Score-based Likelihood Ratios for Handwriting Evidence

Christopher P. Saunders
Assistant Professor of Statistics
South Dakota State University
Disclaimer

This work was supported in part under a Contract Award from the Counterterrorism and Forensic Science Research Unit of the Federal Bureau of Investigation's Laboratory Division.

Names of commercial manufacturers are provided for information only and inclusion does not imply endorsement by the FBI.

Points of view in this presentation are those of the authors and do not necessarily represent the official position of the FBI or the US Government.
The research detailed in this presentation was supported in part by Award No. 2009–DN–BX–K234 awarded by the National Institute of Justice, of Justice Programs, US Department of Justice.

The opinions, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect those of the Department of Justice.
Background

- In 2008, we began exploring likelihood ratios as a tool for presenting handwriting evidence in a forensic context.

- Focus has been on Score-Based Likelihood Ratios (SLRs)
  - Three general classes of SLRs
  - Currently working on characterizing properties of SLRs
This presentation is a summary of results from the fourth in a series of papers related to handwriting evidence.


Overview

We will discuss:

- Bayesian paradigm for interpreting evidence
- Score-based likelihood ratios as approximations to the value of evidence
- Statistical interpretations of forensic hypotheses and the corresponding classes of SLRs
The Bayesian Paradigm

What do we believe the likelihood of observing the questioned document is if the suspect wrote the questioned document, given what we know about the suspect?

vs.

What do we believe the likelihood of observing the questioned document is if it was written by a writer in the alternative source population, given what we know about the alternative source population?

In a formal Bayesian paradigm, this is the comparison we have to address as forensic statisticians with respect to handwriting identification.
Evaluation of Handwriting Evidence

\[
P(H_p | E, I) \cdot \frac{P(E | H_p, I)}{P(E | H_d, I)} = \frac{P(H_p | E, I)}{P(H_d | E, I)}
\]

- **Posterior Odds**
- **Likelihood Ratio and/or Bayes Factor**
- **Prior Odds**

\[E: \text{Evidence}\]
\[H_p: \text{Suspect wrote the questioned document (QD)}\]
\[H_d: \text{Suspect did not write the QD}\]
\[I: \text{Background information}\]

“The evidence is LR (BF) = 100 times more probable if the suspect wrote the QD than if some unknown person wrote it.”
We partition the evidence: 

\[ E = \{E_S, E_U, E_A\} \]

where:

\[ E_S = \text{Sample(s) obtained from the Specific source} \]

\[ E_U = \text{Sample(s) of Unknown source obtained at the crime scene.} \]

\[ E_A = \text{Sample(s) taken from Alternative sources, comprising a relevant database of other potential sources.} \]
Handwriting Assumptions

We assume:

- Every individual has an unobservable *writing profile*, that is constant over time.
- A document is a random sample from a writing profile.

Estimate the profile by collecting many documents, forming a *writing template* (or simply *template*).
Quantification

- A proprietary, automated process developed by Gannon Technologies Group processes handwritten documents.

- This process uniquely associates each parsed character’s skeleton with a graphical isomorphism.
Obtain *measurements* from $E_s$, $E_u$, and each alternate writer’s template in $E_a$.

**Frequency Distribution of Letter/Isocode Usage in a Single Writing Sample**

<table>
<thead>
<tr>
<th>Letter/Isocode</th>
<th>Topology 1</th>
<th>Topology 2</th>
<th>( \cdots )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \cdots )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \cdots )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \cdots )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \cdots )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ X = \text{matrix of counts (MOC)} \]
Likelihood Ratio or Bayes Factor

\[
\frac{P(E|H_p, I)}{P(E|H_d, I)} = \frac{P(X_S, X_U|H_p, I)}{P(X_S, X_U|H_d, X_A, I)}
\]

\(X_S\) = MOC obtained from the Suspect’s template

\(X_U\) = MOC of Unknown source (QD)

\(X_A\) = Collection of MOCs taken from Alternative sources, comprising a relevant database of other potential suspects.
Score Based Likelihood Ratios

- A score-based approach would reduce the dimensionality of the problem.
- The score-based likelihood ratio (SLR) is

\[
SLR = \frac{\Pr(\Delta(X_s, X_u) = \delta | H_p, I)}{\Pr(\Delta(X_s, X_u) = \delta | H_d, X_A, I)}
\]

where \( \Delta \) is a dissimilarity score between two MOCs.
Generic SLR Algorithm

1. Select a statistic, $\Delta$, to assess dissimilarity between two items of evidence.

2. Develop a database of $\Delta$’s for items of known origin, under $H_p$ and $H_d$.

3. Construct an empirical distribution for $\Delta$, under $H_p$ and $H_d$.

4. Evaluate a specific $\delta$ for two items of evidence.

5. Consider in context: evaluate the numerator and denominator distributions from (3) at the point $\delta$.

**Statistical vs. Forensic Propositions**

*Forensic propositions:*

- \( H_p \): Suspect (\( S \)) wrote the QD.
- \( H_d \): \( S \) did not write the QD.

*Statistical propositions:*

- \( H_p \): Evidence score arises from the distribution of scores obtained by pairing a randomly selected QD and a randomly selected template *both written by \( S \).*
Statistical propositions, cont.

Crime scene anchored:

$H_{dl}$: Evidence score arises from the distribution of scores obtained by pairing the QD with a template written by a random individual.
Statistical propositions, cont.

**Crime scene anchored:**

$H_{d1}$: Evidence score arises from the distribution of scores obtained by pairing the QD with a template written by a random individual.

**Suspect anchored:**

$H_{d2}$: Evidence score arises from the distribution of scores obtained by pairing a QD written by a random individual with the template written by S.
**Crime scene anchored:**

\[ H_{d1}: \text{Evidence score arises from the distribution of scores obtained by pairing } \textit{the QD} \text{ with } \textit{a template} \text{ written by a random individual.} \]

**Suspect anchored:**

\[ H_{d2}: \text{Evidence score arises from the distribution of scores obtained by pairing } \textit{a QD} \text{ written by a random individual with } \textit{the template} \text{ written by } S. \]

**General match:**

\[ H_{d3}: \text{Evidence score is a realization from the distribution of scores obtained by pairing } \textit{a QD} \text{ written by a random individual with } \textit{a template} \text{ from a different random individual.} \]
Specific SLR Algorithm

1. **Kullback-Leibler** (KL) divergence
   - Obtain KL, by letter, between a QD and a template.
   - The overall dissimilarity is then a weighted average, over all letters.
     - Weighting accounts for differing numbers of observed characters across letters. (Letter ‘a’ is used more than ‘z’.)
   - One advantage of the KL is that it is non-symmetric:
     - QD is typically much smaller than the template.
   - Many other choices exist – additional research needed.
Specific SLR Algorithm

2. Database of scores: Numerator

- Under $H_p$, QD and template written by suspect
- Assume we have a collection of prior writings obtained from the suspect – suspect’s template.
- Ideally, we would also have a collection of samples from the suspect ‘similar’ to the QD (e.g. same number of characters).
- Then, to generate our database, we would compute the KL between each of the ‘QD-type’ documents and the suspect’s template.
Specific SLR Algorithm

2. Database of scores: Numerator

- Unfortunately, we typically do not have access to a sufficient number of ‘QD-type’ documents written by the suspect.

A subsampling routine:

A. Create:
   - Pseudo-QD: randomly select $n$ characters from template
     ($n = \# \text{ of chars in QD}$)
   - Pseudo-template: remaining characters

B. Compute $KL$ between pseudo-QD and pseudo-template.

C. Repeat A, B many times to generate database of scores.
Specific SLR Algorithm

**Database of scores: Denominator**

- Assume we have access to a collection of templates from a large number of alternative sources.
- One approach would be to compute KL between the actual QD and each template in the collection.

**Histogram Estimator as empirical distribution.**

- Many other choices – additional research needed.

**Evaluate**

\[ \Delta(X_S, X_U) = \delta. \]
Specific SLR Algorithm, cont.

Consider in context:

\[
\hat{S}_{LR} = \frac{\hat{g}(\Delta(X_S, X_U) = \delta | H_p, I)}{\hat{g}(\Delta(X_S, X_U) = \delta | H_d, X_A, I)}
\]
Calculating SLR for handwriting

**SLR numerator:**

- Obtain \((X_{Si}, X_{Ui})\) for \(i = 1, \ldots, 500\) *pseudo-QDs* obtained via *subsampling* (AAFS09) from the suspect’s template.

- Obtain \(\Delta(X_{Si}, X_{Ui})\) for each pseudo-QD.

- Estimate the numerator distribution: \(\hat{g}\).

- Evaluate \(\hat{g}\left(\Delta(X_S, X_U)\right| H_p, I\).\)
Calculating SLR for handwriting

**SLR denominator 1 - crime scene anchored:**

- Obtain \( X_{Ai} \) from templates taken from each writer \( i \) in \( E_A \).
- Obtain \( \Delta(X_{Ai}, X_U) \) for each writer \( i \) in \( E_A \), where \( X_U \) from the QD.
- Estimate the denominator distribution: \( \hat{g} \).
- Evaluate

\[
\hat{g} \left( \Delta(X_S, X_U) \mid H_d, X_A, I \right).
\]
Calculating SLR for handwriting

SLR denominator 2 - suspect anchored:

- Obtain $X_{Ui}$ from 500 pseudo-QD’s sampled from $E_A$.
- Obtain $\Delta(X_S, X_{Ui})$ for each pseudo-QD, where $X_S$ is obtained from suspect’s template.
- Estimate the denominator distribution: $\hat{g}$.
- Evaluate

$$\hat{g}\left(\Delta(X_S, X_U) \mid H_d, X_A, I\right).$$
Calculating SLR for handwriting

SLR denominator 3 – general match:

- Obtain $X_{Ai}$ from 500 templates sampled from $E_A$.
- For each $i$, obtain $X_{Ui}$ by randomly selecting a pseudo-QD from $E_A$, ensuring that writer of $X_{Ui}$ $\neq$ writer of $X_{Ai}$.
- Obtain $\Delta(X_{Ai}, X_{Ui})$.
- Estimate the denominator distribution: $\hat{g}$.
- Evaluate

$$\hat{g}\left(\Delta(X_S, X_U)\mid H_d, X_A, I\right).$$
### Results: Specific Cases

\[ QD = \text{first 60 characters of document} \]

<table>
<thead>
<tr>
<th></th>
<th>( H_p \text{ True} )</th>
<th></th>
<th>( H_d \text{ True} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S \text{ ID:} )</td>
<td>( 0799 )</td>
<td>( 0772 )</td>
<td>( S \text{ ID / True ID:} )</td>
</tr>
<tr>
<td>( SLR_1 )</td>
<td>1858</td>
<td>2370</td>
<td>( SLR_1 )</td>
</tr>
<tr>
<td>( SLR_2 )</td>
<td>1701</td>
<td>6</td>
<td>( SLR_2 )</td>
</tr>
<tr>
<td>( SLR_3 )</td>
<td>15</td>
<td>19</td>
<td>( SLR_3 )</td>
</tr>
</tbody>
</table>
Implications

- SLR’s are starting to be introduced in US legal system with respect to fingerprints.

- Likelihood ratios in general are very hard to defend against.

However-

*To date we have empirically demonstrated that the three standard SLR’s give different values of the Evidence.*
Why are the SLRs different?

- SLR ≠ LR.
- By replacing $E_u$ and $E_s$ with $\Delta(X_s, X_u)$ we are losing information.

\[
SLR = \frac{\Pr(\Delta(X_s, X_u) = \delta| H_p, I)}{\Pr(\Delta(X_s, X_u) = \delta| H_d, I)}
\]

\[
LR = \frac{\Pr(X_s, X_u| H_p, I)}{\Pr(X_s, X_u| H_d, I)}
\]
Why are the SLRs different?

- $\text{SLR}_1 \neq \text{SLR}_2 \neq \text{SLR}_3$.
- The conditioning arguments are different for each SLR-

**Crime scene anchored:**

$$SLR_1 = \frac{\Pr(\Delta(X_s, X_u) = \delta|H_p, I)}{\Pr(\Delta(X_a, X_u) = \delta|H_d, I)}$$

**Suspect anchored:**

$$SLR_2 = \frac{\Pr(\Delta(X_s, X_u) = \delta|H_p, I)}{\Pr(\Delta(X_s, X_a) = \delta|H_d, I)}$$

**General match:**

$$SLR_3 = \frac{\Pr(\Delta(X_s, X_u) = \delta|H_p, I)}{\Pr(\Delta(X_{a1}, X_{a2}) = \delta|H_d, I)}$$
Implications

- $\text{SLR} \neq \text{LR}$.

- It is not a straightforward task to statistically interpret forensic propositions.

- Different statistical interpretations can lead to very different conclusions.
Acknowledgements

- South Dakota State University
  - Douglas Armstrong
- The IC Post Doctoral Fellowship Program
  - Amanda Hepler
  - Eric Kalendra
- George Mason University
  - Linda Davis
  - John J. Miller
  - Don Gantz
- Gannon Technologies Group
- FBI, Counterterrorism and Forensic Science Research Unit
  - JoAnn Buscaglia
- National Institute of Justice
Additional Slides

- A Normal Digression
- Comparative Study- “FBI 500”
- Sub Sampling Data Bases
- Bayes Factors

Please see Helper et al. for details on these slides-
A Normal Digression

In general, it is technically difficult to solve for the exact distributions under the three stated statistical propositions.

- If we use squared distance between the normal random variables we can solve for the three SLR’s in closed form.

- Under $H_p$; Assume that: $X \sim N(\mu_x, \sigma_w^2)$  
  
  $Y \sim N(\mu_x, \sigma_w^2)$

- Under $H_d$; Assume that: $X \sim N(\mu_x, \sigma_w^2)$  
  
  $Y \sim N(\mu_A, \sigma_A^2)$
Comparative study

- Alternative source writing samples: FBI provided 2120 script documents (a convenience sample)
  - 5 documents from 424 individuals
  - Each document has approximately 550 characters
  - Text selected to be representative of English language

- Randomly select a writer to serve as ‘suspect.’

- Obtain SLR1-3 values for both scenarios.
  - $H_p$ True: S wrote QD
  - $H_d$ True: S did not write QD

- Repeat 1000 times.
## Results: $H_p$ True

### Agreement

<table>
<thead>
<tr>
<th></th>
<th>Supports $H_p$</th>
<th>Inconclusive</th>
<th>Supports $H_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SLR &gt; 10$</td>
<td>58%</td>
<td>4%</td>
<td>1%</td>
</tr>
<tr>
<td>$1/10 &lt; SLR &lt; 10$</td>
<td>4%</td>
<td>1%</td>
<td>37%</td>
</tr>
<tr>
<td>$SLR &lt; 1/10$</td>
<td>1%</td>
<td>31%</td>
<td>28%</td>
</tr>
</tbody>
</table>

### Disagreement

<table>
<thead>
<tr>
<th></th>
<th>Supports $H_p$</th>
<th>Inconclusive</th>
<th>Supports $H_d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SLR &gt; 10$</td>
<td>63%</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>$1/10 &lt; SLR &lt; 10$</td>
<td>67%</td>
<td>1%</td>
<td>31%</td>
</tr>
<tr>
<td>$SLR &lt; 1/10$</td>
<td>60%</td>
<td>24%</td>
<td>17%</td>
</tr>
</tbody>
</table>

**Correct**

*QD = first 60 characters of document*
## Results: $H_d$ True

<table>
<thead>
<tr>
<th>Agreement</th>
<th>Disagreement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supports $H_p$</strong></td>
<td><strong>Supports $H_d$</strong></td>
</tr>
<tr>
<td>SLR &gt; 10</td>
<td>SLR &lt; 1/10</td>
</tr>
<tr>
<td><strong>1 vs 2 vs 3</strong></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>16%</td>
</tr>
<tr>
<td><strong>1 vs 2</strong></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>18%</td>
</tr>
<tr>
<td><strong>1 vs 3</strong></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>18%</td>
</tr>
<tr>
<td><strong>2 vs 3</strong></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>19%</td>
</tr>
</tbody>
</table>

### QD = first 60 characters of document
# Results: Rates of Misleading Evidence

## $H_p$ True: Suspect wrote the QD

<table>
<thead>
<tr>
<th>LR Range</th>
<th>QD = 20 Characters</th>
<th>QD = 60 Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$SLR_1$</td>
<td>$SLR_2$</td>
</tr>
<tr>
<td>(0, 0.001]</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>(0.001, 0.01]</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>(0.01, 0.1]</td>
<td>0.004</td>
<td>0.007</td>
</tr>
<tr>
<td>(0.1, 1]</td>
<td>0.074</td>
<td>0.117</td>
</tr>
<tr>
<td><strong>RME</strong></td>
<td><strong>0.087</strong></td>
<td><strong>0.133</strong></td>
</tr>
<tr>
<td>(1, 10]</td>
<td>0.235</td>
<td>0.446</td>
</tr>
<tr>
<td>(10, 100]</td>
<td>0.106</td>
<td>0.114</td>
</tr>
<tr>
<td>(100, 1000]</td>
<td>0.209</td>
<td>0.137</td>
</tr>
<tr>
<td>&gt;1000</td>
<td>0.363</td>
<td>0.170</td>
</tr>
</tbody>
</table>
Results: Rates of Misleading Evidence

**H_d True: Suspect did not write the QD**

<table>
<thead>
<tr>
<th>LR Range</th>
<th>QD = 20 Characters</th>
<th>QD = 60 Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SLR₁</td>
<td>SLR₂</td>
</tr>
<tr>
<td>(0, 0.001]</td>
<td>0.195</td>
<td>0.139</td>
</tr>
<tr>
<td>(0.001, 0.01]</td>
<td>0.114</td>
<td>0.177</td>
</tr>
<tr>
<td>(0.01, 0.1]</td>
<td>0.093</td>
<td>0.050</td>
</tr>
<tr>
<td>(0.1, 1]</td>
<td>0.372</td>
<td>0.377</td>
</tr>
<tr>
<td>(1, 10]</td>
<td>0.201</td>
<td>0.239</td>
</tr>
<tr>
<td>(10, 100]</td>
<td>0.010</td>
<td>0.011</td>
</tr>
<tr>
<td>(100, 1000]</td>
<td>0.005</td>
<td>0.003</td>
</tr>
<tr>
<td>&gt;1000</td>
<td>0.010</td>
<td>0.004</td>
</tr>
<tr>
<td><strong>RME</strong></td>
<td>0.226</td>
<td>0.257</td>
</tr>
</tbody>
</table>
Subsampling Algorithm - Suspect Specific database

For a questioned document of size $n$ characters and a suspect writer template of $N$, we then:

- Randomly divide the template into two sets of characters of sizes $n$ and $N-n$.
- These are the pseudo-document and pseudo-writing template, respectively.
- Compare the pseudo-document and pseudo-writing template with the score function.
- Record the similarity score

Repeat $k$ times to obtain a dataset of size $k$
Parametric Approach

- Conditioning on the letter, we can assume the corresponding row follows a multinomial distribution.

- There are thousands of topologies that can be assigned to a given letter – each row could have thousands of cells.
  - This leads to some ambiguity when specifying the priors necessary for computing the Bayes factor.

- We have explored these issues with limited success.
A note on the Bayes Factor

$$\frac{P(E|H_p, I)}{P(E|H_d, I)} = \frac{P(E_S, E_U, E_A|H_p, I)}{P(E_S, E_U, E_A|H_d, I)}$$

Bayes Factor

$$= \frac{P(E_U, E_S|H_p, I)}{P(E_S|H_d, I)} \frac{P(E_A|H_p, I)}{P(E_S, E_U, E_A|H_d, I)}$$

$$= \frac{P(E_U|E_S, H_p, I)}{P(E_U|E_A, H_d, I)}$$
Theoretical Example

It is not a straightforward task to statistically interpret forensic propositions.

In the traditional setting we have-

\[ LR = \frac{f(x, y|H_p, I)}{f(x, y|H_d, I)} = \frac{f(y|x, H_p, I)}{f(y|H_d, I)} \]

Compared with the SLR-

\[ SLR = \frac{g(\Delta(x, y)|H_p, I)}{g(\Delta(x, y)|H_d, I)} = \int \frac{g(\Delta(x, y)|x, H_p, I)f(x|H_p, I)dx}{\int g(\Delta(x, y)|x, H_d, I)f(x|H_d, I)dx'} \]