Validation of Likelihood Ratio Methods for Forensic Evidence Evaluation Handling Multimodal Score Distributions

Rudolf Haraksim[#], Daniel Ramos[+], Didier Meuwly[*,**]

[#]Signal Processing Laboratory, Swiss Federal Institute of Technology Lausanne, Switzerland

[+]ATVS – Biometric Recognition Group. Escuela Politecnica Superior. Universidad Autonoma de Madrid. C/ Francisco Tomas y Valiente 11. 28049 Madrid, Spain

*Netherlands Forensic Institute, Laan van Ypenburg 6, 2497GB The Hague, The Netherlands

** University of Twente, Drienerlolaan 5, 7522NB Enschede, The Netherlands

Keywords – Validation, Multimodal Score Distribution, Calibration, AFIS, Forensic Evaluation, Fingermark, Likelihood Ratio, Robustness, Reproducibility, Accuracy

ABSTRACT: This article presents a method for computing Likelihood Ratios (LR) from multimodal score distributions produced by an Automated Fingerprint Identification System (AFIS) feature extraction and comparison algorithm. The AFIS algorithm used to compare fingermarks and fingerprints was primarily developed for forensic investigation rather than for forensic evaluation. The computation of the scores is speed-optimized and performed on three different stages, each of which outputs discriminating scores of different magnitudes together forming a multimodal score distribution. It is worthy mentioning that each fingermark to fingerprint comparison performed by the AFIS algorithm results in one single similarity score (e.g. one score per comparison). The multimodal nature of the similarity scores can be typical for other biometric systems and the method proposed in this work can be applied in similar cases, where the multimodal nature in similarity scores is observed. In this work we address some of the problems related to modelling such distributions and propose solutions to issues like data sparsity, dataset shift and over-fitting. The issues mentioned affect the methods traditionally used in the situation when a multimodal nature in the similarity scores is observed (a Kernel Density Functions (KDF) was used to illustrate these issues in our case). Furthermore, the method proposed produces interpretable results in the situations when the similarity scores are sparse and traditional approaches lead to erroneous LRs of huge magnitudes.

1. INTRODUCTION

The commercial "off-the-shelf" AFIS algorithms producing similarity scores[1] are primarily developed to support the process of selection of candidates for forensic investigation and not intended for the use in forensic evidence evaluation [1]. The algorithm selected was speed optimized to perform large number of comparisons in the shortest time possible. Not only is the comparison process[2] speed-optimized, it is performed on three different stages, each of which outputs similarity scores of different magnitudes, together forming a multimodal score distribution. The scores that have been output in the first two stages of the algorithm are referred to as "early outs" and the "full" (profound) comparison is only performed in the final stage. The similarity score outputs of the three different stages of the AFIS algorithm used are illustrated in the Figure 1 below.

[1] The aim is to discriminate the fingermarks originating from the same fingers from the fingermarks originating from different fingers.

[2] Undet the "comparison process" we mean the fingermark to fingerprint comparison.

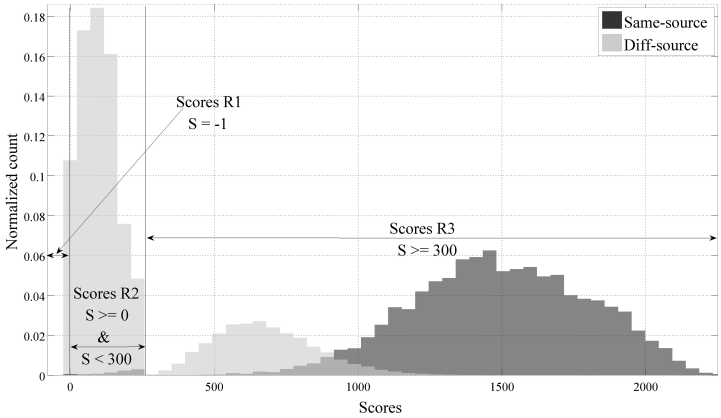

Figure 1 –Different regions of the similarity scores produced by the AFIS algorithm.

The scores provided by the AFIS algorithm used are structured in three regions (R):

- Region 1 (R1) – the first mode of the AFIS – all the comparisons performed at this stage are assigned identical scores value of "-1". The event of observing a similarity score after a fingermark to fingerprpit comparison in this mode occurs when the AFIS algorithm finds no minutiae in agreement between the mark and the print.
- Region 2 (R2) – the second mode of the AFIS – all the comparisons performed in this stage are assigned score values in the range between 0 and ,300. The event of observing a similarity score after a fingermark to fingerprint comparison at the second stage occurs when some similarities are observed, which are however not sufficient to proceed to a full comparison.
- Region 3 (R3) (the third mode of the AFIS) – all the comparisons performed in this stage are assigned score values greater than 300 and a full comparison is performed. The event of observing a similarity score after a fingermark to fingerprint comparison at the third stage occurs when the AFIS algorithm finds the incoming images similar enough to perform a full comparison.

Table 1 sumarizes the scoring process in the three modes of the AFIS algorithm used[3] in more organised way.

Table 1 – different operating stages of the AFIS algorithm

Algorithm processing stage	Score range
1 – early outs 1	-1
2 – early outs 2	[0 , 300]
3 – full comparison	>300

In this work we propose a method to handle the multimodal score distributions resulting from the comparison process, which is also robust to the sparsity in the data. This article is structured in the following way. In section 2 we introduce the datasets used, section 3 is dedicated to the definition of the problem and introduction of the baseline LR method. In section 4 we present our solution to the data sparsity and lack of data in the multimodal score distributions. In section 5 we discuss the performance measures used and present the results in the section 6. Finally we present the discussion and conclusion in section 7.

2. DATASETS USED

Since it is notoriously difficult to find forensically relevant, sufficiently large datasets with the ground truth about the origin of the samples known, we decided to use a set of simulated[4]

[3] It is important to note that the AFIS algorithm uses different scoring methods for the fingermarks containing up to 10 minutiae and for the 11minutiae and above. The fingermarks in 8-minutiae configuration are used in this article. Different AFIS / biometric systems might provide similarity scores of different magnitudes.

[13] [15] fingermarks in 8-minutiae configuration from 6 individuals, paired with their corresponding fingerprints. The fingermarks were obtained by capturing an image sequence of the finger of each individual from an optical live scanner (Smiths Heimann Biometrics ACCO 1394S live scanner) and splitting the frames captured into 8 minutiae configurations.
AFIS scores of simulated fingermarks and the corresponding ground-truth reference fingerprint as training data are used for modelling the SS scores (numerator of the LR), captured from the same individual in controlled conditions. For modelling the DS scores (denominator in LR) we used the fingermark in the case compared against a 200,000 - fingerprint subset of the police database. The values assigned to the parameters of the distributions $\hat{\theta}$ are obtained from the data summarized in the table 2.

3. Table 2: Same and different source scores.

Individual	$\Delta\hat{\theta}_p$ - SS scores	$\Delta\hat{\theta}_d$ - DS scores
Person 1	8,455 marks 1 print	8,455 marks 200,000 prints
Person 2	2,751 marks 1 print	2,751 marks 200,000 prints
Person 3	4,666 marks 1 print	4,666 marks 200,000 prints
Person 4	2,206 marks 1 print	2,206 marks 200,000 prints
Person 5	3,179 marks 1 print	3,179 marks 200,000 prints
Person 6	3,758 marks 1 print	3,758 marks 200,000 prints

For example scores for the Evidence Same Source (E_{SS}) are obtained on a "leave-one-out" basis from the SS score distribution (fingermarks of Person 1 and fingerprint of the Person 1) and scores for the Evidence Different Source (E_{DS}) are obtained from the AFIS scores of the fingermarks of Person 1 with the fingerprints of Persons 2-6. This process is repeated iteratively for each individual. In the "leave-one-out" approach we iteratively sweep through the set of fingermarks. With every iteration we delegate one of the fingermarks to play the role of the crime-scene mark (designated as y^{th} mark abbreviated by "*my*") and maintain the remaining fingermarks to form SS and DS score distributions for training the LR method. The concept of the LR and the method used will be discussed in length in the following sections.

3. PROBLEM DEFINITION

Traditional way to handle the multimodal score distributions is to adopt a holistic approach, e.g. finding a single-function description of the multimodal score distribution (examples of application single-function description of univariate, multimodal score distribution – the KDF can be found in [2,3]). The use of single-function description of such score distributions is well justified, if the scoring mechanism (in our case AFIS comparison algorithm) produces the similarity scores in a uniform or continuous manner.

Our situation is different. The selected AFIS comparison algorithm produces the similarity scores of three different magnitudes and the events of observing the similarity score, e.g. the result of a fingermark to fingerprint comparison, are mutually exclusive – meaning that the similarity score of a certain magnitude can only be observed in a particular region.

The core of the proposed solution therefore rests in treating the three score regions independently. In this approach we do however face a problem of data sparsity, as majority of the SS scores project in the region 3, leaving a few SS observations for the regions 1 and 2. Similar situation occurs with the DS scores, where majority of these scores project into the regions 1 and 2, leaving a few observations for the region 3. A situation where no SS, or no DS score is observed in a particular region poses a significant problem for evaluating the strength of evidence – e.g. resulting in LR values of unreasonably high magnitudes or infinities.

In section 3.1 we present a model used for computing the Likelihood Rations (using the KDF method) and in section 3.2 we highlight the problems with the holistic approach when using the KDF.

---

[4] Simulated fingermarks in this case refer to series of image captions of a finger moving on a glass plate of the fingerprint scanner (the procedure is described in detail in [13]).

3.2 LR COMPUTATION

When computing LR values in forensic applications most of the time we encounter a problem of the choice of the population database. In forensic fingerprints the role of the legacy police databases in individual countries play this role. While comparing fingerprints in forensic scenario we encounter fingerprints of degraded quality, high distortion, partial fingerprints (e.g. fingermarks) etc…

We can shape the prosecution and defense propositions at different levels depending on the investigation scenario – level of the source (where we inquire regarding the source of the fingermark), activity level (where we inquire regarding transfer of the fingermark onto the crime scene) or at the crime level (not commonly addressed in fingerprints as it usually implies transfer of the crime scene material onto the suspected individual).

At the source level for the same source hypothesis we further inquire whether the questioned fingermark is coming from a particular finger of the suspected individual or any finger of the suspected individual (finger / person level propositions). For the different source proposition we can inquire whether the print is coming from a different finger of a suspected individual, from a particular finger (for example the right-hand thumb) of any other individual in the database (conditioning on a particular finger is not common for the DS hypothesis), or any finger of any other individual in the database.

The fingermark primarily address the question of its source, we set the forensic propositions at source level, namely:

Hp: The fingermark and the fingerprint originate from the same finger
Hd: The fingermark and the fingerprint originate from different fingers[5]

The LR's in the case of the KDF are produced from the Same Source (SS) and Different Source (DS) score distributions (examples are shown in the Figure 1) and approximated using the KDF. In our case, the distribution of SS and DS scores observed in each region varies in its shape, mainly due to the three-different-stage scoring process. In most of the cases, the majority of the SS scores projects into the R3 region, because a comparison showing high degree of similarity tends to be a SS comparison resulting in a score > 300. Conversely, the majority of the DS scores projects in the R1 and R2 regions, because a comparison showing low degree of similarity tends to be a DS comparison and results in a score < 300.

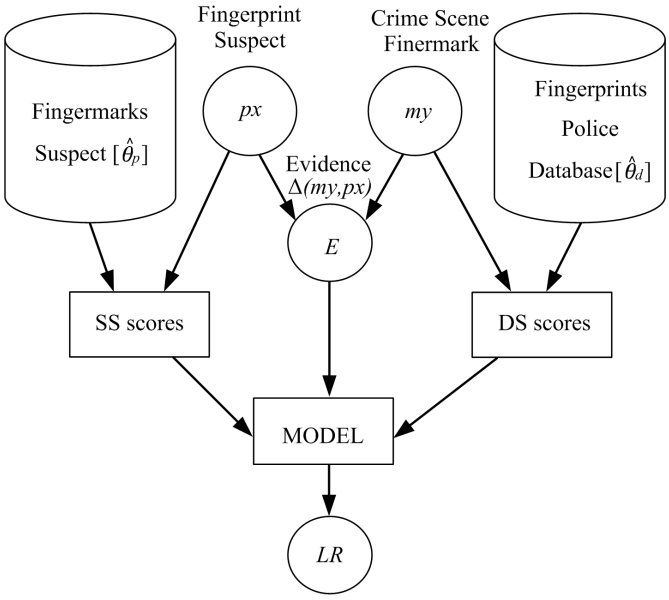

Figure 3 – LR model example

[5] Change at the level of the proposition induces a change in the LR model.

In the forensic literature different strategies have been proposed for calculating LRs from continuously distributed AFIS scores. In the field of score-based biometric recognition [4, 5, 6, 7, 3, 8], the following LR model has been defined (example shown in Figure 3):

$$LR = \frac{f(S \mid H_p)}{f(S \mid H_d)} = \frac{f(\Delta(y,x) \mid H_p, \Delta\hat{\theta}_p)}{f(\Delta(y,x) \mid H_d, \Delta\hat{\theta}_d)} \qquad \text{(eq. 1)}$$

where for the fingerprint evidence evaluation datasets are defined in the following way:
f() – in the equation stands for the probability density function applied to the continuous distribution of the similarity scores
S = $\Delta(y, x)$ – a similarity score between the fingermark *y* found on the crime scene and the fingerprint *x* of the suspect. It is usually referred to as the *evidence score*.
$\Delta\hat{\theta}_p$ – similarity scores obtained from comparing training set of simulated fingermarks of the suspect with the reference fingerprint of the suspect
$\Delta\hat{\theta}_d$ – scores obtained from comparing the crime scene fingermark and a subset[6] of fingerprints from the population database used in the model (in this case a subset of operational 10-print card police database).

Furthermore, we will use below the following notation to refer to the parameters of the models:
θ – represents the parameters of the model (e.g. mean, variance) that need to be trained
$\hat{\theta}$ – represents a value given to the parameters of the model, obtained from the scores of the training set

3.2 BASELINE METHOD - KDF

KDF is typically a first choice to handle univariate, multimodal distributions, as it can be seen in [2, 3]. Therefore, we will use it as a baseline in this work. We do nevertheless anticipate that KDF is a method that can be prone to over-fitting (see Figure 2), an undesired behaviour as it often leads to the Likelihood Ratio (LR) values of erroneously enormous magnitudes, mainly due to the poor description of the tails of the score distribution closely related to the lack of data.

In Figure 2 we see two examples of clearly erroneous behaviour of KDF due to overfitting. The plot on the left-hand-side shows the LR for Same Source evidence (LREss) resulting in an enormous magnitude (numerically infinite), while the plot on the right-hand-side illustrates an example of a LR = 10^{91} for Different Source evidence (LREds), which is supporting the wrong proposition in a very strong way.

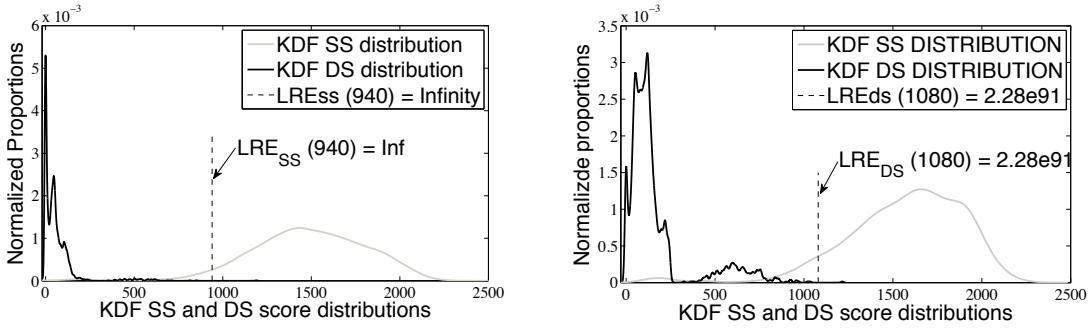

Figure 2 – KDF fit to the SS and DS score distributions – example magnitudes

[6] The subset (rather than the entire set) of the AFIS database, counting 20k individuals (200'000 fingers), was chosen based on the fact, that the features extracted from the fingerprints in this subset have been manually (human) verified.

4. SOLUTION PROPOSED

We do not aim being critical towards the use of the KDF and maintain that in other scenarios it may be useful. The LR's of extreme magnitudes obtained using the KDF presented in Figure 2 however advocate for an alternative solution (this statement is further supported in the results section of this article).

4.1 MULTIMODAL LR METHOD

Each fingermark vs. fingerprint comparison by the AFIS system used results in one single score. This score can either belong to region 1, region 2 or region 3. The events of observing the similarity scores in a particular region (R1, R2 or R3) are therefore *"mutually exclusive"*. This is an intrinsic property of the AFIS algorithm used. The scores in the R1, R2 and R3 for both of the propositions together form a multimodal score distribution. The scores in the three regions *"exhaustively"* cover the multimodal score distribution – R1 ∪ R2 ∪ R3 = multimodal distribution. In the method proposed we will split the SS and DS score distributions into the three regions of interest, since the events of observing an AFIS score in different regions are mutually exclusive and exhaustive (Figure 4).

Due to the fact that the events of observing similarity scores in the three regions are mutually exclusive and exhaustive, we can rewrite equation 1 in the following way:

$$LR = \frac{f(S \mid H_p)}{f(S \mid H_d)} = \frac{f(S \mid R_1, H_p) \times P(R_1 \mid H_p) + f(S \mid R_2, H_p) \times P(R_2 \mid H_p) + f(S \mid R_3, H_p) \times P(R_3 \mid H_p)}{f(S \mid R_1, H_d) \times P(R_1 \mid H_d) + f(S \mid R_2, H_d) \times P(R_2 \mid H_d) + f(S \mid R_3, H_d) \times P(R_3 \mid H_d)}$$

(eq.2)

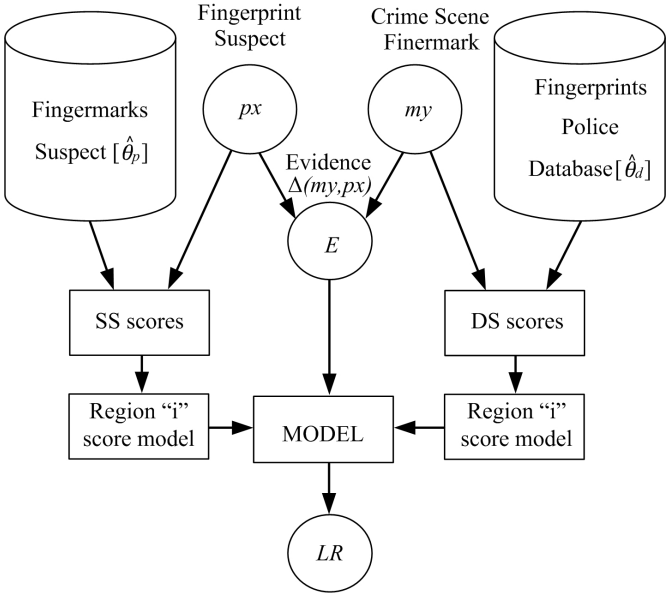

Figure 4 – Multimodal method

By substituting the region numbers in with i (e.g. i = 1, 2, 3) we obtain the following compact representation of the equation 2:

$$LR_i = \frac{f(S \mid R_i, H_p) \times P(R_i \mid H_p)}{f(S \mid R_i, H_d) \times P(R_i \mid H_d)} \qquad \text{(eq. 3)}$$

Where $\frac{P(R_i \mid H_p)}{P(R_i \mid H_d)}$ is the ratio of probabilities of observing R_i scores given that the fingermark and the fingerprint originates from the same finger over the probability of observing R_i scores given that the fingermark and the fingerprint originates from different fingers.

4.1.1 Scores in the Region 3

$$LR_3 = \frac{f(S \mid R_3, H_p) \times P(R_3 \mid H_p)}{f(S \mid R_3, H_d) \times P(R_3 \mid H_d)} \qquad \text{(eq. 4)}$$

From the histograms of the SS and DS score distributions in Figure 5 we consider as a reasonable initial assumption that the scores in the R3 region are following a Gaussian (Normal) distribution.

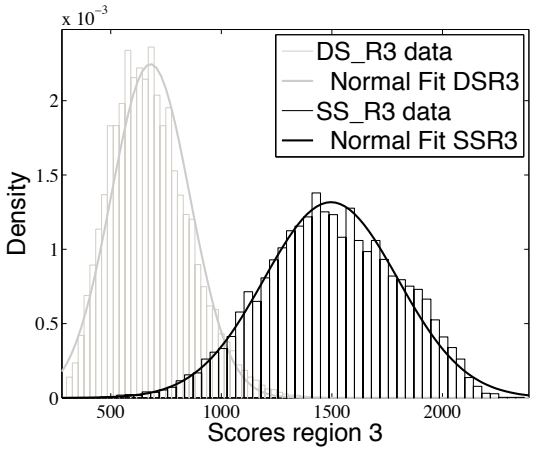

Figure 5 – Gaussian fit to the R3 region score distributions

4.1.2 Scores in the Region 2

$$LR_2 = \frac{f(S \mid R_2, H_p) \times P(R_2 \mid H_p)}{f(S \mid R_2, H_d) \times P(R_2 \mid H_d)} \qquad \text{(eq.5)}$$

The DS score distribution in the R2 region appears to be skewed, and the SS score distribution seems to be monotonically raising in this region. Although different parametric and non-parametric data fits have been tested for the R2 region scores [22], Beta function was chosen mainly due to the modelling simplicity to describe the score distributions in this region.

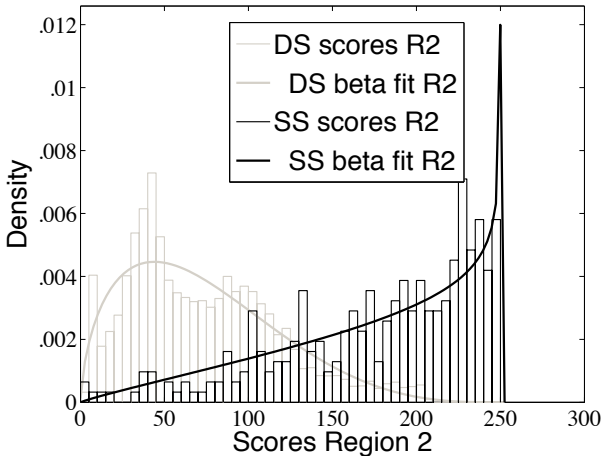

Figure 6 – Beta fit to the R2 score distributions

4.1.3 Scores in the Region 1[7]

As mentioned in the introduction section, all of the scores observed in the R1 region get one particular value of the similarity score (S_{R1} = -1) assigned by the AFIS algorithm. Equation 6 for the R1 region will have this form:

$$LR_1 = \frac{f(S \mid R_1, H_p) \times P(R_1 \mid H_p)}{f(S \mid R_1, H_d) \times P(R_1 \mid H_d)} \qquad \text{(eq. 6)}$$

where $P(S \mid R_1, H_p)$ is the probability of observing a -1 score (S_{R1} = -1) amongst all the scores observed in the R1 region under the H_p, an event, which is always true and we can write $P(S \mid R_1, H_p) = 1$. The same logic applies to the $P(S \mid R_1, H_d)$, which is a probability of observing a -1 score (S_{R1} = -1) amongst all the scores falling into the R1 region under H_d, an event which again is always true and we can write $P(S \mid R_1, H_d) = 1$.

By aplying the above-mentioned conditions, eq. 6 further simplifies to a ratio of probabilities of observing score in the R_1 region under both propositions $\frac{P(R_1 \mid H_p)}{P(R_1 \mid H_d)}$. The scores in region R_1 possess certain evidential value, despite the fact that all of them share the same discrete value.

Let's assume that very few SS are observed in the R_1 region, and that they are mostly DS scores. If we observe a score of -1 (in the R_1 region), the LR should support the defence hypothesis. This happens if $LR = \frac{P(R_1 \mid H_p)}{P(R_1 \mid H_d)}$. Additionally, an apparent solution to the -1 scores of ignoring them because all of them have the same value appears to be a waste of the discriminating information given by the fact that in R_1 there are mostly DS scores.

4.2 ROBUSTNESS TO THE LACK OF DATA

In forensic science, the apparent problem of assigning probabilities when no observations are made in the training data has been studied for example in [9]. In the end the model has to be completed by assigning the following probability ratio for each region R_i:

$$\frac{P(R_i \mid H_p)}{P(R_i \mid H_d)}$$

[7] The similarity scores in this region also carry evidential information, and as such should not be simply ignored.

these are probabilities of observing a score in i-th Region respectively under prosecution and defence propositions.

4.2.1 Example with a simplified binary division

Assigning $P(R_i|H_p)$ and $P(R_i|H_d)$ to each of the different regions R_i, $i=1,2,3$ needs to consider some robustness about the sparsity of the scores in the training set. In order to illustrate this, we start with a simplified example, where we have divided the score axis in two regions R_1 and R_2 – a *binary* division. We consider for illustration the scores under the assumption that H_d is true, but this example can be analogously applied to the scores under the assumption that H_p is true. In order to assign a probability $P(R_i|H_d)$ that a given score will be observed in region R_i, we need previous knowledge regarding the observations of similarity scores falling into different regions – some training observations. Those observations are taken from the training scores $\Delta\hat{\theta}_d$, with N_d – the number of scores observed under the defence proposition, in the following way. Let $R_d = \left\{R_d^1,...,R_d^{N_d}\right\}$ be a sample of random variables, where R_d^j represents the region in which the *j-th* different-source training score was observed. In this *binary* example, the possible outcomes of each R_d^j are R_1 and R_2. Then, the outcome of R_d^j will be the region in which the *j-th* score in $\Delta\hat{\theta}_d$ is observed. Thus, the training observations are the particular values of each of those random variables R_p^j. We assume that variables $R_d = \left\{R_d^1,...,R_d^{N_d}\right\}$ are identically distributed according to a Bernoulli distribution, where the probability that a score is observed in Region *i* is precisely $P(R_i|H_d)$, the parameter of the model. Moreover, we assume that the variables are *conditionally independent given the model*. Then, it can be shown that the *maximum likelihood* rule for the probability that a score will fall into Region *i* is as follows:

$$P\left(R_i\middle|H_d\right) = \frac{M_i}{N_d} \qquad \text{(eq. 7)}$$

where M_i is the number of scores in the training set observed in Region *i* and the N_d is the number of observations of the scores under the defence proposition. If the training scores under H_d for whatever reason contains zero score observations in Region *i*, i.e. $M_i = 0$, we get the following:

$$\frac{P\left(R_i\middle|H_p\right)}{P\left(R_i\middle|H_d\right)} = \frac{P\left(R_i\middle|H_p\right)}{\frac{M_i}{N_d}} = \frac{P\left(R_i\middle|H_p\right)}{\frac{0}{N_d}} = \infty \qquad \text{(eq. 8)}$$

In some cases this might result in a $LR = \infty$. An analogous derivation results in $LR = 0$ for same-source scores falling in a region where no same-source scores have been observed before.

An outcome of $LR = 0$ or ∞ is very likely to occur if a similarity score, either SS or DS, is not observed in one of the regions. The problem arises particularly in the R1 region, where the SS scores are quite rare, but can also occur in the R2 or R3 region as well.

4.2.2 Bayesian solution

In order to avoid "zeroes" in either numerator or denominator of the LR and to assure a valid numerical input, we propose a Bayesian solution to $P(R_i|H_d)$. We start from the above binary example, where a maximum likelihood rule was considered. Under the same assumptions, if we instead consider that the probability $P(R_i|H_d)$, the parameter of the Bernoulli distribution, has a uniform prior distribution (in the [0,1] range), then it can be shown that the solution inferred is the *predictive distribution*, which takes the following form:

$$P\left(R_i|H_d\right) = \frac{M_i + 1}{N_d + 2} \quad \text{(eq. 9)}$$

A full derivation is tractable, and can be found in [10] (Equations (6.66) to (6.73)). This result is known as the *Laplace rule of succession* [11]. For simplicity the application of this rule on our dataset will be demonstrated on R1 region, where all the scores attain a discrete value S = -1. Recall the binary example, where in the R1 region we obtained LR = ∞ because there were no observed scores in that region in the training data. Suppose a number of DS training scores $N_d = 20$ and that none of these scores are observed in the Region 1, thus $M_i = 0$. Then, according to the previously proposed maximum-likelihood rule we would obtain $P(R_1 \mid H_d) = \frac{M}{N} = \frac{0}{20} = 0$ and the LR would be infinite. However, with the Bayesian uniform prior on the model's parameter (Laplace rule of succession) we get following $P(R_1 \mid H_d) = \frac{M+1}{N+2} = \frac{1}{22} \approx 0.05$, which with increasing number of scores will be asymptotically approaching zero, but will still provide a non-zero numerical value. The interpretation of this result is that, additionally to the training data, a uniform prior for the model parameters forces to consider always at least an observation of one score in each of the regions. Therefore, if H_d is true, we have to consider $N_d + 2$ scores, and the scores observed in each region will be at least one. An analogous derivation provides equivalent interpretation for the case when H_p is true.

4.2.3 Generalization to more than 2 regions

The problem addressed in this work requires a generalization with respect to the rule of succession for the binary example, because we are dividing the score range into more than 2 regions. That means that the variables $\left\{R_d^1,...,R_d^{N_d}\right\}$ will now have more than 2 possible outcomes, and therefore their distribution cannot be a Bernoulli distribution. The generalization to more than 2 possible outcomes, say *Q* possible regions, involves the assumption that the variables $\left\{R_d^1,...,R_d^{N_d}\right\}$ follow a multinomial distribution. Moreover, since there are now *Q* parameters for this multinomial model, the prior uniform distribution of the model parameters will be a Dirichlet distribution. Under these conditions, the derivation of the predictive distributions $P\left(R_i|H_d\right)$ for each of the regions can be found in [12], and therefore generalizes the rule for more than 2 regions. That generalization provides the following result for the predictive distribution:

$$P\left(R_i|H_d\right) = \frac{M_i + 1}{N_d + Q} \quad \text{(eq. 10)}$$

or, in the case of 3 regions as in the problem we address in this article, we have:

$$P\left(R_i|H_d\right) = \frac{M_i + 1}{N_d + 3} \quad \text{(eq. 11)}$$

Again, the analogous derivation produces a similar result for the case where H_p is true.

In our model, equation 11 will be used in all three regions to assign all the probabilities $P\left(R_i|H_p\right)$ and $P\left(R_i|H_d\right)$. This is because in cases where there are both SS and DS scores present values, the probabilities do not change significantly with respect to the maximum likelihood solution. In cases where there are zero scores of either SS or DS it will give robustness to the model, avoiding results of *LR* = 0 or *LR* = ∞[8].

[8] One could argue that a system providing LR = 0 or ∞ is the best that can be achieved if always correct. However when the quality of the data is limited, a system providing LR = 0 or ∞ is not desirable, since

The motivations for the use of the Laplace rule of succession and its generalization are thoroughly justified in [10] and [11].

5. PERFORMANCE MEASURES

We will measure the performance of the KDF and the MULTIMODAL LR methods, mainly focusing on their accuracy and the discriminating power. The accuracy was defined in [14] as the closeness of agreement between the decision – driven by a LR computed using a given method – and the ground truth. The discriminating power was defined in [14] as the property of a set of LR's that allows distinguishing between the propositions involved. In the forensic biometric literature [15,16,17] Log Likelihood Ratio Cost (Cllr) is an accepted measure of accuracy. It will be illustrated on an ECE curve [15]. To measure the discriminating power of the two methods we will use the EER (graphically represented in the Detection Error Trade-off DET [18] curve) and the minimum value of the Cllr - the $Cllr^{min}$ and its graphical representation in the ECE^{min} curve [15]. Alongside the ECE and DET plots a Tippett plot [1, 6, 19] showing the rates of misleading evidence for the prosecution and defense [20] (RMEP / RMED) will be presented as well. A detailed description for each graphical representation is provided in the results section.

6. RESULTS

A brief summary of the experimental setup is shown below in table 3.

Table 3. Different methods for LR calculation for the multimodal and baseline method

MULTIMODAL method		
Region 1	**Region 2**	**Region 3**
$\propto (SS/DS)_{Bayesian}$	Beta	Normal
BASELINE method		
KDF baseline for the entire SS and DS score distributions in all regions		

The robustness to the lack of data issue is well visible in the Tippet plots in Figure 7, where the baseline KDF method shows sub-optimal performance in the lower right corner when the inverse cumulative density function of the LR_{SS} fails to converge in the bottom right corner. In extreme cases the LR values reach infinity. Please note that the log(LR) values have been limited for illustration purposes.

Even though identical datasets were used in both methods, the resulting cumulative density functions appear much more refined using the multimodal method. Although we observe similar rates of misleading evidence in both cases, in roughly 3% of the cases the baseline KDF provides unjustifiably high LR values.

---

such an output underestimates or overestimates the quantity of information available particularly in the trace.

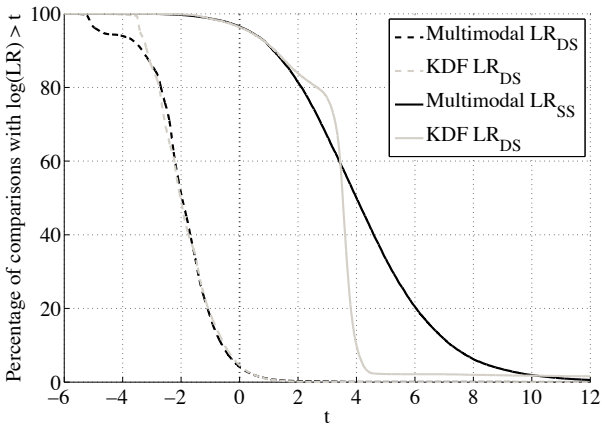

Figure 7 – Tippett plots of the KDF vs. Multimodal LR method

In the Figure 8 we show the Equal Error Rate (EER) and the Detection Error Trade-off[9] (DET) curves [18]. Similarly to the Tippett plots shown earlier, DET plots confirm the deviation from the optimal performance of the KDF method (dashed curve).
Analysing the DET curves we observe, that in some cases the baseline method produces extreme LR values reflecting more an artefact of the modelling approach than expressing the real evidential value of the findings. In the top left corner of the DET curves we clearly see a deviation from otherwise linear distribution of errors for the baseline KDF. This happens because some of the DS evidence scores (roughly 0.1% of the total DS scores) yield an extremely large LR value, strongly supporting the wrong proposition. This is a highly undesirable effect, which has consequences on the reliability of the LRs that are computed using the KDF method. We can measure the discriminating capabilities of a LR method in terms of EER, though as the EERs observed for both of the methods are relatively close to each other it is apparent that measuring the performance of a LR method solely using the discriminating power becomes insufficient. The EERs observed for either of the methods were EER_{KDF} = 3,625 and $EER_{MULTIMODAL}$ = 3,877.

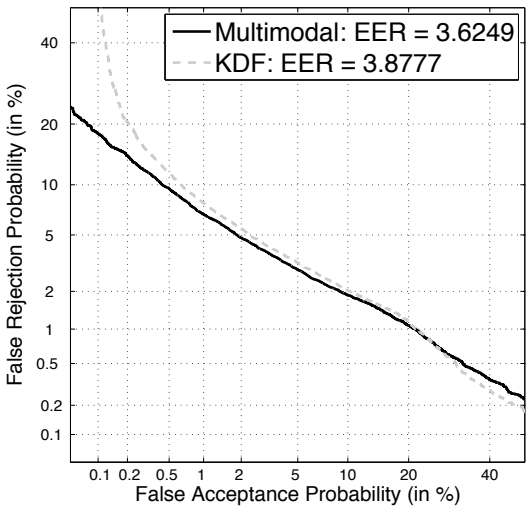

Figure 8 – DET plots of the KDF vs MULTIMODAL LR method

From the ECE plot (Figure 9) we can deduct the performance measures of the two LR methods. The solid line represents the accuracy of a set of LRs (the lower the solid curve, the better the accuracy) and the dashed curve represents the discrimination of the LRs (the lower the dashed curve the better the discriminating power of the LR). The solid minus the dashed curve represents calibration loss of the set of LRs. The smaller the distance between the solid

[9] The DET curve is a 2 dimensional plot of false acceptance and false rejection rates evenly handling both error types plotted on the gaussian wrapped scale. Linearity of the DET curves is due to the assumed "normality" of the LRs. The closer the curve to the coordinate origin, the better the discrimination capabilities of the model [18].

and dashed curves, the better the calibration of the system. The lower the dashed curve, the better the discriminating capabilities of the system. Ideally, both solid and dashed line should be below the black dotted curve, which represents a reference system that continuously returns *LR = 1*. The ECE is directly linked to the Cllr. The Cllr (measure of the accuracy) on the ECE plot lays on the intersection of the solid curve and "zero" prior log odds and the $Cllr^{min}$ (alternative measure of the discriminating power) lays on the intersection of the dashed curve with the "zero" prior log odds line.

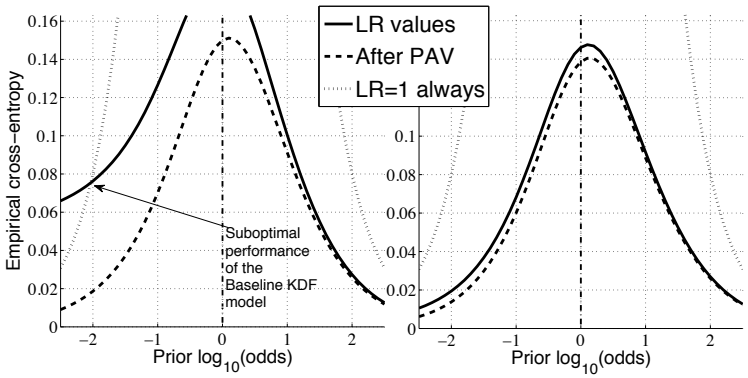

Figure 9 – ECE plots of the Baseline KDF (left) vs. MULTIMODAL LR method (right)

Clearly undesired behaviour of the baseline KDF method is visible at the prior log odds smaller than 10^{-2}, where its performance is worse than that of the reference method which constantly returns LR = 1. The experiments and the visualisation tools used seek to present operational limits and constraints the two methods. This does not mean that the KDF should not be used in similar cases for modelling multimodal distributions; it simply shows the operational constraints of this method. It also warns about the low reliability of the LRs in certain situations, particularly when the prior odds in the case will be low. In other words, it appears to be "safe" to rely on the LR produced by the baseline KDF method only for the prior-log-odds bigger than 10^{-2}. Conversely, the LRs produced by the multimodal method proposed can be deemed as reliable, because the calibration of this method is good for all the regions of the prior odds (see figure 9) and because the solid ECE curve of the multimodal method never remains under the one of the reference method.

Table 4 summarises the results presented graphically above. The improvement of the proposed system compared to the baseline KDF achieved was approximately 21% for the Cllr and 6.5% for the EER.

Table 4. Discriminating capabilities and calibration of the KDF and MULTIMODAL methods

MULTIMODAL Method			**Rate of Misleading Evidence**		**Performance**		
Region 1	**Region 2**	**Region 3**	**RMEP**	**RMED**	**EER**	**$Cllr_{min}$**	**Cllr**
$\propto$ (SS/DS)	Beta	Gauss	3.45	4.6	3.62	0.14	0.15
BASELINE method			**Rate of Misleading Evidence**		**Performance**		
Baseline KDF[10] all regions			3.6	4.16	3.87	0.15	0.19

The objective of this article was not to find the best performing LR method but to deal with similarity scores of a particular AFIS comparison algorithm, which presented multimodal distributions. The methods used to represent data in different regions may vary in different biometric modalities (or different AFIS systems). For example in [21] the AFIS scores are shown to be best modelled by a log-normal distribution.

7. DISCUSSION AND CONCLUSION

The main drawback of the traditionally used KDF method for the multimodal score distributions is its poor description of the tails of the training score distributions, together with

[10] The performance of the baseline KDF method was only possible to measure after removing the extreme outliers (LR = infinity) and setting a hard limit at log(LR) = 30. As such the reader is required to treat the KDF baseline method results with certain amount of moderation in mind.

a tendency to over-fit the underlying score distributions. Using the KDF we observed LRs of enormous magnitude supporting the correct proposition (e.g., $LR_{Ess} = 10^{130}$, $LR = \infty$) in extreme cases, and even supporting the wrong proposition (e.g., $LR_{Eds} = 10^{91}$). This provides an illusion of certainty that transcends reality and leads to a misleading interpretation of forensic evidence. In the ECE plots we observed poor calibration of the baseline KDF method in the low prior-odds region, a problem that is avoided using the multimodal method proposed in this work.

In the multimodal method proposed we have split the SS and DS score distributions into three different regions (R1, R2 and R3), and used a particular score region depending on the magnitude of the evidence score observed. The multimodal method did not dramatically improve the discrimination capabilities of the system in terms of EER (6.5% relative improvement of the multimodal method over the baseline KDF). Measuring performance solely based on the observation of the EER related to the discriminating capabilities of a given LR method appears to be insufficient and measuring other performance characteristics appears to be highly desirable. The relative improvement of the calibration by 25% is considered highly important. Moreover, we have shown that using a multimodal method we can produce well-calibrated LRs for the whole range of the prior odds in a case, as shown on the ECE plots in Figure 9.

Due to its good performance and its computational simplicity, the multimodal LR method was used in [14] to evaluate the coherence of the discriminating scores produced by an AFIS algorithm.

In this article we have highlighted issues – lack of data and over-fitting – related to the modelling multimodal score distributions, which were in our case produced when comparing a fingermark and a fingerprint by an AFIS algorithm. By applying a traditional method we illustrated nuisance LR values when modelling the SS and DS score distributions as whole.
We have proposed an alternative method that is robust against the lack of data and does not over-fit the score distributions. The method proposed can be used in cases where similar multi-modal score distributions are observed.
With the LR method proposed, in future work we will proceed further with the definition of additional validation criteria, apply “real” forensic marks to the method selected (rather than simulated marks) and reproduce the results for 5 – 12 minutiae configurations based on the data from real forensic casework (see [22]).

ACKNOWLEDGEMENTS

The research was conducted in scope of the BBfor2 – Marie Curie Initial Training Network (FP7-PEOPLE-ITN-2008 under the Grant Agreement 238803) at the Netherlands Forensic Institute in cooperation with the ATVS Biometric Recognition Group at the Universidad Autonoma de Madrid and the National Police Services Agency of the Netherlands.

REFERENCES

[1] – D. Meuwly and R.G.F. Veldhuis, *Forensic Biometrics: From two communities to One Discipline*, International Conference of the Biometrics Special Interest Group, Proceedings of the BIOSIG 2012, pp. 207 – 218, 2012
[2] – D. Meuwly, *Reconnaissance de Locuteurs en Sciences Forensiques: L’apport d’une Approche Automatique*, PhD thesis, 2001
[3] – J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro and J. Ortega-Garcia, *Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems*, Forensic Sci. Int., v. 155, n. 2-3, pp. 126-140, 2005
[4] – Amanda B. Hepler et al., *Score-based likelihood ratios for handwriting evidence*, Forensic Sci. Int., 219 (1-3): pp.129-40, 2012
[5] – D. Ramos, *Forensic Evaluation of the Evidence using Automatic Speaker Identification System*, Doctoral Thesis, 2007
[6] – D. Meuwly, *Forensic Individualization from Biometric Data*, Science & Justice, v. 46, pp. 205 – 213, 2006
[7] – N. Egli et al, *Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – Modelling within finger variability*, Forensic Sci. Int., v. 167, pp. 189 – 195, 2007

[8] – J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano and 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, v. 15, n. 7, pp. 2104 – 2115, 2007
[9]– Ch. Brenner, *Fundamental problem of forensic mathematics-The evidential value of a rare halotype*, Forensic Sci. Int.: Genetics 4, pp. 281-291, 2010
[10] – E. T. Jaynes, *Probability Theory: The Logic of Science*, ch. 18, WILEY, 1994
[11] – S. L. Zabell, *The Rule of Succession*, Erkenntnis v. 31, n. 2/3, pp. 283-321, 1989
[12] – W. E. Johnson, *The Logical Foundations of Science*, Cambridge University Press, Logic, Part III, 1924
[13] – C. M. Rodriguez, A. de Jongh, D. Meuwly, *Introducing a semi-automated method to simulate a large number of forensic fingermarks for research on fingerprint identification*, Journal of Forensic Sciences, 57/2, 2012
[14] – R. Haraksim, D. Meuwly, D. Ramos, C. Berger, *Measuring coherence of computer-assisted likelihood ratio methods*, Forensic Sci. Int., Volume 249, pp. 123–132, 2015
[15] – D. Ramos, J. Gonzalez-Rodriguez, G. Zadora and C. Aitken, *Information-theoretical assessment of the performance of likelihood ratio methods*, Journal of Forensic Sciences, 2012
[16] – D. Ramos, J. Gonzales-Rodriguez, *Reliable support: measuring calibration of likelihood ratios*, Forensic Sci. Int., 2013
[17] – N. Brummer and J. du Preez, *Application Independent Evaluation of Speaker Detection*, Computer Speech and Language, 2006
[18] – A. Martin et al., *The DET Curve in Assessment of Detection Task Performance*, National Institute of Standards and Technology (NIST) Gaithersburg, MD 20899 8940; 1997
[19] – Tippett, C., V. Emerson, M. Fereday, F. Lawton, A. Richardson, L. Jones and S. Lampert, *The Evidential Value of the Comparison of Paint Flakes from Sources other than Vehicles*, Medicine, Science and the Law: pp. 61 – 65, 1974
[20] – C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, D. Meuwly, A. Bromage-Griffiths, *Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Three Minutiae*, J Forensic Sci, Vol. 51, n. 6, pp. 1255-66, 2006
[21]– Nicole Egli, *Interpretation of partial fingermarks using an automated fingerprint identification system*, Doctoral Thesis, 2009
[22] – R. Haraksim, *Validation of likelihood ratio methods used for forensic evidence evaluation: application in forensic fingerprints*, (PhD thesis), University of Twente, Enschede, 2014