ROC ANALYSIS IN DIAGNOSTIC MEDICINE

ZHANG YANYU

(Bachelor of Mathematics, Jiangxi Normal University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2010

Acknowledgements

I would like to express my deepest appreciation and thanks to my advisor, Professor Li Jialiang, for his brilliant guidance and invaluable feedback and support, without which this thesis would not be possible. He has been an inspiration to me both professionally and personally. I would also like to thank my loving family for all their support and understanding. They have given me much motivation and encouragement throughout my time in Singapore. My thanks also go out to my classmates and friends for their help and encouragement throughout the writing of this thesis. Special thanks go out to my husband for his motivation and patience. He has been a sounding board for me and given me much love and support during the writing of this thesis. Finally, I would like to dedicate this thesis to the loving memory of my grandfather.

Contents

Summary

The accuracy of a diagnostic test can be quantified by how well the test results classify and predict the true condition status. As such, the diagnostic accuracy of a test is of utmost importance in determining the suitability of implementing the test and is particularly essential in real-world situations. The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are two important summary measures that provide an effective assessment of the overall accuracy of diagnostic tests. Over the years, several parametric, semi-parametric and nonparametric methods have been developed for the estimation of the ROC curve and AUC for two-category classifications.

However, many real-world biomedical classification problems demand the ability to assess more than just two classes. ROC analyses capable of handling multiple classifications are needed to more robustly assess the diagnostic performance. Scurfield (1996) presented the mathematical definition of suitable ROC measures for more than two classes. The ROC curves are extended to ROC surfaces for three-category classification and ROC manifolds for multiple-category classification.

Acquiring the correct order is important for multiple-category ROC analysis when the categories are ordinal. Inference methods that estimate the summary measures have recently been proposed. The volume under the ROC surface (VUS) and the hypervolume under the manifold (HUM) can be estimated for ordered multiple-category problems by applying U-statistic theory. In this thesis, we propose rigorous and automated approaches to sort the multiple categories by using simple summary statistics such as means. We also provide a general discussion regarding the minimum acceptable HUM values in multiple-category classification problems. The analyses presented in this thesis provide insights into how best to screen through the large number of tests available in the health science field. Bootstrap inferences are proposed to account for the variability.

In medical research, evaluating the various factors that can influence the diagnostic performance is also imperative. Recently, statistical regression analysis has been researched to more thoroughly inference about such factors and biomarkers. Statistical methods that combine multiple tests for multiple-category classification can efficiently optimize the accuracy of the combined marker under the criteria of ROC measures. For binary classification, Pepe and Thompson (2000) developed a method based upon maximizing the AUC of the combined biomarkers in genetic studies. Their method is effectively adapted from the maximum rank correlation (MRC) estimation proposed by Han (1987) which is widely practiced. Recently, the MRC estimator has been applied in classification studies due to its close connection with AUC. In this thesis, we explore statistical methods that combine multiple tests for multiple-category classification with the ambition to optimize the accuracy of the combined markers under the criteria

of ROC measures. We develop suitable statistical procedures by extending the MRC estimator to high-dimensional cases and also provide the necessary supporting asymptotic theories. Simulations and examples are provided to demonstrate that significantly higher VUS or HUM can be achieved by combining multiple biomarkers.

List of Tables

List of Figures

Chapter 1

Introduction

Statistical classification is needed in various fields such as computer science, economics, meteorology, biology, biochemistry and medical studies. The diagnosis of the status of a subject is crucial to its accurate classification, and the selection of the statistical methodology applied to the prediction and classification is of utmost importance. Particularly in the field of medicine and in clinical studies, the accurate and timely diagnosis of a patient's condition is crucial to the ultimate treatment of the diseased condition. Detecting these conditions and evaluating the prognosis of patients with disease can be achieved by analyzing the clinical and laboratory data. An inaccurate diagnosis in many real-world biomedical settings carry emotionally stressful and financial consequences.

The classification resulting from a diagnostic test can be as straightforward as the presence or absence of the specific disease-related material or it can yield an entire ar-

ray of non-binary results. For non-binary continuous or ordinal (subjective) scales, the classification can be set by a threshold value with results above or below such threshold classified as positive or negative for disease, as appropriate. The ability to directly predict the multiple stages of a disease rather than to merely distinguish between a disease and non-disease state is often more crucial in real-world situations. For example, in cancer patients in which the progression of the disease is relatively fast, determining the stage of the disease is crucial to applying the appropriate treatment, and earlier detection of the stage of the disease can vastly increase survivability of the patient via the appropriate medical prognosis.

1.1 Diagnostic test

From a technological and procedural perspective, the diagnostic test for the classification can be relatively simple or complex. For example, from a technological standpoint, the test can be a classic bacterial culture test, or it can be a complex application employing the latest in genetic sequencing technologies. From a procedural standpoint, the test may only involve one step which results in one of only two outcomes, positive or negative, or it may involve a vast sequence of procedures that may result in one of an entire spectrum of possible classifications.

The implementation of a diagnostic test should be preconditioned on the practicality and benefit of such a test toward the classification or prediction of the diseased condi-

tion. The key criteria that should be considered before implementing a diagnostic test can be adapted from Wilson and Jungner (1968), Cole and Morrison (1980) and Obuchowski et al. (2001), who discuss criteria for useful screening programs which share similar considerations to the application of diagnostic tests in general. The criteria pertain to the disease (first, second and third criterion), the treatment for the disease (fourth criterion) and to the test itself (fifth and sixth criterion). Firstly, the disease should be serious or potentially so as to merit its use for diagnosis to potentially improve the longevity or quality of life of the subjects. Secondly, the disease should be relatively prevalent in the target population so as to have a potential benefit from testing subjects. Thirdly, the purpose of diagnosing the disease is so that it can be treated, so the disease should be treatable. Fourthly, there must exist an effective treatment to be beneficial for those who test positive. The fifth and sixth criteria pertain to the medical test itself. The fifth criterion is that the test procedure should ideally cause no harm to the individual. However, all tests have more or less negative impact, whether it is financial, physical or emotional discomfort or damage. In practicality, these costs should be reasonably in context and the information from an accurate diagnosis should create potential benefits to be gained by the population or individual being tested. The sixth and final criterion is the accuracy of the test which is discussed in more detail in the next section.

1.2 Diagnostic accuracy

An accurate test is one that correctly classifies its test population according to the disease or non-disease condition. Inaccurate tests cause those with actual disease to be misclassified as non-diseased, also known as a ”false negative”. Conversely, they cause those with no actual disease to be misclassified as diseased, also known as a ”false positive”. False negative errors leave diseased subjects untreated. False positive errors open subjects to being subjected to unnecessary procedures and emotional stress. Both false negatives and false positives may also create disillusionment and distrust within the general subjects towards the medical and diagnostic testing community as a whole, potentially making data collection more difficult, biased and costly. Obviously, such errors must be kept to a minimum. As such, the diagnostic accuracy of a test is of utmost importance and must be thoroughly assessed and understood before such a test can be used in practice.

In order to effectively implement and assess a diagnostic test, we must thoroughly evaluate the test population, the test itself and the resulting observations for many factors which may influence the analysis of the accuracy by applying statistical methodologies. We must make sure that the population taking the tests are not influenced by knowledge of their true disease classifications or that the test itself is not influenced by knowledge of the same which could alter the accuracy of the diagnostic test. The persons administering and assessing the results of the test should also be blind to the population’s true disease classifications so as not to influence the test results. These

situations are more common when assessing more subjective factors of a study.

Many other factors can affect the performance of a diagnostic test for the purpose of detecting disease. These include biased test populations that are not representative of diseased subjects in the general population, inadequate clinical samples that may affect the results of the test, a condition of a repeat testing that results in a positive diseased status which may be counted as tested once rather than twice, the time it takes between when the test is administered and when the results are assessed, patient related factors (demographics, health habits, truthfulness), tester related factors (training, experience), environmental factors (available resources, treatment options, integrity of reporting), etc.

In some cases, statistical methodologies may be enhanced and improved to generate significantly more accurate classification predictions. In other cases, a procedurally simpler statistical methodology may prove to be relatively more efficient than other methodologies, without sacrificing accuracy, especially for computation-heavy studies or for cases in which time is of the essence. The statistical methods discussed in this thesis pertain to assessment of the accuracy of a diagnostic test. The analyses assume that the diagnostic tests are conducted in an appropriately controlled environment. As such, we must keep in mind the many real-world factors, as mentioned above, that may influence the accuracy of such tests, for the benefit of the potential implementation of such methodologies.

1.3 Measures of accuracy

In this section, we introduce and discuss various measures that gauge the accuracy of diagnostic tests. The accuracy is a test's ability to detect a condition correctly when the condition is truly present and to exclude the condition when it is actually absent. The accuracy of a test is always measured by comparing the test results to the true condition status. We assume that the true condition status is either "the condition is present" or "the condition is absent". For example, in medical studies, the true condition status is defined as the disease status. The outcome of test results from the test or tests under evaluation that reveals to us the true condition status of the patient is known as a 'gold standard'. Different gold standards are used for different applications in diagnostic tests.

1.3.1 Sensitivity and specificity

Sensitivity and specificity are two basic measures of diagnostic accuracy. We can illustrate the two definitions using the following contingency table, Table 1.1. Firstly, we denote the true condition status by the indicator variable T, where

$$T = \begin{cases} 1 & \textit{with condition}; \\ 0 & \textit{without condition}. \end{cases}$$

We denote the result of the diagnostic test by the indicator variable X. Test results indicating the condition's presence are called *positive*, denoted as $X = 1$, whereas those

indicating the condition's absence are called *negative*, denoted as $X = 0$, where

$$X = \begin{cases} 1 & postive\ \ test\ \ results; \\ 0 & negative\ \ test\ \ results. \end{cases}$$

Table 1.1 illustrates a basic count table specifying the different numbers under different categories. The total numbers with and without the condition are n_1 and n_0, respectively. The total numbers with the condition whose test result is positive and negative are, p_1 and p_0, respectively. The total numbers without the condition whose test result is positive and negative are, a_1 and a_0, respectively. The total number in the study is N, where $N = p_1 + p_0 + a_1 + a_0$.

Table 1.1: A basic count table

	Test results		
True condition status	Positive(X=1)	Negative(X=0)	Total
Present(T=1)	p_1	p_0	n_1
Absent(T=0)	a_1	a_0	n_0
Total	m_1	m_0	N

The *sensitivity (Se)* is the test's ability to detect the condition when the condition is present. The sensitivity is the probability that the test result is positive($X = 1$), given the presence of the condition ($T = 1$), written as

$$Se = P(X = 1|T = 1). \tag{1.1}$$

In table 1.1, among n_1 numbers with the condition, p_1 test positive. So, $Se = p_1/n_1$.

The *specificity (Sp)* is the test's ability to exclude the condition without the condition. It is the probability that the test result is negative($X = 0$), given the absence of the condition ($T = 0$), written as

$$Sp = P(X = 0|T = 0). \tag{1.2}$$

In table 1.1, among n_0 numbers with the condition, a_0 test positive. Thus, $Sp = a_0/n_0$.

We can also summarize the data by probabilities, as shown in Table 1.2. The consequences associated with the test results are also considered. The test can have two types of errors. One is false positive errors and another one is false negative errors. We define the *true positive fractions*(TPF) and *false positive fractions*(FPF) as follows:

$$false\ \ positive\ \ fraction = FPF = P(X = 1|T = 0), \tag{1.3}$$

$$true\ \ positive\ \ fraction = TPF = P(X = 1|T = 1). \tag{1.4}$$

False negative fraction(FNF) is 1-TPF. *True negative fraction*(TNF) is 1-FPF. The following table illustrates the relationship between them by probabilities.

Table 1.2: Probability table

	Test result		
True condition status	Positive($X = 1$)	Negative($X = 0$)	Total
Present(T=1)	$Se = p_1/n_1$	$FNF = p_0/n_1$	1.0
Absent(T=0)	$FPF = a_1/n_0$	$Sp = a_0/n_0$	1.0

In this usage, sensitivity is known as the TPF and specificity is known as TNF. Under various applications, the terminology for TPF and FPF is often different. In biomedical research, the 'sensitivity' (TPF) and 'specificity' (1-FPF) are often descriptors of test performance. In engineering and audiology, the terminologies 'hit rate' (TPF) and 'false alarm rate' (FPF) are often used. In statistical hypothesis testing, the terms 'significance level' (FPF) and 'statistical power' (TPF) are often used.

1.3.2 Predictive values

The accuracy of a diagnostic test can also be quantified by how well the test results predict the true condition status. As such, another important measure of a diagnostic test is *predictive value*. The predictive values depend on the prevalence of the condition, such as in a disease condition. The predictive values are:

$$positive\ \ predictive\ \ value = PPV = P(T = 1|X = 1), \tag{1.5}$$

$$negative\ \ predictive\ \ value = NPV = P(T = 0|X = 0). \tag{1.6}$$

A perfect test is one that predicts the condition perfectly. That is, PPV=1 and NPV=1. Contrarily, a useless test is one with no information about the true condition status. As such, a test which does not reflect the true condition status very well will result in a low PPV. The predictive values can tell us how likely the condition is given the test result. The values are affected by the prevalence of the condition. Low prevalence of the condition may be a reason for a low PPV. In research studies, both

the classification probability(TPF and FPF) and the predictive values are important and there is a direct relationship between the two. Suppose the prevalence is $\rho = P(T = 1)$. A result can be directly ascertained from the Bayes' theorem:

$$PPV = \frac{\rho TPF}{\rho TPF + (1-\rho)FPF},$$

$$NPV = \frac{(1-\rho)(1-FPF)}{(1-\rho)(1-FPF) + \rho(1-TPF)}.$$

1.3.3 Likelihood ratios

Another way to describe the diagnostic test is the *likelihood ratios*(LR), which is also widely used in research. We define *positive* and *negative* LRs as:

$$postive\ \ LR = LR(+) = \frac{P(X = 1|T = 1)}{P(X = 1|T = 0)}, \tag{1.7}$$

$$negative\ \ LR = LR(-) = \frac{P(X = 0|T = 1)}{P(X = 0|T = 0)}. \tag{1.8}$$

Note that the positive likelihood ratio is the the ratio of sensitivity to the FPF. The negative likelihood ratio is the ratio of the FNF to specificity. The likelihood ratios do not depend on the population prevalence, which are related to the classification probabilities and predictive values. The LR can quantify how much the diagnostic test changes knowledge of the condition status. An LR of 1.0 indicates that the test result is equally likely among the subjects with and without the condition; an LR greater than 1.0 means that the test result is more likely among the subjects with the condition than without the condition; an LR less than 1.0 indicates that the test result is more likely

among the subjects without the condition than with the condition. The higher the LR is, the likelier the test result is among the subjects with the condition relative to the subjects without the condition. We can also consider the odds that a subject has the condition before performing the test which is

$$pre-test\ \ odds = P(T=1)/P(T=0)\ .$$

We can consider the odds of the condition with the knowledge of the test result after performing the test which is

$$post-test\ \ odds = P(T=1|X)/P(T=0|X)\ .$$

We note that the post-test odds can be expressed in terms of the predictive values as:

$$post-test\ \ odds(X=1) = \frac{PPV}{1-PPV}\ ,$$

$$post-test\ \ odds(X=0) = \frac{1-NPV}{NPV}\ .$$

In this case, the likelihood ratios are related to these two odds, where

$$post-test\ \ odds(X=1) = LR(+)\times(pre-test\ \ odds)\ ,$$

$$post-test\ \ odds(X=0) = LR(-)\times(pre-test\ \ odds)\ .$$

1.4 Literature review

The measure of accuracy of a test we introduce is often based upon decision thresholds, which may be difficult to detect. Lusted(1971) illustrated a way in which we could

overcome the limitation of a single sensitivity and specificity pair, which he first applied to psychophysics. Lusted argued that the method could overcome the limitation by considering all of the decision thresholds. By applying the receiver operating characteristic (ROC) curve, we can describe the accuracy of a diagnostic test without the limitations of decision thresholds. Lusted stated that ROC curves offer an ideal means of examining the performance of the diagnostic tests. Subsequently, the ROC curve has been the most valuable and most widely used tool to describe and compare diagnostic tests in various disciplines of medicine.

An ROC curve is a plot of the sensitivity of a diagnostic test versus the false-positive fraction. ROC curves were originally developed for electronic signal-detection theory (Peterson, Birdsall and Fox, 1954). ROC curves and ROC analysis have subsequently formed the basis of statistical decision theory, having been applied to various medical and nonmedical studies, including studies of human perception (Drury and Fox, 1975) and military monitoring (Swets, 1977). Some features of ROC curves, which we discuss below, make them ideal for studying diagnostic tests.

In medical diagnostic testing, we are interested in measuring the observer's abilities for interpreting test results rather than the criteria used for such decisions. As such, Lusted (1971) discussed how in medical diagnostics, a distinction must be made between the observer's cognitive and sensory abilities to interpret the test results for detecting the condition and the observer's criteria used in deciding whether a condition is present or absent.

Swets and Pickett (1982) discussed how ROC curves display all possible cutpoints and thus can estimate the frequency of various outcomes at each cutpoint. Furthermore, ROC curves can apply previously generated probabilities of the condition, as well as calculations of the costs and benefits of correct and incorrect decisions, to determine the optimum cutpoint. They were also the first to study the analysis of multireader studies in which several observers interpret the test results of the same sample of patients. They identified several sources of variability, as well as correlations in multireader studies and then created a methodology for estimating and comparing the test accuracy for such studies.

The first to use the Gaussian model for estimating the ROC curve were Green and Swets (1966). They assumed the numerical value of a sensory event (defined as X) affects the observer's confidence about whether the condition is present or absent. They also assumed a cutpoint (defined as t) such that if $X < t$ and $X > t$, then the observer will choose the hypothesis that the condition is absent and present, respectively. Additionally, they assumed the Gaussian distribution of T under each hypothesis. Furthermore, Dorfman and Alf, Jr (1968, 1969) proposed maximum-likelihood estimates for the parameters of a binormal ROC curve, and provided methodologies for obtaining the variance-covariance matrix and the corresponding confidence intervals.

The most widely used summary measure for the test accuracy of ROC analysis is the area under the ROC curve (AUC). Hanley and McNeil (1982) provided a relatively simple methodology to estimate AUC without having to assume the distribution of the

test results. Interestingly, they noted that AUC is equivalent to the Wilcoxon 2-sample test statistic. They developed a method for calculating sample size for studies that apply the ROC curve area. Several other nonparametric methodologies have subsequently been developed for estimating and comparing ROC curves.

McClish (1989) stated that AUC was a global measure of a test's accuracy. He provided parametric methods for estimating and comparing the partial area under the ROC curve. These parametric methods are based upon a binormal model and parallel the MLEs of the area under the total ROC curve. Many statistical methods were developed shortly after these investigations for the estimation of the ROC analysis for two-way classification.

However, many real-world classification problems involve more than just two categories and the extension of the two-way ROC analysis is needed. Scurfield (1996) first mapped the mathematical definition of a proper ROC measure for more than two categories. Recently, ROC methodology was then extended to multiple-class diagnostic problems by introducing a three-dimensional ROC surface. Mossman (1999) introduced the concept of three-class ROC analysis into medical decision making. Nakas and Yiannoutsos (2004) were the first to consider the estimation of the volume under the ROC surface for ordered three-class problems by using U-statistic theory. Li and Fine (2008) further proposed the estimation of the volume under the ROC surface (VUS) and the hypervolume under the ROC manifold (HUM). They also provided the estimation of the multiple-class ROC measures and applied the multiple-class ROC analysis as a

model of selection criterion in microarray studies. Li and Zhou (2009) considered non-parametric and semiparametric estimation of the ROC surfaces by approximating the asymptotic ROC surfaces with multivariate Brownian bridge processes.

In medical research, it is also important to evaluate the various factors that can influence the medical performance. Great interest has been shown in developing methods for combining biomarkers. Statistical regression analysis has recently been studied to make inferences about such factors and biomarkers.

Han (1987) originally developed the maximum rank correlation estimator (MRC), which was considered as a generalized regression model of nonparametric analysis. It has recently been applied to assess classifications because of its close relationship to the ROC curve. Optimization algorithms that maximize the area under the ROC curve have also recently been proposed. Pepe (2003) developed optimal prognostic scores by applying binary regressions. The optimal linear combination is attained from several available diagnostic biomarkers from which we seek to maximize the area under the ROC curve among all the possible linear combinations in the binary data analysis. Enrique et al. (2004) suggested how to obtain the confidence interval for the generalized ROC criterion, conditional on given covariate values and derived some inferences under the normal distribution assumption. Theory of the consistency of the optimal confidence interval is based upon the argument which comes from Sherman (1993), relying on a general method for establishing the limiting distribution of a maximization estimator.

1.5 Aim and organization of the thesis

Over the last few decades, the most commonly used methods for evaluating the accuracy of numerical diagnostic tests in two-category classification problems have been the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) measure. AUC characterizes the probability that a test can correctly differentiate between two subjects. An effective diagnostic test has an AUC value greater than 1/2. However, as the number and breadth of applications for AUC and its related measures expand in the field of medicine and in clinical studies, we have noticed that the AUC values are at times actually lower than 1/2. Some researchers might ignore such AUC values as trivial data points. But in reality, they may be overlooking important test subjects, such as genes, for the classification. In this thesis, we pointed out a fundamental weakness int the AUC method of interpreting ROC curves, in particular improper ROC curve. We studied and examined the cases when the estimated AUC values are lower than 1/2. A better way to interpret the ROC curves is to examine the ratio of the likelihood of the test results with the condition and without the condition. We suggested to reverse the decision rule and use a screening method, providing significant further insight into the data and the diagnostic test itself.

Identifying the correct classification for multiple-category problems is comparatively more complicated. The volume under the ROC surface (VUS) and the hypervolume under the ROC manifold (HUM) are extensions of the AUC, extended for three or more category classifications. The nonparametric estimation of VUS or HUM is asso-

ciated with calculating the probability that three or more categories are simultaneously ordered correctly by the particular test. However the mathematical procedure to correctly predict the relative order is not as obvious as in the two-class problems. In this thesis, we consider parametric and nonparametric methods to address the elements of the multiple-category issue.

The U-statistic approach for calculating the variance of the non-parametric estimator of the area under the ROC curve has already been proposed. However, as sample sizes increase, the advantage of the U-statistic methodology is heavily diminished, and the U-statistic variance methodology for the case of multiple categories is generally not appropriate. To solve the computational burden as the dimension of the problem increases, we propose bootstrap standard errors for the multiple-category ROC analysis.

In practice, many factors can significantly influence the accuracy performance of a diagnostic test. Various information resources will also be available to assist in the medical prediction. However, at the core is the need to combine multiple biomarkers and factors in order to predict an accurate outcome. As such, great interest in developing methods for combining biomarkers is widespread. Here, we develop an optimization procedure by constructing a linear combination of markers that maximizes the VUS or HUM of the resultant combined marker. We also provide asymptotic theories for our estimators based upon the maximum rank correlation estimation.

Concerning the organization of the various subjects mentioned above, this thesis has been divided into five main chapters. Chapter 1 provides an introduction and review of

some of the basic accuracy measures of statistical ROC analysis.

In Chapter 2, we improve the procedure for the area under the ROC curve in the situation that the estimator of AUC is less than 1/2. In fact, contrary to some prevailing practices, the test with an AUC lower than 1/2 can still be shown to be useful for differentiating the two classes. We present a method which appears to rotate the ROC plot 180 degrees so that it emerges in the upper side of the chance diagonal line. An example is provided which pertains to an ovarian cancer dataset used in a population screening.

In Chapter 3, an extension of the two-class ROC analysis is proposed for three-category classification problems. The relationship between the area under the ROC curve and the volume under the ROC surface is examined. We propose approaches that assess the multiple categories by using simple summary statistics such as the sample mean. Moreover, a general discussion on the minimum acceptable HUM values is applied to multiple-category classification problems. The results of simulation studies we conducted that examine the performance of our proposed methods for sorting the unknown orders of multiple categories is also presented. We use microarray and mass spectrometry datasets to illustrate our methods.

In Chapter 4, we explore statistical methods of combining multiple tests for multiple-category classifications to optimize the accuracy of the combined marker under the criteria of ROC measures. Appropriate statistical procedures are developed by extending the maximum rank correlation estimators to high-dimensional cases. Simulation stud-

ies are then conducted to investigate the performance of the proposed inferences. We also apply our proposed methodology to two examples using data from recent health science studies.

In Chapter 5, we offer concluding remarks and discuss possible paths for future research.

Chapter 2

Two-class ROC Analysis

The ROC curve is considered the most well-developed statistical approach for describing and evaluating the performance of diagnostic tests. ROC curves have been used for a relatively long time. In 1966, Green and Swets developed signal detection theory in psychophysics, which appeared to be a potential method for medical diagnostic testing. In 1971, Lusted pointed out that this method could be adopted for medical decision making and stated that the method could overcome limitations of a single sensitivity and specificity pairs. Since then, this method has been the most valuable and popular tool for describing and comparing diagnostic tests, particularly in medicine.

2.1 The ROC curve

An ROC curve is a plot of the sensitivity of a test which is plotted on the *y* axis versus the test's FPF which is plotted on the *x* axis. Different decision thresholds can generate different points on the graph. Line segments are often used to connect the points from different possible decision thresholds, forming an *empirical ROC curve*. The diagonal line is called a *chance diagonal*.

Figure 2.1: An example of an ROC curve

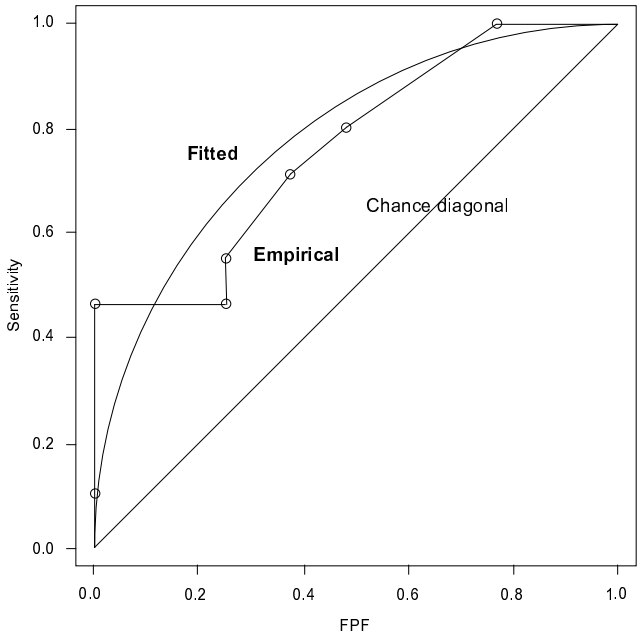


Figure 2.1 illustrates an example of an ROC curve. In this figure, each circle on the empirical ROC curve represents a (FPF, Se) point corresponding to a particular decision threshold. There are seven decision thresholds which provide (FPF, Se) points in addition to the two points, (0,0) and (1,1). Line segments connect all the points generated from the seven possible decision thresholds and then form a *empirical ROC curve*. It is also convenient to connect all the possible points using a smooth curve which is called a *fitted ROC curve*, illustrated in Figure 2.1.

Tests are usually ordinal in nature. For example, the clinical symptoms in medical research are often classified as severe, moderate, mild and not present. But it is often convenient to use a statistical model to fit the test results. Now we discuss the continuous ROC curves. We use a threshold r to define a binary test from the continuous test result X as

$$positive \ \ if \ \ X \geq r,$$

$$negative \ \ if \ \ X < r.$$

The corresponding true positive fraction at the threshold r TPF(r) and false positive fraction at the threshold r FPF(r) are defined as

$$TPF(r) = P(X \geq r|T = 1)\,, \tag{2.1}$$

$$FPF(r) = P(X \geq r|T = 0)\,. \tag{2.2}$$

The set of all possible TPFs and FPFs forms an ROC curve attained by dichotomizing X with different thresholds. That is, the ROC curve can be written as

$$ROC(\cdot) = \{(FPF(r), TPF(r)), r \in (-\infty, \infty)\}. \tag{2.3}$$

When $r = \infty$, then $lim_{r\to\infty} TPF(r) = 0$ and $lim_{r\to\infty} FPF(r) = 0$. When $r = -\infty$, then $lim_{r\to-\infty} TPF(r) = 1$ and $lim_{r\to-\infty} FPF(r) = 1$. We also notice that when the threshold r increases, both FPF(r) and TPF(r) decrease. Thus, the ROC curve is a monotone increasing function. The ROC curve can then be written as:

$$ROC(\cdot) = \{(t, ROC(t)), t \in (0, 1)\}, \tag{2.4}$$

where the ROC function maps t to TPF(r), and r is the threshold corresponding to FPF(r)=t.

Let (FPF(r), TPF(r)) be a point on the ROC curve for X. For any strictly increasing function h of X, we have $P(h(X) \geq h(r)|T = 0) = P(X \geq r|T = 0)$ and $P(h(X) \geq h(r)|T = 1) = P(X \geq r|T = 1)$. Thus, the ROC curve is invariant to strictly increasing transformations of X.

Let S_1 and S_2 denote the survivor functions for X with the condition and without the condition: $S_1(x) = P(X \geq x|T = 1)$ and $S_2(x) = P(X \geq x|T = 0)$. Let $r = S_2^{-1}(t)$ be the threshold corresponding to the FPF=t so that $P(X \geq r|T = 0) = t$. Therefore the ROC curve can also be represented as:

$$ROC(t) = S_1(S_2^{-1}(t)), \quad t \in (0, 1). \tag{2.5}$$

The ROC plot has many advantages compared to other measures of accuracy (Zweig and Campbell, 1993). An ROC curve can visually represent the data's accuracy. The scales of the ROC curve plot are two basic measures of accuracy which can be easily read from the plot. The ROC curve includes all the possible decision thresholds so that there is no requirement to select a particular decision threshold. Because sensitivity and specificity are independent of prevalence, the ROC curve is independent of prevalence as well. The ROC curve is also independent of the scale of the test results. That is, the ROC curve does not vary to any monotonic (e.g., linear, logarithmic) transformations of the test results, which is a useful property (Campbell, 1994). Another advantage of the ROC curve is that it can provide a direct and visual comparison of two or more tests on a single set of scales. It is possible to compare different tests at all decision thresholds by constructing the ROC curves.

2.2 Summary indices

Some summary indices associated with the ROC curve are often used to summarize the accuracy of a diagnostic test and provide important information about the ROC curve. When the ROC curve is not feasible to plot, such summary measures can also provide important information about the ROC curve. *Area under the ROC curve* (AUC) and *partial area under the ROC curve* (PAUC) are two important summary indices which are particularly useful in certain situations.

2.2.1 Area under the ROC curve

ROC curve is a useful measure to summarize the accuracy of a diagnostic test. Another valuable measure associated with the ROC curve is the *area under the ROC curve* (AUC). The area under the ROC curve takes values between 0.0 and 1.0. A perfect diagnostic test is one with an area under the ROC curve of 1.0 and consists of two line segments: (0,0)-(0,1) and (0,1)-(1,1). In contrast, a test with an area of 0.0 is perfectly inaccurate. However, perfect diagnostic tests are rare. The area under the ROC curve can be interpreted as the average of sensitivity for all possible values of specificity. It can also be interpreted as the average value of specificity for all possible values of sensitivity.

The area under the ROC curve is a widely used summary measure for comparing ROC curves which can be defined as (Bamber (1975))

$$AUC = \int_0^1 ROC(t)dt. \tag{2.6}$$

Obviously, if two tests A_1 and A_2 are ordered as

$$ROC_{A_1}(t) \geq ROC_{A_2}(t), \quad \forall t \in (0,1),$$

then the corresponding AUC statistics are also ordered as

$$AUC_{A_1} \geq AUC_{A_2}.$$

However, the converse of the above is not necessarily true.

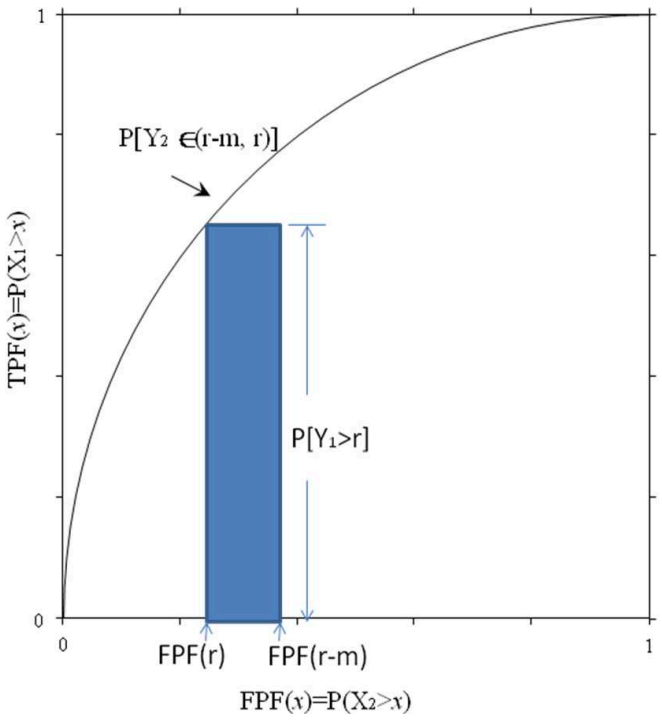
As discussed in the previous section, the ROC curve can be interpreted as

$$ROC(t) = S_1(S_2^{-1}(t)), \quad t \in (0, 1).$$

Here, we denote the test results with the condition as X_1 and the test results without the condition as X_2. Thus, we have

$$AUC = \int_0^1 ROC(t)dt = \int_0^1 S_1(S_2^{-1}(t))dt = \int_{-\infty}^{\infty} S_1(x)dS_2(x) = P(X_1 > X_2).$$

Figure 2.2: AUC=$P(X_1 > X_2)$



The AUC has another interpretation. It is equivalent to the probability that the test

results from randomly selected subjects with the condition and without the condition are correctly ordered, by the form of $P(X_1 > X_2)$, as illustrated in Figure 2.2 (Bamber,1975).

An important link between the area under the ROC curve and the Wilcoxon 2-sample rank-sum statistic or, the Mann-Whitney U-statistic exists. Note that the Mann-Whitney U-statistic is based upon an estimate of $P(X_1 > X_2)$, in which it is exactly the area under the ROC curve. So the properties of the Mann-Whitney U-statistic can be used to predict the statistical properties of the area under the ROC curve.

2.2.2 Partial area under the ROC curve

Another summary measure associated with the ROC curve is the *partial area under the ROC curve* (PAUC). There is particular interest in the area under a portion of the ROC curve. The partial area under the ROC curve is the area between two sensitivities, which can be defined as

$$PAUC(t_0) = \int_0^{t_0} ROC(t)dt,$$

where $t_0 \in (0, 1)$. Its values range from $t_0^2/2$ for a completely uninformative test to t_0 for a perfect test. Dwyer (1997) interpreted the partial area under the ROC curve as the probability that a randomly chosen subject without the condition will be classified correctly from a randomly chosen subject with the condition who tested negative in a diagnostic test. The partial area of test performance is appealing for some special cases and is also well established in many clinical tests.

2.3 The binormal ROC curve

The normal distribution is a classic and widely-used model to describe distribution functions. Now we apply the binormal distribution model to the ROC curve. The binormal ROC curve plays a significant role in ROC analysis. Suppose that the test results are normally distributed in the populations with the condition and without the condition. Assume

$$X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2).$$

For any threshold r, we have

$$TPF(r) = P(X_1 > r) = \Phi(\frac{\mu_1 - r}{\sigma_1})$$

and

$$FPF(r) = P(X_2 > r) = \Phi(\frac{\mu_2 - r}{\sigma_2}),$$

where Φ denotes the standard normal cumulative distribution function. We see that for a FPF t, the corresponding threshold is $r = \mu_2 - \sigma_2\Phi^{-1}(t)$. Hence,

$$ROC(t) = \Phi(\frac{\mu_1 - r}{\sigma_1}) = \Phi(\frac{\mu_1 - \mu_2}{\sigma_1} + \frac{\sigma_2}{\sigma_1}\Phi^{-1}(t)).$$

Then the AUC measure has an analytic form. Recall that $AUC = P(X_1 > X_2) = P(X_1 - X_2 > 0)$. The AUC can be represented with the binormal assumption as

$$AUC = P(X_1 - X_2 > 0) = \Phi(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}).$$

If we define $a_1 = \frac{\mu_1 - \mu_2}{\sigma_1}$ and $a_2 = \frac{\sigma_2}{\sigma_1}$, then the ROC curve and AUC measures can be written as

$$ROC(t) = \Phi(a_1 + a_2\Phi^{-1}(t)) \tag{2.7}$$

and

$$AUC = \Phi(\frac{a_1}{\sqrt{1 + a_2^2}}). \tag{2.8}$$

Recall that the ROC curve is invariant to monotone increasing transformations. If X_1 and X_2 are normally distributed and h is a monotone increasing function, then the ROC curve for the transformations $h(X_1)$ and $h(X_2)$ is also the binormal ROC curve

$$ROC(t) = \Phi(a_1 + a_2\Phi^{-1}(t)).$$

2.4 Estimating summary measures

We defined the ROC curve and introduced its properties in the previous section. We now discuss the statistical methodology for estimating the ROC curve and the summary measures. Firstly, we apply nonparametric empirical approaches to obtain the empirical ROC curve. Then we apply the parametric methods using statistical models to estimate the ROC curve and summary measures. Finally, the nonparametric methods will be introduced.

2.4.1 Empirical estimation

Assume that the numbers for the test results with and without the condition are n_1 and n_2 . X_{1_i} and X_{2_i} are selected randomly from the populations of test results with and without the condition, respectively. $\{X_{1_i}, i = 1, ..., n_1\}$ are identically distributed with the population survivor function $S_1(x) = P(X_{1_i} \geq x)$. Similarly, $\{X_{2_j}, j = 1, ..., n_2\}$ are identically distributed with the population survivor function $S_2(x) = P(X_{2_j} \geq x)$.

The empirical estimator of the ROC curve can easily be derived from the definition of the ROC curve. For each possible threshold c, the empirical TPF and FPF are calculated by

$$\widehat{TPF}(r) = \sum_{i=1}^{n_1} I\{X_{1_i} \geq r\}/n_1$$

and

$$\widehat{FPF}(r) = \sum_{j=1}^{n_2} I\{X_{2_j} \geq r\}/n_2,$$

where I is the indicator function. The empirical ROC curve can be considered as a plot of $\widehat{TPF}(r)$ versus $\widehat{FPF}(r)$ for all $r \in (-\infty, \infty)$. Therefore, the empirical ROC, $\widehat{ROC}$ can be directly obtained from the definition of ROC curve as

$$\widehat{ROC}(t) = \widehat{S}_1(\widehat{S}_2^{-1}(t)), \tag{2.9}$$

where $\widehat{S}_1$ and $\widehat{S}_2$ are the empirical survivor functions for X_1 and X_2, respectively.

Note that the empirical ROC curve is a function of the ranks of the data. It is related to the ordering of the test results and the status of the individuals with and without the condition.

Now we consider the sampling variability for the empirical ROC curve. One of the ways to assess the sampling variability is to assume the test results are continuous. Firstly, we fix the FPF t. Then we determine the estimated threshold corresponding to t. We then determine the proportion of the observations with the condition with test results above the threshold. Hsieh and Turnbull (1996) provided a result of variability of $\widehat{ROC}$ in the case of independent continuous test results. When the numbers of X_1 and X_2, n_1 and n_2, are large, the distribution of $\widehat{ROC}(t)$ is estimated by a normal distribution with mean $\mu_{ROC(t)}$ and variance given by

$$var(\widehat{ROC}(t)) = \frac{\mu_{ROC(t)}(1-\mu_{ROC(t)})}{n_1} + (\frac{g_1(c^*)}{g_2(c^*)})^2 \frac{t(1-t)}{n_2}, \tag{2.10}$$

where $c^* = S_2^{-1}(t)$, g_1 and g_2 denote the probability densities for X_1 and X_2, respectively.

This variance of $\widehat{ROC}(t)$ is broken into the sum of two components. The first component derives from the binomial variability of the estimated TPF when the threshold r is fixed. The second part derives from the estimation of $S_2^{-1}(t)$.

Similarly, the form of the confidence interval for the ROC(t) based upon the asymptotic normal approximation to the distribution of $\widehat{ROC}(t)$ is

$$\widehat{ROC}(t) \pm \Phi^{-1}(1-\frac{\alpha}{2}) \sqrt{\widehat{var}(\widehat{ROC}(t))},$$

where α is the significant level.

2.4.2 The estimation of the area under the ROC curve using parametric model

As defined in the previous section, a general form for the area under the ROC curve is

$$AUC = \int_0^1 ROC(t)dt.$$

When we assume binormality, this integral can be written as

$$AUC = \Phi(\frac{a_1}{\sqrt{1 + a_2^2}}),$$

where a_1 and a_2 are defined in the previous section. The AUC summary measure then is estimated with

$$\widehat{AUC} = \Phi(\frac{\widehat{a_1}}{\sqrt{1 + \widehat{a_2}^2}}) = \Phi(\frac{\widehat{\mu_1} - \widehat{\mu_2}}{\sqrt{\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2}}). \tag{2.11}$$

McClish (1989) derived the variance of AUC as

$$Var(\widehat{AUC}) = f_1^2 Var(\widehat{a_1}) + f_2^2 Var(\widehat{a_2}) + 2 f_1 f_2 Cov(\widehat{a_1}, \widehat{a_2}),$$

where

$$f_1 = \frac{e^{-a_1^2/2(1+a_2^2)}}{\sqrt{2\pi(1 + a_2^2)}} \quad and \quad f_2 = -\frac{a_1 a_2 e^{-a_1^2/2(1+a_2^2)}}{\sqrt{2\pi(1 + a_2^2)^3}}.$$

The variance can be estimated by substituting estimators for the parameters a_1 and a_2.

2.4.3 The estimation of the area under the ROC curve using non-parametric model

AUC can also be estimated directly from the nonparametric method without making any distributional assumptions. The estimation can be directly obtained by summing the trapezoidal areas which are formed by connecting all the possible points of the ROC curve.

Figure 2.3: The trapezoidal rule

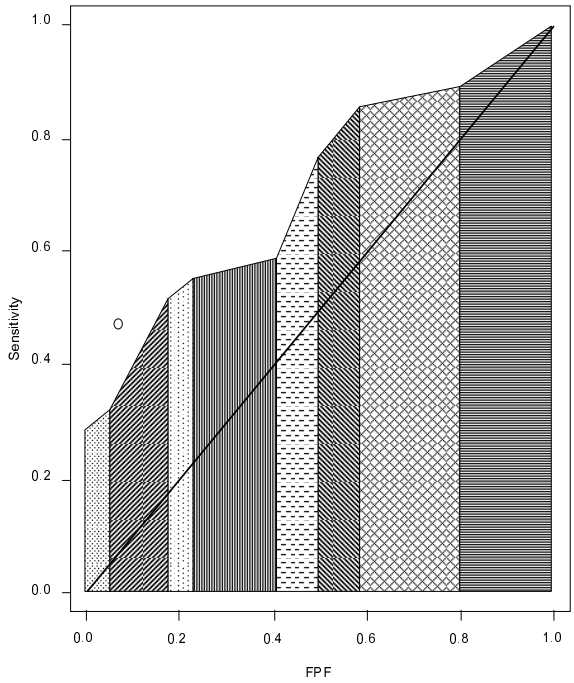


Figure 2.3 illustrates the area calculated by the trapezoidal method formed by connecting all the possible points.

By increasing the number of the possible threshold points, the bias of the estimation can be significantly reduced and make it acceptable for the estimation.

It is noted that AUC is equivalent to the Mann-Whitney U-statistic. Therefore, AUC can be estimated by

$$\widehat{AUC} = \sum_{j=1}^{n_2}\sum_{i=1}^{n_1}[I\{X_{1_i} > X_{2_j}\} + \frac{1}{2}I\{X_{1_i} = X_{2_j}\}]/n_1 n_2. \tag{2.12}$$

The corresponding variance is relatively complicated. A number of methods used to estimate the variance of the nonparametric area have been recommended. One result from Hanley and McNeil (1982) is given by

$$var(\widehat{AUC}) = \frac{AUC(1-AUC) + (n_1-1)(M_1 - AUC^2) + (n_2-1)(M_2 - AUC^2)}{n_1 n_2},$$

where

$$M_1 = P(X_{1_i} \geq X_{2_j}, X_{1_i'} \geq X_{2_j}),$$

$$M_2 = P(X_{1_i} \geq X_{2_j}, X_{1_i} \geq X_{2_j'}),$$

in which $(X_{1_i}, X_{1_i'})$ denotes the randomly selected pair of observations from the population with the condition and $(X_{2_j}, X_{2_j'})$ denotes the randomly selected pair of observations from the population without the condition.

Another nonparametric approach is using the kernel smoothing method to provide a smoothed ROC curve. For the kernel method, there are two parameters that need to be specified; the choice of kernel and the choice of bandwidth.

Zou, Hall and Shapiro (1997) suggested a kernel method to estimate a smooth ROC curve from continuous data. The Gaussian kernel was chosen. They recommended estimating the points on the ROC curve through the integral of the density function with the condition $f_1(x)$ and the density function without the condition $f_2(x)$, where the density functions are estimated as

$$\hat{f}_i(x) = \frac{1}{n_i h_i} \sum_{j=1}^{n_i} k(\frac{x - X_{ij}}{h_i}), \quad i = 1, 2.$$

The function k is called the kernel and h_i is the bandwidth. There can be numerous choices of kernel and bandwidth. They suggested using the kernel

$$k(\frac{x - X_{ij}}{h_i}) = \frac{15}{16}[1 - (\frac{x - X_{ij}}{h_i})^2]^2 \quad for \;\; x \in (X_{ij} - h_i, X_{ij} + h_i) \,,$$

where $k = 0$ otherwise, and the bandwidth

$$h_i = 0.9min(SD, IQR/1.34)/\sqrt[5]{n_i} \,,$$

where SD is the standard deviation and IQR is the interquartile range for the observations of subjects with the condition and without the condition.

The kernel estimator is reasonable when the choice of bandwidth is chosen and the sample size is large. However, it is difficult to prove that the resulting smoothed ROC curve will increase in a monotone manner. Therefore it is not widely applied in the real data analysis.

2.5 Cases when AUC is lower than 1/2

2.5.1 The method

Most ROC curves lie between those of the perfect and useless tests, which is above the diagonal chance line and below the left and upper border of the positive unit quadrant. A useless test corresponds to a test which produces identical distributions for X_1 and X_2. However, a diagnostic test can have an ROC curve with a hook, a portion of the ROC curve lying below the chance diagonal line. These curves are called *improper ROC curves*.

AUC can be interpreted as the probability that a test correctly differentiates between two subjects, one with the condition and one without the condition, which is equivalent to $P(X_1 > X_2)$. A useless test corresponds to a test which produces identical distributions for X_1 and X_2 and has an AUC value equal to 1/2 with an ROC curve on the chance diagonal line. An effective diagnostic test has an AUC value greater than 1/2. The area under the improper ROC curve then will have an AUC value smaller than 1/2. This could happen especially often in large scale microarray studies where thousands of genes are compared for their disease differential abilities according to their AUC values. However, we have sometimes noticed that researchers might overlook this issue and report AUC values lower than 1/2. Without a proper arrangement of the order of the two groups for individual genes and simply reporting $P(X_1 > X_2)$ uniformly for all the genes, it is likely that we might miss some important genes whose AUC should be

defined by $P(X_2 > X_1)$. There can be one fundamental weakness in the AUC method of interpreting ROC curves when the improper ROC curves exists. In fact, the test with an AUC lower than 1/2 can still be useful for differentiating the two classes and should be regarded as a better test than the one with AUC value of 1/2.

An idea for correcting this problem is to rotate the plot by 180 degrees, illustrated in Figure 2.4. Then it will appear in the upper side of the chance diagonal line, from graph (b) to graph (a) in Figure 2.4. A better way to interpret ROC curves is to examine the ratio of the likelihood of X_1 and X_2, in the spirit of Neyman-Pearson. For example, if the support of X_1 and X_2 are disjoint, then we have a perfect test, but the AUC need not to be 1 or 0. In particular, it can take the value of 0.5. This idea leads to a correct AUC definition as the probability $P(X_2 > X_1)$ instead of the rigid stipulation of $P(X_1 > X_2)$.

To make the 180 degree rotation, the ROC curve can easily be changed to appear above the chance diagonal line by reversing the decision rule. This screening method can assure that the ROC curves are correct and useful. Therefore, in practice, if we obtain an AUC value lower than 1/2, we use one minus this value to produce the correct AUC value, which is

$$AUC = \begin{cases} AUC & if \;\; AUC \geq 1/2; \\ 1 - AUC & if \;\; AUC < 1/2. \end{cases}$$

The nonparametric estimation of the improved AUC will be the same as the estimation of AUC in the previous section.

2.5.2 Example

One such example was evidenced in a recent statistical publication. Pepe et al. (2003) analyzed a publicly available ovarian cancer dataset used in a population screening. This dataset was obtained from a gene-expression experiment using glass arrays for 1536 cDNA clones studied by Dr Michel Schummer (Institute for Systems Biology, Seattle). It is a case-control study with 1536 potential diagnostic tests. The scientific objective from the dataset is to identify genes which are differentially expressed in ovarian cancer tissue, compared with the normal ovarian tissue. The experimental data were used to rank potential genes according to some statistical measure characterizing differential expression. They considered statistical methods to rank genes (or proteins) in regards to differential expression between tissues and argued that two measures related to the ROC curve are particularly suitable for their purpose.

In their paper, Pepe et al. focused on the detection of overexpressed genes, whereas the adaptation of the methods for the detection of underexpressed genes is relatively straightforward. Pepe et al. stated that there were many genes overexpressed in cancer tissue making the detection of screening markers difficult. Thus, they suggested to select a sizeable number of overexpressed genes to arrive at a subset which might have potential for screening. Using subsets was effective because clinical assays for some gene products were difficult to develop for technical reasons. In their methods, if one gene proved useless for biomarker development, they pursued yet another that could potentially identify the same cancers. They chose the first 100 genes in the dataset and

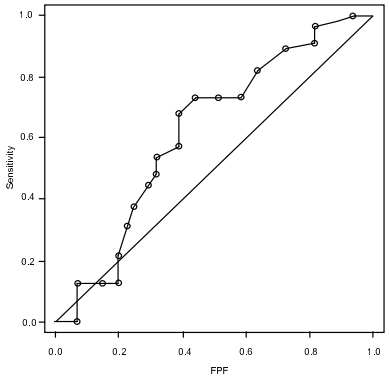
displayed the top 10 genes ranked according to AUC values. Gene 93 was ranked as the optimal one with the largest AUC value of 0.971, with the range of the top 10 gene ranking as (0.736, 0.971).

We did a similar calculation on AUC for these 100 genes with an appropriate adjustment for the order of X_1 and X_2 for each gene. Among the 100 genes, 51 genes have AUC values lower than 1/2, some of which are even close to 0. It might be because of the improper ROC curve existence. One reason is that when the sensitivity and FPR are calculated, the criterion or the decision rule is inappropriate for some of this dataset or the author used a single decision rule at the same time while the size of the variables are large. Another reason may be because of the imperfect laboratory techniques for measuring gene expression with microarrays.

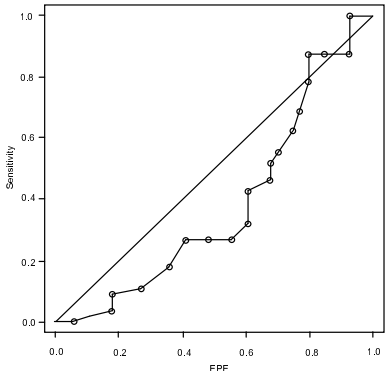
We applied our improved method by rotating the original ROC curve by 180 degrees to correct the ROC curve. After calculating the estimation of AUC using the nonparametric approach we mentioned, this resulted in new AUC values for the first 100 genes. Our results were compared with Pepe et al. (2003) in Figure 2.5. Surprisingly, a totally different ranking appears and only one of the top 10 genes agrees with Pepe et al. The first column in the table is the AUC values of the first top ranking from the paper of Pepe et al. The second column is the AUC values of the first 20 top ranking after the correction in our improved approach. The last column is the corresponding AUC values for each gene. Boxed genes represent the top 10 genes with the largest AUC values in Pepe et al. (2003). Circled genes represent the top 20 genes that were not identified

by Pepe et al. (2003). Our second highest AUC is 0.933 which might have been mistakenly calculated as 0.067 and thus placed at the bottom of the ranked table by the authors. The second highest AUC in Pepe et al. (2003) is only ranked 11th on our list. Nine genes with AUC higher than this one were unnecessarily screened out previously. Consequently the gene ranking from such an analysis may mislead the subsequent medical decision making. Our new improved approach enhances the process of identifying the biomarkers and allows the screening to be more accurate, informative and inclusive.

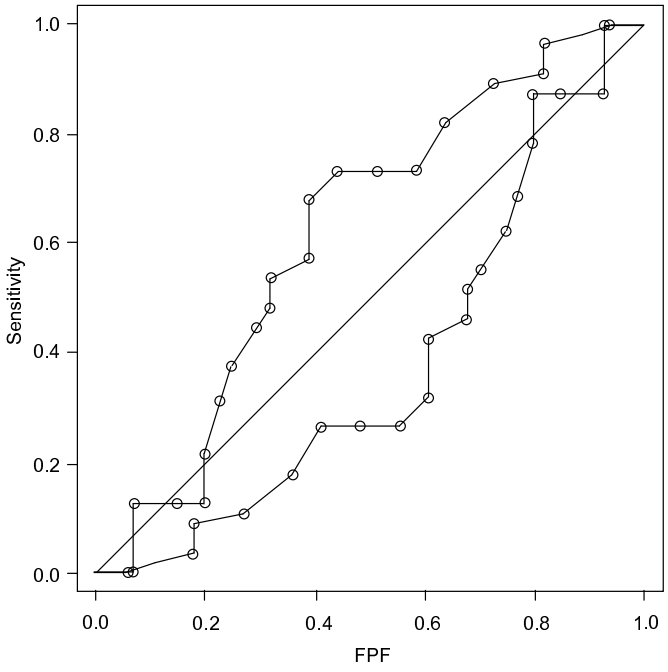
Figure 2.4: Improved method



(a) After rotating by 180 degrees

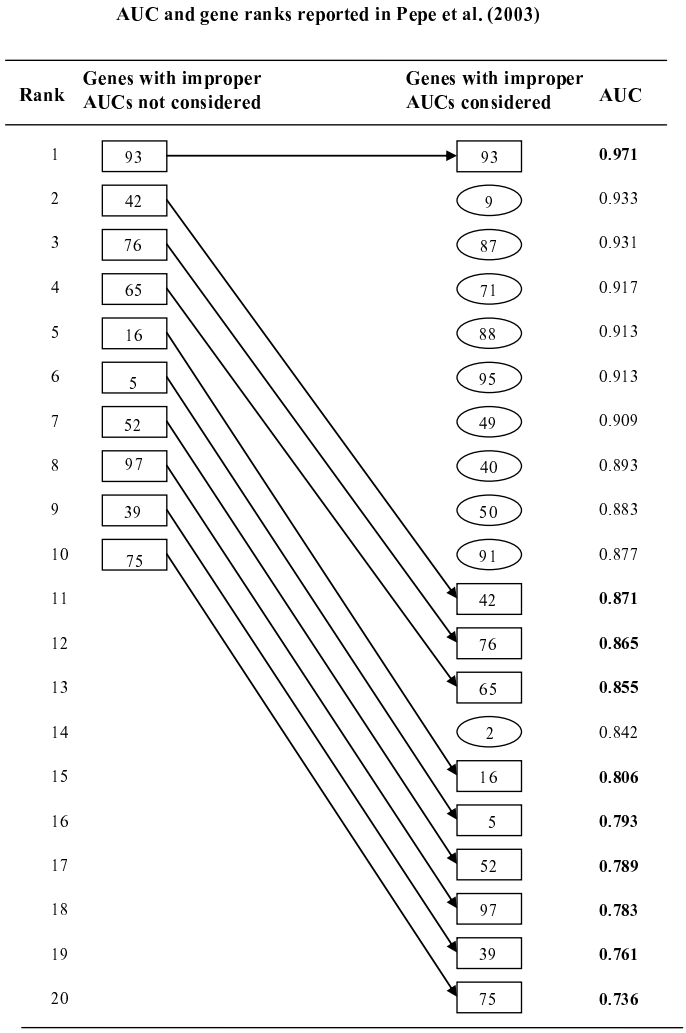


(b) Before rotating



(c) To make the plot be above the chance diagonal by rotating 180 degrees

Figure 2.5: AUC and gene ranks reported in Pepe et al. (2003)



Box: top 10 genes with the largest AUCs in Pepe et al. (2003)
Circle: genes in the top 20 that were not identified in Pepe et al. (2003)
Note: the AUC values of all genes were estimated using the new method where the values for the top 10 genes identified by Pepe et al (shown in bold font) remain unchanged.

Chapter 3

Sorting Multiple Classes in Multiple-category ROC Analysis

As we discussed in the previous chapter, the ROC curve is a useful statistical tool to evaluate the accuracy of continuous diagnostic tests. The ROC curve and AUC are adequate to assess the two-category classifications. However, many real-world biomedical situations have more than two classes. For example, in practice, it is more crucial to predict the stage of a disease rather than to only distinguish between a disease and non-disease state. A major limitation of the two-class ROC analysis is that it can not give a complete picture of how well a test discriminates between more than two classes. Thus, ROC analysis methods capable of handling multiple classes are essential to fully assess diagnostic performance. Unsurprisingly, there is great interest in the medical research field to develop methods for multiple-category classification.

3.1 Assessing three-class problems

3.1.1 ROC surface

Scurfield (1996) proposed the three-class ROC surface which is an extension of the ROC curve. Consider three classes, denoted l_1, l_2, and l_3. The observer's decision for the classification is based upon three decisions, denoted d_1, d_2, and d_3. We consider the three variables X_1, X_2, and X_3 as the test result variables from three classes, say Class I, Class II and Class III. These three variables can be represented as conditional random variable on variable X. Suppose the observation value is x, which is a particular value of the random variable X. Assume that the observer's decision is made with reference to the values of two thresholds, denoted r_1 and r_2 ($r_1 \leq r_2$). The observer uses the two thresholds to partition X into three intervals.

If $r_1 < r_2$, the observer's decision rule is as follows:

$$\begin{cases} if \;\; x < r_1, & then \;\; d_1, \\ if \;\; r_1 < x < r_2 & then \;\; d_2, \\ if \;\; r_2 < x & then \;\; d_3. \end{cases}$$

The values of the thresholds r_1 and r_2 are determined by the prior probabilities of the classes and by the costs associated with each decision outcome as well. For instance, if it is known that the second class l_2 occurs more often than the other two classes, then the width of the interval between r_1 and r_2 should be constructed so as to encompass a significant portion of the X_2 distribution. One assumption is that the observer will guess

whenever x coincides with the value of one of the two thresholds. This guessing will occur only when X is discrete, which can be summarized in Table 3.1.

Table 3.1: Decision probabilities

	Decision probability		
Condition	$P(d_1)$	$P(d_2)$	$P(d_3)$
$x = r_1 < r_2$	p_{11}	p_{12}	0
$r_1 < r_2 = x$	0	p_{22}	p_{23}
$x = r_1 = r_2$	p_{31}	p_{32}	p_{33}

The decision probabilities can be summarized in Table 3.2. The sum of the decision probabilities is equal to one across each row. We notice that when $x = r_1 < r_2$, $P(d_3) = 0$. That is because r_1 is associated with both the decision alternatives d_1 and d_2. Similarly, when $r_1 < r_2 = x$, $P(d_1) = 0$ because r_2 is associated with d_2 and d_3.

Table 3.2: Probability table

	Decision		
Class	d_1	d_2	d_3
l_1	$P(d_1\|l_1)$	$P(d_2\|l_1)$	$P(d_3\|l_1)$
l_2	$P(d_1\|l_2)$	$P(d_2\|l_2)$	$P(d_3\|l_2)$
l_3	$P(d_1\|l_3)$	$P(d_2\|l_3)$	$P(d_3\|l_3)$

In Table 3.2, each entry in the table specifies the probability that a particular decision is made given the presence of a particular class. The table has six degrees of freedom and the sum of all the probabilities across each row is equal to one. The decision rule

is based upon r_1 and r_2. Therefore, the probabilities that a particular decision is made given the presence of a particular class can also be represented as

$$P(d_1|l_{i_1}) = P(X_{i_1} < r_1), \tag{3.1}$$

$$P(d_2|l_{i_2}) = P(r_1 < X_{i_2} < r_2), \tag{3.2}$$

$$P(d_3|l_{i_3}) = P(r_2 < X_{i_3}), \tag{3.3}$$

where $\{i_1, i_2, i_3\}$ is a permutation of $\{1, 2, 3\}$.

The surface generated by these equations, conveyed as the two criteria vary over the domain of X, is called the $i_1i_2i_3$-*ROC surface*. In total, there are six ROC surfaces. All the six ROC surfaces are associated with the three decisions d_1, d_2 and d_3 paired with the three classes l_1, l_2 and l_3, respectively.

If X is discrete, the probabilities of decisions conditional on a particular class will be associated with those in Table 3.1, described as follows:

$$P(d_1|l_{i_1}) = \begin{cases} P(X_{i_1} < r_1) + p_{11}P(X_{i_1} = r_1); & r_1 < r_2 \\ P(X_{i_1} < r_1) + p_{31}P(X_{i_1} = r_1); & r_1 = r_2 \end{cases}, \tag{3.4}$$

$$P(d_2|l_{i_2}) = \begin{cases} P(r_1 < X_{i_2} < r_2) + p_{12}P(X_{i_2} = r_1) + p_{22}P(X_{i_2} = r_2); & r_1 < r_2 \\ p_{32}P(X_{i_2} = r_1); & r_1 = r_2 \end{cases}, \tag{3.5}$$

$$P(d_3|l_{i_3}) = \begin{cases} P(r_2 < X_{i_3}) + p_{23}P(X_{i_3} = r_2); & r_1 < r_2 \\ P(r_2 < X_{i_3}) + p_{33}P(X_{i_3} = r_1); & r_1 = r_2 \end{cases}. \tag{3.6}$$

3.1.2 Volume under the ROC surface

The *volume under each ROC surface* (VUS) is related to the distinctions of the three classes. If X_1, X_2, and X_3 are identically distributed, then equations 3.1 – 3.3 and equations 3.4 – 3.6 indicate that

$$P(d_1|l_{i_1}) + P(d_1|l_{i_2}) + P(d_1|l_{i_3}) = 1.$$

A fundamental result is that the volume under the $i_1 i_2 i_3$ ROC surface will be a sum of probabilities as follows:

$$VUS = P(X_{i_1} > X_{i_2} > X_{i_3}) + \frac{1}{2}P(X_{i_1} > X_{i_2} = X_{i_3}) + \frac{1}{2}P(X_{i_1} = X_{i_2} > X_{i_3}) + \frac{1}{6}P(X_{i_1} = X_{i_2} = X_{i_3}).$$

If X is continuous, then the last three components on the right-hand side are all zero. That is, VUS can be expressed as

$$VUS = P(X_{i_1} > X_{i_2} > X_{i_3}). \tag{3.7}$$

The VUS accounts for six orderings of X_1, X_2, and X_3 when considering all the permutations. The six orderings are mutually exclusive and exhaustive. Hence, it follows that the sum of the six VUSs will be equal to one. That is,

$$\sum_{i_1 i_2 i_3} VUS_{i_1 i_2 i_3} = 1.$$

The ROC surfaces show how well the observer can discriminate between all the three classes and also show how well the observer can discriminate between each pair of the three classes.

The ROC surface and VUS are two measures which are extensions of the two-class ROC curve and AUC. Now we focus on the relationship between VUS and AUC. Traditionally, the ROC curve is a plot of the FPF versus the TPF. Recall that one fundamental result of the theory of signal detectability provided by Bamber stated that the area under the 12-, 13-, 23-ROC curve could be written as

$$AUC = P(X_{i_i} > X_{i_2}) + \frac{1}{2}P(X_{i_1} = X_{i_2}),$$

where $\{i_1, i_2\}$ is $(1, 2)$ or $(1, 3)$ or $(2, 3)$.

As discussed, AUC is related to a particular ordering of X_1 and X_2. If X is continuous, then the second component on the right-hand side is zero. AUC is equal to the probabiity $P(X_{i_1} > X_{i_2})$. There are three ways that X_1, X_2, and X_3 can be ordered such that $X_{i_1} > X_{i_2}$ in the ordering. Either $X_{i_1} > X_{i_2} > X_{i_3}$, or $X_{i_1} > X_{i_3} > X_{i_2}$, or $X_{i_3} > X_{i_1} > X_{i_2}$. Therefore,

$$AUC = P(X_{i_i} > X_{i_2}) = P(X_{i_1} > X_{i_2} > X_{i_3}) + P(X_{i_1} > X_{i_3} > X_{i_2}) + P(X_{i_3} > X_{i_1} > X_{i_2}).$$

It is equal to say that

$$AUC_{i_1 i_2} = VUS_{i_1 i_2 i_3} + VUS_{i_1 i_3 i_2} + VUS_{i_3 i_1 i_2}.$$

In this case, AUC can be determined from the volumes under different ROC surfaces because of their relationship. Each area under the $i_1 i_2$ ROC curve can be represented by a sum of VUSs with a special ordering. Thus, there exists a linear relationship between

them which can be written as

$$\begin{pmatrix} A_{12} \\ A_{13} \\ A_{21} \\ A_{23} \\ A_{31} \\ A_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} VUS_{123} \\ VUS_{132} \\ VUS_{213} \\ VUS_{231} \\ VUS_{312} \\ VUS_{321} \end{pmatrix}.$$

However, generally, VUS cannot be determined from AUC. Notice that the 6×6 matrix on the right side has a determinant of 0. That is, this matrix is singular and has no inverse matrix. Therefore, this linear equation cannot be inverted to express the VUS in terms of AUC.

VUS is equivalent to the probability of correctly classifying the three classes. We know that the probability can be calculated by the integral form of the density function in the continuous case. Here we use integration to express the VUS. The probabilities that a particular decision is made given the presence of a particular class are based upon the two criteria r_1 and r_2. When the two criteria r_1 and r_2 vary over the domain, the volume under the $i_1 i_2 i_3$ ROC surface should be

$$VUS = \int_{-\infty}^{+\infty} \int_{-\infty}^{r_2} p(d_1|l_{i_1})|J| dr_1 dr_2,$$

where

$$|J| = \begin{vmatrix} \dfrac{\partial p(d_2|l_{i_2})}{\partial r_1} & \dfrac{\partial p(d_2|l_{i_2})}{\partial r_2} \\ \dfrac{\partial p(d_3|l_{i_3})}{\partial r_1} & \dfrac{\partial p(d_3|l_{i_3})}{\partial r_2} \end{vmatrix},$$

and (i_1, i_2, i_3) is a permutation of (1,2,3).

The probabilities that a particular decision is made given the presence of a particular class can also be expressed by the integral form of the corresponding density function as follows:

$$p(d_1|l_{i_1}) = \int_{-\infty}^{r_1} f(x|l_{i_1})dx,$$

$$p(d_2|l_{i_2}) = \int_{r_1}^{r_2} f(x|l_{i_2})dx,$$

$$p(d_3|l_{i_3}) = \int_{r_2}^{\infty} f(x|l_{i_3})dx,$$

where f is the probability density function for the continuous case. Then $|J|$ can be written as

$$|J| = \begin{vmatrix} -f(r_1|l_{i_2}) & f(r_2|l_{i_2}) \\ 0 & -f(r_2|l_{i_3}) \end{vmatrix} = f(r_1|l_2)f(r_2|l_3).$$

Therefore, VUS can be expressed by integral form as

$$VUS = \int_{-\infty}^{+\infty} \int_{-\infty}^{r_2} \int_{-\infty}^{r_1} f(x|l_{i_1})f(r_1|l_{i_2})f(r_2|l_{i_3})dxdr_1dr_2 ,$$

where $-\infty < x \leq r_1 \leq r_2 < \infty$.

In the previous chapter we applied the binormal distribution model to the ROC curves which plays a significant role in ROC analysis. In the three-class ROC analysis, we also apply the normal distribution model to explore its properties. Recall that AUC under the binormal distribution assumption has a form of

$$AUC = \Phi(\frac{a_1}{\sqrt{1 + a_2^2}}),$$

where a_1 and a_2 are defined as before.

Similarly, we assume the normal distribution for the three-class case. Here for simplicity, we only consider the case under the 123-ROC surface in which the VUS is the probability $P(X_1 > X_2 > X_3)$. The other five VUSs will have similar forms. Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, and $X_3 \sim N(\mu_3, \sigma_3^2)$. X_1, X_2, and X_3 are mutually independent. Then,

$$\begin{aligned} VUS = P(X_1 > X_2 > X_3) &= \int\int\int_{x_1>x_2>x_3} f_1(x_1) f_2(x_2) f_3(x_3) dx_1 dx_2 dx_3 \\ &= \int_{-\infty}^{\infty} dx_3 \int_{x_3}^{\infty} dx_2 \int_{x_2}^{\infty} f_1(x_1) f_2(x_2) f_3(x_3) dx_1 \\ &= \int_{-\infty}^{\infty} dx_3 \int_{x_3}^{\infty} f_2(x_2) f_3(x_3)[1 - F_1(x_2)] dx_2 \\ &= \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{x_2} f_2(x_2) f_3(x_3)[1 - F_1(x_2)] dx_3 \\ &= \int_{-\infty}^{\infty} F_3(x_2)[1 - F_1(x_2)] f_2(x_2) dx2 \\ &= \int_{-\infty}^{\infty} F_3(y) S_1(y) f_2(y) dy. \end{aligned}$$

Now we write the density function of X_2, $f_2(x_2)$, as the deviation of the probability function F_2. Thus,

$$\begin{aligned} VUS = P(X_1 > X_2 > X_3) &= \int_{-\infty}^{\infty} F_3(y) S_1(y) f_2(y) dy \\ &= \int_{-\infty}^{\infty} F_3(y) S_1(y) [F_2(y)]' dy \\ &= \int_{-\infty}^{\infty} \Phi(\frac{y - \mu_3}{\sigma_3}) \Phi(\frac{-(y - \mu_1)}{\sigma_1}) \varphi(\frac{y - \mu_2}{\sigma_2}) \cdot \frac{1}{\sigma_2} dy. \end{aligned}$$

Let $z = y - \mu_2/\sigma_2$, then $y = \sigma_2 z + \mu_2$. Then,

$$\begin{aligned} VUS = P(X_1 > X_2 > X_3) &= \int_{-\infty}^{\infty} \Phi(\frac{\sigma_2 z + \mu_2 - \mu_3}{\sigma_3}) \Phi(\frac{-(\sigma_2 z + \mu_2 - \mu_1)}{\sigma_1}) \varphi(\frac{\sigma_2 z + \mu_2 - \mu_2}{\sigma_2}) dz \\ &= \int_{-\infty}^{\infty} \Phi(\frac{\sigma_2}{\sigma_3} z + \frac{\mu_2 - \mu_3}{\sigma_3}) \Phi(-\frac{\sigma_2}{\sigma_1} z + \frac{\mu_1 - \mu_2}{\sigma_1}) \varphi(z) dz. \end{aligned}$$

Let $a_1 = \frac{\sigma_2}{\sigma_3}, a_2 = \frac{\mu_2 - \mu_3}{\sigma_3}, a_3 = -\frac{\sigma_2}{\sigma_1}, a_4 = \frac{\mu_1 - \mu_2}{\sigma_1}$, then the VUS can be written as

$$VUS = \int_{-\infty}^{\infty} \Phi(a_1 z + a_2)\Phi(a_3 z + a_4)\varphi(z)dz \ .$$

We will further discuss and examine the multivariate normal distribution assumption in the next section.

In two-class ROC analysis, a useless test is one that produces an identical distribution for X_1 and X_2 and has an AUC value equal to 1/2. Most tests will have a AUC value greater than 1/2. The lower bound for AUC is 1/2 which is the probability that a continuous random variable is greater than an identically distributed random variable. For the three-class ROC analysis, the probability of the three continuous identically distributed random variables, ordered in a special ordering, can also be calculated. We now assume that X_1,X_2 and X_3 are three identically-distributed random variables. The volume under the ROC surface corresponds to the probability that

$$\begin{aligned}
VUS = P(X_1 > X_2 > X_3) &= \int\int_{x_1>x_2>x_3} f(x_1)f(x_2)f(x_3)dx_1dx_2dx_3 \\
&= \int_{-\infty}^{\infty}\int_{x_3}^{\infty}\int_{x_2}^{\infty} f(x_1)f(x_2)f(x_3)dx_1dx_2dx_3 \\
&= \int_{-\infty}^{\infty}\int_{x_3}^{\infty} f(x_3)f(x_2)[1-F(x_2)]dx_2 \\
&= \int_{-\infty}^{\infty} f(x_3)\frac{1}{2}[1-F(x_3)]^2dx_3 \\
&= -\frac{1}{2}\int_{-\infty}^{\infty}[1-F(x_3)]^2 d[1-F(x_3)] \\
&= -\frac{1}{2}\cdot\frac{1}{3}[1-F(x_2)]^3|_{-\infty}^{\infty} = \frac{1}{6}.
\end{aligned}$$

In two-category classification, rejecting the null hypothesis that AUC is equal to 1/2 would imply that the test is able to differentiate between the two classes with a

probability higher than that of a random guess. For a three-category classification, we require the test to have at least some ability to differentiate three categories instead of only two categories. If we reject the null hypothesis that VUS is equal to $1/6$, we can only argue that the test is not the one that completely guesses the three classes. In fact, the test with a VUS greater than $1/6$ might be able to differentially pick out one class but completely guess the other two classes. In that case, the test is still useless for a three-category classification and cannot be recommended for use. For any three-category classifier, it has several pairwise AUCs. We should screen out those tests with any of these pairwise AUC values being too close to 1/2. The lower bound of VUS in three-category ROC analysis should be jointly considered with the lower bound of AUC in pairwise two-category ROC analysis.

3.1.3 Estimation of the volume under the ROC surface

AUC can be predicted by the extensively-studied properties of the Mann-Whitney statistic (or U-statistic). The relationship between AUC and this statistic enables us to estimate the AUC value and its properties without distribution and decision variable assumptions. In this section, we discuss the estimation method for three-class ROC analysis.

Consider that each individual underwent the examination and the test values are recorded. The test results $X_{1,i}$ $(i = 1, \cdots, n_1)$ are i.i.d. with distribution G_1; the test results $X_{2,j}$ $(j = 1, \cdots, n_2)$ are i.i.d. with distributions G_2; and the test results $X_{3,k}$

($k = 1, \cdots, n_3$) are i.i.d. with distributions G_3. G_1, G_2 and G_3 are continuous probability distributions on $\mathbb{R}$. As defined, X_1, X_2 and X_3 are independent to each other as they are obtained from different subjects.

VUS is used to summarize the overall accuracy of the test (Mossman (1999)). A summary index about the distinguishing and discriminatory performance of the test for the three classes is generated using this approach. Here VUS is mathematically equivalent to the probability $P(X_1 > X_2 > X_3)$. Similar to the unbiased nonparametric estimator of AUC, one nonparametric estimator of VUS is suggested with a three-sample U-statistic:

$$\widehat{VUS} = n_1^{-1} n_2^{-1} n_3^{-1} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \sum_{i_3=1}^{n_3} I\{X_{1i_1} > X_{2i_2} > X_{3i_3}\}. \tag{3.8}$$

where I is the indicator function.

Inference for AUC is based upon U-statistic which we have already discussed. Extending to the three-class problem, the developed U-statistic methodology is still feasible. The variance for the estimated VUS will be discussed as a M-category classification case in the next section.

3.2 Sorting multiple classes in multiple-category classification

3.2.1 Hypervolume under the manifold

One theoretical extension of AUC is VUS for three-category classification. However, the dilemma of identifying relative order of tests among groups for M-category classifications is more complicated due to the lack of inferential procedures. *Hypervolume under the manifold* (HUM) has been proposed as an extension of VUS for multiple class diagnosis (Scurfield, 1998). In the case of multiple classes (more than three classes), an ROC hypersurface or an ROC manifold could be constructed by using $M-1$ ordered decision thresholds $r_i (i = 1, 2, ...M-1)$ to define a decision rule, similar to those in the three-class case given in the previous chapter.

Suppose the observer makes a decision using $M-1$ criteria, denoted $r_1, r_2, ..., r_{M-1}$, where $r_1 \leq r_2 \leq ... \leq r_{M-1}$. Let the observer discriminate among M classes (denoted $\{l_i : i = 1, ..., M\}$) by M decisions (denoted $\{d_i : i = 1, ...M\}$) as follows:

$$\begin{cases} if \;\; x < r_1, & then \;\; d_1, \\ if \;\; r_{i-1} < x < r_i, \;\; 2 \leq i \leq M-1, & then \;\; d_i, \\ if \;\; r_{M-1} < x & then \;\; d_M. \end{cases}$$

HUM for multiple-category classification can be determined as an extension of VUS

and can be considered as a summary measure of the accuracy. For the continuous case, the hypervolume under the $i_1 i_2 ... i_M$ ROC-hypersuface can be expressed as

$$V_{i_1 i_2 ... i_M} = \int_{-\infty}^{+\infty} \int_{-\infty}^{r_{M-1}} \int_{-\infty}^{r_{M-2}} ... \int_{-\infty}^{r_2} p(d_1|l_{i_1})|J| dr_1 dr_2 ... dr_{M-1},$$

where

$$|J| = \frac{\partial[p(d_2|l_{i_2}), p(d_3|l_{i_3}), ..., p(d_M|l_{i_M})]}{\partial(r_1, r_2, ... r_{M-1})} = \begin{vmatrix} \frac{\partial p(d_2|l_{i_2})}{\partial r_1} & \frac{\partial p(d_2|l_{i_2})}{\partial r_2} & \cdots & \frac{\partial p(d_2|l_{i_2})}{\partial r_{M-1}} \\ \frac{\partial p(d_3|l_{i_3})}{\partial r_1} & \frac{\partial p(d_3|l_{i_3})}{\partial r_2} & \cdots & \frac{\partial p(d_3|l_{i_3})}{\partial r_{M-1}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial p(d_M|l_{i_M})}{\partial r_1} & \frac{\partial p(d_M|l_{i_M})}{\partial r_2} & \cdots & \frac{\partial p(d_M|l_{i_M})}{\partial r_{M-1}} \end{vmatrix},$$

where $(i_1, i_2, ..., i_M)$ is a permutation of $(1, 2, ...M)$ and

$$p(d_1|l_{i_1}) = \int_{-\infty}^{r_1} f(x|l_{i_1}) dx,$$

$$p(d_j|l_{i_j}) = \int_{r_{j-1}}^{r_j} f(x|l_{i_j}) dx \ , \ 2 \le j \le M-1,$$

$$p(d_M|l_{i_M}) = \int_{r_{M-1}}^{\infty} f(x|l_{i_M}) dx,$$

and where f is the probability density function for the continuous case. Equally, we can use equations $t_i = g_{i-1}(t_1, ..., t_{i-1})$, where $i = 2, ...M$, to denote the probability that a subject from class *i* is correctly classified. Then HUM can be expressed in another form as

$$HUM = \int_0^1 \int_0^{g_1(t_1)} \cdots \int_0^{g_{M-2}(t_1,...,t_{M-2})} g_{M-1}(t_1, ..., t_{M-1}) dt_{M-1} \cdots dt_2 dt_1.$$

As an extension of VUS, HUM is equivalent to the probability that the M categories are correctly classified which is $P(X_{i_1} > X_{i_2} \cdots > X_{i_M})$. In the M-category classification,

there will be $M!$ possible HUMs under the $M!$ manifolds and the sum of all the HUMs which are probabilities of correct classification will be equal to one.

3.2.2 Bootstrap approach for the variability

A non-parametric estimate based upon the U-statistic is

$$\widehat{HUM} = \frac{1}{n_1 n_2 ... n_M} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} ... \sum_{i_M=1}^{n_M} I\{X_{1i_1} > X_{2i_2} > \cdots X_{Mi_M}\}, \tag{3.9}$$

where $\{i_1, i_2, ...i_M\}$ is a permutation of $\{1, 2, ...M\}$ and I is the indicator function. The estimator of HUM can be computed as an M-sample U-statistic, similar to the non-parametric estimator of VUS, after the order of the M classes are determined. The nonparametric estimation of HUM is related to the calculation of the probability that more than three categories are correctly ordered by the test. Among all the possible $M!$ HUMs, the largest one is a sensible measure of the accuracy of the test. For a general M-category problem, we need to evaluate $M!$ HUM measures to identify the largest HUM. In this thesis, we will focus on the largest HUM among all the possible ones.

The U-statistic approach for the calculation of the variance of the non-parametric estimator of AUC has been proposed. However, as the sample size increases, the advantage of the U-statistic methodology is heavily reduced and the methodology becomes inappropriate. Given the computational burden of the U-statistic approach, particularly as the dimension of the problem increases, bootstrap estimation of the standard error is suggested. The bootstrap methodology is used for inference in this thesis. Nakas and

Yiannoutsos (2004) pointed out that the bootstrap approach for the calculation of the nonparametric estimator of VUS and HUM has been shown to be essentially equivalent to the U-statistic. For each of the bootstrap samples, denote the estimators obtained from the estimation formula by $\{\widehat{HUM}_n : n = 1, 2, ...N\}$ where N is the number of samples. Li and Fine (2008) proposed a method to calculate the bootstrap standard error for $\widehat{HUM}$ which is

$$\widehat{se_N}(HUM) = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\widehat{HUM}_n - \widehat{HUM})^2}. \tag{3.10}$$

A $100(1-\alpha)\%$ confidence interval for HUM is

$$\widehat{HUM} \pm z_{\alpha/2}\widehat{se_N}(HUM), \tag{3.11}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile for the standard normal distribution.

The bootstrap methodology is used for inference in this thesis which could overcome the computational burden when the number of the categories is large. When the number of the classes increases, the calculation of the variance based upon the U-theory will become complicated and difficult to evaluate. However, the bootstrap methodology for calculating the standard error of the nonparametric estimator becomes a viable choice.

3.3 Multivariate normal distribution assumption

For a general M-category classification problem, we need to evaluate $M!$ such HUM measures to identify the largest HUM. To avoid extensive calculations, we suggest simple methods in which we only need to report summary statistics for each category at an order $O(M)$ instead of $O(M!)$ to determine the right order. We propose to sort the multiple categories by using simple summary statistics under the normal distribution assumption.

For the two-category problem, $AUC = P(X_1 > X_2)$ under the binormal distribution assumption could be expressed as $\Phi(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}})$, where $\Phi(\cdot)$ is the normal distribution function. If we assume that the test results are normally distributed for the multiple-category classification, we will have the following results based upon the comparison of means.

Theorem 3.3.1. *Assume that the test result variable for the M categories are $X_1, X_2, ...X_M$ and they are mutually independent. Let the test result for the kth category $X_k \sim N(\mu_k, \sigma_k^2)$ for $k = 1, 2, \cdots, M$. If $\mu_1 > \mu_2 > \cdots > \mu_M$, then the greatest HUM corresponds to the probability $P(X_1 > X_2 > \cdots > X_M)$.*

Proof of Theorem 3.3.1. It is easy to show that the theorem holds for $M = 2$. For simplicity of presentation, we prove for $M = 3$ by induction in this section. We need to show that $P(X_1 > X_2 > X_3) \geq P(X_{i_1} > X_{i_2} > X_{i_3})$ for any other permutations (i_1, i_2, i_3) of $(1, 2, 3)$.

Let $S_1 = X_2 - X_1$ and $S_2 = X_3 - X_2$. The distribution assumptions given in the theorem implies that $(\delta_1 S_1, \delta_2 S_2)$ is bivariate normal with mean $(\delta_1(\mu_2 - \mu_1), \delta_2(\mu_3 - \mu_2))$ and covariance $\begin{pmatrix} \sigma_1^2 + \sigma_2^2 & -\delta_1\delta_2\sigma_2^2 \\ -\delta_1\delta_2\sigma_2^2 & \sigma_2^2 + \sigma_3^2 \end{pmatrix}$, where $\delta_i = \pm 1$. For different δ_i values, the absolute value of the correlation remains the same $|\rho| = \dfrac{\sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(\sigma_2^2 + \sigma_3^2)}}$.

We notice $P(X_1 > X_2 > X_3) = P(X_2 - X_1 < 0, X_3 - X_2 < 0) = P(S_1 < 0, S_2 < 0)$. For the other five VUSs, we have

$$\begin{aligned}
P(X_1 > X_3 > X_2) &\leq P(X_2 - X_1 < 0, -(X_3 - X_2) < 0) = P(S_1 < 0, -S_2 < 0), \\
P(X_2 > X_1 > X_3) &\leq P(-(X_2 - X_1) < 0, X_3 - X_2 < 0) = P(-S_1 < 0, S_2 < 0), \\
P(X_2 > X_3 > X_1) &\leq P(-(X_2 - X_1) < 0, X_3 - X_2 < 0) = P(-S_1 < 0, S_2 < 0), \\
P(X_3 > X_1 > X_2) &\leq P(X_2 - X_1 < 0, -(X_3 - X_2) < 0) = P(S_1 < 0, -S_2 < 0), \\
P(X_3 > X_2 > X_1) &= P(-(X_2 - X_1) < 0, -(X_3 - X_2) < 0) = P(-S_1 < 0, -S_2 < 0).
\end{aligned}$$

All of these five versions are bounded by $P(\delta_1 S_1 < 0, \delta_2 S_2 < 0)$ where at least one $\delta_i = -1$.

Write $P(\delta_1 S_1 < 0, \delta_2 S_2 < 0)$ as $P(T_1 < 0, T_2 < 0) = F(t_1, t_2)$, where F is the distribution function. By the well-known properties of the distribution function of bivariate normal (Tong (1990)), we have

$$\begin{aligned}
F(t_1, t_2) &= P[\sqrt{1-\rho}Z_i \leq -\sqrt{\rho}Z_0 + \frac{t_i - \mu_i}{\sigma_i}, i = 1, 2] \\
&= P[Z_i \leq \frac{-\sqrt{\rho}Z_0 + a_i}{\sqrt{1-\rho}}, i = 1, 2],
\end{aligned}$$

where $|\rho| < 1$, $a_i = (t_i - \mu_i)/\sigma_i (i = 1, 2)$, Z_1, Z_2 and Z_0 are independent $N(0, 1)$ variables, and Z_1 and Z_2 are independent $N(0, 1)$ variables under the condition $Z_0 = z$ for all z. Therefore, by conditioning on $Z_0 = z$, then unconditioning, we have

$$\begin{aligned} F(t_1, t_2) &= P[Z_1 \leq \frac{-\sqrt{|\rho|}Z_0 + a_1}{\sqrt{1-|\rho|}}, Z_2 \leq \frac{\sqrt{|\rho|}Z_0 + a_2}{\sqrt{1-|\rho|}}] \\ &= \int_{-\infty}^{\infty} \Phi(\frac{\sqrt{|\rho|}z + a_1}{\sqrt{1-|\rho|}})\Phi(\frac{-\sqrt{|\rho|}z + a_2}{\sqrt{1-|\rho|}})\phi(z)dz \, . \end{aligned}$$

Therefore, it becomes

$$P(\delta_1 S_1 < 0, \delta_2 S_2 < 0) = \int_{-\infty}^{\infty} \Phi(\frac{\sqrt{|\rho|}z + \delta_1 a_1}{\sqrt{1-|\rho|}})\Phi(\frac{-\sqrt{|\rho|}z + \delta_2 a_2}{\sqrt{1-|\rho|}})\phi(z)dz \, , \qquad (3.12)$$

where at least one $\delta_i = -1$ and

$$a_1 = \frac{\mu_1 - \mu_2}{\text{var}(S_1)}, \; a_2 = \frac{\mu_2 - \mu_3}{\text{var}(S_2)}.$$

By induction, we can see easily that the integrand in (3.12) is maximized when $\delta_1 = \delta_2 = 1$ for any z. This completes the proof. □

The theorem is thus very helpful for us to find out the order of multiple classes quickly. In practice, we usually compute the sample mean $\hat{\mu}$ for each class as a simple descriptive statistic at the first step. Since sample mean is strongly consistent to the population mean, the order from sample mean can be used to prescribe the order of the M classes in the calculation of HUM.

We further notice that the results are not just limited to the symmetrical normal distribution. In fact, if we replace the normal distribution with certain skewed distributions such as log-normal, exponential or extreme-value distributions, the same conclusion can

be drawn. The proof for exponential distributions can be found in Chandra and Owen (1975). Since these location-scale families are not as common as the normal distribution in diagnostic medicine, we do not elaborate more on this. Interested researchers may conduct a thorough examination on other familiar statistical distributions. Moreover for most continuous random variables we can consider suitable transformations to make the transformed data appear close to being normally distributed. Therefore we expect the application of this theorem to be broad in practice.

3.4 Simulation studies

We conducted a simulation study to examine the performance of our proposed methods for sorting the unknown orders of multiple categories. We considered two data generation scenarios. In Case I, we generated X_1, X_2 and X_3 from normal distributions with descending means of 4, 2, and 0; and variances of 1, 1 and 2, respectively; in Case II, we generated X_1 and X_3 from the same normal distributions as in Case I but construct X_2 from a positive aging Weibull distribution with shape parameter $a = 1/2$ and scale parameter $b = 1$. The mean of X_2 is $b \times \Gamma(1 + 1/a) = 2$. All assumptions of Theorem 3.3.1 hold for Case I. In Case II, the distribution assumption is violated while the means of the three classes are preserved in the same order. We conducted 1000 simulations and sample sizes were fixed at 30 for each category.

In each simulation, we estimated the sample means $\hat{\mu}$ from the generated samples.

We also estimated the six VUS values extensively by using the nonparametric approach based upon U-statistic theory with six different permutations of three classes, then compared the true order to the order estimated from the sample means $\hat{\mu}$. The computation results showed that in Case I, using sample means $\hat{\mu}$, we could determine the order of the three classes in all 1000 simulations. In Case II, the sample means correctly interpreted the relationship among three classes and yielded correct VUS values for 91.7% of the simulations.

3.5 Applications

3.5.1 Leukemia classification

We analyzed the data from leukemia patients used in Golub et al. (1999). The data came from a study of gene expression of two types of acute leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Two main subclasses are known, those arising from T-cells and those arising from B-cells. The training set contains 8 ALL T-cell samples, 19 ALL B-cell samples and 11 AML samples. Each sample contains 3916 gene expression values obtained from Affymetrix high-density oligonucleotide microarrays. The dataset is publicly available at

http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

We considered evaluating the accuracy of the biomarkers for their ability to differ-

entiate between the three classes. We computed VUS for all 3916 genes and evaluated the six possible orders of VUSs for each gene and the corresponding bootstrap standard error. Here, we only listed the top 20 genes with the highest VUSs among all the VUSs. We used the 500-resampling bootstrap methodology to calculate the standard error. At the same time we calculated the sample means to determine the correct order of the three classes for each gene.

As the number of categories is only three, we were able to evaluate the six possible orders and directly chose the highest VUS. There were 168 genes with VUS greater than 1/2. From the results of the largest VUS values from the exhaustive investigation, 96.4% were correctly identified for the ordering of classes by using means. This resulting subsample of genes was the most vital genes since it could correctly classify the three types of leukemia without much uncertainty.

The results for the 20 genes with the highest VUS values and their associated probability interpretations are summarized in Table 3.3. The means for three classes were also conveyed. In this example, the relative orders of the three classes were quite variable for different genes. The top 1 gene with the highest VUS has a value of 0.832. This indicates that this gene can completely differentiate three subjects each randomly sampled from one of the three classes more than 80% of the time in a long run of repeated experiments. This gene systematically assigns high values for AML, moderate values for ALL-t and low values for All-b. The gene with the second highest VUS is also able to differentiate over 80% of the time, achieved by a gene that systematically

gives high values for All-t, moderate values for All-b and low values for AML. The relative magnitudes as in the definition of VUS are all precisely characterized by the orders of means.

For purposes of comparison with other accuracy criteria, we also included correct classification rate (CCR) values (Li and Fine (2008)) for the top 10 genes in Table 3.4. The CCR can be calculated by

$$\widehat{CCR} = \frac{Number\ \ of\ \ correct\ \ classification}{Total\ \ number\ \ of\ \ subjects}.$$

There appeares to be a relatively moderate-sized correlation between the CCR and HUM, compared to the low correlation between VUS and CCR in the example in Li and Fine (2008). The gene with the highest VUS value has the best overall CCR of 0.842. This gene classifies those in classes 1 and 3 correctly more than ninety percent of the time, and mislabled only half of those in class 2. Note that the second highest CCR value is 0.815, which is achieved by four genes corresponding to VUS rankings 13, 29, 31 and 37 (not shown).

For model construction, we applied a forward selection procedure with these twenty genes, starting with gene 1 and sequentially adding genes which maximize the VUS based on the joint model. That is, the combination of the first two genes which maximizes the VUS can be considered as a 'new' gene and sequentially add new genes which maximize the VUS. Interestingly, we only need to include the gene with the 5th highest VUS value to obtain 100% CCR and VUS. Note that because gene 1 also has the highest CCR, using CCR as the loss function in the forward selection procedure

would result in the same model. Excluding gene 1 and gene 5 and applying the forward selection procedure based upon VUS to the remaining eighteen genes, we were able to attain the best combination of two genes with those genes having the 2nd and 6th highest VUS values. The VUS and CCR for this model are 0.98 and 0.89, respectively, with both diagnostic accuracy measures slightly lower than the model based upon genes 1 and 5. These results suggest that the optimum VUS derived with only two gene expression levels achieves excellent performance in terms of both VUS and CCR. Because of correlation between VUS and CCR across genes, using CCR-based selection methods would yield similar results when applied to this dataset.

There has been considerable prior work pertaining to classification on this dataset. Golub et al. (1999) used an arbitrary number of 50 genes with self-organizing maps in combination with a weighted voting scheme to obtain comparable performance to that of our model. Furey et al. (2000) and Guyon et al. (2002) applied support vector machine techniques with roughly 10 genes to achieve the same accuracy. Albrechet et al. (2003) employed the method of threshold circuits with 9 genes. Li and Yang (2005) and Albrecht (2007) reduced the number of expression levels to 3 by using rigid regression and stochastic local search, respectively. Our findings appear to represent a nontrivial improvement, as it is not entirely obvious that two predictors could be used to perfectly discriminate three categories.

3.5.2 Proteomic study for liver cancer

Another example is based upon a recent mass spectrometry dataset for the detection of Glycan biomarkers for liver cancer (Ressome et al. (2008)). The investigators included 203 participants from Cairo, Egypt; 73 hepatocellular carcinoma (denoted by HC) cases; 52 patients with chronic liver disease (denoted by QC); and 78 healthy individuals (denoted by NC). The spectra were generated by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass analyzer (Applied Biosystems Inc., Framingham, MA). Each spectrum consisted of approximately 121,000 m/z values with the corresponding intensities in the mass range of 1,500-5,500 Da. A Supplementary dataset can be found at the author's public website

http://microarray.georgetown.edu/ressomlab/index_downloads.html

which contains a total of 484 peaks after extensive preprocessing of the raw data (Ressom et al. (2007)).

As in the previous example, we computed VUSs exhaustively for six versions of probability definitions and identified the largest value to be the correct VUS. We used sample means to decide the order of the three classes for each peak and compared with the true order.

Among all the calculated volumes under the 321-ROC surface, the gene 183 has the largest value with 0.647 which indicates that this gene can completely differentiate three subjects each randomly sampled from one of the three classes, nearly 65% of the

time among a long run of repeated experiments. This gene systematically assigns high values for QT, moderate values for NC and low values for HC. The second highest VUS can differentiate nearly 63% of the time, achieved by gene 209. This gene also exhibits a systematical classification which gives high values for QT, moderate values for NC and low values for HC. It is also observed that all the 20 peaks with the highest volume values precisely classify these three groups as in the same order of the corresponding means of the three groups.

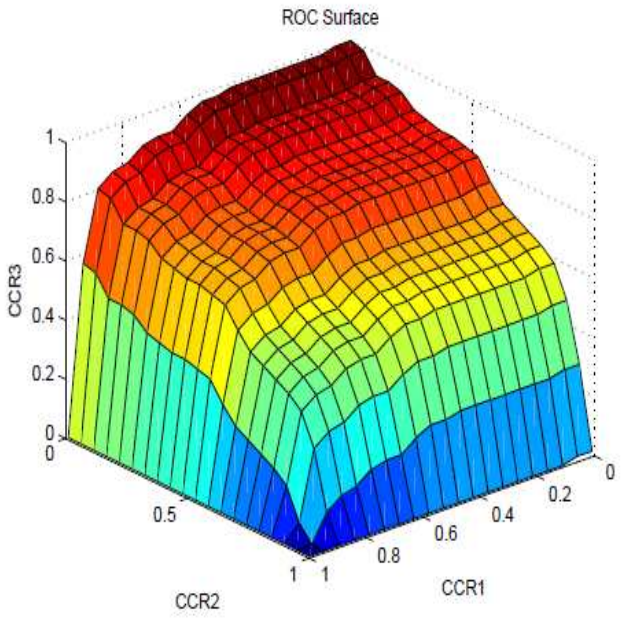
The results for 20 peaks with the highest volume values under the ROC surfaces are shown in Table 3.5. Among all the six volumes for each gene, we noticed the volumes under the 321-ROC surface had more values that were more than 0.5. We also applied the bootstrap methodology as described before to calculate the corresponding standard error with 500 resamples.

Different peaks seemed to maintain the same ordering relationship, except for the 17th peak. For most peaks, healthy subjects (NC) tended to have an intermediate value. Large values tended to lead to chronic liver disease (QT) while low values tended to lead to hepatocellular carcinoma (HC). The 17th peak behaved differently from other peaks where HC patients tended to have the largest peak values relative to the other two groups. Identification of such order information may bring more insights for mass spectrometry studies. In all these 20 cases, the orders in VUS definitions were correctly detected from the orders of means.

In Table 3.6, the corresponding correct rates for the sample means are also reported.

The VUSs of 298 peaks are greater than 0.25 by the sample means with 75.8% correctly identified. The VUSs of 240 are greater than 0.30 by the sample means with 82.2% correctly identified. The VUSs of 110 peaks are greater than 0.4, with 98.2% of them are correctly identified by the sample means. We also noticed that the sample size for each class was not large.

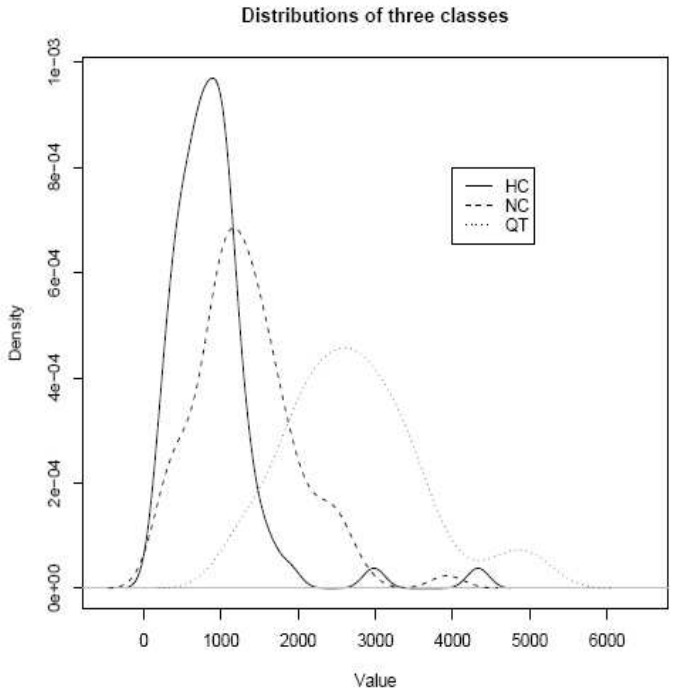
Figure 3.1: ROC surface for the peak with the largest VUS. The three coordinates are the correct classification probabilities for the three classes



The ROC surface for the peak with the largest VUS is plotted in Figure 3.1. One

can choose the appropriate cutoff values r_1 and r_2 for a particular decision to satisfy required correct classification probabilities by locating the corresponding values on this operating surface. The distributions of this peak among three classes are shown in Figure 3.2. The overall shapes of the three empirical density curves are quite close to the normal distribution and justified the assumption in Theorem 3.3.1.

Figure 3.2: The distribution of the peak with the largest VUS among the three groups



3.5.3 Immunohistological data

Another recently studied immunohistological arthritis dataset came from a biomedical company in Germany. All the values in this dataset were positive staining cells which were vessels in the case of fWV. They were measured per 400x microscopic field (0.159mmz). Our interest lies in seven components which are T cells (denoted CD3), B cells (denoted CD20), Plasma cells (denoted CD 38), Mph's subintimal (denoted CD68), Ki67, Total Mononuclear influence cells (denoted TMI), and vWF. The seven components were scored based upon the grading of the immunohistological severity of arthritis. The primary classification outcome involves seven different categories which are 'Normal', 'Orth.A', 'OA', 'Early arthritis', 'RA', 'SeA (disease)', and 'SeA-TKA', respectively. We use $X_1, X_2, X_3, X_4, X_5, X_6$, and X_7 to denote these seven categories. We estimated all the 7! possible HUMs for each component. Among all the possible HUMs, the largest HUM for each component and the corresponding order are reported in Table 3.7.

The largest HUM among all the components comes from the total mononuclear influence cell which has a value of 0.0444. The corresponding probability of the correct ordering is $P(X_1 < X_2 < X_3 < X_7 < X_4 < X_6 < X_5)$. The smallest HUM among all the components is from B cells with the value 0.0034, which has the order $P(X_2 < X_1 < X_7 < X_3 < X_6 < X_4 < X_5)$. To view the correct classification by their means, all the means for the seven categories are also listed in Table 3.8.

The largest HUM value of 0.0444 is from the total mononuclear influence cell with

the probability $P(X_1 < X_2 < X_3 < X_7 < X_4 < X_6 < X_5)$. The corresponding means of the seven different categories are 10.58, 30.52, 42.34, 231.54, 309.50, 272.57, and 158.46, respectively. That is, $mean1 < mean2 < mean3 < mean7 < mean4 < mean6 < mean5$. This indicates that the classification in the total mononuclear influence cell has the correct order based upon the comparison of means. The second largest HUM which comes from the Ki67 component with the value 0.028 has the probability definition of $P(X_2 < X_1 < X_3 < X_5 < X_4 < X_7 < X_6)$. But the corresponding means reveal a different order of the means. The reason for this may be because of the existence of some outliers or extreme observations which may influence the estimation of the means and the distribution of the observations are not normally distributed. Weighted means may be suggested for this case for the order correction instead of only considering the sample means.

Table 3.3: Top 20 gene expression levels ranked by VUS value for Leukemia data. μ_i is the mean for the ith class(i=1,2,3). Classes 1,2 and 3 are ALL-b, ALL-t, and AML respectively.

Rank	VUS	Definition	μ_1	μ_2	μ_3	*s.e*
1	0.832	$P(X_1 < X_2 < X_3)$	-30.684	560.375	7423.545	0.0654
2	0.822	$P(X_3 < X_1 < X_2)$	857.790	2208.50	485.727	0.0732
3	0.788	$P(X_3 < X_1 < X_2)$	666.737	2283.875	129.909	0.0632
4	0.782	$P(X_2 < X_3 < X_1)$	2403.789	145.875	524.0	0.0723
5	0.770	$P(X_3 < X_1 < X_2)$	205.684	3373.125	67.091	0.0831
6	0.763	$P(X_3 < X_2 < X_1)$	1573.632	757.125	310.273	0.0735
7	0.735	$P(X_3 < X_2 < X_1)$	4322.526	2772.625	702.364	0.0687
8	0.724	$P(X_1 < X_2 < X_3)$	479.211	712.0	1439.636	0.0784
9	0.718	$P(X_1 < X_3 < X_2)$	-88.105	1030.875	63.545	0.0823
10	0.708	$P(X_1 < X_3 < X_2)$	8.579	718.125	78.0	0.0764
11	0.705	$P(X_3 < X_2 < X_1)$	4988.579	2371.0	1365.273	0.0784
12	0.704	$P(X_1 < X_3 < X_2)$	380.684	1040.375	503.0	0.0803
13	0.698	$P(X_2 < X_1 < X_3)$	747.895	138.75	1273.091	0.0769
14	0.697	$P(X_1 < X_3 < X_2)$	108.579	806.5	361.636	0.0835
15	0.687	$P(X_1 < X_2 < X_3)$	108.895	229.50	2520.364	0.0753
16	0.681	$P(X_2 < X_3 < X_1)$	2477.789	662.875	1367.909	0.0768
17	0.680	$P(X_2 < X_3 < X_1)$	790.684	183.125	487.455	0.0689
18	0.680	$P(X_3 < X_2 < X_1)$	7974.789	2598.0	801.182	0.0843
19	0.676	$P(X_1 < X_3 < X_2)$	567.947	2695.50	801.181	0.0785
20	0.670	$P(X_3 < X_1 < X_2)$	3437.053	4726.875	1818.273	0.0798

Table 3.4: Top 20 gene expression levels ranked by VUS value for Leukemia data. CCR is the corresponding overall correct classification rate. CCR[i] is the correct classification rate for the ith class(i=1,2,3). Classes 1,2 and 3 are ALL-b, ALL-t, and AML respectively.

Rank	VUS	CCR	CCR[1]	CCR[2]	CCR[3]
1	0.832	0.842	0.947	0.500	0.909
2	0.822	0.657	0.894	0.000	0.727
3	0.788	0.789	0.842	0.750	0.727
4	0.782	0.736	0.684	0.875	0.727
5	0.770	0.736	0.789	1.000	0.454
6	0.763	0.710	0.736	0.875	0.545
7	0.735	0.736	0.842	0.750	0.545
8	0.724	0.789	0.894	0.250	1.000
9	0.718	0.763	0.947	0.250	0.818
10	0.708	0.789	0.842	0.500	0.909

Table 3.5: Top 20 peaks ranked by VUS value for liver cancer data. μ_i is the mean for the ith class(i=1,2,3). Classes 1, 2, and 3 are HC, NC, and QT, respectively.

Rank	VUS	Definition	μ_1	μ_2	μ_3	*s.e*
1	0.647	$P(X_1 < X_2 < X_3)$	896.611	1326.071	2732.444	0.0856
2	0.632	$P(X_1 < X_2 < X_3)$	651.121	985.067	1372.388	0.0886
3	0.623	$P(X_1 < X_2 < X_3)$	1452.321	2010.886	4829.766	0.0902
4	0.584	$P(X_1 < X_2 < X_3)$	124.784	286.132	412.497	0.0846
5	0.563	$P(X_1 < X_2 < X_3)$	481.267	697.342	988.530	0.0856
6	0.558	$P(X_1 < X_2 < X_3)$	544.353	748.769	1122.159	0.0935
7	0.533	$P(X_1 < X_2 < X_3)$	314.048	401.839	607.952	0.0852
8	0.529	$P(X_1 < X_2 < X_3)$	150.320	366.533	553.001	0.0875
9	0.524	$P(X_1 < X_2 < X_3)$	10552.81	21490.65	26878.11	0.0904
10	0.513	$P(X_1 < X_2 < X_3)$	413.769	526.830	772.249	0.0875
11	0.509	$P(X_1 < X_2 < X_3)$	1014.861	1245.426	1928.783	0.0895
12	0.504	$P(X_1 < X_2 < X_3)$	785.593	854.669	1577.023	0.0934
13	0.503	$P(X_1 < X_2 < X_3)$	229.566	285.739	428.804	0.0857
14	0.502	$P(X_1 < X_2 < X_3)$	171.408	226.698	322.948	0.0894
15	0.501	$P(X_1 < X_2 < X_3)$	170.734	281.202	364.334	0.0846
16	0.499	$P(X_1 < X_2 < X_3)$	85.506	126.392	189.421	0.0923
17	0.498	$P(X_2 < X_3 < X_1)$	567.211	198.593	227.376	0.0863
18	0.496	$P(X_1 < X_2 < X_3)$	114.326	170.651	363.178	0.0895
19	0.495	$P(X_1 < X_2 < X_3)$	660.858	949.397	1331.265	0.0935
20	0.491	$P(X_1 < X_2 < X_3)$	333.676	425.174	741.058	0.0964

Table 3.6: Correct identification by the sample means

VUS	Correct Number	Non-correct Number	Total Number	Correct Rate
≥ 0.25	298	95	393	75.8%
≥ 0.26	288	87	375	76.8%
≥ 0.27	269	76	345	77.9%
≥ 0.28	261	70	331	78.9%
≥ 0.29	251	63	314	79.9%
≥ 0.30	240	52	292	82.2%
≥ 0.40	110	2	112	98.2%

Table 3.7: HUMs for immunohistological data

Marker	HUM	Definition
CD3(T Cells)	0.0209	$P(X_2 < X_1 < X_3 < X_7 < X_4 < X_5 < X_6)$
CD20(B Cells)	0.0034	$P(X_2 < X_1 < X_7 < X_3 < X_6 < X_4 < X_5)$
CD38(Plasma Cells)	0.0191	$P(X_1 < X_2 < X_3 < X_7 < X_6 < X_4 < X_5)$
CD68(Mph's Subintimal)	0.0267	$P(X_1 < X_2 < X_3 < X_5 < X_6 < X_7 < X_4)$
Ki67	0.028	$P(X_2 < X_1 < X_3 < X_5 < X_4 < X_7 < X_6)$
TMI (Total Mononuclear influence cell)	0.0444	$P(X_1 < X_2 < X_3 < X_7 < X_4 < X_6 < X_5)$
vWF	0.0072	$P(X_1 < X_2 < X_3 < X_7 < X_5 < X_4 < X_6)$

Table 3.8: Means of the seven categories in immunohistological data

Marker	Mean1	Mean2	Mean3	Mean4	Mean5	Mean6	Mean7
CD3(T Cells)	3.04	8.13	11.10	47.46	91.20	106.47	27.17
CD20(B Cells)	0.30	5.14	3.55	20.67	39.02	28.83	7.59
CD38(Plasma Cells)	0.04	2.45	4.02	51.43	93.06	41.14	11.32
CD68(Mph's Subintimal)	7.2	14.80	23.67	111.98	86.22	96.13	112.39
Ki67	1.23	1.99	4.89	21.04	31.82	64.21	25.50
Total Mononuclear infl cell	10.58	30.52	42.34	231.54	309.50	272.57	158.46
vWF	9.36	13.65	13.43	17.60	18.92	28.82	18.29

Chapter 4

Combining Multiple Markers for Multiple-category Classification

4.1 Introduction

ROC analysis has been the most recommended and effective way to evaluate the accuracy performance of diagnostic tests. Moreover, statistical approaches have been developed for assessing the accuracy of classifications. In practice, multiple factors will influence the accuracy performance and various sources of information are available to assist in predicting medical classification problems. For example, a single biomarker will not be sufficient to assess an optimal result for prognosis or early detection for many diseases. However, multiple biomarkers and various signs and distinctive symptoms of the disease can help detect the disease. A combination of these multiple biomarkers

can potentially detect the disease to a significant extent. Thus, combining multiple biomarkers and factors is needed in order to predict an adequate outcome. So it follows that great interest exists in developing methods for combining biomarkers, especially in medical research.

Recently, methods have been developed for combining multiple biomarkers. Su and Liu (1993) and Pepe and Thompson (2000) considered linear combinations to optimize measures of diagnostic accuracy. Optimal prognostic scores can be determined through binary regressions (Pepe and McIntosh (2003)). Pepe an McIntosh (2003) proposed screening rules based upon logical combinations of biomarker measurements. For binary classification, Pepe and Thompson (2000) developed a method based upon maximizing the AUC to combine biomarkers in genetic studies. Their method was essentially adapted from the *maximum rank correlation* (MRC) estimation which was widely practiced in econometrics. Li and Fine (2008) considered multinomial logistic regression to address multiple-category outcomes. However, it is not clear if their method yields the best combination to maximize VUS or HUM. In this thesis we target maximizing the VUS directly. We will explore statistical methods that combine multiple tests for multiple-category classification to optimize the accuracy of the combined biomarkers under the criteria of ROC measures.

Early discussion about the MRC estimation can be found in Han (1987) and Sherman (1993) where the authors studied the limiting distribution of the MRC estimator. The implementation of the MRC estimation has been applied recently. In the recent

decade the maximum rank correlation (MRC) estimator has been applied in the classification literature for two-class problems due to its close connection with AUC. Wang, H. (2006) further suggested an iterative marginal algorithm which remarkably improved the computation speed. However, none of the previous authors considered the situation in which the number of decision categories exceeds two. We thus aim at developing appropriate statistical procedures by extending the MRC estimators for high-dimensional cases. Necessary asymptotic theories are provided to facilitate the ensuing inference.

4.2 Methods

4.2.1 Methods: extending MRC estimation

Generally, it is natural to expect a monotonic relationship between a response variable and a linear index. To explore the relation between them beyond the linear approximation, the continuous single index model can also be considered, which is a well-known approach in multidimensional cases. This idea of thresholding on a single continuous index for multiple-category classification includes many existing models, such as the smooth transition threshold autoregressive (STAR) model of Chan and Tong and the functional-coefficient autore- gressive (FAR) model of Chen and Tsay. To avoid the dimensionality in multivariate estimations and the specification of the exact nature of the monotonicity, Han (1987) firstly proposed the semiparametric monotonic linear index model.

Let $(Y, \boldsymbol{X})$ be an observation from a distribution P on a set $S \subseteq R \bigotimes R^d$, where Y is a response variable and $\boldsymbol{X}$ is a d-vector of regressor variables. The monotonic linear index model can be proposed as

$$Y = D \circ F(\boldsymbol{X}^T\boldsymbol{\beta}_0, \boldsymbol{\varepsilon}), \tag{4.1}$$

where $\boldsymbol{X}^T\boldsymbol{\beta}_0$ is a linear index with $\boldsymbol{\beta}_0 \in \boldsymbol{B} \subset R^d$, an unknown d-dimensional vector, $\boldsymbol{\varepsilon}$ is a random disturbance, F is a strictly increasing function in each of its arguments, and D is a nonconstant and increasing function. The model is semiparametric in that no parametric assumptions are made about the distribution of $\boldsymbol{\varepsilon}$ or the functional form of $D \circ F$. Previously the sample space for Y is only $\{1, 0\}$. In this thesis, we consider that Y can take values from $\{1, 0, -1\}$.

Suppose we obtain a sample $\{(Y_{i_j}, \boldsymbol{X}_{i_j}); i_j = 1, \cdots, n_j, j = 1, 2, 3\}$, where j indexes the three classes and i_j indexes the observations in the jth class. The MRC estimator of the coefficient parameter $\boldsymbol{\beta}_0$ is obtained from

$$argmax_{\boldsymbol{\beta}\in B} \sum_{i_1,i_2,i_3} I\{Y_{i_1} > Y_{i_2} > Y_{i_3}, \boldsymbol{X}_{i_1}^T\boldsymbol{\beta} > \boldsymbol{X}_{i_2}^T\boldsymbol{\beta} > X_{i_3}^T\boldsymbol{\beta}\}, \tag{4.2}$$

where $I\{\cdot\}$ stands for the indicator function. It has been shown that up to a constant unrelated to $\boldsymbol{\beta}$, the objective function in (4.2) is proportional to VUS defined in Li and Fine (2008),

$$\widehat{VUS} = n_1^{-1}n_2^{-1}n_3^{-1} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \sum_{i_3=1}^{n_3} I\{\boldsymbol{X}_{i_1}^T\boldsymbol{\beta} > \boldsymbol{X}_{i_2}^T\boldsymbol{\beta} > X_{i_3}^T\boldsymbol{\beta}\}.$$

We note that VUS of a diagnostic test can be interpreted as the probability that the marker can simultaneously classify three categories correctly. Therefore using the esti-

mator $\boldsymbol{\beta}_0$ obtained by maximizing the VUS to combine the markers $\boldsymbol{X}$ implies that the resulting accuracy would be optimal for the three-category classification.

Following Han (1987), we consider the MRC estimator in a more general model framework. Let M be a function on R^2 and monotone for either argument when the other argument is fixed. For the real numbers $a_1, ..., a_n$, let $R_n(a_i, a_k)$ denote the number of a_j's between a_i and a_k, i.e.

$$R_n(a_i, a_k) = \sum_j I\{a_i > a_j > a_k\}.$$

We propose to estimate the true parameter $\boldsymbol{\beta}_0$ in (4.1) with

$$\boldsymbol{\beta}_n = argmax_{\boldsymbol{\beta}\in\boldsymbol{B}} \sum_i \sum_k M(Y_i, Y_k) R_n(\boldsymbol{X}_i^T\boldsymbol{\beta}, \boldsymbol{X}_k^T\boldsymbol{\beta}), \tag{4.3}$$

for an appropriate subset $\boldsymbol{B}$ of R^d.

We now show that the estimator from (4.2) is a special case for the general MRC estimator from (4.3). For $l = -1, 0, 1$, define

$$R_n^{(l)}(\boldsymbol{X}_i^T\boldsymbol{\beta}, \boldsymbol{X}_k^T\boldsymbol{\beta}) = \sum_j I\{Y_j = l)\} I\{\boldsymbol{X}_i^T\boldsymbol{\beta} > \boldsymbol{X}_j^T\boldsymbol{\beta} > \boldsymbol{X}_k^T\boldsymbol{\beta}\}.$$

The maximand in (4.2) equals

$$\begin{aligned}
&\sum_i \sum_j \sum_k I\{Y_i > Y_j > Y_k, \boldsymbol{X}_i^T\boldsymbol{\beta} > \boldsymbol{X}_j^T\boldsymbol{\beta} > \boldsymbol{X}_k^T\boldsymbol{\beta}\} \\
&= \sum_i \sum_j \sum_k I\{Y_i = 1\} I\{Y_j = 0\} I\{Y_k = -1\} I\{\boldsymbol{X}_i^T\boldsymbol{\beta} > \boldsymbol{X}_j^T\boldsymbol{\beta} > \boldsymbol{X}_k^T\boldsymbol{\beta}\} \\
&= \sum_i \sum_k I\{Y_i = 1\} I\{Y_k = -1\} R_n^{(0)}(\boldsymbol{X}_i^T\boldsymbol{\beta}, \boldsymbol{X}_j^T\boldsymbol{\beta}),
\end{aligned}$$

which is the maximand in (4.3) with a special choice of M.

In the following we proceed to give the asymptotic results for the more general estimators in (4.3).

We establish the consistency of $\boldsymbol{\beta}_n$ first. Denote

$$G_n(\boldsymbol{\beta}) = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} M(Y_i, Y_k) I\{\boldsymbol{X}_i^T \boldsymbol{\beta} > \boldsymbol{X}_j^T \boldsymbol{\beta} > X_k^T \boldsymbol{\beta}\}. \tag{4.4}$$

One may notice that $\{G_n(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \boldsymbol{B}\}$ is a U-process of order 3.

Define $G(\boldsymbol{\beta}) = E[M(Y_1, Y_3) I\{\boldsymbol{X}_1^T \boldsymbol{\beta} > \boldsymbol{X}_2^T \boldsymbol{\beta} > \boldsymbol{X}_3^T \boldsymbol{\beta}\}]$. We note that $G(\boldsymbol{\beta})$ is the expected value of $G_n(\boldsymbol{\beta})$.

We also define $H(\boldsymbol{X}_i^T \boldsymbol{\beta}) = E[M(Y_1, Y_3) | \boldsymbol{X}_i^T \boldsymbol{\beta}]$ for $i = 1, 2, 3$.

The following sets of technical conditions are needed:

A1. $H(t)$ is a nonconstant monotone real function.

A2. The support of $\boldsymbol{X}_i$ is not contained in a proper linear subspace of R^d, $i = 1, 2, 3$.

A3. The d-th component of $\boldsymbol{X}_i$ has an everywhere positive Lebesgue density, conditional on the other components, $i = 1, 2, 3$.

A4. $\boldsymbol{B}$ is a compact subset of $\{\boldsymbol{\beta} \in R^d : \beta_d = 1\}$.

A5. $E[M(Y_1, Y_3)]^2 < \infty$.

Theorem 4.2.1 (Consistency)**.** *Assume conditions A1 to A5 hold. Then we have*

$$|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0| = o_p(1).$$

Proof of Theorem 4.2.1. Essentially, to establish the consistency of $\boldsymbol{\beta}_n$, it is sufficient to show the following:

(i) $G(\boldsymbol{\beta})$ is uniquely maximized at β_0.

(ii) $sup_{\boldsymbol{B}}|G_n(\boldsymbol{\beta}) - G(\boldsymbol{\beta})| = o_p(1)$.

(iii)$G(\boldsymbol{\beta})$ is continuous.

By symmetry, we may write

$$G(\boldsymbol{\beta}) = \frac{1}{6}E[H(\boldsymbol{X}_1^T\boldsymbol{\beta})I\{\boldsymbol{X}_1^T\boldsymbol{\beta} > \boldsymbol{X}_2^T\boldsymbol{\beta} > \boldsymbol{X}_3^T\boldsymbol{\beta}\} + H(\boldsymbol{X}_1^T\boldsymbol{\beta})I\{\boldsymbol{X}_1^T\boldsymbol{\beta} > \boldsymbol{X}_3^T\boldsymbol{\beta} > \boldsymbol{X}_2^T\boldsymbol{\beta}\}$$

$$+H(\boldsymbol{X}_2^T\boldsymbol{\beta})I\{\boldsymbol{X}_2^T\boldsymbol{\beta} > \boldsymbol{X}_1^T\boldsymbol{\beta} > \boldsymbol{X}_3^T\boldsymbol{\beta}\} + H(\boldsymbol{X}_2^T\boldsymbol{\beta})I\{\boldsymbol{X}_2^T\boldsymbol{\beta} > \boldsymbol{X}_3^T\boldsymbol{\beta} > \boldsymbol{X}_1^T\boldsymbol{\beta}\}$$

$$+H(\boldsymbol{X}_3^T\boldsymbol{\beta})I\{\boldsymbol{X}_3^T\boldsymbol{\beta} > \boldsymbol{X}_1^T\boldsymbol{\beta} > \boldsymbol{X}_3^T\boldsymbol{\beta}\} + H(\boldsymbol{X}_3^T\boldsymbol{\beta})I\{\boldsymbol{X}_3^T\boldsymbol{\beta} > \boldsymbol{X}_2^T\boldsymbol{\beta} > \boldsymbol{X}_1^T\boldsymbol{\beta}\}]. \quad (4.5)$$

If $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, then conditions A1 and A3 ensure that the indicators in (4.5) pick out the largest of $H(\boldsymbol{X}_1^T\boldsymbol{\beta}_0)$, $H(\boldsymbol{X}_2^T\boldsymbol{\beta}_0)$, and $H(\boldsymbol{X}_3^T\boldsymbol{\beta}_0)$ with probability one. Consequently,

$$G(\boldsymbol{\beta}_0) = \frac{1}{3}Emax(H(\boldsymbol{X}_1^T\boldsymbol{\beta}_0), H(\boldsymbol{X}_2^T\boldsymbol{\beta}_0), H(\boldsymbol{X}_3^T\boldsymbol{\beta}_0)). \quad (4.6)$$

Deduce that $G(\boldsymbol{\beta})$ is maximized at $\boldsymbol{\beta}_0$.

We now show that $\boldsymbol{\beta}_0$ is the unique maximizer.

Suppose that for some $\boldsymbol{\beta}$ in $\boldsymbol{B}$,

$$G(\boldsymbol{\beta}) = \frac{1}{3}Emax(H(\boldsymbol{X}_1^T\boldsymbol{\beta}_0), H(\boldsymbol{X}_2^T\boldsymbol{\beta}_0), H(\boldsymbol{X}_3^T\boldsymbol{\beta}_0)). \quad (4.7)$$

Deduce from (4.6) and (4.7) that

$$H(\boldsymbol{X}_1^T\boldsymbol{\beta}_0) \geq H(\boldsymbol{X}_2^T\boldsymbol{\beta}_0) \text{ and } H(\boldsymbol{X}_1^T\boldsymbol{\beta}_0) \geq H(\boldsymbol{X}_3^T\boldsymbol{\beta}_0) \quad \text{when} \quad \boldsymbol{X}_1^T\boldsymbol{\beta} \geq \boldsymbol{X}_2^T\boldsymbol{\beta} \geq \boldsymbol{X}_3^T\boldsymbol{\beta}. \quad (4.8)$$

Let S_χ denote the support of $\chi = (X_1, ...X_{d-1})$ and write CH_χ for the convex surface of S_χ. That is, CH_χ is the smallest convex set containing S_χ. Assumption A2 implies that

CH_χ is a $(d-1)$-dimensional subset of R^{d-1} and so has a nonempty interior. Select a point $\boldsymbol{\mu}$ from this interior and define $I_{\boldsymbol{\mu}} = \{(\boldsymbol{\mu}, t) : t \in R\}$.

Assumption A1 guarantees the existence of two points t_0 and t_1 in the support of $\boldsymbol{X}^T\boldsymbol{\beta}$ for which $H(t_0) < H(t) < H(t_1)$ for $t_0 < t < t_1$.

Choose $\boldsymbol{\tau}_0, \boldsymbol{\tau}_1$ in $I_{\boldsymbol{\mu}}$ for which $\boldsymbol{\tau}_0^T\boldsymbol{\beta}_0 = t_0, \boldsymbol{\tau}_1^T\boldsymbol{\beta}_0 = t_1$. Those points can always be found since A3 and A4 together imply that $\{(\boldsymbol{\tau}_0^T\boldsymbol{\beta}_0, \boldsymbol{\tau}_1^T\boldsymbol{\beta}_0) : (\boldsymbol{\tau}_0, \boldsymbol{\tau}_1) \in I_{\boldsymbol{\mu}}\} \equiv R^2$.

Define the open wedges

$$W_1(\boldsymbol{\beta}) = \{\boldsymbol{x} : \boldsymbol{x}^T\boldsymbol{\beta}_0 < \boldsymbol{\tau}_0^T\boldsymbol{\beta}_0, \boldsymbol{x}^T\boldsymbol{\beta} > \boldsymbol{\tau}_0^T\boldsymbol{\beta}\},$$

$$W_2(\boldsymbol{\beta}) = \{\boldsymbol{x} : \boldsymbol{\tau}_0^T\boldsymbol{\beta}_0 < \boldsymbol{x}^T\boldsymbol{\beta}_0 < \boldsymbol{\tau}_1^T\boldsymbol{\beta}_0, \boldsymbol{\tau}_0^T\boldsymbol{\beta} > \boldsymbol{x}^T\boldsymbol{\beta} > \boldsymbol{\tau}_1^T\boldsymbol{\beta}\},$$

$$W_3(\boldsymbol{\beta}) = \{\boldsymbol{x} : \boldsymbol{x}^T\boldsymbol{\beta}_0 > \boldsymbol{\tau}_1^T\boldsymbol{\beta}_0, \boldsymbol{x}^T\boldsymbol{\beta} < \boldsymbol{\tau}_1^T\boldsymbol{\beta}\}\ .$$

We can replace $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ with their respective unit vector without changing $W_1(\boldsymbol{\beta})$,$W_2(\boldsymbol{\beta})$ and $W_3(\boldsymbol{\beta})$. Thus, for each $\boldsymbol{x}$ in R^d and each $\boldsymbol{\beta}$ in $\boldsymbol{B}$, we may view $\boldsymbol{x}^T\boldsymbol{\beta}$ as the orthogonal projection of $\boldsymbol{x}$ onto the space spanned by $\boldsymbol{\beta}$.

If $\boldsymbol{X}_1 \in W_1(\boldsymbol{\beta})$, $\boldsymbol{X}_2 \in W_2(\boldsymbol{\beta})$ and $\boldsymbol{X}_3 \in W_3(\boldsymbol{\beta})$, then

$$H(\boldsymbol{X}_1^T\boldsymbol{\beta}_0) < H(\boldsymbol{X}_2^T\boldsymbol{\beta}_0) < H(\boldsymbol{X}_3^T\boldsymbol{\beta}_0) \quad \textit{while} \quad \boldsymbol{X}_1^T\boldsymbol{\beta} > \boldsymbol{X}_2^T\boldsymbol{\beta} > \boldsymbol{X}_3^T\boldsymbol{\beta}\ .$$

Then in order for (4.8) to hold, we must have

$$P\{\boldsymbol{X}_1 \in W_1(\boldsymbol{\beta})\}P\{\boldsymbol{X}_2 \in W_2(\boldsymbol{\beta})\}P\{\boldsymbol{X}_3 \in W_3(\boldsymbol{\beta})\} = 0. \tag{4.9}$$

Now we show that (4.9) only holds for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

For each $\boldsymbol{\beta}$ in $\boldsymbol{B}$, define

$$H_{\boldsymbol{\beta}} = \{\boldsymbol{x} : \boldsymbol{\tau}_0^T\boldsymbol{\beta} = \boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{\tau}_1^T\boldsymbol{\beta}\},$$

$$L_{\boldsymbol{\beta}} = H_{\boldsymbol{\beta}} \cap H_{\boldsymbol{\beta}_0}.$$

Consider the projections:

$$P_0(\boldsymbol{\beta}) = \{\boldsymbol{x} \in CH_\chi : (\boldsymbol{x}, t) \in L_\beta \quad for \quad some \quad t \in R\},$$

and for $j = 1, 2, 3$

$$P_j(\boldsymbol{\beta}) = \{\boldsymbol{x} \in CH_\chi : (\boldsymbol{x}, t) \in W_j(\boldsymbol{\beta}) \quad for \quad some \quad t \in R\}.$$

That is, $P_0(\boldsymbol{\beta})$ projects $L_{\boldsymbol{\beta}}$ into CH_χ and $P_j(\boldsymbol{\beta})$ projects $W_j(\boldsymbol{\beta})$ into CH_χ. And $\{P_j(\boldsymbol{\beta}), j = 0, 1, 2, 3\}$ partitions CH_χ.

Since both $H_{\boldsymbol{\beta}}$ and $H_{\boldsymbol{\beta}_0}$ contain $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}_1$, $L_{\boldsymbol{\beta}}$ must contain $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}_1$. Since $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}_1$ are elements of $I_{\boldsymbol{\mu}}$, $P_0(\boldsymbol{\beta})$ must contain $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Since $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are interior points of CH_χ, $P_0(\boldsymbol{\beta})$ cannot contain a face of CH_χ. But each $P_j(\boldsymbol{\beta})$ must contain at least one point of S_χ, implying

$$\int_{P_j(\boldsymbol{\beta}) \cap S_\chi} G_\chi(d\boldsymbol{x}) > 0,$$

where $G_\chi(\cdot)$ denotes the distribution of χ.

For each $\boldsymbol{x}$ in S_χ, write $l_{\boldsymbol{x}}$ for the line through $\boldsymbol{x}$ parallel to the d-th coordinate axis. If $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, then there must be a nonzero angle between H_β and $H_{\boldsymbol{\beta}_0}$. So at least one of $H_{\boldsymbol{\beta}}$ and $H_{\boldsymbol{\beta}_0}$ must intersect at $l_{\boldsymbol{x}}$. Write $t_{\boldsymbol{\beta}}(\boldsymbol{x})$ for the d-th component of $H_{\boldsymbol{\beta}} \cap l_{\boldsymbol{x}}$ and $t_{\boldsymbol{\beta}_0}(\boldsymbol{x})$ for the d-th component of $H_{\boldsymbol{\beta}_0} \cap l_{\boldsymbol{x}}$. If $H_{\boldsymbol{\beta}} \cap l_{\boldsymbol{x}}$ is null, define $t_{\boldsymbol{\beta}}(\boldsymbol{x}) = \infty$ (or $-\infty$). If $H_{\boldsymbol{\beta}_0} \cap l_{\boldsymbol{x}}$ is null, define $t_{\boldsymbol{\beta}_0}(\boldsymbol{x}) = \infty$ (or $-\infty$). Then

$$P(\boldsymbol{X} \in W_j(\boldsymbol{\beta})) = \int_{P_j(\boldsymbol{\beta}) \cap S_\chi} [\int_{min(t_{\boldsymbol{\beta}_0}(\boldsymbol{x}),\ t_{\boldsymbol{\beta}}(\boldsymbol{x}))}^{max(t_{\boldsymbol{\beta}_0}(\boldsymbol{x}),\ t_{\boldsymbol{\beta}}(\boldsymbol{x}))} f(t|\boldsymbol{x})dt] G_\chi(d\boldsymbol{x}),$$

where $f(\cdot|\boldsymbol{x})$ denotes the conditional density of X_d given $\chi = \boldsymbol{x}$. $t_{\boldsymbol{\beta}}(\boldsymbol{x}) \neq t_{\boldsymbol{\beta}_0}(\boldsymbol{x})$ for each $\boldsymbol{x}$ in S_χ because $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. So $P(\boldsymbol{X} \in W_j(\boldsymbol{\beta}))$ can not be 0. That is, $P(\boldsymbol{X} \in W_j(\boldsymbol{\beta})) > 0$, contradicting (4.9). This establishes (i).

For each $\boldsymbol{\beta}$ in $\boldsymbol{B}$ and (z_1, z_2, z_3) in $S \bigotimes S \bigotimes S$, define

$$f(z_1, z_2, z_3, \boldsymbol{\beta}) = M(y_1, y_2) I\{\boldsymbol{x}_1^T\boldsymbol{\beta} > \boldsymbol{x}_2^T\boldsymbol{\beta} > \boldsymbol{x}_3^T\boldsymbol{\beta}\} - G(\boldsymbol{\beta}) .$$

Then

$$G_n(\boldsymbol{\beta}) - G(\boldsymbol{\beta}) = U_n f(.,.,.,\boldsymbol{\beta}),$$

where U_n denotes the random measure putting mass $1/[n(n-1)(n-2)]$ on each pair $(Z_i, Z_j, Z_k), i \neq j \neq k$. That is, $\{U_n f(.,.,.,\boldsymbol{\beta})\}$ is a zero-mean U-process of order 3. From the result of Sherman (1994),

$$sup_{\boldsymbol{B}}|U_n^3 f(.,.,.,\boldsymbol{\beta})| \leq \sum_{i=1}^{3} sup_{\boldsymbol{B}}|U_n^i f(.,.,.,\boldsymbol{\beta})| ,$$

and

$$sup_{\boldsymbol{B}}|n^{3/2} U_n^3 f(.,.,.,\boldsymbol{\beta})| = O_p(1) .$$

Thus,

$$sup_{\boldsymbol{B}}|G_n(\boldsymbol{\beta}) - G(\boldsymbol{\beta})| = o_p(1) .$$

This is enough to establish (ii).

Finally, fix $\boldsymbol{\beta} \in \boldsymbol{B}$ and let $\{\boldsymbol{\beta}(m)\}$ denote a sequence of elements of $\boldsymbol{B}$ converging to $\boldsymbol{\beta}$ as m tends to infinity. Let Q denote the product measure $P \bigotimes P \bigotimes P$.

Then we have

$$QI\{\boldsymbol{x}_1^T\boldsymbol{\beta} = \boldsymbol{x}_2^T\boldsymbol{\beta} = \boldsymbol{x}_3^T\boldsymbol{\beta}\} = 0 \ .$$

This implies that

$$M(y_1, y_2)I\{\boldsymbol{x}_1^T\boldsymbol{\beta}(m) > \boldsymbol{x}_2^T\boldsymbol{\beta}(m) > \boldsymbol{x}_3^T\boldsymbol{\beta}(m)\} - M(y_1, y_2)I\{\boldsymbol{x}_1^T\boldsymbol{\beta} > \boldsymbol{x}_2^T\boldsymbol{\beta} > \boldsymbol{x}_3^T\boldsymbol{\beta}\} \to 0 \ as \ m \to \infty \ ,$$

for Q almost all(z_1, z_2, z_3). Applying the dominated convergence theorem and A5, we can get that $G(\boldsymbol{\beta})$ is continuous which establishes (iii). This proves the theorem.

□

We have denoted that $\boldsymbol{Z} = (Y, \boldsymbol{X})$ denotes an observation from the distribution P on the set $S \subseteq \boldsymbol{R} \bigotimes \boldsymbol{R}^d$, and that the parameter space $\boldsymbol{B}$ is a compact subset of $\{\boldsymbol{\beta} \in \boldsymbol{R}^d : \beta_d = 1\}$. For $\boldsymbol{\beta}$ in $\boldsymbol{B}$, $(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3)$ in $S \bigotimes S \bigotimes S$, (y_1, y_3) in $R \bigotimes R$, and $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ in $R^d \bigotimes R^d \bigotimes R^d$, we define

$$h(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{\beta}) = M(y_1, y_3)I\{\boldsymbol{x}_1^T\boldsymbol{\beta} > \boldsymbol{x}_2^T\boldsymbol{\beta} > \boldsymbol{x}_3^T\boldsymbol{\beta}\} \ .$$

For each $\boldsymbol{z}$ in S, we define the kernel function of the empirical process that drives the asymptotic behavior of $\boldsymbol{\beta}_n$ as

$$\tau(\boldsymbol{z}, \boldsymbol{\beta}) = h(\boldsymbol{z}, P, P, \boldsymbol{\beta}) + h(P, \boldsymbol{z}, P, \boldsymbol{\beta}) + h(P, P, \boldsymbol{z}, \boldsymbol{\beta}) \ ,$$

where $h(\boldsymbol{z}, P, P, \boldsymbol{\beta})$, for example, is short for the conditional expectation of $h(\cdot, \cdot, \cdot, \boldsymbol{\beta})$ given its first argument under $P \bigotimes P$.

Write ∇_m for the m-th partial derivative operator applied to the first $d-1$ components of $\boldsymbol{\beta}$, and

$$\|\nabla_m \tau(\boldsymbol{z}, \boldsymbol{\beta})\| = \sum_{i_1 i_2 \ldots i_m} \|\frac{\partial^m}{\partial\beta_{i_1}\ldots\partial\beta_{i_m}}\tau(\boldsymbol{z}, \boldsymbol{\beta})\| \,,$$

where the symbol $\|\cdot\|$ denotes the modulus of a matrix: $\|(a_{ij})\| = (\Sigma_{i,j} a_{ij}^2)^{1/2}$.

We need a few more assumptions for establishing the asymptotic normality.

A6. The element $\{\beta_1, \beta_2, \ldots \beta_{d-1}\}$ is an interior of a compact subset of $\boldsymbol{R}^{d-1}$.

A7. $\boldsymbol{X}$ and $\boldsymbol{\mu}$ are independent.

A8. On a neighborhood of $\boldsymbol{\beta}_0$, the second partial bounded derivatives of $\tau(\boldsymbol{z}, \boldsymbol{\beta})$ exist. And there exists an integrable function $M(\boldsymbol{z})$ such that

$$\|\nabla_2 \tau(\boldsymbol{z}, \boldsymbol{\beta})) - \nabla_2 \tau(\boldsymbol{z}, \boldsymbol{\beta}_0)\| \leq M(\boldsymbol{z})|\boldsymbol{\beta} - \boldsymbol{\beta}_0| \,,$$

where $E|\nabla_1 \tau(\cdot\boldsymbol{\beta})|^2 < \infty$ and the expectation matrix of $\nabla_2 \tau(\boldsymbol{z}, \boldsymbol{\beta})$ is negative definite.

Theorem 4.2.2 (Asymptotic normality)**.** *If A1-A8 hold, then*

$$\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \Longrightarrow (W^T, 0)^T \,,$$

where $\Longrightarrow$ *denotes convergence in distribution and W has a* $d-1$ *dimensional multivariate normal distribution* $N(\boldsymbol{0}, \boldsymbol{V}A\boldsymbol{V}^{-1})$ *distribution with* $3\boldsymbol{V} = E\nabla_2 \tau(\cdot, \boldsymbol{\beta}_0)$, $A = E\nabla_1 \tau(\cdot, \boldsymbol{\beta}_0)[\nabla_1 \tau(\cdot, \boldsymbol{\beta}_0)]^T$.

Proof of Theorem 4.2.2. Define

$$f(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{\beta}) = h(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{\beta}) - h(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{\beta}_0).$$

Suppose $f(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{\beta})$ is a real-valued function on the product space $S \bigotimes S \bigotimes S$. $G_n(\boldsymbol{\beta}), G(\boldsymbol{\beta})$ are defined in the precious section. Let $\Gamma_n(\boldsymbol{\beta})$ be

$$\Gamma_n(\boldsymbol{\beta}) = G_n(\boldsymbol{\beta}) - G_n(\boldsymbol{\beta}_0),$$

and the expectation $\Gamma(\boldsymbol{\beta})$ be

$$\Gamma(\boldsymbol{\beta}) = G(\boldsymbol{\beta}) - G(\boldsymbol{\beta}_0).$$

We note that $G_n(\boldsymbol{\beta}) - G(\boldsymbol{\beta})$ is a U-statistic of order three. Written $U_n^k f(\cdot, \cdot, \cdot, \boldsymbol{\beta})$ as U-statistic of of order k. Then,

$$G_n(\boldsymbol{\beta}) - G(\boldsymbol{\beta}) = U_n^3 f(\cdot, \cdot, \cdot, \boldsymbol{\beta}).$$

From the properties of the U-statistic, for the U-statistic of order k, there exist functions $f^1(\cdot, \cdot, \cdot, \boldsymbol{\beta}), ... f^k(\cdot, \cdot, \cdot, \boldsymbol{\beta})$ such that for each i, $f^i(\cdot, \cdot, \cdot, \boldsymbol{\beta})$ is P-degenerate on S^i, and

$$U_n^k(\cdot, \cdot, \cdot, \boldsymbol{\beta}) = P_n f^1(\cdot, \cdot, \cdot, \boldsymbol{\beta}) + \sum_{i=2}^{k} U_n^i f^i(\cdot, \cdot, \cdot, \boldsymbol{\beta}),$$

where P_n can be viewed as a random probability measure putting mass $\frac{1}{n}$ at each ordered k-tuple $(Z_{i_1}, ..., Z_{i_k})$ (Serfling (1980).

So,

$$\Gamma_n(\boldsymbol{\beta}) = \Gamma(\boldsymbol{\beta}) + P_n f^1(\cdot, \cdot, \cdot, \boldsymbol{\beta}) + U_n^2 f^2(\cdot, \cdot, \cdot, \boldsymbol{\beta}) + U_n^3 f^3(\cdot, \cdot, \cdot, \boldsymbol{\beta}).$$

Now we apply Taylor expansion of $\tau(\cdot, \boldsymbol{\beta})$ about $\boldsymbol{\beta}_0$:

$$\tau(\cdot, \boldsymbol{\beta}) = \tau(\cdot, \boldsymbol{\beta}_0) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \nabla_1 \tau(\cdot, \beta_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \nabla_2 \tau(\cdot, \boldsymbol{\beta}^\star)(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

for $\boldsymbol{\beta}^\star$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$.

For z in S,

$$\|(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T [\nabla_2 \tau(\boldsymbol{z}, \boldsymbol{\beta}) - \nabla_2 \tau(\boldsymbol{z}, \boldsymbol{\beta}_0)](\boldsymbol{\beta} - \boldsymbol{\beta}_0)\| \leq M(\boldsymbol{z})|\boldsymbol{\beta} - \boldsymbol{\beta}_0|^3.$$

From Theorem 4.2.1 of the consistency and the results from Sherman (1994), we have $E\tau(\boldsymbol{z}, \boldsymbol{\beta}) = 3\Gamma(\boldsymbol{\beta})$, and

$$\Gamma(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^T V\boldsymbol{\beta} + o(|\boldsymbol{\beta}|))^2 \quad as \ \boldsymbol{\beta} \to \boldsymbol{\beta}_0,$$

and

$$P_n f^1(\cdot, \cdot, \cdot, \boldsymbol{\beta}) = \frac{1}{\sqrt{n}}\boldsymbol{\beta}^T W_n + o_p(|\boldsymbol{\beta}|^2),$$

uniformly over $o_p(1)$ neighborhoods of $\boldsymbol{\beta}_0$, where $W_n = \sqrt{n}P_n\nabla_1\tau(\cdot, \boldsymbol{\beta}_0)$.

As a property of the U-statistic of order k (Sherman 1994), it will be true that

$$U_n^k f(\cdot, \cdot, \cdot, \boldsymbol{\beta}) = o_p(1/n^{\frac{k}{2}}).$$

So,

$$U_n^2 f^2(\cdot, \cdot, \cdot, \boldsymbol{\beta}) + U_n^3 f^3(\cdot, \cdot, \cdot, \boldsymbol{\beta}) = o_p(1/n).$$

Thus,

$$\Gamma_n(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^T V\boldsymbol{\beta} + \frac{1}{\sqrt{n}}\boldsymbol{\beta}^T W_n + o_p(|\boldsymbol{\beta}|^2) + o_p(1/n).$$

From the Corollary in Sherman's paper (1994), we can get that

$$\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \Longrightarrow (W, \mathbf{0}),$$

where $\Longrightarrow$ denotes convergence in distribution and W has a $N(\mathbf{0}, \boldsymbol{V}A\boldsymbol{V}^{-1})$ distribution with $3\boldsymbol{V} = E\nabla_2\tau(\cdot, \boldsymbol{\beta}_0)$ and $A = E\nabla_1\tau(\cdot, \boldsymbol{\beta}_0)[\nabla_1\tau(\cdot, \boldsymbol{\beta}_0)]^T$. □

In this thesis, we proposed the estimator extended from the semiparametric monotonic linear index model which has many advantages over other types of methods such

as the maximum linear separation (MLS) measure. The exact nature of the monotonicity is usually difficult to specify even as we often assume a monotonic relationship between a response and a linear index. Therefore, the estimator in the semiparametric monotonic linear index model can directly exploit monotonicity between a response and a linear index without any knowledge about the form of the monotonic relationship, and no parametric assumptions are needed about the error distribution. Another appealing property is that the estimator does not require any subjective bandwidth choice. Moreover, the proposed estimator allows more flexibility in balancing robustness and efficiency objectives for a wider range of models.

In this thesis, the best linear combination is the one which maximizes the VUS=$P(\boldsymbol{X}_{i_1}^T\boldsymbol{\beta} > \boldsymbol{X}_{i_2}^T\boldsymbol{\beta} > X_{i_3}^T\boldsymbol{\beta})$ among all the possible linear combinations. We denote the maximum VUS from the combination as $maxVUS$. Thus the bootstrap standard errors for the estimation of $maxVUS$ and the coefficient vector can be similarly applied as in the previous chapter.

For each of the bootstrap samples, denote the estimators for the maximum VUS by$\{\widehat{maxVUS}_n : n = 1, 2, ...N\}$ where N is the number of samples. The bootstrap standard error for $\widehat{maxVUS}$ is

$$\widehat{se_N}(maxVUS) = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\widehat{maxVUS}_n - \widehat{maxVUS})^2}. \tag{4.10}$$

A $100(1-\alpha)$% confidence interval for $maxVUS$ is

$$\widehat{maxVUS} \pm z_{\alpha/2}\widehat{se_N}(maxVUS), \tag{4.11}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile for the standard normal distribution.

Similarly, for each of the bootstrap samples, denote the estimators for the coefficient vector by $\{\widehat{\boldsymbol{\beta}}_n : n = 1, 2, ...N\}$ where N is the number of samples. The corresponding bootstrap standard error for $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{se_N}(\boldsymbol{\beta}) = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\widehat{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\beta}})^2}. \tag{4.12}$$

A $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} \pm z_{\alpha/2}\widehat{se_N}(\boldsymbol{\beta}), \tag{4.13}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile for the standard normal distribution.

4.2.2 Normal distribution assumption

Diagnostic test data have been modeled under a normal distribution in many studies. Rich literature also exists for combining markers by using multivariate normal properties. Su and Liu (1993) provided classic results developed under the delicate multivariate theories. We also provide a simple parametric result for the optimal combination of which the distribution of the data from the multiple classes are assumed to be normal.

In this section we consider a special parametric case when the random vectors for the three classes follow normal distributions. Under such a tri-normal scenario we can obtain an exact formulation for the combination coefficients.

Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ are from d-dimensional multivariate normal distribution. $\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\boldsymbol{X}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $\boldsymbol{X}_3 \sim N(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$ are mean vectors and $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3$ are variance-covariance matrices for the three classes, respectively. For a coefficient vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_d)^T$,

$$\boldsymbol{\beta}^T \boldsymbol{X}_i \sim N(\boldsymbol{\beta}^T \boldsymbol{\mu}_i, \boldsymbol{\beta}^T \boldsymbol{\Sigma}_i \boldsymbol{\beta}),$$

for $i = 1, 2, 3$.

Then we intend to find a $\boldsymbol{\beta}_0$ that maximizes the following

$$VUS = P(\boldsymbol{\beta}^T \boldsymbol{X}_1 > \boldsymbol{\beta}^T \boldsymbol{X}_2 > \boldsymbol{\beta}^T \boldsymbol{X}_3),$$

which can be expressed by

$$\begin{aligned}
VUS &= \int\int\int_{\boldsymbol{\beta}^T \boldsymbol{x}_1 > \boldsymbol{\beta}^T \boldsymbol{x}_2 > \boldsymbol{\beta}^T \boldsymbol{x}_3} f_{x_1}(\boldsymbol{\beta}^T \boldsymbol{x}_1) f_{x_2}(\boldsymbol{\beta}^T \boldsymbol{x}_2) f_{x_3}(\boldsymbol{\beta}^T \boldsymbol{x}_3) d(\boldsymbol{\beta}^T \boldsymbol{x}_3) d(\boldsymbol{\beta}^T \boldsymbol{x}_2) d(\boldsymbol{\beta}^T \boldsymbol{x}_1) \\
&= \int_{-\infty}^{\infty} d(\boldsymbol{\beta}^T \boldsymbol{x}_3) \int_{\boldsymbol{\beta}^T \boldsymbol{x}_3}^{\infty} d(\boldsymbol{\beta}^T \boldsymbol{x}_2) \int_{\boldsymbol{\beta}^T \boldsymbol{x}_2}^{\infty} f_{x_1}(\boldsymbol{\beta}^T \boldsymbol{x}_1) f_{x_2}(\boldsymbol{\beta}^T \boldsymbol{x}_2) f_{x_3}(\boldsymbol{\beta}^T \boldsymbol{x}_3) d(\boldsymbol{\beta}^T \boldsymbol{x}_1) \\
&= \int_{-\infty}^{\infty} d(\boldsymbol{\beta}^T \boldsymbol{x}_3) \int_{\boldsymbol{\beta}^T \boldsymbol{x}_3}^{\infty} f_{x_2}(\boldsymbol{\beta}^T \boldsymbol{y}) f_{x_3}(\boldsymbol{\beta}^T \boldsymbol{x_3}) [1 - F_{x_1}(\boldsymbol{\beta}^T \boldsymbol{x}_2)] d(\boldsymbol{\beta}^T \boldsymbol{x}_2) \\
&= \int_{-\infty}^{\infty} F_{x_3}(\boldsymbol{\beta}^T \boldsymbol{x}_2) [1 - F_{x_1}(\boldsymbol{\beta}^T \boldsymbol{x}_2)] f_{x_2}(\boldsymbol{\beta}^T \boldsymbol{x}_2) d(\boldsymbol{\beta}^T \boldsymbol{x}_2) \\
&= \int_{-\infty}^{\infty} \Phi\left(\frac{t - \boldsymbol{\beta}^T \boldsymbol{\mu}_3}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_3 \boldsymbol{\beta}}}\right) \cdot \Phi\left(\frac{-(t - \boldsymbol{\beta}^T \boldsymbol{\mu}_1)}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_1 \boldsymbol{\beta}}}\right) \cdot \Phi\left(\frac{t - \boldsymbol{\beta}^T \boldsymbol{\mu}_2}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_2 \boldsymbol{\beta}}}\right) \cdot \frac{1}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_2 \boldsymbol{\beta}}} dt.
\end{aligned}$$

Let $s = \dfrac{t - \boldsymbol{\beta}^T \boldsymbol{\mu}_2}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_2 \boldsymbol{\beta}}}$, then $t = \sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_2 \boldsymbol{\beta}} s + \boldsymbol{\beta}^T \boldsymbol{\mu}_2$. Thus, the VUS can be written as

$$VUS = \int_{-\infty}^{\infty} \Phi\left(\frac{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_2 \boldsymbol{\beta}}}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_3 \boldsymbol{\beta}}} s + \frac{\boldsymbol{\beta}^T (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_3 \boldsymbol{\beta}}}\right) \cdot \Phi\left(-\frac{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_2 \boldsymbol{\beta}}}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_1 \boldsymbol{\beta}}} s + \frac{\boldsymbol{\beta}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_1 \boldsymbol{\beta}}}\right) \varphi(s) ds,$$

where F denotes the distribution function and f denotes the density function, S denotes the survival distribution, Φ denotes the normal distribution function, φ denotes the normal density function.

$$\text{Let } a = \frac{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_2 \boldsymbol{\beta}}}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_3 \boldsymbol{\beta}}}, \quad b = \frac{\boldsymbol{\beta}^T (\boldsymbol{\mu}_3 - \boldsymbol{\mu}_3)}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_3 \boldsymbol{\beta}}}, \quad c = \frac{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_2 \boldsymbol{\beta}}}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_1 \boldsymbol{\beta}}}, \quad d = \frac{\boldsymbol{\beta}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_1 \boldsymbol{\beta}}}.$$

Then, the VUS is a form as

$$VUS = \int_{-\infty}^{\infty} \Phi(as + b)\Phi(-cs + d)\phi(s)ds.$$

Differentiating with respect to $\boldsymbol{\beta}$, we can solve the equation for $\boldsymbol{\beta}$ which maximizing the VUS.

$$\frac{\partial VUS}{\partial \boldsymbol{\beta}} = \int_{-\infty}^{\infty} \Phi(-cs + d)\phi(as + b)\phi(s)sds \cdot \frac{\partial a}{\partial \boldsymbol{\beta}} + \int_{-\infty}^{\infty} \Phi(-cs + d)\phi(as + b)\phi(s)ds \cdot \frac{\partial b}{\partial \boldsymbol{\beta}}$$

$$+ \int_{-\infty}^{\infty} \Phi(as + b)\phi(-cs + d)\phi(s)(-s)ds \cdot \frac{\partial c}{\partial \boldsymbol{\beta}} + \int_{-\infty}^{\infty} \Phi(as + b)\phi(-cs + d)\phi(s)ds \cdot \frac{\partial d}{\partial \boldsymbol{\beta}} = 0$$

$$= A_1 \cdot \frac{\partial a}{\partial \boldsymbol{\beta}} + A_2 \cdot \frac{\partial b}{\partial \boldsymbol{\beta}} + A_3 \cdot \frac{\partial c}{\partial \boldsymbol{\beta}} + A_4 \cdot \frac{\partial d}{\partial \boldsymbol{\beta}}.$$

We now calculate the four parts.

$$\begin{aligned}
A_1 &= \int_{-\infty}^{\infty} \Phi(-cs + d)\phi(as + b)\phi(s) \cdot sds \\
&= \frac{1}{2\pi} exp\{-\frac{1}{2}(\frac{b^2}{a^2 + 1})\} \int_{-\infty}^{\infty} \Phi(-cs + d) exp\{-\frac{a^2 + 1}{2}(s + \frac{ab}{a^2 + 1})^2\} sds \\
&= \frac{1}{2\pi} exp\{-\frac{1}{2}(\frac{b^2}{a^2 + 1})\} \int_{-\infty}^{\infty} \Phi(-cs + d) exp\{-\frac{a^2 + 1}{2}(s + \frac{ab}{a^2 + 1})^2\}(s + \frac{ab}{a^2 + 1})ds \\
&- \frac{1}{2\pi} exp\{-\frac{1}{2}(\frac{b^2}{a^2 + 1})\} \int_{-\infty}^{\infty} \Phi(-cs + d) exp\{-\frac{a^2 + 1}{2}(s + \frac{ab}{a^2 + 1})^2\} \frac{ab}{a^2 + 1} ds \\
&= A_5 - A_6.
\end{aligned}$$

$$\begin{aligned} A_5 &= \frac{1}{4\pi} exp\{-\frac{1}{2}(\frac{b^2}{a^2+1})\} \int_{-\infty}^{\infty} \Phi(-cs+d) exp\{-\frac{a^2+1}{2}(s+\frac{ab}{a^2+1})^2\} d(s+\frac{ab}{a^2+1})^2 \\ &= \frac{1}{2\pi} \cdot \frac{-c}{a^2+1} \cdot \frac{1}{\sqrt{a^2+1+c^2}} \cdot exp\{-\frac{1}{2}[b^2+d^2-\frac{(ab-cd)^2}{a^2+1+c^2}]\}. \end{aligned}$$

$$\begin{aligned} A_2 &= \int_{-\infty}^{\infty} \Phi(-cs+d)\phi(as+b)\phi(s)ds \\ &= \int_{-\infty}^{\infty} \Phi(-cs+d) \cdot \frac{1}{\sqrt{(2\pi}} \cdot exp\{-\frac{1}{2}[\sqrt{a^2+1}s+\frac{ab}{\sqrt{a^2+1}}]^2\} ds. \end{aligned}$$

Let $t = \sqrt{a^2+1}s + \frac{ab}{\sqrt{a^2+1}}$. Then,

$$\begin{aligned} A_2 &= \frac{1}{\sqrt{2\pi}(a^2+1)} exp\{-\frac{b^2}{2(a^2+1)}\} \int_{-\infty}^{\infty} \Phi(-\frac{c}{\sqrt{a^2+1}}t + \frac{abc}{a^2+1} + d)\phi(t)dt \\ &= \frac{1}{\sqrt{2\pi}(a^2+1)} exp\{-\frac{b^2}{2(a^2+1)}\} \Phi(\frac{abc+(a^2+1)d}{\sqrt{(a^2+1)(a^2+1+c^2)}}). \end{aligned}$$

$$\because A_6 = \frac{ab}{a^2+1} \cdot A_2,$$

then,

$$A_2 \cdot \frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\beta}} - A_6 \cdot \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{\beta}} = A_2(\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\beta}} - \frac{ab}{a^2+1}\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{\beta}}) = A_2 \cdot \frac{a^2+1}{2b} \cdot \frac{\partial}{\partial \boldsymbol{\beta}}(\frac{b^2}{a^2+1}).$$

Thus,

$$A_1\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{\beta}} + A_2\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\beta}} = \frac{1}{2\pi} \cdot \frac{-c}{a^2+1} \cdot \frac{1}{\sqrt{a^2+1+c^2}} \cdot exp\{-\frac{1}{2}[b^2+d^2-\frac{(ab-cd)^2}{a^2+1+c^2}]\} \cdot \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{\beta}}$$

$$+\frac{\sqrt{a^2+1}}{2b\sqrt{2\pi}} \cdot exp\{-\frac{1}{2} \cdot \frac{b^2}{a^2+1}\} \cdot \Phi(\frac{abc+(a^2+1)d}{\sqrt{(a^2+1)(a^2+1+c^1)}}) \cdot \frac{\partial}{\partial \boldsymbol{\beta}}(\frac{b^2}{a^2+1}).$$

Similarly,

$$A_3\frac{\partial \boldsymbol{c}}{\partial \boldsymbol{\beta}} + A_4\frac{\partial \boldsymbol{d}}{\partial \boldsymbol{\beta}} = \frac{1}{2\pi} \cdot \frac{-a}{c^2+1} \cdot \frac{1}{\sqrt{a^2+1+c^2}} \cdot exp\{-\frac{1}{2}[b^2+d^2-\frac{(ab-cd)^2}{a^2+1+c^2}]\} \cdot \frac{\partial \boldsymbol{c}}{\partial \boldsymbol{\beta}}$$

$$+\frac{\sqrt{c^2+1}}{2d\sqrt{2\pi}} \cdot exp\{-\frac{1}{2} \cdot \frac{d^2}{c^2+1}\} \cdot \Phi(\frac{adc+(c^2+1)b}{\sqrt{(c^2+1)(a^2+1+c^1)}}) \cdot \frac{\partial}{\partial \boldsymbol{\beta}}(\frac{d^2}{c^2+1}).$$

Denote

$$x_1 = \frac{(a^2+1)d + abc}{\sqrt{(a^2+1)(a^2+1+c^2)}}, \qquad x_2 = \frac{(c^2+1)b + adc}{\sqrt{(c^2+1)(a^2+1+c^2)}},$$

$$C_1 = -\frac{\sqrt{2\pi}}{2b} \cdot \sqrt{(a^2+1)(a^2+1+c^2)} \cdot e^{\frac{1}{2}x_1^2}\Phi(x_1),$$

$$C_2 = -\frac{\sqrt{2\pi}}{2d} \cdot \sqrt{(c^2+1)(a^2+1+c^2)} \cdot e^{\frac{1}{2}x_2^2}\Phi(x_2).$$

Thus the equation can be written as:

$$[\frac{c}{a^2+1}\frac{\partial a}{\partial \boldsymbol{\beta}} + \frac{a}{c^2+1}\frac{\partial c}{\partial \boldsymbol{\beta}}] + C_1\frac{\partial}{\partial \beta}(\frac{b^2}{a^2+1}) + C_2\frac{\partial}{\partial \beta}(\frac{d^2}{c^2+1}) = 0.$$

The analytic solution is not generally attainable. As such, we consider a special case for which $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{I}$, and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_3 = \boldsymbol{\delta}$. That is, we assume a constant covariance $\boldsymbol{I}$ for the three categories and equal distances between the two adjacent categories.

For the results of Liu and Su (1993), we can derive that the coefficients for the best linear combination are proportional to $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$.

4.3 Simulation studies

We conducted simulation studies to assess the performance of the proposed method. Sample sizes of 60, 120 and 150 were considered. In our article, we considered four simulation settings. In each simulation, we fixed $\boldsymbol{\beta} = (\beta_1, \beta_2 ... \beta_d), (d = 2, 3, 4)$ which

maximized $Pr(\boldsymbol{\beta}^T \boldsymbol{X}_1 > \boldsymbol{\beta}^T \boldsymbol{X}_2 > \boldsymbol{\beta}^T \boldsymbol{X}_3)$. We set $\beta_d = 1$ for identifiability and only estimated $(\beta_1, ...\beta_{d-1})$. For the estimation of the standard errors, we applied the standard bootstrap procedure with 500 resamples.

In Case 1, we generated $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ from two-dimensional multivariate normal distributions with mean vectors $(2.2, 2.0)^T$, $(1.1, 1.0)^T$, $(0, 0)^T$, respectively, and covariance matrices being identical as a two-dimensional identity matrix. By using the results in Section 2.2, we derived the best linear combination and obtained the maximal probability $Pr(\boldsymbol{\beta}^T \boldsymbol{X}_1 > \boldsymbol{\beta}^T \boldsymbol{X}_2 > \boldsymbol{\beta}^T \boldsymbol{X}_3)$ to be 0.87.

In Case 2, we generated $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ from three-dimensional multivariate normal distributions with mean vectors $(2.4, 2.2, 2.0)^T$, $(1.2, 1.1, 1.0)^T$, $(0, 0, 0)^T$, respectively, and covariance matrices being identical as a three-dimensional identity matrix. By using the results in Section 2.2, we derived the best linear combination and obtained the maximal probability $Pr(\boldsymbol{\beta}^T \boldsymbol{X}_1 > \boldsymbol{\beta}^T \boldsymbol{X}_2 > \boldsymbol{\beta}^T \boldsymbol{X}_3)$ to be 0.90.

In Case 3, we generated $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ from four-dimensional multivariate normal distributions with mean vectors $(2.6, 2.4, 2.2, 2.0)^T$, $(1.3, 1.2, 1.1, 1.0)^T$, $(0, 0, 0, 0)^T$, respectively, and covariance matrices being identical as a four-dimensional identity matrix. By using the results in Section 2.2, we derived the best linear combination and obtained the maximal probability $Pr(\boldsymbol{\beta}^T \boldsymbol{X}_1 > \boldsymbol{\beta}^T \boldsymbol{X}_2 > \boldsymbol{\beta}^T \boldsymbol{X}_3)$ to be 0.89.

We used the nonparametric MRC estimation to estimate the coefficients for the simulated data in Case 1 to Case 3. The estimation results for the three cases are summarized in Tables 4.1, 4.2, and 4.3. For each case, the coefficients are listed in the column

$\boldsymbol{\beta}$. The average of the estimated coefficients in 1000 simulations are given in the column $\hat{\boldsymbol{\beta}}$. The sample standard deviation of the estimated coefficients are given in the column $sd(\boldsymbol{\beta})$. We applied bootstrap method to account for the variability in this paper. The average of the estimated standard errors are given in the column $\widehat{s.e.}$ To see how well the nonparametric estimation methods performs, we also calculated the coverage rates at the nominal 95% level, given in the column 'coverage rates'. In all cases, the estimated coefficients are consistent to the true coefficients. The results shows a well performance and the performance improves as sample size grows large. Our proposed methods appear to work satisfactorily well for these finite sample studies.

In the three cases, we specified multivariate normality assumptions. In additional to multivariate normal distributions, we considered the wishart distribution as well. In Case 4, we generated $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ from wishart distribution with Σ of $\begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$, $\begin{bmatrix} 2 & 1.5 \\ 1.5 & 2 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, respectively, and degree of 10. We derived the best linear combination and obtained the maximal probability $Pr(\boldsymbol{\beta}^T\boldsymbol{X}_1 > \boldsymbol{\beta}^T\boldsymbol{X}_2 > \boldsymbol{\beta}^T\boldsymbol{X}_3)$ to be 0.72. Results are listed in Table 4.4

4.4 Applications

4.4.1 Proteomic study for liver cancer

We first considered a recent mass spectrometry dataset for the detection of Glycan biomarkers for liver cancer (Ressom et al. (2007, 2008)). The researchers investigated 203 participants from Cairo, Egypt; 73 hepatocellular carcinoma (denoted by HC) cases; 52 patients with chronic liver disease (denoted by QC); and 78 healthy individuals (denoted by NC). The spectra were generated by a matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass analyzer (Applied Biosystems Inc., Frammingham, MA). We downloaded the dataset from the authors' public website and focused on a set of 484 peaks after extensive preprocessing of the raw data.

Each peak may be regarded as a diagnostic test for differentiating the subjects from the three distinctive classes: HC, QC and NC. In this case, the diagnostic task involves more than two categories. Placing an individual into any wrong category may result in adverse consequences. The accuracy of the diagnostic test thus should be reflected by how often the test correctly classifies all three categories. We were interested in studying the diagnostic accuracy of these peaks and identified those peaks with the highest discriminatory ability. Previously, Ressom et al. (2007, 2008) conducted analysis by reducing the number of categories in order to frame a few pairwise two-category classification problems. Pairwise ROC curves and the areas under the ROC curves (AUC) were reported to investigate the differentiability between two classes (eg. HC and QC).

However, such AUC measures cannot summarize the overall accuracy for three categories.

A more appropriate summary measure is VUS as discussed in this thesis. We first estimated VUS values for all the peaks by using the methods in Li and Fine (2008) and then focused on the top twenty peaks among the 484 peaks. They are gene 183, gene 209, gene 147, gene 443, gene 182, gene 262, gene 239, gene 472, gene 368, gene 134, gene 306, gene 188, gene 299, gene 311, gene 361, gene 483, gene 104, gene 425, gene 210, and gene 294, denoted as $D_1, D_2, ..., D_{20}$, respectively. It is noted from our calculation that the largest VUS is only approximately 0.65, indicating that in about 65% of all classification jobs such a peak can correctly sort the three classes of subjects. Evidently, using only a single peak may result in inadequate accuracy. Thus, we then applied the methods introduced in this thesis to build a more accurate classifier by combining multiple peaks.

We considered a selection procedure with these twenty genes, starting with the peak with the highest VUS and sequentially adding peaks which maximized the VUS based upon the joint model. At each step, we estimated the coefficient for optimal combination and then calculated the HUM values. The model selection results are summarized in Table 4.5. We noticed that after including five peaks in our model, the VUS value reached about 86% and no longer significantly increased by adding to the number of peaks. The VUS values no longer increased after the sixth iteration and so we closed at this point. The final model showed a large improvement in VUS values. We also

applied the bootstrap methodology to calculate the corresponding standard error for each combination.

We also estimated the coefficients for the markers in Table 4.5. The size and sign of the coefficients can indicate the relative importance of the marker and the direction of their association with the disease outcome. For the sake of comparison, we also considered a forward selection based upon multinomial logistic regression as in Li and Fine (2008). This approach gave a different combination $D_1 + 3.7D_4 - 3.1750D_{17} + 0.3562D_3 - 1.4D_{10} + 1.0625D_{18}$, with VUS value of 0.843. Compared to the VUS value of 0.860 from our proposed methodology, it seems that our method can provide a higher VUS after combining biomarkers.

4.4.2 Evaluating tissue biomarkers of synovitis

Although the methodology we introduced is contextual to three-category classification problems, there is little difficulty to extend our results to higher dimensional classifications. In this section, we considered an example in which we analyzed five distinct categories.

Krenn et al. (2006) described a three-component score for the grading of the histological severity of synovitis. Each of the three components (lining thickness, inflammatory infiltrates, and stromal density) was graded on a scale from zero to three. In this case, the primary classification outcome involves five different categories. The sample sizes for each category are given in Table 4.6.

We first quantified the diagnostic accuracy for each component and then determined the best linear combination to achieve the highest accuracy.

The estimated HUM are reported as we denote lining thickness as M_1, stromal density as M_2 and inflammatory infiltrates as M_3. We considered a combined score $\beta_1 M_1 + \beta_2 M_2 + M_3$ and estimated the unknown coefficients β_1 and β_2 which maximized the HUM. Stromal density appears to be the most accurate among the three tissue markers with a HUM of 0.0124, followed by lining thickness and inflammatory infiltrates. The estimated coefficients are $\hat{\beta}_1 = 1.03$ and $\hat{\beta}_2 = 1.07$. Clearly individual markers with higher accuracy receive relatively larger weights to build the optimal score. We noticed that the estimated HUM for the optimal linear combination was more than ten times larger than the HUM for any of the three markers. Using information from three markers can thus substantially improve the clinical diagnosis for the multiple categories and stages of inflammatory arthropathies. For the sake of comparison, we also computed the HUM for a naive combination of the three biomarkers by summing them together. The resulting HUM is only 0.0624 which is much lower than the maximum attainable HUM. The results are reported in Table 4.7. We also calculated the corresponding p-value for any two rows in Table 4.8. All the p-values (Table 4.8) are less than 0.05, implying significant differences.

Table 4.1: Estimated $\boldsymbol{\beta}$ which maximizes $P(\boldsymbol{\beta}^\top \boldsymbol{X}_1 > \boldsymbol{\beta}^\top \boldsymbol{X}_2 > \boldsymbol{\beta}^\top \boldsymbol{X}_3)$ in Case 1.

Sample size	$\boldsymbol{\beta}$	$\widehat{\boldsymbol{\beta}}$	$sd(\boldsymbol{\beta})$	$\widehat{s.e}$	coverage rates
60	1.1	1.209	0.0164	0.0243	0.937
120	1.1	1.135	0.0163	0.0258	0.943
150	1.1	1.110	0.0162	0.0199	0.944

Table 4.2: Estimated $\boldsymbol{\beta}$ which maximizes $P(\boldsymbol{\beta}^\top \boldsymbol{X}_1 > \boldsymbol{\beta}^\top \boldsymbol{X}_2 > \boldsymbol{\beta}^\top \boldsymbol{X}_3)$ in Case 2.

Sample size	$\boldsymbol{\beta}$	$\widehat{\boldsymbol{\beta}}$	$sd(\boldsymbol{\beta})$	$\widehat{s.e}$	coverage rates
60	1.2	1.252	0.0191	0.0255	0.935
	1.1	1.167	0.0156	0.0259	0.938
120	1.2	1.245	0.0193	0.0286	0.938
	1.1	1.134	0.0176	0.0247	0.942
150	1.2	1.222	0.0197	0.0239	0.937
	1.1	1.119	0.0163	0.0219	0.940

Table 4.3: Estimated $\boldsymbol{\beta}$ which maximizes $P(\boldsymbol{\beta}^\top \boldsymbol{X}_1 > \boldsymbol{\beta}^\top \boldsymbol{X}_2 > \boldsymbol{\beta}^\top \boldsymbol{X}_3)$ in Case 3.

Sample size	$\boldsymbol{\beta}$	$\widehat{\boldsymbol{\beta}}$	$sd(\boldsymbol{\beta})$	$\widehat{s.e}$	coverage rates
60	1.3	1.255	0.0197	0.0313	0.932
	1.2	1.244	0.0196	0.0280	0.933
	1.1	1.149	0.0175	0.0284	0.937
120	1.3	1.265	0.0198	0.0274	0.939
	1.2	1.243	0.0192	0.0211	0.935
	1.1	1.115	0.0169	0.0209	0.940
150	1.3	1.281	0.0192	0.0224	0.934
	1.2	1.181	0.0176	0.0239	0.937
	1.1	1.087	0.0169	0.0210	0.941

Table 4.4: Estimated $\boldsymbol{\beta}$ which maximizes $P(\boldsymbol{\beta}^\top \boldsymbol{X}_1 > \boldsymbol{\beta}^\top \boldsymbol{X}_2 > \boldsymbol{\beta}^\top \boldsymbol{X}_3)$ in Case 4.

Sample size	$\widehat{\boldsymbol{\beta}}$	$sd(\boldsymbol{\beta})$	$\widehat{s.e}$	coverage rates
60	1.081	0.3721	0.0299	0.939
120	1.113	0.3290	0.0332	0.940
150	1.130	0.2837	0.0262	0.942

Table 4.5: Estimated optimal volume under the ROC surfaces (VUS) for each step of the forward selection. Standard error and P-values are computed by using the bootstrap method.

Step	VUS	Model	s.e	P-value
1	0.647	D_1	0.04	
2	0.750	$0.1250D_3 + D_4$	0.039	< 0.001
3	0.808	$1.3606D_12 - 3.9046D_{17} + D_{18}$	0.037	< 0.001
4	0.850	$7.8778D_3 + 25.3139D_4 - 43.4810D_{17} + D_{20}$	0.036	< 0.001
5	0.859	$1.78D_3 + 6.85D_4 + 6.26D_{14} - 11.75D_{17} + D_{18}$	0.034	< 0.001
6	0.860	$5.76D_3 + 15.33D_4 + 5.23D_{19} - 42.94D_{17} + 10.12D_{18} + D_{20}$	0.028	0.31

Table 4.6: The sample sizes for each category in the synovitis data.

Category	Sample size
Normal healthy control	33
Post-traumatic arthropathy (PtA)	29
Osteoarthritis (OA)	221
Psoriatic arthritis (PsA)	42
Rheumatoid arthritis (RA)	341

Table 4.7: Estimated hypervolume under the ROC manifold (HUM) values for synovitis biomarkers.

Marker	HUM
M_1	0.0085
M_2	0.0124
M_3	0.0011
$M_1 + M_2 + M_3$	0.0624
$1.03M_1 + 1.07M_2 + M_3$	0.1020

Table 4.8: P-values.

P-values	M_1	M_2	M_3	$M_1 + M_2 + M_3$
M_1		4.969×10^{-5}	9.976×10^{-14}	2.513×10^{-67}
M_2	4.969×10^{-5}		1.505×10^{-5}	4.044×10^{-42}
M_3	9.976×10^{-14}	1.505×10^{-5}		1.64×10^{-65}
$M_1 + M_2 + M_3$	2.513×10^{-67}	4.044×10^{-42}	1.64×10^{-65}	
$1.03M_1 + 1.07M_2 + M_3$	2.41×10^{-128}	2.59×10^{-119}	4.63×10^{-156}	1.164×10^{-29}

Chapter 5

Conclusion and Further Research

5.1 Conclusion

Although the multiple-category ROC framework and corresponding HUM were originally introduced by Scurfield (1996), their practical use in empirical analysis was not thoroughly examined. Mossman (1999) simulated statistical work attempting to translate the identified theoretical HUM construct given by Scurfield into practical inferences. Subsequently, a wholly acceptable solution for resolving issues pertaining to multiple tests has not been made fully available. Furthermore, obtaining direct probability assessments from such tests is unfeasible. Simple decision rules are not flexible enough for many applications, like microarray data, where there are many tests and unordered categories.

Our proposed methods overcome this problem by using estimated class probabilities. The main advantage of our proposed method is the simplification of computation required for screening the useless tests and identifying the most useful tests. Due to the uncertainty of the ordering relationship among multiple categories, we need to first determine the correct expression for HUM. Our computation is much lower than the exclusive computation of all possible HUM values. When the number of categories are large, we can provide huge savings in computation time and energy. The correct identification of the ordering relationship among classes prevents us from screening good tests.

Even if the continuous test is not ordered because of the nature of multiple categories, the numeric values can always be ordered. For unordered multiple-category ROC analysis, Li and Fine (2008) used a method based upon Mossman's decision rule and achieved a reasonable estimation of HUM without knowledge of the correct class order. Such a method does not clearly reveal the relative magnitudes of the multiple classes and may not be appealing for interpreting the implications of HUM. Our proposed strategy yields the same estimation of HUM and provides additional information regarding the ordering of numerical test values from different classes.

Distinct diagnostic markers can be sensitive influences to various aspects of the disease being studied. In such cases, applying a linear combination can reveal a 'new' marker comprised of multiple biomarkers which can enhance diagnostic capability. We proposed a new rank estimator and also provided the consistency theorem of the coeffi-

cient estimators. The theorem can be extended to the k-choice-task model under which multiple-dimensional open wedges can be constructed.

Our methodology, which applies the bootstrap method to calculate the variance of the maximum VUS and HUM, was relatively efficient and effective when applied to the computation-heavy simulation results in this paper. The data analysis demonstrates that the best linear combination maximizes the VUS and HUM under a three-class and multiple-class case, respectively. The resulting models based upon the related linear combinations generate further insight into the mass spectrometry dataset.

5.2 Topics for further research

With the increasing number of applications for AUC and related measures in medical field and clinical studies, we have noticed that the AUC values are at times lower than 1/2. Such AUC values are sometimes overlooked or intentionally omitted, especially in large-scale microarray studies. However, they may hold important information about the accuracy of diagnostic tests. In this thesis, we proposed a simple method of rotating 180 degrees to cause the ROC plot to emerge above the chance diagonal line. In future work, we may further consider the concave ROC curve properties and propose nonparametric methods.

Identifying the correct classification for multiple-category classifications is comparatively complicated. Instead of applying the U-statistic approach to calculate the

VUS and HUM, we proposed bootstrap standard errors for the multiple-category ROC analysis, which could significantly remedy the computational burden. In this thesis, we followed the bootstrap approach in Li and Fine (2008) and chose a bootstrap sample size of 500. However, some future work remains to determine the bootstrap sample size. In fact, great interest exists to come up with effective approaches to design and evaluate the bootstrap sample size. The calculation of the corresponding confidence interval of the bootstrap p-values is also complicated, and there is limited literature concerning its calculation. This should open a path for further research.

Sometimes the data distribution could be highly skewed even after the normalization transformation. Outlier or extreme observations might also exist and influence the estimation of distribution means. When distribution conditions are not satisfactorily met, parametric methods may not always indicate the correct ordinal relationship of test results among groups. One might seek distribution-free nonparametric methods to identify the order. Weighted average of the distribution may be another topic for further research.

The MRC estimator has recently attracted much attention from classification literature due to its close relationship with the ROC curve. Combining predictors for classification is discussed in this thesis. We explored statistical methods of a linear combination of multiple tests for multiple-category classifications to optimize the accuracy from the combined markers. Further research may also attempt to solve for non-linear combinations which maximize the VUS or HUM of multiple-category classifications.

A closed-form expression for the best-fitting parameters may sometimes not exist, as there is in a linear combination framework. With the introduction of methods that can solve some of the computational burden of multiple-category problems, the data can be fitted by a method of successive approximations within a viable computational capacity to derive the target nonlinear model.

In this thesis, we applied the nonparametric estimators of HUM and suggested the resampling bootstrap method to calculate the standard errors for the estimators of HUM and the coefficient vectors. This can be viewed as an in-sample estimate. However, when we take an independent sample of the validation data from the same population as the training data, overfitting can sometimes occur; that is, the model does not fit the validation data as well as it fits the training data. This is most likely to occur when the number of parameters is large and the size of the training dataset is very small. Cross-validation is then an applicable way to assess how the results of a statistical analysis will generalize to independent datasets. It involves partitioning a sample of data into complementary subsets, assessing the analysis on the training set and validating the analysis on the testing set. Thus, in particular situations, the application of cross-validation is also of interest for further research.

For binary classification, Pepe and Thompson (2000) developed a method based upon maximizing the AUC to combine biomarkers in genetic studies. Their method was essentially adapted from the maximum rank correlation estimation. In this thesis, we provide statistical approach which yields the best linear combination to maximize

VUS or HUM. Li and Fine (2008) considered multinomial logistic regression to address multi-category outcomes. Further research may also focus on the inferences which yield the most effective multinomial logistic regression to maximize VUS or HUM.

References

Albrecht, A., Vinterbo, S.A., Ohno-Machado, L. (2003). An Epicurean Learning Approach to Gene-expression data Classification. *Artificial Intelligence in Medicine* **97**, 245–271.

Albrecht, A. (2007). Stochastic local search for the feature set problem, with application to microarray data. *Applied Mathematics and Computation* **183**, 1148–1164.

Alonzo, T.A. and Pepe, M.S. (1999). Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* **18**, 2987-3003.

Alonzo, T.A. and Pepe, M.S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics* **3**, 421-432.

Andriy, I.B., Howard, E.R., David, G. (2007). Exact bootstrap variances of the area under ROC curve. *Communications in Statistics-Theory and Methods* **36**, 2443-2461.

Baker, S.G. (1995). Evaluating multiple diagnostic tests with partial verification *Biometrics* **51**, 330-337.

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

Begg, C.B. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine* **6**, 411-423.

Beiden, S.V., Wagner, R.F. and Campbell, G.(2000). Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects receiver operating characteristic analysis. *Academic Radiology* **13**, 414-420.

Birnbaum, Z.W. (1956). On a use of the Mann-Whitney statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability I (J. Neyman, Ed.)*, 13-17.

Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* **13**, 499-508.

Chandra, S. and Owen, D.B. (1975). Estimating reliability of a component subject to several different stresses. *Naval Research Logistics* **22**, 31–39.

Cheng, H., Macajuso, M. and Hardin, J.M. (2000). Validity and coverage of estimates of relative accuracy. *Annals of Epidemiology* **10**, 251-260.

Christopher, C. and Sherman, R.P. (1998). Rank estimators for monotonic index models. *Journal of Econometrics* **84**, 351-381.

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society* **60**, 71-87.

Cole, P. and Morrison, A.S. (1980). Basic issues in population screening for cancer. *Journal of the National Cancer Institute* **64**, 1263-1272.

Delong, E.R., DeLong, D.M. and Clarke Pearson, D.L. (1988). Comparing the areas under the two or more correlated receiver operating characteristic curves: A non-parametric approach. *Biometrics* **44**, 837-845.

Diamond, G.A. (1992). Clinical epistemology of sensitivity and specificity. *Journal of Clinical Epidemiology* **45**, 9-13.

Dorfman, D.D. and Alf, J.E. (1968). Maximum-likelihood estimation of parameters of signal-detection theory-a direct solution. *Psychometrika* **33**, 113-124.

Dorfman, D.D. and Alf, J.E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating method data. *Mathematical Psychometrics* **6**, 487-496.

Dorfman, D.D., Berbaum, K.S. and Lenth, R. (1995). Multireader, multicase receiver operating characteristic methodology: A bootstrap analysis. *Academic Radiology* **2**, 626-633.

Dreiseitl, S., Ohno-Machado, L. and Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making* **20**, 323-331.

Drury, C.G. and Fox, J.G. (1975). Human reliability in quality control. *Halsted, New York.*

Efron, B. and Tibshirani, R.J. (1993). An introduction to the bootstrap. *Chapman and Hall Press, New york.*

Egan, J.P. (1975). Signal detection theory and ROC analysis. *Academic Press, New York.*

Enrique, F.S, David, F. and Benjamin, R.(2004). Adjusting the generalized ROC curve for covariates. *Statistics in Medicine* **23** 3319-3331.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286**, 531–537.

Green, D.M. and Swets, J.A. (1966). Signal detection theory and psychophysics. *Wiley, New York.*

Greenhouse, S.W. and Mantel, N. (1950). The evaluation of diagnostic tests. *Biometrics* **6**, 399-412.

Guyatt, G.H., Tugwell, P.X., Feeny, D.H., Haynes, R.B. and Drummond, M. (1986). A framework for clinical evaluation of diagnostic technologies. *Canadian Medical Association Journal* **134**, 587-594.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389–422.

Hajian-Tilaki, K.O., Hanley, J.A., Joseph, L. and Collet, J.P. (1997). Extension of receiver operating characteristic analysis to data concerning multiple signal detection tasks. *Academic Radiology* **4**, 222-229.

Han, A.K. (1987). Non-parametric analysis of a generalized regression model:the maximum rank correlation estimator. *Journal of Econometrics* **35** 303-316.

Hanley, J.A. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Critical Reviews in Diagnostic Imaging* **29**, 307-335.

Hanley, J.A. (1996). The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine* **15**, 1575-1585.

Hanley, J.A. and Hajian, K.O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves. *Academic Radiology* **4**, 49-58.

Hanley, J.A. and McNeil, B.J. (1982). The meaning and use of the area under an ROC curve. *Radiology* **143**, 29-36.

Heckerling, P.S. (2001). Parametric three-way receiver operating characteristic surface analysis using Mathematica. *Medical Decision Making* **20**, 409-417.

Henkelman, R.M., Kay, I. and Bronskill, M.J. (1990). Receiver operator characteristic (ROC) analysis without truth. *Medical Decision Marking* **10**, 24-29.

Hsieh, F. and Turnbull, B.W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics* **24**, 25-40.

Hui, S.L. and Zhou, X.H. (1998). Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* **7**, 354-370.

Jason, A. (1999). Computation of the maximum rank correlation estimator. *Economics Letters* **62**, 279-285.

Jiang, Y., Metz, C.E. and Nishikawa, R.M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* **201**, 745-750.

John Q. and Jun S.L.(1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350-1355.

Krenn, V., Morawietz, L., Burmester, G.R., Kinne, R.W. et al. (2006). Synovitis score: discrimination between chronic low-grade and high-grade synovitis. *Histopathology* **49**, 358-364.

Kruskal, W.H. and Wallis, W.A. (1952). The use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47**, 583-621.

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

Leisenring, W. and Pepe, M.S. (1998). Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. *Biometrics* **54**, 444-452.

Li, F. and Yang, Y. (2005). Analysis of recursive gene selection approaches from microarray data. *Bioinformatics* **21**, 3741–3747.

Li, J., Fine, J.P., Safdar, N. (2007). Prevalence dependent diagnostic accuracy measures. *Statistics in Medicine* **26**, 3258-3273.

Li, J. and Fine, J.P. (2008). ROC analysis for multiple classes and multiple categories and its application in microarray study. *Biostatistics* **9**, 566–576.

Li, J., Zhou, X.H. (2009). Nonparametric and Semiparametric Estimation of the Three Way Receiver Operating Characteristic Surface. *Journal of Statistical Planning and Inference*. **139**, 4133–4142.

Lusted, L.B. (1971). Signal detectability and medical decision making. *Science* **171**, 1217-1219.

Mantel, N. (1951). Evaluation of a class of diagnostic tests. *Biometrics* **7**, 240-246.

McClish, D.K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**, 190-195.

McCullagh, P. and Nelder, J.A. (1999). Generalized linear models. *Chapman and Hall, London*.

McIntosh, M. and Pepe, M.S. (2002). Combing several screening tests: Optimality of the risk score. *Biometrics* **58**, 657-664.

Metz, C.E. (1989). Some practical issues of experimental design and data analysis in radiologic ROC studies. *Investigative Radiology* **24**, 234-245.

Metz, C.E., Herman, B.A. and Shen, J.H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Statistics in Medicine* **17**, 1033-1053.

Mossman, D. (1995). Resampling techniques in the analysis of non-binormal ROC data. *Medincine Decision Making* **15**, 358-366.

Mossan, D.(1999). Three-way ROCs. *Medical Decision Making* **19**, 78-89.

Nakas, C.T., Yiannoutsos, C.T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine* **23**, 3437-3449.

Obuchowski, N.A., Graham, R.J., Baker, M.E. and Powell, K.A. (2001). Ten criteria for effective screening: Their application to multislice CT screening for pulmonary and colorectal cancers. *American Journal of Roentgenology* **176**, 1357-1362.

Obuchowski, N.A. (2005). Estimating and comparing diagnostic tests accuracy when the gold standard is not binary. *Academic Radiology* **12**, 1198-1204.

Parodi, S., Pistoia, V. and Muselli, M. (2008). Not proper ROC curves as new tool for the analysis of differentially expressed genes in microarray experiments. *BMC*

Bioinformatics **9**, 410-412.

Pepe, M.S. (1997). Aregression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika* **84**, 595-608.

Pepe, M.S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* **54**, 124-135.

Pepe, M.S. and Alonzo, T.A. (2001). comparing disease screening tests when true disease status is ascertained only for screen positives. *Biostatistics* **2**, 249-260.

Pepe, M.S. and Thompson, M.L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123-140.

Pepe, M.S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. *Oxford University Press Oxford*.

Pepe, M.S., Longton, G., Anderson, G.L. and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**, 133-142.

Peterson, W.W., Birdsall, T.G. and Fox, W.C. (1954). The theory of signal detection theory. *Transactions of the IRE Professional Group on Information Theory* **1**, 171-212.

Reiser, B. and Faraggi, D. (1997). Confidence intervals for the generalized ROC criterion. *Biometrics* **53**, 644-652.

Ressom, H.W., Varghese, R.S., Drake, S.K., Hortin, G.L., Abdel-Hamid, M., Loffredo, C.A. and Goldman, R. (2007). Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* **23**, 619–626.

Ressom, H.W., Varghese, R.S., Goldman, L., Loffredo, C.A., Abdel-Hamid, M., Kyselova, Z., Mechref, Y., Novotny, M. and Goldman, R. (2008). Analysis of MALDI-TOF mass spectrometry data for detection of glycan biomarkers. *Pacific Symposium on Biocomputing* **13**, 216–227.

Robins, J.M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122-129.

Scurfield, D.K(1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology* **40**, 253-269.

Serfling, R.J., (1980). Approximation Theory of Mathematical Statistics. *Wiley, New York*.

Shapiro, D.E. (1999). The interpretation of diagnostic tests. *Statistical Methods in Medical Research* **8**, 113-134.

Sherman, R.P (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61**, 123-137.

Sherman, R.P (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Annals of Statistics* **22**, 439-459.

Song, H.H. (1997). Analysis of correlated ROC areas in diagnostic testing. *Biometrics* **53**, 370-382.

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

Swets, J.A. (1977). Vigilance: Relationships among theory, physiological correlates and operational performance. *Plenum, New York.*

Swets, J.A. and Pickett, R.M. (1982). Evaluation of diagnostic systems: Mehods form signal detection theory. *Academic Press, New York.*

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**, 1285-1293.

Thibodeau, L.A. (1981). Evaluating diagnostic tests. *Biometrics* **37**, 801-804.

Thompson, M.L. and Zucchini, W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine* **8**, 1277-1290.

Tong, Y.L. (1990). The Multivariate Normal Distribution. *Springer, New York.*

Tosteson, A.A. and Begg, C.B. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making* **8**, 204-215.

Venkatraman, E.S. and Begg, C.B. (1996). A distribuion-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* **83**, 835-848.

Wang, H. (2006). A note on iterative marginal optimization: a simple algorithm for maximum rank correlation estimation. *Computational Statistics and Data Analysis* **51**, 2803-2812.

Weinstein, M.C. and Fineberg, H.V. (1980). Clinical decision analysis. *Saunders, Philadelphia*.

Wilson, J.M and Jungner, Y.G (1968). Principles and practice of screening for disease. *Public Health Papers 34, Switzerland*.

Zhou, X.H., McClish, D.K. and Obuchowski, N.A. (2002). Statistical methods in diagnostic medicine. *Wiley, New York*.

Zou, K.H., Hall, W.J. and Shapiro, D.E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistic in Medicine* **16**, 2143-2156.

Zou, K.H, Resnic, F.S., Talos, I.F., Goldberg-Zimring, D., Bhagwat, J.G., Haker, S.J., Kikinis, R., Jolesz, F.A. and Ohno-Machado, L. (2005). A global goodness-of-fit test for receiver operating characteristic curve analysis via the bootstrap method. *Biomedicine Informatics* **38**, 395-403

Zweig. M.H. and Campbell, G. (1993). Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**, 561-577.