Improving Score-based Likelihood Ratios via Stacking

February 21, 2020

Abstract

Score-based likelihood ratios are the most promising alternative to feature-based likelihood ratios for the evaluation of the strength of forensic evidence. The construction of effective general score functions, however, has received little, if any, attention. Many scores are measures of

1 Introduction

Score-based likelihood ratios (SLRs) are a popular, tractable alternative to likelihood ratios (LRs) for the quantification of the strength of forensic evidence [Bolck et al., 2009, Hepler et al., 2012, Davis et al., 2012, Bolck et al., 2015, Neijmeijer, 2016, Galbraith and Smyth, 2017, Leegwater et al., 2017, Morrison and Enzinger, 2018, Chen et al., 2018]. Instead of considering distributions of the raw observed features in the data, SLRs instead compute a low-dimensional statistic of the features called a *score*, the distribution of which is then modeled under the competing hypotheses of the prosecution and defense. Often scores are a measure of dissimilarity between two materials of evidence: expected to be small when the two materials share a source and large when they do not. Recent work has shown that supervised machine learning models can successfuly discriminate between two materials of evidence that come from the same versus different sources [Hare et al., 2017, Carriquiry et al., 2019, Park and Carriquiry, 2019], and the predicted probabilities from such models can result in an effective score.

Despite the early successes of score-based methods, there is currently a lack of theoretical justification for their use as surrogates for the true, feature-based LR. Several authors have pointed out issues with certain types of dissimilarity-based SLRs such as their potential *incoherence* [Armstrong, 2017, Neumann and Ausdemore, 2019], as well as the fact that different types of SLRs using distributions conditioned on different pieces of evidence may not agree with each other [Hepler et al., 2012]. Yet, Garton et al. [2020] showed that by constructing an SLR in a natural way, one could derive general probabilistic bounds on the LR given a score. Further, they showed that if the LR is bounded, then the SLR must share the same bounds. This work suggested that score-based methods

often produce SLRs that are less conclusive than the true LR, consistent with observations in Bolck et al. [2009, 2015], but that the SLR and LR are also often directionally consistent.

Choosing a score, even among a predefined set, remains a potentially challenging task. Tippett plots [Evett and Buckleton, 1996, Drygajlo et al., 2003] or receiver operating characteristic (ROC) curves can be used to select between scores, and while useful, these metrics do not have a straightforward relationship to the information contained in the score. Empirical cross-entropy (ECE) plots [Ramos and Gonzalez-Rodriguez, 2013, Ramos et al., 2013] do have connections to information theory and have been shown to be useful in assessing the performance of a score, yet their connection to sufficiency has yet to be elucidated. Moreover, it may be desireable to have a one or two number summary of score performance in addition to a figure.

Additionally, one might imagine that several scores could be combined into a score that is better than any individual score, but there is no existing literature, to our knowledge, that describes how this could be done or that evaluates the advantages of doing so.

The aim of this work is to address these issues. First, we argue that SLRs are not incoherent, and that the valid problem described as incoherence by several authors can be better understood as a problem of choosing an appropriate score function. We propose a class of ways for how this might be done. This further motivates our following work in which we propose a score performance measure that is direcly linked to the sufficiency of the score to select between the prosecution and defense models. We use this performance measure to empirically show how our proposed class of scores typically outperforms scores that are measures of dissimilarity. Finally, we show that we can use a probabilistic classifier to aggregate scores into a single SLR, which is often better than any individual score by our sufficiency performance measure.

The remainder of this paper is organized as follows. Section 2 introduces the general problem of determining whether a piece of evidence with unknown source has the same origin as a piece of evidence with known origin. We define and compare the *common source* and *specific source* problems in our general mathematical setup. In Section 3 we describe our proposed measure of the sufficiency of a given score and provide its derivation in the context of our problem. In Section 4 we argue that the seeming incoherence of SLRs is better described as a problem of selecting an appropriate score function. We also propose a general way of constructing score functions based on measures of dissimilarity that we can empirically show typically outperform typical, dissimilarity-only scores. In Section 5 we propose to partially resolve the problem of choosing among several scores by aggregating them via a probabilistic classifier, which we call *score stacking*. Then, in Section 6, we perform two simulated experiments comparing an aggregated score to individual scores. The second simulated experiment uses a data generating model that was developed in the literature for modeling chemical concentrations on copper wire and found to plausibly approximate the true data generating process [Dettman et al., 2014]. Finally, in Section 7, we discuss our results, challenges to practical implementation, and possible directions for

future research.

2 Common source versus specific source LRs

To our knowledge, the first recognition of the important differences between a *common source* (CS) and *specific source* problem was in Ommen [2017]. The CS problem is to determine whether multiple pieces of evidence, all with unknown origin, have the same, but still unknown, origin. One might be interested in this problem if multiple crimes were thought to be linked, but no suspect has yet been identified. Alternatively, the 'specific source' (SS) problem is to determine whether evidential material coming from an unknown source, such as a shoeprint at a crime scene, has the same origin as evidential material of known origin, such as a shoeprint collected directly from a suspect's shoe.

We denote by $H_{cs} \in \{p, d\}$ the random variable associated with the CS hypotheses. We use $A_i \in \mathcal{A} \equiv \{1, ..., N_A\}$ where $i \in \mathcal{I} \equiv \{1, ..., I\}$ $(I \geq 2)$ to denote discrete random variables representing the sources of evidence. Here, the subscript i indexes the particular piece of evidence. Every piece of evidence is associated with a source random variable, A_i. Note that if multiple pieces of evidence have the same source, then N_A, the number of unique sources, must be less than I, the total number of pieces of evidence. In both the CS and SS problems, there will be at least one source of evidence that is *always* of unknown origin. It will often be useful to use A_u to denote the random variable/vector indicating the source of evidence which has unknown origin *and* for which we are attempting to understand something about the source. That is, we reserve the subscript $u \in \mathcal{I}$ to represent the index which identifies the evidence material whose source we are primarily interested in. The key difference between the CS and SS problem is whether or not we pair materials of unknown origin with those of known origin (SS) or also unknown origin (CS). We will use subscript i to denote the pieces of evidence for which we are uninterested in source information and whose purpose is to aid in modeling.

The distributions for A_i and A_u are defined conditionally based on whether $H_{cs} = p$ or $H_{cs} = d$. In the specific source case, the prosecution defines a statistical hypothesis wherein the source of the unknown evidence, A_u, is the same as one of the known sources. We denote this hypothesis by the conditional random variable $H_{cs}|\{A_i : i \neq u\}$. This is mathematically equivalent to inferring the probability of the *event* $A_u = A_k|\{A_i : i \neq u\}$ where A_k, for $k \in \mathcal{I} \setminus u$, is the source that the prosecution is attributing to E_u. We suppose that $E_i \in \mathbb{R}^d$ are vectors of random variables representing evidence in the form of some data coming from sources A_i. We will also use E_u to denote the evidence coming from the source A_u. In the remainder, we will exclusively concern ourselves with the SS problem.

2.1 Specific Source LR Example

We now provide a hypothetical example of a SS problem and how to define the appropriate statistical models based on our notation. Suppose that a suspect in a crime has been apprehended and is in possession of a shoe which may have been the source of a shoeprint at the crime scene. The forensic scientist may create a print from the suspect's shoe and subsequently produce a 2-D image from the print. Assume that the forensic scientist also has a database of 10 images taken from shoe prints of the identical brand and size of shoe as the suspect's but that each of the 10 images corresponds to distinct shoes. The prosecution lawyers then define a hypothesis that the source of the shoe print image from the crime scene is the same source as that of the image from the suspect's shoe. The defense lawyers alternatively state that the source of the crime scene print image is any one of sources of images in the database of 10 images. In this problem, data comes in the form of shoeprint images, each consisting of the same number of pixels.

Let us reframe this problem mathematically using the notation introduced earlier. First, there are $I = 12$ pieces of evidence: the suspect's shoeprint, the crime scene shoeprint, and the 10 database shoeprints. Thus, $\mathcal{I} \equiv \{1, ..., 12\}$. However, there are only $N_A = 11$ possible values for each A_i corresponding to either the suspect's shoe or one of the 10 database shoes. Let us define A_k (where $k \in \mathcal{I}$) and A_u to be the sources of the evidence resulting from the suspect's shoe and from the crime scene, respectively. In this situation, $A_k = c \in \mathcal{A}$ is known.

Let us define the events implied by the random variable H_{cs}. Let $\{H_{cs} = p\} \equiv \{A_u = A_k\}$ and $\{H_{cs} = d\} \equiv \{A_u \neq A_k\}$. Note that the event $\{H_{cs} = p\}$ does *not*, by itself, imply that $A_u = A_k = c$. That is, the claim that the the source of u-th and k-th shoeprint images have the same source is not the same as stating which is the source of both images.

In this SS example, the prosecution's hypothesis prior to observing any data is that $\{A_u = A_k | A_i \text{ for } i \in \mathcal{I} \setminus u\}$. Note that because $k \in \mathcal{I} \setminus u$ and because we condition on A_i for $i \in \mathcal{I} \setminus u$, the prosecution's hypothesis specifies a specific value for A_u. The defense's hypothesis is the complement of the prosecution's hypothesis. We will denote these *conditional* events using shorter notation involving a new random variable, H_{ss}, representing the specific source hypothesis. That is, we write these two events as $\{H_{ss} = p\} \equiv \{H_{cs} = p | A_i : i \in \mathcal{I} \setminus u\}$ or $\{H_{ss} = d\} \equiv \{H_{cs} = d | A_i : i \in \mathcal{I} \setminus u\}$ in the prosecution and defense's case, respectively.

Finally, we can define the likelihood ratio in terms of the distribution for $\{E_i\}_{i \in \mathcal{I}}$, conditionally on each specific source hypothesis. Let F denote the joint distribution for the entire collection of random variables with corresponding density f. That is, F is a distribution over $(\{E_i\}_{i \in \mathcal{I}}, \{A_i\}_{i \in \mathcal{I}})$. Then the specific source likelihood ratio can be written equivalently as

$$LR_{ss} = \frac{f(\{e_i\}_{i\in\mathcal{I}}|A_i : i \in \mathcal{I}\setminus u, A_u = A_k)}{f(\{e_i\}_{i\in\mathcal{I}}|A_i : i \in \mathcal{I}\setminus u, A_u \neq A_k)} \tag{1}$$

$$= \frac{f(\{e_i\}_{i\in\mathcal{I}}|A_i : i \in \mathcal{I}\setminus u, H_{cs} = p)}{f(\{e_i\}_{i\in\mathcal{I}}|A_i : i \in \mathcal{I}\setminus u, H_{cs} = d)}. \tag{2}$$

We have used e_i to denote the values that the E_i vectors take in the density function. It is reasonable to assume that given all of the A_i, the E_i are all independent. Further, given a single A_i, the distribution of the corresponding E_i does not depend on the other sources A_j where $j \neq i$. We make these assumptions in the rest of this work. The first these two assumptions results in an LR that depends on the data only through E_u.

$$\begin{aligned} LR_{ss} &= \frac{f(e_u|A_u = A_k, A_i : i \in \mathcal{I}\setminus u)}{f(e_u|A_u \neq A_k, A_i : i \in \mathcal{I}\setminus u)} \\ &= \frac{f(e_u|H_{cs} = p, A_i : i \in \mathcal{I}\setminus u))}{f(e_u|H_{cs} = d, A_i : i \in \mathcal{I}\setminus u)} \\ &= \frac{f(e_u|H_{ss} = p)}{f(e_u|H_{ss} = d)}. \end{aligned}$$

Note that we have abused notation in the distributions for E_u given H_{ss}. Using our previous notation, $f(e_u|H_{ss} = h) = f(e_u|\{H_{cs} = h|A_i : i \in \mathcal{I}\setminus u\})(h \in \{p, d\})$, which is notationally ambiguous. Our intention by notationally conditioning on H_{ss} is to, in fact, condition on $\{H_{cs} = h, A_i : i \in \mathcal{I}\setminus u\}$. In words, we say that the prosecution's hypothesis is that the source of the crime scene shoeprint image is the suspect's shoe. The defense's hypothesis is then that the source of the crime scene shoeprint is one of the shoes from the database. Note that the distribution of e_u *does not* depend on any of the A_i's *other than* A_u and A_k under the prosecution's hypothesis. However, the distribution of e_u under the defense's hypothesis depends on *all* the A_i's. This is because, loosely speaking, the probability of the data under the defense hypothesis depends on a mixture of the probabilities of the data assuming its source was each of the A_i's not equal to A_k. Mathematically, this corresponds to the following expression for $f(e_u|A_i : i \in \mathcal{I}\setminus u, H_{cs} = d)$

$$f(e_u|A_i : i \in \mathcal{I}\setminus u, H_{cs} = d) = \sum_{i\in\mathcal{I}\setminus u} f(e_u|A_u = A_i : i \in \mathcal{I}\setminus u)p(A_i),$$

where $p(\cdot)$ is a probability distribution over $\mathcal{A}\setminus c$. This distribution allows the defense to specify which, if any, of the database sources is more likely to be the source of E_u.

3 Information theoretic specific source score sufficiency metric

Consider the specific source problem with arbitrary, but finite, number of possible sources N_A. The following derivations are very similar to those in the 'infinite alternative population' situation considered in Garton et al. [2020]. Recall that we assume mutual independence between E_i conditional on A_i in addition to $E_i \perp A_j | A_i$ for $j \neq i$, the specific source LR is

$$LR_{ss} = \frac{f(e_u|H_{ss} = p)}{f(e_u|H_{ss} = d)}.$$

Thus, we reiterate that the LR depends only on the evidence from the unknown source, E_u. We will now introduce a score function, $s(\cdot)$, which will map the I pieces of evidence to a real number. That is, s is defined as $s : \left(\mathbb{R}^d\right)^I \to \mathbb{R}$ (s is a function of $(E_1, E_2, ..., E_I)$ which are each in $\mathbb{R}^d$). We now show that we can write the LR in terms of both the evidence of unknown origin *and* the score. This will allow us to decompose the LR in a useful way. Note that the LR can be written as,

$$\begin{aligned} LR = \frac{f(e_u|H_{ss} = p)}{f(e_u|H_{ss} = d)} &= \frac{f(s|e_u, H_{ss} = p)f(e_u|H_{ss} = p)}{f(s|e_u, H_{ss} = d)f(e_u|H_{ss} = d)} \\ &= \frac{f(s, e_u|H_{ss} = p)}{f(s, e_u|H_{ss} = d)} \\ &= \frac{f(e_u|s, H_{ss} = p)f(s|H_{ss} = p)}{f(e_u|s, H_{ss} = d)f(s|H_{ss} = d)}. \end{aligned}$$

Because $S|(E_u = e_u, H_{ss})$ is a function only of the evidence of known origin, $\{E_i\}_{i\in\mathcal{I}\setminus u}$, and because $\{E_i\}_{i\in\mathcal{I}\setminus u} \perp H_{cs} | \{A_i\}_{i\in\mathcal{I}\setminus u}$, we have that

$$\begin{aligned} \frac{f(s|e_u, H_{ss} = p)}{f(s|e_u, H_{ss} = d)} &= \frac{f(s|e_u, H_{ss} = p)}{f(s|e_u, H_{ss} = d)} \\ &= \frac{f(s|e_u)}{f(s|e_u)} \\ &= 1. \end{aligned}$$

This justifies the first line of the above. The remaining lines just follow from standard rules regarding conditional and joint distributions.

The fact that the distribution of the score is independent of the SS hypothesis conditioned on the unknown source evidence may, at first, appear strange. Recall, however, that the SS LR is entirely independent of all E_i except for E_u. Thus, for a fixed E_u, a score is dependent only on the evidence from the known sources, which is independent of H_{ss}. One may ask why we bother to state distributions for the known source evidence at all, if they are irrelevant. The reason is that, in practice, samples from the known sources are required

to infer what the distributions for E_u might be under $H_{ss} = p$ and $H_{ss} = d$. In the case of the matching of shoeprint images, the forensic scientist wouldn't need a shoeprint known to be generated from the suspect's shoe if they could be provided with those unique characteristics that are consistently observable on shoeprints from the suspect's shoe across multiple scenarios. For a statistician, even if they could define with certainty the correct family of distributions of relevant characteristics on a shoeprint taken from a suspect's shoe, they would still need samples to choose the distribution from within the family.

Using these facts, we can then decompose the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] of the data under the specific source prosecution hypothesis in the following useful way,

$$
\begin{aligned}
KL\Big(F(\{E_i\}_{i\in\mathcal{I}}|H_{ss}=p)\Big|\Big|F(\{E_i\}_{i\in\mathcal{I}}|H_{ss}=d)\Big) = \\
&= E\left[\log\frac{f(\{e_i\}_{i\in\mathcal{I}}|H_{ss}=p)}{f(\{e_i\}_{i\in\mathcal{I}}|H_{ss}=d)}\middle|H_{ss}=p\right] \\
&= E\left[\log\frac{f(e_u|H_{ss}=p)}{f(e_u|H_{ss}=d)}\middle|H_{ss}=p\right] \\
&= E\left[E\left[\log\frac{f(e_u,s|H_{ss}=p)}{f(e_u,s|H_{ss}=d)}\middle|S,H_{ss}=p\right]\right] \\
&= E\left[E\left[\log\frac{f(e_u|s,H_{ss}=p)}{f(e_u|s,H_{ss}=d)}+\right.\right. \\
&\qquad \left.\left.\log\frac{f(s|H_{ss}=p)}{f(s|H_{ss}=d)}\middle|S,H_{ss}=p\right]\right] \\
&= E\left[E\left[\log\frac{f(e_u|s,H_{ss}=p)}{f(e_u|s,H_{ss}=d)}\middle|S,H_{ss}=p\right]\right]+ \\
&\qquad\qquad E\left[\log\frac{f(s|H_{ss}=p)}{f(s|H_{ss}=d)}\middle|H_{ss}=p\right] \\
&= E\left[KL\Big(F(E_u|S,H_{ss}=p)\Big|\Big|F(E_u|S,H_{ss}=d)\Big)\middle|H_{ss}=p\right] \\
&\qquad\qquad + KL\Big(F(S|H_{ss}=p)\Big|\Big|F(S|H_{ss}=d)\Big).
\end{aligned}
$$

This implies that $KL\Big(F(\{E_i\}_{i\in\mathcal{I}}|H_{ss}=p)\Big|\Big|F(\{E_i\}_{i\in\mathcal{I}}|H_{ss}=d)\Big) \geq KL\Big(F(S|H_{ss}=p)\Big|\Big|F(S|H_{ss}=d)\Big)$[1].

An additional consequence is that for finite, feature-based KL divergences, larger values of $KL\Big(F(S|H_{ss}=p)\Big|\Big|F(S|H_{ss}=d)\Big)$ imply smaller values of $E\left[KL\Big(F(E_u|S,H_{ss}=p)\Big|\Big|F(E_u|S,H_{ss}=d)\Big)\right]$. To see this, note that because $KL\Big(F(E_u|S,H_{ss}=p)\Big|\Big|F(E_u|S,H_{ss}=p)\Big)$ is a nonnegative function in terms of

[1]Note that a more general proof of this inequality was given by Theorem 4.1 in Kullback and Leibler [1951], where equality holds if and only if S is a sufficient statistic.

S, small values of $E\left[KL\Big(F(E_u|S,H_{ss}=p)\Big|\Big|F(E_u|S,H_{ss}=d)\Big)\right]$ imply small values, on average, of $KL\Big(F(E_u|S,H_{ss}=p)\Big|\Big|F(E_u|S,H_{ss}=d)\Big)$. For example, if the expectation is zero, then the (conditional) KL divergence is zero almost everywhere. Zero KL divergence implies that $F(E_u|S,H_{ss}=p)=F(E_u|S,H_{ss}=d)$, i.e. S is a sufficient statistic for the specific source hypothesis.

All of this means that $KL\Big(F(S|H_{ss}=p)\Big|\Big|F(S|H_{ss}=d)\Big)$ and $KL\Big(F(S|H_{ss}=d)\Big|\Big|F(S|H_{ss}=p)\Big)$ are measures of the usefulness of the score which have direct ties to sufficiency. Assuming that $E\left[\left|\log\frac{f(s|H_{ss}=p)}{f(s|H_{ss}=d)}\right|\Big|H_{ss}=p\right]<\infty$ and $E\left[\left|\log\frac{f(s|H_{ss}=d)}{f(s|H_{ss}=p)}\right|\Big|H_{ss}=d\right]<\infty$, consistent score KL divergence estimates are always computable because estimates of the densities $f(s|H_{ss}=p)$ and $f(s|H_{ss}=d)$ (or their ratio directly) are available by assumption. They are also intuitive targets to maximize. For example, if the score is a predicted class probability for 'match', the more discriminative the classifier, the larger the score KL divergences and so the closer the score is to being sufficient. Thus, the score KL divergences are a natural performance metric that can be used to compare multiple scores.

4 Coherence and specific source SLRs

Concern has been raised in the literature on LRs about a desirable property ostensibly absent from SS SLRs. The property, dubbed *coherence*, intuitively says that given two mutually exhaustive hypotheses, H_1 and H_2, the likelihood ratio used to compare hypothesis one to hypothesis two should be the reciprocal of that used to compare hypothesis two to hypothesis one. We will argue that the legitimate problem with SLRs identified by Armstrong [2017] and Neumann and Ausdemore [2019] should not be characterized as a lack of coherence, but rather a subtlety relating to the choice of an appropriate score function. Specifically, we will show that the standard argument as to why SLRs are incoherent can be understood as the comparison of two SLRs based on different score functions.

This line of thought then leads to natural questions about how to construct scores even in the presence of an agreed upon dissimilarity metric. We propose several ways to construct an appropriate score function and demonstrate that the resulting SLRs are both coherent and superior to standard scores via simulations.

We digress for a moment to mention that the necessity of SLRs being coherent differs depending on whether the SLR is going to be used to infer the posterior probability of H_{ss} given the score in a Bayesian way. While this is usually the intention behind the estimation of the SS LR, there is debate in the literature about whether SLRs can or should be used this way in court [NATE: Danica? :) REFS]. To the authors of this work, coherence seems an intuitively desireable property regardless of the intent to use SLRs in a Bayesian way. How-

ever, our main motivation for discussing coherence in this work is that it clearly highlights the challenges associated with selecting a score function, and allows us to examine ways to construct more informative scores.

4.1 Coherence

Denote by $E \equiv (E_1^\top, E_2^\top, ..., E_I^\top)^\top \in \times(\mathbb{R}^d)^I$ the vector of random variables describing *all* of the observed evidence or data which will be used to evaluate the relative likelihood of the two hypotheses. As the ensuing discussion in this section and that in Section 4.2 is applicable to SS and CS LRs and SLRs, we temporarily drop the CS and SS subscripts from the LRs and SLRs. Define by $LR_{i,j} \equiv \frac{f(e|H_i)}{f(e|H_j)}$ the likelihood ratio of hypothesis i to hypothesis j. The coherency principal is satisfied if

$$LR_{i,j} = \frac{1}{LR_{j,i}}.$$

Likelihood ratios are fundamentally coherent, but what about score-based likelihood ratios? Denote by $s : (\mathbb{R}^d)^I \to \mathbb{R}^q$ a score function mapping the original data to Euclidean space of dimension q (typically $q = 1$). Similar to LRs, denote by $SLR_{i,j} \equiv \frac{f(s(e)|H_i)}{f(s(e)|H_j)}$ the score-based likelihood ratio comparing hypothesis i to hypothesis j. We briefly note that in this general context SLRs are also coherent.

4.2 Coherence of specific source SLRs

Let us examine the arguments presented in Armstrong [2017] and Neumann and Ausdemore [2019] for the incoherence of SLRs. These arguments stem from an example where there are two *known* sources of evidence say, source $A_1 = a_1$ and source $A_2 = a_2$, each producing data E_1 and E_2, respectively. Furthermore, assume that we have a third piece of evidence of unknown origin, E_u, which must have come from either A_1 or A_2. We then wish to evaluate the support of the data for H_1 or H_2, which are defined as follows

$$\begin{aligned} H_1 :&\quad E_u \text{ was generated from source } A_1 = a_1 \\ H_2 :&\quad E_u \text{ was generated from source } A_2 = a_2. \end{aligned}$$

Note that in this case, $u = 3$, according to our notational convention. We also have $LR_{1,2} = \frac{f(e_1,e_2,e_u|H_1)}{f(e_1,e_2,e_u|H_2)}$. We make use of *all* available data in the formulation of the numerator and denominator densities. Under our previously stated assumptions, the LR reduces to $LR_{1,2} = \frac{f(e_u|H_1)}{f(e_u|H_2)}$.

Armstrong [2017] and Neumann and Ausdemore [2019] then both consider possible SLRs for this example. Importantly, they define the score so that it is explicitly a function only of two materials of evidence, which are arguably the most common kinds of scores in the literature [Bolck et al., 2009, Hepler et al., 2012, Davis et al., 2012, Bolck et al., 2015, Armstrong, 2017, Leegwater et al., 2017, Galbraith and Smyth, 2017, Chen et al., 2018, Neumann and Ausdemore,

2019]. That is, their score maps $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. A common example of such a score is Euclidean distance, i.e. $s(x, y) = \left[\sum_{i=1}^{k}(x_i - y_i)^2\right]^{1/2}$. Such a score makes perfect sense in a typical SS problem context in which only two materials of evidence are considered: one from the known source and one from the unknown source.

However, when one desires to create an SLR based on this score in this particular example, it is tempting to suggest that the natural definition of the SLR is $SLR_{1,2} = \frac{f(s(e_1,e_u)|H_1)}{f(s(e_1,e_u)|H_2)}$. Yet, the natural SLR if the hypotheses were reversed is $SLR_{2,1} = \frac{f(s(e_2,e_u)|H_2)}{f(s(e_2,e_u)|H_1)}$. Neither of these SLRs is the reciprocal of the other, and so the specific source SLR appears to be incoherent.

The confusion arises due to the fact that that the score is not constructed so as to explicitly be a function of *all* available data. When we consider these SLRs in the more general context of scores depending on all available data, we see that what we have define to be $SLR_{1,2}$ and $SLR_{2,1}$ turn out to be two different SLRs depending on two different scores.

For clarity, we will use $s(\cdot)$ to denote scores which are explicitly functions of *all* observed data, and we will use $\delta(\cdot)$ to denote score functions which are only a function of two materials of evidence/data. We must also define the coordinate mapping function $T_{i,j} : \left(\mathbb{R}^d\right)^3 \to \left(\mathbb{R}^d\right)^2$ to be $T_{i,j}(E_1^\top, E_2^\top, E_u^\top) = (E_i^\top, E_j^\top)^\top$. Specifically, the score in $SLR_{1,2}$ is $s_1(e_1^\top, e_2^\top, e_u^\top) = \delta(T_{1,3}(e_1^\top, e_2^\top, e_u^\top))$ and the score in $SLR_{2,1}$ is $s_2(e_1^\top, e_2^\top, e_u^\top) = \delta(T_{2,3}(e_1^\top, e_2^\top, e_u^\top))$. While the functional form of the score in the two SLRs *appears* to be the same, they are actually two different functions resulting from using two different coordinate maps. Thus, the two SLRs are two distinct options for a single SLR whose relationship needn't be expected to be related any more than if one had decided to use two different functional forms of $\delta(\cdot, \cdot)$ in the two separate SLRs.

It is not immediately obvious, then, how one should go about constructing a score that explicitly depends on all observed data. One possibility would be to consider a vector valued score function $s(e) = (\delta(e_u, e_1), ..., \delta(e_u, e_{I-1}))$ (where we suppose that u is the I-th index). However, such an approach becomes infeasible if I is large. We would like to construct a *univariate* score that explicitly depends on all pieces of evidence. In the two source case, two possible scores would be

$$s_1(e_u, e_1, e_2) = v\left(\frac{\delta(e_u, e_1)}{\delta(e_u, e_2)}\right) \tag{3}$$

$$s_2(e_u, e_1, e_2) = v\left(\delta(e_u, e_1) - \delta(e_u, e_2)\right), \tag{4}$$

where $v : \mathbb{R} \to \mathbb{R}$ is some monotonic function. Intuitively, under H_1, $\delta(e_u, e_1) > \delta(e_u, e_2)$, while under H_2, the opposite should be true. This would mean that both scores would be relatively large under H_1 and small under H_2

4.3 Example of a coherent SLR in the two source problem

Consider the specific, two source problem where $A_1 = a_1$ and $A_2 = a_2$ are both known. Suppose that our hypotheses are defined such that

$$\begin{aligned} H_1: &\quad E_1 \sim \mathcal{N}(0,1), E_2 \sim \mathcal{N}(2,1), E_u \sim \mathcal{N}(0,1) \\ H_2: &\quad E_1 \sim \mathcal{N}(0,1), E_2 \sim \mathcal{N}(2,1), E_u \sim \mathcal{N}(2,1) \end{aligned}$$

where E_u, E_1, E_2 are mutual independent under both H_1 and H_2. Note that, in this example, we know that the specific source LR is given by

$$LR_{ss} = \exp\Big\{ -\frac{1}{2}\left[e_u^2 - (e_u - 2)^2\right] \Big\}.$$

We will examine three different SLRs: $SLR_{ss}^{(1)} \equiv \frac{f(s_1(E)|H_1)}{f(s_1(E)|H_2)}$, $SLR_{ss}^{(2)} \equiv \frac{f(s_2(E)|H_1)}{f(s_2(E)|H_2)}$, and $SLR_{ss}^{(3)} \equiv \frac{f(s_3(E)|H_1)}{f(s_3(E)|H_2)}$, where

$$\begin{aligned} E &= (E_1, E_2, E_u)^\top \\ s_1(E) &= \log\|E_u - E_1\|^2 \\ s_2(E) &= \log\|E_u - E_2\|^2 \\ s_3(E) &= \log \frac{\|E_u - E_1\|^2}{\|E_u - E_2\|^2}. \end{aligned}$$

Figure 1 shows scatterplots of $\log(LR_{ss})$ versus $\log(SLR_{ss})$ for each of the three scores. To calculate each SLR, we estimate densities separately using kernel density estimation, and take their ratio. We see hints that the differences between the SS LR and the SS SLR depend on whether H_1 or H_2 is true for the first two scores, as the distribution of points does not appear similar or in some way symmetric under H_1 as compared to H_2. By contrast, this does not seem to be the case for the third score.

Table 1 provides Monte Carlo estimates of the KL divergence of the raw data, that is $\int \log(LR_{ss}) f(e_u|H_{ss} = p) de_u$ and $\int \log(LR_{ss}^{-1}) f(e_u|H_{ss} = d) de_u$, as well as the KL divergences based on the score only, or $\int \log(SLR_{ss}) f(s|H_{ss} = p) ds$ and $\int \log(SLR_{ss}^{-1}) f(s|H_{ss} = d) ds$. Also provided is the RMSE calculated as

$$RMSE = \sqrt{\frac{1}{10^5} \sum_{i=1}^{10^5} (\log LR_i - \log SLR_i)^2}$$

under each hypothesis. These quantities are calculated for each score under consideration.

Based on Table 1, we see that the score with the largest KL divergence under *both* hypotheses is the third. This is the score that was designed to use all of the observed data. By comparison, the other two scores both perform more strongly under one hypothesis than the other. The third score outperforms even the best performance from either of the first two scores. Using the RMSE

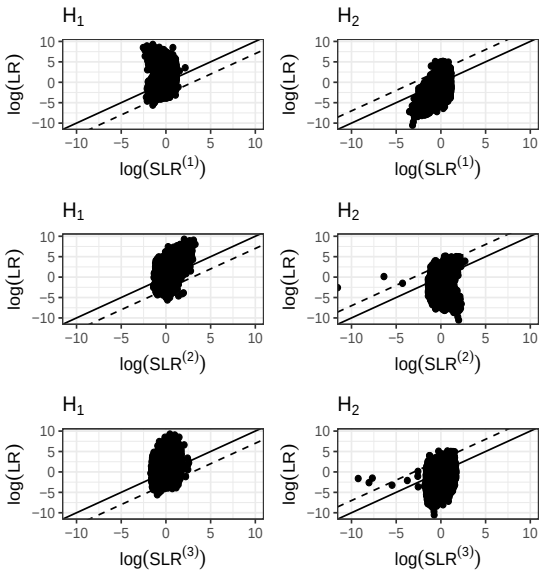

Figure 1: LR_{ss} versus SLR_{ss} scatterplots under hypothesis 1 and 2 using three types of SLRs based on the three presented scores. The first two would be considered by Armstrong [2017], Neumann and Ausdemore [2019] to be incoherent, while the third is one of our proposed scores depending explicitly on all data. Solid lines represent perfect agreement between the LR and the SLR while the dashed line corresponds to a conservative 95% lower confidence bound under H_1 or upper bound under H_2.

Hypothesis	Score	Feature KL	Score KL	RMSE
1	1	4.48 (0.03)	0.35 (0.01)	2.68
1	2	4.48 (0.03)	0.43 (0.01)	2.20
1	3	4.48 (0.03)	0.6 (0.01)	2.27
2	1	6.92 (0.07)	0.4 (0.01)	2.21
2	2	6.92 (0.07)	0.36 (0.01)	2.66
2	3	6.92 (0.07)	0.6 (0.01)	2.25

Table 1: KL divergences for scores one, two, and three as well as for the full data under H_1 and H_2. Monte Carlo standard errors are given in parentheses. Also shown is the RMSE comparing each $\log SLR_{ss}$ to $\log LR_{ss}$. The above results are based on 10,000 simulated data sets under both H_1 and H_2.

as the performance metric, we see that the best score is the second under H_1 and the first under H_2, though the third score is relatively close under both hypotheses. Scores one and two perform notably worse than the third score under hypotheses one and two, respectively. Thus, we see that score three is closer to a sufficient statistic for H_1 and H_2, and this results in a SS SLR that is typically a better estimate, in terms of the RMSE, to the feature-based SS LR.

4.4 Generalizing to the multisource case

It might be nice to, in general, be able to construct a reasonable score given a dissimilarity measure, $\delta(\cdot,\cdot)$, defined in terms of two pieces of evidence. We will provide one general method for constructing such a score. First, we return to the situation where we have N_A sources, one of which is the source of the evidence from an unknown source. The task is to compare the hypothesis that the unknown source evidence was generated by a specific, known source $A_u = A_k = c \in \mathcal{A} \equiv \{1, 2, ..., N_A\}$ to the hypothesis that the unknown source evidence was generated by any one of the other sources $A_u = b \in \mathcal{A} \setminus c$.

Let's consider a class of possible scores based off of an accepted dissimilarity measure, $\delta(\cdot,\cdot) \geq 0$. Let, $g(\delta(e_u, e_1), ..., \delta(e_u, e_{k-1}), \delta(e_u, e_{k+1}), ..., \delta(e_u, e_I))$ be a function mapping dissimilarities between e_u and e_i for $i \in \mathcal{I} \setminus \{u, k\}$ to the real line. Define the score to be

$$s(e_1, ..., e_{N_A}, e_u; g) \equiv \log \frac{\delta(e_u, e_k)}{g(\delta(e_u, e_1), ..., \delta(e_u, e_{k-1}), \delta(e_u, e_{k+1}), ..., \delta(e_u, e_I))}.$$

We could define

$$g_1(\delta(e_u, e_1), ..., \delta(e_u, e_{k-1}), \delta(e_u, e_{k+1}), ..., \delta(e_u, e_I)) = \min_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i),$$

in which case we would get the following score

$$s(e_1, ..., e_{N_A}, e_u; g_1) = \log \frac{\delta(e_u, e_k)}{\min_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}.$$

Another example could be to define

$$g_2(\delta(e_u, e_1), ..., \delta(e_u, e_{k-1}), \delta(e_u, e_{k+1}), ..., \delta(e_u, e_I)) = \sum_{i \in \mathcal{I} \setminus \{u,k\}} w(i)\delta(e_u, e_i),$$

resulting in the following score

$$s(e_1, ..., e_{N_A}, e_u; g_2) = \log \frac{\delta(e_u, e_k)}{\sum_{i \in \mathcal{I} \setminus \{u,k\}} w(i)\delta(e_u, e_i)},$$

where $w(i)$ are weights with $\sum_{i \in \mathcal{I} \setminus \{u,k\}} w(i) = 1$. Intuitively, the first score should perform well. The dissimilarity in the numerator should be compared with the smallest dissimilarity in $\mathcal{A} \setminus c$. Plainly, if the smallest dissimilarity measured between the unknown source evidence and the database evidence is smaller than the dissimilarity between the unknown source evidence and A_k, then that must suggest that A_k was not the source of E_u.

4.5 Multisource example

We now consider a multisource example where we have $N_A = 10$ known sources. For simplicity, we will again assume that all evidence is generated from independent, univariate Gaussian distributions. In this situation, it makes sense to use $H_{ss} = p$ and $H_{ss} = d$ instead of H_1 and H_2 because the 'defense' hypothesis does not now specify which of the alternative $N_A - 1$ sources is that from which E_u was generated. Under both hypotheses, we assume that for $i \in \mathcal{I} \setminus u$, we have $E_i \overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma_i^2)$. We assume the following two data generating distributions for the unknown source evidence, E_u,

$$\begin{aligned} H_{ss} = p: &\quad E_u \sim \mathcal{N}(\mu_k, \sigma_k^2) \\ H_{ss} = d: &\quad E_u \sim \mathcal{GMM}\left(\{\mu_i\}_{i \in \mathcal{I} \setminus \{u,k\}}, \{\sigma_i^2\}_{i \in \mathcal{I} \setminus \{u,k\}}, \{\pi_i\}_{i \in \mathcal{I} \setminus \{u,k\}}\right). \end{aligned}$$

All random variables are assumed to be independent conditional on each hypothesis. We use $\mathcal{GMM}\left(\{\mu_i\}_{i=1}^{N_A-1}, \{\sigma_i^2\}_{i=1}^{N_A-1}, \{\pi_i\}_{i=1}^{N_A-1}\right)$ to denote a Gaussian mixture model with $N_A - 1$ mixture components. The means of the mixing components are $\{\mu_i\}_{i=1}^{N_A-1}$, variances $\{\sigma_i^2\}_{i=1}^{N_A-1}$, and mixture probabilities $\{\pi_i\}_{i=1}^{N_A-1}$ ($\sum_{i=1}^{N_A-1} \pi_i = 1$). The marginal density of this model can be written as

$$f(e_u | H_{ss} = d) = \sum_{i \in \mathcal{I} \setminus \{k,u\}} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{1}{2}\frac{(e_u - \mu_i)^2}{\sigma_i^2}\right\} \pi_i.$$

For this example, we draw $\mu_i \overset{iid}{\sim} \mathcal{N}(0,1)$ for $i \in \mathcal{I} \setminus \{u,k\}$. We set $\mu_k = 4$. All component variances were set to one, that is $\sigma_i^2 = 1$. Finally, we used $\pi_i = \frac{1}{N_A - 1}$.

Like the two source example, we consider three possible scores that appropriately use all of the data. Each score corresponds to using a different summary statistic to aggregate the dissimilarities between the unknown source evidence and the alternative source population. Denote by k, the index in $\mathcal{I}$ corresponding to source N_A. Also, let $E = (E_1, ..., E_{10}, E_u)$

$$s_1(e) = \log \frac{\delta(e_u, e_k)}{\min_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}$$
$$s_2(e) = \log \frac{\delta(e_u, e_k)}{\max_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}$$
$$s_3(e) = \log \frac{\delta(e_u, e_k)}{\frac{1}{N_A - 1} \sum_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}$$

Figure 2 provides scatterplots of $\log LR$ versus $\log SLR$ for each of three scores based on 10,000 simulated observations under $H_{ss} = p$ and $H_{ss} = d$. We again use kernel density estimation to calculate the SLRs. Solid lines represent $\log LR = \log SLR$ and dashed lines correspond to a conservative 95% lower confidence bound under $H_{ss} = p$ or upper bound under $H_{ss} = d$ [Garton et al., 2020]. All three scores visually appear to perform similarly. There appears to be more agreement between the LR and the SLR under $H_{ss} = p$ than $H_{ss} = d$. Under $H_{ss} = d$, we tend to see less agreement for smaller values of the LR (as well as the SLR), where the LR tends to be much smaller than the SLR. This is consistent with observations made in Garton et al. [2020].

Table 2 provides the feature-based KL divergences as well as the score-based KL divergences under both hypotheses. Also provided are RMSEs calculated as in the two-source example. Based on the score KL divergences, the best score under $H_{ss} = p$ is the one which uses the min function in the score denominator, while the best score under $H_{ss} = d$ uses the max. Using the average in the denominator provides a compromise between the two and always performs in the middle. This means that the max is the worst under $H_{ss} = p$ and the min is the worst under $H_{ss} = d$. RMSEs order the performance of scores in the same way as the KL divergences.

5 Score stacking

It is clear that we can devise scores utilizing all available data, which are likely to substantially improve upon similarity based scores of just two materials of evidence. In the two-source case, one can simply compare the two dissimilarities calculated when comparing the known materials of evidence to the unknown material. In the multisource case, one can consider a statistic summarizing the dissimilarities between the known source evidence materials and the unknown

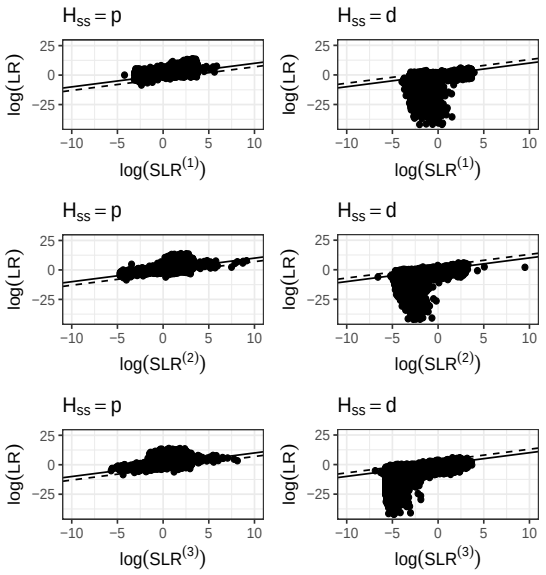

Figure 2: log(LR) versus log(SLR) scatterplots under $H_{ss} = p$ and $H_{ss} = d$ using three types of SLRs which correspond to using different statistics to aggregate dissimilarity scores in the alternative source population. We try min, max, and average, corresponding to rows 1-3, respectively. Solid lines represent perfect agreement between the LR and the SLR while the dashed line corresponds to a conservative 95% lower confidence bound under $H_{ss} = p$ or upper bound under $H_{ss} = d$. Results are based on 10,000 observations for each hypothesis.

Hypothesis	Score	Feature KL	Score KL	RMSE
P	Min	4.48 (0.03)	2.31 (0.02)	3.11
P	Avg	4.48 (0.03)	2.08 (0.01)	3.48
P	Max	4.48 (0.03)	1.75 (0.01)	3.85
D	Min	6.92 (0.07)	1.87 (0.01)	8.39
D	Avg	6.92 (0.07)	2.83 (0.02)	7.53
D	Max	6.92 (0.07)	2.91 (0.02)	6.93

Table 2: KL divergences for scores one, two, and three as well as for the full data under $H_{ss} = p$ and $H_{ss} = d$. Also shown is the RMSE comparing each $\log SLR$ to $\log LR$. The above results are based on 10,000 simulated data sets under both $H_{ss} = p$ and $H_{ss} = d$.

source evidence. However, this raises the question as to the choice of statistic that should be used, as we provided an example showing that the performance of a given score may depend heavily on whether $H_{ss} = p$ or $H_{ss} = d$ is true.

We now describe how multiple scores can be combined via a probabilistic classifier into a single score from which an SLR can be calculated without any density estimation. We will show in the following section that combining scores in such a way often outperforms each individual score on both hypotheses. We call the act of combining multiple scores score *stacking*. This comes from terminology in the machine learning literature where multiple predictive models are combined through a meta-model to improve predictive performance [Hastie et al., 2009, Chapter 8.8]. Using a probabilistic classifier turns out to be an intuitively reasonable method for combining scores as the objective function used to learn the classifier is highly related to the ability of the classifier to separate the data based on the hypothesis. Additionally, the probabilistic classifier allows us to bypass estimating individual score densities in the calculation of SLRs, which is not reliable in high dimensions [Sugiyama et al., 2010].

We are now tasked with choosing viable probabilistic classifiers. We prefer to use nonparametric functions to aggregate scores for their flexibility. We also require that the classifier can be efficiently trained on at least thousands of data points. One candidate that fits both of these criteria is a sparse Gaussian process (GP) classifier (see, for example, Chapter 8 of Rasmussen and Williams [2006] or Quiñonero-Candela and Rasmussen [2005] for more information on sparse Gaussian processes and Chapter 3 of Rasmussen and Williams [2006] for GP classification). We use the algorithm described in [REFS submitted JMLR paper] to select the number and placement of knots, and we also use the same Gaussian posterior approximation. However, when aggregating scores, we use the automatic relevance determination (ARD) covariance function $k_\theta(x, x') = \sigma^2 \exp\left\{-\frac{1}{2}\sum_{i=1}^{q} \frac{(x_i - x_i')^2}{\ell_i^2}\right\}$, where $\theta = (\sigma, \ell_1, ..., \ell_q)$, and q is the number of scores being aggregated.

We assume that we have $J/2$ i.i.d. draws from the score distributions under both $H_{ss} = p$ and $H_{ss} = d$. Let x_i denote the i-th vector of the q score function evaluations. We then form a new dataset by treating $H_{ss} = p$ and $H_{ss} = d$ as class labels with the corresponding predictor variables being the values of each of the score functions. Specifically, we use a response variable where

$$y_i = \begin{cases} 1, & \text{if } x_i \text{ was generated under } H_{ss} = p \\ 0, & \text{if } x_i \text{ was generated under } H_{ss} = d \end{cases}.$$

Once we have a trained classifier, for a new x^* we can then produce estimates of $P(y^* = 1|x^*)$. Note that this has the interpretation of a posterior probability even though we have not directly specified a prior distribution over class labels. This prior is implicitly given in the proportion of class 1 to class 0 labels in our training set. One can calculate an estimate of the new SS SLR by noting that

$$\frac{P(y^* = 1|x^*)}{P(y^* = 0|x^*)} = \frac{P(x^*|y^* = 1)}{P(x^*|y^* = 0)} \frac{P(y^* = 1)}{P(y^* = 0)}.$$

Table 3: Scores used to estimate SLRs for the multivariate normal data.

Score	Abbreviation
$\log \frac{\|\|e_u - e_k\|\|}{\min_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}$	Min
$\log \frac{\|\|e_u - e_k\|\|}{\frac{1}{9} \sum_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}$	Avg
$\log \frac{\|\|e_u - e_k\|\|}{\max_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}$	Min
Stacked Min, Avg, Max	Agg

In order to calculate and estimate of the new SLR, we need only divide the posterior odds by the prior odds.

6 Experiments

We now consider two simulation studies showing that aggregating scores with a probabilistic classifier can result in a score that outperforms individual scores under both hypotheses.

6.1 Multivariate Normal Data

In this example we use the same data generating model as considered in Section 4.5, except that now we sample $\mu_i \overset{iid}{\sim} \mathcal{N}(0, 2)$ for $i \in \mathcal{I} \setminus \{u, k\}$, so that discriminating between $H_{ss} = p$ and $H_{ss} = d$ is a harder task. We calculate each SLR by treating scores as predictor variables in a binary classification as described in the previous section. Note that we can do this even for single individual scores, not just for score aggregation. This has been suggested as a better alternative to density ratio estimation than first individually estimating densities and then taking the ratio [Sugiyama et al., 2010]. However, we use only 1000 observations simulated under each hypothesis due to the training time required for the sparse GP. Table 3 enumerates the scores used to estimate SLRs for this study.

Figure 3 shows scatterplots of the $\log(LR)$ against the $\log(SLR)$ under $H_{ss} = p$ and $H_{ss} = d$. The rows in the figure correspond to using min, max, or average, in the denominator of the score function. The fourth row corresponds to an aggregation of these three scores using a sparse GP classifier. The scatterplots show similar patterns to those in Figure 2. However, now we see that the sparse GP classifier tends result in an SLR that better agrees with the true LR under $H_{ss} = p$. It also appears that the sparse GP classifier is able to better estimate the true LR when it is small under $H_{ss} = d$.

Table 4 shows the estimated full data and score based KL divergences in addition to the RMSEs calculated using the difference between the $\log LR$ and the $\log SLR$. We see that under both hypotheses, the KL divergence of the aggregated score is either the largest or statistically indistinguishable from the

largest KL divergence of the non-aggregated scores, indicating that it is closer to being sufficient than any of the individual scores. The RMSEs tell a similar story.

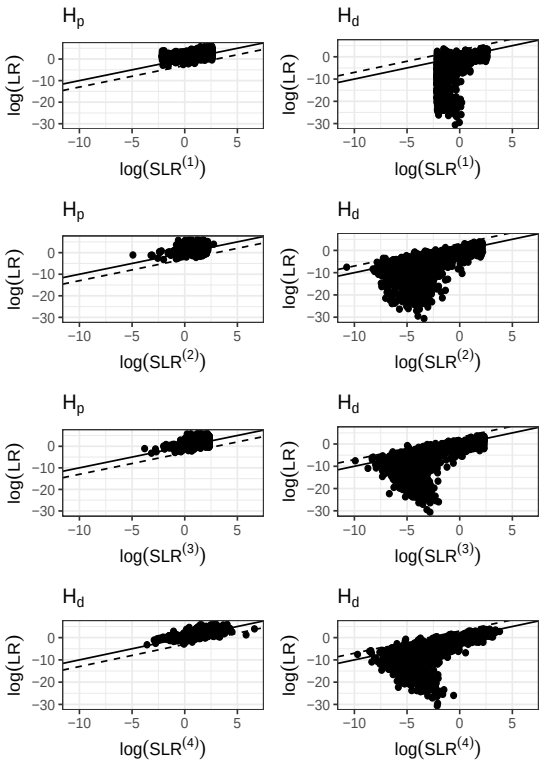

Figure 3: log(LR) versus log(SLR) scatterplots under $H_{ss} = p$ and $H_{ss} = d$ using three types of SLRs which correspond to using different statistics to aggregate dissimilarity scores in the alternative source population. We try min, max, average, and a sparse GP classifier corresponding to rows 1-4, respectively. Solid lines represent perfect agreement between the LR and the SLR while the dashed line corresponds to a conservative 95% lower confidence bound under $H_{ss} = p$ or upper bound under $H_{ss} = d$. Results are based on 1,000 observations for each hypothesis.

6.2 Copper Wire Synthetic Data

We now compare the performance of five different SS SLRs on a more realistic synthetic data set. Dettman et al. [2014] examine whether trace element concentrations in copper wire can be used to discriminate between samples from the same or different sources. In the process of doing this, they develop a plausible generative model for eight trace chemical concentrations within copper wire.

Table 4: KL divergences for four types of scores which correspond to using different statistics to aggregate dissimilarity scores in the alternative source population under $H_{ss} = p$ and $H_{ss} = d$. We try min, max, average, and a sparse GP classifier. We also provide estimates of the true KL divergences based on the true data generating model. Monte Carlo standard errors for estimates of the KL divergences are provided in parentheses. Also shown is the RMSE comparing each $\log SLR$ to $\log LR$. The above results are based on 1,000 simulated data sets under both $H_{ss} = p$ and $H_{ss} = d$.

Hypothesis	Score	Feature KL	Score KL	RMSE
P	Min	2.47 (0.04)	1.42 (0.04)	1.72
P	Max	2.47 (0.04)	1.57 (0.03)	1.64
P	Avg	2.47 (0.04)	1.69 (0.03)	1.50
P	Agg	2.47 (0.04)	1.96 (0.04)	1.20
D	Min	7.69 (0.22)	1.19 (0.04)	9.18
D	Max	7.69 (0.22)	2.93 (0.08)	7.40
D	Avg	7.69 (0.22)	3.06 (0.08)	7.35
D	Agg	7.69 (0.22)	2.98 (0.08)	7.51

The details of this model can be found in their supplementary information. We use this generative model to simulate five data sets where $N_A = 500$. That is, there are 499 database sources which are sampled from the relevant background population, and there is one known source, which, if matched with the copper wire sample found on the suspect, would be incriminating. We simulate five data sets each with 4000 observations, half of which are generated under $H_{ss} = p$ and half of which are generated under $H_{ss} = d$, on which sparse GP models are trained in order to estimate each SLR. Each simulated data set is created by simulating one mean vector for each of the 500 sources, which stays constant for each of the 4000 observations. The 4000 observations per data set represent within source variability. For each training data set, an additional data set, also with 4000 observations and for which the *same* mean vectors as the training set are used, is generated independently in order to estimate KL divergences. The five scores we consider are enumerated in Table 5. We use Agg as an abbreviation for aggregated. This is the stacked score using the Min, Avg, and Max scores.

Figure 4 shows estimated score KL divergences with Monte Carlo standard error bars, though standard errors are so small that they are hard to see. We can see that under $H_{ss} = p$, the Agg score is almost always at least as good as the best of the other scores, and when it is not, as in the first run, it is close enough that the error bars overlap. Under $H_{ss} = d$, the results are mixed. For runs one and two, the stacked score performs about as well as the Min score, which are both better than the others. However, in runs three, four, and five, the Agg and Min scores are the worst.

Table 6 provides estimates of the feature-based KL divergences along with Monte Carlo standard errors in parantheses for each of the five runs. We see

Table 5: Scores used to estimate SLRs for the synthetic copper wire data.

Score	Abbreviation
$\log \|\|e_u - e_k\|\|$	Delta
$\log \frac{\|\|e_u - e_k\|\|}{\min_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}$	Min
$\log \frac{\|\|e_u - e_k\|\|}{\frac{1}{499} \sum_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}$	Avg
$\log \frac{\|\|e_u - e_k\|\|}{\max_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}$	Min
Stacked Min, Avg, Max	Agg

Table 6: Feature-based KL divergences as well as Monte Carlo standard errors in parentheses for five simulated copper wire data sets under $H_{ss} = p$ and $H_{ss} = d$.

Hypothesis	Feature KL	Run
P	113.69 (0.32)	1
D	8366.55 (300.36)	1
P	59.21 (0.22)	2
D	34574.43 (375.96)	2
P	29.04 (0.15)	3
D	8373.05 (331.85)	3
P	35.35 (0.16)	4
D	6858.07 (291.41)	4
P	28.21 (0.15)	5
D	8048.87 (323.79)	5

that the feature-based KL divergence under the defense model tends to take values between approximately 6000 and 35000, while under the prosecution model values of 28 to 113 are typical. Additionally, both of these feature-based KL divergences are much larger than the respective score-based divergences, which tend to be around 9 to 25 under the defense model and 5 to 13 under the prosecution model. Note that the KL divergence is a lower bound on the log expected value of the LR (or SLR). This means that the smallest expected SLRs under the prosecution model are roughly $e^5 \approx 150$ and $e^9 \approx 8000$ under the defense model. However, it would not be unreasonable to observe SLR values of $e^{13} = 440,000$ under the prosecution model and $e^{25} = 7.2 \times 10^{10}$ under the defense model. Thus, while there should be substantial room for improving the given SLRs, they are all able to provide compelling evidence for one hypothesis over the other.

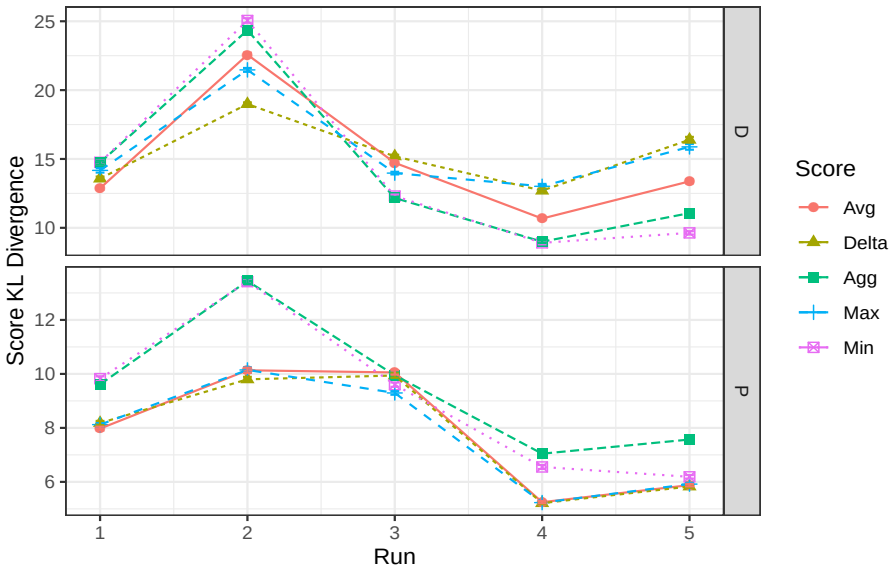

Figure 4: Score KL divergences and Monte Carlo standard errors under $H_{ss} = p$ and $H_{ss} = d$ for five randomly generated synthetic copper wire data sets.

7 Discussion

We've shown that choosing a good score function for score-based likelihood ratios is, perhaps, a more challening and subtle task than previously thought. We've shown that the concerns raised in the literature about the coherence of SLRs can be understood in terms of how the score function mapping all relevant observed data to a, typically univariate, quantity is defined. This gave rise to questions about how to better use all available data, and we have proposed a set of ways for how to construct scores which are based on given dissimilarity measures, but that improve upon purely dissimilarity based scores because they are explicitly a function of all available data. Furthermore, we've proposed a bivariate, continuous measure of the sufficiency of a score for the specific source hypothesis that can be used to compare the usefulness of a collection of scores. This allowed us to experimentally show that our proposed score constructions typically outperform dissimilarity only scores. This measure has an intuitive interpretation as the divergence between the score densities with respect to one or the other score distributions. Finally, we have shown how multiple scores can be aggregated, or stacked, into a single score using a probabilistic classifier, which, in our simulated experiments, performs at least as well as the best individual score. Specifically, we use a sparse Gaussian process classifier, which is useful because it can scale to moderately large data sets and provides the necessary flexibility to learn complicated SLR functions.

Our score sufficiency measure is based on our derivation showing that the

score Kullback-Leibler divergences are lower bounds on the KL divergences using the true data generating distributions, and that when these KL divergences are equal, the score is sufficient for the specific source hypothesis. Despite the direct relationship of this measure to the score property of ultimate importance (the sufficiency of the score for discriminating between $H_{ss} = p$ and $H_{ss} = d$, the fact that it is a bivariate measure complicates its interpretation. If a given score is superior under both $H_{ss} = p$ and $H_{ss} = d$, then it must be that the given score is closer to a sufficient statistic than others considered. However, as we have seen, there are scores that tend to be superior under one hypothesis and inferior under the other. In these situations, the KL divergences still provide useful information, but it is unclear how to use them to make a decision on what score to use. To some extent, we showed experimentally that this can potentially be mitigated by stacking the competing scores. We strongly suspect that, theoretically speaking, stacking scores should produce a score that is no worse than any individual score in terms of the score KL under both prosecution and defense models. However, ensuring that the estimate of the SLR coming from the probabilistic classifier reflects this is potentially difficult and depends on the classifier and the algorithms for training it. In the future, we hope to study other SLR estimation methods summarized in Sugiyama et al. [2010] to assess whether there are more appealing methods in this respect.

One issue that seems generally problematic is, perhaps curiously, when there is an extremely high signal to noise ratio in the raw data, especially if data are scarce; this often appears to be the case with forensic evidence. In this situation, it may be easy to construct a score that, on many training data sets, provides perfect class separation. We observed this when we attempted to run copper wire simulated experiments with small N_A (≤ 50). Without additional assumptions, this makes it impossible to reasonably estimate the SLR. Using a GP (or a sparse GP) classifier imposes assumptions on properties of the SLR such as the smoothness and stationarity of the SLR function. In the case of a sparse GP, the classifier is extremely similar to logistic regression using certain kernel basis functions with an L_2 penalty on the regression coefficients. This avoids the ill-posed optimization that occurs when trying to use logistic regression on classes that are linearly separable. However, results will depend on the kernel hyperparameters. In general, any number of probabilistic classifiers can be used on linearly separable data which produce different SLR estimates which will be impossible to choose between.

One significant hurdle to the practical use of specific source SLRs is that many replicates are needed from the known source, which are often unfeasible to acquire in practice. The situation is potentially worse when one considers the types of scores considered here because many replicates are needed from every single source in the alternative source population. There are at least two possible strategies for dealing with this. The first possibility is to develop realistic data generating algorithms for each relevant type of forensic evidence. Despite the seeming enormity of this challenge, some work has been done to this end for fingerprints with limited success [Abraham et al., 2013, Section 3]. However, one avenue for research that appears yet unexplored is the possibility of using

generative adversarial networks (GANs) [Goodfellow et al., 2014] for the generation of convincing forensic evidence. The success of GANs has been remarkable and frightening in the generation of deepfakes [Korshunov and Marcel, 2018].

The other possible strategy for the specific source data scarcity is to justify a common source SLR in place of the specific source SLR. This would allow for estimation of the score distribution under the prosecutions hypothesis by using sampled scores from many different sources, not just the source in question. This strategy could suffice to justify common source SLR approximations until the relevant data generating algorithms could be shown to work well.

Acknowledgements

This work was partially funded by the 452 Center for Statistics and Applications in Forensic Evidence (CSAFE) 453 through Cooperative Agreement #70NANB15H176 between NIST 454 and Iowa State University, which includes activities carried out at 455 Carnegie Mellon University, University of California Irvine, and 456 University of Virginia.

The authors are also very grateful to Peter Vergeer at the Netherlands Forensic Institute for helpful comments and discussion early on in this work.

Supplementary Information

$$\infty > E\left[(\log LR)^2 | H_{ss} = p\right] \tag{5}$$

$$= E\left[(\log \frac{f(e_u|s, H_{ss} = p)}{f(e_u|s, H_{ss} = p)} + \log \frac{f(s|H_{ss} = p)}{f(s|H_{ss} = p)})^2 | H_{ss} = p\right] \tag{6}$$

$$= E\left[(\log \frac{f(e_u|s, H_{ss} = p)}{f(e_u|s, H_{ss} = p)})^2 + (\log \frac{f(s|H_{ss} = p)}{f(s|H_{ss} = p)})^2 | H_{ss} = p\right]$$
$$+ E\left[2(\log \frac{f(e_u|s, H_{ss} = p)}{f(e_u|s, H_{ss} = d)} \log \frac{f(s|H_{ss} = p)}{f(s|H_{ss} = d)}) | H_{ss} = p\right] \tag{7}$$

$$E\left[(\log\frac{f(e_u|s,H_{ss}=p)}{f(e_u|s,H_{ss}=p)})^2|H_{ss}=p\right]+2E\left[(\log\frac{f(e_u|s,H_{ss}=p)}{f(e_u|s,H_{ss}=d)}\log\frac{f(s|H_{ss}=p)}{f(s|H_{ss}=d)})|H_{ss}=p\right]= \tag{8}$$

$$=E\left[\log\frac{f(e_u|s,H_{ss}=p)}{f(e_u|s,H_{ss}=p)}(\log\frac{f(e_u|s,H_{ss}=p)}{f(e_u|s,H_{ss}=p)}+2\log\frac{f(s|H_{ss}=p)}{f(s|H_{ss}=d)})|H_{ss} \tag{9}$$

$$=E\left[(\log\frac{f(e_u|H_{ss}=p)}{f(e_u|H_{ss}=p)}-\log\frac{f(s|H_{ss}=p)}{f(s|H_{ss}=d)})(\log\frac{f(e_u|H_{ss}=p)}{f(e_u|H_{ss}=p)}+\log\frac{f(s} \tag{10}$$

$$=E\left[(\log\frac{f(e_u|H_{ss}=p)}{f(e_u|H_{ss}=p)})^2|H_{ss}=p\right]-E\left[(\log\frac{f(s|H_{ss}=p)}{f(s|H_{ss}=d)})^2|H_{ss}=p\right] \tag{11}$$

$$=E\left[(\log\frac{f(e_u|H_{ss}=p)}{f(e_u|H_{ss}=p)})^2|H_{ss}=p\right]-E\left[(\log\frac{f(s|H_{ss}=d)}{f(s|H_{ss}=p)})^2|H_{ss}=p\right] \tag{12}$$

$$=E\left[(\log\frac{f(e_u|H_{ss}=p)}{f(e_u|H_{ss}=d)})^2|H_{ss}=p\right]-E\left[(\log\frac{f(s|H_{ss}=p)}{f(s|H_{ss}=d)})^2|H_{ss}=p\right] \tag{13}$$

References

Joshua Abraham, Christophe Champod, Chris Lennard, and Claude Roux. Modern statistical models for forensic fingerprint examinations: a critical review. *Forensic Science International*, 232(1-3):131–150, 2013.

Douglas Armstrong. *Development and Properties of Kernel-based Methods for the Interpretation and Presentation of Forensic Evidence*. PhD thesis, South Dakota State University, 2017.

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

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

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

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

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

Joshua R Dettman, Alyssa A Cassabaum, Christopher P Saunders, Deanna L Snyder, and JoAnn Buscaglia. Forensic discrimination of copper wire using trace element concentrations. *Analytical chemistry*, 86(16):8176–8182, 2014.

Andrzej Drygajlo, Didier Meuwly, and Anil Alexander. Statistical methods and bayesian interpretation of evidence in forensic automatic speaker recognition. In *Eighth European Conference on Speech Communication and Technology*, 2003.

IW Evett and JS Buckleton. Statistical analysis of str data. In *16th Congress of the International Society for Forensic Haemogenetics (Internationale Gesellschaft für forensische Hämogenetik eV), Santiago de Compostela, 12–16 September 1995*, pages 79–86. Springer, 1996.

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

Nathaniel Garton, Danica Ommen, Jarad Niemi, and Alicia Carriquiry. Score-based likelihood ratios to evaluate forensic pattern evidence. *Journal of the Royal Statistical Society: Series A*, 2020. *Submitted*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

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

Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

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

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

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

Cedric Neumann and Madeline A Ausdemore. Defence against the modern arts: the curse of statistics" score-based likelihood ratios". *arXiv preprint arXiv:1910.05240*, 2019.

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

Soyoung Park and Alicia Carriquiry. Learning algorithms to evaluate forensic glass evidence. *The Annals of Applied Statistics*, 13(2):1068–1102, 2019.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6:1939, 2005.

Daniel Ramos and Joaquin Gonzalez-Rodriguez. Reliable support: measuring calibration of likelihood ratios. *Forensic science international*, 230(1-3):156–169, 2013.

Daniel Ramos, Joaquin Gonzalez-Rodriguez, Grzegorz Zadora, and Colin Aitken. Information-theoretical assessment of the performance of likelihood ratio computation methods. *Journal of forensic sciences*, 58(6):1503–1518, 2013.

Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density ratio estimation: A comprehensive review (statistical experiment and its related topics). 2010.