Quantifying Bayes Factors for Forensic Handwriting Evidence

Anyesha Ray and Danica Ommen

Iowa State University

Abstract

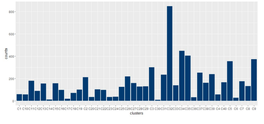
Questioned Document Examiners (QDEs) are tasked with analyzing handwriting evidence to make source (or writership) determinations. The Center for Statistics and Applications of Forensic Evidence (CSAFE) has previously developed computational methods to automatically extract quantifiable handwriting features [1] and statistical methods to analyze handwriting evidence to aid QDEs [2,3]. Our goal is to develop a method that supports feature-based open-set writer identification. We implement an approach to quantify the value of forensic handwriting evidence using Bayes factors and Markov chain Monte Carlo (MCMC) computational techniques like those described in Collins and Ommen [4].

Introduction

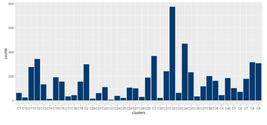
CSAFE has developed many methods to identify writership. The method developed by Crawford et. al uses a K-means clustering algorithm and Bayesian hierarchical model to perform closed-set writer identification [2]. Another method developed by Johnson and Ommen utilized machine learning techniques and score-based likelihood ratios (SLRs) [3]. SLRs have been criticized for including a lack of coherence and ability to incorporate the rarity of the features. Using Bayes Factors enables us to have an open set of writers, while still avoiding these issues.

Data

Handwritten samples were collected by CSAFE. We include data from 240 adults with three data collection sessions resulting in 27 pages of writing from each participant. The CSAFE data is broken down using handwriter. Writing samples are broken into graphs, and each graph is categorized into a cluster group based on their basic shapes (simplified example is that all things that look like an “e” will go into the same cluster). We use 40 clusters. Then, we created a dataset that has the cluster counts for each writer.



Histogram of graph count sums across all pieces of writing from writer 1



Histogram of graph count sums across all pieces of writing from writer 313

Methods and Results

Prosecution Hypothesis H_p: The QD was written by the suspect.
Defense Hypothesis H_d: The QD was written by someone else in the alternative source population.

- u_i is the vector of graph counts from the unknown document will be a vector of length 40 because we have 40 clusters (evidence obtained from crime scene)
- s_j is the vector of graph counts from the control document will be a vector of length 40 because we have 40 clusters. Here we can have multiple documents (material we gather from suspect)
- A_{ij} is vector of graph counts in each of the 40 clusters from document j and alternative writer i in the population where i is number of different writers in alternative and j is number of documents.

The specific source was modeled as a Dirichlet-Multinomial(θ_s) where θ_s is the rate of observing graphs in the clusters for the specific source. Each specific source will have 40 θ_s. For the alternative source population, we use a hierarchical model where θ_s is drawn from a Dirichlet and then a Multinomial. To make the model simpler, we sum all the documents from the specific source into one row s.

$$M_s : s \sim Fs(\cdot|\theta s) = Dirichlet - Multinomial(\theta_s)$$

and

$$M_a : B_i \sim G(\cdot|\theta_a) = Dirichlet(\theta_a)$$
$$A_i|B_i = b_i \sim Fa(\cdot|b_i, \theta_a) = Multinomial(b_i)$$

where marginally

$$A_i|\theta_a \sim Dirichlet - Multinomial(\theta_a)$$

We make the same simplifications for the questioned documents, u, and sum them together by cluster.

$$M_p : u \sim Dirichlet - Multinomial(\theta_s)$$
$$M_d : u \sim Dirichlet - Multinomial(\theta_a)$$

To compute the Bayes Factor, we use the following formula:

$$B_{ss}(E) \approx \frac{\frac{1}{n_p}\sum_{i=1}^{n_p} f(\mathbf{u}|\theta_s^{(i)})}{\frac{1}{n_d}\sum_{i=1}^{n_d} f(\mathbf{u}|\theta_a^{(i)})}$$

Where n_p and n_d are the number of posterior samples and $\theta_s^{(i)}$, $\theta_a^{(i)}$ are samples from the posterior distributions.

Here, we use writer 36 as our specific source. Every other writer in the data set is our alternative source, and we have two questioned documents, one from writer 36 and one from writer 71.

Questioned Document	Bayes Factor
Writer 36 Session 1 LND rep 1	54.35
Writer 71 Session 1 LND rep 1	4.24184e-05

Conclusion

A Bayes Factor of 1 indicates that the prosecution proposition and the defense proposition are equally favored. A Bayes Factor greater than 1 indicates that the H_p is favored and a Bayes Factor less than 1 indicates that the H_d is favored.
From the Bayes Factors shown in the Methods and Results section, we can conclude that our Questioned Document from writer 36 did come from our specific source because it is greater than 1. The Bayes Factor for the second questioned document signifies that the questioned document is from an alternative source as it is a very small number. Since we know the true source is writer 36, these numbers represent the correct conclusion.

References

[1] Nick Berry, James Taylor, and Felix Baez-Santiago. handwriter: Handwriting Analysis in R, 2021. R package version 1.0.1.
[2] A. Crawford, Bayesian hierarchical modeling for the forensic evaluation of handwritten documents, Ph.D. thesis, Iowa State University, 2020.
[3] Madeline Quinn Johnson and Danica M. Ommen. Handwriting identification using random forests and score-based likelihood ratios. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15(3):357–375, 2022.
[4] Collins GL, Ommen DM. Quantifying the Value of Forensic Handwriting Evidence using Open-Source Feature Extraction [thesis]. 2022.
[5] Amy Crawford, Anyesha Ray, Alicia Carriquiry, James Kruse, and Marc Peterson. CSAFE Handwriting Database. 11 2019.
[6] Danica M. Ommen and Christopher P. Saunders. A problem in forensic science highlighting the differences between the bayes factor and likelihood ratio. Statistical Science, 36(3):344–359, Aug 2021.

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 Ne bra ska, Lincoln.