Law, Probability and Risk, 2024, **23**, 1–34
https://doi.org/10.1093/lpr/mgae004
Research Article

OXFORD

Bi-Gaussianized calibration of likelihood ratios

Geoffrey Stewart Morrison [1,2,*]

[1]Forensic Data Science Laboratory, Aston University, Birmingham, United Kingdom
[2]Forensic Evaluation Ltd, Birmingham, United Kingdom
*Corresponding author. E-mail: geoff-morrison@forensic-evaluation.net

Abstract

For a perfectly calibrated forensic evaluation system, the likelihood ratio of the likelihood ratio is the likelihood ratio. Conversion of uncalibrated log-likelihood ratios (scores) to calibrated log-likelihood ratios is often performed using logistic regression. The results, however, may be far from perfectly calibrated. We propose and demonstrate a new calibration method, "bi-Gaussianized calibration," that warps scores toward perfectly calibrated log-likelihood-ratio distributions. Using both synthetic and real data, we demonstrate that bi-Gaussianized calibration leads to better calibration than does logistic regression, that it is robust to score distributions that violate the assumption of two Gaussians with the same variance, and that it is competitive with logistic-regression calibration in terms of performance measured using log-likelihood-ratio cost (C_{llr}). We also demonstrate advantages of bi-Gaussianized calibration over calibration using pool-adjacent violators (PAV). Based on bi-Gaussianized calibration, we also propose a graphical representation that may help explain the meaning of likelihood ratios to triers of fact.

Keywords: calibration; Gaussian distribution; likelihood ratio; logistic regression.

1. Introduction

A set of scales should be well calibrated, otherwise its readout will be misleading. If a set of scales is well calibrated, when an item is placed on the set of scales, the value of its readout will equal the mass of that item. Figure 1 shows, for a perfectly calibrated set of scales, the relationship between the mass placed on the set of scales and the readout of the set of scales—this is the identity function, we will also refer to it as the "perfect-calibration line." The process of calibrating the set of scales involves adjusting its calibration settings so that its readouts over a range of masses are as close as possible to the identity function.

Likewise, a forensic evaluation system that outputs likelihood ratios should be well calibrated, otherwise its output will be misleading (González-Rodríguez et al. 2007; Ramos and González-Rodríguez 2013; Morrison 2013; Brümmer et al. 2014; Meuwly et al. 2017; Vergeer et al. 2020; Morrison et al. 2021). If a forensic evaluation system is well calibrated, the likelihood ratio of the likelihood-ratio value that it outputs will be the same as the likelihood-ratio value that it outputs, or, more pithily, "The likelihood ratio of the likelihood ratio is the likelihood ratio" (Birdsall 1973: 18), or, as an equation, Equation (1), in which Λ is a likelihood ratio, f is a probability-density function, and H_s and H_d are the same-source and different-source hypotheses, respectively.

$$\Lambda = \frac{f(\Lambda|H_s)}{f(\Lambda|H_d)} \tag{1}$$

The following is a perfectly calibrated system: a system for which the distribution of the natural logarithms of the likelihood ratios that it outputs in response to different-source input pairs

Received: 9 November 2023. **Revised:** 13 February 2024. **Accepted:** 10 March 2024

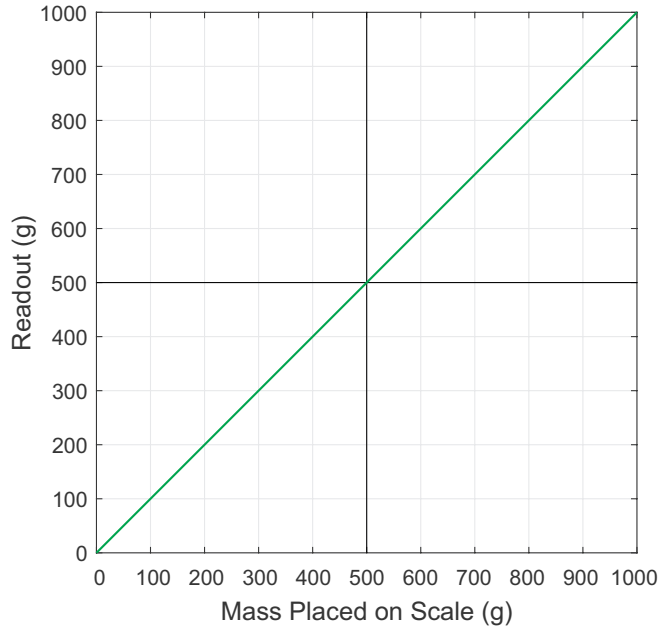

Figure 1. The relationship, for a perfectly calibrated set of scales, between the mass placed on a set of scales and the readout of the set of scales—this is the identity function/the perfect-calibration line.

and the distribution of the natural logarithms of the likelihood ratios that it outputs in response to same-source input pairs are both Gaussian and both have the same variance, $\ln(\Lambda_d) \sim \mathcal{N}(\mu_d, \sigma^2)$ and $\ln(\Lambda_s) \sim \mathcal{N}(\mu_s, \sigma^2)$, and the means of the different-source and same-source distributions are $\mu_d = -\frac{\sigma^2}{2}$ and $\mu_s = +\frac{\sigma^2}{2}$, respectively (Peterson et al. 1954: Sections 4.3 and 4.9; Birdsall 1973: Section 1.3; Good 1985: Section 6; van Leeuwen and Brümmer 2013; Morrison 2021).[1]

Figure 2a shows the different-source and same-source distributions of a perfectly calibrated system with $\sigma = 3$. The fact that the system is perfectly calibrated can be confirmed by selecting any $\ln(\Lambda)$ value on the x-axis, obtaining the probability density of the same-source distribution at that value, obtaining the probability density of the different-source distribution at that value, dividing the former by the latter to obtain a likelihood ratio, and taking the natural logarithm of that likelihood ratio, see Equation (2), in which $f(x|\mu, \sigma)$ is a univariate Gaussian probability-density function. The result is the same as the original $\ln(\Lambda)$ value. Applying Equation (2) across the plotted range of $\ln(\Lambda)$ values results in Fig. 2b, the identity function.

$$\ln(\Lambda) = \ln\left(\frac{f\left(\ln(\Lambda)| + \frac{\sigma^2}{2}, \sigma\right)}{f\left(\ln(\Lambda)| - \frac{\sigma^2}{2}, \sigma\right)}\right) \tag{2}$$

The identity function can be obtained by applying the procedure described above to any perfectly calibrated bi-Gaussian system, irrespective of the value of σ. The panels in the left column of Fig. 3 show examples of different-source and same-source distributions of perfectly calibrated bi-Gaussian systems with different values of σ. All panels represent perfectly calibrated systems, but the performance of systems represented in lower panels is better than the performance of systems represented in higher panels—the overlap between the different-source distribution and the same-source distribution is less. From the top panel to the bottom panel, the corresponding log-

[1] Good (1985) p. 257 attributes the discovery of this relationship to Turing.

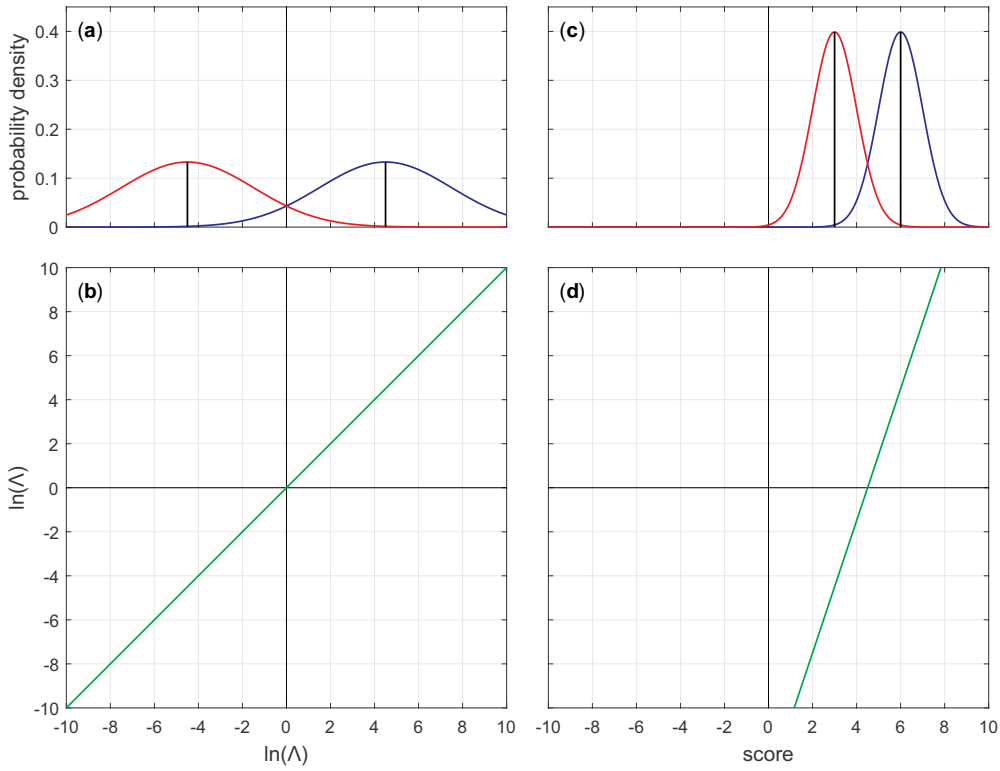

Figure 2. (a) Different-source and same-source Gaussian distributions with, $\mu_d = -4.5$, $\mu_s = 4.5$, $\sigma = 3$. (b) Log-likelihood-ratio-to-log-likelihood-ratio mapping function corresponding to (a). (c) Different-source and same-source Gaussian distributions with, $\mu_d = 3$, $\mu_s = 6$, $\sigma = 1$. (d) Score-to-log-likelihood-ratio mapping function corresponding to (c). Panels (a) and (b) represent a perfectly-calibrated system. Panel (b) shows the identity function.

likelihood-ratio cost (C_{llr}) values are 0.85, 0.52, 0.24, and 0.09. The panels in the right column of Fig. 3 show the Tippett plots corresponding to the panels in the left column.[2]

The process of calibrating a forensic evaluation system that is not already well calibrated involves using a calibration model to calibrate the uncalibrated output of the system. Figure 2c shows an example of the different-source and same-source distributions of the logarithms of likelihood ratios output by a system that is not calibrated, but for which both distributions are Gaussian and they have the same variance. For brevity, we will refer to the logarithms of uncalibrated likelihood ratios as "scores," but note that these scores take into account of both similarity and typicality, they are not similarity-only scores.[3] In order to obtain a mapping function from scores, x, to calibrated $\ln(\Lambda)$, we can, over a range of x values, obtain the probability density of the same-source distribution at each value, obtain the probability density of the different-source distribution at that value, divide the former by the latter to obtain a likelihood ratio, and take the natural logarithm of that likelihood ratio, see Equation (3) and the resulting mapping function in Fig. 2d. Because the scores were not calibrated, the mapping function is not the identity function, but, in this example, it is a linear function (a linear discriminant function, LDF), as described in Equation (4), in which μ_d, μ_s, and σ are the statistics (or parameters) of the score distributions, that is, the statistics of the original values rather than those of the transformed values. For derivation of the equations for the intercept and slope values, a and b, see van Leeuwen and Brümmer (2013) or Morrison (2021).

[2] For explanations of C_{llr} and of Tippett plots, see Morrison et al. (2021) Appendix C.
[3] Morrison & Enzinger (2018), Neumann & Ausdemore (2020), and Neumann et al. (2020) have argued that calculating likelihood ratios based on scores that only take account of similarity does not result in meaningful likelihood-ratio values because they do not take account of typicality with respect to the relevant population for the case. Vergeer (2023), however, argues that use of similarity-score-based systems to calculate likelihood ratios is acceptable if system performance is better than using prior odds.

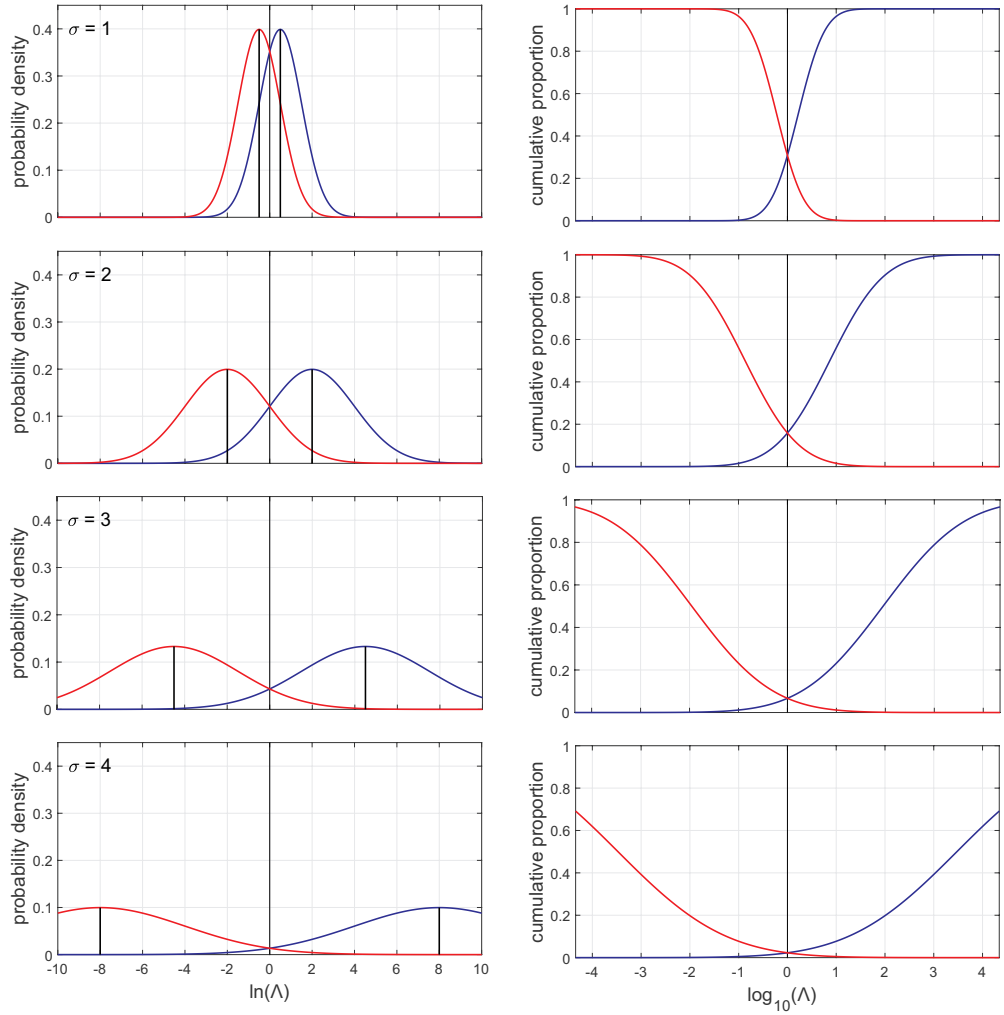

Figure 3. Left column: different-source and same-source distributions of perfectly calibrated bi-Gaussian systems with different values of σ. Right column: Tippett plots corresponding to the distributions in the left column. From the top to the bottom panel, the corresponding C_{llr} values are 0.85, 0.52, 0.24, and 0.09.

$$\ln(\Lambda) = \ln\left(\frac{f(x|\mu_{\text{s}},\sigma)}{f(x|\mu_{\text{d}},\sigma)}\right) \tag{3}$$

$$\ln(\Lambda) = a + bx \tag{4}$$

$$b = \frac{\mu_{\text{s}} - \mu_{\text{d}}}{\sigma^2}$$

$$a = \frac{-\mu_{\text{s}}^2 + \mu_{\text{d}}^2}{2\sigma^2} = -b\frac{\mu_{\text{s}} + \mu_{\text{d}}}{2}$$

The mapping function in the example in Fig. 2d has slope $b = \frac{6-3}{1^2} = 3$ and intercept $a = -3\times\frac{6+3}{2} = -3\times4.5 = -13.5$. If this mapping function is applied to the distributions in Fig. 2c, the result is Fig. 2a. If the same procedure is then applied to the already-calibrated distributions in Fig. 2a, the mapping function in Fig. 2b has slope $b = \frac{4.5+4.5}{3^2} = 1$ and intercept $a = -1\times\frac{4.5-4.5}{2} = 0$, that is, it is the identity function.

In the example above, the different-source score distribution and the same-source score distribution were both Gaussian and they had the same variance. In these circumstances, the calibration model can be based on an LDF as described in Equation (4). In practice, the assumption that the score distributions consist of two Gaussians with the same variance is often violated. Therefore, rather than using an LDF, the intercept and slope, a and b, are usually obtained using logistic regression (LogReg), which is more robust to violations of this assumption (Brümmer and du Preez 2006; González-Rodríguez et al. 2007; Morrison 2013; Morrison et al. 2020, 2023). Notice, however, that since the calibration function is still a linear mapping in the $\ln(\Lambda)$ space, if the score distributions violate the assumption of two Gaussians with the same variance, then the "calibrated" $\ln(\Lambda)$ distributions will not consist of two Gaussians with the same variance, and hence may be far from the same-source and different-source distributions of a perfectly calibrated bi-Gaussian system. The pool-adjacent violators (PAVs) algorithm (Ayer et al. 1955; Zadrozny and Elkan 2002; Brümmer and du Preez 2006), aka isotonic regression, has been used to provide a non-linear but monotonic mapping, but this non-parametric approach overfits on the training data and therefore does not generalize well to new data. Also, it does not extrapolate below the lowest same-source value and above the highest different-source value.[4] Using kernel-density estimates (KDEs) of the same-source scores and of the different-source scores, dividing the former by the latter, and taking the logarithm results in a non-linear mapping, but one which is not monotonic. Conversion of uncalibrated likelihood ratios to calibrated likelihood ratios, and indeed calibration in general, should be monotonic.

This article proposes and demonstrates the use of a new calibration method, "bi-Gaussianized calibration," that uses a non-linear monotonic function to map scores toward a perfectly calibrated bi-Gaussian system.

The remainder of this article is organized as follows:

- Section 2 describes the bi-Gaussianized calibration method.
- Section 3 demonstrates the application of the method on two sets of simulated data and on two sets of real data.
- Section 4 explores the effect of sampling variability on the performance of the method.
- Section 5 describes a graphical representation that may facilitate understanding by the trier of fact of likelihood-ratio values output by the method.
- Section 6 provides a conclusion.

The data and Matlab® code used for this article are available from https://forensic-data-science.net/calibration-and-validation/#biGauss. This includes a function that implements bi-Gaussianized calibration and a function that draws the graphical representation described in Section 5. Python versions of the latter functions will also be made available.

2. Bi-Gaussianized calibration method

2.1 Variants

Below, we describe four variants of bi-Gaussianized calibration. Each variant uses a different method to calculate the σ^2 for a target perfectly calibrated bi-Gaussian system. Three variants include an initial calibration step using LogReg, KDE, or PAV, then calculate the target σ^2 based on C_{llr}, and the other variant calculates the target σ^2 based on equal-error rate (which we will abbreviate as $E_{=}$ or EER depending on context). Scores are then mapped toward the $\ln(\Lambda)$ of the perfectly calibrated bi-Gaussian system with the target σ^2.

In the remainder of this section:

- We describe the steps required to implement each variant of the bi-Gaussianized-calibration method.

[4] Laplace's rule of succession can be used to prevent the values outside this range from being reported as $-\infty$ and $+\infty$, see Brümmer & du Preez (2006) §13.2.1.1.

- We derive, for a perfectly calibrated bi-Gaussian system, the relationship between C_{llr} and σ^2, and the relationship between $E_=$ and σ^2.
- We compare the performance of the EER, LogReg, KDE, and PAV methods for determining the target σ^2.
- We describe the algorithm for mapping from the cumulative score distribution to the cumulative distribution of the perfectly calibrated bi-Gaussian system with the target σ^2.

2.2 Steps for the C_{llr}-based variants

The C_{llr}-based variants of the bi-Gaussianized-calibration method consist of the following steps:

1) Calculate same-source scores and different-source scores (uncalibrated log-likelihood ratios) for a set of training data and a set of test data.
2) Calibrate the training-data output of Step 1 using one of the following methods: LogReg[5]; KDE[6]; PAV.[7]
3) Calculate C_{llr} for the output of Step 2.
4) Determine the σ^2 of the perfectly calibrated system with the same C_{llr} as calculated at Step 3.
5) Ignoring the same-source and different-source labels, determine the mapping function from the empirical cumulative distribution of the training-data output of Step 1 to the cumulative distribution of the perfectly calibrated bi-Gaussian system (a two-Gaussian mixture) with the σ^2 determined at Step 4.
6) Apply the mapping function determined at Step 5 to the test-data output of Step 1, and to the score calculated for the comparison of the actual questioned- and known-source items in the case.

2.3 Relationship between C_{llr} and σ^2

At Step 3 of the C_{llr}-based variants of the bi-Gaussianized-calibration method, C_{llr} is calculated using Equation (5), in which N_{s} and N_{d} are the number of same-source and different-source input pairs, respectively.[8]

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{s}}}\sum_{i=1}^{N_{\text{s}}}\log_2\left(1+\frac{1}{\Lambda_{\text{s}_i}}\right)+\frac{1}{N_{\text{d}}}\sum_{j=1}^{N_{\text{d}}}\log_2\left(1+\Lambda_{\text{d}_j}\right)\right) \tag{5}$$

Although, in general, there is a many-to-one mapping in which different sets of same-source and different-source likelihood ratios can map to the same C_{llr} value, there is a one-to-one mapping between the σ^2 of a perfectly calibrated system and the C_{llr} of that system. Step 4 requires that the latter mapping be known. Given C_{llr}, if one can determine σ^2, then one knows everything about the same-source and different-source distributions of that perfectly calibrated system.

For a perfectly calibrated system, the distributions of $\ln(\Lambda_{\text{d}})$ and $\ln(\Lambda_{\text{s}})$ are reflections of one another about $\ln(\Lambda) = 0$, so we need only consider either the left sum or the right sum within the outer parenthesis of Equation (5). We take the left sum, convert it to a definite integral, and arrive at Equation (6).[9]

[5] We use a regularized version of logistic regression with a regularization weight of $\kappa = 0.01$ relative to the number of sources, see Morrison & Poh (2018) for details. This amount of regularization resolves potential numerical problems, but does not induce substantial shrinkage.

[6] We used Gaussian kernels with the bandwidth determined using the Gaussian-approximation method, aka Silverman's rule of thumb (Silverman, 1986, p. 45).

[7] For PAV, we used Laplace's rule of succession so that score values below the smallest same-source score and score values above the largest different-source score result in finite $\ln(\Lambda)$ values (see note 4 above).

[8] The form of Equation (5) is that given in González-Rodríguez et al. (2007) and thereafter widely repeated in the literature. It can be derived from Brümmer & du Preez (2006) Equation 43.

[9] In general, a sum $\frac{1}{N}\sum_{i=1}^{N} g(x_i)$ converts to an integral $\int_{-\infty}^{\infty} f(x)g(x)dx$, in which $f(x)$ is the probability density function for x, and $g(x)$ is the function of x for which one wishes to integrate out x.

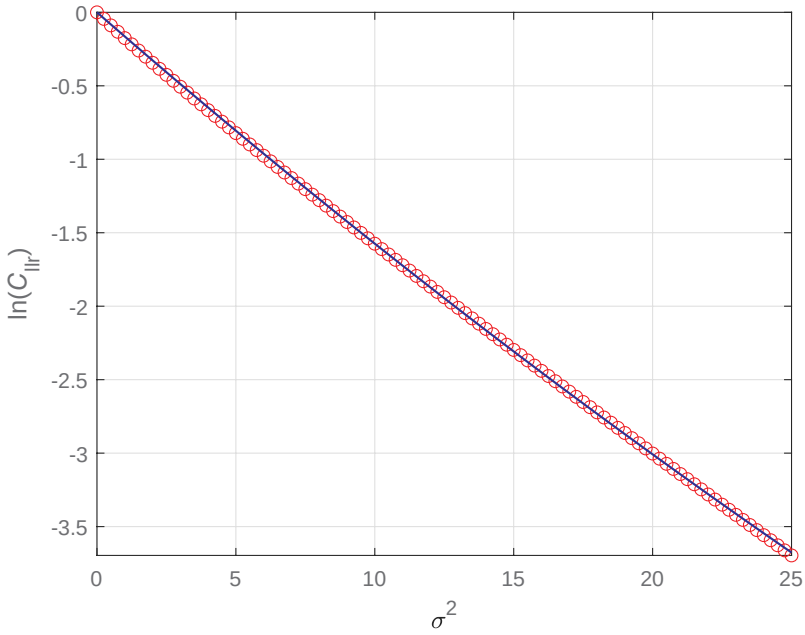

Figure 4. Circles: $\ln(C_{llr})$ values given σ^2 values. $\ln(C_{llr})$ values were calculated using numerical integration applied to Equation (6). Curve: Fitted regression of $\ln(C_{llr})$ on σ^2.

$$C_{llr} = \frac{1}{\ln(2)} \int_{-\infty}^{\infty} f\left(x \middle| \frac{\sigma^2}{2}, \sigma\right) \ln(1+e^{-x}) dx$$
$$x = \ln(\Lambda_s) \tag{6}$$

We can use numerical integration to integrate out x and arrive at the value for the integral.[10] Figure 4 plots the resulting values of $\ln(C_{llr})$ over a range of σ^2 values. Figure 4 also shows a fitted regression, the equation for which is given in Equation (7).[11] The fitted coefficient values for the regression are $b = 17.7$ and $c = 9.33 \times 10^{-3}$. Solving for σ^2, we arrive at Equation (8), which is graphically represented in Fig. 5. Given a C_{llr} value, Equation (8) can be used to calculate the σ^2 of the perfectly calibrated system with that C_{llr} value.

$$\ln(C_{llr}) = b\left(e^{-c\sigma^2} - 1\right) \tag{7}$$

$$\sigma^2 = -\frac{1}{c} \ln\left(\frac{1}{b} \ln(C_{llr}) + 1\right) \tag{8}$$

This subsection has derived the relationship between C_{llr} and σ^2 for a perfectly calibrated bi-Gaussian system. Instead of using C_{llr}, any other strictly proper scoring rule (SPSR) could be used. One would have to derive the relationship between that SPSR and σ^2 for a perfectly calibrated bi-Gaussian system. We use C_{llr}, rather than any other SPSR, because it is commonly used in the forensic-science literature.

2.4 Steps for the $E_=$-based variant

The $E_=$-based variant of the bi-Gaussianized-calibration method consists of the steps listed below. Steps 1, 5, and 6 are identical to the C_{llr}-based variants, and Steps 3 and 4 are parallel. The $E_=$-based variant has no parallel of the C_{llr}-based variants' Step 2.

[10] We used the "integral" function in Matlab®.
[11] The general form of the equation would be $\ln(C_{llr}) = a + be^{-c\sigma^2}$, but at $\sigma^2 = 0$, $\ln(C_{llr}) = 0$, therefore $a = -b$.

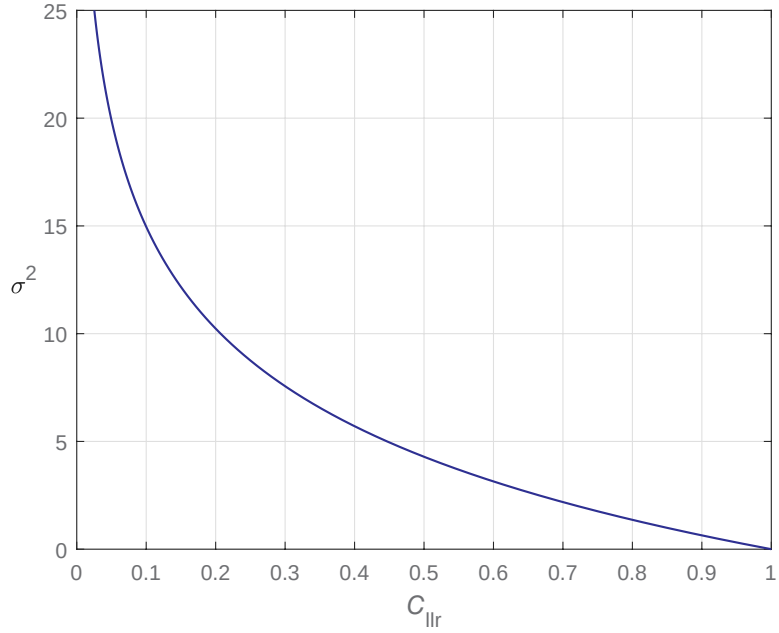

Figure 5. Relationship between C_{llr} and σ^2 for a perfectly calibrated bi-Gaussian system.

1) Calculate same-source scores and different-source scores (uncalibrated log-likelihood ratios) for a set of training data and a set of test data.
2) Skip.
3) Calculate the $E_=$ for the training-data output of Step 1.
4) Determine the σ^2 of the perfectly-calibrated system with the same $E_=$ as calculated at Step 3.
5) Ignoring the same-source and different-source labels, determine the mapping function from the empirical cumulative distribution of the training-data output of Step 1 to the cumulative distribution of the perfectly calibrated bi-Gaussian system (a two-Gaussian mixture) with the σ^2 determined at Step 4.
6) Apply the mapping function determined at Step 5 to the test-data output of Step 1, and to the score calculated for the comparison of the actual questioned- and known-source items in the case.

2.5 Relationship between $E_=$ and σ^2

$E_=$ is the value at which the false-alarm rate (FAR) and miss rate (MR) are equal. To calculate $E_=$ we used an algorithm that sorted the data from Step 1 irrespective of different-source or same-source label, calculated FAR and MR at each datapoint (FAR monotonically decreases and MR monotonically increases), found the first datapoint for which FAR was less than MR, then took the mean of FAR and MR at that datapoint or the mean of FAR and MR at the immediately preceding datapoint, whichever was lower.[12] If FAR was zero, the mean was calculated using the MR obtained as described above and half the lowest non-zero FAR. If MR was zero, the mean was calculated using the FAR obtained as described above and half the lowest non-zero MR. If FAR was never less than MR, the mean was calculated using half the lowest non-zero FAR and half the lowest non-zero MR.

Although, in general, there is a many-to-one mapping in which different sets of same-source and different-source likelihood ratios can map to the same $E_=$ value, there is a one-to-one mapping between the σ^2 of a perfectly calibrated system and the $E_=$ of that system. Step 4 requires

[12] If FAR became less than MR because of a step up in MR, FAR had the same value at that datapoint and at the immediately preceding datapoint. If FAR became less than MR because of a step down in FAR, MR had the same value at that datapoint and at the immediately preceding datapoint.

that the latter mapping be known. Given $E_{=}$, if one can determine σ^2, then one knows everything about the same-source and different-source distributions of that perfectly calibrated system.

An advantage of the $E_{=}$-based variant of the bi-Gaussianized-calibration method over the C_{llr}-based variant is that, unlike for the C_{llr} to σ^2 conversion function (Section 2.3), there is an analytical solution for the $E_{=}$ to σ^2 conversion function. For a perfectly calibrated system (for which $\mu_s = \frac{\sigma^2}{2}$ and the miss rate and false-alarm rate are equal at $\ln(\Lambda) = 0$), the relationship between $E_{=}$ and σ is as given in Equation (9).

$$E_{=} = f(0|\mu_s, \sigma) = f\left(-\frac{\mu_s}{\sigma}|0,1\right) = f\left(-\frac{\sigma^2}{2\sigma}|0,1\right) = f\left(-\frac{\sigma}{2}|0,1\right) \tag{9}$$

Solving Equation (9) with respect to σ results in Equation (10), in which F^{-1} is the inverse cumulative probability function for a Gaussian distribution.

$$\sigma = -2F^{-1}(E_{=}|0,1) \tag{10}$$

Figure 6 shows the relationship between $E_{=}$ and σ^2 for a perfectly calibrated system.

2.6 Comparison of methods for determining the target σ value

In this subsection, we explore the performance of the alternative methods for determining the σ for the perfectly calibrated bi-Gaussian system toward which scores will be mapped (Steps 3 and 4 of the bi-Gaussianized-calibration method).

We generated synthetic data using Monte Carlo simulation. Sample sets consisting of 100 same-source scores and 4950 different-source scores were generated from perfectly calibrated bi-Gaussian systems with $\sigma \in \{1, 2, 3, 4\}$ (the generating distributions are the same as those plotted in Fig. 3). For each value of σ, we generated 1,000 sample sets. For each sample set, we calculated the target σ value obtained from each method: EER, LogReg, KDE, and PAV. Violin plots of results are provided in Fig. 7 through Fig. 10, note that the scale of the y axis differs across figures. In each figure, the solid horizontal line indicates the "true" σ, that is, the parameter value used to generate the Monte Carlo samples. Table 1 gives the root-mean square (RMS) errors between the target σ value and the "true" σ value.

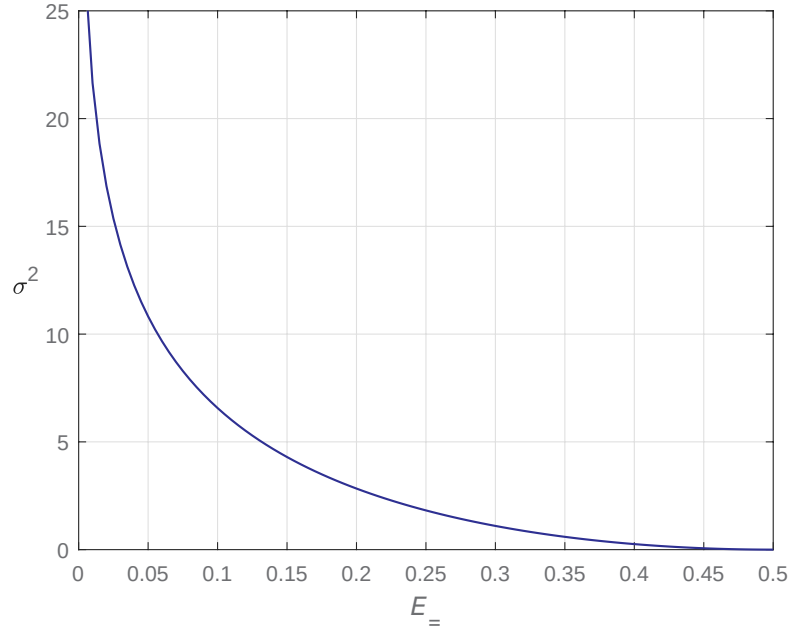

Figure 6. Relationship between $E_{=}$ and σ^2 for a perfectly calibrated system.

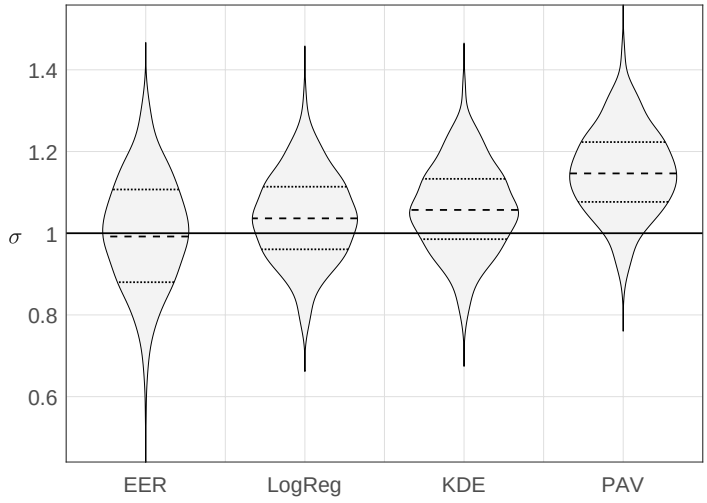

Figure 7. Violin plots of distributions of target σ values estimated by each method. The solid horizontal line indicates the parameter value used to generate the Monte Carlo samples, $\sigma = 1$.

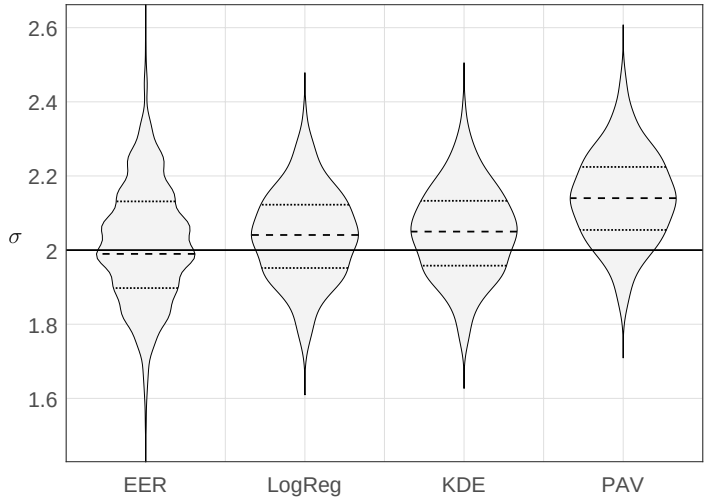

Figure 8. Violin plots of distributions of target σ values estimated by each method. The solid horizontal line indicates the parameter value used to generate the Monte Carlo samples, $\sigma = 2$.

Table 1. RMS errors of the target σ values obtained from each method relative to the "true" σ, the parameter value used to generate the Monte Carlo samples. For each "true" σ value, the RMS value for the best performing method is bolded.

	Methods			
"true" σ	EER	LogReg	KDE	PAV
1	0.133	**0.113**	0.122	0.184
2	0.153	**0.127**	0.131	0.187
3	0.203	0.162	**0.159**	0.217
4	0.309	0.253	**0.232**	0.303

The PAV method had a substantial bias toward larger target σ values than the "true" σ value. This is likely due to PAV overfitting the training data, resulting in smaller C_{llr} values and hence larger target σ values. Except for "true" $\sigma = 4$, for which it had the second highest RMS error, the PAV method always had the highest RMS error. Because of the poor performance of PAV as a method for determining the target σ, in the remainder of the article, we do not make use of the PAV variant of the bi-Gaussianized-calibration method.

For the EER method, as the separation between the different-source distribution and the same-source distribution increases, sparsity of data in the overlapping tails of the distributions accounts for the multimodality in the distributions of target σ, that is, the multiple bulges see in the violin plots for the EER method for $\sigma \in \{2, 3, 4\}$. Each bulge corresponds to a miss rate which, because there are only 100 same-source samples, is discretized into steps of 0.01, for example, for the EER method in Fig. 10, the highest bulge corresponds to a miss rate of 0.01, the middle (widest) bulge to a miss rate of 0.02, and the lowest bulge to a miss rate of 0.03, and in Fig. 9, the two widest bulges correspond to miss rates of 0.06 and 0.07. Except for "true" $\sigma = 4$,

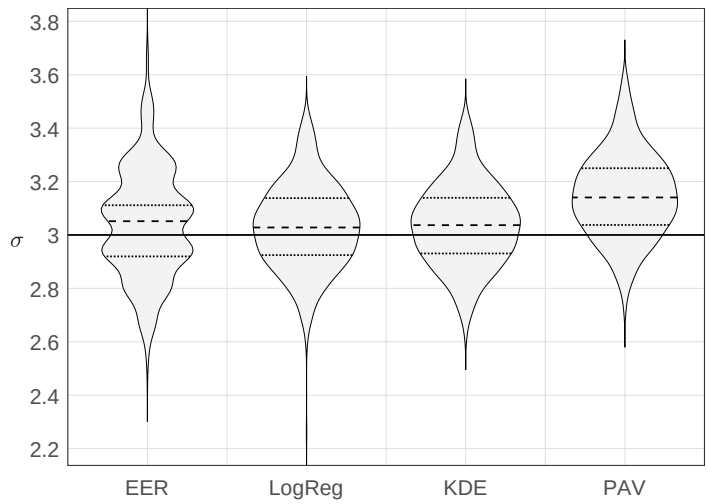

Figure 9. Violin plots of distributions of target σ values estimated by each method. The solid horizontal line indicates the parameter value used to generate the Monte Carlo samples, $\sigma = 3$.

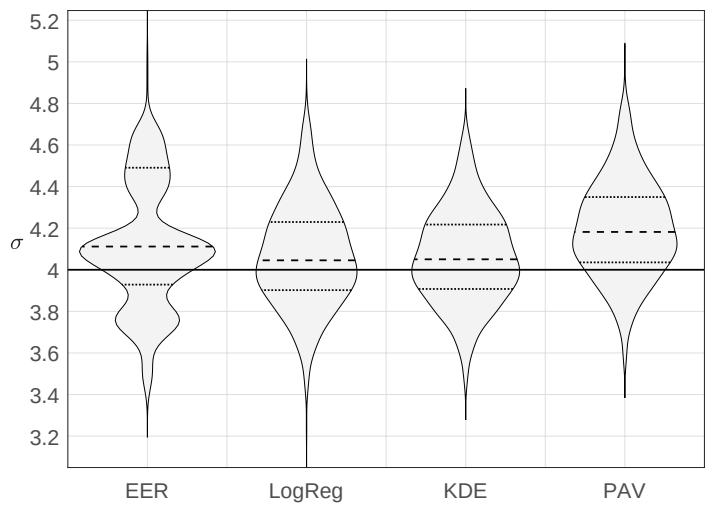

Figure 10. Violin plots of distributions of target σ values estimated by each method. The solid horizontal line indicates the parameter value used to generate the Monte Carlo samples, $\sigma = 4$.

for which it had the highest RMS error, the EER method always had the second highest RMS error.

The RMS errors were consistently lowest for the LogReg and KDE methods. For "true" $\sigma \in \{1, 2\}$ the LogReg method had the lowest RMS error, and for "true" $\sigma \in \{3, 4\}$ the KDE method had the lowest RMS error. Both these methods had a slight bias toward larger target σ values than the "true" value (a somewhat greater bias for the KDE method when "true" $\sigma = 1$). This bias may be due to error in the procedure used to derive the C_{llr} to σ^2 conversion function (Section 2.3), stochastic error, fitting error, or intrinsic underfitting due to an insufficiently complex regression equation.

Because the LogReg and KDE methods for determining the target σ result, on average, in target σ values that are closer to the "true" σ value than the EER and PAV methods, the LogReg and KDE variants of the bi-Gaussianized-calibration method appear to be more promising than the EER and PAV variants of the bi-Gaussianized-calibration method.

For the LogReg method when "true" $\sigma = 3$, there was an outlier with a low σ value, which would result in a conservative calibration, that is, log-likelihood ratios closer to zero than for the "true" perfectly calibrate system. The KDE method does not exhibit an outlier, so this could be a reason to favor the KDE method over the LogReg method.

In order to have a "true" σ against which to compare target σ results, the Monte Carlo simulations used a different-source score distribution and a same-source score distribution which were Gaussian and had the same variance. In these simulations, the LogReg method's assumption of linearity in the score (or $\ln(\Lambda)$) domain was satisfied. When the different-source score distribution and the same-source score distribution are not Gaussian or do not have the same variance, and these assumptions are violated, it could be that the KDE variant of the bi-Gaussianized-calibration method will outperform the LogReg variant. Note that, although KDE as a calibration method would result in a non-linear mapping, when KDE is used to determine the target σ as part of the bi-Gaussianized-calibration method, the bi-Gaussianized-calibration method will result in a monotonic mapping.

2.7 Cumulative-distribution mapping

Step 5 of the bi-Gaussianized-calibration method requires determining the empirical cumulative distribution of the training scores output at Step 1. This is done giving equal weight to the set of different-source scores and the set of same-source scores (the number of different-source scores, N_d, and the number of same-source scores, N_s, usually differ). Each different-source score is given a weight of $\frac{1}{2}\left(\frac{1}{N_d+1}\right)$, and each same-source score is given a weight of $\frac{1}{2}\left(\frac{1}{N_s+1}\right)$. All the scores, $\{x_1 \ldots x_{N_d+N_s}\}$, irrespective of their same-source or different-source labels, are sorted from smallest to largest, and the cumulative sums of the sorted scores' weights are calculated. The resulting values of the cumulative sums of the weights monotonically increase from near 0 for the lowest-value score to near 1 for the highest-value score. In the denominators of the weights, the addition of 1 to N_d and the addition of 1 to N_s prevents the final cumulative-sum value from reaching 1. We define $G(x)$ as the empirical-cumulative-distribution function. $G(x)$ returns the value of the cumulative sum of the weights up to and including score value x. If the score value, x, is not exactly the same value as one of the training-score values, the value of $G(x)$ is linearly interpolated using the closest training-score value below x, the closest training-score value above x, and their corresponding $G(x)$ values. The method does not extrapolate beyond values encountered in the training data: If the value of x is below the smallest score value in the training set or above the largest score value in the training set, the value returned by $G(x)$ corresponds to, respectively, the $G(x)$ value for the smallest score value in the training set or the $G(x)$ value for the largest score value in the training set.

A score, x, is mapped to a calibrated $\ln(\Lambda)$ value using Equation (11), in which F^{-1} is the inverse cumulative distribution for a two-Gaussian mixture with the specified mean, standard-deviation, and weight values.[13] Each Gaussian in the mixture is given the same weight, $w = 0.5$. The value of σ^2 is the target value determined at Step 4 of the bi-Gaussianized-calibration method. The fact that the maximum $G(x)$ value does not reach 1 prevents F^{-1} from returning an infinite value.

[13] F^{-1} is defined differently in Equation (11) than in Equation (10).

$$\ln(\Lambda) = F^{-1}\left(G(x)|[\mu_d, \sigma, w]; [\mu_s, \sigma, w]\right) = F^{-1}\left(G(x)|\left[-\frac{\sigma^2}{2}, \sigma, 0.5\right]; \left[\frac{\sigma^2}{2}, \sigma, 0.5\right]\right) \quad (11)$$

3. Demonstrations of the method

3.1 Introduction

We demonstrate application of the method using:

- Synthetic score data for which the generating distributions are those of a perfectly-calibrated bi-Gaussian system (Section 3.2).
- Synthetic score data for which the generating different-source distribution is Gaussian and the generating same-source distribution is skewed and has a smaller variance than the different-source distribution (Section 3.3).
- Real score data from a forensic-voice-comparison system (Section 3.4). These data moderately deviate from the assumption of equal-variance Gaussians.
- Real score data from a comparison of glass fragments (Section 3.5). These data exhibit extreme deviation from the assumption of equal-variance Gaussians.

For plots below of cumulative distributions, mapping functions, probability-density plots of $\ln(\Lambda)$ values, and Tippett plots, we have always used the target σ obtained from the LogReg variant of bi-Gaussianized calibration. The target σ for the EER and KDE variants were always similar, so would have resulted in similar (often visually indistinguishable) plots.

3.2 Synthetic data: Equal-variance Gaussians

We generated synthetic data using Monte Carlo simulation. The generating different-source distribution was a Gaussian with parameters $\mu_d = -4.5$ and $\sigma = 3$, and the generating same-source distribution was a Gaussian with parameters $\mu_s = 4.5$ and $\sigma = 3$, that is, the perfectly calibrated bi-Gaussian system shown in Fig. 2a and in the third row of Fig. 3. We generated a training-data sample set consisting of 100 same-source scores and 4,950 different-source scores, and a separate test-data sample set of the same size.

In addition to applying bi-Gaussianized calibration (EER, LogReg, and KDE variants), we also applied LDF calibration using the Monte Carlo parameter values (which gives the "true" likelihood-ratio values), LogReg calibration, and PAV calibration.

The parameter value for σ was 3, and the target σ calculated for bi-Gaussianized calibration were 3.10, 2.96, and 2.95 for the EER, LogReg, and KDE variants, respectively. Figure 11 shows the empirical cumulative distribution for the score data, and the target cumulative distribution for a perfectly calibrated bi-Gaussian system with $\sigma = 2.96$.

Figure 12 shows the "true" mapping function from scores to calibrated $\ln(\Lambda)$ given the Monte Carlo population distributions. Since the Monte Carlo population distributions were a perfectly calibrated bi-Gaussian system, the "true" mapping function is the identity function. Figure 12 also shows the mapping functions from scores to calibrated $\ln(\Lambda)$ for PAV, LogReg, and bi-Gaussianized calibration (with target $\sigma = 2.96$). LogReg calibration results in a linear mapping which is close to the perfect-calibration line. PAV calibration results in a stepped mapping function in which some steps are large and have relatively large deviations from the perfect-calibration line. Below the smallest same-source score and above the largest different-source score in the training data, the PAV mapping ceases to change, resulting in even larger deviations from the perfect-calibration line. In contrast to PAV calibration, bi-Gaussianized calibration results in a smoother mapping that generally stays closer to the perfect-calibration line, including below the smallest same-source score and above the largest different-source score in the training data.

Table 2 gives the C_{llr} values for LDF calibration using the Monte Carlo population distributions ("true" values), bi-Gaussianized calibration, LogReg calibration, and PAV calibration. All methods, other then PAV calibration, resulted in similar C_{llr} values (by design, the C_{llr} value for the LogReg variant of bi-Gaussianized calibration should be approximately the same as that for

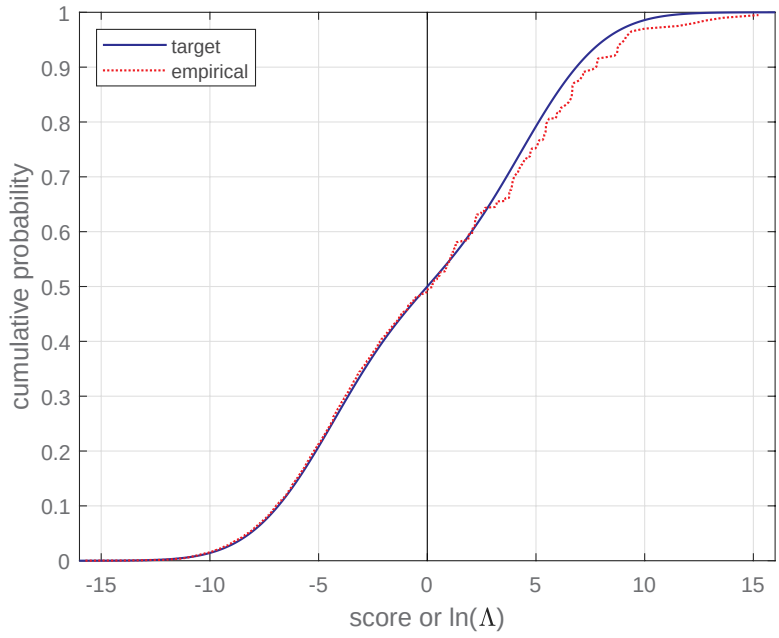

Figure 11. Empirical cumulative distribution of equal-variance Gaussian score data, and target cumulative distribution of a perfectly calibrated bi-Gaussian system with $\sigma = 2.96$.

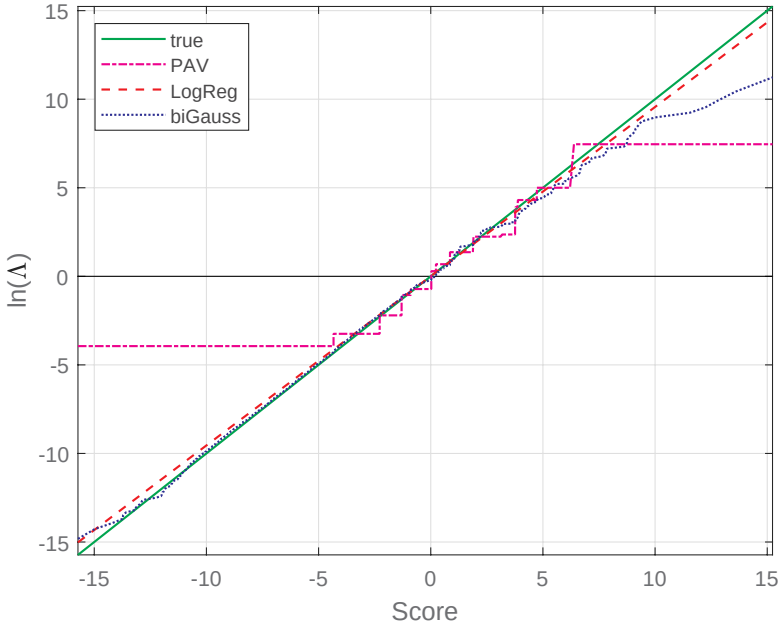

Figure 12. Mapping functions from scores to calibrated $\ln(\Lambda)$ for the synthetic equal-variance Gaussian score data. The bi-Gaussianized-calibration mapping is for target $\sigma = 2.96$.

LogReg calibration). PAV calibration exhibited overfitting on the training data: it had the lowest C_{llr} value on the training data but the highest on the test data. Compared to PAV calibration, bi-Gaussianized calibration exhibited less overfitting.

Table 2. C_{llr} values for different calibration methods applied to the synthetic equal-variance Gaussian data.

	Method					
	"true"	Bi-Gauss			LogReg	PAV
Data		EER	LogReg	KDE		
train	0.251	0.249	0.248	0.248	0.251	0.231
test	0.212	0.213	0.216	0.216	0.213	0.220

3.3 Synthetic data: Gaussian distribution and skewed distribution with smaller variance

We generated synthetic data using Monte Carlo simulation. The generating different-source distribution was a Gaussian with parameters $\mu_d = 0$ and $\sigma_d = 3$, and the generating same-source distribution was a Gumbel distribution with parameters $\mu_s = -9$ and $\sigma_s = 1$ that was mirrored about 0, see Fig. 13. These are similar to distributions sometimes observed for real score data. We generated a training-data sample consisting of 100 same-source scores and 4,950 different-source scores, and a separate test-data sample of the same size.

In addition to applying bi-Gaussianized calibration (EER, LogReg, and KDE variants), we also used the Monte Carlo generating distributions to calculate "true" likelihood-ratio values, and applied LogReg calibration, and PAV calibration.

The calculated target σ for bi-Gaussianized calibration were 3.65, 3.65, and 3.72 for the EER, LogReg, and KDE variants, respectively. Figure 14 shows the empirical cumulative distribution for the score data, and the target cumulative distribution for a perfectly calibrated bi-Gaussian system with $\sigma = 3.65$.

Figure 15 shows the "true" mapping function from scores to calibrated $\ln(\Lambda)$ given the Monte Carlo population distributions. Figure 15 also shows the mapping functions from scores to calibrated $\ln(\Lambda)$ for PAV, LogReg, and bi-Gaussianized calibration (with target $\sigma = 3.65$). For low and high score values, the "true" mapping is non monotonic. This is an artifact of the population distributions chosen. Score to calibrated $\ln(\Lambda)$ mappings for real data should be monotonic. References to "true" mapping in the remainder of this paragraph are with respect to its central monotonically increasing range (between score values of about −8 and +9). LogReg calibration results in a linear mapping which is close to the "true" mapping. PAV calibration results in a stepped mapping function in which the larger steps have relatively large deviations from the "true" mapping. In contrast to PAV calibration, bi-Gaussianized calibration results in a smoother mapping that generally stays closer to the "true" mapping. At low score values, the bi-Gaussianized calibration pulls the $\ln(\Lambda)$ values closer to 0 than does logistic regression, and at high score values, the bi-Gaussianized calibration pushes the $\ln(\Lambda)$ values further from 0 than does logistic regression.

Figures 16 and 17 show the different-source and same-source $\ln(\Lambda)$ distributions for the training data and the test data, respectively. The solid lines show the target distributions, the distributions for a perfectly calibrated bi-Gaussian system with $\sigma = 3.65$. Kernel-density plots are used to draw the LogReg and bi-Gaussianized calibration distributions. LogReg calibration only involves shifting and scaling in the $\ln(\Lambda)$ space, and the resulting distributions are shifted and scaled versions of the distributions of samples taken from the population distributions shown in Fig. 13. These LogReg calibrated distributions are relatively far from those of the perfectly calibrated bi-Gaussian target distributions. In contrast, bi-Gaussianized calibration involves non-linear (but still monotonic) warping, resulting in distributions that are much closer to those of the perfectly calibrated bi-Gaussian target distributions. Although the bi-Gaussianized calibrated distributions for the training data (Fig. 16) are due to training and testing on the same data, they are not perfect matches for the target distributions. This is because the bi-Gaussianized calibration mapping function is trained without the use of same-source and different-source labels. The bi-Gaussianized calibrated distributions for the test data (Fig. 17) are further from those of the perfectly calibrated bi-Gaussian target than are those for the training data (Fig. 16), but they are

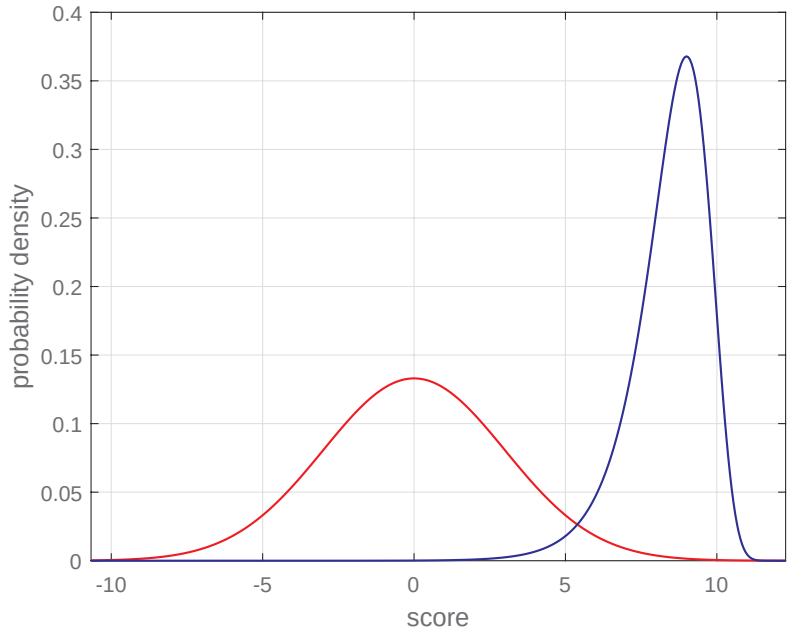

Figure 13. Distributions used for Monte Carlo simulation: Gaussian distribution and skewed distribution with smaller variance.

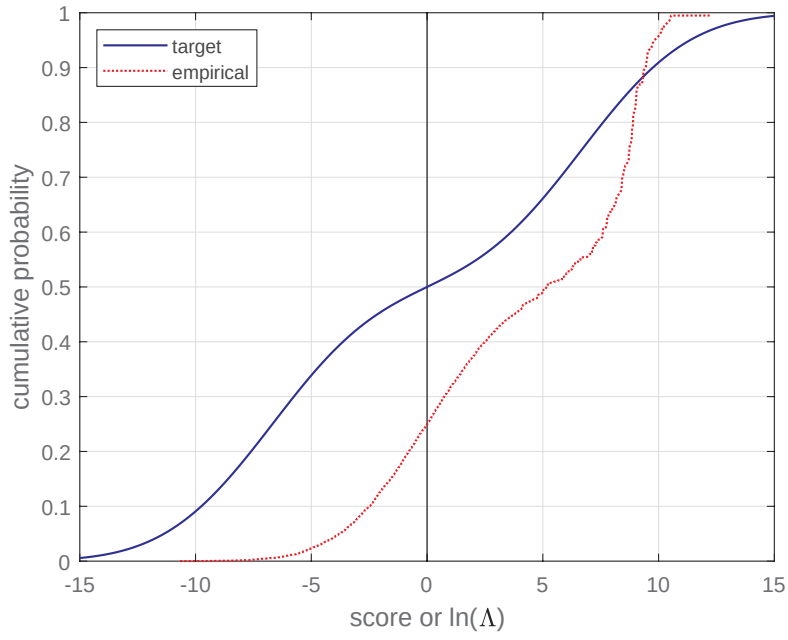

Figure 14. Empirical cumulative distribution of Gaussian and skewed-distribution score data, and target cumulative distribution of a perfectly calibrated bi-Gaussian system with $\sigma = 3.65$.

still much closer to the perfectly calibrated bi-Gaussian target distributions that are the LogReg calibrated distributions.

Table 3 gives the "true" C_{llr} values calculated using the Monte Carlo population distributions, and C_{llr} values for bi-Gaussianized calibration, LogReg calibration, and PAV calibration. All

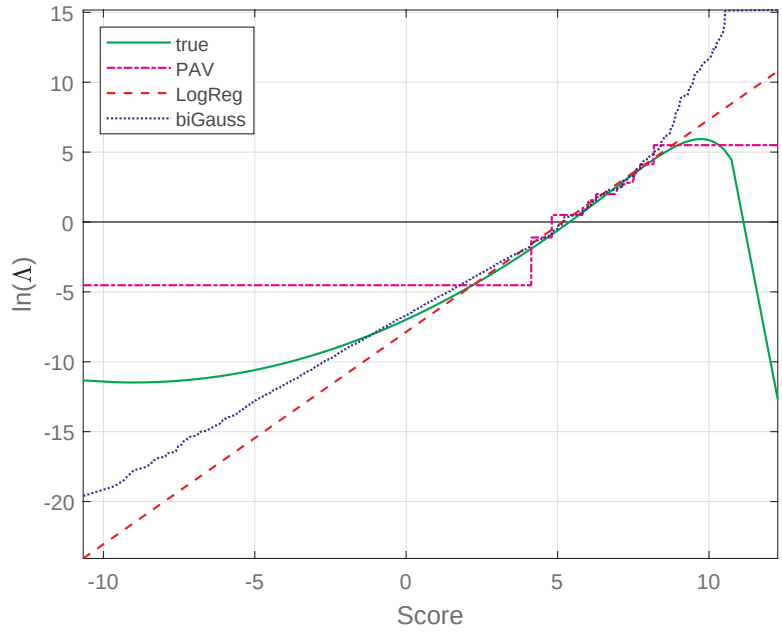

Figure 15. Mapping functions from scores to calibrated $\ln(\Lambda)$ for the synthetic Gaussian and skewed-distribution score data. The bi-Gaussianized-calibration mapping is for target $\sigma = 3.65$. The scale and range of the x-axis on this plot is the same as for the plot of the different-source and same-source distributions in Fig. 13.

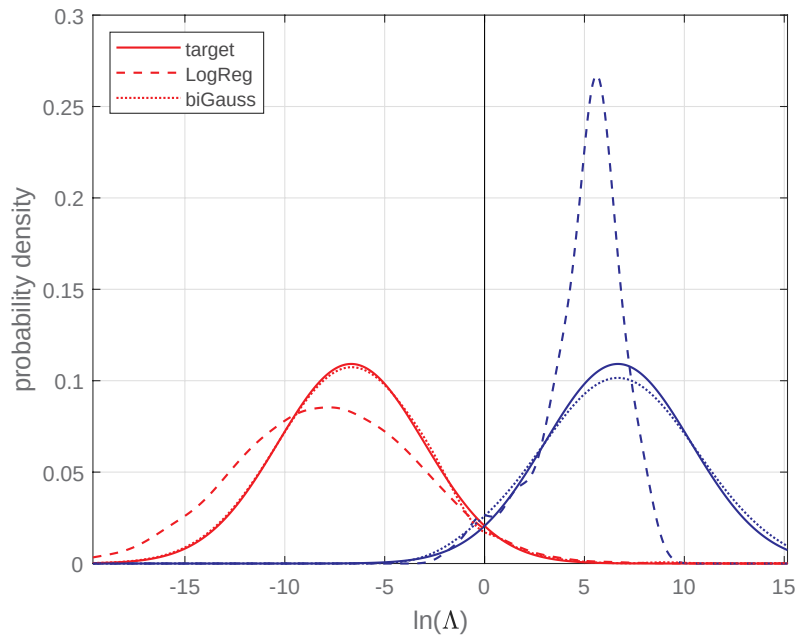

Figure 16. Different-source and same-source distributions of calibrated $\ln(\Lambda)$ for the synthetic Gaussian and skewed-distribution score data. The target and bi-Gaussianized distributions are for target $\sigma = 3.65$. Training data.

methods resulted in similar C_{llr} values, except for PAV calibration, which, due to overfitting, had a lower C_{llr} on the training data.

Figure 18 shows the Tippett plots for the target perfectly calibrated system, and for the logistic-regression and bi-Gaussianized calibrated likelihood-ratio values for the

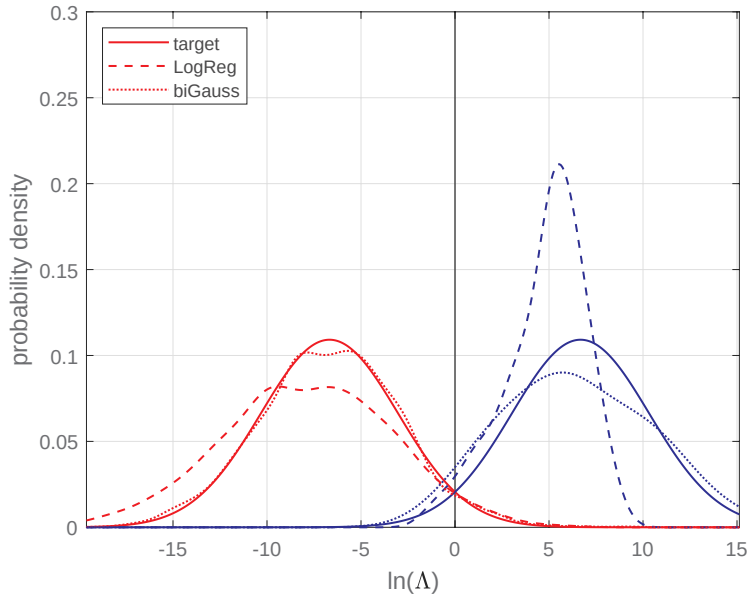

Figure 17. Different-source and same-source distributions of calibrated $\ln(\Lambda)$ for the synthetic Gaussian and skewed-distribution score data. The target and bi-Gaussianized distributions are for target $\sigma = 3.65$. Test data.

Table 3. C_{llr} values for different calibration methods applied to the Gaussian and skewed-distribution data.

		Method				
	"true"	Bi-Gauss			LogReg	PAV
Data		EER	LogReg	KDE		
train	0.127	0.129	0.129	0.129	0.126	0.113
test	0.138	0.140	0.140	0.140	0.134	0.137

test data.[14] The bi-Gaussianized calibrated results are closer to the target perfectly calibrated bi-Gaussian system than are the logistic-regression calibrated results. This is because, as mentioned above in relation to Fig. 15, at low score values, the bi-Gaussianized calibration pulls the $\ln(\Lambda)$ values closer to 0 than does logistic regression, and at high score values, the bi-Gaussianized calibration pushes the $\ln(\Lambda)$ values further from 0 than does logistic regression.

One should note that the curves representing the perfectly calibrated bi-Gaussian system in Figs 16–18 (and parallel figures in subsections below) do not represent the distributions of "true" likelihood ratios, which (except in the context of Monte Carlo simulation) are unknown and unknowable. All three systems represented in these figures (LogReg, bi-Gaussianized, and perfect calibration) have, by design, (approximately) the same C_{llr} for the training data. The comparisons between the different systems that these figures allow are comparisons of how well calibrated the likelihood-ratio outputs of different systems with (approximately) the same C_{llr} are. It is not a comparison of the likelihood-ratio outputs of different systems with "true" likelihood ratios. Given two different systems with (approximately) the same C_{llr}, and so equally good performance on this metric, we argue that the better of the two systems is the system which

[14] For the calculation of the cumulative proportions for Tippett plots, we used denominators of $N_{\text{d}}+1$ and $N_{\text{s}}+1$ rather than N_{d} and N_{s}. This prevents the maximum cumulative proportion from reaching 1, which is appropriate for plotting the cumulative-density functions $F(ln(-\Lambda_{\text{d}})|\mu_{\text{d}},\sigma)$ and $F(ln(\Lambda_{\text{s}})|\mu_{\text{s}},\sigma)$ for the distributions of the perfectly calibrated system. $F(ln(-\Lambda_{\text{d}})|\mu_{\text{d}},\sigma) = 1$ and $F(ln(\Lambda_{\text{s}})|\mu_{\text{s}},\sigma) = 1$ would only occur when $\ln(\Lambda_{\text{d}}) = -\infty$ and $\ln(\Lambda_{\text{s}}) = +\infty$, respectively. The x axes of Tippett plots are scaled in base-ten logarithms rather then natural logarithms, but this is simply a difference in scaling.

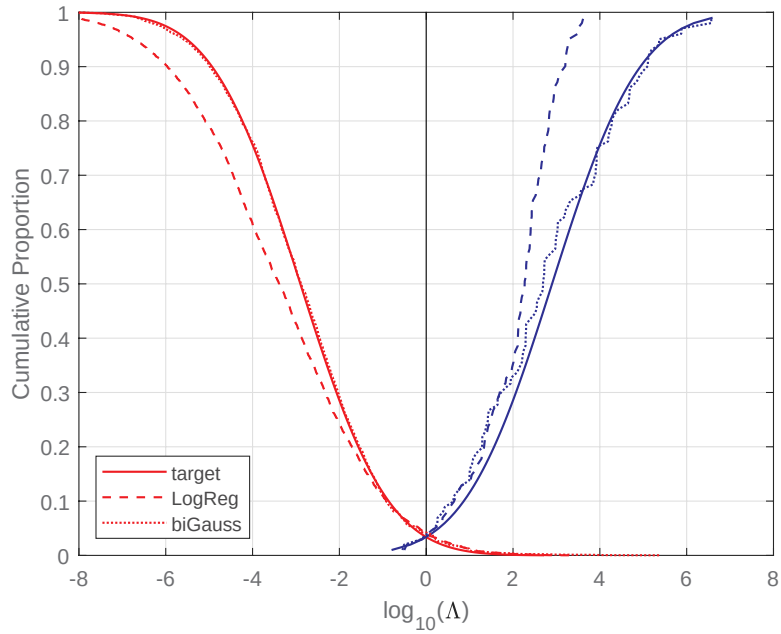

Figure 18. Tippett plots for the target perfectly calibrated system with $\sigma = 3.65$, and for the LogReg calibrated and bi-Gaussianized calibrated likelihood-ratio values. Synthetic Gaussian and skewed-distribution score data. Test data.

is closer to a perfectly calibrated system with that C_{llr}.[15] If one accepts this argument, then Figs 16–18 (and parallel figures in subsections below) show that bi-Gaussianized calibration is better than LogReg calibration.

3.4 Real data: Forensic voice comparison

Real score data were taken from the E^3 Forensic Speech Science System (E^{3}FS^3) and applied to the benchmark *forensic-eval_01* data, see Weber et al. (2022).[16] There were 111 same-source scores and 9719 different-source scores originating from a total of 223 recordings of 61 speakers. Kernel-density plots of the different-source and same-source score distributions are shown in Fig. 19. These data clearly deviate from the assumption of equal-variance Gaussians (although one may consider this a moderate deviation).

We applied bi-Gaussianized calibration (EER, LogReg, and KDE variants), LogReg calibration, and PAV calibration using leave-one-out/leave-two-out cross-validation.[17] Over cross-validation loops, the mean target σ for bi-Gaussianized calibration was 3.75, 3.44, and 3.45 for the EER, LogReg, and KDE variants respectively. Figure 20 shows the empirical cumulative distribution for the score data, and the target cumulative distribution for a perfectly calibrated bi-Gaussian system with $\sigma = 3.44$. The mapping functions are shown in Fig. 21. Relative to the logistic-regression mapping, the bi-Gaussianized calibration mapping pulls both large negative scores and moderate-to-large positive scores closer to $\ln(\Lambda) = 0$.

Figure 22 shows the different-source and same-source $\ln(\Lambda)$ distributions. The solid lines show the target distributions, the distributions for a perfectly calibrated bi-Gaussian system with $\sigma = 3.44$. Kernel-density plots are used to draw the logistic-regression and bi-Gaussianized

[15] The family of perfectly calibrated systems of distributions that we have chosen to use is bi-Gaussian systems. If we had chosen some other family of perfectly calibrated systems, our method would have mapped the test data toward the distributions of the member of that family with the C_{llr} corresponding to that of the training data after the initial calibration step.

[16] The version of E^{3}FS^3 used was a later version than that used in Weber et al. (2022) and had slightly better performance. For display purposes, we rescaled the score values to 2.5% of their raw values. This linear rescaling has no impact on the results of subsequent application of calibration methods.

[17] For a same-source comparison, all scores from comparisons that involved the source contributing to the score being calibrated were excluded from training. For a different-source comparison, all scores from comparisons that involved either or both of the sources contributing to the score being calibrated were excluded from training.

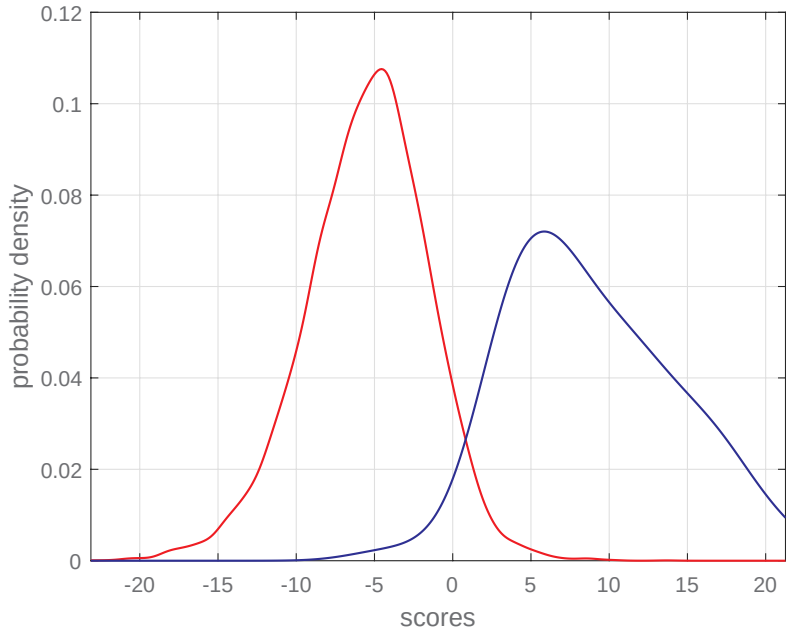

Figure 19. Kernel-density plots of the different-source and same-source score distributions from comparison of voice recordings.

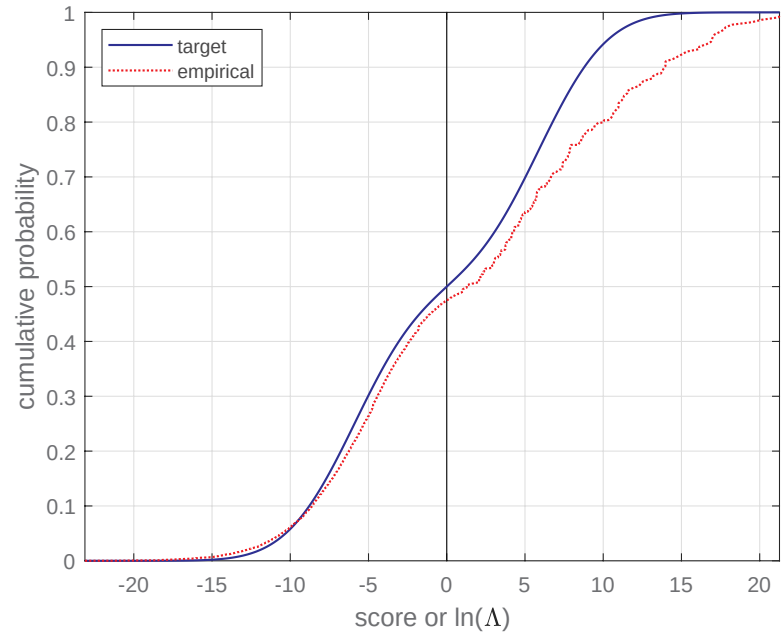

Figure 20. Empirical cumulative distribution of voice-comparison score data, and target cumulative distribution of a perfectly calibrated bi-Gaussian system with $\sigma = 3.44$. The scale and range of the x-axis on this plot is the same as for the plot of the different-source and same-source distributions in Fig. 19.

calibration distributions. Figure 23 shows the corresponding Tippett plots. In both the probability-density plots and the Tippett plots, it can be seen that the bi-Gaussianized calibration results are closer to the perfectly calibrated bi-Gaussian system than are the logistic-regression

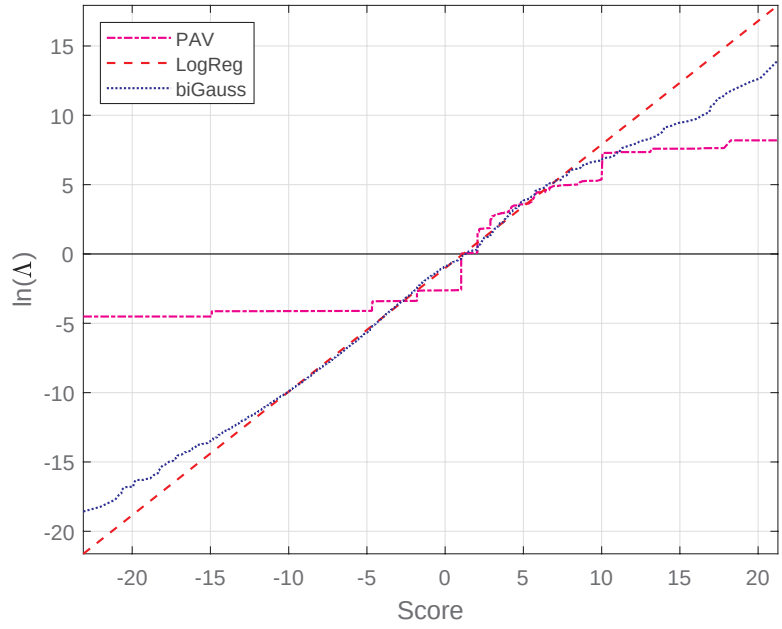

Figure 21. Mapping functions from scores to calibrated $\ln(\Lambda)$ for the voice-comparison score data. The bi-Gaussianized-calibration mapping is for target $\sigma = 3.44$. The scale and range of the *x*-axis on this plot is the same as for the plot of the different-source and same-source distributions in Fig. 19.

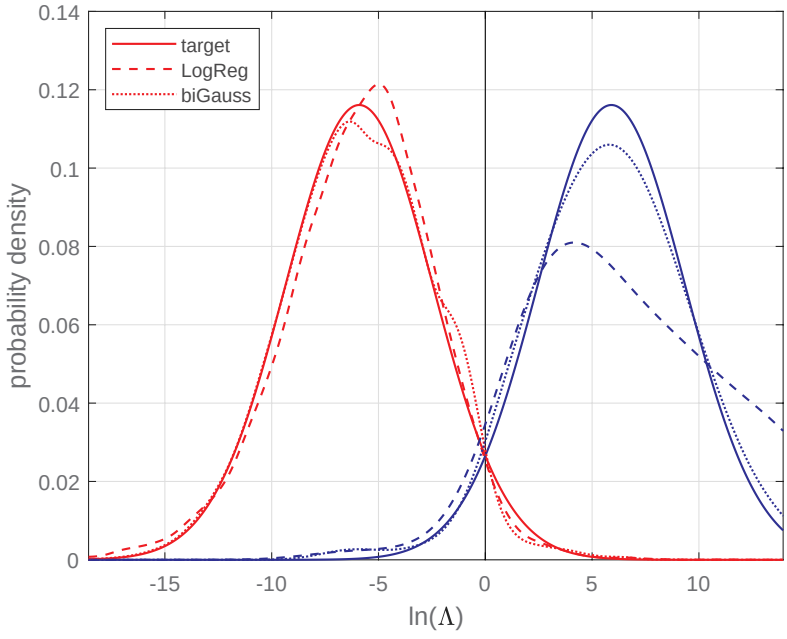

Figure 22. Different-source and same-source distributions of calibrated $\ln(\Lambda)$ for the voice-comparison score data. The target and bi-Gaussianized distributions are for target $\sigma = 3.44$.

calibration results. The logistic-regression calibration results deviate particularly for moderate-to-large log-likelihood-ratio values, these log-likelihood-ratio values are higher than for the perfectly calibrated bi-Gaussian system.

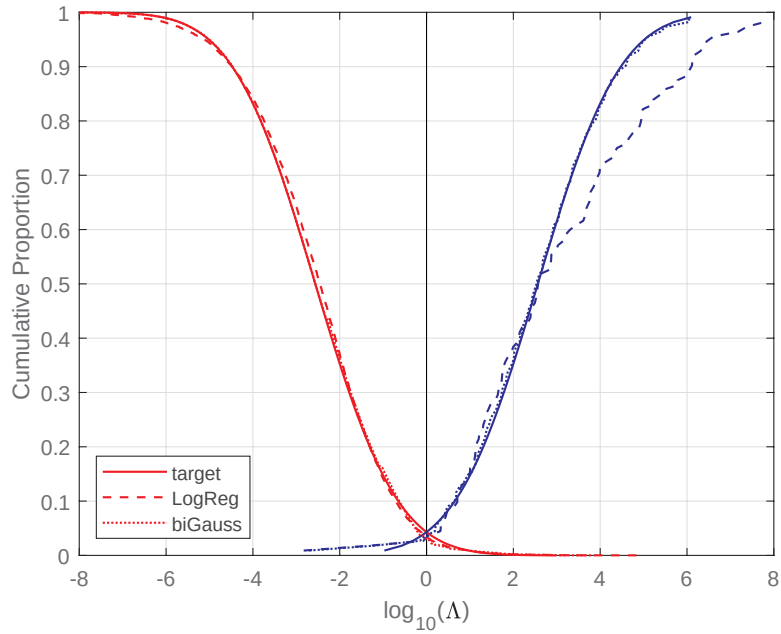

Figure 23. Tippett plots for the target perfectly calibrated system with $\sigma = 3.44$, and for the LogReg and bi-Gaussianized calibrated likelihood-ratio values. Voice-comparison score data.

Table 4. C_{llr} values for different calibration methods applied to the speaker data.

Method				
Bi-Gauss			LogReg	PAV
EER	LogReg	KDE		
0.171	0.172	0.171	0.172	0.168

Table 4 gives the C_{llr} values resulting from the application of bi-Gaussianized calibration (EER, LogReg, and KDE variants), LogReg calibration, and PAV calibration. All the C_{llr} values were approximately the same.

3.5 Real data: Glass fragments

Real score data were taken from comparison of glass fragments by Vergeer et al. (2016) and van Es et al. (2017).[18] The glass-fragment data consisted of multiple fragments from each of 320 sources, resulting in 320 same-source scores and 51,040 different-source scores. Due to numerical limitations in the software that calculated the scores, 41,108 (∼80%) of the different-source scores had a value of $-\infty$. We converted the value of these scores to the lowest finite score value that already existed in the dataset, thus producing a probability mass at that value. Kernel-density plots of the different-source and same-source score distributions are shown in Fig. 24.[19] The different-source distribution has a point mass at approximately –300, and probability density spread between –300 and 6.5. Apart from a few outliers, the same-source density is concentrated in a relatively narrow range of low positive values, with the mode at 9.0.

[18] The glass score data were kindly provided by Peter Vergeer of the Netherlands Forensic Institute. The scores had been calculated using the multivariate-kernel-density method of Aitken & Lucy (2004). The raw data (as opposed to the scores) are available from: https://github.com/NetherlandsForensicInstitute/elemental_composition_glass

[19] To plot the KDEs in Figure 24, the Gaussian-approximation method was used to determine the bandwidth to use for the same-source scores, then that same bandwidth was used for the different-source scores.

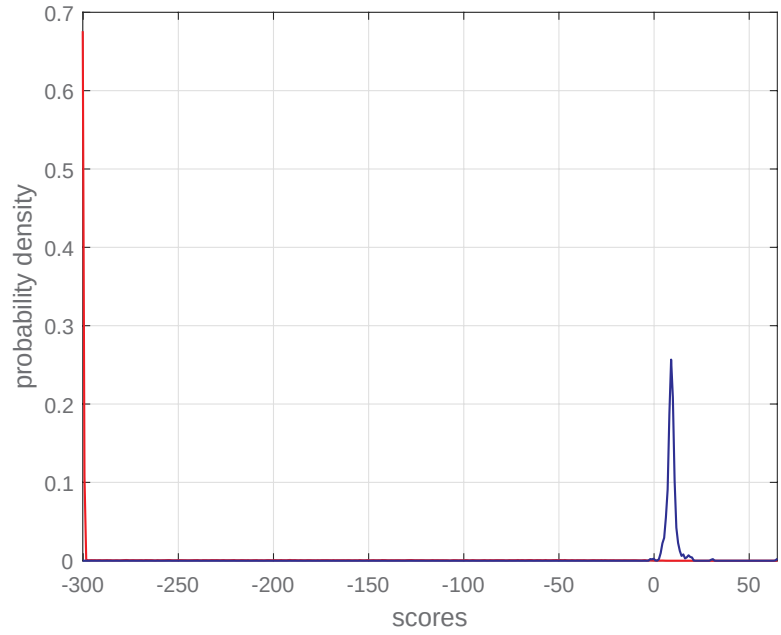

Figure 24. Kernel-density plots of the different-source and same-source score distributions from comparison of glass fragments.

The different-source and same-source scores overlap in the range –1.7 to 6.5. These distributions exhibit extreme deviation from the assumption of equal-variance Gaussians.

We applied bi-Gaussianized, LogReg, and PAV calibration using leave-one-out/leave-two-out cross-validation.[20] Averaged over cross-validation loops, the target σ for bi-Gaussianized calibration were 5.65, 6.02, and 6.01 for the EER, LogReg, and KDE variants respectively. Figure 25 shows the empirical cumulative distribution for the score data, and the target cumulative distribution for a perfectly calibrated bi-Gaussian system with $\sigma = 6.02$. The probability mass accounting for ∼80% of the different-source-scores results in the empirical distribution beginning at ∼0.4. The mapping functions are shown in Figs 26 and 27 (Fig. 27 shows details of Fig. 26 around the range of values in which the different-source and same-source scores overlap). Whereas the LogReg mapping is linear and maps to a broad range of $\ln(\Lambda)$ values, the bi-Gaussianized calibration mapping is sigmoidal and maps to a much narrower range of $\ln(\Lambda)$ values. Whereas, outside the range of values in which the different-source and same-source scores overlap (–1.7 to 6.5), the PAV mapping is flat, the bi-Gaussianized calibration mapping has monotonically increasing slopes which reach below and above the minimum and maximum PAV-mapped $\ln(\Lambda)$ values. Far from the overlap range, those slopes are shallow. In the range of values in which the different-source and same-source scores overlap (–1.7 to 6.5), the bi-Gaussianized calibration mapping and the PAV mapping are close to each other, but the bi-Gaussianized calibration mapping is smoother.[21]

Figure 28 shows the different-source and same-source $\ln(\Lambda)$ distributions. The solid lines show the target distributions, the distributions for a perfectly calibrated bi-Gaussian system with

[20] Because otherwise the cross-validation would have taken an extremely long time, we excluded test scores that had been $-\infty$. Scores that had been $-\infty$ (and had been converted to the lowest finite score value in the dataset) were, however, still included in training. We then assigned the lowest $\ln(\Lambda)$ value calculated for a finite test scores as the $\ln(\Lambda)$ value corresponding to all the test scores that had been $-\infty$. Had we included the test scores that had been $-\infty$ in the cross-validation, because of differences in training data from loop to loop, each would have resulted in a slightly different $\ln(\Lambda)$ value. On the particular machine, we used (using 19 parallel workers), it took 11 hours to run the cross-validation.

[21] The larger steps in the bi-Gaussianized calibration's mapping function are due to sparce same-source scores. The first three same-source scores occur at −1.7, −0.3, and 3.0, which correspond to the end of the rise of the first large step up, the second large step up, and the step up following the second large stepup.

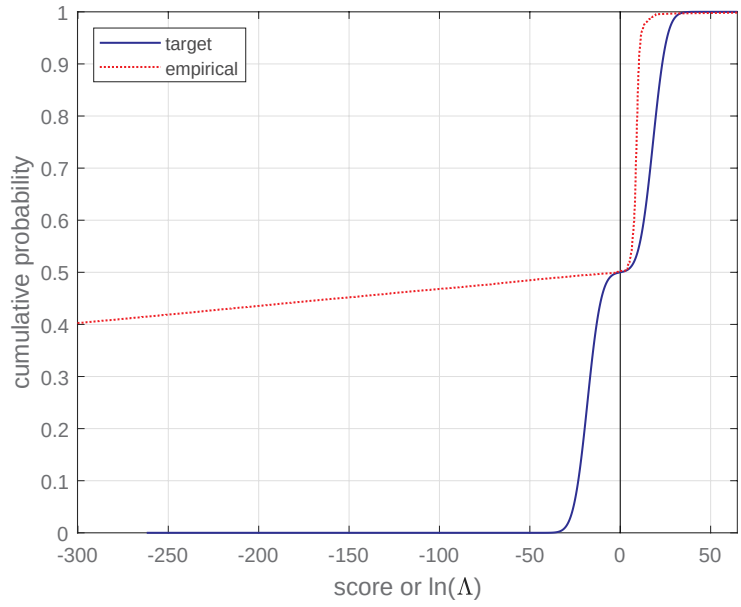

Figure 25. Empirical cumulative distribution of glass-fragment score data, and target cumulative distribution of a perfectly calibrated bi-Gaussian system with $\sigma = 6.02$. The scale and range of the x-axis on this plot is the same as for the plot of the different-source and same-source distributions in Fig. 24.

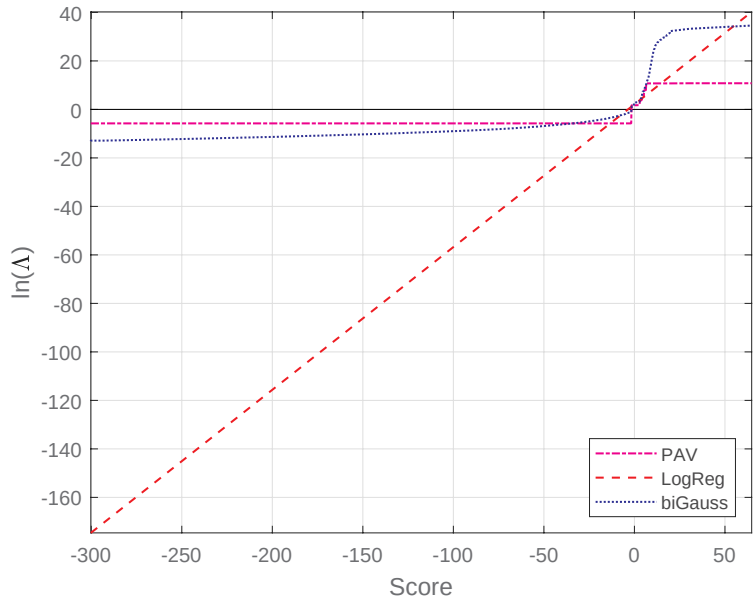

Figure 26. Mapping functions from scores to calibrated $\ln(\Lambda)$ for the glass-fragment score data. The bi-Gaussianized-calibration mapping is for target $\sigma = 6.02$. The scale and range of the x-axis on this plot is the same as for the plot of the different-source and same-source distributions in Fig. 24.

$\sigma = 6.02$. Kernel-density plots were used to draw the LogReg and bi-Gaussianized calibrated distributions. For the different-source bi-Gaussianized calibration distribution, the probability mass, which is at $\ln(\Lambda) = -12.9$, is represented as a spike (the probability-density axis is

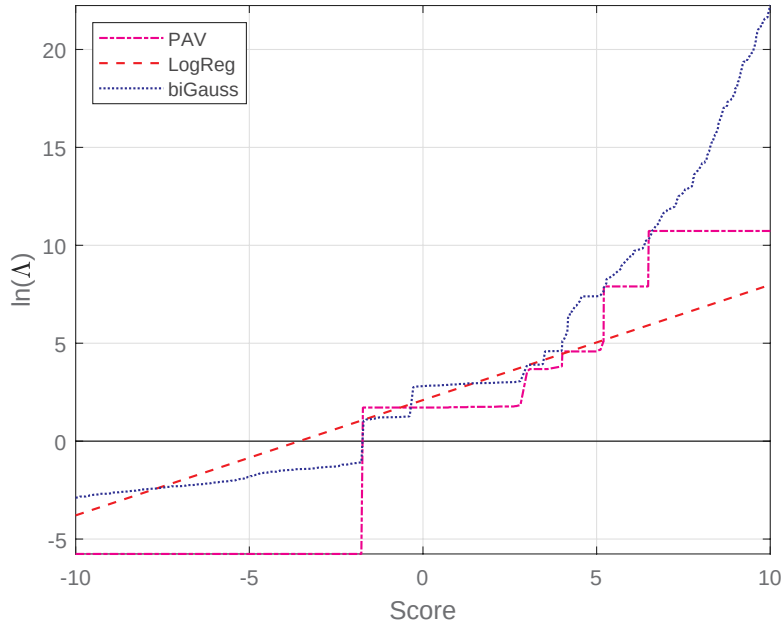

Figure 27. Part of the mapping functions from scores to calibrated $\ln(\Lambda)$ for the glass-fragment score data (part of Fig. 26), showing details around the range of values in which the different-source and same-source scores overlap (–1.7 to 6.5).

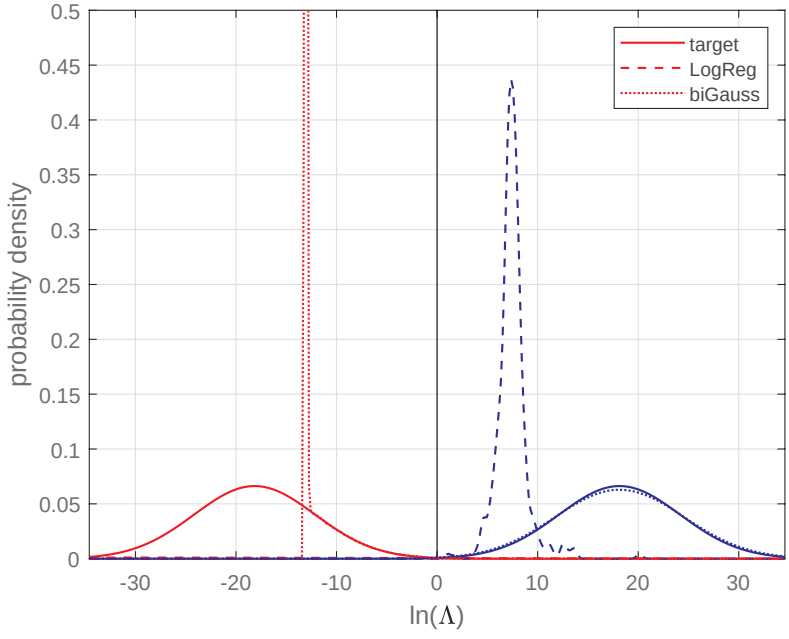

Figure 28. Different-source and same-source distributions of calibrated $\ln(\Lambda)$ for the glass-fragment score data.

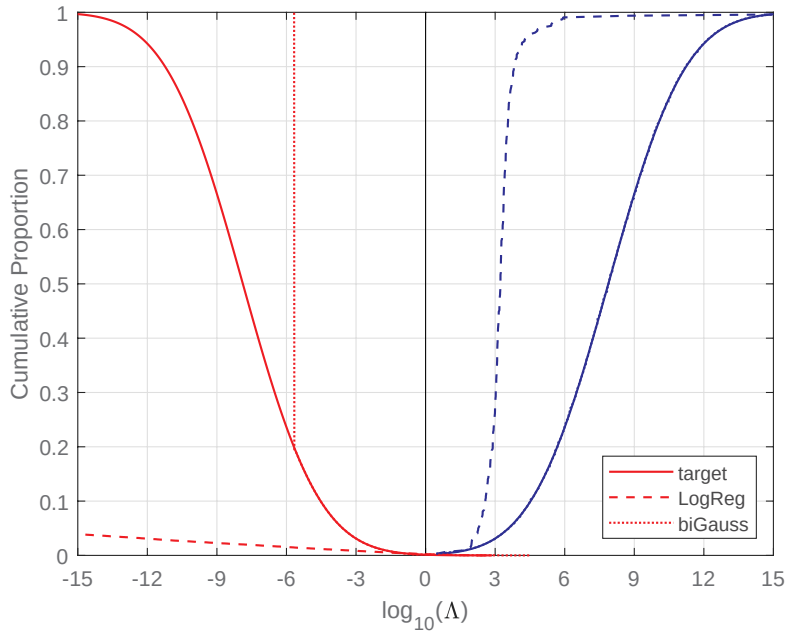

Figure 29. Tippett plots for the target perfectly calibrated system with $\sigma = 6.02$, and for the LogReg and bi-Gaussianized calibrated likelihood-ratio values. Glass-fragment score data.

truncated at 0.5, the spike reaches much higher).[22] The remainder of the different-source bi-Gaussianized distribution is represented as a curve to the right of the spike, which is mostly obscured by the different-source curve of the target bi-Gaussian distribution. The different-source LogReg calibrated distribution rises very slowly, and eventually has a probability mass at $\ln(\Lambda) = -175$ (outside the range of plotted values). Other than at their probability mass, the bi-Gaussianized calibration results are much closer to perfectly calibrated bi-Gaussian target distributions than are the LogReg calibration results.

Figure 29 shows the Tippett plots. The vertical part of the different-source curve for the bi-Gaussianized calibration (at $\log_{10}(\Lambda) = -5.67$) is due to the probability mass, otherwise, the bi-Gaussianized calibration curves lie almost directly on top of the perfectly calibrated bi-Gaussian target curves (and are mostly obscured by the latter). In contrast, the LogReg calibration curves are generally far from the perfectly calibrated bi-Gaussian target curves.

For the bi-Gaussianized calibration, the probability mass occurs at $\log_{10}(\Lambda) = -5.67$, which is a likelihood ratio of approximately 468,000 in favor of the different-source hypothesis. One could report that the likelihood ratio is at least 468,000 in favor of the different-source hypothesis.

Table 5 gives the C_{llr} values resulting from the application of bi-Gaussianized calibration (EER, LogReg, and KDE variants), LogReg calibration, and PAV calibration. All the C_{llr} values were approximately the same.

The maximum likelihood-ratio value generated by bi-Gaussianized calibration was very large, 1.05×10^{15}. Concerns over whether such large likelihood-ratio values are justified given relatively limited amounts of training data have led to proposals that the reported values of likelihood ratios be limited in size or that they be shrunk toward the neutral value of 1. Vergeer et al. (2016) proposed a method to limit likelihood-ratio values, Empirical Lower and Upper Bounds (ELUB), which when applied to the same glass data as in the article limited $\log_{10}(\Lambda)$ to the range

[22] To represent the probability mass as a spike, the bandwidth of the kernel for the different-source distribution was manually set to 0.1.

Table 5. C_{llr} values for different calibration methods applied to the glass data.

Method				
Bi-Gauss			LogReg	PAV
EER	LogReg	KDE		
0.006	0.005	0.005	0.006	0.006

-2.50 to 4.53.[23] Smaller calculated values would be replaced by $\log_{10}(\Lambda) = -2.50$, and larger calculated values would be replaced by $\log_{10}(\Lambda) = 4.53$. Morrison and Poh (2018) proposed a method to shrink log-likelihood-ratio values toward 0 by using a large regularization weight for regularized-logistic-regression calibration. The latter method could be applied in the LogReg variant of bi-Gaussianized calibration and would result in a smaller target σ value than if no regularization were applied or if a small regularization weight were used (this article used a small regularization weight). Bi-Gaussianized calibration potentially offers another method for inducing shrinkage: The magnitude of the target σ value could be reduced from its calculated value, for example, it could be reduced to 90% of the value calculated using Equation (8) or Equation (10). This would result in the bi-Gaussianized calibrated log-likelihood ratios being closer to 0 than would otherwise be the case. As with other methods for inducing shrinkage, one would have to make a choice as to how much shrinkage to induce (a choice which should make reference to the amount of training data used). Note also, that inducing substantial shrinkage leads to larger C_{llr}. We do not explore this shrinkage method in the present paper.

4. Effect of sampling variability

In Section 3, the bi-Gaussianized calibration method was demonstrated using four different data sets (two simulated and two real). Each dataset was treated as a sample from a relevant population. Different samples from the same population would be expected to produce different results. In the present section, we explore the effect of sampling variability on the output of bi-Gaussianized, LogReg, and PAV calibration. Smaller samples would be expected to result in greater sampling variability, and the size of case-relevant samples that can practically be obtained in the context of forensic cases is often small. In addition to testing samples of 100 sources (as for the Monte Carlo simulations in Sections 3.2 and 3.3), we also test samples of 50 sources.

As in Section 3.2, we generated synthetic data using Monte Carlo simulation. The generating different-source distribution was a Gaussian with parameters $\mu_d = -4.5$ and $\sigma = 3$, and the generating same-source distribution was a Gaussian with parameters $\mu_s = 4.5$ and $\sigma = 3$, that is, the perfectly calibrated bi-Gaussian system shown in Fig. 2a and in the third row of Fig. 3. We generated a single test-data sample set consisting of 10,000 same-source scores and 10,000 different-source scores. By using a single large balanced test set, we can attribute any bias and variability in the results to bias in the calibration methods and sampling variability in the training sets. We generated 1,000 training-data sample sets, each consisting of 100 same-source scores and 4,950 different-source scores (the 100 source datasets), and another 1,000 training-data sample sets, each consisting of 50 same-source scores and 1,225 different-source scores (the 50 source datasets). For each training-data sample set, we calibrated the test-data sample set using bi-Gaussianized calibration (EER, LogReg, and KDE variants), LogReg calibration, and PAV calibration, and calculated the C_{llr} for the resulting sets of $\ln(\Lambda)$.

Figures 30 and 31 show violin plots of the distributions of C_{llr} values for the 100 source datasets, and Figs 32 and 33 show violin plots of the distributions of C_{llr} values for the 50 source datasets. The first figure in each pair uses a C_{llr} range of 0 to 1, giving an impression of the distributions of C_{llr} values relative to the possible range of C_{llr} for well-calibrated systems. The second

[23] The bounds are calculated after a calibration model has been applied, so their values will depend on the particular calibration model used. These particular values were the result of calibration that fitted a KDE to the different-source scores and a double-exponential model to the same-source scores.

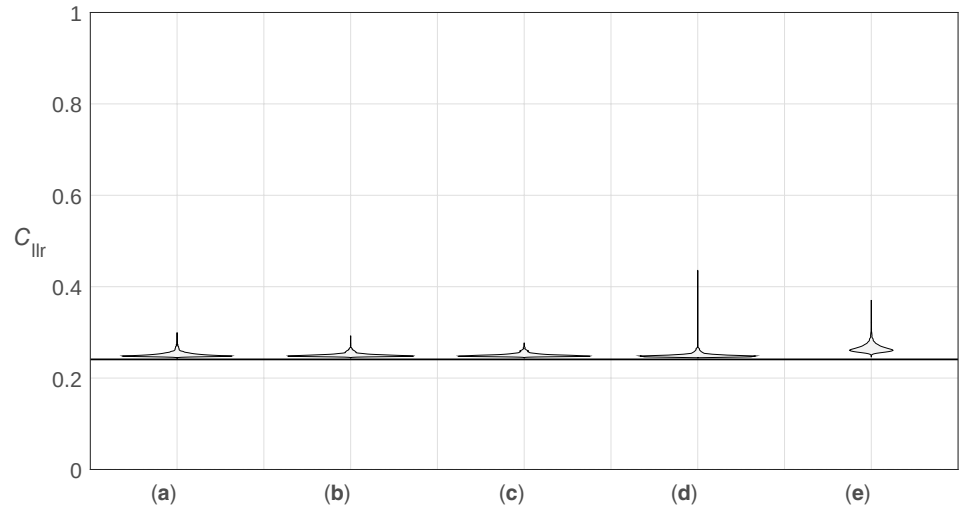

Figure 30. Violin plots of distributions of target C_{llr} values resulting from the application of different calibration methods using 1,000 training-data sample sets (100 source datasets). (a) Bi-Gaussianized calibration EER variant. (b) Bi-Gaussianized calibration LogReg variant. (c) Bi-Gaussianized calibration KDE variant. (d) LogReg calibration. (d) PAV calibration. The solid horizontal line represents the "true" $C_{llr} = 0.241$.

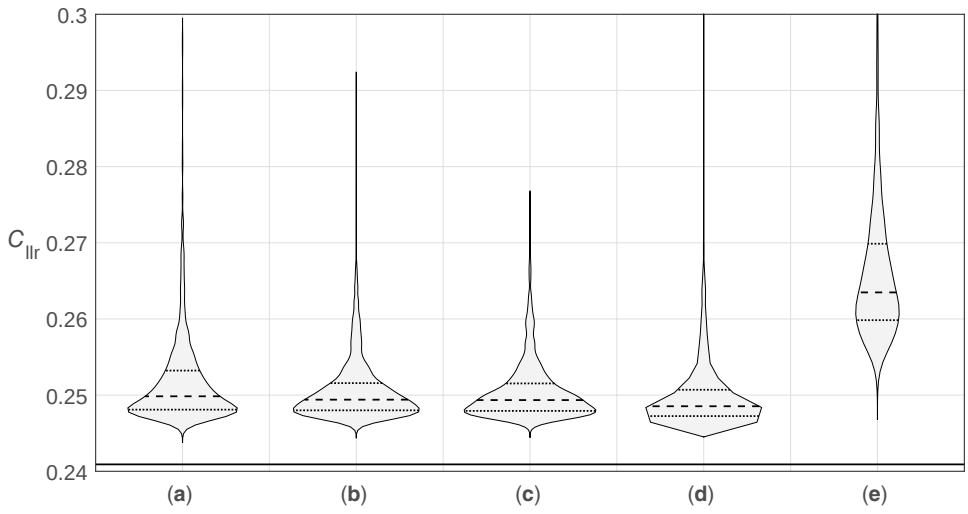

Figure 31. Violin plots of distributions of target C_{llr} values resulting from the application of different calibration methods using 1,000 training-data sample sets (100 source datasets). (a) Bi-Gaussianized calibration EER variant. (b) Bi-Gaussianized calibration LogReg variant. (c) Bi-Gaussianized calibration KDE variant. (d) LogReg calibration. (e) PAV calibration. The solid horizontal line represents the "true" $C_{llr} = 0.241$.

figure in each pair zooms in to show the shapes of the C_{llr} distributions—note that the range of values on the *y* axis of Fig. 33 is twice that of Fig. 31. In each figure, the solid horizontal line represents the "true" C_{llr} value, 0.241, which corresponds to a perfectly calibrated bi-Gaussian system with $\sigma = 3$.[24]

For all calibration methods and all training-data samples, calculated C_{llr} was greater than the "true" C_{llr}, that is, no sample-based method outperformed a calculation based on the population's parameter values. With the exception of PAV calibration and of a few outliers, however,

[24] The C_{llr} value was calculated using Equation (7).

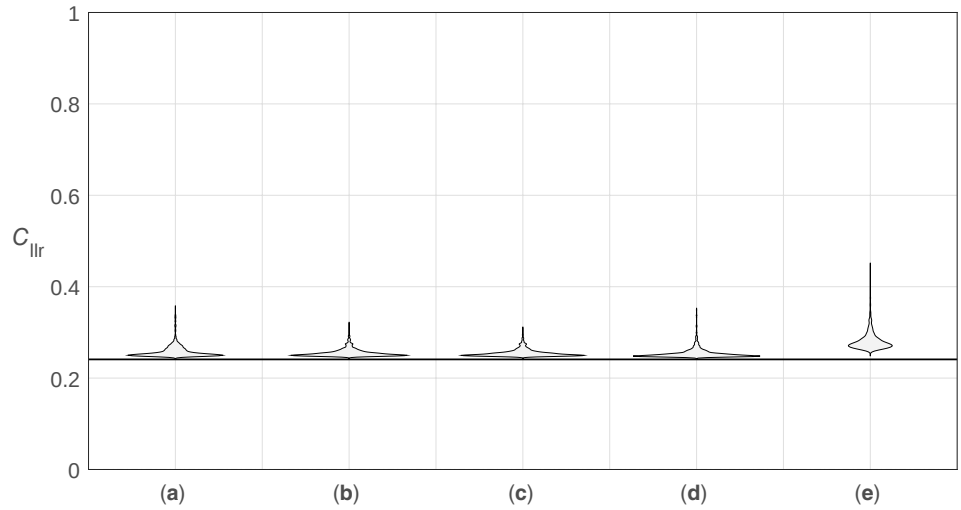

Figure 32. Violin plots of distributions of target C_{llr} values resulting from the application of different calibration methods using 1,000 training-data sample sets (50 source datasets). (a) Bi-Gaussianized calibration EER variant. (b) Bi-Gaussianized calibration LogReg variant. (c) Bi-Gaussianized calibration KDE variant. (d) LogReg calibration. (e) PAV calibration. The solid horizontal line represents the "true" $C_{llr} = 0.241$.

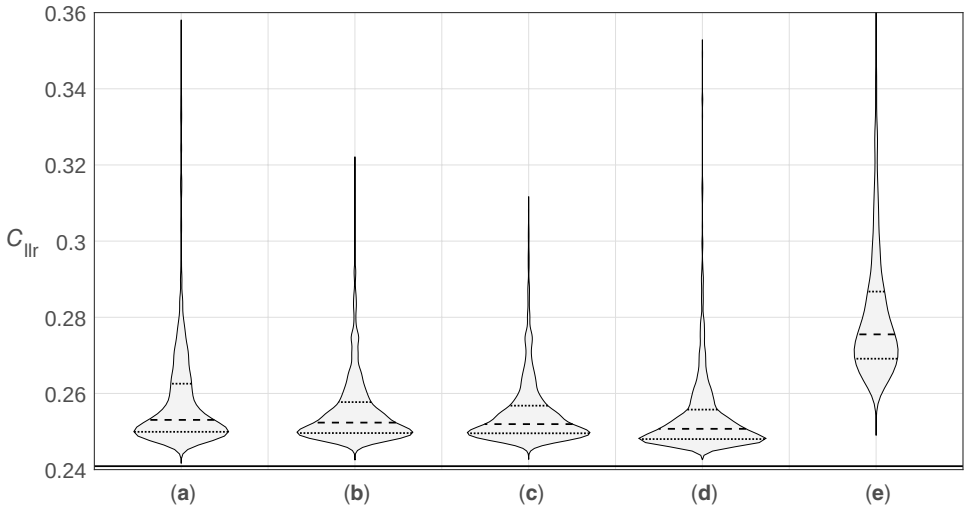

Figure 33. Violin plots of distributions of target C_{llr} values resulting from the application of different calibration methods using 1,000 training-data sample sets (50 source datasets). (a) Bi-Gaussianized calibration EER variant. (b) Bi-Gaussianized calibration LogReg variant. (c) Bi-Gaussianized calibration KDE variant. (d) LogReg calibration. (e) PAV calibration. The solid horizontal line represents the "true" $C_{llr} = 0.241$.

all calculated C_{llr} were only a little higher than the "true" C_{llr}. As would be expected, calculated C_{llr} values tended to be higher for the 50 source datasets than for the 100 source datasets.

The highest C_{llr} values (the worst performance) occurred for PAV calibration—PAV overfits the training data. For bi-Gaussianized calibration, the median and quartile C_{llr} values were similar across variants, but were slightly higher for the EER variant, than for the LogReg variant, and slightly higher for the LogReg variant than for the KDE variant. Also, outliers were most extreme for the EER variant, less extreme for the LogReg variant, and least extreme for the KDE variant. For LogReg calibration, the median and quartile C_{llr} were slightly lower than for the KDE variant of bi-Gaussianized calibration, but it produced much more extreme outliers. Even

though, by design, on the training data, the LogReg variant of bi-Gaussianized calibration should have (approximately) the same C_{llr} as LogReg calibration, on the test data, outliers for the LogReg variant of bi-Gaussianized calibration were not as extreme as for LogReg calibration.

Based on these C_{llr} results, the KDE variant appears to be the best-performing variant of bi-Gaussianized calibration. Taking into account the outliers in these C_{llr} results, the KDE variant of bi-Gaussianized calibration also appears to be a better choice than LogReg calibration.

We separately ran the 1,000 training datasets (100 source versions) for each variant of bi-Gaussianized calibration, and for LogReg calibration, and timed how long each took (including the time to generate the training data). On the particular machine we used (using three parallel workers), the EER variant took 5.8 s, the LogReg variant took 7.8 s, the KDE variant took 131 s, and LogReg calibration took 3.3 s. A disadvantage of the KDE variant of bi-Gaussianized calibration, therefore, is that it takes much longer than the other variants of bi-Gaussianized calibration and much longer than LogReg calibration. If time were an issue, the LogReg variant of bi-Gaussianized calibration might be a better choice than the KDE variant.

5. Graphical representation of likelihood-ratio output

A biproduct of bi-Gaussianized calibration is that it provides a way of graphically representing results which may aid in explaining them to triers of fact or to other interested parties. Since logistic regression is a discriminative method, rather than a generative method, it does not actually calculate the ratio of two likelihoods, thus the output of LogReg calibration cannot be directly graphically represented as the relative heights of two probability-density curves. If one has used bi-Gaussianized calibration, however, one can plot the same-source target probability-density function and the different-source target probability-density function, and graphically show the relative height of the two curves at a value of interest.

Figure 34 provides an example using a perfectly calibrated system with $\sigma = 3$. One can explain that the better the performance of the system under the conditions of the case,[25] the further apart the different-source and the same-source curves will be, that is, the less overlap there will be between the two curves. The x-axis is labeled "likelihood ratio." This axis has a logarithmic scale, but the values along the axis are written in linear form. Imagine that the likelihood-ratio value calculated for the comparison of the questioned- and known-source items was 10. We find the corresponding location on the x-axis, and draw a vertical line that intersects the same-source probability-density curve and the different-source probability-density curve. We highlight the intersect points and draw horizontal lines from the intersect points to the y-axis. The y-axis is labeled "relative likelihood," and is scaled so that the relative likelihood of the lowest of the two aforementioned intersects has a value of 1.[26] In this example, the y-axis is scaled so that the intersect with the different-source curve has a relative-likelihood value of 1. In this example, given this scaling, the intersect with the same-source curve has a relative-likelihood value of 10. It is then easy to explain that, for the reported likelihood-ratio value for the comparison of the questioned- and known-source items, the relative likelihood for obtaining that value if the same-source hypothesis were true is $10/1 = 10$ times greater than the relative likelihood for obtaining that value if the different-source hypothesis were true, which is what the reported likelihood-ratio value on the x axis means. This will work for any likelihood-ratio value selected on the x-axis. We leave it as a task for future research to assess whether graphics of this form will actually be helpful for explaining the meaning of likelihood ratios to triers of fact.

6. Conclusion

For the output of a perfectly calibrated forensic evaluation system, the likelihood ratio of the likelihood ratio is the likelihood ratio. If the distributions of the different-source and same-source natural-log-likelihood ratios, $\ln(\Lambda)$, output by the system are both Gaussian and they have the same variance, σ^2, and the different-source and same-source means are $\mu_d = -\frac{\sigma^2}{2}$ and

[25] The system should have been calibrated and validated using data that are representative of the relevant population for the case and reflective of the conditions of the case (Morrison et al., 2021).

[26] If the intersect were very low, we might scale the y axis so that the intersect value is 1, but draw tick marks on the y axis at 10, 20, 30, etc. or at 100, 200, 300, etc. If both intersects were low, we might add a zoomed-in view.

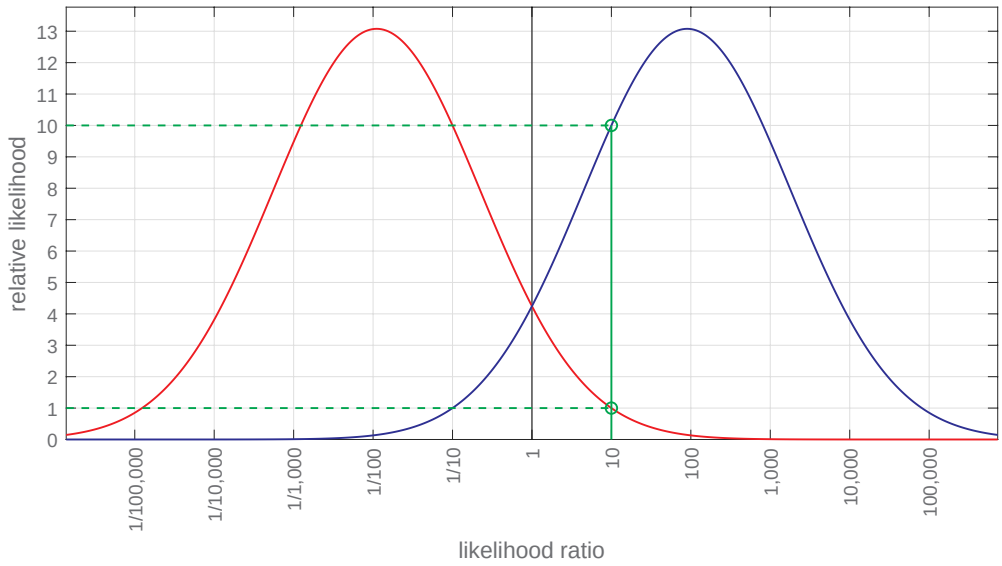

Figure 34. Example graphic designed for communicating the meaning of a likelihood-ratio value.

$\mu_s = +\frac{\sigma^2}{2}$, then the output of the system is perfectly calibrated—this is a perfectly calibrated bi-Gaussian system.

Uncalibrated log-likelihood ratios (scores) are often calibrated using logistic regression (LogReg), but, unless the score distributions consist of two Gaussians with equal variance, the resulting "calibrated" $\ln(\Lambda)$ can be far from perfectly calibrated. We proposed a new calibration method, "bi-Gaussianized calibration," that maps scores toward perfectly calibrated bi-Gaussian $\ln(\Lambda)$ distributions. The particular perfectly calibrated bi-Gaussian system that the scores are mapped toward, is the perfectly calibrated bi-Gaussian system with either the same equal-error rate ($E_=$, EER) as the training data, or the same C_{llr} as the training data after they are calibrated using either LogReg, KDE, or PAVs. The method requires calculating the σ value for the perfectly calibrated bi-Gaussian system that corresponds to the $E_=$ or C_{llr} calculated for the training data. We found that the PAV method resulted in biased target σ values. The EER, LogReg, and KDE methods exhibited less bias and all resulted in similar target σ values.

We demonstrated the application of bi-Gaussianized calibration using two sets of simulated data and two sets of real data, including real data with extreme deviation from the assumption that same-source scores and different-source scores are distributed as two Gaussians with the same variance. The demonstrations showed that:

- Bi-Gaussianized calibration is robust to deviation from the assumption that the scores are distributed as two Gaussians with the same variance.
- Bi-Gaussianized calibration results in smoother score to $\ln(\Lambda)$ mapping functions than PAV calibration, and, for simulated data, bi-Gaussianized calibration results in mapping functions that are closer to "true" mapping functions than does PAV calibration.
- Bi-Gaussianized calibration results in $\ln(\Lambda)$ values that are closer to a perfectly calibrated bi-Gaussian system than is the case for $\ln(\Lambda)$ values output by LogReg calibration.

We introduced an innovation in drawing Tippett plots in which, in addition to the empirical results, we included the cumulative-density functions for the target perfectly calibrated bi-Gaussian system. This allows for comparison of the empirical results (e.g., bi-Gaussianized calibration results or LogReg calibration results) with the perfectly calibrated bi-Gaussian system with (approximately) the same C_{llr}.

We argued that if two calibration methods result in (approximately) the same C_{llr} value when applied to the same data (as is the case for bi-Gaussianized calibration and LogReg calibration),

the better system is the one whose $\ln(\Lambda)$ outputs are closer to the perfectly calibrated system with that C_{llr} value, that is, in our results, in terms of degree of calibration, bi-Gaussianized calibration was better than LogReg calibration.

Using simulated data to explore the effect of sampling variability on the performance of calibration methods, we found that:

- PAV calibration tended to result in higher C_{llr} values (worse performance) than other calibration methods.
- The EER, LogReg, and KDE variants of bi-Gaussianized calibration, and LogReg calibration, all resulted in similar median and quartile C_{llr} values.
- The KDE variant of bi-Gaussianized calibration had less extreme outlier C_{llr} values than the other variants of bi-Gaussianized calibration and than LogReg calibration.

In terms of performance as measured by C_{llr} and in terms of degree of calibration, the KDE variant of bi-Gaussianized calibration therefore appears to be the best of the variants and methods tested. It does, however, take many times longer to run than other variants of bi-Gaussianized calibration and than LogReg calibration. If time were an issue, the LogReg variant of bi-Gaussianized calibration might be a better choice.

As with all calibration methods, it is important to use training data for bi-Gaussianized calibration that are representative of the relevant population and reflective of the conditions for the case. This includes having sufficiently large training sets; otherwise, there is a danger of overfitting the training data and not generalizing well to validation data or to the actual questioned- and known-source items from the case. Depending on one's tolerance for such overfitting, the results of exploring the effect of sampling variability suggested that training data from 50 to 100 items may be acceptable.

We mentioned that, if one were concerned about calculating very large magnitude log-likelihood ratios on the basis of a limited amount of training data, in bi-Gaussianized calibration, one could induce shrinkage of log-likelihood ratios toward the neutral value of 0 by using a smaller σ for the target perfectly calibrated bi-Gaussian system than the calculated target σ value.

Finally, we proposed a graphical representation which may help in explaining the meaning of likelihood ratios to triers of fact. This displays the value of the likelihood ratio of interest on the probability-density plots of the target perfectly calibrated bi-Gaussian system. Whether this actually does assist with explaining the meaning of likelihood ratios is a question for future research.

Disclaimer

All opinions expressed in this article 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.

Declaration of competing interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Funding

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

References

Aitken, C.G.G., and Lucy, D. (2004) 'Evaluation of Trace Evidence in the Form of Multivariate Data', *Applied Statistics*, **53**: 109–22. http://dox.doi.org/10.1046/j.0035-9254.2003.05271.x [Corrigendum: (2004) 53, 665–6. http://dox.doi.org/10.1111/j.1467-9876.2004.02031.x]

Ayer, M. et al. (1955) 'An Empirical Distribution Function for Sampling with Incomplete Information', *The Annals of Mathematical Statistics*, **26**: 641–7. https://www.jstor.org/stable/2236377

Birdsall, T.G. (1973) The theory of signal detectability: ROC curves and their character. Technical Report No. 177. Cooley Electronics Laboratory, Department of Electrical and Computer Engineering, The University of Michigan, Ann Arbor, Michigan.

Brümmer, N., and du Preez, J. (2006) 'Application Independent Evaluation of Speaker Detection', *Computer Speech and Language*, **20**: 230–75. https://doi.org/10.1016/j.csl.2005.08.001

Brümmer, N., Swart, A., and van Leeuwen, D. (2014) A comparison of linear and non-linear calibrations for speaker recognition. *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop* (pp. 14–8). International Speech Communication Association. https://doi.org/10.21437/Odyssey.2014-3

Good I.J. (1985) 'Weight of Evidence: A Brief Survey'. In: Bernardo, J.M. et al. (eds), *Bayesian Statistics 2*, pp. 249–70. Amsterdam: Elsevier.

González-Rodríguez, J. et al. (2007) 'Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition', *IEEE Transactions on Speech and Audio Processing*, **15**: 2104–15. https://doi.org/10.1109/TASL.2007.902747

Meuwly, D., Ramos, D., and Haraksim, R. (2017) 'A Guideline for the Validation of Likelihood Ratio Methods Used for Forensic Evidence Evaluation', *Forensic Science International*, **276**: 142–53. https://doi.org/10.1016/j.forsciint.2016.03.048

Morrison, G.S. (2013) 'Tutorial on Logistic-Regression Calibration and Fusion: Converting a Score to a Likelihood Ratio', *Australian Journal of Forensic Sciences*, **45**: 173–97. https://doi.org/10.1080/00450618.2012.733025

Morrison, G.S. (2021) 'In the Context of Forensic Casework, Are There Meaningful Metrics of the Degree of Calibration?' *Forensic Science International: Synergy*, **3**: 100157. https://doi.org/10.1016/j.fsisyn.2021.100157

Morrison, G.S., and 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. https://doi.org/10.1016/j.scijus.2017.06.005

Morrison, G.S. et al. (2021) 'Consensus on Validation of Forensic Voice Comparison', *Science & Justice*, **61**: 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

Morrison, G.S. et al. (2020) 'Statistical Models in Forensic Voice Comparison', in Banks, D., Kafadar, K., Kaye, D.H., and Tackett, M. (eds) *Handbook of Forensic Statistics*, Ch. 20, pp. 451–97. Boca Raton, FL: CRC. https://doi.org/10.1201/9780367527709

Morrison, G.S., and Poh, N. (2018) 'Avoiding Overstating the Strength of Forensic Evidence: Shrunk Likelihood Ratios/Bayes Factors', *Science & Justice*, **58**: 200–18. http://dx.doi.org/10.1016/j.scijus.2017.12.005

Morrison, G.S. et al. (2023) 'Forensic Voice Comparison—Human-Supervised-Automatic Approach', in Houck, M. et al. (eds) *Encyclopedia of Forensic Sciences* (3rd edn.), vol. **2**, pp. 720–36. Elsevier. https://doi.org/10.1016/B978-0-12-823677-2.00182-3.

Neumann, C., and 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. https://doi.org/10.1093/lpr/mgaa006

Neumann, C., Hendricks, J., and Ausdemore, M. (2020) 'Statistical Support for Conclusions in Fingerprint Examinations', in Banks, D., Kafadar, K., Kaye, D.H., and Tackett, M. (eds) *Handbook of Forensic Statistics*, Ch. 14, pp. 277–324. Boca Raton, FL: CRC. https://doi.org/10.1201/9780367527709

Peterson, W.W., Birdsall, T.G., and Fox W.C. (1954) 'The Theory of Signal Detectability', *Transactions of the IRE Professional Group on Information Theory*, **4**: 171–211. https://doi.org/10.1109/TIT.1954.1057460

Ramos D., and González-Rodríguez, J. (2013) 'Reliable Support: Measuring Calibration of Likelihood Ratios', *Forensic Science International*, **230**: 156–69, https://doi.org/10.1016/j.forsciint.2013.04.014

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London, UK: Chapman & Hall/CRC.

van Es, A. et al. (2017) 'Implementation and Assessment of a Likelihood Ratio Approach for the Evaluation of LA-ICP-MS Evidence in Forensic Glass Analysis', *Science & Justice*, **57**: 181–92. https://doi.org/10.1016/j.scijus.2017.03.002

van Leeuwen, D.A., and Brümmer, N. (2013) The distribution of calibrated likelihood-ratios in speaker recognition. *Proceedings of Biometric Technologies in Forensic Science, BTFS 2013* (pp. 24–9). Nijmegen, The Netherlands: Radboud University. https://cls.ru.nl/staff/dvleeuwen/btfs-2013/proceedings-btfs2013.pdf

Vergeer, P. et al. (2016) 'Numerical Likelihood Ratios Outputted by LR Systems Are Often Based on Extrapolation: When to Stop Extrapolating', *Science & Justice*, **56**: 482–91. https://doi.org/10.1016/j.scijus.2016.06.003

Vergeer, P. et al. (2020) 'Why Calibrating LR-Systems Is Best Practice: A Reaction to "The Evaluation of Evidence for Microspectrophotometry Data Using Functional Data Analysis"', *Forensic Science International*, **314**: 110388. https://doi.org/10.1016/j.forsciint.2020.110388

Vergeer, P. (2023) 'From Specific-Source Feature-Based to Common-Source Score-Based Likelihood-Ratio Systems: Ranking the Stars', *Law, Probability and Risk*, **22**: mgad005. https://doi.org/10.1093/lpr/mgad005

Weber, P. et al. (2022) 'Validation of the Alpha Version of the E^3 Forensic Speech Science System (E^3FS^3) Core Software Tools', *Forensic Science International: Synergy*, **4**: 100223. https://doi.org/10.1016/j.fsisyn.2022.100223

Zadrozny, B., and Elkan, C. (2002) Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–99. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/775047.775151

Law, Probability and Risk, 2024, 23, 1–34
https://doi.org/10.1093/lpr/mgae004
Research Article