Science and Justice xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Sampling variability in forensic likelihood-ratio computation: A simulation study

Tauseef Ali [a,*], Luuk Spreeuwers [a], Raymond Veldhuis [a], Didier Meuwly [b]

[a] Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE, Enschede, The Netherlands
[b] Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB, The Hague, The Netherlands

ARTICLE INFO

Article history:
Received 2 September 2014
Received in revised form 10 April 2015
Accepted 8 May 2015
Available online xxxx

Keywords:
Likelihood-ratio
Forensics
Biometric recognition
Score
Sampling variability

ABSTRACT

Recently, in the forensic biometric community, there is a growing interest to compute a metric called "likelihood-ratio" when a pair of biometric specimens is compared using a biometric recognition system. Generally, a biometric recognition system outputs a score and therefore a likelihood-ratio computation method is used to convert the score to a likelihood-ratio. The likelihood-ratio is the probability of the score given the hypothesis of the prosecution, H_p (the two biometric specimens arose from a same source), divided by the probability of the score given the hypothesis of the defense, H_d (the two biometric specimens arose from different sources). Given a set of training scores under H_p and a set of training scores under H_d, several methods exist to convert a score to a likelihood-ratio. In this work, we focus on the issue of sampling variability in the training sets and carry out a detailed empirical study to quantify its effect on commonly proposed likelihood-ratio computation methods. We study the effect of the sampling variability varying: 1) the shapes of the probability density functions which model the distributions of scores in the two training sets; 2) the sizes of the training sets and 3) the score for which a likelihood-ratio is computed. For this purpose, we introduce a simulation framework which can be used to study several properties of a likelihood-ratio computation method and to quantify the effect of sampling variability in the likelihood-ratio computation. It is empirically shown that the sampling variability can be considerable, particularly when the training sets are small. Furthermore, a given method of likelihood-ratio computation can behave very differently for different shapes of the probability density functions of the scores in the training sets and different scores for which likelihood-ratios are computed.

1. Introduction

For a comparison of a biometric specimen from a known source and a biometric specimen from an unknown source, a metric called *score* can be computed using a biometric recognition system

$$s = g(x, y), \tag{1}$$

where x and y are the two biometric specimens, g is the biometric recognition algorithm (feature extraction and comparison) and s is the computed score. In general, a score quantifies the similarity between the two biometric specimens. The use of biometric recognition systems in applications such as access-control to a building and e-passport gates at some airports require the developer of the system to choose a threshold and consequently any score above the threshold implies a positive decision and vice versa [1]. This strategy works well in such applications; however, it presents several issues in forensic evaluation and reporting of the evidence from biometric recognition systems [2]. In forensics, the known-source biometric specimen can, for example, come from a suspect while the unknown-source biometric specimen can, for example, come from a crime scene and the goal is to give a degree of support for H_p or H_d. The selection of a threshold and therefore making a decision are not the province of a forensic practitioner. Furthermore, in most criminal cases, scientific analysis of the two biometric specimens provides additional information about the case at hand [3] and a threshold-based hard decision cannot be optimally integrated with other evidences in the case [2–4].

1.1. Likelihood-ratio (LR)

There is a growing interest among forensic practitioners to use biometric recognition systems to compare a pair of biometric specimens. The concept of LR can be used to present the output of such a comparison. It has been extensively used for DNA evidence evaluation [5]. In general, given two biometric specimens, one with a known source and another with an unknown source, it is the joint probability of the occurrence of the two biometric specimens given H_p divided by the joint probability of the occurrence of the two biometric specimens given H_d [6–8]. When the two biometric specimens x and y, are compared

* Corresponding author at: Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Building Zilverling 4061, PO Box 217, 7500 AE Enschede, The Netherlands.
E-mail addresses: T.Ali@utwente.nl (T. Ali), L.J.Spreeuwers@utwente.nl (L. Spreeuwers), R.N.J.Veldhuis@utwente.nl (R. Veldhuis), d.meuwly@nfi.minvenj.nl (D. Meuwly).

http://dx.doi.org/10.1016/j.scijus.2015.05.003
1355-0306/

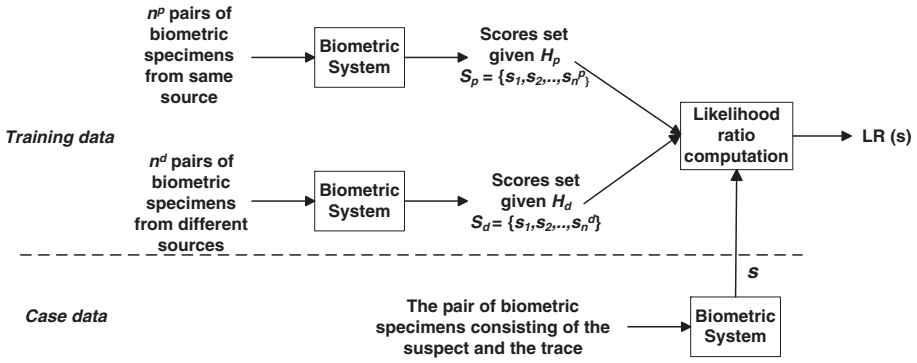

Fig. 1. Computation of a LR for a pair of biometric specimens consisting of the suspect's biometric specimen and the trace biometric specimen.

using a biometric recognition system, the resultant score replaces the joint probability of the occurrence of the two specimens in a score-based LR computation [9,10]

$$LR(x,y) = \frac{P(x,y|H_p,I)}{P(x,y|H_d,I)} = \frac{P(s|H_p,I)}{P(s|H_d,I)}, \quad (2)$$

where I refers to the background information which may or may not be domain specific. Note that here the evidence x, y is redefined into the observation s. Once a forensic practitioner has computed a LR, one way to interpret it is as a multiplicative factor which updates the prior odds (before observing the evidence from a biometric system) to the posterior odds (after observing the evidence from a biometric system) using the Bayesian theorem:

$$\frac{P(H_p|s)}{P(H_d|s)} = \underbrace{\frac{P(s|H_p)}{P(s|H_d)}}_{\text{LR}} \times \frac{P(H_p)}{P(H_d)}, \quad (3)$$

where the background information, I, is omitted for simplicity. This is an appropriate probabilistic framework where the trier of fact is responsible for quantification of the prior beliefs about H_p and H_d while the forensic practitioner is responsible for computation of the LR.

The use of a LR is gradually becoming an accepted manner to report the strength of the evidence computed by biometric recognition systems. This is a more informative, balanced and useful metric than a score for forensic evidence evaluation and reporting [3,11]. A general description of the LR concept for evidence evaluation can be found in [2,3]. It is applied to several biometric modalities including speech [12–15] and fingerprint comparison [16]. Preliminary results of the evidence evaluation using the concept of the LR in the context of face and handwriting recognition systems are presented in [6,9,17,18].

1.2. Computation of a LR

In most cases, the conditional probabilities, $P(s|H_p)$ and $P(s|H_d)$, are unknown and they are computed empirically using a set of training scores under H_p, $s_p = \left\{s_j^p\right\}_{j=1}^{n^p}$ (a set of n^p number of scores given H_p) and a set of training scores under H_d, $s_d = \left\{s_j^d\right\}_{j=1}^{n^d}$ (a set of n^d number of scores given H_d) (see Fig. 1). The s_d scores are computed by comparing pairs of biometric specimens where the two biometric specimens in each pair are obtained from different sources whereas the s_p scores are computed by comparing pairs of biometric specimens where the two biometric specimens in each pair are obtained from the same source. These sets of training scores and the corresponding hypotheses should preferably be case-specific. To compute case-specific different-source scores, either the trace or the suspect's biometric specimen can be compared to the biometric specimens of the potential population (possible potential sources of the trace biometric specimen) [9,10,15,20]. The suspect's biometric specimen used for this purpose should be taken in conditions similar to the trace. Similarly, same-source scores can be obtained by comparing trace-like biometric specimens from the suspect to the reference biometric specimens of the suspect. The effect of using generic instead of the case-specific same-source and different-source scores in the training sets on a LR is studied in the context of handwriting, fingerprint and face recognition systems [9,19,20]. An important condition in LR computation is that the pairs of biometric specimens used for training should reflect the conditions of the pair of biometric specimens for which a LR is computed. Please refer to [21] for an overview of the biometric data set collection in forensic casework for LR computation.

1.3. Sampling variability

Statistically, the training biometric data sets are samples from large populations of biometric data sets. The training biometric data sets, when resampled, lead to slightly different values of the scores in the training sets due to the unavoidable sampling variability. This implies that the sets s_p and s_d consist of random draws from large sets of scores. When the resampling is repeated, slightly different LRs are computed for a given score. This is referred to as the "sampling variability" in a LR. It is desirable that a given LR computation method is less sensitive to the sampling variability in the training sets. If the probability density functions (PDFs) of the scores in the s_p and s_d sets are known, the LR computed using the given sets of training scores can be compared with the LR computed using the two PDFs. The closeness of the two values implies the suitability of a given LR computation method and in this article, we will refer to this performance indicator as "accuracy".

Note that in a given forensic case, the potential population, the trace biometric specimen and the suspect are deterministic inputs to a LR computation procedure. The sampling variability, however, is due to the training scores that are used to compute a LR from a score. This is because these training scores are finite and would vary from one to the next in repeated random sampling. In practice, generation of multiple realizations of the sets of training scores by resampling might not be

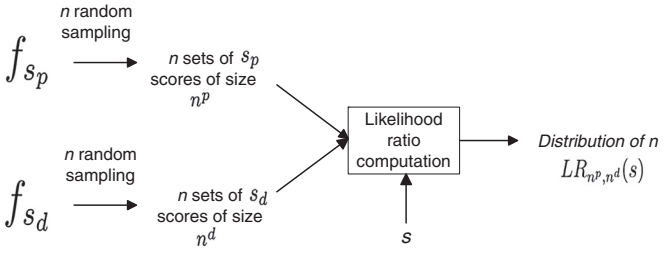

Fig. 2. Generation of n realizations of the training sets by random sampling and computation of n LRs of a given score s. The standard deviation, minimum LR, maximum LR and mean LR follow from the set of n LRs of the score s.

feasible and therefore sampling variability is often not measured. This is the motivation behind this study which provides an assessment of the sampling variability using a simulation framework. In passing, it should be pointed out that different points of view exist on how to treat sampling variability in LR computation [22–26]. This paper adapts one of the views which states that the sampling variability should be considered once a LR is computed.

1.4. Goal and organization of the paper

The purpose of this paper is to explore the sampling variability and accuracy of three commonly proposed LR computation methods. We consider four different shapes of the PDFs of the scores in the training sets, three sizes of the training sets and vary the score for which a LR is computed. A detailed empirical study is carried out in order to understand the merits and demerits of each LR computation method for evidence evaluation.

The rest of the paper is organized as follows. In Section 2, we review existing work on comparison and assessment of different LR computation methods and describe our simulation framework. Section 3 briefly reviews the LR computation methods compared in this paper. Section 4 explains the experimental setup by describing the sets of biometric scores that are used, selection of the PDFs of the scores in the training sets and the proposed strategy to avoid zero and infinite values in LR computation. Results are presented in Section 5 where conclusions and future research directions are stated in Section 6.

2. Comparison of LR computation methods

2.1. Existing work

The training sets, s_p and s_d, can be divided into two subsets so that we have a training subset $\{\{s_{p\_tr}\},\{s_{d\_tr}\}\}$ and a testing subset $\{\{s_{p\_ts}\},\{s_{d\_ts}\}\}$. The most common approach to develop and assess LR computation methods is to learn the mapping function from score to LR using $\{\{s_{p\_tr}\},\{s_{d\_tr}\}\}$. The scores in $\{\{s_{p\_ts}\},\{s_{d\_ts}\}\}$ are used to compute a set of test LRs, $\{\{LR_{p\_ts}\},\{LR_{d\_ts}\}\}$, for performance assessment. We expect large LRs in $\{LR_{p\_ts}\}$ because they are computed for the pairs of biometric specimens obtained from a same source and small LRs in $\{LR_{d\_ts}\}$ because they are computed for the pairs of biometric specimens obtained from different sources. Based on this argument, performance of a LR computation method based on the set of test LRs, $\{\{LR_{p\_ts}\},\{LR_{d\_ts}\}\}$, can be measured in a number of ways including calculation of the rate of misleading evidence in favor of H_p and H_d [9], Tippett plot [27], *Cost of Log LR* (C_{llr}) [28,29] and Empirical Cross Entropy (ECE) plot [30]. This is a 'black-box' approach where performance is assessed using a set of test LRs, $\{\{LR_{p\_ts}\},\{LR_{d\_ts}\}\}$. Such assessments are used extensively and prove useful in practice, however, they do not consider the effect of the sampling variability.

Morrison [31] investigated the sampling variability of a LR that is caused by the sampling variability in the score for which a LR is computed. However, this work did not consider the sampling variability in the training sets. In [24], a strategy based on confidence intervals is used to measure the sampling variability when the frequencies associated with LR computation in DNA evidence are estimated using a sample from the population. For forensic speaker recognition, Alexander [32, Section 3.5] used the "Leave-one-out" strategy in order to assess the variability in a LR due to small change in the training sets. One score is removed per run, with replacement, from the training scores and the LR is computed. This leads to a distribution of LRs for a given score which can be used to assess the variability. However, this procedure does not address the issue of sampling variability where the assumption is that the training sets are samples from populations.

In this work, we carry out a general study, varying the shapes of the PDFs of the scores in the training sets, the sizes of the training sets and the score for which a LR is computed in order to understand their effects on the sampling variability and accuracy of LR computation methods. An analysis of the sampling variability constitutes a measure of reliability of a LR computation method and is important in forensic science as recently discussed by Morrison [22,33].

2.2. Proposed simulation framework

Since the score, output by most biometric systems, is a continuous random variable, computation of a LR ideally requires the PDF of scores in the s_p and s_d sets. However, in practice, these PDFs are not known and the scores in the s_p and s_d sets give a rough idea of what the corresponding PDFs look like. Suppose we have access to the underlying PDFs from which the training sets s_p and s_d are generated, the LR of a score s using the PDFs is computed as:

$$LR_{\infty,\infty}(s) = \frac{f_{s_p}(s)}{f_{s_d}(s)}, \tag{4}$$

where f_{s_p} is the PDF of the s_p scores and f_{s_d} is the PDF of the s_d scores. For ease of presentation, we introduce the notation $LR_{n^p,n^d}(s)$ to represent the LR of a given score s computed using an n^p number of the s_p scores and n^d number of the s_d scores. Using the LR computed from the PDFs as a benchmark, a form of accuracy, based on how close $LR_{n^p,n^d}(s)$ and $LR_{\infty,\infty}(s)$ are, can be measured.[1] Our procedure to measure the sampling variability and the accuracy can be summarized as follows:

- Match, using Maximum Likelihood Estimation (MLE), standard PDFs to large s_p and s_d sets obtained from a biometric recognition system or assume standard PDFs which are similar to commonly observed distributions of the scores in the s_p and the s_d sets.
- Generate n realizations of the s_p and the s_d sets from these standard PDFs by random statistical sampling.
- For a given score, n LRs can be computed using the n random realizations of $\{\{s_p\},\{s_d\}\}$. In each random realization, there is a set of s_p score and a set of s_d scores and these two sets combines to produce a LR.
- The standard deviation and the difference between the minimum and the maximum values of the set of n LRs of a score is used to measure the sampling variability while the mean value of the n LRs of a score along with the $LR_{\infty,\infty}$ of the score is used to measure the accuracy (see Fig. 2).
- A smaller value of the standard deviation and a smaller value of the difference between the maximum and the minimum values imply that the method is less sensitive to the sampling variability in the training sets. Similarly, the closer the mean value of the n LRs and the $LR_{\infty,\infty}$ of a score are, the more accurate a method is. Furthermore, using the mean value, a bias value can be computed as: $bias = LR_{\infty,\infty} - mean$.

Note that random statistical sampling of the two PDFs simulates the random sampling of the biometric data sets from their sources.

It can be argued that the proposed experiment cannot be performed for an exhaustive set of biometric systems or shapes of the distributions of scores. However, based on a few typical PDFs, it can provide a useful insight into the behavior of different methods of LR computation and in the assessment of the sampling variability. For the purpose of performance assessment and comparison of different methods, this procedure has several advantages:

- The benchmark value of the LR of a given score s, $LR_{\infty,\infty}(s)$, is known and can be used to measure the accuracy of different methods of LR computation.

[1] The term "accuracy" is also sometimes used to refer to the performance evaluated using the tools mentioned in Section 2.1 (Tippett plot, C_{llr}, etc.). That measure of accuracy is based on the ground truth information whether a pair of biometric specimens are obtained from the same source or from different sources as the benchmark. The measure of accuracy of interest in this paper is based on the knowledge of the true parameters of the PDFs from which the training sets are generated and uses the LRs obtained from the assumed PDFs as a benchmark.

Please cite this article as: T. Ali, et al., Sampling variability in forensic likelihood-ratio computation: A simulation study, Sci. Justice (2015), http://dx.doi.org/10.1016/j.scijus.2015.05.003

- Multiple realizations of the training sets can be generated by repeated random sampling of the PDFs. These sets simulate the sampling variability when the biometric data sets are repeatedly resampled for computation of the training sets.
- The sizes of the training sets can be increased or decreased easily to quantify its effect on LR computation methods.
- The characteristics of the PDFs can be altered to see how the shapes of the distributions of the scores in the training sets affect a LR computation method.
- The separation between the assumed PDFs can be increased or decreased which is related to the discriminating power of a biometric recognition system and affects different LR computation methods differently.

The choice of the types and the parameters of these PDFs is critical. We consider four pairs of PDFs; two pairs of PDFs are selected based on their best MLE fitting to the s_p and s_d sets of two biometric recognition systems whereas the other two pairs of PDFs are assumed based on their general proximity to the shapes of the distributions of the s_p and s_d sets in the available literature. With this, we cover a variety of the distributions of scores expected from different biometric recognition systems. The details of the selected PDFs and the fitting procedure will follow in a later section.

3. LR computation methods

We consider three LR computation methods commonly proposed for the evaluation of evidence. A brief description of these methods is given in the following. MATLAB scripts along with the biometric scores sets are available online at http://scs.ewi.utwente.nl/other/code_alit/.

Table 1
Parameters of the assumed Normal PDFs.

	μ	σ
f_{s_p}	23	2.3
f_{s_d}	13	4

3.1. Kernel Density Estimation (KDE)

This approach estimates the PDFs of the s_p and the s_d scores using KDE [34]. The estimated PDF of the s_p is divided by the estimated PDF of the s_d at the score location s to compute a LR of s. KDE smooths out the contribution of each observed data point over a local neighborhood. The contribution of a score s_i in the training set to the estimate at score location s depends on how far apart s_i and s are. The extent of this contribution is based on the shape and width of the kernel function. If we denote the kernel function as K and its width by h, the estimated density at score s is

$$f(s) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{(s-s_i)}{h}\right), \tag{5}$$

where for PDF of the s_p scores, n is the total number of scores in the s_p set and for PDF of the s_d scores, n is the total number of scores in the s_d set. In our experiments, a Gaussian kernel is used where the width is optimally chosen as [35]:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right), \tag{6}$$

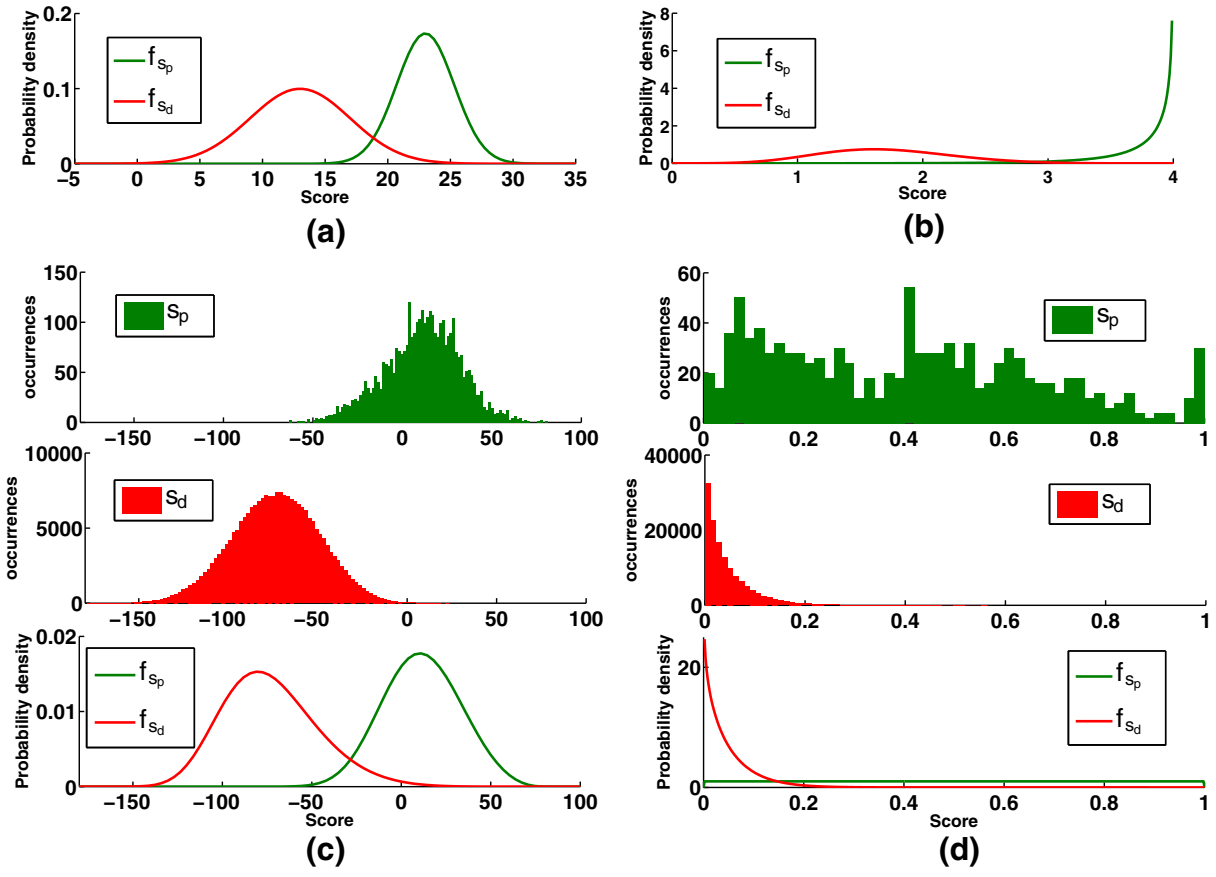

Fig. 3. Pairs of PDFs from which the n realizations of the training sets are generated by random sampling. (a) Assumed Normal PDFs. (b) Assumed reversed Weibull PDFs. (c) Score sets from the speaker recognition system and the fitted reversed Weibull PDFs. (d) Score sets from the Cognitec face recognition system and the fitted Uniform and Beta PDFs.

Please cite this article as: T. Ali, et al., Sampling variability in forensic likelihood-ratio computation: A simulation study, Sci. Justice (2015), http://dx.doi.org/10.1016/j.scijus.2015.05.003

Table 2
Parameters of the assumed Weibull PDFs.

	k	λ
f_{s_p}	0.7	0.2
f_{s_d}	5	2.5

Table 3
Parameters of the Weibull PDFs fitted to the s_p and s_d sets of the speaker recognition system shown in Fig. 3(c).

	k	λ
f_{s_p}	3.59	77.67
f_{s_d}	6.80	165.29

where $\hat{\sigma}$ is the sample standard deviation. A detailed description of this approach to LR computation is presented in [36] in the context of forensic speaker identification and subsequently adapted by [17,18,37] for other biometric modalities.

3.2. Logistic Regression (Log Reg)

Log Reg [38] uses a linear or a quadratic function to map a score s to $log\left(\frac{P(H_p|s)}{1-P(H_p|s)}\right)$. These log odds along with the sizes of the training sets can be used in the Bayesian formula in Eq. (3) to compute a LR. We choose a linear function since it is more common in forensic likelihood-ratio computation [28,38]:

$$log\left(\frac{P(H_p|s)}{1-P(H_p|s)}\right) = \beta_0 + s\beta_1. \quad (7)$$

Parameters β_0 and β_1 are found from the sets of training scores using Iteratively Reweighted Least Squares (IRLS) algorithm [39]. Log Reg is a well-known algorithm in machine learning and statistics and is widely used for LR computation in several biometric modalities including forensic speaker and fingerprint comparison [15,28,38].

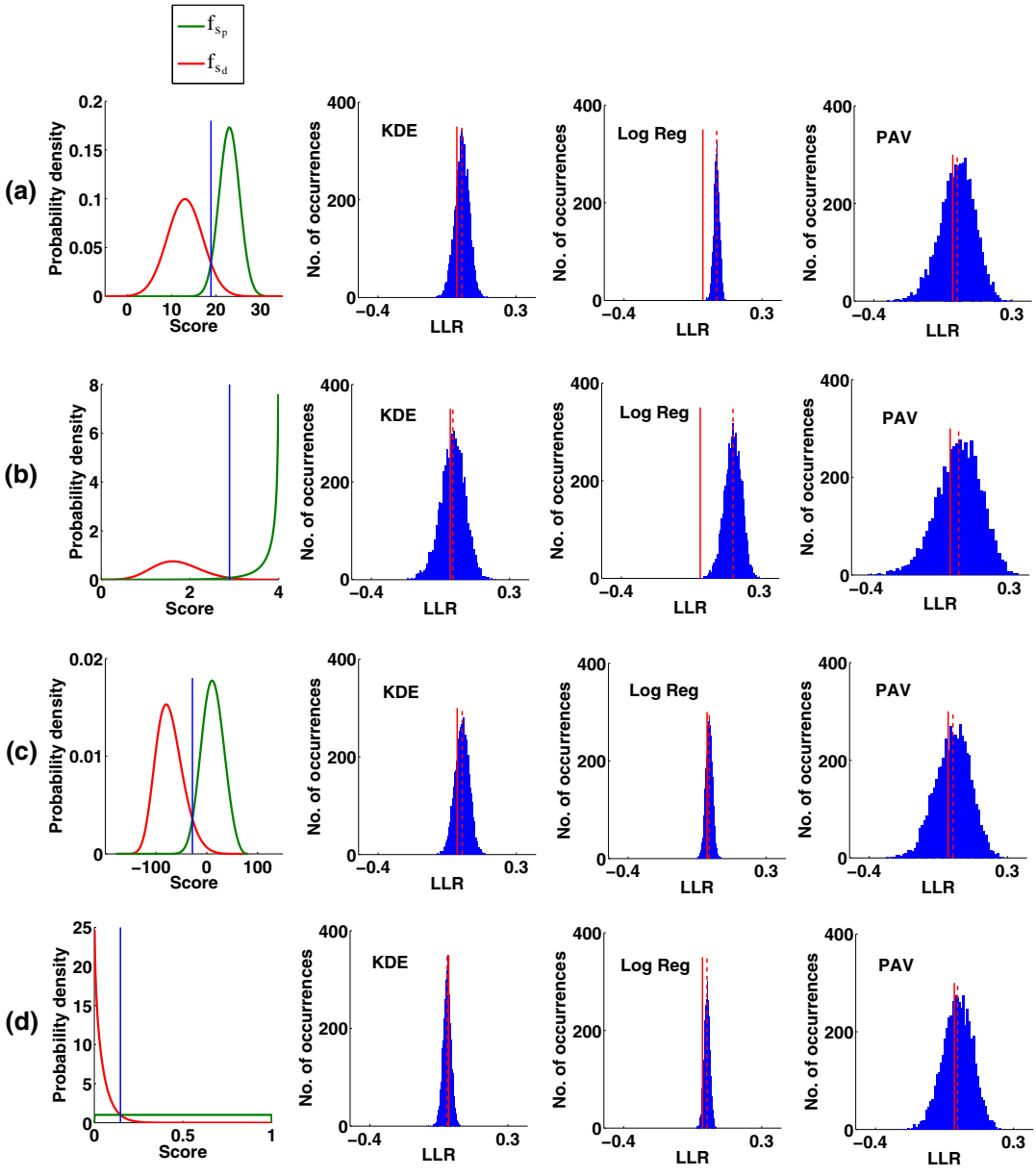

Fig. 4. The leftmost column shows the PDFs with the considered score s shown as a vertical line. The next three columns show histograms of the 5000 $LLR_{2000,\,100{,}000}(s)$ values computed by each method using the training sets tr_1, tr_2, …, tr_{5000} where size of $tr_i = (n^p, n^d) = 2000, 100{,}000$) and are generated using random sampling from the corresponding pairs of PDFs.

Please cite this article as: T. Ali, et al., Sampling variability in forensic likelihood-ratio computation: A simulation study, Sci. Justice (2015), http://dx.doi.org/10.1016/j.scijus.2015.05.003

3.3. Pool Adjacent Violators (PAV)

Given the sets of the s_p and the s_d scores, PAV algorithm (also called Isotonic Regression) [40] combines them into a single set, sorts and assigns a posterior probability $P(H_p|score)$ of 1 to each score of the s_p and 0 to each score of the s_d set. Then, it iteratively looks for an adjacent group of posterior probabilities that violates monotonicity and replaces it with the average of that group. The process of pooling and replacing violator groups' values with the average is continued until the whole sequence is monotonically increasing. The result is a set of posterior probabilities $P(H_p|score)$ where each value corresponds to a score from either the s_p or the s_d set. These posterior probabilities along with the sizes of the training sets are used to obtain LRs by application of the Bayesian formula in Eq. (3). To compute the LR of a new score, linear interpolation is used between the training scores. A detailed description of the PAV algorithm can be found in [40].

It should be mentioned that, another commonly used method to compute LRs from scores is by finding the slope of the Receiver Operating Characteristics Convex Hull (ROCCH). However, it is recently proved that this method is equivalent to PAV [41]. Furthermore, PAV can be considered as an optimized version of the LR computation using histogram binning [42] because PAV chooses optimal bin size depending on the size of the training data in different score locations.

4. Experimental setup

4.1. Selection of PDFs

We consider four pairs of PDFs from which the n realizations of the training sets are generated by random sampling. These training sets are used to assess the sampling variability and accuracy of each LR computation method using the simulation framework discussed in Section 2.2.

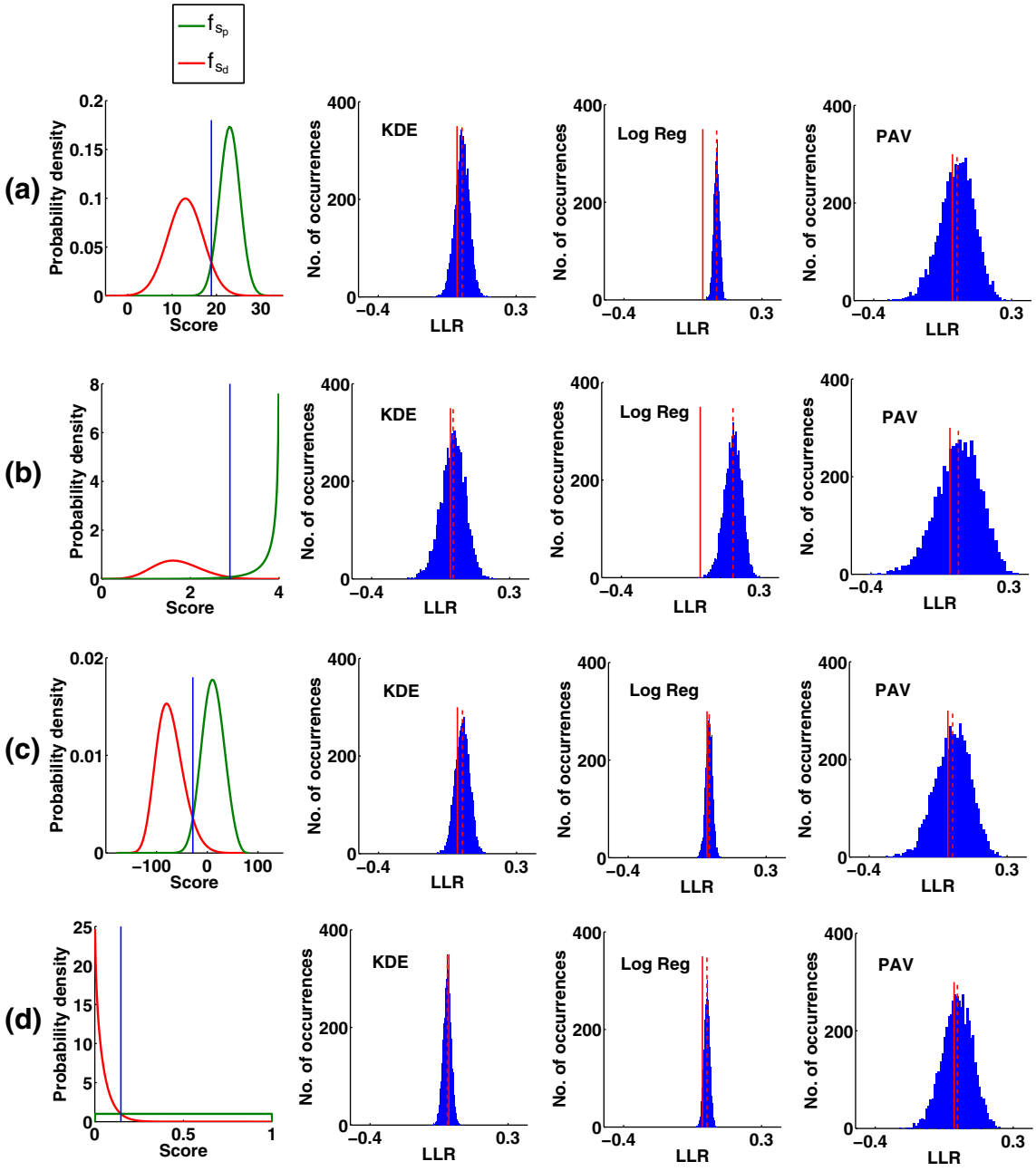

Fig. 5. The leftmost column shows the PDFs with the considered score s shown as a vertical line. The next three columns show histograms of the 5000 $LLR_{20,\ 1000}(s)$ values computed by each method using the training sets tr_1, tr_2, …, tr_{5000} where size of $tr_i = (n^p, n^d) = 20, 1000)$ and are generated using random sampling from the corresponding pairs of PDFs.

Please cite this article as: T. Ali, et al., Sampling variability in forensic likelihood-ratio computation: A simulation study, Sci. Justice (2015), http://dx.doi.org/10.1016/j.scijus.2015.05.003

The first pair of PDFs consists of two Normal PDFs shown in Fig. 3(a). The specific values of the parameters are chosen such that there is some overlap between the two PDFs and the standard deviation of the PDF of the s_d scores is larger than that of the PDF of the s_p scores (see Table 1). The choice of Normal PDFs in this comparative study is motivated by its widespread use to model the distribution of scores in the s_p and s_d sets in various biometric modalities such as handwriting and fingerprint recognition [9,16].

The second pair of PDFs consists of two reversed Weibull PDFs shown in Fig. 3(b) and expressed as:

$$f(s;\lambda,k) = \frac{k}{\lambda}\left(\frac{s_{\max}-s}{\lambda}\right)e^{\frac{(s_{\max}-s)}{\lambda}}, \quad (8)$$

where $k > 0$ is the scale parameter, $\lambda > 0$ is the shape parameter and $s_{\max} = 4$, found experimentally, is the score value along which the PDFs are reversed. Table 2 shows the values of k and λ for the two PDFs.

The choice of these specific PDFs is motivated by the shapes of the distributions of the scores obtained by face recognition systems based on Boosted Linear Discriminant Analysis (BLDA) [18,43].

Next we consider large s_p and s_d sets of scores from a speaker recognition system based on the Probabilistic Linear Discriminant Analysis (PLDA) [44] approach which models the distribution of i-vectors as a multivariate Gaussian. The system is described in [44, Section 2.5] and is used with the data set of the National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) of 2010 [45]. Fig. 3(c) shows the s_p set, the s_d set and the matched pair of PDFs using MLE. These are reversed Weibull PDFs, flipped along $s_{\max}$ as in Eq. (8), with different parameters as shown in Table 3. In this case, $s_{\max}$ is the maximum value of the score in the set $\{\{s_p\},\{s_d\}\}$.

Lastly we consider a state-of-the-art commercial face recognition system developed by Cognitec [46]. It is used with face images captured in extreme surveillance scenario from the SCFace database [47]. There are 5 different qualities of surveillance cameras and 130 subjects in the database. Each subject has a frontal mugshot which is compared to the surveillance camera images. The resultant s_p and s_d score sets are shown in Fig. 3(d). Using MLE, the s_p score set is best fitted by the Uniform PDF and the s_d scores set by the Beta PDF shown in Fig. 3(d) and expressed as:

$$f_{s_p}(s;a,b) = \begin{cases} \frac{1}{b-a} & \text{for } a \le s \le b \\ 0 & \text{otherwise} \end{cases}$$

$$f_{s_d}(s;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} s^{\alpha-1}(1-s)^{\beta-1}$$

where values of a, b, α and β are 0.0020, 0.9995, 0.8532 and 17.3373 respectively found using MLE.

4.2. Avoiding infinite Log_{10}-likelihood-ratios (LLRs)

A LR of 0 or ∞ (LLR of −∞ or ∞) is generally undesirable in forensic evaluation. In order to avoid computation of infinite LLRs, we insert the score s, for which a LR is computed, both in the s_p and in the s_d set. A slightly different strategy has been proposed in [28] where two additional scores are inserted in these sets: one at the maximum possible score location and another at the minimum possible score location. In general, both of these strategies are motivated by Laplace's rule of succession. The inserted scores can be considered to represent training scores which were not encountered in the training sets because there is not enough training biometric data but which could have occurred.

5. Results

Each pair of PDFs in Fig. 3 are randomly sampled and 5000 realizations of the training sets $tr_1 = \{\{s_{p_1}\},\{s_{d_1}\}\}$, $tr_2 = \{\{s_{p_2}\},\{s_{d_2}\}\}$, …, $tr_{5000} = \{\{s_{p_{5000}}\},\{s_{d_{5000}}\}\}$ are generated. In practice, the number of scores in the s_p and s_d sets available for computation of a LR is case-dependent and can vary. We consider three sizes of the training sets (n^p, n^d): (2000, 100,000), (200, 10,000) and (20, 1000). The choice of the small size of the s_p set compared to the s_d set is motivated by the fact that it is generally the case in practice.

In order to understand the effect of the shape of the distribution of the scores in the s_p and in the s_d set and the sizes of these two sets used for training, a fixed score s is considered. This score is shown by the vertical line in each pair of PDFs in the leftmost column in Fig. 4. The value of s is chosen such that its LLR computed from the PDFs is 0; i.e., $LLR_{\infty,\infty}(s) = 0$ or $LR_{\infty,\infty}(s) = 1$. The right three columns of Fig. 4 show, for each method, the histograms of 5000 $LLR_{2000,\ 100{,}000}(s)$ values using the training sets tr_1, tr_2, …, tr_{5000}. The solid vertical line in the histograms shows the $LLR_{\infty,\infty}(s)$ whereas the dotted vertical line shows the mean LLR of s computed from the set of 5000 LLRs. The distance between the solid line and the dotted line gives an estimate of the accuracy whereas the spread of the histograms of the LLRs gives an estimate of the sampling variability. The use of a logarithmic scale is preferred for plotting purposes and it has intuitive appeal for forensic practitioners. In Fig. 4, it can be concluded that, based on these large sizes of the training sets, the Log Reg method is least sensitive to the sampling variability in the training sets followed by the

Table 4
The mean and interval between the maximum (Max) and minimum (Min) LLRs for the three different sizes of the training sets. For each size, the mean LLR closest to the $LLR_{\infty,\infty}$ and the smallest interval is in bold.

	KDE		Log Reg		PAV	
	Mean	[Min, Max]	Mean	[Min, Max]	Mean	[Min, Max]
(a) For PDFs in Figs. 4–6(a)						
(2000, 100,000)	0.0258	[−0.1172, 0.1699]	0.0700	**[0.0101, 0.1204]**	**0.0229**	[−0.3306, 0.2976]
(200, 10,000)	**0.0643**	[−0.2744, 0.3041]	0.0718	**[−0.1471, 0.2299]**	0.0922	[−0.5293, 0.4899]
(20, 1000)	0.1747	**[−0.3841, 0.5735]**	**0.0942**	[−0.7927, 0.4034]	0.3374	[−0.1730, 0.9379]
(b) For PDFs in Figs. 4–6(b)						
(2000, 100,000)	**0.0130**	[−0.2197, 0.2328]	0.1661	**[−0.0110, 0.3139]**	0.0426	[−0.4139, 0.3461]
(200, 10000)	**0.0636**	**[−0.5754, 0.4315]**	0.1622	[−0.6047, 0.4937]	0.1806	[−0.4399, 0.7116]
(20, 1000)	0.3062	**[−0.1645, 0.7999]**	**0.2200**	[−0.6569, 0.9012]	0.6109	[0.1219, 1.3260]
(c) For PDFs in Figs. 4–6(c)						
(2000, 100,000)	0.0254	[−0.1017, 0.1442]	**0.0109**	**[−0.0515, 0.0715]**	0.0240	[−0.3162, 0.2819]
(200, 10,000)	0.0653	[−0.3135, 0.3048]	**0.0140**	**[−0.2543, 0.1829]**	0.1030	[−0.6170, 0.5829]
(20, 1000)	0.1790	**[−0.3659, 0.6237]**	**0.0542**	[−0.9233, 0.4088]	0.3674	[−0.1413, 0.9478]
(d) For PDFs in Figs. 4–6(d)						
(2000, 100,000)	**−0.0078**	[−0.1014, 0.0599]	0.0228	**[−0.0312, 0.0710]**	0.0166	[−0.2779, 0.2368]
(200, 10,000)	−0.0423	[−0.2485, 0.1296]	**0.0234**	**[−0.1688, 0.1769]**	0.0803	[−0.5180, 0.4731]
(20, 1000)	−0.0915	**[−0.8170, 0.3097]**	**0.0353**	[−0.8341, 0.3626]	0.3347	[−0.1543, 0.9792]

Please cite this article as: T. Ali, et al., Sampling variability in forensic likelihood-ratio computation: A simulation study, Sci. Justice (2015), http://dx.doi.org/10.1016/j.scijus.2015.05.003

KDE and then the PAV method. However, the mean LLR of the PAV and KDE is closer to $LLR_{\infty,\infty}$ for distribution pairs in Fig. 4(a), (b) and (d) demonstrating the better generalization ability of these non-parametric approaches across different shapes of the score distributions. The mean values computed by the Log Reg method in Fig. 4(a) and (c) also show the fact that a small difference in the shapes of the distributions of the scores in the training sets can considerably affect the accuracy of the Log Reg method.

To see the effect of reduction in the sizes of training sets, Fig. 5 shows the results when the sizes of the s_p and the s_d set are 20 and 1000 respectively. Note the larger range of the x-axis in the histograms of Fig. 5 compared to Fig. 4, showing, in general, a large standard deviation of the LLRs due to the reduction in the sizes of the training sets. For these small training sets, Log Reg outperforms KDE and PAV in terms of accuracy, however, it is more sensitive to the sampling variability in the training sets compared to KDE.

Sampling variability should not be ignored when reporting the strength of an evidence using a LR. As a specific example of the amount of sampling variability, when the size is (20, 1000), the minimum and maximum LRs computed by KDE in Fig. 5(d) are $\frac{1}{7}$ and 2 respectively, resulting in the closed interval, $[\frac{1}{7}, 2]$. Sampling variability may not be a serious concern in cases where a very large or a very small value of LR is computed. However, it will still be a good practice to perform and include an assessment of the sampling variability when reporting the strength of a biometric evidence in the form of a LR. The corresponding intervals of LRs in Fig. 5(d) in the case of the Log Reg method and the PAV method are $[\frac{1}{7}, 2]$ and [1, 10] respectively. Note that these intervals of LRs are for a score lying in a location which is expected to be less sensitive to the sampling variability in the training sets because of the large number of data points in this score location. We will discuss the effect of varying the score location later.

Fig. 6 shows the effect of the sizes of the training sets on the standard deviation of the 5000 LLRs and on the bias ($mean - LR_{\infty,\infty}$) of each method. Some notes are in the following order:

- In the case of the Log Reg method, the sizes of the training sets have very little effect on the bias. The bias, however, is dependent on the shapes of the distributions of the scores. This is one of the major drawbacks of most parametric approaches, in general; once the model is not appropriate, the sizes of the training sets cannot compensate for it. In contrast to the bias, the sampling variability in the case of the Log Reg method is dependent on both the sizes of the training sets as well as the shapes of the distributions of the scores.
- In the case of the PAV method, the shapes of the distributions of the scores in the training sets have very little effect on the sampling variability. The sampling variability, however, is dependent on the sizes of the training sets. In contrast to the sampling variability, the bias in the PAV method is dependent on both the sizes of the training sets as well as the shapes of the distributions of the scores.
- The KDE method follows a behavior similar to that of the PAV method but it is less biased and has less sampling variability particularly when the sizes of the training sets are reduced significantly, e.g., the case of (20, 1000), as shown in Fig. 6.

Table 4 shows the mean, maximum and minimum values of the 5000 LLRs computed by each method considering the different sizes of the training sets. As also observed in Figs. 4–6, when the sizes of the training sets are large, the Log Reg method has the smallest sampling variability. However, in the case of the small training sets, the KDE method has the smallest sampling variability.

In order to study the effect of the score for which a LR is computed, the sizes of the training sets are kept fixed. Since sampling variability is considerable when the sizes of the training sets are small, we consider the sizes $(n^p, n^d) = (20, 1000)$. A set of 50 equidistant scores are generated in the interval $[\min(s_{d_i})\ \max(s_{p_i})]$ for $i = 1, 2, \ldots, 5000$. Then, as

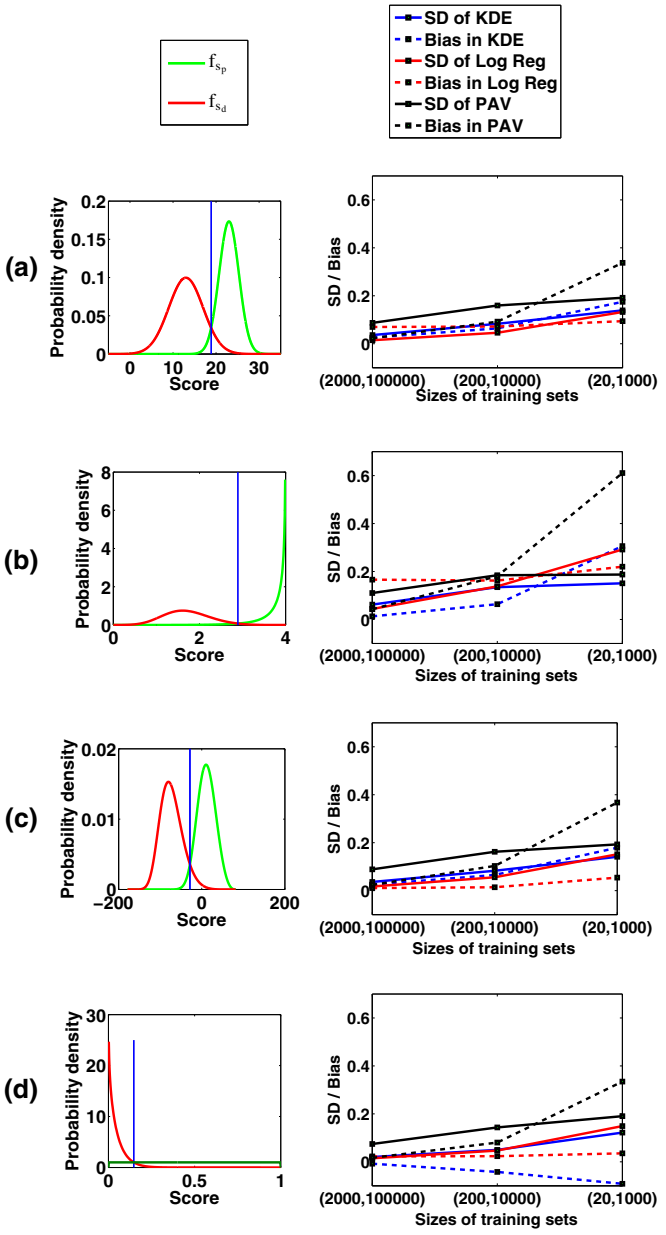

Fig. 6. The left column shows the PDFs with the considered score value shown as a vertical line. The right column shows the standard deviation (SD) and bias values for three different sizes of the training sets.

previously, for each score in this set, 5000 LLRs are computed using the 5000 training sets tr_1, tr_2, …, tr_{5000} generated by random sampling of the PDFs. The mean, maximum and minimum values of the 5000 LLRs are plotted for each score in the set of 50 equidistant scores (see Fig. 7). Only the results for the pairs of PDFs in Figs. 4–6(c) and (d) are shown. The closer the minimum and maximum curves are, the smaller the sampling variability. The closer the mean and the $LR_{\infty,\infty}$ curves are, the smaller the bias. In the case of the KDE method, the function from scores to LLRs may not be monotonically increasing. This is a very undesirable property in evidence evaluation where the score increases with the degree of similarity between the two biometric specimens. A specific regularization procedure such as imposing a condition on s as "$\min(s_p) \le s \le \max(s_d)$ and mapping scores outside this range to the end-points of this range" is needed in the KDE approach. There is a large sampling variability in the Log Reg approach when the score location is not nearby the intersection point of the two PDFs.

Please cite this article as: T. Ali, et al., Sampling variability in forensic likelihood-ratio computation: A simulation study, Sci. Justice (2015), http://dx.doi.org/10.1016/j.scijus.2015.05.003

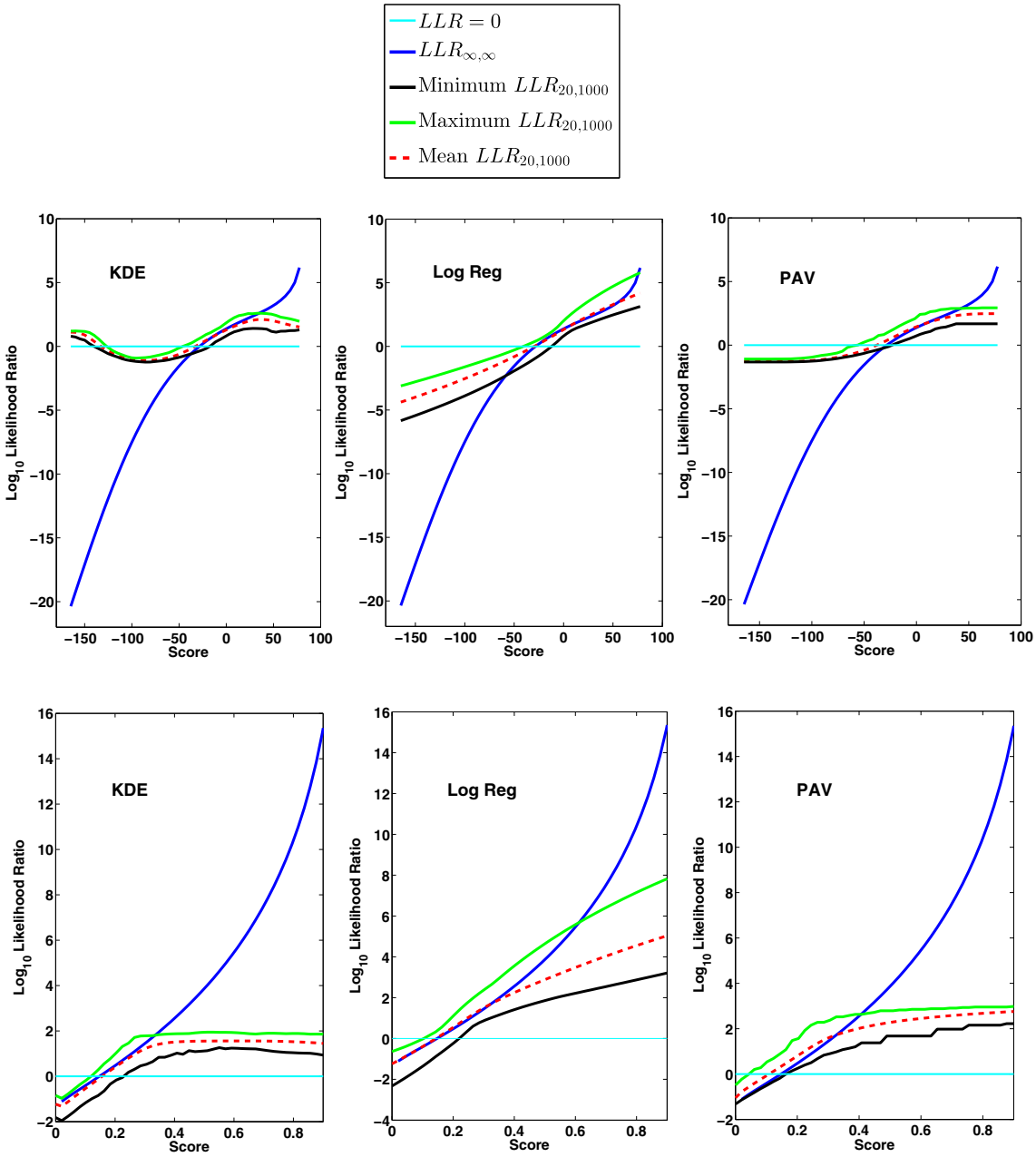

Fig. 7. The mean, maximum and minimum LLRs computed from the set of 5000 LLRs of each of the score in the set of 50 equidistant scores. The top and bottom rows show results based on the pairs of the PDFs in Figs. 4–6(c) and (d) respectively.

It must be pointed out that the effects observed do not hold in general. Particularly, as the shapes of the PDFs vary, we may observe different behaviors of these three methods.

6. Conclusions and future work

In this paper we compared three LR computation methods commonly proposed for evidence evaluation based on assumed distributions and sizes of the databases. The focus was to understand and quantify the effect of the sampling variability in the training sets used. A simulation framework is proposed for this purpose and a comparative study is carried out based on different shapes of the PDFs from which the training sets are generated, the sizes of the training sets and the value of the score for which a LR is computed. It is observed that these three parameters are important and should be considered in order to appropriately select a method of LR computation since all of them affect the sampling variability and accuracy of the computed LR. It is shown that the sampling variability is a serious concern when the sizes of the training sets are small. Especially, the tails of the distributions cause large sampling variability in the LR. The study suggests that given a pair of biometric specimens, a range of LRs should be reported which incorporates the sampling variability instead of reporting a single value of the LR.

For future work, research can be carried out in order to derive the formal relationship between the sizes of the training sets and the sampling variability in the computed LR for different methods of LR computation. LR of a score lying in the tails of the distributions has more sampling variability compared to a score lying in the middle region of the score distributions. Based on the available set of training scores, procedures can be developed for prediction of the location of the score for which a LR is computed. It can be useful for estimation of the sampling variability in a LR. Furthermore, specific strategies can be incorporated in the LR computation methods to make them more robust for cases where small training sets are available for LR computation.

Please cite this article as: T. Ali, et al., Sampling variability in forensic likelihood-ratio computation: A simulation study, Sci. Justice (2015), http://dx.doi.org/10.1016/j.scijus.2015.05.003

References

[1] A.K. Jain, A. Ross, S. Pankanti, Biometrics: a tool for information security, IEEE Trans. Inf. Forensics Secur. 1 (2) (2006) 125–143.
[2] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Forensic Evidence for Forensic Scientists, 2nd ed. Wiley, Chichester, 2004.
[3] B. Robertson, G.A. Vignaux, Interpreting Evidence: Evaluating Forensic Science in the Courtroom, Wiley, Chichester, 1995.
[4] Q. Tao, R. Veldhuis, Robust biometric score fusion by naive likelihood ratio via receiver operating characteristics, IEEE Trans. Inf. Forensics Secur. 8 (2) (2013) 305–313.
[5] D.J. Balding, Weight-of-evidence for Forensic DNA Profiles, Wiley, Chichester, UK, 2005.
[6] Y. Tang, S.N. Srihari, Likelihood ratio estimation in forensic identification using similarity and rarity, Pattern Recognit. 47 (3) (2014) 945–958.
[7] G. Zadora, Evaluation of evidence value of glass fragments by likelihood ratio and Bayesian Network approaches, Anal. Chim. Acta 642 (1) (2009) 279–290.
[8] T. Grant, Quantifying evidence in forensic authorship analysis, Int. J. Speech. Lang. Law 14 (1) (2007) 1–25.
[9] A.B. Hepler, C.P. Saunders, L.J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Forensic Sci. Int. 219 (1) (2012) 129–140.
[10] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, J. Ortega-Garcia, Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems, Forensic Sci. Int. 155 (2) (2005) 126–140.
[11] M.J. Sjerps, C.E.H. Berger, How clear is transparent? Reporting expert reasoning in legal cases, Law Probab. Risk 11 (2012) 317–329.
[12] G.S. Morrison, F. Ochoa, T. Thiruvaran, Database selection for forensic voice comparison, Proceedings of Odyssey: The Language and Speaker Recognition Workshop 2012, pp. 62–77.
[13] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Speech Comm. 31 (2) (2000) 193–203.
[14] P. Rose, Technical forensic speaker recognition: evaluation, types and testing of evidence, Comput. Speech Lang. 20 (2) (2006) 159–191.
[15] D. Ramos-Castro, Forensic evaluation of the evidence using automatic speaker recognition systems(PhD dissertation) Universidad Autonoma de Madrid, 2007.
[16] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc. Ser. A 175 (2) (2012) 1–26.
[17] C. Peacock, A. Goode, A. Brett, Automatic forensic face recognition from digital images, Sci. Justice 44 (1) (2004) 29–34.
[18] T. Ali, L.J. Spreeuwers, R.N.J. Veldhuis, Towards automatic forensic face recognition, International Conference on Informatics Engineering and Information Science (ICIEIS), Communications in Computer and Information Science, 252, Springer Verlag 2011, pp. 47–55.
[19] I. Alberink, A. Jongh, C. Rodriguez, Fingermark evidence evaluation based on automated fingerprint identification system matching scores: the effect of different types of conditioning on likelihood ratios, J. Forensic Sci. 59 (1) (2014) 70–81.
[20] T. Ali, L.J. Spreeuwers, R.N.J. Veldhuis, D. Meuwly, Effect of calibration data on forensic likelihood ratio computation from a face recognition system, Proceedings of: Biometrics: Theory, Application and Systems (BTAS), IEEE, Washington, DC, 2013.
[21] D. Meuwly, Forensic individualisation from biometric data, Sci. Justice 38 (4) (2006) 198–202.
[22] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice 51 (3) (2011) 91–98.
[23] J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Sci. Justice 42 (1) (2002) 29–37.
[24] G.W. Beecham, B.S. Weir, Confidence interval of the likelihood ratio associated with mixed stain DNA evidence, J. Forensic Sci. 56 (1) (2011) 166–171.
[25] N. Brummer, Tutorial for Bayesian forensic likelihood ratio, AGNITIO Research, South Africa, Tech. Rep, 2013 ([Online]. Available: http://arxiv.org/abs/1304.3589).
[26] D. Lindley, Understanding Uncertainty, Wiley, New Jersey, 2006.
[27] C. Tippett, V. Emerson, M. Fereday, F. Lawton, A. Richardson, L. Jones, S. Lampert, The evidential value of the comparison of paint flakes from sources other than vehicles, J. Forensic Sci. Soc. 8 (2) (1968) 61–65.
[28] N. Brummer, J. Preez, Application-independent evaluation of speaker detection, Comput. Speech Lang. 20 (2) (2006) 230–275.
[29] D. Van Leeuwen, N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, Speaker Classification I, Springer, Berlin Heidelberg 2007, pp. 330–353.
[30] D. Ramos-Castro, Reliable support: measuring calibration of likelihood ratios, Forensic Sci. Int. 230 (1) (2013) 156–169.
[31] G.S. Morrison, C. Zhang, P. Rose, An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system, Forensic Sci. Int. 208 (1) (2011) 59–65.
[32] A. Alexander, Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions(PhD Dissertation) École polytechnique fédérale de Lausanne, Lausanne, Switzerland, 2005.
[33] G.S. Morrison, A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: multivariate kernel density (MVKD) versus Gaussian mixture model — universal background model (GMM-UBM), Speech Comm. 53 (2) (2011) 242–256.
[34] E. Parzen, On estimation of probability density function and mode, Ann. Math. Stat. 33 (3) (1962) 267–281.
[35] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall/CRC, London, 1998. 47–55.
[36] D. Meuwly, A. Drygajlo, Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM), Proceedings of Odyssey: The Language and Speaker Recognition Workshop 2011, pp. 145–150.
[37] L.J. Davis, C.P. Saunders, A.B. Hepler, J. Buscaglia, Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios, Forensic Sci. Int. 216 (1) (2012) 146–157.
[38] G.S. Morrison, Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio, Aust. J. Forensic Sci. 45 (2) (2012) 173–197.
[39] R. Chartrand, W. Yin, Iteratively reweighted algorithms for compressive sensing, Proceedings of the 33rd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008.
[40] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining, 2002.
[41] T. Fawcett, A. Niculescu-Mizil, PAV and the ROC convex hull, Mach. Learn. 68 (1) (2007) 97–106.
[42] J.C. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in: A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large Margin Classifiers, MIT Press 1999, pp. 61–74.
[43] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, S.Z. Li, Ensemble-based discriminant learning with boosting for face recognition, IEEE Trans. Neural Netw. 17 (1) (2006) 166–178.
[44] M.I. Mandasari, M.I. McLaren, A. Van Leeuwen, The effect of noise on modern automatic speaker recognition systems, Proceedings of ICASSP, Kyoto, 2012.
[45] NIST Speaker Recognition Evaluation 2010, http://www.itl.nist.gov/iad/mig/tests/sre/2010/.
[46] Cognitec Systems GmbH, FaceVACS C++ SDK, Version 8.4.0, 2010.
[47] M. Grgic, K. Delac, S. Grgic, SCface — surveillance cameras face database, Multimed. Tools Appl. 51 (3) (2011) 863–879.

Please cite this article as: T. Ali, et al., Sampling variability in forensic likelihood-ratio computation: A simulation study, Sci. Justice (2015), http://dx.doi.org/10.1016/j.scijus.2015.05.003