Forensic Science International 219 (2012) 129–140

Contents lists available at SciVerse ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Score-based likelihood ratios for handwriting evidence

Amanda B. Hepler [a,*], Christopher P. Saunders [a], Linda J. Davis [b], JoAnn Buscaglia [c]

[a] Document Forensics Laboratory (MS 1G8), George Mason University, Fairfax, VA 22030, USA
[b] Department of Statistics (MS 4A7), George Mason University, Fairfax, VA, 22030, USA
[c] FBI Laboratory, Counterterrorism & Forensic Science Research Unit, Quantico, VA 22135, USA

ARTICLE INFO

Article history:
Received 12 May 2011
Received in revised form 7 December 2011
Accepted 19 December 2011
Available online 31 January 2012

Keywords:
Forensic science
Likelihood ratio
Handwriting evidence
Statistical evidence evaluation
Forensic statistics
Questioned documents

ABSTRACT

Score-based approaches for computing forensic likelihood ratios are becoming more prevalent in the forensic literature. When two items of evidential value are entangled via a scorefunction, several nuances arise when attempting to model the score behavior under the competing source-level propositions. Specific assumptions must be made in order to appropriately model the numerator and denominator probability distributions. This process is fairly straightforward for the numerator of the score-based likelihood ratio, entailing the generation of a database of scores obtained by pairing items of evidence from the same source. However, this process presents ambiguities for the denominator database generation – in particular, how best to generate a database of scores between two items of different sources.

Many alternatives have appeared in the literature, three of which we will consider in detail. They differ in their approach to generating denominator databases, by pairing (1) the item of known source with randomly selected items from a relevant database; (2) the item of unknown source with randomly generated items from a relevant database; or (3) two randomly generated items. When the two items differ in type, perhaps one having higher information content, these three alternatives can produce very different denominator databases. While each of these alternatives has appeared in the literature, the decision of how to generate the denominator database is often made without calling attention to the subjective nature of this process.

In this paper, we compare each of the three methods (and the resulting score-based likelihood ratios), which can be thought of as three distinct interpretations of the denominator proposition. Our goal in performing these comparisons is to illustrate the effect that subtle modifications of these propositions can have on inferences drawn from the evidence evaluation procedure. The study was performed using a data set composed of cursive writing samples from over 400 writers. We found that, when provided with the same two items of evidence, the three methods often would lead to differing conclusions (with rates of disagreement ranging from 0.005 to 0.48). Rates of misleading evidence and Tippet plots are both used to characterize the range of behavior for the methods over varying sized questioned documents. The appendix shows that the three score-based likelihood ratios are theoretically very different not only from each other, but also from the likelihood ratio, and as a consequence each display drastically different behavior.

1. Introduction

The likelihood ratio paradigm has been proposed as a means for quantifying the strength of evidence for a variety of forensic evidence, including handwriting, speech, earmarks, glass fragments, fingerprints, footwear marks and DNA [1–9]. A body of evidence can be evaluated by calculating the likelihood ratio, which compares the probability of the "evidence" under two competing propositions (or hypotheses), often denoted as the prosecution proposition (H_p) and the defense proposition (H_d). Consider the scenario where two items of evidence are found over the course of a forensic investigation, and the following source-level hypotheses are of interest:

H_p: The two items came from the same source,
H_d: The two items came from different sources.

* Corresponding author. Present address: Innovative Decisions, Inc., 1945 Old Gallows Rd., Suite 207, Vienna, VA 22182, USA. Tel.: +1 919 610 2942; fax: +1 815 550 1617.
E-mail address: abhepler@innovativedecisions.com (A.B. Hepler).

0379-0738/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2011.12.009

Let x denote[1] a measurement obtained from the *source*, or the sample with a known source (e.g., suspect's known writing samples, crime scene window). Let y denote a measurement obtained from the *trace*, or the sample with an unknown source (e.g., bank robbery note, glass fragment obtained from the suspect). If one assumes that x and y are realizations from continuous random variables X and Y, the likelihood ratio is defined by

$$\mathrm{LR} \equiv \frac{f(x,y|H_p,I)}{f(x,y|H_d,I)},$$

where I represents background information, and f denotes the probability distribution associated with the random variables X and Y. When x and y are discrete measurements, f is a probability; when x and y are continuous measurements, f is a continuous probability density function. As stated in [10], the numerator and denominator densities might be very different due to the differing conditioning arguments, but it is common practice to allow the generic symbol f to represent both functions.

In many cases (e.g., when the evidence is represented using a high-dimensional quantification technique [11]), the numerator and denominator of LR are not obtainable directly, without making (perhaps) unfounded assumptions about the underlying processes that generate the evidence [12]. A promising surrogate, which can be applied to virtually any evidence type, is a score-based approach [10,12–18].

In this article, we critically examine three methods appearing in the literature for estimating the score-based likelihood ratio (SLR) in the specific context of natural handwriting evidence. While our illustrations focus on this modality, the concepts apply broadly to the application of these methodologies to any type of evidence for which a meaningful paired score can be defined.

Each methodology makes very different assumptions about the nature of the random variables X and Y, specifically in the denominator (under the defense's proposition). Often these are listed either as assumptions [18], an (often unstated) byproduct of database generation [10,12,16]. The intent of this paper is to illuminate, for both the statistical and non-statistical audience, the assumptions underlying the three different methodologies and how they are in fact subtle changes to the interpretation of H_d. The hope is that once these interpretations are laid bare, the forensic community can then appropriately weigh their merits and applicability. This is particularly important since, as shown in Section 5 and in Appendix A, the three methods can yield drastically different results when given the exact same evidence. It is our belief that these three score-based methods cannot gain mainstream acceptance until this denominator specification problem is resolved by the forensic community.

The outline for this paper is as follows. Section 2 presents each method in a unified notation, while making explicit each of the underlying assumptions and their associated H_d interpretation. Section 3 briefly details the quantification technique used to quantify handwritten documents (more detailed descriptions appear elsewhere [13,19]). Also, Section 3 details the algorithms used to obtain estimates for each SLR, denoted throughout as SLR_1, SLR_2 and SLR_3. Finally, Sections 4 and 5 detail the design and results of a comparison study showing the impact that selecting one SLR over another (i.e., one set of assumptions over another) has on the estimated SLRs.

[1] Throughout this manuscript, the following conventions are used: uppercase bold letters denote *random* matrices or vectors; lowercase bold letters denote *observed or known* matrices or vectors; lowercase letters denote *observed or known* scalars.

2. Score-based likelihood ratios

For many types of forensic evidence, obtaining the likelihood ratio, as defined above, has proved difficult, if not impossible [12]. For some types of evidence, it is rare that the underlying process which generates X and Y is sufficiently understood to make the assumption that the distribution is an element of some common family of distributions. For example, with certain quantifications of the elemental composition of glass fragments it is not necessarily reasonable to make the blanket assumption that X and Y follow a normal distribution [20], as is often done [21]. Even for the most basic forms of DNA evidence there were many years of research and academic discussions before reasonable distributional assumptions were known to an adequate degree of certainty that might be required in a court of law [8]. Even if the distribution is known or can reasonably be assumed, true parameter values are rarely known and are likely difficult to estimate for the more complex quantifications of the evidence. When x and y represent high dimensional measurements, as would be the case if one considers the multifaceted attributes that make up one's full body of handwriting (or *writing profile*), the problem is exacerbated as now we are faced with (1) how to probabilistically characterize each attribute individually and (2) how to capture probabilistic dependencies sure to exist among the attributes.

Score-based approaches seem able to overcome at least some of these challenges. If one can capture similarities or differences between two items via a univariate score function that illuminates as to whether or not the items have a common source, then the dimensionality of the problem is greatly reduced [12,15,16]. Determining (or estimating) the probability distribution of this score function remains a challenge however, as will be highlighted throughout the remaining sections of this paper.

A brief introduction to score-based likelihood ratios is provided here. A more detailed discussion, within the context of handwritten documents can be found in [13]. Let the function which assesses the dissimilarity between x and y be denoted by $\Delta(x, y)$. The score-based likelihood can then be described as a proxy of sorts to the LR,

$$\mathrm{LR} = \frac{f(x,y|H_p,I)}{f(x,y|H_d,I)} \approx \frac{g(\Delta(x,y)|H_p,I)}{g(\Delta(x,y)|H_d,I)}, \quad (1)$$

where g denotes the probability distribution associated with the random variable $\Delta(X, Y)$. Often in the literature the rightmost quantity is also denoted by LR [10,12]. In the interest of transparency and clarity, in this work this quantity is denoted by SLR. Another impetus to keep these quantities distinct is that, as noted by [19], the suitability of the approximation $\mathrm{LR} \approx \mathrm{SLR}$ has not been investigated thoroughly. It is shown in Appendix A for a simplified scenario (where the probability distributions of X, Y, and $\Delta(X, Y)$ are all known) that the three SLRs under consideration here often do not well approximate the LR.

The numerator of the leftmost expression in Eq. (1) can be interpreted in layman's terms as: the likelihood of observing these two measurements if the items come from the same source. Similarly the denominator can be interpreted as the likelihood of observing these two measurements if the items come from different sources. In order to compute this quantity, statisticians typically make the assumptions (1) the marginal distribution of X is independent of whether or not H_p or H_d is true, and (2) measurements on X and Y are independent if H_d is true. Under assumptions (1) and (2), the LR reduces to:

$$\mathrm{LR} = \frac{f(x,y|H_p,I)}{f(x,y|H_d,I)} = \frac{f(y|x,H_p,I)\,f(x|H_p,I)}{f(y|H_d,I)\,f(x|H_d,I)} = \frac{f(y|x,H_p,I)}{f(y|H_d,I)}. \quad (2)$$

The simplification achieved in Eq. (2) is what drives all DNA likelihood ratio calculations, and most non-score based approaches [8,22]. Unfortunately, an analogous development for the SLR (right side of Eq. (1)) is not possible since measurements from the trace and the known source are now tied together via the score function and cannot be disentangled. Conditioning on x is of no use here, since:

$$\mathrm{SLR} = \frac{g(\Delta(x,y)|H_p,I)}{g(\Delta(x,y)|H_d,I)} = \frac{\int g(\Delta(x,y)|x,H_p,I)\, f(x|H_p,I)dx}{\int g(\Delta(x,y)|x,H_d,I)\, f(x|H_d,I)dx}, \quad (3)$$

and, in general, Eq. (3) cannot be simplified in a straightforward manner, if at all. The simplifications leading to Eq. (2) no longer hold – the conditioning on x must remain in the denominator, and the marginal distribution of x no longer cancels out as it appears inside separate integrals in the numerator and denominator.

Despite the fact that the SLR cannot be simplified in any meaningful way to facilitate computation, several score-based methods have emerged in the literature. Many make, either explicitly or implicitly, simplifying assumptions in order to estimate the SLR. The body of literature here is growing, and we restrict our attention to three such methods which serves as a continuation of our work in [13]. Each SLR method makes use of a similar numerator estimation technique previously reviewed in [13], while differing in their approach to estimating the denominator.

The numerator of the simplified LR appearing in Eq. (2) can be interpreted in layman's terms as: the likelihood of observing the trace measurement if it came from the known source. The denominator can be interpreted as: the likelihood of observing the trace measurement if it came from a different source. To compute the denominator directly, an additional assumption must be made regarding the alternate source. The most common, often referred to as the "random man" assumption, is that the source of the trace is randomly selected from some "relevant population" of sources [22]. This leads to the following *statistical interpretation* of the denominator: the likelihood that the trace measurement came from a random source in a relevant population.

The interpretation of the numerator for the SLR is slightly different from that of the LR: the likelihood of observing *this score between the trace and the known source* if they came from the same source. The interpretation of the denominator is: the likelihood of observing *this score between the trace and the known source* if they came from different sources. When one tries to be more specific about the denominator in order to obtain probability distributions, ambiguity arises. As above, some notion of "random source" must come in, but there is subjectivity in how to proceed. Distinct interpretations of H_d motivate the three SLRs under consideration in this paper. The first method contends that the known source is a random selection from the relevant population; the second contends that the source of the trace is a random selection from the relevant population; and the third contends that both the trace and the known source are randomly selected from the relevant population.

2.1. Score-based numerator

All three methods we consider here have considered the following interpretation of the SLR numerator: the likelihood of observing this score if the known source measurement is paired with measurements taken from traces randomly drawn from the known source population. The new specification of the hypothesis being entertained is:

> H_p: $\Delta(x, y)$ arises from the distribution of scores obtained by pairing x with a randomly generated y, where both x and y arise from the same distribution.

While this hypothesis is not necessarily reasonable from the perspective of a prosecution attorney, it is in fact the hypothesis under consideration when one reports one of the three SLRs in court. For clarity, we will refer to the type of proposition which fully specifies the desired probability distribution as a *statistical proposition*, whereas *forensic propositions* refer to those of direct interest to the courts. We prefer this approach over relegating these specifications to the background information or enumerating them as assumptions because we feel those approaches lack transparency and/or clarity, particularly for non-statisticians.

This new specification introduces conditioning upon x the numerator of the SLR, that is $g(\Delta(x,y)|H_p,I) \approx g(\Delta(x,y)|x,H_p,I)$. From Eq. (3) it is clear that this is indeed an approximation. The impact this type of approximation has on the resultant score-based likelihood ratios is investigated in Appendix A for a simplified scenario where all distributions are known.

2.2. SLR_1: trace-anchored

Some researchers [14–16] have considered the following interpretation of the SLR denominator: the likelihood of observing this score if *the* trace measurement is paired with measurements taken from random sources in some relevant population. The statistical proposition being entertained is:

> H_{d1}: $\Delta(x, y)$ arises from the distribution of scores obtained by pairing y with a randomly selected x from the relevant population.

This new interpretation of the denominator of the SLR actually changes the specification of the SLR denominator, $g(\Delta(x,y)|H_d,I) \approx g(\Delta(x,y)|y,H_d,I)$. Noting that conditioning on y in Eq. (3) (rather than x) would also not lead to any simplification, it is clear that these two quantities are not in fact equal. Using this approximation, the first score-based likelihood ratio under consideration is

$$\mathrm{SLR}_1 = \frac{g(\Delta(x,y)|x,H_p,I)}{g(\Delta(x,y)|y,H_d,I)}.$$

Whether or not SLR_1 serves as a reasonable proxy for LR is an open question. The example in Appendix A is aimed at informing this debate.

One issue with conditioning on y in the denominator is that it is asymmetric, in the sense that the numerator and denominator are conditioning on different quantities. Another conceptual issue with SLR_1 is that, in the case of glass evidence (or any type of evidence where the item of unknown source is taken from the suspect), the conditioning in the denominator is on measurements taken from the suspect. Specific properties of the crime scene window are ignored entirely, and it is therefore less informative than if those characteristics had been accounted for [19]. However, in the case of handwriting this type of conditioning seems more plausible, as specific properties of the bank robbery note are informing the denominator probability distribution.

One also might consider the recommended conditioning rules provided in [23]. They advocate conditioning on the sample with greater *information content*, which in the case of handwriting would be y (the suspect's known writing samples). However, for glass the desired conditioning would be x (the window at the scene) which again leads to ambiguous notions of the "correct" conditioning. It should be noted that in [23] this conditioning strategy was aimed at simplifying the computation (much like the arguments in Eq. (2)). This computational advantage is lost for the SLR, as illustrated above in Eq. (3).

2.3. SLR_2: source-anchored

Others [10] have proceeded with following interpretation of the SLR denominator: the likelihood of observing this score if trace measurements taken from randomly selected sources from a relevant population are paired with *the* measurement taken from the known source. This is somewhat analogous to the LR denominator interpretation in that the trace measurement now comes from a random source. This interpretation again changes the specification of the SLR denominator:

$$\mathrm{SLR}_2 = \frac{g(\Delta(x,y)|x,H_p,I)}{g(\Delta(x,y)|x,H_d,I)}.$$

We now have symmetric conditioning, on x in both the numerator and denominator. Again, whether or not this quantity serves as reasonable proxy for LR is considered in Appendix A. This development leads to the following denominator proposition:

> H_{d2}: $\Delta(x, y)$ arises from the distribution of scores obtained by pairing x with a randomly selected y from the relevant population.

This approach succumbs to some of the same criticisms as SLR_1, but in reverse: for glass evidence this conditioning seems reasonable in that specific characteristics of the crime scene evidence are directly relevant to the denominator distribution, but for handwriting specific characteristics of the bank robbery note are ignored (e.g., the writer used all capital letters).

2.4. SLR_3: general match

The final SLR approach considered in this work, often used in biometrics [24], applies the following interpretation of the denominator: the likelihood of observing this score if a trace measurement taken from a randomly selected source from a relevant population is paired with a measurement taken from a different source randomly selected from a relevant population. This makes no changes to the SLR denominator:

$$\mathrm{SLR}_3 = \frac{g(\Delta(x,y)|x,H_p,I)}{g(\Delta(x,y)|H_d,I)}.$$

Again, whether or not this serves as a reasonable proxy for LR is considered in Appendix A. This development leads to the following denominator proposition:

> H_{d3}: $\Delta(x, y)$ arises from the distribution of scores obtained by pairing a randomly selected X from the relevant population with a randomly selected Y from that same relevant population.

This SLR is far less informative in that the denominator distribution depends neither on specific characteristics trace nor on characteristics of the known source [19]. That is, the denominator distribution would remain unchanged if a different trace were observed, or if a different known source is considered.

The next section of the paper shows how to generate each SLR for a specific quantification of handwritten documents.

3. Estimating SLRs in handwriting

Handwriting-specific definitions of the evidence (following the notation introduced in [13]) are as follows: E_S denotes a collection of writings known to have originated from the suspect (henceforth *suspect's template*) and x represents some quantification obtained from those writings. E_U denotes a handwritten questioned document (QD) found at the scene of unknown source, and y represents some quantification obtained from that document. E_A denotes a collection of writing samples taken from alternative sources.

3.1. Handwriting quantification

Selection of an appropriate score will depend heavily on the numeric representation, or quantification technique used to describe a handwritten document. The quantification method used here, developed by Gannon Technologies Group, first scans and skeletonizes the document, which has been manually parsed into characters, as shown for the word "London" in Fig. 1. Subsequent to this segmentation, a proprietary, automated process was used to represent each parsed character's skeleton by an isomorphic class of graphs (a geometric form that remains invariant under certain transformations, e.g., bending or stretching), referred to as an *isocode*. Details of this process are described at length elsewhere [13,25] however a schematic depicting the method appears in Fig. 1.

Define a *writing profile* as the entire body of writing that a writer has written or will ever write. Define a writer's *template* as a collection of writing samples from an individual assumed to be sufficiently rich for characterizing an individual's writing profile. Using this quantification method, E_S is reduced to the matrix of counts computed by combining counts over a large collection of known writing samples obtained from a suspect (*suspect's template*), represented by the random variable $\boldsymbol{X}$. E_U is reduced to the matrix of counts computed from a questioned document, represented by the random variable $\boldsymbol{Y}$.

3.2. Estimating the SLR

3.2.1. Dissimilarity score

We first define a dissimilarity statistic (or score) that can be computed for two documents (or collection of documents). We selected the Kullback–Leibler divergence [26] to capture the difference between the observed matrices of counts for two writing samples, row by row (i.e., letter by letter). These divergences are combined over letters using a weighted average, ensuring that frequently observed letters (across both documents) contribute more to the dissimilarity score. Details appear in Appendix B.

At this point, it is important to emphasize that the procedure that follows does not depend on the selection of this particular score. A multitude of scoring methods can be used in its place (e.g., see [13] for a similar analysis using a similarity score based on Pearson's chi-squared statistic).

3.2.2. Database generation

To estimate the numerator and denominator densities of the SLRs we need to obtain databases of scores generated in several ways. For the numerator, we need a database of scores where both x and y were obtained from documents written by the suspect. This is a fairly straightforward matter in our case, and the reader is referred to [13] for specific details. For the denominator, we need a database of scores where x and y are generated from different sources, according to the conditioning assumptions of SLR_1, SLR_2, and SLR_3.

3.2.2.1. Numerator database. Ideally, a database would exist consisting of scores obtained by comparing "QD-like" documents to the suspect's template. It is unreasonable to expect a large number of "QD-like" documents to be discovered over the course of the investigation. For example, if the QD is a bank robbery note, only in extremely rare cases would, a priori, a collection of such bank robbery notes exist. One might suggest requesting the

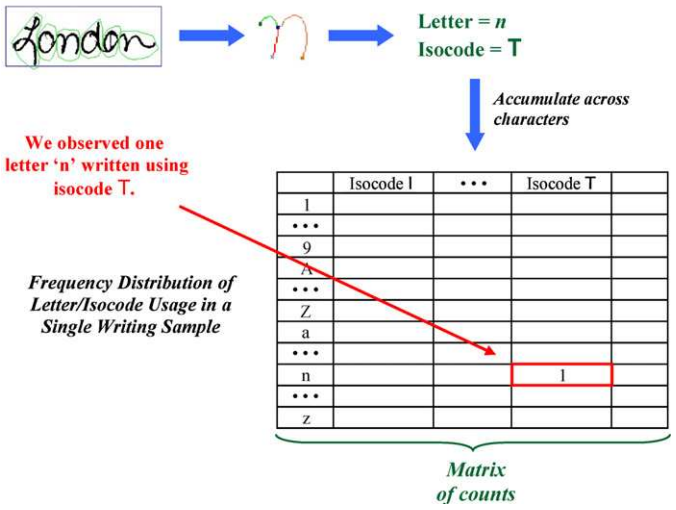

Fig. 1. Schematic of the quantification process.

generation of a collection of "QD-like" documents from the suspect; however, this might not result in the most representative sample, especially in cases where the suspect is indeed the culprit, as there is motivation to disguise his or her writing style. In addition, the number of samples needed to accurately estimate the distribution of scores would be prohibitive.

In light of these challenges, [25] proposed a method of obtaining an arbitrarily large database (size denoted by N) of 'within' scores using a subsampling algorithm. Noting that n_U represents the number of characters in QD and n denotes the total number of characters in the suspect's template, the details of the slightly modified[2] algorithm employed appears below:

Subsampling Algorithm for Generating Numerator Database

For $i = 1, \ldots, N$, where N denotes a sufficiently large number of iterations,

1. Randomly divide the suspect's template into two subsets, with character counts n_U and $n - n_U$ respectively. This is done by randomly selecting a (starting) character from the first $n - n_U$ characters. The selected character along with the next $n_U - 1$ characters is defined as the *pseudo-QD* and from it we obtain the matrix of counts y_i. The remaining characters form a *pseudo-template*, from which we obtain the matrix of counts x_i.
2. Compare the two simulated writing samples, recording the resultant score: $\Delta(x_i, y_i)$.

3.2.2.2. Denominator databases. Before the detailed algorithms are presented, we first must address the challenge of obtaining a representative collection of writing templates from potential alternate sources. Recall, above we denoted this collection of templates by E_A as it is considered part of the evidence collected which may differ from case to case and which, especially when E_A is of limited size, will have a significant impact on the estimation of the score-based likelihood ratio. We make the simplifying assumption that a large, representative collection of templates exists. In future work, we intend to examine more practical scenarios, and investigate the impact typical violations of these assumptions have on the estimation procedure.

Once a large, representative collection of templates E_A is established, the mechanics of generating between scores for each of the three denominator SLR interpretations can be detailed.

SLR$_1$: The trace-anchored interpretation of H_d, tailored to handwriting evidence, is "the evidence score arises from the distribution of scores obtained by pairing *the* QD with *a* template written by a random individual." A detailed illustration of an adaptation of this method for the analysis of handwriting evidence can be found in [13]. The specific algorithm appears below.

Trace-anchored Algorithm for Generating Denominator Database

Obtain a matrix of counts from the QD, denoted by y_U. Then, for $i = 1, \ldots, N_A$, where N_A represents the number of writers in E_A,

1. Select the ith writer from E_A and obtain a matrix of counts from that individual's template, denoted by x_i.
2. Compare the two writing samples, recording the resultant score: $\Delta(x_i, y_U)$.

SLR$_2$: The source-anchored interpretation is "the evidence score arises from the distribution of scores obtained by pairing *a* QD written by a random individual with *the* template written by the suspect." The specific algorithm appears below.

Source-anchored Algorithm for Generating Denominator Database

Obtain a matrix of counts from the suspect's template, denoted by x_S. Then, for $i = 1, \ldots, N$, where N denotes a sufficiently large number of iterations,

1. Select a writer from E_A, and randomly select n_U characters to serve as the pseudo-QD. Obtain the matrix of counts, denoted by y_i.
2. Compare the two writing samples, recording the resultant score: $\Delta(x_S, y_i)$.

It should be noted that while [10] does hold x_S fixed, they do not proceed with their database generation in exactly the same manner. They introduce an extra layer of complexity by generating (what would be the equivalent of) multiple pseudo-QDs from every writer in E_A in order to generate N_A different writer-specific databases. Here, due to computational constraints, only one pseudo-QD is generated per writer.

SLR$_3$: The final interpretation considered, which avoids anchoring all together, is "the evidence score arises from the distribution of scores obtained by pairing *a* QD written by a

[2] In [25], a random selection of n_U characters was chosen, whereas here n_U consecutive characters were chosen. We feel that the use of consecutive characters best aligns with the natural writing that might appear in a QD.

random individual with *a* template written by a different random individual." The specific algorithm appears below.

General Match Algorithm for Generating Denominator Database

For $i = 1, \ldots, N$, where N denotes a sufficiently large number of iterations,

1. Randomly select writer 1 from E_A and randomly select a document of size n_U from his/her template to obtain a pseudo-QD. Obtain the matrix of counts, denoted by y_i.
2. Randomly select writer 2 (distinct from writer 1) from E_A, and obtain a matrix of counts from his/her template, denoted by y_i.
3. Compare the two writing samples, recording the resultant score: $\Delta(x_S, y_i)$.

3.2.3. Distribution estimation

Assuming one of the three denominator algorithms is selected, two collections of scores have been obtained, one under the prosecution's hypothesis and one under the selected defense hypothesis. The probability densities of those scores are rarely known exactly and must be estimated. Denote those estimated densities by $\hat{g}$. Normal probability plots of the "numerator scores" and the three sets of "denominator scores" indicated a normal approximation was reasonable (results not shown). After obtaining the sample mean and variance of our N (or N_A for the trace-anchored approach) generated observations, $\hat{g}$ is defined to be a normal distribution centered at the sample mean, with variance equal to the sample variance estimate. Other methods were considered (e.g. kernel density estimation, as employed in [13] and histogram estimators) but both methods have been shown to poorly model the tail behavior, leading to unwarranted extreme values for the estimated SLR, denoted by $\widehat{SLR}$, both when H_p is true and when H_d is true. The true distributions of scores appear to have light left tails and heavy right tails. Thus, the normal approximation seems the choice of least harm, as it tends to arrive at conservative[3] estimates for $\widehat{SLR}$. Again, it is important to emphasize that the procedure which follows does not depend on the selection of this particular estimation technique for the probability distribution of the scores.

3.2.4. Computing $\widehat{SLR}$

The evidence score, δ, is obtained by comparing the actual QD (specifically the observed matrix of counts denoted by y_U), with the suspect's template (specifically the observed matrix of counts denoted by x_S), using the modified Kullback–Leibler divergence as detailed in Appendix B, $\Delta(x_S, y_U) = \delta$. The final step is to evaluate the estimated distributions at that score: $\hat{g}(\delta|H_p, I)$ and the correct corresponding denominator, $\hat{g}(\delta|x, H_{d1}, I)$, $\hat{g}(\delta|y, H_{d2}, I)$, or $\hat{g}(\delta|H_{d3}, I)$, and then taking their ratio to obtain the estimated score-based likelihood ratio, $\widehat{SLR}$. The next section illustrates that, as expected from the results shown in Appendix A, very different results are obtained for each method.

4. Comparative study

In summary, three methods have been presented for obtaining denominator databases used to estimate the SLR: trace-anchored, source-anchored, and general match. These three databases will necessarily result in three different estimates of SLR, denoted[4] by SLR_1, SLR_2, and SLR_3. It seems prudent to investigate whether or not, given the exact same evidence, the three estimates would differ substantially. To that end, a comparative study was performed.

4.1. Writing samples

The set of writing samples used in the comparative study are those described in detail in [25], collected by the FBI Laboratory over a two-year period. Samples were collected from about 500 different writers. Each writer was asked to provide 10 samples (5 in print and 5 in cursive) of a modified "London Letter" [27] paragraph (533 characters long). In this study, only writing samples in which the writer submitted all five cursive paragraphs were included. This restriction results in 424 writers for a total of 2120 London Letter paragraph writing samples.

4.2. Simulation design

We performed the following simulation:

1. Randomly select two of the 424 writers, denoted by w_1 and w_2. Define E_A to be the remaining 422 writers in the database.
2. Obtain SLR_1, SLR_2, and SLR_3 for two scenarios.
   - **H_p True: The suspect is the culprit**[5] (w_1 = suspect = culprit). One of the five paragraphs written by w_1 is randomly selected, from which a string of size n_U is randomly extracted to serve as QD. We varied n_U to be 20, 40, 60, 80, 100, and 150. The number of scores, N, generated to estimate the numerator distribution was set to 500.
   - **H_d True: The suspect is not the culprit** (w_1 = suspect, w_2 = culprit). QD is obtained in the same manner as the first scenario, except taken from w_2's template rather than w_1's. The number of scores, N, generated to estimate the denominator distribution for SLR_2 and SLR_3 was set to 500.

Repeat steps 1 & 2, 200 times, a computationally feasible number of repetitions.

5. Results and discussion

The estimates obtained for the three methods were highly variable. To illustrate, for one iteration of the above simulation where H_p is true, values obtained were $SLR_1 = 1858$, $SLR_2 = 1701$, $SLR_3 = 15$. Another iteration resulted in the values $SLR_1 = 2370$, $SLR_2 = 6$, $SLR_3 = 19$.

This trend continues over many runs, which are summarized for the H_p true scenario in Table 1. To facilitate the discussion, we arbitrarily assigned a cutoff so that any SLR estimate greater than 100 leads to the conclusion "supports H_p".[6] Similarly for any SLR estimate less than 1/100, we conclude "supports H_d". Finally, for any intermediate values, no conclusion is reached. Results are presented in Table 1. For a QD with 80 characters, we observed a high rate (0.43) of disagreement among the three methods. That is, 43% of the time at least one of the three methods disagreed with the others as to whether or not the evidence supports H_p, supports H_d, or is inconclusive.

Disagreement rates generally increase as the QD gets larger, an indication that most of the agreement that does occur for smaller QDs is due to the majority of values falling in the inconclusive

[3] Conservative in the sense that it protects against Type I errors (errs on the side of innocence) as the estimated SLRs tend to be smaller than the true SLRs.

[4] The 'hat' notation is suppressed for ease of presentation; however the reader should be mindful that these are estimates of the true values of SLR_1, SLR_2, and SLR_3.

[5] Throughout, *culprit* refers to the individual who actually wrote the QD.

[6] The authors are not implying that this is, in any way, a *meaningful* cutoff.

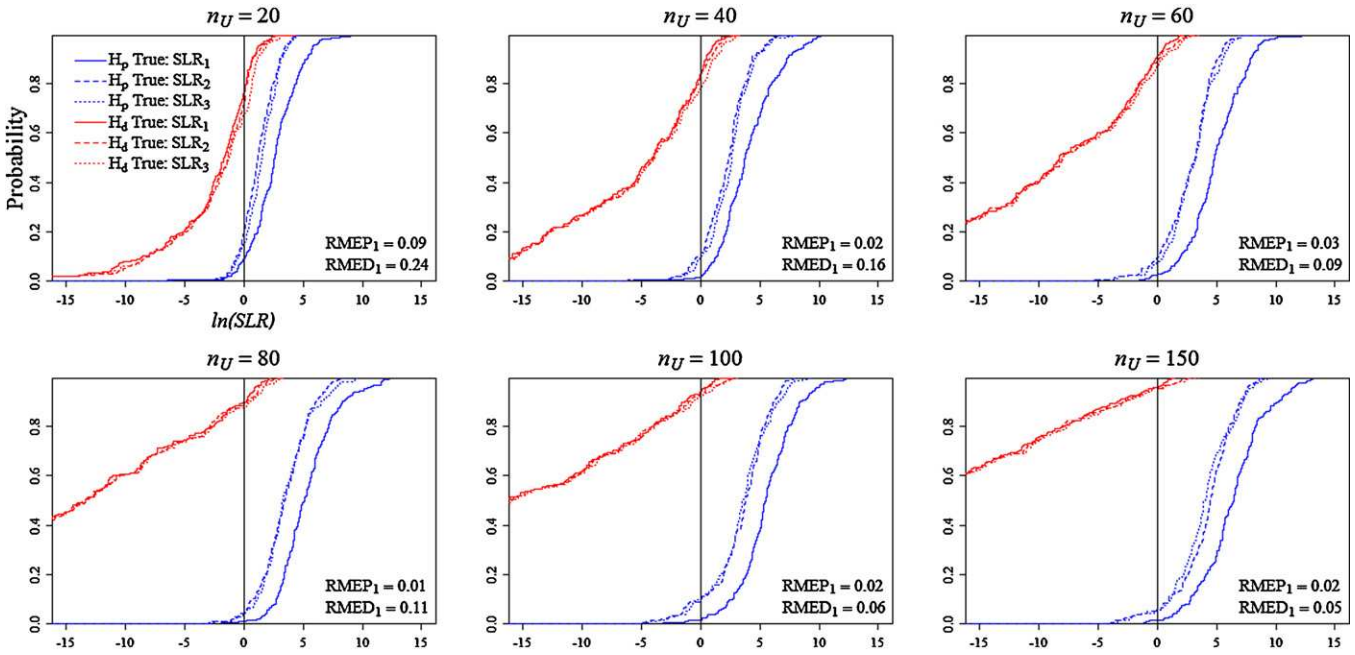

Fig. 2. Tippet Plots for three SLR approaches, under two scenarios: H_p true (black lines) or H_d true (grey lines). Rates of misleading evidence are reported for SLR_1, the method exhibiting the smallest rates.

range. More agreement occurs when H_d is true as seen in Table 2, although there is still some disagreement (3% for $n_U = 80$). From the results in Table 1, it is clear the methods are differing substantially in terms of the conclusions one would draw in cases where H_p is true, and a more detailed analysis of the results is warranted.

Tippet plots (following the conventions described in [28]) are shown for all three SLRs, on the natural log scale, in Fig. 2. The three methods can be compared by the measurement of two "error rates"[7] as described in [12]: RMEP ≡ the rate of misleading evidence in favor of the prosecution, i.e., when H_p is true ($|\ln(SLR)| < 0$) and RMED ≡ the rate of misleading evidence in favor of the defense, i.e., when H_d is true ($|\ln(SLR)| > 0$).

Before proceeding, the reader is reminded that samples were obtained by convenience, all consisting of the exact same cursive text, and under particularly mundane circumstances. These facts most certainly prohibit generalization of results. In addition, the reader must be mindful that selection of a different score or a different distribution estimation technique may lead to very different performances of the three methods. The authors are currently investigating the robustness of the three approaches to alternate scoring and estimation methods.

For each scenario considered, the rates of misleading evidence for SLR_1 were far lower than the other two methods. A full listing of the error rates appears in Table 3. Rates for SLR_2 and SLR_3 are nearly indistinguishable. As expected, the rates decrease as the size of QD increases.

The reporting of this type of error rate is less than ideal, as the possibility of an inconclusive determination is fully ignored. An approach that is more representative of the realities of forensic casework is to impose symmetric cutoffs (e.g., η and $-\eta$ on the natural log scale) so that three intervals are created (e.g., $(-\infty, -\eta]$, $(-\eta, \eta)$, and $[\eta, \infty)$), corresponding to the three common conclusions: exclusion, inconclusive, and source attribution (or match). For a QD with 80 characters, these rates for all three methods and both scenarios are presented in Table 4, for $\eta = 4.61$ (corresponding to SLR ≈ 100).

The results in Table 4 illustrate that additional information is gained from looking at all three intervals, compared to simply reporting RMEP and RMED. The results show that when H_p is true,

Table 1
Rates of agreement and disagreement for the three SLR estimates when H_p is true. To disagree, at least one of the three reached a different conclusion.

	Agreement			
	Supports H_p	*Inconclusive*	*Supports* H_d	
n_U	SLR > 100	1/100 < SLR ≤ 100	SLR ≤ 1/100	Disagreement
20	0.000	0.830	0.005	0.165
40	0.070	0.610	0.005	0.315
60	0.125	0.395	0.000	0.480
80	0.200	0.370	0.000	0.430
100	0.240	0.305	0.000	0.455
150	0.330	0.215	0.000	0.455

Table 2
Rates of agreement for the three SLR estimates when H_d is true. To disagree, at least one of the three reached a different conclusion.

	Agreement			
	Supports H_p	*Inconclusive*	*Supports* H_d	
n_U	SLR > 100	1/100 < SLR ≤ 100	SLR ≤ 1/100	Disagreement
20	0.000	0.760	0.200	0.040
40	0.000	0.515	0.450	0.035
60	0.000	0.395	0.600	0.005
80	0.000	0.240	0.750	0.010
100	0.000	0.205	0.765	0.030
150	0.000	0.120	0.865	0.015

[7] One common critique of likelihood methods is that there is no "error rate" one can report for a given case, as is required by the *Daubert* standard. This is due to the fact that source attribution is not typically reported when LRs are employed. However, in a simulated setting overall error rates can be computed by selecting interval values between which match (or no match) statements might be made.

Table 3
Rates of misleading evidence in favor of the prosecution (RMEP) and in favor of the defense (RMED).

	RMEP			RMED		
n_U	SLR_1	SLR_2	SLR_3	SLR_1	SLR_2	SLR_3
20	0.090	0.200	0.150	0.240	0.290	0.330
40	0.015	0.105	0.095	0.160	0.180	0.220
60	0.025	0.095	0.075	0.090	0.105	0.130
80	0.010	0.045	0.045	0.105	0.115	0.125
100	0.015	0.095	0.105	0.055	0.070	0.080
150	0.015	0.055	0.055	0.045	0.050	0.045

Table 4
Rates of exclusion, inconclusive, and match conclusions.

	H_p true			H_d true		
	Exclusion $(-\infty, -4.61]$	*Inconclusive* $(-4.61, 4.61)$	*Match* $[4.61, \infty)$	*Exclusion* $(-\infty, -4.61]$	*Inconclusive* $(-4.61, 4.61)$	*Match* $[4.61, \infty)$
SLR_1	0.000	0.415	0.585	0.760	0.240	0
SLR_2	0.000	0.710	0.290	0.750	0.250	0
SLR_3	0.000	0.715	0.285	0.750	0.250	0

both SLR_2 and SLR_3 tend toward the inconclusive range far more often than SLR_1.[8]

6. Conclusions

Several methods for obtaining a score-based likelihood ratio for handwriting evidence were illustrated, based on a categorical representation of the feature data produced by the proprietary quantification method developed by Gannon Technologies Group. Regardless of the method selected, the results from Table 4 indicate extremely low false match and false exclusion rates are attained when a moderate conclusion threshold is set ($|\ln(SLR)| \leq 4.61$). Since the categorical representation is an extreme simplification of the entire set of feature data generated by Gannon's quantification method (which includes more detailed information, e.g. segment lengths, angles, etc.), it may be that incorporating this additional information would lead to improved performance. However, preliminary investigations indicate that generating a score that makes use of the full set of high-dimensional data and is also highly discriminating is an elusive task (results not shown). While we feel that these types of quantitative analyses may prove fruitful for document examiners at some point, they should only be employed after careful consideration of the inherently subjective decisions the statistical analyst must make in order to calculate such quantities.

Indeed, the primary purpose of this work is to highlight to the forensic community at large, through an empirical study, that score-based likelihood ratios are not the same as, and cannot be interpreted as, the likelihood ratio. Although one should also note that the comparison of Eqs. (2) and (3) suggest that there is a more basic conflict between the two approaches for calculating the "value" of the evidence. This point has been largely ignored in existing literature. Their interpretations must differ as SLRs are considerably more subjective than LRs, in that an analyst must select and defend (1) the similarity (or dissimilarity) score, (2) the appropriate interpretation of the denominator, and (3) the technique relied upon to estimate the numerator and denominator distributions. Due to these points of subjectivity, SLR values must be interpreted with far more caution than the LR based on a well-defined and known probability model (e.g., simple one-contributor DNA LRs).[9]

Some conclusions could be drawn from the various results presented above as to the best SLR technique to use; however, the authors resist as varying any of the subjective factors enumerated above may affect the outcome. Also, innovative score-based approaches have appeared in the literature since this work was undertaken that also should receive consideration [18]. Due to the nature of density estimation, the performance of all methods will heavily depend on the size and representativeness of the database E_A. To date, no such handwriting database exists. The samples used here are not representative of the general population and the simulated evidence documents are not typical of QDs and templates that might be obtained in real case work. As mentioned earlier, our intention is to simply illustrate the feasibility of obtaining an SLR for handwriting evidence, and to emphasize the ambiguities that arise when calculating this value.

Appendix A. Score-based LRs with known distributions

In this segment, we intend to illustrate the theoretical differences between the three SLRs and the LR by way of a simple illustration. Suppose we have two items of evidence: x, a sample of known source (e.g., suspect's writing template, crime scene window) and y, a sample of unknown source (e.g., bank robbery note, glass fragment obtained from the suspect). Suppose it is known, as a general rule, that samples of this type follow a normal distribution with some mean parameter. Assume the variance parameter representing the within source variability for samples of this type, denoted by σ_w^2, is fixed and known. Also, assume the variance parameter for representing the between source variability for samples of this type, denoted by σ_b^2, is fixed and known.

In this example, we consider the sample x to be one observation from a random process. Let X denote the random variable associated with samples of this type, arising from this specific known source (e.g., writing samples obtained from the suspect, fragments obtained from the crime scene windows). For this illustration, suppose X

[8] This trend can also be gleaned from careful consideration of the Tippet plots in Fig. 1. The authors are simply cautioning against reporting RMEP and RMED as the "error rate" for any likelihood ratio method and illustrating a more meaningful alternative.

[9] Often when calculating a LR, the probability distribution is unknown and must be estimated. In these cases, this estimation process induces subjectivity, just as when estimating the SLR.

follows a normal (Gaussian) distribution with mean μ_X, denoted $X \sim N(\mu_X, \sigma_w^2)$.

We also consider the sample y to be one observation from a random process. Let Y denote the random variable associated with samples of this type, arising from this specific unknown source (e.g., writing samples the culprit could have left at the scene, fragments from a specific, but unknown, window found on the suspect). Suppose $Y \sim N(\mu_Y, \sigma_w^2)$.

One final distribution must be defined, that of samples of this type taken from some broader, 'relevant' population denoted by A. For this illustration, suppose these arise from a normal distribution: $N(\mu_A, \sigma_A^2)$ where $\sigma_A^2 = \sigma_b^2 + \sigma_w^2$.

Suppose we are interested in evaluating the evidence in relation to the following two hypotheses:

H_p: x and y arise from the same source
H_d: x and y arise from different sources

Likelihood ratio

The likelihood ratio, assuming x and y are continuous measurements, is defined by

$$\mathrm{LR} \equiv \frac{f(x, y|H_p)}{f(x, y|H_d)},$$

where f denotes the joint probability density function for the random variables X and Y. The assumptions above imply this will be a bivariate normal density. Thus, in this scenario, we can obtain a closed-form solution for the likelihood ratio.

Numerator

Under the numerator hypothesis H_p, the source of x and y are the same (e.g., the suspect wrote the bank robbery note, the fragment found on the suspect is from the crime scene window). Thus x and y are random (independent) draws from the same distribution, so that $\mu_Y = \mu_X$. Therefore,

$$X \sim N(\mu_X, \sigma_w^2),$$
$$Y \sim N(\mu_X, \sigma_w^2).$$

Noting that the joint density for two independent normal random variables is simply the product of their respective densities, we have

$$f(x, y|H_p) = \frac{1}{\sigma_w^2}\phi\left(\frac{x - \mu_X}{\sigma_w}\right)\phi\left(\frac{y - \mu_X}{\sigma_w}\right),$$

where ϕ denotes the standard normal probability density function.

Denominator

Under the denominator hypothesis H_d, the source of x and y are different (e.g., someone else wrote the bank robbery note, the fragment found on the suspect is from another window). A common assumption made in the forensic literature is that the source of y is a random individual selected from some relevant population, so that $\mu_Y = \mu_A$. Therefore, $Y \sim N(\mu_A, \sigma_A^2)$.

Typically X and Y are assumed to be independent – that is, information about the known source provides no additional information about the unknown source. Therefore, the joint density is again the product of their respective densities,

$$f(x, y|H_d) = \frac{1}{\sigma_A \sigma_w}\phi\left(\frac{x - \mu_X}{\sigma_w}\right)\phi\left(\frac{y - \mu_A}{\sigma_A}\right).$$

Taking the ratio of $f(x, y|H_p)$ and $f(x, y|H_d)$, and noting the second term of each cancel, we find

$$\mathrm{LR} = \frac{\sigma_A \phi((y - \mu_X)/\sigma_w)}{\sigma_w \phi((y - \mu_A)/\sigma_A)}.$$

Score-based likelihood ratios

We now would like to compare the behavior of this likelihood ratio with that of the three SLRs in the ideal case, where we have databases that were of sufficiently large as to completely characterize the relevant probability distributions. Before defining a (dissimilarity) score we first note desired properties:

- If x and y are measurements from the same source, we expect the score to be close to zero.
- If x and y are measurements from different sources, we expect the score to be large.

One reasonable such score for two normal random variables, X and Y, is the square of their differences. Thus define the random variable $\Delta(X,Y) = (X - Y)^2$. Another added advantage of this particular score is that we can exploit the following relationship between squared normal distributions and a chi-squared (χ^2) distribution to obtain exact expressions each SLR.

Property 1. Squared Normal Distributions
If $T \sim N(\mu, \sigma^2)$, then
$\frac{T^2}{\sigma^2} \sim X_{1,\lambda}^2$
where $X_{1,\lambda}^2$ denotes a non-central chi-squared distribution with one degree of freedom and non-centrality parameter $\lambda = \mu^2/\sigma^2$.
It is also true that for any random variable R with probability density function (pdf) f_R and scalar $c > 0$, the random variable $S = cR$ has pdf $f_S = (1/c) f_R(s/c)$. Therefore
$f_{T^2}(t) = \frac{1}{\sigma^2} \chi_{1,\lambda}^2\left(\frac{t}{\sigma^2}\right)$,
with non-centrality parameter $\lambda = \mu^2/\sigma^2$.

To evaluate the evidence, now reduced to $\Delta(x, y) = (x - y)^2 = \delta$, via likelihood ratio, in light of the two hypotheses defined above, H_p and H_d, we are interested in

$$\mathrm{SLR} \equiv \frac{g(\delta|H_p)}{g(\delta|H_d)},$$

where g denotes the probability density function for the random variable $\Delta(X,Y) = (X - Y)^2$.

Numerator

All three SLRs make the exact same assumption regarding the numerator probability distribution, namely Y represents an additional independent draw from the distribution associated with the known source. Thus to evaluate the numerator, we need to derive the distribution of $(X - Y)^2$ conditional on $X = x$, where Y follows a $N(\mu_X, \sigma_w^2)$ distribution. For the difference we find:

$$[(X - Y)|X = x] \sim N(x - \mu_X, \sigma_w^2).$$

Per Property 1, the numerator is then

$$g_{\Delta|X}(\delta|x, H_p) = \frac{1}{\sigma_w^2} \chi_{1,\lambda}^2\left(\frac{\delta}{\sigma_w^2}\right),$$

where $\lambda = (x-\mu_X)^2/\sigma_w^2$. The denominator of SLR will vary, depending on which method you choose (SLR_1, SLR_2, SLR_3).

SLR_1 denominator

The method used to arrive at SLR_1 assumes x is a randomly selected sample from some relevant population. That is,

$$X \sim N(\mu_A, \sigma_A^2).$$

The method of SLR_1 also assumes sample y is fixed and known. Therefore we need to find the distribution $g_{\Delta|Y}(\delta|y, H_d)$. We find

$$[(X-Y)|Y=y] \sim N(\mu_A - y, \sigma_A^2).$$

Per Property 1, we find the denominator for SLR_1 is

$$g_{\Delta|Y}(\delta|y, H_d) = \frac{1}{\sigma_A^2}\chi^2_{1,\lambda_1}\left(\frac{\delta}{\sigma_A^2}\right),$$

where $\lambda_1 = (\mu_A - y)^2/\sigma_A^2$. Thus,

$$SLR_1 = \frac{\sigma_A^2\chi^2_{1,\lambda}(\delta/\sigma_w^2)}{\sigma_w^2\chi^2_{1,\lambda_1}(\delta/\sigma_A^2)}.$$

SLR_2 denominator

The method used to arrive at SLR_2 assumes sample y is a randomly selected sample from relevant population. That is,

$$Y \sim N(\mu_A, \sigma_A^2).$$

The method of SLR_2 also assumes sample x is fixed and known. Therefore we need to find the distribution $g_{\Delta|x}(\delta|x, H_d)$. We find,

$$[(X-Y)|X=x] \sim N(x - \mu_A, \sigma_A^2).$$

Per Property 1, we find the denominator for SLR_2 is

$$g_{\Delta|x}(\delta|x, H_d) = \frac{1}{\sigma_A^2}\chi^2_{1,\lambda_2}\left(\frac{\delta}{\sigma_A^2}\right),$$

where $\lambda_2 = (x-\mu_A)^2/\sigma_A^2$. Thus,

$$SLR_2 = \frac{\sigma_A^2\chi^2_{1,\lambda}(\delta/\sigma_w^2)}{\sigma_w^2\chi^2_{1,\lambda_2}(\delta/\sigma_A^2)}.$$

SLR_3 denominator

Here, we neither condition on x or y, and assume that x and y are independent draws from the distribution associated with the relevant population. Thus both X and Y follow $N(\mu_A, \sigma_A^2)$ with X and Y independent. Therefore, their differences are distributed as:

$$X - Y \sim N(0, 2\sigma_A^2).$$

Per Property 1, we find the denominator for SLR_3 is

$$g_{\Delta}(\delta|H_d) = \frac{1}{2\sigma_A^2}\chi_1^2\left(\frac{\delta}{2\sigma_A^2}\right),$$

where χ_1^2 denotes the *central* chi-squared distribution ($\lambda_3 = 0$). Therefore,

$$SLR_3 = \frac{2\sigma_A^2\chi^2_{1,\lambda}(\delta/\sigma_w^2)}{\sigma_w^2\chi_1^2(\delta/2\sigma_A^2)}.$$

It is very important to note that each of the SLRs have a *different* functional form. While here we are making many simplistic and unrealistic assumptions, it stands to reason that the different methods will *necessarily* provide different answers, providing some insight into the results found in this work. SLR_1 and SLR_2 differ only in their non-centrality parameters in the denominator.

We have laid out a framework where we can easily compare the three SLRs to the LR, a luxury that is not possible in most realistic applications. To help the reader comprehend the differences among the SLRs themselves, and to highlight the deviations of each from the LR (in this contrived example), a graphical illustration is provided below.

Comparison

Rather than inspect the rather complex functional forms of each ratio, we have deferred to illustrating their differences graphically. In Fig. A1, we have plotted the values of SLR_1, SLR_2, SLR_3 and LR given by the formulas above, for various values of σ_b^2 and σ_w^2, and for different μ_A. The x-axis represents a range of possible measurements on sample y. For clarity, we have eliminated one source of variability by making the (unrealistic) assumption x always equals μ_X (i.e., the measurement taken from the known source is always equal to the true mean of its distribution).

Consider the first plot appearing in Fig. A1. Here, the mean of the distribution from which the known source sample arises is 0 (i.e., $\mu_X = 0$). The mean of the distribution from the relevant population is -8 (i.e., $\mu_A = -8$). The black solid line represents the likelihood ratio. As expected, the likelihood ratio takes on positive values as the measurement from the unknown sample (y) approaches the mean of the known source. It continues to increase as the value of x increases, up until it becomes less and less likely to have come from the known source.

Moving on to the SLRs, it is important to note that the functional form of SLR_1 is changing along with y, as the non-centrality parameter in the denominator changes with y. This is stated here to emphasize that we are not looking at the functional form of SLR_1, just the evaluation of SLR_1 at each point y. Each of the SLRs peak at $y = 0$, which is a marked deviation from the LR. There is one segment ($y < -5$) where both SLR_1 and SLR_3 are in fact larger than the LR implying that under these conditions, those SLRs are overstating the value of the evidence in favor of the prosecution (though all values are extremely small, thus providing very strong support for the defense hypothesis). However, in most other places where $\log(LR) > 0$ ($LR > 1$) the SLRs are understating the value of the evidence, in some cases drastically so. For example, when $y = 2.5$, we find $LR = 3.05 \times 10^{13}$ ($\log(LR) = 13.48$) whereas $SLR_1 = 1.23 \times 10^7$, $SLR_2 = 1.60 \times 10^2$, and $SLR_3 = 1.34 \times 10^{-2}$. This shows that, at least in this contrived situation, some amount of evidential value is not being adequately captured by these three methods. It is interesting to note that SLR_2 closely approximates (although slightly overestimating) the LR when $y < 0$, and this property is evident in each graph.

The properties displayed in the first graph are certainly the most extreme. In most cases, the SLRs are fairly well-behaved in comparison to the LR, particularly so when the between variability is much larger than the within variability (looking down the rows in Fig. A1). The approximations are also well behaved when the alternate population mean approaches the mean of the known source (looking across the columns in Fig. A1). In general, SLR_3 tends to underestimate the value of the evidence, very rarely producing log values greater than zero. Several additional interesting features can be observed in these plots, and rather than enumerate them here, the

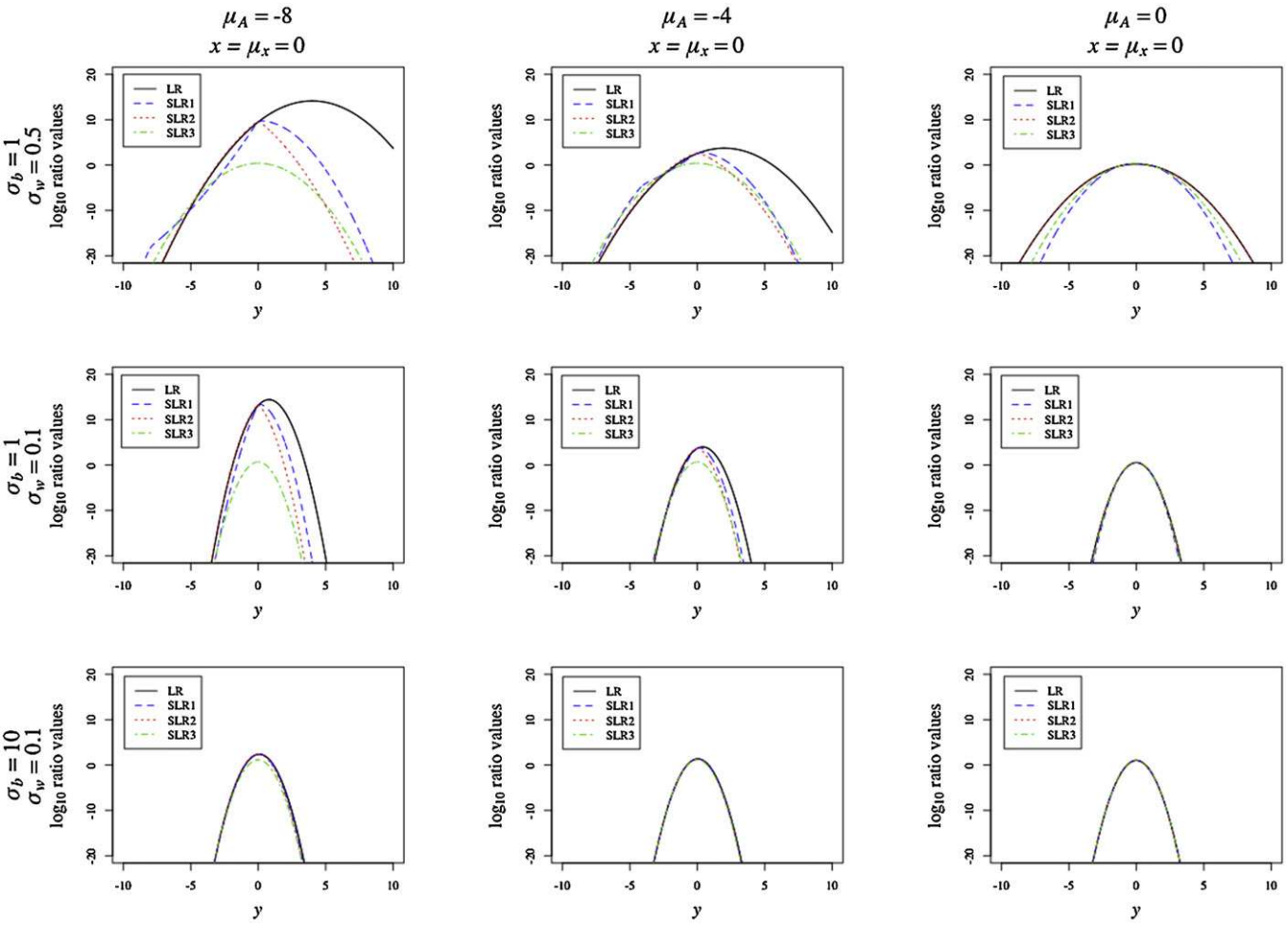

Fig. A1. LR and SLR values, on the log scale, for various μ_A, σ_b, σ_w.

reader is encouraged to study them closely. In addition the R[10] code for this analysis is made freely available as an online supplement to this manuscript.

Appendix B. Dissimilarity score

In this segment, we describe in detail the dissimilarity score used in this study. Suppose we have observed matrices of counts for two writing samples, denoted by $\boldsymbol{x}$ and $\boldsymbol{y}$. For a given letter l (or a given row of $\boldsymbol{x}$ and $\boldsymbol{y}$), define

$$v_{li} = \frac{x_{li} + (1/I_l)}{x_{l\cdot} + 1} \text{ and } \tau_{li} = \frac{y_{li} + (1/I_l)}{y_{l\cdot} + 1},$$

where $x_{l\cdot} = \sum_{i=1}^{I_l} x_{li}, y_{l\cdot} = \sum_{i=1}^{I_l} y_{li}$, and $i = 1, \ldots, I_l$ indexes the distinct isocodes used to write the lth letter in either $\boldsymbol{x}$ or $\boldsymbol{y}$. Then the dissimilarity score for a given letter l is defined as

$$\Delta(\boldsymbol{x}_l, \boldsymbol{y}_l) \equiv \sum_{i=1}^{I_l} \tau_{li} \ln\left(\frac{\tau_{li}}{v_{li}}\right),$$

except when $I_l = 1$ (i.e., when only one isocode is used to write letter l in either $\boldsymbol{x}$ or $\boldsymbol{y}$), in which case $\Delta(\boldsymbol{x}_l, \boldsymbol{y}_l) \equiv 0$.

[10] R: A Language and Environment for Statistical Computing. Vienna, Austria (2011). http://www.R-project.org.

To combine across all letters, $l = 1, \ldots, L$, define a set of weights,

$$\lambda_l \propto \begin{cases} \dfrac{1}{\sqrt{1/x_{l\cdot}} + \sqrt{1/y_{l\cdot}}}, & \min(x_{l\cdot}, y_{l\cdot}) \geq 1 \\ 0, & \text{otherwise} \end{cases},$$

such that $\sum_{i=1}^{L} \lambda_l = 1$. Thus, a letter only receives weight when it appears at least once in both $\boldsymbol{x}$ and $\boldsymbol{y}$. The combined score over all letters is then

$$\Delta(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{L} \lambda_l \Delta(\boldsymbol{x}_l, \boldsymbol{y}_l).$$

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.forsciint.2011.12.009.

References

[1] C.G.G. Aitken, D. Stoney, The Use of Statistics in Forensic Science, Ellis Horwood Limited, Chichester, England, 1991.
[2] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Speech Communication 31 (2-3) (2000) 193–203.
[3] C. Champod, I.W. Evett, B. Kuchler, Earmarks as evidence: a critical review, Journal of Forensic Sciences 46 (6) (2001) 1275–1284.

[4] D.V. Lindley, A problem in forensic science, Biometrika 64 (1977) 207–213.
[5] J.M. Curran, The statistical interpretation of forensic glass evidence, International Statistical Review 71 (3) (2003) 497–520.
[6] C. Champod, I.W. Evett, A probabilistic approach to fingerprint evidence, Journal of Forensic Identification 51 (2) (2001) 101–122.
[7] I.W. Evett, J.A. Lambert, J.S. Buckleton, A Bayesian approach to interpreting footwear marks in forensic casework, Science & Justice 38 (4) (1998) 241–247.
[8] I.W. Evett, B.S. Weir, Interpreting DNA Evidence, Sunderland, MA, Sinauer, 1998.
[9] S. Bozza, F. Taroni, R. Marquis, M. Schmittbuhl, Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship, Applied Statistics 57 (3) (2008) 329–341.
[10] A. Nordgaard, T. Höglund, Assessment of approximate likelihood ratios from continuous distributions: a case study of digital camera identification, Journal of Forensic Sciences 56 (2011) 390–402.
[11] M.A. Walch, D.T. Gantz, J.J. Miller, L.J. Davis, C.P. Saunders, M.L. Lancaster, A.C. Lamas, J. Buscaglia, Evaluation of the individuality of handwriting using FLASH ID – a totally automated, language independent system for handwriting identification, in: Proc. of the 2008 AAFS Annual Meeting, Washington, DC, 2008.
[12] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of three minutiae, Journal of Forensic Sciences 51 (2006) 1255–1266.
[13] L.J. Davis, C.P. Saunders, A.B. Hepler, J. Buscaglia, Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios, Forensic Science International 216 (1–3) (2012) 146–157.
[14] D. Meuwly, Forensic individualisation from biometric data, Science & Justice 46 (2006) 205–213.
[15] J. Gonzalez-Rodriguez, D. Ramos, Forensic automatic speaker classification in the "Coming Paradigm Shift", in: C. Müller (Ed.), Speaker Classification I, vol. 4343, Springer, Berlin/Heidelberg, 2007, pp. 205–217.
[16] N. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – modelling within finger variability, Forensic Science International 167 (2007) 189–195.
[17] C. Neumann, P. Margot, New perspectives in the use of ink evidence in forensic science. Part III. Operational applications and evaluation, Forensic Science International 192 (2009) 29–42.
[18] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, Journal of the Royal Statistical Society: Series A 175 (2) (2012) 1–26.
[19] C. Neuman, New Perspectives in the Use of Ink Evidence in Forensic Science, University of Lausanne, Ph.D. Thesis, 2008.
[20] E.J. Garvin, R.D. Koons, Evaluation of match criteria used for the comparison of refractive index of glass fragments, Journal of Forensic Sciences 56 (2) (2011) 491–500.
[21] C.G.G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, Applied Statistics 53 (1) (2004) 109–122.
[22] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd ed., John Wiley and Sons, Chichester, UK, 2004.
[23] C. Champod, I.W. Evett, G. Jackson, Establishing the most appropriate databases for addressing source level propositions, Science & Justice 44 (3) (2004) 156–164.
[24] A. Ross, K. Nandakumar, A. Jain, Handbook of Multibiometrics, Springer, New York, NY, 2006.
[25] C.P. Saunders, L.J. Davis, A.C. Lamas, J.J. Miller, D.T. Gantz, Construction and evaluation of classifiers for forensic document analysis, Annals of Applied Statistics 5 (1) (2011) 381–399.
[26] S. Kullback, R.A. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 79–86.
[27] A.S. Osborn, Questioned Documents, Boyd Printing Company, Albany, NY, 1929 .
[28] P. Gill, J. Curran, C. Neumann, Interpretation of complex DNA profiles using Tippett plots, Forensic Science International: Genetics Supplement Series 1 (2008) 646–648.