On the Reproducibility and Repeatability of Likelihood Ratio in Forensics: A case study using Face Biometrics

Nik Suki, Norman Poh
Deparment of Computer Science
University of Surrey,
Guildford GU2 7XH United Kingdom
{n.niksuki,n.poh}@surrey.ac.uk

Firham M. Senan,
Nazri A. Zamani,
and M. Zaharudin A. Darus
CyberSecurity Malaysia
{firham,nazri.az,zaharudin}@cybersecurity.my

Abstract

When using biometric technology in forensic applications, it is necessary to compute a Log-likelihood Ratio (LLR) for a given piece of evidence (E) under two competing hypotheses, namely the prosecution and the defence hypotheses. Although LLR is a quantity expressing uncertainty and intuitively quantifying its uncertainty would not make sense, in practice, it is computed under a set of assumptions and methods for a given data set. Therefore, it is essential to ask how well and how repeatable and/or reproducible it is that we can estimate LLR. More specifically, it is desirable to understand the behaviour of the confidence intervals of the estimated LLR for any feasible region since any incorrect estimate may lead to possible condemnation of innocent people. To this end, we have thus tackled the estimate of LLR which is fundamentally a Bayesian concept using a frequentist approach, via bootstraping, using two LLR estimators, namely Logistic Regression (LR) and Kernel Density Estimator (KDE). The experimental results, which are based on seven face recognition systems, show that LLR does have different confidence lengths, thus highlighting that LLR cannot be estimated with the same certainty everywhere. Moreover, for the two LLR estimators investigated, we found that there is a consistent region in which any LLR value can be estimated confidently. To our best knowledge, these two findings have never been systematically reported in literature. They thus advance our understanding of LLR when used in computing the strength of biometric evidence in forensics.

1. Introduction

In the current law practice, there is an increasing need to weigh in evidence more objectively using the biometric technology for the purpose of deciding that a person was present in a crime scene or has committed a crime. Biometric modalities such as fingerprint, DNA, and speech have been used for decades, and due to the widely accepted close-circuit television (CCTV) cameras in public surveillances, face and gait biometrics are also beginning to be used [17, 7, 19].

The strength of evidence required by a court is known as the likelihood ratio [30]. This ratio compares two competing hypotheses, known as the prosecution and defence hypotheses. In this work, we are concerned with the result of this computation. Specifically, we want to know how reliable; or at what precision it is that we can compute likelihood ratio. Since likelihood ratio is estimated numerically, it is reasonable to question its repeatability and reproducibility, that is its level of uncertainty.

1.1. Bayesian or frequentist aproaches to Evidence Interpretation

In defining a probability, there are two approaches commonly practised, namely Bayesian versus frequentist. According to Hochter [18], probability in the frequentist approach is described by sampling certain processes, while in the Bayesian approach, probability is commonly used to model uncertainty or in other word the confidence of a sample.

In theory, likelihood ratio is an estimate of confidence; and therefore, intuitively, this confidence value does not require its own confidence intervals. However, in practice, since we are using data to estimate the confidence value (i.e., the likelihood ratio), the estimand can be affected by the way the data has been sampled. A well known method following the frequentist approach to quantify the uncertainty around an estimand is bootstrap sampling [14]; it relies on a large number of trials from which the data samples could have been drawn. Therefore, it is perfectly feasible to estimate the confidence intervals around the estimand i.e., the likelihood ratio. In other words, while the concept of confidence (in the sense of likelihood ratio) is derived from the Bayesian approach; to put it into practice, we estimate it using bootstrap sampling, which is a frequensist approach.

1.2. Motivations

By combining both Bayesian and frequentist approaches, we seek to answer whether or not likelihood ratio is reproducible and repeatable. According to Bureau International des Poids et Mesures (BIPM) [1], *reproducibility* means the repli-

cability of result if a measurement is taken by another person/operator; whereas *repeatability* refers to the replicability of result of the same person or examiner in another attempt (taking place after a certain time lapse with respect to the first attempt) using the same measurement conditions, i.e., the same procedure, same operator, same measuring system, same operating condition, and same location. We would like to find out the equivalence of these concepts in quantifying likelihood ratio.

Reproducibility. One way to assess the reproducibility of likelihood ratio is to consider the following scenario. Suppose that we use two commonly used likelihood ratio estimators such as Logistic Regression (LR) and Kernel Density Estimator (KDE), we would like to know if they would give consistent likelihood ratio estimates or not. If two independent estimators give consistent estimates of likelihood ratio, then the particular likelihood ratio should be more reliable, because the value is reproducible using two different algorithms.

Repeatability. The question of repeatability in our context could be answered using bootstrap sampling. Using this strategy, the same experimental conditions would be used but with one exception. The actual samples used for estimating the conditional probability constituting the prosecution or defence hypothesis are allowed to vary, which is achieved via bootstrap sampling. If a particular likelihood ratio is repeatable, we would expect the value to remain the same within acceptable confidence intervals despite change in the constituent samples.

Therefore, in order to proceed to answering the above questions, it is necessary to use two estimators; e.g., LR and KDE to be use in this study, in conjunction with a bootstrap sampling procedure.

1.3. Related works

Taroni *et al*. [31] in their study argue on the way evidential value is presented in a court of law. The authors believe that it should be presented in a form of single value derived from Bayesian Factor (BF) rather than an expression based on a distribution over a range of values.

In BF, two competing hypothesises are considered in computing the strength of evidence, namely:

- *Prosecution hypothesis*, H_0, supports a claim that the collected evidence belongs to the suspect, and
- *Defence hypothesis*, H_1, supports a claim that the evidence belongs to someone else.

Given a piece of evidence, E, the Bayes rule suggests that the verdict should be made based on the inference of the posterior probability $P(H_k|E) \propto p(E|H_k)P(H_k)$ for $k \in \{0, 1\}$, which is a consequence of the product rule (the notation of which is to be further explained in the Methodology section). Computing the ratio between $P(H_0|E)$ and $P(H_1|E)$, we have:

$$\underbrace{\frac{P(H_0|E)}{P(H_1|E)}}_{\text{posterior odd}} = \underbrace{\frac{p(E|H_0)}{p(E|H_1)}}_{\text{likelihood ratio or BF}} \times \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{prior odd}} \quad (1)$$

This formulation provides a mathematical representation of the legal process, namely one updates the belief of a suspect being guilty or innocent as evidence is presented [12].

In addition, this formulation limits a forensic expert to weigh in his/her contribution explicitly in the form of the strength of evidence in terms of likelihood ratio, which is based on the forensic evaluation on the evidence rather than the *prior odd ratio* which is the jury's remit. In the context of forensic evidence evaluation, the main focus is to determine the source of the evidence collected at the crime scene for the purposes in the court of law.

Unlike the usual application of a biometrics system, when a biometric sample from a suspect is compared to a piece of biometric evidence collected from the crime scene, the resultant similarity score is not acceptable to be presented in the court of law. Whilst an accept or reject decision can be made in a typical biometric authentication system, such a prescriptive approach is not suitable for forensic applications because the decision is made by the judge; and not forensic practitioners [11, 10, 15]. With that reasoning, in this study, we only work on the likelihood ratio and leave the decision making process to the court of law.

Tauseef *et al*. [5] compare three algorithms to estimating likelihood ratio, namely Logistic Regression, Kernel Density Estimator, and Pull-adjacent Violator algorithms. In each case, they show that the likelihood ratio estimates can be inconsistent with each other, i.e., giving different values.

Despite this finding, it is not clear whether or not the estimated likelihood ratio values are inconsistent *everywhere* in the biometric score space. Our study differs from Tauseef *et al*.'s one in that we are interested to understand the reproducibility of likelihood ratio despite using two different estimators. Our conjecture is that likelihood ratio in some regions of the biometric score space may indeed be consistent. Therefore, the likelihood ratio values in these regions are more credible than regions where the two estimators depart in their estimates.

1.4. Our Contributions

The main contribution in this paper is to provide a better understanding about reproducibility and repeatability of likelihood ratio. To investigate the reproducibility of likelihood ratio, we shall use two common likelihood ratio estimators, namely, Kernel Density Estimator (KDE) and Logistic Regression (LR). To evaluate the repeatability of likelihood ratio, we appeal to bootstrap sampling. In terms of methodology, we advocate the combined use of both Bayesian and frequentist approaches rather than treating both as separate, competing methodologies. We have studied the behaviour of likelihood

ratio using seven face biometric score sets in order to stimulate as close as possible to real forensic conditions.

Although bootstrap sampling has been used by [3, 14] for estimating the uncertainty of likelihood ratio, our work differs in two ways. First, these prior works did not consider the effect of multiple likelihood ratio estimators, and the application domains are different. Second, we measure the confidence intervals of the likelihood ratio by using bootstrap sampling in order to evaluate its uncertainty.

In addition, although our work echoes those of existing literature [27, 13, 3, 21, 4], our study also shows that likelihood ratio has its own uncertainties, in terms of confidence intervals [32].

More importantly, we observe that the confidence intervals of likelihood ratio do not have equal length everywhere. Indeed, the likelihood ratio precision is higher (hence lower variability) when a piece of evidence (as represented by a biometric score output) is near its optimal threshold, that is, near the Equal Error Rate (ERR) operating point. When the evidence (biometric score output) is further away, moving closer to the defence or prosecution hypotheses, i.e, toward either end of the score spectrum, the confidence intervals grow larger.

We have limited our scope of study to the face biometrics and assessment at the score level rather than the general, feature level assessment, using information theory as in [2], for instance. Computing likelihood ratio in the high dimensional feature space, even with dimensionality reduction such as Principal Component Analysis (PCA) or Fisher Discriminant Analysis (FDA), remains a challenging feat, especially on a individual suspect basis where the *relevant* training samples are barely adequate. Furthermore, the state-of-the-art face recognition techniques are not based on PCA or FDA but advanced machine learning algorithms. For this reason, it would seem reasonable to investigate scores derived from a state-of-the-art face recognition classifier and then investigate how likelihood ratio can be derived in this domain.

Last but not least, we do not examine the properties of the accuracy of different likelihood ratio estimators since comparative works have been investigated in [5].

1.5. Paper Organisation

This paper is organised as follow: Section 2 presents the methodology; Section 3 show our experimental approach and results; and finally, the conclusions and future works are found in Section 4.

2. Methodology

This section presents the likelihood ratio framework, which is the currently-acceptable practice for presenting evidence in court. Section 2.1 establishes the commonly-used forensic terms and this is followed by the framework itself.

2.1. Terminology and notation

This section begins by first defining the terminology used in forensics, followed by those used by biometrics. This is necessary because the biometric technology is used to derive the strength of evidence in solving forensic cases.

2.1.1 Forensic terminology

Same vs. different source. In an automatic biometric recognition system, two biometric samples are compared to each other in order to determine whether the two samples belong to the *same source* or two *different sources*. The term "source" here refers to the subject identity.
Prosecution vs. defence. As discussed in Subparagraph 1.3.
Evidence. In the usual forensic sense, a piece of evidence refers to an acquired piece of information, generally collected from a crime scene.
Evidence score. We introduced the term "evidence score" to refer to the scores derived from a biometric matcher, which is a result of comparing two samples. An evidence score is denoted as $E \in \mathbb{R}$.

2.1.2 Biometrics terminology

Match vs. nonmatch comparison The equivalence of the prosecution hypothesis in biometrics is known as the match comparison, whereas the equivalence of the defence hypothesis is known as the non-match comparison. Indeed, the match scores, once generated, are used to estimate the distribution of the evidence when the prosecution hypothesis is true. Similarly, the non-match scores are used to estimate the distribution of the evidence under the defence hypothesis.
Score sets. Let $y \in \mathcal{Y}$ be the comparison score from the score set $\mathcal{Y}$. Since there are two comparison events due to the same versus different sources, we further distinguish two sets of scores; namely, $\mathcal{Y}^0$ under which the prosecution hypothesis (H_0) is true; and, $\mathcal{Y}^1$, where the defence hypothesis (H_1) is true.

2.2. The likelihood ratio framework

Having established the notation in the previous section, we can now revisit the likelihood framework. Let $p(E|H_k)$ represent the likelihood of the evidence score, E, given that the hypothesis H_k is true. The strength of evidence is defined as a ratio between the prosecution hypothesis and the defence hypothesis, also known as likelihood ratio or Bayesian Factor, as shown in (1). However, for computational reason, it is more convenient to do the calculation in the logarithmic domain, leading to the Log-likelihood Ratio, henceforth denoted as $LLR(E)$:

$$LLR(E) = \log \frac{p(E|H_0)}{p(E|H_1)}. \quad (2)$$

Hereinafter, the prior odd term is not discussed as this is the jury's remit.

A number of score-to-likelihood methods, also known as calibration methods, can be found in the literature, namely, Kernel Density Estimation (KDE), Logistic Regression (LR), Histogram Binning, and Pool Adjacent Violators (PAV). In this paper we shall include only two methods in order to demonstrate the reproducibility of likelihood ratio. So, we have chosen the two most commonly used likelihood ratio estimators as reported in [3, 26, 28, 16, 9]. These methods are KDE and LR.

2.3. Kernel Density Estimator (KDE)

The KDE approach directly estimates the likelihood term $p(E|H_k)$ using the KDE algorithm, which is a non-parametric method. This approach is suitable given sufficient a sample size. KDE places a kernel on each data point so that the likelihood evaluated on a given location is a sum of the potentials of the kernel function defined by all the training samples. For our purpose in this paper, it suffices to formalise KDE as an estimator of the form:

$$p(E|H_k) \simeq f(E|\mathcal{Y}^k)$$

which is dependent on the training score set $\mathcal{Y}^k$. The approximated $LLR(E)$ by KDE is therefore,

$$LLR_{KDE}(E) = \log \frac{f(E|\mathcal{Y}^0)}{f(E|\mathcal{Y}^1)}.$$

Because of the use of KDE, the estimate of likelihood on locations where data samples are sparse, e.g., at the extreme tails of the distribution, can be very inaccurate. For this reason, another approach to modelling $LLR(E)$ is by using Logistic Regression (LR), which is described next.

2.4. Logistic Regression (LR)

LR is a commonly used pattern recognition technique for many problems including fusion and calibration [8, 23, 16, 9, 5]. For our purpose here, it is used to directly model the Log-likelihood Ratio, $LLR(E)$ without estimating $p(E|H_k)$ for both hypotheses. LR gives the posterior probability of H_0 given E, i.e., $P(H_0|E)$, which as the following form:

$$P(H_0|E) = \frac{1}{1+\exp(-g(E))} \tag{3}$$

It is a sigmoid function taking the argument $g(E)$ which, in turn, is a linear function of E:

$$g(E) = w_1 E + w_0$$

In order to turn the output of LR into the form that is compatible with $LLR(E)$, we can apply the logit function, which is defined as $\log \frac{P}{1-P}$ when P is a probability. Applying logit to $P(H_0|E)$ gives:

$$LLR_{LR}(E) = \log \frac{P(H_0|E)}{1-P(H_0|E)} \tag{4}$$

Input : Number of bootstraps, B
Score sets, $(\mathcal{Y}^0, \mathcal{Y}^1)$
Output : $\{(\mathcal{Y}^0_b, \mathcal{Y}^1_b)|\forall_{b\in\mathcal{B}}\}$
for $b = 1 \ldots \mathcal{B}$ **do**
  $\mathcal{Y}^0_b = bootstrap(\mathcal{Y}^0)$
  $\mathcal{Y}^1_b = bootstrap(\mathcal{Y}^1)$
end

Algorithm 1: Bootstrap sampling

If we plug (3) into (4) and rearrange the term, we obtain:

$$LLR_{LR}(E) = g(E) = w_1 E + w_0. \tag{5}$$

This implies that $LLR(E)$ can be approximated by taking the raw output of LR, $g(E)$, that is, without applying the sigmoid function. As a result, the $LLR(E)$ as given by LR is simply a linear function of E, that is scaled by w_1 and shifted by w_0. Such an interpretation means that LR is a linear function of the raw biometric matcher output. In comparison, KDE can be viewed as a non-linear transformation of the raw biometric matcher output. For the purpose of subsequent discussion, we shall abstract the LLR estimated by LR using the following equation:

$$LLR_{LR}(E) = f_{LR}(E|\mathcal{Y}_0, \mathcal{Y}_1)$$

in order to highlight the fact that LR needs the training score sets $\mathcal{Y}_0$ and $\mathcal{Y}_1$. After training, we only need to keep w_1 and w_0 because these two parameters are all that is required to calculate LLR.

It should be cautioned that LR may inadvertently model the prior probability of the training data. This situation is particular acute with unbalanced training samples, i.e., $|\mathcal{Y}_0| \ll |\mathcal{Y}_1|$. This can be mitigated during the optimisation process by ensuring that each sample in $\mathcal{Y}_0$ has an associated weight contribution of $1/|\mathcal{Y}_k|$ for both data sets $k \in \{0, 1\}$.

2.5. Bootstrap sampling

A key ingredient to deriving the confidence intervals of LLR is to be able to bootstrap samples. If $y \in \mathcal{Y}$ is a sample drawn from a set $\mathcal{Y}$, bootstrap sampling *with replacement* generates another set of samples of the same size. Let us call this procedure bootstrap, which can be defined as: $\mathcal{Y}' = bootstrap(\mathcal{Y})$. Since, we need to create B bootstraps, this procedure has to be repeated B times:

$$\mathcal{Y}_b = bootstrap(\mathcal{Y}) \tag{6}$$

for $b = \{1 \ldots B\}$.

The bootstrap sampling simply takes each of the score set pairs $(\mathcal{Y}^0, \mathcal{Y}^1)$ and applies the bootstrapping procedure, and doing so B times, as described in Algorithm 1. This results in B pairs of bootstrapped score sets.

Input : Number of bootstraps, B
Score sets, $\{(\mathcal{Y}_b^0, \mathcal{Y}_b^1)|\forall_{b\in\mathcal{B}}\}$
Output : $\{\theta_b|\forall_{b\in\mathcal{B}}\}$
for $b = 1 \ldots \mathcal{B}$ **do**
  $\theta_b = train(\mathcal{Y}_b^0, \mathcal{Y}_b^1)$
end

Algorithm 2: Training procedure

Input : Trained parameters, $\{\theta_b|\forall_{b\in\mathcal{B}}\}$
Evidence score, E
Confidence level, α
Output : Confidence intervals, $LLR_{lower}(E), LLR_{upper}(E)$
for $b = 1 \ldots \mathcal{B}$ **do**
  $LLR^{(b)}(E) = inference(E|\theta_b)$
end
$LLR_{lower}(E)) = \arg\min_{LLR^{(b)}} |Prob(LLR^{(b)}) - \alpha|$
$LLR_{upper}(E)) = \arg\min_{LLR^{(b)}} |Prob(LLR^{(b)}) - (1-\alpha)|$

Algorithm 3: Inference procedure

2.6. Training and Inference based on the Bootstrapped Score Pairs

Let us know turn our attention to dealing with the pairs of score sets $\{(\mathcal{Y}_b^0, \mathcal{Y}_b^1)|\forall_{b\in\mathcal{B}}\}$ which are output of sample-level bootstrapping as discussed.

We shall introduce two abstract functions, namely,

1. Training procedure, i.e., $train : \mathcal{Y}_b^0, \mathcal{Y}_b^1 \rightarrow \theta_b$
2. Inference procedure, i.e., $inference : \theta_b, E \rightarrow LLR^{(b)}(E)$

Training for logistic regression The training procedure takes a pair of score sets and produce a model parameter θ. The training procedure for logistic regression is known as "gradient ascend", which in our case, produces $\theta = [\omega_0, \omega_1]$.

Training for KDE For kernel density estimator, the training involves fitting the KDE model to $\mathcal{Y}_b^k$, for each k and each b:

$$\theta_b^0 = KDE_{train}(\mathcal{Y}_b^0) \quad (7)$$
$$\theta_b^1 = KDE_{train}(\mathcal{Y}_b^1) \quad (8)$$

Inference for logistic regression During inference, for a given E, we repeat this process B times, i.e.,

$$LLR^{(b)}(E) = LLR_{LR}(E|\theta_b), \quad (9)$$

for all $b \in \mathcal{B}$. Refer to (5) for the actual function.

Inference for KDE The inference of KDE is done by repeating the process B times for a given E:

$$LLR_{KDE}(E) = \log \frac{f(E|\theta_b^0)}{f(E|\theta_b^1)} \quad (10)$$

for all $b \in \mathcal{B}$, recalling that f is the KDE function (refer to [6], for instance).

Overall training procedure The overall training procedure first takes B pairs of score sets and produces B model parameters which are stored for inference use later on.

Overall inference procedure The inference procedure derives a set of $LLR(E)$ values from B model parameters. This results in B LLR values from which the confidence intervals of $LLR(E)$ can be derived. The confidence intervals of $LLR(E)$, at the α confidence level, is then given by the values of $\{LLR^{(b)}(E)|\forall_b\}$ at their corresponding $Prob(LLR^{(b)}) \approx \alpha$ and $Prob(LLR^{(b)}) \approx 1-\alpha$ which respectively delineate the lower and uppper confidence intervals, where $Prob(\cdot)$ is a cumulative density function of $\cdot$. The inference procedure is summarised in Algorithm 3.

In short, using each of the two LLR estimators – LR and KDE – in turn, we can invoke Algorithms 2 and 3 in order to study the reproducibility and repeatability properties of likelihood ratio, the experiments of which will be presented next.

3. Experiments and Results

3.1. Database

In order to compare the two variants of methods to estimating confidence intervals of Log-Likelihood Ratio (LLR), ideally, one should use a real forensic database. Unfortunately, due to information governance and privacy issues, real forensic databases are either scant or contain few cases that enable any meaningful evaluation of the proposed model. As a result, following previous studies on LLR computation, e.g., [5, 15] and many authors, we shall use the output of biometric experiments. To this end, we have chosen to use a database of scores taken from experiments carried out on the XM2VTS database [20]. The database contains genuine and impostor scores of seven face systems, across two experimental protocols, known as Lausanne Protocols 1 and 2. The two protocols differ mainly in the way the development (training) data is partitioned to build the baseline systems. The evaluation (test) data in both protocols *remain the same*.

The face system considered in this study is based on the Discrete Cosine Transform (DCT) coefficients [29]. The DCT procedure operates with two image block dimensions, i.e., small (s) or big (b), and is denoted by DCTs or DCTb, respectively. Hence, the matching process is local as opposed to a holistic matching approach such as the Principal Component Analysis.

Table 1 presents a list of baseline experimental scores used in this study. The score data set is publicly available at `http://goo.gl/CdXw9Z` and was reported in [24]. Note that each system can be characterized by a feature representation scheme and a classifier. Two types of classifiers were used, i.e., GMMs and multi-layer Perceptrons (MLPs).

For Lausanne Protocol 1 (LP1), there are 3 genuine scores per subject and there are 200 subjects. Therefore, there are

Table 1. The 13 baseline experiments taken from the XM2VTS benchmark fusion database were considered for studying the user-specific statistics as well as the proposed OR-switcher fusion operator.

Label	Modality	Feature	Classifier
P1:1	face	DCTs	GMM
P1:2	face	DCTb	GMM
P1:6	face	DCTs	MLP
P1:7	face	DCTs	MLPi
P1:8	face	DCTb	MLP
P1:9	face	DCTb	MLPi
P2:1	face	DCTb	GMM

Pm : n denotes the n-th system in the m-th protocol. MLPi denotes the output of MLP converted to LLR using inverse hyperbolic tangent function. P1:6 and P1:7 (resp. P1:8 and P1:9) are the *same* systems except that the scores of the latter have been transformed.

$3 \times 200 = 600$ match scores. The nonmatch score set consists of 40,000 samples which is the result of comparing 8 samples of 25 held-out impostor subjects to the templates of 200 legitimate subjects ($8 \times 25 \times 200$). For LP2, there are 400 (2×200) match scores; and exactly the same number of nonmatch scores for training. Both protocols share the same test set, i.e., $2 \times 200 = 400$ genuine scores and $70 \times 8 \times 200 =$ 112,000 nonmatch scores. However, for the purpose of estimating the length of 'confidence intervals' in LLR, we shall not use the test scores. This is because we are only interested in measuring the confidence intervals around a given LLR value in the score space. So, the test scores serve no purpose as far as the experiment is concerned. More details can be found in [24].

3.2. Experimental hypotheses

From the analyses we have so far, we know that logistic regression, as an LLR estimator, is linear in the score space, as shown by (5). On the contrary, for KDE, the function form as shown by (2) does not provide any guarantee of this. There are, however, two questions that have not yet been answered, which must be evaluated experimentally using real biometric matcher output:

1. Are the estimates of LR and KDE inconsistent everywhere?

2. How does the length of confidence intervals behave across the different E values?

3.3. Results

To answer the first question, we first compare the LLR produced by LR and KDE across the seven face biometric score data sets in XM2VTS database. The curve is produced by setting $E \in \mathcal{Y}$ to a particular value, and then evaluating its confidence intervals using the inference procedure described by Algorithm 3. Throughout the experiments reported here, we shall use 100 bootstrap samples, thus, setting $B = 100$.

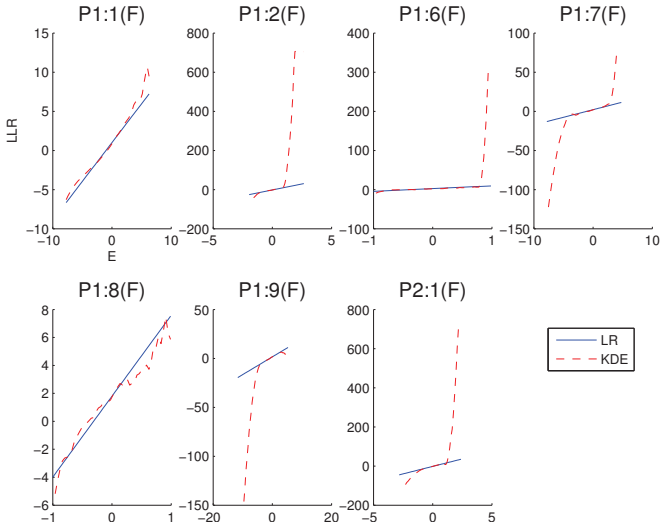

Figure 1. Comparison of LLR estimated using logistic regression (LR, plotted in blue continuous line) and Kernel Density Estimation (KDE, red dashed line) for 7 face biometric systems.

In order to highlight that the two LLR estimators are often consistent with each other, we plot their estimated LLR values as a function of E, for a particular instance of bootstrap, in Figure 1. We can observe that the estimated LLR values are often in agreement with each other, especially around the matching score E near the optimal decision threshold (which happens to be around zero). However, toward the extreme positive or negative ends of E, the estimated LLR for KDE almost always deviates from that of logistic regression.

We can define the *reproducible* LLR values to be those falling in the biometric score space (E) where the two likelihood ratio estimators are consistent. This *reproducible region* can be formally characterised by the following equation for a small ϵ value:

$$\mathcal{Y}_{reproducible} \equiv \{|LLR_{KDE}(E) - LLR_{LR}(E)| < \epsilon, E \in \mathcal{Y}\}$$

Figure 2 shows the absolute difference of LLR in the two LLR estimators, across all bootstrapped samples for each face system. Based on this experiment, we suggest that the error tolerance threshold, ϵ, with a value of about 10, to be somewhat reasonable since it can clearly distinguish a consistent LLR region that is continuous from both the extreme ends of the LLR values where inconsistency exist, across all experiments.

To answer the second question, we shall make use of the bootstrap sampling procedure. The estimated confidence intervals for LR and KDE are shown in Figures 3 and 4, respectively. In both figures, The X-axis shows the index of the biometric score space sampled equally from the negative and positive extreme values of the matcher output. Only the index of these ten sampled values are shown instead of the original matching output, which is not important here in the context of this study here. The index around 5 or 6 is the location where false acceptance or false rejection rates are similar. This

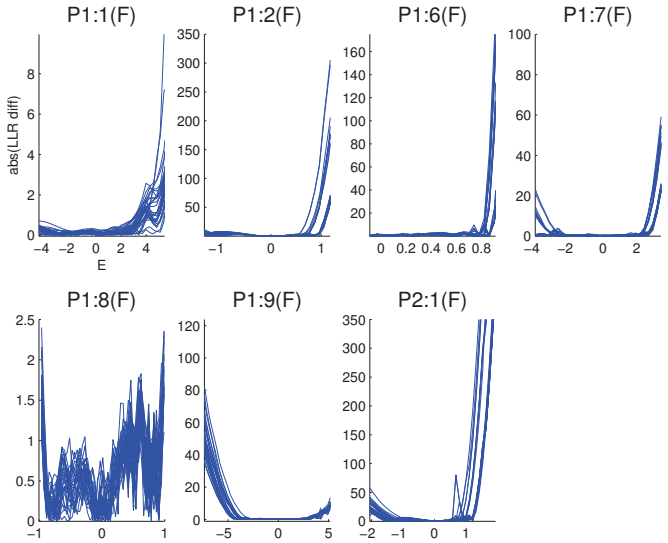

Figure 2. The absolute difference in LLR between LR and KDE for 7 face biometric systems. Each curve in a subplot corresponds to an instance of the bootstrap-sampled absolute LLR difference curve.

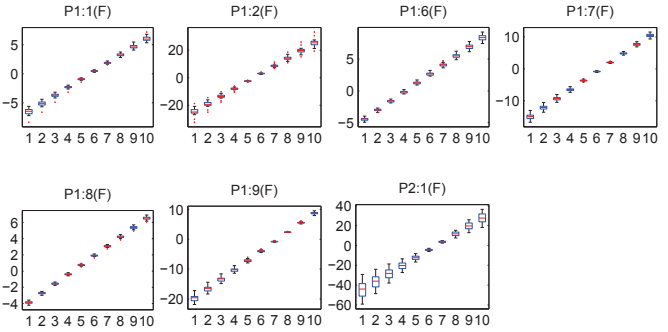

Figure 3. The confidence intervals of Logistic Regression (LR) on seven face biometric systems.

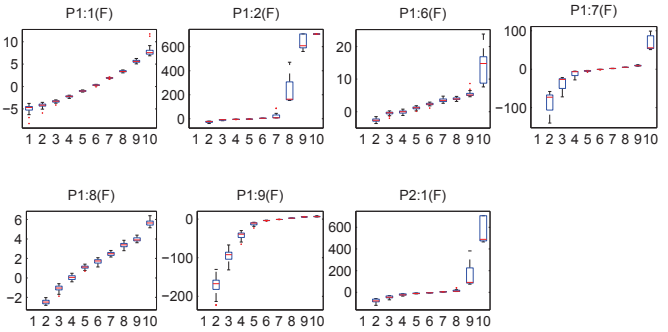

Figure 4. The confidence intervals of Kernel Density Estimation (KDE) on seven face biometric systems.

is where optimal threshold would have been placed if the biometric system is used for authentication.

It is interesting to note that for both KDE and LR, the length of confidence intervals, as shown in the form of boxplot, are not of equal length everywhere. They are smaller when $LLR = 0$ (with index around 5 or 6). This is significant because when the likelihood ratio indicates that there is an equal probability between the prosecution and the defence hypothesis, the certainty to which we can estimate is value is very high, as indicated by the narrow confidence intervals of the LLR around zero values. On the other hand, when the LLR (strength of evidence) leans toward one hypothesis or another – $LLR \gg 0$ (index closer to 10) or $LLR \ll 0$ (index closer to 1) – the procedure tells us that the certainty at which we can compute this LLR is, in fact, relatively low (compared to when LLR is close to zero).

Last but not least, when comparing LR and KDE, we see that the LLR estimated by logistic regression is linear in the score space. It is also more stable and reliable compared to that estimated by KDE. This observation is consistent with [5].

4. Conclusions

In this study, we have advanced our understanding on the reproducibility and repeatability of likelihood ratio. Firstly, our finding suggests that there are regions in the biometric matcher space where the estimated likelihood ratio of logistic regression and Kernel Density Estimator can be consistent with each other. This shows that despite the use of two different algorithms to estimating likelihood ratio, there are regions in which estimated LLR values are reproducible. It is therefore recommended that the *reproducible region* of LLR be used when computing the strength of evidence. Conversely, when the LLR values of intervals is outside this reproducible region, one should interpret this value with care.

Secondly, the length of confidence intervals of LLR is not equal everywhere. The intervals are smaller when LLR is near the operating point where $EER = 0$. This is significant because when the likelihood ratio indicates that there is an equal probability between the prosecution and the defence hypothesis, the certainty to which we can estimate is value is very high. On the other hand, when the LLR (strength of evidence) leans toward one hypothesis or another – $LLR \gg 0$ or $LLR \ll 0$ – there is much discrepancy between the two LLR estimators. This is mainly due the KDE, being sensitive to small sample size, is unable to estimate the density, hence adversely impacting on its ability to accurately estimate LLR at both extreme ends of the spectrum.

Possible future research directions include: (1) using alternative framework of evidence evaluation such as Non-match Probability (NMP) [22]; and (2) investigating other bootstrap procedures such as the bootstrap subset technique which is known to provide more realistic intervals length than bootstrap sampling [25]; and (3) extending this study to other biometrics.

Acknowledgement

Nik Suki would like to acknowledge FSKKP, Universiti Malaysia Pahang (UMP), for the sponsorship on her PhD study. The authors also thank Ministry of Science, Technology and Innovation (MOSTI) through CyberSecurity Malaysia's Technofund project entitled "GPU Enhanced Robust Multi-Dimensional Facial Identification System for CCTV Evidence in Video Forensics Analysis" grant num-

ber: TF0813C268(K2).

References

[1] Joint committee for guides in metrology: International vocabulary of metrology: basic and general concepts and associated terms (vim). *JCGM 200:2008*, 2008.

[2] A. Adler, R. Youmaran, and S. Loyka. Towards a measure of biometric information. In *Electrical and Computer Engineering, 2006. CCECE'06. Canadian Conference on*, pages 210–213. IEEE, 2006.

[3] A. Alexander. Forensic automatic speaker recognition using bayesian interpretation and statistical compensation for mismatched condition. 2005.

[4] T. Ali, L. Spreeuwers, R. Veldhuis, and D. Meuwly. Effect of calibration data on forensic likelihood ratio from a face recognition system. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8, Sept 2013.

[5] T. Ali, L. J. Spreeuwers, and R. N. J. Veldhuis. A review of calibration methods for biometric systems in forensic applications. In *33rd WIC Symposium on Information Theory in the Benelux, Boekelo, Netherlands*, pages 126–133, Enschede, May 2012. WIC.

[6] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1999.

[7] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon. On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56(4):882–889, 2011.

[8] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim. Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):2072–2084, Sept 2007.

[9] N. Brmmer and J. du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(23):230 – 275, 2006. Odyssey 2004: The speaker and Language Recognition Workshop Odyssey-04 Odyssey 2004: The speaker and Language Recognition Workshop.

[10] J. Buckleton, C. Triggs, and C. Champod. An extended likelihood ratio framework for interpreting evidence. *Science & Justice*, 46(2):69 – 78, 2006.

[11] C. Champod and D. Meuwly. The inference of identity in forensic speaker recognition. *Speech Communication*, 31(23):193 – 203, 2000.

[12] J. M. Curran. Statistics in forensic science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(2):141–156, 2009.

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

[14] B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

[15] 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 Science International*, 155(23):126 – 140, 2005.

[16] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. Toledano, and J. Ortega-Garcia. Emulating dna: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):2104–2115, Sept 2007.

[17] R. Hasting. From grainy cctv to a positive id: Recognising the benefits of surveillance. *The Independent*, 2013.

[18] M. Hochster. What is the difference between bayesian and frequentist statisticians? *Available at: www.quora.com/What-is-the-difference-between-Bayesian-and-frequentist-statisticians*, 2012. [Online; accessed 04-April-2016].

[19] P. K. Larsen, E. B. Simonsen, and N. Lynnerup. Gait analysis in forensic medicine*. *Journal of Forensic Sciences*, 53(5):1149–1153, 2008.

[20] J. Luettin and G. Maitre. Evaluation protocol for the XM2VTS database. 1998.

[21] G. S. Morrison. Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2):173–197, 2013.

[22] A. Nagar, H. Choi, and A. Jain. Evidential value of automated latent fingerprint comparison: An empirical approach. *Information Forensics and Security, IEEE Transactions on*, 7(6):1752–1765, Dec 2012.

[23] S. Pigeon, P. Druyts, and P. Verlinde. Applying logistic regression to the fusion of the nist'99 1-speaker submissions. *Digital Signal Processing*, 10(13):237 – 248, 2000.

[24] N. Poh and S. Bengio. A score-level fusion benchmark database for biometric authentication. In T. Kanade, A. Jain, and N. Ratha, editors, *Audio- and Video-Based Biometric Person Authentication*, volume 3546 of *Lecture Notes in Computer Science*, pages 1059–1070. Springer Berlin Heidelberg, 2005.

[25] N. Poh, A. Martin, and S. Bengio. Performance generalization in biometric authentication using joint user-specific and sample bootstraps. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, July 2005.

[26] D. Ramos. Forensic evidence evaluation using automatic speaker recognition forensic evidence evaluation using automatic speaker recognition systems. *PhD Thesis*, 2007.

[27] D. Ramos, J. Franco-Pedroso, and J. Gonzalez-Rodriguez. Calibration and weight of the evidence by human listeners. the atvs-uam submission to nist human-aided speaker recognition 2010. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5908–5911, May 2011.

[28] D. Ramos and J. Gonzalez-Rodriguez. Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, 230(13):156 – 169, 2013. {EAFS} 2012 6th European Academy of Forensic Science Conference The Hague, 20-24 August 2012.

[29] C. Sanderson and K. Paliwal. Fast Features for Face Authentication Under Illumination Direction Changes. *Pattern Recognition Letters*, 24(14):2409–2419, 2003.

[30] S. M. Stigler. In *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap of Harvard UP, 1986.

[31] F. Taroni, S. Bozza, A. Biedermann, and C. Aitken. Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, 2015.

[32] J. Wayman, A. Possolo, and A. Mansfield. Modern statistical and philosophical framework for uncertainty assessment in biometric performance testing. *Biometrics, IET*, 2(3):85–96, September 2013.