Introduction to Forensic Source ID problems

Christopher Saunders
Mathematics and Statistics Department
South Dakota State University

Outline

Background

Most of the work I have done in forensic science has been related to impression and pattern evidence with my main focus on forensic handwriting. These activities have included:

- Building a closed set identification system to suggest the ID of the writer of a given note.
  - Also applications in ink, fingerprints, explosives, body odor etc.
  - These are the most common problems I am asked to work on.
- Developing the statistics for estimating the random match probabilities.
  - fingerprint and handwriting individuality studies.
- Developing methods to construct approximate values of the evidence for complex evidential forms, such as handwriting evidence.
  - Mainly counter examples with fingerprints and handwriting.
- Extending Bayesian evidence interpretation approaches to situations where we have incomplete data about the background population.
  - Specifically, trace evidence such as glass fragments and copper wire samples.

The Forensic Identification of Source Problem

....as I understand it.

> In forensic science a common problem
> *is the identification of the source for an object with an unknown source.*

Examples:
- ► Who wrote this bank robbery note?
- ► Are these glass fragments found on the suspect from the same broken window?
- ► Who is the source of this blood stain?
- ► Were these bullets fired from a common gun?
- ► Is the suspect the source of this latent print?
- ► etc.

Exact Nature of the Question?

In general, when working on these problems, I tend to become concerned with what is being asked by the consumer of my statistics.

With respect to handwriting, some of the questions I have been asked are as follows:

> Q1: In this list of 100 writers, who wrote this short note?
>
> Q2: Did this specific writer write this note?
>
> Q3: Were these two handwritten notes from a common writing profile?
>
> Q4: Did one of these 100 writers write this short note and, if yes, who is the writer?

For those that favor likelihood ratios; there is a different LR for each of these questions... [1][2]

---

[1]This may have to do with the people looking for a general solution that a single statistical approach can solve.

[2]Our groups current NIJ grant is focused on statistical issues associated with Q2 and Q3 for high dimensional quantifications of impression and pattern evidence.

Q1: Closed Set Identification

Closed Set Identification: In this list of 16 windows, which is the source of these glass fragments found on the suspect?

Methods: Statistical Pattern Recognition- Estimation of Bayes Classification Rules.

Interpretation: These methods usually return a short list with a score.

- ▶ In an ideal setting, a special type of LR is used to determine whether or not the source of the trace is on the watch list.

Error: Is the actual window that is the source of the fragments on the short list?

Evidence: Samples collected from each of the windows on the watch list and the object to be classified.

Models: Each window has their own probability model for how they generate evidence (fragments).

These problems and solutions rely on traditional statistical pattern recognition methods dating back a century or so. The modern statistical methods are commonly developed at the intersection of computer science, statistics, and signal processing.

Q2: Specific Source

Are these fragments from this specific window?

Methods: Statistical inference and model selection based approaches.

Interpretation: Depends on the method-

- ▶ Two- Stage Approach-"Can not exclude" and conditional match probability.
- ▶ Bayesian Approach-Value of Evidence.

Error: The glass fragments are implied to have come from the specific window when it has not.

Evidence: The fragments with an unknown source, samples from the specific window, and a sample of sources from a population.

Models: At least two probability models are needed:

1. How the specific window generates evidence.
2. How the alternative source population generates evidence.

Q3: Common Source

Do these two sets of glass fragments have a common source?

Methods: Statistical modeling and various methods from biometric verification applications.

Interpretation: Varies, but usually uses Match/Non-Match language.

- ▶ The match criteria can be expressed in terms of a specialized LR.
- ▶ Receiver Operating Characteristic or Detection Error Tradeoff Curves sometimes Tippet plots.

Error: Two sets of glass fragments are implied to have a common source when they do not.

Evidence: The two sets of glass fragments to be compared and a sample of sources from a population for which we wish to control the error rate.

Models: At least two probability models are needed:

1. Two samples from a common selected source.
2. Two samples each from separately selected sources.

Q4: Open Set Identification

Is one of these known windows the source of this set of window fragments? If yes, which known window?

Methods: To the best of my knowledge, this is an open research area in forensic statistics and biometrics!

Interpretation: This is an open research area! The problems arise in needing to include base rates of the known sources with prior beliefs necessary to work within a rigorous Bayesian paradigm.

Error:
- ▶ Mistakenly concluding that the window that is the source of the fragments is on the watch list.
- ▶ Mistakenly concluding that the window that is the source of the fragments is NOT on the watch list.
- ▶ Correctly concluding that the window that is the source of the fragments is on the watch list, but not including the correct window on the shortlist.

Evidence: The fragments to be assigned a window, the templates from the watch list windows, and a sample of sources from a population for which we wish to control the error rate.

Models: At least two probability models are needed:
1. How each of the specific watch list sources generate evidence.
2. How alternative source population generates evidence.

Why is this distinction important?

Each of these questions can have radically different answers, even when the exact same evidence is used to answer each of the questions in turn....

This issue is due to the following interrelated reasons:

► Methods to solve the problem in an optimal manner
► The interpretation/presentation of the results of the identification process
► The definition of an error
► The evidence that is used to answer the question
► The probability models used to characterize the evidence

In my experience, it very common take a summary statistic used to answer one of the questions (usually a Bayes Factor for Q2) and use the resulting statistic as an answer to each of the other questions.