# Improving Score-based Likelihood Ratios via Stacking

January 23, 2020

## 1 Introduction

## 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 suspected to be linked, but no suspect has yet been identified. Alternatively, the "specific source" (SS) problem is to determine whether a fragment of evidence coming from an unknown source, such as evidence at a crime scene, has the same origin as a fragment of evidence of known origin, such as evidence collected directly from a suspect.

We denote by $H \in \{p, d\}$ 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 fragment whose source we are primarily interested in. In the CS problem *all* pieces of evidence have unknown origin, but some pieces of evidence will likely come from a database and whose purpose is to allow us to model relevant probability distributions. We will use subscript $i$ to denote the fragments 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 $H = p$ or $H = d$. 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 $H|\{A_i : i \neq u\}$. This is mathematically equivalent to inferring the probability of the *event* $A_u = a|\{A_i : i \neq u\}$ where $A_i = a$ for some $i \in \mathcal{I} \setminus u$. 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 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 of 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 $H$. Let $H_p \equiv \{H = p\} \equiv \{A_u = A_k\}$ and $H_d \equiv \{H = d\} \equiv \{A_u \neq A_k\}$. Note that the event $\{H = p\}$ 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 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. That is, $\{A_u \neq c|A_i : i \in \mathcal{I} \setminus u\}$. We equivalently write these two events as $\{H = p|A_i : i \in \mathcal{I} \setminus u\}$ or $\{H = d|A_i : i \in \mathcal{I} \setminus u\}$ in the prosecution and defense's case, respectively.

Finally, we can define the likelihood ratio in terms of the distribution, $E_i$,

conditionally on each 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

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

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

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. This results in a dramatic simplification of the LR to

$$LR = \frac{f(e_u|A_u = A_k, A_i : i \in \mathcal{I}\setminus u)}{f(e_u|A_u \neq A_k, A_i : i \in \mathcal{I}\setminus u)} = \frac{f(e_u|A_k, H = p)}{f(e_u|A_i : i \in \mathcal{I}\setminus u, H = d)}.$$

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 $f(e_u|A_i : i \in \mathcal{I}\setminus u, H = d)$

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

where $p(\cdot)$ is a probability distribution over $\mathcal{A}\setminus c$. Going forward, we will condense notation by suppressing $A_i : i \in \mathcal{I}\setminus u$ after the conditioning bar.

# 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 LR is

$$\begin{aligned} LR &= \frac{f(e_u|H_p, \{A_i : i \in \mathcal{I} \setminus u\})}{f(e_u|H_d, \{A_i : i \in \mathcal{I} \setminus u\})} = \frac{f(e_u|A_u = A_k)}{f(e_u|A_u \neq A_k)} \\ &= \frac{f(e_u|A_u = A_k) \prod_{i \in \mathcal{I} \setminus u} f(e_i|A_i)}{f(e_u|A_u \neq A_k) \prod_{i \in \mathcal{I} \setminus u} f(e_i|A_i)} \\ &= \frac{f(e_u|A_u = A_k)}{f(e_u|A_u \neq A_k)}. \end{aligned}$$

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,

$$\begin{aligned} LR = \frac{f(e_u|A_u = A_k)}{f(e_u|A_u \neq A_k)} &= \frac{f(s|e_u, A_u = A_k) f(e_u|A_u = A_k)}{f(s|e_u, A_u \neq A_k) f(e_u|A_u \neq A_k)} \\ &= \frac{f(s, e_u|A_u = A_k)}{f(s, e_u|A_u \neq A_k)} \\ &= \frac{f(e_u|s, A_u = A_k) f(s|A_u = A_k)}{f(e_u|s, A_u \neq A_k) f(s|A_u \neq A_k)}. \end{aligned}$$

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

$$\begin{aligned} \frac{f(s|e_u, A_u = A_k)}{f(s|e_u, A_u \neq A_k)} &= \frac{f(s|e_u, H = p)}{f(s|e_u, H = d)} \\ &= \frac{f(s|e_u)}{f(s|e_u)} \\ &= 1. \end{aligned}$$

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

Using these facts, we can then decompose the KL divergence of the data under the specific source prosecution hypothesis in the following way,

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

This implies that $KL\Big(F(\{E_i\}_{i\in\mathcal{I}}|A_u = A_k)\Big|\Big|F(\{E_i\}_{i\in\mathcal{I}}|A_u \neq A_k)\Big) \geq KL\Big(F(S|A_u = A_k)\Big|\Big|F(S|A_u \neq A_k)\Big)$.

An additional consequence is that larger values of $KL\Big(F(S|A_u = A_k)\Big|\Big|F(S|A_u \neq A_k)\Big)$ imply smaller values of $E\left[KL\Big(F(E_u|S, A_u = A_k)\Big|\Big|F(E_u|S, A_u \neq A_k)\Big)\right]$. To see this, note that because $KL\Big(F(E_u|S, A_u = A_k)\Big|\Big|F(E_u|S, A_u = A_k)\Big)$ is a nonnegative function in terms of $S$, small values of $E\left[KL\Big(F(E_u|S, A_u = A_k)\Big|\Big|F(E_u|S, A_u \neq A_k)\Big)\right]$ imply small values, on average, of $KL\Big(F(E_u|S, A_u = A_k)\Big|\Big|F(E_u|S, A_u \neq A_k)\Big)$. For example, if the expectation is zero, then the (conditional) KL divergence is zero almost everywhere. Zero KL divergence implies that $F(E_u|S, A_u = A_k) = F(E_u|S, A_u \neq A_k)$, i.e. $S$ is a sufficient statistic for the specific source hypothesis.

All of this means that $KL\Big(F(S|A_u = A_k)\Big|\Big|F(S|A_u \neq A_k)\Big)$ and $KL\Big(F(S|A_u \neq A_k)\Big|\Big|F(S|A_u = A_k)\Big)$ are measures of the usefulness of the score which have direct ties to sufficiency. Estimates of these are always computable in practice as estimates of the densities $f(s|A_u = A_k)$ and $f(s|A_u \neq A_k)$ (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], 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. However, 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.

## 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. 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], 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 $H_2$ 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. In this case, we 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], Neumann and Ausdemore [2019] then both consider possible SLRs for this example. However, they define the score so that it is explicitly a function only of two fragments of evidence. 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 fragments 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 Armstrong [2017], Neumann and Ausdemore [2019] 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 fragments 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 $T_{i,j}(E_1^\top, E_2^\top, ..., E_I^\top) = (E_i^\top, E_j^\top)^\top$. 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.

Scores considered in the literature are almost exclusively measures of (dis)similarity between two fragments on evidence [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]

NATE: I need to double check that the scores in these are measures of dissimilarity. 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 values score function $s(e) = (\delta(e_u, e_1), ..., \delta(e_u, e_{I-1}))$ (where we suppose that $u = I$). 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

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

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

where $g : \mathbb{R} \to \mathbb{R}$ 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

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

where $E_u$, $E_1$, $E_2$ are mutual independent under both $H_1$ and $H_2$. We will examine three different SLRs: $SLR^{(1)} \equiv \frac{f(s_1(E)|H_1)}{f(s_1(E)|H_2)}$, $SLR^{(2)} \equiv \frac{f(s_2(E)|H_1)}{f(s_2(E)|H_2)}$, and $SLR^{(3)} \equiv \frac{f(s_3(E)|H_1)}{f(s_3(E)|H_2)}$, 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 $\log(LR)$ versus $\log(SLR)$ for each of the three scores. We see hints that the differences between the LR and the 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. Additionally, the differences between the LR and the SLR exhibit similar patterns when comparing the plot of $\log(LR)$ versus $\log(SLR^{(1)})$ under $H_1$ to the plot of $\log(LR)$ versus $\log(SLR^{(2)})$ under $H_2$ as well as when comparing the plot of $\log(LR)$ versus $\log(SLR^{(1)})$ under $H_2$ and the plot of $\log(LR)$ versus $\log(SLR^{(2)})$ under $H_1$. The reason for this is that $S_1|H_1 \stackrel{d}{=}$

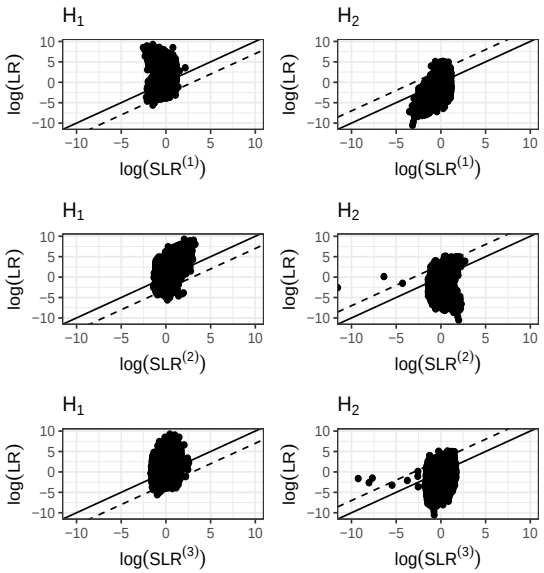


Figure 1: LR versus SLR scatterplots under hypothesis 1 and 2 using three types of SLRs based on the three presented scores. The first two would be considered by Armstrong [2017], Neumann and Ausdemore [2019] to be incoherent, while the third is one of our proposed scores depending explicitly on all 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 $H_2$.

$S_2|H_2$ and $S_1|H_2 \stackrel{d}{=} S_2|H_1$. Thus, $\frac{f(s_1|H_1)}{f(s_1|H_2)}\Big|H_1 \stackrel{d}{=} \frac{f(s_2|H_2)}{f(s_2|H_1)}\Big|H_2$. Equivalently, $\log \frac{f(s_1|H_1)}{f(s_1|H_2)}\Big|H_1 \stackrel{d}{=} -\log \frac{f(s_2|H_1)}{f(s_2|H_2)}\Big|H_2$. Combining this with the fact that, under our data generating models, $\log LR|H_1 \stackrel{d}{=} -\log LR|H_2$ results in

$$\log LR - \log \frac{f(s_1|H_1)}{f(s_1|H_2)}\Big|H_1 \stackrel{d}{=} -\log LR + \log \frac{f(s_2|H_1)}{f(s_2|H_2)}\Big|H_2.$$

In words, the scatterplot of $\log LR$ versus $\log SLR^{(1)}$ under $H_1$ reflected about both the $y$ and $x$-axis should appear the same as the scatterplot of $\log LR$ versus $\log SLR^{(2)}$ under $H_2$.

Table 1 provides Monte Carlo estimates of the KL divergence of the raw data, that is $\int \log(LR) f(e_u|H_p) de_u$ and $\int \log(LR^{-1}) f(e_u|H_d) de_u$, as well as the KL divergences based on the score only, or $\int \log(SLR) f(e_u, e_1, e_2|H_p) d(e_u, e_1, e_u)$ and $\int \log(SLR^{-1}) f(e_u, e_1, e_2|H_d) d(e_u, e_1, e_u)$. 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 as the performance metric, we see that the best scores are either one or two, depending on the hypothesis. However, The third scores performance is competitive under both hypotheses. Contrastingly, scores one and two perform worse 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 an SLR that is typically a better estimate, in terms of the RMSE, to the feature-based LR.

### 4.4 Generalizing to the multisource case

It might be nice to, in general, be able to construct a reasonable score given a "similarity" score, $\delta(\cdot, \cdot)$, 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 two possible scores, both of which will be based off of an accepted dissimilarity metric, $\delta(\cdot, \cdot) \geq 0$. The first score that we will consider is

| Hypothesis | Score | Feature KL | Score KL | RMSE |
|---|---|---|---|---|
| 1 | 1 | 1.98 | 0.35 | 2.66 |
| 1 | 2 | 1.98 | 0.41 | 2.21 |
| 1 | 3 | 1.98 | 0.61 | 2.24 |
| 2 | 1 | 2.02 | 0.40 | 2.23 |
| 2 | 2 | 2.02 | 0.36 | 2.70 |
| 2 | 3 | 2.02 | 0.60 | 2.28 |

Table 1: KL divergences for scores one, two, and three as well as for the full data under $H_1$ and $H_2$. 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 $H_1$ and $H_2$.

$$s_1(e_1, ..., e_{N_A}, e_u) = \log \frac{\delta(e_u, e_k)}{\min_{i \in \mathcal{I} \setminus \{u,k\}} \delta(e_u, e_i)}.$$

The second score is

$$s_2(e_1, ..., e_{N_A}, e_u) = \log \frac{\delta(e_u, e_k)}{\sum_{i \in \mathcal{I} \setminus \{u,k\}} w(i)\delta(e_u, e_i)},$$

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 $E_u$.

In general, there seems to be no reason why a multisource score couldn't be constructed using an arbitrary summary statistic of the "similarity" scores computed between the unknown source evidence and the alternative population.

## 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 $H_p$ and $H_d$ 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 $E_i \overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma^2)$. We assume the following two data generating distributions for the unknown source evidence, $E_u$,

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

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 (therefore $\sum_{i=1}^{N_A-1} \pi_i = 1$) $\{\pi_i\}_{i=1}^{N_A-1}$. The marginal density of this model can be written as

$$f(e_u|H_d) = \sum_{i \in \mathcal{I} \setminus \{k,u\}} \mathcal{N}(\mu_i, \sigma_i^2)\pi_i.$$

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 $\sigma_i = 1$. Finally, we used $\pi_i = \frac{1}{N_A-1}$.

Like the two source example, we consider three possible sources 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 $H_p$ and $H_d$. Solid lines represent $\log LR = \log SLR$ and dashed lines correspond to a conservative 95% lower confidence bound under $H_p$ or upper bound under $H_d$. All three scores visually appear to perform similarly. There appears to be more agreement between the $LR$ and the $SLR$ under $H_p$ than $H_d$. Under $H_d$, 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 $H_p$ is the one which uses the min function in the score denominator, while the best score under $H_d$ uses the max. The 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 $H_p$ and the min is the worst under $H_d$. RMSEs tell the exact same story in terms of the order of the order of performance of each score.

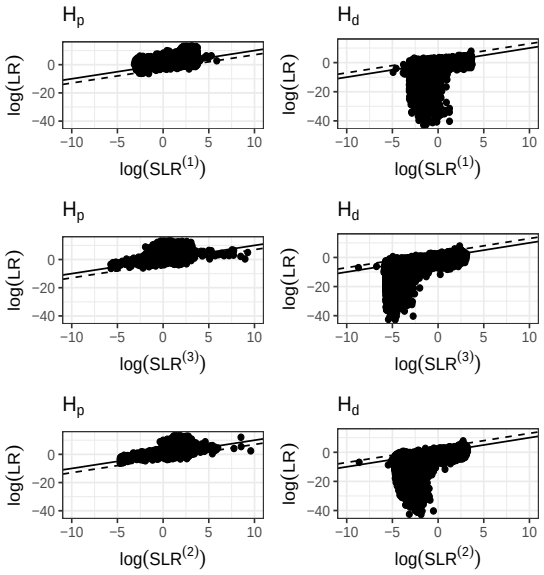


Figure 2: log(LR) versus log(SLR) scatterplots under $H_p$ and $H_d$ using three types of SLRs which correspond to using different statistics to aggregate dissimilarity scores in the alternative source population. We try min, max, and average, corresponding to rows 1-3, respectively. Solid lines represent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

| Hypothesis | Score | Data KL | Score KL | RMSE |
|---|---|---|---|---|
| P | min | 4.41 | 2.29 | 3.06 |
| P | max | 4.41 | 1.73 | 3.77 |
| P | avg | 4.41 | 2.07 | 3.41 |
| D | min | 6.99 | 1.86 | 8.46 |
| D | max | 6.99 | 2.93 | 6.99 |
| D | avg | 6.99 | 2.85 | 7.59 |

Table 2: KL divergences for scores one, two, and three as well as for the full data under $H_p$ and $H_d$. 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 $H_p$ and $H_d$.

# 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 fragments of evidence. In the two-source case, one can simply compare the two dissimilarities calculated when comparing the known fragments of evidence to the unknown fragment. In the multisource case, one can consider a statistic summarizing the dissimilarities between the known source evidence fragments and the unknown 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 $H_p$ or $H_d$ is true.

NATE: This paragraph feels a little awkward 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 computing score densities in the calculation of SLRs. Thus, we are able to eliminate a problematic source of error.

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 [Rasmussen and Williams, 2006, Chapter 8], Quiñonero-Candela and Rasmussen [2005], Snelson and Ghahramani [2006] for more information on sparse Gaussian processes and [Rasmussen and Williams, 2006, Chapter 3] 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.

We assume that we have $J/2$ i.i.d. draws from the score distributions under both $H_p$ and $H_d$. Let $x_i$ denote the vector of score function evaluations at $e_i$, the $i$-th set of evidence. We then form a new dataset by treating $H_p$ and $H_d$ as class labels with the corresponding predictor variables being the values of each of the score functions. Specifically, we use a response variable $y_i = \mathbb{1}[H_i = H_p]$, where

$$\mathbb{1}[H_i = H_p] = \begin{cases} 1, & \text{if } H_i = H_p \\ 0, & \text{otherwise} \end{cases}.$$

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 implicitely given in the proportion of class 1 to class 0 labels in our training set. One can calculate an estimate of the new SLR by noting that

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

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. 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.

Figure 3 shows scatterplots of the $\log(LR)$ against the $\log(SLR)$ under $H_p$ and $H_d$. 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 2. However, now we see that the sparse GP classifier tends result in an SLR that better agrees with the true LR under $H_p$. It also appears that the sparse GP classifier is able to better estimate the true LR when it is small under $H_d$.

Table 3 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.

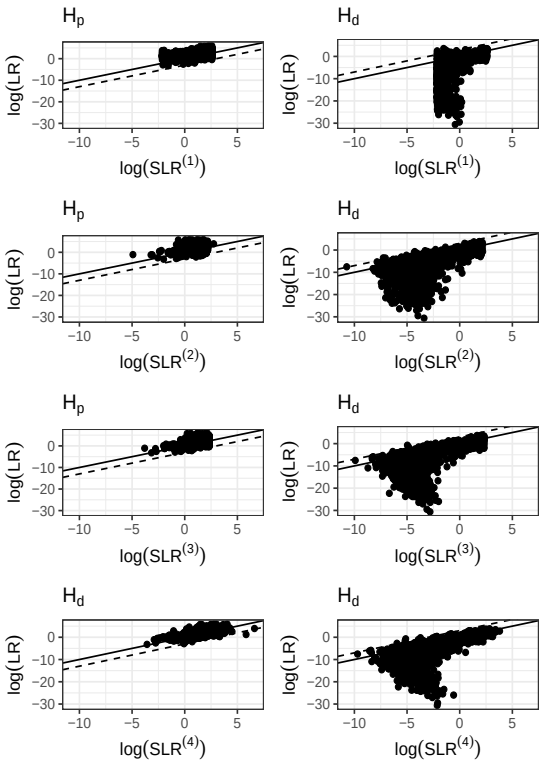


Figure 3: log(LR) versus log(SLR) scatterplots under $H_p$ and $H_d$ using three types of SLRs which correspond to using different statistics to aggregate dissimilarity scores in the alternative source population. We try min, max, average, and a sparse GP classifier corresponding to rows 1-4, respectively.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

| 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 | gp | 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 | gp | 7.69 (0.22) | 2.98 (0.08) | 7.51 |

Table 3: KL divergences for four types of scores which correspond to using different statistics to aggregate dissimilarity scores in the alternative source population under $H_p$ and $H_d$. 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 $H_p$ and $H_d$.

# 7 Discussion

Interestingly, despite the fact that the same data generating model was used in this example as in 4.5, the minimum score actually performed the worst under $H_p$. We believe that this may be due to the different alternative source population means that were generated in the second scenario.

# References


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.

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 hand-

writing 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.

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*.

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.

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.

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.

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

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

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press, 2006. URL http://papers.nips.cc/paper/2857-sparse-gaussian-processes-using-pseudo-inputs.pdf.

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