Resampling methods for score likelihood ratio based inference for source attribution problems

by

Federico A. Veneri Guarch

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Danica M. Ommen, Major Professor
Alicia Carriquiry
Jarad Niemi
Daniel Nordman
Roy Vivekananda

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2024

DEDICATION

To Amelia, who kept stealing my k_ybo_rds k_ys.

And to Cecilia and Tau, who kept getting them back.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I want to express my gratitude to Dr. Danica Ommen, my major professor, who introduced me to the world of forensic statistics. And to my committee members, Dr. Alicia Carriquiry, Dr. Jarad Niemi, Dr. Daniel Nordman, and Dr. Roy Vivekananda, for their guidance throughout the process. I have had the pleasure of attending their lectures and learning from all of them.

I am thankful for the support, comments, and suggestions from fellow students, faculty, and staff at the Department of Statistics and the Center for Statistics and Applications in Forensic Evidence (CSAFE) at Iowa State University. I especially want to thank the Fulbright Commission and ANII [1], and CSAFE [2], who partially funded my graduate journey.

Thank you to friends and family, who, from over nine thousand kilometers away, were always willing to hear me rambling about statistics and how cold it gets in Ames during the winter.

To Cecilia, for believing in me. To Tau, for the long walks. To Amelia, for sharing your immense curiosity with me everyday.

[1] 2019 Fulbright Program

[2] This work was funded (or partially funded) by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University,

ABSTRACT

This dissertation addresses source attribution problems, an inferential task that contrasts two opposing propositions regarding the origin of items. These inferential problems arise in multiple domains but play a key role in forensic science. Due to the complexity of evidence found in practical applications, machine learning has been proposed as an alternative to evaluate the similarity between items when a probabilistic model is not feasible to construct a traditional Likelihood ratio. Score-based likelihood ratio inference hence provides an alternative framework to assess the strength of statistical evidence in this context.

Our work focuses on the common and specific source inferential problems and addresses the dependence structure generated when creating training and estimation sets to develop these inferential systems. We present resampling plans to remedy these shortcomings and how ensemble learning approaches could strengthen the current methods. Chapter 2 introduces Strong Source Resampling (SSR), a source-aware resampling plan for the common source problem. This idea is extended to Weak Source Resampling (WSR) in Chapter 4. These resampling plans are the basis for developing base systems combined into a final value of evidence using an ensemble learning approach proposed in Chapter 2. Chapter 3 focuses on the specific source problem, introducing synthetic source anchoring, which uses synthetic items as data augmentation, allowing the development of specific source score likelihood ratios. Lastly, Chapter 4 introduces discrepancy metrics for score likelihood ratio-based inference that can be used to study model misspecification and the effects of not accounting for dependence. Simulation results and applications in both chapters suggest that combining ensemble learning with a source-aware resampling could provide stronger, more stable statistical evidence value in the correct direction for machine learning and simple score-based likelihood ratios. Chapter 5 provides general conclusions and some avenues for further research.

CHAPTER 1. GENERAL INTRODUCTION

Source attribution problems are a class of inferential tasks where researchers are faced with opposing propositions regarding the origin of items. There are two types of source attribution: common source and specific source. Under the common source framework, the origin of items under consideration is unknown, and the inferential problem is to provide a probabilistic statement if the items share a common source or not. Under the specific source, the origin of a subset of items is known with certainty, and the inferential problem is to assess if the remaining subset of items could have originated from that specific source under consideration [3].

These inferential problems arise in multiple domains but play a key role in forensic science. An example in ballistic examination allows us to illustrate the relevance and the distinctions between the types of source attribution problems. Consider the case where bullet casings (items) were found in two distinct crime scenes. Forensic experts may be asked to assess if the two crime scenes are related via the source of the bullet, in this case, the firearm used. The evidence can be analyzed in the absence of the firearm(s) that shot the bullets. This is an example of a common source problem where the origin of the items is not known.

Following this example, consider that a person of interest was detained after an investigation, and a firearm registered to the individual was recovered (gun). In this scenario, the expert may generate bullet casings (items) under controlled conditions; hence, the source-item relationship is known with certainty for those casings, and experts are tasked with comparing those to the ones found at the crime scene. Under this specific source scenario, experts aim to link the person of interest's firearm to a crime scene.

Professional guidelines have encouraged forensic experts to express their findings in probabilistic terms to achieve balanced, logical, robust, and transparent communication with judges and jurors while incorporating uncertainty about their results [6]. Statisticians have

contributed to this task from different inferential frameworks, including the more prominent Bayesian, Fiducial, and Likelihood perspectives. The latter is the focus of this dissertation, in particular, score-based likelihood ratio inference.

The complexity of evidence found in practice has led to new applications in machine learning to derive scores that can be used to evaluate similarity between items [5, 4, 1, 2], score-based likelihood ratio inference provides a framework to assess the strength of statistical evidence supporting opposing propositions. These propositions are often designed to reflect the two opposing sides in a criminal trial: the prosecution and the defense.

Our work addresses the dependence structure generated when creating training and estimation sets to develop these inferential systems since sources and items are used multiple times. We also explore how resampling plans can remedy this situation and how ensembling learning approaches could strengthen the current methods.

Chapter 2 introduces a source-aware resampling plan for the common source problem, namely Strong Source Resampling (SSR). Chapter 4 extends this idea to Weak Source Resampling (WSR). The strong version enforces that sources are used only once, while the weaker enforces that items are used only once. These resampling plans are the basis for developing weak learners that are combined into a final ensemble value of evidence.

Chapter 3 focuses on the specific source problem. The lack of available data compounds with the dependence structure to reduce system performance. We introduced synthetic anchoring, which used synthetic items as a data augmentation procedure to create learning instances, allowing the development of proper specific source score likelihood ratios.

Simulation results and applications in both chapters suggest that combining ensemble learning with a source-aware resampling could provide stronger, more stable statistical evidence value in the correct direction for machine learning and simple score-based likelihood ratios. While these results are promising for forensic science, our approach could also be applied to source inference problems in other domains.

To further understand the condition under which this improvement could be observed, Chapter 4 introduces discrepancy metrics for score likelihood ratio-based inference. This discrepancy metric can be generally used to study model misspecification, and we focus our work on the estimation stage for the common source problem. We illustrate how they can be used to assess dependence's effect on inference. We provide a simple univariate example that is the basis for our simulation study. Simulation results suggest that while dependence can alter the inference drawn, there may be a tradeoff between thinning out the dependence and retaining more learning instances and that some estimation methods may be more sensitive to dependence.

Lastly, Chapter 5 provides a general conclusion for the dissertation and some avenues for further research.

1.1 References

[1] Carriquiry, A., Hofmann, H., Tai, X. H., and VanderPlas, S. (2019). Machine learning in forensic applications. *Significance*, 16(2):29–35.

[2] Kafadar, K. and Carriquiry, A. L. (2024). Challenges in modeling, interpreting, and drawing conclusions from images as forensic evidence. *Statistics and Data Science in Imaging*, 1(1):2401758.

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

[4] Park, S., Carriquiry, A., et al. (2019). Learning algorithms to evaluate forensic glass evidence. *Annals of Applied Statistics*, 13(2):1068–1102.

[5] Stern, H. S. (2017). Statistical issues in forensic science. *Annual Review of Statistics and Its Application*, 4:225–244.

[6] Willis, S., Aitken, C., Barrett, A., Berger, C., Biedermann, A., Champod, C., Hicks, T., Lucena-Molina, J., Lunt, L., McDermott, S., McKenna, L., Nordgaard, A., O'Donnell, G., Rasmusson, B., Sjerps, M., Taroni, F., and Zadora, G. (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science*. European Network of Forensic Science Institutes, http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

CHAPTER 2. ENSEMBLE LEARNING FOR SCORE LIKELIHOOD RATIOS UNDER THE COMMON SOURCE PROBLEM

Federico Veneri and Danica M. Ommen

Department of Statistics, Iowa State University

Modified from a manuscript published in

Statistical Analysis and Data Mining: The ASA Data Science Journal

2.1 Abstract

Machine learning-based Score Likelihood Ratios (SLRs) have emerged as alternatives to traditional Likelihood Ratios and Bayes Factors to quantify the value of evidence when contrasting two opposing propositions. When developing a conventional statistical model is infeasible, machine learning can be used to construct a (dis)similarity score for complex data and estimate the ratio of the conditional distributions of the scores. Under the common source problem, the opposing propositions address if two items come from the same source. To develop their SLRs, practitioners create data sets using pairwise comparisons from a background population sample. These comparisons result in a complex dependence structure that violates the independence assumption made by many popular methods. We propose a resampling step to remedy this lack of independence and an ensemble approach to enhance the performance of SLR systems. First, we introduce a source-aware resampling plan to construct data sets where the independence assumption is met. Using these newly created sets, we train multiple base SLRs and aggregate their outputs into a final value of evidence. Our experimental results show that this ensemble SLR can outperform a traditional SLR approach in terms of the rate of misleading evidence and discriminatory power and present more consistent results.

2.2 Background

The common source problem refers to an inferential problem where an expert's objective is to provide some probabilistic statement regarding the origin of two items and whether the same source has generated them. While the common source problem may arise in different disciplines, it is particularly interesting to the criminal justice system and forensic experts. In forensic science, the source of the items may refer to a person - in the case of handwriting, the source refers to the writer -, or an object - in the case of bullets, the source refers to firearms [37, 9]. Regardless of the source type, under the common source problem, experts will assess two contrasting propositions after observing some data, in the case of forensic analysis: given the evidence, do these two items share a common unspecified source?

Professional guidelines encourage forensic experts to provide their findings in a balanced, logical, robust, and transparent way so that judges and jurors can assess the strength of the evidence presented. To achieve this requirement, an expert should present their results in terms of probabilities to communicate the measure of uncertainty in their findings. Under this framework, experts can use likelihood ratios to provide a numerical assessment of the strength of the evidence [48]. Using likelihood ratios requires formulating a probabilistic model for the joint distribution of the features that may be challenging, if not infeasible, to estimate, leading researchers to consider score-based likelihood ratios (SLRs) as an alternative [43, 34]. This score measures the (dis)similarity between the feature vector of two items, reducing a complex model to a lower dimensional value [43]. To assess the (dis)similarity, researchers have begun using machine learning to construct these scores and estimate their conditional density to assess the likelihood of the score under the contrasting propositions.

Although the SLR framework has shown promise in different areas of forensics (handwriting: [21, 25, 8, 13], glass: [40, 41], fingerprints: [35, 28, 20], speaker recognition: [19], ink: [36], MDMA tablets: [3, 4], digital: [17], cameras: [42]), concerns have been raised regarding their behavior and use in forensic settings [34, 32], and their evaluation has been the subject of extensive research in the literature [18, 33]. Ishihara and Carne [22] summarize the benefits and shortcomings of using

score-based methods; using a lower dimensional metric reduces the need for complex models, and simpler estimation procedures are required to estimate univariate conditional scores. As a drawback, this dimensional reduction may imply a loss of information and does not address the typicality of the features.

One less-studied aspect is the role of pairwise dependence structure on developing SLRs. It is often assumed that experts have an independent sample of the background population from which they can construct training and estimation sets to develop their algorithms. In practice, pairwise comparisons are used to create these sets, generating a complex dependence structure often overlooked or ignored. However, this dependence violates the independence assumption required by most popular machine learning and density estimation procedures. Our work aims to address this issue.

To remedy the lack of independence, we introduce a source-aware sampling plan to generate samples where the independence assumption is met. This sampling plan is the basis for our proposed ensemble approach for SLRs. By resampling the data, we generate multiple base SLRs mimicking the role of weak learners in ensemble learning and combine their outputs into a final score to measure the probative value of the evidence. By ensembling multiple base learners, we aim to provide stronger and more stable values of evidence, updating prior beliefs in the correct direction.

To illustrate our approach, we explore forensic handwriting data collected by the Center for Statistics and Applications in Forensic Evidence (CSAFE) [11] and the CVL database [26]. Traditionally, document comparison has relied on visual inspection to identify individual characteristics or features. There has been a general call to strengthen the scientific basis and statistical foundations in criminal justice and to push for more objective means of comparison in forensic analysis [27], and handwriting analysis has been previously identified as an area to be strengthened [10]. Previous work has shown the potential of the SLR approach in handwriting analysis [21, 25], and our work contributes in the same direction. The simulation study suggests that an ensemble approach can enhance traditional SLRs. The proposed ensemble version

produced fewer misleading results for known matches at the cost of a slight increase of the rate of misleading evidence for non-matches, and more discriminatory results overall. Furthermore, aggregation methods provided more consistent conclusions for the same hold-out evidence. While these results are promising for handwriting, we believe that our proposed resampling plan and ensembling approach could also be applied to the common source problem in other domains outside of forensic sciences where learning instances are generated using pairwise comparisons.

2.3 Handwriting Data

As part of an ongoing project, CSAFE at Iowa State University has collected handwriting data and made it available to researchers [11] [1]. Participants in the study were tasked with transcribing three prompts across three sessions spaced in time (at least three weeks). Each prompt is transcribed three times in each session, resulting in nine samples per writer of the same prompt at the end of the study. For our work, we used the London Letter (CSAFE-LND), as the sample from the background population to construct our SLR systems. The London Letter is the longest prompt collected in CSAFE's study and is a common exemplar used in handwriting analysis that includes every letter (in both lower and upper case) and numbers [39]. At the time of our analysis, the databases consisted of 241 writers. To construct validation sets, we used the CVL database [26]. The original database consists of 311 writers who were asked to transcribe different texts chosen from literary works in English and German. Traditional feature generation in questioned documents is based on visual inspection by a trained expert who, based on years of training and expertise, can identify distinctive traits. We follow an approach developed by CSAFE authors [12, 2] that decomposes writing samples into graphs, roughly matching letters, and assigns each into one of 40 clusters. Hence for each document, we obtain a forty-dimensional vector of cluster counts. To account for documents of different lengths, we transform the vector of counts into proportions, each entry being between 0 and 1. A zero entry indicates that the writer did not write any graphs that could be categorized into that particular cluster in the documents.

[1] The most up-to-date database can be accessed online https://forensicstats.org/data/

These features have been proven useful for forensic comparison under the common and specific source problem [25] and the closed-set writer identification problem [12] since writers tend to reproduce similar writing patterns. Appendix 2.8 illustrates this fact for selected writers.

2.4 Methods

2.4.1 Likelihood ratios and the common source problem

In a criminal case, forensic scientists examine the evidence and present their findings to jurors (or a judge), who are, in terms, the ones that will combine all the information to deliver a final judgment. Under a probabilistic framework, the jurors are contrasting two propositions traditionally referred to as the prosecutor (H_p) and the defense propositions (H_d) conditional on the evidence observed [1]. Applying Bayes theorem, the ratio of probability can be expressed as:

$$\frac{P(H_p|E)}{P(H_d|E)} = \underbrace{\frac{P(E|H_p)}{P(E|H_d)}}_{\text{Likelihood ratio}} \underbrace{\frac{P(H_p)}{P(H_d)}}_{\text{Prior odds}} . \tag{2.1}$$

In equation 2.1, the juror's prior beliefs regarding the propositions are updated via a likelihood ratio. Forensic experts are advised to present their findings in this manner by scientific and professional organizations [48]. In the case of forensic handwriting, experts may be presented with a pair of questioned documents as evidence $E = (E_x, E_y)$ and asked to evaluate if a common writer wrote the two documents. Under the common source framework [37], we can state the propositions as follows:

- H_p : E_x and E_y were written by the same unknown writer.
- H_d : E_x and E_y were written by two different unknown writers.

To assess these competing propositions, forensic experts can rely on observed features of the questioned document. Let u_i denoted the features of E_i $(i = x, y)$. If the joint distribution of the features under each of the competing propositions, denoted by $f(u_x, u_y \mid H_j)$ $(j = d, p)$, is known,

the likelihood ratio could be computed as:

$$LR = \frac{f(u_x, u_y|H_p)}{f(u_x, u_y|H_d)}, \tag{2.2}$$

and interpreted as follows: a $LR > 1$ would indicate that the priors are being updated towards the prosecutor, meaning the evidence supports the prosecutor's proposition, while a $LR < 1$ would be interpreted as being updated towards the defense. To estimate the joint probability model, researchers use a sample of the background population or reference set composed of information previously collected. Let A denote the reference set, E^A_{ij} an individual item j $(j = 1, \ldots, n_i)$ from source i $(j = 1, \ldots, m)$ and A_{ij} the corresponding measurement from item j from source i. Ommen and Saunders [38] express the forensic proposition and the process that generated the data available to the expert as a sampling model. They consider that the reference set A was generated first by randomly sampling m sources from a reference population and, within each source, sampling n_i items. In the case of handwriting evidence, it would be equivalent to procuring a sample of writers and, within each writer, procuring some handwriting samples from each of them. The authors denote this sampling mechanism as M_a. Under the prosecutor proposition, H_p, in terms of sampling, a single new source was obtained from the population, and two items were subsequently generated E_x and E_y. Under the defense proposition, H_d, two new sources have been generated, one associated with E_x and another associated with E_y [2]. In essence, the experts provide information that allows the decision maker to infer if two sources (H_d) or one source is at play (H_p) [37]. Developing a model can be challenging, especially for complex measurements that often arise in pattern evidence. Even if the model can be formulated, the estimation could prove challenging [21]. Hence, experts have relied on machine learning comparison metrics and density estimation procedures to construct SLRs.

[2] The original common source problem allows for multiple items being generated from the same source, we consider only one for simplicity

2.4.2 The score likelihood ratio approach

An alternative to the LR relies on using a score-based likelihood ratio, often involving popular machine learning algorithms. Adapting the notation presented in [25, 21, 43] , the SLR can be generically defined as:

$$\mathrm{SLR}(\mathrm{u_x}, \mathrm{u_y}) = \frac{g\left(\Delta(u_x, u_y) \mid H_p\right)}{g\left(\Delta(u_x, u_y) \mid H_d\right)} \tag{2.3}$$

where $\Delta()$ is a (dis)similarity metric that allows the comparison of items E_x and E_y via their observed features u_x and u_y, respectively, and the conditional density functions $g\left(\cdot \mid H_p\right)$ and $g\left(\cdot \mid H_d\right)$ allows to assess the likelihood of the score obtained under the alternative propositions. The numerator (denominator) in equation 4.8 can be interpreted as the likelihood of the score under H_p (H_d). Hence, an SLR > 1 (SLR < 1) can be interpreted as evidence toward the prosecutor (defense), resulting in a similar interpretation to likelihood ratios. The development of an SLR system can then be decomposed into two steps, 1) developing a (dis)similarity function and 2) estimating the conditional densities of the scores under both propositions. To complete that endeavor, experts may have at their disposal a reference data set as described in Section 2.4.1, where A_{ij} denotes the measurements from item j and source i. Experts may split the sources into two data sets, one for developing their (dis)similarity metric and a second set to estimate the density functions. In each set, pairwise comparisons are created by taking combinations of items in the reference set.

Without loss of generality, assume that the reference set consists of m sources and n items within each source for a total of $N = n * m$; hence the total pairings are $\binom{mn}{2}$. Consider a particular pairing with two measurements: A_{ik} and A_{jl}. If the pair came from the same source ($i = j$), they are considered a known match (KM), while in the case that the pair do not share the same source ($i \neq j$), they are considered a known non-match (KNM). In total, the comparison dataset consist of $\binom{n}{2}m = n_{KM}$ known matches and $\binom{mn}{2} - \binom{n}{2}m = n_{KNM}$ known non-matches. As in Veneri and Ommen [47], we illustrate generating a comparison set using a diagram. Consider the case of $m = 10$ sources and $n = 3$ items. The total number of items is 30, and

comparisons are 435 ($n_{KM} = 30$, $n_{KNM} = 405$). The first arc diagram of Figure 2.1 illustrates all pairwise comparisons, nodes representing items, and edges comparisons, red for KM and blue for KNM. Consider the case where the measurements are a vector of dimensions P, indexed by p ($p = 1 \ldots, P$). Features can be created for each comparison considering element-wise absolute differences:

$$d_p(u_{x[p]}, u_{y[p]}) = |u_{x[p]} - u_{y[p]}| \tag{2.4}$$

An aggregated feature can be generated considering the L_1 norm of the differences.

$$d_{L1}(u_x, u_y) = \sum_{p=1}^{P} |u_{x[p]} - u_{y[p]}| \tag{2.5}$$

For handwriting data, each element in the vector is the proportion of the graphs in the document that was assigned to one of the $P = 40$ clusters, and the L_1 distance is the sum across all absolute differences resulting in a 41-dimensional vector that can be used as a feature in a classification problem. Pairs from the same source (KM) can be considered as positive cases ($y = 1$), and pairs from different sources (KNM) as negative cases ($y = -1$). The Random Forest classifier [6] has become a popular model in forensic science to construct a similarity metric [25, 40, 41]. Once the Random Forest classifier is trained, it can be used to map from the features space to a univariate score between zero and one, $\widehat{rf} : (\mathbb{R}^+)^{P+1} \mapsto [0, 1]$. We can consider the outputs of the Random Forest as similarity metrics since larger values (closer to 1) would be associated with pairs from the same source.

The SLR framework requires estimating the conditional densities to assess how likely the score is under both propositions. In practice, if the practitioner has available a set for estimation purposes, the newly trained classifier can be used to compute the score of the pairs in the estimation set to generate scores for pairs under both propositions: $\left\{\delta_i^{(KM)}\right\}_{i=1}^{n_{KM}}$ and $\left\{\delta_j^{(KNM)}\right\}_{j=1}^{n_{KNM}}$. This relates to the density ratio estimation problem [44]. Under this framework, let the scores observed be independent and identically distributed (iid) samples from their corresponding conditional distribution, meaning:

$$\left\{\delta_i^{(KM)}\right\}_{i=1}^{n_{KM}} \overset{\text{iid}}{\sim} g(\boldsymbol{\delta} \mid H_p) \quad \text{and} \quad \left\{\delta_j^{(KNM)}\right\}_{j=1}^{n_{KNM}} \overset{\text{iid}}{\sim} g(\boldsymbol{\delta} \mid H_d), \tag{2.6}$$

and popular parametric and non-parametric density estimations have been used to convert the scores to SLRs. The end goal for developing an SLR system is not the estimation of the densities but finding the boundaries between both propositions and using that information to update the conclusions. Morrisson [31] distinguishes between generative methods that explicitly model distribution in each proposition and discriminative methods that focus on the boundary between propositions. Logistic regressions have been widely used in the latter category, especially in forensic voice comparison [31]. Outside of the forensic domain, Sugiyama et al. [44] describe using a probabilistic classifier as a density ratio estimator. The densities of interest can be re written as $g(\boldsymbol{\delta} \mid H_p) = g(\boldsymbol{\delta} \mid y = 1)$ and $g(\boldsymbol{\delta} \mid H_d) = g(\boldsymbol{\delta} \mid y = -1)$. Considering the ratio and applying Bayes Theorem to the density ratio results in

$$\frac{g(\boldsymbol{\delta} \mid H_p)}{g(\boldsymbol{\delta} \mid H_d)} = \frac{p(y=-1)}{p(y=+1)} \frac{p(y=+1 \mid \delta)}{p(y=-1 \mid \delta)}. \tag{2.7}$$

The first ratio in the right-hand side of equation 2.7 can be approximated by the proportion in the sample, meaning:

$$\frac{p(y=-1)}{p(y=+1)} \approx \frac{n_{KNM}}{n_{KM}}. \tag{2.8}$$

The second ratio consists of the probability of belonging to a class given the score, and a logistic regression classifier can be used for this purpose to obtain

$$\hat{\text{SLR}}(u_x, u_y) = \frac{n_{KNM}}{n_{KM}} \exp^{(\beta_0 + \beta_1 \delta)}, \tag{2.9}$$

where $\delta = \Delta(u_x, u_y)$ is the similarity score obtained using the trained random forest. If the estimation sample is balanced, meaning $n_{KM} = n_{KNM}$, the ratio estimator simplifies. Furthermore, it is a well-established fact in the literature that an imbalanced dataset affects the performance of classifiers as the one used to construct the (dis)similarity metric [7, 24] and also affects density estimation procedures [50].

In practice, when developing their models, forensic statisticians can preprocess the data by down-sampling the majority class before developing their (dis)similarity metric and estimating the conditional distribution of the scores. This results in known non-matching pairs being dropped from the sample. We illustrate this in panel B of Figure 2.1, where the comparisons from

Panel A have been down-sampled to have a balanced dataset of 30 observations from known matches and 30 for known non-matched pairs. The diagram allows us to illustrate that even if down-sampling has solved the imbalanced problem, it has not addressed the underlying hierarchical dependence generated by having sources and items. Popular classification and density estimation methods assume independence in the data, even when the data are not iid, but this is not the case in pairwise comparisons where the same source is used in multiple comparisons, and the same items are used multiple times. In our diagram, source one is used in three known match comparisons and multiple known non-match comparisons. Furthermore, the first item of source one is used in four comparisons.

2.4.3 Sampling and ensembling for SLR systems

Utilizing training and testing sets has been broadly adopted as a common practice in machine learning problems. In forensic statistics, a similar idea has been adopted under the likelihood ratio framework by splitting the data used for estimating the joint probability model and for assessing the performance [29]. In the case of SLR evaluation, scores that do not require an additional training stage (e.g., distance-based score) have been featured more prominently (e.g.:[3, 18]). When a new metric is trained using machine learning methods, three sets are required for training, estimation, and validation [47]. The training set is used to develop the (dis)similarity score, an estimation set to estimate the conditional density functions, and a separate validation set to compute the performance. Different authors have addressed how these sets should be constructed. For the common source, Neumann et al. [34], and Ommen and Saunders [38] proposed a thought experiment to establish how the estimation sets could be constructed for a distance-based metric. Under the prosecutor's proposition, the sampling distribution can be studied by considering a large sample of sources and comparing a single pair of items from each source to create a sufficiently large set of independent scores. Under the defense proposition, the distribution of the score can be studied by sampling a large number of independent pairs of sources and comparing an item from the first source to an item from the second source to create a

sufficiently large set of independent scores. In practice, however, pairwise comparison is used to generate learning instances, inducing a dependency among scores. This process generates a dataset from the background sample A that results in the same source and items used for multiple comparisons, as is the case of item one of source one in Panel A and B of Figure 2.1.

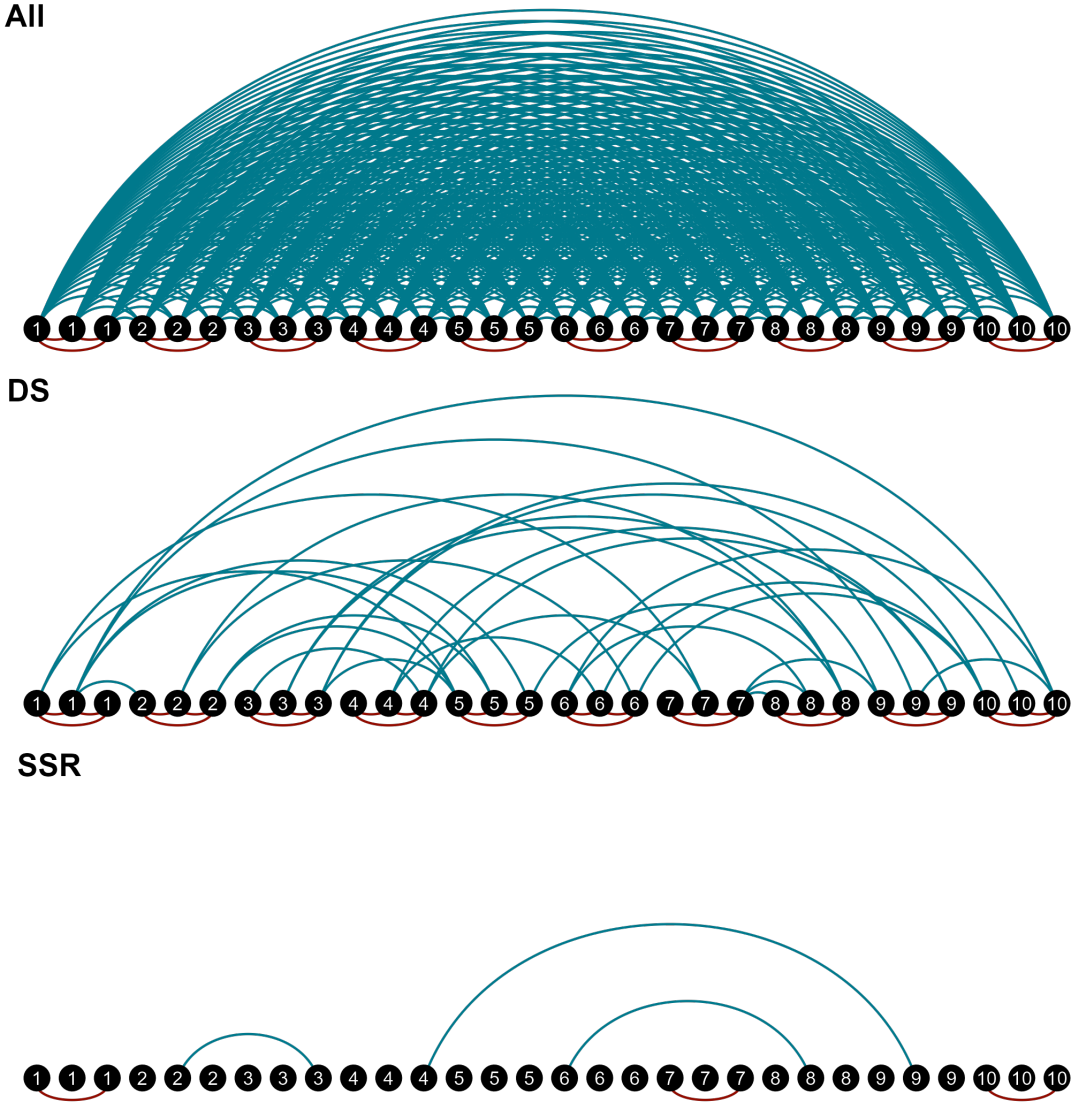
Following the principled way outlined by previous authors, it would be desirable for each source to be used only once, either as a known match or a known non-match comparison. A weaker constraint would be to use items only once, but that would disregard the hierarchical structure of the data. We implemented Strong Source Resampling (SSR, Section 2.4.3.1) to emulate the sampling process suggested by the authors and impose this stronger constraint, ensuring that items and sources are used only once. We illustrate the result of applying SSR to our initial example in panel C of Figure 2.1. While this sampling step remedies the dependence structure, the drawback is a drastic loss of information. To make the most out of the data, we propose using our resampling approach as a preprocessing step within an ensemble learning framework. We propose to train base score-likelihood ratios, which fulfill the role of weak learners, and combine their outputs into an ensemble score likelihood ratio (ESLR, Section 2.4.3.2).

2.4.3.1 Strong Source Resampling

To remedy the dependence structure, we introduced a source-aware sampling plan we denote as Strong Source Resampling (SSR). Our approach can be classified as a resampling plan, a category including methods like a jackknife, cross-validation, and bootstrap sampling, among others [14]. Among these methods, our proposed method shares the most similarities with bootstrap sampling.

Bootstrap sampling [15] was proposed to study the properties of statistics that can be considered a function of an unknown distribution F by using its empirical counterpart $\hat{F}$. In machine learning, it has become a popular approach to emulate the process of generating new training data [23]. A particular application is bagging, a special case of ensemble learning, where sampling without replacement is used to create new data to train base learners and aggregate

Figure 2.1 Pairwise comparison and sampling algorithms



Note: Nodes in the diagram represent the items in the set, indexed by labels to indicate the source. Edges represent comparison for KM (red) and KNM (blue) under different sampling schemes. **All** denote all potential comparisons, **DS** denotes down-sapling, and **SSR** Strong Source Resampling algorithm

their results into a final prediction [5]. By applying bootstrap sampling, the authors approximate the ideal scenario of having independent base learners that when combined would exhibit an increased performance [49]. Our approach is similar in that we try to build weak independent learners by taking a source-aware sampling approach that respects the data's hierarchical structure of sources and items.

Let A denote the original sample from the population of interest. We assume that information about items has been collected, their features created, and the information about their sources is known with certainty. Our initial proposed approach starts by constructing all potential pairwise comparisons. We denote this as the candidate pool of comparison, and the set of independent comparisons can be constructed by iterative removing sources selected in the previous step. The pseudo-code in Algorithm 1 illustrates our approach.

Algorithm 1 Strong Source Resampling Algorithm (SSR)

Construct all pairwise comparisons available.
while $n_{KM} > 0$ & $n_{KNM} > 0$ **do**
Sample randomly one KM pair to be used in the comparison set.
Remove all pairs in the dataset involving sources selected in the previous step from the candidate pool
Sample randomly one KNM pair to be used in the comparison set.
Remove all pairs in the dataset involving sources selected in the previous step from the candidate pool
end while

The pseudo-code depicted in 1, while didactic, implies more computing time if the sample is larger. A faster approach can be implemented as in Algorithm 2 if there are at least two items per source. This approach is not only more efficient but also avoids the step of constructing all possible comparisons in the data, which can become challenging as the data increases.

Algorithm 2 Fast Strong Source Resampling Algorithm

Split sources into three sets: set 1, set 2, and set 3.
For each source in set 1, sample two items. This will generate the KM pairs.
For each source in set 2, sample one item.
For each source in set 3, sample one item.
Pair the items from two previous steps to generate the KNM pairs.

The result of both algorithms is a new set A^* that will be used in the following steps. While SSR and its faster version can be used interchangeably, we will use the second implementation for computational efficiency.

The ensemble learning literature has acknowledged that having more diverse learners results in better model performance under certain conditions. As summarized by Zhou [49], base learners should be both diverse and accurate simultaneously. While we cannot make any claims regarding the necessary conditions that would result in increased performance for ensembled SLR, since it would depend on the density estimation method used and if a scoring function is trained, we provide some back-of-the-envelope calculations in Appendix 2.10 to support some intuition on the effect of data composition in terms of the numbers of sources and items.

In a broad sense, the resampling plan will be mainly constrained by the total number of available sources. The total number of learning instances is limited to $4 \times \lfloor m/3 \rfloor$. The number of items within sources influences the proportion of original instances used, hence the diversity of newly generated data.

2.4.3.2 Base Score Likelihood Ratio and ensembling

The proposed resampling step is the cornerstone of our ensemble approach, and we use the algorithm to create sets where the independence assumption is met to develop a "base score likelihood ratio" (BSLR). The BSLRs fulfill the same role as weak learners in ensemble learning. Algorithm 3 presents the pseudo-code for our approach. If the (dis)similarity metric requires a training stage, a training set is generated using SSR and the metric developed. For the estimation stage, the SSR is also used to generate a data set to study the conditional densities or compute a ratio estimator. In both stages, the data is generated independently.

In the metric training stage, A^* will provide a balanced independent data set of size $4 \times \lfloor m/3 \rfloor$, where sources are used to estimate $\hat{\Delta}^{(m)}$ a new (dis)similarity metric. To study the conditional distribution of the scores (Equation 2.6), a new set A^* is generated, and predictions are made for known matches and known non-matches: $\left\{\hat{\delta}_i^{(KM)}\right\}_{i=1}^{2\times\lfloor m/3 \rfloor}$ and $\left\{\hat{\delta}_j^{(KNM)}\right\}_{j=1}^{2\times\lfloor m/3 \rfloor}$.

These predictions are used to estimate the conditional distribution of the score or a density ratio estimator.

The combination of the (dis)similarity metric and its estimated conditional distribution (or ratio) constitute a single BSLR. This procedure is repeated M times, resulting in M BSLR systems that can be aggregated into a final score; hence our approach can be classified as a parallel ensemble method.

Algorithm 3 Ensemble Score Likelihood Ratio (ESLR) System

```
for m=1:M do
    if Δ(·,·) requires training. then
        Use SSR to generate a pseudo training set.
        Train a comparison metric.
    end if
    Use SSR to generate a pseudo estimation set.
    Predict a comparison score for all cases of the estimation set.
    Estimate conditional densities (or ratio estimator).
    Store comparison metric and estimated densities or ratio estimator.
end for
```

Each BSLR is a function that maps from the features space to $[0, \infty)$. For our work, we consider the base-ten logarithm of the SLR meaning, $\text{BSLR}(u_x, u_y) : (\mathbb{R}^+)^{P+1} \mapsto \mathbb{R}$. We consider three naive approaches to combine their information into a final score: mean, median, and majority vote [3].

The mean ESLR consists of averaging the M numeric outputs into a final value,

$$\text{Mean. ESLR} = \frac{1}{M} \sum_{i=1}^{M} \text{BSLR}^{(i)}(u_x, u_y) \tag{2.10}$$

To reduce the effects of outliers, we explore the median ESLR as an aggregator,

$$\text{Median. ESLR} = \text{median} \left\{ \text{BSLR}^{(i)}(u_x, u_y) \right\}_{i=1}^{M} \tag{2.11}$$

[3] An extra optimization steps can be done to assign differential weights to each base learner and will be addressed in the future.

Our third aggregator, the majority vote, considers the output of each BSLR and maps them into a verbal scale that reflects the strength of the evidence. The most voted category is considered the final outcome. For our work, we consider a ten-level verbal category based on the log10 scale presented in Evett et al. [16].

Let $\text{BSLR}_C(\cdot,\cdot)$ denote a base SLR system where the $\log_{10}$ output has been categorized into one of the ten verbal categories,$\text{BSLR}_C(u_x,u_y):(\mathbb{R}^+)^{P+1}\mapsto\{\mathcal{B}_{10}\}$, where $\mathcal{B}_{10}$ denotes the ten verbal categories[4]:

$$\mathcal{B}_{10}\equiv\{(-\infty,-4),(-4,-3),(-3,-2),(-2,-1),(-1,0),(0,1),(1,2),(2,3),(3,4),(4,\infty)\} \tag{2.12}$$

We consider V. ESLR to denote the aggregated output using majority voting,

$$\text{V. ESLR}=\text{majority vote}\left\{\text{BSLR}_C^{(i)}(u_x,u_y)\right\}_{i=1}^{M} \tag{2.13}$$

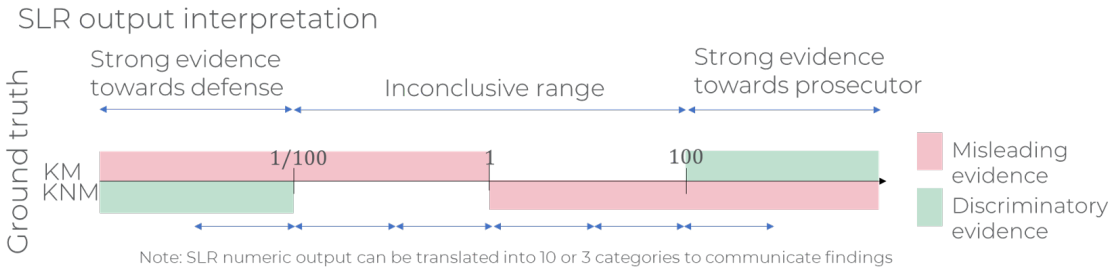
2.4.4 Evaluation Metrics for SLRs

SLRs provide an alternative way to present information to jurors and judges in a criminal case. In this context, specific performance characteristics that may not be part of the traditional machine learning toolbox are more relevant. In a criminal case, we are interested in evaluating if the information provided will lead jurors in the correct direction, obtain a measure of the size of the error committed, and if the evidence presented is strong enough. In addition, we would aim to develop reliable methods. If the training and estimation sets are altered, the conclusion reached for the same validation pair should be similar. Several popular metrics associated with these performance characteristics are discussed in [29], and we present detailed notation in Appendix 2.9.

In the case of an SLR system, a value larger than one (zero in the log10 scale) indicates that the evidence supports the prosecutor's proposition, while the opposite indicates that the evidence

[4] A more detailed discussion of the categories, and their verbal qualifier can also be found in [18, 47]

Figure 2.2 Forensic confusion matrix



supports the defense's. In the case of a known match (known non-match), evidence toward the defense (prosecutor) is considered misleading. Over the validation set, we compute the rate of misleading evidence for known matches (RME_{KM}) as the proportion of cases where the prosecutor proposition is correct, but the system indicates the opposite, and the rate of misleading evidence for known non-matches (RME_{KNM}) as the proportion of cases where the defense proposition is correct but the system output supports the prosecutor.

To bridge the gap in nomenclature between forensic and machine learning performance metrics, we present an extension of the confusion matrix for classification problems (Figure 2.2). If the KM are considered "positive", the rate of misleading evidence for KM is, in essence, the false negative rate, and the rate of misleading evidence for KNM is the false positive rate. However, the difference in nomenclature arises because there is no clear assignment of the positive label in criminal justice.

SLR systems should also provide strong probative value in the correct direction. This translates to outputting "large" ("small") values for known match pairs (known non-matches) above (below) a given threshold C_{KM} (C_{KNM}). While there is no consensus regarding these thresholds, we illustrate this performance characteristic by taking $C_{KM} = 100$ and $C_{KNM} = 1/100$, defining three previously used regions in the literature [21]. Figure 2.2, depicts these regions that can be interpreted as strong evidence for the defense, inconclusive, and strong evidence for the prosecutor. These categories are a less granular version of the 10-category scale introduced for the majority voting scheme.

As an aggregated performance measure, we consider the log-likelihood-ratio cost function (C_{llr}) [30]. This cost function penalizes strong conclusions in the incorrect direction, resulting in smaller values of the metric being associated with better-behaved systems.

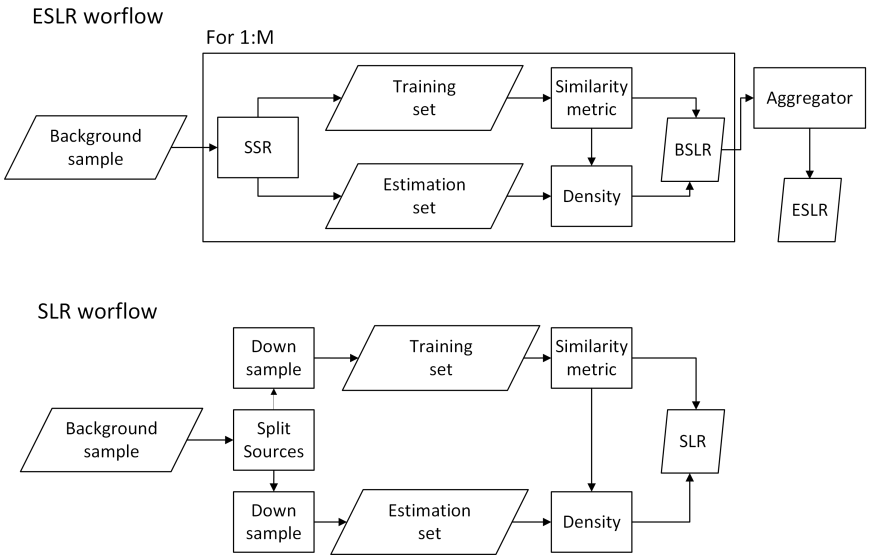
To assess reliability, consider the case of S systems developed using the same methodology, but the learning data available is modified somehow. The variability in our problem is due to splitting the sources and downsampling in the traditional SLR approach and the SSR step in our proposed Ensemble SLR approach. To assess the method's reliability, we trained multiple systems (SLR and ESLR) and compared their conclusions when faced with the same holdout evidence using a distance-based and a consensus metric (Section 2.4.5 provides additional details about the simulation strategy). In the case of a system that outputs a numeric value, we consider the average Euclidean distance in the log10 scale to the mean evidence value obtained for pair t across the S systems. A better-behaved system would present less variability in its results, associated with a smaller average distance.

For systems that output a categorical value or if their results are mapped into the categorical scale, we consider a consensus metric that penalizes more heavily methods that generate more polarized results by considering the ordinal nature of the ten-level verbal categories. Additional details about this metric and its computation are presented in Appendix 2.9.

2.4.5 Simulation strategy

To illustrate our approach, we focus on forensic handwriting data under the common source problem. We used CSAFE's London Letter as the reference sample and the CVL data to construct validation sets; we considered a Random Forest as a similarity metric and a logit-based density ratio estimator (Section 2.4.2). In each iteration of our experiment, we develop a traditional SLR - which will serve as our baseline - by splitting the sources into training and estimation sets and applying a down-sampling step to obtain a balanced sample. We use Strong Source Resampling (SSR) in each iteration to construct 50 base SLRs ($M = 50$) that will be aggregated into a final Ensemble SLR. We present three aggregation approaches: mean, median, and majority vote. We

Figure 2.3 Proposed workflows



introduce two diagrams in Figure 2.3 to summarize and compare our proposed workflow. A new validation set was generated for each iteration by sampling 1000 known matches and 1000 known non-matches from the CVL data set, and used to compute performance metrics for the four proposed methods (SLR, Mean.ESLR, Median.ESLR, V.ESLR). We repeated this process five hundred times to obtain a sample of performance metrics. We consider this our first experiment.

To evaluate the agreement, we run a second experiment. Using the same ensembled SLR and SLRs trained for the first experiment, we held the validation set for the first iteration fixed across iterations. If the method is reliable, we expect to see the same conclusion reached for the same pair in the validation set, even if the training and estimation data is changed.

2.5 Results

We present descriptive statistics for the performance metrics across iterations in Table 2.1 for experiment 1. Traditional SLR delivers a higher rate of misleading evidence for the KM on average than the ensemble approaches (approximately two percentage points more), with the mean ensemble providing the lowest rate. However, our ensemble approach presents about half a percentage point higher rate of misleading evidence for known non-matches. Figure 2.4 further

illustrates the distribution of these performance metrics and shows that their distribution is slightly more spread for traditional SLR.

Regarding strong evidence or discriminatory power, mean and median Ensembled SLR presented larger discriminatory power statistics for known and known-non matches. On average, the mean ensemble achieves almost twice the discriminatory power compared to its traditional counterpart. The majority voting presented smaller discriminatory power for known matches but was on par with the median aggregation for known non-matches, above the conventional approach.

As before, we present Figure 2.5 to illustrate the distribution of the performance metrics across iterations of our experiment. Traditional SLR failed to achieve a positive value for discriminatory power in most iterations, the average being driven by outliers in the case of Known Matches. The mean aggregator may be subject to a similar issue, although to a smaller degree. The median and majority voting achieve similar performance more consistently. This result suggests that while ensembling can improve performance, some aggregators are more robust.

The last lines of Table 2.1 present the cost functions, the aggregated performance metric for SLR systems that output a numeric value. The mean ensemble performs better than its median and traditional counterparts, achieving smaller costs on average. This reduction seems to be due to lower costs for known matches (Figure 2.6).

Table 2.2 presents descriptive statistics for the performance metric for experiment 2, while Figures 2.8 and 2.7 provide additional information on their distributions. In terms of the average distance in the log10 scale, the ensemble methods are associated with smaller values, indicating that the numeric conclusion reached tends to be more similar for the ensemble than the traditional SLR approach. The consensus statistic provides a numeric summary of the level of agreement in the verbal scale and allows for comparing all aggregators. A value of one indicates a perfect agreement, while smaller values indicate deviations. Our ensemble SLRs tend to achieve consensus more consistently than the traditional approach. The classic SLR presents smaller consensus metrics on average, and the results are less concentrated than our ensemble approach.

Figure 2.4 Rate of Misleading Evidence by Match

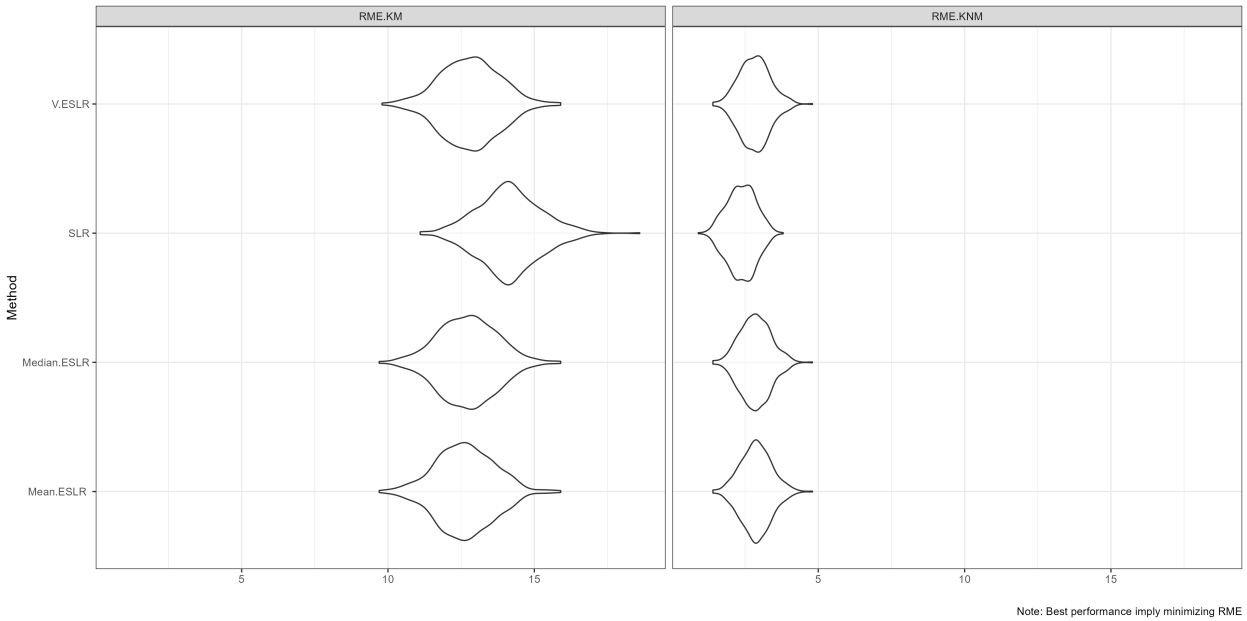


Figure 2.5 Discriminatory Power by Match

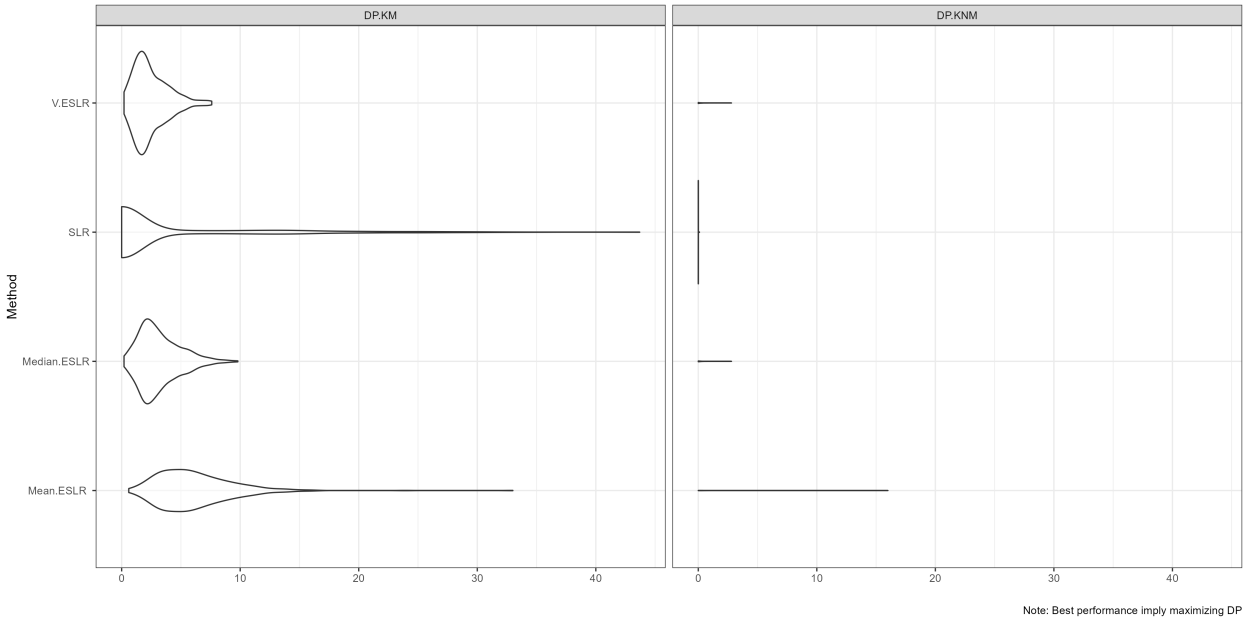
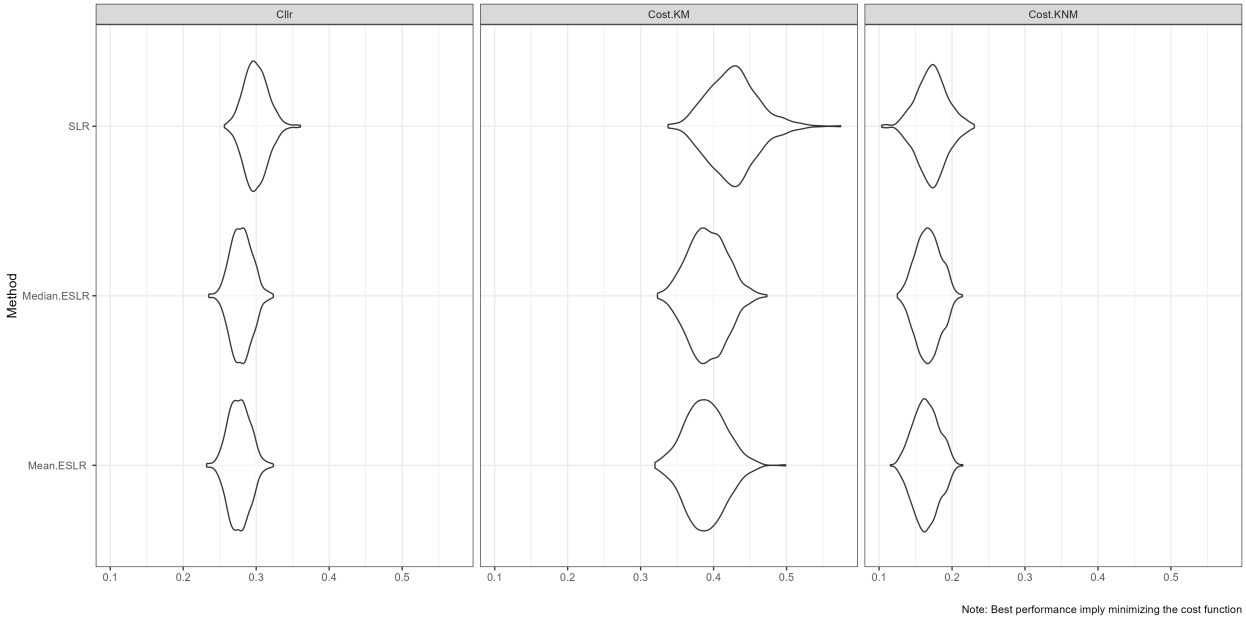


Figure 2.6 C_{llr} and Cost by Match



2.6 Conclusions

Machine learning-based SLRs are gradually playing a more relevant role in the forensic statistics community as an alternative to feature-based likelihood ratios to compute the probative value of evidence.

Under the common source problem, the current procedure to generate training an estimation set relies on creating pairwise comparisons from a sample from the background population. These comparisons are used as instances for statistical learning; however, they can't be considered independent since items (and source) are used in multiple comparisons.

Independence is a common assumption made in popular machine learning and density estimation methods, both cornerstones in developing machine learning-based SLRs. Our work introduces a sampling algorithm to remediate the complex dependence structure in the common source problem. The sampling step can be used as pre-processing to create new samples that will serve as training and estimation sets. While we are unable to provide sufficient conditions where the resampling will guarantee an improvement, Appendix 2.10 provides some intuition on the

Figure 2.7 Average distance in the log10 scale

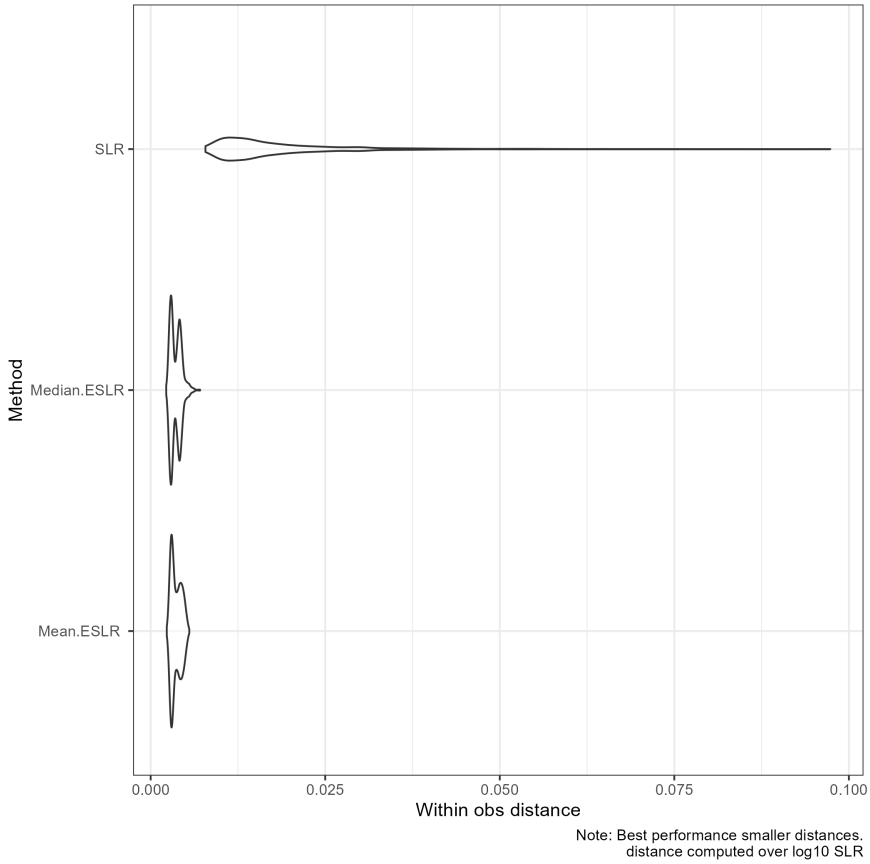
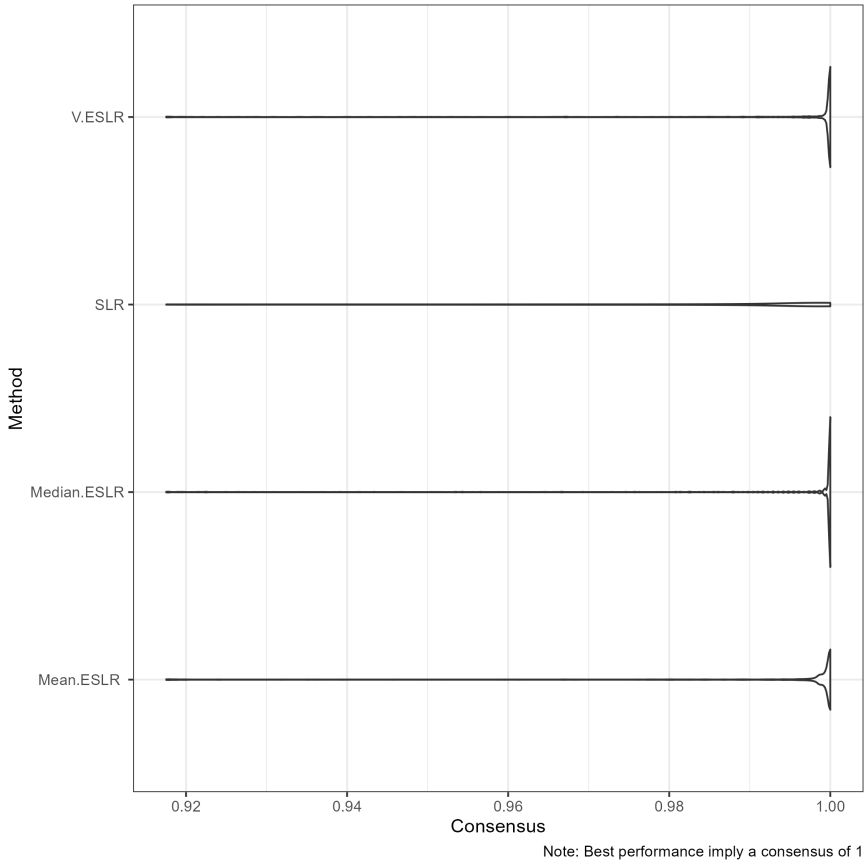


Figure 2.8 Consensus metric



effects of the composition of the original sample. Increasing the number of sources, rather than items within sources, will contribute to more independent learning instances. The number of items within sources contributes to a lesser extent, increasing the diversity of the samples generated. This intuition could potentially be used to guide data collection efforts in the future.

We propose the use of our sampling algorithm as a resampling plan to generate multiple base SLRs that serve as weak learners, learning from a partial view of the data where assumptions are met, and aggregate their outputs into a combined result we denoted as an Ensembled SLR.

Our simulation result suggests that our sampling and ensemble approach is not detrimental to SLR systems; ensemble learning can enhance the performance of traditional SLRs. For the handwriting data used in our experiment, ESLRs presented more discriminatory power and reduced the rate of misleading evidence for known matches at the cost of a slight increase in the rate of misleading evidence for known non-matches.

We explored three aggregation methods: mean, median, and majority vote, to combine base SLRs. Our result suggests that the aggregating method is relevant to the performance metrics. While the mean ESLR presented better results, the median and majority voting aggregators achieved comparable results and did so more consistently. The current aggregation methods followed a similar spirit as bagging but can be considered naïve. In future work, we plan to explore assigning differential weights to each base learner according to their performance on an optimization set or using a sequential procedure, analogous to boosting.

Our secondary concern was assessing if an ensemble approach could result in more stable conclusions for the same hold-out sets, as traditional SLRs are sensitive to perturbation in the training and estimation sets [47].

To assess this, we performed a second experiment which showed that the ensemble approach could provide more consistent conclusions for the same hold-out evidence. This is a less studied characteristic of the SLRs system but is highly relevant for criminal justice.

While our work illustrates ensemble learning to improve traditional Score Likelihood Ratio systems for forensic handwriting, our approach is not limited to handwriting analysis or problems

in the forensic sciences. It could be feasibly applied to other domains for the common source problem or situation where learning instances are generated based on pairwise comparisons.

2.7 References

[1] Aitken, C. G. G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley and Sons, Ltd., West Sussex, UK, 2nd edition.

[2] Báez-Santiago, F., Lundstrom, J., Crawford, A., Berry, N., Escobar, B., Taylor, J., Reinders, S., and Ommen, D. (2021). Handwriter: An r package for statistical writership analysis.

[3] Bolck, A., Ni, H., and Lopatka, M. (2015). Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk*, 14(3):246–266.

[4] Bolck, A., Weyermann, C., Dujourdy, L., Esseiva, P., and van den Berg, J. (2009). Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International*, 191(1):42 – 51.

[5] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24:123–140.

[6] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[7] Chawla, N. V. (2005). *Data Mining for Imbalanced Datasets: An Overview*, pages 853–867. Springer US, Boston, MA.

[8] Chen, X.-H., Champod, C., Yang, X., Shi, S.-P., Luo, Y.-W., Wang, N., Wang, Y.-C., and Lu, Q.-M. (2018). Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features. *Forensic science international*, 282:101—110.

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

[10] Council, N. R. et al. (2009). *Strengthening forensic science in the United States: a path forward*. National Academies Press.

[11] Crawford, A., Ray, A., and Carriquiry, A. (2020). A database of handwriting samples for applications in forensic statistics. *Data in brief*, 28:105059.

[12] Crawford, A. M., Berry, N. S., and Carriquiry, A. L. (2021). A clustering method for graphical handwriting components and statistical writership analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(1):41–60.

[13] Davis, L. J., Saunders, C. P., Hepler, A., and Buscaglia, J. (2012). Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. *Forensic science international*, 216(1-3):146–157.

[14] Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.

[15] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

[16] Evett, I., Jackson, G., Lambert, J., and McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, 40(4):233–239.

[17] Galbraith, C. and Smyth, P. (2017). Analyzing user-event data using score-based likelihood ratios with marked point processes. *Digital Investigation*, 22:S106 – S114.

[18] Garton, N., Ommen, D., Niemi, J., and Carriquiry, A. (2020). Score-based likelihood ratios to evaluate forensic pattern evidence. *arXiv preprint arXiv:2002.09470*.

[19] Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., and Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech & Language*, 20(2):331 – 355. Odyssey 2004: The speaker and Language Recognition Workshop.

[20] Hendricks, J., Neumann, C., and Saunders, C. P. (2021). Quantification of the weight of fingerprint evidence using a ROC-based Approximate Bayesian Computation algorithm for model selection. *Electronic Journal of Statistics*, 15(1):1228 – 1262.

[21] Hepler, A. B., Saunders, C. P., Davis, L. J., and Buscaglia, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic science international*, 219(1-3):129–140.

[22] Ishihara, S. and Carne, M. (2022). Likelihood ratio estimation for authorship text evidence: An empirical comparison of score-and feature-based methods. *Forensic Science International*, 334:111268.

[23] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *Resampling Methods*, pages 197–223. Springer US, New York, NY.

[24] Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

[25] Johnson, M. Q. and Ommen, D. M. (2022). Handwriting identification using random forests and score-based likelihood ratios. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3):357–375.

[26] Kleber, F., Fiel, S., Diem, M., and Sablatnig, R. (2013). Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In *2013 12th international conference on document analysis and recognition*, pages 560–564. IEEE.

[27] Lander, E. S., Group, P. W., et al. (2016). Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods.

[28] Leegwater, A. J., Meuwly, D., Sjerps, M., Vergeer, P., and Alberink, I. (2017). Performance study of a score-based likelihood ratio system for forensic fingermark comparison. *Journal of Forensic Sciences*, 62(3).

[29] Meuwly, D., Ramos, D., and Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic science international*, 276:142–153.

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

[31] Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2):173–197.

[32] Morrison, G. S. and Enzinger, E. (2018). Score based procedures for the calculation of forensic likelihood ratios–scores should take account of both similarity and typicality. *Science & Justice*, 58(1):47–58.

[33] Neijmeijer, R. (2016). Assessing performance of score-based likelihood ratio methods for forensic data. Master's thesis, Leiden University.

[34] Neumann, C. and Ausdemore, M. (2020). Defence against the modern arts: the curse of statistics—Part II: 'Score-based likelihood ratios'. *Law, Probability and Risk*, 19(1):21–42.

[35] Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., Meuwly, D., and Bromage-Griffiths, A. (2006). Computation of likelihood ratios in fingerprint identification for configurations of three minutiæ. *Journal of Forensic Sciences*, 51(6):1255–1266.

[36] Neumann, C. and Margot, P. (2009). New perspectives in the use of ink evidence in forensic science: Part iii: Operational applications and evaluation. *Forensic Science International*, 192(1):29–42.

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

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

[39] Osborn, A. S. (1929). *Questioned Documents, 2nd edn.* Boyd Printing Company, New York, NY.

[40] Park, S., Carriquiry, A., et al. (2019). Learning algorithms to evaluate forensic glass evidence. *Annals of Applied Statistics*, 13(2):1068–1102.

[41] Park, S. and Tyner, S. (2019). Evaluation and comparison of methods for forensic glass source conclusions. *Forensic science international*, 305:110003.

[42] Reinders, S., Guan, Y., Ommen, D., and Newman, J. (2022). Source-anchored, trace-anchored, and general match score-based likelihood ratios for camera device identification. *Journal of Forensic Sciences*, 67(3):975–988.

[43] Stern, H. S. (2017). Statistical issues in forensic science. *Annual Review of Statistics and Its Application*, 4:225–244.

[44] Sugiyama, M., Suzuki, T., and Kanamori, T. (2010). Density ratio estimation: A comprehensive review (statistical experiment and its related topics). *RIMS Kokyuroku*, 1703:10–31.

[45] Tastle, W. J. and Wierman, M. J. (2006). An information theoretic measure for the evaluation of ordinal scale data. *Behavior Research Methods*, 38(3):487–494.

[46] Tastle, W. J. and Wierman, M. J. (2007). Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning*, 45(3):531–545.

[47] Veneri, F. and Ommen, D. (2021). An evaluation of score-based likelihood ratios for glass data. Master's thesis, Iowa State University.

[48] Willis, S., Aitken, C., Barrett, A., Berger, C., Biedermann, A., Champod, C., Hicks, T., Lucena-Molina, J., Lunt, L., McDermott, S., McKenna, L., Nordgaard, A., O'Donnell, G., Rasmusson, B., Sjerps, M., Taroni, F., and Zadora, G. (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science.* European Network of Forensic Science Institutes, http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

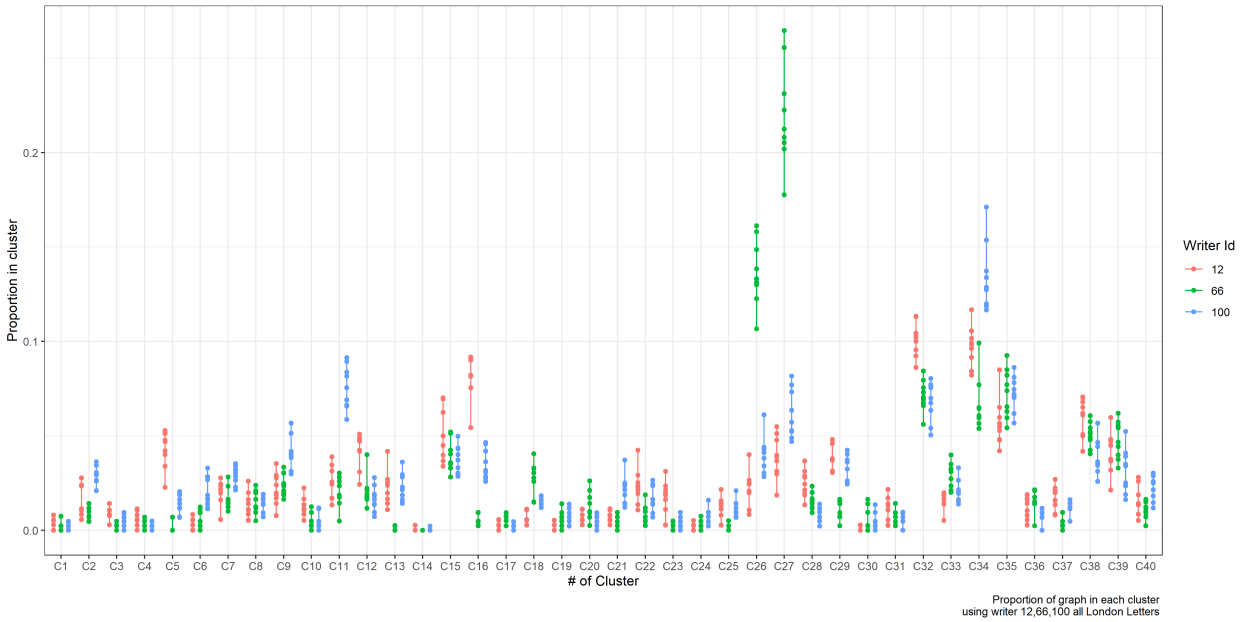
[49] Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms.* CRC press.

[50] Zhu, X., Tang, L., and Tabassi, E. (2017). Repeatability and reproducibility of forensic likelihood ratio methods when sample size ratio varies. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 517–524. IEEE.

2.8 Appendix A: Select descriptive statistics and figures

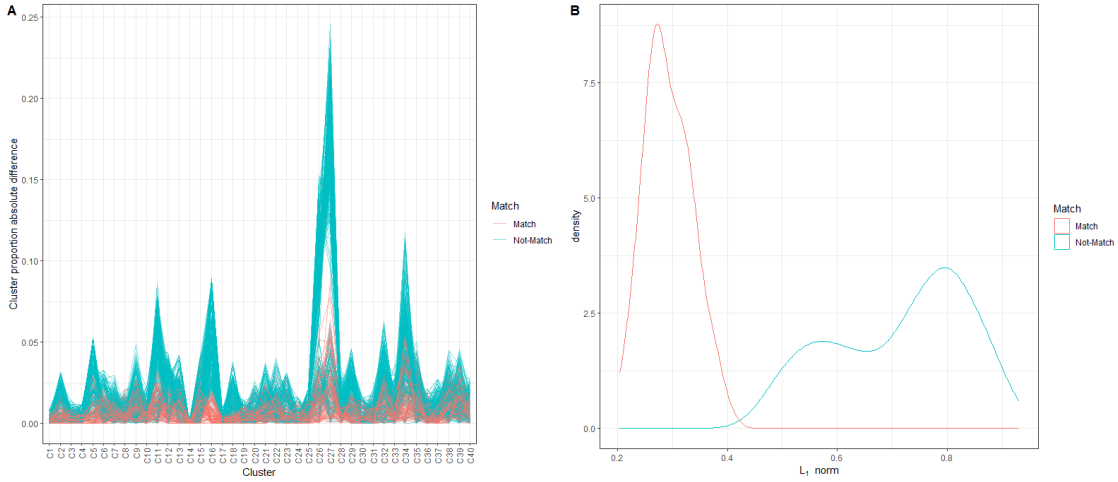
This appendix presents the reader with selected figures to illustrate the results of the feature creation process. Figure 2.9 presents the raw cluster proportions for a subset of writers in the CSAFE-LND data set to illustrate how these features can be used to characterize writers. Each dot represents the cluster proportion for a particular sample writing for one of three writers: writers 12, 66, and 100. Even if some variability exists within writers, the difference between them could be considered a relevant feature for the common source problem. While writer 66 (depicted in green) tended to write characters more frequently assigned to clusters 26 and 27, writer 100 tended to write characters in clusters 34 and 11 relative to the other writers selected.

Figure 2.9 Raw cluster proportion for selected writers



Using the same writers and their prompts, the first panel in figure 2.10 depicts features created using all possible pairwise comparisons for the subset of writers as described in Section 2.4.2. Different source cases (KNM) present larger absolute differences than comparisons from the same source (KM) across the different clusters. The second panel depicts the L_1 distance as an aggregated feature used in the comparisons. Comparison from known non-matches exhibits larger distances than pairs from the same source.

Figure 2.10 Features for known matches and known non-matches for selected writers



2.9 Appendix B: Performance metrics in Forensic science

This appendix presents the interested reader with the notations and formulas for popular performance metrics in forensic statistics and introduces the use of consensus metrics and distance for measuring reliability in forensics. Notation is adapted from [47]. Assume that the validation set consists of T pairs of items, n_{KM} being from the same source and n_{KNM} from different sources ($T = n_{KM} + n_{KNM}$). Let SLR_t denote the output of the SLR or ESLR system for pair t and let y_t a numeric label for pair t, same source or known match ($y_t = 1$) and different source or known non-match ($y_t = 0$). Also, let $1_{\{\}}$ denote the indicator function.

The first performance metric refers to the rate of misleading evidence, which will be independently computed for known and known non-matches. A value smaller than one in the real scale (or smaller than zero in the log10 scale) would be considered misleading for known matches, and the opposite holds for known non-matches. Using the previously introduced notation, we compute the rates as follows:

$$RME_{KM} = \frac{\sum_t^T y_t 1_{\{SLR_t<1\}}}{\sum_t^T y_t} = \frac{\sum_t^T y_t 1_{\{SLR_t<1\}}}{n_{KM}} \tag{2.14}$$

$$RME_{KNM} = \frac{\sum_t^T (1-y_t) 1_{\{SLR_t>1\}}}{\sum_t^T (1-y_t)} = \frac{\sum_t^T (1-y_t) 1_{\{SLR_t>1\}}}{n_{KNM}} \tag{2.15}$$

The discriminatory power aims to answer if the system is providing "strong" evidence, with strong referring to a pre-established threshold. In the case of a known match, ideally, SLR_t would be larger than C_{KM}, and for known non-matches, we would expect to see SLR_t smaller than C_{KNM}. The discriminatory power can be computed as:

$$D_{pKM} = \frac{\sum_t^T y_t 1_{\{SLR_t \geq C_{KM}\}}}{\sum_t^T y_t} = \frac{\sum_t^T y_t 1_{\{SLR_t \geq C_{KM}\}}}{n_{KM}} \tag{2.16}$$

$$D_{pKNM} = \frac{\sum_t^T (1-y_t) 1_{\{SLR_t \leq C_{KNM}\}}}{\sum_t^T (1-y_t)} = \frac{\sum_t^T (1-y_t) 1_{\{SLR_t \leq C_{KNM}\}}}{n_{KNM}} \tag{2.17}$$

As in Hepler et al. [21], we set $C_{KM} = 100$ and $C_{KM} = 1/100$ as cut-off values for illustration purposes. These values define three regions that can be considered as strong support for the defense, inconclusive and strong support for the prosecutor (Figure 2.2).

The previous metrics provide a partial view of the model's performance in terms of correct direction and strength of conclusions. An aggregated measure is given by the Log-Likelihood-Ratio Cost Function ($Cllr$). The $Cllr$ is a popular aggregated metric used in forensic voice comparisons [30]. The function introduces an increasing penalization for conclusions leading to wrong conclusions with stronger output values. As in previous metrics, the cost functions are computed for known and known not matches, but the average is presented as a combined total cost metric.

$$Cllr_{KM} = \frac{1}{n_{KM}} \sum_{t=1}^{T} y_t log_2 \left(1 + \frac{1}{SLR_t}\right) \tag{2.18a}$$

$$Cllr_{KNM} = \frac{1}{n_{KNM}} \sum_{t=1}^{T} (1-y_t) log_2 (1 + SLR_t) \tag{2.18b}$$

$$Cllr = \frac{1}{2}(Cllr_{KM} + Cllr_{KNM}) \tag{2.18c}$$

In terms of interpretation, a smaller cost is associated with a better-performing system.

Lastly, we computed a distance-based and a consensus metric to assess the reliability of the methods. Let SLR_{ts} be the log10 evidence value obtained for the pair t and system s ($s = 1, \ldots, S$), and let d_t the within-pair average distance for comparison t, computed as

$$d_t = \frac{1}{S} \sum_{s=1}^{S} (SLR_{ts} - \overline{SLR}_{t.})^2 \tag{2.19}$$

where $\overline{SLR}_{t\cdot} = \frac{1}{S}\sum_{s=1}^{S} SLR_{ts}$ denotes the mean value obtained for the pair t across the S systems. Since the evidence under consideration remains the same across SLR systems, more reliable methods would output similar values; hence a smaller within-evidence distance would indicate that the conclusions reached in the log10 scale are similar.

In the case of a system that outputs verbal categories (or their outputs are mapped into verbal categories), specific performance metrics can be used to measure agreement. Let SLR^{*}_{Cts} denote the output expressed in one of the ten level verbal categories of $\mathcal{B}_{10}$ for pair t and system s, where consecutive numeric values were attached to each category as in a Likert-type scale. For instance, consider the case of scale ranging from Very Strong evidence for the defense ($SLR^{*}_{Cts} = 1$) to very strong evidence for the prosecutor ($SLR^{*}_{Cts} = 10$). In this context, SLR_{Cst} can be considered as a discrete variable, and its entropy $H(t)$ or consensus Consensus(t) can be computed for a pair t which has been assessed over S systems. This information theory metric has been previously discussed in [45, 46]. The Shannon entropy is given by

$$H(t) = \sum_{s=1}^{S} p(SLR^{*}_{Cts}) \log_2 p(SLR^{*}_{Cts}) \tag{2.20}$$

where $p(SLR^{*}_{Cts})$ is the probability of observing the particular category for pair t across the S systems.

While the Shannon entropy would give a measurement of the agreement of the conclusions reached by the different systems when faced with the same pair t, a value of zero associated with a complete agreement, it does not account for the ordinal nature of the conclusion. Tastle and Wierman [45, 46] proposed a consensus metric to account for the ordering by computing:

$$\text{Consensus}(t) = 1 + \sum_{s=1}^{S} p(SLR^{*}_{C_{ts}}) \log_2 \left(1 - \frac{\left|SLR^{*}_{C_{ts}} - \overline{SLR^{*}}_{C_{t\cdot}}\right|}{\text{range}(SLR^{*}_{C})}\right) \tag{2.21}$$

where $\overline{SLR^{*}}_{C_{t\cdot}}$ and $p(SLR^{*}_{Cts})$ denote the mean and probability as before, and the range(SLR^{*}_{C}) indicates the range of the mapping used to assign numeric values to the verbal scale. The consensus metric is upper bounded by one, in the case of a total agreement, and lower bounded by zero, in the case of polarization. To compare Shannon's entropy and the consensus metric, it is useful to consider a dissent metric (Consensus(t) = 1 − Dissent(t)), where a value of zero is

associated with complete agreement. Figure 2.11 presents ternary plots for simulated proportions using the shorter, three-level scale for illustration purposes.

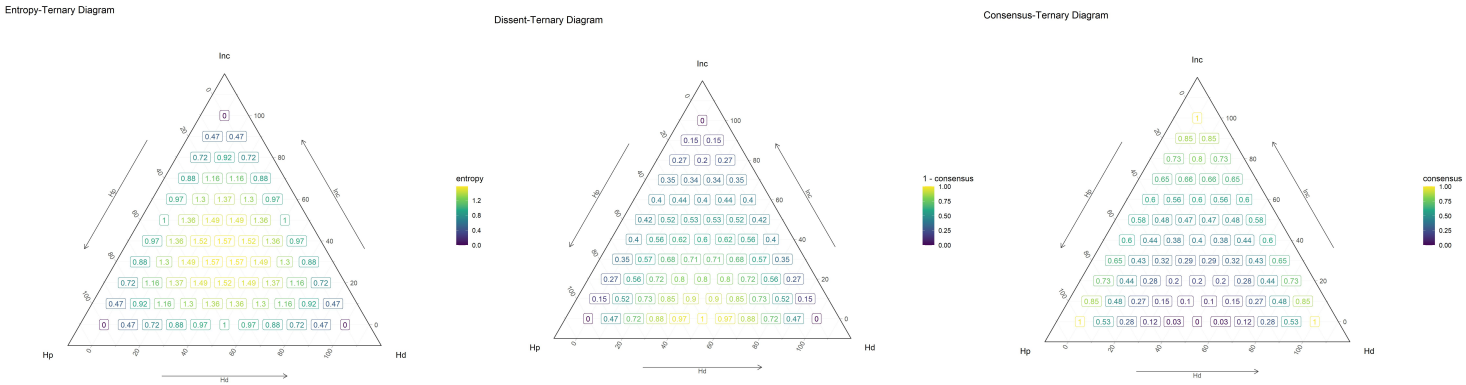
The entropy metric returns higher values, associated with a larger disagreement, along the center of the diagram. This area in the diagram is associated with a more uniform distribution between the three potential categories. In the case of the dissent metric, higher values are associated with more polarized results in the lower part of the diagram, with half the systems outputting values for the defense and the other half towards the prosecutor, not the uniform case. For our main results, we used the consensus metric (last panel of Figure 2.11). A value of one is associated with complete agreement, represented in the vertices of the ternary diagram, where all systems developed agree in their categorical conclusion.

2.10 Appendix C: Sample size constraints

This appendix presents some intuition on how the original data composition in terms of sources and items could affect the performances of the ensemble approach for score likelihood ratios. Let m denote the number of sources in the set A, with n items available for each source. The total number of pair is given by $\binom{mn}{2}$, the total number of known match pair $n_{KM} = \binom{n}{2}m$ and the total number of known non-matches $\binom{mn}{2} - \binom{n}{2}m = n_{KNM}$.

For the calculation, we used our second implementation of the resampling algorithm (Algorithm 2). Under this implementation, the number of sources is split into thirds, $m^* = \lfloor m/3 \rfloor$. For known matches, two items are sampled randomly from n without replacement; hence, the total number of KM learning instances is $2 \times m^*$. For known non-matches, one item is sampled randomly from n for each selected source; hence, the total number of KNM learning instances is also $2 \times m^*$. The new learning sample is balanced and of size $4 \times m^*$. The resulting percentage of data use will differ in terms of known and not known matches. Table 2.3 presents different selected scenarios to illustrate these results. In summary: the number of sources will affect the total number of learning instances used in the sample, while the number of items within

Figure 2.11 Comparison of information metrics to asses agreement



sources will impact on the diversity of the sample as measured by the percentage of cases from matches used.

In all our scenarios, less than 2% of the original data is used. Bootstrap sampling provides a natural benchmark in this regard. For large sample size, the number of times an original case is selected is distributed Poisson($\lambda = 1$), leading to the conclusion that approximately 36.8% of the learning instances is not used to train a particular weak learner [5], hence the weak learners can be more variable under Strong Source Resampling than bootstrap.

Table 2.1 Performance Metrics - Experiment 1

Metric	Statistic	SLR	Mean ESLR	Median ESLR	V. ESLR
RME KM	Mean	14.1450	12.6382	12.7242	12.8096
	Median	14.1000	12.6000	12.7000	12.8000
	Sd	1.0818	0.9984	1.0122	1.0155
RME KNM	Mean	2.3804	2.8772	2.8534	2.8280
	Median	2.4000	2.9000	2.9000	2.8000
	Sd	0.4877	0.5248	0.5244	0.5240
DP KM	Mean	3.1854	6.1000	3.1776	2.4538
	Median	0.0000	5.5000	2.8000	2.1000
	Sd	6.9566	3.2327	1.7540	1.4937
DP KNM	Mean	0.0002	0.9438	0.2806	0.2562
	Median	0.0000	0.5500	0.1000	0.1000
	Sd	0.0045	1.4185	0.4165	0.3935
Cllr	Mean	0.2996	0.2768	0.2796	
	Median	0.2984	0.2767	0.2794	
	Sd	0.0163	0.0149	0.0146	
Cllr KM	Mean	0.4267	0.3892	0.3918	
	Median	0.4262	0.3886	0.3911	
	Sd	0.0341	0.0274	0.0261	
Cllr KNM	Mean	0.1725	0.1644	0.1674	
	Median	0.1727	0.1635	0.1670	
	Sd	0.0208	0.0170	0.0163	

Table 2.2 Performance Metric - Experiment 2

Metric	Statistic	SLR	Mean ESLR	Median ESLR	V.ESLR
Conensus	Mean	0.9816	0.9916	0.9923	0.9922
(10 verbal scale)	Median	0.9948	1.0000	1.0000	1.0000
	Sd	0.0247	0.0197	0.0189	0.0190
Average Distance	Mean	0.0183	0.0036	0.0036	
	Median	0.0147	0.0035	0.0035	
	Sd	0.0104	0.0008	0.0008	

Table 2.3 Sample size implications

m	n	Total	n_{KM}	n_{KNM}	m^*	$4 \times m^*$	$\%n_{KM}$	$\%n_{KNM}$
10	10	4950	450	4500	3	12	1.3333	0.1333
10	20	19900	1900	18000	3	12	0.3158	0.0333
10	30	44850	4350	40500	3	12	0.1379	0.0148
10	50	124750	12250	112500	3	12	0.0490	0.0053
50	10	124750	2250	122500	16	64	1.4222	0.0261
50	20	499500	9500	490000	16	64	0.3368	0.0065
50	30	1124250	21750	1102500	16	64	0.1471	0.0029
50	50	3123750	61250	3062500	16	64	0.0522	0.0010
100	10	499500	4500	495000	33	132	1.4667	0.0133
100	20	1999000	19000	1980000	33	132	0.3474	0.0033
100	30	4498500	43500	4455000	33	132	0.1517	0.0015
100	50	12497500	122500	12375000	33	132	0.0539	0.0005
200	10	1999000	9000	1990000	66	264	1.4667	0.0066
200	20	7998000	38000	7960000	66	264	0.3474	0.0017
200	30	17997000	87000	17910000	66	264	0.1517	0.0007
200	50	49995000	245000	49750000	66	264	0.0539	0.0003
300	10	4498500	13500	4485000	100	400	1.4815	0.0045
300	20	17997000	57000	17940000	100	400	0.3509	0.0011
300	30	40495500	130500	40365000	100	400	0.1533	0.0005
300	50	112492500	367500	112125000	100	400	0.0544	0.0002

CHAPTER 3. SYNTHETIC ANCHORING UNDER THE SPECIFIC SOURCE PROBLEM

Federico Veneri and Danica M. Ommen

Statistics Department, Iowa State University

Modified from a manuscript submitted to *Statistical Analysis and Data Mining: The ASA Data Science Journal*

3.1 Abstract

Source identification is an inferential problem that evaluates the likelihood of opposing propositions regarding the origin of items. The specific source problem refers to a situation where the researcher aims to assess if a particular source originated the items or if they originated from an alternative, unknown source. Score-based likelihood ratios offer an alternative method to assess the relative likelihood of both propositions when formulating a probabilistic model is challenging or infeasible, as in the case of pattern evidence in forensic science. However, the lack of available data and the dependence structure created by the current procedure for generating learning instances can lead to reduced performance of score likelihood ratio systems. To address these issues, we propose a resampling plan that creates synthetic items to generate learning instances under the specific source problem. Simulation results show that our approach achieves a high level of agreement with an ideal scenario where data is not a limitation and learning instances are independent. We also present two applications in forensic sciences -handwriting and glass analysis- illustrating our approach with both a score-based and a machine learning-based score likelihood ratio system. These applications show that our method may outperform current alternatives in the literature.

3.2 Introduction

Source identification problems appear in multiple domains under different names. Assessing the likelihood that particular samples come from a specific batch in manufacturing, verifying whether a transaction was initiated by someone other than the authorized user, establishing if biometric data collected belong to a specific person, or if the person of interest's gun can be associated with a crime scene. All these examples can be considered source identification problems.

Regardless of the domain, source problems can be formalized as a statistical inferential problem where the researcher aims to contrast opposing propositions regarding the origin of some items (or data). Traditionally, in forensic statistics, these opposing propositions have been denoted as the prosecutor (H_p) and defense (H_d) propositions.

Depending on the content of propositions, the problem can be classified into common or specific source problems [33, 31, 42, 15]. In the context of common source problems, presented with items of unknown origin along with their associated features, the propositions seek to contrast what is more likely: that items were generated by the same unknown source (H_p) or that they were generated by two different unknown sources (H_d). Under the specific source problem, a comparison is made between items from unknown origin (recovered items) and items generated by the same known source (control items). The proposition addresses whether the recovered items were generated by the specific source associated with the control items (H_p) or if they were generated by an alternative unknown source (H_d).

When probabilistic models can be formulated for the features of the items, likelihood ratios can be used to compare and contrast propositions and evaluate which is more likely given the information collected. In forensic sciences, while professional guidelines have endorsed this approach [43], formulating a probabilistic model can be challenging or infeasible, leading researchers to focus on developing measures of similarity and constructing score-based likelihood ratios to contrast propositions [38].

Researchers have mainly focused on the common source rather than specific source problem which is particularly relevant in criminal justice where the source under consideration may be associated with the person of interest or a crime scene.

Consider an example in ballistics examination to illustrate the difference. A bullet casing (recovered item) was discovered at the crime scene, and later, a person of interest with a firearm (source) was detained. Ballistic experts may test-fire the person of interest's gun and keep the new casings (control items). Experts would then examine the markings (features) from the control and recovered items to assess the likelihood of the propositions- whether the casing originated from the person of interest's gun (H_p) or an alternative, unknown gun (H_d). This would constitute a specific source problem; another scenario would be if there were no particular gun under consideration, and experts were comparing two recovered items, potentially from two different crime scenes, to evaluate if there was a common firearm linking both crimes.

While both problems are similar, they address fundamentally different questions. From a statistical standpoint, the inference conducted under the specific source problem is conditioned to a source, while under the common source is unconditioned. As noted by Ommen et al. [33], exchanging the paradigms may result in an incorrect interpretation of the results, even in opposing conclusions. In the case of the score-based likelihood ratio systems, Neuman et al. [31] provide evidence that using the common source systems tends to overestimate the specific source results.

Even if the correct paradigm is followed, research may face additional challenges adapting new developments in machine learning and density estimation to score-based likelihood ratios. Conditioning to a specific source implies developing a model for each case, resulting in an increased computational burden and limiting the available data. Further, researchers construct pairwise comparisons using the same items multiple times to develop their system, resulting in a complex dependence structure.

In a previous study, Veneri and Ommen [41] introduced a resampling plan as a pre-processing step to remedy the dependence structure and enhance the performance of score-based likelihood ratio performance for the common source problem. Our current work extends this approach by

introducing a resampling plan to the specific source problem. By creating synthetic items, we generate the necessary large training datasets to develop synthetic anchored specific source score-based likelihood systems.

Using a simulation study, we compare our proposed approach to a theoretically correct system developed under ideal conditions, where an extensive collection of independent scores is available. Results show that while the output from both methods is not perfectly exchangeable, they generally tend to agree. Notably, our proposed method tended to produce results that favor the defense position and have smaller rates of errors for the defense, at the cost of a small increase in the errors when the prosecutor's statement was correct. This implies an overall improvement in the performance.

We provide two applications of our approach in well-known forensic domains: handwriting and forensic glass examination. In the handwriting application, we compared our approach to two alternative resampling plans: (i) a naive approach disregarding the dependence structure and (ii) a previously domain-specific resampling approach for questioned documents [20, 12]. In the forensic glass application, we compare our proposed approach to a common source score-based likelihood ratio, which can be a tempting alternative when samples are not large enough to develop a specific source version. Further, we use this application to illustrate how our approach can be used to develop base systems and combine them into ensemble score-based likelihood ratios [41] suited for the specific source problem.

While our work focuses on a particular type of anchoring suited for our illustration, our proposed approach can be extended to different anchoring proposals to remedy the dependence structure and enhance the performance of specific source score-based likelihood ratio systems.

3.3 Methods

Before introducing our proposed approach, we present the theoretical framework for developing score-based likelihood ratios under the specific source problem (Section 3.3.1) and how different interpretations have led to the proposal of various anchored definitions (Section 3.3.2).

This allows us to illustrate the theoretically correct procedure to generate training data and highlight how this is not achieved in practice. We then introduce synthetic sampling to create independent learning instances and demonstrate how it can be used to develop a score-based likelihood ratio system within a pre-specified anchoring approach (Section 3.3.3).

3.3.1 Score-based likelihood ratios for the specific source problem

Let e_{ij} denote a generic item, where j indexes the item ($j = 1, \ldots, n_i$) from source i ($i = 1, \ldots, m$) and let $u_{ij} \in \mathbb{R}^p$ represent the associated measurement or feature vector for the item. In specific source problems, researchers are interested in comparing recovered items of an unknown source ($i = u, j = 1, \ldots, n_u$) to items that are associated with a specific source ($i = s, j = 1, \ldots, n_s$) and derive the likelihood associated with the two opposing propositions that can be generically stated as:

- H_p : The items e_{uj} were generated from the same specific source that generated e_{sj}.
- H_d : The items e_{uj} were not generated from the specific source that generated e_{sj} but from some other unknown source.

In the case of forensic science, e_{uj} may have been recovered at the crime scene while e_{sj} have been found after an investigative procedure. In the following sections, we will assume that only one item from the unknown source was recovered for simplicity ($n_u = 1$).

A common assumption is that researchers have access to a reference set $A = \{u_{ij} : \text{features from the j-th source an i-th item }\}$, a sample of the background population to develop their systems. The researchers may have collected the elements of A from previous cases or have access to reference databases. Further, it is assumed that elements in A are not associated with the unknown source or the specific source under consideration and that they are suitable to the problem [1].

[1] In the ballistic examination example, elements in A would have been filtered such that they share the same caliber (class characteristics)

The theoretical framework to develop score-based likelihood ratio systems can be formalized by stating the propositions into sampling distributions that generated the data following a two-step hierarchical process [33].

Let B_i denote latent random variables corresponding to the parameters that characterize the distribution of the features for source i. In the first stage, the latent variable is sampled to generate the source from a distribution $F_b(\cdot \mid \theta_b)$ that describes the between source variation, where the parameter θ_b characterizes the between distribution. Conditional on the latent variable representing the source, the measurement for an item j is sampled from $F_w(\cdot \mid b_i, \theta_b)$, F_w being the distribution describing the within source variation.

For $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$ the proces can be sumarized as :

$$B_i \sim F_b\left(\cdot \mid \theta_b\right) \text{ and } u_{ij} \overset{ind}{\sim} F_w\left(\cdot \mid B_i = b_i, \theta_b\right) \tag{3.1}$$

This mechanism formalizes how the data is generated and clarifies the distinction between the propositions. While both propositions agree on the general data-generating process, there is a difference in how the features for e_{uj} and e_{sj} are generated.

Under H_p, a specific source is generated from F_b and $n_s + 1$ items sampled from $F_w\left(\cdot \mid B_i = b_s, \theta_b\right)$. While under H_d, the specific source was generated from F_b and n_s items sampled from $F_w\left(\cdot \mid B_i = b_s, \theta_b\right)$, with a second unknown source generated from F_b and one items was sampled from $F_w\left(\cdot \mid B_i = b_u, \theta_b\right)$.

If both the between and within distribution were known, the joint densities of the features could be derived under both propositions, allowing the calculation of the value of evidence using a likelihood ratio function, as defined in 3.2.

$$LR(u_u, u_{sj} \mid \theta) = \frac{f(u_u, u_{sj}|\theta, H_p)}{f(u_u, u_{sj}|\theta, H_d)} \tag{3.2}$$

If the distribution were unknown, assumptions can be made about the joint distributions of the features. Estimating θ using the reference set A would allow a similar procedure using an estimated likelihood ratio function.

In practical applications, such as comparing shoeprint impressions or markings in a bullet casing, formulating an estimating probabilistic model may not be feasible, leading researchers to use (dis)similarity scores between items and estimate the conditional distribution of the score to develop score-based likelihood ratio systems [38].

Let Δ denote such a (dis)similarity function that maps from the features vector of two generic items being compared (u_{ij}, u_{lk}) to a univariate score δ ($\Delta(u_{ij}, u_{lk}) : \mathbb{R}^P \times \mathbb{R}^P \rightarrow \delta \in \mathbb{R}$).

Let $g(\delta \mid H_j)$, (j =Prosecutor, Defense) denote the conditional distribution of the score under the proposition H_j. Then, given an observed score δ associated with the items being compared, a value of evidence could be estimated by plug-in δ into

$$SLR(\delta) = \frac{g(\delta \mid H_p)}{g(\delta \mid H_d)}, \tag{3.3}$$

to evaluate the likelihood ratio of observing the score under the two alternative propositions.

While it is widely accepted that the numerator (denominator) in equation 3.3 can be interpreted as the likelihood of the score under H_p (H_d), and that ratio larger than one (smaller than one) can be considered as evidence towards the prosecutor (defense); unlike the common source problem where constructing (dis)similarity metrics and estimating conditional density is straightforward, the specific source problem requires additional considerations [22].

Alternate interpretations of the proposition under the specific source suggest different procedures to develop the system, leading to different versions of the anchored score-based likelihood ratio and how their output can be interpreted [20, 30]. This also affects how the score should be obtained to estimate the appropriate conditional densities. Our work considers a source-anchored version as the stepping stone to introduce synthetic anchored score-based likelihood ratios, but it could also be extended to other versions.

3.3.2 Source anchored Score-based Likelihood Ratios

Previous authors have explored different anchoring procedures to address the specific source problem [20]. All definitions agree that the numerator should be interpreted as the likelihood of

the observed score (when comparing control and recovered items) being generated by the same specific source. The logic is that the control and recovered items should achieve similar scores as pairing randomly selected items from the specific source.

Alternative anchoring definitions stem from the difference in the denominator. In the source-anchored version, the denominator should be interpreted as the likelihood of the observed score (comparing control and recovered) when the control items were generated from another alternative source from the population. The logic is that the control and recovered items should achieve similar scores as randomly pairing items from the specific source to items from other sources.

As Hepler et al. [20] noted, the previous interpretations suggest that both densities on Eq 3.3 are conditioned on the specific source under consideration. In contrast to the common source, where models can be pre-trained to address the inference problem, under the specific source, a score-based likelihood ratio system needs to be developed for each source; in practice, this means constructing a new system for each case.

Further, an appropriate sample must be drawn to estimate the conditional densities under this interpretation. Let $\left\{\delta_i^{(j)}\right\}_{i=1}^{n_j}$ $(j = p, d)$ denote a sample of size n_j that has been drawn under H_j, common estimation procedures assumee that $\left\{\delta_i^{(j)}\right\}_{i=1}^{n_p} \overset{\text{iid}}{\sim} g(\boldsymbol{\delta} \mid H_j)$ or at least some level of independency.

Ommen et al.[33] and Neuman et al. [30] propose a thought experiment on how data should be generated. To summarize their ideas under general framework presented by Equation 4.1, let P and D denote a collection of scores for the prosecutor and defense. We can compactly write the data-generating as follows:

$$P = \left\{\delta_i^{(P)}\right\}_{i=1}^{n_p} = \{\delta : \delta = \Delta(u_{sl}, u_{sk}), l \neq k, \tag{3.4}$$

$$u_{sj} \overset{\text{ind}}{\sim} F_w(\cdot \mid B_i = b_s, \theta_b)\} \tag{3.5}$$

$$D = \left\{\delta_i^{(D)}\right\}_{i=1}^{n_d} = \{\delta : \delta = \Delta(u_{sj}, u_{kl}), s \neq k, \tag{3.6}$$

$$u_{sj} \overset{\text{ind}}{\sim} F_w(\cdot \mid B_i = b_s, \theta_a), \tag{3.7}$$

$$b_k \overset{\text{ind}}{\sim} F_b(\cdot \mid \theta_b) \tag{3.8}$$

$$u_{kl} \overset{\text{ind}}{\sim} F_w(\cdot \mid B_i = b_k, \theta_b)\} \tag{3.9}$$

This introduces two prevalent ideas for generating scores that will be used as learning instances: items should be independent and randomly selected. It also highlights the source of randomness since inference is conditioned (anchored) on the specific source held fixed.

In practical application, however, the researcher can not collect a large set of independent items, measure their features, and pair them to construct sets P and D. Due to the limited data, scores are generated using all potential pairwise comparisons from a fixed number of items, resulting in measurements being used multiple times to construct learning instances.

We introduce a small example to illustrate this point in Figure 3.1. The figure describes a scenario where the researchers are faced with one recovered item from an unknown source (u_u), three control items from the specific source ($u_{sj}, j = 1, 2, 3$), and two alternative sources with three items each ($u_{ij}, i = 1, 2, j = 1, 2, 3$).

The control items are paired (dotted lines) to generate three learning instances for set P. Each control item is paired with every item in the reference population (dashed lines) to generate learning instances for set D. Note that each item from the specific source is used eight times (two times for P and six times for set. D), while items in the reference population have been used three times.

The previous example can be extended to larger cases. Without loss of generality, let n_s denote the number of control items, m the number of alternative sources with n items each. The total learning instances for P is $n_P = \binom{n_s}{2}$ and for D is $n_d = n_s \times n \times m$.

This illustration highlights a practical issue: this procedure results in nonindependent learning instances, as noted by Veneri and Ommen [41], and that n_s is one of the strongest limitations for the specific source inference.

From a practical perspective, it might seem reasonable that the researcher can access more data from other sources (increasing m) even if n is small. However, obtaining the measurement from additional items from a specific source, increasing n_s, can be a complex task, especially in forensic sciences. For items generated from the POI (e.g., questioned documents or fingerprints), it could be argued that if the POI refused to collaborate or that data is obtained under duress, the measurement may not follow the same data generating process. Even in types of evidence where the human factor is reduced (e.g., firearms or glass), obtaining additional measurements can be costly, involving complex laboratory procedures, and can alter or even destroy the original items.

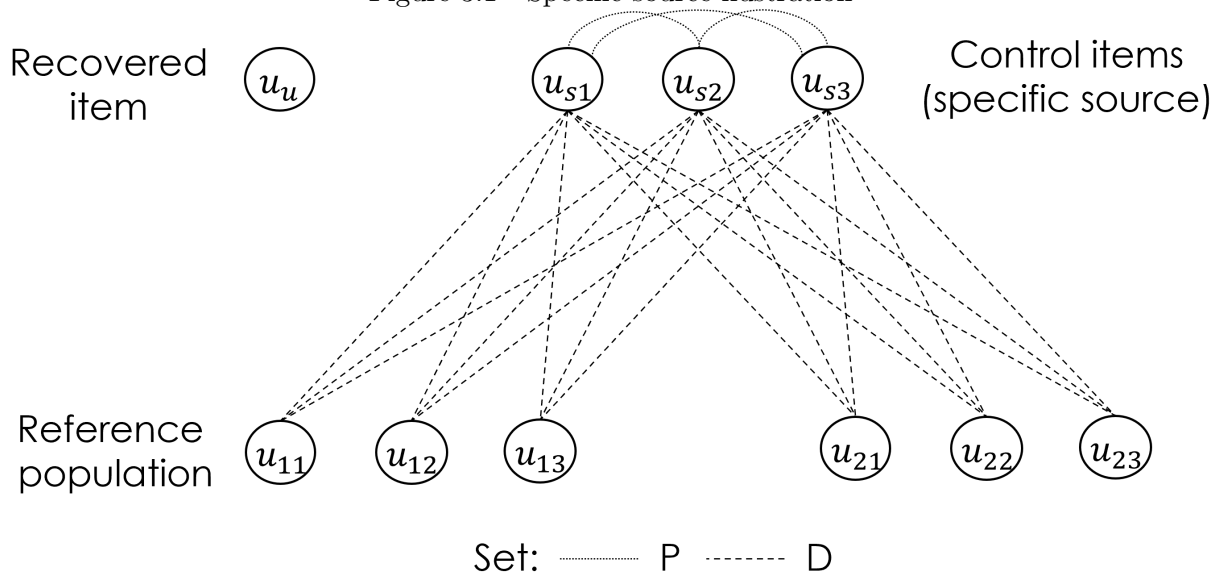
Still, even if alternative sources (m), items within alternative sources (n), and specific source items (n_s) could be freely increased, the dependence would not be directly addressed. The same item in the specific source is used $n \times m$ times to generate learning instances for D, and $n_s - 1$ times for P.

Further, the items from the specific source are also used to compute the final score [2] that will be plugged into the developed score-based likelihood ratio system to derive a conclusion.

In previous works, the lack of data and privacy concerns have led authors in biometrics and forensics to create pseudo items (e.g., fingerprints [25] and question documents [20, 22]) using procedures that are specific to each domain or require a large amount of data. We propose to use synthetic items adapting a popular sampling mechanism in the machine learning literature to develop synthetic anchored score-based likelihood ratio.

[2] Authors have previously computed the mean score $\bar{\delta}$ to accommodate the fact that there are multiple comparisons possible [22]

Figure 3.1 Specific source ilustration



3.3.3 Synthetic items and source anchoring

To address the lack of data and independence, we propose a domain-agnostic procedure to generate a synthetic source-anchored score-based likelihood ratio system that follows the principles outlined in Section 3.3.2.

The first part of the procedure relies on generating synthetic items. Synthetic learning instances have been popularized by resampling algorithms like the Synthetic Minority Oversampling TEchnique (SMOTE) [8]. Originally SMOTE was proposed as a data augmentation method to enhance model performance under imbalanced classes, and different variants have been proposed over the years. Fernandez et al. [14] provide a comprehensive review covering over fifteen years of new developments arising from the initial paper.

A key component of the original SMOTE approach is interpolating between randomly selected learning instances and their nearest neighbors to generate new cases. Our proposed approach follows a similar idea. However, instead of interpolating learning instances, we use interpolation to create new items (and their associated measurements) conditional on a particular source

(Algorithm 4) that are later used to create learning instances to develop a synthetic anchored system (Algorithm 5).

Let $A_i = \{u_{ij} : j = 1, \ldots, n\}$ denote a collection of feature vectors (measurements) associated with items from source i, K the number of candidate neighbors to be used and n^* the requested number of items to be generated.

The initial step computes the K-nearest neighbors for all items in A_i. For each new synthetic item, the feature vectors from one of the original items and one of its neighbors are randomly selected, and data is interpolated. The result of the algorithm is a new database of n^* synthetic items for source i.

Algorithm 4 Synthetic Item (SI)

Require: A_i, K, n^*
  $u_{ij}^{(k)} \leftarrow$ Compute the K-nearest neighbors for each j item, $(k = 1, \ldots, K)$.
  for l=1:n^* **do**
    $j^* \leftarrow$ Randomly select one index from $1 : n$
    $k^* \leftarrow$ Randomly select one index from $1 : K$
    $r \leftarrow$ Sample a random number from $U(0, 1)$
    $u_{il}^* \leftarrow$ Generate a new feature vector $u_{ij^*} + r(u_{ij^*} - u_{ij^*}^{(k^*)})$
  end for
  Output: Set $A_i^* = \{u_{ij}^* : j = 1, \ldots, n^*\}$

Following the principles outlined in 3.3.2, we employ Algorithm 4 to develop a synthetic source-anchored score-based likelihood ratio system. As before, let $A_i = \{u_{ij} : j = 1, \ldots, n\}$ be the subset of measurements associated with a particular source in set $A = \{u_{ij} : i = 1, \ldots, m; j = 1, \ldots, n\}$, and $S = \{u_{sj} : j = 1, \ldots, n_s\}$ denote the set with measurements associated with control items.

The general workflow in Algorithm 5 accommodates two scenarios: first, when researchers aim to use a well-established (dis)similarity metric (e.g., L_1, L_2, cosine); second when researchers prefer to develop a (dis)similarity metric tailored to the specific problem. The latter could mean training a random forest or a neural network to differentiate matches from non-matches. As mentioned before, under the specific source problem, it is not possible to pre-train a model before the source is known. If a tailored metric is preferred, first Algorimth 4 generates synthetic items,

three times the number of sources in the reference population. The first two-thirds are used to obtain matches from the specific source, while the last third is used with newly generated synthetic items from each alternative source to obtain non-matches. This data can be used as the training set in a classification algorithm.

Once the new metric is obtained, or if the researchers selected one that did not require training, the same procedure generates pairs over which scores will be computed. The scores obtained are collected and used to estimate the conditional densities or a ratio density estimator. The output of this procedure is the estimated densities (and metrics) that will be used to compute the value of evidence.

The procedure described follows the principles outlined in Section 3.3.2. Sources in the alternative population are used once to generate synthetic items, and synthetic items from the specific source are generated to augment the number of learning instances while reducing the dependency of using the same data to compute scores. We provide a short illustration of the algorithms in Appendix 3.8. Sections 4.7 presents the result of a simulation study and Section 3.5 provide two realistic applications in forensic science. Limitations of our approach and potential modifications are discussed in Section 3.6.

3.4 Simulation study

Our simulation study aims to address to what extent the inference carried out with synthetic anchoring obtains results similar to those of the traditional anchored systems. This gold standard is unlikely to be achieved in practice but is a natural benchmark to aim for. Both methods were compared using standard performance metrics for Likelihood Ratios [3] and reliability statistics to asses how similar their conclusions were.

Score-based likelihood ratio systems output a numeric value between zero and infinity, where a value larger than one indicates that the evidence supports the prosecutor rather than the defense proposition. The rate of misleading evidence (RME) for the prosecutor (defense) can be

[3] We refer the interested reader to the Appendix B of [41] for detailed notation of these metrics

Algorithm 5 Synthetic source anchored score likelihood ratio

Require: S, A, K
  $m \leftarrow$ Number of sources in A.
  if Δ requires training **then**
    $u^*_{s^h j} \leftarrow SI(S, K, 3 \times m),\ (h = 1, 2, 3; j = 1, \ldots, m)$ ▷ Apply Algorithm 4
    for l=1:m **do**
      $u^*_{l1} \leftarrow SI(A_i, K, 1)$ ▷ Apply Algorithm 4
      $T_l \leftarrow (u^*_{s^1 l}, u^*_{s^2 l}, Match)$
      $T_{2\times l} \leftarrow (u^*_{s^3 l}, u^*_{l1}, Non - match)$
    end for
    $\hat{\Delta} \leftarrow$ Train metric to diferentiate Match and Non-matches in T
  end if
  $u^*_{s^h j} \leftarrow SI(S, K, 3 \times m),\ (h = 1, 2, 3; j = 1, \ldots, m)$ ▷ Apply Algorithm 4
  for l=1:m **do**
    $\delta_l^{(P^*)} \leftarrow \Delta(u^*_{s^1 l}, u^*_{s^2 l})$
    $u^*_{l1} \leftarrow SI(A_l, K, 1)$ ▷ Apply Algorithm 4
    $\delta_l^{(D^*)} \leftarrow \Delta(u^*_{s^3 l}, u^*_{l1})$
  end for
$\hat{g}(\delta \mid H_j) \leftarrow$ Estimate using $\{\delta_i^{(j)}\}_{i=1}^{n^*}, (j = P^*, D^*)$ ▷ $\hat{g}$ can be replaced by a density ratio estimator
Output: $\hat{g}(\delta \mid H_P), \hat{g}(\delta \mid H_D)$ (and $\hat{\Delta}$)

computed as the proportion of cases where the system is expected to output a value larger (smaller) than one, but the opposite is observed over a validation set. We expect that both methods would result in comparable error rates.

Besides the direction, the numerical output of the system provides researchers with a measurement of the strength of evidence. We employ a modified version of the Bland-Altman Plots using the log10 scale to assess our proposed method's agreement with the theoretical gold standard. If methods were exchangeable, results would be aligned along the y-axis in the Bland-Altman plot. Ideally, we would expect that one method does not provide consistently larger values of evidence compared to the other.

In practical applications, forensic experts may present jurors with a verbal interpretation of the numerical output [43]. While previous authors have reviewed the benefits and downfalls of this approach [32, 26], we only focus on the fact that potential discrepancies should be evaluated in terms of their impact on jurors' decisions [15].

Different verbal scales have been proposed to interpret and communicate results; we consider the symmetric ten-level verbal scale [13], which maps the continuous output into ten categories as depicted in Table 3.1.

Over this scale, we present a confusion matrix for the ten-level ordered categories to visualize their agreement. A perfect agreement would be achieved if all elements fall within the diagonal. To provide a summary statistic that accounts for the different levels of disagreement outside the diagonal, we employ a weighted version of the Kappa statistics [9] and Gwest's AC [16], both measures of between-rater reliability, considering ordinal weights [4].

Table 3.1 Ten level verbal scale

Towards	Qualifier	SLR scale	
Defense	Very Strong	0	10^{-4}
	Strong	10^{-4}	10^{-3}
	Moderately strong	10^{-3}	10^{-2}
	Moderate	10^{-2}	10^{-1}
	Limited	10^{-1}	10^{0}
Prosecutor	Limited	10^{0}	10^{1}
	Moderate	10^{1}	10^{2}
	Moderately strong	10^{2}	10^{3}
	Strong	10^{3}	10^{4}
	Very Strong	10^{4}	∞

Note: Ten level verbal scale proposed by Evett, exponents denote the cutt off values in the log10 scale.

3.4.1 Simulation strategy

Our simulation strategy aims to contrast our approach to the gold standard under a well-known data-generating process. Two-level Gaussian models have been extensively used in forensic science to account for within- and between-source variation; section 7.6.2 in Aitken et al. [2] provides a more extensive review of the models.

As before, let u_{ij} denote the features associated with the items j ($j = 1, \ldots, n$) in source i ($i = 1, \ldots, m$). The two-level Gaussian model can be expressed as a random effects model

$$u_{ij} = \mu + a_i + w_{ij}, \tag{3.10}$$

[4] See Chapter 3 of [17] for a review of ordinal agreement coefficients and weights.

where μ represents the overall mean, a_i is a random effect for the i source, and w_{ij} is a random effect for the j item within source i. The two-level Gaussian derives its name from the assumption about the distribution of the random effects, $a_i \sim N(0, \Sigma_b)$ and $w_{ij} \sim N(0, \Sigma_w)$, where Σ_b is the between-source covariance matrix, and Σ_w is the within-source covariance matrix.

Using a two-stage procedure, these models allow straightforward simulation [29, 40]. First, simulating the mean vector (latent variable) for a particular source ($B_i \sim N(\mu, \Sigma_b)$) and conditional on the source, sample the observed features vector ($u_{ij} \sim N(b_i, \Sigma_w)$). Following the notation in Section 3.3, we can write:

$$B_i \sim \mathrm{N}(\mu, \Sigma_b) \tag{3.11}$$

$$u_{ij} \mid B_i = b_i \sim \mathrm{N}(b_i, \Sigma_w) \tag{3.12}$$

where F_b ad F_w are gaussian distributions and $\theta_a = \{\mu, \Sigma_w, \Sigma_b\}$.

To obtain realistic simulations, we consider the previous parameter estimates from forensic glass applications where the data measured consisted of four elemental compositions (Calcium (Ca), Potassium (K), Silicone (Si), and Iron (Fe)) and the features consisted of three log ratios $(log(Ca/K), log(Ca/SI), log(Ca/FE)$ (See section 7.6.4 [2] or [1] [5])

Using these estimates as the true model parameters, we simulated data to create scenarios with different difficulty levels [34].

To ensure a balanced learning set of 200 comparisons evenly split between prosecutor and defense sets, we sampled 104 sources ($i = 1 \ldots, 104$), defined the first randomly generated source as our primary source of interest, and computed the L_1 norm to select three sources for the H_d scenarios. The one associated with the minimum distance (closest non-match), the 5th percentile (5p non-match), and the 10th percentile (10p non-match) were selected, the closest being the hardest to differentiate. The remaining sources (without loss of generality,$i = 5, \ldots, 104$) are set as part of the background population (A).

[5] The authors present rounded values, and we considered more precise estimates for our simulations.

From the source associated with the POI, an item (and its associated features) is sampled to serve as the recovered item. From the same source and the closest, 5p and 10p non-match sources, 300 measurements are generated. Ten additional measurements are sampled for each alternative source in A.

To generate learning instances as depicted in Section 3.3.2 we selected the L_1 norm as a dissimilarity metric. For each scenario, the first two-thirds of the control items are used to create match comparisons, pairing the first and second half one-to-one, resulting in 100 learning instances for P. The remaining third is paired with the first item from the alternative sources in A.

While this approach guarantees balanced learning sets, where scores have been properly created from independent comparisons, the number of required samples per source may be infeasible to obtain in practice.

To bypass this limitation, we consider the alternative approach introduced in Section 3.3.3. We limited the number of samples per source to the first ten items and followed the step described in Algorithm 5 to construct learning instances for P and D.

Since we consider the L_1 norm as a dissimilarity metric for both approaches, a zero value would indicate that the items are identical regarding their observed features. In contrast, larger values indicate that they are more dissimilar. We selected Weibull distributions as the parametric family to characterize the distribution of the scores under both propositions and independently estimate the conditional density.

Once the densities are estimated, the final score is computed between the recovered and the first ten control items, and the average score ($\bar{\delta}$) is plugged into the system to obtain a value of evidence. If the methods introduced to develop specific source systems are exchangeable, we would expect to observe the values of evidence in agreement.

3.4.2 Simulation results

The simulation procedure was repeated 1000 times to obtain a sample of estimated evidence values and compute performance metrics. Results show that while the agreement between the two methods is imperfect, they may be considered adequate.

Let SLR_{DGP} denote the output obtained using the theoretically correct data-generating process, and SLR_{synth} represent the one obtained using our proposed approach. Figure 3.2 present a visualization of the agreement using the Bland Altman plot [6]. If the methods tend to agree, we would observe results aligned along the x-axis. Results show that the difference in the log10 scale is not constant; dispersion tends to increase towards a more extreme value of evidence. Overall, the median difference is negative, ranging from -0.6 to -0.92 across different scenarios (Column 7, Table 3.4), due to our proposed method tending to output smaller evidence values. The probability of observing a larger value of SLR_{synth} compared to a SLR_{DGP} ranges between 25-35% for all scenarios (Column 6 Table, 3.4).

In the context of the evidence interpretation, the synthetic anchored version outputs more conservative evidence values, leaning towards the defense proposition.

This affects the estimated error rates(Columns 4-5, Table 3.4). The rate of misleading evidence for the defense scenario decreased between $5-7$ percentage points across the different scenarios at the expense of a 1.2 percentage point increase in the rate of misleading evidence for the prosecutor. These two errors are not symmetric; the cost of stating that an item originated from a specific source wrongly (providing evidence supporting a guilty verdict) may be weighted more heavily. We address this in the conclusion section.

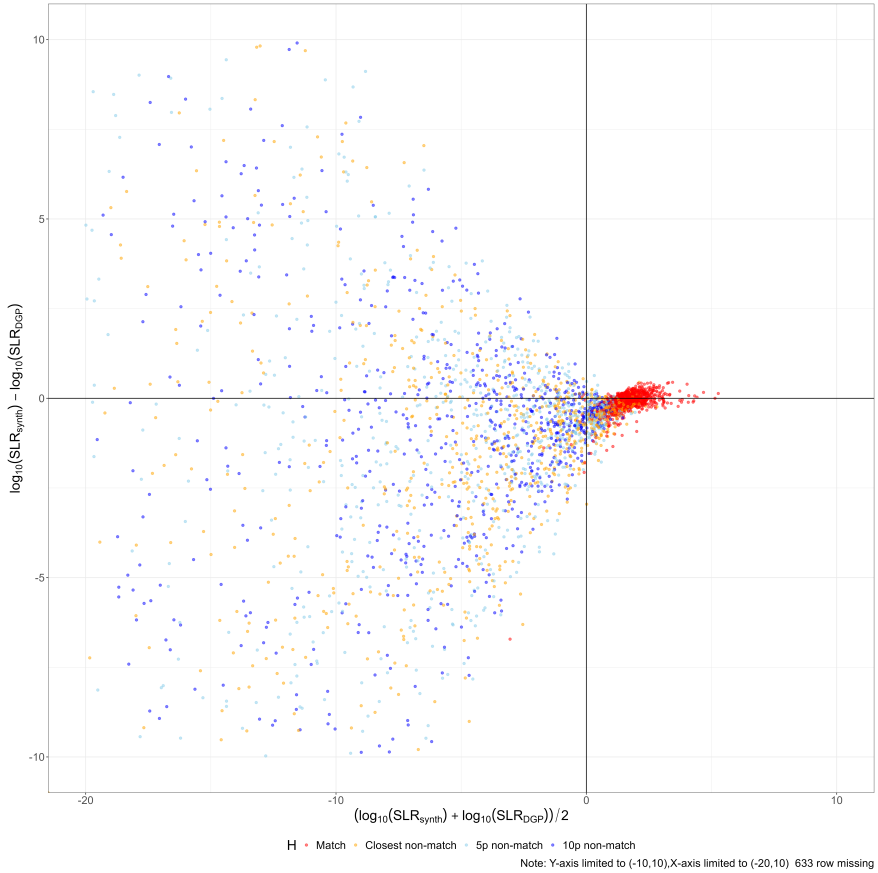
Previous results regarding the level of disagreement may be driven by the extreme value of the evidence observed, which, in practice, would have a small impact on the evidence interpretation. We explore the disagreement over the ten-level verbal scale, where the more extreme values are combined into the categories at the end of the scale. Figure 3.3 presents a visual representation of

[6] The x-axis represents the average in the log10 scale $A = \frac{1}{2}(\log_{10}(SLR_{synth}) + \log_{10}(SLR_{DGP}))$ and the y-axis the difference in the log10 scale $M = \log_{10}(SLR_{synth}) - \log_{10}(SLR_{DGP}) = \log_{10}(SLR_{synth}/SLR_{DGP})$

the confusion matrix, where tiles represent the joint verbal scale distribution and margins the verbal scale distribution of each method.

The output of both methods tends to fall more frequently within the diagonal, ranging from 84.8% in the prosecutor scenario to 64 % for the hardest defense scenario (Column 1, Table 3.4). The previous metric considers elements appearing as off-diagonal errors but does not consider the scale's ordinal nature. We computed the ordinal weighted Kapps and Gwet's agreement coefficient (Columns 2-3, Table 3.4) to account for this. Both statistics suggest a very good agreement between both methods [7].

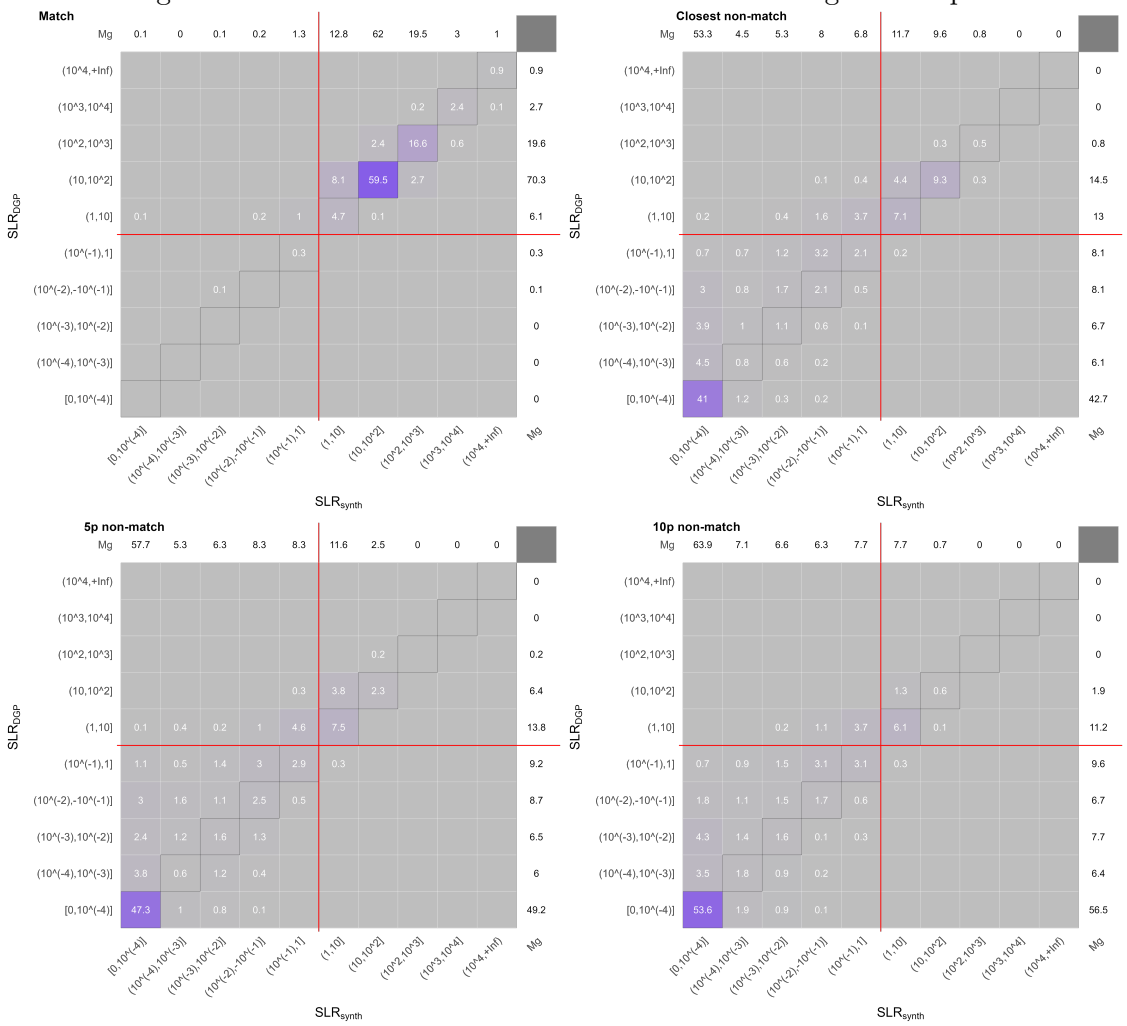
Figure 3.2 Simulation results: Combined Bland-Altman plot



Note: Simulation results for 1000 simulated values for each scenario. Combined Bland-Altman plots (extreme values removed)

[7] Different benchmarks have been proposed in the literature for unweighted statistics: Landis and Koocj's scale would place the values found on the substantial range (0.61-0.8) or almost perfect range (0.81-1), Fleis's scale would place them in the excellent category Excellent (0.75-1), similarly Altman's scale would place them our results in the Very good range (0.81-1). We refer the interested reader to Section 6.2.1 in [17]

Figure 3.3 Simulation results: ten-level verbal scale agreement plot



Note: Simulation results for 1000 simulated values for each scenario. Combined confusion matrix plots using a ten-level verbal scale

Previous analysis suggests that while the agreement is imperfect, our proposed approach may be useful in answering the specific source problem, providing a good level of agreement, and reducing the rate of misleading evidence for the defense.

Table 3.2 Simulation results: Gaussian DGP statistics

Hypothesis	In diagonal (%)	Kappa (Ordinal w)	Gwet's AC (Ordinal w)	RME DGP (%)	RME Synth (%)	$P(SLR_{synth} > SLR_{DGP})$	Median M $\log_{10}(SLR_{synth}/SLR_{DGP})$
Match	84.40	0.80	0.99	0.40	1.70	36.20	-0.06
Closest non-match	64.00	0.88	0.95	28.30	22.10	24.70	-0.70
5p non-match	64.70	0.85	0.95	20.40	14.10	26.20	-0.81
10p non-match	68.50	0.84	0.97	13.10	8.40	29.40	-0.92

3.5 Applications

To evaluate the performance of our approach, we present applications in two domains with a rich history in forensic statistics: question documents (considered pattern evidence) and forensic glass analysis (considered trace evidence) [8] which provide different types of measurements to test our approach. In our application, questioned documents are characterized by counts, later transformed into frequencies. The feature vector can be considered compositional data or a vector of frequencies. In our second application, glass fragments are characterized by the logarithm of chemical composition, which is assumed to be a continuous measurement.

In the handwriting analysis application, we selected the cosine similarity metric and the beta parametric family to estimate the conditional densities. We compare our approach to a naive approach, where comparisons are created disregarding the structure of the data, and previous domain-specific resampling plan for questioned documents [20, 12]. We provide additional details about the previous approach in Appendix 3.8.2.

In the forensic glass analysis application, we selected a random forest to develop a similarity score and a logit-based probabilistic density ratio estimator. In this case, a training and estimation set are required, allowing us to illustrate how this can be achieved under our proposed approach. Further, we illustrate how synthetic anchoring can be used with ensemble learning as proposed in Veneri and Ommen[41] to develop ensemble score likelihood ratios for the specific source problem. To the best of our knowledge, there is no established resampling benchmark for this application, and the small number of learning instances would be further reduced if the learning instances were split into training and estimation. We compare our approach with a

[8] We refer the reader to [35, 21] for an introduction to these types of evidence.

common source score-based likelihood ratio system to illustrate that a pre-trained model for the common source may not perform as expected for the specific source problem.

To create different scenarios for our applications, we follow a similar approach as the one already implemented in Section 3.4.1. In each iteration, we selected one source to play the role of a specific source, defining the first item as the recovered and the remaining as controls. The closest non-match is selected by computing average features for each source and obtaining the nearest neighbor. The remaining sources are defined as the reference set. Additional detail about the simulation strategy is presented in each subsection.

3.5.1 Application in Handwriting Analysis

To motivate this application, consider a scenario where a stalking victim has filed a police report and submitted the threatening note as evidence. The document not yet associated with a specific source will be considered the recovered item. After an investigation, a person of interest was apprehended, and a collection of his previous writings was collected; these items act as control items for which the source is known to be the person of interest.

Traditional methods for questioned document analysis rely on visual inspection by trained examiners to assess the similarity between two items. CSAFE authors [11, 6] have developed a method to decompose writing samples into graphs and assign each to one of 40 pre-established clusters or groups. For each document, the proportion of the graphs falling into each group is defined as the writership profile or feature vector. Each entry of the 40-dimensional vectors is non-negative, and the vector adds up to one.

Writers follow similar patterns, allowing researchers to identify the source of the questioned documents based on these observed features. These features have been previously used under the closed-set problem [11], the common and specific source problem [22], and benchmark an ensembling learning approach to SLRs for the common source problem [41].

Previous applications with similar features have relied on different similarity measurements to assess the similarity between two writership profiles (e.g., Random forest classifiers [22, 41]

chi-square and Kullback-Leibler distance [37, 20, 12]). Khocher and Savoy [24] review other distances used in author profiling. The author reviews the cosine similarity, recently used with deep learning to compare questioned documents [23].

Let u_{ij} be a feature vector with the proportion of graphs classified in each group, and $u_{ij[p]}$ denotes its p-entry $(p = 1, \ldots, 40)$. The similarity between two vectors can be computed as:

$$\Delta(u_{ij}, u_{lk}) = \frac{u_{ij} \cdot u_{lk}}{\|u_{ij}\|\|u_{lk}\|} = \frac{\sum_{p=1}^{P} u_{ij[p]} u_{lk[p]}}{\sqrt{\sum_{p=1}^{P} u_{ij[p]}^2}\sqrt{\sum_{p=1}^{P} u_{lk[p]}^2}}, \tag{3.13}$$

Given that all the entries in the feature vectors are non-negative, the cosine distance maps the features from two documents to a univariate score ($\Delta(u_{ij}, u_{lk}) \to [0, 1]$), one indicating that documents are more similar, the proportions are equal across two vectors.

The cosine distance is an example of a metric that does not require training; estimating the conditional densities is the only requirement to develop a score-based likelihood ratio system. For our application, scores of zero or one did not occur, so we selected the beta exponential family due to its flexibility and domain to estimate the conditional densities.

Our database for this application consists of the first batch of 90 writers from the CSAFE database. We cycle through all 90 writers to develop a specific source system, the first item being defined as recovered and the remaining eight as controls for H_p. We estimated the selected source's nearest neighbors and selected the last eight items as controls for H_d.

Across iterations, we compare the performance of three resampling variants. First, a naive approach described in Section 3.3.2 where all pairwise comparisons are made within the control items ($n_p = \binom{8}{2} = 28$) and all combination between control items and reference populations ($n_d = 8 \times 8 \times 88 = 5632$) are used to create learning instances. Second, a fixed split of a mega document, a method introduced in [20, 12] where items from each source are stacked into a mega document, and random points are selected to split the documents and create new scores for P. We provide the reader with a summary of this approach and our implementation in Appendix 3.8.2. Under this approach, the recovered mega document is compared to the mega document of the alternative source to obtain scores for D ($n_d = 88$). Documents were split 44 times to

generate a score for P ($n_p = 88$) so we have balanced learning instances. And third, we apply algorithm 5 to generate a synthetic source anchored system ($n_d = n_p = 88$).

To analyze our results, we present the distribution of the scores (Figure 3.4) and empirical estimates of the rate of misleading evidence and discriminatory power, metrics associated with the error rate, and the ability to provide strong evidence in the correct direction respectively for each proposition in Table 3.3. Further, we compute the $Cllr$, a forensic cost function, as a composite performance metric.

Synthetic anchoring achieves a smaller cost, as measured by the $Cllr$ (Colum 6, Table 3.3) due to better performance for the defense scenario.

In both scenarios, synthetic anchoring resulted in more conservative evidence values, favoring the defense proposition. This resulted in a smaller rate of misleading evidence for the defense at the expense of increasing misleading evidence for the prosecutor (Colum 1-2 Table 3.3) and a decrease in the ability of the system to provide stronger evidence for the prosecutor case when it is true. Further analysis showed that fixed sampling produced larger values of evidence for the same holdout source.

Table 3.3 Performance metrics of Score Likelihood Ratios for Handwriting analysis

Sampling method	RME_{Hp} $P(SLR < 1\|Hp)$	RME_{Hd} $P(SLR > 1\|Hd)$	DP_{Hp} $P(SLR > 10^2\|Hp)$	DP_{Hd} $P(SLR < 10^{-2}\|Hd)$	$Cllr$	Avg.Cost Hp	Avg.Cost Hd
Naive	3.33	55.56	25.56	16.67	0.95	0.35	1.55
Fix	2.22	63.33	46.67	5.56	0.98	0.17	1.79
Synthetic	11.11	25.56	5.56	11.11	0.60	0.49	0.70

3.5.2 Application in forensic Glass analysis

Glass analysis is a common staple in forensic analysis since it is widely used in daily life (e.g. containers, window panes, and windshields),it is easily transferable if the glass is broken, does not degrade over time, and analysis can be done over small fragments [5]. Consider a scenario where an individual has broken into a car to motivate this application. Upon arrival at the scene, officers apprehended a person a few blocks from the crime scene who had glass fragments

attached to their clothing. Forensic analysis arrived at the scene and recovered glass fragments that will act as a control in their analysis.

In their early inception, glass fragments were characterized by their refractive index, a univariate measurement, while in modern applications, they are characterized by their chemical composition, a multivariate measurement.

It is relevant to highlight the limitation of the inference that can be derived from this analysis. The propositions consider whether the fragments retrieved from the person of interest can be associated with the crime scene. This is a lower-level proposition [10] regarding the origin of the items, not if the person broke the glass. Further, recent work has shown that forensic interpretation may be hindered by the production process, where glass panes are generated in sequence, resulting in similar chemical composition [3]. We address this limitation in the discussion section.

Our application uses the Florida International University (FIU) Glass database [4] [9]. The database consists of glass samples, and the feature vector for each item is a 15-dimensional vector of laser ablation inductively coupled plasma mass spectrometer (LA-ICP-MS) log measurements for the stable isotopes Li7, Mg25,l27, K39, Ti49, Mn55, Fe57, Rb85, Sr88, Zr90, Ba137, La139, Ce140, Nd146, Pb208. The samples were collected from vehicles of different makes and models, totaling 761 sources and nine replicates (items) each. Some sources represent the same windshield's inner and outer sides. There is the possibility that these are more similar and chemically indistinguishable if both sides were produced at the same plant in a short period.

To illustrate the use of a machine learning-based synthetic anchored score-based likelihood ratio system, we followed a similar approach to previous authors [36, 35] where a random forest [7] is trained to be used as a dissimilarity metric. To transform the obtained sample of scores into a score-based likelihood ratio system, we selected a penalized logistic regression [19] as a density ratio estimator. Rather than cycle through all sources in the database, we randomly draw one source and its nearest neighbors across 100 iterations to create validation cases.

[9] The data can also be accessed through the SK4FGA package in R [18]

Developing a dissimilarity metric can be framed as a two-class classification problem (matches vs. non-matches). To construct the learning instances for the random forest, we consider a pair of items and their feature vectors (u_{ij}, u_{lk}), considering a positive case $(y = 1)$ if the two originate from the same source $(i = l)$ and created features associated with the learning instance as follows: let $u_{ij[p]}$ denote the p-entry in the vector original feature vector $(p = 1, \ldots, 40)$, we consider the absolute difference in each entry $(|u_{ij[p]} - u_{kl[p]}|)$ and an aggregated measurement of discrepancy using the Euclidean distance, $d_{L2}(u_{ij}, u_{lk}) = \sum_{p=1}^{P}(u_{ij[p]} - u_{kl[p]})^2$. After the random forest is trained, it is used to map between feature vectors and the score $\hat{\Delta}(u_{ij}, u_{lk}) \rightarrow \delta \in [0, 1]$

Machine-learning dissimilarity metrics tend to separate the two classes exceptionally well, and values of zeros and ones are frequent, making the beta exponential family unsuitable for estimating the scores' conditional densities. We consider a logistic density ratio estimator [10] to obtain:

$$\text{SLR}(u_{ij}, u_{lk}) = \frac{n_d}{n_p} \exp^{(\beta_0 + \beta_1 \Delta(u_{ij}, u_{lk}))} . \tag{3.14}$$

Given the separation observed between the two classes, the regular logistic density ratio estimator would lead to the score-based likelihood ratio systems to produce more extreme outputs $(\beta_1 \rightarrow \infty)$, to avoid this issue, a penalized logistic regression [19] was selected to obtain estimates of β_0 and β_1 to complete the system, meaning $\hat{\Delta}, \hat{\beta}_0$ and $\hat{\beta}_1$.

In this application, eight items were selected as the recovered items; hence, if all pairwise combinations are considered, only $\binom{8}{2} = 28$ potential learning instances are possible for P. Further, only fourteen learning instances are available for each step if the researcher follows standard practices of splitting the data into training and estimation sets.

Researchers may be tempted to use a pre-trained common source score-based likelihood ratio system to avoid this issue and reduce computational time. We consider this as a natural benchmark to compare our approach. Under the common source problem, sources not selected as validation cases were randomly split into two groups of 330 sources, and pairwise comparison used

[10] See Sugiyama et al. [39] for an introduction to density ratio estimators outside the forensic domain, Morrison [28] and Ommen and Veneri [41] for examples of the use of the logistic classifier in forensic applications.

to learning instances. This resulted in $\binom{330\times 8}{2} = 3483480$ total learning instances of which, $\binom{8}{2} \times 330 = 9240$ were from the same source and 3474240 from different sources. We follow the common practice of downsampling to obtain a balanced dataset.

This procedure, while practical, addresses a different inferential problem. Our proposed resampling plan can be used to tackle the correct inferential problem by using Algrotim 5 twice for each specific source. First, we use the algorithm to generate a training set over which the random forest is trained. Then, the algorithm is used again to generate an estimation set, features are constructed, and the random forest developed for the specific source is used to predict new scores and obtain sets P and D. Over these sets, a penalized logit is estimated to obtain a probabilistic classifier to transform the scores into a score-based likelihood ratio system. We repeated this process for each source 30 times to generate a base score-based likelihood ratio that acts as weak learners that could be aggregated into an ensembled score-based likelihood ratio system.

We present the resulting value of evidence for each specific source base systems (black dots), the ensembled value of evidence using the mean as an aggregator (red triangles), and the value of evidence derived from the common source system (blue squares) in Figure 3.5. First, we address the results for the synthetic anchored and ensemble systems. As noted by Veneri and Ommen [41], using a resampling step to develop a score-based likelihood ratio system introduces variability in the results. Still, an aggregator such as the mean can provide more stable results, reducing the effect of outliers.

Comparing the ensembled evidence values to those derived from the common source system suggests that the common source problem may result in larger evidence value, resulting in considerably larger rates of misleading evidence for the defense (Table 3.4). While for the prosecutor scenario, the common source system has reduced the rate of misleading evidence to zero, and provided at least moderately strong evidence in the correct direction consistently, we observed little variability in the evidence values compared to the specific source values of evidence. This is because the random forest scores are close to one; the system was developed to answer if the glass fragments were generated from a common unknown window, not any specific one.

The resulting distribution of the scores for the cases where the prosecutor is correct is highly concentrated in the largest value outputted by the system, while the defense value of evidence seem to have shifted to the prosecutor conclusion region compared to our proposed approach (Figure 3.6). Lastly, we address the forensic cost function (columns 7-8,Table 3.4). The common source system incurred a $Cllr$ of 4.46, mainly driven by the cases where the defense was correct. One is usually set as the threshold for the $Cllr$, values below one indicate the validity of the system [27]. Overall, results suggest that the common source system is not a reasonable answer to the specific source question.

Table 3.4 Performance metrics of Score Likelihood Ratios for Glass

Method	RME_{Hp} $P(SLR < 1\|Hp)$	RME_{Hd} $P(SLR > 1\|Hd)$	DP_{Hp} $P(SLR > 10^2\|Hp)$	DP_{Hd} $P(SLR < 10^{-2}\|Hd)$	$Cllr$	Avg.Cost Hp	Avg.Cost Hd
Synth-ESLR	3.00	24.00	75.00	32.00	0.66	0.16	1.15
CS-SLR	0.00	68.00	100.00	15.00	4.46	0.00	8.91

Note: Synth-ESLR denotes the synthetic anchored ensembled score likelihood ratio system, while CS-SLR denotes the system developed using the common source approach.

3.6 Conclusions

The score-based likelihood ratio has been proposed as an alternative to the traditional likelihood ratio approach for evaluating evidence when formulating and estimating a probabilistic model is not feasible. This is particularly relevant for pattern evidence in criminal justice, where machine learning is increasingly used to assess the value of evidence between two opposing propositions. The current approach relies on constructing pairs of items under scenarios where both propositions held true to train a score-based likelihood ratio. By using pairwise comparisons, researchers create a dependence structure that violates how scores should be created in theory and impacts the performance of machine learning scores and density estimation procedures.

Previous work by Veneri and Ommen [41] sought to remedy the issue by introducing a resampling procedure based on source resampling to thin out the dependence under the common source problem. The methods introduced in their work do not apply to the specific source

problem, the most relevant problem in a court case where the inference is conditional on a specific source.

As noted by Vergeer [42], most work has focused on common source problems. This may be due to the lack of data for machine learning dissimilarity metrics or the computation burden of creating a specific system for each source. However, this aspect should not be disregarded as using pre-trained common source machine learning metrics has been shown by previous authors to underperform [22], with even simpler scores tending to overestimate the evidence value [31]. The results of our forensic glass application seem to indicate similar results.

Our work proposes using synthetic items as a data augmentation tool and a resampling plan to alleviate the dependence structure. Our proposed approach contributes to a line of research in resampling methods for source attribution [20, 12, 25] but is not limited to a particular data type or domain.

Simulation results show that for the well-known two-level Gaussian data-generating process with realistic parameters, our proposed approach and the theoretically correct system tend to agree in terms of the verbal scale used to group evidence value, albeit our proposed methods tend to be conservative and output values that may favor the defense proposition.

In our applications, we illustrated how the method can be used in two forensic domains: handwriting and glass analysis. In handwriting, it was compared to two other resampling plans, showing that our proposed method is viable and outperformed them in terms of misleading evidence for the defense at the expense of an increase in the rate for the prosecutor but in an overall reduction in the cost incurred as measured by the $Cllr$.

Using the forensic glass data application, we illustrated how our proposed approach could be used to develop a machine learning-based score likelihood ratio system for the specific source problem and that the common source systems, although practical, may underperform for specific source inference.

Potential limitations of our proposed approach have been previously mentioned. One concern is that our approach is more lenient toward the defense proposition, producing smaller error rates

in this direction. However, in the context of forensic science, this may be the lesser mistake to be made.

Another concern is that synthetic items are generated conditional on the observed data via interpolating nearest neighbors. Synthetic items are limited to the convex hull in the feature space generated from the original items, meaning there is no interpolation "outside" observed data. The quality and sample size of the candidate pool may impact the results. Further, the number of closest neighbors (K) in Algorithm4 plays a central role. Small values limit the possibility of interpolating to areas where data is unlikely, but it also may affect the variability of the scores. The effects of choosing different values of k should be explored in the future.

In our work we illustrated our approach for source anchoring; forensic statistics literature offers various anchoring proposals that synthetic items could expand upon. Still, our proposed approach is not limited to forensic sciences domains. Synthetic anchoring can be applied to other source attribution problems.

3.7 References

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

[2] Aitken, C. G. G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley and Sons, Ltd., West Sussex, UK, 2nd edition.

[3] Akmeemana, A., Weis, P., Corzo, R., Ramos, D., Zoon, P., Trejos, T., Ernst, T., Pollock, C., Bakowska, E., Neumann, C., et al. (2021). Interpretation of chemical data from glass analysis for forensic purposes. *Journal of Chemometrics*, 35(1):e3267.

[4] Almirall, J. and Akmeemana, A. (2022). Shiny Glass Application.

[5] Almirall, J. and Trejos, T. (2015). *Analysis of glass evidence*, chapter 6, pages 228–272. John Wiley Sons, Ltd.

[6] Báez-Santiago, F., Lundstrom, J., Crawford, A., Berry, N., Escobar, B., Taylor, J., Reinders, S., and Ommen, D. (2021). Handwriter: An r package for statistical writership analysis.

[7] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[9] Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

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

[11] Crawford, A. M., Berry, N. S., and Carriquiry, A. L. (2021). A clustering method for graphical handwriting components and statistical writership analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(1):41–60.

[12] Davis, L. J., Saunders, C. P., Hepler, A., and Buscaglia, J. (2012). Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. *Forensic science international*, 216(1-3):146–157.

[13] Evett, I., Jackson, G., Lambert, J., and McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, 40(4):233–239.

[14] Fernández, A., Garcia, S., Herrera, F., and Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905.

[15] Garton, N., Ommen, D., Niemi, J., and Carriquiry, A. (2020). Score-based likelihood ratios to evaluate forensic pattern evidence. *arXiv preprint arXiv:2002.09470*.

[16] Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

[17] Gwet, K. L. (2014 - 2014). *Handbook of inter-rater reliability : the definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, Gaithersburg, MD, fourth edition. edition.

[18] Hayward, T. and Curran, J. (2023). *SK4FGA: Scott-Knott for Forensic Glass Analysis*. R package version 0.1.1.

[19] Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419.

[20] Hepler, A. B., Saunders, C. P., Davis, L. J., and Buscaglia, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic science international*, 219(1-3):129–140.

[21] Izenman, A. J. (2020). Comparing handwriting in questioned documents. In *Handbook of Forensic Statistics*, pages 341–363. Chapman and Hall/CRC.

[22] Johnson, M. Q. and Ommen, D. M. (2022). Handwriting identification using random forests and score-based likelihood ratios. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3):357–375.

[23] Kim, J., Park, S., and Carriquiry, A. (2024). A deep learning approach for the comparison of handwritten documents using latent feature vectors. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 17(1):e11660.

[24] Kocher, M. and Savoy, J. (2017). Distance measures in author profiling. *Information processing & management*, 53(5):1103–1119.

[25] Maltoni, D., Maio, D., Jain, A. K., and Prabhakar, S. (2009). *Synthetic Fingerprint Generation*, pages 271–302. Springer London, London.

[26] Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., Mazzella, W. D., Taroni, F., and Hicks, T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, 56(5):364–370.

[27] Meuwly, D., Ramos, D., and Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic science international*, 276:142–153.

[28] Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2):173–197.

[29] Morrison, G. S. and Enzinger, E. (2018). Score based procedures for the calculation of forensic likelihood ratios–scores should take account of both similarity and typicality. *Science & Justice*, 58(1):47–58.

[30] Neumann, C. and Ausdemore, M. (2020). Defence against the modern arts: the curse of statistics—Part II: 'Score-based likelihood ratios'. *Law, Probability and Risk*, 19(1):21–42.

[31] Neumann, C., Hendricks, J., and Ausdemore, M. (2020). Statistical support for conclusions in fingerprint examinations. *Handbook of Forensic Statistics, CRC, Boca Raton, FL*, pages 277–324.

[32] Nordgaard, A., Ansell, R., Drotz, W., and Jaeger, L. (2012). Scale of conclusions for the value of evidence. *Law, probability & risk*, 11(1):1–24.

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

[34] Ommen, D. M., Saunders, C. P., and Neumann, C. (2017). The characterization of monte carlo errors for the quantification of the value of forensic evidence. *Journal of Statistical Computation and Simulation*, 87(8):1608–1643.

[35] Pan, K., Chen, J., and Kafadar, K. (2020). Forensic glass evidence. In *Handbook of Forensic Statistics*, pages 411–442. CRC Press 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742.

[36] Park, S. and Tyner, S. (2019). Evaluation and comparison of methods for forensic glass source conclusions. *Forensic science international*, 305:110003.

[37] Saunders, C. P., Davis, L. J., Lamas, A. C., Miller, J. J., and Gantz, D. T. (2011). Construction and evaluation of classifiers for forensic document analysis. *The Annals of Applied Statistics*, 5(1):381 – 399.

[38] Stern, H. S. (2017). Statistical issues in forensic science. *Annual Review of Statistics and Its Application*, 4:225–244.

[39] Sugiyama, M., Suzuki, T., and Kanamori, T. (2010). Density ratio estimation: A comprehensive review (statistical experiment and its related topics). *RIMS Kokyuroku*, 1703:10–31.

[40] Veneri, F. and Ommen, D. (2021). An evaluation of score-based likelihood ratios for glass data. Master's thesis, Iowa State University.

[41] Veneri, F. and Ommen, D. M. (2023). Ensemble learning for score likelihood ratios under the common source problem. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 16(6):528–546.

[42] Vergeer, P. (2023). From specific-source feature-based to common-source score-based likelihood-ratio systems: ranking the stars. *Law, Probability and Risk*, page mgad005.

[43] Willis, S., Aitken, C., Barrett, A., Berger, C., Biedermann, A., Champod, C., Hicks, T., Lucena-Molina, J., Lunt, L., McDermott, S., McKenna, L., Nordgaard, A., O'Donnell, G., Rasmusson, B., Sjerps, M., Taroni, F., and Zadora, G. (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science*. European Network of Forensic Science Institutes, http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

3.8 Appendix: Algorithm Illustrations

To demonstrate the use of both algorithms outlined in section 3.3.2, and compare it to a naive approach disregarding dependence, we provide an illustration of how synthetic learning instances are generated for a simple specific source problem and an introduction to previous resampling plans for questioned documents.

3.8.1 Synthetic source ilustration

Our illustration follows a simplified multivariate two-level Gaussian data-generating process described in Section 3.4.1, with only five sources and ten items each, where the features vector is bi-dimensional.

As before let, $u_{ij} \in R^2$ denote the feature vector of the j^{th} item $(j = 1, \ldots, 5)$ of the i^{th} source $(i = 1, \ldots, 10)$.

Panel A in Figure 3.7 illustrates a realization of the data-generating process outlined in equations 3.11 and 3.12 to simulate the data. We selected $\mu = (0, 0), \Sigma_b = I_2, \Sigma_w = I_2/10$, where I_2 denotes a 2×2 identity matrix for simplicity. The theoretical contours are added for each source as references.

For the following steps, we consider source three as the specific source of interest and the remaining as part of the background population.

Panel B in Figure 3.7 illustrates how pairwise comparisons have traditionally been used in the source anchored specific source problems. This results in items being used multiple times. In particular, the ones from the source of interest.

Panel C shows how Algorithm 4 was used to generate a synthetic item depicted as a triangle. First, an item and one of its nearest neighbors were sampled, and a new measurement was generated by interpolating the measurements from two original items (connected by a line segment).

The third panel depicts the result of using Algorithm 5 to generate synthetic learning instances for the source anchored specific source problem. Twelve new measurements are generated for specific sources, eight are paired to generate learning instances (solid lines) for the prosecutor, while the four remaining are paired with one of the synthetic items generated for each source (dashed lines) to create learning instances for the defense.

In our illustration, the feature vectors are the coordinates in the Euclidean plane ($u_{ij} = (x_{ij}, y_{ij})$). If the Euclidean distances are selected as dissimilarity metrics between two items, Panel D also offers an intuition about the likelihood of the scores. An item is more likely to come from source three, the specific source, if it presents a shorter distance.

3.8.2 Resampling algorithms for questioned documents

Resampling methods in questioned documents have been previously proposed for the specific source problem. For our work, we implemented a modified version of the approach initially introduced by Davis, Saunders, Hepler and Buscaglia [12, 20].

Each questioned document or item can be considered a string of symbols, from which forensic document examiners can extract similarities in handwriting styles or transform symbols into graphs that can be classified into groups. We followed the method proposed by previous CSAFE authors [11, 6] to decompose writing into graphs, but this approach has a longer history in handwriting analysis.

This decomposition approach creates a feature vector for each document u_{ij} consisting of counts or proportions that can be used to assess authorship since writers tend to reproduce similar writing patterns.

The previous author's approach to creating pseudo items can be broadly described as a resampling plan that partitions the string of characters, sampling random position, into two pseudo items, E^1_{ij} E^2_{ij} from which features are created u^1_{ij} u^2_{ij}. To create a large enough questioned document, the different j samples from the writer can be combined into a mega

document E_i [22] that can later be resampled. In contrast, our proposed approach creates synthetic items directly from the features vector without combining questioned documents.

We provide our implementation of the Davis et al. approach in Algorithms 6 and how it can be used to create scores for sets P and D in Algorithm 7.

As before, let E_{ij} be the j question document (item) created by the i writer (source). Let g_{ijk} denote a graph belonging to the question document, where k ($k = 1, \ldots, K$) indexes the position of the graph in the document. A fixed position k^* is randomly selected, splitting the document in two, and features are extracted for each pseudo document (Algorithm 6).

Algorithm 6 Split documents (SD)

Require: E_{ij}
  $k^* \leftarrow$ sample an integer between 1 to K.
  if $k^* + \lfloor K/2 \rfloor \leq H$ **then**
    $E_1 \leftarrow g_{ijk}$ $(k = k^*, \ldots, k^* + \lfloor K/2 \rfloor)$
  else ▷ Cycle to the start of the string
    $E_1 \leftarrow g_{ijk}$ $(k = 1, \ldots, K - (\lfloor K/2 \rfloor - k^*), k^*, \ldots, K)$
  end if
  $E_2 \leftarrow g_{ijk}$ such that $g_{ijk} \notin E_1$ ▷ Graphs not selected in E_1 are selected for E_2
  $u_l \leftarrow (E_l)$ for $l = 1, 2$ ▷ Extract features =0

Note: Algorithm adapted from Hepler et al. [12]. The original version did not cycle to the beginning of the string, resulting in documents with different lengths.

The algorithm is applied multiple times to create sets P and D in Algorithm 7. As before, we will assume that there is only one item from the suspect E_s (or that they have been combined into a mega document), and the same has been done for items in the alternative population $E = \{E_i : i = 1, \ldots, m\}$. If the researcher requires n^* learning instances, Algorithm 6 is applied n^* times to generate pairs of pseudo items over which the dissimilarity metric is computed to create scores for P. From the alternative population, n^* items are selected and compared to E_s to obtain the scores for D

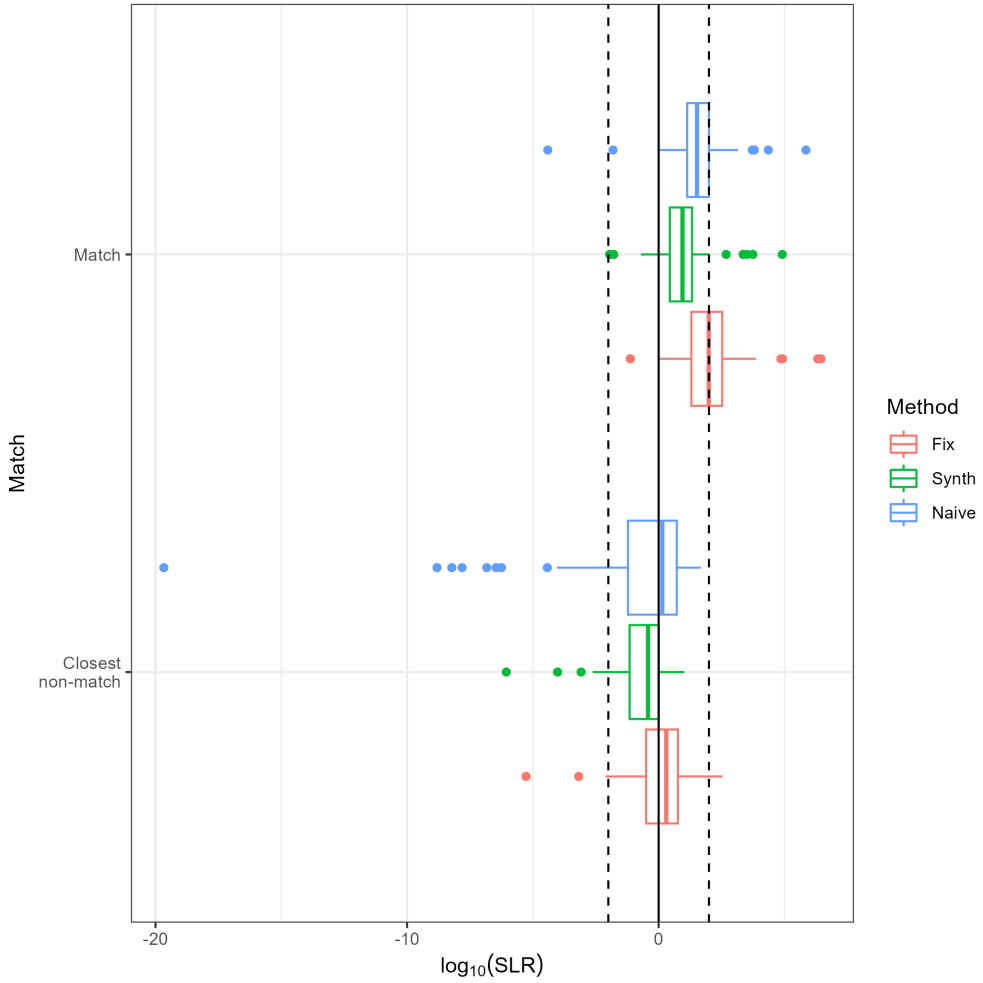
One potential drawback of this approach is that the document's content can impact the sampling process since items are split using a fixed position. A secondary concern is that while Algorithm 7 respects using alternative sources only once, the same questioned document is used for all defense scores. It is worth mentioning that the original algorithm was designed for

Algorithm 7 Subsampling for D, P

Require: E_s, E, n^*
  for $i = 1 : n^*$ **do** ▷ To generate score for P
    $(u_1, u_2) \leftarrow SD(E_s)$.
    $\delta_i^P \leftarrow \Delta(u_1, u_2)$
  end for
  $E_a^* \leftarrow$ Sample n^* items w/reposition from E $(a = 1, \ldots, n^*)$ ▷ To generate score for D
  for $i = 1 : n^*$ **do**
    $\delta_i^D \leftarrow \Delta(f(E_i^*), f(E_s))$ ▷ Extract features and compute disimilarity
  end for

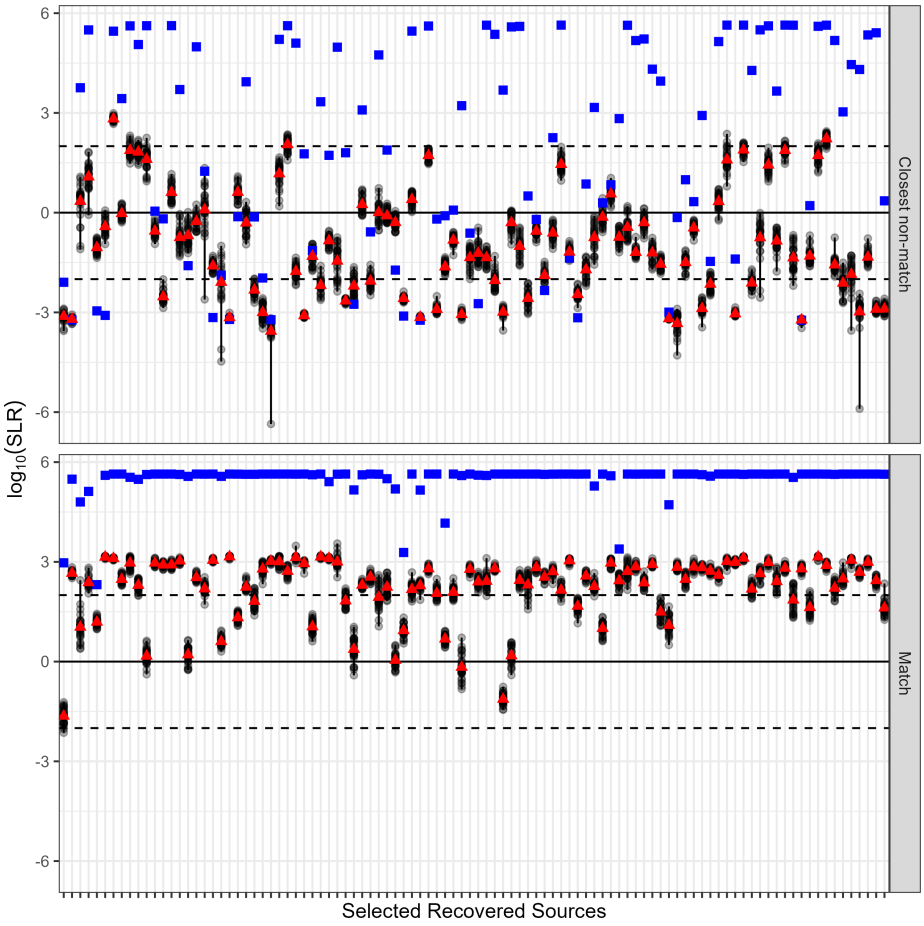
scenarios where only one questioned document was obtained from the person of interest; the approach proposed by our work (Algorithm 4) would not be suitable for this scenario but we believe can be a better alternative to creating a mega document when multiple items are available for the person of interest.

Figure 3.4 Boxplot of Score Likelihood Ratios for Handwriting Analysis



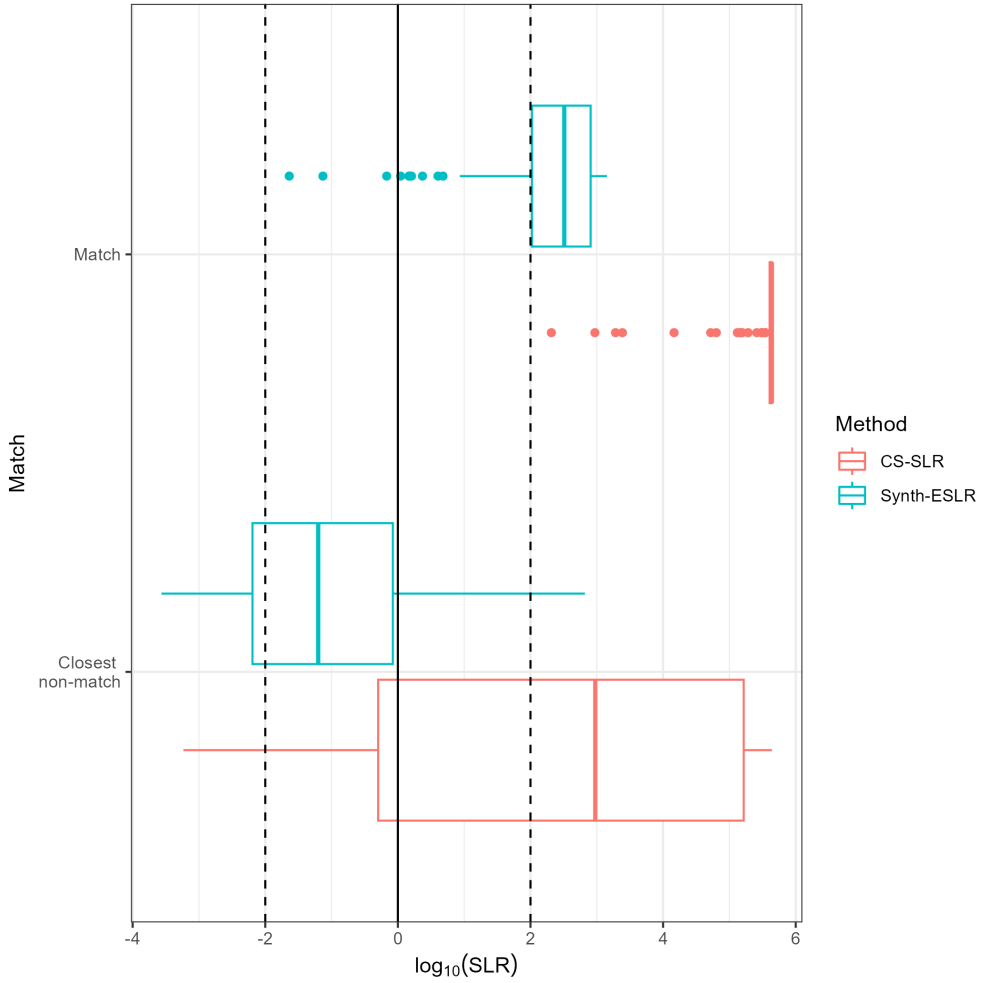
Note: Box plot of score likelihood ratios for the specific source problem on the log10 scale. The solid line represents the standard threshold to classify evidence toward the defense or prosecutor, and the dashed lines are the thresholds associated with at least moderately strong evidence (10^{-2} and 10^{2})

Figure 3.5 Distribution of Score Likelihood Ratios for Glass Analysis



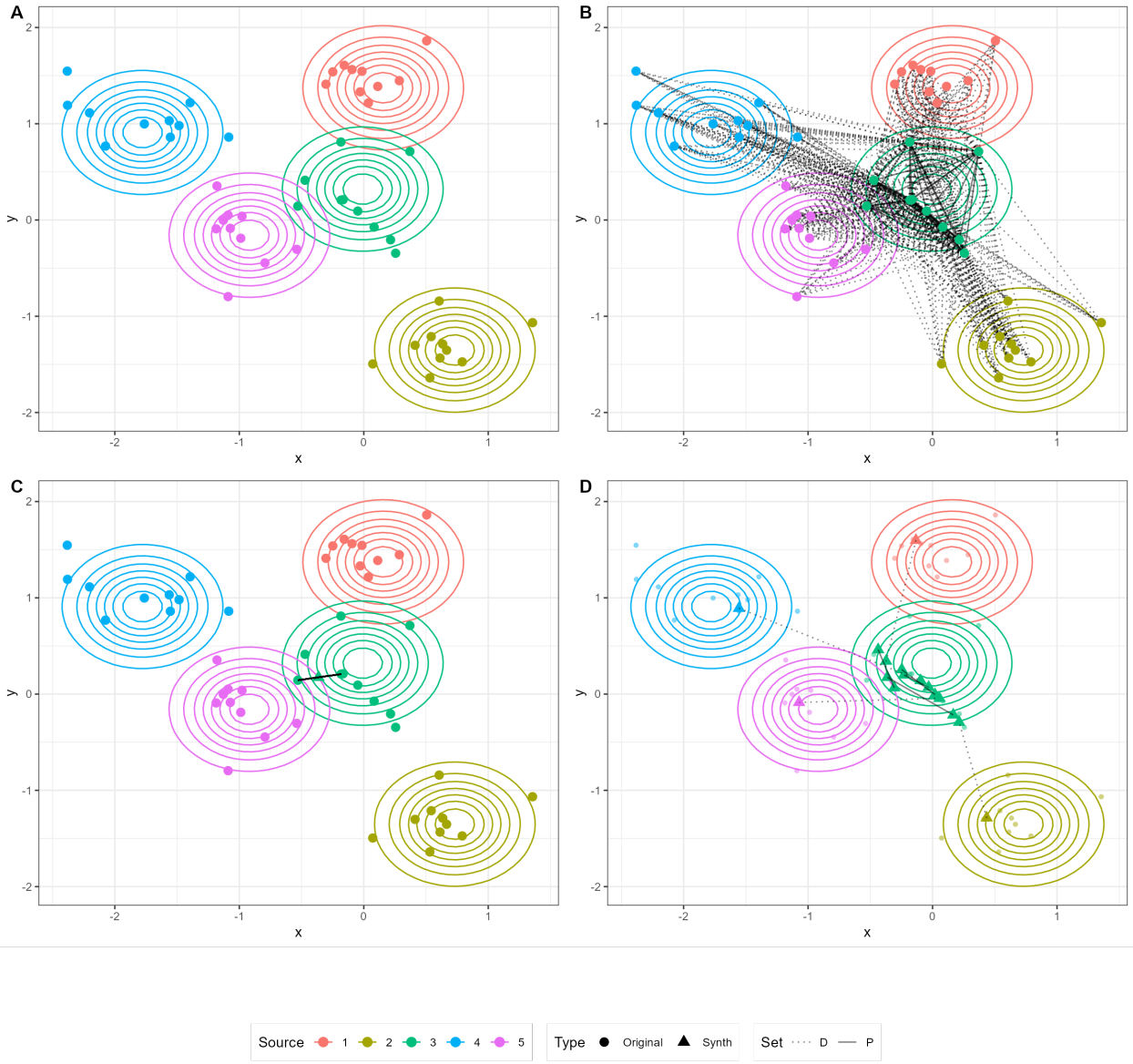
Note: Connected points indicate the output of 30 score likelihood ratios system for the same specific source problem on the log10 scale. The solid line represents the standard threshold to classify evidence toward the defense or prosecutor, and the dashed lines are the thresholds associated with at least moderately strong evidence (10^{-2} and 10^{2}). Red triangles denote the average evidence value for the specific source problem. Blue squares denote the evidence value obtained using a common source system.

Figure 3.6 Boxplot of Score Likelihood Ratios for Glass Analysis



Note: Box plot of score likelihood ratios for the specific source problem on the log10 scale. The solid line represents the standard threshold to classify evidence toward the defense or prosecutor, and the dashed lines are the thresholds associated with at least moderately strong evidence (10^{-2} and 10^{2}). Synth-ESLR denotes the synthetic anchored ensembled score likelihood ratio system, while CS-SLR denotes the system developed using the common source approach.

Figure 3.7 Algorithm Illustration for the Speficic Source problem



Note: Point and triangles represent items within sources. Contour was added as a reference for the DGP conditioned on each source. The specific source under consideration is source three

Panel A: Original data sampled, *Panel B*: Learning instances from pairwise comparisons under the specific source problem,*Panel C*: Generation of a synthetic item via interpolation (Algorithm 4), *Panel D*: Learning instances for synthetic source anchoring (Algorithm 5)

CHAPTER 4. DISCREPANCY METRICS TO EVALUATE MODEL MISSPECIFICATION AND DEPENDENCE EFFECTS IN SCORE-BASED LIKELIHOOD RATIO INFERENCE.

Federico Veneri

Department of Statistics, Iowa State University

4.1 Abstract

Score-based Likelihood Ratios have been proposed as an inferential tool for source attribution problems. In forensic statistics, the estimated likelihood of observing the derived scores under opposing propositions is used to guide judges' and jurors' decisions regarding the origin of items found at crime scenes. Developing a Score-based Likelihood Ratio system relies on creating learning instances where the origin of the items is known with certainty. From a background sample, items are paired, defining a match if the items originated from the same source, and comparison features are generated. This procedure induces a complex dependence structure, as items are used multiple times, violating the independence assumption made by popular estimation methods, which could affect the systems' performance. Our work introduces discrepancy metrics that allow us to study the effect of model misspecification, meaning selecting an estimation method that may not match the target density, failing to account for dependency, or both. We illustrate their use on a univariate example, the basis for our simulation study, where we compare the traditional approach to weak and strong source resampling. Simulation results show that while the induced dependence affects the inference drawn, there is a potential tradeoff between thinning out the dependence and sample size. Weak source resampling performs on par with the traditional approach, while strong source resampling presents mixed results depending on the estimation methods used. These initial results suggest that some estimation methods may be more robust to dependence while others may benefit from resampling strategies.

4.2 Introduction

Source attribution is a statistical inference problem where the researcher aims to assess the likelihood of opposing propositions regarding how items were generated. For instance, under the common source problem, the proposition traditionally denoted as the prosecutor's (H_p) states that the items compared share a common unknown source. In contrast, the defense proposition (H_d) states that the two items were independently generated by two unknown sources.

Score likelihood ratios have gained terrain as an alternative to classical likelihood ratios and Bayes factors for source attribution problems [17]. This is partly due to the challenge of developing a formal probability model for the opposing proposition and the advancements in machine learning to handle complex data [6]. For instance, in forensic analysis of handwriting and shoe impressions, images are collected to characterize the pattern evidence found at crime scenes. Machine learning has been used to extract features and compute a lower dimensional score that can be used as learning instances to develop a score-based likelihood ratio system.

Previous authors have addressed how learning instances should be created. For the common source, Neumann et al. [14] and Ommen and Saunders [15] proposed a thought experiment to establish how the estimation sets could be constructed for a distance-based metric. Under the prosecutor's proposition, the sampling distribution can be studied by considering a sample from sources and comparing a single pair of items from each source. Under the defense proposition, the distribution of the score can be studied by sampling an independent pair of sources and comparing an item from the first source to an item from the second source.

In practical applications, researchers have used pairwise comparisons to generate learning instances using a sample of background population where the sources associated with each item are known with certainty. When two items are paired, new comparison features are generated, and the pair is considered a known match if they share the same source or a known not-match if they don't. As all possible pairs are considered, this induced a complex dependence structure since items and sources are used multiple times, violating the independence assumption made by popular estimation methods and could affect the performance of the systems developed.

Recent work has addressed these issues and proposed resampling plans to account for dependence generated under the common source problem [21]. Strong Source Resampling (SSR) imposes a restriction such that the source and items are used only once to create a learning instance. We expand on this idea by introducing Weak Source Resampling (WSR), which imposes a weaker restriction, that items are used only once. While both proposals aim to emulate the sampling model that generated the observed data under both propositions, following how learning instances should be created more closely, they result in smaller sample sizes than the traditional approach.

To make the most out of the data collected, the authors proposed using resampling plans to develop base score-based likelihood ratio systems equivalent to weak learners, and their output aggregated into a final value of evidence.

While this approach has been shown to improve the inference drawn from score-based likelihood ratio systems, no previous research has explored the conditions that would guarantee a better performance than the traditional approach. In particular, we are interested in examining scenarios associated with different levels of induced dependencies and comparing the performance of resampling methods to the traditional approach.

To answer this question, we first review the inferential problem and sampling associated with the data-generating process for the common source problem in Section 4.3, and how score-based likelihood ratios can be estimated using the traditional approach and resampling methods in Section 4.4. We then introduce discrepancy measures in Section 4.5.1 that allows us to examine the effect of model misspecification. We illustrate their use on an univariate example where the Score Likelihood ratio system can be derived in closed form as the ratio of two half-normal densities in Section 4.6. The proposed example is the basis for the simulation study presented in Section 4.7, where the effect of different levels of dependency is explored for combinations of resampling methods (Traditional, Strong, and weak source resampling) and parametric estimation methods (Half Normal, Lognormal, Weibull, Gamma, and logit based density ratio estimator).

For our first simulation, the total sample size is fixed, and the dependence is augmented by increasing the number of items within sources; results show that scenarios with higher dependence are associated with larger expected discrepancy and worst empirical performance metrics. This result holds for any estimation methods used; smaller discrepancies are observed when a half normal is chosen as the estimation method, which matches the true system in our examples.

Our second simulation allows for a more realistic comparison by allowing the sample size and source-item relationship to vary. Results suggest that there may be a trade-off for resampling methods, as they result in smaller sample sizes. Weak source resampling ensembled approach performed on par with the traditional approach, while strong source resampling was associated with a larger expected discrepancy for some estimation methods. This was particularly true for the half-normal, which matches the theoretical target density in our illustration. Strong source resampling may contribute to smaller expected discrepancies when other estimation methods are used. This suggests that some estimation methods may be more sensible to the dependency introduced by using all potential pairs to create learning instances.

4.3 Sampling models and score likelihood ratio for common source

In source attribution problems, two opposing propositions, traditionally denoted as the prosecutor (H_p) and the defense (H_d), are contrasted. Let $\mathbf{E} = \{\mathbf{E}_x, \mathbf{E}_y\}$ denote the evidence under consideration consisting of two items. The propositions address the data-generating process that resulted in the observed evidence E, and can be generically stated as:

- $H_p : E_x$ and E_y were generated by the same unknown source.
- $H_d : E_x$ and E_y were generated by two different unknown sources.

Consider the case of ballistic examination in forensics. Suppose that two bullet casings (items) were recovered from two different crime scenes. The common source would refer to a common firearm, and the propositions essentially refer to the possibility that the two crime scenes are connected.

The conclusion regarding the likelihood of the propositions is reached by comparing measurements or feature vectors associated with both items denoted as $u_i \in \mathbb{R}^p (i = x, y)$.

To address source attribution problems, Ommen and Saunder [15] formalized the idea of a sampling model that generated the feature vector. Let $u_{ij} \in \mathbb{R}^p$ denote a generic feature vector from item j within source i. The mechanism that generated the observed data can be tough as a two-stage process sampling process. In the first stage, a latent random variable that characterizes the source B_i is sampled from $F_b(\cdot \mid \theta)$, where F_b describes the variation between sources and is characterized by population parameters θ. Conditional on the source, the feature vectors associated with the items are sampled from $F_w(\cdot \mid b_i, \theta)$ a density that characterizes the variation within sources. This can be summarized as:

$$B_i \sim F_b(\cdot \mid \theta) \text{ and } u_{ij} \overset{iid}{\sim} F_w(\cdot \mid b_i, \theta) \tag{4.1}$$

This general mechanism establishes how features are generated and allows the formalization of the prosecutor and defense proposition in terms of sampling models.

Under the prosecutor model, one unknown source common to both items is generated from F_b, and conditional on the source two feature vectors are generated.

$$B_u \sim F_b(\cdot \mid \theta) \tag{4.2}$$

$$\mathbf{u_x} \mid B_u = b_u \sim F_w(\cdot \mid b_u, \theta) \tag{4.3}$$

$$\mathbf{u_y} \mid B_u = b_u \sim F_w(\cdot \mid b_u, \theta) \tag{4.4}$$

$$\tag{4.5}$$

While under the defense model, two unknown sources have been generated from F_b, and conditional on each source, the features generated

$$B_x \sim F_b(\cdot \mid \theta) \text{ and } \qquad \mathbf{u_x} \mid B_x = b_x \sim F_w(\cdot \mid b_x, \theta) \tag{4.6}$$

$$B_y \sim F_b(\cdot \mid \theta) \text{ and } \qquad \mathbf{u_y} \mid B_y = b_y \sim F_w(\cdot \mid b_y, \theta) \tag{4.7}$$

If both densities were known, likelihood ratios could be computed by deriving the joint densities of u_x and u_y under both propositions. In practical applications, researchers can consider similarity metrics between items based on their feature vectors, which summarize how similar two objects are in a univariate score.

Let $\Delta(u_x, u_y)$ denote a similarity metric that allows the comparison of items E_x and E_y via their observed features u_x and u_y; and let $g\left(\delta \mid H_j\right)$ $(j = p, d)$ denote the conditional density of the univariate score under the propositions then a score likelihood ratio is defined as

$$\text{SLR}(\delta) = \frac{g\left(\delta \mid H_p\right)}{g\left(\delta \mid H_d\right)} \tag{4.8}$$

If the data-generating process was known, the conditional densities of the score under both propositions could be derived via the transformation theorem. And if the conditional densities result in a member of the exponential family, meaning the Radon-Nikodym derivative of the probability measure can be written as:

$$g(\delta \mid \theta, H_j) = h_j(\delta) v_j(\theta) \exp\left[\eta_j(\theta) T_j(\delta)\right], \tag{4.9}$$

where $v_j(\cdot)$, $\eta_j(\cdot)$ denotes functions of the parameters associated with the data-generating process and $h_j(\cdot)$ and $T_j(\cdot)$ are functions of the observed data, the score likelihood ratio function can be written in a more tractable form, and the following lemma follows.

Let $g(\delta \mid \theta, H_j)$ denote a density under one of the propositions $(j = p, d)$. If both conditional densities belong to an exponential family, the likelihood ratio for a score δ can be written as:

$$\text{SLR}(\delta \mid \theta_p, \theta_d) = \frac{h_p(\delta)}{h_d(\delta)} \frac{v_p(\theta)}{v_d(\theta)} \exp\left[\eta_p(\theta) T_p(\delta) - \eta_d(\theta) T_d(\delta)\right] \tag{4.10}$$

Both densities are not required to belong to the same exponential family or exponential families in general. Theoretical illustration often results in the same exponential family with different locations or scaling for the densities resulting in $h_p(\cdot) = h_d(\cdot)$ and score likelihood ratio as in the following Lemma.

When both conditional densities belong to the same exponential family, the likelihood ratio for a score δ is simplified as follows.

$$\text{SLR}(\delta \mid \theta_p, \theta_d) = \frac{v_p(\theta)}{v_d(\theta)} \exp\left[T(\delta)(\eta_p(\theta) - \eta_d(\theta))\right] \tag{4.11}$$

where $T(\delta)$ denotes a sufficient statistic.

Previous Equations (4.10 4.11 4.8) can be interpreted as the score likelihood ratio functions, the main tool to derive inference under the score-based likelihood ratio paradigm. For a given score δ, if $\text{SLR}(\delta) > 1$, the score is found to support H_p, meaning that the score is more likely under H_p than H_d. The opposite conclusion is derived when $\text{SLR}(\delta) < 1$. Different thresholds have been proposed to interpret the degree of strength associated with the evidence observed [7].

In the case of Equation 4.11, the inference can be summarized in terms of the statistic $T(\delta)$ and the decision rule stated in terms of $v_j(\cdot)$ and $\eta_j(\cdot)$ $(j = p, d)$.

In practical application, the "true" score likelihood ratio functions are not known, and researchers may apply different methods to estimate these functions.

4.4 Estimating score likelihood ratios

Different methods have been used within the realm of score likelihood ratio systems, a common distinction has been made between generative and discriminative methods [13, 10]. Starting from a sample $\left\{\delta_i^{(j)}\right\}_{i=1}^{n_j}$ generative methods aim to estimate the densities $\hat{g}_j(\delta)$ $(j = p, d)$, such that an estimated score likelihood ratio function is found by taking the ratio

$$\widehat{\text{SLR}}(\delta) = \frac{\hat{g}_p(\delta)}{\hat{g}_d(\delta)} \tag{4.12}$$

To select the appropriate parametric family, the researcher can examine the range of the scores and pre-select a subset of families to explore. If the same exponential family is chosen for both densities, the estimated score likelihood functions can be written as

$$\widehat{\text{SLR}}(\delta) = \frac{\hat{v_p}(\theta)}{\hat{v_d}(\theta)} \exp\left[T(\delta)(\hat{\eta_p}(\theta) - \hat{\eta_d}(\theta))\right] \tag{4.13}$$

In practice, the estimation procedure is done independently for each density.

On the other hand, discriminative methods aim to find the boundary between the densities and provide a density ratio estimator directly. In particular, logistic regression can be used as a probabilistic classifier density ratio estimator [13, 18, 21]

$$\widehat{\mathrm{SLR}}(\delta) = \frac{n_{KNM}}{n_{KM}} e(\hat{\beta}_0 + \hat{\beta}_1 \delta) \tag{4.14}$$

where the first component is the ratio of the sample sizes used to fit a logistic regression, and $\hat{\beta}$ denotes the resulting estimates trained to classify between known matches and non-matches.

Whichever method is selected, once estimates are obtained, the inference is done by plugging the observed score into the estimated score likelihood ratio function, and the interpretation is the same as before.

For the estimation method described, a common assumption is that a random sample of scores from the target densities was available, meaning

$$\left\{\delta_i^{(KM)}\right\}_{i=1}^{n_{KM}} \overset{\text{iid}}{\sim} g(\boldsymbol{\delta} \mid H_p) \quad \text{and} \quad \left\{\delta_j^{(KNM)}\right\}_{j=1}^{n_{KNM}} \overset{\text{iid}}{\sim} g(\boldsymbol{\delta} \mid H_d), \tag{4.15}$$

This matches the theoretical sampling models (Section 4.3) and how the estimation set should be created [14, 15]. However, as noted by Veneri and Ommen [21], researchers create samples using unrestricted pairwise comparison, resulting in multiple data points being used multiple times. This induced dependence structure may have some effects on the estimation methods and the quality of the inference drawn.

To account for this, the author's initial proposal was an iterative algorithm to thin out the dependence structure, ensuring that sources and items are used only once (SSR, Strong Source Resampling). We expand their idea to impose a less stringent condition, that items are used only once (WSR, Weak Source Resampling). We summarize these approaches and present the pseudocode for resampling implementation.

Let $A = \{u_{ij} : i = 1, \ldots, m; j = 1, \ldots, n\}$ denote a background sample or initial dataset available to researchers, where u_{ij} denotes the feature vectors, i index sources, and j index items

within sources. Without loss of generality, we assumed the same number of items within source n, and that m sources are available.

The traditional approach takes all possible pairwise comparisons, resulting in $\binom{mn}{2}$ comparisons. Consider u_{ij} and u_{kl} a potential pair, when $i = k$, meaning the items share the same source they are denoted as a known match (KM), and a known non-match (KNM) if they do not share the same source. This results in $n_{KM} = \binom{n}{2}m$ known matches and the reminder $n_{KNM} = \binom{mn}{2} - \binom{n}{2}m$ are known non-matches.

An alternative to the traditional approach are resampling plans that consider the data's hierarchical nature and thin out the dependence induced by taking all pairwise comparisons.

Strong Source Resampling (Algorithm 8) ensures that sources are used only once. In their main implementation [21] splits the available source in thirds, $m^* = \lfloor m/3 \rfloor$. If the source was selected to create known matches (set 1), two items are resampled within each of them independently and paired. If the sources were selected to create known non-matches (sets 1 and 3), one item is sampled for each source in each set and paired to generate m^* known non-matches. The result of this algorithm is an estimation set where sources (and items) are used only once.

Algorithm 8 Strong Source Resampling

Split sources into three sets: set 1, set 2, and set 3.
For each source in set 1, sample two items. This will generate the KM pairs.
For each source in set 2, sample one item.
For each source in set 3, sample one item.
Pair the items from two previous steps to generate the KNM pairs.

Note: Pseudocode first introduced in Veneri and Ommen [21].

While Algorithm 8 is appropriate in terms of independence and adequacy with the data-generating process (Section 4.3) it may impose a too stringent condition, resulting in a loss of data. In the same spirit, we implemented a weaker version that provides a middle ground between the traditional approach and the correct data-generating process. Weak Source Resampling (Algorithm 9) ensures that items are used only once to create comparisons.

Our simple implementation uses Algorithm 8. After strong source resampling is applied to set A, the items selected are removed. Strong source resampling is implemented again until there are

no more possible comparisons in A [1]. The result of this algorithm is an estimation set where items are used only once, but sources can be used multiple times.

Algorithm 9 Weak Source Resampling

```
while Set A is not empty do
    Apply SSR to set A
    Store comparison scores
    Remove items used from A.
end while
```

The result of both algorithms is an estimation set that can be used to estimate score likelihood ratio functions. Veneri and Ommen [21] proposed an ensemble approach to strengthen inference drawn, we follow the same idea to generate an ensemble approach for both weak and strong resampling methods and compare their performance to the traditional approach in our second simulation study.

4.5 Discrepancy and performance metrics for Score Likelihood Ratio inference

In practical application, the true score likelihood ratio function is unknown, and the evaluation of model performance is based on performance metrics computed over a validation set [11, 10] [2].

A popular metric is the rate of misleading evidence for (not) known matches, which is the percentage of cases where the system should have outputted a value (lower than one) larger than one, but the opposite occurred. This can be considered as the error rates in score-based likelihood ratio inference.

The log-likelihood ratio cost (C_{llr}) provides an aggregated metric of the model performance [19, 12]. Originally introduced for speaker recognition [5], the metric penalizes errors and weak evidence in the correct direction. Smaller values of C_{llr} are associated with better performance, and authors have established a threshold of $C_{llr} = 1$ to consider a system as uninformative. Still, there is no consensus on interpreting other values [19].

[1] We determine that there are no more comparisons available if the number of available sources is less than four, the minimum required for Algorithm 8

[2] For a more formal notation, we refer the reader to Appendix B of [21]

These two metrics evaluate the performance from a decision rule perspective, the last stage in the inference pipeline. However, fewer studies have concentrated on the theoretical aspect of Score Likelihood ratio-based inference.

Seminal Work by Royall [16] established the groundwork for interpreting evidential values and computing the probability of misleading evidence for likelihood ratios. Garton et.al. [8] addressed a probabilistic bound for the discrepancy between likelihood and Score Likelihood ratios. Our work follows this line of research, focusing on the discrepancy between a true score likelihood ratio and the estimated counterpart. We first derive discrepancy metrics in Section 4.5.1 and address computing the probability of misleading evidence as benchmarks in theoretical studies of score likelihood ratios. We illustrate their use in Section 4.6 with a univariate example, which is the basis of a simulation study in Section 4.7

4.5.1 Discrepancy metrics

In this section, we propose discrepancy metrics that can be used to evaluate the estimated score likelihood ratio function to their theoretical counterpart.

We assume that the scores under both propositions follow a known conditional density as in Equation 4.15 or can be derived from the data-generating process for the features. Let $\widehat{SLR}(\delta)$ denote an estimated score likelihood ratio function using a combination of a resampling and a parametric estimation method. And let $SLR(\delta)$ denote the true score likelihood ratio function. We will consider three discrepancy metrics based on the discrepancy function on Equation 4.16.

$$h(\delta) = \left|\log_{10}(\widehat{SLR}(\delta)) - \log_{10}(SLR(\delta))\right| = \left|\log_{10}\frac{\widehat{SLR}(\delta)}{SLR(\delta)}\right| \tag{4.16}$$

A common practice in score likelihood-based inference is considering the log10 version, resulting in evidential value in $(-\infty, \infty)$. Where the zero threshold is associated with the change of support towards one of the propositions, negative values are associated with evidence supporting the defense, and positive values support the prosecutor.

When the true SLR function can be expressed as the ratio of two conditional densities as in Eq 4.8 and their empirical counterpart were also independently estimated, the discrepancy function can be rewritten as

$$h(\delta) = \left|\log_{10} \hat{SLR}(\delta) - \log_{10} SLR(\delta)\right| \tag{4.17}$$

$$= |\log_{10} g_p(\delta) - \log_{10} g_d(\delta) - \log_{10} \hat{g}_p(\delta) + \log_{10} \hat{g}_d(\delta)| \tag{4.18}$$

$$= \left|\log_{10} \frac{g_p(\delta)}{\hat{g}_p(\delta)} + \log_{10} \frac{\hat{g}_d(\delta)}{g_d(\delta)}\right| \tag{4.19}$$

$$= \left|\log_{10} \frac{g_p(\delta)}{\hat{g}_p(\delta)} - \log_{10} \frac{g_d(\delta)}{\hat{g}_d(\delta)}\right| \tag{4.20}$$

to highlight the discrepancy between the estimated and true conditional densities under the prosecutor and defense propositions.

By considering the triangle inequality, Upper and lower bounds can be derived,

$$\left|\left|\log_{10} \frac{g_p}{\hat{g}_p}\right| - \left|\log_{10} \frac{\hat{g}_d}{g_d}\right|\right| \qquad \text{(From triangle ineq)} \tag{4.21}$$

$$\leq \left|\log_{10} \frac{g_p(\delta)}{\hat{g}_p(\delta)} - \log_{10} \frac{g_d(\delta)}{\hat{g}_d(\delta)}\right| \tag{4.22}$$

$$= \left|\log_{10(\delta)} \frac{g_p(\delta)}{\hat{g}_p(\delta)} + \log_{10} \frac{\hat{g}_d(\delta)}{g_d(\delta)}\right| \tag{4.23}$$

$$\leq \left|\log_{10} \frac{g_p(\delta)}{\hat{g}_p(\delta)}\right| + \left|\log_{10} \frac{\hat{g}_d(\delta)}{g_d(\delta)}\right| \qquad \text{(From triangle ineq)} \tag{4.24}$$

Since the $h(\delta)$ is lower bounded by zero when the estimated and true densities match, we will consider only the upper bound for our discrepancy metrics.

The first metric considered (Equation 4.25) evaluates the maximum discrepancy observed between the two functions, similar to Kolmogorov-Smirnov type discrepancy statistic.

$$KS = \sup_{\delta} h(\delta) \tag{4.25}$$

$$\delta^{KS} = \arg\max_{\delta} h(\delta) \tag{4.26}$$

where KS denotes the maximum observed and δ^{KS} the associated score. While it is straightforward to compute, the point of maximum discrepancy can be found using a simple grid

search, a desirable property of a score likelihood system is monotonicity. In the case of a similarity score, smaller values should be associated with more similar items and more likely under the prosecutor. Larger values of the score depict more dissimilar items and should be more likely under the defense. This holds for the parametric estimation methods described in Section 4.4; hence, it may not be an informative metric since we expect a more significant discrepancy to be found for extreme values of the scores. However, this metric can be informative when non-parametric methods (e.g. kernel density estimation) are used, as they have been shown to result in nonmonotonic score likelihood ratio functions [20, 10].

Another drawback of the Kolmogorov-Smirnov-inspired metric is that it fails to consider which proposition is true. The discrepancy could affect one proposition more than the other. Our second metric is inspired by Von-misses type statistics, and we denote it as the expected discrepancy under the prosecutor and the defense (Equation 4.27).

$$E_j[h(\delta)] = \int h(\delta) g(\delta \mid H_j) d\delta \tag{4.27}$$

Were $E_j[h(\delta)]$ indicates the expected discrepancy under the proposition H_j being true. If the score likelihood ratio and its empirical counterpart are estimated independently, the previous inequality in Equation 4.24 can be used to find an upper bound of the expected discrepancy.

$$E_j[h(\delta)] \leq E_j\left[\left|\log_{10}\frac{g_p(\delta)}{\hat{g}_p(\delta)}\right| + \left|\log_{10}\frac{\hat{g}_d(\delta)}{g_d(\delta)}\right|\right] \tag{4.28}$$

$$= E_j\left|\log_{10}\frac{g_p(\delta)}{\hat{g}_p(\delta)}\right| + E_j\left|\log_{10}\frac{\hat{g}_d(\delta)}{g_d(\delta)}\right| \tag{4.29}$$

$$= \int \left|\log_{10}\frac{g_p(\delta)}{\hat{g}_p(\delta)}\right| g_j(\delta) d\delta \tag{4.30}$$

$$+ \int \left|\log_{10}\frac{\hat{g}_d(\delta)}{g_d(\delta)}\right| g_j(\delta) d\delta \tag{4.31}$$

Hence, the expected discrepancy under one of the propositions can be upper bounded by two expected divergences that can be interpreted as the discrepancy between estimated and true densities weighted by a density associated with the true proposition.

Our last discrepancy statistic proposes an alternative by considering an integration area critical for the system. Previous authors [9, 21] have considered cut-off values for the score likelihood ratio functions that can be interpreted as weaker evidence towards the prosecutor or the defense [22] and define an inconclusive range.

Let $\delta_t = \{\delta \mid \log_{10} SLR(\delta) = t\}$ denote the scores that achieve a specific threshold t; for instance in the log10 scale $t \in (-2, 2)$ have been considered as weak or moderate evidence range. We can define A_δ as the score that falls in a particular range associated with weak or moderate evidence, and define the discrepancy over the inconclusive area as

$$I[h(\delta)] = \int h(\delta) \mathbb{1}_{A_\delta} d\delta \tag{4.32}$$

As before, an upper bound for the discrepancy function can be found using Equation4.24

$$I[h(\delta)] \leq \int \left| \log_{10} \frac{g_p(\delta)}{\hat{g}_p(\delta)} \right| \mathbb{1}_{A_\delta} d\delta \tag{4.33}$$

$$+ \int \left| \log_{10} \frac{\hat{g}_d(\delta)}{g_d(\delta)} \right| \mathbb{1}_{A_\delta} d\delta \tag{4.34}$$

While this metric focuses on an area deemed critical to inference, some drawbacks should be considered. As previously mentioned, monotonicity is a desirable property for a score likelihood ratio function. Since A_δ is defined over the true function, which we can assume fulfills this property, the integration area is not affected by the estimation method. However, the thresholds are not always achieved in some problems which can complicate the interpratation of this metric.

We further discuss the application and limitation of the different metrics in our illustration in Section 4.6. Overall, we argue that the expected discrepancy is the most informative and has a clear interpretation in terms of an expectation under the prosecutor or defense being correct.

4.5.2 Theoretical rate of misleading evidence and thresholds

In the context of studying properties of score likelihood ratio functions, when $SLR(\delta)$ is known, rather than studying the empirical rate of misleading evidence as described in Section 4.5,

it is worthwhile to study the theoretical error rate which provides a natural benchmark to assess the empirical performance of estimated score likelihood ratio functions.

Misleading evidence can be observed even if a model has been correctly specified and all assumptions are met. Royal [16] addresses the probability of observing strong misleading evidence for likelihood ratio inference, the existence of universal bounds, and illustrates how the probability can be considerably smaller than the bounds.

We follow a similar approach to compute the probability of observing misleading evidence in terms of model parameters. In the context of score-based inference, misleading evidence for the prosecutor (defense) means that δ is more likely under H_p (H_d) than H_d (H_p) resulting in a ratio larger than one (smaller than one). We can denote these conditional probabilities as

$$P(SLR(\delta) > 1 \mid H_d) = P_d(SLR(\delta) > 1) \tag{4.35}$$

$$P(SLR(\delta) < 1 \mid H_p) = P_p(SLR(\delta) < 1) \tag{4.36}$$

where $SLR(\delta)$ denotes the true function. When Lemma 4.3 holds, and the parameters are known, the probability of misleading evidence can be written as a function of the statistic $T(\delta)$, and its distribution under the prosecutor or defense used to compute the desired conditional probability.

A similar argument can be used to compute the probability of observing a score over any threshold selected. We provide an illustration in Section 4.6.

4.6 An univariate Illustration

To illustrate the discrepancy metrics, we will consider an example in forensic glass analysis: the refractive index. The refractive index is a common feature that can be measured in glass. The sources could be window panes, and the items could be glass fragments found during a criminal investigation [3]. Hence u_{ij} denotes a univariate measurement taken from $j-th$ fragment (item) that came from the $i-th$ window (source).

[3] See Section 1.3.3 in [1] for more details and historical overview

The logic behind the inference is that sources are associated with an overall refractive index value. While there is variability within sources, resulting in different refractive indices observed in items from the same source, the variability between sources is larger. This allows us to conclude that if two fragments (items) share a similar refractive index, it is more likely the same source has generated them.

We will assume that the measurements follow a two-level Gaussian distribution following. The data-generating process described in Equation 4.1 can written as :

$$B_i \sim N\left(\cdot \mid \mu_a, \sigma_b\right) \text{ and } u_{ij} \overset{iid}{\sim} N\left(\cdot \mid b_i, \sigma_w\right) \tag{4.37}$$

for this illustration, where μ_a denotes an overall refractive index, σ_b the parameter associated with the between source variation, and σ_w the within source variation. Both F_b and F_w are assumed to be Gaussian.

In the case of the common source problem, the two measurements u_x and u_y, will follow a joint multivariate distribution with the covariance structure depending on the proposition (Equation 4.38 and 4.39). The measurements are independent under H_d , while under H_p they have a covariance σ_b^2 since they share the same source.

$$\begin{pmatrix} u_x \\ u_y \end{pmatrix} \mid H_p \sim N\left[\begin{pmatrix} \mu_a \\ \mu_a \end{pmatrix}, \begin{pmatrix} \sigma_w^2 + \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_w^2 + \sigma_b^2 \end{pmatrix}\right] \tag{4.38}$$

$$\begin{pmatrix} u_x \\ u_y \end{pmatrix} \mid H_d \sim N\left[\begin{pmatrix} \mu_a \\ \mu_a \end{pmatrix}, \begin{pmatrix} \sigma_w^2 + \sigma_b^2 & 0 \\ 0 & \sigma_w^2 + \sigma_b^2 \end{pmatrix}\right] \tag{4.39}$$

In practical applications, it has been noted that the between variation is considerably larger than the within variation, $\sigma_b^2 \gg \sigma_w^2$.

In this illustration, likelihood ratios can be easily computed from the ratio of the joint distribution under the prosecutor and defense. We will consider distance as a score to develop a score likelihood ratio system instead.

Several scores have been used in forensic statistics [8, 4, 3]. We will focus on three classical distances

$$\Delta_{2d}(u_x, u_y) = (u_x - u_y)^2, \tag{4.40}$$

$$\Delta_{L_2}(u_x, u_y) = \sqrt{(u_x - u_y)^2} \tag{4.41}$$

$$\Delta_{L_1}(u_x, u_y) = |(u_x - u_y)| \tag{4.42}$$

In a univariate setting, the L_1 and L_2 are equivalent, but we will treat them differently for the moment. The following lemma from Hepler et al. [9] on the properties of squared normal distribution and the additional two extra lemmas allow us to find the conditional distribution under propositions and distances. Note that under this setting, the choice of distance results in equivalent score likelihood functions; some will be more useful for our simulations later.

Let $X \sim N(\mu, \sigma^2)$ then $\frac{X^2}{\sigma^2} \sim \chi^2_{1,\lambda}$ a chi-square distribution with one degree of freedom and a non-centrality parameter $\lambda = \frac{\mu^2}{\sigma^2}$. The density function of X^2 can be written as a function

$$f_{X^2}(t) = \frac{1}{\sigma^2} f_{\chi^2_{1,\lambda}}\left(\frac{t}{\sigma}\right), \text{ with } \lambda = \frac{\mu^2}{\sigma^2} \tag{4.43}$$

Let $X \sim \chi^2_n$, then $\sqrt{X} \sim \chi_n$ meaning a chi-distribution with n degrees of freedom.

Let $X \sim \chi_1$, then $\sigma X \sim HN(\sigma)$, meaning a half-normal

Let $X \sim N(0, \sigma^2)$, then $|X| \sim HN(\sigma)$, meaning a half-normal

For all our scoring functions in 4.40, since we are considering the difference between two Gaussian random variables, it follows from properties of Gaussian distribution that the difference is Gaussian random variables with mean zero under both propositions, variance $2\sigma_w^2$ under H_p and $2\sigma_a^2$ under H_d.

For our first scoring function, Δ_{2d}. Applying Lemma 4.6 results in

$$g\left(\Delta_{2d}(u_x, u_y) = \delta \mid H_p\right) = \frac{1}{2\sigma_w^2} f_{\chi_1^2}\left(\frac{\delta}{2\sigma_w^2}\right) \tag{4.44}$$

$$g\left(\Delta_{2d}(u_x, u_y) = \delta \mid H_d\right) = \frac{1}{2\sigma_a^2} f_{\chi_1^2}\left(\frac{\delta}{2\sigma_a^2}\right) \tag{4.45}$$

Where $f_{\chi_1^2}$ denotes the density of a chi-square distribution with one degree of freedom,

$$f_{\chi_1^2}(x) = \frac{x^{-1/2} e^{-x/2}}{2^{1/2}\Gamma\left(\frac{1}{2}\right)}, \quad \delta > 0. \tag{4.46}$$

For our second distribution Δ_{L_2}, let σ denote generically σ_a or σ_w, we know that $((u_x - u_y)^2/2\sigma) \sim \chi_1^2$ and by lemma 4.6 it follows that

$$g\left(\Delta_{L2}(u_x, u_y) = \delta \mid H_p\right) = f_{HN}\left(\delta \mid \sigma = \sqrt{2\sigma_w^2}\right) \tag{4.47}$$

$$g\left(\Delta_{L2}(u_x, u_y) = \delta \mid H_d\right) = f_{HN}\left(\delta \mid \sigma = \sqrt{2\sigma_a^2}\right) \tag{4.48}$$

For our third scoring function, Δ_{L_1} applying lemma 4.6 to the difference of Gaussian distributions resutls in:

$$g\left(\Delta_{L1}(u_x, u_y) = \delta \mid H_p\right) = f_{HN}\left(\delta \mid \sigma = \sqrt{2\sigma_w^2}\right) \tag{4.49}$$

$$g\left(\Delta_{L1}(u_x, u_y) = \delta \mid H_d\right) = f_{HN}\left(\delta \mid \sigma = \sqrt{2\sigma_a^2}\right) \tag{4.50}$$

For both results, f_{HN} denotes the density of a half normal distribution with scaling parameter σ.

$$f_{HN}(x \mid \sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad \delta > 0 \tag{4.51}$$

As noted before, L_1 and L_2 will match in our illustration. In the following sections, we will focus on the L_1 norm and the squared difference to derive rates of misleading evidence and divergence results.

4.6.1 True Score Likelihood ratio functions, thresholds, and probability of misleading evidence

From the conditional densities obtained in Section 4.6 or applying Lema 4.3, the score likelihood function can be written as in Equation 4.11. Different thresholds and the probability of misleading evidence can be derived as a function of $T(\delta)$.

In the case of the L_1 and L_2 norms,

$$\mathrm{SLR}_{\mathrm{L1}}(\delta \mid \sigma^2_{\mathrm{w}}, \sigma^2_{\mathrm{b}}) = \frac{\frac{1}{\sqrt{2\sigma^2_w}} e^{\frac{-\delta^2}{4\sigma^2_w}}}{\frac{1}{\sqrt{2\sigma^2_a}} e^{\frac{-\delta^2}{4\sigma^2_a}}} = \left(\frac{\sigma^2_a}{\sigma^2_w}\right)^{1/2} e^{\frac{-\delta^2}{4}\left(\frac{\sigma^2_b}{\sigma^2_w \sigma^2_a}\right)} \tag{4.52}$$

And for the squared difference,

$$\mathrm{SLR}_{\mathrm{2d}}(\delta \mid \sigma^2_{\mathrm{w}}, \sigma^2_{\mathrm{b}}) = \frac{\sigma^2_a}{\sigma^2_w} \frac{f_{\chi^2_1}\left(\frac{\delta}{2\sigma^2_w}\right)}{f_{\chi^2_1}\left(\frac{\delta}{2\sigma^2_a}\right)} = \left(\frac{\sigma^2_a}{\sigma^2_w}\right)^{1/2} e^{-\frac{\delta}{4}\left(\frac{\sigma^2_b}{\sigma^2_w \sigma^2_a}\right)}. \tag{4.53}$$

Note that δ in Equation 4.53 is the square of the original δ from Equation 4.52, meaning that inference drawn by either system will match. In what follows, we will consider the L_1 score likelihood representation, where $T(\delta) = \delta^2$, for brevity. This case explicitly relates the initial parameters from the data-generating process to the scaling parameters of the half normal, which can be estimated using maximum likelihood when a sample of scores is obtained.

First, consider the $\log_{10}$ of the equation 4.52, we can write

$$\log_{10} \mathrm{SLR}_{L1}(\delta = \delta_t \mid \sigma^2_w, \sigma^2_b) = \frac{1}{2} \log_{10}\left(\frac{\sigma^2_a}{\sigma^2_w}\right) - \log_{10}(e) \frac{\delta^2_t}{4} \left(\frac{\sigma^2_b}{\sigma^2_w \sigma^2_a}\right) \tag{4.54}$$

Note that σ^2_w and σ^2_b are model parameters and only δ can be considered a random variable. The conditional distribution for the statistic $T(\delta) = \delta^2$ can be derived depending on the propositions considered.

Rates of misleading evidence and different thresholds can be derived from Equation 4.52, in close form. First we derive the asociated score δ_t to a generic threshold t in terms of the data generating process paramters an estimable model parameters following Eq 4.54.

$$\delta_t = \sqrt{\frac{4}{\log_{10}(e)}\left(\frac{\sigma_w^2\sigma_a^2}{\sigma_b^2}\right)\left[\frac{1}{2}\log_{10}\left(\frac{\sigma_a^2}{\sigma_w^2}\right) - t\right]} \tag{4.55}$$

$$= \sqrt{\frac{4}{\log_{10}(e)}\left(\frac{\sigma_p^2\sigma_d^2}{2(\sigma_d^2 - \sigma_p^2)}\right)\left[\frac{1}{2}\log_{10}\left(\frac{\sigma_d^2}{\sigma_p^2}\right) - t\right]} \tag{4.56}$$

The last equality results from noting that $\sigma_w^2 = \sigma_p^2/2$, $\sigma_b^2 = (\sigma_d^2 - \sigma_p^2)/2$,$\sigma_a^2 = \sigma_d^2/2$, and can be used to find δ_t by plug-in σ_j^2 $(j = p, d)$ or its estimates to find estimated threshold.

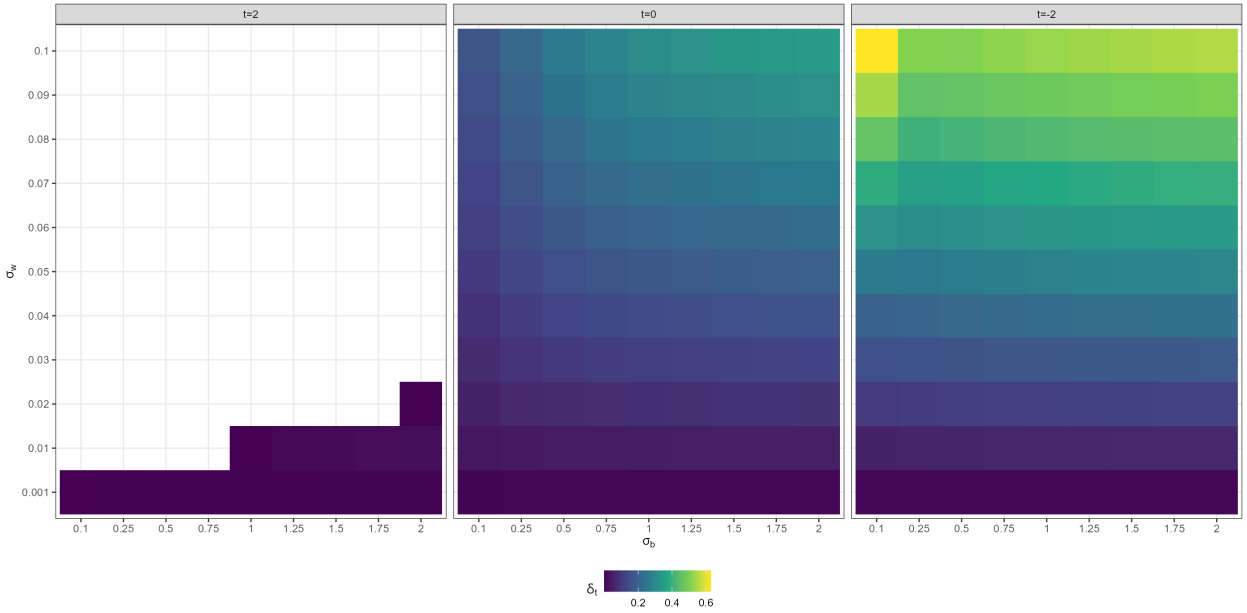
We illustrate the value δ_t for selected parameter values and t in Figure 4.1. The central panel depicts the scores associated with $t = 0$, the threshold that determines whether an SLR is leaning towards the prosecutor or defense, while the left panel $t = 2$ depicts the score associated with strong evidence toward the prosecutor and right panel $t = -2$ strong evidence toward the defense.

For a fixed value of σ_w, incresing σ_b, would result in a larger treshold required to distinguish between prosecutor and defense propositions. Intuitively, as there is more variability between sources, refractive indices will tend to overlap, requiring a larger score to decide between propositions. It is relevant to note that strong evidence towards the prosecutor is not always achieved since the scores are lower bounded by zero. Achieving a strong conclusion for the prosecutor would require a larger contribution of the between variance to the total variability. Achieving a strong conclusion for the defense does not face this limitation as the scores are not upper bounded, and in general, a smaller threshold is required when sources are more separable.

The previous result illustrates a potential weakness of these scores as the prosecutor may never observe strong evidence in his direction given certain model parameters and metrics like Equation 4.32 could be harder to interpret.

Besides these thresholds, a key aspect of score likelihood ratio functions is the probability of observing misleading evidence for the true system as described in Section 4.5.2. For this illustration, the probability for the defense can be computed as

Figure 4.1 Score associated with different thresholds by model parameters



Note: δ_t denotes the score associated with threshold t

$$P_d(SLR > 1) = P_d\left(\left(\frac{\sigma_a^2}{\sigma_w^2}\right)^{1/2} e^{-\frac{\delta^2}{4}\left(\frac{\sigma_b^2}{\sigma_w^2\sigma_a^2}\right)} > 1\right) \tag{4.57}$$

$$= P_p\left(\delta^2 \leq -2\ln\left(\frac{\sigma_w^2}{\sigma_a^2}\right)\frac{\sigma_a^2\sigma_w^2}{\sigma_b^2}\right) \tag{4.58}$$

$$= F_{\chi_1^2}\left(-\ln\left(\frac{\sigma_w^2}{\sigma_a^2}\right)\frac{\sigma_w^2}{\sigma_b^2}\right) \tag{4.59}$$

$$\tag{4.60}$$

and the probability for the prosecutor can be computed as

$$P_p(SLR < 1) = P_p\left(\left(\frac{\sigma_a^2}{\sigma_w^2}\right)^{1/2} e^{-\frac{\delta^2}{4}\left(\frac{\sigma_b^2}{\sigma_w^2\sigma_a^2}\right)} < 1\right) \tag{4.61}$$

$$= 1 - P_p\left(\delta^2 \leq -2\ln\left(\frac{\sigma_w^2}{\sigma_a^2}\right)\frac{\sigma_a^2\sigma_w^2}{\sigma_b^2}\right) \tag{4.62}$$

$$= 1 - F_{\chi_1^2}\left(-\ln\left(\frac{\sigma_w^2}{\sigma_a^2}\right)\frac{\sigma_a^2}{\sigma_b^2}\right) \tag{4.63}$$

$$\tag{4.64}$$

For both cases, the last equality comes from the fact that δ follows a half-normal distribution hence scaling and taking its square result in a Chi-square distribution.

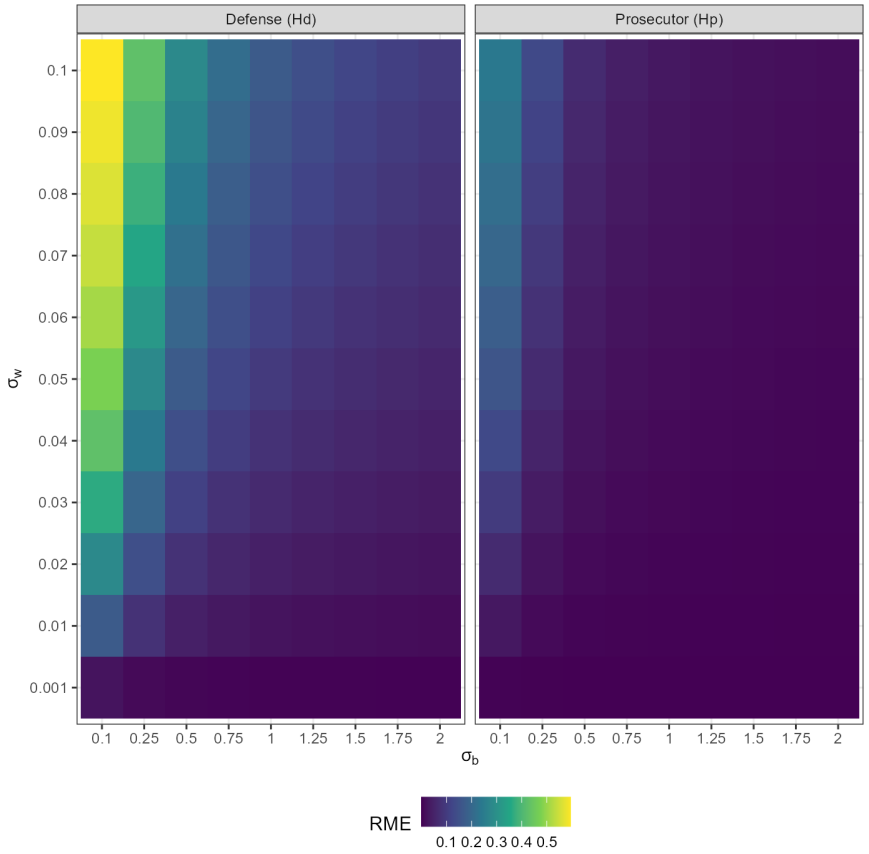
We illustrate the probability or rate of misleading evidence for selected model parameters (σ_b and σ_w) in Figure 4.2. When the between variability is a larger portion of the total variability ($\sigma_a^2 = \sigma_b^2 + \sigma_w^2$), both errors decrease as sources are easier to identify. However, the errors are not symmetric (Figure 4.3). There is a larger probability of obtaining misleading evidence for the defense, meaning concluding that the prosecutor is correct when the defense is correct, compared to the rate of misleading evidence for the prosecutor.

4.6.2 Discrepancy metrics

When the L_1 norm is considered, the target densities are half normal with their scaling parameters a function of the parameters from the process: $\sigma_p^2 = 2\sigma_w^2$ under the prosecutor and $\sigma_d^2 = 2\sigma_a^2$ under the defense. We illustrate the discrepancy measures derived in Section 4.5.1 for our example and provide a close form when the correct parametric family is chosen.

In practice, target densities are not known in advance. An alternative is to select within a collection of potential flexible parametric families that will adapt to the domain to obtain systems described by Equation 4.12. In our simulations, we explored the Weibull, Gamma, and LogNormal families since their support aligns with the potential values of the score. We also explore using a logit-based density ratio estimator to obtain systems described by 4.14. In the first case, discrepancy measure and their upper bounds can be computed numerically, while for

Figure 4.2 Rate of misleading evidence under different model parameters



Note: σ_b^2 denotes the between variance, σ_w^2 the within variance. RME denotes the rate of misleading evidence for the prosecutor (H_p) or defense (H_d) proposition

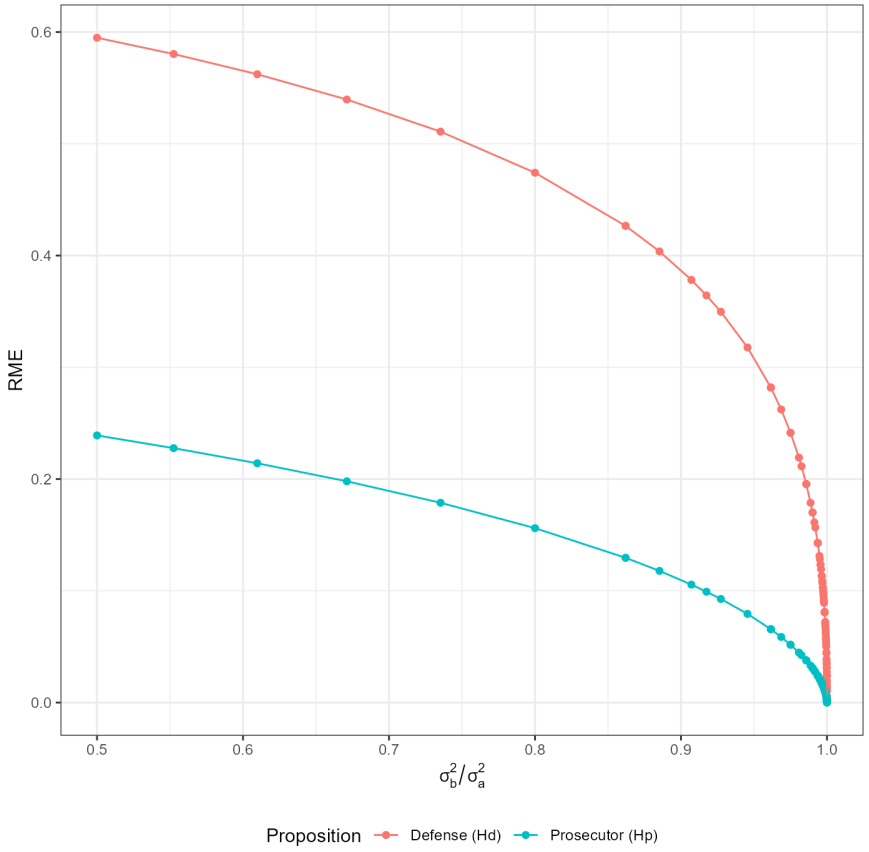
the density ratio estimator only the original metrics can be computed. We do not provide a close form for the discrepancy statistics under these alternatives.

If the correct family was chosen, in our case a half normal, but the model parameters are unknown. Reasonable estimates would result in a good approximation to the true densities resulting in small discrepancies.

When a sample of scores under a particular proposition is obtained, maximum likelihood estimation can be used to obtain the scaling parameter $\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} \delta_{ji}^2}{n} (j = h, p)$.

The resulting absolute values of the log densities that acted as the components of the upper bound of our discrepancy function (Equation 4.24) can be written as

Figure 4.3 Rate of misleading evidence by ratio of variability



Note: σ_b^2 denotes the between variance, σ_w^2 the within variance and $\sigma_a^2 = \sigma_w^2 + \sigma_b^2$ the total variance. RME denotes the rate of misleading evidence for the prosecutor (H_p) or defense (H_d) proposition

$$\left|\log_{10}\frac{g_j}{\hat{g}_j}\right| = \left|\delta^2\frac{\log_{10}(e)}{2}\left(\frac{\sigma_j^2-\hat{\sigma_j}^2}{\sigma_j^2\hat{\sigma_j}^2}\right)+\log_{10}\left(\frac{\hat{\sigma}_j}{\sigma_j}\right)\right| \tag{4.65}$$

$$\leq \delta^2\frac{\log_{10}(e)}{2}\left|\left(\frac{\sigma_j^2-\hat{\sigma_j}^2}{\sigma_j^2\hat{\sigma_j}^2}\right)\right|+\left|\log_{10}\left(\frac{\hat{\sigma}_j}{\sigma_j}\right)\right| \tag{4.66}$$

These results can be used to summarize several discrepancy statistics, as they provide an upper bound to the integrands.

Under regularity conditions and iid sample, the MLE estimator is consistent for the true parameters[4], meaning $\hat{\sigma}_j^2 \xrightarrow{p} \sigma_j^2$ (j=h,p), hence $\hat{\sigma}_j^2 - \sigma_j^2 \xrightarrow{p} 0$ and by the continuous mapping theorem $\log_{10}\left(\frac{\hat{\sigma}_j}{\sigma_j}\right) \xrightarrow{p} 0$.

Hence, the components of the upper bound will converge to zero if the estimators are consistent, and the discrepancy will go to zero. For completition, we present the closed form and upper bounds for the different discrepancy metrics in our illustration.

First we apply results found in Equation 4.66 to Equation 4.24 resulting in

$$h(\delta) = \left|\log_{10}\frac{g_p(\delta)}{\hat{g}_p(\delta)}+\log_{10}\frac{\hat{g}_d(\delta)}{g_d(\delta)}\right| \tag{4.67}$$

$$\leq \left|\log_{10}\frac{g_p(\delta)}{\hat{g}_p(\delta)}\right|+\left|\log_{10}\frac{\hat{g}_d(\delta)}{g_d(\delta)}\right| \quad \text{(From)} \tag{4.68}$$

$$\leq \delta^2\frac{\log_{10}(e)}{2}\left(\left|\left(\frac{\sigma_p^2-\hat{\sigma_p}^2}{\sigma_p^2\hat{\sigma}_p^2}\right)\right|+\left|\left(\frac{\hat{\sigma}_d^2-\sigma_d{}^2}{\sigma_d^2\hat{\sigma_d}^2}\right)\right|\right)+\left|\log_{10}\left(\frac{\hat{\sigma}_p}{\sigma_p}\right)\right|+\left|\log_{10}\left(\frac{\sigma_d}{\hat{\sigma}_d}\right)\right| \tag{4.69}$$

Hence, for our illustration, the supremum of $h(\delta)$ is bounded, and the bound increases as $\delta \to \infty$. For a fixed delta, the bound archives its smallest value if the estimates are consistent.

In the case of the expected discrepancy,

[4] We refer the reader to classical books like [2]

$$E_j[h(\delta)] = \int h(\delta) g(\delta \mid H_j) d\delta \tag{4.70}$$

$$\leq \int g(\delta \mid H_j) \left[\delta^2 \frac{\log_{10}(e)}{2} \left(\left| \left(\frac{\sigma_p^2 - \hat{\sigma_p}^2}{\sigma_p^2 \hat{\sigma}_p^2} \right) \right| + \left| \left(\frac{\hat{\sigma}_d^2 - {\sigma_d}^2}{\sigma_d^2 \hat{\sigma_d}^2} \right) \right| \right) + \left| \log_{10} \left(\frac{\hat{\sigma}_p}{\sigma_p} \right) \right| + \left| \log_{10} \left(\frac{\sigma_d}{\hat{\sigma}_d} \right) \right| \right] \tag{4.71}$$

$$= \frac{\log 10(e)}{2} E_j(\delta^2) \left(\left| \left(\frac{\sigma_p^2 - \hat{\sigma_p}^2}{\sigma_p^2 \hat{\sigma}_p^2} \right) \right| + \left| \left(\frac{\hat{\sigma}_d^2 - {\sigma_d}^2}{\sigma_d^2 \hat{\sigma_d}^2} \right) \right| \right) + \left| \log_{10} \left(\frac{\hat{\sigma}_p}{\sigma_p} \right) \right| + \left| \log_{10} \left(\frac{\sigma_d}{\hat{\sigma}_d} \right) \right| \tag{4.72}$$

$$= \frac{\sigma_j^2 \log 10(e)}{2} \left(\left| \left(\frac{\sigma_p^2 - \hat{\sigma_p}^2}{\sigma_p^2 \hat{\sigma}_p^2} \right) \right| + \left| \left(\frac{\hat{\sigma}_d^2 - {\sigma_d}^2}{\sigma_d^2 \hat{\sigma_d}^2} \right) \right| \right) + \left| \log_{10} \left(\frac{\hat{\sigma}_p}{\sigma_p} \right) \right| + \left| \log_{10} \left(\frac{\sigma_d}{\hat{\sigma}_d} \right) \right| \tag{4.73}$$

The second to last inequality comes from the definition of expectation. And the fact that we are considering the estimated parameters as fixed. The last equality comes from the variance $(V(\delta) = \sigma_j^2(1 - \frac{2}{\pi}))$ and the square of expected value $(E(\delta)^2 = \sigma_j^2 \frac{2}{\pi})$,hence $E(\delta^2) = V(\delta) + E(\delta)^2 = \sigma_j^2$.

Given an estimated parameter value, the expected discrepancy will depend on the proposition considered via the σ_j^2 in Equation 4.73. Given that $\sigma_d = \sqrt{2(\sigma_b^2 + \sigma_w)}$ and $\sigma_p = \sqrt{2\sigma_w}$, the expected discrepancy will be larger for the defense than for the prosecutor.

Similar arguments can be used to derive Equation 4.32 in close form. Results are skipped for brevity.

4.7 Simulation study

To evaluate the effect of model misspecification, selecting the incorrect parametric family and not accounting for the dependence structure, during the estimation step to develop a score likelihood ratio system, we propose a simulation study based on the illustration presented in Section 4.6. Two scenarios are considered to identify potential channels that affect inference.

In both scenarios, we run 500 iterations, creating a sample of δ's, and compared results in terms of expected discrepancy and empirical performance metrics. When the half-normal is

selected for estimation, the maximum likelihood estimates are compared to the true model parameters based on their bias, variance, MSE, and MAPE.

For the first simulation study (Section 4.7.1), we created a scenario that would result in the number of matches roughly held constant at 500 cases. For each iteration of our simulation, 1500 sources with 10 items each are generated, but the available data is restricted to generate varying degrees of dependence. Table 4.1 presents the overall setup. Let I denote the number of sources used and J the number of items available. For the first three cases: $I11 - J10$, $I24 - J7$, and $I50 - J5$, all pairs are generated, resulting in known match and non-match pairs. The first case is the one with more induced dependence, as more items are available from fewer sources. The latter provides a smaller dependence structure. After all pairs are created, downsamplig is applied to balance the classes. We employ Algortithm 8 on the simulated data to generate a scenario we denote as $I1500 - J1$. This scenario emulates more closely a theoretically correct set of scores.

Table 4.1 Restriction imposed over the sample

	Sources (I)	Items (J)	KM	KNM*
All pairs	11	10	495	5500
	24	7	504	13524
	50	5	495	30625
SSR	1500	10	500	500

Note: KM denotes known matches, and KNM* denotes the number of cases before applying downsampling to balance cases.
SSR denotes strong source resampling

Our second simulation (Section 4.7.2) does not impose restrictions over the data, it considers a fixed number of sources ($I = 150, 300, 600$) and items within the source J ($J = 10, 20, 50$). For each iteration, a sample is drawn and comparisons are created using either the traditional approach, strong or weak source resampling.

One drawback of applying a resampling step is reducing the sample size available for estimation. Consider the case of $I = 150$ and $J = 10$, the traditional method would result in $\binom{10}{2} \times 150 = 111750$ known matches, while strong source resampling would imply only 50 known matches. To account for this drawback, when resampling methods are used, they are applied

$M = 20$ time to generate base learners whose output is aggregated into a final value of evidence [21].

This is the most realistic scenario from a researcher's perspective since when faced with a given sample, their decision is what is the appropiate estimation method and if a resampling step should be applied.

In both scenarios, data was simulated following a univariate hierarchical Gaussian process (Eq 4.37) with $\sigma_w = 1$ and $\sigma_b = \sqrt{2}$, and comparisons are generated using the L_1 norm. An additional validation set was generated for each scenario by drawing 1000 observations from each half-normal target density. These observations are fixed across iterations, allowing us to compute the variability of the output obtained as a measurement of consensus. In criminal justice, score likelihood ratios are not only expected to provide strong evidence in the correct direction, but ideally, they should be less sensible to changes in the training data.

4.7.1 Simulation 1 results-Fixed sample size

Our first simulation scenario provides results when the estimation sample size is fixed and the degree of dependence changes. We first present the results when the correct estimation method (half-normal family) is chosen and examine the maximum likelihood estimates in terms of Bias, Variance, MSE, and MAPE in Table 4.2.

For scenarios associated with larger dependence,i.e., more items per source, estimates present larger MSE, the major contribution being an increase in variance. In the case of known (non-)matches, MAPE ranged from 5.79 (12.36) % under the most dependent scenario to 2.54 (2.48) % when a theoretically correct sample is used.

We present the results for the expected discrepancy in Figure 4.4 by estimation method and dependence structure. Each observation represents the outcome of one of the 500 simulations. Scenarios closer to the theoretically correct sample are associated with smaller expected discrepancies for both the prosecutor and defense. Further, the expected discrepancy is smaller

Table 4.2 Parameter estimates half-normal distribution. Simulation 1

Match	Scenario	Bias	Variance	MSE	MAPE
KM	I11-J10	0.0820	0.0107	0.0107	5.79
	I24-J7	0.0659	0.0067	0.0066	4.66
	I50-J5	0.0547	0.0049	0.0049	3.87
	I1500-J1	0.0359	0.0020	0.0020	2.54
KNM	I11-J10	0.3028	0.1399	0.1410	12.36
	I24-J7	0.2067	0.0684	0.0683	8.44
	I50-J5	0.1606	0.0392	0.0391	6.55
	I1500-J1	0.0606	0.0058	0.0058	2.48

when the correct parametric family is chosen. This suggests that the dependence structure and the selected estimation method can affect the inference drawn.

Additional empirical performance metrics are computed for each simulation using the validation set. Results are presented in Figure 4.5.

The cost likelihood functions (C_{llr}) present the system's overall performance, a smaller value associated with better performance. Choosing the correct parametric family is associated with smaller costs, and within a chosen estimation method, the median cost is smaller for the theoretically correct sample. In the case of the rate of misleading evidence, we observed that for more dependent samples, the variability of the metric increases. More dependent scenarios are also associated with more variability in systems output. Choosing the correct estimation method results in a closer empirical rate of misleading to the theoretical probability of misleading evidence; however, this is not always associated with the smaller rate of misleading evidence. Choosing an alternative estimation method may result in smaller errors for one of the propositions.

While this simulation illustrates that the estimation method and dependence structure affect the estimation and inference, this exercise compares different starting sample sizes of sources and items that result in a similar number of learning instances. In practical applications, the researcher will start with a fixed sample size and decide between estimation and resampling methods. Our second simulation aim to provide additional information in this regard.

Figure 4.4 Expected discrepancy by proposition. Simulation 1

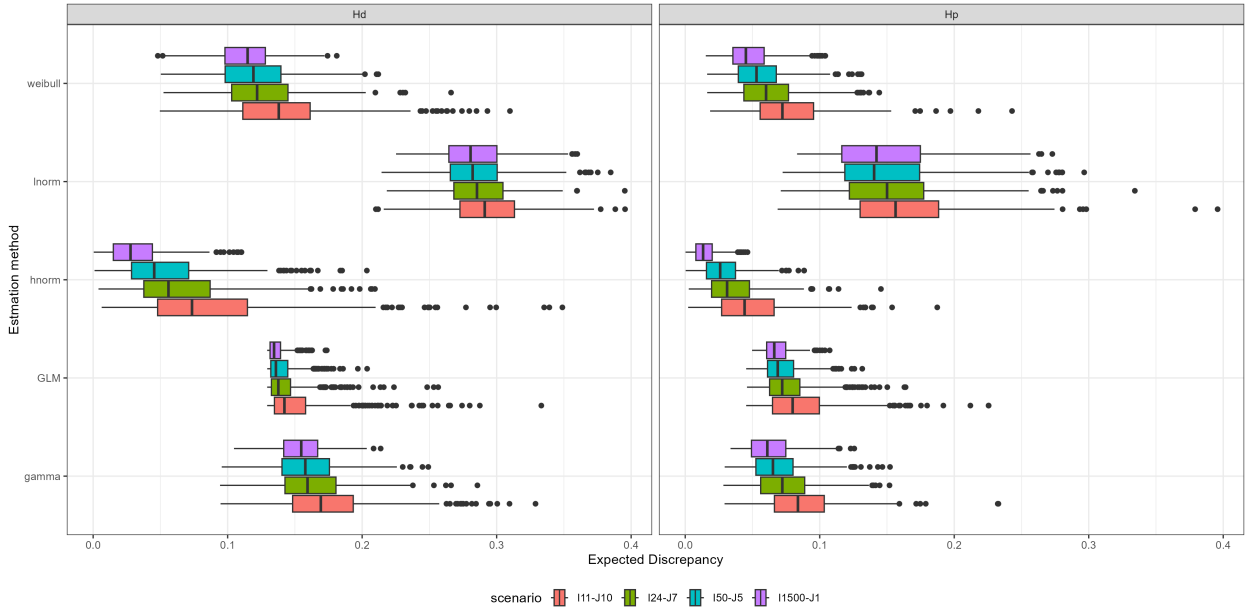
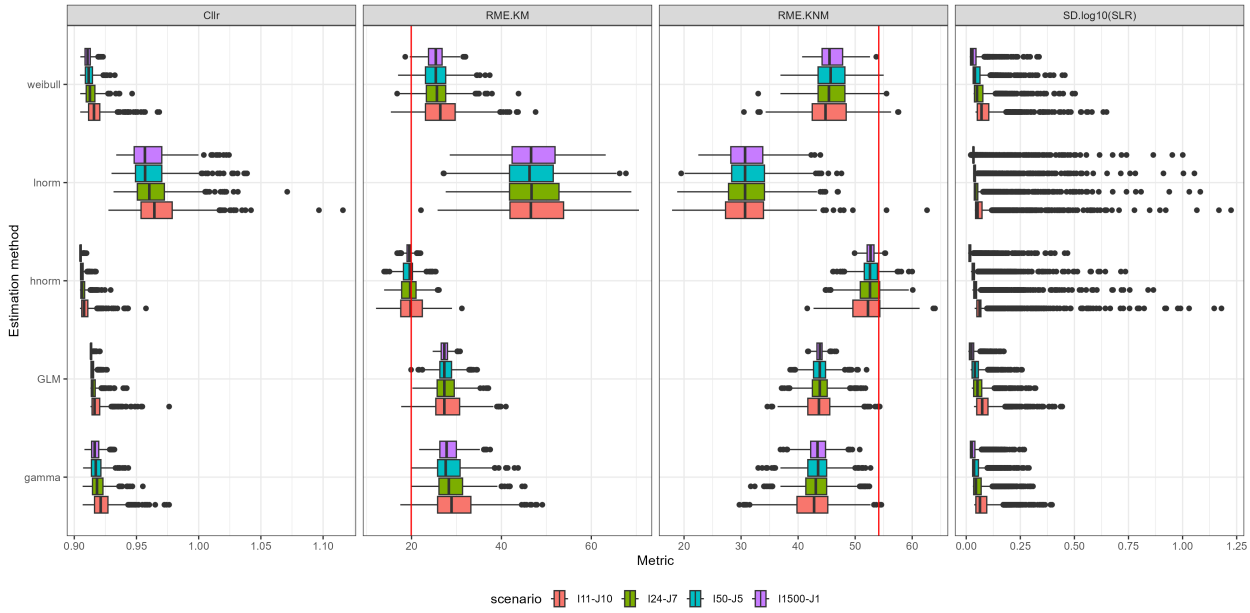


Figure 4.5 Empirical Performance. Simulation 1



Note: *Cllr* denotes the log-likelihood ratio cost, RME the rate of misleading evidence for known matches (KM) and not known matches (KNM). SD denotes the standard deviation. Red vertical lines present the theoretical probability of observing misleading evidence.

4.7.2 Simulation 2 results - Varying the number of sources and items within

Our second simulation presents a more realistic scenario, where the different resampling plans and estimation methods are compared over a fixed number of sources (I) and items within the source (J).

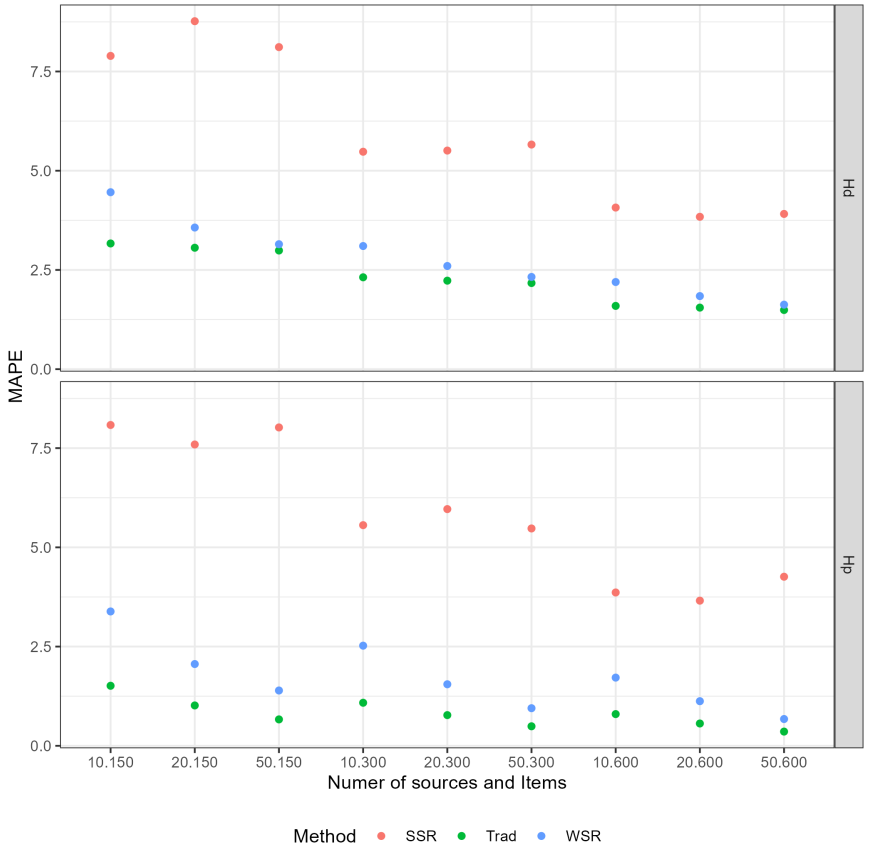
As in our previous simulation, we first present the results for the scaling parameter of the half-normal density comparing the traditional approach, weak and strong resampling. We provide the full results in Table 4.3 in the Appendix and focus on the MAPE (Figure 4.6) for the discussion.

Across the different sample compositions, the traditional approach resulted in better estimates, while strong source resampling produced the worst estimates. This result was expected as the limited amount of sources acts as a strong limit for strong source resampling, and increasing the number of items per source does not increase the number of cases available, resulting in a smaller sample size available for estimation compared to the other methods. Weak resampling performed better, close to the traditional approach. We also observe that point estimates improve faster for weak source resampling than the traditional approach as the diversity of items per source increases.

These results suggest that while dependence may affect the estimation result, as seen in Section 4.7.1), there is a trade-off when applying a resampling step. Resampling plans alone may not result in better estimates; more stringent restrictions can result in smaller sample sizes that outweigh the independence generated.

Given the previous results, when we extend our analysis of expected discrepancy and empirical performance metrics, rather than compare estimated Score Likelihood Ratio systems obtained by applying only one resampling step, we followed [21] to create an ensemble system that has been shown to perform better. We created 20 base systems for each iteration and aggregated their output by taking a simple average.

Figure 4.6 MAPE for a half-normal distribution by number of sources, items and methods. Simulation 2



Note: SSR denotes Strong Source Resampling, WSR Weak Source Resampling, and Trad denotes the traditional approach.

The first number on the x-axis denotes the number of sources I, and the second the number of items J. H_p denotes the prosecutor proposition, H_d the defense.

The average expected discrepancy for the different propositions, estimation methods, and resampling plans are presented in Table 4.4 in the Appendix. We focus our discussion on two estimation methods: half-normal and Weibull (Figure 4.7).

When the half-normal density is chosen, the weak source resampling ensembled system performs similarly to the traditional approach. Strong source resampling ensembles presented larger expected discrepancies for the prosecutor and defense and more variability. The difference between the sampling methods and the overall variability of the metric is reduced as the number of sources is increased.

When an alternative method is chosen, the expected discrepancies are larger (Table 4.4). Showing the importance of selecting the correct density.

The Weibull density presented the smallest discrepancy within the alternative options. For smaller samples with more dependency, strong source resampling ensembled systems resulted in a (larger) smaller expected discrepancy for the (prosecutor) defense. Similar patterns were observed for the other estimation methods, but the gains were smaller.

Lastly, we present some empirical performance metrics for score-likelihood ratio systems. Table 4.5 in the Appendix presents the average metrics by estimation and resampling method, while figure 4.8 focuses on the results obtained for the half-normal and Weibull densities.

Half-normal densities are associated with smaller costs, a smaller rate of misleading evidence for the prosecutor, and a larger rate of misleading evidence for the defense compared to other estimation methods. While weak source resampling performed similarly to traditional methods, strong source resampling ensembles under the half normal density were associated with larger costs and larger rates of misleading evidence for the prosecutor; however, it was associated with a smaller rate of misleading evidence for the defense.

When resampling methods are used for alternative parametric estimation methods, results are mixed. While there is more variability in the metrics for all methods when strong source resampling ensembles are used, for the Weibull and Gamma, smaller average costs are observed.

In the case of the Weibull, there is no effect of the resampling methods on the median rate of misleading evidence. As in Simulation 1, the rate of misleading evidence for defense is smaller for all numbers of sources, items, and resampling plans compared to the half-normal.

Across all estimation methods, resampling plans were associated with more variability as measured by the standard deviation of the log10 output of the systems.

These simulation results present mixed results, suggesting that some methods can benefit more from resampling; the gains seem greater when a small sample and larger dependence are present. As before, we would like to highlight that an alternative estimation method may result in a smaller empirical error rate for the defense than the theoretical probability of observing misleading evidence.

4.8 Conclusions

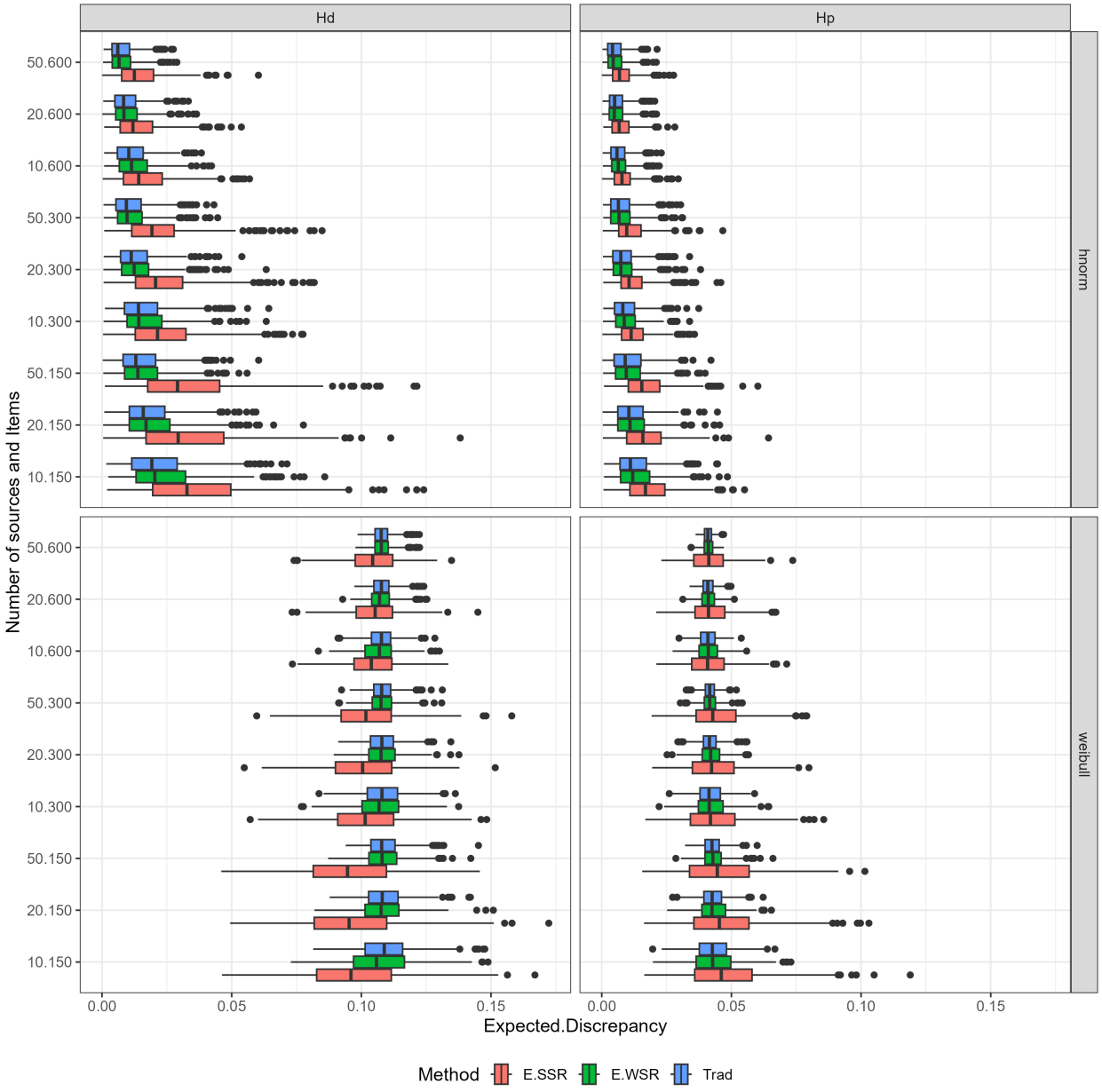
Score likelihood ratio-based inference is becoming more prevalent in source inference problems. This is the case in forensic evidence where machine learning-derived scores are used for complex pattern evidence [6].

Our work examines the effect of model misspecification: selecting an incorrect density and not accounting for the dependence structure, on score-based likelihood ratio inference. We continue previous work on resampling plans to alleviate dependence structure [21] by introducing Weak Source Resampling. This approach imposes that items are used only once, generating more learning instances than Strong Source Resampling.

We introduced discrepancy metrics that can be used to study the effect of model misspecification and illustrate our results in a simplified scenario where only conditional density estimation is required. This univariate example is the basis of our simulations, where alternative estimation methods and resampling plans are compared.

In both our simulations, selecting the incorrect parametric estimation method was associated with a larger expected discrepancy. However, alternative estimation methods may produce better

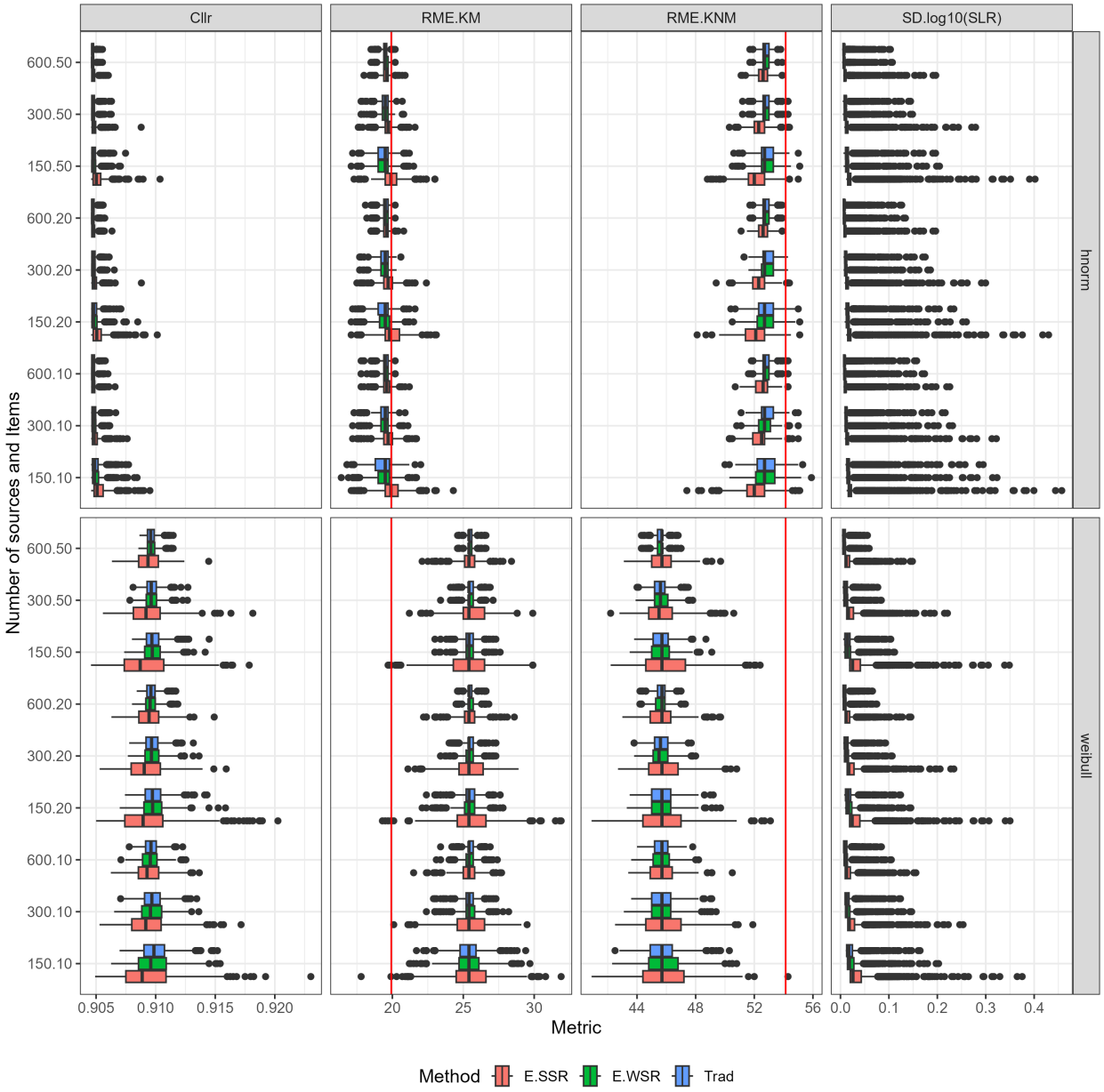
Figure 4.7 Expected discrepancy by proposition, select densities. Simulation 2



Note: E.SSR denotes Strong Source Resampling ensemble, E.WSR Weak Source Resampling ensemble, and Trad denotes the traditional approach.

The first number on the y-axis denotes the number of sources I, and the second the number of items J. H_p denotes the prosecutor proposition, H_d the defense.

Figure 4.8 Performance metric by proposition, select densities. Simulation 2



Note: E.SSR denotes Strong Source Resampling ensemble, E.WSR Weak Source Resampling ensemble, and Trad denotes the traditional approach.

The first number on the y-axis denotes the number of sources I, the second the number of items J. *Cllr* denotes the log-likelihood ratio cost, RME the rate of misleading evidence for known matches (KM) and not known matches (KNM). SD denotes the standard deviation.

empirical performance metrics. We observe that the associated rate of misleading evidence may be smaller than the probability of observing misleading evidence for our scenario.

Results for the effect of dependence structure were mixed. While our first simulation shows that there is a detrimental effect of using nonindependent data to estimate score likelihood ratio functions across the different estimation methods, our second simulation suggests that there could be a tradeoff between thinning out dependence to obtain theoretically correct learning instances and sample size available for estimation.

Weak Source Resampling performed comparably to the traditional approach, while Strong Source Resampling showed mixed results depending on the estimation methods used. This suggests that some methods can benefit from a resampling step, but further research is needed to establish the conditions that guarantee improvements.

Further, our work examined simple distances and the dependence effect on the estimation step. Analogous scenarios are needed to examine the effect on more complex scores that also require a training stage.

4.9 References

[1] Aitken, C. G. G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley and Sons, Ltd., West Sussex, UK, 2nd edition.

[2] Berger, R. and Casella, G. (2001). *Statistical Inference*. Duxbury Press, Florence, AL, 2 edition.

[3] Bolck, A., Ni, H., and Lopatka, M. (2015). Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk*, 14(3):246–266.

[4] Bolck, A., Weyermann, C., Dujourdy, L., Esseiva, P., and van den Berg, J. (2009). Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International*, 191(1):42 – 51.

[5] Brümmer, N. and Du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275.

[6] Carriquiry, A., Hofmann, H., Tai, X. H., and VanderPlas, S. (2019). Machine learning in forensic applications. *Significance*, 16(2):29–35.

[7] Evett, I., Jackson, G., Lambert, J., and McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, 40(4):233–239.

[8] Garton, N., Ommen, D., Niemi, J., and Carriquiry, A. (2020). Score-based likelihood ratios to evaluate forensic pattern evidence. *arXiv preprint arXiv:2002.09470*.

[9] Hepler, A. B., Saunders, C. P., Davis, L. J., and Buscaglia, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic science international*, 219(1-3):129–140.

[10] Leegwater, A. J., Vergeer, P., Alberink, I., van der Ham, L. V., van de Wetering, J., El Harchaoui, R., Bosma, W., Ypma, R. J., and Sjerps, M. J. (2024). From data to a validated score-based lr system: A practitioner's guide. *Forensic Science International*, 357:111994.

[11] Meuwly, D., Ramos, D., and Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic science international*, 276:142–153.

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

[13] Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2):173–197.

[14] Neumann, C. and Ausdemore, M. (2020). Defence against the modern arts: the curse of statistics—Part II: 'Score-based likelihood ratios'. *Law, Probability and Risk*, 19(1):21–42.

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

[16] Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the american statistical association*, 95(451):760–768.

[17] Stern, H. S. (2017). Statistical issues in forensic science. *Annual Review of Statistics and Its Application*, 4:225–244.

[18] Sugiyama, M., Suzuki, T., and Kanamori, T. (2010). Density ratio estimation: A comprehensive review (statistical experiment and its related topics). *RIMS Kokyuroku*, 1703:10–31.

[19] van Lierop, S., Ramos, D., Sjerps, M., and Ypma, R. (2024). An overview of log likelihood ratio cost in forensic science–where is it used and what values can we expect? *Forensic science international: synergy*, 8:100466.

[20] Veneri, F. and Ommen, D. (2021). An evaluation of score-based likelihood ratios for glass data. Master's thesis, Iowa State University.

[21] Veneri, F. and Ommen, D. M. (2023). Ensemble learning for score likelihood ratios under the common source problem. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 16(6):528–546.

[22] Willis, S., Aitken, C., Barrett, A., Berger, C., Biedermann, A., Champod, C., Hicks, T., Lucena-Molina, J., Lunt, L., McDermott, S., McKenna, L., Nordgaard, A., O'Donnell, G., Rasmusson, B., Sjerps, M., Taroni, F., and Zadora, G. (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science*. European Network of Forensic Science Institutes, http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

4.10 Appendix: Additional simulation results

Table 4.3 Parameter estimation for a half-normal distribution by number of sources, items, and methods. Simulation 2

			Bias		Variance		MSE		MAPE	
I	J	Method	KM	KNM	KM	KNM	KM	KNM	KM	KNM
150	10	SSR	0.1143	0.1933	0.0200	0.0595	0.0201	0.0597	8.08	7.89
		WSR	0.0479	0.1092	0.0037	0.0189	0.0036	0.0189	3.38	4.46
		Trad	0.0214	0.0776	0.0007	0.0098	0.0007	0.0098	1.51	3.17
	20	SSR	0.1074	0.2147	0.0181	0.0710	0.0182	0.0709	7.59	8.77
		WSR	0.0291	0.0874	0.0013	0.0122	0.0013	0.0123	2.06	3.57
		Trad	0.0144	0.0749	0.0003	0.0089	0.0003	0.0089	1.02	3.06
	50	SSR	0.1134	0.1988	0.0203	0.0651	0.0206	0.0651	8.02	8.11
		WSR	0.0197	0.0771	0.0006	0.0093	0.0006	0.0094	1.39	3.15
		Trad	0.0094	0.0732	0.0001	0.0083	0.0001	0.0084	0.67	2.99
300	10	SSR	0.0786	0.1342	0.0100	0.0281	0.0100	0.0281	5.56	5.48
		WSR	0.0357	0.0760	0.0019	0.0093	0.0019	0.0093	2.52	3.10
		Trad	0.0153	0.0567	0.0004	0.0052	0.0004	0.0052	1.08	2.31
	20	SSR	0.0843	0.1349	0.0112	0.0288	0.0113	0.0288	5.96	5.51
		WSR	0.0219	0.0637	0.0008	0.0065	0.0008	0.0065	1.55	2.60
		Trad	0.0109	0.0546	0.0002	0.0048	0.0002	0.0048	0.77	2.23
	50	SSR	0.0774	0.1386	0.0092	0.0297	0.0092	0.0296	5.48	5.66
		WSR	0.0134	0.0569	0.0003	0.0052	0.0003	0.0052	0.95	2.32
		Trad	0.0070	0.0532	0.0001	0.0045	0.0001	0.0045	0.49	2.17
600	10	SSR	0.0546	0.0997	0.0048	0.0153	0.0048	0.0154	3.86	4.07
		WSR	0.0243	0.0538	0.0009	0.0048	0.0009	0.0048	1.72	2.20
		Trad	0.0113	0.0390	0.0002	0.0024	0.0002	0.0024	0.80	1.59
	20	SSR	0.0517	0.0940	0.0042	0.0139	0.0042	0.0139	3.66	3.84
		WSR	0.0159	0.0451	0.0004	0.0032	0.0004	0.0032	1.12	1.84
		Trad	0.0079	0.0380	0.0001	0.0023	0.0001	0.0023	0.56	1.55
	50	SSR	0.0602	0.0958	0.0056	0.0142	0.0056	0.0142	4.26	3.91
		WSR	0.0095	0.0398	0.0001	0.0026	0.0001	0.0026	0.67	1.63
		Trad	0.0051	0.0365	0.0000	0.0022	0.0000	0.0022	0.36	1.49

Note: SSR denotes strong Source Resampling, WSR Weak Source Resampling, and Trad denotes the traditional approach. I denotes sources and J items within the source.

Table 4.4 Average Expected discrepancy by estimation and resampling method. Simulation 2

			$E_d h(\delta)$			$E_p h(\delta)$		
I	J	Estimation	E.SSR	E.WSR	Trad	E.SSR	E.WSR	Trad
150	10	GLM	0.1342	0.1348	0.1349	0.0733	0.0677	0.0671
		gamma	0.1467	0.1503	0.1515	0.0642	0.0582	0.0580
		hnorm	0.0371	0.0239	0.0220	0.0182	0.0136	0.0128
		lnorm	0.2781	0.2787	0.2802	0.1511	0.1425	0.1428
		weibull	0.0973	0.1070	0.1094	0.0481	0.0430	0.0430
	20	GLM	0.1340	0.1346	0.1345	0.0726	0.0675	0.0672
		gamma	0.1464	0.1507	0.1510	0.0631	0.0580	0.0578
		hnorm	0.0344	0.0196	0.0180	0.0169	0.0122	0.0116
		lnorm	0.2770	0.2798	0.2799	0.1490	0.1433	0.1427
		weibull	0.0970	0.1080	0.1089	0.0473	0.0429	0.0428
	50	GLM	0.1340	0.1344	0.1344	0.0726	0.0673	0.0671
		gamma	0.1449	0.1509	0.1509	0.0621	0.0580	0.0578
		hnorm	0.0334	0.0159	0.0151	0.0169	0.0110	0.0107
		lnorm	0.2752	0.2801	0.2801	0.1467	0.1433	0.1430
		weibull	0.0950	0.1087	0.1088	0.0463	0.0430	0.0427
300	10	GLM	0.1343	0.1344	0.1344	0.0688	0.0670	0.0666
		gamma	0.1484	0.1504	0.1509	0.0591	0.0575	0.0572
		hnorm	0.0243	0.0172	0.0160	0.0124	0.0098	0.0093
		lnorm	0.2781	0.2797	0.2801	0.1447	0.1429	0.1425
		weibull	0.1020	0.1071	0.1084	0.0430	0.0419	0.0417
	20	GLM	0.1339	0.1342	0.1343	0.0693	0.0668	0.0665
		gamma	0.1477	0.1505	0.1508	0.0593	0.0574	0.0571
		hnorm	0.0234	0.0139	0.0132	0.0121	0.0086	0.0084
		lnorm	0.2777	0.2798	0.2800	0.1446	0.1427	0.1425
		weibull	0.1011	0.1080	0.1083	0.0432	0.0420	0.0416
	50	GLM	0.1337	0.1342	0.1342	0.0695	0.0666	0.0665
		gamma	0.1483	0.1505	0.1507	0.0602	0.0571	0.0571
		hnorm	0.0220	0.0114	0.0111	0.0113	0.0078	0.0077
		lnorm	0.2791	0.2798	0.2800	0.1468	0.1423	0.1425
		weibull	0.1017	0.1080	0.1082	0.0441	0.0417	0.0416
600	10	GLM	0.1340	0.1342	0.1343	0.0675	0.0665	0.0662
		gamma	0.1489	0.1502	0.1506	0.0574	0.0568	0.0566
		hnorm	0.0165	0.0127	0.0117	0.0085	0.0071	0.0066
		lnorm	0.2785	0.2796	0.2800	0.1428	0.1423	0.1423
		weibull	0.1041	0.1072	0.1078	0.0414	0.0411	0.0409
	20	GLM	0.1340	0.1342	0.1342	0.0672	0.0662	0.0661
		gamma	0.1496	0.1504	0.1505	0.0578	0.0567	0.0566
		hnorm	0.0143	0.0099	0.0095	0.0077	0.0060	0.0059
		lnorm	0.2797	0.2798	0.2798	0.1444	0.1423	0.1421
		weibull	0.1051	0.1077	0.1080	0.0417	0.0410	0.0409
	50	GLM	0.1340	0.1341	0.1342	0.0673	0.0662	0.0661
		gamma	0.1488	0.1506	0.1505	0.0572	0.0568	0.0566
		hnorm	0.0145	0.0081	0.0077	0.0078	0.0055	0.0053
		lnorm	0.2782	0.2800	0.2799	0.1424	0.1425	0.1422
		weibull	0.1044	0.1080	0.1079	0.0414	0.0411	0.0410

Note: E.SSR denotes Strong Source Resampling ensemble, E.WSR Weak Source Resampling ensemble, and Trad denotes the traditional approach.

I denotes sources and J items within the source. $E_j h(\delta)$ denotes the expected discrepancy under the propositions.

Table 4.5 Average performance metric by estimation and resampling method. Simulation 2

			Cllr			RME.KM			RME.KM			SD.log10(SLR)		
I	J	Estimation	E.SSR	E.WSR	Trad	E.SSR	E.WSR	Trad	E.SSR	E.WSR	Trad	E.SSR	E.WSR	Trad
	10	GLM	0.9138	0.9134	0.9134	27.51	27.31	27.30	43.71	43.82	43.82	0.0294	0.0213	0.0197
		gamma	0.9155	0.9156	0.9158	28.29	27.84	27.84	43.28	43.50	43.46	0.0345	0.0224	0.0195
		hnorm	0.9053	0.9051	0.9050	19.99	19.40	19.31	52.05	52.74	52.84	0.0315	0.0238	0.0221
		lnorm	0.9601	0.9568	0.9569	45.58	46.51	46.96	31.75	31.09	30.78	0.0429	0.0241	0.0194
		weibull	0.9095	0.9098	0.9100	25.41	25.33	25.42	45.88	45.88	45.76	0.0381	0.0248	0.0218
	20	GLM	0.9137	0.9133	0.9133	27.52	27.28	27.28	43.68	43.80	43.80	0.0272	0.0186	0.0170
		gamma	0.9153	0.9156	0.9156	28.25	27.83	27.80	43.29	43.45	43.45	0.0323	0.0182	0.0161
		hnorm	0.9052	0.9049	0.9049	20.01	19.38	19.34	52.05	52.79	52.83	0.0295	0.0201	0.0186
		lnorm	0.9591	0.9569	0.9567	45.53	46.81	46.85	31.81	30.82	30.73	0.0409	0.0182	0.0147
		weibull	0.9094	0.9098	0.9099	25.45	25.43	25.46	45.82	45.72	45.67	0.0356	0.0202	0.0179
	50	GLM	0.9137	0.9133	0.9132	27.50	27.27	27.27	43.69	43.80	43.80	0.0274	0.0162	0.0154
		gamma	0.9150	0.9156	0.9156	28.04	27.81	27.79	43.45	43.42	43.44	0.0331	0.0152	0.0143
		hnorm	0.9052	0.9049	0.9049	20.00	19.38	19.36	52.05	52.81	52.82	0.0283	0.0168	0.0162
		lnorm	0.9582	0.9568	0.9567	45.00	46.87	46.87	32.09	30.70	30.67	0.0427	0.0144	0.0126
		weibull	0.9091	0.9098	0.9098	25.25	25.49	25.48	46.04	45.63	45.63	0.0364	0.0169	0.0158
300	10	GLM	0.9134	0.9132	0.9132	27.38	27.32	27.28	43.73	43.76	43.77	0.0203	0.0152	0.0141
		gamma	0.9154	0.9155	0.9156	28.04	27.86	27.82	43.41	43.42	43.44	0.0241	0.0156	0.0142
		hnorm	0.9050	0.9049	0.9049	19.65	19.44	19.39	52.44	52.73	52.80	0.0222	0.0171	0.0162
		lnorm	0.9575	0.9567	0.9567	46.24	46.86	46.94	31.31	30.81	30.73	0.0295	0.0161	0.0142
		weibull	0.9094	0.9097	0.9098	25.37	25.45	25.46	45.88	45.70	45.68	0.0265	0.0176	0.0158
	20	GLM	0.9133	0.9132	0.9132	27.41	27.29	27.28	43.70	43.76	43.77	0.0196	0.0131	0.0125
		gamma	0.9152	0.9155	0.9155	28.03	27.82	27.77	43.39	43.41	43.43	0.0230	0.0129	0.0119
		hnorm	0.9049	0.9048	0.9048	19.68	19.43	19.41	52.41	52.77	52.79	0.0210	0.0144	0.0138
		lnorm	0.9574	0.9566	0.9566	46.15	46.86	46.88	31.36	30.72	30.67	0.0287	0.0127	0.0109
		weibull	0.9093	0.9097	0.9097	25.40	25.50	25.48	45.82	45.61	45.62	0.0252	0.0143	0.0132
	50	GLM	0.9133	0.9132	0.9131	27.41	27.27	27.27	43.68	43.76	43.76	0.0183	0.0116	0.0112
		gamma	0.9153	0.9155	0.9155	28.15	27.76	27.76	43.29	43.43	43.43	0.0222	0.0111	0.0104
		hnorm	0.9049	0.9048	0.9048	19.70	19.43	19.43	52.38	52.78	52.79	0.0196	0.0122	0.0120
		lnorm	0.9582	0.9564	0.9565	46.49	46.82	46.85	31.14	30.68	30.66	0.0286	0.0102	0.0091
		weibull	0.9094	0.9097	0.9097	25.51	25.49	25.49	45.67	45.61	45.60	0.0241	0.0123	0.0115
600	10	GLM	0.9132	0.9131	0.9131	27.34	27.27	27.26	43.71	43.75	43.76	0.0140	0.0109	0.0101
		gamma	0.9153	0.9154	0.9155	27.89	27.79	27.75	43.41	43.43	43.44	0.0164	0.0115	0.0102
		hnorm	0.9048	0.9048	0.9048	19.57	19.47	19.44	52.58	52.74	52.78	0.0155	0.0125	0.0116
		lnorm	0.9566	0.9565	0.9565	46.59	46.85	46.88	31.03	30.77	30.70	0.0200	0.0119	0.0102
		weibull	0.9094	0.9096	0.9096	25.43	25.47	25.46	45.74	45.66	45.66	0.0180	0.0128	0.0113
	20	GLM	0.9132	0.9131	0.9131	27.29	27.25	27.25	43.75	43.76	43.76	0.0133	0.0091	0.0088
		gamma	0.9154	0.9154	0.9154	27.91	27.74	27.71	43.41	43.43	43.45	0.0157	0.0090	0.0084
		hnorm	0.9048	0.9048	0.9048	19.55	19.46	19.46	52.63	52.77	52.78	0.0138	0.0103	0.0098
		lnorm	0.9573	0.9564	0.9564	46.80	46.84	46.82	30.91	30.68	30.66	0.0198	0.0088	0.0075
		weibull	0.9095	0.9096	0.9096	25.44	25.47	25.47	45.72	45.62	45.61	0.0171	0.0099	0.0094
	50	GLM	0.9132	0.9131	0.9131	27.30	27.24	27.25	43.73	43.76	43.75	0.0132	0.0081	0.0079
		gamma	0.9152	0.9154	0.9154	27.84	27.74	27.72	43.44	43.43	43.44	0.0154	0.0077	0.0073
		hnorm	0.9048	0.9047	0.9047	19.57	19.47	19.47	52.60	52.77	52.77	0.0139	0.0087	0.0084
		lnorm	0.9564	0.9565	0.9564	46.45	46.85	46.82	31.08	30.66	30.67	0.0192	0.0072	0.0063
		weibull	0.9094	0.9096	0.9096	25.41	25.48	25.46	45.74	45.58	45.59	0.0170	0.0085	0.0081

Note: E.SSR denotes Strong Source Resampling ensemble, E.WSR Weak Source Resampling ensemble, and Trad denotes the traditional approach. I denotes sources and J items within the source. $Cllr$ denotes the log-likelihood ratio cost, RME the rate of misleading evidence for known matches (KM) and not known matches (KNM). SD denotes the standard deviation.

CHAPTER 5. GENERAL CONCLUSION

5.1 Conclusion

This dissertation focuses on source attribution problems and the use of Score based Likelihood ratio to draw inferences for these problems. Score Likelihood ratios have been proposed as an alternative to traditional methods for their ability to address open set problems and accommodate scores derived from modern machine learning methods required for the complex features found in forensic science [1, 4]

While there have been recent advances in the use of score likelihood ratio in this domain there are still open challenges and critics to adopt their use fully.

Scores act as a dimensionality reduction technique [11, 3], implying a loss of information. Further, scores consider the similarity between items, not typically [7]. Ideally, both should be considered to provide a correct value of evidential strength. Other lines of research have focused on how appropriate they are within a Bayesian framework [10, 8], how closely they approximate Likelihood ratio systems [2, 8], and to what extent can specific and common source system be used interchangeably [12, 9].

From a practitioner's perspective, other authors have centered their attention on the validation of the systems developed; see, for example, [6, 5].

We focus on a less studied aspect, system misspecification, particularly the effect of not considering the dependence structure created when generating learning instances.

While there are proposals on how the scores should be generated following the inference problem [9, 8], for the specific source, the lack of available data has led research to use common source systems, leading to incorrect conclusions. Further, practical applications for the common source have used all potential pairwise comparisons to create learning instances for the common source problem, generating a complex dependence structure.

We propose resampling plans for the common (Chapter 2 and 4) specific source (Chapter 3) to address these issues. These resampling plans can be used as the building block to create weak learners that can be aggregated to obtain better-performing systems (Chapter 2).

While our simulation and application for these chapters suggest that using resampling and ensembling can strengthen inference, there are still some avenues of research to explore.

Chapter 4, introduced divergence metrics used to study the effect of model misspecification on score likelihood ratio-based inference. Our univariate illustration for a simple score suggests that while there is an effect associated with unaccounted dependence, there may be a trade-off when applying a too-strict resampling plan for some density estimation procedures. Further research is needed to delimit the condition under which a gain is expected.

Similarly, our work proposes ensembling M-weak learners using different aggregation methods (Chapter2). The number of weak learners and how results are combined could be further studied to boost the system's performance.

Similarly, for Chapter 3 the number of neighbors to create synthetic items is an input from the user. We want to extend recommendations associated with the data available and how this can impact performance. Further, other extrapolation methods could be explored.

Lastly, current score likelihood ratio inference is used to compute the evidential value of the evidence observed. We believe that our resampling methods could also be used to quantify the uncertainty associated with the estimates provided to judges and jurors.

5.2 References

[1] Carriquiry, A., Hofmann, H., Tai, X. H., and VanderPlas, S. (2019). Machine learning in forensic applications. *Significance*, 16(2):29–35.

[2] Garton, N., Ommen, D., Niemi, J., and Carriquiry, A. (2020). Score-based likelihood ratios to evaluate forensic pattern evidence. *arXiv preprint arXiv:2002.09470*.

[3] Ishihara, S. and Carne, M. (2022). Likelihood ratio estimation for authorship text evidence: An empirical comparison of score-and feature-based methods. *Forensic Science International*, 334:111268.

[4] Kafadar, K. and Carriquiry, A. L. (2024). Challenges in modeling, interpreting, and drawing conclusions from images as forensic evidence. *Statistics and Data Science in Imaging*, 1(1):2401758.

[5] Leegwater, A. J., Vergeer, P., Alberink, I., van der Ham, L. V., van de Wetering, J., El Harchaoui, R., Bosma, W., Ypma, R. J., and Sjerps, M. J. (2024). From data to a validated score-based lr system: A practitioner's guide. *Forensic Science International*, 357:111994.

[6] Meuwly, D., Ramos, D., and Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic science international*, 276:142–153.

[7] Morrison, G. S. and Enzinger, E. (2018). Score based procedures for the calculation of forensic likelihood ratios–scores should take account of both similarity and typicality. *Science & Justice*, 58(1):47–58.

[8] Neumann, C. and Ausdemore, M. (2020). Defence against the modern arts: the curse of statistics—Part II: 'Score-based likelihood ratios'. *Law, Probability and Risk*, 19(1):21–42.

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

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

[11] Stern, H. S. (2017). Statistical issues in forensic science. *Annual Review of Statistics and Its Application*, 4:225–244.

[12] Vergeer, P. (2023). From specific-source feature-based to common-source score-based likelihood-ratio systems: ranking the stars. *Law, Probability and Risk*, page mgad005.