RESEARCH ARTICLE

Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data

Javier Franco-Pedroso*, Daniel Ramos, Joaquin Gonzalez-Rodriguez

ATVS-Biometric Recognition Group, Universidad Autonoma de Madrid, Madrid, Spain

* javier.franco@uam.es

OPEN ACCESS

Citation: Franco-Pedroso J, Ramos D, Gonzalez-Rodriguez J (2016) Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data. PLoS ONE 11 (2): e0149958. doi:10.1371/journal.pone.0149958

Editor: Gang Han, Texas A&M University, UNITED STATES

Received: November 25, 2015

Accepted: January 27, 2016

Published: February 22, 2016

Copyright: © 2016 Franco-Pedroso et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Glass-fragments dataset is from the article "Aitken, C. G. G., Lucy, D. Evaluation of trace evidence in the form of multivariate data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2004;53:109–122. doi: [10.1046/j.0035-9254.2003.05271.x](https://doi.org/10.1046/j.0035-9254.2003.05271.x)" and can be downloaded from: [http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1467-9876/homepage/glass-data.txt](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9876/homepage/glass-data.txt) Inks and car-paints datasets are from the book "Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, Colin Aitken. Statistical 515 Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical 516 Data. Wiley; January 2014."

Abstract

In forensic science, trace evidence found at a crime scene and on suspect has to be evaluated from the measurements performed on them, usually in the form of multivariate data (for example, several chemical compound or physical characteristics). In order to assess the strength of that evidence, the likelihood ratio framework is being increasingly adopted. Several methods have been derived in order to obtain likelihood ratios directly from univariate or multivariate data by modelling both the variation appearing between observations (or features) coming from the same source (within-source variation) and that appearing between observations coming from different sources (between-source variation). In the widely used multivariate kernel likelihood-ratio, the within-source distribution is assumed to be normally distributed and constant among different sources and the between-source variation is modelled through a kernel density function (KDF). In order to better fit the observed distribution of the between-source variation, this paper presents a different approach in which a Gaussian mixture model (GMM) is used instead of a KDF. As it will be shown, this approach provides better-calibrated likelihood ratios as measured by the log-likelihood ratio cost (C_{llr}) in experiments performed on freely available forensic datasets involving different trace evidences: inks, glass fragments and car paints.

Introduction

A likelihood ratio represents a ratio of likelihoods between two competing hypothesis. In the context of forensic science, these two hypotheses are that of the prosecution, H_p (for instance, the suspect originated the crime scene mark), and that of the defence, H_d (for instance, the suspect is not the origin of the crime scene mark). If some samples of a given material coming from a known source (*control* data) and some others coming from an unknown source (*recovered* data) are given, both known as *the evidence* (E), and some other information (I) related to the crime is available, the trier of fact (judge or jury) looks for the ratio between the probabilities

and can be downloaded from: <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470972106.html>.

Funding: JFP received funding from "Ministerio de Economía y Competitividad (ES)" (<http://www.mineco.gob.es/>) through the project "CMC-V2: Caracterización, Modelado y Compensación de Variabilidad en la Señal de Voz", with grant number TEC2012-37585-C02-01. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

of the H_p and H_d hypotheses given by

$$\frac{P(H_p|E, I)}{P(H_d|E, I)} \quad (1)$$

expressing the relative strength of one hypothesis versus the other.

However, the role of the forensic scientist must be restricted to evaluate the likelihood of the evidence assuming that any of the competing hypothesis is true, and it is not the evaluation of any other information different from that needed to evaluate the strength of the evidence. Using Bayesian theory, the above described ratio can be decomposed in the following way:

$$\frac{P(H_p|E, I)}{P(H_d|E, I)} = \frac{P(H_p|I)}{P(H_d|I)} \cdot \frac{P(E|H_p)}{P(E|H_d)} = \frac{P(H_p|I)}{P(H_d|I)} \cdot LR \quad (2)$$

making a clear separation of the role of the forensic scientist and that of the judge or jury. Thus, the likelihood ratio (LR) strengthens ($LR > 1$) or weakens ($LR < 1$) the probabilities of the propositions, in the light of the newly observed evidence. In the process of assigning/computing the LR , additional data, usually known in forensics as *background population*, is needed to obtain the likelihood of the parameters for the model used.

A possible statement of the hypotheses at the source level [1] is:

- H_p : the samples found at the crime scene and those obtained from the suspect come from a common source.
- H_d : the samples found at the crime scene and those obtained from the suspect come from different sources.

Other forms of the hypotheses are possible [1], but the analysis is outside the scope of this paper.

Likelihood ratios can be either directly derived from the data through the application of some probabilistic models (also known as feature-based LRs) or by transforming simple raw scores from a recognition system through a calibration step [2] (also known as score-based LRs). The score-based approach has been mainly used for biometric systems [3], in which the pattern recognition process does not follow a probabilistic model but a pattern matching procedure [4], the assumed conditions does not exactly hold (e.g. observations are not i.i.d. or do not follow a normal distribution), or the number of dimensions in the feature space makes the problem intractable (e.g. image vectors [5] or GMM-means supervectors [6]). However, recent approaches in face and speaker recognition modalities have begun to apply probabilistic methods with the aid of dimensionality reduction techniques [7–9]. On the other hand, the feature-based approach is usually followed in applied statistics to forensic science [10–12], where the observations are quite stable features whose within-source variation can be modelled by a normal distribution (for instance, measurements of the concentration of some chemical compounds).

A widely used approach within forensics [12–14] is that presented in [10], where the likelihood ratio is computed from multivariate data through the application of a two-level random effect model taking into account the variation *i*) between samples coming from the same source, known as *within-source* variation, and *ii*) between samples coming from different sources, known as *between-source* variation. Within-source variation is taken to be constant and normally distributed, and expressions for both normal and non-normal distribution for the between-source variation are given. When a normal distribution can not be assumed for the between-source variation, a kernel density function (KDF) [15] is used. However, as it will

be shown, this KDF approach overestimates the between-source density function in some areas of the feature space for datasets where sources are grouped in several clusters.

In order to avoid this problem, an alternative approach is presented in this work, in which the between-source distribution is represented by means of a Gaussian mixture model (GMM) [16, 17], whose parameters are obtained through a maximum-likelihood (ML) criterion, with the aim of obtaining a better representation of how the parameter being modelled (sources mean) varies across the different sources observed in the background population. As being also a probabilistic method for clustering data, GMMs provide a better representation of such kind of datasets, which leads to obtain better calibrated likelihood ratios.

The rest of the paper is organized as follows. In Section [Likelihood ratio computation], the likelihood ratio computation method is presented and the generative model defined. Section [Models for between-source distribution] describes the expressions to be used for a normally distributed between-source variation and those to be used when it is represented by means of a Gaussian mixture; for this latter case, the KDF expression used in [10] is also shown. In Section [GMMs for non-normal between-source distributions], the GMM training process is described, and the differences between using the KDF and the GMM approaches are highlighted. Section [Experimental framework] describes the forensic databases, the experimental protocols and the evaluation metrics, while the results are presented and discussed in Section [Results and Discussion]. Finally, conclusions are drawn in Section [Conclusions].

Likelihood ratio computation

In order to compute the likelihood ratio, the probability of the evidence has to be evaluated under the two competing hypothesis, H_p and H_d , where the evidence consists in both the control (y_1) and the recovered (y_2) datasets (see the mathematical notation given in the [Appendix]). If H_p is assumed true, the joint probability of both datasets has to be evaluated; on the other hand, if H_d is assumed true, each dataset is generated from a different source and hence they are independent.

$$LR = \frac{P(E|H_p)}{P(E|H_d)} = \frac{P(y_1, y_2|H_p)}{P(y_1, y_2|H_d)} = \frac{p(y_1, y_2)}{p(y_1) \cdot p(y_2)} \quad (3)$$

If a generative model with parameters Λ for the observed samples is assumed, the Bayesian solution is obtained by integrating out these parameters (if they vary from one source to another) for a given distribution which is usually obtained from a background population dataset, $p(\Lambda|\mathbf{X})$.

$$\frac{p(y_1, y_2)}{p(y_1) \cdot p(y_2)} = \frac{\int_{\Lambda} p(y_1, y_2|\Lambda) p(\Lambda|\mathbf{X}) d\Lambda}{\int_{\Lambda} p(y_1|\Lambda) p(y_2|\Lambda) p(\Lambda|\mathbf{X}) d\Lambda} \quad (4)$$

Final expressions for the numerator and denominator of the likelihood ratio will depend on the assumed generative model, which defines both the parameters Λ and the specific density functions. In this Section, we will describe the generative model used in [10], and the within-source distribution will be defined.

The generative model

The two-level random effect model [18] used in [10] can be seen as a generative model in which a particular observed feature vector $\mathbf{x}_{ij}$ coming from source i is generated through

$$\mathbf{x}_{ij} = \theta_i + \psi_j \quad (5)$$

where θ_i is a realization of the source random variable Θ and ψ_j is a realization of the additive random noise Ψ representing its within-source variation. This noisy term is taken to be constant among different sources and randomly distributed following

$$\Psi \sim \mathcal{N}(0, \mathbf{W}) \quad (6)$$

where $\mathbf{W}$ is the within-source covariance matrix. Thus, the conditional distribution of the random variable X_i (from which $\mathbf{x}_{ij}$ is drawn), given a particular source i , follows a normal distribution with mean θ_i and covariance matrix $\mathbf{W}$

$$X_{ij} | (\Theta = \theta_i) \sim \mathcal{N}(\theta_i, \mathbf{W}) \quad (7)$$

Within-source covariance matrix can be computed from a background population dataset, comprising $N = m \cdot n$ samples coming from m different sources, through

$$\mathbf{W} = \frac{\mathbf{S}_w}{N - m} \quad (8)$$

being $\mathbf{S}_w$ the within-source scatter matrix given by

$$\mathbf{S}_w = \sum_{i=1}^m \sum_{j=1}^n (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T \quad (9)$$

where $\bar{\mathbf{x}}_i$ is the average of a set of n feature vectors from source i .

As the assumed generative model has only one varying parameter, θ , characterizing the particular source, and the observed samples are assumed i.i.d. conditioned on the knowledge of θ , the numerator and the denominator of the likelihood ratio given in Eq 4 can be expressed, respectively, by

$$p(y_1, y_2) = \int_{\theta} p(y_1 | \theta, \mathbf{W}) p(y_2 | \theta, \mathbf{W}) p(\theta | \mathbf{X}) d\theta \quad (10)$$

where the parameter θ jointly varies for both control and recovered conditional probabilities, as they are assumed to come from the same source (say $\theta_1 = \theta_2 = \theta$), and

$$p(y_1) \cdot p(y_2) = \int_{\theta} p(y_1 | \theta, \mathbf{W}) p(\theta | \mathbf{X}) d\theta \times \int_{\theta} p(y_2 | \theta, \mathbf{W}) p(\theta | \mathbf{X}) d\theta \quad (11)$$

where these conditional probabilities can be integrated out independently as they are assumed to come from different sources (say $\theta_1 \neq \theta_2$).

Similarly to the random variable X_{ij} , the conditional distribution of a random variable $\bar{X}_i$ representing the average of a set of n feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ coming from a particular source i is given by

$$\bar{X}_i | (\Theta = \theta_i) \sim \mathcal{N}(\theta_i, \mathbf{D}), \quad \mathbf{D} = \frac{\mathbf{W}}{n} \quad (12)$$

Thus, when evaluating the conditional probability of a set of n_1 control samples, $\mathbf{y}_1$, or a set of n_2 recovered samples, $\mathbf{y}_2$, they will be evaluated in terms of their sample mean. That is,

$$p(y_l | \theta_l, \mathbf{W}) = p(\bar{\mathbf{y}}_l | \theta_l, \frac{\mathbf{W}}{n_l}) = N(\bar{\mathbf{y}}_l; \theta_l, \mathbf{D}_l), \quad l = 1, 2 \quad (13)$$

This leads to the following expressions for the previously shown integrals:

$$p(y_1, y_2) = \int_{\theta} N(\bar{y}_1; \theta, \mathbf{D}_1) N(\bar{y}_2; \theta, \mathbf{D}_2) p(\theta|\mathbf{X}) d\theta \quad (14)$$

and

$$p(y_l) = \int_{\theta} N(\bar{y}_l; \theta, \mathbf{D}_l) p(\theta|\mathbf{X}) d\theta, \quad l = 1, 2 \quad (15)$$

where only the distribution of the parameter θ remains undefined.

Models for between-source distribution

Regarding the distribution $p(\theta|\mathbf{X})$ from which the parameter characterizing the source θ is drawn, its shape depends on how the between-source variation is modelled. In this Section, two different types of distribution of such parameter, obtained from a background population, are shown. First, we will describe the expressions for a normally distributed between-source variation. While this is not the case under analysis in this work, it will serve to derive the expressions for the non-normal case, which is expressed in terms of a weighted sum of Gaussian densities.

Normal case

If sources means can be assumed normally distributed, $\Theta \sim \mathcal{N}(\mu, \mathbf{B})$, then

$$p(\theta|\mathbf{X}) = N(\theta; \mu, \mathbf{B}) \quad (16)$$

where μ and $\mathbf{B}$ are, respectively, the mean vector and the covariance matrix of the between-source distribution. These *hyperparameters* can be obtained from a background population (with m sources, n samples per source and N total samples) through

$$\mu = \frac{1}{m} \sum_{i=1}^m \bar{x}_i \quad (17)$$

and

$$\mathbf{B} = \frac{\mathbf{S}_b}{m-1} - \frac{\mathbf{S}_w}{n(N-m)} \quad (18)$$

where the between-source scatter matrix, $\mathbf{S}_b$, is given by

$$\mathbf{S}_b = \sum_{i=1}^m (\bar{x}_i - \mu)(\bar{x}_i - \mu)^T \quad (19)$$

Using this distribution for the parameter θ of the generative model, the integrals involved in the likelihood ratio computation can be written

$$p(y_1, y_2) = \int_{\theta} N(\bar{y}_1; \theta, \mathbf{D}_1) N(\bar{y}_2; \theta, \mathbf{D}_2) N(\theta; \mu, \mathbf{B}) d\theta \quad (20)$$

and

$$p(y_l) = \int_{\theta} N(\bar{y}_l; \theta, \mathbf{D}_l) N(\theta; \mu, \mathbf{B}) d\theta, \quad l = 1, 2 \quad (21)$$

Using the Gaussian identities given in the Appendix, the numerator of the likelihood ratio can be shown to be equal to:

$$p(y_1, y_2) = N(\bar{y}_1; \bar{y}_2, \mathbf{D}_1 + \mathbf{D}_2) \cdot N(\bar{y}^*; \mu, \mathbf{D}^* + \mathbf{B}) \quad (22)$$

where

$$\bar{y}^* = (\mathbf{D}_1 + \mathbf{D}_2)^{-1} (\mathbf{D}_2 \bar{y}_1 + \mathbf{D}_1 \bar{y}_2) \quad (23)$$

and

$$\mathbf{D}^* = \mathbf{D}_1 (\mathbf{D}_1 + \mathbf{D}_2)^{-1} \mathbf{D}_2 \quad (24)$$

Finally, each of the integrals in the denominator is given by

$$p(y_l) = N(\bar{y}_l; \mu, \mathbf{D}_l + \mathbf{B}), \quad l = 1, 2 \quad (25)$$

Non-normal case

When the normal assumption does not hold for the distribution of sources means among the background population data, the between-source distribution $p(\theta|\mathbf{X})$ can be approximated by a weighted sum of C Gaussian densities in the following form:

$$p(\theta|\mathbf{X}) = \sum_{c=1}^C \pi_c N(\theta; \mu_c, \Sigma_c) \quad (26)$$

where $\{\pi_k\}_{k=1, \dots, C}$ are the weighting factors and have the following constraints

$$0 \leq \pi_c \leq 1, \quad \sum_{c=1}^C \pi_c = 1 \quad (27)$$

With this distribution as the prior probability for the parameter θ of the generative model, the integrals involved in the likelihood ratio computation can be written

$$\begin{aligned} p(y_1, y_2) &= \int_{\theta} \{N(\bar{y}_1; \theta, \mathbf{D}_1) N(\bar{y}_2; \theta, \mathbf{D}_2) \sum_{c=1}^C \pi_c N(\theta; \mu_c, \Sigma_c)\} d\theta \\ &= \sum_{c=1}^C \pi_c \int_{\theta} \{N(\bar{y}_1; \theta, \mathbf{D}_1) N(\bar{y}_2; \theta, \mathbf{D}_2) N(\theta; \mu_c, \Sigma_c)\} d\theta \end{aligned} \quad (28)$$

and

$$\begin{aligned} p(y_l) &= \int_{\theta} \{N(\bar{y}_l; \theta, \mathbf{D}_l) \sum_{c=1}^C \pi_c N(\theta; \mu_c, \Sigma_c)\} d\theta \\ &= \sum_{c=1}^C \pi_c \int_{\theta} \{N(\bar{y}_l; \theta, \mathbf{D}_l) N(\theta; \mu_c, \Sigma_c)\} d\theta, \quad l = 1, 2 \end{aligned} \quad (29)$$

As it can be seen, the Gaussian mixture expressions become a weighted sum of the expressions given for the normal case, and so the probabilities involved in the likelihood ratio computation can be easily derived, resulting in

$$p(y_1, y_2) = N(\bar{y}_1; \bar{y}_2, \mathbf{D}_1 + \mathbf{D}_2) \cdot \sum_{c=1}^C \pi_c N(\bar{y}^*; \mu_c, \mathbf{D}^* + \Sigma_c) \quad (30)$$

and

$$p(y_l) = \sum_{c=1}^C \pi_c N(\bar{y}_l; \mu_c, \mathbf{D}_l + \Sigma_c), \quad l = 1, 2 \quad (31)$$

In [10], between-source distribution $p(\theta|\mathbf{X})$ is approximated through a KDF [15] where the kernel function $K(\cdot)$ is taken to be a multivariate normal function with smoothing parameter, or *bandwidth*, $\mathbf{H} = h^2 \mathbf{B}$:

$$p(\theta|\mathbf{X}) = \frac{1}{m|\mathbf{H}|^{1/2}} \sum_{i=1}^m K\left(\frac{\theta - \bar{x}_i}{\mathbf{H}^{1/2}}\right) = \frac{1}{m} \sum_{i=1}^m N(\theta; \bar{x}_i, h^2 \mathbf{B}) \quad (32)$$

where

$$h = \left(\frac{4}{2d+1}\right)^{\frac{1}{d+4}} m^{-1/(d+4)} \quad (33)$$

As it can be seen, this Gaussian KDF is in fact a Gaussian mixture whose parameters, equating terms in Eq 26, are given by

$$C = m, \quad \pi_c = \frac{1}{m}, \quad \mu_c = \bar{x}_i, \quad \Sigma_c = h^2 \mathbf{B} \quad (34)$$

Thus, the between-source variation is approximated by an equally weighted sum of multivariate Gaussian functions placed at every source mean present in the background population, $\bar{x}_i$, being their covariance matrices given by $h^2 \mathbf{B}$. That is, a weighted version of the between-source variation is *translated* to each source mean present in the background. As we will show later on, this will lead to overestimations of the between-source density in some areas of the feature space.

GMMs for non-normal between-source distributions

In this work, we propose to use a Gaussian Mixture Model (GMM) trained by means of a maximum-likelihood (ML) criterion in order to represent the distribution of the parameter θ characterizing the source. This model assumes that the observations are generated from a mixture of a finite number of Gaussian densities with unknown *hyperparameters*. Thus, it has been widely used to model the distribution of datasets in which the observations are grouped in several clusters, being each of them represented by a Gaussian density. In the case at hand, the observations are the means of the sources ($\bar{x}_i$) present in the background population dataset ($\mathbf{X}$), from which the distribution $p(\theta|\mathbf{X})$ is going to be modelled.

GMM training

Maximum likelihood (ML) is a method of determining the parameters Φ of a model that makes the observed samples the most probable given that model. Conversely to KDF, where the parameters ($\bar{x}_i, \mathbf{H}$) are first established and the density function $p(\theta|\mathbf{X})$ arises from them, in the GMM approach the density function is obtained by maximizing the likelihood of the observed data given the model, $p(\mathbf{X}|\Phi)$, from which the optimum parameters of the model are derived. In the case of a GMM of C components in the form of Eq 26, the ML parameters of the model, $\Phi = \{\pi_c, \mu_c, \Sigma_c\}_{c=1, \dots, C}$, are obtained [17] by maximizing the following

log-likelihood:

$$\ln p(\mathbf{X}|\Phi) = \sum_{i=1}^m \ln \left\{ \sum_{c=1}^C \pi_c N(\bar{\mathbf{x}}_i; \mu_c, \Sigma_c) \right\} \quad (35)$$

This can be done through the well known expectation-maximization (EM) algorithm [17, 19], which is an iterative method that successively updates the parameters Φ of the model until convergence. A recipe for this iterative process can be found in [17].

For a faster convergence of the algorithm, usually some steps of the *k-means* algorithm [17, 20] are previously iterated in order to obtain a good initialization of the GMM, as this clustering method provides the mean vectors $\{\mu_c\}_{c=1, \dots, C}$ (known as *centroids*) and the initial assignment of samples to clusters, from which $\{\pi_c\}_{c=1, \dots, C}$ and $\{\Sigma_c\}_{c=1, \dots, C}$ can be obtained.

The specific number of components, C , can be set by different methods. If the feature vectors are low-dimensional, the number of components can be visually estimated by inspecting a 2-D or 3-D projection of the background population data; however, depending on the structure of the data, there can be a lot of ambiguity in this process. Another option is to apply the *elbow method* [21] in the initial clustering stage, in which the cost function is plotted for different (increasing) number of clusters; for the first number of clusters there will be a great change when increasing the number of clusters, but at some point the marginal gain will drop indicating the proper number of clusters. A similar method can be applied by training GMMs for different numbers of components and evaluating the gain in terms of likelihood when increasing the number of them. Finally, similarly to the previous approach, if different GMMs for different number of components are trained, some model selection methods, like the Bayesian information criterion (BIC) [22] or the Akaike information criterion (AIC) [23], can be applied.

In this work, results are reported for several number of components in order to analyse how the evaluation metrics vary depending on this parameter, and the proper number of components related to the log-likelihood of the background data given the between-source density. For a given number of components, the *k-means* algorithm is iterated until convergence previously to the EM algorithm. In order to avoid local minima in *k-means* clustering, 100 random initializations are performed for a given number of components.

GMM versus KDF

For the purpose of illustrating the differences between KDF and GMM approaches, a synthetic 2-dimensional dataset has been generated (see Fig 1), in which 10 samples from 50 sources are drawn from normal distributions with the same covariance matrix (having then the same within-source variation). Sources means are drawn from 2 different normal distributions (25 sources each), each centred at a different separated point of the feature space, and one having a larger variance than the other in one of the dimensions. As a consequence, samples coming from different sources are grouped in two clearly separated clusters, one of them having a larger local intra-cluster between-source variation than the other. Also, the overall between-source variation is higher in one of the dimensions.

As already shown in Section [Models for between-source distribution], the density function $p(\theta|\mathbf{X})$ given by KDF approach is an equally weighted sum of Gaussian densities centred at each background source mean with covariance matrices $h^2 \mathbf{B}$ (see Eq 32). Thus, a weighted version of the overall between-source variation is *translated* to every source mean, reproducing this variation locally at each source mean. The resulting density function $p(\theta|\mathbf{X})$ for our synthetic dataset can be seen in Fig 2, where it is shown that the local intra-cluster between-source variation in dimension 1 is highly overestimated for both clusters, and slightly overestimated in dimension 2 for one of them due to the larger variation in the other one.

Fig 1. Synthetic dataset. Samples from a 2-dimensional synthetic dataset in which sources are grouped in two separate clusters.

doi:10.1371/journal.pone.0149958.g001

Conversely to KDF, in the GMM approach the Gaussian components are not forced to be centred at each source mean present in the background population, but a smaller number of components can be established allowing different sources means being generated from the same Gaussian component. Moreover, covariance matrices are neither fixed in advance, allowing to be locally learned for each component. As a consequence, the resulting density function can better fit the local between-source variation and the clustered nature of the dataset, as it is shown in Fig 3 for a 2-component GMM.

However, care must be taken in order to avoid *overfitting* when computing the density function through maximum likelihood. For a ML-trained GMM, the degree of fitting to the background data can be controlled through both the number of components C of the mixture and the number of EM iterations. In this work, for a given number of components, only two EM iterations are performed in order to avoid *overfitting*.

Accounting for within-source variation in the background population

When training a GMM from background sources means by maximizing the log-likelihood in Eq 35, it is assumed that there is no uncertainty in these mean values. However, the number of samples per source in the background population can be limited in forensic scenarios, and so these means cannot be reliably computed. In order to account for the uncertainty in these mean values, every observation belonging to those sources can be used to train a GMM by

Fig 2. KDF modelling of between-source variation in the synthetic dataset. (Above) Sources means and level contours of the between-source density function. (Below) 3-dimensional representation of the between-source density function.

doi:10.1371/journal.pone.0149958.g002

maximizing the following log-likelihood:

$$\ln p(\mathbf{X}|\Phi) = \sum_{i=1}^m \sum_{j=1}^n \ln \left\{ \sum_{c=1}^C \pi_c N(\mathbf{x}_{ij}; \mu_c, \Sigma_c) \right\} \quad (36)$$

While there can be not much difference in the values obtained for components means μ_c in a well balanced background dataset (same number of samples per source), taking into account the variation of the samples from each source around its mean value through Eq 36 provides a more conservative background density, as every background sample is considered as a possible mean value of a source. Furthermore, this also helps to avoid Gaussian collapsing when a reduced number of sources are assigned to a particular component. The effect on our synthetic dataset is shown in Fig 4, where the Gaussian densities are placed at the same locations as in Fig 3 but larger variances and covariances are obtained, specially for the cluster with lower intra-cluster between-source variation.

Experimental framework

Forensic datasets

In order to test the approach proposed in this work, several types of forensic datasets have been used, being one of them the glass-fragments dataset also used in [10], which can be downloaded

Fig 3. GMM modelling of between-source variation in the synthetic dataset. (Above) Sources means and level contours of the between-source density function. (Below) 3-dimensional representation of the between-source density function.

doi:10.1371/journal.pone.0149958.g003

from [http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1467-9876/homepage/glass-data.txt](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9876/homepage/glass-data.txt). A detailed description of the other two datasets can be found in [12], and can be downloaded from <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470972106.html>.

- *Inks*. For this dataset, the features are the measurements of the $d = 3$ chromaticity coordinates r , g and b (being $r + g + b = 1$) taken on samples of blue inks. The dataset comprises the measurements on $n = 10$ samples for each of the $m = 40$ different ink sources.
- *Glass fragments*. For this dataset, the features are the measurements of the concentrations in $d = 3$ elemental ratios taken on glass fragments: $\log(\text{Ca/K})$, $\log(\text{Ca/Si})$ and $\log(\text{Ca/Fe})$. The dataset comprises the measurements on $n = 5$ fragments for each of the $m = 62$ different glass sources.
- *Car paints*. For this dataset, the features are the measurements of $d = 7$ organic components present in the top layer of different acrylic car paintings. The dataset comprises the measurements on $n = 3$ samples for each of the $m = 36$ different car-paint sources.

Table 1 gathers the already mentioned characteristics of these three datasets, while Figs 5, 6 and 7 show 2-dimensional projections of their sources means. As it can be seen, sources means in the last two datasets (glass fragments and car paints) present a clustered nature, while those in the first one (inks) are normally distributed [12].

Fig 4. GMM modelling of between-source variation in the synthetic dataset when taking into account the background within-source variation. (Above) Sources means and level contours of the between-source density function. (Below) 3-dimensional representation of the between-source density function.

doi:10.1371/journal.pone.0149958.g004

Protocols

The protocol followed in [10] used the whole glass-fragment dataset in order to obtain the between-source probability density function $p(\theta|\mathbf{X})$. Then, for each source, the first 3 samples (out of 5) were used as control data and the last 3 were used as recovered data, having so both datasets one sample in common. While this *non-partitioning* protocol alleviates the lack of data due to the small size of the dataset, it may lead to overoptimistic results as the different subsets (background, control and recovered) are overlapped.

In this work, a *cross-validation* protocol is also used in order to avoid overoptimistic results, in which the dataset is divided into two non-overlapping subsets devoted to:

- obtain the between-source distribution $p(\theta|\mathbf{X})$ (*known data* or *training subset*), and

Table 1. Datasets summary.

	m	n	d
Inks	40	10	3
Glass fragments	62	5	3
Car paints	36	3	7

m, number of sources; n, number of samples per source; d, number of features.

doi:10.1371/journal.pone.0149958.t001

Fig 5. Sources means in the inks dataset. The three 2-dimensional projections of the sources means.

doi:10.1371/journal.pone.0149958.g005

- compute same-source and different-source likelihood ratios (*unknown data* or *testing subset*). This subset is further divided into two non-overlapping halves acting as control and recovered data.

In order to alleviate the lack of data, this procedure is carried out in the following way. For each of the $m(m-1)/2$ possible pairs of sources in the dataset, all the samples belonging to those two sources are taken apart from the dataset in order to be used as the *testing subset*, being the remaining sources used as the *training subset*. Each of the two sources in the *testing subset* is divided into two non-overlapping halves ($\{1a, 1b\}$ and $\{2a, 2b\}$) that can be used either as control or recovered data to perform 2 same-source comparisons (1a-1b, 2a-2b) and 4 different-source comparisons (1a-2a, 1a-2b, 1b-2a, 1b-2b). Although the same control and recovered data from a particular source is used in all the different pairs in which it is involved, as the remaining sources change for each different pair, different between-source distributions $p(\theta|X)$ are involved in likelihood ratio computations. This procedure allow us to perform a total number of $m(m-1)$ same-source comparisons and $2 \times m(m-1)$ different-source comparisons for a given dataset, instead of the m same-source comparisons and $m(m-1)/2$ different-source comparisons performed in [10], while the between-source distribution $p(\theta|X)$ used in every

Fig 6. Sources means in the glass-fragments dataset. The three 2-dimensional projections of the sources means.

doi:10.1371/journal.pone.0149958.g006

Fig 7. Sources means in the car-paints dataset. Three 2-dimensional projections of the sources means.

doi:10.1371/journal.pone.0149958.g007

comparison is obtained from $m - 2$ different sources instead of m . The specific number of comparisons for each evaluation protocol on the different datasets are given in Table 2.

Evaluation Metrics

The main evaluation metric used in order to compare the different approaches is the log-likelihood ratio cost function (C_{llr}) [2, 24], which evaluates both the *discrimination* abilities of the computed log-likelihood ratios and the goodness of their *calibration*. Given a set of log-likelihood ratios $\{\mathcal{L}\} = \{\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_C\}$ obtained from C comparisons, the C_{llr} can be computed in the following way:

$$C_{llr}(\{\mathcal{L}\}) = \frac{1}{2 \log 2} \left(\frac{1}{N_{ss}} \sum_{c \in ss} \log(1 + e^{-\mathcal{L}_c}) + \frac{1}{N_{ds}} \sum_{c \in ds} \log(1 + e^{\mathcal{L}_c}) \right) \quad (37)$$

where 'ss' is the set of N_{ss} same-source comparisons and 'ds' is the set of N_{ds} different-source comparisons. As it is a cost function, the larger the C_{llr} value, the worse the verification method, being $C_{llr} = 0$ the minimum achievable cost. Note also that this metric allows to define a *neutral reference* which does not provide support for any of the two hypothesis (that is, $\mathcal{L}_c = 0$ for every comparison), providing a reference value of $C_{llr} = 1$. Thus, a verification method for which C_{llr} is larger than 1 means that it is providing misleading likelihood ratios.

An important aspect of the C_{llr} is that it can be decomposed into two additive terms, one due to the discrimination abilities (C_{llr}^{min}) and another one due to the calibration of the verification method (C_{llr}^{cal}) where

$$C_{llr}^{cal} = C_{llr} - C_{llr}^{min} \quad (38)$$

Table 2. Number of same-source and different-source comparisons in each dataset for the non-partitioning and cross-validation protocols.

	Non-partitioning		Cross-validation	
	Same-source	Different-source	Same-source	Different-source
Glass fragments	62	1891	3782	7564
Inks	40	780	1560	3120
Car paints	36	630	1260	2520

doi:10.1371/journal.pone.0149958.t002

and C_{lr}^{min} is obtained by means of the *Pool Adjacent Violators* (PAV) algorithm [25, 26] and represents the minimum achievable C_{lr} in the case of having an optimally calibrated log-likelihood ratios set $\{\mathcal{L}'\}$ (details can be found in [24]).

In order to show the performance over a wide range of prior probabilities, the Empirical Croos-Entropy (ECE) plots [27, 28] will be used. These figures (see, for example, Fig 8) graphically represent what would be the accuracy (solid curve) when using the set of logLR values $\{\mathcal{L}\}$ for each of the prior probabilities (represented as logarithmic odds) in the given range. Additionally, the discriminating power is also plotted (dashed curve) for the optimally calibrated (ideal) logLRs set $\{\mathcal{L}'\}$, along with the neutral reference (dotted curve).

Fig 8. ECE plots for the KDF and GMM approaches on the inks dataset when applying the cross-validation protocol. GMM is trained by maximizing Eq 35.

doi:10.1371/journal.pone.0149958.g008

Results and Discussion

Inks dataset

For this dataset, as the background sources means are normally distributed, GMMs with a single component has been trained by maximizing either Eq 35 or Eq 36. Table 3 shows the detailed results (C_{llr} , C_{llr}^{min} and C_{llr}^{cal}) for KDF and GMM approaches (Eq 35 and Eq 36) when applying both the *non-partitioning* and the *cross-validation* protocols.

First, it should be noted that results in the *non-partitioning* protocol are slightly better for every method as it is an overoptimistic framework where data is shared between training and testing subsets. Regarding the comparison between methods, it can be seen that no significant improvement is obtained by the GMM approach as the sources means for this dataset do not present a clustered nature. Moreover, among the two GMM variants, the results obtained when maximizing Eq 35 are slightly better, presumably due to the fact that enough number of samples per source are available ($n = 10$), compared to the number of features ($d = 3$), to compute reliable sources means, and further uncertainty accounted for Eq 36 seems to be counter-productive.

Finally, Fig 8 show ECE plots for KDF and GMM (Eq 35) approaches when applying the *cross-validation* protocol, where it can be seen that both present similar performance for a wide range of prior probabilities.

Glass-fragments dataset

For this dataset, several GMMs have been trained, by maximizing Eq 35, in order to analyse how the main evaluation metric (C_{llr}) varies as a function of the number of components, C . In the experiments carried out, the maximum number of components has been limited to 6 in order to avoid Gaussian collapsing due to a reduced number of observations (sources means) per component (62 total sources in the whole dataset). Results for the *non-partitioning* protocol can be seen in Fig 9 for both KDF and GMM (Eq 35) approaches, where also the log-likelihood of the background data (sources means) given the between-source density has been plotted.

As it was expected for this *non-partitioning* protocol, C_{llr} decreases as the number of components increases, due to the shared data between training and testing subsets, which can lead to overfit the background density. However, as soon as the log-likelihood for the GMM surpass that obtained for the KDF density, better results are obtained with the GMM approach. It is also worth noting that this happens for a number of components (2–3) around that which could be expected from visual inspection of the 2-dimensional projections shown in Fig 6.

Fig 10 show the same analysis for the *cross-validation* protocol. In this case, the log-likelihood is not plotted as the GMM change for every testing sources-pair (being trained on the remaining sources). Similar conclusions than before can be drawn, but here the overfitting problem affecting the *non-partitioning* protocol is revealed, as the C_{llr} for the *cross-validation*

Table 3. Performance of KDF and GMM approaches on the inks dataset for the *non-partitioning* and *cross-validation* protocols.

	Non-partitioning			Cross-validation		
	C_{llr}^{min}	C_{llr}^{cal}	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}	C_{llr}
KDF	0.1459	0.0224	0.1684	0.1558	0.0214	0.1778
GMM (Eq 35)	0.1430	0.0223	0.1653	0.1533	0.0223	0.1756
GMM (Eq 36)	0.1453	0.0271	0.1724	0.1569	0.0286	0.1855

$$C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$$

doi:10.1371/journal.pone.0149958.t003

Fig 9. Analysis of the number of GMM components for the glass-fragments dataset when applying the *non-partitioning* protocol. GMM is trained by maximizing Eq 35. (Above) Log-likelihood ratio cost. (Below) Log-likelihood of the background data given the between-source density function.

doi:10.1371/journal.pone.0149958.g009

protocol reaches a minimum value for a given number of components ($C = 4$) and then increases. Results are also shown for GMMs trained by maximizing Eq 36, with similar conclusions but slightly better results, presumably due to the small number of samples per source ($n = 5$) compared to the number of features ($d = 3$).

Table 4 shows the detailed results (C_{lr} , C_{lr}^{min} and C_{lr}^{cal}) for KDF and GMM approaches (Eq 35 and Eq 36) when applying both the *non-partitioning* and the *cross-validation* protocols. For GMM approaches, results are given for the optimum number of components ($C = 4$) when the *cross-validation* protocol is applied. Again, as the *non-partitioning* protocol constitutes an over-optimistic framework, results are slightly better for every method compared to the *cross-validation* protocol. This is also the reason of obtaining better results when GMMs are trained by maximizing Eq 36, as the same sources are present in both training and testing subsets. However, when the *cross-validation* protocol is applied, there is no shared data between those subsets, and so the additional uncertainty accounted by Eq 36 provides slightly better results. In any case, both GMM approaches outperform the KDF one due to their better calibration properties for this clustered dataset.

Finally, Fig 11 shows the comparative results between KDF and GMM (Eq 36) in the form of ECE plots when the *cross-validation* protocol is applied.

Car-paints dataset

An equivalent analysis to that shown for the glass-fragments dataset has been performed for the car-paints one. Fig 12 shows both the C_{lr} and the log-likelihood of the background data

Fig 10. Analysis of the number of GMM components for the glass-fragments dataset when applying the cross-validation protocol.

doi:10.1371/journal.pone.0149958.g010

given the model (trained by maximizing Eq 35) as a function of the number of components for the *non-partitioning* protocol. Similarly to what happened with the previous dataset, C_{lr} decreases as the number of components increases, and as soon as the log-likelihood for the GMM surpasses that obtained for the KDF density, better results are obtained with the GMM approach. Again, this happens for a number of components (3–4) around that which could be expected from visual inspection of some of the 2-dimensional projections shown in Fig 7.

Fig 13 show the same analysis for the *cross-validation* protocol (without showing the log-likelihood plot), where it can be seen (solid line) that, similarly to what happened with the glass-fragments dataset, a minimum C_{lr} value is reached for a particular number of components ($C = 3$) and then it increases. However, when plotting results for GMMs trained by maximizing Eq 36 instead (dot-dashed line), the number of components for which the minimum

Table 4. Performance of KDF and GMM approaches on the glass-fragments dataset for the *non-partitioning* and *cross-validation* protocols.

	Non-partitioning			Cross-validation		
	C_{lr}^{min}	C_{lr}^{cal}	C_{lr}	C_{lr}^{min}	C_{lr}^{cal}	C_{lr}
KDF	0.0787	0.0394	0.1182	0.0850	0.0410	0.1260
GMM (Eq 35)	0.0785	0.0223	0.1008	0.0863	0.0291	0.1154
GMM (Eq 36)	0.0785	0.0229	0.1013	0.0862	0.0282	0.1144

$$C_{lr} = C_{lr}^{min} + C_{lr}^{cal}$$

doi:10.1371/journal.pone.0149958.t004

Fig 11. ECE plots for the KDF and GMM approaches on the glass-fragments dataset when applying the *cross-validation* protocol. GMM is trained by maximizing [Eq 36](#).

doi:10.1371/journal.pone.0149958.g011

C_{llr} value is reached is slightly larger ($C = 5$); this also happens for the *non-partition* protocol, as the log-likelihood of the training data (observations) given the model for the GMM do not surpass that of the KDF until a larger number of components ($C = 4$) is reached.

[Table 5](#) shows the detailed results (C_{llr} , C_{llr}^{min} and C_{llr}^{cal}) for KDF and GMM approaches ([Eq 35](#) and [Eq 36](#)) when applying both the *non-partitioning* and the *cross-validation* protocols. For GMM approaches, results are given for the optimum number of components ($C = 4$ for [Eq 35](#), $C = 5$ for [Eq 36](#)) when the *cross-validation* protocol is applied. Similar conclusions to those obtained for the glass-fragments dataset can be drawn, but much better results are obtained by GMMs approaches presumably due to the distance among clusters, which lead to KDF densities which overestimate the between-source distribution in some areas of the feature space (as shown in [Fig 2](#) for the synthetic dataset). Among GMM approaches, the maximization of [Eq 36](#) leads to much better results for the *cross-validation* protocol due to the small number of samples per source ($n = 3$) compared to the number of features ($d = 7$), which lead to unreliably computed sources means when training GMMs by maximizing [Eq 35](#).

Fig 12. Analysis of the number of GMM components for the car-paints dataset when applying the *non-partitioning* protocol. GMM is trained by maximizing Eq 35. (Above) Log-likelihood ratio cost. (Below) Log-likelihood of the background data given the between-source density function.

doi:10.1371/journal.pone.0149958.g012

Finally, Fig 14 shows the comparative results between KDF and GMM (Eq 36) in the form of ECE plots when the *cross-validation* protocol is applied.

Conclusions

In this work, we present a new approach for computing likelihood ratios from multivariate data in which the between-source distribution is obtained through ML training of the parameters of a GMM. Using the same generative model as in [10], a common derivation of the LR expressions is presented for both Gaussian KDF and GMM, in which the between-source distribution is represented in terms of a weighted sum of Gaussian densities. Then, differences between KDF and GMM approaches are highlighted, and the effects on the obtained probability density are shown for a synthetic dataset. Furthermore, a variant in GMM training has been tested in order to account for the uncertainty in sources means when few samples per source are available in the background data.

The proposed approach has been tested on three different forensic datasets and compared with the KDF approach. Additionally to the *non-partitioning* protocol applied in [10], a more realistic *cross-validation* protocol is applied in order to avoid overoptimistic results, as ML-trained GMMs can overfit the background population density. Performance is evaluated in terms of the log-likelihood ratio cost function (C_{lr}), which allows to decompose the performance in a term due to the discrimination abilities and another one due to the calibration

Fig 13. Analysis of the number of GMM components for the car-paints dataset when applying the cross-validation protocol.

doi:10.1371/journal.pone.0149958.g013

properties. ECE plots have been used to show the behaviour in a wide range of prior probabilities, which is needed in forensic science.

Results show that, although KDF and GMM approaches present similar discrimination abilities, when the datasets have a *clustered* nature, the between-source distribution is better described by a GMM, leading to better calibrated likelihood ratios. If clusters are not easily distinguishable, the between-source distribution still can be modelled by one single component, obtaining similar results to the KDF approach. Specially remarkable are the results obtained for the car-paints dataset, where $\sim 50\%$ improvement in terms of calibration performance is obtained.

Table 5. Performance of KDF and GMM approaches on the car-paints dataset for the non-partitioning and cross-validation protocols.

	Non-partitioning			Cross-validation		
	C_{lr}^{min}	C_{lr}^{cal}	C_{lr}	C_{lr}^{min}	C_{lr}^{cal}	C_{lr}
KDF	0.0819	0.1388	0.2208	0.0972	0.1786	0.2759
GMM (Eq 35)	0.0715	0.0671	0.1386	0.0968	0.1769	0.2737
GMM (Eq 36)	0.0715	0.0729	0.1443	0.0899	0.0934	0.1833

$$C_{lr} = C_{lr}^{min} + C_{lr}^{cal}$$

doi:10.1371/journal.pone.0149958.t005

Fig 14. ECE plots for the KDF and GMM approaches on the car-paints dataset when applying the *cross-validation* protocol. GMM is trained by maximizing [Eq 36](#).

doi:10.1371/journal.pone.0149958.g014

Appendix

Mathematical notation

Throughout this work we consider multivariate data in the form of d -dimensional column vectors $\mathbf{x} = (x_1, x_2, \dots, x_d)^T$. Following the same notation as in [10], a set of n elements of such data belonging to the same particular source i are denoted by $\mathbf{x}_i = \{\mathbf{x}_{ij}\}_{j=1, \dots, n} = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{in}\}$, while their sample mean is denoted by $\bar{\mathbf{x}}_i$. Similarly, $\mathbf{x}_i$ is used to denote background data while $\mathbf{y}_i$ is used to denote either control ($\mathbf{y}_1$) or recovered data ($\mathbf{y}_2$). The set of feature vectors coming from different sources present in the background data is denoted by $\mathbf{X}$.

In general, column vectors are denoted by bold lower-case letters and matrices by bold upper-case letters, while scalar quantities are denoted by lower-case italic letters. Random variables are denoted by upper-case non-italic letters. $P(\cdot)$ is used to indicate the probability of a certain event, while $p(\cdot)$ denotes a probability density function. We denote a d -dimensional Gaussian distribution with mean μ and covariance matrix Σ by $\mathcal{N}(\mu, \Sigma)$ and the corresponding probability density function by $N(\mathbf{x}; \mu, \Sigma)$ ($\mathbf{x} \in \mathbb{R}^d$).

Multivariate Gaussian function

$$N(\mathbf{x}; \mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\} = N(\mu; \mathbf{x}, \Sigma) \quad (39)$$

Gaussian identities

Product of two multivariate Gaussian functions.

$$N(\mathbf{x}; \mathbf{a}, \mathbf{A}) \cdot N(\mathbf{x}; \mathbf{b}, \mathbf{B}) = N(\mathbf{a}; \mathbf{b}, \mathbf{a} + \mathbf{B}) \cdot N(\mathbf{x}; \mathbf{c}, \mathbf{C}) \quad (40)$$

$$\mathbf{c} = \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1} \mathbf{a} + \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1} \mathbf{b} \quad (41)$$

$$\mathbf{C} = \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1} \mathbf{B} \quad (42)$$

Convolution of two multivariate Gaussian functions.

$$\int_{\mathbf{x}} N(\mathbf{x}; \mathbf{a}, \mathbf{A}) N(\mathbf{y} - \mathbf{x}; \mathbf{b}, \mathbf{B}) d\mathbf{x} = N(\mathbf{y}; \mathbf{a} + \mathbf{b}, \mathbf{A} + \mathbf{B}) \quad (43)$$

Expressions for a normal between-source distribution

Derivation of the numerator. First, we solve the product of the two Gaussian functions depending on either the control or the recovered data means, obtaining the following expression

$$\begin{aligned} p(y_1, y_2) &= \int_{\theta} \{N(\bar{y}_1; \theta, \mathbf{D}_1) N(\bar{y}_2; \theta, \mathbf{D}_2) N(\theta; \mu, \mathbf{B})\} d\theta \\ &= \int_{\theta} \{N(\theta; \bar{y}_1, \mathbf{D}_1) N(\theta; \bar{y}_2, \mathbf{D}_2) N(\theta; \mu, \mathbf{B})\} d\theta \\ &= \int_{\theta} \{N(\theta; \mathbf{z}, \mathbf{Z}) N(\bar{y}_1; \bar{y}_2, \mathbf{D}_1 + \mathbf{D}_2) N(\theta; \mu, \mathbf{B})\} d\theta \end{aligned} \quad (44)$$

where

$$\mathbf{z} = (\mathbf{D}_1 + \mathbf{D}_2)^{-1} (\mathbf{D}_2 \bar{y}_1 + \mathbf{D}_1 \bar{y}_2) \quad (45)$$

and

$$\mathbf{Z} = \mathbf{D}_1(\mathbf{D}_1 + \mathbf{D}_2)^{-1} \mathbf{D}_2 \quad (46)$$

Being $N(\bar{y}_1; \bar{y}_2, \mathbf{D}_1 + \mathbf{D}_2)$ independent of θ , we can solve the remaining integral as a convolution of two Gaussian functions:

$$\begin{aligned} p(y_1, y_2) &= N(\bar{y}_1; \bar{y}_2, \mathbf{D}_1 + \mathbf{D}_2) \int_{\theta} \{N(\mathbf{z}; \theta, \mathbf{Z}) N(\theta; \mu, \mathbf{B})\} d\theta \\ &= N(\bar{y}_1; \bar{y}_2, \mathbf{D}_1 + \mathbf{D}_2) \int_{\theta} \{N(\mathbf{z} - \theta; 0, \mathbf{Z}) N(\theta; \mu, \mathbf{B})\} d\theta \\ &= N(\bar{y}_1; \bar{y}_2, \mathbf{D}_1 + \mathbf{D}_2) \cdot N(\mathbf{z}; \mu, \mathbf{Z} + \mathbf{B}) \end{aligned} \quad (47)$$

Finally, replacing $\mathbf{D}_l = \mathbf{W}/n_l$, $l = 1, 2$, in $\mathbf{z}$ and $\mathbf{Z}$

$$\mathbf{z} = \left(\frac{\mathbf{W}}{n_1} + \frac{\mathbf{W}}{n_2} \right)^{-1} \left(\frac{\mathbf{W}}{n_2} \bar{y}_1 + \frac{\mathbf{W}}{n_1} \bar{y}_2 \right) = \frac{\frac{1}{n_2} \bar{y}_1 + \frac{1}{n_1} \bar{y}_2}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_2 + n_1} = \bar{y}^* \quad (48)$$

$$\mathbf{Z} = \frac{\mathbf{W}}{n_1} \left(\frac{\mathbf{W}}{n_1} + \frac{\mathbf{W}}{n_2} \right)^{-1} \frac{\mathbf{W}}{n_2} = \frac{\frac{\mathbf{W}}{n_1 n_2}}{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{\mathbf{W}}{n_2 + n_1} \quad (49)$$

we obtain

$$p(y_1, y_2) = N(\bar{y}_1; \bar{y}_2, \frac{\mathbf{W}}{n_1} + \frac{\mathbf{W}}{n_2}) \cdot N(\bar{y}^*; \mu, \frac{\mathbf{W}}{n_1 + n_2} + \mathbf{B}) \quad (50)$$

Derivation of the denominator. Each of the integrals in the denominator of the LR can be solved by the convolution of two Gaussian functions

$$\begin{aligned} p(y_l) &= \int_{\theta} \{N(\bar{y}_l; \theta, \mathbf{D}_l) N(\theta; \mu, \mathbf{B})\} d\theta \\ &= \int_{\theta} \{N(\bar{y}_l - \theta; 0, \mathbf{D}_l) N(\theta; \mu, \mathbf{B})\} d\theta \\ &= N(\bar{y}_l; \mu, \mathbf{D}_l + \mathbf{B}) = N(\bar{y}_l; \mu, \frac{\mathbf{W}}{n_l} + \mathbf{B}) \end{aligned} \quad (51)$$

giving the following final expression for the denominator of the LR under the between-source normal assumption:

$$p(y_1) \cdot p(y_2) = N(\bar{y}_1; \mu, \frac{\mathbf{W}}{n_1} + \mathbf{B}) \cdot N(\bar{y}_2; \mu, \frac{\mathbf{W}}{n_2} + \mathbf{B}) \quad (52)$$

Author Contributions

Conceived and designed the experiments: JFP DR JGR. Performed the experiments: JFP. Analyzed the data: JFP DR. Contributed reagents/materials/analysis tools: JFP DR. Wrote the paper: JFP DR JGR.

References

1. Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A hierarchy of propositions: deciding which level to address in casework. *Science and Justice*. 1998; 38(4):231–239. doi: [10.1016/S1355-0306\(98\)72117-3](https://doi.org/10.1016/S1355-0306(98)72117-3)
2. van Leeuwen DA, Brümmer N. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. *Speaker Classification I: Lecture Notes in Computer Science*. 2007; 4343:330–353. doi: [10.1007/978-3-540-74200-5_19](https://doi.org/10.1007/978-3-540-74200-5_19)
3. Gonzalez-Rodriguez J, Rose P, Ramos D, Toledano DT, Ortega-Garcia J. Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*. 2007; 15(7):2104–2115. doi: [10.1109/TASL.2007.902747](https://doi.org/10.1109/TASL.2007.902747)
4. Jain AK, Ross A, Pankanti S. Biometrics: a tool for information security. *IEEE Transactions on Information Forensics and Security*. 2006; 1(2):125–143. doi: [10.1109/TIFS.2006.873653](https://doi.org/10.1109/TIFS.2006.873653)
5. Turk MA, Pentland AP. Face recognition using eigenfaces. *Proceedings of the 1991 CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1991;586–591.

6. Campbell WM, Sturim DE, Reynolds DA. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*. 2006; 13(5):308–311. doi: [10.1109/LSP.2006.870086](https://doi.org/10.1109/LSP.2006.870086)
7. Li P, Fu Y, Mohammed U, Elder JH, Prince SJD. Probabilistic Models for Inference about Identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. January 2012; 34(1):144–157.
8. Sizov A, Lee KA, Kinnunen T. Unifying Probabilistic Linear Discriminant Analysis Variants in Biometric Authentication. Structural, Syntactic, and Statistical Pattern Recognition. *Lecture Notes in Computer Science*. 2014; 8621:464–475.
9. Borgstrom BJ, McCree A. Supervector Bayesian speaker comparison. *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2013;7693–7697.
10. Aitken CGG, Lucy D. Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2004; 53:109–122. doi: [10.1046/j.0035-9254.2003.05271.x](https://doi.org/10.1046/j.0035-9254.2003.05271.x)
11. Aitken CGG, Taroni F. *Statistics and the Evaluation of Evidence for Forensic Scientists*, 2nd Edition. Wiley; July 2004.
12. Zadora G, Martyna A, Ramos D, Aitken C. *Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data*. Wiley; January 2014.
13. Rose P. Forensic Voice Comparison with Monophthongal Formant Trajectories—a likelihood ratio-based discrimination of “Schwa” vowel acoustics in a close social group of young Australian females. *Proceedings of the 40th ICASSP International Conference on Acoustics, Speech and Signal Processing*. 2015;4819–4823.
14. Bolck A, Weyermann C, Dujourdy L, Esseiva P, van den Berg J. Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International*. 2009; 191(1–3):42–51. doi: [10.1016/j.forsciint.2009.06.006](https://doi.org/10.1016/j.forsciint.2009.06.006) PMID: [19608360](https://pubmed.ncbi.nlm.nih.gov/19608360/)
15. Epanechnikov VA. Non-Parametric Estimation of a Multivariate Probability Density. *Theory Probab. Appl.*, 14(1):153–158. doi: [10.1137/1114019](https://doi.org/10.1137/1114019)
16. McLachlan GJ, Basford KE. *Mixture models: Inference and applications to clustering*. Applied Statistics. 1988.
17. Bishop CM. *Pattern Recognition and Machine Learning*. Springer; 2006.
18. Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. *Biometrics*. 1982; 38(4):963–974. doi: [10.2307/2529876](https://doi.org/10.2307/2529876) PMID: [7168798](https://pubmed.ncbi.nlm.nih.gov/7168798/)
19. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*. 1977; 39(1):1–38.
20. MacQueen JB. *Some Methods for classification and Analysis of Multivariate Observations*. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1*. University of California Press. 1967;281–297.
21. Ketchen DJ, Shook CL. The application of cluster analysis in Strategic Management Research: An analysis and critique. *Strategic Management Journal*. 1996; 17(6):441–458. doi: [10.1002/\(SICI\)1097-0266\(199606\)17:6%3C441::AID-SMJ819%3E3.0.CO;2-G](https://doi.org/10.1002/(SICI)1097-0266(199606)17:6%3C441::AID-SMJ819%3E3.0.CO;2-G)
22. Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6(2):461–464. doi: [10.1214/aos/1176344136](https://doi.org/10.1214/aos/1176344136)
23. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974; 19(6):716–723. doi: [10.1109/TAC.1974.1100705](https://doi.org/10.1109/TAC.1974.1100705)
24. Brümmer N, du Preez J. Application-independent evaluation of speaker detection. *Computer Speech and Language*. 2006; 20:230–275. doi: [10.1016/j.csl.2005.08.001](https://doi.org/10.1016/j.csl.2005.08.001)
25. Ahuja RK, Orlin JB. A fast scaling algorithm for minimizing separable convex functions subject to chain constraints. *Operations Research*. 2001; 49:784–789. doi: [10.1287/opre.49.5.784.10601](https://doi.org/10.1287/opre.49.5.784.10601)
26. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2002;694–699.
27. Ramos D, Gonzalez-Rodriguez J, Zadora G, Aitken C. Information-Theoretical Assessment of the Performance of Likelihood Ratio Models. *Journal of Forensic Sciences*. November 2013; 58(6):1503–1518.
28. Ramos D, Gonzalez-Rodriguez J. Reliable support: measuring calibration of likelihood ratios. *Forensic Science International*. 2013; 230:156–169. doi: [10.1016/j.forsciint.2013.04.014](https://doi.org/10.1016/j.forsciint.2013.04.014) PMID: [23664798](https://pubmed.ncbi.nlm.nih.gov/23664798/)