Document Title: Statistical Infrastructure for the Use of Error Rate Studies in the Interpretation of Forensic Evidence

Author(s): Liansheng (Larry) Tang, Ph.D., Danica Ommen, Ph.D.

Document Number: 304762

Date Received: May 2022

Award Number: 2018-DU-BX-4228

National Institute of Justice

Research and Development in Forensic Science
for Criminal Justice Purposes

Award #: 2018-DU-BX-4228

Statistical Infrastructure for the Use of Error Rate Studies in the Interpretation of Forensic Evidence

Final Research Report

Award Amount: $ 197,669
Project Period: Jan 1st – Dec 31st 2019 & September 21st 2020 – May 31st 2021

Principal Investigators:

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

Dr. Danica Ommen
Assistant Professor
Iowa State University
Phone: (515) 294-8865
Email: dmommen@iastate.edu

Recipient Organization:

Original recipient:
George Mason University
4400 University Drive
Fairfax, VA 22030

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

Signature of Submitting Official: Mary Davis Digitally signed by Mary Davis
Date: 2021.10.27 13:00:50 -04'00'

Table of Contents

1 Project Summary

1.1 Major Goals and Objectives

During this one and a half years' foundation research project, we aimed to relate matching probabilities to the error rates arising in common individuality studies and propose a paradigm for evidence interpretation based on error rate studies. In this research program we worked within the classical paradigm for evidence interpretation based on conditional match probabilities.

For many researchers, a number of recent recommendations concerning the use of error rates in forensic science have shifted the focus of forensic evidence interpretation away from the formal subjective Bayesian approach long advocated by the research community. Some of the concerns related to error rates of forensic science methods are mentioned in the Congressionally mandated 2009 National Academy of Science (NAS) report entitled "Strengthening Forensic Science in the United States: A Path Forward" [29], the 2016 President's Council of Advisors on Science and Technology (PCAST) report entitled "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" [36] and the 2016 NIJ Forensic Science Technology Working Group Operation Requirements[1]. These recommendations have paralleled a string of papers expressing concerns about the formal Bayesian methods and the score-based approaches (see, for example, Iyer and Lund [19]; Morrison [27]). In response to these recommendations, as well as the success of the black box and white box studies in latent print analysis [44], a large number of other forensic disciplines have proposals for similar studies. These types of studies report an average error rate across a population of examiners for a given set of tasks related to identification of source problems. Although this is not the intention, these studies are often used to justify the conclusions that an examiner has made in a specific case. As statisticians focused on the identification of specific source problems, it is our view that it is unclear what these studies imply about a given

[1] https://nij.ojp.gov/library/publications/forensic-science-technology-working-group-operational-requirements

source identification problem.

1.2 Research Questions

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

1. What are the formal sampling and statistical experiments for source and sub-source propositions for questioned document, fingerprints, and facial recognition evidence?

2. Can a paradigm for reasoning about the source of traces based only on error rates used to characterize the performance of automated identification systems be proposed?

3. Can a set of methods for characterizing the uncertainty about estimated error rates be developed?

4. How can the results for the uncertainty associated with error rate based methods for quantifying evidence best be presented?

1.3 Research Design & Methods

1.3.1 Research Question 1: Formalize Sampling Models

The forensic identification of source problem is an inferential analysis to support answering the question of where a collection of forensic evidence originated. (It should be noted that we use the term "identification" in this context to mean something less strong than is typically considered an "identification" in forensic science, i.e. the source of bullet is this gun, to the exclusion of all other guns.) The point of origin may be a person, as is the case for DNA and handwriting evidence, or a specific object or collection of objects, as is the case with firearms and glass evidence [22]. This type of problem is typically of interest to the criminal justice system. In supporting the quest for the answer to this problem, the evidence interpretation expert is expected to summarize the observed evidence relative to two competing propositions, often referred to as the prosecution and defense propositions, for how the evidence was generated [11, 1]. Typically when considering forensic evidence, the forensic scientist is concerned with source or sub-source level propositions or hypotheses, although activity level propositions might be considered in some cases. However, the court system is typically concerned with offense level propositions concerning the guilt or innocence of the defendant (for a detailed description of the hierarchy of propositions, see [7, 22, 11]). The focus of this research project is on a particular class of source-level identification problems.

In the identification of source problems, it is often of interest to determine whether a suspect can be linked to the evidence found at the scene of the crime. In a sense, this type of identification of the source problem is defined relative to a specified source population, and referred to as the identification of the specific source problem [32]. In this research project, we consider sources to be defined as generators or creators of the objects of interest (for example, a person is a generator of DNA and handwriting profiles, and a window is a generator of glass fragments). This means that all the evidential objects considered in a given case can be split into three different subsets:

e_s: Set of objects associated with or generated by a specified source (denote the number of objects by n_s)

e_a: Collection of sets of objects each associated with a source of traces in an alternative source population (denote the number of sources by n_a)

e_u: A set of trace objects that are all from the same unknown source (denote the number of trace objects by n_u)

We are then tasked with summarizing and presenting the evidence in e_s, e_a, and e_u so that a decision maker can decide between two propositions for how the evidence has arisen. The two propositions can be stated as:

H_p: The unknown source evidence e_u and the specific source evidence es 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.

Following Ommen and Saunders [32], the two competing propositions imply two competing statistical sampling models for defining the source of the trace evidence. These two models are expressed in the context of sampling models below.

M_1**:** The traces in e_u are a simple random sample from the population of traces associated with the specified source

M_2**:** The traces in e_u are a simple random sample from a randomly selected source in the alternative source population of sources.

For the identification of specific source problem, the first model implies that e_u has been generated according to the model for the specific source, implying that there are n_s+n_u traces from the specific source. In contrast, the second model implies that e_u has been generated according to the model for the alternative source population, implying that there are $n_a + 1$ sources sampled from the background population. *In this project, the investigators will build on these propositions, and formalize the underlying sampling and statistical experiments for pattern evidence such as questioned documents and latent fingerprints, see Section 4.2.1 for details.*

A common quantification of the weight of evidence for one proposition over another is the Bayesian likelihood ratio (LR) [23, 12, 24]. This approach is based on the theory that the examiner should report their relative belief concerning how reasonable it is to observe the evidence under two competing propositions for how the evidence has arisen. This statistic is commonly known as the Bayes Factor, or what is often referred to as the likelihood ratio in the field of forensic science. Once the Bayes Factor has been assigned, then the examiner can determine the relative merit of the two models given the information about the evidence by multiplying the Bayes Factor by his/her/their own prior odds. The prior odds is the a priori belief about the relative merit of the two models before the evidence has been observed. When determining probabilistic beliefs under the formal subjective Bayesian approach, it is important that all parties involved exhibit logic and coherence in their reasoning (see Schum [38] for an overview.) This approach to reasoning about evidence has been demonstrated to be exceptionally powerful in evaluating and interpreting forensic evidence for simple DNA and the analytical measures concerning certain types of trace evidence. However, it is unclear how to define an error rate under this paradigm, since all probabilities are personal, and therefore dependent on a variety of factors external to the case at hand.

1.3.2 Research Question 2: Evidence Interpretation via Error Rates

The two-stage approach is the predominant approach to evidence interpretation in the United States and was developed by [21]. As the name suggests, it is based on two steps or stages. The first step considers whether the specific (sometimes called "putative") source can be excluded as the actual source of the trace evidence. If the putative source cannot be excluded in the first step, the examiner is then expected to make an assessment concerning how many alternative sources (in some relevant population) can be excluded as the actual source of the trace evidence for the second step. In the ideal scenario, the second step would exclude all of the other potential sources, leaving the specific source as the only likely source of the trace evidence. This two stage approach to evidence interpretation was formalized by a statistician, J.B. Parker, working in the United Kingdom Atomic Energy Authority in the 1960's based on methods first developed in forensic science by Paul Kirk. These methods have been used throughout forensic science and many formally trained criminalists have training in this form of statistics or evidence interpretation. Under this paradigm, error rates are often characterized by random match probabilities.

To interpret the value of evidence using the two-stage approach, the examiner must first define a scoring rule for comparing control samples from a known source (denoted by e_s) to a trace with an unknown source (denoted by e_u), say $C(e_u, e_s)$. This scoring rule basically establishes a criteria for assessing the similarity of two sets of characteristics. Following Parker [34], assume that $C(\cdot, \cdot)$ is a dissimilarity score, i.e. the bigger the value of $C(\cdot, \cdot)$ the less similar e_u and e_s are. If the value of the comparison $C(e_u, e_s)$ is less than some threshold, say τ, then the source of e_s cannot be excluded as the source of e_u. It is important to note that even if e_u and e_s are realizations of the samples from the same source, they will not typically have the same amount of detail or information concerning their respective source(s). This can be illustrated by latent print examination, where the control print is a full-rolled print taken under controlled conditions and the latent (or recovered print) is a partial, incomplete print with less information. The comparison method $C(\cdot, \cdot)$ will need to account for this differing amount of information and/or complexity.

There are several methods of defining the scoring rule or the comparison metric. One popular method, especially for pattern and impression evidence like fingerprints, is to use an automated identification system to compute the score. Generally speaking, automated identification systems are designed to input a trace, compare it to a finite list of candidate sources, and output a ranking of which source is most likely to have generated the trace sample. Most of the algorithms for comparing the evidential items are black-box and proprietary. Nevertheless, the ranking is built off a score (which in the simplest case could be something like a Bayes classification rule). The outputted scores can then be used as $C(\cdot, \cdot)$ in the two-stage approach.

For the first stage, if the examiner cannot exclude the trace as having arisen from source of the known control samples, then the examiner will state an association exists between the controls and the trace (e.g., "match," "cannot exclude the source as the actual source of the traces," "analytically indistinguishable in all measured properties," etc.), and will proceed to the second stage. For the second stage, the examiner will need to present some measure of the strength of the association. There are various methods for measuring the strength of the association, with one of the more popular being "at what rate would alternative sources (in some specified and hopefully relevant population) not be excluded as the source of the traces?" This rate is commonly referred to as the coincidence probability or the random match probability. There are various other methods that have been suggested as appropriate for measuring the significance of a "match", but we will focus on variations of the random match probability. When two samples provided by different sources are declared to "match" by the comparison methodology, then a false match error has occurred, and the probability of this type of error is the random match probability (RMP). Similarly, when two samples provided by the same source are declared a "non-match", then a false non-match error has occurred, and the probability of this error is the random non-match probability (RNMP). These RMPs and RNMPs are typically unknown quantities, and often need to be estimated from collected data (a recent MS thesis by Fuglsby [14] from SDSU presents algorithms for estimating RMPs and RNMPs for handwriting). To the best of our knowledge, there are three different approaches for conducting the second stage using variations of the random match probability.

The first approach is to estimate of the probability of observing a randomly selected set of control samples from a randomly selected source (in some relevant population of sources) that is sufficiently similar to the trace samples that we would conclude a "match". This approach is referred to as the trace-anchored approach. Then, the error rate of interest is the rate at which we would not exclude sources in the relevant background population, with samples collected in a manner similar to e_s, when compared to the observed trace samples in e_u. This error rate can be estimated by

$$rmp_1(e_u, e_a) = \frac{1}{n_a} \sum_{i=1}^{n_a} I\left(C(e_u, e_{a_i}) < \tau \right),$$

where e_{a_i} denotes a set of control samples from the i^{th} alternative source. Smaller values of this error rate correspond to stronger evidence for associating the trace to the known source.

The second approach is to estimate the probability of observing a randomly selected set of pseudo-trace samples under similar conditions to those under which e_u was generated, from a randomly selected source (in some relevant population of sources), that is sufficiently similar to the specific source's control samples that we would conclude a "match". Pseudo-traces are control samples which mimic the same level of quality and detail to the observed trace samples. This approach is referred to as the source-anchored approach. Now, the error rate of interest is the rate at which we would not exclude pseudo-traces from sources in the alternative source population different from the specific source when compared to the observed samples from the specific source. This error rate can be estimated by

$$rmp_2(e_u, e_a) = \frac{1}{n_a} \sum_{i=1}^{n_a} I\left(C\left(e_s, e_{u_i} \right) < \tau \right),$$

where e_{u_i} denote a pseudo-trace generated from the i^{th} alternative source. Again, smaller values of this error rate correspond to stronger evidence for associating the trace to the known source.

The third, and final, approach is to estimate the probability of observing a "match" when comparing a randomly selected set of pseudo-trace samples, under similar conditions that e_u was generated, from a randomly selected source with control samples from another randomly selected source. This approach is commonly referred to as the general match approach. The error rate of interest here is the rate at which we would not exclude traces (similar to e_u) from one source when compared to control samples (collected in a manner similar to how e_s was collected) from a different source, and is estimated by

$$rmp_3(e_u, e_a) = \frac{1}{n_a\left(n_a - 1\right)} \sum_{i=1}^{n_a} \sum_{j=i} I\left(C\left(e_{a_i}, e_{u_j} \right) < \tau \right),$$

where e_{a_i} denotes a set of control samples and e_{u_i} denotes the pseudo-trace from the i^{th} alternative source. Note that neither the control samples from the specific source, e_s, or the traces with an unknown source, e_u, have been used in defining this error rate. Yet again, smaller values of this error rate correspond to stronger evidence for associating the trace to the known source.

Now that all the relevant forensic error rates have been defined, we need to figure out a good way of presenting them to a fact-finder. The receiver operating characteristic (ROC) curve, which is commonly used in medical diagnostic studies and biometric system evaluation 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) at different possible thresholds. 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 indicates the trade-off between the TPR and FPR under different thresholds. It has many advantages and overcomes the limitation of using isolated measurements of 1-TPR and FPR. The ROC curve is plotted by connecting all the points generated by possible thresholds [49]. *In this project, we propose to develop ROC curves for the Two-Stage approach using the relevant error rates for forensic evidence interpretation in place of the true and false positive error rates, see Section 4.2.2 for details.*

1.3.3 Research Question 3: Uncertainty Quantification

Given its obvious importance, there has been an ongoing debate about how to properly express the forensic value of evidence [28]. Some advocate for the use of a single number (for example [41]), while others advocate for some sort of interval quantification that would provide the decision-maker with an idea of the uncertainty in the analysis (for example [19] and [40]). Many other researchers have provided their opinions on how to deal with uncertainty when quantifying the value of evidence, particularly in a special edition of Science and Justice [45, 4, 5, 8, 9, 26, 33, 42]. In keeping with this theme, *we propose to quantify the uncertainty associated with an SLR system by providing a measure of the variability of the SLRs, see Section 4.2.3 for details.*

1.3.4 Research Question 4: Visualization

Recently, there have been a handful of studies performed to determine how lay-persons (like jurors) perceive numerical and statistical results [43]. In an effort to increase lay-persons' understanding of the strength of forensic evidence, the European Network of Forensic Science Institutes (ENFSI) has recommended the use of a verbal equivalent scale for interpreting the results of likelihood ratios (or Bayes Factors) [11]. The ENFSI likelihood ratio (Bayes Factor) verbal equivalent scale is in seven ordered categories ranging from "no support" to "extremely strong support" of one proposition over the other proposition. However, these likelihood ratio scales are most useful in the Bayesian paradigm, and are not designed to work within the Two-Stage process. Conclusion scales more closely aligned with the Two-Stage approach are not new; forensic document examiners use a 9-point scale for expressing conclusions regarding handwritten documents (ASTM: Standard Terminology for Expressing Conclusion of Forensic Document Examiners), and fingerprint examiners use a 3-point scale.

As an alternative to these scales, graphics and visualizations are often an effective and efficient way to communicate statistical results to both experts and lay audiences. *In this project, we propose to develop visualization methods for ROC curves and forensic error rates, see Section 4.2.4 for details.* We will explore the use of interactive graphics, small multiple charts, and the use of additional graphical features (e.g. color, shading) to link these methods in an intuitive manner. ROC curve visualizations are fairly common, however, the comparison of multiple ROC curves frequently triggers the sine illusion, which affects the perception of differences in the curves [46, 10]. Methods for visualizing the difference in ROC curves will be examined, with the goal of identifying guidelines for the visual comparison of two ROC curves. The methods developed for visualizing probabilistic evidence assessment will initially focus on practitioners, but with the additional goal to develop methods which can be adapted to explain these methods to lay audiences in an intuitive manner.

1.4 Expected Applicability

The 2009 NAS report and 2016 PCAST report state some forensic science disciplines are supported by little rigorous systematic research to validate the discipline's basic premises and techniques and more federal funding is needed to support research in universities and private laboratories committed to such work [29]. The statistical objectives outlined in this project are at the center of investigations into the role of statistics in the evaluation of evidence. These objectives hold great potential for addressing some of the concerns expressed in the

2009 NAS report and 2016 PCAST report. The successful completion of the goals proposed in this project will shift forensic practice paradigms in several important ways. First, formalizing source and sub-source propositions will build the foundation for forensic evidence interpretation. Second, the proposed paradigm based on error rates involves the interpretation of forensic evidence in an exclusion stage and an atypicality stage. The proposed paradigm will be more intuitive for forensic scientists to understand than giving one single number based on the likelihood ratio. Interpreting the forensic evidence in separate stages will also provide the jury more information than the likelihood ratio. Third, interpretation of forensic evidence often involves uncertainty. The methods developed will quantify the uncertainty about estimated error rates. Finally, a set of visualization tools to be developed will present forensic examiners intuitive statistical graphics about the uncertainty associated with error rate based methods for quantifying evidence. The computer codes for visualizing uncertainty of error rates based small sample were made publicly available on investigators' webpages: 1) R codes and the datasets at `https://sites.google.com/view/larrytang/software?authuser=0`, 2) Shiny app at `https://forensicaccuracy.shinyapps.io/order_constrained_ROC_calculation/`. The user guides were written in plain language so that forensic scientists will be able to implement the developed tool.

2 Participants & Collaborators

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 (PI - 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 proposed 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. Susan Vanderplas (University of Nebraska, Lincoln): Dr. Susan Vanderplas spent several years working in industry as a data scientist. Her dissertation research focused on visualization of large data sets, statistical computing, and the perception of statistical graphics. She was responsible for researching and developing methods of graphical and other visual representations for comparing ROC curves and other probability related functions. Dr. Vanderplas used her expertise in statistical graphics to develop visualization methods for ROC curves for error-based methods of evidence interpretation. Additionally, she provided computational support necessary for various aspects of the project.

Dr. Donald Gantz (George Mason University): Dr. Gantz has had many years of successful research in applications to forensic science of pairwise methods of algorithmic data analysis and statistical modeling. He has lectured and published concerning the data analysis and statistical concepts that underpin this research project. Together with statistics professors, postdocs and graduate students in the forensics research group at George Mason founded by Dr. Gantz in 2006, they have published and presented their research internationally.

Ms. Elham Tabassi (National Institute of Standards and Technology): Ms. Tabassi's research area is machine leaning, computer vision and pattern recognition with application in biometrics in general and friction ridge pattern recognition in particular. She is the principal architect of NIST Fingerprint Image Quality (NFIQ) which has become the defacto standard for measuring fingerprint image quality and is currently deployed in some of US Government and EU biometric applications. Her background was important for developing the practical Two-Stage Approach for latent prints.

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

- Ms. Cami Fuglsby (South Dakota State University): advised by Dr. Saunders
- Dr. Xiaochen Zhu (George Mason University): advised by Dr. Larry Tang
- Ms. Mengling He (University of Central Florida): advised by Dr. Larry Tang

3 Changes and Justification

One of the investigators, Dr. John Miller, retired from George Mason University in 2019 and was unable to contribute to the project. Dr. Miller was initially budgeted for implementing the likelihood ratio approach on fingerprint matching scores. This planned activity was carried out by other investigators on the project, Dr. Larry Tang and Dr. Danica Ommen. No replacement personnel were requested for the project. The budgets for the project awards were reallocated to Dr. Larry Tang and Dr. Danica Ommen due to departure of Dr. Miller from the project.

One of the PIs, Dr. Tang, left George Mason University and started his new position at University of Central Florida starting Aug. 12, 2019. The grant was transferred from George Mason University to University of Central Florida. The grant was successfully transferred in November of 2020 and the corresponding sub-awards were successfully set up in December. Due to the change of institution of Dr. Tang, the PIs have spent considerable time reorganizing the grant research and reassessing the current capabilities in early 2019. Then, they requested and received a No-Cost Extension to the proposal. The PIs have prepared a new time-line for the research objectives associated with the grant.

Dr. Saunders requested the change in the travel budget to support the international travel to the SimStat 2019 workshop in Salzburg, Austria. The request was approved. He chaired and organized an invited session on the research associated with this award.

4 Outcomes

4.1 Activities & Accomplishments

During the period of performance, the lead investigators (Tang, Ommen, Saunders) engaged in conference calls via Zoom every other week to update the other participants on research projects and to coordinate and conduct collaborative efforts. Overall, this award resulted in the training of 2 graduate students in the interpretation of forensic evidence, including 1 PhD graduate and 1 MS graduate. This award directly resulted in 1 PhD dissertation, 1 published paper with R codes, a R-Shiny app and 1 submitted paper. For a detailed list of research products and conference presentations, see Section 5.

4.2 Results and Findings

4.2.1 Research Question 1: Formalize Sampling Models

We considered a formal set-up of the identification of source problem for three different evidence types in this project: 1) questioned documents, 2) latent fingerprints, 3) facial recognition.

Questioned Documents

The handwriting dataset we used was collected for NIJ award #2017-DN-BX-0148 lead by PI Mike Caligiuri. The dataset consists of measurements of several kinematic features for writing from 33 individuals and was described in Fuglsby et al. [15]. Following Ommen and Saunders [32], we considered the common source problem formulation. This particular formulation was chosen due to the nature of the data collection as short phrases rather than the subjects writing long, full-paragraph prompts. The common source propositions for handwriting evidence can be stated as

H_p**:** The two questioned documents originate from the same unknown writer.

H_d**:** The two questioned documents originate from two different unknown writers.

The formal sampling models corresponding to the handwriting data deal with the generation of the kinematic features, and are equivalent to the common source sampling models provided in Ommen and Saunders [32]. However, it is not known at this time what the sampling distributions, denoted by F_a and G, for the kinematic data look like. For this reason, we chose to work with a score that captures the Wasserstein distance between kinematic features when developing ROC curves for this evidence (see Ommen et al. [30] for details of the score and see the Sept 2019 presentation by Fuglsby for further details of the ROC curve development).

Latent Fingerprints

The fingerprint data we used in the project came from the National Institute of Standards and Technology Special Database 4 (NIST SD4). The process of comparing fingerprints most often involves comparing a latent fingermark from a crime scene to a full-rolled fingerprint from a suspect. For this reason, we chose to use the specific source development of the sampling models given in Ommen and Saunders [32].

In the fingerprint context, possible sets of propositions include:

1. H_p: the fingermark and fingerprint originate from the same person, vs.

   H_d: the fingermark and fingerprint originate from different persons.

2. H_p: the fingermark and fingerprint originate from the same finger, vs.

   H_d: the fingermark and fingerprint originate from different fingers from the same person.

3. H_p: the fingermark and fingerprint originate from the same finger, vs.

   H_d: the fingermark and fingerprint originate from different fingers from different persons.

The formal sampling models corresponding to the fingerprint data deal with the generation of minutiae, and are equivalent to the specific source sampling models provided in Ommen and Saunders [32]. Similar to the questioned documents example, it is not known at this time what the sampling distributions, denoted by F_s, F_a, and G, for the minutiae look like. For this reason, we chose to work with AFIS scores instead.

Facial Recognition

The facial recognition data that we used in the project came from the publicly available "Good, Bad, and Ugly" dataset. The common source propositions for facial recognition can be stated as

H_p**:** The two facial pictures originate from the same unknown person.

H_d**:** The two facial pictures originate from two different unknown people.

The formal sampling models corresponding to the facial recognition data deal with the generation of facial images, and are equivalent to the common source sampling models provided in Ommen and Saunders [32]. However, it is not known at this time what the sampling distributions, denoted by F_a and G, for facial images look like. For this reason, we chose to work with comparison scores. The comparison scores represent measurement of the characteristic difference, and a smaller distance indicates higher similarity. So a low score represent a pair of pictures with high similarity.

4.2.2 Research Question 2: Evidence Interpretation via Error Rates

In situations where the features are too high-dimensional and complex, the score-based likelihood ratio (SLR) is used to provide some information about the value of evidence. Rather than modelling the original measurements, this approach models "scores" resulting from applying a distance function to the pair (X, Y). The definition of the SLR is

$$SLR(S_{X,Y}) = \frac{Pr(S_{X,Y}|H_p)}{Pr(S_{X,Y}|H_d)},$$

where $S_{X,Y} = S(X_1, ..., X_m, Y_1, ..., Y_n)$ is the (dis)similarity score, a function of X and Y, H_p is the proposition that the pair X and Y come from the same source, and H_d is the proposition that the pair X and Y come from different sources. In contrast to the specific source propositions for the LR, the propositions for the SLR are those for the common source problem [31]. Due to usually small sample sizes of X and Y, the reference population database is valuable for the estimation of the distributions of scores needed to compute the SLR. Typically, this is done by performing all pairwise comparisons of objects in the reference database. The score from the i^{th} pair of objects from the same subject, $T_{p,i}, i = 1, \ldots, M$, has a cumulative distribution function (CDF), F_p, and the probability density function (pdf), f_p. The score from the j^{th} pair of subjects, $T_{d,j}, j = 1, \ldots, N$, has a CDF, F_d, and the pdf,

f_d. Here, the sample sizes for the pairwise comparisons are larger than the original sample sizes.

Evaluating the accuracy of diagnostic biomarkers is important in diagnostic medicine research. In diagnostic medicine, biomarkers are evaluated for their accuracy to distinguish a case who is truly diseased from a control who is not diseased. Diagnostic biomarker results can be binary, ordinal and continuous. Some biomarkers have binary results. Some biomarkers results have some ordered values such as 1, 2, 3, which are called ordinal data [3]. Most of the biomarkers in proteomics and genetics, are on a continuous scale [39]. The receiver operating characteristic (ROC) curve is commonly used to summarize the accuracy of biomarkers with continuous or ordinal outcomes at different chosen thresholds.

The ROC curve indicates the trade-off between the true positive rate (TPR) (i.e. probability of identifying a case when the subject is truly diseased) and the false positive rate (FPR) (i.e. probability of identifying a case when the subject is not diseased). The ROC curve is plotted by connecting all the points generated by a variety of possible thresholds [49]. The ROC curve is also widely used in radiology, psychophysical and medical imaging research for detection performance, military monitoring, and industrial quality control [20]. The ROC curve has many advantages and overcomes the limitation of using isolated measurements of TPR and FPR.

In the mathematical notation, TPR is given by $P(T > c|D = 1)$ and FPR is given by $P(T > c|D = 0)$, where c denotes the threshold, T denotes the biomarker outcome and D is the indicator for disease status with 1 being a case and 0 being a control. A biomarker with 100% TPR and 0% FPR is a perfect predictor.

The commonly used ROC measures are the diagnostic likelihood ratios (not to be confused with the LR weight of evidence defined above), the area under the ROC curve (AUC), the TPR at a fixed FPR, and the partial area under the ROC curve (pAUC). Most ROC curves are concave and above the chance diagonal which is the line segment between $(0, 0)$

and $(1, 1)$. However, some of them are below the chance diagonal and are called improper curves [17]. The AUC between 0.5 and 1 indicates that the diagnostic biomarker has a good performance on detecting the case condition and control condition. The closer the curve is to the left upper corner, the larger the ROC curve area is and the better ability of the diagnostic biomarker has. The perfect biomarker has an AUC of 1.

Order-Restricted ROC Curve Estimation

Consider V classification markers measured on continuous scales to distinguish individuals between diseased and non-diseased groups. In biometric recognition, a classification marker is a matching algorithm used to recognize an individual from others. The diseased and non-diseased observations correspond to genuine and imposter scores, respectively.

Without loss of generality, we assume that the outcome of a classification marker is from the diseased (non-diseased) group if its value is greater (smaller) than a given threshold. Let F_v and G_v be the distribution function of the diseased and non-diseased observations for the vth marker, where $v = 1, \cdots, V$. The ROC curve of the vth marker at a threshold value u is then $R_v(u) = 1 - F_v\{G_v^{-1}(1-u)\}$, where $G_v^{-1}(u) = \inf\{t : G_v(t) \geq u\}$ and $u \in [0, 1]$. The AUC and pAUC over the range $(0, \tau)$ of the vth marker are then $\mathrm{AUC}_v = \int_0^1 [1 - F_v\{G_v^{-1}(1-u)\}]du$ and $\mathrm{pAUC}_v = \int_0^\tau [1 - F_v\{G_v^{-1}(1-u)\}]du$. In various applications, AUC or pAUC are used to compare the performance among markers. As noted earlier, a natural stochastic ordering commonly occurs among observations collected under different conditions, which is particularly evident in fingerprint data. To elaborate, consider a classification marker v with observations Y from the non-diseased group and X from the diseased group. Then Y is said to be stochastically smaller than X, denoted by $Y \leq_{\mathrm{st}} X$, if $F_v(x) \leq G_v(x)$ for $x \in \mathcal{R}$. Our aim is to model $R_v(u)$ as an empirical process, and obtain the estimators for AUC_v and pAUC_v, $v = 1, \cdots, V$, while taking such order constraints into account. The estimators constructed in this way are referred to as order-restricted estimators.

Suppose that the V classification markers are applied to m subjects. Corresponding to each subject i $(= 1, \cdots, m)$ and marker v $(= 1, \cdots, V)$, the observations from F_v and G_v are denoted by $\{X_{vip} : p = 1, \cdots, m_{vi}\}$ and $\{Y_{viq} : q = 1, \cdots, n_{vi}\}$, respectively. For each marker v, the observations $\{X_{vip}, Y_{viq} : p = 1, \cdots, m_{vi}, q = 1, \cdots, n_{vi}\}$ within a subject are clustered. Moreover, between-marker correlation also exists among the observations for different markers. With such a data structure, both within-cluster and between-marker correlations need to be accounted for.

To accommodate such a complex correlation structure, we introduce the weighted ROC curve estimation which assigns different weights to the observations from different clusters. For $v = 1, \cdots, V$, write $m_v = {}_{i=1}^{m} m_{vi}$ and $n_v = {}_{i=1}^{m} n_{vi}$. Let $\{w_{vi}, i = 1, \cdots, m\}$ and $\{w_{vi}, i = 1, \cdots, m\}$ be two sequences of weights satisfying $(1/m_v) \; {}_{i=1}^{m} m_{vi} w_{vi} = 1$ and $(1/n_v) \; {}_{i=1}^{m} n_{vi} w_{vi} = 1$. The weighted ROC curve estimators are established based on the following weighted empirical estimates of distribution functions

$$F_v(x) = \frac{1}{m_v} {}_{i=1}^{m} w_{vi} {}_{p=1}^{m_{vi}} I(X_{vip} \leq x) \;\; \text{and} \;\; G_v(x) = \frac{1}{n_v} {}_{i=1}^{m} w_{vi} {}_{q=1}^{n_{vi}} I(Y_{viq} \leq x). \qquad (1)$$

Here the weights w_{vi} and w_{vi} are used to account for within-cluster correlations. Choices of weights will be discussed in detail later. With F_v and G_v, the weighted ROC curve estimator is then given by $R_v(u) = 1 - F_v\{G_v^{-1}(1 - u)\}$, which subsequently yields an AUC estimator $\text{AUC}_v = {}_0^1 R_v(u)du$ and pAUC estimator $\text{pAUC}_v = {}_0^\tau R_v(u)du$ over $(0, \tau)$.

By incorporating the order restriction, we consider the order-restricted estimators for F_v and G_v defined as $F_v(x) = \min\{F_v(x), Q_v(x)\}$ and $G_v(x) = \max\{G_v(x), Q_v(x)\}$, respectively, where $Q_v(x) = \eta_v F_v(x) + (1 - \eta_v)G_v(x)$ with $0 \leq \eta_v \leq 1$ is an estimator of the distribution function that generates the pooled observations $\{X_{vi1}, \cdots, X_{vim_{vi}}, Y_{vi1}, \cdots, Y_{vin_{vi}}, \; i = 1, \cdots, m\}$. Notice that $F_v(x)$ and $G_v(x)$ are "order-preserving" in the sense that $F_v(x) \leq G_v(x)$ for any $x \in \mathcal{R}$. A natural choice of η_v is the proportion of sample sizes, namely $m_v/(m_v + n_v)$. Certainly, other alternatives can be used, such as the MSE-based weights [48] and proportion of two samples' mean value. In this article, we derive theoretical properties for the order-restricted estimators with a general η_v, provided that $0 \leq \eta_v \leq 1$; in the simulation studies, we use η_v as the proportion of sample sizes.

Subsequently, $R_v(u)$ is estimated by the empirical process $R_v(u) = 1 - F_v\{G_v^{-1}(1-u)\}$, where $0 < u < 1$. The summary statistics of the order-restricted ROC curve, AUC_v and pAUC_v for the vth marker, are given by

$$\text{AUC}_v = \int_0^1 [1 - F_v\{G_v^{-1}(1-u)\}]du \;\text{ and }\; \text{pAUC}_v = \int_0^\tau [1 - F_v\{G_v^{-1}(1-u)\}]du. \tag{2}$$

Proposition 1. The statistic AUC_v is equal to

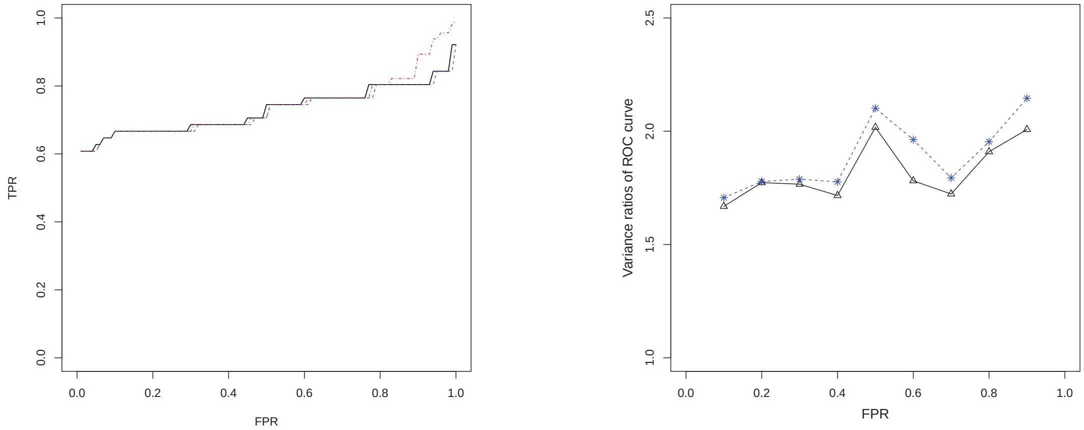
$$1 - \sum_{k=1}^{m_v+n_v} \eta_v(1-\eta_v)A_kB_k, \tag{3}$$

where $A_k = \eta_v F_v(Z_k^{(v)})/(1-\eta_v) + F_v(Z_k^{(v)}) \wedge G_v(Z_k^{(v)})$, and $B_k = \{(1-\eta_v)G_v(Z_k^{(v)})/\eta_v + F_v(Z_k^{(v)}) \vee G_v(Z_k^{(v)})\} - \{(1-\eta_v)G_v(Z_{k-1}^{(v)})/\eta_v + F_v(Z_{k-1}^{(v)}) \vee G_v(Z_{k-1}^{(v)})\}$.

We exemplify our proposed method with the NIST SD4 dataset, which is established to evaluate the accuracy of fingerprint matching algorithms in the NIST Biometric Image Software package [47]. According to Henry classification system [25], fingerprint images can be classified into five coarse-level classes: "Arch", "Left Loop", "Right Loop", "Tented Arch", and "Whorl". The coarse-level classification is mainly used for excluding an individual, and not for identification. For the identification purpose, fingerprint features such as ridge endings and bifurcation provide a finer level classification, which is referred to as a minutiae. One of the widely used minutiae-based matcher is the NIST's Bozorth matcher, which was developed to match minutiae's locations and orientation of two fingerprints, and give a score based on how well they match [47]. The Bozorth matcher was run on all pairs of fingerprints from SD4 database.

It is worth noting that the imposter scores of different subjects can be correlated, since they may be obtained by being matched with the same subject. To maintain independence among scores from different subjects, instead of using all subjects in the sample, we first randomly divided the sample into two groups and chose the first group of subjects as the data sample for analysis. For these selected subjects, only the imposter scores obtained by matching their fingerprints to those of subjects in the second group were considered. According to the ACE-V process of fingerprint recognition [2], we took the maximum values of all imposter scores of each subject as its final imposter score. If the maximum imposter scores of two subjects were obtained by matching to the same subject, then the subject with the largest imposter score is retained. Since each finger has two rolled fingerprints, if the scores of the matched finger which yields the maximum imposter score are very close (both larger than 95% quantile of all imposter scores), we then used both scores. On the other hand, for the genuine score, two types of matching scores were obtained for each subject: matching an image to itself and matching an image to its rolled version. Here we used the latter. As a result, the genuine and imposter scores of different subjects are independent.

To evaluate the discrimination accuracy of the Bozorth matcher, we apply the conventional, weighted empirical, and proposed method to these scores. For illustration, we first focus on the fingerprint matching scores of all female subjects in the "Arch" class. This subgroup includes 102 subjects. The intraclass correlation coefficients for the genuine and imposter scores are estimated as $_1 = 1$ (since each subject has only one genuine score) and $_1 = 0.285$, respectively. The estimated ROC curves by three methods are displayed in Figure 1a. Figure 1b presents the ratios of variance estimates of the conventional and weighted empirical ROC curve estimators to the proposed estimators, calculated based on 1000 Monte Carlo replications. This figure shows that the proposed ROC curves estimators always have smaller variances than the unrestricted estimators, since the ratios are all larger than 1. Moreover, we also compare the performance of the three methods in estimating AUC and pAUC based on the variance and p-values for testing the null hypotheses $H_0 : \text{AUC} = \theta_1$ and $H_0 : \text{pAUC} = \theta_2$, where $\theta_1 \in \{0.6, 0.7, 0.8\}$ and $\theta_2 \in \{0.2, 0.3, 0.4\}$. The results are displayed



(a)

(b)

Figure 1: (a) ROC curve estimator for female fingerprint data: the conventional estimator (black solid line); the weighted empirical estimator (blue dashed line); the proposed estimator (red dotted line); (b) MSE ratios of the conventional and weighted empirical ROC estimators to the proposed estimators.

in Table 1. From this table, we can see that the variances of the proposed AUC and pAUC estimators are much smaller than the corresponding estimators in comparison, and the p-values of the proposed method are always the smallest among the three methods.

Relationship between SLR and ROC Curve

In this section, we explore the relationship between the SLR and the ROC curve. Recently, there has been a lot of debate surrounding the 2016 PCAST report [36]. In the report, there was a lot of attention on "feature-comparison" methods in forensic science. These methods refer to the process by which examiners perform visual comparisons of evidence, such as fingerprints, firearms, footwear, hair, and bite marks. The foremost recommendation of the report was that, in an effort to strengthen the scientific foundations of forensic examinations, there needs to be a comprehensive study of the error rates associated with each one of these forensic fields. To follow this recommendation, it is logical to apply the binary classification techniques associated with the ROC curve to get the associated error rates (TPR, FPR). However, this contradicts other recommendations to compute the LR

Table 1: Performance of the conventional (M_c), weighted empirical (M_w), and the proposed (M_r) method on the estimation and hypothesis testing of AUC and pAUC for female fingerprint data. The left panel is for estimation of the logit transformation of AUC and pAUC; the right panel are the p-values for the null hypothesis AUC $= 0.6, 0.7, 0.8$ and pAUC $= 0.2, 0.3, 0.4$.

		Estimation			Test		
		Estimate	Variance	CI	AUC=0.6	AUC=0.7	AUC=0.8
$\log \frac{\text{AUC}}{1-\text{AUC}}$	M_c	0.982	0.084	[0.964, 1.000]	0.046	0.642	0.162
	M_w	1.085	0.086	[1.076, 1.094]	0.047	0.633	0.173
	M_r	0.987	0.021	[0.969, 1.006]	$< .001$	0.090	0.032
		Estimate	Variance	CI	pAUC=0.2	pAUC=0.3	pAUC=0.4
$\log \frac{\text{pAUC}}{1-\text{pAUC}}$	M_c	-1.023	0.091	[-1.042, -1.004]	0.230	0.560	0.041
	M_w	-1.018	0.092	[-1.038, -0.999]	0.226	0.572	0.043
	M_r	-1.024	0.023	[-1.034, -1.015]	0.017	0.245	$< .001$

(or SLR) and provide its value without going so far as to say which of the two propositions to choose. Any relationship between the SLR and ROC curve will bridge the gap between the binary classification role of the ROC curve (under which error rates are clearly defined) and the "weight of evidence"-style role of the SLR (under which error rates are ambiguously defined, at best). If successful, our method will satisfy both the recommendation of the PCAST report (relying heavily on error rates) as well as the recommendation to compute likelihood ratios.

Denote continuous similarity scores for the ith pair of mated evidence measurements as $T_{p,i}, i = 1, \ldots, M$, which follow a distribution, F_p, and continuous similarity scores for the jth pair of non-mated evidence measurement as $T_{d,j}, j = 1, \ldots, N$, which follow a distribution, F_d. In the forensic context, there are two common types of "error rates," the random match probability (RMP) and the random non-match probability (RNMP). The RMP and RNMP are defined relative to a threshold c; when a score T exceeds c then the pair the produced T is declared a "match" and when T is smaller than c then the pair that produced T is declared a "non-match." Therefore, the RMP is defined as $RMP(c) = P(T_d > c)$ for any non-mated score T_d, and is interpreted as the probability that a non-mated score will be declared a "match" by chance. Similarly, the RNMP is defined as $RNMP(c) = P(T_p < c)$ for any mated score T_p, and is interpreted as the probability that a mated score will be declared a "non-match" by chance.

The ROC curve plots a pair of points $(FPR(c), TPR(c))$, where c is the possible threshold, true positive rate $TPR(c) = 1 - F_p(c)$ and false positive rate $FPR(c) = 1 - F_d(c)$. The $TPR(c)$ is also denoted as a survivor function $TPR(c) = P(T_p > c)$ and $FPR(c)$ is denoted as a survivor function $FPR(c) = P(T_d > c)$. Therefore, in the forensic context the ROC curve plots $1 - RNMP(c)$ against $RMP(c)$ for a variety of thresholds, c. Let u be $FPR(c)$, and let $R(u)$ be $TPR(c)$, and $R(u)$ is given by $R(u) = 1 - F_p(F_d^{-1}(1 - u))$, where u is the false positive rate.

The first derivative of the ROC curve has been shown to be closely related to likelihood ratio (Choi, 1998). Specifically, the tangent at a point, u, of the ROC curve is written as $R\,(u) = F_p(F_d^{-1}(1-u))/F_d(F_d^{-1}(1-u))$. For a realized comparison score $t_{x,y}$ based on evidence measurements, x and y, we write $t_{x,y} = F_d^{-1}(1 - u)$. Since it follows that $u = 1 - F_d(t_{x,y})$, we then have the mathematical relationship between the score-based likelihood ratio and the the tangent at a point u of the ROC curve $SLR(t_{x,y}) = R\,(1 - F_d(t_{x,y}))$. This applies to comparison scores on a continuous scale, which is commonly the case in fingerprint matching. The SLR can be interpreted as the instantaneous change in the true positive rate in a unit change of $1 - F_d(t_{x,y})$.

Estimating the Score-based Likelihood Ratio

The relationship described provides a way to take advantage of both the LR-style and error rate-based approaches to forensic science, provided that the ROC curve can be estimated. Several methods of estimating ROC curves exist, including parametric, nonparametric and semiparametric methods. The parametric methods usually assume parametric distributions for diagnostic similarity scores and yield a smooth ROC curves. The nonparametric ROC methods do not have distribution at requirements. The semiparametric ROC methods could generate smooth ROC curves without distribution assumptions for the similarity scores. We will use parametric methods of estimating the ROC curve for the purpose of deriving SLR values based on the ROC curve. Then, we will compare those approaches to a popular method of obtaining the SLR without the use of an ROC curve.

Parametric ROC Curve Method. In a simple setting, after some monotone transformation, the mated and non-mated scores follow normal distributions $F_p \sim N(\mu_p, \sigma_p^2)$ and $F_d \sim N(\mu_d, \sigma_d^2)$, respectively. Since the mated scores are more likely to be larger than the non-mated scores, we have $\mu_p > \mu_d$. The resulting ROC curve is referred to as the binormal ROC method [13]. Normality of the original scores is checked through quantile-quantile plots for mated and non-mated groups, separately. If the normality assumption is invalid, Zou et al. [50] suggest that before estimating the normal parameters, the Box-Cox power transformation should be used to transform the original score using

$$\psi_{\lambda_1}(T_{p,i}) = \frac{(T_{p,i})^{\lambda_1} - 1}{\lambda_1}, \quad \psi_{\lambda_1}(T_{d,j}) = \frac{(T_{d,j})^{\lambda_1} - 1}{\lambda_1}, \tag{4}$$

where λ_1, λ_2 are the parameters of Box-Cox transformation, $\lambda_1 = 0$ and could be estimated by maximum likelihood estimator. It is worth noting that the monotone transformation should be the same for the two groups so that the underlying ROC curve remains unchanged. This is due to the transformation invariance of the ROC curve. For simplicity, we still use $T_{p,i}$ and $T_{d,j}$ to denote the transformed normal scores. Without loss of generality, we assume that $T_{p,i}$ has a larger mean than $T_{d,j}$. Then, we have a parametric estimate (PE) of the FPR given by $FPR_{PE}(c) = 1 - \Phi((\hat{\mu}_p - c)/\hat{\sigma}_p)$ and a parametric estimate (PE) of the TPR given by $TPR_{PE}(c) = 1 - \Phi((\hat{\mu}_d - c)/\hat{\sigma}_d)$ where the sample means, $\hat{\mu}_p = \bar{T}_p$ and $\hat{\mu}_d = \bar{T}_d$ are estimators for the Normal population means and the sample standard deviations $\hat{\sigma}_p = s_p$ and $\hat{\sigma}_d = s_d$ are estimators for the Normal population standard deviations. The ROC curve is plotted for all possible values of c and is given by $R_{PE}(u) = \Phi(a + b\Phi^{-1}(u))$, where $a = (\mu_p - \mu_d)/\sigma_p$

and $b = \sigma_d/\sigma_p$. The first derivative of the ROC curve is given by

$$R_{PE}(u) = \frac{b\phi(a + b\Phi^{-1}(u))}{\phi(\Phi^{-1}(u)))}.$$

This gives a function of the TPR, u, instead of a function of a score. For a score, $t_{x,y}$, the associated TPR is given by $u = P(T > t_{x,y}|H_d)$, or the probability of having a score greater than the observed score $t_{x,y}$ when the defence hypothesis is true. Hanley and Hajian-Tilaki [18] recognize that $100 \times (1 - u)$ is the percentile of $t_{x,y}$ in the non-mated scores.

Percentiles are commonly used to standardize growth and lung function measurements for children and to standardize many laboratory measures. By substituting u with a placement value $1 - \Phi((t_{x,y} - \mu_d)/\sigma_d)$, we have the SLR using a parametric estimate (PE) under the binormal model

$$SLR_{PE}(t_{x,y}) = \frac{b\phi(a + b\Phi^{-1}(1 - \Phi((t_{x,y} - \mu_d)/\sigma_d))}{\phi(\Phi^{-1}(1 - \Phi((t_{x,y} - \mu_d)/\sigma_d))}. \tag{5}$$

It follows from the symmetry of the standard normal density that the numerator of (5) can be simplified to be $b\phi(a + b\Phi^{-1}(\Phi((-t_{x,y} + \mu_d)/\sigma_d))$, or $b\phi(\mu_p/\sigma_p - t_{x,y}/\sigma_p)$, and the denominator can be simplified to $\phi((\mu_d - t_{x,y})/\sigma_d)$. The $\log SLR_{PE}$ at a score $t_{x,y}$ is then given by

$$\log SLR_{PE}(t_{x,y}) = \log \sigma_d/\sigma_p + \log \phi((\mu_p - t_{x,y})/\sigma_p) - \log \phi((\mu_d - t_{x,y})/\sigma_d). \tag{6}$$

The estimators for a and b are obtained by substituting the sample means, $\hat{\mu}_p = \bar{T}_p$ and $\hat{\mu}_d = \bar{T}_d$ and sample standard deviations, $\hat{\sigma}_p = s_p$ and $\hat{\sigma}_d = s_d$ for the true means and standard deviations: $\hat{a} = (\hat{\mu}_p - \hat{\mu}_d)/\hat{\sigma}_p$ and $\hat{b} = \hat{\sigma}_d/\hat{\sigma}_p$. Then, $\log SLR_{PE}(t_{x,y})$ is estimated by plugging in the corresponding estimates of mean and standard deviation.

The estimated SLR needs the estimators for the mean and variances separately for both groups. Denote the parameter vector $\theta = (\mu_p, \sigma_p, \mu_d, \sigma_d)^T$ and its estimator $\hat{\theta} = (\hat{\mu}_p, \hat{\sigma}_p, \hat{\mu}_d, \hat{\sigma}_d)^T$.

The first order Taylor expansion on the logarithm of the likelihood ratio is written as

$$\log SLR_{PE}(t_{x,y}) = \log SLR_{PE}(t_{x,y}) + \nabla^T \log SLR_{PE}(t_{x,y})(\hat{\theta} - \theta).$$

where

$$\nabla \log SLR_{PE} = (\frac{\partial \log SLR_{PE}}{\partial \mu_p}, \frac{\partial \log SLR_{PE}}{\partial \sigma_p}, \frac{\partial \log SLR_{PE}}{\partial \mu_d}, \frac{\partial \log SLR_{PE}}{\partial \sigma_d})^T$$

and the explicit expression for $\nabla \log SLR_{PE}(t_{x,y})$ is given by

$$\begin{pmatrix} \phi\left((\mu_p - t_{x,y})/\sigma_p\right)/(\sigma_p \phi((\mu_p - t_{x,y})/\sigma_p)) \\ -1/\sigma_p - (\mu_p - t_{x,y})\phi\left((\mu_p - t_{x,y})/\sigma_p\right)/(\sigma_p^2 \phi((\mu_p - t_{x,y})/\sigma_p))) \\ -\phi\left((\mu_d - t_{x,y})/\sigma_d\right)/(\sigma_p \phi((\mu_d - t_{x,y})/\sigma_d)) \\ 1/\sigma_d + (\mu_d - t_{x,y})\phi\left((\mu_d - t_{x,y})/\sigma_d\right)/((\sigma_p^2 \phi((\mu_d - t_{x,y})/\sigma_d))) \end{pmatrix}^T.$$

The variance of $\log SLR_{PE}(t_{x,y})$ is derived from the first order Taylor expansion on the parameter vector (or the multivariate delta method):

$$var(\log SLR_{PE}(t_{x,y})) = \nabla^T \log SLR_{PE}(t_{x,y}) cov(\hat{\theta}) \nabla \log SLR_{PE}(t_{x,y}). \tag{7}$$

The variance and covariance elements in $cov(\hat{\theta})$ follow standard expressions. We have $var(\hat{\mu}_p) = \sigma_p^2/M$, $var(\hat{\mu}_d) = \sigma_d^2/N$. With the normal distributions, the variance formulas are simplified to $var(\hat{\sigma}_p^2) = 2\sigma_p^4/(M-1)$ and $var(\hat{\sigma}_d^2) = 2\sigma_d^4/(N-1)$. The delta method gives the variance expressions for sample standard deviation: $var(\hat{\sigma}_p) = 1/(4\sigma_p^2) var(\sigma^2) = \sigma_p^2/(2(M-1))$, and $var(\hat{\sigma}_d) = \sigma_d^2/(2(N-1))$. With these expressions, the covariance matrix of $\hat{\theta}$ is

$$cov(\hat{\theta}) = \begin{pmatrix} \sigma_p^2/M & 0 & 0 & 0 \\ 0 & \sigma_p^4/2(M-1) & 0 & 0 \\ 0 & 0 & \sigma_d^2/N & 0 \\ 0 & 0 & 0 & \sigma_d^4/2(N-1) \end{pmatrix}$$

Facial Recognition Example

We use a facial recognition data set, and apply PE, logistic regression estimation (LRE) and kernel density estimation (KDE) to investigate the variance, repeatability, and reproducibility of these methods. The biometric images were frontal face images taken with a digital single-lense reflex camera. The similarity scores were extracted from the picture comparison, and used in our study as the score. The data set has three categories, which are "good," "bad," and "ugly," based on the quality of the images [35]. We only consider the category "good" in our study. The comparison scores represent measurement of the characteristic difference, and a smaller distance indicates higher similarity. So a low score represent a pair of pictures with high similarity. Then, a genuine comparison score is measured by comparing two pictures of the same individual, and is generally a smaller value than an imposter score, which is measured by comparing pictures of different people. Scores in both groups have extremely large outliers, so we remove all the outlier samples before applying the methods.

We randomly select 2000 samples from the genuine group, and various numbers of samples from imposter group to vary the log sample size ratios of data from -2 to 2 by the increment of 0.1. We use this as training data for both PE, LRE, and KDE. Then, we calculate the LLR at the score of 25 for all the three methods. For LRE and KDE, we repeat the random selection 1000 times to get the empirical variances and then generate 95% confidence intervals based on the variances. For PE, we use our variance estimate approach given in Equation (7) to get the estimator and also the 95% confidence interval.

Figure 2 shows the LLR values for the PE, KDE, and LRE methods when the sample size ratio varies. We see that both PE and KDE have good repeatability since they are not sensitive to varying sample size ratios. This is similar to our simulation findings which also show good repeatability of the PE and KDE methods. The repeatability of the LRE method is unsatisfactory because the LLR from the method takes on a wide range of values. For reproducibility, the PE method generates larger LLR values than the KDE method. The LLR from the PE and KDE methods takes on all positive values, while the LLR from the LRE method takes on both positive and negative values. If one uses zero as a decision threshold to decide whether the score 25 comes from H_p or H_d, both PE and KDE can arrive at the same conclusion that the score of 25 likely supports H_p with all positive LLR values. With LLR value ranging from negative values to positive values, the decision by LRE depends on the sample size ratio. When the sample size ratio is as small as -2, the LRE concludes that the score of 25 supports H_d, and with the log ratio is as large as 2, the LRE supports H_p instead.

Figure 2 also shows the confidence interval of LLR values for the PE, KDE, and LRE methods when the sample size ratio varies. Note that the confidence interval is increasing as the log sample size ratio increases. That is because we fixed the sample size of mated group and vary numbers of samples in non-mated group to increase the log sample size ratios of data from -2 to 2 by the increment of 0.1, so the total sample size is decreasing. Since the confidence interval is dependent on sample size, this explains why the confidence intervals become wider as sampling ratio increases.

Table 2 gives the ranges of the confidence intervals along with the variance of the LLR for the PE, LRE, and KDE methods applied to facial recognition data. We only select three typical log sampling ratio for each method from the pool, which are -1.0, 0, and 1.0, and list the estimated LLR (LLR), the estimated lower bound (LB) and upper bound (UB) of the confidence intervals, and the variance of the estimated LLR ($Var(LLR)$). Note that the variance and the width of the confidence interval of PE and LRE are similar. The KDE method has the largest variance among all methods. All of the estimated LLR values and confidence intervals are positive except for one LRE result when log sampling ratio is -1.0, which implies that for this situation we will conclude that the score belongs to a group different from other decisions which were made with same score but different methods or different log sampling ratios. Therefore, the LRE method is less reliable than the PE and KDE methods.

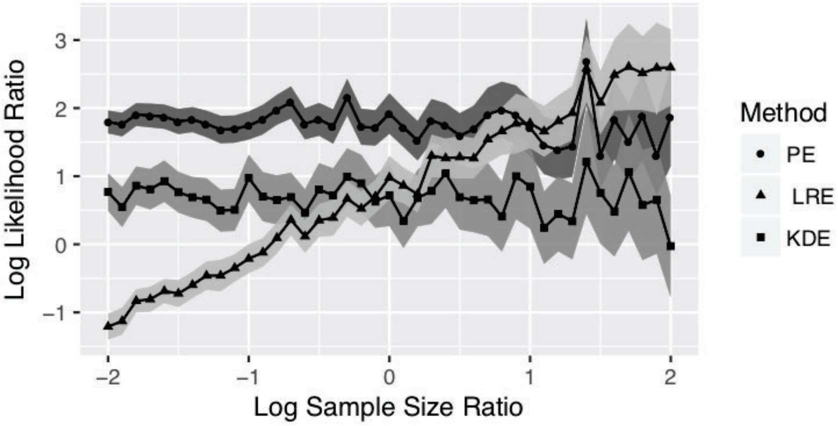


Figure 2: Confidence interval of estimated LLR in facial recognition data using PE, LRE, and KDE.

Table 2: Ranges of the confidence interval for different methods in facial recognition data

Method	Log Sampling Ratio	LLR	LB	UB	$Var(LLR)$
PE	-1.0	1.744	1.538	1.951	0.01113
	0.0	1.920	1.623	2.217	0.02294
	1.0	1.713	1.295	2.131	0.04548
LRE	-1.0	-0.211	-0.424	0.002	0.01180
	0.0	0.980	0.665	1.296	0.02597
	1.0	1.777	1.414	2.141	0.03441
KDE	-1.0	0.973	0.629	1.316	0.03067
	0.0	0.721	0.268	1.173	0.05335
	1.0	0.847	0.232	1.462	0.09855

Fingerprint Matching Example

We also apply the PE, LRE, and KDE methods to a set of fingerprint comparison scores to study their reproducibility and repeatability for fingerprints. The genuine and impostor comparison scores were generated by applying a fingerprint comparison algorithm using NIST Biometric Image Software to National Institute of Standards and Technology Special Database 4. Genuine scores were obtained by comparing two patches of the same rolled print of the same finger, and imposter scores were obtained by comparing patches of rolled prints from two different fingers.

The patch sizes are 128 by 128, 192 by 192 or 256 by 256. The neighboring patches with the same x-coordinate are shifted by the half of the patch width. This way, half of the area in the neighboring patches are overlapped. For the patches of 128 by 128, the patch starts at the coordinate (1,1), and the next patch starts at (1,65). Every patch has the same size of 128 by 128. The comparison scores, the numbers of matching minutiae, and the distance to the singularity point are recorded. The average and the standard deviation of all comparison scores are computed.

The scores in the genuine group are generally greater than the score in the imposter group. The sample means and sample standard deviations are 350.9 and 293.6 for the genuine group, and 7.5 and 2.5 for the imposter group. In our computation of the LLR values using all three methods, we randomly select 4000 genuine scores and various numbers of imposter scores, so that the log sample size ratio ranges from -2 to 2 by the increment of 0.1. When the sample size ratio changes, we repeat the sampling procedure to select genuine and imposter scores before the LLR methods are applied.

To get the empirical variances and the 95% confidence intervals for the LRE and KDE methods, 1000 iterations are adopted for each log sampling ratio value. For the PE method, our variance estimate approach given by Equation (7) was used to get the estimated variance and the 95% confidence interval. We estimate the LLR at the score of 10 with all the LRE and KDE methods. Note that the PE method assumes that the data are normally distributed.

Since the fingerprint scores data do not follow normal distributions, we use the Box-Cox power transformation given in Equation (4) to obtain the normality for both groups. We use a λ to transform both mated and non-mated data which is equal to the average of λ estimated from both groups separately. We also use the same λ value to transform the score 10.

Figure 3 shows the LLR values for the PE, KDE, and LRE methods when the sample size ratio varies. In terms of the repeatability, the LLR values from the PE and KDE methods have small fluctuations when the sample size ratio varies. However, the LLR values from these two methods differ by approximately 1. All the LLR values from these three methods are negative. If one uses zero as a decision threshold to decide whether the score 10 comes from H_p or H_d, all three methods should arrive at the same conclusion that the score of 10 likely supports H_d with all negative LLR values. Again, the LLR values from LRE have a linear relationship with the log sample size ratio, and thus, the repeatability of the LRE method is unsatisfactory. The LLR values from the KDE and PE methods are similar, indicating reproducibility between the two methods is high.

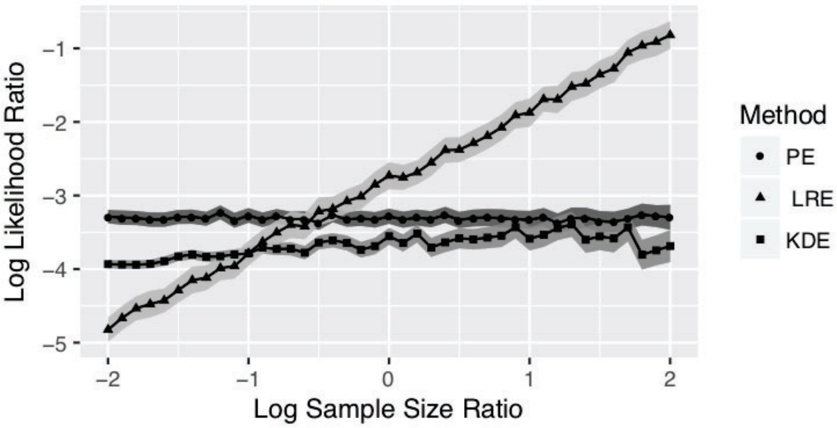


Figure 3: Confidence interval of estimated LLR in fingerprint identification data using PE, LRE, and KDE.

Table 3: Ranges of the confidence interval for different methods in fingerprint identification data

Method	Log Sampling Ratio	LLR	LB	UB	$Var(LLR)$
PE	-1.0	-3.453	-3.552	-3.354	0.002556
	0.0	-3.369	-3.471	-3.226	0.002742
	1.0	-3.429	-3.550	-3.307	0.003849
LRE	-1.0	-3.786	-3.953	-3.618	0.007291
	0.0	-2.731	-2.913	-2.548	0.008708
	1.0	-1.875	-2.067	-1.682	0.009629
KDE	-1.0	-3.788	-3.863	-3.712	0.001476
	0.0	-3.554	-3.668	-3.440	0.003372
	1.0	-3.588	-3.747	-3.429	0.006574

4.2.3 Research Question 3: Uncertainty Quantification

The uncertainty quantification of the SLR is explored through a simulation study using a variety of models for the data.

Binormal Data The datasets are generated using functions in R. Let S_p and S_d denote similarity scores simulated under the matching and non-matching groups, respectively, where $S_p \sim N(20, 9)$ and $S_d \sim N(10, 25)$. Then, we investigate the impact of the sample size ratio on the logarithm of the score-based likelihood ratio (LLR) values for a particular comparison score. We chose the score s_0, as the score at which the true genuine (matching) and impostor (non-matching) probability density functions intersect. The true value of the logarithm of the score-based likelihood ratio at s_0 is zero (the ratio of the two probabilities is 1 and so the LLR is zero). We then estimate the LLR values at s_0 as the sample size ratio varies.

Let M and N represent the sample sizes of genuine group and imposter group, respectively. To examine the variance, fix the total sample size to be $M + N = 10000$, and vary the log sampling ratio $\log(M/N)$ from -2 to 2 by 0.1, so we used 41 pairs of (M, N). For each pair of sample sizes, we simulate 1000 sets of simulated scores, and we denote the true variance ($Var(LLR)$) as the variance of the 1000 LLR. The estimated variance ($Var(LLR)$) is given by Equation (7).

Next, the coverage is the percentage of the 1000 LLR covered by the 95% confidence interval which is:

$$0 + Z_{0.95} \times \overline{Var(LLR)}, 0 - Z_{0.95} \times \overline{Var(LLR)})$$

since the true LLR for s_0 is 0. The resulting variances and coverages for the simulated data are given in Figure 4 below.

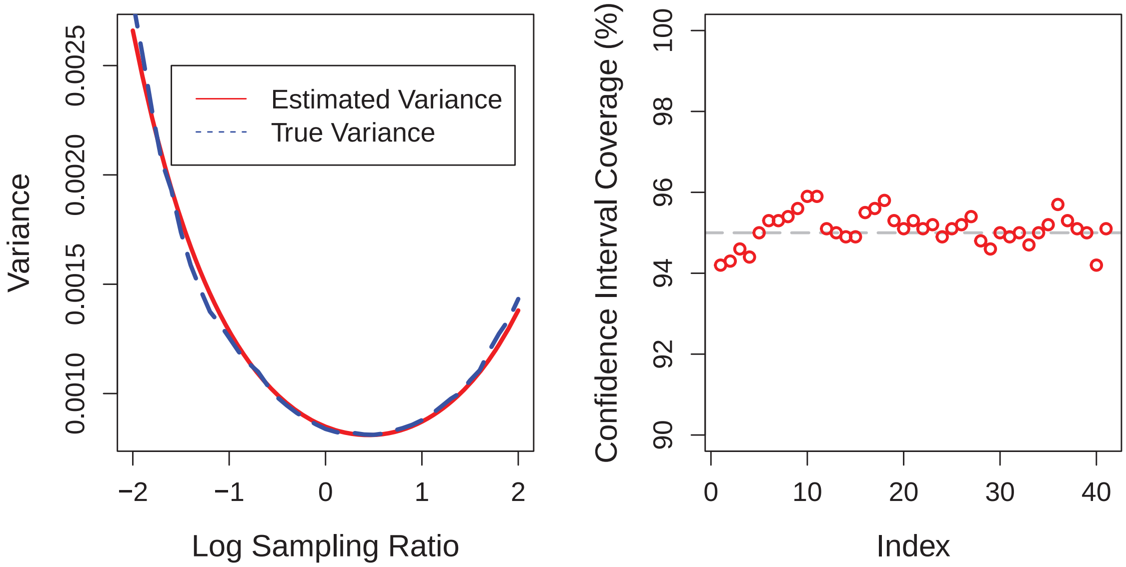


Figure 4: True variance and estimated variance (left) and Confidence interval coverage (right).

As seen in Figure 4 (left), the estimated variance agrees well with the true variance. Both of them reach their minimum with log sampling ratio equal to around 0.5. Moreover, it is shown in the Figure 4 (right), the confidence interval coverage varies about 95% with amplitude of 1%.

Other Data Types The data sets are generated using functions in R [37]. We compare PE, KDE and LRE in three simulation studies with different distributions for genuine and imposter groups. The data sets from each group follow different distributions described in Table 4. The distributions and parameters in each study are obtained from real data sets [16].

10000 data for each group is generated for each set up, and we randomly select data set from the total data based on the sample size (M, N). After this, we repeat the random selection step 1000 times, and calculate the empirical coverage. That is, we generate the confidence interval based on the variance, and check if the estimated LLR from the 1000 iterations falls within the confidence interval. We find the value of s_0 for each set up, which is the score for the true genuine and impostor probability density functions cross. The true value of LLR at s_0 is zero and the confidence interval is given by

$$0 + Z_{0.95} \times \quad \overline{Var(LLR)}, 0 - Z_{0.95} \times \quad \overline{Var(LLR)})$$

since the true LLR in s_0 is 0. The Bias in the table is given as

$$\text{Bias} \quad SLR \quad = |SLR - SLR| = |1 - SLR|.$$

We use SLR here (instead of the logarithm of the SLR) because the difference of the LLR values will give us very small numbers and is also hard to interpret.

Table 4: Distributions and parameters in the three simulation studies

Study	$f_p(s_{x,y})$		$f_d(s_{x,y})$	
	Distribution	Parameters	Distribution	Parameters

1	Normal	Mean = 20
		Variance = 9
			Normal	Mean = 10
				Variance = 25

2	Uniform	Min = 0
		Max = 1
			Beta	$shape_1 = 0.8$
				$shape_2 = 17$

3	Normal	mean = 2
		variance = 4
			t	Degrees of Freedom = 2

Figure 5 displays the bias and coverage estimation using the PE, KDE, and LRE methods. For the results of the PE method applied to the binormal dataset, the bias is small and the coverage is close to 95%. But when the data are not from normal distributions, the bias increases and the coverage is far different from 95%. This makes sense since the PE method relies on the distributional assumption. The results of the LRE method are heavily

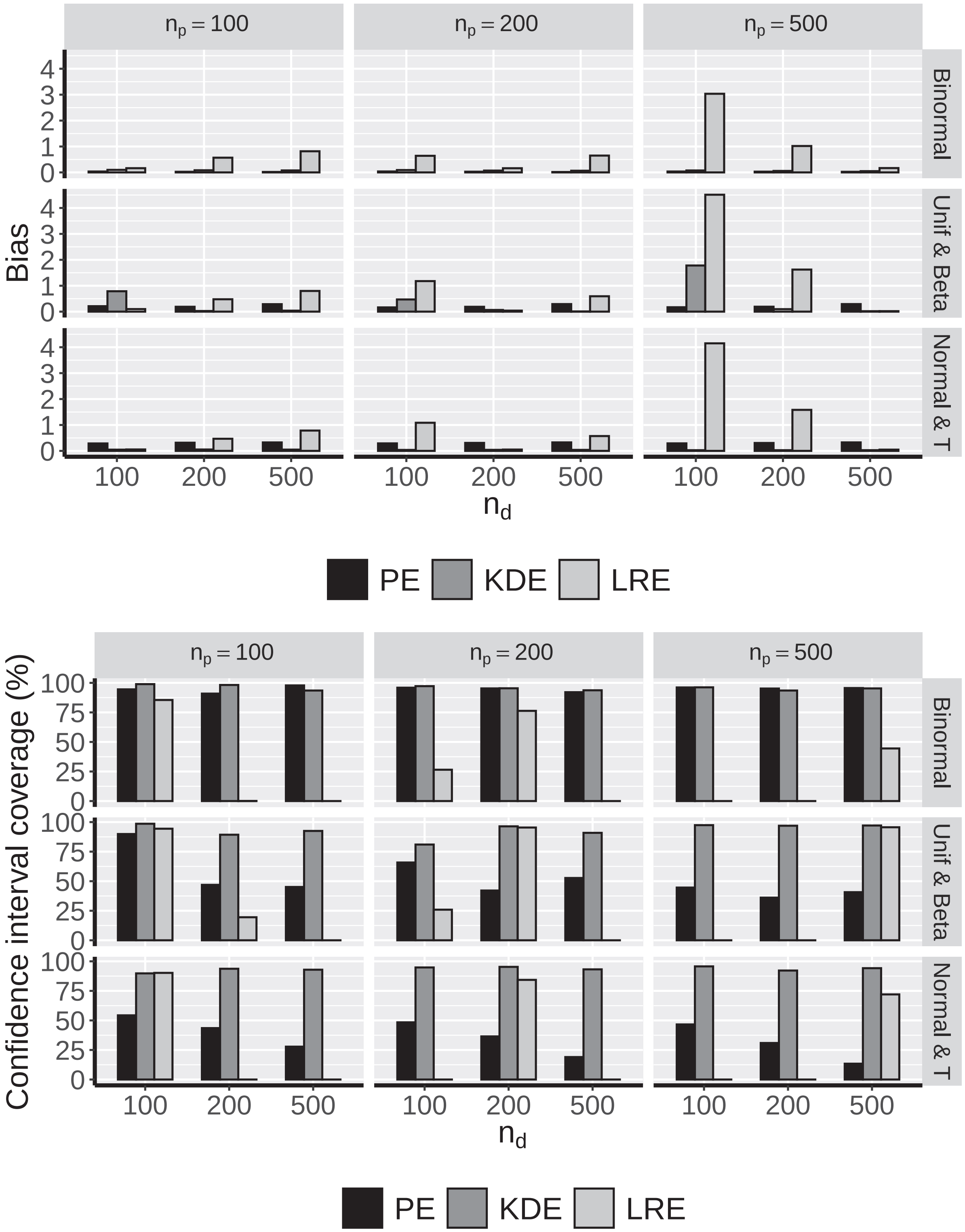


Figure 5: Bias and Confidence Interval Coverage with PE, KDE, and LRE method.

influenced by the sample size. Note that when we increase the difference between n_p and n_d, the bias of the LRE method increases and the coverage of LRE decreases dramatically. Overall, when the sample size difference is larger than 100, the coverage is no more than 30%. The KDE method produces small biases and the coverage is close to 95% through all the settings. Generally, when the sample size gets larger, the bias becomes smaller and the coverage gets closer to 95%.

4.2.4 Research Question 4: Visualization

One visual solution we explored in the project is developed from the relationship between the ROC curve for the Two-Stage approach and the likelihood ratio (LR). The derivative of the ROC curve is shown to be closely related to likelihood ratio [6]. Specifically, the LR is interpreted as the instantaneous change in the 1-RNMP in a unit change of RMP. An illustration of the relationship between ROC and LR is given in Figure 6 for simulated scores from normal distributions.

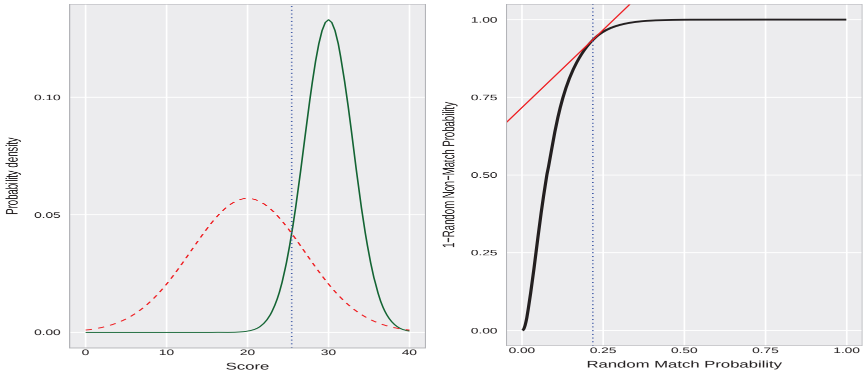


Figure 6: Left panel: dash curve– normal density of different-source scores, solid curve– normal density of same-source scores; right panel: solid black curve – ROC curve, the slope of the red line is likelihood ratio at the false match rate (or RMP)=0.2181

Additionally, the figures in Sections 4.2.1-4.2.3 illustrate the visualization tools for uncertainties associated with error rates and likelihood ratios.

4.3 Impact

The project has implications in shifting forensic practice paradigms in several important ways. First, formalizing source and sub-source propositions builds the foundation for forensic evidence interpretation. Second, the proposed paradigm based on error rates involves the interpretation of forensic evidence in a similarity stage and an exclusion stage. The paradigm is more intuitive for forensic scientists to understand than giving one single number based on likelihood ratio. Interpreting the forensic evidence in separate stages also provides the jury more information than the likelihood ratio. Finally, a set of visualization tools developed in the project present forensic examiners intuitive statistical graphics about the uncertainty associated with error rate based methods for quantifying evidence.

The work performed for this project has supported the Federal Bureau of Investigation Laboratory Division on research projects related to the interpretation of forensic evidence from handwriting and improvised explosive devices. The work performed has also supported National Institute of Standards and Technology on research projects related to accuracy evaluation of biometrics algorithms.

5 Artifacts

5.1 List of Products

1. Zhang, W, **Tang, LL**, Li, Q, Liu, A, Lee, M-LT. Order-restricted inference for clustered ROC data with application to fingerprint matching accuracy. *Biometrics.* 2020; 76: 863– 873. https://doi.org/10.1111/biom.13177

2. **Larry Tang**, Xiaochen Zhu, Ty Nguyen, **Danica M. Ommen**, **Elham Tabassi**. Score-based Likelihood Ratios based on ROC Curve Analysis and the Variabilities of the Likelihood Ratios, submitted to *Science and Justice.*

3. Dr. Larry Tang and his student Mengling He developed a Shiny app for evaluating the error rates and providing the variability of the error rates. The link to the Shiny app is https://forensicaccuracy.shinyapps.io/order_constrained_ROC_calculation/.

4. The R codes and the datasets are provided at: https://sites.google.com/view/larrytang/software?authuser=0

5.2 Data Sets Generated

None.

5.3 Dissemination Activities

Conference Presentations

Feb 2019 - Drs. Chris Saunders and Danica Ommen presented "On the Development of Score Rules for the Pairwise Sample Comparison of Particle Micromorphometry of Aluminum (Al) Powders" at the 2019 American Academy of Forensic Sciences Conference.

Sept 2019 - Dr. Danica Ommen presented "Which Forensic Likelihood Ratio Approach is Better?" at the 10th International Workshop on Simulation and Statistics.

Sept 2019 - Dr. Chris Saunders presented (with Dr. Danica Ommen as coauthor) "The Incorporation of U-processes for Bayesian Approaches to Pattern Recognition with Application to Forensic Source Identification" at the 10th International Workshop on Simulation and Statistics.

Sept 2019 - Dr. Larry Tang presented (with Dr. Danica Ommen as coauthor) "The Confidence Interval for the Likelihood Ratio with Application to Biometrics" at the 10th International Workshop on Simulation and Statistics.

Sept 2019 - Ms. Cami Fuglsby presented (with Drs. Chris Saunders and Danica Ommen as coauthors) the poster "A Class of Score Functions for the Analysis of Kinematic Handwriting Data" at the 10th International Workshop on Simulation and Statistics.

Feb 2020 - Ms. Cami Fuglsby presented (with Drs. Chris Saunders and Danica Ommen as coauthors) "The Interaction of Writing Profiles and Automated Scoring Rules" at the 2020 American Academy of Forensic Sciences Conference.

Mar 2020 - Dr. Xiaochen Zhu presented (with Dr. Larry Tang as coauthor) "ROC Methodology For Estimating Source-matching Likelihood Ratios and Evaluating Demographic Effects" at the Pittcon 2020 conference.

Aug 2020 - Dr. Danica Ommen presented (with Drs. Larry Tang and Christopher Saunders as coauthors) "A Method of Forensic Evidence Interpretation Using Error Rates" at the Joint Statistical Meetings .

Dec 2020 - Dr. Danica Ommen presented (with Drs. Larry Tang and Christopher Saunders as coauthors) "A Method of Forensic Evidence Interpretation Using Error Rates" at the International Chinese Statistical Association (ICSA) Applied Statistics Symposium.

Dec 2020 - Dr. Chris Saunders presented (with Dr. Danica Ommen as coauthor) "Bayesian Characterizations Of U-processes Used In Pattern Recognition With Application To Forensic Source Identification" at the International Chinese Statistical Association (ICSA) Applied Statistics Symposium.

Dec 2020 - Dr. Xiaochen Zhu presented (with Dr. Larry Tang as coauthor) "Order-Constrained ROC Regression with Application to Facial Recognition" at the International Chinese Statistical Association (ICSA) Applied Statistics Symposium.

Seminars/Workshops

Jan 2019 - Dr. Chris Saunders organized an invited session "Forensic Statistics" for the 10th International Workshop on Simulation and Statistics in Salzburg, Austria.

Feb 2019 - Dr. Larry Tang presented "Order-Restricted Inference for Evaluating Error Rates with Application to Fingerprint Matching" in the Department of Biostatistics, Bioinformatics & Biomathematics at Georgetown University.

Mar 2019 - Dr. Larry Tang presented "Order-Restricted Inference for Evaluating Error Rates with Application to Fingerprint Matching" in the Department of Statistics at University of Central Florida.

Sept 2019 - Dr. Tang gave a tutorial titled "Estimation of Soft-biometrics from fingerprints" at 10th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)

Dec 2019 - Dr. Larry Tang organized an invited session “Current advances in forensic statistics” for the International Chinese Statistical Association (ICSA) Applied Statistics Symposium.

Jan 2021 - Dr. Chris Saunders organized a topic contributed session “Bias and Interpretability in Biometrics for Forensic Science” for the 2021 Joint Statistical Meetings 2021.

References

[1] Colin G. G. Aitken, Paul Roberts, and Graham Jackson. *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings; Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses.* Royal Statistical Society’s Working Group on Statistics and the Law, London, UK, 1st edition, 2010.

[2] D. Ashbaugh. *Quantitative-qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology.* Florida: CRC Press, 1999.

[3] D. Bamber. The area above the ordinal dominance graph and the area below the receive operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415, 1975.

[4] Charles E.H. Berger and Klaas Slooten. The LR does not exist. *Science and Justice*, 56(5):388–391, 2016.

[5] A. Biedermann, S. Bozza, F. Taroni, and C. G. G. Aitken. Reframing the debate: A question of probability, not of likelihood ratio. *Science and Justice*, 56(5):392–396, 2016.

[6] 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.

[7] R Cook, I W Evett, G Jackson, P J Jones, and J A Lambert. A hierarchy of propositions: deciding which level to address in casework. *Science and Justice*, 38(4):231–239, 1998.

[8] James M. Curran. Admitting to uncertainty in the LR. *Science and Justice*, 56(5):380–382, 2016.

[9] A. Philip Dawid. Forensic likelihood ratio: Statistical problems and pitfalls. *Science and Justice*, 57:73–75, 2017.

[10] Ross H Day and Erica J Stecher. Sine of an illusion. *Perception*, 20(1):49–55, 1991.

[11] European Network of Forensic Science Institutes. *ENFSI Guideline for Evaluative Reporting in Forensic Science*, 2015.

[12] IW Evett and JA Lambert. The interpretation of refractive index measurements. iii. *Forensic Science International*, 20(3):237–245, 1982.

[13] David Faraggi and Benjamin Reiser. Estimation of the area under the ROC curve. *Statistics in medicine*, 21(20):3093–3106, 2002.

[14] Cami Fuglsby. U-statistics for characterizing forensic sufficiency studies. *South Dakota State University, MS Thesis*, 2017.

[15] Cami Fuglsby, Christopher Saunders, Danica M. Ommen, and Michael P. Caligiuri. Use of an automated system to evaluate feature dissimilarities in handwriting under a two-stage evaluative process. *Journal of Forensic Sciences*, 65(6):2080–2086, 2020.

[16] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011.

[17] J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[18] James A. Hanley and Karim O. Hajian-Tilaki. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology*, 4(1):49 – 58, 1997.

[19] Hariharan K Iyer and Steven P Lund. Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research (NIST JRES)-*, 122(Journal of Research (NIST JRES)-), 2017.

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

[21] Paul L Kirk. The ontogeny of criminalistics. *The Journal of Criminal Law, Criminology, and Police Science*, 54(2):235–238, 1963.

[22] Quon Yin Kwan. *Inference of Identify of Source.* Ph.D. Dissertation in Criminology, University of California, Berkeley, 1977.

[23] Dennis V Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

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

[25] D. Maltoni, M. Dario, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition.* London: Springer-Verlag, 2009.

[26] Geoffrey S. Morrison and Ewald Enzinger. What should a forensic practitioner's likelihood ratio be? *Science and Justice*, 56(5):374–379, 2016.

[27] Geoffrey Stewart Morrison. Measuring the validity and reliability of forensic likelihood-ratio systems. *Science and Justice*, 51(3):91–98, 2011.

[28] National Institute of Standards and Technology. Technical Colloquium on the Weight of Evidence. https://www.nist.gov/news-events/events/2017/06/technical-colloquium-weight-evidence, June 2017. U.S. Department of Commerce.

[29] National Research Council Committee on Identifying the Needs of the Forensic Sciences Community. *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, D.C., USA, 2009.

[30] Danica M. Ommen, Cami Fuglsby, and Michael P. Caligiuri. Advances toward validating examiner writership opinion based on handwriting kinematics. *Forensic Science International*, 318:110644, 2021.

[31] Danica M Ommen and Christopher P Saunders. Building a unified statistical framework for the forensic identification of source problems. *Law, Probability and Risk*, 17(2):179–197, 05 2018.

[32] Danica M. Ommen and Christopher P. Saunders. A problem in forensic science highlighting the differences between the bayes factor and likelihood ratio. *Statist. Sci.*, 36(3):344–359, 2021.

[33] Danica M. Ommen, Christopher P. Saunders, and Cedric Neumann. An argument against presenting interval quantifications as a surrogate for the value of evidence. *Science and Justice*, 56(5):383–387, 2016.

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

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

[36] President's Council of Advisors on Science and Technology. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods, 2016.

[37] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[38] David A Schum. *The evidential foundations of probabilistic reasoning.* Northwestern University Press, 1994.

[39] David E Shapiro. The interpretation of diagnostic tests. *Statistical Methods in Medical Research*, 8(2):113–134, 1999.

[40] M. J. Sjerps, I. Alberink, A. Bolck, R. Stoel, P. Vergeer, and J. H. van Zanten. Uncertainty and LR; to integrate or not to integrate, that's the question. *Law, Probability, and Risk*, 15(1):23–29, March 2016.

[41] Franco Taroni, Silvia Bozza, Alex Biedermann, and Colin G. G. Aitken. Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability, and Risk*, 15(1):1–16, March 2016.

[42] Duncan Taylor, Tacha Hicks, and Christophe Champod. Using sensitivity analyses in Bayesian Networks to highlight the impact of data paucity and direct future analyses: a contribution to the debate on measuring and reporting the precision of likelihood ratios. *Science and Justice*, 56(5):402–410, 2016.

[43] William C Thompson, Nicholas Scurich, Rachel Dioso-Villa, and Brenda Velazquez. Evaluating negative forensic evidence: When do jurors treat absence of evidence as evidence of absence? *Journal of Empirical Legal Studies*, 14(3):569–591, 2017.

[44] 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.

[45] Ardo van den Hout and Ivo Alberink. Posterior distributions for likelihood ratios in forensic science. *Science and Justice*, 56(5):397–401, 2016.

[46] Susan VanderPlas and Heike Hofmann. Signs of the sine illusion?why we need to care. *Journal of Computational and Graphical Statistics*, 24(4):1170–1190, 2015.

[47] C. I. Watson, M. D. Garris, E. Tabassi, C. L. Wilson, R. M. Mccabe, S. Janet, and et al. *User's Guide to NIST Biometric Image Software (NBIS).* Gaithersburg, MD, 2007.

[48] H. Zhong and R. L. Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, 9:621–634, 2008.

[49] X. H. Zhou, D. K. McClish, and N.A. Obuchowski. *Statistical Methods in Diagnostic Medicine*. Wiley, New York, 2002.

[50] Kelly H. Zou, Clare M. Tempany, Julia R. Fielding, and Stuart G. Silverman. Original smooth receiver operating characteristic curve estimation from continuous data: Statistical methods for analyzing the predictive value of spiral ct of ureteral stones. *Academic Radiology*, 5(10):680 – 687, 1998.