National Institute of Justice

Research and Development in Forensic Science
for Criminal Justice Purposes

Award #: 2019-DU-BX-4011

Accounting for Covariates in Forensic Error Rate

Final Research Report

Award Amount: $495,056
Project Period: 9/21/2020 – 4/30/2023

Principal Investigator:

Dr. Liansheng (Larry) Tang
Professor
University of Central Florida
Phone: (407) 823-0638
Email: liansheng.tang@ucf.edu

Recipient Organization:

University of Central Florida
12201 Research Parkway, Suite 501
Orlando, FL 32826

Signature of Submitting Official: Shannon M. Callahan, AOR

Digitally signed by Shannon M. Callahan, AOR
Date: 2023.07.28 16:05:50 -04'00'

Table of Contents

1 Summary of the project

1.1 Major Goals and Objectives

In this research project, we aimed to develop error rate interpretation tools using regression on decision scores and using receiver operating characteristic (ROC) regression models.

In forensics, quantification of error rates in pattern and trace evidence interpretation has been a concern raised by the congressionally mandated 2009 National Academy of Science (NAS) report *Strengthening Forensic Science in the United States: A Path Forward*, and more recently, the 2016 President's Council of Advisors on Science and Technology (PCAST) report *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. The 2009 NAS report highlights the need for developing quantifiable measures of uncertainty in forensic analyses. This report discusses the determination of error rates in forensic evidence, and the 2016 PCAST report discusses the validity and reliability of seven feature-based methods. Both reports emphasize the importance of measuring accuracy and performance, and transforming subjective feature-comparison methods into objective methods. The errors in evidence interpretation are mainly errors related to individualization and exclusion decisions. Since the presentation by forensic scientists in courts has influence over decisions made by the judge and the jury, as consequences of these errors, an innocent person could be wrongly accused and a criminal could be mistakenly claimed innocent.

In response to these recommendations, the black box and white box studies in latent print examination and face recognition have been successfully conducted to systematically quantify errors in these forensic disciplines. The latent print study by [36] mainly studied examiners' binary decisions of either individualization or exclusion, and presented a systematic study on the error rates including false positive rate (FPR) that is the probability of incorrect individualization on imposter pairs and false negative rate (FNR) that is the probability of incorrect exclusion from the same source. Besides binary decisions, [32] asked examiners to provide an ordinal-scale decision score for a pair of images based on their belief on whether

the images come from the same source. A score obtained by comparing two images of the same source is often referred to as a genuine score, and that obtained by comparing two different sources as an imposter score. For the ordinal decision, error rates such as the FPR and the FNR are calculated for all possible thresholds. For a specific threshold point, the FPR is the percentage of imposter scores greater than the threshold in the non-genuine pairs, and the FNR is the percentage of genuine scores less than or equal to the threshold in the genuine pairs. As the threshold increases, the FPR decreases while the FNR increases. The pairs of 1-FPR and FNR are plotted as the receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) is commonly used to summarize the ROC curve, and the discussion of its application in diagnostic medicine and medical imaging can be found in textbooks [45, 28] The AUC was used by [32] to assess the examiners error rates. The error rates from these latent print and face recognition studies provide a rigorous way to validate practices in their respective disciplines.

These error rates and the ROC curve are obtained by pooling all the decisions from examiners or computer algorithms with same-source or different-source pairs. These measures report the average error rates across a population of examiners for evidence sources. Ideally, the error rates tend to provide guidance for the error rates for interpreting a given evidence source if error rates are consistent for sources with various aspects and for examiners with various background. However, error rates may not be consistent as [36] found that "... the false positive errors involved latents on the most complex combination of processing and substrate included in the study. ... Further research is necessary to identify the attributes of prints associated with false positive or false negative errors...". A more recent study on an operational fingerprint database by [39] also found that fingerprint decision scores vary with covariates such as subjects' demographic information (age and gender, etc). In other biometrics disciplines such as face recognition, if examining facial images from male subjects do not yield the same error rates as examining images from female subjects, it is unclear whether the pooled error rates lead to the the same error rates in subgroups of source subjects

with different demographics. Besides these covariates for the source subjects, the other set of covariates related to forensic examiners may also play an important role in the error rates. Fro example, forensic examiners with different training and demographics may not result in the same error rates. Hypothetically, more experienced examiners may tend to have higher confidence and lower individualization error rates than newly recruited examiners. It would be ideal to account for both sets of covariates such as 1) source subjects' covariate information including their demographics and/or source images' attributes and quality, and 2) examiners' covariate information such as their training background and demographics. The methods to be developed in the project will provide general statistical tools to incorporate these two sets of covariates in the interpretation of error rates so that the error rates can be personalized to source subjects with specific demographic information and examiners with specific training background.

Appropriately accounting for covariates in error rate assessment and evidence interpretation requires sophisticated statistical analyses with modern statistical concepts and methods. The recent 2018 NIJ Forensic Science Technology Working Group (TWG) Operation Requirements report specifically calls for research and development in "Determination of accuracy and reliability of forensic analyses and conclusions, including potential sources of error", and "Practical statistical approaches for the interpretation of forensic evidence" in impression, pattern and trace evidence. In this research program we will work within the ROC regression framework for error rate quantification by allowing covariates specific to source subjects and examiners. We will study statistical techniques by 1) fitting regression models in order to relate covariates to decision scores, and 2) by fitting ROC regression in order to relate covariates to error rates quantified by the ROC curve. The resulting covariate-specific ROC curves in face recognition, handwriting, and latent print databases will model the relationship between covariates and decision scores, give the error rates for specific values of covariates. The resulting covariate-adjusted ROC curve will provide error rates by accounting for covariates. These ROC curves will be compared with the pooled

ROC curves studied in the forensic literature. We will then relate these ROC methods to LR in terms of trace and pattern evidence interpretation by accounting for covariates.

1.2 Research questions

To achieve our main objective, we set out to explore three main research questions:

1. Develop error rate interpretation tools using regression on decision scores;
2. Develop error rate interpretation tools using ROC regression models;
3. Develop evidence interpretation tools based on covariate-specific likelihood ratios.

1.3 Research design, methods, analytical and data analysis techniques

1.3.1 Research Question 1: Develop error rate interpretation tools using regression on decision scores

The ROC curve, which is commonly used in biometric system evaluation studies, and more recently, forensic error rate studies, is a plot of the true positive rate (TPR) (i.e., probability of identifying a case when the subject is truly diseased) versus false positive rate (FPR) (i.e., probability of identifying a case when the subject is not diseased) in dependency of the decision threshold. The ROC curve is widely used in radiology, psychophysical and medical imaging research for detection performance, military monitoring, and industrial quality control [20]. The ROC curve graphs the trade-off between the TPR and FPR under different thresholds. It has a variety of appealing properties, and overcomes the limitations associated with isolated measurements of TPR and FPR. The ROC curve displays the (FPR, TPR)-pairs corresponding to all possible decision thresholds [45, 28].

The accuracy of biometric systems or humans in source identification problems can be assessed with the ROC curve when the decision scores are ordinal or continuous. The source identification problems aim to determine the link between the known source (suspect) and an unknown source (evidence from the crime scene). The identification relates to a specified source population. In this project, sources are defined as generators of the objects of interest (i.e., a person generates of face and handwriting profiles, and a ignitable liquid is

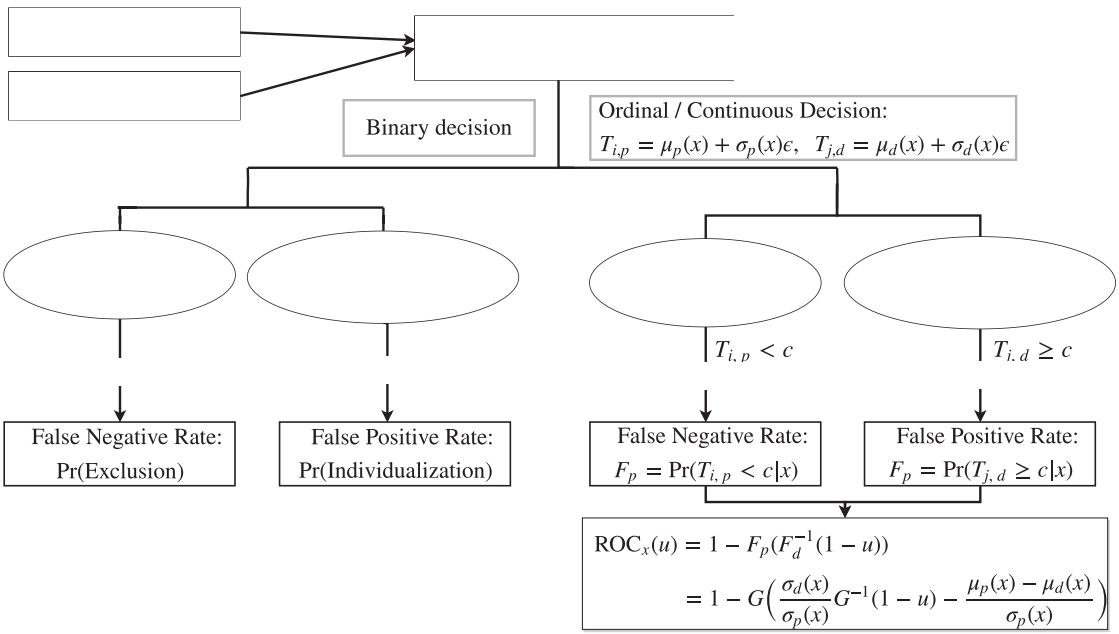

Figure 1: Schematic overview on notation used for decision scores, error rates, and ROC curve.

a generator of fire debris). Thus, all the evidential objects considered in a given case are divided into three subsets as follows: e_s – a set of objects associated with or generated by a specified source; e_a – collection of sets of objects each associated with a source of traces in an alternative source population; and e_u – a set of trace objects that are all from the same unknown source. The evidence from the specified source e_s also includes covariates of the source such as the source subject's demographic information and other source covariates. The error rates are related to the decision on the following two propositions for how the evidence has arisen:

H_p**:** The unknown source evidence e_u and the specific source evidence e_s both originate from the specific source;

H_d**:** The unknown source evidence e_u does not originate from the specific source, but from some other source in the alternative source population.

For each decision score, black-box studies often have a set of multiple covariates capturing various attributes, $X = (X_s, \ldots, X_d)$, such as 1) demographic properties of the underlying subjects (e.g., gender, age, race, ...), 2) measurement-specific contextual information (e.g.,

image quality), and 3) information regarding examiners' background in terms of training and experience. The covariate-specific ROC conditional on $X = x$, i.e., observing a certain combination of values for the covariates is defined by $ROC_x(u) = 1 - F_p(F_d^{-1}(1-u)|x)$, where F_p and F_d denote the distribution functions of genuine and imposter scores conditional on $X = x$. The diagram in Figure 1 depicts the relationships.

In case of a single categorical covariate such as gender, age group, or image quality ('good', 'bad', 'ugly'), the covariate-specific ROC curve boils down to stratum-specific ROC curves, where the strata are defined by the category levels of the covariate. Examples are provided in Figure 2 which contrasts stratum-specific empirical ROC curves to their pooled counterparts without distinction between strata. The figure shows that ROC curves can differ markedly across strata. Subgroup analysis as displayed in Figure 2 is essentially limited to the case of

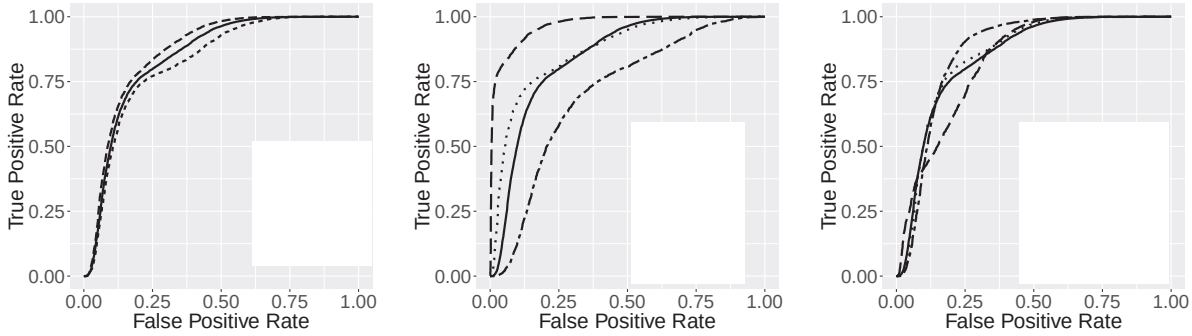

Figure 2: ROC curves for the GBU data. Left panel: Gender-specific ROC; Middle panel: quality-specific ROC; Right panel: age-specific ROC (in terms of year of birth).

a single covariate with a small number of category levels. Partitioning continuous covariates such as age into age categories is problematic since subsequent results depend on the choice of the number and range of age subgroups. Similarly, when considering several categorical covariates simultaneously, the number of subgroups grows exponentially with the number of covariates d. As a result, subgroup is analysis is no longer a suitable approach.

In addition to having potentially different ROC curves with and without covariates, pooled error rates or the pooled ROC curve may also tend to overestimate the error rates or have a lower ROC curve. Such an example can be found in Figure 3. When covariate X changes,

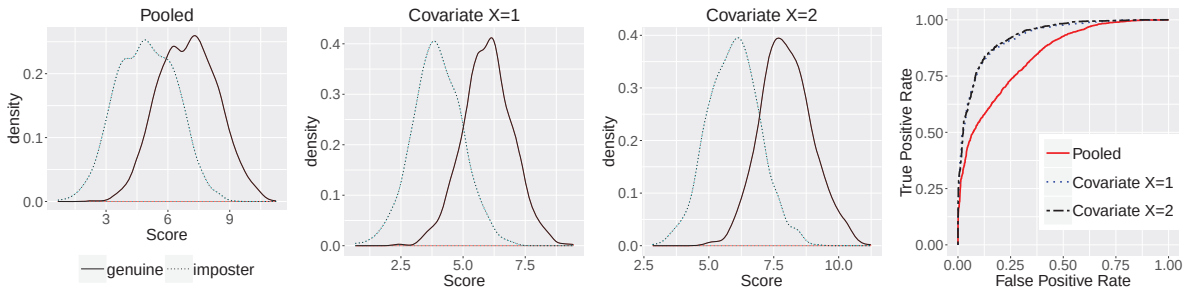

Figure 3: Covariate-specific ROC vs. pooled ROC. First three panels (from the left): distribution of pooled scores and unpooled scores; rightmost panel: corresponding ROC curves.

the covariate-specific ROC curves remain the same. Interestingly, the ROC curve by pooling all the data together is below both covariate-specific ROC curves.

It is clear from these figures that ignoring important covariates in the error rate assessment may lead to somehow different error rate or ROC interpretation. Ideally, error rates or ROC curves should be provided with and without covariates. Regression techniques for the ROC curve will be studied in the project.

1.3.2 Research Question 2: Develop error rate interpretation tools using ROC regression models

Regression modeling can often be enhanced by incorporating additional structural properties such as bounds on the regression coefficients, smoothness, or shape constraints. Taking advantage of such constraints can often lead to substantial reductions in estimation variance, particularly with small samples. In the above setting, it is often appropriate to assume that $T_{i,p}$ is stochastically larger than $T_{j,d}$, which is known as *stochastic ordering* [9, 27, 19]. This ensures that the resulting ROC curve is always above the diagonal line. The latter constraint is implied by a monotonically increasing likelihood ratio [11, 6, 2]. A sufficient condition to ensure stochastic ordering in the location-scale model is $\mu_p(x) \geq \mu_d(x)$ and $\sigma_p(x) = \sigma_d(x)$ for all x. If the location and scale functions are assumed to be linear in a set of unknown parameters, these constraints reduce to linear inequality constraints for the parameters, and hence are computationally tractable. To give an example, consider a single covariate x with range $[a, b]$ for numbers $a < b$. The constraint $\mu_p(x) \geq \mu_d(x)$ is then equivalent to

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

$(\alpha_p - \alpha_d) + (\beta_p - \beta_d)z \geq 0$ for all $z \in [a, b]$, which by convexity is equivalent to the two linear inequality constraints $(\alpha_p - \alpha_d) + a(\beta_p - \beta_d) \geq 0, \qquad (\alpha_p - \alpha_d) + b(\beta_p - \beta_d) \geq 0.$ which can be integrated seamlessly into parameter estimation for improved statistical efficiency. The rationale directly extends to the case of multiple bounded covariates.

Order constraints

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 larger scores for genuine pairs than for imposter pairs. Similarly, in biomarker studies, the level of a biomarker indicating the presence of a disease is supposed to be larger among diseased than healthy patients. This ordering property can be integrated into the location-scale model by requiring that that the locations of the two score distributions associated with the status variable D are ordered accordingly, regardless of the specific values observed for the covariates. This yields the constraint

$$\mu_1(\mathbf{x}; b_1^*) \geq \mu_0(\mathbf{x}; b_0^*) \;\Leftrightarrow\; \mathbf{x}\;\; (b_1^* - b_0^*) + (b_{01}^* - b_{00}^*) \geq 0 \quad \text{for all } \mathbf{x} \in \mathcal{X}, \tag{1}$$

where the equivalence is according to (2) and the associated comment.

The constraint (1) has recently been studied in [47]. If the random variables e_0 and e_1 in are symmetric about zero, (1) is equivalent to *stochastic precedence ordering* [2] of the random variables $T|\{D = 0, X = \mathbf{x}\}$ and $T|\{D = 1, X = \mathbf{x}\}$ for all $\mathbf{x} \in \mathcal{X}$. Moreover, in the case of identical scale functions in the two status groups, the constraint (1) is equivalent to (ordinary) stochastic ordering, i.e., $P(T \geq t|D = 1, X = \mathbf{x}) \geq P(T \geq t|D = 0, X = \mathbf{x})$ for all $t \in \mathbb{R}$ and all $\mathbf{x} \in \mathcal{X}$. Order constraints have received considerable attention in recent literature, e.g., [7, 40, 35, 37], including papers discussing such constraints in the context of ROC curve modeling [19, 43]. Their incorporation can be beneficial for at least two reasons: first, they yield more interpretable results in applications in which those constraints are known to be satisfied; second, as shown in [47], they can improve statistical efficiency in settings with low

sample size or weak separation of the score distributions in the two populations.

1.3.3 Research Question 3: Develop evidence interpretation tools based on covariate-specific likelihood ratios

The likelihood ratio (LR) is $LR = Pr(Y_u, Y_p|H_p)/Pr(Y_u, Y_p|H_d)$ under the propositions. Here the numerator is the joint probability mass function or probability density function of evidence measurements Y_u from unknown source objects and Y_p from a known source, under the prosecutor's hypothesis H_p that Y_u and Y_p come from the same source. The denominator is the joint probability or density of Y_u and Y_p under the defendant's hypothesis H_d that Y_u comes from a difference source from Y_p. The Bayes factor updates the prior odds as follows $Pr(H_p|Y_u, Y_p)/Pr(H_d|Y_u, Y_p) = LR \times Pr(H_p)/Pr(H_d)$, with the last term being the prior odds of favoring the prosecutor's hypothesis H_p relative to the defendant's hypothesis H_d without the knowledge of any evidence. [21] provides an earlier review of the relevant statistical methods including LR.

We use the upper-case letters such as Y_u and Y_p to denote random variables with probability distributions. *X denotes covariates characterizing demographics of the known source and properties of relevant population.* Let these original trace evidence measurements be Y_u from the unknown source, and Y_p from the known source with covariates, Y_u and Y_p are either univariate measurements or vectors of multivariate measurements. Besides the evidence measurements, another important dataset, ideally, is e_a, in the reference population database. The reference population database provides the estimates of marginal density distributions and joint distributions of Y_u and Y_p conditional on the covariates X. Here the covariates X can be background information such as demographics on the known source. In the context of fire debris, the covariates from the known source could be the container material for the ignitable liquid. With the container information, examiners will likely have more precise understanding of the fire debris mixture. Thus, the LR conditional on the container type will provide a better way to quantify the weigh of the evidence than the LR ignoring the container type. Figure 4 provides the diagram of the notations used for evidence, ROC and LR. Note

that for trace evidence, the LR can be calculated directly from trace observations, while for impression and pattern evidence, the LR is calculated mostly from matching scores.

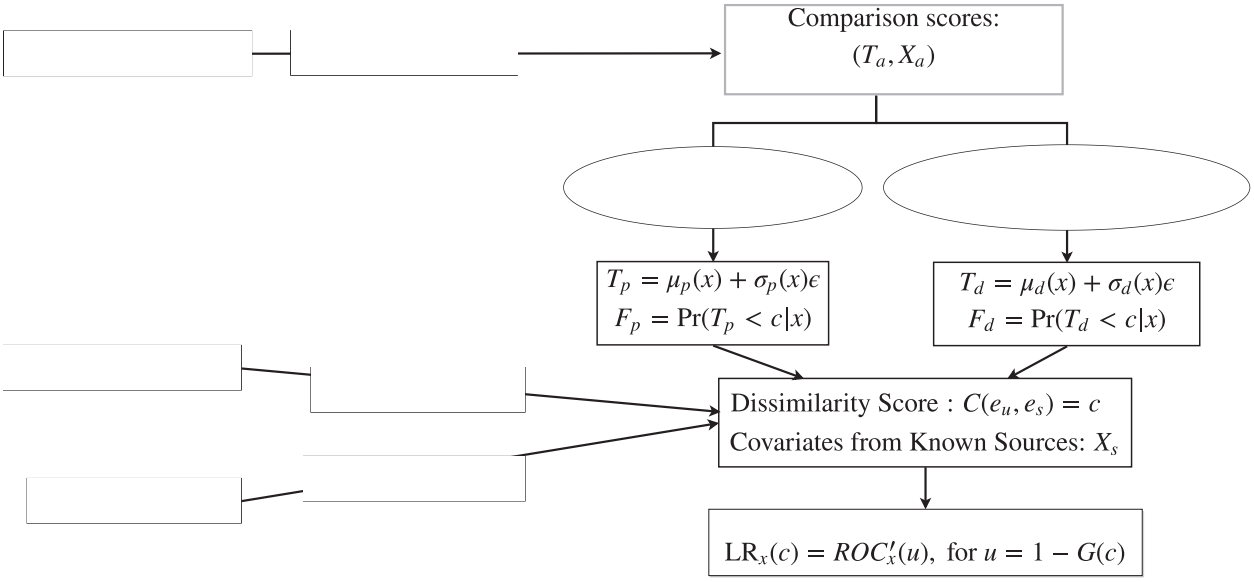

Figure 4: Schematic overview on notations used for evidence, ROC, and LR.

In this project we will work with the multivariate LR which handles multi-dimensional trace observations.

Without conditioning on the covariate X_s, this LR is the same as the Lindley's LR for univariate data. It is commonly assumed that within-source distributions conditional on Y_p are normal, that is, $Y_{u,i} \sim N(\theta_u, \Sigma_u^2)$. It is also reasonable to assume that $Y_{s,j}|(X_s = x) \sim N(\mu_p(x), \Sigma_p^2)$, where $\mu_p(x)$ is a covariate-specific mean. Note that the covariate information is available for the known source and not for the unknown source unless covariates of the unknown source are witnessed by bystanders. The prior distribution for θ_u and $\mu_p(x)$ can be a normal distribution $N(\tau, \Sigma_\theta)$, or an exponential distribution. The sample conditional mean $\bar{Y}_u$ is sufficient statistics for θ_u. When $Y_{u,i}$ and $Y_{s,j}$ are from a common source, $\theta_u = \mu_p(x) = \theta$. The LR is then simplified as: $LR_x = \frac{\int f(\bar{Y}_u, \bar{Y}_s|\theta) f(\theta) d\theta}{\int f(\bar{Y}_u|\theta_u) f(\theta_u) d\theta_u \int f(\bar{Y}_p|\mu_p(x)) f(\mu_p(x)) d\mu_p(x)}$. Without covariates, several authors [1, 4] let $\mu_p(x) = \theta_p$, and give the explicit form of the LR based on the multivariate normal distributions for $Y_{u,i}$ and $Y_{s,j}$, and θ's when the covariance matrices are not the same. Without covariates, a kernel estimation can be substituted to replace $f(\theta)$ in the integration of the numerator and denominator for the LR [23, 1, 4] if

the normal assumption for θ is not true. In this project, the investigators will estimate the covariate-specific mean $\mu_p(x)$ for subpopulations, and obtain the covariate-specific LR. Specifically, $\mu_p(x)$ has a prior distribution such as multivariate normal distribution with its mean and covariance parameters calculated from the relevant database. The posterior distribution of $\mu_p(x)$ will then estimated from the prior and the evidence measurements from the known source. The maximum a posteriori probability estimate of $\mu_p(x)$ will be the mode of the posterior distribution.

The score-based LR is commonly used for the decision scores from the same subjects and different subjects in impression and pattern evidence. For impression and pattern evidence, [26] developed a summary statistic or comparison methodology, denoted as $C(e_u, e_s)$. The realized value of the comparison statistic $C(e_u, e_s) = c$ is obtained by comparing the control samples from a specified (known) source (denoted by e_s) to the recovered traces from a questioned or unknown source (denoted by e_u). Similar to Parker, we assume that $C(\cdot, \cdot)$ is a dissimilarity score, with the bigger value of $C(\cdot, \cdot)$ indicating the less similarity between e_u and e_s. It is known that the first derivative of the ROC curve gives the LR when covariates are not considered [14, 8]. The connection is similar with covariates. Define f_p and f_d be the density functions of F_p and F_d, respectively. Note that based on definition, the densities of $T_{i,p}$ and $T_{j,d}$ conditional on the covariates such as subpopulations $X = x$ are $f_p(t|x) = F_p(t|x)$, and $f_d(t|x) = F_d(t|x)$. These densities will be estimated from the relevant database e_a – collection of sets of objects each associated with a source of evidence in an alternative source population.

1.4 Expected applicability of the research

The research provides forensic researchers ready-to-use statistical tools so that covariate-specific and covariate-adjusted error rate assessment can be implemented for black box studies in various forensic disciplines. Also, implementing the ordinal decision scores as in [32] have more intrinsic information for the error rates than dichotomized binary decisions. Availability of the computer packages for calculating the ROC curve with and without covariates

will provide a useful tool to forensic scientists for the analysis of their black box studies with ordinal decision scores. In addition, given that the PI and investigators on the project are currently supporting a number of local crime labs, this research will directly impact the guidance they are providing on the error rate interpretation and evidence interpretation.

2 Participants and other collaborating organizations

2.1 Individuals involved in the project

Dr. Larry Tang (PI - University of Central Florida): Dr. Tang's research background in statistics in forensics and criminology, biometrics, and nonparametric methodology in high-dimensional settings was crucial to successful completion of the aim involving the relationship between ROC curves and likelihood ratios. His work with NIST in biometrics on developing statistical methodology to advance the evaluation of fingerprint matching algorithms and to advance the understanding of forensic methods in biometric matching provided him with the necessary background to supervise completion of the project.

Dr. Danica Ommen (Iowa State University): Dr. Ommen has extensive training and expertise in forensic statistics, and computational statistics. Her doctoral research concerned the use of Bayesian likelihood ratio and frequentist likelihood ratio in a forensic setting. Her past experiences deriving and evaluating likelihood ratios within complex scenarios aided the research group in developing and assessing novel methodologies developed for complex cases of handwriting evidence. Her expertise in programming and Bayesian methodology, as well as her forensic background was especially important to the successful completion of the paradigms of evidence interpretation.

Dr. Christopher Saunders (South Dakota State University): Dr. Saunders has past experience with NIH funded projects and Intelligence Community (IC) research fellowships. Since completing his dissertation, Dr. Saunders has focused on providing statistical support to the Intelligence Community, first as an IC Postdoctoral Research Fellow

and then as a Research Assistant Professor with the Document Forensics Laboratory at George Mason University. In an ongoing collaboration with Gannon Technologies Group, he contributed to the development of a highly accurate handwriting based identification tool, known as FLASH ID. Dr. Saunders was specifically responsible for investigating the accuracy of the handwriting based biometric identification procedures as a function of the amount of handwritten text available. Recently Dr. Saunders has been focused on the development of forensic likelihood ratios for assessing the strength of handwriting evidence. Dr. Saunders' background in statistical approximation theory was highly important in the development of conclusion scales.

Dr. Martin Slawski (George Mason University): Dr. Slawski has extensive research experience in statistical modeling, machine learning, and mathematical optimization in statistical settings. His efforts on massive data inference was funded by major government agencies. In the project, Dr. Slawski will be developing methods for ROC and likelihood ratio estimation with a specific focus on regression modeling, particularly quantile regression and stochastic ordering constraints. He will provide additional computational support for various aspects of the project.

Dr. Emanuela Marasco (George Mason University): Dr. Marasco's research experience involves Pattern Recognition, Machine Learning, Computer Vision and Biometrics. Her main contribution has focused on design and evaluation of anti-spoofing countermeasures in fingerprint recognition systems, and automatic estimators of soft biometrics from fingerprints. Dr. Marasco has collaborated on several projects funded by the major government funding agencies. Dr. Marasco's background will be important in the implementation of the methods in biometrics data.

Dr. Semhar Michael (South Dakota State University): Dr. Semhar Michael is an applied statistician by training. Her research focuses on computational statistics with an emphasis on developing novel methodologies for analyzing datasets in challenging

forms. She has addressed problems in clustering of time series, forensic, and text data. Her work has been published in peer-reviewed statistics journals and led to national paper competition awards. With unstructured data, she worked on a building clustering, classification, and sentiment analysis models. In the health sciences, she has worked on mixture modeling, spatial clustering, and forecasting projects on datasets from Electronic Health Records and South Dakota Department of Health.

Funded Ph.D. Graduate Student: Dr. Ty Nguyen (University of Central Florida), advised by Dr. Larry Tang

Unfunded Graduate Students: Ph.D. students whose research was related to this funded project

- Dr. Xiaochen Zhu (George Mason University): advised by Drs. Larry Tang and Martin Slawski
- Ms. He Qi (George Mason University): advised by Dr. Martin Slawski
- Mr. Andrew Simpson (South Dakota State University): advised by Drs. Saunders and Michael
- Mr. Dylan Borchert (South Dakota State University): advised by Drs. Saunders and Michael

2.2 Organizations involved in the project

The work performed for this project has supported the Federal Bureau of Investigation Laboratory Division on research projects related to the analysis of forensic evidence from improvised explosive devices. NIST has provided rating scores of forensic examiners on NIST blackbox facial recognition study for us to implement the developed methods. We are fortunate to be able to benefit from their expertise in these areas.

3 Outcomes

3.1 Activities/accomplishments

During the period of performance, the PI and investigators engaged in virtual meetings to update the other participants on research projects, and met at conferences to coordinate and conduct collaborative efforts. Overall, this award resulted in the training of 5 graduate students in the interpretation of forensic evidence, including 5 PhD graduate students and 1 MS graduate student. This award directly resulted in 1 PhD dissertation, 4 peer-reviewed journal articles, 2 peer-reviewed conference papers, an R-Shiny app and 21 conference presentations. For a detailed list of research products and conference presentations, see Section 4.

3.2 Results and findings

3.2.1 Research Question 1: Develop error rate interpretation tools using regression on decision scores

We denote the upper-case letter X as data which contains covariates as columns and samples as rows. $X = x$ is then understood as a specific value of covariates in the data. In the data, upper-case R is used to describe the ordinal score whose values are from 1 to L where L is called ordinal scale. Also, a binary status is denoted as D where $D = 1$ or $D = 0$ splits observations into two sub-classes. Upper-case G is the number of rater groups who give ordinal scores as assessing subjects.

In this project, we use ROC curve to characterize the accuracy of performances. Let Y denote a continuous random variable related to scores in evaluation. The general formula of a ROC curve is expressed as a function of FPR as $ROC\left(t\right) = S_1 \ \ S_0^{-1}\left(t\right) \ \ , \quad t \in (0,1)$ where $S_0\left(c\right) = P\left(Y \geq c | D = 0\right)$ and $S_1\left(c\right) = P\left(Y \geq c | D = 1\right)$ are FPR and TPR with threshold c. If S_0 and S_1 follow normal distributions, it follows $ROC\left(t\right) = \Phi \ \ \frac{\mu_1 - \mu_0}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\Phi^{-1}\left(t\right) \ \ , \quad t \in (0,1)$ where $\mu_0, \mu_1, \sigma_0, \sigma_1$ are the means and the standard deviations of two sub-populations, respectively. Then, the AUC also has a explicit form as $AUC = \Phi \ \ (\mu_1 - \mu_0)/ \ \ \overline{\sigma_1^2 + \sigma_0^2} \ \ .$ In the present study, $ROC_{\mathbf{x},g}(t)$ and $AUC_{\mathbf{x},g}$ are the ROC curve and the corresponding AUC

at a specific covariates $\mathbf{x}$ of group g with $g = 1, \cdots, G$. Covariates in face recognition data are raters' group, age, gender. Also, because ROC curves are built within framework of the ordinal regression, their variances are determined by variance of parameters in the model. For the sake of making inference easily, we denote $ROC_{\mathbf{x},g}(t)$ and $AUC_{\mathbf{x},g}$ be the estimated ROC curve and AUC at covariate $\mathbf{x}$ of group g and estimated parameter $\hat{\boldsymbol{\gamma}}$ of the ordinal regression.

A Homogeneity Test

In this section, we introduce a homogeneity test for ROC curves. Assume that there are G rater groups each of which includes $J_1, J_2, \cdots, J_G$ members assessing $K = K_0 + K_1$ subjects such as images in medical diagnostics or image pairs in fingerprint or facial recognition. Out of K, there are K_0 non-diseased subjects in medical diagnostic or K_0 different sources image pairs in fingerprint or facial recognition and K_1 diseased or same source ones. Accuracies of groups are characterized by $ROC_{x,1}(t), \cdots, ROC_{x,G}(t)$ or $AUC_{x,1}, \cdots, AUC_{x,G}$ respectively. The goal is to test homogeneity among groups. The null hypothesis of the test is stated as all groups have the same accuracy while the alternative is supported if there exist differences among groups. It is noteworthy to mention that the ROC curves are functions of TPR with respect to FPR. Therefore, the test is conducted at each fixed FPR. We define a vector as $\boldsymbol{\Lambda} = (ROC_{x,1}(t), ROC_{x,2}(t), \cdots, ROC_{x,G}(t))$. If a new vector $\boldsymbol{\Lambda}_C$ is defined by subtracting $ROC_{x,G}(t)$ from $\boldsymbol{\Lambda}$ as $\boldsymbol{\Lambda}_C = (ROC_{x,1}(t) - ROC_{x,G}(t), \cdots, ROC_{x,G-1}(t) - ROC_{x,G}(t))$, the null hypothesis is now formulated as $H_0 : \boldsymbol{\Lambda}_C = \mathbf{0}$ vs. $H_a : \boldsymbol{\Lambda}_C = \mathbf{0}$. The relationship between $\boldsymbol{\Lambda}_C$ and $\boldsymbol{\Lambda}$ can be expressed as $\boldsymbol{\Lambda}_C = \mathcal{K}\boldsymbol{\Lambda}$ where $\mathcal{K} = (I_{G-1}, -\mathbf{1}_{G-1})$ with an identity matrix I_{G-1} and a vector of one's $\mathbf{1}_{G-1}$. In this case, we use $ROC_G(t)$ as a reference for comparison purpose. In fact, any group can be in charge of the role. With a given data, we need to estimate the ROC curves to proceed the test. An estimate of a ROC curve, denoted with a hat, can be retrieved nonparametrically or parametrically [44]. With the first approach, an empirical ROC curve is obtained. Alternatively, parametric methods need to assume

distributional forms for two populations. Thus, ROC curve can be derived analytically. In this project, we use the later technique with binormality assumption for scores within framework of ordinal regression discussed in the next section.

ROC Estimators based on Ordinal ROC Regression

Assume that we want to bridge L-scale ordinal scores R with observable variables comprised in a matrix X. Without loss of generality, we denote the first column of X as D which is a binary variable of 0 or 1. Then, D splits observations into two sub-groups such as diseased and non-diseased status in medical diagnostics, genuine and imposter scores in facial recognition.

We use a location-scale model to estimate the covariate-specific ROC curve. In the model, each of outcomes $R_i, i = 1, 2, \cdots N$ links to an example which is described by a vector of covariates, $X_i = \{D_i, x_{i1}, ..., x_{ip}\}$ or $\{D_i, \mathbf{x}_i\}$ where p is the number of covariates and N is the total number of observations. It is noteworthy that out of p covariates, one represents for group status of raters. The ordinal ROC regression starts by supposing that discrete outcomes R belong to a latent continuous variable T which can be partitioned into sub-regions by thresholds $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{L-1} < \tau_L = \infty$. The outcome R receives the value r_l if $\tau_{l-1} < T \leq \tau_l$. The general formula of ordinal regression can be expressed as

$$g\left[\phi_l\left(R \leq l|\mathbf{x}\right)\right] = \frac{\tau_l - \left(\alpha_0 D + \boldsymbol{\alpha}_1\mathbf{x} + D\boldsymbol{\alpha}_2\mathbf{x}\right)}{\exp\left(\beta_0 D + \boldsymbol{\beta}_1\mathbf{x} + D\boldsymbol{\beta}_2\mathbf{x}\right)},$$

where $l = 1, 2, ..., L-1$ and $\mathbf{x}$ denote for any of $\{\mathbf{x}_i\}_{i=1}^{N}$ where $N = K \sum_{g=1}^{G} J_g$ is the total number of observations, $g\left(\cdot\right)$ is the link function, $\phi_l\left(R \leq l|\mathbf{x}\right)$ is the cumulative probability that $R \leq l$, a vector production, for example $\boldsymbol{\alpha}_1\mathbf{x}$, is written as $\boldsymbol{\alpha}_1\mathbf{x} = \alpha_{11}x_1 + \cdots + \alpha_{1p}x_p$. With the probit link, the model is rewritten as

$$\phi_l\left(R \leq l|\mathbf{x}\right) = \Phi\left(\frac{\tau_l - \left(\alpha_0 D + \boldsymbol{\alpha}_1\mathbf{x} + D\boldsymbol{\alpha}_2\mathbf{x}\right)}{\exp\left(\beta_0 D + \boldsymbol{\beta}_1\mathbf{x} + D\boldsymbol{\beta}_2\mathbf{x}\right)}\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. With this approach, the latent variables for a particular covariate $\mathbf{x}$ are normally distributed with means and standard deviations described in Table 1.

Table 1: Ordinal Regression ROC Parameters

	$D=1$	$D=0$
Mean	$\mu_1 = \alpha_0 + (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2)\mathbf{x}$	$\mu_0 = \boldsymbol{\alpha}_1\mathbf{x}$
Standard dev.	$\sigma_1 = \exp\{\beta_0 + (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)\mathbf{x}\}$	$\sigma_0 = \exp\{\boldsymbol{\beta}_1\mathbf{x}\}$

Substitute the means and standard deviations in Table 1, the ROC curve within the framework of ordinal regression for a specific covariate $\mathbf{x}$ of group g is finalized as

$$ROC_{\mathbf{x},g}(t) = \Phi\left(\frac{\alpha_0 + \boldsymbol{\alpha}_2\mathbf{x}}{\exp(\beta_0 + \boldsymbol{\beta}_1\mathbf{x} + \boldsymbol{\beta}_2\mathbf{x})} + \frac{1}{\exp(\beta_0 + \boldsymbol{\beta}_2\mathbf{x})}\Phi^{-1}(t)\right), \quad t \in (0,1).$$

The corresponding AUC is also expressed as

$$AUC_{\mathbf{x},g} = \Phi\left(\frac{\alpha_0 + \boldsymbol{\alpha}_2\mathbf{x}}{\sqrt{\exp(2\beta_0 + 2\boldsymbol{\beta}_1\mathbf{x} + 2\boldsymbol{\beta}_2\mathbf{x}) + \exp(2\boldsymbol{\beta}_1\mathbf{x})}}\right).$$

To simplify notations, we use $\hat{\boldsymbol{\gamma}} \equiv \left(\hat{\boldsymbol{\tau}}, \hat{\alpha}_0, \hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, \hat{\beta}_0, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2\right)$.

Statistical Property of the Test

Let $\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Lambda}}_C$ be estimators of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_C$ which can be written as $\hat{\boldsymbol{\Lambda}} = \left(ROC_{\mathbf{x},1}(t), \cdots, ROC_{\mathbf{x},G}(t)\right)$, and $\hat{\boldsymbol{\Lambda}}_C = \left(ROC_{\mathbf{x},1}(t) - ROC_{\mathbf{x},G}(t), \cdots, ROC_{\mathbf{x},G-1}(t) - ROC_{\mathbf{x},G}(t)\right)$. The relationship $\hat{\boldsymbol{\Lambda}}_C = \mathcal{K}\hat{\boldsymbol{\Lambda}}$ still holds for estimators. The test statistic is defined as

$$\Psi = \boldsymbol{\Lambda}_C \left(\mathrm{Var}\boldsymbol{\Lambda}_C\right)^{-1} \boldsymbol{\Lambda}_C.$$

The variance $\mathrm{Var}\boldsymbol{\Lambda}_C$ is dependent on covariance matrix $\Sigma_{\hat{\boldsymbol{\gamma}}}$ which is asymptotically approx-

imated as $\Sigma_{\boldsymbol{\gamma}_0}$. Concatenating all $ROC_{\mathbf{x},g}(t)$ for $g = 1, \cdots, G$ yields to

$$\hat{\Lambda} = \Lambda + \mathrm{F}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\right) + o\ \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\right)^2\ ,$$

where $F = \ J_{ROC_{\mathbf{x},1}(t)}, \cdots, J_{ROC_{\mathbf{x},G}(t)}\ $. Using $\hat{\boldsymbol{\Lambda}}_C = K\hat{\boldsymbol{\Lambda}}$ and taking variance both sides yields to

$$\mathrm{Var}\hat{\boldsymbol{\Lambda}}_C = \mathrm{KF}\Sigma_{\hat{\boldsymbol{\gamma}}}\mathrm{F\ K}\ ,$$

where $\Sigma_{\hat{\boldsymbol{\gamma}}}$ depends on the sample size. Employing Lemma 1, one can see that $\hat{\Lambda}_C$ asymptotically follows a multinormal distribution given by $N_{G-1}\ \ \Lambda_C, \Sigma_{G-1} = KF\Sigma_{\boldsymbol{\gamma}_0}F\ K\ \ $.

Theorem: Under the null hypothesis, Ψ converges in distribution to a Chi-square distribution with $G-1$ degrees of freedom χ^2_{G-1} and under the alternative, Ψ still converges to a Chi-square distribution with the same degrees of freedom but with a non-centrality parameter $\eta = \boldsymbol{\Lambda}_C \left(\mathrm{Var}\boldsymbol{\Lambda}_C\right)^{-1} \boldsymbol{\Lambda}_C$ as $N \to \infty$.

Given a significance level α, the null hypothesis is rejected if $\Psi > \chi^2_{G-1,\alpha}$ where $\chi^2_{G-1,\alpha}$ is the critical value of a Chi-square distribution with $G-1$ degrees of freedom. Determining non-centrality parameter occurs in various statistical analysis, such as the analysis of variance for tests of homogeneity, Chi squared test for goodness of fit, power analysis. Since the power analysis usually relates to the sample size problem, the non-centrality parameter can be used to determine the minimum sample size provided the power is supplied. Solution of a power problem replies on the availability of the non-centrality of a Chi-squared distribution. Early, [16] prepared tables for the non-centrality parameter of a Chi-squared distribution with some given values of degree of freedom, significance level and power. Then, [15] calculated the minimum sample size for the three most frequently used tests at given power using those tables. Next, [34] estimated the non-centrality parameter of a Chi-squared distribution by employing the maximum likelihood technique. However, only were the lower and upper bounds derived instead of a closed form for the parameter. Thanks to developing of computer technology, nowadays we can numerically compute the non-centrality parameter. The power

$1-\beta$ where β is the probability of a type II error is defined as $1-\beta = P\ \ \chi^2_{G-1}(\eta) > \chi^2_{G-1,\alpha}\ \ .$ With given values of α, β and G, the non-centrality parameter η can be determined by solving the equation above. Denote $\eta_{\beta,\alpha}$ be the solution, using definition of η yields to $\eta_{\beta,\alpha} = \mathbf{\Lambda}_C \left(\mathrm{KF}\Sigma_{\gamma_0}\mathrm{F\ K}\right)^{-1} \mathbf{\Lambda}_C$. The minimum sample size is determined to obtain $\eta_{\alpha,\beta}$ numerically and scanning sample size until the equality is satisfied.

In this part, we describe our design for simulation. Our data includes ordinal scores, a continuous variable X_1 and discrete covariates representing for rater groups. The latent scores of g^{th} group are normally distributed as $T_g|(X_1, D=1) \sim N\left(1+2X_1+\psi+a_g, \phi\mathrm{Var}\left(e_1\right)\right)$, and $T_g|(X_1, D=0) \sim N\left(1+X_1, \mathrm{Var}\left(e_0\right)\right)$, where X_1 is uniformly distributed in $[0,1]$, e_0, e_1 are standard normal distributions. In those equations , parameters ψ, a_g and ϕ control the distance and difference in variances between two normal distributions. Moreover, parameter a_g, which only depends on the group label is utilized to adjust differences in ROC curves among groups. On the other hand, parameter a_g is able to control the null and alternative hypothesis. With those distributions, the true ROC curve and the corresponding AUC are expressed as $ROC^{true}_{\mathbf{x},g}(t) = \Phi\ \ \frac{x_1+\psi+a_g}{\sqrt{\phi}} + \frac{1}{\sqrt{\phi}}\Phi^{-1}(t) \qquad t \in (0,1)$, First, we examine the consistency of estimated ROC curves and AUCs. We assume that all groups have the same number of members, i.e. $J_1 = J_2 = \cdots = J_G = J$. Hence, term "sample size" K should be understood as the number of samples assigned to each rater. Setting 1 with 10000 data sets are simulated in this subsection. In Figure 5, estimated ROC curves and AUCs are depicted with some sample sizes where $G=5$ and $L=7$ are used. Value of other parameters can be seen in the caption. In Fig.5, estimated ROC curves of four selected sample sizes are depicted along with the true curve. It is obvious that the larger the sample size is, the closer the estimated curve approaches the exact one. In this case, a sample size between 10 and 20 is a reasonably optimal value leading to a consistent ROC curve. Next, in Fig.5b, estimated AUCs with sample sizes varying from 5 to 50 is presented. One can see that estimated AUC asymptotically converges to the true value. Indeed, the bias of estimators are just around 2% at sample size of 5 and less than 1% at 15. Thus, 15 could be consider as asymptotic value of the sample

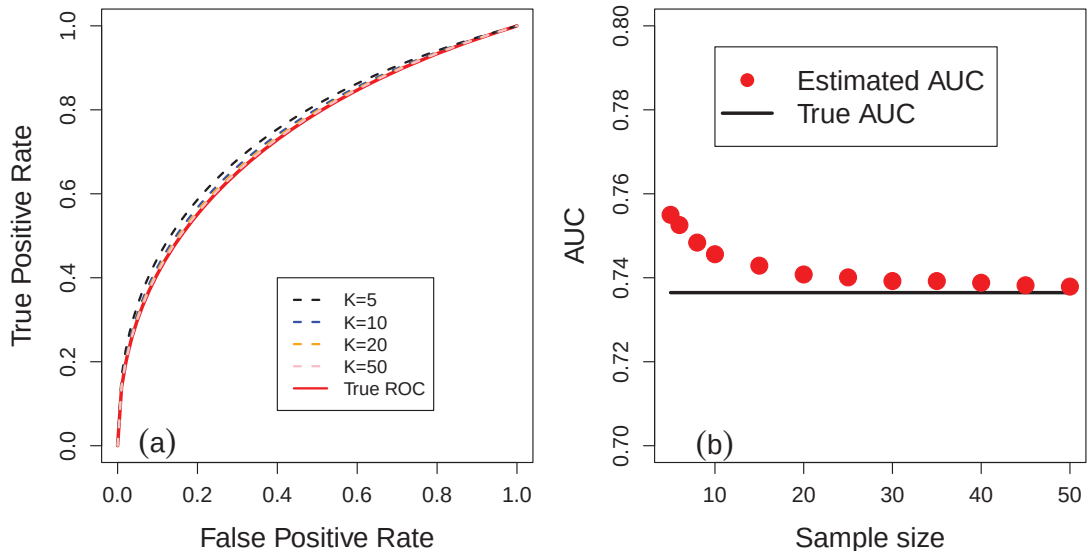

Figure 5: Convergence of ROC curves, AUCs when the sample size changes. G=5, J=10, L=7 and $\psi = 0.5$, and $\phi = 1.5$, $x_1 = 0.5$ are used. (a): Estimated ROC curves for four selected sample sizes of 5 (dashed black), and 10 (dashed blue), 20 (dashed orange) and 50 (dahsed pink). Red solid line is the true ROC curve. (b): Dots are estimated AUCs and black solid line is the exact AUC of 0.736.

size. Futhermore, we validate the quality of estimators and their variance by calculating the confidence interval coverages of difference in ROC curves and AUCs. Let $\Delta ROC_{12}(t) = ROC_2(t) - ROC_1(t), \Delta AUC_{12} = AUC_2 - AUC_1$ be the difference in ROC curves or in AUCs between the second and the first group. The coverage of $\Delta ROC_{12}(t)$ curves is the portion of the 10000 curves bounded by the $(100-\alpha)\%$ confidence interval that is $\Delta ROC_{12}(t) - Z_{\frac{\alpha}{2}} \times \sqrt{\text{var}\ \ \Delta ROC_{12}(t)}\ , \Delta ROC_{12}(t) + Z_{\frac{\alpha}{2}} \times \sqrt{\text{var}\ \ \Delta ROC_{12}(t)}$. The coverage of $\Delta ROC_{12}(t)$ at $t_1 = 0.3$ and ΔAUC_{12} with different number of groups and sample sizes are presented in Figure 6.

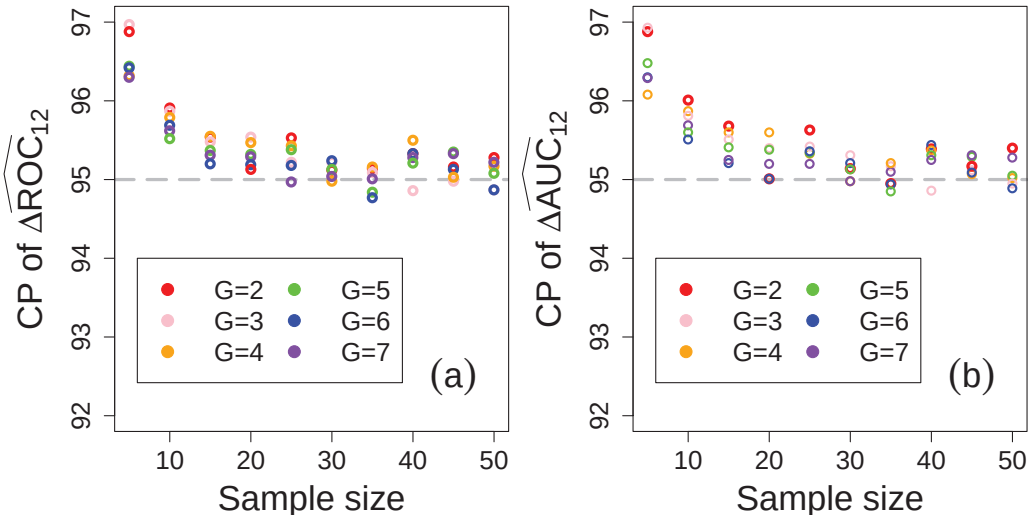

Figure 6: Coverage probabilities of the 95% confidence intervals of ΔROC_{12} curve at FPR of 0.3 (a) and of ΔAUC_{12} (b). The dashed grey line is the nominal level. J=10, L=7 and $\psi = 0.5$, and $\phi = 1.5$, $x_1 = 0.5$ are used.

In Fig.6a, confidence interval coverage of ΔROC_{12} at FPR of 0.3 is illustrated and those of ΔAUC_{12} are shown in Fig.6b. In both figures, one can see that the portions approach to 95% starting from the sample size of 100 and get closer when the size increases regardless of the number of groups. It is noticeable that the result for ΔROC_{12} is presented at one value of FPR but the similar ones are also obtained at different points on the ROC curves. It implies that the convergence occurs for the entire ΔROC_{12}. Moreover, results for other pairs are also found analogous to that of ΔROC_{12}. In addition, we compute the probability of Type I error of the homogeneity test. In Figure 7, we depict Type I error rate of the test for ROC curves at FPR of 0.3 (a) and for AUCs on (b). As seen in Fig.7, regardless of the number of group, the Type I error approaches 5% for both of tests using ROC curves and AUCs. Analogous results are also retrieved for different values of FPR.

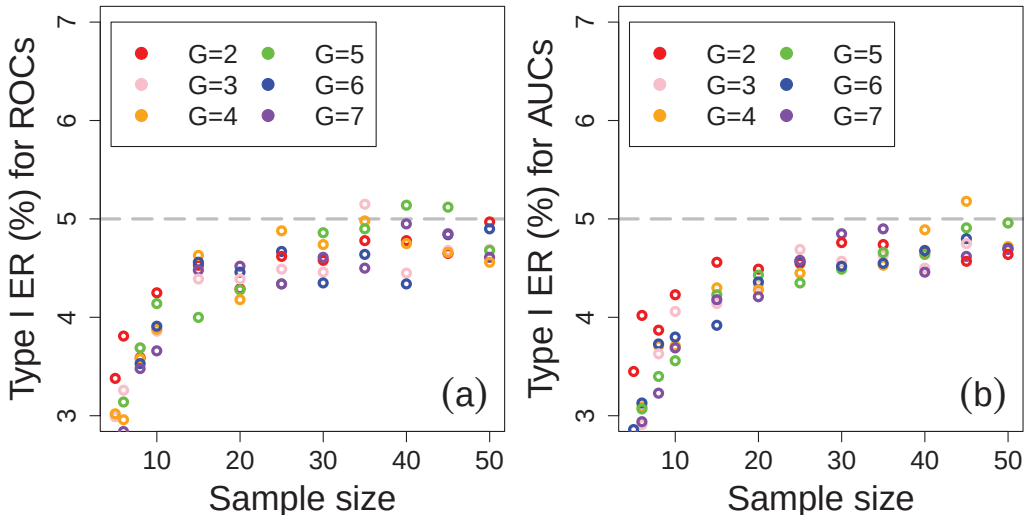

Figure 7: Type I error rate for the test for ROCs at FPR of 0.3 (a) and for AUCs (b). The dashed grey line is 5% significance level. Parameters are fixed the same as in Figure 6.

We illustrate calculation of minimum sample sizes given α and β with settings supporting for the alternative hypothesis. First, with evenly distributed samples, minimum sample sizes to reach a probability of a Type I error of 5% and a power of 80% is demonstrated in Table 2. With each setting, the minimum sample size is calculated by using TPR at three different FPRs, denoted at t_1, t_2, t_3. Estimated sample size using AUCs is also provided. With setting 2, a vector of AUC values is $\Lambda = (0.736, 0.736, \cdots, 0.736, 0.829)$ whose last entry is larger than others. With setting 3, $\Lambda = (0.736, 0.736..., 0.736, 0.829, 0.829)$ where last two identical elements are larger than others. With setting 4 where all elements are different, $\Lambda = (0.736, 0.776, 0.812, 0.844, 0.873, 0.897, 0.918)$ for seven groups. If less groups are needed, for instance five groups, first five elements are used.

Table 2: Minimum sample sizes with $\alpha = 0.05, \beta = 0.2$. J=10, L=7 and $\psi = 0.5$, and $\phi = 1.5$, $x_1 = 0.5$ are used. FPRs at $t_1 = 0.3, t_2 = 0.5, t_3 = 0.7$ are selected.

	Setting 2				Setting 3				Setting 4			
G	t_1	t_2	t_3	AUC	t_1	t_2	t_3	AUC	t_1	t_2	t_3	AUC
3	41	47	58	41	42	49	68	43	80	89	115	80
4	35	37	47	35	30	35	46	30	37	43	56	37
5	33	35	41	33	24	28	34	24	19	24	30	22
6	31	34	38	31	21	23	30	25	12	14	19	12
7	29	31	36	29	20	22	26	20	8	10	13	8

As seen in Table 2, with each setting, a larger sample size is needed if a higher FPR is used. This can be explained as at higher FPR, the gap between curves are narrower. Furthermore, with three settings, the minimum sample sizes retrieved from ROC curves at FPR of 0.3 are similar to those from AUCs. That could be because at FPR of 0.3, the gaps in TPRs, $\mathbf{\Lambda}_C^{\{ROC\}}$ and difference in AUCs, $\mathbf{\Lambda}_C^{\{AUC\}}$, among rater groups are similar. Next, we investigate the scenario in which groups have different number of raters. Denote $n_1 : n_2 : \cdots : n_G$ be the ratio of the number of raters among groups, i.e. $J_1 = n_1 J, \cdots, J_G = n_G J$. We use setting 4 with four groups for following calculations. The minimum sample size presented in Table 3 are the number of samples assigned for each rater. As seen in the table, the minimum sample size is sensitive to changes of the ratio. Assume that among four groups, one has twice samples than others which is described in first four rows in Table 3. It is obvious that which group has more samples influences the total minimum sample size. This finding is also seen with different ratios of samples. Once again, the minimum sample sizes retrieved from higher FPRs are larger than that from a lower one. Additionally, sample sizes obtained by using FPRs of 0.3 is still similar to those from AUCs.

Table 3: Minimum sample sizes with different number of raters in groups. $\alpha = 0.05, \beta = 0.2$ J=10, L=7 and $\psi = 0.5$, and $\phi = 1.5$, $x_1 = 0.5$ are used. FPRs at $t_1 = 0.3, t_2 = 0.5, t_3 = 0.7$ are selected.

Ratio	t_1	t_2	t_3	AUC	Ratio	t_1	t_2	t_3	AUC
1:1:1:2	28	33	45	29	2:1:1:2	20	23	30	20
1:1:2:1	37	42	57	37	2:1:2:1	25	28	38	25
1:2:1:1	35	39	51	35	1:2:3:4	20	24	33	20
2:1:1:1	26	29	38	26	4:3:2:1	18	19	24	18
1:1:2:2	29	32	46	29	4:2:2:1	17	19	25	17
1:2:1:2	26	30	40	26	1:2:2:4	20	24	32	20
2:2:1:1	25	28	37	25	4:2:1:4	10	11	15	10
1:2:2:1	34	37	50	34	4:1:2:4	10	12	16	10

Application of the proposed method to facial recognition Forensic facial examiners perform detailed comparisons between images of two faces and determine if the faces are from the same person or different people. Examiners' extensive training and qualifications allow them to give expert opinion in court proceedings. Because of facial examiners' detailed comparisons, the field of facial forensics is a pattern-based forensic discipline. Two reports identified the necessity to empirically measure error rates for pattern-based disciplines in forensics [24, 33]. [32] provided the needed scientific evidence of facial examiners' ability by conducting a study that measured examiners' accuracy when they performed forensic comparisons. To assess examiners' ability relative to other groups, the study measured the accuracy of forensic facial reviewers, super-recognizers, fingerprint examiners, and students. Forensic facial reviewers are trained to perform facial comparisons faster than examiners. Super-recognizers possess a natural ability to recognize faces. Fingerprint examiners specializing in comparing latent fingerprints. Students served as a proxy for the general population. Next we give an overview of the methods in [32]. The participants consisted of 57 facial ex-

aminers, 30 facial reviewers, 13 super-recognizers, 53 fingerprint examiners and 31 students. Each participate judged the similarity of the same 20 face-pairs. For each face-pair, participants judged the similarity of the two faces on a 7-point scale, with +3 for the highest confidence of same person to –3 for the highest confidence of different people. [32] computed accuracy at the individual level by computing the AUC for each participant. They reported overall group accuracy with the median AUC of the group and compared two groups with the Mann-Whitney test. In our analysis, we pool participants for each of the five groups and we assume members within the same group have the same accuracy. Using scores as outcomes and group status as covariates, we estimate the ROC curves and the corresponding AUCs for each group. The two methods produce slightly different results, but overall the results from the two studies are consistent.

We start our analysis by applying our ordinal regression technique to facial recognition ratings and estimating the ROCs and AUCs for each of the first subject groups. For the ROCs we compute the 95% confidence bands and for the AUC we compute the 95% confidence intervals. Figure 8 shows estimated ROCs and AUCs with corresponding 95% confidence bands and intervals for each group. Based on the AUC estimates, the facial examiners has the highest AUC followed by super-recognizers, facial reviewers, fingerprint examiners, and students. This order agrees with [32].

Next, we check if the the AUCs and ROCs for five groups are statistically different. We formulate this question as a hypothesis test with the null hypothesis that the AUCs (respectively ROCs) for all five groups are statistically the same. If the null hypothesis is not true, then at least one of group's AUCs (respectively ROCs) are statistically different than other four groups. First, we test AUCs and for the five groups, then the ROCs. For AUCs, we compute the homogeneity test statistic $\Psi = 129.2$. Since this test statistic is larger than $\chi^2\left(0.95, df = 4\right) = 9.49$, the null hypothesis is rejected with a 95% confidence level, and the AUCs are not the same for all five groups. We perform the homogeneity test for the

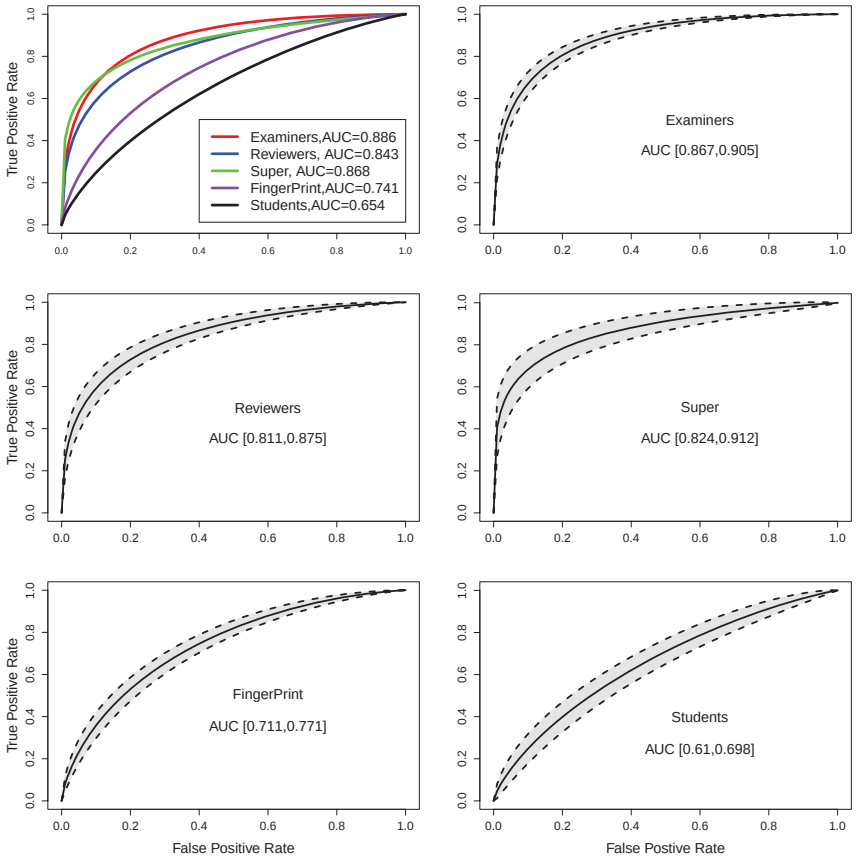

Figure 8: Plot of estimated ROCs and their 95% confidence bands for the five participant groups. The upper left panel shows the estimated ROCs for all five groups and the legend reports the estimated AUC for each group. The remaining panels plot the ROC and confidence bands for each group individually. In each panel, the solid lines is the estimated ROC, the dashed lines are upper and lower bound of the 95% confidence bands, and the legend states the group and reports the 95% lower and upper confidence intervals for the AUC.

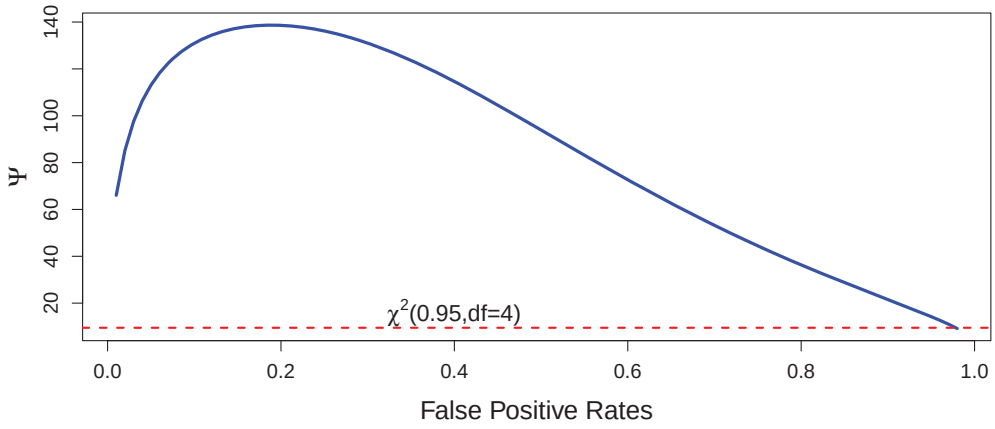

Figure 9: Plots the values of the test statistic Ψ for ROC curves (the blue solid line) versus FPR. The dashed line is the critical value for the 95% confidence of the Chi-square distribution with df=4.

ROCs, which requires computing the test statistic Ψ for all FPR values. In Figure 9 we plot the value of test Ψ as a function of FPR. The test statistic is larger than the critical value $\chi^2\,(0.95, df = 4)$, except for FPR values close to 1. Thus, the ROCs are different for FPRs smaller than 0.95 and the same for FPRs greater than 0.95.

Since the homogeneity tests showed differences among the AUCs and ROCs of the five groups, we perform post hoc pairwise comparison. This allows us to identify which groups have different AUCs or ROCs. We assess statistical differences between two ROCs, by comparing the ROCs at each FPR. For comparing two ROCs, there are three possible conclusions: the two ROCs are statistical the same for all FPR, they are statistical different for all FPR, or for some FPRs the two ROCs are the same and for some FPRs they are different. In our analysis we found all three cases. In most applications, systems operate a low FPRs, and our technique allows engineers to focus on the FPR relevant to their applications. Figure 10 shows the pairwise comparison for four groups: examiners, reviewers, super-recognizers, and fingerprint examiners. We start by looking at the pairwise comparison of facial examiners and fingerprint examiners, upper-left-hand plot in Figure 10. The horizontal axis corresponds to FPR, and the vertical axis reports the ΔTPR, the difference between the two ROCs at

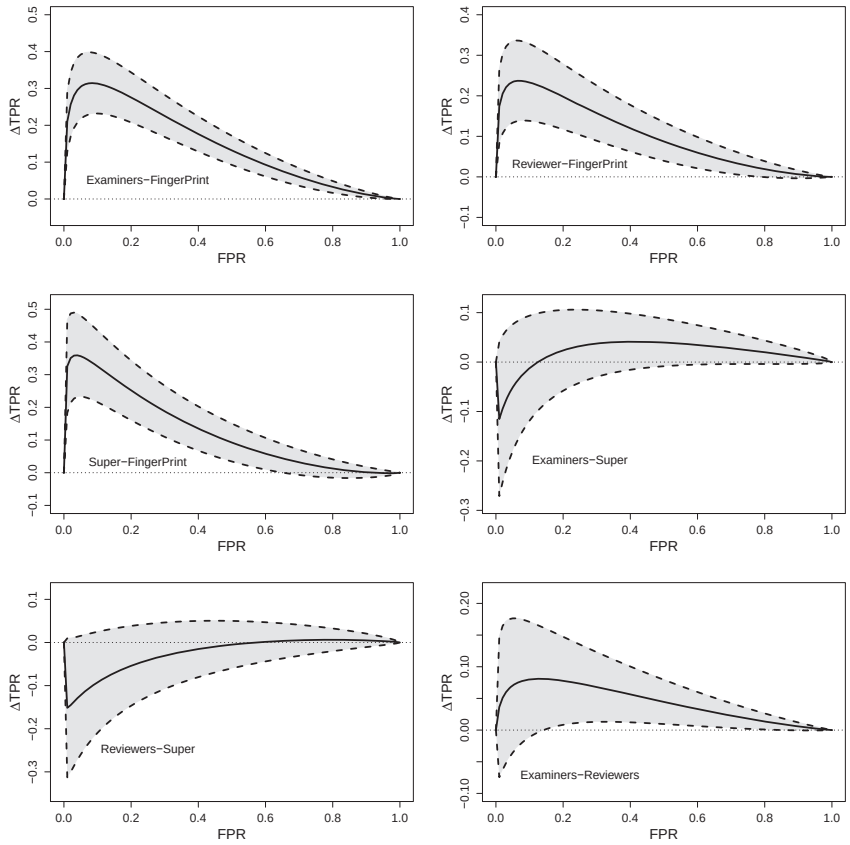

Figure 10: Differences between ROC curves of four groups in facial recognition data. The horizontal axis corresponds to FPR. The vertical axis reports the ΔTPR, the difference between the two ROCs at each FPR. The solid line shows the estimated difference between the ROCs. Dashed lines are upper and lower bounds of the 95% confidence band.

each FPR. The solid line shows the estimated difference between the ROCs' for the face examiners minus the fingerprint examiners. Dashed lines are upper and lower bounds of the 95% confidence band. For all FPRs, the 95% confidence band, the gray region, is above the $\Delta TPR = 0$ line. Thus, for the entire ROCs, the face examiners and fingerprint examiners are statistical different with 95% confidence.

For facial examiners and super-recognizers, the 95% confidence band contains the $\Delta TPR = 0$ line, therefore, the differences between the ROCs is not statistical significant with 95% confidence for all FPRs. We get the same findings when comparing facial reviewer and super-recognizers. These results are consist with the previous ad-hoc analysis for AUCs that found no statistical difference with 95% confidence. For facial examiners and reviewers, the

confidence band is not above the $\Delta TPR = 0$ line, nor does the band contain the $\Delta TPR = 0$ line. Instead, for $FPR \leq 0.15$ and $FPR > 0.8$, the the band contains the $\Delta TPR = 0$, and for $0.15 < FPR \leq 0.8$, the the band contains the $\Delta TPR = 0$ line. Thus, for $FPR \leq 0.15$ and $FPR > 0.8$, the examiners and reviewers have the same accuracy with a 95% confidence, and for $0.15 < FPR \leq 0.8$, the examiners and reviewers have different accuracy with a 95% confidence. In the majority of applications, the operating point requires a low FPR. Systems general operate at a low FPR to minimize false accusations. The comparison between super-recognizers and fingerprint examiners has a similar pattern. For $FPR < 0.6$, the difference is significant, and for $FPR > 0.6$, the difference is not significant–both with 95% confidence.

Overall, our conclusions are consistent with [32] , with each having difference strengths. [32] concentrated on the accuracy of individual participants and permitted examination of the range of accuracy for members of each group. Our analysis treat groups as covariates, and analysis produced ROCs with confidence bands and AUCs with confidence intervals . One key strength of out approach is the ability to produce results at operationally relevant decision thresholds. Since the majority of applications operate at low FPRs, producing results with error bands for ROCs will enable examiners, analysts and engineers to concentrate on the appropriate FPRs.

3.2.2 Research Question 2: Develop error rate interpretation tools using ROC regression models

The location-scale model [12] is an established approach for modeling covariate-specific ROC curves, which we summarize here. 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 $T = D\{\mu_1(X; b_1^*) + \sigma_1(X; a_1^*) \cdot e_1\} + (1-D)\{\mu_0(X; b_0^*) + \sigma_0(X; a_0^*) \cdot e_0\}$. The above equation involves the following quantities:

- e_0 and e_1 are random variables with "location" zero and unit "scale", where the measure of location can be, for example, the population mean or median, and the measure of scale can be the standard deviation or the median absolute value of the differences from the median (MAD).

- $\mu_j(\cdot)$ and $\sigma_j(\cdot)$, $j \in \{0,1\}$, are referred to as *location* and *scale functions*, respectively. The latter are assumed to be known functions of the covariates X, and depend on unknown parameters b_j^* and a_j^*, $j \in \{0,1\}$. In this project, we confine ourselves to functions having identical form for both values of D and that are linear functions of unknown parameters, i.e.,

$$\mu_j(\mathbf{x}; b_j^*) = \phi(\mathbf{x}) \;\; b_j^* + b_{0j}^*, \qquad \sigma_j(\mathbf{x}; a_j^*) = \Psi(\mathbf{x}) \;\; a_j^* + a_{0j}^*, \qquad j \in \{0,1\}, \quad (2)$$

  so that $b_j^* = (b_{0j}^*, [b_j^*]\ )$ and $a_j^* = (a_{0j}^*, [a_j^*]\ )$. To avoid notational clutter, we henceforth assume without loss of generality that $\phi(\mathbf{x}) = \mathbf{x}$ since this can be achieved 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 σ_1 and σ_0 do not depend on X; otherwise, we shall speak of *heteroscedasticity*.

Proposed approach

A major shortcoming of the composite quantile regression (CQR) is the requirement of i.i.d. errors. Consequently, the model associated with CQR is misspecified under heteroscedasticity, and the resulting estimates can exhibit substantial bias. Even though CQR may be arbitrarily more efficient than plain median regression in the i.i.d. case, the gain in efficiency can be rather moderate for common error distributions: for the family of Gaussian scale mixtures, which includes the Laplacian, logistic, and t-distribution among many others, the relative efficiency cannot exceed 1.5 (with the upper bound being attained by

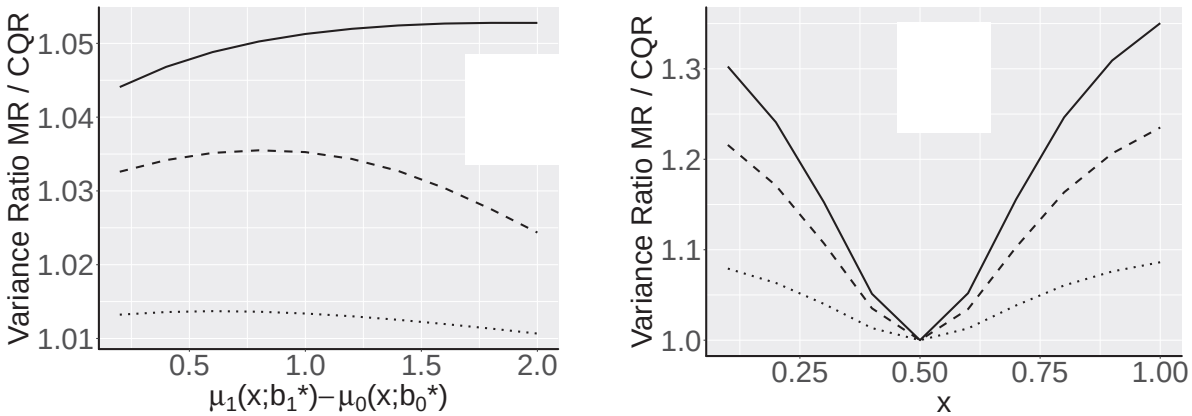

Figure 11: Ratios of the asymptotic variances of $ROC_{\mathbf{x}}(u)$ [maximized over $u \in (0,1)$] based on median regression (MR) and composite quantile regression (CQR) according to Theorem 2 in [10] for a single covariate with domain $\mathcal{X} = [0,1]$ and different error distributions (N – Normal, CN – Contaminated Normal, T – t-distribution, cf. setup).

the Gaussian distribution) as proved in the supplement [46]. Apart from these statistical aspects, CQR is also computationally more involved than median regression. In light of the above considerations, the use of the latter is proposed in this project, complemented by suitable additions to incorporate 1) the ordering constraint (1) discussed in the previous section, and 2) possible heteroscedasticity. Specifically, our approach is based on two-fold median regression, the first of which is used to estimate the location functions. Following He's method [17], the second median regression is employed to estimate the scale functions by regressing the absolute residuals of the preceding median regression on the given covariates. Subsequently, the base CDFs G_0 and G_1 are estimated by the corresponding empirical CDFs of the resulting rescaled residuals. In combination, this scheme yields estimates of all components in the expression for the covariate-specific ROC. In the sequel, we provide more detailed accounts of the individual steps.

For what follows, we suppose that we are given a sample of N triplets $\{(T_i, D_i, \mathbf{x}_i)\}_{i=1}^N$ each consisting of a continuous score, a $\{0,1\}$-valued status indicator, and covariates. Specifically, it assumed that $T_i|D_i, \mathbf{x}_i$, $1 \leq i \leq N$, are independent random variables distributed according to the location-scale model. Without loss of generality, we assume that the triplets are ordered by the value of their status, i.e., $D_i = 0$ for $i = 1, \ldots, n = |\{1 \leq i \leq N : \ D_i = 0\}|$,

and accordingly $D_i = 1$ for the remaining indices $i = n+1, \ldots, n+m = N$.

Stage I: Solve the median regression problem $\min_{\beta \in \mathbb{R}^d} \ \mathbf{y} - \mathbf{X}\beta \ _1$ subject to $\mathbf{A}\beta \geq \mathbf{0}$, where $\mathbf{y} = (T_i)_{i=1}^N$ and $\mathbf{X}$ is an $N \times d$ matrix, $d = 2(p+1)$, whose i-th row is given by $\mathbf{X}_{i\bullet} = [1 \ \ \mathbf{x}_i \ \ D_i \ \ (\mathbf{x}_i \cdot D_i) \ ]$ containing an intercept, status indicator, covariates, and interaction terms between the previous two, $1 \leq i \leq N$. The (system of) linear inequality constraints expressed by $\mathbf{A}\beta \geq \mathbf{0}$ serve as proxy for the ordering constraint imposed on the two location functions associated with the status indicator, i.e., $\mu_1(\mathbf{x}; b_1^*) \geq \mu_0(\mathbf{x}; b_0^*)$ for $\mathbf{x} \in \mathcal{X}$ in Eq. (1). While the constraint (1) is linear in the parameters for any fixed $\mathbf{x}$, the set $\mathcal{X}$ could be complex[1], and as a result, would render the constraint difficult to implement. There are two possible proxies that yield reductions to linear inequality constraints (cf. Figure 12 for illustrations):

(1) *Outer approximation*: suppose that $[l_j, u_j]$ are known upper and lower bounds for the j-th covariate, $1 \leq j \leq p$. 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 (1) holds if it holds for the vertices $\{v \ \}_{\ell=1}^q = \{l_1, u_1\} \times \ldots \times \{l_p, u_p\}$ of $\overline{\mathcal{X}}$, where $q = 2^p$. Accordingly, the ℓ-th row of $\mathbf{A}$ is given by $[\mathbf{0}_{p+1} \ \ 1 \ \ v \ \ ]$, $\ell = 1, \ldots, q$.

(2) *Inner approximation*: the constraint is imposed for the observed covariates $\{\mathbf{x}_i\}_{i=1}^N$ or a suitable subset thereof. Accordingly, each row of $\mathbf{A}$ is of the form $[\mathbf{0}_{p+1} \ \ 1 \ \ \mathbf{x}_i \ ]$. For more examples and a discussion of the merits of the two approximation schemes, we refer to [47].

Stage II: Let β^{M} denote the minimizer, and let further $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta^{\mathrm{M}}$ denote the resulting residuals. In view of the location-scale model according to (2), the median of $|T - \mu_j(\mathbf{x}; b_j^*)|$ given $D = j, X = \mathbf{x}$ equals $\sigma_j(\mathbf{x}; a_j^*) = \Psi(\mathbf{x}) \ \ a_j^* + a_{0j}^*$, $j \in \{0, 1\}$. This suggests that the scales can be inferred from median regression of the absolute values of the residuals $|\mathbf{r}| = (|r_i|)_{i=1}^N$ on $\mathbf{Z}$ whose rows are given by $[1 \ \ \mathbf{z}_i \ \ D_i \ \ (\mathbf{z}_i \cdot D_i) \ ]$ with $\mathbf{z}_i = \Psi(\mathbf{x}_i)$, $1 \leq i \leq N$.

[1] By "complex", we here mean difficult to characterize in terms of computationally tractable (convex) constraints [3].

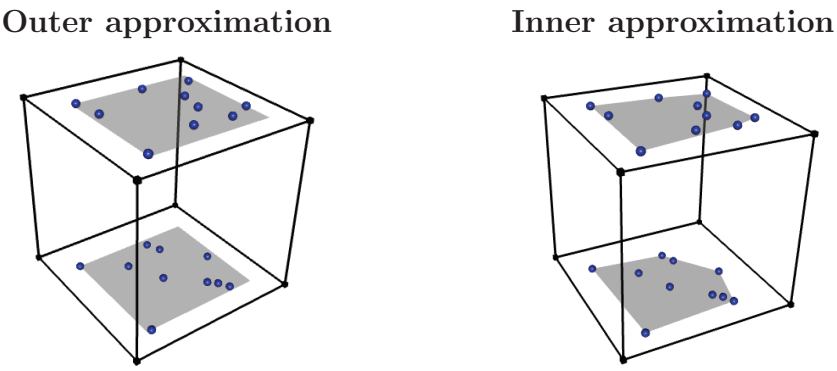

Figure 12: Illustration of the two constraint set approximation schemes described in the text. The dots represent the observed covariate values $\{\mathbf{x}_i\}_{i=1}^N$, here taking values in $[0,1] \times [0,1] \times \{0,1\}$, corresponding to two continuous covariates and one binary covariate. The gray-shaded areas depict the outer (left) and inner (right) approximations.

Specifically, we solve the following second median regression problem:

$$\min_{\gamma \in \mathbb{R}^{2(s+1)}} \| |\mathbf{r}| - \mathbf{Z}\gamma \|_1, \tag{3}$$

where s denotes the dimension of the $\{\mathbf{z}_i\}_{i=1}^N$. This approach for estimating scale functions in the presence of heteroscedasticity originates in the quantile regression literature [17], and will henceforth be referred to as "He's method"[2].

Stage III: Given the output from the previous two stages, we obtain the standardized residuals

$$e_{0i} = \frac{r_i}{\sigma_0(\mathbf{x}_i)} = \frac{T_i - \beta_0^{\mathrm{M}} - \mathbf{x}_i\, \beta_X^{\mathrm{M}}}{\gamma_0^{\mathrm{M}} + \mathbf{z}_i\, \gamma_X^{\mathrm{M}}}, \quad i = 1, \ldots, n,$$
$$e_{1i} = \frac{r_i}{\sigma_1(\mathbf{x}_i)} = \frac{T_i - \beta_0^{\mathrm{M}} - \beta_D^{\mathrm{M}} - \mathbf{x}_i\, (\beta_X^{\mathrm{M}} + \beta_{XD}^{\mathrm{M}})}{\gamma_0^{\mathrm{M}} + \gamma_D^{\mathrm{M}} + \mathbf{z}_i\, (\gamma_X^{\mathrm{M}} + \gamma_{XD}^{\mathrm{M}})}, \quad i = (n+1), \ldots, N,$$

where γ^{M} denotes the minimizer of (3). The standardized residuals serve as proxy for the centered and scaled errors $\{e_{0i}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} G_0$ and $\{e_{1i}\}_{i=n+1}^N \overset{\text{i.i.d.}}{\sim} G_1$ according to the location-scale model for the scores $\{T_i\}_{i=1}^N$. Let G_0 and G_1 denote the empirical CDFs of the $\{e_{0i}\}_{i=1}^n$

[2]We note that the approach does not explicitly enforce non-negative fitted values. However, we did not encounter any instances with negative fitted values neither with simulated nor real data.

and $\{e_{1i}\}_{i=n+1}^{N}$, respectively. Finally, the covariate-specific ROC for $X = \mathbf{x}$ is estimated as

$$\mathrm{ROC}_{\mathbf{x}}(u) = 1 - G_1 \quad \frac{\sigma_0(\mathbf{x})}{\sigma_1(\mathbf{x})} \quad G_0^{-1}(1-u) - \frac{\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})}{\sigma_0(\mathbf{x})} \quad , \quad u \in (0,1),$$

$$\text{where} \quad \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) = \beta_D^{\mathrm{M}} + \mathbf{x} \;\; \beta_{XD}^{\mathrm{M}}, \qquad \sigma_0(\mathbf{x}) = \gamma_0^{\mathrm{M}} + \Psi(\mathbf{x}) \;\; \gamma_X^{\mathrm{M}},$$

$$\sigma_1(\mathbf{x}) = \sigma_0(\mathbf{x}) + \gamma_D^{\mathrm{M}} + \Psi(\mathbf{x}) \;\; \gamma_{XD}^{\mathrm{M}}.$$

Modifications

We here outline two modifications of the approach outlined in the preceding subsection. The first modification integrates ordering constraints into CQR and represents a possible alternative to Stage I above in homoscedastic settings. The second modification concerns non-linear modeling of covariates based on basis functions.

Order-constrained CQR. We here briefly sketch how the the integration of the order constraints in Stage I can be accomplished in a similar way when using the DZ method based on CQR. For $\tau \in (0,1)$, consider the so-called check loss $\rho_\tau : \mathbb{R} \to \mathbb{R}_+$ defined by $u \mapsto \rho_\tau(u) = u\{\tau - I(u \leq 0)\}$, and let further $\tau_k = k/(K+1)$ for $K \geq 1$ odd. A CQR-based counterpart to the constrained median regression problem is then given by the formulation

$$\min \sum_{k=1}^{K} \sum_{i=1}^{N} \rho_{\tau_k}(T_i - \beta_{0k} - \mathbf{x}_i \;\beta_X - D_i \cdot \beta_{Dk} - D_i \cdot \mathbf{x}_i \;\beta_{XD}) \tag{4}$$

$$\text{subject to } \mathbf{A}\beta_{(K+1)/2} \geq \mathbf{0}, \quad \text{where } \beta_k = (\beta_{0k} \;\; \beta_X \;\; \beta_{Dk} \;\; \beta_{XD}) \;\; , \quad 1 \leq k \leq K. \tag{5}$$

The above optimization problem can be expressed as a linear program [22]. Note that by imposing the constraints $\mathbf{A}\beta_k \geq \mathbf{0}$ for all $1 \leq k \leq K$, median ordering as represented by (5) can be strengthened further to incorporate ordering for all quantiles under consideration, i.e., $Q(\tau_k|D=1, X=\mathbf{x}) \geq Q(\tau_k|D=0, X=\mathbf{x})$ for all $1 \leq k \leq K$ given the homoscedastic model that underlies CQR. approach in conjunction with the expanded sets of constraints

will be referred to as "Order-constrained CQR".

Non-linear modeling of covariates. We here provide an outline showing how the order-constrained median regression formulation in Stage I in the previous subsection can be extended to allow for non-linear covariate effects. To keep the exposition simple, we confine ourselves to a single continuous covariate with domain $\mathcal{X} = [0, 1]$, and present technical details in the supplement [46]. The basic idea is to expand the location functions μ_0 and μ_1 in a suitable set of basis functions (in the sequel, we use cubic B-splines given their popularity, but this choice is not essential), employ a roughness penalty to enforce smoothness, and approximate the order constraint $\mu_1(x) \geq \mu_0(x)$, $x \in \mathcal{X}$, by imposing it over a finite grid of points $\mathcal{X} \subset \mathcal{X}$. With a quadratic penalty as typically used for splines, the smoothing parameter can be chosen efficiently via generalized cross validation-type criteria, building on ideas in [38] and [41]. Specifically, letting $\{h_\ell\}_{\ell=1}^{L}$ denote the basis functions, the following optimization problem is solved:

$$\min_{\substack{\mathbf{b}_0=(b_{0\ell}) \\ \mathbf{b}_1=(b_{1\ell})}} \left\{ \sum_{i:\, D_i=0} \left| T_i - \sum_{\ell=1}^{L} h_\ell(x_i) b_{0\ell} \right| + \sum_{i:\, D_i=1} \left| T_i - \sum_{\ell=1}^{L} h_\ell(x_i) b_{1\ell} \right| + \lambda(\mathbf{b}_0' \mathbf{P} \mathbf{b}_0 + \mathbf{b}_1' \mathbf{P} \mathbf{b}_1) \right\}$$
$$\text{subject to } \sum_{\ell=1}^{L} (b_{1\ell} - b_{0\ell}) h_\ell(x) \geq 0, \quad x \in \mathcal{X},$$

where $\mathbf{P}$ is a symmetric positive semidefinite matrix representing a roughness penalty such as the integrated squared derivatives of the associated basis function expansions (cf., e.g., [13]), $\lambda > 0$ is a smoothing parameter, and the constraint serves as proxy for the order constraint under consideration. The above optimization problem reduces to quadratic programming and is straightforward to solve via modern optimization techniques.

Applicaiton to biometric data

In this section, we present an application of our framework to biometric data including fingerprint data in the FBI Biometric Collection [18] and the Face Recognition Vendor Test [31]. The fingerprint data set is a subset of the FBI Biometric Collection of People (BioCoP)

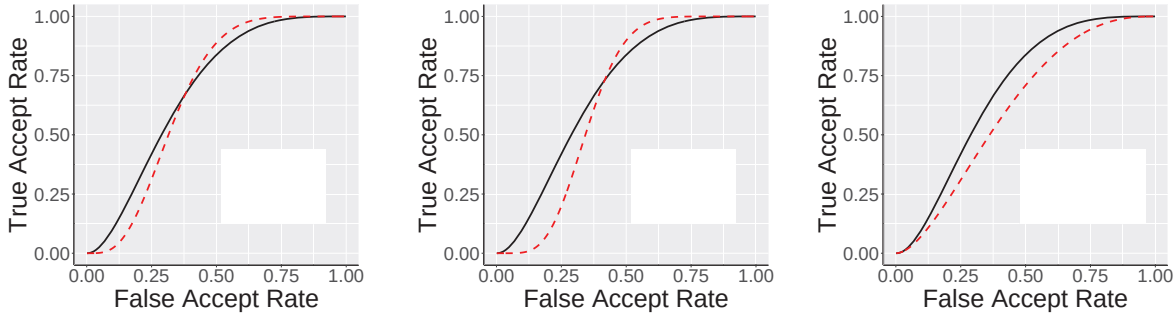

Figure 13: Age-stratified Binormal and pooled ROC curves for the fingerprint dataset. Stratification by age is based on the age of the query subject.

Next Generation Identification Phase 1 between 2008 and 2009 [18]. Data collection involved the acquisition of latent and exemplar fingerprints. The latent fingerprints were friction ridge impressions deposited on common materials. Higher-quality exemplar fingerprints were acquired under controlled conditions using standard ink and paper methods. The comparison scores were generated by comparing latent prints to exemplar prints. The matching scores were obtained using the end-to-end latent fingerprint search system recently published by [5]. The algorithm does include automated ridge structure cropping, latent image pre-processing, feature extraction, feature comparison, and outputs a candidate list. The underlying model is robust to poor quality latent fingerprints since it generates a set of virtual minutiae to construct a texture template. The resulting data set contains $n = 193$ genuine pairs and $m = 40,304$ imposter pairs.

In Figure 13, we display binormal ROC curves fitted to different age strata (determined by the age of the query subject only) in the fingerprint dataset under consideration. The "pooled" ROC curve based on the entire data serves as a reference. Figure 13 indicates that the age-stratified ROC curves are visibly different across strata, and it is therefore appropriate to consider covariate-specific curves, with (logarithm of) age being one of the covariates. In addition, we also use gender (binary, with "1" representing "male") and race (a binary indicator for individuals who identify as "white") as covariates. We also considered pairwise interactions of these two variables; after model selection based on the AIC, the

interaction between age and race was dropped. Specifically, we propose the following linear model.

$$\begin{aligned} T_{ij} =& \beta_0 + \beta_1 D_{ij} + \beta_2 \cdot \mathtt{A}_i + \beta_3 \cdot \mathtt{A}_j + \beta_4 \cdot \mathtt{G}_i + \beta_5 \cdot \mathtt{G}_j + \beta_6 \cdot \mathtt{R}_i + \beta_7 \cdot \mathtt{R}_j \\ &+ \beta_8(\mathtt{A}_i \cdot \mathtt{G}_i) + \beta_9(\mathtt{R}_i \cdot \mathtt{G}_i) + \beta_{10}(\mathtt{A}_j \cdot \mathtt{G}_j) + \beta_{11}(\mathtt{R}_j \cdot \mathtt{G}_j) + \beta_{12}(\mathtt{A}_i \cdot D_{ij}) \\ &+ \beta_{13}(\mathtt{G}_i \cdot D_{ij}) + \beta_{14}(\mathtt{R}_i \cdot D_{ij}) + \beta_{15}(\mathtt{A}_i \cdot \mathtt{G}_i \cdot D_{ij}) + \beta_{16}(\mathtt{R}_i \cdot \mathtt{G}_i \cdot D_{ij}) \\ &+ D_{ij}\epsilon_{ij1} + (1 - D_{ij})\epsilon_{ij0}. \end{aligned} \tag{6}$$

Here, we have $D_{ij} = 1$ when the subject belongs to the genuine group and $D_{ij} = 0$ otherwise. The symbols $\mathtt{A}$, $\mathtt{G}$, and $\mathtt{R}$ represent the logarithm of age, gender, and race, respectively. In the above equation, the index i equals the ID of the query subject, and the index j equals the index of the gallery subject. Different regression coefficients are assumed in the genuine and in the imposter group, hence interaction terms with D are included. Note that only interactions for terms associated with the query subject are needed, e.g., $\mathtt{A}_i \cdot D_{ij}$ but not $\mathtt{A}_j \cdot D_{ij}$ (since if $D_{ij} = 1$ we must have $\mathtt{A}_i = \mathtt{A}_j$ and in turn $\mathtt{A}_i \cdot D_{ij} = \mathtt{A}_j \cdot D_{ij}$). The random error terms $\{\epsilon_{ij0}\}$ and $\{\epsilon_{ij1}\}$ are supposed to be i.i.d. (in particular, independent of all covariates), i.e., we consider a homoscedastic model. Examination of the residuals from the regression model in (6) did not reveal a departure from homoscedasticity.

We compare the methods for estimating covariate-specific ROC curves based on model (6). Since genuine scores tend to be larger than imposter scores, specific attention is paid to the use of the order constraints for MR and CQR[3], respectively, and their effectiveness in stabilizing estimation (i.e., achieving variance reduction) compared to the unconstrained counterparts. Specifically, we impose the constraint that the linear predictor for genuine pairs exceeds the linear predictor of the imposter pairs, uniformly in the observed range for age (18 to 73) and all gender-race combinations (cf. supplement [46] the precise form of the

[3]For this analysis, we consider a partial CQR model in which the regression coefficients involving the model terms $\mathtt{G}_i$ and $\mathtt{G}_j$ depend on the quantile τ while the coefficients of all other terms are independent of τ.

corresponding constraint matrix).

Table 4 reports means, standard deviations, and associated variance ratios of the estimated covariate-specific ROC curves for $\mathbf{x} = (\mathtt{A}_i, \mathtt{A}_j, \mathtt{G}_i, \mathtt{G}_j, \mathtt{R}_i, \mathtt{R}_j) = (\log x, \log x, 1, 1, 1, 1)$ and $x \in \{30, 40, 50\}$ over 1000 bootstrap samples drawn from the original data.

Table 4: Mean and standard deviation (SD) of $ROC_x(u)$ for different ages x and false accept rates u for the facial recognition data based on 1k bootstrap iterations. *MR*: median regression; *OMR*: order-constrained median regression; *CQR*: composite quantile regression; *OCQR*: order-constrained composite quantile regression. The column VR contains the ratios of the variance of *MR* relative to one of the three other methods (larger values correspond to improved efficiency relative to *MR*).

x		30			40			50		
u	Method	mean	SE	VR	mean	SE	VR	mean	SE	VR
0.05	MR	0.454	0.088	–	0.542	0.100	–	0.612	0.108	–
	OMR	0.445	0.080	1.21	0.520	0.081	1.54	0.586	0.083	1.69
	CQR	0.455	0.085	1.09	0.540	0.096	1.07	0.608	0.108	1.00
	OCQR	0.449	0.079	1.27	0.523	0.078	1.62	0.587	0.082	1.69
0.1	MR	0.542	0.094	–	0.640	0.087	–	0.700	0.099	–
	OMR	0.528	0.088	1.15	0.618	0.070	1.52	0.680	0.080	1.54
	CQR	0.541	0.089	1.11	0.636	0.087	0.99	0.696	0.100	0.97
	OCQR	0.530	0.085	1.22	0.618	0.071	1.51	0.681	0.081	1.50
0.15	MR	0.609	0.080	–	0.694	0.083	–	0.745	0.095	–
	OMR	0.594	0.075	1.13	0.676	0.068	1.50	0.730	0.078	1.46
	CQR	0.605	0.079	1.02	0.691	0.084	0.97	0.744	0.096	0.97
	OCQR	0.595	0.075	1.13	0.677	0.068	1.47	0.732	0.079	1.45
0.2	MR	0.655	0.072	–	0.729	0.081	–	0.777	0.091	–
	OMR	0.638	0.069	1.08	0.715	0.065	1.52	0.765	0.076	1.44
	CQR	0.651	0.073	0.96	0.727	0.082	0.97	0.777	0.092	0.97
	OCQR	0.640	0.070	1.05	0.717	0.066	1.49	0.768	0.076	1.43
0.25	MR	0.690	0.067	–	0.760	0.077	–	0.803	0.088	–
	OMR	0.675	0.064	1.10	0.747	0.063	1.52	0.794	0.073	1.45
	CQR	0.687	0.068	0.96	0.759	0.079	0.97	0.803	0.089	0.98
	OCQR	0.677	0.064	1.08	0.750	0.063	1.51	0.797	0.073	1.44

Figure 14 shows contour plots of the variance ratios of the estimated covariate-specific curves. This observation is consistent with similar results in our simulations: the variance reduction achieved by the order constraint particularly concerns regions of the covariate domain where less data is observed.

The face recognition data set contains similarity scores of human face pairs along with several covariates including the study subjects' gender and age as well as image quality (a combined

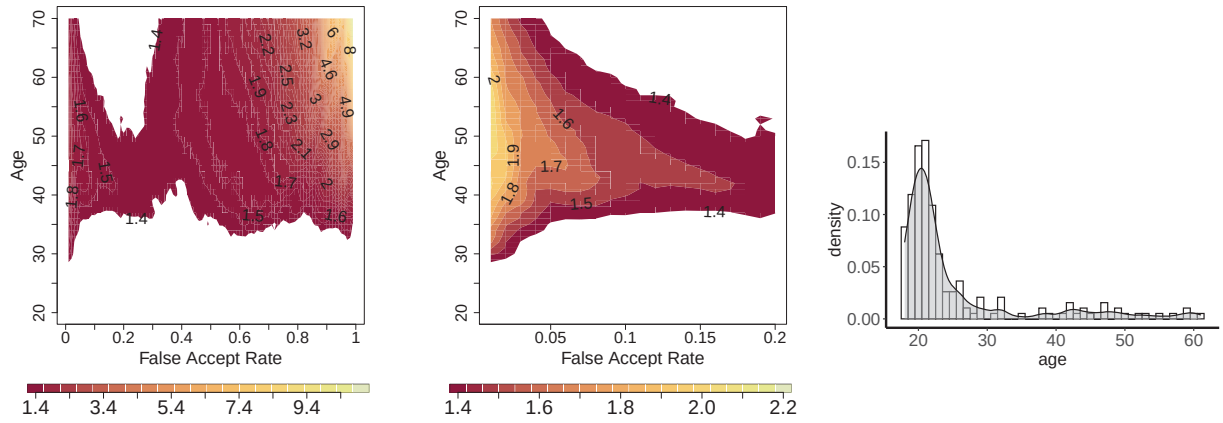

Figure 14: Contour plots of the bootstrap variance ratios of *MR* vs. *OMR* for the estimated covariate-specific ROC in different ranges of FAR (Left: FAR range in $[0, 1]$, and Middle: "zoom-in" for the FAR range $[0, 0.2]$. Contour lines corresponding to relative efficiencies less than 1.4 are not shown; in the white regions, the variance ratios take values in $(0.9, 1.4)$. Right: Histogram and kernel density estimate of the age distribution in the underlying finger print dataset.

rating for each pair of images) according to the categories "good", "bad", and "ugly". The data set has been used for various purposes such as the validation of facial recognition algorithms [25, 42] and the study of the influence of image quality on accuracy [30].

The covariates observed with each matching score are given by age, gender, and image quality. As mentioned in the introduction of this section, the latter takes values according to the three categories "good", "bad", and "ugly". We note that different from the other covariates, the variable image quality is recorded only once for each comparison rather than in terms of a covariate pair. Accordingly, for each image pair, we let $\mathtt{B}_{ij}$ and $\mathtt{U}_{ij}$ denote the two dummy variables associated with the image quality categories "bad" and "ugly", respectively. As in the previous subsection, $\mathtt{A}_i$ and $\mathtt{A}_j$ denote the logarithm of the ages of the query and gallery subject, respectively; $\mathtt{G}_i$ and $\mathtt{G}_j$ denote the corresponding values of gender (binary, with "1" representing "male"). Examination of the residuals resulting from the fit of the regression model given below suggests that a heteroscedastic model depending on image quality is more appropriate than a homoscedastic model. Specifically, the model

under consideration can be expressed as

$$\begin{aligned} T_{ij} =& \beta_0 + \beta_1 D_{ij} + \beta_2 \cdot \mathtt{A}_i + \beta_3 \cdot \mathtt{A}_j + \beta_4 \cdot \mathtt{G}_i + \beta_5 \cdot \mathtt{G}_j + \beta_6 \cdot \mathtt{B}_{ij} + \beta_7 \cdot \mathtt{U}_{ij} \\ &+ \beta_8(\mathtt{A}_i \cdot D_{ij}) + \beta_9(\mathtt{G}_i \cdot D_{ij}) + \beta_{10}(\mathtt{B}_{ij} \cdot D_{ij}) + \beta_{11}(\mathtt{U}_{ij} \cdot D_{ij}) \\ &+ \alpha_1(\mathtt{B}_{ij}, \mathtt{U}_{ij}) D_{ij} \epsilon_{ij1} + \alpha_0(\mathtt{B}_{ij}, \mathtt{U}_{ij})(1 - D_{ij}) \epsilon_{ij0}, \end{aligned} \tag{7}$$

where the $\{\epsilon_{ij0}\}$, $\{\epsilon_{ij1}\}$ are each i.i.d. errors and $\alpha_0(0,0)$, $\alpha_0(1,0)$, $\alpha_0(0,1)$, $\alpha_1(0,0)$, $\alpha_1(1,0)$, $\alpha_1(0,1)$ are non-negative scale coefficients depending on genuine/imposter status (subscript) and the two values for the above dummy variables (parentheses).

Model (7) and the resulting covariate-specific ROC curves are estimated based on our three-stage approach *without* the order constraints; the order constraint is omitted in order to be able to study the impact of the proposed heteroscedastic adjustment alone. For comparison, we also fit the corresponding homoscedastic model in which α_1 and α_0 do not depend on image quality. The corresponding comparison *with* order constraints can be found in the supplement [46].

In order to investigate the sensitivity of the homoscedasticity assumption, we evaluate the differences of the estimated ROC curves under a heteroscedastic and homoscedastic model, respectively. Table 8 lists both relative and absolute differences of the averages values (over 10k bootstrap samples) of the ROC curves for selected values of the FAR (0.05, 0.1, and 0.2) and covariates $\mathbf{x} = (\mathtt{A}_i, \mathtt{A}_j, \mathtt{G}_i, \mathtt{G}_j, \mathtt{B}_{ij}, \mathtt{U}_{ij}) = (\log x, \log x, g, g, b, u)$ with $x = 50$ and $g, b, u \in \{0, 1\}$. Table 8 shows that a homoscedastic model tends to produce larger ROC values. Differences are most pronounced for the image quality categories "good" and "bad", with relative differences exceeding 30% for FAR = 0.05, and are less noticeable for the category "ugly". Figure 16 provides a visual comparison on the logit scale. The plots highlight the differences of the ROC curves for small FAR. While the use of the heteroscedastic approach generic entails higher variability, Table 8 shows that the increase in variance is moderate, with MR^+ having an efficiency of at least .63 relative to MR^+.

Table 5 gives the results of the estimated mean difference for different qualities. We note that the results of WLS and CQR are similar and RRQ has a larger SD than other methods.

Table 5: Bias and SD of the estimated mean difference for the facial recognition data for different qualities. (WLS: weighted least square; RRQ: restricted regression quantiles; CQR: composite quantile regression; GCQR: grouped composite quantile regression.)

quality	good			bad			ugly		
method	bias	SD	VR	bias	SD	VR	bias	SD	VR
WLS	0.14	0.90	1.00	0.13	0.93	1.00	0.01	1.15	1.00
RRQ	0.52	1.21	0.55	0.02	1.27	0.54	1.15	1.44	0.63
CQR	-0.03	0.89	1.02	0.01	0.91	1.04	0.29	1.19	0.93

Table 8 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 WLS_x is moderate.

Table 6: 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
0.1	WLS_x	-0.034	0.119	1.00	-0.033	0.089	1.00	0.037	0.088	1.00
	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
0.3	WLS_x	-0.010	0.050	1.00	0.051	0.121	1.00	0.026	0.098	1.00
	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
0.5	WLS_x	0.009	0.024	1.00	0.011	0.066	1.00	-0.084	0.119	1.00
	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
0.7	WLS_x	0.021	0.019	1.00	0.003	0.058	1.00	-0.022	0.115	1.00
	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
0.9	WLS_x	-0.003	0.008	1.00	-0.033	0.035	1.00	0.024	0.064	1.00
	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

In addition to the values of the ROC curves, we also compare their (approximate) derivatives, which bear a close relationship with likelihood ratios (i.e., the ratio of the density of the scores in the genuine population over the density of the scores in the imposter population). Since the

Table 7: Mean, standard errors (SE), absolute differences (AD) of the mean, relative differences (RD) of the mean, and variance ratio (VR) of the covariate-specific ROC values based on median regression (MR) vs. median regression followed by He's method (MR^+) for $A_i = A_j = \log 20$ based on 10k bootstrap sample.

gender			male					female				
quality	FAR	method	mean	SE	AD	RD	VR	mean	SE	AD	RD	VR
good	0.05	MR	0.644	0.149	0.118	0.22	0.52	0.570	0.167	0.101	0.22	0.67
		MR^+	0.527	0.207				0.469	0.204			
	0.1	MR	0.731	0.126	0.078	0.12	0.58	0.665	0.152	0.068	0.11	0.76
		MR^+	0.653	0.165				0.597	0.175			
	0.2	MR	0.829	0.082	0.046	0.06	0.57	0.787	0.104	0.045	0.06	0.71
		MR^+	0.783	0.109				0.743	0.124			
bad	0.05	MR	0.139	0.067	0.030	0.28	0.49	0.113	0.070	0.025	0.28	0.64
		MR^+	0.109	0.095				0.089	0.088			
	0.1	MR	0.199	0.087	0.027	0.16	0.56	0.161	0.093	0.023	0.16	0.71
		MR^+	0.172	0.117				0.139	0.110			
	0.2	MR	0.337	0.118	0.023	0.08	0.67	0.273	0.127	0.023	0.09	0.77
		MR^+	0.314	0.145				0.249	0.145			
ugly	0.05	MR	0.094	0.050	0.003	0.04	0.48	0.072	0.041	0.004	0.07	0.55
		MR^+	0.091	0.072				0.068	0.055			
	0.1	MR	0.135	0.067	0.009	0.07	0.43	0.104	0.056	0.004	0.04	0.52
		MR^+	0.144	0.102				0.108	0.079			
	0.2	MR	0.236	0.102	0.015	0.06	0.53	0.187	0.086	0.002	0.01	0.61
		MR^+	0.251	0.140				0.189	0.110			

Table 8: Mean, standard errors (SE), absolute differences (AD) of the mean, relative differences (RD) of the mean, and variance ratio (VR) of the covariate-specific ROC values based on median regression (MR) vs. median regression followed by He's method (MR^+) for $A_i = A_j = \log 50$ based on 10k bootstrap sample.

gender			male					female				
quality	FAR	method	mean	SE	AD	RD	VR	mean	SE	AD	RD	VR
good	0.05	MR	0.642	0.009	0.097	0.18	1.01	0.595	0.167	0.095	0.19	1.06
		MR^+	0.545	0.009				0.499	0.009			
	0.1	MR	0.702	0.008	0.055	0.08	1.05	0.661	0.152	0.054	0.09	1.08
		MR^+	0.648	0.008				0.607	0.008			
	0.2	MR	0.787	0.007	0.033	0.04	1.08	0.755	0.104	0.033	0.05	1.10
		MR^+	0.754	0.007				0.722	0.007			
bad	0.05	MR	0.261	0.008	0.066	0.34	1.10	0.213	0.009	0.057	0.36	1.14
		MR^+	0.195	0.007				0.157	0.006			
	0.1	MR	0.326	0.008	0.039	0.14	0.99	0.274	0.008	0.038	0.16	1.02
		MR^+	0.287	0.008				0.237	0.008			
	0.2	MR	0.442	0.009	0.010	0.02	0.93	0.387	0.008	0.014	0.04	0.94
		MR^+	0.431	0.010				0.373	0.009			
ugly	0.05	MR	0.191	0.006	0.018	0.10	0.69	0.149	0.005	0.014	0.09	0.66
		MR^+	0.209	0.008				0.162	0.006			
	0.1	MR	0.249	0.007	0.036	0.15	0.68	0.200	0.006	0.033	0.16	0.63
		MR^+	0.286	0.009				0.233	0.008			
	0.2	MR	0.360	0.009	0.035	0.10	0.77	0.305	0.008	0.034	0.11	0.73
		MR^+	0.395	0.010				0.338	0.009			

estimated ROC curves are not differentiable, its derivatives are approximated via difference quotients. Specifically, we compute the (approximate) log-likelihood ratios (LLRs)

$$\mathrm{LLR}_{\mathbf{x}}(u) = \log \quad \frac{\mathrm{ROC}_{\mathbf{x}}(u+h) - \mathrm{ROC}_{\mathbf{x}}(u-h)}{2h} \qquad u \in (0,1),$$

for $h = 0.01$. In the above equation, u corresponds to the FAR.

The ROC in this function is estimated using covariate-specific ROC curve (average over 10,000 bootstrap iterations) based on the heteroscedastic or homoscedastic fits.

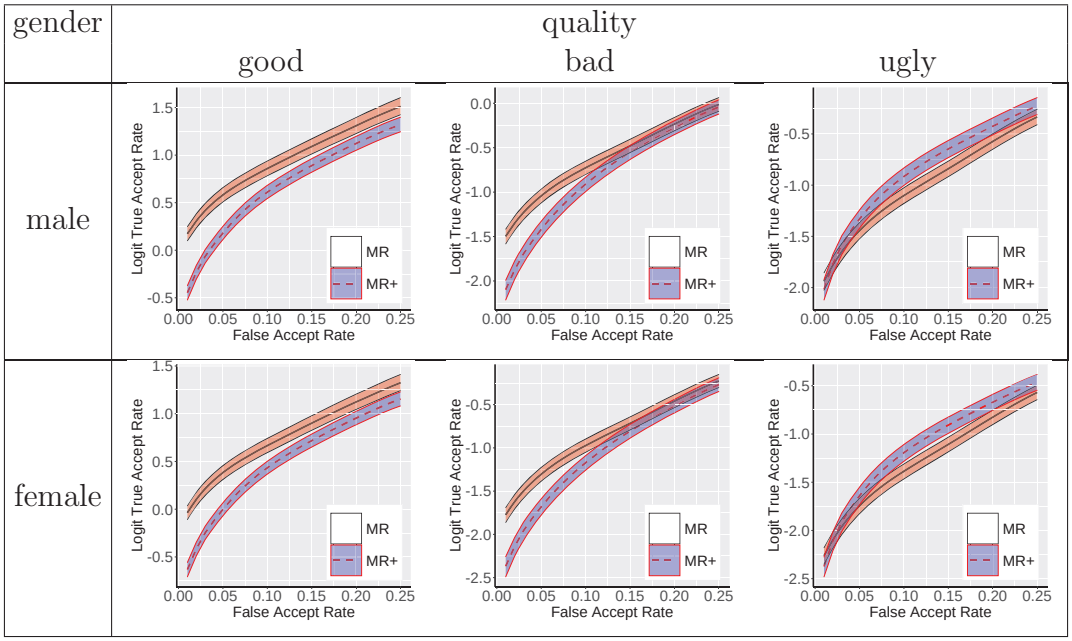

Figure 15: Covariate-specific ROC curves (after applying the logit transformation to the TAR) with 95% pointwise uncertainty intervals (shaded areas) based on *MR* (solid line) vs. MR^{+} (dashed line) for age 50 over 10k bootstrap iterations. *MR*: median regression; MR^{+}: median regression followed by He's method. Best seen in color.

3.2.3 Research Question 3: Develop evidence interpretation tools based on covariate-specific likelihood ratios

In order to access the likelihood ratio, we use the relationship between ROC curve and likelihood ratio In order to access the likelihood ratio, we use the relationship between ROC curve and likelihood ratio. Accordingly, we first calculate the derivative of the covariate-

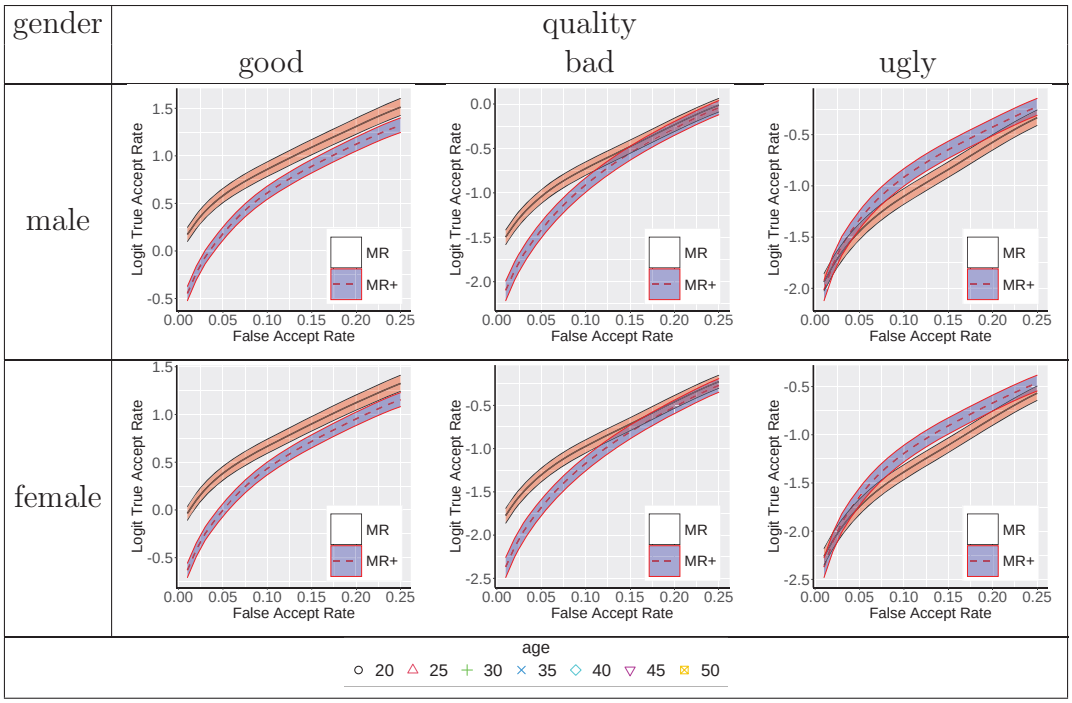

Figure 16: The absolute mean difference for the estimation of covariate-specific ROC curves of *MR* vs. *HM* at age $\{20, 25, 30, 35, 40, 45, 50\}$ over 10k bootstrap iterations. *MR*: median regression – homoscedastic fit; *HM*: He's method – heteroscedastic fit. Best seen in color.

specific ROC-curve:

$$\mathrm{ROC}_x(u) = \frac{\frac{1}{\sigma_1(x;\alpha_1^*)} G_1 \quad \frac{F_{0,x}^{-1}(1-u)-\mu(x,1;\beta^*)}{\sigma_1(x;\alpha_1^*)}}{\frac{1}{\sigma_0(x;\alpha_0^*)} G_0 \quad \frac{F_{0,x}^{-1}(1-u)-\mu(x,0;\beta^*)}{\sigma_0(x;\alpha_0^*)}} = \frac{\frac{1}{\sigma_1(x;\alpha_1^*)} G_1 \quad \frac{S_{0,x}^{-1}(u)-\mu(x,1;\beta^*)}{\sigma_1(x;\alpha_1^*)}}{\frac{1}{\sigma_0(x;\alpha_0^*)} G_0 \quad \frac{S_{0,x}^{-1}(u)-\mu(x,0;\beta^*)}{\sigma_0(x;\alpha_0^*)}}, \; u \in (0,1), \tag{8}$$

where $g_0 = G_0$ and $g_1 = G_1$ denote the PDFs of e_0 and e_1.

Evaluating (8) at $S_{0,x}(t)$, $t \in \mathbb{R}$, we obtain the covariate-specific likelihood ratio conditional on $X = x$:

$$\mathrm{LR}_x(t) = \mathrm{ROC}_x(S_{0,x}(t)) = \frac{\sigma_0(x;\alpha_0^*)}{\sigma_1(x;\alpha_1^*)} \frac{g_1 \quad \frac{t-\mu(x,1;\beta^*)}{\sigma_1(x;\alpha^*)}}{g_0 \quad \frac{t-\mu(x,0;\beta^*)}{\sigma_0(x;\alpha^*)}}, \quad t \in \mathbb{R}. \tag{9}$$

Estimator based on the Location-Scale Model

In particular, the equation (9) suggests the following approach for estimation:

(i) Estimate $\mu(x, 0; \beta^*)$ and $\mu(x, 1; \beta^*)$:

In the simplest setting, X represents a single continuous covariate and $XD = (X_1 \cdot D, \cdots, X_p \cdot D)$, $\mu(X, D, \beta^*)$ is the following linear function in $\beta^* = (\beta_0^*, \beta_D^*, \beta_X^*, \beta_{XD}^*)$

$$\mu(X, D; \beta^*) = \beta_0^* + \beta_D^* D + \beta_X^* X + \beta_{XD}^* XD \tag{10}$$

Define $\epsilon_0 = \sigma_0 e_0$ and $\epsilon_1 = \sigma_1 e_1$, then $T = \mu(X, D; \beta^*) + D\sigma_1(X; \alpha_1^*)e_1 + (1 - D)\sigma_0(x; \sigma_0^*)e_0$ can be written as $T = \mu(X, D; \beta^*) + D\sigma_1 e_1 + (1 - D)\sigma_0 e_0$, substituting $\mu(X, D; \beta^*)$ by $\mu(X, D; \beta)$, we get the residual:

$$\epsilon_i = (1 - D_i) \quad T_i - \mu(x_i, 0; \beta) \quad + D_i \quad T_i - \mu(x_i, 1; \beta) \quad , \ i = 1, \cdots, N, \tag{11}$$

(ii) Estimate $\sigma_0(x, \alpha_0^*)$ and $\sigma_1(x; \alpha_1^*)$:

Without loss of generality, we assume that $D_i = 0$, for $i = 1, \cdots, n_0$; $D_i = 1$, for $i = n_0 + 1, \cdots, N$. Since $\epsilon_0 = \sigma_0 e_0$ and $\epsilon_1 = \sigma_1 e_1$, the corresponding CDFs $\Gamma_0(\cdot) = G_0(\cdot/\sigma_0)$ and $\Gamma_1(\cdot) = G_1(\cdot/\sigma_1)$ are estimated by the following equations, respectively: $\Gamma_0(\cdot) = \frac{1}{n_0 h} \quad \sum_{i=1}^{n_0} K \quad \frac{\cdot - \widehat{\epsilon}_{0i}}{h}$, and $\Gamma_1(\cdot) = \frac{1}{n_1 h} \quad \sum_{i=n_0+1}^{N} K \quad \frac{\cdot - \widehat{\epsilon}_{1i}}{h}$.

In summary, for location-scale model, we have $F_{0,x}$ and $F_{1,x}$, $F_{1,x}(t) = F_1 \quad \frac{t - \mu_1(x)}{\sigma_1(x)}$, and $F_{0,x}(t) = F_0 \quad \frac{t - \mu_0(x)}{\sigma_0(x)}$ Assume $\sigma_1(x) \equiv c_1 > 0$ independent of x, and $\sigma_0(x) \equiv c_0 > 0$ independent of x, then, $F_{1,x}(t) = F_1 \quad \frac{t - \mu_1(x)}{\sigma_1(x)} \quad = G_1(t - \mu_1(x))$, and $F_{0,x}(t) = F_0 \quad \frac{t - \mu_0(x)}{\sigma_0(x)} \quad = G_0(t - \mu_0(x))$. So, the estimate process was estimate μ_1, μ_0 using regression(least squares), compute the residual from regression, estimate Γ_0 and Γ_1 from the (standardized) residuals.

Logistic Regression Model

We estimate the covariate-specific likelihood ratio using other models other than the location-scale model, like the logistic regression model. We compare of location-scale model and the logistic regression model to see whether the logistic regression model is more robust than

the location-scale model to estimate the covariate-specific likelihood ratio.

$$LR_x(t) = \exp \quad h(t) + x^T\Psi - \log\left[\frac{P(D=1|X=x)}{P(D=0|X=x)}\right]\Bigg). \tag{12}$$

where, h is a smooth baseline function, $\Psi \in \mathbb{R}^p$ are regression coefficients associated with the covariates, $X^T\Psi$ is some factor depending on covariate X, $\log \frac{P(D=1|X=x)}{P(D=0|X=x)}$ will henceforth referred to as **"correction term"**. If the probability of $D|X$ is independent of X, $P(D = 1|X = x) = P(D = 0|X = x) = c$, for $c \in (0, 1)$, then $LR_x(t) = e^{h(t)} \cdot e^{x^T\Psi+\log(c)}$. In general, this can be written as $LR_x(t) = e^{h(t)} \cdot e^{x^T\Psi+c'}$, the correction term is constant. We are considering this model because it is a convenient model. There is no interaction between X and t. If we look at the ratio of likelihood ratio $\frac{LR_x(t)}{LR_{x'(t)}}$, x and x represent two values of covariance, this ratio does not depend on t, it only depends on x and x , so for any given value of t the ratio will be the same. This has non-crossing property, which is also one of the major criticism of the Cox model. The non-crossing property has been indicated in the figure 17.

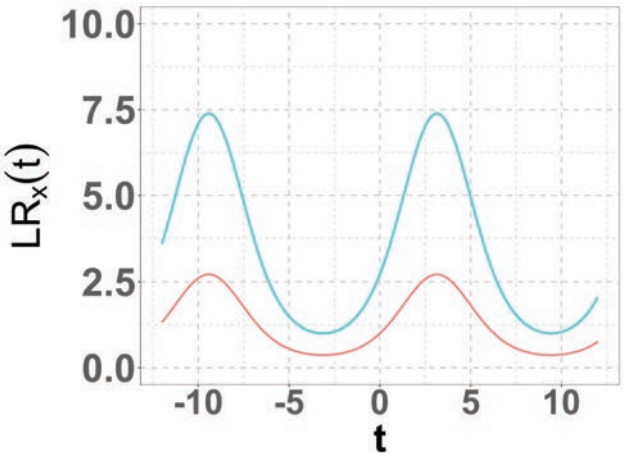

Figure 17: Imposter and genuine scores separated by a threshold based on continuous test scores.

This implies the following logistic regression model:

$$\log\left[\frac{P(D=1|T=t,X=x)}{P(D=0|T=t,X=x)}\right] = h(t) + x^T\Psi =: \eta(t,x). \tag{13}$$

So, $D|T,X \sim Bernoulli\ \ \text{expit}(h(t)+x^T\Psi)\ \ $, where $\text{expit}(h(t)+x^T\Psi) = \frac{e^{h(t)+x^T\Psi}}{1+e^{h(t)+x^T\Psi}}$.

$$P(D=1|X=x) = \frac{e^{h(t)+x^T\Psi}}{1+e^{h(t)+x^T\Psi}} f_{T|X=x}(t)dtdt.$$

Real Data Analysis

We evaluate facial recognition data with covariates age (2004-yob) and gender using the proposed method. The facial recognition dataset from Face Recognition Vendor Test (FRVT) by Phillips [29].

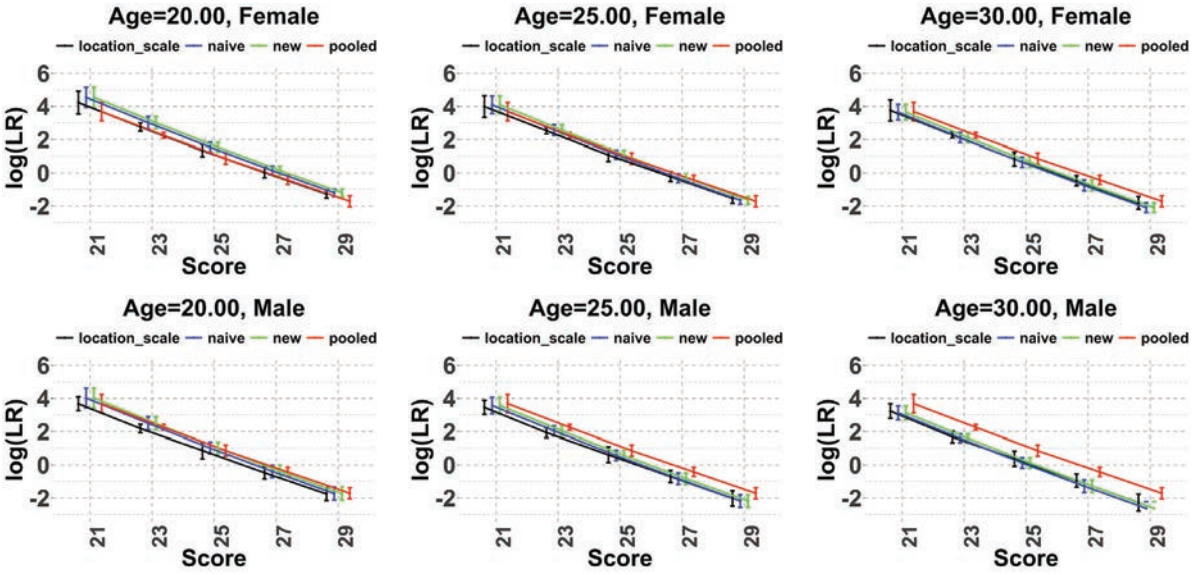

Figure 18: Bootstrap mean and SD of LR based on real data analysis. Results of Covariate-specific LR using location-scale model, naive logistic regression, and generalized linear mixed model and compare with Pooled LR. new is represent the generalized linear mixed model.

In figure 18 we plot the log-likelihood ratio and also add the mean and standard deviations over the 1000 bootstrap replications. In each bootstrap replication, the value of Likelihood ratio is estimated using the location-scale model, naïve logistic regression, generalized linear

mixed model which are considered covariates, and compared with Pooled LR, which is ignore covariates. Age and gender are considered as multiple covariates in our study. We can denote the naive logistic regression as the old method and the generalized linear mixed model as the new method. For the new method, we use glmer function in R to fit the model. Instead of estimating the parameters (β and σ) separately, we estimate the value of $\log(LR)$ directly. The sample sizes in both groups are the same, i.e., $n = m = 1000$.

After fitting the regression models, we have compared the location-scale model, naïve logistic regression, generalized linear mixed model, and the pooled likelihood ratio in figure 18. We have presented three approaches for estimating (score-based) likelihood ratios to account for covariates. These covariate-specific likelihood ratios can be much more appropriate than pooled likelihood ratios. The plot shows that the likelihood ratio can be noticeably influenced by the covariate.

4 Artifacts

4.1 List of products (e.g., publications, conference papers, technologies, websites, databases), including locations of these products on the Internet or in other archives or databases

1. The statistical package for the ROC regression methods was made publicly available as a R Shiny app at https://tynguyen.shinyapps.io/Ordinal_ROC/. User guides are written in plain language so that forensic scientists will be able to implement the developed tools.

2. Zhu, X., Slawski, M., Phillips, P.J. and Tang, L.L., 2021. Order-constrained roc regression with application to facial recognition. *Technometrics*, 63:3, 343-353, https://www.tandfonline.com/doi/abs/10.1080/00401706.2020.1785549.

3. Marasco, E., He, M., Tang, L., Sriram, S. (2021). Accounting for Demographic Differentials in Forensic Error Rate Assessment of Latent Prints via Covariate-Specific ROC Regression. In *Computer Vision and Image Processing: 5th International Conference,*

CVIP 2020, Prayagraj, India, December 4-6, 2020, (pp. 338-350). Springer Singapore. (The paper won the Best Paper Award).

4. Zhu, X., Slawski, M., Phillips, P.J. and Tang, L.L., 2023. A Framework for Covariate-Specific ROC Curve Estimation, with Application to Biometric Recognition. *Annals of Applied Statistics*, Accepted.

5. Hahn, C.A., Tang, L.L., Yates, A.N. and Phillips, P.J., 2022. Forensic facial examiners versus super-recognizers: Evaluating behavior beyond accuracy. *Applied Cognitive Psychology*, 36(6), pp.1209-1218. https://doi.org/10.1002/acp.4003.

6. Marasco, E., He, M., Tang, L. and Sriram, S., 2022. Demographic-Adapted ROC Curve for Assessing Automated Matching of Latent Fingerprints. *Springer Nature Computer Science)*, 3(3), p.190.

7. Marasco, E., He, M., Tang, L. and Tao, Y., 2022, March. Demographic Effects in Latent Fingerprint Matching and their Relation to Image Quality. In *2022 7th International Conference on Machine Learning Technologies (ICMLT)* (pp. 170-179).

8. Simpson, A., Michael, S., Borchet, D., Saunders, C., and Tang, L., 2023, Modeling subpopulations for hierarchically structured data. *Under revision at Statistical Analysis and Data Mining.*

9. Borchet, D., Michael, S., Simpson, A., Saunders, C., and Tang, L., 2023, Effects of prescreening for likelihood ratio approaches in the forensic identification of source problem. *To be submitted to Science and Justice.*

4.2 Data sets generated

Not applicable

4.3 Dissemination activities

Conference Presentations

December 2020 - Dr. Martin Slowski gave an invited talk, titled "Univariate Likelihood

Ratio Estimation via Mixture of Beta Distributions" at 2020 ICSA Applied Statistics Symposium

December 2020 - Dr. Xiaochen Zhu, with Drs Tang and Slawski as co-authors, gave a presentation, titled "Order-Constrained ROC Regression with Application to Facial Recognition" at 2020 ICSA Applied Statistics Symposium

June 2021 - Ty Nguyen, the GRA supported by this grant, gave an oral presentation, titled "Quantifying Uncertainty in Classification Accuracy", at the Crossing Forensic Borders Global Lecture Series.

August 2021 - Dr. Tang gave an oral presentation, titled "Covariate-Adjusted ROC Curves: An Introduction and Application to Characterizing Hidden Behavior in Biometric Matching System" at 2021 Joint Statistical Meetings

August 2021 - Dr. Ommen gave an oral presentation, titled "Machine Learning Methods for Dependent Data Resulting from Forensic Evidence Comparisons" at 2021 Joint Statistical Meetings

August 2021 - Dr. Saunders served as a discussant in the session, titled "Bias and Interpretability in Biometrics for Forensic Science" at 2021 Joint Statistical Meetings

September 2021 - Dr. Saunders gave an oral presentation, titled "The Effect of Latent Structures on Forensic Values of Evidence", at 2021 ICSA Applied Statistics Symposium in September 2021

September 2021 - Ms. He Qi, who is supervised by Dr. Slawski, gave an oral presentation, titled "Approaches to Likelihood Ratio Estimation for Forensic Evidence Interpretation" at 2021 ICSA Applied Statistics Symposium

September 2021 - Ty Nguyen who is supported by this grant, gave an oral presentation, titled "Homogeneity test for ordinal ROC regression and application to facial recognition", at 2021 ICSA Applied Statistics Symposium

September 2021 - Dr. Ommen gave an oral presentation, titled "Constructing Coherent Score-Based Likelihood Ratios that Account for Rarity" at 2021 ICSA Applied Statistics Symposium

Feburary 2022 - Dr. Tang gave an oral presentation, titled "Assessing Error Rates in Multiple Examiner Groups Using Regression Methods" at 2022 Forensic Science Research and Development (RD) Symposium

Feburary 2022 - Dr. Tang and his Ph.D. student, Ngoc-Ty Nguyen, gave an oral presentation, titled "Ordinal Regression for Error Rates in a Black-Box Face Recognition Study" at AAFS

Feburary 2022 - Dr. Tang gave a workshop on "Determining Sufficiency for the Identification of Gasoline" at AAFS in Feburary 2022

March 2022 - Drs. Marasco and Tang's PhD students gave an oral presentation titled "Demographic Effects in Latent Fingerprint Matching and their Relation to Image Quality" at the ACM International Conference on Machine Learning Technologies in March 2022

March 2022 - Dr. Saunders gave an invited keynote presentation titled "An Overview Of The Forensic Identification Of Source Problem" at the XIII COLOQUIO NACIONAL DE ESTADÍSTICA Escuela de Estadística - Facultad de Ciencias

February 2022 - Dylan Borchert and Andrew Simpson presented two poster presentations at the 2022 South Dakota State University Data Science Symposium titled "An Alpha-based Prescreening Methodology for a Common but Unknown Source Likelihood Ratio with Different Subpopulation Structures" and "Identifying Subpopulations of a Hierarchical Structured Data using a Semi-Supervised Mixture Modeling Approach".

August 2022 - Dr. Tang and his student gave an oral presentation, titled "A Survey of Likelihood Ratio Method Development and Implementation Across Multiple Forensic

Disciplines" at Joint Statistical Meetings

August 2022 - Ms. He Qi, who is supervised by Dr. Slawski, gave an oral presentation, titled "Approaches to Likelihood Ratio Estimation for Forensic Evidence Interpretation" at 2022 Joint Statistical Meetings

August 2022 - Dr. Tang and his student gave an oral presentation, titled "Homogeneity Test for Ordinal ROC Regression and Application to Facial Recognition" at Joint Statistical Meetings

August 2022 - Drs. Ommen and Tang gave an oral presentation, titled "Interpretation of Handwriting Evidence Using Error Rates and Score-Based Likelihood Ratios" at Joint Statistical Meetings

Feburary 2023 - Drs. Ommen and Saunders gave an oral presentation, titled " Statistical Discrimination Methods for Forensic Source Interpretation: The Application to Micromorphometric Feature Measurement of Aluminum Powders Used in Explosives" at the American Academy of Forensic Sciences meeting

June 2023 - Dr. Michael gave an oral presentation, titled "Modeling heterogeneity in hierarchically structured data for source identification problems" at the 2023 International Indian Statistical Association Annual Conference

June 2023 - Dylan Borchert gave a poster presentation, titled "A prescreening methodology for the use of likelihood ratios with subpopulation structures in the alternative source population" at the 2023 International Indian Statistical Association Annual Conference

June 2023 - Dr. Michael gave an oral presentation, titled "Detection and Characterization of Subpopulations and the Study of Algorithmic Bias in Forensic Identification of Source Problems" at the 2023 International Conference on Forensic Inference and Statistics

Seminars/Workshops

August 2021 - Dr. Saunders organized a topic-contributed session at 2021 Joint Statistical Meetings for this project

September 2021 - Dr. Slawski organized and chaired an invited session "Advances in Forensic Statistics" at 2021 ICSA Applied Statistics Symposium in September 2021.

References

[1] Colin GG Aitken and David Lucy. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):109–122, 2004.

[2] M. A. Arcones, A. Miguel, H. K. Paul, and J. S. Francisco. Nonparametric estimation of a distribution subject to a stochastic precedence constraint. *Journal of the American Statistical Association*, 97:170–182, 2002.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] Silvia Bozza, Franco Taroni, Raymond Marquis, and Matthieu Schmittbuhl. Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3):329–341, 2008.

[5] Kai Cao, Dinh-Luan Nguyen, Cori Tymoszek, and Anil K Jain. End-to-end latent fingerprint search. *IEEE Transactions on Information Forensics and Security*, 15:880–894, 2019.

[6] C. A. Carolan and J. M. Tebbs. Nonparametric tests for and against likelihood ratio ordering in the two-sample problem. *Biometrika*, 92:159–171, 2005.

[7] Baojiang Chen, Pengfei Li, Jing Qin, and Tao Yu. Using a monotonic density ratio model

to find the asymptotically optimal combination of multiple diagnostic tests. *Journal of the American Statistical Association*, 111(514):861–874, 2016.

[8] Bernard CK Choi. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American Journal of Epidemiology*, 148(11):1127–1132, 1998.

[9] O. Davidov and A. Herman. Ordinal dominance curve based inference for stochastically ordered distributions. *Journal of the Royal Statistical Society, Series B*, 74:825–847, 2012.

[10] Xiaogang Duan and Xiao-Hua Zhou. Composite quantile regression for the receiver operating characteristic curve. *Biometrika*, 100(4):889–900, 2013.

[11] R. Dykstra, S. Kochar, and T. Robertson. Inference for likelihood ratio ordering in the two-sample problem. *Journal of the American Statistical Association*, 90:1034–1040, 1995.

[12] Wenceslao González-Manteiga, Juan Carlos Pardo-Fernández, and Ingrid van Keilegom. ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics*, 38(1):169–184, 2011.

[13] P. Green and B. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1993.

[14] Wen Gu and Margaret Sullivan Pepe. Estimating the diagnostic likelihood ratio of a continuous marker. *Biostatistics*, 12(1):87–101, 2010.

[15] William C Guenther. Power and sample size for approximate chi-square tests. *The American Statistician*, 31(2):83–85, 1977.

[16] Z. Govindarajulu Haynam, G.E. and F.C. Leone. *Tables of the Cumulative Non-*

central Chi-Square Distribution. In: Selected Tables in Mathematical Statistics, volume 1. Markham Publishing Co., Chicago, 1970.

[17] Xuming He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997.

[18] Lawrence Hornak, LaRue Williams, Bojan Cukic, Arun Ross, Keith Morris, Jeremy Dawson, Simona Crihalmeanu, Nathan Kalka, and Nicole Kayal. FBI biometric collection of people (biocop) - next generation identification - phase 1 (2008 - 2009). *2008 Biometric Collection Project 08-06-2008 to 12-31-2009 FINAL REPORT*, 2009.

[19] Beom Seuk Hwang and Zhen 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*, 110(511):923–934, 2015.

[20] Yulei Jiang, Charles E Metz, and Robert M Nishikawa. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, 201(3):745–750, 1996.

[21] David Kaye. Statistical evidence of discrimination. *Journal of the American Statistical Association*, 77(380):773–783, 1982.

[22] Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[23] DV Lindley. A problem in forensic science. *Biometrika*, 64(2):207–213, 1977.

[24] National Research Council. Strengthening forensic science in the united states: A path forward. 2009.

[25] Alice J O'Toole, P Jonathon Phillips, Xiaobo An, and Joseph Dunlop. Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, 30(3):169–176, 2012.

[26] J. B. Parker. A statistical treatment of identification problems. *Journal of the Forensic Science Society*, 6(1):33–39, 1966.

[27] S. D. Peddada, G. Dinse, and G. Kissling. Incorporating historical control data when comparing tumor incidence rates. *Journal of the American Statistical Association*, 102:1212–1220, 2007.

[28] Margaret Sullivan Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

[29] P Jonathon Phillips. A cross benchmark assessment of a deep convolutional neural network for face recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 705–710. IEEE, 2017.

[30] P Jonathon Phillips, J Ross Beveridge, Bruce A Draper, Geof Givens, Alice J O'Toole, David Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada, and Samuel Weimer. The good, the bad, and the ugly face challenge problem. *Image and Vision Computing*, 30(3):177–185, 2012.

[31] P Jonathon Phillips, Kevin W Boyer, Alice J Flynn, Patrick J O'Toole, Cathy L Schott, W Todd Scruggs, and Matthew Sharpe. Face recognition vendor test (FRVT) 2006 and iris challenge evaluation (ICE) 2006 large-scale results. *NIST Interagency/Internal Report*, 2007.

[32] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O'Toole. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[33] President's Council of Advisors on Science and Technology. Forensic science in criminal

courts: Ensuring scientific validity of feature-comparison methods. 2016.

[34] KM Lal Saxena and Khursheed Alam. Estimation of the non-centrality parameter of a chi squared distribution. *The Annals of Statistics*, pages 1012–1016, 1982.

[35] Chuan-Fa Tang, Dewei Wang, and Joshua M Tebbs. Nonparametric goodness-of-fit tests for uniform stochastic ordering. *Annals of Statistics*, 45(6):2565–2589, 2017.

[36] Bradford T Ulery, R Austin Hicklin, JoAnn Buscaglia, and Maria Antonia Roberts. Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences*, 108(19):7733–7738, 2011.

[37] Ted Westling, Kevin J Downes, and Dylan S Small. Nonparametric maximum likelihood estimation under a likelihood ratio order. *arXiv preprint arXiv:1904.12321*, 2019.

[38] S. Wood. Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, 15(5):1126–1133, 1994.

[39] Soweon Yoon and Anil K Jain. Longitudinal study of fingerprint recognition. *Proceedings of the National Academy of Sciences*, 112(28):8555–8560, 2015.

[40] Tao Yu, Pengfei Li, and Jing Qin. Density estimation in the two-sample problem with likelihood ratio ordering. *Biometrika*, 104(1):141–152, 2017.

[41] M. Yuan. Gacv for quantile smoothing splines. *Computational Statistics & Data analysis*, 50(3):813–829, 2006.

[42] Hao Zhang, J Ross Beveridge, Bruce A Draper, and P Jonathon Phillips. On the effectiveness of soft biometrics for increasing face verification rates. *Computer Vision and Image Understanding*, 137:50–62, 2015.

[43] Wei Zhang, Larry L. Tang, Qizhai Li, Aiyi Liu, and Mei-Ling Ting Lee. Order-restricted inference for clustered ROC data with application to fingerprint matching accuracy. *Biometrics*, 76(3):863–873, 2020.

[44] Zheng Zhang and Ying Huang. A linear regression framework for the receiver operating characteristic (roc) curve analysis. *Journal of biometrics & biostatistics*, 3(2), 2012.

[45] Xiao-Hua Zhou, Donna K McClish, and Nancy A Obuchowski. *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, 2009.

[46] X. Zhu, M. Slawski, and L. Tang. Supplement to "A Framework for Covariate-Specific ROC Curve Estimation with Application to Biometric Recognition". 2022.

[47] Xiaochen Zhu, Martin Slawski, P Jonathon Phillips, and Liansheng Larry Tang. Order-constrained roc regression with application to facial recognition. *Technometrics*, pages 1–11, 2020.