Ensemble of Score Likelihood Ratios under the common source problem

Federico Veneri and Danica Ommen

Iowa State University

February, 2023

csafe

Center for Statistics and Applications in Forensic Evidence

Acknowledgement

This work was funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

Agenda for today

- The source problem in forensics comparison.
- Developing SLR: the issue with using pairwise comparisons.
- A remedy through sampling.
- Strengthening SLR system using an ensemble approach.

The Common Source Problem

- 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 given the observed evidence.
- In forensics we usually contrast two propositions
  - H_p : Item 1 and Item 2 come from the same source.
  - H_d : Item 1 and Item 2 come from two different sources.
- Our conclusions will be based on measurements (features) taken from each item.
  - u_x : Measurement(s) from Item 1
  - u_y : Measurement(s) from Item 2

Score Likelihood Ratios

- We can use score-based likelihood ratios are an alternative way to measure the evidence's weight.

$$SLR(\delta) = \frac{g(\delta|H_p)}{g(\delta|H_d)}$$

  - δ : A comparison computed between u_x and u_y
  - g : conditional probability density function.

- Interpretation:
  - Top: the probability of the score under H_p
  - Bottom: the probability of the score under H_d
- Hence:
  - $SLR(\delta) > 1$ supports H_p (Score more likely under H_p than H_d)
  - $SLR(\delta) < 1$ supports H_d (Score more likely under H_d than H_p)

Developing an SLR

- Developing a Score Likelihood Ratio consists of two steps:
  - Constructing a comparison metric: $\delta(.,.)$.
  - Estimating the distribution of the score under both propositions $g(\delta \mid .)$ or a density ratio estimator $r(\delta)$

- **Popular methods used for creating the comparison metric and density estimation required an independence assumption that is not met in forensic comparison.**

The Background Population

- It is often assumed that a set of background population samples (A) is available to construct the SLR system

$$A = \{A_{ij} : i^{th} \text{ Source}, j^{th} \text{ Item}\}$$

- Researchers create all possible pairwise combinations from A.
- Categorize them into known matches (KM) when two items originate from the same source or known non-matches (KNM) when two items originate from different sources.

The issue with the pairwise comparison.

- We have an imbalanced sample. Known Non-Matches are going to outnumber Known Matches.
  - For a set A with 10 sources (N_s) and 3 items each (N_i).

$$N_{KM} = N_s \binom{N_i}{2} \qquad N_{KNM} = \binom{N_s N_i}{2} - N_{KM}$$

    There are $N_{KM} = 30$ pairs and $N_{KNM} = 405$ pairs
  - We can accommodate this, for example, by downsampling the KNM.
- However, we aren't acknowledging how the data was generated (Sources $\rightarrow$ Items). We are learning from dependent data, assuming they are independent

The Dependence Problem

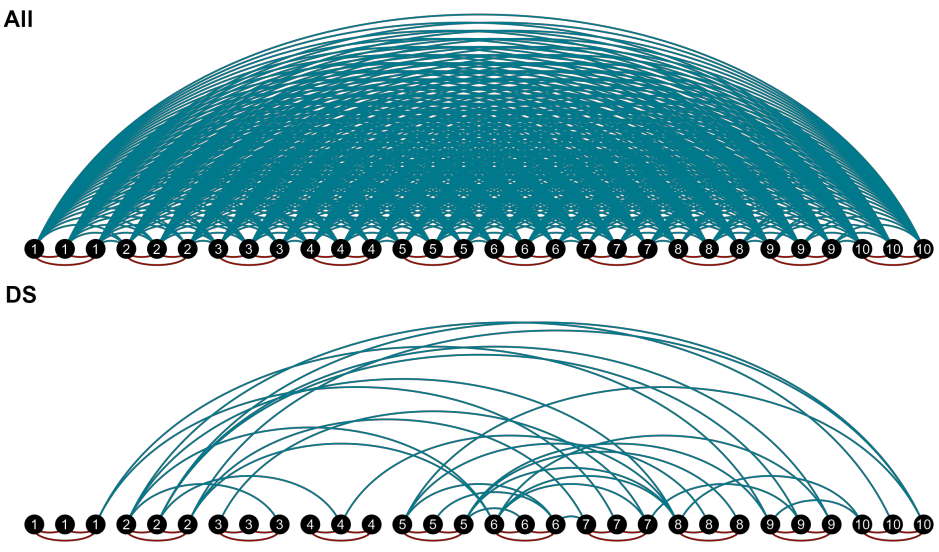


Figure 1: Dependence structure

At source level:

- Multiple KM comparisons use the same source.
- Multiple KNM comparisons use the same source.

At item level:

- Items are used multiple times.

Methodology: A sampling remedy

- To solve this issue, we introduce a sampling step to make sure that the training/estimation sets have independent observations
- **SSSA**: Strong Source Sampling Algorithm

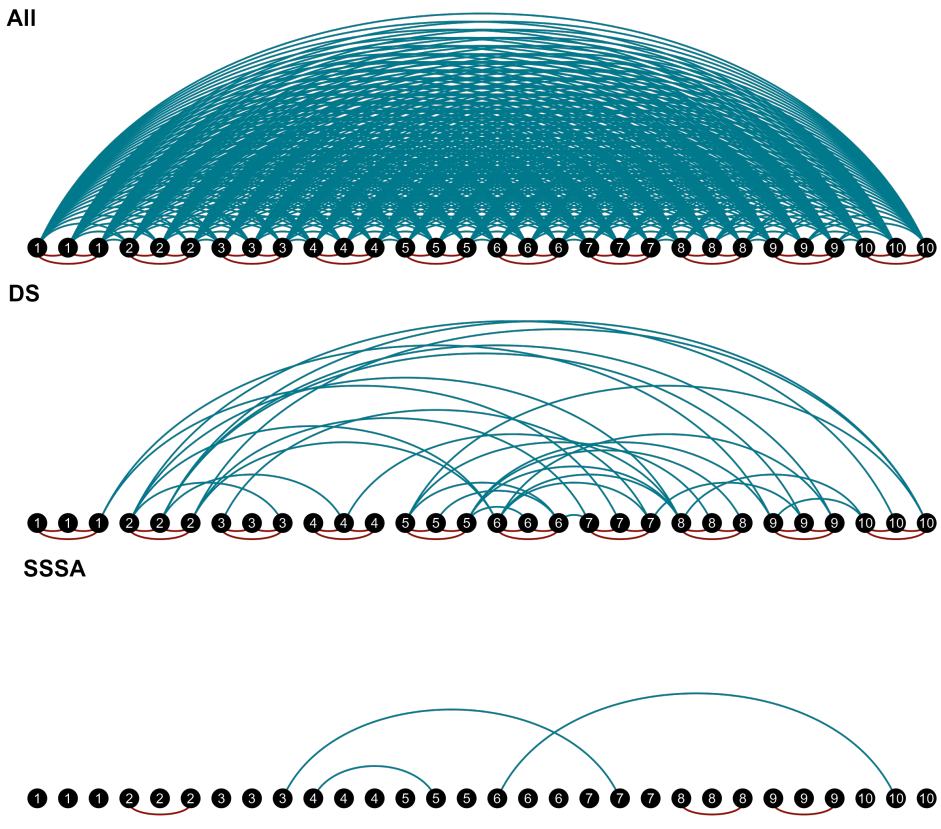


Figure 2: Dependence structure

Methodology: Creating multiple experts

- To make the most out of our data, we propose an ensembling learning approach for SLRs
- We build multiple Base SLRs (BSLR) over a set where assumptions are met.
- Expert's opinions (BSLR) are aggregated into a final SLR score.

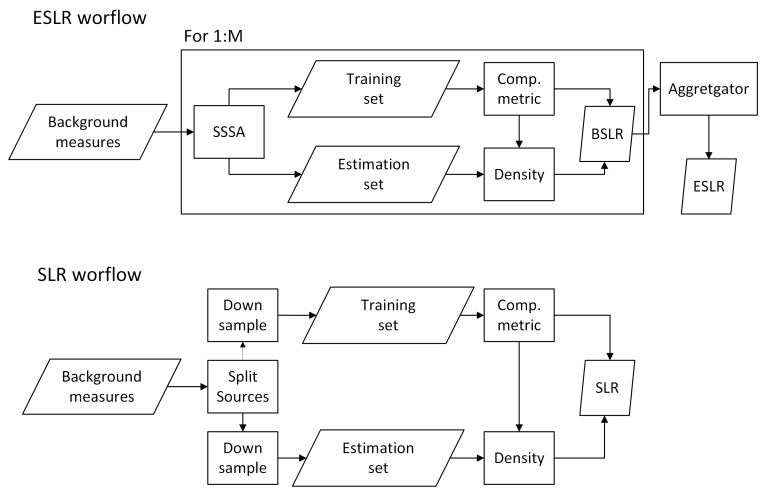
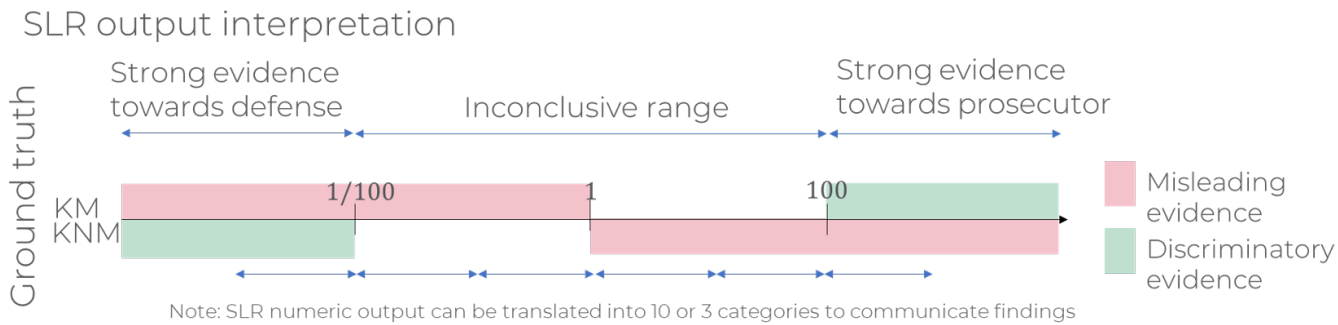


Figure 3: Workflow structure

Testing Our Approach

- Our Data and methods:
  - CSAFE London Letter as background measurements.
  - CVL (Computer Vision Lab) set for validation.
  - Random forest-based comparison metric.
  - Logit-based density ratio estimator.
  - Naive Aggregators: Mean, median, majority voting.
- Our evaluation metrics:
  - Rate of misleading evidence.
  - Discriminatory power.



  - Consensus -over verbal sets- and distance in value of evidence.
  - *Cllr*, as a forensic cost function

Testing Our Approach

- We repeated 500 times the following experiments:
  - Generate 50 BSLR following our ESLR workflow.
    - SSSA+ ensemble step.
  - Generate a traditional SLR following the traditional workflow.
    - Splitting sources+ Down sampling step.
  - Generate a validation set of 1000 known matches and 1000 known non-matches from the CVL dataset.
  - Computed performance metrics to evaluate traditional SLR and ESLRs (Exp. 1); and held the same validation set across repetition to evaluate consensus (Exp. 2).

Performance Metric - 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	

- Main takeaways:
  - ESLR resulted in less misleading evidence for KM (2%) at the cost of a small increase of misleading evidence for KNM (.5%).
  - More discriminatory evidence for KM and KNM.
  - Overall reduction in the cost incurred.

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	
(log10 scale)	Median	0.0147	0.0035	0.0035	
	Sd	0.0104	0.0008	0.0008	

- Main takeaways:
  - Ensemble SLR resulted in similar conclusions for the same holdout evidence.
  - Higher consensus: ESLR conclusions tend to fall in the same category.
  - Smaller distance: ESLR showed smaller distances in terms of probative value for the same hold-out pair.

Summary

- Our results show that ESLRs (SSSA+Ensembling):
  - Reduced the rate of misleading evidence for KM at the cost of a small increase in misleading evidence for KNM.
  - Enhanced the discriminatory power, as traditional SLR tends to produce more inconclusive results
  - Can reduce the overall error, as measured by the cost function.
  - Produced more consistent conclusion, more similar probative value.
- Our approach is not limited to handwriting or forensics; it can be used whenever SLRs are built over pairwise comparisons.

Thank you

Questions? Email us at
fveneri@iastate.edu
dmommen@iastate.edu