KU LEUVEN

# Estimation of ROC curve in the presence of Measurement Error

**Author:**
YOUHEE KIL

Thesis submitted for the degree of
Master of Science in
Statistics and Data Science

**Thesis supervisor:**
Dr. INGRID VAN KEILEGOM

**Mentor:**
ELIF AKÇA

Academic year 2020-2021

# *Preface*

This thesis concludes my Master of Statistics and Data Science at the University of Leuven (KU Leuven). This master thesis kept me busy for the last year of the program and half of the last year was full of frustration. The research gave me the opportunity to learn how to work on a challenging statistical modeling problem. My research question was guided and formulated together with my supervisor, Dr. Ingrid Van Keilegom. The research was complicated to me, but conducting an extensive investigation has allowed me to answer the question that we identified.

I owe my deepest gratitude to my supervisor Dr. Ingrid Van Keilegom and co-supervisor Elif Akça. Mercifully, both of them were always available, willing to answer my questions, and support me with warm-hearted words. This thesis would not have been completed without their encouragement, support, and their guidance during this process. I also wish to thank my family, friends and all of the respondents, without whose cooperation I would not have been able to conduct this research and complete my master thesis.

I hope you enjoy your reading.

Youhee Kil
August 18, 2021

# *Abstract*

**Estimation of ROC curves in the presence of measurement errors**

by YOUHEE KIL

The Receiver operating characteristic (ROC) curve is one of the most popular statistical tool for describing the accuracy of a diagnostic test which is graphical representation of false positive against true positive rates. This document lays out methodologies and application for the estimation of the ROC curve in the presence of measurement errors. We adopt maximum likelihood estimator and Bernstein polynomial model for the estimation of the contaminated nonparametric density. The result of this deconvolution density estimation using parameters interest of the proposed model is extended to use for the estimation of the ROC curve in the presence of measurement error. The EM algorithm is mainly used to obtain the maximum approximate Bernstein likelihood estimates. Simulation studies are conducted to show the performance of the deconvolution density estimation and compare our proposed estimator with Empirical estimator of the ROC curve. Finally, the proposed method is illustrated by real data example.

**Keyword:** ROC curves, Measurement error model, Bernstein polynomial model, Density Estimation, Beta Mixture model

# List of Abbreviations

| | |
|---|---|
| **ROC** | **R**eceiver **O**perating **C**haracteristic |
| **PTP** | **P**robability of a **T**rue **P**ositive |
| **PFP** | **P**robability of a **F**alse **P**ositive |
| **AUC** | **A**rea **U**nder the **C**urve |
| **ML** | **M**aximum **L**ikelihood |
| **MBLE** | **M**aximum **B**ernstein **L**ikelihood **E**stimation |
| **BIC** | **B**ayesian **I**nformation **C**riterion |

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Receiving operating characteristic (ROC) curve is the most popular graphical tool for evaluating the accuracy of a diagnostic examination for diagnostic markers to classify individuals into one of two groups. It provides a visual description of classifier performance by graphically representing the trade-off between the probability of a true positive (PTP) and the probability of a false positive (PFP). The ability of diagnostic procedure estimation is measured by the performance of sorting out observations accurately into each group. The Area Under the Curve (AUC) denoted as $\theta$ is widely used as distinctively as ROC curves as a summary measure of diagnostic accuracy, therefore AUC ($\theta$) can measure globally how well the separator variable distinguishes between cases and control.

## 1.1 Context

The estimators of ROC curves have been extensively researched and developed from parametrical, nonparametrical, semiparametrical, and Bayesian statistical approaches. The bi-normal model is the most widely considered parametric method in combination with Box-Cox transformation when both of two categories such as results of the diseased and non-diseased test follow normal distributions (Faraggi and Reiser, 2002). The other models, for example, bi-gamma, bi-beta, bi-logistic, bi-exponential, bi-lognormal, and bi-Rayleigh, are available as an alternate choice of parametric models (Gonçalves et al., 2014). The maximum likelihood method on a bi-normal estimator of the ROC curve was introduced by Cai and Moskowitz, 2004. The parametric estimator of the ROC curve is still in an active field of research. In addition, several researchers have proposed nonparametric approaches to obtain smoothed ROC curves. Among those, the empirical method and the kernel method are the most commonly used and the simplest nonparametric estimators of the ROC curve (Hsieh and Turnbull, 1996; Lloyd, 1998; Lloyd

and Yong, 1999). Even though the empirical estimator of the ROC curve uniformly converges to the theoretical curve, the estimator has a significant drawback such as large variability with small sample sizes. The Kernel estimators of the ROC curve were suggested and developed to overcome the lack of smoothness of the empirical estimator of ROC curves (Zou, Hall, and Shapiro, 1997; Lloyd, 1998; Lloyd and Yong, 1999). The selection of the 'optimal' bandwidth is the most complex aspect of the kernel estimation of the ROC curve, hence, many researchers have proposed and improved the optimal bandwidth selection choices methods.

In this paper, the estimation of the ROC curve in the presence of the measurement error is the main concern. Coffin and Sukhatme, 1996 stated that medical literature comprehensively reported that diagnostic indicators might be prone to be affected by errors of measurement. The measurements are highly susceptible to measurement error for example the measuring can be attributed to either the laboratory equipment or the technician who uses it (Krzanowski and Hand, 2009). As Carroll et al., 2006 stated failing to take into account the measurement error will lead to serious consequences such as bias in the estimators of ROC curves and AUC in nonparametric cases and even in parametric cases. Hence, the measurement error has to be considered to derive well-grounded inferences by some bias-correction methods. This topic has generated some recent interest (Coffin and Sukhatme, 1996; Coffin and Sukhatme, 1997; Faraggi, 2000; Reiser, 2000; Schisterman et al., 2001; Tosteson et al., 2005; Vexler, Schisterman, and Liu, 2008) with focus on inferences for ROC curves and AUC with the assumption of measurement errors and the impact of measurement error on ROC curves and AUC.

Initially, bias-corrected estimators for normal and exponential models and non-parametric models were introduced when AUC is estimated (Coffin and Sukhatme, 1996; Coffin and Sukhatme, 1997). The effect of random measurement error on the confidence interval for AUC in parametric normal models has been considered over the last few decades. To cite an example, Faraggi, 2000 introduced an adjusted confidence interval when a parametric normal model was assumed. A simulation was performed based on the adjusted confidence interval to show the effect of not taking measurement error into account. Schisterman et al., 2001 and Reiser, 2000 discussed and developed adjusting confidence intervals for the AUC taking measurement error into account from the external experiment and internal experimental study, respectively.

However, those adjusted confidence intervals have some limitations in

implementation. To cite an example, it will be expensive to carry out the analyses and eliminating the impacts of the measurement errors at the same time in a study with large sample size. To overcome this limitation, an external experiment with a small-sized sample is usually conducted initially to handle measurement errors. Unfortunately, future measuring on this marker will still be contaminated. Alternatively, as one of the means to reduce the measurement error, taking a number of repeated measurements and using their average to represent the 'true' marker value can be a solution. On the other hand, Tosteson et al., 2005 developed confidence intervals and regions for specific points on the bi-normal ROC curves with an adjustment of heteroscedastic and possibly non-normal measurement errors. This method allows for inferences on the features of the ROC curves rather than the AUC and uses the individual estimates of the measurement error variances.

The aforementioned methods to correct measurement errors in AUC and ROC curves are mostly based on the normality assumption of the diagnostic biomarkers. However, there are methods in the literature for the non-normal biomarkers. For example, Rosner, Tworoger, and Qiu, 2015 extended the method of Reiser, 2000 by relaxing the normality assumption. Then,Rosner, Tworoger, and Qiu, 2015 presented a new method without requiring the normality assumption to derive approximated confidence limits for AUC curves in the presence of measurement errors.

Although the forenamed methods have contributed well to the literature, an approach to estimate ROC and AUC curves in the presence of measurement errors with weak assumptions has not been reported yet. The exact error distribution is frequently required for most of proposed ROC curves and AUC bias-correction methods, however, as Bertrand, Van Keilegom, and Legrand, 2019 mentioned it is almost impossible to carry out the exact error distribution (variance of measurement error) when neither validation nor auxiliary data are available due to complexity. In this paper, we propose a smooth nonparametric ROC curve derived from Bernstein type polynomial estimates to obtain the ROC curves and AUC. Bernstein polynomials are able to handle the drawbacks of non-parametric ROC curves such as slow convergence rate. Wang et al., 2019 proposed a non-parametric ROC curve Bernstein polynomial estimator without considering the presence of measurement errors. In that paper, asymptotic properties are also examined with the aim of improving the choice of the optimal tuning parameter. Following that, Wang and Cai, 2021 enhanced that estimator by providing an inherent bandwidth. In this case, the same asymptotic properties hold.

In this thesis, the main focus is on taking those estimators one step further. In other words, a ROC curve maximum approximate Bernstein likelihood estimator is proposed when measurement errors are present. To this end, the main references are Guan, 2016, Bertrand, Van Keilegom, and Legrand, 2019 and Guan, 2021. Therefore, under weak distributional assumptions, ROC and AUC curves are obtained based on a maximum approximate Bernstein likelihood estimator.

## 1.2 Organization

In **Chapter 2**, we described the basics of the ROC curves and measurement error problems. In **Chapter 3**, the Bernstein polynomial model for the non-parametric density estimation is preliminarily delineated. The approach of obtaining the maximum Bernstein likelihood estimates for the estimation of ROC curves in the presence of measurement errors is introduced. Moreover, a method of selecting maximum approximate likelihood estimates and optimal model degrees are also illustrated. The simulation results of the estimation of ROC curve with measurement errors are reported in **Chapter 4**. In **Chapter 5**, the proposed method is performed by analyzing real data sets. We conclude this paper with some remarks in **Chapter 6**.

# Chapter 2

# Basic Methods

## 2.1 Receiver Operating Characteristics (ROC) curve

The *Receiver Operating Characteristic* (ROC) curve is an important tool when the outcome is binary for the evaluation, and the comparison of predictive models (Liao, Wu, and Yu, 2017). The estimation methods of the ROC Curve have been constantly developed, adjusted, and extended.

The emergence of the ROC curve was during World War II for identifying enemy objects on battlefields and it was developed to enhance early radar signals to detect bombers (Collinson, 1998). The use of the ROC curve was expanded to many different fields such as psychology for perceptual detection of stimuli, finance, geosciences, and sociology, economics, environmental science and so on (Hanley and McNeil, 1982; Collinson, 1998; Baulch, 2002; Gonçalves et al., 2014). Machine learning and Data mining are recent and emerging fields for the ROC curve and AUC. Many different domains have committed to improving the ROC curve. The usefulness of estimating the ROC curve is differed by purpose, the focused intention of the ROC curve in this paper is an ability of a continuous biomarker score to distinguish two-group correctly. As many different estimators of the ROC curve have been proposed, the definition and estimation methods of the ROC curve will be described in this chapter.

### 2.1.1 Definitions and modeling framework of the ROC Curve

Diagnostic tests play a crucial role in medical field because their ability of diagnostic procedure estimation is assessed by the performance of classifying patients accurately into each group those with diseased or non-diseased. The diagnostic test accuracy is expressed in terms of sensitivity and specificity. The definition of sensitivity is the ability to detect the diseased population,

| | New test result positive | New test result negative |
|---|---|---|
| Diseased | True positive (TP) | False negative (FN) |
| Non-diseased | False positive (FP) | True negative (TN) |

TABLE 2.1: Standard 2X2 Contingency table used for calculating sensitivity and specificity

on the other hand, the definition of specificity is the ability to exclude correctly the non-diseased population (Collinson, 1998). There will be a trade off without an exception when calculating sensitivity and specificity. As sensitivity increases, specificity will decrease, and vice versa. Therefore, the level of threshold point for the test must be critically selected, not only to balance sensitivity and specificity but also to take into account the underlying bias imposed by the studied population (Collinson, 1998). The technique of ROC curves can address the diagnostic ability of a binary classifier system, simply, the ROC curve illustrates the trade-off between sensitivity (or TPR) and specificity (or 1-FPR). In short, the representation of ROC curve (the estimation target of interest) shows graphically the relationship between TPR and FPR or a plot of sensitivity against specificity at various threshold settings as shown in Figure 2.1.

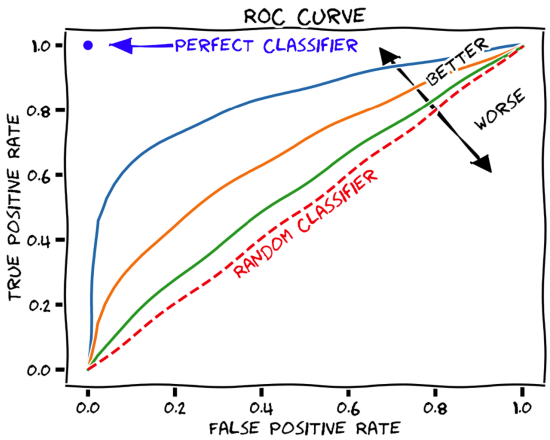



FIGURE 2.1: The representation of ROC curves. The original image is from MartinThoma, 2020

The ROC curve is needed to provide an evaluation of the classifier over the whole range of value $t$ in a particular classification rule rather at just a single chosen one where $t$ is the classifier threshold value, hence a direct evaluation of the test power is allowed (Krzanowski and Hand, 2009). The ROC

curve must lie within the border of $[(0,0),(0,1)]$ and $[(0,1),(1,1)]$. Therefore, in practice, the ROC curve will be a continuous curve lying in the upper triangle of the graph such as between two extremes points of the graph (Krzanowski and Hand, 2009). The nearer the ROC curve is to the top left-hand corner of the graph, the better the test which indicates it's closer to a perfect classifier. On the other hand, a $45^\circ$ line indicates its exact to a random classifier which is a useless test (Collinson, 1998).

Let the test variable $X$ denote the continuous scores, $F_1$ and $F_0$ be distribution functions of $X$ for the diseased group (1) and non-diseased group (0). For the cut point (threshold value) $t$ where the test result is positive if it is greater than $t$ and negative otherwise, Sensitivity (TPR) denoted as $SE(t) = P(X > t|group = diseased(1)) = 1 - F_1(t)$ while specificity$(s)$ is denoted as $SP(t) = P(X < t|group = nondiseased(0)) = F_0(t)$, and false positive rate is denoted as $1 - SP(t)$. The ROC curve is a plot of the sensitivity against 1-specificity $(1 - s)$, in other words, the true positive fraction (TPF) against false positive fraction (FPF) using a threshold $t$. The equation of the ROC curve is

$$\begin{aligned} ROC(\cdot) &= (FPF(t), TPF(t)) \qquad t \in (-\inf, \inf) \\ ROC(s) &= 1 - F_1[F_0^{-1}(1-s)] \qquad (0 \leq s \leq 1) \end{aligned} \tag{2.1}$$

There are two prevailing ROC Curve summary metrics of the discriminatory accuracy of a test, the area under the curve (AUC) and the Youdan index $(max_t se(t) + sp(t) - 1)$ (Gonçalves et al., 2014). The equation of AUC is

$$AUC = \int_0^1 ROC(u)du \tag{2.2}$$

where AUC's closer value to 1 indicates the high diagnostic accuracy of the test.

### 2.1.2 Estimation methods of the ROC Curve

The straightforward approach to estimate the ROC Curve is empirical approach where the data modeled samples from the relevant population without assumptions (Krzanowski and Hand, 2009). Then the ROC curve is estimated based on the data sampled from sensitivity ($SE(t) = 1 - F_1(t)$), specificity ($SP(t) = F_0(t)$). However, the result of the empirical estimator of the ROC curve is seemingly jagged and not accurate enough for the smaller sample sizes. Therefore, many other methods have been explored to estimate

the ROC curve by researchers with different approaches. The proposed approaches for the estimation of the ROC curve are a fully parametric approach, a fully nonparametric approach, and a semiparametric approach. This section will explore estimators of the ROC curve of preambles.

**Parametric Approach**

The parametric estimation approach for the ROC curve remained a crucial area of research. But, it can be used only when the population of the true $F_0$ and $F_1$ in non-diseased and diseased groups are known. Faraggi and Reiser, 2002 stated that the most common parametric approach is the bi-normal model, which is used when the assumption of the marker values for both diseased and non-diseased populations follow a normal distribution. The parametric estimator of the ROC curve can be denoted in the following way,

$$ROC(s) = \Phi[(\mu_1 - \mu_0)/\sigma_1 + (\sigma_0/\sigma_1)\Phi^{-1}(s)] \tag{2.3}$$

where $\Phi$ are standard normal distribution function with mean values $\mu_0, \mu_1$ and variances $\sigma_0^2, \sigma_1^2$. The bi-normal model allows handy maximum likelihood estimates of the ROC curve parameters, but many researchers researched to develop and propose alternative models since problems of using bi-normal were noted. Mainly, some authors insisted that the binormal estimator of ROC curve is robust regarding model mis-specification deriving from a different distribution.

**Nonparametric Approach**

The empirical estimator is the simplest nonparametric method where there is no assumption. The equation of the empirical estimate of the ROC is given by

$$R\tilde{O}C(s) = 1 - \tilde{F}_1[{\tilde{F}_0}^{-1}(1 - s)] \tag{2.4}$$

where $\tilde{F}_1$ and ${\tilde{F}_0}^{-1}$ are denoted as the empirical distribution and the empirical quantile function representing diseased and non-diseased groups, respectively. Gonçalves et al., 2014 described empirical distribution as the percentage of sample points smaller or equal to $s$, for any given value $s$. The empirical estimation has a significant drawback in that small sample sizes suffer from large variability. Due to this, it is not an adequate method for

medical research since small samples are not inevitable in clinical research. Another nonparametric estimator model of ROC curves was introduced by Zou, Hall, and Shapiro, 1997 based on kernel density methods to overcome the drawback of the empirical estimator, the lack of smoothness. Basic Kernel estimation of the density function is given by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where K(·) is the kernel function in population and $h$ is the bandwidth. The random sample drawn from $f$ is denoted as $X_1, X_2, ..., X_n$ in here.

Kernel density estimators for non-diseased and diseased population, $\hat{f}_0$ and $\hat{f}_1$, are given

$$\begin{aligned} \widehat{f_0}(x_0) &= \frac{1}{nh_0} \sum_{i=1}^{n} K_0\left(\frac{x_0 - x_{0i}}{h_0}\right) \\ \widehat{f_1}(x_1) &= \frac{1}{nh_1} \sum_{i=1}^{m} K_1\left(\frac{x_1 - x_{1i}}{h_1}\right) \end{aligned} \tag{2.5}$$

where $(x_{01}, ..., x_{0n})$ and $(x_{11}, ..., x_{1n})$ be two independent samples from $X_0$ and $X_1$. The amount of smoothness is controlled by using $h_i$ $(i = 0, 1)$. The cumulative distribution function can be estimated as

$$\begin{aligned} \widehat{F_0}(x) &= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{x} \frac{1}{h_0} K_0\left(\frac{u - x_i}{h_0}\right) du \\ \widehat{F_1}(y) &= \frac{1}{m} \sum_{i=1}^{m} \int_{-\infty}^{x} \frac{1}{h_1} K_1\left(\frac{v - y_i}{h_1}\right) dv \end{aligned} \tag{2.6}$$

Then, the Kernel-based ROC curve estimator can be represented as

$$\widehat{ROC}_K(s) = 1 - \widehat{F}_1[\widehat{F}_0^{-1}(1 - s)] \tag{2.7}$$

where the equation 2.6 was plugged into the equation 2.1. As mentioned above, the selection of optimal choice is the most complex facet of the kernel based estimator of the ROC curve. Researchers proposed and developed kernel based estimators to obtain a smooth ROC curve with an improved way of choosing optimal bandwidths (Zou, Hall, and Shapiro, 1997; Lloyd, 1998; Lloyd and Yong, 1999). Unfortunately, according to Peng and Zhou, 2004, the

resulting estimators from those standard kernel estimations methods have some drawbacks in the case where the support of the density function $f$ to be estimated is compact support.

In a recent study, Wang et al., 2019 applied Bernstein polynomial to the estimation of the ROC curve. Nonparametric estimation of the ROC curve based on the Bernstein polynomial performs more accurately than the empirical ROC estimator and the kernel based estimator. The Bernstein polynomial of the positive model degree $m$ for the ROC curve is defined as

$$B_m(t) = \sum_{k=0}^{m} ROC(\frac{k}{m}) P_{k,m}(t)$$

where $P_{k,m}(t) = C_m^k t^k (1-t)^{m-k}$. Since $ROC(t)$ is continuous within compact support,

$$\lim_{m \to +\infty} B_m(t) = ROC(t)$$

proposed by Lorentz, 2013. Therefore, Nonparametric estimation of the ROC curve based on the Bernstein polynomial can be obtained by

$$ROC_m(t) = \sum_{k=0}^{m} ROC_e(\frac{k}{m}) P_{k,m}(t)$$

where $m$ is positive order of Bernstein estimator and $ROC_e$ is the empirical ROC estimator. More details about Bernstein Polynomial will be described in the next chapter.

**Semi-parametric Approach**

Researchers used specific nonlinear logistic regression models to estimate ROC curves (Lloyd, 2002a; Lloyd, 2002b). The logistic regression model for a given test and defined status is

$$P(D = 1|X = x) = \frac{\exp[\alpha^* + \beta^T \gamma(x)]}{1 + \exp[\alpha^* + \beta^T \gamma(x)]}$$

where $\alpha^*$ is a scalar parameter, $\beta$ is a vector parameter, and $\gamma(x)$ is smooth vector function of $x$. Qin and Zhang, 1997 proved that the logistic regression model is equivalent to a two-sample semiparametric density ratio model where two density functions $f(x)$ and $g(x)$ are independent, unknown, and linked by an exponential tilt ($\exp[\alpha + \beta^T \gamma(x)]$). Furthermore, the two-sample semiparametric density ratio model denoted as $f(x) = \exp[\alpha + \beta^T \gamma(x)] g(x)$ was considered for estimating the ROC curve (Qin and Zhang, 2003). Then

the semiparametric density ratio model estimator of the ROC curve and AUC can be stated as

$$\widehat{ROC}_{dr}(s) = \int_0^s \exp[\alpha + \beta^T \gamma(x)(G^{(1-t)})]dt \tag{2.8}$$

Zhang, 2006 proposed a semiparametric approach to testing the null hypothesis based on the AUC to test whether a diagnostic test is capable of discriminating between diseased and non-diseased groups.

## 2.2 Measurement Error

The classical measurement error is one of the most commonly used measurement errors where the observed variables are measured with an additive error. The classical measurement error model states that $W_{ij} = X_i + U_{ij}$ where $W_{ij}$ is an unbiased measure of $X_i$, and $U_{ij}$ is mean-zero error which could be homoscedastic or heteroscedatic. This classical measurement error model can be interpreted as the observed value ($W_{ij}$) equals the true value ($X_i$) plus classical measurement error ($U_{ij}$).

### 2.2.1 What is Measurement Error

There are many reasons to have measurement errors in statistical analysis. It occurs whenever the exact variables are not able to be observed. The most common measurement errors are instrument error and sampling error (Buonaccorsi, 2010). There is an increasing awareness that measurement error must be taken into account for accurate statistical analysis, especially with epidemiological data. Buonaccorsi, 2010 exemplified the binary outcome variable of different disease type tests is frequently assessed through an imperfect diagnostic procedure, such as a blood test or an imaging technique, which can lead to either false positives or false negatives. And, in this case, the ROC curve is the most effective and common classification technique, simply the ROC curve depends on the distribution of the variable of each group or means or variance if each group is under the normal distribution (Buonaccorsi, 2010). But, as Tosteson et al., 2005 described, the induced measurement error from sampling or imaging technique where commonly occurred in epidemiology will distort the estimated ROC curves. Tosteson et al., 2005 stated, "in general, the introduction of measurement error will

reduce the estimated diagnostic accuracy of test based on error-free data and bias estimates of the AUC"( Tosteson et al., 2005)

The measurement error can occur due to many reasons and it could provoke critical errors in the statistical analysis as discussed. The following subsection is organized to discuss, (i) the notation and models for the true (unobserved) values of the measurement error, (ii) the consequences of ignoring the measurement error.

Historically, there are two major defining characteristics of the taxonomy of measurement error. Prevalent measurement error literature is based on *classical measurement error* where the truth (unobserved) is measured with additive error. Therefore, the classical measurement error model is that we make an assumption about the observed values' distribution given the true values or vice versa (Tosteson et al., 2005; Buonaccorsi, 2010). In the classical measurement error model, the true (unobserved) values for covariate are denoted as $X$. The error values are denoted as $U$, and its mean-zero could be homoscedastic or heteroscedastic. Therefore, the classical measurement error model is :

$$W = X + U \tag{2.9}$$

where $W$ is the observed (mismeasured) covariate, besides, $X$ and $U$ are assumed to be independent of each other. We further assume that the classical measurement error model with a Gaussian error of unknown variance $\tau^2$, so that $U \sim N(0, \tau^2)$.

On the other hand, the Berkson error model is formulated as

$$X = W + U \tag{2.10}$$

where $X$, $W$ and $U$ indicate true value, the estimated value, and measurement error, furthermore, it satisfies $E(U_i|W_i) = 0$. One of the highlighted differences between Equation 2.9 and Equation 2.10 is that, unlike the classical measurement error, true value ($X$) in the Berkson Error has more variability than the estimated value ($W$). It is crucial to understand these elusive differences between the Berkson and Classical measurement errors. In the following subsections, classical measurement error will be mainly described.

### 2.2.2 Effects of ignoring measurement error

Measuring data with errors happen frequently in many different fields, such as medicine, bioinformatics, chemistry, astronomy, and econometric field. Wang and Wang, 2011 described real examples of measuring data with error and briefly introduced a solution to the listed example problem. The effect of ignoring measurement error is ranged from nothingness to dramatic. Measurement errors on analyses were often forgotten or ignored, first and foremost, the majority of researchers do not take account of measurement error, even if they are aware of the presence of measurement error and its possible implications.

The effect of ignoring measurement error is caused by three reasons, (i) disguising the crucial characteristics of the data which eventually confuses graphical model analysis, (ii) loss power of relationship detection among variables, (iii) produce bias in the estimation of parameters (Wang and Wang, 2011).

The effect of correcting the bias in measurement error problems can be observed in the following simple example. When we have a sample of $w$, we assume $w$ is observed with additive noise and $x$ is unobserved. In Nonparametric Estimation of the density of random variable $X$, no assumptions are made on $w$ or $x$ but only on noise ($u$). For example, a density of random variable $X$ is assumed to follow Weibull distribution with a shape of 5 and a scale of 1 between a range of [0, 2], and it is displayed as true density in Figure 2.2.

In order to show the effect of correcting the bias in measurement error problems, two different R packages, decon (Wang and Wang, 2011) and deamer (Stirnemann et al., 2012), are performed for deconvolution estimation by adding classical additive measurement error assuming to follow N(0, 0.2) into the dataset of the example. Figure 2.2 shows the difference between true density and measurement error corrected density. It is visually certified that if the researcher ignored the measurement error, in this case, the shown difference in Figure 2.2 might affect the accurate statistical analyses and inferences. Therefore, the variation may explain the effects of ignoring measurement error which can be impactful.

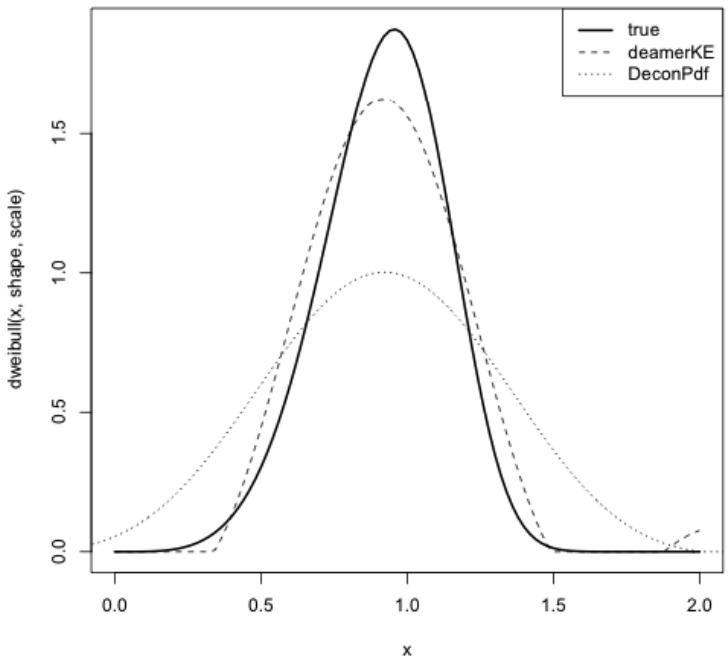



FIGURE 2.2: Effects of Measurement Error

### 2.2.3 Measurement Error Estimation

Statistical models of measurement error problems can be addressed from the perspectives of parametric and nonparametric. Our attention is on nonparametric measurement error models and their estimation. Carroll et al., 2006 explained two kinds of nonparametric measurement error models. The first kind of nonparametric measurement error model is the additive measurement error model, and the second one is known as regression with errors-in-variables. The (*additive measurement error model*) where the problem of estimating $f_X$ is known as the deconvolution problem was mainly applied in this paper. The deconvolution kernel density estimator employing the kernel idea and the Fourier inverse was proposed as a method for correcting the effect of measurement error in the additive measurement error model to recover the unknown density ($f_X$) from contaminated data ($W$) (Carroll and Hall, 1988; Stefanski and Carroll, 1990). Extensive research about the deconvolution kernel approach was studied after then. In this paper, a maximum approximate likelihood method for deconvoluting a density in additive measurement error models was used by using the approximate Bernstein polynomial model which was proposed by Guan, 2021. Guan, 2021 proved that the proposed estimator performs better than the deconvolution kernel density

estimator in a small sample.

## 2.3 Estimation of ROC curves in the presence of measurement errors

The maximum approximate likelihood method for deconvoluting a density in additive measurement error models will be applied to estimate the ROC curve in the presence of measurement errors by using the approximate Bernstein polynomial model. This approach was built on the work of Schwarz and Van Bellegem, 2010, Bertrand, Van Keilegom, and Legrand, 2019, Guan, 2016, and Guan, 2021. We firstly review several results on Bernstein distribution function estimators and measurement error, then present the precise definition of the maximum Bernstein likelihood density estimation of the measurement error of the ROC curve. Following studies examined asymptotic properties and weak convergence of the Bernstein polynomial estimator. Babu, Canty, and Chaubey, 2002 implemented Bernstein polynomial estimation to the univariate distribution and density function, later Leblanc, 2010 showed the Bernstein estimator outperforms the classical empirical distribution regarding the asymptotic variance by proving point-wise asymptotic normality. Guan, 2016 proposed maximum Bernstein likelihood density estimation method. This Bernstein polynomial method relying on only one regularization parameter was possible because it is the same as using a mixture of Beta density function with known parameters. Besides, it approximates any continuous function with compact support. Guan, 2016 mentioned the flexible approximate parametric estimation is favored over nonparametric estimation due to a lower rate of convergence to the true density function. A likelihood-based approach for estimating error variance with only the weak assumption was derived by Schwarz and Van Bellegem, 2010. Bertrand, Van Keilegom, and Legrand, 2019 introduced a flexible parametric approach for classical measurement error variance estimation when auxiliary data are not available. To prove the identifiability which is a model property that must fulfill for precise inference, they followed the work of Schwarz and Van Bellegem, 2010.

The classical measurement error model is considered with a Gaussian error of unknown variance $\tau^2$, $U \sim N(0, \tau^2)$, from equation 2.9 where X is only assumed to be continuous and to have compact support. The objective

is to obtain an estimate of $\tau$ which will then be available for use in a bias-correction method. Therefore, the key idea of this is to build the probability density function of W with only weak assumptions on the distribution of X. Therefore, we can assume that

$$X = \alpha S + \beta \tag{2.11}$$

where $\alpha$ and $\beta$ are two unknown constants with $\alpha > 0$ for identifiability reasons, and $S$ is a continuous random variable taking values in $[0, 1]$. Let $f_X, f_S, f_U$ and $f_W$ be the density functions of X, S, U, and W, respectively, under the addictive measurement error model( 2.9), besides, value of $X$ are in $[\beta, \alpha + \beta]$ according to the equation of ( 2.11). The classical additive measurement error model( 2.9 ) with normally distributed error is $\frac{1}{\tau}\phi\left(\frac{w-x}{\tau}\right)$ where $\phi(\cdot)$ is a normal distribution with zero mean and unit variance. Carroll et al., 2006 noted that the density function $f_W$ is the convolution of $f_X$ and $f_U$.

$$\begin{aligned} f_W(w) &= \frac{1}{\tau}\int_{x=\beta}^{\alpha+\beta} f_X(x)\phi\left(\frac{w-x}{\tau}\right)dx \\ &= \frac{1}{\tau}\int_{u=0}^{u=1} f_S(u)\phi\left(\frac{w-x}{\tau}\right)du \qquad \textit{where } u = \frac{x-\beta}{\alpha}, du = \frac{1}{\alpha}dx \\ &= \frac{1}{\alpha\tau}\int_{x=\beta}^{\alpha+\beta} f_S\left(\frac{x-\beta}{\alpha}\right)\phi\left(\frac{w-x}{\tau}\right)dx \end{aligned} \tag{2.12}$$

where the probability density function of S is denoted as $f_s$ and it is specified with minimal assumption.

# Chapter 3

# Methodology

The methodology for the proposed estimator of the ROC curve in the presence of measurement error will be explained in this chapter. The chapter outlined the description of the Maximum Bernstein likelihood density estimation (MBLE) on the ROC curve in the presence of measurement error. In the first section, the approximate Bernstein polynomial model is preliminarily described with the explanation of the nonparametric maximum likelihood estimator and the Bernstein polynomial Model. The proposed method, Maximum Bernstein Likelihood Estimation (MBLE), is introduced succeeding and explained based on the preliminary knowledge. The application of the MBLE on the estimation of the ROC curve in the presence of measurement error will be explained. Furthermore, in the following section, the EM algorithm approach for obtaining maximum approximate likelihood estimates and the model selection criterion that helps to find optimal model degrees are also demonstrated.

## 3.1 Maximum Bernstein Likelihood Estimation

### 3.1.1 Nonparametric maximum likelihood estimator

The nonparametric approach is getting more attention. It is a generically suitable and undoubtedly beneficial tool for the ROC curve estimation. Nevertheless, the nonparametric has some limitations, for example, the boundary effect and slow convergence rate. These limitations of the nonparametric method led us to apply a maximum Bernstein likelihood density estimation method proposed on the estimation of the ROC curve and AUC. The maximum Bernstein likelihood density estimation method is an approximate

parametric model for a nonparametric model to estimate the underlying density by using the maximum likelihood method. It is a new approach proposed by Guan, 2016 to overcome the limitations of nonparametric estimation because it can inference better on the population parameters with a better underlying population density estimator.

The empirical distribution function is often called a nonparametric maximum likelihood estimator which is confirmed by Owen, 2001. Guan, 2016 demonstrated that there is a unique $\hat{f}_m$ that maximizes likelihood $\ell(f_m) = \sum_{j=1}^{n} \log f_m(x_j)$ for each number of dimensions ($m$). Therefore, if the empirical distribution converges to $f$ as fast as number of dimension ($m$) and observation ($n$) go to infinity, then an approximate parametric density estimation ($\hat{f}_m$) converges to $f$ at a faster rate than other nonparametric density such as kernel density (Guan, 2016).

### 3.1.2 Bernstein polynomial Model

The Bernstein polynomial model is known as a very smooth estimator with acceptable behavior at the boundaries (Leblanc, 2010). As Bernstein, 1912 defined if $f$ is any continuous density function in the closed unit interval $[0, 1]$, the Bernstein polynomial of degree $m > 0$ can approximate it. The equation of the Bernstein polynomial is

$$f_m(x) = B_m f(x) = \sum_{k=0}^{m} f\left(\frac{k}{m}\right) p_{m,k}(x), \qquad 0 \leq x \leq 1 \tag{3.1}$$

where $p_{m,k}(x) = \binom{m}{k} x^k (1-x)^{(m-k)} (k = 1, 2, ..., m)$ is the Bernstein basis polynomial and $B_m$ is the Bernstein operator. Under some conditions, the Bernstein polynomial, $B_m f(x)$, converges to $f(x)$ with its the best convergence rate. As the theorem presented by Berens, Lorentz, and MacKenzie, 1972 and Lorentz, 2013, any continuous $f$ is defined on [0,1],

$$\lim_{m \to +\infty} B_m f(x) = f(x) \tag{3.2}$$

the Bernstein polynomial, $B_m f(x)$, converges to $f(x)$ with its best convergence rate under the conditions where $f$ is approximated by Bernstein polynomials. Vitale, 1975 stated form of the Bernstein polynomial estimate of $f(x)$ is one of a linear combination of beta densities with random coefficients based on the observation. In parallel, Ghosal et al., 2001 explained that as mixtures of beta densities form a very flexible model for a density on the unit interval, the class of Bernstein density is a much smaller subclass of the

beta mixtures defined by Bernstein polynomial, which can approximate any continuous density.

### 3.1.3 Approximate Bernstein polynomial model

The condition to converge with the best rate is when $f$ has bounded second or even higher-order derivatives. It produces a better approximation. Lorentz, 2013 further proved that if density $f$ has higher derivatives and positive lower bound which is passed that there are no values lower than 0 and can be approximated better. It is called a polynomial with positive coefficients denoted as $P_m^f(x) = \sum_{k=0}^{m} f(k/m)/(m+1)p_{m,k}(x)$. Therefore, Guan, 2021 defined an approximate Bernstein polynomial model as

$$f(x;\theta_m) = \sum_{k=0}^{m} \bar{\theta}_{k,m} p_{m,k}(x) \tag{3.3}$$

where $\bar{\theta}_m = (\theta_{0,m}, ..., \theta_{m,m})^T$ and meet the conditions, $\theta_i \geq 0$, and $\sum_{i=1}^{m} \theta_i = 1$. The density $f_x$ can be approximately modeled and parameterized by $f_B(x;\bar{\theta}_m)$ as a mixture of the beta distribution and estimate the $\theta_m$ as parameters using the maximum likelihood method (Guan, 2016). Maximum Bernstein Likelihood Estimation (MBLE) of $\theta_m$ is the maximizer of the set of estimated parameters of $\ell_B(\theta_m)$. Then we can obtain maximum Bernstein likelihood estimator $\widehat{f_B}(x) = f_m(x;\widehat{\theta_m})$ where $\theta_m$ are unknown parameters. Details will be discussed later with the application of the ROC curve estimation in the presence of the measurement error.

### 3.1.4 The approximation of the density function

As aforementioned in the equation 3.1 and 3.2, the Bernstein polynomial converges to $f(x)$. The Bernstein polynomial of order $m > 0$ for the $f_s$ can be defined as

$$B_m(s) = \sum_{k=0}^{m} f\left(\frac{k}{m}\right) p_{m,k}(s) \tag{3.4}$$

where $P_{m,k}(x) = \binom{m}{k} x^k (1-x)^{(m-k)}$, for $k = 0, ..., m$. The theorem presented by Lorentz, 2013, any continuous $f$ is defined on unit interval [0,1],

$$\lim_{m \to +\infty} B_m(s) = f(s) \tag{3.5}$$

One of the advantages of the Bernstein polynomial method is approximating any continuous function with compact support (Guan, 2016). The features of the Bernstein polynomial method allowed us to apply the maximum Bernstein likelihood density estimation approach to the estimation of $\tau$ and the approximation of $f_s$. The approximation of $f_s$ was introduced by Bertrand, Van Keilegom, and Legrand, 2019 and applied these on the proposed method by Guan, 2016, maximum Bernstein likelihood density estimate. The equation of the approximating $f_s$ is

$$\begin{aligned}
\tilde{f}_{s,m}(s;\bar{\theta}_m) &= \sum_{k=0}^{m} f_s\left(\frac{k}{m}\right) p_{k,m}(s) \\
&= \sum_{k=0}^{m} f_s\left(\frac{k}{m}\right) \binom{m}{k} s^k (1-s)^{(m-k)} \\
&= \sum_{k=0}^{m} \theta_{k,m} \binom{m}{k} s^k (1-s)^{(m-k)} \\
&= \sum_{k=0}^{m} f_s\left(\frac{k}{m}\right) \binom{m}{k} \frac{k!(m-k)!}{(m+1)!} s^k (1-s)^{(m-k)} \\
&= \frac{1}{m+1} \sum_{k=0}^{m} f_s\left(\frac{k}{m}\right) s^k (1-s)^{(m-k)}
\end{aligned} \tag{3.6}$$

where $\bar{\theta} = (\theta_{0,m}, ..., \theta_{m,m})$ and $s \in [0, 1]$. The approximation of $f_s$ is done by a mixture of $Beta(k+1, m-k+1)$ densities with known parameters. The probability density function of X, $f_x(\cdot; \alpha, \beta)$, from the equation (2.11) can be approximated by implementing the context of density estimation of the equation (3.6).

$$\begin{aligned}
X &= \alpha S + \beta \\
S &= \frac{X - \beta}{\alpha}
\end{aligned} \tag{3.7}$$

The support $S$ of $f$ has to be within the range between [0, 1], but if the support $S$ of $f$ is different from [0, 1], then we can find a finite interval of a sample data of size $n$ from $f$. Since $f$ is a nonparametric model which is totally unspecified and don't provide information about the support of $f$. A finite interval of a sample of size $n$ from $f$ can be the minimum and maximum order statistics, $[y_{(1)}, y_{(n)}] \in [a, b]$. Then we let $\alpha$ and $\beta$ as $\alpha = (b - a)$ and $\beta = a$. The probability density function of X can be approximated with a $beta_{\alpha,\beta}(\cdot)$ which is probability density function of a Beta-distributed random

variables by using the approximated Bernstein polynomial (3.3 based on the parameter $\alpha, \beta$ where $x \in [\beta, \alpha + \beta]$.

$$\begin{aligned}
\tilde{f}_{x,m}(x;\alpha,\beta,\bar{\theta}_m) &= \frac{1}{(\alpha+\beta)-\beta}\tilde{f}_{s,m}\left(\frac{X-\beta}{\alpha};\bar{\theta}_m\right) \\
&= \frac{1}{\alpha}\sum_{k=0}^{m}\theta_{k,m}\binom{m}{k}\left(\frac{X-\beta}{\alpha}\right)^{k}\left(1-\left(\frac{X-\beta}{\alpha}\right)\right)^{(m-k)} \\
&= \frac{1}{\alpha}\sum_{k=0}^{m}\theta_{k,m}beta_{(k+1,m-k+1)}\left(\frac{X-\beta}{\alpha}\right)
\end{aligned} \tag{3.8}$$

The density of the classical measurement error model (2.9) can be approximated and applied on the equation (2.12) based on the result of the approximated probability density function of X denoted as $\tilde{f}_{x,m}(x;\alpha,\beta,\bar{\theta}_m)$.

$$\begin{aligned}
\tilde{f}_{w,m}(w;\tau,\alpha,\beta,\bar{\theta}_m) &= \frac{1}{\tau}\int f_{x,m}(x;\alpha,\beta,\bar{\theta}_m)\phi\left(\frac{w-x}{\tau}\right)dx \\
&= \frac{1}{\tau}\int\left\{\frac{1}{\alpha}\sum_{k=0}^{m}\theta_{k,m}beta_{(k+1,m-k+1)}\left(\frac{x-\beta}{\alpha}\right)\right\}\phi\left(\frac{w-x}{\tau}\right)dx \\
&= \frac{1}{\alpha\tau}\sum_{k=0}^{m}\theta_{k,m}\int beta_{(k+1,m-k+1)}\left(\frac{x-\beta}{\alpha}\right)\phi\left(\frac{w-x}{\tau}\right)dx
\end{aligned} \tag{3.9}$$

The theorem 3.5 by Lorentz, 2013 is also applied to $\tilde{f}_{w,m}(w;\tau,\alpha,\beta,\bar{\theta}_m)$ since $f_s$ is continuous. Hence, we apply the Bernstein log-likelihood of the set of unknown parameters $\tau, \alpha, \beta, \bar{\theta}_m$ in following sections.

Guan, 2016 defined Bernstein log-likelihood function of the set of parameters ($\theta$) given the observed data as

$$\ell(\theta_m) = \sum_{j=1}^{n}\log f_B(x_j,\theta_m) \tag{3.10}$$

In this case, the Bernstein log-likelihood function of the set of parameters $(\tau, \alpha, \beta, \bar{\theta})$ given the observed data is

$$\ell_n(\tau, \alpha, \beta, \bar{\theta}_m) = \sum_{i=1}^{n} \log f_{w,m}(W_i; \tau, \alpha, \beta, \bar{\theta}_m)$$
$$= \sum_{i=1}^{n} \log \left[ \frac{1}{\alpha\tau} \sum_{k=0}^{m} \theta_{k,m} \int beta_{(k+1,m-k+1)} \left( \frac{x-\beta}{\alpha} \right) \phi \left( \frac{W_i - x}{\tau} \right) dx \right]. \tag{3.11}$$

where a collection of sample of W are independent and identically distributed, $W_1, W_2, ..., W_n \sim f(w)$. Therefore, the maximizer $\hat{\tau}, \hat{\alpha}, \hat{\beta}, \hat{\theta}$ of $\ell(\tau, \alpha, \beta, \bar{\theta_m})$ is called the MBLE of $\tau, \alpha, \beta, \bar{\theta}_m$ and the MBLEs $\hat{f}_w(w) = f_w(w; \hat{\tau}_m, \hat{\alpha}_m, \hat{\beta}_m, \hat{\theta}_m)$ and $\hat{F}_w(w) = F_w(w; \hat{\tau}_m, \hat{\alpha}_m, \hat{\beta}_m, \hat{\theta}_m)$ of $f(w)$ and $F(w)$ are called "the Bernstein probability density function" and "the Bernstein cumulative distribution function" respectively (Guan, 2016).

## 3.2 Maximum Bernstein likelihood density estimator of the ROC curve in the presence of measurement error

In this section, the Maximum Bernstein likelihood density estimation procedure will be performed on the ROC curve in the presence of measurement error given in the equation 2.1 above. As we presumed that diagnostic test results $x_1^1, x_2^1, ...x_m^1$ and $x_1^0, x_2^0, ..., x_n^0$ denoted as $X_1$ and $X_0$ are from the diseased and non-diseased population having cumulative distribution functions $F_1$ and $F_0$ respectively.

The proposed refining nonparametric approach (MBLE) is to provide a smooth ROC curve in the presence of measurement error using the mixture of flexible approximate parametric estimation with Bernstein type polynomials. Hence, the proposed estimation method for the ROC curve involves replacing $F_1$ and $F_0$ by their distribution functions $\hat{F}_{1m}(s_1)$ and $\hat{F}_{0n}(s_0)$, respectively, by estimating the $\bar{\theta}$ parameters via maximum Bernstein likelihood due to the presence of the measurement error in the estimator of $F$.

In this context, density $f_{s1}$ and $f_{s0}$ can be approximately modeled and parameterized by $f_{s_1,m}(s_1; \bar{\theta_1}_m)$ and $f_{s0,n}(s_0; \bar{\theta_0}_n)$, respectively, as a mixture of the beta distribution and estimate the $\bar{\theta}_m$ parameters via the maximum likelihood method. Therefore, by applying the equation 3.3, $f_1$ can be estimated as

$$\widehat{f_1}(x;\bar{\theta_{1m}}) = \sum_{k=0}^{m} \theta_{k,m} p_{m,k}(x) \tag{3.12}$$

where $\bar{\theta}_m = (\theta_{1,m}, ..., \theta_{m,m})^T$. Clearly, $F_1(x) = \int_{-\infty}^{x} f_1(t)dt$ is corresponding continuous cumulative density function. The $F_1$ also can be obtained approximately as

$$\widehat{F_1}(x;\bar{\theta_{1m}}) = \sum_{k=0}^{m} \theta_{k,m} \int_0^x p_{m,k}(u)du \tag{3.13}$$

similarly, $\tilde{f_0}(x)$ and $\tilde{F_0}(x)$ can obtained similarly with a polynomial with $n$ positive coefficients $\bar{\theta_{0n}}$. After obtaining the estimated $\widehat{F_1}$ and $\widehat{F_0}$ respectively, $\widehat{F_0}^{-1}$ can be acquired by doing inverse of $\widehat{F_0}$. Eventually, the ROC curve can be estimated by plugging in the estimated $\widehat{F_1}$, $\widehat{F_0}^{-1}$ into the equation 2.1.

Each probability density function of $X_1$ and $X_0$ can be approximated with a $beta_{\alpha,\beta}(\cdot)$ which is probability density function of a beta-distributed random variables based on the parameters $\alpha_1, \beta_1$ and $\alpha_0, \beta_0$ respectively from the assumption equation (2.11). Hence, the equations of the the approximated probability density function of $X_1$ and $X_0$ denoted as $\tilde{f}_{1_{x,m}}(x_1;\alpha_1,\beta_1,\bar{\theta}_{1m})$ and $\tilde{f}_{0_{x,n}}(x_0;\alpha_0,\beta_0,\bar{\theta}_{0n})$, respectively, are

$$\begin{aligned} f_{1_{x,m}}(x;\alpha_1,\beta_1,\bar{\theta}_{1m}) &= \frac{1}{(\alpha_1+\beta_1)-\beta_1} \tilde{f}_{s_1,m}\left(\frac{X-\beta_1}{\alpha_1};\bar{\theta}_{1m}\right) \\ &= \frac{1}{\alpha_1}\sum_{k=0}^{m} \theta_{k,m} beta_{(k+1,m-k+1)}\left(\frac{X-\beta_1}{\alpha_1}\right) \end{aligned} \tag{3.14}$$

As Lorentz, 2013 stated since both of each $f_{s1}$ and $f_{s2}$ are continuous, the theorem 3.5 applied to the approximated density of $f_{w_1,m}(w_1,\tau_1,\alpha_1,\beta_1,\bar{\theta}_{1m})$

$$\begin{aligned}\tilde{f}_{w_1,m}(w_1;\tau_1,\alpha_1,\beta_1,\bar{\theta}_{1m}) &= \frac{1}{\tau_1}\int f_{x_{(1)},m}(x_{(1)};\alpha_1,\beta_1,\bar{\theta}_{1m})\phi\left(\frac{w_1-x_1}{\tau_1}\right)dx_1 \\ &= \frac{1}{\tau_1}\int\left\{\frac{1}{\alpha_1}\sum_{k=0}^{m}\theta_{k,m}beta_{(k+1,m-k+1)}\left(\frac{x_1-\beta_1}{\alpha_1}\right)\right\}\phi\left(\frac{w_1-x_1}{\tau_1}\right)dx_1 \\ &= \frac{1}{\alpha_1\tau_1}\sum_{k=0}^{m}\theta_{k,m}\int beta_{(k+1,m-k+1)}\left(\frac{x_1-\beta_1}{\alpha_1}\right)\phi\left(\frac{w_1-x_1}{\tau_1}\right)dx_1\end{aligned} \tag{3.15}$$

which results similarly to $f_{w_0,n}(w_0,\tau_0,\alpha_0,\beta_0,\bar{\theta}_{0n})$. Finally, the Bernstein log-likelihood function as defined as Guan, 2016 of the set of parameters, $(\tau_1,\alpha_1,\beta_1,\bar{\theta}_{1m})$ given the observed data is

$$\begin{aligned}\ell_{n1}(\tau_1,\alpha_1,\beta_1,\bar{\theta}_{1m}) &= \sum_{i=1}^{n_1}\log f_{W_{1i},m}(w_1;\tau_1,\alpha_1,\beta_1,\bar{\theta}_{1m}) \\ &= \sum_{i=1}^{n_1}\log\left[\frac{1}{\alpha_1\tau_1}\sum_{k=0}^{m}\theta_{k,m}\int beta_{(k+1,m-k+1)}\left(\frac{x_1-\beta_1}{\alpha_1}\right)\phi\left(\frac{W_{1i}-x_1}{\tau_1}\right)dx_1\right].\end{aligned} \tag{3.16}$$

where a collection of sample of $W_1$ are independent and identically distributed, $W_{11},W_{12},W_{13},...,W_{1p} \sim W_1$. Besides, it applies to the Bernstein log-likelihood function of the set of parameters, $(\tau_0,\alpha_0,\beta_0,\bar{\theta}_{0m})$ denoted as $l_{n0}(\tau_0,\alpha_0,\beta_0,\bar{\theta}_{0n})$ similarly where $W_0$ is a collection of sample with i.i.d , $W_{01},W_{02},W_{03},...,W_{0_{n0}} \sim W_0$.

$$\begin{aligned}\ell_{n0}(\tau_0,\alpha_0,\beta_0,\bar{\theta}_{0n}) &= \sum_{i=1}^{n_0}\log f_{W_{0i},n}(w_0;\tau_0,\alpha_0,\beta_0,\bar{\theta}_{0n}) \\ &= \sum_{i=1}^{n_0}\log\left[\frac{1}{\alpha_0\tau_0}\sum_{k=0}^{n}\theta_{k,n}\int beta_{(k+1,n-k+1)}\left(\frac{x_0-\beta_0}{\alpha_0}\right)\phi\left(\frac{W_{0i}-x_0}{\tau_0}\right)dx_0\right].\end{aligned} \tag{3.17}$$

The degree of the Bernstein polynomial, $m$ and $n$, respectively for each distribution is crucial to determine the optimal value because it determines the model of the Bernstein polynomial. An aforementioned advantage of the maximum Bernstein likelihood density estimation is only one regularization parameter for each density to choose. Such as the choice of optimal bandwidth and kernel function are difficult, selecting tuning parameters of the Bayesian approach is even more complicated (Guan, 2016). In this case, the

Bernstein polynomial model is only determined by each positive integer $m$ and $n$, respectively. Even more importantly, the constraints of $\theta$ must satisfy to have valid densities for $w_1$ and $w_0$, (i) sum of $\bar{\theta}_{1m}$ and $\bar{\theta}_{0n}$ add up to 1, (ii) they take values in between 0 and 1.

The choice of the each optimal value, $m$ and $n$, is based on the result of optimization. There are two proposed means of finding the maximum likelihood estimate $(\hat{\tau}_m, \hat{\alpha}_m, \hat{\beta}_m, \hat{\theta}_m)$ of $(\tau_1, \alpha_1, \beta_1, \bar{\theta}_{1m})$ and $(\hat{\tau}_n, \hat{\alpha}_n, \hat{\beta}_n, \hat{\theta}_n)$ of $(\tau_0, \alpha_0, \beta_0, \bar{\theta}_{0n})$. The EM algorithm is applied to estimate the listed parameters, iteratively, to find the maximum. The EM algorithm is relatively simple to apply on the Bernstein model and produces reasonably fast results.

## 3.3 The EM Algorithm

The Expectation-Maximization (EM) Algorithm has computational advantages for the maximum likelihood parameter estimation. To find parameters, $\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\theta}$, the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) was applied and used as a way of iterative method to solve and estimate the set of parameters. The EM algorithm was developed to assist maximum likelihood estimation with missing or incomplete data, but it can be used in other situations where procedures are necessary. The EM algorithm is directed at finding a value of parameters, $\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\theta}$, which maximize sampling densities of $f(w; \hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\theta})$ given an observed $w$.

The EM algorithm has two steps, the expectation step (E-step) and the maximization step (M-step). In order to find the optimal value of parameters by the EM approach, $w$ was considered as the *incomplete* component labeled data. The information of these incomplete component labeled data in the mixture models can be estimated given the initial parameters in the E-step. In M-step, the ML was used to derive better parameters estimates for each component separately. Schröder and Rahmann, 2017 stated these two steps are repeatedly run until getting optimal values of parameters where "the EM converges to a local optimum of the log-likelihood function" (Schröder and Rahmann, 2017). The iterations of the EM algorithm for $\hat{\alpha}, \hat{\beta}, \hat{\tau}, \theta_m$ are described.

**Step of Algorithm for finding parameters $(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \theta_m)$ based on $m$**

**step 0.** Initialize $\hat{\alpha}, \hat{\beta}, \hat{\tau}$ based on the information of the data.

**step 1.** Set $m$ and initialize $\theta$ based on the $m$:

$$\theta = \mathbf{1}^T/(m+1)$$

**step 2.** Set $s = 0$ and start with positive initial $\theta_k^{(0)}$ with convergence ($s \hookrightarrow \infty$) of $\theta^{(s)} = (\theta_0^{(s)}, \cdots, \theta_m^{(s)})^T$ to $\bar{\theta}$. The iteration of the EM algorithm for $\hat{\theta}$ is

$$\theta_k^{(s+1)} = \theta_k^{(s)} S_{mn}^{(k)}(\theta^{(s)}), \;\; k \in I_0^{(m)};\; s \in I_0^{\infty}$$

It can be equated as

$$\theta_k^{(s+1)} = \theta_k^{(s)} \frac{1}{n} \sum_{j=1}^{n} \frac{\int_0^1 beta_{(k+1,m-k+1)}(\frac{x-\beta}{\alpha})\phi(\frac{W_j-x}{\tau})dx}{\sum_{k=0}^{m} \theta_k \int_0^1 beta_{(k+1,m-k+1)}(\frac{x-\beta}{\alpha})\phi(\frac{W_j-x}{\tau})dx}$$

The log-likelihood function of the set of parameters $(\tau, \alpha, \beta, \theta)$ given the observed data is defined as

$$\ell(\tau, \alpha, \beta, \theta) = \sum_{i=1}^{n} \log f_{w,m}(W_i; \tau, \alpha, \beta, \theta)$$

**step 3.** repeat step 1 and 2 until satisfying optimization options.

This step of the algorithm is given for $w_1$ and the same applies to perform maximum likelihood estimation for $w_0$.

## 3.4 Model Selection

The degrees of the Bernstein polynomial, $m$ and $n$, for each distribution, is crucial to determine the optimal value because it determines the model of the Bernstein polynomial. As previously mentioned, an advantage of the ML Bernstein density estimation is only one regularization parameter for each density to choose. In this case, the Bernstein polynomial model is only determined by each positive integer $m$ and $n$ for each distribution, it is called the optimal degrees $m$ and $n$. To find the optimal degree $m$ and $n$, the selection method of Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are considered. Claeskens, Hjort, et al., 2008 emphasized the importance of the choice of model selection criterion, "most of the selection methods are defined in terms of an appropriate *information criterion*, a mechanism that uses data to give each candidate model a certain score; this

then leads to a fully ranked list of candidate models, from the best to the worst" (Claeskens, Hjort, et al., 2008). The general formula of the AIC is

$$AIC(M) = -2\log_{max}(Model) + 2length(Model) \tag{3.18}$$

where $M$ indicates for each candidate model and the length($M$) indicates the number of parameters you estimate in the model.

Bayesian Information Criterion (BIC) is another approach, the candidate model with the highest probability is selected given the data (Claeskens, Hjort, et al., 2008). The general formula of the BIC is

$$BIC(f(\cdot;\phi)) = -2\log L(\hat{\phi}) + \log(n)length(\phi) \tag{3.19}$$

where $\phi$ represents parameters, $n$ is the sample size, and length($\phi$) is number of parameters ($\phi$). BIC is the primary model selection in the following chapters because it produced better results since BIC punishes the number of parameters in the model.

# Chapter 4

# Simulation Study

To evaluate the performance of the proposed estimator of the ROC curve in the presence of measurement errors, we conducted simulation studies using the MBLE estimation method for which the details are given in Chapter 3. A comprehensive simulation study is performed under three various scenarios that $F_0$ and $F_1$ are either Exponential or Beta or Normal distributions. We generated the simulated data $w_{01}, ..., w_{0n_0}$ and $w_{11}, ..., w_{1n_1}$ with three different sample sizes, $(n_1, n_0)$ = (30, 30), (100, 30), (300, 300), based on the measurement error model with six different distributions for $X$ ($X = \alpha S + \beta$) using four different values of the noise-to-signal (NSR), $\frac{\tau}{\sigma_X} = (0.1, 0.25, 0.50, 0.75)$, where $\sigma_X$ is the standard deviation of $X$. The details of simulation setting is presented in the forthcoming section.

## 4.1 Data generation

Three simulation scenarios are considered to generate the sample data.

- Scenario I: $X_1 \sim 20Exp(10, 20) + 1$ and $X_0 \sim 4Exp(0.5, 4) + 1$

- Scenario II: $X_1 \sim 7Beta(1, 1) - 7$ and $X_0 \sim 4Beta(2, 2) - 4$

- Scenario III: $X_1 \sim Normal(0, 1, -7, 7)$ and $X_0 \sim Normal(2, 1, -7, 7)$

In the first simulation study, Exponential distribution with $(\mu, t_U) = (10, 20)$ and $(0.5, 4)$ of mean $\mu$ and truncated at $t_U$ are used for $X$. $F_1$ and $F_0$ are generated with $X_1 \sim 20Exp(10, 20) + 1$ and $X_0 \sim 4Exp(0.5, 4) + 1$ with true AUC of each sample size (0.986, 0.979, 0.987) for the evaluation of excellent

biomarker. Since $Exp(10, 20)$ and $Exp(0.5, 4)$ are not within compact unit interval, $\alpha$ and $\beta$ parameters $(\alpha, \beta)$ for each distribution, $20Exp(10, 20) + 1$ and $4Exp(0.5, 4) + 1$, were $(400, 1)$ and $(10, 1)$, respectively.

In the following second simulation study, Beta$(a, b)$ where $a$ and $b$ are shape parameters of the beta distribution is generated with $(1, 1)$ and $(2, 2)$. The parameters, $\alpha$ and $\beta$, were assigned $(7, -7)$ and $(4, -4)$, respectively, since the generated sample data from Beta(1, 1) and Beta(2, 2) is within support range. Therefore, $F_1$ and $F_0$ were generated with $X_1 \sim 7Beta(1, 1) - 7$ and $X_0 \sim 4Beta(2, 2) - 4$ with true AUC of each sample size (0.312, 0.282, 0.316) for the evaluation of worse biomarker.

Lastly, the distributions used for $X$ in the third simulation study are generated with Normal$(0, 1)$ and Normal$(2, 1)$ truncated by the interval [-7, 7]. Hence, $F_1$ and $F_0$ were generated with $X_1 \sim Normal(0, 1, -7, 7)$ and $X_0 \sim Normal(2, 1, -7, 7)$ where two last numbers (-7, 7) indicate the truncated interval with true AUC of each sample size (0.938, 0.876, 0.927) for the evaluation of fairly good biomarker. Since $Normal(0, 1)$ and $Normal(2, 1)$ are not within compact unit interval, the parameters, $\alpha, \beta$, were $(5, -1)$ and $(5, -3)$, respectively, after truncating each distribution.

The density functions corresponding to these cases are displayed in Figure 4.1. The figure 4.1 is illustrated that MBLE accurately estimates the density distribution in various scenarios under weak assumption.

Initially, parameters of interest $(\tau, \alpha, \beta, \bar{\theta})$ and the optimal degrees $m$ and $n$ for each contaminated distributions $W_1$ and $W_0$, respectively, are selected by using EM Algorithm with the optimal degree range $M = (1, 10)$ for each simulation. The optimal degree range $M = (1, 10)$ indicates each simulation was conducted with the optimal degrees $m = 1, 2, ..., 10$ and $n = 1, 2, ..., 10$, respectively. The algorithm for the estimation of $\tau$ was performed with different noise-to-signal values to produce the largest value of the maximum log-likelihood of the objective function. The best parameters of a model of interest, $\hat{\tau}, \hat{\alpha}, \hat{\beta}, \bar{\theta}_m$, are chosen using BIC criterion based on the estimated $\hat{\tau}$. The proposed estimator of the ROC curve for each generated sample data from the desired distributions given sample sizes are calculated at the grid points $p_i = 0.01$ to $0.99$, $i = 1, 2, ..., 512$. Two assessments are performed to assess the performance of the proposed estimator of the ROC curve and AUC in the presence of the measurement error. The process is replicated $N = 150$ times.

The first assessment of the estimator of the ROC curve is the integrated mean bias.

$$MIB(\widehat{ROC}) = \frac{1}{N}\sum_{s=1}^{N}\int(\widehat{ROC_s}(p) - ROC_s(p))dp$$

Mean Integrated Square Error (MISE) is performed as the second assessment to assess the overall performance of the proposed estimator (Zhou and Harezlak, 2002; Wang and Cai, 2021).

$$MISE(\widehat{ROC}) = \frac{1}{N}\sum_{s=1}^{N}\int(\widehat{ROC_s}(p) - ROC_s(p))^2dp$$

where the empirical estimator of ROC curves is used as a reference for both assessment. Because empirical estimator of ROC curves was performed under exactly same condition with the proposed estimator.

## 4.2 Simulation Results

The simulation results below are organized into three parts. First, we compare the estimation of the parameters of a model interest. Second, the overall performance of the MBLE estimator of the ROC curve in the presence of measurement errors given different sample sizes is assessed by comparing the Empirical estimator of the ROC curve ($AUC_E$). Lastly, we compare distribution of the selected the optimal degrees $m$ and $n$ for each distribution.

Tables 4.1, 4.3, and 4.5 summarize the estimation results of parameters of the model interest for the three simulation scenarios. Each table contains the results of the parameter estimations with standard deviation (SD) by different NSR settings. The results regarding the estimation of the R0C curve in the presence of measurement error based on the estimation of $\tau$ for $(n_1, n_0) = (30, 30), (30, 100), (300, 300)$ can be found in Table 4.1, Table 4.3, Table 4.5, respectively. When there is small amount of measurement error ($NSR = 0.1$), difference between the true $\tau$ and the estimated $\hat{\tau}$ is relatively small. For other NSR settings, $\hat{\tau}$ is always underestimated. This difference between the true $\tau$ and the estimated $\hat{\tau}$ tends to increase as NSR increases. The proposed model tends to depreciate the estimation of $\hat{\tau}$, therefore, it produces almost perfectly when there is low standard error except $20Exp(10, 20) + 1$ and $4Beta(2, 2) - 4$ cases. In general, an increase in the NSR is associated with the overestimation of $\alpha$ and $\beta$. The estimation of $\tau$ for $20Exp(10, 20) + 1$ was affected by number of sample size and different NSR setting. For instance, when the setting of $NSR$ is 0.1, the estimated $\hat{\tau}$ values of $20Exp(10, 20) + 1$ for $n_1 = (30, 100, 300)$ are approximately $(31, 39, 31)$ which

is differed a lot from the true $\tau$ value for three different sample sizes. On the other hand, those didn't affect much for the estimation of $\hat{\tau}$ in Simulation I (only $4Exp(0.5, 4) + 1$ case), Simulation II (only $4Beta(2, 2) - 4$ case), and Simulation III. It estimated more or less similar value of $\hat{\tau}$ in different sample sizes especially in Simulation III.

The assessments of two different estimators of the ROC curve are summarized in Tables 4.2, 4.4, and 4.6. We observed that (1) the uncorrected AUC estimates ($AUC_E$) underestimated the true $AUC$ for all 4 NSR settings in Simulation study I, II and III. The MBLE estimator underestimated even more than $AUC_E$ in every Simulation study. (2) The MIB and MISE was larger as a value of NSR increases and sample size increases. (3) In Simulation study I and III, result values of MISE were pretty much similar in different NSR setting. Repeatedly, according to the results of two assessments, MIB and MISE, the proposed estimator of the ROC curve tends to yield slightly bigger MIB and MISE values as NSR increases. The larger sample size produces moderately larger MIB and MISE as well.

Distribution of the selected optimal degrees, $m$ and $n$, using EM algorithm with the estimated variance of measurement error when range of optimal degree is in between $[1, 10]$ is displayed in Table 4.7. It shows that distributions of the selected optimal degrees, $m$ and $n$, are fairly dispersed in each different NSR setting. But, there are some trends observed, for example, lower value of the optimal degrees $m$ was frequently selected in Simulation I for diseased group with lower NSR setting. In simulation III, the selected optimal degrees are evenly distributed within the range of $[4, 7]$. Simulation code can be found here.

| | $n_1 = 30$ | | | | $n_0 = 30$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $X_1 \sim 20Exp(10,20)+1$ | | | | $X_0 \sim 4Exp(0.5,4)+1$ | | | |
| **NSR** | **0.1** | **0.25** | **0.5** | **0.75** | **0.1** | **0.25** | **0.5** | **0.75** |
| $\tau$ | 10.482 | 27.782 | 58.265 | 97.056 | 0.192 | 0.458 | 1.154 | 1.993 |
| $\hat{\tau}$ **(SD** $\hat{\tau}$**)** | 30.933 (24.732) | 36.897 (25.375) | 34.229 (26.749) | 42.9 (29.971) | 0.198 (0.051) | 0.201 (0.045) | 0.214 (0.038) | 0.241 (0.038) |
| $\alpha$ | 400 | 400 | 400 | 400 | 10 | 10 | 10 | 10 |
| $\hat{\alpha}$ **(SD** $\hat{\alpha}$**)** | 425.606 (62.108) | 461.098 (69.366) | 515.993 (85.775) | 607.628 (94.758) | 8.362 (2.145) | 8.799 (2.212) | 9.762 (2.337) | 10.944 (2.441) |
| $\beta$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $\hat{\beta}$ **(SD** $\hat{\beta}$**)** | -29.442 (26.421) | -51.656 (31.639) | -85.955 (40.42) | -136.928 (50.439) | 0.702 (0.153) | 0.303 (0.317) | -0.494 (0.569) | -1.369 (0.802) |
| | $X_1 \sim 7Beta(1,1)-7$ | | | | $X_0 \sim 4Beta(2,2)-4$ | | | |
| **NSR** | **0.1** | **0.25** | **0.5** | **0.75** | **0.1** | **0.25** | **0.5** | **0.75** |
| $\tau$ | 0.207 | 0.544 | 1.126 | 1.899 | 0.095 | 0.228 | 0.519 | 0.829 |
| $\hat{\tau}$ **(SD** $\hat{\tau}$**)** | 0.339 (0.218) | 0.307 (0.2) | 0.27 (0.144) | 0.267 (0.083) | 0.374 (0.168) | 0.325 (0.188) | 0.211 (0.156) | 0.168 (0.136) |
| $\alpha$ | 7 | 7 | 7 | 7 | 4 | 4 | 4 | 4 |
| $\hat{\alpha}$ **(SD** $\hat{\alpha}$**)** | 7.336 (0.422) | 7.769 (0.644) | 8.886 (1.108) | 10.289 (1.541) | 4.014 (0.405) | 4.088 (0.404) | 4.32 (0.486) | 4.824 (0.637) |
| $\beta$ | -7 | -7 | -7 | -7 | -4 | -4 | -4 | -4 |
| $\hat{\beta}$ **(SD** $\hat{\beta}$**)** | -7.164 (0.407) | -7.372 (0.56) | -7.645 (0.819) | -8.616 (1.083) | -4.02 (0.303) | -4.061 (0.342) | -3.963 (0.379) | -4.426 (0.52) |
| | $X_1 \sim Normal(0,1,-7,7)$ | | | | $X_0 \sim Normal(2,1,-7,7)$ | | | |
| | **0.1** | **0.25** | **0.5** | **0.75** | **0.1** | **0.25** | **0.5** | **0.75** |
| $\tau$ | 0.103 | 0.264 | 0.514 | 0.704 | 0.1 | 0.248 | 0.485 | 0.779 |
| $\hat{\tau}$ **(SD** $\hat{\tau}$**)** | 0.099 (0.014) | 0.099 (0.014) | 0.101 (0.014) | 0.102 (0.017) | 0.099 (0.014) | 0.099 (0.014) | 0.1 (0.014) | 0.101 (0.014) |
| $\alpha$ | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| $\hat{\alpha}$ **(SD** $\hat{\alpha}$**)** | 4.327 (0.779) | 4.334 (0.771) | 4.364 (0.756) | 4.41 (0.746) | 4.26 (0.688) | 4.267 (0.689) | 4.296 (0.695) | 4.349 (0.697) |
| $\beta$ | -1 | -1 | -1 | -1 | -3 | -3 | -3 | -3 |
| $\hat{\beta}$ **(SD** $\hat{\beta}$**)** | -0.137 (0.464) | -0.14 (0.462) | -0.155 (0.456) | -0.177 (0.455) | -2.155 (0.48) | -2.158 (0.479) | -2.17 (0.48) | -2.194 (0.481) |

TABLE 4.1: Simulation results respecting the estimation of $\tau, \alpha, \beta, \bar{\theta}$ for sample size $n_0 = 30$ and $n_1 = 30$

| | $n_1 = 30$ | | $n_0 = 30$ | |
|---|---|---|---|---|
| | $X_1 \sim 20Exp(10,20)+1$ | | $X_0 \sim 4Exp(0.5,4)+1$ | |
| **NSR** | $\widehat{AUC}_E$ | $\widehat{AUC}_{MBLE.decon}$ | **MIB** | **MISE** |
| **0.1** | 0.905 | 0.837 | -0.072 | 0.021 |
| **0.25** | 0.906 | 0.836 | -0.073 | 0.021 |
| **0.5** | 0.904 | 0.835 | -0.074 | 0.021 |
| **0.75** | 0.902 | 0.832 | -0.077 | 0.022 |
| | $X_1 \sim 7Beta(1,1)-7$ | | $X_0 \sim 4Beta(2,2)-4$ | |
| **NSR** | $\widehat{AUC}_E$ | $\widehat{AUC}_{MBLE.decon}$ | **MIB** | **MISE** |
| **0.1** | 0.281 | 0.282 | 0.003 | 0.010 |
| **0.25** | 0.283 | 0.285 | 0.006 | 0.009 |
| **0.5** | 0.29 | 0.298 | 0.016 | 0.009 |
| **0.75** | 0.302 | 0.319 | 0.025 | 0.011 |
| | $X_1 \sim Normal(0,1,-7,7)$ | | $X_0 \sim Normal(2,1,-7,7)$ | |
| **NSR** | $\widehat{AUC}_E$ | $\widehat{AUC}_{MBLE.decon}$ | **MIB** | **MISE** |
| **0.1** | 0.905 | 0.837 | -0.072 | 0.021 |
| **0.25** | 0.906 | 0.836 | -0.073 | 0.021 |
| **0.5** | 0.904 | 0.835 | -0.074 | 0.021 |
| **0.75** | 0.902 | 0.832 | -0.077 | 0.022 |

TABLE 4.2: Simulation results of the Empirical estimator of AUC ($\widehat{AUC}_E$), Maximum Bernstein Likelihood Estimator (MBLE) of AUC for deconvolution ($\widehat{AUC}_{MBLE.decon}$), MIB, and MISE based on the estimated $\hat{\tau}, \hat{\alpha}, \hat{\beta}, \hat{\bar{\theta}}$ for sample size $n_1 = 30$ and $n_0 = 30$

| | $n_1 = 100$ | | | | $n_0 = 30$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $X_1 \sim 20Exp(10,20)+1$ | | | | $X_0 \sim 4Exp(0.5,4)+1$ | | | |
| **NSR** | **0.1** | **0.25** | **0.5** | **0.75** | **0.1** | **0.25** | **0.5** | **0.75** |
| $\tau$ | 10.482 | 27.782 | 58.265 | 97.056 | 0.192 | 0.458 | 1.154 | 1.993 |
| $\hat{\tau}$ **(SD** $\hat{\tau}$**)** | 38.766 (25.894) | 34.456 (23.7) | 35.109 (25.612) | 43.954 (34.426) | 0.198 (0.051) | 0.201 (0.045) | 0.214 (0.038) | 0.241 (0.038) |
| $\alpha$ | 400 | 400 | 400 | 400 | 10 | 10 | 10 | 10 |
| $\hat{\alpha}$ **(SD** $\hat{\alpha}$**)** | 477.598 (57.392) | 507.196 (57.036) | 588.644 (68.723) | 699.522 (96.618) | 8.362 (2.145) | 8.799 (2.212) | 9.762 (2.337) | 10.944 (2.441) |
| $\beta$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $\hat{\beta}$ **(SD** $\hat{\beta}$**)** | -46.268 (27.158) | -67.115 (27.074) | -115.298 (37.397) | -176.633 (54.163) | 0.702 (0.153) | 0.303 (0.317) | -0.494 (0.569) | -1.369 (0.802) |
| | $X_1 \sim 7Beta(1,1)-7$ | | | | $X_0 \sim 4Beta(2,2)-4$ | | | |
| **NSR** | **0.1** | **0.25** | **0.5** | **0.75** | **0.1** | **0.25** | **0.5** | **0.75** |
| $\tau$ | 0.207 | 0.544 | 1.126 | 1.899 | 0.095 | 0.228 | 0.519 | 0.829 |
| $\hat{\tau}$ **(SD** $\hat{\tau}$**)** | 0.321 (0.209) | 0.296 (0.195) | 0.269 (0.151) | 0.271 (0.108) | 0.227 (0.12) | 0.174 (0.12) | 0.125 (0.079) | 0.118 (0.034) |
| $\alpha$ | 7 | 7 | 7 | 7 | 4 | 4 | 4 | 4 |
| $\hat{\alpha}$ **(SD** $\hat{\alpha}$**)** | 7.344 (0.361) | 7.807 (0.557) | 8.968 (0.975) | 10.397 (1.46) | 4.115 (0.17) | 4.272 (0.247) | 4.799 (0.429) | 5.539 (0.602) |
| $\beta$ | -7 | -7 | -7 | -7 | -4 | -4 | -4 | -4 |
| $\hat{\beta}$ **(SD** $\hat{\beta}$**)** | -7.143 (0.359) | -7.366 (0.462) | -7.927 (0.68) | -8.632 (0.945) | -4.059 (0.163) | -4.146 (0.233) | -4.424 (0.345) | -4.804 (0.459) |
| | $X_1 \sim Normal(0,1,-7,7)$ | | | | $X_0 \sim Normal(2,1,-7,7)$ | | | |
| **NSR** | **0.1** | **0.25** | **0.5** | **0.75** | **0.1** | **0.25** | **0.5** | **0.75** |
| $\tau$ | 0.103 | 0.264 | 0.514 | 0.704 | 0.1 | 0.248 | 0.485 | 0.779 |
| $\hat{\tau}$ **(SD** $\hat{\tau}$**)** | 0.099 (0.014) | 0.1 (0.007) | 0.101 (0.007) | 0.102 (0.007) | 0.099 (0.014) | 0.099 (0.014) | 0.1 (0.014) | 0.101 (0.014) |
| $\alpha$ | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| $\hat{\alpha}$ **(SD** $\hat{\alpha}$**)** | 5.228 (0.574) | 5.244 (0.571) | 5.287 (0.569) | 5.346 (0.579) | 4.26 (0.688) | 4.267 (0.689) | 4.296 (0.695) | 4.349 (0.697) |
| $\beta$ | -1 | -1 | -1 | -1 | -3 | -3 | -3 | -3 |
| $\hat{\beta}$ **(SD** $\hat{\beta}$**)** | -0.644 (0.437) | -0.655 (0.43) | -0.682 (0.416) | -0.719 (0.405) | -2.155 (0.48) | -2.158 (0.479) | -2.17 (0.48) | -2.194 (0.481) |

TABLE 4.3: Simulation results respecting the estimation of $\tau, \alpha, \beta, \bar{\theta}$ for sample size $n_1 = 100$ and $n_0 = 30$

| | $n_1 = 100$ | | $n_0 = 30$ | |
|---|---|---|---|---|
| | $X_1 \sim 20Exp(10,20)+1$ | | $X_0 \sim 4Exp(0.5,4)+1$ | |
| **NSR** | $\widehat{AUC}_E$ | $\widehat{AUC}_{MBLE.decon}$ | **MIB** | **MISE** |
| **0.1** | 0.951 | 0.907 | -0.034 | 0.002 |
| **0.25** | 0.921 | 0.892 | -0.034 | 0.002 |
| **0.5** | 0.873 | 0.849 | -0.017 | 0.001 |
| **0.75** | 0.834 | 0.807 | -0.024 | 0.002 |
| | $X_1 \sim 7Beta(1,1)-7$ | | $X_0 \sim 4Beta(2,2)-4$ | |
| **NSR** | $\widehat{AUC}_E$ | $\widehat{AUC}_{MBLE.decon}$ | **MIB** | **MISE** |
| **0.1** | 0.274 | 0.28 | 0.004 | 0.009 |
| **0.25** | 0.274 | 0.284 | 0.009 | 0.009 |
| **0.5** | 0.279 | 0.3 | 0.018 | 0.010 |
| **0.75** | 0.291 | 0.323 | 0.027 | 0.013 |
| | $X_1 \sim Normal(0,1,-7,7)$ | | $X_0 \sim Normal(2,1,-7,7)$ | |
| **NSR** | $\widehat{AUC}_E$ | $\widehat{AUC}_{MBLE.decon}$ | **MIB** | **MISE** |
| **0.1** | 0.905 | 0.835 | -0.073 | 0.021 |
| **0.25** | 0.905 | 0.833 | -0.072 | 0.020 |
| **0.5** | 0.903 | 0.831 | -0.072 | 0.019 |
| **0.75** | 0.902 | 0.832 | -0.077 | 0.022 |

TABLE 4.4: Simulation results of the Empirical estimator of AUC ($\widehat{AUC}_E$), Maximum Bernstein Likelihood estimator (MBLE) of AUC for deconvolution ($\widehat{AUC}_{MBLE.decon}$), MIB, and MISE based on the estimated $\hat{\tau}, \hat{\alpha}, \hat{\beta}, \hat{\bar{\theta}}$ for sample size $n_1 = 100$ and $n_0 = 30$

| | $n_1 = 300$ | | | | $n_0 = 300$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $X_1 \sim 20Exp(10,20)+1$ | | | | $X_0 \sim 4Exp(0.5,4)+1$ | | | |
| **NSR** | **0.1** | **0.25** | **0.5** | **0.75** | **0.1** | **0.25** | **0.5** | **0.75** |
| $\tau$ | 10.482 | 27.782 | 58.265 | 97.056 | 0.192 | 0.458 | 1.154 | 1.993 |
| $\hat{\tau}$ **(SD** $\hat{\tau}$**)** | 41.894 (27.956) | 31.583 (21.438) | 42.194 (31.498) | 51.132 (35.318) | 0.197 (0.014) | 0.203 (0.014) | 0.22 (0.013) | 0.247 (0.013) |
| $\alpha$ | 400 | 400 | 400 | 400 | 10 | 10 | 10 | 10 |
| $\hat{\alpha}$ **(SD** $\hat{\alpha}$**)** | 502.127 (56.314) | 535.822 (45.133) | 661.587 (70.99) | 793.119 (85.461) | 12.679 (1.919) | 13.431 (1.943) | 14.935 (2.002) | 16.647 (2.097) |
| $\beta$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $\hat{\beta}$ **(SD** $\hat{\beta}$**)** | -56.296 (27.83) | -78.782 (23.782) | -151.335 (39.814) | -226.111 (48.694) | 0.483 (0.099) | -0.185 (0.231) | -1.426 (0.45) | -2.744 (0.651) |
| | $X_1 \sim 7Beta(1,1)-7$ | | | | $X_0 \sim 4Beta(2,2)-4$ | | | |
| **NSR** | **0.1** | **0.25** | **0.5** | **0.75** | **0.1** | **0.25** | **0.5** | **0.75** |
| $\tau$ | 0.207 | 0.544 | 1.126 | 1.899 | 0.095 | 0.228 | 0.519 | 0.829 |
| $\hat{\tau}$ **(SD** $\hat{\tau}$**)** | 0.202 (0.005) | 0.207 (0.006) | 0.225 (0.007) | 0.251 (0.009) | 0.143 (0.072) | 0.101 (0.032) | 0.101 (0.007) | 0.112 (0.005) |
| $\alpha$ | 7 | 7 | 7 | 7 | 4 | 4 | 4 | 4 |
| $\hat{\alpha}$ **(SD** $\hat{\alpha}$**)** | 7.812 (0.17) | 8.905 (0.355) | 11.068 (0.641) | 13.427 (0.955) | 4.141 (0.125) | 4.408 (0.181) | 5.195 (0.318) | 6.148 (0.436) |
| $\beta$ | -7 | -7 | -7 | -7 | -4 | -4 | -4 | -4 |
| $\hat{\beta}$ **(SD** $\hat{\beta}$**)** | -7.401, (0.119) | -7.944, (0.248) | -9.037 (0.46) | -10.239 (0.682) | -4.075 (0.123) | -4.21 (0.155) | -4.602 (0.244) | -5.076 (0.328) |
| | $X_1 \sim Normal(0,1,-7,7)$ | | | | $X_0 \sim Normal(2,1,-7,7)$ | | | |
| | **0.1** | **0.25** | **0.5** | **0.75** | **0.1** | **0.25** | **0.5** | **0.75** |
| $\tau$ | 0.103 | 0.264 | 0.514 | 0.704 | 0.1 | 0.248 | 0.485 | 0.779 |
| $\hat{\tau}$ **(SD** $\hat{\tau}$**)** | 0.1 (0.004) | 0.1 (0.004) | 0.101 (0.004) | 0.102 (0.004) | 0.101 (0.004) | 0.101 (0.004) | 0.102 (0.004) | 0.103 (0.004) |
| $\alpha$ | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| $\hat{\alpha}$ **(SD** $\hat{\alpha}$**)** | 6.016 (0.556) | 6.024 (0.557) | 6.064 (0.558) | 6.13 (0.565) | 5.959 (0.57) | 5.975 (0.571) | 6.025 (0.575) | 6.106 (0.576) |
| $\beta$ | -1 | -1 | -1 | -1 | -3 | -3 | -3 | -3 |
| $\hat{\beta}$ **(SD** $\hat{\beta}$**)** | -1.052 (0.44) | -1.055 (0.44) | -1.071 (0.441) | -1.099 (0.445) | -3.019 (0.394) | -3.026 (0.397) | -3.05 (0.401) | -3.089 (0.404) |

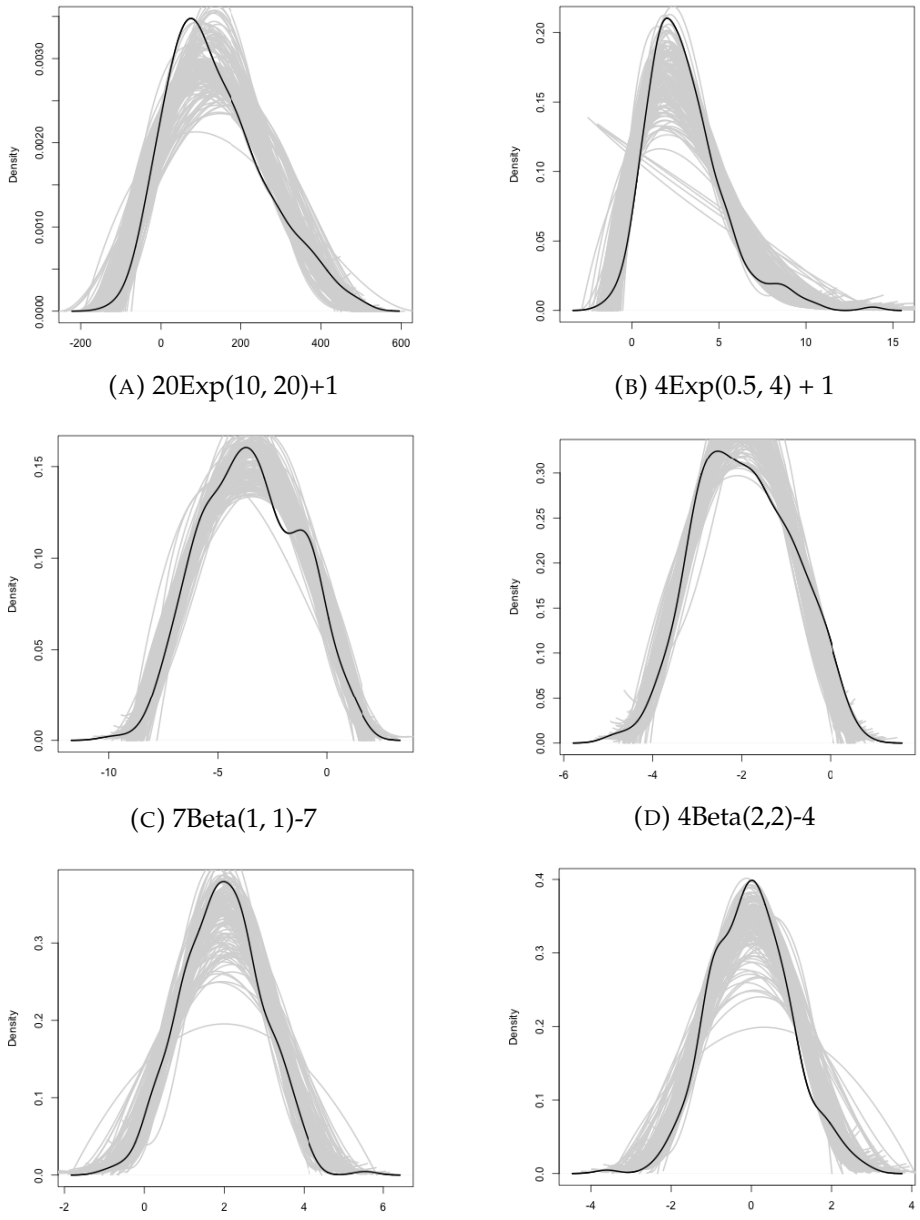
TABLE 4.5: Simulation results respecting the estimation of $\tau, \alpha, \beta, \bar{\theta}$ for sample size $n_1 = 300$ and $n_0 = 300$

| | $n_1 = 300$ | | $n_0 = 300$ | |
|---|---|---|---|---|
| | $X_1 \sim 20Exp(10,20)+1$ | | $X_0 \sim 4Exp(0.5,4)+1$ | |
| **NSR** | $\widehat{AUC}_E$ | $\widehat{AUC}_{MBLE.decon}$ | **MIB** | **MISE** |
| **0.1** | 0.97 | 0.907 | -0.042 | 0.003 |
| **0.25** | 0.939 | 0.88 | -0.044 | 0.003 |
| **0.5** | 0.892 | 0.834 | -0.038 | 0.002 |
| **0.75** | 0.851 | 0.795 | -0.044 | 0.004 |
| | $X_1 \sim 7Beta(1,1)-7$ | | $X_0 \sim 4Beta(2,2)-4$ | |
| **NSR** | $\widehat{AUC}_E$ | $\widehat{AUC}_{MBLE.decon}$ | **MIB** | **MISE** |
| **0.1** | 0.277 | 0.278 | 0.001 | 0.008 |
| **0.25** | 0.284 | 0.285 | 0.006 | 0.010 |
| **0.5** | 0.291 | 0.312 | 0.030 | 0.017 |
| **0.75** | 0.304 | 0.336 | 0.039 | 0.023 |
| | $X_1 \sim Normal(0,1,-7,7)$ | | $X_0 \sim Normal(2,1,-7,7)$ | |
| **NSR** | $\widehat{AUC}_E$ | $\widehat{AUC}_{MBLE.decon}$ | **MIB** | **MISE** |
| **0.1** | 0.919 | 0.717 | -0.188 | 0.088 |
| **0.25** | 0.919 | 0.717 | -0.188 | 0.088 |
| **0.5** | 0.917 | 0.715 | -0.191 | 0.090 |
| **0.75** | 0.915 | 0.711 | -0.192 | 0.092 |

TABLE 4.6: Simulation results of the Empirical estimator of AUC ($\widehat{AUC}_E$), Maximum Bernstein Likelihood Estimator (MBLE) of AUC for deconvolution ($\widehat{AUC}_{MBLE.decon}$), MIB, and MISE based on the estimated $\hat{\tau}, \hat{\alpha}, \hat{\beta}, \hat{\bar{\theta}}$ for sample size $n_1 = 300$ and $n_0 = 300$

| Distribution | NSR | Distribution (in %) of the selected optimal degree | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **m = 2** | **m = 3** | **m = 4** | **m = 5** | **m = 6** | **m = 7** | **m = 8** | **m = 9** |
| $X_1 \sim 20Exp(10,20)+1$ | **0.1** | 2.67 | 41.33 | 12.67 | 18.67 | 13.33 | 6.00 | 4.00 | 1.33 |
| | **0.25** | 2.00 | 64.00 | 11.33 | 16.00 | 1.33 | 0.67 | 4.00 | 0.67 |
| | **0.5** | 2.67 | 33.33 | 14.67 | 38.00 | 2.00 | 1.33 | 0.00 | 8.00 |
| | **0.75** | 6.00 | 4.67 | 32.67 | 46.67 | 1.33 | 6.00 | 0.00 | 2.67 |
| | | **n = 2** | **n = 3** | **n = 4** | **n = 5** | **n = 6** | **n = 7** | **n = 8** | **n = 9** |
| $X_0 \sim 4Exp(0.5,4)+1$ | **0.1** | 12.00 | 60.00 | 10.00 | 0.00 | 0.67 | 1.33 | 12.67 | 3.33 |
| | **0.25** | 1.33 | 5.33 | 0.00 | 19.33 | 21.33 | 26.00 | 20.67 | 6.00 |
| | **0.5** | 2.67 | 0.67 | 22.00 | 26.67 | 22.67 | 6.00 | 12.67 | 6.67 |
| | **0.75** | 8.67 | 4.00 | 9.33 | 18.00 | 28.67 | 9.33 | 16.00 | 6.00 |
| | | **m = 2** | **m = 3** | **m = 4** | **m = 5** | **m = 6** | **m = 7** | **m = 8** | **m = 9** |
| $X_1 \sim 7Beta(1,1)-7$ | **0.1** | 12.00 | 0.00 | 20.67 | 52.00 | 11.33 | 2.67 | 0.67 | 0.67 |
| | **0.25** | 82.67 | 3.33 | 10.00 | 2.67 | 1.33 | 0.00 | 0.00 | 0.00 |
| | **0.5** | 32.67 | 3.33 | 56.00 | 6.00 | 2.00 | 0.00 | 0.00 | 0.00 |
| | **0.75** | 2.00 | 2.00 | 48.00 | 16.00 | 29.33 | 2.67 | 0.00 | 0.00 |
| | | **n = 2** | **n = 3** | **n = 4** | **n = 5** | **n = 6** | **n = 7** | **n = 8** | **n = 9** |
| $X_0 \sim 4Beta(2,2)-4$ | **0.1** | 66.67 | 8.67 | 23.33 | 0.67 | 0.67 | 0.00 | 0.00 | 0.00 |
| | **0.25** | 38.67 | 9.33 | 49.33 | 1.33 | 1.33 | 0.00 | 0.00 | 0.00 |
| | **0.5** | 4.67 | 1.33 | 59.33 | 15.33 | 18.00 | 0.67 | 0.67 | 0.00 |
| | **0.75** | 0.67 | 0.00 | 39.33 | 17.33 | 30.67 | 9.33 | 2.67 | 0.00 |
| | | **m = 2** | **m = 3** | **m = 4** | **m = 5** | **m = 6** | **m = 7** | **m = 8** | **m = 9** |
| $X_1 \sim Normal(0,1,-7,7)$ | **0.1** | 0.67 | 2.00 | 27.33 | 19.33 | 28.00 | 18.00 | 1.33 | 3.33 |
| | **0.25** | 0.67 | 2.00 | 32.00 | 19.33 | 26.00 | 15.33 | 2.00 | 2.67 |
| | **0.5** | 1.33 | 0.00 | 26.67 | 18.67 | 28.67 | 20.00 | 4.00 | 0.67 |
| | **0.75** | 1.33 | 0.67 | 31.33 | 21.33 | 28.00 | 14.67 | 2.00 | 0.67 |
| | | **n = 2** | **n = 3** | **n = 4** | **n = 5** | **n = 6** | **n = 7** | **n = 8** | **n = 9** |
| $X_0 \sim Normal(2,1,-7,7)$ | **0.1** | 1.33 | 0.00 | 22.67 | 20.67 | 28.00 | 19.33 | 6.67 | 1.33 |
| | **0.25** | 2.00 | 0.00 | 20.67 | 22.00 | 28.67 | 19.33 | 6.67 | 0.67 |
| | **0.5** | 1.33 | 0.00 | 22.67 | 22.67 | 32.67 | 16.67 | 3.33 | 0.67 |
| | **0.75** | 0.67 | 0.00 | 24.67 | 24.00 | 28.67 | 16.67 | 4.67 | 0.67 |

TABLE 4.7: Simulation results regarding the selection of the optimal degree of the model ($m$ and $n$) for samples of size $n_1 = 300$ and $n_0 = 300$




(A) 20Exp(10, 20)+1

(B) 4Exp(0.5, 4) + 1

(C) 7Beta(1, 1)-7

(D) 4Beta(2,2)-4

(E) Normal(0, 1, -7, 7)

(F) Normal(2, 1, -7, 7)

FIGURE 4.1: For each distribution of the contaminated data, $\widehat{f_w}$ obtained for 150 datasets (in gray) and $f_w$ (in black), in the case $n_1 = 300$ and $n_0 = 300$ and NSR = 0.5

# Chapter 5

# Application

To illustrate an application of the proposed estimator of the ROC curve in the presence of the measurement error, Subarachnoid hemorrhage data (aSAH) published by Turck et al., 2010 was analyzed to evaluate the capacity of the combination of clinical scores together with brain-injury related biomarkers to distinguish patients with an aneurysmal subarachnoid hemorrhage (aSAH) ($n_1 = 41$) from those who are not ($n_0 = 72$). It remains difficult for acute early prediction of long-term irreversible brain damage during the acute phase of patients with aSAH (Turck et al., 2010). They provided blood samples of 131 aSAH patients from a cohort that were analyzed. The available covariates are gender (37%), age (between 18 and 81, median: 51, standard deviation:), wfns (1-5 categories), s100b (between 0.030 and 2.07, median: 0.247, standard deviation: 0.272), ndka (between 3.01 and 419.19, median: 12.22, standard deviation: 40.216). Continuous variables are the main focus to investigate and analyze in this case. The simulation Extrapolation (SIMEX) algorithm, known for the correction method, was also performed to ensure the use of the proposed method by showing differences between no correction and correction with SIMEX for each variable. The SIMEX algorithm results can be checked in Table 4.7. Estimates of the measurement error variance were 0.2177 for S100b and 16.086 for NDKA. It can be differed by removing outliers of the covariate. When these measurement errors were taken into account, the significance of the S100b remained highly significant and the significance level of the NDKA slightly increased. The objective is to study the effect of S100b and NDKA when those variables are measured with error. We analyze this data with the proposed method to estimate the ROC curve in the presence of measurement error.

The proposed ROC curve estimator in the presence of measurement error was performed based on the estimated parameters $\phi(\hat{\tau}, \hat{\alpha}, \hat{\beta}, \bar{\theta_m})$ using Maximum Approximate Bernstein Likelihood. Four different estimators of the

| | | S100b | NDKA |
|---|---|---|---|
| No correction | Estimate | 5.334 | 0.031 |
| | SE | 1.263 | 0.016 |
| SIMEX | Estimate | 10.013 | 0.061 |
| | SE | 2.057 | 0.025 |

TABLE 5.1: Regression coefficient estimates based on the estimated variance of measurement error without and with correction for the measurement error

| | Area Under the Curve | | | | |
|---|---|---|---|---|---|
| **Estimator** | **Empirical** | **Binormal** | **Nonparametric** | **Bernstein** | **Maximum Approximate Bernstein Likelihood Estimate (in the presence of the measurement error)** |
| **S100b** | 0.731 | 0.726 | 0.713 | 0.644 | 0.704 |
| **NDKA** | 0.612 | 0.581 | 0.601 | 0.583 | 0.522 |

TABLE 5.2: The estimation of the ROC Curve in the presence of measurement error of variable NLDK of aSAH data

ROC curves, Empirical Estimator, Binormal estimator, Kernel based Nonparametric estimator, and Bernstein Polynomial based Nonparametric estimator which proposed by Wang et al., 2019 of the ROC curve without adjusting measurement error were performed for comparison. The estimators of the ROC curves for S100b and NDKA as covariates are plotted in Figure 5.1 and 5.2. Table 5.3 contains optimal model degree ($m$), $\hat{\tau}$, and BIC using $m = 1, ..., 10$ according to various values of NSR given in Chapter 4, (0.1, 0.2, ..., 0.8), for each variable. The estimated parameters $\bar{\theta}_{m_1}$ are 0.9747, 0.0000, 0.0000, 0.0000, 0.0253 for diseased group of S100b, and the parameters $\bar{\theta}_{m_0}$ for non-diseased group of S100b are 0.000000, 0.999999, 0.000001, 0.000000, 0.000000. The estimated values of the AUC for Empirical Estimator, Binormal estimator, Kernel based Nonparametric estimator, and Bernstein Polynomial based Nonparametric estimator are 0.731, 0.726, 0.713, 0.644 (when $m$ is assigned as 4) for S100b and 0.612, 0.581, 0.601, 0.583 (when $m$ is assigned as 5) for NDKA, respectively.

Figure 5.1 and 5.2 show the proposed Maximum Approximate Bernstein Likelihood Estimator (MABLE which is equivalent to MBLE) of the ROC curve in the presence of the measurement error. The MABLE estimator of the ROC curve is below the diagonal line in Figure 5.2, the Bernstein estimator of the ROC curve performed a similar trend as well. Generally, when the ROC curve gets closer to the diagonal line and being close to the line, it usually means that we do not separate groups very well at that threshold. It could be a data collection issue as well. In this case, the ROC curve is below diagonal and crossing diagonal. In principle, there is nothing wrong with

crossing diagonally. The MABLE and Bernstein estimators of the ROC curve show these estimators think negative samples are actually positive at certain threshold points. A further step for the researcher will be finding the threshold point where the compromise between FPR and TPR is satisfactory in this case.

In conclusion, the results demonstrate visually the effect of taking into account measurement error on the estimation of the ROC curve. This may explain the apparent difference between the estimation of the ROC curve corrected by measurement error and not corrected by measurement error.

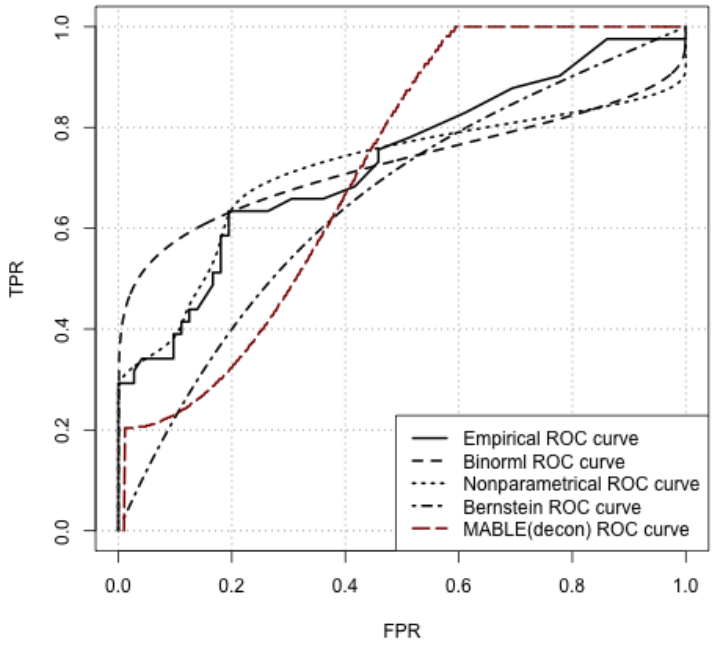



FIGURE 5.1: The estimation of the ROC Curve in the presence of measurement error of variable s100B of aSAH data

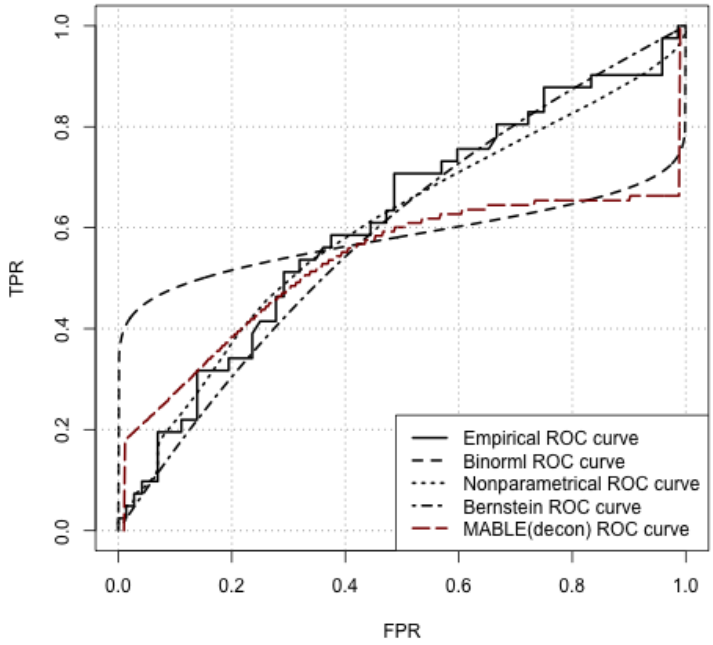



FIGURE 5.2: The estimation of the ROC Curve in the presence of measurement error of variable NLDK of aSAH data

| Variable | Group | Result | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **NSR** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** |
| **S100B** | **diseased** | m | 4 | 4 | **4** | 4 | 4 | 4 | 4 | 4 |
| | | BIC | 28.780 | 41.309 | **27.045** | 41.177 | 27.130 | 41.002 | 27.336 | 40.817 |
| | | $\tau$ | 0.037 | 0.075 | **0.113** | 0.150 | 0.188 | 0.225 | 0.263 | 0.300 |
| | **non-diseased** | m | 3 | 4 | **4** | 4 | 4 | 4 | 4 | 4 |
| | | BIC | 33.303 | 45.814 | **31.550** | 45.682 | 31.635 | 45.506 | 31.841 | 45.322 |
| | | $\tau$ | 0.013 | 0.0262 | **0.039** | 0.052 | 0.065 | 0.078 | 0.092 | 0.105 |
| **NDKA** | **diseased** | m | 5 | 5 | 5 | 5 | 5 | 5 | **5** | 5 |
| | | BIC | 31.655 | 35.376 | 31.318 | 35.521 | 34.811 | 39.127 | **30.886** | 35.396 |
| | | $\tau$ | 6.421 | 12.843 | 19.264 | 25.685 | 32.106 | 38.528 | **44.949** | 51.370 |
| | **non-diseased** | m | **4** | 5 | 5 | 6 | 6 | 5 | 5 | 6 |
| | | BIC | **32.446** | 40.444 | 36.386 | 44.865 | 40.441 | 40.481 | 35.954 | 44.741 |
| | | $\tau$ | **1.280** | 2.561 | 3.841 | 5.122 | 6.402 | 7.683 | 8.963 | 10.244 |

TABLE 5.3: Results of the estimation of the measurement error standard deviation for the two mismeasured assumed covariates. Results referring to the selected values of optimal degrees ($m$ and $n$) and NSR are identified in bold based on BIC selection method

# Chapter 6

# Conclusion

This master's thesis proposed to apply new methodology to estimate the ROC curve when the data is contaminated. We showed how existing MBLE estimator can be applied to estimate ROC curve in the presence of the measurement errors under weak assumption. This last chapter summarizes the most important results, limitations of the proposed methodology and opportunities for further research.

## 6.1 Results Summary

Measuring data with errors are commonly taken place in many scientific fields, however, the effect of ignoring measurement errors is often overlooked by many researchers. Similarly, measurement errors in the estimation of the ROC curve and AUC are largely disregarded. It is important to take into account the measurement errors for the estimation of the ROC curve and AUC especially for the medical domain, a serious consequence will be followed by failing consideration of the measurement errors for the estimation of the ROC curve. In order to acquire valid inference, the measurement errors have to be considered by some bias-correction methods. The limitations of the available few bias-correction methods for the ROC curve and AUC are existed, such as the assumption of density distributions and the availability of validation and auxiliary data. When there is no knowledge of distributional assumption and no availability of validation and auxiliary data, it is usually impossible to estimate the variance of the estimation of the density and the ROC curve. Therefore, in this paper, Maximum approximate Bernstein Likelihood Estimator (MBLE) which is known as a remarkably flexible approximate parametric approach and introduced by *Guan*, 2016 will be applied to the density estimation of the contaminated data and it will be extended to the estimation of the ROC curve in the presence of measurement errors. The simplicity of tuning parameter is the advantage of the proposed method concurrently

with weak assumption, searching an optimal model degree, $m$ and $n$, using EM Algorithm under the assumption that underlying unknown density has a positive lower bound on known compact support and the measurement errors distribution follows Gaussian normal distribution. Then the parameters interest of the model will be chosen by Bayesian Information Criterion (BIC). Once parameters of the model are selected, the density estimation can be obtained by using Bernstein base polynomials which can be extended to the estimation of the ROC curve and AUC.

Our approach performs quite well in many different situations as shown in the results in chapter 4. Especially, when true density has consisted of low standard errors, the model performed quite well. The performance of the proposed model is shown using the real data analysis, the model was conducted without any distributional assumption. The result value of the corrected AUC by the proposed method was lesser than by the other estimators of the ROC curve, the difference between the Empirical estimator of the AUC and the MBLE estimator of the AUC for deconvolution can be shown as the effect of ignoring measurement errors. One concern of the model based on the simulation study is that it had a tendency of underestimation of the standard errors in large NSR settings. Regardless of the lack of distributional assumption, replicated data, and auxiliary information, the proposed estimator allowed us to take into account the measurement errors for the estimation of the ROC curve.

The proposed method in this paper contributes to the easier use of methods for the deconvolution of the density estimation which allowed us to extend to the estimation of the ROC curve in the presence of the measurement errors. Most existing methods in the literature on measurement errors assume that the error variance is known, whereas the proposed method does not require the error variance to be known. However, we assumed that underlying unobserved density has bounded compact support and the distribution of the errors is assumed to be Gaussian. This assumption should be considered before utilizing the proposed method. These restrictions might limit the scope of the application since the assumptions might not hold in practice.

## 6.2 Limitations

The limitations of the model are primarily two things. First, the model depreciates the standard errors when the data sample consisted of large standard

errors. Second, the model might estimate less accurately since the model is conducted by estimating the density $f_1$ and $f_0$ with the nonparametric approach for both densities. In Statistics, nonparametric density estimation is tough work. It will be even more difficult to estimate precisely two densities with nonparametric approach for small data.

## 6.3 Future Work

In order to cope with the previously mentioned limitations, several improvements can be applied in the future. Alternative optimization algorithms could be a way to obtain more accurate variance estimation and the estimation of the ROC curve in the presence of measurement errors. To overcome the second limitation, the two-sample semiparametric density ratio model is an option for further work to apply using the proposed method.

The aforementioned advantage of the proposed method is that information of the measurement error is not required. Also, these advantage allows us to estimate ROC curve with contaminated data under the weak distributional assumption. It provides a huge benefit of conducting a simple, speedy, and flexible estimation of the ROC curve in the presence of measurement error. However, the advantages challenge dubiousness and instability. Therefore, future work has to focus on surmounting those challenges.

# Bibliography

Babu, G Jogesh, Angelo J Canty, and Yogendra P Chaubey (2002). "Application of Bernstein polynomials for smooth estimation of a distribution and density function". In: *Journal of Statistical Planning and Inference* 105.2, pp. 377–392.

Baulch, Bob (2002). "Poverty monitoring and targeting using ROC curves: examples from Vietnam". In:

Berens, Hubert, George G Lorentz, and Robert E MacKenzie (1972). "Inverse theorems for Bernstein polynomials". In: *Indiana University Mathematics Journal* 21.8, pp. 693–708.

Bernstein, Serge (1912). "Demo istration du th'eoreme de Weierstrass fondee sur le calcul des probabilites". In: *i* 13.1, pp. 1–2.

Bertrand, Aurélie, Ingrid Van Keilegom, and Catherine Legrand (2019). "Flexible parametric approach to classical measurement error variance estimation without auxiliary data". In: *Biometrics* 75.1, pp. 297–307.

Buonaccorsi, John P (2010). *Measurement error: models, methods, and applications*. CRC press.

Cai, T. and C. S. Moskowitz (2004). "Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test". In: *Biostatistics* 5.4, 573–586. DOI: 10.1093/biostatistics/kxh009.

Carroll, Raymond J and Peter Hall (1988). "Optimal rates of convergence for deconvolving a density". In: *Journal of the American Statistical Association* 83.404, pp. 1184–1186.

Carroll, Raymond J et al. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Claeskens, Gerda, Nils Lid Hjort, et al. (2008). "Model selection and model averaging". In: *Cambridge Books*.

Coffin, Marie and Shashikala Sukhatme (1996). "A parametric approach to measurement errors in receiver operating characteristic studies". In: *Lifetime data: Models in reliability and survival analysis*. Springer, pp. 71–75.

— (1997). "Receiver operating characteristic studies and measurement errors". In: *Biometrics*, pp. 823–837.

Collinson, P (1998). "Of bombers, radiologists, and cardiologists: time to ROC". In: *Heart* 80.3, pp. 215–217.

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

Faraggi, David (2000). "The effect of random measurement error on receiver operating characteristic (ROC) curves". In: *Statistics in Medicine* 19.1, pp. 61–70.

Faraggi, David and Benjamin Reiser (2002). "Estimation of the area under the ROC curve". In: *Statistics in medicine* 21.20, pp. 3093–3106.

Ghosal, Subhashis et al. (2001). "Convergence rates for density estimation with Bernstein polynomials". In: *The Annals of Statistics* 29.5, pp. 1264–1280.

Gonçalves, Luzia et al. (2014). "ROC curve estimation: An overview". In: *REVSTAT–Statistical Journal* 12.1, pp. 1–20.

Guan, Zhong (2016). "Efficient and robust density estimation using Bernstein type polynomials". In: *Journal of Nonparametric Statistics* 28.2, pp. 250–271.

— (2021). "Fast nonparametric maximum likelihood density deconvolution using Bernstein polynomials". In: *Statistica Sinica* 31.2.

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

Hsieh, Fushing and Bruce W Turnbull (1996). "Nonparametric and semiparametric estimation of the receiver operating characteristic curve". In: *The annals of statistics* 24.1, pp. 25–40.

Krzanowski, Wojtek J and David J Hand (2009). *ROC curves for continuous data*. Crc Press.

Leblanc, Alexandre (2010). "A bias-reduced approach to density estimation using Bernstein polynomials". In: *Journal of Nonparametric Statistics* 22.4, pp. 459–475.

Liao, Peizhou, Hao Wu, and Tianwei Yu (2017). "ROC curve analysis in the presence of imperfect reference standards". In: *Statistics in biosciences* 9.1, pp. 91–104.

Lloyd, Chris J (1998). "Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems". In: *Journal of the American Statistical Association* 93.444, pp. 1356–1364.

— (2002a). "Estimation of a convex ROC curve". In: *Statistics & probability letters* 59.1, pp. 99–111.

Lloyd, Chris J (2002b). "Theory & Methods: Semi-parametric estimation of ROC curves based on binomial regression modelling". In: *Australian & New Zealand Journal of Statistics* 44.1, pp. 75–86.

Lloyd, Chris J and Zhou Yong (1999). "Kernel estimators of the ROC curve are better than empirical". In: *Statistics & Probability Letters* 44.3, pp. 221–228.

Lorentz, George G (2013). *Bernstein polynomials*. American Mathematical Soc.

MartinThoma (2020). *File:Roc-draft-xkcd-style.svg — Wikimedia Commons, the free media repository*. [Online; accessed 20-March-2021]. URL: `https://commons.wikimedia.org/w/index.php?title=File:Roc-draft-xkcd-style.svg&oldid=491003296`.

Owen, Art B (2001). *Empirical likelihood*. CRC press.

Peng, Liang and Xiao-Hua Zhou (2004). "Local linear smoothing of receiver operating characteristic (ROC) curves". In: *Journal of Statistical Planning and Inference* 118.1-2, pp. 129–143.

Qin, Jing and Biao Zhang (1997). "A goodness-of-fit test for logistic regression models based on case-control data". In: *Biometrika* 84.3, pp. 609–618.

— (2003). "Using logistic regression procedures for estimating receiver operating characteristic curves". In: *Biometrika* 90.3, pp. 585–596.

Reiser, Benjamin (2000). "Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves". In: *Statistics in medicine* 19.16, pp. 2115–2129.

Rosner, Bernard, Shelley Tworoger, and Weiliang Qiu (2015). "Correcting AUC for measurement error". In: *Journal of biometrics & biostatistics* 6.5.

Schisterman, Enrique F et al. (2001). "Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error". In: *American Journal of Epidemiology* 154.2, pp. 174–179.

Schröder, Christopher and Sven Rahmann (2017). "A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification". In: *Algorithms for Molecular Biology* 12.1, pp. 1–12.

Schwarz, Maik and Sébastien Van Bellegem (2010). "Consistent density deconvolution under partially known error distribution". In: *Statistics & probability letters* 80.3-4, pp. 236–241.

Stefanski, Leonard A and Raymond J Carroll (1990). "Deconvolving kernel density estimators". In: *Statistics* 21.2, pp. 169–184.

Stirnemann, J. et al. (2012). "Deconvolution density estimation with adaptive methods for a variable prone to measurement error". In: URL: `https://cran.r-project.org/web/packages/deamer/index.html`.

Tosteson, Tor D et al. (2005). "Measurement error and confidence intervals for ROC curves". In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 47.4, pp. 409–416.

Turck, Natacha et al. (2010). "A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage". In: *Intensive care medicine* 36.1, pp. 107–115.

Vexler, Albert, Enrique F Schisterman, and Aiyi Liu (2008). "Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures". In: *Statistics in medicine* 27.2, pp. 280–296.

Vitale, Richard A (1975). "A Bernstein polynomial approach to density function estimation". In: *Statistical inference and related topics*. Elsevier, pp. 87–99.

Wang, Dongliang and Xueya Cai (2021). "Smooth ROC curve estimation via Bernstein polynomials". In: *Plos one* 16.5, e0251959.

Wang, Xiao-Feng and Bin Wang (2011). "Deconvolution Estimation in Measurement Error Models: The R Package decon". In: *Journal of Statistical Software* 39.i10. DOI: `http://hdl.handle.net/10.`. URL: `https://ideas.repec.org/a/jss/jstsof/v039i10.html`.

Wang, Xiaoguang et al. (2019). "Nonparametric estimation of the ROC curve based on the Bernstein polynomial". In: *Journal of Statistical Planning and Inference* 203, pp. 39–56. ISSN: 0378-3758. DOI: `https://doi.org/10.1016/j.jspi.2019.02.004`. URL: `https://www.sciencedirect.com/science/article/pii/S0378375819300187`.

Zhang, Biao (2006). "A semiparametric hypothesis testing procedure for the ROC curve area under a density ratio model". In: *Computational statistics & data analysis* 50.7, pp. 1855–1876.

Zhou, Xiao-Hua and Jaroslaw Harezlak (2002). "Comparison of bandwidth selection methods for kernel smoothing of ROC curves". In: *Statistics in medicine* 21.14, pp. 2045–2055.

Zou, Kelly H, WJ Hall, and David E Shapiro (1997). "Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests". In: *Statistics in medicine* 16.19, pp. 2143–2156.