CONTRIBUTIONS TO ROC CURVE AND LIKELIHOOD RATIO ESTIMATION
FOR FORENSIC EVIDENCE INTERPRETATION

by

Xiaochen Zhu
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Statistical Science

Committee:

_____     Dr. Martin P. Slawski, Dissertation Co-Director

_____     Dr. Liansheng L. Tang, Dissertation
                                     Co-Director

_____     Dr. Anand N. Vidyashankar, Committee
                                     Member

_____     Dr. Christopher P. Saunders, Committee
                                     Member

_____     Dr. Wanli Qiao, Co-Director of PhD
                                     Program

_____     Dr. Kenneth S. Ball, Dean, The Volgenau
                                     School of Engineering

Date: _____       Fall Semester 2020
                                     George Mason University
                                     Fairfax, VA

Contributions to ROC Curve and Likelihood Ratio Estimation for Forensic Evidence
Interpretation

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at George Mason University

By

Xiaochen Zhu
Master of Science
George Mason University, 2017
Bachelor of Engineering
Beijing University of Chemical Technology, 2010

Co-Directors: Dr. Martin P. Slawski, Professor & Dr. Liansheng L. Tang, Professor
Department of Statistics

Fall Semester 2020
George Mason University
Fairfax, VA

# Acknowledgments

First of all, I am grateful to my advisors, Dr. Martin P. Slawski and Dr. Liansheng L. Tang, for their invaluable guidance and patient through my research. This dissertation would not have been possible without their guidance during my graduate education. It was a great privilege and honor to work with them and study under their guidance.

Besides, I want to express my sincere gratitude to my committee members, Dr. Anand N. Vidyashankar and Dr. Christopher P. Saunders for their guidance and suggestion.

I am deeply indebted to my lovely wife, Chao Liu, and my daughter, Abigail Z. Zhu for their support, understanding, and encouragement.

I am extremely grateful to my family for their love and support.

# Table of Contents

# List of Tables

# List of Figures

# Abstract

CONTRIBUTIONS TO ROC CURVE AND LIKELIHOOD RATIO ESTIMATION FOR FORENSIC EVIDENCE INTERPRETATION

Xiaochen Zhu, PhD

George Mason University, 2020

Dissertation Co-Directors: Dr. Martin P. Slawski & Dr. Liansheng L. Tang

Biometric traits such as faces and fingerprints are critical in forensic evidence interpretation. There has been a growing interest to study the repeatability, reproducibility, and accuracy of modalities for forensic evidence interpretation since the 2009 National Research Council report and the more recent report from the 2016 President's Council of Advisors on Science and Technology. In this dissertation, we investigate two statistical methods for forensic evidence interpretation, which are score-based likelihood ratio (SLR) and receiver operating characteristic (ROC) curve.

In the first part of this dissertation, we investigate the repeatability and reproducibility of three existing statistical methods for estimating the SLR including parametric estimation, kernel density estimation, and recently adopted logistic regression estimation. We perform extensive simulations and used different face and fingerprint biometric datasets to investigate the repeatability and reproducibility of the existing SLR estimation methods. We also provide a parametric way to estimate the variance of the SLR based on the ROC curves. Simulation studies and real studies are provided to indicate the usefulness of the variance estimation method.

In the second part, we consider modeling of ROC curves using both the order constraint and covariates associated with each score given that the latter (e.g., demographic characteristics of the underlying subjects) often have a substantial impact on discriminative accuracy. The proposed method is based on the indirect ROC regression approach using a location-scale model, and quadratic optimization is used to implement the order constraint. The statistical properties of the proposed order-constrained least squares estimator are studied. Several situations are discussed in the simulation studies, including multiple covariates, non-Gaussian random samples, and heteroscedastic modeling. The results of the simulation studies corroborate the superior performance of the proposed approach. Its practical usefulness is demonstrated applying face recognition data from the "Good, Bad, and Ugly" face challenge, a domain in which accounting for covariates has hardly been studied.

## Notation Table

| | |
|---|---|
| $X \sim P$ | random variable $F$ follows the distribution $M$ |
| $N(\mu, \sigma^2)$ | normal distribution with mean $\mu$ and variance $\sigma^2$ |
| $U(a,b)$ | uniform distribution with the lower and upper boundaries $a$ and $b$, respectively |
| $\mathrm{Beta}(a,b)$ | Beta distribution with two shape parameters $a$ and $b$ |
| $\mathrm{Bernoulli}(p)$ | Bernoulli distribution with probability $p$ |
| $t(d)$ | t distribution with degree of freedom $d$ |
| $\Phi(\cdot)$ | the cumulative density function of a standard normal distribution |
| $\phi(\cdot)$ | the probability density function of a standard normal distribution |
| $\phi'(\cdot)$ | first derivative of the probability density function of a standard normal distribution |
| $N$ | total sample size |
| $m$ | sample size for the genuine group |
| $n$ | sample size for the imposter group |
| $T$ | test scores |
| $X$ | set of covariates |
| $\mathbf{x}$ | observation of $X$ |
| $p$ | number of covariates |
| $q$ | number of constraints in the order-constrained weighted least square regression, $q = 2^p$ |
| $\mathbf{A}$ | $q$ by $d$ constraint matrix, where $d = 2(p+1)$ |
| $\mathbf{y}$ | dependent variable in the regression, $\mathbf{y} = \{T_i\}_{i=1}^N$ in our study |
| $H_0$ | defense hypothesis; a pair of samples are from different sources |
| $H_1$ | prosecutor hypothesis a pair of samples are from the same source |
| $D$ | binary status with $D = 1$ indicating $H_1$ is true and $D = 0$ indicating $H_0$ is true |
| $\mathbf{X}$ | $N$ by $d$ design matrix, $\mathbf{X} = (1, \mathbf{x}^\top, D, (\mathbf{x} * D)^\top)^\top$ |
| $\mathbf{W}$ | weight matrix in the weighted least square regression |
| $\boldsymbol{\beta}$ | vector of regression coefficients |
| $\mathbf{b}_0, \mathbf{b}_1$ | coefficient parameters for the location function of the imposter and genuine group, respectively |
| $\mathbf{a}_0, \mathbf{a}_1$ | coefficient parameters for the scale function of the imposter and genuine group, respectively |
| $\Delta$ | mean difference of two random variables |
| $||a||_2$ | $l_2-$norm of $a$, i.e. $||a||_2 = \sqrt{\sum_{i=1}^N a_i^2}$ |

# Chapter 1: Introduction

In this chapter, we start by introducing the background knowledge about forensic investigations and identifications. We will discuss the log-likelihood ratio method, which is an important biometric technique for measuring the strength of the forensic evidence. Furthermore, we will introduce the receiver operating characteristic curve which is a graphical tool to evaluate the discriminative accuracy of classification methodologies.

## 1.1 Forensic Evidence Interpretation

Forensic problems aim to determine the link between the known evidence (e.g., suspect) and an unknown evidence (e.g., evidence from the crime scene). The common-source problem relates to whether two evidences are from the same source. A decision will be provided on the following two propositions for how the evidence has arisen:

$H_0$: The unknown and known evidence are from different sources;

$H_1$: The unknown and known evidence are both from a common source.

Biometrics, such as DNA, fingerprint, facial features, and voice are characteristics that are different from person to person, and have been used for automatic human identification and recognition. Since the 19th century, law enforcement gradually adopted biometrics to determine the identity of criminals [2]. Studies in fingerprint and face recognition have been conducted to systematically quantify the accuracy of forensic examiners and computer algorithms. Ulery *et al.* [3, 4] generated binary decisions of individualization or exclusion on *genuine* print pairs (a pair of biometric measurements from the same source, e.g., a pair of fingerprints of the same individual) and *imposter* pairs (a pair of biometric measurements from different sources). Phillips *et al.* [1] compared the accuracy of facial recognition algorithms with forensic examiners, students, as well as untrained residents.

Typically, a biometric score is generated from the biometric recognition algorithm to help forensic scientists drop their decision. The score indicates the degree of similarity between measurements from a crime scene evidence and measurements from a person of interest. The higher the comparison score is, the higher the probability that two samples are from the same source. Take the recent black box[1] study by Phillips *et al.* [1] as an example. In this facial recognition study, image pairs (see Figure 1.1) from males and females were selected and a biometric algorithm was adopted to grade each image pair with a similar score based on the likelihood of whether the pairs belong to the same person.



Figure 1.1: Left pair: same identity; right pair: different identities [1].

The most common way to make the decision when using the biometric score is to find a threshold to biometric score values: scores greater than the cut point imply that the prosecution hypothesis, $H_1$, is likely to be true and false otherwise (see Figure 1.2) [5]. However, selecting the score threshold is not a trivial matter and it depends on the scale and distribution of the genuine and imposter scores. Alternatively, the log-likelihood ratio has been proposed as a way of normalizing comparison scores and quantifying the relationship between two samples being compared, which are evidence or unknown samples from a crime scene and known samples from a person of interest.

---

[1]The term "black box" here refers to the fact that the way in which biometric measurements are converted into scores by computer-based algorithms or a human examiner is not always fully transparent.
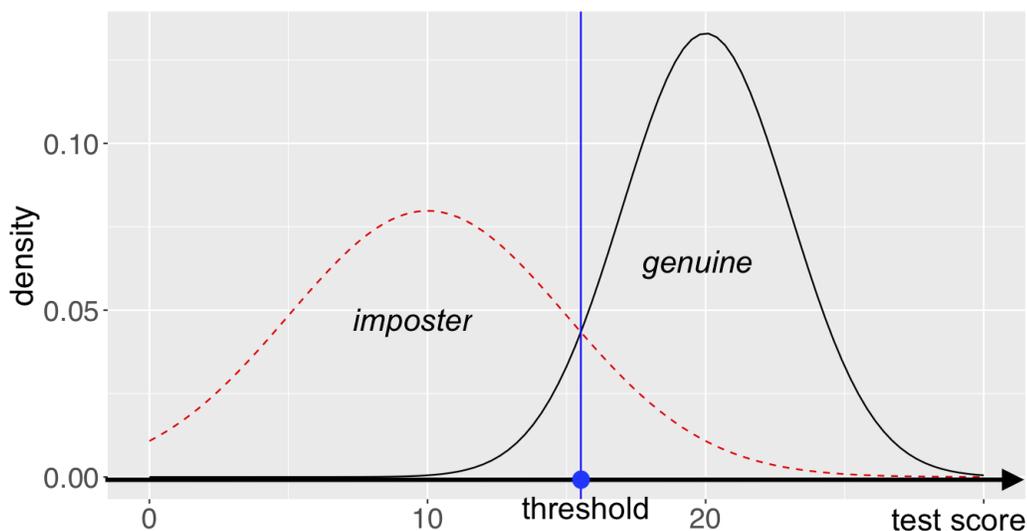
Figure 1.2: Continuous test scores and a threshold to separate the genuine and imposter scores.

## 1.2 Log-Likelihood Ratio Method

Log-likelihood ratio (LLR) is the log of the ratio of two joint probabilities if the measurements are available [6]. Throughout this thesis, the log-transformation is the natural logarithm. When the measurements are continuous comparison scores (e.g., Figure 1.2), the LLR is the ratio of the probability density function of the genuine comparison scores against the imposter comparison scores [7,8]. LLR methods have been proposed as a measure of the strength of evidence [9] in several forensic disciplines, such as DNA, fingerprint, facial recognition, and voice identification [10–13].

The increased visibility of likelihood ratio (LR) methods grows the need to further study the accuracy as well as repeatability and reproducibility of LLR methods. Both the 2009 National Research Council report [14] and the more recent report from the 2016 President's Council of Advisors on Science and Technology [15] emphasized these required properties for a valid forensic method. The repeatability and reproducibility of examiners from papers

3

[3, 4] are defined as follows.

- Repeatability: intra-method agreement; that is, a statistical forensic method provides consistent likelihood ratio values for the same distributions of mated and non-mated scores with various sample size ratios from mated and non-mated groups;

- Reproducibility: inter-method agreement; that is, two statistical forensic methods provide consistent likelihood ratio values on the same set of data.

Since we want to minimize the number of defendants who had been wrongfully convicted based on evidence, recent papers put their focus on how to improve the results in different biometric identification systems. Mandasari *et al.* [16] developed an improvement method for voice identification to deal with background noise. They also pointed out that adopting this voice filtering software to clean the single performs better than using a set of noisy background training data. Neumann *et al.* [17] worked on the feasibility and accuracy of the latent print comparison and the effect of sample sizes, but this study has not discussed the effect of varied sample size ratios between mated and non-mated groups in the training data set. Poh *et al.* [18] defined a trustworthy LLR which considered the probability of the trustworthiness term, but further discussion about how to improve the LLR calculation methods based on the trustworthiness term is not included in his study.

Whereas these studies developed new methodologies to improve the accuracy of the identification procedures, they talk little about the assessment of the discriminative ability. This is crucial since the evaluation report gives a chance for people to glance at the reliability of the biometric techniques. The accuracy of biometric systems or humans in source matching problems can be assessed with the receiver operating characteristic (ROC) curve when the decision scores are ordinal or continuous.

## 1.3 Receiver Operating Characteristic Curves

The receiver operating characteristic (ROC, see Figure 1.3) curve plays an important role in characterizing errors for binary decisions. While ROC analysis originated in signal detection

in radar, the analysis has been adopted and extended in the fields of psychology, diagnostic medicine, biometrics, and forensic sciences. In the context of medical diagnostics, the false positive rate (FPR) denotes the probability of a positive test result given the individual is healthy, and the true positive rate (TPR) denotes the probability of a positive test result given the individual is diseased. The ROC curve graphs the trade-off between the FPRs and TPRs resulting from binarization of a continuous test result in dependence of different thresholds. Discussions of applications in diagnostic medicine, medical imaging and signal detection can be found in various textbooks [19, 20] and research papers [21–26].
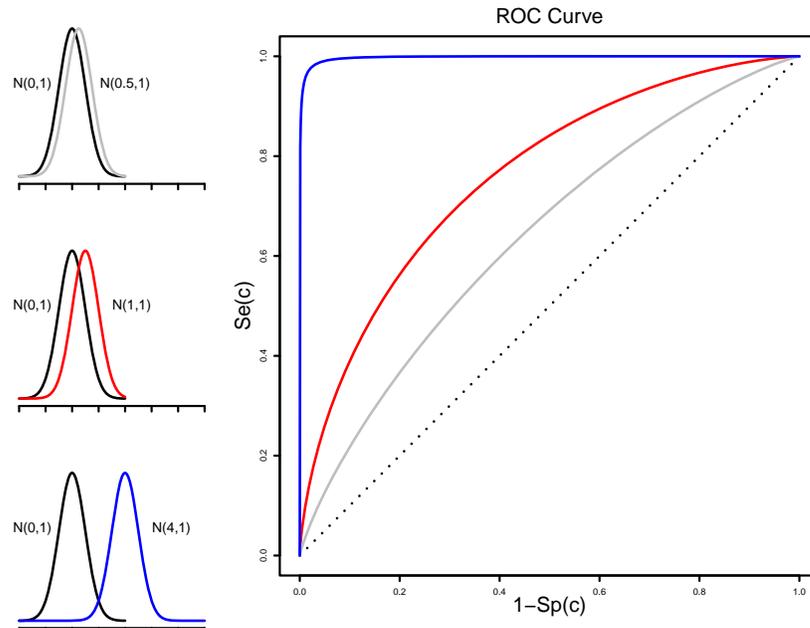


Figure 1.3: Relationship between densities (left panel) and ROC curves (right panel).

The error rates and the ROC curves can be applied to the latent print, facial recognition, or other biometric studies. The FPR and TPR in the medical studies are analogous to the false accept rate (FAR) and true accept rate (TAR), respectively. Hendricks *et al.* [27] derived the likelihood ratio from the ROC curve to handle some common issues in the

fingerprint data such as the curse of dimensionality. Srihari *et al.* [28] used the ROC curve to determine the optimal threshold of the comparison scores in the fingerprint verification study.

The discriminative accuracy of diagnostic tests is potentially influenced by characteristics of individual patients or by specific properties of the process in which the test result was obtained. For example, (i) the severity of a patient's disease might affect the accuracy of a medical test, (ii) the reader's performance is known to be a key factor in medical imaging studies. In light of this, the "raw" ROC curve without incorporating such factors can be of limited use.

### 1.3.1 Covariate-Specific ROC Curves

The covariate-specific receiver operation characteristic curve is a ROC curve that depends on the value of covariate. In the real world, covariates can affect the result of tests, which makes the ROC hard to evaluate without being given the covariates. To address this shortcoming, Pepe [29] introduced the notion of *covariate-specific* ROC curves. Currently, there are various strategies for modeling covariate-specific ROC curves. Those methods include nonparametric or semiparametric [30, 31] Bayesian methods and induced or direct-regression methods [32, 33], etc. González-Manteiga *et al.* [34] used the kernel method along with empirical distribution function to estimate the conditional ROC curves. Duan & Zhou [35] provided a induced-regression method with the location function estimated by the composite quantile regression. Tang & Zhou [36] introduced a semiparametric method to estimate the covariate-specific ROC curves.

One particularly convenient as well as popular approach is the location-scale model [34, 35]. In a nutshell, the distribution of the score is modeled via location-scale transformation of base distributions associated with the two populations, where potentially both the location and scale transformation depend on covariates. A common way to estimate the latter is via regression techniques with least squares (LS) regression as the simplest approach. However, a well-known shortcoming of LS is sensitivity to heavy-tailed error

distributions, and composite quantile regression (CQR) [37] is proposed as an alternative by Duan & Zhou [35].

A recent study on an operational fingerprint database [38] found that fingerprint decision scores vary with subjects' demographic information such as age and gender. In face recognition, demographic variables also affect the discrimination accuracy of face recognition algorithms. Phillips *et al.* [1] studied face identification accuracy of examiners from ordinal decision scores on genuine and imposter face image pairs, and another facial recognition study [39, 40] on the "Good, Bad, and Ugly" (GBU) Face challenge presented by NIST investigated the accuracy of continuous decision scores generated by the Vision Geometry Group (VGG)-face algorithm [41] under various imaging conditions. It is thus of interest to account for source subjects' covariate information including their demographics when estimating the ROC curve. Figure 1.4 provides the ROC curves with pooled data and the covariate-specific empirical ROC curves. This figure illustrates that the ROC curves can be noticeably influenced by the covariate.
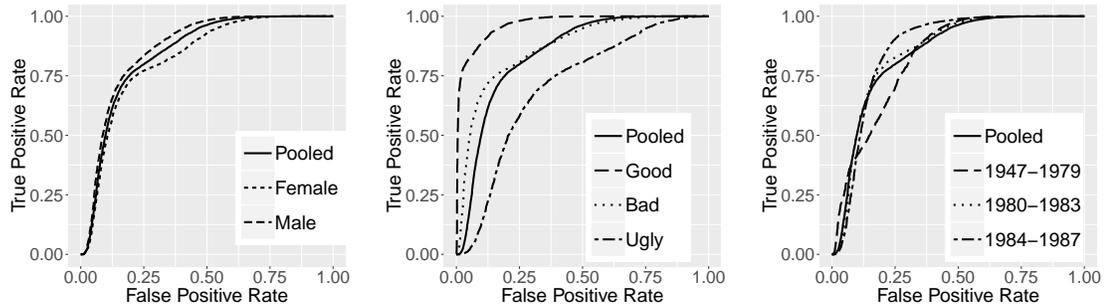


Figure 1.4: ROC curves for the GBU data. Left: Gender-specific ROC; Middle: quality-specific ROC; Right: age-specific ROC.

Additionally, using pooled error rates to estimate the ROC curve instead of considering the covariate might elevate the error and underestimate the accuracy of the modality. Figure 1.5 provides an example of using the covariate-specific ROC and the pooled ROC to evaluate

a classification algorithm. In this study, two important covariates are incorporated as factors that influence the discriminate accuracy of the algorithms. In Figure 1.5, we note that the pooled ROC curve is lower than both of the covariate-specific ROC curves. This implies that the covariates lead to a different evaluation and interpretation of the algorithms.
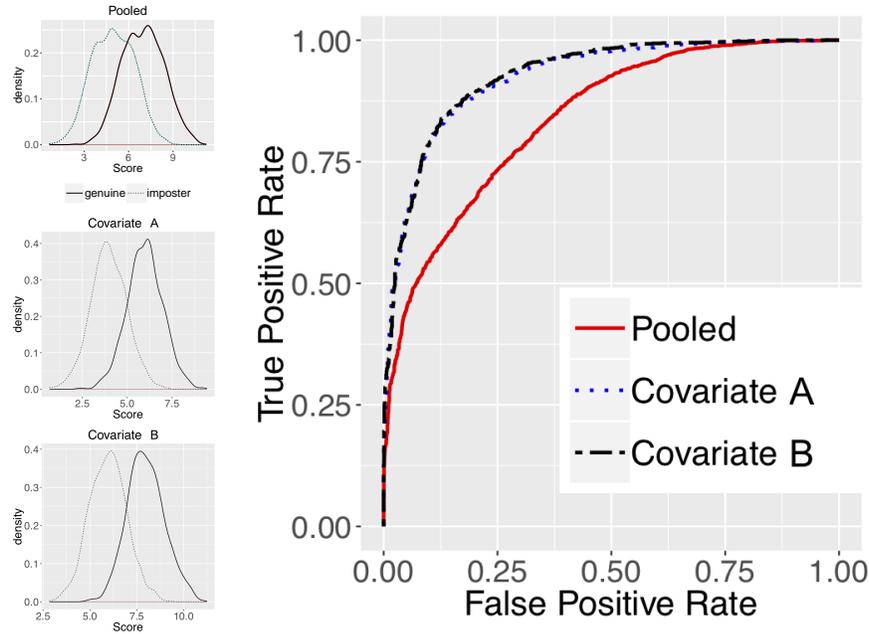


Figure 1.5: Covariate-specific ROC curves and pooled ROC curve. Left panel: distribution of pooled scores and unpooled scores; right panel: corresponding ROC curves.

### 1.3.2 Stochastic Ordering

In addition to the need to account for covariate information, another important property of biometric data is that genuine scores are usually larger than imposter scores in the sense of a *stochastic ordering* [42, 43]. Research on ordering constraints has mainly been done for testing equality in distribution against the alternative hypothesis in which these distributions are stochastically ordered [44–48]. Davidov & Herman [49] focused on the estimation for the area under the ordinal dominance curve (ODC) by imposing stochastic

ordering to improve estimation efficiency. Chen *et al.* [50], Yu *et al.* [51], and Westling *et al.* [52] studied estimation under *likelihood ratio ordering*, a stronger notion than stochastic ordering. The stochastic ordering assumption is reasonable in biometric evaluation since the computer algorithms used in the GBU study and other biometric matching studies have been developed to ensure larger genuine scores than imposter scores. In the case of human examiners, the ordering assumption is also reasonable since examiners are trained to match biometrics. However, to the best of our knowledge, the discussion of ROC regression methods with order restriction has been scarce. The method developed in this thesis will fill this methodology gap concerning order-restricted inference for ROC curves while accounting for covariate information.

The rest of this thesis is structured as three chapters as follows:

Chapter 2 compares existing statistical methods for estimating the log-likelihood ratio from biometric scores including parametric estimation, kernel density estimation, and recently adopted logistic regression estimation. We discuss the repeatability and reproducibility of these methods on biometric datasets. We perform extensive simulations and use different face and fingerprint biometric datasets to investigate the repeatability and reproducibility of the existing log-likelihood ratio estimation methods.

Chapter 3 considers modeling of ROC curves using both the order constraint and covariates associated with each score given that the latter (e.g., demographic characteristics of the underlying subjects) often have a substantial impact on discriminative accuracy. The proposed method in this chapter is based on the indirect ROC regression approach using a location-scale model, and quadratic optimization is used to implement the order constraint. The statistical properties of the proposed order-constrained least squares estimator are studied. Simulation studies are provided to show the superior performance of the proposed approach. The practical usefulness of the order-constrained method is demonstrated in an application of face recognition data from the "Good, Bad, and Ugly" face challenge, a domain in which accounting for covariates has hardly been studied.

Chapter 4 provides an approach of covariate-specific ROC curve estimation with a

stochastic ordering constraint. This method uses the location-scale model with an ordered restriction on the location function, and the location parameter is estimated based on the composite quantile regression. The proposed method is not only insensitive to the heavy-tailed non-Gaussian distribution data, but also has improved statistical efficiency in terms of the mean squared error. We also discuss the covariate-specific ROC estimation in the existence of the heteroscedastic model and provide a comparison of several methods. Simulation studies demonstrate the advantage of our methods and we apply all the methods to the data from a facial recognition study.

## Chapter 2: Repeatability and Reproducibility of Forensic Likelihood Ratio Methods when Sample Size Ratio Varies

## 2.1 Introduction

Biometrics techniques are widely implemented in forensic investigations including face recognition, gait analysis, iris identification, and fingerprint recognition, etc.. The biometric features are different from person to person and have been used for forensic identifications and verifications. A biometric recognition algorithm compares a pair of samples and generates a comparison score that indicates the similarity between the two samples [53]. When a pair of images are from the same source, for example, the same finger, we call the comparison score either a genuine or mated score. When a pair of images are from different sources, we call the comparison score either an impostor or non-mated score. These comparison scores are separated into either a genuine group or an imposter group.

Biometric comparison scores play an important role in forensic determination on whether and how a crime scene evidence is related to a suspect. Two types of hypotheses are considered in forensic determination: prosecutor hypothesis ($H_1$) that the crime scene evidence is from the person of interest, and defense hypothesis ($H_0$) that the crime scene evidence is not from the person of interest. $D$ denotes a binary status with $D = 1$ indicating $H_1$ is true and $D = 0$ indicating $H_0$ is true A comparison score $t = s(I_1, I_2)$ generated by a biometric recognition algorithm $s(\cdot)$ indicates the degree of similarity between measurements from a crime scene evidence, $I_1$, and measurements from a person of interest, $I_2$. The higher the comparison score is, the higher the probability of the two samples are from the same source.

Generally, forensic scientists implement the log-likelihood ratio to calibrate test scores and describe the similarity between the samples from a crime scene and the samples from the suspect. Log-likelihood ratio (LLR) is the natural logarithm of the ratio of two joint

probabilities $\text{LLR}(I_1, I_2) = \log(P(I_1, I_2|D = 1)/P(I_1, I_2|D = 0))$, if the measurements $I_1$ and $I_2$ are available [6]. If the only available measurements from $I_1$ and $I_2$ are their continuous comparison scores, then LLR is the ratio of probability density function of genuine comparison scores and that of imposter comparison scores: $\text{LLR}(t) = \log(P(t|D = 1)/P(t|D = 0))$ [7,8].

Note that in this Chapter, we adopt the score-based likelihood ratios [54]. This method is to avoid considering the unknown nuisance parameters by using the marginal likelihood ratio. An alternative approach is called Bayesian integrated likelihood ratios which replaces the unknown parameters by using the prior probability of the parameters. We should note that the score-based likelihood ratios, likelihood ratios, and the Bayesian integrated likelihood ratios are three distinct likelihood methods, with the score based methods not necessarily having a direct correspondence to formal Bayesian definition of evidential value. The prior distribution in the Bayes factor comes from an assumption of the nuisance parameter before the observation of the evidence to be interpreted. The score-based likelihood ratio method treats the comparison score as an ad-hoc approach to dealing with the nuisance parameters in the style of Likelihoodist paradigm of statistics. (See Davis *et al.*for a discussion of these issues with respect to forensic evidence [55].) In general, the score-based likelihood ratio is estimated only based on a subset of the observed data instead of the entire evidence.

After the 2009 National Research Council report [14] and the 2016 President's Council of Advisors on Science and Technology [15] emphasized the requirement of the accuracy, repeatability, and reproducibility for a forensic procedure, there is a noticeable growth of studies about those properties of LLR methods. This work aims to address that need. Repeatability and reproducibility are defined as follows. Repeatability is intra-method agreement; that is, a given statistical forensic evaluation method provides consistent likelihood ratio values for the same distributions of genuine and imposter groups but various sample size ratios from the two groups. Reproducibility is inter-method agreement; that is, two statistical forensic evaluation methods provide consistent likelihood ratio values on the

same set of data. We use the same definition of reproducibility and repeatability as Suki *et al.* [56] to consider the sample size ratio between genuine and imposter groups.

Our contributions include several aspects. First, we discuss the repeatability and reproducibility of parametric estimation, kernel density estimation, and logistic regression estimation in extensive simulation studies. We compare the log-likelihood ratio values generated by those methods with various sample size ratios, various total sample sizes, and various probability distributions. Second, we discuss and provide statistical reasoning for the inconsistencies between existing methods base on the results of our simulation studies. In addition, we compare the performance of the parametric estimation, kernel density estimation, and logistic regression estimation methods in two biometric examples, including facial recognition and fingerprint matching.

The rest of this chapter is structured as follows. Section 2.2 introduces existing methods for log-likelihood ratio estimation. Section 2.2.4 compares the repeatability and reproducibility of kernel density estimation and logistic regression estimation methods using simulation studies, and provides reasons for inconsistencies among different methods, especially for the logistic regression estimation method. Further studies on the repeatability and reproducibility of the log-likelihood ratio methods using real biometric datasets are given in Section 2.4. The discussion is provided in Chapter 5.

## 2.2 Existing Likelihood Ratio Methods

Recall that a log-likelihood ratio (LLR) is the ratio of two probability density functions of continuous genuine and imposter comparison scores: $\text{LLR}(t) = \log(f(t|D = 1)/f(t|D = 0))$, where $f(t|D = 1)$ and $f(t|D = 0)$ are probability density functions (PDF) of genuine and imposter groups, respectively. The current methods for estimating LLRs from comparison scores mainly focus on the estimation of these density functions, and include parametric estimation [57], kernel density estimation (KDE) [11], and logistic regression estimation (LRE) methods [13]. Some recent papers have studied repeatability and reproducibility of

these methods through simulated studies [53] and biometric datasets [56]. Ali *et al.* [53] suggested that the LRE method is the least sensitive to the sampling variability in large sample sizes, and is more accurate in small sample sizes than the other two methods. However, the sensitivity of LRE to the sampling variability increases when the sample size decreases. Also, the paper shows that the bias of LLR only depends on the shapes of the distributions and not on sample sizes – changing the sample size did not affected the LLR values. The limitation of their study is that it only focuses on a constant sample size ratio, but the ratio may be a crucial factor affecting the resulting LLR values. Suki *et al.* [56] focused on repeatability and reproducibility of LLR methods using KDE and LRE in biometric systems.Their result shows that the estimated LLRs from KDE and LRE are similar when the biometric score is close to zero, and are quite different otherwise. Furthermore, confidence intervals for the result of the two methods are narrower when the biometric score is close to zero, and become wider otherwise. The authors conclude that LLR methods using KDE and LRE are repeatable and reproducible only when the biometric scores are within a certain range. They also did not consider varying sample size ratios.

Suppose we have two groups of scores, $T_1$ and $T_0$, as scores under prosecutor hypothesis and defense hypothesis, respectively. $m$ is the sample size for the genuine group or under the prosecutor hypothesis, and $t_{1,i}$ is the $i$-th score from the genuine group. Similarly, $n$ is the sample size for the imposter group or under the defense hypothesis, and $t_{0,j}$ is the $j$-th score from the imposter group. A detailed discussion of log-likelihood ratio (LLR) estimation methods base on parametric estimation, kernel density estimation and logistic regression estimation follows.

### 2.2.1 Parametric Estimation

Parametric estimation (PE) is based on the distribution assumptions. The distribution parameters are estimated from the training samples [57]. If we can assume that the data follows specific distributions such as normal distributions, mixture normal distributions, or other types of distributions, the parameters can be estimated for the specific distributions

using standard statistical methods. For example, assuming the data are from normal distributions, that is, $T_1 \sim N(\mu_1, \sigma_1^2)$ and $T_0 \sim N(\mu_0, \sigma_0^2)$, then we can estimate means $(\mu_1, \mu_0)$ and variances $(\sigma_1^2, \sigma_0^2)$ with sample means and sample variances for $T_1$ and $T_0$, respectively. $f(t|D = 1)$ and $f(t|D = 0)$ can then be estimated and the log likelihood ratio (LLR) is written as

$$\text{LLR}(t) = \log\left(\frac{\hat{\sigma}_0}{\hat{\sigma}_1}\right) + \log\phi\left(\frac{\hat{\mu}_1 - t}{\hat{\sigma}_1}\right) - \log\phi\left(\frac{\hat{\mu}_0 - t}{\hat{\sigma}_0}\right),$$

where $\phi$ is the standard normal PDF. Since parameter estimates are obtained separately from genuine and imposter groups, the sample size ratio, $m/n$, does not affect the consistency of LLR calculation.

### 2.2.2   Kernel Density Estimation

A kernel density estimation (KDE) method first estimates kernel density functions $f(t|D = 1)$ and $f(t|D = 0)$, for genuine and imposter groups, and then calculates the LLR by simply using the ratio of $f(t|D = 1)$ and $f(t|D = 0)$ [58, 59]. The density at a score $t$ is estimated using the the frequency counts at $t$ and the neighboring region. The contribution of scores in the neighboring region to the density depends on the kernel function. The KDE for $f(t|D = 1)$ under $H_1$ can be written as: $\widehat{f}(t|D = 1) = 1/(hn_1)\sum_{i=1}^{m} K((t - t_{1,i})/h)$. Here $K(\cdot)$ is a kernel function with the kernel bandwidth $h$. A Gaussian kernel function is written as $K(a) = \exp(-\frac{1}{2}a^2)/\sqrt{2\pi}$, and $K((t - t_{1,i})/h))$ gives more weights to the scores $t_{1,i}$ in the neighboring region of $t$. The KDE under $H_0$ can be written as: $\widehat{f}(t|D = 0) = 1/(hn_0)\sum_{j=1}^{n} K((t - t_{0,j})/h)$. Then the estimated LLR is given by $\text{LLR}(t) = \log(\widehat{f}(t|D = 1)/\widehat{f}(t|D = 0))$. Similar to parametric estimation methods, these kernel density functions are estimated separately from genuine and imposter groups, and sample size ratio has little impact on consistency of LLR calculation.

### 2.2.3 Logistic Regression Estimation

Logistic regression estimation (LRE) is a recently adopted method in speaker recognition for calculation of LLR from comparison scores [13, 60–62]. The expression for the LRE is

$$\text{LLR}(t) = \log\left(\frac{f(t|D=1)}{f(t|D=0)}\right) = \beta_0 + \beta_1 s, \tag{2.1}$$

where $\beta_0$ and $\beta_1$ are the intercept and the slope [56, 61].

In statistics it is well known that if the observations are from a case-control study in which the sample size ratio between cases and controls is fixed by the study design, then the intercept $\beta_0$ becomes a function of the sample size ratio [63]. Relating to biometric studies, the sample size ratio between cases and controls is analogous to the sample size ratio between genuine and imposter groups. Thus, the estimate for $\beta_0$ is likely to change when the sample size ratio varies, and so does the LLR values from LRE methods. It is also known that estimates for the slope parameter, $\beta_1$, remain consistent and valid for different sample size ratios. Prentice & Pyke [64] is one of the first papers to study the properties of parameter estimation in LRE methods in case-control studies. The discussion from that article can help us further understand appropriateness of the LRE method for calculating LLR. Their paper pointed out that in the case-control studies, only $P(t|D=1)$ and $P(t|D=0)$ are valid since we can no longer estimate $P(D=1|t)$ and $P(D=0|t)$. Then $P(t|D=1) = c_1 \exp(\gamma(t) + t\beta_1)$, and $P(t|D=0) = c_2 \exp(\gamma(t))$, where $c_1$ and $c_2$ are functions of the sample size ratio $m/n$ and $\beta_1$. The LLR is thus given by $\text{LLR}(t) = \log(P(t|D=1)/P(t|D=0)) = \log(c_1/c_2) + \log\beta_1 s$. Prentice & Pyke show that the slope $\beta_1$ can be estimated consistently regardless of if or how sample size ratio varies, but $\beta_0 = \log(c_1/c_2)$ depends on the sample size ratio between the genuine and imposter groups. Prentice & Pyke has similar conclusion that the estimated $\beta_0$ is not accurate from the LRE method since it relates to the sample size ratio. As a result, the LLR calculated from a LRE method is likely to differ when the sample size ratio from the genuine and imposter groups changes, and therefore LRE methods may not be repeatable.

Table 2.1: Distributions and their parameters in the three simulation studies

| Study | $F(t|D=1)$ | $F(t|D=0)$ |
|:---:|:---:|:---:|
| 1 | $N(20,3)$ | $N(10,5)$ |
| 2 | $U(0,1)$ | $\text{Beta}(0.8,17)$ |
| 3 | $N(2,2)$ | $t(2)$ |

## 2.2.4 Repeatability and Reproducibility Based on Simulated Datasets

This section describes the simulation studies for repeatability and reproducibility of PE, KDE, and LRE. Since the PE based on correct distribution assumptions should give similar LLR results as KDE, we only compare the results between LRE and KDE to show the impact of sample size ratio on the LLR values. We use several distributions to generate scores aiming to mimic real biometric comparison scores in reality, and use different sample size ratios to evaluate the repeatability and reproducibility of existing statistical methods. We first describe how the simulation data are generated, and then present the comparison results.

## 2.2.5 Simulated Datasets

The datasets are generated using functions in R [65]. We compare LRE and KDE in three simulation studies with different distributions for genuine and imposter groups. Table 2.1 summarizes the different distributions we simulated. The distributions and parameters in each study are obtained from real data sets [66].

Without loss of generality, we investigate the impact of the sample size ratio on LLR values for a particular comparison score. Similar to Ali *et al.* [53], we chose the score $t_0$, as the score at which the true genuine and impostor probability density functions cross. The true value of LLR at $t_0$ is zero (the ratio of the two probabilities is 1 and so the log likelihood ratio is zero). We then estimate the LLR values at $t_0$ as the sample size ratio

varies.

Let $m$ and $n$ represent the sample sizes of genuine group and imposter group, respectively. Similar to Ali *et al.* [53], we used five different pairs of sample sizes for plotting the histograms of the LLRs. To examine the effect of sample size ratio on LLR results, unlike Ali *et al.* [53], our sample size ratios vary over the five pairs. We use $(m, n)$ pairs of (400, 2000), (2000, 10000), (2000, 2000), (2000, 400), and (10000, 2000), varying the sample size ratio $m/n$ from 1/5, to 5/1. A random sample is generated 5000 times for each pair, and for each set the estimated LLR at $t_0$ is calculated for method evaluation.

We investigate the reproducibility of the LRE and KDE methods using different sample size ratios when genuine and imposter groups follow distributions outlined in Table 2.1. The bias between the true LLR value and estimated LLR values from these two methods are provided. This is a more extensive simulation studies than Ali *et al.* [53].

We consider two settings for varying sample size ratios. First, we fixed the sample size of the imposter group and varied the size of the genuine group from 100 to 10000 in thirty-five steps, and we calculate the biases of LRE and KDE for each sample size ratio. In a second setting, we consider a fixed total (genuine and impostor) sample size of 10000 with seven different sample size ratios.

## 2.2.6    Results

For each set of simulated genuine and imposter comparison scores, KDE and LRE methods are applied to calculate LLRs at $t_0$. The histograms of LLR values from 5000 sets of simulated scores are plotted side by side in Figures 2.1 – 2.3 according to Studies 1 – 3 in Table 2.1. For all three studies and 5 pairs of sample sizes, we notice that KDE and LRE have a large discrepancy in terms of frequency distributions of the resulting LLR values. When the impostor group has a larger sample size than the genuine group ($m/n < 1$), LRE method tends to generate smaller LLRs than KDE method. When the genuine group has a larger sample size ($m/n > 1$), LRE method tends to generate larger LLRs than KDE method. In terms of the variability of LLR values, the three studies show a larger variation

18

for the KDE method than the LRE method. For equal sample sizes ($m/n = 1$), we observe that the overlap of the LLR histograms is larger than when $m/n \neq 1$, and that the pattern of the histograms are different for the three distributions studied as illustrated in Figures 2.1, 2.2, and 2.3. In Study 1 which assumes normal distribution of genuine and impostor comparison scores, LRE gives smaller LLRs than KDE, but there is some overlap in the histograms. In Study 2, which assumes uniform distribution for genuine comparison scores and beta distribution for imposter comparison scores, both methods give similar estimated LLR values, but unlike the KDE method, the estimated LLRs from the LRE method does not center around the true LLR value. In Study 3, with normal genuine comparison scores and T-distributed imposter comparison scores, the center of the histogram for KDE method tends to be shifted to the left of the histogram by the LRE method, indicating that the LLR values tend to be smaller for KDE method.

In addition, for all three different studies, we note that the estimated LLRs from LRE method follow symmetric distributions, even when the genuine or imposter group follows a non-normal distribution. But the LLR histograms generated from KDE method may have strong skewness when the scores are not normally distributed. Particularly, as shown in Figure 2.3, increasing the total sample size, increases the degree of the skewness.

The large discrepancy in the center and variation of the LLR histograms in the three studies, suggest that reproducibility of LLR estimation is not guaranteed for LRE nor for KDE methods.

The repeatability of the PE, KDE and LRE methods is examined by calculating the biases of the estimated LLR values at $t_0$. Specifically, we compute the difference between the calculated LLR values and the true LLR value (which is zero) for different sample size ratios using the simulated comparison scores. The biases as a function of the sample size ratio are plotted in Figures 2.4 and 2.5 for Studies 1 and 3, respectively. A method has good repeatability if the bias is shown to stay close to zero when sample size ratio varies, and does not have good repeatability if the bias tends to deviate from zero with different
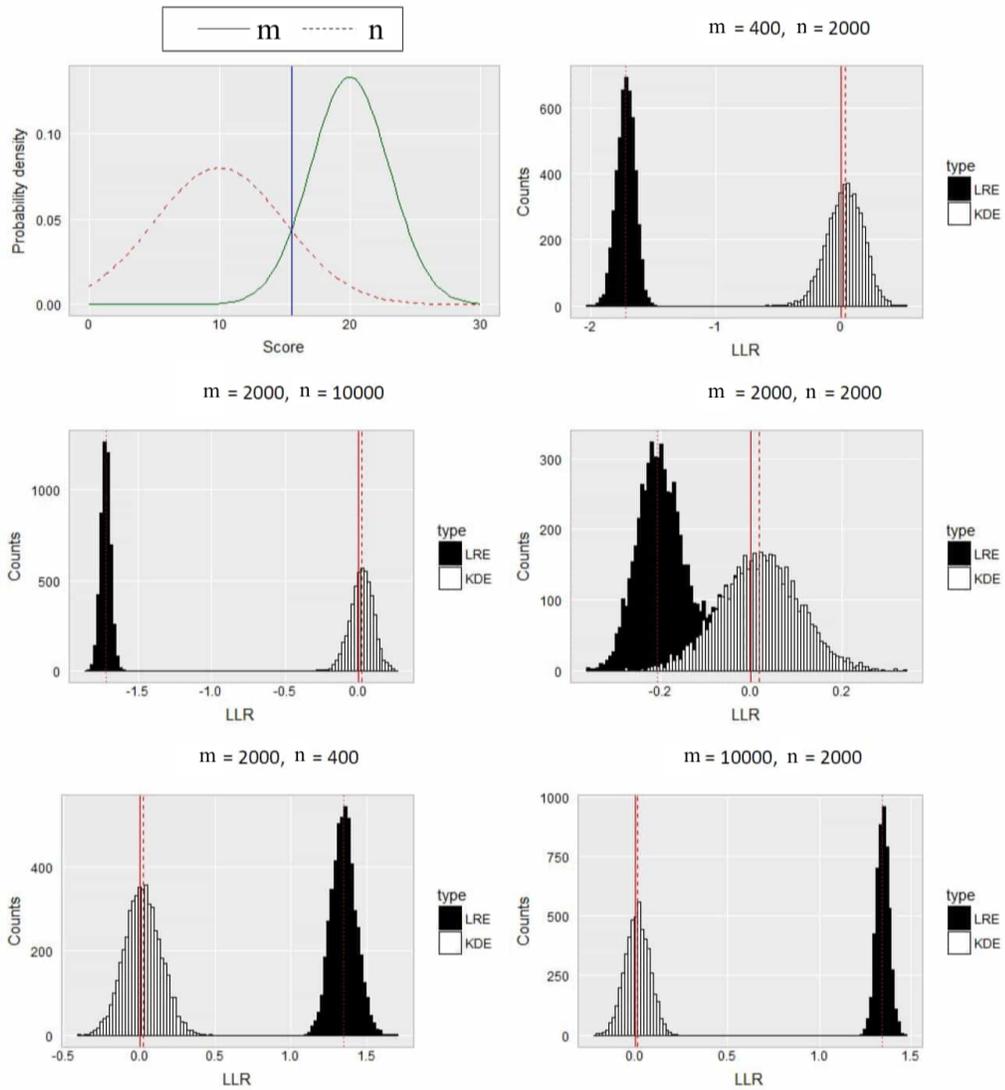
Figure 2.1: Histograms of LLRs at $t_0$ estimated by LRE and KDE for Study 1. The solid vertical line is at score value $t_0$.
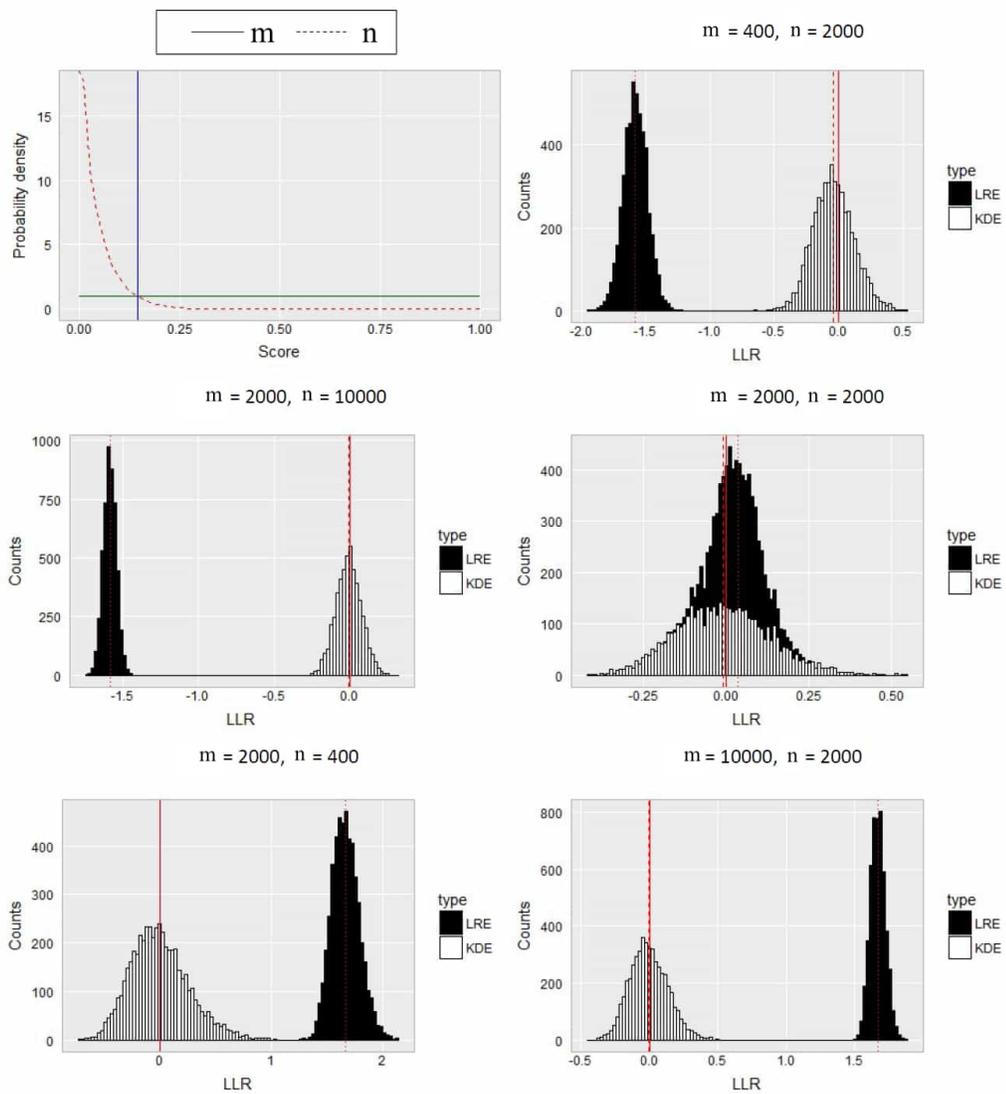
Figure 2.2: Histograms of LLRs at $t_0$ estimated by LRE and KDE for Study 2. The solid vertical line is at score value $t_0$.
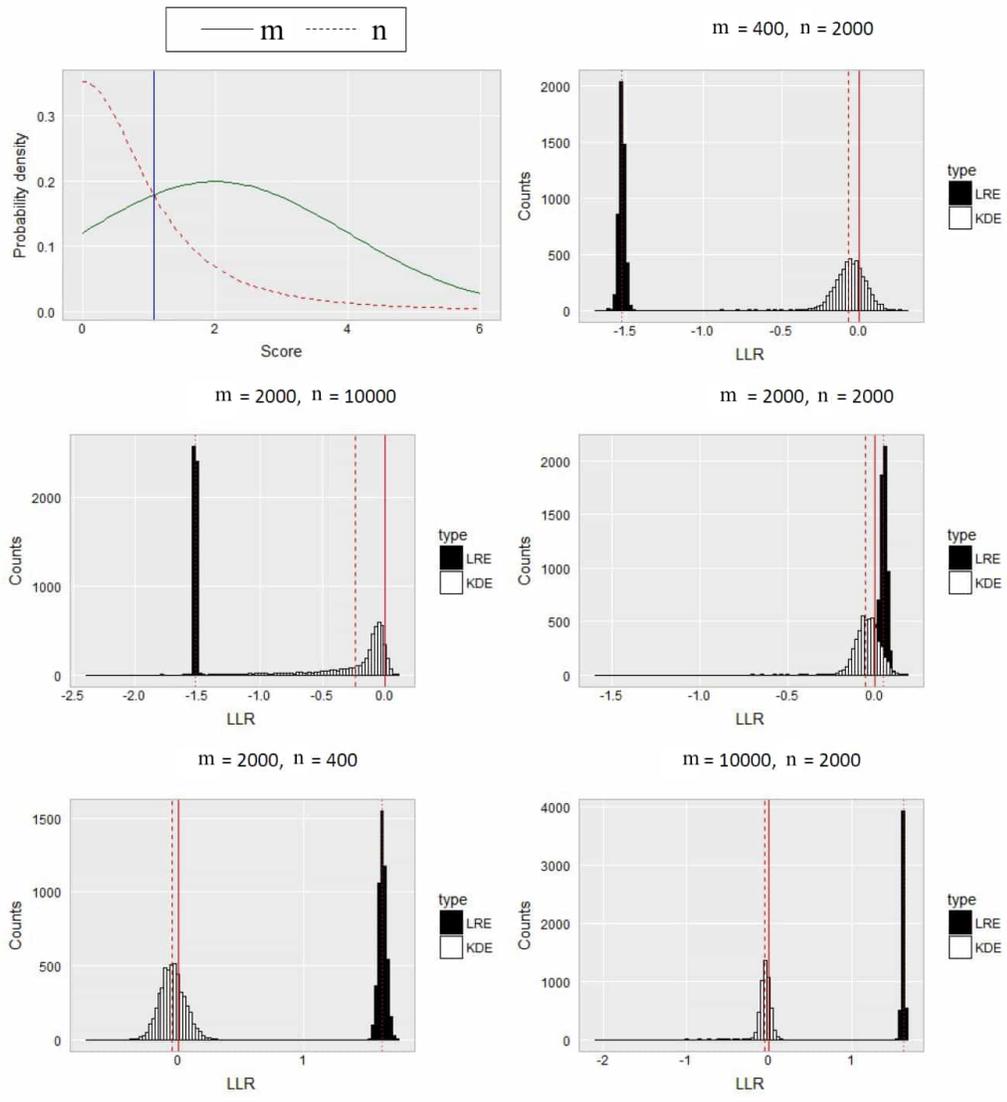
Figure 2.3: Histograms of LLRs at $t_0$ estimated by LRE and KDE for Study 3. The solid vertical line is at score value $t_0$.

sample size ratios.

In Study 1, PE method assumes normal distributions for both the genuine and impostor comparison scores, and LLR values are estimated as outlined in Section 2.2. Figure 2.4 shows that the biases for PE and KDE methods tend to stay close to zero for normal-distributed data in Study 1. This indicates the repeatability of KDE method and PE method with appropriate distribution assumption. More importantly, the consistent biases from KDE and PE methods with various sample size ratios show good repeatability of each of these two methods when the distributions for genuine and imposter scores remain unchanged. The biases for the LRE method show a linear trend as log of the sample size ratio changes from -2 to 2. Thus, the repeatability of LLR values cannot be guaranteed for the LRE method when the sample size ratio varies. The main reason is that the intercept estimate in the LRE method is a function of the sample size ratio, and the resulting LLR is also a function of the sample size ratio.

In Study 3 with a non-normal distribution for imposter comparison scores, the PE method incorrectly assumes a normal distribution for the imposter scores. We see from Figure 2.5 that the PE method tends to give larger biases than the KDE method. The biases by the PE method show departure from zero, indicating that the LLR values are not accurately estimated. However, the biases by the PE method are consistent with different sample size ratios, which indicates repeatability of the PE method even though the method may not yield accurate LLR values. The KDE method gives biases close to zero, indicating accurate LLR estimates no matter how the sample size ratio varies. This shows the repeatability of the KDE method. Biases by the LRE method tend to follow a linear trend as the log of sample ratio changes. The repeatability of LRE method may be unsatisfactory for various sample size ratios even when the distributions for the scores remain unchanged because LLR itself is a function of the sample size ratio.

23
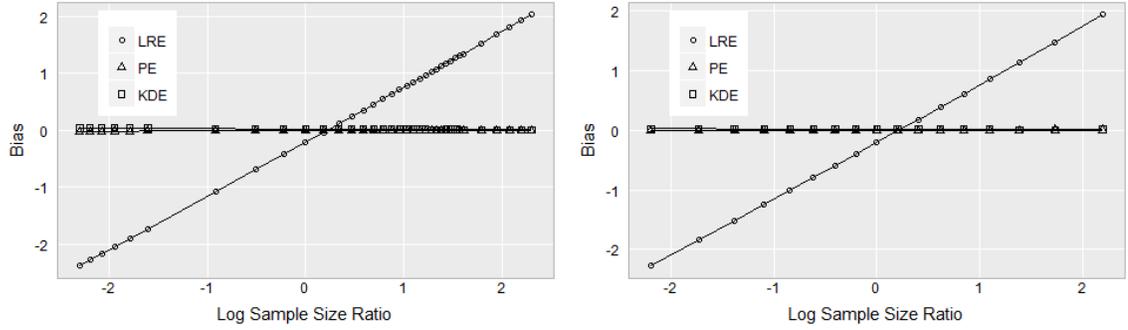
Figure 2.4: Bias of PE, KDE, and LRE methods versus log sample size ratio using different (Left) and same (Right) total sample sizes in Study 1.
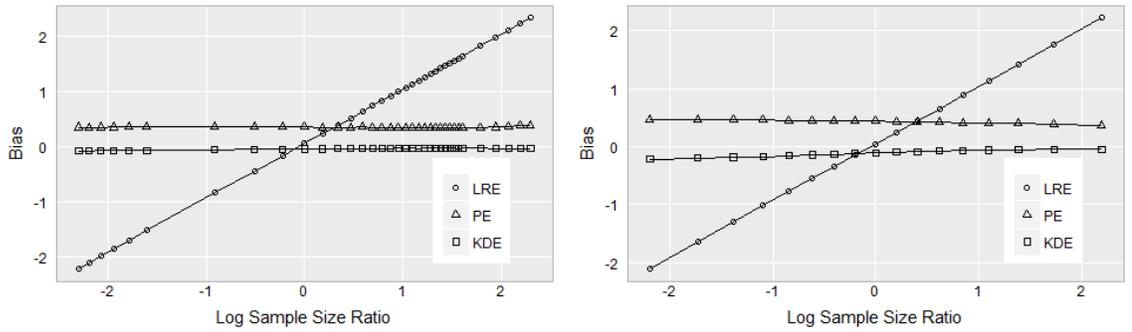


Figure 2.5: Bias of PE, KDE, and LRE methods versus log sample size ratio using different (Left) and same (Right) total sample sizes in Study 3.

### 2.2.7 Unsatisfactory Repeatability of LRE Method

We now discuss the reason that the repeatability of LRE is unsatisfactory when the sample size ratio varied in our studies. Figure 2.6 is the visualization of the variation of the coefficients of the LRE model in Study 1. As the log sample size ratio varies from -2 to 2, the slope $\beta_1$ has little change, while the intercept $\beta_0$ increases in a linear pattern. This is consistent with the equations in Section 2.2.3. Thus, the change of the estimated LLR is mainly due to the difference between estimated $\beta_0$ and the true value of $\beta_0$. Figure 2.6 shows that as the sample size ratio changes, the slope estimate stays close to the true value, but the difference between true value and estimated value of intercept estimate increases. Such a difference is much larger when the log of the sample size ratio takes on values further away from 0.
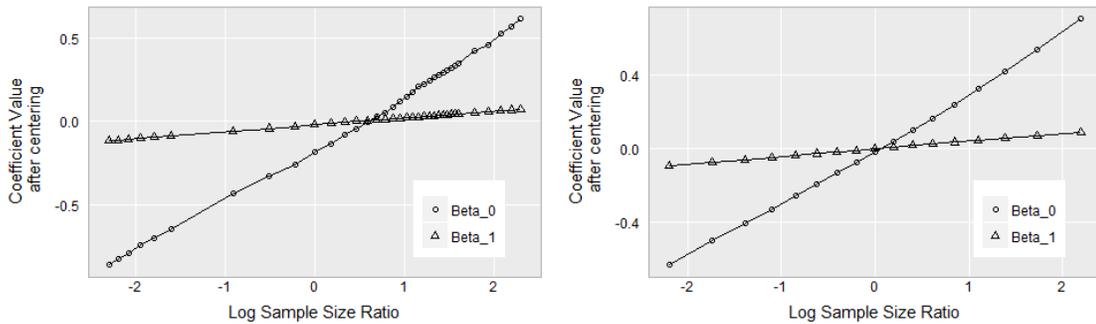


Figure 2.6: The value of coefficients of LRE versus log sample size ratio using different (Left) and same (Right) total sample sizes in Study 1.

## 2.3 Variance of Log-likelihood Ratio Estimation Using Parametric Method Based on Receiver Operating Characteristic Curve

The receiver operating characteristic (ROC) curve is a graphical tool that plots the sensitivity (i.e. probability of identifying a case when the subject is truly diseased) versus one minus specificity (i.e. probability of identifying a case when the subject is not diseased) at different possible thresholds. The ROC curve has been wildly applied in many fields such as radiology and medical imaging research to evaluate the performance of the imaging devices, examiners, and artificial intelligent algorithms. In this section, we will introduce the variance estimation for the log-likelihood ratio based on the receiver operating characteristic (ROC) curve analysis.

### 2.3.1 Estimating the Score-Based Likelihood Ratio Based on ROC Curve

Denote continuous similarity scores for the $i$-th pair of fingerprints as $T_i$, where $i = 1, \ldots, N$. We assume that the first $n$ observations are from imposter group and the remaining $m = N - n$ observations are from genuine group. The test scores $T_i$ in the imposter and genuine group follow distributions $F_0$ and $F_1$, respectively. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) for various thresholds. Let $u$ be the FPR at a given threshold $c$, then the ROC curve is given by

$$\text{ROC}(u) = 1 - F_1(F_0^{-1}(1 - u)),$$

where $u$ is the false positive rate, and $F_j(t) := P(T \leq t | D = j)$, $F_j^{-1}(u) := \inf\{t \in \mathbb{R} : F_j(t) \geq u\}$, for $j \in \{0, 1\}$.

Choi [67] showed that the first derivative of the ROC curve is closely related to the likelihood ratio (LR). Specifically, the tangent of the ROC curve at point $u$ can be written

as

$$\text{ROC}'(u) = \frac{F_1'(F_0^{-1}(1-u))}{F_0'(F_0^{-1}(1-u))} = LR(F_0^{-1}(1-u)),$$

Let $t$ be a realized matching score. If we write $t = F_0^{-1}(1-u)$, it follows that $u = 1 - F_0(t)$, and we then have the mathematical relationship between the LR and the tangent at a point $u$ of the ROC curve that $\text{LR}(t) = \text{ROC}'(1 - F_0(t))$. This applies to companion scores on a continuous scale, which is commonly the case in fingerprint matching and face recognition. The LR can be interpreted as the instantaneous change in the true positive rate in a unit change of $1 - F_0(t)$.

### 2.3.2 Variance Estimation Using Parametric Methods

In a simple setting, after some monotone transformation such as the Box-Cox power transformation [68], the genuine and imposter scores follow the normal distributions $F_1 \sim N(\mu_1, \sigma_1^2)$ and $F_0 \sim N(\mu_0, \sigma_0^2)$, respectively. The ROC curve thus is referred to as the binormal ROC method [22]. It is worth noting that the monotone transformation should be the same for the two groups so that the underlying ROC curve remains unchanged. Without loss of generality, we can write the ROC curve as

$$\text{ROC}(u) = \Phi(a + b\Phi^{-1}(u)),$$

where $a = (\mu_1 - \mu_0)/\sigma_1$ and $b = \sigma_0/\sigma_1$, $\Phi$ is the CDF of the standard normal distribution. Therefore, the first derivative of the ROC curve is given by

$$\text{ROC}'(u) = \frac{b\phi(a + b\Phi^{-1}(u))}{\phi(\Phi^{-1}(u))},$$

where $\phi$ denote the PDF of the standard normal distribution. Note that this gives a function of $u$ instead of $t$. By substituting $u$ with a placement value $1 - \Phi((t - \mu_0)/\sigma_0)$, we have the

27

LR under the binormal model such that

$$\text{LR}(t) = \frac{b\phi(a + b\Phi^{-1}(1 - \Phi((t - \mu_0)/\sigma_0)))}{\phi(\Phi^{-1}(1 - \Phi((t - \mu_0)/\sigma_0)))}. \tag{2.2}$$

It follows from the symmetry of the standards normal density that the numerator of (2.2) can be simplified to be $b\phi(a+b\Phi^{-1}(\Phi((-t+\mu_0)/\sigma_0)))$, or $b\phi(\mu_1/\sigma_1 - t/\mu_1)$, and the denominator can be simplified to $\phi((\mu_0 - t)/\sigma_0)$. The log-likelihood ratio (LLR) at a score $t$ is then given by

$$\text{LLR}(t) = \log \sigma_0/\sigma_1 + \log \phi((\mu_1 - t)/\sigma_1) - \log \phi((\mu_0 - t)/\sigma_0).$$

The estimators for $a$ and $b$ are obtained by substituting the sample means, $\widehat{\mu}_0$ and $\widehat{\mu}_1$, and sample standard deviations, $\widehat{\sigma}_0$ and $\widehat{\sigma}_1$, for the true means and standard deviations: $\widehat{a} = (\widehat{\mu}_1 - \widehat{\mu}_0)/\widehat{\sigma}_1$ and $b = \widehat{\sigma}_0/\widehat{\sigma}_1$. Therefore the LLR is estimated by plugging in the corresponding estimates of mean and standard deviation.

The estimated LLR needs the estimators for the mean and variances separately for the genuine and imposter groups. Denote the parameter vector $\boldsymbol{\theta} = (\mu_1, \sigma_1, \mu_0, \sigma_0)^\top$ and its estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_1, \widehat{\sigma}_1, \widehat{\mu}_0, \widehat{\sigma}_0)^\top$. The first order Taylor expansion on the LLR is written as

$$\widehat{\text{LLR}}(t) \approx \text{LLR}(t) + \nabla^\top \text{LLR}(t)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

where $\nabla^\top \text{LLR} = (\partial \text{LLR}/\partial \mu_1, \partial \text{LLR}/\partial \sigma_1, \partial \text{LLR}/\partial \mu_0, \partial \text{LLR}/\partial \sigma_0)^\top$. The variance of the estimated LLR can be derived using the first order Taylor expansion on the parameter vector or the multivariate Delta method,

$$var(\widehat{\text{LLR}}(t)) \approx \nabla^\top \text{LLR}(t)cov(\widehat{\boldsymbol{\theta}})\nabla \text{LLR}(t), \tag{2.3}$$

where the variance and covariance elements in $cov(\widehat{\boldsymbol{\theta}})$ follow standard expressions. We can

explicit expression of the derivative of LLR which is given by

$$
\nabla \text{LLR}(t) = \begin{pmatrix} \phi'((\mu_1 - t)/\sigma_1/(\sigma_1 \phi((\mu_1 - t)/\sigma_1)) \\ -1/\sigma_1 - (\mu_1 - t)\phi'((\mu_1 - t)/\sigma_1/(\sigma_1^2 \phi((\mu_1 - t)/\sigma_1)) \\ -\phi'((\mu_0 - t)/\sigma_0/(\sigma_1 \phi((\mu_0 - t)/\sigma_0)) \\ 1/\sigma_0 - (\mu_0 - t)\phi'((\mu_0 - t)/\sigma_0/(\sigma_1^2 \phi((\mu_0 - t)/\sigma_0)) \end{pmatrix}^{\top} .
$$

The covariance matrix of $\widehat{\boldsymbol{\theta}}$ under the normal assumption is a diagonal matrix with diagonal elements $(\sigma_1^2/m, \sigma_1^2/(2m-2), \sigma_0^2/n, \sigma_0^2/(2n-2))$. Furthermore, the pointwise $(1-\alpha)\%$ confidence interval at score $t$ is given by

$$
\left( \widehat{\text{LLR}}(t) + \Phi^{-1}(\alpha/2)\sqrt{var(\widehat{\text{LLR}}(t))}, \ \widehat{\text{LLR}}(t) - \Phi^{-1}(\alpha/2)\sqrt{var(\widehat{\text{LLR}}(t))} \right) .
$$

### 2.3.3 Simulation Study

We generate the test scores for the genuine and imposter groups follow the normal distributions $F_1 \sim N(20, 9)$ and $F_0 \sim N(10, 25)$, respectively. Denote $m$ and $n$ be the sample sizes of the genuine group and the imposter group, respectively. We fix the total sample size $N = 10000$ and variate the log sample size ratio $\log(m/n)$ from -2 to 2. For each sample size ratio, we generate the test scores 1000 times. Let $t_0$ be the score value such that $\text{LLR}(t_0) = 0$, which implies that the distributions $F_1$ and $F_0$ across at score $t_0$. For each iteration of the test score generation, we calculate the estimated LLR at $t_0$ using the aforementioned PE method.

We investigate the variance and the coverage of the 95% confidence interval as the log sample size ratio changes. Denote $var_{pe}$ as the true variance that directly calculated using the values of the LLR for the 1000 iterations, and let $var_{roc}$ be the estimated variance using the parametric method based on the ROC curve. The coverage is defined as the ratio of the count of the estimated LLRs that located in the estimated confidence interval to the total

number of iterations. The results are given in Table 2.2.

From Table 2.2, we note that the result using our proposed method based on the ROC curve is close to the true variance of the LLRs. The coverage is close to 95%, which is the expected coverage of our confidence interval, throughout all log sampling ratios in our study. This implies that our proposed method can accurately estimate the variance along with the pointwise confidence interval for the LLR.

Table 2.2: Variance and confidence interval coverage for different values of log sampling ratio.

| $\log(m/n)$ | $var_{pe}$ $(\times 10^{-3})$ | $var_{roc}$ $(\times 10^{-3})$ | coverage (%) | $\log(m/n)$ | $var_{pe}$ $(\times 10^{-3})$ | $var_{roc}$ $(\times 10^{-3})$ | coverage (%) |
|---|---|---|---|---|---|---|---|
| -1.9 | 2.45 | 2.55 | 94.3 | 0.1 | 0.83 | 0.83 | 95.1 |
| -1.8 | 2.26 | 2.29 | 94.6 | 0.2 | 0.82 | 0.81 | 95.2 |
| -1.7 | 2.09 | 2.06 | 94.4 | 0.3 | 0.81 | 0.82 | 94.9 |
| -1.6 | 1.94 | 1.93 | 95.0 | 0.4 | 0.81 | 0.81 | 95.1 |
| -1.5 | 1.80 | 1.73 | 95.3 | 0.5 | 0.81 | 0.81 | 95.2 |
| -1.4 | 1.67 | 1.59 | 95.3 | 0.6 | 0.81 | 0.82 | 95.4 |
| -1.3 | 1.55 | 1.48 | 95.4 | 0.7 | 0.82 | 0.83 | 94.8 |
| -1.2 | 1.46 | 1.37 | 95.6 | 0.8 | 0.83 | 0.84 | 94.6 |
| -1.1 | 1.37 | 1.32 | 95.9 | 0.9 | 0.85 | 0.86 | 95.0 |
| -1.0 | 1.29 | 1.26 | 95.9 | 1.0 | 0.87 | 0.88 | 94.9 |
| -0.9 | 1.21 | 1.19 | 95.1 | 1.1 | 0.90 | 0.91 | 95.0 |
| -0.8 | 1.15 | 1.14 | 95.0 | 1.2 | 0.93 | 0.94 | 94.7 |
| -0.7 | 1.09 | 1.10 | 94.9 | 1.3 | 0.96 | 0.98 | 95.0 |
| -0.6 | 1.04 | 1.04 | 94.9 | 1.4 | 1.00 | 1.00 | 95.2 |
| -0.5 | 0.99 | 0.98 | 95.5 | 1.5 | 1.05 | 1.06 | 95.7 |
| -0.4 | 0.96 | 0.94 | 95.6 | 1.6 | 1.10 | 1.11 | 95.3 |
| -0.3 | 0.92 | 0.91 | 95.8 | 1.7 | 1.10 | 1.20 | 95.1 |
| -0.2 | 0.89 | 0.88 | 95.3 | 1.8 | 1.22 | 1.27 | 95.0 |
| -0.1 | 0.87 | 0.86 | 95.1 | 1.9 | 1.30 | 1.34 | 94.2 |
| 0.0 | 0.85 | 0.84 | 95.3 | 2.0 | 1.38 | 1.43 | 95.1 |

## 2.4 Repeatability and Reproducibility Based on Real Biometric Data

In this section, we investigate the impact of sample size ratios on the LLR calculation from the existing methods. The genuine and imposter scores from facial recognition and

fingerprint matching are used to calculate the LLRs with different sample size ratios.

## 2.4.1 Facial Recognition Data and Results

We use a facial recognition data set, and apply PE, LRE and KDE to investigate the repeatability and reproducibility of these methods. For the PE method, we use the sample means and sample variances of each training data to estimate the means and variances for genuine and imposter groups. We generate two normal distribution density functions base on the estimated parameters. Base on the estimated density functions, we calculate LLR values for each data in the test group. We then observe the number of miscounts as an evaluation of this method. The biometric images were frontal face images taken with a digital single-lense reflex camera. The similarity scores were extracted from the picture comparison, and used in our study as the score. The data set has three categories, which are "good", "bad", and "ugly", based on the quality of the images [69]. We only consider the category "good" in our study. The comparison scores represent measurement of the characteristic difference, and a smaller distance indicates higher similarity. So a low score represent a pair of pictures with high similarity. Then a genuine comparison score in genuine group, measured by comparing two pictures of the same individual, is generally a smaller value than an imposter score, which is measured by comparing pictures of different people. Scores in both groups have extremely large outliers, so we remove all the outlier samples before applying the methods. We use normal distributions for the genuine and imposter comparison scores and obtain the sample mean of 15.2 and the sample standard deviation of 4.6 for the genuine group, and the sample mean of 38.9 and the sample standard deviation of 4.9 for the imposter group. We randomly select 2000 samples from the genuine group, and various numbers of samples from imposter group

Figure 2.7 shows the LLR values for PE, KDE, and LRE methods when the sample size ratio varies. We see that both PE and KDE have good repeatability since they are not sensitive to varying sample size ratios. This is similar to our simulation findings which also show the repeatability of PE and KDE. The repeatability of the LRE method is unsatisfactory

since the LLR from the method takes on a wide range of values. For the reproducibility, the PE method generates much larger LLR values than KDE and LRE methods. For the variance and confidence band, we note that the variance and the width of the confidence interval for the three methods are close. The LLR from PE and KDE methods takes on all positive values, while the LLR from the LRE method takes on both positive and negative values. If one uses zero as a decision threshold to decide whether the score 25 comes from $H_1$ or $H_0$, both PE and KDE can arrive at the same conclusion that the score of 25 likely supports $H_1$ with all positive LLR values. With LLR value ranging from negative values to positive values, the decision by LRE depends on the sample size ratio. When the ratio is as small as -2, the LRE concludes that the score of 25 supports $H_0$, and with the ratio is as large as 2, the LRE supports $H_1$ instead.
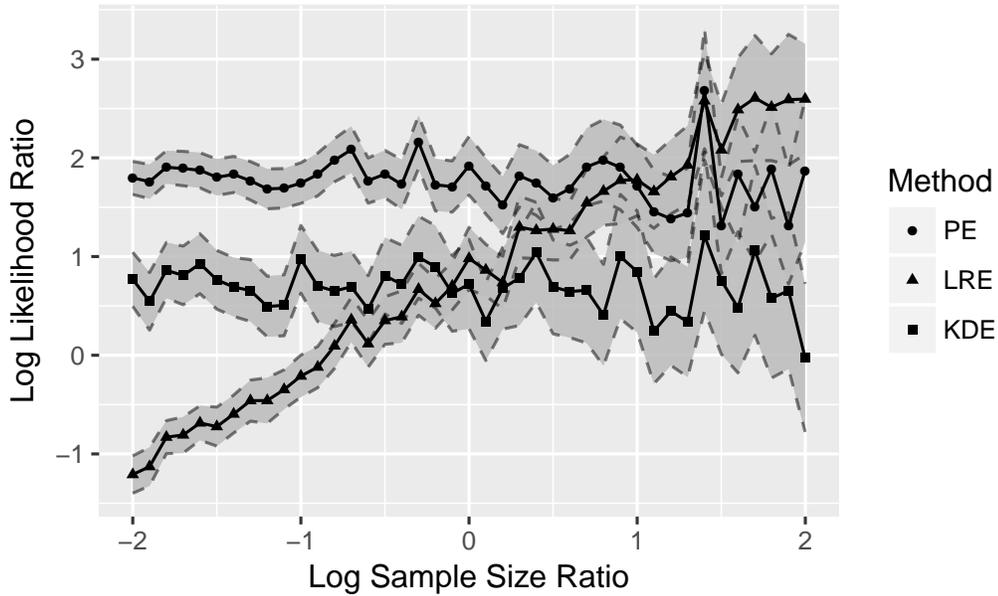


Figure 2.7: Estimated LLR with confidence interval in facial recognize data using PE, LRE, and KDE.

### 2.4.2 Fingerprint Matching Data and Results

We also apply PE, LRE, and KDE to a set of fingerprint comparison scores to study their reproducibility and repeatability for fingerprint recognition. The genuine and impostor comparison scores were generated by applying a fingerprint comparison algorithm in NIST Biometric Image Software (`https://www.nist.gov/services-resources/software/nist-biometric-image-software-nbis`) to National Institute of Standards and Technology Special Database 4. Genuine scores were obtained by comparing the two rolled fingerprints of the same finger, and imposter scores were obtained by comparing rolled fingerprints from two different fingers. The scores in the genuine group are generally greater than the score in the imposter group. The sample means and sample standard deviations are 350.9 and 293.6 for the genuine group, and 7.5 and 2.5 for the imposter group. In our computation of the LLR values using all three methods, we randomly select 4000 genuine scores and various numbers of imposter scores, so that the log sample size ratio ranges from -3 to 3 by the increment of 0.1. When the sample size ratio changes, we repeat the sampling procedure to select genuine and imposter scores before the LLR methods are applied. We calculate the LLR at the score of 10 with all the three methods.

Figure 2.8 shows the LLR values for PE, KDE, and LRE methods when the size ratio varies. In terms of the repeatability, LLR values from PE and KDE methods has small fluctuation when the sample size ratio varies although LLR values from these two methods differ by around 1. The confidence interval also reveals that these two methods have a close result. All the LLR values from these methods are negative. If one uses zero as a decision threshold to decide whether the score 10 comes from $H_1$ or $H_0$, all three methods should arrive at the same conclusion that the score of 10 likely supports $H_0$ with all negative LLR values. Again, the LLR values from LRE have a linear relationship with the log sample size ratio, and thus, the repeatability of the LRE is unsatisfactory. As the log sample size is close to -2, the LLR values from KDE and LRE are similar, indicating reproducibility between the two methods only at this point. The LLR values vary in other sample size ratios, and thus the reproducibility among all three methods are poor.
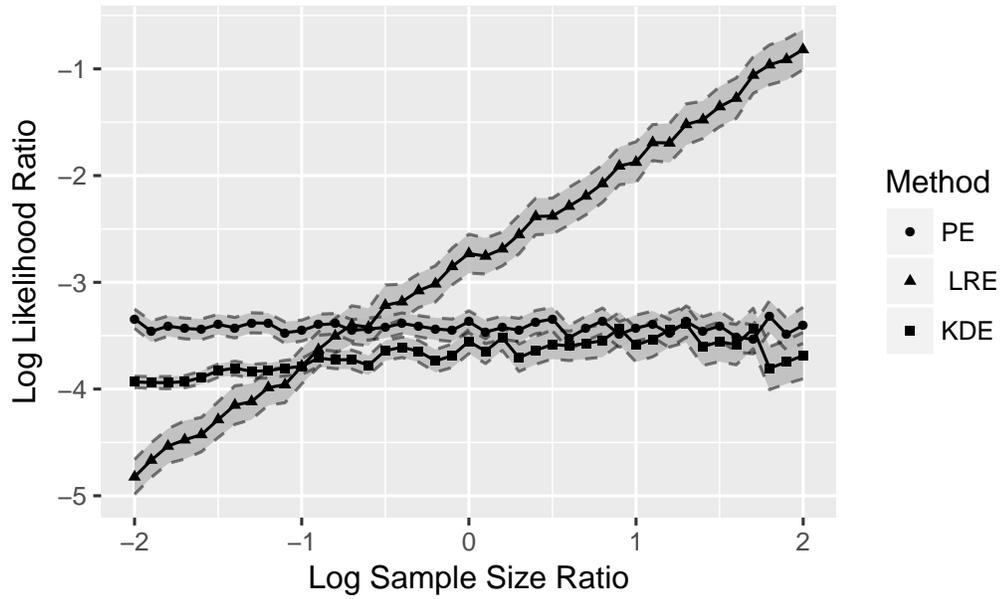
Figure 2.8: Estimated LLR with confidence interval in fingerprint matching data using PE, LRE, and KDE

Although the normal distribution assumption in the PE method for face and finger comparison scores may not be valid, it was useful for demonstrating the issues of repeatability and reproducibility of PE, KDE and LRE methods for estimating LLR, or at a minimum, for illustrating how to evaluate the repeatability and reproducibility of a method.

# Chapter 3: Order-Constrained ROC Regression with Application to Facial Recognition

## 3.1 Introduction

The receiver operating characteristic (ROC) curve is widely used to assess discriminative accuracy of two groups based on a continuous score. The ROC curve graphs the trade-off between sensitivity and specificity in dependence of the decision threshold associated with the score. However, the classification accuracy of diagnostic tests might be impacted by covariates of subjects. The diagnostic accuracy of medical tests is analogous to identification accuracy in biometric traits such as faces and fingerprints. The latter are critical in forensic evidence interpretation. Studies in fingerprint and face recognition have been conducted to systematically quantify the accuracy of forensic examiners and computer algorithms.

The error rates associated with binary decisions are given by the false accept rate (FAR), i.e., the probability of incorrect individualization on imposter pairs, and the false rejection rate (FRR), i.e., the probability of incorrect exclusion from the same source. The FAR and the true accept rate (TAR) are used in analogy to the FPR and TPR, respectively, in medical studies. The error rates and the ROC curves from the aforementioned latent print and face recognition studies are commonly based on a population of subjects with varying demographics. The latter as well the quality of biometric measurements play an important role in identification accuracy.

Another important property of biometric data is that genuine scores are usually larger than imposter scores in the sense of a *stochastic ordering* [42, 43]. In biometric evaluation, the stochastic ordering assumption makes sense because the genuine scores is always larger than the imposter scores.

The rest of this chapter is organized as follows: Section 3.2 introduces the covariate-specific ROC curve and order-constrained modeling. Section 3.3 discusses estimation of covariate-specific ROC curves under the location-scale model with linear location functions subject to order constraints. We then derive statistical properties of the proposed order-constrained least squares estimator in Section 3.4. In Section 3.5 we present the results of simulation studies under different settings. We also present a real data analysis in Section 3.6 in which the proposed method is applied. The conclusion of this chapter and the discussion about potential directions of future research are provided in Chapter 5. All proofs and additional simulation results are contained in the Appendix.

## 3.2 Order-Constrained Modeling for the Covariate-Specific ROC Curve

In this section, we first provide background on the covariate-specific ROC curve, specifically the location-scale model [34], before introducing and motivating an order-constrained modification of this model to be studied in greater depth in Section 3.3. Along the way, we introduce essential notation used throughout this chapter.

### 3.2.1 Covariate-Specific ROC for Source Matching

The accuracy of biometric systems or humans in source matching problems can be assessed with the ROC curve when the decision scores are ordinal or continuous. Source identification problems aim to determine the link between the known evidence (e.g., suspect) and an unknown evidence (e.g., evidence from the crime scene). Let $T$ denote the real-valued random variable associated with a continuous biometric measurement (score) for assessing the above two propositions, and let further $D$ be a $\{0, 1\}$-valued status variable, with $D = 1$ indicating a genuine pair ($H_1$ is true), and $D = 0$ indicating an imposter pair ($H_0$ is true). Let $F_0(t) = P(T \leq t | D = 0)$ and $F_1(t) = P(T \leq t | D = 1)$, $t \in \mathbb{R}$, denote the cumulative distribution functions of $T$ conditional on $D = 0$ and $D = 1$, respectively,

and let further $S_j(t) = 1 - F_j(t)$, $j \in \{0,1\}$, and $F_j^{-1}(u) := \inf\{t \in \mathbb{R} : F_j(t) \geq u\}$, $j \in \{0,1\}$, denote the corresponding survivor and quantile functions, respectively. The receiver operator characteristic (ROC) curve associated with $T$ and $D$ is then defined by

$$\text{ROC}(u) = 1 - F_1(F_0^{-1}(1 - u)), \quad u \in (0,1),$$

where the argument $u$ represents a false accept rate (FAR), so that $\text{ROC}(u)$ returns the true accept rate (TAR) at $u$.

Next, we consider the situation when each measurement of $T$ is accompanied by a set of covariates $X = (X_1, \ldots, X_p)^\top$ that may represent subject demographics, information on quality of the measurement on the process, etc. In this case it is often of interest to examine how the ROC curve varies conditional on observed covariates $X = \mathbf{x}$. Accordingly, we define the covariate-specific ROC curve by

$$\text{ROC}_{\mathbf{x}}(u) = 1 - F_{1,\mathbf{x}}(F_{0,\mathbf{x}}^{-1}(1 - u)), \quad u \in (0,1), \tag{3.1}$$

where $F_{j,\mathbf{x}}(t) := P(T \leq t | D = j, X = \mathbf{x})$, $F_{j,\mathbf{x}}^{-1}(u) := \inf\{t \in \mathbb{R} : F_{j,\mathbf{x}}(t) \geq u\}$, for $j \in \{0,1\}$. In the sequel, we discuss modeling and estimation of the covariate-specific ROC curve (3.1).

## 3.2.2 Location-Scale Model

Various modeling strategies have been proposed for the covariate-specific ROC curve (3.1) [30, 32–34]. In this chapter, we follow the location-scale model proposed in Duan & Zhou's paper [35] that is particularly convenient in view of its modular structure and connection to conventional regression modeling. In this model, the score $T$ is modeled as a location-scale transformation of a base distribution that only depends on the status variable $D$. The location-scale transformation is a function of both the covariates $X$ and the status variable

$D$. Specifically, in its most general form, the model is given by

$$T = \mu(X, D; \boldsymbol{\beta}^*) + D\sigma_1(X; \boldsymbol{\alpha}_1^*)e_1 + (1 - D)\sigma_0(X; \boldsymbol{\alpha}_0^*)e_0. \tag{3.2}$$

The constituents of the above model equation are as follows:

- $e_0$ and $e_1$ are zero-mean and unit variance random variables representing the base distributions for $D = 0$ and $D = 1$, respectively.

- $\mu(X, D; \boldsymbol{\beta}^*)$ is a location function depending on $X, D$, and an unknown parameter $\boldsymbol{\beta}^*$.

- $\sigma_0(X; \boldsymbol{\alpha}_0^*)$ and $\sigma_1(X; \boldsymbol{\alpha}_1^*)$ are non-negative scale functions depending on $X$ and an unknown parameter $\boldsymbol{\alpha}_0^*$ and $\boldsymbol{\alpha}_1^*$, respectively.

It is easy to derive that model (3.2) implies the following for the covariate-specific ROC curve conditional on $\{X = \mathbf{x}\}$ [20]:

$$\mathrm{ROC}_{\mathbf{x}}(u) = 1 - G_1 \left( \frac{\sigma_0(\mathbf{x}; \boldsymbol{\alpha}_0^*)}{\sigma_1(\mathbf{x}; \boldsymbol{\alpha}_1^*)} \left( G_0^{-1}(1 - u) - \frac{\mu(\mathbf{x}, 1; \boldsymbol{\beta}^*) - \mu(\mathbf{x}, 0; \boldsymbol{\beta}^*)}{\sigma_0(\mathbf{x}; \boldsymbol{\alpha}_0^*)} \right) \right), \quad u \in (0, 1),$$

$$\tag{3.3}$$

where $G_0$ and $G_1$ denote the CDFs of $e_0$ and $e_1$, respectively.

Since $\mu(X, D; \boldsymbol{\beta}^*)$ specifies the conditional mean of $T$ given $X$ and $D$, model (3.2) is directly linked to regression modeling. In the simplest setting, $X$ represents a single continuous covariate and $\mu$ is the following linear function in $\boldsymbol{\beta}^* = (\beta_0^*, \beta_D^*, \beta_X^*, \beta_{XD}^*)^\top$:

$$\mu(X, D; \boldsymbol{\beta}^*) = \beta_0^* + \beta_D^* D + \beta_X^* X + \beta_{XD}^* XD. \tag{3.4}$$

Equivalently, the location is modeled via separate straight lines for the two groups defined by the status variable $D$ whose intercept-slope pairs are given by $(\beta_0^*, \beta_X^*)$ and $(\beta_0^* + \beta_D^*, \beta_X^* + \beta_{XD}^*)$, respectively. If the errors of those two regressions are additionally assumed to follow two normal distributions $N(0, \sigma^2)$ and $N(0, \tau^2)$ for $D = 0$ and $D = 1$, respectively, i.e.,

$\sigma_0(X; \boldsymbol{\alpha}_0^*) \equiv \sigma$ and $\sigma_1(X; \boldsymbol{\alpha}_1^*) \equiv \tau$ and $e_0 \sim e_1 \sim N(0,1)$ in terms of (3.2), the covariate-specific ROC curve (3.1) results as

$$\text{ROC}_x(u) = 1 - \Phi\left(\frac{\sigma}{\tau}\left(\Phi^{-1}(1-u) - \frac{\Delta^*(x)}{\sigma}\right)\right), \quad \Delta^*(x) := \beta_D^* + \beta_{XD}^* x. \qquad (3.5)$$

which is a specific bi-normal ROC curve [22]. In (3.5), $\Phi$ denotes the CDF of the $N(0,1)$-distribution.

The above example highlights the modular structure of (3.2) and the fact that the location-scale model integrates popular parametric models for the ROC curve like the bi-normal model. Within model (3.2), it is also straightforward to avoid explicit parametric assumptions regarding the distribution of $T$ even though approximate normality of the latter can often be ensured in practice by suitable data transformations such as the Box-Cox power transformation [70]. In fact, the location-scale model offers a convenient compromise between fully non-parametric and parametric models. It uses parametric models for the location and scale functions which can be estimated by estimating equations [19], while leaving the base distributions, or equivalently, the distributions for the errors of the regression of $T$ on $X$ and $D$ unspecified. Given estimators for the location and scale functions, the covariate-specific ROC can be estimated from the residuals of that regression, which yields empirical CDFs $\widehat{G}_0$ and $\widehat{G}_1$ to be plugged into (3.3).

### 3.2.3 Order-Constrained Modeling

It is often reasonable to assume that the score $T$ is stochastically larger in the population $D = 1$ than in the population $D = 0$, irrespective of covariates. A necessary condition for stochastic ordering conditional on $\{X = \mathbf{x}\}$ is *ordering in mean* (existence provided), i.e., $\mathbf{E}[T|D = 1, X = \mathbf{x}] \geq \mathbf{E}[T|D = 0, X = \mathbf{x}]$, or referred to as the stochastic precedence ordering when comparing normal distributions [45–47,71]. As mentioned above, we consider stochastic ordering uniformly in $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^p$ is the range of the vector of

covariates $X$. In the context of the location-scale model (3.2), this translates to $\mu(\mathbf{x}, 1; \boldsymbol{\beta}^*) \geq \mu(\mathbf{x}, 0; \boldsymbol{\beta}^*)$ for all $\mathbf{x} \in \mathcal{X}$.

*Linear location functions.* In this, we consider the basic case in which $\boldsymbol{\beta}^* = (\theta_0^*, \eta_0^*, \boldsymbol{\theta}^{*\top}, \boldsymbol{\eta}^{*\top})^\top$, $\mu(\mathbf{x}, 0; \boldsymbol{\beta}^*) = \theta_0^* + \mathbf{x}^\top \boldsymbol{\theta}^*$, and $\mu(\mathbf{x}, 1; \boldsymbol{\beta}^*) = \eta_0^* + \mathbf{x}^\top \boldsymbol{\eta}^*$, i.e., the location functions are linear functions in unknown regression parameters $\theta_0^*, \eta_0^* \subset \mathbb{R}$ and $\boldsymbol{\theta}^*, \boldsymbol{\eta}^* \subseteq \mathbb{R}^p$. This corresponds to separate linear regressions of $T$ on $X$ according to the status variable $D$, or equivalently, a single regression model with regressors $X$, $D$, and $X * D :=$ $(X_1 \cdot D, \ldots, X_p \cdot D)^\top$, i.e., the full set of interaction terms between $X$ and $D$; the example (3.4) in the previous subsection corresponds to $p = 1$. Accordingly, with a slight abuse of notation, $\boldsymbol{\beta}^*$ can be re-parameterized as $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\beta}_X^{*\top}, \beta_D^*, \boldsymbol{\beta}_{XD}^{*\top})^\top$ with

$$\beta_0^* = \theta_0^*, \qquad \boldsymbol{\beta}_X^* = \boldsymbol{\theta}^*, \qquad \beta_D^* = \eta_0^* - \theta_0^*, \qquad \boldsymbol{\beta}_{XD}^* = \boldsymbol{\eta}^* - \boldsymbol{\theta}^*.$$

In this setting, the ordering constraint $\mu(\mathbf{x}, 1; \boldsymbol{\beta}^*) \geq \mu(\mathbf{x}, 0; \boldsymbol{\beta}^*)$ for all $\mathbf{x} \in \mathcal{X}$ becomes

$$\beta_D^* + \boldsymbol{\beta}_{XD}^{*\top} \mathbf{x} \geq 0 \quad \text{for all } \mathbf{x} \in \mathcal{X}.$$

To ensure that estimation subject to the above constraint remains computationally tractable, it is appropriate to replace $\mathcal{X}$ by a hyperrectangle $\mathcal{X} \subseteq \mathcal{B} = [l_1, u_1] \times \ldots \times [l_p, u_p]$, where $l_j$ and $u_j$ are lower and upper bounds on the range of the $j$-th covariate, $j = 1, \ldots, p$. By convexity, the constraint $\beta_D^* + \boldsymbol{\beta}_{XD}^{*\top} \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathcal{B}$ is then equivalent to

$$\beta_D^* + \boldsymbol{\beta}_{XD}^{*\top} \mathbf{v}_\ell \geq 0, \quad \ell = 1, \ldots, q = 2^p, \tag{3.6}$$

where the $\{\mathbf{v}_\ell\}_{\ell=1}^q$ denote the $q = 2^p$ vertices of $\mathcal{B}$. Eq. (3.6) can in turn be expressed

equivalently as

$$\mathbf{A}\boldsymbol{\beta}^* \geq \mathbf{0}, \qquad \mathbf{A} := \begin{pmatrix} \mathbf{0}_{p+1} & 1 & \mathbf{v}_1^\top \\ \vdots & \vdots & \vdots \\ \mathbf{0}_{p+1} & 1 & \mathbf{v}_q^\top \end{pmatrix}, \tag{3.7}$$

where $\mathbf{A}$ has $d = 2(p+1)$ columns, and for a vector $\mathbf{u} = (u_j)$, the relation $\mathbf{u} \geq \mathbf{0}$ is understood component-wise, i.e., $\mathbf{u} \geq \mathbf{0} \Leftrightarrow u_j \geq 0$ for all $j$.

*Example 1.* Consider a single covariate ($p = 1$) with range $\mathcal{X} = [0,1]$. In this case, $\mathcal{B}$ has two vertices given by $v_1 = 0$ and $v_2 = 1$. The matrix $\mathbf{A}$ in (3.7) thus becomes

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

which corresponds to the constraints $\beta_D^* \geq 0$ and $\beta_D^* + \beta_{XD}^* \geq 0$.

*Example 2.* Consider a categorical covariate having $L$ levels. Using an encoding by $L-1$ dummy variables, we note that

$$\mathcal{X} = \{\underbrace{(0,\ldots,0)^\top}_{(L-1)\text{ times}}, \mathbf{e}_1,\ldots,\mathbf{e}_{L-1}\}, \quad \mathbf{e}_\ell := (0,\ldots,0,\underbrace{1}_{\text{position } \ell},0,\ldots,0)^\top, \quad \ell = 1,\ldots,L-1.$$

Accordingly, the matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{pmatrix} \mathbf{0}_L & 1 & 0\ldots0 \\ \mathbf{0}_L & 1 & \mathbf{e}_1^\top \\ \vdots & \vdots & \vdots \\ \mathbf{0}_L & 1 & \mathbf{e}_{L-1}^\top \end{pmatrix},$$

41

which corresponds to the constraints

$$\beta_D^* \geq 0, \quad \beta_D^* + \beta_{D*\ell} \geq 0, \;\; \ell = 1, \ldots, L - 1, \tag{3.8}$$

where $\beta_{D*\ell}$ denotes the regression coefficient of the interaction between $D$ and the $\ell$-th dummy variable, $\ell = 1, \ldots, L - 1$.

*Example 3.* Consider the combination of the two previous examples, i.e., one continuous covariate with range $[0, 1]$ and one categorical covariate having $L$ levels. Compared with Example 2, the number of constraints simply doubles according to the two vertices of $[0, 1]$: in addition to (3.8), we obtain the constraints

$$\beta_D^* + \beta_{XD}^* \geq 0, \quad \beta_D^* + \beta_{XD}^* + \beta_{D*\ell} \geq 0, \;\; \ell = 1, \ldots, L - 1.$$

The matrix $\mathbf{A}$ is obtained analogously, and hence omitted for the sake of brevity.

The above three basic examples illustrate how the order constraint translates into simple linear inequality constraints. Considerably more complex configurations of covariates can be treated using the same underlying principles, even though it has to be pointed out that the number of covariates $p$ is assumed to be moderate, since the resulting number of linear inequality constraints generally grows exponentially in $p$. As suggested by a reviewer, an alternative for larger values of $p$ is to impose the constraints in (3.6) for a given set of points $\mathcal{V} \subset \mathcal{X}$ rather than for all $2^p$ vertices of $\mathcal{B}$. A natural candidate for $\mathcal{V}$ is the set of $\mathbf{x}$'s observed in a given sample, or a suitable subset thereof. Note, however, that this alternative approach only ensures that the constraint holds for the convex hull of $\mathcal{V}$ which does not necessarily contain $\mathcal{X}$.

## 3.3 Inference

Building on the previous subsection, we next discuss estimation of covariate-specific ROC curves under the location-scale model with linear location functions subject to the order constraint discussed in §3.2.3.

### 3.3.1 Estimation of Linear Location Functions Subject to Ordering Constraints

Let us recall the regression perspective on the location-scale model (3.2) discussed previously. Assuming linear location functions as in the previous subsection §3.2.3 and constant variances within the two groups defined by $D$, i.e., $\sigma_0(X; \boldsymbol{\alpha}_0^*) \equiv \sigma$ and $\sigma_1(X; \boldsymbol{\alpha}_1^*) \equiv \tau$, the most straightforward approach of estimating the parameter $\boldsymbol{\beta}^*$ given a sample of (score, status, covariate)-triples $(T_i, D_i, \mathbf{x}_i)_{i=1}^N$ is weighted least squares estimation as stated in (3.9) below. Without loss of generality, it is assumed that the first $n$ observations have status $D = 0$, while the remaining $m = N - n$ observations have status $D = 1$. We thus consider

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_2^2, \qquad \mathbf{W} = \text{diag}\Big( \underbrace{1/\sigma^2, \ldots, 1/\sigma^2}_{n \text{ times}}, \underbrace{1/\tau^2, \ldots, 1/\tau^2}_{m \text{ times}} \Big). \qquad (3.9)$$

Here, we have $\mathbf{y} = (T_i)_{i=1}^N$ for the response, $\mathbf{X}_{i\bullet} = (1, \mathbf{x}_i^\top, D_i, (\mathbf{x}_i * D_i)^\top)$ denotes the $i$-th row of the $N \times d$ design matrix $\mathbf{X}$, $i = 1, \ldots, N$, and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_X^\top, \beta_D, \boldsymbol{\beta}_{XD}^\top)^\top$ has $d = 2(p+1)$ components. Throughout the remainder of the , we assume that $\mathbf{X}$ is non-singular. The order constraint in §3.2.3 can be incorporated by imposing suitable linear inequality constraints, i.e.,

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_2^2 \quad \text{subject to } \mathbf{A}\boldsymbol{\beta} \geq \mathbf{0}, \qquad (3.10)$$

which is a quadratic program. Using Lagrangian duality in §5 of Boyd's book [72], the dual of (3.10) is given by Liew [73]

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^q} \frac{1}{2} \boldsymbol{\lambda}^\top \mathbf{H} \boldsymbol{\lambda} + \boldsymbol{\lambda}^\top \mathbf{A} \widehat{\boldsymbol{\beta}}^{\mathrm{WLS}}, \tag{3.11}$$

where $\mathbf{H} = \mathbf{A}(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{A}^\top$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}$ denotes the weighted least squares estimator, i.e., the minimizer of (3.9). Denote by $\widehat{\boldsymbol{\beta}}$ the minimizer of the primal optimization problem (3.10), and by $\widehat{\boldsymbol{\lambda}}$ the minimizer of the dual optimization problem (3.11). The Karush-Kuhn-Tucker (KKT) optimality conditions [72] for (3.10) imply that $\widehat{\boldsymbol{\beta}}$ can be obtained from $\widehat{\boldsymbol{\lambda}}$ via the following relation:

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{A}^\top \widehat{\boldsymbol{\lambda}}. \tag{3.12}$$

This suggests the following scheme for computing $\widehat{\boldsymbol{\beta}}$:

1. Compute the weighted least squares estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}}$.
2. If $\mathbf{A} \widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} \geq \mathbf{0}$, return $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{WLS}}$. Otherwise, compute the minimizer $\widehat{\boldsymbol{\lambda}}$ of the dual problem (3.11) via a non-negative least squares solver [74, 75].
3. Given $\widehat{\boldsymbol{\lambda}}$, compute $\widehat{\boldsymbol{\beta}}$ from (3.12).

Given the availability of numerous highly scalable solvers for non-negative least squares problem that can easily handle thousands of variables, computation becomes straightforward. In the case of a single continuous covariate (cf. Example 1 in §3.2.3), $\widehat{\boldsymbol{\lambda}}$ even has a closed form solution.

*Remark.* In practice, $\sigma^2$ and $\tau^2$ in $\mathbf{W}$ are not known, and thus need to be estimated based on the residuals of an ordinary least squares fit , cf. Eq. (3.14) below. As shown in the Appendix, given the specific properties of $\mathbf{X}$ and $\mathbf{W}$ under consideration here, the minimizer of the *unconstrained* weighted least squares criterion (3.9) equals the ordinary $\sigma^2$

and $\tau^2$ are estimated. However, a corresponding property does *not* hold for the *constrained* counterpart (3.10), i.e., the constrained least squares solution and the constrained, weighted least squares solution differ in general.

### 3.3.2 Estimation of Covariate-Specific ROC Curves

In this subsection, we discuss estimation of the covariate-specific ROC curve based on the location-scale model (3.2) given the estimator $\widehat{\boldsymbol{\beta}}$ in (3.12). We present both a non-parametric and parametric approach. As above, we suppose that the scale functions are constant within the groups defined by $D$, i.e., $\sigma_0(X; \boldsymbol{\alpha}_0^*) \equiv \sigma$ and $\sigma_1(X; \boldsymbol{\alpha}_1^*) \equiv \tau$.

*Non-parametric approach.* Define $\epsilon_0 = \sigma e_0$ and $\epsilon_1 = \tau e_1$, and denote the corresponding CDFs by $H_0(\cdot) = G_0(\cdot/\sigma)$ and $H_1 = G_1(\cdot/\tau)$, respectively. Accordingly, model (3.2) can be written as

$$T = \mu(X, D; \boldsymbol{\beta}^*) + D\epsilon_1 + (1 - D)\epsilon_0.$$

Substituting $\mu(\cdot, \cdot; \boldsymbol{\beta}^*)$ by $\mu(\cdot, \cdot; \widehat{\boldsymbol{\beta}})$, we set

$$\widehat{\epsilon}_i = (1 - D_i)(T_i - \mu(\mathbf{x}_i, 0; \widehat{\boldsymbol{\beta}})) + D_i(T_i - \mu(\mathbf{x}_i, 1; \widehat{\boldsymbol{\beta}})), \ i = 1, \dots, N, \quad (3.13)$$

and estimate $H_0$ and $H_1$ by their empirical counterparts

$$\widehat{H}_0(e) = \frac{1}{n} \sum_{i=1}^{n} I(\widehat{\epsilon}_i \leq e), \qquad \widehat{H}_1(e) = \frac{1}{N - n} \sum_{i=n+1}^{N} I(\widehat{\epsilon}_i \leq e), \ e \in \mathbb{R},$$

where we recall that $D_i = 0$, $1 \leq i \leq n$, and $D_i = 1$ for $n + 1 \leq i \leq N$. Plug-in of $\widehat{H}_0$, $\widehat{H}_1$ and $\mu(\cdot, 1; \widehat{\boldsymbol{\beta}}) - \mu(\cdot, 0; \widehat{\boldsymbol{\beta}})$ into (3.3) yields

$$\widehat{\mathrm{ROC}}_{\mathbf{x}}(u) = 1 - \widehat{H}_1(\widehat{H}_0^{-1}(1 - u) + \mu(\mathbf{x}, 0; \widehat{\boldsymbol{\beta}}) - \mu(\mathbf{x}, 1; \widehat{\boldsymbol{\beta}})),$$

where $\widehat{H}_0^{-1}(u) := \inf\{e \in \mathbb{R} : \widehat{H}_0(e) \geq u\}$.

*Parametric approach.* Alternatively, $G_0$ and $G_1$ may be specified in terms of an explicit parametric model with $G_0(\cdot) = G(\cdot/\sigma)$ and $G_1(\cdot) = G(\cdot/\tau)$, where the choice $G(\cdot) = \Phi(\cdot)$ would correspond to a binormal model. In this case, $\sigma$ and $\tau$ can be estimated by the empirical standard deviations of the residuals (3.13):

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_i^2, \qquad \widehat{\tau}^2 = \frac{1}{N-n}\sum_{i=n+1}^{N}\widehat{\epsilon}_i^2. \tag{3.14}$$

The latter can be plugged into (3.3) along with $\mu(\cdot, 1; \widehat{\boldsymbol{\beta}}) - \mu(\cdot, 0; \widehat{\boldsymbol{\beta}})$ to obain an estimator of the covariate-specific ROC curve. Specifically, in the binormal case (cf. (3.5)) one obtains

$$\widehat{\mathrm{ROC}}_{\mathbf{x}}(u) = 1 - \Phi\left(\frac{\widehat{\sigma}}{\widehat{\tau}}\left(\Phi^{-1}(1-u) - \frac{\widehat{\Delta}(\mathbf{x})}{\widehat{\sigma}}\right)\right), \quad \widehat{\Delta}(\mathbf{x}) := \widehat{\beta}_D + \widehat{\boldsymbol{\beta}}_{XD}^{\top}\mathbf{x}. \tag{3.15}$$

## 3.4  Statistical Properties

In this section, we present basic statistical properties of the order-constrained least squares estimator introduced in §3.3.1 for fixed (non-random) covariates under a bi-normal model with linear location functions, i.e.,

$$T|D = 0, X = \mathbf{x} \sim N(\beta_0^* + \boldsymbol{\beta}_X^{*\top}\mathbf{x},\, \sigma^2), \quad T|D = 1, X = \mathbf{x} \sim N(\beta_0^* + \beta_D^* + (\boldsymbol{\beta}_X^* + \boldsymbol{\beta}_{XD}^*)^{\top}\mathbf{x},\, \tau^2).$$
$$\tag{3.16}$$

Note that for any $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_X^{\top}, \beta_D, \boldsymbol{\beta}_{XD}^{\top})^{\top}$, we have

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}\mathbf{L}\boldsymbol{\beta}, \quad \text{where } \mathbf{Z} = \begin{bmatrix} \mathbf{1}_n & \mathbf{Z}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_m & \mathbf{Z}_1 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} \mathbf{I}_{p+1} & \mathbf{0} \\ \mathbf{I}_{p+1} & \mathbf{I}_{p+1} \end{bmatrix}. \tag{3.17}$$

In the above equation $\mathbf{Z}_0 \in \mathbb{R}^{n \times p}$ has rows $\mathbf{X}_{i\bullet}$, $i = 1, \ldots, n$, corresponding to the covariates of subjects with status $D = 0$, and $\mathbf{Z}_1 \in \mathbb{R}^{m \times p}$ has rows $\mathbf{X}_{i\bullet}$, $i = (n+1), \ldots, N$, corresponding to the covariates of subjects with status $D = 1$. The relations (3.12), (3.16), and (3.17) imply the following result for the distribution of the order-constrained least squares estimator $\widehat{\boldsymbol{\beta}}$.

**Proposition 3.1.** *Suppose that (3.16) holds. Consider the minimizer $\widehat{\boldsymbol{\beta}}$ of the order-constrained weighted least squares problem (3.10) with constraint matrix $\mathbf{A}$. We have that*

$$\widehat{\boldsymbol{\beta}} \stackrel{\mathcal{D}}{=} \boldsymbol{\beta}^* + \sigma \left( \frac{\boldsymbol{\zeta}}{\sqrt{N}} + \boldsymbol{\Omega} \mathbf{A}^\top \operatorname*{argmin}_{\boldsymbol{\lambda} \in \mathbb{R}_+^q} \left\{ \boldsymbol{\lambda}^\top \left( \frac{\boldsymbol{\Delta}^*}{\sigma} + \frac{\boldsymbol{\xi}}{\sqrt{N}} \right) + \frac{1}{2} \boldsymbol{\lambda}^\top \mathbf{A} \boldsymbol{\Omega} \mathbf{A}^\top \boldsymbol{\lambda} \right\} \right), \qquad (3.18)$$

*where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution, $\boldsymbol{\Delta}^* = \mathbf{A}\boldsymbol{\beta}^*$ and*

$$\boldsymbol{\zeta} \sim N_d(\mathbf{0}, \boldsymbol{\Omega}), \quad \boldsymbol{\xi} = \mathbf{A}\boldsymbol{\zeta} \sim N_q(\mathbf{0}, \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top), \quad \boldsymbol{\Omega} = \begin{bmatrix} \frac{N}{n}\mathbf{C}_0^{-1} & -\frac{N}{n}\mathbf{C}_0^{-1} \\ -\frac{N}{n}\mathbf{C}_0^{-1} & \frac{N}{n}\mathbf{C}_0^{-1} + \frac{N}{m} \cdot \frac{\tau^2}{\sigma^2}\mathbf{C}_1^{-1} \end{bmatrix}.$$

*According to (3.17), the matrices $\mathbf{C}_0$ and $\mathbf{C}_1$ are given by*

$$\mathbf{C}_0 = \frac{1}{n}[\mathbf{1}_n \ \mathbf{Z}_0]^\top [\mathbf{1}_n \ \mathbf{Z}_0], \quad \mathbf{C}_1 = \frac{1}{m}[\mathbf{1}_m \ \mathbf{Z}_1]^\top [\mathbf{1}_m \ \mathbf{Z}_1].$$

Proposition 3.1 asserts that the distribution of the order-constrained least squares estimator $\widehat{\boldsymbol{\beta}}$ differs from that of the unconstrained estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} = \widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ (cf. remark at the end of §3.3.1) by one extra term, whose distribution resembles that of truncated Gaussian random variables. To establish that connection, substitute $\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top$ inside the "argmin" in (3.18) by the identity matrix. With that substitution, it easy to see that the minimizer $\widehat{\boldsymbol{\lambda}}$ results as

$$\widehat{\lambda}_k = \max\{N^{-1/2}\xi_k - \Delta_k^*/\sigma, 0\}, \quad \xi_k \sim N(0,1), \quad k = 1, \ldots, q.$$

Assuming that $\boldsymbol{\Delta}^* \geq \mathbf{0}$ (i.e., the order constraint holds true), Proposition 3.1 thus also immediately implies that the bias of $\widehat{\boldsymbol{\beta}}$ is at most of the order $N^{-1/2}$. Depending on the magnitude of the entries of $\boldsymbol{\Delta}^*/\sigma$, the bias can be considerably smaller, which is an implication of the corollary below.

**Corollary 3.1.** *Under the conditions of Proposition 3.1, we have*

$$P(\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{LS}}) \geq 1 - \sum_{k=1}^{q} \exp\left(-N \cdot \frac{\Delta_k^{*2}}{2\sigma^2 \cdot \mathbf{a}_k^\top \boldsymbol{\Omega} \mathbf{a}_k}\right),$$

*where* $\mathbf{a}_k = [\mathbf{1} \ \mathbf{v}_k^\top]^\top$ *with* $\mathbf{v}_k$ *as in (3.7),* $k = 1, \ldots, q$, *and* $\boldsymbol{\Omega}$ *as defined in Proposition 3.1.*

The corollary confirms the intuition that in certain regimes the order-constrained estimator $\widehat{\boldsymbol{\beta}}$ coincides with the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ with overwhelming probability.

- *Large number of samples.* As the number of samples in both populations defined by the status variable $D$ increases, the least squares estimator the distribution of $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ concentrates more and more tightly around the true parameter $\boldsymbol{\beta}^*$. Since the latter satisfies the order constraint by assumption, it is intuitive that this carries over to $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ to an increasing extent as the sample size grows.

- *Large signal.* The ratio $\min_k \Delta_k^{*2}/\sigma^2$ can be interpreted as strength of evidence for the order constraint. As this ratio increases, the probability that the sampling variation of $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ leads to a violation of the order constraint becomes negligible.

Corollary 3.1 indicates that the order constraint only has an impact in case of small to moderate sample sizes and small to moderate differences in the conditional means $\mathbf{E}[T|D = 1, X = \mathbf{x}]$ and $\mathbf{E}[T|D = 0, X = \mathbf{x}]$. Otherwise, the order constraint is satisfied with overwhelming probability even if it is not imposed in estimation.

The second corollary of Proposition 3.1 yields a more explicit expression for the order-constrained estimator in the case of a single covariate ($p = 1$) with range $\mathcal{X} = [0, 1]$.

**Corollary 3.2.** *Under the conditions of Proposition 3.1, suppose additionally that $p = 1$ and $\mathcal{X} = [0, 1]$, i.e., $X$ is a single continuous covariate with range $[0, 1]$. We then have*

$$\widehat{\boldsymbol{\beta}} \overset{\mathcal{D}}{=} \boldsymbol{\beta}^* + \sigma\boldsymbol{\zeta} + \sigma\boldsymbol{\Omega}\mathbf{A}^\top \Bigg\{ \mathbf{G}^{-1}\boldsymbol{\nu} \cdot I\left(\mathbf{G}^{-1}\boldsymbol{\nu} \geq \mathbf{0}\right)$$

$$+ \begin{pmatrix} \frac{\nu_1}{G_{11}} \\ 0 \end{pmatrix} I\left(\nu_1 > 0, \frac{G_{12}\nu_1}{G_{11}} - \nu_2 > 0\right)$$

$$+ \begin{pmatrix} 0 \\ \frac{\nu_2}{G_{22}} \end{pmatrix} I\left(\nu_2 > 0, \frac{G_{12}\nu_2}{G_{22}} - \nu_1 > 0\right) \Bigg\},$$

*with $\boldsymbol{\zeta}, \boldsymbol{\Omega}$ as in Proposition 1, $\boldsymbol{\nu} = -\left(\frac{\boldsymbol{\Delta}^*}{\sigma} + \frac{1}{\sqrt{N}}\boldsymbol{\xi}\right)$, and $\boldsymbol{\xi} = (\zeta_3, \zeta_3 + \zeta_4)^\top \sim N_2(\mathbf{0}, \mathbf{G})$, where the entries of the 2-by-2 matrix $\mathbf{G}$ are given by*

$$G_{11} = \frac{N}{n}\left(\frac{\bar{x}^2}{s_x^2} + 1\right) + \frac{N}{m} \cdot \frac{\tau^2}{\sigma^2}\left(\frac{\bar{z}^2}{s_z^2} + 1\right), \quad G_{12} = G_{11} - \frac{N}{n}\frac{\bar{x}}{s_x^2} - \frac{N}{m} \cdot \frac{\tau^2}{\sigma^2}\frac{\bar{z}}{s_z^2},$$

$$G_{22} = 2G_{12} - G_{11} + \frac{N}{n}\frac{1}{s_x^2} + \frac{N}{m} \cdot \frac{\tau^2}{\sigma^2}\frac{1}{s_z^2}.$$

*The quantities $\bar{x}, s_x^2$ and $\bar{z}, s_z^2$ are in turn given by*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \bar{z} = \frac{1}{m}\sum_{i=n+1}^{N} x_i, \qquad s_z^2 = \frac{1}{m}\sum_{i=n+1}^{N}(x_i - \bar{z})^2.$$

This expression enables extremely rapid sampling from the distribution of $\widehat{\boldsymbol{\beta}}$ ($10^6$ samples in less than 0.1 seconds on a 2016 MacBook Pro) and in turn highly accurate numerical evaluation of important statistical properties including subsequent use in parametric ROC curve estimation (cf. §3.3.2) over grids of quantities of interest. The plots in Figure 3.1 were generated in this fashion. Specifically, we set $n = m = N/2$ for varying $N$ and $\sigma^2 = \tau^2 = 1$,

where the choice 1 can be made without loss of generality if the interest concerns the ROC curve since the latter only depends on the ratio

$$\frac{\boldsymbol{\Delta}^*}{\sigma} = \frac{1}{\sigma}\begin{pmatrix} \mu(x=0, D=1; \boldsymbol{\beta}^*) - \mu(x=0, D=0; \boldsymbol{\beta}^*) \\ \mu(x=1, D=1; \boldsymbol{\beta}^*) - \mu(x=1, D=0; \boldsymbol{\beta}^*) \end{pmatrix}.$$

The matrix $\boldsymbol{\Omega}$ is chosen as the population version corresponding to a uniform distribution of $X$ on $[0,1]$ for both groups $D=0$ and $D=1$, in which case it is easy to compute that

$$\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top = \begin{pmatrix} 16 & -8 \\ -8 & 16 \end{pmatrix}.$$

With the above choices, the distribution of $\widehat{\boldsymbol{\Delta}} = \mathbf{A}\widehat{\boldsymbol{\beta}}$ only varies with $\boldsymbol{\Delta}^* \in \mathbb{R}_+^2$ and the total sample size $N$. For simplicity, we let $\boldsymbol{\Delta}^* = \delta \cdot (1\ 1)^\top$, for $\delta > 0$.

Figure 3.1 confirms that order-constrained estimation differs visibly from unconstrained estimation if $N$ or the "signal" (i.e., the magnitude of the entries of $\boldsymbol{\Delta}^*/\sigma$, here quantified by $\delta$) is small. In that case, order-constrained estimation yields noticeable benefits by trading a slight bias for significant reductions in variance – by a factor up to 5. The order constraint can thus be a valuable addition in the small sample size/small signal regimes.

It is natural to expect that the order-constrained estimator is guaranteed to yield lower mean squared estimation error than the unconstrained estimator whenever the constraint holds, i.e., $\mathbf{A}\boldsymbol{\beta}^* \geq \mathbf{0}$. A theoretical result of this flavor is presented in the Appendix.

## 3.5 Simulation Studies

In this section, we outline the design of our simulation studies and their evaluation in terms of bias and mean square error. We discuss how the results vary with the chosen error distributions, sample sizes, location and scale differences across the two populations, and
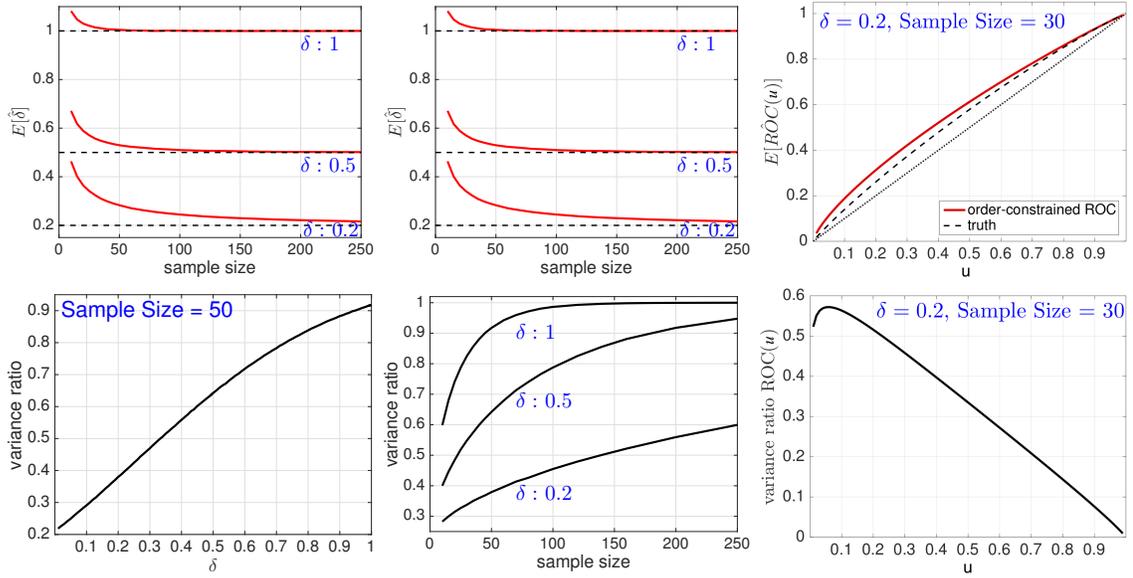
Figure 3.1: (T,L): Expectation of $\widehat{\delta} = \widehat{\Delta}_2$ as a function of $\delta$ for fixed sample size $N = 50$; straight line corresponds to zero bias. (T,M): Expectation of $\widehat{\delta}$ as a function of $N$ for fixed $\delta \in \{0.2, 0.5, 1\}$; horizontal line corresponds to zero bias. (T,R): Expectation of $\widehat{\mathrm{ROC}}(u)$ based on (3.15), $u \in [0,1]$ for $\delta = 0.2, N = 30$. (B,L): Variance ratio $\mathrm{var}(\widehat{\delta})/\mathrm{var}(\widehat{\delta}^{\mathrm{LS}})$ as a function of $\delta$ for fixed $N$, where $\widehat{\delta}^{\mathrm{LS}}$ denotes the estimator based on (weighted) least squares without constraint. (B,M): Variance ratio as a function of $N$ for fixed $\delta \in \{0.2, 0.5, 1\}$. (B,R): Pointwise variance ratios for the ROC curve for $\delta = 0.2, N = 30$.

assess the performance of the method proposed in this .

Let $X_i$, $D_i$, and $T_i$ be the covariate, binary status, and score of the $i$-th subject, respectively, where $i = 1, 2, \ldots, N$. Let further $e_{ij}$ be zero mean random error, $1 \leq i \leq N$, $j \in \{0, 1\}$. Data are generated according to

$$T_i = 1 + X_i + D_i \times \psi + X_i \cdot D_i + e_{i1} D_i \times \sqrt{\phi} + e_{i0}(1 - D_i), \qquad (3.19)$$

where $X_i \sim U(0, 1)$, $D_i \sim \text{Bernoulli}(0.5)$. The parameter $\psi$ is the location parameter and $\phi$ is the scale parameter, where $\psi \in (0, \infty]$, $\phi \in (0, \infty]$.

The following four settings are considered for the error terms $\{e_{ij}\}$:

- Study 1: $\{e_{i0}\}$ and $\{e_{i1}\}$ are i.i.d. from the $N(0, 1)$-distribution.
- Study 2: $\{e_{i0}\}$ and $\{e_{i1}\}$ are i.i.d. from a contaminated normal distribution $0.95N(0, 1)$ $+ 0.05N(0, 100)$.
- Study 3: $\{e_{i0}\}$ and $\{e_{i1}\}$ are i.i.d. from the t-distribution with four degrees of freedom $t(4)$.
- Study 4: $\{e_{i0}\}$ and $\{e_{i1}\}$ are from centered log-normal distribution with variance $[\exp(0.5^2) - 1] \exp(0.5^2) \approx 0.365$.

Note that the underlying mean difference $\Delta^*(x)$ given $\{X = x\}$ is equal to

$$\Delta^*(x) = \mathbf{E}[T|D = 1, X = x] - \mathbf{E}[T|D = 0, X = x] = \psi + x.$$

The underlying covariate-specific ROC at $u \in (0, 1)$ and $X = x$ is given by

$$\text{ROC}_x^*(u) = 1 - G\left(\frac{1}{\sqrt{\phi}} \left(G^{-1}(1 - u) - \Delta^*(x)\right)\right),$$

where $G$ denotes the CDF of the $\{e_{i0}\}$ and the $\{e_{i1}\}$. Let $\text{MSE}_1$ and $\text{MSE}_2$ denote the mean squared estimation error without and with constraint, respectively, and let further RE represent the relative efficiency, which is given by $\text{RE} = \text{MSE}_1/\text{MSE}_2$; an RE exceeding

one indicates better performance with the order constraint. Our results are based on 1000 iterations. The results for the mean difference are for $x = 0.5$, and the results for the covariate-specific ROC are for $x = 0.5$ and FAR $u = 0.5$. A more comprehensive account including graphical representations, results on other combinations of $x$ and $u$, as well as the case of multiple $(p > 1)$ correlated covariates is contained the Appendix.

### 3.5.1 Bias and MSE of the Mean Difference and Covariate-Specific ROC Curves

In this subsection, we investigate the bias and the MSE in estimating $\Delta^*(x)$ and $\text{ROC}_x^*(u)$. Specifically, we fix $x = 0.5$ for the mean difference, and we fix $u = 0.5$ and $x = 0.5$ for the covariate-specific ROC curves. Figure 3.2 depicts the results in studies 1 and 3; the complete simulation results can be found in §A.5.2 of the Appendix. In Figure 3.2, the bias of the constrained method exceeds slightly that without constraint. On the other hand, the MSE of the constrained method is smaller, which offsets the slight increase in bias. Consequently, the relative efficiency (RE) is larger than one throughout. As the sample size increases, the RE get closer to one, which is anticipated in light of Corollary 3.1. Furthermore, it can be seen that gains in efficiency persist for a wide range of values for $\phi$.

## 3.6 Real Data Example

The proposed method is illustrated through its application to the Face Recognition Vendor Test (FRVT) 2006 by [76]. The data from FRVT have been used on the accuracy of existing facial recognition algorithms by integrating demographic factors [77,78], and on the relationship between image quality and classification accuracy [39]. Although those papers are among the very few in the literature discussing the impact of categorical covariates such as gender, race, and image quality on the accuracy of facial recognition tests, the potential influence of continuous covariates such as age could not be taken into account with the methods in those papers.
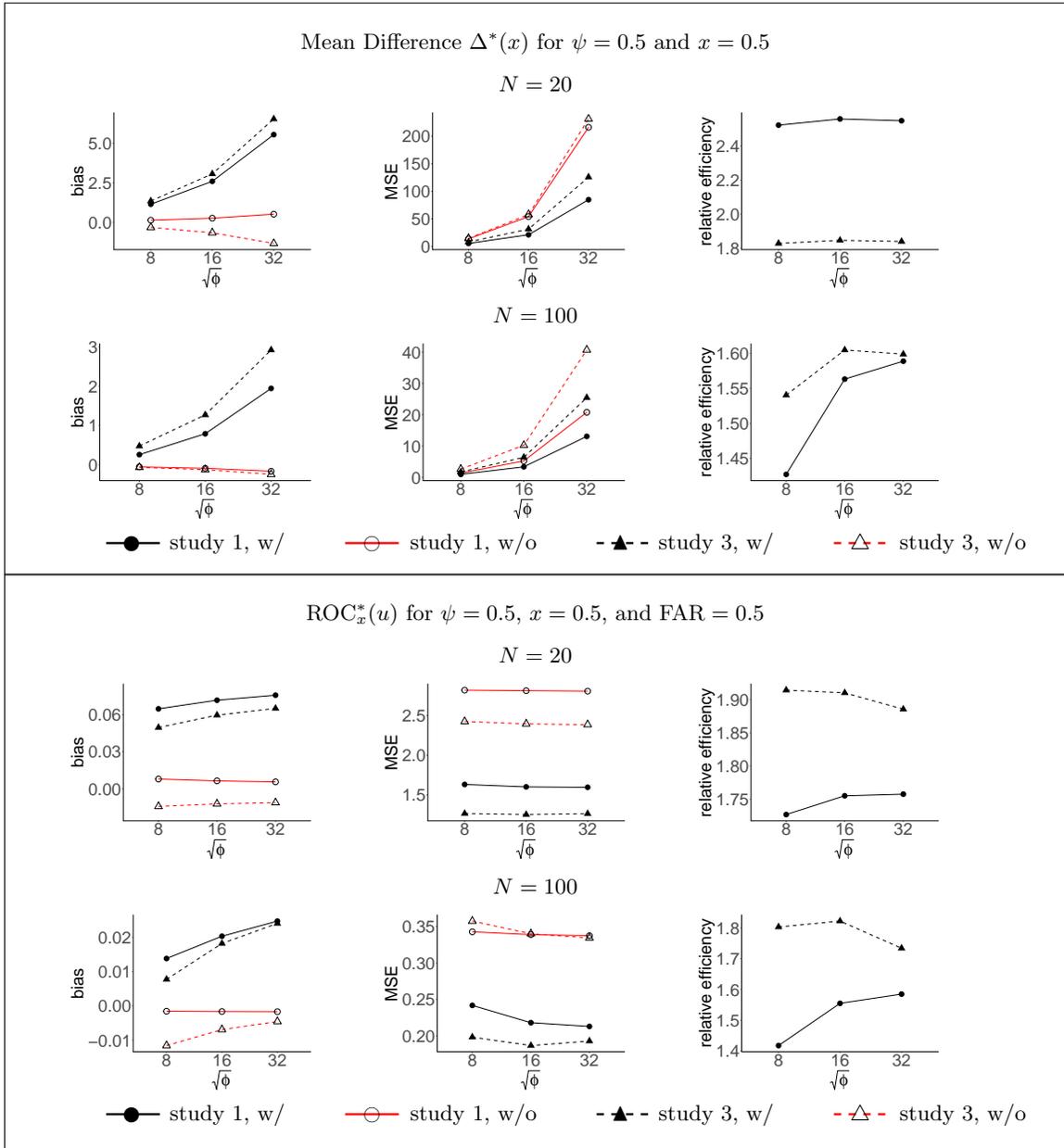
Figure 3.2: Selected results of the simulation study. Left column: bias, middle: mean squared error, right: relative efficiency (values larger than 1 correspond to better performance). w/o: linear regression without order constraint; w/: linear regression with order constraint.

We start by laying out the notations formally. We are given scores $S_{ij} = s(I_i, I_j)$ quantifying the similarity of a pair of facial images $(I_i, I_j)$, where $i$ and $j$ are the IDs of two specific subjects. A score is called genuine if two images are from the same source, i.e., $i = j$; otherwise, it is called imposter. To appropriately apply the proposed method, the given similarity scores were reduced to a set of independent scores for which $\{i, j\} \cap \{k, l\} = \emptyset$ as follows. The set of subject IDs were first randomly divided into two subsets $\mathcal{I}_1$ and $\mathcal{I}_2$ of equal size. Genuine scores $\{S_{ii} | D_{ii} = 1, i \in \mathcal{I}_1\}$ were selected from $\mathcal{I}_1$, and imposter scores $\{S_{ij} | D_{ij} = 0, i, j \in \mathcal{I}_2\}$ were selected from $\mathcal{I}_2$. The selected set of genuine scores does not contain more than one score from each subject. Subset $\mathcal{I}_2$ was further randomly split into halves, and imposter scores were obtained by pairing IDs in the resulting two subsets based on lexicographical ordering. In this way, independence among the selected imposter scores is ensured. We only consider scores whose image quality category equal "good" since the assumed order constraint turns out to be most adequate for this subset of images.
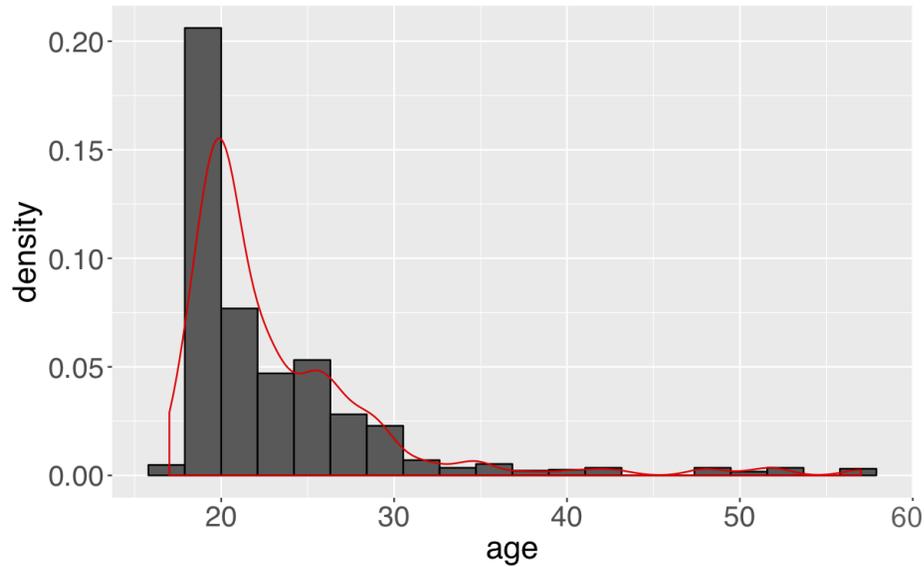


Figure 3.3: Histogram and kernel density estimate of the covariate "age".

The sample mean and sample standard deviation (SD) for the genuine group are -22.02

and 3.47, respectively, and -25.23 and 3.97 for the imposter group; the sample sizes in both groups are the same, i.e., $n = m = 228$. The scores $S_{ij}$ are transformed by the Box-Cox transformation. The parameter $\lambda = 0.5$ of the transformation is determined based on the scores in the entire data set; note that while the population ROC curve is invariant under monotone transformations, the use of the Box-Cox transformation can improve the fit of the regression model (cf. Eq. (3.20) below), in which the transformed scores $T_{ij} = S_{ij}^{0.5}$ are used as responses.

The age of each subject was considered as a covariate in our study. A histogram and a kernel density estimate are shown in Figure 3.3. It can be seen that the density is right-skewed. The range of ages in this study is between 17 to 57, with a mode close to 20 and relatively little mass in the range of 30 to 60. The variable age is considered as a continuous random variable throughout our study.

The transformed scores are modeled according to the linear regression model

$$T_{ij} = \beta_0 + \beta_1 D_{ij} + \beta_2 \cdot \mathtt{A}_i + \beta_3 \cdot \mathtt{A}_j + \beta_4(\mathtt{A}_i \cdot D_{ij}) + D_{ij}\epsilon_{ij1} + (1 - D_{ij})\epsilon_{ij0}, \qquad (3.20)$$

where the binary status variable $D_{ij}$ equals one if the score $T_{ij}$ is from the genuine group and zero else, and $\mathtt{A}_i$ and $\mathtt{A}_j$ represent the ages of the subjects in images $I_i$ and $I_j$, respectively. In the regression model, we only consider the interaction term $\mathtt{A}_i \cdot D_{ij}$ because if $D_{ij} = 1$, then the underlying score $T_{ij}$ is from the same person, and therefore $\mathtt{A}_i = \mathtt{A}_j$; when $D_{ij} = 0$, then the interaction term is also equal to 0. It is hence not meaningful to include both interaction terms in the model. In the same vein, we only consider the condition $\{\mathtt{A}_i = \mathtt{A}_j = x\}$ for the covariate-specific ROC curves.

After fitting the regression model and obtaining the age-specific ROC curve as in (3.15), we compare the variabilities of the traditional unconstrained and proposed constrained regression methods. We are particularly interested in the estimated ROC curve at three specific ages (17, 23, and 30) as suggested by [79] since these three age groups represent younger through older subjects, respectively.

The unconstrained and constrained regression methods are run for 1000 bootstrap samples drawn from the original data set obtained after pre-processing. The means and SDs are calculated over all bootstrap replications. The summary measure for comparing efficiency of two methods is calculated using variance ratios (VR) over all bootstrap replications as VR $= var_1/var_2$, where $var_1$ is the variance of traditional linear regression method and $var_2$ is the variance of the linear regression with ordering constraint. In each bootstrap replication, we calculate the estimated mean difference given by

$$\widehat{\Delta}(x) = \mu(\mathtt{A}_i = x, \mathtt{A}_j = x, D_{ij} = 1; \widehat{\boldsymbol{\beta}}) - \mu(\mathtt{A}_i = x, \mathtt{A}_j = x, D_{ij} = 0; \widehat{\boldsymbol{\beta}}) = \widehat{\beta}_1 + \widehat{\beta}_4 x \quad (3.21)$$

Note that $\widehat{\beta}$ here denotes the estimates of the regression coefficients obtained from either the unconstrained or constrained approach. Accordingly, we estimate $\mathrm{ROC}_x(u)$ based on (3.15) for selected values of FAR, using the parametric approach in §3.3.2.

Figure 3.5 (a) shows the averages and the SDs of the mean difference over the 1000 bootstrap replications when using the traditional regression ROC method and the proposed constrained method. Our method produces smaller variances than the traditional method for all ages. For the age of 23, the mean and SD of our method are similar to the results produced when using linear regression without ordering constraint, but the VR is still slightly larger than 1. Based on the histogram of our data in Figure 3.3, we note that the majority of the population is aged between 18 and 26. This means that the sample size is relatively large when $x = 23$, and the large sample size leads to similar variability for our method and the method without ordering constraint.

Figure 3.4 compares pointwise averages for the covariate-specific ROC curves of both approaches for different ages. The pointwise confidence intervals of the ROC curves from the proposed method mostly overlap with those from the traditional method without constraints. The confidence intervals show the variabilities at every point on the estimated ROC curves. For ages of 17 and 30, the proposed method has narrower confidence intervals than the traditional method, while for age 23, the difference in the confidence intervals is

not noticeable.
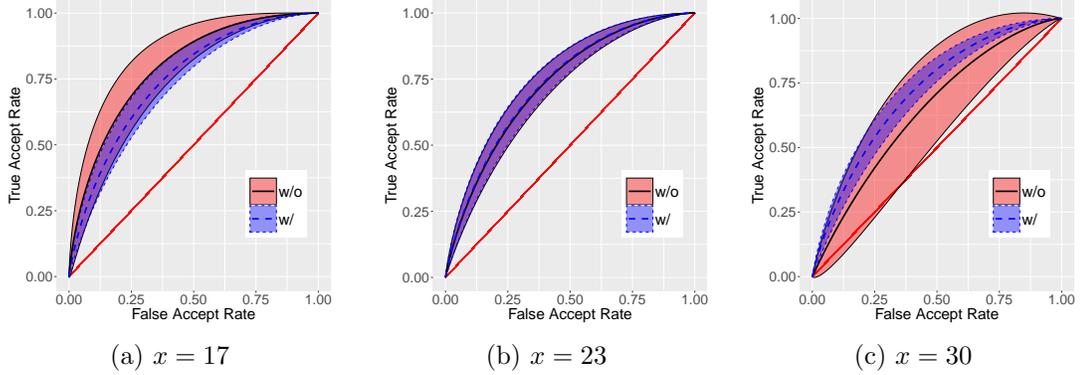


Figure 3.4: Covariate-specific ROC curves for different ages. (w/o: linear regression without order constraint; w/: linear regression with order constraint. The shaded area represents pointwise 95% confidence intervals; best seen in color)

Figure 3.5 (b,c) displays the bootstrap variance of the ROC curves in dependence of different values of FAR when $x = 30$ and the covariate age when FAR $u = 0.5$, respectively. We can see that the order-constrained method consistently reduces the variance. For Figure 3.5b, the largest difference occurs for FAR near 0.5. As FAR tends to 0 or 1, the ROC curve will tend to 0 or 1 by definition. Figure 3.5c shows that the difference in the variances across the two methods increases with the value for the covariate age. Note that when the value of the age is about 20, the variances of the covariate-specific ROC for two methods are roughly the same.

Mean and SD of $\widehat{\Delta}(x)$

| $x$ | method | mean | SD | VR |
|-----|--------|------|------|------|
| 30 | w/o | 0.28 | 0.07 | 3.28 |
|    | w/ | 0.32 | 0.04 | |
| 23 | w/o | 0.34 | 0.04 | 1.01 |
|    | w/ | 0.34 | 0.04 | |
| 17 | w/o | 0.40 | 0.06 | 2.08 |
|    | w/ | 0.36 | 0.04 | |

(a)
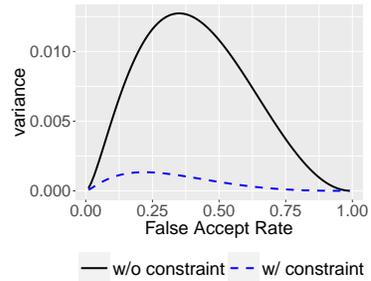


(b)
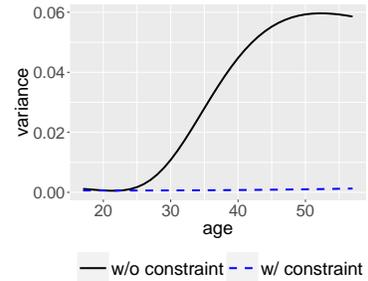


(c)

Figure 3.5: Bootstrap mean and SD of $\widehat{\Delta}(x)$ (a), bootstrap variance of the estimated covariate-specific ROC in dependence of FAR when the covariate age is 30 (b), and different values of the covariate age when FAR is 0.5 (c). (w/o: linear regression without order constraint; w/: linear regression with order constraint.)

# Chapter 4: Order-Restricted ROC Curve and Heteroscedastic Modeling Using Quantile Regressions

## 4.1 Introduction

The receiver operating characteristic (ROC) curve is an important tool for assessing how well two populations can be discriminated based on a continuous score like the measured level of a specific biomarker. The ROC curve graphs the trade-off between sensitivity and specificity in dependence of the decision threshold associated with the score. In practice, it is common that the accuracy of a score in discriminating between two populations depends noticeably on additional covariates. For example, the discriminative accuracy of a medical imaging diagnostic devices is potentially influenced by the evaluators's performance; the performance of a biomarker measurement is frequently impacted by the severity of a patient's disease. An approach to account for covariates is the so-called *covariate-specific* ROC curve introduced by Pepe [29].

There are various strategies for modeling covariate-specific ROC curves such as non-parametric or semiparametric Bayesian methods and induced or direct-regression methods [30–33]. One particularly convenient as well as popular approach is the location-scale model [34, 35]. In a nutshell, the distribution of the score is modeled via location-scale transformation of base distributions associated with the two populations, where potentially both the location and scale transformation depend on covariates. A common way to estimate the latter is via regression techniques with least squares (LS) regression as the simplest approach. However, a well-known shortcoming of LS is sensitivity to heavy-tailed error distributions, and composite quantile regression (CQR) [37] is proposed as an alternative in the paper by Duan & Zhou [35].

In the present chapter, we discuss two innovations of the method by Duan & Zhou (henceforth DZ) [35], motivated by an application from the field of biometric recognition as illustrated in Figure 4.1a. First, we propose an extension of DZ to incorporate order constraints for the (covariate-specific) location parameters associated with each population. If the scale parameters and base distributions are the same across the two populations, such constraints are equivalent to a stochastic ordering constraint. Their use is often appropriate for biometric recognition systems since those are typically calibrated to deliver larger scores for matching (so-called *genuine*) scores than for non-matching (i.e., *imposter*) scores [3, 4]. In addition to that specific application domain that is focused on herein, the proposed methodology is likely to be useful in other fields such as medical diagnostics or biomarker studies.

Second, the use of the CQR method for covariate-specific ROC curve modeling adopted in DZ assumes a regression model with heteroscedastic errors, i.e., that the errors do not depend on covariates. This assumption can be limiting in applications. As a remedy, we suggest the use of He's method [80] that consists of a two-fold application of median regression to estimate both location and scale transformation. The result enables ROC curve estimation in a robust (i.e., insensitive to heavy-tailed distributions) fashion in the situation that *both* location and scale function are dependent on covariates. In fact, this scenario is not uncommon in biometrics: in the facial recognition study displayed in Figure 1, the variance of the scores varies considerably with the variable "image quality" (see Figure 4.1b and Figure 4.1c).

The rest of this chapter is organized as follows. Section 4.2 introduces key concepts such as the covariate-specific ROC curve and the location-scale model. In Section 4.3, we lay out the details of our technical contributions with regard to order-constrained and heteroscedastic modeling. We present the results of simulation studies including a detailed comparison to existing methods in Section 4.4. In Section 4.5, we discuss the application of our method to biometric data. A conclusion and discussion are provided in Chapter 5.

| (a) density | (b) non-matched | (c) matched |
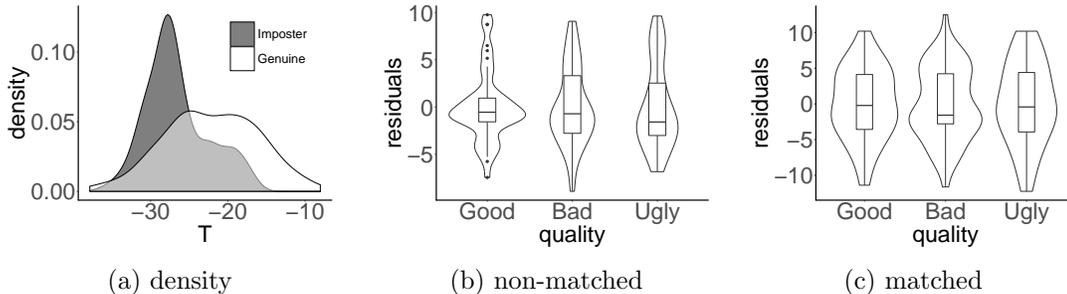
Figure 4.1: Density of the facial recognition data and box-plots of the residuals in the location-scale model conditional on status and quality

## 4.2 Modeling Covariate-Specific ROC Curves via the Location-Scale Model

In this section, we introduce the notion of covariate-specific ROC curves and their modeling based on the location-scale model [34] and regression techniques, in particular composite quantile regression (CQR) [35].

### 4.2.1 Covariate-Specific ROC Curve

The receiver operator characteristic (ROC) curve is a tool to evaluate the accuracy of a continuous score $T$ in discriminating between two different populations. The score is dichotomized depending on a threshold $t_0$, i.e., an assignment to one of the two populations is made depending on whether $T \geq t_0$, and the ROC curve graphs the trade-off between sensitivity (true positive rate, TPR) and $1-$specificity (false positive rate, FPR) as $t_0$ is varied. A common example is medical diagnosis based on the level of a biomarker known to be indicative of a certain disease; here, the two populations refer to diseased and healthy patients, respectively. Many techniques in forensic science are based on the biometric scores. For example, the fingerprint identification is developed according to the comparison scores [81] and the gait analysis is constructed associate with the likelihood scores [82].

Let $D$ denote the $\{0, 1\}$-valued variable that indicates the population a corresponding

score $T$ stems from, for example $D = 1$ may refer to a diseased subject and accordingly $D = 0$ to a healthy subject. The TPR and FPR for threshold $t_0$ are hence given by $P(T \geq t_0 | D = 1)$ and $P(T \geq t_0 | D = 0)$. Denote the conditional cumulative distribution functions of $T$ given $D = 0$ and $D = 1$ by $F_0$ and $F_1$, respectively, i.e, $F_0(t) = P(T \leq t | D = 0)$ and $F_1(t) = P(T \leq t | D = 1)$, $t \in \mathbb{R}$. The ROC curve then results as

$$\text{ROC}(u) = 1 - F_1(F_0^{-1}(1 - u)), \quad u \in (0, 1),$$

where $F_0^{-1} := \inf\{t \in \mathbb{R} : F_0(t) \geq u\}$. The argument $u$ corresponds to a given value of the FPR, and $\text{ROC}(u)$ returns the value of the TPR at $u$.

As pointed out in the introduction, the discriminate accuracy of $T$ might strongly depend on covariates $X = (X_1, \ldots, X_p)^\top$ observed with $T$ and $D$. Let $\mathcal{X} = \text{range}(X)$. Loosely speaking, the *covariate-specific* ROC curve [29] is obtained from the previous definition of the ROC by working conditional on the event $\{X = \mathbf{x}\}$ for $\mathbf{x} \in \mathcal{X}$:

$$\text{ROC}_{\mathbf{x}}(u) = 1 - F_{1,\mathbf{x}}(F_{0,\mathbf{x}}^{-1}(1 - u)), \quad u \in (0, 1), \tag{4.1}$$

where $F_{j,\mathbf{x}}(t) := P(T \leq t | D = j, X = \mathbf{x})$ and $F_{j,\mathbf{x}}^{-1}(u) := \inf\{t \in \mathbb{R} : F_{j,\mathbf{x}}(t) \geq u\}$ are the conditional CDFs of $F_j$ and its generalized inverse, respectively, $j \in \{0, 1\}$.

## 4.2.2 Location-Scale Model

A basic modeling strategy for the covariate-specific ROC curve inducing additional structure that is amenable to estimation in the presence of multiple covariates based on regression techniques is the location-scale model [34]. In essence, the location-scale model assumes that in each of the two populations indicated by the binary status variable $D$, the score $T$ can be expressed as a location-scale transformation that depends on the covariates. Specifically,

the location-scale model postulates that

$$T = D\big(\mu_1(X; \boldsymbol{b}_1^*) + \sigma_1(X; \boldsymbol{a}_1^*)e_1\big) + (1 - D)\big(\mu_0(X; \boldsymbol{b}_0^*) + \sigma_0(X; \boldsymbol{a}_0^*)e_0\big). \qquad (4.2)$$

In the following, let us comment on the quantities appearing in Eq. (4.2).

- $e_0$ and $e_1$ are random variables with "location" zero and unit "scale", where as measure of location one may take the expectation or the median (which coincide if $e_0$ and $e_1$ are symmetric), and as measure of scale one may take the variance or the median of the absolute values. Note that the second alternatives for location and scale extend (4.2) to a broader class of distributions.

- $\mu_j$ and $\sigma_j$, $j \in \{0, 1\}$, are referred to as location and scale functions, respectively. The latter are assumed to be known functions of the covariates $X$ and unknown parameters $\boldsymbol{b}_j^*$ and $\boldsymbol{a}_j^*$, $j \in \{0, 1\}$. In this chapter, we confine ourselves to functions having identical form for both values of $D$ and that are affine in the unknown parameters, i.e.,

$$\mu_j(\mathbf{x}; \boldsymbol{b}_j^*) = \phi(\mathbf{x})^\top \widetilde{\boldsymbol{b}}_j^* + b_{0j}^*, \qquad \sigma_j(\mathbf{x}; \boldsymbol{a}_j^*) = \psi(\mathbf{x})^\top \widetilde{\boldsymbol{a}}_j^* + a_{0j}^*, \qquad j \in \{0, 1\}, \quad (4.3)$$

where $\boldsymbol{b}_j^* = (b_{0j}^*, [\widetilde{\boldsymbol{b}}_j^*]^\top)^\top$ and $\boldsymbol{a}_j^* = (a_{0j}^*, [\widetilde{\boldsymbol{a}}_j^*]^\top)^\top$. To avoid notational clutter, we may assume without loss of generality that $\phi(\mathbf{x}) = \psi(\mathbf{x}) = \mathbf{x}$ since this can simply be accomplished by augmenting $\mathbf{x}$ as needed to include additional transformations of the original set of covariates.

In the sequel, *homoscedasticity* will refer to the case in which $\sigma_1$ and $\sigma_0$ are both independent of $X$; otherwise, we shall speak of *heteroscedasticity*.

Direct calculations show that under the location-scale model (4.2), the covariate-specific ROC curve defined in (4.1) at any $\mathbf{x} \in \text{range}(X)$ results as

$$\text{ROC}_{\mathbf{x}}(u) = 1 - G_1\left(\frac{\sigma_0(\mathbf{x}; \boldsymbol{a}_0^*)}{\sigma_1(\mathbf{x}; \boldsymbol{a}_1^*)}\left(G_0^{-1}(1-u) - \frac{\mu_1(\mathbf{x}; \boldsymbol{b}_1^*) - \mu_0(\mathbf{x}; \boldsymbol{b}_0^*)}{\sigma_0(\mathbf{x}; \boldsymbol{a}_0^*)}\right)\right), \quad u \in (0,1), \quad (4.4)$$

where $G_0$ and $G_1$ denote the CDFs of $e_0$ and $e_1$, respectively.

### 4.2.3   Order Constraints

As mentioned in the introduction, in some applications it is common that the distribution of the score $T$ in the population $D = 1$ is associated with a larger location than the distribution of $T$ in the population $D = 0$ regardless of the specific values $\mathbf{x}$ observed for the covariates. For example, the level of a biomarker indicating the presence of a disease is supposed to be larger among diseased than healthy patients. In the analysis of biometric traits, computer algorithms providing scores that assess agreement between pairs of measurements (e.g., fingerprints or facial images) are typically calibrated to deliver large scores for matching (*genuine*) pairs than for non-matching (*imposter*) pairs. This yields the constraint

$$\mu_1(\mathbf{x}; \boldsymbol{b}_1^*) \geq \mu_0(\mathbf{x}; \boldsymbol{b}_0^*) \iff \mathbf{x}^\top(\widetilde{\boldsymbol{b}}_1^* - \widetilde{\boldsymbol{b}}_0^*) + (b_{01}^* - b_{00}^*) \geq 0 \quad \text{for all } \mathbf{x} \in \mathcal{X}, \qquad (4.5)$$

where the equivalence is according to (4.3) and the subsequent comment. In the context of ROC estimation, this constraint is discussed in Chapter 3. It amounts to an ordering of $T|D = 1$ and $T|D = 0$ *in mean* if the latter exists and if $e_0$ and $e_1$ in (4.2) are symmetric. Moreover, in the case of identical scale functions, i.e., $\boldsymbol{a}_0^* = \boldsymbol{a}_1^*$, the constraint (4.5) is equivalent of a stochastic ordering of $T|D = 1$ and $T|D = 0$.

Order constraints have received considerable attention in recent literature including papers discussing such constraints in the context of ROC curve modeling [83, 84]. Their incorporation can be beneficial for at least two reasons: first, as explained above, they yield more interpretable results in applications in which those constraints are known to be

satisfied; second, they can yield improve statistical efficiency in low sample size or weak signal (in the sense of weak separation of the score distribution in the two poplations) contexts (see Chapter 3).

## 4.3 Inference

Building on methodology developed in DZ's paper [35], Chapter 3 discussed least squares-based estimation of the ROC based on the location-scale model and subject to the order constraint (4.5) in the homoscedastic setting, i.e., $\sigma_0(\mathbf{x}; \boldsymbol{a}_0) \equiv \sigma$, $\sigma_1(\mathbf{x}; \boldsymbol{a}_1) \equiv \tau$. In the sequel, we briefly review the approach taken in Chapter 3 and identify some of its limitations. We then present two extensions intended to address those limitations and to widen the scope of the method considerably.

### 4.3.1 Least Squares Regression

Let us review the method in Chapter 3. There is an obvious connection between the location-scale model (4.2) and a regression model that suggests estimation via a two-stage procedure [20,35].Given a sample of size $N$ of the form $\{(T_i, D_i, \mathbf{x}_i)\}_{i=1}^N$ in which the $T_i | D_i, \mathbf{x}_i$ are independent and distributed according to (4.2), the first stage amounts to (weighted) least squares regression of the $\mathbf{y} = \{T_i\}_{i=1}^N$ on $\{D_i, \mathbf{x}_i, \mathbf{x}_i * D_i\}_{i=1}^N$, where $\mathbf{x}_i * D_i = \mathbf{x}_i \cdot D_i$, $1 \le i \le N$, refers to the interaction terms between covariates and the status variable $D$. The second stage amounts to estimation of the CDFs $G_0$ and $G_1$ from the residuals of the preceding regression. As shown in Chapter 3, a close proxy of the constraint (4.5) can be incorporated in the first stage by means of linear inequality constraints on the regression parameter. Details are given below; without loss of generality, it is assumed that $D_i = 0$, $1 \le i \le n$, and $D_i = 1$, $i = n+1, \ldots, n+m = N$.

*Stage I*: Solve the optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_2^2 \quad \text{subject to } \mathbf{A}\boldsymbol{\beta} \ge \mathbf{0}, \tag{4.6}$$

66

where

- $\mathbf{W} = \mathrm{diag}\Big( \underbrace{1/\sigma^2, \ldots, 1/\sigma^2}_{n \text{ times}}, \underbrace{1/\tau^2, \ldots, 1/\tau^2}_{m \text{ times}} \Big)$ is a diagonal weight matrix, where the weights are inversely proportional to the status-specific variances $\sigma^2$ for $D = 0$ and $\tau^2$ for $D = 1$, respectively. While in (4.6) the latter are assumed to be known, they are straightforward to estimate based on an ordinary least squares of $\mathbf{y}$ on $\mathbf{X}$.

- $\mathbf{X}$ is an $N \times d$ design matrix, $d = 2(p + 1)$, whose $i$-th row is given by $\mathbf{X}_{i\bullet} = (1,\ \mathbf{x}_i^\top,\ D_i,\ (\mathbf{x}_i * D_i)^\top)$, $1 \le i \le N$. We assume that $\mathbf{X}$ is non-singular throughout this chapter.

- $\boldsymbol{\beta} = (\beta_0,\ \boldsymbol{\beta}_X^\top,\ \beta_D,\ \boldsymbol{\beta}_{XD}^\top)^\top$ represents a vector of regression coefficients. The estimator $\widehat{\boldsymbol{\beta}}$ is defined as the minimizer of (4.6), and is partitioned accordingly. Given $\widehat{\boldsymbol{\beta}}$, we use $\widehat{\boldsymbol{b}}_0 = (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}_X^\top)^\top$ and $\widehat{\boldsymbol{b}}_1 = (\widehat{\beta}_0 + \widehat{\beta}_D, (\widehat{\boldsymbol{\beta}}_X + \widehat{\boldsymbol{\beta}}_{XD})^\top)^\top$ as estimators of $\boldsymbol{b}_0^*$ and $\boldsymbol{b}_1^*$, respectively. Please refer to Figure 4.2 for a diagram visualizing the relationships between the different parameters.

- $\mathbf{A}$ is a $q \times d$ constraint matrix, where each row of $\mathbf{A}$ represents one linear constraint imposed on $\boldsymbol{\beta}$ in order to integrate (a proxy of) the order constraint (4.5) into estimation. While the constraint is linear in the parameters for each fixed $\mathbf{x}$, the set $\mathcal{X}$ could be complex and as a result would render the constraint difficult to implement. There are two possible proxies that yield reductions to linear inequality constraints: (1) *Outer approximation*: suppose that $[l_j, u_j]$ are known upper and lower bounds for the $j$-th covariate, $1 \le j \le p$. Then, the hyperrectangle $\overline{\mathcal{X}} := [l_1, u_1] \times \ldots \times [l_p, u_p]$ includes $\mathcal{X}$, and by convexity the constraint (4.5) holds if it holds for the vertices $\{v_\ell\}_{\ell=1}^q = \{l_1, u_1\} \times \ldots \times \{l_p, u_p\}$ of $\overline{\mathcal{X}}$, where $q = 2^p$. Accordingly, the $\ell$-th row of $\mathbf{A}$ is given by $[\mathbf{0}_{p+1}\ 1\ v_\ell^\top]$, $\ell = 1, \ldots, q$.

  (2) *Inner approximation*: the constraint is imposed with respect to a given sample $\{(T_i, D_i, \mathbf{x}_i)\}_{i=1}^N$ or a suitable subset thereof. Accordingly, each row of $\mathbf{A}$ is of the form $[\mathbf{0}_{p+1}\ 1\ \mathbf{x}_i^\top]$. For a more detailed explanation, specific examples, and a discussion of

the merits of the two approximation schemes, we refer to Chapter 3.

Optimization problem (4.6) is a quadratic program, and can be handled by a host of algorithms available for problems of this form.

*Stage II*: Estimation of the ROC curve

Let $\widehat{e}_{i0} = (T_i - \widehat{\beta}_0 - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_X)/\sigma$, $i = 1, \ldots, n$, and $\widehat{e}_{i1} = (T_i - \widehat{\beta}_0 - \widehat{\beta}_D - \mathbf{x}_i^\top (\widehat{\boldsymbol{\beta}}_X + \widehat{\boldsymbol{\beta}}_{XD}))/\tau$, $i = n+1, \ldots, N$, denote the (standardized) residuals associated with the solution of (4.6). Their empirical CDFs $\widehat{G}_0$ and $\widehat{G}_1$ and their generalized inverses naturally serve as estimators of $G_0$ and $G_1$ and their inverses, respectively, and can be plugged into the expression (4.4) for the covariate-specific ROC curve. In summary, this yields the following estimator:

$$\widehat{\mathrm{ROC}}_{\mathbf{x}}(u) = 1 - \widehat{G}_1 \left( \frac{\sigma}{\tau} \left( \widehat{G}_0^{-1}(1-u) - \frac{\widehat{\beta}_D + \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{XD}}{\sigma} \right) \right), \quad u \in (0,1). \qquad (4.7)$$

**Remark**. As pointed out in DZ, estimation of the ROC curve in the flavor of (4.7) can be done without knowledge of $\sigma$ and/or $\tau$ or separate estimation of these quantities. Specifically, observe that the re-scaled error $\sigma e_0$ and $\tau e_1$ have CDFs $H_0(\cdot) = \frac{1}{\sigma} G_0(\cdot/\sigma)$ and $H_1(\cdot) = \frac{1}{\tau} G_1(\cdot/\tau)$ and accordingly $\mathrm{ROC}_{\mathbf{x}}^* = 1 - H_1(H_0^{-1}(1-u) - \mathbf{x}^\top (\boldsymbol{b}_1^* - \boldsymbol{b}_0^*))$. The CDFs $H_0$ and $H_1$ can be estimated as the empirical CDFs $\widehat{H}_0$ and $\widehat{H}_1$ of the "raw" residuals (i.e., without scaling by the inverse standard deviations as above). In summary, this yields

$$\widehat{\mathrm{ROC}}_{\mathbf{x}}(u) = 1 - \widehat{H}_1 \left( \left( \widehat{H}_0^{-1}(1-u) - \widehat{\beta}_D - \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{XD} \right) \right), \quad u \in (0,1). \qquad (4.8)$$

### 4.3.2 Composite Quantile Regression

An apparent shortcoming of the previous approach is its sensitivity to heavy tails in the distribution of the error terms $e_0$ and/or $e_1$. Composite quantile regression (CQR) [37] was shown to outperform least squares regression for a variety of error distributions including
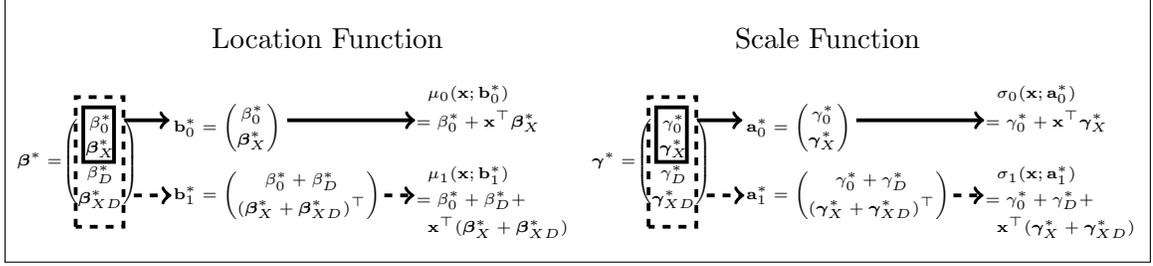
Figure 4.2: Diagrams summarizing the relationship between regression parameters and the corresponding parameters of the location-scale model.

contaminated normal distribution, Cauchy distribution, log-normal distribution, etc., while significantly boosting statistical efficiency in the case of normal errors in comparison to median regression. Given these appealing properties, CQR has been adopted as regression method to be applied in the first stage of estimation in the location-scale model [35]. The results therein confirm the superiority of CQR over least squares regression for a wide variety of distributions for the error terms. The use of the constraint (4.5) in conjunction with the approach in DZ has not been discussed in the previous chapter, and we consider this extension in the following. Building on Chapter 3 as outlined in the previous subsection, the least squares objective is replaced by the loss function employed in CQR while the linear constraints on the coefficients are modified appropriately. The use of CQR enables a more faithful representation of stochastic ordering constraints as opposed to a mere ordering of the location parameters as in (4.5) since two distributions are stochastically ordered if and only their corresponding quantiles are. Details are presented in the sequel. Let $\{\tau_k\}_{k=1}^K \subset (0,1)$ be a grid of quantiles; a reasonable default is $\tau_k = \frac{k}{K+1}$, $1 \leq k \leq K$, and odd $K \geq 3$, with larger values of $K$ yielding a more fine-grained modeling at the expense of additional computational effort required. Accordingly, we define the so-called check loss functions $\rho_{\tau_k} : \mathbb{R} \to \mathbb{R}_+$ by $u \mapsto \rho_{\tau_k}(u) = u\{\tau_k - I(u \leq 0)\}$, $1 \leq k \leq K$ [85]. We are thus in position

to present a counterpart of the least squares-based formulation (4.6).

$$\min_{\substack{\{\beta_{0k}\}_{k=1}^K, \boldsymbol{\beta}_X \in \mathbb{R}^p \\ \{\beta_{Dk}\}_{k=1}^K, \boldsymbol{\beta}_{XD} \in \mathbb{R}^p}} \sum_{k=1}^K \sum_{i=1}^N \rho_{\tau_k}(T_i - \beta_{0k} - \mathbf{x}_i^\top \boldsymbol{\beta}_X - D_i \cdot \beta_{Dk} - D_i \cdot \mathbf{x}_i^\top \boldsymbol{\beta}_{XD}) \qquad (4.9)$$

$$\text{subject to } \mathbf{A}\boldsymbol{\beta}_k \geq \mathbf{0}, \quad \boldsymbol{\beta}_k := (\beta_{0k} \ \boldsymbol{\beta}_X^\top \ \beta_{Dk} \ \boldsymbol{\beta}_{XD}^\top)^\top, \quad 1 \leq k \leq K, \qquad (4.10)$$

where the matrix $\mathbf{A}$ takes the same form as for the least squares-based approach (4.6). Note, however, that the total number of constraints multiplies by a factor of $K$ since each of the chosen quantiles $\tau_k$ is associated with its individual set of constraints. Formulation (4.9) arises from the model

$$Q(\tau_k | D = 0, X = \mathbf{x}) = \beta_{0k} + \mathbf{x}^\top \boldsymbol{\beta}_X,$$

$$(4.11)$$

$$Q(\tau_k | D = 1, X = \mathbf{x}) = \beta_{0k} + \beta_{Dk} + \mathbf{x}^\top (\boldsymbol{\beta}_X + \boldsymbol{\beta}_{XD}),$$

where for $\tau \in (0,1)$, $Q(\tau | D = 0, X = \mathbf{x})$ and $Q(\tau | D = 1, X = \mathbf{x})$ denote the $\tau$-quantile of $T$ conditional on the value of the status $D$ and covariates equal to $\mathbf{x}$. Accordingly, the constraint $\mathbf{A}\boldsymbol{\beta}_k \geq \mathbf{0}$ reflects the constraint $Q(\tau_k | D = 1, X = \mathbf{x}) \geq Q(\tau_k | D = 0, X = \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, $1 \leq k \leq K$, with the understanding that $\mathcal{X}$ gets replaced by a suitable proxy as discussed in the previous subsection. The above optimization problem thus reduces to a linear program given the well-known reformulation of the check loss function via linear constraints [85].

**Remarks.**

(i) The parameters of the location-scale model and the ROC curve are estimated analogously to the least squares-based approach, cf. Eq. (4.7).

(ii) The order constraints (4.10) can be relaxed to ordering of the medians in alignment with the earlier constraint (4.5) by imposing only one set of constraint for $k = (K + 1)/2$ which corresponds to $\tau = 0.5$.

70

(iii) Note that appears natural to impose the constraints that the intercepts $\{\beta_{0k}\}$ corresponding to $\{\tau_k\}$ are ordered accordingly. However, it can be shown that for any minimizer of (4.9), such ordering will be satisfied, thus it is not necessary to add this constraint to the optimization problem.

### 4.3.3 Heteroscedastic Modeling

Even though the CQR approach in the previous section yields improvements over the least squares-based methods with regard to increased robustness and additional flexibility regarding order-constrained modeling as pointed out in the preceding remark (ii), the CQR approach exhibits a major limitation. Namely, CQR assumes homoscedasticity, i.e., the scale functions $\sigma_0$ and $\sigma_1$ are required to be independent of $X$. This is also implicit in the specific form of the quantile functions (4.11) in which only the intercepts in the two groups defined by $D$ depend on the individual quantiles $\{\tau_k\}$.

An alternative to CQR is the use of a three-stage method consisting of estimation of the location functions based on median regression in stage (I), estimation of the scale functions $\sigma_0$ and $\sigma_1$ in stage (II), and estimation of the ROC curve in stage (III) based on the standardized results obtained from applying the results of the previous two stages. Stage (II) follows a strategy developed in the literature on quantile regression [80]. The details of the three stages are presented in the following. For the sake of clarity, the use of order constraints is addressed separately in a subsequent discussion.

*Stage I*: Solve the median regression problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1, \tag{4.12}$$

and estimate the location parameters $\boldsymbol{b}_0^*$ and $\boldsymbol{b}_1^*$ from the minimizer $\widehat{\boldsymbol{\beta}}^{\mathrm{M}}$ of (4.12) in analogy to the least-squared based method.

*Stage II*: Given $\widehat{\boldsymbol{\beta}}^{\mathrm{M}}$, compute the corresponding "raw" residuals $\mathbf{r} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\mathrm{M}}$, and define

by $|\mathbf{r}| = (|r_i|)_{i=1}^N$ their absolute values. Solve the median regression problem

$$\min_{\boldsymbol{\gamma} \in \mathbb{R}^d} \| |\mathbf{r}| - \mathbf{X}\boldsymbol{\gamma} \|_1 \quad \text{subject to } \mathbf{C}\boldsymbol{\gamma} \geq 0. \tag{4.13}$$

Let $\widehat{\boldsymbol{\gamma}}^{\mathrm{M}} = (\widehat{\gamma}_0^{\mathrm{M}}, (\widehat{\boldsymbol{\gamma}}_X^{\mathrm{M}})^\top, \widehat{\gamma}_D^{\mathrm{M}}, (\widehat{\boldsymbol{\gamma}}_{XD}^{\mathrm{M}})^\top)$ denote the minimizer of (4.13). The parameters $\boldsymbol{a}_0^*$ and $\boldsymbol{a}_1^*$ of the scale functions $\sigma_0(\mathbf{x}; \boldsymbol{a}_0^*)$ and $\sigma_1(\mathbf{x}; \boldsymbol{a}_1^*)$ are estimated as $\widehat{\boldsymbol{a}}_0 = (\widehat{\gamma}_0^{\mathrm{M}}, (\widehat{\boldsymbol{\gamma}}_X^{\mathrm{M}})^\top)^\top$ and $\widehat{\boldsymbol{a}}_1 = (\widehat{\gamma}_0^{\mathrm{M}} + \widehat{\gamma}_D^{\mathrm{M}}, (\widehat{\boldsymbol{\gamma}}_X^{\mathrm{M}} + \widehat{\boldsymbol{\gamma}}_{XD}^{\mathrm{M}})^\top)^\top$, respectively (cf. Figure 4.2). The linear constraint $\mathbf{C}\boldsymbol{\gamma}$ is added to enforce the non-negativity of the two-scale functions, i.e., $\sigma_0(\mathbf{x}; \boldsymbol{a}_0^*) \geq 0$ and $\sigma_1(\mathbf{x}; \boldsymbol{a}_1^*) \geq 0$ for all $\mathbf{x} \in \mathcal{X}$. In order to ensure computational tractability, the constraint is formulated in terms of an enclosing hyperrectangle $\overline{\mathcal{X}}$ or the $\{\mathbf{x}_i\}_{i=1}^N$ as elaborated above. In the former case, $\mathbf{C}$ has $2q$ rows, two per vertex of $\overline{\mathcal{X}}$: $[1 \; v_\ell^\top \; \mathbf{0}_{p+1}]$ and $[\mathbf{0}_{p+1} \; 1 \; v_\ell^\top]$, $\ell = 1, \ldots, q$; in the latter case, the $\{\mathbf{x}_i\}_{i=1}^N$ are assume the roles of the vertices, and accordingly $\mathbf{C}$ has $2N$ rows.

Let us comment on the rationale behind the formulation (4.13) that can be found in He's paper [80]. First, note that without loss of generality, we may assume that the error terms $e_0$ and $e_1$ in (4.2) are scaled such that the median of $|e_0|$ and $|e_1|$ equal one. Observe further that according to (4.2)

$$|T - \mu_j(X; \boldsymbol{\beta}_0^*)| \, \big| D = j \overset{\mathcal{D}}{=} |e_j| \sigma_j(X; \boldsymbol{a}_j^*), \; j \in \{0, 1\}, \tag{4.14}$$

where $\overset{\mathcal{D}}{=}$ denotes equality in distribution. In particular, the medians of those distributions are given by $\sigma_j(X; \boldsymbol{a}_j^*)$, $j \in \{0, 1\}$. With the absolute values of the residuals $|\mathbf{r}|$ obtained from the median regression (4.12) in Stage I serving as counterpart to (4.14), the use of the subsequent median regression (4.13) in Stage II becomes clear.

*Stage III*: Given the results from (4.12) and (4.13), we compute the standardized residuals

$$\widehat{e}_{i0} = \frac{T_i - \widehat{\beta}_0^{\mathrm{M}} - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_X^{\mathrm{M}}}{\widehat{\gamma}_0^{\mathrm{M}} + \mathbf{x}_i^\top \widehat{\boldsymbol{\gamma}}_X^{\mathrm{M}}}, \ i = 1, \ldots, n,$$

$$\widehat{e}_{i1} = \frac{T_i - \widehat{\beta}_0^{\mathrm{M}} - \widehat{\beta}_D^{\mathrm{M}} - \mathbf{x}_i^\top (\widehat{\boldsymbol{\beta}}_X^{\mathrm{M}} + \widehat{\boldsymbol{\beta}}_{XD}^{\mathrm{M}})}{\widehat{\gamma}_0^{\mathrm{M}} + \widehat{\gamma}_D^{\mathrm{M}} + \mathbf{x}_i^\top (\widehat{\boldsymbol{\gamma}}_X^{\mathrm{M}} + \widehat{\boldsymbol{\gamma}}_{XD}^{\mathrm{M}})}, \ i = (n+1), \ldots, N,$$

and obtain $\widehat{G}_0$ and $\widehat{G}_1$ as the empirical CDFs of the $\{\widehat{e}_{i0}\}$ and $\{\widehat{e}_{i1}\}$, respectively. Finally, in alignment with (4.4), the covariate-specific ROC is estimated as

$$\widehat{\mathrm{ROC}}_{\mathbf{x}}(u) = 1 - \widehat{G}_1 \left( \frac{\widehat{\gamma}_0^{\mathrm{M}} + \mathbf{x}^\top \widehat{\boldsymbol{\gamma}}_X^{\mathrm{M}}}{\widehat{\gamma}_0^{\mathrm{M}} + \widehat{\gamma}_D^{\mathrm{M}} + \mathbf{x}^\top (\widehat{\boldsymbol{\gamma}}_X^{\mathrm{M}} + \widehat{\boldsymbol{\gamma}}_{XD}^{\mathrm{M}})} \left( \widehat{G}_0^{-1}(1 - u) - \frac{\widehat{\beta}_D^{\mathrm{M}} + \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{XD}^{\mathrm{M}}}{\widehat{\gamma}_0^{\mathrm{M}} + \mathbf{x}^\top \widehat{\boldsymbol{\gamma}}_X^{\mathrm{M}}} \right) \right).$$

*Order constraints.* The order constraint (4.5) can be incorporated into Stage I by adding the linear constraint $\mathbf{A}\boldsymbol{\beta} \geq \mathbf{0}$ to the median regression problem (4.12), where $\mathbf{A}$ has the same form as for weighted least squares regression (cf. (3.9) and the subsequent discussion). We note that a constraint similar to ordering of the locations (4.5) can be formulated for the scale functions, i.e.,

$$\sigma_1(\mathbf{x}; \boldsymbol{a}_1^*) \geq \sigma_0(\mathbf{x}; \boldsymbol{a}_0^*) \ \Leftrightarrow \ \mathbf{x}^\top (\widetilde{\boldsymbol{a}}_1^* - \widetilde{\boldsymbol{a}}_0^*) + (a_{01}^* - a_{00}^*) \geq 0 \quad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{4.15}$$

The latter constraint can be implemented analogously as the constraint for the locations by means of a linear constraint on $\boldsymbol{\gamma}$ in the median regression problem (4.13) for the scale function. While such constraint and its implementation are not investigated further in the present paper, it is worth noting that it can be treated within a common framework. In addition, the combination of both (4.5) and (4.15) can be a practically convenient way of achieving a stochastic ordering of $T|D = 1, X = \mathbf{x}$ and $T|D = 0, X = \mathbf{x}$ for all $\mathbf{x} \in \mathcal{X}$.

## 4.4  Simulation Studies

In this section, we compare the value of bias, standard error (SE), and mean squared error (MSE) of the covariate-specific ROC curve using different approaches with different sample sizes. Let $T_i$ be the score of the $i$-th subject, where $i = 1, 2, \ldots, N$. Let $X_i \sim U(0, 1)$ be the covariate, and $D_i \sim$ Bernoulli(0.5) be the binary status. Let $e_{ij}$ be zero mean random error, $1 \leq i \leq N$, $j = \{0, 1\}$. The simulation data is generated with the model

$$T_i = 1 + X_i + D_i + X_i \cdot D_i + e_{i1}D_i \times \phi(X_i) + e_{i0}(1 - D_i),$$

where $\phi(X_i) = a + b \cdot X_i$ is a function of $X_i$ and controls the variance of the diseased group. Four settings are considered for the error terms $\{e_{ij}\}$ as following:

- Study N : $\{e_{i0}\}$ and $\{e_{i1}\}$ are i.i.d. from the $N(0, 1)$-distribution.
- Study CN: $\{e_{i0}\}$ and $\{e_{i1}\}$ are i.i.d. from a contaminated normal distribution $0.95N(0, 1) + 0.05N(0, 100)$.
- Study T: $\{e_{i0}\}$ and $\{e_{i1}\}$ are i.i.d. from the t distribution with degree of freedom 2.
- Study Cauchy: $\{e_{i0}\}$ and $\{e_{i1}\}$ are i.i.d. from the Cauchy distribution with location parameter 1 and scale parameter 0.

The methods that are applied in this study are:

> WLS: weighted least square method
>
> cWLS: order-constrained weighted least square method
>
> HM: He's regression method
>
> CQR: composite quantile regression method
>
> ORCQR: order-restricted composite quantile regression method

The cWLS method is the coder-constrained method discussed in Chapter 3. The CQR follows the composite quantile regression discussed in §4.3.2 with $K = 5$. The HM method is the abovementioned three-stage approach which can be found in §4.3.3, and the ORCQR method is our proposed method. Sample sizes 20 and 100 are considered in the study, and

bias, SE, and MSE are derived base on the average of 1000 iterations.

Several values of $\phi(X)$ are considered in the study to explore the performance of our pro-posed method. The values of the parameters $(a, b)$ of the function $\phi(X)$ are $\{(1, 0), (4, 0), (0, 5)\}$ in the simulation study, and thus the corresponding values of $\phi(X)$ are $\{1, 4, 5X\}$. Note that when $\phi(X) = 1$, the CDFs of $e_0$ and $e_1$ are equivalent.

Table 4.1 provides the results of bias, standard error (SE), and mean squared error (MSE) with different values of $\phi(X)$ and different sample sizes. We note that as the sample size increases, the values of SE and MSE decrease, but the bias does not change much. We emphasized the smallest MSE within each $N$ and $\phi(X)$ in boldface. It can be seen from the table that the MSE of our proposed method is the smallest except when $G_0$ and $G_1$ are from normal distributions (Study N).

Table 4.1: Bias, standard error (SE), and mean squared error (MSE) of $\widehat{\mathrm{ROC}}_x(u)$ using different methods with $x = 0.5$ and $u = 0.5$. All values of bias and SE have been multiplied by 100.

| Study | | | N | | | CN | | | T | | | Cauchy | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | $\phi(X)$ | Method | Bias | SE | MSE | Bias | SE | MSE | Bias | SE | MSE | Bias | SE | MSE |
| 50 | 1 | WLS | 0.04 | 6.1 | 0.37 | -17.91 | 22.5 | 8.26 | -1.34 | 9.5 | 0.92 | -9.38 | 18.3 | 4.24 |
| | | cWLS | 0.04 | 6.1 | **0.37** | -10.67 | 16.5 | 3.85 | -0.72 | 8.4 | 0.72 | -5.16 | 13.5 | 2.09 |
| | | HM | -0.74 | 6.1 | 0.37 | -0.80 | 6.7 | 0.46 | -0.51 | 8.1 | 0.66 | -0.76 | 9.5 | 0.90 |
| | | CQR | -0.06 | 6.2 | 0.38 | -0.13 | 6.8 | 0.46 | -0.12 | 8.1 | 0.66 | -0.47 | 9.6 | 0.93 |
| | | ORCQR | -0.03 | 6.2 | 0.38 | -0.07 | 6.7 | **0.45** | -0.03 | 7.9 | **0.63** | -0.12 | 9.1 | **0.82** |
| | 4 | WLS | 0.51 | 10.3 | 1.06 | -5.59 | 14.4 | 2.40 | -2.04 | 18.3 | 4.39 | -4.02 | 13.4 | 1.95 |
| | | cWLS | 0.53 | 10.2 | **1.05** | -2.21 | 12.1 | 1.52 | 0.06 | 10.4 | 1.07 | -1.49 | 11.4 | 1.33 |
| | | HM | 0.77 | 10.6 | 1.18 | 0.78 | 10.5 | 1.12 | 0.30 | 10.6 | 1.12 | -0.16 | 10.6 | 1.13 |
| | | CQR | 0.60 | 10.5 | 1.11 | 0.56 | 10.6 | 1.12 | 0.18 | 10.7 | 1.14 | -0.29 | 10.9 | 1.19 |
| | | ORCQR | 0.54 | 10.3 | 1.06 | 0.64 | 10.2 | **1.05** | 0.46 | 10.2 | **1.05** | 0.37 | 10.3 | **1.06** |
| 100 | 1 | WLS | -0.01 | 4.3 | 0.18 | -17.44 | 17.3 | 6.03 | -0.78 | 6.4 | 0.41 | -9.17 | 15.6 | 3.27 |
| | | cWLS | -0.01 | 4.3 | 0.18 | -12.10 | 13.6 | 3.32 | -0.46 | 5.8 | 0.34 | -5.37 | 11.3 | 1.56 |
| | | HM | -0.40 | 4.2 | 0.18 | -0.57 | 4.7 | 0.22 | -0.22 | 5.5 | 0.30 | -0.24 | 6.2 | 0.39 |
| | | CQR | -0.04 | 4.3 | 0.18 | -0.17 | 4.8 | 0.23 | 0.01 | 5.5 | 0.30 | -0.08 | 6.2 | 0.38 |
| | | ORCQR | -0.03 | 4.3 | **0.18** | -0.16 | 4.7 | **0.22** | 0.03 | 5.5 | **0.30** | -0.01 | 6.1 | **0.37** |
| | 4 | WLS | 0.20 | 7.1 | 0.50 | -6.46 | 9.7 | 1.36 | -0.64 | 7.4 | 0.55 | -4.00 | 10.2 | 1.21 |
| | | cWLS | 0.22 | 7.0 | 0.50 | -3.24 | 8.6 | 0.85 | -0.18 | 7.2 | 0.51 | -1.83 | 8.5 | 0.75 |
| | | HM | 0.42 | 7.0 | 0.50 | 0.25 | 7.0 | 0.49 | 0.15 | 7.2 | 0.51 | 0.02 | 7.2 | 0.52 |
| | | CQR | 0.25 | 7.1 | 0.51 | 0.13 | 7.1 | 0.51 | 0.07 | 7.3 | 0.53 | -0.17 | 7.4 | 0.55 |
| | | ORCQR | 0.18 | 7.1 | **0.50** | 0.10 | 7.0 | **0.49** | 0.17 | 7.2 | **0.51** | 0.13 | 7.2 | **0.52** |

Table 4.2 shows the results when the value of $\phi$ depends on the covariate $X$. In the exception of Study N, our proposed method has the smallest MSE which emphasized in

boldface when the sample size is 50. However, when the sample size increases to 100, the MSE of He's method becomes the smallest in most of the studies. Moreover, the smallest bias is highlighted using italicize and boldface text. We recommend He's method when the data with heteroscedastic errors since He's method leads to a small bias and even a small MSE when the sample size is large.

Table 4.2: Bias, standard error (SE), and mean squared error (MSE) of $\widehat{\mathrm{ROC}}_x(u)$ using different methods with $x = 0.5$ and $u = 0.5$ in the heteroscedastic model. All values of bias and SE have been multiplied by 100.

| Study | | | N | | | CN | | | T | | | Cauchy | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | $\phi(X)$ | Method | Bias | SE | MSE | Bias | SE | MSE | Bias | SE | MSE | Bias | SE | MSE |
| 50 | 5$X$ | WLS | 3.50 | 10.4 | 1.20 | -7.65 | 18.2 | 3.91 | 1.19 | 12.7 | 1.62 | -5.64 | 16.6 | 3.06 |
| | | cWLS | 4.17 | 9.3 | **1.03** | -2.72 | 14.3 | 2.13 | 2.54 | 10.4 | 1.17 | -1.47 | 13.5 | 1.83 |
| | | HM | *0.87* | 10.8 | 1.18 | *0.59* | 11.0 | 1.21 | *0.34* | 11.1 | 1.24 | *-0.07* | 11.4 | 1.30 |
| | | CQR | 3.74 | 10.8 | 1.30 | 3.15 | 11.2 | 1.36 | 2.89 | 11.7 | 1.46 | 1.69 | 12.6 | 1.62 |
| | | ORCQR | 4.58 | 9.4 | 1.09 | 4.20 | 9.6 | **1.09** | 3.97 | 10.1 | **1.17** | 3.07 | 10.8 | **1.27** |
| 100 | 5$X$ | WLS | 4.08 | 6.8 | 0.63 | -8.77 | 13.5 | 2.60 | 1.31 | 9.5 | 0.92 | -5.34 | 13.6 | 2.13 |
| | | cWLS | 4.26 | 6.4 | 0.60 | -4.46 | 11.5 | 1.52 | 2.99 | 7.4 | 0.64 | -1.46 | 10.7 | 1.16 |
| | | HM | *0.33* | 7.7 | **0.59** | *-0.25* | 7.9 | 0.63 | *-0.20* | 7.7 | **0.60** | *0.25* | 7.6 | **0.58** |
| | | CQR | 4.29 | 6.9 | 0.66 | 3.86 | 7.2 | 0.66 | 3.74 | 7.8 | 0.75 | 3.10 | 8.2 | 0.77 |
| | | ORCQR | 4.57 | 6.4 | 0.62 | 4.22 | 6.6 | **0.61** | 4.19 | 7.0 | 0.62 | 3.73 | 7.2 | 0.66 |

Figure 4.3 displays the bias and MSE of the ROC curves with different sample sizes $N$ when $x = 5$ and $u = 0.5$ by using 10000 replicates. In Figure 4.3, we investigate the bias and MSE using different methods when the scale function $\phi(X)$ are 1 and 5$X$. Note that when $\phi(X) = 5X$, the data is heteroscedastic. We see that the bias and MSE of WLS and cWLS methods are larger than other methods in Study CN, Study T, and Study Cauchy, but close to others in Study N. When $\phi(X) = 5X$, we find that the bias of HM is constantly close to 0, while other methods are highly biased. The MSE of HM is slightly smaller than ORCQR when the sample size is large.
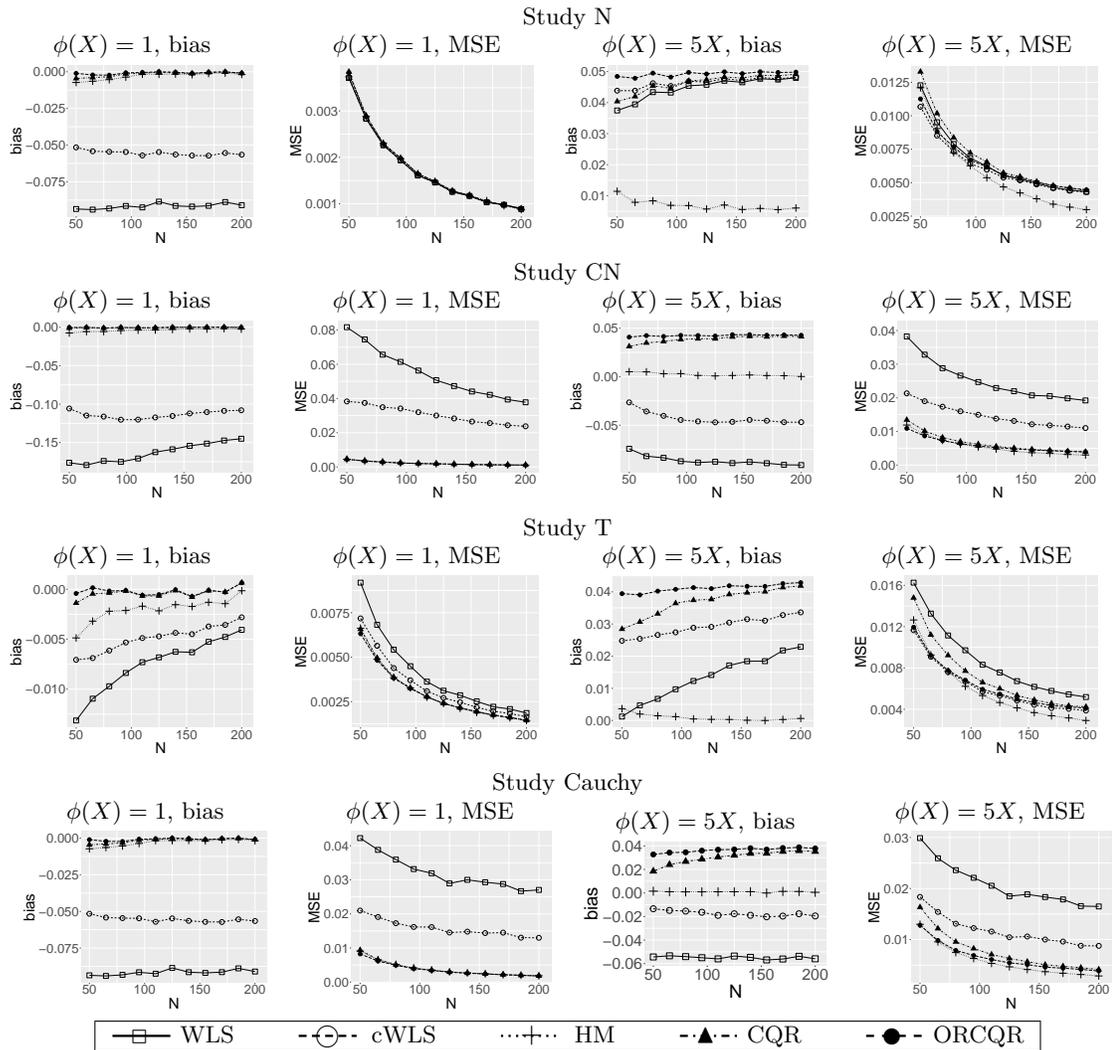
Figure 4.3: Bias and mean squared error (MSE) of $\widehat{\mathrm{ROC}}_x(u)$ in dependence of sample size $N$ using different methods with $x = 0.5$ and $u = 0.5$.

## 4.5 Real Examples

In this section, we report the results of the real study in the application of our proposed method. We compare the mean, SD, and variance ration (VR) of the estimated mean difference and covariate-specific ROC curves using weighted least square (WLS) method, He's regression (HM) method, composite quantile regression (CQR), and order-restricted composite quantile regression (ORCQR) using biometric data.

The Face Recognition Vendor Test (FRVT) 2006 by [76] includes biometric scores for face-pairs and the features of each image. Features, for example, are including race, gender, and image quality, etc. This face recognition data has been used for many studies such as facial recognition algorithm studies [77, 78] and the study about the influence of image quality on the classification accuracy [39]. However, the discussion about the impact of covariates has been scarce. In this section, we first apply the order-restricted composite quantile regression ROC curve on the data, and then discuss the heteroscedastic model using the data. We first define the notations used in the rest of this section. Let $T_{ij} = s(I_i, I_j)$ be the biometric score which reflects the similarity of the facial image pair $(I_i, I_j)$, where $i$ and $j$ are the IDs of two subjects. $T_{ij}$ belongs to the genuine group if two images are from the same source, i.e., $i = j$; otherwise, it belongs to the imposter group. The error rates associated with binary decisions in forensic analysis are given by the false accept rate (FAR) and the true accept rate (TAR) in analogy to the FPR and TPR, respectively, in medical studies.

### 4.5.1 Facial Recognition Data with Order-Restricted Regression

In this section, we demonstrate the application of ORCQR method and compare it with the traditional CQR method. To apply the ORCQR method, we need to remove the correlation of the scores. We divided the subject IDs into two subgroups, $\mathcal{I}_1$ and $\mathcal{I}_2$, randomly. We then selected the genuine scores with the ID from $\mathcal{I}_1$, i.e. $\{T_{ii} | D_{ii} = 1, i \in \mathcal{I}_1\}$, and selected the imposter score with the ID from $\mathcal{I}_2$, i.e. $\{T_{ij} | D_{ij} = 0, i, j \in \mathcal{I}_2\}$. This approach removed

scores with duplicate IDs across the genuine and imposter groups. After the first step, we further filtered the data by separating $\mathcal{I}_2$ in two equal size subgroups. Then we obtained the imposter score IDs by pairing the two subgroups. In this way, all the IDs in the imposter group only appears once. Furthermore, We removed the scores belongs to the image quality categories "bad" and "ugly" since scores in this two categories perform poor discriminative ability. The observed data are modeled using the linear model as

$$T_{ij} = \beta_0 + \beta_1 D_{ij} + \beta_2 \cdot \mathtt{A}_i + \beta_3 \cdot \mathtt{A}_j + \beta_4(\mathtt{A}_i \cdot D_{ij}) + D_{ij}\epsilon_{ij1} + (1 - D_{ij})\epsilon_{ij0}. \qquad (4.16)$$

The status variable $D_{ij} = 1$ when the subject belongs to the genuine group and 0 other wise. The two covariates $\mathtt{A}_i$ and $\mathtt{A}_j$ are ages of the two subjects in each image pair. Note that there is only one interaction term $\mathtt{A}_i \cdot D_{ij}$ being considered because the interaction term dose not exist when $D_{ij} = 0$, and this leads to $\mathtt{A}_i \cdot D_{ij} = \mathtt{A}_j \cdot D_{ij}$. We follow the suggestion of Givens *et al.* [79] and specific the ages covariates to be (17, 23, and 30) since these three age groups represent younger through older subjects, respectively.

The results including mean and SD are generated using 1000 bootstraps. The variance ratios (VR), which defended as VR $= var_1/var_2$, is calculated over all bootstrap replications. In the equation of VR, $var_1$ is the variance of CQR method and $var_2$ is the variance of the proposed method. Therefore, a VR larger than 1 implies that the performance of the proposed method is better than the conventional method. In each bootstrap replication, the value of $\mathrm{ROC}_x(u)$ is estimated using the empirical method.

Table 4.3 shows the mean and SD of the covariate-specific ROC over 1000 iterations for different values $x$ of the covariate age and different values of $p$ when using different methods. The VR is larger than one throughout. Although these two methods generate a similar mean result, the performance of ORCQR is better than CQR in terms of SD. At the ages of 23, all the VRs are close to 1.0.

Table 4.3: Mean and SD of $ROC_x(u)$ for different values of $x$ and different values of $u$ for the facial recognition data. (CQR: composite quantile regression; ORCQR: order-restricted composite quantile regression.)

| $x$ | | 17 | | | 23 | | | 30 | | |
|-----|--------|-------|------|------|-------|------|------|-------|------|------|
| $u$ | Method | mean | SD | VR | mean | SD | VR | mean | SD | VR |
| 0.1 | CQR | 0.537 | 0.103 | 1.45 | 0.609 | 0.071 | 1.02 | 0.680 | 0.091 | 1.57 |
| | ORCQR | 0.543 | 0.085 | | 0.607 | 0.070 | | 0.688 | 0.077 | |
| 0.3 | CQR | 0.897 | 0.046 | 1.12 | 0.925 | 0.031 | 1.13 | 0.936 | 0.049 | 2.87 |
| | ORCQR | 0.896 | 0.044 | | 0.927 | 0.030 | | 0.945 | 0.029 | |
| 0.5 | CQR | 0.948 | 0.022 | 1.14 | 0.963 | 0.015 | 1.01 | 0.967 | 0.022 | 2.03 |
| | ORCQR | 0.948 | 0.021 | | 0.964 | 0.015 | | 0.970 | 0.015 | |
| 0.7 | CQR | 0.969 | 0.016 | 1.06 | 0.977 | 0.013 | 1.02 | 0.979 | 0.014 | 1.29 |
| | ORCQR | 0.969 | 0.015 | | 0.977 | 0.012 | | 0.980 | 0.012 | |
| 0.9 | CQR | 0.980 | 0.012 | 1.04 | 0.984 | 0.011 | 1.11 | 0.988 | 0.010 | 1.17 |
| | ORCQR | 0.980 | 0.012 | | 0.984 | 0.010 | | 0.989 | 0.009 | |

Figure 4.4 compares the estimated covariate-specific ROC curves using CQR and OR-CQR methods for different ages. The SD of the ROC was calculated using the bootstrap method, and we generate the pointwise confidence intervals based on the SD for these two approaches. The estimated values are similar to each other but the error bands are variate. When $x = 23$, the covariate specific ROC curves using the method with order constraint has a similar confidence interval to the method without order constraint. However, when $x = 17$ and $x = 30$, the confidence intervals using our ORCQR are narrower than the conventional method, especially when $x = 30$.

Figure 4.5 graphs the variance results of the ROC curves against FAR with $x = 30$ and the covariate age with FAR $u = 0.5$ using CQR and ORCQR methods. The variance generates using the bootstrap over 1000 iterations. The variance of ORCQR is smaller than CQR thoroughly Figure 4.5a shows that the order-constraint noticeable reduces the variance when FAR less than 0.5, and then the ORCQR method converges to CQR method as FAR increases. In Figure 4.5b, we note that the variances of the conditional ROC for the two methods are similar to each other when the value of age is small. The variance of the CQR method increases fast as age increases but the variance of the ORCQR method is insensitive to the age increases.

|  |  |  |
| :---: | :---: | :---: |
| (a) $x = 17$ | (b) $x = 23$ | (c) $x = 30$ |
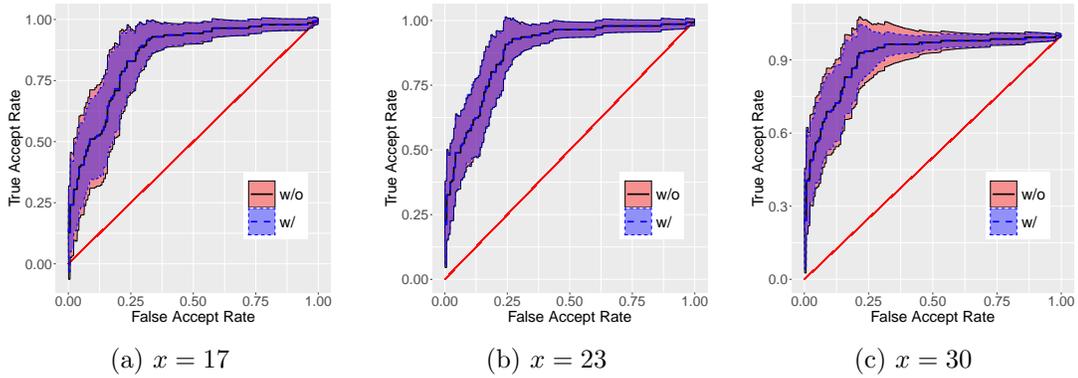
Figure 4.4: Covariate-specific ROC curves for different ages. (w/o: composite quantile regression without order constraint; w/: composite quantile regression with order constraint. The shaded area represents pointwise 95% confidence intervals; best seen in color)



|  |  |
| :---: | :---: |
| (a) | (b) |
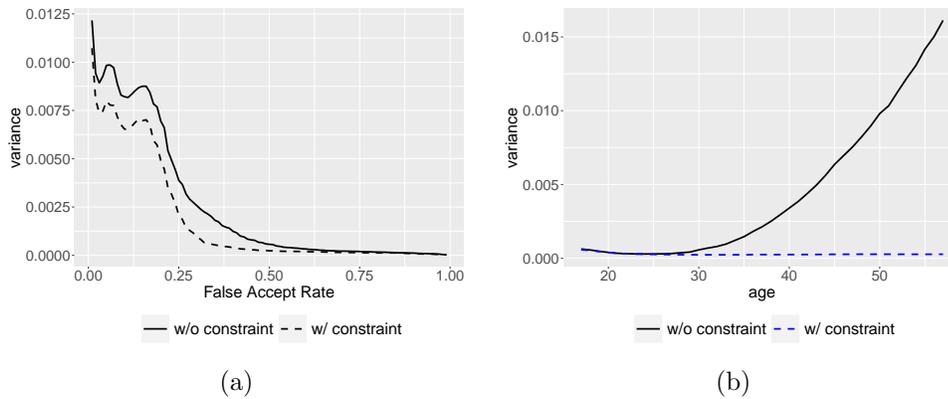
Figure 4.5: Bootstrap variance of the estimated covariate-specific ROC in dependence of $u$ when the covariate age is 30 (a), and different values of the covariate age when $u$ is 0.5 (b). (w/o: composite quantile regression without order constraint; w/: composite quantile regression with order constraint.)

### 4.5.2 Facial Recognition Data with Heteroscedastic Modeling

In this section, we compare the performance of He's method (HM) with the CQR and the weighted least square using grouped data (WLS$_x$). We use the facial recognition data but consider all three image quality categories "good", "bad", and "ugly". As we introduced in §4.1, the imaging quality influences the variance and thus estimating the scale function

with traditional methods will bring extra bias. Let $\mathtt{B}_{ij}$ and $\mathtt{U}_{ij}$ denote two dummy variables for quality "bad" and "ugly", respectively. Therefore the quality can be expressed using the duplet $(\mathtt{B}_{ij}, \mathtt{U}_{ij})$. The linear model in this study is given by

$$
\begin{aligned}
T_{ij} =& \beta_0 + \beta_1 D_{ij} + \beta_2 \cdot \mathtt{B}_{ij} + \beta_3 \cdot \mathtt{U}_{ij} + \beta_4(\mathtt{B}_{ij} \cdot D_{ij}) + \beta_4(\mathtt{U}_{ij} \cdot D_{ij}) \\
& + \alpha_1(\mathtt{B}_{ij}, \mathtt{U}_{ij})D_{ij}\epsilon_{ij1} + \alpha_0(\mathtt{B}_{ij}, \mathtt{U}_{ij})(1 - D_{ij})\epsilon_{ij0}.
\end{aligned} \tag{4.17}
$$

We consider the covariate duplet $(\mathtt{B}_{ij}, \mathtt{U}_{ij})$ to be $(0,0)$, $(1,0)$, and $(0,1)$ which stand for "good", "bad", and "ugly" quality. We use grouped composite quantile regression $(\mathrm{CQR}_x)$ as a reference. The $\mathrm{CQR}_x$ method estimates the location-scale model based on the CQR within each subgroup which depends on the covariate quality. Similarly, the $\mathrm{WLS}_x$ method estimates the scale function and the weight for each subgroup depending on the quality of image, but estimates the location function with WLS regression. The CQR method is the method that we introduced in §4.3.2 with $K = 5$. We here compare the performance of the $\mathrm{WLS}_x$, He's method (HM), and CQR methods. We apply the bootstrap with 1000 iterations and provide the SD and VR of the bootstrap results. The VR is the ratio of the variance of a given method with respect to the variance of $\mathrm{WLS}_x$.

Table 4.4 shows the bias and SD of the covariate-specific ROC over 1000 bootstrap iterations for different qualities and different FPR. We do not recommend using the CQR method in a heteroscedastic model since it is highly bias. On the contrary, the bias using HM is relatively small, and the loss in statistical efficiency relative to $\mathrm{WLS}_x$ is moderate.

Table 4.4: Bias and SD of the covariate-specific ROC for different qualities and different values of $u$ for the facial recognition data. (WLS$_x$: grouped weighted least square; HM: He's method; CQR: composite quantile regression)

| quality | | good | | | bad | | | ugly | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $u$ | Method | bias | SD | VR | bias | SD | VR | bias | SD | VR |
| | WLS$_x$ | -0.034 | 0.119 | 1.00 | -0.033 | 0.089 | 1.00 | 0.037 | 0.088 | 1.00 |
| 0.1 | HM | 0.102 | 0.080 | 2.17 | -0.114 | 0.103 | 0.74 | 0.048 | 0.104 | 0.72 |
| | CQR | -0.300 | 0.089 | 1.77 | -0.262 | 0.081 | 1.20 | -0.108 | 0.035 | 6.36 |
| | WLS$_x$ | -0.010 | 0.050 | 1.00 | 0.051 | 0.121 | 1.00 | 0.026 | 0.098 | 1.00 |
| 0.3 | HM | 0.010 | 0.036 | 1.91 | -0.007 | 0.134 | 0.81 | 0.078 | 0.138 | 0.51 |
| | CQR | -0.182 | 0.095 | 0.28 | -0.251 | 0.085 | 1.99 | -0.197 | 0.096 | 1.04 |
| | WLS$_x$ | 0.009 | 0.024 | 1.00 | 0.011 | 0.066 | 1.00 | -0.084 | 0.119 | 1.00 |
| 0.5 | HM | 0.009 | 0.024 | 1.03 | -0.010 | 0.076 | 0.76 | 0.015 | 0.118 | 1.02 |
| | CQR | -0.101 | 0.059 | 0.17 | -0.291 | 0.103 | 0.41 | -0.315 | 0.084 | 2.03 |
| | WLS$_x$ | 0.021 | 0.019 | 1.00 | 0.003 | 0.058 | 1.00 | -0.022 | 0.115 | 1.00 |
| 0.7 | HM | 0.018 | 0.021 | 0.78 | 0.001 | 0.063 | 0.86 | 0.068 | 0.092 | 1.56 |
| | CQR | -0.015 | 0.042 | 0.20 | -0.137 | 0.101 | 0.33 | -0.203 | 0.099 | 1.35 |
| | WLS$_x$ | -0.003 | 0.008 | 1.00 | -0.033 | 0.035 | 1.00 | 0.024 | 0.064 | 1.00 |
| 0.9 | HM | -0.010 | 0.015 | 0.28 | -0.030 | 0.032 | 1.24 | 0.094 | 0.060 | 1.15 |
| | CQR | -0.019 | 0.019 | 0.74 | -0.112 | 0.065 | 0.30 | 0.075 | 0.109 | 0.35 |

# Chapter 5: Concluding Remarks and Future Research

In this thesis, we have presented several applications of the ROC curve and the likelihood ratio to data from forensic evidence analysis. First, we discussed and compared three popular approaches for the log-likelihood ratio (LLR) estimation, which are parametric estimation (PE), kernel density estimation (KDE), and logistic regression estimation (LRE). Furthermore, we proposed a parametric method to estimate the variance of the LLR based on the ROC curve. We discussed the application of the ROC curve when the discriminative power of the considered test depends on covariates. We provided a location-scale model procedure using linear regression or composite quantile regression (CQR) to incorporate the influence of the covariates. As a consequence, this approach addresses the problem when the accuracy of the identification test is potentially influenced by some covariates. Ultimately, we proposed the order-constrained method to consider the stochastic ordering information and introduced the restricted regression quantiles (RRQ) method to delta with the heteroscedasticity issue in a biometric example. Both the order-constrained ROC and the RRQ method help us improve the statistical efficiency of the covariate-specific ROC curves.

We compared the performances of these methods, PE, KDE, and LRE, using simulation studies and two biometric studies in Chapter 2. In the simulation studies, the KDE outperforms LRE when the sample sizes of the two groups are not equal. The LLR generated by LRE has a smaller variance than the one generated by KDE. Each study justifies the repeatability of the PE and KDE as the LLR values have little fluctuation for various sample size ratios. The LRE shows unsatisfactory repeatability as the LLR values tend to have linear relationship with the log sample size ratios. If the distribution assumption is valid for PE, then the PE and KDE methods have similar LLR values, indicating reproducibility between PE and KDE. Our results showed the reproducibility among the three methods

is poor when the distribution assumption is invalid for PE, since the LLR values from PE and KDE are quite different. Also, since LLR values from the LRE method depend on the sample size ratio, the LLR values are inconsistent of this three methods. Based on the assessment from simulation datasets and the real datasets, the PE method is recommended for estimating the LLR if appropriate distribution assumptions can be made for the datasets. Otherwise, the KDE method would be a better alternative for estimating the LLR. The inconsistency of the LRE method with varying sample size ratios leads to our recommendation against the usage of the method unless the true ratio between genuine and imposter groups is justified.

In Chapter 3, we studied the proposed order-constrained linear regression for estimating covariate-specific ROC curves based on the location-scale model by González-Manteiga *et al.* [34], with a focus on application to data from biometric recognition systems. In the context of diagnostic medicine, Pepe [20] points out the importance of the covariate-specific ROC curves when covariates impact the performance of a diagnostic test. Many works have been done afterwards in various aspects of covariate-specific ROC curves [33, 34, 86]. The main technical innovation in the present dissertation is the use of an ordering constraint that potentially plays an important role in fields such as medical diagnostics, toxicology [87], and biometrics, in order to yield more precise estimators of distribution functions [71]. In this thesis, we developed an order-constrained ROC regression method and apply the method to estimate age-specific ROC curves for a facial recognition study. While the covariates in the example section are demographic variables of the source subjects, the proposed framework also allows for the integration of information about forensic examiners if the accuracy of examiners is of primary interest. Theoretical analysis, simulation studies, and the case study unanimously show that the presence of the order constraint yields a more favorable bias-variance trade-off as a consequence of variance reductions that can be rather substantial.

Chapter 4 incorporates the stochastic ordering condition into the covariate-specific ROC estimation using the composite quantile regression (CQR). The proposed method acquires

the advantages of CQR and obtains a reduction of the mean squared error concerning the unconstrained version. We compared the proposed method with other methods in estimating the ROC curves depending on the age variable using facial recognition study. The results show that the proposed method has a smaller variance than other approaches. This advantage has also been proved in the simulation study.

We also adopted He's [80] method (HM) to estimate the scale function in the heteroscedastic data. In the simulation study, HM performs its advantage in terms of bias when the variance is a function of the covariates. We also compared the HM with grouped weighted least square ($WLS_x$) and CQR methods in the facial recognition data whose variance is influenced by the image quality. The results demonstrate that the bias of HM is lower than $WLS_x$ and CQR. As a conclusion, we recommend the HM when the variance of the data depends on the covariates.

There are several research directions for us to work on in the future. The first can be made from Chapter 2 where the LLR values were calculated based on the datasets in facial recognition and fingerprint matching. Since the LLR values may change if multiple face and fingerprint datasets are used in the study, it is interesting to assess the reproducibility using multiple face and fingerprint datasets. The issue of choice of dataset has been highlighted by Jain *et al.* [88, 89]. If the LLR values are not consistent across multiple datasets for a single biometric modality, careful consideration on selecting an appropriate dataset for LLR estimation is necessary to ensure that the characteristics of the selected dataset represent the background information of forensic cases.

Another important direction for these covariate-specific ROC curve studies concerns the development of approaches that can systematically deal with dependencies among biometric scores arising from pairwise comparisons of individuals, a scenario that we have encountered in the case study presented herein. Furthermore, In the real study, more information can be acquired from the picture as the quality of the image improves, and thus the variability in the scores is getting smaller. Therefore, an alternative topic is to encounter the order constraint of the variance additional to the mean in the future to improve the statistical

efficiency.

# Appendix A: Appendix to Chapter 3

This supplement is organized as follows. Sections A.1 and A.2 contain proofs of the two statements presented in the body of the paper. Some supplementary statements and their proofs are contained in Sections A.3 and A.4. Section A.5 is dedicated to additional simulation results.

## A.1   Proof of Proposition 3.1

We start by showing that

$$
\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} \overset{\mathcal{D}}{=} \boldsymbol{\beta}^* + \frac{\sigma}{\sqrt{N}}\boldsymbol{\zeta}, \qquad \boldsymbol{\zeta} \sim N_d(\mathbf{0}, \boldsymbol{\Omega}), \qquad \boldsymbol{\Omega} = \begin{bmatrix} \frac{N}{n}\mathbf{C}_0^{-1} & -\frac{N}{n}\mathbf{C}_0^{-1} \\ -\frac{N}{n}\mathbf{C}_0^{-1} & \frac{N}{n}\mathbf{C}_0^{-1} + \frac{N}{m}\cdot\frac{\tau^2}{\sigma^2}\mathbf{C}_1^{-1} \end{bmatrix},
$$

$$
\mathbf{C}_0 = \tfrac{1}{n}[\mathbf{1}_n\ \mathbf{Z}_0]^\top[\mathbf{1}_n\ \mathbf{Z}_0],\ \mathbf{Z}_0 = [\mathbf{X}_{1\bullet};\ldots;\mathbf{X}_{n\bullet}], \quad \mathbf{C}_1 = \tfrac{1}{m}[\mathbf{1}_m\ \mathbf{Z}_1]^\top[\mathbf{1}_m\ \mathbf{Z}_1],\ \mathbf{Z}_1 = [\mathbf{X}_{n+1\bullet};\ldots;\mathbf{X}_{N\bullet}],
$$

where ";" here denotes row-wise concatenation. Under model (3.16), $\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} \sim N_d(\boldsymbol{\beta}^*, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}$, and it remains to show that $\frac{N}{\sigma^2}\boldsymbol{\Sigma} = \boldsymbol{\Omega}$. Invoking relation (3.17), we obtain

$$
\boldsymbol{\Sigma} = \mathbf{L}^{-1}\begin{bmatrix} \sigma^2\frac{1}{n}\mathbf{C}_0^{-1} & \mathbf{0} \\ \mathbf{0} & \tau^2\frac{1}{m}\mathbf{C}_1^{-1} \end{bmatrix}(\mathbf{L}^\top)^{-1}, \qquad \mathbf{L}^{-1} = \begin{pmatrix} \mathbf{I}_{p+1} & \mathbf{0} \\ -\mathbf{I}_{p+1} & \mathbf{I}_{p+1} \end{pmatrix}, \qquad (\mathbf{L}^\top)^{-1} = (\mathbf{L}^{-1})^\top.
$$

Performing the above matrix multiplication confirms that $\frac{N}{\sigma^2}\cdot\boldsymbol{\Sigma} = \boldsymbol{\Omega}$. To conclude the proof of the proposition, we make use of the primal-dual relation (3.12). According to the previous display, we use that $\frac{N}{\sigma^2}(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1} = \boldsymbol{\Omega}$, and then re-write the minimizer of the

dual problem:

$$
\widehat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}\in\mathbb{R}^q_+}{\operatorname{argmin}} \left\{ \frac{1}{2}\boldsymbol{\lambda}^\top \mathbf{A}(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{A}^\top\boldsymbol{\lambda} + \boldsymbol{\lambda}^\top\mathbf{A}\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} \right\}
$$

$$
= \underset{\boldsymbol{\lambda}\in\mathbb{R}^q_+}{\operatorname{argmin}} \left\{ \frac{1}{2}\boldsymbol{\lambda}^\top \frac{\mathbf{A}(N^{-1}\sigma^2\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{A}^\top}{N/\sigma}\boldsymbol{\lambda} + \boldsymbol{\lambda}^\top\left(\frac{\boldsymbol{\Delta}^*}{\sigma} + \frac{\boldsymbol{\xi}}{\sqrt{N}}\right) \right\}, \ \boldsymbol{\xi}\sim N(\mathbf{0},\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top)
$$

$$
= \underset{\boldsymbol{\lambda}\in\mathbb{R}^q_+}{\operatorname{argmin}} \left\{ \frac{1}{2}\frac{\boldsymbol{\lambda}^\top\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top}{N/\sigma}\boldsymbol{\lambda} + \boldsymbol{\lambda}^\top\left(\frac{\boldsymbol{\Delta}^*}{\sigma} + \frac{\boldsymbol{\xi}}{\sqrt{N}}\right) \right\}.
$$

Using Lemma A.3 below yields that the "argmin" in (3.18) results as

$$
\underset{\boldsymbol{\lambda}\in\mathbb{R}^q_+}{\operatorname{argmin}} \left\{ \frac{1}{2}\boldsymbol{\lambda}^\top\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top\boldsymbol{\lambda} + \boldsymbol{\lambda}^\top\left(\frac{\boldsymbol{\Delta}^*}{\sigma} + \frac{\boldsymbol{\xi}}{\sqrt{N}}\right) \right\} = \frac{\widehat{\boldsymbol{\lambda}}}{N/\sigma}.
$$

Putting together the pieces and simplifying terms yields the assertion.

## A.2   Proof of Corollary 3.1

In view of (3.11) and (3.12), it is clear that $\{\mathbf{A}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}\geq\mathbf{0}\} = \{\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{LS}}\}$. We have $\{\mathbf{A}\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}\geq\mathbf{0}\} = \bigcap_{k=1}^q\{\mathbf{a}_k^\top\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}\geq 0\}$. Using the Gaussian tail bound $P(Z<\mu-t)\leq\exp\left(-\frac{t^2}{2\sigma^2}\right)$, $t>0$ for $Z\sim N(\mu,\sigma^2)$, the fact that $N^{1/2}\mathbf{a}_k^\top\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}\sim N(\Delta_k^*,\sigma^2\mathbf{a}_k^\top\boldsymbol{\Omega}\mathbf{a}_k)$, $k=1,\dots,q$, and a union bound, we obtain that

$$
P\left(\bigcup_{k=1}^q\left\{\mathbf{a}_k^\top\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}<0\right\}\right) \leq \sum_{k=1}^q \exp\left(-N\cdot\frac{\Delta_k^{*2}}{2\sigma^2\cdot\mathbf{a}_k^\top\boldsymbol{\Omega}\mathbf{a}_k}\right).
$$

This concludes the proof of the corollary.

## A.3 Proof of Corollary 3.2

*Proof.* The above corollary follows from Proposition 3.1 by evaluating the "argmin" in (3.18), which reduces to a simple expression. Let $\mathbf{G} = \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top$ and $\boldsymbol{\nu} = -\left(\frac{\boldsymbol{\Delta}^*}{\sigma} + \frac{1}{\sqrt{N}}\boldsymbol{\xi}\right)$ with $\boldsymbol{\xi} \sim N_2(\mathbf{0}, \mathbf{G})$. The optimization problem inside the curly brackets in (3.18) then becomes

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^2} \frac{1}{2}\boldsymbol{\lambda}^\top \mathbf{G}\boldsymbol{\lambda} - \boldsymbol{\nu}^\top \boldsymbol{\lambda} \tag{A.1}$$

Note that we can have (i) $\{\widehat{\lambda}_1 > 0, \widehat{\lambda}_2 > 0\}$, (ii) $\{\widehat{\lambda}_1 > 0, \widehat{\lambda}_2 = 0\}$, (iii) $\{\widehat{\lambda}_1 = 0, \widehat{\lambda}_2 > 0\}$ and (iv) $\{\widehat{\lambda}_1 = 0, \widehat{\lambda}_2 = 0\}$. If case (iv) occurs, the result immediately follows. The optimality conditions of (A.1) imply that case (i) requires that $\mathbf{G}^{-1}\boldsymbol{\nu} \geq \mathbf{0}$ in which case $\widehat{\boldsymbol{\lambda}} = \mathbf{G}^{-1}\boldsymbol{\nu}$. Similarly, case (ii) requires $G_{11}\widehat{\lambda}_1 = \nu_1$ and $G_{12}\widehat{\lambda}_1 \geq \nu_2 \Leftrightarrow (G_{12}/G_{11})\nu_1 - \nu_2 \geq 0$. Case (iii) is analogous to case (ii). It remains to calculate the entries of the matrix $\mathbf{G} = \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top$ with $\boldsymbol{\Omega}$ as in Corollary 3.2 and

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

It suffices to compute the bottom 2-by-2 diagonal block of $\boldsymbol{\Omega}$. Direct calculations show that

$$\mathbf{C}_0^{-1} = \begin{pmatrix} \frac{\bar{x}^2}{s_x^2} + 1 & -\frac{\bar{x}}{s_x^2} \\ -\frac{\bar{x}}{s_x^2} & \frac{1}{s_x^2} \end{pmatrix}, \qquad \mathbf{C}_1^{-1} = \begin{pmatrix} \frac{\bar{z}^2}{s_z^2} + 1 & -\frac{\bar{z}}{s_z^2} \\ -\frac{\bar{z}}{s_z^2} & \frac{1}{s_z^2} \end{pmatrix},$$

and the given expressions for $G_{11}$, $G_{12}$, and $G_{22}$ are then obtained by straightforward computations. $\square$

## A.4 Additional Lemmas

The following lemma states that in the case of constant variances within the two groups defined by $D$, i.e., $\sigma_0(X; \boldsymbol{\alpha}_0^*) \equiv \sigma$ and $\sigma_1(X; \boldsymbol{\alpha}_1^*) \equiv \tau$, least squares and weighted least squares (3.9) yield identical solutions.

**Lemma A.1.** *Consider the weighted least squares criterion*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_2^2$$

*with $\mathbf{X}$ as defined below (3.9) and weight matrix*

$$\mathbf{W} = \mathrm{diag}\left(\underbrace{\frac{1}{\sigma^2}, \dots, \frac{1}{\sigma^2}}_{n \text{ times}}, \underbrace{\frac{1}{\tau^2}, \dots, \frac{1}{\tau^2}}_{m \text{ times}}\right).$$

*We then have $\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} = \widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$, where $\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}}$ denotes the minimizer of the above weighted least squares problem, and $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ denotes the ordinary least squares solution.*

*Proof.* We have

$$\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}$$

$$= (\mathbf{L}^\top \mathbf{Z}^\top \mathbf{W} \mathbf{Z} \mathbf{L})^{-1} \mathbf{L}^\top \mathbf{Z}^\top \mathbf{W} \mathbf{y} \qquad (\text{cf. } (3.17))$$

$$= \mathbf{L}^{-1} \left( \mathbf{Z}^\top \mathbf{W} \mathbf{Z} \right)^{-1} \mathbf{Z}^\top \mathbf{W} \mathbf{y}$$

$$= \mathbf{L}^{-1} \begin{pmatrix} \frac{1}{\sigma^2} [\mathbf{1}_n \ \mathbf{Z}_0]^\top [\mathbf{1}_n \ \mathbf{Z}_0] & \mathbf{0} \\ \mathbf{0} & \frac{1}{\tau^2} [\mathbf{1}_m \ \mathbf{Z}_1]^\top [\mathbf{1}_m \ \mathbf{Z}_1] \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sigma^2} [\mathbf{1}_n \ \mathbf{Z}_0]^\top & \mathbf{0} \\ \mathbf{0} & \frac{1}{\tau^2} [\mathbf{1}_m \ \mathbf{Z}_1]^\top \end{pmatrix} \mathbf{y}$$

$$= \mathbf{L}^{-1} \begin{pmatrix} ([\mathbf{1}_n \ \mathbf{Z}_0]^\top [\mathbf{1}_n \ \mathbf{Z}_0])^{-1} [\mathbf{1}_n \ \mathbf{Z}_0]^\top \\ ([\mathbf{1}_m \ \mathbf{Z}_1]^\top [\mathbf{1}_m \ \mathbf{Z}_1])^{-1} [\mathbf{1}_m \ \mathbf{Z}_1]^\top \end{pmatrix} \mathbf{y}$$

$$= \mathbf{L}^{-1} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}$$

$$= (\mathbf{Z}^\top \mathbf{Z} \mathbf{L})^{-1} \mathbf{Z}^\top \mathbf{y}$$

$$= (\mathbf{L}^\top \mathbf{Z}^\top \mathbf{Z} \mathbf{L})^{-1} \mathbf{L}^\top \mathbf{Z}^\top \mathbf{y}$$

$$= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \widehat{\boldsymbol{\beta}}^{\mathrm{LS}}.$$

$\square$

The second lemma shows that the constrained estimator of the mean difference $\widehat{\boldsymbol{\Delta}} = \mathbf{A}\widehat{\boldsymbol{\beta}}$ is closer to $\boldsymbol{\Delta}^* = \mathbf{A}\boldsymbol{\beta}^*$ than the unconstrained solution $\widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}} = \mathbf{A}\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}}$ with respect to the norm $\|\cdot\|_{\mathbf{H}^{-1/2}} := \|\mathbf{H}^{-1/2}\cdot\|_2$, with $\mathbf{H} = \mathbf{A}(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{A}^\top$ as defined below (3.11).

**Lemma A.2.** *With probability one, we have* $\|\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*\|_{\mathbf{H}^{-1/2}}^2 \leq \|\widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}} - \boldsymbol{\Delta}^*\|_{\mathbf{H}^{-1/2}}^2$, *with equality holding if and only if* $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}}$.

*Proof.* We have

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{A}^\top \widehat{\boldsymbol{\lambda}}.$$

This and the fact that $\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} \sim N_d(\boldsymbol{\beta}^*, (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1})$ implies that

$$\mathbf{A}\widehat{\boldsymbol{\beta}} = \mathbf{A}\widehat{\boldsymbol{\beta}}^{\mathrm{WLS}} + \mathbf{A}(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}\mathbf{A}^\top \widehat{\boldsymbol{\lambda}} \Rightarrow \widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}} + \mathbf{H}\widehat{\boldsymbol{\lambda}}$$

$$\Rightarrow \widehat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}^* + \boldsymbol{\varsigma} + \mathbf{H}\widehat{\boldsymbol{\lambda}}, \quad \boldsymbol{\varsigma} \sim N_q(\mathbf{0}, \mathbf{H}). \qquad (\mathrm{A.2})$$

It follows that

$$\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^* = \boldsymbol{\varsigma} + \mathbf{H}\widehat{\boldsymbol{\lambda}}$$

$$\Rightarrow \quad \mathbf{H}^{-1/2}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*) = \mathbf{g} + \mathbf{H}^{1/2}\widehat{\boldsymbol{\lambda}}, \quad \mathbf{g} \sim N_q(\mathbf{0}, \mathbf{I}). \qquad (\mathrm{A.3})$$

Now note that the dual optimization problem (3.11) can be rewritten as

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^q} \frac{1}{2}\boldsymbol{\lambda}^\top \mathbf{H}\boldsymbol{\lambda} + \boldsymbol{\lambda}^\top \widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}}.$$

Using the variable transformation $\boldsymbol{\gamma} = \mathbf{H}^{1/2}\boldsymbol{\lambda} \Leftrightarrow \boldsymbol{\lambda} = \mathbf{H}^{-1/2}\boldsymbol{\gamma}$, we obtain the equivalent optimization problem

$$\min_{\mathbf{H}^{-1/2}\boldsymbol{\gamma} \in \mathbb{R}_+^q} \frac{1}{2}\|\boldsymbol{\gamma}\|_2^2 + \boldsymbol{\gamma}^\top \mathbf{H}^{-1/2}\widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}}.$$

Using (A.2) and (A.3), we obtain another equivalent optimization problem

$$\min_{\mathbf{H}^{-1/2}\boldsymbol{\gamma} \in \mathbb{R}_+^q} \left\{ \frac{1}{2}\|\boldsymbol{\gamma}\|_2^2 + \boldsymbol{\gamma}^\top \mathbf{H}^{-1/2}\boldsymbol{\Delta}^* + \boldsymbol{\gamma}^\top \mathbf{g} + \frac{1}{2}\|\mathbf{g}\|_2^2 \right\}$$

$$= \min_{\mathbf{H}^{-1/2}\boldsymbol{\gamma} \in \mathbb{R}_+^q} \left\{ \boldsymbol{\gamma}^\top \mathbf{H}^{-1/2}\boldsymbol{\Delta}^* + \frac{1}{2}\|\mathbf{g} + \boldsymbol{\gamma}\|_2^2 \right\}.$$

Let $\widehat{\boldsymbol{\gamma}}$ denote the corresponding minimizer. Note that $\widehat{\boldsymbol{\gamma}}^\top \mathbf{g} < 0$ unless $\widehat{\boldsymbol{\gamma}} = \mathbf{0}$ (otherwise, $\widehat{\boldsymbol{\gamma}} = \mathbf{0}$ would be the optimal solution, since $\boldsymbol{\Delta}^* > \mathbf{0}$ and hence for any feasible non-zero $\boldsymbol{\gamma}$, we must have $\boldsymbol{\gamma}^\top \mathbf{H}^{-1/2}\boldsymbol{\Delta}^* > 0$). The latter observation implies that $\|\widehat{\boldsymbol{\gamma}} + \mathbf{g}\|_2^2 \leq \|\mathbf{g}\|_2^2$. At

the same time, we note that in view of (A.2) and (A.3)

$$\mathbf{H}^{-1/2}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*) = \mathbf{g} + \widehat{\boldsymbol{\gamma}}, \qquad \mathbf{H}^{-1/2}(\widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}} - \boldsymbol{\Delta}^*) = \mathbf{g}.$$

Since $\|\mathbf{g} + \widehat{\boldsymbol{\gamma}}\|_2^2 \leq \|\mathbf{g}\|_2^2$ as shown above, with equality holding (with probability one) if and only if $\widehat{\boldsymbol{\gamma}} = \mathbf{0} \Leftrightarrow \widehat{\boldsymbol{\lambda}} = \mathbf{0}$ (and thus $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}}$), we conclude that

$$\|\mathbf{H}^{-1/2}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}^*)\|_2^2 \leq \|\mathbf{H}^{-1/2}(\widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}} - \boldsymbol{\Delta}^*)\|_2^2,$$

with equality holding if and only if $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Delta}}^{\mathrm{WLS}}$. $\qquad\square$

The third lemma is used in the Proof of Proposition 3.1.

**Lemma A.3.** *For $s > 0$, consider the quadratic program*

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^d} \frac{1}{2}\boldsymbol{\lambda}^\top \mathbf{Q}\boldsymbol{\lambda} - s \cdot \mathbf{h}^\top \boldsymbol{\lambda}, \tag{A.4}$$

*for a symmetric positive definite matrix $\mathbf{Q}$, and let $\widehat{\boldsymbol{\lambda}}(s)$ be its minimizer. We then have* $\widehat{\boldsymbol{\lambda}}(s) = s\widehat{\boldsymbol{\lambda}}(1)$.

*Proof.* Let $\mathcal{A}(1) = \{1 \leq j \leq d : \widehat{\lambda}_j(1) > 0\}$ be the active set of $\widehat{\boldsymbol{\lambda}}(1)$. In particular, $\widehat{\boldsymbol{\lambda}}_{\mathcal{A}(1)}(1) = (\mathbf{Q}_{\mathcal{A}(1)\mathcal{A}(1)})^{-1}\mathbf{h}_{\mathcal{A}(1)}$, where the subscripts $\mathcal{A}(1)\mathcal{A}(1)$ and $\mathcal{A}(1)$ refer to the principal submatrix and subvector, respectively, corresponding to $\mathcal{A}(1)$. We shall demonstrate that for any $s > 0$, it holds that $\mathcal{A}(s) = \mathcal{A}(1)$. For this purpose, observe that

$$\mathbf{Q}_{\mathcal{A}(1)\mathcal{A}(1)}(\widehat{\boldsymbol{\lambda}}_{\mathcal{A}(1)}(1) \cdot s) = s\mathbf{h}_{\mathcal{A}(1)} \quad \Leftrightarrow \quad \mathbf{Q}_{\mathcal{A}(1)\mathcal{A}(1)}\widehat{\boldsymbol{\lambda}}_{\mathcal{A}(1)}(1) = \mathbf{h}_{\mathcal{A}(1)},$$

$$\mathbf{Q}_{\mathcal{A}(1)^c\mathcal{A}(1)}(\widehat{\boldsymbol{\lambda}}_{\mathcal{A}(1)}(1) \cdot s) \geq s\mathbf{h}_{\mathcal{A}(1)^c} \quad \Leftrightarrow \quad \mathbf{Q}_{\mathcal{A}(1)^c\mathcal{A}(1)}\widehat{\boldsymbol{\lambda}}_{\mathcal{A}(1)}(1) \geq \mathbf{h}_{\mathcal{A}(1)^c},$$

Noting that the left hand sides in the above display constitute the necessary and sufficient

optimality conditions for optimization problem (A.4) (cf. p. 5[75]) we conclude that $\mathcal{A}(s) = \mathcal{A}(1)$ and $\widehat{\boldsymbol{\lambda}}(s) = s\widehat{\boldsymbol{\lambda}}(1)$. □

## A.5    Additional Simulation Studies

In this section, we complement the results presented in §3.5 in the following ways.

**A.5.1** We examine the relative efficiency (RE) concerning estimation of the mean difference $\Delta^*(x)$ and the covariate-specific ROC curve $\mathrm{ROC}_x^*(u)$ in dependence of *various combinations* of (i) the specific value taken by the covariate $x$ and (ii) the FAR $u$.

**A.5.2** We complete the result of our simulation for the mean difference and the covariate-specific ROC curves in §3.5 using Table.

**A.5.3** We visualize the results of the simulation study in §3.5 using multiple figures as an alternative to the Tables A.1 and A.2.

**A.5.4** We shed light on the performance of our method in the presence of *multiple*, possibly highly correlated covariates.

We recall that values of RE larger than one are equivalent to a smaller MSE of the proposed order-constrained method relative to the unconstrained method.

## A.5.1    Relative Efficiency of the Mean Differences and the Covariate-Specific ROC Curves

*RE of mean differences.* In Figure A.1, the RE attains its minimum when $x = 0.5$ for both choices of the sample size ($N = 20$ and $N = 100$) under consideration. For all four studies, the RE for $N = 100$ is always smaller than for $N = 20$; in all cases, the RE exceeds the baseline of one except for study 4 in which the RE is close to one.

*RE of covariate-specific ROC curves.* Figures A.2 to A.5 display the RE for the estimated covariate-specific ROC curves. Figures A.2 and A.3 graph RE versus $x$ for a set of fixed values of the FAR $u$, different sample sizes, and different error distributions. The RE attains its minimum when $x = 0.5$. Moreover, the RE increases with $u$. In Figures A.4 and A.5, the

95

roles of $x$ and $u$ are swapped compared to the two previous figures. It should be emphasized that the RE is above one in all figures except for two cases in study 4 in which the RE is slightly below one.
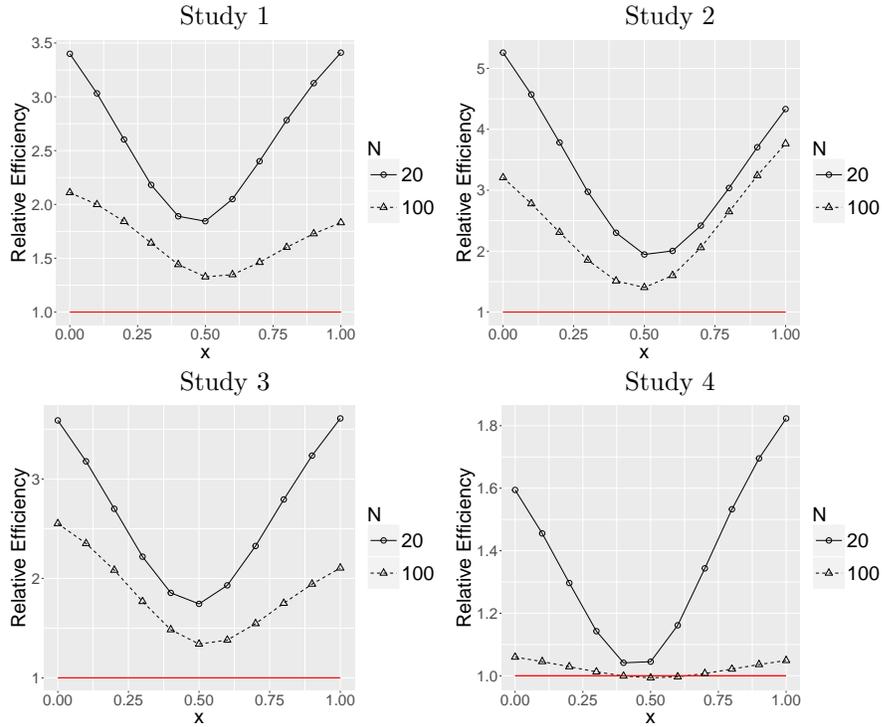


Figure A.1: Relative efficiency (RE) of the estimated mean difference in dependence of $x$. The red horizontal line corresponds to an RE of one.
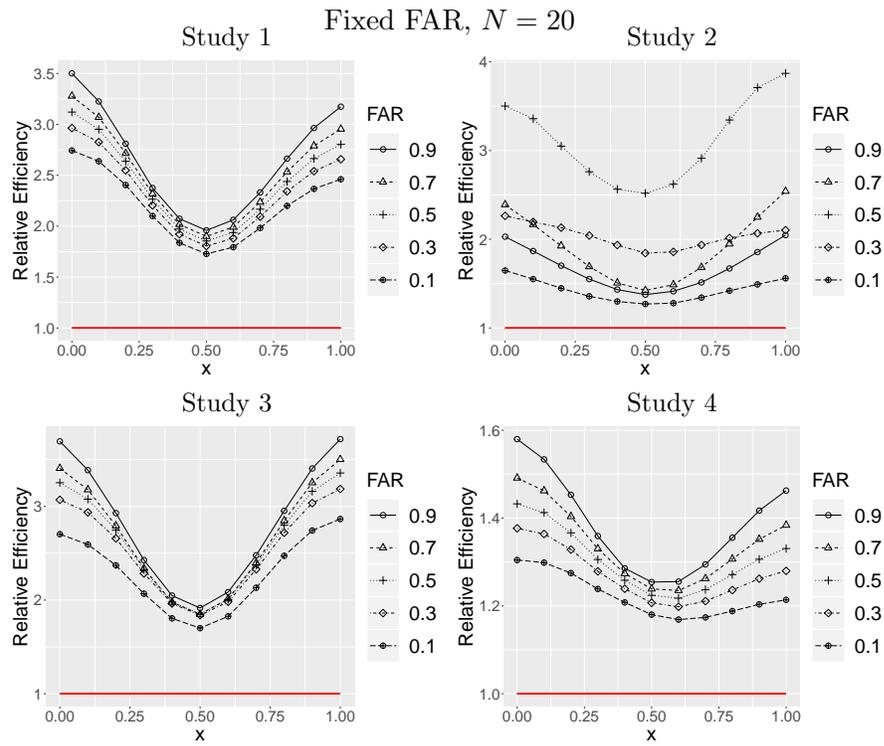
Figure A.2: Relative efficiency (RE) of the estimated ROC in dependence of $x$ for different values of FAR when $N = 20$. The red horizontal line corresponds to an RE of one.
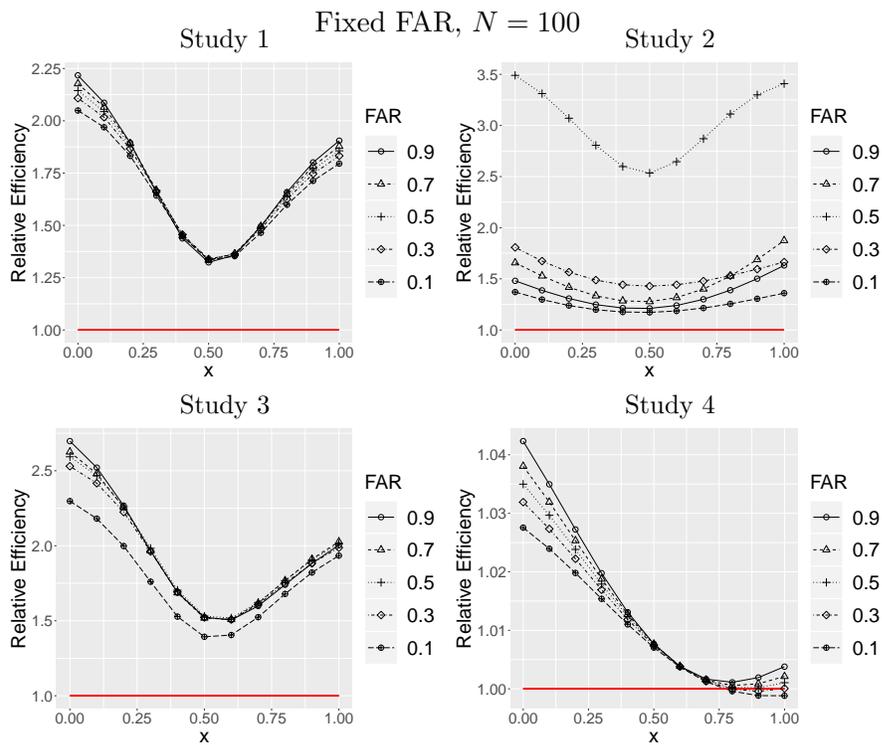
Figure A.3: Relative efficiency (RE) of the estimated ROC in dependence of $x$ for different values of FAR when $N = 100$. The red horizontal line corresponds to an RE of one.
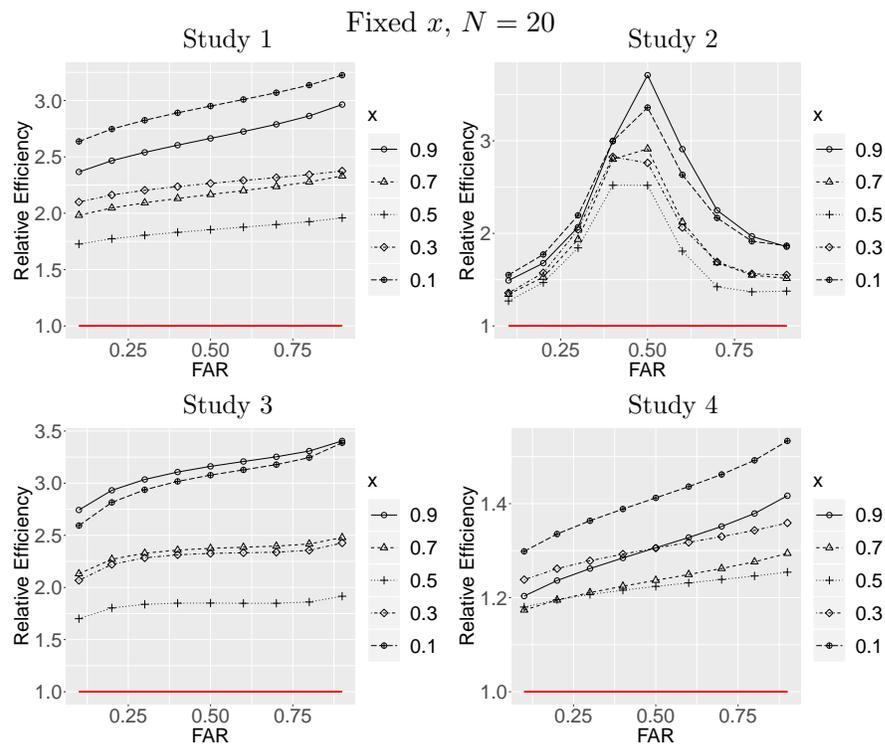
Figure A.4: Relative efficiency (RE) of the estimated ROC in dependence of FAR for different values of $x$ when $N = 20$. The red horizontal line corresponds to an RE of one.
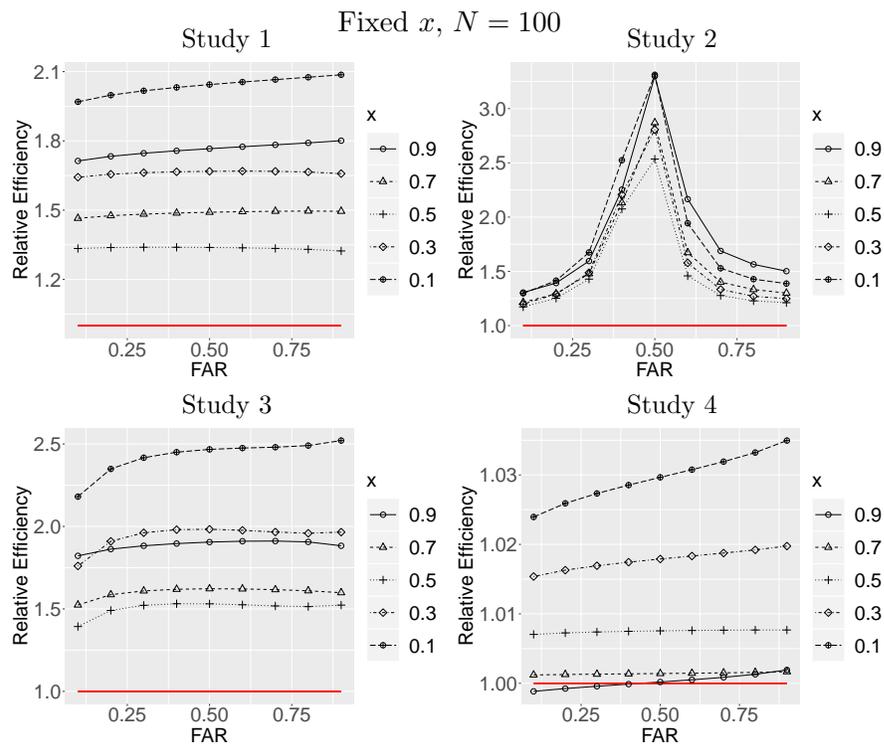
Figure A.5: Relative efficiency (RE) of the estimated ROC in dependence of FAR for different values of $x$ when $N = 100$. The red horizontal line corresponds to an RE of one.

## A.5.2 Bias and MSE of the Mean Differences and the Covariate-Specific ROC Curves

Table A.1 shows the bias and MSE in estimating the mean difference $\Delta^*(x)$ for $x = 0.5$ for all four studies, in dependence of the sample size and different values for $\psi$ and $\phi$. As expected, the bias of the constrained method exceeds that without constraint. On the other hand, the MSE of the former is smaller, i.e., the relative efficiency (RE) is larger than one throughout. As the sample size increases, the RE gets closer to one, which is anticipated in light of Corollary 3.1. Furthermore, it can be seen that small values of $\psi$ and large values of $\phi$ lead to an increase of the RE. The largest REs tend to be attained in Study 1 (normal errors), which is not surprising given the optimality of least squares regression in this case.

| $N$ | $\psi$ | $\sqrt{\phi}$ | Method | Study 1 $B$ | MSE | RE(SE) | Study 2 $B$ | MSE | RE(SE) | Study 3 $B$ | MSE | RE(SE) | Study 4 $B$ | MSE | RE(SE) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | | | | | | | | | | | | | | | |
| | 0.5 | 8 | w/o | 0.13 | 13.70 | 2.52(0.79) | -1.48 | 4759.14 | 2.07(0.49) | -0.33 | 14.61 | 1.83(0.18) | 0.08 | 4.44 | 1.79(0.42) |
| | | | w/ | 1.14 | 5.43 | | 18.43 | 2299.81 | | 1.35 | 8.00 | | 0.54 | 2.49 | |
| | | 16 | w/o | 0.26 | 52.25 | 2.56(0.81) | -2.75 | 18878.61 | 2.07(0.49) | -0.67 | 57.93 | 1.85(0.18) | 0.16 | 17.58 | 1.85(0.44) |
| | | | w/ | 2.59 | 20.20 | | 36.06 | 9106.17 | | 3.06 | 31.38 | | 1.39 | 9.52 | |
| | | 32 | w/o | 0.51 | 216.32 | 2.55(0.80) | -5.30 | 75390.17 | 2.08(0.50) | -1.34 | 231.48 | 1.84(0.18) | 0.33 | 70.11 | 1.85(0.44) |
| | | | w/ | 5.55 | 84.88 | | 71.46 | 36257.42 | | 6.78 | 125.88 | | 3.17 | 37.95 | |
| | 1 | 8 | w/o | 0.13 | 13.70 | 2.45(0.76) | -1.48 | 4759.14 | 2.07(0.49) | -0.33 | 14.61 | 1.80(0.17) | 0.08 | 4.44 | 1.69(0.39) |
| | | | w/ | 0.99 | 5.59 | | 18.22 | 2297.45 | | 1.16 | 8.12 | | 0.41 | 2.63 | |
| | | 16 | w/o | 0.26 | 54.25 | 2.55(0.80) | -2.75 | 18878.61 | 2.07(0.49) | -0.67 | 57.93 | 1.85(0.18) | 0.16 | 17.58 | 1.82(0.43) |
| | | | w/ | 2.41 | 21.25 | | 35.83 | 9101.16 | | 2.85 | 31.32 | | 1.21 | 9.67 | |
| | | 32 | w/o | 0.51 | 216.32 | 2.56(0.81) | -5.30 | 75390.17 | 2.08(0.50) | -1.34 | 231.48 | 1.85(0.18) | 0.33 | 70.11 | 1.85(0.44) |
| | | | w/ | 5.35 | 84.65 | | 71.23 | 36246.83 | | 6.32 | 125.31 | | 2.96 | 37.94 | |
| 100 | | | | | | | | | | | | | | | |
| | 0.5 | 8 | w/o | -0.05 | 1.33 | 1.43(0.10) | 1.01 | 682.21 | 1.43(0.13) | -0.06 | 2.60 | 1.54(0.11) | -0.02 | 0.48 | 1.16(0.08) |
| | | | w/ | 0.26 | 0.93 | | 12.44 | 476.73 | | 0.48 | 1.69 | | 0.09 | 0.42 | |
| | | 16 | w/o | -0.08 | 5.22 | 1.56(0.11) | 2.09 | 2687.60 | 1.42(0.13) | -0.13 | 10.23 | 1.60(0.12) | -0.04 | 1.89 | 1.33(0.10) |
| | | | w/ | 0.79 | 3.34 | | 25.00 | 1887.30 | | 1.27 | 6.38 | | 0.35 | 1.42 | |
| | | 32 | w/o | -0.16 | 20.77 | 1.59(0.12) | 4.24 | 10698.28 | 1.42(0.13) | -0.24 | 40.69 | 1.60(0.12) | -0.08 | 7.54 | 1.43(0.11) |
| | | | w/ | 1.94 | 13.08 | | 50.26 | 7526.15 | | 2.92 | 25.45 | | 1.02 | 5.28 | |
| | 1 | 8 | w/o | -0.05 | 1.33 | 1.29(0.08) | 1.01 | 682.21 | 1.44(0.13) | -0.06 | 2.60 | 1.45(0.10) | -0.02 | 0.48 | 1.07(0.07) |
| | | | w/ | 0.16 | 1.03 | | 12.21 | 475.33 | | 0.36 | 1.79 | | 0.03 | 0.45 | |
| | | 16 | w/o | -0.08 | 5.22 | 1.50(0.11) | 2.09 | 2687.60 | 1.43(0.13) | -0.12 | 10.23 | 1.58(0.11) | -0.04 | 1.89 | 1.23(0.09) |
| | | | w/ | 0.64 | 3.48 | | 24.77 | 1884.01 | | 1.10 | 6.47 | | 0.23 | 1.54 | |
| | | 32 | w/o | -0.16 | 20.77 | 1.58(0.12) | 4.24 | 10698.28 | 1.42(0.13) | -0.24 | 40.69 | 1.61(0.12) | -0.08 | 7.54 | 1.39(0.11) |
| | | | w/ | 1.75 | 13.13 | | 50.03 | 7519.18 | | 2.72 | 25.35 | | 0.84 | 5.42 | |

Table A.1: Bias ($B$) and MSE of the mean difference $\Delta^*(x)$ for $x = 0.5$. (w/o: linear regression without order constraint; w/: linear regression with order constraint.)

We further investigate the bias and the MSE in estimating $\mathrm{ROC}_x^*(u)$. Specifically, we fix $u = 0.5$ and $x = 0.5$. Table A.2 depicts the results in studies 1 to 4, respectively. Small sample sizes yield large values for the RE in alignment with the results in Table A.1. The RE noticeably exceeds one, showing advantages of the proposed use of order constraints.

In Study 4, unconstrained estimation performs on par with constrained estimation.

| | | | | Study 1 | | | Study 2 | | | Study 3 | | | Study 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | $\psi$ | $\sqrt{\phi}$ | Method | $B$ | MSE | RE(SE) | $B$ | MSE | RE(SE) | $B$ | MSE | RE(SE) | $B$ | MSE | RE(SE) |
| 20 | 0.5 | 8 | w/o | 0.78 | 2.82 | 1.73(0.13) | -1.89 | 4.88 | 2.21(0.17) | -1.44 | 2.43 | 1.91(0.15) | -43.01 | 21.15 | 1.39(0.04) |
| | | | w/ | 6.48 | 1.63 | | 6.89 | 2.21 | | 4.96 | 1.27 | | -37.69 | 15.22 | |
| | | 16 | w/o | 0.63 | 2.81 | 1.76(0.14) | -0.87 | 3.64 | 1.96(0.15) | -1.23 | 2.40 | 1.91(0.15) | -47.88 | 25.77 | 1.47(0.04) |
| | | | w/ | 7.18 | 1.60 | | 7.54 | 1.86 | | 5.97 | 1.25 | | -40.77 | 17.48 | |
| | | 32 | w/o | 0.55 | 2.81 | 1.76(0.14) | -0.37 | 2.89 | 1.85(0.14) | -1.33 | 2.39 | 1.89(0.15) | -50.29 | 28.24 | 1.53(0.04) |
| | | | w/ | 7.59 | 1.60 | | 7.59 | 1.57 | | 6.52 | 1.27 | | -42.09 | 18.49 | |
| | 1 | 8 | w/o | 0.91 | 2.77 | 1.68(0.12) | -2.81 | 4.98 | 2.36(0.18) | -1.64 | 2.42 | 1.89(0.14) | -38.21 | 16.97 | 1.31(0.04) |
| | | | w/ | 5.75 | 1.65 | | 5.44 | 2.11 | | 3.96 | 1.28 | | -34.40 | 12.94 | |
| | | 16 | w/o | 0.71 | 2.80 | 1.74(0.13) | -1.26 | 3.67 | 2.03(0.16) | -1.34 | 2.40 | 1.93(0.15) | -45.44 | 23.38 | 1.42(0.04) |
| | | | w/ | 6.76 | 1.61 | | 6.82 | 1.80 | | 5.41 | 1.25 | | -39.37 | 16.38 | |
| | | 32 | w/o | 0.59 | 2.81 | 1.67(0.14) | -0.54 | 2.90 | 1.89(0.14) | -1.19 | 2.39 | 1.90(0.15) | -49.09 | 26.99 | 1.50(0.04) |
| | | | w/ | 7.36 | 1.60 | | 7.23 | 1.54 | | 6.22 | 1.25 | | -41.47 | 18.01 | |
| 100 | 0.5 | 8 | w/o | -0.16 | 0.34 | 1.42(0.10) | -3.84 | 0.71 | 2.66(0.32) | -1.16 | 0.36 | 1.80(0.11) | -42.27 | 18.13 | 1.05(0.01) |
| | | | w/ | 1.38 | 0.24 | | -0.49 | 0.27 | | 0.77 | 0.20 | | -41.29 | 17.25 | |
| | | 16 | w/o | -0.16 | 0.34 | 1.55(0.12) | -1.84 | 0.44 | 2.37(0.20) | -0.69 | 0.34 | 1.82(0.13) | -46.54 | 21.97 | 1.09(0.01) |
| | | | w/ | 2.03 | 0.22 | | 1.37 | 0.19 | | 1.82 | 0.19 | | -44.70 | 20.15 | |
| | | 32 | w/o | -0.17 | 0.34 | 1.58(0.12) | -0.85 | 0.37 | 1.95(0.12) | -0.46 | 0.33 | 1.73(0.12) | -48.69 | 24.04 | 1.12(0.01) |
| | | | w/ | 2.47 | 0.21 | | 2.31 | 0.19 | | 2.41 | 0.19 | | -46.11 | 21.40 | |
| | 1 | 8 | w/o | -0.12 | 0.34 | 1.28(0.09) | -5.80 | 0.92 | 2.55(0.22) | -1.61 | 0.37 | 1.68(0.10) | -38.05 | 14.70 | 1.03(0.01) |
| | | | w/ | 0.90 | 0.27 | | -2.60 | 0.36 | | -0.13 | 0.22 | | -37.60 | 14.33 | |
| | | 16 | w/o | -0.15 | 0.34 | 1.49(0.11) | -2.83 | 0.50 | 2.77(0.23) | -0.92 | 0.34 | 1.84(0.12) | -44.39 | 19.98 | 1.06(0.01) |
| | | | w/ | 1.65 | 0.23 | | 0.30 | 0.18 | | 1.26 | 0.19 | | -43.13 | 18.78 | |
| | | 32 | w/o | -0.16 | 0.34 | 1.58(0.12) | -1.38 | 0.38 | 2.24(0.14) | -0.57 | 0.34 | 1.79(0.13) | -47.61 | 22.98 | 1.10(0.01) |
| | | | w/ | 2.24 | 0.21 | | 1.77 | 0.17 | | 2.10 | 0.19 | | -45.45 | 20.81 | |

Table A.2: Bias ($B$) and MSE of $\text{ROC}_x^*(u)$ for $x = 0.5$ and FAR $= 0.5$. (w/o: linear regression without order constraint; w/: linear regression with order constraint). All values of $B$ and MSE have been multiplied by 100.

### A.5.3   Visualizations of the Results

Figure A.6 and Figure A.7 provide an alternative representation of the results reported in Table A.1 and A.2, respectively. Recall that according to the data-generating model (3.19) underlying the simulations, we denote by $\psi$ the location parameter and by $\phi$ the scale parameter. It can be seen from Figure A.6 that the RE decreases and approaches one as the sample size increases. The REs for the mean difference also increase with small values of $\psi$ and large values of $\phi$. The RE exceeds one in each figure, which shows that our method reduces the MSE.
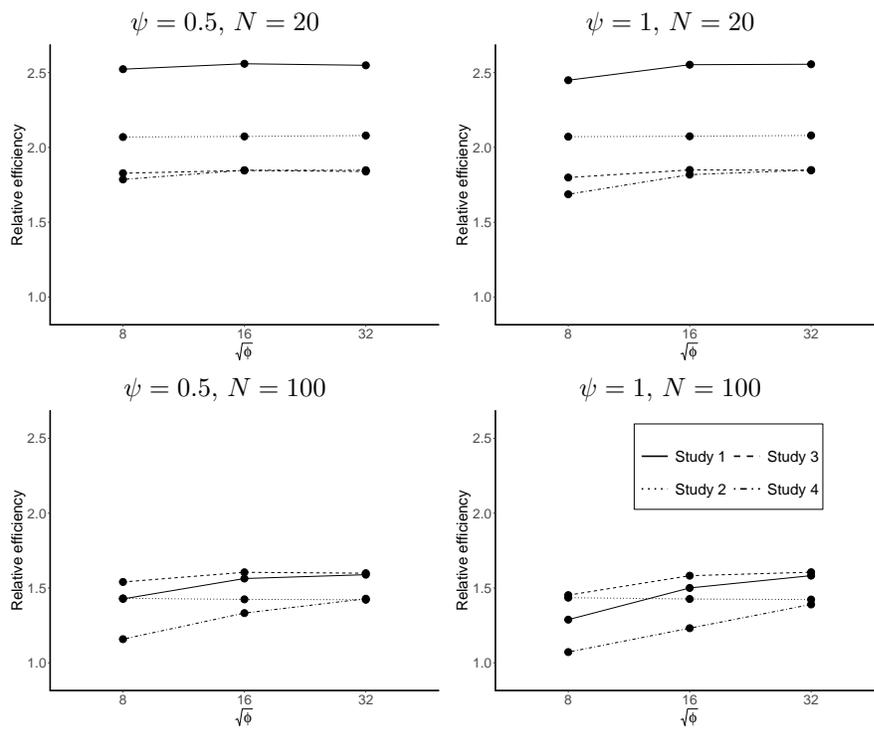
Figure A.6: Relative efficiency (RE) of the estimated mean difference in dependence of $\psi$ for different values of $\phi$ when $x = 0.5$.
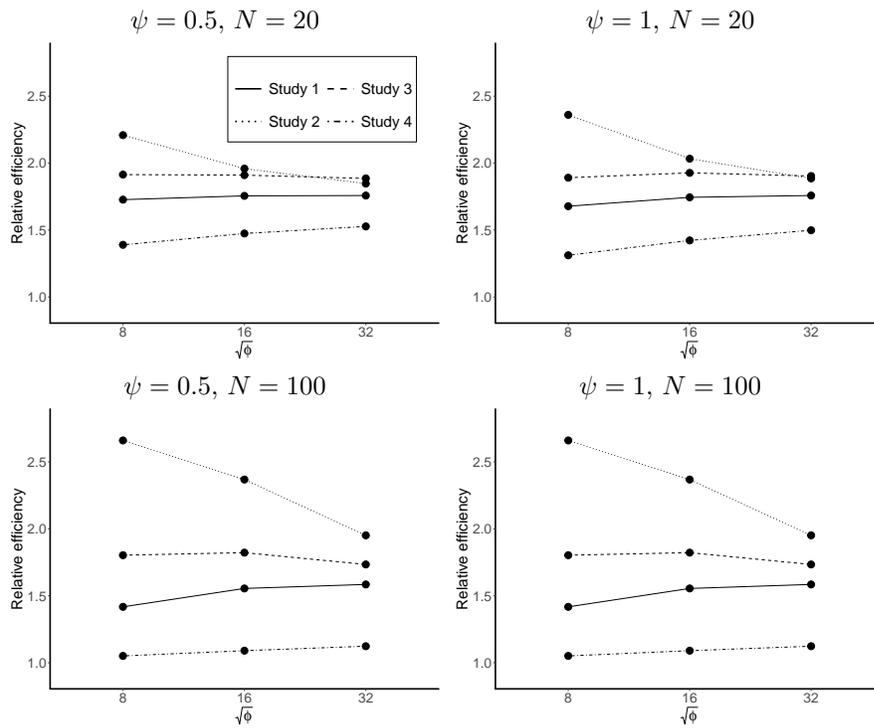
Figure A.7: Relative efficiency (RE) of the estimated ROC in dependence of $\psi$ for different values of $\phi$ when $x = 0.5$ and FAR $= 0.5$.

### A.5.4 Bias and MSE of the Mean Difference and the Covariate-Specific ROC Curves with Multiple Predictors

We here consider the case of multiple $(p > 1)$, possibly highly correlated covariates. Extending the simulation model (3.19), we generate data as follows:

$$T_i = 1 + \mathbf{x}_i^\top \boldsymbol{\theta} + D_i \times \psi + D_i \times \mathbf{x}_i^\top \boldsymbol{\theta} + e_{1i} D_i \times \sqrt{\phi} + e_{0i}(1 - D_i),$$

- $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is such that $x_{ij} = \Phi(g_{ij})$ follows a uniform distribution, $j = 1, \ldots, p$, where

$$\mathbf{g}_i = (g_{i1}, \ldots, g_{ip})^\top, \quad \mathbf{g}_i \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}_\rho), \quad 1 \leq i \leq N.$$

  The covariance matrix $\boldsymbol{\Sigma}_\rho$ has entries one on its diagonal, and $0 \leq \rho < 1$ otherwise.
- The entries of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ are all set to $1/p$.
- All other quantities remain unchanged compared to the simulations for $p = 1$.

The results of the bias and MSE with regard to estimation of the mean difference and the covariate-specific ROC curve for (i) $\mathbf{x} = (0.5, \ldots, 0.5)^\top$ and (ii) $\mathbf{x} = (1, 0, \ldots, 0)^\top$ in dependence of different values of the sample size $N$, the number of covariates $p$, and the correlation coefficient $\rho$.

Table A.3 shows the bias and MSE in estimating the mean difference for $\mathbf{x} = (0.5, \ldots, 0.5)^\top$. The RE increases with the number of predictors and $\rho$ but decreases with the sample size. All REs are larger than the baseline 1, which implies that the proposed method outperforms the traditional method. Table A.4 shows the bias and MSE in estimating the conditional ROC for $\mathbf{x} = (0.5, \ldots, 0.5)^\top$. In Table A.4, with the exception of the two cases in study 1 when the sample size is 50, most of the REs are larger than 1.

Table A.5 and Table A.6 demonstrate the bias and MSE in estimating the mean difference and ROC for $\mathbf{x} = (1, 0, \ldots, 0)^\top$. Comparing with Table A.3 and A.4, it can be seen that the values of both bias and MSE of the proposed method change slightly as $\mathbf{x}$ changes,

where the MSE of the conventional method is changes dramatically when **x** changes. All values of RE are larger than 1 in these two tables.

| N | p | ρ | Method | Study 1 | | | Study 2 | | | Study 3 | | | Study 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | B | MSE | RE(SE) | B | MSE | RE (SE) | B | MSE | RE(SE) | B | MSE | RE(SE) |
| 50 | 1 | - | w/o | -0.03 | 2.73 | 1.41(0.10) | 0.87 | 1441.26 | 1.40(0.19) | 0.13 | 6.09 | 1.48(0.12) | -0.01 | 1.04 | 1.15(0.08) |
| | | | w/ | 0.39 | 1.93 | | 15.37 | 1026.54 | | 0.82 | 4.11 | | 0.13 | 0.90 | |
| | 2 | 0.5 | w/o | 0.01 | 3.20 | 1.36(0.09) | -2.05 | 1422.08 | 1.66(0.19) | -0.09 | 6.15 | 1.47(0.11) | 0.02 | 1.20 | 1.15(0.10) |
| | | | w/ | 0.51 | 2.36 | | 13.93 | 854.38 | | 0.75 | 4.20 | | 0.22 | 1.05 | |
| | | 0.9 | w/o | 0.01 | 3.19 | 1.39(0.09) | -2.03 | 1427.52 | 1.78(0.20) | -0.09 | 6.14 | 1.53(0.11) | 0.02 | 1.19 | 1.18(0.10) |
| | | | w/ | 0.46 | 2.29 | | 13.37 | 802.42 | | 0.68 | 4.02 | | 0.19 | 1.01 | |
| | 5 | 0.5 | w/o | -0.03 | 3.17 | 1.41(0.09) | -1.50 | 1560.73 | 1.57(0.17) | 0.10 | 7.03 | 1.55(0.12) | -0.02 | 1.16 | 1.18(0.09) |
| | | | w/ | 0.60 | 2.25 | | 16.11 | 996.03 | | 0.96 | 4.52 | | 0.27 | 1.00 | |
| | | 0.9 | w/o | -0.03 | 3.17 | 1.53(0.10) | -1.24 | 1550.42 | 1.79(0.20) | 0.10 | 6.99 | 1.72(0.13) | -0.02 | 1.16 | 1.26(0.09) |
| | | | w/ | 0.46 | 2.07 | | 14.59 | 864.54 | | 0.76 | 4.07 | | 0.18 | 0.92 | |
| | 10 | 0.5 | w/o | 0.06 | 5.34 | 2.04(0.14) | -0.01 | 2263.82 | 1.77(0.20) | -0.02 | 9.30 | 1.87(0.16) | 0.04 | 1.92 | 1.82(0.14) |
| | | | w/ | 0.75 | 2.62 | | 19.15 | 1279.04 | | 1.07 | 4.97 | | 0.35 | 1.06 | |
| | | 0.9 | w/o | 0.04 | 5.30 | 2.31(0.15) | -0.41 | 2335.25 | 2.37(0.27) | 0.02 | 9.30 | 2.19(0.20) | 0.03 | 1.91 | 2.01(0.14) |
| | | | w/ | 0.54 | 2.29 | | 16.37 | 985.06 | | 0.77 | 4.26 | | 0.23 | 0.95 | |
| 100 | 1 | - | w/o | -0.05 | 1.33 | 1.29(0.08) | 1.01 | 682.21 | 1.44(0.13) | -0.06 | 2.60 | 1.45(0.10) | -0.02 | 0.48 | 1.07(0.07) |
| | | | w/ | 0.16 | 1.03 | | 12.21 | 475.33 | | 0.36 | 1.79 | | 0.03 | 0.45 | |
| | 2 | 0.5 | w/o | -0.02 | 1.42 | 1.23(0.08) | 1.08 | 762.43 | 1.46(0.13) | 0.07 | 2.70 | 1.23(0.09) | -0.01 | 0.52 | 1.06(0.07) |
| | | | w/ | 0.27 | 1.16 | | 12.99 | 520.79 | | 0.53 | 2.21 | | 0.10 | 0.49 | |
| | | 0.9 | w/o | -0.02 | 1.43 | 1.24(0.08) | 1.10 | 759.93 | 1.54(0.14) | 0.07 | 2.70 | 1.28(0.10) | -0.01 | 0.52 | 1.07(0.07) |
| | | | w/ | 0.23 | 1.15 | | 12.43 | 493.24 | | 0.47 | 2.12 | | 0.09 | 0.48 | |
| | 5 | 0.5 | w/o | 0.07 | 1.43 | 1.19(0.08) | 0.19 | 736.11 | 1.36(0.13) | 0.02 | 2.92 | 1.31(0.09) | 0.03 | 0.51 | 1.06(0.07) |
| | | | w/ | 0.40 | 1.20 | | 13.28 | 540.22 | | 0.59 | 2.22 | | 0.17 | 0.48 | |
| | | 0.9 | w/o | 0.07 | 1.44 | 1.24(0.08) | 0.09 | 743.19 | 1.62(0.16) | 0.02 | 2.91 | 1.43(0.10) | 0.03 | 0.51 | 1.09(0.08) |
| | | | w/ | 0.32 | 1.16 | | 11.77 | 459.01 | | 0.46 | 2.04 | | 0.13 | 0.47 | |
| | 10 | 0.5 | w/o | -0.01 | 1.64 | 1.37(0.09) | -0.02 | 815.33 | 1.46(0.14) | 0.10 | 3.48 | 1.42(0.10) | -0.01 | 0.59 | 1.26(0.09) |
| | | | w/ | 0.42 | 1.20 | | 13.74 | 559.49 | | 0.73 | 2.46 | | 0.18 | 0.47 | |
| | | 0.9 | w/o | -0.02 | 1.62 | 1.50(0.10) | -0.15 | 819.44 | 1.88(0.19) | 0.09 | 3.42 | 1.57(0.11) | -0.01 | 0.59 | 1.34(0.09) |
| | | | w/ | 0.27 | 1.08 | | 11.40 | 436.93 | | 0.52 | 2.17 | | 0.10 | 0.44 | |

Table A.3: Bias ($B$) and MSE of the mean difference $\Delta^*(\mathbf{x})$ for $\mathbf{x} = (0.5,\ldots,0.5)^\top$, where $\mathbf{x} \in \mathbb{R}^p$. (w/o: linear regression without order constraint; w/: linear regression with order constraint.)

|  |  |  |  | Study 1 |  |  | Study 2 |  |  | Study 3 |  |  | Study 4 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | $p$ | $\rho$ | Method | $B$ | MSE | RE(SE) | $B$ | MSE | RE(SE) | $B$ | MSE | RE(SE) | $B$ | MSE | RE(SE) |
| 50 | 1 | - | w/o | -0.01 | 0.71 | 1.43(0.09) | -4.50 | 2.25 | 2.40(0.23) | -0.78 | 0.77 | 1.67(0.11) | -38.12 | 15.05 | 1.08(0.02) |
|  |  |  | w/ | 2.10 | 0.50 |  | 0.89 | 0.94 |  | 1.69 | 0.46 |  | -36.85 | 13.95 |  |
|  | 2 | 0.5 | w/o | -1.00 | 0.62 | 1.18(0.08) | -5.35 | 1.72 | 2.83(0.27) | -2.40 | 0.64 | 1.41(0.09) | -39.82 | 16.32 | 1.18(0.02) |
|  |  |  | w/ | 2.27 | 0.52 |  | -0.54 | 0.61 |  | 1.15 | 0.45 |  | -36.72 | 13.84 |  |
|  |  | 0.9 | w/o | -2.56 | 0.47 | 0.96(0.06) | -5.75 | 1.32 | 2.83(0.25) | -3.46 | 0.51 | 1.18(0.08) | -42.23 | 18.20 | 1.28(0.02) |
|  |  |  | w/ | 1.86 | 0.49 |  | -1.46 | 0.47 |  | 0.75 | 0.43 |  | -37.19 | 14.18 |  |
|  | 5 | 0.5 | w/o | -4.40 | 0.35 | 0.96(0.06) | -6.20 | 0.82 | 2.70(0.19) | -4.46 | 0.37 | 1.09(0.06) | -45.19 | 20.57 | 1.40(0.01) |
|  |  |  | w/ | 1.65 | 0.36 |  | -2.11 | 0.31 |  | 0.95 | 0.34 |  | -37.99 | 14.70 |  |
|  |  | 0.9 | w/o | -5.95 | 0.40 | 1.37(0.07) | -6.59 | 0.57 | 1.92(0.08) | -5.76 | 0.38 | 1.38(0.06) | -47.63 | 22.73 | 1.41(0.01) |
|  |  |  | w/ | 0.09 | 0.29 |  | -3.37 | 0.30 |  | -0.60 | 0.27 |  | -39.76 | 16.07 |  |
|  | 10 | 0.5 | w/o | -6.13 | 0.41 | 1.88(0.09) | -6.73 | 0.55 | 2.05(0.10) | -6.03 | 0.40 | 1.87(0.11) | -48.00 | 23.08 | 1.40(0.01) |
|  |  |  | w/ | 0.01 | 0.22 |  | -3.07 | 0.27 |  | -0.64 | 0.21 |  | -40.32 | 16.45 |  |
|  |  | 0.9 | w/o | -6.89 | 0.48 | 2.36(0.08) | -6.94 | 0.50 | 1.63(0.05) | -6.58 | 0.44 | 2.02(0.06) | -49.15 | 24.16 | 1.26(0.01) |
|  |  |  | w/ | -2.79 | 0.20 |  | -4.48 | 0.31 |  | -3.15 | 0.22 |  | -43.56 | 19.12 |  |
| 100 | 1 | - | w/o | -0.12 | 0.34 | 1.28(0.09) | -5.80 | 0.92 | 2.55(0.22) | -1.61 | 0.37 | 1.68(0.10) | -38.05 | 14.70 | 1.03(0.01) |
|  |  |  | w/ | 0.90 | 0.27 |  | -2.60 | 0.36 |  | -0.13 | 0.22 |  | -37.60 | 14.33 |  |
|  | 2 | 0.5 | w/o | -0.74 | 0.31 | 1.16(0.07) | -5.98 | 0.81 | 2.91(0.17) | -1.80 | 0.31 | 1.38(0.08) | -39.17 | 15.55 | 1.10(0.01) |
|  |  |  | w/ | 1.18 | 0.26 |  | -2.83 | 0.28 |  | 0.26 | 0.22 |  | -37.43 | 14.18 |  |
|  |  | 0.9 | w/o | -1.95 | 0.28 | 1.08(0.06) | -6.16 | 0.70 | 2.61(0.13) | -2.77 | 0.28 | 1.29(0.08) | -41.13 | 17.12 | 1.19(0.01) |
|  |  |  | w/ | 0.91 | 0.25 |  | -3.23 | 0.27 |  | 0.01 | 0.22 |  | -37.72 | 14.40 |  |
|  | 5 | 0.5 | w/o | -3.15 | 0.23 | 1.04(0.07) | -6.31 | 0.60 | 2.40(0.11) | -3.85 | 0.28 | 1.42(0.08) | -43.19 | 18.77 | 1.29(0.01) |
|  |  |  | w/ | 1.22 | 0.22 |  | -3.34 | 0.25 |  | 0.11 | 0.20 |  | -37.88 | 14.49 |  |
|  |  | 0.9 | w/o | -5.17 | 0.31 | 1.57(0.09) | -6.64 | 0.51 | 1.85(0.06) | -5.29 | 0.33 | 1.75(0.08) | -46.38 | 21.55 | 1.41(0.01) |
|  |  |  | w/ | 0.32 | 0.20 |  | -4.18 | 0.27 |  | -0.75 | 0.19 |  | -38.94 | 15.32 |  |
|  | 10 | 0.5 | w/o | -5.47 | 0.33 | 2.45(0.12) | -6.77 | 0.51 | 2.04(0.06) | -5.38 | 0.32 | 2.25(0.10) | -46.78 | 21.91 | 1.38(0.01) |
|  |  |  | w/ | 0.02 | 0.14 |  | -3.94 | 0.25 |  | -0.54 | 0.14 |  | -39.72 | 15.88 |  |
|  |  | 0.9 | w/o | -6.57 | 0.44 | 3.11(0.12) | -6.92 | 0.49 | 1.59(0.03) | -6.26 | 0.40 | 2.55(0.09) | -48.57 | 23.59 | 1.33(0.01) |
|  |  |  | w/ | -1.93 | 0.14 |  | -4.95 | 0.31 |  | -2.34 | 0.16 |  | -41.97 | 17.71 |  |

Table A.4: Bias ($B$) and MSE of $\mathrm{ROC}_{\mathbf{x}}^*(u)$ for $\mathbf{x} = (0.5, \ldots, 0.5)^\top$ where $\mathbf{x} \in \mathbb{R}^p$, and FAR = 0.5. (w/o: linear regression without order constraint; w/: linear regression with order constraint. All values of $B$ and MSE have been multiplied by 100.)

|  |  |  |  | Study 1 |  |  | Study 2 |  |  | Study 3 |  |  | Study 4 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | $p$ | $\rho$ | Method | $B$ | MSE | RE(SE) | $B$ | MSE | RE (SE) | $B$ | MSE | RE(SE) | $B$ | MSE | RE(SE) |
| 50 | 1 | - | w/o | -0.08 | 11.5 | 2.22(0.16) | -0.53 | 5848.5 | 3.19(0.63) | 0.09 | 23.1 | 2.45(0.22) | -0.06 | 4.4 | 1.49(0.12) |
|  |  |  | w/ | 0.09 | 5.2 |  | 14.99 | 1026.5 |  | 0.53 | 9.5 |  | -0.09 | 2.9 |  |
|  | 2 | 0.5 | w/o | 0.66 | 39.8 | 7.55(0.65) | 1.05 | 17610.0 | 12.36(1.75) | -1.51 | 77.7 | 10.07(0.99) | 0.42 | 14.8 | 5.24(0.52) |
|  |  |  | w/ | 0.57 | 5.3 |  | 15.04 | 1425.0 |  | 0.67 | 7.7 |  | 0.29 | 2.8 |  |
|  |  | 0.9 | w/o | 1.42 | 186.1 | 35.77(3.12) | 3.87 | 80450.9 | 61.87(8.99) | -1.00 | 363.3 | 47.74(4.82) | 0.90 | 67.5 | 23.49(2.09) |
|  |  |  | w/ | 0.50 | 5.2 |  | 14.13 | 1300.4 |  | 0.60 | 7.6 |  | 0.24 | 2.9 |  |
|  | 5 | 0.5 | w/o | -0.63 | 75.4 | 25.19(2.29) | -7.62 | 40461.7 | 47.37(8.82) | 0.85 | 160.1 | 34.32(3.42) | -0.32 | 26.3 | 15.18(1.41) |
|  |  |  | w/ | 0.01 | 3.0 |  | 8.96 | 854.2 |  | 0.29 | 4.7 |  | -0.16 | 1.7 |  |
|  |  | 0.9 | w/o | -1.29 | 344.1 | 117.32(10.76) | -11.03 | 174154.7 | 210.68(37.12) | 1.84 | 746.5 | 157.16(16.50) | -0.66 | 120.4 | 73.30(6.52) |
|  |  |  | w/ | 0.13 | 2.9 |  | 9.39 | 826.6 |  | 0.43 | 3.4 |  | -0.06 | 1.6 |  |
|  | 10 | 0.5 | w/o | 0.22 | 132.6 | 55.88(5.39) | -10.25 | 74462.3 | 188.71(39.49) | -0.13 | 265.7 | 77.58(10.24) | 0.15 | 47.6 | 31.85(2.83) |
|  |  |  | w/ | -0.27 | 2.4 |  | 5.01 | 394.6 |  | -0.20 | 3.4 |  | -0.40 | 1.5 |  |
|  |  | 0.9 | w/o | 0.28 | 613.0 | 236.57(22.26) | -15.03 | 343743.7 | 721.07(137.15) | 0.11 | 1297.9 | 330.54(47.27) | 0.21 | 219.3 | 152.08(13.42) |
|  |  |  | w/ | -0.01 | 2.6 |  | 7.04 | 476.7 |  | 0.11 | 3.9 |  | -0.18 | 1.4 |  |
| 100 | 1 | - | w/o | 0.05 | 5.2 | 1.75(0.11) | -0.59 | 2500.4 | 3.25(0.41) | -0.08 | 11.1 | 2.25(0.16) | 0.05 | 1.9 | 1.24(0.09) |
|  |  |  | w/ | 0.03 | 2.96 |  | 11.04 | 768.61 |  | 0.09 | 4.93 |  | -0.01 | 1.54 |  |
|  | 2 | 0.5 | w/o | -0.15 | 18.2 | 6.39(0.48) | -5.16 | 8271.2 | 9.68(1.16) | -0.15 | 37.1 | 7.70(0.67) | -0.07 | 6.5 | 4.02(0.29) |
|  |  |  | w/ | 0.24 | 2.9 |  | 13.56 | 854.8 |  | 0.51 | 4.8 |  | 0.10 | 1.6 |  |
|  |  | 0.9 | w/o | -0.30 | 82.1 | 27.77(2.12) | 9.20 | 36276.9 | 45.27(5.63) | -0.38 | 167.7 | 35.25(3.15) | -0.13 | 29.6 | 16.64(1.18) |
|  |  |  | w/ | 0.19 | 3.0 |  | 12.84 | 801.3 |  | 0.46 | 4.8 |  | 0.08 | 1.8 |  |
|  | 5 | 0.5 | w/o | 0.11 | 31.9 | 14.59(1.12) | -0.01 | 16781.2 | 39.30(5.57) | 0.10 | 65.5 | 21.43(1.80) | 0.11 | 12.0 | 9.30(0.71) |
|  |  |  | w/ | 0.60 | 2.2 |  | 7.16 | 427.0 |  | 0.11 | 3.1 |  | -0.09 | 1.3 |  |
|  |  | 0.9 | w/o | 0.08 | 141.3 | 65.04(4.90) | -0.87 | 74247.6 | 192.56(27.69) | 0.08 | 293.6 | 99.50(8.64) | 0.14 | 53.5 | 42.12(3.16) |
|  |  |  | w/ | 0.15 | 2.2 |  | 7.29 | 385.6 |  | 0.20 | 3.0 |  | 0.03 | 1.3 |  |
|  | 10 | 0.5 | w/o | -0.24 | 41.5 | 24.86(2.03) | -6.04 | 25492.4 | 149.22(26.39) | 0.36 | 84.7 | 35.79(3.99) | -0.12 | 15.4 | 13.99(0.98) |
|  |  |  | w/ | -0.40 | 1.7 |  | 3.22 | 170.8 |  | -0.32 | 2.4 |  | -0.45 | 1.1 |  |
|  |  | 0.9 | w/o | -0.25 | 179.8 | 117.23(8.48) | -10.55 | 107392.8 | 502.03(79.35) | 0.74 | 363.8 | 145.88(15.32) | -0.13 | 66.6 | 66.82(4.81) |
|  |  |  | w/ | -0.19 | 1.5 |  | 4.90 | 213.9 |  | -0.02 | 2.5 |  | -0.21 | 1.0 |  |

Table A.5: Bias ($B$) and MSE of the mean difference $\Delta^*(\mathbf{x})$ for $\mathbf{x} = (1, 0, \ldots, 0)^\top$, where $\mathbf{x} \in \mathbb{R}^p$. (w/o: linear regression without order constraint; w/: linear regression with order constraint.)

| N | p | ρ | Method | Study 1 B | MSE | RE(SE) | Study 2 B | MSE | RE(SE) | Study 3 B | MSE | RE(SE) | Study 4 B | MSE | RE(SE) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | - | w/o | -0.92 | 2.52 | 2.30(0.14) | -6.57 | 5.00 | 3.50(0.25) | -1.86 | 2.62 | 2.52(0.16) | -35.44 | 14.74 | 1.07(0.03) |
| | | | w/ | 0.20 | 1.10 | | -1.78 | 1.43 | | -0.14 | 1.04 | | -35.37 | 13.80 | |
| | 2 | 0.5 | w/o | 0.28 | 3.69 | 3.61(0.22) | -5.71 | 6.09 | 6.50(0.50) | -3.51 | 3.84 | 4.96(0.31) | -38.86 | 18.61 | 1.28(0.04) |
| | | | w/ | 2.30 | 1.02 | | -0.60 | 0.94 | | 0.68 | 0.77 | | -36.70 | 14.54 | |
| | | 0.9 | w/o | -1.15 | 5.83 | 5.94(0.35) | -6.14 | 8.34 | 10.84(0.79) | -4.95 | 6.11 | 8.12(0.50) | -41.15 | 22.65 | 1.51(0.05) |
| | | | w/ | 1.81 | 0.98 | | -1.40 | 0.77 | | 0.30 | 0.75 | | -37.31 | 15.01 | |
| 50 | 5 | 0.5 | w/o | -4.66 | 2.21 | 4.21(0.26) | -5.41 | 3.75 | 10.30(0.68) | -2.73 | 2.19 | 4.70(0.28) | -47.55 | 24.61 | 1.30(0.03) |
| | | | w/ | -0.65 | 0.52 | | -2.84 | 0.36 | | -0.63 | 0.47 | | -42.82 | 18.98 | |
| | | 0.9 | w/o | -5.82 | 2.20 | 5.19(0.31) | -5.55 | 3.58 | 11.68(0.65) | -3.70 | 2.12 | 5.71(0.35) | -49.24 | 26.13 | 1.38(0.03) |
| | | | w/ | -0.80 | 0.42 | | -3.25 | 0.31 | | -0.92 | 0.37 | | -42.90 | 18.95 | |
| | 10 | 0.5 | w/o | -4.38 | 0.86 | 2.64(0.15) | -5.49 | 1.72 | 6.45(0.47) | -4.64 | 0.89 | 3.29(0.19) | -48.45 | 24.15 | 1.12(0.01) |
| | | | w/ | -2.55 | 0.33 | | -3.82 | 0.27 | | -2.89 | 0.27 | | -46.15 | 21.64 | |
| | | 0.9 | w/o | -5.01 | 0.76 | 3.12(0.17) | -5.52 | 1.53 | 6.71(0.48) | -5.09 | 0.81 | 3.95(0.19) | -49.31 | 24.84 | 1.14(0.01) |
| | | | w/ | -2.87 | 0.24 | | -3.87 | 0.23 | | -3.08 | 0.20 | | -46.44 | 21.78 | |
| | 1 | - | w/o | 0.04 | 1.20 | 1.74(0.10) | -7.86 | 2.46 | 2.99(0.20) | -2.30 | 1.39 | 2.15(0.12) | -33.96 | 12.49 | 1.00(0.02) |
| | | | w/ | 0.09 | 0.69 | | -4.92 | 0.82 | | -1.66 | 0.65 | | -34.34 | 12.54 | |
| | 2 | 0.5 | w/o | -1.65 | 2.46 | 4.13(0.24) | -5.41 | 3.61 | 8.20(0.48) | -2.58 | 2.51 | 4.88(0.29) | -40.18 | 18.47 | 1.23(0.03) |
| | | | w/ | 0.88 | 0.59 | | -2.78 | 0.44 | | 0.13 | 0.51 | | -37.81 | 15.02 | |
| | | 0.9 | w/o | -3.26 | 4.71 | 7.83(0.43) | -5.51 | 6.00 | 14.63(0.66) | -3.89 | 4.79 | 9.29(0.51) | -42.59 | 22.69 | 1.48(0.04) |
| | | | w/ | 0.59 | 0.60 | | -3.21 | 0.41 | | -0.12 | 0.52 | | -38.12 | 15.33 | |
| 100 | 5 | 0.5 | w/o | -2.61 | 1.80 | 4.24(0.24) | -5.90 | 2.60 | 9.35(0.52) | -2.98 | 1.84 | 5.39(0.29) | -44.68 | 21.70 | 1.19(0.02) |
| | | | w/ | -0.38 | 0.42 | | -3.66 | 0.28 | | -1.14 | 0.34 | | -41.98 | 18.17 | |
| | | 0.9 | w/o | -4.17 | 2.03 | 5.44(0.31) | -6.26 | 2.75 | 10.89(0.51) | -4.08 | 1.96 | 6.66(0.35) | -47.07 | 24.03 | 1.35(0.03) |
| | | | w/ | -0.17 | 0.37 | | -3.85 | 0.25 | | -1.15 | 0.29 | | -41.68 | 17.86 | |
| | 10 | 0.5 | w/o | -4.39 | 0.79 | 2.74(0.14) | -5.04 | 1.52 | 6.50(0.41) | -3.83 | 0.76 | 3.02(0.16) | -48.04 | 23.69 | 1.11(0.01) |
| | | | w/ | -2.69 | 0.29 | | -4.26 | 0.23 | | -2.82 | 0.25 | | -45.88 | 21.36 | |
| | | 0.9 | w/o | -5.06 | 0.75 | 3.38(0.17) | -5.19 | 1.33 | 6.05(0.40) | -4.42 | 0.70 | 3.54(0.18) | -49.17 | 24.68 | 1.18(0.01) |
| | | | w/ | -2.50 | 0.22 | | -4.17 | 0.22 | | -2.63 | 0.20 | | -45.38 | 20.84 | |

Table A.6: Bias $(B)$ and MSE of $\mathrm{ROC}^*_{\mathbf{x}}(u)$ for $\mathbf{x} = (1, 0, \ldots, 0)^\top$ where $\mathbf{x} \in \mathbb{R}^p$, and FAR = 0.5. (w/o: linear regression without order constraint; w/: linear regression with order constraint. All values of $B$ and MSE have been multiplied by 100.)

# Bibliography

# Bibliography

[1] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan *et al.*, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," *Proceedings of the National Academy of Sciences*, vol. 115, no. 24, pp. 6171–6176, 2018.

[2] C. R. Rao, "Advanced statistical methods in biometric research," *Advanced Statistical Methods in Biometric Research.*, 1952.

[3] B. T. Ulery, R. A. Hicklin, J. Buscaglia, and M. A. Roberts, "Accuracy and reliability of forensic latent fingerprint decisions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 19, pp. 7733–7738, 2011.

[4] B. T. Ulery, R. A. Hicklin, J. Buscaglia, and M. A. Roberts, "Repeatability and reproducibility of decisions by latent fingerprint examiners," *Public Library of Science one*, vol. 7, no. 3, p. e32800, 2012.

[5] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, 2006.

[6] D. Lindley, "A problem in forensic science," *Biometrika*, vol. 64, no. 2, pp. 207–213, 1977.

[7] S. M. Stigler, *The history of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, 1986.

[8] B. Robertson, G. A. Vignaux, and C. E. Berger, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. John Wiley & Sons, 2016.

[9] D. Meuwly and R. Veldhuis, "Forensic biometrics: From two communities to one discipline," in *2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*. IEEE, 2012, pp. 1–12.

[10] J. F. Brookfield, "The effect of relatives on the likelihood ratio associated with dna profile evidence in criminal cases," *Journal of the Forensic Science Society*, vol. 34, no. 3, pp. 193–197, 1994.

[11] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, and A. Bromage-Griffiths, "Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae," *Journal of Forensic Sciences*, vol. 52, no. 1, pp. 54–64, 2007.

[12] T. Ali, L. Spreeuwers, R. Veldhuis, and D. Meuwly, "Effect of calibration data on forensic likelihood ratio from a face recognition system," in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*. IEEE, 2013, pp. 1–8.

[13] G. S. Morrison, "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2387–2397, 2009.

[14] H. Edwards, C. Gotsonis *et al.*, "Strengthening forensic science in the united states: A path forward," *Statement before the United State Senate Committee on the Judiciary*, pp. 1–328, 2009.

[15] J. Holdren, E. Lander, W. Press, M. Savitz, W. Austin, C. Chyba *et al.*, "Report to the president forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods," *Subcommittee on the Social and Behavioral Sciences Team: United States Government*, 2016.

[16] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4249–4252.

[17] C. Neumann, D. E. Armstrong, and T. Wu, "Determination of afis sufficiency in friction ridge examination," *Forensic Science International*, vol. 263, pp. 114–125, 2016.

[18] N. Poh, N. Suki, A. Iorliam, and A. T. Ho, "Anti-forensic resistant likelihood ratio computation: A case study using fingerprint biometrics," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 1377–1381.

[19] X.-H. Zhou, D. K. McClish, and N. A. Obuchowski, *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, 2009.

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

[21] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[22] D. Faraggi and B. Reiser, "Estimation of the area under the ROC curve," *Statistics in Medicine*, vol. 21, no. 20, pp. 3093–3106, 2002.

[23] L. Tang, S. S. Emerson, and X.-H. Zhou, "Nonparametric and semiparametric group sequential methods for comparing accuracy of diagnostic tests," *Biometrics*, vol. 64, pp. 1137–1145, 2008.

[24] L. Tang and A. Liu, "Sample size recalculation in sequential diagnostic trials," *Biostatistics*, vol. 11, no. 1, pp. 151–163, 2009.

[25] T. Dong, L. Tang, and W. F. Rosenberger, "Optimal sampling ratios in comparative diagnostic trials," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 63, no. 3, pp. 499–514, 2014.

[26] L. Tang, L. Kang, C. Liu, E. F. Schisterman, and A. Liu, "An additive selection of markers to improve diagnostic accuracy based on a discriminatory measure," *Academic Radiology*, vol. 20, no. 7, pp. 854–862, 2013.

[27] J. Hendricks, C. Neumann, and C. Saunders, "A ROC-based approximate bayesian computation algorithm for model selection: application to fingerprint comparisons in forensic science," *arXiv Dated: 2019*, 2019.

[28] S. N. Srihari and H. Srinivasan, "Comparison of ROC and likelihood decision methods in automatic fingerprint verification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 22, no. 03, pp. 535–553, 2008.

[29] M. S. Pepe, "A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing," *Biometrika*, vol. 84, no. 3, pp. 595–608, 1997.

[30] V. I. de Carvalho, A. Jara, T. E. Hanson, and M. de Carvalho, "Bayesian nonparametric ROC regression modeling," *Bayesian Analysis*, vol. 8, no. 3, pp. 623–646, 2013.

[31] A. Rodríguez and J. C. Martínez, "Bayesian semiparametric estimation of covariate-dependent ROC curves," *Biostatistics*, vol. 15, no. 2, pp. 353–369, 2014.

[32] M. S. Pepe, "Three approaches to regression analysis of receiver operating characteristic curves for continuous test results," *Biometrics*, vol. 54, no. 1, pp. 124–135, 1998.

[33] T. A. Alonzo and M. S. Pepe, "Distribution-free ROC analysis using binary regression techniques," *Biostatistics*, vol. 3, no. 3, pp. 421–432, 2002.

[34] W. González-Manteiga, J. C. Pardo-Fernández, and I. v. Keilegom, "ROC curves in non-parametric location-scale regression models," *Scandinavian Journal of Statistics*, vol. 38, no. 1, pp. 169–184, 2011.

[35] X. Duan and X.-H. Zhou, "Composite quantile regression for the receiver operating characteristic curve," *Biometrika*, vol. 100, no. 4, pp. 889–900, 2013.

[36] L. Tang and X.-H. Zhou, "Semiparametric inferential procedures for comparing multivariate ROC curves with interaction terms," *Statistica Sinica*, pp. 1203–1221, 2009.

[37] H. Zou and M. Yuan, "Composite quantile regression and the oracle model selection theory," *The Annals of Statistics*, vol. 36, no. 3, pp. 1108–1126, 2008.

[38] S. Yoon and A. K. Jain, "Longitudinal study of fingerprint recognition," *Proceedings of the National Academy of Sciences*, vol. 112, no. 28, pp. 8555–8560, 2015.

[39] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "The good, the bad, and the ugly face challenge problem," *Image and Vision Computing*, vol. 30, no. 3, pp. 177–185, 2012.

[40] P. J. Phillips, P. J. Flynn, and K. W. Bowyer, "Lessons from collecting a million biometric samples," *Image and Vision Computing*, vol. 58, pp. 96–107, 2017.

[41] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." in *British Machine Vision Conference*, vol. 1, 2015, p. 6.

[42] E. L. Lehmann, "Ordered families of distributions," *The Annals of Mathematical Statistics*, vol. 26, no. 3, pp. 399–419, 1955.

[43] A. R. Jonckheere, "A distribution-free k-sample test against ordered alternatives," *Biometrika*, vol. 41, no. 1/2, pp. 133–145, 1954.

[44] P. J. Boland and F. Proschan, "Stochastic order in system reliability theory," *Stochastic Orders and Their Applications*, pp. 485–508, 1994.

[45] M. Kijima and M. Ohnishi, "Stochastic orders and their applications in financial optimization," *Mathematical Methods of Operations Research*, vol. 50, no. 2, pp. 351–372, 1999.

[46] S. D. Peddada, G. E. Dinse, and G. E. Kissling, "Incorporating historical control data when comparing tumor incidence rates," *Journal of the American Statistical Association*, vol. 102, no. 480, pp. 1212–1220, 2007.

[47] Y. Park, J. M. Taylor, and J. D. Kalbfleisch, "Pointwise nonparametric maximum likelihood estimator of stochastically ordered survivor functions," *Biometrika*, vol. 99, no. 2, pp. 327–343, 2012.

[48] C.-F. Tang, D. Wang, and J. M. Tebbs, "Nonparametric goodness-of-fit tests for uniform stochastic ordering," *Annals of Statistics*, vol. 45, no. 6, pp. 2565–2589, 2017.

[49] O. Davidov and A. Herman, "Ordinal dominance curve based inference for stochastically ordered distributions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 5, pp. 825–847, 2012.

[50] B. Chen, P. Li, J. Qin, and T. Yu, "Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 861–874, 2016.

[51] T. Yu, P. Li, and J. Qin, "Density estimation in the two-sample problem with likelihood ratio ordering," *Biometrika*, vol. 104, no. 1, pp. 141–152, 2017.

[52] T. Westling, K. J. Downes, and D. S. Small, "Nonparametric maximum likelihood estimation under a likelihood ratio order," *arXiv preprint arXiv:1904.12321*, 2019.

[53] T. Ali, L. Spreeuwers, R. Veldhuis, and D. Meuwly, "Sampling variability in forensic likelihood-ratio computation: A simulation study," *Science & justice*, vol. 55, no. 6, pp. 499–508, 2015.

[54] R. Royall, *Statistical Evidence: a Likelihood Paradigm.* CRC press, 1997, vol. 71.

[55] L. J. Davis, C. P. Saunders, A. Hepler, and J. Buscaglia, "Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios," *Forensic Science International*, vol. 216, no. 1-3, pp. 146–157, 2012.

[56] N. Suki, N. Poh, F. M. Senan, N. A. Zamani, and M. Z. A. Darus, "On the reproducibility and repeatability of likelihood ratio in forensics: A case study using face biometrics," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2016, pp. 1–8.

[57] R. Azencott and D. Dacunha-Castelle, *Series of Irregular Observations: Forecasting and Model Building*. Springer Science & Business Media, 2012, vol. 2.

[58] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[59] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. CRC press, 1986, vol. 26.

[60] P. Rose, "More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends," *International Journal of Speech Language and the Law*, vol. 20, no. 1, pp. 77–116, 2013.

[61] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio," *Australian Journal of Forensic Sciences*, vol. 45, no. 2, pp. 173–197, 2013.

[62] E. Enzinger, G. S. Morrison, and F. Ochoa, "A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case," *Science & Justice*, vol. 56, no. 1, pp. 42–57, 2016.

[63] A. Agresti and M. Kateri, *Categorical Data Analysis*. Springer, 2011.

[64] R. L. Prentice and R. Pyke, "Logistic disease incidence models and case-control studies," *Biometrika*, vol. 66, no. 3, pp. 403–411, 1979.

[65] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[66] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, no. 3, pp. 863–879, 2011.

[67] B. C. Choi, "Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test," *American Journal of Epidemiology*, vol. 148, no. 11, pp. 1127–1132, 1998.

[68] K. H. Zou, C. M. Tempany, J. R. Fielding, and S. G. Silverman, "Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones," *Academic Radiology*, vol. 5, no. 10, pp. 680–687, 1998.

[69] P. J. Phillips, "A cross benchmark assessment of a deep convolutional neural network for face recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2017.

[70] A. J. O'Malley, K. H. Zou, J. R. Fielding, and C. M. Tempany, "Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: prostate biopsy and spiral CT of ureteral stones," *Academic Radiology*, vol. 8, no. 8, pp. 713–725, 2001.

[71] M. A. Arcones, P. H. Kvam, and F. J. Samaniego, "Nonparametric estimation of a distribution subject to a stochastic precedence constraint," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 170–182, 2002.

[72] S. Boyd and L. Vandenberghe, *Convex Optimization*.   Cambridge University Press, 2004.

[73] C. K. Liew, "Inequality constrained least-squares estimation," *Journal of the American Statistical Association*, vol. 71, no. 355, pp. 746–751, 1976.

[74] D. Kim, S. Sra, and I. S. Dhillon, "Tackling box-constrained optimization via a new projected quasi-Newton approach," *SIAM Journal on Scientific Computing*, vol. 32, no. 6, pp. 3548–3563, 2010.

[75] D. Chen and R. J. Plemmons, "Nonnegativity constraints in numerical analysis," in *The Birth of Numerical Analysis*.   World Scientific, 2010, pp. 109–139.

[76] P. J. Phillips, K. W. Boyer, P. J. Flynn, A. J. O'Toole, C. L. Schott, W. T. Scruggs, and M. Sharpe, "Face recognition vendor test (FRVT) 2006 and iris challenge evaluation (ICE) 2006 large-scale results," *NIST Interagency/Internal Report*, 2007.

[77] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop, "Demographic effects on estimates of automatic face recognition performance," *Image and Vision Computing*, vol. 30, no. 3, pp. 169–176, 2012.

[78] H. Zhang, J. R. Beveridge, B. A. Draper, and P. J. Phillips, "On the effectiveness of soft biometrics for increasing face verification rates," *Computer Vision and Image Understanding*, vol. 137, pp. 50–62, 2015.

[79] G. H. Givens, J. R. Beveridge, P. J. Phillips, B. Draper, Y. M. Lui, and D. Bolme, "Introduction to face recognition and evaluation of algorithm performance," *Computational Statistics & Data Analysis*, vol. 67, pp. 236–247, 2013.

[80] X. He, "Quantile curves without crossing," *The American Statistician*, vol. 51, no. 2, pp. 186–192, 1997.

[81] S. A. Cole, "History of fingerprint pattern recognition," in *Automatic Fingerprint Recognition Systems*.   Springer, 2004, pp. 1–25.

[82] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi, "Gait-based person-verification system for forensics," in *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*.   IEEE, 2012, pp. 113–120.

[83] B. S. Hwang and Z. Chen, "An integrated bayesian nonparametric approach for stochastic and variability orders in roc curve estimation: an application to endometriosis diagnosis," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 923–934, 2015.

[84] W. Zhang, L. L. Tang, Q. Li, A. Liu, and M.-L. T. Lee, "Order-restricted inference for clustered roc data with application to fingerprint matching accuracy," *Biometrics*, 2019.

[85] R. Koenker, *Quantile Regression.* Cambridge University Press, 2005.

[86] M. X. Rodríguez-Álvarez, J. Roca-Pardiñas, and C. Cadarso-Suárez, "ROC curve and covariates: extending induced methodology to the non-parametric framework," *Statistics and Computing*, vol. 21, no. 4, pp. 483–499, 2011.

[87] O. Davidov and S. Peddada, "Order-restricted inference for multivariate binary data with application to toxicology," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1394–1404, 2011.

[88] A. Jain, B. Klare, and A. Ross, "Guidelines for best practices in biometrics research," in *Biometrics (ICB), 2015 International Conference on.* IEEE, 2015, pp. 541–545.

[89] A. K. Jain and A. Ross, "Bridging the gap: from biometrics to forensics," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1674, p. 20140254, 2015.

# Curriculum Vitae

Xiaochen Zhu graduated from Beijing University of Chemical Technology in 2014 with the Bachelor's degree in bioengineering and the Master's degree in chemical engineering.

After his graduation, he decided to join George Mason University for his Master's degree in biostatistics in 2015 and his Doctoral degree in statistical science in 2017. During his graduate education, he worked for Volgenau School of Engineering, Department of Health Administration and Policy, and Center for Applied Proteomics and Molecular Medicine at George Mason University as a research assistant. He also worked at Inova Fairfax Hospital Medical Campus and Food and Drug Administration as an internship during his summertime.

He was awarded the title of Distinguished Academic Achievement M.S. Programs, Department of Statistics in 2017, and the Washington Statistical Society Outstanding Graduate Student Award in 2020. He was also awarded the 1st Place of 2020 American Statistical Association Medical Devices and Diagnostic Student Paper Award for his paper "Order-Constrained ROC Regression with Application to Facial Recognition" in 2020.

From 2017 to 2020, he worked with Dr. Martin P. Slawski and Dr. Liansheng L. Tang in the Ph.D. program of statistic department at George Mason University.