{v_y}{v_y - 2}\Sigma$$

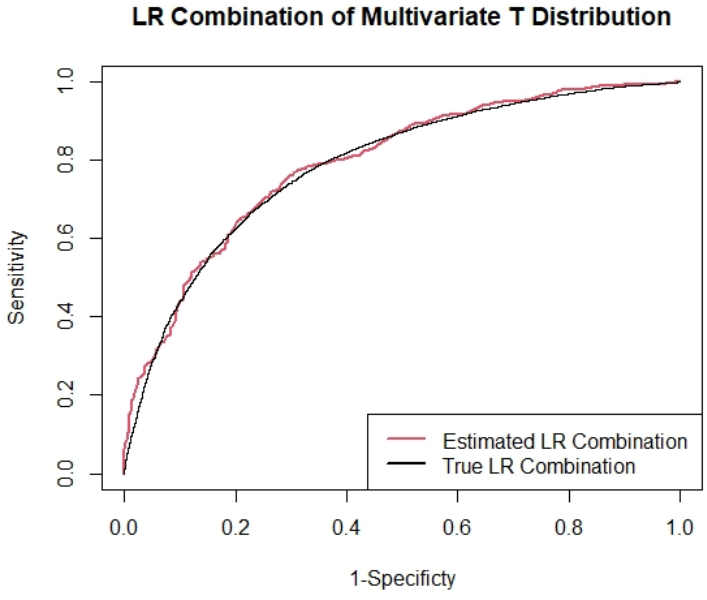

Figure 4.2: Multivariate T Distribution

Let $v_x = 4$, $v_y = 6$, $\mu_x = \begin{pmatrix} 2 \\ 1.6 \\ 1.3 \end{pmatrix}$, $\mu_y = \begin{pmatrix} 1.4 \\ 1 \\ 1 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 0.15 & 0.1 \\ 0.15 & 0.26 & 0.02 \\ 0.1 & 0.02 & 1 \end{pmatrix}$.

Generate 500 observations in each group and estimate the likelihood ratio combination ROC curve. Figure 4.2 shows the simulation result. The black line is the true ROC curve which is smooth, and the red line is the estimated ROC curve obtained by previous discussed approach. We can see that the red line is approximately around the black line.

4.2.3 Multivariate Laplace Distribution

Multivariate Laplace distribution is an extension of the Laplace distribution to vectors (Kotz et al., 2001). Suppose case and control group follow the multivariate Laplace distribution with the expectation μ_x and μ_y, separately. Assume they share the same scalar matrix Σ:

$$X \sim ML(\mu_x, \Sigma), \quad Y \sim ML(\mu_y, \Sigma)$$

where $\mu_x = \begin{pmatrix} 3 \\ 1.5 \\ 1 \end{pmatrix}$, $\mu_y = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 0.3 & 0.1 \\ 0.3 & 2 & 0.02 \\ 0.1 & 0.02 & 1 \end{pmatrix}$.

Generate 500 observations in each group and estimate the likelihood ratio combination ROC curve. Figure 4.3 shows the simulation result. The black line is the true ROC curve and red line is the estimated ROC curve obtained by previous discussed approach. We can see that the red line is approximately around the black line.

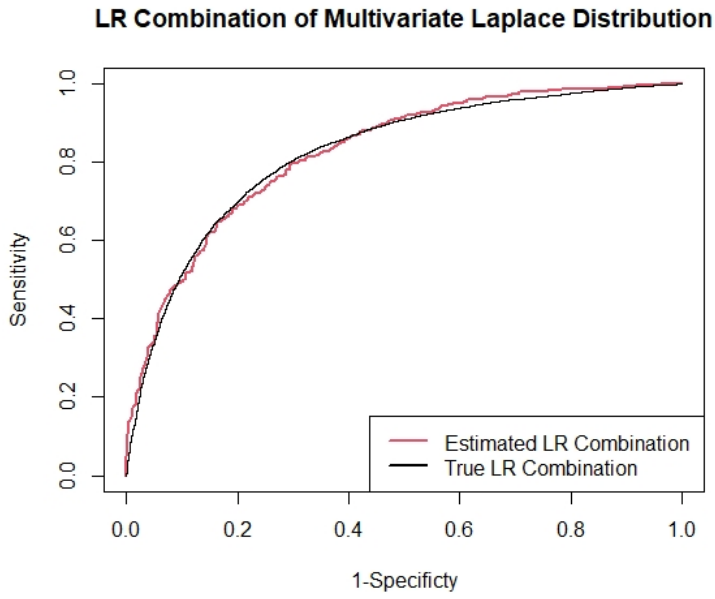

Figure 4.3: Multivariate Laplace Distribution

4.3 Linear Combination of Biomarkers with Elliptical Distribution

Consider the linear combination of d biomarkers. Let $l(\mathbf{M}) = \lambda^{\mathrm{T}}\mathbf{M}$ where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)^{\mathrm{T}}$ is the combination coefficients. A subject will be diagnosed as diseased if $l(\mathbf{M})$ is larger than a cut point. In 1993, Su and Liu derived the linear combined ROC curve under multivariate normal distribution. If $X \sim N_d(\mu_x, \Sigma_x)$ and $Y \sim N_d(\mu_y, \Sigma_y)$ are distribution of case and control group respectively, given any specificity p, the corresponding sensitivity s could be represented by $S(p, \lambda) = \Phi\left(\frac{\lambda^{\mathrm{T}}(\mu_x - \mu_y) - \Phi^{-1}(p)\sqrt{\lambda^{\mathrm{T}}\Sigma_x\lambda}}{\sqrt{\lambda^{\mathrm{T}}\Sigma_y\lambda}}\right)$, where λ is the linear combination coefficient (Su and Liu, 1993). Here we obtained the linear combined ROC curve based on elliptical distribution.

Theorem 4.3.1. *Suppose* $X \sim \mathcal{E}_d(\mu_x, \Sigma_x; \phi_x)$ *and* $Y \sim \mathcal{E}_d(\mu_y, \Sigma_y; \phi_y)$ *are distribution of case*

and control group respectively, then the linear combined ROC Curve could be expressed as

$$s(t,\lambda) = F_0\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}\Sigma_x\lambda}} - \frac{\sqrt{\lambda^{\mathrm{T}}\Sigma_y\lambda}}{\sqrt{\lambda^{\mathrm{T}}\Sigma_x\lambda}}G_0^{-1}(1-t)\right) \tag{4.3.1}$$

where s is the sensitivity (TPR) and t is the 1-specificity (FPR). F_0 is the c.d.f of $\frac{\lambda^{\mathrm{T}}X-\lambda^{\mathrm{T}}\mu_x}{\lambda^{\mathrm{T}}\Sigma_x\lambda}$ and G_0 is the c.d.f of $\frac{\lambda^{\mathrm{T}}Y-\lambda^{\mathrm{T}}\mu_y}{\lambda^{\mathrm{T}}\Sigma_y\lambda}$

Proof. Let $X_0 = \frac{X-\lambda^{\mathrm{T}}\mu_x}{\lambda^{\mathrm{T}}\Sigma_x\lambda} \sim \mathcal{E}_1(0,1;\phi_x)$ with p.d.f f_0 and c.d.f F_0, $Y_0 = \frac{Y-\lambda^{\mathrm{T}}\mu_y}{\lambda^{\mathrm{T}}\Sigma_y\lambda} \sim \mathcal{E}_1(0,1;\phi_y)$ with p.d.f g_0 and c.d.f G_0. Since f_0, g_0 are both symmetric functions, we have $f_0(x) = f_0(-x)$, $g_0(y) = g_0(-y)$, $1-F_0(x) = F_0(-x)$ and $1-G_0(y) = G_0(-y)$.

For each cut point $-\infty < c < \infty$, the false positive rate (1-specificity) t could be represented by:

$$t = P(Y > c) = P\left(\frac{Y-\lambda^{\mathrm{T}}\mu_y}{\sqrt{\lambda^{\mathrm{T}}\Sigma_y\lambda}} > \frac{c-\lambda^{\mathrm{T}}\mu_y}{\sqrt{\lambda^{\mathrm{T}}\Sigma_y\lambda}}\right) = G_0\left(\frac{\lambda^{\mathrm{T}}\mu_y - c}{\sqrt{\lambda^{\mathrm{T}}\Sigma_y\lambda}}\right)$$

So $c = \lambda^{\mathrm{T}}\mu_y - \sqrt{\lambda^{\mathrm{T}}\Sigma_y\lambda}G_0^{-1}(t)$. Then the sensitivity s is given by:

$$\begin{aligned}
s(t,\lambda) &= P(X > c) \\
&= F_0\left(\frac{\lambda^{\mathrm{T}}\mu_x - c}{\sqrt{\lambda^{\mathrm{T}}\Sigma_x\lambda}}\right) \\
&= F_0\left(\frac{\lambda^{\mathrm{T}}\mu_x - \lambda^{\mathrm{T}}\mu_y + \sqrt{\lambda^{\mathrm{T}}\Sigma_y\lambda}G_0^{-1}(t)}{\sqrt{\lambda^{\mathrm{T}}\Sigma_x\lambda}}\right) \\
&= F_0\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}\Sigma_x\lambda}} - \frac{\sqrt{\lambda^{\mathrm{T}}\Sigma_y\lambda}}{\sqrt{\lambda^{\mathrm{T}}\Sigma_x\lambda}}G_0^{-1}(1-t)\right)
\end{aligned}$$

□

In particular, when $\Sigma_x = \Sigma_y = \Sigma$, the ROC curve function could be simplified to $s(t,\lambda) = F_0\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}} - G_0^{-1}(1-t)\right)$. After getting the ROC function, the AUC value

is the integration of t from 0 to 1. What we need to do is finding the best linear combination which maximizes the AUC among all possible values of λ.

Chapter 5

Maximum Linear AUC Estimation under Elliptical Distribution

A lot of measurement indices are obtained from ROC curves. One of the most popular index is the AUC value which is the area under the ROC curve. The AUC value is belongs to 0 to 1 but usually larger than 0.5. A better biomarker should have a larger AUC value. When there are multiple biomarkers, we prefer a combination which has the largest AUC value. For example, considering the linear combination, we would choose a linear coefficient λ which could maximize AUC value.

5.1 Empirical Estimation

According to Bamber (1975), the AUC of the one biomarker can be represented by the probability that rank of a randomly selected positive instance higher than a negative one, which is equivalent to:

$$AUC = P(Y < X) \tag{5.1.1}$$

where Y and X are samples from control and case group, respectively.

Suppose there are d biomarkers, Let $l(\mathbf{M}) = \lambda^{\mathrm{T}}\mathbf{M}$ where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)^{\mathrm{T}}$ is the combination coefficients. The optimal linear combined AUC could be calculated by $AUC = P(l(Y) < l(X)) = P\left(\lambda^{\mathrm{T}}Y < \lambda^{\mathrm{T}}X\right)$. Suppose there are n independent observations in case group and m independent observations in control group. Given a $\lambda \in \mathbb{R}^d$, the empirical AUC would be:

$$AUC(\lambda) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} I\left(l(Y_j) < l(X_i)\right) \tag{5.1.2}$$

where I is the indicator function.

The best linear combination coefficient is the one that could maximize the AUC in equation 5.1.2 among all possible values of $\lambda \in \mathbb{R}^d$. That is, the best empirical estimator $\hat{\lambda}$ should be:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} I\left(l(Y_j) < l(X_i)\right) \tag{5.1.3}$$

5.2 Optimal Linear AUC of Elliptical Distribution

If $X \sim N_d(\mu_x, \Sigma_x)$ and $Y \sim N_d(\mu_y, \Sigma_y)$, Su and Liu (1993) proved that the best linear combination coefficient is $\hat{\lambda}_{max} = (\Sigma_x + \Sigma_y)^{-1}(\mu_x - \mu_y)$. They also proved that if case and control populations follow multivariate normal distributions with a same covariance matrix, the linear combination is equivalent to the optimal combination. Now we will extend the best linear combination to the elliptical distribution.

Theorem 5.2.1. *Suppose $X \sim \mathcal{E}_d(\mu_x, \Sigma, \phi_x)$, $Y \sim \mathcal{E}_d(\mu_y, \Sigma, \phi_y)$ are distribution of case and control group, respectively. Here two group have a same covariance generating matrix. The best λ which could maximize the linear combined AUC value is $\hat{\lambda}_{max} = \Sigma^{-1}(\mu_x - \mu_y)$. The largest linear combined AUC is:*

$$\int_0^1 F_0\left(\sqrt{(\mu_x - \mu_y)^{\mathrm{T}}\Sigma^{-1}(\mu_x - \mu_y)} - G_0^{-1}(1-t)\right) dt \tag{5.2.1}$$

where F_0 is the c.d.f of $\frac{(\mu_x-\mu_y)^{\mathrm{T}}\Sigma^{-1}(X-\mu_x)}{\sqrt{(\mu_x-\mu_y)^{\mathrm{T}}\Sigma^{-1}(\mu_x-\mu_y)}}$ and G_0 is the c.d.f of $\frac{(\mu_x-\mu_y)^{\mathrm{T}}\Sigma^{-1}(Y-\mu_y)}{\sqrt{(\mu_x-\mu_y)^{\mathrm{T}}\Sigma^{-1}(\mu_x-\mu_y)}}$.

Proof. From equation 4.3.1, when $\Sigma_x = \Sigma_y$:

$$s(t,\lambda) = F_0\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}} - G_0^{-1}(1-t)\right)$$

So the AUC is:

$$AUC(\lambda) = \int_0^1 R(t,\lambda)dt = \int_0^1 F_0\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}} - G_0^{-1}(1-t)\right)dt$$

Since F_0 is a non-decreasing function and G_0 part is independent of λ, maximizing AUC is equivalent to maximizing $\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}$ for $\forall t \in (0,1)$. Seber (1977) derived the solution for the quadratic form by Cauchy-Schwarz inequality. Therefore, $\hat{\lambda}_{max} = \Sigma^{-1}(\mu_x-\mu_y)$.

Then the maximum AUC is $\int_0^1 F_0\left(\sqrt{(\mu_x-\mu_y)^{\mathrm{T}}\Sigma^{-1}(\mu_x-\mu_y)} - G_0^{-1}(1-t)\right)dt$.

□

Here our result $\hat{\lambda}_{max}$ is consistent with the result of multivariate normal distribution from Su and Liu (1993), but we just require the same covariance generating matrix instead of the same covariance matrix. In fact, $\sqrt{(\mu_x-\mu_y)^{\mathrm{T}}\Sigma^{-1}(\mu_x-\mu_y)}$ is the Mahalanobis distance of elliptical distributions (Mitchell and Krzanowski, 1985). So the best linear combined AUC only depends on the Mahalanobis distance of elliptical distributions and the populations of case and control group.

Following is the another condition we got the optimal AUC estimation.

Theorem 5.2.2. *If case group and control group are jointly elliptical distributed, i.e.*

$$\begin{pmatrix} X_d \\ Y_d \end{pmatrix} \sim \varepsilon_{2d}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_x & 0 \\ 0 & \Sigma_y \end{pmatrix}, \phi\right)$$

Then the best λ which could maximize the linear combined AUC value is $\hat{\lambda}_{max} = (\Sigma_x + \Sigma_y)^{-1}(\mu_x - \mu_y)$. The largest linear combined AUC is:

$$\frac{1}{2}\left(F_R\left(\sqrt{(\mu_x-\mu_y)^{\mathrm{T}}(\Sigma_x+\Sigma_y)^{-1}(\mu_x-\mu_y)}\right)+1\right)$$

where R is the generating variates of jointly elliptical distribution.

Proof. For any $\lambda \in \mathbb{R}^d$

$$\begin{aligned} AUC(\lambda) &= P\left(\lambda^{\mathrm{T}}X \geq \lambda^{\mathrm{T}}X\right) \\ &= P\left(\lambda^{\mathrm{T}}(X-Y) \geq 0\right) \end{aligned}$$

Let $Z \triangleq \lambda^{\mathrm{T}}X - \lambda^{\mathrm{T}}Y$, then according to the theorem 3.2.2, Z follows a one dimensional elliptical distribution with mean $\lambda^{\mathrm{T}}\mu_x - \lambda^{\mathrm{T}}\mu_y$ and scalar parameter $\lambda^{\mathrm{T}}\Sigma_x\lambda + \lambda^{\mathrm{T}}\Sigma_y\lambda$.

Therefore, $f_Z(z) = \frac{1}{2\sqrt{\lambda^{\mathrm{T}}(\Sigma_x+\Sigma_y)\lambda}} f_R\left(\frac{|z-\lambda(\mu_x-\mu_y)|}{\sqrt{\lambda^{\mathrm{T}}(\Sigma_x+\Sigma_y)\lambda}}\right)$. Then the AUC could be expressed as:

$$\begin{aligned} AUC(\lambda) &= P(Z \geq 0) \\ &= \int_0^\infty f_Z(z)dz \\ &= \int_0^\infty \frac{1}{2\sqrt{\lambda^{\mathrm{T}}(\Sigma_x+\Sigma_y)\lambda}} f_R\left(\frac{|z-\lambda(\mu_x-\mu_y)|}{\sqrt{\lambda^{\mathrm{T}}(\Sigma_x+\Sigma_y)\lambda}}\right)dz \\ &= \frac{1}{2}\left(F_R\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}\Sigma_x\lambda+\lambda^{\mathrm{T}}\Sigma_y\lambda}}\right)+1\right) \end{aligned}$$

Since F_R is a cumulative function which is non-decreasing, maximizing AUC is equivalent to maximizing $\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}\Sigma_x\lambda+\lambda^{\mathrm{T}}\Sigma_y\lambda}}\right)$. Seber (1977) derived the solution for the quadratic form by Cauchy-Schwarz inequality.

Thus, $\hat{\lambda}_{max} = (\Sigma_x+\Sigma_y)^{-1}(\mu_x-\mu_y)$ and corresponding optimal AUC value is $AUC(\hat{\lambda}_{max}) = \frac{1}{2}\left(F_R\left(\sqrt{(\mu_x-\mu_y)^{\mathrm{T}}(\Sigma_x+\Sigma_y)^{-1}(\mu_x-\mu_y)}\right)+1\right)$.

□

The joint covariance generating matrix is a block diagonal matrix because the distributions of case and control group are usually independent.

Lemma 5.2.1. *Assume the conditions in Theorem 5.2.1 or Theorem 5.2.2 holds, the covariance matrix is expressed by* $V_(x) = \frac{E(\mathcal{R}_x^2)}{2}\Sigma_x$ *and* $V_(y) = \frac{E(\mathcal{R}_y^2)}{2}\Sigma_y$ ($\Sigma_x = \Sigma_y$ *under the conditions in Theorem 3.1), then the best linear combination coefficient could also be calculated by* $\hat{\lambda}_{max} = (V_x + V_y)^{-1}(\mu_x - \mu_y)$.

Proof. $\hat{\lambda}_{max} = (V_x + V_y)^{-1}(\mu_x - \mu_y) = c(\Sigma_x + \Sigma_y)^{-1}(\mu_x - \mu_y)$, which is proportional of the estimation of λ in theorem 3.1 and theorem 3.2. Here $c = \left(\frac{E(\mathcal{R}_x^2)+E(\mathcal{R}_y^2)}{d}\right)$ is a constant. □

5.3 Numerical Results for Combination of Elliptical Distribution

The main reason that we consider the distribution of biomarkers is that the combined AUC values are different for different distributions, even if they have the same location vectors and scalar matrices. Firstly, we will show the difference of optimal combinations and best linear combinations among different elliptical distributions.

5.3.1 AUC for Different Distributions

Three Dimensional Biomarkers with Equal Scalar Matrix

Suppose there are 3 biomarkers in case group and control group which follow the elliptical distribution with the expectation μ_x and μ_y, separately. Assume that they share the same scalar matrix Σ. The likelihood ratio combination is estimated by the theorem 4.1.1 and the best linear combination is calculated by the theorem 5.2.1.

$$\text{Assume } \mu_x = \begin{pmatrix} 2 \\ 1.6 \\ 1.3 \end{pmatrix}, \mu_y = \begin{pmatrix} 1.4 \\ 1 \\ 1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.15 & 0.1 \\ 0.15 & 0.26 & 0.02 \\ 0.1 & 0.02 & 1 \end{pmatrix}.$$

Combined AUC for Different Distributions				
Case Group	Control Group	LR Combination	Linear Combination	Ratio
Multivariate Normal	Multivariate Normal	0.8074	0.8074	1
Multivariate T_3	Multivariate T_{11}	0.8014	0.7715	0.9627
Multivariate T_4	Multivariate T_6	0.7841	0.7725	0.9852
Multivariate Laplace	Multivariate Laplace	0.8602	0.8355	0.9713
Multivariate T_4	Multivariate Laplace	0.8076	0.7963	0.9860
Multivariate T_{11}	Multivariate Laplace	0.8156	0.8114	0.9948
Multivariate Laplace	Multivariate T_5	0.8075	0.8009	0.9919

Table 5.1: Combined AUC for Different Distributions with the Same Location Vector and Scalar Matrix

The best linear combined AUC values for different distribution types are shown in table 5.1. The first two columns are the distribution populations of case and control group. The third column expresses the AUC values given by the likelihood ratio combination in the theorem 4.1.1. The best linear combined AUC values are shown in the third column. The last column shows the ratio of best linear combination and likelihood ratio combination.

The ratio of multivariate normal distribution is 1 which is consistent with the conclusion that the best linear combination is equivalent to the likelihood ratio combination when the scalar metrics are same in case and control group.

Four Dimensional Biomarkers with Equal Scalar Matrix

Table 5.2 shows another simulation results for 4 dimensional biomarkers with parameters:

$$\mu_x = \begin{pmatrix} 4 \\ 3.6 \\ 8 \\ 5 \end{pmatrix}, \mu_y = \begin{pmatrix} 3 \\ 3.5 \\ 7.5 \\ 4 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.2 & 0.3 & 0.1 \\ 0.2 & 2 & 0.2 & 0.02 \\ 0.3 & 0.2 & 0.6 & 0.1 \\ 0.1 & 0.02 & 0.1 & 1 \end{pmatrix}$$

Combined AUC for Different Distributions				
Case Group	Control Group	LR Combination	Linear Combination	Ratio
Multivariate Normal	Multivariate Normal	0.8329	0.8329	1
Multivariate T_3	Multivariate T_{11}	0.8309	0.7947	0.9564
Multivariate T_4	Multivariate T_6	0.8107	0.7958	0.9817
Multivariate Laplace	Multivariate Laplace	0.8870	0.8575	0.9668
Multivariate T_4	Multivariate Laplace	0.8399	0.8194	0.9756
Multivariate T_{11}	Multivariate Laplace	0.8451	0.8355	0.9887
Multivariate Laplace	Multivariate T_5	0.8396	0.8245	0.9820

Table 5.2: Combined AUC for Different Distributions with the Same Location Vector and Scalar Matrix

From table 5.1 and 5.2 we can see that the ratios of linear combination and likelihood ratio combination are all larger than 0.9. This means when the biomarkers follow elliptical distributions with a same scalar matrix in case and control populations, the loss between best linear combination and likelihood ratio combination is relatively small.

Three Dimensional Biomarkers with Unequal Scalar Matrices

When case and control group have different scalar matrices, we can still use $\hat{\lambda}_{elliptical} = (\Sigma_x + \Sigma_y)^{-1}(\mu_x - \mu_y)$ to calculate the linear combined AUC although there may be some loss. The combined AUC values for different distributions are shown in table 5.3 with

parameters:

$$\mu_x = \begin{pmatrix} 2 \\ 3 \\ 1.5 \end{pmatrix}, \mu_y = \begin{pmatrix} 1.4 \\ 1 \\ 1 \end{pmatrix}, \Sigma_x = \begin{pmatrix} 1 & 0.15 & 0.1 \\ 0.15 & 0.26 & 0.02 \\ 0.1 & 0.02 & 1 \end{pmatrix}, \Sigma_y = \begin{pmatrix} 1 & 0.2 & 0.15 \\ 0.2 & 3 & 0.1 \\ 0.15 & 0.1 & 1.5 \end{pmatrix}$$

Combined AUC for Different Distributions				
Case Group	Control Group	LR Combination	Linear Combination	Ratio
Multivariate Normal	Multivariate Normal	0.8783	0.8783	1
Multivariate T_3	Multivariate T_{11}	0.9131	0.8514	0.9324
Multivariate T_4	Multivariate T_6	0.9155	0.8472	0.9254
Multivariate Laplace	Multivariate Laplace	0.9562	0.9005	0.9418
Multivariate T_4	Multivariate Laplace	0.9307	0.8882	0.9543
Multivariate T_{11}	Multivariate Laplace	0.9424	0.8972	0.9521
Multivariate Laplace	Multivariate T_5	0.9393	0.8537	0.9088

Table 5.3: Combined AUC for Different Distributions with the Same Location Vector and Scalar Matrix

Four Dimensional Biomarkers with Unequal Scalar Matrices

Table 5.2 shows another simulation results for 4 dimensional biomarkers with paremeters:

$$\mu_x = \begin{pmatrix} 4 \\ 3.6 \\ 8 \\ 5 \end{pmatrix}, \mu_y = \begin{pmatrix} 3 \\ 3.5 \\ 7.5 \\ 4 \end{pmatrix}$$

$$\Sigma_x = \begin{pmatrix} 1 & 0.2 & 0.3 & 0.1 \\ 0.2 & 2 & 0.2 & 0.02 \\ 0.3 & 0.2 & 0.6 & 0.1 \\ 0.1 & 0.02 & 0.1 & 1 \end{pmatrix} \text{ and } \Sigma_y = \begin{pmatrix} 1 & 0.15 & 0.1 & 0.6 \\ 0.15 & 1.5 & 0.2 & 0.15 \\ 0.1 & 0.2 & 2 & 0.6 \\ 0.6 & 0.15 & 0.6 & 1 \end{pmatrix}$$

Combined AUC for Different Distributions				
Case Group	Control Group	LR Combination	Linear Combination	Ratio
Multivariate Normal	Multivariate Normal	0.8063	0.8063	1
Multivariate T_3	Multivariate T_{11}	0.8800	0.7722	0.8775
Multivariate T_4	Multivariate T_6	0.8705	0.7724	0.8873
Multivariate Laplace	Multivariate Laplace	0.9147	0.8352	0.9131
Multivariate T_4	Multivariate Laplace	0.8848	0.8014	0.9058
Multivariate T_{11}	Multivariate Laplace	0.8919	0.8160	0.9148
Multivariate Laplace	Multivariate T_5	0.8839	0.7952	0.8996

Table 5.4: Combined AUC for Different Distributions with the Same Location Vector and Scalar Matrix

From table 5.3 and 5.4 we can see that the loss between best linear combination and likelihood ratio combination is relatively small.

In the next section, we will show the details about simulation results.

5.3.2 Simulation for the Same Scalar Matrix

Multivariate T-Distribution with the Same Scalar Matrix

Multivariate t-distribution is a generalization to random vectors of the Student's t-distribution. It would converge to multivariate normal distribution as the degree of freedom $v \to \infty$. The real dataset which follows multivariate t-distribution could easily be mistaken for multivariate normal distribution, especially when the degree of freedom is relatively high. This kind of misjudgment may reduce diagnostic accuracy obviously although multivariate t-distribution has similar features as multivariate normal distribution. In order to obtain the loss of combined AUC caused by misjudgment, we did the simulation under the multivariate t-distribution.

Suppose case and control group follow the multivariate-t distribution with the expectation μ_x and μ_y, separately. Assume that they share the same covariance generating matrix

Σ:

$$X \sim T_{v_x}(\mu_x, \Sigma), \quad Y \sim T_{v_y}(\mu_y, \Sigma)$$

Notice the covariance matrices V of case and control group are different when they have different degrees of freedom.

$$V_x = \frac{v_x}{v_x - 2}\Sigma, \quad V_y = \frac{v_y}{v_y - 2}\Sigma$$

$$\text{Let } v_x = 3,\ v_y = 11,\ \mu_x = \begin{pmatrix} 2 \\ 1.6 \\ 1.3 \end{pmatrix},\ \mu_y = \begin{pmatrix} 1.4 \\ 1 \\ 1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.15 & 0.1 \\ 0.15 & 0.26 & 0.02 \\ 0.1 & 0.02 & 1 \end{pmatrix}.$$

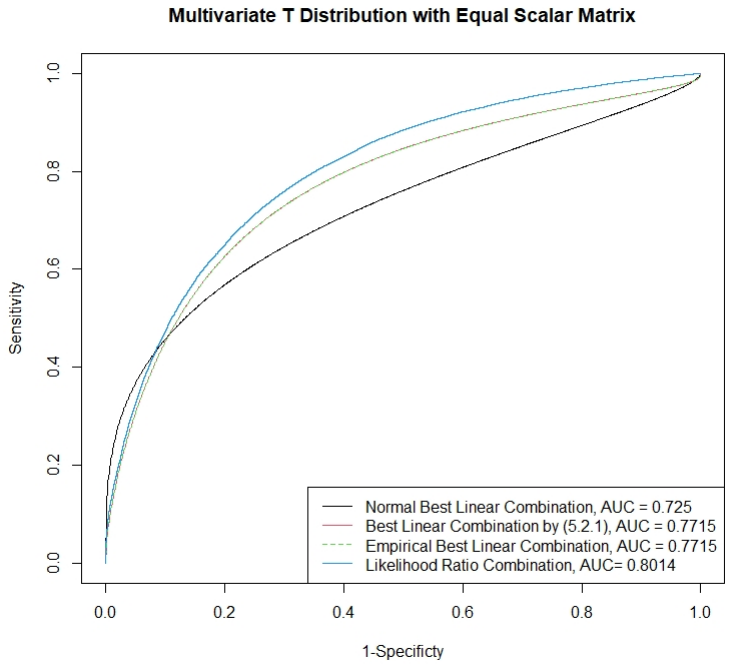

Figure 5.1: Multivariate T Distribution with Same Scalar Matrix

We generated 50000 observations in each group and the simulation result is shown in Figure 5.1. The empirical maximum linear combined ROC is shown as the green dotted line and the corresponding AUC is 0.7715. From Theorem 5.2.1 we could calculate the

optimal linear coefficient $\hat{\lambda}_{max} = \Sigma^{-1}(\mu_x - \mu_y) = \begin{pmatrix} 0.2555 & 2.1425 & 0.2316 \end{pmatrix}^{\mathrm{T}}$ and the corresponding optimal linear combined AUC value is 0.7715. The ROC curve obtained by Theorem 5.2.1 is shown as the red line which almost coincides with the empirical estimation in Figure 5.1

However, if the data is misjudged as the multivariate normal distribution with different covariance matrix, then the linear combined ROC curve could be expressed as $R(t) = \Phi\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}V_x\lambda}} - \frac{V_y}{V_x}\Phi^{-1}(1-t)\right)$, where Φ is the cumulative distribution function of the standard normal distribution (Su and Liu, 1993). And the corresponding AUC value under normal assumption would be 0.725, which is less than optimal linear combined AUC value.

We also show the likelihood ratio combination ROC proposed in chapter 4.1. It is obvious that the AUC of likelihood ratio combination is larger than any other combinations.

Multivariate Laplace Distribution with the Same Scalar Matrix

Another popular elliptical distribution is multivariate Laplace distribution (Kotz et al., 2001). Assume that the case and control populations follow the multivariate Laplace distribution. The location vectors μ_x and μ_y and the scalar matrix Σ are same with the previous settings. We still generate 50000 observations in each group.

Figure 5.2 shows the simulation results. Similar with the multivariate T distribution, the AUC of likelihood ratio combination is largest which equals to 0.8602. If we misjudge the data as multivariate normal distribution, the linear combined AUC will reduce from 0.8355 to 0.8072.

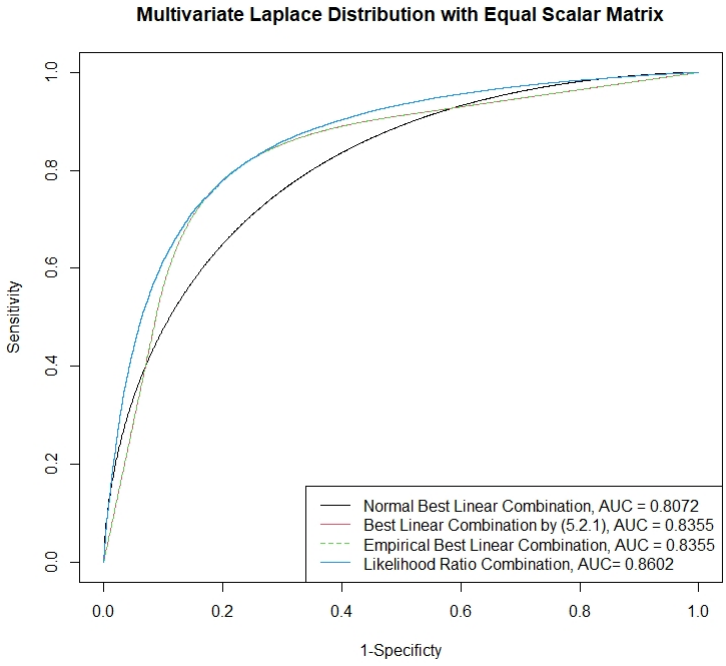

Figure 5.2: Multivariate Laplace Distribution with Same Scalar Matrix

Other Elliptical Distributions with the Same Scalar Matrix

Instead of popular elliptical distributions like multivariate-t distribution, we can generate the random variables according to the definition of elliptical distribution. Let $U_x^{(3)}$, $U_y^{(3)}$ are uniformly distributed on the unit hypersphere with 2 topological dimension:

$$\mathcal{S}^2 := \left\{ x \in IR^3 : \|x\|_2 = 1 \right\}$$

Let $\mathcal{R}_x \sim U(1,3)$, $\mathcal{R}_y \sim Exponential(1)$,

$$\mu_x = \begin{pmatrix} 2 \\ 1.8 \\ 1.8 \end{pmatrix}, \mu_y = \begin{pmatrix} 1.5 \\ 1 \\ 1.6 \end{pmatrix}, \Lambda = \begin{pmatrix} 1 & 0.1 & 0.3 \\ 0.1 & 0.5 & 0.02 \\ 0.3 & 0.02 & 0.7 \end{pmatrix}.$$

Then we have:

$$X \overset{d}{=} \mu_x + \mathcal{R}_x \Lambda U_x^{(2)} \sim \mathcal{E}_3(\mu_x, \Sigma, \phi_x)$$

$$Y \overset{d}{=} \mu_y + \mathcal{R}_y \Lambda U_y^{(2)} \sim \mathcal{E}_3(\mu_y, \Sigma, \phi_y)$$

where $\Sigma = \Lambda\Lambda^{\mathrm{T}}$. So X and Y both follow elliptical distribution with the same covariance generating matrix. We still generated 50000 observations in each group. The simulation results are shown in Figure 5.3.

The likelihood ratio combined AUC is 0.9437 which is largest among all combinations. The best linear combination estimated by Theorem 5.2.1 is $\hat{\lambda}_{max} = \Sigma^{-1}(\mu_x - \mu_y) = \begin{pmatrix} -0.0133 & 3.0654 & 0.0711 \end{pmatrix}^{\mathrm{T}}$ and the corresponding optimal linear combined AUC value is 0.8624. Using the method proposed by Su and Liu (1993), we can get the linear combined AUC under normality assumption which is 0.8598. Another thing we can discover is that the normal curve is not always lower that the best linear combination.

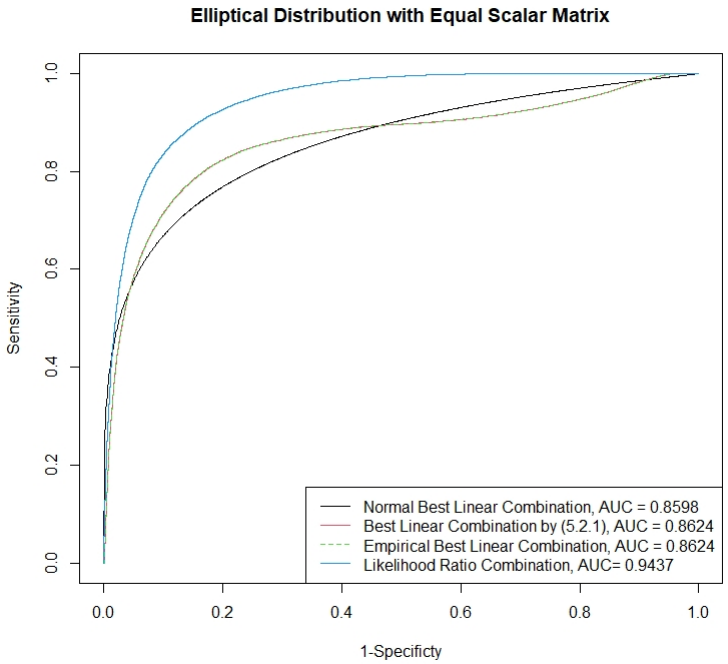

Figure 5.3: Elliptical Distribution with Same Scalar Matrix

From above simulation results, we can conclude that the combination obtained from Theorem 5.2.1 is the best linear combination.

5.3.3 Simulation for Different Scalar Matrices

When case and control group neither share a same covariance generating matrix nor jointly elliptical distributed, we can still use $\hat{\lambda}_{elliptical} = (\Sigma_x + \Sigma_y)^{-1}(\mu_x - \mu_y)$ to calculate the linear combined AUC although there may be some loss. One main question here is that we need to find out the difference of AUC values between different linear combinations. We will calculate four combination coefficients and corresponding AUC values:

$$\hat{\lambda}_{elliptical} = (\Sigma_x + \Sigma_y)^{-1}(\mu_x - \mu_y), \quad \hat{AUC}_{elliptical} = P\left(\hat{\lambda}_{elliptical}^{\mathrm{T}} Y < \hat{\lambda}_{elliptical}^{\mathrm{T}} X\right)$$

$$\hat{\lambda}_{var} = (V_x + V_y)^{-1}(\mu_x - \mu_y), \quad \hat{AUC}_{normal} = \Phi\left(\frac{\hat{\lambda}_{var}^{\mathrm{T}}(\mu_x - \mu_y)}{\left(\hat{\lambda}_{var}^{\mathrm{T}}(V_x + V_y)^{-1}\hat{\lambda}_{var}\right)^{1/2}}\right)$$

$$\hat{\lambda}_{var} = (V_x + V_y)^{-1}(\mu_x - \mu_y), \quad \hat{AUC}_{ell.var} = P\left(\hat{\lambda}_{var}^{\mathrm{T}} Y < \hat{\lambda}_{var}^{\mathrm{T}} X\right)$$

$$\hat{\lambda}_{best} = \underset{\lambda}{\operatorname{argmax}}\left(P\left(\lambda^{\mathrm{T}} Y < \lambda^{\mathrm{T}} X\right)\right) \quad \hat{AUC}_{best} = P\left(\hat{\lambda}_{best}^{\mathrm{T}} Y < \hat{\lambda}_{best}^{\mathrm{T}} X\right)$$

The main purpose of this part is to find out how much loss there would be, compared with the best linear combination.

Multivariate T-Distribution with Different Scalar Matrices

Assume biomarkers in case and control group follow multivariate-t distribution:

$$X \sim T_{v_x}(\mu_x, \Sigma_x), \quad Y \sim T_{v_y}(\mu_y, \Sigma_y)$$

Let $v_x = 3$, $v_y = 11$, $\mu_x = \begin{pmatrix} 2 \\ 3 \\ 1.5 \end{pmatrix}$, $\mu_y = \begin{pmatrix} 1.4 \\ 1 \\ 1 \end{pmatrix}$, $\Sigma_x = \begin{pmatrix} 1 & 0.15 & 0.1 \\ 0.15 & 0.26 & 0.02 \\ 0.1 & 0.02 & 1 \end{pmatrix}$ and $\Sigma_y = \begin{pmatrix} 1 & 0.2 & 0.15 \\ 0.2 & 3 & 0.1 \\ 0.15 & 0.1 & 1.5 \end{pmatrix}$.

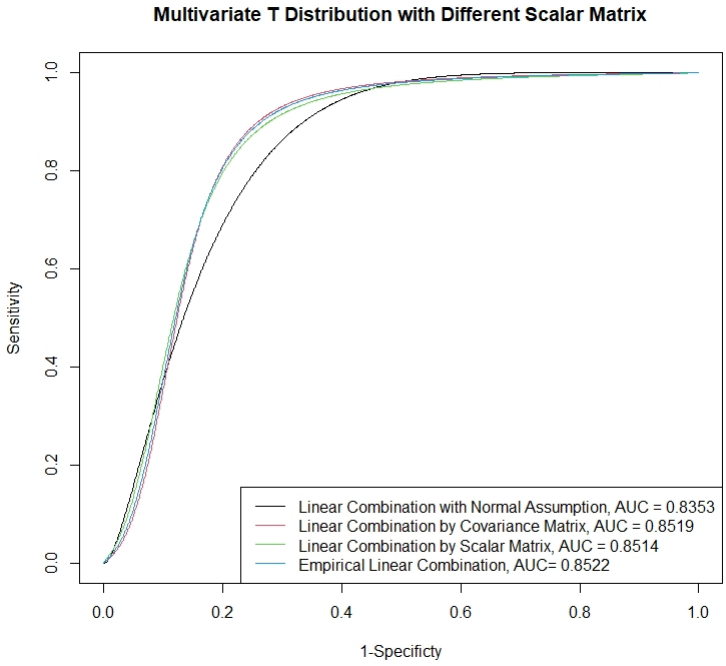

Figure 5.4: Multivariate T Distribution with Different Scalar Matrices

Generating 50000 observations in each group and the simulation results are shown in Figure 3. $\hat{AUC}_{normal} = 0.8353$, which is the smallest AUC values. $\hat{AUC}_{ell.var} = 0.8519$, $\hat{AUC}_{elliptical} = 0.8514$ and $\hat{AUC}_{best} = 0.8522$. From Figure 5.4 we can see that except the linear combination with normal assumption, the other three linear combined ROC curves are close to each other. Although the AUC value is smallest under normality assumption, it has largest accuracy over a range of high specificity. This is one of the drawbacks of AUC index. Small value of AUC doesn't always mean a bad performance. If we consider partial area, the normality line would have a better performance than others.

Multivariate Laplace Distribution with Different Scalar Matrices

Assume biomarkers in case and control group follow multivariate Laplace distribution:

$$X \sim ML_x(\mu_x, \Sigma_x), \quad Y \sim ML_y(\mu_y, \Sigma_y)$$

$$\mu_x = \begin{pmatrix} 2 \\ 3 \\ 1.5 \end{pmatrix}, \mu_y = \begin{pmatrix} 1.4 \\ 1 \\ 1 \end{pmatrix}, \Sigma_x = \begin{pmatrix} 1 & 0.15 & 0.1 \\ 0.15 & 5 & 0.02 \\ 0.1 & 0.02 & 1 \end{pmatrix}, \Sigma_y = \begin{pmatrix} 1 & 0.2 & 0.15 \\ 0.2 & 3 & 0.1 \\ 0.15 & 0.1 & 1.5 \end{pmatrix}$$

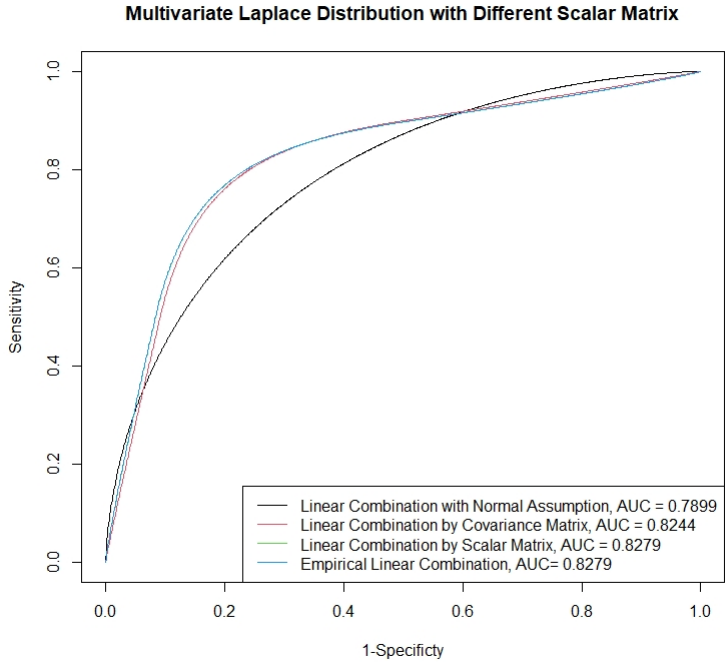

Figure 5.5: Multivariate Laplace Distribution with Different Scalar Matrices

Generating 50000 observations in each group and the simulation results are shown in Figure 3. $\hat{AUC}_{normal} = 0.7988$, which is the smallest AUC values. $\hat{AUC}_{ell.var} = 0.8244$, $\hat{AUC}_{elliptical} = 0.8279$ and $\hat{AUC}_{best} = 0.8279$. From Figure 5.5 we can see that except the linear combination with normal assumption, the other three linear combined ROC curves are close to each other. Especially, the ROC curve estimated by the elliptical distribution almost coincides with the empirical linear ROC curve.

Other Elliptical Distributions with Different Scalar Matrices

The simulation in this chapter is similar with previous settings in chapter 5.3.2 except that the covariance generating matrices in case and control group are different. Let $U_x^{(3)}$,

$U_y^{(3)}$ are uniformly distributed on the unit hypersphere with 2 topological dimension:

$$\mathcal{S}^2 := \{x \in IR^3 : \|x\|_2 = 1\}$$

Let $\mathcal{R}_x \sim U(1,3)$, $\mathcal{R}_y \sim Exponential(1)$.

$$\mu_x = \begin{pmatrix} 2 \\ 1.8 \\ 1.8 \end{pmatrix}, \mu_y = \begin{pmatrix} 1.5 \\ 1 \\ 1.6 \end{pmatrix}, \Lambda_x = \begin{pmatrix} 1 & 0.1 & 0.3 \\ 0.1 & 0.5 & 0.02 \\ 0.3 & 0.02 & 0.7 \end{pmatrix}, \Lambda_y = \begin{pmatrix} 1 & 0.05 & 0.2 \\ 0.05 & 1.5 & 0.6 \\ 0.2 & 0.6 & 1 \end{pmatrix}$$

.

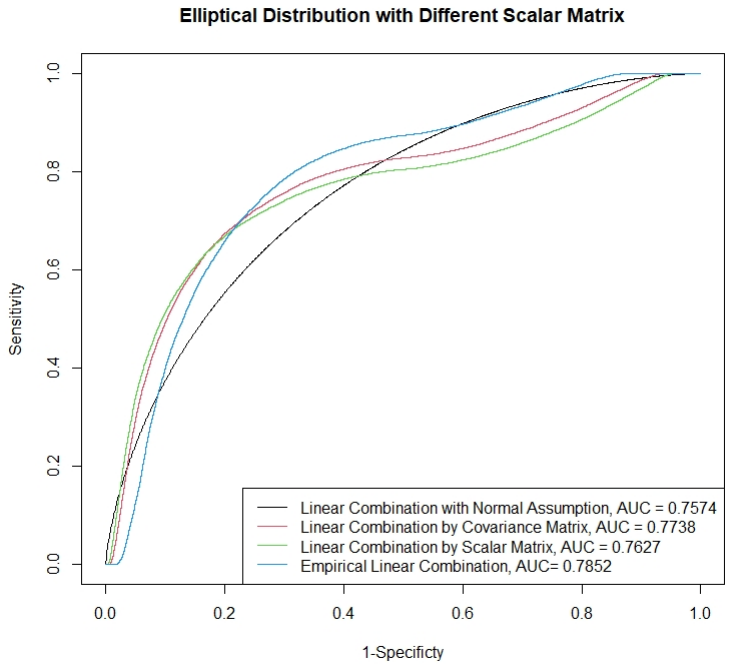

Figure 5.6: Elliptical Distribution with Different Scalar Matrices

Then we have:

$$X \overset{d}{=} \mu_x + \mathcal{R}_x \Lambda_x U_x^{(2)} \sim \mathcal{E}_3(\mu_x, \Sigma_x)$$

$$Y \overset{d}{=} \mu_y + \mathcal{R}_y \Lambda_y U_y^{(2)} \sim \mathcal{E}_3(\mu_y, \Sigma_y)$$

where $\Sigma_x = \Lambda_x \Lambda_x^{\mathrm{T}}$ and $\Sigma_y = \Lambda_y \Lambda_y^{\mathrm{T}}$.

We still generated 50000 observations in each group. The simulation results are presented in Figure 5.6. $\hat{AUC}_{normal} = 0.7574$, which is the smallest AUC values. $\hat{AUC}_{ell.var} = 0.7738$, $\hat{AUC}_{elliptical} = 0.7627$ and $\hat{AUC}_{best} = 0.7852$. From Figure 5.6 we can see that the ROC curves are close to each other when using covariance matrix and covariance generating matrix to estimate the linear combination coefficient. Although the AUC value of the best linear combination is the largest, it has lower accuracy over a range of high specificity. This expresses the drawback of AUC index again. Small value of AUC doesn't always mean a bad performance. If we consider partial area, the normality line would have a better performance than others.

Chapter 6

Combination of Projected Length

From previous simulation, we can find some drawbacks of AUC. Besides our previous results, here are some other conditions that AUC may not be a great index. Another example could show the drawbacks of auc value is here. If we have two curves, both auc values are 0.5. But we can't say they have same performance because the orange line indeed has some diagnosis ability especially for a range of high specificity. To offset the disadvantages of auc value, Lee and Hsiao (1996) proposed another index - projected length.

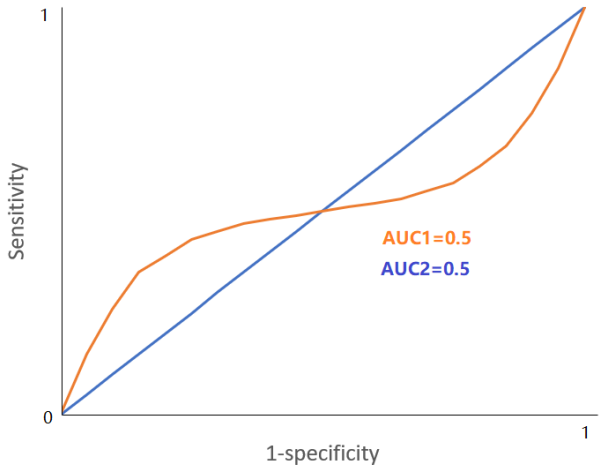

Figure 6.1: An Example of Drawbacks of AUC

6.1 Definition of Projected Length

The definition of "projected length of the ROC curve" (PLC) is proposed by Lee and Hsiao (1996). It is the summation of all of the projected lengths of the ROC curve onto the "negative" diagonal line. For example, here is a ROC curve which is composed of three segments of straight lines a, b, and c. Then The PLC index is the summation of all of the projected lengths of these three segments, i.e. $(a' + b' + c')$.

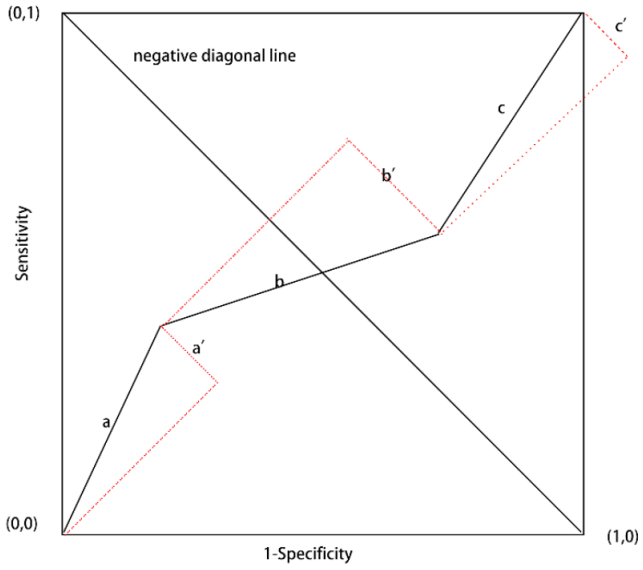

Figure 6.2: Example: Projected Length of the ROC Curve

Example 6.1.1. Suppose a ROC is composed of five points with coordinates (0, 0), (0.1, 0.5), (0.6, 0.7), (0.8, 0.9) and (1,1). Then is PLC is:

$$\begin{aligned}\text{PLC} = &\left|(0.1,\ 0.5)\cdot\left(\frac{1}{\sqrt{2}},\frac{-1}{\sqrt{2}}\right)\right| + \left|(0.6-0.1,\ 0.7-0.5)\cdot\left(\frac{1}{\sqrt{2}},\frac{-1}{\sqrt{2}}\right)\right| \\ &+ \left|(0.8-0.6,\ 0.9-0.7)\cdot\left(\frac{1}{\sqrt{2}},\frac{-1}{\sqrt{2}}\right)\right| + \left|(1-0.8,\ 1-0.9)\cdot\left(\frac{1}{\sqrt{2}},\frac{-1}{\sqrt{2}}\right)\right| \\ =&0.5657\end{aligned}$$

Definition 6.1.1. Lee and Hsiao (1996) defined that the PLC of a continuous biomarker could be represented as: $\frac{1}{2} + \frac{\sqrt{2}}{4}\text{PLC} = \int_{D\cup N} \max\left\{\frac{f_X(t)}{2}, \frac{f_Y(t)}{2}\right\} dt$, where the $f_X(t)$ and

$f_Y(t)$ represent the density functions of the biomarker in the diseased and the non-diseased populations, respectively, and

$$D = \left\{ t : \frac{f_Y(t)}{f_Y(t)+f_X(t)} \leq \frac{f_X(t)}{f_Y(t)+f_X(t)} \right\}$$
$$N = \left\{ t : \frac{f_Y(t)}{f_Y(t)+f_X(t)} > \frac{f_X(t)}{f_Y(t)+f_X(t)} \right\}$$

In other words, the PLC is a function of the sum of probability of correctly diagnosing a diseased subject and probability of correctly diagnosing a non-diseased subject. The maximum value of PLC is $\sqrt{2}$. Similarly with AUC value, higher PLC means better diagnosis performance. But the PLC is more sensitive than the AUC value for detecting differences between the distributions of the diseased and the non-diseased groups(Lee and Hsiao, 1996). Zhang et al. (2009) discussed basic properties of PLC and derived approximate variance for the PLC estimator.

6.2 Optimal Linear Combination of PLC Under Normal Distribution

Similar with the combined AUC value, we could derive the best linear combination which could maximize the combined PLC of d biomarkers. Firstly we begin with the normal assumption:

Theorem 6.2.1. *Suppose here are d biormarkers following multivariate normal distribution with the same covariance matrix in case and control group, denoted by* $X \sim MN_d(\mu_x, \Sigma)$, $Y \sim MN_d(\mu_y, \Sigma)$, *respectively. The best linear combination coefficient* λ *which could maximize the linear combined PLC is* $\hat{\lambda}_{max} = \Sigma^{-1}(\mu_x - \mu_y)$, *and the corresponding linear combined PLC is:*

$$PLC_{max} = 2\sqrt{2}\Phi\left(\frac{1}{2}\sqrt{(\mu_x-\mu_y)^T\Sigma^{-1}(\mu_x-\mu_y)}\right) - \sqrt{2} \tag{6.2.1}$$

where Φ *is the c.d.f of standard normal distribution.*

Proof. Suppose $X \sim MN_d(\mu_x, \Sigma)$, $Y \sim MN_d(\mu_y, \Sigma)$, then for any $\lambda \in \mathbb{R}^d$ we have:

$$X^* \triangleq \lambda^{\mathrm{T}} X \sim N(\lambda^{\mathrm{T}}\mu_x, \lambda^{\mathrm{T}}\Sigma\lambda),\ Y^* \triangleq \lambda^{\mathrm{T}} Y \sim N(\lambda^{\mathrm{T}}\mu_y, \lambda^{\mathrm{T}}\Sigma\lambda)$$

The linear combined PLC is given by: $\frac{1}{2} + \frac{\sqrt{2}}{4}\mathrm{PLC}(\lambda) = \int_{D\cup N} \max\left\{\frac{f_X^*(t)}{2}, \frac{f_Y^*(t)}{2}\right\} \mathrm{d}t$, where the $f_X^*(t)$ and $f_Y^*(t)$ are density functions of X^* and Y^*.

$$\begin{aligned} f_Y^*(t) \le f_X^*(t) &\iff -\frac{1}{2}\left(\frac{t-\lambda^{\mathrm{T}}\mu_y}{\lambda^{\mathrm{T}}\Sigma\lambda}\right)^2 \le -\frac{1}{2}\left(\frac{t-\lambda^{\mathrm{T}}\mu_x}{\lambda^{\mathrm{T}}\Sigma\lambda}\right)^2 \\ &\iff \left(t-\lambda^{\mathrm{T}}\mu_x\right)^2 - \left(t-\lambda^{\mathrm{T}}\mu_y\right)^2 \le 0 \\ &\iff t \ge \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2} \end{aligned}$$

Therefore $D = \left\{t : \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2} \le t < \infty\right\}$ and $N = \left\{t : -\infty < t < \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2}\right\}$. Then we have:

$$\begin{aligned} \frac{1}{2} + \frac{\sqrt{2}}{4}\mathrm{PLC}(\lambda) &= \frac{1}{2}\left(\int_{\lambda^{\mathrm{T}}(\mu_x+\mu_y)/2}^{\infty} f_X^*(t)dt + \int_{-\infty}^{\lambda^{\mathrm{T}}(\mu_x+\mu_y)/2} f_Y^*(t)dt\right) \\ &= \frac{1}{2}\left(P(X^* \ge \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2}) + P(Y^* \le \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2})\right) \\ &= \Phi\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{2\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}\right) \end{aligned}$$

Maximizing PLC is equivalent to maximizing $\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{2\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}$. From Seber (1977), the solution of this optimal question is $\hat{\lambda}_{max} = \Sigma^{-1}(\mu_x - \mu_y)$. The corresponding maximized PLC is given by:

$$\frac{1}{2}+\frac{\sqrt{2}}{4}\text{PLC}_{max}=\Phi\left(\frac{1}{2}\sqrt{(\mu_x-\mu_y)^T\Sigma^{-1}(\mu_x-\mu_y)}\right)$$
$$\Longrightarrow \text{PLC}_{max}=2\sqrt{2}\Phi\left(\frac{1}{2}\sqrt{(\mu_x-\mu_y)^T\Sigma^{-1}(\mu_x-\mu_y)}\right)-\sqrt{2}$$

□

We can see that for the multivariate normal distribution, the best linear combination coefficients of AUC and PLC are equal when the covariance matrices in case and groups are equal. This result verifies the conclusion again that the best linear combination is equivalent to the optimal combination for the multivariate normal distributions with the same covariance matrices in case and control populations.

6.3 Optimal Linear Combination of PLC Under Elliptical Distribution

In this section, we will extend the linear combination of PLC from multivariate normal distribution to the elliptical distribution family. We would like to derive the best linear combination coefficient which could maximize the linear combined PLC for the elliptical unimodal.

Theorem 6.3.1. *Suppose d biomarkers follow elliptical unimodal distribution with a same scalar matrix and a same generating variate in case and control populations, denoted as* $X \sim \mathcal{E}_d(\mu_x,\Sigma,\phi_x)$, $Y \sim \mathcal{E}_d(\mu_y,\Sigma,\phi_y)$ *respectively. Then the best* λ *which could maximize the linear combined PLC value is* $\hat{\lambda}_{max}=\Sigma^{-1}(\mu_x-\mu_y)$ *and the corresponding maximized PLC is:*

$$\text{PLC}_{max}=\sqrt{2}\left(F_0(a)+G_0(a)\right)-\sqrt{2} \tag{6.3.1}$$

where $a=\frac{\sqrt{(\mu_x-\mu_y)^T\Sigma^{-1}(\mu_x-\mu_y)}}{2}$.

Proof. Suppose $X \sim \mathcal{E}_d(\mu_x, \Sigma, \phi_x)$, $Y \sim \mathcal{E}_d(\mu_y, \Sigma, \phi_y)$. Then for any $\lambda \in \mathbb{R}^d$ we have:

$$X^* \triangleq \lambda^{\mathrm{T}} X \sim \mathcal{E}_1(\lambda^{\mathrm{T}}\mu_x, \lambda^{\mathrm{T}}\Sigma\lambda, \phi_x),\ Y^* \triangleq \lambda^{\mathrm{T}} Y \sim \mathcal{E}_1(\lambda^{\mathrm{T}}\mu_y, \lambda^{\mathrm{T}}\Sigma\lambda, \phi_y)$$

The linear combined PLC is given by: $\frac{1}{2} + \frac{\sqrt{2}}{4}\mathrm{PLC}(\lambda) = \int_{D\cup N} \max\left\{\frac{f_X^*(t)}{2}, \frac{f_Y^*(t)}{2}\right\} \mathrm{d}t$, where the $f_X^*(t)$ and $f_Y^*(t)$ are density functions of X^* and Y^*.

According to theorem 3.1.1:

$$f_{X^*}(x) = \frac{1}{\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}} g_{\mathcal{R}}\left(\sqrt{\frac{(x-\lambda^{\mathrm{T}}\mu_x)^2}{\lambda^{\mathrm{T}}\Sigma\lambda}}\right)$$

and

$$f_{Y^*}(y) = \frac{1}{\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}} g_{\mathcal{R}}\left(\sqrt{\frac{(y-\lambda^{\mathrm{T}}\mu_y)^2}{\lambda^{\mathrm{T}}\Sigma\lambda}}\right)$$

where $g_{\mathcal{R}}$ is monotonically non-increasing since vectors X and Y are elliptical unimodal.

$$\begin{aligned} f_Y^*(t) \le f_X^*(t) &\iff g_{\mathcal{R}}\left(\sqrt{\frac{(t-\lambda^{\mathrm{T}}\mu_y)^2}{\lambda^{\mathrm{T}}\Sigma\lambda}}\right) \le g_{\mathcal{R}}\left(\sqrt{\frac{(t-\lambda^{\mathrm{T}}\mu_x)^2}{\lambda^{\mathrm{T}}\Sigma\lambda}}\right) \\ &\iff \left(t-\lambda^{\mathrm{T}}\mu_x\right)^2 - \left(t-\lambda^{\mathrm{T}}\mu_y\right)^2 \le 0 \\ &\iff t \ge \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2} \end{aligned}$$

Therefore $D = \left\{t : \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2} \le t < \infty\right\}$ and $N = \left\{t : -\infty < t < \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2}\right\}$. Then we have:

$$\begin{aligned} \frac{1}{2} + \frac{\sqrt{2}}{4}\mathrm{PLC}(\lambda) &= \frac{1}{2}\left(\int_{\lambda^{\mathrm{T}}(\mu_x+\mu_y)/2}^{\infty} f_X^*(t)dt + \int_{-\infty}^{\lambda^{\mathrm{T}}(\mu_x+\mu_y)/2} f_Y^*(t)dt\right) \\ &= \frac{1}{2}\left(P(X^* \ge \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2}) + P(Y^* \le \frac{\lambda^{\mathrm{T}}(\mu_x+\mu_y)}{2})\right) \\ &= \frac{1}{2}\left(\left(1 - F_0\left(\frac{\lambda^{\mathrm{T}}(\mu_y-\mu_x)}{2\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}\right)\right) + G_0\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{2\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}\right)\right) \end{aligned}$$

where F_0 is the c.d.f of $\frac{\lambda^{\mathrm{T}}(X-\mu_x)}{\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}$ and G_0 is the c.d.f of $\frac{\lambda^{\mathrm{T}}(Y-\mu_y)}{\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}$. Since F_0 is a symmetric function and $1-F_0(t)=F_0(-t)$, above equation could be simplified to:

$$\frac{1}{2}+\frac{\sqrt{2}}{4}\mathrm{PLC}(\lambda)=\frac{1}{2}\left(F_0\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{2\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}\right)+G_0\left(\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{2\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}\right)\right)$$

Maximizing PLC is equivalent to maximizing $\frac{\lambda^{\mathrm{T}}(\mu_x-\mu_y)}{2\sqrt{\lambda^{\mathrm{T}}\Sigma\lambda}}$ since both F_0 and G_0 are non-decreasing. From Seber (1977), the solution of this optimal question is $\hat{\lambda}_{max}=\Sigma^{-1}(\mu_x-\mu_y)$. The corresponding maximized PLC is given by:

$$\begin{aligned}\frac{1}{2}+\frac{\sqrt{2}}{4}\mathrm{PLC}_{max}&=\frac{1}{2}\left(F_0(a)+G_0(a)\right)\\ \Longrightarrow\ \mathrm{PLC}_{max}&=\sqrt{2}\left(F_0(a)+G_0(a)\right)-\sqrt{2}\end{aligned}$$

where $a=\frac{\sqrt{(\mu_x-\mu_y)^T\Sigma^{-1}(\mu_x-\mu_y)}}{2}$ □

It is obvious that when biomarkers are elliptical umimodal in case and control populations with the same scalar matrices and generating variates, the best linear combination coefficient λ which maximizes the combined AUC could also maximize the combined PLC. In the following section, we will show some simulation results to verify this.

In addition, from equation 6.3.1 we can see that best linear combined PLC only depends on the Mahalanobis distance of elliptical distributions ($\sqrt{(\mu_x-\mu_y)^T\Sigma^{-1}(\mu_x-\mu_y)}$) and the populations of case and control group (Mitchell and Krzanowski, 1985). Table 6.1 indicates the different best linear combined PLC values under different Mahalanobis distances and distribution populations. The * in table 6.1 means any elliptical distribution types. We can see that no matter which distribution types that the biomarkers follow, the PLC will converge to $\sqrt{2}$ as the Mahalanobis distance of case and control elliptical populations goes to infinity.

Combined PLC for Elliptical Distributions with Equal Covariance Matrix			
Mahalanobis Distance	Case Group	Control Group	PLC
0	*	*	0
1	Multivariate Normal	Multivariate Normal	0.5415
	Multivariate T_3	Multivariate T_{11}	0.5103
	Multivariate T_4	Multivariate T_6	0.5104
	Multivariate Laplace	Multivariate Laplace	0.7169
	Multivariate T_4	Multivariate Laplace	0.6107
	Multivariate T_{11}	Multivariate Laplace	0.6223
	Multivariate Laplace	Multivariate T_5	0.6142
2	Multivariate Normal	Multivariate Normal	0.9655
	Multivariate T_3	Multivariate T_{11}	0.8982
	Multivariate T_4	Multivariate T_6	0.8982
	Multivariate Laplace	Multivariate Laplace	1.0704
	Multivariate T_4	Multivariate Laplace	0.9779
	Multivariate T_{11}	Multivariate Laplace	1.0027
	Multivariate Laplace	Multivariate T_5	0.9855
3	Multivariate Normal	Multivariate Normal	1.2253
	Multivariate T_3	Multivariate T_{11}	1.1368
	Multivariate T_4	Multivariate T_6	1.1368
	Multivariate Laplace	Multivariate Laplace	1.2447
	Multivariate T_4	Multivariate Laplace	1.1824
	Multivariate T_{11}	Multivariate Laplace	1.2151
	Multivariate Laplace	Multivariate T_5	1.1923
5	Multivariate Normal	Multivariate Normal	1.3966
	Multivariate T_3	Multivariate T_{11}	1.3313
	Multivariate T_4	Multivariate T_6	1.3341
	Multivariate Laplace	Multivariate Laplace	1.3730
	Multivariate T_4	Multivariate Laplace	1.3464
	Multivariate T_{11}	Multivariate Laplace	1.3727
	Multivariate Laplace	Multivariate T_5	1.3551
10	Multivariate Normal	Multivariate Normal	1.4142
	Multivariate T_3	Multivariate T_{11}	1.4030
	Multivariate T_4	Multivariate T_6	1.4072
	Multivariate Laplace	Multivariate Laplace	1.4130
	Multivariate T_4	Multivariate Laplace	1.4083
	Multivariate T_{11}	Multivariate Laplace	1.4133
	Multivariate Laplace	Multivariate T_5	1.4107
Infinity	*	*	1.4142

Table 6.1: Combined PLC for Different Distributions and Mahalanobis Distances

6.4 Simulation

In this section, we will demonstrate that the equation 6.3.1 could obtain the optimal linear combined PLC through simulation.

Consider the linear combination of d biomarkers. Suppose there are n independent observations $(X_1,\ X_2,\ \ldots, X_n) \sim \mathcal{E}_d(\mu_x, \Sigma; \phi_x)$ in the case group, and m independent observations $(Y_1,\ Y_2,\ \ldots, Y_m) \sim \mathcal{E}_d(\mu_y, \Sigma; \phi_y)$ in the control group. The density generators in both groups are monotonically non-increasing which means these biomarkers are elliptical unimodal distributed. Following are main steps to get the empirical estimation of best linear combined PLC.

1. For any $\lambda \in \mathbb{R}^d$ and cut point value $c \in \mathbb{R}$, let $n_c = \sum_{i=1}^{n} I(\lambda^{\mathrm{T}} X_i > c)$ and $m_c = \sum_{j=1}^{m} I(\lambda^{\mathrm{T}} Y_j > c)$, where I is the indicator function.
2. Estimate the ROC curve by the equation 1.1.4.
3. In general the empirical ROC curve is composed of $m+n+1$ points including the coordinates (0,0) and (1,1). Smooth the ROC curve (Dümbgen and Rufibach, 2011). Suppose the coordinates of smoothed ROC curve is denoted as $(u_i,\ v_i)$ where $i \in (1,\ 2,\ \ldots,\ m+n+1)$.
4. For $i \in (2,\ 3,\ \ldots,\ m+n+1)$, let $(u^*_{i-1},\ v^*_{i-1}) = (u_i,\ v_i) - (u_{i-1},\ v_{i-1})$.
5. $\mathrm{PLC}(\lambda) = \sum_{i=1}^{m+n} \left| (u^*_i,\ v^*_i) \cdot \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right) \right|$.
6. The best linear coefficient is the one which could maximize all possible combined PLC, i.e. $\hat{\lambda}_{max} = \underset{\lambda}{\mathrm{argmax}} \mathrm{PLC}(\lambda)$

The reason of smoothing in step 3 is that the PLC of a stepwise ROC curve is always $\sqrt{2}$. Here we use the Log-Concave smooth method proposed by Dümbgen and Rufibach (2011).

6.4.1 Multivariate Normal Distribution

Firstly, we begin with the simplest condition - multivariate normal distribution. Similar with the setting of simulation in chapter 5, assume three biomarkers follow multivariate normal distribution in the case and control group:

$$X \sim MN_y(\mu_x, \Sigma), \quad Y \sim MN_y(\mu_y, \Sigma)$$

$$\text{Let } \mu_x = \begin{pmatrix} 2 \\ 1.6 \\ 1.3 \end{pmatrix}, \mu_y = \begin{pmatrix} 1.4 \\ 1 \\ 1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.15 & 0.1 \\ 0.15 & 0.26 & 0.02 \\ 0.1 & 0.02 & 1 \end{pmatrix}.$$

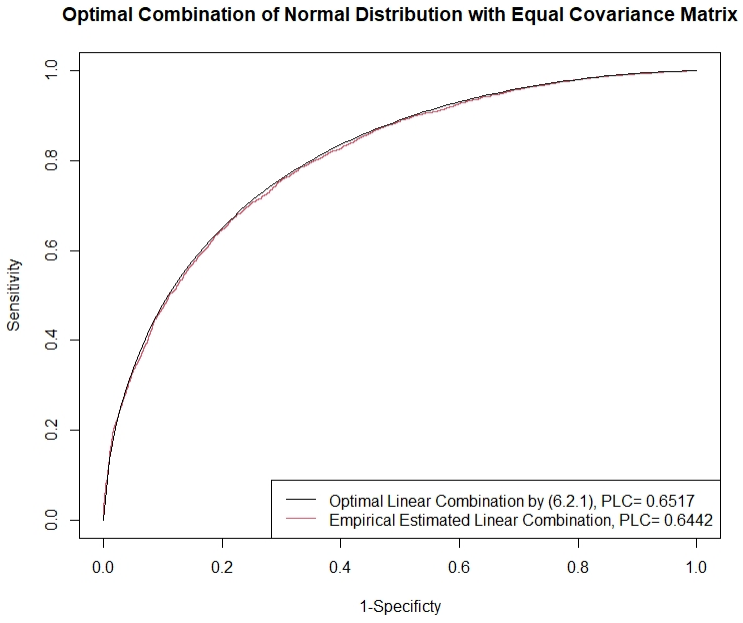

Figure 6.3: Projected Length of the Normal Distribution ROC Curve

Generate 5000 observations in each group and estimate the empirical linear combined ROC curve which could maximize the PLC. Figure 6.3 shows the simulation results. The empirical ROC curve (red line) which almost coincides with the ROC curve obtained by equation 6.2.1 (black line). The empirical PLC estimation is 0.6442 which is close to 0.6517 but slightly underestimated.

6.4.2 Multivariate T Distribution

Now assume that the biomarkers in case and control group both follow multivariate t distributions. The parameters are same with the simulation of multivariate t distribution in section 5.3.2.

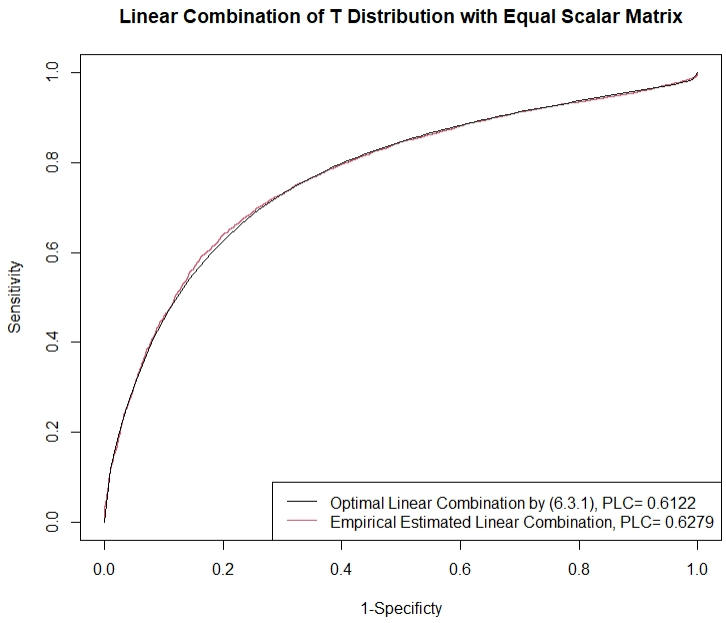

Figure 6.4: Projected Length of the Multivariate T Distribution ROC Curve

The simulation results are shown in Figure 6.4. The empirical PLC estimation is 0.6279 which is close to 0.6122 but slightly overestimated.

6.4.3 Multivariate Laplace Distribution

Now assume that the biomarkers in the case and control group both follow multivariate Laplace distributions. The parameters are same with the simulation of multivariate Laplace distribution in section 5.3.2. The simulation results are illustrated in Figure 6.5. The empirical PLC estimation is 0.8312 which is shown as the red line and the PLC obtained by equation 6.3.1 is 0.8208 which is shown as the black line.

From previous simulations we can see that sometimes the smoothing will underesti-

mate the ROC curve while sometimes overestimate. But the errors are relatively small.

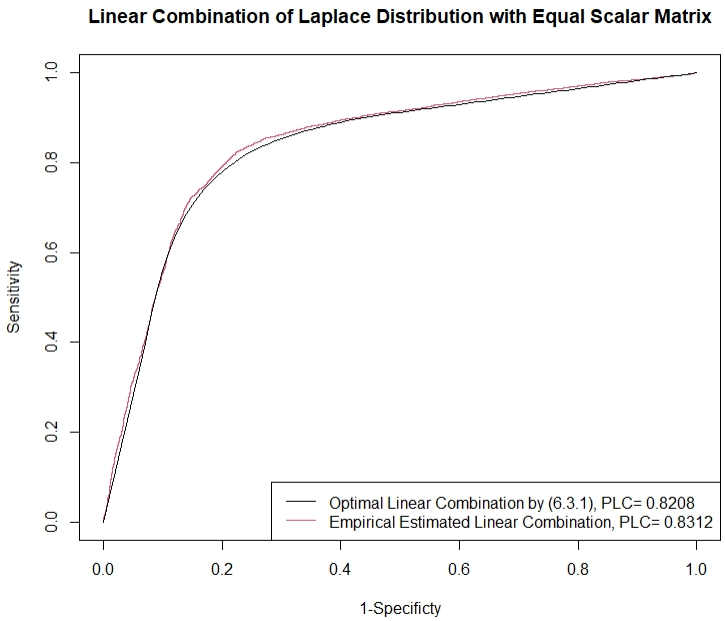

Figure 6.5: Projected Length of the Multivariate Laplace Distribution ROC Curve

Chapter 7

Application

We will apply the proposed methods to some case-control studies and compare with the results from other methods. The expected result is that our combinations could have larger AUC and PLC value than other combinations. At first we elucidate the application of the combination of biomarkers to the diagnosis of autism.

7.1 Biomarkers for Autism/Autistic Disorder

Autism/autism spectrum disorder (ASD) is a collection of developmental disorders that affects communication and stereotyped behavior, and increases irritability (Pinto-Martin et al., 2008). The diagnosing of autism/autism spectrum disorders (ASD) has gained higher attentions because of its diagnostic challenge. Failing to correctly diagnosing children with autism/ASD may lead to missed opportunities for appropriate treatments, while misdiagnosis of autism/ASD also has potentially harmful consequences for children (Falkmer et al., 2013).

In 2002-2005, the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development conducted a case-control study to investigate whether growth-related

hormones may play part, if any, in the diagnosis of autism for children 4-8 years old (Mills et al., 2007). The data consists of 59 subjects in the control group and 71 subjects in the case group. Each participant was assayed for multiple groups of diagnostics biomarkers, including insulin-like growth factors (IGF1, IGF2, IGFBP3), as well as dehydroepiandrosterone (DHEA) and DHEA sulfate (DHEAS) and so on. Mills et al. (2007) described the details of the subject enrollment and data collection. We will combine five biomarkers including IGF1, IGF2, IGFBP3, DHEA and DHEAS in the data to diagnose the autism/ASD and calculate combined AUC values and PLC values.

The insulin-like growth factors (IGFs), as a group of biomarkers of autism/ASD, are parts of a complex system. They consist of two ligands: the Insulin-like growth factor 1 (IGF-1) and the Insulin-like growth factor 2 (IGF-2), some binding proteins including IGFBP3 and others. Evidence shows that IGF-1 is a growth factor required for achieving maximal growth while IGF-2 expression is required for early development (Gunnell et al., 2005). Since the binding of IGFBP-3 to cells will reduce its own affinity, IGF-1 combined with cell-related IGFBP-3 is in a more dominant beneficial balance with the receptor (Jameson and Groot, 2015). Dehydroepiandrosterone sulfate (DHEAS) is a metabolite of dehydroepiandrosterone (DHEA) that circulates in far greater relative concentrations. Both DHEAS and DHEA are produced in the zona reticularis of the adrenal cortex under the control of adrenocorticotropic hormone (ACTH) (Prough et al., 2016).

We compare diagnosis performances of different combinations of these five biomarkers, including likelihood ratio combination and linear combination. Suppose the observations are elliptically distributed in case and control group, denoted as $(X_1,\ X_2,\ \ldots, X_{71}) \sim \mathcal{E}_5\left(\mu_x, \Sigma_x; \phi_x\right)$ and $(Y_1,\ Y_2,\ \ldots, Y_{59}) \sim \mathcal{E}_5\left(\mu_y, \Sigma_y; \phi_y\right)$ respectively.

Firstly, we assume the observations follow multivariate normal distribution in the case and control group. We use the sample mean and sample variance to estimate the mean and covariance matrix. The best normal linear combination coefficient obtained by the equation

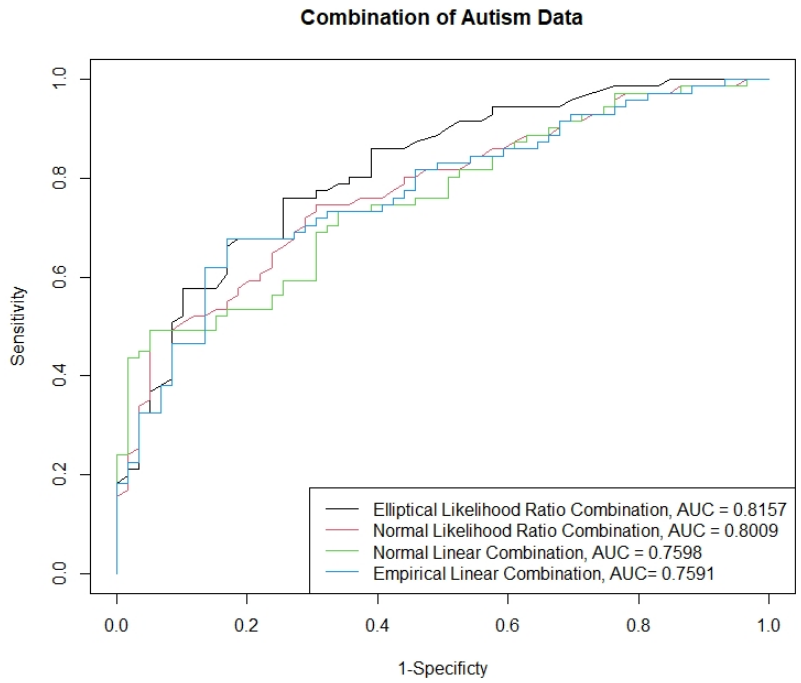

Figure 7.1: Combinations of Autism Data and Corresponding AUC Values

1.3.5 is $\hat{\lambda}_{normal} = (-0.0059,\ 0.0158,\ 7.4308,\ -0.0193,\ 0.1214)^{\mathrm{T}}$ and corresponding AUC value is 0.7598. The empirical linear combination coefficient obtained by equation 1.3.4 is $\hat{\lambda}_{normal} = (0.0101,\ 0.0186,\ 0.0117,\ -0.0041,\ 0.0100)^{\mathrm{T}}$ and corresponding AUC value is 0.7591. It is obvious that the normal linear combined AUC value is close to the empirical linear combination. However, if we assume elliptical distributions and use the equations 3.3.1 and 3.3.2 to estimate the empirical density generators of case and control group, the empirical likelihood ratio combination AUC obtained by theorem 4.1.1 is 0.8157 which is significantly improved.

Figure 7.1 shows the ROC curves of different combinations. The ROC curve of elliptical likelihood ratio combination (black line) has the largest AUC value. In addition, the black line is obviously higher than all other ROC curves when the specificity is small. We then smooth ROC curves and compare the PLC values. Figure 7.2 shows the smoothed ROC curves of different combinations. The smoothing process almost doesn't change the

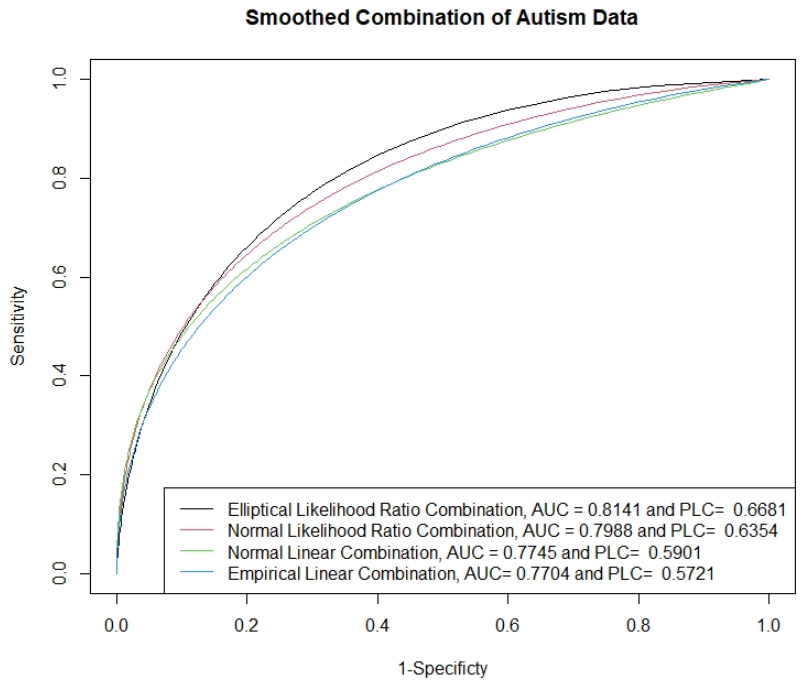

Figure 7.2: Smoothed Combinations of Autism Data and Corresponding AUC and PLC Values

AUC values. The PLC value of normal linear combination is 0.5901 which is slightly larger than the empirical linear combination 0.5721. The PLC of elliptical likelihood ratio combination is 0.6681 which is obviously larger than the normal likelihood ratio combination which is 0.6354. Therefore, according to the combined AUC and PLC values, we can safely conclude that the elliptical combination significantly improved the diagnosis performance of biomarkers related to the autism.

7.2 Biomarkers for Neural Tube Defects (NTDs)

Another example is diagnosing neural tube defects (NTDs) of the infants. Neural tube defects (NTDs) are a group of birth defects including spina bifida, anencephaly and others (National Institute of Child Health and Human Development, 2017). NTDs are one of the most common birth defects. Maternal biomarkers play a crucial role in early diagnosing

and prevention of NTDs.

A study of the Irish National Rubella Screening Program collected data from blood samples from pregnant women and recorded their neural tube defects. We choose four biomarkers related to the diagnosis of NTDs: vitamin B12, reduced red cell folate, plasma total folate and free choline. After deleting missing values, there are 71 observations in the case group and 230 observations in the control group. At first, we test the multivariate normality of data. Both p-values of Shapiro-Wilk test of case and control observations are less than 0.05 and close to 0, so we can conclude that the date are not multivariate normal distributed (Shapiro and Wilk, 1965).

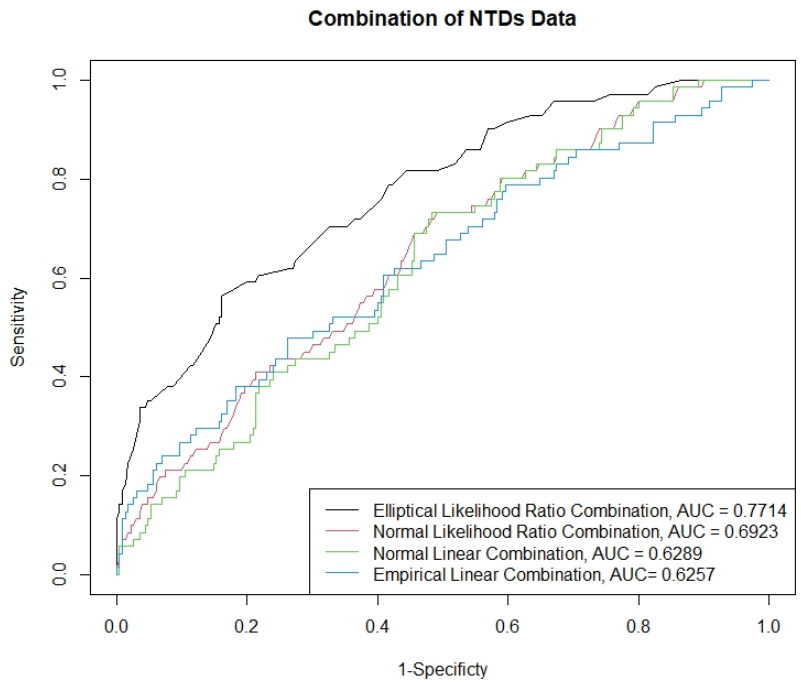

Figure 7.3: Combinations of NTDs Data and Corresponding AUC Values

Figure 7.3 shows the combination results. The best normal linear combination coefficient obtained by the equation 1.3.5 is $\hat{\lambda}_{normal} = (-0.0395,\ 0.0020,\ 0.0214,\ 0.8176)^{\mathrm{T}}$ and corresponding AUC value is 0.6289. The empirical linear combination coefficient obtained by equation 1.3.4 is $\hat{\lambda}_{normal} = (-0.2593,\ 0.9770,\ -0.3823,\ -0.7983)^{\mathrm{T}}$ and correspond-

ing AUC value is 0.6257. Although the normal linear combined AUC value is close to the empirical linear combination, we can see that the empirical ROC curve (blue line) is higher than the normal linear combination ROC curve (green line) when the specificity is large. The AUC value of normal likelihood ratio combination is 0.6923 which is larger than the linear combinations, but the improvement is small. If we assume elliptical distributions and use the equations 3.3.1 and 3.3.2 to estimate the empirical density generators of case and control group, the empirical likelihood ratio combination AUC obtained by theorem 4.1.1 is 0.7714 which is significantly improved. In addition, the elliptical likelihood ratio combination ROC curve (black line) is higher than all other combinations for any specificity value.

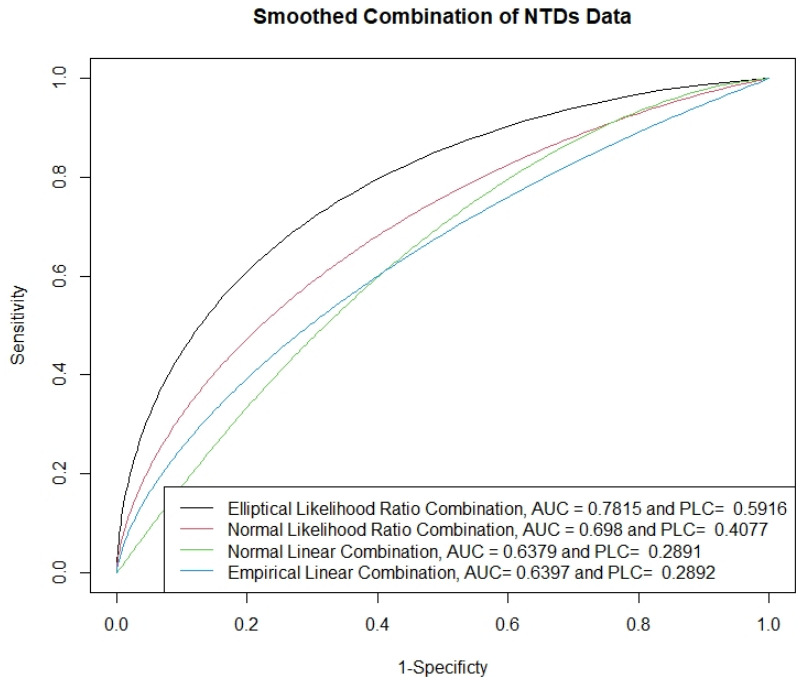

Figure 7.4: Smoothed Combinations of NTDs Data and Corresponding AUC and PLC Values

Figure 7.4 shows the smoothed ROC curves of different combinations. The smoothing process almost doesn't change the AUC values. The PLC value of normal linear combination is 0.2891 which is almost equivalent to the empirical linear combination 0.2892. The

PLC of elliptical likelihood ratio combination is 0.5916 which is obviously larger than the normal likelihood ratio combination which is 0.4077. Therefore, according to the combined AUC and PLC values, we can safely conclude that the elliptical combination significantly improved the diagnosis performance of biomarkers related to the NTDs.

Chapter 8

Discussion and Future Research

8.1 Summary of Results

In chapter 2, we discuss some properties and features of the multivariate normal combination. We find that if the case and control group have different covariance matrices for the binormal model, there would always exist a segment of ROC curve which is lower than the chance line. Through the properties and estimation methods introduced in the chapter 3, the elliptical likelihood ratio combination and linear combination of multiple biomarkers are proposed. The likelihood ratio combination is optimal but generally complicated to calculate. To solve this, we describe the empirical estimation of elliptical likelihood ratio combination without specifying exact distribution types. Simulation results show that the empirical ROC curves of likelihood ratio combination are close to the true combination ROC curves.

We also derive the ROC curve function of linear combination of biomarkers under the elliptical distribution. In chapter 5 and chapter 6, we develop the method to estimate the linear combination coefficient according to maximizing the AUC value and PLC value. The best linear coefficient of PLC is as same as the combination coefficient of AUC under ellip-

tical unimodal assumption. In addition, we find that the best linear combination coefficient only depends on the Mahalanobis distance of elliptical distributions. We then apply proposed methods to two case-control studies. One example is the diagnosing of autism/ASD and another one is the diagnosing of NTDs. The application results show that our methods could significantly improve the diagnostic performance of multiple biomarkers.

8.2 Future Research

8.2.1 Alternative Index - Youden's index

The simulation results in chapter 5 show the drawbacks of AUC value. To offset the disadvantages of auc value, we develop the combination of PLC values. However, the PLC index still has shortcomings. Future research is needed to strengthen and improve the combination performance.

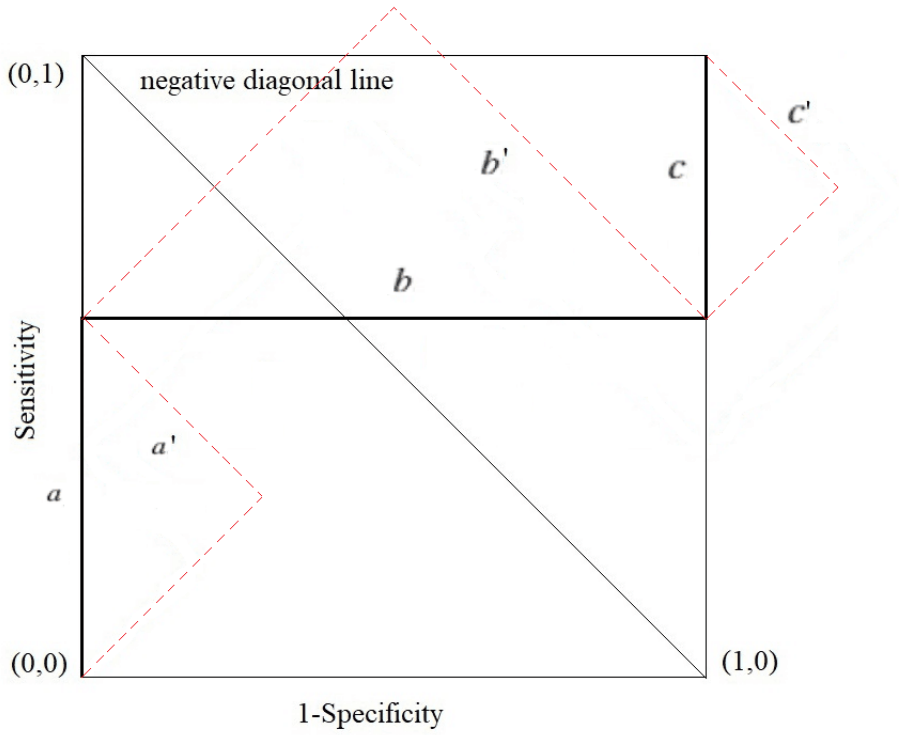

Figure 8.1: An Example of Drawbacks of PLC

Figure 8.1 represents an extreme example of the drawbacks of PLC index. There is a ROC curve which is composed of three segments of straight lines a, b, and c. Then The PLC index is the summation of all of the projected lengths of these three segments, i.e. $(a' + b' + c')$. It's easy to show that $a' + b' + c' = \sqrt{2}$ which is the maximum value of PLC index. In fact, for any stepwise ROC curve, the PCL would always be $\sqrt{2}$. And the PLC is more sensitive than the AUC value for detecting diseases.

Except for the AUC and PLC, another measurement index which could be obtained from the ROC curve is the Youden's index defined as $J = sensitivity + specificity - 1$. Its value ranges from 0 through 1 (inclusive). Yin and Tian (2014) proposed empirical methods for the optimal linear combination of Youden's index. They also developed the best linear combination under multivariate normal distribution. It is feasible to extend the linear combination to the elliptical distribution. Following theorem we derived expresses the best linear combination of Youden's index when case and control population have the same scalar matrix.

Theorem 8.2.1. *Suppose $X \sim \mathcal{E}_d(\mu_x, \Sigma, \phi_x)$, $Y \sim \mathcal{E}_d(\mu_y, \Sigma, \phi_y)$ are distribution of case and control group, respectively. Here two groups have a same scalar matrix. In addition, X and Y have the same generating variate R, then the best λ which could maximize the linear combined Youden's Index value is $\hat{\lambda}_{max} = \Sigma^{-1}(\mu_x - \mu_y)$.*

Up to now, the main limitation is the requirement of equivalent scalar matrix and generating variate. Further research on the linear combination of multiple biomarkers, which follow elliptical distribution with different scalar matrices and generating variates in case and control group, is needed.

8.2.2 Comparing Two Groups of Biomarkers

Considerable amount of research has been devoted to the comparison of two or more biomarkers in terms of their ROC curves or AUC values. Methods have been developed to

test the equality of two or more AUC values, both parametrically or nonparametrically. Especially under normal distributed assumptions, we could derive a Likelihood Ratio Test for two biomarkers in our study. However, the issue of comparing the diagnostic performance between two groups of biomarkers, where biomarkers in the same group share common or similar biological functions, has not been addressed in literature.

Suppose two groups of biomarkers are available, $\{M_1,\ldots,M_k\}$ and $\{M_{k+1},\ldots,M_p\}$. We are interested in whether, as a whole, the first group of biomarkers has higher diagnostic accuracy than the second group. We propose addressing this issue by testing the hypotheses:

$$H_0 : AUC_1 = AUC_2, H_1 : AUC_1 > AUC_2$$

where AUC_1 is the linear combination AUC of the first group of biomarkers, and AUC_2 is the linear combination AUC of the second group of biomarkers.

Let $X_i = (X_{1i}, X_{2i}, \ldots, X_{ki}, X_{k+1}, \ldots, X_{pi})\,(i = 1,2,\ldots,n)$ be the observed values from the biomarkers from a random sample of size n in the diseased population, and $Y_i = \left(Y_{1j}, Y_{2j}, \ldots, Y_{kj}, Y_{k+1j}, \ldots, Y_{pj}\right)\,(j = 1,2,\ldots,m)$ be the observed values from the biomarkers from a random sample of size m in the non-diseased population. According to empirical estimation of AUC, we have:

$$\begin{aligned} \hat{AUC}_1 &= \frac{1}{mn}\sum_{i,j} I\left(\lambda_1 Y_1 + \lambda_2 Y_2 + \ldots + \lambda_k Y_k < \lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k\right) \\ \hat{AUC}_2 &= \frac{1}{mn}\sum_{i,j} I\left(\lambda_{k+1} Y_{k+1} + \ldots + \lambda_p Y_p < \lambda_{\mathrm{k}+1} X_{k+1} + \ldots + \lambda_p X_p\right) \end{aligned} \tag{8.2.1}$$

Particularly, we can use the proposed method to estimate the linear combination AUC under multivariate elliptical distribution. After estimating the linear combination AUC of the two groups of biomarkers, we can calculate:

$$T = \frac{\hat{AUC}_1 - \hat{AUC}_2}{\mathrm{se}\left\{\hat{AUC}_1 - \hat{AUC}_2\right\}} \tag{8.2.2}$$

If $|T| > 1.96$, we will reject the null hypothesis at the significant level of 0.05.

In order to evaluate the performance of a test, we need to measure two aspects: the probability of type I error and the power of the test.

Simulation under Multivariate Normal Condition

If the joint distribution of two group biomarkers follows the multivariate normal distribution, the test statistics which using the estimator from equation 1.3.6 should be better than the empirical estimator since the linear combination is the global optimal combination under the normality assumption.

Type I Error Control

Firstly, we simulate the test to see whether it can control the type I error. Consider three biomarkers, which follow two multivariate normal distributions with different mean vectors and the same covariance matrix:

$$\mu_x = \begin{pmatrix} 0.5 \\ 0.7 \\ 0.343848403 \end{pmatrix}, \ \mu_y = \begin{pmatrix} 0.15 \\ 0.25 \\ 0 \end{pmatrix}, \ \Sigma_x = \Sigma_y = \begin{pmatrix} 1 & 0.22 & 0.0001 \\ 0.22 & 1 & 0.0001 \\ 0.0001 & 0.001 & 0.44 \end{pmatrix}$$

representing the non-diseased and diseased population, respectively. From equation 1.3.6, we can get $AUC_1 = AUC_2 = 0.64301947$, so the null hypothesis is true in this setting. Now we wonder whether the first two biomarkers are better than the third one $(k = 2, p = 3)$. Creating a test:

$$H_0 : AUC_1 = AUC_2, \ H_1 : AUC_1 > AUC_2$$

We can calculate the test statistics from 8.2.2:

$$T = \frac{\hat{AUC}_1 - \hat{AUC}_2}{\text{se}\left\{\hat{AUC}_1 - \hat{AUC}_2\right\}}$$

If $|T| > 1.96$, we will reject the null hypothesis at the significant level of 0.05 otherwise, we can't reject the null hypothesis.

Generating 200 random samples from the non-diseased population, that is, $m = n = 200$. Using Jackknife method, we can estimate the standard error of $\{\hat{AUC}_1 - \hat{AUC}_2\}$

Repeating this process 500 times and getting the normal linear combination estimation of AUC by function 1.3.6, the results show that in 500 repeats, 94.4% absolute values of test statistics are less than 1.96. So the probability of type I error is 5.6%.

If using function 1.3.4 to estimate the empirical linear combination test statistics, the results show that 94% absolute values of test statistics are less than 1.96 in 500 repeats. So the probability of type I error is 6%, which is a little bit higher than that from multivariate normal estimation.

Therefore, we can conclude that this test based on the linear combination can control type I error effectively.

Power of the Test under Linear Combination

Now we consider the power of the test. Suppose the two multivariate normal distributions have mean vectors and a covariance matrix as:

$$\mu_x = \begin{pmatrix} 0.5 \\ 0.7 \\ 0.343846403 \end{pmatrix}, \ \mu_y = \begin{pmatrix} 0.3 \\ 0.4 \\ 0 \end{pmatrix}, \ \Sigma_x = \Sigma_y = \begin{pmatrix} 1 & 0.22 & 0.0001 \\ 0.22 & 1 & 0.0001 \\ 0.0001 & 0.001 & 0.44 \end{pmatrix}$$

From function 1.3.6, we can get $AUC_1 = 0.59224055$, $AUC_2 = 0.7211637$, so the alternative hypothesis is true in this setting.

Creating a test:

$$H_0 : AUC_1 = AUC_2,\ H_1 : AUC_1 > AUC_2$$

We still use the Jackknife method to calculate the test statistics. Generating 200 random samples from the non-diseased population, and 200 random samples from the diseased population.

Repeating this process 500 times, the results show that in 500 repeats of multivariate normal estimation, 93.4% absolute values of test statistics are greater than 1.96 . So the power of this test is 0.934. If using empirical linear combination function 1.3.4 to estimate the test statistics, we can find that the power of this test is 0.91.

Therefore, the test using multivariate normal estimation has a slightly higher power than the empiric function when the biomarkers follow multivariate normal distribution.

In conclusion, the numerical test statistics obtained from Jackknife method could control type I error and have great power. However, more studies about the theoretical test statistics are needed, especially under the elliptical distribution.

Bibliography

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristics graph. *Journal of Mathematical Psychology*, 12:387–415.

Barry, M. J. (2001). Clinical practice. prostate-specific-antigen testing for early diagnosis of prostate cancer. *The New England journal of medicine*, 344(18):1373–1377.

Bhattacharyya, S. and Bickel, P. (2013). Adaptive estimation in elliptical distributions with extensions to high dimensions.

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, 69(3):89–95.

Chia, N. and Nakano, J. (2008). M-decomposability, elliptical unimodal densities, and applications to clustering and kernel density estimation. *arXiv: Methodology*.

Dümbgen, L. and Rufibach, K. (2011). logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software, Articles*, 39(6):1–28.

Falkmer, T., Anderson, K., Falkmer, M., and Horlin, C. (2013). Diagnostic procedures in autism spectrum disorders: A systematic literature review. *European Child and Adolescent Psychiatry*, 22:329–340.

Fang, K., Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related distributions*. New York: Chapman and Hall.

Frahm, G. (2004). Generalized elliptical distributions: Theory and applications.

Gunnell, D., Miller, L. L., Rogers, I., Holly, J. M. P., and Team, A. S. (2005). Association of insulin-like growth factor i and insulin-like growth factor-binding protein-3 with intelligence quotient among 8- to 9-year-old children in the avon longitudinal study of parents and children. *Pediatrics*, 116(5):e681–e686.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36.

Hult, H. and Lindskog, F. (2002). Lindskog, f.: Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34:587–608.

Jameson, J. L. and Groot, L. D. (2015). *Endocrinology: Adult and Pediatric, 2-Volume Set*. W.B. Saunders, Philadelphia, 7th edition.

Kermali, M., Khalsa, R. K., Pillai, K., Ismail, Z., and Harky, A. (2020). The role of biomarkers in diagnosis of covid-19 – a systematic review. *Life sciences*, 254:117788.

Kotz, S., Kozubowski, T., and Podgorski, K. (2001). *The Laplace Distribution and Generalizations*.

Kotz, S. and Nadarajah, S. (2004). *Multivariate T-Distributions and Their Applications*. Cambridge University Press.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*. New York: Chapman and Hall, 1st edition.

Lee, W.-C. and Hsiao, C. K. (1996). Alternative summary indices for the receiver operating characteristic curve. *Epidemiology*, 7(6):605–611.

Liu, A., Schisterman, E. F., and Zhu, Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, 24:37–47.

Liu, C., Liu, A., and Halabi, S. (2011). A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, 30:2005–2014.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics*, 58(3):657–664.

Mills, J., Hediger, M., Molloy, C., Chrousos, G., Manning-Courtney, P., Yu, K., Brasington, M., and England, L. (2007). Elevated levels of growth-related hormones in autism and autism spectrum disorder. *Clinical endocrinology*, 67:230–237.

Mitchell, A. F. S. and Krzanowski, W. J. (1985). The mahalanobis distance and elliptic distributions. *Biometrika*, 72(2):464–467.

National Institute of Child Health and Human Development (2017). About neural tube defects (ntds). `https://www.nichd.nih.gov/health/topics/ntds/conditioninfo/`.

Neyman, J. and Pearson, E. S. (1993). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337.

Pepe, M. and Thompson, M. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, 1:123–40.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62:221–229.

Pinto-Martin, J. A., Young, L. M., Mandell, D. S., Poghosyan, L., Giarelli, E., and Levy, S. E. (2008). Screening strategies for autism spectrum disorders in pediatric primary care. *Journal of Developmental and Behavioral Pediatrics*, 29(5):345–350.

Prough, R. A., Clark, B. J., and Klinge, C. M. (2016). Novel mechanisms for dhea action. *Journal of Molecular Endocrinology*, 56(3):R139–R155.

Seber, G. A. F. (1977). *Linear Regression Analysis*. Wiley Press, New York.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88:1350 –1355.

Tyler, D. E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika*, 70:411–420.

Tyler, D. E. (1987). A distribution-free m-estimator of multivariate scatter. *The Annals of Statistics*, 15(1):234–251.

World Health Organization (2011). *Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus: Abbreviated Report of a WHO Consultation*. World Health Organization.

Yin, J. and Tian, L. (2014). Optimal linear combinations of multiple diagnostic biomarkers based on youden index. *Statistics in medicine*, 33(8):1426–1440.

Zhang, L. and Guo, H. (2020). Biomarkers of covid-19 and technologies to combat sars-cov-2. *Advances in biomarker sciences and technology*, 2:1–23.

Zhang, X., Walter, S. D., and Agnihotram, R. V. (2009). Properties of the projected length of the curve (plc) and area swept out by the curve (asc) indices for the receiver operating characteristic (sroc) curve. *The International Journal of Biostatistics*, 5(1).