2012

Making diagnoses with multiple tests under no gold standard

Jingyang Zhang
University of Iowa

MAKING DIAGNOSES WITH MULTIPLE TESTS UNDER NO GOLD STANDARD

by

Jingyang Zhang

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics
in the Graduate College of
The University of Iowa

May 2012

Thesis Supervisors: Professor Ying Zhang
Professor Kathryn Chaloner

ABSTRACT

In many applications, it is common to have multiple diagnostic tests on each subject. When there are multiple tests available, combining tests to incorporate information from various aspects in subjects may be necessary in order to obtain a better diagnostic. For continuous tests, in the presence of a gold standard, we could combine the tests linearly [59] or sequentially [64], or using some risk score as studied in [36]. The gold standard, however, is not always available in practice. This dissertation concentrates on deriving classification methods based on multiple tests in the absence of a gold standard.

Motivated by a lab data set consisting of two tests testing for an antibody in 100 blood samples, we first develop a mixture model of four bivariate normal distributions with the mixture probabilities depending on a two-stage latent structure. The proposed two-stage latent structure is based on the biological mechanism of the tests. A Bayesian classification method incorporating the available prior information is derived utilizing Bayesian decision theory. The proposed method is illustrated by the motivating example, and the properties of the estimation and the classification are described via simulation studies. Sensitivity to the choice of the prior distribution is also studied.

We also investigate a general problem of combining multiple continuous tests without any gold standard or a reference test. We thoroughly study the existing methods for combining multiple tests and develop optimal classification rules cor-

responding to the methods accommodating the situation without a gold standard. We justify the proposed methods both theoretically and numerically through extensive simulation studies and illustrate the methods with the motivating example. In the end, we conclude the thesis with remarks and some interesting open questions extended from the dissertation.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

_____
Thesis Supervisor

_____
Title and Department

_____
Date

MAKING DIAGNOSES WITH MULTIPLE TESTS UNDER NO GOLD STANDARD

by

Jingyang Zhang

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics
in the Graduate College of
The University of Iowa

May 2012

Thesis Supervisors: Professor Ying Zhang
Professor Kathryn Chaloner

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Jingyang Zhang

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biostatistics at the May 2012 graduation.

Thesis committee: _____
Ying Zhang, Thesis Supervisor

_____
Kathryn Chaloner, Thesis Supervisor

_____
Joseph Cavanaugh

_____
Joseph Lang

_____
Jack T. Stapleton

ACKNOWLEDGEMENTS

This thesis would not be possible without the guidance of my committee members, help from professors and fellow students in the department, and support from my family.

Firstly, I would like to show my gratitude to my principle advisor, Dr. Ying Zhang, for his excellent guidance, encouragement, understanding and patience during my dissertation research. His perpetual energy and enthusiasm in research have motivated me all the time. I would also like to thank my secondary advisor, Dr. Kathryn Chaloner, for her insightful advice, encouragement and financial support during my research and study at the University of Iowa. Her immense knowledge and experience have enriched my understanding of biostatistics. I would not be able to finish my thesis without their guidance and inspiration, .

Furthermore, I owe my gratitude to my committee members, Dr. Joseph Cavanaugh, Dr. Joseph Lang and Dr. Jack Stapleton, for their encouraging words, thoughtful criticism, and time and attention during busy semesters.

I am very grateful to my professors in Department of Biostatistics and Department of Statistics for showing me by examples and through challenging coursework how to think, teach and teach others to think. I also appreciate my fellow students in the department for sharing their enthusiasm and comments on my work. Special thanks goes to Ms. Terry Kirk and Ms. Ann Weber, for assisting me with the administrative tasks necessary for completing the doctoral program.

Finally, I am indebted to my parents, Jinhong Han and Xinyi Zhang, my elder sister, Binbin Zhang and my husband, Dr. Hongbo Dong for their never-ending support and love. To them, I dedicate this thesis.

ABSTRACT

In many applications, it is common to have multiple diagnostic tests on each subject. When there are multiple tests available, combining tests to incorporate information from various aspects in subjects may be necessary in order to obtain a better diagnostic. For continuous tests, in the presence of a gold standard, we could combine the tests linearly [59] or sequentially [64], or using some risk score as studied in [36]. The gold standard, however, is not always available in practice. This dissertation concentrates on deriving classification methods based on multiple tests in the absence of a gold standard.

Motivated by a lab data set consisting of two tests testing for an antibody in 100 blood samples, we first develop a mixture model of four bivariate normal distributions with the mixture probabilities depending on a two-stage latent structure. The proposed two-stage latent structure is based on the biological mechanism of the tests. A Bayesian classification method incorporating the available prior information is derived utilizing Bayesian decision theory. The proposed method is illustrated by the motivating example, and the properties of the estimation and the classification are described via simulation studies. Sensitivity to the choice of the prior distribution is also studied.

We also investigate a general problem of combining multiple continuous tests without any gold standard or a reference test. We thoroughly study the existing methods for combining multiple tests and develop optimal classification rules cor-

responding to the methods accommodating the situation without a gold standard. We justify the proposed methods both theoretically and numerically through extensive simulation studies and illustrate the methods with the motivating example. In the end, we conclude the thesis with remarks and some interesting open questions extended from the dissertation.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

1.1 Literature Review

1.1.1 Diagnostic Tests

A diagnostic test is a medical test that aids to make the diagnosis or detection of a certain case. Here the case is a general term, which could either mean a disease in the common sense, or a specific symptom, like the existence of some antibody in blood sample. The outcome of a diagnostic test, depending on the mechanism of the test, can be qualitative or quantitative. The result of a qualitative test comes from two possible responses, hence is easy to interpret. The result of a quantitative test is considered to be a random variable with values distributed on a continuous or an ordinal scale. For a quantitative test, a fixed cut-off value c is often facilitated and a subject is classified as "case" if the test result is greater than c, otherwise is classified as "non-case" or "control". Therefore, the classifications from the same quantitative test may differ due to different thresholds.

The study subjects of a diagnostic test can be selected in different ways. Subjects can be selected from a *case-control* study, where a fixed number of case subjects and a fixed number of non-case subjects (controls) are selected. Alternatively, subjects can come from the population of interest and the true case status is determined by a *gold standard test*. Such study is usually referred to as a *cohort* study.

1.1.2 Evaluating Diagnostic Accuracy of Binary Tests

In medical applications, an accurate diagnosis is the first step to study the disease. Therefore, it is important to evaluate the accuracy of a diagnostic test. There are various ways to define the accuracy depending on different types of tests. For qualitative tests which yield a binary result, true positive rate (TPR) and false positive rate (FPR) are the two classification probabilities commonly used in practice [47, 81]. TPR is defined as the probability that a case subject has a positive test result and FPR is defined as the probability that a control subject has a positive test result. Let the binary variable D denote the true condition as:

$$D = \begin{cases} 1, & \text{case}, \\ 0, & \text{control}; \end{cases}$$

and let the variable Y denote the test result. Here Y is binary:

$$Y = \begin{cases} 1, & \text{positive}, \\ 0, & \text{negative}. \end{cases}$$

So

$$\text{TPR} = \Pr\left(Y = 1 | D = 1\right), \tag{1.1}$$

$$\text{FPR} = \Pr\left(Y = 1 | D = 0\right). \tag{1.2}$$

$(1 - \text{TPR})$ is accordingly equal to the false negative rate (FNR), hence the pair (FPR, FNR) is usually used to represent the error probabilities for a test. In biomedical research, TPR is known as the sensitivity and $(1 - \text{FPR})$ is known as the specificity.

In addition to the classification probabilities, predictive values reflecting how well the test results predict the case status are also an option to assess the accuracy

of a test. The positive predictive value (PPV) and negative predictive value (NPV) are defined as:

$$\mathrm{PPV} = \Pr\left(D = 1 | Y = 1\right), \tag{1.3}$$

$$\mathrm{NPV} = \Pr\left(D = 0 | Y = 0\right). \tag{1.4}$$

A perfect test will have zero misclassification probabilities with TPR = 1 and FPR = 0, and predict the condition impeccably, with PPV = NPV = 1.

In practice, for a study, based on the test results and the true condition, the data can be summarized in a 2×2 contigency table (Table 1.1). The empirical estimates of FPR and TPR are the proportions of positive results for control and case subjects, respectively. Similarly, the empirical estimates of PPV and NPV are the proportions of case and control subjects for subjects with positive and negative results, respectively. Namely,

$$\widehat{\mathrm{TPR}} = \frac{n_{11}}{n_{1+}},\ \widehat{\mathrm{FPR}} = \frac{n_{21}}{n_{2+}};$$
$$\widehat{\mathrm{PPV}} = \frac{n_{11}}{n_{+1}},\ \widehat{\mathrm{NPV}} = \frac{n_{22}}{n_{+2}}.$$

Under the independent observations, the estimators are proportions in binomial distributions and the inference can be made based on the binomial distributions (see Chapter 2 of [47]).

Table 1.1: Summary of the test results in a cohort study.

	Test Result		
Disease Status	Positive	Negative	Total
True	n_{11}	n_{12}	n_{1+}
False	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

1.1.3 Evaluating Diagnostic Accuracy of Quantitative Tests

1.1.3.1 The ROC Curve

For the tests with results not simply dichotomized by positive or negative, but in an ordinal or continuous scale, a binary test can be defined based on a pre-specified threshold c, as

$$\begin{cases} \text{positive if} & Y \geq c, \\ \text{negative if} & Y < c. \end{cases}$$

Hence the TPR and FPR are functions of the threshold value c:

$$\text{TPR}(c) = \Pr\left(Y \geq c | D = 1\right), \tag{1.5}$$

$$\text{FPR}(c) = \Pr\left(Y \geq c | D = 0\right). \tag{1.6}$$

For c ranging over all possible values, the pairs (FPR(c), TPR(c)) form a curve, the receiver operating characteristics (ROC) curve [20, 47, 60, 81], and it is the most commonly used statistical tool for evaluating quantitative tests [20, 44, 47, 81, 82]. The ROC curve was first developed by engineers working in the signal detection theory to detect enemy objects in battle fields [20], then was gradually introduced to psychology [20], medicine [47, 81, 82], radiology [44], and other areas. From the

definition, the ROC curve can be expressed as

$$\mathrm{ROC}(\cdot) = \{(\mathrm{FPR}(c), \mathrm{TPR}(c)), c \in \mathbb{R}\} \tag{1.7}$$

It is easy to see that both FPR and TPR are a monotone decreasing function of c, with $\lim_{c\to\infty} \mathrm{FPR}(c) = \lim_{c\to\infty} \mathrm{TPR}(c) = 0$ and $\lim_{c\to-\infty} \mathrm{FPR}(c) = \lim_{c\to-\infty} \mathrm{TPR}(c) = 1$. Therefore, the ROC curve is a monotone increasing curve that lies in the first quadrant:

$$\mathrm{ROC}(t) = \left\{\left(t, \mathrm{TPR}\left(\mathrm{FPR}^{-1}(t)\right)\right),\ t \in (0,1)\right\}, \tag{1.8}$$

where $c = \mathrm{FPR}^{-1}(t)$ is the threshold values such that $t = \mathrm{FPR}(c)$. It is invariant to any monotone transformation on Y according to (1.7) and (1.8).

An uninformative test has an ROC curve as the diagonal line of the first quadrant, given by TPR = FPR for any threshold c. In that case, the probability distribution of Y does not vary between the case and control populations, thus Y is unrelated to the case status. The perfect test, on the other hand, has the ROC curve along the left and upper border of the first unit quadrant, i.e., TPR = 1 and FPR = 0 for any c. Most tests have ROC curves between those two extreme ROC curves as shown in Figure 1.1.

For a continuous test, denote $S_1(c) = \Pr(Y \geq c|D = 1)$ and $S_0(c) = \Pr(Y \geq c|D = 0)$, then the ROC curve can be easily expressed in a function form as

$$\mathrm{ROC}(t) = S_1\left(S_0^{-1}(t)\right),\ t \in (0,1) \text{ (Result 4.2 of [47])}. \tag{1.9}$$

If Y is distributed as a normal distribution in both populations, then the ROC curve

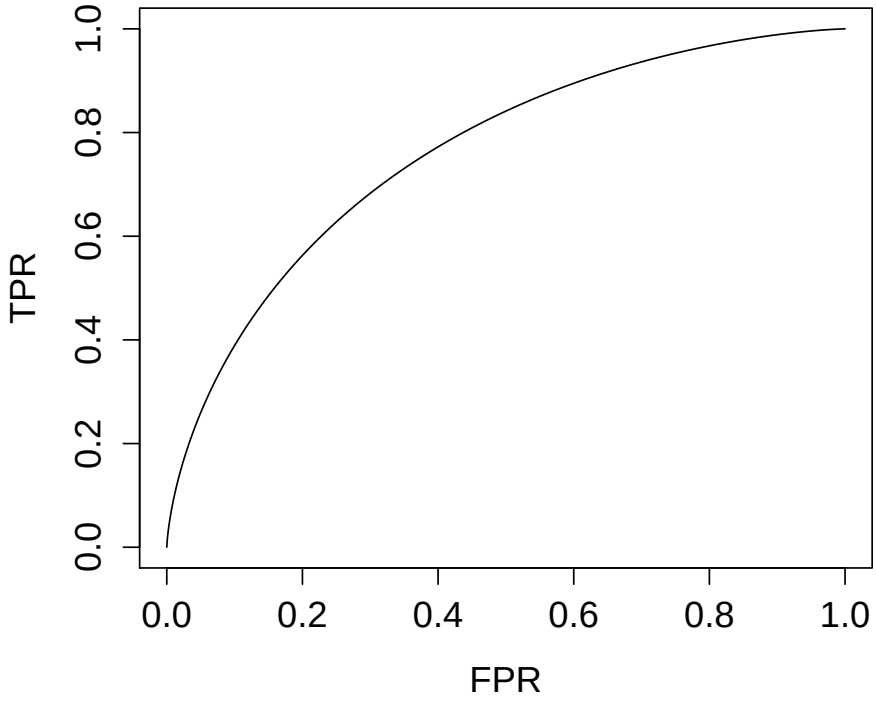

Figure 1.1: An example of an ROC curve.

is called the binormal ROC curve. Particularly, as stated in Result 4.7 of [47], if

$$Y|D=1 \sim N(\mu_1, \sigma_1^2),\ Y|D=0 \sim N(\mu_0, \sigma_0^2)$$

then

$$\mathrm{ROC}(t) = \Phi\left(a + b\Phi^{-1}(t)\right), \tag{1.10}$$

where

$$a = \frac{\mu_1 - \mu_0}{\sigma_1},\ b = \frac{\sigma_0}{\sigma_1}$$

According to Result 4.9 of [47], there exists some monotone transformation of Y such that the distributions of the transformed test results are normal in both $D = 1$ and $D = 0$ populations. Hence, the data are not necessarily normal to hold a binormal ROC curve. The binormal ROC curve is a common function form of the ROC curves.

For an ordinal test, since the test result can only be a discrete number of values, the ROC curve could still be the function form of (1.8), but in a discrete form as well,

$$\mathrm{ROC} = \left\{\left(t_y, \mathrm{TPR}\left(\mathrm{FPR}^{-1}(t_y)\right)\right), y = 1, \cdots, P+1\right\}, \tag{1.11}$$

where $c_y = \mathrm{FPR}^{-1}(t_y)$ such that $t_y = \mathrm{FPR}(c_y)$ (Note that $c_1 = -\infty$ and $c_{P+1} = \infty$.). The parametric model, for example the binormal model (1.10), can also be adopted for the discrete ROC functions. In that case, the ROC points fall upon a continuous curve. The interpretation of the discrete ROC function here is different from that of the continuous ROC function. The discrete ROC curves are only interpretable at the points of the domain. Even though the ROC points of two discrete ROC functions lie

on the same continuous curve, the two ROC functions differ because their domains differ. Section 4.5.5 of [47] has more discussion on the discrete ROC curve.

1.1.3.2 The Area Under the ROC Curve

To summarize the accuracy of a quantitative test using a number, one could use the area under the ROC curve (AUC) as an index. It is defined as

$$\text{AUC} = \int_0^1 \text{ROC}(t)dt. \tag{1.12}$$

The range of AUC is between 0 and 1. For a perfect test, AUC= 1 since $\text{ROC}(t) = 1$ for any $t \in (0, 1)$, and for an uninformative test, AUC= 0.5 since $\text{ROC}(t) = t$ for any $t \in (0, 1)$.

AUC can be interpreted as the probability that the test result from a case subject is greater or equal to the result from a control subject, namely, $\Pr(Y_D \geq Y_{\bar{D}})$ for continuous tests and $\Pr(Y_D > Y_{\bar{D}}) + \frac{1}{2}\Pr(Y_D = Y_{\bar{D}})$ for ordinal tests [4, 22], where Y_D and $Y_{\bar{D}}$ are random variables representing test results from case and control populations, respectively. The proof for continuous tests is Result 4.6 and 4.10 of [47].

As shown in Result 4.8 of [47], the AUC for the binormal ROC curve has an explicit function form of (1.13).

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right), \tag{1.13}$$

where a and b are defined as in (1.9).

Sometimes, AUC can be misleading as a measure of accuracy [47, 81], especially when comparing two tests with crossing ROC curves like shown in Figure 1.2.

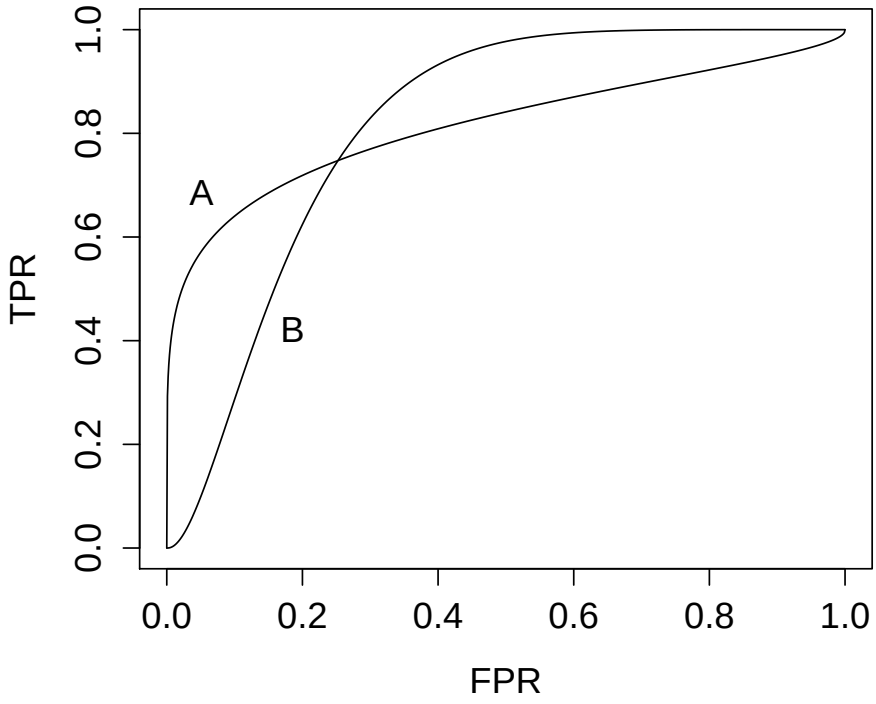

Figure 1.2: Two tests with crossing ROC curves.

In Figure 1.2, two ROC curves cross at FPR around 0.25 with the same AUC (0.814). If we use AUC as the measure, there is no difference in the diagnostic accuracy, but if the primary clinical interest is in the region of low FPRs, Test A is apparently preferred to Test B. This also implies that other alternative summary indices are necessary.

1.1.3.3 Other Summary Indices

One alternative summary index is the sensitivity at a fixed FPR. The test with a higher value of sensitivity estimated from the ROC curve at the fixed common FPR is preferable. Using the two tests shown in Figure 1.2 as an illustration, at FPR = 0.1, the TPR, or the sensitivity, is 0.64 and 0.29 for Test A and B, respectively. This indicates that 64% of the case subjects have a positive result from Test A given the threshold corresponding to the FPR= 0.1, and similarly, 29% of the cases have a positive result from Test B. From this point of view, Test A is preferred to Test B. This measure, however, also has some drawbacks as thoroughly discussed in [47, 81].

Another summary measure of diagnostic accuracy is the partial area under the ROC curve (pAUC). It is defined as the area under the ROC curve within a given region of FPR $(0, t_0)$:

$$\text{pAUC}(t_0) = \int_0^{t_0} \text{ROC}(t)dt, \tag{1.14}$$

and it can be interpreted as the average sensitivity over the relevant range of FPR, or equivalently, specificity.

1.1.3.4 Estimating the ROC Curve

Suppose the data consist of n_1 case subjects and n_0 control subjects

$$\{Y_{1i}, i = 1, \ldots, n_1\} \text{ and } \{Y_{0j}, j = 1, \ldots, n_0\}.$$

$S_1(y) = \Pr(Y_{1i} \geq y)$ and $S_0(y) = \Pr(Y_{0j} \geq y)$ stand for the survival functions of $\{Y_{1i}, i = 1, \ldots, n_1\}$ and $\{Y_{0j}, j = 1, \ldots, n_0\}$, respectively.

By (1.9), the ROC function can be expressed as $S_1\left(S_0^{-1}(t)\right)$, so intuitively, the empirical estimate of the ROC curve, denoted by $\widehat{\mathrm{ROC}}_e$, is derived by plugging the empirical survival functions for Y_1 and Y_0 in (1.9).

$$\widehat{\mathrm{ROC}}_e(t) = \hat{S}_1\left(\hat{S}_0^{-1}(t)\right). \tag{1.15}$$

The AUC under the empirical ROC curve is just the Mann-Whitney U-statistic [47]

$$\widehat{\mathrm{AUC}}_e = \sum_{j=1}^{n_0}\sum_{i=1}^{n_1}\left\{\mathrm{I}[Y_{1i} > Y_{0j}] + \frac{1}{2}\mathrm{I}[Y_{1i} = Y_{0j}]\right\} \Big/ n_1 n_0. \tag{1.16}$$

The statistical inferences of the empirical ROC and AUC are discussed in Section 5.2 of [47].

If the distributions of Y_1 and Y_0 are known and determined by parameters α_1 and α_0, respectively, one could estimate the parameters α_1 and α_0 using maximum likelihood (ML) method, which immediately results in ML estimate (MLE) of ROC by

$$\widehat{\mathrm{ROC}}_{\hat{\alpha}_1,\hat{\alpha}_0}(t) = \hat{S}_{1,\hat{\alpha}_1}\left(\hat{S}_{0,\hat{\alpha}_0}^{-1}(t)\right). \tag{1.17}$$

As a specific example, for $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_0 \sim N(\mu_0, \sigma_0^2)$, and ($\hat{\mu}_1$, $\hat{\mu}_0$, $\hat{\sigma}_1^2, \hat{\sigma}_0^2$) are the MLEs of $(\mu_1, \mu_0, \sigma_1^2, \sigma_1^2)$ for a given random sample, then according to (1.10) and (1.13), the ROC curve and AUC are estimated by

$$\widehat{\mathrm{ROC}}(t) = \Phi\left(\frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}_1} + \left(\frac{\hat{\sigma}_0}{\hat{\sigma}_1}\right)\Phi^{-1}(t)\right), \tag{1.18}$$

$$\widehat{\mathrm{AUC}} = \Phi\left(\frac{\hat{\mu}_1 - \hat{\mu}_0}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_0^2}}\right). \tag{1.19}$$

The standard errors of $\widehat{\mathrm{ROC}}(t)$ and $\widehat{\mathrm{AUC}}$ can be calculated by the Delta method.

In Chapter 5 of [47], the ROC-GLM methodology is introduced. It parameterizes the ROC curve without any distributional assumptions on the data. This approach was first proposed in [45, 46] and has been discussed and extended by several articles [7, 8, 9, 48].

1.1.4 Combining Multiple Tests

Diagnostic tests are usually imperfect. When there are multiple tests available, a composite test incorporating information from various tests may yield a better diagnostic accuracy than any single test. There are numerous examples of combining multiple tests. For example, there are several biomarkers available to screen one certain cancer. Each biomarker alone may not be sufficiently sensitive and specific to detect the cancer. Doctors usually need to look at more than one biomarkers to make a diagnosis. Therefore, it becomes important to find the optimal combination of multiple tests.

For the sake of simplicity, we consider combining two tests throughout this dissertation.

1.1.4.1 Combining Binary Tests

Two binary tests can be combined by the "believe the positive" (BP) rule or the "believe the negative" (BN) rule [35, 51] defined as follows.

Definition 1.1. *For two binary tests:*

- *The BP rule means that the classification is positive if either test is positive.*
- *The BN rule means that the classification is positive if both tests are positive.*

According to Result 9.2 of [47], the BP rule increases the TPR relative to the component tests, as well as the FPR by no more than the sum of the FPRs of the two tests, therefore this rule is preferred when there is low FPRs but inadequate TPRs. The BN rule decreases the FPR and TPR, but the TPR is maintained above the difference between the sum of the TPRs of the two tests and 1. So the BN rule should be considered when both tests have high TPR but also have too high FPR. The BN rule helps to reduce the FPR without losing too much TPR.

A special case of multiple tests is the repeated applications of one certain test. Politser [51] discussed the importance of repeating some unreliable tests, and besides the simple BP and BN rules, he also proposed a "formal rule" that sequentially integrates all the information of the repeated tests based on the likelihood ratio of the results. Lau [32] considered a composite test consisting of a series of Bernoulli trials. Each Bernoulli trial is to apply one given screening test on a subject and this trial is repeated n times for the given screening test on the same subject. If k ($k \leq n$) individual tests are positive, the composite test is positive; otherwise, the composite test is negative. In this paper, Lau studied the dependence of the results from two trials and its effect on the diagnostic accuracy of the composite test. The value of n and k can be chosen in order to achieve the anticipated level of sensitivity and specificity. Ten Have and Bixler [62] proposed a model-based approach to model the population heterogeneity in measures of diagnostic accuracy, sensitivity and specificity, of a multi-stage screening test. In the multi-stage screening procedure, subjects can dropout early if they have a positive test result, but the method needs to know the

definitive diagnosis on such dropout subjects in order to use their information to calculate sensitivity and specificity.

When there are more than two binary tests to be combined, the BP and BN rules could be extended, but more complicated. For example, if there are n tests, then there are 2^n possible combinations and an enormous number of rules can be made based on the 2^n possibilities. For a more general question of combining multiple binary tests, Baker [3] introduced the likelihood function as a combination of multiple tests and he applied the methodology to multiple binary tests by approximating FPR and TPR non-parametrically.

1.1.4.2 Combining Continuous Tests

Similar to the repeated measures on one binary test, Tolley et al. [67] and Murtaugh [40] considered the scenario of repeated applications of the same continuous test. At each test application, the cutoff remains the same. For a set of different tests, Su and Liu [59] justified that the linear discriminant function is the optimal linear combination that produces the maximum AUC when the test results are multivariate normally distributed in both case and control groups. Pepe and Thompson [49] relaxed the distributional assumption by maximizing the distribution-free rank-based estimate of AUC. Thompson [64] considered the combination of a sequence of tests. The sequence of tests can be the repeated applications of the same test on the same subject, or different tests simultaneously. The sequential rule does not limit to the linear combinations, nor does it require the application of all tests on each subject.

McIntosh and Pepe [36] generalized the likelihood ratio methodology to continuous tests as the optimal risk score method. In the framework of hypothesis testing, the risk score, which is the scalar function of all the n tests results $\mathbf{Y} = (Y_1, \ldots, Y_n)$, namely

$$\mathrm{LR}(\mathbf{Y}) = \frac{\Pr(\mathbf{Y}|D=1)}{\Pr(\mathbf{Y}|D=0)},$$

yields an optimal decision rule:

$$\begin{cases} \text{Positive} & \text{if } LR(\mathbf{Y}) > c; \\ \text{Negative} & \text{otherwise,} \end{cases}$$

where $\Pr(\mathbf{Y}|D=1)$ and $\Pr(\mathbf{Y}|D=0)$ are the probability distribution functions of $\mathbf{Y}$ under $D=1$ and $D=0$. This rule maximizes the TPR at any fixed level of FPR by the Neyman-Pearson lemma [42] as shown in Result 9.3 of [47]. The threshold c is usually chosen depending on the choice of FPR.

In view of practice, the risk score $LR(\mathbf{Y})$ is not easy to utilize if there is no knowledge about the distribution of $\mathbf{Y}$. Alternatively, they suggested using an equivalent risk score to combine multiple tests as:

$$\begin{aligned} \mathrm{RS}(\mathbf{Y}) &= \Pr(D=1|\mathbf{Y}) \\ &= \frac{\Pr(\mathbf{Y}|D=1)\Pr(D=1)}{\Pr(\mathbf{Y})} \\ &= \frac{\Pr(\mathbf{Y}|D=1)\Pr(D=1)}{\Pr(\mathbf{Y}|D=1)\Pr(D=1) + \Pr(\mathbf{Y}|D=0)\Pr(D=0)} \\ &= \frac{\mathrm{LR}(\mathbf{Y})\Pr(D=1)}{\mathrm{LR}(\mathbf{Y})\Pr(D=1) + \Pr(D=0)}, \end{aligned}$$

which is a monotone increasing function of $\mathrm{LR}(\mathbf{Y})$. This alternative risk score can be

estimated by a logistic regression model

$$logit\left\{\Pr\left(D=1|\mathbf{Y}\right)\right\}=\alpha+h(\mathbf{Y},\beta),$$

where h is a parametric function of $\mathbf{Y}$ with parameters β.

1.1.5 Lack of a Gold Standard

Estimating the aforementioned measures of diagnostic accuracy requires a knowledge of true case status, or a gold standard from a perfect test, which has zero error rates. However, in practice, such information is not always available, because it may be difficult or even impossible to determine the true status, and even the available reference test against which new tests are compared is subject to errors. There are numerous examples of imperfect gold standard tests. For example, the diagnosis of Alzheimer's disease is made based on certain symptoms, but the diagnosis is not definitive until the brain tissue has been examined after death. Another example would be diagnosing the myocarditis. Myocarditis is hard to diagnose because it resembles many other diseases. The definitive diagnostic test is the heart biopsy, which is also subject to laboratory and other errors. Kraemer [29] brought up the opinion that the true case status is almost never ascertained. The absence of a perfect reference test adds more complexity to the evaluation of new tests.

If one considers the reference test as perfect whether or not it is true, the perceived accuracy of new tests may be seriously biased due to ignoring the error in the reference test [58, 69]. If the error rates of the reference test are known, then the accuracy of the new test could still be well estimated [18, 58, 63]. Hui and

Walter [26] generalized the problem to the case where the error rates of both the new test and the reference test are unknown. By applying two tests simultaneously to individuals from two populations with different prevalence of case, and further assuming the conditional independence, the sensitivity and specificity of both tests, together with the true prevalence in two populations, could be estimated by the ML method. The Hui-Walter model has been extensively discussed and extended since it was proposed. Vacek [68] discussed the impact of the conditional independence on the estimates of the error rates in the model, and Walter an Irwig [72] provided a thorough discussion of the method in different settings. Joseph et al. [28] developed Bayesian methods for the evaluation and implementation of the conditional independent tests. Hui and Zhou [27] summarized many available methods for qualitative diagnostic test evaluation, with special focus on estimating sensitivity and specificity without assuming the conditional independence. The Hui-Walter model and its extensions have also been applied to animal research as discussed by Enøe et al. [17]. The conditional dependence is accommodated via either maximum likelihood approach or Bayesian approach, for example, by Qu et al. [52], Yang and Becker [77], Dendukuri and Joseph [14], Black and Craig [5]. All of these methods are applicable only to imperfect binary tests, and cannot be directly utilized for the continuous tests.

For ordinal and continuous-scale tests, the sensitivity and specificity are computed based on a certain classification rule with a specific threshold value. The two accuracy indices depend on the choice of the classification rule. Therefore, when the true case status is unknown and there is no gold standard or even an imperfect binary

reference test, it is necessary to reconcile the results of multiple imperfect tests to establish a classification [47]. Nielsen et al. [43] proposed to estimate the sensitivity and specificity pointwise over the whole range of cutoff values by the ML method of the Hui-Walter model. But the estimated ROC curve by connecting all the estimated (sensitivity, 1-specificity) is not necessarily monotone. Henkelman et al. [24] used a mixture of multivariate normal latent model to estimate the ROC curve for ordinal-scale tests, and Choi et al. [10] adopted the same parametric model and used the Bayesian method to estimate the ROC curve for continuous-scale tests. Both estimated curves are guaranteed to be monotone increasing. To relax the normality assumption, Hall and Zhou [21] proposed a nonparametric estimator for the ROC curves of continuous tests based on the conditional independence assumption. Zhou et al. [80] applied this estimator to estimate the ROC curves for ordinal tests in the absence of a gold standard. If there is one imperfect binary test available, the ROC curve of the new continuous test can be estimated by a Bayesian approach using the binary test [73]. This approach assures the monotonicity of the ROC curve without any assumptions regarding to the distributions of the test results.

The methods above primarily focus on the evaluation of diagnostic tests when there is no definitive diagnosis or a gold standard. For binary tests, Alonzo and Pepe [1] proposed a composite reference standard to assess the accuracy of the new test. The composite standard is a combination of several reference tests. The aforementioned methods of combining multiple continuous tests may be extended under some parametric distributional assumptions. For example, Su-Liu's linear discrimi-

nant method is still applicable with the parameters in the normal distributions estimated through the ML method using the EM algorithm [13]. Yu et al. [78] applied the Su-Liu's linear combination method to combining multiple measures of chronic kidney disease. They proposed a Bayesian latent disease model incorporating other covariates, like age and gender, and the estimation is done using the Markov Chain Monte-Carlo (MCMC) algorithm.

1.2 Content

The remaining of the thesis is organized as follows. A motivating example is introduced in Chapter 2. In Chapter 3, we develop a screening rule based on this specific motivating example using Bayesian statistical decision theory. The Bayesian screening rule is applied to the example and evaluated via extensive simulation studies. The general problem that making diagnoses on multiple continuous tests without any gold standard or reference tests is explored in Chapter 4. The extension of existing methods and the classification rules based on those methods are presented with their numerical implementations and theoretical justifications. We also compare those classification methods by a series of simulation studies, together with the illustration with the motivating example. Some concluding remarks are provided in Chapter 5 and the computer programs are included in the Appendix.

CHAPTER 2
MOTIVATING EXAMPLE

2.1 GBV-C and E2 antibody

GB Virus C (GBV-C), formerly named as hepatitis G virus (HGV), is a human RNA virus, not currently known to definitely cause any disease although a recent observational study suggested a potential link between GBV-C and non-Hodgkins lymphoma [30]. There is also evidence that people with HIV disease who are co-infected with GBV-C have prolonged survival [79]. In addition, some studies found an association between GBV-C and improved response to HIV therapy [57]. The mechanism for these mechanisms is under investigation [38, 39, 76]. GBV-C infection are shown to be cleared in a great portion of patients from several months to several years [33], and antibodies develop that are directed against the viral envelope glycoprotein 2 (E2) [50]. The E2 antibodies are a marker of past GBV-C infection [15, 61]. Subjects without GBV-C who have GBV-C E2 antibodies are found to survive longer from HIV infection than those who do not have the antibodies in a study [74].

2.2 ELISA Testing Mechanisms

To detect the presence of E2 antibodies in human serum samples, there is no commercial and validated test available. One commonly used method is through the Enzyme Linked Immunosorbent Assays (ELISAs). ELISAs can be designed in several ways, but all GBV-C assays reported to date use E2 Monoclonal antibodies (MAb) which bind to the E2 protein at a specific site.

One test was developed by Roche Laboratories and is denoted the μPlate Anti-HGenv test [61]. It is a variation of a "sandwich capture assay". It uses full-length recombinant E2 protein in a Chinese Hamster Ovary (CHO) cell lysate (this contains other cellular material in addition to E2 protein). This lysate is treated with a specific murine monoclonal antibody (MAb #1) which binds to the E2 protein. MAb #1 is biotinylated and binds to the E2 in the lysate but supposedly not to the other cellular materials present. After MAb #1 is mixed with the E2, it is added to wells on a microtitre plate together with the human sample. The wells are coated with streptavidin which binds to the biotin on the MAb #1 (which has E2 protein attached). If there are GBV-C E2 antibodies in the human sample, these human antibodies (denoted as Ab #2) in Figure 2.1 will bind to the E2 protein. When the plate is subseqently washed, the E2 protein-biotinylated MAb complex remains on the plate. In some samples however, the human antibodies will not bind because they are directed against the same region on E2 recognized by MAb #1 and their access is therefore blocked; blocking may also occur because of the additional cellular material in the lysate. This blocking is the mechanism for false negatives. Anti-human IgG antibodies conjugated to an enzyme are then added to the wells, which attach to Ab#2. A colorimetric substrate for the enzyme is added afterward to allow determination of the concentration of enzyme present in the well, reflecting the amount of human anti-E2 antibody. Control wells to which no human serum is added are present on each plate to measure nonspecific material that may stick to Ab #2 and give rise to background fluorescence. The ultraviolet absorbance of color in the

wells is measured and compared to the fluorescence of the control wells.

A second test (denoted M5), was developed in the Stapleton laboratory [37], and is a more common variation on the sandwich capture assay . The end result of the test is the same, as in Figure 2.1, but the procedure to get there differs from the μPlate Anti-HGenv ELISA. A murine MAb #1 specific for E2 protein is attached to microtiter plate wells. This MAb was provided by Dr. Alfred Engel, Roche Diagnostics, Penzburg, Germany. The MAb used may be the same as the MAb used in the μPlate Anti-HGenv test; however, this information is proprietary. This antibody is not biotinylated. Semi-purified recombinant E2 protein for which the C-terminal membrane spanning domain is not included is added to wells. The plate is then washed and human serum samples applied. Human antibodies against E2 (Ab #2) will bind to the E2 protein, again unless they have the same specificity as the murine capture MAb #1. Anti-Human IgG conjugated to an enzyme is added, and the colorimetric substrate to measure human IgG uses the same methods as the μPlate Anti-HGenv assay. The result of both the μPlate Anti-HGenv and the M5 test is quantitative; however, due to differences in the capture antibody, recombinant E2 protein, the quantitative results can not be directly compared.

2.3 The Data Set

In the example, a total of 100 independent blood specimens obtained from HIV infected subjects were tested with both μPlate Anti-HGenv and M5 assays. As explained above, false negatives occur in both tests when the binding site of Ab #2 is

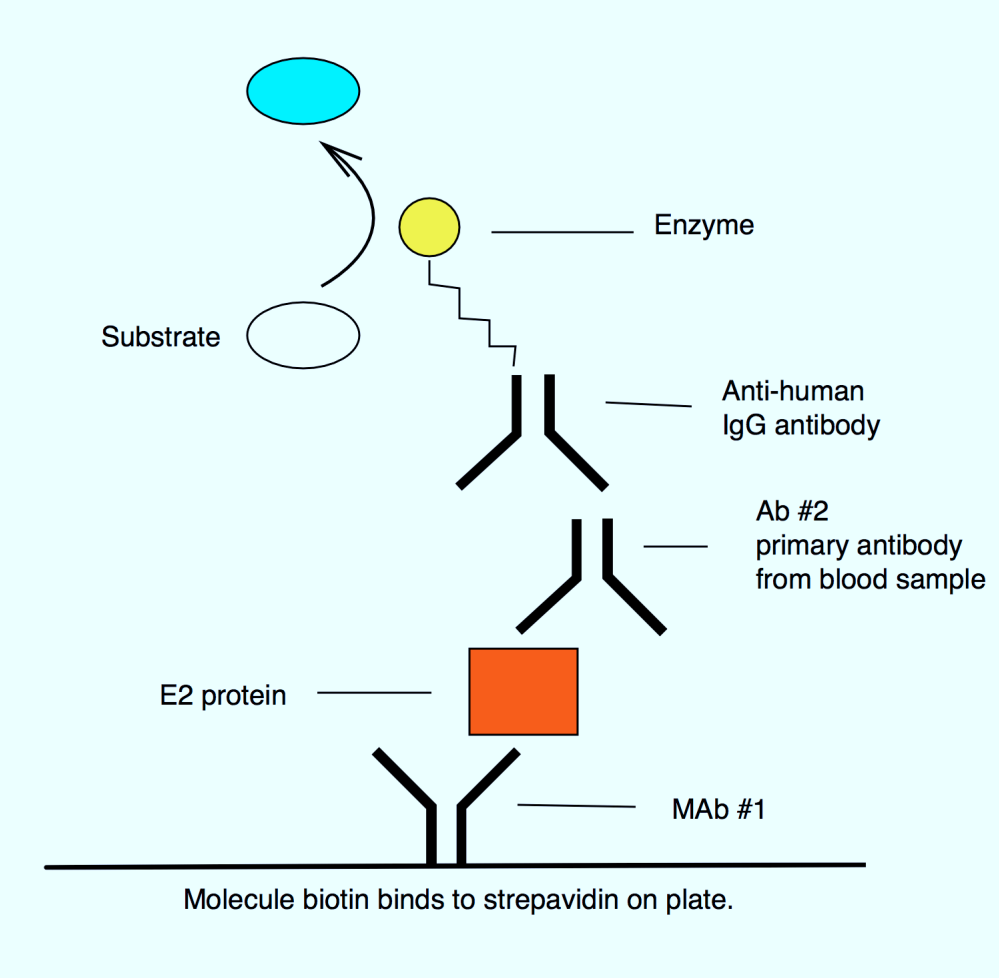

Figure 2.1: Diagram of sandwich capture ELISA test.

blocked by MAb #1. The additional material introduced in the μPlate Anti-HGenv through the lysate may add additional noise. Neither test is perfect and false negatives may occur approximately 10% of the time. The goal of this study is to develop a classification rule based on the results of the two tests.

CHAPTER 3
BAYESIAN CLASSIFICATION WITH MULTIPLE TESTS UNDER NO GOLD STANDARD

In this chapter, the main purpose is to reconcile the results from the two imperfect ELISA tests described in Chapter 2 in order to acquire a classification with a higher accuracy than each test alone. A parametric approach is adopted here. Based on the biological mechanism underlying the tests, a two-stage latent model is proposed, together with a Bayesian classification incorporating prior information on the prevalence of the condition, sensitivity and specificity of each test, and the dependence between the tests.

3.1 Notation, Assumptions and Model

For the k^{th} sample, $k = 1, 2, \ldots, n$, and the i^{th} test, $i = 1, 2$, let Y_{ik} be the observed test result. X_k, X_{ik} are binary latent variables as below:

$$
\begin{aligned}
X_k &= \begin{cases} 1 & \text{if the GBV-C E2 antibodies are present in the blood sample } k. \\ 0 & \text{if the GBV-C E2 antibodies are absent in the blood sample } k. \end{cases} \\
X_{ik} &= \begin{cases} 1 & \text{if } X_k = 1 \text{ and the binding site for test } i \text{ on sample } k \text{ is accessible.} \\ 0 & \text{if } X_k = 1 \text{ and the binding site for test } i \text{ on sample } k \text{ is blocked.} \end{cases} \\
X_{ik} &= 0, \text{ if } X_k = 0.
\end{aligned}
$$

Assume that if E2 antibodies are present ($X_k = 1$) and both tests have accessible binding sites ($X_{1k} = X_{2k} = 1$), then Y_{1k} and Y_{2k} are positively correlated. If antibodies are present but at least one binding site is inaccessible, then Y_{1k} and Y_{2k} are independent. Similarly, if there are no antibodies present, $X_k = 0$, then Y_{1k} and Y_{2k} are independent and have the identical distribution as when antibodies are

present but both binding sites are inaccessible ($X_k = 1$ and $X_{1k} = X_{2k} = 0$). The joint distribution of Y_{1k} and Y_{2k} conditioning on any combination of X_{1k} and X_{2k} is assumed to be bivariate normal. Hence Y_{1k} and Y_{2k} are jointly distributed as a mixture of four bivariate normal distributions conditioning on X_{1k} and X_{2k}, $k = 1, \ldots, n$ as shown in Figure 3.1. The four distributions are defined:

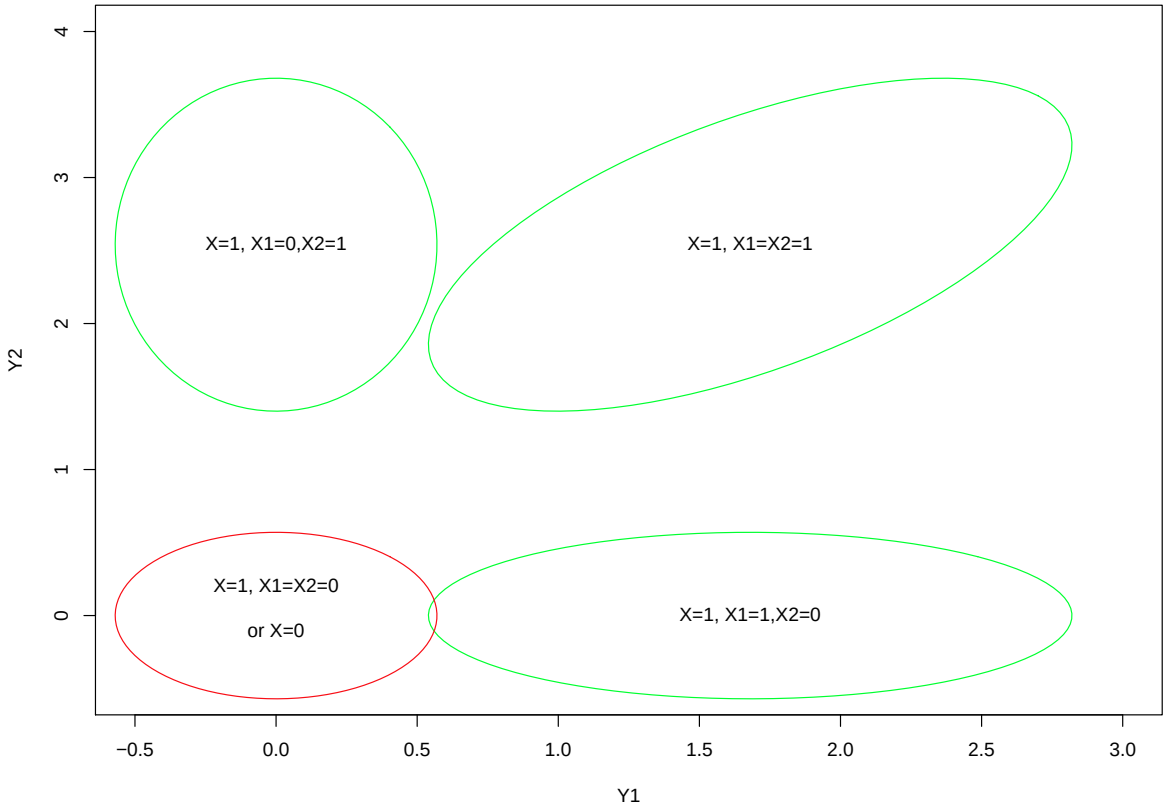

Figure 3.1: Diagram of the mixture of four bivariate normal distributions for Y_1 and Y_2.

$$
\begin{aligned}
\left(\begin{pmatrix} Y_{1k} \\ Y_{2k} \end{pmatrix} \middle| X_{1k} = X_{2k} = 1,\ X_k = 1 \right) &\sim N\left(\begin{pmatrix} \mu_{1P} \\ \mu_{2P} \end{pmatrix}, \begin{pmatrix} \sigma_{1P}^2 & \rho\sigma_{1P}\sigma_{2P} \\ \rho\sigma_{1P}\sigma_{2P} & \sigma_{2P}^2 \end{pmatrix} \right), \\
\left(\begin{pmatrix} Y_{1k} \\ Y_{2k} \end{pmatrix} \middle| X_{1k} = X_{2k} = 0 \right) &\sim N\left(\begin{pmatrix} \mu_{1N} \\ \mu_{2N} \end{pmatrix}, \begin{pmatrix} \sigma_{1N}^2 & 0 \\ 0 & \sigma_{2N}^2 \end{pmatrix} \right), \\
\left(\begin{pmatrix} Y_{1k} \\ Y_{2k} \end{pmatrix} \middle| X_{1k} = 1,\ X_{2k} = 0,\ X_k = 1 \right) &\sim N\left(\begin{pmatrix} \mu_{1P} \\ \mu_{2N} \end{pmatrix}, \begin{pmatrix} \sigma_{1P}^2 & 0 \\ 0 & \sigma_{2N}^2 \end{pmatrix} \right), \\
\left(\begin{pmatrix} Y_{1k} \\ Y_{2k} \end{pmatrix} \middle| X_{1k} = 0,\ X_{2k} = 1,\ X_k = 1 \right) &\sim N\left(\begin{pmatrix} \mu_{1N} \\ \mu_{2P} \end{pmatrix}, \begin{pmatrix} \sigma_{1N}^2 & 0 \\ 0 & \sigma_{2P}^2 \end{pmatrix} \right),
\end{aligned}
$$

where the means μ_{1N} and μ_{2N} denote the mean of Y_{1k} and Y_{2k} when either antibodies are absent (true negatives) or antibodies are present but binding site 1 or 2 respectively is inaccessible (false negatives). The means μ_{1P} and μ_{2P} denote the mean response when antibodies are present and can bind. Based on the biological mechanism, the high test result values should correspond to a higher chance of being "positive". Hence, we set a constraint that $\mu_{1P} \geq \mu_{1N}$ and $\mu_{2P} \geq \mu_{2N}$. To guarantee that the constraint holds, define new parameters $\beta_i = \log(\mu_{iP} - \mu_{iN})$ $(i = 1, 2)$. Parameters σ_{1N}^2, σ_{2N}^2, σ_{1P}^2 and σ_{2P}^2 are variances, constrained to be positive. The positive correlation between Y_{1k} and Y_{2k}, if both binding site are accessible, is denoted by ρ with $0 < \rho < 1$.

Denote the prevalence of E2 antibodies $\phi = \Pr(X_k = 1)$, and denote the probability of the binding site being accessible in test i $(i = 1, 2)$ if E2 antibodies are present as $\phi_i = \Pr(X_{ik} = 1 | X_k = 1)$. Then assuming latent variables X_{1k} and X_{2k}

are independent conditionally on $X_k = 1$, the mixture proportions are:

$$\begin{aligned}
\Pr(X_{1k} = X_{2k} = 1,\ X_k = 1) &= \phi_1\phi_2\phi, \\
\Pr(X_{1k} = X_{2k} = 0) &= (1-\phi_1)(1-\phi_2)\phi + (1-\phi), \\
\Pr(X_{1k} = 1,\ X_{2k} = 0,\ X_k = 1) &= \phi_1(1-\phi_2)\phi, \\
\Pr(X_{1k} = 0,\ X_{2k} = 1,\ X_k = 1) &= (1-\phi_1)\phi_2\phi.
\end{aligned}$$

The unknown parameters are denoted as

$$\underline{\psi} = (\phi, \phi_1, \phi_2, \mu_{1N}, \mu_{2N}, \beta_1, \beta_2, \sigma^2_{1N}, \sigma^2_{2N}, \sigma^2_{1P}, \sigma^2_{2P}, \rho)^T.$$

The values ϕ, ϕ_1, ϕ_2 are probabilities and are between 0 and 1, as is the correlation ρ. The values of σ^2_{1N}, σ^2_{2N}, σ^2_{1P} and σ^2_{2P} are above 0.

3.2 Parameters Estimation

3.2.1 Maximum Likelihood Estimation

The parameters $\underline{\psi}$ can be estimated by maximizing the log-likelihood function (3.1).

$$\begin{aligned}
l(\underline{\psi}) = \sum_{k=1}^{n} \log \big\{ &(\phi_1\phi_2\phi) \cdot f_1(Y_{1k}, Y_{2k} | \mu_{1P}, \mu_{2P}, \sigma^2_{1P}, \sigma^2_{2P}, \rho) \\
&+ ((1-\phi_1)(1-\phi_2)\phi + (1-\phi)) \cdot f_2(Y_{1k}, Y_{2k} | \mu_{1N}, \mu_{2N}, \sigma^2_{1N}, \sigma^2_{2N}) \\
&+ (\phi_1(1-\phi_2)\phi) \cdot f_3(Y_{1k}, Y_{2k} | \mu_{1P}, \mu_{2N}, \sigma^2_{1P}, \sigma^2_{2N}) \\
&+ ((1-\phi_1)\phi_2\phi) \cdot f_4(Y_{1k}, Y_{2k} | \mu_{1N}, \mu_{2P}, \sigma^2_{1N}, \sigma^2_{2P}) \big\},
\end{aligned} \tag{3.1}$$

where f_1, f_2, f_3 and f_4 are probability density functions of the four mixing bivariate normal distributions, respectively.

The estimates (MLE) can be found using numerical optimization and an iterative approach as follows:

1. Choose a starting value for $\underline{\psi}_1 = (\phi, \phi_1, \phi_2)^T$.
2. Maximize the log-likelihood as a function of

$$\underline{\psi}_2 = (\mu_{1N}, \mu_{2N}, \beta_1, \beta_2, \sigma^2_{1N}, \sigma^2_{2N}, \sigma^2_{1P}, \sigma^2_{2P}, \rho)^T$$

   for that fixed $\underline{\psi}_1$.
3. Denote the results are $\hat{\underline{\psi}}_2|\underline{\psi}_1$ and then maximize the log-likelihood as a function of $\underline{\psi}_1$ for fixed $\underline{\psi}_2 = \hat{\underline{\psi}}_2|\underline{\psi}_1$.
4. Denote the results as $\hat{\underline{\psi}}_1$ and use that as a starting value to repeat the steps above until the estimates converge.

The procedure of computing the MLEs, though no obstacle for the example data, may have some issues due to the constraints on the parameters ϕ, ϕ_1, ϕ_2, σ^2_{1N}, σ^2_{2N}, σ^2_{1P}, σ^2_{2P}, and ρ. To avoid such computation difficulties caused by the constraints, one may consider reparameterization on those parameters, like the logistic transformation on the parameters restricted between 0 and 1 and logarithm transformation on the variance parameters.

Without the constraint that $\mu_{iP} \geq \mu_{iN}$ for $i = 1, 2$, the log-likelihood function may be multimodal. Figure 3.2 demonstrates the plots of profile log-likelihood of μ_{iN} and μ_{iP} for $i = 1, 2$ on the example data. Note that there are a ridge in the plots, indicating that there is a lack of identifiability without the constraint: the constraint requires high values of either test to be "positive" and low values to be "negative".

The values of μ_{iP} and μ_{iN} can be interchanged without changing the log-likelihood much for $i = 1, 2$.

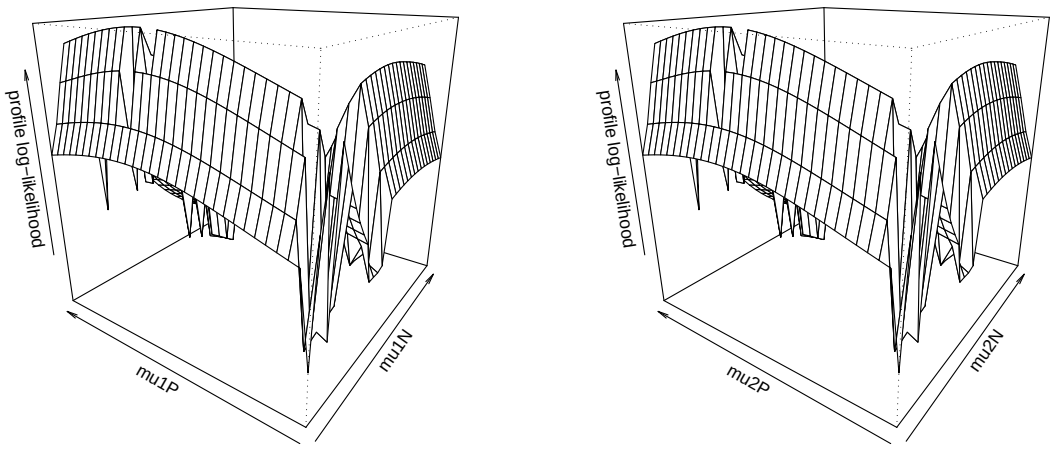

Figure 3.2: Profile log-likelihood plots for (μ_{1N}, μ_{1P}) (left) and (μ_{2N}, μ_{2P}) (right) .

3.2.2 Bayesian Estimation

For the ELISA tests, there is some prior information available and this is used in constructing the prior distribution. This prior distribution incorporates the constraint that $\mu_{iP} \geq \mu_{iN}$ for $i = 1, 2$. Because of the complexity of the model, it is

impossible to obtain the marginal posterior distribution for parameters analytically. The MCMC method is utilized to simulate samples from the marginal posterior distribution of each parameter. We use the software WinBUGS [34] to implement the MCMC method and use the R package `R2WinBUGS` [53] to call WinBUGS. The code is available in the Appendix.

3.3 Bayesian Decision Rule

The classification decision is chosen after observing the values of the random variables Y_1 and Y_2 and computing the posterior distribution, denoted $p(\underline{\psi}|data)$. The observed quantitative test results Y_1 and Y_2 provide information about the parameters $\underline{\psi}$. For a new sample with test results (Z_1, Z_2), let the loss of classifying this sample as negative if it is in fact positive be L_1 and the loss of classifying this sample as positive if it is negative be L_2, as illustrated in Table 3.1.

Table 3.1: Loss function in the decision function.

	Positive	Negative
Antibodies are present	0	L_1
Antibodies are absent	L_2	0

The posterior probability of E2 antibodies being present for (Z_1, Z_2), $Pr(X = 1|Z_1, Z_2, data)$, abbreviated to PPP, is:

$$PPP = \Pr(X = 1|Z_1, Z_2, data) = \int \Pr(X = 1|Z_1, Z_2, \underline{\psi}) p(\underline{\psi}|data) d\underline{\psi}$$
$$= \int \frac{f(Z_1, Z_2|X = 1, \underline{\psi}) \Pr(X = 1|\underline{\psi})}{f(Z_1, Z_2|X = 1) \Pr(X = 1|\underline{\psi}) + f(Z_1, Z_2|X = 0, \underline{\psi}) \Pr(X = 0|\underline{\psi})} p(\underline{\psi}|data) d\underline{\psi}. \tag{3.2}$$

Under the Bayesian decision theory [12], the risk under the negative classification is $\{PPP \cdot L_1\}$, and the risk under the positive classification is $\{(1 - PPP) \cdot L_2\}$.

The optimal Bayes decision for (Z_1, Z_2) based on the observed data is the one that has the smaller risk. Hence, (Z_1, Z_2) is classified as positive if $(1 - PPP) \cdot L_2 < PPP \cdot L_1$, which is equivalent to $PPP > C$, where $C = 1/(1 + L_1/L_2)$. The value $C = 0.5$ corresponds to $L_1 = L_2$ and represents a symmetric loss of misclassification. In many applications $L_1 \neq L_2$ and any value of C between 0 and 1 can be obtained by choosing different values. For example, false negatives in disease screening may lead to no treatment and subsequently worse consequences of the disease: in this case it may be appropriate to choose $L_1 > L_2$. Alternatively if the treatment subsequent to a positive result is toxic it may be appropriate to choose $L_2 > L_1$.

3.4 Illustration with the Motivating Example

3.4.1 Bayesian Classification

In the motivating example, a total of 100 blood specimens obtained from HIV infected subjects were tested with each of two tests: called the μPlate Anti-HGenv ($i = 1$) and M5 ($i = 2$) assays. The two assays are variations on the sandwich ELISA and the differences between them have been explained in more detail in Chapter 2. False negatives occur in both tests when the binding site of the human antibody Ab #2 to the E2 protein is blocked by MAb #1. The additional material introduced in the μPlate Anti-HGenv through the lysate may add additional noise that causes blocking. Neither test is perfect and false negatives are thought to occur approximately 10% of the time. Moreover, no commercial and validated test is available for the antibody, which means that there is no gold standard in the data.

In this example, the prevalence of the target antibodies varies between populations but the average is about 50% in HIV-infected populations and less than 5% in general blood donors based on previous study [6, 11, 25, 55, 56, 66]. Additionally, we know that the chance of blocking (false negatives) for each test is thought to be around 10%. Note however that these studies used imperfect tests. Based on such information, we choose the prior distribution for the prevalence ϕ with an expectation of 0.5, and the prior distribution for the probability of binding sites being accessible for each test ϕ_1 and ϕ_2 with an expectation of 0.9. Moreover, to guarantee the constraint on the mean values of results for each test, the prior distribution is set on β_1 and β_2 instead of μ_{1P} and μ_{2P}. The prior distribution for parameters other than ϕ, ϕ_1 and ϕ_2 are set as diffuse enough. The prior distribution (prior A) is listed as follows:

$$\begin{aligned}
\phi &\sim Beta(5,5) \\
\phi_1, \phi_2 &\sim Beta(18,2) \\
\mu_{1N}, \mu_{2N} &\sim N(0,100) \\
\beta_1, \beta_2 &\sim N(0,10000) \\
\sigma_{1N}^{-2}, \sigma_{2N}^{-2}, \sigma_{1P}^{-2}, \sigma_{2P}^{-2} &\sim \Gamma(0.01, 0.01). \\
\rho &\sim U(0,1),
\end{aligned}$$

with all of the above assumed to be independent.

The observed data are plotted on the right panel in Figure 3.3 and the PPP for each of the 100 blood samples are shown as a histogram in the left panel. Because

the classification was to be used in an analysis comparing antibody positive subjects to antibody negative subjects a value $C = 0.5$ was used for classification: positively classified samples are in green, and negatively classified samples are in red. The samples with low results on both tests are classified as E2 antibody negative, and samples with a high results on at least one test are classified as positive. This is consistent with the biological mechanism.

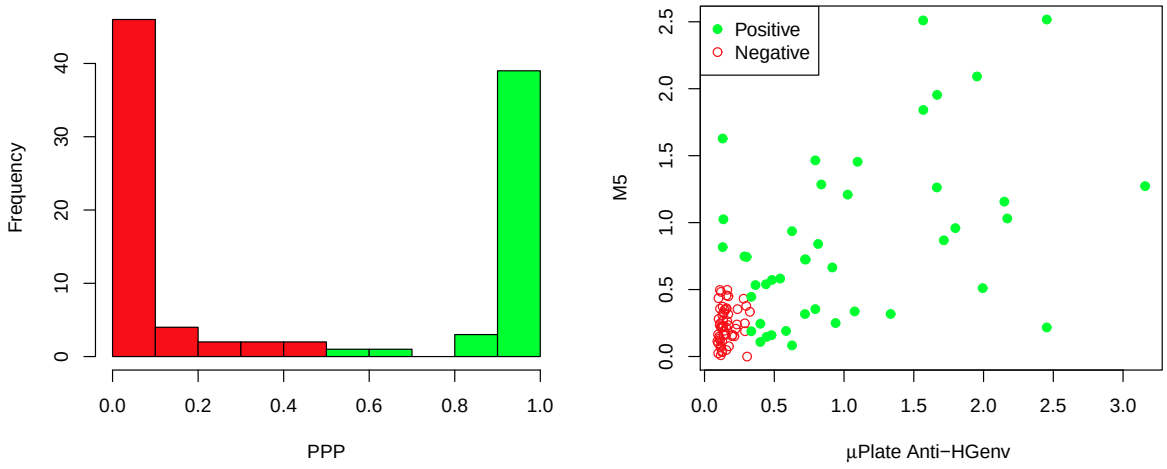

Figure 3.3: Histograms of posterior probability of E2 antibodies being present and the classifications of the 100 blood samples with $C = 0.5$.

The convergence is examined by Geweke's diagnostic [19], Heidelberger and Welch's diagnostic [23] and Raftery and Lewis's diagnostic [54]. The chain converges. Details of the posterior distribution are listed in Table 3.2, and Figure 3.4 shows the

prior and posterior probability density of each parameter marginally. The posterior estimates of the parameters in marginal distributions are close to the MLE's for parameters in the model. As shown in the Figure 3.4, the prior distributions for ϕ, ϕ_1 and ϕ_2 are informative, while the prior distributions for other parameters are quite diffuse.

Table 3.2: Summary statistics of the posterior distribution and the MLE for each parameter.

Parameter	Mean	SD	Median	MLE
ϕ	0.478	0.063	0.477	0.559
ϕ_1	0.907	0.045	0.913	0.907
ϕ_2	0.842	0.066	0.846	0.689
μ_{1N}	0.157	0.015	0.156	0.136
μ_{2N}	0.237	0.019	0.237	0.235
μ_{1P}	1.029	0.137	1.029	0.936
μ_{2P}	0.916	0.125	0.913	0.897
σ^2_{1N}	0.004	0.002	0.004	0.001
σ^2_{2N}	0.018	0.004	0.017	0.016
σ^2_{1P}	0.564	0.128	0.546	0.520
σ^2_{2P}	0.417	0.104	0.402	0.379
ρ	0.555	0.126	0.569	0.643

3.4.2 Sensitivity Analysis

A sensitivity analysis using six additional prior distributions (priors B, C, D, E, F and G in Table 3.3) is carried out to accesses the sensitivity of the estimation and classification to the prior distribution. Prior B, C and D use prior distributions that express different extent of information for ϕ_1 and ϕ_2, and E and F replace the prior

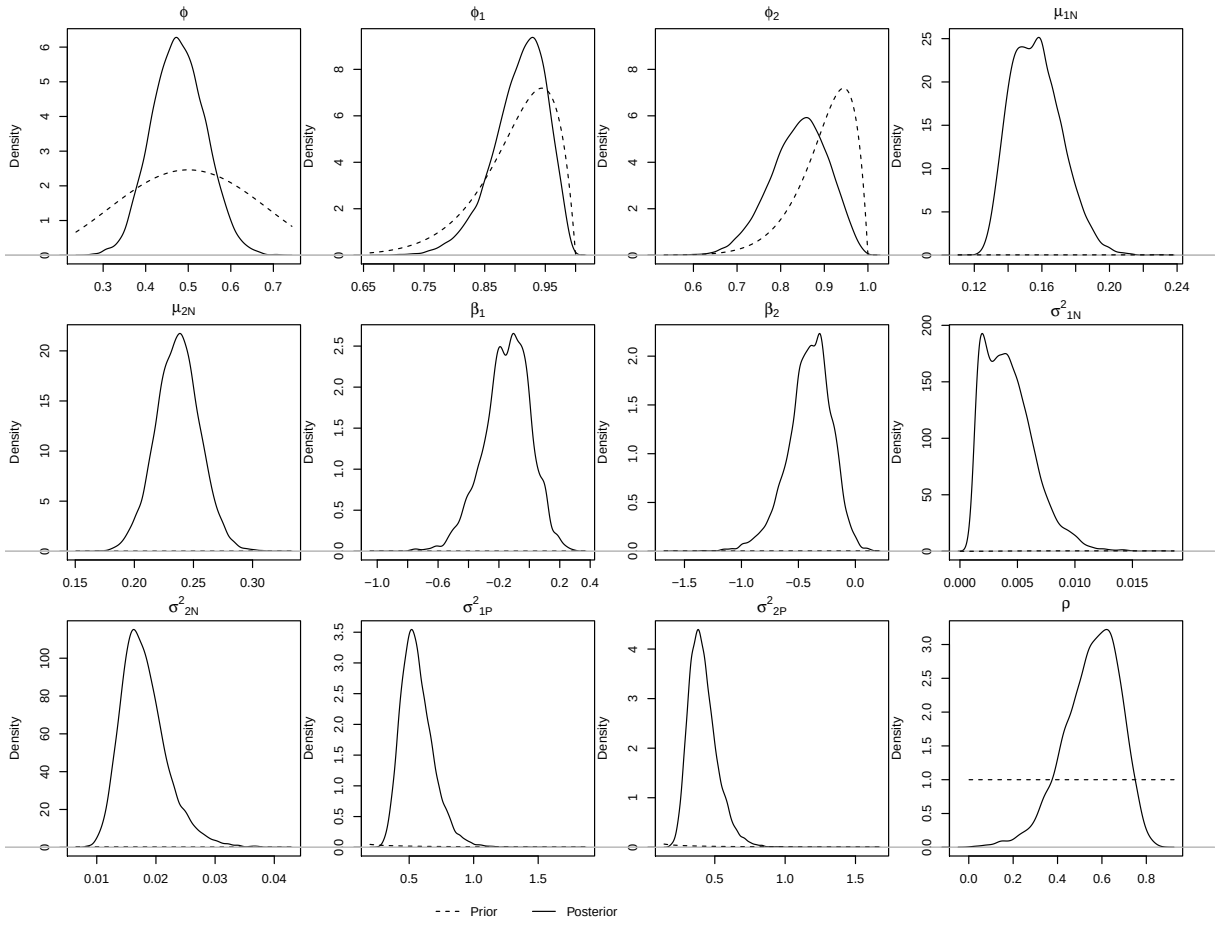

Figure 3.4: Marginal density plot of the prior (dashed line) and posterior (solid line) distribution for each parameter.

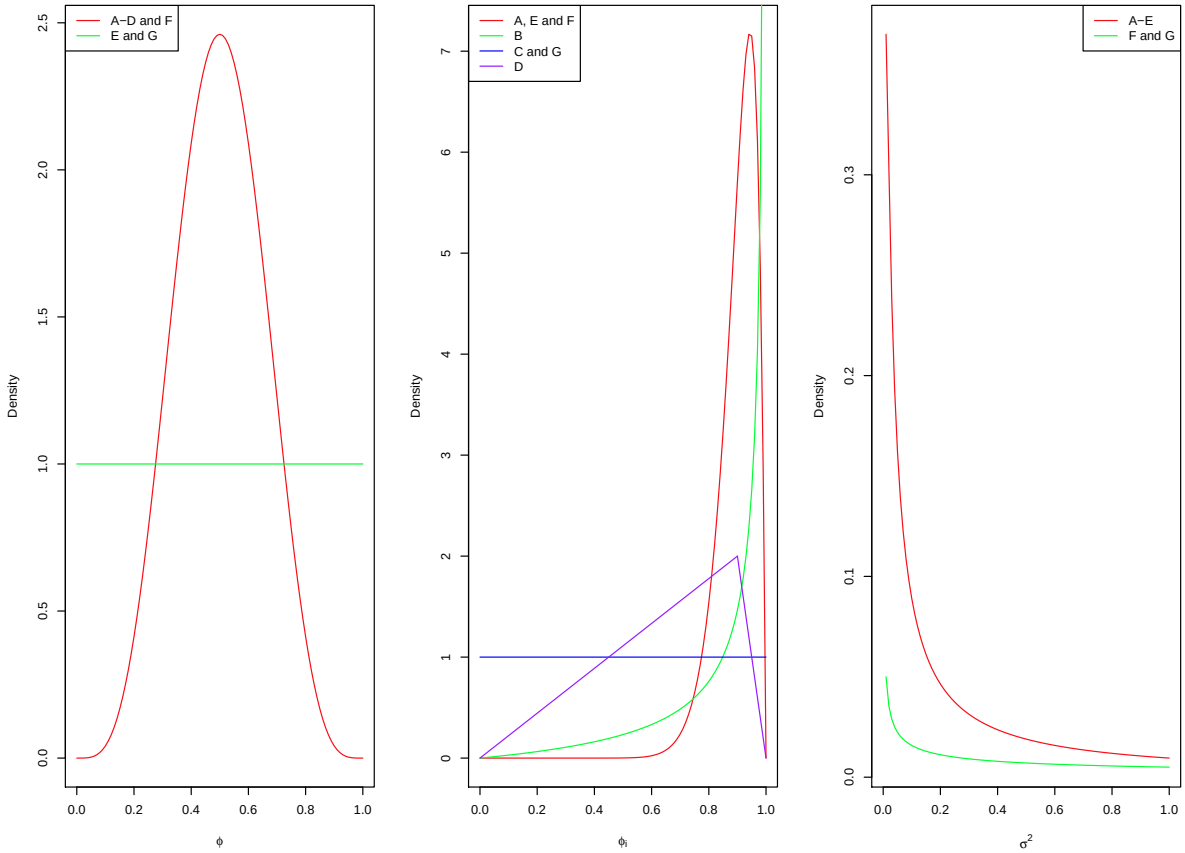

Figure 3.5: Plots of different prior distributions for ϕ, ϕ_i and σ^2.

distribution for ϕ and σ's in prior A by a less informative one, respectively. Prior G sets a less informative prior distribution for all parameters. Plots of the different priors are illustrated in Figure 3.5.

All the six chains converge. As for the parameter estimations, the posterior mean, standard deviation and median for all parameters listed in Table 3.4 look insensitive, but the marginal posterior densities shown in Figure 3.6 are sensitive to different choice of prior distributions.

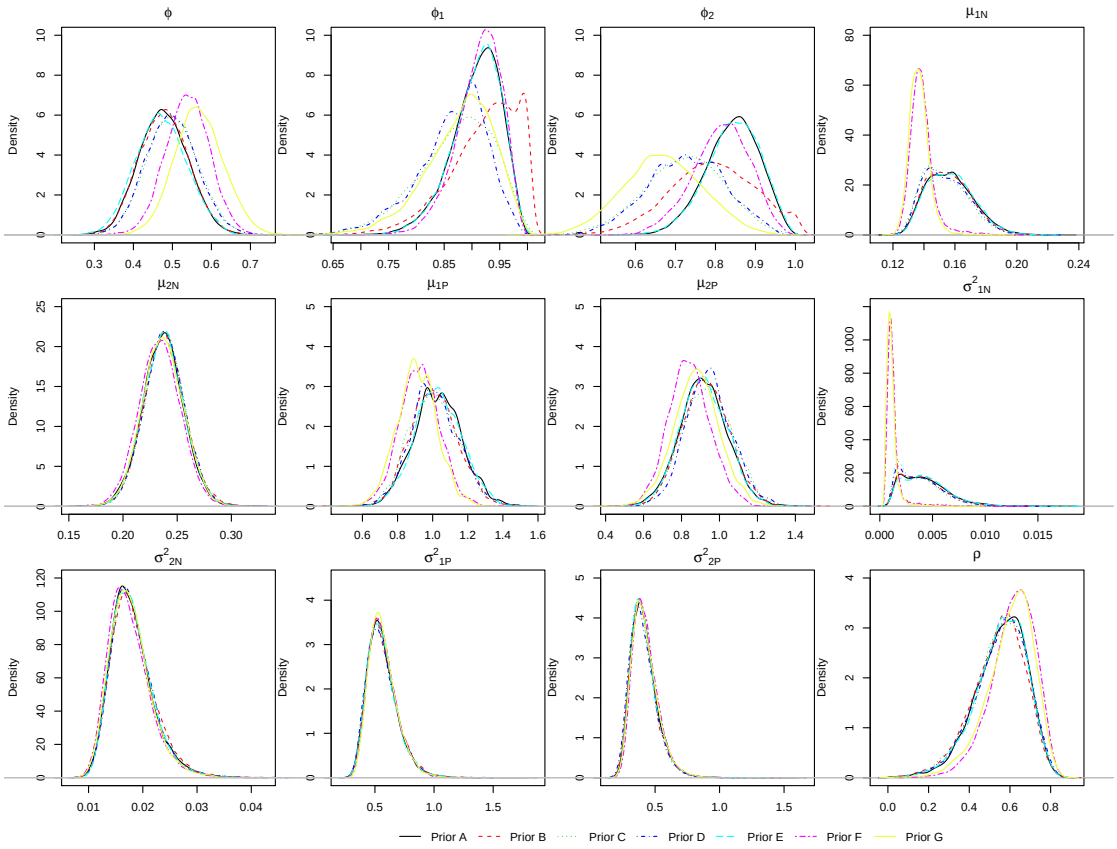

Figure 3.6: Marginal posterior densities of parameters for Prior A (solid), Prior B (dashed), Prior C (dotted), Prior D (dotdash), Prior E (longdash), Prior F (twodash) and Prior G(yellow solid).

The classification with $C = 0.5$ is quite stable under each prior distribution. The classification under priors B, C, D and E differs from the classification under prior A for only 1, 3, 2 and 1 samples respectively, and the classification under F and G differs for 7 and 8 samples respectively. Note that F and G are the least informative prior informations. Figure 3.7 illustrates the classifications.

The analyses are repeated without the constraint that $\mu_{iP} \geq \mu_{iN}$ for $i =$

1, 2, and if starting value is chosen that does not satisfy the constraint, the analysis sometimes converges to a local mode at which the constraint does not hold.

3.4.3 Classification by ML Approach

The classification can also be achieved in the ML approach. Since from the ML aspect, the unknown parameters are fixed, not random variables from some prior distributions, the PPP for each sample is calculated by the $\Pr(X = 1|Z_1, Z_2, \hat{\underline{\psi}})$ in the integrand of (3.2), where $\hat{\underline{\psi}}$ are the MLE of $\underline{\psi}$. Figure 3.8 illustrates the histogram of PPP estimated by ML approach and the corresponding classification, which has the same classification as using prior G.

3.5 Simulation Study

A simulation study, in a 4×3 factorial structure, is designed in order to assess the accuracy of the Bayesian classification rule developed above. The data are assumed to arise from either the mixture of four bivariate normal distributions or a similar mixture of four bivariate t distributions with 4 degrees of freedom. Skewed versions of the distributions are also used: the bivariate skew-normal and skew-t distributions with shape parameter -3 (right-skewed) [2]. The values of the parameters in the model are close to the posterior means from the motivating example in Table 3.2. The posterior distribution is calculated using the mixture of bivariate normal distributions. The classification under $C = 0.5$, 0.7 and 0.9 is implemented. Because the simulated data is generated with a known classification of each sample (gold standard), a linear discriminant analysis is also carried out; this assumes the model is a

mixture of two bivariate normal distributions. The empirical measures of the diagnostic accuracy are computed based on 500 simulated data sets for the three Bayesian classification rules and the linear discriminant classifier.

All the analyses converged and results are summarized in Table 3.5 and Table 3.6. Results in Table 3.5, indicate that even though the linear discriminant classifier uses more information, it assumes an incorrect distribution and it generally performs worse than the Bayesian classification method. Among the three Bayesian classification rules with different cutoff values C, for any kind of data, a higher cutoff value C leads to a lower sensitivity and a higher specificity (the higher C is, the fewer samples are classified as positive). At any fixed C, the sensitivity for the t data is slightly higher than that for the normal data, while the specificity for the normal data is much greater than the t data. This is reasonable considering that the t distribution has fatter tails, hence the true negative group has more overlap with the true positives. The PPV and NPV have similar comparisons as the specificity and sensitivity, implying that the mis-specified model tends to overestimate the PPP for the t data, and hence more samples are classified as positive. Adding the skewness to the data does not affect the performance of the classification much but Table 3.6 indicates that the coverage probabilities of the 95% highest posterior density intervals for some parameters are very low in many cases. Note that the parameter estimation is biased under the mis-specified model, especially for the location and scale parameters when the true underlying marginal distribution is a mixture of skew-normal or skew-t, according to Table 3.6.

3.6 Summary of the Bayesian Classification Method

In this chapter, a two-level latent model is proposed. The analysis of our motivating data set, combined with the simulation study, indicates that the classification rule using Bayesian decision theory classifies the 100 blood samples into E2 antibody positive and negative consistently with the biological background for the examples simulated. If the data are from the assumed bivariate normal mixture distribution, or from a similar t-distribution, with or without skewness, the classification has a robust discriminating capability.

The model is developed based on two ELISA tests for the E2 antibodies, but it can be extended easily to an arbitrary number of tests, or modified to accommodate different kinds of testing problem. For example, in a real time polymerase chain reaction (PCR) test, part of a virus genome is amplified and quantified. If a mutation occurs in that part of the genome, the primer does not detect the virus, and a false negative results. In RNA viruses especially, errors in transcription result frequently, and mutations (and hence false negatives) result.

The model assumes that conditioning on the antibodies being present ($X_k = 1$) and both binding sites being accessible ($X_{k1} = X_{k2} = 1$), the measurements are positively correlated. This is reasonable as they both measure the concentration of the E2 antibody in the sample. If either the antibody is absent ($X_k = 0$), or it is present but in one of the tests the binding site is blocked, then the responses are independent. This conditional independence assumption can be criticized, but in this case seems biologically very plausible. The two tests are carried out separately on

different plates, so if the antibody is present in the sample, the blocking of the binding site for one test is independent from the blocking for the other test. Therefore, the responses that reflect the concentration of the antibody in the sample from the two tests are consequently independent from each other. The conditional independence is reasonable.

In the biological mechanism, high values of a test result should correspond to "positive" classifications and low values to "negative". The constraint $\mu_{iP} \geq \mu_{iN}$ for $i = 1, 2$ is implemented by defining $\beta_i = \log(\mu_{iP} - \mu_{iN})$ $(i = 1, 2)$. Without the constraint, there is an identifiability question. Plots of profile log-likelihood of μ_{iN} and μ_{iP} $(i = 1, 2)$ indicate very well the issue in the parameter estimation for the example data set. The profile likelihoods for $i = 1, 2$ have a ridge, symmetric around the axis $\mu_{iP} = \mu_{iN}$ where the values of μ_{iP} and μ_{iN}can be interchanged without changing the likelihood much for each $i = 1, 2$. Omitting the constraint may lead to a classification that is inconsistent with the biological mechanism. The sensitivity analysis was also repeated without the constraint, and if starting value is chosen that does not satisfy the constraint, there exists an issue of identifiability.

To summarize, this method provides a reasonable method for combining the results of quantitative tests when there is no gold standard and false negatives may occur fairly frequently, independently on each test, and the probability of a false negative does not depend on the underlying value of the quantitative variable. It provides a systematic way of combining the results so that sufficiently high values of any one test lead to a positive classification.

Table 3.3: Six alternative sets of prior distributions for the mixture model.

Parameter	Prior B	Prior C	Prior D	Prior E	Prior F	Prior G
ϕ	$Beta(5,5)$	$Beta(5,5)$	$Beta(5,5)$	$Beta(1,1)$	$Beta(5,5)$	$Beta(1,1)$
ϕ_1, ϕ_2	$Beta(2, \frac{2}{9})$	$Beta(1,1)$	$\Delta(0.9)^*$	$Beta(18,2)$	$Beta(18,2)$	$Beta(1,1)$
μ_{1N}, μ_{2N}	$N(0,100)$	$N(0,100)$	$N(0,100)$	$N(0,100)$	$N(0,100)$	$N(0,100)$
β_1, β_2	$N(0,10^4)$	$N(0,10^4)$	$N(0,10^4)$	$N(0,10^4)$	$N(0,10^4)$	$N(0,10^4)$
$\sigma_{1N}^{-2}, \sigma_{2N}^{-2}, \sigma_{1P}^{-2}, \sigma_{2P}^{-2}$	$\Gamma(0.01,0.01)$	$\Gamma(0.01,0.01)$	$\Gamma(0.01,0.01)$	$\Gamma(0.01,0.01)$	$U(0,100)^\dagger$	$U(0,100)^\dagger$
ρ	$Beta(1,1)$	$Beta(1,1)$	$Beta(1,1)$	$Beta(1,1)$	$Beta(1,1)$	$Beta(1,1)$

* $\Delta(0.9)$ denotes a triangular distribution with the mode at 0.9.

† The prior distribution is on σ instead of σ^{-2}.

Table 3.4: Summary statistics of posterior distributions under six sets of prior distributions for sensitivity analysis

	Prior B			Prior C			Prior D		
	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
ϕ	0.479	0.063	0.479	0.500	0.065	0.499	0.498	0.065	0.497
ϕ_1	0.921	0.063	0.932	0.871	0.069	0.880	0.864	0.063	0.873
ϕ_2	0.792	0.106	0.794	0.721	0.099	0.725	0.727	0.096	0.728
μ_{1N}	0.156	0.014	0.155	0.153	0.014	0.152	0.155	0.015	0.153
μ_{2N}	0.238	0.019	0.238	0.239	0.019	0.239	0.238	0.019	0.238
μ_{1P}	1.007	0.134	1.003	1.007	0.140	1.004	1.013	0.137	1.007
μ_{2P}	0.927	0.125	0.926	0.936	0.129	0.934	0.944	0.121	0.941
σ^2_{1N}	0.004	0.002	0.004	0.004	0.002	0.003	0.004	0.002	0.004
σ^2_{2N}	0.018	0.004	0.018	0.018	0.004	0.018	0.018	0.004	0.018
σ^2_{1P}	0.568	0.131	0.547	0.560	0.126	0.544	0.561	0.129	0.544
σ^2_{2P}	0.412	0.103	0.397	0.406	0.101	0.392	0.407	0.105	0.392
ρ	0.542	0.129	0.557	0.550	0.130	0.563	0.554	0.129	0.569
	Prior E			Prior F			Prior G		
	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
ϕ	0.472	0.065	0.470	0.538	0.056	0.539	0.472	0.065	0.470
ϕ_1	0.908	0.045	0.914	0.912	0.041	0.917	0.908	0.045	0.914
ϕ_2	0.844	0.067	0.848	0.814	0.071	0.818	0.844	0.067	0.848
μ_{1N}	0.157	0.015	0.157	0.139	0.008	0.138	0.157	0.015	0.157
μ_{2N}	0.237	0.019	0.237	0.234	0.019	0.234	0.237	0.019	0.237
μ_{1P}	1.031	0.135	1.028	0.925	0.113	0.925	1.031	0.135	1.028
μ_{2P}	0.917	0.123	0.914	0.842	0.110	0.839	0.917	0.123	0.914
σ^2_{1N}	0.004	0.002	0.004	0.001	0.001	0.001	0.004	0.002	0.004
σ^2_{2N}	0.018	0.004	0.018	0.017	0.004	0.017	0.018	0.004	0.018
σ^2_{1P}	0.563	0.127	0.545	0.566	0.119	0.551	0.563	0.127	0.545
σ^2_{2P}	0.418	0.102	0.402	0.425	0.099	0.411	0.418	0.102	0.402
ρ	0.554	0.126	0.567	0.614	0.109	0.626	0.554	0.126	0.567

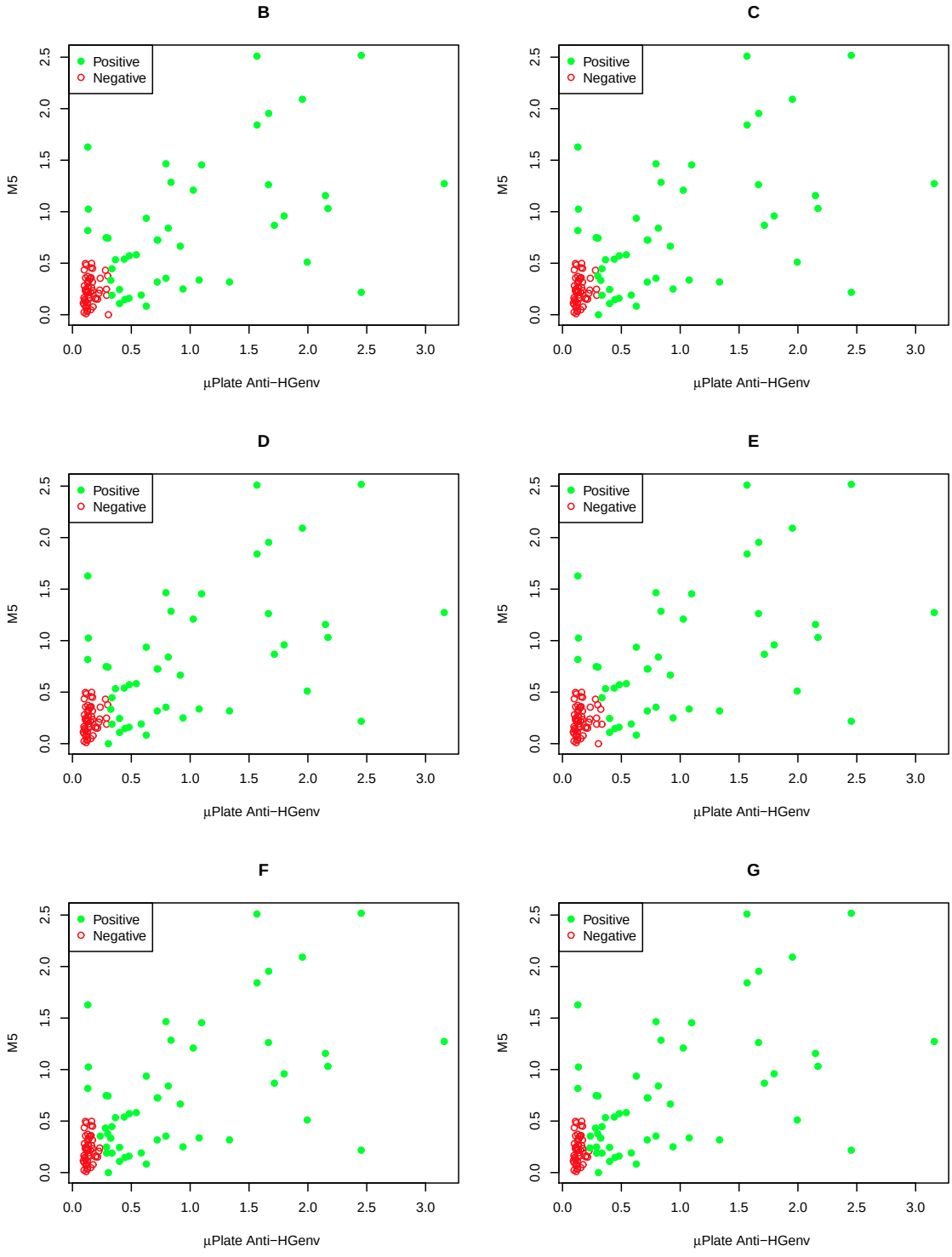

Figure 3.7: Plots of the classifications using six prior distributions B, C, D, E, F and G with $C = 0.5$.

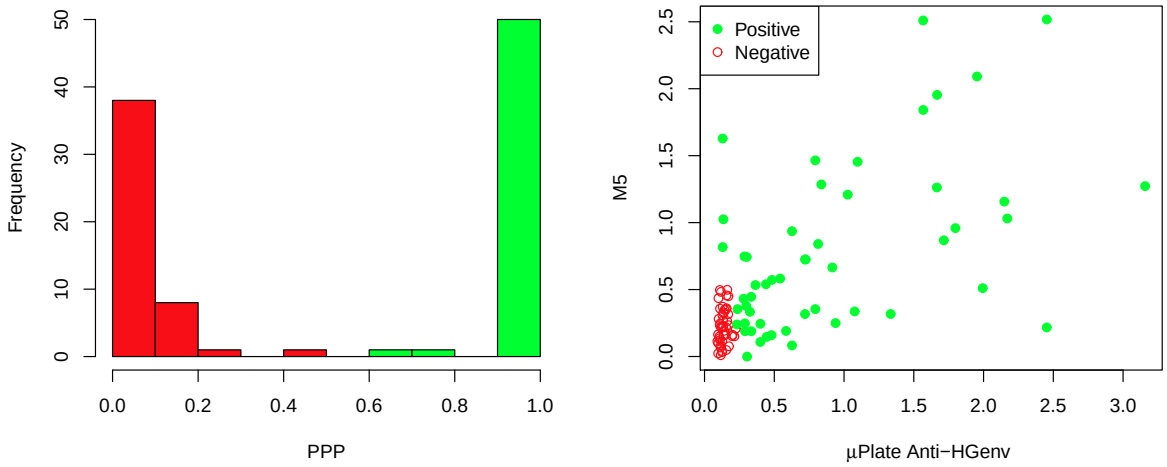

Figure 3.8: Histograms of posterior probability of E2 antibodies being present and the classifications of the 100 blood samples with $C = 0.5$ estimated by the ML approach.

Table 3.5: Empirical sensitivity, specificity, positive predictive value and negative predictive value of the Bayesian classification rule and the linear discriminant classifier based on 500 simulated data sets.

	Normal	t_4	Skew-Normal	Skew-t_4
		$C = 0.5$		
Sen (MCSE)	0.899(0.051)	0.931(0.039)	0.893(0.048)	0.964(0.025)
Spe (MCSE)	0.991(0.014)	0.948(0.034)	0.974(0.025)	0.945(0.037)
PPV (MCSE)	0.989(0.016)	0.944(0.036)	0.971(0.027)	0.944(0.036)
NPV (MCSE)	0.914(0.042)	0.937(0.035)	0.909(0.040)	0.965(0.024)
		$C = 0.7$		
Sen (MCSE)	0.886(0.054)	0.922(0.043)	0.875(0.052)	0.961(0.026)
Spe (MCSE)	0.996(0.009)	0.962(0.029)	0.988(0.018)	0.956(0.032)
PPV (MCSE)	0.996(0.010)	0.958(0.032)	0.986(0.020)	0.955(0.032)
NPV (MCSE)	0.904(0.044)	0.931(0.038)	0.896(0.043)	0.963(0.025)
		$C = 0.9$		
Sen (MCSE)	0.868(0.059)	0.909(0.045)	0.849(0.058)	0.956(0.028)
Spe (MCSE)	0.999(0.004)	0.977(0.022)	0.996(0.010)	0.970(0.028)
PPV (MCSE)	0.999(0.004)	0.973(0.026)	0.995(0.011)	0.968(0.028)
NPV (MCSE)	0.891(0.047)	0.921(0.040)	0.879(0.046)	0.959(0.026)
		Linear Discriminant Classifier		
Sen (MCSE)	0.688(0.062)	0.725(0.063)	0.668(0.068)	0.614(0.152)
Spe (MCSE)	0.999(0.005)	0.999(0.007)	0.984(0.066)	0.848(0.289)
PPV (MCSE)	0.999(0.005)	0.998(0.009)	0.986(0.049)	0.893(0.164)
NPV (MCSE)	0.775(0.032)	0.799(0.035)	0.765(0.043)	0.669(0.174)

Table 3.6: Empirical properties of posterior estimates based on 500 simulated data sets.

	Normal			t_4		
Parameter	Mean (MCSE)	SD (MCSE)	CP (%)	Mean (MCSE)	SD (MCSE)	CP (%)
ϕ (0.480)	0.476(0.049)	0.053(0.002)	97.2	0.503(0.048)	0.053(0.002)	95.0
ϕ_1 (0.900)	0.904(0.032)	0.047(0.008)	99.6	0.900(0.031)	0.046(0.007)	99.0
ϕ_2 (0.800)	0.864(0.040)	0.061(0.009)	82.0	0.876(0.040)	0.055(0.009)	69.0
μ_{1N} (0.150)	0.150(0.009)	0.010(0.001)	96.2	0.151(0.008)	0.008(0.001)	96.6
μ_{2N} (0.240)	0.242(0.021)	0.021(0.003)	94.2	0.242(0.017)	0.018(0.003)	95.6
μ_{1P} (1.000)	0.982(0.129)	0.128(0.018)	94.2	0.924(0.152)	0.123(0.029)	88.4
μ_{2P} (0.900)	0.857(0.131)	0.120(0.021)	93.0	0.784(0.142)	0.112(0.030)	81.6
σ^2_{1N} (0.004)	0.005(0.001)	0.001(0.000)	95.0	0.003(0.001)	0.001(0.000)	78.0
σ^2_{2N} (0.020)	0.021(0.005)	0.005(0.001)	95.0	0.014(0.004)	0.004(0.001)	55.4
σ^2_{1P} (0.570)	0.595(0.133)	0.139(0.034)	94.2	0.614(0.358)	0.140(0.088)	72.6
σ^2_{2P} (0.400)	0.425(0.101)	0.107(0.028)	95.0	0.437(0.218)	0.105(0.058)	78.8
ρ (0.540)	0.470(0.122)	0.135(0.019)	91.2	0.500(0.163)	0.123(0.027)	79.8
	Skew-Normal			Skew-t_4		
Parameter	Mean (MCSE)	SD (MCSE)	CP (%)	Mean (MCSE)	SD (MCSE)	CP (%)
ϕ (0.480)	0.488(0.049)	0.056(0.003)	97.8	0.512(0.046)	0.051(0.001)	93.2
ϕ_1 (0.900)	0.898(0.030)	0.053(0.009)	100	0.906(0.032)	0.042(0.006)	97.8
ϕ_2 (0.800)	0.882(0.033)	0.063(0.010)	78.2	0.882(0.035)	0.051(0.007)	61.0
μ_{1N} (0.150)	0.117(0.008)	0.008(0.001)	3.00	0.190(0.009)	0.009(0.001)	0.8
μ_{2N} (0.240)	0.171(0.019)	0.018(0.003)	6.6	0.326(0.021)	0.020(0.003)	1.2
μ_{1P} (1.000)	0.397(0.161)	0.106(0.032)	0.00	1.525(0.139)	0.143(0.046)	2.0
μ_{2P} (0.900)	0.349(0.133)	0.085(0.031)	0.00	1.295(0.108)	0.128(0.026)	8.0
σ^2_{1N} (0.004)	0.003(0.001)	0.001(0.000)	82.2	0.004(0.001)	0.001(0.000)	95.2
σ^2_{2N} (0.020)	0.014(0.003)	0.004(0.001)	56.0	0.018(0.005)	0.005(0.001)	86.4
σ^2_{1P} (0.570)	0.368(0.097)	0.092(0.029)	40.0	0.911(1.211)	0.207(0.308)	64.2
σ^2_{2P} (0.400)	0.242(0.065)	0.064(0.020)	36.0	0.595(0.317)	0.140(0.068)	70.8
ρ (0.540)	0.210(0.098)	0.122(0.027)	22.2	0.327(0.172)	0.125(0.029)	50.8

CHAPTER 4
FREQUENTIST CLASSIFICATION WITH MULTIPLE TESTS UNDER NO GOLD STANDARD

In this chapter, we develop methods for diagnoses with multiple tests under no gold standard motivated by the application to ELISA data. We generalize Su-Liu's linear discriminant method and MacIntosh-Pepe's risk score method introduced in Chapter 1 to combine multiple tests under the situation that no gold standard exists. The methods are developed by fitting a two-term multivariate normal mixture model to unclassified data on the results of diagnostic tests [27] with two terms corresponding to "case" and "non-case" respectively. The parameters of multivariate normal distributions and the event prevalence are estimated using maximum likelihood estimation with EM algorithm. Then Su-Liu's test and MnIntosh-Pepe's test are easily evaluated as the tests can be expressed as some functions of the parameters of the multivariate normal mixture distribution. We also develop a diagnostic method in a sequential fashion in which the second test is implemented only if the first test does not result in a positive diagnosis. The idea of combining tests in a sequence was first discussed by Kraemer [29], and Thompson [64] discussed theoretical accuracy of a sequence of tests. This method is practically sound when people know one test is definitely more sensitive to the case than the other and do not wish to conduct unnecessary multiple diagnostic tests either due to potential complication resulted from the tests or high cost associated with the tests.

4.1 Methods

For the simplicity in illustration, suppose that there are two quantitative diagnostic tests on each subject and for each test, a greater value of the test result indicates a larger chance of case. Denote X_i as the random variable representing the result from test i for $i = 1, 2$ and D as the random variable indicating the case presence, with $D = 1$ meaning case present and $D = 0$ meaning case absent. Moreover, F_1 and F_0 are the joint distribution functions of $\mathbf{X} = (X_1, X_2)$ for the case and non-case populations, respectively, and f_1 and f_0 are the corresponding probability density functions.

4.1.1 The Optimal Linear Composite Method

Suppose $\mathbf{X}$ is normally distributed under both $D = 1$ and $D = 0$ with different model parameters i.e., $\mathbf{X}|D = 1 \sim N(\mu_1, V_1)$ and $\mathbf{X}|D = 0 \sim N(\mu_0, V_0)$. Su and Liu [59] considered a linear combination $U = \mathbf{a}^T\mathbf{X}$ of the two diagnostic tests as a composite diagnostic test. Under the normality assumption, the coefficients $\mathbf{a}$ corresponding to the optimal linear-composite test, which provides the highest sensitivity uniformly at any specificity among all possible linear-composite tests, are given by (4.1).

$$\begin{aligned}\mathbf{a}_0 &\propto V_0^{-1/2}\left(I + V_0^{-1/2}V_1V_0^{-1/2}\right)^{-1}V_0^{-1/2}(\mu_1 - \mu_0)\\ &= (V_0 + V_1)^{-1}(\mu_1 - \mu_0),\end{aligned} \tag{4.1}$$

where I is the 2×2 identity matrix.

The ROC curve and the AUC of the optimal linear-composite test are easily calculated as

$$ROC(u) = \Phi\left(\frac{\mathbf{a}_0^T(\mu_1 - \mu_0) + \Phi^{-1}(u)\sqrt{\mathbf{a}_0^T V_0 \mathbf{a}_0}}{\sqrt{\mathbf{a}_0^T V_1 \mathbf{a_0}}}\right), \tag{4.2}$$

$$AUC = \Phi\left(\sqrt{(\mu_1 - \mu_0)^T (V_0 + V_1)^{-1}(\mu_1 - \mu_0)}\right), \tag{4.3}$$

where Φ is the cumulative distribution function of the standard normal distribution.

For diagnosis at a given specificity p_0, the threshold for the optimal linear-composite test is $\mathbf{a}_0^T\mu_0 + \Phi^{-1}(p_0)\sqrt{\mathbf{a}_0^T V_0 \mathbf{a}_0}$, and the corresponding sensitivity is

$$\text{sen}_A = \Phi\left(\frac{\mathbf{a}_0^T(\mu_1 - \mu_0) - \Phi^{-1}(p_0)\sqrt{\mathbf{a}_0^T V_0 \mathbf{a}_0}}{\sqrt{\mathbf{a}_0^T V_1 \mathbf{a}_0}}\right) \tag{4.4}$$

according to (4.2). Su-Liu's method can be automatically applied to the situation when no "gold standard" exists, as long as the model parameters can be consistently estimated.

4.1.2 The Optimal Risk Score Composite Method

McIntosh and Pepe [36] developed a composite diagnostic test in the framework of hypothesis testing in which $D = 0$ and $D = 1$ represent the null and alternative hypotheses, respectively. The decision rule to classify D as one is analogous to the rule for rejecting the null hypothesis in favor of the alternative. With this analogy, type I error corresponds to the FPR and the power corresponds to the TPR. Based on the Neyman-Pearson lemma [42] for the likelihood ratio test, it is easily established that the decision rule defined as

$$LR(\mathbf{X}) = f_1(\mathbf{X})/f_0(\mathbf{X}) > c(p_0), \text{where } \Pr(LR(\mathbf{X}) > c(p_0)|D = 0) = 1 - p_0.$$

is the uniformly most sensitive (UMS) diagnostic test among all tests with FPR= $1 - p_0$.

To evaluate the risk score $LR(\mathbf{X})$, one needs to estimate the distributions of the diagnostic markers X in advance. Using Bayes rule, they demonstrated that the decision rule can be equivalently postulated based on an alternative risk score $p(\mathbf{X}) = \Pr(D = 1|\mathbf{X})$ as

$$p(\mathbf{X}) = \frac{LR(\mathbf{X})q}{LR(\mathbf{X})q + 1} > c^*(p_0),$$

where $q = \Pr(D = 1)/\Pr(D = 0)$, and $\Pr(p(\mathbf{X}) > c^*(p_0)|D = 0) = 1 - p_0$. This risk score can be easily estimated via logistic regression $logit(p(\mathbf{X})) = \beta_0 + h(\beta, \mathbf{X})$ without the knowledge of the markers' distribution. But MnIntosh-Pepe's risk score is only applicable when a gold standard exists so that subjects can be classified without error.

When the diagnostic markers' distributions are estimable, the aforementioned risk score based composite test is also valid even though no "gold standard" is available. For continuous diagnostic markers, we use the original risk score

$$LR(\mathbf{X}) = f_1(\mathbf{X})/f_0(\mathbf{X}).$$

The threshold $c(p_0)$ under the specificity p_0 is in fact the p_0 percentile of $LR(\mathbf{X})$ evaluated for the control group. Let H_1 and H_0 denote the distribution function of $LR(X)$ under $X \sim F_1$ and $X \sim F_0$, respectively. Then the sensitivity for the threshold $c(p_0)$ is simply

$$\text{sen}_B = 1 - H_1\left(H_0^{-1}(p_0)\right). \tag{4.5}$$

To determine the decision rule and its sensitivity corresponding to a fixed specificity, one just needs to have some consistent estimates for the distribution function H_1 and the quantile of H_0.

4.1.3 The Optimal Sequential Composite Method

The composite tests described above require that subjects undergo all the diagnostic tests, which may not be desirable in practice in view of potential risks associated with the diagnostic tests or the excessive financial burden with the extra tests, particularly if the tests are all expensive. In many medical diagnosis procedures, one may start with the most sensitive test among all available diagnostic tools and continue to the second test only if the diagnosis based on the first test is not conclusive. This practical diagnostic procedure motivates us to design a new optimal composite test in a sequential fashion, to study its statistical properties, and to compare its performance with the existing composite tests.

Suppose Test 1 is superior to Test 2 judged by a greater value of AUC. The decision rule driven by the sequential composite test is determined by a pair of cut-off values (C_1, C_2) such that:

1. if $X_1 > C_1$, then this subject is classified as positive for the study event; else,
2. if $X_2 > C_2$, then classified as positive;
3. otherwise, classified as negative.

Figure 4.1 depicts the classification partition in the diagnostic test domain with the 2 cut-off sequential composite test.

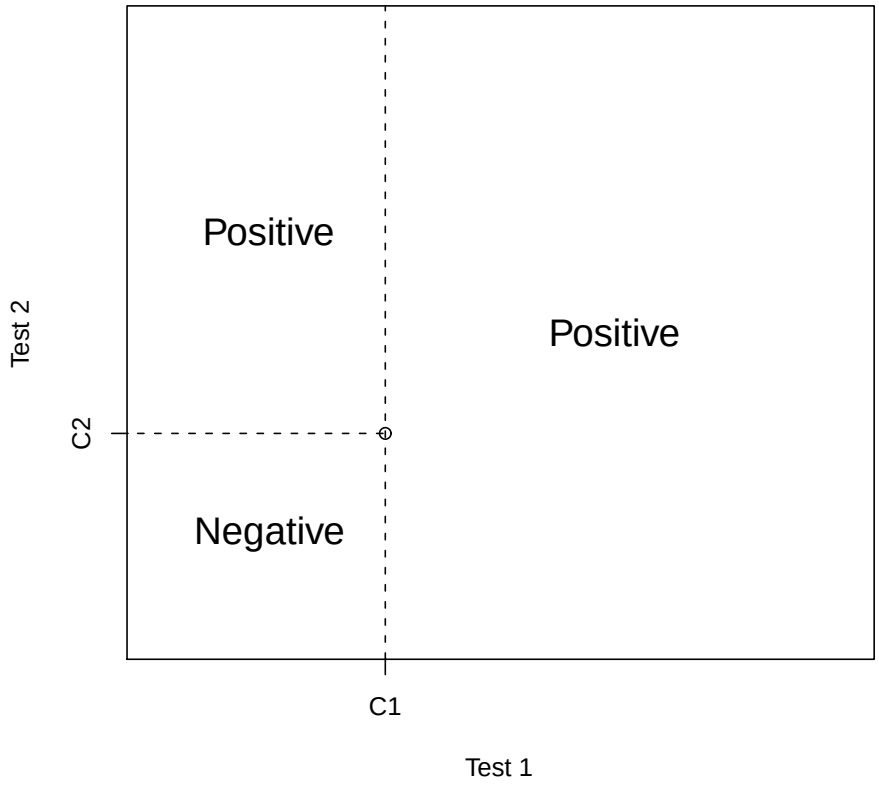

Figure 4.1: Illustration of the 2-cutoff sequential classification method.

Given the cut-off (C_1, C_2), the sensitivity and specificity for evaluating this composite test can be expressed as follows:

$$\begin{aligned}\text{Sensitivity} &= \Pr\left(\text{Positive classification}|\text{case}\right)\\ &= \Pr\left(X_1 > C_1|D=1\right) + \Pr\left(X_1 \le C_1,\ X_2 > C_2|D=1\right)\\ &= 1 - F_1(C_1, C_2). \qquad (4.6)\\ \text{Specificity} &= \Pr\left(\text{Negative classification}|\text{control}\right)\\ &= \Pr\left(X_1 \le C_1,\ X_2 \le C_2|D=0\right)\\ &= F_0(C_1, C_2). \qquad (4.7)\end{aligned}$$

We are searching for the optimal sequential composite test in the sense that it achieves the maximum sensitivity among all the sequential composite tests whose specificity is fixed at p_0 . Based on (4.6) and (4.7), this task can be converted to a constrained non-linear optimization problem:

$$\min_{F_0(C_1,C_2)=p_0} F_1(C_1, C_2). \qquad (4.8)$$

An efficient algorithm for finding the optimal (C_1, C_2) in (4.8) is essential in the development of this sequential method.

4.2 Computation

4.2.1 MLE of Multivariate Normal Mixture Model

Suppose we have a sample of diagnostic markers $X_1, X_2, \cdots, X_n$ that are assumed to be independent and identically distributed copies of X with distribution F. The implementation of all the foregoing methods requires estimation of F_1 and

F_0 from observed data in the first place. Here we follow the set-up of [59] for the distribution of the diagnostic markers X, i.e. $X|D = 1 \sim F_1 \equiv N(\mu_1, V_1)$ and $X|D = 0 \sim F_0 \equiv N(\mu_0, V_0)$. We adopt the mixture distribution of F_1 and F_0 to model the observed data, that is

$$F_\theta(\cdot) = \pi F_{1,\boldsymbol{\theta_1}}(\cdot) + (1-\pi)F_{0,\boldsymbol{\theta_0}}(\cdot), \tag{4.9}$$

where π is an unknown parameter indicating the mixture proportion, or equivalently, the case prevalence, and $\theta = (\pi, \theta_1, \theta_0) = (\pi, (\mu_1, V_1), (\mu_0, V_0))$ denotes the model parameters. The log-likelihood of the observed data can be expressed as:

$$\begin{aligned} l(\theta) &= \sum_{k=1}^{n} \log f_\theta(X_{1k}, X_{2k}) \\ &= \sum_{k=1}^{n} \log \left[\pi f_{1,\boldsymbol{\theta_1}}(X_{1k}, X_{2k}) + (1-\pi) f_{0,\boldsymbol{\theta_0}}(X_{1k}, X_{2k})\right]. \end{aligned}$$

In principal, the MLE of the parameters $\hat{\theta}_n$ can be computed by directly maximizing the log likelihood $l(\theta)$. It is however, found that this approach is not numerically stable. We note that if the gold standard does exist so that the exact memberships $\mathcal{D} = (D_1, \ldots, D_n)$ are known, the log likelihood for the augmented data $\{(X_1, D_1), \cdots, (X_n, D_n)\}$ is given by

$$l_a(\theta) = \sum_{k=1}^{n} D_k \log \pi f_{1,\theta_1}(X_{1k}, X_{2k}) + (1-D_k)\log(1-\pi) f_{0,\theta_0}(X_{1k}, X_{2k}) \tag{4.10}$$

and

$$Pr(D_k = 1|(X_1, \cdots, X_n); \theta) = \frac{\pi f_{1,\theta_1}(X_{1k}, X_{2k})}{\pi f_{1,\theta_1}(X_{1k}, X_{2k}) + (1-\pi) f_{0,\theta_0}(X_{1k}, X_{2k})}.$$

Hence the MLE of the model parameters $\hat{\theta}_n$ is easily computed using the EM algorithm [13] due to its numerical stability and algorithmic convenience for this problem.

The EM algorithm treats $\mathcal{D}$ as missing. Therefore, the complete data consist of $(\mathcal{X}, \mathcal{D})$, and the complete-data log-likelihood is given by (4.10).

Let $\theta^{(i)}$ denote the estimate of θ after ith iteration of the EM algorithm.

- **E step:** The E step computes the conditional expectation of $l_c(\theta)$ given the observed data $\mathcal{X}$ and the current estimates of $\theta = \theta^{(i)}$,

$$E\left(l_c(\theta)|\mathcal{X}, \theta = \theta^{(i)}\right) = \sum_{k=1}^{n} \left\{ \Pr(D_k = 1|X_{1k}, X_{2k}, \theta^{(i)}) \log \pi f_1(X_{1k}, X_{2k}|\theta_1) + \right.$$
$$\left. \Pr(D_k = 0|X_{1k}, X_{2k}, \theta^{(i)}) \log(1-\pi) f_0(X_{1k}, X_{2k}|\theta_0) \right\}.$$

If we write

$$\tilde{\pi}_k^{(i)} = \Pr(D_k = 1|X_{1k}, X_{2k}, \theta^{(i)}),$$
$$f_1^{(i)}(X_{1k}, X_{2k}) = f_1(X_{1k}, X_{2k}|\theta_1^{(i)}),$$
$$f_0^{(i)}(X_{1k}, X_{2k}) = f_0(X_{1k}, X_{2k}|\theta_0^{(i)}),$$

it is easy to show that

$$\tilde{\pi}_k^{(i)} = \frac{\pi^{(i)} f_1^{(i)}(X_{1k}, X_{2k})}{\pi^{(i)} f_1^{(i)}(X_{1k}, X_{2k}) + (1-\pi^{(i)}) f_0^{(i)}(X_{1k}, X_{2k})}, \tag{4.11}$$

and

$$E\left(l_c(\theta)|\mathcal{X}, \theta = \theta^{(i)}\right) = \sum_{k=1}^{n} \tilde{\pi}_k^{(i)} \log \pi f_1(X_{1k}, X_{2k}|\theta_1) + (1-\tilde{\pi}_k^{(i)}) \log(1-\pi) f_0(X_{1k}, X_{2k}|\theta_0). \tag{4.12}$$

- **M step:** The M step updates the estimate $\theta^{(i+1)}$ for θ by maximizing $E\left(l_c(\theta)|\mathcal{X}, \theta = \theta^{(i)}\right)$ in (4.12) with respect to θ. We can show that $\theta^{(i+1)}$ has the following explicit

expression:

$$\pi^{(i+1)} = \frac{1}{n}\sum_{k=1}^{n}\tilde{\pi}_k^{(i)}, \tag{4.13}$$

$$\mu_1^{(i+1)} = \frac{1}{n\pi^{(i+1)}}\sum_{k=1}^{n}\tilde{\pi}_k^{(i)}\mathbf{X}_k, \tag{4.14}$$

$$V_1^{(i+1)} = \frac{1}{n\pi^{(i+1)}}\sum_{k=1}^{n}\tilde{\pi}_k^{(i)}(\mathbf{X}_k - \mu_1^{(i+1)})(\mathbf{X}_k - \mu_1^{(i+1)})^T, \tag{4.15}$$

$$\mu_0^{(i+1)} = \frac{1}{n(1-\pi^{(i+1)})}\sum_{k=1}^{n}(1-\tilde{\pi}_k^{(i)})\mathbf{X}_k, \tag{4.16}$$

$$V_0^{(i+1)} = \frac{1}{n(1-\pi^{(i+1)})}\sum_{k=1}^{n}(1-\tilde{\pi}_k^{(i)})(\mathbf{X}_k - \mu_0^{(i+1)})(\mathbf{X}_k - \mu_0^{(i+1)})^T, \tag{4.17}$$

where $\mathbf{X}_k = (X_{1k}, X_{2k})^T$.

4.2.2 Computation of the Optimal Linear Composite Test

Obtaining $\hat{\theta}_n$, the estimates of the coefficients of the optimal linear composite test are directly computed by plugging $\hat{\theta}_n$ to (4.1) as

$$\hat{\mathbf{a}}_0 \propto (\hat{V}_0 + \hat{V}_1)^{-1}(\hat{\mu}_1 - \hat{\mu}_0), . \tag{4.18}$$

Accordingly, for the fixed specificity p_0, the threshold is estimated by

$$\hat{\mathbf{a}}_0^T\hat{\mu}_0 + \Phi^{-1}(p_0)\sqrt{\hat{\mathbf{a}}_0^T\hat{V}_0\hat{\mathbf{a}}_0}$$

, and the sensitivity by

$$\widehat{\text{sen}}_A = \Phi\left(\frac{\hat{\mathbf{a}}_0^T(\hat{\mu}_1 - \hat{\mu}_0) - \Phi^{-1}(p_0)\sqrt{\hat{\mathbf{a}}_0^T\hat{V}_0\hat{\mathbf{a}}_0}}{\sqrt{\hat{\mathbf{a}}_0^T\hat{V}_1\hat{\mathbf{a}}_0}}\right). \tag{4.19}$$

4.2.3 Computation of the Optimal Risk Score Composite Test

With the multivariate normal distribution assumption for the diagnostic markers X, the distribution functions of the risk score $LR(X)$ under $D = 0$ and $D = 1$

can be estimated empirically by drawing a random sample from $F_{0,\hat{\theta}_{0n}}$ and $F_{1,\hat{\theta}_{1n}}$, respectively. The estimate of the threshold for the decision rule and its corresponding sensitivity can be computed in the following subsequent steps:

- First draw a random sample of $V_{n,1}, \ldots, V_{n,m}$ from $F_{0,\hat{\theta}_{0n}}$, then form a sample of $\{Z_{n,i} = LR_{\hat{\theta}_n}(V_{n,i}) = f_{1,\hat{\theta}_{1n}}(V_{n,i})/f_{0,\hat{\theta}_{0n}}(V_{n,i}),\ i = 1, \ldots, m\}$

- Compute the empirical distribution of

$$\mathbb{H}_{0,\hat{\theta}_n,m}(Z) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}_{[Z_{n,i} \leq Z]}$$

  to estimate the distribution function of H_0. Then the threshold $\hat{c}(p_0)$ can be estimated by the sample p_0 percentile of $Z_{n,1}, \ldots, Z_{n,m}$, denoted as $\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(p_0)$, which is defined as

$$Z_{n,(i)},\ \text{for } p_0 \in \left(\frac{i-1}{m}, \frac{i}{m}\right].$$

- To estimate the sensitivity for the threshold $\hat{c}(p_0)$, draw a random sample of $W_{n,1}, \ldots, W_{n,m}$ from $F_{1,\hat{\theta}_{1n}}$, then form a sample of $\{Y_{n,i} = LR_{\hat{\theta}_n}(W_{n,i}),\ i = 1, \ldots, m\}$.

- Similarly, compute the empirical distribution of

$$\mathbb{H}_{1,\hat{\theta}_n,m}(Y) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}_{[Y_{n,i} \leq Y]},$$

  and estimate the sensitivity by

$$\widehat{\mathrm{sen}}_B = 1 - \mathbb{H}_{1,\hat{\theta}_n,m}\left(\hat{c}(p_0)\right) = 1 - \mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(p_0)\right). \tag{4.20}$$

4.2.4 Computation of the Optimal Sequential Composite Test

Under the normality assumption, the feasible set of (C_1, C_2) defined by a given specificity $F_0(C_1, C_2) = p_0$ constitutes a convex contour curve [65]. When the diagnostic markers are more variant for the case subjects, it is expected that the contour given by $F_1(C_1, C_2) = t$ is also convex but with less curvature and moves towards the origin of (C_1, C_2) domain as t decreases. The optimization problem (4.8) can be illustrated geometrically in Figure 4.2.

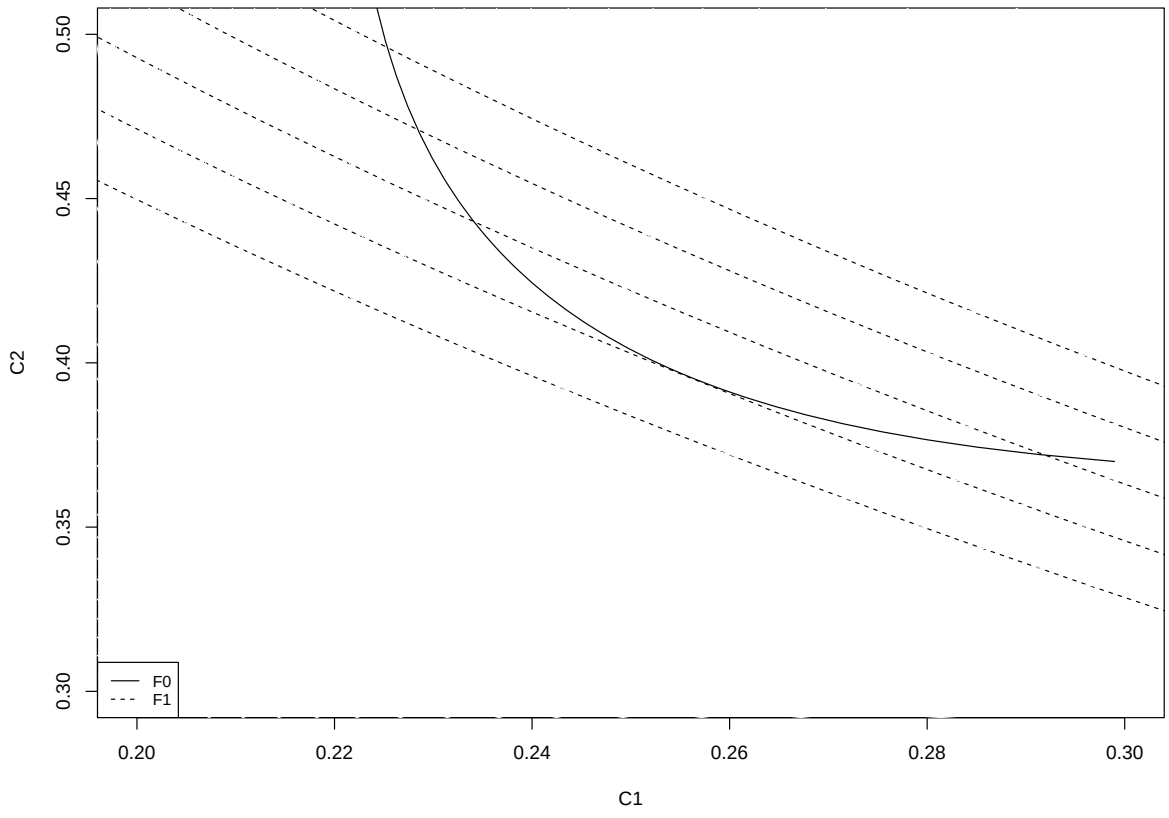

Figure 4.2: Illustration of the search for the optimal (C_1, C_2) at a given specificity p_0.

As seen in Figure 4.2, the constrained optimal value t corresponds to the value given by the contour that touches the contour of $F_0(C_1, C_2) = p_0$. The threshold (C_1, C_2) for the decision rule is simply the tangent point of the two contour lines and can be uniquely determined. Therefore, the original optimization problem (4.8) is converted to solving the system of bivariate nonlinear equations (4.21) for the tangent point of the contour lines of F_1 and F_0.

$$G(\mathbf{C}, \theta) = \begin{cases} F_{0,\theta_0}(C_1, C_2) = p_0 \\ \dfrac{\partial F_{1,\theta_1}}{\partial C_1}(C_1, C_2)\dfrac{\partial F_{0,\theta_0}}{\partial C_2}(C_1, C_2) - \dfrac{\partial F_{1,\theta_1}}{\partial C_2}(C_1, C_2)\dfrac{\partial F_{0,\theta_0}}{\partial C_1}(C_1, C_2) = 0. \end{cases} \tag{4.21}$$

The first equation represents the constraint given by the fixed specificity and the second equation reflects that the two contour lines have the same gradient at the tangent point. The Newton-Raphson method with the step-halving line search procedure is utilized to solve the system.

Let $\hat{\mathbf{C}}_n = (\hat{C}_{1n}, \hat{C}_{2n})$ denote the solution of (4.21) with the MLE of θ, $\hat{\theta}_n = (\hat{\theta}_{1n}, \hat{\theta}_{0n})$, then the sensitivity is estimated by $\widehat{\text{sen}}_C = 1 - F_{1,\hat{\theta}_{1n}}(\hat{C}_{1n}, \hat{C}_{2n})$.

4.3 Asymptotic Properties

Suppose θ is the true vector of the model parameters under the mixture of bivariate normal distribution. By the MLE properties [70], it is known that as $n \to \infty$, $\hat{\theta}_n \to_P \theta$, and

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to_d N(0, \mathcal{I}^{-1}),$$

where $\mathcal{I}$ is the Fisher information matrix given by $-E\left[\frac{\partial^2}{\partial\theta^2}l(\theta)\middle|\theta\right]$.

For the optimal linear composite method, the estimated sensitivity $\widehat{\text{sen}}_A$ as given in (4.19), is a continuous function of $\hat{\theta}_n$, hence it is consistent and asymptotically normally distributed by the ordinary continuous mapping theorem and the delta method.

For the optimal risk score method, the true sensitivity under the mixture bivariate normal distribution with the true model parameters θ is given by

$$\text{sen}_B = 1 - H_{1,\theta}\left(H_{0,\theta}^{-1}(p_0)\right).$$

The estimated sensitivity $\widehat{\text{sen}}_B$ can be shown consistent and asymptotically normal using the result of Theorem 4.1. The proof of this theorem is deferred to 4.4.1.

Theorem 4.1. *If m satisfies that $m/n \to \infty$ as $n \to \infty$, and furthermore, if $H_{1,\theta}(x)$ and $H_{0,\theta}(x)$ are continuously differentiable with respect to both θ and x and their first derivatives are bounded, then for any fixed $p_0 \in (0,1)$,*

$$\sqrt{n}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}_{0,\hat{\theta}_n,m}^{-1}(p_0)\right) - H_{1,\theta}\left(H_{0,\theta}^{-1}(p_0)\right)\right]$$

converges weakly to a normal distribution with mean 0 and variance given by (4.27).

For the optimal sequential composite test, let $\mathbf{C}_0 = (C_{10}, C_{20})$ denote the solution of the system (4.21) under the true parameters θ, then the true sensitivity is $\text{sen}_C = 1 - F_{1,\theta}(C_{10}, C_{20})$. The estimated sensitivity $\widehat{\text{sen}}_C$ is also consistent and asymptotically normal under the mild condition (4.22) given in Theorem 4.2. The proof of the theorem is also deferred to 4.4.2

Theorem 4.2. *If F_0 and F_1 are continuously differentiable with respect to $\mathbf{C} = (C_1, C_2)$ and θ and satisfy the following inequality (4.22) at C_0 and θ,*

$$\begin{aligned}&\left[\frac{\partial^2 F_1}{\partial C_1 \partial C_2}\frac{\partial F_0}{\partial C_2} + \frac{\partial F_1}{\partial C_1}\frac{\partial^2 F_0}{\partial {C_2}^2} - \frac{\partial^2 F_1}{\partial {C_2}^2}\frac{\partial F_0}{\partial C_1} - \frac{\partial F_1}{\partial C_2}\frac{\partial^2 F_0}{\partial C_1 \partial C_2}\right]\frac{\partial F_0}{\partial C_1} \\ &- \left[\frac{\partial^2 F_1}{\partial {C_1}^2}\frac{\partial F_0}{\partial C_2} + \frac{\partial F_1}{\partial C_1}\frac{\partial^2 F_0}{\partial C_1 \partial C_2} - \frac{\partial^2 F_1}{\partial C_1 \partial C_2}\frac{\partial F_0}{\partial C_1} - \frac{\partial F_1}{\partial C_2}\frac{\partial^2 F_0}{\partial {C_1}^2}\right]\frac{\partial F_0}{\partial C_2} \neq 0\end{aligned} \tag{4.22}$$

then as sample size $n \to \infty$, $\sqrt{n}\,(\widehat{\text{sen}}_C - \text{sen}_C)$ converges to a normal distribution with mean 0 and variance given by (4.28).

Remark 4.1. Based on the conditions in Theorem 4.1, in practice, when calculating the optimal sensitivity for the risk score composite test, the corresponding m can be selected as any polynomial function of n with an order greater than 1. For example, $m = n^2$.

Remark 4.2. Condition (4.22) can be justified algebraically for bivariate normal random variables when F_1 and F_0 have a different covariance matrix. In fact, the left side is the determinant of the Jacobian matrix of (4.21).

Remark 4.3. Although the asymptotic normality holds for all the three estimators under fairly mild conditions, the asymptotic variances of the sensitivities are hard to estimate directly. Therefore for the inference, their standard errors are estimated using the nonparametric bootstrap method [16]. Specifically, 200 samples with the same size are drawn from the original data with replacement. For each of the three diagnostic tests, each sample yields an estimated sensitivity at the given specificity, and the standard error is then estimated by the standard deviation of the 200 estimated sensitivities.

4.4 Theoretical Results

4.4.1 Proof of Theorem 4.1

Throughout Section 4.4.1, K is denoted as a universal constant that may vary from place to place. For each n, $X_{n,1}, X_{n,2}, \ldots, X_{n,m(n)}$, where $m(n)/n \to \infty$ as $n \to \infty$, are *i.i.d.* random variables according to a probablility measure P_n, which converges to a measure P_0 in a suitable sense. Let $\mathcal{F}$ be the collection of all indicator functions of form $f_t(x) = \mathrm{I}_{[x \le t]}$, with t ranging over $\mathbb{R}$. Define the semimetric ρ_P on $\mathcal{F}$ as

$$\begin{aligned}\rho_P(f) &= \left(P|f|^2\right)^{1/2} \\ &= \left(\int |f(x)|^2 dP\left[X \le x\right]\right)^{1/2}, \forall f \in \mathcal{F},\end{aligned}$$

and define $l^\infty(\mathcal{F})$ as the set of all uniformly bounded functionals on $\mathcal{F}$: $z : \mathcal{F} \mapsto \mathbb{R}$ such that

$$\sup_{f \in \mathcal{F}} |z(f)| < \infty.$$

$\mathbb{P}_n$ is the ordinary empirical measure based on the sample of $X_{n,1}, \cdots, X_{n,m(n)}$, that is

$$\mathbb{P}_n = \sum_{i=1}^{m(n)} \delta_{X_{n,i}},$$

and the centered empirical measure of $\mathbb{P}_n$ is defined as

$$\mathbb{G}_{n,P_n} = \sqrt{m(n)}(\mathbb{P}_n - P_n),$$

given in Chapter 2 of [71] (page 173).

The following lemmas are needed in order to prove Theorem 4.1.

Lemma 4.3. *If the semimetric ρ_{P_n} converges uniformly to ρ_{P_0} in the sense that*

$$\sup_{f,g\in\mathcal{F}} |\rho_{P_n}(f,g) - \rho_{P_0}(f,g)| \to 0, \quad as \;\; n \to \infty \tag{4.23}$$

then $\mathbb{G}_{n,P_n}$ converges in distribution to $\mathbb{G}_{P_0}$ in $l^\infty(\mathcal{F})$, where $\mathbb{G}_{P_0}$ is a tight Borel measurable element in $l^\infty(\mathcal{F})$, i.e., the limit process $\{\mathbb{G}_{P_0}f, f \in \mathcal{F}\}$ is a Gaussian process with zero mean and covariance functions given in the proceeding display (see Chapter 19, page 269 of [70]).

$$E\mathbb{G}_{P_0}f_i\mathbb{G}_{P_0}f_j = P_0f_if_j - P_0f_iP_0f_j.$$

Proof. We prove this lemma using Theorem 2.8.10 of [71]. $\mathcal{F}$ contains all the indicator functions of form $f_t(x) = \mathrm{I}_{[x\le t]}$ for $t \in \mathbb{R}$, so it is a class of measurable functions. Since the indicator function takes values at only 0 and 1, the constant function $F = 1$ is an envelope function for $\mathcal{F}$, which is measurable and totally bounded for ρ_{P_0}. Moreover, $\forall\epsilon > 0$, $\exists N$, such that $\forall n \ge N$, $\epsilon\sqrt{m(n)} > 1$ and $F^2\left\{F \ge \epsilon\sqrt{m(n)}\right\} = 0$. Hence, the condition that $\limsup_{n\to\infty} P_nF^2\left\{F \ge \epsilon\sqrt{m(n)}\right\} = 0$ is satisfied.

Next, we show that the class $\mathcal{F}$ is P_n-Donsker. For each n and for any $\epsilon > 0$, assuming that H_n is the distribution function induced by the probability measure P_n, we construct the brackets of the form $[\mathrm{I}_{(-\infty,t_{i-1}]}, \mathrm{I}_{(-\infty,t_i]}]$ with a grid of points $-\infty = t_0 < t_1 < \ldots < t_k = \infty$ satisfying $H_n(t_i-) - H_n(t_{i-1}) < \epsilon$ for each i. This can be achieved by the fact that P_n is a probability measure that converges to P_0. These brackets have $L_1(P_n)$-size ϵ, and the total number k is bounded by $1/\epsilon$. Because $P_nf^2 \le P_nf$ for every $0 \le f \le 1$, the $L_2(P_n)$-size of the brackets is bounded by $\sqrt{\epsilon}$. Thus we have the bracketing number $N_{[\,]}(\sqrt{\epsilon}, \mathcal{F}, L_2(P_n)) \le (1/\epsilon)$. Equivalently,

$N_{[\,]}(\epsilon, \mathcal{F}, L_2(P_n)) \leq (1/\epsilon^2)$. The bracketing entropy of $\mathcal{F}$ is of the order of $\log(\epsilon)$, which is $o(1/\epsilon^\tau)$ for any $\tau \in (0,1)$ since $\lim_{\epsilon \to 0} \epsilon^\tau \log(\epsilon) = 0$. Therefore, the bracketing integral

$$\begin{aligned} J_{[\,]}(\delta, \mathcal{F}, L_2(P_n)) &= \int_0^\delta \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P_n))} d\epsilon \\ &\leq K \int_0^\delta \epsilon^{-\tau/2} d\epsilon = \frac{K}{1-\tau/2} \delta^{1-\tau/2} \\ &< \infty, \text{ for any } \delta > 0. \end{aligned}$$

This indicates that $\mathcal{F}$ is a P_n-Donsker class, and it also indicates that

$$\lim_{\delta_n \to 0} J_{[\,]}(\delta_n, \mathcal{F}, L_2(P_n)) = 0.$$

Therefore, all the conditions in Theorem 2.8.10 of [71] are satisfied, so $\mathbb{G}_{n,P_n} \rightsquigarrow \mathbb{G}_{P_0}$ in $l^\infty(\mathcal{F})$.

Lemma 4.4. *Suppose that the condition* (4.23) *in Lemma 4.3 is satisfied. If f_n is a sequence of functions in $\mathcal{F}$ such that $\int (f_n - f_0)^2 dP_0$ converges to 0 in probability for some $f_0 \in \mathcal{F}$, then $\mathbb{G}_{n,P_n}(f_n - f_0) \to_P 0$ and $\mathbb{G}_{n,P_n} f_n \rightsquigarrow \mathbb{G}_{P_0} f_0$.*

Proof. The proof of this lemma is similar to the proof of Lemma 19.24 of [70]. Define a functional $g : l^\infty(\mathcal{F}) \times \mathcal{F} \mapsto \mathbb{R}$ by $g(z, f) = z(f) - z(f_0)$. The semimetrics on $l^\infty(\mathcal{F})$ and $\mathcal{F}$ are denoted by $\|\cdot\|_\infty$ and $\|\cdot\|_2$, respectively. We need to show that g is continuous with respect to the product semimetric at every (z, f) such that $f \mapsto z(f)$ is continuous.

Actually, if $(z_n, f_n) \to (z, f)$ in $l^\infty(\mathcal{F}) \times \mathcal{F}$, then $\|z_n - z\|_\infty \to 0$, and $\|f_n -$

$f\|_2 \to 0$. Hence,

$$\begin{aligned}
|g(z_n, f_n) - g(z, f)| &= |\,(z_n(f_n) - z_n(f_0)) - (z(f) - z(f_0))\,| \\
&= |z_n(f_n) - z(f_n) + z(f_n) - z(f) - (z_n(f_0) - z(f_0))\,| \\
&\leq |z_n(f_n) - z(f_n)| + |z(f_n) - z(f)| + |z_n(f_0) - z(f_0)| \\
&\leq \|z_n - z\|_\infty + |z(f_n) - z(f)| + \|z_n - z\|_\infty \\
&= o(1) \text{ if } z \text{ is continuous at } f.
\end{aligned}$$

Now let $z_n = \mathbb{G}_{n,P_n}$, and $z = \mathbb{G}_{P_0}$. By the assumption, $f_n \to_P f_0$ in metric space $\mathcal{F}$, and by Lemma 4.3, $\mathbb{G}_{n,P_n} \rightsquigarrow \mathbb{G}_{P_0}$ in $l^\infty(\mathcal{F})$. Then $(\mathbb{G}_{n,P_n}, f_n) \rightsquigarrow (\mathbb{G}_{P_0}, f_0)$ in the space $l^\infty(\mathcal{F}) \times \mathcal{F}$. By Lemma 18.15 of [70], $\mathbb{G}_{P_0}$ is continuous on $\mathcal{F}$ for almost all sample paths, thus g is continuous at almost every point of $(\mathbb{G}_{P_0}, f_0)$. By the continuous mapping theorem, $\mathbb{G}_{n,P_n}(f_n - f_0) = g(\mathbb{G}_{n,P_n}, f_n) \rightsquigarrow g(\mathbb{G}_{P_0}, f_0) = 0$. It is equivalent to $\mathbb{G}_{n,P_n}(f_n - f_0) \to_P 0$.

The second assertion follows because $\mathbb{G}_{n,P_n} f_n = o_P(1) + \mathbb{G}_{n,P_n} f_0 \rightsquigarrow \mathbb{G}_{P_0} f_0$ due to the result of Lemma 4.3.

Lemma 4.5. *Suppose F_n is the underlying distribution function of $X_{n,1}, \ldots, X_{n,m(n)}$ with $F_n \rightsquigarrow F_0$[1], and $\mathbb{F}_n$ is the empirical distribution function of $X_{n,1}, \ldots, X_{n,m(n)}$. If F_n and F_0 are continuously differentiable at $F_n^{-1}(c)$ and $F_0^{-1}(c)$ with bounded and strictly positive derivative $f_n\left(F_n^{-1}(c)\right)$ and $f_0\left(F_0^{-1}(c)\right)$ for $\forall c \in (0,1)$, respectively, then, $\sqrt{m(n)}\left(\mathbb{F}_n^{-1}(c) - F_n^{-1}(c)\right)$ converges in distribution to a normal distribution with mean 0 and variance $c(1-c)/f_0^2\left(F_0^{-1}(c)\right)$.*

[1] $F_n \rightsquigarrow F_0$ if and only if $F_n(t) \to F_0(t)$ at every t where F_0 is continuous.

Proof. Define a function ϕ as $\phi(P) = F^{-1}(c)$, then $\phi(P_n) = F_n^{-1}(c)$ and $\phi(\mathbb{P}_n) = \mathbb{F}_n^{-1}(c)$. By the *von Mises expansion* in Chapter 20 of [70],

$$\phi(\mathbb{P}_n) - \phi(P_n) \approx \frac{1}{\sqrt{m(n)}}\phi'(\mathbb{G}_{n,P_n}) = \frac{1}{m(n)}\sum_{i=1}^{m(n)} \phi'_{P_n}(\delta_{X_{n,i}} - P_n),$$

where the influence function $\phi'_{P_n}(\delta_x - P_n)$ can be computed as

$$\phi'_{P_n}(\delta_x - P_n) = \frac{d}{dt}_{|t=0} \phi\left((1-t)P_n + t\delta_x\right).$$

Let $I_n(x) = \phi'_{P_n}(\delta_x - P_n)$ and $P_{ntx} = (1-t)P_n + t\delta_x$. For any t, $F_{P_{ntx}}\left(\phi(P_{ntx})\right) = (1-t)F_n\left(\phi(P_{ntx})\right) + tG_x\left(\phi(P_{ntx})\right) = c$, where $G_x(t) = \mathrm{I}_{[x,\infty)}(t)$. Taking the derivative with respect to t and evaluating it at $t = 0$ on both sides, we get

$$\begin{aligned} &-F_n\left(\phi(P_n)\right) + f_n\left(\phi(P_n)\right) I_n(x) + G_x\left(\phi(P_n)\right) \\ &\qquad = -c + f_n\left(\phi(P_n)\right) I_n(x) + \mathrm{I}_{[x \le \phi(P_n)]} = 0. \end{aligned}$$

So the influence function is given by

$$I_n(x) = -\frac{\mathrm{I}_{[x \le \phi(P_n)]} - c}{f_n\left(\phi(P_n)\right)},$$

and

$$\begin{aligned} \sqrt{m(n)}\left(\mathbb{F}_n^{-1}(c) - F_n^{-1}(c)\right) &= \sqrt{m(n)}\left(\phi(\mathbb{P}_n) - \phi(P_n)\right) = \frac{1}{\sqrt{m(n)}} I_n(X_{n,i}) + o_P(1) \\ &= \frac{1}{f_n\left(F_n^{-1}(c)\right)} \cdot \frac{1}{\sqrt{m(n)}} \sum_{i=1}^{m(n)} \left(\mathrm{I}_{[X_{n,i} \le F_n^{-1}(c)]} - c\right) + o_P(1). \end{aligned}$$

Note that $f_n = F_n'$ and $f_0 = F_0'$ are the density functions of F_n and F_0, respectively. $F_n^{-1}(c) \to F_0^{-1}(c)$, because $F_n \rightsquigarrow F_0$ by Lemma 21.2 of [70]. Hence $f_n\left(F_n^{-1}(c)\right) \to f_0\left(F_0^{-1}(c)\right)$.

Furthermore, $\mathrm{I}_{(-\infty,F_n^{-1}(c)]}$ is a sequence of functions in $\mathcal{F}$ that converges to $\mathrm{I}_{(-\infty,F_0^{-1}(c)]}$ in the sense that

$$\int \left(\mathrm{I}_{(-\infty,F_n^{-1}(c)]} - \mathrm{I}_{[(-\infty,F_0^{-1}(c)]}\right)^2 dP_0 = \left| \int_{F_0^{-1}(c)}^{F_n^{-1}(c)} dP_0 \right| \to 0.$$

Thus by Lemma 4.4,

$$\frac{1}{\sqrt{m(n)}} \sum_{i=1}^{m(n)} \left(\mathrm{I}_{[X_{n,i}\le F_n^{-1}(c)]} - c\right) = \mathbb{G}_{n,P_n}\mathrm{I}_{[(-\infty,F_n^{-1}(c)]} \rightsquigarrow \mathbb{G}_{P_0}\mathrm{I}_{[(-\infty,F_0^{-1}(c)]}.$$

Taking together, $\sqrt{m(n)}\left(\mathbb{F}_n^{-1}(c) - F_n^{-1}(c)\right) \rightsquigarrow \frac{1}{f_0^2\left(F_0^{-1}(c)\right)}\mathbb{G}_{P_0}\mathrm{I}_{[(-\infty,F_0^{-1}(c)]}$, which is the normal distribution with mean 0 and variance $c(1-c)/f_0^2\left(F_0^{-1}(c)\right)$.

Now using the preceding lemmas, Theorem 4.1 is proved as follows.

It is noted that

$$\begin{aligned}
&\sqrt{n}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right] \\
&= \frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\hat{\theta}_n}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\right] \qquad (4.24)\\
&+ \frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}\left[H_{1,\hat{\theta}_n}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n}(q_0)\right)\right] \qquad (4.25)\\
&+ \sqrt{n}\left[H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n}(q_0)\right) - H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right]. \qquad (4.26)
\end{aligned}$$

Now we examine the asymptotic properties of (4.24) - (4.26) one at a time.

First we show that (4.24) $= o_P(1)$. Let θ be the true model parameters, so the MLE $\hat{\theta}_n \to_P \theta$, and $\sup_{t\in\mathbb{R}} |H_{1,\hat{\theta}_n}(t) - H_{1,\theta}(t)| \to_P 0$ due to the condition that the derivative of $H_{1,\theta}$ with respect to θ is uniformly bounded in θ. Suppose $g_1(x) = \mathrm{I}_{[x\le t_1]}$ and $g_2(x) = \mathrm{I}_{[x\le t_2]}$ $(\forall t_1, t_2 \in \mathbb{R})$ are two indicator functions from $\mathcal{F}$ and without loss

of generality, $t_1 < t_2$, then under the $L_2(P)$-metric as the semimetric on $\mathcal{F}$, namely,

$$\rho_{P_n}(g_1, g_2) = \sqrt{\int (g_1 - g_2)^2 dP_n} = \sqrt{\int \left(\mathrm{I}_{[t_1 \le x \le t_2]}\right)^2 dH_{1,\hat{\theta}_n}(x)}$$
$$= \sqrt{H_{1,\hat{\theta}_n}(t_2) - H_{1,\hat{\theta}_n}(t_1)},$$

and

$$\rho_{P_0}(g_1, g_2) = \sqrt{\int (g_1 - g_2)^2 dP_0} = \sqrt{\int \left(\mathrm{I}_{[t_1 \le x \le t_2]}\right)^2 dH_{1,\theta}(x)}$$
$$= \sqrt{H_{1,\theta}(t_2) - H_{1,\theta}(t_1)},$$

it is easily seen that,

$$\sup_{g_1, g_2 \in \mathcal{F}} |\rho_{P_n}(g_1, g_2) - \rho_{P_0}(g_1, g_2)| = \sup_{t_1, t_2 \in \mathbb{R}} \left| \sqrt{H_{1,\hat{\theta}_n}(t_2) - H_{1,\hat{\theta}_n}(t_1)} - \sqrt{H_{1,\theta}(t_2) - H_{1,\theta}(t_1)} \right|$$
$$\to 0 \text{ as } n \to \infty.$$

This justifies Lemma 4.3.

Similarly, it is easily shown that $\sup_{t \in \mathbb{R}} |H_{0,\hat{\theta}_n}(t) - H_{0,\theta}(t)| \to_P 0$, so by Lemma 4.5,

$$\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right) \to_d N\left(0, q_0(1-q_0)/h^2_{0,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right),$$

where $h_{0,\theta}$ is the density function of $H_{0,\theta}$. This implies that $\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0) = o_P(1)$, and hence

$$\int \left(\mathrm{I}_{[x \le \mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)]} - \mathrm{I}_{[x \le H^{-1}_{0,\theta}(q_0)]}\right)^2 dH_{1,\theta_0}(x) = \left| H_{1,\theta}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right) \right| \to_P 0,$$

by the continuous mapping theorem.

Let $f_n(x) = \mathrm{I}_{\left[x \le \mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right]}$, and $f_0(x) = \mathrm{I}_{\left[x \le H^{-1}_{0,\theta}(q_0)\right]}$. Therefore, by Lemma 4.4,

$$\begin{aligned}
&\sqrt{m}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\hat{\theta}_n}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\right] \\
&= \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^{m}\mathrm{I}_{\left[Y_{n,i} \le \mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right]} - \int \mathrm{I}_{[x \le \mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)]} dH_{1,\hat{\theta}_n}(x)\right] \\
&= \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^{m} f_n(Y_{n,i}) - \int f_n(x) dH_{1,\hat{\theta}_n}(x)\right] = \mathbb{G}_{n,P_n} f_n \rightsquigarrow \mathbb{G}_{P_0} f_0 = O_P(1).
\end{aligned}$$

It results in

$$\begin{aligned}
(4.24) &= \sqrt{n/m}\sqrt{m}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\hat{\theta}_n}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\right] \\
&= \sqrt{n/m} O_P(1) = o_P(1),
\end{aligned}$$

since $\lim_{n\to\infty}(n/m) = 0$.

Next we show that (4.25) is also $o_P(1)$.

$$\begin{aligned}
(4.25) &= \sqrt{\frac{n}{m}}\sqrt{m}\left[H_{1,\hat{\theta}_n}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\right] \\
&= \sqrt{\frac{n}{m}}\sqrt{m}\left[h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)\right. \\
&\qquad \left. + O_P\left(\left|\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right|^2\right)\right] \\
&= \sqrt{\frac{n}{m}}\left[h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)\right. \\
&\qquad \left. + \sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)^2 O_P(1)\right] \\
&= \sqrt{\frac{n}{m}}\left(O_P(1) + o_P(1)\right) = o_P(1),
\end{aligned}$$

since

$$h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n,m}(q_0)\right)\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right) \to_d N\left(0, q_0(1-q_0)\frac{h^2_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)}{h^2_{0,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)}\right),$$

by Lemma 4.5 and

$$\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)^2 = \left[\sqrt{m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)\right]\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0) - H^{-1}_{0,\hat{\theta}_n}(q_0)\right)$$
$$= O_P(1)\cdot o_P(1) = o_P(1).$$

Finally,

$$\begin{aligned}
(4.26) &= \sqrt{n}\left[H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n}(q_0)\right) - H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right] \\
&= \sqrt{n}\left\{H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n}(q_0)\right) - H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right) + H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right) - H_{1,\theta_0}\left(H^{-1}_{0,\theta}(q_0)\right)\right\} \\
&= \sqrt{n}\left\{h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right)\left(H^{-1}_{0,\hat{\theta}_n}(q_0) - H^{-1}_{0,\theta}(q_0)\right) + \nabla_\theta H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)(\hat{\theta}_n - \theta)\right\} + o_P(1) \\
&= \sqrt{n}\left\{h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right)\nabla_\theta H^{-1}_{0,\theta}(q_0)(\hat{\theta}_n - \theta) + \nabla_\theta H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)(\hat{\theta}_n - \theta)\right\} + o_P(1) \\
&= \left\{h_{1,\hat{\theta}_n}\left(H^{-1}_{0,\theta}(q_0)\right)\nabla_\theta H^{-1}_{0,\theta}(q_0) + \nabla_\theta H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right\}\sqrt{n}(\hat{\theta}_n - \theta) + o_P(1).
\end{aligned}$$

By the continuous mapping theorem and the delta method, (4.26) is asymptotically normal with mean 0 and covariance matrix $V = A\mathcal{I}^{-1}A^T$, where

$$A = h_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\nabla_\theta H^{-1}_{0,\theta}(q_0) + \nabla_\theta H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right). \tag{4.27}$$

In summary, as $n \to \infty$,

$$\begin{aligned}
&\sqrt{n}\left[\mathbb{H}_{1,\hat{\theta}_n,m}\left(\mathbb{H}^{-1}_{0,\hat{\theta}_n,m}(q_0)\right) - H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right] = (4.24) + (4.25) + (4.26) \\
&= o_P(1) + o_P(1) + \sqrt{n}\left[H_{1,\hat{\theta}_n}\left(H^{-1}_{0,\hat{\theta}_n}(q_0)\right) - H_{1,\theta}\left(H^{-1}_{0,\theta}(q_0)\right)\right] \\
&\to_d N(0, V).
\end{aligned}$$

4.4.2 Proof of Theorem 4.2

Since F_1 and F_0 are the cumulative distribution function of bivariate normal distributions, the function $G(\mathbf{C}, \theta)$ is continuously differentiable with respect to $\mathbf{C}$

and θ. Condition (4.22) is equivalent to the statement that the Jacobian matrix $\nabla_{\mathbf{C}} G(\mathbf{C}_0, \theta)$ is invertible by deriving the determinant of $\nabla_{\mathbf{C}} G(\mathbf{C}_0, \theta)$ and setting it not equal to zero. Hence, according to the implicit function theorem [31], there exists an open set U containing θ, an open set V containing $\mathbf{C}_0$, and a unique continuous differentiable function $g : U \to V$ such that $\mathbf{C} = g(\theta)$ and $G(g(\theta), \theta) = 0$ for all $\theta \in U$.

Based on the MLE properties, it is known that $\hat{\theta}_n \to_p \theta$ and $\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, \mathcal{I}^{-1})$. So for any $\epsilon > 0$ and $\delta > 0$, there exists an N, such that $n > N$, $\Pr(|\hat{\theta}_n - \theta| > \delta) < \epsilon$. This implies that for any $n > N$, $\hat{\theta}_n \in U$ in probability, and hence the proposed method for finding the cut-off $\hat{\mathbf{C}}_n = (\hat{C}_{n,1}, \hat{C}_{n,2})$ through solving for $G(\hat{\mathbf{C}}_n, \hat{\theta}_n) = 0$ results in $\hat{\mathbf{C}}_n = g(\hat{\theta}_n)$ in probability.

Further note that $F_1(\mathbf{C}, \theta) = F_1(g(\theta), \theta)$ is a continuously differentiable function of θ, and consequently, by the continuous mapping theorem and the delta method, we have

$$
\begin{aligned}
\sqrt{n}\,(\widehat{\mathrm{sen}}_C - \mathrm{sen}_C) &= \sqrt{n}\left(F_1(\hat{\mathbf{C}}_n, \hat{\theta}_n) - F_1(\mathbf{C}_0, \theta)\right) \\
&= \sqrt{n}\left(F_1(\hat{\mathbf{C}}_n, \hat{\theta}_n) - F_1(\mathbf{C}_0, \hat{\theta}_n) + F_1(\mathbf{C}_0, \hat{\theta}_n) - F_1(\mathbf{C}_0, \theta)\right) \\
&= \sqrt{n}\left(\nabla_{\mathbf{C}} F_1(\mathbf{C}_0, \hat{\theta}_n)(\hat{\mathbf{C}}_n - \mathbf{C}_0) + \nabla_{\theta} F_1(\mathbf{C}_0, \theta)(\hat{\theta}_n - \theta)\right) + o_p(1) \\
&= \sqrt{n}\left(\nabla_{\mathbf{C}} F_1(\mathbf{C}_0, \hat{\theta}_n)\nabla_{\theta} g(\theta)(\hat{\theta}_n - \theta) + \nabla_{\theta} F_1(\mathbf{C}_0, \theta)(\hat{\theta}_n - \theta)\right) + o_p(1) \\
&= \left(\nabla_{\mathbf{C}} F_1(\mathbf{C}_0, \hat{\theta}_n)\nabla_{\theta} g(\theta) + \nabla_{\theta} F_1(\mathbf{C}_0, \theta)\right)\sqrt{n}(\hat{\theta}_n - \theta) \\
&\to_d N(0, B\mathcal{I}^{-1}B^T),
\end{aligned}
$$

where

$$B = \nabla_{\mathbf{C}} F_1(\mathbf{C}_0, \theta) \nabla_\theta g(\theta) + \nabla_\theta F_1(\mathbf{C}_0, \theta). \tag{4.28}$$

4.5 Numerical Results

4.5.1 Application: ELISA Data

In this section, we applied all three methods to the motivating data set from the E2 antibody study example described in Chapter 2 with the goal of detection of the antibody presence in the blood sample of 100 HIV infected study participants. The scatter plots of Figure 4.3 presents the results from the two ELISA tests. The data are fitted by the mixture of two bivariate normal distributions: $N(\mu_1, V_1)$ and $N(\mu_0, V_0)$. The maximum likelihood estimation gives $\hat{\mu}_1 = (1.01, 0.84)^T$, $\hat{\mu}_0 = (0.16, 0.24)^T$, and

$$\hat{V}_1 = \begin{pmatrix} 0.54 & 0.22 \\ 0.22 & 0.40 \end{pmatrix}, \hat{V}_0 = \begin{pmatrix} 0.004 & 0.001 \\ 0.001 & 0.017 \end{pmatrix}.$$

The estimated ROC curves based on the two individual tests are depicted in Figure 4.4. It appears that Test 1 is superior to Test 2 as it is more sensitive to the antibody presence for any given specificity and hence it is chosen to be the initial test for our proposed sequential method.

The decision rules for the three optimal composite tests with specificity = 0.90 are superimposed to the scatter plot of Figure 4.3 to characterize the composite tests. Their ROC curves are also plotted in Figure 4.4 along with those based on the individual tests. Figure 4.4 indicates that all the optimal composite tests provide a better discriminant capability than the two individual tests for this application. Both the optimal sequential and optimal risk score composite tests are substantially

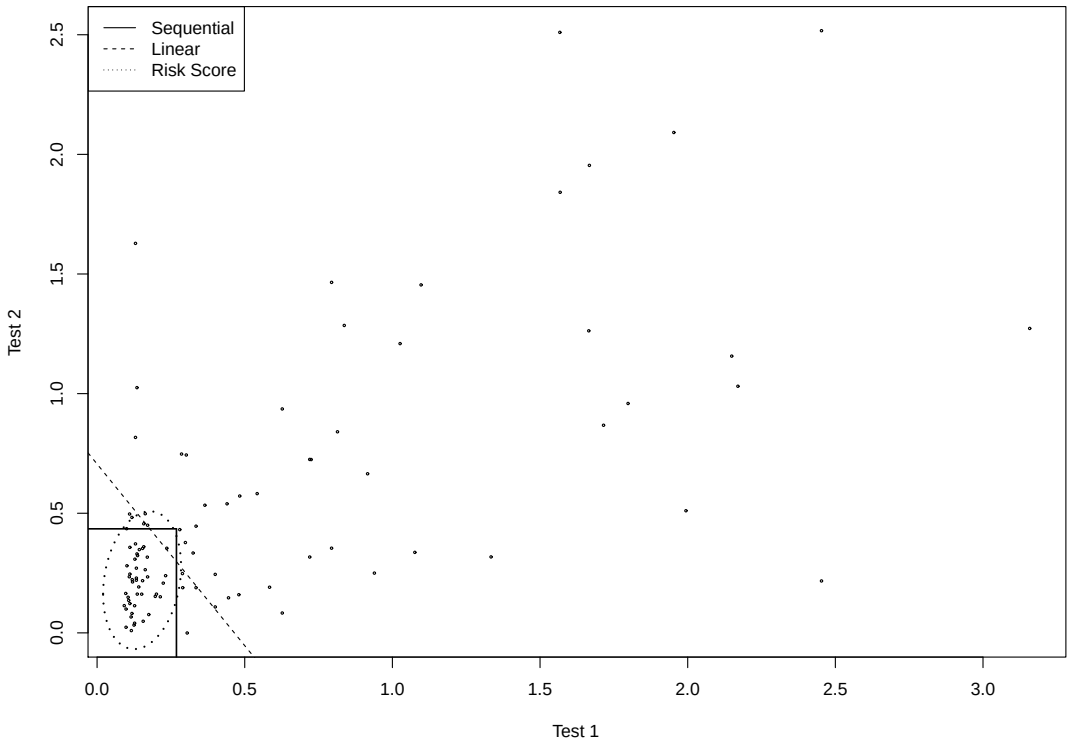

Figure 4.3: Results from the two tests in 100 blood samples along with the optimal linear composite test, the optimal risk score composite test and the optimal sequential composite test at specificity = 0.90.

better than the individual tests, but the optimal linear composite test only adds little discriminant power to the best individual test. While the optimal risk score composite test is superior to the optimal sequential composite test as anticipated due to Neyman-Pearson theorem of likelihood ratio test, the optimal sequential composite test only needs 53% of the blood samples for the second test.

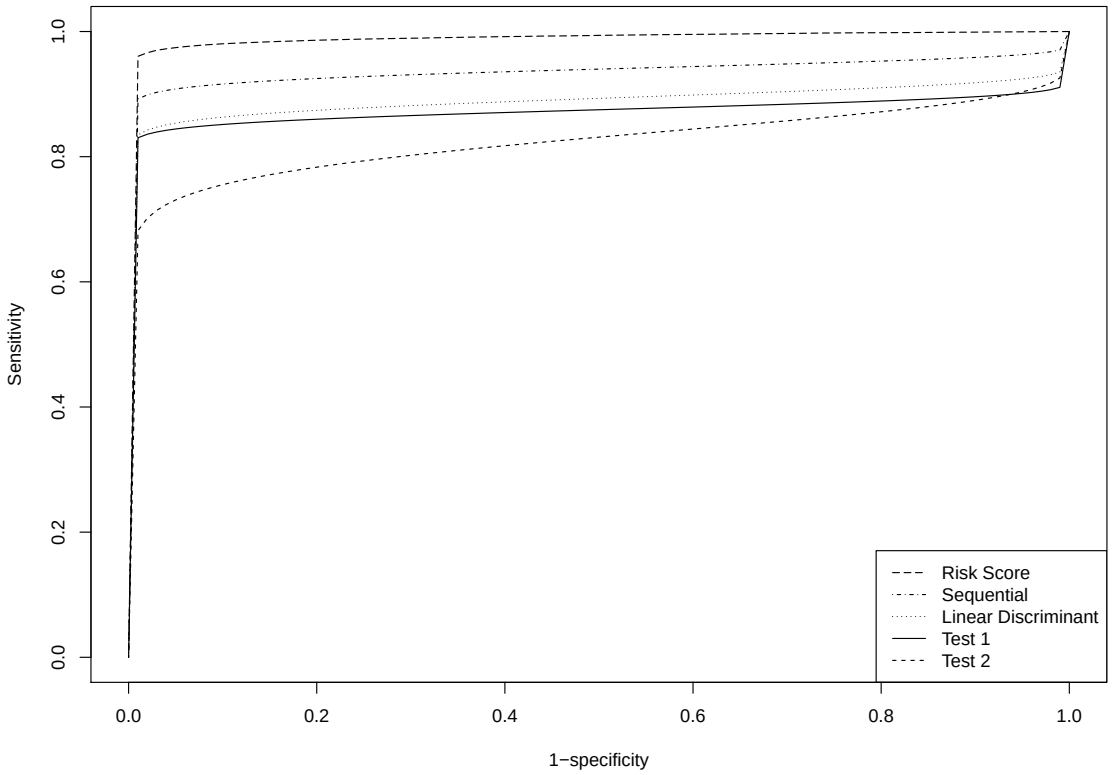

Figure 4.4: ROC curves for Test 2, Test 1, the optimal linear, sequential and risk score composite tests (from bottom to top).

4.5.2 Simulation Study

In this section, we conduct simulation studies to assess the performance of the three methods.

In the first study, we generate two diagnostic markers for the case group from a bivariate normal distribution $N(\mu_1, V_1)$ of

$$\mu_1 = (3.77, 1.51)^T \text{ and } V_1 = \begin{pmatrix} 3.97 & 0.69 \\ 0.69 & 1.42 \end{pmatrix},$$

and the markers for the non-case group from a bivariate normal distribution $N(\mu_0, V_0)$ of

$$\mu_0 = (2, 0.81)^T \text{ and } V_0 = \begin{pmatrix} 0.68 & 0.03 \\ 0.03 & 0.18 \end{pmatrix}.$$

The values of the parameters in the model are selected to mimic the ELISA data example. The MLE of (μ_1, V_1, μ_0, V_0) obtained from the example data is fairly small, so we amplify the scale of the values but keep the structure of the data as shown in Figure 4.5.

Because it is important to evaluate the effects of sample size and latent case prevalence on the performance of the methods. Three sample sizes (100, 200 and 400) for both groups with case prevalence of 0.25 and 0.5, respectively, are examined. For each combination of the sample size and case prevalence, 1000 Monte-Carlo samples are generated from the two designed bivariate normal distributions, $N(\mu_1, V_1)$ and $N(\mu_0, V_0)$. For inference, the standard error of the estimated sensitivity is estimated via the nonparametric bootstrap method aforementioned and its 95% Wald confidence interval is constructed using the bootstrap standard error as well.

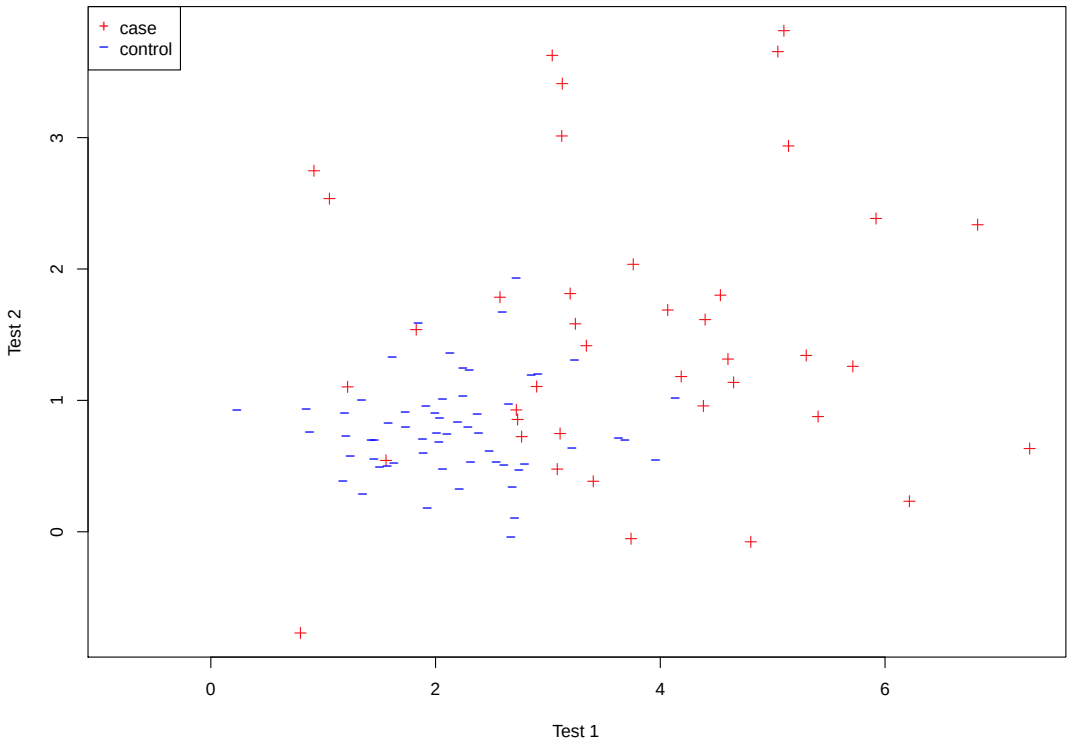

Figure 4.5: Scatter plot of a simulated data set of 100 subjects from the mixture model of two bivariate normal distributions with case prevalence as 0.5.

The exact sensitivity at a given specificity can be determined for the composite tests given the bivariate normal distributions F_1 and F_0. For the linear composite test, the exact sensitivity is calculated via (4.4) and for the sequential composite test, the exact sensitivity is obtained via solving the nonlinear system (4.21). For the risk score composite test, it is not straightforward to derive the distribution functions of the likelihood ratio H_1 and H_0. Hence the calculation of the exact sensitivity is achieved through the Monte-Carlo method described in 4.2.3.

For each simulated sample, we can similarly estimate model-based specificity and sensitivity with F_1 and F_0 replaced by their corresponding ML estimates of $\hat{F}_1$ and $\hat{F}_0$ and hence the bias, root mean square error (RMSE) and coverage probability of the 95% confidence interval (CP) can be estimated with 1000 replicative Monte-Carlo samples. In addition, since the true case status is known for the simulated data, the empirical specificity (Espe) and sensitivity (Esen) can be also directly computed for the three composite tests. Table 4.1 and 4.2 summarize this simulation study based on the 1000 Monte-Carlo samples.

As seen in the table, the composite tests are generally better than the best individual test (Test 1) with the optimal risk score composite test having the largest sensitivity for a given specificity. Though slightly less sensitive than the optimal risk score composite test, the optimal sequential composite test has a better sensitivity than the optimal linear composite test. The model-based estimated sensitivities appear to be unbiased with decreasing RMSE and right coverage probability when sample size increases, which justifies the asymptotic normality properties stated in Section 4.3. By comparing the RMSE, it can be also inferred that the estimated sensitivities may be more accurate when the case prevalence increases. Moreover, the empirical specificity and sensitivity for the composite tests are all in-line with their designed values for the simulation study, indicating these tests working properly.

The second simulation study is designed to evaluate the robustness of the tests against the normality assumption. Since the multivariate normal distribution is a special form of the Gaussian copula [41], with marginally normal distributed

Table 4.1: Summary of the simulation study at the given specificities based on 1000 Monte-Carlo samples with different total sample size N and the case prevalence of $\pi = 0.25$.

$\pi = 0.25$										
Test 1 alone										
Specificity=80%, Sensitivity=0.705						Specificity=90%, Sensitivity=0.640				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.038	0.145	0.701	0.797	0.836	0.044	0.138	0.634	0.898	0.852
200	0.030	0.113	0.710	0.799	0.874	0.032	0.099	0.641	0.900	0.900
400	0.010	0.077	0.705	0.799	0.945	0.011	0.070	0.640	0.899	0.948
Optimal Linear Composite Test										
Specificity=80%, Sensitivity=0.738						Specificity=90%, Sensitivity=0.682				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.058	0.149	0.742	0.793	0.741	0.064	0.141	0.685	0.892	0.771
200	0.039	0.113	0.743	0.797	0.844	0.042	0.100	0.686	0.897	0.867
400	0.014	0.074	0.739	0.798	0.942	0.014	0.066	0.684	0.898	0.946
Optimal Sequential Composite Test										
Specificity=80%, Sensitivity=0.802						Specificity=90%, Sensitivity=0.750				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.037	0.119	0.794	0.794	0.777	0.044	0.116	0.741	0.895	0.822
200	0.026	0.092	0.800	0.798	0.856	0.029	0.084	0.747	0.899	0.872
400	0.009	0.063	0.800	0.798	0.941	0.010	0.059	0.749	0.899	0.947
Optimal Risk Score Composite Test										
Specificity=80%, Sensitivity=0.864						Specificity=90%, Sensitivity=0.809				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.033	0.087	0.843	0.790	0.765	0.045	0.080	0.787	0.891	0.781
200	0.020	0.068	0.853	0.798	0.867	0.027	0.060	0.798	0.899	0.878
400	0.007	0.050	0.861	0.799	0.943	0.011	0.045	0.807	0.898	0.945

Table 4.2: Summary of the simulation study at the given specificities based on 1000 Monte-Carlo samples with different total sample size N and the case prevalence of $\pi = 0.5$.

$\pi = 0.5$										
Test 1 alone										
Specificity=80%, Sensitivity=0.705						Specificity=90%, Sensitivity=0.640				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.031	0.120	0.698	0.800	0.882	0.034	0.111	0.630	0.896	0.903
200	0.018	0.084	0.708	0.799	0.929	0.019	0.072	0.642	0.899	0.939
400	0.006	0.065	0.704	0.800	0.957	0.007	0.060	0.637	0.900	0.961
Optimal Linear Composite Test										
Specificity=80%, Sensitivity=0.738						Specificity=90%, Sensitivity=0.682				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.044	0.122	0.732	0.792	0.818	0.048	0.111	0.674	0.891	0.839
200	0.023	0.081	0.741	0.795	0.917	0.024	0.067	0.684	0.896	0.932
400	0.009	0.061	0.737	0.799	0.959	0.010	0.054	0.681	0.898	0.960
Optimal Sequential Composite Test										
Specificity=80%, Sensitivity=0.802						Specificity=90%, Sensitivity=0.750				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.029	0.098	0.788	0.795	0.848	0.034	0.091	0.730	0.893	0.894
200	0.016	0.066	0.800	0.796	0.927	0.018	0.056	0.747	0.896	0.951
400	0.007	0.053	0.800	0.799	0.947	0.007	0.048	0.747	0.898	0.961
Optimal Risk Score Composite Test										
Specificity=80%, Sensitivity=0.863						Specificity=90%, Sensitivity=0.811				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.031	0.076	0.836	0.786	0.765	0.039	0.064	0.780	0.881	0.788
200	0.014	0.054	0.855	0.794	0.908	0.017	0.043	0.800	0.891	0.920
400	0.007	0.045	0.859	0.798	0.932	0.007	0.040	0.805	0.896	0.941

random variables, we generate the data from a mixture of two Gaussian copulas with the same correlation in each group as in the previous simulation study. However, the marginal distributions of the two markers are set to be Student-t with 4 degrees of freedom and scaled to have the same means and variances as in the previous simulation study. Although the data have the same means and variances as in the first study, the distribution of the diagnostic markers are misspecified from the mixture of bivariate normal distribution. Figure 4.6 displays one data set of 100 subjects with the case prevalence 0.5 that also appears similar to the situation presented in the motivating example.

The data are still fitted using the mixture of two bivariate normal distributions and the parallel results are summarized in Table 4.3 and 4.4. We note that, except for the optimal linear composite, the exact sensitivities for the individual test (Test 1), both the optimal risk score and optimal sequential composite tests are still able to be determined based on the true Student-t distributions, F_0 and F_1. (4.4) is not applicable to calculate the exact sensitivity for the optimal linear composite test due to the violation of the normality assumption.

It is interesting to observe that although the (normal) model-based estimated sensitivity from Test 1 alone is seemingly biased from the designed value, the biases of both the optimal risk score and sequential composite methods tend to be negligible. The empirical specificity and sensitivity of the composite tests are also close to their corresponding designed values indicating the composite tests are fairly robust in terms of accurately classifying the study subjects even though the underlying statistical

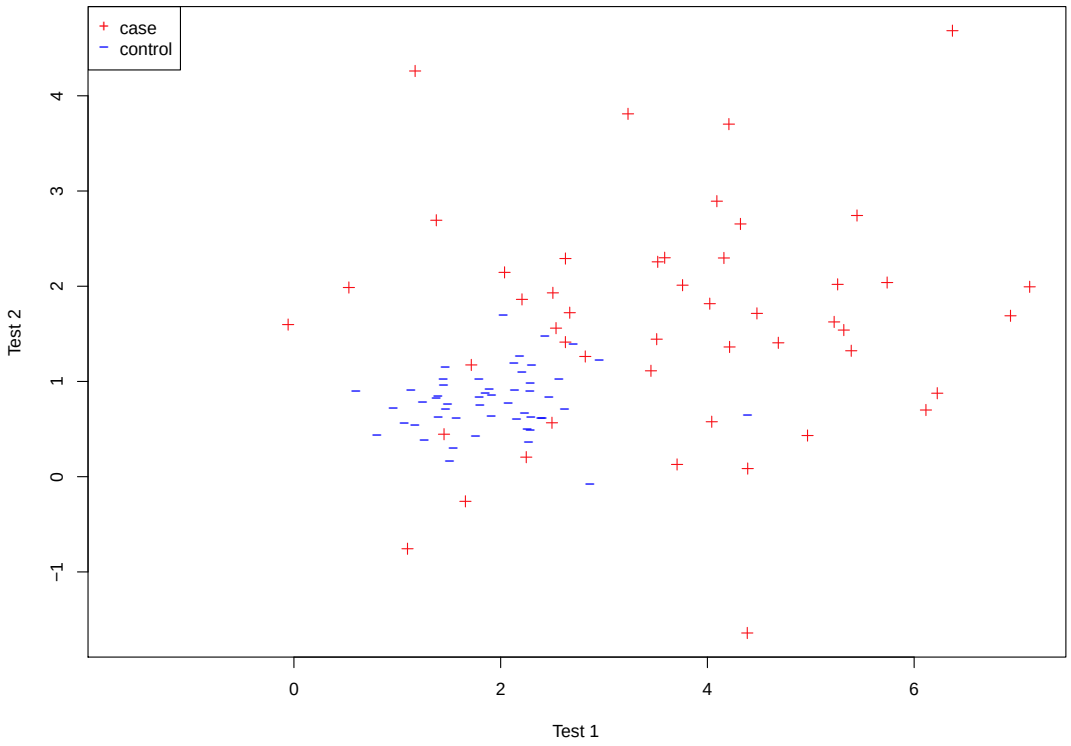

Figure 4.6: Scatter plot of a simulated data set of 100 subjects from the Gaussian copula model with the Student-t marginal with case prevalence of 0.5.

model for developing the tests are misspecified.

While the estimated sensitivities based on the composite tests appear consistent, the asymptotic normality of the estimates are no longer valid as indicated by the coverage probability of the confidence interval. This is not surprising, because the asymptotic normality of the estimated model parameters depends on the assumption that the underlying statistical model is correctly specified. Therefore, it should be cautious in making distributional inference about the estimated sensitivity. The risk

score composite test is relatively more robust compared to the sequential composite test. This can be explained by their different estimating procedures. The sensitivity of the risk score composite test is estimated by the Monte-Carlo method. Instead of estimating the sensitivity directly from the model as for the sequential composite test, the sensitivity is estimated by simulating random samples from the estimated bivariate normal distributions for the risk score composite test. In this simulation study, the bivariate normal distribution is not very different from the true distribution, Gaussian copula with student-t margins, so the inference is not biased much for the risk score composite test.

In the simulation studies, although it is less accurate in classification than the optimal risk score composite test, the optimal sequential composite test does not require all subjects to take both tests. As listed in Table 4.5, when the marginal distribution of the two assays are normal, using the optimal sequential rule, around 27% of subjects in the sample only needs to take one test to obtain the diagnosis of the case, which prevalence is 25%. This percentage is higher, around 40%, for the case with a prevalence of 50%. The percentages are similar for the data with student-t marginal distributions since the patterns of the data are similar.

4.6 Summary of the Frequentists' Classification Methods

In this chapter, we have overviewed and extended some existing methods for combining multiple quantitative tests to classify subjects without a gold standard. In addition, we also implement an alternative classification method based on sequen-

Table 4.3: Summary of the simulation study at the given specificities based on 1000 Monte-Carlo samples of Gaussian copula with the Student-t marginal distributions under different total sample size N and the case prevalence $\pi = 0.25$.

$\pi = 0.25$										
Test 1 alone										
Specificity=80%, Sensitivity=0.783						Specificity=90%, Sensitivity=0.716				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	-0.106	0.134	0.775	0.806	0.809	-0.092	0.170	0.717	0.890	0.854
200	-0.123	0.118	0.781	0.808	0.586	-0.110	0.163	0.724	0.893	0.663
400	-0.131	0.113	0.778	0.810	0.200	-0.118	0.161	0.720	0.893	0.295
Optimal Linear Composite Test										
Specificity=80%, Sensitivity=NA†						Specificity=90%, Sensitivity=NA†				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	NA	NA	0.790	0.790	NA	NA	NA	0.742	0.877	NA
200	NA	NA	0.796	0.796	NA	NA	NA	0.747	0.884	NA
400	NA	NA	0.794	0.798	NA	NA	NA	0.747	0.886	NA
Optimal Sequential Composite Method										
Specificity=80%, Sensitivity=0.856						Specificity=90%, Sensitivity=0.786				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	-0.061	0.094	0.853	0.780	0.894	-0.034	0.119	0.807	0.864	0.947
200	-0.079	0.077	0.859	0.784	0.714	-0.052	0.110	0.812	0.868	0.940
400	-0.085	0.070	0.858	0.786	0.321	-0.059	0.106	0.814	0.868	0.744
Optimal Risk Score Composite Test										
Specificity=80%, Sensitivity=0.896						Specificity=90%, Sensitivity=0.830				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.008	0.064	0.898	0.739	0.896	0.034	0.058	0.850	0.830	0.807
200	-0.002	0.052	0.904	0.739	0.953	0.022	0.049	0.859	0.829	0.860
400	-0.004	0.047	0.907	0.739	0.944	0.020	0.043	0.866	0.827	0.815

† The sensitivity of the optimal linear composite test cannot be calculated directly from the true model because the joint distribution of two tests in each group is not multivariate normal.

Table 4.4: Summary of the simulation study at the given specificities based on 1000 Monte-Carlo samples of Gaussian copula with the Student-t marginal distributions under different total sample size N and the case prevalence $\pi = 0.5$.

$\pi = 0.5$										
Test 1 alone										
Specificity=80%, Sensitivity=0.783						Specificity=90%, Sensitivity=0.716				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	-0.092	0.121	0.739	0.835	0.902	-0.081	0.159	0.668	0.909	0.942
200	-0.099	0.107	0.746	0.838	0.727	-0.088	0.147	0.678	0.910	0.844
400	-0.105	0.093	0.751	0.839	0.347	-0.093	0.141	0.684	0.911	0.564
Optimal Linear Composite Test										
Specificity=80%, Sensitivity=NA†						Specificity=90%, Sensitivity=NA†				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	NA	NA	0.757	0.824	NA	NA	NA	0.697	0.901	NA
200	NA	NA	0.762	0.828	NA	NA	NA	0.705	0.903	NA
400	NA	NA	0.770	0.832	NA	NA	NA	0.715	0.907	NA
Optimal Sequential Composite Method										
Specificity=80%, Sensitivity=0.856						Specificity=90%, Sensitivity=0.786				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	-0.049	0.085	0.819	0.811	0.952	-0.022	0.106	0.758	0.885	0.961
200	-0.056	0.071	0.826	0.810	0.899	-0.029	0.095	0.770	0.884	0.986
400	-0.059	0.056	0.836	0.813	0.612	-0.032	0.086	0.782	0.886	0.968
Optimal Risk Score Composite Test										
Specificity=80%, Sensitivity=0.895						Specificity=90%, Sensitivity=0.831				
N	Bias	RMSE	ESen	ESpe	CP	Bias	RMSE	ESen	ESpe	CP
100	0.014	0.067	0.859	0.756	0.822	0.041	0.057	0.802	0.837	0.704
200	0.008	0.057	0.868	0.757	0.912	0.033	0.047	0.816	0.837	0.773
400	0.004	0.052	0.883	0.757	0.920	0.028	0.041	0.834	0.839	0.740

† The sensitivity of the optimal linear composite test cannot be calculated directly from the true model because the joint distribution of two tests in each group is not multivariate normal.

Table 4.5: Average of percentage of subjects taking only one test in simulation studies (%) under different sample sizes, case prevalences, and pre-specified specificities.

	π =0.25				π =0.5			
	Spe=80%		Spe=90%		Spe=80%		Spe=90%	
	Normal	t	Normal	t	Normal	t	Normal	t
N=100	28.1	29.6	20.9	23.4	40.7	41.6	33.9	35.6
N=200	27.5	28.8	20.4	23.0	40.6	41.2	33.9	35.4
N=400	27.1	28.3	20.2	22.7	40.2	41.2	33.5	35.5

tial tests under no gold standard circumstance. All the classification methods are developed under the model assumption that the diagnostic markers for the tests for both case and non-case populations coming from multivariate normal distribution. A mixture model of two multivariate normal distributions is fitted to the unclassified data and the optimal decision rule for each method is determined with the fitted model. All the methods are illustrated with the real data set for two ELISA tests on E2 antibody from an HIV study and their performance is assessed through simulation studies. For the data presented in Figure 4.5 and Figure 4.6, the simulation studies demonstrate that the optimal risk score composite test has the most discriminant power to distinguish case from non-case in view of AUC and is most sensitive among the three composite tests. The optimal sequential composite test, though little less sensitive than its risk score counterpart, outperforms the optimal linear composite test and has an additional advantage of engaging less tests. This advantage of the sequential method is particularly desired when the tests are costly or not applicable to all study subjects in some applications. The optimality of the tests in this article

is purely based on the classification accuracy without considering risk or cost associated with the tests. If the risk or cost ought to be considered for determining an optimal decision rule in some applications, the risk score composite test may not be the optimal test among the three.

The optimal sequential composite test is statistically equivalent to the implementation of a sequence of tests discussed by Thompson [64], who discussed the statistical properties of the sequential composite test without addressing the choice of the optimal threshold and statistical inferences on the estimated ROC curves. In addition, she was not concerned about the issue of no gold standard.

There are multiple ways to design a sequential composite test according to the data that reflect the study nature. The sequential test designed in this paper is evidently supported by the biological mechanism in E2 antibody. For a blood sample, it is known that each ELISA test returns a value reflecting the concentration of the E2 antibody and a larger value implies a greater chance of the antibody presence. Because each test uses a different protein preparation for the antibody and the binding site is occasionally inaccessible even when the antibody is there, it is anticipated that the variance of the ELISA tests for the blood samples with the antibody is greater than the variance for the blood samples without the antibody. Therefore, a cluster of small values of the ELISA tests represents the group of absence of E2 antibody. So the sequential method starting with classification for case is the right option. If, for some applications, there is a small cluster of points present in the upper right corner of the scatter plot that likely represents the group of case, it would be more reasonable

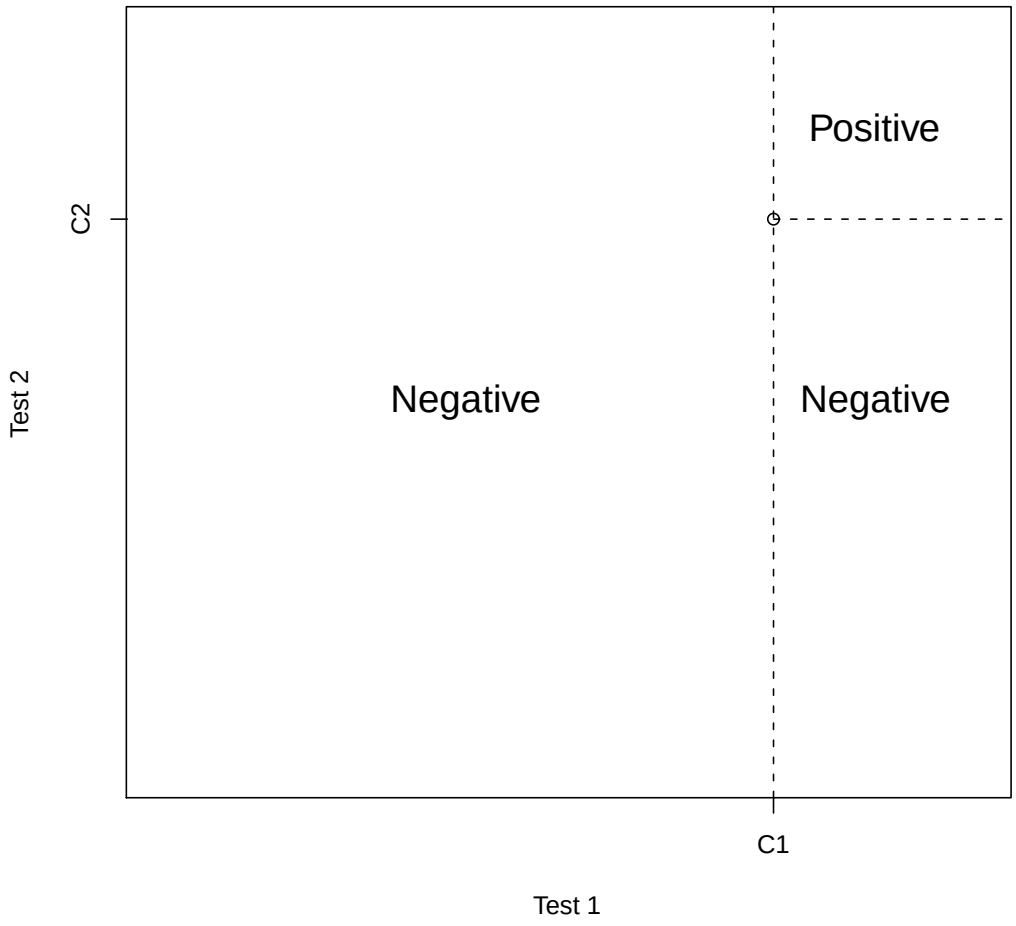

Figure 4.7: Illustration of the alternative 2-cutoff sequential classification method.

to start with classification of non-case as shown in Figure 4.7 for the sequential test.

CHAPTER 5
DISCUSSION AND FUTURE WORK

5.1 Conclusions

In summary, we have investigated classification methods with multiple diagnostic tests under no gold standard. In Chapter 3, the problem is viewed from the Bayesian perspective. Motivated from the ELISA data, we propose a two-level latent model based on the biological background and obtain the Bayesian estimates of parameters by incorporating the available prior information on the prevalence and binding probability for each test. The posterior positive probability is estimated as the classifier and the corresponding Bayesian classification on the motivating data set is consistent with the test mechanisms. The model can be extended to accommodate problems with more than two tests, or problems from various testing mechanisms, such as PCR tests.

As stated in Chapter 3, the parameter estimation can also be obtained via ML approach, and accordingly, the classifier is estimated by treating the parameters as a fixed unknown vector instead of random variables. Although on the motivating data set, the classification result is not quite different from the Bayesian classification result, there are twelve parameters in the model to be estimated from the data of size 100, which is relatively small to make asymptotic inferences based on MLE's. The Bayesian approach is preferred in Chapter 3 because there does exist some prior information on parameters, such as prevalence, to obtain a more stable estimation.

Of course, for a different population, a different prior distribution will be needed and the parameters may be different. For example, because HIV and GBV-C share the same modes of infection, the prevalence of the E2 antibody in an HIV-infected population is high, about 50%, whereas the prevalence in the general population of blood donors is much lower, about 5%. The prior distribution on the prevalence should be different for the two populations. Caution should also be used in using prior information from one population to extrapolate to a different population in constructing the prior. It is possible that the underlying level of what is being tested for (concentration of antibody in an ELISA, or concentration of virus in a PCR) is different in different infected populations and therefore the parameters of the mixture components involving true positive measurements will differ. It may be reasonable to assume that the true negative responses on a test are similar across populations and perhaps also the probability of a false negative. Different prior distributions for different populations are easily incorporated. Different populations may share the same model structure, but with different parameter values (perhaps, for example, with a hierarchical structure between populations).

The methods studied in Chapter 4 provide solutions to the problem from the Frequentists' perspective. Not only applied to the ELISA testing problem, Chapter 4 also focuses on a more general problem, which is to make diagnostic decisions based on multiple tests without a gold standard. We first overview and extend existing optimal linear combination method and optimal risk score method for combining and evaluating multiple tests to incorporate the no gold standard situation. We also

study a sequential diagnostic method, which has its own merits when the risk or cost of the tests are considered in the decision rule. Indicated in the motivating example and simulation study, the optimal linear combination is easy to implement and make inference on the estimate sensitivity for the classification. The optimal risk score composite test, with no doubt, has the best discriminant capability among the three composite tests, yet it is not easy to performed if the parametric assumption is released. Moreover, even when a gold standard exists, using the equivalent risk score still needs some caution when choosing the mean function in the logistic regression model for estimating the risk score. To illustrate, for one simulated data set with 400 subjects from the normal model, the classification at 80% specificity using the likelihood ratio is 98.5%, while is only 87.8% using the risk score estimated from the logistic regression with a linear mean function. The sequential composite test provides an alternative procedure which outperforms the optimal linear composite test and does not lose much sensitivity from the optimal risk score composite test. When applying the classification rules generated from the existing data to a new data set, the optimal risk score rule may identify the subjects with low results on both tests but outside the elliptical classification region as case positive, while the optimal linear rule and the optimal sequential rule do not have such issue.

To conclude, guidelines to the researchers applying these methods in practice would be: the Bayesian classification method is straightforward when there is prior knowledge of some parameters, hence it can be employed as a start. But one drawback of the Bayesian method is that it is not easy to evaluate the test uniformly (ROC curve

is difficulty to estimate). The optimal linear composite test can also serve as a starting point. The optimal risk score composite test is recommended as the first choice due to its optimality of distinguishing case from non-case with the most sensitivity. Besides, learned from the simulation studies, it has the most robust inference on the estimated sensitivity if that is also an study interest. But when the tests are expensive, or not applicable to all study subjects, the sequential composite test could be a better choice with an advantage of engaging less tests.

5.2 Future Work

All the methods in Chapter 3 and Chapter 4 are developed based on some parametric assumptions. The Bayesian two-level latent model assumes that the test results under every combination of the accessibility of the binding sites for both tests follow a multivariate normal distribution, and the three Frequentists' methods are derived on the basis of mixture model assuming that the test results in both case and control populations are multivariate normal. Multivariate normal distribution is a widely used distribution for continuous variables in practice, but it is difficult to check the validity of this distributional assumption on the unclassified data. In application, one may consider possible monotone transformation to convert the original data to normal-like multivariate data and apply the methods to the transformed data. But this still cannot guarantee the normality. Alternatively, one may consider nonparametric approaches. Bayesian nonparametric modeling using a Dirichlet process mixture based on a multivariate Gaussian kernel may be a substitute in the

Bayesian classification method. For the Frequentists' methods, the mixing probability distributions can be estimated nonparametrically according to Hall and Zhou [21]. But their estimation method is very complicated to implement and is restrictive. The tensor spline-based sieve maximum likelihood estimation [75] of the multivariate distribution function is a compromise to the Hall-Zhou's nonparametric estimation of mixture distribution. Although the optimal risk score and sequential composite tests can still be computed with the tensor spline-based sieve estimation, the numerical implementation of the tests is much more demanding and challenging than the multivariate normal model. Moreover, the statistical properties of the tests using the spline-based model are much harder to study. Study and implementation of the spline-based model for releasing the normality assumption in this context are still an open problem for further investigation.

The sequential classification method can be similarly designed for the situation with more than two tests, it is, however, a mathematically challenging problem because finding the optimal cut-off values cannot be equivalently converted to the problem of solving a nonlinear system as it does for the two-test case. The grid search is a straightforward option but it can be very numerically inefficient, especially for high dimensional data. There is still a space for improving the numerical algorithm in order to accommodate an arbitrary number of tests.

Finally, in both approaches, both the statistical models assume a constant variance-covariance structure for each cluster; however, this assumption may be violated in some cases. For instance, if the testing target is a certain virus in blood

samples, the virus can change over time or be affected by other factors. Therefore, how to adapt the variance-covariance structure over time, and how to incorporate other covariates in the model are also a direction for future work.

APPENDIX A
WINBUGS AND R PROGRAMS FOR CHAPTER 3

```
# ========================================================================
#      WinBUGS function for the Bayesian analysis of the example data
# ========================================================================

# Notations:
#
# phi: the prevalence of the E2 antibody
# phi1 and phi2: the accessible probability of the binding site for each test
# mu1N, mu2N, mu1P, mu2P: the means of the normal marginal distributions
# mud1=log(mu1P - mu1N); mud2=log(mu2P - mu2N)
# sigma2_1N, sigma2_2N, sigma2_1P, sigma2_2P:
#  the variances of the normal marginal distributions
# tau1N, tau2N, tau1P, tau2P:
#  the precisions of the normal marginal distributions
# rho: the positive correlation of the two tests results
#  when the antibody is present and both test bind
# N: the sample size
# y: the data matrix (N*2)
# C: vector of indicators for the 4 mixture elements
# p: vector of 4 mixture probabilities
#
# Seven prior distributions are all listed.
# Comment out the unnecessary priors when use.
# The model is saved as ‘‘model.txt’’

model
{
# prior A
phi ~ dbeta(5,5)
phi1 ~ dbeta(18, 2)
phi2 ~ dbeta(18, 2)

mu1N ~ dnorm(0, 1.0E-2)
mu2N ~ dnorm(0, 1.0E-2)
mud1 ~ dnorm(0, 1.0E-4)
mud2 ~ dnorm(0, 1.0E-4)
mu1P <- mu1N + exp(mud1)
mu2P <- mu2N + exp(mud2)

tau1N ~ dgamma(0.01, 0.01)
```

```
tau2N ~ dgamma(0.01, 0.01)
tau1P ~ dgamma(0.01, 0.01)
tau2P ~ dgamma(0.01, 0.01)
sigma2_1N <- 1/tau1N
sigma2_1P <- 1/tau1P
sigma2_2N <- 1/tau2N
sigma2_2P <- 1/tau2P

rho ~ dunif(0,1)

# prior B
phi ~ dbeta(5,5)
phi1 ~ dbeta(2, b)
phi2 ~ dbeta(2, b)
b <- 2/9

mu1N ~ dnorm(0, 1.0E-2)
mu2N ~ dnorm(0, 1.0E-2)
mud1 ~ dnorm(0, 1.0E-4)
mud2 ~ dnorm(0, 1.0E-4)
mu1P <- mu1N + exp(mud1)
mu2P <- mu2N + exp(mud2)

tau1N ~ dgamma(0.01, 0.01)
tau2N ~ dgamma(0.01, 0.01)
tau1P ~ dgamma(0.01, 0.01)
tau2P ~ dgamma(0.01, 0.01)
sigma2_1N <- 1/tau1N
sigma2_1P <- 1/tau1P
sigma2_2N <- 1/tau2N
sigma2_2P <- 1/tau2P

rho ~ dunif(0,1)

# prior C
phi ~ dbeta(5,5)
phi1 ~ dbeta(1, 1)
phi2 ~ dbeta(1, 1)

mu1N ~ dnorm(0, 1.0E-2)
mu2N ~ dnorm(0, 1.0E-2)
mud1 ~ dnorm(0, 1.0E-4)
mud2 ~ dnorm(0, 1.0E-4)
mu1P <- mu1N + exp(mud1)
```

```
mu2P <- mu2N + exp(mud2)

tau1N ~ dgamma(0.01, 0.01)
tau2N ~ dgamma(0.01, 0.01)
tau1P ~ dgamma(0.01, 0.01)
tau2P ~ dgamma(0.01, 0.01)
sigma2_1N <- 1/tau1N
sigma2_1P <- 1/tau1P
sigma2_2N <- 1/tau2N
sigma2_2P <- 1/tau2P

rho ~ dunif(0,1)

# prior D
phi ~ dbeta(5,5)

Const <- 10000
b <- 20/9

zero1 <- 0
phi1 ~ dflat()
theta1 <- -log(b*phi1*step(phi1)*step(0.9-phi1)+
              (20-20*phi1)*step(phi1-0.9)*step(1-phi1)) + Const
zero1 ~ dpois(theta1)

zero2 <- 0
phi2 ~ dflat()
theta2 <- -log(b*phi2*step(phi2)*step(0.9-phi2)+
              (20-20*phi2)*step(phi2-0.9)*step(1-phi2)) + Const
zero2 ~ dpois(theta2)

mu1N ~ dnorm(0, 1.0E-2)
mu2N ~ dnorm(0, 1.0E-2)
mud1 ~ dnorm(0, 1.0E-4)
mud2 ~ dnorm(0, 1.0E-4)
mu1P <- mu1N + exp(mud1)
mu2P <- mu2N + exp(mud2)

tau1N ~ dgamma(0.01, 0.01)
tau2N ~ dgamma(0.01, 0.01)
tau1P ~ dgamma(0.01, 0.01)
tau2P ~ dgamma(0.01, 0.01)
sigma2_1N <- 1/tau1N
sigma2_1P <- 1/tau1P
```

```
sigma2_2N <- 1/tau2N
sigma2_2P <- 1/tau2P

rho ~ dunif(0,1)

# prior E
phi ~ dbeta(1,1)
phi1 ~ dbeta(18, 2)
phi2 ~ dbeta(18, 2)

mu1N ~ dnorm(0, 1.0E-2)
mu2N ~ dnorm(0, 1.0E-2)
mud1 ~ dnorm(0, 1.0E-4)
mud2 ~ dnorm(0, 1.0E-4)
mu1P <- mu1N + exp(mud1)
mu2P <- mu2N + exp(mud2)

tau1N ~ dgamma(0.01, 0.01)
tau2N ~ dgamma(0.01, 0.01)
tau1P ~ dgamma(0.01, 0.01)
tau2P ~ dgamma(0.01, 0.01)
sigma2_1N <- 1/tau1N
sigma2_1P <- 1/tau1P
sigma2_2N <- 1/tau2N
sigma2_2P <- 1/tau2P

rho ~ dunif(0,1)

# prior F
phi ~ dbeta(5,5)
phi1 ~ dbeta(18, 2)
phi2 ~ dbeta(18, 2)

mu1N ~ dnorm(0, 1.0E-2)
mu2N ~ dnorm(0, 1.0E-2)
mud1 ~ dnorm(0, 1.0E-4)
mud2 ~ dnorm(0, 1.0E-4)
mu1P <- mu1N + exp(mud1)
mu2P <- mu2N + exp(mud2)

sigma1N ~ dunif(0, 100)
sigma2N ~ dunif(0, 100)
sigma1P ~ dunif(0, 100)
sigma2P ~ dunif(0, 100)
```

```
sigma2_1N <- pow(sigma1N,2)
sigma2_1P <- pow(sigma1P,2)
sigma2_2N <- pow(sigma2N,2)
sigma2_2P <- pow(sigma2P,2)

tau1N <- 1/sigma2_1N
tau2N <- 1/sigma2_2N
tau1P <- 1/sigma2_1P
tau2P <- 1/sigma2_2P

rho ~ dunif(0,1)

# prior G
phi ~ dbeta(1,1)
phi1 ~ dbeta(1, 1)
phi2 ~ dbeta(1, 1)

mu1N ~ dnorm(0, 1.0E-2)
mu2N ~ dnorm(0, 1.0E-2)
mud1 ~ dnorm(0, 1.0E-4)
mud2 ~ dnorm(0, 1.0E-4)
mu1P <- mu1N + exp(mud1)
mu2P <- mu2N + exp(mud2)

sigma1N ~ dunif(0, 100)
sigma2N ~ dunif(0, 100)
sigma1P ~ dunif(0, 100)
sigma2P ~ dunif(0, 100)
sigma2_1N <- pow(sigma1N,2)
sigma2_1P <- pow(sigma1P,2)
sigma2_2N <- pow(sigma2N,2)
sigma2_2P <- pow(sigma2P,2)

tau1N <- 1/sigma2_1N
tau2N <- 1/sigma2_2N
tau1P <- 1/sigma2_1P
tau2P <- 1/sigma2_2P

rho ~ dunif(0,1)

# likelihood of the ith data
for ( i in 1:N)
{
  y[i, 1:2 ] ~ dmnorm(mu[ C[i], 1:2 ], T[ C[i], 1:2 , 1:2 ] )
```

```
  C[i] ~ dcat(p[ 1:4])
}

p[1] <- phi * phi1 * phi2
p[2] <- phi * phi1 * (1 - phi2)
p[3] <- phi * (1 - phi1) * phi2
p[4] <- phi * (1 - phi1) * (1 - phi2) + 1-phi

mu[1, 1 ] <- mu1P
mu[1, 2 ] <- mu2P
mu[2, 1 ] <- mu1P
mu[2, 2 ] <- mu2N
mu[3, 1 ] <- mu1N
mu[3, 2 ] <- mu2P
mu[4, 1 ] <- mu1N
mu[4, 2 ] <- mu2N

sigma1[1, 1] <- 1/tau1P
sigma1[1, 2] <- rho * pow(tau1P * tau2P, -0.5)
sigma1[2, 1] <- rho * pow(tau1P * tau2P, -0.5)
sigma1[2, 2] <- 1/tau2P
T[1, 1:2, 1:2 ] <- inverse(sigma1[ , ])

sigma2[1, 1] <- 1/tau1P
sigma2[1, 2] <- 0
sigma2[2, 1] <- 0
sigma2[2, 2] <- 1/tau2N
T[2, 1:2, 1:2 ] <- inverse(sigma2[ , ])

sigma3[1, 1] <- 1/tau1N
sigma3[1, 2] <- 0
sigma3[2, 1] <- 0
sigma3[2, 2] <- 1/tau2P
T[3, 1:2, 1:2 ] <- inverse(sigma3[ , ])

sigma4[1, 1] <- 1/tau1N
sigma4[1, 2] <- 0
sigma4[2, 1] <- 0
sigma4[2, 2] <- 1/tau2N
T[4, 1:2, 1:2 ] <- inverse(sigma4[ , ])
}

# ============================================================================
#      Calling WinBUGS function in R
```

```
# ========================================================================

library(R2WinBUGS)
library(mvtnorm)

# N = sample size
# Y: N*2 data matrix
data <- list(
                N = 100,
                y = dput(Y, control="showAttributes")
              )

inits <- function()
{
  list( C = c(2, 3, 1, 4, 4, 2, 4, 2, 1, 2, 2, 2, 1, 3, 1, 4, 3, 3, 2, 2,
2, 4, 3, 2, 4, 2, 3, 2, 2, 3, 1, 3, 2, 1, 3, 3, 1, 3, 2, 2, 4,
3, 2, 3, 1, 1, 2, 1, 4, 4, 3, 4, 4, 3, 3, 2, 3, 1, 2, 3, 3, 1,
2, 2, 1, 3, 3, 1, 4, 3, 4, 4, 4, 4, 3, 3, 2, 1, 2, 2, 2, 1, 2,
2, 4, 2, 1, 1, 4, 2, 4, 1, 1, 1, 3, 4, 4, 3, 2, 1),
     phi = 0.5,phi1 = 0.5,phi2 = 0.5,mu1N = 0,mu2N = 0,mud1 = 0,
       mud2 = 0,tau1N = 1,tau2N = 1,tau1P = 1,tau2P = 1,rho = 0.5)
}

sim <- bugs(data, inits,
            model.file = "model.txt",
            n.iter = 15000, n.chains=1, n.thin=1, n.burnin=5000, digits=5,
            parameters.to.save = c("phi", "phi1", "phi2", "mu1N", "mu2N",
            "mu1P", "mu2P", "sigma2_1N", "sigma2_2N", "sigma2_1P", "sigma2_2P",
            "rho","mud1","mud2"),
            bugs.directory = "C:/Program Files/WinBUGS14/"
            )

# ========================================================================
#     function to calculate posterior positive prob. (PPP)
# ========================================================================

prob.1 <- function(y,para)
# para is the parameters saved from the ``bugs()'' function
  {
    phi <- para[1]
    phi1 <- para[2]
    phi2 <- para[3]
    mu1N <- para[4]
    mu2N <- para[5]
```

```
  mu1P <- para[6]
  mu2P <- para[7]
  var1N <- para[8]
  var2N <- para[9]
  var1P <- para[10]
  var2P <- para[11]
  rho <- para[12]

  p1 <- phi1 * phi2
  p2 <- phi1 * ( 1 - phi2 )
  p3 <- ( 1 - phi1 ) * phi2
  p4 <- ( 1 - phi1 ) * ( 1 - phi2 )

  meanNN <- c(mu1N, mu2N)
  covNN <- diag(c(var1N, var2N))

  meanPN <- c(mu1P, mu2N)
  covPN <- diag(c(var1P, var2N))

  meanNP <- c(mu1N, mu2P)
  covNP <- diag(c(var1N, var2P))

  meanPP <- c(mu1P, mu2P)
  covPP <- matrix(c(var1P, (rho*sqrt(var1P)*sqrt(var2P)),
   (rho*sqrt(var1P)*sqrt(var2P)), var2P),nrow=2)

  f1 <- dmvnorm(y, mean = meanPP, sigma = covPP)
  f2 <- dmvnorm(y, mean = meanPN, sigma = covPN)
  f3 <- dmvnorm(y, mean = meanNP, sigma = covNP)
  f4 <- dmvnorm(y, mean = meanNN, sigma = covNN)

  res <- (phi * (f1 * p1 + f2 * p2 + f3 * p3 + f4 * p4)) /
   (phi * (f1 * p1 + f2 * p2 + f3 * p3 + f4 * p4) + f4 * (1 - phi))
  return (res)
}
```

APPENDIX B
R PROGRAMS FOR CHAPTER 4

```
# ========================================================================
#      Likelihood function
# ========================================================================

library(mvtnorm)

dmvt <- function (x, mu, Sigma, df=Inf, log = FALSE) {
     if (!is.matrix(x))
         x <- rbind(x)
     p <- nrow(Sigma)
     ed <- eigen(Sigma, symmetric = TRUE)
     ev <- ed$values
     if (!all(ev >= -1e-06 * abs(ev[1])))
         stop("'Sigma' is not positive definite")
     ss <- if (!is.matrix(mu)) {
         x - rep(mu, each = nrow(x))
     } else {
         x - mu
     }
     inv.Sigma <- ed$vectors %*% (t(ed$vectors)/ev)
     quad <- 0.5 * rowSums((ss %*% inv.Sigma) * ss)
     fact <- -0.5 * (p * log(2 * pi) + sum(log(ev)))
     if (log)
         as.vector(fact - quad)
     else
         as.vector(exp(fact - quad))
}
fi <- function(x, para)
{
  dmvt(x,para[1:2],matrix(c(para[3],para[4],para[4],para[5]),2,2))
}

f <- function(x, para)
{
  p <- para[1];
  para1 <- para[2:6]; para0 <- para[7:11]
  p * fi(x,para1) + (1-p) * fi(x,para0)
}
```

```
# log-likelihood function
l <- function(data, para)
{
  ft <- f(data, para)

  sum(log(ft))
}

# ========================================================================
#     Binormal ROC function
# ========================================================================

ROC <- function(x,para)
# x = FPR = 1 - specificity
{
  muD <- para[1]; sigmaD <- sqrt(para[2])
  muDb <- para[3]; sigmaDb <- sqrt(para[4])
  a <- (muD-muDb)/sigmaD; b <- sigmaDb/sigmaD

  pnorm(a + b*qnorm(x))
}

# ========================================================================
#      Binormal AUC
# ========================================================================

AUC <- function(para)
{
  muD <- para[1]; sigmaD <- sqrt(para[2])
  muDb <- para[3]; sigmaDb <- sqrt(para[4])
  a <- (muD-muDb)/sigmaD; b <- sigmaDb/sigmaD

  pnorm(a/sqrt(1+b^2))
}

# ========================================================================
#      ROC function for the optimal linear composite test
# ========================================================================

ROC.SL <- function(x,para)
# x = FPR = 1 - specificity
{
  muD <- para[1:2]; muDb <- para[6:7]
  VD <- matrix(c(para[3],para[4],para[4],para[5]),2,2)
```

```
  VDb <- matrix(c(para[8],para[9],para[9],para[10]),2,2)

  a0 <- solve(VD+VDb)%*%(muD-muDb)

  para.lc <- c(t(a0)%*%muD, t(a0)%*%VD%*%a0, t(a0)%*%muDb, t(a0)%*%VDb%*%a0)

  ROC(x,para.lc)
}

# ========================================================================
#      EM algorithm for the MLE
# ========================================================================

EM <- function(dat,init,eps,maxit)

# dat: the data matrix
# init: vector of initial values
# eps: tolerance level
# maxit: maximum number of iterations

{
  post.p <- function(x,para)
  {
    p <- para[1]; para1 <- para[2:6]
    p*fi(x,para1)/f(x,para)
  }

  para.new <- init ; n <- nrow(dat)
  dif <- 1; iter <- 0 ; converge <- 1; error <- 0
  tran <- matrix(c(1,0,0,
                   0,1,0,
                   0,0,0,
                   0,0,1),3,4)

  while (dif>eps)
  {
    para.old <- para.new
    post.p.old <- try(post.p(dat,para=para.old),TRUE)

    if(!inherits(post.p.old,'try-error')){
    p.new <- rep(1/n, n) %*% post.p.old
    mu1.new <- c(t(dat) %*% post.p.old) / (n * p.new)
    V1.new <- c((t(dat)-mu1.new) %*% diag(post.p.old) %*% t(t(dat)-mu1.new)) /
    (n * p.new)
```

```
    mu0.new <- c(t(dat) %*% (1-post.p.old)) / (n * (1-p.new))
    V0.new <- c((t(dat)-mu0.new) %*% diag(1-post.p.old) %*% t(t(dat)-mu0.new)) /
    (n * (1-p.new))

    para.new <- c(p.new, mu1.new, tran%*%V1.new, mu0.new, tran%*%V0.new)

    dif <- max(abs(para.new-para.old))
    iter <- iter + 1
    if (iter > maxit)
    {
      converge <- 0
      cat("More iterations needed.\n")
      break
    }} else {
    error <- 1
    cat("Error when calculating the posterior probabilities.\n")
    break}
  }

  list(par=para.new,convg=converge,error=error)
}

# ========================================================================
#    Sensitivity function of the optimal risk score composite test
# ========================================================================

# Likelihood Ratio Risk Score
LR <- function(z, para)
{
  para1 <- para[2:6]; para0 <- para[7:11]
  fi(z,para1)/fi(z,para0)
}

MC_sen <- function(para, Spes, Nsim)

# para: vector of the parameters in the two-term mixture model
# Spes: pre-specified specificity(s)
# Nsim: number of simulated samples in the Monte-Carlo method

{
  # find c
  para.sim <- para[7:11]
  mu.sim <- para.sim[1:2]
  cov.sim <- matrix(c(para.sim[3],rep(para.sim[4],2),para.sim[5]),2,2)
```

```
  simY <- rmvnorm(Nsim, mu.sim, cov.sim)

  fsim <- LR(simY, para)

  # thresholds
  cf0 <- quantile(fsim, prob=Spes)

  # estimate the sensitivity
  para.sim <- para[2:6]
  mu.sim <- para.sim[1:2]
  cov.sim <- matrix(c(para.sim[3],rep(para.sim[4],2),para.sim[5]),2,2)
  simY <- rmvnorm(Nsim, mu.sim, cov.sim)

  fsim <- LR(simY, para)

  est.sen <- sapply(cf0, Fun <- function(c0) mean(fsim>c0))

  return(list(cf0=cf0, est.sen=est.sen))
}

# ============================================================================
#    Sensitivity function of the optimal risk score composite test
# ============================================================================

# CDF of the bivariate normal distribution
F <- function(x, para)
{
  library(mvtnorm)

  m <- para[1:2]
  v <- matrix(c(para[3:4],para[4:5]),2,2)

  pmvnorm(lower=c(-Inf,-Inf),upper=x,mean=m,sigma=v)
}

# first-order partial derivatives of the CDF
F1 <- function(x, para)
{
  x1 <- x[1]; x2 <- x[2]
  m1 <- para[1]; m2 <- para[2]
  v1 <- para[3]; v2 <- para[5]; rho <- para[4]/sqrt(v1*v2)

  m.new <- m2 + sqrt(v2/v1) * rho * (x1 - m1)
  v.new <- (1 - rho^2) * v2
```

```
  dnorm(x1, m1, sqrt(v1)) * pnorm(x2, m.new, sqrt(v.new))
}

F2 <- function(x, para)
{
  x1 <- x[1]; x2 <- x[2]
  m1 <- para[1]; m2 <- para[2]
  v1 <- para[3]; v2 <- para[5]; rho <- para[4]/sqrt(v1*v2)

  m.new <- m1 + sqrt(v1/v2) * rho * (x2 - m2)
  v.new <- (1 - rho^2) * v1

  pnorm(x1, m.new, sqrt(v.new)) * dnorm(x2, m2, sqrt(v2))
}

# second-order partial derivatives of the CDF
F11 <- function(x,para)
{
  x1 <- x[1]; x2 <- x[2]
  m1 <- para[1]; m2 <- para[2]
  v1 <- para[3]; v2 <- para[5]; rho <- para[4]/sqrt(v1*v2)

  m.new <- m2 + sqrt(v2/v1) * rho * (x1 - m1)
  v.new <- (1 - rho^2) * v2

  res <- -((x1-m1)/v1) * dnorm(x1,m1,sqrt(v1)) * pnorm(x2,m.new,sqrt(v.new)) -
   sqrt(v2/v1)*rho * dnorm(x1,m1,sqrt(v1)) * dnorm(x2,m.new,sqrt(v.new))

  return(res)
}

F22 <- function(x,para)
{
  x1 <- x[1]; x2 <- x[2]
  m1 <- para[1]; m2 <- para[2]
  v1 <- para[3]; v2 <- para[5]; rho <- para[4]/sqrt(v1*v2)

  m.new <- m1 + sqrt(v1/v2) * rho * (x2 - m2)
  v.new <- (1 - rho^2) * v1

  res <- -((x2-m2)/v2) * dnorm(x2,m2,sqrt(v2)) * pnorm(x1,m.new,sqrt(v.new)) -
   sqrt(v1/v2)*rho * dnorm(x2,m2,sqrt(v2)) * dnorm(x1,m.new,sqrt(v.new))
```

```
  return(res)
}

F12 <- function(x,para)
{
  library(mvtnorm)

  m <- para[1:2]
  v <- matrix(c(para[3:4],para[4:5]),2,2)

  dmvnorm(x,mean=m,sigma=v)
}

# the second equation of (4.20) and its first partial derivatives
S <- function(x,p1,p0)
{
  F1(x,p0) * F2(x,p1) - F2(x,p0) * F1(x,p1)
}

S1 <- function(x,p1,p0)
{
  return(F11(x,p0) * F2(x,p1) + F1(x,p0) * F12(x,p1) -
   F12(x,p0) * F1(x,p1) - F2(x,p0) * F11(x,p1))
}

S2 <- function(x,p1,p0)
{
  return(F12(x,p0) * F2(x,p1) + F1(x,p0) * F22(x,p1) -
   F22(x,p0) * F1(x,p1) - F2(x,p0) * F12(x,p1))
}

# the nonlinear system (4.20) and its Jacobian matrix
T <- function(x, p1, p0, pr)
# pr: the p_{0} in (4.20)
{
  y1 <- (F(x,p0) - pr)
  y2 <- S(x,p1,p0)

  c(y1,y2)
}

DT <- function(x, p1, p0)
{
  e11 <- F1(x,p0); e12 <- F2(x,p0)
```

```
  e21 <- S1(x,p1,p0); e22 <- S2(x,p1,p0)

  matrix(c(e11,e21,e12,e22),2,2)
}

# function of solving the optimal cutoffs via New-Raphson's method
NR <- function(x0, para, p)
{
  para1 <- para[1:5]; para0 <- para[6:10]
  eps <- 1e-6
  d1 <- d2 <- 1
  i <- 0
  x.old <- x0
  while((d1>eps | d2>eps) & i<1000)
  {
    DTmat <- DT(x.old, para1, para0)
    Tvec <- T(x.old, para1, para0, p)

    dx <- try(qr.solve(DTmat,Tvec),TRUE)
    if (inherits(dx,'try-error')) dx <- qr.solve(DTmat+diag(0.1,2),Tvec)

    x.new <- x.old - dx
    Tvec.new <- T(x.new,para1,para0,p)
    i.line <- 0; alpha <- 1
    while( crossprod(Tvec.new) > crossprod(Tvec) & i.line < 1000)
    {
      alpha <- alpha/2
      x.new <- x.old - alpha*dx
      i.line <- i.line + 1
      Tvec.new <- T(x.new,para1,para0,p)
    }

    i <- i + 1
    d1 <- sqrt(sum((x.new - x.old)^2))
    d2 <- sqrt(sum(Tvec.new^2))
    x.old <- x.new

  }
  list(opt.cuts=x.new, steps=i)
}
```

REFERENCES

[1] T. A. Alonzo and M. S. Pepe. Using a combination of refence tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine*, 18:2987–3003, 1999.

[2] A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985.

[3] S. G. Baker. Evaluation multiple diagnostic tests with partial verification. *Biometrics*, 51(1):330–337, 1995.

[4] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psycology*, 12:387–415, 1975.

[5] M. A. Black and B. A. Craig. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine*, 21:2653–2669, 2002.

[6] T. Bourlet, C. Guglielminotti, M. Evrard, P. Berthelot, F. Grattard, A. Frésard, F. R. Lucht, and B. Pozzetto. Prevalence of GBV-C/hepatitis G virus RNA and E2 antibody among subjects infected with human immunodeficiency virus type 1 after parenteral or sexual exposure. *Journal of Medical Virology*, 58:373–377, 1999.

[7] T. Cai. Semi-parametric ROC regression analysis with placement values. *Biostatistics*, 5:45–60, 2004.

[8] T. Cai and C. Moskovitz. Semiparametric estimation of the binormal ROC curve. *Biostatistics*, 5:573–86, 2004.

[9] T. Cai and M. S. Pepe. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of American Statistical Association*, 97:1099–1107, 2002.

[10] Y. W. Choi, M. Collins, and I. Gardner. Bayesian inferences for receiver opterating characteristic curves in the absence of a gold standard. *Journal of Agricultural Biological and Environmental Statistics*, 11(19), 2006.

[11] P. Clevenbergh, J. Durant, P. Halfon, A. Tran, T. Manos, V. Rahelinirina, G. Yang, S. Benzaken, D. Ouzan, P. Rampal, and P. Dellamonica. High prevalence of GB virus C/hepatitis G virus infection in different risk groups of HIV-infected patients. *Clinical Microbiology and Infection*, 4:644–647, 1998.

[12] M. H. DeGroot. *Optimal Statistical Decisions*. Wiley Classics Library, 2004.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977.

[14] N. Dendukuri and L. Joseph. Baeysian approach to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*, 57:158–167, 2001.

[15] B. J. Dille, T. K. Surowy, R. A. Gutierrez, P. F. Coleman, M. F. Knigge, R. J. Carrick, R. D. Aach, F. B. Hollinger, C. E. Stevens, L. H. Barbosa, G. J. Nemo, J. W. Mosley, G.J. Dawson, and I.K. Mushahwar. An ELISA for detection of antibodies to the E2 protein of GB virus C. *Journal of Infectious Diseases*, 175:458–461, 1997.

[16] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

[17] C. Enøe, M. P. Georgiadis, and W. O. Johnson. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine*, 45:61–81, 2000.

[18] J. J. Gart and A. A. Buck. Comparison of a screening test and a reference test in epidemiologic studies. ii. a probabilistic model for the comparison of diagnostic tests. *Am. J. Epidemiol.*, 83(593-602), 1966.

[19] J. Geweke. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics*, volume 4. Clarendon Press: Oxford, 1992.

[20] D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

[21] P. Hall and X. H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics*, 31:201–224, 2003.

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

[23] P. Heidelberger and P. D. Welch. Simulation run length control in the presence of an initial transient. *Operationg Research*, 31:1109–1144, 1993.

[24] R. M. Henkelman, I. Kay, and M. J. Bronskill. Receiver operating characteristics (ROC) analysis without truth. *Medical Decision Making*, 10:24–29, 1990.

[25] S. Heringlake, J. Ockenga, H. L. Tillmann, C. Trautwein, D. Meissner, M. Stoll, J. Hunt, C. Jou, N. Solomon, R. E. Schmidt, and M. P. Manns. GB Virus C/Hepatitis G Virus Infection: A Favorable Prognostic Factor in Human Immunodeficiency Virus-Infected Patients. *The Journal of Infectious Diseases*, 177(6):1723–1726, 1998.

[26] S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1):167–171, March 1980.

[27] S. L. Hui and X. H. Zhou. Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research*, 7:354–370, 1998.

[28] L. Joseph, T. W. Gyorkos, and L. Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–272, 1995.

[29] H. C. Kraemer. *Evaluation medical tests: objective and quantitative guidelines*. Sage Publications, 1992.

[30] M. Krajden, A. Yu, H. Braybrook, A. S. Lai, A. Mak, R. Chow, D. Cook, R. Tellier, M. Petric, R. D. Gascoyne, J. M. Connors, A. R. Brooks-Wilson, R. P. Gallagher, and J. J. Spinelli. GBV-C/hepatitis G virus infection and non-Hodgkin lymphoma: a case control study. *International Journal of Cancer*, 126:2759–2761, 2010.

[31] L. D. Kudryavtsev. Implicit function. In Michiel Hazewinkel, editor, *Encyclopaedia of Mathematics*. Springer, http://eom.springer.de/i/i050310.htm, 2001.

[32] T. S. Lau. On dependent repeated screening tests. *Biometrics*, 47:77–86, 1991.

[33] J. J. Lefrére, P. Loiseau, J. Maury, J. Lasserre, N. Mariotti, N. Ravera, J. Lerable, G. Lefévre, T. Morand-Joubert, and R. Girot. Natural history of GBV-C/hepatitis G virus infection through the follow-up of GBV-C/hepatitis G virus-infected blood donors and recipients studied by RNA polymerase chain reaction and anti-E2 serology. *Blood*, 90:2776–2780, 1997.

[34] D. J. Lunn, H. Thomas, N. Best, and D. Spiegelhalter. Winbugs – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337, 2000.

[35] R. J. Marshall. The predictive value of simple rules for combining two diagnostic tests. *Biometrics*, 45:1213–22, 1989.

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

[37] J. H. McLinden, T. M. Kaufman, J. Xiang, Q. Chang, D. Klinzman, A. M. Engel, G. Hess, U. Schmidt, M. Houghton, and J. T. Stapleton. Characterization of an immunodominant antigenic site on GB virus C glycoprotein E2 that is involved in cell binding. *Journal of Virology*, 80:2131–12140, 2006.

[38] E. L. Mohr and J. T. Stapleton. GB virus type C interactions with HIV: the role of envelope glycoproteins. *Journal of Viral Hepatitis*, 16:757–768, 2009.

[39] E. L. Mohr, J. Xiang, J. H. McLinden, T. M. Kaufman, Q. Chang, D. C. Montefiori, D. Klinzman, and J. T. Stapleton. GB virus type C envelope protein E2 elicits antibodies that react with a cellular antigen on HIV-1 particles and neutralize diverse HIV-1 isolates. *Journal of Immunology*, 185:4496–4505, 2010.

[40] R. J. Murtaugh. ROC curves with multiple marker measurements. *Biometrics*, 51:1514–1522, 1991.

[41] R. B. Nelsen. *An Introduction to Copulas*. New York: Springer, 1999.

[42] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statisitcal hypothesis. *Philosophical Transactions of the Royal Society of London, Series A*, 231:289–337, 1933.

[43] S. S. Nielsen, C. Gronbaek, J. F. Agger, and H. Houe. Maximum-likelihood estimation of sensitivity and specificity of ELISAs and faecal culture for diagnosis of paratuberculosis. *Preventive Veterinary Medicine*, 53:191–204, 2002.

[44] N. A. Obuchowski. Receiver operating characteristics curves and their use in radiology. *Radiology*, 229(1):3–8, 2003.

[45] M. S. Pepe. A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84:595–608, 1997.

[46] M. S. Pepe. An interpretation for the ROC curve and inference using glm procedures. *Biometrics*, 56(2):352–259, 2000.

[47] M. S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press:New York, 2003.

[48] M. S. Pepe and T. Cai. The analysis of placement values for evaluating discriminatory measures. *Biometrics*, 60:528–535, 2004.

[49] M. S. Pepe and M. L. Thompson. Combining diagnostic tests to increase accuracy. *Biostatistics*, 1(2):123–140, 2000.

[50] T. J. Pilot-Matias, R. J. Carrick, P. F. Coleman, T. P. Leary, T. K. Surowy, J. N. Simons, A. S. Muerhoff, S. L. Buijk, M. L. Chalmers, G. J. Dawson, S. M. Desai, and I. K. Mushahwar. Expression of the GB virus C E2 glycoprotein using the Semliki Forest virus vector system and its utility as a serology marker. *Virology*, 225:282–292, 1996.

[51] P. Politser. Reliability, decision rules, and the value of repeated tests. *Medical Decision Making*, 2:47–69, 1982.

[52] Y. Qu, M. Tan, and M. K. Kutner. Random effects models for evaluating accuracy of diagnostic tests. *Biometrics*, 52:797–810, 1996.

[53] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[54] A. E. Raftery and S. M. Lewis. One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7:493–497, 1992.

[55] D. Rey, J. Vidinic-Moularde, P. Meyer, C. Schmitt, S. Fritsch, J. M. Lang, and F. Stoll-Keller. High Prevalence of GB Virus C/Hepatitis G Virus RNA and Antibodies in Patients Infected with Human Immunodeficiency Virus Type 1. *European Journal of Clinical Microbiology & Infectious Diseases*, 19:721–724, 2000.

[56] S. Smith, M. Donio, M. Singh, J. Fallon, L. Jitendranath, N. Chkrebtii, J. Slim, D. Finkel, and G. Perez. Prevalence of GB virus type C in urban Americans infected with human immunodeficiency virus type 1. *Retrovirology*, 2:38, 2005.

[57] J. T. Stapleton, K. Chaloner, J. Zhang, D. Klinzman, I. E. Souza, J. Xiang, Fahey Landay, A., Pollard J., R., and R Mitsuyasu. B virus C viremia is associated with reduced CD4 expansion following interleukin 2 therapy in HIV-infected people receiving HAART. *AIDS*, 23:605–610, 2009.

[58] M. Staquet, M. Rozencweig, Y. J. Lee, and F. M. Muggia. Diagnostic performance of two tests and fecal culture for subclinical paratuberculosis and associations with production. *Journal of Chronic Diseases*, 34:599–610, 1981.

[59] J. Q. Su and J. S. Liu. Linear combination of multiple diagnostic markers. *Journal of the American Statistical Association*, 88:1350–1355, 1993.

[60] J. A. Swets. *Signal detection theory and ROC analysis in psychology and diagnostics : collected papers*. Lawrence Erlbaum Associates, Mahwah, NJ, 1996.

[61] M. Tacke, K. Kiyosawa, K. Stark, V. Schlueter, B. Ofenloch-Haehnle, G. Hess, and A. M. Engel. Detection of antibodies to a putative hepatitis G virus envelope protein. *Lancet*, 349:318–320, 1997.

[62] T. R. Ten Have and E. O. Bixler. Modelling population heterogeneity in sensitivity and specificity of a multi-stage screen for obstructive sleep apnoea. *Statistics in Medicine*, 16:1995–2008, 1997.

[63] L. A. Thibodeau. Evaluating diagnostic tests. *Biometrics*, 37:801–804, 1981.

[64] M. L. Thompson. Assessing the diagnostic accuracy of a sequence of tests. *Biostatistics*, 4(3):341–351, 2003.

[65] D. P. Tihansky. Properties of the bivariate normal cumulative distribution. *Journal of the American Statistical Association*, 67(340):903–905, 1972.

[66] H. L. Tillmann, H. Heiken, A. Knapik-Botor, S. Heringlake, J. Ockenga, J. C. Wilber, B. Goergen, J. Detmer, M. McMorrow, M. Stoll, R. E. Schmidt, and M. P. Manns. Infection with gb virus c and reduced mortality among hiv-infected patients. *New England Journal of Medicine*, 345(10):715–724, 2001.

[67] E. A. Tolley, G. W. Somes, and E. S. Willey. Determining efficacy of monitoring devices: evaluating new technologies. *Statistics in Medicine*, 10:351–360, 1991.

[68] P. M. Vacek. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41:959–968, 1985.

[69] P. N. Valenstein. Evaluating diagnostic tests with imperfect standards. *American Journal of Clinical Pathology*, 93:252–258, 1990.

[70] A. W. van der Vaart. *Asymptotic statistics*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, 2000.

[71] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer series in statistics. Springer, 1996.

[72] S. D. Walter and L. M. Irwig. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemology*, 41:923–937, 1988.

[73] C. Wang, B. W. Turnbull, Y. T. Grohn, and S. S. Nielsen. Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown. *Journal of Agricultural, Biological, and Environmental Statistics*, 12:128–146, 2007.

[74] C. F. Williams, D. Klinzman, T. E. Yamashita, J. Xiang, P. M. Polgreen, C. Rinaldo, C. Liu, J. Phair, J. B. Margolick, D. Zdunek, G. Hess, and J. T. Stapleton. Persistent GB Virus C Infection and Survival in HIV-Infected Men. *New England Journal of Medicine*, 350(10):981–990, 2004.

[75] Y. Wu. *The Partially Monotone Tensor Spline Estimation of Joint Distribution Function with Bivariate Current Status Data*. PhD thesis, University of Iowa, 2010.

[76] J. Xiang, J. H. McLinden, R. A. Rydze, Q. Chang, T. M. Kaufman, D. Klinzman, and J. T. Stapleton. Viruses within the Flaviviridae decrease CD4 expression and inhibit HIV replication in human CD4+ cells. *Journal of Immunology*, 183:7860–7869, 2009.

[77] I. Yang and M. P. Becker. Latent variable modeling of diagnostic accuracy. *Biometrics*, 53:948–958, 1997.

[78] B. Yu, C. Zhou, and S. Bandinelli. Combining multiple continuous tests for the diagnosis of kidney impairment in the absence of a gold standard. *Statistics in Medicine*, 30:1712–1721, 2011.

[79] W. Zhang, K. Chaloner, H. L. Tillmann, C. F. Williams, and J. T. Stapleton. Effect of early and late GB virus C viraemia on survival of HIV-infected individuals: a meta-analysis. *HIV Medicine*, 7:173–180, 2006.

[80] X. H. Zhou, P. Castelluccio, and C. Zhou. Nonparametric estimation of ROC curves in the absence of a gold standard. *Biometrics*, 61:600–9, 2005.

[81] X. H. Zhou, D. K. McClish, and N. A. Obuchowski. *Statistical Methods in Diagnostic Medicine*. Wiley-Interscience, 2002.

[82] M. H. Zweig and G. Campbell. Receiver operating characteristics (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(8):561–577, 1993.