# Calculating likelihood ratios
# (lecture notes)

## Available at:

https://geoff-morrison.net/#calcLRs_lecture_notes

## Author and affiliations:

Geoffrey Stewart Morrison [1,2]

[1] Forensic Data Science Laboratory, Aston University, Birmingham, UK

[2] Forensic Evaluation Ltd, Birmingham, UK

e-mail: geoff-morrison@forensic-evaluation.net

ORCID: 0000-0001-8608-8207

## Acknowledgements:

This research was supported by Research England's Expanding Excellence in England Fund as part of funding for the Aston Institute for Forensic Linguistics 2019–2024.

## Disclaimer:

All opinions expressed in the present paper are those of the author, and, unless explicitly stated otherwise, should not be construed as representing the policies or positions of any organizations with which the author is associated.

**Abstract**

These lecture notes describe calculation of specific-source, common-source, and similarity-score-based likelihood ratios using Gaussian-distribution models. They demonstrate that, since it does not take account of typicality with respect the relevant population, the similarity-score approach does not result in appropriate likelihood-ratio values.

**Keywords**

calculation; forensic inference; likelihood ratio; score

## 1 Introduction

In the context of source attribution, when calculating a likelihood ratio, one must take account not only of the similarity between the items of interest, but also of their typicality with respect to the relevant population. The relevant population is the population from which, in the context of the legal case, the item of questioned origin could conceivably have come had it not come from the known source.[1] The data used to train the likelihood-ratio models (particularly the model in the denominator) must

[1] The relevant population could be explicitly proposed by the defence, but in common-law systems the defence in a criminal trial is under no obligation to propose an alternative to the prosecution's proposition. A forensic practitioner usually has to decide what to adopt as the relevant population and communicate that decision to the court so that the court can later decide whether the forensic practitioner's decision was appropriate, i.e., will it result in a likelihood ratio that answers a question of interest to the court.

be representative of the relevant population for the case, otherwise the calculated likelihood ratio will not answer the question of interest for the case.

Typicality with respect to the relevant population can be incorporated into the calculation of a likelihood ratio using either a specific-source or a common-source approach (Ommen & Saunders, 2021). Another approach, similarity-score-based calculation of likelihood ratios does not properly incorporate typicality with respect to the relevant population. We discuss each of these approaches below.

## 2 Specific-source approach

A specific-source likelihood ratio answers the two-part question:

1. What is the probability of obtaining the measured properties of the item of questioned-source if it came from the specific known source?

   versus

2. What is the probability of obtaining the measured properties of the item of questioned-source if it came not from the specific known source but from some other source selected at random from the relevant population?

A specific-source calculation has the form given in Equation (1), in which $\Lambda$ is the likelihood ratio, $f(x|M)$ is a probability-density function, $x_{\mathrm{q}}$ is the feature (or feature vector) extracted from the questioned-source item, and $M_{\mathrm{k}}$ and $M_{\mathrm{r}}$ are the specific-known-source model and the relevant-population model respectively. The specific-known-source model is trained using feature values extracted from multiple items sampled from the specific known source, and the relevant-population model is trained using feature values extracted from multiple items sampled from the relevant population.

(1)

$$\Lambda = \frac{f\left(x_\mathrm{q} \middle| M_\mathrm{k}\right)}{f\left(x_\mathrm{q} \middle| M_\mathrm{r}\right)}$$

Equation (2) provides an example of a simple model for calculating specific-source likelihood ratios. The model assumes univariate Gaussian distributions in both the numerator and the denominator, and assumes that all sources have the same within-source variance $\sigma_\mathrm{w}^2$. In Equation (2), $f(x|\mu,\sigma^2)$, is a Gaussian probability-density distribution (parametrized using mean and variance), $\mu_\mathrm{k}$ and $\mu_\mathrm{r}$ are the specific-known-source mean and the relevant-population mean respectively, and $\sigma_\mathrm{w}^2$ and $\sigma_\mathrm{b}^2$ are the within-source variance and between-source variance respectively.

(2)

$$\Lambda = \frac{f\left(x_\mathrm{q} \middle| \mu_\mathrm{k}, \sigma_\mathrm{w}^2\right)}{f\left(x_\mathrm{q} \middle| \mu_\mathrm{r}, \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2\right)}$$

Figure 1 shows examples of the calculation of two specific-source likelihood ratios using Equation (2). The wider distribution represents the relevant-population model (denominator model), and the two peakier distributions represent two different specific-source models (numerator models). For these examples: $\mu_\mathrm{r} = 0$, $\sigma_\mathrm{b}^2 = 100$, and $\sigma_\mathrm{w}^2 = 1$; for the filled circles $\mu_\mathrm{k} = 0$ and $x_\mathrm{q} = -1$, and for the unfilled circles $\mu_\mathrm{k} = 20$ and $x_\mathrm{q} = 19$.[2] The resulting likelihood-ratio values, corresponding to the filled and unfilled circles respectively, are $0.24/0.040 = 6$ and $0.24/0.0066 = 36$.

[2] The values in these examples, and in other examples below, were chosen purely for illustrative purposes.

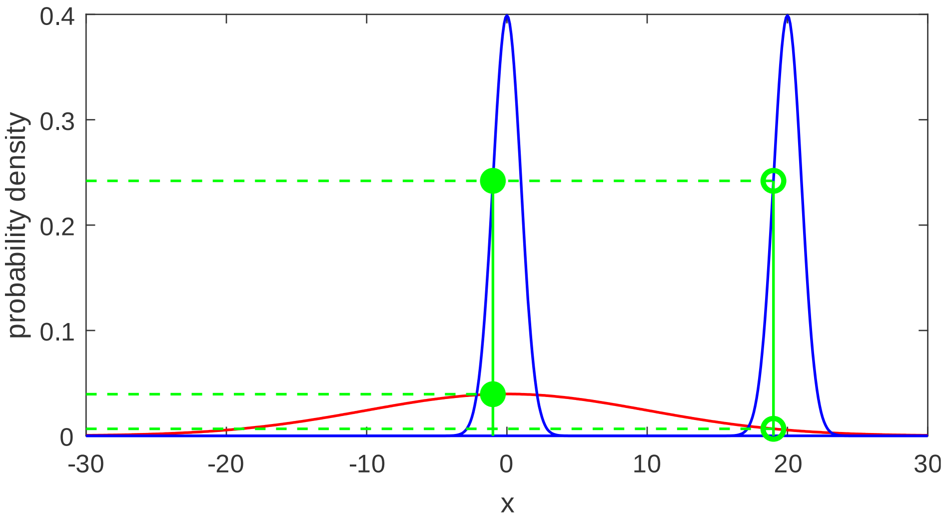


**Figure 1.** Examples of the calculation of specific-source likelihood ratios.

## 3 Common-source approach

A common-source likelihood ratio answers the two-part question:

1. What is the probability of obtaining the measured properties of the items of questioned- and known-source if they both came from the same source (a source selected at random from the relevant population)?

   versus

2. What is the probability of obtaining the measured properties of the items of questioned- and known-source if they each came from a different source (each a source selected at random from the relevant population)?

A common-source calculation has the form given in Equation (3), in which $\Lambda$ is the likelihood ratio, $f\left(x_{\mathrm{q}}, x_{\mathrm{k}} \middle| M\right)$ is a joint probability-density function, $x_{\mathrm{q}}$ and $x_{\mathrm{k}}$ are the features (or feature vectors) extracted from the questioned- and known-source items respectively, and $M_{\mathrm{s}}$ and $M_{\mathrm{d}}$ are the same-source and different-source models respectively.

(3)

$$\Lambda = \frac{f\left(x_{\mathrm{q}}, x_{\mathrm{k}} \middle| M_{\mathrm{s}}\right)}{f\left(x_{\mathrm{q}} \middle| M_{\mathrm{d}}\right) f\left(x_{\mathrm{k}} | M_{\mathrm{d}}\right)}$$

Equation (4) provides an example of a simple model for calculating common-source likelihood ratios.[3] The model assumes univariate Gaussian distributions in both the numerator and the denominator, and assumes that all sources have the same within-source variance $\sigma_{\mathrm{w}}^2$. The numerator of Equation (4) integrates over all possible values for source means given the between-source distribution, with the constraint that $x_{\mathrm{q}}$ and $x_{\mathrm{k}}$ come from the same source. The denominator of Equation (4) integrates over all possible values for source means given the between-source distribution, but does so independently for $x_{\mathrm{q}}$ and for $x_{\mathrm{k}}$. The solutions to the integrals can be expressed as bivariate Gaussian distributions in which for the same-source model (the numerator model) the covariances equal the between-source variance, $\sigma_{\mathrm{b}}^2$, but for the different-source model (the denominator model) the covariances are zero, $0$. This reflects the logic that if a source is selected at random from the population and the mean of the source is high, then two items selected at random from that source will both be expected to have high values. Likewise, if the mean of the source is low, the values of both items will be expected to be low. In general, the values of two items selected from the same source are expected to be correlated. In contrast, if two sources are selected at random from the population and the mean of one is high, there is no expectation that the mean of the other source will also be high. The mean of the other source is more likely to the average, and it is equally likely to be low as to be high. The values of two items each selected from a different source are not expected to be correlated.

[3] This is the univariate version of the model described Aitken & Lucy (2014) as the “multivariate normal (MVN) procedure”, and in the automatic-speaker-recognition and forensic-voice-comparison literature (e.g., Prince & Elder, 2007; Kenny, 2010; Brümmer & de Villiers, 2010; Sizov et al., 2014; Morrison, Enzinger, et al., 2020) as the “two-covariance model” for “probabilistic linear discriminant analysis (PLDA)”.

(4)

$$\Lambda = \frac{\int f\left(x_\mathrm{q} \middle| \mu_\mathrm{i}, \sigma_\mathrm{w}^2\right) f\left(x_\mathrm{k} | \mu_\mathrm{i}, \sigma_\mathrm{w}^2\right) f\left(\mu_\mathrm{i} \middle| \mu_\mathrm{r}, \sigma_\mathrm{b}^2\right) d\mu_\mathrm{i}}{\int f\left(x_\mathrm{q} \middle| \mu_\mathrm{i}, \sigma_\mathrm{w}^2\right) f\left(\mu_\mathrm{i} \middle| \mu_\mathrm{r}, \sigma_\mathrm{b}^2\right) d\mu_\mathrm{i} \int f\left(x_\mathrm{k} \middle| \mu_\mathrm{j}, \sigma_\mathrm{w}^2\right) f\left(\mu_\mathrm{j} \middle| \mu_\mathrm{r}, \sigma_\mathrm{b}^2\right) d\mu_\mathrm{j}}$$

$$= \frac{f\left(\begin{bmatrix} x_\mathrm{q} \\ x_\mathrm{k} \end{bmatrix} \middle| \begin{bmatrix} \mu_\mathrm{r} \\ \mu_\mathrm{r} \end{bmatrix}, \begin{bmatrix} \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 & \sigma_\mathrm{b}^2 \\ \sigma_\mathrm{b}^2 & \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 \end{bmatrix}\right)}{f\left(x_\mathrm{q} \middle| \mu_\mathrm{r}, \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2\right) f\left(x_\mathrm{k} \middle| \mu_\mathrm{r}, \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2\right)}$$

$$= \frac{f\left(\begin{bmatrix} x_\mathrm{q} \\ x_\mathrm{k} \end{bmatrix} \middle| \begin{bmatrix} \mu_\mathrm{r} \\ \mu_\mathrm{r} \end{bmatrix}, \begin{bmatrix} \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 & \sigma_\mathrm{b}^2 \\ \sigma_\mathrm{b}^2 & \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 \end{bmatrix}\right)}{f\left(\begin{bmatrix} x_\mathrm{q} \\ x_\mathrm{k} \end{bmatrix} \middle| \begin{bmatrix} \mu_\mathrm{r} \\ \mu_\mathrm{r} \end{bmatrix}, \begin{bmatrix} \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 & 0 \\ 0 & \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 \end{bmatrix}\right)}$$

Figure 2 shows examples of the calculation of two common-source likelihood ratios using Equation (4). In Figure 2a the univariate distribution represents the relevant-population model with $\mu_\mathrm{r} = 0$, and $\sigma_\mathrm{r}^2 = \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2$ ($\sigma_\mathrm{b}^2 = 100$, and $\sigma_\mathrm{w}^2 = 1$), and the pairs of circles represent pairs of $x_\mathrm{q}$ and $x_\mathrm{k}$ feature values. For the filled circles $x_\mathrm{q} = -1$ and $x_\mathrm{k} = 1$, and for the unfilled circles $x_\mathrm{q} = 19$ and $x_\mathrm{k} = 21$. Figure 2b shows the projection of the $x_\mathrm{q}$ and $x_\mathrm{k}$ feature values into the two-dimensional space of Equation (4), and shows the peakier same-source model (the numerator model) and the flatter different-source model (the denominator model). The resulting likelihood-ratio values, corresponding to the filled and unfilled circles respectively, are $(41 \times 10^{-4})/(16 \times 10^{-4}) = 2.6$ and $(56 \times 10^{-5})/(3.0 \times 10^{-5}) = 19$.

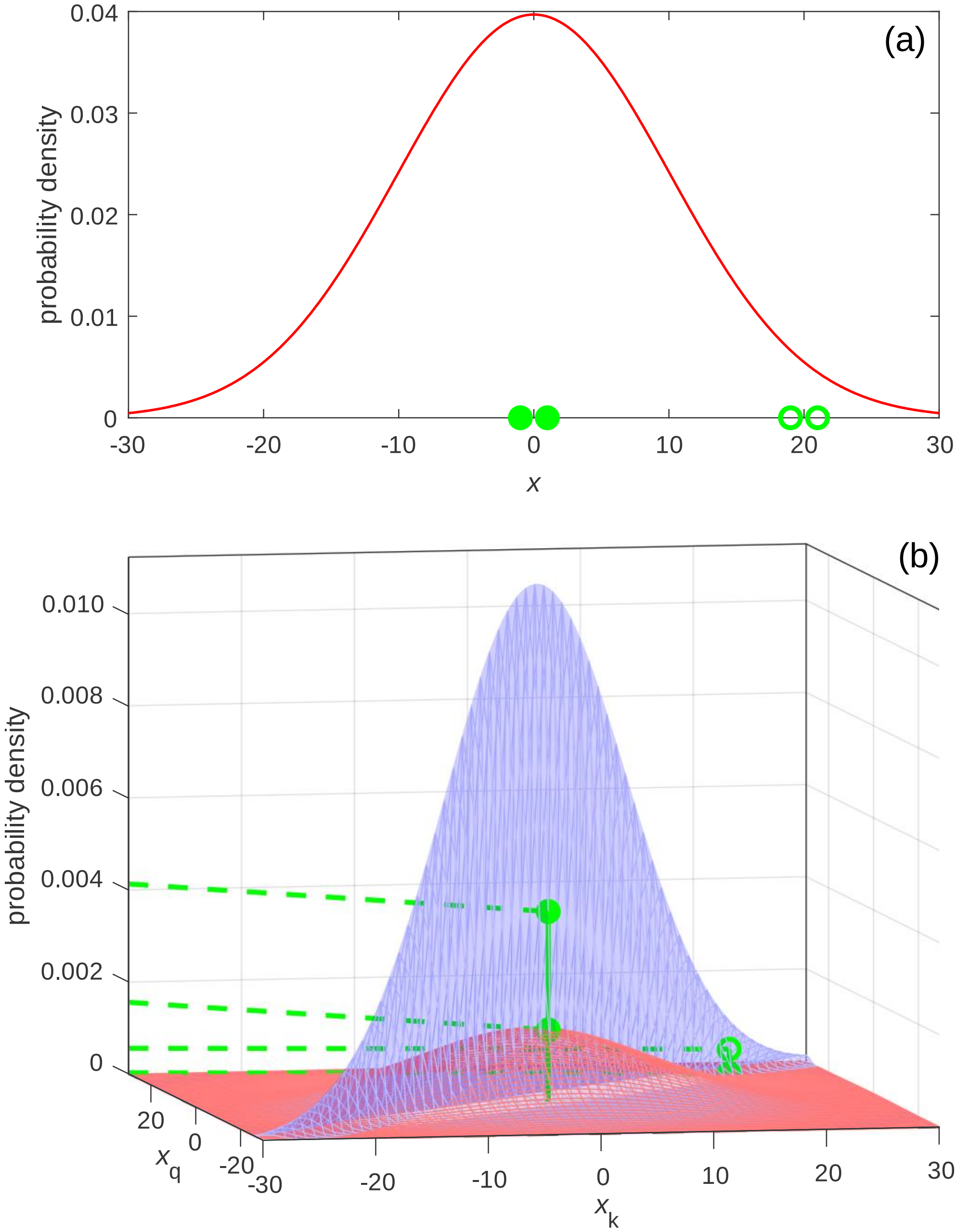


**Figure 2.** Examples of the calculation of common-source likelihood ratios.

Although the example specific-source and common-source models above assume univariate Gaussian distributions, both specific-source and common-source models can fit more complex multivariate distributions. The latter can potentially exploit more useful information resulting in better performance, but they require larger amounts of data in order to estimate larger numbers of parameter values. In forensic casework, the amount of case-relevant data is often limited. Solutions may include use of dimension-reduction techniques, extraction of features using functional data analysis, and use of embeddings from deep neural networks (DNNs) as feature vectors.

## 4 Similarity-score approach

Across multiple branches of forensic science, there have been repeated proposals to calculate likelihood ratios using similarity-score-based approaches (see Morrison & Enzinger, 2018). Such approaches may be motivated by situations in which a specific-source approach cannot be adopted because the data only allow for a comparison between one feature vector and one other feature vector (and this constraint may be due to casework conditions, hence collecting more data may not be a possible solution). They may also be motivated by the difficulties of trying to fit potentially complex distributions in high-dimensional spaces using limited amounts of training data. Instead of fitting models to feature vectors, as for the specific-source and common-source models described above, models are fitted to scores which quantify degree of similarity (or, inversely, degree of difference) between pairs of items. Similarity-scores are scalar values that can be based, for example, on Manhattan distance, Euclidian distance, or a correlation coefficient.[4] Similarity-scores are calculated for pairs of items that are known to come from the same source and for pairs of items known to come from different sources, resulting in a set of same-source scores and a set of different-source scores. The latter scores are used to train a model of the form given in Equation (5), in which $\delta(x_{\mathrm{q}}, x_{\mathrm{k}})$ is the function that calculates the score, and $M_{\delta,\mathrm{s}}$ and $M_{\delta,\mathrm{d}}$ are univariate models trained on same-source scores and different-source scores

[4] Some published literature has failed to distinguish between the calculation of likelihood ratios based on similarity-scores and the calibration of uncalibrated likelihood ratios (see Morrison, 2013, for an introduction to the latter). The confusion may stem from the fact that uncalibrated log likelihood ratios have usually been called “scores” (and the calibration process has been called “score to likelihood-ratio conversion”), but these scores are not similarity-scores, they are scores that take account of both similarity and typicality with respect to the relevant population.

respectively.

(5)

$$\Lambda = \frac{f\left(\delta\left(x_{\mathrm{q}}, x_{\mathrm{k}}\right)\middle|M_{\delta,\mathrm{s}}\right)}{f\left(\delta\left(x_{\mathrm{q}}, x_{\mathrm{k}}\right)\middle|M_{\delta,\mathrm{d}}\right)}$$

A similarity-score-based likelihood ratio answers the two-part question:

1. What is the probability of obtaining the measured degree of similarity between the items of questioned- and known-source if they both came from the same source (a source selected at random from the relevant population)?

   versus

2. What is the probability of obtaining the measured degree of similarity between the items of questioned- and known-source if they each came from a different source (each a source selected at random from the relevant population)?

It may not be immediately obvious, but similarity-score-based approaches do not properly account for typicality with respect to the relevant population. They are therefore not appropriate for calculating likelihood ratios addressing questions of interest for a court. Below, we briefly demonstrate that similarity-scores do not properly account for typicality with respect to the relevant population (more extensive arguments and demonstrations are presented in Morrison & Enzinger, 2018; Neumann & Ausdemore, 2020; and Neumann et al., 2020). In Figure 2a the distance between the two filled circles and the distance between the two non-filled circles is the same, 2 in each case, hence the two pairs of circles will both have the same similarity-score value. The two curves in Figure 3 represent Weibull distributions fitted to same-source scores and to different-source scores, which were obtained by drawing Monte Carlo samples of pairs of same-source feature values and pairs of different-source feature values from the population distribution (consisting of within- and between-source Gaussian distributions with $\mu_{\mathrm{r}} = 0$, $\sigma_{\mathrm{b}}^2 = 100$, and $\sigma_{\mathrm{w}}^2 = 1$), and using unsigned difference as the score function, $\delta(x_{\mathrm{q}}, x_{\mathrm{k}}) = |x_{\mathrm{q}} - x_{\mathrm{k}}|$. Since the two pairs of circles in Figure 2a have the same score value $\delta(x_{\mathrm{q}}, x_{\mathrm{k}}) = 2$, when score-based calculation of likelihood ratios is applied (as graphically illustrated in Figure 3) they both result in the same likelihood-ratio value, which in this example is $0.18/0.058 = 3.1$. Note, however, that in the feature space in Figure 2a, the probability of obtaining the pair of filled-circle feature values if each came from one of two different sources each selected at random from the population is higher than the probability of obtaining the pair of non-filled-circle feature values if each of them came from one of two different sources each selected at random from the population, because the former pair are more typical than the latter pair. Selecting two sources at random, the probability that they will both be in the middle of the distribution is relatively high, whereas the probability that they will both be on the same tail of the distribution is very low. Hence, the likelihood-ratio value associated with the unfilled circles should be higher than the likelihood-ratio value associated with the filled circles, which was the case when the common-source approach was used (19 versus 2.6), but was not the case when the similarity-score-based approach was used (both 3.1), *quod erat demonstandum*: similarity-score-based approaches do not properly account for typicality with respect to the relevant population.

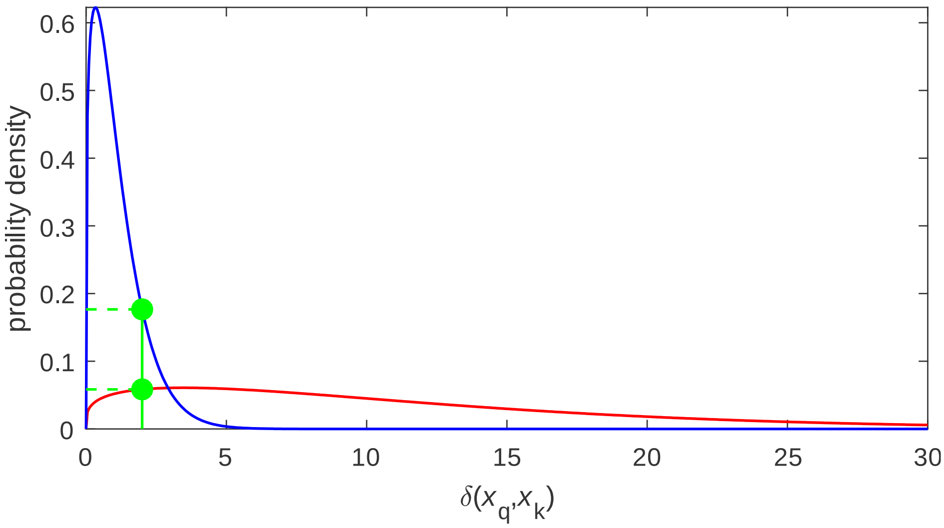


**Figure 3.** Example of the calculation of a similarity-score-based likelihood ratio.

Some published research reports have cited publications such as Morrison & Enzinger (2018), Neumann & Ausdemore (2020), and Neumann et al. (2020), but have then proceeded to use similarity-score-based approaches anyway. Sometimes, they characterize the use of similarity-score-based approaches as entailing "some" loss of information, but we consider the loss of information about typicality with respect to the relevant population to be a loss of essential information, hence we take the position that the use of similarity-score-based approaches are not appropriate for evaluating strength of forensic evidence.

## 5 References

Aitken C.G.G., Lucy D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53, 109–122. http://dox.doi.org/10.1046/j.0035-9254.2003.05271.x [Corrigendum: (2004) 53, 665–666. http://dox.doi.org/10.1111/j.1467-9876.2004.02031.x]

Brümmer N., du Preez J. (2006). Application independent evaluation of speaker detection. *Computer Speech and Language*, 20, 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Kenny P. (2010). Bayesian speaker verification with heavy tailed priors. In *Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop*, paper 014. https://www.isca-speech.org/archive_open/odyssey_2010/od10_014.html

Morrison G.S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45, 173–197. http://dx.doi.org/10.1080/00450618.2012.733025

Morrison G.S., Enzinger E. (2018). Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. *Science & Justice*, 58, 47–58. http://dx.doi.org/10.1016/j.scijus.2017.06.005

Morrison G.S., Enzinger E., Ramos D., González-Rodríguez J., Lozano-Díez A. (2020). Statistical models in forensic voice comparison. In Banks D., Kafadar K., Kaye D.H., Tackett M. (Eds.), *Handbook of Forensic Statistics*, pp. 451–497. Boca Raton, FL: CRC. https://doi.org/10.1201/9780367527709

Neumann C., Ausdemore M. (2020) Defence against the modern arts: The curse of statistics – Part II: 'Score-based likelihood ratios'. *Law, Probability and Risk*, 19, 21–42. http://dx.doi.org/10.1093/lpr/mgaa006

Neumann C., Hendricks J., Ausdemore M. (2020). Statistical support for conclusions in fingerprint examinations. In Banks D., Kafadar K., Kaye D.H., Tackett M. (Eds.), *Handbook of Forensic Statistics*, pp. 277–324. Boca Raton, FL: CRC. https://doi.org/10.1201/9780367527709

Ommen D.M., Saunders C.P. (2021). A problem in forensic science highlighting the differences between the Bayes factor and likelihood ratio. *Statistical Science*, 36, 344–359. https://doi.org/10.1214/20-STS805

Prince S.J.D., Elder J.H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Proceedings of the IEEE 11th International Conference on Computer Vision*, pp. 1–8. https://doi.org/10.1109/ICCV.2007.4409052

Sizov, A., Lee, K.A., Kinnunen, T. (2014). Unifying probabilistic linear discriminant analysis variants in biometric authentication. In Fränti P., Brown G., Loog M., Escolano F., Pelillo M. (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 464–475. Berlin: Springer. https://doi.org/10.1007/978-3-662-44415-3_47