# Likelihood Ratio as Weight of Forensic Evidence: A Closer Look

**Steven P. Lund and Hari Iyer**

Statistical Engineering Division, Information Technology Laboratory,
National Institute of Standards and Technology,
Gaithersburg, MD 20899, USA

steven.lund@nist.gov
hari@nist.gov

The forensic science community has increasingly sought quantitative methods for conveying the weight of evidence. Experts from many forensic laboratories summarize their findings in terms of a likelihood ratio. Several proponents of this approach have argued that Bayesian reasoning proves it to be normative. We find this likelihood ratio paradigm to be unsupported by arguments of Bayesian decision theory, which applies only to personal decision making and not to the transfer of information from an expert to a separate decision maker. We further argue that decision theory does not exempt the presentation of a likelihood ratio from uncertainty characterization, which is required to assess the fitness for purpose of any transferred quantity. We propose the concept of a lattice of assumptions leading to an uncertainty pyramid as a framework for assessing the uncertainty in an evaluation of a likelihood ratio. We demonstrate the use of these concepts with illustrative examples regarding the refractive index of glass and automated comparison scores for fingerprints.

**Key words:** assumptions lattice; Bayes' factor; Bayes' rule; Bayesian decision theory; subjective probability; uncertainty; uncertainty pyramid.

## Executive Summary

In response to calls from the broader scientific community [1, 2] and concerns of the general public, experts in many disciplines of forensic science have increasingly sought to develop and use objective or quantitative methods to convey the meaning of evidence to others, such as an attorney or members of a jury. Support is growing, especially in Europe [3, 4], for a recommendation that forensic experts communicate their findings using a "likelihood ratio" (see Appendix A for an introduction to likelihood ratios). Proponents of this approach [5–11] appear to believe that it is supported by Bayesian reasoning, a paradigm often viewed as normative (*i.e.*, the *right* way; what someone *should* use) for making decisions when uncertainty exists [12–14].

Individuals following Bayesian reasoning may establish their personal degrees of belief regarding the truth of a claim in the form of odds (*i.e.*, ratio of their probability that the claim is true to their probability that the claim is false), taking into account all information currently available to them. Upon encountering new evidence, individuals quantify their "weight of evidence" as a personal likelihood ratio. Following Bayes' rule, individuals multiply their previous (or prior) odds by their respective likelihood ratios to obtain their updated (or posterior) odds, reflecting their revised degrees of belief regarding the claim in question. Because the likelihood ratio is subjective and personal, we find that the proposed framework in which a forensic expert provides a likelihood ratio for others to use in Bayes' equation is unsupported by Bayesian

decision theory, which applies only to personal decision making and not to the transfer of information from an expert to a separate decision maker, such as a juror.

Nevertheless, a likelihood ratio may be viewed as a potential tool for experts in their communications to triers of fact. If a likelihood ratio is reported, however, experts should also provide information to enable triers of fact to assess its fitness for the intended purpose. A primary concern should be the extent to which a reported likelihood ratio value depends on personal choices made during its assessment. Even career statisticians cannot objectively identify one model as authoritatively appropriate for translating data into probabilities, nor can they state what modeling assumptions one should accept. Rather, they may suggest criteria for assessing whether a given model is reasonable. We describe a framework that explores the range of likelihood ratio values attainable by models that satisfy stated criteria for reasonableness. The exploration of several such ranges, each corresponding to different criteria, provides the opportunity to better understand the relationships among interpretation, data, and assumptions. We propose the concept of a lattice of assumptions leading to an uncertainty pyramid as a framework for such an analysis.

Recent reports from the U.S. National Research Council and the President's Council of Advisors on Science and Technology [1, 2] primarily focus on the scientific validity of expert testimony, requiring empirically demonstrable error rates. In particular, they promote the value of "black-box" studies [15] in which practitioners from a particular discipline assess constructed control cases where ground truth is known (to researchers, but not the participating practitioners) as surrogates for casework in order to evaluate the collective performance of the discipline. Although we are primarily focused on the use of likelihood ratios, which these reports only tangentially consider, the concerns identified in this article also apply to subjectively selecting the pool of control scenarios required to estimate case-specific error rates. Practitioners adhering to Bayesian principles appear to consider likelihood ratio to be the only logical approach for expert communication, and they seek to implement its use in all forensic disciplines. We acknowledge that likelihood ratios provide a potential tool but emphasize that an extensive uncertainty analysis is critical for assessing when and how likelihood ratios should be used.

In the absence of an uncertainty assessment, likelihood ratios may still be useful as metrics for differentiating between competing claims when adequate empirical information is available to provide some meaning to the quantity offered by the expert. Free of normative claims requiring the use of likelihood ratios, forensic experts may openly consider what communication methods are scientifically valid and most effective for each forensic discipline.

## 1. Introduction

In criminal and civil cases alike, the judicial system involves many individuals making decisions after consideration of some form of evidence (*e.g.*, district attorneys deciding whether or not to file criminal charges, prosecution or defense attorneys deciding or advising their clients whether to accept a plea agreement or proceed to trial, jurors voting guilty or not guilty). These decision makers (DMs) often rely on the findings of forensic experts, whether expressed as a written report or through testimony at a trial, to help inform their decision. How experts express their findings and how DMs factor that information into their ultimate decisions remain areas of great public importance and current research; see, for example, in Refs. [16–17].

Lindley [18] presented a subjective Bayesian perspective for evaluating the weight of evidence[1] in forensic science. Within this framework, the odds form of the Bayes' rule, namely,

$$\text{Posterior Odds}_{\text{DM}} = \text{Prior Odds}_{\text{DM}} \times LR_{\text{DM}} \tag{1}$$

separates the ultimate degree of doubt a DM feels regarding the guilt of a defendant, as expressed via posterior odds (*i.e.*, probability of guilt after considering the evidence divided by probability of innocence

---

[1]The term "weight of evidence" appears in the book *Probability and the Weighing of Evidence* by I. J. Good [19], much earlier than Lindley's *Biometrika* paper [18]. In fact, Chapter 6 in this book is entirely devoted to weighing evidence.

after considering the evidence), into degree of doubt felt before consideration of the evidence at hand (prior odds) and the influence or weight of the newly considered evidence expressed as a likelihood ratio for the DM ($LR_{DM}$). A brief introduction to likelihood ratios is given in Appendix A. For a general exposure to the potential role of probability and statistics in the law, the reader may consult Fienberg [20], Dawid [21], and Kaye and Freedman [22].

In theory, the subjective Bayesian framework provides a uniquely rational and coherent[2] approach for an individual to make decisions in the presence of uncertainty. As such, it has garnered much attention among the statistical forensics community, with many scholars advocating that forensic experts summarize their findings by presenting their own personal *LR* to DMs, who could then apply (or envision others applying) Bayes' rule to modify their respective prior odds by the reported *LR* and arrive at their posterior odds as to the guilt or innocence of the defendant and choose actions accordingly (*e.g.*, a district attorney decides to file criminal charges, a juror decides to vote not guilty, etc.). This proposed hybrid adaptation can be expressed by the equation

$$\text{Posterior Odds}_{DM} = \text{Prior Odds}_{DM} \times LR_{Expert}. \tag{2}$$

See Aitken and Taroni [5] (chapter 3) or the European Network of Forensic Science Institutes (ENFSI) guidance document [4] for several examples illustrating how forensic examiners may use subjective probabilities to arrive at an *LR* value, which they can then use to convey to the DMs the strength of the evidence they examined. Furthermore, this guidance document also indicates that forensic examiners may convert the numerical *LR* value into a verbal equivalent following some scale of conclusions. Verbal expressions, however, cannot be multiplied by prior odds to obtain posterior odds.

The proclaimed appeal of the hybrid approach in Eq. (2) is that an impartial expert examiner could determine and convey the meaning of the evidence by computing a likelihood ratio (*LR*), while leaving strictly subjective initial perspectives regarding the guilt or innocence of the defendant to the DM. This adaptation has been embraced by many forensic scientists in several European countries and is currently being evaluated as a candidate framework for adoption in the United States. Kadane [23], Lindley [13], and others, however, clearly state that the *LR* in Bayes' formula is the personal *LR* of the DM due to the inescapable subjectivity required to assess its value.

Many researchers before us, privately and publicly, have considered whether or not it is appropriate to associate an uncertainty with an *LR* value offered as weight of evidence. The reader may refer to a special issue in *Science & Justice* [24] that is wholly devoted to this debate. Some of those who adhere to Bayesian decision theory have asserted that it is nonsensical to try to associate an uncertainty to an *LR* since its computation has already taken into account all the evaluator's uncertainty. Others who acknowledge sampling variability, measurement errors, and variability in choice of assumptions and choice of models have felt a need to express the effect of such variabilities on an *LR* value by offering an interval estimate (either a frequentist confidence interval or a Bayesian credible interval) or a posterior distribution.

Our paper explicitly identifies the swap from Eq. (1) to Eq. (2) as having no basis in Bayesian decision theory; this also applies to any related claims suggesting that the use of an *LR* to transfer knowledge from an expert to a DM is somehow normative. We further suggest it is necessary to conduct an uncertainty evaluation regarding the potential difference between $LR_{DM}$ and $LR_{Expert}$, requiring consideration of the range of results attainable under a wide-ranging and explicitly defined class of models. This is a broad and systematic view of uncertainty, for which limited sensitivity analyses or use of weighting tools such as Bayesian model averaging will generally be inadequate. Instead, we propose using an assumptions lattice and uncertainty pyramid to enable an audience to evaluate whether an *LR* characterization is well fitted for the intended purpose.

We begin by outlining the general steps required to theoretically evaluate an *LR*.

- The DM constructs a collection of scenarios to consider (*i.e.*, possible sequences of acts of those who may have been involved in the event that is the focus of the legal proceedings or its investigation).

---

[2]For a systematic introduction to the statistical meanings of "rational" and "coherent," see Lindley [13].

Constructing an *LR* requires partitioning this collection of considered scenarios into two sets. Suppose the DM is a juror who will cast a vote of either 'guilty' or 'not guilty' at the conclusion of a trial. The DM may assign any considered scenario to one of two categories, guilty and not guilty, according to how he or she would vote if that scenario were known to be exactly true. Suppose there are *a* mutually exclusive scenarios under which the DM would declare the defendant to be guilty. For notational convenience, we refer to this set as $H_p = \{H_{pi}\}_{i=1}^a$. Similarly, we refer to the collection of *b* mutually exclusive scenarios under which the DM would declare the defendant to be not guilty as $H_d = \{H_{dj}\}_{j=1}^b$.

- After sorting the set of considered scenarios, the DM assigns his or her (prior) degree of belief in each scenario before considering the totality of trial evidence, $E$. This is done by assigning a probability to each scenario such that the sum of all the probabilities is one. Let the probability assigned to scenarios $H_{pi}$ and $H_{dj}$ be denoted by $\pi_{pi}$ and $\pi_{dj}$, respectively. Denote the sum $\sum_{i=1}^a \pi_{pi}$ by $\pi_0$. Then the sum $\sum_{j=1}^b \pi_{dj}$ equals $1 - \pi_0$. Here, $\pi_0$ is the prior probability from the perspective of the DM that the defendant is *guilty*, and $1 - \pi_0$ is the corresponding prior probability that the defendant is *not guilty*. The conditional probability of scenario $H_{pi}$ given that the defendant is guilty is $w_{pi} = Pr[H_{pi}|H_p] = \dfrac{\pi_{pi}}{\pi_0}$. Similarly, the conditional probability of scenario $H_{dj}$ given that the defendant is not guilty is $w_{dj} = Pr[H_{dj}|H_d] = \dfrac{\pi_{dj}}{1 - \pi_0}$. Note that any scenario not explicitly given a positive prior weight is given a prior weight of zero, where a prior weight of zero indicates that the DM would never consider the scenario as plausible regardless of what data were presented. This hardline stance would seem more likely to be taken unintentionally or as a matter of convenience rather than conviction. (Here, convenience occurs from the fact that the entire collection of scenarios with an assigned prior weight of zero can be removed from further consideration to produce a manageable problem.) Additionally, even the most outlandish scenarios could become seemingly irrefutable, provided sufficient data. By this notion, it seems unlikely that any prior probability is rigid and exactly zero.

- For each scenario with nonzero prior weight, the DM is to assess the probability of the presented evidence $E$ occurring among all outcomes that could result from the described scenario. Let $L_{pi}$ denote the probability of observing the evidence under scenario $H_{pi}$. (Some may find it more natural to denote this quantity as $Pr[E|H_{pi}]$, but we will use $L_{pi}$ for succinctness.) Similarly, let $L_{dj}$ denote the probability of observing the evidence under scenario $H_{dj}$.

- Once a weight and a likelihood have been determined for each scenario of the observed evidence, the likelihood ratio is given as the sum of the products of the likelihood and the corresponding prior weight for each scenario in the *guilty set* divided by the sum of the products of the likelihood and the corresponding prior weight for each scenario in the *not guilty set*. This may be expressed algebraically as follows:

$$LR = \frac{\sum_{i=1}^a L_{pi} w_{pi}}{\sum_{j=1}^b L_{dj} w_{dj}}. \tag{3}$$

This formulation highlights that computing an *LR* is generally not free from prior probability assignment at the level of specific scenarios. The *LR* is insensitive to the redistribution of prior weights among scenarios that share a common likelihood within the *guilty set* (or within the *not guilty set*). In the context of source attribution, for instance, the DM may believe the alternative sources are a random sample from a particular population and not have any additional information that would lead to assigning different likelihoods among the alternative sources. In this instance, the DM might assign each alternative source a likelihood representing the probability of observing the evidence by random selection from that population, and the denominator becomes that same probability, regardless of what weights $w_{dj}$ would be chosen.

As a more concrete and narrowly focused example, suppose that evidence $y$ has been recovered from a crime scene and that, for simplicity, the DM is only interested in the identity of its source. Further suppose that, given which potential source actually produced $y$, there are no further relevant and unknown details from the perspective of the DM.[3] Let $S_0, S_1, \ldots, S_N$ denote the totality of potential sources, one of which is responsible for $y$. The actual source of $y$ is denoted by $S_q$, where $q$ is unknown. The source $S_0$ is of particular interest to the DM because it is attributed to the defendant. Thus, the primary proposition in question is

$$H_0 : \ S_0 \text{ is the source of } y \ (i.e., q = 0).$$

The complement of the proposition $H_0$ is $H_d = H_0^c$, given by

$$H_d : \ S_1 \text{ or } S_2 \text{ or } \ldots S_N \text{ is the source of } y \ (i.e., q \in \{1, 2, \ldots, N\}).$$

In addition to $y$, suppose one or more control samples (that is, samples from known sources) are available from one or more of the sources $S_j$, $j = 0, \ldots, N$. Denote these, collectively, by $x$.

Suppose $I$ denotes the totality of information available to the DM prior to being exposed to the information supplied by $y$ and $x$. According to the framework Lindley presents, a DM has prior probability $\pi_0 = Pr[H_0|I]$ for the proposition $H_0$ based on whatever information $I$ is available to him or her apart from $y$ and $x$. After being informed about the available new information $y$ and $x$, the DM would like to update his or her belief concerning $H_0$ in a *rational* and *coherent* manner.

The DM is interested in $Pr[H_0|y, x, I]$, the probability that $S_0$ is the source of $y$ given all the information available in the crime scene evidence ($y$), the control samples ($x$), and whatever else ($I$). Using the odds form of Bayes' rule, and following Lindley [18], Neumann *et al.* [26], and others, we get

$$LR = \frac{Pr[y|x, H_0, I]}{Pr[y|x, H_d, I]}. \tag{4}$$

In the context of this example, there is only one scenario under which $S_0$ is considered the source of $y$. Hence, the *LR* numerator requires only the conditional probability of $y$ given $x$, $H_0$, and $I$. Suppose this is denoted by $Pr[y|x, H_0]$. For simplicity of presentation, we have dropped the term $I$ with the proviso that all probabilities mentioned are *conditional on I*. Furthermore, it is to be understood that expressions such as $Pr[y]$ (or $Pr[y|x]$) refer to marginal (or conditional) probabilities or probability densities depending on whether $y$ is treated as discrete or continuous.

When the number of possible alternative sources is greater than one, evaluating the *LR* denominator, which corresponds to scenarios under which $S_0$ is not the source of $y$, is more complex. The proposition $H_d$ does not say anything about which of $S_1, \ldots, S_N$ is in fact the source. We can decompose $H_d$ as the union of the propositions $H_j$, $j = 1, \ldots, N$, where

$$H_j : S_j \text{ is the source of } y.$$

Because $H_d$ involves multiple scenarios, computing the *LR* denominator requires both a weight and conditional probability of $y$ given $x$ for each $H_j$. Suppose $\pi_0, \pi_1, \ldots, \pi_N$ are the *prior* probabilities, from the perspective of the DM, associated with the propositions $H_j$, $j = 0, 1, \ldots, N$, respectively. Then the denominator of the *LR* takes the form

$$Pr[y|x, H_d] = \sum_{j=1}^{N} w_j Pr[y|x, H_j],$$

---

[3]For example, if $y$ were a fingerprint, suppose the only relevant component of uncertainty to the DM is which person, or more specifically which finger, left the impression, or, if $y$ consisted of striation marks on a bullet fragment, suppose the DM is only concerned about identifying the gun from which the bullet was fired. For real situations involving multiple pieces of evidence and multiple experts, some forensic scientists suggest the use of Bayesian networks. See, for instance, Taroni *et al.* [25].

where $Pr[y|x,H_j]$ is the probability of $y$ given $x$ and $H_j$ (and $I$) and $w_j = \dfrac{\pi_j}{1-\pi_0}$. Thus, $w_j$ are the prior probabilities of the DM associated with $H_1, \ldots, H_N$, given $H_0$ is false.

Given the quantities $Pr[y|x,H_j]$, $j = 0, 1, \ldots, N$, and $\pi_0, \pi_1, \ldots, \pi_N$, the $LR$ corresponding to $H_0$ is computed as

$$LR = \frac{Pr[y|x,H_0]}{\sum_{j=1}^{N} w_j Pr[y|x,H_j]}.$$

## 1.1 List of Concerns

The recommendation that an individual substitute someone else's $LR$ for his or her own, as represented in Eq. (2), is indefensible, rather than normative, under the subjective Bayesian paradigm. Nevertheless, if it can be argued that $LR_{Expert}$ is sufficiently close to $LR_{DM}$, then such a substitution may be acceptable to the DM and fit for his or her purpose. However, there are many reasons why an $LR$ value offered by the expert may differ from that of the DM.

The following considerations are intended to highlight some of the more prominent subjective choices influencing the value of an $LR$.

### 1.1.1 Whose Scenarios?

According to the definition of the $LR$, any scenario given a nonzero prior probability by the DM can influence the value of the $LR$ and is therefore relevant; scenarios given a prior probability of zero cannot influence the value of the $LR$ regardless of the value of the corresponding likelihood $Pr[y|x,H_j]$, $j = 1, \ldots, N$, and are therefore irrelevant to the DM. Even if $Pr[y|x,H_j]$ is exactly known for any scenario proposed, the $LR$ still depends upon the collection of scenarios that are considered as well as the corresponding weights given to them by the DM, neither of which is known to the expert.

As in the source attribution example above, the set of sources with positive prior probability forms the *relevant population* for the DM. If there are several DMs, then each one could have their own set of weights $w_j$ and hence their own relevant population. Given a particular relevant population, the weights assigned to elements of that population can affect the $LR$ unless the assigned likelihoods are constant across all members of the population. In particular, $w_j = \frac{1}{N}$ is a special case, not a mandate. The question remains: How sensitive is the $LR$ value to any particular definition of a relevant population?

### 1.1.2 Whose Likelihoods?

In practice, probability functions $Pr[y|x,H_j]$ ($j = 0, 1, \ldots, N$) are rarely known in any authoritative sense. A forensic analyst will commonly begin with a prior distribution over a class of models that will then be updated by consideration of empirical data.[4] That is, crime scene data $y$ and control data $x$ are assumed to be conditionally independent, given the parameter $\theta$ and the event $H_j$, with known distributions $g(y|\theta,H_j)$ and $h(x|\theta,H_j)$ ($j = 0, 1, \ldots, N$), respectively. Given $H_j$, $\theta$ is assumed to have a distribution described by the probability function $f(\theta|H_j)$, which is used to express a prior belief about likelihood functions for $x$ and $y$ given $H_j$ (not to be confused with the prior $\pi_j$, which reflects prior belief in the proposition $H_j$). Hence, the joint distribution of $y$, $x$, and $\theta$, given the proposition $H_j$, is described by the probability function

$$a(y,x,\theta|H_j) = g(y|\theta,H_j)h(x|\theta,H_j)f(\theta|H_j). \tag{5}$$

---

[4]This framework includes Bayesian model averaging (BMA) (see Hoeting *et al.* [27]), whereby the DM specifies a collection of probability model families along with his or her personal probabilities attached to each model. Other DMs implementing BMA may choose differently, leading to different model averaging results. Thus, BMA does not remove the need to examine how assumptions affect uncertainty if it is to represent or inform interpretations of multiple individuals. Depending on what is considered to be a *reasonable* class of priors on the model space, the corresponding range of plausible $LR$ values may tend to be narrower when using BMA than otherwise.

The quantity $Pr[y|x,H_j]$ can be expressed as

$$Pr[y|x,H_j] = \frac{\int h(y,x,\theta|H_j)d\theta}{\int\int h(y,x,\theta|H_j)d\theta dy}. \tag{6}$$

Thus, the distribution of interest for source $j$, $Pr[y|x,H_j]$, has been *exactly* specified through the choice of $f$, $g$ and $h$. Asymptotically, as the number of control observations goes to infinity for each potential source $j = 0,1,\ldots,N$, the value of $Pr[y|x,H_j]$ may converge to the same answer for many different choices of $f$, $g$ and $h$. In real applications with finite data, however, subjective choices of $f$, $g$ and $h$ remain influential.

Support for particular choices of $f$, $g$ and $h$ is sometimes given by showing them to be consistent (as defined by some user-selected process for evaluating such things) with empirical data from similar situations. Even when all DMs agree on what data are appropriate to consider for the case at hand and the criteria to use in assessing whether or not a model is consistent with those data, multiple choices of $f$, $g$, and $h$ will satisfy that requirement. The question remains: How sensitive is the result to any particular modeling choice?

### 1.1.3 Approximation

When following a subjective Bayesian approach, one uses a definition of *personal* probability that could be viewed as an individual's assessment of a fair value for a bet of $H_0$ versus its complement. It is assumed that for any required probability, such a value exists and is unique, and that the individual is able to identify this value without any doubts. Some authors have considered the practical difficulties associated with precisely identifying fair values for bets, and this has led to the consideration of *imprecise probabilities*. For a systematic introduction to this topic see, for instance, Walley [28]. This field remains an active area of research (Augustin *et al.* [29]). Moreover, it is assumed that the collection of specified probabilities satisfies the requirement of *coherence* (*i.e.*, the standard rules of probability are obeyed). Lindley *et al.* [30] discussed the practical issues one must address in order to reconcile the generally incoherent probability assessments by an individual. They considered several different approaches that one could use in such a reconciliation process. See also Kadane and Winkler [31]. The fact that such reconciliation efforts are necessary points to uncertainties associated with subjective probability assessments. Nevertheless, results derived using such probability models are sometimes treated as free from uncertainties (see, *e.g.*, Taroni *et al.* [32]).

Computing an *LR* for anything but the simplest of problems will involve approximations. Rather than assign prior weights that exactly and genuinely reflect one's personal belief, tractable and familiar substitutions are made. In the absence of a rigorous uncertainty analysis demonstrating that the resulting value is sufficiently insensitive to such replacements, the computed value can only provide an approximation of unknown accuracy for the rational and coherent ratio between posterior and prior odds of the DM. Although any DM only needs to be personally satisfied regarding the suitability of using any given *LR* in Bayes' formula, guiding the probabilistic interpretation of others requires greater care.

We note that the considerations listed here are not addressed by explaining the assumptions that underlie a given statistical interpretation. Stating assumptions promotes *transparency*, enabling a trained audience to assess whether a presented analysis seems reasonable, much like a statistical hypothesis test. It does not, however, even begin to inform the range of results attainable under alternative analyses that may also be deemed reasonable, the analog of a statistical confidence interval. The *transferability* of an analyst's statistical interpretation (*i.e.*, its value as a surrogate for that of a DM) depends on its robustness across the set of analyses that the DMs would deem plausible. The book by Morgenthaler and Tukey [33] titled *Configural Polysampling: A Route to Practical Robustness* provides an interesting discussion of the need for considering multiple plausible models and emphasizes the development of robust methods of statistical analysis of data and approaches for assessing small sample robustness of statistical inference procedures.

To assess robustness in a systematic manner, an analyst must first define the space of models to be considered, possibly by providing an explicit *plausibility criterion*, so that robustness has a precise meaning.

When extensively characterizing uncertainty, justifying why models in the defined space are reasonable seems less important than justifying why models not in the defined space are unreasonable. The analyst then explores the corresponding range of attainable results by fitting multiple models from within the defined space.[5] In instances where this exploration is incomplete, the full range of plausible results, and thus the suitability of relying on any one particular interpretation, is unknown. To begin to explore the relationships among data, assumptions, and interpretations, we consider multiple assumption sets in a form we refer to as the *lattice of assumptions* and present the resulting ranges of *LR*s as an *uncertainty pyramid*. This approach is intended to encourage analysts to explicitly recognize and systematically evaluate the influence of their subjective modeling choices and is illustrated in the following section.

## 2. The Influence of Modeling Assumptions

We are concerned about seemingly innocuous modeling assumptions latently constraining the space of plausible interpretations as might be presented by a forensic expert. In this section, we demonstrate a process for evaluating the restrictive influence of unsubstantiated information that can creep in solely on the basis of distributional assumptions made by an analyst. It should be noted that the data and modeling approaches used in this section are not exhaustive and are not intended to represent analyses generally undertaken by any particular forensic practice. As such, the actual numerical results obtained in this section are not of primary interest. Our intention is to illustrate a process for assessing the influence of modeling assumptions on concrete examples.

Evaluating the influence of a given assumption set (say, assumption set *A*) requires considering the results of multiple analyses, one in which assumption set *A* was made and others in which different assumption sets (say, assumption sets $B_i, i = 1, 2, \ldots$), *each consistent with empirically observed data*, were made. The influence of assumption set *A* is reflected by the differences among the conclusions drawn upon evaluation of each set of results. In cases where the differences are considered to be substantial, assumption set *A* has played a critical role, and the conclusion reached from results of the analysis in which assumption set *A* was made stretches beyond what the data used in the analysis can in fact support. In such a case, it may be inappropriate to rely on any particular assumption set.

### 2.1 Illustration 1: Glass Example

In an illustrative example discussing the use of RI values in the interpretation of glass evidence, Evett [42] considers the following scenario. Suppose a window is believed to have been broken during the commission of a crime and fragments of glass are recovered from the crime scene. Suppose also that fragments of glass were found on a suspect.[6] Denote by $x_1, \ldots, x_m$ the RIs of the crime scene fragments (bulk sample) and by $y_1, \ldots, y_n$ the RIs of suspect fragments (receptor sample). The two propositions of interest are

$H_p$ : The receptor sample is from the same source as the bulk sample

$H_d$ : The receptor sample is from a different source than the bulk sample.

In Evett's example, $m = 10$ and $n = 5$, and the RI values of the corresponding glass fragments are given in Table 1.

---

[5]When a plausibility criterion pertains to a theoretical probability distribution used to model empirical data, the collection of all plausible models defines a region in the space of all cumulative distribution functions (CDFs). This region is sometimes referred to as a *Probability Box* (p-box, for short). See for instance, Williamson [34], Williamson and Downs [35], Hoffman and Hammonds [36], Ferson *et al.* [37], Zhang and Berleant [38, 39], Ferson and Siegrist [40], and references contained therein. These and other authors have investigated methods for propagating uncertainties when component distributions are specified in terms of p-boxes.

[6]The fact that fragments of glass were found on the defendant's clothes is in itself of evidential value. However, for our illustration, we focus only on the source question.

**Table 1.** Refractive index (RI) measurements from the window and from the suspect.
(Source: Evett [42])

| Measurement Location | RI | | | | |
|---|---|---|---|---|---|
| Measurements from the window | 1.51844 | 1.51848 | 1.51844 | 1.51850 | 1.51840 |
| | 1.51848 | 1.51846 | 1.51846 | 1.51844 | 1.51848 |
| Measurements from the suspect | 1.51848 | 1.51850 | 1.51848 | 1.51844 | 1.51846 |

### 2.1.1 Within-Source and Between-Source Distributions

Interpreting the information contained in the observed RIs regarding these two propositions requires understanding the distribution of RIs within each source and how that distribution varies from one source to the next. (Note that if the RI distribution did not vary across sources, then the RI observations would not provide any useful information about their source.) Considering how the RI distribution varies from one glass pane to the next results in a distribution of distributions. The collection of possible descriptions or models for the distribution of distributions is overwhelmingly vast. The tendency is to limit the class of potential descriptions by specifying properties of RI distributions that are assumed to remain constant from one window to the next. In particular, the RI distributions across glass panes are often assumed to be identical except for their location (*e.g.*, mean or median). That is, the RI distribution for every potential source is assumed to have exactly the same shape and exactly the same scale (or spread). Such a family of distributions is referred to as a *location family*. This assumption implies that the distribution for the difference between the RIs of each fragment within a glass pane and the median RI of all fragments from that glass pane (*i.e.*, $x - \text{median}(x)$) is exactly the same for any glass pane in the considered relevant population.

In general, the results of analyses (*e.g.*, *LR*) can be highly sensitive to deviations from the assumption that RI distributions differ only by their median from one glass pane to another. Generating empirical confidence in such a strong assumption would require collecting RI data from many windows with enough measurements from each window so as to convince oneself that strictly limiting the set of plausible distributions to a location family will have only a *negligible effect* on the interpretation of the analysis results compared to, for instance, when the shape and scale of the presumed location family are allowed to vary from one source to another. Even with such a vast and consistent data set, the possibility remains that the RI distribution of any unexamined window differs substantially from the observed characteristics of the other windows. Further illustration of the potential influence of assuming a location family on the interpretation of the observed RI from a particular case is beyond the scope of this paper. That is, the notion of uncertainty we portray in these examples is incomplete. The uncertainty resulting from a more complete examination is expected to be greater than what is illustrated here.

For the sake of simplicity, we proceed by supposing that the informed DM is willing to make the location family assumption. To compute an *LR* for this scenario, let us first introduce some notation. Suppose the cumulative distribution function (*CDF*) of RI values from any single window belongs to the location family of distributions $G(y; \theta) = G_0(y - \theta)$ for some continuous distribution with *CDF* $G_0$ for which the median value is zero. Denote the corresponding probability density function (*PDF*) by $g_0$. Furthermore, suppose that, across the (relevant) population of windows, the median RIs $\theta^{(j)}, j = 0, 1, \ldots, N$, are independently and identically distributed (iid) with an unknown *PDF* $f(\theta)$ and corresponding *CDF* $F(\theta)$. That is, we have assumed that $f(\theta|H_j) = f(\theta)$ for all $j = 0, 1, \ldots, N$. For completeness, we display

the expression for the resulting *LR* in Eq. (7).

$$LR = \frac{\int \left( \prod_{i=1}^{m} g_0(x_i - \theta) \prod_{j=1}^{n} g_0(y_j - \theta) \right) dF(\theta)}{\int \left( \prod_{i=1}^{m} g_0(x_i - \theta) \right) dF(\theta) \int \left( \prod_{j=1}^{n} g_0(y_j - \theta) \right) dF(\theta)}. \tag{7}$$

This example provides an illustration where there is no information available for us to justify assigning different likelihoods to each particular potential source. Hence, we consider the probabilities in the numerator and the denominator of the *LR* from the perspective of a population of windows rather than weighting likelihoods from individual windows according to their prior probability; see related comments following Eq. (3).

### 2.1.2 Illustrative Analyses

In the educational example provided in chapter 10 (section 10.4.2) of Aitken and Taroni [5], it is assumed that $g_0$ is the *PDF* of a normal distribution with a standard deviation equal to 0.00004. That is, RI values that could be observed from window $j$ are iid according to a normal distribution with unknown window-specific mean $\theta^{(j)}$, $j = 1, \ldots, N$, and known standard deviation $\sigma$ equal to 0.00004. The distribution of $\{\theta^{(j)}\}$ is modeled using data from Table 10.5 of Lambert and Evett [41], which gives the average RI measurements from 2269 different samples of float glass. These sample data are assumed to be representative of the mean RIs associated with sources $S_j$, $j = 0, 1, \ldots, N$, and the density $f$ (or the *CDF F*) is estimated via kernel density estimation (KDE), using a Gaussian kernel with varying bandwidths. The resulting estimates are then used to evaluate the *LR* corresponding to various hypothetical pairs of average RI measurements from the source (window) and receptor (suspect). See Table 10.6 of Aitken and Taroni [5].

### 2.1.3 Multiple Plausible Models

The consideration of multiple kernel bandwidths for estimating $f$ begins to illustrate the potential uncertainty due to the influence of modeling choices. A more complete evaluation may be obtained by considering how variable the computed *LR* is across the set of all combinations of $g_0$ and $f$ that might be considered plausible. The criteria for establishing the plausibility of a prosed model is personal and likely to vary from one person to the next. However, it is possible for the criteria of a specific individual to be expressed in an objective manner. Analogous to selecting prior distributions when conducting Bayesian inference, the choice of a plausibility criterion should not be guided by the set of *LR* values it permits, but by the information available before application to the case at hand. When criteria for plausibility have been established, the objective intention is to characterize the range of results attainable by any model meeting those criteria rather than identifying a single plausible model (or a narrow set of closely related models in the case of multiple kernel density estimates of $f$ obtained from different bandwidths) and proceeding as though it is the only plausible model or representative of all plausible models.

### 2.1.4 Goodness-of-Fit Tests and Plausibility Criteria

We note that it is common practice for a data analyst to use a statistical test of goodness-of-fit to assess plausibility of one or more models. In our example, the data modeler could assess the plausibility of a proposed distribution pair ($g_0$ and $f$), given sample data, using any of a number of goodness-of-fit statistical testing procedures. Some well-known methods are: (1) Kolmogorov-Smirnov (KS) test, (2) Cramer-von Mises test, and (3) Anderson-Darling test. For related other approaches the interested reader should also consult Owen [43], Frey [44], Liu and Tewfik [45], and Goldman and Kaplan [46]. The concept is the same for each criterion: the data sample itself cannot reduce the space of plausible models to a single *CDF*.

Here, we consider the KS test for illustrative purposes. Any other procedure can be used in place of the KS test, but the computations can be more challenging. The KS test leads to a confidence band consisting of a family of *CDF*s, each of which is consistent with the data at a prescribed level of confidence, say 95 %. When the KS test is used to assess plausibility, any *CDF* that lies entirely within the confidence band would be deemed plausible given the sample data. As the number of observations in the data set increases, the confidence band narrows, and the set of plausible distributions is reduced.

We now consider the influence of two data sets on plausible choices for $g_0$ and $f$, or, equivalently, *CDF*s $G_0$ and $F$.

### 2.1.5   Float Glass Data

The first data set (see page 16, Lambert and Evett [41]) contains a collection of average RI measurements obtained from various within-window samples collected from different manufactured pieces of float glass. The number of observations contained in each sample is not provided, so sample sizes may vary across the samples, and there is some uncertainty as to how these data should be viewed during evidence evaluation. If each sample contained a single observation, the KS confidence band might be used to restrict the marginal distribution of a single RI measurement obtained from a randomly selected window in the population. This marginal distribution is determined by the choice of $g_0$ and $f$ as $h(y) = \int g_0(y - \theta) dF(\theta)$. If the samples consisted only of means of many replicate observations, the KS bounds could serve to restrict the class of plausible choices for $f$, but would not provide much insight for the choice of $g_0$.

For illustrative purposes, we treat the data from this set as providing median RI values for a sample of 2269 windows representative of the relevant population. These data are displayed in Table 2.
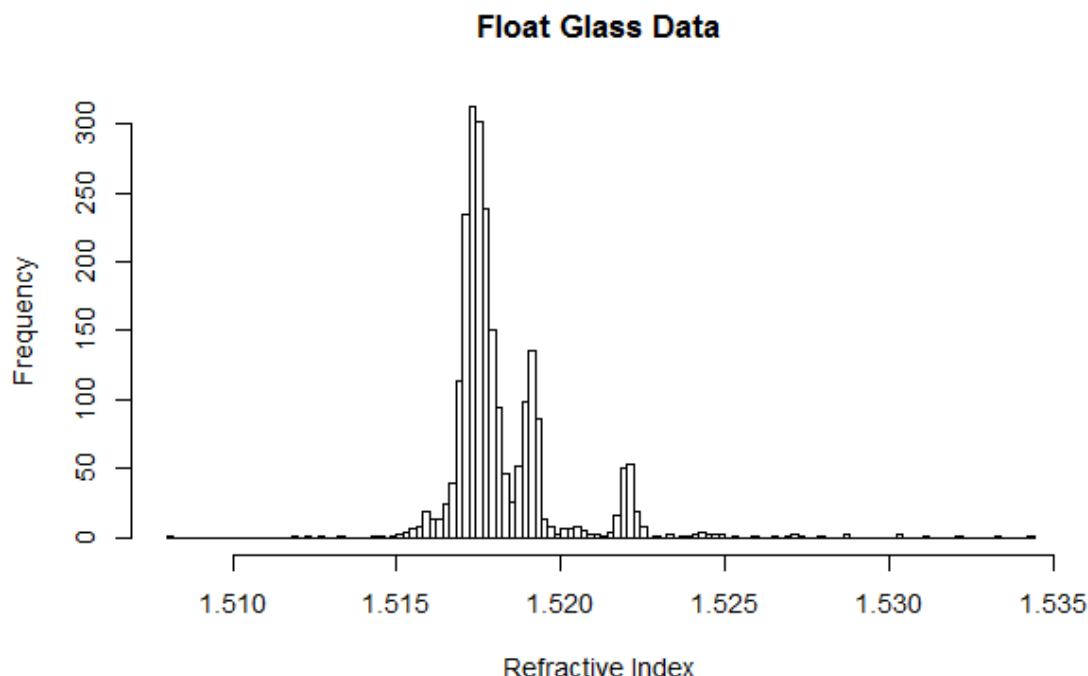


**Fig. 1.** Histogram of float glass data from Lambert and Evett [41].

11

**Table 2.** Refractive index (RI) measurements for 2269 glass fragments given in Lambert and Evett [41].

| RI | Count | RI | Count | RI | Count | RI | Count |
|---|---|---|---|---|---|---|---|
| 1.5081 | 1 | 1.5170 | 65 | 1.5197 | 7 | 1.5230 | 1 |
| 1.5119 | 1 | 1.5171 | 93 | 1.5198 | 1 | 1.5233 | 1 |
| 1.5124 | 1 | 1.5172 | 142 | 1.5199 | 2 | 1.5234 | 1 |
| 1.5128 | 1 | 1.5173 | 145 | 1.5201 | 4 | 1.5237 | 1 |
| 1.5134 | 1 | 1.5174 | 167 | 1.5202 | 2 | 1.5240 | 1 |
| 1.5143 | 1 | 1.5175 | 173 | 1.5203 | 4 | 1.5241 | 1 |
| 1.5146 | 1 | 1.5176 | 128 | 1.5204 | 2 | 1.5242 | 1 |
| 1.5149 | 1 | 1.5177 | 127 | 1.5205 | 3 | 1.5243 | 3 |
| 1.5151 | 1 | 1.5178 | 111 | 1.5206 | 5 | 1.5244 | 1 |
| 1.5152 | 1 | 1.5179 | 81 | 1.5207 | 2 | 1.5246 | 2 |
| 1.5153 | 1 | 1.5180 | 70 | 1.5208 | 3 | 1.5247 | 2 |
| 1.5154 | 3 | 1.5181 | 55 | 1.5209 | 2 | 1.5249 | 1 |
| 1.5155 | 5 | 1.5182 | 40 | 1.5211 | 1 | 1.5250 | 1 |
| 1.5156 | 2 | 1.5183 | 28 | 1.5212 | 1 | 1.5254 | 1 |
| 1.5157 | 1 | 1.5184 | 18 | 1.5213 | 1 | 1.5259 | 1 |
| 1.5158 | 7 | 1.5185 | 15 | 1.5215 | 1 | 1.5265 | 1 |
| 1.5159 | 13 | 1.5186 | 11 | 1.5216 | 3 | 1.5269 | 1 |
| 1.5160 | 6 | 1.5187 | 19 | 1.5217 | 4 | 1.5272 | 2 |
| 1.5161 | 6 | 1.5188 | 33 | 1.5218 | 12 | 1.5274 | 1 |
| 1.5162 | 7 | 1.5189 | 47 | 1.5219 | 21 | 1.5280 | 1 |
| 1.5163 | 6 | 1.5190 | 51 | 1.5220 | 30 | 1.5287 | 2 |
| 1.5164 | 8 | 1.5191 | 64 | 1.5221 | 25 | 1.5288 | 1 |
| 1.5165 | 9 | 1.5192 | 72 | 1.5222 | 28 | 1.5303 | 2 |
| 1.5166 | 16 | 1.5193 | 56 | 1.5223 | 13 | 1.5312 | 1 |
| 1.5167 | 15 | 1.5194 | 30 | 1.5224 | 6 | 1.5322 | 1 |
| 1.5168 | 25 | 1.5195 | 11 | 1.5225 | 3 | 1.5333 | 1 |
| 1.5169 | 49 | 1.5196 | 3 | 1.5226 | 5 | 1.5343 | 1 |

A histogram of these data is shown in Fig. 1. We use the median rather than the mean to reduce the sensitivity of the location parameter $\theta$ to the tails of the distribution $g_0$, which cannot be well estimated from sample data used in this example. Figure 2 shows the empirical *CDF* (*eCDF*) for these data along with the lower and upper boundaries of a KS 95 % confidence band used to define which choices for $f$ will be considered plausible given the *eCDF*. In the lattice of assumptions illustration, we consider several estimates for $f$ based on Gaussian kernel density estimates fit to the 2269 observations with bandwidths spanning from 0 (which corresponds to the *eCDF*) to $2.155 \times 10^{-4}$, which is the maximum bandwidth for which the corresponding discrete distribution obtained by accounting for the reported measurements being interval censored (to plus or minus $1 \times 10^{-4}$) remains entirely within the KS confidence band. *CDF*s for the discrete distributions obtained by accounting for interval censored measurements and the corresponding underlying continuous distributions are shown in Fig. 2 for both the *eCDF* and the smoothest kernel density estimate. Kernel density estimates resulting from the intermediate bandwidths of $10^{-5}$, $2 \times 10^{-5}$, $5 \times 10^{-5}$, and $10^{-4}$ were considered during computation but are not displayed. For illustration only, we also include a *CDF* not produced by kernel density estimation. This *CDF*, referred to as Jump, follows the lower KS bound for values less than the mean RI value $m = \dfrac{\sum_{i=1}^{10} y_i + \sum_{j=1}^{5} x_j}{15}$ for the 15 sample fragments, and the upper KS bound for values greater than $m$, with a jump at $m$. This *CDF* is shown in blue in Fig. 2. An analyst might feel that the jump distribution is unrealistic and should not be considered. Our point in including it is to emphasize that once a plausibility criterion has been laid down, we must attempt to consider as broad a collection of candidate distributions meeting the criterion as possible; if not, the plausibility criterion and corresponding uncertainty characterization become moving targets.
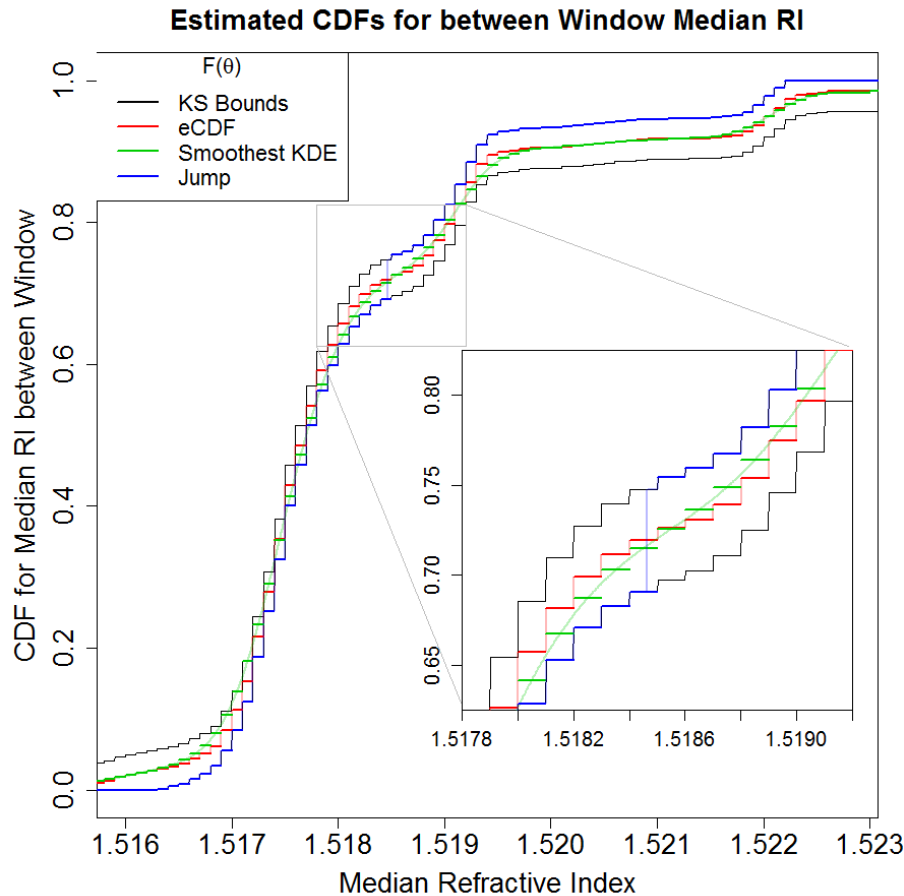
**Fig. 2.** 95% Kolmogorov-Smirnov Confidence Band for the Lambert and Evett Glass Data [41]. The bold line segments portray the discrete distribution obtained by accounting for the reported data being interval censored to $\pm$ 0.0001. The faded lines display the *CDF* of the underlying continuous distribution.

### 2.1.6   Bennett Data

The second data set consists of 49 RI measurements on samples of fragments from 49 different locations on a single window and is used to evaluate the plausibility of within-window distribution choices. These data were collected by Bennett *et al.* [47] and are also mentioned in Curran [48] (see page 42).[7] They are publicly available in the dafs package in R [49]. The original data set consists of RI measurements for a sample of 10 fragments from each of 49 locations on a single window pane for a total of 490 readings. We have selected a single fragment from each of the 49 locations (the listed value in the first row of the bennett.df data frame in dafs). These data are reproduced in Table 3 for the convenience of the reader. For illustrative purposes, we treat these 49 RI values as representative of the RI distribution within a single window, providing guidance for choosing $g_0$. The empirical *CDF* and corresponding KS 95 % confidence band for these 49 RI measurements are shown in Fig. 3.

---

[7]Although not explicitly mentioned in Bennett *et al.* [47], these data appear to be interval-censored with variable interval half-widths (approximately) equal to $1.5 \times 10^{-6}$. Consequently, all of our analyses based on these data take this interval-censoring into account.

**Table 3.** Refractive index (RI) measurements from 49 different locations from a single window. (Data from Curran [47])

| RI | | | | | | |
|---|---|---|---|---|---|---|
| 1.519788 | 1.519901 | 1.519941 | 1.519941 | 1.519941 | 1.519963 | 1.519970 |
| 1.519974 | 1.519974 | 1.519974 | 1.519974 | 1.519974 | 1.519978 | 1.519978 |
| 1.519978 | 1.519981 | 1.519981 | 1.519981 | 1.519981 | 1.519985 | 1.519989 |
| 1.519989 | 1.519992 | 1.519992 | 1.519996 | 1.519996 | 1.519996 | 1.519996 |
| 1.520000 | 1.520000 | 1.520003 | 1.520007 | 1.520007 | 1.520007 | 1.520007 |
| 1.520010 | 1.520010 | 1.520014 | 1.520014 | 1.520014 | 1.520014 | 1.520025 |
| 1.520025 | 1.520029 | 1.520040 | 1.520043 | 1.520047 | 1.520047 | 1.520069 |

In the lattice of assumptions, we consider several distributional shapes, including normal distributions, $t$ distributions with 1 and 0.5 degrees of freedom, respectively, and $\chi^2$ distributions with 2 and 3 degrees of freedom. We also consider a small simulated collection of $CDF$s not belonging to any particular parametric family. Some of these $CDF$s fulfill additional constraints of unimodality and/or symmetry. For each considered distributional shape, we identify the range of scale parameters such that the discrete distribution obtained by accounting for the interval-censoring of the reported within-window measurements is contained entirely within the confidence band. For each shape, we consider estimates of $g_0$ obtained at 15 evenly spaced scale values spanning this range. For a given shape and scale parameter, the $LR$ is evaluated for each pairing of $g_0$ with each of the choices for $f$ described above. Figure 3 provides a visual summary of the analysis when $g_0$ is assumed to have the shape of a normal distribution. Analogous displays for subsets of other considered shapes are provided in Appendix B.

### 2.1.7 Assumptions Lattice

When modeling the distribution of RI values for fragments from any single window, Lindley [18] assumed normality, as did Aitken and Taroni [5]. We recognize this was done for illustrative purposes only. Nevertheless, it is worth noting that normal distributions represent a tiny fraction of $CDF$s meeting the KS criteria, and the impact of exclusively assuming a normal distribution is not clear until the sets of $LR$ values obtainable by using other distributions lying within the KS bounds have been investigated. Because we recognize that a given individual's criterion for a distribution to be plausible may include conditions beyond a KS test, in this section we examine the $LR$ values obtainable by distributions satisfying a variety of assumption sets. These assumption sets are displayed in the form of a lattice diagram (Grätzer, [50]) as shown in Fig. 4. In the figure, when a line segment connects two assumption statements, the assumption appearing lower in the lattice diagram is nested within (*i.e.*, more restrictive than) the assumption appearing higher in the diagram. In Fig. 5, we report interval summaries of the range of $LR$ values over the considered subset of the space of all possible models satisfying the criteria for a subset of nodes in Fig. 4.

### 2.1.8 Discussion of Results

Results in Fig. 5 clearly demonstrate that within this particular educational example, the distributional assumptions made regarding the data-generating process can have a substantial effect on the $LR$ values that would be reported. Keep in mind that we examined only a small subset of all possible $CDF$s that would be deemed by the KS confidence band to be consistent with the considered RI data. As such, the uncertainty pyramid portrayed in Fig. 5 is likely to under represent the influence of choices of $f$ and $g_0$ within this example. Once again, the point is that reporting a single $LR$ value after an examination of available forensic evidence fails to correctly communicate to the DM the information actually contained in the data. Personal choices strongly permeate every model. If expert testimony is to include the computation of an $LR$, we feel an assumptions lattice and corresponding $LR$ uncertainty pyramid provide a more accurate assessment of the information in the evidence itself and better enable an audience to assess the fitness-for-purpose of the evaluation. The proposal to present an uncertainty pyramid is neither intended to replace, nor intended to
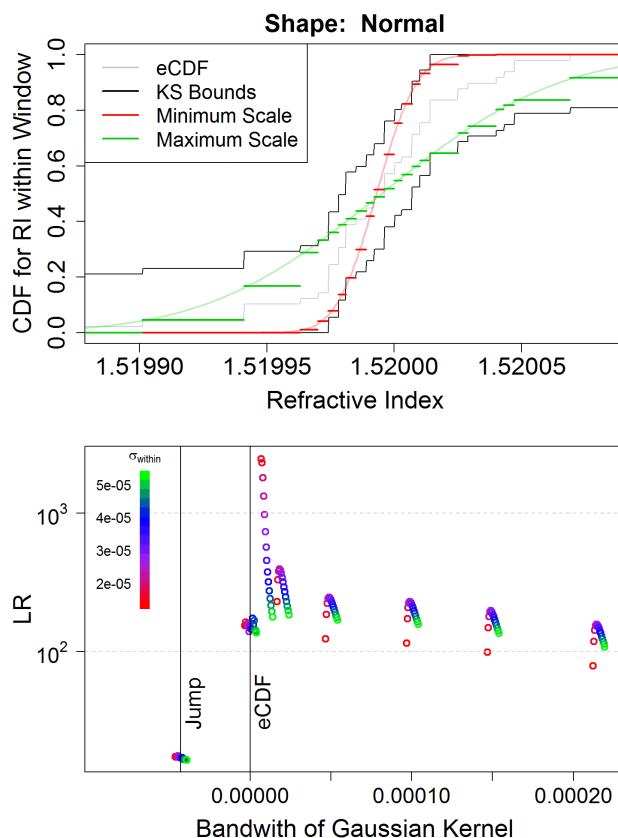
**Fig. 3.** Top: 95% Kolmogorov-Smirnov confidence band for the *CDF* of refractive indices from 49 fragments from a single window (Bennett Data). The empirical *CDF* is shown in gray. The faded red and green smooth curves respectively correspond to normal distributions with the smallest and largest scale (standard deviation) parameters such that the discrete distributions obtained, to account for interval-censoring in the reported data (shown using solid red and solid green line segments, respectively), are entirely contained within the confidence band.

Bottom: *LR* values corresponding to various choices of *F*, reflected by position along the *x*-axis, and the scale factor for the shape corresponding to a normal distribution. The left-most results correspond to the estimate of *F* labeled as Jump, which is displayed in Fig. 2. The remaining positions reflect the bandwidth of the Gaussian kernel leading to the estimate of *F* used in computing the *LR*. Within each choice of *F*, the *LR* values are staggered in order of the scale parameter used to define $g_0$ to emphasize the potential non-monotonic relationship between *LR* and scale parameter. The points are color coded to indicate the associated scale parameter values.

lessen, the importance of providing objective descriptions of empirical results from analysis and investigation.

## 2.2 Illustration 2: Score-based Likelihood Ratio based on Simulated Fingerprints

For this illustration we used a collection of simulated fingerprints to avoid confidentiality issues associated with using real finger marks from actual casework. To be clear, this example does not reflect or assess the behavior or performance of trained latent print examiners. Rather, it is intended to examine the influence of assumptions when forming a score-based likelihood ratio (*SLR*). The ideas expressed in this
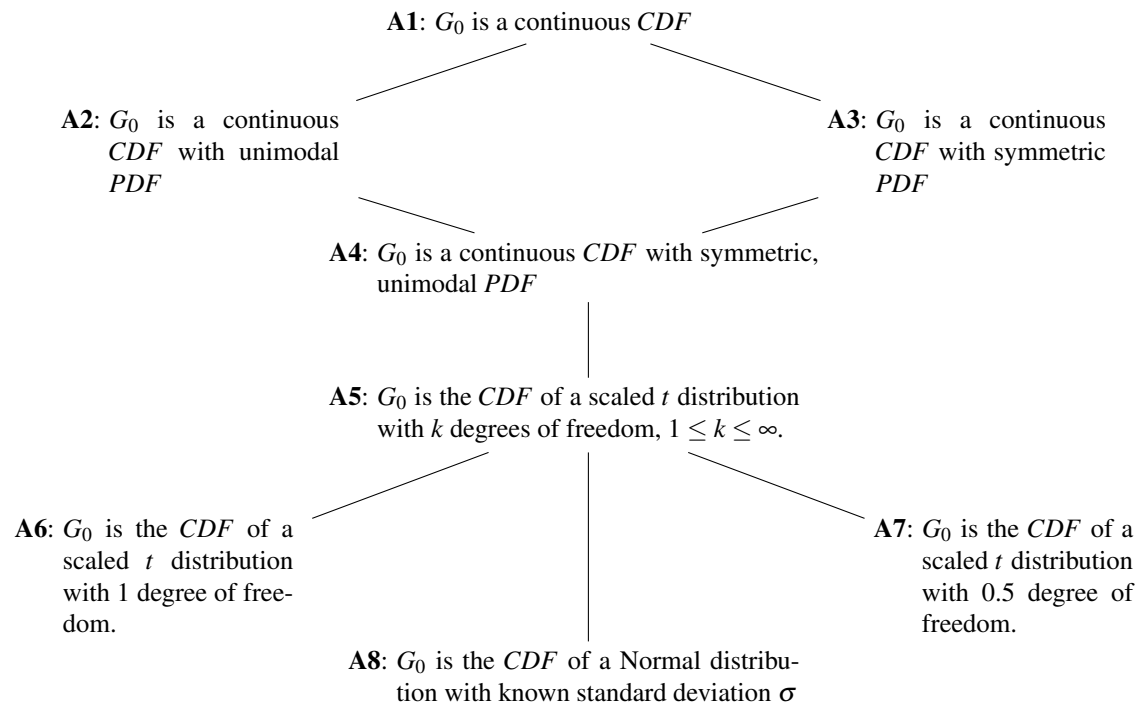
Journal of Research of National Institute of Standards and Technology

**A1**: $G_0$ is a continuous *CDF*

**A2**: $G_0$ is a continuous
*CDF* with unimodal
*PDF*

**A3**: $G_0$ is a continuous
*CDF* with symmetric
*PDF*

**A4**: $G_0$ is a continuous *CDF* with symmetric,
unimodal *PDF*

**A5**: $G_0$ is the *CDF* of a scaled $t$ distribution
with $k$ degrees of freedom, $1 \leq k \leq \infty$.

**A6**: $G_0$ is the *CDF* of a
scaled $t$ distribution
with 1 degree of free-
dom.

**A7**: $G_0$ is the *CDF* of a
scaled $t$ distribution
with 0.5 degree of
freedom.

**A8**: $G_0$ is the *CDF* of a Normal distribu-
tion with known standard deviation $\sigma$

**Fig. 4.** Assumptions lattice for the glass example.
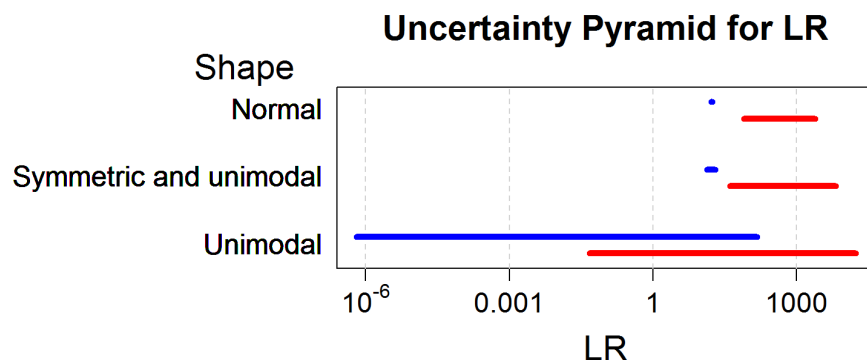


**Fig. 5.** Ranges of *LR* values corresponding to a subset of choices from the assumptions lattice for $G_0$ combined
with either the Jump distribution (blue) or Gaussian kernel density estimates (red) for $F$.

example are not limited to applications of *SLR*s for friction ridge evaluations; they apply to *SLR*
formulations for any comparison discipline.

The software system **Anguli** (see http://dsl.cds.iisc.ac.in/projects/Anguli/index.html; Jadhav
[51]) was used to generate a pair of exemplar-like impressions for 10,000 simulated fingers. One impression
from each pair was blurred, occluded, distorted, and overlaid on a background image to represent a
questioned impression. Minutia were automatically marked in each image using the automatic minutia
detecting program MINDTCT (NBIS [52]) from the National Institute of Standards and Technology (NIST).

Figure 6 displays two pairs of simulated images along with the minutia identified by MINDTCT. We retained all detected minutia with a quality score of at least 20. As seen in Fig. 6, this threshold allows erroneous minutia detections; stricter thresholds, however, were found to remove true minutia detections. As the focus here is on the uncertainty in interpreting a given set of scores and not on obtaining the best scores, no formal optimization was performed to select a minutia quality threshold. The BOZORTH3 algorithm (NBIS [52]) was used to automatically assign a similarity score between two lists of marked minutia.
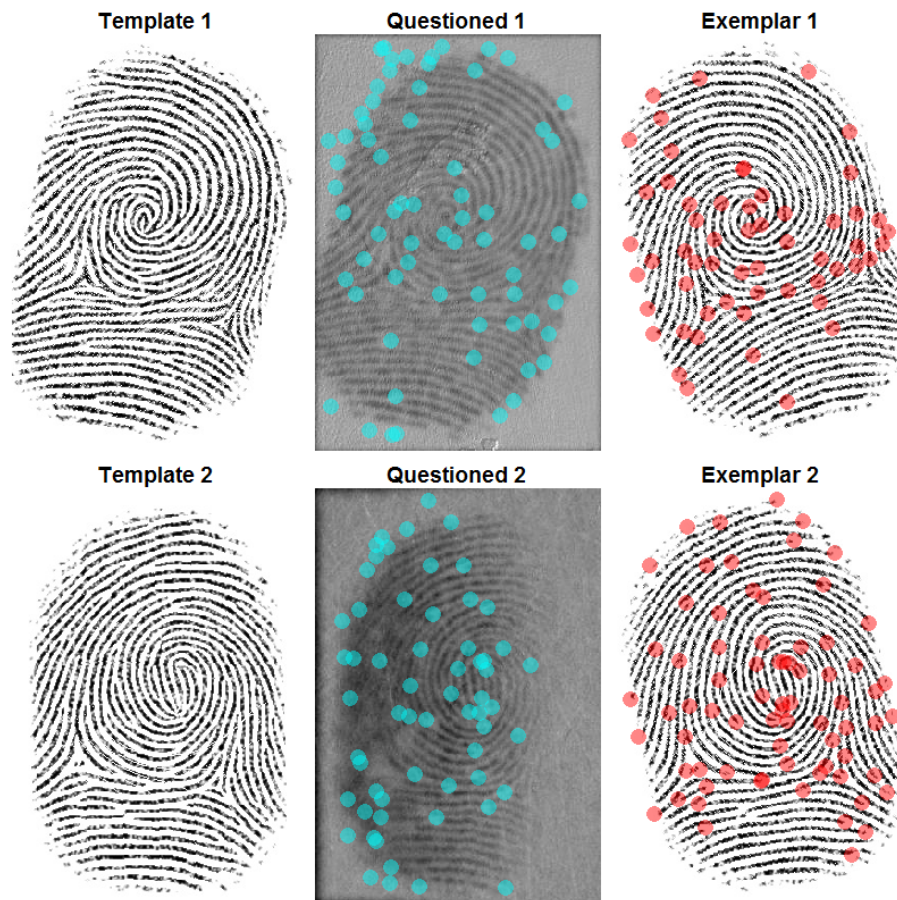


**Fig. 6.** Left: Two simulated exemplars used as templates to construct questioned impressions.
Center: Simulated questioned impression. Cyan dots indicate minutia detected by MINDTCT.
Right: Simulated exemplar used for comparison with questioned impressions. Red dots indicate minutia detected by MINDTCT.

Suppose a questioned impression ($Q$) from an unknown source is compared to a test impression ($T_i$) from source $i$ using algorithm $C(\ ,\ )$, resulting in score $s = C(Q, T_i)$. Let $F(s)$ denote the probability of observing score $s$ when comparing two images from a common source, and let $G(s)$ denote the probability of observing score $s$ when comparing two images from two different sources. Interest lies in the ratio

$$SLR(s) = \frac{F(s)}{G(s)}.$$

To inform possible choices of $F$ and $G$, we consider a collection of scores obtained from comparisons for which we know whether or not the compared images originated from a common finger. We refer to

comparisons between images generated from the same simulated finger (*e.g.*, comparing Questioned 1 to Exemplar 1 from Fig. 6) as "mated." Comparisons between images originating from different simulated fingers (*e.g.*, comparing Questioned 1 to Exemplar 2 from Fig. 6) are referred to as "nonmated." Comparing the questioned and exemplar images within each simulated finger produced a collection of $10^4$ mated scores, $S_{M,i}$ ($i = 1, \ldots, 10^4$). We compared questioned and exemplar images independently sampled from their respective collections of $10^4$ images, subject to the constraint that the selected images did not originate from the same simulated finger, to produce a collection of $10^5$ nonmated scores $S_{NM,i}$ ($i = 1, \ldots, 10^5$). The similarity scores output by BOZORTH3 are always nonnegative integers. In our simulation, mated scores ranged from 0 to 227, and nonmated scores ranged from 0 to 36. Scores of 1 and 2 did not occur among any of the mated or nonmated evaluations. The number of occurrences of each integer from 0 to 250 was tabulated for mated and nonmated scores, respectively, and is portrayed in Fig. 7. Let $M = [M_0, \ldots, M_{250}]$ and $NM = [NM_0, \ldots, NM_{250}]$ denote the corresponding vectors of occurrences, where $M_j = \sum_{i=1}^{10^4} I_{[S_{M,i}=j]}$ and $NM_j = \sum_{i=1}^{10^5} I_{[S_{NM,i}=j]}$. Here, the term $I_{[S=j]}$ is equal to 1 when $S = j$, and it is 0 otherwise.
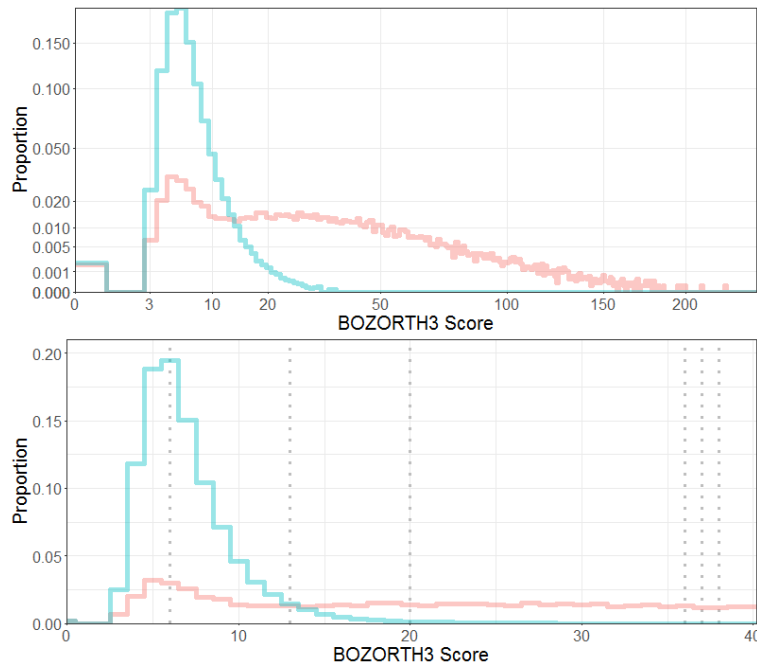


**Fig. 7.** Top: Histogram reflecting proportion of simulations resulting in each score for the mated (red) and nonmated (blue) pairs. Note that *x* and *y* axes are provided on a square root scale.
Bottom: A zoomed-in view, using linear scales. The dotted lines indicate the scores for which the *SLR* range was examined.

The choice of reference comparisons that are suitable for informing the score distributions *F* and *G* for a particular case is subjective and influential. In this illustration, we ignore this choice as a potential source of uncertainty and operate as though all DMs have agreed on the simulated collection of scores as being exclusively appropriate for informing their beliefs. That is, we suppose that all relevant DMs consider the mated scores to be iid from *F* (*i.e., $S_{M,i} \sim F$*) and the nonmated scores to be iid from *G* (*i.e., $S_{NM,i} \sim G$*).

The *SLR* value corresponding to the score observed for a particular comparison varies as one considers various plausible sets of assumptions used to evaluate *F* and *G*. In this exercise, we examine *SLR* ranges for scores of 6, 13, 20, 36, 37, and 38. The scores 6, 13, and 20 were chosen because the corresponding ratios of relative frequencies were near 0.1, 1, and 10, respectively. The scores 36, 37, and 38 were chosen to

examine the robustness of the *SLR* at and just beyond the most extreme observed nonmated score (36 in our illustration).

We consider a different plausibility criterion here than was used in the glass example. Suppose the proposed mated probability mass function (*PMF*) is given by $\boldsymbol{F}' = [F'_0, \ldots, F'_{250}]$, where $F'_j = Pr(S_M = j)$. Similarly, let the proposed nonmated *PMF* be given by $\boldsymbol{G}' = [G'_0, \ldots, G'_{250}]$, where $G'_j = Pr(S_{NM} = j)$. We consider the test statistic

$$Z_{\boldsymbol{F}',\boldsymbol{G}'} = \sum_{i=0}^{250} \left[ \left( \frac{(E_{M,i} - M_i)^2}{E_{M,i}} + \frac{(E_{NM,i} - NM_i)^2}{E_{NM,i}} \right], \right. \tag{8}$$

where $E_{M,i} = F'_i \times 10^4$ and $E_{NM,i} = G'_i \times 10^5$ are the expected counts associated with a score of $i$ under $\boldsymbol{F}'$ and $\boldsymbol{G}'$, respectively. The tables of observed counts include many cells with small values, so we estimate the sampling distribution of this test statistic under proposed distributions $\boldsymbol{F}'$ and $\boldsymbol{G}'$ using simulation rather than relying on an asymptotic chi-squared approximation. That is, in each of many iterations, we draw $\boldsymbol{M}^* \sim \text{multinomial}(10^4, \boldsymbol{F}')$ and $\boldsymbol{NM}^* \sim \text{multinomial}(10^5, \boldsymbol{G}')$ and use the simulated values to obtain $Z^*_{\boldsymbol{F}',\boldsymbol{G}'}$, computed from Eq. (8) with $\boldsymbol{M}^*$ and $\boldsymbol{NM}^*$ in place of $\boldsymbol{M}$ and $\boldsymbol{NM}$, respectively. The collection of $Z^*_{\boldsymbol{F}',\boldsymbol{G}'}$ values is used to asses whether $Z_{\boldsymbol{F}',\boldsymbol{G}'}$ is lower than the 95th percentile of the test statistic in Eq. (8) under the null distribution where $\boldsymbol{F}'$ and $\boldsymbol{G}'$ are exactly correct. If so, then $\boldsymbol{F}'$ and $\boldsymbol{G}'$ are considered plausible.
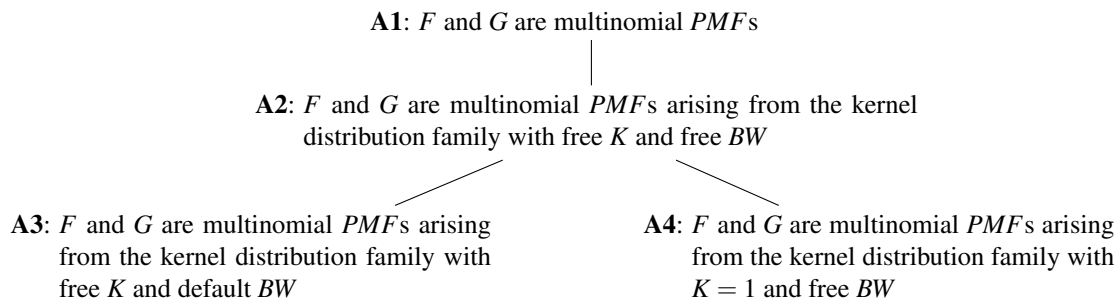
We evaluate the range of *SLR* values attainable from distributions meeting the criteria described above, first while considering any $\boldsymbol{F}'$ and $\boldsymbol{G}'$ as candidates and then considering only those belonging to various classes of Gaussian kernel distribution estimates applied to power transformations of the observed scores. More precisely, we consider kernel distribution estimates of the form

$$F'_{K_1,BW_1}(s) = 10^{-4} \sum_{i=3}^{227} M_i \times \frac{\Phi\left( \frac{(s+0.5)^{K_1} - i^{K_1}}{BW_1} \right) - \Phi\left( \frac{(s-0.5)^{K_1} - i^{K_1}}{BW_1} \right)}{1 - \Phi\left( \frac{2.5^{K_1} - i^{K_1}}{BW_1} \right)} \left( \right.$$

and

$$G'_{K_2,BW_2}(s) = 10^{-5} \sum_{i=3}^{36} NM_i \times \frac{\Phi\left( \frac{(s+0.5)^{K_2} - i^{K_2}}{BW_1} \right) - \Phi\left( \frac{(s-0.5)^{K_2} - i^{K_2}}{BW_2} \right)}{1 - \Phi\left( \frac{2.5^{K_2} - i^{K_2}}{BW_2} \right)} \left( \right.$$

for $s \geq 3$, where $0 < K_1, K_2 \leq 1$; $BW_1, BW_2 \geq 0$; and $\Phi(\cdot)$ denotes the *CDF* of a standard normal distribution. For completeness, define $F'_{K_1,BW_1}(0) = M_0 \times 10^{-4}$, $F'_{K_1,BW_1}(1) = 0$, and $F'_{K_1,BW_1}(2) = 0$. Similarly, define $G'_{K_2,BW_2}(0) = NM_0 \times 10^{-5}$, $G'_{K_2,BW_2}(1) = 0$, and $G'_{K_2,BW_2}(2) = 0$. Note $BW_1 = 0$ corresponds to $\boldsymbol{F}'$ being the empirical *PMF* for the mated scores, and $BW_2 = 0$ corresponds to $\boldsymbol{G}'$ being the empirical *PMF* for the nonmated scores. We also consider the class of distributions where $K_1$ and $K_2$ are fixed at 1 (still allowing $BW_1, BW_2 \geq 0$), and the class of distributions where $BW_1$ and $BW_2$ are the bandwidth selections produced by applying the R function `density` [49] with default settings to the sets $\left\{ M_i^{K_1} \mid 1 \leq i \leq 10^4 \right\}$ and $\left\{ NM_i^{K_2} \mid 1 \leq i \leq 10^5 \right\}$, respectively (allowing $0 < K_1, K_2 \leq 1$). The distributions produced using $K_1 = K_2 = 1$ and default bandwidths did not pass the plausibility criterion as the corresponding value of $Z_{\boldsymbol{F}',\boldsymbol{G}'}$ was near the 99th percentile of the null distribution. The assumptions lattice for the considered classes of distributions is shown in Fig. 8. Corresponding *SLR* ranges are presented as uncertainty pyramids in Fig. 9.

**A1**: *F* and *G* are multinomial *PMF*s

**A2**: *F* and *G* are multinomial *PMF*s arising from the kernel distribution family with free *K* and free *BW*

**A3**: *F* and *G* are multinomial *PMF*s arising from the kernel distribution family with free *K* and default *BW*

**A4**: *F* and *G* are multinomial *PMF*s arising from the kernel distribution family with $K = 1$ and free *BW*

**Fig. 8.** Assumptions lattice for the fingerprint example.

## 3.  Discussion

The viewpoints expressed in this paper are largely motivated by considerations of standard practices in measurement science, a discipline for which a fundamental purpose is to facilitate meaningful communication regarding properties of an object or system among interested parties. From the perspective of metrology, the hybrid *LR* framework asks a forensic expert to measure the weight of evidence on behalf of the DM and report its value for subsequent use in Bayes' formula. As a measurement, any provided *LR* value would require an accompanying uncertainty statement (JCGM [53]; Possolo [54]) characterizing the analyst's belief regarding its deviation from the "true value," which the Bayesian paradigm defines as the *LR* value a given DM would arrive at following careful review of the complete body of evidence considered by the expert. Overlooking or dismissing the relevant uncertainty would treat the value obtained by an expert as though it is a perfect measurement of weight of evidence, universally and exactly accurate. This directly contradicts the Bayesian paradigm, where no such value can be assumed to exist, as the *LR* is a personal and subjective entity.

Although our discussion of the *LR* has centered around the perspective of Bayesian decision theory, our concerns apply to any framework motivating the use of an *LR* as a means for experts to communicate their findings. Whether a probability is intended to be personal or communal, it is not empirical in the sense that it is not directly observable. A model is required in order to translate data into a probability, and the question of how robust the translation is among reasonable model choices remains central.

We do not make a recommendation regarding when an uncertainty characterization yields a particular *LR* result that may be considered to be fit for the intended purpose. Our hope is that policy makers will assess the adequacy of relying on *LR* characterizations in the context of the framework presented here, mindful of the range of alternative results that might be reasonably obtained and of the criteria used to make that assessment. One might expect to find the least degree of uncertainty in applications of probabilistic evaluation of high-template, low-contributor DNA samples, and we recognize that the community may be well founded in its use of probability to facilitate knowledge transfer in such cases. We do not view this as an exception to the framework we present, but rather as a scenario in which extensive uncertainty evaluations would likely yield a degree of consensus leading most people to conclude an offered *LR* value is fit for the intended purpose. Forming a lattice of assumptions and uncertainty pyramid, including explicitly identifying what data will be considered, for applications in the field of high-template, low-contributor DNA evaluations could help to provide clarity to other forensic disciplines seeking to demonstrate or develop a basis for using a similar *LR* framework. In absence of a suitable uncertainty characterization, or when the uncertainty is deemed too large, *LR* values may require less literal interpretations.

When an *LR* value is the output of a computer algorithm, one may reasonably assume that, given the inputs, it is highly reproducible. In this sense, an *LR* value may be transferable as a discriminant score rather than the ratio of two probabilities. In this context, a discriminant score attempts to produce an optimal
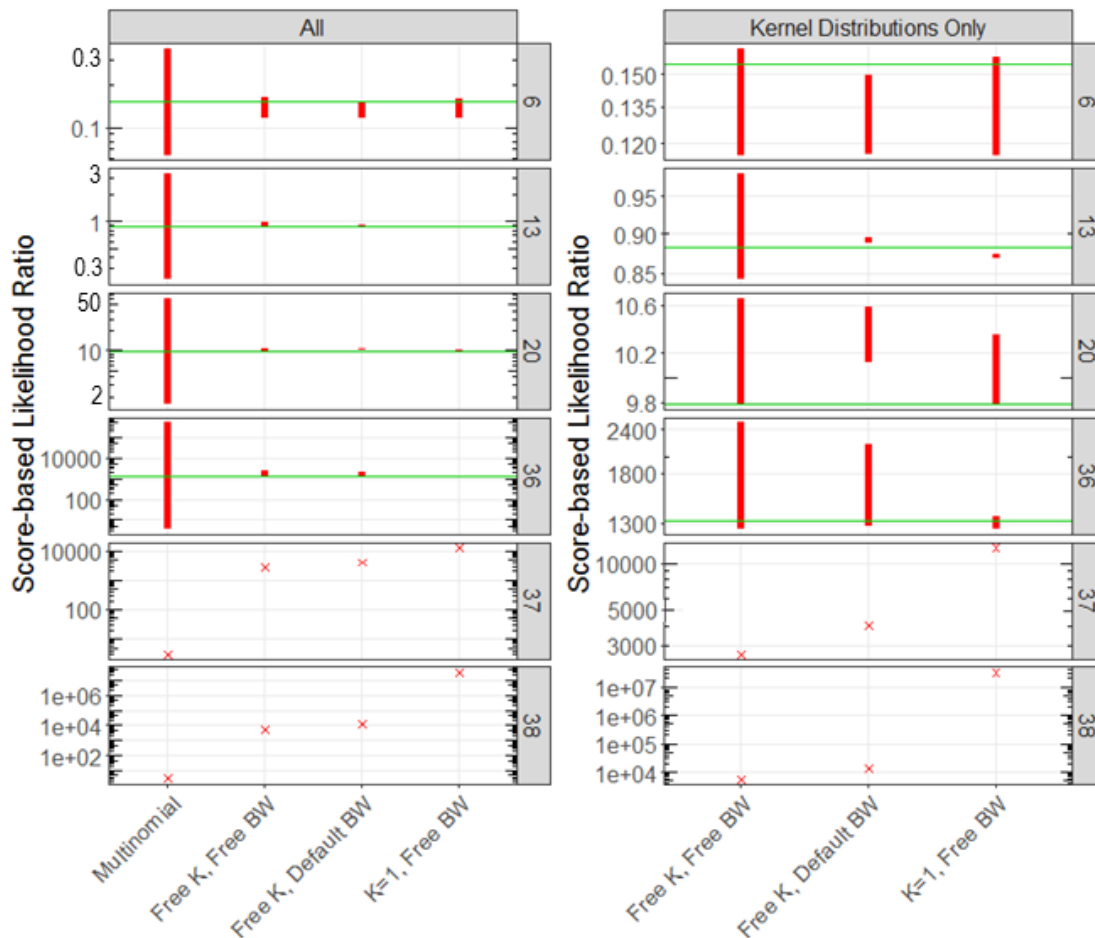
**Fig. 9.** *SLR* uncertainty pyramids for various scores. Panels are vertically arranged according to the score for which the *SLR* is computed. The right panel excludes results from the general multinomial class in order to better depict the results from the classes of kernel distribution estimates. Green horizontal lines depict the ratio of relative frequencies, $10 \times \dfrac{M_s}{NM_s}$. Vertical line segments depict the range of *SLR* attainable by distributions satisfying the selected plausibility criterion and belonging to the class indicated along the *x*-axis. Points shown in the bottom two panels corresponding to scores of 37 and 38 indicate the lower bound of the *SLR* range. The corresponding upper limits and ratio of relative frequencies are all positive infinity.

ordering among a collection of independent scenarios that may originate from either $H_p$ or $H_d$. For a given ordering, a decision rule is indicated by a threshold, with all scenarios having a score to one side of the threshold being ascribed to $H_d$ and scenarios with scores on the other side of the threshold ascribed to $H_p$. The ordering is optimal when any chosen threshold corresponds to the best attainable error rates given the total number of scenarios that will be ascribed to $H_p$ and $H_d$. In a theoretical scenario where the true *LR* is known for each scenario, the *LR* is the optimum discriminant score. When viewed as a discriminant score, an *LR* value would not have direct, probabilistic interpretation, because its meaning only becomes apparent from its positioning relative to *LR* determinations for other scenarios, evaluated by the same process, including suitable, controlled reference applications. The effectiveness of a given scoring method can be empirically assessed using Receiver operating characteristic (ROC) plots (Peterson and Birdsall [55]; Green and Swets [56–57]).

Relying on a given scoring method, an expert could provide demonstrations or scientifically sound descriptions to answer many helpful questions. For instance, in a source-level evaluation, an expert might address:

- How were the scores produced and why? What collection of reference scenarios were used to evaluate the performance of the considered scoring methods? How were these chosen in light of the considered case?

- What score was obtained corresponding to the source of interest?

- What alternative sources were considered, and what were the corresponding scores?

- How do the scores from this particular case compare to the scores obtained among the reference collection used to evaluate method performance?

More broadly, objective descriptions of procedures followed and outcomes obtained throughout investigation of the case and broader experience may present a promising path to ensuring transferability of information from a forensic expert to DMs.

## 4.  Summary

The *LR* framework has been portrayed by some as having an exclusive, normative role in forensic expert communication on the basis of arguments centered around mathematical definitions of rationality and coherence (*e.g.*, Biedermann *et* al. [58]). These arguments are aimed at ensuring a form of self-consistency of a single, autonomous decision maker.[8] Decision theory, however, does not consider the transfer of information among multiple parties, such as that occurring throughout the judicial process when one or more DMs rely on forensic experts to help inform their decisions. Thus, while decision theory may have a normative role in how a DM processes information presented during a case or trial in accordance with his or her own personal beliefs and preferences, it does not dictate that a forensic expert should communicate information to be considered in the form of an *LR*.

Some may argue that because any given DM is likely unfamiliar with formal decision theory, a trained expert should act on their behalf to form an *LR*. As expounded throughout this paper, the interpretation of evidence in the form of an *LR* is personal and subjective. We have not encountered any basis for the presumption that the surrogate *LR* of an expert will reflect a truer implementation of decision theory than will the unquantified perception of the DM following effective presentation of the information upon which the expert's *LR* is based.

Bayesian decision theory neither mandates nor proves appropriate the acceptance of a subjective interpretation of another,[9] regardless of training, expertise, or common practice. It does not recognize one person's subjective inputs as superior to those of another, and therefore it does not support any one particular *LR* value. Validation efforts can demonstrate that the interpretation corresponding to a particular model is reasonable, but this should not be misunderstood to mean the model is accurate or authoritatively appropriate.[10]

Validation efforts can also inspire an explicit plausibility criterion. By conducting multiple analyses attempting to span the space of assumptions meeting a specified plausibility criterion, an analyst can purposefully explore the robustness of an interpretation. Presenting an uncertainty pyramid, along with an explanation of the corresponding plausibility criterion and a description of the data, may provide the audience the opportunity for greater understanding of the interactions among data, assumptions, and

---

[8]More specifically, the Dutch book arguments (Hájek [59])

[9]"I emphasize that the answers you give to the questions I ask you about your uncertainty are yours alone, and need not be the same as what someone else would say, even someone with the same information as you have, and facing the same decisions." – Kadane [23]

[10]As a result of this common misunderstanding, we prefer phrases that use "plausible" in place of "validated."

interpretation. The audience may then, more reasonably, assess whether any particular result is fit for the intended purpose.

If such uncertainty characterizations are considered untenable for a given application, one may be forced to conclude that the *hybrid plan* [see Eq. (2)], though appealing, is impractical to implement. It does not mean that, just because one is unable to calculate the required value, one should accept the value that can be calculated.

We hope this paper will encourage the forensic science community to be mindful of the many subjective components involved in any interpretation. Correspondingly, we hope best-practice guidance will address how to avoid overstating the authority or rigor underlying any particular interpretation of evidence and require a presentation of uncertainty. Additionally, we hope the forensic science community comes to view the *LR* as one *possible*, not normative or necessarily optimum, tool for communicating to DMs. We hope such viewpoints will increase the priority given to developing tools for descriptive presentations that meet the strict standards of scientific validity by focusing on empirical and reproducible results, assisting the DMs in directly establishing their own respective interpretations of the weight of evidence.

## 5.    Appendix A: Likelihood Ratio Introduction

The concept of likelihood ratio (*LR*) arises naturally when one is faced with the problem of deciding whether an observation $x$ came from one of two populations. Consider a simple situation involving two urns, urn 1 and urn 2. Urn 1 has 99 red balls and one green ball, and urn 2 has 99 green balls and one red ball. One of the urns is chosen (we do not know which one or the process used to make the choice), and, after thoroughly mixing the balls in it, one ball is selected, and its color is noted. Suppose the ball is red. We would like to know whether the ball is from urn 1 or urn 2.

One may proceed as follows. Let us assume that every ball from the chosen urn had an equal chance of being chosen. Then, if urn 1 was chosen, the probability of drawing a red ball is 99 %. If urn 2 was chosen then the probability of drawing a red ball is 1 %. Thus, a red ball is 99 times more likely to be drawn if urn 1 was chosen than if urn 2 was chosen. That is, the ratio

$$\frac{\text{Probability of drawing a red ball given urn 1 was chosen}}{\text{Probability of drawing a red ball given urn 2 was chosen}} = 99. \qquad (A.1)$$

Whatever the initial belief might have been of an individual regarding whether urn 1 or 2 was selected, the effect of observing a red ball is likely to encourage the individual to update their beliefs by increasing the probability they initially assigned to the scenario that urn 1 was selected.

The above example provides the beginnings of the concept of weight of evidence. It also suggests that the ratio of probabilities of an observed occurrence under each of the two considered scenarios must play a role in adjusting one's prior beliefs regarding which scenario is true. The ratio in Eq. (A.1) is called the likelihood ratio for urn 1 corresponding to the observation of a red ball. More generally, if $x$ denotes data observed from one of two distributions, $f_1$ or $f_2$, then the ratio

$$\frac{\text{Probability of observing } x \text{ when sampling from } f_1}{\text{Probability of observing } x \text{ when sampling from } f_2}$$

is called the likelihood ratio for $f_1$ corresponding to the observation $x$. This simple example might help the reader understand why *LR* is a quantity of importance when one faces the problem of discriminating between two populations.

More formal mathematical justifications are available for the use of *LR* for assessing the added value provided by new information $x$ when faced with discriminating between two situations. These justifications are based on ideal applications where the needed probabilities are exactly known. We give a brief outline of two theoretical justifications often given in the literature.

## 5.1 Discriminating between Two Simple Hypotheses

Neyman and Pearson are perhaps the most recognized as the first to give a formal explanation for the role of the likelihood ratio in discriminating between two hypotheses, populations, or propositions. Suppose, in each of many repeated trials $T_i$, resulting in observations $x_i$ ($i = 1, \ldots, n$), one is tasked with deciding from which of two known distributions ($f_1$ or $f_2$) the observation $x_i$ is drawn. That is, in each trial one must decide between the hypotheses

$$H_{1i} : x_i \text{ came from } f_1,$$
$$\text{or,} \quad H_{2i} : x_i \text{ came from } f_2.$$

The Neyman-Pearson fundamental lemma [60] essentially states that these outcomes are optimally ordered according to the ratio

$$LR_i = \frac{f_1(x_i)}{f_2(x_i)},$$

in the sense that $x_i$ should be considered as more strongly favoring $H_{1i}$ than $x_{i'}$ favors $H_{1i'}$ if and only if $LR_i > LR_{i'}$. Given any rule $\mathcal{R}$ for discriminating between $H_{1i}$ and $H_{2i}$ that is based on an observation $x_i$ (i.e., conclude $H_{1i}$ if $x_i$ satisfies some given condition and conclude $H_{2i}$ otherwise), one can always find an $LR$ rule $\mathcal{R}_{LR}$ (i.e., for a given $\tau \geq 0$, conclude $H_{1i}$ if $LR_i \geq \tau$ and conclude $H_{2i}$ if $LR_i < \tau$) that will, in the long run, correctly decide $H_{1i}$ to be true, when it is in fact true, for at least as many trials as $\mathcal{R}$ will, and will wrongly decide $H_{1i}$ to be true, when it is in fact false, in no more trials than $\mathcal{R}$ will.

Note that we have assumed $f_1$ and $f_2$ to be completely known. That is, no modeling was necessary and no distribution was fit to empirical data. The Neyman-Pearson fundamental lemma is applicable primarily in such *ideal* situations. Real situations are more complex, and optimality of any particular *LR*-based rule cannot be universally guaranteed.

## 5.2 *LR* in a Bayesian framework.

The Bayesian framework is based on the philosophical viewpoint that all probabilities are personal and quantify one's state of uncertainty regarding the truth of propositions. Given the problem of discriminating between $H_1$ and $H_2$ as above, one first quantifies one's uncertainties associated with the truth of $H_1$ and of $H_2$ by (prior) probabilities $\pi_1$ and $\pi_2 = 1 - \pi_1$. These describe the levels of uncertainty experienced by an individual prior to seeing the data $x$.

After seeing $x$, one's uncertainties regarding the truth of $H_1$ and of $H_2$ may change. Uncertainty experienced after seeing the data $x$ is called a posterior probability. Posterior probabilities are written as $P(H|x)$, which is read as "the probability of $H$ given $x$." Symbols appearing to the right of the vertical line represent quantities known to the individual and used in evaluating the probability. In the considered scenario, interest lies in the posterior probabilities $P(H_1|x)$ and $P(H_2|x)$ (note that $P(H_1|x) + P(H_2|x) = 1$) or, equivalently, the posterior odds

$$\text{Posterior Odds} = \frac{P(H_1|x)}{P(H_2|x)} = \frac{P(H_1|x)}{1 - P(H_1|x)}.$$

An application of Bayes' rule for updating one's prior personal probabilities after having observed new information leads to the equation

$$\text{Posterior Odds for } H_1 = \frac{P(H_1|x)}{P(H_2|x)} = \frac{P(x|H_1)}{P(x|H_2)} \times \frac{P(H_1)}{P(H_2)} = \frac{f_1(x)}{f_2(x)} \times \text{Prior Odds for } H_1.$$

If we define *weight of evidence* associated with $x$ for a particular individual to be the ratio of posterior odds (given $x$) of that individual to his or her prior odds (before observing $x$), then the above equation implies that

$LR = \dfrac{f_1(x)}{f_2(x)}$ is to be viewed as the weight of the evidence provided by $x$ for $H_1$ for the individual making the probability assessments.

### 5.3  Surrogate *LR*s as Discriminant Scores

The Neyman-Pearson fundamental lemma tells us that the theoretical *LR* is the best summary of the information in $x$ for discriminating between $H_1$ and $H_2$. In this sense, we can say that, when $f_1$ and $f_2$ are known, *LR* is the best *discriminant score*. When $f_1$ and $f_2$ are not known, it is customary to use empirical information to find surrogates for $f_1$ and $f_2$ (*i.e.*, models) and use these to construct a surrogate *LR* corresponding to an observed value $x$. Different models based on different sets of assumptions will lead to different *LR*s. These can all be helpful, some more than others, in discriminating between $H_1$ and $H_2$. We continue to refer to these surrogate *LR* values as discriminant scores. The performance characteristics of competing discriminant scores may be evaluated empirically using suitable, ground-truth known data through the use of receiver operating characteristic (ROC) plots. For a detailed discussion of ROC plots, the reader is referred to Peterson and Birdsall [55] and Green and Swets [56-57].

### 5.4  Summary

The study of *LR* in theoretical settings provides useful guidance when dealing with problems of discriminating between two or more populations in real-life applications. However, since we never really know $f_1$ or $f_2$, we have to rely on available data and statistical models to develop surrogates for the theoretical *LR*s, and no theoretical optimality properties may be claimed in the Neyman-Pearson setting. Even under the Bayesian framework, there is no unique *LR*. A main thrust of this paper is to bring to the attention of the community that these surrogate *LR*s can have substantial disagreements with one another, and no unique authoritative model from which to derive an *LR* for public consumption exists. The usefulness of any particular surrogate *LR* as a discriminant score (sometimes referred to as an *LR* system; see Leegwater *et al.* [61]) has to be demonstrated empirically using tools such as ROC plots.

## 6.  Appendix B: Additional Results from the Glass Example

In this section of the appendix, we display results for additional choices of $F$ and $G_0$. Choices considered here for $G_0$ are $\chi^2$ distribution with 3 degrees of freedom (Fig. 10); an example symmetric, unimodal distribution (Fig. 11), and an example asymmetric, unimodal distribution (Fig. 12).

The top plot in each figure shows the 95 % Kolmogorov-Smirnov confidence band for the *CDF* of RI values from 49 fragments from a single window (Bennett Data). The empirical *CDF* is shown in gray. The faded red and green smooth curves, respectively, correspond to members of the chosen scale family with the smallest and largest scaling factors such that the discrete distributions obtained by accounting for interval-censoring in the reported data (shown using solid red and solid green line segments, respectively) are entirely contained within the confidence band.

The bottom plot in each figure displays the *LR* values corresponding to various choices for $F$, reflected by position along the $x$-axis, and the scale factor used with the shape chosen for $G_0$. The left-most results correspond to the estimate of $F$ labeled as Jump, which is displayed in Fig. 2. The remaining positions reflect the bandwidth of the Gaussian kernel leading to the estimate of $F$ used in computing the *LR*. Within each choice of $F$, the *LR* values are staggered in order of the scale parameter used with $G_0$ to emphasize the potential non-monotonic relationship between *LR* and the scale parameter. The points are color-coded to indicate the associated scale parameter values in accordance with the legend titled $\sigma_{\text{within}}$.
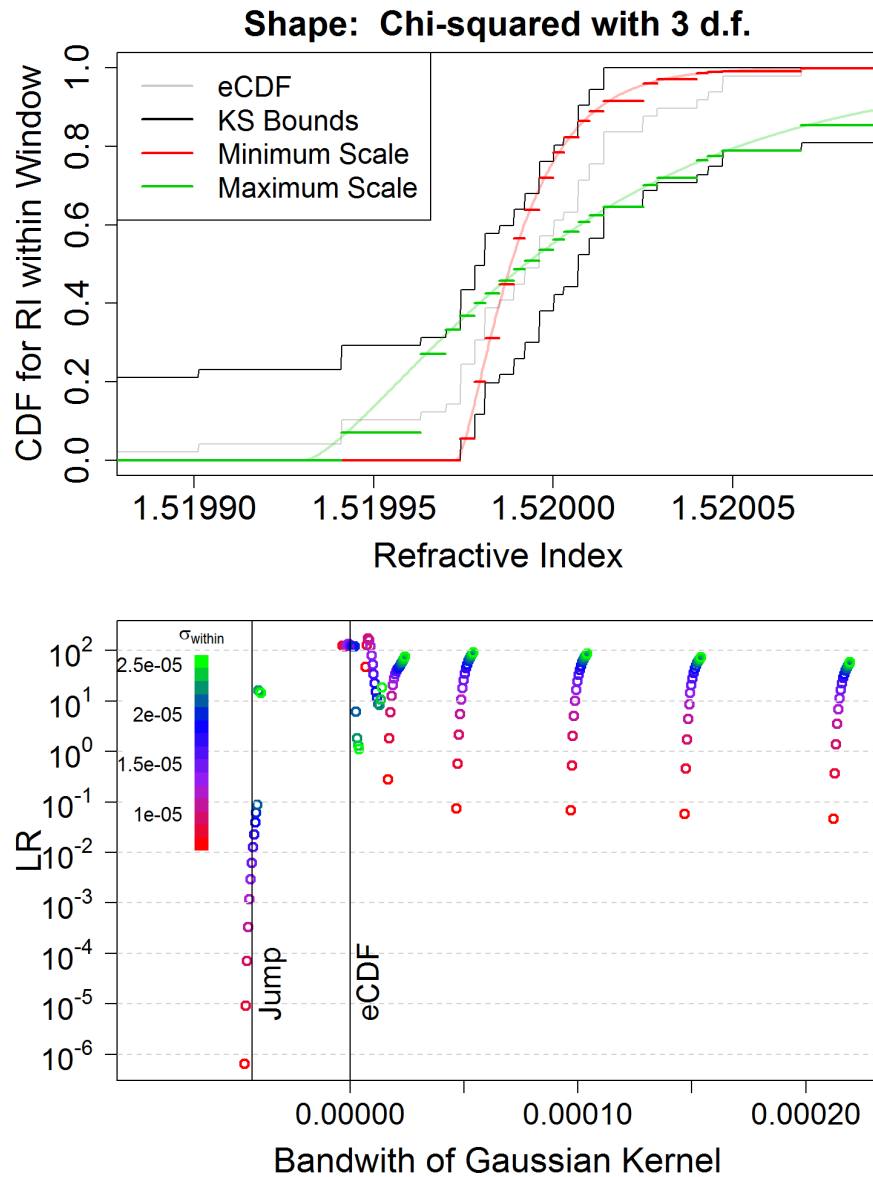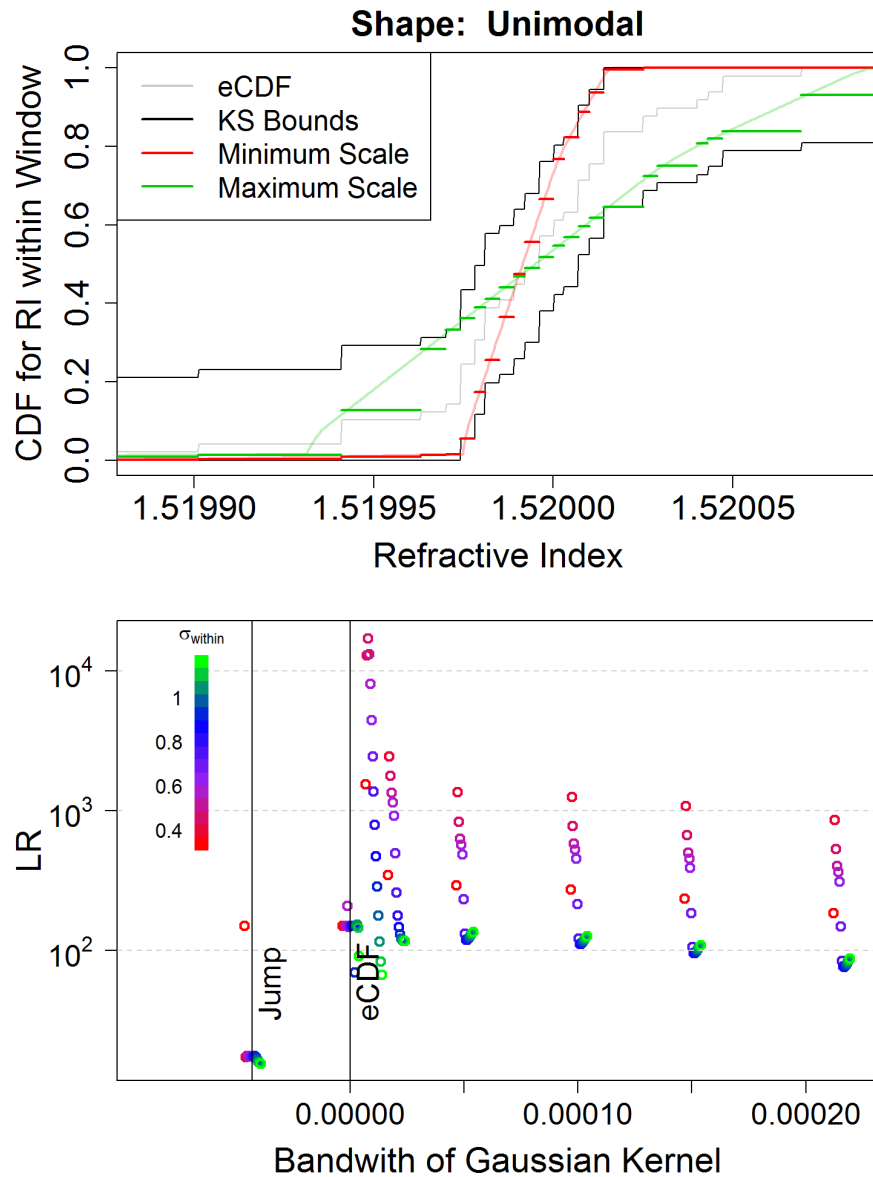
**Fig. 10.** *LR* values when $G_0$ is a $\chi^2$ distribution with 3 degrees of freedom.
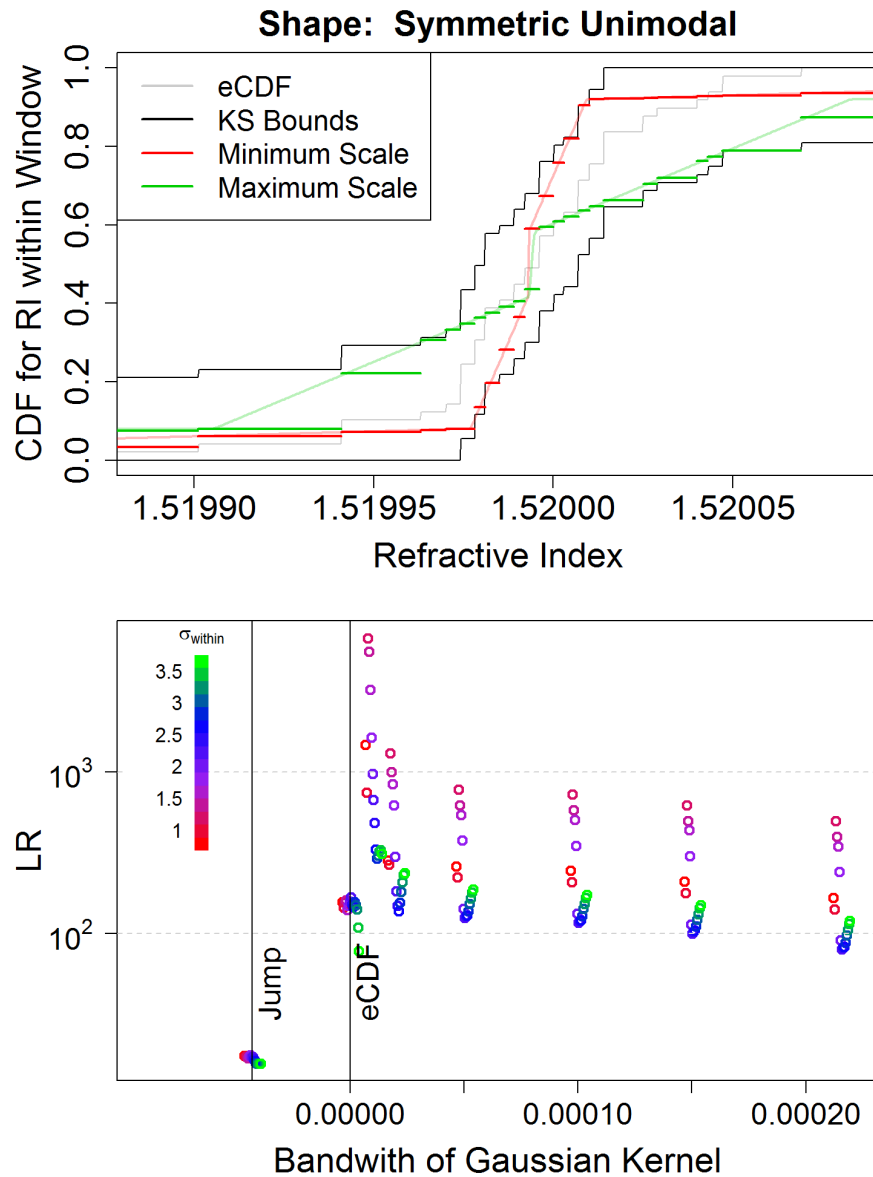
**Fig. 11.** *LR* values when $G_0$ has the shape of the symmetric unimodal distribution shown.

**Fig. 12.** *LR* values when $G_0$ has the shape of the unimodal distribution shown.

## 7.   References

[1] National Research Council (2009) Strengthening Forensic Science in the United States: A Path Forward. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, Washington, D.C., Document No. 228091, National Academy of Sciences. https://doi.org/10.17226/12589.

[2] President's Council of Advisors on Science and Technology (2016) Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. President's Council of Advisors on Science and Technology, Washington, D.C.

[3] Aitken CGG, Roberts P, Jackson G (2010) Practitioner Guide No 1: Fundamentals of Probability and Statistical Evidence in Criminal Proceedings-Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses. Communicating and Interpreting Statistical Evidence in the Administration of Criminal Justice. Prepared under the auspices of the Royal Statistical Society's Working Group on Statistics and the Law, London, United Kingdom.

[4] European Network of Forensic Science Institutes (2015) Strengthening the Evaluation of Forensic Results across Europe (STEOFRAE), ENFSI guideline for evaluative reporting in forensic science, approved version 3.0. European Network of Forensic Science Institutes, Wiesbaden, Germany.

[5] Aitken CGG, Taroni F (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists* (John Wiley and Sons, New York). https://doi.org/10.1002/0470011238.

[6] Meester R, Sjerps M (2004) Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence. *Law, Probability, and Risk* 3:61-62. https://doi.org/10.1093/lpr/3.1.51.

[7] de Keijser J, Elffers H (2012) Understanding of forensic expert reports by judges, defense lawyers and forensic professionals. *Psychology, Crime & Law* 18(2):191-207. https://doi.org/10.1080/10683161003736744.

[8] Martire KA, Kemp RI, Sayle M, Newell BR (2014) On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International* 240:61-8. https://doi.org/10.1016/j.forsciint.2014.04.005.

[9] Taroni F, Biedermann A, Bozza S, Garbolino P, Aitken CGG (2014) *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science* (Wiley, New York), 2nd ed. https://doi.org/10.1002/9781118914762.

[10] Zadora G, Martyna A, Ramos D, Aitken CGG (2014) *Statistical Analysis in Forensic Science Evidential Value of Multivariate Physicochemical Data* (Wiley, New York). https://doi.org/10.1002/9781118763155.

[11] Moreya RD, Romeijna J-W, Rouderc JN (2016) The philosophy of Bayes' factors and the quantification of statistical evidence. *Journal of Mathematical Psychology* 72:6–18. https://doi.org/10.1016/j.jmp.2015.11.001.

[12] Savage LJ (1972) *The Foundations of Statistics* (Dover, Mineola, NY) 2nd revised ed. https://doi.org/10.1002/nav.3800010316.

[13] Lindley DV (2014) *Understanding Uncertainty* (Wiley, New York), revised ed., Wiley Series in Probability and Statistics. https://doi.org/10.1002/0470055480.

[14] Gilboa I (2009) Theory of Decision under Uncertainty (Cambridge University Press, Cambridge, United Kingdom), Econometric Society Monographs 45. https://doi.org/10.1017/CBO9780511840203.

[15] Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2011) Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences of the USA* 108(19):7733–7738. https://doi.org/10.1073/pnas.1018707108.

[16] Thompson WC, Kaasa SO, Peterson T (2013) Do jurors give appropriate weight to forensic identification evidence? *Journal of Empirical Legal Studies* 10(2):359–397. https://doi.org/10.1111/jels.12013.

[17] Thompson WC, Newman EJ (2015) Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior* 39(4):332–349. https://doi.org/10.1037/lhb0000134.

[18] Lindley DV (1977) A problem in forensic science. *Biometrika* 64(2):207–213. https://doi.org/10.1093/biomet/64.2.207.

[19] Good IJ (1950) *Probability and the Weighing of Evidence* (Charles Griffin and Company Limited, London, United Kingdom). https://doi.org/10.1086/398369.

[20] Fienberg SE (1989) *The Evolving Role of Statistical Assessments as Evidence in the Courts* (Springer-Verlag, New York). https://doi.org/10.1007/978-1-4612-3604-7.

[21] Dawid AP (2002) Bayes's theorem and weighing evidence by juries. *Proceedings of the British Academy* 113:71–90. https://doi.org/10.5871/bacad/9780197263419.001.0001.

[22] Kaye DH, Freedman DA (2011) Reference guide on statistics. *Reference Manual on Scientific Evidence* (National Academy Press, Washington, D.C.), 3rd ed., pp 211–302. https://doi.org/10.17226/13163.

[23] Kadane J (2011) *Principles of Uncertainty* (Chapman and Hall/CRC Texts in Statistical Science, Boca Raton, FL). https://doi.org/10.1201/b11322.

[24] Science & Justice (2016) Special Issue on Measuring and Reporting the Precision of Forensic Likelihood Ratios, GS Morrison, Guest Editor. *Science & Justice* 56(5). https://doi.org/10.1016/j.scijus.2016.05.002.

[25] Taroni F, Aitken CGG, Garbolino P, Biedermann A (2006) *Bayesian Networks and Probabilistic Inference in Forensic Science* (John Wiley & Sons, Ltd., New York). https://doi.org/10.1002/0470091754.

[26] Neumann C, Evett IW, Skerrett J (2012) Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm. *Journal of the Royal Statistical Society A* 175(2):371–415. https://doi.org/10.1111/j.1467-985X.2011.01027.x.

[27] Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. *Statistical Science* 14(4):382–417. https://doi.org/10.7916/D84M92N7.

[28] Walley P (1991) *Statistical Reasoning with Imprecise Probabilities* (Springer-Science+Business Media, B.Y., Berlin, Germany). https://doi.org/10.1007/978-1-4899-3472-7 .

[29] Augustin T, Doria S, Marinacci M (2016) Special Issue: Ninth International Symposium on Imprecise Probability: Theory and Applications (ISIPTA'15). *International Journal of Approximate Reasoning* 83:1–280. https://doi.org/10.1016/j.ijar.2016.01.004.

[30] Lindley DV, Tversky A, Brown RV (1979) On the reconciliation of probability assessments. *Journal of the Royal Statistical Society A* 142(2):146–180. https://doi.org/10.2307/2345078.

[31] Kadane JB, Winkler RL (1988) Separating probability elicitation from utilities. *Journal of the American Statistical Association* 83(402):357–363. https://doi.org/10.2307/2288850.

[32] Taroni F, Bozza S, Biedermann A, Aitken CGG (2016) Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. Law, Probability and Risk 15:1–16. https://doi.org/10.1093/lpr/mgv008.

[33] Morgenthaler S, Tukey JW (1991) *Configural Polysampling: A Route to Practical Robustness* (Wiley-Interscience, New York).

[34] Williamson RC (1989) *Probabilistic Arithmetic*, Ph.D. dissertation. Department of Electrical Engineering, University of Queensland, Brisbane, Australia. https://doi.org/10.14264/uql.2015.241.

[35] Williamson R, Downs T (1990) Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning* 4(2):89–158. https://doi.org/10.1016/0888-613X(90)90022-T.

[36] Hoffman FO, Hammonds JS (1994) Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. Risk Analysis 14(5):707–712. https://doi.org/10.1111/j.1539-6924.1994.tb00281.x.

[37] Ferson S, Kreinovich V, Ginzburg L, Myers DS, Sentz K (2002) Constructing Probability Boxes and Dempster-Shafer Structures. Sandia National Laboratory, Albuquerque, NM, Report SAND2002-4015, printed January 2003. https://doi.org/10.2172/809606.

[38] Zhang J, Berleant D (2003) Envelopes around cumulative distribution functions from interval parameters of standard continuous distributions. *Proceedings of North American Fuzzy Information Processing Society (NAFIPS 2003)*, Chicago, IL, pp 407–412. https://doi.org/10.1109/NAFIPS.2003.1226819.

[39] Zhang J, Berleant D (2005) Arithmetic on random variables: squeezing the envelopes with new joint distribution constraints. *4th International Symposium on Imprecise Probabilities and Their Applications*, Pittsburgh, PA.

[40] Ferson S, Siegrist J (2011) Verified computation with probabilities. *Uncertainty Quantification in Scientific Computing*, 10th IFIPWG2.5 Working Conference, Boulder, CO, eds Dienstfrey AM, Boisvert RF (Springer, Berlin), pp. 95–122. https://doi.org/10.1007/978-3-642-32677-6_7.

[41] Lambert JA, Evett IW (1984) The refractive index distribution of control glass samples examined by the forensic science laboratories in the United Kingdom. *Forensic Science International*, 26:1–23. https://doi.org/10.1016/0379-0738(84)90207-X.

[42] Evett IW (1977) The interpretation of refractive index measurements. *Forensic Science* 9:209–217. https://doi.org/10.1016/0300-9432(77)90093-0.

[43] Owen AB (1995) Nonparametric likelihood confidence bands for a distribution function. *Journal of the American Statistical Association* 90(430):516–521. https://doi.org/10.2307/2291062.

[44] Frey J (2008) Optimal distribution-free confidence bands for a distribution function. *Journal of Statistical Planning and Inference* 138:3086–3098. https://doi.org/10.1016/j.jspi.2007.12.001.

[45] Liu Y, Tewfik A (2013) Empirical likelihood ratio test with distribution function constraints. *IEEE Transactions on Signal Processing* 61(18):4463–4472. https://doi.org/10.1109/icassp.2013.6638886.

[46] Goldman M, Kaplan DM (2015) Evenly sensitive KS-type inference on distributions: new computational, Bayesian, and two-sample contributions. University of California–San Diego Working Paper. http://econweb.ucsd.edu/ mrgoldman/KG2013.pdf.

[47] Bennett RL, Kim ND, Curran JM, Coulson SA, Newton AWN (2003) Spatial variation of refractive index in a pane of float glass. *Science & Justice* 43(2):71–76. https://doi.org/10.1016/S1355-0306(03)71746-8.

[48] Curran JM (2011) *Introduction to Data Analysis with* R *for Forensic Scientists* (CRC Press, Boca Raton, FL). https://doi.org/10.1201/9781420088274.

[49] R Core Team (2017) R*: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria). http://www.R-project.org/.

[50] Grätzer G (2011) *Lattice Theory: Foundation* (Birkhäuser, Basel, Switzerland). https://doi.org/10.1007/978-3-0348-0018-1.

[51] Jadhav SN (2011) *Generating, Classifying and Indexing Large Scale Fingerprints*. Master of Engineering Report, Computer Science and Engineering, Indian Institute of Science, Bangalore, India.

[52] NBIS (2015) NIST Biometric Image Software, Version 5.0.0, NIST authors: Kenneth Ko and Wayne J. Salamon, last accessed on March 24, 2017, from https://www.nist.gov/services-resources/software/nist-biometric-image-software-nbis.

[53] JCGM (2008) *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement*. Joint Committee for Guides in Metrology, International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008; BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections. http://www.bipm.org/en/publications/guides/gum.html.

[54] Possolo A (2015) Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results. U.S. Department of Commerce, Washington, D.C., NIST Technical Note 1900. https://doi.org/10.6028/nist.tn.1900.

[55] Peterson WW, Birdsall TG (1953) The Theory of Signal Detectability -. Part I. General Theory. Department of Electrical Engineering, Engineering Research Institute, University of Michigan, Ann Arbor, MI. Electronic Defense Group Technical Report 13. http://hdl.handle.net/2027.42/7068.

[56] Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics* (Wiley, New York).

[57] Green DM, Swets JA (1974) *Signal Detection Theory and Psychophysics* (Robert E. Krieger Publishing Co.,

Huntington, NY), a reprint with corrections of the original 1966 edition.

[58] Biedermann A, Taroni F, Aitken CGG (2015) Liberties and constraints of the normative approach to evaluation and decision in forensic science: a discussion towards overcoming some common misconceptions. *Law, Probability and Risk* 13(2):181–191. https://doi.org/10.1093/lpr/mgu009.

[59] Hájek A (2008) Dutch book arguments. *The Oxford Handbook of Rational and Social Choice*, eds Anand P, Prasanta, P, Clemens P (Oxford University Press, Oxford, United Kingdom). https://doi.org/10.1093/acprof:oso/9780199290420.001.0001.

[60] Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A* 231:289–337. https://doi.org/10.1098/rsta.1933.0009.

[61] Leegwater AJ, Didier M, Sjerps M, Vergeer P, Alberink I (2017) Performance study of a score-based likelihood ratio system for forensic fingermark comparison. *Journal of Forensic Sciences* 62(3):626–640. https://doi.org/10.1111/1556-4029.13339.

***About the authors:*** *Steve Lund and Hari Iyer are mathematical statisticians in the Statistical Engineering Division of the Information Technology Laboratory at NIST. The National Institute of Standards and Technology is an agency of the U.S. Department of Commerce.*