Science & Justice xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Science & Justice

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

# Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors

Geoffrey Stewart Morrison[a,b,*], Norman Poh[c,d]

[a] Forensic Speech Science Laboratory, Centre for Forensic Linguistics, Aston University, Birmingham, England, United Kingdom
[b] Isaac Newton Institute for Mathematical Sciences, Cambridge, England, United Kingdom
[c] Department of Computer Science, University of Surrey, Guildford, England, United Kingdom
[d] QuintilesIMS, London, England, United Kingdom

ARTICLE INFO

*Keywords:*
Likelihood ratio
Bayes factor
Shrinkage
Conservative
Regularize
Logistic regression

ABSTRACT

When strength of forensic evidence is quantified using sample data and statistical models, a concern may be raised as to whether the output of a model overestimates the strength of evidence. This is particularly the case when the amount of sample data is small, and hence sampling variability is high. This concern is related to concern about precision. This paper describes, explores, and tests three procedures which shrink the value of the likelihood ratio or Bayes factor toward the neutral value of one. The procedures are: (1) a Bayesian procedure with uninformative priors, (2) use of empirical lower and upper bounds (ELUB), and (3) a novel form of regularized logistic regression. As a benchmark, they are compared with linear discriminant analysis, and in some instances with non-regularized logistic regression. The behaviours of the procedures are explored using Monte Carlo simulated data, and tested on real data from comparisons of voice recordings, face images, and glass fragments.

## 1. Introduction

When strength of forensic evidence is quantified using sample data and statistical models, a concern may be raised as to whether the output of the model overestimates the strength of evidence, particularly when the amount of sample data is small. The present paper explores three different statistical procedures for addressing this concern, and discusses their advantages and disadvantages.

From a frequentist perspective, the models used to calculate the numerator and denominator of a likelihood ratio are sample-based estimates of probability density functions that have true but unknown population distributions. Given a large amount of sample data and assuming models appropriate for fitting the population distributions, the modelled sample distributions will be reasonable approximations of the population distributions, and a calculated likelihood ratio value will be a reasonable estimate of the true but unknown likelihood ratio value. When the amount of data sampled is small, sampling variability may result in modelled sample distributions that deviate substantially from the population distributions, and hence a calculated likelihood ratio could be a poor estimate of the true but unknown likelihood ratio value. This raises a concern about imprecision in general, about whether the estimate is too high or too low, but practical proposals for dealing with imprecision are usually also driven by a desire not to overstate strength of evidence. Hence, rather than calculate and report an upper and lower bound for a coverage interval around a calculated likelihood ratio value, what may be reported is the bound closest to the neutral likelihood ratio value of 1 (assuming both upper and lower bounds are greater than 1 or both are less than 1), e.g., rather than report a two-sided 90% interval of 100 to 10,000, report a one-sided 95% bound of at least 100. The practice may be to interpret a confidence interval as if it were a credible interval, or actually formally calculate a Bayesian credible interval, or to simply adjust the value of the reported likelihood ratio to that of a bound.[1] If the amount of sample data is large, the coverage interval will be small, hence there will be little adjustment to the reported value, but if the amount of sample data is small, the coverage interval will be large, hence the reported value will be substantially closer to 1. For discussion and examples of approaches of this sort, see [1–9].

From a subjectivist Bayesian perspective, there are no true but unknown population distributions, and the value of a Bayes factor (which is the Bayesian counterpart of the frequentist likelihood ratio) is a state of belief, not an estimate of a true but unknown value (in the context of evaluation of forensic evidence, this position is espoused in, for example [10,11]). Calculation of a Bayesian posterior predictive

[*] Corresponding author at: Forensic Speech Science Laboratory, Centre for Forensic Linguistics, Aston University, Birmingham, England, United Kingdom.
*E-mail address:* geoff-morrison@forensic-evaluation.net (G.S. Morrison).

[1] The latter practice is common in DNA, see [1], and is also attested elsewhere, e.g., for footwear [2].

https://doi.org/10.1016/j.scijus.2017.12.005
Received 8 August 2017; Received in revised form 6 October 2017; Accepted 20 December 2017

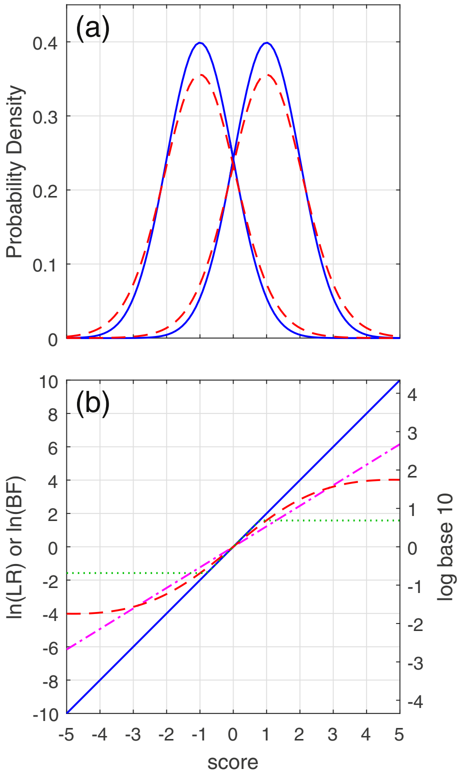

**Fig. 1.** (a) Sample-based Gaussian distributions (solid blue lines), and Bayesian posterior predictive distributions (dashed red lines). (b) Score to log likelihood ratio or score to log Bayes factor transformation functions based on: Sample-based Gaussian distributions (solid blue line), Bayesian posterior predictive distributions (dashed red line), ELUB (dotted green line), and regularized logistic regression (dot-dashed magenta line). For interpretation of the references to colour in figure captions, the reader is referred to the electronic version of this article.

distribution is based not only on sample statistics but also on prior distributions for parameter values. If the amount of sample data is small the effect of the priors will be greater. Two Bayesians starting with different priors will arrive at different posterior predictive distributions when the amount of sample data is small, but their posterior predictive distributions will converge on the sample distribution as the amount of sample data increases.

If a Bayesian does not have strong prior beliefs about the distribution of the parameter values, they will use (relatively) uninformative priors. Uninformative priors have broad flat distributions, such that when the amount of sample data is small the posterior predictive distributions will be substantially broader and flatter than the sample distributions. The solid blue lines in Fig. 1a show two *Gaussian distributions*. The means for the leftmost and rightmost distributions are −1 and +1 respectively, and they both have the same variance, which is 1. These can be thought of as representing sample distributions. In contrast, the dashed red lines show *Bayesian posterior predictive distributions* given the same sample means and variance, assuming 10 sample data points were used to calculate each mean, a pooled calculation of the variance using all 20 data points, and uninformative Jeffreys reference priors.[2] With such a small amount of data, the posterior predictive distributions are substantially broader and flatter than the sample distributions.

We take the rightmost Gaussian distribution in Fig. 1a as a model for calculating the numerator of a likelihood ratio, and the leftmost distribution as a model for calculating the denominator. This model, one Gaussian for each category with both Gaussians having the same variance, is known as *linear discriminant analysis* (LDA).[3] We also take the rightmost Bayesian posterior predictive distribution in Fig. 1a as a model for calculating the numerator of a Bayes factor, and the leftmost distribution as a model for calculating the denominator. We calculate log likelihood ratios or log Bayes factors for the range of values on the $x$ axis, and plot the resulting values on the $y$ axis in Fig. 1b. The solid blue line corresponds to log likelihood ratios calculated using LDA, and the dashed red line corresponds to log Bayes factors calculated using the posterior predictive distributions. Note that the broader flatter posterior predictive distributions result in log Bayes factors that are closer to 0 than the log likelihood ratios calculated using the narrower peakier sample distributions (if the amount of sample data were larger, the log Bayes factor values would be closer to the log likelihood ratio values).

Although philosophically very different, a frequentist procedure using the bound of the coverage interval closest to 1 and a Bayesian procedure using uninformative priors would have the same practical effect: If the amount of sample data is small, the strength of evidence value will be closer to the neutral value of 1 (the log strength of evidence value will be closer to the neutral value of 0) than it would be if the amount of sample data were larger. Both procedures exhibit a property known as *shrinkage*. As a potential practical solution to the philosophical impasse, [12] proposed that those on both sides of the debate agree to use procedures involving shrinkage, without having to adopt the other side's philosophical interpretation of probability. We hope that those on both sides will give this proposal serious consideration and not dismiss it out of hand. Imprecision or sensitivity would still be assessed, but the results of such an assessment would only be used to help decide whether the performance of the system was good enough to be used in casework, and would not be used to further reduce the magnitude of the reported strength of evidence.

[13] presents another proposal for avoiding overestimating the strength of evidence. Large magnitude log likelihood ratio values are desirable, as they give strong support for one hypothesis over the other, but large magnitude log likelihood ratio values result from the questioned-origin data being on the tail of at least one of the distributions. The tails of distributions are intrinsically sparsely sampled, hence probability density estimates in tail areas are susceptible to large fluctuations resulting from even small changes due to sampling variability. [13] proposes the imposition of upper and lower bounds on the value of the likelihood ratio. Values beyond the bounds are replaced by the values at the bounds. The bounds are empirically determined as the value at which adding one misleading log likelihood ratio value[4] would result in worse performance in the trier of fact's decision making process than if the system were replaced by a system that always outputs a log likelihood ratio value of 0 (a likelihood ratio of 1). This is called a *consequential misleading likelihood ratio*. The dotted green line in Fig. 1b shows an example of the result of applying the *empirical lower and upper bound* (ELUB) procedure to the likelihood ratios calculated using the LDA procedure (assuming 10 data points for training each sample mean and a pooled sample variance based on all 20 data points). All else being equal, as the amount of sample data increases the number of data points falling in the tails of the distributions will increase, and the deviation of the upper and lower bounds from the neutral log likelihood

[2] This procedure and the other procedures outlined in the introduction are described in greater detail in Section 3.

[3] LDA is generalizable to any number of categories, but in the present paper (in the context of score to likelihood ratio conversion, and in one case feature to likelihood ratio conversion) we only use binomial LDA. In the automatic speaker recognition and automatic face recognition literature LDA is used to refer to a procedure for reducing the number of dimensions prior to applying some other probabilistic classification model. In the statistics literature, however, LDA itself refers to a probabilistic classification model. In the present paper we use the term LDA in the latter sense. The former sense might be described as using canonical linear discriminant functions for dimension reduction without actually performing a linear discriminant analysis.

[4] A misleading log likelihood ratio is one which is less than 0 when the hypothesis in the numerator is actually true, or greater than 0 when the hypothesis in the denominator is actually true.

ratio value of 0 will increase.

Another model which explicitly includes shrinkage is *regularized logistic regression*. In score-based approaches to likelihood ratio calculation, logistic regression is a popular model for *score to likelihood ratio conversion* (aka *calibration*; e.g., [14–16]). Small amounts of regularization can be applied to logistic regression to avoid numerical problems in parameter estimation. Larger amounts of regularization can be applied to deliberately reduce the slope of the fitted logistic regression model, and hence shrink the log likelihood ratio output. The present paper introduces a novel variant of regularized logistic regression inspired by uninformative Bayesian prior distributions. Regularization is achieved by adding to the sample data a uniform "prior" distribution with a weight equivalent to a specified number of pseudodata points. The dot-dashed magenta line in Fig. 1b shows an example of the result of applying regularized logistic regression to the same sample data as before (10 sample points from each of two categories) using regularization based on a uniform prior with a weight of 5 pseudodata points. The log likelihood ratio values resulting from the regularized logistic regression procedure are closer to 0 than the log likelihood ratio values resulting from LDA.

We have outlined four different procedures for calculating likelihood ratios or Bayes factors:

1. linear discriminant analysis (LDA),
2. a Bayesian procedure using uninformative priors,[5]
3. empirical lower and upper bounds (ELUB),[6] and
4. regularized logistic regression (LogReg).

The latter three involve some form of shrinkage. LDA is included as a baseline against which to compare the other procedures. Also for comparison purposes, on some datasets we will also test non-regularized logistic regression.[7]

In the remainder of the present paper:

- Section 2 briefly outlines score-based approaches for the calculation of likelihood ratios/Bayes factors.
- Section 3 provides additional details related to each of the procedures outlined above.
- Section 4 explores the behaviour of each procedure on simulated score data.
- Section 5 tests the performance of each procedure when applied to real data from comparisons of voice recordings, face images, and glass fragments.
- Section 6 provides discussion and conclusion.

In order to more quickly understand the main arguments of the present paper, readers may wish to skip Section 3 on first reading.

Software that implements the procedures described in Section 3, and the analyses reported in Section 4 and Section 5.2 (along with the score data for the latter), is provided at http://geoff-morrison.net/#shrunk_LRs. The score data for the analysis reported in Section 5.3 is available from http://personal.ee.surrey.ac.uk/Personal/Norman.Poh/web/fusion/.

## 2. Score-based approaches for the calculation of likelihood ratios/Bayes factors

Score-based approaches are increasingly popular across multiple branches of forensic science, e.g., [17–25]. Quantitative measurements made on objects of interest such as voice recordings, face images, and glass fragments usually result in multivariate data with complex distributions. Models fitted to these feature data may be, for example, kernel density models or Gaussian mixture models, the former semi-parametric and the latter parametric requiring a large number of parameter values to be estimated. For example, in forensic voice comparison it is not uncommon to fit a Gaussian mixture model with 1024 Gaussian components to feature data with around 28 dimensions, requiring estimates for 85,988 parameter values. The amount of training data is seldom sufficient to obtain good estimates for such a large number of parameters, hence the values output by such models cannot be safely interpreted as the ratios of likelihoods answering the same-origin and different-origin hypotheses specified for the case. Instead, the outputs of such (first-stage) models are treated as *scores* which must be converted to likelihood ratios (or calibrated) before their values can be interpreted.

The first stage of a score-based procedure can be considered a procedure for extracting information about the similarity of the questioned-origin data with respect to the known source and their typicality with respect to the relevant population,[8] and projecting the complex multidimensional feature space down to a univariate score space. The second stage then fits simple models to the univariate scores. These second-stage models only require estimation of a few parameter values, and there are a relatively large number of training scores available for estimating the values of those parameters. The parameter estimates for the second-stage models are therefore good estimates, and the outputs of these models are naturally well calibrated. The projection from a complex multidimensional feature space to a simple univariate score space involves a loss of information, but, given the difficulty of fitting good models in the original feature space, score-based procedures may empirically outperform procedures which attempt to directly estimate likelihood ratio values in the original complex multidimensional feature space.

The absolute value of a score is not interpretable, but a score has the form of a log likelihood ratio, it quantifies the similarity of the questioned-origin data with respect to the known-origin sample and the typicality of the questioned-origin data with respect to a sample of the relevant population. Hence, given two scores, the absolute values of the scores and the absolute difference between the scores are not interpretable, but the higher valued score should correspond to a higher valued likelihood ratio than does the lower valued score. An appropriate score to likelihood ratio conversion model should, therefore, be monotonic, at least within the foreseeable operating range of the system.[9]

The second-stage score to likelihood conversion model should be trained using training data that are distinct from the training data used to train the first-stage model. The second-stage training data should consist of feature values extracted from pairs of objects of interest (pairs of voice recordings, facial images, glass fragments, etc.). Some pairs

[5] In the present paper, we will always use Jeffreys reference priors as the uninformative priors for the Bayesian procedure.

[6] ELUB is not actually a procedure for calculating likelihood ratios, but a procedure for limiting the size of the likelihood ratios generated by another procedure. In the present paper, with one exception, we will always apply ELUB to the output of LDA. ELUB could also be applied to the output of other procedures.

[7] In the present paper we do not explore the performance of procedures based on kernel density models or Gaussian mixture models because they do not induce shrinkage. When data are sparse, even if only locally sparse, this increases concern regarding imprecision and/or overestimating strength of evidence, and in such circumstances non-parametric, semi-parametric, and low-bias parametric procedures are prone to overfitting the training data and exacerbating the problem. Such procedures are also potentially problematic because they are not monotonic and do not naturally produce well calibrated results by estimating a small number of parameters using a relatively large amount of data.

[8] Scores should be calculated in a manner which captures information about both similarity and typicality, i.e., the procedure for calculating a score should be an attempt to estimate a log likelihood ratio value. Scores based only on similarity lack information about the typicality of the questioned-origin data with respect to the relevant population, and that information cannot be adequately incorporated as part of the score to likelihood ratio conversion stage [26].

[9] [27] describes some models which are in principle non-monotonic, but which in practice may be monotonic (but not linear) within the operating range of the system. [28] illustrates problems with models that are not monotonic within the operating range of the system.

must be known to be same-origin pairs and some pairs must be known to be different-origin pairs. Each pair is input to the already-trained first-stage model, and the output is a set of same-origin scores and a set of different-origin scores. These same- and different-origin scores are then used to train the second-stage model.

The training data for the second-stage model should be sampled from the relevant population specified in the different-origin hypothesis. One member of each pair must have conditions reflecting those of the known-origin data in the case and the other member of the pair must have conditions reflecting those of the questioned-origin data in the case. For example, in a forensic voice comparison case the relevant population may be specified to be speakers of a given sex speaking a particular language with a particular accent, and the questioned-speaker recording may be of a lively mobile telephone conversation (mobile telephone codecs distort and discard information from the speech signal) whereas the known-speaker recording may be of subdued answers to a police interview made in a room with substantial reverberation and ventilation system noise. If the second-stage training data differ substantially from the relevant population or conditions, this will produce miscalibrated results [29,30].

## 3. Details regarding each of the four procedures

### 3.1. Linear discriminant analysis (LDA)

A simple model for score to likelihood ratio conversion is linear discriminant analysis [31], i.e., two Gaussian distributions with the mean of one calculated using the different-origin training scores, the mean of the other calculated using the same-origin training scores, and both using the same pooled variance calculated using all the training scores. Use of the same variance for both Gaussians ensures that the score to likelihood ratio conversion function is linear, and hence monotonic, [12,18 §6.5.2.1, 28,32]. Using different variances for each Gaussian (quadratic discriminant analysis) would result in a non-linear non-monotonic function.

Eq.1 represents a linear discriminant analysis model, in which: $\Lambda^{\mathrm{LDA}}$ is the likelihood ratio corresponding to score value $x$; $\hat{\mu}_s$ and $\hat{\mu}_d$ are the sample means calculated from the same-origin training scores and different-origin training scores respectively; $\hat{\sigma}^2$ is the pooled sample variance calculated using data from both same-origin and different-origin training scores; $f(x \mid \hat{\mu}, \hat{\sigma}^2)$ is the Gaussian probability density function; and $\ln(\cdot)$ are natural logarithms.

$$\ln(\Lambda^{\mathrm{LDA}}) = \ln(f(x \mid \hat{\mu}_s, \hat{\sigma}^2)) - \ln(f(x \mid \hat{\mu}_d, \hat{\sigma}^2)) \tag{1a}$$

$$f(x \mid \hat{\mu}, \hat{\sigma}^2) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{\frac{-(x-\hat{\mu})^2}{2\hat{\sigma}^2}} \tag{1b}$$

The linear form given in Eq. (2) can be derived from Eq. (1).

$$\ln(\Lambda^{\mathrm{LDA}}) = a + bx \tag{2a}$$

$$a = -\frac{\hat{\mu}_s^2 - \hat{\mu}_d^2}{2\hat{\sigma}^2} = -b\frac{\hat{\mu}_s + \hat{\mu}_d}{2} \tag{2b}$$

$$b = \frac{\hat{\mu}_s - \hat{\mu}_d}{\hat{\sigma}^2} \tag{2c}$$

Note that the slope $b$ is the difference between the means divided by the variance. In Fig. 1 the means were 2 variance units apart, hence the slope was 2.

The linear discriminant analysis model does not include shrinkage, but it will be tested below to provide results against which the performance of the other models can be compared.

### 3.2. Bayesian model using uninformative Jeffreys reference priors

For a Bayesian calculation of likelihood using sample data assumed to be from a Gaussian distribution with unknown mean and variance, the conjugate prior is a Gaussian-gamma distribution, and the posterior predictive distribution is a $t$ location and scale distribution [33 §9.6, 34]. The calculations are described in Eq. (3), in which: $\lambda^{\mathrm{B}}$ is the likelihood of the Bayesian model evaluated at a score value $x$; $n$ is the number of sample data points used for calculating the sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$; $\mu^{\mathrm{prior}}$ is the prior belief regarding the mean, and $\kappa^{\mathrm{prior}}$ is how strong that belief is (quantifiable as the number of pseudodata points); and $\beta^{\mathrm{prior}}$ is the prior belief regarding the variance, and $\alpha^{\mathrm{prior}}$ is how strong that belief is. In Eq. (3f), $\nu$ is the degrees of freedom and $\Gamma(\cdot)$ is the gamma function.

$$\lambda^{\mathrm{B}} = t_{2\alpha^{\mathrm{post}}}\left(x \mid \mu^{\mathrm{post}}, \frac{\kappa^{\mathrm{post}} + 1}{\kappa^{\mathrm{post}}\alpha^{\mathrm{post}}}\beta^{\mathrm{post}}\right) \tag{3a}$$

$$\mu^{\mathrm{post}} = \frac{\kappa^{\mathrm{prior}}\mu^{\mathrm{prior}} + n\hat{\mu}}{\kappa^{\mathrm{prior}} + n} \tag{3b}$$

$$\kappa^{\mathrm{post}} = \kappa^{\mathrm{prior}} + n \tag{3c}$$

$$\alpha^{\mathrm{post}} = \alpha^{\mathrm{prior}} + \frac{n}{2} \tag{3d}$$

$$\beta^{\mathrm{post}} = \beta^{\mathrm{prior}} + \frac{n\hat{\sigma}^2}{2} + \frac{\kappa^{\mathrm{prior}} n(\hat{\mu} - \mu^{\mathrm{prior}})^2}{2(\kappa^{\mathrm{prior}} + n)} \tag{3e}$$

$$t_\nu(x \mid \mu, \sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\nu + \left(\frac{x-\mu}{\sigma}\right)^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)} \tag{3f}$$

In general, values for the hyperparameters $\mu^{\mathrm{prior}}$, $\kappa^{\mathrm{prior}}$, $\beta^{\mathrm{prior}}$, and $\alpha^{\mathrm{prior}}$ must be selected, either reflecting informative prior beliefs or giving (relatively) uninformative wide flat prior distributions. This model was previously described in [35], and used in conjunction with relatively uninformative priors to convert scores to Bayes factors. This included a demonstration of the effect of varying the amount of sample data on the average magnitude of log Bayes factors: as the amount of sample data decreases the average magnitude of the log Bayes factors decreases. See [12] for a graphical example of calculating Bayes factors using this model and relatively uninformative priors. [36] includes results of applying a similar (but multivariate) model using informative priors and using relatively uninformative priors. The latter resulted in a log Bayes factor that was much closer to zero.

An option for uninformative priors is to use Jeffreys reference priors, for which the values for the hyperparameters $\kappa^{\mathrm{prior}}$, $\beta^{\mathrm{prior}}$, and $\alpha^{\mathrm{prior}}$ are 0, 0, and $-½$ respectively (the value of $\mu^{\mathrm{prior}}$ is irrelevant since it is always multiplied by $\kappa^{\mathrm{prior}} = 0$).[10] Substituting these hyperparameter values into Eq. (3) and simplifying leads to Eq. (4).

$$\lambda^{\mathrm{B}} = t_{n-1}\left(x \mid \hat{\mu}, \frac{n+1}{n-1}\hat{\sigma}^2\right) \tag{4}$$

A Bayes factor $\Lambda^{\mathrm{BF}}$ can then be calculated as in Eq. (5), in which the $s$ and $d$ subscripts indicate values derived from same-origin training scores and different-origin training scores respectively.

$$\ln(\Lambda^{\mathrm{BF}}) = \ln\left(t_{n_s-1}\left(x \mid \hat{\mu}_s, \frac{n_s+1}{n_s-1}\hat{\sigma}_s^2\right)\right) - \ln\left(t_{n_d-1}\left(x \mid \hat{\mu}_d, \frac{n_d+1}{n_d-1}\hat{\sigma}_d^2\right)\right) \tag{5}$$

Because of the use of $t$ distributions rather than Gaussian distributions, using the same variance in both numerator and denominator will not guarantee monotonicity, but we impose this constraint to reduce the potential extent of non-monotonicity. This leads to a modification of Eq. (5), as shown in Eq. (6), in which $\hat{\sigma}^2$ is the pooled sample variance

[10] Note that the Jeffreys reference priors are improper priors with distributions that do not integrate to 1 and that cannot be graphically represented. The pseudodata interpretation also breaks down.

and the degrees of freedom have been adjusted to take account of the pooled variance calculation (the adjusted degrees of freedom imply that $\alpha^{\text{prior}}$ was $-1$).

$$\ln(\Lambda^{\text{BF}}) = \ln\left(t_{n_s+n_d-2}\left(x \mid \widehat{\mu}_s, \frac{\overline{n}+1}{\overline{n}-1}\widehat{\sigma}^2\right)\right) - \ln\left(t_{n_s+n_d-2}\left(x \mid \widehat{\mu}_d, \frac{\overline{n}+1}{\overline{n}-1}\widehat{\sigma}^2\right)\right) \tag{6a}$$

$$\overline{n} = \frac{n_s + n_d}{2} \tag{6b}$$

When we calculate scores below, we often adopt a procedure of the following form: If we have 2 recordings from each of 10 speakers ($n_{spk} = 10$), we compare every speaker's first recording with their own second recording and with every other speaker's second recording. This results in a total of 100 scores, 10 same-speaker scores ($n_s = 10$) and 90 different-speaker scores ($n_d = 90$). The calculation of different-speaker scores reuses each recording multiple times, and thus the score values are not statistically independent. Rather than base the expansion of the variance and the degrees of freedom on $n_s + n_d$, we therefore base them on $n_{spk}$, i.e., as in Eq. (5) but with $n_s = n_d = n_{spk}$ and using the pooled variance $\widehat{\sigma}^2$. Mutatis mutandis when the comparisons are of face images or glass fragments.

### 3.3. Empirical lower and upper bound (ELUB)

The ELUB procedure is a secondary procedure applied to the output from a procedure that calculates likelihood ratios. The procedure and its rationale are described in [13]. To implement the procedure, we need training data consisting of a set of likelihood ratio values known to be from same-origin comparisons and a set of likelihood ratio values known to be from different-origin comparisons (note that these are likelihood ratio values, not score values). We sort the likelihood ratio values from the training set in ascending order, and at each likelihood ratio value in the training set (a threshold value $LR_{th}$) we calculate the expected utility ratio EUratio(ELUB) using Eq. (7), in which: $n_{LR_s}$ and $n_{LR_d}$ are the total number of same-origin and different-origin likelihood ratios in the training set; $n_{LR_s} \leq LR_{th}$ is the number of same-origin likelihood ratios in the training set that are less than or equal to $LR_{th}$; and $n_{LR_d} > LR_{th}$ is the number of different-origin likelihood ratios in the training set that are greater than $LR_{th}$. The numerator reads: If $LR_{th} > 1$ then 1, else $LR_{th}$.[11]

$$\text{EUratio(ELUB)} = \frac{\begin{cases} 1 & \text{if } LR_{th} > 1 \\ LR_{th} & \text{if } LR_{th} \leq 1 \end{cases}}{\frac{n_{LR_s \leq LR_{th}} + 1}{n_{LR_s} + 1} + LR_{th} \times \frac{n_{LR_d > LR_{th}} + 1}{n_{LR_d} + 1}} \tag{7}$$

The first value of $LR_{th}$ for which EUratio(ELUB) is greater than 1 is the empirical lower bound, and the last value of $LR_{th}$ for which EUratio (ELUB) is greater than 1 is the empirical upper bound. When the ELUB procedure is applied, all likelihood ratios below the empirical lower bound are replaced by the value of the empirical lower bound, and all likelihood ratios above the empirical upper bound are replaced by the value of the empirical upper bound.

In the present paper, we will apply the ELUB procedure to the output of the LDA procedure. As same- and different-origin likelihood ratios for training, we use likelihood ratio values calculated from the same same- and different-origin scores that were used to train the LDA model. The upper and lower bounds are then imposed on the LDA output obtained when the model is applied to test scores.

### 3.4. Regularized logistic regression

If the assumptions for LDA hold that the data have Gaussian distributions with equal variance, and there are sufficient training data, logistic regression will give the same results as LDA. Logistic regression, however, is not dependent on these assumptions, and hence is more robust to violations of these assumptions than is LDA, [37] §4.4.5. A logistic regression model is fitted using an iterative maximum-likelihood algorithm, descriptions of which can be found in multiple sources including [37–41]. Logistic regression is usually thought of as calculating posterior probabilities, but if equal priors are used log posterior odds outputs are interpretable as log likelihood ratios. The model is linear in the logistic space (i.e., the logged odds space), and the fitted intercept and slope values in the logistic space can therefore be used to convert a score $x$ to a log likelihood ratio $\ln(\Lambda^{\text{LogReg}})$ as in Eq. (8). Eq. (8a) is identical to Eq. (2a), except that a different algorithm is used to calculate $a$ and $b$.

$$\ln(\Lambda^{\text{LogReg}}) = a + bx \tag{8a}$$

$$\begin{aligned} \ln(\Lambda^{\text{LogReg}}) &= \ln\left(\frac{p(x \mid H_s)}{p(x \mid H_d)}\right) = \ln\left(\frac{p(H_s \mid x)}{p(H_d \mid x)} \times \frac{p(H_d)}{p(H_s)}\right) \\ &= \ln\left(\frac{p(H_s \mid x)}{1 - p(H_s \mid x)} \times \frac{1 - p(H_s)}{p(H_s)}\right) \end{aligned} \tag{8b}$$

$$p(H_s \mid x) = \frac{e^{a+bx}}{1 + e^{a+bx}} \tag{8c}$$

$$p(H_s) = p(H_d) = 0.5 \tag{8d}$$

To fit a logistic regression model for $p(H_s|x)$, we code each different-origin training score $x_{d_i}$ as $p_{d_i} = 0$ and each same-origin training score $x_{s_j}$ as $p_{s_j} = 1$ (see Fig. 2 in which $[x_{d_i}, p_{d_i}]$ are plotted as large magenta triangles and $[x_{s_j}, p_{s_j}]$ as large blue circles). The iterative training algorithm searches for $a$ and $b$ values that will maximize the likelihood of the model averaged over all the $[x_{d_i}, p_{d_i}]$ and $[x_{s_j}, p_{s_j}]$ values (and taking into account any priors on the two categories). In the probability space shown in Fig. 2, $p(H_s|x)$ is plotted as the green sigmoidal curve. The slope of a tangent to the steepest part of the sigmoidal curve (which occurs at $p(H_s|x) = 0.5$) is $b/4$ ([42] p 24).

A numerical problem may occur in fitting the model if there is complete separation between the two categories in the training data (e.g., same-origin versus different-origin categories). The likelihood is maximized when the slope $b$ is infinite and the intercept $a$ is anywhere between the highest different-origin training score and the lowest same-origin training score. In practice, if the algorithm is stopped after a set number of iterations or when the change in the coefficients from one iteration to the next is less than a specified threshold, the slope will not actually reach infinity. The results from the last iteration, however, are unlikely to produce a model that is a good predictor for new data that fall between or close to the highest different-origin score and the lowest same-origin score in the training data. A solution proposed in [40] ch7 (and previously in [43] §4.3) for categorical predictor variables when some cells have zero count, is to add one observation to each cell. [44] extended this idea to continuous predictor variables by adding a copy of the same-origin training data (coded as $p_{s_j} = 1$) and recoding the copy as $p_{s_j}^{\text{copy}} = 0$ but giving it a small weight relative to the original same-origin training data, and mutatis mutandis for the different-origin training data (coded as $p_{d_i} = 0$) with the copy coded as $p_{d_i}^{\text{copy}} = 1$ but with a small weight. Adding the copies removes the complete separation in the training data and the algorithm converges on a non-infinite slope. The model has been regularized.[12] Whatever the data, the slope fitted using regularization will be shallower than for a model not including regularization, but, apart from in circumstances involving complete or near complete separation, the difference in slope will be slight if the regularization weight is small.

Here we propose a revision to this form of regularization which can

[11] We have simplified by assuming that we only add one consequential misleading likelihood ratio at $LR_{th}$.

[12] Note that this is a different form of regularization than the more common form described in [37] §4.4.4.

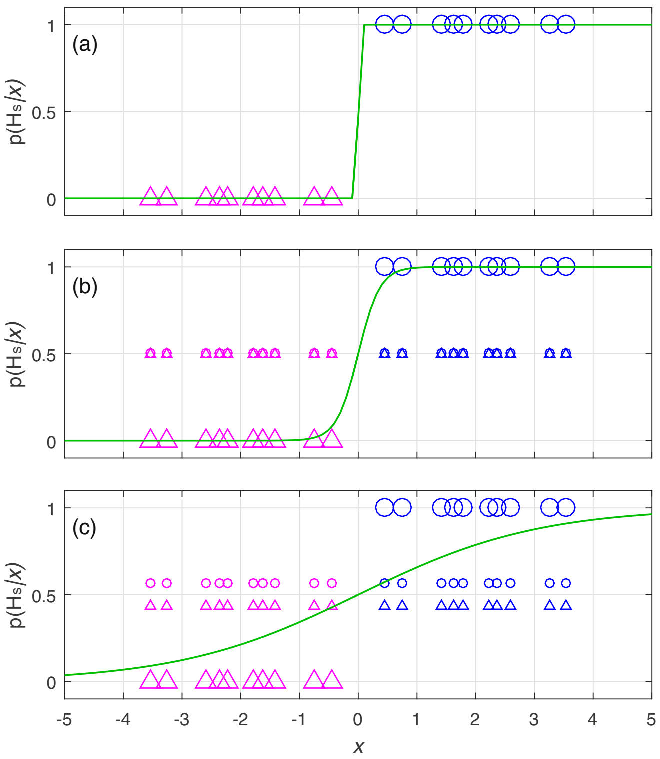

**Fig. 2.** (a) Example of logistic regression fitted without regularization. (b) Example of logistic regression fitted with a small amount of regularization to avoid numerical problems. (c) Example of logistic regression fitted with a large amount of regularization to induce shrinkage. The large symbols represent sample data, and the small symbols represent weighted pseudodata with an uninformative uniform distribution.

be used to substantially reduce the fitted slope and hence shrink the log likelihood ratio output. The proposed procedure is inspired by and mimics uninformative priors from Bayesian analyses. A "prior" distribution is added consisting of pseudodata points that have an uninformative uniform distribution. At each data point, $x_{d_i}$ and $x_{s_j}$, we add two pseudodata points which we code as ${p_{d_i}}^{\psi_0} = 0$ and ${p_{d_i}}^{\psi_1} = 1$ or as ${p_{s_j}}^{\psi_1} = 1$ and ${p_{s_j}}^{\psi_0} = 0$. We set the weights on these pseudodata points such that the sum of all the weights equals $\kappa^{\psi}$. The overall strength of the "prior" distribution is equivalent to $\kappa^{\psi}$ data points, as was the case in the Bayesian procedure described in Section 3.2 above in which $\kappa^{\text{prior}}$ was the strength of belief for the mean quantified as a number of pseudodata points. If the number of pseudodata points is fixed, the relative effect of regularization will decrease as the amount of real training data increases.

The weighting procedure we use is the same as the weighting procedure used for dealing with priors and imbalances in the amount of training data from each category. For example, if the priors for each category are equal, i.e., $p(H_s) = p(H_d)$, but the training data are imbalanced, e.g., $n_d = 3n_s$, then the $p_{d_i}$ are weighted by $w_d = n_s/(n_d + n_s) = 1/4$ and the $p_{s_j}$ are weighted by $w_s = n_d/(n_d + n_s) = 3/4$, such that $n_d w_d = n_s w_s$.[13] $x_{d_i}$ and $x_{s_j}$ with low weights will have $p_{d_i}$ and $p_{s_j}$ values close to 0.5, whereas $x_{d_i}$ and $x_{s_j}$ with high weights will have $p_{d_i}$ and $p_{s_j}$ values close to 0 and 1 respectively.

We apply the same procedure as described in the previous paragraph to weight the pseudodata points, weighting each point by $w^{\psi} = \kappa^{\psi}/2(n_d + n_s)$. Fig. 2b and c show the weighted pseudodata points as the small symbols near $p(H_s|x) = 0.5$ (triangles represent weighted ${p_{d_i}}^{\psi_0}$ and ${p_{s_j}}^{\psi_0}$, circles represent weighted ${p_{d_i}}^{\psi_1}$ and ${p_{s_j}}^{\psi_1}$). For Fig. 2b, $\kappa^{\psi} = 0.1$, and for Fig. 2c, $\kappa^{\psi} = 5$ (the amount of training data was 10 points for each category, generated using equal-variance Gaussians whose means were 4 variance units apart). As a heuristic, we recommend values of $\kappa^{\psi} \leq 0.1$ to avoid numerical problems, and $\kappa^{\psi} \geq 1$ to induce shrinkage.

For the remainder of the present paper, we will fix the value of $\kappa^{\psi}$ at 5. This value was chosen based on some preliminary exploration of performance using simulated data. Ultimately the choice of the value for $\kappa^{\psi}$ is arbitrary, which is a disadvantage of this procedure, although one should note that prior distributions also have to be selected for Bayesian procedures (even if the choice is to use Jeffreys reference priors that is still a choice).

When we calculate scores below, we often adopt a procedure of the following form: If we have 2 recordings from each of 10 speakers ($n_{spk} = 10$), we compare every speaker's first recording with their own second recording and with every other speaker's second recording. This results in a total of 100 scores, 10 same-speaker scores ($n_s = 10$) and 90 different-speaker scores ($n_d = 90$). The calculation of different-speaker scores reuses each recording multiple times, and thus the score values are not statistically independent. Rather than weight the pseudodata points by $w^{\psi} = \kappa^{\psi}/2(n_d + n_s)$, we will therefore weight them according to the number of speakers used to generate the scores $w^{\psi} = \kappa^{\psi}/2n_{spk}$. Mutatis mutandis when the comparisons are of face images or glass fragments.

## 4. Exploration of the behaviour of the four procedures using simulated data

As a preliminary exploration of the behaviour of the four procedures, we use simulated data. In the context of Monte Carlo simulations, one specifies the parameter values for the population distributions, and one can therefore calculate what in frequentist terms would be true likelihood ratio values. In subjectivist Bayesian terms, these would be the values towards which all individuals' beliefs regarding Bayes factors should converge given enough data. We will henceforth call these values the *reference values*.

We specified same-origin score and different-origin score Monte Carlo distributions as Gaussians with means of $\mu_d = -1$ and $\mu_s = +1$, and a common variance $\sigma^2$ of 1, i.e., the distributions shown as the solid blue lines in Fig. 1a.[14] We then used a pseudorandom-number generator (based on Mersenne Twister with a seed of 0) to generate samples from the specified Monte Carlo distributions. For each set of training data, 10 same-origin sample scores and 10 different-origin sample scores were generated. The sample generation process was repeated 1000 times to generate 1000 sets of training data.

For each set of simulated training data, we fitted an LDA model, a Bayesian model using Jeffreys reference priors, ELUB applied to the LDA output (LDA was used to calculate likelihood ratios on the training data and these were used to derive the upper and lower bounds), and regularized logistic regression with a uniform "prior" weighted by $\kappa^{\psi} = 5$ pseudodata points. Fig. 3 shows the resulting score to likelihood ratio conversion functions. The $x$ axis represents the score values and the $y$ axis the corresponding log likelihood ratio or log Bayes factor values. The $y$ axis is scaled in natural logarithms (equivalent log base 10 values are given on the right). Each thin line represents a function trained on one sample set. The thick line indicates the score to reference value function (this is the same as the solid blue line in Fig. 1b).

The LDA functions (Fig. 3a) have a substantial variation in log likelihood ratio values (vertical spread), which increases as the

[13] The weighting is implemented by linearly transforming the probabilistic space from the range 0 to 1, to the range − 1 to + 1, and multiplying the transformed $p_{d_i}$ and $p_{s_j}$ by the weights $w_d$ and $w_s$. The space is subsequently transformed back to the usual range of 0 to 1.

[14] For this preliminary exploration of the behaviour of the procedures, we generate data based on a model which meets the distributional assumptions of the LDA and Bayesian procedures, and are thus not a priori biassed against these procedures. Most of the sets of real data used in Section 5 clearly violate these assumptions.

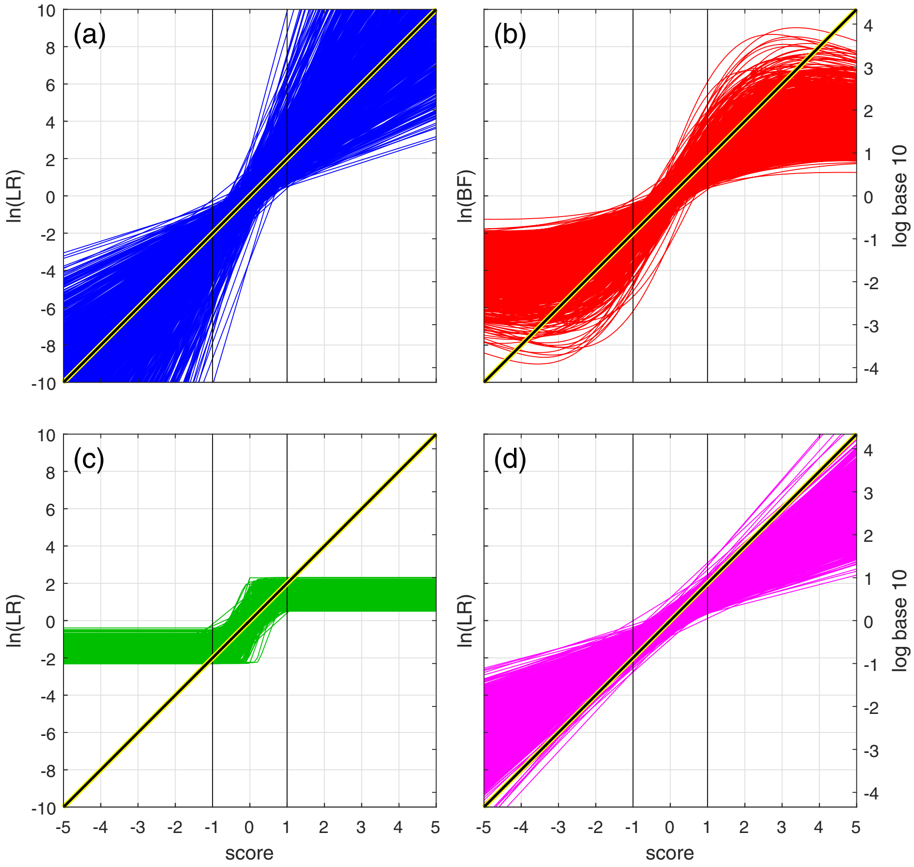

**Fig. 3.** Score to log likelihood ratio or score to log Bayes factor transformation functions fitted to Monte Carlo simulated data: 10 same-origin sample scores and 10 different-origin sample scores, same-origin and different-origin population means separated by 2 variance units. (a) LDA. (b) Bayesian procedure with uninformative Jeffreys reference priors. (c) ELUB applied to LDA output. (d) Regularized logistic regression.

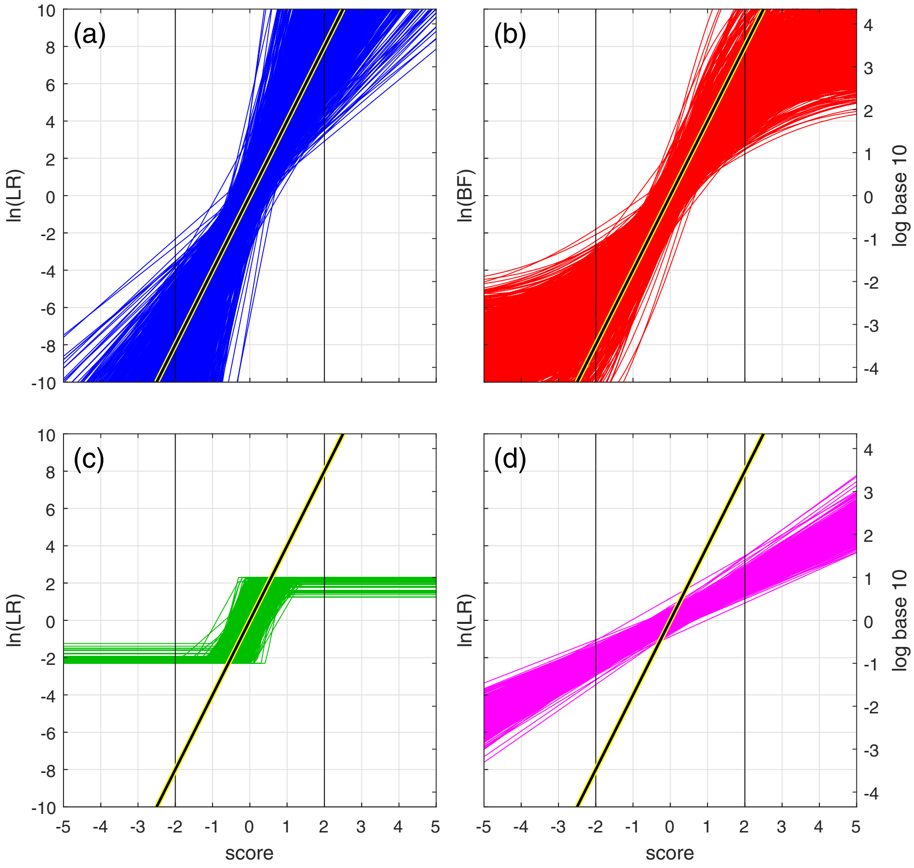

**Fig. 4.** Score to log likelihood ratio or score to log Bayes factor transformation functions fitted to Monte Carlo simulated data: 10 same-origin sample scores and 10 different-origin sample scores, same-origin and different-origin population means separated by 4 variance units. (a) LDA. (b) Bayesian procedure with uninformative Jeffreys reference priors. (c) ELUB applied to LDA output. (d) Regularized logistic regression.

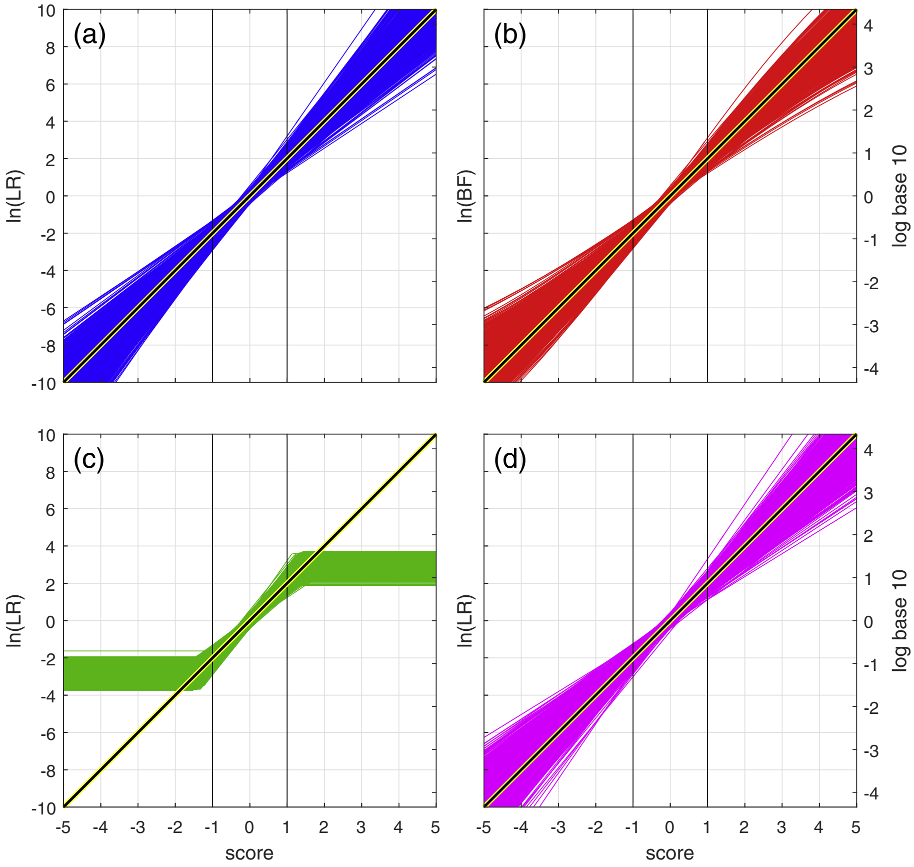

**Fig. 5.** Score to log likelihood ratio or score to log Bayes factor transformation functions fitted to Monte Carlo simulated data: 100 same-origin sample scores and 100 different-origin sample scores, same-origin and different-origin population means separated by 2 variance units. (a) LDA. (b) Bayesian procedure with uninformative Jeffreys reference priors. (c) ELUB applied to LDA output. (d) Regularized logistic regression.

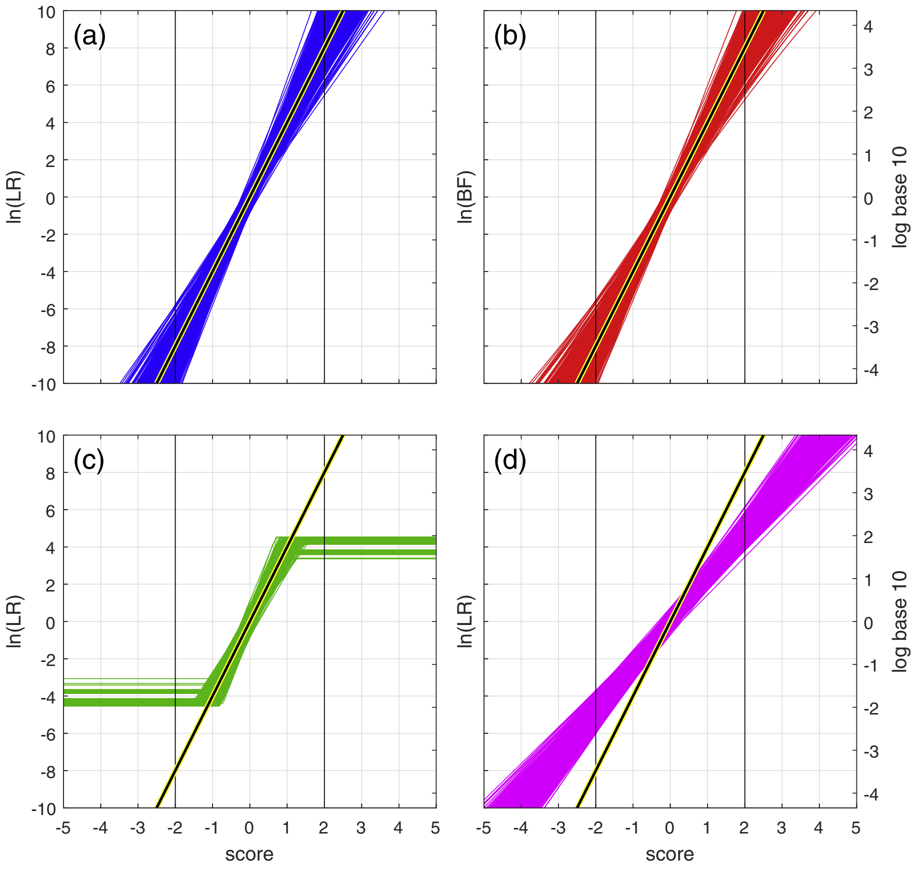

**Fig. 6.** Score to log likelihood ratio or score to log Bayes factor transformation functions fitted to Monte Carlo simulated data: 100 same-origin sample scores and 100 different-origin sample scores, same-origin and different-origin population means separated by 4 variance units. (a) LDA. (b) Bayesian procedure with uninformative Jeffreys reference priors. (c) ELUB applied to LDA output. (d) Regularized logistic regression.

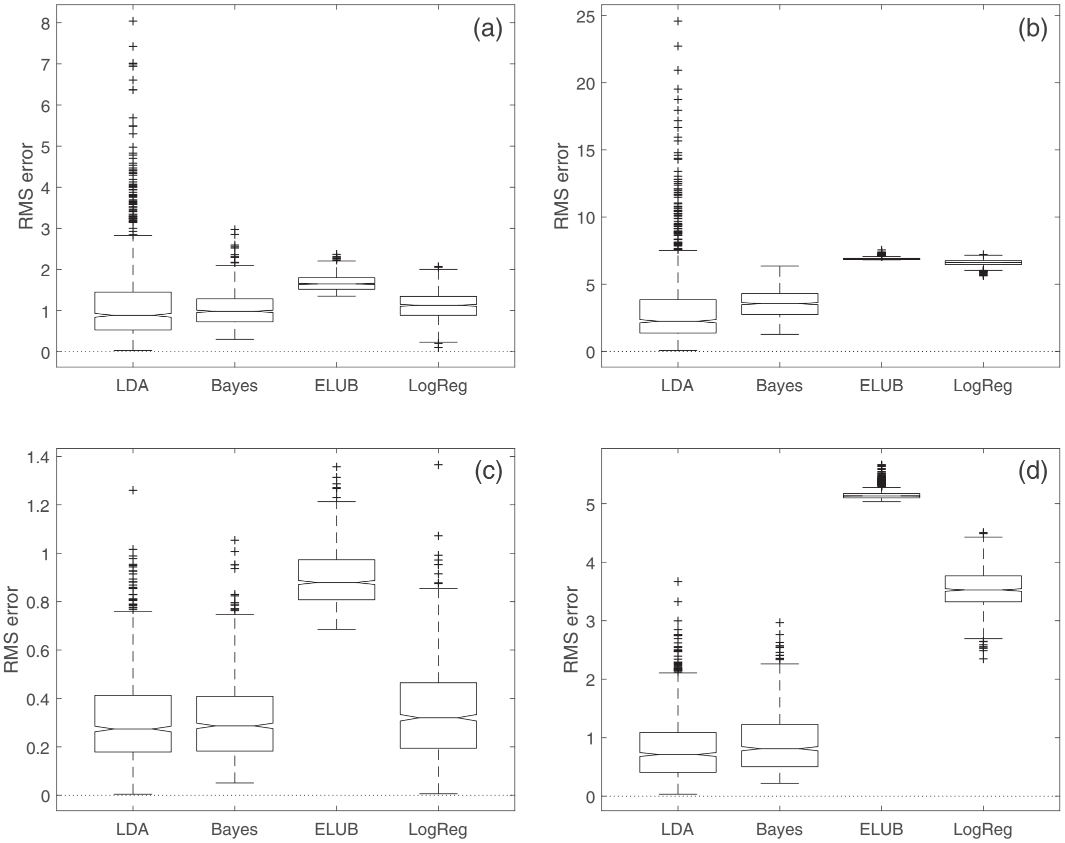

**Fig. 7.** Boxplots of RMS error values for each procedure in each condition: (a) 10 training samples per category and separation between the means in the Monte Carlo population distributions of 2 variance units. (a) 10 training samples per category and separation between the means in the Monte Carlo population distributions of 4 variance units. (a) 100 training samples per category and separation between the means in the Monte Carlo population distributions of 2 variance units. (a) 100 training samples per category and separation between the means in the Monte Carlo population distributions of 4 variance units. Note that the *y*-axis scale is different for each panel.

reference values deviate from 0. This imprecision is symmetrical about the reference values. Values that are substantially further from 0 than the reference values are of practical concern given that they overestimate the strength of evidence.

The Bayesian functions (Fig. 3b) have a substantial variation in log Bayes factor values (vertical spread), but on average less than the variation of log likelihood ratio values for the LDA functions. The Bayesian functions also tend to be closer to 0 than the reference value function – shrinkage has been effected. A substantial proportion of the Bayesian functions are, however, further from 0 than the reference value function. The Bayesian functions are sigmoidal and the rate at which they deviate from 0 is close to that of the reference function for low reference values, but decreases as the magnitude of the reference values increase. This may be considered desirable. Non-monotonicity can be observed in some Bayesian functions. This is not desirable.

The ELUB functions (Fig. 3c) follow the LDA functions for low magnitude reference values (by design), and are then suddenly truncated. The ELUB procedure is clearly the most conservative of those already considered. One could conclude that it is overly conservative. One could argue that limiting the log likelihood ratio values is desirable, but that the sudden truncation (cliff-edge effect) is not, and that score values substantially beyond a bound should correspond to log likelihood ratios further from 0 than scores just at the bound.[15] The better precision of ELUB compared to the other procedures is an artefact of the fact that all likelihood ratios beyond the upper or lower bound are replaced by the value at the bound and therefore all have the same value.

The regularized logistic regression functions (Fig. 3d) have substantial variation in log likelihood ratio values (vertical spread), but substantially less than that for the LDA functions. The regularized logistic regression functions exhibit shrinkage, they are almost always closer to 0 than the reference value functions. Unlike the Bayesian and ELUB functions, the regularized logistic regression functions are linear. One could consider this desirable.

Figs. 4–6 show results from simulations using larger amounts of training data and/or greater separation between the means used to generate the simulated same-origin scores versus different-origin scores. Fig. 4 is based on 10 data points per category, but with a separation between the means of 4 variance units. Figs. 5 and 6 are based on 100 points per category, and with separations between the means of 2 and 4 variance units respectively.[16]

Given a larger amount of training data (Figs. 5 and 6), the relative effect of shrinkage is less. All procedures give more precise results, and the procedures involving shrinkage also give more accurate results.

Given greater separation between the different-origin and same-origin scores (Figs. 4 and 6), the slopes of the reference value function, LDA functions, and Bayesian functions increase. When the separation between the categories is greater and the amount of data is larger, the ELUB bounds are further from 0 and the slopes of the regularized logistic regression functions increase (compare Fig. 5c and d with Fig. 6c and d). In contrast, when the separation between the categories is greater and the amount of data is small, the increases in the ELUB

[15] A potential alternative that avoids the sudden truncation could be to fit a sigmoidal function in the logistic space [45].

[16] We believe that this range of amount of sample data and range of separation between the categories is sufficient to gain an understanding of the relative behaviour of the procedures and to conceptually interpolate and extrapolate within and beyond these ranges. Other values can easily be substituted into the software we provide.

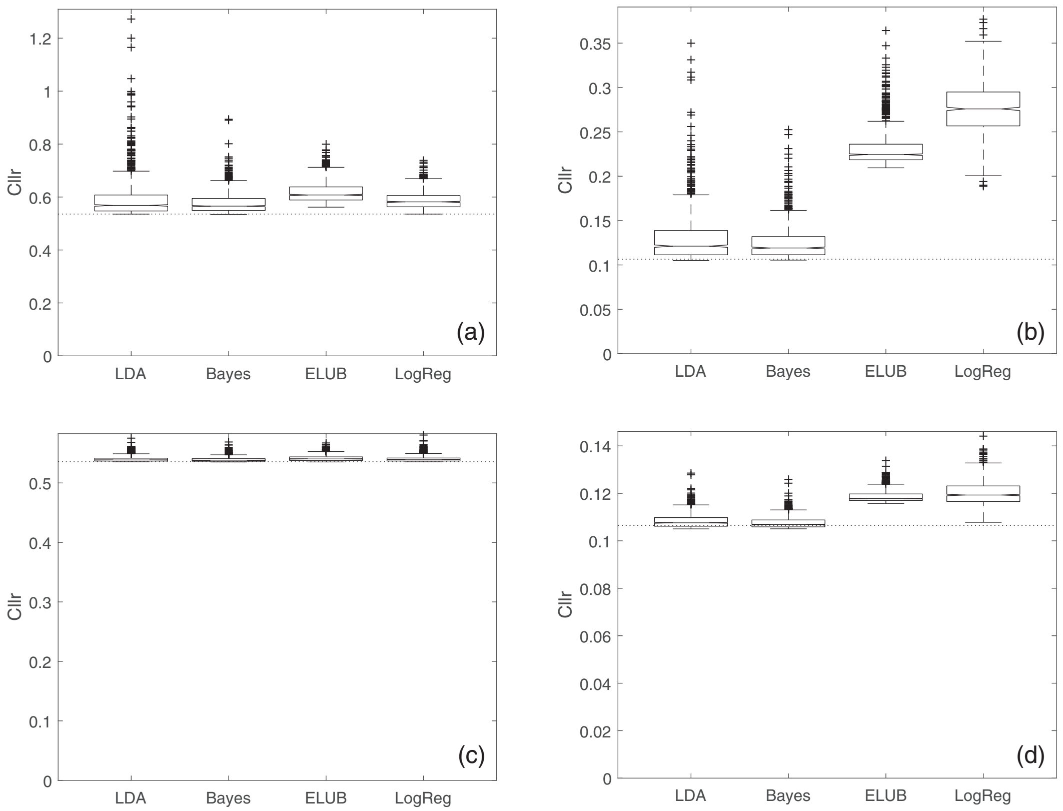

**Fig. 8.** Boxplots of $C_{\text{llr}}$ values for each procedure in each condition: (a) 10 training samples per category and separation between the means in the Monte Carlo population distributions of 2 variance units. (a) 10 training samples per category and separation between the means in the Monte Carlo population distributions of 4 variance units. (a) 100 training samples per category and separation between the means in the Monte Carlo population distributions of 2 variance units. (a) 100 training samples per category and separation between the means in the Monte Carlo population distributions of 4 variance units. The dotted lines indicate the $C_{\text{llr}}$ values given the reference likelihood ratio values. Note that the $y$-axis scale is different for each panel.

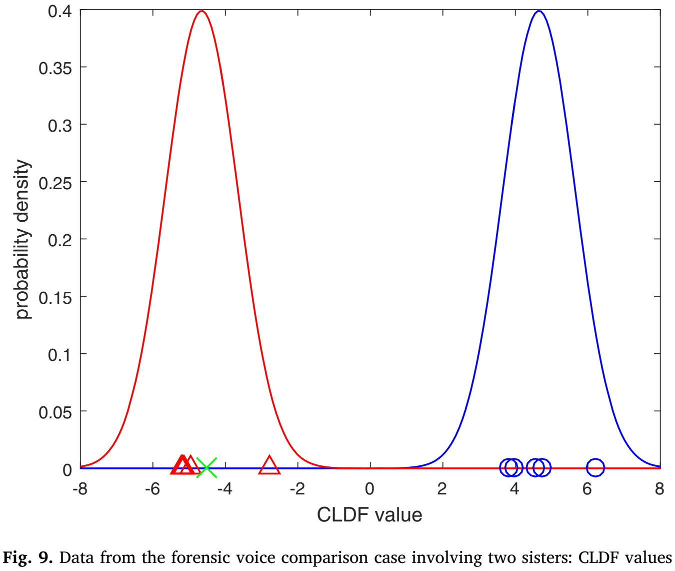

**Fig. 9.** Data from the forensic voice comparison case involving two sisters: CLDF values for the elder-sister recordings (blue circles), younger-sister recordings (red triangles), and the questioned-speaker recording (green cross), along with pooled-variance Gaussian distributions fitted to the known-speaker data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

bounds and the slopes of the regularized logistic regression functions are relatively small, although they both become more precise (compare Fig. 3c and d with Fig. 4c and d). The latter procedures are conservative in not increasing the magnitudes of log likelihood ratios unless there is sufficient data to support those larger magnitude values. Given a larger separation even with a relatively large amount of data (Fig. 6), the ELUB and regularized logistic regression procedures are more conservative than the LDA and Bayesian procedures, either by imposing bounds or by having a substantially shallower slope. In these examples we known that the greater separation is due to a change in the Monte Carlo population, not due to sampling variability, but in a single instance of real sample data we would not know whether the sample distributions were close to or far from the population distributions. The conservativeness of the ELUB and regularized logistic regression procedures may therefore be considered desirable.

In addition to the graphical representations of system performance described above, numeric metrics of system performance were calculated. A single test set consisting of 1000 simulated data points from each category was generated for each condition, and likelihood ratios or Bayes factors were calculated for these test data using each training sample set and each of the four procedures. The root mean square (RMS) error between the calculated log likelihood ratio or log Bayes factors values and the log reference values was calculated, as was the log likelihood ratio cost ($C_{\text{llr}}$, [6,15,46–48]). Each metric can be considered a quantifier of a different form of accuracy. For both metrics the better the performance, the lower the value. The boxplots in Figs. 7 and 8 show the distributions of the RMS error and $C_{\text{llr}}$ metrics respectively, calculated over the 1000 sets of training data.

Across all condition, and for both metrics, the Bayesian procedure outperformed the other procedures or was not substantially worse than the best performing system for a condition. When the amount of training data was large, LDA outperformed the other procedures on both metrics or was not substantially worse than the best performing system for a condition. The performance of the regularized logistic regression procedure on these metrics was generally not as good as for the

**Table 1**
Likelihood ratio and Bayes factor results $p(x|H_E)/p(x|H_Y)$ for the four procedures applied to the data from the elder- versus younger-sister case.

| Speaker | Y | Y | Y | Y | Y | E | E | E | E | E | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LDA | $2 \times 10^{-20}$ | $3 \times 10^{-20}$ | $6 \times 10^{-24}$ | $5 \times 10^{-40}$ | $2 \times 10^{-22}$ | $1 \times 10^{12}$ | $2 \times 10^{8}$ | $1 \times 10^{16}$ | $5 \times 10^{16}$ | $3 \times 10^{40}$ | $6 \times 10^{-19}$ |
| Bayesian | $\frac{1}{1369}$ | $\frac{1}{1349}$ | $\frac{1}{1738}$ | $\frac{1}{567}$ | $\frac{1}{300}$ | 130 | 25 | 774 | 878 | 4876 | $\frac{1}{2439}$ |
| ELUB | $\frac{1}{5}$ | $\frac{1}{5}$ | $\frac{1}{5}$ | $\frac{1}{5}$ | $\frac{1}{5}$ | 5 | 5 | 5 | 5 | 5 | $\frac{1}{5}$ |
| LogReg | $\frac{1}{5.7}$ | $\frac{1}{5.7}$ | $\frac{1}{7.4}$ | $\frac{1}{26.6}$ | $\frac{1}{2.0}$ | 2.2 | 1.6 | 4.2 | 4.6 | 9.6 | $\frac{1}{4.6}$ |

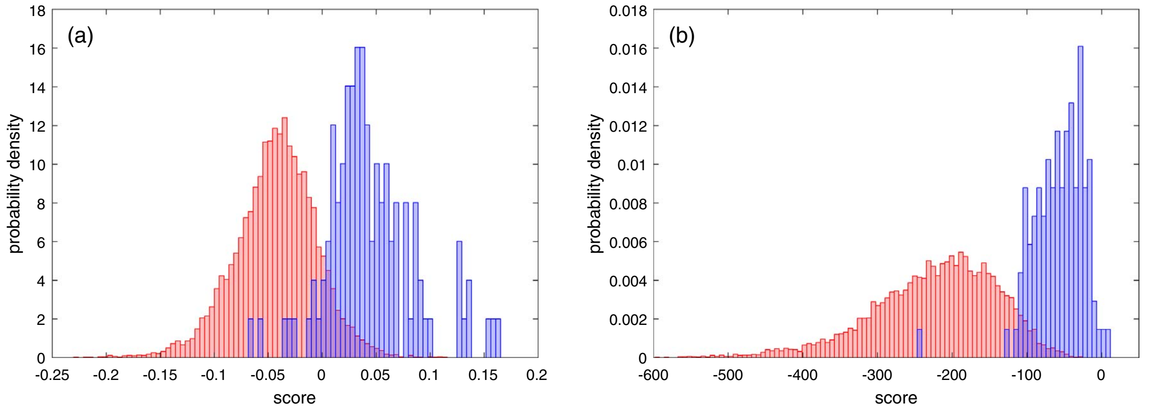

**Fig. 10.** Histograms of score data from (a) the GMM-UBM forensic voice comparison system, and (b) the i-vector PLDA forensic voice comparison system. Both systems were applied to the same voice-recording data.

Bayesian procedure. The ELUB procedure usually performed the worst on these metrics.

It should, however, be borne in mind that the training and test data were drawn from Monte Carlo population distributions that met the LDA and Bayesian procedures' assumption of Gaussian distributions with equal variance. The regularized logistic regression procedure could have advantages when these assumptions are violated. Based on the reported results using these metrics, the Bayesian procedure would be preferred. Ones utilities for casework, however, could warrant the acceptance of some decrease in accuracy in exchange for decreasing the probability of overestimating strength of evidence, leading one to prefer the ELUB or the regularized logistic regression procedure.

## 5. Tests using real data

The previous section explored the behaviour of the different score to likelihood ratio conversion procedures using simulated data. In order to explore the generalizability to conditions more reflective of casework, the present section applies the procedures to real data from comparisons of voice recordings, face images, and glass fragments. In order to keep the focus on the score to likelihood ratio conversion procedures, we will only briefly describe the data and systems which were used to generate the scores. Readers interested in more detailed descriptions are directed to the cited papers. For simplicity, in describing the performance of the different procedures on the different data sets, we will focus on accuracy and not report results related to precision.

### *5.1. Forensic comparison of voice recordings: sisters case*

The first set of real data comes from a forensic voice comparison case previously reported in [49]. These data are not actually scores, but features (mel frequency cepstral coefficients, MFCCs) projected down to a single dimension using a canonical linear discriminant function (CLDF). The hypotheses in this case were that the voice on the questioned-speaker mobile-telephone recording (Q) was either the elder (E) or the younger (Y) of two sisters: $p(x_Q|H_E)/p(x_Q|H_Y)$. The sisters were cooperative, and 5 recordings were made of each sister using the same mobile telephone as had been used to make the questioned-speaker recording.

Fig. 9 shows the CLDF values for the elder sister's recordings (blue circles), younger sister's recordings (red triangles), and the questioned-speaker recording (green cross), along with pooled-variance Gaussian distributions fitted to the known-speaker data.

Table 1 shows the results of using the different procedures to calculate likelihood ratios and Bayes factors on these feature data. Results for recordings known to be of either the elder or the younger sister were obtained using leave-one-out cross validation. The LDA procedure resulted in extremely large magnitude log likelihood ratio values, that would be difficult to justify give any realistic amount of training data, let alone 5 recordings per speaker. The Bayesian procedure resulted in smaller log Bayes factor values,[17] but still reaching values that may be difficult to justify given the small amount of training data. ELUB gave very conservative results. Since there was complete separation in the training data, the bounds simplified to the number of training data points for the least likely category. Regularized logistic regression resulted in likelihood ratio values around the same size as those from ELUB, but which could be up to several times larger or several times smaller. Given the small amount of training data (5 data points per category), the relative effect of the regularization (5 pseudodata points) was large.

### *5.2. Forensic comparison of voice recordings: scores from GMM-UBM and i-vector PLDA systems*

The next sets of real data come from a forensic voice comparison case previously reported in [50,51 ch 4]. The questioned-speaker recording was of a landline telephone call. It included background office noise and was saved in a compressed format. The known-speaker

[17] The Bayes factor values reported here are closer to 1 than those reported in [46]. The version of the formula used for calculating the Bayes factors in [46] included what we now believe to be an error with respect to how to account for the pooled calculation of variance. The version used in the present paper (Eq. 6) is, we believe, correct.

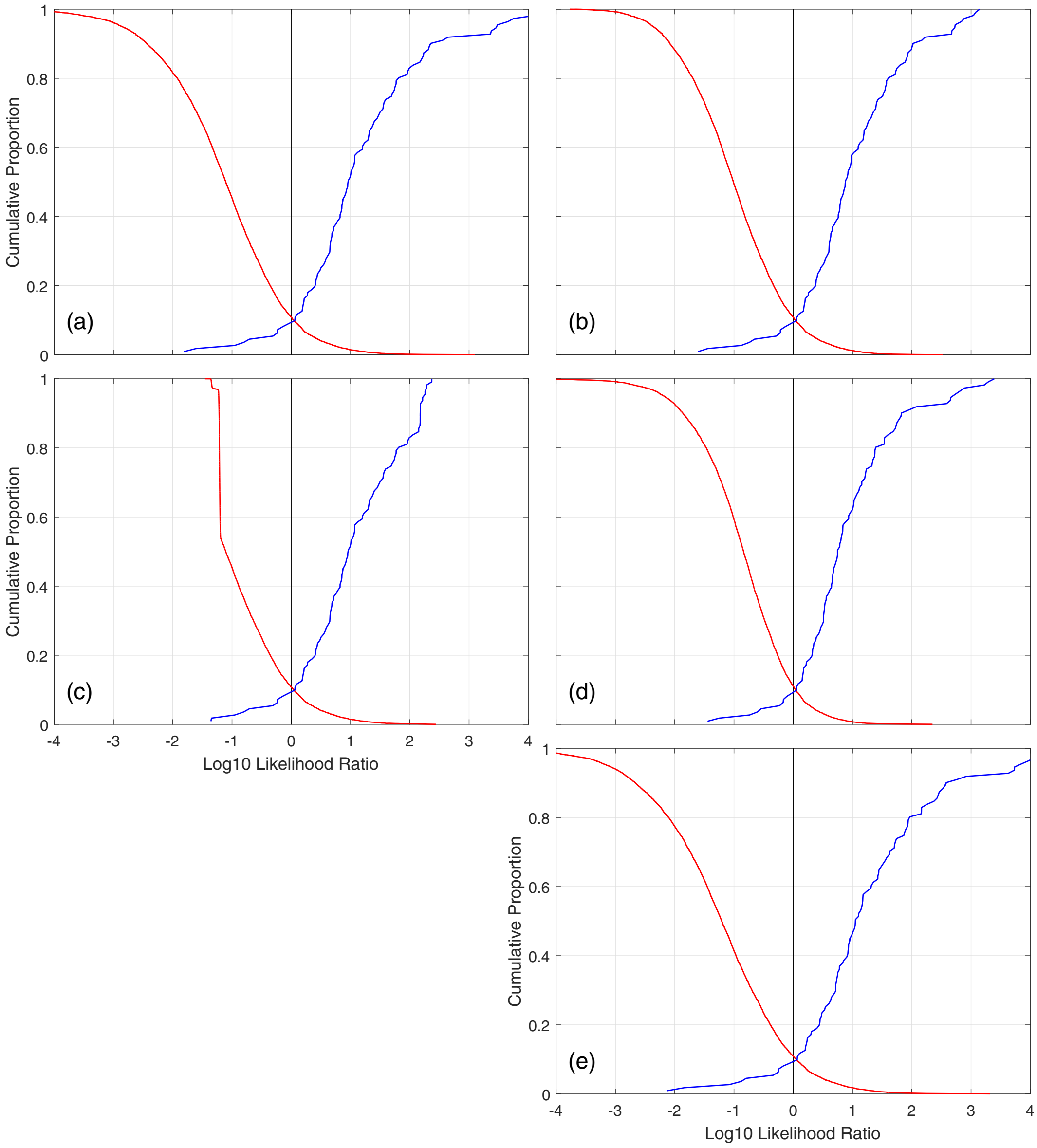

**Fig. 11.** Tippett plots resulting from fitting the score to log likelihood ratio or score to log Bayes factor transformation functions to the score data from the GMM-UBM forensic voice comparison system. (a) LDA. (b) Bayesian procedure with uninformative Jeffreys reference priors. (c) ELUB applied to LDA output. (d) Regularized logistic regression. (e) Non-regularized logistic regression.

recording was of a police interview. It had substantial reverberation and background ventilation system noise. One set of scores was generated using a Gaussian mixture model universal background model (GMM-UBM) system, and the other set was generated using an i-vector PLDA system. Both systems were trained and tested using the same feature data. Feature data were MFCC + deltas. Data from recordings of 105 speakers were used for training the first-stage model, and data from recordings of an additional 61 speakers were used to train the second-stage model and for testing. There were multiple recordings of each speaker resulting in a total of 111 same-speaker scores and 9720 different-speaker scores. A cross-validation procedure was used to avoid training and testing on the same data.[18]

Histograms of the different-speaker and same-speaker scores from the GMM-UBM and the i-vector PLDA systems are shown in Fig. 10a and Fig. 10b respectively. For the scores from the GMM-UBM system, the assumptions of normality and equal variance appear to be met. For the scores from the i-vector PLDA system, the different-speaker scores have a slight skew, and the assumption of equal variance is clearly violated.

The results of using the different procedures to convert the scores to likelihood ratios are shown as Tippett plots [48,52,53] in Fig. 11 (GMM-UBM scores) and Fig. 12 (i-vector PLDA scores), and as $C_{\text{llr}}$ values in Table 2. Results from a non-regularized logistic regression model are also included.

For the GMM-UBM scores, which met the assumptions of normality and equal variance, LDA gave the best results in terms of $C_{\text{llr}}$. The Bayesian procedure and regularized logistic regression effected shrinkage, which was greatest for the latter. Shrinkage came at the cost of a decrease in accuracy (an increase in $C_{\text{llr}}$).

For the i-vector PLDA scores, which violated the assumptions of normality and equal variance, logistic regression outperformed LDA

[18] For example, to test a same-speaker comparison for Speaker 1, all scores based on comparisons which included a recording of Speaker 1 were held out from the training data, and to test a different-speaker comparison for Speaker 1 versus Speaker 2, all scores based on comparisons which included a recording of either Speaker 1 or Speaker 2 were held out from the training data. Mutatis mutandis for cross-validated testing of the face-image and glass-fragment scores below.

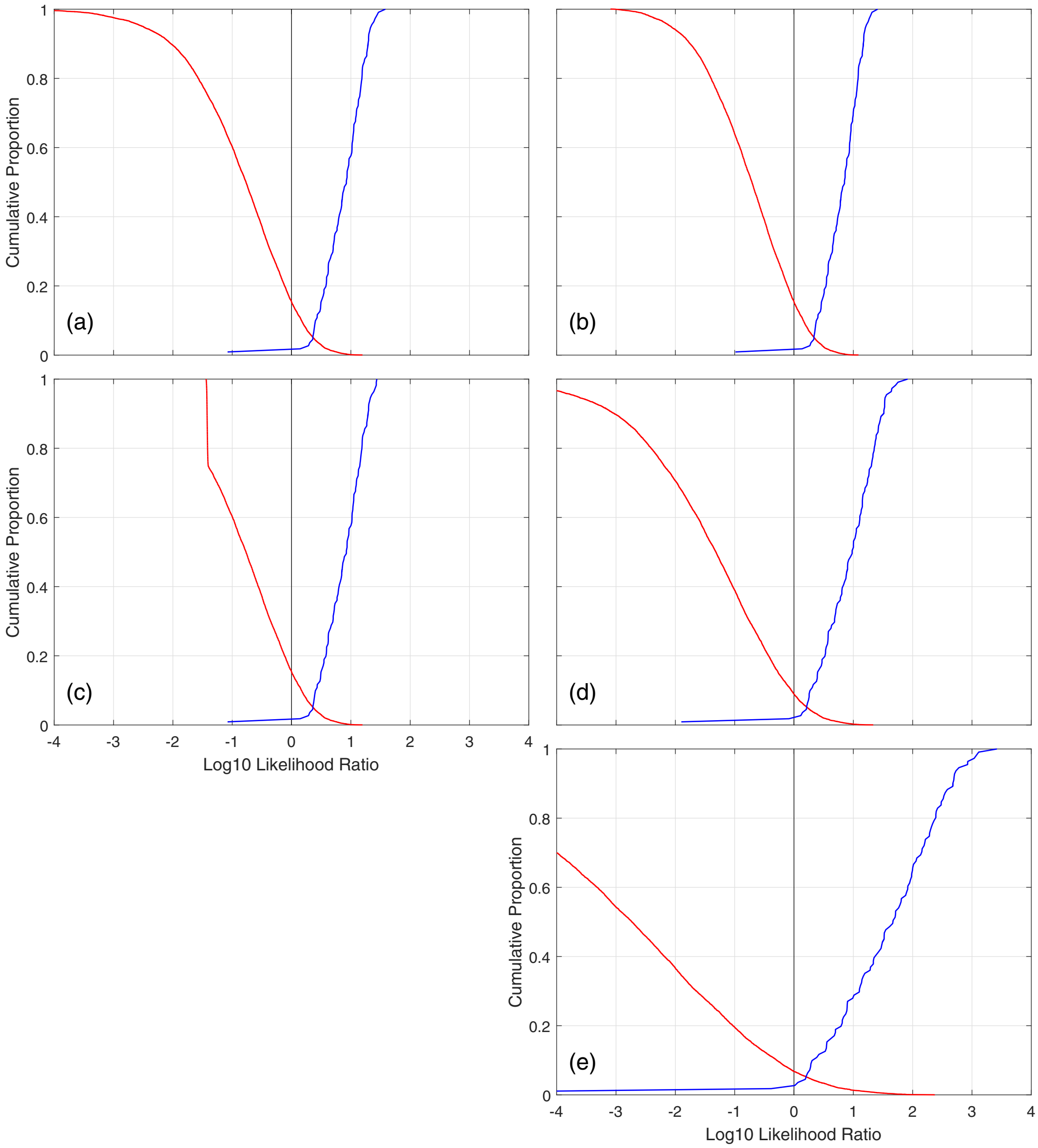

**Fig. 12.** Tippett plots resulting from fitting the score to log likelihood ratio or score to log Bayes factor transformation functions to the score data from the i-vector PLDA forensic voice comparison system. (a) LDA. (b) Bayesian procedure with uninformative Jeffreys reference priors. (c) ELUB applied to LDA output. (d) Regularized logistic regression. (e) Non-regularized logistic regression.

**Table 2**
$C_{\text{llr}}$ values for the results of each procedure when applied to scores from the GMM-UBM and the i-vector PLDA forensic voice comparison systems.

| System | GMM-UBM | i-vector PLDA |
|---|---|---|
| LDA | 0.410 | 0.357 |
| Bayesian | 0.413 | 0.374 |
| ELUB | 0.412 | 0.361 |
| LogReg regularized | 0.430 | 0.290 |
| LogReg non-regularized | 0.413 | 0.261 |

and the Bayesian procedure in terms of $C_{\text{llr}}$. Regularizing the logistic regression model effected substantial shrinkage. Shrinkage came at the cost of a decrease in accuracy (an increase in $C_{\text{llr}}$).

We also tested regularized and non-regularized logistic regression fusion of the GMM-UBM scores and the i-vector PLDA scores.[19] Performance was very slightly better than that of the i-vector system alone: $C_{\text{llr}}$ was 0.286 and 0.241 for the regularized and non-regularized procedures respectively.

### 5.3. Forensic comparison of face images

The next set of real data is scores from comparisons of face images previously described in [54,55]. The images of 225 faces (8 images per face) came from the XM2VTS database [56]. These were frontal face images from high-definition video sequences with controlled lighting and background. Such conditions could potentially exist at a police station or an immigration post, and could be realistic for comparison with previously collected mug shots or passport photographs, but these data are not reflective of more challenging forensic conditions such as CCTV surveillance images.

Feature data were extracted from 40 × 32 pixel images, and consisted of two-dimensional discrete cosine transform coefficient values (2D-DCTs) fitted within an 8 × 8 pixel sliding window. This resulted in a sequence of 35 18-dimensional feature vectors per image. Scores were

[19] Each set of scores was first independently normalized to a mean of 0 and variance of 1 (without reference to category).

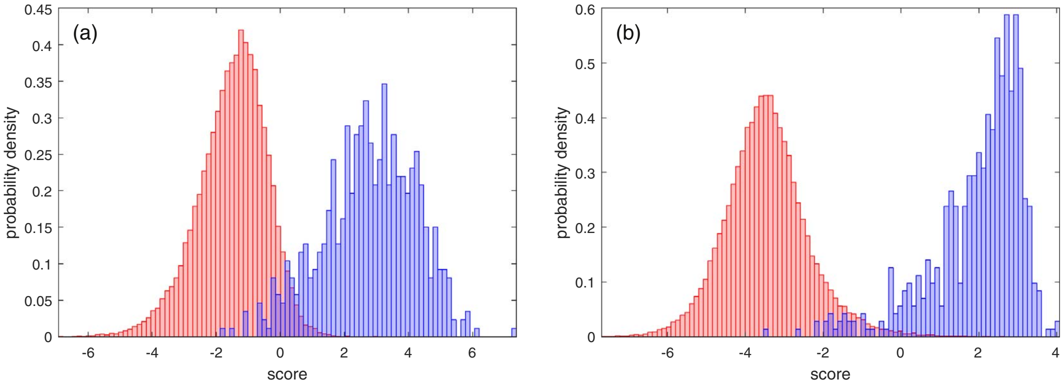

**Fig. 13.** Histograms of score data from (a) the GMM-UBM face-image comparison system, and (b) the MLP face-image comparison system. Both systems were applied to the same face-image data.

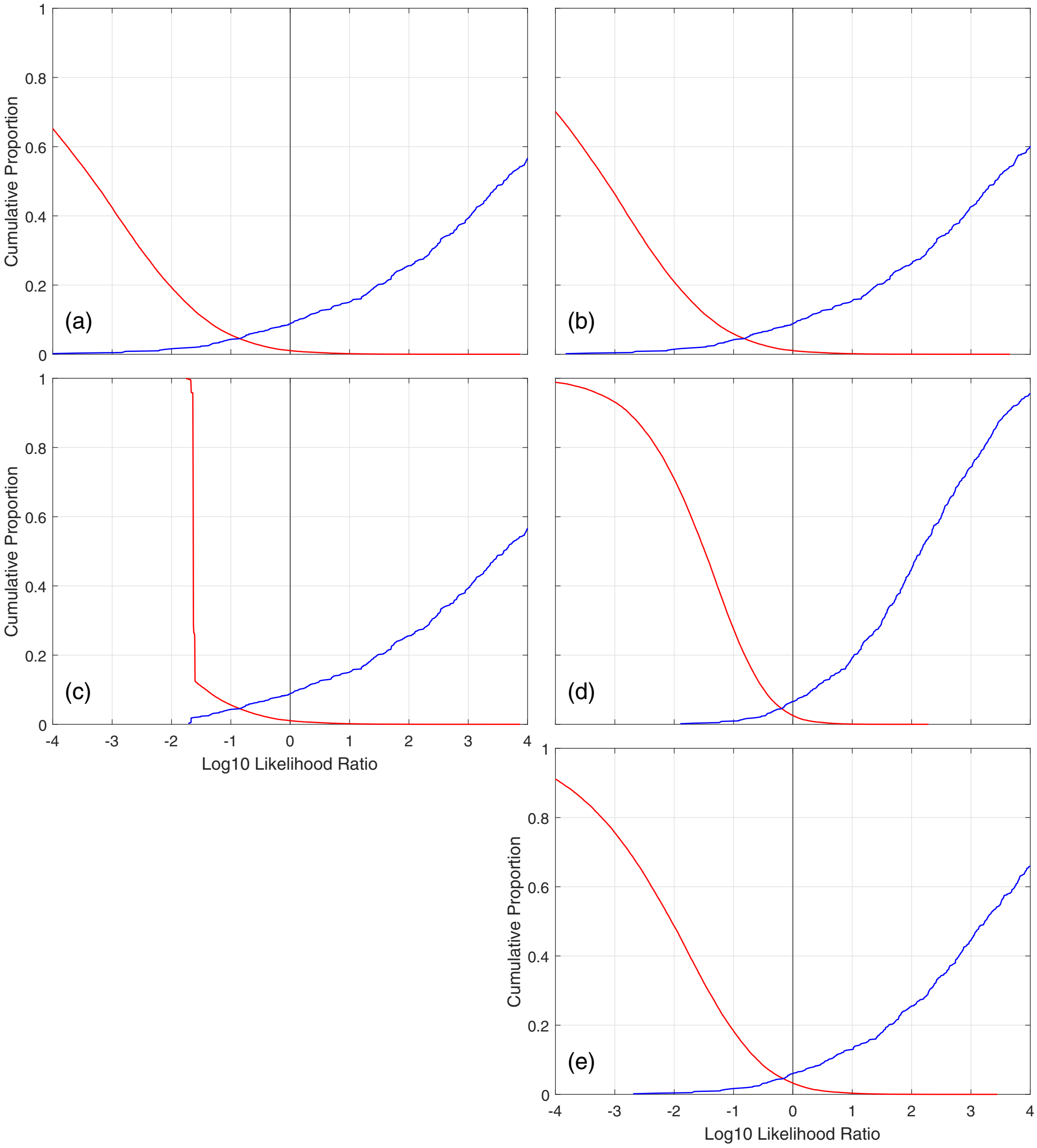

**Fig. 14.** Tippett plots resulting from fitting the score to log likelihood ratio or score to log Bayes factor transformation functions to the score data from the GMM-UBM face-image comparison system. (a) LDA. (b) Bayesian procedure with uninformative Jeffreys reference priors. (c) ELUB applied to LDA output. (d) Regularized logistic regression. (e) Non-regularized logistic regression.

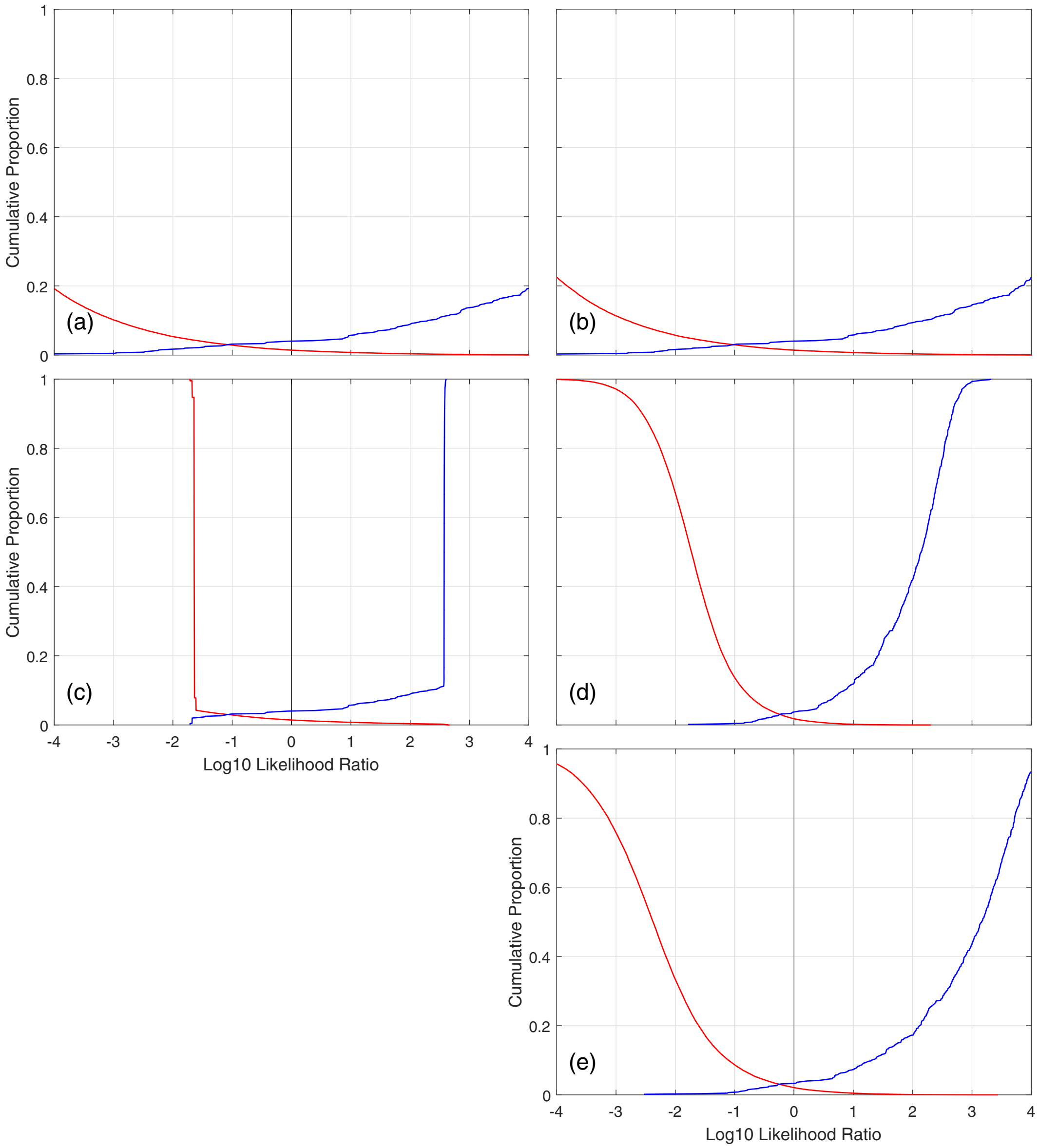

**Fig. 15.** Tippett plots resulting from fitting the score to log likelihood ratio or score to log Bayes factor transformation functions to the score data from the MLP face-image comparison system. (a) LDA. (b) Bayesian procedure with uninformative Jeffreys reference priors. (c) ELUB applied to LDA output. (d) Regularized logistic regression. (e) Non-regularized logistic regression.

**Table 3**
$C_{\text{llr}}$ values for the results of each procedure when applied to scores from the GMM-UBM and the MLP face-image comparison systems.

| System: | GMM-UBM | MLP |
|---|---|---|
| LDA | 0.220 | 0.167 |
| Bayesian | 0.215 | 0.159 |
| ELUB | 0.206 | 0.143 |
| LogReg regularized | 0.181 | 0.115 |
| LogReg non-regularized | 0.169 | 0.104 |

generated using two systems. The first system was a GMM-UBM system with 64 Gaussian components. The second system was a multi-layer perceptron (MLP), three layers with 32 nodes in the hidden (middle) layer.[20] The systems are described in [57]. The protocol for model training and score calculation is described in [56].[21] It resulted in 600 same-origin scores (from 200 faces) and 40,000 different-origin scores (comparing 200 face models with images of 25 other faces). A cross-validation procedure was used to avoid training and testing on the same data.

Histograms of the different-face and same-face scores from the GMM-UBM and the MLP systems are shown in Fig. 13a and Fig. 13b respectively. Both deviate from the assumptions of Gaussian distributions with equal variance. The assumption of equal variance is clearly violated, and the same-origin scores are skewed, the skewness being more pronounced for the MLP system.

The results of using the different procedures to convert the scores to likelihood ratios are shown as Tippett plots in Fig. 14 (GMM-UBM scores) and Fig. 15 (MLP scores), and as $C_{\text{llr}}$ values in Table 3. Results

[20] The MLP system was a closed box with score output scaled to the range $-1$ to $+1$. An inverse hyperbolic tangent function was applied to rescale the scores in the form of log likelihood ratios.

[21] The protocol was not designed for forensic application, and we would have used the data differently if starting from scratch, but it suffices for the purpose of examining the behaviour of the different score to likelihood ratio conversion procedures on different score distributions.

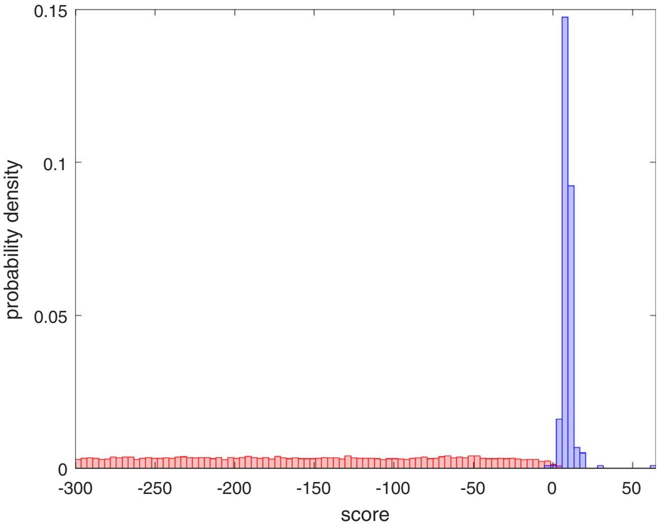

**Fig. 16.** Histograms of score data from the 2-level MVKD system applied to the glass-fragment data.

from a non-regularized logistic regression model are also included.

From the Tippett plots it can be seen that the LDA and Bayesian procedures, which depend on the assumption of Gaussian distributions with equal variance, produced clearly biassed results (both the same-origin and different-origin curves are too far to the left). Both these procedures applied to both sets of scores also produced some very large magnitude log likelihood ratio values (beyond the −4 to +4 range plotted on the Tippett plots) which would be difficult to justify given the amount of training data. Bias was much smaller for the logistic regression results, and the range of log likelihood ratio values produced were more reasonable. Logistic regression gave the best results in terms of $C_{\mathrm{llr}}$, and regularizing the logistic regression model effected shrinkage, with a concomitant decrease in accuracy (an increase in $C_{\mathrm{llr}}$).

### 5.4. Forensic comparison of glass fragments

The final set of real data is scores from comparisons of glass fragments previously reported in [58]. The feature data were measurements of the concentrations of 10 trace elements made using Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS). The hypotheses tested were $H_s$: the glass fragment found on the suspect's garment is from the window at the crime scene, versus $H_d$: the glass fragment found on the suspect's garment is from a different source in the relevant population. The sample representing the relevant population was a sample of known-source glass fragments collected during casework over a 10 year period. The sample contained multiple fragments from each of 979 sources. Data from 659 sources were used for training the first-stage feature to score model, and data from 320 sources used for training and testing the second-stage score to likelihood ratio conversion procedures. A cross-validation procedure was used to avoid training and testing on the same data.

The feature to score model was a two-level multivariate kernel density (MVKD) model [59,60]. A total of 320 same-origin scores and 51,040 different-origin scores were calculated. 41,108 of the different-origin scores had values which were smaller than the smallest encodable value in the software used to calculate them (hereinafter “very small score values”, note that these were not smaller than the smallest encodable value in the software used for the second-stage model). We ran two versions of the tests on the procedures for converting scores to likelihood ratios. In one version the very small score values were excluded from both training and testing, and only the remaining 10,032 different-origin scores were used. In the other version, for both training and testing, the very small score values were replaced with the same value as the smallest score value in the data that had been encoded. The latter version was included to allow for comparison with the results reported in [58]. In the present research, the interest is in the relative performance of the different score to likelihood ratio conversion procedures, not on the best way to deal with this particular data set.

Histograms of the different-speaker and same-speaker scores are shown in Fig. 16 (excluding the very small score values). These scores exhibit an extreme case of non-equal variance. The results of using the different procedures to convert the scores to likelihood ratios are shown as Tippett plots in Fig. 17 and as $C_{\mathrm{llr}}$ values in Table 4. Results from a non-regularized logistic regression model are also included. A Tippett plot and $C_{\mathrm{llr}}$ value based on the results reported in [58] are also included. In [58] ELUB was applied to a second-stage model that was tailored to be a good fit to these particular data. A kernel density model was used for the different-origin scores and a double exponential decay was used for the same-origin scores.

The difference in the lower bound visible in panels (c) and (e) ELUB of Fig. 17, ELUB applied to LDA and ELUB applied to the output of the tailored model respectively, is due to the former being based on calculations excluding the very small score values and the latter being based on calculations including the very small score values.

LDA and the Bayesian procedure depend on assumptions of Gaussian distributions with equal variance and, as expected, performed poorly on these data. Non-regularized logistic regression gave by far the best performance in terms of $C_{\mathrm{llr}}$, and regularization effected substantial shrinkage with a concomitant loss in accuracy (although still a substantially lower $C_{\mathrm{llr}}$ than the other procedures). All procedures except non-regularized logistic regression and the tailored model had a substantial bias in their output.

## 6. Discussion and conclusion

We have tested several procedures for converting scores to interpretable likelihood ratios or Bayes factors. Three of these procedures (a Bayesian procedure with uninformative Jeffreys reference priors, a procedure which imposes empirical lower and upper bounds ELUB, and regularized logistic regression) involve some form of shrinkage so as to address concerns about imprecision and overstating strength of evidence. Which score to likelihood ratio conversion procedure one prefers will depend on one's utilities. Based on the results of fitting the different procedures to real data, we prefer regularized logistic regression for the following reasons.

When the amount of sample data is small, we prefer regularized logistic regression because it induces a degree of shrinkage which results in likelihood ratio values which would seem defensible given the amount of training data. The values are around the same magnitude as the bounds from the ELUB procedure, but do not suffer from ELUB's cliff-edge effects. We prefer regularized logistic regression over the Bayesian procedure as the latter was non-monotonic and produced relatively large Bayes factor values that may be difficult to justify given the amount of sample data. When the distributions of the same-origin and different-origin scores deviate from Gaussian distributions with equal variance, we prefer logistic regression since, unlike the Bayesian procedure, it does not depend on such distributional assumptions. Since deviations from these distributional assumptions are common in score data, we prefer regularized logistic regression in general.

It may be that for a Bayesian approach other choices with respect to the model and uninformative priors used would have produced more favourable results. For simplicity in the present paper, we only tested a model which assumed equal-variance Gaussian distributions and used Jeffreys reference priors.

Given the extreme difference in variance between the same-origin and different-origin scores for the glass data, only the non-regularized logistic regression procedure and the tailored model produced unbiased output. The fact that the output of the regularized logistic regression procedure was biassed indicates that in this extreme case the form of

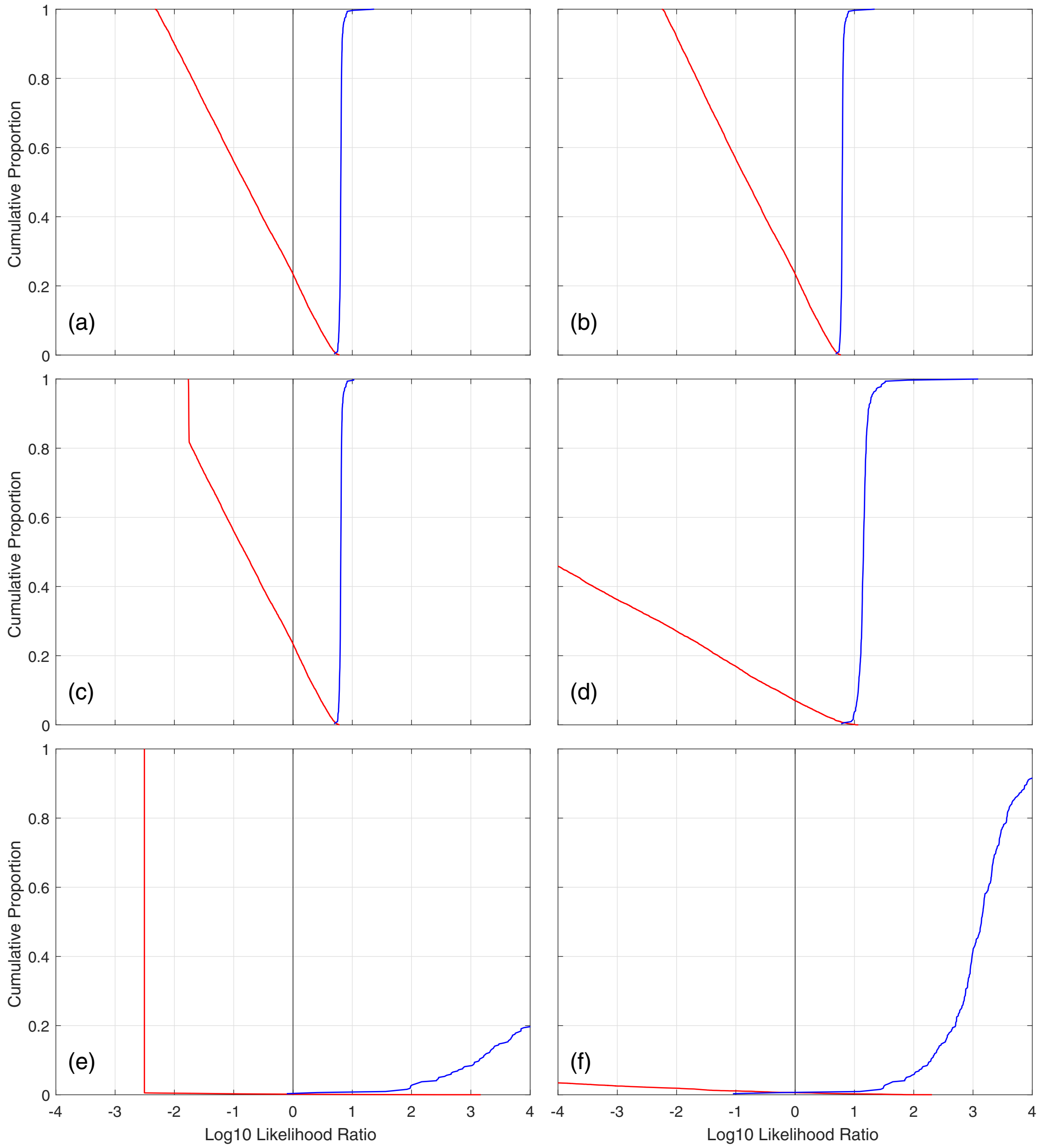

**Fig. 17.** Tippett plots resulting from fitting the score to log likelihood ratio or score to log Bayes factor transformation functions to the score data from the 2-level MVKD system applied to the glass-fragment data. (a) LDA. (b) Bayesian procedure with uninformative Jeffreys reference priors. (c) ELUB applied to LDA output. (d) Regularized logistic regression. (e) ELUB applied to the output of a tailored model. (f) Non-regularized logistic regression. For all panels except (e) the results presented are those when very small score values were excluded from both training and testing.

**Table 4**

$C_{\text{llr}}$ values for the results of each procedure when applied to MVKD scores from the comparison of glass fragments. Plus ELUB applied to output from the tailored model used in [58]. Very small scores were either included or excluded from both the training and the test data.

| Very small scores | Excluded | Included |
|---|---|---|
| LDA | 0.391 | 0.242 |
| Bayesian | 0.413 | 0.236 |
| ELUB on LDA | 0.389 | 0.244 |
| ELUB on tailored model | – | 0.009 |
| LogReg regularized | 0.104 | 0.111 |
| LogReg non-regularized | 0.021 | 0.007 |

regularization used affected not only the slope but also the intercept of the score to likelihood ratio transformation model. A potential solution could be to use a more common form of regularization which only targets the slope coefficient and does so explicitly (see [37] §4.4.4). Such a model would not be applicable for addressing numerical problems associated with complete or near complete separation (and would not directly have a pseudodata interpretation), but may offer a better solution when complete separation is not an issue and there is a relatively large amount of data but same-origin and different-origin scores have very different variances.[22]

By design, the procedures including shrinkage reduced overestimation of the strength of evidence, and hence on average

[22] Postscript: Upon reflection, we realize that the results are as would be expected given logistic regression's minimization of the deviance statistic. The different-origin log-base-10 likelihood ratio values have a wide range, but the same-origin log-base-10 likelihood ratio values are in a narrow range mostly between +1 and +1.5. Moving them to say 0 to +0.5 would not represent better calibration. Looking at the histogram of score distributions one would actually expect higher values for same-speaker likelihood ratios. The fundamental problem lies in the disparate distributions of the same-origin and different-origin scores, but we think our practical error in attempting to address this problem was actually choosing too high a degree of regularization, and that it would have been better to use a smaller degree of regularization giving results somewhere between those shown in Fig. 17d and f.

underestimated the strength of evidence (see the Monte Carlo simulation results). Thus, these procedures did not produce results that were as accurate as their non-shrinkage counterparts. The procedures including shrinkage did not produce $C_{\text{llr}}$ values that were as low as their non-shrinkage counterparts. The $C_{\text{llr}}$ metric, however, does not take account of the utility of not overstating the strength of evidence. A metric which does take this into account could potentially be developed, but we have not attempted to do so here. The challenge would be to develop a principled metric which did not simply arbitrarily favour the procedure one preferred a priori.

A major disadvantage of the regularized logistic regression procedure is that one has to specify the weight ($\kappa^{\psi}$) of the uninformative "prior" distribution, and the choice is essentially arbitrary. As pointed out, however, the specification of particular uninformative priors for a Bayesian procedure also requires a choice which arguably is also arbitrary. One should make it clear what weight one has used, one should choose the weight before looking at the data, and one should not change the weight after having examined the data and especially not after one has obtained the likelihood ratio results. In hindsight, in the results reported in the present paper using $\kappa^{\psi} = 5$ may have induced a greater degree of shrinkage than one might prefer. Following our own advice, we have not rerun the experiments using a smaller value for $\kappa^{\psi}$. Before commencing future research or casework, one could decide to use a smaller value (e.g., $\kappa^{\psi} = 2$, or $\kappa^{\psi} = 1$), declare that value before beginning and then proceed to use that value. We are not presently in a position to develop an optimization strategy for $\kappa^{\psi}$ since that would require an optimization objective such as the metric discussed in the previous paragraph.

In the research reported in the present paper, we have tested a number of different procedures for converting scores to likelihood ratio values. In casework, one should not test multiple procedures and then use the one which gives the best results. Doing so would over-optimize on the test data, and the best performing procedure on the particular test data may not be the best performing procedure on new data, e.g., on the actual known- and questioned-origin data in the case. With respect to admissibility, the issue is not whether the system used is better than all other systems, but whether the performance of the system used is adequate under the conditions of the case. One should choose a system before looking at the data, and one should test and use that system. One should choose a system which based on previous research one believes will have an adequate level of performance. One may optimize the system using data other than the test data, but one should not over-optimize on those data. After optimization, one should empirically assess the performance of the chosen system using previously-unseen test data that reflect the conditions of the known- and questioned-origin data in the case under investigation.

## Acknowledgements

Thanks to the following who provided or helped provide data: Cuiling Zhang, Ewald Enzinger, Andrew van Es, Peter Vergeer.

This work was partially funded by a Simons Foundation Visiting Fellowship. The first author would like to thank the Isaac Newton Institute for Mathematical Sciences for its hospitality during the programme Probability and Statistics in Forensic Science which was supported by EPSRC Grant Number EP/K032208/1. Thanks to several participants in the programme who provided advice and feedback.

All opinions expressed are those of the authors and, unless explicitly stated otherwise, do not necessarily reflect the opinions or policies of any other individuals or organizations.

## References

[1] J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Sci. Justice 42 (2002) 29–37, http://dx.doi.org/10.1016/S1355-0306(02)71794-2.

[2] S. Hancock, R. Morgan-Smith, J.S. Buckleton, The interpretation of shoeprint comparison class correspondences, Sci. Justice 52 (2012) 243–248, http://dx.doi.org/10.1016/j.scijus.2012.06.002.

[3] D.J. Balding, Estimating products in forensic identification using DNA profiles, J. Am. Stat. Assoc. 90 (1995) 839–844, http://dx.doi.org/10.2307/2291317.

[4] J.M. Curran, An introduction to Bayesian credible intervals for sampling error in DNA profiles, Law Prob. Risk 4 (2005) 115–126, http://dx.doi.org/10.1093/lpr/mgi009.

[5] G.W. Beecham, B.S. Weir, Confidence interval of the likelihood ratio associated with mixed stain DNA evidence, J. Forensic Sci. 56 (2011) S166–S171, http://dx.doi.org/10.1111/j.1556-4029.2010.01600.x.

[6] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice 51 (2011) 91–98, http://dx.doi.org/10.1016/j.scijus.2011.03.002.

[7] A. Nordgaard, T. Höglund, Assessment of approximate likelihood ratios from continuous distributions: a case study of digital camera identification, J. Forensic Sci. 56 (2011) 390–402, http://dx.doi.org/10.1111/j.1556-4029.2010.01665.x.

[8] I. Alberink, A. Bolck, S. Menges, Posterior likelihood ratios for evaluation of forensic trace evidence given a two-level model on the data, J. Appl. Stat. 40 (2013) 2579–2600, http://dx.doi.org/10.1080/02664763.2013.822056.

[9] A.J. Leegwater, D. Meuwly, M. Sjerps, P. Vergeer, I. Alberink, Performance study of a score-based likelihood ratio system for forensic fingermark comparison, J. Forensic Sci. 62 (2017) 626–640, http://dx.doi.org/10.1111/1556-4029.13339.

[10] F. Taroni, S. Bozza, A. Biedermann, C.G.G. Aitken, Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio, Law Prob. Risk 15 (2016) 1–16, http://dx.doi.org/10.1093/lpr/mgv008.

[11] C.E.H. Berger, K. Slooten, The LR does not exist, Sci. Justice 56 (2016) 388–391, http://dx.doi.org/10.1016/j.scijus.2016.06.005.

[12] G.S. Morrison, What should a forensic practitioner's likelihood ratio be? II, Sci. Justice 57 (2017) 472–476, http://dx.doi.org/10.1016/j.scijus.2017.08.004.

[13] P. Vergeer, A. van Es, A. de Jongh, I. Alberink, R.D. Stoel, Numerical likelihood ratios outputted by LR systems are often based on extrapolation: when to stop extrapolating? Sci. Justice 56 (2016) 482–491, http://dx.doi.org/10.1016/j.scijus.2016.06.003.

[14] S. Pigeon, P. Druyts, P. Verlinde, Applying logistic regression to the fusion of the NIST99 1-speaker submissions, Digital Signal. Process. 10 (2000) 237–248, http://dx.doi.org/10.1006/dspr.1999.0358.

[15] J. González-Rodríguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-García, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Transact. Audio Speech Lang. Process. 15 (2007) 2104–2115, http://dx.doi.org/10.1109/TASL.2007.902747.

[16] G.S. Morrison, Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio, Aust. J. Forensic Sci. 45 (2013) 173–197, http://dx.doi.org/10.1080/00450618.2012.733025.

[17] A. Alexander, Drygajlo, Scoring and direct methods for the interpretation of evidence in forensic speaker recognition, Proceedings of Interspeech, 2004, pp. 2397–2400 http://www.isca-speech.org/archive/interspeech_2004/i04_2397.html.

[18] D. Ramos-Castro, Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems, PhD dissertation Universidad Autónoma de Madrid, Madrid, Spain, 2007, http://atvs.ii.uam.es/files/2007_11_28_thesis_daniel_ramos_searchable_v1.pdf.

[19] W. van Houten, I. Alberink, Z. Geradts, Implementation of the likelihood ratio framework for camera identification based on sensor noise patterns, Law, Prob. Risk 10 (2011) 149–159, http://dx.doi.org/10.1093/lpr/mgr006.

[20] A.B. Hepler, C.P. Saunders, L.J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Forensic Sci. Int. 219 (2012) 129–140, http://dx.doi.org/10.1016/j.forsciint.2011.12.009.

[21] J. Abraham, C. Champod, C. Lennard, C. Roux, Modern statistical models for forensic fingerprint examinations: a critical review, Forensic Sci. Int. 232 (2013) 131–150, http://dx.doi.org/10.1016/j.forsciint.2013.07.005.

[22] I. Alberink, A. de Jongh, C. Rodríguez, 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 (2014) 70–81, http://dx.doi.org/10.1111/1556-4029.12105.

[23] T. Ali, L.J. Spreeuwers, R.N.J. Veldhuis, D. Meuwly, Effect of calibration data on forensic likelihood ratio from a face recognition system, IET Biom. 3 (2014) 335–346, http://dx.doi.org/10.1049/iet-bmt.2014.0009.

[24] M.I. Mandasari, M. Günther, R. Wallace, R. Saeidi, S. Marcel, D.A. van Leeuwen, Score calibration in face recognition, IET Biom. 3 (2014) 246–256, http://dx.doi.org/10.1049/iet-bmt.2013.0066.

[25] A. Bolk, H. Ni, M. Lopatka, Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison, Law, Prob. Risk 14 (2015) 243–266, http://dx.doi.org/10.1093/lpr/mgv009.

[26] G.S. Morrison, E. Enzinger, Score based procedures for the calculation of forensic likelihood ratios – scores should take account of both similarity and typicality, Sci. Justice (2017), http://dx.doi.org/10.1016/j.scijus.2017.06.005.

[27] N. Brümmer, A. Swart, D. van Leeuwen, A comparison of linear and non-linear calibrations for speaker recognition, Proceedings of Odyssey, The Language and Speaker Recognition Workshop, 2014, pp. 14–18 http://www.isca-speech.org/archive/odyssey_2014/pdfs/31.pdf.

[28] G.S. Morrison, Calculation of forensic likelihood ratios: Use of Monte Carlo simulations to compare the output of score-based approaches with true likelihood-ratio values, (2015) http://arxiv.org/abs/1612.08165 (arXiv:1612.08165).

[29] J.J. Lucena-Molina, D. Ramos-Castro, J. González-Rodríguez, Performance of likelihood ratios considering bounds on the probability of observing misleading evidence, Law Prob. Risk 14 (2015) 175–192, http://dx.doi.org/10.1093/lpr/mgu022.

[30] G.S. Morrison, E. Enzinger, C. Zhang, Refining the relevant population in forensic voice comparison – a response to Hicks et alii (2015) the importance of distinguishing information from evidence/observations when formulating propositions, Sci. Justice 56 (2016) 492–497, http://dx.doi.org/10.1016/j.scijus.2016.07.002.

[31] W.R. Klecka, Discriminant Analysis, Sage, Beverly Hills CA, 1980, http://dx.doi.org/10.4135/9781412983938.

[32] D.A. van Leuwen, N. Brümmer, The distribution of calibrated likelihood-ratios in speaker recognition, Proceedings of Biometric Technologies in Forensic Science, 2013, pp. 24–29 http://www.ru.nl/clst/btfs/proceedings/ (last visited 18 May 2017).

[33] M.H. DeGroot, Optimal Statistical Decisions, Wiley, Hoboken NJ, 1970, http://dx.doi.org/10.1002/0471729000.

[34] K.P. Murphy, Conjugate Bayesian analysis of the Gaussian distribution, (2007) https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf (last visited 18 May 2017).

[35] N. Brümmer, A. Swart, Bayesian calibration for forensic evidence reporting, Proceedings of Interspeech, 2014, pp. 388–392 http://www.isca-speech.org/archive/interspeech_2014/i14_0388.html (last visited 18 May 2017).

[36] G.S. Morrison, J. Lindh, J.M. Curran, Likelihood ratio calculation for a disputed-utterance analysis with limited available data, Speech Comm. 58 (2014) 81–90, http://dx.doi.org/10.1016/j.specom.2013.11.004.

[37] T. Hastie, R. Tibshirani, J. Freidman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, 2nd ed., Springer, New York, 2009, http://dx.doi.org/10.1007/978-0-387-84858-7.

[38] P. McCullagh, J.A. Nelder, Generalized Linear Models, 2nd ed., Chapman & Hall, London, 1989.

[39] A. Agresti, Categorical Data Analysis, 3rd ed., Wiley, New York, 2013.

[40] S. Menard, Logistic Regression: From Introductory to Advanced Concepts and Applications, Sage, Thousand Oaks CA, 2010, http://dx.doi.org/10.4135/9781483348964.

[41] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, 3rd ed., Wiley, Hoboken NJ, 2013, http://dx.doi.org/10.1002/9781118548387.

[42] F.C. Pampel, Logistic Regression: A Primer, Sage, Thousand Oaks CA, 2002, http://dx.doi.org/10.4135/9781412984805.

[43] S. Menard, Applied Logistic Regression 2nd Ed, Sage, Thousand Oaks CA, 2002, http://dx.doi.org/10.4135/9781412983433.

[44] G.S. Morrison, A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: multvariate kernel density (MVKD) versus Gaussian mixture model – universal background model (GMM-UBM), Speech Comm. 53 (2011) 242–256, http://dx.doi.org/10.1016/j.specom.2010.09.005.

[45] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grézl, M. Karafiát, D.A. van Leeuwen, P. Matějka, P. Schwarz, A. Strasheim, Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006, IEEE Trans. Audio Speech Lang. Process. 15 (2007) 2072–2084, http://dx.doi.org/10.1109/TASL.2007.902870.

[46] N. Brümmer, J. du Preez, Application-independent evaluation of speaker detection, Comput. Speech Lang. 20 (2006) 230–275, http://dx.doi.org/10.1016/j.csl.2005.08.001.

[47] 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, http://dx.doi.org/10.1007/978-3-540-74200-5_19.

[48] D. Meuwly, D. Ramos, R. Haraksim, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Sci. Int. 276 (2017) 142–153, http://dx.doi.org/10.1016/j.forsciint.2016.03.048.

[49] C. Zhang, G.S. Morrison, E. Enzinger, Use of relevant data, quantitative measurements, and statistical models to calculate a likelihood ratio for a Chinese forensic voice comparison case involving two sisters, Forensic Sci. Int. 267 (2016) 115–124, http://dx.doi.org/10.1016/j.forsciint.2016.08.017.

[50] E. Enzinger, G.S. Morrison, F. Ochoa, A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case, Sci. Justice 56 (2016) 42–57, http://dx.doi.org/10.1016/j.scijus.2015.06.005.

[51] E. Enzinger, Implementation of Forensic Voice Comparison Within the New Paradigm for the Evaluation of Forensic Evidence, PhD dissertation University of New South Wales, Sydney, New South Wales, Australia, 2016, http://handle.unsw.edu.au/1959.4/55772.

[52] D. Meuwly, Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique, PhD dissertation University of Lausanne, 2001.

[53] G.S. Morrison, E. Enzinger, Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_eval_01*) – Introduction, Speech Comm. 85 (2016) 119–126, http://dx.doi.org/10.1016/j.specom.2016.07.006.

[54] N. Poh, A. Ross, W. Lee, J. Kittler, A user-specific and selective multimodal biometric fusion strategy by ranking subjects, Pattern Recogn. 46 (2013) 3341–3357, http://dx.doi.org/10.1016/j.patcog.2013.03.018.

[55] N. Poh, A. Ross, W. Lee, J. Kittler, Corrigendum to “a user-specific and selective multimodal biometric fusion strategy by ranking subjects” [pattern recognition 46 (2013) 3341–3357], Pattern Recogn. 47 (2014) 493, http://dx.doi.org/10.1016/j.patcog.2013.08.001.

[56] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, E. Mayoraz, Comparison of face verification results on the XM2VTS database, Proceedings of the 15th International Conference on Pattern Recognition, 4 2000, pp. 858–863, , http://dx.doi.org/10.1109/ICPR.2000.903052.

[57] F. Cardinaux, C. Sanderson, S. Marcel, Comparison of MLP and GMM classifiers for face verification on XM2VTS, Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication, 2003, pp. 911–920, , http://dx.doi.org/10.1007/–3-540-44887-X_106.

[58] A. van Es, W. Wiarda, M. Hordijk, I. Alberink, P. Vergeer, Implementation and assessment of a likelihood ratio approach for the evaluation of LA-ICP-MS evidence in forensic glass analysis, Sci. Justice 57 (2017) 181–192, http://dx.doi.org/10.1016/j.scijus.2017.03.002.

[59] C.G.G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, J. R. Stat. Soc.: Ser. C: Appl. Stat. 53 (2004) 109–122, http://dx.doi.org/10.1046/j.0035-9254.2003.05271.x.

[60] C.G.G. Aitken, D. Lucy, Corrigendum: evaluation of trace evidence in the form of multivariate data, J. R. Stat. Soc.: Ser. C: Appl. Stat. 53 (2004) 665–666, http://dx.doi.org/10.1111/j.1467-9876.2004.02031.x.