Forensic Science International 230.1-3 (2013): 156 – 169

DOI: http://dx.doi.org/10.1016/j.forsciint.2013.04.014

Title Page (with authors and addresses)

TITLE:
"Reliable Support: Measuring Calibration of Likelihood Ratios."

AUTHORS:
Daniel Ramos* and Joaquin Gonzalez-Rodriguez

ADDRESS AND AFFILIATIONS OF BOTH AUTHORS:
Research Institute on Forensic Science (ICFS); ATVS - Biometric Recognition Group
Escuela Politecnica Superior, Universidad Autonoma de Madrid
C/ Francisco Tomas y Valiente 11, E-28049 Madrid, Spain.

* Corresponding author. Tel.: +34-91-4976206; fax: +34-91-4972235. Email address: daniel.ramos@uam.es (Daniel Ramos).

Reliable Support:

Measuring Calibration of Likelihood Ratios.

XXX

Abstract

Calculation of likelihood ratios (LR) in evidence evaluation still presents major challenges in many forensic disciplines: for instance, an incorrect selection of databases, a bad choice of statistical models, low quantity and bad quality of the evidence are factors that may lead to likelihood ratios supporting the wrong proposition in a given case. However, measuring performance of LR values is not straightforward, and adequate metrics should be defined and used. With this objective, in this work we describe the concept of calibration, a property of a set of LR values. We highlight that some desirable behavior of LR values happens if they are well calibrated. Moreover, we propose a tool for representing performance, the Empirical Cross-Entropy (ECE) plot, showing that it can explicitly measure calibration of LR values. We finally describe some examples using speech evidence, where the usefulness of ECE plots and the measurement of calibration is shown.

Key words: calibration, empirical cross-entropy, accuracy, likelihood ratio, performance, evidence evaluation

1 Introduction

Despite the increasing acceptance of the likelihood ratio (LR) approach of evidence evaluation in forensic science [1], computation of LR values still remains a challenge. There are many factors that may lead to values of the LR supporting the wrong proposition in a case, an effect known as *misleading evidence* [2]. If this happens, the LR values are said to present bad performance. Those factors may include sparsity of the databases used as populations [3,4], mismatch in the conditions of the elements in the population databases and in the evidence [5,6], degraded quality or quantity of the evidential materials [7–9], and so forth.

Good performance of the LR is essential in casework. Otherwise, misleading LR values in court may lead fact finders to wrong decisions. This idea is the main motivation behind the establishment of validation procedures for evidence evaluation methods, as a way to establish procedures to control and allow the use of LR models in casework. These validation procedures of evidence evaluation methods should be based on a careful process of performance measurement.

Motivated by this critical problem, in this work we adopt a methodology for the measurement of performance of LR methods in forensic science based on so-called Strictly Proper Scoring Rules (SPSR) [10–12] that has solid grounds on Bayesian statistics. The main contribution of this work is highlighting the importance of a property of a set of LR values called *calibration*, and its relationship with the desirable behavior that the LR should have. Also, although the SPSR methodology is not new, we adapt it to the LR framework

for forensic evaluation inference; and we describe a useful representation of the performance of LR values in terms of SPSR and calibration: the Empirical Cross-Entropy (ECE) plot. This methodology for measuring calibration is not intended to replace other methods for measuring performance of the LR, based on *e.g.* Tippett plots or other measurements over the numerator and the denominator of the LR separately. Conversely, we show in this article that measuring the calibration of the LR is an excellent complement to all those methods, in order to have a deep analysis of the performance of the LR with views to a validation process of LR computation in forensic science. In this sense, the example shown in this article illustrate the adequacy and complementarity of using ECE plots in addition to Tippett plots.

Calibration is understood here as a property of a set of LR values, which can be measured. Although the term *calibration* has been recently used to denote a process for obtaining likelihood ratios, we do not follow that meaning in this article. Therefore, our proposal in this article is not about methods to compute the LR, but a methodology to measure the performance and the calibration of a set of LR values, no matter how they were computed. Thus, LR values can be computed using *e.g.* widely accepted models which assign probabilities separately to the numerator and the denominator of the LR (such as the ones described in [13]), and the calibration of the LR values can be measured for those LR values using the methodology proposed in this article.

The article is organized as follows. First, we present the performance assessment methodology based on SPSR, particularizing in the classical example of weather forecasting. Then, we intuitively define and describe the concept of calibration. After that, we give reasons that reveal that it is not straightforward to directly apply this methodology to forensic science, and we describe

the ECE plot as a solution to overcome those difficulties. Finally, we present experimental examples in forensic speaker recognition where the properties of well-calibrated likelihood ratios are highlighted, after which we draw some conclusions.

2 Measuring performance of probabilistic assessments

In this work, we start by adopting a methodology for measuring performance based on Strictly Proper Scoring Rules (SPSR) [10,12], which is not new and has been studied for decades in Bayesian statistics. We begin with a classical example that has motivated abundant research: the elicitation of probabilistic assessments for weather forecasting [14,11].

2.1 Probabilistic Weather Forecasting

Consider an unknown variable, say θ, whose value we want to know. Let θ be binary, which means that it only can take one out of 2 values: either $\theta = \theta_p$ or $\theta = \theta_d$ [1]. In the weather forecasting example we are going to assume that the unknown variable θ refers to a particular day in the future. We therefore denote $\theta^{(i)}$ as the corresponding variable θ for day i. Thus, in that context the values of $\theta^{(i)} \in \{\theta_p, \theta_d\}$, with the following meaning for day i:

- θ_p: it rains in day i.
- θ_d: it does not rain in day i.

[1] We adopt this notation intentionally, because we ultimately aim at the forensic inference problem

A probabilistic weather forecaster, or simply a forecaster, is defined as someone who assigns probabilities for $\theta^{(i)} = \theta_p$ or $\theta^{(i)} = \theta_d$ *before* the value of $\theta^{(i)}$ is known, aiming at predicting its value. The mechanism by which the forecaster assigns probabilities does not need to be known, but it can be said that, as any other probabilistic assignment, it must consider all the knowledge available to the forecaster, say K [15]. The probability that $\theta^{(i)} = \theta_p$ given K is then denoted as $P\left(\theta^{(i)} = \theta_p \middle| K\right)$ which, in words of the forecaster, should be read *the probability that it rains in day i in the future, given all my available knowledge K*. We denote K, the available knowledge, as an *observed* value, in the sense that it is known and fixed. It may include the education, experience and preferences of the forecaster; some data in which the forecaster is basing their assessment; a statistical model; etc. All the resources that are known to the forecaster and used in some way for the elicitation of the probabilistic forecast are included in K, no matter their origin.

For simplicity and convenience, we will eliminate the reference to the day i from the notation when it is clear from the context. Therefore, in those cases we will denote $\theta^{(i)} \equiv \theta$ and $P\left(\theta^{(i)} = \theta_p \middle| K\right) \equiv P\left(\theta_p | K\right)$. Moreover, by definition of θ_p and θ_d, both values have to be complementary, *i.e.*, $P\left(\theta_p | K\right) = 1 - P\left(\theta_d | K\right)$.

We assume that at the end of day i the actual value of $\theta^{(i)}$ in day i and all past days will be known. In other words, at the end of the current day the fact of whether it rained or not in any day in the past will be known. Thus, the forecaster will elicit forecasts for future days from day i, when θ is actually unknown.

Notice that $P\left(\theta^{(i)} = \theta_p \middle| K\right)$ denotes a probability of the value of the variable of interest (θ) given all the available, *observed* knowledge K. In Bayesian inference, this is known as a *posterior* probability, and therefore probabilistic weather forecasters assign posterior probabilities.

2.2 Performance of Probabilistic Assessments: Strictly Proper Scoring Rules

During decades, Bayesian statisticians have been seriously concerned about the elicitation of probabilistic assessments [10,16,17], which can be understood given the Bayesian interpretation of probability as a degree of belief [18,19]. In this topic of research, one of the main questions under study has always been the performance of the probabilistic assessments, that can be summarized as follows: if someone is eliciting probability assessments (according to a given model and data, or based on personal experience), how can we evaluate how they perform?

Contextualizing to our weather forecasting example, we can get some intuition about how to evaluate the performance of one single probabilistic assessment of the forecaster. Imagine that the forecaster assigns a probability of raining for tomorrow (day i) as $P(\theta_p | K) = 0.9$. Then, after two days it turns out that it did actually rain in day i, i.e. $\theta = \theta_p$. As the probability given by the forecaster to the value of θ that actually occured (θ_p) is fairly high, then for that particular probabilistic assessment the forecaster did a good work. Therefore, if an external evaluator would assign a cost (or penalty) to that particular forecast, that penalty should be low. However, if the forecaster would have assigned $P(\theta_p | K) = 0.1$, then that forecast would not have been a good one, since the probability for what it actually happened (it rained,

θ_p) would have been low. These examples suggest that, in order to evaluate a single forecast, two elements are needed: the probability distribution of θ as assigned by the forecaster (the probability of rain in day i, $P(\theta_p|\, K)$), and the actual value of the variable θ, that was unknown by the forecaster, but it is known when performance is to be measured.

According to this intuition in Bayesian statistics the performance of probabilistic assessments has been classically addressed by the use of Strictly Proper Scoring Rules (SPSR) [10–12]. A SPSR is a function both of a probability distribution assigned to a given unknown variable, *and* the actual value of the variable. The value of the SPSR will be interpreted as a *loss* or a *cost* given to the probability distribution depending on the actual value of the variable. In this work we will use the *logarithmic* SPSR, which is defined as follows[2]:

$$\mathcal{C}\left(P\left(\theta_p|\, K\right), \theta\right) = \begin{cases} -\log_2\left(P\left(\theta_p|\, K\right)\right) & \text{if } \theta = \theta_p; \\ -\log_2\left(1 - P\left(\theta_p|\, K\right)\right) & \text{if } \theta = \theta_d. \end{cases} \tag{1}$$

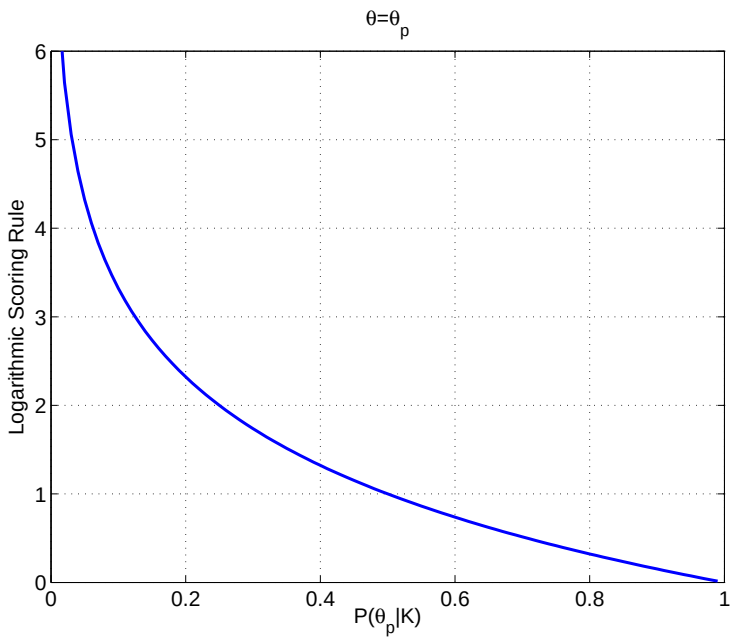
where $\mathcal{C}\left(P\left(\theta_p|\, K\right), \theta\right)$ represents the SPSR as a function of $P\left(\theta_p|\, K\right)$ and the actual value of θ. The intuition behind SPSR will be exemplified with the representation of the logarithmic SPSR in Figure 1. The figure shows the two possible values of the logarithmic SPSR depending on the actual value of θ, as a function of $P\left(\theta_d|\, K\right)$. According to Equation 1, if $\theta = \theta_p$ (it actually rained in day i), the SPSR assigns a high penalty to low values of

[2] There are strong reasons to prefer the logarithmic scoring rule to other SPSR, but they are out of the scope of this work, see [20,21] for details. The base of the logarithm is irrelevant for the expositions. We use base-2 logarithms for information-theoretical reasons, that are explained in [22].

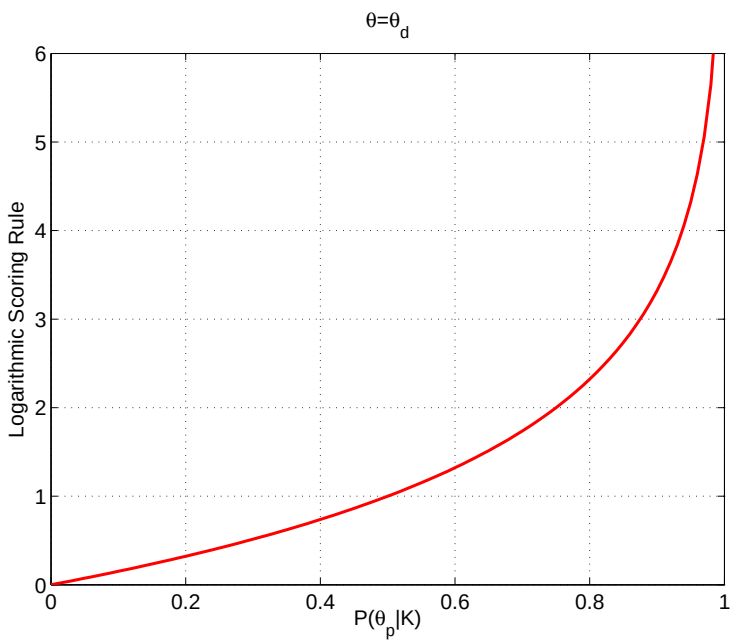
$P(\theta_d|\, K)$, and vice-versa. This corresponds to the fact that, if the weather forecaster expressed a high probability of rain in day i (high $P(\theta_d|\, K)$), and it actually rained ($\theta = \theta_p$), then the penalty should be low, and vice-versa. In the limit, if the forecaster expressed a categorical probability of $P(\theta_d|\, K) = 0$ (*i.e., it is impossible that tomorrow it will rain*), and it actually rained, the penalty will be *infinite* for the logarithmic SPSR[3]. From Figure 1, an analogous reasoning can be followed for the case where $\theta = \theta_d$ (it did not rain in day i), where forecasts expressing high probability of rain (high $P(\theta_d|\, K)$) are severely penalized by the logarithmic SPSR, and vice-versa.

Stricly Proper Scoring Rules measure the goodness of a single forecast from the forecaster. However, an overall performance of a given set of forecasts from a given forecaster should be given. In order to do that, we will firstly need a set of forecasts from the forecaster, with their corresponding actual values of θ for each forecast in the set. We denote the latter the *ground-truth* labels. The whole set of forecasts with their corresponding ground-truth labels will be denoted a *validation* set of *posterior* probabilities. In the context of weather forecasting, the validation set of posterior probabilities can be drawn from an archive of past forecasts from a given forecaster, and the ground-truth labels can be obtained from weather databases where registries of whether it rained those days or not will be available. We use the term *validation* for this set because we assume that the measurement of performance will lead to a validation process where it will be decided whether or not something or someone should be valid for their purpose. In our weather forecaster example,

[3] This is, in fact, one desirable property of the logarithmis SPSR, if it is assumed that someone who categorically expresses a wrong judgement should be the worst possible forecaster.



(a)



(b)

Fig. 1. Representation of the logarithmic Strictly Proper Scoring Rule.

the *validation* process would yield a decision of whether a given forecaster is valid for a given task or not.

In order to measure the accuracy of a validation set of posterior probabilities from the forecaster, the following procedure has been proposed [11]:

$$\mathcal{L}_{\mathcal{C}} = -\frac{1}{N_p} \sum_{\theta^{(i)}=\theta_p} \log_2 \left[P\left(\theta^{(i)} = \theta_p \middle| K\right)\right] - \frac{1}{N_d} \sum_{\theta^{(j)}=\theta_d} \log_2 \left[1 - P\left(\theta^{(j)} = \theta_p \middle| K\right)\right] \tag{2}$$

where $\mathcal{L}_{\mathcal{C}}$ is the accuracy of the validation set of posterior probabilities, and N_p and N_d are the number of forecasts where θ is respectively θ_p or θ_d. Roughly speaking, for the wheather forecaster example, the accuracy of the validation set is the average of the SPSR for the days where it rained, namely $-\frac{1}{N_p} \sum_{\theta^{(i)}=\theta_p} \log_2 \left[P\left(\theta^{(i)} = \theta_p \middle| K\right)\right]$; plus the average of the SPSR for the days where it did not rain, namely $-\frac{1}{N_d} \sum_{\theta^{(j)}=\theta_d} \log_2 \left[1 - P\left(\theta^{(j)} = \theta_p \middle| K\right)\right]$.

2.3 Interpretation as Accuracy

Accuracy is defined as the closeness of a given magnitude to its true value. We illustrate the accuracy interpretation of $\mathcal{L}_{\mathcal{C}}$ in Equation 2 as follows. We define a *perfect* or *oracle* forecaster as the one who assigns probability distributions to θ each day in the light of the true value of θ. Thus, such an oracle forecaster assigns $P(\theta_p| K) = 1$ if $\theta = \theta_p$ and $P(\theta_p| K) = 0$ if $\theta = \theta_d$. For each forecast, the accuracy of the oracle forecaster will be the best possible, and because of that we call it *perfect* accuracy. Also, the SPSR for the oracle forecaster would be always 0 (see Equation 1 and Figure 1). For a forecaster that is not the oracle one, the true value of θ will be unknown for a given day i, and therefore their forecast $P(\theta_p| K)$ will not present perfect accuracy as explained before. Then, the penalty assigned by the SPSR depends on the deviation of the non-

perfect forecaster from the behavior of the oracle forecaster, and therefore the latter it is a measure of the accuracy for the given forecast.

There are other desirable properties of SPSR as measures of accuracy that are out of the scope of this article. Interested readers will find an appropriate formal introduction to the topic in [10,12].

2.4 Calibration

The so-called property of *calibration* of a set of probabilistic assessments has been extensively studied in the past by Bayesian statisticians [14,11]. An intuitive definition of calibration can be reproduced from [16] in the same context as in our weather forecaster example:

> [...] If the metheorologist is using the scale properly, however, we would expect that rain would occur in two-thirds of the days to which he assigns a rain probability of 2/3. This criterion is called calibration [...].

Considering our notation, calibration can be defined as follows. If a forecaster elicits probabilistic assessments about raining in some days, namely $P\left(\theta^{(i)} = \theta_p \middle| K\right)$ with $i \in \{1, 2, ..., N\}$, then, for all probabilistic assessments for which $P\left(\theta^{(i)} = \theta_p \middle| K\right) = p$ the proportion of days where $\theta = \theta_p$ will be p.

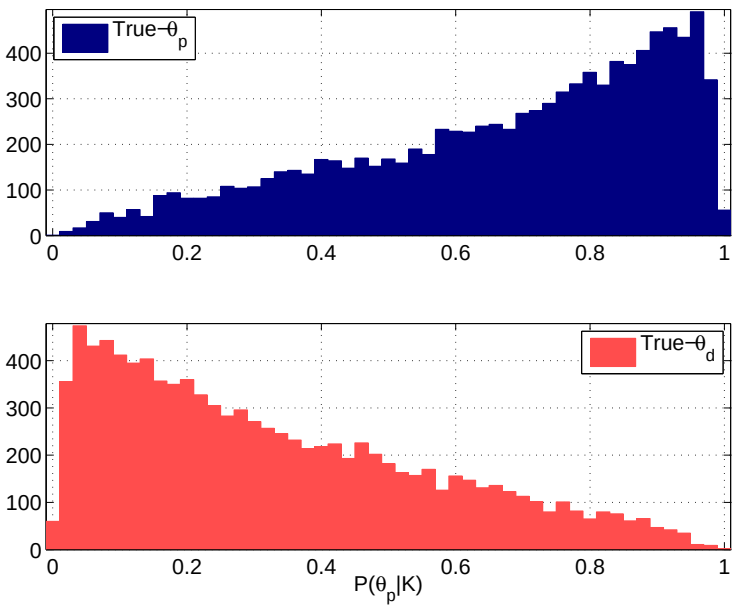
This definition is only useful if it is possible to compute proportions of cases where θ_p is true in the validation set of posterior probabilities for each of the value of such probabilities. That is generally possible if the values of the probabilities that the forecaster can assign is discrete (as it happens in [11]). If the forecaster can elicit any continuous value in the $[0, 1]$ range, then the

definition should consider some kind of partition (or *binning*) of such a range. Then, we can say that a set of probabilistic assessments is well calibrated[4] if, for all assessments $P\left(\theta^{(i)} = \theta_p \middle| K\right)$ with their values within a region (or *bin*) defined by $p \pm \Delta$, with Δ not too large, the proportion of cases where $\theta = \theta_p$ is close to p.

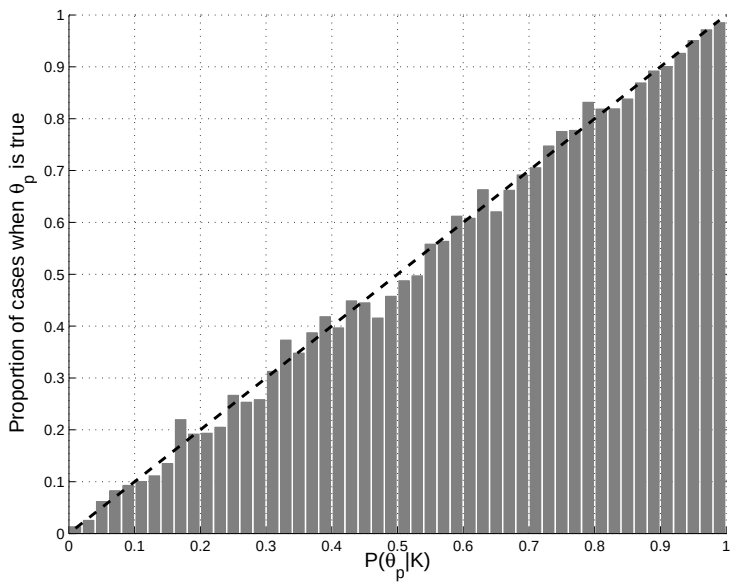
According to the definition of calibration, we can visualize whether the calibration of a set of probabilistic assessments is good or bad by means of so-called *empirical calibration plots*. Figure 2(b) shows one of this plots. It represents a histogram of the proportion of cases where it actually rained with respect to the total number of cases, as a function of the value of the probabilistic assessment. For the sake of illustration, in Figure 2(a) the histograms of the probabilistic assessments in the set respectively when $\theta = \theta_p$ and when $\theta = \theta_d$ are represented. Notice that the empirical calibration plot in Figure 2(b) is obtained by bin-by-bin dividing the number of cases in the histogram where $\theta = \theta_p$ over the total number of cases in both histograms in Figure 2(a).

Thus, in the empirical calibration plot, for all the days where the probabilistic assessment of the forecaster fell into a given range of values in the x-axis (*i.e.* a given *bin*), the y-axis represents the proportion of days where it actually rained. From the definition of calibration, for a well-calibrated set of probabilistic assessments, the value of the x- and y-axes should tend to be equal, and then the $x = y$ line is represented in the empirical calibration plot. Figure 2 shows an example of a well calibrated validation set of probabilities, whereas

[4] Notice that, due to this definition of calibration for continuous probabilistic assessments, calibration is not treated as something that is present or absent. Conversely, calibration is used as a continuous metric, and we will rather talk about probabilistic assessments that are *well calibrated* or *badly calibrated*.



(a)

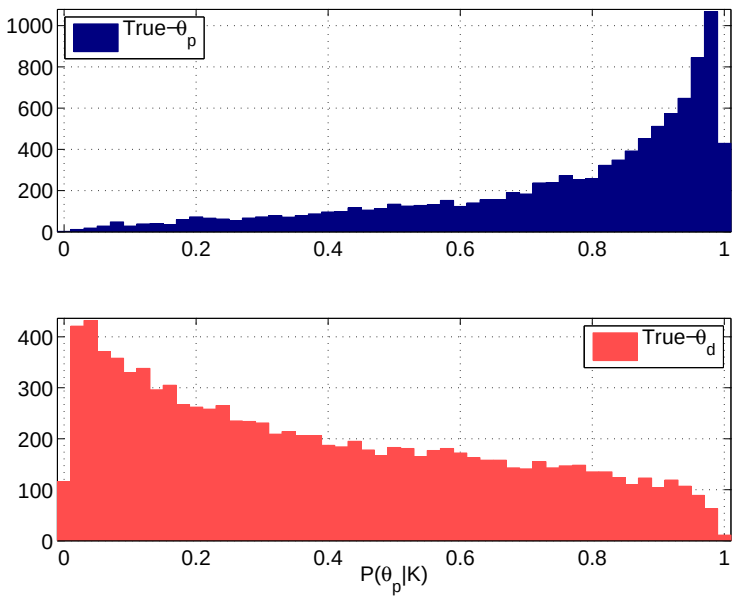


(b)

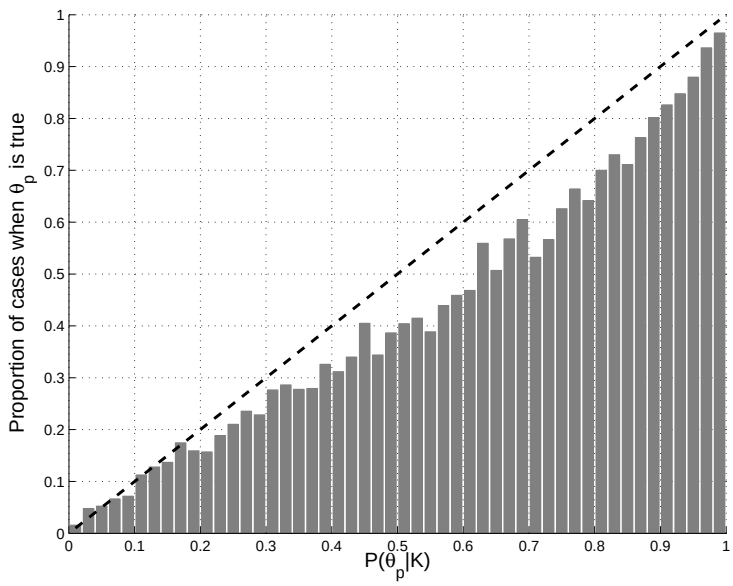
Fig. 2. Well-calibrated set of posterior probabilities. (a): Histograms of probabilities when $\theta = \theta_p$ (top) and when $\theta = \theta_d$ (bottom). (b): Empirical calibration plot.

a badly-calibrated set is represented in Figure 3.

Some of the good properties of calibration can be guessed intuitively now: if a forecaster elicits well-calibrated probabilistic assessments, they will be



(a)



(b)

Fig. 3. Badly-calibrated set of posterior probabilities. (a): Histograms of probabilities when $\theta = \theta_p$ (top) and when $\theta = \theta_d$ (bottom). (b): Empirical calibration plot.

constraining their opinions to the actual proportion of occurrence of events, a behavior that seems reasonable according to [16].

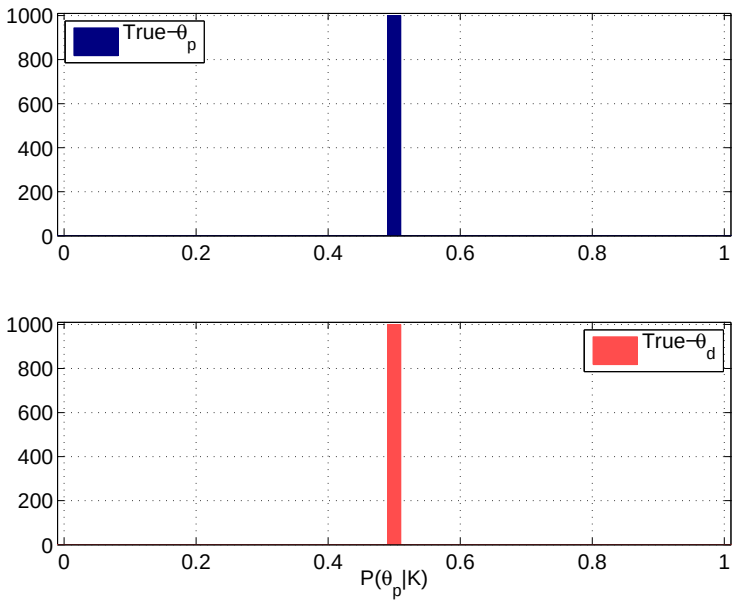
In the context of weather forecasting, we define the *discriminating power* of a set of posterior probabilities as the ability to distinguish between days where it will rain ($\theta = \theta_p$) and days where it will not rain ($\theta = \theta_p$). Discriminating power is also seen as the ability of the probability assessments to give *information* about the true value of θ. An example of a set of probabilities presenting good discriminating power is the one represented in the histograms in Figure 2(a), where it can be seen that the forecasts in days where it rains tend to be higher than the forecasts in days where it does not rain. Thus, a single forecast in that set of probabilities will give information about whether tomorrow it will rain or not, because if it is high, it will tend to indicate that in the following day it will rain, and vice-versa.

Although calibration has been described as a desirable property, a set of posterior probabilities needs not only to be well calibrated, but also to be *discriminating*. We exemplify this as follows:

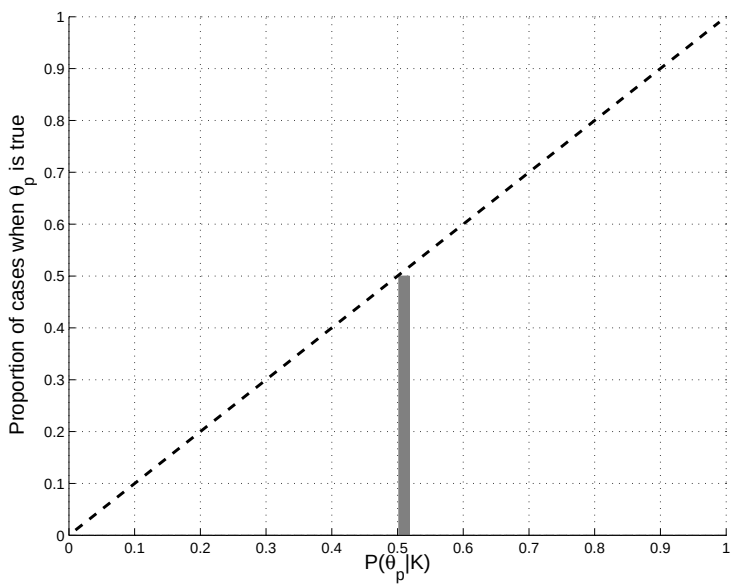
- It is possible that a well-calibrated set of posterior probabilities would give little or no information about the value of θ. For instance, a weather forecaster assigning well-calibrated forecasts can be useless in order to determine whether I should take my umbrella or not before going out. An example is a weather forecaster that is eliciting probabilities about raining in a region where the average probability of rain is 50%, and they always assign $P(\theta_p | K) = 0.5$. In the long-term, they will be well calibrated, because for the forecasts of value 0.5, the proportion of cases where it rains is 0.5, the average proportion of rainy days. However, such a forecaster does not give

any information whatsoever about whether it will rain or not in a given day, and therefore I will obtain no advice in any day about whether I have to take my umbrella with me or not. In other words, the forecaster does not give any information about the value of θ, the variable of interest, and they are useless in this sense. This type of forecaster is said to present no discriminating power in order to distinguish between days where it rains or not, but on the other hand they assign extremely well-calibrated probabilities in the long-term. Such a forecaster is represented in Figure 4.

- Even a forecaster that perfectly separates days in which it rains or not can be badly calibrated. We say that such a forecaster presents perfect discriminating power. For instance, imagine a weather forecaster that assigned probabilistic assessments during a period of time in a way that all the days where it actually rained the probability of rain was 0.6, and all the days where it actually did not rain the probability of rain was 0.4. The forecaster had the ability of perfectly separating the days where $\theta = \theta_p$ (with probabity of 0.6 for all of them) and the days where $\theta = \theta_d$ (with probabity of 0.4 for all of them), and therefore the forecaster presented perfect discriminating power in that period of time. However, those forecasts seem quite imperfect, because if the forecaster has such an amazing ability, a stronger opinion (*i.e.*, respectively closer to probability of 1 or 0) would have been more convincing in order to decide if I have to take my umbrella or not that day. Thus, even with perfect *discriminating power*, the forecaster is not assigning the best possible probabilities for my decision of taking my umbrella. Such a set of posterior probabilities is represented in Figure 5, where the empirical calibration plot shows bad calibration.



(a)

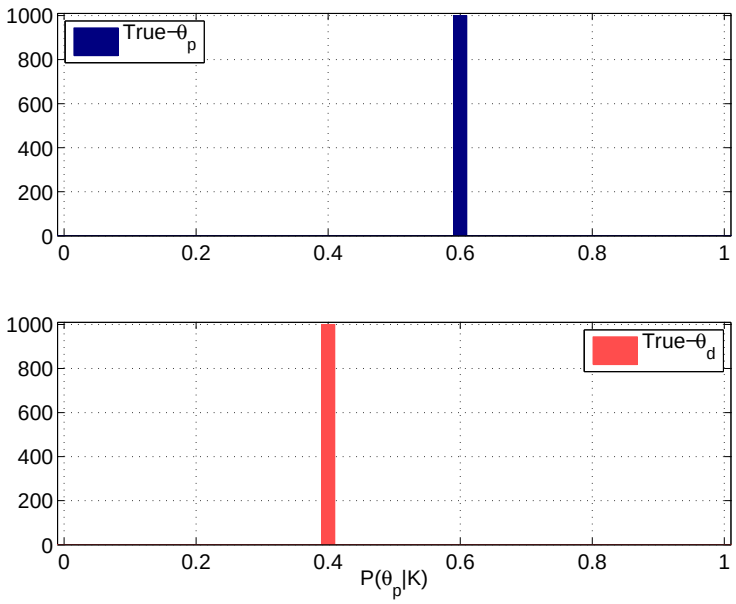


(b)

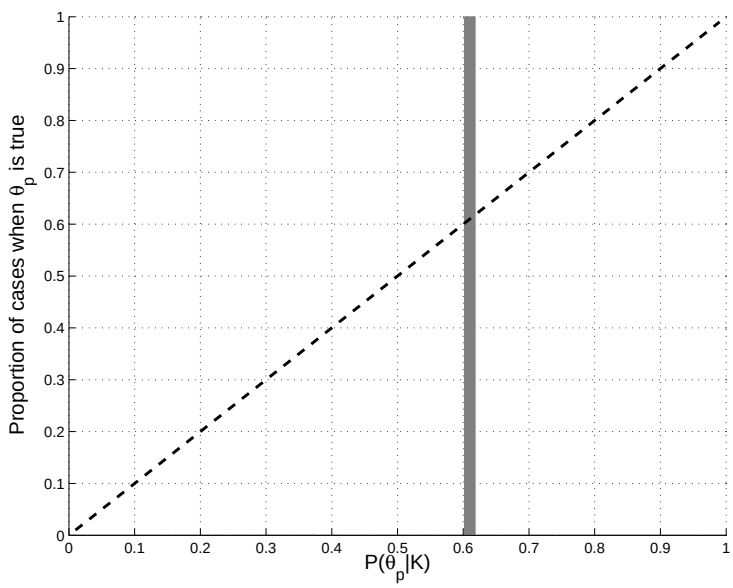
Fig. 4. A set of posterior probabilities showing null discriminating power but perfect calibration. (a): Histograms of probabilities when $\theta = \theta_p$ (top) and when $\theta = \theta_d$ (bottom). (b): Empirical calibration plot.

2.6 Calibration and Strictly Proper Scoring Rules

From our previous descriptions, now we can measure the accuracy of a set of probabilistic assessments by means of average values of SPSR (see Equation



(a)



(b)

Fig. 5. A set of posterior probabilities showing perfect discriminating power but bad calibration. (a): Histograms of probabilities when $\theta = \theta_p$ (top) and when $\theta = \theta_d$ (bottom). (b): Empirical calibration plot.

2). Additionally, we have introduced two desirable properties of the set of probabilities: calibration and discriminating power. The question is *are these two properties related to our definition of accuracy?*

The answer to the former question is *yes*, and was given by [20,21] in relation to the work in [11]. In those works, it is shown that:

$$\mathcal{L}_{\mathcal{C}} = \mathcal{L}_{\mathcal{C}}^{\text{disc}} + \mathcal{L}_{\mathcal{C}}^{\text{cal}} \tag{3}$$

where $\mathcal{L}_{\mathcal{C}}$ is the average cost or penalty due to a lack of accuracy of the set of posterior probabilities (Equation 2), $\mathcal{L}_{\mathcal{C}}^{\text{disc}}$ is the fraction of this average cost due to a lack of discrimination, and $\mathcal{L}_{\mathcal{C}}^{\text{cal}}$ is the fraction of the average cost due to a lack of calibration. Thus, by means of this decomposition, the accuracy of a set of forecasts will be good if and only if the discriminating power *and* the calibration of the probabilities are also both of them good.

Exemplifying in our weather forecasting example, we understand that a forecaster is accurate if their predictions allow us to efficiently decide whether to take an umbrella or not a given day. As we highlighted in Section 2.5, if the discriminating power of the probabilities elicited by the forecaster is good and their calibration is bad, then the accuracy can be very bad, because the forecaster can be under-confident or even biased in their probabilistic assessments, and our decision about the umbrella can be incorrect because of those reasons. On the other hand, if the calibration is good but the discriminating power is bad, we will probably have a forecaster that assigns so weak probabilistic assessments that give us little information about whether tomorrow it will rain or not, and therefore accuracy will be also poor because we will not find them useful in our decision about taking an umbrella or not.

The decomposition in Equation 3 allows to *explicitly* measuring calibration as the average cost given by $\mathcal{L}_{\mathcal{C}}^{\text{cal}}$: the lower the value of $\mathcal{L}_{\mathcal{C}}^{\text{cal}}$, the better the

calibration, and vice-versa. However, arriving to such a decomposition is not trivial. We have adopted the solution proposed in [20,21] by means of the Pool Adjacent Violators algorith (PAV). As described in [20,21,23], PAV receives as an input a *validation* set of posterior probabilities [5], and a new probability is assigned to each of the posterior probabilities of the input set, yielding a set of *optimally-calibrated* posterior probabilities. Thus, if the accuracy of the set of posterior probabilities *before* the application of PAV is $\mathcal{L}_{\mathcal{C}}$, then the accuracy of the set *after* the application of PAV will be $\mathcal{L}_{\mathcal{C}}^{\text{disc}}$, because the calibration loss of the probabilities transformed by PAV $\mathcal{L}_{\mathcal{C}}^{\text{cal}}$ will be reduced to zero. Therefore, if we have a validation set of posterior probabilities, we can measure the following magnitudes:

- Accuracy, the global measure of performance of the set, as the average value of the SPSR, namely $\mathcal{L}_{\mathcal{C}}$ (Equation 2).
- Discriminating power, as the accuracy of the set after being transformed by the PAV algorithm, namely $\mathcal{L}_{\mathcal{C}}^{\text{disc}}$.
- Calibration, as $\mathcal{L}_{\mathcal{C}}^{\text{cal}} = \mathcal{L}_{\mathcal{C}} - \mathcal{L}_{\mathcal{C}}^{\text{disc}}$.

It is important to highlight here that the use of PAV in the proposed methodology aims at establishing just a reference of performance. In this article, we are not proposing PAV as a method for computing LR values. Thus, a set of LR values can be computed using common LR computation methods (*e.g.* such as the ones described in [13]), and then the LR values obtained can be compared to the reference LR values obtained by the application of PAV. See also [24] for details.

[5] Notice that the ground truth labels are also in the validation set.

3 Probabilistic Assessments in Forensic Science

In many aspects, the Bayesian probabilistic framework in forensic science is analogous to the one described above for weather forecasting. However, although it is possible to apply the assessment methodology described above to forensic science, the steps are not straightforward. We need to take into account the particular inferential context of forensic interpretation and the competences and roles involved, in order to arrive to a satisfactory solution, that we develop in this section.

3.1 The Likelihood Ratio Approach

In order to understand the problems arising from the use of the SPSR methodology in forensic science, it is needed that we get deeper into the Bayesian inferential mechanism in forensic science, namely the LR framework for forensic evidence evaluation [13].

Consider the forensic evidence E, which includes two types of *evidential materials*: recovered materials of unknown origin and control materials whose origin is known. Such so-called *alternative* propositions can be stated at several levels [25,26]. An example of a pair of propositions at source level is:

- θ_p (also known as the *prosecution* proposition): the recovered and control materials come from the same source.
- θ_d (also known as the *defence* proposition, or the *alternative* proposition): the recovered materials come from a population of potential sources, which does not include the source of the control materials.

In a forensic case, the unobserved variable of interest is the proposition that is actually true in the case, namely θ, which may take the values θ_p or θ_d. Bayes' theorem relates probabilities of the values of θ before and after the analysis of the evidence:

$$\frac{P\left(\theta_p \mid E, I\right)}{P\left(\theta_d \mid E, I\right)} = LR \times \frac{P\left(\theta_p \mid I\right)}{P\left(\theta_d \mid I\right)} \tag{4}$$

where I is the background information available in the case apart from the evidence E. The propositions must be mutually exclusive and exhaustive[6]. Equation 4 is the so-called *odds form* of Bayes' theorem, because the *odds* O are defined as $O\left(\theta_p \mid E, I\right) = \frac{P(\theta_p \mid E, I)}{P(\theta_d \mid E, I)}$. The likelihood ratio (LR) is therefore defined as:

$$LR = \frac{P\left(E \mid \theta_p, I\right)}{P\left(E \mid \theta_d, I\right)} \tag{5}$$

and expresses the degree of support given by the evidence to any of the propositions in the case.

From Equations 4 and 5, the relationship between the posterior probability that $\theta = \theta_p$ and the LR value can be stated as:

[6] Although in this work we assume that the propositions fulfill both requirements, some authors do not consider exhaustion as a strict requirement. For instance, in [27] when the propositions are not exhaustive but the probability of any of them to be true is close to 1, they define *pseudo-odds* as a way of performing the inferential process

$$P\left(\theta_p \middle| E, I\right) = \frac{LR \times O\left(\theta_p\right)}{1 + LR \times O\left(\theta_p\right)} \tag{6}$$

It is important to highlight here that in casework the forensic examiner should assess the value of the LR, but not the value of the prior probabilities. This is the one of the main advantages of the use of likelihood ratios: in a given case, the fact finder evaluates all the possible information but the evidence, whose strength is contributed by the forensic examiner by the reporting of a LR. In this way, the forensic examiner should not need to assess the prior probabilities in casework.

3.2 Weather Forecasting and Forensic Science

In order to apply in forensic science the performance assessment methodology based on SPSR, we establish some analogies between the previously described weather forecasting example and the forensic inference framework:

- The binary variable of interest, $\theta^{(i)} \equiv \theta$. In weather forecasting, its value indicates if it rained or not in day i. In forensic inference, its value indicates the proposition (prosecution or defence) that is true in a given case indexed as i.
- The possible values of θ. In weather forecasting, θ_p means that it rained in day i and θ_d means that it did not rain in day i. In forensic inference, θ_p means that the prosecution proposition is true in case i and θ_d means that the defence proposition is true in case i.
- The available knowledge. In weather forecasting, all the knowledge available to the forecaster is referred to as K. In forensic inference, all the available

knowledge is split into E and I. If we understand these as sets, we can define $K \equiv E \cup I$ for forensic inference, and K would have an analogous meaning than in weather forecasting: all the available *observed* knowledge in the problem.

- The assessment of probabilities. Given the analogies above, we can state the problem of forensic inference in the same terms as in weather forecasting. On the one hand, the aim of the weather forecaster is assigning a posterior probability distribution $P(\theta_p \mid K)$ that represents their opinion about the value of the unknown variable θ, from the information obtained from K. On the other hand, in forensic inference the aim of the fact finder is obtaining a posterior probability distribution $P(\theta_p \mid E, I)$ that represents their opinion about the value of θ from the information obtained from the evidence E and the rest of iformation in the case I.

Despite all the analogies before, using the methodology of performance assessment based on SPSR in forensic science is not straightforward, because there is a substantial difference with respect to the weather forecasting example. And that difference comes from *what* we want to measure in forensic evaluation of the evidence.

In weather forecasting, we wanted to measure the performance of the probabilistic assessments of the forecaster, namely $P(\theta_p \mid K)$. As the responsible of such assessments is the forecaster, then we can measure performance directly from a validation set of posterior probabilities assigned by the forecaster (and their corresponding ground-truth labels).

However, in the case of forensic inference, this is not the case, for several reasons:

- In forensic inference we only want to measure the performance of the probabilistic assessments of the forensic examiner. Therefore, we want to focus on measuring the performance of the LR, not the performance of the posterior probability $P(\theta_p \mid E, I)$. This is because the posterior probability depends both on the LR *and* on the prior probability $P(\theta_p \mid I)$, the latter not being the province of the forensic examiner. However, the SPSR framework is based on measuring the performance of the posterior probabilities, and therefore if we apply it directly to forensic science, we would be partially measuring performance of the assessments of the fact finder, not just the LR given by the forensic examiner.
- It is currently unrealistic to imagine a scenario where fact finders assign prior probabilities in a Bayesian context. Thus, it may not be possible in general to arrive to posterior probabilities in a case, and therefore the SPSR methodology could not be applied to measure performance of posterior probabilities in a realistic scenario.
- The measurement of the performance of the methods is typically conducted in the forensic laboratories, by means of simulated experiments where the evidence evaluation methods are tested empirically. This is typically done before deciding whether a method is ready to be used in casework or not. As the forensic examiner cannot assign prior probabilities, it is not straightforward to arrive to posterior probabilities from the values of LR that may be computed in those simulated experiments.

4 Measuring Accuracy in Forensic Science

Despite all the reasons given in the previous section, we give a solution that allows the use of the SPSR assessment methodology to LR values computed by forensic examiners.

4.1 Avoiding the Prior Odds in Performance Measurement

Imagine that we work in a forensic laboratory where we want to assess the performance of a given LR computation method. In order to apply the SPSR methodology, our proposed solution firstly considers the set-up of a simulated experiment, where a *validation set of LR values* will be computed from a so-called *validation database*. Each LR value from such a validation set, say LR_i, would be generated by the simulation of a case, say case i, where the evidential materials E_i (control and recovered) will be taken from the validation database. The experimental protocol considers that several values of LR are computed following this procedure, each one by means of a different simulated case. As the ground-truth labels are known in the validation database, we can follow a comparison protocol that yields a number N_p of LR values from simulated cases where we know that θ_p is actually true, and a number N_d of LR values from simulated cases where we know that θ_d is true. Summarizing, the validation set of LR values will consist of the set of $N = N_p + N_d$ LR values with their corresponding ground-truth labels.

It is important to remark that if any of these LR values would have been computed in a real case, the forensic scientist should report the LR value to court, but not accompanied by any assessment of prior probabilities [25], and

not usurping the role of the court and reporting a posterior probability (a phenomenon that is called the *prosecutor's fallacy* [28]).

However, measuring performance by SPSR is focused on posterior probabilities, but forensic examiners only have the values of the LR in the validation set. If a value of the prior would be fixed in the experiment, a validation set of posterior probabilities could be obtained from that prior and the LR values, and those posteriors could be used to assess accuracy with the SPSR methodology. In that case, we would be measuring performance for the particular case when the fact finder would set the prior to that given value. However, the forensic examiner cannot fix a value of the prior, even in their experiment, because that value is not their competence.

Instead of that, we propose to compute accuracy in our experiment for a *wide range* of prior probabilities, following the same procedure as in [20]. That way, in the experiment, we vary the prior probabilities, and we compute what would be the accuracy when we use the validation set of LR values for each of the the prior probabilities in that range. Thus, the forensic examiner will not fix a value for the prior probabilities in the experiment, but they will be able to know the performance (accuracy) if someone would be using their LR values in a correct Bayesian framework.

Notice the differences we establish between the use of a LR in casework and our proposed procedure to measure performance of a set of LR values in a controlled experiment, typically prior to casework. First, in casework the ground truth is unknown, but in the controlled experiment the ground-truth labels are available. Second, in casework it is totally unrealistic to think that the fact finder will be assessing a prior probability, at least in current practice. How-

ever, the duty of our methodology for measuring performance is to consider the LR in a formal way, as part of a Bayesian inference process. Therefore, in our experiment, we assume that the fact finder will assess a prior, but we will never know its particular value. This is the reason to consider a wide range of prior probabilities, and not a particular value of the prior.

4.2 Proposed Performance Representation: ECE plots

Following the procedure described in the previous section, we propose a way of representing the performance of a set of LR values, which measures accuracy as an explicit decomposition between discriminating power and calibration. It is out of the scope of this work to go into deep details, that can be found in [24,22].

First, we define our measure of accuracy, namely *Empirical Cross-Entropy* (ECE) as the average value of the logarithmic scoring rule, weighted in the following way:

$$\begin{aligned} ECE = & -\frac{P\left(\theta_p \middle| I\right)}{N_p} \sum_{\theta^{(i)}=\theta_p} \log_2 P\left(\theta_p \middle| E_i, I\right) \\ & -\frac{P\left(\theta_d \middle| I\right)}{N_d} \sum_{\theta^{(j)}=\theta_d} \log_2 P\left(\theta_d \middle| E_j, I\right). \end{aligned} \tag{7}$$

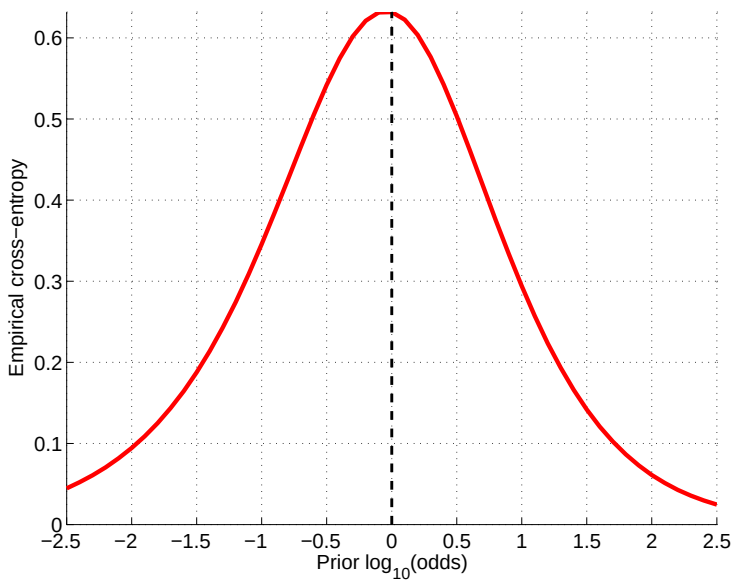
where E_i and E_j denote the evidence in each of the comparisons (cases) in the validation set where θ_p or θ_d are respectively true. It is illustrating to express ECE explicitly in terms of the prior odds and the LR using Equations 6 and 7:

$$ECE = \frac{P(\theta_p|\, I)}{N_p} \sum_{\theta^{(i)}=\theta_p} \log_2 \left(1 + \frac{1}{LR_i \times O(\theta_p)}\right) + \frac{P(\theta_d|\, I)}{N_d} \sum_{\theta^{(j)}=\theta_d} \log_2 (1 + LR_i \times O(\theta_p)). \quad (8)$$

As it can be seen in Equations 7 and 8, the averages in ECE are weighted by the value of the prior probabilities. This weighting allows ECE to be interpreted in an information-theoretical way, but this topic is out of the scope of this work (see [24,22] for details). However, it can be shown that the interpretation of ECE as accuracy and its properties related to calibration remain the same as for $\mathcal{L}_{\mathcal{C}}$ (see, e.g., [21]).

Equation 8 shows that ECE depends on the validation set of LR values in the experiment (*i.e.*, the LR values and their corresponding ground-truth labels). However, ECE also depends on the value of the prior odds $O(\theta_p|\, I)$, since a SPSR depends on the posterior probabilities. Thus, following the procedure described in Section 4.1 ECE can be represented as a function of the logarithm of the prior odds. An example of such a representation can be seen in Figure 6. We use base-10 logarithms for the prior odds because they are typically used for evidence evaluation. However, base-2 logarithms will be used for computation of ECE because of its information-theoretical interpretation (see [24,22] and [29] for details).

ECE in Figure 6 represents the accuracy for all the possible values of the prior probability, but calibration is not explicitly represented. Therefore, we give an explicit measurement of discriminating power and calibration in a so-called ECE plot ([24]), which shows three comparative performance curves together (Figure 7):



(a)

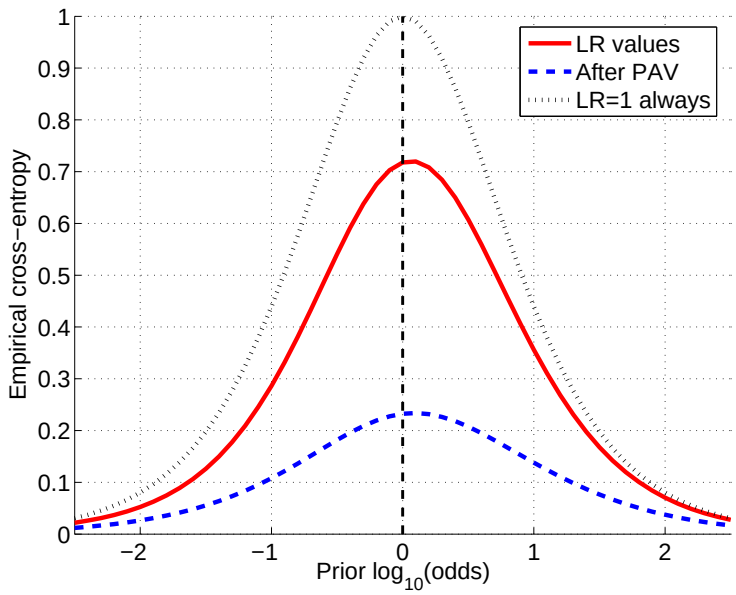
Fig. 6. ECE as a function of the logarithm of the prior odds (prior log-odds).

(1) Solid curve: accuracy. This curve is the ECE of the LR values in the validation set, as a function of the prior log-odds. The lower this curve, the better the accuracy. This is the same representation as shown in Figure 6.

(2) Dashed curve: accuracy after PAV. This curve is the ECE of the validation set of LR values after being transformed using the PAV algorithm, as a function of the prior log-odds. Therefore, this shows the performance of a validation set of optimally-calibrated LR values, according to Section 2.5 [7].

(3) Dotted curve: neutral reference. It represents the comparative performance of a so-called neutral evidence evaluation method, defined as the one which always delivers LR=1 for each case. This neutral method is taken as a performance bound: the accuracy should be always better than

[7] Recall that PAV is used in the proposed methodology as a reference for measuring performance, and not as a method of obtaining LR values in casework.



(a)

Fig. 7. Example of an ECE plot.

the neutral reference.

Thus, in ECE plots we can observe the following:

- Accuracy: solid curve. The lower the curve, the better the accuracy.
- Discriminating power: dashed curve. The lower the curve, the better the discriminating power.
- Calibration: difference between the solid and dashed curves. The closer the blue and the red curves, the better the calibration.

With this representation, the calibration of a validation set of LR values can be explicitly measured.

In order to facilitate the use of these tools, there is a freely available software for the computation of ECE plots for Octave and Matlab^{TM}, that can be downloaded from http://atvs.ii.uam.es/software/ECE_plots_SW.zip. Moreover, the R package *comparison* by David Lucy also includes functions for drawing ECE plots, and can be downloaded from the CRAN repository (http://cran.r-project.org).

An example of the use of the software in Matlab^{TM} is shown as follows, where it can be seen that with the LR values separated in the cases where θ_p and θ_d are respectively true, it is easy to use the software to draw ECE plots:

```
>> figureS1=ECE_plot_10({'Example ECE plot',{logLRss1,logLRds1}});
>> % Variables (two vectors) containing the LR values:
>> % logLRss1: LR where prosecutor proposition is true
>> % logLRds1: LR where defence proposition is true
```

5 Why *reliable*? A Desirable Property of Well-calibrated Likelihood Ratios

Until now, we have been focusing on describing and explaining the concept of calibration, an we have provided tools to measure accuracy and calibration. In this section, we justify why a set of LR values *should* be well calibrated, by means of an important property of calibration: for a well-calibrated set of LR values, the higher their discriminating power, the stronger the support they

will tend to yield, and *vice-versa*. In other words, if we have a method that yields well-calibrated LR values which also present high discriminating power, the weight of the evidence given by that method, expressed as the value of $|\log(LR)|$, will tend to be high, and *vice-versa* [8].

We illustrate this property with an example. Figure 8 shows the Tippett plots of two validation sets of LR values, namely set S_1 and set S_2. In the y-axis Tippett plots represent the proportion of cases where the $\log_{10}(LR)$ exceeds the value in the x-axis in the experimental set. Moreover, two curves are represented: one for the LR values in the set where $\theta = \theta_p$, and one for the LR values in the set where $\theta = \theta_d$ (see *e.g.* [31] for details about Tippett plots). These two validation sets of LR values have been generated synthetically for the sake of illustration.

From the Tippett plots, one can figure out what is the range of LR values in each of the sets S_1 and S_2, because the curves represent an empirical cumulative distribution of the LR values in the set. Thus, as a rule of thumb, the further are the curves from the value of $\log_{10}(LR) = 0$, the higher the value of $|\log_{10}(LR)|$. The latter is a measure of the strength (or *weight*) of the evidence, because the higher the value of $|\log_{10}(LR)|$ the stronger the support of the evidence to a given proposition in the case. Therefore, Tippett plots in Figure 8 show that the strength of the LR values in S_1 tends to be much higher than for the LR values in S_2. Also, in Tippett plots the discriminating power is measured as the vertical separation of both curves in the graph, measured at a given value of the $x - axis$. Therefore, if that separation is measured *e.g.*

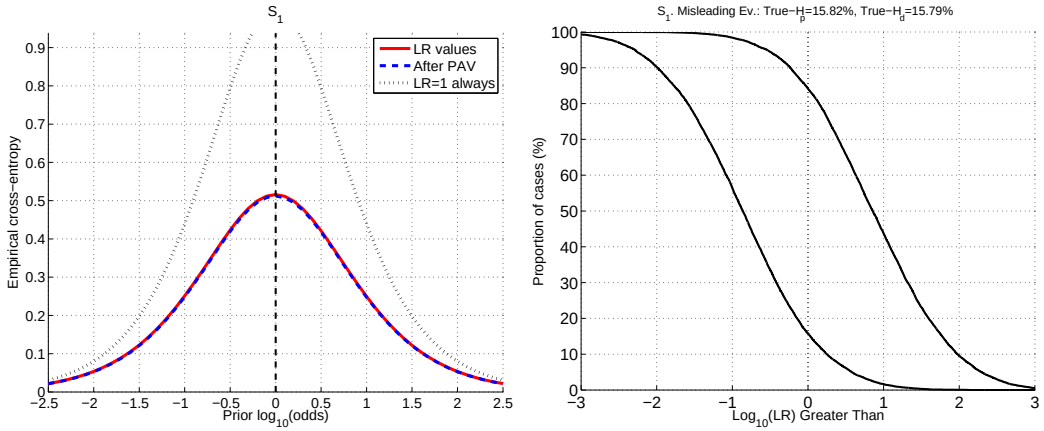
[8] Although well-calibrated probabilistic assessments present other desirable properties, that are not described here for limits in the extension of this work. See [14,30] for more details.

at $x = 0$, it can be also be seen from Tippett plots that the discriminating power of S_1 is higher than the discriminating power of S_2.

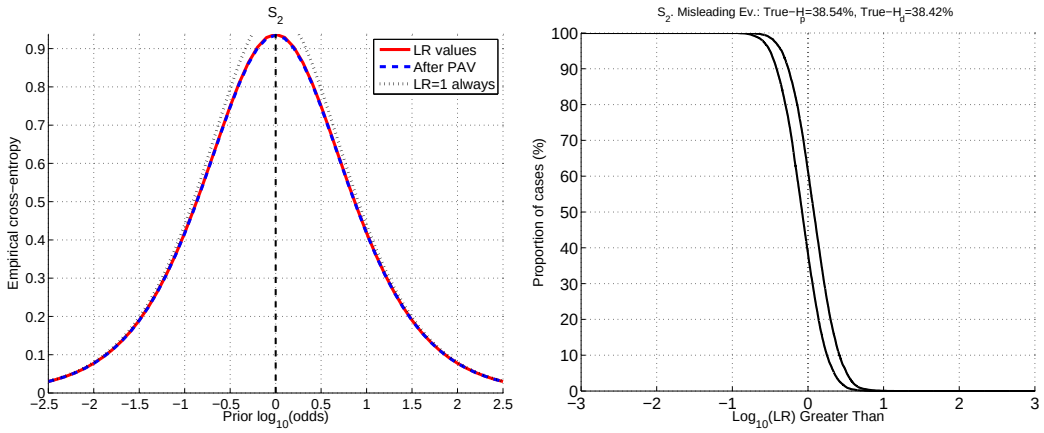
Also, ECE plots are represented in Figure 8. ECE plots reveal that the calibration of the LR values in S_1 and S_2 is good in both cases, because the separation between the solid and dashed curves is small for both sets. This could not be seen in Tippett plots, where calibration is not explicitly measured. Moreover, the discriminating power is better in S_1 than in S_2, because the dashed curve is lower for S_1. Therefore, the example illustrates that if S_1 and S_2 are well calibrated, the method with better discriminating power will yield stronger support to the propositions (higher weight of the evidence, *i.e.* $|\log_{10}(LR)|$).

This relationship between the discriminating power and the weight of the evidence does not necessarily happen for badly-calibrated sets of LR values. In figure 9, the performance of two badly-calibrated LR sets is shown, namely S_1^u and S_2^u. ECE shows the miscalibration, because there is a considerable distance between the solid and dashed curves in both cases. Also, the discriminating power is better for S_1^u than for S_2^u. However, the values of $|\log_{10}(LR)|$ are in most cases much higher for S_2^u than for S_1^u, and therefore the relationship between discrimination and the magnitude of $|\log_{10}(LR)|$ does not hold.

This property has strong implications, mostly because it agrees with common sense. In daily life, it seems reasonable that people expresses strong opinions if and only if they have a considerable amount of information about what they are talking about. As a consequence, if someone has little information about some variable of interest, they should express a weak opinion about that variable. For instance, for probabilistic assessments in weather forecasting, if



(a)

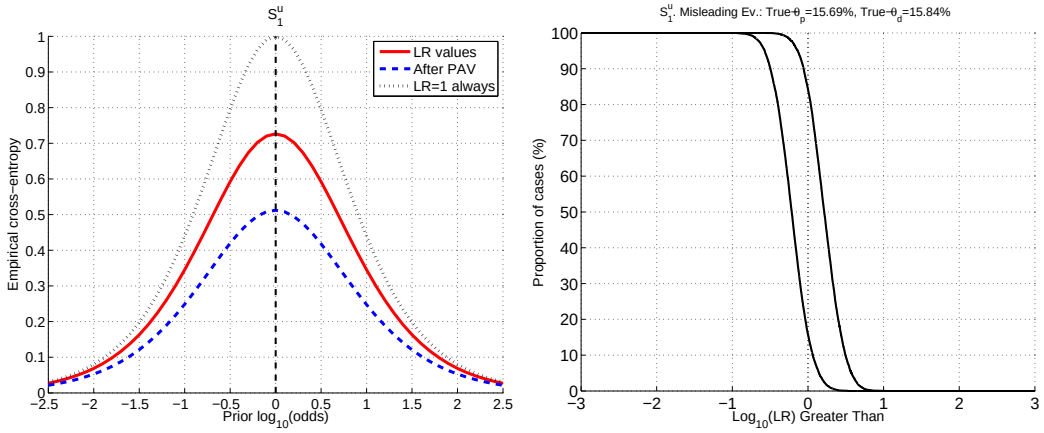


(b)

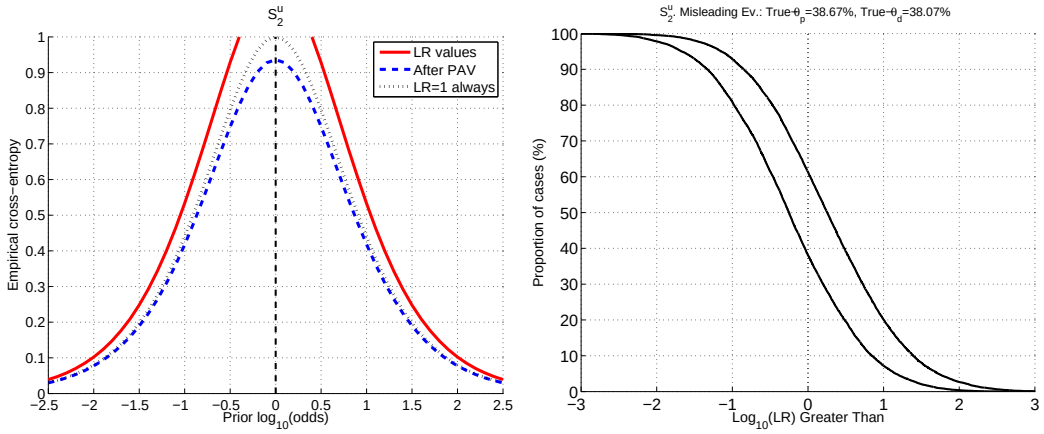
Fig. 8. Example with two well-calibrated sets of LR values. ECE plots on the left column and Tippett plots on the right column. (a): Set S_1. (b): Set S_2

a weather forecaster is able to gather lots of information about whether it will rain or not, then their probabilistic assessments should be strong, close to 1 or 0. On the other hand, if a forecaster is not proficient, then the best they can do is expressing a weak opinion, typically very close to the annual average probability of rain.

In forensic science, imagine for instance a comparison between DNA and speech evidence. Assume a forensic scenario where the quality of the evidence and the operational conditions are good for both disciplines. It seems



(a)



(b)

Fig. 9. Example with two badly-calibrated sets of LR values. ECE plots on the left column and Tippett plots on the right column. (a): Set S_1^u. (b): Set S_2^u

reasonable that the current methods for computing LR values for DNA evidence yield highly discriminating LR values in those conditions. Moreover, we could also affirm that the discriminating power of LR values from DNA models is much higher than the discriminating power obtained with a forensic automatic speaker recognition system in those conditions [32]. Then, common sense suggests that the strength of the evidence in DNA should be in general much higher than the strength of LR values from forensic automatic speaker recognition systems. If the DNA models and the speaker recognition models

yield well-calibrated LR values, then this property will hold. In this sense, methods yielding well-calibrated LR values will help to prevent the calculation of very strong LR values in fields where the discriminating power can be shown to be limited.

Mainly because of the property described above, calibration has been dubbed *reliability* in the context of Bayesian probabilistic assessment [14,11]. According to the dictionary, *reliability* is a property of something that can be trusted [33]. In fact, it is also common sense that, because of this property, a probabilistic assessor eliciting well-calibrated probabilities can be trusted. On the one hand, if the probabilistic assessor is able to gather lots of information about the variable of interest, their opinions will be strong. On the other hand, if the assessor has little information about the value of interest, their opinions will be weak.

The same idea of reliability can be applied to well-calibrated LR values in forensic science. Therefore, if the LR values are discriminating, then they will tend to be strong and therefore they will have lots to do in the final decisions by the fact finder. However, if their discriminating power is low, the LR will tend to be weak, and therefore the prior probabilities of the fact finder will not suffer too much change. Thus, in general one can trust well-calibrated LR values in the inferential process.

6 Experimental Examples

In this section we show two experimental examples in the context of speaker recognition. The aim is to highlight the importance of calibration in the per-

formance measurement of LR computation methods, and to exemplify the property described in Section 5.

At is is described below, the models for LR computation used in this experimental example are based on assignments of probability distributions separately to the numerator and the denominator of the LR, following other approaches in the literature such as *e.g.* [34–39]. Then, the performance of the LR computed is measured by means of the proposed methodology. Therefore, it is again highlighted that our proposal is focused on performance measurement, not on the methods for computing LR values.

6.1 Human-listener speaker recognition at NIST HASR 2010

The example is presented in the context of a bi-annual speaker recognition evaluation conducted by the American National Institute of Standards and Technology (NIST). These evaluations constitute one of the most important scientific fora in order to foster the improvement and development of automatic speaker recognition technology. Participation is free, and the number of participants increase year by year.

The first experiment presented in this section has been obtained from our participation in NIST Human-Assisted Speaker Recognition evaluation 2010 (NIST HASR 2010). The main idea is to blindly compare the performance of different speaker recognition systems from different participants. This is done by consecutively conducting a given amount of comparisons of two speech segments according to a given protocol, without knowledge of the speaker that generated each speech segment. The aim is to obtain information about

whether both segments belong to the same person or not. The conditions of the speech are rather uncontrolled, presenting strong variability in the acoustic environment, the transmission channel, the acquisition device, the emotional state of the speaker, the dialectal variation of English, etc. Thus, the comparison protocol resembles some kind of forensic scenario, with poor quality of the recordings, and strong variability of the conditions. Once the results of the comparisons are submitted to NIST for evaluation, the organizers make the ground-truth labels public, and benchmark results are disseminated among participants to reveal which systems have performed better.

The NIST HASR 2010 dataset is a subset of the NIST Speaker Recognition Evaluation 2010 (NIST SRE 2010), the latter containing hundreds of speakers and a number of comparisons in the range of the hundreds of thousands. However, as opposed to the NIST SRE 2010, where only automatic means are allowed for comparison, the NIST HASR 2010 allowed human intervention in the comparison process. Thus, in order to make the task feasible, the NIST HASR dataset considered a small subset of 150 comparisons.

More details about the data and protocols in NIST HASR 2010 can be found in [40].

6.1.1 Methods to compare

Using the comparison protocol of NIST HASR 2010, we will generate two different validation sets of LR values for two different LR computation methods, and we will compare them in order to illustrate ECE plots and calibration.

For both models presented, it is assumed that the evidential materials are com-

pared in order to obtain a score s_E. Then, the probability density functions of that score under the competing propositions θ_p and θ_d will be obtained from scores under both assumptions, obtained using suitable databases. Thus, we will denote $\mathbf{s_p} = \left\{s_p^{(1)}, \cdots, s_p^{(M_p)}\right\}$ the set of M_p scores used to obtain the probability density function in the numerator of the LR, which will be computed under the assumption that θ_p is true, *i.e.*, comparing speech segments coming from the same speaker. Analogously, we will denote $\mathbf{s_d} = \left\{s_d^{(1)}, \cdots, s_d^{(M_d)}\right\}$ the set of M_d scores used to obtain the probability density function in the denominator of the LR, which will be computed under the assumption that θ_d is true, *i.e.*, comparing speech segments coming from different speakers.

The first method to compare is a fully-automatic speaker recognition system, which outputs a score from the comparison of two speech segments, from which a LR will be obtained. It is out of the scope of this article to deeply describe the system, but its details can be found in [41]. From the scores of the automatic speaker recognition system, in this article a model for LR computation is used based on assigning probabilistic distributions to the numerator and the denominator of the LR separately, according to a Kernel Density Function (KDF). Formally, the model is as follows:

$$\frac{P\left(s_E|\,\theta_p\right)}{P\left(s_E|\,\theta_d\right)} = \frac{f\left(s_E|\,\mathbf{s}_p\right)}{f\left(s_E|\,\mathbf{s}_d\right)} \tag{9}$$

where $f\left(s_E|\,\mathbf{s}_p\right)$ and $f\left(s_E|\,\mathbf{s}_d\right)$ are probability density functions obtained with Kernel Density Functions using Gaussian kernels with optimal kernel widths. Details about KDF and the selection of the kernel width can be found in [42].

The second method to compare is based on a panel of 11 lay listeners not

trained in forensic speaker recognition, and not English native speakers either. Each of the listeners was assigned part of the 150 comparisons of the HASR evaluation. For each comparison, each participant was plenty of time to hear the two speech recordings to compare, and several tools were allowed in order to assist them to do basic speech representation and analysis. After this process, a score was given by the listener, which would be higher than 0 as more support was given to the prosecutor proposition (θ_p, meaning *same-source* in this case) and lower than 0 as more support was given to the defence proposition (θ_d, meaning *different-sources* in this case). The details of this process can be found in [40]. Then, LR values were computed according to a Gaussian model trained with scores from comparisons using data from past NIST evaluations (namely $\mathbf{s_p}$ and $\mathbf{s_d}$). Thus, for each comparison yielding a score s_E by the listener, a LR value was computed from s_E in the following way:

$$\frac{P\left(s_E|\,\theta_p\right)}{P\left(s_E|\,\theta_d\right)} = \frac{f\left(s_E|\,\hat{\mu}_p, \hat{\sigma}_p\right)}{f\left(s_E|\,\hat{\mu}_d, \hat{\sigma}_d\right)} \tag{10}$$

where $f\left(s_E|\,\hat{\mu}_p, \hat{\sigma}_p\right)$ and $f\left(s_E|\,\hat{\mu}_d, \hat{\sigma}_d\right)$ are Gaussian probability density functions. The parameters for the numerator are $\hat{\mu}_p, \hat{\sigma}_p$, the sample mean and variance of $\mathbf{s}_p$. The parameters for the denominator are $\hat{\mu}_d, \hat{\sigma}_d$, the sample mean and variance of $\mathbf{s}_d$

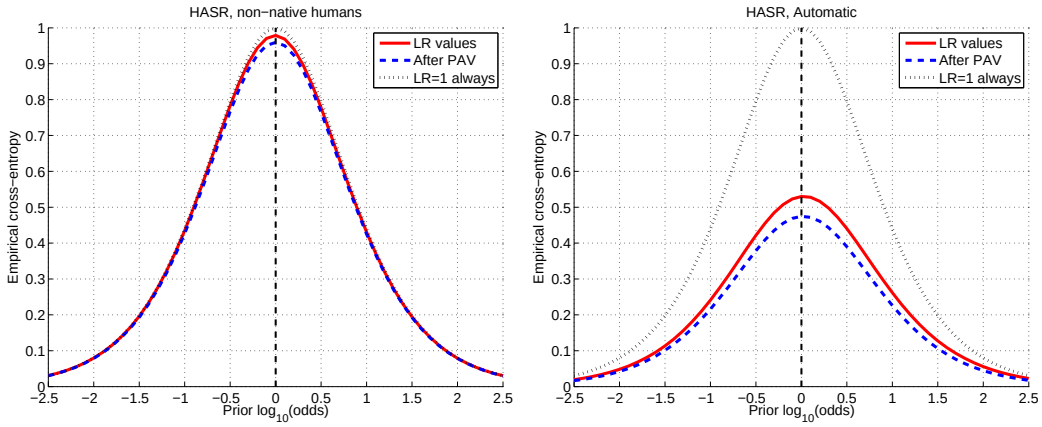
6.1.2 Results: the importance of calibration

Figure 10 shows ECE plots for the validation set of LR values for each method. From the figure, the following can be observed:

- It is seen that the accuracy is good for the automatic method, but not for the human listener method. The former presents an accuracy (ECE, solid curve) much better than the neutral reference (dotted curve). However, the accuracy of the latter is really close to the neutral reference.
- The discriminating power of the automatic method is good, because the dashed curve is much lower than the neutral reference. However, for the human listeners the dashed curve is pretty close to the neutral reference, which means that the human listeners have similar discriminating power than a set of LR values that are all equal to 1. In other words, the LR values computed from the human listener scores have almost no discriminating power.
- For both methods the calibration is good, because the solid and dashed lines are really close. According to the property described in Section 5, this will mean that for the automatic method, the LR values will tend to be be moderately strong, because they present moderate disrcriminating power. However, for the human listeners, the LR values will tend to be weak, because their discriminating power is poor.

Tippett plots in Figure 11 confirm the conclusions above concerning calibration. It is observed that for the automatic method, the ranges of the LR values are moderate, mostly in the order of magnitude of 100 or even 1000 when the prosecution proposition is true (and inversely between 0.001 and 0.01 when the defence proposition is true). However, for the human listener method, the values of the LR are very weak, with $|\log_{10}(LR)| < 1$ in all cases.

From the results we can say that both methods present the desired behavior. On the one hand, the discriminating power of the automatic method can assist the court in their inferential process, but only moderately, because it is not so



(a) (b)

Fig. 10. ECE plots for the human listeners (a) and automatic (b) methods in NIST HASR 2010 evaluation.

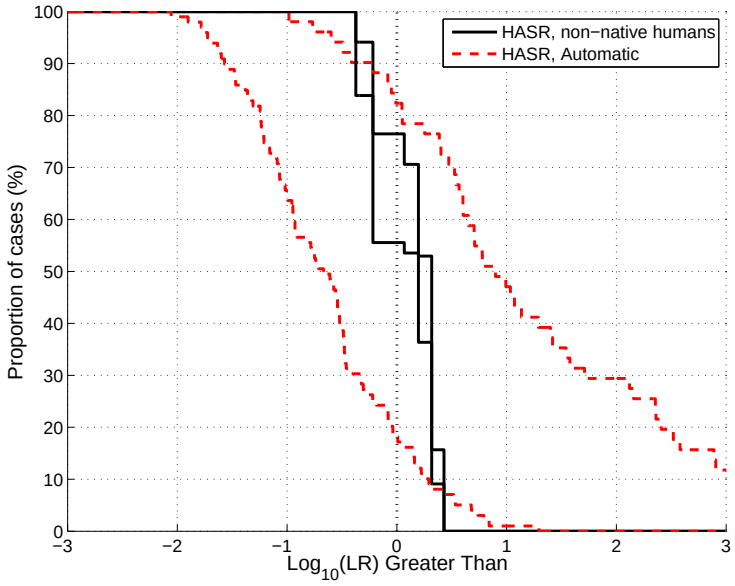


Fig. 11. Tippett plots for the human listeners (solid) and automatic (dashed) methods in NIST HASR 2010 evaluation.

discriminating. However, as the discriminating power of the human listeners is poor, the best they can do is almost not affecting the prior opinion of the fact finder, which means a weak LR value. This desirable effect is achieved if the LR values from the evidence evaluation methods are well calibrated.

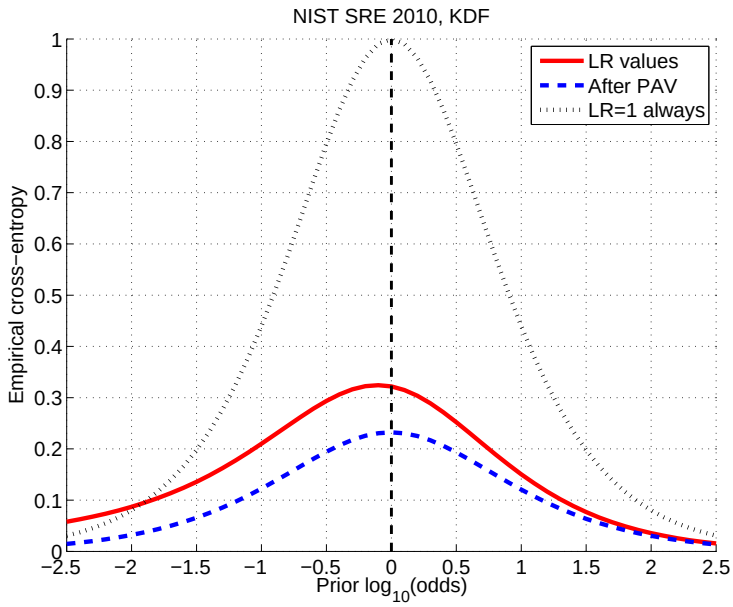
In this example, we present results using the ATVS-UAM automatic speaker recognition system used in Section 6.1.1 and deeply described in [41]. This section presents the outcome of the participation of the ATVS group of the Universidad Autonoma de Madrid in NIST SRE 2010.

The NIST SRE 2010 database and protocol is described as follows [9] . The task is comparing a big amount of pairs of speaker utterances from a so-called *testing database*, provided by NIST for the evaluation (and previously unreleased). All available speech databases from past NIST evaluations can be used for what is called a *development dataset*, which includes speech data used to build the models in the system, population data for LR computation, and validation databases in order to test the system prior to the evaluation. The participants must submit their results without knowing the ground-truth labels of the comparisons and without hearing the audios, as opposed to the previously described NIST HASR evaluation. In the results presented here, the recordings to compare (*i.e.,* the evidence) consisted of two utterances of roughly 5 minutes duration each, that may come from a variety of microphone or telephone origins. The variability of the rest of conditions is strong, comparable to the NIST HASR 2010 evaluation described above, and therefore it is assumed that the comparisons in the evaluation can be as challenging as in many forensic scenarios.

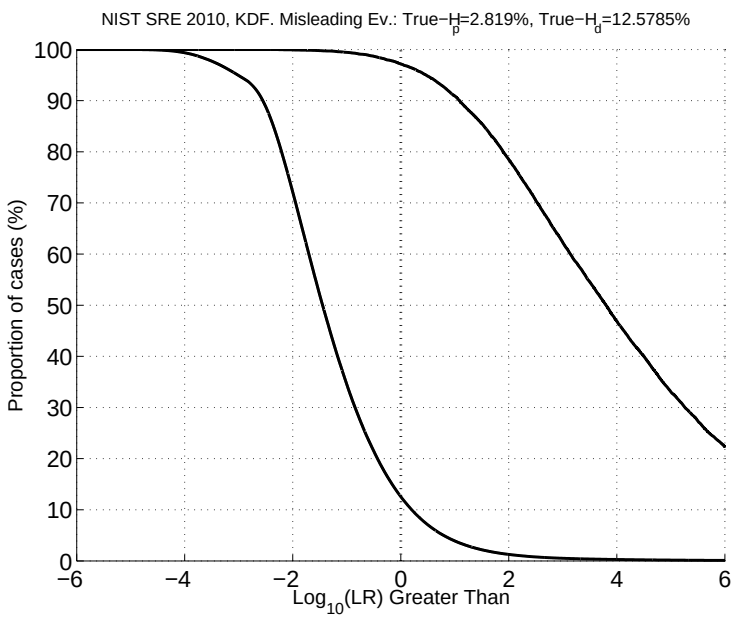
[9] A complete description of the NIST SRE 2010 dataset and protocol is available in http://www.itl.nist.gov/iad/mig//tests/sre/2010/NIST_SRE10_evalplan.r6.pdf (last accessed 28 September 2012).

The evaluation protocol considered the comparison of roughly 6000 control speech recordings with roughly 25000 recovered recordings. The number of comparisons to be conducted was roughly 700000. For each of the comparisons, the automatic system was able to give a score in a blind way. Then, for this article a LR was obtained from the score using the same Kernel Density Function (KDF) model as in Section 6.1.1 (Equation 9), which assigns a probability density to the numerator and the denominator of the LR separately.

After the submission of the LR values for the evaluation, NIST released the ground-truth labels, and the participants could check the performance of their systems. Figure 12 shows the performance of the scores of the system submitted by us, with LR values computed using the described KDF model. Performance is represented in the form of ECE and Tippett plots. Several effects are observed. First, since the solid and dashed curves in the ECE plot are reasonably close, we can say that the LR values blindly submitted to the evaluation are well calibrated, a valuable result given the challenging nature of the evaluation. Second, the accuracy is also good, since the red curve is much lower than the neutral reference in a wide range of prior probabilities. This also means that the discriminating power of the system is also good, because the blue, dashed curve is also low. Third, the strength of the LR values is also moderately high, confirming again that a method that yields LR values that have good discriminating power will give strong LR values if the calibration is also good. However, there is still some room for improvement, because the accuracy of the method is slightly worse than the neutral reference for some extreme values of the prior odds (lower than 10^{-2}).



(a)



(b)

Fig. 12. ECE plots and Tippett plots of the ATVS automatic speaker recognition method in NIST SRE 2010. Note that the range of the x-axis in Tippett plots is from −6 to 6.

7 Conclusions

This work has highlighted the importance of calibration as a desirable property of likelihood ratios (LR) computed by an evidence evaluation method.

It has also presented Empirical Cross-Entropy (ECE) plots as a valuable tool to measure the performance of LR values, including calibration, following a methodology based on Strictly Proper Scoring Rules from Bayesian statistics. The proposed methodology is not intended to replace other measures of performance of the LR previously proposed, but to be complementary to them, introducing the measurement of the calibration of the LR values as a way of improving the analysis of its performance in a validation process.

After describing some desirable properties of the LR, we have introduced calibration focusing more on intuitions rather than on complex mathematics, showing that some of those properties are achieved if the LR values are well calibrated. In particular, we have remarked that if the LR values present good discriminating power, they should express strong support, and vice-versa. We also show that this property holds if the LR values are well calibrated, but does not have to be the case if they are badly calibrated. Some examples in the context of forensic speaker recognition have illustrated how ECE plots are useful to represent and measure the performance of LR values, and how LR values present a desirable behavior if they are well calibrated.

References

[1] I. Evett, Expressing evaluative opinions: A position statement, Science and Justice 51 (2011) 1–2, several signatories.

[2] R. Royall, On the probability of observing statistical misleading evidence, Journal of the American Statistical Association 95 (451) (2000) 760–768.

[3] J. M. Curran, J. S. Buckleton, C. M. Triggs, B. S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Science and Justice 42 (1) (2002)

29–37.

[4] C. Saunders, D. T. Gantz, J. Buscaglia, E. Kalendra, The effect of uncertainty about the background population on the forensic value of the evidence, in: European Academy of Forensic Science Conference, EAFS 2012, 2012.

[5] G. Zadora, D. Ramos, Evaluation of glass samples for forensic purposes – an application of likelihood ratio model and information-theoretical approach, Chemometrics and Intelligent Laboratory Systems 102 (2010) 62–63.

[6] G. S. Morrison, F. Ochoa, T. Thiruvaran, Database selection for forensic voice comparison, in: Proc. of Odyssey 2012, Singapore, 2012.

[7] N. Gilbert, DNA's identity crisis, Nature 464 (2010) 347–348.

[8] P. Gill, J. Whitaker, C. Flaxman, N. Brown, J. Buckleton, An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA, Forensic Science International 112 (2000) 17–40.

[9] S. Perez-Gomez, D. Ramos, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, Score-level compensation of extreme speech duration variability in speaker verification, in: Proc. of Interspeech 2010, Makuhari, Japan, 2010, pp. 374–377.

[10] L. Savage, The elicitation of personal probabilities and expectations, Journal of the American Statistical Association 66 (336) (1971) 783–801.

[11] M. H. deGroot, S. E. Fienberg, The comparison and evaluation of forecasters, The Statistician 32 (1983) 12–22.

[12] T. Gneiting, A. Raftery, Strictly proper scoring rules, prediction and estimation, Journal of the American Statistical Association 102 (2007) 359–378.

[13] C. G. G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, John Wiley & Sons, Chichester, 2004.

[14] A. P. Dawid, The well-calibrated Bayesian, Journal of the American Statistical Association 77 (379) (1982) 605–610.

[15] E. T. Jaynes, Probability theory: the logic of science, Cambridge University Press, 2003.

[16] D. V. Lindley, A. Tversky, R. V. Brown, On the reconciliation of probability assessments, Journal of the Royal Statistical Society A 142 (2) (1979) 146–180.

[17] P. H. Garthwaite, J. B. Kadane, A. O'Hagan, Statistical methods for eliciting probability distributions, Journal of the American Statistical Association 100 (470) (2005) 680–701.

[18] D. V. Lindley, Understanding Uncertainty, Wiley, 2006.

[19] T. O'Hagan, Dicing with the unknown, Significance 1 (3) (2004) 132–133.

[20] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Computer Speech and Language 20 (2-3) (2006) 230–275.

[21] N. Brümmer, Measuring, refining and calibrating speaker and language information extracted from speech, Ph.D. thesis, School of Electrical Engineering, University of Stellenbosch, Stellenbosch, South Africa, available at http://sites.google.com/site/nikobrummer/ (2010).

[22] D. Ramos, Forensic evaluation of the evidence using automatic speaker recognition systems, Ph.D. thesis, Depto. de Ingenieria Informatica, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain, available at http://atvs.ii.uam.es (2007).

[23] D. vanLeeuwen, N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, in: C. Müller (Ed.), Speaker Classification, Vol. 4343 of Lecture Notes in Computer Science / Artificial Intelligence, Springer, Heidelberg - Berlin - New York, 2007.

[24] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, Information-theoretical assessment of the performance of likelihood ratio computation methods, Journal of Forensic SciencesIn press.

[25] R. Cook, I. W. Evett, G. Jackson, P. J. Jones, J. A. Lambert, A model for case assessment and interpretation, Science and Justice 38 (1998) 151–156.

[26] R. Cook, I. W. Evett, G. Jackson, P. J. Jones, J. A. Lambert, A hierarchy of propositions: deciding which level to address in casework, Science and Justice 38 (4) (1998) 231–239.

[27] B. Robertson, G. A. Vignaux, Interpreting Evidence Evaluating Forensic Science in the Courtroom, Wiley, UK, 1995.

[28] I. W. Evett, Avoiding the transposed conditional, Science and Justice 35 (1995) 127–131.

[29] T. M. Cover, J. A. Thomas, Elements of Information Theory, 2nd ed., Wiley Interscience, 2006.

[30] I. Cohen, M. Goldszmidt, Properties and benefits of calibrated classifiers, in: Proc. of European Conference on Machine Learning ECML/PKDD, 2004.

[31] I. W. Evett, J. Buckleton, Statistical analysis of STR data, Advances in Forensic Haemogenetics, Springer-Verlag, Heilderberg 6 (1996) 79–86.

[32] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, J. Ortega-Garcia, Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Transactions on Audio, Speech and Language Processing 15 (7) (2007) 2072–2084.

[33] P. Procter (Ed.), Cambridge international dictionary of English, Cambridge University Press, 1995.

[34] D. Meuwly, Reconaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approache Automatique, Ph.D. thesis, IPSC-Universite de Lausanne, 2001.

[35] D. Meuwly, Forensic individualisation from biometric data, Science and justice 46 (4) (2007) 205–213.

[36] N. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems - modelling within finger variability, Forensic Science International 167 (2-3) (2007) 189–195.

[37] N. Egli, Interpretation of partial fingermarks using an automated fingerprint identification system, Ph.D. thesis, Institute de Police Scientifique, Ecole de Sciences Criminelles (2009).

[38] C. Neumann, I. Evett, J. Skerret, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, Journal of the Royal Statistical Society, Series A: Statistics in Society 175 (2) (2012) 371–415.

[39] A. B. Hepler, C. P. Saunders, L. J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Forensic Science International 219 (1-3) (2012) 129–140.

[40] D. Ramos, J. Franco-Pedroso, J. Gonzalez-Rodriguez, Calibration and weight of the evidence by human listeners: the ATVS-UAM submission to NIST human-aided speaker recognition 2010, in: Proc. of ICASSP 2011, Pague, Czeck Republic, 2011, pp. 5908 – 5911.

[41] J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano, J. Gonzalez-Rodriguez, ATVS-UAM NIST SRE 2010 system, in: Proc. of FALA 2010, Vigo, Spain, 2010.

[42] B. W. Silverman, Density estimation for statistics and data analysis, Chapman and Hall, 1986.

Elsevier Editorial System(tm) for Forensic Science International

Manuscript Draft

Manuscript Number: FSI-D-12-00808R2

Title: Reliable Support: Measuring Calibration of Likelihood Ratios.

Article Type: Special Issue: EAFS 2012

Keywords: calibration; empirical cross-entropy; accuracy; likelihood ratio; performance; evidence evaluation

Corresponding Author: Dr. Daniel Ramos,

Corresponding Author's Institution: Universidad Autonoma de Madrid

First Author: Daniel Ramos

Order of Authors: Daniel Ramos; Joaquin Gonzalez-Rodriguez

Abstract: Calculation of likelihood ratios (LR) in evidence evaluation still presents major challenges in many forensic disciplines: for instance, an incorrect selection of databases, a bad choice of statistical models, low quantity and bad quality of the evidence are factors that may lead to likelihood ratios supporting the wrong proposition in a given case. However, measuring performance of LR values is not straightforward, and adequate metrics should be defined and used. With this objective, in this work we describe the concept of calibration, a property of a set of LR values. We highlight that some desirable behavior of LR values happens if they are well calibrated. Moreover, we propose a tool for representing performance, the Empirical Cross-Entropy (ECE) plot, showing that it can explicitly measure calibration of LR values. We finally describe some examples using speech evidence, where the usefulness of ECE plots and the measurement of calibration is shown.

Suggested Reviewers: Raymond Veldhuis
r.n.j.veldhuis@twente.nl
Expert in forensic biometrics, with solid background about calibration, the main topic of the article.

David Lucy
d.lucy@lancaster.ac.uk
His work on forensic statistics is known by his introductory book. He has also worked on calibration in recent published work.

Acknowledgements

This article is based on a keynote presentation of D. R. at the EAFS 2012 coference in The Hague, The Netherlands. We thank the organizers of the conference for the invitation. Also, the authors wish to thank Niko Brümmer, David van Leeuwen, Colin Aitken, Grzegorz Zadora, Charles Berger and Didier Meuwly for inspiring advice and discussions. Finally, we thank the reviewers for their constructive comments that have significantly improved the quality of the final work.

UNIVERSIDAD AUTONOMA DE MADRID

Dear Editors:

The manuscript is submitted for the Special Issue of Forensic Science International about the 6th European Academy of Forensic Science Conference (EAFS 2012) in The Hague, The Netherlands, in August 2012. The contents of the manuscript are based on a keynote presentation that Daniel Ramos, the corresponding author of this work, has presented at the EAFS 2012 conference. Joaquin Gonzalez-Rodriguez is co-author of the work, because of his important contributions in the contents of that keynote, and also of the present manuscript.

The contributions of the work are as follows. First, the importance of calibration of likelihood ratios is described and highlighted. Second, a methodology based on information theory in order to assess the performance of likelihood ratios considering their calibration is proposed. Third, a software package is presented in Matlab™ in order to assist forensic examiners in the use of the proposed performance assessment tools. Although calibration of probabilistic assessments is a concept that has been studied for decades and it is not new, its application to the particular framework of forensic inference is seen as the main contribution of this manuscript.

The contact information of the corresponding author is the following:

Daniel Ramos
Research Institute of Forensic Science and Security (ICFS). ATVS – Biometric Recognition Group
Escuela Politecnica Superior. Universidad Autonoma de Madrid
E-28049 Madrid
Tel: (+34) 914976206
Fax: (+34) 914972235
E-mail: daniel.ramos@uam.es

We look forward for having early news from you.

Best Regards,

Daniel Ramos Castro.