Ensemble of SLR systems for forensic evidence

Federico Veneri, MS ; Danica M. Ommen, Phd.

Iowa State University – Department of Statistics

Iowa State University – Center for Statistics and Applications in Forensic Evidence

1. Objectives.

- Score likelihood ratios (SLR) are an alternative way to provide a numerical assessment of evidential strength when contrasting two propositions.
- The SLR approach focuses on a lower-dimensional (dis)similarity metric and avoids distributional assumptions regarding the features [1].
- SLR can be developed in two steps:
  1. Constructing comparison metric $C(.)$.
  2. Estimating the distribution of the metric under both propositions (or the density ratio).
- Popular methods used in these steps assume independence between observations. However, the independence assumption is not met and is often overlooked.
- We introduce an ensemble approach to remedy this lack of independence and improve SLR performance.
- To illustrate our approach, we use handwriting data [2,3] under the common source problem [4].
- We use a Random Forest (RF) based score as a comparison metric and a logistic classifier to estimate the density ratio.

2. Forensic propositions and assumptions in SLR.

- Consider the common source problem with two questioned documents (QD) E_x, E_y.
- Forensics science considers two propositions:

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.

Traditional and CSAFE approaches:

- The traditional approach is based on visual inspection by a trained expert who identifies distinctive traits.
- The CSAFE approach [6,7] decomposes writing samples into graphs, roughly matching letters, and assigns each into one of 40 clusters.
- Cluster frequency has been used as a feature to answer the common source problem [7] since writers are expected to write following similar patterns.

Data and notation:

- Let u_x and u_y cluster frequencies from E_x and E_y respectively.
- $\boldsymbol{A}$ a set of background measurements used to construct the SLR system.

$$A = \{A_{ij}: i^{th}\ writer, j^{th}\ document\}$$

- Pairwise comparisons are created from the set $\boldsymbol{A}$ and classified as known match ("KM") or known non match ("KNM").

- The forensic proposition can be translated into sampling models that generated the data M_p and M_d respectively to define the training and estimation set [3].

Under M_p, KM are used:

$$C_{CS_P} = \{C(A_{ij}, A_{kl}): i = k\}$$

The set of pairwise comparisons with the same source.

Under M_d, KNM are used:

$$C_{CS_D} = \{C(A_{ij}, A_{kl}): i \neq k\}$$

The set of pairwise comparisons between two different sources.

- Machine learning-based comparison metrics and density estimation procedures rely on the independence assumption, but this assumption is not met.
- At a source level: Sources are compared multiple times.
  - In C_{cs_P}: multiple within comparisons uses the same source.
  - In C_{cs_D}: multiple between comparison use the same sources.
  - Same sources appear in both C_{cs_P} and C_{cs_D}
- At an item level: QD are compared multiple times. e.g., $C(A_{11}, A_{21}), C(A_{11}, A_{31})$

3. Sampling and ensembling algorithms.

- We propose a sampling approach to generate multiple base SLR.

Strong Source Sampling Algorithm (SSSA)

1. Construct all pairwise comparisons.
2. For KM pairs:
   Sample one pair to be used in the final database.
   Remove from candidate pool pairs involving selected sources.
3. For KNM pairs:
   Sample one pair to be used in the final database.
   Remove from candidate pool pairs involving selected sources.
4. Repeat 2 and 3 until data is exhausted.

Result: A pairwise database where sources are used only once.

Base Score Likelihood Ratio (BSLR)

For 1:M

1. Use SSSA to generate a training set.
2. Train a machine learning comparison score.
3. Use SSSA to generate an estimation set.
4. Compute scores for cases in the estimation set.
5. Estimate the distribution of scores under both propositions (or ratio estimator).
6. Store the comparison metrics and distributions

Result: M- base score likelihood ratios (BSLR).

- The M base SLR can be combined into a final ensembled SLR.
  - Naïve: mean, median, or majority voting.
  - Optimized: weight BSLR outputs or votes by performance in an optimization set.

Optimization of BSLR

1. Use SSSA to generate an optimization set.
2. Compute a performance metric.
3. Derive weights according to performance.

Result: M weights associated to each BSLR

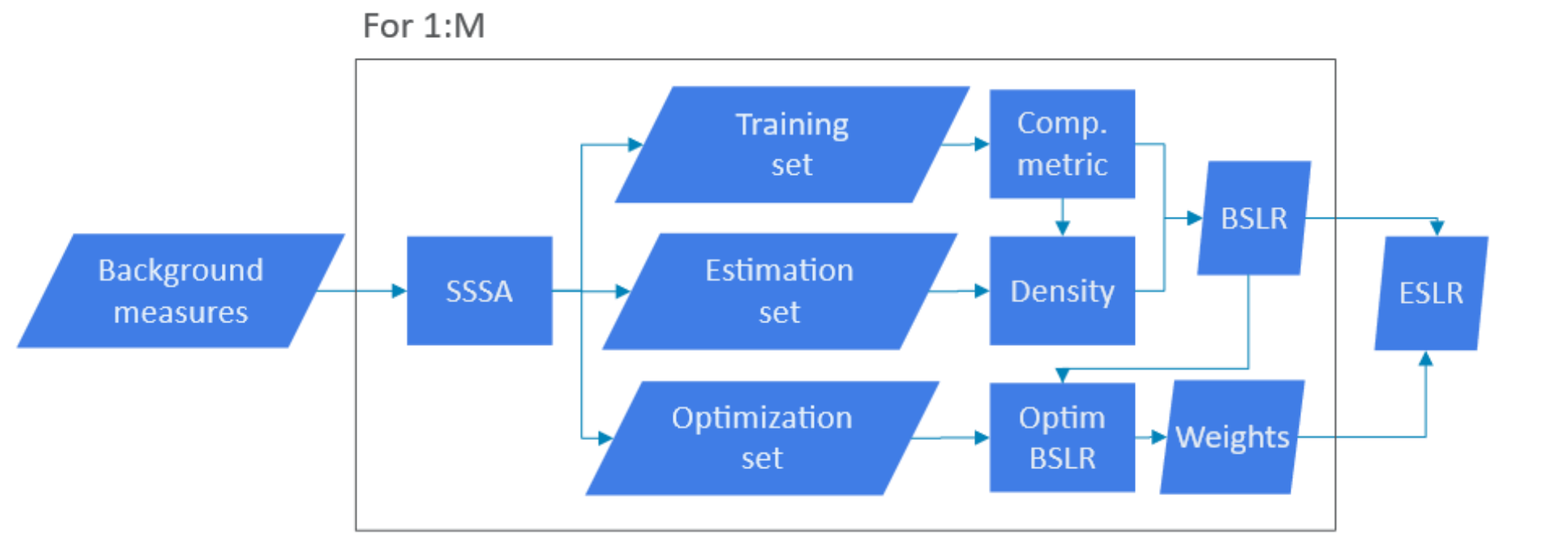
Ensembled Score Likelihood Ratio (ESLR)

For a new forensic comparison

1. Compute individual BSLR scores
2. Ensemble the BSLR into a final ESLR.

Result: Final value of evidence.

Proposed workflow for the ESLR



4. Simulation and results.

- To illustrate our approach, we used CSAFE'S London Letter as background measurements and the CVL set for validation and repeated 500 times the following:
  - We generated 50 BSLR and corresponding weights following our ESLR workflow.
  - We generated a traditional SLR by splitting the sources in the background population and downsampling to obtain balanced training and estimation sets.
  - We generated a validation set of 1000 known matches and 1000 known non-matches.
- We computed performance metrics to evaluate traditional SLR and ESLRs (Exp. 1); and held the same validation set across repetition to evaluate consensus (Exp. 2).

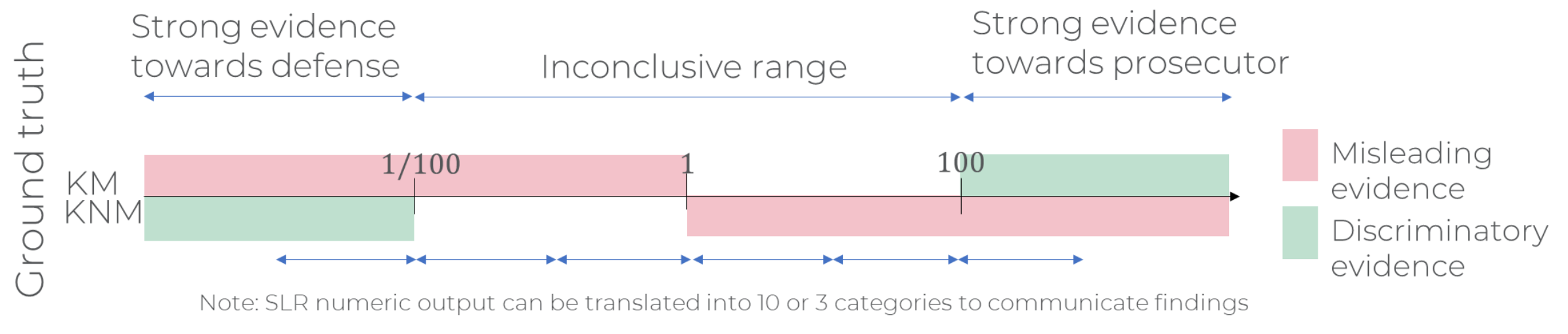
Metrics used.

RME KM (KNM): The percentage of true KM (KNM) that falls in the misleading range.

DP KM (KNM): The percentage of true KM (KNM) that falls in the discriminatory range.

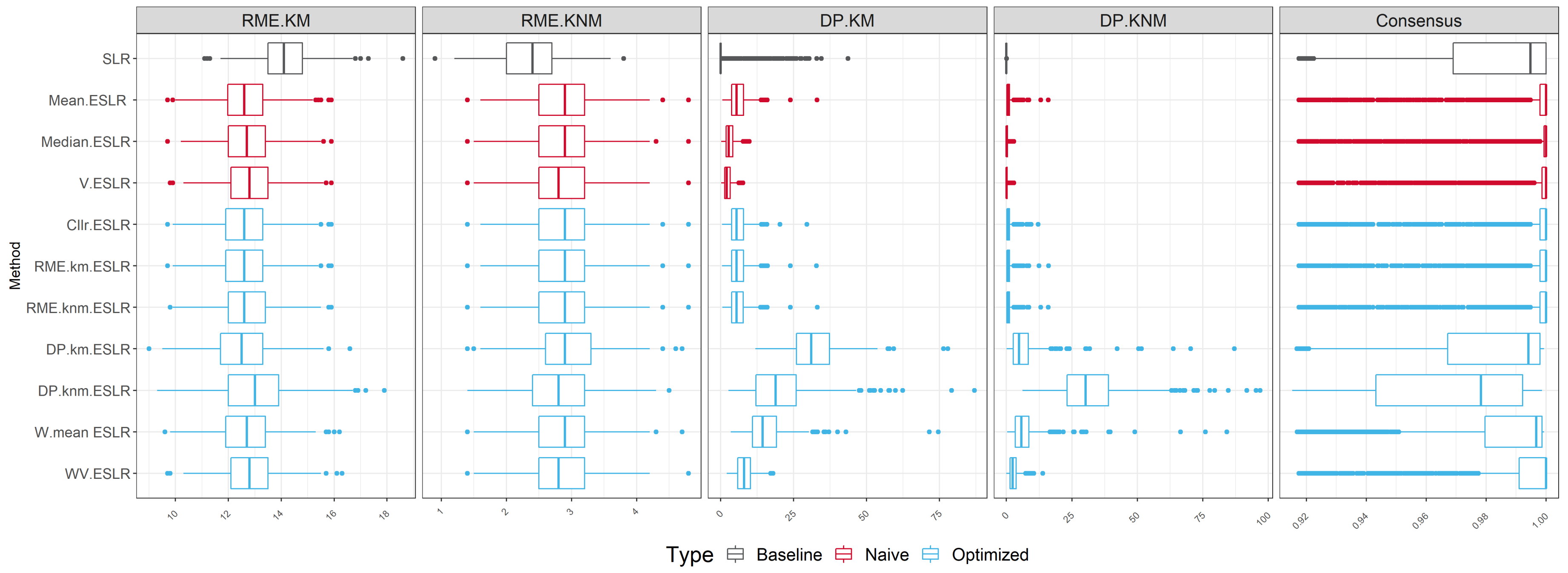
Consensus: Agreement is measured in the ten verbal category scale.

SLR output interpretation



Note: SLR numeric output can be translated into 10 or 3 categories to communicate findings

Performance evaluation: traditional SLR and ensemble alternatives.



Note: SLR denotes the traditional approach. Mean, Median combines the numeric output of BSLR while Vote transforms the numeric output into verbal categories and combine result using majority voting.

Optimized ESLRs are computed using weights derived from a cost function (Cllr), rate of misleading evidence (RME) for KM and KNM, discriminatory power (DP) for KM and KNM, and a multivariate weight (W) combining RME and DP. Weighted vote (WV) combines verbal categories using multivariate weights.

Best performers minimize the rate of misleading evidence and maximize discriminatory power. RME and DP are computed over Experiment 1. Consensus is computed over Experiment 2.

5. Discussion.

- Independence is a crucial assumption for machine learning and density estimation procedures, both cornerstones in machine learning-based score likelihood ratios.
- We introduce a sampling algorithm that remediates the dependence structure in forensic comparisons.
- Over these new data sets, popular methods that require independence can be applied.
- As in ensemble learning, multiple base SLR systems can be developed, allowing them to learn from a partial view of the data and aggregate their conclusion into a final SLR score.
- Our results show that ESLRs can perform better than traditional SLRs:
  - For this data, traditional SLR tends to produce inconclusive results.
  - Naïve and optimized ESLR increased discriminatory power and reduced the rate of misleading evidence for KM at the cost of a small increase in the rate of misleading evidence for KNM.
  - Traditional SLRs are affected by changes in the training and estimation set. Our more naïve approaches reduce the sensitivity to changes in the training and estimation set.

6. Further research.

- Our current work illustrates how ensemble learning can improve Score Likelihood Ratio systems.
- This approach is not limited to handwriting analysis, it can be extended to other forensic domains. Furthermore, our approach could be used for any common source problem that requires the creation of a multiple comparisons database.
- The methods presented are more akin to bagging. In the future, we plan to explore the use of boosting as an ensemble method to construct a final ESLR.

7. References.

[1] Stern, H., "Statistical Issues in Forensic Science," Annual Review of Statistics and Its Application 4 (2017): 225–244.

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

[3]Kleber, F., Fiel, S., Diem, M., & 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 (pp. 560-564).

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

[5]Crawford, A. M., Berry, N. S., & 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.

[6] Berry, N., Taylor, J., Baez-Santiago, F.. (2021) Handwriter: Handwriting Analysis in R.

[7] Johnson, M. Q., & Ommen, D. M. (2021). Handwriting identification using random forests and score-based likelihood ratios. Statistical Analysis and Data Mining: The ASA Data Science Journal

8. Access the data.

The CSAFE Handwriting Database is made publicly available.

The database allows researchers to develop models aimed at handwriting evaluation in forensics.

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, 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.