Defence Against the Modern Arts: the Curse of Statistics
“Score-based likelihood ratios”

C. Neumann and M.A. Ausdemore
Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, USA.
E-mail: Cedric.Neumann@me.com

October 2019: Version 1

Summary. For several decades, legal and scientific scholars have argued that conclusions from forensic examinations should be supported by statistical data and reported within a probabilistic framework. Multiple models have been proposed to quantify the probative value of forensic evidence. Unfortunately, several of these models rely on ad-hoc strategies that are not scientifically sound. The opacity of the technical jargon used to present these models and their results, and the complexity of the techniques involved make it very difficult for the untrained user to separate the wheat from the chaff. This series of papers is intended to help forensic scientists and lawyers recognise limitations and issues in tools proposed to interpret the results of forensic examinations. This paper focuses on tools that have been proposed to leverage the use of similarity scores to assess the probative value of forensic findings. We call this family of tools “score-based likelihood ratios”. In this paper, we present the fundamental concepts on which these tools are built, we describe some specific members of this family of tools, and we explore their convergence to the Bayes factor through an intuitive geometrical approach and through simulations. Finally, we discuss their validation and their potential usefulness as a decision-making tool in forensic science.

Keywords: Bayes factor; Weight of evidence; Pattern evidence; Trace evidence; Biometry; Score-based likelihood ratios; distance/similarity measures
1. Introduction

For more than half a century, legal and scientific scholars have widely advocated Bayesian reasoning for handling the uncertainty in the determination of the source of forensic evidence (see Evett (1998) for a historical reference, Aitken and Taroni (2004) for a general introduction). Bayesian inference revolves around the use of the Bayes factor to update one’s prior beliefs about two competing propositions related to the source of the evidence. The updated beliefs are often called posterior beliefs. Posterior beliefs are probabilities and do not equate to categorical decisions. The path leading from a posterior probability to a decision involves the use of loss functions and has been described, in the forensic context, by Biedermann et al. (2008). Proponents of Bayesian reasoning argue that it is the only coherent and logical manner for performing inferences in forensic science. They further argue that, in casework, forensic scientists do not possess the information that would allow them to assign prior beliefs to the propositions that are considered. Consequently, forensic scientists should limit themselves to reporting Bayes factors and let others (e.g., fact-finders, jurors, judges) complete the inference process. Therefore, the challenge for forensic scientists is to assign values to the Bayes factors for various evidence types (e.g., fibre, paint, glass, fingerprints, footwear impressions, handwriting, toolmarks, etc.).

Forensic scientists have been able to assign Bayes factors to simple forms of forensic evidence for many years. For example, the statistical models used to quantify the weight of single DNA profiles or simple mixtures of DNA profiles are well understood. Conversely, only anecdotal attempts have been made to assign Bayes factors to complex forms of forensic evidence, such as handwriting and fingerprint evidence (Bozza et al., 2008; Forbes et al., 2014; Neumann et al., 2015; Tackett, 2018).

Assigning Bayes factors to complex evidence forms requires defining reasonable likelihood functions to represent the joint distributions of heterogenous and high-dimensional feature vectors. To bypass the need to work with intractable likelihood functions, researchers have concentrated on the use of (dis)similarity metrics or kernel functions to reduce the complexity and dimensionality of the problem. These attempts have given rise to a family of ad-hoc methods aimed at describing the probative value of forensic evidence. We call these methods “score-based likelihood ratios”.

In this paper, we show that these ad-hoc tools offered to support Bayesian inference of the source of complex forms of evidence may have some merits as deterministic decision tools; however, their use within a Bayesian paradigm is not appropriate. As a result,

---

1 The adjective “simple” refers to the level of complexity of the mathematical representation of the evidence and of the probabilistic models involved; it is not used to qualify how the evidence is transferred, recovered or analysed.

2 A single DNA profile is usually represented by a small set of independent bivariate categorical vectors, which joint distributions under the two competing propositions are usually trivial to model.

3 For example, in the case of fingerprint evidence, a single minutia can be represented by its Euclidean coordinates (bivariate continuous variable), its type (nominal variable) and its direction in the ridge flow (circular variable). An impression where $n$ minutiae are observed is then represented by a $4n$-long vector, which contains three different types of variables.
they cannot be used to update prior beliefs on the source of a finger impression as part of Bayesian reasoning, and they are not fulfilling the requirements set forth by the legal and scientific scholars who advocate for a move towards a more formal Bayesian approach in forensic science. Specifically, we show that:

(a) Some tools do not address the question of interest;
(b) Some tools can induce incoherent inference, in the sense that they can support both mutually exclusive propositions using the same information obtained from the evidence;
(c) Some tools may unpredictably over- or underrepresent the weight of the evidence represented by a set of trace and control objects.

2. Common source vs. specific source scenarios

The inference of the identity of the donor of a trace from its comparison with control material from a known source requires considering two mutually exclusive hypotheses, denoted $H_0$ and $H_1$ below. A certain lack of formalism in the formulation of these hypotheses has resulted in the development of models and the collection of data that mismatch the needs of the criminal justice system.

The next sections briefly develop two formal scenarios that frame the inference of the source of forensic evidence: the common source scenario and the specific source scenario (Ommen et al., 2017). These scenarios are often confused with one another. This results in the development of models under one scenario to answer the question considered by the other one. Thus, understanding their differences is important and helps assess the potential and limitations of the different inference frameworks for forensic evidence.

2.1. Common source scenario

The common source scenario considers whether two pieces of forensic evidence originate from the same source or from different sources without formally specifying which sources are considered. This scenario typically relates to inference of the source of two trace samples, $e_{u_1}$ and $e_{u_2}$ (e.g., two finger impressions recovered on two different crime scenes or even on the same crime scene), with the goal of determining if they originate from the same unknown source (e.g., determining whether the two scenes are linked or whether there were one or more perpetrators).

The hypotheses considered in the common source scenario can be stated as follows:

$H_{0_{CS}}$: $e_{u_1}$ and $e_{u_2}$ originate from the same, unknown, source;

$H_{1_{CS}}$: $e_{u_1}$ and $e_{u_2}$ originate from two different, unknown, sources.

In this scenario, the true source of each piece of evidence is considered to be a random source from a population of potential sources. Under $H_{0_{CS}}$, the source of the two pieces of evidence is the same random source, while the evidence material originate from two different random sources under $H_{1_{CS}}$. These hypotheses are commonly called the prosecution hypothesis and defence hypothesis (Aitken and Taroni, 2004).
2.2. Specific source scenario

Contrary to the previous scenario, the specific source scenario typically involves the comparison of trace material, \( e_u \), with control material from a known source, \( e_s \), with the goal of determining if the trace was made by the considered source. The hypotheses considered in the specific source scenario can be stated as follows:

\[
H_{0:SS}: e_u \text{ and } e_s \text{ were made by Source X}.
\]

\[
H_{1:SS}: e_u \text{ was made by another source than Source X}.
\]

In this scenario, Source X is identified. It can be considered fixed. Under \( H_{1:SS} \), the true source of \( e_u \) is unknown and is considered to be a random source from a population of potential sources, while Source X remains the undisputed donor of \( e_s \).

The distinction between both scenarios is not merely theoretical. Each scenario results in different likelihood functions for the same information, and in different interpretations of the results of forensic examinations.

In the vast majority of cases, the inference questions of greatest interest to the criminal justice system fall under the umbrella of the specific source scenario. Nevertheless, the determination that two pieces of evidence were made by the same unknown source may be relevant to some investigations (e.g., for forensic intelligence-led investigations). Since these two scenarios are different and consider two radically different pairs of hypotheses, it seems intuitive that they should not be interchanged. Unfortunately, they are often confused.

2.3. Generative models

This paper explores the convergence of different models and inference frameworks partly through simulations. The simulations rely on generative models that give simplified representations of how the data arise under the different hypotheses laid out in Sections 2.1 and 2.2. These simplified models are used throughout the paper and are introduced below.

We consider a simple univariate setting to explore the construction and convergence of the different Bayes factors in the common and specific source scenarios. The common source scenario considers whether two traces originate from the same, unknown, source; thus, the generative models under both common source hypotheses can be represented by two hierarchical random effects models:

\[
e_{u1} = \mu + d_1 + u_1, \text{ where } d_1 \sim N(0, \sigma_d^2) \text{ and } u_1 \sim N(0, \sigma_u^2);
\]

\[
e_{u2} = \mu + d_2 + u_2, \text{ where } d_2 \sim N(0, \sigma_d^2) \text{ and } u_2 \sim N(0, \sigma_u^2);
\]

where \( \mu \) is the mean of the population of sources, \( d_1 \) and \( d_2 \) are random effects due to sources, and \( u_1 \) and \( u_2 \) are random effects due to objects within sources.\(^5\)

\(^5\)If we consider the practical example of fingerprint evidence, \( \mu \) represents the mean of the distribution of the characteristics of all friction ridge skin in a population; \( d_1 \) and \( d_2 \) represent the deviations between the overall mean of the population, \( \mu \), and the friction ridge characteristics of the first and second sources; \( u_1 \) and \( u_2 \) are random effects that affect the final appearance (after
Under $H_{0_{CS}}$, the two pieces of evidence originate from the same source and, thus, have the same value for $d_1$ and $d_2$. Under $H_{1_{CS}}$, the two pieces of evidence originate from two different sources and are therefore independent. Thus, the respective joint distributions of $e_{u_1}$ and $e_{u_2}$ are:

$$(e_{u_1}, e_{u_2})|H_{0_{CS}} \sim MVN\left((\mu, \beta), \begin{pmatrix} \sigma^2_d + \sigma^2_{u_1} & \sigma^2_d \\ \sigma^2_d & \sigma^2_d + \sigma^2_{u_2} \end{pmatrix}\right);$$

$$(e_{u_1}, e_{u_2})|H_{1_{CS}} \sim MVN\left((\mu, \beta), \begin{pmatrix} \sigma^2_d & 0 \\ 0 & \sigma^2_d + \sigma^2_{u_2} \end{pmatrix}\right).$$ (1)

The generative models in the specific source scenario differ depending on whether $H_{0_{SS}}$ or $H_{1_{SS}}$ is considered. Under $H_{0_{SS}}$, when all evidence originate from the same source, the models are two simple random effects models:

$$e_u = \mu_d + u, \text{ where } u \sim N(0, \sigma^2_u);$$

$$e_s = \mu_d + s, \text{ where } s \sim N(0, \sigma^2_s);$$

and, where $\mu_d$ represents the mean for the considered specific source, and $u$ and $s$ are random effects respectively corresponding to trace and control samples.

Under $H_{1_{SS}}$, the generative model for the control material from the specific source is the same as under $H_{0_{SS}}$ (indeed, there is no dispute that $e_s$ originates from the known source). However, the model for the trace material, $e_u$, is a hierarchical random effects model to reflect that its true source is an unknown source in the population of potential sources:

$$e_u = \mu + d + u, \text{ where } d \sim N(0, \sigma^2_d) \text{ and } u \sim N(0, \sigma^2_u);$$

$$e_s = \mu_d + s, \text{ where } s \sim N(0, \sigma^2_s);$$

and where $\mu, \mu_d, d, u$ and $s$ are defined as above.\(^7\)

Under $H_{0_{SS}}$, trace and control materials are independent given $\mu_d$, and their joint distribution is multivariate normal. Under $H_{1_{SS}}$, trace and control materials are independent since they are not from the same source, and their joint distribution is also multivariate normal. We have:

development, transfer, photography, etc.) of fingerprints resulting from different impressions of the fingers represented by $u_1$ and $u_2$ on various surfaces. The effects $u_1$ and $u_2$ may be distinct as two impressions may be affected by different sets of factors.

\(^6\)Note that they do not necessarily have the same value for $\sigma^2_u$ if the different pieces of evidence were left under different conditions.

\(^7\)A similar analogy to the one made in Footnote 5 can be made here. The constant $\mu_d$ represents the characteristics of the friction ridge skin of a specific finger from a known individual (e.g., a suspect). The effect $d$ represents the characteristics of the friction ridge skin of a specific finger from an unknown individual (e.g., the true donor of the latent print). The random effects $u$ and $s$ affect the final appearance (after development, transfer, photography, etc.) of fingerprints resulting from different impressions of the fingers represented by $\mu_d$ and $d$ on various surfaces. The variance terms, $\sigma^2_u$ and $\sigma^2_s$, may be distinct as latent and control prints are acquired under different sets of conditions.
If we take the view that forensic evidence must be evaluated within a Bayesian paradigm, then we are interested in quantifying the weight of the evidence using Bayes factors (or, when the parameters are known, likelihood ratios). In the common source framework, the likelihood ratio for $e_u$ and $e_s$ is (Ommen et al., 2017):

$$LR_{CS} = \frac{f(e_u, e_s | H_{0CS})}{f(e_u, e_s | H_{1CS})} = \frac{f(e_u | H_{0CS}) f(e_s | H_{1CS})}{f(e_u | H_{1CS}) f(e_s | H_{0CS})};$$

(3)

while the likelihood ratio for $e_u$ and $e_s$ in the specific source framework is (Ommen et al., 2017):

$$LR_{SS} = \frac{f(e_u, e_s | H_{0SS})}{f(e_u, e_s | H_{1SS})} = \frac{f(e_u | H_{0SS})}{f(e_u | H_{1SS})}.$$

(4)

2.4. Convergence of specific and common source Bayes factors

We already mentioned that, in many cases, forensic scientists are working within the context of the specific source scenario. They are provided with trace material and they want to infer whether it originates from the same specific source that was used to obtain the control material. Using the toy examples in Equations (1) and (2), we can study the convergence of the common source likelihood ratio in Equation (3) to the specific source likelihood ratio in Equation (4) that should be used to quantify appropriately the weight of the evidence.

To compare these likelihood ratios, we consider pairs of $e_u$ and $e_s$ generated by the model in Equation (2) under $H_{0SS}$ or $H_{1SS}$ and we calculate the likelihood ratios in Equations (3) and (4). To calculate the common source likelihood ratio using the data generated under the specific source model, we set $e_{u1} = e_u$, $e_{u2} = e_s$, $\sigma^2_{u1} = \sigma^2_u$ and $\sigma^2_{u2} = \sigma^2_s$.

Figure 1 presents the results of three experiments. In all three experiments, $\mu = 10$, $\sigma^2_d = 10$ and $\sigma^2_u = 2$. All simulations were repeated 1,000 times. In the first experiment, the characteristics of the source of $e_s$ were chosen to be relatively common with respect to the population of sources ($\mu_d = 9$) but also quite variable ($\sigma^2_s = 1$). In the second experiment, the characteristics of the source of $e_s$ were chosen to be rare with respect to the population of sources ($\mu_d = 0$) but remained variable ($\sigma^2_s = 1$). In the last experiment, the variability of the characteristics of the known source of the control material was chosen to be virtually negligible ($\sigma^2_s = 10^{-5}$[8]).

The values for the models’ parameters are chosen to represent different situations: paint, glass and fibres are material with somewhat large within-source variability and their specificity is variable; control finger impressions have low within-source variability and contain a large number of very discriminative features; finally, DNA profiles obtained directly from individuals have virtually not within-source variability (in terms of allelic designation) and are highly specific to these individuals.
Fig. 1. Comparisons between LRs in the common source (y-axis) and specific source scenarios (x-axis). Columns: the left column reports the results when $e_u$ and $e_s$ have been sampled under $H_{0SS}$; the right column reports the results under $H_{1SS}$. Rows: (a) the source of the control material has common characteristics; (b) the source of the control material has rare characteristics; (c) the source of the control material has common characteristics, however it has virtually no variance.

The results of the experiments show that likelihood ratios for the common and the
specific source scenarios do not converge unless the variability of the source of the control material is negligible\(^9\). Importantly, the results for the first two experiments in Figure 1 show that the common source likelihood ratio unpredictably over- or underestimates the value of the specific source likelihood ratio. That said, while assigning a common source likelihood ratio when \(H_1\) is true may underestimate the corresponding specific source likelihood ratio, Figure 1 shows that common source likelihood ratios have a marked tendency to overestimate their counterparts (we have not found a situation where common source likelihood ratios consistently underestimate specific source likelihood ratios).

The lack of convergence raises issues regardless of whether \(H_0\) or \(H_1\) is true. When \(H_0\) is true, underestimating the value of the specific source likelihood ratio may result in the erroneous exclusion of the source of the control impressions as the source of the trace impression. While this is an obvious issue, the criminal justice system currently considers this to be a better outcome than the erroneous identification of an innocent. Furthermore, when \(H_0\) is true, overestimating the value of the specific source likelihood ratio only results in being overconfident in the support of the correct conclusion that the source of the control material is also the source of the trace material; thus, the impact of the overestimation may be considered minimal. Unfortunately, when \(H_1\) is true, overestimating the value of the specific source likelihood ratio may result in exculpatory evidence not being given the appropriate weight in favour of an innocent, yet suspected, source. In fact, Figures 1(a) and (b) show that some pieces of evidence result in values of the specific source likelihood ratios that are less than one and values of the common source likelihood ratios that are greater than one.

Ultimately, miss-specifying the interpretation framework results in answering the wrong question, and may result in serious miscarriages of justice when the common source likelihood ratio is used instead of the specific source one.

### 3. Score-based likelihood ratios

The use of scores to calculate score-based likelihood ratios can be tracked back to the late 1990s and early 2000s and the field of speaker recognition, fingerprint and other types of evidence (Meuwly and Drygajlo, 2001; Champod et al., 2001; Gonzalez-Rodriguez et al., 2003, 2005, 2006; Egli et al., 2006; Meuwly, 2006; Neumann et al., 2007; Neumann and Margot, 2009; Egli-Anthonioz and Champod, 2014). The natural proximity of these forensic sub-disciplines and biometry have led researchers to quickly realise the benefits of modelling the (dis)similarity between pairs of observations, rather than modelling complex feature vectors in their original space. This enabled them to bypass the need to work with the intractable likelihood functions associated with complex forms of pattern and trace evidence, and, instead, allowed them to model univariate continuous data.

Different constructions of score-based likelihood ratios have been proposed over the

---

\(^9\)This is typically the case for forensic DNA analysis when single full DNA profiles are considered. Since the allelic designation of a full DNA profile is extremely reproducible, the inference of the identity of source of a pair of full DNA profiles will be the same under both common and specific source scenarios. This may explain why the distinction between common and specific source scenarios was not discussed until recently by Ommen et al. (2017).
years and, despite their limitations, their use in casework is advocated (at least in Europe by the European Network of Forensic Science Institutes (ENFSI) (2016)). The concept behind most models rests on the comparison of the likelihood of the score calculated between a single trace object, $e_u$, and a single control object from a known source, $e_s$, evaluated in two different density functions. These density functions are based on the sampling distributions of the score under two mutually exclusive propositions. The concept is illustrated in the left panel of Figure 2 which shows the ratio of $f(\delta(e_u, e_s)|H_0)$ and $f(\delta(e_u, e_s)|H_1)$, where $\delta(e_u, e_s)$ is the score between the single trace and single control objects, and $f(\cdot|H_0)$ and $f(\cdot|H_1)$ represents the sampling distributions of interest.

We warn the reader that other models involving scores have been proposed (Armstrong et al., 2017; Swofford et al., 2018; Ausdemore et al., 2019; Hendricks et al., 2019) but are not considered to be score-based likelihood ratios. These models do not rely on the ratio of the likelihoods of the score in two sampling distributions. For example, the right panel of Figure 2 shows the concept underlying a model called FRStat (Swofford et al., 2018, Neumann, 2019) which relies on the ratio of two tail probabilities bounded by the score between a trace and a control object.

---

Fig. 2. Comparison between two different concepts for the use of summary statistics/kernel functions to provide some information on the probative value of fingerprint evidence. Left panel: score-based likelihood ratio obtained by calculating the ratio of the density of a summary statistic for an observed pair of trace and control objects, $\delta(e_u, e_s)$, in its sampling distribution under the first proposition and in its sampling distribution under the second proposition. Right panel: FRStat-like (Swofford et al., 2018; Neumann, 2019) ratio obtained by calculating the ratio of $\alpha$- and $\beta$-error types associated with a decision of identification or exclusion at an observed level of dissimilarity of a pair of trace and control impressions, $\delta(e_u, e_s)$, using the two sampling distributions under the two propositions.

See Neumann (2019) for a critic of FRStat.
Neumann and Ausdemore

3.1. Similarity metrics and kernel functions

A score can have two interpretations: it can be seen as a summary statistic resulting from the comparison of two objects, or it can be seen as the scalar projection resulting from the inner product of two vectors. In the first case, we talk about \textit{(dis)similarity metrics}, while in the second case we talk about \textit{kernel functions}. Both functions map complex random vectors from their natural space to the real line, $\mathbb{R}$, and both offer great flexibility to researchers. First, researchers can design algorithms that measure the distance between two objects, such that the value representing that distance is minimised when the two objects originate from the same source, and is maximised when they originate from different sources\footnote{Some algorithms maximise the value of the score when the objects originate from the same source, and minimise it when the objects originate from different sources. The upcoming discussion on the use of scores is not affected by their interpretation as similarity scores or distances.}. Secondly, the level of (dis)similarity between pairs of objects can be expressed as a univariate continuous random variable, which probability distribution is significantly more convenient to model than the distribution of original vectors representing the observations made on the impressions. However, the benefits of being able to work in a continuous univariate space are not without limitations, which are explored below.

When the function used to calculate scores is considered as a summary statistic, we can discuss the sampling distributions of the score under various situations. When the score function is considered to be a kernel function, the score has a geometric interpretation. Formally, a score interpreted as a summary statistic of the (dis)similarity between two objects $e_i$ and $e_j$ can be defined as $\delta(e_i, e_j)$, where $\delta$ is any function with a real-valued output. A score interpreted as the inner product of two vectors can be similarly defined as $\kappa(e_i, e_j) = \langle \eta(e_i), \eta(e_j) \rangle$, where $\kappa$ is a kernel function, $\eta$ is a set of basis expansions and $\langle \cdot, \cdot \rangle$ is the inner product. The main difference between $\delta$ and $\kappa$ is that $\kappa$ has to be a positive semi-definite symmetric function, while there is no requirement for the construction of $\delta$.

These two different perspectives on the score function are used to investigate the different score-based models in the next sections. The generative models described in Section 2.3 are used to discuss the convergence of these models to the specific source likelihood ratio of interest in Equation (4) as we did in Section 2.4. In order to perform these simulations, both $\delta$ and $\kappa$ are defined as the squared Euclidean distance, which is both a summary statistic and a valid kernel function, and which also has tractable distributions for the chosen generative models (Hepler et al., 2012).

3.2. Common source score-based models

The models in the following sections are best introduced through experiments that allow to study the sampling distributions under the two alternative propositions.

The first type of score-based model is based on results obtained in biometry (Section 4.3 in Ross et al. (2006))\footnote{Early papers on the use of scores to approximate Bayes factors in forensic science lack clarity on how the sampling distributions of the scores were studied. While it seems that the work by Champod et al. (2001) and Gonzalez-Rodriguez et al. (2005) describes asymmetric score-based likelihood ratios (see Section 3.5), it may be that they are in fact common source score-based}. The score, $\delta(e_u, e_s)$, is evaluated using sampling distributions...
Curse of Statistics: “Score-based likelihood ratio”

based on the following thought experiments:

(a) When the prosecutor proposition is correct, $\delta(e_u, e_s)$ is a score that is calculated by comparing trace and control material from the same, random, source. The sampling distribution of $\delta(e_u, e_s)$ under $H_0$ can be studied by considering a sample of sources from a relevant population, and by sampling and comparing a single trace and a single control object from each source.

(b) When the defence proposition is correct, $\delta(e_u, e_s)$ is a score that is calculated by comparing trace and control objects from different sources. The sampling distribution of $\delta(e_u, e_s)$ under $H_1$ can be studied by sampling independent pairs of objects from a relevant population, and by comparing a trace object from the first source to a control object from the second source.

This type of model has one main advantage: both sampling distributions can be learned ahead of time based on a large sample of sources from a relevant population. Once learned, the two sampling distributions can be used for any new case. It also has two main limitations. First, it is only reporting the average density of $\delta(e_u, e_s)$ under both propositions. Neither sampling distribution is specific to the donor of $e_s$. This type of model clearly addresses the common source pair of propositions and is not relevant to a specific case involving the comparison of a trace object to known control material from a given source. Second, Bayes factors can roughly be viewed as the ratio between some measure of similarity between the characteristics of the trace and control objects, and some measure of the rarity of the characteristics of the trace. However, the model described above does not account for the rarity of the trace characteristics at all. This type of model only accounts for the rarity of the level of similarity.

To compare this type of model to the specific source likelihood ratio in Eq. 4, we use the generative models proposed in Eq. 1. By defining $\delta(e_u, e_s) = (e_{u_1} - e_{u_2})^2$, we have that:

\[
(e_{u_1} - e_{u_2})^2 | H_{0CS} \sim \frac{1}{\sigma_{u_1}^2 + \sigma_{u_2}^2} \chi^2 \left( \frac{(e_{u_1} - e_{u_2})^2}{\sigma_{u_1}^2 + \sigma_{u_2}^2} \right),
\]

\[
(e_{u_1} - e_{u_2})^2 | H_{1CS} \sim \frac{1}{\sigma_{u_1}^2 + \sigma_{u_2}^2 + 2\sigma_p^2} \chi^2 \left( \frac{(e_{u_1} - e_{u_2})^2}{\sigma_{u_1}^2 + \sigma_{u_2}^2 + 2\sigma_p^2} \right), \tag{5}
\]

and that:

\[
SLR_{CS} = \frac{f((e_{u_1} - e_{u_2})^2 | H_{0CS})}{f((e_{u_1} - e_{u_2})^2 | H_{1CS})}, \tag{6}
\]

likelihood ratios (or at least calculated as such).

For example, consider a bloodstain recovered at a crime scene. A suspect, who has blood of the same type as the bloodstain, is considered. Clearly, the information that the type of the blood recovered at a crime scene is the same as the one of the suspect will be a lot more helpful to support the inference that the blood comes from the suspect if the blood type is AB$^{-}$ (less than 1% of the population) than if the blood type is O$^{+}$ (approx. 40% of the population). Yet, under the defence proposition, the model described above only assigns a probability to the event that the two blood types correspond by chance without accounting for the specific type of the blood at the crime scene.
The results of the comparison of Equations (4) and (6) using our toy example are presented in Figure 3. To study the convergence of $LR_{SS}$ and $SLR_{CS}$, we set $e_{u_1} = e_{u}$, $e_{u_2} = e_{s}$, $\sigma^2_{u_1} = \sigma^2_{u}$ and $\sigma^2_{u_2} = \sigma^2_s$. In both models, $\mu = 10$, $\sigma^2_d = 10$ and $\sigma^2_u = 2$. All simulations were repeated 1,000 times. In the first experiment, the characteristics of the donor of $e_s$ were chosen to be relatively common with respect to the population of donors ($\mu_d = 9$) but also quite variable ($\sigma^2_s = 1$). In the second experiment, the characteristics of the donor of $e_s$ were chosen to be rare with respect to the population of donors ($\mu_d = 0$) with a virtually negligible variability ($\sigma^2_s = 10^{-5}$).

The same general conclusions drawn from the data presented in Figure 1 can be reached when observing the data presented in Figure 3. Figure 3 shows that the $SLR_{CS}$ have a marked tendency to overestimate their $LR_{SS}$ counterparts, which may not necessarily be a problem when $H_{0 SS}$ is true; however, the use of $SLR_{CS}$ to report forensic evidence in court may be very detrimental to innocent suspects. We also note the particular behaviour of the relationship between $LR_{SS}$ and $SLR_{CS}$ when the variance of the control impressions is very small.

Overall, common source score-based models may be convenient to implement, but are not relevant to most examinations of forensic interest, and do not converge to the weight of forensic evidence. The lack of convergence between $LR_{SS}$ and $SLR_{CS}$ is not only a by-product of the use of a potentially non-sufficient summary statistic as the score between a pair of impressions, but also results from the inadequacy of Equation (6) to account for the rarity of the characteristics observed on the trace impression.
Curse of Statistics: “Score-based likelihood ratio”

Fig. 3. Comparisons between SLRs in the common source scenario with the LR in the specific source scenario. Columns: the left column reports the results when e_u and e_s have been sampled under $H_0$, the right column reports the results under $H_1$. Rows: (a) the source of the control impression is common and has some variance; (b) the source of the control impression is rare and has virtually no variance.

3.3. Suspect-centred score-based models

A second type of score-based model was proposed to be more relevant to the case at hand (Hepler et al., 2012; Alberink et al., 2014). According to this type of model, the sampling distributions of $\delta(e_u, e_s)$ can be represented by the following thought experiments:

(a) When the prosecutor proposition is correct, $\delta(e_u, e_s)$ is a score that is calculated by comparing a trace and a control object that have been both obtained from the source of $e_s$. Thus, the sampling distribution of $\delta(e_u, e_s)$ under $H_0$ can be studied by sampling, and comparing, independent pairs of trace and control objects from the source of $e_s$.

(b) When the defence proposition is correct, $\delta(e_u, e_s)$ is a score that is calculated by comparing trace objects sampled from random sources in a relevant population to control objects from the source of $e_s$.

This may seem counterintuitive, and the reader may wonder why trace objects, rather than
This type of score-based model has been designed to address the specific source pair of propositions since it is “anchored” on the control material obtained from the putative source. Nonetheless, it does not address the issue of the rarity of the trace characteristics as it only estimates the probability of the control material using a sample of trace objects from the population. Furthermore, it may not be trivial to repeatedly sample trace and control objects from sources under controlled conditions (i.e., repeatedly resampling fingerprints from an uncooperative suspect may be tricky). To overcome the latter issue, authors have proposed to generate pseudo-fingerprints (Neumann et al., 2012; Rodriguez et al., 2012) or the use of parametric models for the score distributions (Egli et al., 2006; Egli-Anthonioz and Champod, 2014). To avoid repeatedly sampling control impressions from the donor of \( e_s \), it is possible to condition the score-based model on \( e_s \). The difference between the unconditioned suspect-centred score-based model described in the previous paragraph and the conditioned model is that \( \delta(e_u, e_s) \) and both sampling distributions use the same fixed \( e_s \). Mathematically:

\[
\begin{align*}
SLR_{SS|e_s} &= \frac{f(\delta(e_u, e_s) \mid H_0) f(e_s \mid H_0)}{f(\delta(e_u, e_s) \mid H_1) f(e_s \mid H_1)} \\
&= \frac{f(\delta(e_u, e_s)|e_s, H_0) f(e_s \mid H_0)}{f(\delta(e_u, e_s)|e_s, H_1) f(e_s \mid H_1)}.
\end{align*}
\]

The second ratio in Equation (7) cancels out since the characteristics of the control material have the same density irrespective of whether the source of \( e_s \) is also the source of the trace material. From the generative models proposed in Equation (2), and with \( \delta(e_u, e_s) = (e_u - e_s)^2 \), we obtain the following sampling distributions for \( \delta(e_u, e_s) \):

\[
\begin{align*}
(e_u - e_s)^2 \mid e_s, H_0 &\sim \frac{1}{\sigma_u^2} \chi^2 \left( \frac{(e_u - e_s)^2}{\sigma_u^2} , \lambda = \frac{(\mu - e_s)^2}{\sigma_u^2} \right) \\
(e_u - e_s)^2 \mid e_s, H_1 &\sim \frac{1}{\sigma_u^2 + \sigma_p^2} \chi^2 \left( \frac{(e_u - e_s)^2}{\sigma_u^2 + \sigma_p^2} , \lambda = \frac{(\mu - e_s)^2}{\sigma_u^2 + \sigma_p^2} \right).
\end{align*}
\]

These sampling distributions enable us to compare the likelihood ratio of interest, \( LR_{SS} \), with its proxy, \( SLR_{SS|e_s} \). This comparison is reported in Figure 4 using the same parameter values as those used to generate the results presented in Figures 1 and 3.

The model proposed in Equation (7) certainly seems to be a reasonable ad-hoc solution: it is specific to the suspected source; the required sampling/simulation of trace objects from relevant sources can be achieved by using a suitable parametric model of the score distributions; furthermore, under the reasonable assumption, for some evidence types, that control objects have very limited variance, it will mostly converge to the specific source likelihood ratio of interest (Figure 4(c)). Unfortunately, this type of model is plagued by a fundamental lack of coherence: indeed, with these models, a given piece control objects, are randomly sampled from sources in the relevant population. This sampling model is rooted in the definition of the generative model in Equation (2): in the specific source scenario, there is no dispute that \( e_s \) originates from the suspected source.

15 Note that the results presented in Figure 4 are highly dependent of the value chosen for \( e_s \). In particular, the patterns in Figures 4(a) and (b) are very sensitive to the value of \( e_s \).
Fig. 4. Comparisons between SLRs conditioned on $e_s$ in the specific source scenario with the LR in the specific source scenario. Columns: the left column reports the results when $e_u$ and $e_s$ have been sampled under $H_{0SS}$; the right column reports the results under $H_{1SS}$. Rows: (a) the source of the control object is common and has some variance; (b) the source of the control object is rare and has some variance; (c) the source of the control object is common and has virtually no variance.
of evidence can provide support for either of the alternative propositions, depending on which proposition is considered first.

To demonstrate this lack of coherence, consider a model designed to address the two following specific source propositions:

- $H_A$: $e_u$ originates from Source A;
- $H_B$: $e_u$ originates from Source B.

We observe the trace object as well as two control objects, one from Source A and one from Source B. The specific source generative models under $H_A$ and $H_B$ are described below.

Under $H_A$, we have:

- $e_a = \mu_a + a$, where $a \sim N(0, \sigma_a^2)$;
- $e_b = \mu_b + b$, where $b \sim N(0, \sigma_b^2)$.

Under $H_B$, we have:

- $e_a = \mu_a + a$, where $a \sim N(0, \sigma_a^2)$;
- $e_b = \mu_b + b$, where $b \sim N(0, \sigma_b^2)$;
- $e_u = \mu_b + u$, where $u \sim N(0, \sigma_u^2)$.

The specific source likelihood ratio that addresses $H_A$ and $H_B$ is:

$$LR_{SS} = \frac{f(e_u, e_a, e_b|H_A)}{f(e_u, e_a, e_b|H_B)} = \frac{f(e_u|H_A)}{f(e_u|H_B)} = \left( \frac{f(e_u|H_B)}{f(e_u|H_A)} \right)^{-1}. \quad (9)$$

Thus, $LR_{SS}$ coherently supports the same proposition irrespective of which one is considered first. However, the specific source $SLR_{SS}$ conditioned on the control impression considered by the first proposition is:

$$SLR_{SS|e_s} = \frac{f(\delta(e_u, e_s)|e_s, H_A)}{f(\delta(e_u, e_s)|e_s, H_B)} = \left( \frac{f(\delta(e_u, e_s)|e_s, H_B)}{f(\delta(e_u, e_s)|e_s, H_A)} \right)^{-1}. \quad (10)$$

Equation (10) shows that $SLR_{SS|e_s}$ is not coherent in general since it potentially does not support the same proposition depending on which one is considered first. This lack of coherence can similarly be demonstrated for the unconditioned suspect-centred score-based likelihood ratio.

The conditioning of $\delta(e_u, e_s)$ on $e_s$ has an interesting geometric interpretation. When $e_s$ is fixed, all trace objects from the source considered under $H_{0SS}$ and from the sources from the population considered under $H_{1SS}$ are compared to the same control object. It is thus possible to consider that all scores considered in our thought experiment are equivalent to the scalar projections of the vectors representing all trace objects onto a vector space defined by $e_s$. Figure 5 illustrates this interpretation.

In Figure 5, a vector $e_u$, representing a trace object recovered in connection with a crime, is compared to objects from sources $A$ and $B$, represented by two mean vectors, $e_a$ and $e_b$. The left panel shows the orthogonal projection of $e_u$ onto $e_a$ and $e_b$. The resulting scalars are the scores calculated by $\delta(e_u, e_a)$ and $\delta(e_u, e_b)$, which are, in this case, equivalent to $\kappa(e_u, e_a)$ and $\kappa(e_u, e_b)$.
Fig. 5. Left panel: vector projection of $e_a$ onto $e_a$ and $e_b$. Middle panel: vector projection of all (pseudo-)trace objects, including $e_u$, onto $e_a$. Right panel: vector projection of all (pseudo-)trace objects, including $e_u$, onto $e_b$.

The middle panel shows the orthogonal projections of (1) $e_u$, (2) multiple pseudo-trace objects sampled from the source of $e_b$ (represented by dots near the tip of $e_b$), and (3) multiple pseudo-trace objects sampled from the source of $e_a$ (represented by dots near the tip of $e_a$) onto $e_a$. The two density functions represent the distributions of the scalar projections of these vectors onto $e_a$. According to the middle panel, $SLR_{SS|e_a}$, in Equation (10), is equivalent to the ratio of the likelihoods of the scalar projection of $e_u$ onto $e_a$ evaluated using the two distributions of the scalar projections of the trace objects from both suspects onto $e_a$. We see that, in this case, $SLR_{SS|e_a}$ would support the proposition that $e_u$ was made by Source B. The right panel shows the same information as in the middle panel, but this time, projected onto $e_b$. We see that in this case, $SLR_{SS|e_b}$ would support the proposition that $e_u$ was made by Source A. This geometric interpretation holds in the general case, when the alternative hypothesis is not specific to a single donor, but considers a population of sources as in Section 2.2. In the general case, all pseudo-trace objects from all sources of the relevant population are projected onto the vector representing a single source under $H_{0,SS}$ or under $H_{1,SS}$. This results in the same lack of coherence in the support of the evidence for alternative propositions representing different sources.

3.4. Trace-centred score-based models

The two types of models presented above lack the ability to account for the rarity of the characteristics of the trace object, which is crucial to properly quantify the weight of forensic evidence. To remedy this shortcoming, the use of trace-centred score-based models is found in Alberink et al. (2014). This type of model is somewhat similar to the family of suspect-centred models. However, since it is not possible to sample more trace and control objects from the true source of $e_u$ (since it is unknown), these models must
be conditioned on the observed trace, \( e_u \). Mathematically, we have:

\[
SLR_{SS|e_u} = \frac{f(\delta(e_u, e_s), e_u|H_{0SS})}{f(\delta(e_u, e_s), e_u|H_{1SS})} = \frac{f(\delta(e_u, e_s)|e_u, H_{0SS}) f(e_u|H_{0SS})}{f(\delta(e_u, e_s)|e_u, H_{1SS}) f(e_u|H_{1SS})} = \frac{f(e_u|H_{0SS})}{f(e_u|H_{1SS})}.
\]

Interestingly, the second ratio in Equation (11) does not cancel out. Indeed, the likelihood of observing the trace object, \( e_u \), is very different under \( H_{0SS} \) and \( H_{1SS} \). In fact, this ratio corresponds exactly to the likelihood ratio of interest presented in Equation 4. On the contrary, the first ratio, which includes the score, does cancel since \( e_u \) is fixed under both propositions due to conditioning, and \( e_s \) has the same distribution under both propositions in the specific source scenario. This can be seen when using the generative model in Equation (2) with \( \delta(e_u, e_s) = (e_u - e_s)^2 \), which results in the same sampling distributions under both propositions:

\[
(e_u - e_s)^2|e_u, H_{0SS} \sim \frac{1}{\sigma_s^2} \chi^2 \left( \frac{(e_u - e_s)^2}{\sigma_s^2} \right), \lambda = \frac{(e_u - \mu_p)^2}{\sigma_p^2} ;
\]

\[
(e_u - e_s)^2|e_u, H_{1SS} \sim \frac{1}{\sigma_s^2} \chi^2 \left( \frac{(e_u - e_s)^2}{\sigma_s^2} \right), \lambda = \frac{(e_u - \mu_p)^2}{\sigma_p^2} .
\]

The results in Equations (11) and (12) may seem suspicious at first. Some readers may consider that, under \( H_{1SS} \), the sampling distribution should involve control objects from sources in the relevant population. However, \( H_{1SS} \) is very clear on the origins of \( e_s \); its source is undisputed and it is the same specific source considered in \( H_{0SS} \) (see Section 2.2 and Equation (3)). Geometrically, this type of model has a similar interpretation as the suspect-centred model. The trace-centred model can be understood as the projection of all control objects onto a vector space defined by \( e_u \). However, as mentioned above, the only control objects available in this type of model are control objects of the source of \( e_u \) under both alternative propositions. Therefore, the first ratio in \( SLR_{SS|e_u} \) will always be one. In conclusion, it appears that trace-centred score-based likelihood ratios are not very useful.

3.5. Asymmetric score-based models

A last type of models, which historically happened at the very early stages of the development of score-based likelihood ratios and seems to be the most prevalent in the literature (Champod et al., 2001; Meuwly and Drygajlo, 2001; Meuwly, 2006; Egli et al., 2006; Gonzalez-Rodriguez et al., 2003, 2005, 2006; Neumann et al., 2007; Neumann and Margot, 2009), focuses on the putative source in its numerator and on some measure of the rarity of the characteristics of the trace object in the denominator.

According to this type of model, the sampling distributions of \( \delta(e_u, e_s) \) can be represented by the following thought experiments:

(a) When the prosecutor proposition is correct, \( \delta(e_u, e_s) \) is a score that is calculated by comparing a trace and a control object that have been both obtained from the source of \( e_s \). Thus, the sampling distribution of \( \delta(e_u, e_s) \) under \( H_0 \) can be studied
by sampling, and comparing, independent pairs of trace and control objects from the source of $e_s$. This experiment is similar to the one described for the numerator of the suspect-centred score-based likelihood ratio in Section 3.3. It may or may not be conditioned on $e_s$.

(b) When the defence proposition is correct, $\delta(e_u, e_s)$ is a score that is calculated by comparing the observed trace objects to a random control object in the population of potential sources. Thus, the sampling distribution of $\delta(e_u, e_s)$ under $H_1$ can be studied by sampling control objects from the sources in the population of potential sources and comparing them to $e_u$. This experiment is somewhat similar to the one that is done to address the denominator of the common source score-based likelihood ratio, with the modification that control objects are sampled from the population of sources (instead of trace objects in Section 3.2) and that the denominator is conditioned on the observed trace, $e_u$.

Mathematically, the asymmetric score-based likelihood ratio could be represented as:

$$SLR_{ASY} = \frac{f(\delta(e_u, e_s), e_s | H_{0SS})}{f(\delta(e_u, e_s), e_u | H_{1SS})} \cdot \frac{f(e_s | H_{0SS})}{f(e_u | H_{1SS})}. \quad (13)$$

While the first ratio in the right-hand part of Equation 13 seems appealing, at first, in the sense that it is both suspect and trace anchored, it is clear from Equation 13 that the ratio does not consider the same evidence in the numerator and in the denominator, which is a logical violation of the concept of likelihood ratio. Furthermore, the second ratio in the right-hand part of Equation 13 does not cancel. Thus, we do not see that $SLR_{ASY}$ can possibly converge to the specific source Bayes factor of interest.

4. Discussion and conclusion

Various attempts have been made to quantify the weight of fingerprint evidence. Most of these attempts suffer from severe shortcomings, which result in unpredictable bias with respect to the Bayesian inference framework. Some of these shortcomings include addressing the common source scenario instead of being specific to a suspected donor, failing to account for the rarity of the features observed on the latent impression, or providing incoherent evidence which may support both of two mutually exclusive propositions.

A Bayes factor is the ratio between two probabilities. Following Good (1950), Jeffreys (1961), Savage (1972), Jaynes (2003), Lindley (2006) and many others (for a recent review see (Taroni et al., 2016)), we took the view throughout this paper that probabilities can only represent the degree of belief of an individual about an event and are influenced by the information that he has about the event. Two individuals considering a particular event from two different perspectives may very well have different degrees of belief about that event. Thus, probabilities are subjective in the sense that they represent the personal relationship between the subject and the event.

The Bayes factor is not an intrinsic property of the evidence in itself, and we want to be very clear that we do not claim that there is such thing as a true or universal Bayes factor for a given piece of evidence. Different scientists may assign different weights
to forensic evidence if they characterise the evidence material using different types of features or measure it using different analytical techniques if they organise the data in different ways or if they choose different models to represent the data.

Nonetheless, the adjectives subjective or personal are not meant to suggest, or justify, that probabilities can be assigned arbitrarily, or reflect sloppy thinking. Bayes factors have fundamental properties, which should be satisfied by any method designed to quantify the weight of forensic evidence. These general properties are applicable to any model designed to assign Bayes factors.

By definition, a Bayes factor provides two pieces of information: which one of the two competitive propositions is favoured by the evidence, and the amount of support provided by the evidence. While it may be relatively easy to test the accuracy of a probabilistic model in large scale simulation settings (i.e., whether the correct proposition is consistently favoured), determining the appropriateness of the amount of support is an open problem.

Concentrating on the accuracy of a probabilistic model is arguably equivalent to considering the method as a deterministic decision-making engine with known error rates. The use of this type of techniques to infer the source of forensic traces has been explicitly discouraged by some but encouraged by others. We agree with in that error rates are only an average measure of performance over a population and do not provide information regarding the support of the evidence in individual cases.

Accuracy does not inform on whether a particular method supports a given proposition with the appropriate magnitude. Yet, in the legal context, the magnitude of the Bayes factor is critical. Grossly under- or overestimating the weight of the evidence can seriously distort the fact-finding process and be prejudicial to the accused. Some authors have proposed methods to study the magnitude of the values outputted by probabilistic models. We believe that some of these methods have merit and we will discuss them in a future paper. However, none

For example, glass fragments may be characterised by their refractive index or by their elemental composition.

Neumann et al. describe a method to characterise the spatial relationships between fingerprint landmarks (i.e., minutiae) using triangles and used these triangles to assign probability distributions to minutiae constellations. However, it is certainly possible to characterise the spatial relationships between minutiae in many other ways.

Consider that a partial fingerprint is recovered at a crime scene and is compared to the friction ridge skin of an accused. A jury will perceive the probative value of the evidence differently and may reach different conclusions if the reported Bayes factor is one thousand, or one billion. Depending on the case circumstances, the defence may be able to argue that the other evidence against the defendant is sufficiently weak that a Bayes factor of one thousand is not sufficient to reach a conclusion beyond reasonable doubt. A similar argument will be excessively difficult to make if the reported Bayes factor for the forensic evidence turns out to be one billion.
of these methods answers the question of the appropriateness of these magnitudes or addresses the soundness of the scientific foundations of a given probabilistic model.

It is also always possible to consider that the magnitude of the value produced by the model is important as a rank statistic, but not as a value in itself. In this case, we fall within the realm of likelihoodist inference ([Royall](#) 1997; [Kaye](#) 2012), which is not necessarily compatible with Bayesian inference.

Therefore, our conclusion is that none of the score-based models proposed to date can be considered as suitable proxies of the Bayes factor of interest. We are not arguing that these ad-hoc methods are not useful in their own way, but the harsh reality is that if one wants to abide by the idea that forensic evidence should be reported within a Bayesian paradigm, then one cannot use score-based likelihood ratios. We appreciate that the use of scores may be the only viable method to reduce the complexity of forensic evidence, but more efforts should be spent in the development of more rigorous models for handling these scores (see [Armstrong et al.](#) 2017; [Ausdemore et al.](#) 2019; [Hendricks et al.](#) 2019) for some early work on these models).

**References**

Aitken, C. and F. Taroni (2004). *Evaluation of Evidence for Forensic Scientists* (2nd ed.). Wiley and Sons Ltd, Chichester.

Alberink, I., A. de Jongh, and C. Rodriguez (2014). Fingermark evidence evaluation based on automated fingerprint identification system matching scores: The effect of different types of conditioning on likelihood ratios. *Journal of Forensic Sciences* 59, 70–81.

Armstrong, D., C. Neumann, C. Saunders, D. Gantz, J. Miller, and D. Stoney (2017). Kernel-based methods for source identification using very small particles from carpet fibers. *Chemometrics and Intelligent Laboratory Systems* 160, 99–209.

Ausdemore, M., C. Neumann, C. Saunders, D. Armstrong, and C. Muehlethaler (2019). Two-stage approach for the inference of the source of high-dimension and complex chemical data in forensic science. Technical report. https://arxiv.org/abs/1804.01049.

Biedermann, A., S. Bozza, and F. Taroni (2008). Decision theoretic properties of forensic identification: underlying logic and argumentative implications. *Forensic Science International* 177, 120–132.

Bozza, S., F. Taroni, R. Marquis, and M. Schmittbuhl (2008). Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *Journal of the Royal Statistical Society C.* 57, 329–341.

Brummer, N. and J. du Preez (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language* 20, 230–275.

Champod, C. (2015). Fingerprint identification: advances since the 2009 National Research Council report. *Philosophical Transactions of the Royal Society B* 370, 123–132. [http://dx.doi.org/10.1098/rstb.2014.0259](http://dx.doi.org/10.1098/rstb.2014.0259)
Champod, C., I. Evett, and B. Kuchler (2001). Earmarks as evidence: a critical review. *Journal of Forensic Sciences* 46, 1275–1284.

Egli, N., C. Champod, and P. Margot (2006). Evidence evaluation in fingerprint comparison and automated fingerprint identification systems - modelling within finger variability. *Forensic Science International* 176, 189–195.

Egli-Anthonioz, N. and C. Champod (2014). Evidence evaluation in fingerprint comparison and automated fingerprint identification systems - modeling between finger variability. *Forensic Science International* 235, 86–101.

European Network of Forensic Science Institutes (ENFSI) (2016). *ENFSI Guideline for Evaluative Reporting in Forensic Science*. European Network of Forensic Science Institutes (ENFSI). http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

Evett, I. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice* 38(3), 1198–202.

Evett, I., C. Berger, J. Buckleton, C. Champod, and G. Jackson (2017). Finding the way forward for forensic science in the USA: Commentary on the PCAST report. *Forensic Science International* 278, 16–23.

Forbes, P., S. Lauritzen, and J. Moller (2014). Fingerprint analysis with marked point processes. Technical report. https://arxiv.org/abs/1407.5809.

Gonzalez-Rodriguez, J., A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language* 20, 331–355.

Gonzalez-Rodriguez, J., J. Fierrez-Aguilar, and J. Ortega-Garcia (2003). Forensic identification reporting using automatic speaker recognition systems. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03)*, 11–93.

Gonzalez-Rodriguez, J., J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia (2005). Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. *Forensic Science International* 155, 126–140.

Good, I. (1950). *Probability and the Weighting of Evidence*. Charles Griffin & Co., London, U.K.

Haraksim, R., D. Ramos-Castro, D. Meuwly, and C. Berger (2015). Measuring coherence of computer-assisted likelihood ratio methods. *Forensic Science International*, 123–132.

Hendricks, J., C. Neumann, and C. Saunders (2019). A ROC-based Approximate Bayesian Computation algorithm for model selection: application to fingerprint comparisons in forensic science. Technical report. https://arxiv.org/abs/1803.10121.
Curse of Statistics: “Score-based likelihood ratio”

Hepler, A., C. Saunders, L. Davis, and J. Buscaglia (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Science International* 219, 129–140.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford University Press.

Kaye, D. H. (2012). Likelihoodism, bayesianism, and a pair of shoes. *Jurimetrics* 53, 1–9.

Leegwater, A., D. Meuwly, M. Sjerps, P. Vergeer, and I. Alberink (2017). Performance study of a score-based likelihood ratio system for forensic fingerprint comparison. *Journal of Forensic Sciences*, 1–15.

Lindley, D. (2006). *Understanding Uncertainty*. John Wiley & Sons.

Meuwly, D. (2006). Forensic individualisation from biometric data. *Science and Justice* 46, 205–213.

Meuwly, D. and A. Drygajlo (2001). Forensic speaker recognition based on a bayesian framework and gaussian mixture modeling. *Proc. of Odyssey 2001 Speaker Recognition Workshop, Crete (Greece)*.

Morrison, G., D. Kaye, D. Balding, D. Taylor, P. Dawid, C. Aitken, S. Gittelson, G. Zadora, B. Robertson, S. Pope, M. Neil, K. Martire, A. Hepler, R. Gill, A. Jamieson, J. de Zoete, R. Ostrum, and A. Caliebe (2017). A comment on the PCAST report: Skip the match/non-match stage. *Forensic Science International* 272, e7–e9.

National Research Council (NRC) of the National Academies (2009). *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, D.C.

Neumann, C. (2019). Defence Against the Modern Arts: the Curse of Statistics - FRStat. Technical report. https://arxiv.org/abs/1908.01408.

Neumann, C., C. Champod, R. Puch-Solís, N. Egli, A. Anthonioz, and A. Bromage-Griffiths (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences* 52, 54–64.

Neumann, C., C. Champod, M. Yoo, T. Genessay, and G. Langenburg (2015). Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingerprints. *Forensic Science International* 248, 154–171.

Neumann, C., I. Evett, and J. Skerrett (2012). Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm. *Journal of the Royal Statistical Society (Series A)* 175, 1–26.
Neumann, C. and P. Margot (2009). New perspectives in the use of ink evidence in forensic science: Part iii: Operational applications and evaluation. *Forensic Science International* 192, 29–42.

Ommen, D., C. Saunders, and C. Neumann (2017). The characterization of Monte Carlo errors for the quantification of the value of forensic evidence. *Journal of Statistical Computation and Simulation* 87(8), 1608–1643.

President’s Council of Advisors on Science and Technology (PCAST) (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Washington, D.C.: Executive Office of the President’s Council of Advisors on Science and Technology.

Ramos, D. and J. Gonzalez-Rodriguez (2013). Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International* 230, 156–169.

Ramos, D., J. Gonzalez-Rodriguez, G. Zadora, and C. Aitken (2013). Information-theoretical assessment of the performance of likelihood ratio computation methods. *Journal of Forensic Sciences* 58(6), 1503–1518.

Rodriguez, C., A. de Jongh, and M. D. (2012). Introducing a quick and simple approach to simulate large numbers of crime scene fingermarks for research on fingerprint identification. *Journal of Forensic Sciences* 57, 334–342.

Ross, A., K. Nandakumar, and A. Jain (2006). *Handbook of Multibiometrics*. International Series on Biometrics. Springer.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall.

Savage, L. (1972). *The Foundations of Statistics* (2nd ed.). Dover Publications, Inc. New York.

Swofford, H., A. Koertner, F. Zemp, M. Ausdemore, A. Liu, and M. Salyards (2018). A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation. *Forensic Science International* 287, 113–126.

Tackett, M. (2018). *Creating Fingerprint Databases and a Bayesian Approach to Quantify Dependencies in Evidence*. Ph. D. thesis, University of Virginia.

Taroni, F., S. Bozza, A. Biedermann, and C. Aitken (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk* 15, 1–16.