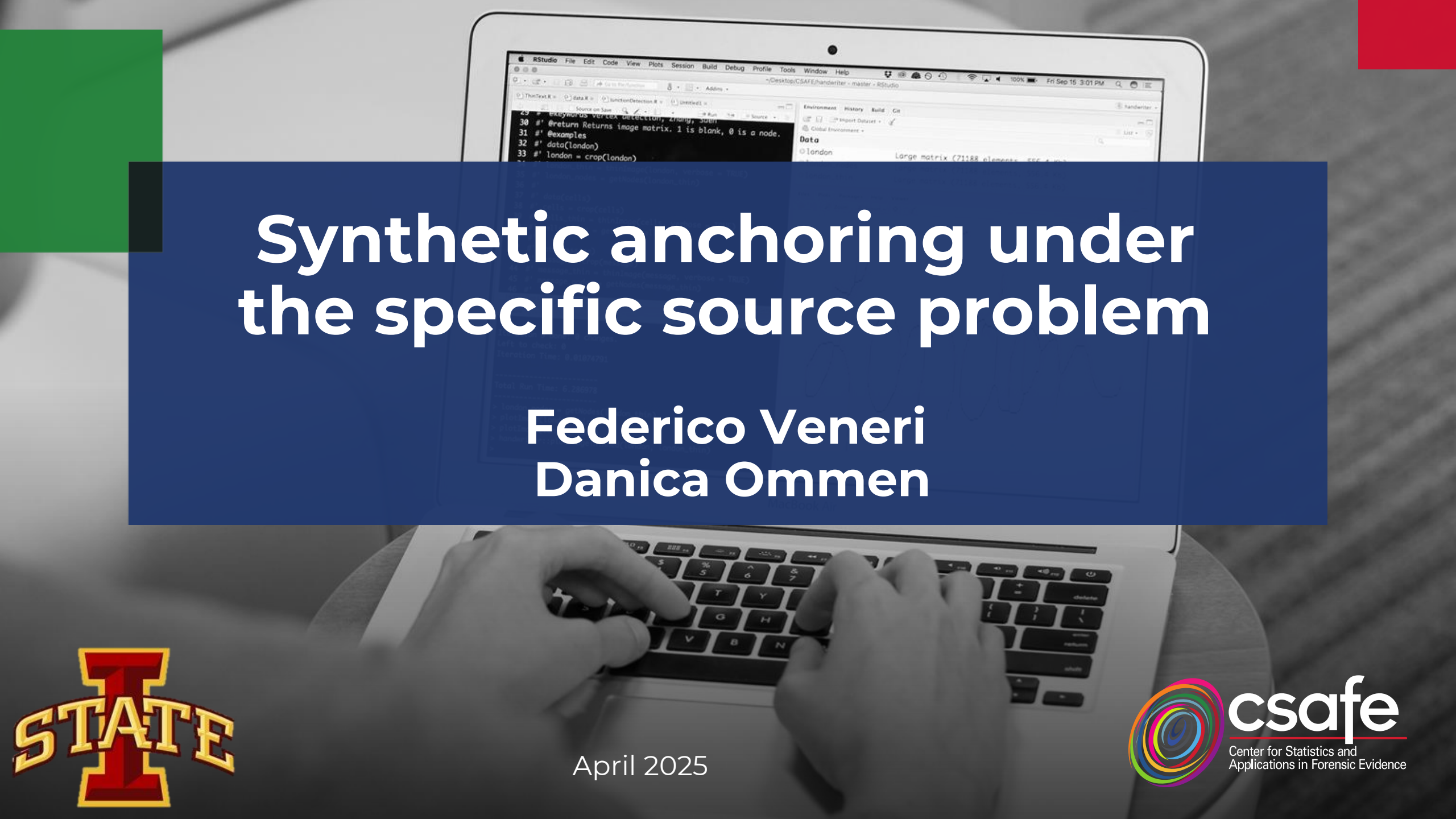


Synthetic anchoring under the specific source problem

Federico Veneri
Danica Ommen

April 2025

Center for Statistics and Applications in Forensic Evidence

Acknowledgments

Funding statement

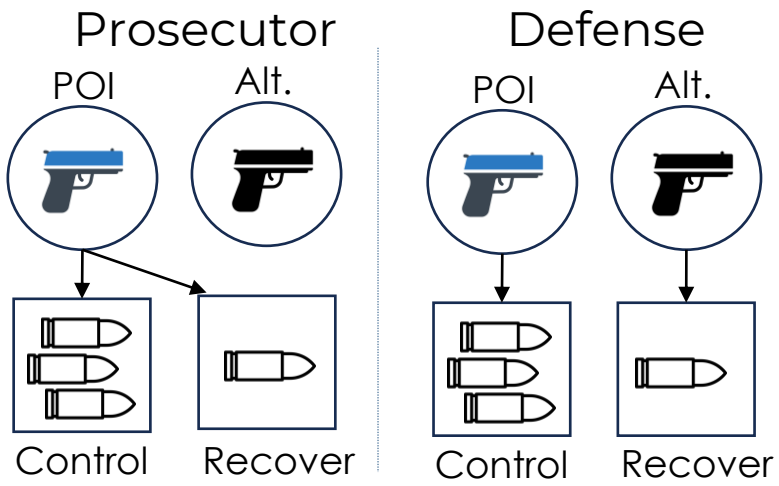
This work was funded (or partially funded) by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement 70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of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

Statistics in forensics

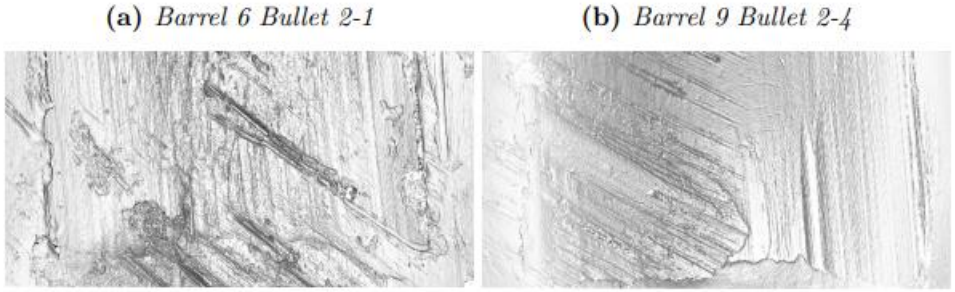
- In criminal cases, forensic experts examine the evidence and present their findings to judges and jurors who make an ultimate decision regarding the guilt/innocence.

- Professional guidelines recommend that findings should be based on sound statistical foundations and a probabilistic framework that allows to communicate uncertainty [22].

- One of the primary task faced by experts are source attribution problems [1]:



Your data may look like this:



Can we pose a probabilistic models for the features and use Likelihood ratios?

Verbal scale

Score Likelihood Ratios (SLRs)

- Statistician have explored the use of machine learning (score) based SLRs as a **likelihood free** approach to contrast propositions [4,21].

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

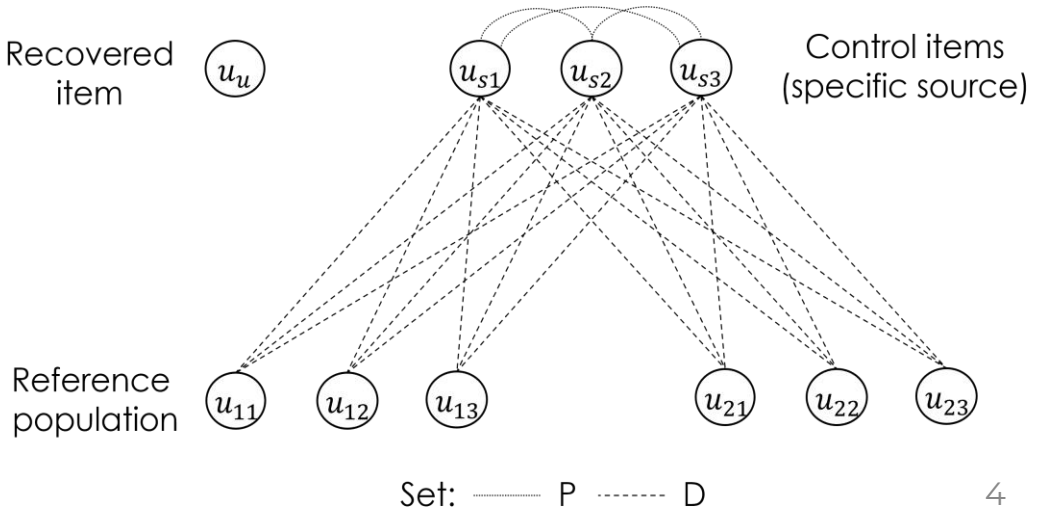
Where:

δ is an observed score between two features vectors

g denote conditional densities under a propositions

- Interpretation same as LR but focusing on the likelihood of observing a given score.

- In practice, researchers may need to:
  - Train $\Delta(x, y) = \delta$
  - Estimate $g(\delta \mid H_j)$
  - Evaluate their methods
- However, data is limited.



Our proposal: Synthetic anchoring.

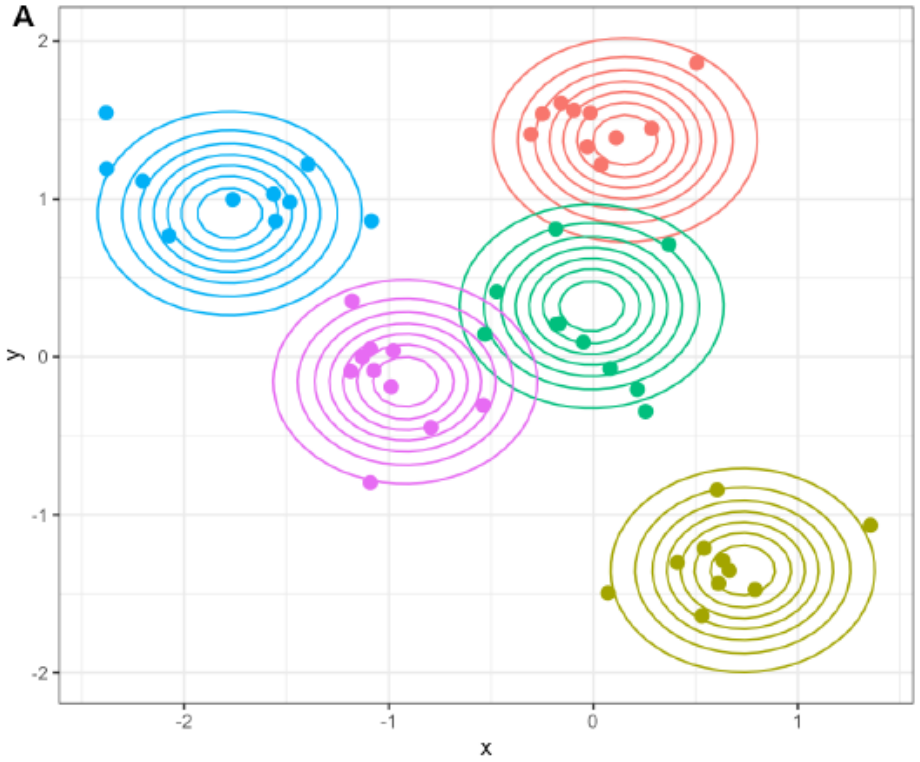
- We first create **synthetic items**.
  - This idea has been already been popularized by Synthetic minority Oversampling (SMOTE) in classification for class imbalance problems [29,30].
  - The Key component in the original algorithm is linear interpolation between learning instances.
  - Our approach is similar but instead of interpolating between learning instances we interpolate between observed items to generate new synthetic items.

- To generate **learning instances**, we anchor on the **specific source**
  - This synthetic items are used to emulate how data should be generated [1,9].
  - We create new sets: P and D, that can be used for estimation.

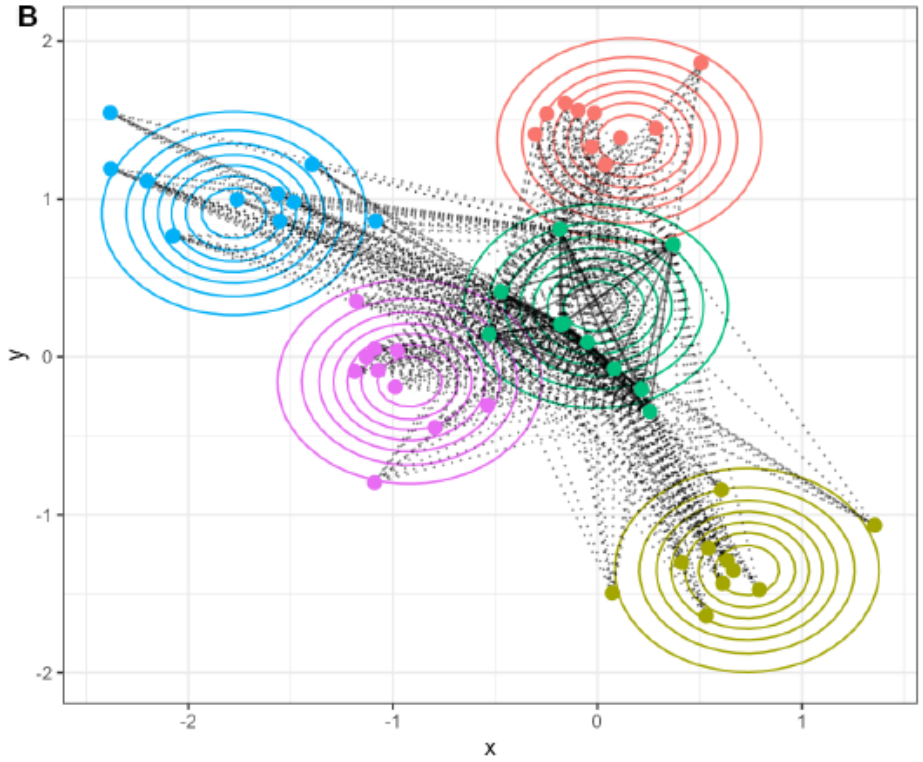
A Gaussian bivariate illustration.

- Let's consider 5 sources. Green is the specific source.

Simulated data



All pairs (Naïve)

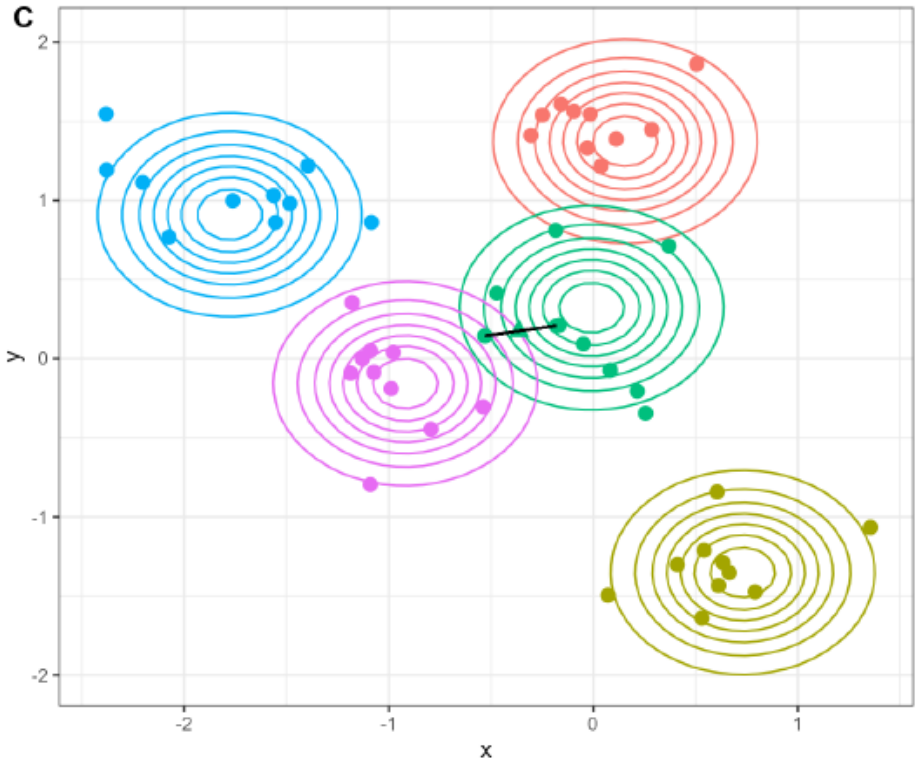


- Panel B: All pairs uses specific source items multiple times.

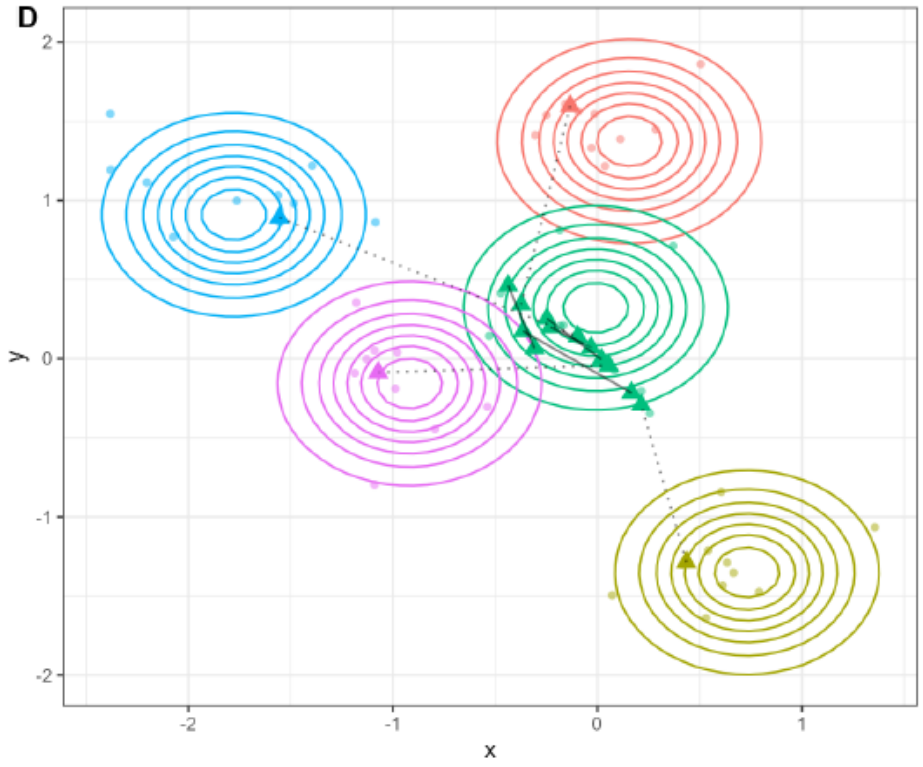
A Gaussian bivariate illustration.

- Let's consider 5 sources. Green is the specific source.

Simulated data



Synthetic anchoring



- Panel C: Generating a synthetic item.
- Panel D: Data to develop an SLR

In our work:

Simulation study:

- Comparing a scenario were the DGP is known, and we can generate data vs our synthetic approach

Applications:

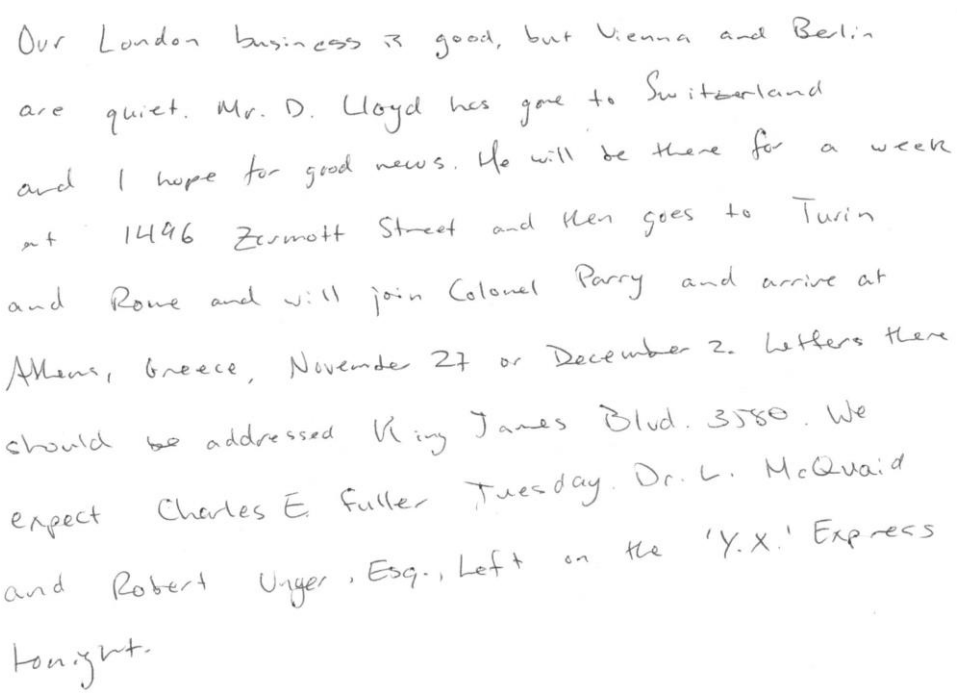
- Glass analysis , FIU dataset (Trace evidence)
- **Handwriting, CSAFE (Pattern evidence)**

- A motivating (toy) example:
  - A handwritten threatening note (recovered) was sent to a Statistics professor (Victim).
  - An unhappy student who did not pass Stats 101, is suspected of sending the note (POI).
  - A collection of his handwriting notes (control) was collected and compared.

- Statistical question:
  - Did the writer (source) who wrote the collected handwriting notes (control) also wrote the threatening notes (recovered)?

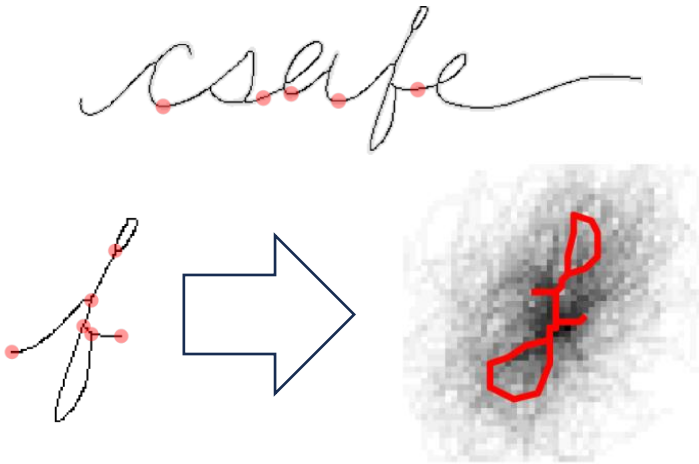
An application in Handwriting

Writer 12, rep 1 (i=12,j=1)



- We follow an approach developed by CSAFE authors [16] that decomposes writing samples into graphs (roughly matching letters) and classify them into one of 40 groups from a previously developed clustering template.

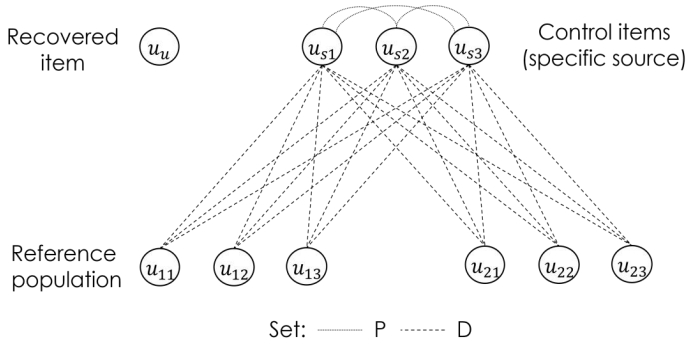
- Cluster proportions constitute writership profiles that have been used to characterize writers in close-set problems [17] and open-set problems [18,25].



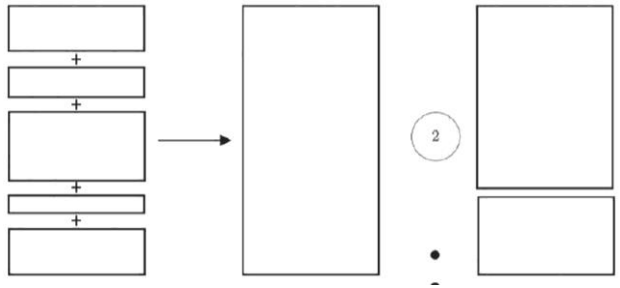
Handwriting: Application details

- This applications considers the writership profile of 90 writers (9 items each)
- Additional details:
  - We cycle over each writer:
  - Designate the first item as recovered item,
  - H_p:Remaining QD 2-9 as controls.
  - H_d: Find the closest non match, QD 2-9 as controls.
  - We consider cosine distances and the Beta parametric family.
- Across iterations we compare the performance of three resampling plans.

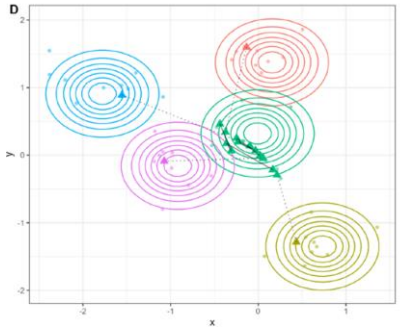
Naïve



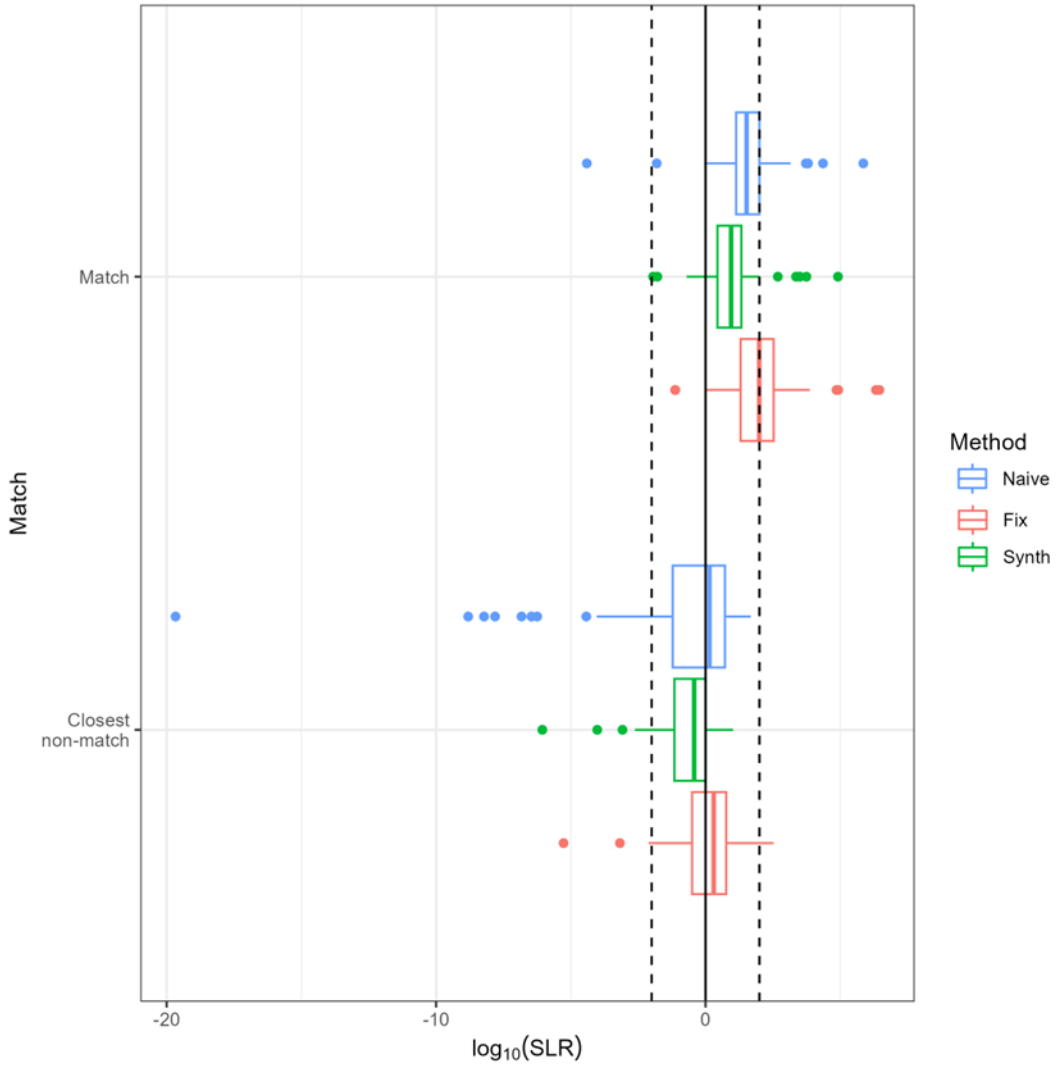
Fix split



Synthetic anchoring



An application in Handwriting



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

- Synth achieved smaller cost, better performance (<1).
- Synth presented more conservative SLRs
  - Smaller RME_{Hd} at the cost of larger RME_{Hp}
  - Smaller capacity of providing stronger evidence.
- Conclusion:
  Overall, better performance but more conservative

Conclusions

- SLRs are proposed as an alternative for evaluating complex evidence.
  - Current approach relies on pairing items, creating dependence.
  - Further, the specific source is conditioned on POI and data is limited.
- Our work proposes using synthetic items as a data augmentation tool and a resampling plan to alleviate the dependence structure.
- Simulation results show that for the well-known data-generating process with realistic parameters, the proposed approach and the theoretically correct system tend to agree, albeit our proposed methods tend to be more conservative.
- For our application in handwriting, we compared our approach to two other resampling plans. Our proposed method outperformed them in terms of the rate of misleading evidence for the defense at the expense of a small increase in the rate for the prosecutor.
- Overall, we observe a reduction in the cost incurred as measured by the Cllr.

Comments or suggestions

- Thank you for your time today,

- We appreciate any comments or suggestions you may have about the current (and future) work.