ORIGINAL ARTICLE

Field of the paper

Score-based likelihood ratios to evaluate forensic pattern evidence

Nathaniel Garton[1] | **Danica Ommen**[1] | **Jarad Niemi**[1] | **Alicia Carriquiry**[1]

[1]Iowa State University

Correspondence
Nathaniel Garton, Snedecor Hall
2438 Osborn Dr, Iowa State University,
Ames, IA, 50010, U.S.A.
Email: nmgarton@iastate.edu

Present address
Snedecor Hall 2438 Osborn Dr, Iowa State University, Ames, IA, 50010, U.S.A.

Funding information
Authors Garton, Ommen, and Carriquiry were partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia

In 2016, the European Network of Forensic Science Institutes (ENFSI) published guidelines for the evaluation, interpretation and reporting of scientific evidence. In the guidelines, ENFSI endorsed the use of the likelihood ratio (LR) as a means to represent the probative value of most types of evidence. While computing the value of a LR is practical in several forensic disciplines, calculating an LR for pattern evidence such as fingerprints, firearm and other toolmarks is particularly challenging because standard statistical approaches are not applicable. Recent research suggests that machine learning algorithms can summarize a potentially large set of *features* into a single score which can then be used to quantify the similarity between pattern samples. It is then possible to compute a score-based likelihood ratio (SLR) and obtain an approximation to the value of the evidence, but research has shown that the SLR can be quite different from the LR not only in size but also in direction. We provide theoretical and empirical arguments that under reasonable assumptions, the SLR can be a practical tool for forensic evaluations.

Keywords — Forensic science, likelihood ratio, similarity score, value of evidence, machine learning

Introduction

Score-based likelihood ratios (SLRs) are one of the most popular methods for evaluating the strength of forensic *pattern* evidence (Bolck et al., 2009; Hepler et al., 2012; Davis et al., 2012; Bolck et al., 2015; Neijmeijer, 2016; Galbraith and Smyth, 2017; Leegwater et al., 2017; Morrison and Enzinger, 2018; Chen et al., 2018). We consider the use of SLRs for quantifying the strength of forensic *pattern* evidence to determine whether two pieces of evidence share a specific, known source. In practice, one piece of evidence will have an unknown source, for example a crime scene shoe print, while another piece of evidence will have a known origin, such as a suspect's shoe print. The SLR is formed by taking the ratio of distributions of a low-dimensional statistic calculated from the evidence, called a *score*, under two competing hypotheses. The numerator hypothesis considers the unknown source evidence to have been generated by the suspect while the denominator hypothesis considers the same evidence to have arisen from a source other than the suspect. The numerator hypothesis, therefore, is consistent with a prosecuting attorney's position that the suspect is guilty. Thus, we call the numerator hypothesis the *prosecution* hypothesis. Alternatively, the denominator hypothesis is consistent with the defense attorney's position that the suspect is innocent and did not generate the crime scene evidence in question. Therefore, we call the denominator hypothesis the *defense* hypothesis.

Scores are often a measure of dissimilarity between the piece of evidence known to come from the suspect and that which has unknown origin. However, this needn't always be the case (Morrison and Enzinger, 2018). In the shoe print example, one possible score function could be the Euclidean distance between the vector of pixel values in an image of the suspect's shoe print and an image of the crime scene shoe print. Using data sampled under the two competing hypotheses, the distribution of the score is modeled under both hypotheses and "score-based" likelihood ratios (SLRs) are *estimated*. For example, one could compute Euclidean distances between images of shoe prints created from repeated impressions of the suspect's shoe. This would be used to model the score under the prosecution hypothesis. One could similarly compute Euclidean distances between images of the suspect's shoe prints and images of shoe prints taken from impressions of other sets of shoes. These distances could then be used to model the score under the defense hypothesis. This type of data generation and modeling procedure is essentially that which is used in any paper studying SLRs. Note that there is a pair of implicitly assumed data generating models for the raw data which is unknown, but can be sampled from. For any data generating distributions under the two competing hypotheses, there exists (under the technical condition of measurability), an almost surely unique SLR. This SLR is rarely, if ever, known in practice, and thus SLRs can only be estimated. However, because score distributions are low dimensional, it is reasonable to assume that techniques like density estimation would be accurate. Recently, researchers have successfully applied "black box" machine learning classification algorithms to learn the score function by optimizing some objective (e.g. misclassification rates on a training data set) (Hare et al., 2017; Carriquiry et al., 2019).

Unfortunately, there is a lack of principled justification for the use of SLRs in court. In fact, several authors have raised concerning issues with SLRs. (Hepler et al., 2012) showed that SLRs can be constructed in multiple ways which do not always agree on the strength or the directionality of the evidence. A compelling way to establish the validity of SLRs would be to compare them to the ideal *likelihood ratio* (LR) that one would get if it were possible to correctly specify distributions under the prosecution and defense hypotheses for the original, highly complex data prior to reduction through the score function. Note that these distributions are those implicitly assumed by the data generating procedure necessary to sample scores under the competing hypotheses. Unfortunately, (Neumann and Ausdemore, 2019) suggests that some types of SLRs can poorly approximate these LRs in unpredictable ways. At the same time, (Bolck et al., 2009, 2015) found that, when doing MDMA tablet comparisons, SLRs tended to be more "stable" and $|\log SLR|$ was commonly smaller than $|\log LR|$. In a very similar comparison, (Robert et al., 2011) showed

that using summary statistics to do model selection in ABC can result in inconsistent (in the data sample size) Bayes factors when the statistics being used are not jointly sufficient for the model and model parameters. More importantly for our purposes, they showed that even in nearly ideal situations, the discrepancy between the true Bayes factor and the approximation based on summary statistics is equivalent to a ratio of probability densities with sample size of order n, the number of random variables describing the observed data. The problems identified by (Robert et al., 2011) are fundamentally the same as the issues that arise when comparing an SLR to the ideal LR.

If SLRs (or estimates of them) are to be used in court as a method of calculating the strength forensic evidence, then it is necessary to verify that SLRs will not misrepresent the strength of evidence according to the ideal, but unknowable, LR; this is the focus in our paper. We specifically assume that the prosecution's hypothesis, H_p, is that the evidence with unknown origin was generated by the "distribution of the known source". In other words, the evidence of unknown origin and the evidence of known origin can be considered to be random draws from the *same* probability distribution. An example of this might be that a fingerprint at a crime scene and a fingerprint collected from a suspect are both random draws of fingerprints generated by a particular finger from the suspect. Under the defense hypothesis, H_d, the crime scene evidence is assumed to be generated by a distribution differing from the distribution that of the evidence of known origin. One intuitive, but by no means defining, example of such a hypothesis is that the crime scene evidence was generated by a random draw from a "relevant population". Note that the random draw could be according to any discrete or continuous probability distribution, not necessarily uniform. We assume that the prosecution and defense need only be able to sample evidence from the probability distribution that matches with their hypothesis for how the data was physically generated. The ratio of probabilities of the observed evidence under these true, but unknowable, distributions is the quantity that we define to be the *likelihood ratio*. This definition is the same as that considered in (Royall, 1997, p. 3). We acknowledge, however, that this may differ from other common uses of the term "likelihood ratio" in forensic science. For example, if a forensic statistician were to translate the competing attorneys' physical data generating hypotheses into a pair of parametric models for the observed evidence, any resulting LR would, at best, be an estimate of the ideal LR according to our definition.

Note that in the context of DNA evidence, one may be able to confidently and accurately estimate LRs due to established biological science. Furthermore, measurement error, or any kind of "within source" variation, is likely negligible. This is *not* the case for forensic pattern evidence. (Stern, 2017) discusses many of the challenges that would need to be overcome to develop LRs for pattern evidence. One of the main challenges is defining a probability model for a given person that could accurately describe the variability of, for example, fingerprints or shoe prints across repeated impressions (Stern, 2017).

In this paper, we explore the degree to which any given SLR approximates the ideal LR. We first examine a small example to illustrate when and how an SLR may misrepresent the strength of forensic evidence. We see that, even in this simple example with an intuitively reasonable score, $|\log(SLR) - \log(LR)|$ is unbounded. However, we observe that empirical probabilities of a juror making different decisions depending on whether they are provided with an SLR as opposed to an LR behave reasonably. We then generalize those ideas through the development of probabilistic bounds on the LR and argue that meaningful discrepancies between an SLR and an LR are unlikely. To our knowledge, our results provide the first non-asymptotic theoretical explanation for patterns noticed in (Bolck et al., 2009, 2015) in terms of the "stability" and magnitude of SLRs compared to LRs. Further, this shows that, for at least the type we consider, SLRs tend to underestimate the value of evidence in a predictable way. Thus, we address one of the chief problems identified by (Neumann and Ausdemore, 2019). We also show results from simulation studies designed to reflect more realistic settings, which corroborate our theoretical findings and the observations of (Bolck et al., 2009, 2015). Finally, we conclude by discussing some implications of our results.

A Simple Example

The example that we study uses the same data generating model as in (Lindley, 1977) and (Grove, 1980), but we use the notation from (Hepler et al., 2012). We let X and Y denote random draws from the distribution that generated the evidence with known source and from the distribution that generated the evidence at the crime scene, respectively. The model assumes that there are two sources of variability for evidence: within-source variance, σ_w^2, and between-source variance, σ_b^2. Within-source variance can be thought of as measurement error. Between-source variance can be thought of as the variability of the noise-free characteristic that defines the source. The model we consider is defined as follows,

$$\begin{aligned} H_p : X &\sim N(\mu_x, \sigma_w^2), \quad Y \sim N(\mu_x, \sigma_w^2) \\ H_d : X &\sim N(\mu_x, \sigma_w^2), \quad Y \sim N(\mu_b, \sigma_w^2 + \sigma_b^2), \end{aligned}$$

where $X \perp Y$ under both models. The intuition for this model is that each source in the the population uniquely corresponds to the mean of a Gaussian distribution, and the distribution of means within the broader population is itself distributed according to $N(\mu_b, \sigma_b^2)$. Thus, under H_d the generative procedure for Y, considered to be sampled from a random source, can be written hierarchically as

$$\begin{aligned} Y &\sim N(\mu_y, \sigma_w^2) \\ \mu_y &\sim N(\mu_b, \sigma_b^2). \end{aligned}$$

Some may find it strange that the unknown source evidence and the known source evidence are independent under H_p. However, the prosecution hypothesis specifies that the source of Y is the same as that of X. This means that under H_p, $\mu_x = \mu_y$ is *nonrandom*, i.e. it is conditioned upon. Unless one intends to put a subjective Bayesian prior on μ_x, independence of measurements from within a source is a very common assumption in the literature (Grove, 1980; Aitken and Lucy, 2004). Note that this conditional independence would *not* reasonable in a *common source* LR (Ommen, 2017, Section 3.1). In this simple example, the score is $s(x, y) = (x - y)^2$. These distributional assumptions in combination with this score result in tractable score distributions. We note that this pair of data generating models is almost identical to those considered in the specific source scenario in (Neumann and Ausdemore, 2019, Section 2.3) except we assume the variances of X and Y are equal under H_p. It is straightforward to show, using normal distribution theory, that under H_p, $\frac{s(X,Y)}{2\sigma_w^2} \sim \chi_1^2$, and under H_d, $\frac{s(X,Y)}{2\sigma_w^2+\sigma_b^2} \sim \chi_1^2\left(\frac{[\mu_x-\mu_b]^2}{2\sigma_w^2+\sigma_b^2}\right)$, where $\frac{[\mu_x-\mu_b]^2}{2\sigma_w^2+\sigma_b^2}$ is the non-centrality parameter.

We note that this example, as well as the situations to which our following results apply, are "specific source" LRs (Ommen, 2017, Section 3.2) in the sense that one of the pieces of evidence, X, comes from a known, as opposed to an unknown source. A general set of "specific source" models may neither require that the distribution of the unknown source evidence comes from the same family under H_p and H_d nor that any parameters are shared between the two distributions. Our particular data generating model, however, assumes a shared "within source" variance, σ_w^2, for each piece of evidence. Intuitively, one might imagine that the only variation between observations arising from a fixed source is due to the process of taking measurements, i.e. measurement error.

| SLRs May Be Poor Approximations to LRs

We first consider how the SLR compares to the LR for a grid of possible values for X and Y. We suppose that $\mu_x = 0 = \mu_b$, so that the known source typically produces evidence commonly observed within the broader population. Further, we will fix $\sigma_b = 1$ and $\sigma_w = 0.2$. Figure 1 compares the contour lines of $\log(LR)$ and $\log(SLR)$ on an even grid of possible (x, y) values ranging from -2 to 2. Comparing the LR to the SLR reveals a key difference. Because we have assumed that $X \perp Y$, the LR depends only on Y, but the SLR depends on both X and Y. This is not surprising, but it is important to note that this causes a potential problem. For example, if we restrict ourselves to the values of (X, Y) shown in Figure 1, the LR achieves its minimum values at $Y = -2$ and $Y = 2$, and the SLR achieves its minimum values at $(X, Y) = (2, -2)$ and $(X, Y) = (-2, 2)$. Comparing these minimum values shows that the ratio of the LR to the SLR (or vice versa) can become very large. In this case, when $(X, Y) = (2, -2)$ the LR is roughly 3×10^{19} times larger than the SLR.

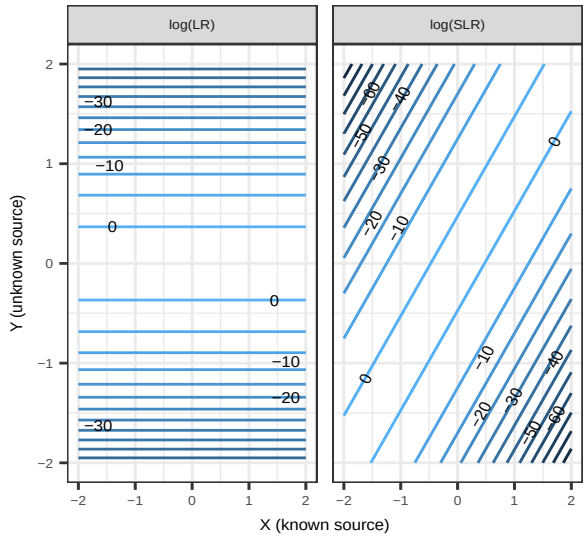

FIGURE 1 Contour plots of $\log(LR)$ and $\log(SLR)$ at an even grid of values from -2 to 2 for $\mu_x = \mu_b = 0, \sigma_w = 0.2, \sigma_b = 1$. Contour lines are horizontal for the $\log(LR)$, but they are of the form $y = x + b$ for the $\log(SLR)$.

In this example, both X and Y are one dimensional, the score is intuitive and simple, yet clearly large discrepancies between the LR and SLR are possible. Furthermore, while it is true that the most troubling inconsistencies between the LR and SLR occur when the observed evidence is rare with respect to the known source distribution (and thus our example discrepancy is improbable), it is not true that the known source itself is highly unusual; we have ensured that evidence generated from the known source is often very similar to that observed from the background population. Thus, it seems that such inconsistencies would be possible in most actual trials.

Perhaps the biggest problem, however, with using an SLR to approximate an LR in this example results from the differences in the contour lines between the SLR and LR. The LR contour lines are horizontal, meaning that the LR only changes with the observed values of the unknown source evidence, but the contour lines for the SLR are diagonal with slope equal to one. This is because the score distributions only depend on (X, Y) through the score function. This

implies that the score densities, and therefore the SLR, are constant for any given fixed score. The score is constant along lines where $y = x + b$ because when $y - x = b$ is fixed, $s(x, y) = b^2$. Because it is possible to fix the score and make Y arbitrarily large or small by simply changing X accordingly, we see that we can make $|\log(SLR) - \log(LR)|$ arbitrarily large. Worse, there is nothing to prevent situations where $\log(SLR)$ is positive and the $\log(LR)$ is negative, and vice versa. This implies there are situations where not only is the discrepancy between an SLR and LR large, but they are directionally inconsistent.

A final remark on the above examples is that it is impossible to determine, based solely on the score, whether there is a large discrepancy between the SLR and the LR. In this specific and seemingly reasonable case, fixing the score does not restrict the range of Y-values that we might obtain. This means that any given score (and consequently SLR) value can be associated with any true LR value.

| Probability of Large Discrepancies

The worst problems described in the previous section involved fixing either the score or the value of Y and manipulating either (X, Y) or X, respectively, such that their values were unlikely under either H_p or H_d. We now show that though the probability of large discrepancies may be high, most large discrepancies will not likely affect jurors' decisions.

Figure 2 shows two histograms of $\log(SLR) - \log(LR)$ generated from 5000 data sets simulated under H_d and H_p when $\sigma_w = 0.2, \sigma_b = 1, \mu_x = \mu_b = 0$. The left panel shows the empirical distribution of $\log(SLR) - \log(LR)$ under H_d, and the right panel shows the empirical distribution of the same quantity under H_p. We see that the distribution under H_d is highly skewed right, and the smallest values that $\log(SLR) - \log(LR)$ can take on are near zero. Under H_p, the distribution is fairly symmetric and unimodal, and most values are between -3 and 3. This difference implies that the directionality and the severity of the discrepancy between the SLR and the LR may be highly dependent on whether or not H_p or H_d is actually true. It also shows that the probability of large discrepancies can be very high. For example, a large fraction of data sets generated under H_d result in values of $\log(SLR) - \log(LR)$ larger than 10.

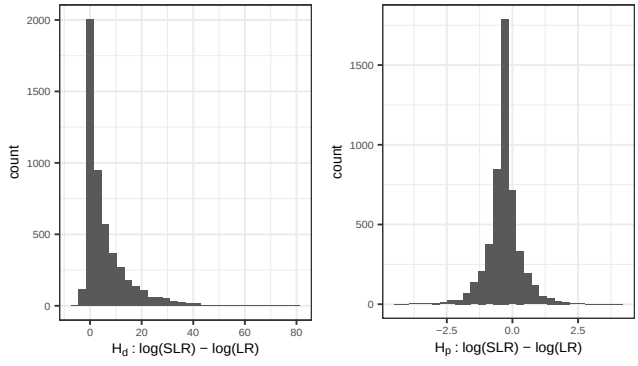

FIGURE 2 Histograms of $\log(LR) - \log(SLR)$ generated from 5000 samples of (X, Y) values under H_p (right panel) and H_d (left panel).

| Impact of Discrepancies on Jurors' Decisions

Such discrepancies arguably only matter insofar as they have the potential to impact a juror's decision. With this in mind, we consider a set of bins for values of the LR and assume that a juror's decision is only impacted by the bin in which the LR falls, not its exact value. The notion that LRs should be presented in such a way has already been advocated for in, for example, (Nordgaard and Rasmusson, 2012; European Network of Forensic Science Institutes (ENFSI), 2016). We use the ranges proposed in (Evett et al., 2000) and similar to the proposal in (Marquis et al., 2016). The scale proposed in (Evett et al., 2000) is as follows:

LR Range	Evidence to support H_p
$1 < LR \leq 10$	Limited
$10 < LR \leq 100$	Moderate
$100 < LR \leq 1000$	Moderately strong
$1000 < LR \leq 10000$	Strong
$10000 < LR$	Very strong.

It will be convenient to define this particular collection of sets as

$$\mathcal{B} \equiv \{(0, 10^{-4}), (10^{-4}, 10^{-3}), ..., (1, 10), ..., (10^4, \infty)\}.$$

For $B \in \mathcal{B}$, figure 3 shows heatmaps of empirical conditional probabilities $P(LR \in B | SLR \in B, H_p)$ on the left and $P(LR \in B | SLR \in B, H_d)$ on the right. The probabilities were computed based on 10^5 simulated observations for each parameter setting. The grey areas on the plots correspond to ranges of values for which no SLR was observed. We see that, under both hypotheses, only when the SLR is observed in the lowest or highest attainable bins are the probabilities of agreement between the LR and SLR close to 1.

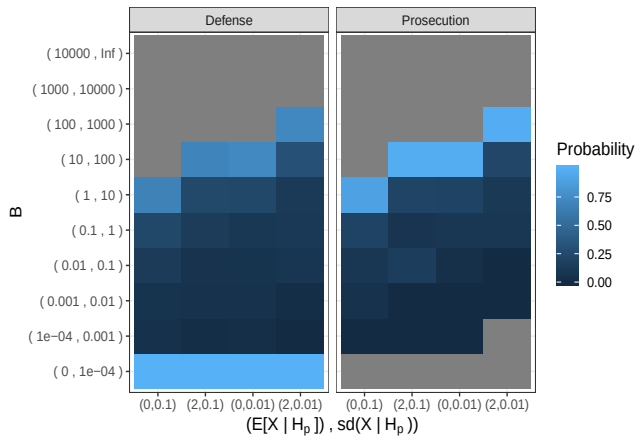

FIGURE 3 Heatmap of empirical estimates of $P(LR \in B | SLR \in B, H_d)$ on the left and $P(LR \in B | SLR \in B, H_p)$ on the right based on 10^5 simulated data sets. That is, for data generated under the prosecution hypothesis, this shows conditional probabilities that the LR is in the same set as the SLR. Grey areas correspond to bins in which an SLR was not observed.

If we examine estimates of the probabilities of agreement averaged over all sets in $\mathcal{B}$, we see that they are encouragingly large. We compute these by approximating

$$\sum_{B \in \mathcal{B}} P(LR \in B | SLR \in B, H_d) P(SLR \in B | H_d)$$

and

$$\sum_{B \in \mathcal{B}} P(LR \in B | SLR \in B, H_p) P(SLR \in B | H_p)$$

with empirical probabilities. These probabilities are shown in Table 1. As the difference between μ_x and μ_b grows or the σ_b/σ_w increases, these probabilities increase as well. This means that, roughly speaking, as the difference between the known and unknown source distributions grows, the probabilities of making an error in the sense of the LR being in a different bin than the SLR decrease.

	H_d	H_p
$(\mu_x = 0, \sigma_w = 0.1)$	0.69	0.87
$(\mu_x = 2, \sigma_w = 0.1)$	0.93	0.91
$(\mu_x = 0, \sigma_w = 0.01)$	0.96	0.93
$(\mu_x = 2, \sigma_w = 0.01)$	0.99	0.92

TABLE 1 Empirical probabilities that the LR is in the same bin as the SLR.

Thus, we see that while probabilities of large discrepancies may be large, the actual probability of arriving at a categorically different decision when faced with the SLR as opposed to the LR is, at worst, moderate and shrinks as there is more signal in the data to discriminate the known source from a random draw from the relevant population. This suggests that the largest and most common discrepancies occur for the most extreme values of the LR and the SLR. In the next section, we show that this pattern occurs more generally.

Probabilistic Bounds on the LR

By constructing probabilistic bounds on the LR conditional on the score, we demonstrate that the patterns observed in the previous section will generalize to realistic settings. The bounds we develop are typically highly conservative, and only one side of each bound can be computed with only knowledge of the SLR. However, we find that these bounds provide enough insight to explain much of the behavior that we have observed up to this point.

Denote by $p(x, y|H_i)$ the joint probability density of the known source evidence, $X \in \mathbb{R}^{q_1}$, and the unknown source evidence, $Y \in \mathbb{R}^{q_2}$, under hypothesis H_i. We will use $S = s(X, Y) \in \mathbb{R}$ to denote the score random variable. We require the following assumptions for our inequalities to hold.

Assumption 1 *$p(x, y|H_i) = p(x|H_i)p(y|H_i)$ for $i \in \{p, d\}$.*

Assumption 1 means that under both the prosecution and defense models, the known and unknown source evidence are generated independently. Any similarity between the two fragments of evidence arises only due to the similarity of the distributions of the known and unknown source evidence.

Assumption 2 $p(x|H_p) = p(x|H_d)$.

Assumption 2 means that regardless of whether the prosecution or defense hypothesis is true, the distribution for the known source data is the same. This assumption is usually reasonable if, for example, the known source data is sampled from the suspect after they are in custody.

Assumption 3 *Given a fixed value $Y = y$, $S(X, y)$ is a nondegenerate random variable.*

This final assumption forces the score to depend meaningfully on the known source evidence. A score function that is constant for any fixed value of $Y = y$, such as the true likelihood ratio, is forbidden. To our knowledge, scores violating this assumption are not used in practice.

Under 1, 2, 3 and for $\alpha \in (1, \infty)$, we have that

$$P\left(LR > SLR/\alpha | s, H_p\right) \geq 1 - \frac{1}{\alpha}, \tag{1}$$

$$P(LR < \alpha SLR | s, H_p) \geq \left(1 - \frac{1}{\alpha}\right)^2 \frac{SLR^{-1}}{E_{Y|s,H_d}\left[LR^{-1}\right]}, \tag{2}$$

$$P\left(LR < \alpha SLR | s, H_d\right) \geq 1 - \frac{1}{\alpha}, \tag{3}$$

$$P\left(LR > SLR/\alpha | s, H_d\right) \geq \left(1 - \frac{1}{\alpha}\right)^2 \frac{SLR}{E_{Y|s,H_p}\left[LR\right]}. \tag{4}$$

The derivations of these inequalities are provided in the supplementary information. They are fairly straightforward applications of Markov's and Cauchy-Scwhartz's inequalities. We note that inequalities 1 and 3 are very similar to two inequalities derived in (Royall, 1997, p. 7) for discrete probability distributions though the derivation differs from ours. It is worth noting that, in 2 and 4, the ratios multiplying $\left(1 - \frac{1}{\alpha}\right)^2$ are less than or equal to 1. To see this, note that

$$\begin{aligned} \frac{SLR}{E_{Y|s,H_p}\left[LR\right]} &= \frac{1}{E_{Y|s,H_p}\left[\frac{p(y|s,H_p)}{p(y|s,H_d)}\right]} && (5) \\ &\leq E_{Y|s,H_p}\left[\frac{p(y|s,H_d)}{p(y|s,H_p)}\right] && (6) \\ &= 1, && (7) \end{aligned}$$

by Jensen's inequality. A similar argument applies to $\frac{SLR^{-1}}{E_{Y|s,H_p}\left[LR^{-1}\right]}$. A consequence of this is that, if the LR is bounded, then the SLR must also be bounded. This is because for every score s, given an upper bound M for the LR, $SLR \leq E_{Y|s,H_p}\left[LR\right] \leq M$. A similar argument applies to the lower bound.

Inequalities 1 and 3 provide a partial explanation for Figure 3. We know that, in general, there is no reason to think that the SLR is close to the LR, and so the fact that $P(LR \in B | SLR \in B, H_i)$ may be small is no surprise. However, if the SLR is sufficiently small and the defense hypothesis is true, it is highly likely that both the SLR and LR will be in the same bin. For example, supposing that we observe a score such that $SLR = 10^{-5}$, then inequality 3 implies that $LR < 10^{-4}$ with at least 0.9 probability.

A similar reasoning can be used to understand the situation when the prosecution hypothesis is true. Supposing that the SLR is in the highest (observed) bin, the LR will be at least in the second highest bin at least 90% of the time. And, because the LR is bounded above when all of of the data distributions are Gaussian, it turns out that the LR doesn't often take a value in a higher bin than the SLR.

One practical consequence of these bounds is that if estimates of score densities are sufficiently accurate and one observes extremely large or small SLR values, it will be likely that the LR will similarly be extremely large or small, provided that large SLRs correspond to situations in which the prosecution hypothesis is true and small SLRs correspond to situations in which the defense hypothesis is true. Finally, inequalities 2 and 4 imply that if the SLR is a good approximation of the expected value of the LR, we can establish bounds similar to those resulting from inequalities 1 and 3. For example, assuming that $SLR^{-1} \approx E_{Y|s,H_d}\left[LR^{-1}\right]$, we can say that $P(LR < 10SLR|s, H_p) \gtrapprox 0.81$.

Simulation Studies

| Multivariate Normal Data

We now consider a simulated example where the score is learned from a "black box" machine learning classifier. We specifically consider the case where the score is a predicted class "probability" from a trained random forest (RF). We briefly note that often the "probabilities" commonly provided by popular implementations of random forest packages are *not* directly interpretable as estimates of posterior probabilities as one might expect (Pudlo et al., 2015). We treat H_p and H_d as the class labels we wish to predict given the observed data, (X, Y). The multivariate Gaussian example is as follows,

$$\begin{aligned} H_p : X &\sim N_5(\mu_x, \Sigma_w), \quad Y \sim N_5(\mu_x, \Sigma_w) \\ H_d : X &\sim N_5(\mu_x, \Sigma_w), \quad Y \sim N_5(\mu_b, \Sigma_w + \Sigma_b). \end{aligned}$$

We specifically consider the case when $\mu_x = (0.5, ..., 0.5)^\top$, $\mu_b = (0, ..., 0)^\top$, $\Sigma_w = 0.5 I_{5\times5}$, $\Sigma_b = I_{5\times5}$. Figure 4 shows histograms of 10000 scores generated under each hypothesis. The random forest was trained on 10000 data sets generated under both hypotheses which are different and independent from the data shown in these histograms. We then use kernel density estimation on the data shown in the histograms to compute score densities and SLRs. Note that it is not always necessary to model score densities directly if the score is an estimate of the posterior class probability in which case one can simply multiply the estimated posterior odds by the inverse prior odds to get an estimate of the likelihood ratio. However, this is not possible with the random forest scores. Therefore, we resort to density estimation here.

Figure 5 shows a scatterplot of the LR versus the SLR for 10000 simulated data sets under H_p and H_d (20000 total). Note that this figure is similar to ones considered in (Neumann and Ausdemore, 2019, Section 3), but (Neumann and Ausdemore, 2019) either compare common source SLRs to specific source LRs, or they compare "anchored" specific source SLRs to specific source LRs. An anchored SLR uses score densities that are conditioned on a value of either the

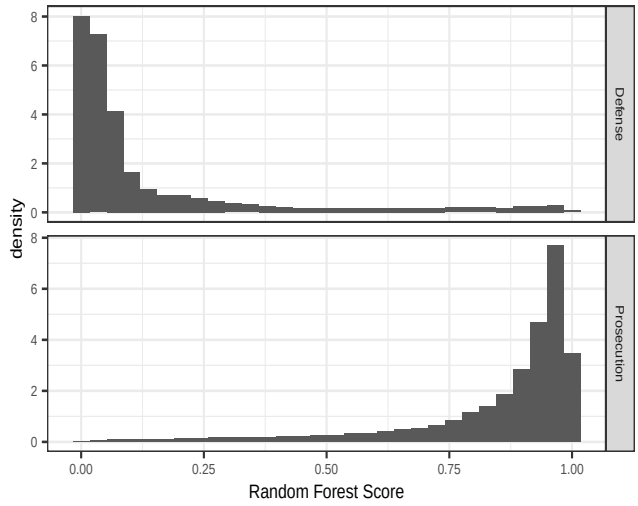

FIGURE 4 Histograms of scores (random forest predictions) for 10000 simulated data sets under both the prosecution and defense hypothesis. Scores are generated from a random forest trained on 20000 simulated data sets, half of which correspond to the prosecution and defense hypotheses. ($\mu_x = (0.5, ..., 0.5)^\top$, $\mu_b = (0, ..., 0)^\top$, $\Sigma_w = 0.5I_{5\times5}$, $\Sigma_b = I_{5\times5}$)

known or unknown source evidence. The red line corresponds to $SLR = LR$ and the red dashed line corresponds to the conservative 95% upper bound on the LR under H_d and lower bound under H_p resulting from inequalities 3 and 1, respectively. Because the other set of bounds require knowledge of the conditional distribution of the LR given the score, we are unable to plot them.

We observe the same patterns in Figure 5 as we did in the bivariate normal example. We see that under both hypotheses, the SLR and LR largely agree as long as the $\log(SLR)$ is not too small. Furthermore, it appears that the conditional expectation of the LR given the score is close to the SLR in this case. However, as the SLR gets smaller, which typically only happens under H_d, we see a wider range of possible LR values. Many LR values are much smaller than the SLR when the SLR is small itself. It is in this situation that the bounds based on inequalities 2 and 4 would be largely useless even if the required conditional expectations were known.

We also see that the bounds resulting from inequalities 1 and 3 are typically overly conservative, with far fewer than 5% of LRs violating the bound. Table 2 provides empirical estimates of $P(LR < \alpha SLR|H_d)$ and $P(LR > SLR/\alpha|H_p)$ for six different levels of α. The bounds based on inequalities 1 and 3 imply that all of these empirical probabilities should be greater than $1 - \frac{1}{\alpha}$, and in most cases they are much greater.

α	100	50	20	10	5	2
H_d	1.00	1.00	0.99	0.98	0.96	0.87
H_p	1.00	1.00	1.00	1.00	1.00	0.96

TABLE 2 Empirical estimates of $P(LR < \alpha SLR|s, H_d)$ and $P(LR > SLR/\alpha|s, H_p)$ averaged across scores for three different levels of α.

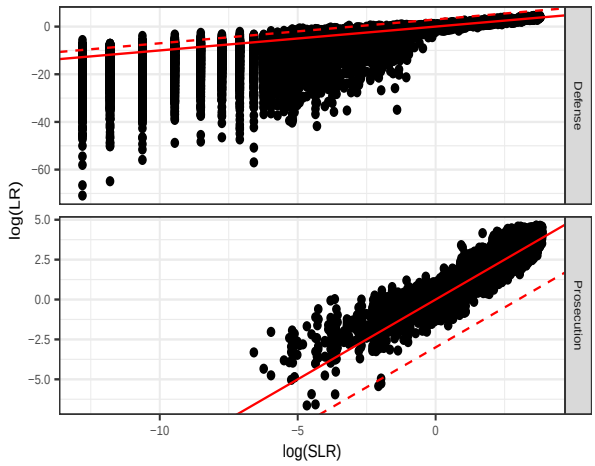

FIGURE 5 Scatterplot of $\log(LR)$ versus $\log(SLR)$ for 10000 simulated 10 dimensional Gaussian data sets under both the prosecution and defense hypothesis. Scores are generated from a random forest trained on 20000 simulated data sets, half of which correspond to the prosecution and defense hypotheses. Score densities are estimated via kernel density estimation. The red lines correspond to what would happen if the SLR and LR were perfectly correlated and the red dashed lines correspond to 95% probability bounds resulting from inequalities 1 and 3. ($\mu_x = (0.5, ..., 0.5)^\top$, $\mu_b = (0, ..., 0)^\top$, $\Sigma_w = 0.5 I_{5\times5}$, $\Sigma_b = I_{5\times5}$)

| Beta Data Simulation

In the previous examples, the LR, and hence the SLR, were bounded above. This was because the data distributions under H_p and H_d were both Gaussian and the variance under H_d was larger than under H_p. This makes sense in our context as variability under H_p is due exclusively to measurement error, whereas the variability of Y under H_d is due both to measurement error and variability between different sources. The consequence of this was that large discrepancies between the SLR and the LR tended to occur only under H_d. We now provide an example where large discrepancies are possible under both hypotheses. We consider the following pair of models,

$$\begin{aligned} H_p &: X_i \overset{iid}{\sim} Beta(\alpha_x, \beta_x), \quad Y_i \overset{iid}{\sim} Beta(\alpha_x, \beta_x) \\ H_d &: X_i \overset{iid}{\sim} Beta(\alpha_x, \beta_x), \quad Y_i \overset{iid}{\sim} Beta(\alpha_y, \beta_y), \end{aligned}$$

where $i = 1, ..., 5$. We specifically consider the case where $(\alpha_x, \beta_x) = (2, 1)$ and $(\alpha_y, \beta_y) = (2, 1)$.

Figure 6 shows score histograms and Figure 7 shows scatterplots of the LR vs the SLR. Patterns in the histograms and scatterplots in this example are similar to those in the multivariate normal example. One major difference here is that the score distributions are more peaked near 1 when H_p is true and near 0 when H_d is true. The second major difference is that the the conditional distribution of the LR given the SLR tends to be skewed both when the SLR is small and H_d is true and when the SLR is large and H_p is true. We still see that the bounds from inequalities 1 and 3 hold, but the bounds from inequalities 2 and 4 would not be very useful outside of $-4 < \log(SLR) < 4$.

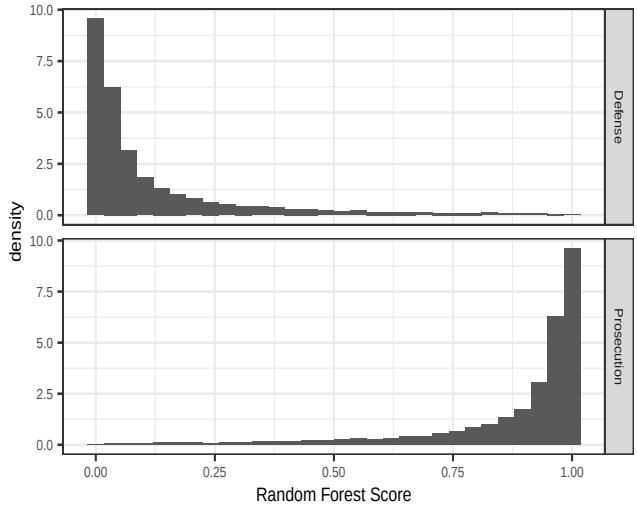

FIGURE 6 Histograms of scores (random forest predictions) for 10000 simulated 10 dimensional Beta distributed data sets under both the prosecution and defense hypothesis. Scores are generated from a random forest trained on 20000 simulated data sets, half of which correspond to the prosecution and defense hypotheses. $((\alpha_x, \beta_x) = (2, 1)$ and $(\alpha_y, \beta_y) = (2, 1))$

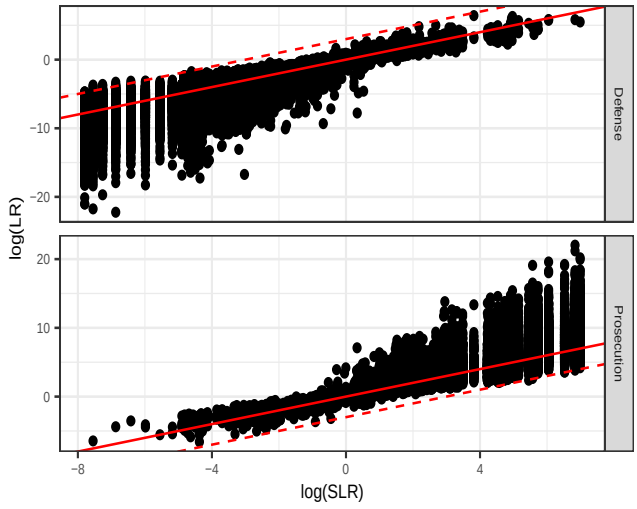

FIGURE 7 Scatterplot of $\log(LR)$ versus $\log(SLR)$ for 10000 simulated 10 dimentional Beta distributed data sets under both the prosecution and defense hypothesis. Scores are generated from a random forest trained on 20000 simulated data sets, half of which correspond to the prosecution and defense hypotheses. Score densities are estimated via kernel density estimation. The red lines correspond to what would happen if the SLR and LR were perfectly correlated and the red dashed lines correspond to 95% probability bounds resulting from inequalities 1 and 3. $((\alpha_x, \beta_x) = (2, 1)$ and $(\alpha_y, \beta_y) = (2, 1))$

Discussion

weave this paragraph into the discussion Readers familiar with approximate Bayesian computation (ABC) (Tavaré et al., 1997; Pritchard et al., 1999; Beaumont, 2010; Lopes and Beaumont, 2010) may recognize that, like ABC, using a score-based approach in our forensic context involves replacing intractable full data likelihoods with likelihoods based on summary statistics. Therefore, recent discussions surrounding theoretical issues encountered when using ABC for model selection are relevant here (Robert et al., 2011; Barnes et al., 2011; Marin et al., 2013), and we will touch more on this shortly. We also note, however, that while model selection is one objective of ABC (and the principal objective in the matching of pattern evidence), the practical details of, and contexts surrounding, ABC and SLRs tend to be somewhat different. ABC is typically used when the full data likelihood is mathematically intractable but simulation from the model is possible. Summary statistics come into play to circumvent the possibly very large computational cost incurred by the need to simulate many observations from the data model, and the entire distribution of the summary statistics needn't be known. On the other hand, there do not generally exist models from which realistic forensic pattern evidence can be simulated. Instead, we collect as many samples of actual data as possible and directly model the entire score distributions, at which point all possible SLR values, including that one which is observed in an actual trial, are available.

(Robert et al., 2011) showed that in general there is no direct connection between a score-based likelihood ratio and the true likelihood ratio. Other research, including our simple bivariate normal example, has further indicated that SLRs need not always be close approximations to the true LR. The requirement, however, that the SLR be close to the LR with probability 1 is stronger than we believe is reasonable to hope for. We have shown instead that for a typical set of statistical hypotheses considered in forensic science, it is possible to establish probabilistic bounds on the LR given an observed score. These bounds are, perhaps, too loose to be used to construct interval estimates of the LR in court, but they support the use of SLRs in place of LRs.

The bounds that we develop in combination with our simulation studies suggest that the largest and most common discrepancies between SLRs and LRs occur when the SLR is either very large, but the LR is much larger, or the SLR is very small, but the LR is much smaller. Among possible discrepancies, these are arguably the least troubling because SLRs are conservative – a property that should favor the defense. Furthermore, these types of discrepancies will only very rarely involve directional inconsistencies between the SLR and LR.

Our simulations involved data that was relatively low in dimension as compared to what is typically encountered in practice. It quickly becomes computationally prohibitive, with high dimensional data, to accurately model the tails of the score distributions nonparametrically, and so we presented no such high dimensional experiments here. However, we derive in the supplementary material the following results

$$\mathcal{D}(p(y|H_p)||p(y||H_d) \geq \mathcal{D}(p(s|H_p)||p(s|H_d)) \tag{8}$$

$$\mathcal{D}(p(y|H_d)||p(y||H_p) \geq \mathcal{D}(p(s|H_d)||p(s|H_p)), \tag{9}$$

where $\mathcal{D}(p(x)||q(x)) \equiv \int \log \frac{p(x)}{q(x)} p(x)dx$ is the Kullback-Leibler (KL) divergence between distributions P and Q having densities p and q, respectively. The KL divergence is a measure of discrepancy between two probability distributions. Larger values of which intuitively imply that larger values of the LR under H_p are common and smaller values of the LR under H_d are common.

As the data dimension increases, one would expect the KL divergence on the left hand side of the above inequalities to grow. The behavior of the right hand side, however, is not obvious. Even if the right hand side grows, it may

do so slower than the left, and so the above bounds become looser and looser. This would result in the same relative behavior of the LR and the SLR shown thus far, but would likely become more extreme.

Unfortunately, one of our sets of bounds involves expectations based on the conditional distribution of the data given the score, which are unavailable. One might worry, then, about cases when the SLR is very large but the defense hypothesis is true. In this case, it is practically impossible to use our lower bound on the LR, and our experiments suggest assuming $SLR^{-1} \approx E_{Y|s,H_d}\left[LR^{-1}\right]$ will likely be unjustifiable. Thus we can say nothing about how small the true LR might be. However, we note that

$$P(LR < SLR/\alpha, SLR > \beta | H_d) \leq P(SLR > \beta | H_d). \tag{10}$$

It is possible to estimate $P(SLR > \beta|H_d)$, and for large β this probability should be very small in the first place. So while it might not be possible to provide a lower bound on the LR in this situation, we can verify that such occasions are rare to begin with.

As a practical note, while the statistical hypotheses and assumptions we have utilized seem reasonable, score distributions may not always be generated in an appropriate way as to make the above results directly applicable. To elaborate, many score distributions are generated using samples that are necessarily dependent. One example of this might be looking at scores for all pairwise comparisons of two shoeprint images created from a suspect's shoe to generate samples from $p(s|H_p)$. Using this empirical score distribution to estimate both probability densities and probabilities of the form $P(SLR \in B|H_i)$ requires some additional assumptions, which require further investigation.

Acknowledgement

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

References

Colin GG Aitken and David Lucy. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):109–122, 2004.

Chris Barnes, Sarah Filippi, Michael Stumpf, and Tom Thorne. Considerate approaches to achieving sufficiency for ABC model selection. *Nature Precedings*, 2011. doi: 10.1038/npre.2011.5952.1.

Mark A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406, 2010. doi: 10.1146/annurev-ecolsys-102209-144621. URL https://doi.org/10.1146/annurev-ecolsys-102209-144621.

Annabel Bolck, Céline Weyermann, Laurence Dujourdy, Pierre Esseiva, and Jorrit van den Berg. Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International*, 191(1):42 – 51, 2009. ISSN 0379-0738. doi: https://doi.org/10.1016/j.forsciint.2009.06.006. URL http://www.sciencedirect.com/science/article/pii/S0379073809002692.

Annabel Bolck, Haifang Ni, and Martin Lopatka. Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk*, 14(3):246–266, 2015. doi: 10.1093/lpr/mgv009.

Alicia Carriquiry, Heike Hofmann, Xiao Hui Tai, and Susan VanderPlas. Machine learning in forensic applications. *Significance*, 16(2):29–35, 2019.

Xiao-Hong Chen, Christophe Champod, Xu Yang, Shao-Pei Shi, Yi-Wen Luo, Nan Wang, Ya-Chen Wang, and Qi-Meng Lu. Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features. *Forensic science international*, 282:101—110, January 2018. ISSN 0379-0738. doi: 10.1016/j.forsciint.2017.11.022. URL https://doi.org/10.1016/j.forsciint.2017.11.022.

Linda J. Davis, Christopher P. Saunders, Amanda Hepler, and JoAnn Buscaglia. Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. *Forensic Science International*, 2012. doi: 10.1016/j.forsciint.2011.09.013.

European Network of Forensic Science Institutes (ENFSI). ENFSI guideline for evaluate reporting in forensic science, 2016. URL http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

IW Evett, G Jackson, JA Lambert, and S McCrossan. The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, 40(4):233–239, October 2000. URL https://doi.org/10.1016/S1355-0306(00)71993-9.

Christopher Galbraith and Padhraic Smyth. Analyzing user-event data using score-based likelihood ratios with marked point processes. *Digital Investigation*, 22:S106 – S114, 2017. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2017.06.009. URL http://www.sciencedirect.com/science/article/pii/S1742287617301962.

Daniel M. Grove. The interpretation of forensic evidence using a likelihood ratio. *Biometrika*, 67(1):243–246, April 1980.

Eric Hare, Heike Hofmann, and Alicia Carriquiry. Automatic matching of bullet land impressions. *The Annals of Applied Statistics*, 11(4):2332–2356, December 2017.

Amanda B. Hepler, Christopher P. Saunders, Linda J. Davis, and JoAnn Buscaglia. Score-based likelihood ratios for handwriting evidence. *Forensic Science International*, 219:129–140, 2012.

Anna Jeannette Leegwater, Didier Meuwly, Majan Sjerps, Peter Vergeer, and Iwo Alberink. Performance study of a score-based likelihood ratio system for forensic fingermark comparison. *Journal of Forensic Sciences*, 62(3), 2017. doi: 10.1111/1556-4029.13339.

Dennis V. Lindley. A problem in forensic science. *Biometrika*, 64(2):207–213, August 1977.

J.S. Lopes and M.A. Beaumont. ABC: A useful Bayesian tool for the analysis of population data. *Infection, Genetics and Evolution*, 10(6):825–832, 2010. doi: 10.1016/j.meegid.2009.10.010.

Jean-Michel Marin, Natesh S. Pillai, Christian P. Robert, and Judith Rousseau. Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):833–859, 2013. doi: 10.1111/rssb.12056. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12056.

Raymond Marquis, Alex Biedermann, Liv Cadola, Christophe Champod, Line Gueissaz, Geneviève Massonet, Williams David Mazzella, Franco Taroni, and Tacha Hicks. Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science and Justice*, 56(5):364–370, September 2016. URL https://doi.org/10.1016/j.scijus.2016.05.009.

Geoffrey Stewart Morrison and Ewald Enzinger. Score based procedures for the calculation of forensic likelihood ratios - scores should take account of both similarity and typicality. *Science and Justice*, 58:47–58, 2018.

Renè Neijmeijer. Assessing performance of score-based likelihood ratio methods for forensic data. Master's thesis, Leiden University, 2016. URL https://openaccess.leidenuniv.nl/bitstream/handle/1887/44582/Neijmeijer%2C%20Ren%C3%A9-s1436643-MA%20Thesis%20MS-2016.pdf?sequence=1.

Cedric Neumann and Madeline A. Ausdemore. Defence against the modern arts: the curse of statistics "score-based likelihood ratios", 2019. URL https://arxiv.org/abs/1910.05240.

Anders Nordgaard and Birgitta Rasmusson. The likelihood ratio as value of evidence - more than a question of numbers. *Law, Probability and Risk*, 11(4):303–315, July 2012. doi: 10.1093/lpr/mgs019.

Danica Ommen. *Approximate statistical solutions to the forensic identification of source problem*. PhD thesis, South Dakota State University, 2017. URL https://openprairie.sdstate.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=2780&context=etd.

Jonathon K. Pritchard, Mark T. Seielstad, Anna Perez-Lezaun, and Marcus W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999. doi: 10.1093/oxfordjournals.molbev.a026091.

Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P Robert. Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866, 2015.

Christian P. Robert, Jean-Marie Cornuet, Jean-Michel Marin, and Natesh S. Pillai. Lack of confidence in approximate Bayesian computation of model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117, September 2011. URL https://doi.org/10.1073/pnas.1102900108.

Richard M. Royall. *Statistical evidence: a likelihood paradigm*. Chapman & Hall, 1997.

Hal S. Stern. Statistical issues in forensic science. *Annual Review of Statistics and Its Applications*, 4:2, 2017.

Simon Tavaré, David J. Balding, R.C. Griffiths, and Peter Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.