Statistical methods for combining diagnostic tests and performance evaluation metrics

by

Chengning Zhang

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

University of Wisconsin-Madison

2022

Date of Final Oral Exam: 06/24/2022

The dissertation is approved by the following members of the Final Oral Committee:
Lu Mao, Associate Professor, Department of Biostatistics and Medical Informatics
Jun Zhu, Professor, Department of Statistics
Richard Chappell, Professor, Department of Biostatistics and Medical Informatics
Guanhua Chen, Assistant Professor, Department of Biostatistics and Medical Informatics
Yeonhee Park, Assistant Professor, Department of Biostatistics and Medical Informatics

Statistical methods for combining diagnostic tests and evaluation metrics

Chengning Zhang

Abstract

In biomedical studies, it is usually the case that several diagnostic tests can be performed on an individual or multiple disease markers are available simultaneously, and that many of them may be associated with the clinical outcome. In practice, a single test or marker often has limited diagnostic performance. Therefore, it is important to combine multiple sources of information available to achieve higher classification performance. This dissertation focuses on statistical methods for combining multiple diagnostic tests and the corresponding performance evaluation metrics. In the first project, we provide a survey of the current state of the art in methods for combining multiple tests. We categorize existing methods into three general groups and conduct extensive simulation studies to compare the performance of different combination methods. The reviewed methods serve as benchmark for developing new combination approaches in the following projects.

In the second project, we consider the problem of combining multiple tests whose values are missing at random (MAR). In addition, we aim to exploit the known monotonicity relationship between the input variables and the disease outcome for gains in diagnostic accuracy. We develop a novel likelihood-based approach to monotone classification that accounts for missing inputs in a natural and principled way. The risk score function is obtained through the nonparametric maximum likelihood estimation (NPMLE). A novel expectation-maximization (EM)-type algorithm is devised to compute the NPMLE by treating the monotonicity-constrained risk score function as a cumulative distribution for a latent random vector. Through simulation studies and a real data example, we demonstrate that the proposed method outperforms state-of-the-art methods for combining multiple inputs under monotonic assumption, especially when the inputs contain

missing data. We illustrate our approach with a dataset from a recent nonalcoholic fatty liver disease (NALFD) study.

In the third project, our approach established in the second part is extended to the scenario where one covariate is randomly censored. The proposed approach consists of two steps. In step one, we use a Cox proportional hazards model for the distribution of the censored covariate given other covariates in the model, this conditional distribution is used for calculating the observed likelihood of data. In step two, a similar expectation-maximization (EM)-type algorithm is devised, based on observed data likelihood from step one, to compute the NPMLE of the monotonicity-constrained risk score function. Through simulation studies, we demonstrate that the proposed method outperforms the simple but inefficient complete-case analysis as well as the substitution methods. We apply our method to the data set from a primary biliary cirrhosis (PBC) study conducted at Mayo Clinic.

The proposed methods in part two and three can be extended to multi-class cases, where the labels have an inherent order but no meaningful numeric distance between them. A natural question arises as to how to evaluate the classification performance under such setting. Therefore, in the fourth project, we consider the problem of performance evaluation metrics for ordinal classification. We propose three novel performance evaluation metrics that better capture the ordinality of the outcomes. The first metric is adapted from the area under the receiver operating characteristic (ROC) curve (AUC), while the latter two are simple and interpretable generalizations of the Harrell's concordance index (C-INDEX). Moreover, we show the optimality of the AUC based metrics through Neyman-Pearson lemma. We conduct extensive simulation studies to confirm the usefulness of the proposed performance metrics for ordinal classification.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Lu Mao, for his constant guidance, support and encouragement during my Ph.D. study. He guided me throughout this project. It is my great honor to be supervised by him.

Next, I would like to thank my defense committee members, Professor Jun Zhu, Professor Richard Chappell, Professor Guanhua Chen and Professor Yeonhee Park, for their valuable time, comments and support.

I want to express my gratitude to my internship managers Santiago Eduardo and Roger Ding for their guidance during my internship. I really appreciate their help and support.

In addition, I would like to thank my girlfriend Muen Chen and my friend duobao Jia, for always being there for me to support me in the hard times as well as the easy times.

Furthermore, I would like to thank all my friends who always support me and help me. My heartfelt thanks also go to all the faculty members, staff and graduate students in the Department of Statistics.

Lastly, I would like to give my special thanks to my parents Shouyan Zhang and Yuxiang Zhang for their tremendous love, support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the main problems in medical research is the use of clinical and laboratory data to classify health conditions and predict disease outcomes, especially in the era of advancements in biotechnology that promise accurate noninvasive testing. A notable example of such advancements is the development of quantitative imaging and the use of imaging biomarkers for the assessment of patients' status.

It is usually the case that several diagnostic tests can be performed on an individual or multiple disease markers are available simultaneously, and that many of them may be associated with the clinical outcome. A single test or disease marker, on the other hand, often has limited classification/prediction performance. Consider for example the Wisconsin Breast Cancer data (Wolberg and Mangasarian, 1990), where nine disease markers associated with benign or malignant samples for 699 subjects are measured. If only a single marker is used for classification, the classification performance measured by either sensitivity or specificity is not as satisfactory as when multiple markers are used simultaneously. Therefore, it is important to combine multiple sources of information available to achieve higher classification performance (Pepe, Cai, and Longton, 2006).

There is no shortage of methods for building classification models out of multiple input variables. In binary classification, a simple and common approach is to model the outcome directly against the covariates using, e.g., logistic regression (Richards, Hammitt, and

Tsevat, 1996), support vector machine (SVM) (Cristianini, Shawe-Taylor, et al., 2000), or classification trees (Breiman et al., 2017). These methods, however, cannot guarantee optimality in terms of the area under the receiver operating characteristic (ROC) curve (or AUC) (Pepe, 2003; Zhou, McClish, and Obuchowski, 2009), a common metric of diagnostic accuracy. Alternatively, a number of authors have studied classification rules that aim explicitly to maximize the AUC. Under the assumption of multivariate normality for the input vector, for example, Su and Liu (1993) derived the AUC-optimal linear classification rule in closed form. In the general setting, Pepe and Thompson (2000) proposed to construct linear classifiers by directly maximizing the empirical AUC with respect to the coefficients of combination. The first part of this dissertation conducts an extensive review of existing statistical methods for combining multiple diagnostic tests to improve classification performance.

In term of building classification models, it often proves fruitful to exploit established domain knowledge for gains in diagnostic accuracy. One important instance of such domain knowledge is monotonicity between the input variables and the disease outcome. For example, larger tumors as measured by magnetic resonance imaging (MRI) volumetry (Soutter et al., 2004; desouza et al., 2006) or higher cerebral β-amyloid levels as measured by positron emission tomography (PET) (Vlassenko, Benzinger, and Morris, 2012; Huynh and Mohan, 2017; Aschenbrenner et al., 2018) usually indicate greater risks of cancer or Alzheimer's disease, respectively. In fact, monotone patterns of this kind are common with general quantitative biomarkers derived from imaging modalities. Such biomarkers are typically designed to measure the degree of certain anatomical anomaly and, as a result, are likely to correlate monotonically with the probability of disease (Sullivan et al., 2015; Kessler et al., 2015; Raunig et al., 2015; Obuchowski et al., 2015a; Obuchowski et al., 2015b; Huang et al., 2015).

A simple way to draw on the monotonicity relationships is to embed suitable constraints in standard classification models such as the SVM (Chen and Li, 2014; Li and Chen, 2014). The main weakness of this approach is that the modified classifier inher-

its the same assumptions from the parent model and may thus perform poorly if the parent model is wrongly specified. In the interest of robustness, some authors, mostly computer scientists, have attempted at monotone classification in a completely nonparametric setting. To date, the most successful examples are the ordinal stochastic dominance learner (OSDL) (Lievens, De Baets, and Cao-Van, 2008) and MOCA (presumably short for "monotone classifier") (Barile and Feelders, 2008). Both methods start off by nonparametrically estimating the conditional distribution of an ordinal outcome subject to a stochastic monotonicity constraint with the covariates, with only minor difference in the method of estimation. Then, the classified label is taken to be the median of the estimated conditional distribution. This leads to monotone classifiers whose output is non-decreasing with respect to each component of the covariates.

A major limitation of OSDL and MOCA is that they cannot easily accommodate missing covariates, which are the norm rather than exception in studies of diagnostic medicine. In a recent study involving 186 liver disease patients at the University of Wisconsin (UW) Hospitals, for example, 16.2% of the participants have missing values in at least one of several computed tomography (CT) biomarkers or Fibrosis-4 scores needed for the diagnosis of non-alcoholic steatohepatitis (NASH). In "monotonizing" the conditional outcome distribution, both OSDL and MOCA rely on traditional isotonic regression techniques (Barlow and Brunk, 1972), with no natural strategy to handle missing data other than to exclude them. If the missingness is not completely at random, such "complete case analysis" will incur bias in the estimated outcome distribution, which in turn worsens the diagnostic performance of the associated classifier.

We address this limitation in second part of the thesis by developing a novel likelihood-based approach to nonparametric monotone classification that accounts for missing covariates in a natural and principled way. For ease of exposition, we focus on the case of binary classification. The risk score function is obtained through the nonparametric maximum likelihood estimation (NPMLE). A novel expectation-maximization (EM)-type algorithm (Dempster, Laird, and Rubin, 1977) is devised to compute the NPMLE by treating the

monotonicity-constrained risk score function as a cumulative distribution for a latent random vector. Not only is this EM-type algorithm numerically more stable and efficient than traditional quadratic programming in the presence of many covariates, but more importantly it provides a natural platform to handle missing data — we simply add another E-step to compute the conditional expectation of the coarsened input given the observed data.

Apart from missing covariate data, data with censored covariates is also commonly seen in medical research. The situation of randomly-censored covariates arises, for instance, when modeling the value of a diagnostic marker as a function of the time lag between the measurement and the occurrence of disease. Models of this type are utilized in Cai et al. (2006) and Tsimikas, Bantis, and Georgiou (2012), where the marker sensitivity is considered as a function of survival time. The time-to-event covariates may be censored due to loss to follow-up, study termination or detection limits. Another scenario where randomly censored covariates are encountered frequently is the study of associations between parental risk factors and the onset of disease in their offspring. Often, the commonly used parental risk factor, age-of-onset of disease in parents, is right-censored, meaning that the study either terminates prior to the event being observed or a patient is lost to follow-up prior to the event. For example, Allport et al. (2016) and Atem, Matsouaka, and Zimmern (2019) studied associations between parental age of onset of cardiovascular disease and offspring age of onset of cardiovascular disease, Atem et al. (2017) and Maye et al. (2016) investigated the association between maternal age of onset of dementia and beta-amyloid deposition (measured by in vivo PET imaging) in cognitively normal older offspring.

While missing and censored covariate observations share some similarities, they are fundamentally different. An observation is missing when the observed value of some variable is unknown. On the other hand, an observation is censored when the true value is only partially observed due to varying reasons. As a result, the established approach in the second project is not readily applicable to censored covariate because it does not fully

utilize the partial information contained in censored observations. Moreover, the fully non-parametric approach may not be feasible when covariate are continuous. In the third part of the thesis, we extend our approach established in the second part to the scenario where one covariate is randomly censored. The proposed approach consists of two steps. In step one, we use a Cox proportional hazards model for the distribution of the censored covariate given other covariates in the model, this conditional distribution is used for calculating the observed likelihood of data. In step two, a similar expectation-maximization (EM)-type algorithm is devised, based on observed data likelihood from step one, to compute the NPMLE of the monotonicity-constrained risk score function.

The proposed methods in part two and three can be extended to multi-class cases, where the labels have an inherent order but no meaningful numeric distance between them. This presence of ordering information separates ordinal classification from nominal classification problems. As a result, conventional performance evaluation metrics appropriate for nominal classes are unsuitable for ordinal classification, in that they do not account for the ordinality of the target classes. A natural question arises as to how to evaluate the classification performance under such setting.

In recent years, there have been growing interest in designing proper evaluation metrics for ordinal classification models. Baccianella, Esuli, and Sebastiani (2009) addressed the adaption of exiting measures into ones robust to imbalance. Cardoso and Sousa (2011) proposed an alternative metric defined directly in the confusion matrix, which can capture how much the result diverges from true labels and how "inconsistent" the classifier is in regard to the relative order of the labels. The use of rank order measures are introduced in Lee and Liu (2002) and Vanbelle and Albert (2009), and the adaption of the ROC curve is discussed in Waegeman, De Baets, and Boullart (2006). To date, most evaluation metrics of ordinal data are devised for "ordinary" classifier, whose output is a single label. Sometimes, however, we have probabilistic classifiers for which the output is a probability distribution over a set of labels, rather than a single label. To use these evaluation metrics of ordinal data, one has to convert probabilistic classifiers into "ordinary" classifier

using a given threshold or decision rule, at the risk of losing information. Therefore, our main goal, in fourth part of the thesis, is to propose novel metrics specifically designed to ordinal classification, which can evaluate probabilistic classifiers directly, such that the information from predicted probability distributions can be fully utilized. The first metric is adapted from the area under the receiver operating characteristic (ROC) curve (AUC), while the latter two are simple and interpretable generalizations of the Harrell's concordance index (C-INDEX). In addition, we show the optimality of all AUC based metrics through Neyman-Pearson lemma.

The rest of the dissertation is organized as follows. In Chapter 2, we provide a survey of the current state of the art in methods for combining multiple tests. We categorize the methods into three general groups and conduct extensive simulation studies to compare the performance of different combination methods. In Chapter 3, we develop a novel likelihood-based approach to monotone classification that accounts for missing inputs in a natural and principled way. The risk score function is obtained through the nonparametric maximum likelihood estimation (NPMLE). A novel expectation-maximization (EM)-type algorithm is devised to compute the NPMLE by treating the monotonicity-constrained risk score function as a cumulative distribution for a latent random vector. In Chapter 4, our approach established in the Chapter 3 is extended to the scenario where one covariate is randomly censored. In Chapter 5, we consider the problem of performance evaluation metrics for ordinal classification. We propose three novel performance evaluation metrics and confirm their usefulness through extensive simulation studies. Discussion and concluding remarks are given in Chapter 6.

Chapter 2

Statistical methods for combining multiple diagnostic tests: a review

2.1 Introduction

Classification is one of the most important and typical applications of medical research. In recent years, there has been an increasing interest in using clinical and laboratory data to classify health conditions and predict disease outcomes, in part because of biotechnologic advancements that promise accurate noninvasive testing. One notable example of such advancements is the development of quantitative imaging biomarkers for the assessment of patients' status.

It is often the case that a single diagnostic test or biomarker has limited classification performance. For example, Wolberg and Mangasarian (1990) studied the association between breast cancer risk and nine disease markers. They showed that if only a single marker is used for classification, the classification performance measured by either sensitivity or specificity is not as satisfactory as when multiple markers are used simultaneously. As another example, it was shown in Pepe et al. (2001) that multiple biomarkers will be needed for detecting subclinical cancer with adequate sensitivity and specificity. Therefore, investigators find it necessary to combine multiple sources of information available

to achieve higher diagnostic performance.

Numerous methods have been proposed for building classification models out of multiple test results. Among these methods, we focus on those which can accommodate case-control designs, a commonly employed study design where study subjects are selected on the basis of the outcome. Such retrospective designs usually require far smaller sample sizes than the prospective studies and are often employed at early stages of classifier development (Pepe et al., 2001; Pepe, 2003).

To assess the performance of a binary diagnostic test, which yields a binary result with either negative or positive outcome, we make use of the commonly used metrics in medicine: the true- and false-positive fractions (TPF and FPF). Such two-dimensional measures of accuracy are more frequently used than one-dimensional summary measures including the overall mis-classification rate and odds ratio due to the fact that the consequence of false-negative and false-positive errors is often very different and hard to quantify. In fact, one-dimensional summary measures are rarely used in practice in that they often lead to spurious results (Pepe et al., 2004). Another popular two-dimensional measures of accuracy are PPV and NPV. However, they depend on the disease prevalence and thus cannot be assessed directly from case-control studies. As a result, we only focus on TPF and FPF, which are also known as sensitivity and 1 - specificity.

When multiple diagnostic predictors are available, one seeks to find a classification rule mapping the predictors into a scalar value. It predicts a subject to be positive if the scalar-valued function is greater than or equal to a threshold value, and predict negative otherwise. The performance evaluation tool, in that case, is the receiver operating characteristic (ROC) curve which generalizes the notions of (TPF, FPF) from binary classifiers to scalar-valued classifiers. The ROC curve is the most popular way to evaluate the overall performance of scalar-valued classifiers (Baker, 2003). In addition, numerical indices are often used to summarize the ROC curves. The commonly used ones are the area under the ROC curve (AUC), the partial area under the ROC curve (pAUC) and Youden's index.

Although there are an increasing number of peer-reviewed articles regarding methods for combining multiple tests, at a high level, the key research questions can be formulated rather straightforwardly as: (1) what assumptions are typically made to facilitate the methods of combining multiple tests? (2) whether the combination method can handle case-control study design? (3) whether the scalar-valued function is a linear or non-linear combination of predictors? (4) whether statistical inference can be performed? (5) whether variable selection can be performed? The survey is structured around giving a comprehensive overview about how the community is tackling each of these questions.

In summary, this chapter reviews methods of combining multiple test results into a score, i.e, a scalar-valued function, when the goal is to use that score for classification. The evaluation of classification accuracy is based on summary indices of the ROC curve. In Section 2.2, we review core concepts of ROC curve for single diagnostic test and the optimality of likelihood ratio and risk score, which will build solid foundation when we comprehensively review methods for combining multiple tests in Section 2.3. In Section 2.4 we conduct extensive simulation studies to investigate the performance of different combination approaches, and provide practical suggestions on how to choose methods of combining multiple tests. Existing approaches are applied, in Section 2.5, to a real data set of 186 patients with NALFD to combine four markers to increase diagnostic accuracy. This article also reviews, in Section 2.6, available computing packages. We build a Python and R wrapper for most existing methods, which is available on the author's Github. In Section 2.7, some concluding remarks are given about how the community is tackling the central research questions posed at the end of the introduction, discussion of future research directions regarding that topic is also included.

2.2 Preliminaries

2.2.1 Basic concepts of the ROC curve

We first consider diagnostic tests that yield a binary result T with either negative or positive diagnosis. The diagnostic accuracy of T is characterized by the probability that the test correctly classifies a healthy subject as healthy (TNF), as well as the probability that the test correctly classifies a diseased subject as diseased (TPF). Table 2.1 provides a 2×2 table of T versus D.

Table 2.1: Confusion matrix of a binary outcome D and a binary test T.

		True status D		
		Positive	Negative	Total
Test T	Positive	$a = \#\{TN\}$	$b = \#\{FN\}$	$a + b$
	Negative	$c = \#\{FP\}$	$d = \#\{TP\}$	$c + d$
	Total	$a + c$	$b + d$	N

When the outcome of diagnostic tests is a scalar value, still labeled as T, we can dichotomize the test based on a threshold c. In that case, the ROC curve is a simple and yet useful tool that generalizes the notions of (TPF, FPF). Instead of choosing a fixed threshold c, we vary the threshold from $-\infty$ to ∞ and obtain all possible pairs of TPF(c) and FPF(c). The curve can be constructed by plotting all possible pairs of FPF on the x axis and TPF on the y axis, at every common threshold value of c. From another perspective, TPF(c) and FPF(c) are determined by the common threshold c and the survival functions of T in the diseased and healthy populations. Therefore, by varying the threshold c, the curve can be determined jointly by the survival functions of T in the diseased and healthy subjects, where the survival function is one minus the cumulative distribution function (CDF).

The ROC curve has several desirable properties. First, it does not depend on the disease prevalence and thus is able to accommodate case-control studies. Second, it is not influenced by the choice of the threshold c. Third, the ROC curve is invariant of strictly increasing transformation of the test result T.

Numerical indices are often used to summarize the ROC curves. For example, the area

under the curve (AUC) is an overall summary index of the ROC curve. The AUC ranges from 0.5 for uninformative test to one for perfect test. It is possible for the AUC to be smaller than 0.5, in that case the rating can be reversed by multiplying all test results by minus one. Additionally, the AUC has an interesting interpretation. That is, it is equal to the probability that test results from a randomly selected pair of diseased and non-diseased subjects are correctly ordered (Bamber, 1975; Hanley and McNeil, 1982).

As mentioned, the ROC curve is fully determined by the survival functions of T in the diseased and healthy subjects. In addition, the survival functions can be estimated nonparametrically (e.g., Kaplan–Meier estimation). As a result, the empirical ROC curve can be estimated in a nonparametric fashion. Interestingly, the area under the empirical ROC curve is the well-known Mann-Whitney U-statistics or, equivalently, the Wilcoxon rank sum statistic.

A major limitation of the full AUC is that it gives equal attention to the entire range of FPF and TPF, while in practice only a limited range (e.g., either high specificity or high sensitivity) may be of interest. Alternatively, the partial AUC, which is the area under the ROC curve between two fixed values of specificity, can be employed. Partial AUC can be estimated nonparametrically in a similar fashion as AUC. Another summary index of the ROC curve is Youden's index, which is defined as the maximum vertical distance between the ROC curve and the 45° line. Youden's index is an indicator of how far the curve is from that of the uninformative test and thus can be used as evaluation metric.

2.2.2 Optimality of likelihood ratio and risk score

McIntosh and Pepe (2002) pointed out that the likelihood ratio function, defined as the ratio of probability distribution function of the test result in the diseased populations to that in the non-diseased populations, yields all optimal decision rules in the sense that it maximizes the sensitivity over the entire specificity range uniformly. Essentially, this result is the well-known Neyman-Pearson lemma developed in the context of statistical hypothesis testing (Neyman and Pearson, 1933), but applied to medical diagnosis.

In order to calculate the likelihood ratio function, the multivariate probability distributions for multiple test results in the diseased and non-diseased populations must be known or estimated, which often needs strong assumptions. Alternatively, it was shown in McIntosh and Pepe (2002) and Baker (2000) that the risk score, defined as the probability of disease given multiple test results, is also the optimal combination function in the same sense as the likelihood ratio function. The equivalence between optimality of risk score and likelihood ratio function lies in the fact that the risk score function is a monotone increasing function of the likelihood ratio function.

2.3 Methods for combining multiple test results

Statistical methods of combining multiple tests have been widely developed. This section provides a comprehensive overview of the methods and their development history. We group them into three broad categories, which are parametric methods, semiparametric methods and other methods that not specially designed for combining multiple tests but can be applied equally well in practice. Within each category, we further divide them into subcategories based on certain property of the methods.

2.3.1 Parametric methods

As the name implies, methods in this category require parametric modeling. Depending on either probability distribution for test results in the diseased and non-diseased populations or the risk score function is modeled, we further divide parametric methods into two subcategories. The goal for different combination methods is to improve diagnostic accuracy by maximizing the area or partial area under the ROC curve.

Modeling probability distribution function for test results

Su and Liu (1993) derived the best linear combination function that maximize the area under the receiver operating characteristic curve (AUC) under normal assumptions. They considered two scenarios. The first scenario is that when the two covariance matrices of

diseased and non-diseased subjects are proportional to each other, then the best linear combination function derived by them maximizes sensitivity uniformly over the entire range of specificity. They argued that proportionality among covariance matrices seems to be a sensible assumption when the diseased and non-diseased populations have similar performances in different aspects, but with shifted means and different scales of variance. The second scenario is for the case of non-proportional covariance matrices, there generally does not exist a dominating combination rule that is uniformly better than any others. Therefore, they derived the best linear combination coefficients that maximize AUC among all possible linear combinations. The multivariate normality assumption is crucial in that the optimality of their method is only valid when the test results for both diseased and non-diseased subjects follow multivariate normal distributions.

When comparing the diagnostic accuracy of several classifiers, it is often of interest to the investigators to confine attention to only part of the ROC curve with higher specificity. Liu, Schisterman, and Zhu (2005) pointed out that the optimal linear combination of Su and Liu may give an unsatisfactory low sensitivity on either high or low specificity areas, which is undesirable in practice. Hence, Liu, Schisterman, and Zhu (2005) started from deriving the sufficient condition for one linear combination dominating the other in terms of sensitivity over a range of specificity. They proposed a linear combination, which has higher sensitivity than the other linear combination in some specificity region. However, the dominance region depends on the linear combination being compared.

In many cases, investigators often confine their attention to a predefined clinical relevant region, say $(0, t)$ on false positive fraction where t is between zero and one. The method in Liu, Schisterman, and Zhu (2005) cannot handle this situation since the dominance region of their optimal combination depends on the linear combination being compared. Therefore, Hsu and Hsueh (2013) proposed a method to find the linear combination that maximizes the partial AUC for the pre-defined range. The analytic solution of the partial AUC can be derived under normal assumption, and they derived the first derivative of the partial AUC with respect to the linear combination coefficients. They proposed a

new algorithm to solve this fixed point problem.

Yu and Park (2015) used first derivative condition for the linear combination that maximizes partial AUC derived in Hsu and Hsueh (2013), they proposed two simple algorithms to solve the fixed point problem.

All methods, which model the probability distribution function, build on the normal assumptions. Therefore, in practice, assessing the multivariate normality is a prerequisite for these combination methods. Korkmaz, Göksülük, and Zararsiz (2014) provides a useful R package to check normality assumptions. In general, however, checking normality is not an easy task, especially for multivariate cases. It is recommended that some transformations, such as Box-Cox transformation, be applied to the input variables to improve normality.

Modeling risk score function

The other subcategory of parametric methods aims to model the risk score function. The most well known example is the generalized linear models. There are several popular link functions we can choose from, such as logit and probit link functions. Note that the generalized linear model for the risk score, is a much weaker assumption than the multivariate normal distribution assumed for the probability distribution function. That is, the generalized linear model presupposes a statistical relation such that once the multiple test results are given, then the probability that this individual belongs to diseased group is determined. The distribution of multiple test results are, therefore, irrelevant. This feature renders the generalized linear model more robust than the multivariate normal models (Su and Liu, 1993; Liu, Schisterman, and Zhu, 2005; Hsu and Hsueh, 2013; Yu and Park, 2015), but may also be less efficient due to its weaker assumptions.

Even though we do not promote one technique over another for estimating the risk score, we should recognize one important advantage of logistic regression (Walker and Duncan, 1967) for this problem, namely its ability to handle case-control study designs (Prentice and Pyke, 1979). Logistic regression is a special case of the generalized linear

model, where logit link function is used for modeling the risk score. The unique form of its logit link function is the reason that logistic regression can handle case-control study designs, and this property is not shared by other link functions. Estimates can be derived by maximizing the likelihood.

Therefore, logistic regression is the most often employed generalized linear model to combine markers for classification. It is statistically efficient if the link function is correctly specified. However, misspecification can lead to non-optimal combination even when the sample size is very large. As a result, flexible methods are therefore appealing. For example, the modified logistic regression approach of McIntosh and Pepe (2002), requires only that the risk score is modeled appropriately over a relevant subregion of the marker space and thus reduces modeling requirements. In addition, Li and Duan (1989) indicated that logistic regression is itself quite robust. That is, under certain conditions stated in their paper, logistic regression performs well even when the link function is not logit. However, in practice there is no guarantee that the conditions will be met.

2.3.2 Semiparametric methods

From a practical perspective, we do not need to assume any model for methods reviewed in this subsection. In order to conduct statistical inference of the resulting combination coefficients, however, assumptions of the risk score function will be needed. That is, a generalized linear model is usually assumed for the risk score but the link function can be left unspecified, which is why we call these methods semiparametric. Under such assumption, the estimated combination coefficients, obtained from maximizing the empirical AUC, are consistent and asymptotically normal. As a result, standard inference procedure such as confidence interval and hypothesis testing can be implemented based on its asymptotic properties.

Again, the goal is to combine multiple test results to improve diagnostic accuracy by maximizing the area or partial area under the empirical ROC curve. In addition, depending on whether the methods employ linear or nonlinear combination of the multiple tests, we

further group semiparametric methods into linear methods and nonlinear methods.

Linear methods

Pepe and Thompson (2000) proposed an approach to obtain optimal linear combination of multiple test results by maximizing the area under the empirical ROC curve, which is the well-known Mann-Whitney statistic, among all linear combinations. They also extend their approach to maximize the partial area under the empirical ROC curve, in order to confine attention to only part of the ROC curve with higher specificity. As they pointed out, the Mann-Whitney statistic is neither continuous nor concave function of its linear combination coefficients, thus a grid search rather than a derivative-based method is needed for the optimization procedure. When the number of tests is relatively large, this optimization process can be quite computational intensive.

To address the computational difficulty, Liu, Liu, and Halabi (2011) proposed a min-max combination approach that linearly combines only the minimum and maximum values of test results to maximize the Mann-Whitney statistic of AUC. Note that the ROC curve is invariant of monotone increasing transformation, therefore min-max approach only involves searching for a single coefficient and thus is computational feasible. Ma, Halabi, and Liu (2019) extended min-max approach to maximize the partial area under the ROC curve for the range of specificity of clinical interest. Although this procedure is easy to implement, it has several drawbacks as noted in Liu, Liu, and Halabi (2011). Firstly, the measurements from multiple tests need to be standardized before using this min-max approach when the tests are measured with different units. Secondly, during the implementation, information may get lost in that it only uses the minimum and maximum values of the test results. Lastly, the interpretation of the estimated linear combination coefficient can be difficult since the minimum and maximum of multiple tests may come from different test results for different subjects.

Kang, Liu, and Tian (2016) proposed a stepwise approach for linearly combining multiple tests to maximize the Mann-Whitney statistic of AUC. They combine all the diagnostic

tests in a stepwise fashion, such that information from all tests will be used and it is still computational feasible.

Ma and Huang (2007) tackled the computational difficulty of Pepe and Thompson (2000) by using a sigmoid approximation of the Mann-Whitney statistic of AUC. Though still non-concave, the approximated Mann-Whitney statistic of AUC is a continuous function of its linear combination coefficients. Maximization is achieved by gradient descend search, the gradient search will be continued until the estimate converges. Computation of the gradient is simple and fast due to the nice derivative property of sigmoid function.

When a large set of diagnostic tests is collected, one important objective is to select a subset of tests that are most significantly associated with the outcome. To enable variable selection, Zhou et al. (2012) added a penalty term in their model. They used a smooth function to approximate the Mann-Whitney statistic of AUC in a similar manner to Ma and Huang (2007).

Nonlinear methods

The linear combination methods of test results are the mainstream research direction regarding this topic in that they are easier to implement and interpret. However, they may not perform well when there is strong nonlinearity in the data. Therefore, we turn to methods that can capture informative nonlinear structures. Komori (2011) proposed a boosting method for maximization of the area under the ROC curve. In the proposed iterative procedure, various simple classifiers are combined flexibly into a single strong classifier. They also considered adding a penalty term to prevent overfitting to data. Following the same framework of boosting procedure, Komori and Eguchi (2010) extended their method for maximizing the partial area under the ROC curve.

Fong, Yin, and Huang (2016) proposed another approach to approximating the Mann-Whitney statistic of AUC to make the maximization computationally affordable. Instead of using sigmoid function in Ma and Huang (2007), they used a ramp function, which can be decomposed as the difference between two convex functions. They used a special

optimization algorithm called difference of convex functions algorithm. They showed in their simulations that it is less likely to be stuck in local optima than the gradient-based optimization used by Ma and Huang (2007). Another merit of their approach is that, by employing the "kernel trick", a commonly seen term in the SVM literature, they can map the input vector to a high dimensional feature space without having to specify the mapping explicitly, hence capture nonlinear relationship.

2.3.3 Other methods

In this part, we review methods that are that not specially designed for combining multiple tests, but can be applied equally well in practice.

Machine learning methods

A simple and common approach is to model the outcome directly against the test results and use the estimated risk score function as the combination rule for test results. Numerous machine learning algorithms can be used such as logic regression (Ruczinski, Kooperberg, and LeBlanc, 2003), classification trees (Breiman et al., 2017), neural network, support vector machines (Cristianini, Shawe-Taylor, et al., 2000) and boosting (Bartlett et al., 1998; Friedman, Hastie, and Tibshirani, 2000), to obtain the resulting estimated risk score function. Ripley (2007) provides a complete description and comparison for these algorithms.

However, there are several drawbacks from machine learning methods. Firstly, they cannot necessarily accommodate case-control designs. That is, it is possible that machine learning methods estimate different risk score function in case-control and cohort study depending on whether study subjects are selected on the basis of the outcome. Secondly, It is hard to quantify sampling variability for machine learning methods. Therefore, standard statistical inference is not applicable. Other limitations of machine learning methods include computational cost for tuning parameters, data inefficiency and suboptimality.

Maximizing Youden's index

We have reviewed parametric methods, semiparametric methods, machine learning methods in previous sections, all these methods aim at maximizing AUC or pAUC either explicitly or implicitly. The majority of articles regarding combination of multiple tests fall in this category. However, these methods only produce an optimal combination rule without specifying the diagnostic threshold. Alternatively, Yin and Tian (2014) proposed the idea of using Youden's index as an objective function for searching the optimal linear combination. The combined score directly achieves the maximum overall correct classification rate at the diagnostic threshold corresponding to Youden's index. In other words, it is the optimal linear combination score for making the disease diagnosis.

2.4 Simulation studies

Re-substitution or cross Validation?

The evaluation of combination methods in past literature has been done in two fashions: re-substitution and cross validation. The re-substitution method consists of the following steps: (1) combination rule is learned from a particular data set. (2) a composite score is calculated using the estimated combination rule on the same data set. (3) the evaluation metric is calculated based on the composite score. However, as pointed out by a few researchers (Kang, Liu, and Tian, 2016; Copas and Corbett, 2002), the estimated performance metric using the re-substitution method usually is overoptimistic for estimating the diagnostic accuracy on future observations. This is also a common phenomenon between training set and validation set in the field of machine learning.

On the other hand, cross-validation method is considered as the simplest and most widely used method for assessing how a combination rule would generalise to an independent data set and how accurately it will perform in practice. Some researchers advocated using a "leave one pair out" cross validation (LOPO CV) procedure for performance evaluation (Kang, Liu, and Tian, 2016; Huang, Qin, and Fang, 2011). The authors also pointed

out that alternative 5-fold cross validation and 10-fold cross validation can be applied instead of LOPO CV to gain computational efficiency.

We compared the performance of eight approaches, namely, Su and Liu's method (SULIU), logistic regression approach with l_2 penalty (LOGISTIC), Kang et al's stepwise approach (SW), Liu et al's min-max approach (MIN-MAX), Huang et al's RAUC approach (RAUC), and machine learning methods including random forest (RF), SVM using linear kernel (SVMl) and radial basis function (rbf) kernel (SVMr). Table 2.2 summarizes all methods that are being compared.

Table 2.2: Eight methods that are being compared in simulation studies and real data examples. l_2 penalty is applied to LOGISTIC.

	SULIU	LOGISTIC	SW	MIN-MAX	RF	SVMl	SVMr	RAUC
Category	Parametric	Parametric	Semi-parametric	Semi-parametric	Machine learning	Machine learning	Machine learning	Semi-parametric
Linear/Nonlinear	Linear	Linear	Linear	Linear	Nonlinear	Linear	Nonlinear	Nonlinear
Assumption	Normality	GLM with logit link	None	None	None	None	None	GLM
Accommodate Case-control?	Yes	Yes	Yes	Yes	No	No	No	Yes

The performance of all approaches to obtaining the largest AUC was investigated through extensive simulation studies. In particular, we implemented two simulation scenarios commonly used throughout the literature. The first scenario is to generate observations for diseased and non-diseased populations separately from different underlying distributions. The second scenario is to generate observations for all populations simultaneously, then generate outcome based on a pre-specified risk score function given its covariates. we call the first scenario as scenario A and second as scenario B.

2.4.1 Scenario A

In Scenario A, six different settings of the joint distributions of four markers were considered. For each setting, observations were generated from the underlying distribution with different sample sizes. The AUC of the combination score function was estimated from 10-fold cross validation for comparison purpose. For each setting, 1000 Monte Carlo samples were generated to calculate the mean AUC of the combination rule and its standard error (SE).

Multivariate normal distributions with equal variance

Data from multivariate normal distributions with different mean vectors and equal variance matrices for non-diseased and diseased populations were generated with the following two settings. Results are shown in Table 2.3.

$$A: u_1 = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}, u_2 = \begin{bmatrix} 0.6 \\ 0.8 \\ 1.0 \\ 1.2 \end{bmatrix} \text{ and } \Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix} \tag{2.1}$$

$$B: u_1 = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}, u_2 = \begin{bmatrix} 1.1 \\ 1.4 \\ 1.7 \\ 2.0 \end{bmatrix} \text{ and } \Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix} \tag{2.2}$$

Table 2.3 shows that, under these two settings, SULIU and LOGISTIC always produce the largest AUC on validation set. This is because both of them are guaranteed to have optimal linear combination asymptotically under normal distributions with equal variance. Note that LOGISTIC has slightly better performance than SULIU when the sample size is small. This is due to the fact that l_2 penalty is applied to LOGISTIC and, as a result, improves its generalization ability. The ROC curves measuring the performance of different methods based on 10-fold cross validation for configuration A and B are illustrated in Figure 2.2 and Figure 2.3 respectively.

Table 2.3: Mean area under receiver operating characteristic curve AUC (SE) under setting "multivariate normal with equal variance".

Sample size	Mean config	SULIU	LOGISTIC	SW	MIN-MAX	RF	SVMl	SVMr	RAUC
(20,20)	A	0.763 (0.10)	0.771 (0.10)*	0.752 (0.10)	0.734 (0.10)	0.742 (0.11)	0.729 (0.16)	0.664 (0.20)	0.643 (0.13)
	B	0.907 (0.06)	0.914 (0.06)*	0.896 (0.06)	0.884 (0.07)	0.898 (0.06)	0.908 (0.06)	0.895 (0.07)	0.812 (0.11)
(20,30)	A	0.769 (0.09)	0.775 (0.08)*	0.757 (0.09)	0.739 (0.09)	0.749 (0.09)	0.753 (0.12)	0.723 (0.13)	0.645 (0.11)
	B	0.913 (0.05)	0.917 (0.05)*	0.906 (0.05)	0.889 (0.05)	0.901 (0.05)	0.914 (0.05)	0.898 (0.05)	0.817 (0.09)
(50,50)	A	0.786 (0.05)	0.788 (0.05)*	0.780 (0.05)	0.753 (0.05)	0.77 (0.05)	0.785 (0.05)	0.753 (0.07)	0.659 (0.07)
	B	0.921 (0.03)	0.923 (0.03)*	0.919 (0.03)	0.899 (0.03)	0.912 (0.03)	0.92 (0.03)	0.906 (0.04)	0.83 (0.06)
(100,100)	A	0.797 (0.04)*	0.797 (0.04)*	0.794 (0.04)	0.761 (0.04)	0.784 (0.04)	0.796 (0.04)	0.771 (0.04)	0.671 (0.04)
	B	0.927 (0.02)*	0.927 (0.02)*	0.926 (0.02)	0.903 (0.02)	0.918 (0.02)	0.927 (0.02)	0.911 (0.02)	0.838 (0.02)

Multivariate normal distribution with unequal variance

Now we consider multivariate normal distributions with different mean vectors and unequal variance matrices for non-diseased and diseased populations. The mean configurations A and B are the same as in previous setting, with variance matrices as follows,

$$\Sigma_1 = \begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 1 & 0.7 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 & 0.7 \\ 0.7 & 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 0.7 & 1 \end{bmatrix} \tag{2.3}$$

From Table 2.4, It is interesting to observe that MIN-MAX and SVMr are superior to other methods in yielding the largest AUC under mean configuration A, while SULIU and LOGISTIC have better performance under mean configuration B. The outstanding performance of MIN-MAX under mean configuration A, suggests that it would filter out the best linear combination when non-diseased and diseased populations are not far apart and the variance of the two populations are not same. SVMr shows superior performance under mean configuration A when the sample sizes are big enough. The corresponding ROC curves for mean configuration A and B are illustrated in Figure 2.4 and Figure 2.5 respectively.

Table 2.4: Mean area under receiver operating characteristic curve AUC (SE) under setting "multivariate normal with unequal variance".

Sample size	Mean config	SULIU	LOGISTIC	SW	MIN-MAX	RF	SVMl	SVMr	RAUC
(20,20)	A	0.758 (0.10)	0.769 (0.09)	0.746 (0.10)	0.79 (0.08)*	0.755 (0.10)	0.724 (0.16)	0.723 (0.18)	0.703 (0.13)
	B	0.906 (0.06)	0.913 (0.05)*	0.896 (0.06)	0.89 (0.07)	0.898 (0.06)	0.908 (0.06)	0.904 (0.06)	0.845 (0.11)
(20,30)	A	0.766 (0.09)	0.771 (0.09)	0.754 (0.09)	0.795 (0.07)*	0.771 (0.08)	0.742 (0.13)	0.774 (0.11)	0.715 (0.12)
	B	0.911 (0.05)	0.916 (0.05)*	0.902 (0.05)	0.894 (0.05)	0.907 (0.05)	0.909 (0.05)	0.909 (0.05)	0.843 (0.09)
(50,50)	A	0.787 (0.05)	0.789 (0.05)	0.781 (0.06)	0.803 (0.05)	0.786 (0.05)	0.786 (0.05)	0.806 (0.05)*	0.725 (0.07)
	B	0.921 (0.03)	0.923 (0.03)*	0.919 (0.03)	0.903 (0.03)	0.913 (0.03)	0.921 (0.03)	0.918 (0.03)	0.860 (0.06)
(100,100)	A	0.799 (0.03)	0.799 (0.03)	0.796 (0.03)	0.811 (0.03)	0.801 (0.03)	0.798 (0.03)	0.828 (0.03)*	0.754 (0.05)
	B	0.928 (0.02)	0.928 (0.02)*	0.926 (0.02)	0.908 (0.02)	0.921 (0.02)	0.927 (0.02)	0.928 (0.02)	0.881 (0.03)

Multivariate log-normal distribution with unequal variance

In this part, we investigated the performance of different combination methods, assuming that the markers follow multivariate log-normal distributions, that is, the log-transformed markers are normal distributed. The mean configurations A and B as well as the variance

matrices are the same as in previous setting.

From Table 2.5, It is clearly shown that MIN-MAX has the best performance in yielding the largest AUC under mean configuration A, while RF is dominant under mean configuration B. It suggests that, under highly skewed probability distributions for non-diseased and diseased populations, RF would most likely produce a composite score that has the best discriminatory ability on future observations when non-diseased and diseased populations are far apart, while MIN-MAX shows superiority when the two populations are not. In contrast, SULIU, whose optimality is based on multivariate normality, has much worse performance compared with previous settings as expected. On the other hand, LOGISTIC shows robustness compared with SULIU when the working model is mis-specified. The ROC curves for mean configuration A and B are illustrated in Figure 2.6 and Figure 2.7 respectively.

Table 2.5: Mean area under receiver operating characteristic curve AUC (SE) under setting "multivariate log-normal with unequal variance".

Sample size	Mean config	SULIU	LOGISTIC	SW	MIN-MAX	RF	SVMl	SVMr	RAUC
(20,20)	A	0.719 (0.11)	0.739 (0.11)	0.735 (0.11)	0.791 (0.08)*	0.754 (0.10)	0.697 (0.16)	0.579 (0.22)	0.617 (0.13)
	B	0.847 (0.08)	0.89 (0.07)	0.879 (0.07)	0.888 (0.07)	0.90 (0.06)*	0.883 (0.07)	0.881 (0.10)	0.77 (0.12)
(20,30)	A	0.73 (0.08)	0.747 (0.09)	0.743 (0.09)	0.795 (0.07)*	0.771 (0.08)	0.713 (0.14)	0.683 (0.16)	0.614 (0.12)
	B	0.86 (0.06)	0.894 (0.06)	0.89 (0.06)	0.894 (0.06)	0.907 (0.05)*	0.889 (0.06)	0.895 (0.05)	0.753 (0.10)
(50,50)	A	0.748 (0.05)	0.763 (0.06)	0.766 (0.06)	0.804 (0.05)*	0.786 (0.05)	0.766 (0.06)	0.75 (0.09)	0.636 (0.09)
	B	0.877 (0.04)	0.905 (0.04)	0.906 (0.04)	0.901 (0.03)	0.913 (0.03)*	0.906 (0.04)	0.903 (0.04)	0.792 (0.06)
(100,100)	A	0.767 (0.04)	0.779 (0.04)	0.783 (0.04)	0.812 (0.03)*	0.801 (0.03)	0.783 (0.03)	0.781 (0.03)	0.637 (0.02)
	B	0.892 (0.03)	0.914 (0.02)	0.915 (0.02)	0.907 (0.02)	0.92 (0.02)*	0.915 (0.02)	0.911 (0.02)	0.818 (0.02)

2.4.2 Scenario B

Nonlinear logit link simulation

Instead of generating covariates for diseased and non-diseased populations separately, we generate four covariates, namely (y_1, y_2, y_3, y_4), whose components follow independent standard normal distribution, and the outcome variable D follows Bernoulli distribution with probability given by

$$\text{logit}(P(D = 1|Y)) = y_1 - y_2 - y_3 + (y_1 - y_2)^2 - y_4^4 \tag{2.4}$$

From Table 2.6, we can observe that all linear combination methods including SULIU, LOGISTIC, MIN-MAX, SW and SVMl do not perform well due to the strong nonlinear structure in data. On the other hand, nonlinear combination methods such as RF, SVMr and RAUC excel in producing larger AUC under such setting. The corresponding ROC curves are shown in Figure 2.8

Table 2.6: Mean area under receiver operating characteristic curve AUC (SE) under nonlinear logit link, with $\text{logit}(P(D=1|Y)) = y_1 - y_2 - y_3 + (y_1 - y_2)^2 - y_4^4$. The components of Y follow independent standard normal distribution.

Sample size	SULIU	LOGISTIC	SW	MIN-MAX	RF	SVMl	SVMr	RAUC
40	0.614 (0.13)	0.614 (0.13)	0.588 (0.14)	0.490 (0.15)	0.712 (0.12)*	0.523 (0.18)	0.667 (0.18)	0.709 (0.11)
50	0.631 (0.11)	0.6325 (0.11)	0.612 (0.11)	0.51 (0.13)	0.73 (0.08)	0.564 (0.17)	0.76 (0.11)*	0.712 (0.11)
100	0.65 (0.07)	0.649 (0.07)	0.638 (0.08)	0.499 (0.09)	0.774 (0.05)	0.603 (0.13)	0.811 (0.05)*	0.746 (0.08)
200	0.655 (0.05)	0.654 (0.05)	0.652 (0.05)	0.491 (0.06)	0.80 (0.03)	0.642 (0.07)	0.85 (0.03)*	0.757 (0.04)

In summary, SULIU method was developed under multivariate normality with equal or proportional variance and hence it works well when the diseased and non-diseased populations follow normal distributions with equal variance. The performance of SULIU deteriorates as the true distributions deviate from normality. LOGISTIC is more robust compared to SULIU method, it performs well even under settings where the working model is slightly mis-specified. However, when there is strong non-linearity in the data, LOGISTIC method will deteriorate as well. The MIN-MAX method excels in some scenarios involving highly skewed multivariate data and multivariate normal data with relatively close population means and unequal variance matrices. However, this method is not very stable in general. SVMr and RAUC excel in settings where the data shows strong nonlinearity, but it performs badly when the true relationship is relatively simple, in part because they overfit the model and hence have bad generalizability on future observations. On the other hand, RF is the most robust method in that it shows comparable performance consistently across different scenarios, and its performance dominates when the data shows strong nonlinearity.

2.5 Analysis of NASH data: An example

In this Section, the approaches compared in simulation studies were applied to a real data set. 186 patients (74 Male, 112 Female, mean age 49 yrs) with nonalcoholic fatty liver disease (NALFD) were included in the study. We aim to combine several non-invasive measurements to improve diagnostic capacity for advanced fibrosis. These variables include a composite clinical score Fibrosis-4 (FIB-4), ordinal assessment of computed tomography (CT) images by two readers (from 1 to 5), and quantitative imaging biomarkers (QIBs) such as liver segmental volume ratio (LSVR) and liver surface nodularity (LSN). The four markers to be combined are "R1-NASH", "R2-NASH", "FIB4" and "LSVR". The data was processed by a log transformation to increase normality.

The performance of different methods were first evaluated by re-substitution method. To investigate the performance of different methods on future observations, we also conducted 10-fold cross validation for comparison purpose. Results are shown in Table 2.7.

Table 2.7: Re-substitution and 10-fold cross validation results for NASH data.

		SULIU	LOGISTIC	SW	MIN-MAX	RF	SVMl	SVMr	RAUC
Re-substitution	Empirical AUC	0.73	0.72	0.74	0.64	0.79*	0.73	0.74	0.73
10-fold CV	Mean AUC	0.713	0.716*	0.710	0.576	0.659	0.705	0.676	0.560

It is clearly shown that RF method gives largest re-substitution AUC, however, as we have discussed, the estimated AUC using re-substitution usually is overoptimistic for estimating the diagnostic accuracy on future observations, which is verified by results from 10-fold cross validation. On the other hand, we can see that LOGISTIC method produces largest 10-fold cross validation mean AUC, and SULIU also shows comparable performance. This suggests that the true relationship in data is more or less linear and normality is approximately satisfied after log transformation. The ROC curve measuring the performance of different methods based on 10-fold cross validation is illustrated in Figure 2.1.

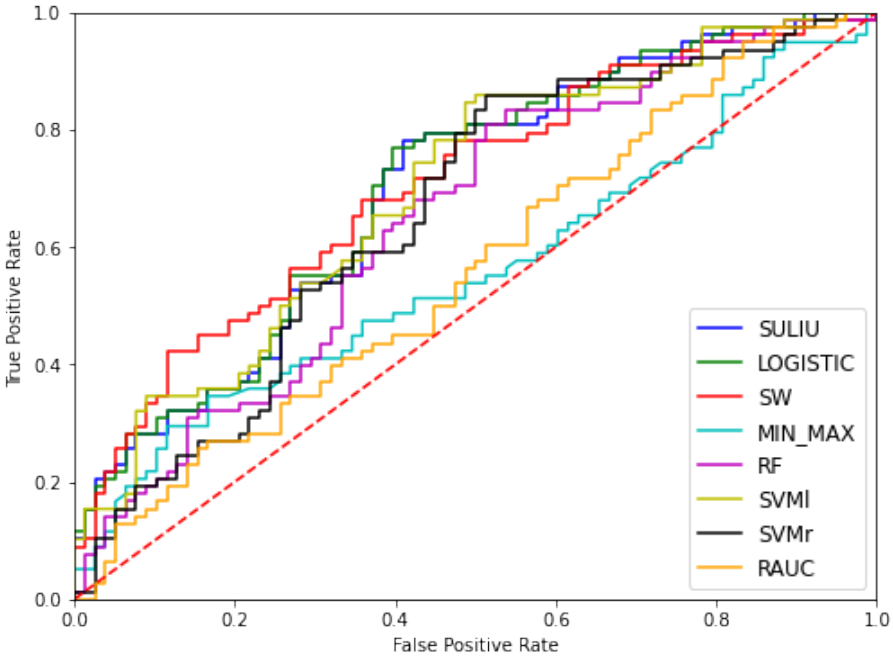

Figure 2.1: ROC curves for NASH data based on 10-fold cross validation. Methods used are "SULIU", "LOGISTIC", "SW", "MIN-MAX", "RF", "SVMl", "SVMr", "RAUC".

2.6 Useful Libraries and Resources

We review available computing packages for methods of combining multiple tests. Unfortunately, not many articles provide specific computing package to implement their methods. RAUC (Fong, Yin, and Huang, 2016) and SAUC (Ma and Huang, 2007) can be implemented in R package called "aucm", which is available on CRAN `https://cran.r-project.org/web/packages/aucm`. The code of AUCBoost (Komori, 2011) can be downloaded from `https://github.com/moriiism/hsc`. The R code to implement SW can be found in the Appendix of Kang, Liu, and Tian (2016). However, until now the computation packages and codes for implementation of different methods have been scattered across different platforms. To solve this problem, we developed an useful wrapper for most state of the art methods, which is R and Python friendly. The wrapper can be downloaded at `https://github.com/chengning-zhang/Combining-multiple-tests-Wrapper-`

2.7 Conclusions and Perspective

Statistical method for combining multiple diagnostic tests is a very active area of research in epidemiology, medicine and radiology. We will end this chapter by tying the survey back to the central research questions that posed at the end of the introduction.

What assumptions are typically made to facilitate the methods of combining multiple tests? For parametric methods, researchers model either the probability distribution for test results in the diseased and non-diseased populations or the risk score function. On the other hand, semiparametric methods do not require modeling from a practical perspective. However, in order to conduct statistical inference of the estimators from semiparametric methods, a generalized linear model is often assumed for the risk score, but the link function can be left unspecified.

Whether the combination method can handle case-control study design? AUC-based methods such RAUC, SAUC, SW, and MIN-MAX can accommodate case-control designs, in that these methods are derived from the ROC curve and the ROC curve does not depend on disease prevalence. SULIU can also handle case-control designs in that it models the probability distributions of test results which do not depend on disease prevalence as well. Logistic regression is another approach that excels in handling case-control study designs. However, this desired property is not shared for other generalized linear models. In general, machine learning methods cannot accommodate case-control designs, and this is one of the reasons that they are not the research mainstream regarding this topic.

Whether the scalar-valued function is a linear or non-linear combination of predictors? Most approaches aim at obtaining a linearly combination of multiple test results in that they are easier to implement and interpret. RAUC and boosting method are able to capture non-linear structures but they require careful hyper-parameter tuning. Machine learning methods such as RF and SVM can also learn non-linear relationships.

Whether statistical inference can be performed? In general, statistical inference cannot be conducted for machine learning methods. Methods such as MIN-MAX and SW are purely motivated from computational perspective, hence no statistical inference can be

performed as well. On the other hand, some methods require certain degree of modeling or assumptions to enable statistical inference. For example, normality is needed for SULIU. Generalized linear model with link function unspecified is assumed for SAUC and RAUC. For logistic regression, one has to assume the underlying link function is logit.

Whether variable selection can be performed? Most methods do not focus on variable selection, Zhou et al. (2012) is the first paper with main focus on variable selection by adding a penalty term to achieve sparsity.

Given that methods for combining multiple tests are gaining popularity in different fields due to biotechnologic advancements, it should continue to be an active area of medical research. The key open questions will revolve around making sure the assumptions and settings considered align with real-world tasks. Therefore, there are several key directions that research about methods for combining multiple tests could take, which we now expand.

As shown in the simulation studies, machine learning methods work fairly well in combining multiple test results. However, these methods generally suffer from lack of inference and inability to accommodate case-control study designs. There are already many studies discussing how to draw inference from machine learning models and thus the first problem can be mitigated. On the other hand, few approaches are available for machine learning methods to handle case-control scenarios. Therefore, a potential future direction could be designing machine learning methods which can accommodate case-control scenarios by making reasonable assumptions.

It is shown in this survey that the majority of papers consider linear combination of multiple test results in that they are easier to implement and interpret. However, linear combination may not work well when there is strong nonlinearity in the data. While some researchers proposed methods which are able to learn non-linear structure in the data, statistical inference is not readily available. Therefore, another future direction could be to enable the inference of nonlinear combination methods.

To date, most combination methods aim at maximizing a particular summary index

of the ROC curve (AUC/pAUC/Youden's index). Majority of papers consider AUC and pAUC while only one paper aims at maximizing Youden's index. However, methods aiming at maximizing AUC/pAUC only produce an optimal combination rule without specifying the diagnostic threshold. On the other hand, using Youden's index as an objective function for searching the optimal linear combination could achieve the maximum overall correct classification rate at the diagnostic threshold corresponding to Youden's index. Therefore, another future direction could be using Youden's index as objective function for searching the optimal linear combination.

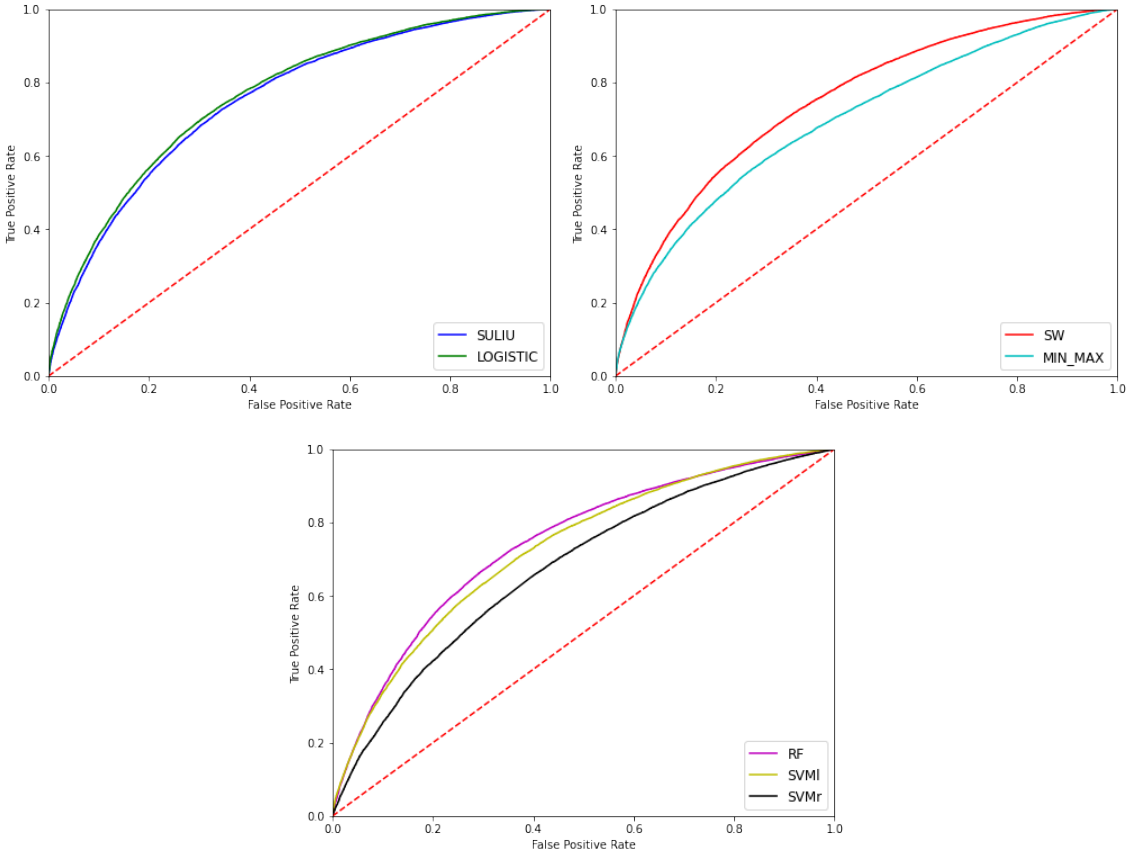

Figure 2.2: ROC curves for simulation data under multivariate normal distribution with equal variance and mean configuration A, based on 10-fold cross validation. Methods used are "SULIU", "LOGISTIC", "SW", "MIN-MAX", "RF", "SVMl", "SVMr". The numbers of positive examples and negative examples are 20 and 20.

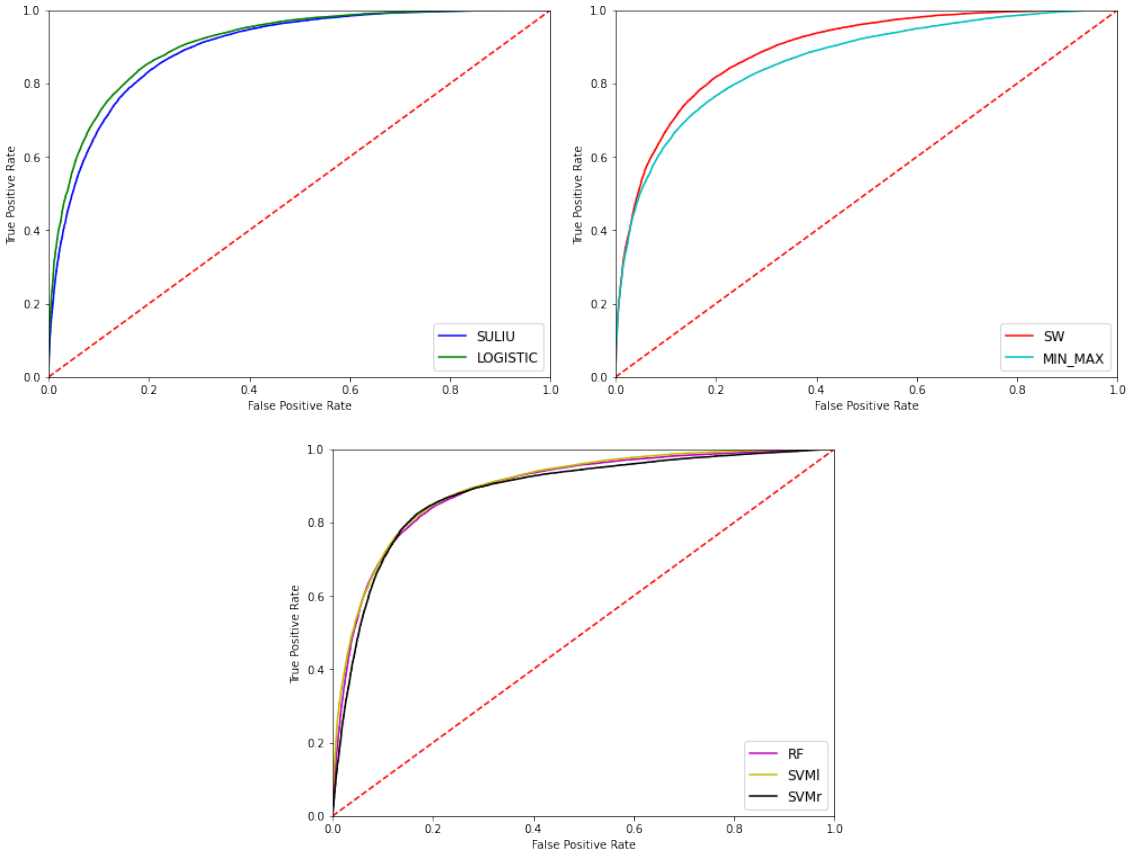

Figure 2.3: ROC curves for simulation data under multivariate normal distribution with equal variance and mean configuration B, based on 10-fold cross validation. Methods used are "SULIU", "LOGISTIC", "SW", "MIN-MAX", "RF", "SVMl", "SVMr". The numbers of positive examples and negative examples are 20 and 20.

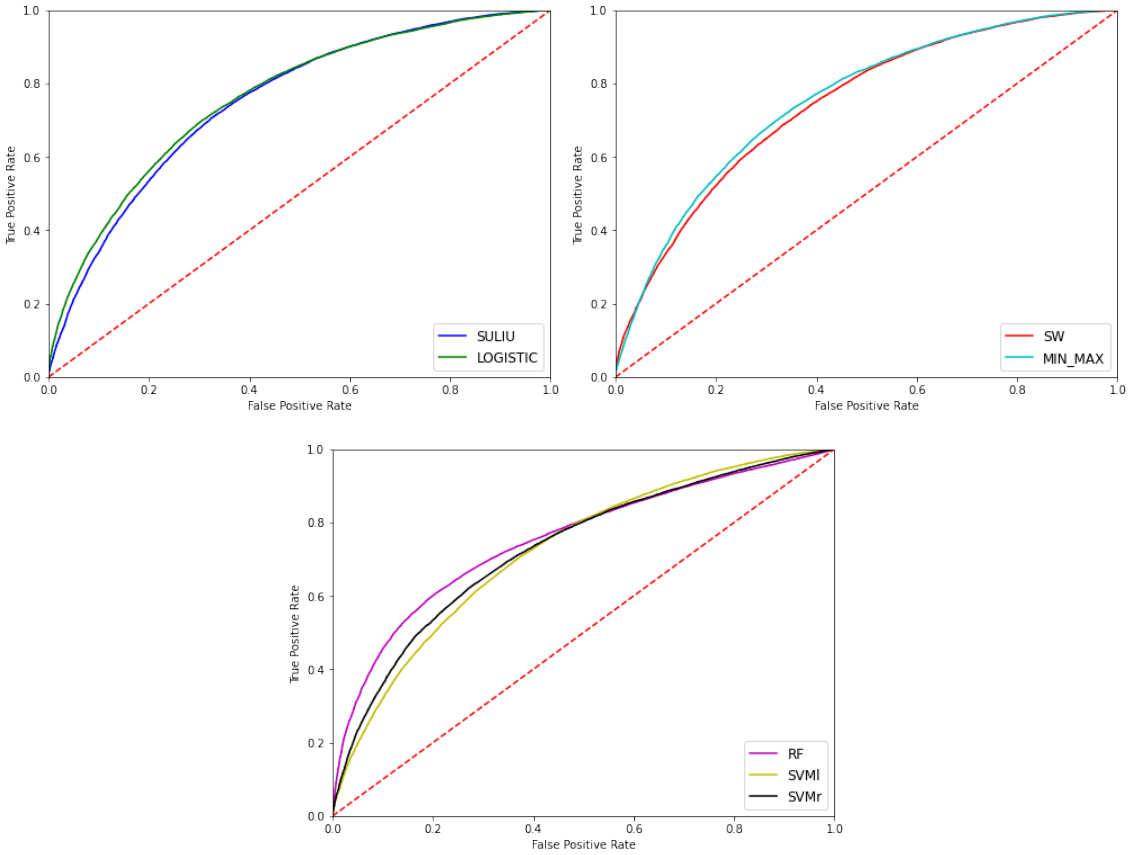

Figure 2.4: ROC curves for simulation data under multivariate normal distribution with unequal variance and mean configuration A, based on 10-fold cross validation. Methods used are "SULIU", "LOGISTIC", "SW", "MIN-MAX", "RF", "SVMl", "SVMr". The numbers of positive examples and negative examples are 20 and 20.

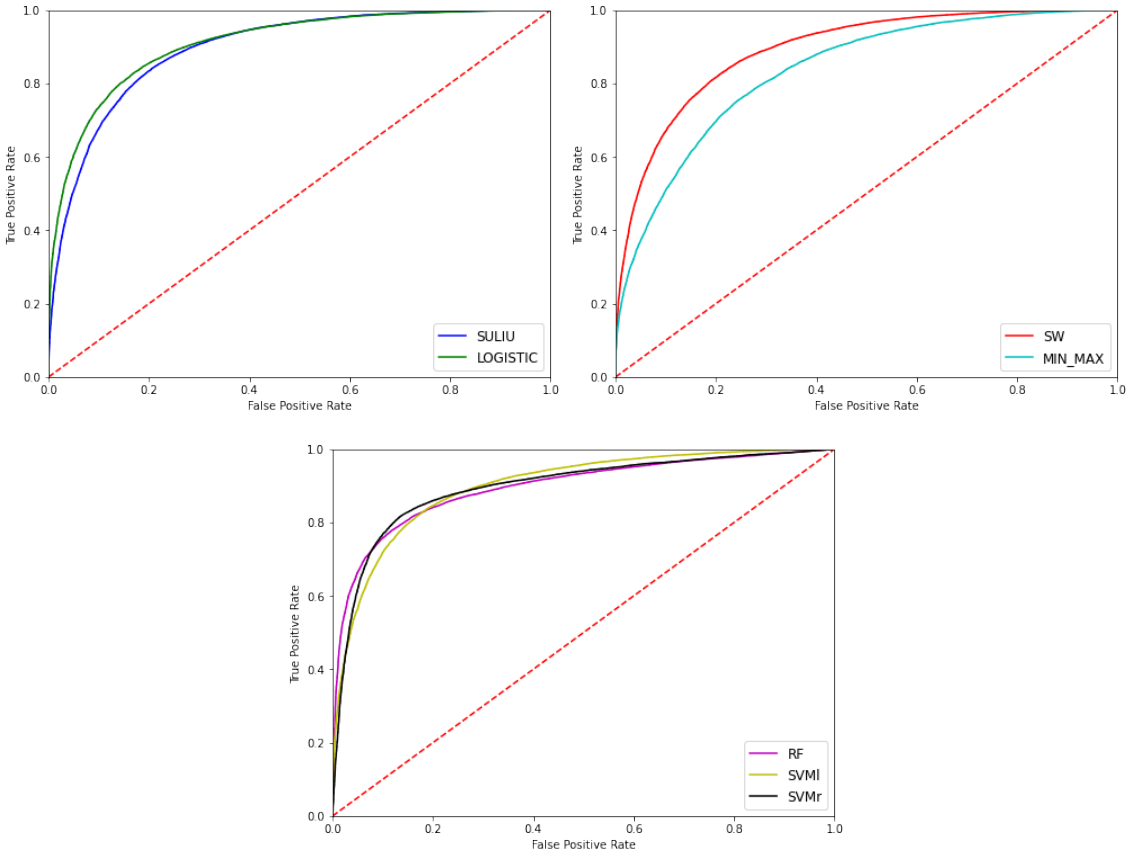

Figure 2.5: ROC curves for simulation data under multivariate normal distribution with unequal variance and mean configuration B, based on 10-fold cross validation. Methods used are "SULIU", "LOGISTIC", "SW", "MIN-MAX", "RF", "SVMl", "SVMr". The numbers of positive examples and negative examples are 20 and 20.

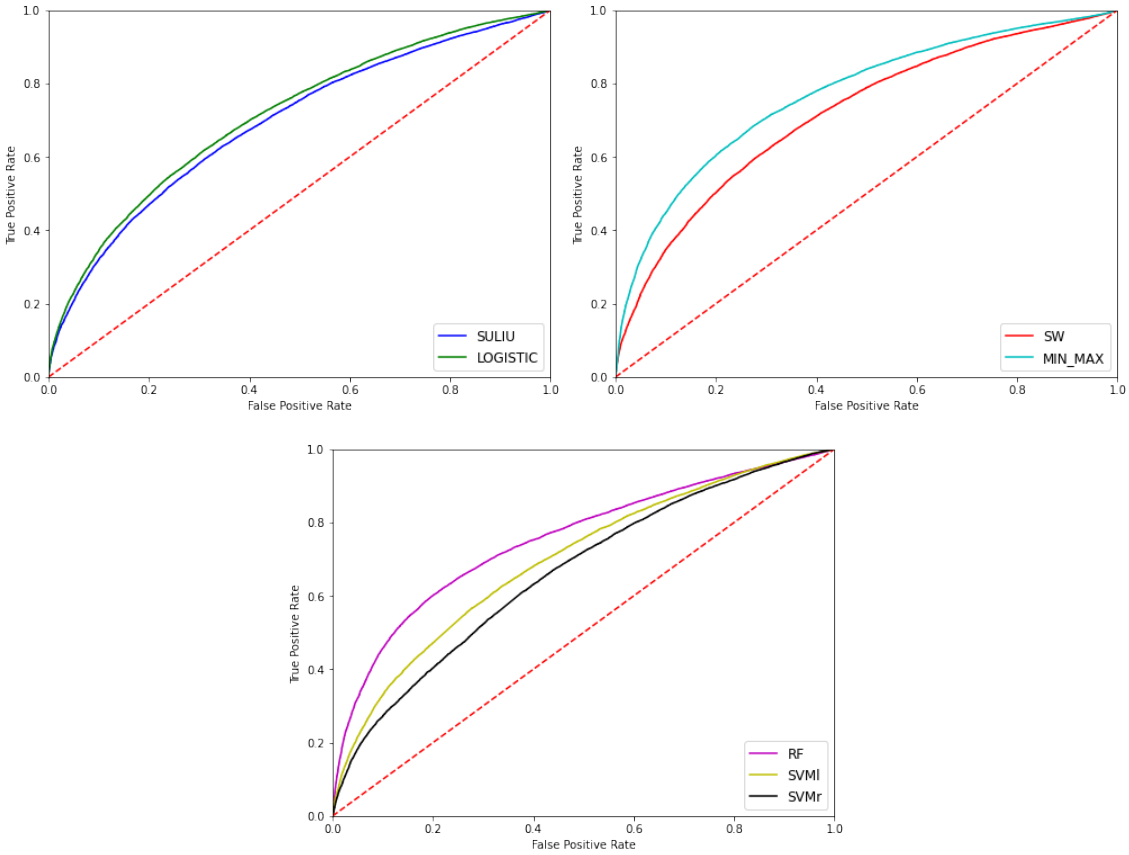

Figure 2.6: ROC curves for simulation data under log-normal distribution with unequal variance and mean configuration A, based on 10-fold cross validation. Methods used are "SULIU", "LOGISTIC", "SW", "MIN-MAX", "RF", "SVMl", "SVMr". The numbers of positive examples and negative examples are 20 and 20.

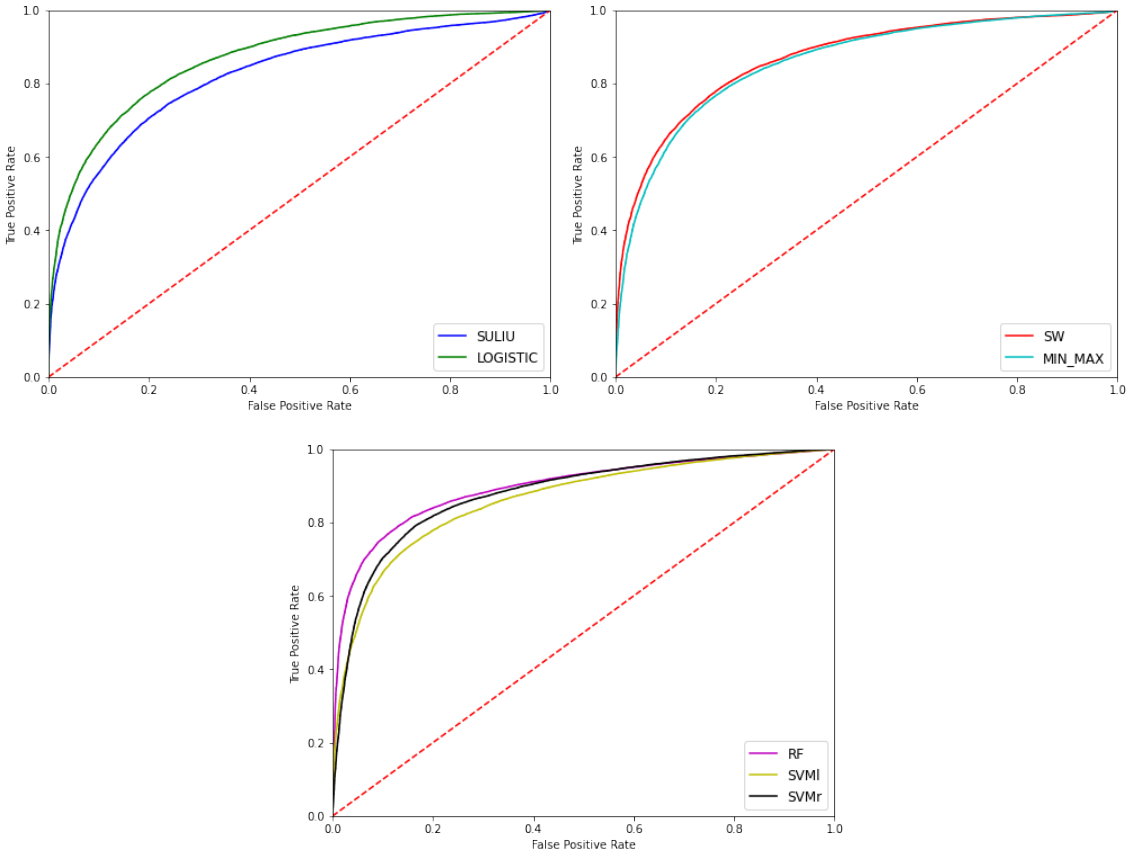

Figure 2.7: ROC curves for simulation data under log-normal distribution with unequal variance and mean configuration B, based on 10-fold cross validation. Methods used are "SULIU", "LOGISTIC", "SW", "MIN-MAX", "RF", "SVMl", "SVMr". The numbers of positive examples and negative examples are 20 and 20.

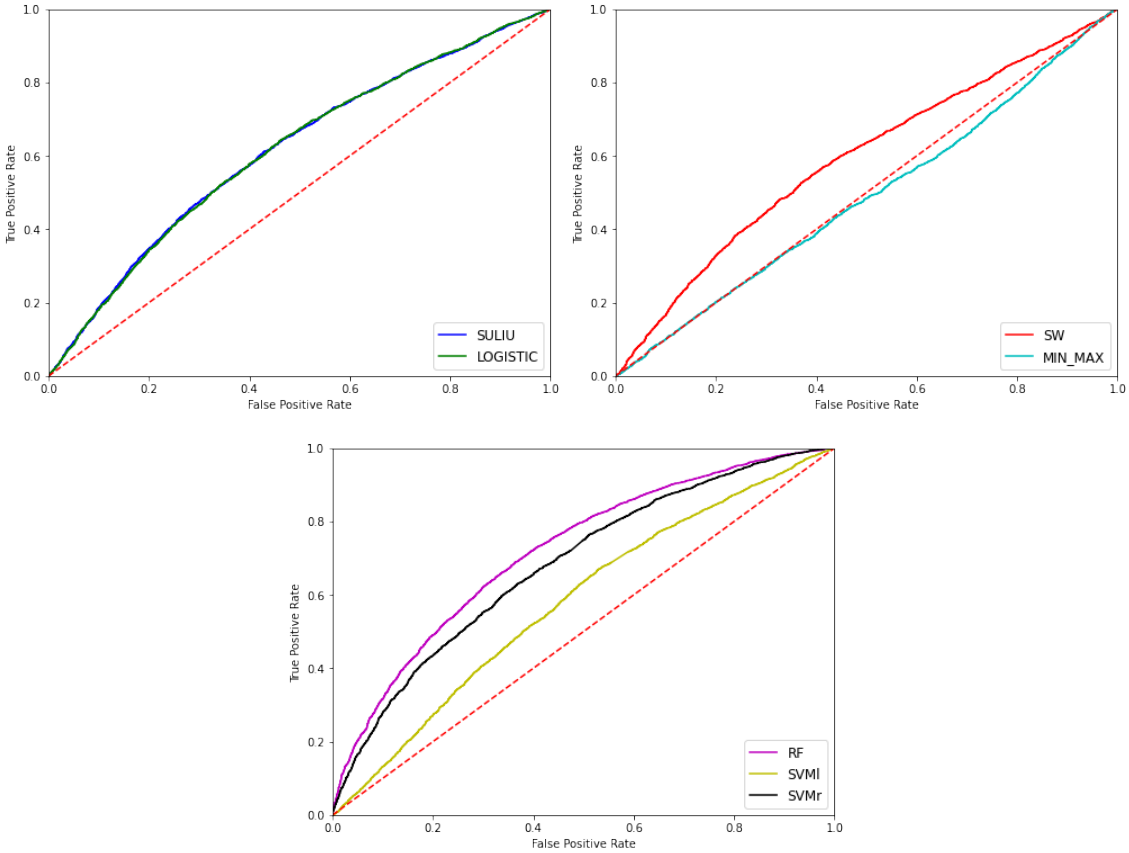

Figure 2.8: ROC curves for simulation data where the components of Y follow independent standard normal distribution and the risk score function follows $\text{logit}(P(D = 1|Y)) = y_1 - y_2 - y_3 + (y_1 - y_2)^2 - y_4^4$, based on 10-fold cross validation. Methods used are "SULIU", "LOGISTIC", "SW", "MIN-MAX", "RF", "SVMl", "SVMr". The number of examples is 40.

Chapter 3

Monotone classification with discrete missing at random covariates

3.1 Introduction

Disease diagnosis and screening have long been a central topic of medical research. In many cases, a single medical test may not be accurate enough to be used alone in practice. As a result, investigators find it necessary to combine multiple tests or biomarkers to boost diagnostic performance (Pepe, Cai, and Longton, 2006).

There is no shortage of statistical methods for building classification models out of multiple input variables. In binary classification, a simple and common approach is to model the outcome directly against the covariates using, e.g., logistic regression (Richards, Hammitt, and Tsevat, 1996), support vector machine (SVM) (Cristianini, Shawe-Taylor, et al., 2000), or classification trees (Breiman et al., 2017). These methods, however, cannot guarantee optimality in terms of the area under the receiver operating characteristic (ROC) curve (or AUC) (Pepe, 2003), a common metric of diagnostic accuracy. Alternatively, a number of authors have studied classification rules that aim explicitly to maximize the

AUC. Under the assumption of multivariate normality for the input vector, for example, Su and Liu (1993) derived the AUC-optimal linear classification rule in closed form, thereby allowing it to be easily estimated through the means and variances of the input distributions. In the general setting, Pepe and Thompson (2000) proposed to construct linear classifiers by directly maximizing the empirical AUC with respect to the coefficients of combination. This optimization problem, however, is numerically challenging as the empirical AUC is a non-smooth function of the coefficients. To alleviate the computational burden, Liu, Liu, and Halabi (2011) reduced the dimension of the input variables by using only their maximum and minimum values to fit the model (i.e., the "min-max" method). A more refined and yet numerically feasible approach was proposed by Kang, Liu, and Tian (2016), whose model incorporates input variables in a sequential, stepwise fashion, with each step solving a one-dimensional optimization problem.

All the aforementioned methods rely on linear models, in one way or another, that are mathematically expedient rather than substantively driven. In practice, it often proves fruitful to exploit established domain knowledge for gains in diagnostic accuracy. One important instance of such domain knowledge is monotonicity between the input variables and the disease outcome. For example, larger tumors as measured by magnetic resonance imaging (MRI) volumetry (Soutter et al., 2004; desouza et al., 2006) or higher cerebral β-amyloid levels as measured by positron emission tomography (PET) (Vlassenko, Benzinger, and Morris, 2012; Huynh and Mohan, 2017; Aschenbrenner et al., 2018) usually indicate greater risks of cancer or Alzheimer's disease, respectively. In fact, monotone patterns of this kind are common with general quantitative biomarkers derived from imaging modalities. Such biomarkers are typically designed to measure degree of certain anatomical anomaly and, as a result, are likely to correlate monotonically with the probability of disease (Sullivan et al., 2015; Kessler et al., 2015; Raunig et al., 2015; Obuchowski et al., 2015a; Obuchowski et al., 2015b; Huang et al., 2015).

A simple way to draw on monotone relationships is to embed suitable constraints in standard classification models such as the SVM (Chen and Li, 2014; Li and Chen, 2014).

The main weakness of this approach is that the modified classifier inherits the same assumptions from the parent model and may thus perform poorly if the parent model is wrongly specified. In the interest of robustness, some authors, mostly computer scientists, have attempted at monotone classification in a completely nonparametric setting. To date, the most successful examples are the ordinal stochastic dominance learner (OSDL) (Lievens, De Baets, and Cao-Van, 2008) and MOCA (presumably short for "monotone classifier") (Barile and Feelders, 2008). Both methods start off by nonparametrically estimating the conditional distribution of an ordinal outcome subject to a stochastic monotonicity constraint with the covariates, with only minor difference in the method of estimation. Then, the classified label is taken to be the median of the estimated conditional distribution. This leads to monotone classifiers whose output is nondecreasing with respect to each component of the covariates. Moreover, it is shown that the classifiers are asymptotically L_1-optimal in the sense of minimizing the expected absolute deviation between the assigned and true labels (Dykstra, Hewett, and Robertson, 1999).

A major limitation of OSDL and MOCA is that they cannot easily accommodate missing covariates, which are the norm rather than exception in studies of diagnostic medicine. In a recent study involving 186 liver disease patients at the University of Wisconsin (UW) Hospitals, for example, 16.2% of the participants have missing values in at least one of several computed tomography (CT) biomarkers or Fibrosis-4 scores needed for the diagnosis of non-alcoholic steatohepatitis (NASH). In "monotonizing" the conditional outcome distribution, both OSDL and MOCA rely on traditional isotonic regression techniques (Barlow and Brunk, 1972), with no natural strategy to handle missing data other than to exclude them. If the missingness is not completely at random, such "complete-case analysis" will incur bias in the estimated outcome distribution, which in turn worsens the diagnostic performance of the associated classifier.

To address this limitation, we develop a novel likelihood-based approach to nonparametric monotone classification that accounts for missing covariates in a natural and principled way. For ease of exposition, we focus on the case of binary classification, though

the proposed approach can be easily extended to multiclass cases via strategy introduced in Frank and Hall (2001). Like OSDL and MOCA, our approach starts with nonparametric estimation of the outcome distribution, or equivalently in the binary case, the risk score function; Unlike OSDL and MOCA, however, we obtain the estimates through the nonparametric maximum likelihood estimation (NPMLE) rather than isotonic regression. A novel expectation-maximization (EM)-type algorithm (Dempster, Laird, and Rubin, 1977) is devised to compute the NPMLE by treating the monotonicity-constrained risk score function as a cumulative distribution for a latent random vector. Not only is this EM-type algorithm numerically more stable and efficient than traditional quadratic programming in the presence of many covariates, but more importantly it provides a natural platform to handle missing data — we simply add another E-step to compute the conditional expectation of the coarsened input given the observed data.

The rest of the chapter is organized as follows. In Section 3.2, we introduce the NPMLE of the risk score function and present its EM-type algorithm. In Section 3.3 simulation studies are conducted to evaluate the proposed approach against state-of-art methods for combining multiple inputs. A real data example from a recent nonalcoholic fatty liver disease (NALFD) study is presented in Section 3.4. Concluding remarks are given in Section 3.5.

3.2 Theory and methods

3.2.1 EM algorithm for MC with fully observed covariates

We shall first consider the situation where covariates are fully observed. Let D denote the binary status of disease, i.e., $D = 1$ if diseased and $D = 0$ if otherwise. We use $\boldsymbol{X} \in \mathscr{X} \subseteq \mathscr{R}^p$ to denote a p-dimensional random vector of input variables. The observed data are $\{D_i, \boldsymbol{X}_i\},\ i = 1, \cdots, n$, where n is the sample size. Let $f(.\ ; c) : \mathscr{X} \to \{1, 0\}$ denote a generic classifier indexed by a threshold c, and R denote the risk score function defined as $R(\boldsymbol{X}) = P(D = 1|\boldsymbol{X})$.

In order to exploit the monotonicity relationship between input variables and the outcome variable, we restrict our attention to monotonic classifiers satisfying the following condition:

$$f(\boldsymbol{X}_1\,;c) \geq f(\boldsymbol{X}_2\,;c) \qquad \forall \boldsymbol{X}_1 \geq \boldsymbol{X}_2 \tag{3.1}$$

with the latter inequality operates component-wise. That is, a monotonic classifier is constrained by the sole requirement that, if input $\boldsymbol{X}$ is labeled as disease, then an increase in any of its components must not reverse the label. Clearly, if the risk score function R is a monotonic function, then the binary classifiers constructed by thresholding R, i.e., $I[R(\boldsymbol{X}) > c]$, are also monotonic for any fixed c. Therefore, it is sufficient to focus on classifiers with monotonic risk score function, which is

$$R(\boldsymbol{X}_1) \geq R(\boldsymbol{X}_2) \qquad \forall \boldsymbol{X}_1 \geq \boldsymbol{X}_2 \tag{3.2}$$

with the latter inequality operates component-wise. Such monotonicity assumption on the risk score reflects the a prior belief that the likelihood of disease increases with the input values. In addition, McIntosh and Pepe (2002) pointed out that classifiers of the form $I[R(\boldsymbol{X}) > c]$ yield optimal decision rules based on $\boldsymbol{X}$ in the sense that they maximize the sensitivity over the entire specificity range uniformly. Thus, the problem of (optimal) monotone classification becomes tantamount to the estimation of R under the monotonicity constraint.

The joint likelihood of observed data $(D_i, \boldsymbol{X}_i)$ can be factorized as

$$\begin{aligned} P(D_i, \boldsymbol{X}_i) &= P(D_i|\boldsymbol{X}_i)q(\boldsymbol{X}_i) \\ &= R(\boldsymbol{X}_i)^{D_i}\bar{R}(\boldsymbol{X}_i)^{1-D_i}q(\boldsymbol{X}_i) \\ &\propto R(\boldsymbol{X}_i)^{D_i}\bar{R}(\boldsymbol{X}_i)^{1-D_i} \end{aligned} \tag{3.3}$$

where q denotes the density function of $\boldsymbol{X}$ and $\bar{R}(.) = 1 - R(.)$. We aim to estimate the risk score R by maximizing the observed data likelihood under monotonicity constraint

given by Equation 3.2, that is,

$$\begin{aligned}\hat{R} &= \arg\max_{R\in M_p} \sum_{i=1}^{n} \log P(D_i, \boldsymbol{X}_i) \\ &= \arg\max_{R\in M_p} \sum_{i=1}^{n} D_i \log R(\boldsymbol{X}_i) + (1-D_i)\log \bar{R}(\boldsymbol{X}_i)\end{aligned} \tag{3.4}$$

where M_p is the space of all non-decreasing functions on input space bounded in $[0,\ 1]$. The factorization of observed data likelihood shows the inference of R has nothing to do with the density of $\boldsymbol{X}$.

Direct optimization under the monotonicity constraint, however, is not a trivial task. Fortunately, note that M_p is also precisely the space of all cumulative distribution functions (CDF) on $\mathscr{X}$. This leads us to construct an independent latent random vector $\boldsymbol{U}$ with $P(\boldsymbol{U} \le \boldsymbol{x}) = R(\boldsymbol{x})$. In other words, we introduce a latent random vector $\boldsymbol{U}$ and treat the monotonic risk score function as cumulative distribution function (CDF) of $\boldsymbol{U}$, thereby making the likelihood in Equation 3.3 identical to that of $\{D_i = I(\boldsymbol{U}_i \le \boldsymbol{X}_i), \boldsymbol{X}_i\},\ i = 1, \cdots, n$. Now we can derive an EM-type algorithm with the $\boldsymbol{U}_i$ treated as missing data.

Without loss of generality, we assume that both $\boldsymbol{U}$ and $\boldsymbol{X}$ are supported on $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m$ and write $R(d\boldsymbol{x}_k) = P(\boldsymbol{U} = \boldsymbol{x}_k),\ k = 1, \cdots, m$. Then, the nonparametric estimation of R reduces to that of the m parameters $R(d\boldsymbol{x}_k),\ k = 1, \cdots, m$ with the understanding that $R(\boldsymbol{x}) = \sum_{\boldsymbol{x}_k \le \boldsymbol{x}} R(d\boldsymbol{x}_k)$. First we consider the M step. Suppose that we can observe the full data $(\boldsymbol{U}, \boldsymbol{X})$, the full data log-likelihood of $(\boldsymbol{U}, \boldsymbol{X})$ is

$$\sum_{i=1}^{n} \log P(\boldsymbol{U} = \boldsymbol{U}_i) + \sum_{i=1}^{n} \log P(\boldsymbol{X} = \boldsymbol{X}_i) \tag{3.5}$$

The full data MLE of the point mass of $\boldsymbol{U}$ can be easily derived, which is $\hat{R}(d\boldsymbol{x}_k) = \sum_{i=1}^{n} I(\boldsymbol{U}_i = \boldsymbol{x}_k)/n,\ k = 1, \cdots, m$. Then, for the E step at the $(j+1)$th iteration, we take expectation of the full data MLE of the point mass of $\boldsymbol{U}$ given observed data $(D_i, \boldsymbol{X}_i)$

and the current estimates of the point mass of $\boldsymbol{U}$. That is,

$$R^{(j+1)}(d\boldsymbol{x}_k) = \frac{\sum_{i=1}^{n} E[I(\boldsymbol{U}_i = \boldsymbol{x}_k)|D_i, \boldsymbol{X}_i, R^{(j)}]}{n} \tag{3.6}$$

By standard derivation, the $(j+1)$th iteration of the algorithm becomes

$$R^{(j+1)}(d\boldsymbol{x}_k) = \frac{R^{(j)}(d\boldsymbol{x}_k)}{n} \sum_{i=1}^{n} \Big\{ \frac{D_i I(\boldsymbol{x}_k \leq \boldsymbol{X}_i)}{R^{(j)}(\boldsymbol{X}_i)} + \frac{(1-D_i) I(\boldsymbol{x}_k \not\leq \boldsymbol{X}_i)}{\bar{R}^{(j)}(\boldsymbol{X}_i)} \Big\} \tag{3.7}$$

3.2.2 EM algorithm for MC with discrete missing at random covariates

To make our algorithm to allow for discrete missing at random (MAR) covariates, we introduce, for the ith subject, M_i representing the missing pattern of covariates, so that the observed data are $\{D_i, M_i, M_i(\boldsymbol{X}_i)\},\ i = 1, \cdots, n$. The joint likelihood of observed data can be factorized as

$$\begin{aligned} P(D_i, M_i, M_i(\boldsymbol{X}_i)) &= \sum_{\boldsymbol{x}_l \in \mathscr{X}_i} P(D_i, M_i, \boldsymbol{X}_i = \boldsymbol{x}_l) \\ &= \sum_{\boldsymbol{x}_l \in \mathscr{X}_i} P(D_i, \boldsymbol{X}_i = \boldsymbol{x}_l) P(M_i | D_i, \boldsymbol{X}_i = \boldsymbol{x}_l) \\ &= P(M_i | D_i, M_i(\boldsymbol{X}_i)) \sum_{\boldsymbol{x}_l \in \mathscr{X}_i} R(\boldsymbol{x}_l)^{D_i} \bar{R}(\boldsymbol{x}_l)^{1-D_i} q(\boldsymbol{x}_l) \\ &\propto \sum_{\boldsymbol{x}_l \in \mathscr{X}_i} R(\boldsymbol{x}_l)^{D_i} \bar{R}(\boldsymbol{x}_l)^{1-D_i} q(\boldsymbol{x}_l) \end{aligned} \tag{3.8}$$

where $\mathscr{X}_i$ denotes the set of all possible covariates values that are "compatible" with the observed values $M_i(\boldsymbol{X}_i)$ of the ith subject, namely, $\mathscr{X}_i = \{\boldsymbol{x} \in \mathscr{X} : M_i(\boldsymbol{x}) = M_i(\boldsymbol{X}_i)\}$. The third equality holds by assuming covariates are MAR. Again, we aim to estimate the risk score R by maximizing the observed data likelihood under monotonicity constraint,

that is

$$\begin{aligned}\hat{R} &= \arg\max_{R\in M_p} \sum_{i=1}^{n} \log P(D_i, M_i, M_i(\boldsymbol{X}_i)) \\ &= \arg\max_{R\in M_p} \sum_{i=1}^{n} \Big\{ D_i \log\big[\sum_{\boldsymbol{x}_l\in\mathscr{X}_i} R(\boldsymbol{x}_l)q(\boldsymbol{x}_l)\big] + (1-D_i)\log\big[\sum_{\boldsymbol{x}_l\in\mathscr{X}_i} \bar{R}(\boldsymbol{x}_l)q(\boldsymbol{x}_l)\big]\Big\}\end{aligned} \tag{3.9}$$

The factorization of observed data likelihood shows that, with MAR covariates, q is entangled with the joint density and thus cannot be ignored.

Similar to that of Section 3.2.1, it is difficult to directly solve this constrained optimization problem. Again, we can construct an independent latent random vector $\boldsymbol{U}$ with $P(\boldsymbol{U} \le \boldsymbol{x}) = R(\boldsymbol{x})$. Note that the log-likelihood in Equation 3.8 is identical to that of $\{D_i = I(\boldsymbol{U}_i \le \boldsymbol{X}_i), M_i, M_i(\boldsymbol{X}_i)\},\ i = 1, \cdots, n$. Now we can derive an EM-type algorithm with the $\boldsymbol{U}_i$ and $\boldsymbol{X}_i$ treated as missing data.

Suppose again that both $\boldsymbol{U}$ and $\boldsymbol{X}$ are supported on $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m$ and write $q_k = P(\boldsymbol{X} = \boldsymbol{x}_k),\ R(d\boldsymbol{x}_k) = P(\boldsymbol{U} = \boldsymbol{x}_k),\ k = 1, \cdots, m$. Then, the nonparametric estimation of R requires estimation of $2m$ parameters, namely, q_k and $R(d\boldsymbol{x}_k)$. The M step involves separate maximization regarding the point mass of $\boldsymbol{U}$ and $\boldsymbol{X}$, respectively, with the full data log-likelihood of $(\boldsymbol{U}, \boldsymbol{X})$ shown in Equation 3.5. That is, $\hat{R}(d\boldsymbol{x}_k) = \sum_{i=1}^{n} I(\boldsymbol{U}_i = \boldsymbol{x}_k)/n$ and $\hat{q}_k = \sum_{i=1}^{n} I(\boldsymbol{X}_i = \boldsymbol{x}_k)/n,\ k = 1, \cdots, m$. For the E step at $(j+1)$th iteration, we need to take expectation of the full data MLE of the point mass of $\boldsymbol{U}$ and $\boldsymbol{X}$, given $\{D_i, M_i, M_i(\boldsymbol{X}_i)\}$ and the current estimates of the point mass of $\boldsymbol{U}$ and $\boldsymbol{X}$.

$$\begin{aligned} R^{(j+1)}(d\boldsymbol{x}_k) &= \frac{\sum_{i=1}^{n} E[I(\boldsymbol{U}_i = \boldsymbol{x}_k)|D_i, M_i, M_i(\boldsymbol{X}_i), R^{(j)}, q^{(j)}]}{n} \\ q_k^{(j+1)} &= \frac{\sum_{i=1}^{n} E[I(\boldsymbol{X}_i = \boldsymbol{x}_k)|D_i, M_i, M_i(\boldsymbol{X}_i), R^{(j)}, q^{(j)}]}{n} \end{aligned} \tag{3.10}$$

Under MAR assumption, the $(j+1)$th iteration of the EM can be derived as

$$R^{(j+1)}(d\boldsymbol{x}_k) = \frac{R^{(j)}(d\boldsymbol{x}_k)}{n}\sum_{i=1}^{n}\Big\{D_i\frac{\sum_{\boldsymbol{x}_l\in\mathscr{X}_i} I(\boldsymbol{x}_k \le \boldsymbol{x}_l)q_l^{(j)}}{\sum_{\boldsymbol{x}_l\in\mathscr{X}_i} R^{(j)}(\boldsymbol{x}_l)q_l^{(j)}} + (1-D_i)\frac{\sum_{\boldsymbol{x}_l\in\mathscr{X}_i} I(\boldsymbol{x}_k \nleq \boldsymbol{x}_l)q_l^{(j)}}{\sum_{\boldsymbol{x}_l\in\mathscr{X}_i} \bar{R}^{(j)}(\boldsymbol{x}_l)q_l^{(j)}}\Big\}$$
$$q_k^{(j+1)} = \frac{q_k^{(j)}}{n}\sum_{i=1}^{n} I(\boldsymbol{x}_k \in \mathscr{X}_i)\Big\{D_i\frac{R^{(j)}(\boldsymbol{x}_k)}{\sum_{\boldsymbol{x}_l\in\mathscr{X}_i} R^{(j)}(\boldsymbol{x}_l)q_l^{(j)}} + (1-D_i)\frac{\bar{R}^{(j)}(\boldsymbol{x}_k)}{\sum_{\boldsymbol{x}_l\in\mathscr{X}_i} \bar{R}^{(j)}(\boldsymbol{x}_l)q_l^{(j)}}\Big\} \tag{3.11}$$

In practice, a non-trivial question arises as to how the supporting points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m$ should be chosen. Our strategy is as follows. Let $S = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m^{'}\}$ be the set of unique values of the covariates $\boldsymbol{X}$ from fully observed subjects. Then for each observation i with partially observed covariates, we "impute" missing component of observation i with all possible values of $\boldsymbol{X}$ that have been observed. For instance, with $p = 2$ and $\boldsymbol{x}_k = (x_{k1}, x_{k2})^T$ $(k = 1, \cdots, m')$, if $\boldsymbol{X}_i = (X_{i1}, .)^T$ with the second component missing, then we can pick $(X_{i1}, x_{k2})^T$, $(k = 1, \cdots, m')$ from $\mathscr{X}_i$ to add to the support. Note that this approach may not be easily extendable to continuous covariates, in which case an unmanageably large number of supporting points may result.

3.3 Simulation studies

Simulation studies were conducted to evaluate and compare the performance of the proposed MC against state-of-art methods, namely, Su and Liu's method (SULIU) (Su and Liu, 1993), logistic regression approach (LOGISTIC) (Richards, Hammitt, and Tsevat, 1996), Kang et al.'s stepwise approach (SW) (Kang, Liu, and Tian, 2016), Liu et al.'s min-max approach (MIN-MAX) (Liu, Liu, and Halabi, 2011), and machine learning methods including random forest (RF) (Breiman et al., 2017) and SVM (Cristianini, Shawe-Taylor, et al., 2000) with linear kernel (SVMl). We used the area under the ROC curve (AUC), a common metric of diagnostic accuracy, as the performance evaluation metric. In addition, we considered two general settings: 1) covariates are fully observed, and 2) covariates are missing at random. In each scenario, we varied sample size and covariates dimensionality

under different true risk score functions.

3.3.1 Fully observed covariates

The simulation procedure is shown as follows: (1) generate the vector of input variables $\boldsymbol{X}$ with p components and each component taking integer values from 1 to k; (2) calculate the true risk score $R_0(\boldsymbol{X})$ in two different ways: (a) $R_0(\boldsymbol{X}) = \mathrm{S}(\boldsymbol{\beta}_0^T \boldsymbol{X})$, where $\boldsymbol{\beta}_0 = [1, \cdots, 1]$ and $S(t) = 1/(1 + e^{-t})$. (b) $R_0(\boldsymbol{X}) = \sum_{\boldsymbol{x}:\boldsymbol{x} \leq \boldsymbol{X}} p_{\boldsymbol{x}}$, where $p_{\boldsymbol{x}}$ is predefined probability point mass and $\boldsymbol{x} \leq \boldsymbol{X}$ operates component-wise. We set a few $p_{\boldsymbol{x}}$ very large to generate non-linearity in $R_0(\boldsymbol{X})$; (3) generate the outcome variable D which follows Bernoulli distribution with probability $R_0(\boldsymbol{X})$; (4) split the $(\boldsymbol{X}, D)$ into training part and testing part, models are trained on the training part while AUC is calculated from testing part. It should be noted that both risk score functions defined in step two satisfy the monotonic assumption given in Equation 3.2. Simulation results are based on 1000 replicates. Table 3.1 shows the results for $R_0(\boldsymbol{X}) = \sum_{\boldsymbol{x}:\boldsymbol{x} \leq \boldsymbol{X}} p_{\boldsymbol{x}}$, while Table 3.2 shows the results for $R_0(\boldsymbol{X}) = \mathrm{S}(\boldsymbol{\beta}_0^T \boldsymbol{X})$.

From Table 3.1, it is clearly shown that MC is dominant in obtaining the largest AUC, whereas LOGISTIC shows comparable performance to MC only when $k = 2$. Note that $R_0(\boldsymbol{X}) = \sum_{\boldsymbol{x}:\boldsymbol{x} \leq \boldsymbol{X}} p_{\boldsymbol{x}}$ is proportional to linear combination of $\boldsymbol{X}$ when $k = 2$, but it becomes nonlinear when $k > 2$ and that non-linearity gets stronger as k increases. MC always has the best performance in terms of AUC due to the fact that the monotonicity assumption on the risk score is satisfied, thereby $\hat{R}$ estimated by MC converges to the true risk score function R_0. On the other hand, $\hat{R}$ estimated by LOGISTIC converges to R_0 only when $k = 2$ because of its robustness to link violation under linear model assumption introduced in Li and Duan (1989). Figure 3.2 plots the predicted risk score from MC and LOGISTIC when $p = 2, k = 3$, it confirms that as sample size increases, the predicted risk score from MC converges to the true risk score while LOGISTIC does not.

When the true risk score function $R_0(\boldsymbol{X}) = \mathrm{S}(\boldsymbol{\beta}_0^T \boldsymbol{X})$, it is intuitive to observe from Table 3.2 that LOGISTIC outperforms other methods most of the time, due to the fact

Table 3.1: Mean area under receiver operating characteristic curve AUC (SE) under non-sigmoidal risk score function.

$(p\ ,\ k)$	n	LOGISTIC	MC	SULIU	SW	MIN-MAX	RF	SVMl
	100	0.829 (0.071)	0.830* (0.071)	0.830 (0.071)	0.830 (0.072)	0.816 (0.077)	0.822 (0.073)	0.779 (0.121)
(2, 2)	600	0.837 (0.031)	0.838* (0.031)	0.838 (0.031)	0.838 (0.031)	0.828 (0.031)	0.838 (0.031)	0.826 (0.033)
	1000	0.837 (0.025)	0.837* (0.025)	0.837 (0.025)	0.837 (0.025)	0.827 (0.025)	0.837 (0.025)	0.827 (0.025)
	100	0.946 (0.035)	0.976* (0.026)	0.946 (0.036)	0.946 (0.036)	0.969 (0.033)	0.972 (0.033)	0.940 (0.043)
(2, 3)	600	0.945 (0.013)	0.978* (0.008)	0.943 (0.014)	0.944 (0.014)	0.975 (0.010)	0.977 (0.008)	0.938 (0.018)
	1000	0.945 (0.011)	0.978* (0.007)	0.944 (0.011)	0.944 (0.011)	0.974 (0.008)	0.977 (0.007)	0.939 (0.015)
	100	0.893 (0.057)	0.935* (0.042)	0.890 (0.056)	0.892 (0.057)	0.902 (0.054)	0.935 (0.042)	0.890 (0.061)
(2, 4)	600	0.895 (0.022)	0.936* (0.015)	0.895 (0.023)	0.894 (0.024)	0.907 (0.020)	0.936 (0.016)	0.889 (0.029)
	1000	0.895 (0.020)	0.936* (0.012)	0.895 (0.020)	0.895 (0.019)	0.907 (0.015)	0.936 (0.013)	0.892 (0.022)
	100	0.981 (0.026)	0.982* (0.020)	0.979 (0.024)	0.955 (0.031)	0.839 (0.059)	0.974 (0.029)	0.967 (0.032)
(3, 2)	600	0.983 (0.008)	0.984* (0.007)	0.983 (0.008)	0.958 (0.014)	0.836 (0.026)	0.980 (0.009)	0.969 (0.012)
	1000	0.984 (0.005)	0.985* (0.005)	0.983 (0.005)	0.961 (0.014)	0.836 (0.020)	0.983 (0.006)	0.970 (0.008)
	100	0.941 (0.040)	0.999* (0.001)	0.937 (0.042)	0.934 (0.041)	0.996 (0.013)	0.999 (0.001)	0.938 (0.039)
(3, 3)	600	0.937 (0.015)	0.999* (0.0008)	0.938 (0.015)	0.940 (0.014)	0.999 (0.001)	0.999 (0.0008)	0.936 (0.020)
	1000	0.936 (0.011)	0.999* (0.0006)	0.936 (0.012)	0.939 (0.011)	0.999 (0.001)	0.999 (0.0005)	0.935 (0.019)
	100	0.942 (0.061)	0.971* (0.054)	0.938 (0.065)	0.936 (0.064)	0.965 (0.054)	0.963 (0.059)	0.943 (0.060)
(3, 4)	600	0.949 (0.019)	0.974* (0.017)	0.950 (0.019)	0.949 (0.019)	0.972 (0.017)	0.972 (0.018)	0.950 (0.020)
	1000	0.948 (0.012)	0.973* (0.012)	0.948 (0.012)	0.949 (0.012)	0.971 (0.012)	0.972 (0.012)	0.948 (0.014)
	100	0.871 (0.085)	0.882* (0.079)	0.865 (0.085)	0.855 (0.092)	0.804 (0.083)	0.859 (0.089)	0.850 (0.148)
(4, 3)	600	0.885 (0.042)	0.899* (0.038)	0.883 (0.041)	0.876 (0.044)	0.809 (0.046)	0.879 (0.044)	0.878 (0.055)
	1000	0.889 (0.028)	0.903* (0.028)	0.888 (0.028)	0.880 (0.030)	0.814 (0.031)	0.879 (0.031)	0.887 (0.029)

that the estimated coefficients from LOGISTIC converge to the true parameters and thus LOGISTIC approximates the optimal classifier under large samples. It is worth noting that MC is slightly inferior to LOGISTIC when p and k are small but the difference gets bigger when p and k increase. The reason is that the estimated risk score from MC is also consistent since the monotonic assumption still holds. The convergence rate of MC, however, gets slower as the covariates dimensionality increases. In other words, both MC and LOGISTIC approximate the optimal classifier under large samples but LOGISTIC converges at a faster rate when p and k are large. Figure 3.3 plots the predicted risk score from MC and LOGISTIC when $p = 2, k = 3$, it confirms that as sample size increases, the predicted risk scores from LOGISTIC and MC converge to the true risk score while LOGISTIC converges at a faster rate.

Table 3.2: Mean area under receiver operating characteristic curve AUC (SE) under sigmoid risk score function.

$(p\ ,\ k)$	n	LOGISTIC	MC	SULIU	SW	MIN-MAX	RF	SVMl
	100	0.688 (0.097)	0.696* (0.091)	0.687 (0.097)	0.687 (0.097)	0.682 (0.094)	0.673 (0.098)	0.641 (0.125)
(2, 2)	600	0.698* (0.037)	0.697 (0.036)	0.698 (0.037)	0.698 (0.037)	0.697 (0.035)	0.698 (0.037)	0.661 (0.044)
	1000	0.697* (0.029)	0.697 (0.028)	0.697 (0.029)	0.697 (0.029)	0.697 (0.027)	0.697 (0.029)	0.662 (0.038)
	100	0.788 (0.080)	0.788* (0.081)	0.787 (0.080)	0.787 (0.080)	0.788 (0.082)	0.780 (0.084)	0.773 (0.085)
(2, 3)	600	0.793* (0.031)	0.792 (0.031)	0.793 (0.031)	0.793 (0.031)	0.793 (0.031)	0.792 (0.031)	0.789 (0.034)
	1000	0.793* (0.023)	0.792 (0.023)	0.793 (0.023)	0.793 (0.023)	0.793 (0.024)	0.792 (0.024)	0.791 (0.024)
	100	0.852* (0.065)	0.849 (0.068)	0.851 (0.065)	0.851 (0.063)	0.850 (0.065)	0.848 (0.070)	0.848 (0.066)
(2, 4)	600	0.858* (0.026)	0.857 (0.027)	0.858 (0.025)	0.858 (0.026)	0.858 (0.025)	0.858 (0.029)	0.858 (0.026)
	1000	0.857* (0.022)	0.856 (0.023)	0.857 (0.022)	0.857 (0.022)	0.857 (0.022)	0.855 (0.025)	0.857 (0.022)
	100	0.718 (0.091)	0.723* (0.088)	0.717 (0.091)	0.712 (0.090)	0.675 (0.079)	0.716 (0.092)	0.670 (0.123)
(3, 2)	600	0.733* (0.035)	0.733 (0.035)	0.733 (0.034)	0.728 (0.035)	0.682 (0.030)	0.733 (0.035)	0.712 (0.041)
	1000	0.732* (0.027)	0.732 (0.027)	0.732 (0.027)	0.728 (0.027)	0.681 (0.023)	0.732 (0.029)	0.713 (0.033)
	100	0.825* (0.076)	0.813 (0.079)	0.823 (0.075)	0.819 (0.073)	0.800 (0.077)	0.810 (0.079)	0.814 (0.075)
(3, 3)	600	0.832* (0.031)	0.827 (0.033)	0.832 (0.031)	0.831 (0.031)	0.803 (0.031)	0.827 (0.035)	0.831 (0.033)
	1000	0.830* (0.023)	0.828 (0.025)	0.830 (0.023)	0.829 (0.023)	0.801 (0.023)	0.827 (0.027)	0.830 (0.023)
	100	0.748* (0.085)	0.747 (0.087)	0.744 (0.090)	0.734 (0.094)	0.617 (0.071)	0.737 (0.096)	0.707 (0.122)
(4, 2)	600	0.757* (0.035)	0.756 (0.037)	0.756 (0.035)	0.747 (0.036)	0.625 (0.025)	0.749 (0.040)	0.754 (0.034)
	1000	0.759* (0.031)	0.759 (0.034)	0.758 (0.031)	0.755 (0.031)	0.630 (0.019)	0.752 (0.037)	0.756 (0.031)
	100	0.835* (0.083)	0.792 (0.092)	0.830 (0.085)	0.823 (0.093)	0.779 (0.085)	0.800 (0.087)	0.829 (0.088)
(4, 3)	600	0.846* (0.028)	0.831 (0.030)	0.846 (0.028)	0.842 (0.028)	0.766 (0.032)	0.820 (0.035)	0.844 (0.028)
	1000	0.849 (0.021)	0.839 (0.025)	0.849 (0.021)	0.846 (0.023)	0.770 (0.021)	0.820 (0.027)	0.850* (0.022)

3.3.2 Missing at random covariates

We further investigated the performance of MC when covariates are missing at random. In particular, we allowed the probability of the occurrence of a missing value in a covariate to depend on the value of the outcome variable. This situation is typical for case-control studies, where cases are often better documented or more willing to answer questions than controls. The simulation strategy is same as in Section 3.3.1 except the extra coarsening step of covariates in the training data. The coarsening step is shown as follows: the vector of input variables $\boldsymbol{X}$ can be expanded as $(\boldsymbol{X}_{obs}, \boldsymbol{X}_{mis})$ where $\boldsymbol{X}_{obs}$ is always observed while $\boldsymbol{X}_{mis}$ is not. For simplicity, we treated the last covariate as $\boldsymbol{X}_{mis}$ and the rest as $\boldsymbol{X}_{obs}$. In addition, we denote Γ as the missing data indicator, i.e., $\Gamma = 1$ if $(\boldsymbol{X}_{obs}, \boldsymbol{X}_{mis})$ is observed and $\Gamma = 0$ if $\boldsymbol{X}_{obs}$ is observed. We set $P(\Gamma = 0|\boldsymbol{X}_{obs}, D = 0) = \mathrm{S}(\boldsymbol{\beta}_1^T \boldsymbol{X}_{obs})$ and $P(\Gamma = 0|\boldsymbol{X}_{obs}, D = 1) = 0$, where $\boldsymbol{\beta}_1$ is deliberately calibrated to obtain the desired missing rate. As a result, the missing probability of $\boldsymbol{X}_{mis}$ only depends on the observed

data $(\boldsymbol{X}_{obs}, D)$, the covariates are, therefore, missing at random.

The proposed MC can handle MAR covariates through EM-type algorithm while other combination methods cannot. A natural strategy is to use complete cases of data for those methods. In addition, we applied complete-case analysis to the proposed MC, and as a result, there were two versions of MC, with the first version (MC-missing) using the coarsened training data and second version (MC-complete) using complete cases of training data. The sample size is 1000 and results are based on 1000 replicates. Table 3.3 shows the results for $R_0(\boldsymbol{X}) = \sum_{\boldsymbol{x}:\boldsymbol{x}\leq\boldsymbol{X}} p_{\boldsymbol{x}}$ and Table 3.4 shows the results for $R_0(\boldsymbol{X}) = \mathrm{S}(\boldsymbol{\beta}_0^T\boldsymbol{X})$.

Table 3.3: Mean area under receiver operating characteristic curve AUC (SE) under non-sigmoidal risk score function with MAR covariates.

$(p\ ,\ k)$	Missing rate	MC-missing	MC-complete	LOGISTIC	SULIU	SW	MIN-MAX	RF	SVMl
(2, 2)	0.16	0.837* (0.026)	0.833 (0.027)	0.831 (0.027)	0.831 (0.027)	0.832 (0.027)	0.826 (0.025)	0.832 (0.027)	0.684 (0.164)
	0.24	0.837* (0.026)	0.828 (0.029)	0.826 (0.029)	0.825 (0.029)	0.826 (0.029)	0.827 (0.025)	0.826 (0.029)	0.505 (0.049)
(2, 3)	0.20	0.804* (0.026)	0.802 (0.027)	0.799 (0.029)	0.798 (0.029)	0.799 (0.032)	0.804 (0.029)	0.800 (0.032)	0.784 (0.046)
	0.42	0.803* (0.026)	0.792 (0.027)	0.786 (0.031)	0.789 (0.030)	0.785 (0.028)	0.803 (0.029)	0.771 (0.050)	0.503 (0.047)
(2, 4)	0.27	0.827* (0.024)	0.821 (0.025)	0.811 (0.028)	0.811 (0.027)	0.811 (0.028)	0.824 (0.024)	0.816 (0.029)	0.792 (0.062)
	0.45	0.827* (0.024)	0.808 (0.028)	0.760 (0.064)	0.775 (0.055)	0.775 (0.053)	0.805 (0.036)	0.779 (0.049)	0.501 (0.034)
(3, 2)	0.33	0.889* (0.019)	0.883 (0.021)	0.877 (0.023)	0.875 (0.023)	0.869 (0.027)	0.816 (0.019)	0.877 (0.024)	0.875 (0.025)
	0.46	0.889* (0.019)	0.874 (0.023)	0.874 (0.023)	0.870 (0.024)	0.862 (0.026)	0.816 (0.019)	0.875 (0.023)	0.881 (0.023)
(3, 3)	0.27	0.802* (0.028)	0.792 (0.029)	0.786 (0.030)	0.783 (0.030)	0.780 (0.034)	0.779 (0.028)	0.788 (0.030)	0.773 (0.040)
	0.45	0.801* (0.028)	0.783 (0.030)	0.759 (0.036)	0.758 (0.035)	0.760 (0.033)	0.779 (0.031)	0.781 (0.031)	0.749 (0.044)
	0.55	0.801* (0.028)	0.781 (0.029)	0.742 (0.040)	0.746 (0.036)	0.742 (0.045)	0.777 (0.031)	0.769 (0.037)	0.712 (0.045)
(3, 4)	0.29	0.902* (0.030)	0.901 (0.031)	0.888 (0.030)	0.881 (0.032)	0.888 (0.029)	0.866 (0.031)	0.881 (0.033)	0.886 (0.031)
	0.59	0.901* (0.031)	0.892 (0.032)	0.862 (0.033)	0.848 (0.036)	0.851 (0.043)	0.867 (0.030)	0.867 (0.036)	0.865 (0.034)
	0.79	0.898* (0.030)	0.879 (0.034)	0.815 (0.046)	0.812 (0.043)	0.816 (0.048)	0.864 (0.031)	0.869 (0.036)	0.807 (0.057)
(3, 5)	0.26	0.920* (0.015)	0.916 (0.016)	0.887 (0.021)	0.882 (0.021)	0.883 (0.022)	0.885 (0.020)	0.893 (0.020)	0.883 (0.022)
	0.50	0.919* (0.015)	0.901 (0.019)	0.857 (0.025)	0.843 (0.026)	0.851 (0.026)	0.886 (0.020)	0.881 (0.024)	0.855 (0.027)
	0.70	0.917* (0.016)	0.886 (0.024)	0.799 (0.042)	0.799 (0.038)	0.797 (0.048)	0.870 (0.034)	0.868 (0.045)	0.794 (0.046)
(4, 2)	0.26	0.785* (0.033)	0.780 (0.033)	0.779 (0.033)	0.778 (0.033)	0.764 (0.035)	0.665 (0.018)	0.773 (0.037)	0.735 (0.055)
	0.39	0.785* (0.034)	0.770 (0.036)	0.769 (0.035)	0.768 (0.035)	0.755 (0.037)	0.665 (0.020)	0.764 (0.039)	0.711 (0.064)
	0.50	0.784* (0.034)	0.763 (0.035)	0.748 (0.042)	0.752 (0.039)	0.744 (0.040)	0.661 (0.025)	0.750 (0.040)	0.690 (0.088)
(4, 3)	0.27	0.816* (0.033)	0.808 (0.032)	0.808 (0.030)	0.806 (0.031)	0.801 (0.033)	0.772 (0.030)	0.785 (0.034)	0.804 (0.031)
	0.51	0.815* (0.033)	0.796 (0.036)	0.776 (0.042)	0.772 (0.042)	0.767 (0.044)	0.771 (0.030)	0.770 (0.036)	0.764 (0.047)
	0.73	0.815* (0.033)	0.796 (0.034)	0.703 (0.051)	0.709 (0.049)	0.698 (0.060)	0.744 (0.051)	0.734 (0.045)	0.681 (0.073)

Table 3.3 and Table 3.4 show that, regardless of which R_0 is used, MC-missing consistently outperforms other methods when there are MAR covariates. Additionally, the performance of MC-missing stays almost unchanged as missing rate increases, whereas the performance of other methods deteriorates quickly. The predicted risk scores from LOGISTIC, MC-missing and MC-complete for $p = 2, k = 2$ are shown in Figure 3.4-3.5, respectively. This confirms that the estimation of risk score by MC-missing is still consis-

tent, while the estimation by other methods based on complete cases is associated with serious bias when covarites are MAR. In other words, only MC-missing will approximate the optimal classifier under large samples.

Table 3.4: Mean area under receiver operating characteristic curve AUC (SE) under sigmoid risk score function with MAR covariates.

$(p\ ,\ k)$	Missing rate	MC-missing	MC-complete	LOGISTIC	SULIU	SW	MIN-MAX	RF	SVMl
(2, 2)	0.20	0.700 (0.027)	0.700 (0.028)	0.702 (0.038)	0.702 (0.038)	0.702 (0.038)	0.702* (0.030)	0.701 (0.040)	0.642 (0.089)
	0.32	0.700 (0.028)	0.701 (0.028)	0.700 (0.042)	0.700 (0.042)	0.701 (0.042)	0.701* (0.032)	0.688 (0.048)	0.504 (0.041)
(2, 3)	0.19	0.794 (0.024)	0.793 (0.026)	0.793 (0.027)	0.793 (0.027)	0.793 (0.027)	0.796* (0.031)	0.794 (0.033)	0.776 (0.071)
	0.38	0.793 (0.024)	0.787 (0.028)	0.781 (0.035)	0.782 (0.031)	0.781 (0.036)	0.795* (0.033)	0.778 (0.039)	0.505 (0.033)
(2, 4)	0.26	0.856* (0.023)	0.849 (0.024)	0.853 (0.024)	0.851 (0.025)	0.852 (0.023)	0.855 (0.022)	0.850 (0.026)	0.842 (0.033)
	0.42	0.856* (0.023)	0.840 (0.024)	0.824 (0.042)	0.824 (0.043)	0.826 (0.044)	0.852 (0.029)	0.825 (0.037)	0.505 (0.035)
(3, 2)	0.25	0.735* (0.027)	0.731 (0.027)	0.731 (0.027)	0.730 (0.027)	0.724 (0.028)	0.684 (0.023)	0.733 (0.037)	0.710 (0.060)
	0.35	0.734* (0.027)	0.727 (0.028)	0.724 (0.033)	0.718 (0.035)	0.709 (0.032)	0.684 (0.023)	0.721 (0.033)	0.588 (0.093)
(3, 3)	0.20	0.827* (0.025)	0.813 (0.027)	0.823 (0.026)	0.819 (0.027)	0.818 (0.027)	0.800 (0.023)	0.817 (0.028)	0.820 (0.041)
	0.35	0.827* (0.025)	0.797 (0.026)	0.806 (0.030)	0.801 (0.028)	0.801 (0.030)	0.800 (0.024)	0.803 (0.028)	0.796 (0.037)
	0.42	0.826* (0.026)	0.796 (0.024)	0.793 (0.033)	0.786 (0.035)	0.786 (0.032)	0.800 (0.024)	0.789 (0.032)	0.661 (0.144)
(3, 4)	0.16	0.875 (0.018)	0.862 (0.021)	0.880* (0.018)	0.876 (0.018)	0.877 (0.018)	0.860 (0.020)	0.865 (0.021)	0.879 (0.018)
	0.33	0.874* (0.018)	0.839 (0.024)	0.860 (0.021)	0.851 (0.021)	0.856 (0.023)	0.860 (0.020)	0.852 (0.023)	0.862 (0.024)
	0.45	0.869* (0.018)	0.837 (0.024)	0.839 (0.027)	0.825 (0.028)	0.830 (0.029)	0.857 (0.021)	0.819 (0.033)	0.767 (0.146)
(3, 5)	0.16	0.904 (0.015)	0.890 (0.018)	0.914* (0.016)	0.910 (0.017)	0.912 (0.017)	0.899 (0.017)	0.898 (0.018)	0.914 (0.015)
	0.33	0.903* (0.015)	0.872 (0.019)	0.898 (0.018)	0.887 (0.019)	0.893 (0.020)	0.900 (0.017)	0.881 (0.019)	0.898 (0.019)
	0.47	0.895* (0.017)	0.870 (0.021)	0.868 (0.028)	0.856 (0.028)	0.861 (0.031)	0.888 (0.035)	0.833 (0.043)	0.842 (0.11)
(4, 2)	0.21	0.758* (0.030)	0.745 (0.033)	0.749 (0.030)	0.747 (0.030)	0.733 (0.032)	0.630 (0.018)	0.741 (0.033)	0.738 (0.033)
	0.31	0.757* (0.030)	0.733 (0.032)	0.735 (0.033)	0.732 (0.032)	0.721 (0.033)	0.630 (0.018)	0.730 (0.035)	0.654 (0.100)
	0.40	0.757* (0.030)	0.728 (0.031)	0.714 (0.037)	0.711 (0.041)	0.702 (0.038)	0.629 (0.020)	0.711 (0.037)	0.522 (0.074)
(4, 3)	0.18	0.835 (0.024)	0.817 (0.027)	0.841* (0.024)	0.838 (0.025)	0.834 (0.026)	0.769 (0.021)	0.807 (0.028)	0.838 (0.025)
	0.33	0.833* (0.025)	0.792 (0.027)	0.815 (0.028)	0.807 (0.029)	0.799 (0.033)	0.768 (0.022)	0.790 (0.029)	0.816 (0.033)
	0.46	0.825* (0.026)	0.791 (0.028)	0.768 (0.045)	0.760 (0.049)	0.751 (0.053)	0.748 (0.066)	0.750 (0.044)	0.549 (0.115)

3.4 Analysis of NASH data: An example

In this Section, the mentioned approaches in simulation studies were applied to a real data set from a recent nonalcoholic fatty liver disease (NALFD) study. 186 patients (74 Male, 112 Female, mean age 49 yrs) with nonalcoholic fatty liver disease (NALFD) were included in the study. We aim to combine several non-invasive measurements to improve diagnostic capacity for advanced fibrosis. These variables include a composite clinical score Fibrosis-4 (FIB-4), ordinal assessment of computed tomography (CT) images by two readers (from 1 to 5), and quantitative imaging biomarkers (QIBs) such as liver segmental volume ratio (LSVR) and liver surface nodularity (LSN). The four markers to be combined are "R1-NASH", "R2-NASH", "FIB4" and "LSVR". To reduce the cardinality

of the covariates support and thereby decreasing the computational burden of MC, we discretized continuous markers "FIB4" and "LSVR" by rounding "FIB4" to integers and "LSVR" to one digit. See Table 3.5 for patient characteristics in the NALFD study.

Table 3.5: Quantitative variables are summarized by median (inter-quartile range) and categorical variables by N (%).

		Non-NASH (N = 99)	NASH (N = 87)	Overall (N = 186)
R1-NASH	1	17 (17.1%)	8 (9.1%)	25 (13.4%)
	2	29 (29.2%)	26 (29.8%)	55 (29.5%)
	3	25 (25.2%)	24 (27.5%)	49 (26.3%)
	4	19 (19.1%)	17 (19.5%)	36 (19.3%)
	5	8 (8.0%)	12 (13.7%)	20 (10.7%)
R2-NASH	1	30 (30.3%)	19 (21.8%)	49 (26.3%)
	2	33 (33.3%)	37 (42.5%)	70 (37.6%)
	3	20 (20.2%)	10 (11.4%)	30 (16.1%)
	4	12 (12.1%)	12 (13.7%)	24 (12.9%)
	5	3 (3.0%)	9 (10.3%)	12 (6.4%)
FIB4		1.17 (0.67, 2.41)	2.17 (1.30, 3.67)	1.58 (0.83, 3.22)
LSVR		0.28 (0.23, 0.35)	0.28 (0.21, 0.38)	0.28 (0.22, 0.36)

In this real data, we have to deal with subjects with missing covariates. To be more specific, 9 out of 186 subjects had missing "FIB-4" values, 22 subjects had missing "LSVR" values and 1 subject had missing "R1-NASH" value. Furthermore, it is reasonable to assume that the input variables are monotonically associated with the risk score. For instance, a higher ordinal assessment by radiologists should correspond to a higher risk of disease. We used 10-fold cross validation to calculate the mean AUC for each method. That is, in each iteration, the real data was split into training part and testing part, with MC-missing fitting on the full training data which may contain missing values, and other methods using complete cases of the training data. The AUC was calculated from complete cases of testing data. In addition, the data was analyzed with/without log-transformation on covariates to see the effect on different approaches. Results are summarized in Table 3.6 and their corresponding ROC curves without log-transformation on covariates are

presented in Figure 3.1.

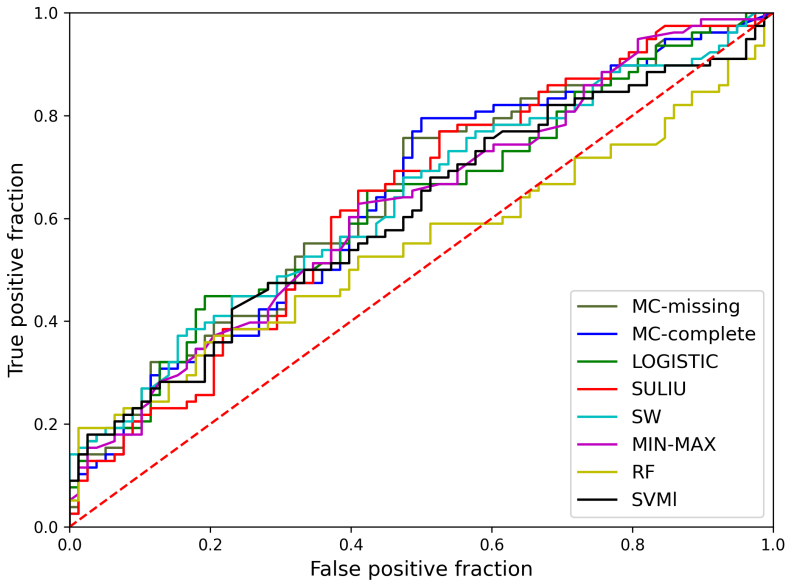

Figure 3.1: ROC curves for NASH data based on 10-fold cross validation without log transformation on covariates. Methods used are "MC-missing", "MC-complete", "SULIU", "LOGISTIC", "SW", "MIN-MAX", "RF", "SVMl".

From Table 3.6 we can conclude that MC-missing produces the largest AUC, which may be due to the fact that MC-missing does not neglect the information from subjects with missing values in covariates, and the missing at random assumption approximately holds in this study. MC-complete also shows comparable performance, probably because the four markers to be combined are monotonically associated with the risk score. Another observation is that MC-missing and MC-complete are invariant to monotone transformation of the covariates in that they only utilize the ordering information rather than the actual value of covariates, whereas LOGISTIC and SULIU benefit from the transformations.

Table 3.6: Mean area under receiver operating characteristic curve AUC for NASH data with/without log-transformation on covariates.

	MC-missing	MC-complete	LOGISTIC	SULIU	SW	MIN-MAX	RF	SVMl
Log transformation	0.670*	0.648	0.640	0.640	0.638	0.628	0.578	0.661
No transformation	0.670*	0.648	0.632	0.632	0.640	0.630	0.580	0.641

3.5 Conclusions and Perspective

The main contribution of the current chapter is that we developed a novel likelihood-based approach to nonparametric monotone classification that accounts for missing covariates in a natural and principled way. The first advantage of our method is that it does not require specifying the form of the risk score function. Instead, the risk score is estimated nonparametrically by an EM-type algorithm. Therefore, it is more robust than parametric models such as logistic regression. The second advantage of the proposed method is the ability to handle missing covariates appropriately. Unlike complete-case analysis, MC handles the missing covariates via EM-type algorithm and thereby utilizing information from subjects with missing values in covariates. Another merit of MC is the incorporation of monotonicity relationship between the input variables and the disease outcome. Exploitation of such established domain knowledge usually leads to gains in diagnostic accuracy. Lastly, in order to provide researchers an easy access to our approach, we developed a custom implementation of Monotone Classifier available in both Python and R, which can be found at `https://github.com/chengning-zhang/Monotone-classifier`. Through simulation studies and a real data example, we demonstrated that MC outperforms other statistical methods for combining multiple inputs under monotonic assumption, especially when the inputs contain missing data.

There is work to be done in the future before the approach can be routinely applied in practice. The first thing is to develop statistical procedure to check validity of monotonic assumption. The superiority of our approach rests on the assumption that the inputs are monotonically associated with the risk score function. When the assumption is violated, the trained classifiers may not converge to the optimal classifier under large samples. In fact, it may even underperform other standard methods. Therefore, we need to develop statistical procedure to check monotonic assumption both theoretically and empirically; The second thing is to extend the proposed MC to handling continuous missing covariates, which are more common than discrete missing covariates in studies of diagnostic medicine. For continuous missing covariates, the cardinality of the support could far exceed the

sample size and thus increase the computational burden. To alleviate the burden, one potential future direction is to consider various regularized NPMLE approaches under the monotonic constraint such as kernel smoothing (Eggermont and LaRiccia, 2000; Van Der Vaart and Van Der Laan, 2003) and penalization (Eggermont, LaRiccia, and LaRiccia, 2001; Groeneboom and Jongbloed, 2013).

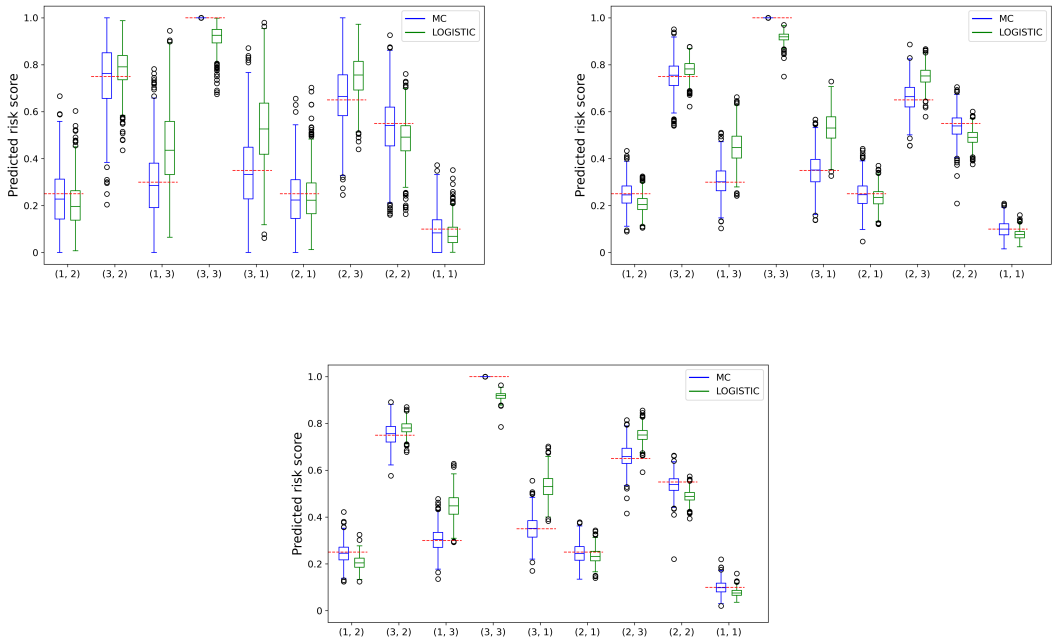

Figure 3.2: Box plots of predicted risk scores from MC and LOGISTIC under non-sigmoidal risk score function with $p = 2, k = 3$. The red dotted lines are the true risk scores. The sample sizes from top left to bottom are 100, 600, 1000 respectively.

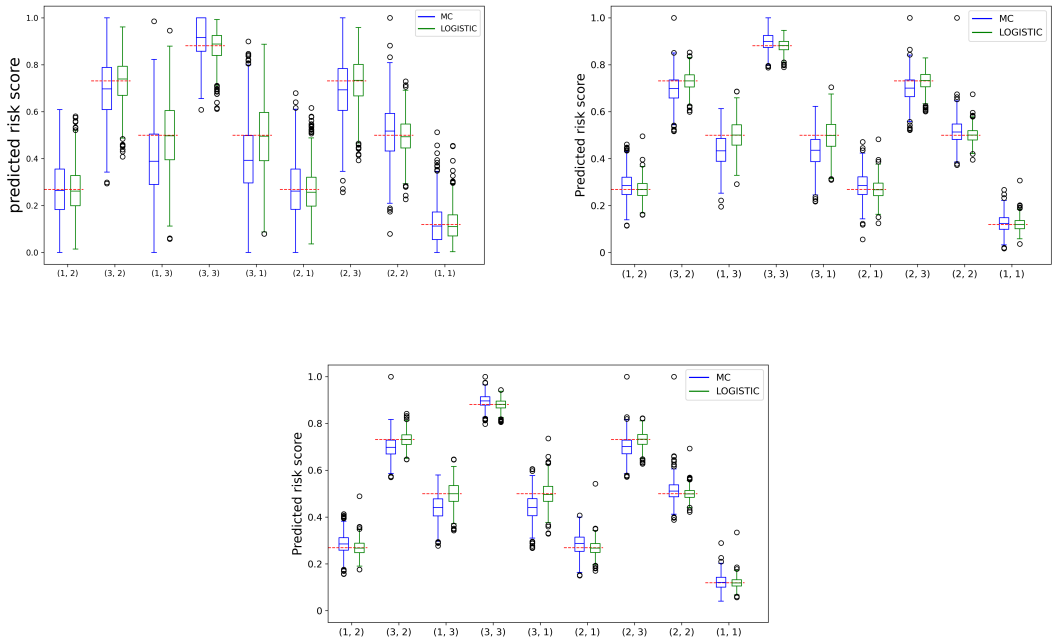

Figure 3.3: Box plots of predicted risk scores from MC and LOGISTIC under sigmoid risk score function with $p = 2, k = 3$. The red dotted lines are the true risk scores. The sample sizes from top left to bottom are 100, 600, 1000 respectively.

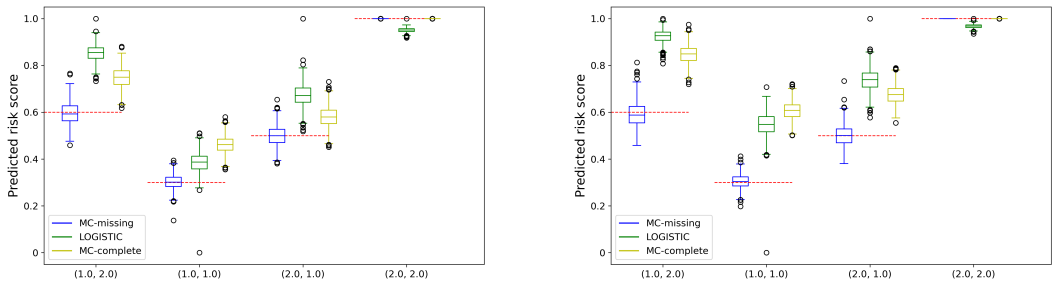

Figure 3.4: Box plots of predicted risk scores from MC-missing, MC-complete and LOGISTIC under non-sigmoidal risk score function with $p = 2, k = 2$. The red dotted lines are the true risk scores. The missing rates from left to right are 0.16, 0.24 respectively.

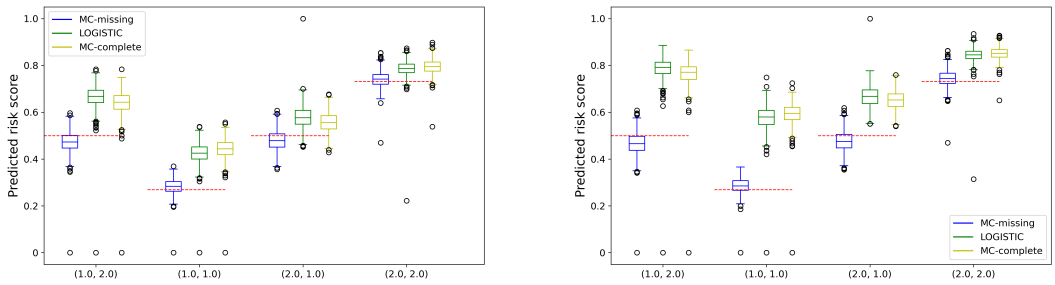

Figure 3.5: Box plots of predicted risk scores from MC-missing, MC-complete and LOGISTIC under sigmoid risk score function with $p = 2, k = 2$. The red dotted lines are the true risk scores. The missing rates from left to right are 0.20, 0.32 respectively.

Chapter 4

Monotone classification with a randomly censored covariate

4.1 Introduction

Censored data are ubiquitous in social, behavioral, and medical studies. Numerous studies have investigated the problem of censored data. Nevertheless, most studies focus on censored outcomes rather than censored covariates.

The situation of censored covariates arises, for instance, when modeling the value of a diagnostic marker as a function of the time lag between the measurement and the occurrence of disease. Models of this type are utilized in a number of studies (Cai et al., 2006; Tsimikas, Bantis, and Georgiou, 2012), where the marker sensitivity is considered as a function of survival time. The time-to-event covariates may be censored due to loss to follow-up, study termination or detection limits. Another scenario where censored covariates are encountered frequently, is the study of associations between parental risk factors and the onset of disease in their offspring. Often, the commonly used parental risk factor, age-of-onset of disease in parents, is right-censored, meaning that the study either terminates prior to the event being observed or a patient is lost to follow-up prior to the event. For example, Allport et al. (2016) and Atem, Matsouaka, and Zimmern

(2019) studied associations between parental age of onset of cardiovascular disease and offspring age of onset of cardiovascular disease, Atem et al. (2017) and Maye et al. (2016) investigated the association between maternal age of onset of dementia and β-amyloid deposition (measured by in vivo PET imaging) in cognitively normal older offspring.

One often-used approach for censored covariates is the complete-case analysis, which consists in deleting all censored observations in the analysis. However, deleting censored observations can reduce the sample size dramatically and thereby make this approach highly inefficient. Another commonly used approach is the substitution methods. In this approach, the censored covariate is used as if it were the true value. Though this approach can use all the observations in the analysis, it misestimates the value of censored covariate, and the extent of the misestimation depends on the extent and severity of censoring.

Along with the limited studies on the issue of covariates subject to type I, type II, or limit-of-detection censoring, an even smaller number of studies investigated the issue of randomly censored covariates. For example, under the linear model assumption, Atem et al. (2017) proposed a method of replacing the censored covariate with conditional mean values predicted either by a Kaplan–Meier method or a Cox proportional hazards model; Under the Cox proportional hazards model, Atem, Matsouaka, and Zimmern (2019) considered the problem of censored covariates in the context of a censored outcome. They investigated the effects of covariates on a time-to-event outcome, with the censored covariate imputed based on either Kaplan–Meier estimates or a Cox proportional hazards model; Tsimikas, Bantis, and Georgiou (2012) studied the problem of censored covariates in the generalized linear model. They proposed a quasi-score based method to estimate the parameters of interest. When no censoring of the covariate occurs, their proposed estimating function reduces to the well known estimating function used in generalized linear models.

Dealing with censored covariates under the known monotonicity relationship between the input variables and the disease outcome is even more complicated. All existing regression methods with censored covariates rely on classical linear model or other generalized

linear models, making it difficult to utilize the monotone relationship between the input variables and outcome variable. Yet, even with such monotone constraints embedded, these methods may perform poorly if the model is wrongly specified. For robustness, a preferred approach should therefore build the monotone classification with censored covariates without specifying the form of the risk score function. Note that the established monotone classification approach for missing covariates in Chapter 3 is not readily applicable here because it does not naturally utilize the partial information contained in censored observations. Moreover, the fully nonparametric approach may not be feasible when covariates are continuous.

In order to allow the use of information from subjects with a randomly censored covariate and also leverage the monotonicity between the input variables and the disease outcome, we develop a two-step likelihood-based approach to monotone classification that accounts for the randomly censored covariate in a natural and principled way. The proposed approach consists of two steps. In step one, to reduce the number of supporting points for continuous covariates, we use a Cox proportional hazards model for the distribution of the censored covariate given other covariates in the model. This conditional distribution is used for calculating the observed likelihood of data. In step two, the EM algorithm devised in Chapter 3 is used, based on observed data likelihood from step one, to compute the NPMLE by treating the monotonicity-constrained risk score function as a cumulative distribution for a latent random vector. The randomly censored covariate is handled by adding another E-step to compute the conditional expectation of the censored input given the observed data.

This rest of the Chapter is organized as follows. In Section 4.2, we introduce the two-step approach for estimation of the risk score function and present its EM-type algorithm. In Section 4.3 simulation studies are conducted to evaluate the proposed approach against complete-case analysis and substitution methods. A real data example from a primary biliary cirrhosis (PBC) study conducted at Mayo Clinic is shown in Section 4.4. Concluding remarks are given in Section 4.5.

4.2 Theory and methods

For ease of exposition, we focus on the case of binary classification. The proposed approach can be easily extended to multi-class cases via strategy introduced in Frank and Hall (2001). Let D denote the binary status of disease, i.e., $D = 1$ if diseased and $D = 0$ if otherwise. We use T to denote the potentially censored covariate, and $\boldsymbol{Z}$ to denote the vector of other covariates. Again, for the sake of simplicity, we consider the situation where T is potentially right censored, that is, T is observed only if $T < C$, where C is the censoring variable, which we assume is independent of T and D given $\boldsymbol{Z}$. While we consider right censoring, all of our developments can be adapted to the case of a left or interval censored covariate. The observed data are $\{D_i, X_i, \boldsymbol{Z}_i, \delta_i\},\ i = 1, \cdots, n$, where n is the sample size, $X_i = \min(T_i, C_i)$ and $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ if $T_i > C_i$. Our primary goal is to make valid estimation of the risk score function, i.e., $R(T, \boldsymbol{Z}) = P(D = 1|T, \boldsymbol{Z})$.

In order to exploit the monotonicity relationship between the covariates and the disease outcome, we restrict our attention to monotonic classifiers satisfying the following condition:

$$R(T_1, \boldsymbol{Z}_1) \geq R(T_2, \boldsymbol{Z}_2) \qquad \forall (T_1, \boldsymbol{Z}_1) \geq (T_2, \boldsymbol{Z}_2) \tag{4.1}$$

with the latter inequality operates component-wise. That is, a monotonic classifier is constrained by the sole requirement that, an increase in any of its components must not decrease the probability of the underlying disease.

The joint likelihood of observed data can be factorized as

$$\begin{aligned} P(D, \delta, X = x, \boldsymbol{Z}) &= P(\boldsymbol{Z})P(D, T = x, C > x|\boldsymbol{Z})^{\delta} P(D, T > x, C = x|\boldsymbol{Z})^{1-\delta} \qquad (4.2) \\ &= P(\boldsymbol{Z})[P(D, T = x|\boldsymbol{Z})P(C > x|\boldsymbol{Z})]^{\delta}[P(D, T > x|\boldsymbol{Z})P(C = x|Z)]^{1-\delta} \\ &= P(\boldsymbol{Z})P(C > x|\boldsymbol{Z})^{\delta} P(C = x|\boldsymbol{Z})^{1-\delta} \\ &\big[R(x, \boldsymbol{Z})^{D} \bar{R}(x, \boldsymbol{Z})^{1-D} P(T = x|\boldsymbol{Z})\big]^{\delta} \big[\sum_{t>x} R(t, \boldsymbol{Z})^{D} \bar{R}(t, \boldsymbol{Z})^{1-D} P(T = t|\boldsymbol{Z})\big]^{1-\delta} \\ &\propto \big[R(x, \boldsymbol{Z})^{D} \bar{R}(x, \boldsymbol{Z})^{1-D} P(T = x|\boldsymbol{Z})\big]^{\delta} \big[\sum_{t>x} R(t, \boldsymbol{Z})^{D} \bar{R}(t, \boldsymbol{Z})^{1-D} P(T = t|\boldsymbol{Z})\big]^{1-\delta} \end{aligned}$$

where the second equality holds by assuming $C \perp (T, D)|\boldsymbol{Z}$, the last item in third equality is the likelihood derived by summing over all possible realizations of the right censored covariate T and $\bar{R}(.) = 1 - R(.)$. We aim to estimate the risk score R by maximizing the observed data likelihood under monotonicity constraint given by Equation 4.1. That is,

$$\begin{aligned}\hat{R} &= \underset{R \in M_p}{\arg\max} \sum_{i=1}^{n} \log P(D_i, \delta_i, X_i, \boldsymbol{Z}_i) \qquad (4.3)\\ &= \underset{R \in M_p}{\arg\max} \sum_{i=1}^{n} D_i \Big\{ \delta_i \log \big[R(X_i, Z_i) P(T = X_i|Z_i)\big] + (1-\delta_i) \log \big[\sum_{t > X_i} R(t, Z_i) P(T = t|Z_i)\big] \Big\} \\ &+ (1 - D_i) \Big\{ \delta_i \log \big[\bar{R}(X_i, Z_i) P(T = X_i|Z_i)\big] + (1-\delta_i) \log \big[\sum_{t > X_i} \bar{R}(t, Z_i) P(T = t|Z_i)\big] \Big\}\end{aligned}$$

where M_p is the space of all non-decreasing functions on input space bounded in $[0,\ 1]$. The factorization of observed data likelihood shows that, the inference of R depends on conditional density of T given $\boldsymbol{Z}$, but has nothing to do with density of $\boldsymbol{Z}$.

The proposed approach is carried out in two stages. In step one, we use a Cox proportional hazards model to estimate the conditional distribution of T given $\boldsymbol{Z}$. In particular, we assume that $h(t|\boldsymbol{z}) = h_0(t) \exp(\beta \boldsymbol{z})$, where $h(t|\boldsymbol{z})$ is the hazard function for T given $\boldsymbol{Z} = \boldsymbol{z}$ evaluated at t, $h_0(t)$ is the baseline hazard function for T at $\boldsymbol{Z} = 0$. Let $X_{(0)} = -\infty < X_{(1)} < X_{(2)} \cdots < X_{(n)}$ denote the ordered, unique values of X from data, the conditional distribution of T given $\boldsymbol{Z}$ evaluated at $X_{(i)}$ is given by $P(T = X_{(i)}|\boldsymbol{Z}) = S(X_{(i-1)}|\boldsymbol{Z}) - S(X_{(i)}|\boldsymbol{Z}),\ i = 1, \cdots, n$, where $S(.|\boldsymbol{Z})$ is the survivor function of T estimated using the method of Breslow (Breslow, 1972) and $S(X_{(0)}|\boldsymbol{Z}) = 1$. It should be noted that the estimated survival probability from Breslow estimator will never reach zero, that is, $S(X_{(n)}|\boldsymbol{Z}) > 0$ for any $\boldsymbol{Z}$. To use the estimated survival probability of the last value of X, we create a dummy value $X_{(n+1)} = X_{(n)} + 1$, such that $P(T = X_{(n+1)}|\boldsymbol{Z}) = S(X_{(n)}|\boldsymbol{Z})$. We denote the estimated conditional distribution of T as $q^*(T|\boldsymbol{Z})$.

In step two we plug $q^*(T|\boldsymbol{Z})$ into Equation (4.3) and reconsider the maximization problem. Direct optimization under the monotonicity constraint is not a trivial task.

However, note that M_p is also precisely the space of all cumulative distribution functions on the input space. This leads us to construct an independent latent random vector $\boldsymbol{U}$ with $P\big(\boldsymbol{U} \leq (T, \boldsymbol{Z})\big) = R(T, \boldsymbol{Z})$. That is, we introduce a latent random vector $\boldsymbol{U}$ and treat the monotonic risk score function as cumulative distribution function of $\boldsymbol{U}$. Note that the log-likelihood in Equation (4.2) is identical to that of $\{D = I\big(\boldsymbol{U} \leq (T, \boldsymbol{Z})\big), \delta, X, \boldsymbol{Z}\}$. Now we can derive an EM-type algorithm with the $\boldsymbol{U}$ and T treated as missing data.

Let $(t_1, \boldsymbol{z}_1), (t_2, \boldsymbol{z}_2), \cdots, (t_m, \boldsymbol{z}_m)$ denote the support of $(T, \boldsymbol{Z})$. We write $R\big(d(t_k, \boldsymbol{z}_k)\big) = P\big(\boldsymbol{U} = (t_k, \boldsymbol{z}_k)\big),\ k = 1, \cdots, m$. Then, the nonparametric estimation of R reduces to that of the m parameters $R\big(d(t_k, \boldsymbol{z}_k)\big)$ with the understanding that $R(t, \boldsymbol{z}) = \sum_{(t_k, \boldsymbol{z}_k) \leq (t, \boldsymbol{z})} R\big(d(t_k, \boldsymbol{z}_k)\big)$. First we consider the M step. Suppose that we can observe the full data $(\boldsymbol{U}, T, \boldsymbol{Z})$, the full data log-likelihood of $(\boldsymbol{U}, T, \boldsymbol{Z})$ is

$$\sum_{i=1}^{n} \log P(\boldsymbol{U} = \boldsymbol{U}_i) + \sum_{i=1}^{n} \log P\big((T, \boldsymbol{Z}) = (T_i, \boldsymbol{Z}_i)\big) \tag{4.4}$$

The full data MLE of the point mass of $\boldsymbol{U}$ is given by $\hat{R}\big(d(t_k, \boldsymbol{z}_k)\big) = \sum_{i=1}^{n} I\big(\boldsymbol{U}_i = (t_k, \boldsymbol{z}_k)\big)/n$. Then, for the E step at the $(j+1)$th iteration, we take expectation of the full data MLE of the point mass of $\boldsymbol{U}$ given observed data $\{D_i, \delta_i, X_i, \boldsymbol{Z}_i\}$, the current estimates of the point mass of $\boldsymbol{U}$ and estimated conditional distribution of T from step one. That is,

$$R^{(j+1)}\big(d(t_k, \boldsymbol{z}_k)\big) = \frac{\sum_{i=1}^{n} E\big[I\big(\boldsymbol{U}_i = (t_k, \boldsymbol{z}_k)\big) | D_i, \delta_i, X_i, \boldsymbol{Z}_i; R^{(j)}, q^*(t|\boldsymbol{z})\big]}{n} \tag{4.5}$$

By standard derivation, the $(j+1)$th iteration of the algorithm becomes

$$R^{(j+1)}\big(d(t_k, \boldsymbol{z}_k)\big) = \frac{R^{(j)}\big(d(t_k, \boldsymbol{z}_k)\big)}{n} \Big\{ \sum_{i:\delta_i=1} \big[D_i \frac{I\big((t_k, \boldsymbol{z}_k) \leq (X_i, \boldsymbol{Z}_i)\big)}{R^{(j)}(X_i, \boldsymbol{Z}_i)} + (1 - D_i) \frac{I\big((t_k, \boldsymbol{z}_k) \not\leq (X_i, \boldsymbol{Z}_i)\big)}{\bar{R}^{(j)}(X_i, \boldsymbol{Z}_i)} \big] \tag{4.6}$$

$$+ \sum_{i:\delta_i=0} \big[D_i \frac{\sum_{t>X_i} I\big((t_k, \boldsymbol{z}_k) \leq (t, \boldsymbol{Z}_i)\big) q^*(t|\boldsymbol{Z}_i)}{\sum_{t>X_i} R^{(j)}(t, \boldsymbol{Z}_i) q^*(t|\boldsymbol{Z}_i)} + (1 - D_i) \frac{\sum_{t>X_i} I\big((t_k, \boldsymbol{z}_k) \not\leq (t, \boldsymbol{Z}_i)\big) q^*(t|\boldsymbol{Z}_i)}{\sum_{t>X_i} \bar{R}^{(j)}(t, \boldsymbol{Z}_i) q^*(t|\boldsymbol{Z}_i)} \big] \Big\}$$

In practice, a non-trivial question arises as to how the supporting points $(t_1, \boldsymbol{z}_1), \cdots, (t_m, \boldsymbol{z}_m)$ should be created. Our strategy is as follows. For fully observed observation with $\delta_i = 1$, we add $(T_i, \boldsymbol{Z}_i)$ into the support; For censored observation with $\delta_i = 0$, X_i is observed with the understanding that $T_i > X_i$. We use the observed values of T in the data which are greater than X_i and the dummy value $X_{(n+1)}$ to guess the true value of T_i. That is, we add $\{(T_j, \boldsymbol{Z}_i)|j : T_j > X_i \ \wedge \ \delta_j = 1\} \cup \{(X_{(n+1)}, \boldsymbol{Z}_i)\}$ into the support. Note that the augmented set will never be empty due to the dummy value $X_{(n+1)}$ we created in step one.

4.3 Numeric Study

Simulation studies are conducted to evaluate and compare the performance of the proposed two-step likelihood-based approach (MC) against the complete-case analysis and substitution methods based on logistic regression (LR). While method of dealing with a censored covariate in general linear models has been discussed in a number of studies (Tsimikas, Bantis, and Georgiou, 2012), their method requires additional assumption for the distribution of the censored covariate. As a result, instead of using their method, we apply complete-case analysis and substitution methods for handling censored covariate. Additionally, in order to verify the superiority of the proposed approach for handling censored covariate, we also apply complete-case analysis and substitution methods to MC. Therefore, there are five methods being compared in total, which are denoted as MC, MC-substitution, MC-complete, LR-substitution and LR-complete.

Various classification evaluation metrics might be used. Here we focus on the area under the ROC (receiver operating characteristic) curve as the performance measure, which is the most widely used index in diagnostic medicine (Begg, 1991). We consider two different conditional distributions of T given $\boldsymbol{Z}$: exponential and Weibull distributions. In addition, we study the effect of sample size and censoring rate on their classification performance.

4.3.1 Exponential distribution

For the exponential distribution as the conditional distributions of T given $\boldsymbol{Z}$, the simulation procedure is shown as follows: (1) generate the fully observed covariate $\boldsymbol{Z} \sim$ chisquare(df $= 10$). (2) simulate $T|\boldsymbol{Z} \sim$ Exponential(rate $= \lambda_0 e^{\beta_0 \boldsymbol{Z}}$) where $\lambda_0 = 0.1$ and $\beta_0 = 0.01$. (3) generate the censoring variable $C|\boldsymbol{Z} \sim$ Uniform$(0, \gamma\sqrt{\boldsymbol{Z}})$, γ is set to be 30, 10, 3 to obtain the desired censoring rates around 0.24, 0.50, 0.69, respectively. (4) calculate the true risk score as $R_0(T, \boldsymbol{Z}) = \sum_{(t,\boldsymbol{z}) \leq (T,\boldsymbol{Z})} p_{(t,\boldsymbol{z})}$, where $p_{(t,\boldsymbol{z})}$ is predefined probability point mass and $(t, \boldsymbol{z}) \leq (T, \boldsymbol{Z})$ operates component-wise. We set a few $p_{(t,\boldsymbol{z})}$ very large to generate nonlinearity in $R_0(T, \boldsymbol{Z})$. (5) generate the outcome variable $D|T, \boldsymbol{Z} \sim$ Bernoulli$(R_0(T, \boldsymbol{Z}))$. It should be noted that the risk score function satisfies the monotonic assumption, that is, the probability of $D = 1$ increases with covariates values. In addition, we let the sample size $n = 300, 500, 1000$ and we randomly select 66% of the samples as training part and the rest as testing part. The models are fitted on the training part and AUCs are calculated from testing part. Simulation results are based on 1000 replicates. Table 4.1 shows the results for exponential distribution.

From Table 4.1, We can see that MC based methods generally have better performance than LR based methods. This is because the true risk score function is a non-sigmoidal function, that is, the assumption of LR is not satisfied and thus render the bad performance of LR. Secondly, it is clearly shown that MC is dominant in obtaining the largest AUC across different sample sizes and censoring rate, and the dominance of MC becomes more evident when the censoring rate increases. As a consequence, it verifies the superiority of the proposed approach for handling censored covariate over traditional naive methods. Another interesting observation from Table 4.1 is that complete-case analysis has better performance than substitution methods when censoring rate is low or medium, but it has worse performance when censoring rate is high. It is known that complete-case analysis produces unbiased but inefficient estimation of the risk score function when the censoring is independent of the outcome. Our hypothesis is that when censoring rate is low or medium, complete-case analysis produces decent estimation in that it still uses reasonable

number of samples. Its performance, however, deteriorates quickly because its inefficiency nature dominates as censoring rate increases.

Table 4.1: Mean area under receiver operating characteristic curve AUC (SE) under Exponential distribution.

n	Censoring rate	MC	MC-substitution	MC-complete	LR-substitution	LR-complete
300	0.24	0.894* (0.028)	0.886 (0.029)	0.890 (0.028)	0.837 (0.041)	0.841 (0.040)
	0.50	0.893* (0.027)	0.876 (0.031)	0.877 (0.031)	0.842 (0.040)	0.843 (0.039)
	0.67	0.885* (0.030)	0.831 (0.040)	0.816 (0.042)	0.843 (0.040)	0.839 (0.041)
500	0.24	0.902* (0.021)	0.898 (0.021)	0.900 (0.022)	0.839 (0.035)	0.841 (0.035)
	0.50	0.901* (0.021)	0.890 (0.023)	0.892 (0.024)	0.842 (0.035)	0.839 (0.034)
	0.69	0.893* (0.023)	0.832 (0.032)	0.814 (0.043)	0.841 (0.034)	0.834 (0.037)
1000	0.24	0.902* (0.014)	0.899 (0.015)	0.901 (0.015)	0.843 (0.020)	0.846 (0.020)
	0.50	0.902* (0.014)	0.890 (0.016)	0.897 (0.015)	0.847 (0.020)	0.847 (0.021)
	0.69	0.896* (0.015)	0.835 (0.022)	0.822 (0.040)	0.847 (0.020)	0.841 (0.024)

4.3.2 Weibull distribution

For Weibull distribution as the conditional distributions of T given $\boldsymbol{Z}$, the simulation procedure is similar to that from Section 4.3.1: (1) generate the fully observed covariate $\boldsymbol{Z} \sim$ chisquare(df = 10). (2) simulate $T|\boldsymbol{Z} \sim$ Weibull(scale = $(\theta_0 e^{\beta_0 \boldsymbol{Z}})^{-1/q}$, shape = q) where $\theta_0 = 0.1$, $q = 0.9$ and $\beta_0 = 0.001$. (3) generate the censoring variable $C|\boldsymbol{Z} \sim$ Uniform$(0, \gamma\sqrt{\boldsymbol{Z}})$, γ is set to be 30, 15, 6 to obtain the desired censoring rate around 0.34, 0.50, 0.68, respectively. (4) calculate the true risk score as $R_0(T, \boldsymbol{Z}) = \sum_{(t,\boldsymbol{z}) \leq (T,\boldsymbol{Z})} p_{(t,\boldsymbol{z})}$, where $p_{(t,\boldsymbol{z})}$ is predefined probability point mass and $(t, \boldsymbol{z}) \leq (T, \boldsymbol{Z})$ operates componentwise. (5) generate the outcome variable $D|T, \boldsymbol{Z} \sim$ Bernoulli$(R_0(T, \boldsymbol{Z}))$. Simulation results are based on 1000 replicates. Table 4.2 shows the results for Weibull distribution.

Similar conclusions can be drawn from Table 4.2. MC always has the best performance in terms of AUC across different sample sizes and censoring rate. The superiority of MC becomes more evident when the censoring rate and sample size increase. Another observation is that MC has larger AUC values under Weibull scenario than that from exponential scenario.

Table 4.2: Mean area under receiver operating characteristic curve AUC (SE) under Weibull distribution.

n	Censoring rate	MC	MC-substitution	MC-complete	LR-substitution	LR-complete
300	0.34	0.913* (0.023)	0.904 (0.026)	0.911 (0.023)	0.840 (0.042)	0.846 (0.041)
	0.50	0.913* (0.023)	0.898 (0.030)	0.905 (0.027)	0.844 (0.041)	0.846 (0.041)
	0.68	0.910* (0.024)	0.879 (0.034)	0.868 (0.044)	0.846 (0.041)	0.842 (0.042)
500	0.34	0.921* (0.020)	0.915 (0.021)	0.919 (0.020)	0.838 (0.037)	0.842 (0.035)
	0.50	0.921* (0.019)	0.909 (0.021)	0.917 (0.020)	0.841 (0.036)	0.840 (0.035)
	0.67	0.918* (0.019)	0.893 (0.025)	0.881 (0.032)	0.842 (0.035)	0.838 (0.035)
1000	0.34	0.922* (0.011)	0.916 (0.013)	0.920 (0.012)	0.844 (0.022)	0.849 (0.021)
	0.50	0.922* (0.011)	0.909 (0.014)	0.918 (0.013)	0.848 (0.021)	0.849 (0.021)
	0.69	0.918* (0.013)	0.881 (0.017)	0.868 (0.034)	0.850 (0.020)	0.844 (0.025)

4.4 Analysis of PBC data: An example

In this Section, we considered the implementation of the mentioned approaches to a survival setting that involves marker measurements. From a biological point of view, we are interested in modeling the value of a diagnostic marker as a function of the time lag between the measurement and the occurrence of disease. Models of this type are utilized in a number of studies (Cai et al., 2006; Tsimikas, Bantis, and Georgiou, 2012). The data are from a double-blinded randomized placebo controlled clinical trial of the drug D-penicillamine (DPCA) used for the treatment of primary biliary cirrhosis (PBC). The trial was conducted at the Mayo Clinic during 1974–1984 (Fleming and Harrington, 2011) and involved 312 subjects. We used a threshold to dichotomize bilirubin such that positive examples and negative examples are approximately equal. Our model considers dichotomized bilirubin as a outcome (D), versus negative survival time (T), adjusted for age (Z).

In this real data, we have to deal with right-censored covariate. To be more specific, 187 out of 312 subjects have censored survival time. Furthermore, it is reasonable to assume that the input variables are monotonically associated with the risk score, in that a shorter survival time or larger age should correspond to a higher value of bilirubin, respectively. We used 10-fold cross validation to calculate the mean AUC for each method. In each iteration, the real data was split into training part and testing part. All five methods

were fitted on training data which may contain censored observations. The AUC was calculated from complete cases of testing data. Results are summarized in Table 4.3 and their corresponding ROC curves are presented in Figure 4.1.

Table 4.3: Mean area under receiver operating characteristic curve AUC for PBC data under 10-fold cross validation.

MC	MC-substitution	MC-complete	LR-substitution	LR-complete
0.755* (0.064)	0.736 (0.064)	0.732 (0.110)	0.726 (0.042)	0.734 (0.120)

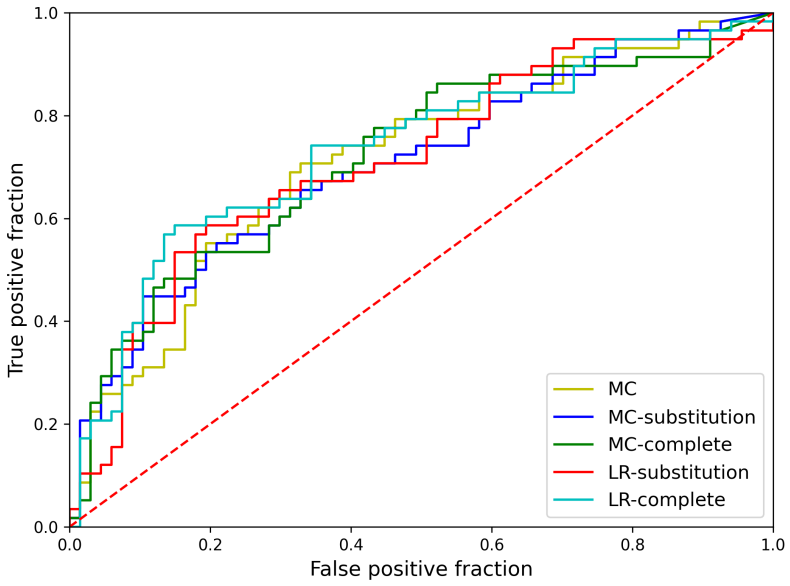

Figure 4.1: ROC curves for PBC data based on 10-fold cross validation. Methods used are "MC", "MC-substitution", "MC-complete", "LR-substitution", "LR-complete".

From Table 4.3 we can conclude that MC has the best performance, which may due to the fact that MC appropriately leverage the information from subjects with censored covariate and the monotonicity assumption is approximately satisfied. LR based methods have slightly worse performance than MC based methods, possibly because the true risk score function is a non-sigmoidal function.

4.5 Conclusion

The main contribution of the current chapter is that we extended the likelihood-based approach to monotone classification to account for randomly censored covariates in a natural and principled way. The first advantage of our method is that it does not require specifying the form of the risk score function. Instead, the risk score is estimated nonparametrically by an EM-type algorithm. Therefore, it is more robust than parametric models such as logistic regression. The second advantage of the proposed method is the ability to handle the censored covariate appropriately. Unlike complete-case analysis and substitution methods, MC handles the censored covariate via EM-type algorithm without neglecting information from censored observations or misestimating the value of censored covariate. Another merit of MC is the incorporation of monotonicity relationship between the input variables and the disease outcome. Exploitation of such established domain knowledge usually leads to gains in diagnostic accuracy. Through simulation studies and a real data example, we demonstrate that MC outperforms the simple but inefficient complete-case analysis as well as the substitution methods.

The proposed approach in this Chapter is different from the one in Chapter 3. Firstly, here we consider censored covariates instead of missing covariates. While missing and censored covariate observations share some similarities, they are fundamentally different. An observation is missing when the observed value of some variable is unknown. On the other hand, an observation is censored when the true value is only partially observed due to varying reasons. As a result, the established approach in Chapter 3 is not readily applicable to censored covariates in that it does not fully utilize the partial information contained in censored observations. Secondly, the fully nonparametric approach in Chapter 3 may not be feasible when covariates are continuous. To reduce the number of supporting points for continuous covariates, we use a Cox proportional hazards model for the distribution of the censored covariate given other covariates. This conditional distribution is then used for calculating the observed likelihood of data and an EM algorithm can be devised to compute the NPMLE of the risk score function.

Chapter 5

Evaluation metrics for ordinal classification

5.1 Introduction

Classification is one of the most important and typical tasks for different applications, involving estimation of a mapping from the feature space into a finite class space. Depending on the cardinality of the finite class space we are left with binary or multi-class classification problems. In many real life problems humans are required to compare or rate items or objects into naturally ordered classes, where there is no meaningful numeric distance between them. For example, a doctor could rate the severity of a patient's disease on a scale of minor, moderate, major, and extreme. This presence of ordering information, therefore, separates ordinal classification from nominal classification problems.

In fact, ordinal classification is a class of problems between multi-class classification and regression. As a result, one often used strategy for ordinal classification is to transform the ordinal class values into numeric quantities and then apply a regression model to the transformed data. A disadvantage of such method is that the regression model may misinterpret the ordering information in that the numeric values chosen to represent the ordinal classes are arbitrary, and thus may perform poorly in practice. Another

commonly employed strategy is to apply standard classification algorithms for nominal classes. However, the ordering information is lost when this is done, and thereby potentially deteriorating the predictive performance of a classifier. To appropriately utilize the ordering information, numerous predictive methods for ordinal data have been developed throughout the literature. For example, McCullagh (1980) developed a regression model incorporating ordinal information on the data eliminating the need for assigning labels. Following his work, Tutz (2003) presented an extension of it through the generalization of the additive model (Hastie and Tibshirani, 2017) by incorporating nonparametric terms. In the general setting, Frank and Hall (2001) introduced a simple process enabling standard binary classification algorithms to exploit the ordering information. The original ordinal classification problem was converted into a series of binary classification problems with the ordering information encoded, and thereby making it possible for standard binary classification algorithms. Cardoso and Da Costa (2007) proposed a similar method to reduce the problem to the standard two-class setting, using a nonparametric procedure called data replication method. In their paper this method was mapped into neural networks and support vector machines. Other techniques for ordinal classification can be found in Costa, Alonso, and Cardoso (2008), Costa, Sousa, and Cardoso (2010), Shashua, Levin, et al. (2003), and Kotsiantis and Kanellopoulos (2010).

There is no shortage of performance evaluation metrics for nominal multi-class classification and regression models. Conventional performance evaluation metrics appropriate for nominal classes or regression, however, are unsuitable for ordinal classification problems in that they do not account for the ordinality of the target classes. In recent years, there have been growing interest in designing proper evaluation metrics for ordinal classification models. For instance, Baccianella, Esuli, and Sebastiani (2009) proposed a method to adapt exiting evaluation measures into ones robust to imbalance, due to the fact that ordinal classification often involves highly imbalanced datasets. In fact, their method was inspired by the well-known distinction between the microaveraged and macroaveraged versions of F_1. The transformed metrics is obtained by averaging over the classes instead of

samples and thereby making it robust to class imbalance. To capture the ordering information, some investigators have advocated the use of rank order measures (Lee and Liu, 2002; Vanbelle and Albert, 2009) in that these measures only look at the order relation between the true and predicted class numbers. However, such rank order measures cannot capture how much the predictions diverge from true labels. Alternatively, Cardoso and Sousa (2011) proposed a novel metric defined directly in the confusion matrix, which can capture how much the predictions diverge from true labels and how "inconsistent" the classifier is in regard to the relative order of the labels. Motivated from the probability interpretation of the area under the receiver operating characteristic (ROC) curve in the binary case, Waegeman, De Baets, and Boullart (2006) extended the volume under the ROC surface (VUS) to ordinal setting and proposed three approximations to this measure to alleviate the computational burden.

All the aforementioned metrics for ordinal data are devised for "ordinary" classifier, whose output is a single label or a scalar-valued quantity. However, most of the time we have probabilistic classifiers, e.g., logistic regression (Richards, Hammitt, and Tsevat, 1996), support vector machine (SVM) (Cristianini, Shawe-Taylor, et al., 2000), or classification trees (Breiman et al., 2017), with the output being a probability distribution over a set of labels, rather than a single label. To use the aforementioned evaluation metrics of ordinal data, one has to convert probabilistic classifiers into "ordinary" classifier using a given threshold or decision rule. Such conversion may potentially lose the information from predicted results and as a result, are likely to misevaluate the performance of a classifier.

In this chapter our main goal is to propose novel performance evaluation metrics specifically designed for ordinal classification, which can not only appropriately capture the ordering information but also evaluate probabilistic classifiers directly, such that the information from predicted probability distributions can be fully utilized. The first proposed metric is adapted from the area under the ROC curve (AUC). It requires partitioning the true labels and predicted values into a series of binary data, with the ordering information

encoded. The AUC for each binary series is obtained and the overall performance is the average over all levels. The latter two proposed metrics are simple and interpretable generalizations of the Harrell's concordance index (C-INDEX) (Harrell et al., 1982; Harrell et al., 2015). In addition, we show the optimality of all AUC based metrics through Neyman-Pearson lemma (Neyman and Pearson, 1933). Through extensive simulation studies, we confirm the usefulness of proposed metrics and argue that conventional nominal classification metrics do not adequately take into account for the ordinality of the target classes and thus should not be used.

5.2 Common evaluation metrics for classification

In this Section, we review the commonly used performance evaluation metrics. Note that the survey cannot be exhaustive due to limited scope of the article. Let $Y \in \{1, 2, \cdots, c\}$ denote the outcome variable consisting of c classes, with the labels reflecting intrinsic order to the classes under ordinal setting and otherwise under nominal setting. Let $\boldsymbol{X} \in \mathscr{X}$ denote the input vector. The data are given by $(Y_i, \boldsymbol{X}_i),\ i = 1, \cdots, n$, where n is the sample size. We let $n_j = \sum_{i=1}^{n} \mathbb{1}(Y_i = j)$ denote the number of examples of class j and $p(j) = n_j/n$ denote the class prevalence of class j. A probabilistic classifier is a function that maps a predictor $\boldsymbol{x}$ to its c-tuple of estimated conditional probabilities $\hat{\boldsymbol{P}}(\boldsymbol{x}) = \{\hat{P}_1(\boldsymbol{x}), \cdots, \hat{P}_c(\boldsymbol{x})\}$, where $P_j(\boldsymbol{x}) = P(Y = j|\boldsymbol{X} = \boldsymbol{x}),\ j = 1, \cdots, c$. That is, $\hat{\boldsymbol{P}} : \mathscr{X} \longrightarrow \mathscr{S}^c$, where $\mathscr{S}^c$ is the c dimensional unit simplex $\{(p_1, \cdots, p_c) : \sum_{j=1}^{c} p_j = 1 \text{ and } p_j \geq 0, j = 1, \cdots, c\}$. On the other hand, an "ordinary" classifier is a function mapping a predictor $\boldsymbol{x}$ to a scalar-valued output, that is, $\gamma : \mathscr{X} \longrightarrow \{1, 2, \cdots, c\}$. A probabilistic classifier can be converted into a "ordinary" classifier using a given threshold or decision rule. For example, one often employed conversion strategy is to assign the class of largest conditional probability to a new example. The resulting classifier is called the Bayes classifier, which is optimal in the sense that it minimizes the unweighted misclassifcation error. However, such conversion is specific to the choice of the threshold or decision rule and may potentially lose the information from predicted probabilities.

In general, performance measures can be categorized by many aspects, in this work, we focus on two aspects in particular, that is, (1) whether performance measures are devised for "ordinary" classifiers or probabilistic classifiers. (2) whether performance measures are utilizing the ordering information. We first review metrics appropriate for nominal data in that they do not assume or utilize any ordering information of the classes. The first three qualitative metrics are designed for "ordinary" classifiers in that they just use the term $\gamma(\boldsymbol{X}_i)$ in definition, while the other six metrics directly evaluate the predicted conditional probabilities $\hat{P}_j(\boldsymbol{X}_i)$.

- **Accuracy**: (Acc).
$$Acc = \frac{\sum_{i=1}^{n} \mathbb{1}(\gamma(\boldsymbol{X}_i) = Y_i)}{n}$$

- **Kappa statistic**: (KapS).
$$KapS = \frac{P(A) - P(E)}{1 - P(E)}$$
Where $P(A) = Acc$ and
$$P(E) = \frac{\sum_{j=1}^{c} \sum_{i=1}^{n} \sum_{t=1}^{n} \mathbb{1}(\gamma(\boldsymbol{X}_i) = j)\mathbb{1}(Y_t = j)}{n^2}$$

- **Mean F-measure**: (MFM).
$$\mathrm{F}_1 = 2\frac{\text{precision.recall}}{\text{precision} + \text{recall}}$$

- **AUC of each class against the rest, using the uniform class distribution**: (AUNU)
$$\text{AUC}(j,k) = \frac{\sum_{i=1}^{n} \sum_{t=1}^{n} \mathbb{1}(Y_i = j, Y_t = k) I(\hat{P}_j(\boldsymbol{X}_i), \hat{P}_j(\boldsymbol{X}_t))}{n_j.n_k}$$
$I(.)$ is a comparison function $I(a,b) = 1$ iff $a > b$, $I(a,b) = 0$ iff $a < b$ and $I(a,b) =$

0.5 iff $a = b$.

$$\text{AUNU} = \frac{\sum_{j=1}^{c} \text{AUC}(j, \text{rest}_j)}{c}$$

- **AUC of each class against the rest, using the priori class distribution**: (AUNP).

$$\text{AUNP} = \sum_{j=1}^{c} p(j)\text{AUC}(j, \text{rest}_j)$$

- **AUC of each class against each other, using the uniform class distribution**: (AU1U). (Hand and Till, 2001)

$$\text{AU1U} = \frac{1}{c(c-1)} \sum_{j=1}^{c} \sum_{k \neq j} \text{AUC}(j, k)$$

- **Mean Squared Error or Brier score**: (MSE-p). (Brier et al., 1950)

$$\text{MSE-p} = \frac{\sum_{j=1}^{c} \sum_{i=1}^{n} \left[\mathbb{1}(Y_i = j) - \hat{P}_j(\boldsymbol{X}_i)\right]^2}{n.c}$$

- **LogLoss**: (LogL).

$$\text{LogL} = -\frac{\sum_{j=1}^{c} \sum_{i=1}^{n} \mathbb{1}(Y_i = j) \log_2 \hat{P}_j(\boldsymbol{X}_i)}{n}$$

The previously reviewed metrics are unsuitable for ordinal classification problems because no ordering information of Y is needed in their definition. Some papers on ordinal classification, however, simply use these metrics as evaluation measure. For example, Frank and Hall (2001) simply reported accuracy as an error measure. In the following, we review evaluation metrics which utilize the ordering information of the classes. Note that these metrics are all designed for "ordinary" classifiers in that only the term $\gamma(\boldsymbol{X}_i)$ is used.

- **Mean Absolute Error**: (MAE).

$$\text{MAE} = \frac{\sum_{i=1}^{n} |\gamma(\boldsymbol{X}_i) - Y_i|}{n}$$

- **Mean Square Error**: (MSE).

$$\text{MSE} = \frac{\sum_{i=1}^{n} |\gamma(\boldsymbol{X}_i) - Y_i|^2}{n}$$

- **Spearman's Rank Correlation Coefficient**: (r_s).

$$r_s = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(\gamma(\boldsymbol{X}_i) - \overline{\gamma(\boldsymbol{X})})}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}\sqrt{\sum_{i=1}^{n}(\gamma(\boldsymbol{X}_i) - \overline{\gamma(\boldsymbol{X})})^2}}$$

  where $\overline{Y}$ and $\overline{\gamma(\boldsymbol{X})}$ are the average of Y_i and $\gamma(\boldsymbol{X}_i)$, $i = 1, \cdots, n$.

- **Harrell's Concordance Index**: (C-Index).

$$\text{C-Index} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{n} \mathbb{1}(Y_i > Y_t) I(\gamma(\boldsymbol{X}_i), \gamma(\boldsymbol{X}_t))}{\sum_{i=1}^{n}\sum_{t=1}^{n} \mathbb{1}(Y_i > Y_t)}$$

- **Kendall's tau-b**: (τ_b).

$$\tau_b = \frac{\sum_{i=1}^{n}\sum_{t=1}^{n} c^*_{i,t} c_{i,t}}{\sqrt{(\sum_{i=1}^{n}\sum_{t=1}^{n} c^2_{i,t})(\sum_{i=1}^{n}\sum_{t=1}^{n} c^{*2}_{i,t})}}$$

  where $c_{i,t}$ is 1 if $\gamma(\boldsymbol{X}_i) > \gamma(\boldsymbol{X}_t)$, 0 if $\gamma(\boldsymbol{X}_i) = \gamma(\boldsymbol{X}_t)$ and -1 if $\gamma(\boldsymbol{X}_i) < \gamma(\boldsymbol{X}_t)$, where $i, t = 1, \cdots, n$. Similarly for $c_{i,t}$.

These metrics are more appropriate for ordinal classification problem as all of them utilize the ordering information of the classes. It should be noted MAE, MSE and r_s are actually metrics for regression task, and the numeric values chosen to represent the ordinal classes will influence the performance measurement given by MAE, MSE and r_s. On the other hand, C-Index and τ_b only look at the order relation between the true and

predicted classes numbers, and thus avoid the influence of the number chosen to represent the classes.

To the best of the author's knowledge, most the ordinal evaluation metrics including these recently introduced in Baccianella, Esuli, and Sebastiani (2009), Cardoso and Sousa (2011), and Vanbelle and Albert (2009) are devised for "ordinary" classifier, i.e., they just use the term $\gamma(\boldsymbol{X}_i)$ rather than $\hat{P}_j(\boldsymbol{X}_i)$. However, sometimes we may have probabilistic classifiers, such as logistic regression, random forest, naive Bayes etc. To use such metrics, one has to convert probabilistic classifiers into "ordinary" classifier using a given threshold or decision rule. Such conversion may potentially lose the information from predicted results and as a result, are likely to misevaluate the performance of a classifier. To solve this problem, we proposed metrics for ordinal data which directly use $\hat{P}_j(\boldsymbol{X}_i)$ in the definition such that the information from predicted probability distributions can be fully utilized.

5.3 Proposed ordinal classification metrics

5.3.1 Ordinal AUC

The first proposed metric is adapted from AUNU and AU1U (Hand and Till, 2001). For nominal data, there are two ways to partition the data based on Y. That is, the data can be partitioned either using each class against the rest or each class against each other, with the first one corresponding to AUNU and second one corresponding to AU1U, respectively. In contrast, with the ordering information, we can naturally partition ordinal data at each level. To be more specific, for each $j \in \{1, 2, \cdots, c-1\}$, the data can be partitioned into $\{(\boldsymbol{X}_i, Y_i) : Y_i \leq j\}$ and $\{(\boldsymbol{X}_i, Y_i) : Y_i > j\}$. A probabilistic classifier will give us estimates of the probability that each test point belongs to each class $\hat{P}_j(\boldsymbol{x})$, or equivalently the probability of each test point belongs to the two partitioned fragments $\hat{F}_j(\boldsymbol{x})$ and $1 - \hat{F}_j(\boldsymbol{x})$, where $\hat{F}_j(\boldsymbol{x})$ is the estimated cumulative distribution function of level j, i.e., $\hat{F}_j(\boldsymbol{x}) = \sum_{k=1}^{j} \hat{P}_k(\boldsymbol{x})$.

To evaluate the estimated probabilities, we are interested in how well the predicted probabilities from the first partitioned fragment are separated from the other. Therefore, we define the measure AUC(j) as the probability that a randomly chosen example from $\{(\boldsymbol{X}_i, Y_i) : Y_i \leq j\}$ will have a bigger estimated probability of belonging to $\{(\boldsymbol{X}_i, Y_i) : Y_i \leq j\}$ than a randomly chosen example from $\{(\boldsymbol{X}_i, Y_i) : Y_i > j\}$ for each $j = 1, \cdots, c-1$. The empirical AUC for level j is defined as

$$\text{AUC}(j) = \frac{\sum_{i=1}^{n} \sum_{t=1}^{n} \mathbb{1}(Y_i \leq j, Y_t > j) I(\hat{F}_j(\boldsymbol{X}_i), \hat{F}_j(\boldsymbol{X}_t))}{n_{\leq j}.n_{>j}} \tag{5.1}$$

This measure reduces to the Mann-Whitney statistic in the binary case. Then the overall performance is the average of this over all levels:

$$\text{OAUC} = \frac{\sum_{j=1}^{c-1} \text{AUC}(j)}{c-1} \tag{5.2}$$

Optimality of likelihood ratio and risk score

Now we present the optimality of likelihood ratio and risk score for all variants of AUC based metrics. Firstly, let us consider the nominal metrics. Given a hypothesis test $H_0 : Y = k$, $H_1 : Y = j$. From Neyman–Pearson lemma (Neyman and Pearson, 1933) we know that $\text{LR}_{j,k}(\mathbf{X}) = P(\mathbf{X}|Y = j)/P(\mathbf{X}|Y = k)$ will generate uniformly better ROC curve than any other decision rule based on $\mathbf{X}$. Therefore, $LR_{j,k}(\mathbf{X})$ has the largest AUC(j, k) among all decision rules based on $\mathbf{X}$. Furthermore, we can show that $P(Y = j|\mathbf{X})$ is a monotone increasing function of $LR_{j,t}(\mathbf{X})$:

$$P(Y = j|\mathbf{X}) = \frac{LR_{j,t}(\mathbf{X})P(Y = j)}{LR_{j,t}(\mathbf{X})P(Y = j) + P(Y = t) + \text{constant}} \tag{5.3}$$

Therefore, $P(Y = j|\mathbf{X})$ has the same ROC curve as $LR_{j,i}(\mathbf{X})$ and it yields the same optimal decision rule as $LR_{j,i}(\mathbf{X})$. If a learning algorithm has the estimated conditional probabilities $\hat{\mathbf{P}}$ converging to the true underlying conditional probabilities $\mathbf{P}$, then it guarantees to produce largest AU1U, AUNU, AUNP among other learning algorithms

in large samples.

Similarly for ordinal data. Consider a hypothesis test $H_0 : Y > j$, $H_1 : Y \leq j$ for $j = \{1, 2, \cdots, c-1\}$. From Neyman–Pearson lemma we know that $LR_j(\mathbf{X}) = P(\mathbf{X}|Y \leq j)/P(\mathbf{X}|Y > j)$ will generate uniformly better ROC curve than any other decision rule based on $\mathbf{X}$. Furthermore, we can show that $F_j(\mathbf{X}) = P(Y \leq j|\mathbf{X})$ is a monotone increasing function of $LR_j(\mathbf{X})$:

$$F_j(\mathbf{X}) = P(Y \leq j|\mathbf{X}) = \frac{LR_j(\mathbf{X})P(Y \leq j)}{LR_j(\mathbf{X})P(Y \leq j) + P(Y > j)} \tag{5.4}$$

Therefore, $F_j(\mathbf{X})$ has the same ROC curve as $LR_j(\mathbf{X})$ and it yields the same optimal decision rule. If a learning algorithm has the estimated conditional probabilities $\hat{\mathbf{P}}$ converging to the true underlying conditional probabilities $\mathbf{P}$, or equivalently $\hat{\mathbf{F}}$ converging to $\mathbf{F}$, then it guarantees to produce largest OAUC among other learning algorithms in large samples.

5.3.2 Randomized C-Index and Quantile C-Index

The important point to note here is that even though OAUC directly evaluates estimated probabilities, it only takes into account the ranking, which corresponds to the term $I(\hat{F}_j(\boldsymbol{X}_i), \hat{F}_j(\boldsymbol{X}_t))$, but not the direct predicted values of probability. Alternatively, we consider another way to generalize the AUC to ordinal data, which is based on the Harrell's concordance index (C-Index).

C-Index is defined as the probability that $\gamma(\boldsymbol{X})$ and Y are concordant among the "comparable" pairs, i.e., $C(\gamma) = P(\gamma(\boldsymbol{X}_i) < \gamma(\boldsymbol{X}_t)|Y_i < Y_t) + 0.5P(\gamma(\boldsymbol{X}_i) = \gamma(\boldsymbol{X}_t)|Y_i < Y_t)$, and it reduces to the AUC when Y is binary. However, C-index requires that the output is a single label. With simplex-valued output $\hat{\boldsymbol{P}}(\boldsymbol{x})$ from a probabilistic classifier, we can pick a mapping $\kappa : \mathscr{S}^c \longrightarrow \{1, 2, \cdots, c\}$ to convert the probabilistic classifier into "ordinary" classifier, that is, $\kappa \; o \; \hat{\boldsymbol{P}} : \mathscr{X} \longrightarrow \{1, 2, \cdots, c\}$. One commonly used κ is constructed by taking varying quantiles of $\hat{\boldsymbol{P}}(\boldsymbol{x})$. That is,

$$\kappa_q \; o \; \hat{\boldsymbol{P}} := \min\{j = 1, \cdots, c : \hat{F}_j(\boldsymbol{x}) \geq q\}, \; 0 \leq q \leq 1 \tag{5.5}$$

Then C-index can be applied to $\kappa_q \, o \, \hat{\boldsymbol{P}}$. Nevertheless, this metric is specific to the choice of q and falls short of measuring the intrinsic value of $\hat{\boldsymbol{P}}$.

A simple way to fix this is to plot $C(\kappa_q \, o \, \hat{\boldsymbol{P}})$ against $0 \leq q \leq 1$ and calculate the area under the curve. This is the first generalization of C-index for simplex-valued output and we call this metric Quantile C-index (Q-C-index). One disadvantage of this measure is that it is unclear if this area has any intuitive meaning. Alternatively, we propose another simple and interpretable generalization of C-Index. Think of the predicted probability distributions $\hat{\boldsymbol{P}}$ as a mechanism of generating random predicted outcomes $\hat{Y}|\boldsymbol{X} \sim \hat{\boldsymbol{P}}(\boldsymbol{X})$. Following the definition of C-Index, consider

$$C(\hat{\boldsymbol{P}}) = P(\hat{Y}_i < \hat{Y}_t | Y_i < Y_t) + \frac{1}{2} P(\hat{Y}_i = \hat{Y}_t | Y_i < Y_t) = E\{L[\hat{\boldsymbol{P}}(\boldsymbol{X}_i), \hat{\boldsymbol{P}}(\boldsymbol{X}_t)] | Y_i < Y_t\} \quad (5.6)$$

where $L\{\hat{\boldsymbol{P}}(\boldsymbol{X}_i), \hat{\boldsymbol{P}}(\boldsymbol{X}_t)\} = \sum_{l=2}^{c} \sum_{k=1}^{l-1} \hat{P}_k(\boldsymbol{X}_i) \hat{P}_l(\boldsymbol{X}_t) + 2^{-1} \sum_{k=1}^{c} \hat{P}_k(\boldsymbol{X}_i) \hat{P}_k(\boldsymbol{X}_t)$ is the randomized inner probability of $\hat{Y}_i < \hat{Y}_t$ under fixed $\boldsymbol{X}_i$ and $\boldsymbol{X}_t$. The empirical version can be obtained by:

$$\hat{C}(\hat{\boldsymbol{P}}) = \frac{\sum_{i=1}^{n} \sum_{t=1}^{n} \mathbb{1}(Y_i < Y_t) L\{\hat{\boldsymbol{P}}(\boldsymbol{X}_i), \hat{\boldsymbol{P}}(\boldsymbol{X}_t)\}}{\sum_{i=1}^{n} \sum_{t=1}^{n} \mathbf{1}(Y_i < Y_t)} \quad (5.7)$$

5.4 Numeric Study

In this Section we evaluate the behaviour of the different performance metrics, where it is possible to define a reasonable reference behaviour. We constrain our attention to metrics for probabilistic classifier including AUNU, AUNP, AU1U, MSE-p, LogL for nominal data and the proposed OAUC, R-C-Index, Q-C-Index for ordinal data.

5.4.1 Ordering information of classes

To illustrate the necessarity of utilizing ordering information, we created the following synthetic classification results. Suppose that two probabilistic classifiers produce the following probability distributions ($c = 3, n = 3$) where the true labels are $(1, 2, 3)$.

$$C1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad C2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad Y_{\text{true}} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

The only difference between the two probabilistic classifiers is that, for the first example C1 predicts class 1 to be class 3 with probability one while C2 predicts it to be class 2 with probability one. One would expect that a valid performance evaluation metric would output the performance of C2 superior to that of C1 in that predicting class 1 to be class 3 is a more severe error than predicting it to be class 2. Table 5.1 presents the results for the different metrics.

Table 5.1: Results for ordering information of classes

Classifier	MSE-p	LogL	AU1U	OAUC	R-C-Index	Q-C-Index
C1	0.2222	11.51	0.75	0.625	0.5	0.5
C2	0.2222	11.51	0.75	0.75	0.83333	0.8245

Table 5.1 shows that AU1U, MSE-p, LogL are unable to detect any performance difference between C1 and C2. This results from the fact that they are not leveraging the ordering information of the classes and every misclassification is considered equally costly. In comparison, all proposed ordinal metrics can capture the superiority of C2 over C1.

5.4.2 Calibration and ranking

In this toy example we investigate the calibration and ranking properties of these metrics. Suppose that two probabilistic classifiers produce the following probability distributions $(c = 3, n = 3)$ where the true labels are $(1, 2, 3)$.

$$C3 = \begin{bmatrix} 0.998 & 0.001 & 0.001 \\ 0.001 & 0.998 & 0.001 \\ 0.001 & 0.001 & 0.998 \end{bmatrix} \quad C4 = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.3 & 0.1 & 0.6 \end{bmatrix} \quad Y_{\text{true}} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

From Table 5.2 it shows that AU1U and OAUC are unable to detect the performance difference between C3 and C4. It is due to the fact that AUC based metrics only take

the ranking into account, which corresponds to the term $I(\hat{F}_j(\boldsymbol{X}_i), \hat{F}_j(\boldsymbol{X}_t))$, but not the actual values of predicted conditional probabilities. In contrast, MSE-p, LogL, R-C-Index and Q-C-Index use the actual values of predicted conditional probabilities, and all of them consider C3 as a better classifier in that it is more "confident" for every correct prediction.

Table 5.2: Results for calibration and ranking

Classifier	MSE	LogL	AU1U	OAUC	R-C-Index	Q-C-Index
C3	0.000002	0.002	1.0	1.0	0.998	0.974
C4	0.091	0.520	1.0	1.0	0.668	0.728

5.4.3 Robustness to imbalance

In the following example, we present the robustness of these metrics against class imbalance. Consider an imbalanced data with n examples of class 1, one example of class 2 and one example of class 3. Suppose there is a trivial probabilistic classifier, which always assigns the majority class (class 1) with probability one and other classes with probability zero. A preferred performance metric should not reward such trivial classifier. Table 5.3 shows that MSE-p and LogL go to 0 as n goes to infinity, while AU1U, OAUC, R-C-Index, Q-C-Index stay constant as n increases. This results from the fact MSE and LogL are based on a sum of the classification errors across examples, while other metrics reply on "pairwise" comparison and thus are robust against class imbalance.

$$C5 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ . & . & . \\ 1 & 0 & 0 \end{bmatrix} \quad Y_{\text{true}} = \begin{bmatrix} 2 \\ 3 \\ 1 \\ . \\ 1 \end{bmatrix}$$

5.4.4 Experiments with real classifiers

To fully confirm the usefulness of the proposed metrics for ordinal data, we conducted experiments with real classifiers. Following a similar experiment setting with Cardoso and

Table 5.3: Results for robustness to imbalance

Classifier	n	MSE	LogL	AU1U	OAUC	R-C-Index	Q-C-Index
C5	100	0.0130	0.677	0.5	0.5	0.5	0.5
	1000	0.0013	0.0689	0.5	0.5	0.5	0.5
	2000	0.0006	0.0345	0.5	0.5	0.5	0.5

Da Costa (2007) and Cardoso and Sousa (2011), our synthetic dataset is generated in the following steps:

1. Generate covariate matrix with sample size 500 where $\mathbf{x} = [x_1, x_2, x_3]$ are in the unit hypercube according to a uniform distribution.

2. Assign each example a score corresponding to

$$\text{y-score} = \prod_{i=1}^{3}(x_i - 0.5)$$

3. Divide the range of y-score into a given number of intervals. The cutoff values of each interval is calculated in such a way that the number of instances in each interval is approximately constant.

4. Split the synthetic dataset into 33% for testing and 67% for training. Denote the training data as {x-train, y-score-train} and testing data as {x-test, y-score-test}.

5. Create corrupted version of training data by adding noise to y-score-train. That is, y-noisy-score-train = y-score-train + ϵ, where $\epsilon \sim N(0, \sigma^2)$.

6. Discretize the y-score-train, y-score-test and y-noisy-score-train into ordinal quantities using the cutoff values from step 3. Use y-train, y-test and y-train-noisy to denote the discretized ordinal quantities.

We applied the ordinal classification algorithm developed by Frank and Hall (2001) to the synthetic dataset. The baseline binary classifier we used is the binary SVM with RBF kernel. The Frank & Hall's method is fitted using non-corrupted training data {x-train, y-train } as well as the corrupted training data {x-train, y-train-noisy }. In addition,

the training data is considered as mild, moderate and severe corrupted depending on the standard error of the noise ϵ, which are $0.005, 0.01, 0.05$ respectively. The proposed ordinal classification metrics were calculated based on test data. In addition, in order to evaluate the influence of different numbers of classes on the classification performance, we discretized the target value into three, five, and ten intervals respectively. This procedure was repeated one hundred times to obtain more stable results for performance estimation.

Table 5.4 shows that all proposed evaluation metrics are in agreement with the expected conclusion that the performance of Frank & Hall's approach fitted on non-corrupted training data is better than that fitted on corrupted training data, and the performance difference becomes more significant as severity of corruption increases. Another interesting observation is that the degradation of OAUC is not as great as that from R-C-Index and Q-C-Index, which may due to the fact that OAUC only uses the ranking instead of direct values of the predicted probabilities and, as a result, it is less sensitive to the noise in the data.

Table 5.4: Performance average (std. dev.) results for the synthetic datasets

Number of classes	Corruption level	OAUC	R-C-Index	Q-C-Index
3	No	0.989* (0.004)	0.897* (0.012)	0.893* (0.010)
	Mild	0.988 (0.005)	0.829 (0.016)	0.850 (0.013)
	Moderate	0.980 (0.008)	0.758 (0.018)	0.792 (0.016)
	Severe	0.857 (0.049)	0.555 (0.023)	0.584 (0.033)
5	No	0.987* (0.003)	0.890* (0.011)	0.898* (0.009)
	Mild	0.985 (0.004)	0.834 (0.015)	0.877 (0.010)
	Moderate	0.978 (0.007)	0.761 (0.019)	0.845 (0.013)
	Severe	0.879 (0.042)	0.545 (0.017)	0.611 (0.037)
10	No	0.988* (0.003)	0.874* (0.012)	0.901* (0.009)
	Mild	0.987 (0.003)	0.810 (0.014)	0.887 (0.009)
	Moderate	0.983 (0.005)	0.731 (0.018)	0.863 (0.011)
	Severe	0.887 (0.038)	0.530 (0.011)	0.638 (0.040

5.4.5 Experiments with real datasets

The superiority of Frank & Hall's algorithm over the conventional classification algorithms for nominal classes has been shown in a number of previous studies (Frank and Hall, 2001;

Cardoso and Da Costa, 2007; Cardoso and Sousa, 2011). However, these studies simply reported evaluation measures for nominal classes. To further confirm the superiority of Frank & Hall's algorithm, the following experiments were conducted with sets of real ordinal data using the proposed metrics. The first dataset, SWD, contains real-world assessments of qualified social workers regarding the risk facing children if they stayed with their families at home and is composed by 10 features and 4 classes. The second dataset, LEV, contains examples of anonymous lecturer evaluations, taken at the end of MBA courses and is composed by 4 features and 5 classes. The third dataset which we worked on was the ESL dataset containing 488 profiles of applicants for certain industrial jobs. Features are based on psychometric tests results and interviews with the candidates performed by expert psychologists. The class assigned to each applicant was an overall score corresponding to the degree of fitness for the type of job.

We compared two different multiclass classification strategies:

- A conventional multiclass classifier based on the one-against-rest strategy. The baseline binary classifier is the binary SVM with RBF kernel.
- The multiclass classifier adapted for ordinal data by Frank and Hall. The baseline binary classifier is again the binary SVM with RBF kernel.

The performance estimation results were calculated based on 10-fold cross validation, which are presented in Table 5.5. From Table 5.5, it confirms the superiority of Frank & Hall's algorithm specific to ordinal data over the conventional approaches. Secondly, the superiority of Frank & Hall's algorithm based on OAUC is not as obvious as that from R-C-Index and Q-C-Index. Again, this is due to the fact that OAUC only used the ranking instead of direct values of the predicted probabilities.

5.5 Conclusion

In this work, we conducted systematic review of common evaluation metrics for classification and provided taxonomy of measures according to two aspects: (1) whether per-

Table 5.5: Performance average (std. dev.) results for the real datasets

Dataset	Strategy	OAUC	R-C-Index	Q-C-Index
ESL	OVR	0.917 (0.049)	0.774 (0.013)	0.818 (0.015)
	Frank&Hall	0.918* (0.043)	0.819* (0.015)	0.855* (0.017)
LEV	OVR	0.783* (0.038)	0.662 (0.011)	0.706 (0.013)
	Frank&Hall	0.714 (0.048)	0.689* (0.013)	0.726* (0.016)
SWD	OVR	0.772 (0.050)	0.600 (0.012)	0.633 (0.016)
	Frank&Hall	0.803* (0.032)	0.625* (0.012)	0.653* (0.016)

formance measures are devised for "ordinary" classifier or probabilistic classifiers. (2) whether performance measures are designed for nominal or ordinal data. To date, most ordinal evaluation metrics are devised for "ordinary" classifier with a single label as its output. However, most of the time we have probabilistic classifiers with simplex-valued output. In order to fill the gap of evaluation metrics for probabilistic classifier under ordinal setting, we have proposed three novel metrics. The first proposed metric is adapted from the area under the ROC curve (AUC), which requires partitioning the true labels and predicted values into a series of binary data, with the ordering information encoded. The AUC for each level is obtained and the overall performance is the average across all levels. The latter two proposed metrics are simple and interpretable generalizations of the Harrell's concordance index (C-INDEX). To the best knowledge of the authors, this work is the first to introduce performance evaluation metrics for probabilistic classifier under ordinal setting.

As is known, it is a challenging task to argue a new metric is better than current ones. To overcome this difficulty and illustrate the usefulness of the new metrics, we created toy classification examples where it is possible to define a reasonable reference behaviour. In addition, the study is further completed with an experiment with real classifiers as well as an experiment in real datasets. We further confirmed the superiority of Frank & Hall's algorithm specific to ordinal data over the conventional approaches.

Chapter 6

Conclusion

In this dissertation, we studied the statistical methods for combining diagnostic tests and performance evaluation metrics for ordinal data. In biomedical studies, it is usually the case that several diagnostic tests can be performed on an individual or multiple disease markers are available simultaneously, and that many of them may be associated with the clinical outcome. In practice, a single test or marker often has limited diagnostic performance. Therefore, it is important to combine multiple sources of information available to achieve higher classification performance. In the first project, we provided a survey of the current state of the art in methods for combining multiple tests. We categorized existing methods into three general groups and conducted extensive simulation studies to compare the performance of different combination methods. The reviewed methods serve as benchmark for developing new combination approaches in the following projects.

In the second project, we considered the problem of combining multiple tests whose values are missing at random (MAR). In addition, we aimed to exploit the known monotonicity relationship between the input variables and the disease outcome for gains in diagnostic accuracy. We developed a novel likelihood-based approach to monotone classification that accounts for missing inputs in a natural and principled way. The risk score function is obtained through the nonparametric maximum likelihood estimation (NPMLE). A novel expectation-maximization (EM)-type algorithm was devised to compute the NPMLE by

treating the monotonicity-constrained risk score function as a cumulative distribution for a latent random vector. Through simulation studies and a real data example, we demonstrated that the proposed method outperforms state of the art in methods for combining multiple inputs under monotonic assumption, especially when the inputs contain missing data. We illustrated our approach with a dataset from a recent nonalcoholic fatty liver disease (NALFD) study.

In the third project, our approach established in the second part was extended to the scenario where one covariate is randomly censored. The proposed approach consists of two steps. In step one, we use a Cox proportional hazards model for the distribution of the censored covariate given other covariates in the model, this conditional distribution is used for calculating the observed likelihood of data. In step two, a similar expectation-maximization (EM)-type algorithm is devised, based on observed data likelihood from step one, to compute the NPMLE of the monotonicity-constrained risk score function. Through simulation studies, we demonstrated that the proposed method outperforms the simple but inefficient complete-case analysis as well as the substitution methods. We applied our method to the data set from a primary biliary cirrhosis (PBC) study conducted at Mayo Clinic.

The proposed methods in part two and three can be extended to multi-class cases, where the labels have an inherent order but no meaningful numeric distance between them. A natural question arises as to how to evaluate the classification performance under such setting. Therefore, in the fourth project, we considered the problem of performance evaluation metrics for ordinal classification. We proposed three novel performance evaluation metrics that better capture the ordinality of the outcomes. The first proposed metric was adapted from the area under the receiver operating characteristic (ROC) curve (AUC), which requires partitioning the true labels and predicted values into a series of binary data, with the ordering information encoded. The AUC for each level is obtained and the overall performance is the average across all levels. The latter two proposed metrics are simple and interpretable generalizations of the Harrell's concordance index (C-INDEX).

Moreover, we showed the optimality of the AUC based metrics through Neyman-Pearson lemma. We conducted extensive simulation studies to confirm the usefulness of the proposed performance metrics for ordinal classification.

Bibliography

[1] Shannon Anjelica Allport et al. "Parental age of onset of cardiovascular disease as a predictor for offspring age of onset of cardiovascular disease". In: *PloS one* 11.12 (2016), e0163334.

[2] Andrew J Aschenbrenner et al. "Influence of tau PET, amyloid PET, and hippocampal volume on cognition in Alzheimer disease". In: *Neurology* 91.9 (2018), e859–e866.

[3] Folefac D Atem, Roland A Matsouaka, and Vincent E Zimmern. "Cox regression model with randomly censored covariates". In: *Biometrical Journal* 61.4 (2019), pp. 1020–1032.

[4] Folefac D Atem et al. "Linear regression with a randomly censored covariate: application to an Alzheimer's study". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66.2 (2017), pp. 313–328.

[5] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. "Evaluation measures for ordinal regression". In: *2009 Ninth international conference on intelligent systems design and applications*. IEEE. 2009, pp. 283–287.

[6] Stuart G Baker. "Identifying combinations of cancer markers for further study as triggers of early intervention". In: *Biometrics* 56.4 (2000), pp. 1082–1087.

[7] Stuart G Baker. "The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer". In: *Journal of the National Cancer Institute* 95.7 (2003), pp. 511–515.

[8] Donald Bamber. "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph". In: *Journal of mathematical psychology* 12.4 (1975), pp. 387–415.

[9] Nicola Barile and Ad Feelders. "Nonparametric monotone classification with MOCA". In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE. 2008, pp. 731–736.

[10] Richard E Barlow and Hugh D Brunk. "The isotonic regression problem and its dual". In: *Journal of the American Statistical Association* 67.337 (1972), pp. 140–147.

[11] Peter Bartlett et al. "Boosting the margin: A new explanation for the effectiveness of voting methods". In: *The annals of statistics* 26.5 (1998), pp. 1651–1686.

[12] Colini B Begg. "Advances in statistical methodology for diagnostic medecine ni the 1980's". In: *Statistics in medicine* 10.12 (1991), pp. 1887–1895.

[13] Leo Breiman et al. *Classification and regression trees*. Routledge, 2017.

[14] Norman E Breslow. "Discussion of Professor Cox's paper". In: *J Royal Stat Soc B* 34 (1972), pp. 216–217.

[15] Glenn W Brier et al. "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[16] Tianxi Cai et al. "The sensitivity and specificity of markers for event times". In: *Biostatistics* 7.2 (2006), pp. 182–197.

[17] Jaime S Cardoso and Joaquim F Pinto Da Costa. "Learning to classify ordinal data: the data replication method." In: *Journal of Machine Learning Research* 8.7 (2007).

[18] Jaime S Cardoso and Ricardo Sousa. "Measuring the performance of ordinal classification". In: *International Journal of Pattern Recognition and Artificial Intelligence* 25.08 (2011), pp. 1173–1195.

[19] Chih-Chuan Chen and Sheng-Tun Li. "Credit rating with a monotonicity-constrained support vector machine model". In: *Expert Systems with Applications* 41.16 (2014), pp. 7235–7247.

[20] JB Copas and P Corbett. "Overestimation of the receiver operating characteristic curve for logistic regression". In: *Biometrika* 89.2 (2002), pp. 315–331.

[21] Joaquim F Pinto da Costa, Hugo Alonso, and Jaime S Cardoso. "The unimodal model for the classification of ordinal data". In: *Neural Networks* 21.1 (2008), pp. 78–91.

[22] Joaquim F Pinto da Costa, Ricardo Sousa, and Jaime S Cardoso. "An all-at-once unimodal svm approach for ordinal classification". In: *2010 Ninth International Conference on Machine Learning and Applications*. IEEE. 2010, pp. 59–64.

[23] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[24] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[25] Nandita M desouza et al. "Cervical cancer: value of an endovaginal coil magnetic resonance imaging technique in detecting small volume disease and assessing parametrial extension". In: *Gynecologic oncology* 102.1 (2006), pp. 80–85.

[26] Richard Dykstra, John Hewett, and Tim Robertson. "Nonparametric, isotonic discriminant procedures". In: *Biometrika* 86.2 (1999), pp. 429–438.

[27] Paulus Petrus Bernardus Eggermont, Vincent N LaRiccia, and VN LaRiccia. *Maximum penalized likelihood estimation*. Vol. 1. Springer, 2001.

[28] PPB Eggermont and VN LaRiccia. "Maximum likelihood estimation of smooth monotone and unimodal densities". In: *Annals of statistics* (2000), pp. 922–947.

[29] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*. John Wiley & Sons, 2011.

[30] Youyi Fong, Shuxin Yin, and Ying Huang. "Combining biomarkers linearly and nonlinearly for classification using the area under the ROC curve". In: *Statistics in medicine* 35.21 (2016), pp. 3792–3809.

[31] Eibe Frank and Mark Hall. "A simple approach to ordinal classification". In: *European conference on machine learning*. Springer. 2001, pp. 145–156.

[32] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)". In: *The annals of statistics* 28.2 (2000), pp. 337–407.

[33] Piet Groeneboom and Geurt Jongbloed. "Smooth and non-smooth estimates of a monotone hazard". In: *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*. Institute of Mathematical Statistics, 2013, pp. 174–196.

[34] David J Hand and Robert J Till. "A simple generalisation of the area under the ROC curve for multiple class classification problems". In: *Machine learning* 45.2 (2001), pp. 171–186.

[35] James A Hanley and Barbara J McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1 (1982), pp. 29–36.

[36] Frank E Harrell et al. "Evaluating the yield of medical tests". In: *Jama* 247.18 (1982), pp. 2543–2546.

[37] Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Vol. 3. Springer, 2015.

[38] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

[39] Man-Jen Hsu and Huey-Miin Hsueh. "The linear combinations of biomarkers which maximize the partial area under the ROC curves". In: *Computational Statistics* 28.2 (2013), pp. 647–666.

[40] Erich P Huang et al. "Meta-analysis of the technical performance of an imaging procedure: guidelines and statistical methodology". In: *Statistical methods in medical research* 24.1 (2015), pp. 141–174.

[41] Xin Huang, Gengsheng Qin, and Yixin Fang. "Optimal combinations of diagnostic tests based on AUC". In: *Biometrics* 67.2 (2011), pp. 568–576.

[42] Rose Ann Huynh and Chandra Mohan. "Alzheimer's disease: biomarkers in the genome, blood, and cerebrospinal fluid". In: *Frontiers in neurology* 8 (2017), p. 102.

[43] Le Kang, Aiyi Liu, and Lili Tian. "Linear combination methods to improve diagnostic/prognostic accuracy on future observations". In: *Statistical methods in medical research* 25.4 (2016), pp. 1359–1380.

[44] Larry G Kessler et al. "The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions". In: *Statistical methods in medical research* 24.1 (2015), pp. 9–26.

[45] Osamu Komori. "A boosting method for maximization of the area under the ROC curve". In: *Annals of the Institute of Statistical Mathematics* 63.5 (2011), pp. 961–979.

[46] Osamu Komori and Shinto Eguchi. "A boosting method for maximizing the partial area under the ROC curve". In: *BMC bioinformatics* 11.1 (2010), pp. 1–17.

[47] Selcuk Korkmaz, Dinçer Göksülük, and GÖKMEN Zararsiz. "MVN: An R package for assessing multivariate normality". In: *R JOURNAL* 6.2 (2014).

[48] Sotiris Kotsiantis and Dimitris Kanellopoulos. "Cascade generalisation for ordinal problems". In: *International Journal of Artificial Intelligence and Soft Computing* 2.1-2 (2010), pp. 46–57.

[49] JWT Lee and Da-Zhong Liu. "Induction of ordinal decision trees". In: *Proceedings. International Conference on Machine Learning and Cybernetics*. Vol. 4. IEEE. 2002, pp. 2220–2224.

[50] Ker-Chau Li and Naihua Duan. "Regression analysis under link violation". In: *The Annals of Statistics* 17.3 (1989), pp. 1009–1052.

[51] Sheng-Tun Li and Chih-Chuan Chen. "A regularized monotonic fuzzy support vector machine model for data mining with prior knowledge". In: *IEEE Transactions on Fuzzy Systems* 23.5 (2014), pp. 1713–1727.

[52] Stijn Lievens, Bernard De Baets, and Kim Cao-Van. "A probabilistic framework for the design of instance-based supervised ranking algorithms in an ordinal setting". In: *Annals of Operations Research* 163.1 (2008), pp. 115–142.

[53] Aiyi Liu, Enrique F Schisterman, and Yan Zhu. "On linear combinations of biomarkers to improve diagnostic accuracy". In: *Statistics in medicine* 24.1 (2005), pp. 37–47.

[54] Chunling Liu, Aiyi Liu, and Susan Halabi. "A min–max combination of biomarkers to improve diagnostic accuracy". In: *Statistics in medicine* 30.16 (2011), pp. 2005–2014.

[55] Hua Ma, Susan Halabi, and Aiyi Liu. "On the use of min-max combination of biomarkers to maximize the partial area under the ROC curve". In: *Journal of probability and statistics* 2019 (2019).

[56] Shuangge Ma and Jian Huang. "Combining multiple markers for classification using ROC". In: *Biometrics* 63.3 (2007), pp. 751–757.

[57] Jacqueline E Maye et al. "Maternal dementia age at onset in relation to amyloid burden in non-demented elderly offspring". In: *Neurobiology of aging* 40 (2016), pp. 61–67.

[58] Peter McCullagh. "Regression models for ordinal data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2 (1980), pp. 109–127.

[59] Martin W McIntosh and Margaret Sullivan Pepe. "Combining several screening tests: optimality of the risk score". In: *Biometrics* 58.3 (2002), pp. 657–664.

[60] Jerzy Neyman and Egon Sharpe Pearson. “IX. On the problem of the most efficient tests of statistical hypotheses”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[61] Nancy A Obuchowski et al. “Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons”. In: *Statistical methods in medical research* 24.1 (2015), pp. 68–106.

[62] Nancy A Obuchowski et al. “Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example”. In: *Statistical methods in medical research* 24.1 (2015), pp. 107–140.

[63] Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA, 2003.

[64] Margaret Sullivan Pepe, Tianxi Cai, and Gary Longton. “Combining predictors for classification using the area under the receiver operating characteristic curve”. In: *Biometrics* 62.1 (2006), pp. 221–229.

[65] Margaret Sullivan Pepe and Mary Lou Thompson. “Combining diagnostic test results to increase accuracy”. In: *Biostatistics* 1.2 (2000), pp. 123–140.

[66] Margaret Sullivan Pepe et al. “Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker”. In: *American journal of epidemiology* 159.9 (2004), pp. 882–890.

[67] Margaret Sullivan Pepe et al. “Phases of biomarker development for early detection of cancer”. In: *Journal of the National Cancer Institute* 93.14 (2001), pp. 1054–1061.

[68] Ross L Prentice and Ronald Pyke. “Logistic disease incidence models and case-control studies”. In: *Biometrika* 66.3 (1979), pp. 403–411.

[69] David L Raunig et al. “Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment”. In: *Statistical methods in medical research* 24.1 (2015), pp. 27–67.

[70] Robert J Richards, James K Hammitt, and Joel Tsevat. "Finding the optimal multiple-test strategy using a method analogous to logistic regression: the diagnosis of hepatolenticular degeneration (Wilson's disease)". In: *Medical decision making* 16.4 (1996), pp. 367–375.

[71] Brian D Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007.

[72] Ingo Ruczinski, Charles Kooperberg, and Michael LeBlanc. "Logic regression". In: *Journal of Computational and graphical Statistics* 12.3 (2003), pp. 475–511.

[73] Amnon Shashua, Anat Levin, et al. "Ranking with large margin principle: Two approaches". In: *Advances in neural information processing systems* (2003), pp. 961–968.

[74] W Patrick Soutter et al. "Pretreatment tumour volume measurement on high-resolution magnetic resonance imaging as a predictor of survival in cervical cancer". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 111.7 (2004), pp. 741–747.

[75] John Q Su and Jun S Liu. "Linear combinations of multiple diagnostic markers". In: *Journal of the American Statistical Association* 88.424 (1993), pp. 1350–1355.

[76] Daniel C Sullivan et al. "Metrology standards for quantitative imaging biomarkers". In: *Radiology* 277.3 (2015), pp. 813–825.

[77] John V Tsimikas, Leonidas E Bantis, and Stelios D Georgiou. "Inference in generalized linear regression models with a censored covariate". In: *Computational Statistics & Data Analysis* 56.6 (2012), pp. 1854–1868.

[78] Gerhard Tutz. "Generalized semiparametrically structured ordinal models". In: *Biometrics* 59.2 (2003), pp. 263–273.

[79] Aad Van Der Vaart and Mark Van Der Laan. "Smooth estimation of a monotone density". In: *Statistics: A Journal of Theoretical and Applied Statistics* 37.3 (2003), pp. 189–203.

[80] Sophie Vanbelle and Adelin Albert. “A note on the linearly weighted kappa coefficient for ordinal scales”. In: *Statistical Methodology* 6.2 (2009), pp. 157–163.

[81] Andrei G Vlassenko, Tammie LS Benzinger, and John C Morris. “PET amyloid-beta imaging in preclinical Alzheimer’s disease”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1822.3 (2012), pp. 370–379.

[82] Willem Waegeman, Bernard De Baets, and Luc Boullart. “A comparison of different ROC measures for ordinal regression”. In: *Proceedings of the CML 2006 workshop on ROC Analysis in Machine Learning*. Citeseer. 2006.

[83] Strother H Walker and David B Duncan. “Estimation of the probability of an event as a function of several independent variables”. In: *Biometrika* 54.1-2 (1967), pp. 167–179.

[84] William H Wolberg and Olvi L Mangasarian. “Multisurface method of pattern separation for medical diagnosis applied to breast cytology.” In: *Proceedings of the national academy of sciences* 87.23 (1990), pp. 9193–9196.

[85] Jingjing Yin and Lili Tian. “Optimal linear combinations of multiple diagnostic biomarkers based on Youden index”. In: *Statistics in medicine* 33.8 (2014), pp. 1426–1440.

[86] Wenbao Yu and Taesung Park. “Two simple algorithms on linear combination of multiple biomarkers to maximize partial area under the ROC curve”. In: *Computational Statistics & Data Analysis* 88 (2015), pp. 15–27.

[87] XH Zhou et al. “Variable selection using the optimal ROC curve: An application to a traditional Chinese medicine study on osteoporosis disease”. In: *Statistics in Medicine* 31.7 (2012), pp. 628–635.

[88] Xiao-Hua Zhou, Donna K McClish, and Nancy A Obuchowski. *Statistical methods in diagnostic medicine*. John Wiley & Sons, 2009.