Latent Class and Latent Profile Analysis in Medical Diagnosis and Prognosis

Zheyu Wang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Xiao-Hua A. Zhou, Chair

Patrick J. Heagerty

Kathleen F. Kerr

Brian P. Flaherty

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Latent Class and Latent Profile Analysis
in Medical Diagnosis and Prognosis

Zheyu Wang

Chair of the Supervisory Committee:
Professor Xiao-Hua A. Zhou
Department of Biostatistics

Evaluating test accuracy is an important topic in medical diagnosis and prognosis. Accuracy information is necessary for care-givers to make well-informed decisions; it also helps researchers to select better diagnostic tools, either from new techniques or based on combinations of current information. This field has recently been re-energized, due to advances in diagnostic techniques and the discovery of novel biomarkers. However, assessment becomes difficult when the underlying medical condition, or the gold standard, is unknown due to time or cost constraints, lack of biotechnology, or concerns over the invasive nature of a diagnostic procedure. This issue is becoming more common and pressing with the growing interest in, and emphasis on, preclinical diagnosis and prevention. Moreover, with improvements in clinical practice, there is now a need to go beyond a traditional binary disease status approach and incorporate an ordinal gold standard. Additionally, the ability to take subjects individual characteristics, which may affect disease prevalence and test performance, into consideration, will allow care-givers to provide their patients with more accurate and personalized diagnoses.

This dissertation views the unobserved gold standard as a latent variable, and proposes models in the latent class and latent profile framework to solve the above mentioned problems. For categorical tests, a latent class approach is adopted to nonparametrically model the conditional distributions of the tests within different disease groups. Additionally, a

random effect method is introduced to relax the classic conditional independence assumption in latent class models, so that the model can then be applied to more general situations. A likelihood ratio test on the conditional independence assumption is also discussed. For continuous tests, a latent profile model is proposed, which allows for the inclusion of a set of covariates that may be associated with disease prevalence, and a set of covariates that may influence test performance. Therefore, the model also relaxes the conditional independence assumption by explicitly explaining correlations among the tests within each disease category. Moveover, it can provide information about risk factors' impacts and about a tests properties within subpopulations. Additionally, the model proposed here allows for a transformation on the test results to take into account possible skewness in the data.

This dissertation also proposes that a summary measure and graphical presentation of the results in terms of the commonly used receiver operating characteristic (ROC) curve cannot be directly apply to data with an ordinal gold standard. This dissertation extends the concept of the ROC curve into a high dimensional volume and provides corresponding interpretations.

Extensive simulations have been performed to assess the consistency and robustness of the proposed methods. Moreover, this dissertation carefully discusses the local and global identifiabilities of latent class and latent profile models, and is the first to provide sufficient conditions for establishing local and global identifiability for latent class and latent profile models in the general form. These results provide theoretical justification of the proposed methods and guidance for practical applications of these models.

The proposed methods are illustrated using data from a traditional Chinese medicine practice to evaluate doctors' diagnostic accuracy for symptom diagnosis, and in a data set from a study on Alzheimer's disease to select and combine biomarkers that can help with early detection.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my dissertation advisor, Andrew Zhou, who has been supportive and encouraging throughout my course of study here, who is patient and tolerant with me, and who provided great guidance and inspiration, while giving me the freedom to pursue my own ideas, so that I could become an independent researcher.

I would like to thank the members of my supervisory committee, Patrick Heagerty, Kathleen Kerr, Brian Flaherty and Walter Kukull, and the other professors who provided important support during my study. Patrick Heagerty has had a great influence on me. He taught me to recognize the fundamental idea behind a complicated problem. Your advice and support was invaluable. I would also like to thank Brian Flaherty for his generosity in sharing his knowledge and insights with me, and for being so responsive to my questions. I owe a special thanks to Kathleen Kerr, who introduced me to the research of risk prediction, and who has been more than an advisor to me. She is very considerate and thinks always of my best interest. She has devoted a lot of time to advising me. I could not have asked for a more supportive advisor. My gratitude is also due to Walter Kukull, who served as my graduate school representative, and who provided many suggestions about the Alzheimer's disease application of my dissertation, and to Thomas Montine, my biology advisor, who helped me to have a better understanding of Alzheimer's disease.

I also wish to thank Margaret Pepe, for her guidance outside my dissertation research. She has deepened my understanding of diagnostic research and widened my research area. It was a pleasure to work with you. I would like to thank Jon Wellner for his helpful suggestions and comments inside and outside of my dissertation, and for his always heart-warming smile. I am grateful to my advisor, Zhi Geng in Peking University, who first led me into the field of medical diagnosis and remains very supportive of me across the Pacific Ocean. Many thanks to the faculty, staff and my fellow students in the Department of

Biostatistics; their friendship and continued support have made my time here enjoyable and memorable.

I owe my deepest gratitude to my parents, who have always supported and believed in me, and who let their only daughter pursue her dream anywhere in the world. Their love and blessings give me the strength to overcome any challenge in my life. This dissertation is dedicated to you.

Chapter 1

INTRODUCTION

1.1 Medical Tests

Medical tests play an important role in the practice of health care. They are performed to help detect, diagnose, or monitor diseases, disease processes, and susceptibility, and to determine a course of treatment. Medical tests can be classified as either diagnostic or prognostic in the classic sense, depending on how they are utilized, as discussed below.

1.1.1 Diagnostic Tests

Diagnosis is the process of attempting to determine or identify a disease or a medical condition. It can be thought of as a classification–a way to distinguish between diseased subjects and healthy subjects, to determine or rule out certain conditions. A diagnostic test has at least three purposes (Sox, et al., 1989; McNeil and Adelstein, 1976): (1) to provide reliable information about the patient's condition, (2) to influence the treatment plan for the patient, and (3) to understand disease mechanisms and natural history. The first purpose is the most straightforward. For example, we use X-rays to examine bone fractures, take a complete blood count to check for a bacterial infection, or use MRI technology to detect pathologic tissue, such as that arises from a brain tumor. The second purpose points out the relationship between a diagnostic test and treatment. Certainly, a diagnostic test is only useful when the disease is treatable and treatment is available to those who test positive. In most situations where multiple treatments are available, a diagnostic test should provide the care-giver with information to help them select the most appropriate treatment plan. For example, based on the specific type, location and stage of a tumor, breast cancer can be treated with surgery, medications (hormonal therapy and/or chemotherapy), radiation, immunotherapy, or with a combined method. Because of the influence that a diagnostic test has on a treatment plan, in most situations, what we are interested in is more than a binary

result; not only do we want to know whether a test is positive or negative, but also about severity and other important information. This prompts us to consider ordinal or nominal results in our work. The third purpose points out the longitudinal value of a diagnostic test – for example, with repeated examinations of patients who have chronic conditions to evaluate progression, or with monitoring subjects who might have recurrent events.

1.1.2 Prognostic Tests

Prognosis is the process of predicting the most likely outcome or course of a disease – for example, it frequently involves predicting the likelihood of developing a certain condition in a high risk group, or for an individual with a certain condition, predicting their expected duration, function, and course of the disease (e.g. progressive decline, an intermittent crisis, or a sudden and unpredictable crisis). It can be considered as a prediction for a future event. Prognostic tests provide information to help with such prediction. For example, prognostic scoring is used to make cancer outcome predictions. A Manchester score is considered an indicator of prognosis in small-cell lung cancer. Likewise, for Non-Hodgkin lymphoma, physicians have developed the International Prognostic Index to predict patient outcomes.

1.1.3 A Broader Context

Despite the differences between diagnosis and prognosis in a classic medical sense, the two are quite similar from a statistical point of view. The same statistical methodologies used to evaluate, compare, and combine tests can be applied to both diagnostic and prognostic tests. In fact, in medical diagnoses, the outcome of interest can be a disease, a medical condition, or any other medical event. In this broader context, prognosis can be viewed as a special type of diagnosis, in which the condition of interest is a future event and the goal is to diagnose or predict the risk of developing such an event, or to know the duration or other related information. For example, we may use clinical and laboratory data to predict a subject's risk of kidney failure within a certain time period. In this case, kidney failure within a certain time period can be considered as a medical condition to be diagnosed. Moreover, most medical conditions are processes rather than sudden events.

They can include prolonged preclinical and clinical stages. Prognosis can be considered as diagnosing preclinical signs that can help to predict a future event. One such example is screening tests. These tests are usually performed in subjects with apparently good health, to identify previously unrecognized diseases, in the same way that a Pap smear or liquid-based cytology is used to detect potentially precancerous lesions and prevent cervical cancer, or that a mammography is used to detect breast cancer. They are used to reveal developing conditions, and to help predict future events. Depending on how a condition or disease is defined, these tests can be viewed as either diagnosic or prognosic. As another example, genetic markers can also be used to help diagnose and prevent diseases, such as screening for trisomy 21 in Down's syndrome, amyloid precursor proteins for early onset Alzheimer's disease and Apolipoprotein E-e4 for high risk of late onset Alzheimer's disease. With an increasing emphasis on early detection and prevention, diagnosis and prognosis are becoming more unified. By considering the outcome of interest to be any medical condition or event, either present or under development, I henceforth refer to all such tests as diagnostic tests in my work, and will use the term "disease" to collectively denote any disease, medical condition or event of interest in this dissertation.

In addition to the outcomes of interest, diagnostic tests can be further generalized according to their means of diagnosis. Traditional diagnostic techniques include radiographic imaging, biochemical testing, bacterial culture, etc. Additionally, doctors' diagnostic opinions, radiologists' reading of tests, and scores obtained from various administrated or self-administrated questionnaires can all be considered diagnostic tests, as they provide information that helps to detect, diagnose, or monitor a medical condition. Similarly, measures of risk factors, biomarkers and their combinations, such as risk models and composite scores, can also be thought of as diagnostic tests. In other words, one could consider a diagnostic test to be any technique that provides information for detecting, diagnosing or monitoring any medical condition or event of interest, as I will do here in my work. In this point of view, the methodology for evaluating and comparing diagnostic tests can also be applied to evaluate doctors' performance, to assess various measures and questionnaires, to compare risk models, and to select and combine biomarkers.

1.2 Measures of Diagnostic Accuracy

Diagnostic tests provide information about subjects' medical conditions. However, most diagnostic tests are imperfect. Information about the accuracy of the tests, or the abilities of the tests to discriminate among alternative states of health (Zweig and Campbell, 1993), is important for allowing health care providers to make well-informed decisions.

The accuracy of diagnostic tests can be evaluated in many ways. Various metrics have been proposed to quantify different aspects. Most of these metrics are defined with a focus on binary disease status (i.e. the presence or absence of a disease), and reflect the tests' abilities to distinguish between these two groups. In this section, I consider only binary disease status and give a brief review on the commonly used sensitivity, specificity and the receiver operating characteristic (ROC) curve. Their generalizations to ordinal disease status, such as severity level, are discussed in Chapter 3.

1.2.1 Notation

Here and in later chapters, I use D to denote the true disease status, also referred to as the *gold standard*. For binary disease status, $D = 0$ indicates a subject without disease, and $D = 1$ a subject with disease. For ordinal disease status, $D = 0, \ldots, L-1$ denotes subjects with differing levels of severity, with the higher value of D indicating a more severe status. The methods proposed in this dissertation are subsequently applied to data sets with an ordinal latent disease category. However, with additional information, these methods can also be used when D is nominal (refer to Chapter 7).

I use T to denote the result of a diagnostic test. T can be binary, with a value of 0 or 1 to indicate a negative or positive test result, ordinal with values $0, \ldots, J-1$, or continuous. By convention, I assume here that a higher value of T is more indicative of disease in the binary case, or that it indicates a more severe status in the ordinal case.

The number of possible values for T should be more than the number of possible values for D. For binary disease status, a cutoff point, c, is selected to dichotomize T into positive or negative results to help classify diseased and non-diseased groups: one considers a test positive (denote as "$T+$") if $T \geq c$, and negative (denote as "$T-$") if $T < c$. Similarly,

a series of cutoff points can be used to categorize T into corresponding ordinal disease statuses.

1.2.2 Sensitivity and Specificity

Sensitivity and specificity are two basic measures of diagnostic accuracy. Before introducing them, I first look at different classification probabilities of the test results by disease status. When the disease status is binary, comparing test results with true disease status, I classified the performance into four categories as shown in Table 1.1.

Table 1.1: Classification table of T and D

	True Disease Status	
Test Results	Disease present ($D = 1$)	Disease absent ($D = 0$)
Test positive ($T+$)	True positive	False positive
Test negative ($T-$)	False negative	True negative

The two cells on the main diagonal line, true positive and true negative, are correct diagnoses, whereas the two cells off the main diagonal line, false positive and false negative, represent two types of diagnostic error. Consequently, the corresponding probabilities of obtaining positive or negative results among disease and non-diseased groups reflect some aspects of the accuracy of a test. Specifically, these probabilities are called the true positive rate (TPR), the true negative rate (TNR), the false positive rate (FPR) and the false negative rate (FNR), respectively, and are defined as follows:

$$\begin{aligned} \text{TPR} &= P(T+|D=1), \qquad \text{FPR} = P(T+|D=0), \\ \text{FNR} &= P(T-|D=1), \qquad \text{TNR} = P(T-|D=0). \end{aligned}$$

It is easy to see that these measures satisfy the following relationship

$$\text{TPR} + \text{FNR} = 1 \quad \text{and} \quad \text{FPR} + \text{TNR} = 1.$$

Therefore, the value of only one quantity from each pair is needed. For example, the two probabilities of correct diagnosis, TPR and TNR, are often used. In fact, they are usually referred to as the sensitivity and specificity of a test, i.e.,

$$\text{Sensitivity} = \text{TPR} = P(T+|D=1),$$
$$\text{Specificity} = \text{TNR} = 1 - \text{FPR} = P(T-|D=0).$$

Sensitivity measures a test's performance among diseased subjects; in other words, it is the ability of a test to identify a disease when the disease is present. Specificity measures a test's performance among non-diseased subjects, and therefore represents the ability of a test to rule out a disease when it is not present.

Sensitivity and specificity, and similarly TPR, FPR, TNR and FNR, are conditional quantities among the diseased or non-diseased groups. Their values are intrinsic to the diagnostic test and are not affected by the prevalence of the disease.

1.2.3 Receiver Operating Characteristic Curve

As previously mentioned in section 1.2.1, with ordinal or continuous tests a threshold c is usually selected to dichotomize test results as being either positive or negative. In this case, sensitivity and specificity can be considered functions of c.

$$\text{Sensitivity(c)} = P(T \geq c|D=1), \tag{1.1}$$
$$\text{Specificity(c)} = P(T < c|D=0). \tag{1.2}$$

Different choices of c lead to different sensitivity and specificity pairs. Usually, there is a trade-off between high sensitivity and high specificity. In equations (1.1) and (1.2), it is evident that a higher cutoff point c leads to a lower sensitivity and a higher specificity. For example, if a test only looks at subjects with extremely high blood pressure, diagnosed as hypertension, then individuals with mild conditions may be missed, leading to lower sensitivity. On the other hand, using such a test assures that fewer healthy subjects are misdiagnosed, yielding a higher specificity. The receiver operating characteristic (ROC) curve reflects this trade-off between sensitivity and specificity pairs. It is a plot of sensitivity,

or TPR, versus 1−specificity, or FPR, across all possible cutoff points, i.e.,

$$\begin{aligned} \mathrm{ROC}(\cdot) &= \{\big(1-\text{specificity}(c),\ \text{sensitivity}(c)\big),\ c\in(-\infty,+\infty)\}, \\ \text{or equivalently,} &= \{\big(\mathrm{FPR}(c),\ \mathrm{TPR}(c)\big),\ c\in(-\infty,+\infty)\}. \end{aligned}$$

By noting that both $\mathrm{FPR}(c)$ and $\mathrm{TPR}(c)$ are monotonic functions, the ROC curve can be written explicitly as the following,

$$\mathrm{ROC}(t)=\mathrm{TPR}\big(\mathrm{FPR}^{-1}(t)\big),\quad t\in(0,1)$$

where $\mathrm{FPR}^{-1}(\cdot)$ is the inverse function of $\mathrm{FPR}(\cdot)$.

Figure 1.1 shows some examples of ROC curves for tests with different accuracies. A good test has high sensitivity and high specificity, so its ROC curve tends to be close to the upper left-hand corner of the unit quadrant. A perfect test has its ROC curve along the left and upper borders of the unit quadrant, as shown in the gold curve on the plot. On the other hand, a useless test, such as that which corresponds to a random guess, cannot distinguish diseased groups from healthy groups. In other words, it has the same conditional distribution in these two groups. Therefore, for any threshold c, the corresponding we have $\mathrm{FPR}(c)=\mathrm{TPR}(c)$, and the ROC curve is a line with unit slope $\mathrm{ROC}(t)=t$, as shown by the black curve on the plot. Most tests have ROC curves that lie between these two extremes, like the red and blue lines in this plot. In this case, test A (blue) is uniformly better than test B (red) since, for any fixed specificity, test A has higher sensitivity. When two ROC curves cross, the comparison will depend on which region of the plot is more of interest.

The ROC curve provides a visualization of the trade-off between sensitivity and specificity. It can be used as guidance in choosing an appropriate cutoff point. Another nice property of the ROC curve is that it is invariant to any monotonic transformation of the test result. In other words, the curve only depends on the rank of the test, not on its actual magnitude (Zweig and Campbell, 1993; Campbell, 1994). This makes it very convenient to compare ROC curves of tests with different scales.

In addition, a quantity derived from the ROC curve, the area under the ROC curve (AUC), is a commonly used single number index that summarizes the accuracy of a test. It

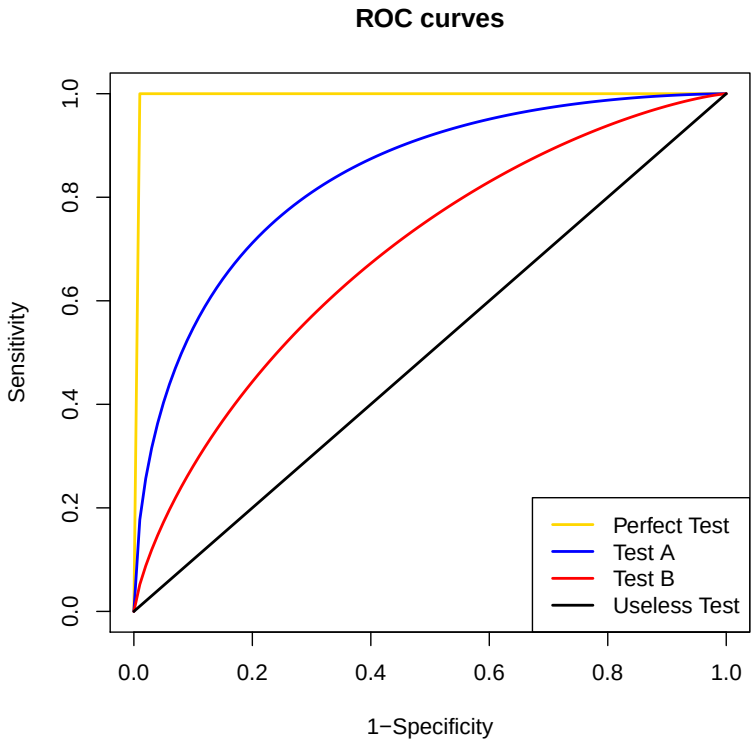

Figure 1.1: Examples of ROC curves

is defined as follows:

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

A perfect test has an AUC equal to 1, and a useless test has an AUC equal to 0.5. It can be shown that the value of AUC is equivalent to "the probability that a randomly selected case (patient with the condition) has a greater test result than a randomly selected control (patient without the condition)" (Hanley and McNeil, 1982). Therefore, an alternative representation of the AUC is

$$\text{AUC} = P(T_2 > T_1 | D_2 = 1, D_1 = 0) + 1/2P(T_2 = T_1 | D_2 = 1, D_1 = 0).$$

The second term appears to adjust for possible ties in the test results. AUC is also equivalent to the *concordance-index*, or *c-index*, in comparing predictions or prognostic models. It is defined as "the proportion of all usable patient pairs in which the predictions and outcomes

are concordant". It describes how well the model discriminates between subjects with or without disease (Harrell et al., 1996). This also suggests some similarities between diagnosis and prognosis from a statistical point of view.

Because an ROC is invariant to monotonic transformations of the test results, similar to sensitivity and specificity, the ROC curve and the AUC are also intrinsic properties of a diagnostic test.

1.2.4 Other Accuracy Measures

In addition to sensitivity, specificity and the ROC curve, there are other measures that capture different aspects of a test's accuracy, such as predictive values, likelihood ratio, total gain, etc. They are outside the scope of this dissertation, but readers can refer to Pepe, 2003 and Gu et al., 2009 for more information.

1.3 Challenges in Diagnostic Accuracy Study

1.3.1 The Absence of a Gold Standard

All measures of diagnostic accuracy describe a relationship between diagnostic results and true disease status – the gold standard. In a diagnostic accuracy study, the most common approach to evaluating a test's performance is to compare its diagnostic result with the gold standard, using various measures that summarize accuracy. As a result, most methods require obtaining gold standard information for the study population, or at least for a portion of the study population. However, in many cases, verification of the gold standard may be hard to obtain, due to cost constraints, concerns about the invasive nature of the diagnostic procedure, or a lack of biotechnology needed to obtain a definitive result. For example, the diagnosis of Alzheimer's disease (AD) cannot be established until a patient has died and a neuropathological examination has been conducted. In some situations, a method for establishing the true disease status may not exist or has not been widely accepted. Research on traditional Chinese medicine (TCM) usually encounters such a difficulty. The diagnosis in TCM is usually carried out through the classic "four diagnostic methods", including observation, inquiry, smelling/listening and palpation", which are relatively ob-

jective. Despite its good performance, TCM has not been well explained by modern science. Many concepts of symptoms are unique to TCM; they can hardly be found in the theories of western medicine or measured by any other medical instrument. As a result, a gold standard cannot be obtained due to the incomplete understanding of a disease, as the disease is defined by TCM. In addition, even with a good understanding of the disease, measurement error can still cause the absence of a gold standard. For example, the "definitive" diagnosis of a well-defined condition, such as an infection by a known agent, requires a culture of the organism or other detection methods, any of which may be subject to laboratory and other errors.

When gold standard information is not available, sometimes in practice an imperfect reference test may be adopted as the gold standard. This test is usually the best available test under reasonable conditions and has relatively good accuracy performance. For example, clinical AD diagnosis is often used when autopsy information is not available. However, diagnostic error can occur with an imperfect reference test. Zhou et al., 2011 provided real and hypothetical examples illustrating that these errors can lead to underestimation as well as overestimation of the diagnostic accuracy measures. Therefore, statistical methods that can account for these errors and evaluate diagnostic test accuracy are needed when a gold standard is not available. Especially when thinking more carefully, true gold standard information is rarely available. Most so-called "gold standard tests" are, in fact, merely the best available reference tests. For example, the gold standard in cancer diagnosis is often the pathologic review of a biopsy specimen. If the biopsy is not taken from exactly the right location, the diagnosis can be inaccurate. A similar situation may occur when blood or tissue culture is used to detect the presence of bacterial infection. Moreover, a lack of gold standard tests is becoming more pressing and common with the growing interest in, and emphasis on, preclinical stage diagnosis and prevention, in which case, even if a gold standard test exists a long follow-up period may be required.

1.3.2 Ordinal Disease Status

Most accuracy measures and methodologies for diagnostic test accuracy focus on situations in which the disease status is binary. However, with recent improvements in clinical practice, ordinal diagnoses have become more common and are preferred, since they reflect the severity or progression of a disease. For example, in the diagnosis of Alzheimer's disease (AD), subjects are often classified into normal or mild cognitive impairment (MCI) and dementia groups. Similarly, regions of a stroke patient's heart are classified using an ordinal scale: normal, hibernating, ischemic or necrotic, to reflect the severity of the condition. Traditionally, this ordinal information is combined into a binary status, such as dementia or non-dementia, or normal or abnormal region. However, information on disease severity and progression can help caregivers choose the best intervention plan, such as in cancer, where the stage of cancer progression is an ordinal scale, ranging from localized cancer to distant metastases. Different quantities of medications and other methods of therapy might be assigned to patients with different gradations of illness, in order to receive better curative results and to reduce side-effects at the same time.

Therefore, accurate information regarding disease severity or subtypes in addition to the presence or absence of a disease is also important. This requires an examination of the ability of a diagnostic test to distinguish among subjects with different severity levels. In other words, statistical methods to assess diagnostic accuracy when the disease status, or the gold standard, are needed to accomplish this task . Likewise,measures that can incorporate ordinal disease status in the evaluation of diagnostic accuracy since common measures such as sensitivity, specificity and the ROC curve are only defined for a binary disease status are needed.

1.4 Dissertation Outline

In this dissertation I focus on the methodology for diagnostic test accuracy study when a gold standard is not available. The proposed models are in the latent variable modeling framework and allow for ordinal disease status. The scale of a diagnostic test can be ordinal or continuous. Specifically, Chapter 2 gives a review of latent variable models, with an

emphasis on latent class models and latent profile models, which provide the basic framework of this dissertation. Chapter 3 considers latent class models for evaluation and comparison of multiple conditional independent ordinal tests without a gold standard. It is motivated by TCM research and used to evaluate doctors' diagnostic performance. Summary measures and graphical representations of diagnostic accuracy are also discussed. Chapter 4 extends the method delineated in Chapter 3 by relaxing the conditional independence assumption. A random effect model is then adopted to capture the possible dependent structure among the tests. Tests for model fit and for assessing the conditional independence assumption are also discussed. In Chapter 5, I consider evaluation of continuous diagnostic tests with latent profile models. This method is explored in AD research to select, evaluate and combine biomarkers that can be informative for early AD detection. Identifiability issues of latent class and latent profile models are discussed in Chapter 6, as a way of providing justification of the proposed methods and guidance for employment of latent variable models in practice, and chapter 7 concludes the dissertation and discusses topics for future research.

Chapter 2

LATENT VARIABLE MODELS

2.1 Overview

Latent variable modeling is widely used in many disciplines, including psychometrics, psychology, behavioral sciences, social sciences, diagnostic medicine, bioinformatics, natural language processing, economics and machine learning/artificial intelligence. It contains a broad category of statistical models that relate a set of observed variables (or manifest variables) to a set of unobserved latent variables. The concept of latent variables was first introduced in the psychiatric and behavioral sciences to represent entities that may be regarded as existing, but that cannot be measured directly. For instance, in psychometric research, intelligence is considered a latent variable that cannot be directly observed but can be inferred from participants' performance on tests. The purpose of latent variable models is to use underlying latent variables to explain the observed correlation among test results and to reduce dimensionality or to infer, based on the observed variables, the structure of the more fundamental variables that cannot be directly observed. In the latter case, they are also referred to as latent structure models, or structural equation models. Since the models estimate associations between manifest variables and latent variables, they can also be used to evaluate the extent to which the observed variables reflect the underlying structure, i.e., how well they measure the construct of interest, such as in item response theory (IRT) or diagnostic medicine.

Latent variable analysis allows for different types of variables, for both manifest and the latent variables, such as binary, nominal, ordered-categorical, or interval/continuous, resulting in various models (Table 2.1). For example, in factor analysis, both the manifest variables and the latent variables are continuous. The purpose of the analysis is to describe variability among observed, correlated variables in terms of potentially smaller numbers of latent factors. A latent trait model, which is also known as item response theory, considers

continuous latent variables (or traits) and categorical manifest variables (or items). It focuses on the scoring of tests, questionnaires, and similar instruments that measure abilities, attitudes, or other latent traits of a subject. Items can be multiple choice questions that have correct or incorrect responses, or statements on questionnaires that allow respondents to indicate their levels of agreement. When the latent variable is categorical, the corresponding models are called latent class models. They are used to identify groups or subtypes in which observations are homogeneous. These subtypes are called "latent classes". The manifest variables are often discrete, but can also be continuous, although the latter models are usually referred to as latent profile models.

Table 2.1: Latent variable models grouped by types of latent and manifest variables.

	Latent Variables	
Manifest Variables	Discrete	Continuous
Discrete	Latent Class Models	Latent Trait Models/IRT
Continuous	Latent Profile Models	Factor Analysis

2.1.1 Modeling Assumptions and Bayesian Foundation

The basic modeling assumptions of latent variable models are:

- Manifest variables can be modeled as a mathematical function of the latent variable(s).
- Manifest variables have nothing in common after controlling for the latent variable(s) (local independence, or conditional independence).

The first assumption indicates that the manifest variables are surrogates or results of the latent variables. The second assumption suggests that all correlations among the observed variables can be attributed to the latent variables. The local independence assumption can

be relaxed by introducing covariates or random effects, in which case the observed variables are independent, conditional on the latent variables and covariates or the random effects.

In addition, latent variable analysis implicitly uses the Bayesian formula in model implementation. It aims to determine the most likely structure of latent variables, given information about manifest variables. In a sense, the method does this by considering the probability distribution of all possible states of latent variables, given observed data, and selecting for the most likely latent structure. Specifically, it expresses

$$\Pr(\text{latent variables structure} \mid \text{manifest variables})$$

for all possible latent structure states, utilizing Bayes' rule and the probability model constructs to describe how the manifest variables are affected by the latent variables, i.e.,

$$\Pr(\text{manifest variables} \mid \text{latent variables structure}).$$

Parameter estimation can be obtained by constructing a marginal likelihood function of the manifest variables over all possible latent structures and calculating the maximum likelihood estimates.

2.1.2 Advantages and Criticisms

The latent variable models provide a general framework that unifies many disparate methods. This framework provides a flexible approach to statistical analysis where models can be specifically tailored to meet specific research needs, and is thus increasing in popularity. It can summarize multiple manifest measures parsimoniously and infer the underlying structure. The ability to infer unobserved entities is of great value. For example, since many health issues are directly related to health behaviors, researchers in sociology are interested in building theories to describe different aspects of personal and social/familial networks that influence peoples behavior. However, many of these aspects cannot be measured directly. Latent variable models offer a means to study such unobserved entities. Moreover, by including latent variables in the model, this approach acknowledges measurement problems and avoids biases due to imperfect reference measurements. It also allows researchers

to investigate error rates and differential reporting problem. In addition, especially with categorical latent variables, population heterogeneity can be described and studied.

However, the latent variable approach has also received many criticisms. The first involves a philosophical debate about the fundamental existence or non-existence of conceptual latent variables. In some applications, the latent variable is a real world entity with a clear definition, such as whether a subject has a particular disease, but that we cannot observe directly due to measurement error, cost, or ethical issues. However, in applications such as those used in the social sciences, the latent variable is usually an abstract concept, which may not have a clear definition. For example, entities such as quality of life, socioeconomic status, or intelligence may not be regarded as concrete entities in the sense that length and weight are. They are merely a way to summarize some complex features. In these situations, is it reasonable to use methods for prediction or establishing relationships as if they are concrete entities? Moreover, model interpretation can be ambiguous without a clear definition of the latent variable, or when the definition of the latent variable changes as our knowledge increases.

In addition, most criticism relates to the modeling assumptions that pertain to latent variables. Since the latent variable is unobservable, these assumptions cannot be tested. However, the conclusion that one can draw from the model usually greatly depends on whether these assumptions are met. In this type of situation there is no basis for arguing whether the inferred latent variable is the truth or instead something that results from the incorrect modeling assumptions, and yet latent variable analysis provides a powerful, model based approach to investigating phenomena that cannot be directly observed. Furthermore, all statistical models rely on assumptions that may or may not hold in any given circumstance. Muthén (2002) argued that the skepticism about latent variable analysis in mainstream statistics is at least partially due to the separate development of psychometrics and statistics. He pointed out that many statistical analyses implicitly utilize the idea of latent variables in the form of random effects, components of variation, missing data, mixture components and clusters. Consequently, as in other statistical methods, assumptions made in latent variable models need to be considered carefully based on scientific foundations, and acknowledged when making inferences.

2.2 Latent Class Models and Latent Profile Models

2.2.1 Overview

The latent class model, sometimes referred to as the finite mixture model, was initially introduced by Lazarsfeld and Henry (1968) as a way of formulating latent attitudinal variables from dichotomous survey items. It is a type of latent variable model in which the unobserved variable is categorical. The manifest variables can be either categorical or continuous. The latter is also termed as the latent profile model (Lazarsfeld and Henry, 1968). In this Chapter I will not distinguish between the two, and will refer them collectively as latent class models. On the other hand, the choice of using a categorical latent variable instead of a continuous one is more fundamental than the corresponding choice of the proper scale type for observed outcomes (Muthén, 2002). The categorical latent variable can be used to represent unobserved heterogeneity, and thus greatly enlarges modeling capabilities. With categorical underlying latent variables, fewer distributional assumptions are needed. The model is arguably most robust for summarizing data whose basic structure is patterns of categorical response. The methodology was formalized and extended to nominal variables by Goodman (1974a, 1974b), who also developed the maximum likelihood (ML) algorithm for parameter estimation.

Traditional latent class analysis (i.e., Goodman, 1974a) assumes that each observation is a member of one and only one of L latent (unobservable) classes, and that a sufficient number of latent classes for the categorical latent variable can fully account for the unobserved heterogeneity and results in local independence among the manifest variables. That is, conditional on latent class membership, the manifest variables are mutually independent of each other.

For example, suppose that there are three categorical manifest variables, or response items A, B and C. Item A has I categories, indexed by $i = 1, \ldots, I$; item B has J categories, indexed by $j = 1, \ldots, J$; and item C has K categories, indexed by $k = 1, \ldots, K$. The group variable, denoted by D takes on the values $d = 0, \ldots, L-1$, indicating the class to which each subject belongs. Then with Goodman's notation, the model can be expressed

as follows:

$$\pi_{ijkd}^{\bar{A}\bar{B}\bar{C}D} = \pi_d^D \pi_{id}^{\bar{A}D} \pi_{jd}^{\bar{B}D} \pi_{kd}^{\bar{C}D},$$

where $\pi_d^D = P(D = d)$ denotes the (unconditional) probability that a subject belongs in latent class d, and $\pi_{id}^{\bar{A}D} = P(A = i|D = d)$, $\pi_{jd}^{\bar{B}D} = P(B = j|D = d)$, $\pi_{kd}^{\bar{C}D} = P(C = k|D = d)$ denote the conditional probability of obtaining the response i from item A, the response j from item B, and the response k from item C among group d, respectively.

By summing over all possible values for group variable G, I obtained the joint marginal probability of manifest variables A, B and C as follows:

$$\pi_{ijk}^{ABC} = \sum_{d=0}^{L-1} \pi_{ijkd}^{\bar{A}\bar{B}\bar{C}D}.$$

The maximum likelihood estimation can be obtained by solving the likelihood equation for the observed data. Computationally, this can be carried out by the Newton-Raphson approach as used in Haberman's computer program LAT (Haberman, 1979), or most commonly by applying an EM algorithm to the complete data likelihood in which the unobserved disease status is treated as missing data (Goodman, 1974; Dempster, Laird, and Rubin, 1977).

2.2.2 Goodness of Fit

In some situations previous knowledge of the number of latent classes is unavailable, and the goal is to determine the smallest number of latent classes that is sufficient to explain the associations observed among the manifest variables. In such situations, the most common approach is to fit latent class models with different numbers of classes, and then to assess model fit. In some other situations there may be uncertainty about whether some of the items, say item A, have different distributions among classes, and therefore want to fit a restricted model with the constraint:

$$\pi_{i0}^{\bar{A}D} = \pi_{i1}^{\bar{A}D} = \ldots = \pi_{i(L-1)}^{\bar{A}D}, \quad \text{for } i = 1, 2, \ldots, I,$$

or with similar constraints only for selected classes, and then to compare model fit.

The likelihood ratio statistic, and measures such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), are the most popular choices for assessing model fit. The likelihood ratio statistic evaluates the extent to which the expected cell frequencies $\hat{\pi}_{ijk}^{ABC}$, based on the maximum likelihood estimates, agree with the corresponding observed frequencies π_{ijk}^{ABC}. It is defined as,

$$LR = 2N \sum_{i,j,k} \Big[\pi_{ijk}^{ABC} log\Big(\frac{\pi_{ijk}^{ABC}}{\hat{\pi}_{ijk}^{ABC}} \Big) \Big],$$

where N is the total sample size. When N is sufficiently large, LR follows a centered χ^2 distribution with degrees of freedom equal to the total number of cells in the muti-way table ($I \times J \times K$ in our example) minus the number of distinct parameters ($L \times [(I-1)+(J-1)+(K-1)]+L$ in our example). A zero value for LR indicates a perfect fit, while the extent to which LR exceeds 0 measures the lack of model fit. When the total number of cells in the resulting multi-way frequency table is very large, an insufficient sample size will result in sparse data or many empty cells. In this situation, the χ^2 distribution should not be used to compute the p-value, as asymptotic theory does not apply. Instead, the bootstrap approach can be used (Langeheine, Pannekoek, and Van de Pol, 1996).

Alternatively, measures such as AIC (Akaike, 1974) or BIC (Schwarz, 1978) were developed especially for model selection. In addition to comparing the maximum value of the likelihood function, they take the number of parameters into consideration and add a penalty term. Let h denote the total number of free parameters in the model and L_M denote the maximum value of the likelihood function. AIC is defined as $AIC = 2h - 2log(L_M)$, with smaller values indicating better models. On the other hand, BIC puts a bigger penalty on the number of parameters in the model. It is defined as $BIC = hlogN - 2log(L_M)$. Similarly, models with smaller BIC values are preferred.

2.2.3 Relation to Cluster Analysis

Traditional cluster analysis usually uses an unsupervised classification algorithm that groups together observations that are near each other according to some definition of distance. Latent class models or latent profile models assume that the subjects belong to one of the

L latent classes and investigate the relationship between the manifest variables and the latent variables. Since the posterior probability of a subject's class membership can be obtained by utilizing Bayes' rule and estimates from the model, these models can also be used as classification tools. The first explicit connection between latent class analysis and cluster analysis was made by Wolfe in 1970. Since then, the latent class/profile approach has become increasingly popular in cluster analysis, and have taken on various names, including mixture likelihood clustering (McLachlan and Basford 1988; McLachlan et al. 1999), model-based clustering (Baneld and Raftery 1993; Bensmail et. al. 1997), Bayesian classification (Cheeseman and Stutz 1995), unsupervised learning (McLachlan and Peel 1996), latent discriminant analysis (Dillon and Mulani, 1999) and latent class cluster analysis (Vermunt and Magidson 2000).

Although latent class cluster analysis is often used in practice, most work has in fact focused on models with continuous manifest variables, or latent profile models. Generally, the continuous manifest variables are assumed to be normally distributed within latent classes, possibly after applying an appropriate non-linear transformation (Baseld and Raftery 1993; McLachlan 1988; McLachlan et. al. 1999).

An important difference between standard cluster analysis and latent class clustering is that the latter is a model-based approach. It allows researchers to specify a statistical model for the population from which the sample is drawn. This has several advantages. First, the method is very flexible, since both simple and complicated distributional forms can be assumed and covariates can be included in the model. Restrictions on parameters can be easily imposed, and statistical tests may be developed to check their validities. Secondly, the maximum likelihood approach can be used for parameter estimation. Compared to other choices of cluster criterion involving an *ad hoc* defined distance, maximizing the likelihood may be less arbitrary. This also guarantees that the results are the same, regardless of whether the variables are normalized or not, so no decisions have to be made about whether and how to scale the observed variables. In contrast, scaling is always an issue in standard cluster analysis. Finally, latent class analysis yields posterior probabilities of a subject's class membership, which represent some uncertainty. This allows one to perform a "soft" classification as used in the mixed membership models or fuzzy clustering techniques

(Ruspini, 1969). An important difference between these two approaches is that, in fuzzy clustering a subject's grades of membership are parameters to be estimated, while in latent class clustering the posterior probabilities of a subject's class membership are computed from the estimated model parameters and the observed variables. This makes it possible to classify other objects from the same population, which is not possible with standard fuzzy cluster techniques.

2.2.4 Relation to Factor Analysis

In psychometrics, researchers have found that a covariance matrix generated by a latent profile model can be perfectly fitted by a factor analysis model. This has been shown analytically by Bartholomew in his book "Latent variable models and factor analysis". To see this, let π_s be the probability of a subject belonging to latent class d, $d = 0, \ldots, L-1$. Let $\mu_i(d)$ be the mean of manifest variable y_i for a subject in latent class d and $\sigma_i^2(d)$ be its variance. Then,

$$\begin{aligned} E(y_i) &= \sum_{d=0}^{L-1} \pi_d \mu_i(d) \\ var(y_i) &= \sum_{d=0}^{L-1} \pi_d \sigma_i^2(d) + \sum_{d=0}^{L-1} \pi_d [\mu_i(d) - \bar{\mu}_i]^2 \\ cov(y_i, y_j) &= \sum_{d=0}^{L-1} \pi_d [\mu_i(d) - \bar{\mu}_i][\mu_j(d) - \bar{\mu}_j] \quad \text{for } i \neq j, \end{aligned}$$

where $\bar{\mu}_i = \sum_{d=0}^{L-1} \pi_d \mu_i(d)$. The covariance matrix can be written as

$$cov(\mathbf{Y}) = \mathbf{L}\mathbf{L}' + \mathbf{\Psi},$$

where $\mathbf{\Psi}$ is a diagonal matrix with (i,i)th element $\sum_{d=1}^{L-1} \pi_d \sigma_i^2(d)$ and the (i,d)th element of $\mathbf{\Psi}$ is given by $l_{id} = \sqrt{\pi_{d-1}}[\mu_i(d-1) - \bar{\mu}_i]$ when $i \neq d$. "$d-1$" in the formula is due to the fact that d start with a value of 0.

Therefore, $cov(\mathbf{Y})$ has the same form as the covariance matrix for the linear factor model with a normally distributed latent variable, except that the columns of $\mathbf{L}$ are linearly

dependent because

$$\sum_{d=0}^{L-1} \sqrt{\pi_d} l_{id} = 0 \quad \text{for } i = 1, \ldots, N.$$

However, there exists a $N \times (L-1)$ matrix $\mathbf{\Lambda}$ with linearly independent columns such that $\mathbf{LL}' = \mathbf{\Lambda\Lambda}'$. This means that the covariance matrix from the latent profile model with L classes is indistinguishable from a factor model with $L-1$ factors. This result has been taken further by Molenaar and von Eye (1994), who showed that the converse is also true. That is, for every covariance matrix generated from a linear factor model there exists a latent class model with the same covariance structure.

As a result, at the level of second moments, a latent profile model with L classes is equivalent to a factor model with $L-1$ factors. This phenomenon raises many criticisms against factor analysis and latent class models. As pointed out by Molenaar (2007), "analyses in inter-individual variation appear to be entirely insensitive to the actual presence of substantial heterogeneity (unique individuality) in the population". Should the individual differences be attributated to their distinctions among common dimensions (such as the factors in factor analysis) or should to various types (such as the classes in latent class models) in the population?

Molenaar (2007) commented on this problem by stating that there may not be a fundamental distinction between dimensions and types. He pointed out the uniqueness in every individual and quoted Toomela (2007)'s remark: "the problem with an idiographic (individual-focused) approach is that the number of possible models describing an individual is potentially very large. Every individual is in some sense unique; unique models may also be necessary for every different situation, setting, developmental phase, historical period, even a mood swing." Hence, the scientifically relevant question is, as Molenaar argued, how to conceptualize the situation; "in which relevant senses individuals are unique and in which other relevant senses they are not unique". Muthén endorsed this argument and noted that this problem should be seen as merely "two ways of looking at the same reality. The factor analysis provides information about underlying dimensions and how they are measured by the items, whereas the latent profile analysis sorts individuals into clusters of individuals who are homogeneous with respect to the item responses. The two analyses are

not competing but complementary."

2.2.5 Modeling with Covariates

One advantage of latent class modeling over cluster analysis is that it is a model based approach. It offers an opportunity to include covariate effects. Both the latent class membership and the manifest variable within each class can be influenced by covariates. The former allows the latent class membership prevalence to vary with subjects who have different characteristics, whereas the latter relaxes the homogeneity assumption within classes, or the local independence assumption, in the sense that the manifest variables can depend on covariates in addition to their latent class membership. Theoretically, parameter estimates can be obtained by a maximum likelihood approach. However, partially due to computational difficulties, latent class analysis with covariates is often performed in an *ad hoc* three step procedure:

- Step 1: traditional latent class analysis without covariates;
- Step 2: classifying individuals into classes based on posterior probabilities
- Step 3: logistic regression relating predicted class membership to covariates.

The regression model in the third step ignores the prediction error and treats the predicted class membership as the true membership. Bolck, Croon and Hagenaars (2004) demonstrated that such a three step approach can underestimate covariate effects. To see this, let D denote the true class membership and $\hat{D}$ denote the predicted value. Let X denote the covariates and T denote the manifest variable. The relationship between these four variables can be expressed as in the following diagram:

$$X \longrightarrow D \longrightarrow T \longrightarrow \hat{D}$$

As a side note, this diagram suggests that the three step approach essentially only considers covariate effects when modeling class membership. The conditional distribution of T in each class is not affected by X, given the true class membership.

The quantity of interest is how the latent class membership is associated with covariates. The three step approach uses the predicted membership to explore these associations. Based on the model shown above, the following equation shows that the conditional probability of the predicted class membership, given covariates $P(\hat{D}|X)$ is, in general, not equal to the conditional probability of the true class membership given covariates $P(D|X)$:

$$\begin{aligned} P(\hat{D}|T) &= \sum_d \sum_t P(\hat{D}, T, D|X) = \sum_d \sum_t P(\hat{D}, T|D, X)P(D|X) \\ &= \sum_d P(\hat{D}|D, X)P(D|X) = \sum_d P(\hat{D}|D)P(D|X) \end{aligned}$$

In fact, they are only identical when there is no prediction error (i.e., when $P(\hat{D} = d|D = d\prime)$ equals 1 when $d = d\prime$, and equals 0 otherwise). Consequently, the three step approach will lead to biased results. Bolck, Croon and Hagenaars (2004) proposed a correction in the third step by introducing a weighting matrix of classification error.

Another way to obtain unbiased results with the three step approach is proposed by Petersen, Bandeen-Roche, Budtz-Jørgensen and Larsen (2012). They considered latent class models with binary manifest variables and adopted the "Bartlett" method (Bartlett, 1937, 1938) in step two. The "Bartlett" method has been used in factor analysis with a similar three step approach. It obtains the latent factors by performing a linear regression of the manifest variable on the estimated factor loading. Skrondal and Laake (2001) showed that the "Bartlett" method yields consistent estimators in step three for the covariate effects on the factors. Petersen and his colleagues (2012) modified this approach to use in latent class models. They considered J binary manifest variables for each individual and proposed to obtain the class membership D by regressing $\mathbf{T} = (T_1, \ldots, T_J)$ on the estimated distribution matrix of the manifest variables given class membership $P(\mathbf{T} = \vec{1}|D = d)$. Specifically, they created an indicator vector for each subject's latent class membership

$$\mathbf{G}(D_i) = \big[g_0(D_i), \ldots, g_{L-1}(D_i)\big]^t = (1_{D_i=0}, \ldots, 1_{D_i=L-1}).$$

This results in a linear relationship, as required by the "Bartlett" method, between an individual's binary item response probabilities and its class membership, $E(T_{ij}|D_i) = \pi_j^t \mathbf{G}(D_i)$, where $\pi_j = (\pi_{j0}, \ldots, \pi_{j(L-1)})$ with $\pi_{jd} = P(Y_{ij} = 1|D_i = d)$, and the superscript t denotes the transpose. Then, analogous to the "Bartlett" method, individuals' latent class

memberships are treated as if they were fixed parameters and estimated from the following equation:

$$E(\mathbf{T}_i|D_i) = \pi\mathbf{G}(D_i) = \sum_{d=0}^{L-1} \pi_d g_d(D_i),$$

where $\pi = (\pi_1, \ldots, \pi_J)^t$. If π were known. Theory about weighted least square estimators guarantees the consistency of the estimated latent class membership $P(D_i = d) = P_d(X_i\beta)$, which in turn leads to consistent estimates for covariate effects β in step three. For the estimated value $\hat{\pi}$, Bandeen-Roche and her colleagues (1997) showed that $\hat{\pi} \to^p \pi$ by marginalizing over the covariates and applying theory about maximum likelihood estimators. Then the uniform integrability completes the proof.

On the other hand, attempts to simultaneously estimate class membership and covariate effects have been made dating back to the mid-1980's (Clogg and Goodman, 1984; Formann 1985, 1992). These methods incorporate covariates by stratifying combinations of risk factors, and are thus highly constrained when applying them to categorical covariates. Dayton and Macread (1988) proposed a "concomitant variable" model, which allows class membership to depend on continuous variables as follows:

$$P(T_1 = t_1, \ldots, T_J = t_J) = \sum_{d=0}^{L-1} \big\{ P(D = d|X) \prod_{j=1}^{J} P(T_j = 1|D = d)^{t_j} [1 - P(T_j = 1|D = d)]^{1-t_j} \big\}.$$

Here, the term $P(D = d|X)$ defines a latent polytomous regression of class membership D on covariates X. This method has been formalized by Bandeen-Roche, Miglioretti, Zeger and Rathouz (1997), and extended under the term "latent class regression" to allow for a more general relationship between class membership and covariates.

Up to the present time, most research on latent class modeling with covariates has focused on allowing the latent class membership to depend on covariates, while little work has been done to allow for the dependence between manifest variables and covariates within each class. This is partially due to the fact that the primary research interest is usually the structure of the unobserved variables and how they are associated with covariates, such as how health status is related to risk factors. In addition, simultaneously modeling latent class membership as well as manifest variables with covariates can be computationally

challenging. It also raises substantial identification questions. Huang and Bandeen-Roche (2004) gave a discussion on the identifiability issue when both the latent variable and the manifest variable are categorical and the covariate effects are modeled with a logit link. One of the conditions necessary to ensure the identifiability in their paper is that the covariate effects are constrained to be equal across classes. Intuitively, if the covariate effects can vary across classes, we may not be able to distinguish this effect of the covariates from the heterogeneity introduced by another class.

2.3 Latent Variable Modeling in Medical Diagnosis

2.3.1 Latent Gold Standard

An important task in medical research, as well as in many other fields, is to evaluate and compare the accuracy performance of multiple diagnostic tests. This is usually done by comparing the test results with true disease status. As mentioned before, the evaluation becomes challenging when gold standard information is not available. Latent variable models can be helpful in such situations. In fact, if we consider the absent gold standard as a latent variable indicating the presence or absence of a disease or condition, the problem with multiple diagnostic tests without a gold standard can be studied in the framework of latent class and latent profile models. However, this important application of the latent class models in diagnostic medicine has received little appreciation in the early development of latent class models. The paper by Dawid and Skene (1979) is probably the first modern paper using a latent class-like approach to analyze rater agreement data. In the epidemiological literature, Hui and Walter (1980) developed mathematical models for estimating the accuracy of two diagnostic tests without a gold standard under special conditions, perhaps not at first seeing how it relates to the earlier work of Lazarsfeld, Goodman, and others. Later, Walter gradually extended this work to consider multiple diagnostic tests (Walter, 1984; Walter and Irwig, 1988) and point out its connection to latent class models.

In a classic diagnostic test setting, the test results are binary, indicating whether a subject is diagnosed positively or negatively with respect to an attribute of interest. The primary interest is usually to quantify the diagnostic performance of the tests, summarized

by quantities such as sensitivity and specificity. Consequently, Walter's models are parameterized with respect to these quantities and developed for binary data. In their notation, θ is the proportion of subjects who are actually "positive" for an attribute of interest, or the prevalence. K is the total number of tests, and α_k is the false positive rate for the kth test, making $1-\alpha_k$ the corresponding specificity. The false negative rate of the kth test is denoted by β_k, so $1-\beta_k$ is the corresponding sensitivity. Let Y_k denote the binary result given by the kth test, and then the model can be written as follows,

$$P(Y_1 = y_1, \ldots, Y_k = y_k) = \theta \prod_{k=1}^{K} \beta_k^{1-x_k}(1-\beta_k)^{x_k} + (1-\theta) \prod_{k=1}^{K} \alpha_k^{x_k}(1-x_k)^{1-x_k}.$$

Walter and Irwig (1988) also considered the heterogeneity in the population and allowed for data drawn from several subpopulations with different prevalences, as well as for different sensitivities and specificities of each test. This approach, in a loose sense, is similar to the idea of incorporating covariates in latent class models by stratifying combinations of risk factors, as in the work of Clogg and Goodman (1984), Formann (1985, 1992) and others.

Since then, latent class analysis has been widely used in existing literature on estimating diagnostic accuracy without a gold standard (Espeland and Handelman, 1989; Henkelman et al, 1990; Uebersax 1993; Hui and Zhou 1998; Albert et al, 2004; Zhou et al, 2005). Espeland and Handelman (1989) extended Haberman's approach and used loglinear modeling in latent class analysis to study radiographic diagnosis data from five dentists. Henkelman, et al. (1990) and Uebersax (1993) focused more on developing a model to explain how ratings or diagnoses are made. They considered the discrete diagnoses as categorized results based on an underlying mixture of two continuous traits, one from the diseased population and one from the healthy population. Normality is assumed for both distributions. Zhou and his colleagues (1998, 2005) used nonparametric methods to model the discrete diagnostic results directly without assuming a parametric mixture distribution. Albert, et al. (2004) cautioned about model selection with latent class models applying to binary diagnosis tests. Most of the work is derived when both the disease status and the test results are binary. Although some models, such as in Uebersax (1993) and in Zhou, et al. (2005), have extensions to categorical test results, the disease status, or the gold standard is still binary.

2.3.2 Dependence among Diagnostic Tests

In classic latent class analysis, the manifest variables within each class are assumed independent with each other. In the diagnostic test application of the latent class models, this requires the conditional independence assumption among the test results, that is, the test results are independent with each other given the true status of the disease or other observed attributes of interest. This assumption is often made when applying latent class models to evaluate diagnostic tests, and it can be arguably true when the tests are performed separately and based on different biological processes. However, as pointed out by Vacek (1985), this assumption is not always justifiable. The misclassification errors can be correlated when tests are based on the same physiologic phenomenon, such as a particular antibody reaction. Something which inhibits the reaction or causes a false reaction for one of the tests may have a similar effect on others. Another example, given by Dunn (1989) is when there are overlapping items included in two psychiatric screening questionnaires.

Many researchers have developed models to describe the dependence structure among diagnostic tests. In fact, many of the methods mentioned in the last section have natural extensions to dependent tests, as pointed out by their authors. For example, Henkelman, et al. (1990) proposed that discrete diagnoses were made based on an underlying continuous trait with a mixture of normal distributions resulting from the mixture of diseased and healthy groups. With a multivariate normal distribution, this model allows for dependence among test results, although the results can be highly dependent on the parametric assumption. Without parametric assumptions, loglinear models provide a comprehensive framework for investigating the correlation structure, and have been used in latent class modeling. For example, Espeland and Handelman (1989) summarized the discrete test results in multi-way contingency tables and used loglinear modeling with two-way interactions. However, including higher interactions may cause an identifiability issue. Similarly, if we use a different parameterization as in Hui and Walter (1980), dependence among tests can be characterized by additional parameters of the correlation among sensitivities and specificities of different tests. Dendukuri and Joseph (2001) regard this model as a fixed effect model, as the covariances among tests are introduced directly as parameters. This model is studied by Vacek

(1985) and Hagenaars (1988) with two correlated tests. Likewise, when multiple tests are involved, one may run into an identifiability issue without constraints on the correlations. However, it is usually hard to consider proper constraints for correlations among sensitivities and specificities of different tests. Instead, Branscum et al (2005) applied a Bayesian approach to this model, supplying informative priors to sensitivities and specificities of some tests that have been studied previously, rather than to the correlations.

In contrast to the fixed effect models, Qu, Tan and Kutner (1996) proposed a random effect latent class model to reflect the dependence in binary data. They considered that correlation among test results are introduced by another latent variable – a subject level random effect. Specifically, let y_k denote the binary test result from the kth test, D denote the disease status and R denote the random effect. The probability of getting a positive test result conditioned on the disease status and the random effect is given by

$$P(Y_k = 1|D = d, R = r) = \Phi(a_{kd} + b_{kd}r), \quad d = 0, 1, \quad R \sim N(0, 1),$$

where Φ is the cumulative distribution function of a standard normal variable. The sensitivity and specificity of the kth test are obtained by integrating over the random effect R as follows,

$$\begin{aligned} \text{Sensitivity}_k &= P(Y_k = 1|D = 1) = \int_{-\infty}^{+\infty} \Phi(a_{k1} + b_{k1}r)d\Phi(r) = \Phi(\frac{a_{k1}}{\sqrt{1 + b_{k1}^2}}) \\ \text{Specificity}_k &= P(Y_k = 0|D = 0) = 1 - P(Y_k = 1|D = 0) = \Phi(\frac{-a_{k0}}{\sqrt{1 + b_{k0}^2}}). \end{aligned}$$

Albert, et al. (2001) extended this approach and used mixture models to characterize the heterogeneity when analyzing immunohistochemical assays in bladder tumors. They argued that a specimen from the tumor positive group may consist of two subgroups. One is severe enough and always tests positive, while the other group of subjects is prone to diagnostic error. Similarly, the specimen from the tumor negative group also consists of two subgroups, one of which is always tests negative. This is essentially a four class latent class model. This model can also be regarded as an extension of the model proposed by Walter and Irwig (1988), which assumes that the data are drawn from several subpopulations.

2.3.3 Covariate effects

Covariates may influence the magnitude of the test results, and affect test accuracy in different subgroups. For example, in psychological diagnosis, it has been shown that women are more likely to be given positive diagnosis than men (Munch, 2004). The minimental state examination (MMSE) for cognitive impairment, which is commonly used to screen for dementia, has lower (worse) scores in older subjects. Ignoring covariate effects in a diagnostic accuracy study can lead to biased results. To illustrate this, we consider a hypothesized example.

Let Z denote a binary covariate, such as gender. Suppose that for subjects with $Z = 0$, test results follow a normal distribution $N(0, 2^2)$ among controls and $N(2, 2^2)$ among cases, while as for subjects with $Z = 1$, test results follow a normal distribution $N(3, 2^2)$ among controls and $N(5, 2^2)$ among cases. That is, covariate Z shifts the test results. Assuming equal numbers of cases and controls, the distributions of test results for cases and for controls stratified by Z are shown in the top left panels of Figure 2.1. In each stratum, the test distribution for cases is two units to the right compared to that for controls. The separation of test distributions among cases and among controls are the same when comparing these two strata. Since the ROC curve is invariant to monotonic transformation, the ROC curve is conditional on each stratum, also called the covariate specific ROC curves, are the same for $Z = 0$ and $Z = 1$.

However, if we ignore covariate Z and calculate the ROC curve based on pooled data, the results may change. We consider two scenarios here. In the first scenario, covariate Z has the same distribution among cases and controls: $P(Z = 1|D = 0) = P(Z = 1|D = 1) = 0.5$. In the second scenario, there are more subjects with $Z = 1$ in cases, and less subjects with $Z = 1$ in controls: $P(Z = 1|D = 0) = 0.3$ and $P(Z = 1|D = 1) = 0.7$. The distributions of test results for cases and for controls in pooled data are shown in the bottom left panels of Figure 2.1. The covariate-specific ROC curve and the ROC curves ignoring covariate Z for both scenarios are shown on the right side on Figure 2.1. We can see that, in scenario 1, the ROC curve for pooled data is uniformly below the covariate-specific ROC curve, and in scenario 2, the ROC curve for pooled data is uniformly above the covariate-specific ROC

curve.

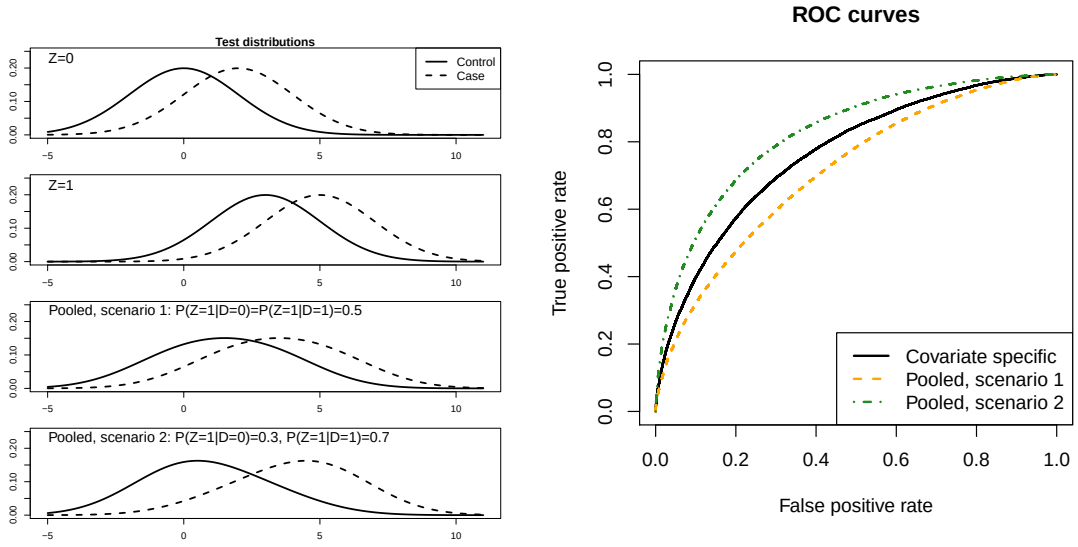

Figure 2.1: Hypothesized example showing covariate effect on ROC curve.

This example illustrates the importance of considering covariate effects in the assessment of diagnostic test accuracy. In addition, including covariates in the model to allow for dependence between test results and subjects' characteristics within disease classes is another way to relax the conditional independence assumption for the test results. In fact, this approach is often preferred if we have covariate information, because compared to the random effect modeling approach, it can provide more specific information on how the covariates influence test performance. In contrast to the classical settings of latent class models, where more interest is in the relationship between latent class membership, and in covariates such as the relationship between an individual's intelligence/success and education, or between general health and health behaviors, in the studies of diagnostic tests we focus more on the relationship between the observed test results and the covariates within each disease class. In other words, the focus is on how the test performance varies for subjects with different characteristic.

However, probably because methods with covariates in the general latent class analysis

have not been fully developed, limited work has been done on incorporating covariate information in the evaluation of diagnostic tests without a gold standard. Models allowing data to be drawn from several subpopulations, as described by Walter and Irwig (1988), inexplicitly used covariate information and stratified populations into groups. Qu and Hadgu (1998) discussed random effect latent class models with binary covariates. Both models can be viewed as a reformulation of models proposed by Clogg and Goodman (1984), Formann (1985, 1992), wherein they incorporated covariates by stratifying combinations of risk factors.

Allowing disease prevalence to vary with covariates can also be useful. Although this part of the work is limited in diagnostic test setting, the methodologies in general latent class models discussed in section 2.2.5 can be easily extended to this area. However, due to computational difficulty and substantial identifiability issues, little work has been done to simultaneously include covariate effects on test performance in addition to disease prevalence. An exception can be found in Huang and Bandeen-Roche (2004), who proposed a model where both the latent variable and the manifest variable are categorical and the covariate effects are modeled with a logit link. They also discussed the identifiability issue for this model.

2.3.4 Indentifiability

While latent class models arise in many statistical applications, conditions for their identifiability have not been established for general cases. Theoretically, latent class models are not strictly identifiable, as they suffer from the "label switching" problem – that is, the distribution is identical if class labels were switched. Practically speaking, the "label switching" problem of the latent class models is not an pressing issue because the group labels are generally apparent, for example, when indicating which group is the diseased group and which is the healthy one. However, aside from "label switching", it is often still hard to establish the global identifiability of the models because the likelihood functions of these models usually have multiple local maximas. Research on the identifiability issue of the latent class models has largely been focused on the local identifiability (McHugh 1956;

Goodman 1974; Formann 1992). By definition, a distribution F is locally identifiable at γ_0 if F is invertible in a neighborhood of γ_0. It is often believed that when the degrees of freedom in the data meet or exceed the number of parameters in the model, the model is identifiable. Goodman (1974) showed that this is not the case. Elmore, Hall and Neeman (2005) pointed this out again in the diagnostic test setting. Consider Hui and Walter (1980)'s model, for example. A model with two binary tests and three subpopulations has degrees of freedom $(2^2-1)\times 3 = 9$. The total number of parameters is seven (two sensitivities, two specificities and three prevalences). Despite the higher value of degrees of freedom, this model lacks identifiability, as explained by Johnson and Hanson (2005).

In latent class models, McHugh (1956) proposed sufficient conditions for the local identifiability of models with dichotomous observed variables, and Goodman (1974) extended these conditions to polytomous variables. In the special application of latent class models in diagnostic tests, more work on the identifiability issue has been accomplished. Hall and Zhou (2003) and Hall, et al. (2005) studied the nonparametric identifiability of models of K tests with a conditional independence assumption within each of the M subpopulations, and provided sufficient conditions for model identification of $K \geq (1+o(1))6M\log M$. For an M-component mixture of binomial distributions, $B(K, p_m)$, it has been shown that $K \geq 2M-1$ is both necessary and sufficient for model identification (Teicher, 1961, 1963; Blischke 1964). Based on this result, with an additional assumption of identically distributed tests, Hettmansperger and Thomas (2000) and CruzMedina et al (2004) dichotomized the test results and concluded that $K \geq 2M-1$ is sufficient for identification of such models. Other work has been focused on finding sufficient conditions to guarantee that the Jacobian matrix of the mixture distribution has full column rank (Huang and Bandeen-Roche, 2004; Kasahara and Shimotsu, 2008). However, none of the conditions are both necessary and sufficient. In addition, sufficient conditions for more general models, such as models that allow for dependence structure among test results, have not been established.

Chapter 3

INDEPENDENT ORDINAL DIAGNOSTIC TESTS WITH UNKNOWN ORDINAL GOLD STANDARD

In this chapter I consider evaluations of multiple ordinal diagnostic tests when a gold standard is unknown and possibly on an ordinal scale, and further discuss how to adopt a latent class modeling approach and use a nonparametric maximum likelihood method for estimating and comparing the accuracy of the tests. The proposed model relies on a classic conditional independence assumption. In other words, I assume that, given true disease status, the test results are conditionally independent of each other. Methods that do not make this assumption will be discussed in the next chapter. In addition, I propose an accuracy measure which can be viewed as a high dimensional counterpart to the area under the ROC curve, and discuss some graphical representations of the high dimensional results.

3.1 Introduction

3.1.1 Motivating Example: TCM

Traditional Chinese medicine dates back more than 5,000 years. It is still heavily practiced in many hospitals and clinics in China, along with western medicine. It has also become popular in Europe, America and many other places around the world. Some possible beneficial effects of TCM on pain management, depression, diabetes, cancer, etc. have been observed. The effort put into studying its clinical efficacy (e.g. Eisenberg et al. 2002; Allen 2006; Shang 2007) and into developing appropriate statistical methods in has increased in recent years (e.g. Feng, 2006; Zhang, et al., 2004, 2007, 2008). While the exact number of people who use TCM in the United States is unknown, it was estimated in 1997 that more than 1 million patients receive TCM each year. According to the 2007 National Health Interview Survey, an estimated 3.1 million U.S. adults had used acupuncture in the previous

year. In addition, according to this same survey, approximately 17 percent of adults use natural products, including herbs, making TCM the most commonly used therapy.

Despite of popularity of TCM, it has not been well explained or accepted by modern sciences. One primary concern is that TCM practice is based on its empirical diagnosis of a symptom or a disease. Specifically, in TCM, the cause of illness is explained by five main theories, the Yin and Yang theory, the five paths (elemental energies) theory, the vital organs theory, the theory of fundamental substances, and the theory of the meridians (Kaptchuk, 2000). Diagnosis is based on examining the interaction of the five main theories and the human body, which is usually carried out through the classic "four diagnostic methods", including observing (especially the tongue), hearing/smelling, asking/interviewing, and touching/palpating (especially the pulse). Consequently, the diagnosis in TCM is usually subjective and can be influenced by the doctor's knowledge and clinical experience. Thus, knowledge about the performance of the doctors is useful for patients in choosing the right doctor, as well as for doctors in making improvements and preventing misdiagnosis. Nevertheless, some current research focuses on developing standard and more objective diagnostic guidelines for TCM practitioners. Doing so requires the ability to compare and evaluate TCM doctors' diagnostic accuracy. More immediately and importantly, however, it necessitates the ability to understand and examine diagnoses on a symptom level, since symptoms are the basic component of TCM diagnosis.

In evaluating TCM doctors' accuracies for symptom diagnosis, the first challenge is that the true symptom status is often unavailable. First of all, TCM theory has not been fully accepted by modern science, so efforts to understand and verify TCM are limited as we cannot use a gold standard defined according to TCM theory. Secondly, at least currently, a gold standard cannot be obtained using modern techniques. This is because many concepts of symptoms and syndromes are unique to TCM. They can hardly be found in the theories of modern medicine or measured by any other medical instrument. In fact, TCM has a unique view of the world and the human body that is very different from western medicine concepts (O'Brien and Xue, 2003; Ernst 2006). This view is based on the ancient Chinese perception of humans as microcosms of the larger, surrounding universe – interconnected with nature and subject to its forces. The human body is regarded as an organic entity in which the

various organs, tissues, and other parts have distinct functions but are all interdependent. Health and disease are related to the balance of the functions. Therefore, the symptoms and syndromes in TCM are not exactly the same as those in western medicine, and so far cannot be measured directly.

Another issue in analyzing TCM data is that most of the diagnostic results are given on an ordinal scale, usually with four categories, denoting the symptom severity of the patients. Collapsing the categories can result in a loss of information. In many situations it is unaffordable, because a binary result is not adequate for further diagnosis, and for choosing the right therapy. Similarly, ordinal diagnosis is also common in other areas of medical practice. For example, the stage of cancer progression at detection is described on an ordinal scale, ranging from localized cancer to distant metastases. Different types and quantities of medication might be assigned to patients with different gradations of illness, in order to achieve better curative results while simultaneously reducing noxious side-effects. The ability to differentiate among different severities of disease status, in addition to the ability to detect the presence or absence of the disease, is therefore important. A standard ROC curve methodology cannot be applied directly to evaluate the diagnostic accuracy of tests when the true disease status is ordinal.

The methodology work in this chapter is motivated by these two challenges in assessing the diagnostic accuracy of TCM doctors who wish to detect a particular symptom–the absence of a gold standard and ordinal-scale symptom status.

3.1.2 Literature Review

When the gold standard is binary, e.g. disease or non-disease, many methods have been discussed on how to estimate the accuracy of multiple binary scaled tests without a gold standard. In a review article, Hui and Zhou 2008 note that almost all available statistical methods focus on binary test results. Henkelman, Kay and Bronskill (1990) proposed a maximum likelihood estimation method for the ROC curves of 5-point rating scale tests without a gold standard using a multivariate normal mixture latent model. Their method assumes that the ordinal rating is a categorized version of some underlying latent variables

that follow a multivariate normal distribution. However, this assumption may not be satisfied in practice. Zhou, Castelluccio and Zhou (2005) proposed a nonparametric method of estimating ROC curves of ordinal-scale tests in the absence of a gold standard, without a distributional assumption, but still with a focus on binary disease status. This method can be generalized in the latent class modeling framework to accommodate situations with ordinal disease status, and is a basis of the method that I propose in this chapter.

On the other hand, commonly used accuracy measures, such as sensitivity, specificity and the ROC curve cannot be directly applied when the true disease status is ordinal. In the special case of three ordered categories, several authors have proposed the three-way ROC surface as a graph built on the three true classification rates (Mossman, 1999; Nakas and Yiannoutsos, 2004). In addition, a nonparametric likelihood-based approach has been proposed by Chi and Zhou (2010) to construct the empirical ROC surface in the presence of verification bias. Analogous to the area under the traditional ROC curve, the volume under the surface (VUS) has also been proposed to assess the overall accuracy of diagnostic tests. Fawcett (2001) investigated the performance of different classification strategies for multiple classes. Based on the relationship of the AUC to the Mann-Whitney-Wilcoxon test statistic, Hand and Till (2001) extended the definition the AUC to the case of more than two classes by averaging pair-wise comparisons. However, although the AUC can be viewed as a special case of this summary statistic, the VUS cannot, and the interpretation of this summary statistic is not clear.

In this Chapter I discuss a nonparametric maximum likelihood method for estimating and comparing the accuracy of ordered-scale tests when the true disease status is also ordinal and unavailable. I also extend the volume under the three-way ROC surface into high dimensions and discuss other graphical representations of the results with an application in TCM.

3.2 Estimation of Diagnostic Accuracy

3.2.1 Notation and Setting

Assume that there are N patients and K tests, I let D_i be the missing or unobservable true disease status for the ith patient, taking a value on an ordinal scale from 0 to $L-1$, representing the severity level of the disease status from absent to severe. A vector of K imperfect test results, or diagnoses from K doctors, is denoted by $T_i = (T_{i1}, \ldots, T_{iK})$, where T_{ij} also has an ordinal scale from 0 to $J-1$. The possible values of test results can be more than those of the disease status, i.e. $J \geq L$. The diagnosis is made by choosing a sequence of cut points $\{j_0 = 0, j_1, \ldots, j_L\}$, where $j_d \leq j_{d+1}$ for every d=0, $\ldots$,$L-1$ to classify patients into a certain severity group. That is, the kth test or doctor will diagnose a patient as having the disease with severity level d if the test result is between cut point j_d and j_{d+1} ($j_d \leq T_{ik} < j_{d+1}$).

I assumed that the test results reflect the true disease status of a patient and that, given the true disease status, the test results are independent of each other. That is, I assumed that the conditional independence assumption holds. I further assumed that each doctor's performance was consistent, and that the diagnostic probabilities $P(T_{ik} = j|D_i = d)$ did not vary with patients. In the case with a set of cut points this means that, for each given doctor, the cut points do not vary with different patients. However, the cut points of different tests/doctors can be different, i.e. the cut points of the kth test/doctor are $j_{k_d}, d = 0, \ldots, L-1$, where $j_{k_0} = 0$, and $j_{k_d} \leq j_{k_{d+1}}$. I omited the subscript k in the notation for the cut points, for the sake of simplicity in the presentation.

In addition, to facilitate my model, I created a dummy variable for the test results. Specifically, I let $\mathbf{Y}$ be a $N \times K \times J$ array with binary entries y_{ikj}, such that $y_{ikj} = 1$ if the test result of the kth diagnostic test for the ith patient is j (i.e. $T_{ik} = j$), and y_{ikj} =0 otherwise, where $i = 1, 2, \ldots, N, k = 1, 2, \ldots, K$, and $j = 0, 1, \ldots, J-1$. Then I could construct a $K \times J$ dimensional vector of binary variables $\mathbf{y_i} = (y_{i10}, \ldots, y_{i1j}, \ldots, y_{ik0}, \ldots, y_{ikj-1})$, that reflects the test results of the ith patient.

3.2.2 A Latent Class Model

The quantities of interest are essentially the prevalence of each disease severity group $P_d = P(D = d)$ and the conditional diagnostic probabilities of each test within a severity group $\phi_{dkj} = P(T_{ik} = j|D_i = d)$ for every $d = 0, 1, \ldots, L-1, k = 1, 2, \ldots, K$, and $j = 0, 1, \ldots, J-1$. If one considers the gold standard as a latent variable, which divides the population into L disease severity groups, then the observed data can be described by a latent class model as shown in Figure 3.1.

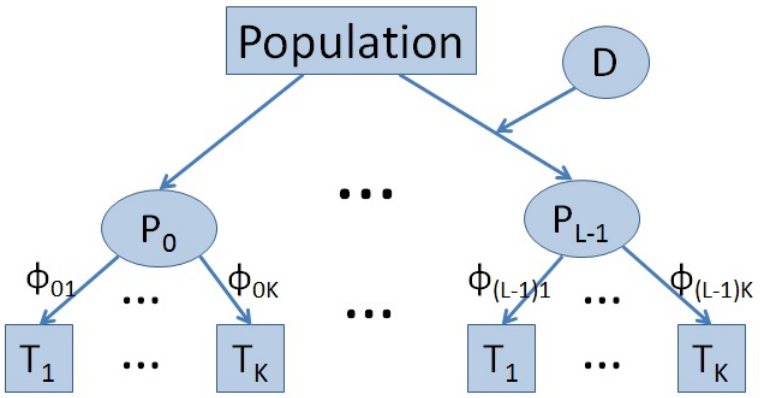

Figure 3.1: A latent class model

In accordance with the notation commonly used in graphical models, I used circles to denote unobserved variables, and rectangles to denote observed variables. In this graph, the first level is the latent structural $P(D = d)$, which describes disease prevalence. The second level is the observed test results. It was modeled conditional on each disease group in the measurement model $P(T_{ik} = j|D_i = d)$. The observed test results were then a mixture of the component distributions:

$$P(\vec{T}) = \sum_{d=0}^{L-1} P(D = d)P(\vec{T}|D = d)$$

Essentially, this approach attributes the heterogeneity of the test results to disease severity groups and models the test performance conditioned on each group, then chooses the most likely latent structure, given the observed data.

Since the tests were discrete, their distributions within disease groups were modeled nonparametrically in their empirical forms. Under conditional independence, the conditional probability of having ith patient's diagnostic result y_i, given their disease status, D_i can be written as follows:

$$\begin{aligned} g_d(y_i) &= P(y_i \,|D = d) \\ &= \prod_{k=1}^{k} \prod_{j=0}^{J-1} \big[P(T_k = j \,|D = d)\big]^{y_{ikj}} \\ &= \prod_{k=1}^{K} \prod_{j=0}^{J-1} \big[\phi_{dkj}\big]^{y_{ikj}}. \end{aligned}$$

The likelihood contributed by the ith patient has a mixture form of

$$\begin{aligned} P(y_i) &= \sum_{d=0}^{L-1} P(y_i \,|D_i = d) P(D_i = d) \\ &= \sum_{d=0}^{L-1} p_d g_d(y_i), \end{aligned}$$

Hence, the joint log likelihood of the observed data of all N patients, $y = (y_1, \ldots, y_N)'$, is given by:

$$l(p, \phi) = \sum_{i=1}^{N} log\big[\sum_{d=0}^{L-1} p_d g_d(y_i)\big],$$

where $p = (p_0, ..., p_{L-1})$, $\phi = (\phi_0, ..., \phi_{L-1})$, and $\phi_d = (\phi_{d10}, ..., \phi_{d1(J-1)}, ..., \phi_{dk0}, ..., \phi_{dk(J-1)})$.

3.2.3 Estimation with the EM Algorithm

With the likelihood function, the maximum likelihood approach can be used to obtain the parameter estimates for $p = (p_0, \ldots, p_{L-1})$, and $\phi = (\phi_0, \ldots, \phi_{L-1})$, subjects to the normalizing conditions $\sum_{j=0}^{J-1} \phi_{dkj} = 1$ for every $d = 0, 1, \ldots, L-1$ and $k = 1, 2, \ldots, K$. Here, employing the EM algorithm by considering D as missing data can be more computationally efficient than direct maximization. Specifically, the complete data were (y, D), and the complete-data log likelihood was as follows:

$$l_c(p, \phi) = \sum_{i=1}^{N} \sum_{d=0}^{L-1} P(D_i = d) log p_d g_d(y_i).$$

I let $p^{(t)} = (p_0^{(t)}, \ldots, p_{L-1}^{(t)})$ and $\phi^{(t)} = (\phi_0^{(t)}, \ldots, \phi_{L-1}^{(t)})$ indicate the estimates of p and ϕ after the tth iteration of the EM algorithm. The conditional expectation of complete-data log likelihood, given the observed data y and the current parameter estimates $p = p^{(t)}, \phi = \phi^{(t)}$, computed in the tth step, is given as follows:

$$E(l_c(p,\phi)\,|y, p^{(t)}, \phi^{(t)}) = \sum_{i=1}^{N}\sum_{d=0}^{L-1} P(D_i = d\,|y_i, p^{(t)}, \phi_i^{(t)}) log p_d^{(t)} g_d^{(t)}(y_i),$$

where

$$g_d^{(t)}(y_i) \;=\; \prod_{k=1}^{K}\prod_{j=0}^{J-1}[\phi_{dkj}^{(t)}]^{ikj},$$

and

$$P(D_i = d\,|y_i, p^{(t)}, \phi_i^{(t)}) \;\;=\;\; \frac{p_d^{(t)} g_d^{(t)}(y_i)}{\sum_{d=0}^{L-1} p_d^{(t)} g_d^{(t)}(y_i)}.$$

In the M-step, we the updated estimates of the parameters were found by maximizing the conditional expectation of complete-data log likelihood, computed in E-step, with respect to p and ϕ, which result in the following explicit expressions:

$$p_d^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} P(D_i = d\,|y_i, p^{(t)}, \phi_i^{(t)})$$

and

$$\phi_{dkj}^{(t+1)} = \frac{\sum_{i=1}^{N} P(D_i = d\,|y_i, p^{(t)}\phi_i^{(t)}) y_{ikj}}{\sum_{i=1}^{N} P(D_i = d\,|y_i, p^{(t)}, \phi_i^{(t)})}.$$

According to the theory about the EM algorithm, the convergent values from the above EM loop yield a local maximum of the likelihood function. Since the likelihood function of a latent class model usually has multiple local maxima, these estimates may not be the MLEs of the parameters. This was also pointed out by Zhou, Castelluccio and Zhou (2005) that, these estimates may be sensitive to the selection of the initial parameter estimates. Their recommendations were to avoid taking equal $\phi_{0kj} = \phi_{1kj} = \ldots = \phi_{(L-1)kj}$ for all k and j

as initial values, to try a set of reasonable initial parameter estimates, and to compare the local log-likelihood maxima obtained.

3.3 Summary Measure and Graphical Representation

3.3.1 True Positive Rates

Similarly, in the situation with binary disease status, each doctor's true positive rates for each disease status under a given sequence of cut points can be defined as follows:

$$\begin{aligned} TPR_k(d) &= P(j_d \leq T_k < j_{d+1} \,|D = d) \\ &= \sum_{j=j_d}^{j_{d+1}-1} \phi_{dkj}, \end{aligned}$$

They can then be used to evaluate the diagnostic accuracy of different doctors.

There are also other ways to define true positive rates when disease status is ordinal. For example, some literature defines the true positive rate as the probability that a test result is greater than or equal to the true severity level $TPR_k(d) = P(T_k \geq j_d \,|D = d)$. It essentially still considers the test results in a dichotomized setting and does not use the full ordinal information. This definition makes sense when the decision-making remains the same once the test results are above a given cut point. However, I argue here that it is important to distinguish between different severity levels, and consider the term "true positive" as "correct". I therefore defined the true positive rate as the probability of correctly identifying a subject severity level in the corresponding group. This definition also assists in the development of a straight-forward definition for an accuracy summary measure.

3.3.2 An Accuracy Summary Measure

Although the true positive rates and the diagnostic probabilities $\phi_{dkj} = P(T_k = j \,|D = d)$ obtained from the latent class model can provide information regarding the diagnostic accuracy of different tests/doctors, the results can be overwhelming and difficult to digest due to their ordinal scales. When the disease status is binary, AUC is also used as a summary measure to reflect the diagnostic ability of a test. It is of interest to define an accuracy summary measure in the ordinal situation.

I have already defined the true positive rate for each severity level as the probability of correct diagnosis of the patients in that severity group. A straight-forward way to summarize diagnostic accuracy is to consider correct diagnoses for all severity groups and take the average across all possible cut-points sequences. Therefore, I defined the overall accuracy V of the kth test as

$$\begin{aligned} V_k &= \underset{0=j_0\le j_1\le\ldots\le j_L=J}{\text{aver}} \Big\{ \prod_{d=0}^{L-1} P(j_d \le T_k < j_{d+1} \,|D=d) \Big\} \\ &= \underset{0=j_0\le j_1\le\ldots\le j_L=J}{\text{aver}} \Big\{ \prod_{d=0}^{L-1} TPR_k(d) \Big\}, \end{aligned}$$

where "aver" denotes an average operator.

Based on this definition, V is the averaged probability of making a correct diagnosis in all disease severity groups across all possible sequences of cutoff points. Recalling that AUC has an interpretation of averaged true positive rates over all false positive rates, or over all possible cutoff points, it is not hard to find that the accuracy summary measure V is an equivalent concept to AUC in a high dimensional space. In fact, the formulae can be rewritten as follows:

$$\begin{aligned} V_k &= \underset{0=j_0\le j_1\le\ldots\le j_D=J}{\text{aver}} \Big\{ \prod_{d=0}^{L-1} TPR_k(d) \Big\} \\ &= \underset{0=j_0\le j_1\le\ldots\le j_D=J}{\text{aver}} \Big\{ \prod_{d=0}^{L-1} \sum_{j=j_d}^{j_{d+1}-1} \phi_{dkj} \Big\} \\ &= \sum_{j_1=0}^{J} \Big\{ \phi_{0kj_1} \sum_{j_2=j_1}^{J} \Big[\phi_{1kj_2} \quad \ldots \sum_{j_L=j_{L-1}}^{J} \phi_{(L-1)kj_D} \Big] \Big\} \end{aligned} \tag{3.1}$$

The last expression suggests that V_k has a form of high dimensional volume. In fact, if doctors' responses are continuous, $P(T_k = t|D = d)$ are smooth functions $f_{kd}(t)$, and V_k can be written as follows:

$$\begin{aligned} V_k &= \int_{t_1=0}^{J} f_{k0}(t_1)dt_1 \int_{t_2=t_1}^{J} f_{k1}(t_2)dt_2 \quad \cdots \int_{t_L=t_{L-1}}^{J} f_{k(L-1)}(t_L)dt_L \\ &= \int\int\cdots\int f_k(t_1,\ldots,t_L)dt_1\cdots dt_L, \end{aligned}$$

where f_k is the function of a smooth hyper-surface of the diagnostic results in L-dimensional space.

If I depict the points $(TPR_k(0), ..., TPR_k(L-1))$ in L-dimensional coordinates, where the projection on the ith axis is $TPR_k(i-1)$, then the volume under the terrace-shaped hyper-surface generated by these points can be computed exactly as (3.1). The proof for the equivalence of the accuracy summary measure V and the AUC in the case of binary disease status can be found in Appendix A.

Therefore, the overall accuracy measure V is a high dimensional counterpart of AUC, or a generalized AUC. Recall that, in the case of binary disease status, AUC can be interpreted as the probability of correctly ranking the test results, one being randomly chosen from the diseased population and the other being randomly chosen from the non-diseased population. Mossman (1999) extended this concept to the three-way ROC surface, where the volume under the ROC surface (VUS) was equal to the probability that a triplet randomly chosen from three different groups will be correctly ranked by the decision-maker. In the current case, by randomly selecting L subjects with one from each severity group, V can be interpreted as the probability of correctly ranking their test results according to severity levels.

In the special case of binary disease status where the diagnosis probabilities $P(T_k = j|S = d)$ are discrete step functions, the accuracy summary measure is reduced to the following expression:

$$V_k = \sum_{j=0}^{J} \Big[\phi_{0kj} \sum_{l=j}^{J} \phi_{1kl} \Big],$$

which is equal to the AUC when terrace-shaped lines are used to connect the points. If inclined lines, rather than terrace-shaped lines, are used to connect the points, meaning that $P(T_k = t|S = d)$ are linear functions, then the accuracy summary measure is:

$$\begin{aligned} V_k &= \sum_{j=0}^{J-1} \Big[\phi_{0kj} \sum_{l=j+1}^{J} \phi_{1kl} \Big] + \frac{1}{2} \sum_{j=0}^{J} \phi_{0kj} \phi_{1kj} \\ &= \sum_{j=0}^{J} \Big[\phi_{0kj} (\frac{1}{2} \phi_{1kj} + \sum_{l=j+1}^{J} \phi_{1kl}) \Big], \end{aligned}$$

which is equal to the AUC under the trapezoidal rule.

In addition, the result for a three-class diagnostic task with inclined planes connecting

the points is given by

$$\begin{aligned} V_k &= \sum_{j_1=0}^{J}\Big[\phi_{0kj}\sum_{j_2=j_1}^{J}\phi_{lkj}\Big(\frac{1}{2}\phi_{2kj_2}+\frac{2}{3}\times\frac{1}{2}\phi_{2kj_2}+\frac{1}{3}\times\frac{1}{2}\phi_{2kj_2+1}+\sum_{l=j_2+2}^{J}\phi_{2kj}\Big)\Big] \\ &= \sum_{j_1=0}^{J}\Big[\phi_{0kj}\sum_{j_2=j_1}^{J}\phi_{1kj}\Big(\frac{5}{6}\phi_{2kj_2}+\frac{1}{6}\phi_{2kj_2}+\sum_{j=j_2+2}^{J}\phi_{2kj}\Big)\Big]. \end{aligned}$$

It equals the VUS of the three-way ROC surface.

Also, when evaluating the diagnostic accuracy of different doctors, it should be noted that the accuracy indicator defined above is $1/D!$ for a random guess, rather than $1/2$ as in dichotomous cases.

3.3.3 Graphical Representation

Although it can be shown that the accuracy summary measure proposed in the last section is equal to the volume under the L-dimensional ROC surface, this high-dimensional ROC surface is difficult to depict graphically. Furthermore, beyond the overall accuracy measures, the specific probabilities $P(T_{ik}=j|D_i=d)$ are also of interest. I use two forest graphs to represent these quantities:

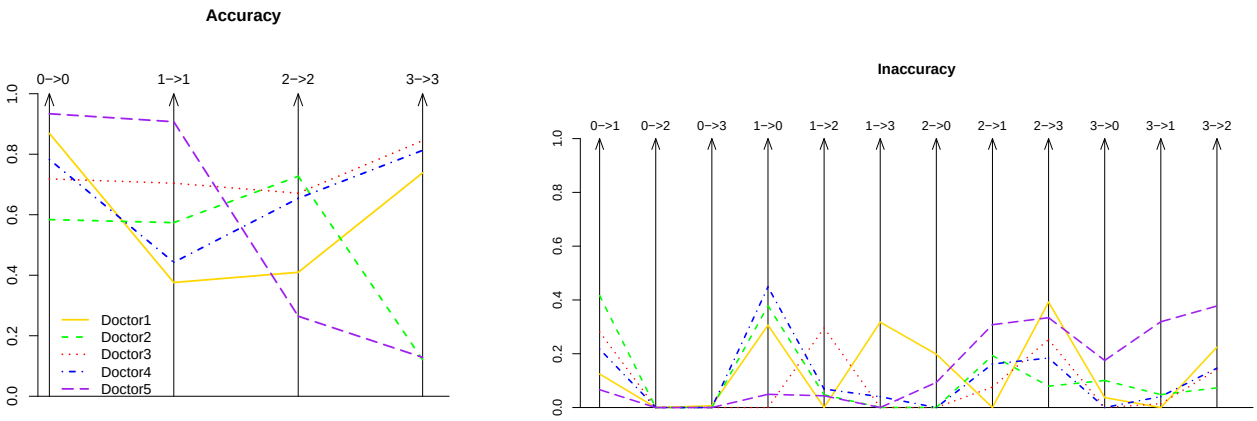

Figure 3.2: Forest plots representing tests/doctors' diagnostic abilities.

The graph of accuracy represents the probabilities of correct diagnosis. It contains L axes, each representing one disease severity status, and the scale on each of the axes denotes the probability of correctly classifying a patient according to his/her true disease status. An example, when K=5 and L=J=3, is shown in Figure 3.2. The first bar from the left, labeled 0 to 0, indicates diagnosed subjects in the first category, no disease by no disease, followed by diagnosed subjects with a mild condition by a mild condition and so on. The graph of inaccuracy reflects the probabilities of drawing incorrect conclusions with different types of errors. Specifically, it contains $L \times (L-1)$ axes labeled as $d \to j$. A point on the axis $d \to j$ indicates the probability that the conclusion based on the test result T is j when the true disease status D is d, i.e. the value of $\phi_{dkj} = P(T_k = j \,|\, D = d\,)$, while different colors and line types are used to identify different tests/doctors. Consequently, a better test/doctor tends to occupy a larger area in the graph of accuracy and a smaller area in the graph of inaccuracy. These graphs allow one to visualize the diagnostic abilities of different tests/doctors.

The accuracy and inaccuracy plots show the probability of each type of correct or incorrect diagnosis. However, sometimes for misdiagnosis, one may want to combine some categories and look at the over-diagnosis and under-diagnosis. I borrowed the idea of the classification profile (CP) plot used in process control for this purpose. An example with $L = J = 3$ is shown in Figure 3.3. The vertical axis is still on the probability scale, and variable k on the horizontal axis is the difference between the diagnostic result and the true disease severity level, i.e., $k = t - d$, so a positive k means over-diagnosis and a negative k means under-diagnosis, and the middle bar with $k = 0$ is the probability of accurate diagnosis. The shape of the classification profile graphically reflects overall diagnostic performance. A CP plot with a relatively high and narrow peak at $k = 0$ corresponds to a test/doctor with high diagnostic accuracy and a relatively flat plot corresponds to a test/doctor with low diagnostic accuracy. In addition, a CP plot skewed towards the left indicates a test/doctor that tends to under-diagnose, while a CP plot skewed towards the right indicates a test/doctor that tends to over-diagnose.

The bar labels in the CP plot in Figure 3.3 show how to compute the height of each bar based on a contingency table of test results and the true disease status as table 3.1. It

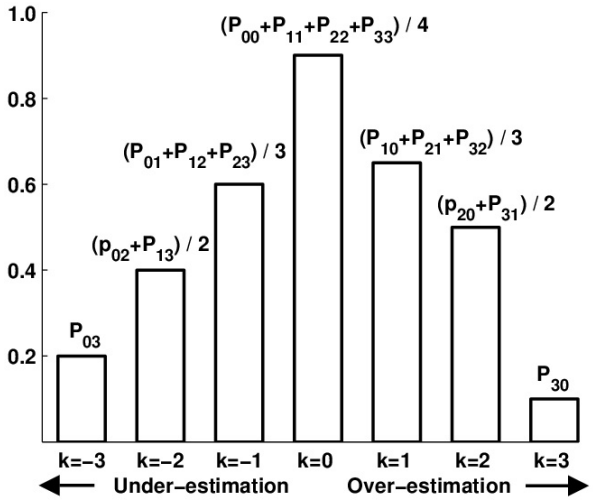

Figure 3.3: Sketch map of a CP plot.

suggests that the CP plot is essentially pressing the contingency table along the diagonal lines. The mean probabilities on the main diagonal line are the height of the middle bar with $k = 0$. Cell probabilities P_{10}, P_{21} and P_{32} above the main diagonal line contribute to the bar with $k = 1$, and so on.

Table 3.1: Contingency table of test results and the true disease status.

	True disease status			
Test results	$d = 0$	$d = 1$	$d = 2$	$d = 3$
$t = 0$	P_{00}	P_{01}	P_{02}	P_{03}
$t = 1$	P_{10}	P_{11}	P_{12}	P_{13}
$t = 2$	P_{20}	P_{21}	P_{22}	P_{23}
$t = 3$	P_{30}	P_{31}	P_{32}	P_{33}

3.4 Simulation Studies

In this section I report the results of some simulation studies carried out under various settings, including different levels of the diagnostic performance of the tests and different prevalence rates of the severity groups, to assess the performance of the nonparametric method and the overall diagnostic accuracy described in the previous section. The results presented here use K=5 and L=J=3 as in the real data example.

3.4.1 Small Sample Property

In practice, it is usually difficult to collect data where one subject is diagnosed by many tests or doctors. Therefore, the total number of patients N in multiple test data may not be very large. I evaluated the small sample performance of the proposed model in this section. In my simulation, I chose as the numbers of patients 20 and 40. I considered three different prevalence rates for each sample size. The first scenario assumes equal portions for patients with different disease severities. The second one assumes that are more patients with higher severity status, and the last one assumes a U-shaped prevalence rate, where most patients are either in the asymptomatic group or in the most severe groups and fewer patients are in the two groups with mild or moderate disease.

I simulated 5,000 data sets under each setting, and then implemented the proposed method to obtain the parameter estimates. As for the choice of initial values, I avoided having $\phi_{0kj} = \phi_{1kj} = \cdots = \phi_{Lkj}$ for all k and j as mentioned previously. In addition, to avoid reaching a local maximum, I used a number of initial values, and then compared the local log-likelihood maxima obtained for each of the chosen initial values to obtain the final ML estimates. With ordinal disease status and test results, there are many parameters for diagnostic probabilities. Ive only presented the mean, minimum and maximum of the bias and mean squared errors (MSE) of the diagnostic probabilities and focused more on the accuracy summary measure. The results are shown in Table 3.2.

The results suggest that the proposed method yields ML estimates for the overall diagnostic accuracy, with small biases and MSEs, regardless of the true diagnostic performances of doctors and true prevalence rates of the disease. The summary result of all ϕ_{djk} also

supports this method. In general, and as expected, the bias is smaller when disease groups are relatively balanced in size, but in general the bias is acceptable for even the small sample size of 20. The mean bias is about 1% for a sample size of 40.

3.4.2 Violation of the Conditional Independence Assumption

The conditional independence assumption is crucial in the proposed method. Although it is a reasonable assumption in the real data problem, it is helpful to evaluate the robustness of the method in situations where this assumption is violated. Suppose that, except for the true symptom status, there are some doctor-level covariates R that would influence the doctors' decisions. As a result, instead of conditional independence giving the true symptom status, the diagnosis results were correlated, unless providing both true disease status D and covariates R. In these simulations, I supposed that doctors' decisions were influenced by the covariate R via the coefficient b through a probit model, where R follows a normal distribution with mean 0 and standard deviation 0.5. Specifically, $P(T_{ik} = j|D_i = d, R_k = r) = \Phi(a_{dkj} + br)$, for $i = 1, ..., N$, $d = 0, 1, ..., L-1$ and $k = 1, 2, ..., K$. I chose a sample size of $N = 34$, as in the real data set, and simulated 5,000 data sets under the three prevalence scenarios as before, i.e., equal portions with $p_0 = 0.25, p_1 = 0.25, p_3 = 0.25, p_4 = 0.25$; increasing prevalence $p_0 = 0.1, p_1 = 0.2, p_3 = 0.3, p_4 = 0.4$; and U-shaped prevalence $p_0 = 0.25, p_1 = 0.1, p_3 = 0.15, p_4 = 0.5$. In each setting, I considered three different correlation levels by setting b equal to 0.2, 0.5 and 1. The results for bias in the estimates are shown in Table 3.3, with MSEs in parenthesis.

The results show that, with correlations induced by covariates, both the biases and MSEs are larger than when the conditional independence assumption holds. However, when the influence of the covariates on the diagnosis results was not very large (i.e. when the dependence was mild) the results were still acceptable. Especially, the impact of correlation on the accuracy summary measure V_k is not as big as that on the diagnostic probabilities ϕ_{dkj}. However, as expected, when the covariates had bigger influences the biases and MSEs increased, suggesting that the conditional assumption is necessary for the proposed method, and that it is not suitable in the situation where diagnostic results are highly correlated

even within the groups with the same disease severity level.

3.4.3 Boundary Effects

I performed additional simulations to assess the model performance on the boundary of the parameter space (i.e. when some of the diagnosis probabilities ϕ_{dkj} were zero). This question was first raised by the results of the real data analysis, where some of these quantities were estimated as zero. The parameter set-up in the simulation used the estimated values of these parameters from the real data study. The results are shown in Appendix B; they do not suggest that the proposed method performs any differently in these settings than in the earlier settings.

3.5 Real Data Analysis

In this section I apply the latent class model proposed in previous sections to the motivating example of TCM. The goal is to assess 5 TCM doctors' diagnostic abilities on symptom level, since this is the basic level of diagnosis. In this example, the gold standard is the true symptom level, and the diagnostic tests are doctors' diagnoses.

3.5.1 The TCM Data

The data used here are for the syndrome differentiation results for Chills disease collected in 2005 in the Xiaoyudong Village of Sichuan Province of China. Chills disease is a condition studied in TCM. A similar disease, called Raynauds phenomenon, (also referred to as Raynaud's disease or sometimes Raynaud's syndrome) is recognized by western medicine. This condition is caused by irregular dietary and life-style habits and hormonal imbalance; it can lead to abnormal vascular contraction and expansion, decreased blood circulation function and peripheral nerve excretion not fully discharged. Patients who suffer from Chills disease often have the following symptoms: cold hands and feet, headaches or shoulder stiffness, fatigue, irritability, insomnia, and so on. For this condition, TCM doctors usually have good agreement on disease level, but much variability on symptom level. This provides a good opportunity to examine TCM symptom diagnosis.

In this data set, 34 patients were diagnosed by 5 doctors separately. The data set contained the diagnostic results of the five doctors on twelve symptoms, including wind-phobia, cold extremities, cold body, sniffling, abdominal pain, warmphile, crouchinness, a preference for knead, inappetence, a preference for mild temperature, cold-sensitiveness, and cold-phobia, which are all related to the Chills disease. For convenience, I denoted the symptoms as symptom 1 through symptom 12, respectively. The diagnostic results are given on an ordinal scale, 0, 1, 2, and 3, representing the doctors' conclusions regarding the symptom severity of the patients: asymptomatic, slight, middle or severe.

The conditional independence assumption is necessary for my method (i.e., given the true symptom statuses, doctors' diagnoses on the symptoms should be independent of each other).

This assumption does not usually hold when considering a disease, since doctors may all have learned the same theory and been trained to look for the same symptoms. In the present case this assumption is reasonable, because what I have considered here are basic symptoms without further surrogates, and I have assumed that doctors do not communicate during diagnosis. The diagnoses of the doctors reflect their own opinions on whether a particular symptom is mild or severe, based on their own judgments of cut-points rather than on any common principles or common variables. It was unlikely that there were any common covariates that influenced their decisions. Taking symptom 1, wind-phobia, as an example, each doctor gave his/her own opinion on how sensitive a patient is to the wind. The doctors do not know the diagnosis results of others, and their conclusions are reached according to their own perceptions about the patients' symptoms. Thus, when the true symptom status is given, their diagnoses are likely to be independent.

3.5.2 Analysis Results

I applied the proposed latent class modeling approach to these data. Again, in the implementation of the EM algorithm, I used a number of initial values and compared the local maxima to obtain the MLEs. In addition, I used the initial values given by experts of traditional Chinese medicine and achieved the same results. The results of symptom 1

are shown as an example in Table 3.4. Each column represents a different doctor. The diagnostic probabilities are denoted as $\phi_{dkj} = P(T_k = j \,|S = d)$, with S indicating the true symptom level. The last row in table 3.4 is the accuracy summary measure for each doctor. The results, with standard error estimates for all twelve symptoms, are shown in Appendix C.

In these data, with 4 symptom severity levels, the accuracy summary measure is equal to 1 over 24, which is about .05 for random diagnoses. The results here suggest good performance of the doctors, especially for Doctor 3, followed by Doctor 4, among these five TCM doctors.

In order to give a more clear perception of the doctors' diagnostic accuracies, I drew two forest plots based on Table 3.4 to show the accuracy and inaccuracy as discussed in Section 3.3.3. The plots are displayed in Figure 3.1. It is evident that Doctor 3, indicated by the red line, has good and relatively stable performance among all symptom levels and Doctor 5, indicated by the purple line, has the lowest accuracy summary measure among the doctors, despite having especially good performance among subjects with mild symptoms. However, based on his performance for subjects with severe symptoms, it is clear that he tends to under-diagnose. Doctor 1, indicated by the yellow line, also has good performance on the two extreme cases–subjects with no symptoms or subjects with severe symptoms.

I also drew a CP plot in Figure 3.3. From the graphs it is evident that, in general, the diagnostic ability of Doctor 3 on symptom 1 was better than that of the other doctors, as the most weight concentrates on the middle bar, although he has a slight tendency to over-diagnose. On the other hand, the incorrect diagnoses of doctor 1 and doctor 4 are relatively balanced, but doctor 4 is better since his diagnoses are closer to the true symptom levels.

3.6 Summary

In this Chapter I propose a nonparametric maximum likelihood method in a latent class modeling framework to estimate the accuracy of ordinal-scale diagnostic tests when the gold standard on the true severity of the medical condition has an ordinal scale and is not available. The maximization is facilitated by the EM algorithm. The explicit solution in the M step makes the implementation easy and fast. I also examined the performance of

the proposed method in various settings in simulation studies.

I proposed an accuracy summary measure, which is shown to be a high dimensional counterpart of AUC for ordinal data. I used forest plots and CP plots to visually present the diagnostic ability of each doctor in distinguishing the severity levels of a symptom. The graphs are easier to digest compared with the series of estimators, and can provide more information about diagnostic performance than the single accuracy summary measure V_k. Specifically, the inclination of underestimation or overestimation of the diagnostic test on the disease severity level can be assessed.

The proposed method requires that the conditional independence assumption holds. Although this assumption is reasonable in the example used here, there are many other situations in which this assumption is questionable, and the conditional dependence structure among different doctors should be considered. One area of further research strives to develop a method without assuming conditional independence. For example, a log-linear model without higher-order interactions may be used to relax this assumption. Or, one might consider utilizing a random-effect modeling approach, which has become very popular for correlated tests, and is discussed by Hadgu and Qu (1998), Qu and Hadgu (1998) in the context of binary disease status. This approach is discussed in the next chapter.

Another extension is to situations where a symptom or disease status has a nominal scale. In the present case, symptom status has an ordinal scale ranging from 0 to 3 that denotes the severity of symptoms from asymptomatic to severe. However, there are some situations in which discrete values are used to indicate different sub-diseases. For example, in the diagnosis of cancer, a doctor may be interested in determining whether it is a lung cancer, colon cancer, or another type. My definition of TPR and the overall accuracy indicator did not include information about severity. Rather, the TPR only includes the situations where the diagnosis results agree with the true symptom status after excluding both under-diagnosed cases and over-diagnosed cases. This essentially gives equal weight to all types of errors, although one might consider giving different weights when some errors are known to be more harmful. Using my definitions, the approach could easily be extended to diseases with a nominal scale, once the diagnosis strategy for the nominal status is determined. However, without ordering, this kind of diagnosis problem is harder to handle. Nakas and

Alonzo (2007) discussed the case when the disease classes have an umbrella ordering. They essentially viewed this problem as a combination of the two possible sequences. A similar method of considering a nominal scale as a combination of all possible ordering may offer one approach to addressing the diagnosis problem on a nominal scale.

Table 3.2: Results from 5,000 simulations with L=3, K=5, J=3 under various prevalence rates and parameter settings.

Patient number N=20

True prevalence	Statistics	V_1	V_2	V_3	V_4	V_5	ϕ^*_{dkj}		
rates	True values	0.356	0.408	0.505	0.606	0.703	max	min	mean
p_0=0.25,p_1=0.25	Bias	-0.014	-0.014	-0.024	-0.026	-0.035	0.046	0.000	0.017
p_2=0.25,p_3=0.25	MSE	0.045	0.050	0.057	0.069	0.059	0.123	0.003	0.046
p_0=0.10,p_1=0.20	Bias	-0.016	-0.018	-0.024	-0.019	-0.030	0.098	0.000	0.024
p_2=0.30,p_3=0.40	MSE	0.050	0.062	0.064	0.060	0.078	0.168	0.006	0.051
p_0=0.25,p_1=0.10	Bias	-0.031	-0.021	-0.015	-0.023	-0.058	0.357	0.001	0.042
p_2=0.15,p_3=0.50	MSE	0.050	0.058	0.064	0.065	0.087	0.199	0.003	0.066

Patient number N=40

True prevalence	Statistics	V_1	V_2	V_3	V_4	V_5	ϕ_{dkj}		
rates	True values	0.356	0.408	0.505	0.606	0.703	max	min	mean
p_0=0.25,p_1=0.25	Bias	-0.010	-0.008	-0.008	-0.003	-0.030	0.033	0.000	0.008
p_2=0.25,p_3=0.25	MSE	0.019	0.022	0.024	0.026	0.029	0.045	0.004	0.023
p_0=0.10,p_1=0.20	Bias	-0.014	-0.013	-0.013	-0.010	-0.039	0.042	0.000	0.011
p_2=0.30,p_3=0.40	MSE	0.024	0.028	0.031	0.030	0.039	0.077	0.003	0.028
p_0=0.25,p_1=0.10	Bias	-0.015	-0.014	-0.015	-0.016	-0.044	0.045	0.001	0.012
p_2=0.15,p_3=0.50	MSE	0.021	0.024	0.028	0.028	0.036	0.073	0.003	0.028

* Results for all ϕ_{dkj} with $d = 0, 1, ..., 3$, $k = 1, 2, ..., 5$, and $j = 0, 1, ..., 3$.

Table 3.3: Robustness performance from 5,000 simulations when L=3, K=5, J=3 and patient number N=34. Covariate R follows a normal distribution with mean 0 and standard deviation 0.5, and influent diagnosis with coefficient b.

True prevalence	Statistics	V_1	V_2	V_3	V_4	V_5	ϕ_{dkj}		
rates	True values	0.356	0.408	0.505	0.606	0.703	max	min	mean
		-0.008*	-0.012	-0.019	-0.032	-0.023	0.067	0.000	0.017
	b=0.2	(0.020)**	(0.020)	(0.023)	(0.026)	(0.027)	(0.040)	(0.003)	(0.021)
p_0=0.25,p_1=0.25		-0.009	-0.003	-0.003	-0.018	-0.018	0.129	0.000	0.030
p_2=0.25,p_3=0.25	b=0.5	(0.015)	(0.018)	(0.019)	(0.023)	(0.020)	(0.035)	(0.002)	(0.020)
		-0.030	-0.010	-0.039	-0.042	-0.031	0.271	0.003	0.054
	b=1.0	(0.018)	(0.020)	(0.024)	(0.026)	(0.027)	(0.048)	(0.001)	(0.022)
		-0.015	-0.023	-0.034	-0.047	-0.053	0.132	0.000	0.025
	b=0.2	(0.026)	(0.034)	(0.037)	(0.039)	(0.052)	(0.106)	(0.002)	(0.030)
p_0=0.10,p_1=0.20		-0.012	-0.016	-0.013	-0.027	-0.011	0.146	0.000	0.028
p_2=0.30,p_3=0.40	b=0.5	(0.016)	(0.021)	(0.026)	(0.025)	(0.025)	(0.062)	(0.001)	(0.023)
		-0.023	-0.020	-0.036	-0.038	-0.021	0.293	0.002	0.053
	b=1.0	(0.015)	(0.025)	(0.032)	(0.023)	(0.026)	(0.068)	(0.001)	(0.024)
		-0.018	-0.018	-0.009	-0.018	-0.028	0.084	0.000	0.019
	b=0.2	(0.018)	(0.021)	(0.021)	(0.022)	(0.027)	(0.066)	(0.002)	(0.024)
p_0=0.25,p_1=0.10		-0.019	-0.008	-0.019	-0.02	-0.03	0.151	0.000	0.036
p_2=0.15,p_3=0.50	b=0.5	(0.017)	(0.018)	(0.022)	(0.021)	(0.026)	(0.074)	(0.001)	(0.026)
		-0.038	-0.016	-0.026	-0.028	-0.051	0.236	0.000	0.062
	b=1.0	(0.016)	(0.022)	(0.027)	(0.026)	(0.028)	(0.096)	(0.001)	(0.029)

* Bias ** MSE

Table 3.4: Parameter estimates and corresponding overall accuracy estimates of the doctors in detecting symptom 1.

	Doctor1	**Doctor2**	**Doctor3**	**Doctor4**	**Doctor5**
P(T=0|S=0)	0.869	0.584	0.718	0.783	0.934
P(T=1|S=0)	0.125	0.416	0.282	0.217	0.066
P(T=2|S=0)	0.000	0.000	0.000	0.000	0.000
P(T=3|S=0)	0.006	0.000	0.000	0.000	0.000
P(T=0|S=1)	0.306	0.378	0.000	0.448	0.049
P(T=1|S=1)	0.376	0.574	0.704	0.443	0.907
P(T=2|S=1)	0.000	0.047	0.296	0.069	0.044
P(T=3|S=1)	0.317	0.000	0.000	0.040	0.000
P(T=0|S=2)	0.198	0.000	0.000	0.000	0.093
P(T=1|S=2)	0.000	0.193	0.076	0.162	0.308
P(T=2|S=2)	0.410	0.727	0.671	0.654	0.265
P(T=3|S=2)	0.392	0.079	0.253	0.184	0.334
P(T=0|S=3)	0.037	0.101	0.000	0.000	0.175
P(T=1|S=3)	0.000	0.048	0.014	0.041	0.319
P(T=2|S=3)	0.225	0.731	0.141	0.146	0.378
P(T=3|S=3)	0.739	0.121	0.845	0.813	0.128
V_k	0.550	0.650	0.906	0.786	0.389

Chapter 4

DEPENDENT ORDINAL DIAGNOSTIC TESTS WITH UNKNOWN ORDINAL GOLD STANDARD

A key assumption made by the latent class model discussed in the previous section is the conditional independence assumption. Although this assumption is reasonable in the TCM example, it can be violated in many practical situations. For example, test results can be correlated if they are based on a similar physiological phenomenon, such as a particular antibody reaction. Any disturbance or contamination may have similar effects on several tests. In addition, tests may be designed to capture the same aspect, for example, when using questionnaires some of the items may overlap among different questionnaires.

In this chapter I incorporate a random effect model into the latent class model to describe the correlation among test results within each disease group. The method is illustrated on the same TCM data used in the previous chapter. In addition, I discuss a likelihood ratio test for the conditional independence assumption and an *ad hoc* test for model fit.

4.1 Literature Review

Many researchers have discussed various models to describe the possible dependence structures among diagnostic tests, including the parametric model, the log linear model, the direct effect model and the random effect model.

In the latent class framework, test results are modeled within each group. Instead of having several independent univariate distributions, a perhaps straight forward way to include correlations is to use a multivariate parametric distribution as the component distribution, such as the multivariate normal distribution. The parametric assumption largely reduces the number of parameters needed for the correlation and makes the estimation feasible. To apply this idea when using discrete test results, Henkelman et al. (1990) proposed that discrete diagnoses were actually a categorized version of an underlying trait that follows a multivariate normal distribution within the diseased group, and also within the healthy

group. One disadvantage of this method is that it relies on a parametric assumption. Additional limitations of this model were discussed by Begg and Metz (1990).

Without this parametric assumption, Espeland and Handelman (1989) used a latent log linear model to explain the dependence among the tests within each disease class. Essentially, they consider a contingency table for all test results within each disease class, and model the cell probability with a log linear model. However, including too many interaction terms can cause an identifiability problem. Another drawback of this model is that, with the additional interactions among the tests, the main effect and other parameters in the model no longer have straight forward interpretations in terms of accuracy measures of the tests.

One way to deal with the second limitation of the log linear model is to introduce sensitivity and specificity as model parameters, and directly model the correlations among them. Vacek (1985) and Hagenaars (1988) discussed this method with two binary correlated tests. This approach was later referred to by Dendukuri and Joseph (2001) as direct effect models. Similarly to the log linear model, these models can run into identifiability problems when too many interaction terms are assumed. In such a situation researches usually introduce additional constraints or use a Bayesian approach and impose informative prior distributions.

Somewhat combining the parametric approach and the log linear idea is the random effect method. Qu et al. (1996) use a random effect term to summarize correlation among the tests within the disease groups. The test results are then modeled conditionally on both the disease status and the random effect. The drawback of using this model is that the integration over the random effect makes it computationally intensive. Qu et al. discussed its application for binary disease status. In this chapter I extend the method to ordinal disease status and discuss the computational issues that arise from taking this approach.

4.2 A Random Effect Model

I continue using the latent class modeling framework. In the previous chapter I described heterogeneity in the test results by the true disease status D. The observed test results are

modeled as a mixture from all disease groups, as follows:

$$P(\vec{T}) = \sum_{d=0}^{L-1} P(D = d)P(\vec{T}|D = d)$$

In this basic framework, the conditional independence assumption is not required. This assumption is only used when modeling the joint test distribution within a disease group $P(\vec{T}|D = d)$. Instead of assuming conditional independence and modeling this distribution as a product of the univariate distributions, other models can be used to describe the joint distribution. Here, I consider the additional correlation among tests within a disease group to be caused by a subject level random effect R, such as some unobserved patient characteristic. The model for this case is shown in Figure 4.1.

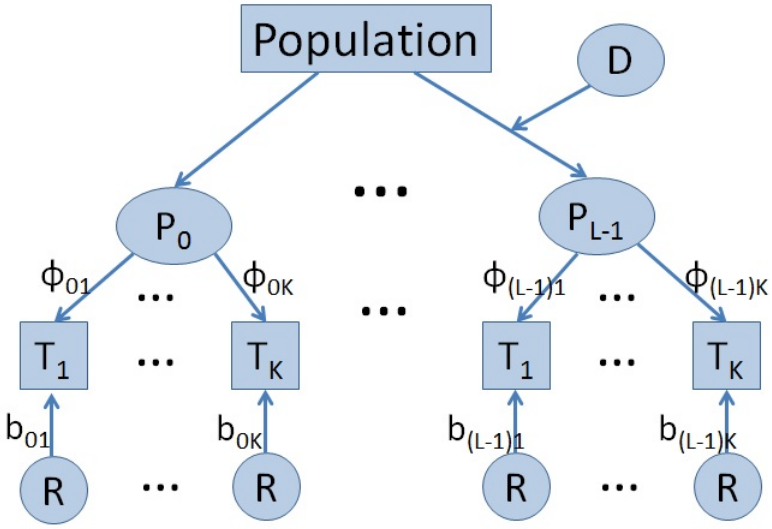

Figure 4.1: A latent class model with random effect

4.2.1 The Model

Now the test results depend on both disease status D_i and the latent variable R_i. They can be modeled independently after conditioning on both D_i and R_i.

In this model I assume that the subject level random effect follows a standard normal distribution $N(0,1)$, but that other distributions can be adopted as well. I parameterize $P(T_k|D)$ with a cumulative probit random effects model. Specifically, the conditional distribution of test results can be written as follows:

$$
\begin{aligned}
& P(T_{ik} \leq j-1 | D_i = d, R_i = r) = \Phi(a_{kdj} + b_{kdj} r), \\
\text{where} \quad & R_i \sim N(0,1), \quad b_{kdj} \geq 0 \\
& a_{kd0} \leq a_{kd1} \leq \ldots \leq a_{kd(J-2)} \quad \text{for each given } k = 1, 2 \ldots, K \\
& d = 0, 1, ..., L-1, \ k = 1, 2, ..., K, \text{ and } j = 1, 2, ..., J-1.
\end{aligned}
$$

Here, Φ is the cumulative distribution function of a standard normal variable. The two unobserved random variables, D_i and R_i, are assumed to be independent of each other. Condition $a_{kd1} \leq a_{kd2} \leq \ldots \leq a_{kd(J-1)}$ guarantees that the probability that the test takes any values in its domain is non-negative. In addition, I define $a_{kd0} = -\infty$ and $a_{kdJ} = +\infty$. The condition $b_{kdj} \geq 0$ is required for model identifiability, but in this way b_{kdj} also characterizes the variance of the subject-specific random variable R_i within the kth test when the true disease status is d. Other notation is the same as in the previous chapter.

Note that the choice of the probit link is not essential here, since the test results and true disease status are discrete. With other link functions, one can still model the same set of diagnosis probabilities but with a different set of cut points a_{kdj}, b_{kdj}, $d = 0, 1, ..., L-1$, $k = 1, 2, ..., K$, $j = 1, 2, ..., J-1$. However, the probit link can simplify the computation, as it makes the estimated probabilities within the proper range and sum up to 1 without imposing additional constraints. Computationally, parameters a_{kdj} that range from $-\infty$ to $+\infty$ and b_{kdj} ranging from 0 to $+\infty$ can be more stably estimated with this model than by directly estimating the corresponding probabilities within the narrow domain of 0 and 1. In addition, the cumulative probit random effects models have a natural interpretation corresponding to the thresholds of a continuous underlying latent variable, which is modeled by a linear mixed model. Thus, the cumulative probit link function has been commonly used for regression modeling of univariate and multivariate ordinal data. Further discussion can be found in McCullagh and Nelder (1989).

I have modeled $P(T_{ik} \leq j|D_i = d, R_i = r)$. Then, by integrating over the subject-specific random effects, it follows that:

$$P(T_{ik} \leq j|D_i = d) = \int_{-\infty}^{+\infty} \Phi(a_{kdj} + b_{kdj}r)d\Phi(r) = \Phi\big(\frac{a_{kdj}}{\sqrt{1 + b_{kdj}^2}}\big). \tag{4.1}$$

Also, the following equality holds:

$$\begin{aligned} & P(y_{ijk} = 1|D_i = d, R_i = r) = P(T_{ik} = j|D_i = d, R_i = r) \\ = \;& P(T_{ik} \leq j|D_i = d, R_i = r) - P(T_{ik} \leq j-1|D_i = d, R_i = r) \\ = \;& \Phi(a_{kdj} + b_{kdj}r) - \Phi(a_{kdj-1} + b_{kdj-1}r). \end{aligned}$$

Since the model assumes that the test results are independent conditionally on both the disease status D_i and the latent variable R_i, it follows that:

$$\begin{aligned} P(\mathbf{y_i}|D_i = d, R_i = r) &= \prod_{k=1}^{K}\prod_{j=0}^{J-1}[P(y_{ijk}|D_i = d, R_i = r)]^{y_{ikj}} \\ &= \prod_{k=1}^{K}\prod_{j=0}^{J-1}[\Phi(a_{kdj} + b_{kdj}r) - \Phi(a_{kdj-1} + b_{kdj-1}r)]^{y_{ikj}}. \end{aligned}$$

By integrating over the subject specific random effects again, I obtained the following expression for $P(\mathbf{y_i}|D_i = d)$:

$$\begin{aligned} & g_d(\mathbf{y_i}) = P(\mathbf{y_i}|D_i = d) \\ = \;& \int_{-\infty}^{+\infty} \prod_{k=1}^{K}\prod_{j=0}^{J-1}[P(y_{ijk}|D_i = d, R_i = r)]^{y_{ikj}} d\Phi(r) \\ = \;& \int_{-\infty}^{+\infty} \prod_{k=1}^{K}\prod_{j=0}^{J-1}[\Phi(a_{kdj} + b_{kdj}r) - \Phi(a_{kdj-1} + b_{kdj-1}r)]^{y_{ikj}} d\Phi(r). \end{aligned} \tag{4.2}$$

The integration in equation (3.1) can be evaluated numerically. I used Gaussian quadrature (Abramowitz and Stegun, 1972). This method replaces the integration by a summation over a finite number of mass points $(r_1, \ldots, r_M)$. I chose $M = 50$ Gaussian quadrature points in the estimation, but in the examples I tried the approximation was sufficient and stabilized with less than 10 quadrature points. The integration in (3.1) then becomes the following

expression:

$$\begin{aligned} g_d(\mathbf{y_i}) &= \sum_{m=1}^{M} \omega_m \prod_{k=1}^{K} \prod_{j=0}^{J-1} [\Phi(a_{kdj} + b_{kdj} r_m) - \Phi(a_{kdj-1} + b_{kdj-1} r_m)]^{y_{ikj}} \\ &= \sum_{m=1}^{M} \omega_m g_{dm}(\mathbf{y_i}), \end{aligned}$$

where $g_{dm}(\mathbf{y_i}) = P(\mathbf{y_i}|D_i = d, R_i = r_m)$, and ω_m is the mass of point r_m $(m = 1, \ldots, M)$, computed by the Gaussian quadrature method.

The observed data for the ith patient is $\mathbf{y_i}$. Its probability, or contribution to the likelihood, can be written as $P(\mathbf{y_i})$, which is equal to $\sum_{d=0}^{D-1} p_d g_d(\mathbf{y_i})$. Hence, the joint log likelihood of the observed data for all N patients is given by:

$$l(\theta) = \sum_{i=1}^{N} \log \big[\sum_{d=0}^{L-1} p_d g_d(\mathbf{y_i}) \big],$$

where $p_d = P(D_i = d)$, and θ contains all parameters p_d, a_{kdj}, and b_{kdj}.

4.2.2 Estimation

To find the MLE for θ, subject to the monotonicity constraint that $a_{kd0} \leq a_{kd1} \leq \ldots \leq a_{kdJ-2}$ for $k = 1, 2, ..., K$ and $d = 0, 1, ..., L-1$, I again employed the EM algorithm to the complete data $(\mathbf{y_i}, D_i), i = 1, ...N$.

The expected value of the complete data log likelihood is given as follows:

$$E\big[l_c(\theta)\big] = \sum_{i=1}^{N} \sum_{d=0}^{L-1} \sum_{m=1}^{M} h_{idm} \log \big[p_d \omega_m g_{dm}(\mathbf{y_i}) \big],$$

where $h_{idm} = P(D_i = d, R_i = r_m)$.

To maximize the expected value of the complete data log likelihood with respect to θ, I took the first order derivative of $E\big[l_c(\theta)\big]$ with respect to θ and solved the following equations:

$$\sum_{i=1}^{N} \sum_{d=0}^{L-1} \sum_{m=1}^{M} h_{idm} \frac{\partial \log p_d \omega_m}{\partial \theta} + \sum_{i=1}^{N} \sum_{d=0}^{L-1} \sum_{m=1}^{M} h_{idm} \frac{\partial g_{dm}(\mathbf{y_i})}{\partial \theta} = 0. \tag{4.3}$$

Notice that the first term in (4.2) is a function of p_d only, and that the second term is a function of β only, which contains all parameters a_{kdj} and b_{kdj}. Consequently, the equation

can be broken into two parts:

$$\sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{m=1}^{M} h_{idm}\frac{\partial \log p_d\omega_m}{\partial p_d} + \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{m=1}^{M} h_{idm}\frac{\partial g_{dm}(\mathbf{y_i})}{\partial p_d} = \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{m=1}^{M} h_{idm}\frac{\partial \log p_d}{\partial p_d} = 0$$

and $$\sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{m=1}^{M} h_{idm}\frac{\partial \log p_d\omega_m}{\partial \beta} + \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{m=1}^{M} h_{idm}\frac{\partial g_{dm}(\mathbf{y_i})}{\partial \beta} = \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{m=1}^{M} h_{idm}\frac{\partial g_{dm}(\mathbf{y_i})}{\partial \beta} = 0.$$

The first equation results in the explicit solution for $p_d^{(t+1)}$, below:

$$p_d^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} h_{idm}^{(t)},$$

$$\text{where} \qquad h_{idm}^{(t)} = P(D_i = d, R_i = r_m|\theta^{(t)}) = \frac{p_d^{(t)}\omega_m g_{dm}^{(t)}(\mathbf{y_i})}{\sum_{d=0}^{L-1}\sum_{m=1}^{M} p_d^{(t)}\omega_m g_{dm}^{(t)}(\mathbf{y_i})},$$

On the other hand, $\beta^{(t+1)} = \{a_{kdj}^{(t)}, b_{kdj}^{(t)}\}$ was obtained by solving the second equation:

$$\sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{m=1}^{M} h_{idm}^{(t)}\frac{\partial g_{dm}^{(t)}(\mathbf{y_i})}{\partial \beta} = 0, \tag{4.4}$$

subject to the monotonically nondecreasing constraint that $a_{kd1} \leq a_{kdJ-1} \leq \ldots \leq a_{kdJ-2}$ for $k = 1, 2, ..., K$ and $d = 0, 1, ..., L-1$.

Equation (4.3) is essentially a score equation for a generalized linear model. To see this more clearly, I defined

$$\begin{aligned}\mu_{kdmj} &= P(y_{ik} = j|D_i = d, R_i = r_m)\\ &= \Phi(a_{kdj} + b_{kdj}r_m) - \Phi(a_{kdj-1} + b_{kdj-1}r_m)\\ &\quad j = 1, 2, \ldots, J-1.\end{aligned}$$

Then,

$$g_{dm}(\mathbf{y_i}) = \prod_{k=1}^{K} \mu_{kdm1}^{y_{ik1}}\mu_{kdm2}^{y_{ik2}}\cdots\mu_{kdm(J-1)}^{y_{ik(J-1)}}\big(1 - \sum_{j=1}^{J-1}\mu_{kdmj}\big)^{y_{ikJ}}.$$

Therefore, $g_{dm}(\mathbf{y_i})$ can be considered as a likelihood function for K independent multinomial observations, and $\frac{\partial g_{dm}^{(t)}(\mathbf{y_i})}{\partial \beta}$ is the corresponding score function. Consequently, equation (4.3) can be considered a score equation for a generalized linear model with $NDMK$ independent multinomial observations and weight h_{idm}, which can be solved iteratively by many software packages.

4.2.3 Testing the Conditional Independence Assumption

Adding a random effect into the model can relax the conditional independence assumption and explain some possible dependence structures among the tests within each disease group. However, it is of interest to know how much this random effect improves the model fit, and whether it is really necessary to include the random effect term for a given data set. Another side of this question is, can we develop a test to assess the conditional independence assumption? The null hypothesis here is: H_0: there is no random effect (i.e. the conditional independence assumption holds).

In the model with a random effect, the test distribution within a disease group is modeled as a function of disease status and the random effect $P(T_{ik} \leq j | D_i = d, R_i = r) = \Phi(a_{kdj} + b_{kdj} r)$. If the conditional independence assumption holds, the test results are independent, only conditional on the true disease status. In other words, the random effect term becomes a fixed term. Equivalently, the variance of the random effect component is zero. Therefore, testing the conditional independence assumption for the test distribution is equivalent to testing the variance of the random effect $b_{kdj} = 0$.

I can use a likelihood ratio test. The test statistic is given as follows:

$$\Lambda(\mathbf{Y}) = -2 \log \frac{\sup_{p_d, a_{kdj}, b_{kdj} \geq 0} L(p_d, a_{kdj}, b_{kdj})}{\sup_{p_d, a_{kdj}, b_{kdj} = 0} L(p_d, a_{kdj}, b_{kdj})},$$

where $L(p_d, a_{kdj}, b_{kdj})$ is the observed data likelihood.

This test statistic is not regular since $b_{kdj} = 0$ is on the boundary of the parameter space. Under the null hypothesis, $\Lambda(\mathbf{Y})$ is a 50:50 mixture of χ^2_0 and χ^2_g, where g is the number of free parameters among b_{kdj}. Further discussion on this can be found in Self and Liang (1987).

4.3 Simulation Studies

In this section I evaluate the performance of the proposed model with simulations.

4.3.1 Small Sample Property

I chose a small sample size $N = 40$ and considered three different prevalence rates, as in the previous chapter. The first scenario assumes equal proportions for patients with different

disease severities: $p_0 = 0.25, p_1 = 0.25, p_3 = 0.25, p_4 = 0.25$. The second one assumes that there are more patients with higher severity status: $p_0 = 0.1, p_1 = 0.2, p_3 = 0.3, p_4 = 0.4$. And the last one assumes a U-shaped prevalence rate, where most patients are either in the asymptomatic or the most severe groups, while fewer patients are in the two groups with mild or moderate disease: $p_0 = 0.25, p_1 = 0.1, p_3 = 0.15, p_4 = 0.5$. I assumed that the subject-specific random effect ha the same impact on different tests in different disease groups: $b = 1.0$. A summary of results based on 500 simulations is shown in Table 4.1.

Table 4.1: Results from a random effect model based on 500 simulations with $N = 40$, L=3, K=5, J=3 under various prevalence rates and parameter settings.

True prevalence	Statistics	V_1	V_2	V_3	V_4	V_5	ϕ^*_{dkj}			b
rates	True values	0.356	0.408	0.505	0.606	0.703	max	min	mean	1.0
p_0=0.25,p_1=0.25	Bias	-0.062	-0.080	-0.114	-0.144	0.057	-0.109	0.001	0.033	0.058
p_2=0.25,p_3=0.25	MSE	0.008	0.008	0.012	0.021	0.027	0.030	0.003	0.014	0.069
p_0=0.10,p_1=0.20	Bias	-0.060	-0.074	-0.096	-0.132	-0.160	0.107	0.001	0.038	0.116
p_2=0.30,p_3=0.40	MSE	0.009	0.011	0.016	0.025	0.033	0.042	0.004	0.025	0.092
p_0=0.25,p_1=0.10	Bias	-0.071	-0.077	-0.103	-0.136	-0.163	0.149	0.001	0.044	0.128
p_2=0.15,p_3=0.50	MSE	0.011	0.011	0.017	0.026	0.034	0.049	0.005	0.017	0.095

* Results for all ϕ_{dkj} with $d = 0, 1, ..., 3$, $k = 1, 2, ..., 5$, and $j = 0, 1, ..., 3$.

For diagnostic probabilities, the estimation is better when the disease groups are relatively balanced in size. Compared to the results in the previous chapter, the bias of the estimates increased, but was still acceptable. For a sample size of 40, the average bias was less than 0.05 for ϕ_{dkj} and about 0.1 for V_k. As a reference, the results for the diagnostic probabilities for the first scenario are shown in Table 4.2. The true diagnostic probability matrix for the simulations is shown in Appendix D.

Comparable simulation results under different parameter settings can be found in Xie, et al. 2013.

4.3.2 Misspecification of the Random Effect Distribution

The additional random effect component allows for some dependence structure among the tests, but assumes a parametric distribution for the random effect. A misspecified distribution may result in biased estimates.

In the situation with binary disease status, Albert, et al. (2001) demonstrated that inferences about sensitivities and specificities can be sensitive to assumptions about the random effect distribution. Specifically, they showed that the Gaussian random effects model can be sensitive to the assumption of a symmetric random effects distribution. Albert and Dodd (2004) showed that, for a limited number of binary tests it is very difficult to distinguish between competing models with respect to the dependence between tests.

With the additional information gained from the ordinal gold standard, I expected that this problem would not be as severe as when the disease is binary. In fact, in a recently published paper, Xie, et al. (2013) evaluated the models performance under different random effect distributions. Their results suggest that, in contrast to a binary setting, penalized likelihood criteria can select the correct model for situations wherein the gold standard and test results are both ordinal.

The authors also investigated the impact of misspecification when the random effect component is assumed normal, but in fact follows a mixture normal distribution; their model has both a patient-level random effect and a physician-level random effect. Their results suggest that this type of model misspecification has a minimal impact on diagnostic probabilities, which are the primary interest of the diagnostic study.

4.4 Real Data Analysis

In this section I apply the proposed method to the same TCM data set used in the previous chapter and assess the diagnostic performances of the doctors without assuming conditional independence. Here I use symptom 1, wind-phobia, as an example. For illustrative purposes, the variances of the random effects are assumed to be the same for all the doctors and all the true symptom statuses, i.e., the conditional distributions of the diagnosis results are

modeled as:

$$P(T_{ik} \leq j | D_i = d, R_i = r) = \Phi(a_{kdj} + br), \qquad R_i \sim N(0,1)$$
$$d = 0, 1, ..., L-1, \ k = 1, 2, ..., K, \quad \text{and} \quad j = 1, 1, ..., J-1,$$

where a_{kdj} are monotonically nondecreasing cutpoints, $a_{kd1} \leq a_{kd2} \leq \ldots \leq a_{kdJ-1}$ for $k = 1, 2, ..., K$ and $d = 0, 1, ..., L-1$. Here, L=3, K=5, and J=3.

4.4.1 Results

The results are shown in table 4.1. The rows are the true symptom statuses, and the columns are the cumulative diagnostic probabilities. The accuracy summary measure described in Section 3.3 is also given to provide a general evaluation of doctors' performances.

The results indicate that doctor 3 had the best diagnostic ability for this symptom. Doctor 4 also had good performance, except that his diagnoses tended to be a little understated when compared to the true symptom statuses. The same tendency was also apparent for doctor 2, who was estimated to rank 3rd among the five doctors. The diagnosis performance was not so good for doctor 5, who also tended to understate the symptom statuses. Compared with the results given by Wang, et al. under the conditional independence assumption, the correct diagnostic probabilities here were slightly lower, as well as the overall accuracy measures for all the doctors. However, in general, the results were comparable to the ones obtained from a model without the random effect term. Especially, the ranking of these five doctors, based on the accuracy summary measures, remained the same, because the conditional independence assumption was reasonable, as I argued previously, and adding random effects to the model did not cause too much change. The estimated variance of the patient-specific random effects was small ($b = 0.007$), which also suggested that the conditional independence assumption was reasonable for this data set.

Two forest plots representing the doctors' diagnostic accuracies are shown in Figure 4.2. In general, it is evident that the diagnostic abilities of Doctor 3 (the red line) and Doctor 4 (the blue line) were better than those of the other doctors, since their lines are at a relatively higher position in the accuracy graph and at a relatively lower position in the inaccuracy graph. From the accuracy plot, it is evident that Doctor 3 had exceptionally good

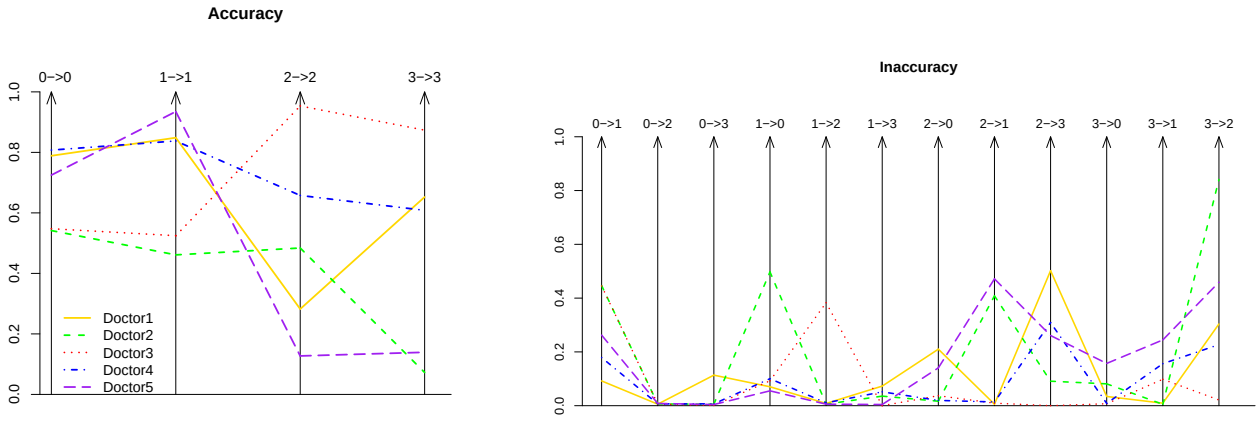

Figure 4.2: Forest plots for doctors' diagnostic abilities based on model with random effect.

performance among the subjects with medium or severe symptoms. In contrast, Doctor 4 had better performance among the subjects with no or mild symptoms. I also found that the accuracy performance of Doctor 4 was relatively stable across symptom severity levels. However, he had a relatively higher misdiagnosis rate than Doctor 3, as suggested by the inaccuracy plot, which resulted in a slightly lower accuracy summary measure than Doctor 3. Similarly, it is evident that the diagnostic ability of Doctor 2 is not so good. Again, the results agree with those in the previous chapter.

4.4.2 Goodness of fit

Now I will consider the evaluation of the goodness of fit of the model. Note that, since the diagnosic results $(T_1, \ldots, T_5)$ are correlated, a test based on the joint, rather than the marginal distribution of the diagnostic results, should be considered. One potential choice for assessing the goodness of fit is to compare the estimates of $(T_1 = j_1, \ldots, T_5 = j_5)$ with the observed numbers of the same quantities for all $j_1, \ldots, j_5 = 0, 1, 2, 3$. In other words, one can construct a five dimensional contingency table with numbers of $(T_1 = j_1, \ldots, T_5 = j_5)$ as cells, $j_1, \ldots, j_5 = 0, 1, 2, 3$, and perform a χ^2 test. However, for 5 doctors with 4 diagnosis

categories, the table contains 4^5 cells in total. Consequently, a lot of cells would result in zero observations, which will compromise the validity of the test. Due to this difficulty, I performed tests on the marginal results of each pair of doctors. Specifically, for each pair of doctors (e.g. doctor 1 and doctor 2), I constructed a 2 dimensional table of the marginal test results ($T_1 = j_1, T_2 = j_2$), $j_1, j_2 = 0, 1, 2, 3$ and performed a χ^2 test with 15 degrees of freedom. The results are shown in Table 4.2.

Note that the multiple comparisons increase the probability of getting a significant p-value, since I performed 10 tests. Thus, in order to maintain an overall 0.05 type I error rate of the tests, a modified significance level should be less than 0.05. However, even without modification, none of the tests were significant. Thus, the tests did not suggest significant differences between the estimated diagnosic results based on the model and the observed data.

In addition, I performed the likelihood ratio test discussed previously, in Section 4.2.3, to assess the conditional independence assumption. Here, I only had one parameter b for the random effect component, so the test statistic

$$\Lambda(\mathbf{Y}) = -2 \log \frac{\sup_{p_d, a_{kdj}, b \geq 0} L(p_d, a_{kdj}, b)}{\sup_{p_d, a_{kdj}, b=0} L(p_d, a_{kdj}, b)},$$

follows a 50:50 mixture of χ^2_0 and χ^2_1.

For the TCM data, I computed $\Lambda(\mathbf{Y}) = 0.373$, and p-value$= 1 - \frac{1}{2} - \frac{1}{2} P(\chi^2_1 \leq 0.373) = 0.27$. In conclusion, the test did not reject the conditional independence assumption in favor of the existence of a random effect. It can also be considered an explanation of the similarity between the results based on random effects modeling and the results that rely on the conditional independence assumption.

4.5 Summary and Further Models

In this chapter I applied the random effects model to account for the correlation structure among different tests or doctors when estimating the diagnostic performance without a gold standard. This method is still in the latent class modeling framework and is an extension to the nonparametric maximum likelihood method discussed in the previous chapter. The dependence among the tests/doctors was modeled by a patient-specific random effect with

probit link, which extended the latent class analysis proposed by Qu, Tan, and Kutner (1996) to ordinary outcomes. This approach can be applied to estimate the diagnostic accuracy of multiple tests or raters when the gold standard is not available. It has two advantages. First, it allows both the test results and the disease statuses to have ordinal scales, and second, it does not rely on the conditional independence assumption, and so allows for dependence among the tests through patient level random effects.

I introduced an unobserved random effect, R, to characterize the correlation structure among tests results. I assumed that the random effect is independent of the true disease status D. However, this assumption was only used in equation (3). The model can be extended to situations without this assumption if necessary. For example, if the true disease status D is believed to have a "shifting" effect on R and assume $(R|D = d) \sim N(\mu_d, 1)$, then equation (3) can be written as follows:

$$\begin{aligned} & g_d(\mathbf{y_i}) = P(\mathbf{y_i}|D_i = d) \\ = & \int_{-\infty}^{+\infty} \prod_{k=1}^{K} \prod_{j=0}^{J-1} [P(y_{ijk}|D_i = d, R_i = r)]^{y_{ikj}} dP(r|D = d) \\ = & \int_{-\infty}^{+\infty} \prod_{k=1}^{K} \prod_{j=0}^{J-1} \Big\{ \Phi\big[a_{kdj} + b_{kdj}(r - \mu_d)\big] - \Phi\big[a_{kdj-1} + b_{kdj-1}(r - \mu_d)\big] \Big\}^{y_{ikj}} d\Phi(r). \end{aligned}$$

and the derivations can still carry through.

However, similar to any latent model for estimating diagnostic accuracies, one must recognize the fact that, if the latent structure is misspecified, the performance of the proposed estimates can be poor. Compared with the binary gold standard case, this problem is less of an issue because, with additional information from ordinal data, the correct distribution is more likely to be identified. Nonetheless, results based on the latent structure should still be used with caution. Future research can consider different models for the latent structure, and can further investigate the model checking problem in the ordinal case.

The dependence structure among tests relaxes the conditional independence assumption, and thus accommodates more situations in practice. However, it assumes that the entirety of the correlation structure is accounted for by some unobserved patient-specific variables, which can be summarized by a continuous Gaussian random effect. This may not be true in some cases. For example, different patients' characteristics may have different impacts

on the diagnostic probabilities. When covariate information is available, it can be utilized to improve the model fit. Additionally, sometimes it is of interest to examine the test performance for subjects with different characteristics. Models that accommodate covariate effects will be discussed in the next chapter. Another limitation of this random effect model is that the dependence among the tests is explained by a regression model. As a result, although the model relaxes the conditional independence assumption, it does not model all possible dependence structures. Also, it is now a semi-parametric model rather than a non-parametric model, and compared to the model without a random effect, the EM algorithm for this approach does not yield an explicit solution for β, which makes the computation more difficult and time consuming.

Table 4.2: Bias and standard deviation estimate (in parentheses) of the diagnostic probabilities in scenario 1 based on 500 simulations.

		$T_k = 0$	$T_k = 1$	$T_k = 2$	$T_k = 3$
Doctor 1	$D = 0$	0.031* (0.13)**	-0.038 (0.16)	-0.035 (0.13)	0.042 (0.07)
	$D = 1$	0.041 (0.12)	-0.050 (0.17)	-0.020 (0.13)	0.028 (0.07)
	$D = 2$	0.048 (0.08)	-0.001 (0.14)	-0.053 (0.15)	0.006 (0.10)
	$D = 3$	0.043 (0.09)	0.002 (0.14)	-0.034 (0.16)	-0.011 (0.13)
Doctor 2	$D = 0$	0.018 (0.11)	-0.035 (0.14)	-0.020 (0.12)	0.037 (0.07)
	$D = 1$	0.031 (0.12)	-0.045 (0.16)	-0.012 (0.12)	0.027 (0.07)
	$D = 2$	0.051 (0.08)	-0.003 (0.14)	-0.054 (0.15)	0.006 (0.10)
	$D = 3$	0.049 (0.08)	0.007 (0.12)	-0.013 (0.14)	-0.043 (0.12)
Doctor 3	$D = 0$	0.011 (0.12)	-0.045 (0.13)	-0.017 (0.10)	0.050 (0.06)
	$D = 1$	0.039 (0.11)	-0.052 (0.15)	-0.024 (0.12)	0.036 (0.06)
	$D = 2$	0.043 (0.08)	0.005 (0.14)	-0.042 (0.15)	-0.006 (0.10)
	$D = 3$	0.045 (0.06)	0.022 (0.12)	-0.027 (0.14)	-0.039 (0.12)
Doctor 4	$D = 0$	0.005 (0.11)	-0.048 (0.13)	-0.013 (0.07)	0.055 (0.04)
	$D = 1$	0.037 (0.11)	-0.057 (0.15)	-0.020 (0.10)	0.040 (0.06)
	$D = 2$	0.054 (0.07)	0.011 (0.12)	-0.071 (0.14)	0.007 (0.09)
	$D = 3$	0.046 (0.07)	0.022 (0.12)	-0.019 (0.14)	-0.050 (0.12)
Doctor 5	$D = 0$	-0.028 (0.09)	-0.018 (0.09)	-0.008 (0.07)	0.054 (0.05)
	$D = 1$	0.041 (0.10)	-0.085 (0.12)	-0.006 (0.07)	0.050 (0.04)
	$D = 2$	0.057 (0.06)	0.021 (0.10)	-0.019 (0.12)	0.031 (0.07)
	$D = 3$	0.050 (0.06)	0.032 (0.06)	-0.015 (0.08)	-0.067 (0.08)

* Bias **Standard deviation

Table 4.3: Parameter estimates and corresponding overall accuracy measures of the five doctors from a latent class model with random effect.

True symptom status		Cumulated diagnosis probabilities			
		$T_k \leq 0^*$	$T_k \leq 1$	$T_k \leq 2$	$T_k \leq 3$
Doctor 1	d=0	0.789	0.881	0.886	1.000
	d=1	0.070	0.919	0.927	1.000
	d=2	0.210	0.216	0.498	1.000
	d=3	0.034	0.044	0.347	1.000
Doctor 2	d=0	0.542	0.988	0.994	1.000
	d=1	0.497	0.959	0.964	1.000
	d=2	0.017	0.425	0.909	1.000
	d=3	0.081	0.086	0.928	1.000
Doctor 3	d=0	0.548	0.993	0.999	1.000
	d=1	0.092	0.617	1.000	1.000
	d=2	0.037	0.047	1.000	1.000
	d=3	0.007	0.105	0.127	1.000
Doctor 4	d=0	0.807	0.986	0.993	1.000
	d=1	0.100	0.938	0.949	1.000
	d=2	0.020	0.033	0.691	1.000
	d=3	0.008	0.165	0.392	1.000
Doctor 5	d=0	0.725	0.987	0.996	1.000
	d=1	0.055	0.990	0.986	1.000
	d=2	0.140	0.612	0.739	1.000
	d=3	0.157	0.402	0.861	1.000
Variance of patient-specific random effect: $b = 0.007$					
Overall accuracy measures: $V_1 = 0.524$ $V_2 = 0.605$					
$V_3 = 0.821$ $V_4 = 0.700$ $V_5 = 0.497$					

* $T_k \leq j$ denote the cumulated conditional diagnosis probabilities $P(T_k \leq j | D_i = d)$

Table 4.4: χ^2 tests based on the marginal diagnosis results of each pair of the doctors.

df=15	χ^2 statistics	p-value
Doctor 1 and 2	15.58	0.59
Doctor 1 and 3	12.83	0.38
Doctor 1 and 4	12.45	0.36
Doctor 1 and 5	14.30	0.50
Doctor 2 and 3	12.33	0.35
Doctor 2 and 4	13.99	0.47
Doctor 2 and 5	15.34	0.57
Doctor 3 and 4	14.95	0.54
Doctor 3 and 5	14.15	0.49
Doctor 4 and 5	10.05	0.18

Chapter 5

CONTINUOUS TESTS AND BIOMARKER ASSESSMENT WITHOUT A GOLD STANDARD

5.1 Introduction

5.1.1 Continuous Tests and Biomarkers

Traditionally, most diagnostic tests result in direct conclusions about the presence or absence of a condition, and its severity level or subtype. Therefore, methods for studying diagnostic test accuracy often assume that test results are based on a binary or categorical scale, as in the models discussed in Chapters 3 and 4 of this dissertation. However, as I mentioned in Chapter 1, a broader definition of a diagnostic test includes any technique that provides information used for detecting, diagnosing or monitoring any medical condition or event of interest. Therefore, various clinical tests, including biomarker measures and some statistical models that combine a series of risk factors, and that usually have continuous values, such as the Framingham risk score for a cardiovascular event and the Gail model for breast cancer, can all be regarded as diagnostic tests. In this Chapter latent profile models are utilized and a method for comparing and evaluating multiple continuous diagnostic tests without a gold standard are discussed. The method can be used to compare and evaluate these tests or risk models, to assess and select various novel biomarkers, and possibly to combine several measures as a means of obtaining a better diagnosis.

5.1.2 Modeling with Covariates

In Chapter 2 I constructed hypothetical examples to illustrate the impact of ignoring the covariate effect on diagnostic accuracy estimates. In practice, disease prevalence can vary greatly among groups that have different sets of risk factors. Test performance can also change with patients characteristics. Incorporating covariate information can improve model fit and facilitate a better understanding of a test's properties among various subpopulations.

Such methods have great value, especially with the increasing interest on personalized health care.

One advantage of a latent class/profile model, compared with cluster analysis, is that it employs model based approaches. Its methodology offers a basic and flexible structure, where different models can be adopted for different problems to better describe the underlying latent groups and the manifest variables within each group; incorporating covariates is a relatively straightforward process.

In fact, a large body of literature discusses various methods for incorporating covariate effects into the latent class or latent profile models. Perhaps more development originates from work on the latent class models, which are usually referred to as latent class models with covariates, or regression extended latent class models, even when the manifest variables are continuous. Covariates can be included in a logistic or polytomous regression in the latent structure part of the model to allow for dependence between the construct summarized by the latent variable on other factors (Dayton and Macready, 1988; Bandeen-Roche, et al., 1997; Bartolucci and Forcina, 2006; Peterson, et al., 2012), or included in a regression in the measurement part of the model to describe the relationship between the manifest variables and other covariates within each latent group (Melton, Liang and Pulver, 1994). The former is more common. This is because latent variable modeling is most active in social and psychiatric studies, where the latent construct usually varies with covariates. For example, health status can depend on socioeconomic status, life habits, etc., and the goal of these studies is usually to understand how an abstract construct, such as mental health status or behavior type relates to some manifest factors. Including covariates in the latent structure model therefore seeks to estimate the effect of the manifest variables on the latent construct. On the other hand, including covariates in the measurement model strives to understand how the manifest variables within each latent group vary with other covariates – for example, does a questionnaire tend to yield higher scores for women than for it does for men with the same disease severity? In a diagnostic accuracy study, both approaches are relevant. First, disease prevalence varies with risk factors, so it is of interest to understand how they relate to each other. Second, it is important to understand how the test performance changes among subjects with different characteristics.

Attempts to simultaneously include covariates in latent structure models and measurement models date back to 1980's (Clogg and Goodman, 1984, 1985; Formann, 1985, 1992). These methods incorporate covariates by stratifying combinations of risk factors; thus, these methods are highly restricted to categorical covariates. Due to computational difficulty and the identifiability issue, more general methods are still limited. Huang and Bandeen-Roche (2004) proposed a method that can simultaneously include covariates in both models, and allows for the use of categorical and continuous covariables. However, they assume that the covariate effects are constant across all disease severity groups. More research is needed to develop methods that do not require this restriction.

5.1.3 Motivating Example: Biomarkers for Preclinical Alzheimer's Disease

Alzheimer's disease (AD) is a progressive and fatal neurodegenerative disorder, and is the most common form of dementia. Prevalence studies suggest that AD affected 4.5 million people in the United States in 2004. Without advanced therapy, this number is predicted to rise to 13.2 million by 2050 (Hebert, et al., 2003). The percentage of persons with AD increases by a factor of two with approximately every five years of age; 1 percent of 60-year-olds and about 30 percent of 85-year-olds have the disease (Jorm, 1991).

Typical clinical symptoms include memory and cognitive impairment, declines in language and visuospatial function, as well as altered behavior. Figure 5.1 shows cross-section brain images comparing a normal subject to an AD patient. We can see that in the AD patient, the grooves in the brain, called sulci, are widened and that the well-developed folds of the brain's outer layer, called gyrus is reduced; the ventricles area is also enlarged. In the early stages of Alzheimer's disease, short-term memory begins to fade when the cells in the hippocampus degenerate. Then, as Alzheimer's disease spreads through the outer layer of the brain, judgment declines and language is impaired. Consequently, patients have difficulties in daily living activities, such as writing checks or using public transportation. As the disease progresses, more nerve cells die, leading to changes in affected persons behaviors. Patients may then have trouble with performing some more basic activities of daily living, such as eating and grooming.

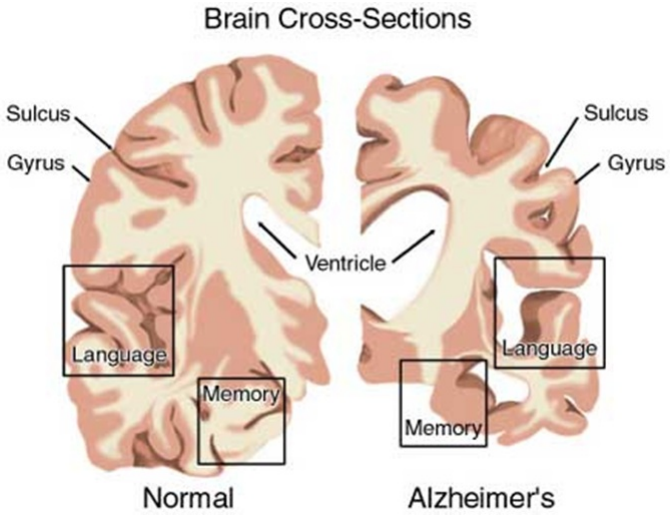

Figure 5.1: Cross-section brain images comparing a normal subject and an AD patient.

Recent AD research emphasizes the preclinical stage of AD (Sperling, et al., 2011). This is because the AD-related pathological changes are believed to begin 10 years or more before any detectable clinical symptom is apparent, and decades before there is sufficient cognitive impairment to warrant a clinical diagnosis (Jack, et al., 2009). If detected early, this long "preclinical" phase of AD can provide a critical opportunity for therapeutic intervention, since disease-modifying therapies for AD are likely to be most effective when used before significant neurodegeneration has occurred. However, the AD pathological process is unobservable without a brain autopsy. Guided by the amyloid hypothesis (Selkoe, 1991), researchers utilize biomarkers to help detect and monitor AD pathological changes. Amyloid$-\beta$ (Aβ) and tau proteins are currently among the most promising biomarkers; they are briefly introduced below.

Amyloid-β

The amyloid β peptide is a normal byproduct of neuronal activities. After its generation, the Aβ is moved to the spinal fluid for breakdown and disposal; it can then be detected in both the cerebrospinal fluid (CSF) and the plasma of healthy subjects throughout their lives. However, in AD patients, there is an imbalance of Aβ production and clearance. This imbalance causes Aβ to undergo a conformational change, such that it self-aggregates into

toxic plaques that deposit inside the brain. This is one of the two classic lesions in an AD brain (Figure 5.2). It leads to decreased Aβ levels in the CSF and increased Aβ levels in the brain (Hardy and Selkoe, 2002; Cummings, 2004). An increased amount of Aβ in the brain tissue is considered the hallmark of AD, and is specific to AD. In late-onset AD, the total amount of Aβ is about 100- to 200-fold higher in homogenates from Alzheimer's disease brains than from healthy brains (Gravina, et al., 1995). The high level of Aβ in brain then initiates a cascade of events eventually leads to Alzheimer's disease.

Tau protein

In addition to amyloid plaques, a neurofibrillary tangle is another classic lesion in an AD brain (Figure 5.2). The formation of this structure is related to effected tau proteins. Normally, tau proteins help to form and stabilize microtubules in neurons. These microtubules provide platforms for intracellular transport and are involved in a variety of cellular processes. In AD subjects, tau proteins are abnormally hyperphosphorylated and the microtubules twist together. This eventually causes the structure to collapse. Dystrophic tau neurites and tangles are released from damaged and dying neurons, resulting in increased CSF tau levels (Binder, et al., 2005). In contrast to amyloid plaques, defective tau proteins are also involved in other neurological diseases besides AD, such as progressive supranuclear palsy and Parkinson's disease. In current AD research, both total tau (t-tau) and hyperphosphorylated tau (p-tau) are used for detecting AD-related pathological changes.

Therefore, reduced CSF Aβ levels and increased CSF t-tau and p-tau levels can indicate AD-related pathological changes and be helpful for detecting preclinical AD. They provide an opportunity for early therapeutic intervention. In addition, this information can be useful for monitoring AD progress, evaluating new AD treatments and recruiting specific subsets in AD clinical trials. Current research often uses CSF Aβ or tau separately (Schoonenboom, et al. 2004; Shaw, et al. 2009) or the CSF tau/Aβ ratio (Li et al., 2007; Fagan, et al. 2007) as measures that reflect underlying AD pathological changes. Due to inadequate follow-up and limited resources to obtain autopsy data, the cut-off points used to classify normal or non-normal subjects, and the corresponding sensitivities and specificities, are generally obtained using clinical diagnosis as the gold standard. However, bias can occur with this approach because it ignores possible error in the clinical diagnosis. In

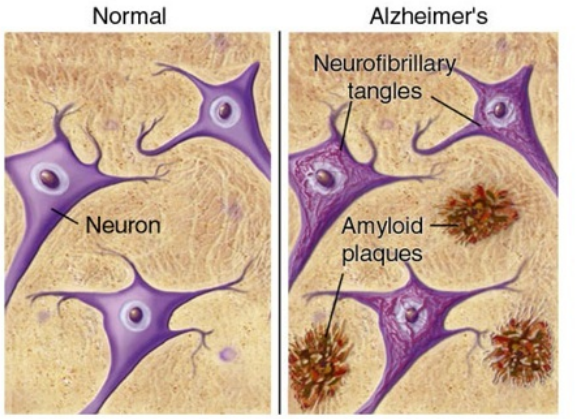

Figure 5.2: Two lesions in AD brain.

addition, due to the large time lag between the unobservable initiation of AD pathological processes and the onset of clinical symptoms, studying the progression from normal to mild cognitive impairment (MCI) requires a long follow-up period. Consequently, most research on early detection actually considers progression from MCI to AD, and some considers the transition to MCI in a not entirely normal group, but there is only limited research that investigates the use of biomarkers for early detection of the progression from normal to MCI. However, as mentioned above, a preclinical diagnosis is more important. Perhaps a better role for biomarkers is to help identify subjects that are clinically normal but with AD related pathological changes from subjects without AD pathologies. In this case, the gold standard is unobservable.

An additional challenge is that AD prevalence changes greatly with other risk factors, such as age, family history, depression, and cardiovascular disease. When developing a biomarker test, this information should be taken into account. Moreover, biomarkers may have different levels and diagnostic capacities for subgroups with different characteristics. Obtaining this information can be helpful in reaching a more informative, and perhaps more personalized, conclusion.

In this chapter I consider using latent profile models to account for this issue and to evaluate and combine biomarkers for preclinical AD detection. In section 2 I propose the

model and discuss its estimation. Section 3 investigates the properties of the proposed model by simulation studies. Real data analysis is given in Section 4, and finally, I summarize my findings and discuss avenues for future research in section 5.

5.2 A Latent Profile Model

Here I adopt the notation used before: T_k denotes results from the kth diagnostic test or biomarker, and D denotes the unobserved and possibly ordinal disease status. In addition, $\mathbf{X}$ denotes the covariates that may affect test performance and $\mathbf{Z}$ denotes the covariates related to disease prevalence. The elements in $\mathbf{X}$ and $\mathbf{Z}$ can be overlapping or mutually exclusive. Figure 5.3 gives a graphical representation of the model to be considered here.

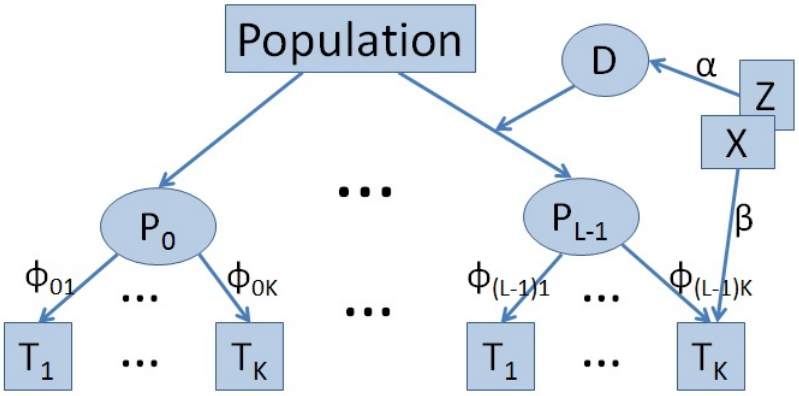

Figure 5.3: A latent profile model with two sets of covariates.

5.2.1 The Model

A latent profile model without covariates has the same form as the corresponding latent class model as below,

$$P(\vec{T}) = \sum_{d=0}^{L-1} P(D = d)P(\vec{T}|D = d),$$

except that the observed test results $\vec{T} = (T_1, \ldots, T_K)$ take continuous values.

Here I extend this model to incorporate covariate information. I assumed that $\mathbf{Z}$ was the set of covariates that affected disease prevalence and $\mathbf{X}$ was the set of covariates that affected test performance within each disease severity group. I further assumed that the test results were conditionally independent, given the true disease status D and the covariates $\mathbf{X}$. In other words, I assumed that the possible dependence among test results within each disease severity group was due to covariates $\mathbf{X}$, such as some subjects' characteristics. Compared with the random effect modeling in the previous chapter, I used specific covariates instead of a random effect to explain the possible dependence among test results more explicitly. Under this assumption, the latent profile model with covariates is given as follows,

$$P(\vec{T}|\mathbf{X},\mathbf{Z}) = \sum_{d=0}^{L-1} P(\vec{T}, D|\mathbf{X},\mathbf{Z}) = \sum_{d=0}^{L-1} \big[P(D=d|\mathbf{Z}) \prod_{k=1}^{K} P(T_k|D=d,\mathbf{X})\big].$$

Considering the true disease status as a missing value, the complete data log likelihood was:

$$l_c = \sum_{i=1}^{N}\sum_{d=0}^{L-1} I(D_i=d)\log P(D_i=d|\vec{Z}_i) + \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} I(D_i=d)\log P(T_{ik}|D_i=d,\vec{X}_i),$$

where $I(D_i=d)$ is an indicator function. It equals 1 if $D_i=d$ and equals 0 otherwise.

The next question is how to model the latent structure part $P(D|\mathbf{Z})$ and the measurement part $P(T|D,\mathbf{X})$. The former can be modeled with a latent polytomous regression since the latent variable D is ordinal. Let $\eta(\vec{Z}_i^T\alpha_d) = P(D_i=d|\vec{Z}_i=\vec{z})$ as in the generalized linear model framework, subject to the normalizing conditions $\sum_{d=0}^{L-1}\eta(\vec{z}^T\alpha_d)=1, \forall\vec{z}\in\mathcal{X}$ with $\mathcal{X}$ denoting the covariate space. Then the model for the latent structure part is,

$$\begin{aligned}
\eta(\vec{Z}_i^T\alpha_d) &= P(D_i=d|\vec{Z}_i=\vec{z}) = \frac{exp(\vec{z}^T\alpha_d)}{1+\sum_{l=1}^{L-1} exp(\vec{z}^T\alpha_l)},\\
log\frac{\eta(\vec{Z}_i^T\alpha_d)}{\eta(\vec{Z}_i^T\alpha_0)} &= \alpha_{d0}+\alpha_{d1}z_{i1}+...+\alpha_{dp}z_{ip}, \quad d=1,...,L-1,
\end{aligned}$$

where $D=0$ is the baseline group with parameters $\alpha_0=(\alpha_{00},\ldots,\alpha_{0p})=(0,\ldots,0)$.

As for $P(T_k|D,\mathbf{X})$, I assumed a transformation model as follows,

$$H_k(T_{ik}) = \vec{X}_i^T\beta_{kd}+\epsilon_{ik}, \quad \epsilon_{ik}\sim^{i.i.d.} G(\nu_k).$$

Here I chose $G(\nu_k)$ to be a normal distribution function $N(0,\sigma_k^2)$, but it can have any specified parametric distribution. $H_k(\cdot)$ is a monotonic transformation function. When the

error distribution is assumed normal, it accounts for possible skewness of the test results within each severity group. In this chapter, I assume that $H_k(\cdot)$ has the form of a Box-Cox power transformation (Box and Cox, 1964):

$$H_k(T_k, \lambda_k) = T_k^{(\lambda_k)} = \begin{cases} \frac{T_k^{\lambda_k}-1}{\lambda_k} & \lambda_k \neq 0 \\ \log T_k & \lambda_k = 0 \end{cases}.$$

Therefore, the latent profile model with covariate is given as follows,

$$P(\vec{T}|\mathbf{X},\mathbf{Z}) = \sum_{d=0}^{L-1} P(\vec{T}, D|\mathbf{X},\mathbf{Z}) = \sum_{d=0}^{L-1} \Big[P(D=d|\mathbf{Z}) \prod_{k=1}^{K} J(\lambda_k) P(T_k^{(\lambda_k)}|D=d,\mathbf{X}) \Big],$$

where $J(\lambda_k)$ is the Jacobian of the transformation $T_k \to T_k^{(\lambda_k)}$: $J(\lambda_k) = T_k^{\lambda_k - 1}$.

The maximum likelihood estimates of $\theta = \{\alpha_d, \beta_{kd}, \lambda_k, \nu_k \mid d = 0, \ldots, L-1 \text{ and } k = 1, \ldots, K\}$ can be obtained by directly maximizing the observed data likelihood function, or by applying the EM algorithm to complete the data likelihood function below

$$\begin{aligned} l_c(\theta) &= \sum_{i=1}^{N}\sum_{d=0}^{L-1} I(D_i=d) \log P(D_i=d|\vec{Z}_i) + \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} I(D_i=d) \log P(T_{ik}^{(\lambda_k)}|D_i=d,\vec{X}_i) \\ &\quad + \sum_{i=1}^{N}\sum_{k=1}^{K} (\lambda_k - 1) \log T_{ik}. \end{aligned}$$

It is worth noting that, in this model I the allow covariate effect β to depend on both test index k and disease status d. This is more general than current available methods, such as the one proposed by Huang and Bandeen-Roche (2004), where the covariate effects are assumed to be constant across disease groups. To better understand this improvement, I consider a special case when D is binary. The measurement model is given by

$$H_k(T_{ik}) = \vec{X}_i^T \beta_{kd} + \epsilon_{id}, \quad \text{where } \epsilon_{i0} \sim G_0 \text{ and } \epsilon_{i1} \sim G_1, \quad d = 0 \text{ or } 1.$$

For any cut-point c, the covariate-specific sensitivity and specificity of test k are as follows,

$$\begin{aligned} Sens_k(c,\vec{X}) &= P(T_k > c|D=1) = 1 - G_1(H_k(c) - \vec{X}^T\beta_{k1}), \\ 1 - Spec_k(c,\vec{X}) &= P(T_k > c|D=0) = 1 - G_0(H_k(c) - \vec{X}^T\beta_{k0}). \end{aligned}$$

Consequently, the covariate-specific ROC curve is given by:

$$ROC_k(t, \vec{X}) = P(T_{ik} > T_{jk} | D_i = 1, D_j = 0) = 1 - G_1(G_0^{-1}(1-t) + \vec{X}^T \beta_{k0} - \vec{X}^T \beta_{k1}).$$

If covariate effects are assumed the same for every disease group with only intercept terms differing, i.e., $\beta_{kd} = \{\gamma_{kd0}, \gamma_{k1}, \ldots, \gamma_{kq}\}$, the ROC curve no longer depends on covariates. In other words, when D is binary, assuming a constant covariate effect across disease groups is equivalent to assuming that the covariate-specific ROC curves are the same as the marginal ROC curves. This is rarely the case. Previously, the constraint of constant covariate effect across disease groups was imposed because of identifiability concerns, but I show that the model is still identifiable without this assumption. Chapter 6 gives a more detailed discussion on this subject.

5.2.2 Estimation

In this section, I apply the EM algorithm to obtain the maximum likelihood estimates of the parameters.

In the E step, I computed the expected value of the complete data log likelihood,

$$\begin{aligned} E[l_c^{(t)}(\theta) | \vec{T}_i, \vec{X}_i, \vec{Z}_i, \theta^{(t)}] &= \sum_{i=1}^{N} \sum_{d=0}^{L-1} P_i^{(t)}(d) \log \eta(\vec{z}_i^T \alpha_d^{(t)}) \\ &+ \sum_{i=1}^{N} \sum_{d=0}^{L-1} \sum_{k=1}^{K} P_i^{(t)}(d) \log g_k^{(t)}(t_{ik}^{(\lambda_k^{(t)})} - \vec{x}_i^T \beta_{kd}^{(t)}) + \sum_{i=1}^{N} \sum_{k=1}^{K} (\lambda_k^{(t)} - 1) \log t_{ik}, \end{aligned} \tag{5.1}$$

where superscript (t) denoted the iteration number, $g_k^{(t)}$ denoted the density function for $G(\nu_k^{(t)})$, and $P^{(t)}(d)$ was the expected value of $I(D_i = d)$ at current parameter values,

$$\begin{aligned} P_i^{(t)}(d) &\equiv E\big[I(D_i = d) | \vec{T}_i, \vec{X}_i, \vec{Z}_i, \theta^{(t)}\big] \\ &= P(D_i = d | \vec{T}_i, \vec{X}_i, \vec{Z}_i, \theta^{(t)}) \\ &= \frac{P(\vec{T}_i | D_i = d, \vec{X}_i) P(D_i = d | \vec{Z}_i)}{\sum_{d=0}^{L-1} P(\vec{T}_i | D_i = d, \vec{X}_i) P(D_i = d | \vec{Z}_i)} \\ &= \frac{\big[\prod_{k=1}^{K} g_k^{(t)}\big(H_k^{(t)}(t_{ik}) - \vec{x}_i^T \beta_{kd}^{(t)}\big)\big] \eta_d(\vec{z}_i^T \alpha^{(t)})}{\sum_{d=0}^{L-1} \big\{ \prod_{k=1}^{K} \big[g_k^{(t)}\big(H_k^{(t)}(t_{ik}) - \vec{x}_i^T \beta_{kd}^{(t)}\big)\big] \eta_d(\vec{z}_i^T \alpha^{(t)}) \big\}}. \end{aligned}$$

In the M step, I maximized the expected complete data log likelihood given in (5.1). Let

$$l_1(\alpha_d^{(t)}) = \sum_{i=1}^{N}\sum_{d=0}^{L-1} P_i^{(t)}(d)\log\eta(\vec{z}_i^T\alpha_d^{(t)}), \tag{5.2}$$

$$\begin{aligned} l_2(\lambda_k^{(t)},\beta_{kd}^{(t)},\nu_k^{(t)}) &= \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d)\log g_k^{(t)}(t_{ik}^{(\lambda_k^{(t)})} - \vec{x}_i^T\beta_{kd}^{(t)}) \\ &\quad + \sum_{i=1}^{N}\sum_{k=1}^{K}(\lambda_k^{(t)}-1)\log t_{ik}. \end{aligned} \tag{5.3}$$

Then, the function to be maximized was

$$E[l_c^{(t)}(\theta)|\vec{T}_i,\vec{X}_i,\vec{Z}_i,\theta^{(t)}] = l_1(\alpha_d^{(t)}) + l_2(\lambda_k^{(t)},\beta_{kd}^{(t)}).$$

Noting that parameters $\alpha_d^{(t)}$ were only involved in $l_1(\alpha_d^{(t)})$, and that parameters $\lambda_k^{(t)}$, $\beta_{kd}^{(t)}$ and $\nu_k^{(t)}$ were only involved in $l_2(\lambda_k^{(t)},\beta_{kd}^{(t)},\nu_k^{(t)})$, the maximization can be broken down into two parts:

$$\begin{aligned} \{\alpha_d^{(t+1)}\}_{d=0,\ldots,L-1} &= \text{argmax } l_1(\alpha_d^{(t)}), \\ \{\lambda_k^{(t+1)},\beta_{kd}^{(t+1)},\nu_k^{(t+1)}\}_{d=0,\ldots,L-1,k=1,\ldots,K} &= \text{argmax } l_2(\lambda_k^{(t+1)},\beta_{kd}^{(t)},\nu_k^{(t)}). \end{aligned}$$

Examining equation (5.2), one can see that $l_1(\alpha_d^{(t)})$ had the same form of the log likelihood function as a polytomous, or multinomial logit regression with N observation L categories, except that the group indicator function $I(D_i = d)$ is replaced by its expected value $P^{(t)}(d)$, since $\sum_{d=0}^{L-1} P^{(t)}(d) = 1$, $l_1(\alpha_d^{(t)})$ satisfies all of the properties for a log likelihood function. Nevertheless, $l_1(\alpha_d^{(t)})$ were concave and the maximization could be carried out in the same way, such as using the Newton-Raphson method. Its implementation could borrow the polytomous regression routine in many statistical software packages. My computation utilized the R function "multinom" in package "nnet".

The second part of the maximization function, $l_2(\lambda_k^{(t)},\beta_{kd}^{(t)},\nu_k^{(t)})$, given in equation (5.3), is also a log likelihood function. In fact, it is a log likelihood function of a transformation regression model with $N\times L\times K$ observations, weight $P_i^{(t)}(d)$, Box-Cox transformation and error distribution $G_k(\cdot)$. In fact, the original $N\times K$ (each of the N subjects were diagnosed by each of the K tests) can be considered replicated L copies with 1 in each disease severity group. This log likelihood function took the total $N\times L\times K$ observations, but weighted

them according to the posterior probability of a subject belonging to a disease severity group $P_i^{(t)}(d)$, computed in the E step. This is clearer with matrix notation. Assume that $\mathbb{X}(D)$ was the design matrix defined by the measurement model $P(T_k|D, \mathbf{X})$, including all covariates $\mathbf{X}$ and their interactions with D. The design matrix then slowed the covariate effects to depend on the latent disease status D. I created a stacked design matrix $\mathbb{X}_{st}$ with $N \times L$ rows,

$$\mathbb{X}_{st} = \begin{pmatrix} \mathbb{X}(D=0) \\ \dots \\ \mathbb{X}(D=L-1) \end{pmatrix}$$

Let $Y_k^{(t)}$ be the vector of the transformed outcomes with N elements at the tth EM iteration, $Y_{ik}^{(t)} = T_{ik}^{(\lambda_k^{(t)})} = (T_{ik}^{\lambda_k^{(t)}} - 1)/\lambda_k^{(t)},\ i = 1, \dots, N$. I also created a stacked outcome vector $Y_{k,st}^{(t)}$ with $N \times L$ elements,

$$Y_{k,st}^{(t)} = (\underbrace{Y_k^{(t)'}, \dots, Y_k^{(t)'}}_{L \text{ times}})',$$

where superscript $'$ denotes matrix transpose. In addition, let $\mathbb{W}$ be a $N \times L$ by $N \times L$ diagonal matrix $\mathbb{W}^{(t)} = diag\{P^{(t)}(0), \dots, P^{(t)}(L-1)\}$, where $P^{(t)}(d) = (P_1^{(t)}(d), \dots, P_N^{(t)}(d)),\ d = 0, \dots, L-1$.

Then equation (5.3) can be rewritten as follows,

$$l_2(\lambda_k^{(t)}, \beta_{kd}^{(t)}, \nu_k^{(t)}) \quad = \quad \sum_{k=1}^{K} \mathbb{W}^{(t)} \log g_k^{(t)} (Y_{k,st}^{(t)} - \mathbb{X}_{st}\beta_{kd}^{(t)}) + \sum_{k=1}^{K} (\lambda_k^{(t)} - 1) \log T_k. \qquad (5.4)$$

Maximization can be done using these stacked design matrix and stacked outcomes as if D is known. A more detailed procedure and derivations for homoscedastic error distribution $N(0, \sigma_k^2)$ and heteroscedastic error distribution $N(0, \sigma_{kd}^2)$ are given in Appendix E.

The above formulation with a design matrix also makes it easy to impose constraints on the covariate effect β_{kd}. For example, the model in Huang, et al. (2004) assumes that covariate effects are the same across disease groups. If one wants to impose this constraint on some of the covariates, they can simply remove the corresponding interaction terms between D and these covariates in the design matrix. Similarly, if the design matrix and outcome vectors for each k are stacked, and include the interaction terms between the covariate effects and the tests, then constraints on the covariate effects across different

tests can easily be included. Here, I do not further complicate the design matrix as the constraints on covariate effects across different tests are used less commonly, and without these constraints, the maximization in (5.4) can be performed separately for each k.

5.2.3 Computational Issues

Initial parameter values

The EM algorithm needs appropriate initial parameter values to start. Ideally, one chooses the most likely parameter values based on prior information to serve as the initial values. However, such information may not be available in many cases, and the specification of the initial values may be difficult, especially with multiple disease groups and many covariates. I used two methods for specifying the initial parameter values in my computations.

Initial values by clustering:

All parameters can be easily estimated if the true disease status D is known. To obtain some sensible label, D, for computing the initial parameters, a crude clustering on the multivariate test results, which ignores the covariates, can be performed. I used K-means clustering in my program, but other clustering methods can also be used. To determine which cluster corresponds to each disease group, I ranked the univariate results for each test separately among the clusters. Recall that I assumed that higher test results are more indicative of a more severe disease status. I labeled the clusters such that higher values of D are assigned to the clusters with higher summations of the test ranks. If ties occur the clusters are further ordered by the summation of the means of the standardized test results. I used D_0 to denote the resulting labeling for each subject, and then performed a polytomous regression with outcome D_0 and covariate $\mathbf{Z}$, as specified in the structure model $P(D|\mathbf{Z})$. The resulting regression coefficients can be used as initial values for α_d. Similarly, initial values for β_{kd}, and the transformation parameters λ_k and σ_k can be obtained based on a Box-Cox regression with outcome $\vec{T}$ and covariate $\mathbf{X}$ and D_0, as specified in the measurement model $P(\vec{T}|\mathbf{Z}, D)$.

Random starting values:

The EM algorithm for latent variable models can have different results depending on the

choices of initial parameter values, because the likelihood functions for these models usually have multiple local maxima. More discussions can be found in Seidel, et al. (2000). Therefore, I ran a large set of random initial values and compared the resulting likelihood maxima to determine the final results. The random starting values were generated by the following procedures. First, if prior information on disease severity group was available, the prevalence was generated accordingly. Otherwise, I randomly generated $L-1$ numbers $a_1, \ldots, a_{L-1}$ from the standard uniform distribution. I set the prevalence as $\vec{p} = \{a_1, a_2 - a_1, \ldots, a_{L-1} - a_{L-2}, 1 - a_{L-1}\}$. I then randomly generated disease label D_0 for each subject from a multinomial distribution with probability $\vec{p}$. Finally, I obtained the initial values α_d, β_{kd} and the transformation parameters as in the clustering method by polytomous regression and transformation regression.

Scaling the test results

The target function $l_2(\lambda_k^{(t)}, \beta_{kd}^{(t)}, \sigma_k^{(t)})$ in equation (5.3) for obtaining $\beta_{kd}^{(t+1)}$, $\lambda_k^{(t+1)}$ and $\sigma_k^{(t+1)}$ can be simplified when the outcomes are scaled by their geometric means. This scaling was originally proposed for the Box-Cox regression by Zarembka in 1968. Specifically, I divided the test results T_k by its geometric mean,

$$Z_k = \frac{T_k}{\dot{T}_k}, \quad \text{where } \dot{T}_k = \sqrt[N]{\prod_{i=1}^{N} T_{ik}} = \exp\big(\frac{1}{N}\sum_{i=1}^{N} \log T_{ik}\big), \quad k = 1, \ldots, K.$$

Since $\dot{T}_k$ is a constant, the same power transformation that makes T_k on a normal scale also makes Z_k on a normal scale. Therefore, the transformation parameter λ_k does not change. However, if Z_k is used, the Jacobian term in equation (5.3) disappears because

$$\sum_{i=1}^{N} \log Z_{ik} = \sum_{i=1}^{N} \log T_{ik} - N \times \frac{1}{N}\sum_{i=1}^{N} \log T_{ik} = 0.$$

I use the superscript * to denote the corresponding parameters when scaled outcomes Z_k are used. The target function becomes,

$$l_2(\lambda_k^{*(t)}, \beta_{kd}^{*(t)}, \sigma_k^{*(t)}) \;=\; \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d) \log g_k^{*(t)}(z_{ik}^{(\lambda_k^{*(t)})} - \vec{x}_i^T \beta_{kd}^{*(t)}).$$

Then the estimates for the original parameters can be recovered as follows,

$$\begin{aligned}
\lambda_k &= \lambda_k^*, \qquad \sigma_k = \dot{T}_k^{\lambda} \sigma_k^*, \\
\beta_{kl} &= \begin{cases} \dot{T}_k^{\lambda} \beta_{kl}^* + \dfrac{\dot{T}_k^{\lambda} - 1}{\lambda} & l = 0 \\ \dot{T}_k^{\lambda} \beta_{kl}^* & l = 2, \ldots, q \end{cases} \\
& \qquad k = 1, \ldots, K.
\end{aligned}$$

This scaling makes the maximization slightly easier. More importantly, because the test results can have very different ranges, I found that scaling by their geometric means also makes the computation more stable.

Spurious local maximizers

If the error distribution in the measurement model is allowed to be heteroscedastic among disease groups, that is

$$H_k(T_{ik}) = \vec{X}_i^T \beta_{kd} + \epsilon_{ikd}, \quad \epsilon_{ikd} \sim^{i.i.d.} N(0, \sigma_{kd}^2)$$

a relatively large local maximum can occur as a consequence of fitting a disease group having only a few and very close observations (therefore the variance in the denominator of the likelihood function becomes very small). The resulting parameter estimates are usually called spurious local maximizers. This problem is due to the fact that the likelihood function for a heteroscedastic mixture model can be unbounded. As a simple example, the likelihood function of a mixture model with two heteroscedastic normal distributions is unbounded (Lehmann and Casella, 1998, Example 6.10 in Chapter 6).

When the likelihood is unbounded, the MLE for the parameters does not exist as a global maximizer. However, this does not invalidate my method, because maximizing the likelihood function is not its goal; my goal is, essentially, to find some estimates that are consistent. In fact, in this situation, a sequence of roots of the likelihood function still exists as local maxima for my model, and they are consistent, efficient and asymptotically normal. Further discussion can be found in Peters and Walker (1978), Redner and Walker (1984), Cheng and Traylor (1995).

Since the consistent estimates are still local maximizers, they can be obtained via the EM algorithm. However, in addition to comparing the log likelihood maxima obtained, one needs to examine the relative size or the variance of the fitted component in order to exclude possible spurious local maximizers. In practice, they are usually easily identified, as some of the components have extremely small sizes and small variances compared to other components (see example 3.6 in McLachlan and Peel, 2000). This problem can also be handled by imposing constraints in the maximization, such as requiring that. for each given k,

$$\begin{aligned} & \sigma_{kd}/\sigma_{kh} \leq R, \quad \forall d, h = 0, \ldots, L-1 \\ \text{or} \quad & \sigma_{kd}/\sigma_{kh} \geq 1/R, \quad \forall d, h = 0, \ldots, L-1 \\ \text{or} \quad & 1/R \leq \sigma_{kd}/\sigma_{kh} \leq R, \quad \forall d, h = 0, \ldots, L-1 \\ \text{or} \quad & \sigma_{kd}/\sum_{h=0}^{D} \sigma_{kh} \leq R/(D+1), \quad \forall d = 0, \ldots, L-1 \end{aligned}$$

where R is a pre-specified positive constant that restricts the relative size of the component variance. I do not use this method in my computation because (1) in diagnostic settings spurious local maximizers can be easily identified with very minimal information about prevalence; (2) unconstrained maximization is easier and faster than constrained ones; and (3) it can be useful to obtain and examine all results because sometimes a result with dissimilar components is not a spurious solution, see example 3.9 in McLachlan and Peel, 2000 on galaxy data. In addition, it can be used to detect a nuclear explosion, an earthquake, or other situations where one component is expected to be very small (Wang, et al., 1997; Sain, et al., 1999).

5.3 Simulation Studies

In this section I assess the performance of the proposed method with simulation studies. In all simulations, I chose 3 tests ($K = 3$) and 3 disease categories ($L = 2$). In the latent structure model, I assumed that a binary variable $Z \sim Bern(0.5)$ affects the disease prevalence $P(D_i = d|Z_i = z) = \eta(\alpha_{d0} + \alpha_{d1}z)$, $d = 1, 2$. I chose $\alpha = (\alpha_{10}, \alpha_{11}, \alpha_{20}, \alpha_{21}) = (-0.5, 1, -1, 1.5)$. This parameter value resulted in prevalence $\vec{p} \approx (0.51, 0.31, 0.19)$ among

subjects with $Z = 0$ and $\vec{p} \approx (0.23, 0.38, 0.38)$ among subjects with $Z = 1$, where $\vec{p} = (P(D = 0|Z), P(D = 1|Z), P(D = 2|Z))$. As a result, among subjects with $Z = 0$, most were healthy,,and about 20% had severe conditions, with the risk factor $Z = 1$ leading to more subjects having mild or severe conditions.

In the measurement model I assumed that a normally distributed variable $X \sim N(0, 1)$ affects the test performance $H_k(T_k) = \tilde{\beta}_{kd0} + \tilde{\beta}_{kd1} X_i + \epsilon_{ik}$, $d = 0, 1, 2$. This can be parameterized as

$$\begin{aligned} H_k(T_k) &= \beta_{k0} + \beta_{k1} X_i + \beta_{k2} I(D = 1) + \beta_{k3} I(D = 2) \\ &+ \beta_{k4} X I(D = 1) + \beta_{k5} X I(D = 2) + \epsilon_{ik}, \qquad k = 1, \ldots, K. \end{aligned}$$

It is easy to find that the relationship between $\tilde{\beta}$ and β is as follows,

$$\begin{aligned} &\tilde{\beta}_{k00} = \beta_{k0}, \quad \tilde{\beta}_{k01} = \beta_{k1}, \quad \tilde{\beta}_{k10} = \beta_{k0} + \beta_{k2}, \quad \tilde{\beta}_{k11} = \beta_{k0} + \beta_{k4} \\ &\tilde{\beta}_{k20} = \beta_{k0} + \beta_{k3}, \quad \tilde{\beta}_{k21} = \beta_{k0} + \beta_{k5}, \qquad k = 1, \ldots, K. \end{aligned}$$

I used $\epsilon_{ik} \sim N(0, 0.5^2)$ as the error distribution for all tests. I considered identity, square root and log transformations, $\lambda = 1$, 0.5 or 0. I chose two sets of values in the simulations: $\beta_k = (5, -1, 1, 2.5, 0.5, 1.5)$ representing tests with good discriminating ability, and $\beta_k = (5, -1, 0.5, 1, 0.5, 1.5)$ representing tests with fair discriminating ability. For the first scenario, the covariate-specific AUC for discriminating $D = 0$ versus $D = 1$ is $\Phi(2+X)$, where Φ is the standard normal cumulate distribution function. The average value of this covariate-specific AUC in this population is $\Phi(\frac{1}{\sqrt{0.5^2 + 0.5^2}}) \approx 0.92$. A more detailed derivation can be found in Appendix F. The covariate-specific AUC for discriminating $D = 1$ versus $D = 2$ is $\Phi(1.5+X)$, with the average value in this population being $\Phi(\frac{1.5}{\sqrt{1 + 0.5^2}}) \approx 0.91$. Similarly, for the second scenario, the covariate-specific AUC for discriminating $D = 0$ versus $D = 1$ is $\Phi(1+X)$, with the average value in this population being $\Phi(\frac{0.5}{\sqrt{0.5^2 + 0.5^2}}) \approx 0.76$. The covariate-specific AUC for discriminating $D = 1$ versus $D = 2$ is $\Phi(1 + 2X)$, with the average value in this population being $\Phi(\frac{1}{\sqrt{0.5^2 + 2^2}}) \approx 0.69$.

Simulation results for tests that have good or fair diagnostic performances are shown in Table 5.1 and 5.2. In each scenario, I considered sample sizes $N = 500$ and $N = 800$. All results are based on 500 simulation replicates.

Table 5.1: Mean and standard error (in parentheses) estimates based on 500 simulations for tests with good performance.

$N = 500$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.499 (0.12)	1.010 (0.18)	-1.008 (0.14)	1.512 (0.19)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 1$	$\beta_{k3} = 2.5$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.109 (0.64)	-1.036 (0.19)	1.044 (0.20)	2.629 (0.54)	0.507 (0.10)	1.568 (0.31)
$k = 2$	5.089 (0.45)	-1.032 (0.15)	1.034 (0.16)	2.606 (0.41)	0.511 (0.08)	1.557 (0.24)
$k = 3$	5.008 (0.22)	-1.007 (0.09)	1.010 (0.10)	2.526 (0.26)	0.501 (0.06)	1.511 (0.15)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	1.002 (0.11)	0.505 (0.06)	0.000 (0.02)	0.523 (0.11)	0.519 (0.08)	0.502 (0.05)
$N = 800$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.500 (0.12)	1.009 (0.18)	-0.997 (0.14)	1.502 (0.19)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 1$	$\beta_{k3} = 2.5$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.046 (0.64)	-1.016 (0.19)	1.019 (0.20)	2.558 (0.54)	0.504 (0.10)	1.529 (0.31)
$k = 2$	5.016 (0.45)	-1.007 (0.15)	1.008 (0.16)	2.526 (0.41)	0.502 (0.08)	1.515 (0.24)
$k = 3$	5.004 (0.22)	-1.002 (0.09)	1.001 (0.10)	2.513 (0.26)	0.502 (0.06)	1.504 (0.15)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	1.000 (0.11)	0.499 (0.06)	0.000 (0.02)	0.510 (0.11)	0.504 (0.08)	0.501 (0.05)

These results suggest that the proposed method converges to the true parameter values. Parameters λ_k and α_d can be well estimated with a sample size of 500. Increasing the sample size to 800 does not change their results much. On the other hand, bias in the estimates for β_{kd} decreases when the sample size increases. This is because, in my model β_{kd} are assumed to differ across tests and disease groups. Relatively, the effective sample size needed to estimate β_{kd} is less than that needed to estimate λ_k and α_d. The standard deviation σ_k for the error distribution in the measurement model is assumed to be the same across disease groups, but the standard deviation is relatively harder to estimate than the regression coefficient. Therefore, the bias is slightly smaller when the sample size increases. In addition, I found that the bias for β_{kd} and σ_k seemed to be smaller with smaller λ_k.

Table 5.2: Mean and standard error (in parentheses) estimates based on 500 simulations for tests with fair performance.

$N = 500$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.499 (0.12)	1.017 (0.18)	-0.993 (0.14)	1.501 (0.19)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 0.5$	$\beta_{k3} = 1$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.132 (0.64)	-1.048 (0.19)	0.525 (0.20)	1.054 (0.54)	0.521 (0.10)	1.580 (0.31)
$k = 2$	5.071 (0.45)	-1.025 (0.15)	0.510 (0.16)	1.029 (0.41)	0.507 (0.08)	1.546 (0.24)
$k = 3$	5.054 (0.22)	-1.020 (0.09)	0.508 (0.10)	1.024 (0.26)	0.504 (0.06)	1.537 (0.15)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	1.001 (0.11)	0.498 (0.06)	0.002 (0.02)	0.525 (0.11)	0.513 (0.08)	0.510 (0.05)
$N = 800$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.506 (0.12)	1.024 (0.18)	-1.010 (0.14)	1.516 (0.19)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 0.5$	$\beta_{k3} = 1$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.059 (0.64)	-1.022 (0.19)	0.509 (0.20)	1.026 (0.54)	0.512 (0.10)	1.536 (0.31)
$k = 2$	5.034 (0.45)	-1.015 (0.15)	0.506 (0.16)	1.013 (0.41)	0.507 (0.08)	1.522 (0.24)
$k = 3$	5.017 (0.22)	-1.007 (0.09)	0.502 (0.10)	1.009 (0.26)	0.503 (0.06)	1.513 (0.15)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	0.999 (0.11)	0.499 (0.06)	0.000 (0.02)	0.512 (0.11)	0.507 (0.08)	0.503 (0.05)

Comparing these two tables, the performances of the tests did not seem to affect the biases and the standard errors of the estimates.

In the simulations for tables 5.1 and 5.2, although I used the same parameters to generate test results for all three tests, I allowed β_{kd} to be different in the estimations. Therefore, the results displayed here represent the models performance when the tests had different performances. The same parameters β_{kd} in the data generation were used for the sake of clarity in the presentation, and to help examine the possible effects of different transformations on the estimates. As a reference, Table 5.3 shows the results when the first test had good performance and the other two tests had fair performances. It is evident that the model performance, in terms of bias and standard error, is similar to that found in previous results.

Table 5.3: Mean and standard error (in parentheses) estimates based on 500 simulations for tests with different performances.

$N = 500$		$\alpha_{10} = -0.5$	$\alpha_{11} = 1$	$\alpha_{20} = -1$	$\alpha_{21} = 1.5$	
		-0.516 (0.12)	1.050 (0.18)	-1.005 (0.14)	1.520 (0.19)	
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 1$	$\beta_{k3} = 2.5$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 1$	5.092 (0.64)	-1.032 (0.19)	1.034 (0.2)	2.607 (0.54)	0.516 (0.1)	1.555 (0.31)
	$\beta_{k0} = 5$	$\beta_{k1} = -1$	$\beta_{k2} = 0.5$	$\beta_{k3} = 1$	$\beta_{k4} = 0.5$	$\beta_{k5} = 1.5$
$k = 2$	5.006 (0.45)	-1.008 (0.15)	0.512 (0.16)	1.015 (0.41)	0.497 (0.08)	1.513 (0.24)
$k = 3$	5.021 (0.22)	-1.008 (0.09)	0.504 (0.10)	1.009 (0.26)	0.495 (0.06)	1.517 (0.15)
	$\lambda_1 = 1$	$\lambda_2 = 0.5$	$\lambda_3 = 0$	$\sigma_1 = 0.5$	$\sigma_2 = 0.5$	$\sigma_3 = 0.5$
	0.997 (0.11)	0.492 (0.06)	0.000 (0.02)	0.52 (0.11)	0.504 (0.08)	0.504 (0.05)

5.4 Real Data Study

In this section I applyiedthe proposed method to examine the diagnostic performance of CSF biomarkers in detecting Alzheimer's disease-related pathological changes. I also examined the impact of risk factors and subjects' characteristics on the prevalence of AD-related pathology and on biomarker levels. The data for my analyses were obtained from the Alzheimer's Disease Neuroimaging Initiative (ANDI) database (http://adni.loni.ucla.edu/).

5.4.1 Background

The pathological changes that lead to Alzheimer's disease are believed to happen decades before the onset of clinical symptoms, and even longer before there is sufficient cognitive impairment to warrant a clinical diagnosis (Price, et al., 1999; Jack, et al., 2009; Bateman, et al., 2012). This long preclinical period provides an opportunity for early detection and prevention. Biomarkers for AD are therefore widely studied as indicators of disease pathology, or as predictors for the progression from normal cognition to MCI due to AD, and to AD dementia. Currently, CSF concentrations of total tau (t-tau), phosphorylated tau (p-tau$_{181p}$), and β-amyloid 1-42 (Aβ_{42}) are attracting the most attention, because they are

particularly relevant to tracking the pathological onset and preclinical stages of AD; they may also, ultimately, serve as intermediate endpoints for clinical trials of disease modifying therapies.

The great time lag between the initiation of AD pathology and the onset of clinical symptoms makes the information from biomarkers potentially valuable, but it also makes demonstrating or evaluating these markers' utilities for AD prediction difficult. Limited data on autopsies, multiple coexisting pathologies at the time of death, and errors in clinical AD diagnoses cause further difficulty with, and potential biases in, assessing the sensitivities and specificities of the antemortem markers. Nevertheless, most current studies ignore these issues and use subjects' last available clinical diagnoses as the gold standard to investigate biomarkers. An exception is the study conducted by De Meyer, et al. (2010), in which the investigators used a mixture modeling approach and assumed that the biomarker values (or the log transformed values for t-tau and p-tau) follow a two component mixture normal distribution without using any clinical information. However, the normality assumption can be too restrictive in practice. In addition, their analyses only used biomarker data. Many factors that are known to be associated with AD prevalence are not considered in their model, even though such information can be used to construct a better model. Meanwhile, it is of interest to examine the impact of risk factors on the prevalence of AD-related pathology. Likewise, biomarker levels can be associated with subjects' characteristics. Understanding these relationships can help clinicians to better utilize biomarker information to monitor and predict AD pathology. Additionally, although the levels of CSF Aβ_{42}, t-tau and p-tau$_{181p}$ are all related to AD pathology, they are believed to reflect different processes (Storandt, et al. 2012). Therefore, combining biomarkers may yield higher diagnostic power. My analyses addresses these questions.

5.4.2 The ADNI Data

The ADNI study is a multicenter longitudinal study launched in 2004. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological

assessments can be combined to measure the progression of MCI and early AD. The determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, and to lessen the time and cost of clinical trials. ANDI has three phases, ANDI1, ANDI GO and ADNI2. The initial enrollment target of ADNI (ADNI1) was to recruit 800 subjects from 59 centers in the United States and Canada: 200 subjects with normal cognition (NC), 400 subjects with MCI and 200 subjects with early AD. It was later extended to ADNI GO and ADNI2 to continue follow-up and to recruit new subjects. Detailed information about the study's design and inclusion and exclusion criteria can be found at http://www.adni-info.org.

Briefly, all participants were recruited between the ages of 55 and 90 years and had at least 6 years of education. Screening criteria for entry into the study included the Mini-Mental State Examination score, the Clinical Dementia Rating scale, and an education-adjusted cutoff score on delayed recall. NC subjects had MMSE scores between 24-30 (inclusive), a CDR of 0, absence of depression, MCI, and dementia. MCI subjects had MMSE scores between 24-30 (inclusive), a memory complaint, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains and dementia. Early AD subjects have MMSE scores between 20-26 (inclusive), a CDR of 0.5 or 1.0, and meet NINCDS/ADRDA criteria for probable AD. Participants who took specific psychoactive medications, or who had other neurological disorders, were excluded. After the baseline visit, subsequent visits occurred at 6- or 12-month intervals. Participants with NC or MCI were followed up for 3 years, whereas those with AD were followed up for 2 years at maximum. More detailed information on the study population can be found in Hansson, et al. (2006) and Engelborghs, et al. (2008).

CSF biomarkers, Aβ_{42}, t-tau, and p-tau$_{181p}$, were measured at baseline using the multiplex xMAP Luminex platform (Luminex Corp, Austin, TX) with Innogenetics (INNO-BIA AlzBio3; Ghent, Belgium; for research use only reagents) immunoassay kit-based reagents. More details about CSF biomarker collection and measurement can be found in Shaw, et al. (2009).

5.4.3 Descriptive Results

I used a cross sectional dataset at baseline, which included all currently available ADNI subjects (ADNI1, ADNIGO and ADNI2) as of June 1st, 2013. There are a total of 805 subjects (221 CN, 455 MCI and 127 early AD) with at least one CSF biomarker measurement at baseline. Subjects' demographics, cognitive assessment and biomarker measurements, stratified by baseline clinical diagnosis are summarized in Table 5.4.

Table 5.4: Baseline characteristics (mean and standard deviation (in parentheses) for continuous variables, counts and percentage (in parentheses) for categorical variables)

		Baseline clinical diagnosis			
Characteristics		CN (N=221)	MCI (N=455)	AD (N=127)	Total (N=805)
Gender	Female	106 (47.5%)	184 (40.3%)	52 (40.3%)	342 (42.5%)
	Male	115 (51.6%)	271 (59.3%)	75 (58.1%)	461 (57.3%)
Age (years)		75.10 (±5.5)	72.75 (±7.6)	75.04 (±8.5)	73.76 (±7.3)
Race	White	206 (92.4%)	422 (92.3%)	126 (97.7%)	754 (93.7%)
	Black	13 (5.8%)	13 (2.8%)	1 (0.8%)	27 (3.4%)
	Others	2 (0.9%)	20 (4.4%)	0 (0.0%)	22 (2.7%)
Education (years)		16.12 (±2.8)	15.91 (±2.8)	15.28 (±3.2)	15.87 (±2.9)
MMSE		29.05 (±1.1)	27.67 (±1.8)	23.38 (±1.9)	27.37 (±2.5)
CDR Sum of Boxes		0.03 (±0.1)	1.43 (±0.9)	4.34 (±1.5)	1.51 (±1.6)
Aβ_{42}		218.76 (±64.7)	194.41 (±71.7)	145.17 (±42.8)	193.21 (±70)
T-tau		71.28 (±31.9)	94.03 (±57.9)	124.26 (±57.6)	92.42 (±54.6)
P-tau$_{181p}$		23.16 (±12.0)	29.38 (±16.3)	39.94 (±18.9)	29.33 (±16.6)
Hippocampus ($\times 10^{-3}$ mm^3)		7.32 (±0.8)	6.85 (±1.1)	5.74 (±1.1)	6.83 (±1.2)
Ventricles ($\times 10^{-3}$ mm^3)		33.32 (±17.2)	39.39 (±23.4)	48.27 (±24.3)	38.91 (±22.4)
Whole Brain ($\times 10^{-4}$ mm^3)		102.33 (±10.4)	104.41 (±11.2)	96.89 (±12.3)	102.74 (±11.4)
ApoE4	0 copy	170 (76.2%)	237 (51.9%)	41 (31.8%)	448 (55.7%)
	1 copy	45 (20.2%)	176 (38.5%)	59 (45.7%)	280 (34.8%)
	2 copies	6 (2.7%)	41 (9.0%)	27 (20.9%)	74 (9.2%)

Most of the subjects were white (93.4%). The percentage of male subjects (57.3%) was slightly higher than that of female subjects (42.5%), especially for MCI and AD groups.

The average age and years of education were similar for all three groups, partially due to the study's inclusion criteria. As expected, the AD group had lower MMSE scores and higher CDR sum of boxes scores. Moreover, their biomarker levels were consistent with current belief that, as subjects progress from NC to MCI and then to AD, their Aβ_{42} level decreases, t-tau and p-tau$_{181p}$ levels increase, the hippocampus shrinks and the ventricle area enlarges. The average whole brain volume is similar for subjects in all three groups. Additionally, most subjects (76.2%) in the CN group did not carry an ApoE 4 allele. The percentage of subjects with one copy of ApoE 4 allele was higher in MCI group than in CN group, and even higher in AD group. This trend was much stronger for the percentage of subjects with two copies of ApoE 4 alleles.

5.4.4 Analysis Results

I applied the proposed profile model to compare and combine CSF biomarkers for detecting preclinical AD pathology. There were two disease groups ($D = 0$ or 1) for this problem, defined by the presence or absence of an AD related pathology signature in the subjects' brains. Covariates in the latent structure model were age (continuous, in years), ApoE 4 (binary, indicating the presence or absence of any ApoE 4 alleles) and education (continuous, in years). Because covariates in the measurement model were age and gender I also included the interaction term between age and D. Thus, age could have different impacts among subjects with or without an AD pathology signature. There were a total of 789 subjects with complete biomarker and covariate information included in the analysis. Results are shown in Table 5.5, with bold indicating a significant effect.

These results suggest that the log odds of having an AD pathology signature is about 0.07 (95% CI: 0.04, 0.1) higher for a 1-year increase in age, adjusting for ApoE 4 and education. Subjects with ApoE 4 alleles have about exp(2.55)=12.8 fold higher odds of developing an AD pathology signature (95% CI: 8.76, 20.1), adjusting for age and education. Additionally, education seems to have a protective effect against AD pathology, but this effect was not significant based on these data. On the biomarker side, I found that subjects with AD pathology have lower levels of A*beta* and higher levels of t-tau and p-tau$_{181p}$. All of these

Table 5.5: Estimates and 95% CI in parentheses for CSF Aβ_{42}, t-tau and p-tau$_{181p}$ (Estimates in bold indicate a significant effect).

Structural (prevalence) model $P(D\|Z)$				
	(Intercept)	Age	ApoE4	Education
$D = 1$	**-5.49** (-8.07, -3.33)	**0.07** (0.04, 0.10)	**2.55** (2.17, 3.00)	-0.03 (-0.10, 0.03)
Measurement model $P(T\|D, X)$				
		$A\beta_{42}$	T-tau	P-tau$_{181p}$
(Intercept)		**12.37** (7.38, 20.58)	**2.54** (2.17, 2.99)	**1.91** (1.66, 2.23)
D1		**-2.77** (-6.63, -1.06)	**0.90** (0.52, 1.56)	**0.74** (0.45, 1.13)
Age(×10)		-0.09 (-0.31, 0.03)	**0.06** (0.03, 0.11)	**0.02** (0.00, 0.05)
Gender=Male		-0.11 (-0.31, 0.04)	**-0.03** (-0.08, -0.01)	0.003 (-0.02, 0.03)
D1×Age(×10)		0.11 (-0.07, 0.36)	**-0.08** (-0.14, -0.04)	**-0.05** (-0.01, -0.02)
λ		0.25 (0.09, 0.39)	-0.17 (-0.26, -0.06)	-0.24 (-0.34, -0.14)
σ		0.88 (0.38, 1.82)	0.19 (0.12, 0.29)	0.16 (0.12, 0.22)

associations were statistically significant. Levels of t-tau and p-tau$_{181p}$ were significantly higher among older subjects. However, age did not seem to have a significant association with Aβ. This age effect was smaller among subjects with AD pathology than among subjects without. Moreover, males seemed to have lower levels for all three biomarkers, but only the one with t-tau was marginally significant. In addition, the estimates for λ were about 0.25 for Aβ, -0.17 and -0.24 for t-tau and p-tau$_{181p}$, which challenges the conventional wisdom that Aβ is normal, and that tau measures are log normal.

Hippocampus volume and other brain region volumes are suspected to represent a different process in AD pathology than the ones represented by CSF biomarkers. Therefore, I performed additional an analysis with the ratio of hippocampus to whole brain volume as the fourth biomarker. The ratio was used to adjust for variations in brain size among subjects. This analysis was performed in a subset of 642 subjects with hippocampus and whole brain volume measures. The results are shown in Table 5.6.

These results suggest that, although both the hippocampus and the whole brain areas shrink with the progression of AD pathology, the hippocampus shrinks faster than the whole

Table 5.6: Estimates and 95% CI in parentheses for CSF Aβ_{42}, t-tau, p-tau$_{181p}$ and the ratio of hippocampus to whole brain volume (Estimates in bold indicate a significant effect).

	Structural (prevalence) model $P(D\|Z)$			
	(Intercept)	Age	ApoE4	Education
$D=1$	**-5.56** (-8.19, -2.85)	**0.07** (0.04, 0.10)	**2.54** (2.12, 3.00)	-0.05 (-0.12, 0.02)
	Measurement model $P(T\|D,X)$			
	$A\beta_{42}$	T-tau	P-tau$_{181p}$	Hippo/WB*
(Intercept)	**12.44** (7.12, 21.77)	**2.41** (1.99, 2.91)	**1.89** (1.61, 2.25)	**-0.08** (-0.15, -0.01)
D1	**-2.90** (-7.38, -1.02)	**0.71** (0.35, 1.3)	**0.76** (0.43, 1.22)	**-0.14** (-0.25, -0.02)
Age(×10)	-0.09 (-0.3, 0.06)	**0.05** (0.02, 0.09)	**0.02** (0.00, 0.04)	**-0.03** (-0.04, -0.01)
Gender=Male	-0.13 (-0.4, 0.02)	-0.02 (-0.06, 0.00)	0.02 (-0.01, 0.04)	**-0.02** (-0.04, -0.01)
D1×Age(×10)	0.12 (-0.06, 0.49)	**-0.06** (-0.12, -0.02)	**-0.06** (-0.1, -0.02)	0.01 (-0.00, 0.03)
λ	0.25 (0.09, 0.40)	-0.21 (-0.33, -0.09)	-0.25 (-0.36, -0.13)	1.47 (1.06, 1.87)
σ	0.9 (0.38, 1.91)	0.16 (0.09, 0.25)	0.16 (0.11, 0.23)	0.07 (0.06, 0.08)

* Volume ratio: Hippocampus/Whole Brain ($\times 10^2$)

brain. Additionally, the hippocampus/whole brain ratio is lower in older subjects and in males. All of these associations were statistically significant. All other results were similar as before, except that the association between male and t-tau was no longer significant.

Based on the model, one can compute the covariate-specific ROC curves. As an example, I plot the age-specific ROC curves for CSF biomarker Aβ_{42}, t-tau and p-tau$_{181p}$ among age groups 65 years or younger, 65-80 years and 80 years or older in Figure 5.4. From this it is evident that the ROC curves differ greatly among age groups for both t-tau and p-tau$_{181p}$, with higher AUCs for younger groups. This result also suggests that including covariates in the model and allowing for the covariate effect to vary among disease groups is very important. On the other hand, the age-specific ROC curves for Aβ do not vary much among these age groups. One explanation is that Aβ changes much earlier than tau in the pathology; it is possible that, at this stage it is already relatively stabilized.

Other than looking at each biomarker separately, one may wish to combine information from multiple biomarkers to achieve better diagnostic accuracy. However, due to the

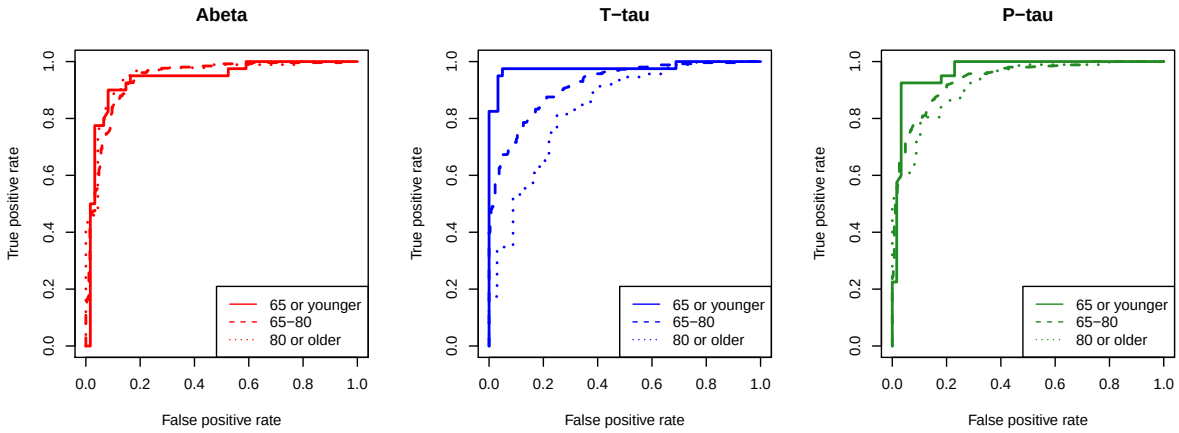

Figure 5.4: Age-specific ROC curves for CSF Aβ_{42}, t-tau and p-tau$_{181p}$ in detecting AD pathology

unobserved AD pathology, combinations of CSF biomarkers have not been well examined. The model I used provides one way of combining different markers. Specifically, the model based on posterior risk $P(D_i|\vec{T}_i, \vec{X}_i, \vec{Z}_i, \theta)$ can be used as a score for diagnosing a subject as having started AD pathology or not. An ROC curve comparing a model based score with all three single biomarkers is shown in Figure 5.5. Because the true AD pathology status is not available, this ROC curve uses clinical diagnosis as the gold standard. The graph suggests an advantage of using model based risk to combine biomarkers. The AUC is about 0.69 for the model based combined score, and about 0.65, 0.65 and 0.66 for Aβ_{42}, t-tau and p-tau$_{181p}$, respectively. This advantage is more visible at the lower left-hand corner of the graph, where the false positive rate is low and the true positive rate is much higher when combining the biomarkers. Keep in mind that this plot uses clinical diagnosis as the gold standard; some subjects with preclinical AD may appear to be normal. If the subjects are followed long enough to have better clinical assessment or even autopsy, it is possible that the advantage of combining these biomarkers will be even bigger.

With the model-based posterior risks, I examined the proportions of subjects that have AD pathology signatures in groups defined by clinical diagnosis. The results are shown in Table 5.7. I assign a subject into the group with AD pathology if the posterior risk of AD

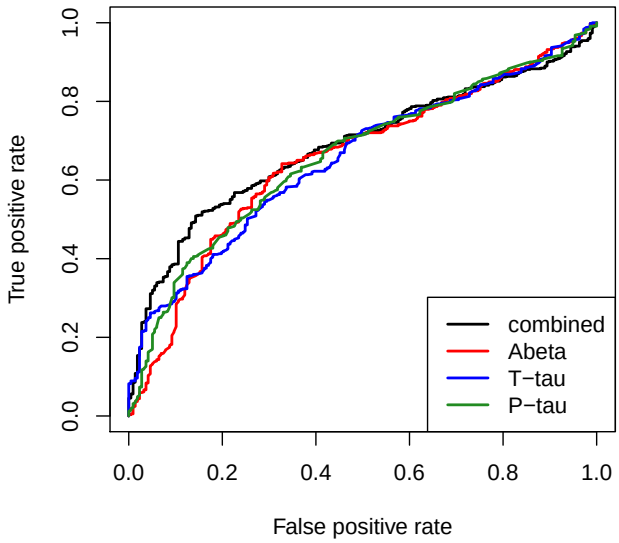

Figure 5.5: ROC curves for model based combined score and single biomarkers using clinical diagnosis as the gold standard

pathology present is higher than that of AD pathology absent for this subject.

It is evident that the proportion of subjects with an AD signature increased from about 31% in the early MCI group to 86% in the AD group, representing a progression. Additionally, there were about 26% of subjects in the CN group that had an AD pathology signature according to the model. This portion likely indicates that the subjects exhibit preclinical AD pathological changes but still appear to have normal cognition.

Another advantage of using a model-based risk as a score for preclinical AD diagnosis is that this score is well-separated in the population, as shown in the histogram on the top left panel in Figure 5.6. As shown in this plot, this score puts most of the subjects either in the very low risk end or in the very high risk end of the distribution. Only 9% of the subjects had an estimated risk between 0.2 and 0.8. This means that the model based-risk had good stratification ability the strata are well separated between subjects with and without AD pathology. By contrast, the histograms for single biomarkers put most subjects in the

Table 5.7: Numbers and proportions of subjects that have an AD pathology signature in groups defined by clinical diagnosis.

Clinical diagnosis	AD signature present	AD signature absent
CN	57 (26.3%)	160 (73.7%)
EMCI[1]	58 (30.7%)	131 (69.3%)
LMCI[2]	168 (64.9%)	91 (35.1%)
AD	107 (86.3%)	17 (13.7%)

1/2: Early/Late MCI–meet ADNI MCI criteria, objective memory loss measured by education adjusted scores on delayed recall of one paragraph from Wechsler Memory Scale Logical Memory II ($\geq$16 years: 9-11; 8-15 years: 5-9; 0-7 years: 3-6) for EMCI, ($\geq$16 years:$\leq$8; 8-15 years: $\leq$4; 0-7 years: $\leq$2) for LMCI.

middle range. Consequently, if only a single cut-off point is used to classify subjects, as in most current research, a high false positive rate or a high false negative rate may occur. If one chooses two cut-off points, a big portion of the subjects will fall in-between them and become unclassified.

5.4.5 Conclusions

This example illustrates the use of latent profile models for assessing and combining biomarkers without using any clinical information. The approach described in this chapter can be helpful for detecting possible AD related pathological changes in the preclinical stage, and therefore enable early intervention or serve as an alternative endpoint for clinical studies.

This approach can also be used to study risk factors for the unobserved AD pathology. The results suggest a strong association between AD pathology with age and the ApoE 4 allele, consistent with previous findings. These results do not use any clinical or autopsy information, and thus can be considered as a separate validation of previous results. Moreover, they are less affected by errors in clinical diagnoses, and do not require a long follow-up period or autopsy results.

All three CSF biomarkers, Aβ_{42}, t-tau and p-tau$_{181p}$, are significantly associated with an

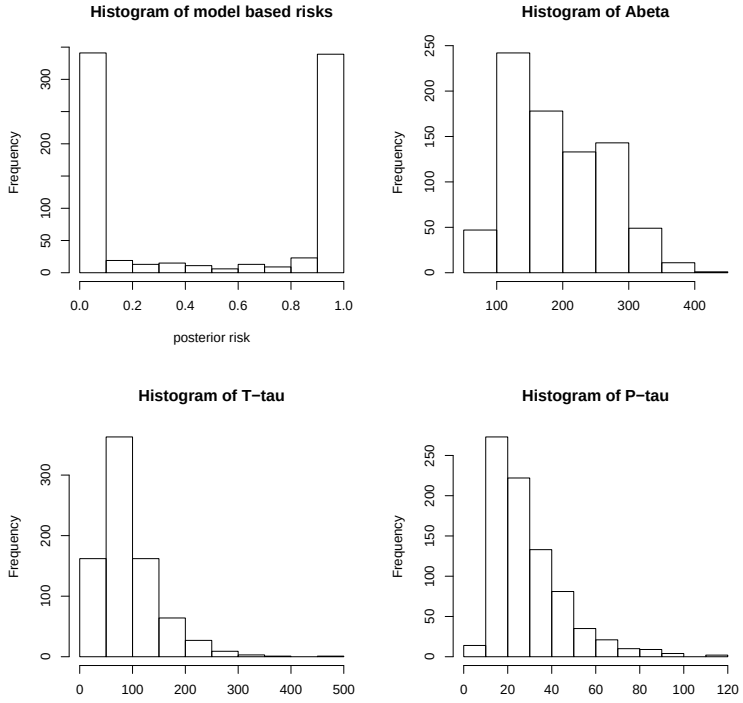

Figure 5.6: Histograms of model based combined score and single biomarkers

AD pathology signature. Thus, they can be used for detecting preclinical AD subjects. However, the biomarkers' levels, and therefore the diagnostic performance, vary with subjects' characteristics. This information needs to be taken into account when using biomarkers to make diagnostic conclusions. Specifically, my analyses suggests that age has a significant positive association with t-tau and p-tau$_{181p}$ levels and that the association between AD pathology is smaller among older subjects. Consequently, the diagnostic performance for both tau measures is different among different age groups, with better performance in the younger groups. On the other hand, the level of CSF Aβ does not seem to vary with age. It is possible that the effect of age on CSF Aβ happens early in the pathology process (Jack, et al. 2010) and is no longer visible in the ADNI cohort. The brain hippocampus volume, adjusted by whole brain volume, is also predictive for AD pathology.

A result that was unexpected but that cannot be ignored in the analysis relates to the

transformation parameters for the biomarkers. The conventional assumption in many AD biomarker studies is that CSF Aβ is normal, CSF t-tau and p-tau$_{181p}$ are log normal within disease groups, but this assumption may not hold and needs to be further investigated.

The model also provides a means to combining biomarkers and can achieve better diagnostic performance. Moreover, the combined score based on my model exhibits great bimodel separation in the population. Therefore, fewer subjects lie in the "gray area" with undetermined disease status compared with the situation in which a single biomarker is used.

The proportions of subjects with an AD signature based on this model increases in groups with progressing clinical diagnosis (CN, EMCI, LMCI and AD), which partially validates this method. The subjects in the CN group and classified as an AD signature present by the model are especially important, as they may represent the subjects with preclinical AD. Compared with earlier results by De Mayer, et al. (2010), the proportion of having an AD signature is similar in the clinical AD group, but is smaller (26% versus 39%) in the CN group. This is possibly because the proposed model includes covariates, and so accounts for the variation of biomarker level due to, for example, a normal aging process.

My analyses was performed on cross-sectional data. Therefore, although they provided a promising way to evaluate and combine AD biomarkers, the results need to be validated with longitudinal follow-up and autopsy data. Another limitation is that the ADNI study enrolls subjects with strict criteria, so the data are more ideal than those used in practice, as subjects with comorbidity are excluded. The biomarkers' diagnostic performances need further investigation in a population-based study.

5.5 Summary and Discussion

In this chapter I propose a latent profile model for assessing the accuracy of continuous diagnostic tests without a gold standard. Compared with current available methods, this model has several advantages. First, this model can include covariates in both the latent structure model and the measurement model. The former can help to examine the impact of risk factors on constructs that are difficult to measure, such as the pathological process in a preclinical AD brain. The latter can be helpful in understanding the tests' performances in

subgroups of patients with different characteristics. This is especially valuable for obtaining a more personalized diagnostic procedure and for achieving higher diagnostic accuracy for each individual subject. Second, the model for covariate effect is much more flexible – there is no constraint on covariate type (i.e., categorical or continuous); additionally, covariates can have different impacts on different tests and within different disease severity groups. Third, the transformation on the test results accounts for possible skewness in the data. Last, the method can deal with ordinal disease status. The computation of the proposed method has been written in the R package "lpreg" (stands for "latent profile regression") so that it can be easily accessed by other researches. The package is currently in the final modification and testing stage and will be online soon.

In addition to diagnostic test and biomarker evaluation, there are many areas for these methods to be applied and developed further. For example, they can be used to study unobserved factors in mental/behavioral research (the area where latent variable modeling is most frequently applied), such as in the study of the impacts of alcohol/drug use on HIV progression and transmission, maternal stress on low birth weight, and in adherence. Because latent profile models can summarize multiple measures simultaneously, they can be useful for risk prediction, such as for detecting HHV-8 infection, when only simple immunoassay measures are available (Pfeiffer, et al. 2008). They can also be useful for learning about measurement problems and describing population heterogeneity.

One possible extension to the proposed method is to use a semi-parametric transformation with an unspecified H instead of a Box-Cox transformation. In other words, it can be assumed that the tests are normally distributed after an unknown nonparametric transformation H. This transformation can allow for more general relationships between the test results and covariates. In this case, the transformation H was a nuisance parameter to be estimated. However, since H was nonparametric with infinite dimensions, the maximum likelihood approach did not apply. One possible choice is to use estimating equations in each M step to estimate H, as originally proposed by Cheng, et al. (1995) for transformation regression when the outcome is observed. Briefly, I solved H and β with the following

equations:

$$\sum_i \sum_j \Big\{\omega(X_{ij}\beta_k)X_{ij}\big[I(T_{ik}-T_{jk}\geq 0)-G(-X_{ij}\beta_k)\big]\Big\}=0$$
$$\frac{1}{n}\sum_i \big[I(T_{ik}\leq t)-F(H_{nk}(t)-X_i\beta_k)\big]=0$$
$$F=\Phi, G(z)=\int_{-\infty}^{\infty}\big[1-F(u+z)\big]dF(u), \quad X_{ij}=X_i-X_j,\ i\neq j.$$

In the above, $\omega(\cdot)$ is a weight function. In the paper by Cheng, et al (1995), $\omega(\cdot)=1$ or, or $\omega(\cdot)=\dfrac{G'(\cdot)}{G(\cdot)(1-G(\cdot))}$, which mimics the quasi-likelihood approach for independent observation. Additionally, $G(z)=\int_{-\infty}^{\infty}\big[1-F(u+z)\big]dF(u)=1-\Phi\big(\frac{z}{\sqrt{2}}\big)=\Phi\big(-\frac{z}{\sqrt{2}}\big)$, since $\int_{-\infty}^{+\infty}\Phi(a+bx)d\Phi(x)=\Phi\big(\frac{a}{\sqrt{1+b^2}}\big)$.

However, the nonparametric estimators are obtained by a different criterion than maximizing the likelihood. Therefore, the resulting estimates no longer guarantee the non-decreasing property in a typical EM algorithm, which is a condition used to prove that the EM algorithm converges to a stationary point of the likelihood function (Wu, 1983). Consequently, it is unclear whether the algorithm can result in a consistent estimate, or even whether it converges. However, this type of algorithm (sometimes referred to as an EM-like algorithm) is used in practice and appears to lead to reasonable results (Benaglia, et al. 2009). Levine, et al. (2011) proved that when the likelihood is replaced by its smoothed version, a majorization-minimization idea can retain the desired non-decreasing property of an EM algorithm, and therefore guarantee its convergence. Additionally, they showed that the resulting estimates minimize a penalized Kullback-Leibler divergence between the targeted distribution and the smoothed likelihood. However, the consistency of the estimators from this type of algorithm has not been proven.

The flexibility of the model that Ive proposed for including covariate effects, and for allowing a transformation on the test distribution, may raise concerns about identifiability. This will be discussed in the next chapter, where I show that the nonparametric identifiability for the proposed models, which applies even if an unknown nonparametric transformation replaces the Box-Cox transformation in the measurement model.

The proposed model assumes the test results are conditionally independent given the

true disease status and a set of defined covariates. Essentially, this model uses covariates to explicitly explain the correlation among the tests that was previously explained by a random effect in Chapter 4. Additional random effects can still be introduced if one suspects that there are remaining correlations among the tests due to other unobserved covariates. However, one should be cautious when building more complicated model structures that are hard to validate. Additionally, the identifiability issue for such models must be resolved before adopting them in practice.

Chapter 6

THE INDENTIFIABILITY ISSUE IN LATENT CLASS AND LATENT PROFILE MODELS

Latent class and latent profile models both assume that the latent variable is categorical. They are also referred to as finite mixture models because the likelihoods of these models always have a mixture form. In this section I discuss the identifiability issue of these models and use the term "finite mixture models" to refer to both of types, collectively.

6.1 Introduction

Finite models provide a flexible approach to modeling unobserved constructs and to explaining population heterogeneity. These models can be specifically tailored to meet the needs of the researchers using them, and are thus gaining popularity in many disciplines. In addition to their applications to medical diagnosis, as discussed in previous chapters, finite mixture models are also actively applied in psychometrics, the social sciences, economics and other fields whose respective research problems involve entities that are hard to measure directly. For example, researchers are sometimes interested in examining behaviors that connect to various health issues, describing personal, familial and social networks that influence one's decision, or understanding factors that relate to consumer patterns and preferences. Finite mixture models offer a means to study these unobserved entities.

The flexibility of finite mixture models makes them a powerful tool, but it also renders them vulnerable to problems associated with identifiability. Loosely speaking, a non-identifiable model can have multiple parameter values that all correspond to the same likelihood, which means that one cannot identify the true parameter values among these choices based only on the observed data. Although identifiability is not the same as consistency, estimates from a model lacking identifiability will not be consistent. Thus, establishing identifiability is a crucial issue for establishing the validity of the study when using a finite mixture approach, and for the interpretation of its results. However, while finite mixture

models arise in many statistical applications, conditions for their identifiability have not been established for the general case.

6.1.1 "Label Switching" and Local Maxima

Theoretically, finite mixture models are not strictly identifiable, as they suffer from a "label switching" problem–that is, the distribution remains identical if the labels of the mixture components are switched. In a sense, finite mixture models only help researchers to "group" together subjects that are similar; they cannot further "label" the groups. For this reason, I previously assumed that higher test results correspond to more severe conditions. Practically speaking, this "label switching" problem of the finite mixture models is not an issue of great importance, because it is usually easy to correctly associate each mixture component with its label. For example, when a finite mixture approach is used in diagnostic medicine to model the diagnostic test results from subjects with or without a certain medical condition, it is usually apparent in the results which group is the diseased group and which is the healthy one.

In addition to the "label switching" problem, another difficulty in proving the identifiability of the finite mixture model is that the likelihood functions of such models usually have multiple local maxima. In other words, one cannot show that the likelihood is a bijection–a common approach to establish model identification.

6.1.2 Previous Results

Due to the difficulties caused by "label switching" and multiple local maxima, research on the identifiability issue of finite mixture models has mainly focused on local identifiability (McHugh 1956; Goodman 1974; Formann 1992). As discussed in Chapter 2, the commonly used criterion that the degrees of freedom in the data must meet or exceed the number of parameters in the model is not enough to guarantee model identifiability (Goodman, 1974; Elmore et al. 2005). McHugh (1956) proposed sufficient conditions for local identifiability of models with dichotomous observed variables. Goodman (1974) extended these conditions to polytomous variables, and other work has focused on finding sufficient conditions to

guarantee that the Jacobian matrix of the mixture distribution has full column rank (Huang and Bandeen-Roche, 2004; Kasahara and Shimotsu, 2008). However, conditions for ensuring identifiability for more general models, such as models that allow for covariate effects or continuous test results have not been established.

Moreover, although local identifiability is necessary for the implementation of the model and for establishing the validity of the asymptotic approximation, it only addresses the problem at an infinitely small neighborhood in the entire parameter space. No conclusion can be made about the same model with a different set of parameters. The identifiability of the model for one researcher cannot be carried on to another researcher who wants to repeat the experiment. Local identifiability looks at model identification at a given point of parameters. In a sense, it is more specific to a given problem in which the true parameters are fixed (even though they are unknown) than to an unrestricted model itself. Additionally, since the true parameters are unknown, current methods evaluate local identifiability using the estimated parameter values. This means that identifiability cannot be assessed until after data collection has occurred. However, in many cases the ability to establish the identifiability before collecting the data is necessary or desirable, as the study may only be meaningful when the model is identifiable.

In the special application of finite mixture models to diagnostic tests, more work on the identifiability issue has been accomplished. Hall and Zhou (2003) and Hall, et al. (2005) studied the nonparametric identifiability of models of K tests with conditional independence assumptions within each of the M subpopulations, and provided sufficient conditions for model identification of $K \geq (1+o(1))6M \log M$. For the M-component mixture of binomial distributions, $B(K, p_m)$, it has been shown that $K \geq 2M-1$ is both necessary and sufficient for model identification (Teicher, 1961, 1963; Blischke 1964). Based on this result, with an additional assumption of identically distributed tests, Hettmansperger and Thomas (2000) and CruzMedina et al (2004) dichotomized the test results and concluded that $K \geq 2M-1$ is sufficient for the identification of such models. However, conditions for more general models have not yet been established.

6.2 Revisiting the Finite Mixture models

A finite mixture model is characterized by its component distributions, its number of components and its mixing proportion. Consider a model with L components with label $d = 0, \ldots, L-1$ (such as disease severity), and K manifest variables $\vec{T} = (T_1, \ldots, T_K)$ (such as diagnostic tests). Assuming conditional independence among manifest variables T_k within a given component, the model can be expressed as follows:

$$P(\vec{T}) = \sum_{d=0}^{L-1} P(\vec{T}, D) = \sum_{d=0}^{L-1} \big[P(D = d) \prod_{k=1}^{K} P(T_k|D = d)\big],$$

where $P(.)$ is the marginal distribution function of manifest variable $\vec{T}$, $P(D = d)$ is the mixture proportion of the dth subpopulation, and $P(T_k|D = d)$ is the conditional probability of T_k in the dth subpopulation, either as a conditional density function $f_{T_k|D}(.)$ for continuous T_k, or as a probability mass function for categorical T_k. For example, when all of the manifest variables are categorical and take on the values $\{1, \ldots, J_k\}$, $k = 1, \ldots, K$, the model is a latent class model as follows,

$$P(T_1 = t_1, \ldots, T_K = t_K) = \sum_{d=0}^{L-1} \big\{P(D = d) \prod_{k=1}^{K} \prod_{j=1}^{J_k} \big[P(T_k = j|D = d)\big]^{I[t_k=j]}\big\}, \qquad (6.1)$$

where $I[t_k = j]$ is an indicator, which equals 1 if $t_k = j$ and 0 otherwise.

Model (6.1) is a finite mixture model without covariates. When covariates are involved, the extended model can be expressed as

$$P(\vec{T}|\vec{X}, \vec{Z}) = \sum_{d=0}^{L-1} P(\vec{T}, D|\vec{X}, \vec{Z}) = \sum_{d=0}^{L-1} \big[P(D = d|\vec{Z}) \prod_{k=1}^{K} P(T_k|D = d, \vec{X})\big],$$

where $\vec{Z}$ is the set of covariates associated with mixture membership and $\vec{Z}$ is the set of covariates associated with manifest variables within each group. The two sets of covariates may be mutually exclusive or overlapping, and they can include continuous and categorical variables.

Similarly as before, $P(T_k|D = d, \vec{X})$ can be the conditional density function or probability mass function. In the latter case, the model is,

$$P(T_1 = t_1, \ldots, T_K = t_K|\vec{X}, \vec{Z}) = \sum_{d=0}^{L-1} \big\{P(D = d|\vec{Z}) \prod_{k=1}^{K} \prod_{j=1}^{J_k} \big[P(T_k = j|D = d, \vec{X})\big]^{I[t_k=j]}\big\} \qquad (6.2)$$

6.3 Local Identifiability

Local identifiability is important for the implementation of the model and the validity of its asymptotic approximation. Although, as mentioned earlier, it only address the identifiability of the model at a given point and thus may not be adequate in some situations; I nevertheless discuss local identifiability here.

I argue here that it is sufficient to consider the identifiability issues for models with discrete manifest variables, as is the case in model (6.1) and model (6.2). This is because local identifiability essentially considers model identification at a given data set. When some or all of the manifest variables are continuous, their distributions can be modeled empirically on the observed data points, which are discrete. In other words, the same techniques used to establish the local identifiability of a latent class model can also be used to establish the nonparametric local identifiability of a finite mixture model with some or all of its manifest variables being continuous. The parametric assumption for the conditional distribution of a manifest variable in a latent profile model, if any, can be viewed as an additional constraint in the estimation procedure.

6.3.1 Definition

By definition, a function F is locally identifiable at parameter $\theta_0 \in \Theta$ if there exists some neighborhood U_{θ_0} of θ, such that

$$F(\theta) \neq F(\theta_0) \qquad \forall\ \theta \in U_{\theta_0} \setminus \{\theta_0\}.$$

This suggests that F is a one-to-one map, or locally invertable in U_{θ_0}.

As argued before, without loss of generality, I consider a situation in which all of the manifest variables are categorical. Let $\vec{t_h} = (t_{h1}, \ldots, t_{hK})$ be the hth possible in lexicographic order among $(\prod_{k=1}^{K} J_k) - 1$ distinct response patterns of the manifest variables, excluding a reference pattern. I stacked the probability $P(\vec{T} = \vec{t_h})$ into a vector p of length $(\prod_{k=1}^{K} J_k) - 1$. A given model specifies a function F, which determines how p is calculated from parameters θ,

$$p = F(\theta).$$

The model is locally identifiable at θ_0 if F is invertible in a neighborhood of θ_0. When the number of parameters is less than $(\prod_{k=1}^{K} J_k) - 1$, F is potentially invertible, and local invertibility at θ_0 can be evaluated by examining the Jacobian matrix of F at θ_0, $J(\theta_0) = \frac{\partial F}{\partial \theta}\Big|_{\theta=\theta_0}$. By the weak inversion theorem, if $J(\theta_0)$ has full column rank, F is locally invertible at θ_0, and thus the model was locally identifiable at θ_0.

6.3.2 Models without Covariates

In this section I consider local identifiability for models without covariates, using model (6.1). The basic idea is to find sufficient conditions that guarantee the local invertibility of the Jacobian matrix of the model.

Conditions for Local Identifiability

Let $\pi_d = P(D = d)$, $g_{kjd} = P(T_k = j|D = d)$, Ψ_d be a vector of length $(\prod_{k=1}^{K} J_k) - 1$, $d = 0, \ldots, L-1$, with the hth element

$$\psi_{dh} = P(\vec{T} = \vec{t}_h | D = d) = \prod_{k=1}^{K} g_{kt_{hk}d}.$$

Further, let $\eta_d = \Psi_d - \Psi_0, \quad d = 1, \ldots, L-1$, and Γ_{kjd} be a vector of length $(\prod_{k=1}^{K} J_k) - 1$, $d = 0, \ldots, L-1$, $k = 1, \ldots K$ and $j = 1, \ldots, J_k - 1$, with the hth element

$$\gamma_{kjdh} = \pi_d \psi_{hd} \Big[\frac{I(t_{hk} = j)}{g_{kjd}} - \frac{I(t_{hk} = J_k)(g_{kjd} - \sum_{j=1}^{J_k - 1} g_{kjd})}{g_{kJ_k d}} \Big],$$

where $I(t_{hk} = j)$ is an indicator function that equals 1 if $t_{hk} = j$ and 0 if otherwise. Then, I claim the following theorem:

Theorem 1:

The finite mixture model (6.1) is locally identifiable at parameter $\theta = \{\pi_d, g_{kjd} \mid d = 0, \ldots, L-1; k = 1, \ldots K; j = 1, \ldots, J_k\}$ if the following conditions hold.

(i) $(\prod_{k=1}^{K} J_k) - 1 \geq L \times \sum_{k=1}^{K} (J_k - 1) + L - 1$;

(ii) $P(\vec{T} = \vec{t}_h) = \sum_{d=0}^{L-1} \pi_d \psi_{dh} > 0, \quad \forall h$; and

(iii) vectors $\{\eta_d | d = 1, \ldots, L-1\}$, $\{\Gamma_{kjd} | d = 0, \ldots, L-1; k = 1, \ldots K; j = 1, \ldots, J_k\}$ are linearly independent.

Proof:

Condition (i) requires that the degrees of freedom in the data are greater than the number of parameters. Moreover, $(\prod_{k=1}^{K} J_k) - 1$ is the number of rows of the Jacobian matrix of model (6.1), and $L \times \sum_{k=1}^{K}(J_k - 1) + L - 1$ is the number of columns of the Jacobian matrix. When condition (i) is satisfied, the Jacobian matrix can potentially have full column rank. Condition (ii) is included to ensure that the probability of observing every response pattern is positive. It is the third condition in Theorem 1 of McHugh (1956) that pertains to the local identifiability of latent class models with binary manifest variables. Here, I only need to prove that Condition (iii) is equivalent to requiring that the Jacobian matrix of model (6.1) has full column rank.

Based on model (6.1), the function between parameter θ and all possible response patterns of the manifest variables $\vec{t}_h$ (excluding a reference pattern) can be expressed as

$$F(\vec{t}_h;\theta) = P_\theta(\vec{T} = \vec{t}_h) = \sum_{d=0}^{L-1} \pi_d \prod_{k=1}^{K} g_{kt_{hk}d}, \quad h = 1, \ldots, H,$$

where $H = (\prod_{k=1}^{K} J_k) - 1$. The component probabilities sum up to 1. Therefore, $\pi_0 = 1 - \sum_{d=1}^{L-1} \pi_d$, and

$$F(\vec{t}_h;\theta) = \sum_{d=1}^{L-1} \pi_d \prod_{k=1}^{K} g_{kt_{hk}d} + (1 - \sum_{d=1}^{L-1} \pi_d) \prod_{k=1}^{K} g_{kt_{hk}0}, \quad h = 1, \ldots, H.$$

Taking the derivative of F with respect to free component probability parameters π_d, $d = 1, \ldots, L-1$, I have that

$$\frac{\partial F(\vec{t}_h;\theta)}{\partial \pi_d} = \prod_{k=1}^{K} g_{kt_{hk}d} - \prod_{k=1}^{K} g_{kt_{hk}0} = \psi_{dh} - \psi_{0h}.$$

Therefore, the first $L-1$ columns of the Jacobian matrix of model (6.1) are,

$$\frac{\partial F}{\partial \pi_d} = \Psi_d - \Psi_0 = \eta_d, \quad d = 1, \ldots, L-1.$$

Meanwhile, since $\sum_{j=1}^{J_k} g_{kjd} = 1$, the function F can be rewritten as follows,

$$\begin{aligned} & F(\vec{t}_h;\theta) = \sum_{d=0}^{L-1} \pi_d \prod_{k=1}^{K} g_{kt_{hk}d} \\ = & \sum_{d=0}^{L-1} \pi_d \prod_{k=1}^{K} \Big[I(t_{hk} \neq J_k) g_{kt_{hk}d} + I(t_{hk} = J_k)\big(1 - \sum_{j=1}^{J_k - 1} g_{kjd}\big) \Big], \quad h = 1, \ldots, H. \end{aligned}$$

Taking the derivative of F with respect to the free parameter g_{kjd}, $d = 0, \ldots, L-1$, $k = 1, \ldots K$ and $j = 1, \ldots, J_k - 1$, I have

$$\begin{aligned}\frac{\partial F(\vec{t}_h;\theta)}{\partial g_{kjd}} &= \pi_d \psi_{hd}\Big[\frac{I(t_{hk}=j)}{g_{kjd}} - \frac{I(t_{hk}=J_k)(1-\sum_{j=1}^{J_k-1} g_{kjd}+g_{kjd}-1)}{g_{kJ_kd}}\Big] \\ &= \pi_d \psi_{hd}\Big[\frac{I(t_{hk}=j)}{g_{kjd}} - \frac{I(t_{hk}=J_k)(g_{kjd}-\sum_{j=1}^{J_k-1} g_{kjd})}{g_{kJ_kd}}\Big].\end{aligned}$$

Thus, the last $L \times \sum_{k=1}^{K}(J_k - 1)$ columns of the Jacobian matrix of model (6.1) are,

$$\frac{\partial F}{\partial g_{kjd}} = \Gamma_{kjd}, \quad d = 0, \ldots, L-1,\ k = 1, \ldots K,\ j = 1, \ldots, J_k - 1.$$

Therefore, the Jacobian matrix of model (6.1) is a $(\prod_{k=1}^{K} J_k) - 1$ by $L \times \sum_{k=1}^{K}(J_k - 1) + L - 1$ matrix $J(\theta)$ with columns $\big(\eta_1, \ldots, \eta_{L-1}, \Gamma_{111}, \ldots, \Gamma_{KJ_K-1L-1}\big)$. As a result, condition (iii) is equivalent to requiring that the Jacobian matrix of model (6.1) has full column rank, which in turn guarantees that the finite mixture model (6.1) is locally identifiable at parameter $\theta = \{\pi_d, g_{kjd} \mid d = 0, \ldots, L-1; k = 1, \ldots K; j = 1, \ldots, J_k\}$. □

More on Condition (iii)

Conditions (i) and (ii) in Theorem 1 are relatively straight forward to examine, condition (iii) requires more attention. Some efforts have been made to provide equivalent but simplified conditions to condition (iii) with additional model specifications, such as constraining all of the manifest variables to be binary.

Huang and Bandeen-Roche (2004) gave a nice discussion on the identifiability issue of latent class models. They made the first attempt to construct a simplified version of condition (iii) in the general formulation of model (6.1) without further specification. In doing so, they claimed that

> Jacobian matrix of model (6.1) of full rank $\Leftrightarrow$ $\Psi_0, \ldots, \Psi_{L-1}$ were linearly independent.

However, it is unclear why equation (A.4)=0 in their proof deduces $\mathbf{P_1 a_1}^* = \ldots, = \mathbf{P_J a_J}^* = 0$, since equation (A.4) involves element-wise multiplication.

In fact, I believe that having $\Psi_0, \ldots, \Psi_D$ being linearly independent is not sufficient for the Jacobian matrix of model (6.1) to have full rank. For example, consider the situation

where all but one of the manifest variables are totally non-informative about the latent subgroup membership.

Without loss of generality, I assumed that the last manifest variable is the only informative one. Let $\mathbf{g_{Kd}}$ be the vector of probability mass of this last manifest variable in subgroup d,

$$\mathbf{g}_{Kd} = \big(P(T_K = 1|D = d), \ldots, P(T_K = J_K|D = d)\big)^T = \big(g_{K1d}, \ldots, g_{KJ_Kd}\big)^T, \quad d = 0, \ldots, L-1.$$

Suppose that $J_K \geq L$ and vectors $\mathbf{g}_{K0}, \ldots, \mathbf{g}_{K(L-1)}$ are linearly independent. Additionally, suppose that all other manifest variables are uniformly distributed among the subgroups, and are thus non-informative about subgroup membership:

$$P(T_k = j|d = 0) = \ldots = P(T_k = j|d = D), \quad \forall\ k = 1, \ldots, K-1;\ j = 1, \ldots, J_k,$$

or equivalently, $$g_{kj0} = \ldots = g_{kj(L-1)} \equiv \bar{g}_{kj} \neq 0, \quad \forall\ k = 1, \ldots, K-1;\ j = 1, \ldots, J_k.$$

Then the first J_K elements of Ψ_d is

$$\begin{pmatrix} P(\vec{T} = (1, 1, \ldots, 1) \mid D = d) \\ P(\vec{T} = (1, 1, \ldots, 2) \mid D = d) \\ \vdots \\ P(\vec{T} = (1, 1, \ldots, J_K) \mid D = d) \end{pmatrix} = \begin{pmatrix} (\prod_{k=1}^{K-1} \bar{g}_{k1}) g_{K1d} \\ (\prod_{k=1}^{K-1} \bar{g}_{k1}) g_{K2d} \\ \vdots \\ (\prod_{k=1}^{K-1} \bar{g}_{k1}) g_{KJ_Kd} \end{pmatrix} = (\prod_{k=1}^{K-1} \bar{g}_{k1}) \mathbf{g}_{Kd}.$$

Because $\prod_{k=1}^{K-1} \bar{g}_{k1} \neq 0$ and $\mathbf{g}_{K0}, \ldots, \mathbf{g}_{K(L-1)}$ are linearly independent, I have that

$$\text{vectors } (\prod_{k=1}^{K-1} \bar{g}_{k1}) \mathbf{g}_{K0}\ , \ldots,\ (\prod_{k=1}^{K-1} \bar{g}_{k1}) \mathbf{g}_{K(L-1)} \text{ are linearly independent.} \tag{6.3}$$

With more elements appended below each vector in (6.3), the extension groups $\Psi_0, \ldots, \Psi_{(L-1)}$ are linearly independent.

Consequently, if having $\Psi_0, \ldots, \Psi_{(L-1)}$ being linearly independent is sufficient for the Jacobian matrix of model (6.1) to have full rank, the above example suggests that having only one "good" manifest variable with several non-informative manifest variables is sufficient to achieve local identifiability. This is certainly not the case. Otherwise, in diagnostic testing settings with binary disease groups, for example, this would mean that one informative binary test with two random guesses is sufficient for model identifiability. However,

the only informative test needs to provide estimates for disease prevalence as well as for its own sensitivity and specificity. There are three parameters but only two degrees of freedom for the results of the first binary test. Therefore, the model is not identifiable.

In fact, let π be the disease prevalence and Se_k and Sp_k be the sensitivity and specificity of the kth binary test, respectively. The Jacobian for model (6.1) with a binary disease group and 3 binary tests is a 7 by 7 matrix with determinant:

$$|J| = \pi^3(\pi-1)^3(Se_1+Sp_1-1)^2(Se_2+Sp_2-1)^2(Se_3+Sp_3-1)^2. \tag{6.4}$$

When a test is non-informative, I have that

$$Se = P(T+|D+) = P(T+|D-) = 1-Sp.$$

Consequently, it is easy to see that when one or more of the tests are non-informative, Jacobian (6.4) is singular. However, by the same construction described above, I still have that Ψ_0 and Ψ_1 are linearly independent as long as one test is informative.

An Algebraic Geometry Point of View

It is interesting to revisit this problem from an algebraic geometry point of view. Model (6.1) can be expressed as follows,

$$p = F(\theta) = \sum_{d=0}^{L-1} \pi_d \Psi_d. \tag{6.5}$$

Therefore, p is a linear combination of vectors $\Psi_0, \ldots, \Psi_{L-1}$. It may be natural to consider that if $\Psi_0, \ldots, \Psi_{L-1}$ are linearly independent, it becomes a basis of its span over the real field. Thus, the decomposition is unique and the model is identifiable. In fact, this is the essential idea of Yakowitz and Spragins (1968) work, when they studied the identifiability of finite mixture models. However, they required that the component distributions belonged to a pre-specified family $\mathcal{F}$, and that all elements in $\mathcal{F}$ were linearly independent over the field of real numbers. For example, as they showed, the exponential family and the Gaussian family are such families.

However, their results do not apply when considering the identifiability of model (6.1), where all manifest variables are categorical and are not constrained as reasoned below.

In this case, component distributions belong to a multinomial family. For any m variate multinomial family, the probability mass function can be expressed as a factor with $m-1$ elements. Therefore, at most m such vectors will be linearly dependent. Meanwhile, there are an infinite number of such m element probability vectors in the m variate multinomial family, so the multinomial family does not satisfy the conditions set forth in Yakowitz and Spragins (1968).

Another way to understand this problem is by directly examining equation (6.5). In this model, $\Psi_0, \ldots, \Psi_{L-1}$ may only expand a subspace in $\mathcal{F}$, and therefore is not its basis, and even if it were, this is not the only basis of the vector space of $F(\theta)$. In fact, any simultaneous rotation of $\Psi_0, \ldots, \Psi_{L-1}$ along vector p can lead to a different decomposition of p while maintaining the necessary property for probability mass that the sum of all of the elements is equal to1.

For example, one can consider the problem in a 3-dimensional space for simplicity. Define the length of a vector $\vec{a} = (a_1, a_2, a_3)$ as $|\vec{a}| = a_1 + a_2 + a_3$. Then, as vectors of probability mass, Ψ_1, Ψ_2, Ψ_3 are all unit vectors with nonnegative elements, suppose that they are linearly independent and that,

$$p = F(\theta) = \pi_1\Psi_1 + \pi_2\Psi_2 + \pi_3\Psi_3.$$

A rotation of angle ϕ along vector $p = (p_1, p_2, p_3)$ has a rotation matrix as follows:

$$R(\phi) =$$
$$\begin{pmatrix} [\cos\phi + p_1(1-\cos\phi)]^2 & [\sqrt{p_1p_2}(1-\cos\phi) - \sqrt{p_3}\sin\phi]^2 & [\sqrt{p_1p_3}(1-\cos\phi) + \sqrt{p_2}\sin\phi]^2 \\ [\sqrt{p_1p_2}(1-\cos\phi) + \sqrt{p_3}\sin\phi]^2 & [\cos\phi + p_2(1-\cos\phi)]^2 & [\sqrt{p_2p_3}(1-\cos\phi) - \sqrt{p_1}\sin\phi]^2 \\ [\sqrt{p_1p_3}(1-\cos\phi) - \sqrt{p_2}\sin\phi]^2 & [\sqrt{p_2p_3}(1-\cos\phi) + \sqrt{p_1}\sin\phi]^2 & [\cos\phi + p_3(1-\cos\phi)]^2 \end{pmatrix}.$$

Note that R is different from the rotation matrix of angle ϕ along vector $a = (a_1, a_2, a_3)$ in Euclidean space

$$\begin{pmatrix} \cos\phi + a_1^2(1-\cos\phi) & a_1a_2(1-\cos\phi) - a_3\sin\phi & a_1a_3(1-\cos\phi) + a_2\sin\phi \\ a_1a_2(1-\cos\phi) + a_3\sin\phi & \cos\phi + a_2^2(1-\cos\phi) & a_2a_3(1-\cos\phi) - a_1\sin\phi \\ a_1a_3(1-\cos\phi) - a_2\sin\phi & a_2a_3(1-\cos\phi) + a_1\sin\phi & \cos\phi + a_3^2(1-\cos\phi) \end{pmatrix}$$

because of the different definition of length .

Then, for any angle $\phi \in [0, 2\pi)$, I have that,

$$p = F(\theta) = \pi_1 \Psi_1^*(\phi) + \pi_2 \Psi_2^*(\phi) + \pi_3 \Psi_3^*(\phi),$$

where $\Psi_k^*(\phi) = R(\phi)\Psi_k$, $k = 1, 2, 3$, remain unit vectors with nonnegative elements, and thus are vectors of probability mass.

This example further illustrates that having $\Psi_0, \ldots, \Psi_{L-1}$ being linearly independent is not in itself a sufficient condition to guarantee the identifiability of finite mixture model (6.1).

6.3.3 Models with Covariates

Now I will consider local identifiability for latent class models with covariates, model (6.2), as follows:

$$P(T_1 = t_1, \ldots, T_K = t_K | \vec{X}, \vec{Z}) = \sum_{d=0}^{L-1} \Big\{ P(D = d|\vec{Z}) \prod_{k=1}^{K} \prod_{j=1}^{J_k} \big[P(T_k = j | D = d, \vec{X}) \big]^{I[t_k = j]} \Big\}.$$

I further assume that the covariate effects are linear under certain pre-specified transformations, such as under a logit transformation. For example, one may use polytomous regression models for both the latent group membership and the manifest variables, as both are categorical,

$$\pi_d(\vec{z}^T \alpha_d) = P(D_i = d | \vec{Z}_i = \vec{z}), \quad \log \frac{\pi_d(\vec{Z}_i^T \alpha_d)}{\pi_0(\vec{Z}_i^T \alpha_0)} = \vec{Z}_i^T \alpha_d \quad d = 1, ..., L-1$$

$$g_{kjd}(\vec{x}^T \beta_{kjd}) = P(T_{ik} = j | D_i = d, \vec{X}_i = \vec{x}), \quad \log \frac{g_{kjd}(\vec{X}_i^T \beta_{kjd})}{g_{kJ_kd}(\vec{X}_i^T \beta_{kJ_kd})} = X^T \beta_{kd}$$
$$d = 0, ..., L-1; \ k = 1, \ldots, K; \ j = 1, \ldots, J_k - 1.$$

Under the linear covariate effect assumption, model (6.2) can be rewritten in the following form:

$$P(\vec{T} = \vec{t} \,| \vec{X}, \vec{Z}) = \sum_{d=0}^{L-1} \pi_d(\vec{Z}^T \alpha_d) \prod_{k=1}^{K} \prod_{j=1}^{J_k} g_{kjd}(\vec{X}^T \beta_{kjd})^{I[t_k = j]}, \tag{6.6}$$

where $\quad \pi_d(\vec{Z}^T \alpha_d) = P(D = d | \vec{Z}), \quad g_{kjd}(\vec{X}^T \beta_{kjd}) = P(T_{ik} = j | D_i = d, \vec{X}).$

We can see that the function above has a similar form to the model without covariates, except that π_d and g_{kjd} are predefined functions while the parameters to be estimated are

α_d and β_{kdj}, $d = 0, \ldots, L-1$; $k = 1, \ldots K$; $j = 1, \ldots, J_k$. In fact, a similar procedure can be used when examining the local invertibility of this function.

Let N by q matrix $\mathbf{X}$ and N by p matrix $\mathbf{Z}$ be the design matrix in model (6.6), with the ith row $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively. Let $\eta_d(\mathbf{X}_i) = \Psi_d(\mathbf{X}_i) - \Psi_0(\mathbf{X}_i), \quad d = 1, \ldots, L-1$, where $\Psi_d(\mathbf{X}_i)$ is a vector of length $(\prod_{k=1}^{K} J_k) - 1$, with the hth element

$$\psi_{dh}(\mathbf{X}_i) = P(\vec{T} = \vec{t}_h | D = d, \mathbf{X}_i) = \prod_{k=1}^{K} g_{kt_{hk}d}(\mathbf{X}_i \beta_{kt_{hk}d}).$$

Additionally, let $\Gamma_{kjd}(\mathbf{X}_i, \mathbf{Z}_i)$ be a vector of length $(\prod_{k=1}^{K} J_k) - 1$, $d = 0, \ldots, L-1$, $k = 1, \ldots K$ and $j = 1, \ldots, J_k - 1$, with the hth element

$$\gamma_{kjdh}(\mathbf{X}_i, \mathbf{Z}_i) = \pi_d(\mathbf{Z}_i \alpha_d)\psi_{hd}(\mathbf{X}_i)\Big[\frac{I(t_{hk} = j)}{g_{kjd}(\mathbf{X}_i \beta_{kjd})} - \frac{I(t_{hk} = J_k)\big(g_{kjd}(\mathbf{X}_i \beta_{kjd}) - \sum_{j=1}^{J_k - 1} g_{kjd}(\mathbf{X}_i \beta_{kjd})\big)}{g_{kJ_kd}(\mathbf{X}_i \beta_{kJ_kd})}\Big]$$

where $I(t_{hk} = j)$ is an indicator function, which equals 1 if $t_{hk} = j$ and 0 otherwise.

Further, define $N\big[(\prod_{k=1}^{K} J_k) - 1\big]$ by p matrix A_d, $d = 1, \ldots, L-1$, and $N \times ((\prod_{k=1}^{K} J_k) - 1)$ by q matrix B_{kjd}, $d = 0, \ldots, L-1$, $k = 1, \ldots K$, $j = 1, \ldots, J_k$, as follows,

$$A_d = \begin{pmatrix} \eta_d(\mathbf{X}_1)\mathbf{Z}_1 \\ \vdots \\ \eta_d(\mathbf{X}_N)\mathbf{Z}_N \end{pmatrix}, \quad B_{kjd} = \begin{pmatrix} \Gamma_{kjd}(\mathbf{X}_i, \mathbf{Z}_i)\mathbf{X}_1 \\ \vdots \\ \Gamma_{kjd}(\mathbf{X}_N, \mathbf{Z}_N)\mathbf{X}_N \end{pmatrix}.$$

Theorem 2:

Finite mixture model (6.6) is locally identifiable at parameter $\theta = \{\alpha_d, \beta_{kjd} \mid d = 0, \ldots, L-1; k = 1, \ldots K; j = 1, \ldots, J_k\}$ if the following conditions hold.

(i) $N\big[(\prod_{k=1}^{K} J_k) - 1\big] \geq q \times L[\sum_{k=1}^{K}(J_k - 1)] + p(L-1)$;

(ii) $P(\vec{T} = \vec{t}_h | \mathbf{X}_i, \mathbf{Z}_i) = \sum_{d=0}^{L-1} \pi_d \psi_{dh}(\mathbf{X}_i) > 0, \quad \forall h$; and

(iii) column vectors in matrices $\{A_d | d = 1, \ldots, L-1\}$, $\{B_{kjd} | d = 0, \ldots, L-1; k = 1, \ldots K; j = 1, \ldots, J_k\}$ all together are linearly independent.

To prove Theorem 2, I will first prove the following lemma.

Lemma 1:

Suppose that A is a matrix, and that A^* is a matrix obtained by stacking some of the rows

in A vertically l times,

$$A* = \begin{pmatrix} A_1 \\ A_2^* \end{pmatrix}, \qquad A_2^* = \left.\begin{pmatrix} A_2 \\ \vdots \\ A_2 \end{pmatrix}\right\} l \text{ times}$$

where A_1, A_2 are sub-matrices of A, such that $A = (A_1^T, A_2^T)^T$, l is a given positive integer. Then the column vectors of A being linearly independent is equivalent to the column vectors of A^*, linearly independent.

Proof of Lemma 1:

Denote the set of column vectors of A by A_{col}, and the set of column vectors of A^* by A^*_{col}. Because each of the vectors in A^*_{col} append more elements to each of the vectors in A_{col}, A^*_{col} is an extension group of A_{col}. Therefore, if the vectors in A_{col} are linearly independent, I have that the vectors in A^*_{col} are linearly independent.

Now I prove that the reverse is also true by contradiction. Suppose that the vectors in A^*_{col} are linearly independent, but that the vectors in A_{col} are linearly dependent. Then there exists a vector α, such that $A\alpha = 0$. Thus, $A_1\alpha = 0$ and $A_2\alpha = 0$. It follows that,

$$A^*\alpha = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_2 \end{pmatrix} \alpha = \begin{pmatrix} A_1\alpha \\ A_2\alpha \\ \vdots \\ A_2\alpha \end{pmatrix} = 0.$$

However, this contradicts my supposition that the vectors in A_{col} are linearly independent. □

Proof of Theorem 2:

Conditions (i) and (ii) are similar as before. I only need to prove condition (iii) here.

Based on model (6.6), the function between parameter θ and all possible response pat-

terns of the manifest variables $\vec{t}_h$ (excluding a reference pattern) is,

$$\begin{aligned} p &= F(\vec{t}_h;\theta|\vec{X},\vec{Z}) = P_\theta(\vec{T}=\vec{t}\,|\vec{X},\vec{Z}) \\ &= \sum_{d=0}^{L-1}\pi_d(\vec{Z}^T\alpha_d)\prod_{k=1}^{K} g_{kjd}(\vec{X}^T\beta_{kt_{hk}d}), \end{aligned}$$

where $h=1,\ldots,H$ with $H=(\prod_{k=1}^{K}J_k)-1$.

First, I consider the situation where all of the covariate vectors $(\vec{X}_i^T,\vec{Z}_i^T)^T$ are different for $i=1,\ldots,N$. In this case, F defines a mapping between all H possible response patterns of $\vec{t}$ for each of the covariate vectors $(\vec{X}_i^T,\vec{Z}_i^T)^T$. Therefore, the Jacobian matrix has $N\times H$ rows. It can be divided into N blocks, each containing H contiguous rows. Then the ith block is the derivative of F with respect to each of the parameters when the covariate vectors are $(\vec{X}_i^T,\vec{Z}_i^T)^T$. In other words, it is the Jacobian matrix of F when there is only a single observation with covariates $(\vec{X}_i^T,\vec{Z}_i^T)^T$. Since each block can compute each block separately, I only compute the ith block with covariates $(\vec{X}_i^T,\vec{Z}_i^T)^T$.

Take the derivative of F with respect to free parameters α_d, $d=1,\ldots,L-1$, and by the chain rule, I have

$$\begin{aligned} \frac{\partial F(\vec{t}_h;\theta|\mathbf{X}_i,\mathbf{Z}_i)}{\partial\alpha_d} &= \frac{\partial F(\vec{t}_h;\theta|\mathbf{X}_i,\mathbf{Z}_i)}{\partial\pi_d(\mathbf{Z}_i\alpha_d)}\cdot\frac{\partial\pi_d(\mathbf{Z}_i\alpha_d)}{\partial\alpha_d} \\ &= \Big[\prod_{k=1}^{K} g_{kt_{hk}d}(\mathbf{X}_i\beta_{kt_{hk}d}) - \prod_{k=1}^{K} g_{kt_{hk}0}(\mathbf{X}_i\beta_{kt_{hk}0})\Big]\mathbf{Z}_i \\ &= \Big[\psi_{dh}(\mathbf{X}_i)-\psi_{0h}(\mathbf{X}_i)\Big]\mathbf{Z}_i. \end{aligned}$$

Therefore, the first $p\times(L-1)$ columns of the ith block of the Jacobian matrix of model (6.6) is,

$$\frac{\partial F}{\partial\alpha_d} = \Big[\Psi_d(\mathbf{X}_i)-\Psi_0(\mathbf{X}_i)\Big]\mathbf{Z}_i = \eta_d(\mathbf{X}_i)\mathbf{Z}_i, \quad d=1,\ldots,L-1.$$

Then, taking the derivative of F with respect to free parameter β_{kjd}, $d=0,\ldots,L-1$, $k=1,\ldots K$ and $j=1,\ldots,J_k-1$, I have

$$\begin{aligned} \frac{\partial F(\vec{t}_h;\theta|\mathbf{X}_i,\mathbf{Z}_i)}{\partial\beta_{kjd}} &= \frac{\partial F(\vec{t}_h;\theta|\mathbf{X}_i,\mathbf{Z}_i)}{\partial g_{kjd}(\mathbf{X}_i^T\beta_{kjd})}\cdot\frac{\partial g_{kjd}(\mathbf{X}_i^T\beta_{kjd})}{\partial\beta_{kjd}} \\ &= \Big\{\pi_d(\mathbf{Z}_i\alpha_d)\psi_{dh}(\mathbf{X}_i)\Big[\frac{I(t_{hk}=j)}{g_{kjd}(\mathbf{X}_i^T\beta_{kjd})} - \frac{I(t_{hk}=J_k)(g_{kjd}(\mathbf{X}_i^T\beta_{kjd})-\sum_{j=1}^{J_k-1}g_{kjd}(\mathbf{X}_i^T\beta_{kjd}))}{g_{kJ_kd}(\mathbf{X}_i^T\beta_{kJ_kd})}\Big]\Big\}\mathbf{X}_i \\ &= \gamma_{kjdh}(\mathbf{X}_i,\mathbf{Z}_i)\mathbf{X}_i \end{aligned}$$

Thus, the last $q \times L[\sum_{k=1}^{K}(J_k - 1)]$ columns of the ith block of the Jacobian matrix of model (6.6) are,

$$\frac{\partial F}{\partial \beta_{kjd}} = \Gamma_{kjd}(\mathbf{X}_i, \mathbf{Z}_i)\mathbf{X}_i, \quad d = 0, \ldots, L-1, \ k = 1, \ldots K, \ j = 1, \ldots, J_k - 1.$$

Therefore, the Jacobian matrix of model (6.6) is a $N \times \left[(\prod_{k=1}^{K} J_k) - 1\right]$ by $q \times L[\sum_{k=1}^{K}(J_k - 1)] + p(L-1)$ matrix

$$J^*(\theta) = (A_1, \ldots, A_{L-1}, B_{111}, \ldots, B_{KJ_K-1L-1}).$$

As a result, condition (iii) is equivalent to requiring that the Jacobian matrix of model (6.6) has full column rank, which in turn guarantees that the finite mixture model (6.6) is locally identifiable at parameter $\theta = \{\alpha_d, \beta_{kjd} \mid d = 0, \ldots, L-1; k = 1, \ldots K; j = 1, \ldots, J_k\}$.

Now suppose that some of the covariate vectors $(\vec{X}_i^T, \vec{Z}_i^T)^T$ are the same. Then the Jacobian matrix of model (6.6) is a sub-matrix of $J(\theta)$, obtained by excluding the repeated blocks. By Lemma 1, column vectors linearly independent are equivalent for these two matrices. □

Note that, in the above proof,,the number of rows in the Jacobian matrix $J^*(\theta)$ for a finite mixture model with covariates is N-fold of that in the Jacobian matrix $J(\theta)$ for a finite mixture model without covariates, where N is the total sample size. If $J^*(\theta)$ is divided vertically into N blocks with equal sizes, each block is essentially $J(\theta)$ for subjects with the same covariate $\mathbf{X}_i$, $\mathbf{Z}_i$ and then multiplied by the corresponding design matrix. As a result, if one of these N blocks has full column rank, the longer matrix $J^*(\theta)$ will have full column rank. Moreover, even if $J(\theta)$ does not have full column rank for any of the covariate patterns, when the design matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ have full column rank, they may help restore full column rank of $J^*(\theta)$. This result is very interesting since it suggests that finite mixture models with covariates may be easier to identify than models without covariates. The difficulty with including covariates in finite mixture models is that it increases the number of parameters to be estimated, especially when the covariates are categorical, and especially when it is of interest to consider their interactions with each manifest variable and each latent group. Because of this, researchers adopted various constraints to ensure model identification, such as assuming that either the latent structure model or the measurement model is indexed

with covariates, or assuming that covariate effects are constant across latent groups to remove some interaction terms, etc. In contrast to previous beliefs, these results suggest that for models considered here, a model with covariates is more likely to be identifiable than a model without covariates. This result is not surprising when considering that the additional degrees of freedom in the data introduced by covariates are much higher than the increase in the number of parameters – although having sufficient degrees of freedom does not guarantee model identification, it is a necessary condition. As a simple example, consider that a model with two independent binary tests for binary disease status is not identifiable, but Hui and Walter (1980) showed that with two populations, this model can be identifiable. The population here can be viewed as a covariate that they included in their latent structure model.

Specially, I have the following corollary.

Corollary 1:

Finite mixture model (6.6) is locally identifiable at parameter $\theta = \{\alpha_d, \beta_{kjd} \mid d = 0, \ldots, L-1; k = 1, \ldots K; j = 1, \ldots, J_k\}$ if the following conditions hold.

(i) $N\left[(\prod_{k=1}^{K} J_k) - 1\right] \geq q \times L[\sum_{k=1}^{K}(J_k - 1)] + p(L-1)$;

(ii) $P(\vec{T} = \vec{t}_h | \mathbf{X}_i, \mathbf{Z}_i) = \sum_{d=0}^{L-1} \pi_d \psi_{dh}(\mathbf{X}_i) > 0, \quad \forall h$;

(iii) column vectors in matrices $\{\eta_d(\mathbf{X}_0) | d = 1, \ldots, L-1\}$, $\{\Gamma_{kjd}(\mathbf{X}_0, \mathbf{Z}_0) | d = 0, \ldots, L-1; k = 1, \ldots K; j = 1, \ldots, J_k\}$ all together are linearly independent for some $\mathbf{X}_0 \in \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and $\mathbf{Z}_0 \in \{\mathbf{Z}_1, \ldots, \mathbf{Z}_N\}$;and

(iv) design matrix $\mathbf{X}$ and $\mathbf{Z}$ both have full rank.

Proof:

We only need to show that conditions (iii) and (iv) in Corollary 1 $\Rightarrow$ condition (iii) in Theorem 2. Let

$$\Theta = \begin{pmatrix} \eta_1(\mathbf{X}_1) & \ldots & \eta_{L-1}(\mathbf{X}_1) & \Gamma_{111}(\mathbf{X}_1, \mathbf{Z}_1) & \ldots & \Gamma_{KJ_K-1L-1}(\mathbf{X}_1, \mathbf{Z}_1) \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ \eta_1(\mathbf{X}_1) & \ldots & \eta_{L-1}(\mathbf{X}_N) & \Gamma_{111}(\mathbf{X}_N, \mathbf{Z}_N) & \ldots & \Gamma_{KJ_K-1L-1}(\mathbf{X}_N, \mathbf{Z}_N) \end{pmatrix},$$

and $\mathbb{X}$ be a block diagonal matrix,

$$\mathbb{X} = \text{diag}\{\ \underbrace{\mathbf{Z}, \ \ldots, \ \mathbf{Z}}_{L-1 \text{ times}}, \ \underbrace{\mathbf{X}, \ \ldots, \ \mathbf{Z}}_{L[\sum_{k=1}^{K}(J_k-1)] \text{ times}}\ \}.$$

Then the Jacobian matrix of model (6.6) can be rewritten as follows,

$$J(\theta) = (A_1, \ldots, A_{L-1}, B_{111}, \ldots, B_{KJ_K-1L-1}) = \Theta\mathbb{X}.$$

Condition (iii) guarantees that Θ has full column rank, since the column vectors in Θ are an extension group of the vectors in condition (iii). Meanwhile, condition (iv) guarantees that $\mathbb{X}$ has full rank. Therefore, $J(\theta) = \Theta\mathbb{X}$ has full column rank. □

6.4 Global Identifiability

Global identifiability considers the identifiability of a model in its entire parameter space. It does not just focus on a single parameter value that may result from a particular sample, but instead considers all possible parameter choices for a given model structure, and is thus more fundamental than local identifiablity. However, proving the global identifiability of a function is generally very hard; with the additional difficulty introduced by "label switching" in a finite mixture model, this problem has not been well addressed in the literature.

6.4.1 Definition

The global identifiability of a model is defined equivalently as the global invertibility of the model induced function $p = F(\theta)$. In the classic definition, a function F is globally invertible for all parameter $\theta \in \Theta$, if

$$F(\theta) = F(\theta^*) \Rightarrow \theta = \theta^*$$

$$\text{or equivalently,} \qquad F(\theta) \neq F(\theta^*) \qquad \forall\ \theta \neq \theta^*,\ \theta \in \Theta,\ \theta^* \in \Theta.$$

This suggests that F is a one-to-one map on its domain Θ.

Due to the "label switching" problem mentioned in the introduction, finite mixture models are not strictly identifiable. However, since group labels are usually not difficult to determine in practice, it is still worthwhile to consider whether the model induced function $p = F(\theta)$ can uniquely determine a parameter value, up to permutations of group labels.

To express this idea clearly, I first define an equivalent class and an equivalent relationship. Let $\theta = \{\theta_1, \ldots, \theta_{L-1}\}$, where θ_d, $d = 1, \ldots, L-1$, contain the parameters related to the dth group. I define an equivalent class of θ, denoted by $[\theta]$, as

$$[\theta] = \{\theta_{\sigma(1)}, \ldots, \theta_{\sigma(L-1)} \mid \sigma(1), \ldots, \sigma(L-1) \text{ is a permutation of } 1, \ldots, L-1.\}.$$

Moreover, define an equivalent relationship on Θ, denoted by $\sim$, as

$$\theta \sim \theta^* \text{ if } \theta \in [\theta^*], \text{ or equivalently } [\theta] = [\theta^*].$$

The global identifiability of a finite mixture model can then be defined as follows. A finite mixture model is globally identifiable if its induced function $p = F(\theta)$ satisfies that,

$$F(\theta) = F(\theta^*) \Rightarrow \theta \sim \theta^*$$

$$\text{or equivalently,} \qquad F(\theta) \neq F(\theta^*) \qquad \forall\ \theta \notin [\theta^*],\ \theta \in \Theta,\ \theta^* \in \Theta.$$

In other words, if a finite mixture model is globally identifiable, its parameter θ can be uniquely determined by the model induced function $p = F(\theta)$, regardless the value of θ, which is in contrast to local identifiability. Clearly, if a model is globally identifiable on Θ, then it is locally identifiable at each $\theta \in \Theta$. Consequently, the Jacobian matrix $F(\theta)$ has full column rank for all $\theta \in \Theta$. However, the reverse is not true. This is because having a Jacobian of full column rank in a region does not guarantee F is invertible in that region. As an example, we can consider the following function on the unit circle $D = \{(x, y) \mid 0 < x^2 + y^2 < 1\}$,

$$F(x, y) = \begin{cases} x^2 - y^2 = u \\ 2xy = v \end{cases}$$

The Jacobian is

$$\frac{\partial(u, v)}{\partial x, y} = \begin{vmatrix} 2x & -2y \\ 2y & 2x \end{vmatrix} = 4(x^2 + y^2) > 0.$$

However, $F(0, \frac{1}{2}) = F(0, -\frac{1}{2})$, so F is not globally invertible. Therefore, examining the Jacobian in the entire parameter space does not lead to global identifiability of the model. Consequently, methods in the previous section cannot be extended to show global identifiability of a finite mixture model.

Because of the difficulty in establishing global identifiability, an alternative concept that some researchers have considered, is generic identifiability, which requires that the set of parameter values on which the model is not identifiable has measure zero. More formally, a model is called generically identifiable if its induced function $p = F(\theta)$ satisfies that,

$$\theta \in \Theta \setminus U,\ \theta^* \in \Theta \setminus U,\ F(\theta) = F(\theta^*) \Rightarrow \theta = \theta^*$$
$$\text{or equivalently,}\ \ F(\theta) \neq F(\theta^*) \qquad \forall\ \ \theta \neq \theta^*,\ \theta \in \Theta \setminus U,\ \theta^* \in \Theta \setminus U,$$

where U is a set of measure zero.

Due to the "label switching" problem, the definition for generic identifiability of a finite mixture model is in fact as follows: a finite mixture model is called generically identifiable if its induced function $p = F(\theta)$ satisfies that,

$$\theta \in \Theta \setminus U,\ \theta^* \in \Theta \setminus U,\ F(\theta) = F(\theta^*) \Rightarrow \theta \sim \theta^*$$
$$\text{or equivalently,}\ \ F(\theta) \neq F(\theta^*) \qquad \forall\ \theta \notin [\theta^*],\ \theta \in \Theta \setminus U,\ \theta^* \in \Theta \setminus U,$$

where U is a set of measure zero.

This concept is useful because, with generic identifiability, one has probability one to reach an identifiable model. Thus, the parameter values can be uniquely determined up to label switching. However, this conclusion is weaker than the global identifiability considered in my work.

6.4.2 Models without Covariates

In this section I consider global identifiability for a finite mixture model without covariates. An important result that I incorporate into my proof establishes the uniqueness of trilinear decomposition (Kruskal, 1977). I summarize Kruskal's work on this subject below. I then proceed with my proof, which follows the idea of Allman, et al. (2009), except that I will focus on global identifiability instead of generic identifiability of mixture models.

Kruskal's Result

Kruskal's result originated from the decomposition of a three-way contingency table. To summarize the result, I first introduce some notation and definitions.

Let M_k be a L by J_k matrix, $k = 1, 2, 3$, with the dth row $\mathbf{m_d^k} = (m_{d1}^k, \ldots, m_{dJ_k}^k)$. Let $\big[M_1, M_2, M_3\big]$ denote a $J_1 \times J_2 \times J_3$ tensor defined as follows,

$$\big[M_1, M_2, M_3\big] = \sum_{d=0}^{L-1} \mathbf{m_d^1} \otimes \mathbf{m_d^2} \otimes \mathbf{m_d^3}.$$

In other words, $\big[M_1, M_2, M_3\big]$ is a three-dimensional array with the (u, v, w) element

$$\big[M_1, M_2, M_3\big](u, v, w) = \sum_{d=0}^{L-1} m_{du}^1 m_{dv}^2 m_{dw}^3.$$

Additionally, define the Kruskal rank of a matrix M, denoted by $rank_K M$, as the largest integer I such that every set of I rows of M are linearly independent. Consequently, $rank_K M$ is less than or equal to the row rank of M, with equality if and only if M has full row rank. Kruskal showed the following result.

Lemma 2

If $rank_K M_1 + rank_K M_2 + rank_K M_3 \geq 2L + 2$, then tensor $\big[M_1, M_2, M_3\big]$ uniquely determines M_k, up to simultaneous permutation and re-scaling of the rows.

I first consider the identifiability of model (6.1), where all the manifest variables are discrete. If $K = 3$, the following theorem holds.

Theorem 3

Let M_k be a L by J_k matrix, with the $(d + 1, j)$ element

$$M_k(d + 1, j) = P(T_k = j | D = d), \quad k = 1, 2, 3, \ d = 0, \ldots, L - 1, \ j = 1, \ldots, J_k.$$

Then finite mixture model (6.1) with $K = 3$ is globally identifiable if

$$rank_K M_1 + rank_K M_2 + rank_K M_3 \geq 2L + 2.$$

Proof:

Let L be a diagonal matrix, $L = diag\{P(D=0), \ldots, P(D=L-1)\}$. Let $\tilde{M}_1 = LM_1$, then the (u,v,w) element of tensor $\big[\tilde{M}_1, M_2, M_3\big]$ is

$$\begin{aligned}\big[M_1, M_2, M_3\big](u,v,w) &= \sum_{d=0}^{L-1} P(D=d) M_1(d,u) M_2(d,v) M_3(d,w) \\ &= P(T_1 = u, T_2 = v, T_3 = w).\end{aligned}$$

According to Lemma 2, the finite mixture model (6.1) uniquely determines $\tilde{M}_1$, M_2 and M_3 up to a simultaneous permutation and re-scaling of the rows. Meanwhile, since $\sum_{D=0}^{L-1} P(D = d) = 1$, $\sum_{j=1}^{J_k} P(T_k = j|D=d) = 1$ for all $d = 0, \ldots, L-1$, it follows that: $\sum_{d=0}^{L-1}\sum_{j=1}^{J_1} \tilde{M}_1 = 1$, $\sum_{d=0}^{L-1}\sum_{j=1}^{J_2} M_2 = L$ and $\sum_{d=0}^{L-1}\sum_{j=1}^{J_3} M_3 = L$. Thus, the scaling of the rows is uniquely determined. Moreover, suppose that $\tilde{M}_1$ is properly scaled, then $P(D=d) = \sum_{j=1}^{J_1} \tilde{M}_1(d,j)$, $d = 0, \ldots, L-1$, and $M_1 = L^{-1}\tilde{M}_1$. Therefore, M_1, M_2 and M_3 are uniquely determined up to simultaneous permutation of the rows, and thus finite mixture model (6.1) is globally identifiable. □

When $K \geq 3$ in model (6.1), several univariate tests can be viewed as a single multivariate test and Theorem 3 can be used to examine the global identifiability of the model. Specifically, I have the following theorem,

Theorem 4

Suppose $K \geq 3$. Let A, B, C be a partition of $\{1, \ldots, K\}$, where $A = \{a(1), \ldots, a(p)\}$, $B = \{b(1), \ldots, b(q)\}$ and $C = \{c(1), \ldots, c(r)\}$, with $p \geq 1$, $q \geq 1$,$r \geq 1$, and $p + q + r = K$. Let M_A be a L by $\prod_{i=1}^{p} J_{a(i)}$ matrix with the $(d+1, j)$ element

$$M_A(d+1, j) = P\big((T_{a(1)}, \ldots, T_{a(p)}) = \vec{t}_{Aj} | D = d\big),$$

where $\vec{t}_{Aj}$ is the jth possible in lexicographic order among $\prod_{i=1}^{p} J_{a(i)}$ distinct response patterns of $(T_{a(1)}, \ldots, T_{a(p)})$. Similarly, define M_B as a L by $\prod_{i=1}^{q} J_{b(i)}$ matrix with the $(d+1, j)$ element $M_B(d+1, j) = P\big((T_{b(1)}, \ldots, T_{b(q)}) = \vec{t}_{Bj} | D = d\big)$, and M_C as a L by $\prod_{i=1}^{r} J_{c(i)}$ matrix with the $(d+1, j)$ element $M_C(d+1, j) = P\big((T_{c(1)}, \ldots, T_{c(r)}) = \vec{t}_{Bj} | D = d\big)$. Then the finite mixture model (6.1) is globally identifiable if $\text{rank}_K M_A + \text{rank}_K M_B + \text{rank}_K M_C \geq 2L + 2$.

Proof:

Applying Theorem 3, parameters $P(D = d)$, $P((T_{a(1)}, \ldots, T_{a(p)})|D = d)$, $P((T_{b(1)}, \ldots, T_{b(q)})|D = d)$ and $P((T_{c(1)}, \ldots, T_{c(r)})|D = d)$, $D = 0, \ldots, L-1$, can be uniquely identified up to label switching. Moreover, the marginal distributions $P(T_k = j|D = d)$, $k = 1, \ldots, K$, $d = 0, \ldots, L-1$, $j = 1, \ldots, J_k$ can be obtained from these joint distributions, thus the theorem holds. □

Theorem 4 also illustrates why information from additional tests is helpful for model identification, since the matrix for multiple manifest variables has a larger Kruskal rank than the one for a single manifest variable. To see this more clearly, let M_1 be a L by J_1 matrix, with the $(d+1, j)$ element $M_1(d+1, j) = P(T_1 = j|D = d)$, and M_2 be a L by J_2 matrix, with the $(d+1, j)$ element $M_2(d+1, j) = P(T_2 = j|D = d)$. Additionally, let L_j be a diagonal matrix $L_j = diag\{M_2(0, j), \ldots, M_2(L-1, j)\}$, $j = 1 \ldots, J_2$. Then the matrix according to the joint distribution of T_1 and T_2 is a L by $J_1 J_2$ matrix as follows,

$$M = (L_1 M_1, \ldots, L_{J_2} M_1)\,.$$

Because L_j is a diagonal matrix with all diagonal elements positive, if any of the row vectors in M_1 are linearly independent, the same rows of the vectors in $L_1 M_1$ are also linearly independent, and vice versa. Additionally, the row vectors in M are an extension group of the row vectors in $L_1 M_1$, thus $rank_K M \geq rank_K M_1$. This argument can be generalized to several manifest variables. Consequently, I have the following corollary.

Corollary 2

Let M_k be a L by J_k matrix, with the $(d+1, j)$ element

$$M_k(d+1, j) = P(T_k = j|D = d), \quad k = 1, \ldots, K,\ d = 0, \ldots, L-1,\ j = 1, \ldots, J_k.$$

If there exist $k_1, k_2, k_3 \in \{1, \ldots, K\}$ such that

$$rank_K M_{k_1} + rank_K M_{k_2} + rank_K M_{k_3} \geq 2L + 2,$$

finite mixture model (6.1) is globally identifiable.

Moreover, if a L by J_k matrix M has full row rank, $rank_K M = L$, this gives the following corollary.

Corollary 3

Let M_k be a L by J_k matrix, with the $(d+1, j)$ element

$$M_k(d+1, j) = P(T_k = j | D = d), \quad k = 1, \ldots, K, \ d = 0, \ldots, L-1, \ j = 1, \ldots, J_k.$$

If there exist $k_1, k_2, k_3 \in \{1, \ldots, K\}$ such that M_{k_1}, M_{k_2}, M_{k_3} have full row rank, then the finite mixture model (6.1) is globally identifiable.

Now I consider the identifiability of a finite mixture model with continuous manifest variables. Let $f_{kd}(\cdot) = P(T_k | D = d)$ be the conditional density function of the kth manifest variable in group d, and $F_{kd}(\cdot)$ be the corresponding cumulative distribution function (CDF), $k = 1, \ldots, K$, $D = 0, \ldots, L-1$. The model can be expressed as follows,

$$P(T_1 = t_1, \ldots, T_K = t_k) = \sum_{d=0}^{L-1} P(\vec{T}, D) = \sum_{d=0}^{L-1} \big[P(D = d) \prod_{k=1}^{K} f_{kd}(t_k) \big]. \tag{6.7}$$

When $K = 3$, I claim that the following theorem holds.

Theorem 5

If there exists integer $J_1, J_2, J_3 \geq L$ and points $t_{11}, \ldots, t_{1(J_1-1)}$, $t_{21}, \ldots, t_{2(J_2-1)}$ and $t_{31}, \ldots, t_{3(J_3-1)}$, such that matrix M_1, M_2 and M_3 satisfy that,

$$rank_K M_1 + rank_K M_2 + rank_K M_3 \geq 2L + 2,$$

where the $(d+1, j)$ element in M_k is $F_{kd}(t_{kj})$, $j = 1, \ldots, J_k - 1$, and $F_{kd}(t_{J_k}) = 1$, $k = 1, 2, 3$, $d = 0, \ldots, L-1$, then the finite mixture model (6.7) is globally identifiable.

To prove Theorem 5, I first show that the following lemma holds.

Lemma 3

Let A and A^* be two P by Q matrices, defined below,

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1Q} \\ \vdots & \vdots & & \vdots \\ a_{P1} & a_{P2} & \dots & a_{PQ} \end{pmatrix} \quad A^* = \begin{pmatrix} a_{11} & a_{12} - a_{11} & \dots & a_{1Q} - a_{1Q-1} \\ \vdots & \vdots & & \vdots \\ a_{P1} & a_{P2} - a_{P1} & \dots & a_{PQ} - a_{PQ-1} \end{pmatrix},$$

then $rank_K A = rank_K A^*$.

Proof of Lemma 3

We only need to show that if any rows in A are linearly independent, the corresponding rows in A^* are also linearly independent, and vice versa. Equivalently, we can show that if any rows in A are linearly dependent, the corresponding rows in A^* are also linearly dependent, and vice versa. Without loss of generality, I assume the first $p \leq P$ rows, denoted by A_p, are linearly dependent. Then there exist $k_1, \dots, k_p$, such that $(k_1, \dots, k_p)A_p = 0$, in other words, the following equations hold

$$\begin{cases} k_1 a_{11} + \ldots + k_p a_{p1} = 0 \\ k_1 a_{12} + \ldots + k_p a_{p2} = 0 \\ \dots \\ k_1 a_{1Q} + \ldots + k_p a_{pQ} = 0 \end{cases}.$$

These are equivalent to the following equations

$$\begin{cases} k_1 a_{11} + \ldots + k_p a_{p1} = 0 \\ k_1 (a_{12} - a_{11}) + \ldots + k_p (a_{p2} - a_{p1}) = 0 \\ \dots \\ k_1 (a_{1Q} - a_{1Q-1}) + \ldots + k_p (a_{1Q} - a_{1Q-1}) = 0 \end{cases}.$$

Thus $(k_1, \dots, k_p)A_p^* = 0$, where A_p^* are the first p rows of A^*. As a result, if any rows in A are linearly independent, the corresponding rows in A^* are also linearly independent, and vice versa. Therefore, $rank_K A = rank_K A^*$. $\square$

Proof of Theorem 5

Matrix M_k, $k = 1, 2, 3$, in theorem 5 is

$$M_k = \begin{pmatrix} F_{k0}(t_{k1}) & F_{k0}(t_{k2}) & \dots & F_{k0}(t_{kJ_k-1}) & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ F_{kL-1}(t_{k1}) & F_{kL-1}(t_{k2}) & \dots & F_{kL-1}(t_{kJ_k-1}) & 1 \end{pmatrix}.$$

Create new matrices M_k^*,$k = 1, 2, 3$, as follows:

$$M_k^* = \begin{pmatrix} F_{k0}(t_{k1}) & F_{k0}(t_{k2}) - F_{k0}(t_{k1}) & \dots & 1 - F_{k0}(t_{kJ_k-1}) \\ \vdots & \vdots & & \vdots \\ F_{kL-1}(t_{k1}) & F_{kL-1}(t_{k2}) - F_{kL-1}(t_{k1}) & \dots & 1 - F_{kL-1}(t_{kJ_k-1}) \end{pmatrix}.$$

Then according to Lemma 3,

$$rank_K M_1^* + rank_K M_2^* + rank_K M_3^* = rank_K M_1 + rank_K M_2 + rank_K M_3 \geq 2L + 2.$$

Meanwhile, create categorical variables T_k^*,$k = 1, 2, 3$, such that $T_k^* = j$ if $t_{kj-1} \leq T_k \leq t_{kj}$, where $j = 1, \dots, J_k$ and $a_0 = -\infty$. Then according to Theorem 3, the finite mixture model (6.1) with T_1^*, T_2^* and T_3^* as manifest variables is globally identifiable. Specially, parameters $P(D = d)$, $F_{kd}(t_{k1})$, $F_{kd}(t_{k2}) - F_{kd}(t_{k1}), \dots, 1 - F_{kd}(t_{kJ_k-1})$ are globally identifiable, which in turn leads to parameters $P(D = d)$, $F_{kd}(t_{k1})$, $F_{kd}(t_{k2}), \dots, F_{kd}(t_{kJ_k-1})$ being globally identifiable, $d = 0 \dots, D$ and $k = 1, 2, 3$.

Let t_k^* be any number in the domain of $F_{kd}(\cdot)$. Without loss of generality, I assume $t_k^* < t_{k1}$, and create matrices $\tilde{M}_k$, $k = 1, 2, 3$ as follows:

$$\tilde{M}_k = \begin{pmatrix} F_{k0}(t_k^*) & F_{k0}(t_{k1}) & \dots & F_{k0}(t_{kJ_k-1}) & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ F_{kL-1}(t_k^*) & F_{kL-1}(t_{k1}) & \dots & F_{kL-1}(t_{kJ_k-1}) & 1 \end{pmatrix}.$$

The row vectors in $\tilde{M}_k$ are extension groups of the row vectors in M_k, $k = 1, 2, 3$. Therefore,

$$rank_K \tilde{M}_1 + rank_K \tilde{M}_2 + rank_K \tilde{M}_3 \geq rank_K M_1 + rank_K M_2 + rank_K M_3 \geq 2L + 2.$$

Constructing the corresponding finite mixture model and following the same argument above, parameters $P(D = d)$, $F_{kd}(t_k^*)$, $F_{kd}(t_{k1}), \dots, F_{kd}(t_{kJ_k-1})$ are globally identifiable,

$d = 0 \ldots, D$ and $k = 1, 2, 3$. Since t_k^* is any number in the domain of $F_{kd}(\cdot)$, the function $F_{kd}(\cdot)$ is globally identifiable. As a result, the finite mixture model (6.7) is globally identifiable. □

Similarly as in Theorem 4,several univariate tests as a multivariate test to obtain global identifiability of model (6.7) when $K \geq 3$. Specifically, we have the following theorem.

Theorem 6

Suppose $K \geq 3$. Let A, B, C be a partition of $\{1, \ldots, K\}$, where $A = \{a(1), \ldots, a(p)\}$, $B = \{b(1), \ldots, b(q)\}$ and $C = \{c(1), \ldots, c(r)\}$, with $p \geq 1$, $q \geq 1$,$r \geq 1$, and $p + q + r = K$. Let G_{1d} be the joint CDF for manifest variables $(T_{a(1)}, \ldots, T_{a(p)})$ in group d, G_{2d} be the joint CDF for manifest variables $(T_{b(1)}, \ldots, T_{b(q)})$ in group d and G_{3d} be the joint CDF for manifest variables $(T_{c(1)}, \ldots, T_{c(r)})$ in group d, $d = 0, \ldots, L-1$.

If there exists an integer $J_1, J_2, J_3 \geq L$ and points $t_{11}, \ldots, t_{1(J_1-1)}$, $t_{21}, \ldots, t_{2(J_2-1)}$ and $t_{31}, \ldots, t_{3(J_3-1)}$, such that matrices M_1, M_2 and M_3 satisfy

$$rank_K M_1 + rank_K M_2 + rank_K M_3 \geq 2L + 2,$$

where the $(d+1, j)$ element in M_k is $G_{kd}(t_{kj})$, $j = 1, \ldots, J_k - 1$, and $G_{kd}(t_{J_k}) = 1$, $k = 1, 2, 3$, $d = 0, \ldots, L-1$, then the finite mixture model (6.7) is globally identifiable. □

Moreover, with the same argument for proving Corollary 2 and 3, we know that if at least three of the manifest variables satisfy the condition in Theorem 5, then the finite mixture model (6.7) is globally identifiable. Therefore, we have the following corollaries.

Corollary 4

If there exists $s_1, s_2, s_3 \in \{1, \ldots, K\}$, integers $J_1, J_2, J_3 \geq L$ and points $t_{11}, \ldots, t_{1(J_1-1)}$, $t_{21}, \ldots, t_{2(J_2-1)}$ and $t_{31}, \ldots, t_{3(J_3-1)}$, such that matrices M_1, M_2 and M_3 satisfy that,

$$rank_K M_1 + rank_K M_2 + rank_K M_3 \geq 2L + 2,$$

where the $(d+1, j)$ element in M_k is $F_{s_k d}(t_{kj})$, $j = 1, \ldots, J_k - 1$, and $F_{s_k d}(t_{J_k}) = 1$, $k = 1, 2, 3$, $d = 0, \ldots, L-1$, then the finite mixture model (6.7) is globally identifiable. □

Corollary 5

If there exists $s_1, s_2, s_3 \in \{1, \ldots, K\}$, integers $J_1, J_2, J_3 \geq L$ and points $t_{11}, \ldots, t_{1(J_1-1)}$, $t_{21}, \ldots, t_{2(J_2-1)}$ and $t_{31}, \ldots, t_{3(J_3-1)}$, such that matrices M_1, M_2 and M_3 have full row rank, where the $(d+1, j)$ element in M_k is $F_{s_k d}(t_{kj})$, $j = 1, \ldots, J_k - 1$, and $F_{s_k d}(t_{J_k}) = 1$, $k = 1, 2, 3$, $d = 0, \ldots, L-1$, then the finite mixture model (6.7) is globally identifiable. □

From the theorem above, we can see that one important condition to guarantee global identifiability of a finite mixture model is about the row ranks of a matrix whose $(d+1, j)$ element is $P(T = j|D = d)$. I give an intuitive explanation here before moving to be next section. Each row of this matrix is about the conditional distribution of manifest variable T in a certain group. If the rows are linearly independent, it means the manifest variable T can reveal some nontrivial differences among the latent groups that can then be used to distinguish between them. Identifying latent groups is more fundamental, as once the group labels are determined, other parameters can be obtained by, in some sense, regular regression. Another explanation can be obtained by observing that the marginal distribution of T is a linear combination of its conditional distributions in each of the latent groups. The factorization is unique only when the conditional distributions are linearly independent. However, since we have multiple manifest variables, we may not need to require that every one of them be informative for all latent groups – for example, if some manifest variables are informative for all groups except groups 1 and 2, the model may still be identifiable if there is a manifest variable that can distinguish between these two groups. Clearly, the task is harder if the number of latent groups is bigger. The theorems here show a balance between the information needed from the manifest variables and the number of latent groups needed for the model to be identifiable.

6.4.3 Models with Covariates

In this section I examine the global identifiability of a finite mixture model with covariates. Again, I assume that covariate effects are linear on a certain transformed scale. First I consider model (6.6) – a model in which all manifest variables are categorical. The model

is given as follows,

$$P(\vec{T}=\vec{t}\,|\vec{X},\vec{Z}) = \sum_{d=0}^{L-1} \pi_d(\vec{Z}^T\alpha_d) \prod_{k=1}^{K}\prod_{j=1}^{J_k} g_{kjd}(\vec{X}^T\beta_{kjd})^{I[t_k=j]},$$

$$\text{where} \quad \pi_d(\vec{Z}^T\alpha_d) = P(D=d|\vec{Z}), \quad g_{kjd}(\vec{X}^T\beta_{kjd}) = P(T_{ik}=j|D_i=d,\vec{X}).$$

When $K=3$ I claim the following theorem is true.

Theorem 7

In model (6.6), suppose α_d has $u \leq N$ elements and β_{kd} has $v \leq N$ elements. Let $max\{u,v\} \leq W \leq N$ and $\{i(1),\ldots,i(W)\}$ be a subset of $\{1,\ldots,N\}$. Let $M_{k[w]}$ be a L by J_k matrix with the $(d+1,j)$ element $g_{kjd}(\mathbf{X}_{i(w)}\beta_{kd})$, where $\mathbf{X}_{i(w)}$ is the $i(w)$th row of the design matrix $\mathbf{X}$, $d=0,\ldots,L-1$, $k=1,2,3$, $j=1,\ldots,J_k$, $w=1,\ldots,W$. Let M_k be a block diagonal matrix defined as follows,

$$M_k = \begin{pmatrix} M_{k[1]} & 0 & \ldots & 0 \\ 0 & M_{k[2]} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M_{k[W]} \end{pmatrix}, \quad k=1,2,3.$$

Let $\mathbf{X}_w = (\mathbf{X}_{i(1)},\ldots,\mathbf{X}_{i(W)})^T$ and $\mathbf{Z}_w = (\mathbf{Z}_{i(1)},\ldots,\mathbf{Z}_{i(W)})^T$. Then, the finite mixture model (6.6), with $K=3$, is globally identifiable if the following conditions hold.

(i) $rank_K M_1 + rank_K M_2 + rank_K M_3 \geq 2WL+2$;

(ii) $rank\ \mathbf{X}_w \geq u$ and $rank\ \mathbf{Z}_w \geq v$.

Proof:

The idea is to "absorb" the observed covariate pattern into manifest variables, and then use the results about models without covariates to complete the proof.

Create a new categorical variable D^* based on pairs $(D,\mathbf{x}_k)$ such that $D^* = L(w-1)+d$ if $D=d$ and $\mathbf{x}_k = \mathbf{X}_{i(w)}$, $d=0,\ldots,L-1$, $w=1,\ldots,W$. Additionally, create new categorical variables T_k^* based on pairs $(T_k,\mathbf{x}_k)$ such that, $T_k^* = J_k(w-1)+j$ if $T_i=j$ and $\mathbf{x}_k = \mathbf{X}_{i(w)}$, $k=1,2,3$, $j=1,\ldots,J_k$, $w=1,\ldots,W$. Then the marginal probability of

triplet (T_1^*, T_2^*, T_3^*) is

$$P(T_1^*, T_2^*, T_3^*) = \begin{cases} P(T_1, T_2, T_3 \mid \mathbf{X}_{i(w)}) & \text{if } \mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_3 = \mathbf{X}_{i(w)} \\ 0 & \text{Otherwise} \end{cases}.$$

Construct a finite mixture model of form (6.2) with manifest variables T_1^*, T_2^*, T_3^* and latent variable D^*. Then, based on Theorem 3, condition (i) guarantees that this new model is globally identifiable. It follows that quantities $\pi_d(\mathbf{Z}_{i(w)}^T \alpha_d)$, $g_{kjd}(\mathbf{X}_{i(w)}^T \beta_{kjd})$ in the original model are globally identifiable, $d = 0, \ldots, L-1$, $k = 1, 2, 3$, $j = 1, \ldots, J_k$, $w = 1, \ldots, W$. Moreover, since $\pi_d(\cdot)$, $g_{kjd}(\cdot)$ are pre-specified monotone link functions, the following equations hold

$$\begin{cases} \mathbf{Z}_{i(1)}^T \alpha_d = \pi_d^{-1}(a_{d1}) \\ \vdots \\ \mathbf{Z}_{i(W)}^T \alpha_d = \pi_d^{-1}(a_{dW}) \end{cases} \quad \text{and} \quad \begin{cases} \mathbf{X}_{i(1)}^T \beta_{kd} = g_{kjd}^{-1}(b_{kjd1}) \\ \vdots \\ \mathbf{X}_{i(W)}^T \beta_{kd} = g_{kjd}^{-1}(b_{kjdW}) \end{cases},$$

where a_{dw} is the value of $\pi_d(\mathbf{Z}_{i(w)}^T \alpha_d)$ and b_{kjdw} is the value of $g_{kjd}(\mathbf{X}_{i(w)}^T \beta_{kjd})$, $d = 0, \ldots, L-1$, $k = 1, 2, 3$, $j = 1, \ldots, J_k$, $w = 1, \ldots, W$. Then, condition (ii) guarantees that at least u equations on the left hand side are linearly independent, and at least v equations on the right hand side are linearly independent. As a result, α_d, β_{kjd} have unique solutions. Therefore, the finite mixture model (6.6) with $K = 3$ is globally identifiable. □

Since multiple univariate manifest variables can be combined into a single multivariate manifest variable, I have the following theorem.

Theorem 8

Suppose $K \geq 3$, α_d has $u \leq N$ elements and β_{kd} has $v \leq N$ elements. Let $max\{u, v\} \leq W \leq N$, and $\{i(1), \ldots, i(W)\}$ be a subset of $\{1, \ldots, N\}$. Let A, B, C be a partition of $\{1, \ldots, K\}$, where $A = \{a(1), \ldots, a(p)\}$, $B = \{b(1), \ldots, b(q)\}$ and $C = \{c(1), \ldots, c(r)\}$, with $p \geq 1$, $q \geq 1$,$r \geq 1$, and $p + q + r = K$. Let $M_{A[w]}$ be a L by $\prod_{i=1}^{p} J_{a(i)}$ matrix with the $(d+1, j)$ element

$$M_{A[w]}(d+1, j) = P\big((T_{a(1)}, \ldots, T_{a(p)}) = \vec{t}_{Aj} | D = d, \mathbf{x} = \mathbf{X}_{i(w)}\big),$$

where $\vec{t}_{Aj}$ is the jth possible in lexicographic order among $\prod_{i=1}^{p} J_{a(i)}$ distinct response

patterns of $(T_{a(1)}, \ldots, T_{a(p)})$. Let M_k be a block diagonal matrix defined as follows,

$$M_A = \begin{pmatrix} M_{A[1]} & 0 & \ldots & 0 \\ 0 & M_{A[2]} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M_{A[W]} \end{pmatrix}.$$

Define M_B and M_C similarly; then, the finite mixture model (6.6) is globally identifiable if the following conditions hold.

(i) $rank_K M_A + rank_K M_B + rank_K M_C \geq 2WL + 2$.

(ii) $rank\ \mathbf{X}_w \geq u$ and $rank\ \mathbf{Z}_w \geq v$. □

Observing that the row vectors in the "M" matrix are generated by multiple manifest variables is an extension group of the row vectors in the "M" matrix, as generated by any one of these manifest variables, I have the following corollary.

Corollary 6

When $K \geq 3$, if there are at least three manifest variables that satisfy the conditions in Theorem 7, then the finite mixture model (6.6) is globally identifiable. □

Moreover, for a block diagonal matrix, the row vectors that intersect with different blocks are clearly linearly independent. Therefore, the Kruskal rank of a block diagonal matrix equals the summation of the Kruskal rank of each of the block. The following corollary results.

Corollary 7

Suppose $K \geq 3$. In the same set up as in Theorem 7, if there exist $k_1, k_2, k_3 \in \{1, \ldots, K\}$ such that

(i) $rank_K M_{k_1[w]} + rank_K M_{k_2[w]} + rank_K M_{k_3[w]} \geq 2L + 2$, for all $w = 1, \ldots, W$;

(ii) $rank\ \mathbf{X}_w \geq u$ and $rank\ \mathbf{Z}_w \geq v$,

then the finite mixture model (6.6) is globally identifiable. □

Now I consider the identifiability of a finite mixture model with continuous manifest variables. Again, I assumed that the covariate effects are linear on a certain transformed

scale. Let $f_{kd}(t|\mathbf{X}_i\beta_{kd}) = P(T_k = t|D = d, \mathbf{X}_i)$ be the conditional density function of the kth manifest variable in group d and $F_{kd}(t|\mathbf{X}_i\beta_{kd})$ be the corresponding CDF, $k = 1, \ldots, K$, $D = 0, \ldots, L-1$. I further assumed that after an unknown transformation H_{kd}, manifest variable T_k satisfied

$$\big(H_{kd}(T_k) \mid \mathbf{X}_i\big) = \mathbf{X}_i\beta_{kd} + \epsilon_{ikd}, \quad \epsilon_{ikd} \sim G_{kd}(\cdot),$$

where $G_{kd}(\cdot)$ is a pre-specified distribution function with corresponding density function $g_{kd}(\cdot)$, $k = 1, \ldots, K$, $D = 0, \ldots, L-1$. The model can be expressed as follows,

$$P(T_1 = t_1, \ldots, T_K = t_k|\mathbf{X}_i, \mathbf{Z}_i) = \sum_{d=0}^{L-1} \pi_d(\mathbf{Z}_i^T\alpha_d) \prod_{k=1}^{K} f_{kd}(t_k|\mathbf{X}_i\beta_{kd}). \tag{6.8}$$

When $K = 3$ I claim the following theorem holds.

Theorem 9

Suppose α_d has $u \le N$ elements and β_{kd} has $v \le N$ elements. Let $max\{u, v+1\} \le W \le N$ and $\{i(1), \ldots, i(W)\}$ be a subset of $\{1, \ldots, N\}$. Let J_1, J_2, J_3 be some positive integers. Let $M_{k[w]}$ be a L by J_k matrix with the $(d+1, j)$ element $F_{kd}(t_{kj}|\mathbf{X}_i\beta_{kd})$, where $t_{k1}, \ldots, t_{kJ_k}$ are a set of points in the domain of $F_{kd}(t|\mathbf{X}_i\beta_{kd})$, and $\mathbf{X}_{i(w)}$ is the $i(w)$th row of the design matrix $\mathbf{X}$, $d = 0, \ldots, L-1$, $k = 1, 2, 3$, $j = 1, \ldots, J_k$, $w = 1, \ldots, W$. Let M_k be a block diagonal matrix defined as follows,

$$M_k = \begin{pmatrix} M_{k[1]} & 0 & \ldots & 0 \\ 0 & M_{k[2]} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M_{k[W]} \end{pmatrix}, \quad k = 1, 2, 3.$$

Let $\mathbf{X}_w = (\mathbf{X}_{i(1)}, \ldots, \mathbf{X}_{i(W)})^T$ and $\mathbf{Z}_w = (\mathbf{Z}_{i(1)}, \ldots, \mathbf{Z}_{i(W)})^T$. Then, finite mixture model (6.8) with $K = 3$ is globally identifiable if the following conditions hold.

(i) $rank_K M_1 + rank_K M_2 + rank_K M_3 \ge 2WL + 2$;

(ii) $rank\ \mathbf{X}_w \ge u$ and $rank\ \mathbf{Z}_w \ge v + 1$.

Proof:

Using Lemma 3 and following the same logic as in the proof of Theorem 5, I can show that for any t_k^* in the domain of $F_{kd}(t|\mathbf{X}_i\beta_{kd})$, quantities $\pi_d(\mathbf{Z}_{i(w)}^T\alpha_d)$ and $F_{kd}(t_k^*|\mathbf{X}_i\beta_{kd})$, $F_{kd}(t_{kj}|\mathbf{X}_i\beta_{kd})$ are identifiable, $d = 0,\ldots,L-1$, $k = 1,2,3$, $j = 1,\ldots,J_k$, $w = 1,\ldots,W$. Moreover, since

$$\begin{aligned} & F_{kd}(t_k^*|\mathbf{X}_i\beta_{kd}) = P(T_k \le t_k^*|\mathbf{X}_i\beta_{kd}) \\ = \;& P(H_{kd}(T_k) \le H_{kd}(t_k^*)|\mathbf{X}_i\beta_{kd}) = G_{kd}\big(H_{kd}(t_k^*) - \mathbf{X}_i\beta_{kd}\big), \end{aligned}$$

I have

$$\left\{\begin{array}{c} \mathbf{Z}_{i(1)}^T\alpha_d = \pi_d^{-1}(a_{d1}) \\ \vdots \\ \mathbf{Z}_{i(W)}^T\alpha_d = \pi_d^{-1}(a_{dW}) \end{array}\right. \quad \text{and} \quad \left\{\begin{array}{c} H_{kd}(t_k^*) - \mathbf{X}_{i(1)}^T\beta_{kd} = G_{kd}^{-1}(b_{kjd1}) \\ \vdots \\ H_{kd}(t_k^*) - \mathbf{X}_{i(W)}^T\beta_{kd} = G_{kd}^{-1}(b_{kjdW}) \end{array}\right. ,$$

where a_{dw} is the value of $\pi_d(\mathbf{Z}_{i(w)}^T\alpha_d)$ and b_{kjdw} is the value of $F_{kd}(t_k^*|\mathbf{X}_i\beta_{kd})$, $d = 0,\ldots,L-1$, $k = 1,2,3$, $j = 1,\ldots,J_k$, $w = 1,\ldots,W$. Note that the equations on the right hand side are still linear equations of $\mathbf{X}_i$, so condition (ii) guarantees that α_d, β_{kjd} and $H_{kd}(t_k^*)$ have unique solutions. Since t_k^* is arbitrary, the finite mixture model (6.8) with $K = 3$ is globally identifiable. □

Comparing Theorem 9 to Theorem 7, the additional rank of $\mathbf{Z}_w$ was used for identification of $H_{kd}(\cdot)$. Applying the same techniques as before, it was straightforward to obtain the following theorem and corollaries.

Theorem 10

Suppose that $K \ge 3$, α_d has $u \le N$ elements and β_{kd} has $v \le N$ elements. Let $max\{u, v+1\} \le W \le N$, and $\{i(1),\ldots,i(W)\}$ be a subset of $\{1,\ldots,N\}$. Let J_1, J_2, J_3 be some positive integers and $t_{k1},\ldots,t_{kJ_k}$ be a set of points in the domain of $F_{kd}(t|\mathbf{X}_i\beta_{kd})$, $k = 1,\ldots,K$. Let A, B, C be a partition of $\{1,\ldots,K\}$, where $A = \{a(1),\ldots,a(p)\}$, $B = \{b(1),\ldots,b(q)\}$ and $C = \{c(1),\ldots,c(r)\}$, with $p \ge 1$, $q \ge 1$,$r \ge 1$, and $p+q+r = K$. Let $M_{A[w]}$ be a L by $\prod_{i=1}^p J_{a(i)}$ matrix with the $(d+1, j)$ element $\mathbb{F}_{kd}(\vec{t}_{Aj}|\mathbf{X}_i\beta_{kd})$, where $\mathbb{F}_{kd}(\cdot|\mathbf{X}_i\beta_{kd})$ is the CDF of $(T_{a(1)},\ldots,T_{a(p)})$ conditional on $\mathbf{X}_i$, and $\vec{t}_{Aj}$ is the jth possible in lexicographic order among $\prod_{i=1}^p J_{a(i)}$ distinct response patterns of $(T_{a(1)},\ldots,T_{a(p)})$, generated by point

$t_{a(s)1}, \ldots, t_{a(s)J_{a(s)}}$, $s = 1, \ldots, p$. Let M_k be a block diagonal matrix defined as follows,

$$M_A = \begin{pmatrix} M_{A[1]} & 0 & \ldots & 0 \\ 0 & M_{A[2]} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M_{A[W]} \end{pmatrix}.$$

Define M_B and M_C similarly; then the finite mixture model (6.8) is globally identifiable if the following conditions hold.

(i) $rank_K M_A + rank_K M_B + rank_K M_C \geq 2WL + 2$.

(ii) $rank\ \mathbf{X}_w \geq u$ and $rank\ \mathbf{Z}_w \geq v + 1$. □

Corollary 8

When $K \geq 3$, if there are at least three manifest variables that satisfy the conditions in Theorem 9, then the finite mixture model (6.8) is globally identifiable. □

Corollary 9

Suppose $K \geq 3$. In the same set-up as used in Theorem 9, if there exist $k_1, k_2, k_3 \in \{1, \ldots, K\}$, such that

(i) $rank_K M_{k_1[w]} + rank_K M_{k_2[w]} + rank_K M_{k_3[w]} \geq 2L + 2$, for all $w = 1, \ldots, W$;

(ii) $rank\ \mathbf{X}_w \geq u$ and $rank\ \mathbf{Z}_w \geq v + 1$,

then the finite mixture model (6.8) is globally identifiable. □

6.5 Summary

In this chapter I provided some conditions under which a finite mixture model is locally identifiable or globally identifiable. The results can be applied to a wide range of models, including, but not restricted to, all of the models discussed in previous chapters. Specifically, I studied models with categorical manifest variables, models with continuous manifest variables, and both types of models when covariates are included. One of the assumptions that I made here is that the covariate effects are linear on some transformed scales. This assumption is common in the literature on finite mixture models. However, this transforma-

tion may be unknown in some of the models discussed here. Additionally, I did not impose any constraints about the covariate effects among different groups or on different manifest variables. Therefore, the model considered here represents many general situations.

For local identifiability, a key idea in the proofs was to show that the Jacobian matrix of the model induced function had full column rank. The results suggest that, contrary to common belief, including covariates in the model may in fact help model identification. Additionally, by considering a continuous distribution in its empirical form for a given data set, I unified the proofs for models with categorical manifest variables or with continuous manifest variables, and showed nonparametric identifiability of the models. Consequently, the results can also be applied to cases when manifest variables have mixed types.

For global identifiability, the proof is established on a previous result about the uniqueness of trilinear decomposition (Kruskal, 1977). I only discussed the situation where the number of manifest variables K is greater than or equal to 3. When $K = 2$, Hall and Zhou (2003) showed that the model is not nonparametrically identifiable. The proof for models with continuous manifest variables used the results about models with categorical variables, and was accomplished by showing that the CDF was globally identifiable at every point in its domain. The results easily extend to models with mixed types of manifest variables.

Chapter 7

SUMMARY AND DISCUSSION

7.1 Summary of the Dissertation

7.1.1 Summary of the Dissertation

This dissertation focused on the evaluation of diagnostic and prognostic tests when gold standard information was not available. The term "test" was used quite generally here – it referred to any technique that provides information used for detecting, diagnosing, or monitoring any medical condition or event of interest, such as radiologists' readings, biomarker values, questionnaire scores, and combinations of these measures. The models studied here belong to the latent class and latent profile framework and can easily handle situations where the unobserved gold standard has an ordinal scale.

One classic assumption in the use of latent variable models is the conditional independence assumption–that the observed variables (tests) are independent conditional on the latent variable (disease status). I proposed two approaches to relaxing this assumption: the first approach used random effects to summarize the correlations among the tests that were caused by some unobserved factors; the second approach utilized observed covariate information to explicitly explain the sources and magnitudes of these correlations. The proposed model with covariates is very flexible; both covariates that affect disease prevalence and covariates that influence test performance can be included in the model. Additionally, there are no constraints on the covariate effect. Therefore, the model can be applied to many research questions, such as studying risk factors for an unobserved medical condition and understanding a test's performance for individuals with different characteristics.

Another assumption that is often made by latent variable models is that the observed variables within each latent group follow a pre-specified parametric distribution, such as a normal distribution. I proposed models that relax this assumption. When tests were categorical, their conditional distributions within each disease status were modeled non-

parametrically. When tests were continuous I introduced a transformation model to allow for more general shapes of the test result distributions.

The flexibility of latent variable models may lead to serious identifiability issues. However, this problem has not been well studied and is often overlooked in applications of these models. An important goal of this dissertation is to emphasize this potential problem and to provide conditions that guarantee local identifiability and global identifiability of the models. To my knowledge, it is the first attempt to establish identifiability conditions for latent class and latent profile models in a general form, and the possible inclusion of covariates. The results provide a theoretical justification of the proposed methods and can be used in practice to facilitate valid study results and interpretations.

The proposed models are presented in the situation where the latent disease status is ordinal. However, this assumption is only used to identify a unique solution among several equivalent ones caused by the "label switching" problem in latent class and latent profile models. There is no additional constraint imposed for the estimation. Therefore, this assumption is not essential if there are other pieces of information that can help to determine the labeling of the latent groups – for example, the group labels can be determined if the sizes of the latent groups are known. As a result, the proposed methods can be applied to situations where the latent disease status is nominal, as long as additional information is available to label the groups identified by the models. For methods that further explore the ordinal nature of the data, one may refer to Rost (1988, 1991), who assumed the latent classes corresponding to an uni-dimensional score on a latent continuum, and Croon (1990, 1991), who imposed inequality constraints that $P(T \geq t|D = d_1) \geq P(T \geq t|D = d_2), \forall d_1 > d_2$ to ensure the ordinality of the classes.

Additionally, the proposed methods provide a way to summarize information from multiple tests simultaneously. These models essentially use the observed variable to infer the most likely value of the unobserved gold standard. Therefore, in addition to diagnostic test evaluation, these methods can be used for marker combination and for obtaining potentially better diagnoses when results from multiple tests are available as discussed in Chapter 5.

When using latent class and latent profile models, one should realize that all these models merely offer a means to "cluster" the population into groups in which the manifest variables

(tests) are relatively homogenous. They cannot directly identify disease groups *per se*. In this dissertation, I implicitly assumed that the tests weremdesigned to reflect disease status rather than some other characteristics among the subjects, so that disease status was the fundamental factor that lead to the heterogeneity of the test results. However, one should carefully evaluate this assumption in each application.

I did not discuss problems related to assessing the number of components in latent variable models, since in a diagnostic testing study, the number of components is usually apparent. Moreover, the components usually have specific medical interpretations, and thus, the number of them should be determined by the study goal – for example, do the researchers want to assess a test's ability to identify diseased and non-disease subjects, or do they want to assess its ability to distinguish among severity levels for a certain condition? Nevertheless, in cases where the number of components cannot be determined by the scientific question itself, the commonly used likelihood criteria such as AIC and BIC appear to sufficiently accomplish this task in most situations (Leroux, 1992; Biernacki, et al., 1998; Solka, et al., 1998). Additional discussion on this issue and on other techniques for determining the latent component number can be found in Chapter 6 of McLachlan and Peel, 2000.

Aside from the potential computational burden, it is relatively easy and perhaps tempting to build more complicated structures in latent variable models – for example, introducing multiple levels of random effects (e.g. hospital level, test level, subject level) in addition to covariates to further account for possible dependence among the tests. I did not explore this direction because in my opinion, when dealing with an unobserved latent structure, many modeling assumptions can hardly be verified. The possible danger of model misspecification or over-fitting results from building a complicated model may compromise its potential gains relative to using a simpler model. A simple but adequate model is preferable, since it can be more readily interpreted. If a more complicated model were to be used, careful examination of the identifiability issue would be necessary.

7.2 Discussion on Potential Future Directions

Some relatively direct extensions of the proposed methods have been discussed in previous chapters, such as replacing the Box-Cox transformation in the latent profile model proposed

in Chapter 6 with a semiparametric transformation. I will not repeat them here, but discuss some additional topics for future research.

7.2.1 Partially Verified Gold Standard

One interesting direction is to connect the work on absent gold standards to methods that deal with verification bias. The models discussed in this dissertation assume that gold standard information is completely unavailable. However, sometimes, even for a hard to obtain gold standard, a small portion of the subjects that are verified may be available, especially among those with severe conditions. Another scenario is one in which subjects with extremely severe conditions or with some specific characteristics are unlikely to be misdiagnosed by some or all of the tests considered. Incorporating this information can potentially improve the accuracy and efficiency of the assessment. These models can still fit into the finite mixture model framework, as they add an additional component for the verified subgroup into the mixture. However, techniques such as the ones used in a verification bias problem need to be adopted, since the verified subgroup is usually different from the remaining population.

7.2.2 Pseudo Gold Standard Test

Sometimes gold standard information on any of the subjects does not exist; however, it may still be known that one or more of the tests have a higher accuracy than the others. For example, the Consortium to Establish a Registry for Alzheimer Disease (CERAD) criteria is often used as a gold standard in many AD studies when autopsy data are not available. It has higher accuracy for identifying AD than some other criteria, such as MMSE, CDR or biomarker values. This information can also be utilized to improve model performance. A Bayesian approach with appropriate priors is an option. In a frequentist method, employing different weights is another choice.

7.2.3 Missing Values and Ceiling/Floor Effects

The proposed methods assumed that, except for the latent disease status and possibly some random effects, all other variables in the model had no missing values. However, real data are rarely complete. Exploring the advantages and disadvantages of different missing data techniques in this particular framework can provide valuable guidance for practice. A special case of a missing value type is ceiling or floor effects. Many diagnostic tests and especially biomarkers can only detect variations within a certain range. For example, one of the neuropsychological tests for AD detection is trail making; it measures the time that a subject takes to complete a visual task, such as connecting 25 dots in order. If a subject cannot complete the task in a certain time, this maximum allowance time will be assigned. However, the "true" completion time of this subject is higher than the maximum value, and two subjects with this same maximum value may have different completion times if they were followed long enough. A similar problem is common in using biomarker measurements, as most assays or instruments have a finite detectable range. Examining the impact of this type of missing data on the proposed methods, and searching for possible ways to address it, poses another interesting question for future research.

7.2.4 Longitudinal Approaches

Still another interesting direction that research could consider is the study of longitudinal aspects of diagnostic tests and biomarker assessments. With growing interest in, and emphasis on, the preclinical stage and prevention of many diseases, many diagnostic tests are not only used to ascertain a subject's current condition, but also to predict a future event. Using a time-to-event variable, as in survival analysis, instead of a cross-sectional gold standard to reflect the changes of the condition over time may be more attractive for this purpose. Moreover, researchers may increasingly have repeated test results for the same individual over time because of advances in electronic medical records. Longitudinal information, such as the temporal trend in these repeated measures, can be informative for making diagnoses. Mixture growth curve analysis offers one direction for this topic. Another direction is to view disease progress as a continuous process and to use, for example, a

hidden Markov chain model to represent the time-dependent disease progression. Difficulties include developing a model that will allow for non-identically distributed measures at each time point.

7.2.5 Improving Prediction

In addition to evaluating the accuracies of different diagnostic tests, another important question in diagnostic medicine is how best to combine available information for a better prediction. As mentioned before, model-based posterior risks provide one way of combining different tests. However, many questions remain, such as what the statistical property of a prediction based on these posterior risks is, in what sense the resulting prediction is better or worse than that given by a method that first infers the latent gold standard then applies prediction techniques developed with the gold standard present, such as the optimal risk score method proposed by McIntosh and Pepe (2002), and how to test and quantify the prediction improvement by adding an additional test or through other means.

BIBLIOGRAPHY

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6):716–723.

[2] Albert, P. S. and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 60:427–435.

[3] Albert, P. S., McShane, L. M., Shih, J. H., et al. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* 57(2):610–619.

[4] Allen, J. J., Schnyer, R. N., Chambers, A. S., et al. (2006). Acupuncture for depression: a randomized controlled trial. *J Clin Psychiatry* 67(11):1665–73.

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

[6] Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* 92:1375–1386.

[7] Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* 49:803–821.

[8] Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. John Wiley and Sons.

[9] Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology* 28(1):97–104.

[10] Bateman, R. J., Xiong, C., Benzinger, T. L. S., Fagan, A. M., Goate, A., Fox, N. C., Marcus, D. S., Cairns, N. J., Xie, X., Blazey, T. M., Holtzman, D. M., Santacruz, A., Buckles, V., Oliver, A., Moulder, K., Aisen, P. S., Ghetti, B., Klunk, W. E., McDade, E., Martins, R. N., Masters, C. L., Mayeux, R., Ringman, J. M., Rossor, M. N., Schofield, P. R., Sperling, R. A., Salloway, S., and Morris, J. C. (2012). Clinical and biomarker changes in dominantly inherited alzheimer's disease. *N Engl J Med* 367(9):795–804.

[11] Begg, C. B. and Metz, C. E. (1990). Consensus diagnoses and "gold standards". *Medical Decision Making* 10:29–30.

[12] Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997). Inference in model based clustering. *Statistics and Computing* 7:1–10.

[13] Biernacki, C., Celeux, G., and Govaert, G. (1998). Assessing a mixture model for clustering with the integrated classification likelihood. Technical Report No. 3521 Rhône-Alpes: INRIA.

[14] Binder, L. I., Guillozet-Bongaarts, A. L., Garcia-Sierra, F., and Berry, R. W. (2005). Tau, tangles, and alzheimer's disease. *BiochimBiophysActa* 1739(2-3):216–23.

[15] Bolck, A., Croon, M. A., and Hagenaars, J. A. (2004). Estimating latent structure models with categorical variables: one-step versus three-step estimators. *Political Analysis* 12(1):3–27.

[16] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Royal Statistical Society. Series B* 26(2):211–252.

[17] Branscum, A. J., Gardner, I. A., and Johnson, W. O. (2005). Estimation of diagnostic test sensitivity and specificity through bayesian modeling. *Preventive Veterinary Medicine* 68:145–163.

[18] Cheeseman, P. and Stutz, J. (1996). Bayesian classification (autoclass): theory and results. In *Advances in knowledge discovery and data mining*. The AAAI press.

[19] Cheng, R. C. H. and Traylor, L. (1995). Non-regular maximum likelihood problems (with discussion). *Journal of the Royal Statistical Society B* 57:3–44.

[20] Chi, Y. Y. and Zhou, X. H. (2010). Roc surfaces in the presence of verification bias. *Journal of Royal Statistical Association, Series C, Applied Statistics* 57(1):1–23.

[21] Choi, Y. K., Johnson, W. O., Collins, M. T., and Gardner, I. A. (2006). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics* 11(2):210–229.

[22] Clogg, C. C. and Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency table. *Jounal of the American Statistical Association* 79:762–771.

[23] Croon, M. A. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology* 43:171–192.

[24] Croon, M. A. (1991). Investigating mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology* 44:315–331.

[25] Cruz-Medina, I. R., Hettmansperger, T. P., and Thomas, H. (2004). Semiparametric mixture models and repeated measures: the multinomial cut point model. *Royal Statistical Society: Series C(Applied Statistics)* 53:463–474.

[26] Cummings, J. L. (2004). Alzheimer's disease. *N Engl J Med* 351:56–67.

[27] Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics* 28(1):20–28.

[28] Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association* 83:173–178.

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

[30] Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 57(1):158–167.

[31] Dillon, W. R. and Mulani, N. (1989). LADI: A latent discriminant model for analyzing marketing research data. *Journal of Marketing Research* 26(1):15–29.

[32] Dunn, G. (1989). *The Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors*. London: Edward Arnold.

[33] Eisenberg, D. M., Cohen, M. H., Hrbek, A., et al. (2002). Credentialing complementary and alternative medical providers. *Annals of Internal Medicine* 137(12):965–973.

[34] Elmore, R., Hall, P., and Neeman, A. (2005). An application of classical invariant theory to identifiability in nonparametric mixtures. *Annales de l'institut Fourier* 55(1):1–28.

[35] Engelborghs, S., Vreese, K. D., de Casteele, T. V., et al. (2008). Diagnostic performance of a CSF-biomarker panel in autopsy-confirmed dementia. *Neurobiol Aging* 29:1143–1159.

[36] Ernst, E. (2006). Methodological aspects of traditional chinese medicine (TCM). *Annals, Academy of Medicine, Singapore* 35(11):773–4.

[37] Espeland, M. A. and Handelman, S. L. (1989). Using latent class models to characterize and assess relative error in discrete masurements. *Biometrics* 45:587–599.

[38] Everitt, B. S. (1993). *Cluster analysis*. London: Edward Arnold 3rd edition.

[39] Fagan, A. M., Roe, C. M., Xiong, C., et al. (2007). Cerebrospinal fluid tau/-amyloid42 ratio as a prediction of cognitive decline in nondemented older adults. *Arch Neurol* 64:343–349.

[40] Fawcett, T. (2001). Using rule sets to maximize roc performance. In *Proceedings IEEE International Conference on Data Mining* pages 131–138 San Jose, CA. IEEE Computer Society. Print ISBN: 0-7695-1119-8.

[41] Feng, Y., Wu, Z., Zhou, X., Zhou, Z., and Fan, W. (2006). Methodological review: Knowledge discovery in traditional chinese medicine: State of the art and perspectives. *Arificial Intelligence in Medicine* 38(3):219–236.

[42] Ferri, C., Hernndez-Orallo, J., and Salido, M. A. (2003). Volume under the roc surface for multi-class problems. In *14th European Conference on Machine Learning:ECML 2003* volume 2837 pages 108–120. LNAI Springer Verlag.

[43] Formann, A. K. (1985). Constrained latent class models: theory and applications. *British Jounal of Mathematical and Statistical Psychology* 38:87–111.

[44] Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association* 87:476–486.

[45] Geoffrey, J., Wesley, O. J., Timothy, E. H., and Ronald, C. (2010). Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics* 66(3):855–863.

[46] Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variable are unobserved, Part I: A modified latent structure approach. *American Journal of Sociology* 79:1179–1259.

[47] Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61(2):215–231.

[48] Gravina, S. A. et al. (1995). Amyloid beta protein (abeta) in alzheimer's disease brain. *Biological Chemstry* 270:7013–7016.

[49] Haberman, S. J. (1979). *Analysis of qualitative data* volume 2. Academic Press, NewYork.

[50] Hadgu, A. and Qu, Y. (1998). A biomedical application of latent class models with random effects. *Applied Statistics* 47(4):603–616.

[51] Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: local dependence models. *Sociological Methods and Research* 16(3):379–405.

[52] Hall, P., Neeman, A., Pakyari, R., and Elmore, R. (2005). Nonparametric inference in multivariate mixture. *Biometrika* 92:667–678.

[53] Hall, P. and Zhou, X. H. (2003). Nonparametric estimation of component distribution in a multivariate mixture. *Annals of Statistics* 31(1):201–224.

[54] Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning* 45(2):171–186.

[55] Hansson, O., Zetterberg, H., Buchhave, P., Londos, E., Blennow, K., and Minthon, L. (2006). Association between CSF biomarkers and incipient alzheimer's disease in patients with mild cognitive impairment: a follow-up study. *Lancet Neurol* 5(3):228–234.

[56] Hardy, J. and Selkoe, D. J. (2002). The amyloid hypothesis of alzheimer's disease: progress and problems on the road to therapeutics. *Science* 297(5580):353–6.

[57] Hebert, L. E., Scherr, P. A., Bienias, J. L., Bennett, D. A., and Evans, D. A. (2003). Alzheimer disease in the us population: prevalence estimates using the 2000 census. *Arch Neurol* 60:1119–22.

[58] Henkelman, R. M., Kay, I., and Bronskill, M. J. (1990). Receiver operator characteristic (ROC) analysis without truth. *Medical Decision Making* 10(1):24–29.

[59] Hettmansperger, T. P. and Thomas, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society: Series B(Statistical Methodology)* 62:811–825.

[60] Huang, G. H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika* 69(1):5–32.

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

[62] Hui, S. L. and Zhou, X. H. (1998). Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* 7(4):354–370.

[63] Johnson, W. O. and Hanson, T. E. (2005). Comment on "on model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables". *Statistical Science* 20(2):111–140.

[64] Jorm, A. F. (1991). Cross-national comparisons of the occurrence of alzheimer's and vascular dementias. *Eur Arch Psychiatry ClinNeurosci* 240:218–22.

[65] Jr, C. R. J., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *Lancet Neurol* 9:119–28.

[66] Kaptchuk, T. J. (2001). *The Web that has no Weaver: understanding Chinese Medicine*. McGraw-Hill Professional. Chicago, IL.

[67] Kasahara, H. and Shimotsu, K. (2008). Nonparametric identification and estimation of multivariate mixtures. Unpublished manuscript Queen's University Working Papers.

[68] Langeheine, R., Pannekoek, J., and de Pol, F. V. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods Research* 24(4):492–516.

[69] Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. New York: Houghton Mifflin.

[70] Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annal of Statistics* 20:1350–1360.

[71] Li, G., Sokal, I., Quinn, J. F., et al. (2007). CSF tau/Abeta 42 ratio for increased risk of mild cognitive impairment: a follow up study. *Neurology* 69:631–639.

[72] Magidson, J. and Vermunt, J. K. (2002). Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing Research* 20:37–44.

[73] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. New York: Chapman and Hall 2nd edition.

[74] McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* 21(4):331–347.

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

[76] McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and Application to Clustering*. New York: Marcel Dekker.

[77] McLachlan, G. J. and Peel, D. (1996). An algorithm for unsupervised learning via normal mixture models. In *Information, statistics and induction in science*. edited by Dowe, D. L., Korb, K. B. and Oliver, J. J. Singapore: World Scientific Publishing.

[78] McLachlan, G. J. and Peel, D. (1999). Modelling nonlinearity by mixtures of factor analysers via extension of the EM algorithm. Technical report Australia: Center for Statistics, University of Queensland.

[79] Melton, B., Liang, K. Y., and Pulver, A. E. (1994). Extended latent class approach to the study of familial/sporadic forms of a disease: Its application to the study of the heterogeneity of schizophrenia. *Genetic Epidemiology* 11(4):311–327.

[80] Meyer, G. D., Shapiro, F., Vanderstichele, H., Vanmechelen, E., Engelborghs, S., Deyn, P. P. D., Coart, E., Hansson, O., Minthon, L., Zetterberg, H., Blennow, K., Shaw, L., and Trojanowski, J. (2010). Alzheimer's disease neuroimaging initiative. Diagnosis-independent alzheimer disease biomarker signature in cognitively normal elderly people. *Arch Neurol* 67(8):949–56.

[81] Molenaar, P. C. M. (2007). Psychological methodology will change profoundly due to the necessity to focus on intra-individual variation. *Integrative Psychological and Behavioral Science* 41(1):35–40.

[82] Molenaar, P. C. M. and von Eye, A. (1994). On the arbitrary nature of latent variables. In *Analysis of latent variables in developmental research* pages 226–242. Newbury Park, CA: Sage.

[83] Mossman, D. (1999). Three-way ROCs. *Medical Decision Making* 19(1):78–89.

[84] Munch, S. (2004). Gender-biased diagnosing of women's medical complaints: Contributions of feminist thought. *Women and Health* 40(1):1970–1995.

[85] Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika* 29(1):81–117.

[86] Muthén, B. O. (2003). Statistical and substantive checking in growth mixture modeling: comment on Bauer and Curran. *Psychological Methods* 8(3):369–377.

[87] Nakas, C. T. and Alonzo, T. A. (2007). ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. *Biometrics* 63(2):603–609.

[88] Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class roc analysis with continuous measurements. *Statistical in Medicine* 23(22):3437–3449.

[89] O'Brien, K. A. and Xue, C. C. (2003). The theoretical framework of chinese medicine. In *Leung PC, Xue CC, Cheng YC, eds. A Comprehensive Guide to Chinese Medicine.* River Edge, NJ: World Scientific Publishing Co.

[90] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Predication.* New York: Oxford Univerisity Press 1st edition.

[91] Peters, B. C. and Walker, H. F. (1978). An iterative procedure for obtaining maximum likelihood estimators of the parameters for a mixture of normal distributions. *SIAM Journal on Applied Mathematics* 35:362–378.

[92] Petersen, J., Bandeen-Roche, K., Budtz-Jørgensen, E., and Larsen, K. G. (2012). Predicting latent class scores for subsequent analysis. *Psychometrika* 77(2):244–262.

[93] Price, J. L. and Morris, J. C. (1999). Tangles and plaques in nondemented aging and "preclinical" alzheimer's disease. *Ann Neurol* 45:358–68.

[94] Qu, Y. and Hadgu, A. (1998). A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association* 93(443):920–928.

[95] Qu, Y., Tan, M., and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuary of diagnostic tests. *Biometrics* 52(3):797–810.

[96] Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review* 26:195–239.

[97] Rost, J. (1988). Rating scale analysis with latent class models. *Psychometrika* 53:327–348.

[98] Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology* 44:7592.

[99] Ruspini, E. H. (1969). A new approach to clustering. *Information and Control* 15:22–32.

[100] Sain, S. R., Gray, H. L., Woodward, W. A., and Fisk, M. D. (1999). Outlier detection from a mixture distribution when training data are unlabeled. *Bulletin of the Seismological Society of America* 89:294–304.

[101] Schoonenboom, N. S. M., Pijnenburg, Y. A. L., Mulder, C., Rosso, S. M., Elk, E. J. V., Kamp, G. J. V., Swieten, J. C. V., and Scheltens, P. (2004). Amyloid beta (1-42) and phosphorylated tau in CSF as markers for early-onset Alzheimer disease. *Neurology* 62(9):1580–1584.

[102] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2):461–464.

[103] Seidel, W., Mosler, K., and Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics* 52(3):481–487.

[104] Self, S. G. and Liang, K. Y. (1987). Asympototic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82(398):605–610.

[105] Selkoe, D. J. (1991). The molecular pathology of alzheimer's disease. *Neuron* 6:487–498.

[106] Shang, A., Huwiler, K., Nartey, L., et al. (2007). Placebo-controlled trials of chinese herbal medicine and conventional medicine comparative study. *International Journal of Epidemiology* 36(5):1086–1092.

[107] Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., and et al.; Alzheimer's Disease Neuroimaging Initiative (2009). Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects. *Ann Neurol* 65(4):403–413.

[108] Skrondal, A. and Laake, P. (2001). Regression among factor scores. *Psychometrika* 66(4):563–575.

[109] Solka, J. L., Wegman, E. J., Priebe, C. E., Poston, W. L., and Rogers, W. (1998). Mixture structure analysis using the Akaike criterion and the bootstrap. *Statistics and Computing* 8:177–188.

[110] Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jr, J. C., Kaye, J., Montine, T. J., Park, D. C., Reiman, E. M., Rowe, C. C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M. C., Thies, B., Morrison-Bogorad, M., Wagster, M. V., and Phelps, C. H. (2011). Toward defining the preclinical stages of alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimers Dement* 7(3):280–92.

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

[112] Thomson, G. H. (1938). Methods of estimating mental factors. *Nature* 141(3562):246.

[113] Toomela, A. (2007). Culture of science: Strange history of the methodological thinking in psychology. *Integrative Psychological and Behavioral Science* 41(1):6–20.

[114] Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association* 88(422):421–427.

[115] Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* 41(4):959–968.

[116] Walter, S. D. (1984). Measuring the reliability of clinical data: the case for using three observers. *Revue d'Epidemiologie et de Sante Publique* 32(3-4):206–11.

[117] Walter, S. D. and Irwig, L. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology* 41(9):923–37.

[118] Wang, S. J., Woodward, W. A., Gray, H. L., Wiechecki, S., and Sain, S. R. (1997). A new test for outlier detection from a multivariate mixture distribution. *Journal of Computational and Graphical Statistics* 6:285–299.

[119] Xie, Y., Chen, Z., and Albert, P. (2013). *A crossed random effects modeling approach for estimating diagnostic accuracy from ordinal ratings without a gold standard.* Statistics in Medicine. DOI: 10.1002/sim.5784.

[120] Zarembka, P. (1968). Functional form in the demand for money. *American Statistical Association* 6(3):502–511.

[121] Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research* 5(6):697–723.

[122] Zhang, N. L., Yuan, S., Tao, C., and Wang, Y. (2007). Hierarchical latent class models and statistical foundation for traditional chinese medicine. *Artificial Intelligence in Medicine* 4594:139–143.

[123] Zhou, X. H., Castelluccio, P., and Zhou, C. (2005). Nonparametric estimation of ROC curves in the absence of a gold standard. *Biometrics* 61:600–609.

Appendix A

THE EQUIVALENCE OF THE OVERALL ACCURACY INDICATOR AND THE AUC IN TWO DIMENSION SITUATIONS

If there are only two disease statuses, without loss of generality, suppose that $J = 3$. The overall accuracy measure V, defined by averaging the correct diagnoses, or TPRs, over all possible choices of cutpoints was,

$$
\begin{aligned}
V &= \underset{0=j_0\le j_1\le j_2=J}{\text{aver}} \Big\{ \prod_{d=0}^{1} TPR(d) \Big\} = \underset{0=j_0\le j_1\le j_2=J}{\text{aver}} \Big\{ \prod_{d=0}^{1} \sum_{j=j_d}^{j_{d+1}-1} \phi_{dj} \Big\} \\
&= \lim_{\triangle t\to 0} \Big[\big(\sum_{j=0}^{3} \phi_{1j}\big)\triangle t \times \frac{\phi_{00}}{\triangle t} + \big(\sum_{j=1}^{3} \phi_{1j}\big)\triangle t \times \frac{\sum_{j=0}^{1}\phi_{0j} - \phi_{00}}{\triangle t} \\
&\quad + \big(\sum_{j=2}^{3} \phi_{1j}\big)\triangle t \times \frac{\sum_{j=0}^{2}\phi_{0j} - \sum_{j=0}^{1}\phi_{0j}}{\triangle t} + \phi_{13}\triangle t \times \frac{\sum_{j=0}^{3}\phi_{0j} - \sum_{j=0}^{2}\phi_{0j}}{\triangle t} \Big] \\
&= \phi_{00}\sum_{j=0}^{3}\phi_{1j} + \phi_{01}\sum_{j=1}^{3}\phi_{1j} + \phi_{02}\sum_{j=2}^{3}\phi_{1j} + \phi_{03}\phi_{1j} \\
&= \text{Area under the terrace-shaped ROC curve.}
\end{aligned}
$$

Appendix B

SIMULATION RESULTS FOR MODEL IN CHAPTER 3 WHEN PARAMETERS WERE ON THE BOUNDARY

Table B.1: Results from 5,000 simulations with patient number N=34, and L=3, K=5, J=3 on the boundary of the parameter space where some of the diagnostic probabilities are zero.

True prevalence	Statistics	V_1	V_2	V_3	V_4	V_5	ϕ_{dkj}		
rates	True values	0.356	0.408	0.505	0.606	0.703	max	min	mean
p_0=0.25,p_1=0.25	Bias	-0.036	-0.011	-0.020	-0.029	-0.022	0.206	0.001	0.043
p_2=0.25,p_3=0.25	MSE	0.029	0.024	0.036	0.032	0.020	0.084	0.000	0.026
p_0=0.10,p_1=0.20	Bias	-0.030	-0.019	-0.031	-0.023	-0.021	0.209	0.001	0.046
p_2=0.30,p_3=0.40	MSE	0.033	0.027	0.030	0.029	0.018	0.100	0.000	0.028
p_0=0.25,p_1=0.10	Bias	-0.015	-0.012	-0.004	-0.008	-0.065	0.211	0.000	0.037
p_2=0.15,p_3=0.50	MSE	0.026	0.021	0.021	0.022	0.026	0.067	0.000	0.024

Appendix C

RESULTS OF FIVE TCM DOCTORS BASED ON THE LATENT CLASS MODEL IN CHAPTER 3

Table C.1: Results of five TCM doctors' diagnostic performances for all twelve symptoms.

Symptoms		**Accuracy summary measure**				
		Doctor1	**Doctor2**	**Doctor3**	**Doctor4**	**Doctor5**
Symptom1	Estimation	0.550	0.650	0.906	0.786	0.389
	SE*	0.045	0.043	0.015	0.026	0.037
Symptom2	Estimation	0.627	0.449	0.532	0.600	0.482
	SE	0.044	0.067	0.068	0.070	0.059
Symptom3	Estimation	0.492	0.564	0.596	0.917	0.477
	SE	0.119	0.124	0.128	0.007	0.085
Symptom4	Estimation	0.932	0.537	0.846	0.759	0.704
	SE	0.050	0.030	0.031	0.025	0.042
Symptom5	Estimation	0.851	0.578	0.839	0.213	0.536
	SE	0.028	0.130	0.025	0.141	0.096
Symptom6	Estimation	0.448	0.579	0.627	0.444	0.534
	SE	0.033	0.044	0.069	0.073	0.048
Symptom7	Estimation	0.394	0.571	0.801	0.626	0.532
	SE	0.049	0.053	0.031	0.057	0.024
Symptom8	Estimation	0.548	0.384	0.646	0.663	0.173
	SE	0.048	0.057	0.050	0.075	0.025
Symptom9	Estimation	0.694	0.570	0.833	0.607	0.667
	SE	0.033	0.032	0.025	0.038	0.025
Symptom10	Estimation	0.305	0.429	0.534	0.754	0.402
	SE	0.044	0.060	0.028	0.030	0.057
Symptom11	Estimation	0.381	0.677	0.825	0.476	0.629
	SE	0.032	0.030	0.017	0.070	0.033
Symptom12	Estimation	0.597	0.342	0.749	0.846	0.337
	SE	0.062	0.030	0.053	0.047	0.035

*Standard error estimate was based on a bootstrap method.

Appendix D

TRUE DIAGNOSTIC PROBABILITY MATRIX IN CHAPTER 4'S SIMULATIONS

Table D.1: True diagnostic probability matrix in Chapter 4's simulations.

		$T_k = 0$	$T_k = 1$	$T_k = 2$	$T_k = 3$
Doctor 1	$D = 0$	0.40	0.30	0.20	0.10
	$D = 1$	0.20	0.50	0.20	0.10
	$D = 2$	0.10	0.20	0.50	0.20
	$D = 3$	0.10	0.20	0.30	0.40
Doctor 2	$D = 0$	0.50	0.20	0.20	0.10
	$D = 1$	0.20	0.50	0.20	0.10
	$D = 2$	0.10	0.20	0.50	0.20
	$D = 3$	0.10	0.15	0.20	0.55
Doctor 3	$D = 0$	0.60	0.20	0.15	0.05
	$D = 1$	0.20	0.53	0.20	0.07
	$D = 2$	0.10	0.20	0.50	0.20
	$D = 3$	0.05	0.15	0.20	0.60
Doctor 4	$D = 0$	0.68	0.20	0.10	0.02
	$D = 1$	0.15	0.65	0.15	0.05
	$D = 2$	0.05	0.15	0.65	0.15
	$D = 3$	0.05	0.15	0.20	0.60
Doctor 5	$D = 0$	0.75	0.10	0.10	0.05
	$D = 1$	0.15	0.75	0.08	0.02
	$D = 2$	0.05	0.10	0.75	0.10
	$D = 3$	0.05	0.05	0.10	0.80

Appendix E

EM ALGORITHM FOR THE LATENT PROFILE MODEL IN CHAPTER 5

First, I considered the homoscedastic error distribution in the measurement model $\epsilon_{ik} \sim^{i.i.d} N(0, \sigma_k^2)$.

In the E step, I computed the expected value of the complete data log likelihood,

$$\begin{aligned} P_i^{(t)}(d) &= \frac{P(\vec{T}_i|D_i = d, X_i)P(D_i = d|Z_i)}{\sum_{d=0}^{L-1} P(\vec{T}_i|D_i = d, X_i)P(D_i = d|Z_i)} \\ &= \frac{\big[\prod_{k=1}^{K} \phi\big(\frac{H_k^{(t)}(t_{ik}) - \vec{x}_i^T\beta_{kd}^{(t)}}{\sigma_k^{(t)}}\big)\big]\eta_d(\vec{z}_i^T\alpha^{(t)})}{\sum_{d=0}^{L-1}\big\{\prod_{k=1}^{K}\big[\phi\big(\frac{H_k^{(t)}(t_{ik}) - \vec{x}_i^T\beta_{kd}^{(t)}}{\sigma_k^{(t)}}\big)\big]\eta_d(\vec{z}_i^T\alpha^{(t)})\big\}}, \end{aligned}$$

where $\phi(\cdot)$ is the standard normal density function.

In the M step,

$$\{\alpha_d^{(t+1)}\}_{d=0,\ldots,L-1} = \text{argmax } l_1(\alpha_d^{(t)})$$

was obtained by performing a polytomous regression with outcome $P_i^{(t)}(d)$, $i = 1, \ldots, N$ and $d = 0, \ldots, L-1$ and covariate $\mathbf{Z}$ for $N \times L$ observations.

Now I will discuss how to maximize $l_2(\lambda_k^{(t)}, \beta_{kd}^{(t)}, \sigma_k^{(t)})$ to obtain estimates for $\lambda_k^{(t+1)}$, $\beta^{(t+1)}$ and $\sigma_k^{(t+1)}$.

$$\begin{aligned} l_2(\lambda_k, \beta_{kd}, \sigma_k) &= \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d)\log P(T_{ik}^{(\lambda_k)}|D_i = d, X_i) + \sum_{i=1}^{N}\sum_{k=1}^{K}(\lambda_k - 1)\log T_{ik} \\ &= \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d)\log\Big[\frac{1}{\sqrt{2\pi\sigma_k^2}}e^{-\frac{(T_{ik}^{(\lambda_k)} - \vec{X}_i^T\beta_{kd})^2}{2\sigma_k^2}}\Big] + \sum_{i=1}^{N}\sum_{k=1}^{K}(\lambda_k - 1)\log T_{ik} \\ &= \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d)\Big[-\frac{(T_{ik}^{(\lambda_k)} - \vec{X}_i^T\beta_{kd})^2}{2\sigma_k^2} - \log\sqrt{2\pi\sigma_k^2}\Big] + \sum_{i=1}^{N}\sum_{k=1}^{K}(\lambda_k - 1)\log T_{ik}. \end{aligned}$$

Taking the derivative of $l_2(\lambda_k, \beta_{kd}, \sigma_k)$ with respect to β_{kd},

$$\begin{aligned}
\frac{\partial l_2(\lambda_k, \beta_{kd}, \sigma_k|\vec{T}, \mathbf{X})}{\partial \beta_{kd}} &= \sum_{i=1}^{N}\sum_{d=0}^{L-1}\Big[\frac{2P_i^{(t)}(d)(T_{ik}^{(\lambda_k)} - \vec{X}_i^T\beta_{kd})X_i}{2\sigma_k^2}\Big] \\
&= \frac{1}{\sigma_k^2}\mathbb{X}'_{st}\mathbb{W}^{(t)}(Y_{k,st}^{(t)} - \mathbb{X}_{st}\beta_{kd}) =^{set} 0 \\
\Rightarrow \quad & \mathbb{X}'_{st}\mathbb{W}^{(t)}\mathbb{X}_{st}\beta_{kd} = \mathbb{X}'_{st}\mathbb{W}^{(t)}Y_{k,st}^{(t)} \\
\Rightarrow \quad & \beta_{kd}^{(t+1)} = (\mathbb{X}'_{st}\mathbb{W}^{(t)}\mathbb{X}_{st})^{-1}\mathbb{X}'_{st}\mathbb{W}^{(t)}Y_{k,st}^{(t)},
\end{aligned}$$

where $\mathbb{X}_{st}$, $Y_{k,st}^{(t)}$ and $\mathbb{W}^{(t)}$ were defined in Chapter 5.2 as follows,

$$\mathbb{X}_{st} = \begin{pmatrix} \mathbb{X}(D=0) \\ \dots \\ \mathbb{X}(D=L-1) \end{pmatrix}, \quad Y_{k,st}^{(t)} = (\underbrace{Y_k^{(t)'}, \dots, Y_k^{(t)'}}_{L \text{ times}})',$$

$$\text{and} \quad \mathbb{W}^{(t)} = diag\{P^{(t)}(0), \dots, P^{(t)}(L-1)\},$$

where $\mathbb{X}(D)$ is the design matrix defined by the measurement model $P(T_k|D, \mathbf{X})$, including all covariates $\mathbf{X}$ and their interactions with D, and $Y_k^{(t)}$ is the vector of transformed outcomes with N elements at the tth EM iteration, $Y_{ik}^{(t)} = T_{ik}^{(\lambda_k^{(t)})} = (T_{ik}^{\lambda_k^{(t)}} - 1)/\lambda_k^{(t)},\ i = 1, \dots, N$, and $P^{(t)}(d) = (P_1^{(t)}(d), \dots, P_N^{(t)}(d)),\ d = 0, \dots, L-1$.

Take the derivative of $l_2(\lambda_k, \beta_{kd}, \sigma_k)$ with respect to σ_k^2,

$$\begin{aligned}
& \frac{\partial l_2(\lambda_k, \beta_{kd}, \nu_k|\vec{T}, \mathbf{X})}{\partial \sigma_k^2} = \sum_{i=1}^{N}\sum_{d=0}^{L-1}\Big[\frac{P_i^{(t)}(d)(T_{ik}^{(\lambda_k)} - \vec{X}_i^T\beta_{kd})^2}{2\sigma_k^4} - \frac{P_i^{(t)}(d)}{2\sigma_k^2}\Big] =^{set} 0 \\
\Rightarrow \quad & \frac{\sum_{i=1}^{N}\sum_{d=0}^{L-1}P_i^{(t)}(d)(T_{ik}^{(\lambda_k)} - \vec{X}_i^T\beta_{kd})^2}{\sigma_k^2} = \sum_{i=1}^{N}\sum_{d=0}^{L-1}P_i^{(t)}(d) = N \\
\Rightarrow \quad & \hat{\sigma}_k^{2(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=0}^{L-1}P^{(t)}(d)(T_{ik}^{(\lambda_k^{(t)})} - \vec{X}_i^T\beta_{kd}^{(t)})^2 \\
&= \frac{1}{N}(Y_{k,st}^{(t)} - \mathbb{X}_{st}\beta_{kd})'\mathbb{W}^{(t)}(Y_{k,st}^{(t)} - \mathbb{X}_{st}\beta_{kd}) \\
&= \frac{1}{N}Y_{k,st}^{*(t)'}P_{\mathbb{X}_{st}^{*(t)}}Y_{k,st}^{*(t)},
\end{aligned}$$

where the superscript $*$ denotes a weighted version, $\mathbb{X}_{st}^{*(t)} = \mathbb{W}^{(t)\frac{1}{2}}\mathbb{X}_{st}$, $Y_{k,st}^{*(t)} = \mathbb{W}^{(t)\frac{1}{2}}Y_{k,st}^{(t)}$, and P_X is a projection matrix, $P_X = I - X(X'X)^{-1}X'$.

To get λ_k, I rewrote $l_2(\lambda_k, \beta_{kd}, \sigma_k)$ in its concentrated form (the profile likelihood) as

$$\begin{aligned} l_2(\lambda_k, \beta_{kd}, \sigma_k) &= -\sum_{k=1}^{K} \frac{N\hat{\sigma}_k^2}{2\hat{\sigma}_k^2} - \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d) \log\sqrt{2\pi\hat{\sigma}_k^2} + \sum_{i=1}^{N}\sum_{k=1}^{K}(\lambda_k - 1)\log T_{ik} \\ &= \sum_{k=1}^{K}\Big[-\frac{N}{2} - N\log\sqrt{2\pi\hat{\sigma}_k^2} + (\lambda_k - 1)\sum_{i=1}^{N}\log T_{ik}\Big] \\ &= \sum_{k=1}^{K}\Big[-N\log\hat{\sigma}_k + (\lambda_k - 1)\sum_{i=1}^{N}\log T_{ik} + C\Big], \end{aligned}$$

where $C = -\frac{N}{2} - N\log\sqrt{2\pi}$ was a constant.

Thus, $\lambda_k^{(t+1)}$ was obtained by maximizing the following function:

$$l_3(\lambda_k) = -N\log\hat{\sigma}_k + (\lambda_k - 1)\sum_{i=1}^{N}\log T_{ik}, \quad k = 1, \ldots, K.$$

or, equivalently, $\lambda_k^{(t+1)}$ was the root of the following equation:

$$\frac{\partial l_3(\lambda_k^{(t+1)})}{\partial \lambda_k^{(t+1)}} = 0 \quad k = 1, \ldots, K.$$

Because the gradient of the Box-Cox transformation is

$$\frac{\partial Y^\lambda}{\partial \lambda} = \begin{cases} \dfrac{\lambda Y^\lambda \log Y - Y^\lambda + 1}{\lambda^2} & \text{if } \lambda^{(t+1)} \neq 0 \\ \dfrac{(\log Y)^2}{2} & \text{if } \lambda^{(t+1)} = 0 \end{cases} \qquad k = 1, \ldots, K,$$

I obtained that $\lambda_k^{(t+1)}$ satisfied the following equation:

$$\frac{\partial l_3(\lambda_k^{(t+1)})}{\partial \lambda_k^{(t+1)}} = \begin{cases} -N\dfrac{Y_{k,st}^{*\prime}(\lambda_k^{(t+1)}) P_{\mathbb{X}_{st}^{*(t)}} U_k(\lambda_k^{(t+1)})}{Y_{k,st}^{*\prime}(\lambda_k^{(t+1)}) P_{\mathbb{X}_{st}^{*(t)}} Y_{k,st}^{*}(\lambda_k^{(t+1)})} + \dfrac{N}{\lambda_k^{(t+1)}} + \vec{1}\log T_k = 0, & \text{if } \lambda^{(t+1)} \neq 0 \\ -N\dfrac{Y_{k,st}^{*\prime}(\lambda_k^{(t+1)}) P_{\mathbb{X}_{st}^{*(t)}} U_k(0)}{Y_{k,st}^{*\prime}(\lambda_k^{(t+1)}) P_{\mathbb{X}_{st}^{*(t)}} Y_{k,st}^{*}(\lambda_k^{(t+1)})} + \vec{1}\log T_k = 0, & \text{if } \lambda^{(t+1)} = 0 \end{cases}$$

where $Y_{k,st}^{*}(\lambda_k^{(t+1)}) = Y_{k,st}^{*(t+1)}$. I used this notation to emphasize that $Y_{k,st}^{*(t+1)}$ was a function of $\lambda_k^{(t+1)}$. And $U_k(\lambda_k^{(t+1)})$ was a $N \times L$ by 1 vector with elements $W_{ii}^{\frac{1}{2}}(T_{ik}^{\lambda_k^{(t+1)}} \log T_{ik})/\lambda_k^{(t+1)}$ when $\lambda_k^{(t+1)} \neq 0$ and with element $W_{ii}^{\frac{1}{2}}(\log T_{ik})^2/2$ when $\lambda_k^{(t+1)} = 0$, where W_{ii} was the (i,i) entry of $\mathbb{W}$.

Then, I considered the heteroscedastic error distribution in the measurement model $\epsilon_{ikd} \sim^{i.i.d} N(0, \sigma^2_{kd})$.

In the E step, I computed the expected value of the complete data log likelihood,

$$\begin{aligned} P_i^{(t)}(d) &= \frac{P(\vec{T}_i|D_i=d,X_i)P(D_i=d|Z_i)}{\sum_{d=0}^{L-1} P(\vec{T}_i|D_i=d,X_i)P(D_i=d|Z_i)} \\ &= \frac{\big[\prod_{k=1}^{K} \phi\big(\frac{H_k^{(t)}(t_{ik})-\vec{x}_i^T\beta_{kd}^{(t)}}{\sigma_{kd}^{(t)}}\big)\big]\eta_d(\vec{z}_i^T\alpha^{(t)})}{\sum_{d=0}^{L-1}\big\{\prod_{k=1}^{K}\big[\phi\big(\frac{H_k^{(t)}(t_{ik})-\vec{x}_i^T\beta_{kd}^{(t)}}{\sigma_{kd}^{(t)}}\big)\big]\eta_d(\vec{z}_i^T\alpha^{(t)})\big\}}. \end{aligned}$$

In the M step, $\alpha_d^{(t+1)}$ was obtained as before by maximizing $l_1(\alpha_d^{(t)})$. The log "likelihood" function $l_2(\lambda_k^{(t)}, \beta_{kd}^{(t)}, \sigma_k^{(t)})$ for obtaining estimates for $\lambda_k^{(t+1)}$, $\beta^{(t+1)}$, $\sigma_k^{(t+1)}$ was,

$$\begin{aligned} l_2(\lambda_k,\beta_{kd},\sigma_{kd}) &= \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d)\log P(T_{ik}^{(\lambda_k)}|D_i=d,X_i) + \sum_{i=1}^{N}\sum_{k=1}^{K}(\lambda_k-1)\log T_{ik} \\ &= \sum_{i=1}^{N}\sum_{d=0}^{L-1}\sum_{k=1}^{K} P_i^{(t)}(d)\Big[-\frac{(T_{ik}^{(\lambda_k)}-\vec{X}_i^T\beta_{kd})^2}{2\sigma_{kd}^2} - \log\sqrt{2\pi\sigma_{kd}^2}\Big] + \sum_{i=1}^{N}\sum_{k=1}^{K}(\lambda_k-1)\log T_{ik} \end{aligned}$$

I took the derivative of $l_2(\lambda_k, \beta_{kd}, \sigma_{kd})$ with respect to β_{kd},

$$\begin{aligned} \frac{\partial l_2(\lambda_k,\beta_{kd},\sigma_{kd}|\vec{T},\mathbf{X})}{\partial\beta_{kd}} &= \sum_{i=1}^{N}\sum_{d=0}^{L-1}\Big[\frac{2P_i^{(t)}(d)(T_{ik}^{(\lambda_k)}-\vec{X}_i^T\beta_{kd})X_i}{2\sigma_{kd}^2}\Big] \\ &= \mathbb{X}'_{st}\mathbb{W}^{*(t)}(Y_{k,st}^{(t)}-\mathbb{X}_{st}\beta_{kd}) =^{set} 0 \\ \Rightarrow \qquad \beta_{kd}^{(t+1)} &= (\mathbb{X}'_{st}\mathbb{W}^{*(t)}\mathbb{X}_{st})^{-1}\mathbb{X}'_{st}\mathbb{W}^{*(t)}Y_{k,st}^{(t)}, \end{aligned}$$

where $W^{*(t)}$ was a $N\times L$ by $N\times L$ diagonal matrix, $W^{*(t)} = diag\Big\{\frac{P^{(t)}(0)}{\sigma_{k0}^{(t)}},\ldots,\frac{P^{(t)}(L-1)}{\sigma_{k(L-1)}^{(t)}}\Big\}$.

I also took the derivative of $l_2(\lambda_k, \beta_{kd}, \sigma_{kd})$ with respect to σ_{kd}^2,

$$\frac{\partial l_2(\lambda_k,\beta_{kd},\sigma_{kd}|\vec{T},\mathbf{X})}{\partial\sigma_{kd}^2} = \sum_{i=1}^{N}\Big[\frac{P_i^{(t)}(d)(T_{ik}^{(\lambda_k)}-\vec{X}_i^T\beta_{kd})^2}{2\sigma_{kd}^4} - \frac{P_i^{(t)}(d)}{2\sigma_{kd}^2}\Big] =^{set} 0$$

$$\begin{aligned} \Rightarrow \qquad \hat{\sigma}_{kd}^{2(t+1)} &= \frac{1}{\sum_{i=1}^{N}P_i^{(t)}(d)}\sum_{i=1}^{N}P^{(t)}(d)(T_{ik}^{(\lambda_k^{(t)})}-\vec{X}_i^T\beta_{kd}^{(t)})^2 \\ &= \frac{1}{\sum_{i=1}^{N}P_i^{(t)}(d)}(Y_{k,st}^{(t)}-\mathbb{X}_{st}\beta_{kd}^{(t)})_{[d]}^T\mathbb{W}_{[d]}^{(t)}(Y_{k,st}^{(t)}-\mathbb{X}_{st}\beta_{kd}^{(t)})_{[d]} \end{aligned}$$

where the subscript $[d]$ refers to the corresponding d "block" of the vector or matrix.

And $l_2(\lambda_k, \beta_{kd}, \sigma_{kd})$, in its corresponding concentrated Form, was

$$\begin{aligned} l_2(\lambda_k, \beta_{kd}, \sigma_{kd}) &= \sum_{k=1}^{K}\Big[-\sum_{d=0}^{L-1}\frac{\sum_{i=1}^{N}P_i^{(t)}(d)\hat{\sigma}_{kd}^2}{2\hat{\sigma}_{kd}^2} - \sum_{i=1}^{N}\sum_{d=0}^{L-1}P_i^{(t)}(d)\log\sqrt{2\pi\hat{\sigma}_{kd}^2} + (\lambda_k - 1)\sum_{i=1}^{N}\log T_{ik}\Big] \\ &= \sum_{k=1}^{K}\Big[-\frac{N}{2} - \frac{N}{2}\log 2\pi - \sum_{d=0}^{L-1}\sum_{i=1}^{N}P_i^{(t)}(d)\log\hat{\sigma}_{kd} + (\lambda_k - 1)\sum_{i=1}^{N}\log T_{ik}\Big] \\ &= \sum_{k=1}^{K}\Big[-\sum_{d=0}^{L-1}\sum_{i=1}^{N}P_i^{(t)}(d)\log\hat{\sigma}_{kd} + (\lambda_k - 1)\sum_{i=1}^{N}\log T_{ik} + C\Big], \end{aligned}$$

where $C = -\dfrac{N}{2} - N\log\sqrt{2\pi}$ was a constant.

Thus, $\lambda_k^{(t+1)}$ was obtained by maximizing the following function:

$$l_3(\lambda_k) = -\sum_{d=0}^{L-1}\sum_{i=1}^{N}P_i^{(t)}(d)\log\hat{\sigma}_{kd} + (\lambda_k - 1)\sum_{i=1}^{N}\log T_{ik},$$

or, equivalently, $\lambda_k^{(t+1)}$ satisfied the following equation:

$$\frac{\partial l_3(\lambda_k^{(t+1)})}{\partial\lambda_k^{(t+1)}} = \begin{cases} -\sum_{d=0}^{L-1}\Big\{[\sum_{i=1}^{N}P_i^{(t)}(d)]\dfrac{\sum_{i=1}^{N}P_i^{(t)}(d)(T_{ik}^{(\lambda_k^{(t+1)})} - \vec{X}_i^T\beta_{kd}^{(t)})V_{ik}(\lambda_k^{(t+1)})}{\sum_{i=1}^{N}P_i^{(t)}(d)(T_{ik}^{(\lambda_k^{(t+1)})} - \vec{X}_i^T\beta_{kd}^{(t)})^2}\Big\} \\ \quad + \dfrac{N}{\lambda_k^{(t+1)}} + \vec{\mathbb{1}}\log T_k = 0, & \text{if } \lambda^{(t+1)} \neq 0 \\ -\sum_{d=0}^{L-1}\Big\{[\sum_{i=1}^{N}P_i^{(t)}(d)]\dfrac{\sum_{i=1}^{N}P_i^{(t)}(d)(T_{ik}^{(\lambda_k^{(t+1)})} - \vec{X}_i^T\beta_{kd}^{(t)})V_{ik}(0)}{\sum_{i=1}^{N}P_i^{(t)}(d)(T_{ik}^{(\lambda_k^{(t+1)})} - \vec{X}_i^T\beta_{kd}^{(t)})^2}\Big\} \\ \quad + \vec{\mathbb{1}}\log T_k = 0, & \text{if } \lambda^{(t+1)} = 0 \end{cases}$$

where $V_k(\lambda_k^{(t+1)})$ is a N by 1 vector with the ith element $(T_{ik}^{\lambda_k^{(t+1)}}\log T_{ik})/\lambda_k^{(t+1)}$ when $\lambda_k^{(t+1)} \neq 0$ and with the ith element $(\log T_{ik})^2/2$ when $\lambda_k^{(t+1)} = 0$.

Appendix F

DERIVATION OF THE COVARIATE-SPECIFIC AUC IN CHAPTER 5 SIMULATIONS

The measurement model was given by $H_k(T_k) = \tilde{\beta}_{kd0} + \tilde{\beta}_{kd1}X_i + \epsilon_{ik}$, $d = 0, 1, 2$, or equivalently,

$$\begin{aligned} H_k(T_k) &= \beta_{k0} + \beta_{k1}X_i + \beta_{k2}I(D=1) + \beta_{k3}I(D=2) \\ &+ \beta_{k4}XI(D=1) + \beta_{k5}XI(D=2) + \epsilon_{ik}, \\ & k = 1, \ldots, K, \qquad \epsilon_{ik} \sim^{i.i.d} N(0, 0.5^2), \end{aligned}$$

with the following relation between $\tilde{\beta}$ and β,

$$\tilde{\beta}_{k00} = \beta_{k0}, \quad \tilde{\beta}_{k01} = \beta_{k1}, \quad \tilde{\beta}_{k10} = \beta_{k0} + \beta_{k2}, \quad \tilde{\beta}_{k11} = \beta_{k0} + \beta_{k4}$$
$$\tilde{\beta}_{k20} = \beta_{k0} + \beta_{k3}, \quad \tilde{\beta}_{k21} = \beta_{k0} + \beta_{k5}, \qquad k = 1, \ldots, K.$$

I derived the covariate-specific ROC curve of test T_k for distinguishing between two disease groups. Without loss of generality, I considered distinguishing between $D = 0$ and $D = 1$.

$$\begin{aligned} Sens_k(c, X) &= P(T_k > c|D=1) = 1 - \Phi\big(\frac{H_k(c) - \tilde{\beta}_{k10} - \tilde{\beta}_{k11}X}{\sigma_{k1}}\big), \\ 1 - Spec_k(c, X) &= P(T_k > c|D=0) = 1 - \Phi\big(\frac{H_k(c) - \tilde{\beta}_{k00} - \tilde{\beta}_{k01}X}{\sigma_{k0}}\big). \end{aligned}$$

Consequently, the covariate-specific ROC curve was given by:

$$\begin{aligned} ROC_k(t, X) &= 1 - \Phi\big(\frac{\sigma_{k0}\Phi^{-1}(1-t) + \tilde{\beta}_{k00} + \tilde{\beta}_{k01}X - \tilde{\beta}_{k10} - \tilde{\beta}_{k11}X}{\sigma_{k1}}\big) \\ &= 1 - \Phi\big(\frac{\sigma_k\Phi^{-1}(1-t) - \beta_{k2} - \beta_{k4}X}{\sigma_k}\big) \\ &= 1 - \Phi\big(-\frac{\beta_{k2} + \beta_{k4}X}{\sigma_k} + \Phi^{-1}(1-t)\big). \end{aligned} \tag{F.1}$$

To obtain the AUC value, I first showed that the following equations hold,

$$\int_{-\infty}^{+\infty} \Phi(a + bx)d\Phi(x) = \Phi\big(\frac{a}{\sqrt{1+b^2}}\big), \tag{F.2}$$

$$\int_0^1 1 - \Phi\big(a + b\Phi^{-1}(1-t)\big)dt = \Phi\big(\frac{-a}{\sqrt{1+b^2}}\big). \tag{F.3}$$

The derivation for equation (F.2) is given below,

$$\int_{-\infty}^{+\infty} \Phi(a+bx)d\Phi(x) = \int_{-\infty}^{+\infty} \Big(\int_{-\infty}^{a+bx} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \Big) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

let $u = t - bx$

$$\begin{aligned}
&= \int_{-\infty}^{+\infty} \Big(\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(u+bx)^2}{2}} e^{-\frac{x^2}{2}} du \Big) dx \\
&= \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \Big(\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(u+bx)^2+x^2}{2}} dx \Big) du \\
&= \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \Big(\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{1+b^2}x + \frac{bu}{\sqrt{1+b^2}})^2 - \frac{u^2}{2(1+b^2)}} dx \Big) du \\
&= \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2(1+b^2)}} \Big(\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{1+b^2}x + \frac{bu}{\sqrt{1+b^2}})^2} dx \Big) du \\
&= \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2(1+b^2)}} \frac{1}{\sqrt{1+b^2}} du = \Phi\big(\frac{a}{\sqrt{1+b^2}}\big)
\end{aligned}$$

Next, I derived equation (F.3).

$$\int_0^1 1 - \Phi\big(a + b\Phi^{-1}(1-t)\big) dt$$

let $x = \Phi^{-1}(1-t)$, then $t = 1 - \Phi(x)$

$$\begin{aligned}
&= \int_{+\infty}^{-\infty} 1 - \Phi(a+bx) d\big(1 - \Phi(x)\big) = \int_{-\infty}^{+\infty} 1 - \Phi(a+bx) d\Phi(x) \\
&= 1 - \int_{-\infty}^{+\infty} \Phi(a+bx) d\Phi(x) = 1 - \Phi\big(\frac{a}{\sqrt{1+b^2}}\big) \qquad \text{using (F.2)} \\
&= \Phi\big(\frac{-a}{\sqrt{1+b^2}}\big) \qquad \text{because } 1 - \Phi(x) = \Phi(-x).
\end{aligned}$$

So that the covariate-specific AUC for the ROC curve given by equation (F.1) was,

$$\begin{aligned}
AUC_k(X) &= \int_0^1 ROC_k(t, X) dt \\
&= \int_0^1 1 - \Phi\big(- \frac{\beta_{k2} + \beta_{k4} X}{\sigma_k} + \Phi^{-1}(1-t)\big) dt \\
&= \Phi\big(\frac{\beta_{k2} + \beta_{k4} X}{\sigma_k}\big).
\end{aligned}$$

In addition, because X was a standard normal variable, the average AUC for this pop-

ulation was,

$$\begin{aligned} AUC_k &= \int_{-\infty}^{+\infty} AUC_k(X) d\Phi(X) \\ &= \int_{-\infty}^{+\infty} \Phi\Big(\frac{\beta_{k2} + \beta_{k4} X}{\sigma_k}\Big) d\Phi(X) \\ &= \Phi\Big(\frac{\beta_{k2}}{\sqrt{\sigma_k^2 + \beta_{k4}^2}}\Big). \end{aligned}$$

With similar derivations, the average AUC of each test for discriminating between any two disease groups in the simulations can be computed.