

**USING THE EXPONENTIAL TILTING MODEL AND
THE MONOTONIC DENSITY RATIO MODEL TO FIND
THE EMPIRICAL DISTRIBUTION FOR PARAMETERS**

AUTHOR: HUANG YI (A0123864A)

SUPERVISOR: ASSOC. PROF YU TAO

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

**DEPARTMENT OF STATISTICS AND APPLIED
PROBABILITY**

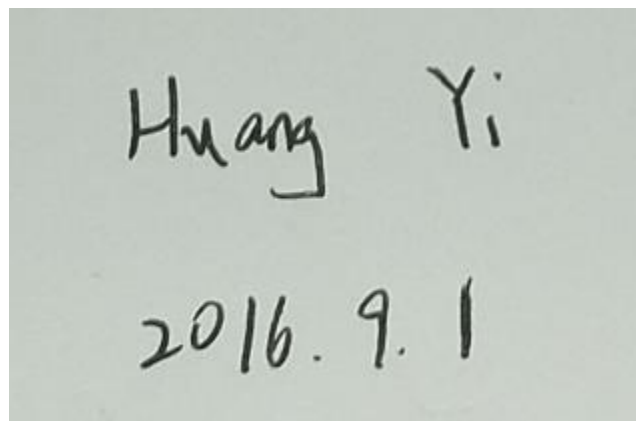
NATIONAL UNIVERSITY OF SINGAPORE

2016.8.31

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.



Huang Yi
2016.9.1

Acknowledgement

I would like to extend my sincere thanks to everyone who has provided me with support and guidance throughout my working on this thesis.

First and foremost, I would like to express my appreciation to my supervisor Professor Yu Tao. He offered me such an interesting topic to work on and guided me step by step to complete this project. His idea and knowledge always inspired me a lot.

I also feel grateful to the department to give me this opportunity to study a project by myself. It was quite a beneficial experience for me.

Abstract

With the development of technology, the clinical study related to disease detection has become a popular topic for biostatisticians. For those studies, the combination of multiple diagnostic tests seems to be the most general tool to achieve optimal sensitivity and specificity. In this thesis, we apply the exponential tilting model proposed by Qin and Zhang (2010) and the monotonic density ratio model proposed by Chen et al. (2015) to estimate the asymptotic distribution for β and AUC using three bootstrapping methods and give an evaluation and comparison towards their performances. We give a good estimation for the distribution of β no matter whether the robustness is taken into consideration. And we also have a good thinking for estimating the distribution of AUC .

Keywords: Multiple diagnostic tests; Monotonic density ratio model; Exponential tilting model; ROC curve; AUC; Sensitivity; Specificity; Asymptotic distribution estimation; Bootstrapping

TABLE OF CONTENTS

Declaration	1
Acknowledgement	2
Abstract	3
Chapter 1 Introduction	1
Chapter 2 Basic concepts	6
2.1 ROC and AUC	6
2.2 Bootstrapping and resampling.....	8
Chapter 3 Optimal combination of multiple diagnostic tests	11
3.1 Optimal combination based on exponential titling model	11
3.2 Optimal combination based on semiparametric monotonic density ratio model	14
Chapter 4 Method description	20
Chapter 5 Simulation studies	23
5.1 Biomarkers follow exponential distribution.....	23
5.2 Distribution estimation and comparison	25
Chapter 6 Conclusion and discussion	37
Reference	39

Chapter 1

Introduction

With the medical technology developing rapidly, people become more and more enthusiastic to the clinical study related to disease detection in the early stage when the patient in fact doesn't have any obvious symptoms yet. This kind of early detection, or in other words, screening, grows to be a popular topic for biostatisticians because these trials can find the possible disease earlier with less cost and significantly reduce the death rate.

Among those clinical researches, the studies about biomarkers towards disease detections and classifications seem to be the essential part and have considerable research results according to Henson et al. (1999) and Srinivas et al. (2001). However, those tumor biomarkers, for example, CA-125 for diagnosing ovarian cancer, are accurately not perfect in performance for disease detecting. Many diseased individuals may have normal tumor biomarker concentrations, causing false negative diagnostic tests, while many non-diseased individuals may also have strange biomarker concentrations, leading to unnecessary diagnostic work-up and possible further treatments. Therefore, diagnostic tests, especially multiple diagnostic tests, are usually used for screening and diagnostic program, while the specificity and sensitivity of a

single diagnostic test cannot meet the researchers' needs in practice. And the combination of multiple diagnostic tests can further increase the accuracy of testing and then obtain an optimal testing method, so that those hidden diseases can be detected earlier and those diseased and non-diseased individuals can be distinguished more easily and accurately.

In clinical practice, many diagnostic testing methods are available for detecting possible diseases. And researchers can find that different diagnostic tests are sensitive to different aspects of the disease. Therefore, in recent several years, researchers are keen to find the best combination of multiple diagnostic tests for different assumptions. Kay and Little (1987) discussed various versions of the density ratio model by transforming the variables in the logistic regression model for binary data. If they assume that the variable satisfies a multivariate normal distribution, Su and Liu (1993) found the best way to combine different multiple diagnostic tests. By applying the Neyman-Pearson fundamental lemma, Eguchi and Copas (2002), Copas and Corbett (2002), and McIntosh and Pepe (2002) found the best combination of multiple diagnostic tests using log density ratio statistic for diseased and non-diseased individuals. Etzioni et al. (2003) took logistic combinations towards biomarkers to detect disease for cancers. Yuan and Ghosh (2008) came out with a novel model-combining algorithm when they combined multiple biomarker models in logistic regression. Liu and Zhou (2013) took the covariate adjustment into account and then studied the optimal combination in this case. Qin and Zhang (2010) assumed that the

diseased and non-diseased population has a log density ratio and proposed an exponential titling model as a combination of multiple diagnostic tests. Chen et al. (2015) considered a semiparametric monotonic model by directly modeling the density ratio as an unspecified monotonic non-decreasing function of a combination of multiple tests between those two groups of individuals. More development about the combination of multiple diagnostic tests can be found in the papers written by Barreno et al. (2008), and Kim et al. (2013). The last two models proposed by Qin and Zhang (2010) and Chen et al. (2015) will be introduced in detail and applied in future chapters.

And when the sampling distribution of the diagnostic variable is continuous, the ROC curve, or to be more precisely, the receiver operating characteristic curve, is one of the most widely used techniques for assessing the diagnostic accuracy in disease detection. To further classify the accuracy of the proposed method, *AUC* (the area under the ROC curve) is often applied for estimating the performance. Back to the whole research history, the estimations related to the ROC curve and the corresponding *AUC* are always based on a parametric model, a semiparametric model, or a fully nonparametric model. If we take some reading of the papers written by Begg (1991), Hsieh and Turnbull (1996), Zhou et al. (2002), Krzanowski and Hand (2009), Pepe (2003) and Zou et al. (2011), it is supposed to have a more comprehensive reviews of the development of the applications based on the ROC curve and the corresponding *AUC* in recent years.

This thesis is going to mainly discuss the application of the ROC curve to estimate the accuracy of the optimal combinations proposed by Qin and Zhang (2010) and Chen et al. (2015). Moreover, in Chen et al. (2015), they have proceeded to establish the convergence rate of $\boldsymbol{\beta}$ and the *AUC* for the proposed method. Here $\boldsymbol{\beta}$ is the parameter of the semiparametric monotonic density ratio model which satisfies:

$$\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} = \psi(v(\boldsymbol{x}, \boldsymbol{\beta}))$$

where $\psi(\cdot)$ is an unknown monotonic non-decreasing function, and $v(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^T h(\boldsymbol{x})$.

But in Chen et al. (2015), it only gave an estimation of the convergence rates for these two parameters, while their asymptotic distributions are actually still in mystery, inspiring us to have deeper exploration towards this problem. Therefore, this thesis will apply bootstrapping and resampling methods to show the empirical distributions for $\boldsymbol{\beta}$ and *AUC* and estimate the accuracy with different parameter settings, which is the main part of this thesis.

This thesis will be organized as follows. In Chapter 2, we will introduce some basic concepts which will be applied later in this thesis. In Chapter 3, we will describe those two optimal combinations of multiple diagnostic tests proposed by Qin and Zhang (2010) and Chen et al. (2015) and some related asymptotic results. In Chapter 4, we will describe the methods for estimating the asymptotic distribution of $\boldsymbol{\beta}$ and *AUC* in detail. In Chapter 5, we will do some simulation studies to give an evaluation of the methods of estimating distributions we have already described in Chapter 4, with those

biomarkers following bivariate exponential distribution. A conclusion remark and future discussion will be given in Chapter 6.

Chapter 2

Basic concepts

2.1 ROC and AUC

The ROC (receiver operating characteristic) curve is a graphical plot that can evaluate the performance of a diagnostic test. It was first used during the period of World War II when the United States army tried to analyze the radar signals and predict the routes of Japanese aircrafts. Its employment in signal detection extended to medicine in 1950s to assess human detection of weak signals according to Green and Swets (1966). The ROC curve was then extensively applied in medical research, epidemiology, machine learning and the evaluation of diagnostic tests, which we can find more details from Zweig and Campbell (1993). It also became a common technique to evaluate radiology technique.

The ROC curve can be created by plotting the true positive rate (TPR), also known as sensitivity, against the false positive rate (FPR), known as 1-specificity, at various threshold settings.

		Predicted condition	
		Predicted Condition positive	Predicted Condition negative
True condition	Condition positive	Ture positive rate (sensitivity)	False negative rate (Type II error)
	Condition negative	False positive rate (Type I error)	True negative rate (1-specificity)

Table 1. 2×2 contingency table

From Swets (2014), only the rates $TPR(T)$ and $FPR(T)$ are needed for plotting the ROC curve. The TPR is the ratio describing how many correct positive results occur among all positive samples appeared in the test. On the other hand, the FPR is the ratio describing how many incorrect positive results occur among all negative samples appeared in the test. For a single diagnostic test, if F and G denote the cumulative distribution functions for the test result in the diseased and non-diseased individuals, and f and g denote the probability density functions for them, then $TPR(T)$ and $FPR(T)$ can be represented as:

$$TPR(T) = \int_T^{\infty} f(x)dx$$

$$FPR(T) = \int_T^{\infty} g(x)dx$$

And the ROC curve has the following representation:

$$ROC(s) = 1 - F(G^{-1}(1 - s)), s \in (0,1).$$

As a classifier to measure the efficacy, *AUC* (the area under the curve) will often be applied, which is given by:

$$AUC = \int_0^1 ROC(s) ds.$$

If we define *AUC* directly from the rates $TPR(T)$ and $FPR(T)$, then it can be represented as:

$$\begin{aligned} AUC &= \int_{-\infty}^{\infty} TPR(T) FPR'(T) dT \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(T' > T) f(T') g(T) dT' dT = P(X_1 > X_0) \end{aligned}$$

where X_1 is the score for the positive distance and X_0 is the score for the negative distance. It should be noted that those two expressions above are equivalent in mathematics.

This ROC-AUC statistic is widely applied in many different fields nowadays. Sometimes researchers will link this statistic to a number of other performance metrics such as Brier score described in Hernández (2012) to reduce the noise when it is applied as a classification measure, which will not be considered in this thesis due to the application complexity.

2.2 Bootstrapping and resampling

Bootstrapping refers to some tests based on random sampling with replacement in statistics. This method was first published by Bradley Efron (1992). It is a popular and straightforward technology that can be applied to assign measures of accuracy, such as bias, variance, confidence intervals and prediction error for some complex estimators of complex parameters of the distribution. It may also be used for constructing hypothesis tests.

When bootstrapping is used to calculate confidence intervals for the population-parameter, we can first approximate the distribution by referring to the empirical distribution function of the observed data. In the case where the observations are assumed to be an independent and identically distributed dataset, this distribution can be obtained by constructing a number of resampling with replacement from the observed dataset. The size of the new dataset can be smaller or equal to the initial dataset. If we set the confidence level equal to α and using percentiles of the bootstrapping distribution, a confidence interval can be obtained as follows: $(\theta_{\alpha/2}^*, \theta_{1-\alpha/2}^*)$, where $\theta_{\alpha/2}^*$ denotes the $\alpha/2$ percentile of the bootstrapped coefficients θ^* .

Except this kind of classical way of bootstrapping, various alternatives are available for regression problems such as Bayesian bootstrapping, smooth bootstrapping, wild bootstrapping described by Wu (1986) and block bootstrapping described by Hernández (1989).

Bootstrapping method work well in those cases where the bootstrapping distribution is symmetrical and centered on the observed statistics and where the sample statistic is median-unbiased and has maximum concentration. And due to the development of computing power, it is recommended to increase the number of bootstrapping samples as many as possible, so that the effects of random sampling errors which arise from the bootstrapping method itself can be reduced.

Chapter 3

Optimal combination of multiple diagnostic tests

Suppose that there are K tests in total, with the k th test denoted as \mathfrak{X}_k for $k = 1, \dots, K$. Let $\{\mathbf{T}_1, \dots, \mathbf{T}_n\}$ denote the combined K -vector $\{\mathbf{X}_1, \dots, \mathbf{X}_{n_0}; \mathbf{Y}_1, \dots, \mathbf{Y}_{n_1}\}$, where $\mathbf{X}_1, \dots, \mathbf{X}_{n_0}$ represent independent and identically distributed results from the non-diseased individuals, and $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_1}$ represent independent and identically distributed results from the diseased individuals. Here $\mathbf{X}_i, \mathbf{Y}_i$ are all K -vector represented as $(\mathfrak{X}_1, \dots, \mathfrak{X}_K)^T$, and n_0, n_1 are the sample size for non-diseased and diseased individuals with $n = n_0 + n_1$, n refers to the size of the whole sample dataset. Let $D = 1$ and $D = 0$ refer to the diseased and non-diseased status, and let $F(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x} | D = 1)$ and $G(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x} | D = 0)$ represent the cumulative distribution functions of \mathbf{X}_i and \mathbf{Y}_i , $f(\mathbf{x})$ and $g(\mathbf{x})$ are the corresponding possibility density functions.

3.1 Optimal combination based on exponential titling model

In Qin and Zhang (2010), they consider an exponential titling model, which we will call it model (1) in short. For a given K in $\mathbf{X} = \mathbf{x}$, the conditional distribution $P(D = 1 | \mathbf{X} = \mathbf{x})$ is given by the logistic regression model:

$$P(D = 1 | \mathbf{X} = \mathbf{x}) = \frac{\exp(\alpha^* + \boldsymbol{\beta}^\tau h(\mathbf{x}))}{1 + \exp(\alpha^* + \boldsymbol{\beta}^\tau h(\mathbf{x}))}$$

And this model is equivalent to represent the density ratio as followed:

$$\frac{f(\mathbf{x})}{g(\mathbf{x})} = \exp(\alpha + \boldsymbol{\beta}^\tau h(\mathbf{x}))$$

where α is a scalar parameter, $\boldsymbol{\beta}$ is a $p \times 1$ vector parameter, and $h(\mathbf{x})$ is a $p \times 1$ smooth vector function of \mathbf{x} .

Here $U = \alpha + \boldsymbol{\beta}^\tau h(\mathbf{x})$ is the optimal combination. It can be evaluated by plotting the receiver of characteristic $ROC_C(s)$:

$$ROC_C(s) = 1 - F_c(G_c^{-1}(1 - s)), s \in (0,1)$$

where $F_c(u) = P(U \leq u | D = 1)$ and $G_c(u) = P(U \leq u | D = 0)$, the corresponding area under the optimal curve $ROC_C(s)$ is given by

$$AUC_C = \int_0^1 ROC(s) ds.$$

To be more specific, let $\{\mathbf{t}_1, \dots, \mathbf{t}_n\}$ denote the observed value of $\{\mathbf{T}_1, \dots, \mathbf{T}_n\}$, and because the optimal combination of different diagnostic tests is the likelihood ratio, the semiparametric log likelihood function of $(\alpha, \boldsymbol{\beta})$ is given by:

$$\ell(\alpha, \boldsymbol{\beta}) = \sum_{j=1}^{n_1} [\alpha + \boldsymbol{\beta}^\tau h(\mathbf{y}_j)] - \sum_{i=1}^n \log(1 + \rho \exp(\alpha + \boldsymbol{\beta}^\tau h(\mathbf{t}_i))) - n \log n_0$$

where $\rho = n_1/n_0$ is assumed to converge when $n \rightarrow \infty$. And the maximum semiparametric likelihood estimator $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\boldsymbol{\beta}}^\tau)$ of $\boldsymbol{\theta}$ is the solution of the following equations:

$$\frac{\partial \ell(\alpha, \boldsymbol{\beta})}{\partial \alpha} = n_1 - \sum_{i=1}^n \frac{\rho \exp(\alpha + \boldsymbol{\beta}^\tau h(\mathbf{t}_i))}{1 + \exp(\alpha + \boldsymbol{\beta}^\tau h(\mathbf{t}_i))} = 0$$

$$\frac{\partial \ell(\alpha, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{j=1}^{n_1} h(\mathbf{y}_j) - \sum_{i=1}^n \frac{\rho \exp(\alpha + \boldsymbol{\beta}^\tau h(\mathbf{t}_i))}{1 + \exp(\alpha + \boldsymbol{\beta}^\tau h(\mathbf{t}_i))} h(\mathbf{t}_i) = 0$$

Under this model, the maximum semiparametric likelihood estimator of F and G are given by:

$$\tilde{F}(t) = \frac{1}{n_0} \sum_{i=1}^n \frac{\exp(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^\tau h(\mathbf{T}_i)) I(\mathbf{T}_i \leq t)}{1 + \rho \exp(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^\tau h(\mathbf{T}_i))}$$

$$\tilde{G}(t) = \frac{1}{n_0} \sum_{i=1}^n \frac{I(\mathbf{T}_i \leq t)}{1 + \rho \exp(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^\tau h(\mathbf{T}_i))}$$

Now let $\tilde{U}_k = \tilde{\alpha} + \tilde{\boldsymbol{\beta}}^\tau h(\mathbf{T}_k)$ for $k = 1, \dots, n$. Similar to the estimation about F and G above, $F_C(u)$ and $G_C(u)$ can be estimated by the following expressions:

$$\tilde{F}_C(u) = \frac{1}{n_0} \sum_{i=1}^n \frac{\exp(\tilde{U}_i) I(\tilde{U}_i \leq u)}{1 + \rho \exp(\tilde{U}_i)}$$

$$\tilde{G}_C(u) = \frac{1}{n_0} \sum_{i=1}^n \frac{I(\tilde{U}_i \leq u)}{1 + \rho \exp(\tilde{U}_i)}$$

It may also be proposed that the estimated optimal ROC and its area (AUC) can be represented as:

$$\overline{ROC}_C(s) = 1 - \tilde{F}_C(\tilde{G}_C^{-1}(1-s)), s \in (0,1)$$

$$\overline{AUC}_C = \int_0^1 \overline{ROC}_C(s) ds.$$

In addition, this semiparametric estimators $(\widetilde{F}_C(u), \widetilde{G}_C(u))$ and the area under the estimated optimal ROC curve \widetilde{AUC}_C has the asymptotic behaviors as the following theorem:

Theorem 1

Under the model (1), for the estimator $(\widetilde{F}_C(u), \widetilde{G}_C(u))$, $\sqrt{n}(\widetilde{F}_C - F_C) \rightarrow W_F$ and $\sqrt{n}(\widetilde{G}_C - G_C) \rightarrow W_G$ weakly in $\wp[-\infty, \infty]$ as $n \rightarrow \infty$, where $\wp[-\infty, \infty]$ refers to the set of all real-valued functions that are right continuous and has left-hand limits for all $x \in [-\infty, \infty]$. And let $0 < a < b < 1$, suppose that f_C and g_C are the corresponding density functions that are continuous on $[G_C^{-1}(a) - \varepsilon, G_C^{-1}(b) + \varepsilon]$ for some $\varepsilon > 0$, then $\sqrt{n}(\widetilde{ROC}_C(s) - ROC_C(s)) \rightarrow W[G_C^{-1}(1 - s)]$ weakly in $\wp[1 - b, 1 - a]$ and $\sqrt{n}(\widetilde{AUC}_C - AUC_C) \rightarrow N(0, \sigma^2(\alpha_0, \beta_0, G_C))$. Here in these relations, W_F , W_G and W should satisfies some conditions.

More details and proofs of the asymptotic results can be found in Qin and Zhang (2010).

3.2 Optimal combination based on semiparametric monotonic density ratio model

In most papers, the combination needs to specify the density ratio or the distribution for multiple diagnostic tests correctly to obtain the optimal combination, which is

difficult to realize in practice, especially when the data is high-dimensional. Therefore, Chen et al. (2015) directly model the density ratio as a nonparametric function of a combination of multiple diagnostic tests, which they think can greatly improve the robustness of the combination proposed by Qin and Zhang (2010). They consider a semiparametric monotonic density ratio model, which we will call it model (2) in future pages:

$$\frac{f(\mathbf{x})}{g(\mathbf{x})} = \psi(v(\mathbf{x}, \boldsymbol{\beta}))$$

where $\psi(\cdot)$ is an undefined monotonic non-decreasing function, and $v(\mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^T h(\mathbf{x})$. While ROC is an invariant property, the optimal ROC curve of model (2) is based on the combination $v(\mathbf{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}^T h(\mathbf{x})$.

Let $\{\mathbf{t}_1, \dots, \mathbf{t}_n\}$ denote the observed value of $\{\mathbf{T}_1, \dots, \mathbf{T}_n\}$. And:

$$p_i = dF(\mathbf{t}_i), q_i = dG(\mathbf{t}_i), i = 1, 2, \dots, n,$$

$$m_i = I(t_i \in \{Y_1, \dots, Y_{n_1}\}), r_i = I(t_i \in \{X_1, \dots, X_{n_0}\}).$$

The corresponding likelihood function is

$$L = \prod_{i=1}^n p_i^{m_i} q_i^{r_i}$$

where $p_i, q_i \geq 0$, $\sum_{i=1}^n p_i = 1$, $\sum_{i=1}^n q_i = 1$, $p_i/q_i = \psi(v(\mathbf{t}_i, \boldsymbol{\beta}))$.

And the maximum empirical likelihood estimators of p_i and q_i can be defined to be:

$$\{\widehat{p}_1, \dots, \widehat{p}_n, \widehat{q}_1, \dots, \widehat{q}_n\} = \arg \max_{p_1, \dots, p_n, q_1, \dots, q_n} L$$

Then we assume $n_1/(n_0 + n_1) \rightarrow \lambda \in (0,1)$ as $n \rightarrow \infty$. p_i and q_i can be reparameterized to $\theta(v(\mathbf{t}_i; \boldsymbol{\beta}))$ and ϕ_i in this way:

$$\theta(v(\mathbf{t}_i; \boldsymbol{\beta})) = \frac{\lambda p_i}{\lambda p_i + (1 - \lambda) q_i} = \frac{\lambda \psi(v(\mathbf{t}_i; \boldsymbol{\beta}))}{\lambda \psi(v(\mathbf{t}_i; \boldsymbol{\beta})) + (1 - \lambda)}$$

$$\phi_i = \lambda p_i + (1 - \lambda) q_i$$

which is equivalent to:

$$p_i = \phi_i \theta(v(\mathbf{t}_i; \boldsymbol{\beta})) / \lambda$$

$$q_i = \frac{\phi_i \{1 - \theta(v(\mathbf{t}_i; \boldsymbol{\beta}))\}}{1 - \lambda}$$

If we apply the new-reparametrized parameters $\theta(v(\mathbf{t}_i; \boldsymbol{\beta}))$ and ϕ_i , the empirical likelihood function can be revised to $L = L_1 L_2$, with L_1 , and L_2 be expressed as:

$$L_1 = \lambda^{-n_1} (1 - \lambda)^{-n_0} \prod_{i=1}^n \{\theta(v(\mathbf{t}_i; \boldsymbol{\beta}))\}^{m_i} \{1 - \theta(v(\mathbf{t}_i; \boldsymbol{\beta}))\}^{r_i}$$

$$L_2 = \prod_{i=1}^n \phi_i^{m_i + r_i}$$

The maximum empirical likelihood estimators $\widehat{\phi}_i$ of ϕ_i are calculated by Dykstra et al. (1995) when $\boldsymbol{\beta}$ is fixed:

$$\widehat{\phi}_i = \frac{m_i + r_i}{n}$$

Then the maximum empirical likelihood estimators for $\theta(\cdot)$ and $\boldsymbol{\beta}$ can be calculated by maximizing L_1 :

$$\{\hat{\theta}(\cdot), \hat{\boldsymbol{\beta}}\} = \arg \min_{\theta(\cdot) \in \Theta, \boldsymbol{\beta} \in \mathbf{B}} M_n(\theta, \boldsymbol{\beta})$$

with

$$M_n(\theta, \boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^n [m_i \log\{\theta(v(\mathbf{t}_i; \boldsymbol{\beta}))\} + r_i \log(1 - \theta(v(\mathbf{t}_i; \boldsymbol{\beta})))]$$

$$\Theta = \{\theta(\cdot): \theta(\cdot) \in [0,1] \text{ and is monotonic increasing}\}$$

$$\mathbf{B} = \{1\} \times \mathbf{B}_{-1}$$

It should be noted that for this semiparametric monotonic density ratio model, we would apply PAVA method described in Ayer (1955) to do the minimization with respect to θ .

Let $F_C(u; \boldsymbol{\beta})$ and $G_C(u; \boldsymbol{\beta})$ be the cumulative distribution function (cdf) of $v(\mathbf{Y}; \boldsymbol{\beta})$ and $v(\mathbf{X}; \boldsymbol{\beta})$. Those cumulative distribution functions, the optimal ROC curve and its area can be estimated by:

$$\widehat{F}_C(u) = \sum_{i=1}^n \widehat{p}_i I((v(\mathbf{t}_i; \widehat{\boldsymbol{\beta}}) < u)$$

$$\widehat{G}_C(u) = \sum_{i=1}^n \widehat{q}_i I((v(\mathbf{t}_i; \widehat{\boldsymbol{\beta}}) < u)$$

$$\widehat{ROC}_C(s) = 1 - \widehat{F}_C(\widehat{G}_C^{-1}(1 - s)), s \in (0,1)$$

$$\widehat{AUC}_C = \int_0^1 \widehat{ROC}_C(s) ds.$$

And the following theorem shows the convergence rate of the previous estimated parameters if they satisfy some conditions:

$$(a) \ d\left(\widehat{\theta}(v(\mathbf{x}; \widehat{\boldsymbol{\beta}}), \theta_0(v(\mathbf{x}; \boldsymbol{\beta}_0)))\right) = O_p(n^{-1/3})$$

$$(b) \ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p(n^{-1/3})$$

$$(c) \ \sup_{u \in \mathbf{R}} |\widehat{F}_C(u) - F_C(u; \boldsymbol{\beta}_0)| = O_p(n^{-1/3}) \text{ and } \sup_{u \in \mathbf{R}} |\widehat{G}_C(u) - G_C(u; \boldsymbol{\beta}_0)| = O_p(n^{-1/3})$$

$$(d) \ |\widehat{ROCC}_C(s) - ROC_C(s)| = O_p(n^{-1/3}) \text{ for every } s \in (0,1)$$

$$(e) \ \widehat{AUC}_C - AUC_C = O_p(n^{-1/3})$$

The conditions that the asymptotic estimations above should satisfy and the detailed proofs could be found in Chen et al. (2015). We will not include all these parts because of the length limit and presentational continuity.

Chapter 4

Method description

In this chapter, we will describe the methods we are going to estimate the distribution for the estimated parameters β and AUC . We compare the performance of the following methods: ①Method I based on model (2) (semiparametric monotonic density ratio model); ②Method II based on model (2); ③Method II based on model (1) (exponential titling method).

Here Method I has the following steps:

- (a) Generate a dataset $\{X_1, \dots, X_{n_0}; Y_1, \dots, Y_{n_1}\}$ with n_0 non-diseased individuals and n_1 diseased ones.
- (b) Do B resamplings with replacement from the previous dataset $\{X_1, \dots, X_{n_0}; Y_1, \dots, Y_{n_1}\}$, form a ‘new’ dataset $\{X_1^B, \dots, X_{n_0}^B; Y_1^B, \dots, Y_{n_1}^B\}$, here $\{X_1^B, \dots, X_{n_0}^B\}$ are from $\{X_1, \dots, X_{n_0}\}$, $\{Y_1^B, \dots, Y_{n_1}^B\}$ are from $\{Y_1, \dots, Y_{n_1}\}$ and then calculate the corresponding $\widehat{\beta}^B$ and \widehat{AUC}_C^B for each resampling.
- (c) Collect B estimated $\widehat{\beta}^B$ and \widehat{AUC}_C^B calculated from each resampling in step (b), establish the asymptotic distribution of $\widehat{\beta}$ and \widehat{AUC}_C from the resampling results and do the hypothesis test to find whether the true value β_0 and AUC_C are in the confidence interval $(\alpha/2, 1 - \alpha/2)$ with confidence level α . If the true value of

β_0 or AUC_C is in this confidence interval, then the corresponding counting parameter $C = C + 1$.

- (d) Repeat step (a)~(c) for N times, and calculate the corresponding Type I error and Type II error.

Method II has the following steps:

- (a) Generate a dataset $\{\mathbf{X}_1, \dots, \mathbf{X}_{n_0}; \mathbf{Y}_1, \dots, \mathbf{Y}_{n_1}\}$ with n_0 non-diseased individuals and n_1 diseased ones.

- (b) Calculate the estimated cumulative density functions $\widehat{F}_C(u)$, $\widehat{G}_C(u)$ and \widehat{AUC}_C for the dataset $\{\mathbf{X}_1, \dots, \mathbf{X}_{n_0}; \mathbf{Y}_1, \dots, \mathbf{Y}_{n_1}\}$.

- (c) Generate a ‘new’ dataset $\{\mathbf{X}_1^N, \dots, \mathbf{X}_{n_0}^N; \mathbf{Y}_1^N, \dots, \mathbf{Y}_{n_1}^N\}$ based on the dataset $\{\mathbf{X}_1, \dots, \mathbf{X}_{n_0}; \mathbf{Y}_1, \dots, \mathbf{Y}_{n_1}\}$ and the estimated $\widehat{F}_C(u)$, $\widehat{G}_C(u)$ calculated in step (b).

Here ‘based on’ means that we assume that the distribution of diseased and non-diseased individuals are known and discrete which take values among $\{\mathbf{X}_1^N, \dots, \mathbf{X}_{n_0}^N; \mathbf{Y}_1^N, \dots, \mathbf{Y}_{n_1}^N\}$. They follow the cumulative distribution function $\widehat{F}_C(u)$ and $\widehat{G}_C(u)$. And different from Method I, here $\{\mathbf{X}_1^N, \dots, \mathbf{X}_{n_0}^N\}$ and $\{\mathbf{Y}_1^N, \dots, \mathbf{Y}_{n_1}^N\}$ are in fact from the whole dataset $\{\mathbf{X}_1, \dots, \mathbf{X}_{n_0}; \mathbf{Y}_1, \dots, \mathbf{Y}_{n_1}\}$.

- (d) Calculate the estimated parameters $\widehat{\beta}^N$ and \widehat{AUC}_C^N for the ‘new’ dataset $\{\mathbf{X}_1^N, \dots, \mathbf{X}_{n_0}^N; \mathbf{Y}_1^N, \dots, \mathbf{Y}_{n_1}^N\}$

- (e) Repeat this bootstrapping steps (c)~(d) for \mathbf{B} times, collect \mathbf{B} estimated $\hat{\boldsymbol{\beta}}^N$ and \widehat{AUC}_C^N calculated in step (d), establish the distribution of $\hat{\boldsymbol{\beta}}$ and \widehat{AUC}_C from the bootstrapping results and do the hypothesis test to find whether the true value $\boldsymbol{\beta}_0$ and AUC_C are in the confidence interval $(\alpha/2, 1 - \alpha/2)$ with confidence level α . If the true value $\boldsymbol{\beta}_0$ or AUC_C is in this confidence interval, then the corresponding counting parameter $C = C + 1$.
- (f) Repeat step (a)~(e) for \mathbf{N} times, and calculate the corresponding Type I error and Type II error.

Here Type I error for method I and II both have the expression:

$$\text{error } I_{\boldsymbol{\beta}} = C_{\boldsymbol{\beta}}/\mathbf{N}$$

$$\text{error } I_{AUC} = C_{AUC}/\mathbf{N}$$

For Type II error of $\boldsymbol{\beta}$, since we will standardize $\|\boldsymbol{\beta}\|_2 = 1$ in the simulation study in next chapter, we will set the value for hypothesis test vary from 0 to 1, with an interval of 0.05 and find out the value of Type II error.

And for AUC , which takes value from 0.5 to 1 if it is defined correctly, we will set the varying interval value equal to 0.01 around the true value because AUC has a more concentrated distribution compared with $\boldsymbol{\beta}$. We can observe the change of Type II error more precisely if we set the interval smaller for AUC .

About the process of calculating Type II error, we will count the number which the testing value is in the confidence interval and Type II error of that value is the counting number divided by N .

Chapter 5

Simulation study

In this chapter, we will estimate the distribution of $\hat{\boldsymbol{\beta}}$ and \widehat{AUC}_C and do the corresponding hypothesis tests to find out Type I error and Type II error while applying the methods described in the Chapter 4. Due to the time limit, we consider the sample size (n_0, n_1) equals to (600,300) and consider a combination of two biomarkers $\mathbf{X} = (X_1, X_2)$. For each method, we perform $\mathbf{B} = 100$ resamplings and $\mathbf{N} = 1000$ replications.

In this simulation, we assume $\mathbf{X} = (X_1, X_2)$ follows a bivariate exponential distribution, which we will describe it in detail in the next section. And we should also note that our estimation can also be applied for other distributions.

5.1 Biomarkers follow exponential distribution

First, we study the case when the two biomarkers follow the bivariate exponential distribution. We posit that $X_1|D = 1 \sim \exp(\xi_1)$, $X_2|D = 1 \sim \exp(\xi_2)$, and correlation $\text{Corr}(X_1, X_2) = \xi_0 / (\xi_1 + \xi_2 - \xi_0)$, where $0 \leq \xi_0 \leq \min(\xi_1, \xi_2)$, here ξ_1 and ξ_2 are rates of exponential distribution. The process for generating data is as follows:

- (a) Generate Y_1 from the univariate exponential distribution with rate $\xi_1 - \xi_0$ and Y_2 from the univariate exponential distribution with rate $\xi_2 - \xi_0$.
- (b) Generate Z from the univariate exponential distribution with rate ξ_0 .
- (c) Let $X_1 = \min(Y_1, Z)$ and $X_2 = \min(Y_2, Z)$.

From the process above, the joint density function of $\mathbf{X}|D = 1$ for diseased individuals can be easily verified:

$$\begin{cases} f(\mathbf{x}) = \xi_1(\xi_2 - \xi_0)e^{-\xi_1 x_1 - (\xi_2 - \xi_0)x_2} \text{ if } x_2 < x_1 \\ f(\mathbf{x}) = \xi_2(\xi_1 - \xi_0)e^{-\xi_2 x_2 - (\xi_1 - \xi_0)x_1} \text{ if } x_1 < x_2 \end{cases}$$

For $\mathbf{X}|D = 0$, we similarly generate \mathbf{X} for non-diseased individuals. We apply different rates for different groups. For example, we set $\xi_1^D = \xi_2^D = 2$, $\xi_0^D = 1$ for diseased group and $\xi_1^N = \xi_2^N = 10$, $\xi_0^N = 1$ for non-diseased group. Then suppose $f(\mathbf{x})$ and $g(\mathbf{x})$ are the probability density function for diseased and non-diseased groups, the log density ratio can be easily verified:

$$\log \frac{f(\mathbf{x})}{g(\mathbf{x})} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Let $U = \alpha + \beta_1 x_1 + \beta_2 x_2$. It is easy to verify that U is always the optimal combination of X_1 and X_2 . Here if the rates $\xi_1^D, \xi_2^D, \xi_0^D, \xi_1^N, \xi_2^N, \xi_0^N$ are given, then α , β_1 and β_2 would have an explicit expression:

$$\alpha = \log \left(\frac{\xi_1^D (\xi_2^D - \xi_0^D)}{\xi_1^N (\xi_2^N - \xi_0^N)} \right)$$

$$\beta_1 = \xi_1^N - \xi_1^D$$

$$\beta_2 = \xi_2^N - \xi_2^D$$

Next we study the robustness of the methods. It means that the “diseased group” would include some non-diseased individuals and the “non-diseased group” would include some diseased individuals. We posit that the diseased group satisfies the distribution:

$$f(\mathbf{x}) = \lambda_1 f_1(\mathbf{x}) + (1 - \lambda_1) f_0(\mathbf{x})$$

and the non-diseased group satisfies the distribution:

$$g(\mathbf{x}) = (1 - \lambda_0) f_1(\mathbf{x}) + \lambda_0 f_0(\mathbf{x})$$

where $1 - \lambda_1$ is the proportion of non-diseased individuals contained in the diseased group and $1 - \lambda_0$ is the proportion of diseased individuals contained in the non-diseased group. Here $f_1(\mathbf{x})$ and $f_0(\mathbf{x})$ refers to the probability density function of the bivariate exponential distribution for diseased and non-diseased populations described in this section.

5.2 Distribution estimation and comparison

In this section, we will compare the distribution estimation of $\boldsymbol{\beta}$ and AUC while applying different methods described in Chapter 4. For all methods, we standardize $\|\boldsymbol{\beta}\|_2 = 1$, while $\|\cdot\|_2$ is the L_2 norm. We fix $\xi_1^D = \xi_2^D = 2$, $\xi_0^D = 1$ for the

diseased group, and vary $\xi_1^N = \xi_2^N = 3/5/10$, $\xi_0^N = 1$ for the non-diseased group to simulate different magnitudes between these two groups.

First, we display the estimating results of the models proposed by Qin and Zhang (2010) and Chen et al. (2015) when $\lambda = 1$ and $\lambda < 1$:

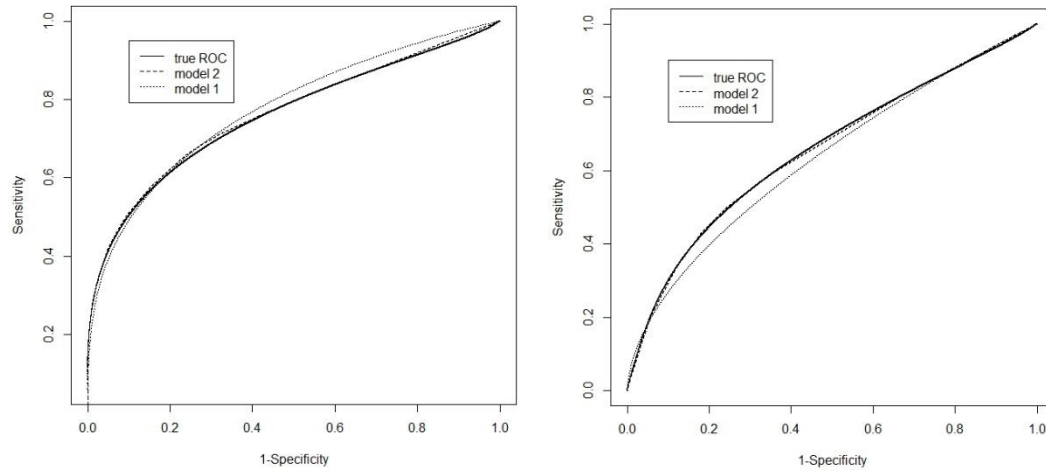


Figure 1. The ROC estimation results for $\lambda = 1$ (left panel) and $\lambda = 0.8$ (right panel) when $\xi_1^D = \xi_2^D = 2$ and $\xi_1^N = \xi_2^N = 5$

From the Figure 1 in the left panel, the estimations based on two models are both acceptable for non-robust case. For $\lambda = 0.8$, the estimations are still not bad for both models. If we take a comparison, Model 2 performs comparably better than Model 1 for both robust and non-robust cases.

Then, we could focus on the distribution estimation for β . Here we only test for β_2 since we have already standardized β . And we set confidence level $\alpha = 0.05$ for

all methods. Since we set $\xi_1^D = \xi_2^D$, $\xi_1^N = \xi_2^N$, the true value for β_2 is equal to $1/\sqrt{2} \approx 0.707$.

We take $\xi_1^D = \xi_2^D = 2$ and $\xi_1^N = \xi_2^N = 3$ as an example to describe the results in detail. Other parameter settings can be analyzed in a similar way:

			①			②			③		
λ_0	λ_1	Par.	Bias	MSE	Type I	Bias	MSE	Type I	Bias	MSE	Type I
1	1	β_2	-0.021	0.025	0.016	-0.017	0.023	0.127	-0.005	0.016	0.063
0.8	0.8	β_2	-0.063	0.073	0.020	-0.083	0.081	0.006	-0.036	0.053	0.001
0.6	0.6	β_2	-0.172	0.165	0.019	-0.151	0.165	0.000	-0.251	0.355	0.000

Table 2. Method performances for estimating β_2 while choosing different λ_0 , λ_1

$$\text{when } \xi_1^D = \xi_2^D = 2 \text{ and } \xi_1^N = \xi_2^N = 3$$

Table 2 shows the bias, MSE and Type I error for three methods in different λ_0 , λ_1 settings. If we do not consider the robustness, that is $\lambda_0 = \lambda_1 = 1$, all three proposed methods have an acceptable bias and MSE. If we compare these parameters more precisely, method ③ is comparably better than method ① and method ②. And from the Type I error, we compare the performance of the distribution estimation. From these three methods, Type I error of method ③ (0.063) is closest to the confidence level $\alpha = 0.05$, therefore this method has the best performance in

distribution estimation among three methods. And Type I error of method ① (0.016) is comparably small and Type I error of method ② (0.127) is comparably big, means that the distribution estimation of method ① is a little rough and the distribution estimation of method ② is a little specific compared with the true distribution. Here are their distribution estimations for a dataset showed by histograms, which is consistent to the analysis above:

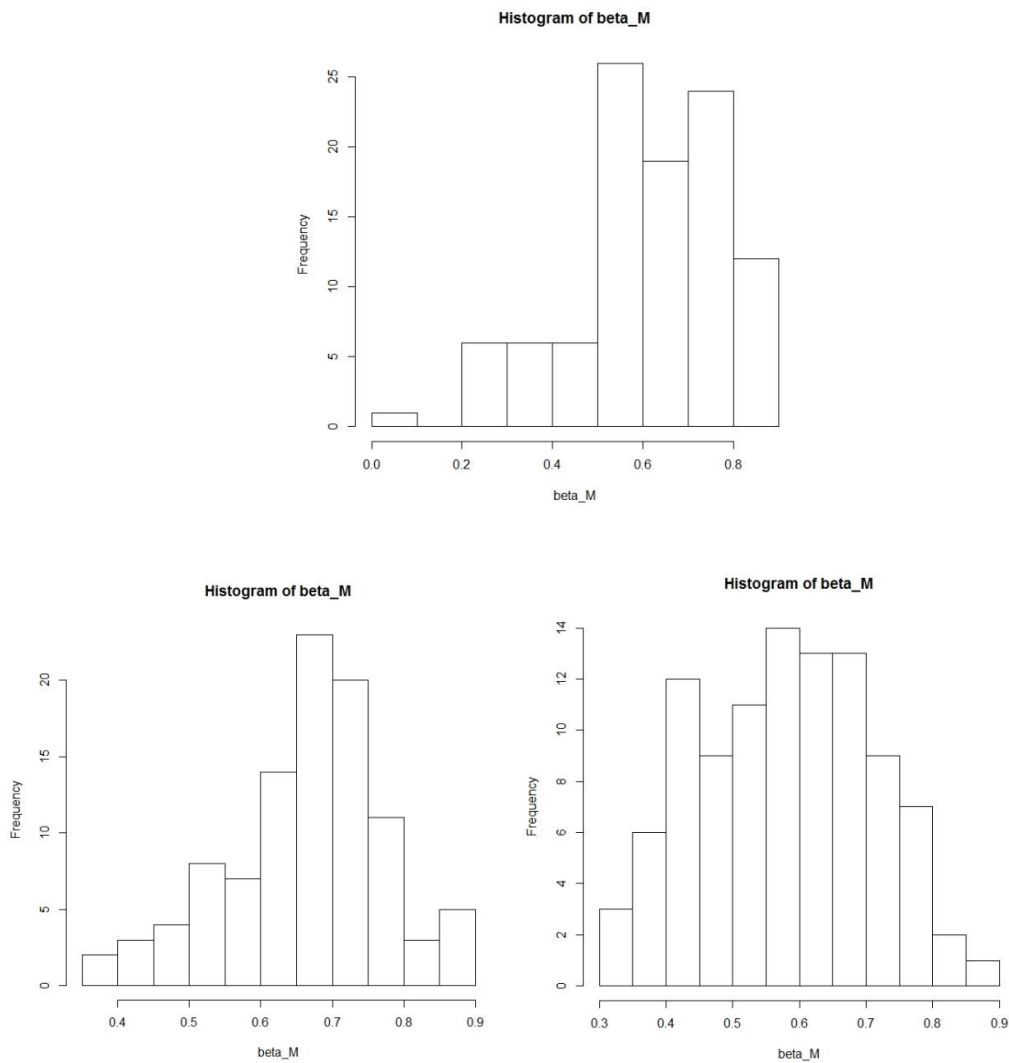


Figure 2. Histograms of the distribution estimation for three methods (graphs are listed by the order method ①、 ② and ③)

And the Table 3 below shows Type II error for non-robust case. We roughly find that ③>②>① in performance because Type II error of method ③ has the fastest decreasing speed compared with method ① and method ② when β_2 is away from the true value. But since the difference of ξ^D and ξ^N is not great, even when $\beta_2 = 0.15$ or another side $\beta_2 = 0.95$, the Type II error for all three methods are still bigger than 0.050, showing that the Type II error for all three methods can not be controlled quite well. But this can be controlled much better when the magnitude is greater.

β_2	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
①	0.121	0.186	0.256	0.316	0.399	0.482	0.581	0.667	0.761	0.834
②	0.038	0.066	0.093	0.133	0.197	0.256	0.344	0.432	0.518	0.627
③	0.021	0.032	0.050	0.086	0.140	0.216	0.297	0.406	0.548	0.664
β_2	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
①	0.902	0.947	0.976	0.987	0.967	0.942	0.858	0.690	0.437	0.000
②	0.726	0.803	0.851	0.875	0.866	0.794	0.666	0.483	0.238	0.000
③	0.768	0.859	0.907	0.933	0.928	0.867	0.755	0.547	0.257	0.000

Table 3. Comparison of Type II error for distribution estimation while $\lambda_0 = \lambda_1 = 1$

$$\xi_1^D = \xi_2^D = 2 \text{ and } \xi_1^N = \xi_2^N = 3 \text{ for different methods}$$

Back to Table 2, we then focus on the robustness. It can be found the absolute value of bias for method ③ increase extremely quickly (from 0.005 to 0.251) and its Type I error suddenly decreases to 0, showing that method ③ is not suitable for robust case. Similar results and conclusions can be found for method ②, with its absolute value of bias increases from 0.017 to 0.151. And method ① seems to be the most stable method that even if its absolute value of bias increases a little bit (from 0.021 to 0.171) when λ_0, λ_1 are decreasing, its Type I error is still at the same level (~ 0.02).

- And here are the estimating results for three method when $\xi_1^D = \xi_2^D = 2$, $\xi_1^N = \xi_2^N = 5$ and 10:

			①			②			③		
λ_0	λ_1	Par.	Bias	MSE	Type I	Bias	MSE	Type I	Bias	MSE	Type I
1	1	β_2	-0.001	0.007	0.020	-0.010	0.007	0.097	-0.004	0.004	0.052
0.8	0.8	β_2	-0.020	0.027	0.026	-0.027	0.029	0.002	-0.014	0.020	0.003
0.6	0.6	β_2	-0.114	0.131	0.019	-0.121	0.134	0.000	-0.113	0.152	0.000

Table 4. Method performances for estimating β_2 while choosing different λ_0, λ_1

when $\xi_1^D = \xi_2^D = 2$ and $\xi_1^N = \xi_2^N = 5$

β_2	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
①	0.000	0.000	0.000	0.001	0.003	0.008	0.028	0.092	0.216	0.413
②	0.000	0.000	0.000	0.000	0.002	0.007	0.019	0.052	0.145	0.300
③	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.016	0.083	0.191
β_2	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
①	0.634	0.816	0.933	0.979	0.965	0.821	0.523	0.152	0.011	0.000
②	0.481	0.679	0.835	0.898	0.815	0.615	0.334	0.098	0.007	0.000
③	0.408	0.643	0.868	0.946	0.882	0.663	0.301	0.041	0.000	0.000

Table 5. Comparison of Type II error for distribution estimation while $\lambda_0 = \lambda_1 = 1$

$$\xi_1^D = \xi_2^D = 2 \text{ and } \xi_1^N = \xi_2^N = 5 \text{ for different methods}$$

			①			②			③		
λ_0	λ_1	Par.	Bias	MSE	Type I	Bias	MSE	Type I	Bias	MSE	Type I
1	1	β_2	-0.003	0.005	0.023	-0.009	0.005	0.108	-0.001	0.003	0.057
0.8	0.8	β_2	-0.019	0.020	0.018	-0.015	0.020	0.003	-0.013	0.020	0.004
0.6	0.6	β_2	-0.111	0.126	0.022	-0.120	0.130	0.000	-0.077	0.111	0.000

Table 6. Method performances for estimating β_2 while choosing different λ_0, λ_1

$$\text{when } \xi_1^D = \xi_2^D = 2 \text{ and } \xi_1^N = \xi_2^N = 10$$

β_2	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
①	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.007	0.068	0.215
②	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.014	0.060	0.165
③	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.068
β_2	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
①	0.460	0.735	0.919	0.975	0.950	0.763	0.364	0.053	0.000	0.000
②	0.383	0.637	0.829	0.896	0.814	0.547	0.218	0.027	0.001	0.000
③	0.246	0.545	0.821	0.945	0.870	0.536	0.160	0.008	0.000	0.000

Table 7. Comparison of Type II error for distribution estimation while $\lambda_0 = \lambda_1 = 1$

$$\xi_1^D = \xi_2^D = 2 \text{ and } \xi_1^N = \xi_2^N = 10 \text{ for different methods}$$

From Table 2-7, we can have a look at the performances of these three methods for different ξ_1^N, ξ_2^N settings in robust and non-robust case. We can analyze these results in a similar way to the previous setting. From all these settings, we conclude that the method ③ estimate the distribution best if we do not consider the robustness. It has the smallest MSE and bias and its Type I error is always closest to confidence level $\alpha = 0.05$ even if magnitudes are different. For the other two methods, Type I error for method ① is comparably a little small (~ 0.02) while Type I error for method ② is comparably a little big (~ 0.10). Then if we consider the robust case, we find that method ① is the most stable method that its Type I error for distribution estimation doesn't change too much (~ 0.02). Method ② and method ③ perform not so well if we take

robustness into account, especially method ③ for small magnitude, its bias and MSE increase quickly when λ_0, λ_1 decrease. We also find similar opinion in Chen et al. (2015)'s paper that the exponential titling method cannot be correctly specified if $\lambda < 1$, corresponding to our analysis towards method ③.

If we focus on Type II error, we conclude that ③>②>① in performance. We can look at another parameter setting different from the discussion above. If we choose to take $\xi_1^D = \xi_2^D = 2$, $\xi_1^N = \xi_2^N = 5$ and Type II error=0.05 as the judgement point as an example, we find that for method ③, we have a Type II error smaller than 0.05 if $\beta_2 < 0.430$ or $\beta_2 > 0.898$, while for method ②, we achieve this when $\beta_2 < 0.397$ or $\beta_2 > 0.916$, and for method ①, β_2 should satisfy $\beta_2 < 0.368$ or $\beta_2 > 0.942$.

Then, we turn to the distribution estimation for AUC . While the biomarkers follow bivariate exponential distribution, the direct calculation for the true value AUC would be not be a difficult task. Another approach is to numerically calculate $P(U_D > U_N)$, here U_D is the optimal combination of the data from the diseased group and U_N is for the non-diseased group. The result of these two approaches should be equivalent. It should also be noticed that when we consider the robustness, the true value of AUC will change greatly. Here are the true value of AUC for different ξ and λ and the estimation of AUC for three methods:

λ_0	λ_1	Par.	$\xi_1^N = \xi_2^N = 3$	$\xi_1^N = \xi_2^N = 5$	$\xi_1^N = \xi_2^N = 10$
1	1	<i>AUC</i>	0.61975	0.75735	0.87943
0.8	0.8	<i>AUC</i>	0.57137	0.65120	0.72908
0.6	0.6	<i>AUC</i>	0.52448	0.55094	0.57750

Table 8. True value for *AUC*

			①		②		③	
λ_0	λ_1	Par.	Bias	MSE	Bias	MSE	Bias	MSE
$\xi_1^N = \xi_2^N = 3$								
1	1	<i>AUC</i>	0.018	7×10^{-4}	0.018	7×10^{-4}	0.013	5×10^{-4}
0.8	0.8	<i>AUC</i>	0.023	9×10^{-4}	0.022	8×10^{-4}	0.005	3×10^{-4}
0.6	0.6	<i>AUC</i>	0.029	0.001	0.028	0.001	0.007	3×10^{-4}
$\xi_1^N = \xi_2^N = 5$								
1	1	<i>AUC</i>	0.012	5×10^{-4}	0.011	4×10^{-4}	0.013	4×10^{-4}
0.8	0.8	<i>AUC</i>	0.020	8×10^{-4}	0.020	7×10^{-4}	-0.013	5×10^{-4}
0.6	0.6	<i>AUC</i>	0.026	0.001	0.024	9×10^{-4}	-0.004	3×10^{-4}
$\xi_1^N = \xi_2^N = 10$								
1	1	<i>AUC</i>	0.010	3×10^{-4}	0.010	3×10^{-4}	0.012	3×10^{-4}
0.8	0.8	<i>AUC</i>	0.014	5×10^{-4}	0.014	5×10^{-4}	-0.055	0.003
0.6	0.6	<i>AUC</i>	0.021	8×10^{-4}	0.020	7×10^{-4}	-0.022	8×10^{-4}

Table 9. Estimation for *AUC* for three methods

Table 8 shows the true value of AUC in different cases and Table 9 shows the estimation results for AUC for three methods. It is not surprising that method ① and ② have almost the same bias and MSE, since these two methods have the same data-generating and initial estimating process. The results in Table 9 are consistent to those results showed in Figure 1. However, from the bias for non-robust case, we also notice that the estimation for AUC is not so good as the estimation for β_2 for this sample size. Actually, the estimated distribution of AUC is always more concentrated than β_2 . Therefore, taking those three existing methods to give a distribution estimation for AUC do not have quite good performance. Taking $\xi_1^N = \xi_2^N = 5$ as an example, confidence level $\alpha = 0.05$:

			①		②		③	
λ_0	λ_1	Par.	Bias	Type I	Bias	Type I	Bias	Type I
1	1	AUC	0.012	0.227	0.011	0.299	0.013	0.167
0.8	0.8	AUC	0.020	0.372	0.020	0.300	-0.013	0.999
0.6	0.6	AUC	0.026	0.532	0.024	0.000	-0.004	0.763

Table 10. Method performances for estimating AUC while choosing different λ_0 ,

$$\lambda_1 \text{ when } \xi_1^D = \xi_2^D = 2 \text{ and } \xi_1^N = \xi_2^N = 5$$

From Table 10, we find that the hypothesis test perform badly for AUC estimation for all these three methods. Although their bias seem acceptable for both robust and non-robust cases, their Type I errors are a little big compared with the confidence level $\alpha = 0.05$.

In short, the methods we have come out with are suitable for estimating the asymptotic distribution of β . For non-robust case, Method ③ is found to be the best method and for robust case, Method ① has the best performance. But these three methods do not have a good performance for estimating AUC . We probably need to think of some new methods if we want to estimate its asymptotic distribution.

Chapter 6

Conclusion and discussion

In this study, we apply the optimal combination of multiple diagnostic test while using exponential tilting method proposed by Qin and Zhang (2010) and the monotone density ratio method proposed by Chen et al. (2015) to establish the asymptotic distribution for β and AUC . By applying bootstrapping and resampling, we come out with three methods and make comparison to their performances.

From the simulation results, we find that the estimated distributions of β are not so central-concentrated than AUC but have better estimating accuracy for their estimations. If we do not consider the robustness and if we set the confidence level $\alpha = 0.05$, method ③ has the best approximation for the distribution of β with about 0.05~0.06 Type I error and also perform best for Type II error. And if we take the robustness into account, method ① seems to be the most stable one. Its Type I error remains to be about 0.02 for all λ we have tested.

As we have concluded above, those three methods we come out have acceptable results for estimating the asymptotic distribution for β and each method has its advantages. However, if we apply our methods to estimate AUC , those methods work not so well as expected. Maybe it is because that the asymptotic distribution of AUC

is more central-concentrated. Therefore, the hypothesis test towards AUC may be more sensitive to the bias, and it needs a method with higher accuracy for estimating the distribution than estimating for β .

In conclusion, we have successfully given an estimation towards the asymptotic distribution of β with good accuracy. And about AUC , the estimating part still remains to be an open question for future studies. We can perhaps turn to discover some methods with higher accuracy in estimation which may work for estimating AUC . More thinking can be taken into this interesting topic.

Reference

Ayer M, Brunk H D, Ewing G M, et al. An empirical distribution function for sampling with incomplete information[J]. *The annals of mathematical statistics*, 1955, 26(4): 641-647.

Barreno M, Cardenas A, Tygar J D. Optimal ROC curve for a combination of classifiers[C]//*Advances in Neural Information Processing Systems*. 2008: 57-64.

Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine* 1991; **10**:1887--1895.

Chen B, Li P, Qin J, et al. Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests[J]. *Journal of the American Statistical Association*, 2015.

Copas J, Corbett P. Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* 2002; **89**:315--331.

Dykstra R, Kocher S, Robertson T. Inference for likelihood ratio ordering in the two-sample problem[J]. *Journal of the American Statistical Association*, 1995, 90(431): 1034-1040.

Efron B. *Bootstrap methods: another look at the jackknife*[M]. Springer New York, 1992.

Eguchi S, Copas J. A class of logistic-type discriminant functions. *Biometrika* 2002; **89**:1--22.

Etzioni R, Kooperberg C, Pepe M, Smith R, Gann PH. Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics* 2003; **4**:523--538.

Green D M, Swets J A. Signal detection theory and psychophysics[J]. Society, 1966, 1: 521.

Henson D E, Srivastava S, Kramer B S. Molecular and genetic targets in early detection[J]. *Current opinion in oncology*, 1999, 11(5): 419.

Hernández-Orallo J, Flach P, Ferri C. A unified view of performance metrics: translating threshold choice into expected classification loss[J]. *The Journal of Machine Learning Research*, 2012, 13(1): 2813-2869.

Hsieh F, Turnbull B W. Nonparametric and semiparametric estimation of the receiver operating characteristic curve[J]. *The annals of statistics*, 1996, 24(1): 25-40.

Kay R, Little S. Transformations of the explanatory variables in the logistic regression model for binary data[J]. *Biometrika*, 1987, **74**(3): 495-501.

Kim Y S, Jang M K, Park C Y, et al. Exploring multiple biomarker combination by logistic regression for early screening of ovarian cancer[J]. *Int J Bio-Sci Bio-Tech*, 2013, 5: 67.

- Krzanowski W J, Hand D J. ROC curves for continuous data[M]. CRC Press, 2009.
- Liu D, Zhou X H. ROC analysis in biomarker combination with covariate adjustment[J]. *Academic radiology*, 2013, 20(7): 874-882.
- Kunsch H R. The jackknife and the bootstrap for general stationary observations[J]. *The Annals of Statistics*, 1989: 1217-1241.
- McIntosh MW, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics* 2002; **58**:657--664.
- Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: New York, 2003.
- Qin J, Zhang B. Best combination of multiple diagnostic tests for screening purposes[J]. *Statistics in medicine*, 2010, **29**(28): 2905-2919.
- Srinivas P R, Kramer B S, Srivastava S. Trends in biomarker research for cancer detection[J]. *The lancet oncology*, 2001, 2(11): 698-704.
- Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* 1993; **88**:1350--1355.
- Swets J A. Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers[M]. Psychology Press, 2014.
- Wu C F J. Jackknife, bootstrap and other resampling methods in regression analysis[J]. *the Annals of Statistics*, 1986: 1261-1295.

Yuan Z, Ghosh D. Combining multiple biomarker models in logistic regression[J].

Biometrics, 2008, 64(2): 431-439.

Zhou XH, McClish DK, Obuchowski NA. *Statistical Methods in Diagnostic Medicine*.

Wiley: New York, 2002.

Zou K H, Liu A, Bandos A I, et al. Statistical evaluation of diagnostic performance:

topics in ROC analysis[M]. CRC Press, 2011.

Zweig M H, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental

evaluation tool in clinical medicine[J]. Clinical chemistry, 1993, 39(4): 561-577.