



Gaussianization of LA-ICP-MS features to improve calibration in forensic glass comparison

Pablo Ramirez-Hereza ^{a,*}, Daniel Ramos ^a, Juan Maroñas ^a, Sergio A. Balanya ^a, Jose Almirall ^b

^a AUDIAS Laboratory - Audio, Data Intelligence and Speech, Escuela Politecnica Superior, Universidad Autónoma de Madrid, Calle Francisco Tomás y Valiente 11, 28049 Madrid, Spain

^b Center for Advanced Research in Forensic Science, Department of Chemistry and Biochemistry, Florida International University, USA

ARTICLE INFO

Article history:

Received 23 August 2022

Received in revised form 30 March 2023

Accepted 17 May 2023

Available online 19 May 2023

Keywords:

Forensic Glass Comparison

LA-ICP-MS

Likelihood Ratio

Gaussianization

Normalization flows

Bayesian models

ABSTRACT

The forensic comparison of glass aims to compare a glass sample of an unknown source with a control glass sample of a known source. In this work, we use multi-elemental features from Laser Ablation Inductively Coupled Plasma with Mass Spectrometry (LA-ICP-MS) to compute a likelihood ratio. This calculation is a complex procedure that generally requires a probabilistic model including the within-source and between-source variabilities of the features. Assuming the within-source variability to be normally distributed is a practical premise with the available data. However, the between-source variability is generally assumed to follow a much more complex distribution, typically described with a kernel density function. In this work, instead of modeling distributions with complex densities, we propose the use of simpler models and the introduction of a data pre-processing step consisting on the Gaussianization of the glass features. In this context, to obtain a better fit of the features with the Gaussian model assumptions, we explore the use of different normalization techniques of the LA-ICP-MS glass features, namely marginal Gaussianization based on histogram matching, marginal Gaussianization based on Yeo-Johnson transformation and a more complex joint Gaussianization using normalizing flows. We report an improvement in the performance of the Likelihood Ratios computed with the previously Gaussianized feature vectors, particularly relevant in their calibration, which implies a more reliable forensic glass comparison.

© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In forensic science, multivariate quantitative physicochemical data is extracted from the analysis of various kinds of trace evidence. The role of the forensic expert is to evaluate such physicochemical data (also called evidence, E) in the context of two competing propositions, H_1 and H_2 [1]. An example of forensic science discipline is the comparison of glasses found in cases such as house robberies or hit-and-run car accidents.

In the forensic glass comparison problem, fragments of glass are found at a crime scene and over a certain person of interest. Those fragments whose source is unknown are defined as *recovered* fragments, while those fragments whose source is known are defined as *control* glasses. As a matter of example, in hit-and-run car accidents, the *recovered* fragments are those found in the victim's clothes, and the *control* fragments are those extracted from the windshield of the

suspect's car. Then, physicochemical features are extracted from both fragments using analytical chemistry methods, generating a set of so-called *recovered* and *control* features. Multiple features can be extracted from glass fragments. For example, the glass refractive index measurement (GRIM) and the concentration of oxides obtained by the scanning electron microscopy coupled with an energy dispersive X-ray spectrometer (SEM-EDX) are often used in many forensic institutions for the investigation of glass and other trace evidences [2]. Nevertheless, isotope concentrations extracted from the Laser Ablation with Inductive Coupled Plasma and Mass Spectrometry (LA-ICP-MS) method have proven to have a higher discrimination power than other features [3–6].

Once these features are taken from the glass fragments, the forensic scientist must evaluate and report this evidence in the context of propositions: H_1 , also called in literature H_p after *prosecution*, which states that the glass samples being compared came from the same source; and H_2 , or H_d after *defense*, which states that these glass samples came from different sources. Regarding this evaluation and reporting, the guideline released by the European Network of Forensic Science Institutes (ENFSI) recommends rigorous

* Corresponding author.

E-mail address: pablo.ramirez@estudiante.uam.es (P. Ramirez-Hereza).

¹ <http://audias.ii.uam.es>

probabilistic evidence evaluation for all forensic disciplines in Europe, including the glass comparison. The ENFSI guideline recommends the determination of the value of the evidence by a likelihood ratio (LR). The LR is the probability of the evidence if H_p is true, $P(E|H_p)$, divided by the probability of the evidence if H_d is true, $P(E|H_d)$. This LR takes into account the similarity between the control and recovered samples, the within-source and between-source variation, the rarity of the elemental profile, and the correlation between variables [2].

The computation of the LR is a complex procedure that requires modeling distributions $P(E|H_p)$ and $P(E|H_d)$, giving rise to what we will refer to as *LR models*. Moreover, following the typical machine learning terminology, we will refer to *training* as this modeling process. Thus, the training process of the LR models requires a sufficiently large and representative *Background* database. However, despite the growing interest in increasing and facilitating access to glass databases, data scarcity is a clear obstacle to the glass comparison problem. This limitation is amplified by the need for additional data to define a rigorous validation process for the correct evaluation of the discrimination power and the calibration of the LR models [7].

In this context, two LR-based strategies for the forensic glass comparison problem are found in the literature. Both approaches are schematized in Fig. 1. On the one hand, *feature-based LR models* compute the ratio directly from the multivariate features extracted from the evidence glasses. We will denote the features of the control sample as x_c , and those of the recovered samples as x_r . In our problem, both feature vectors are of dimension 17. In this way, the computation of LR requires the probabilistic modeling of complex multivariate distributions. In order to simplify this computation, and even obtain closed-form equations of the LR, most of the feature-based approaches define some assumptions over the within-source and between-source variabilities. As a matter of example, Aitken and Lucy [8] propose an LR model assuming that the within-source variability follows a Gaussian distribution and the between-source variability is either a Gaussian distribution (MVN model) or a more complex distribution described with a kernel density function (MVK model). Despite the high discriminating power of these models, these assumptions lead to the computation of badly calibrated LR values, which represents their strongest limitation.

On the other hand, *score-based LR models* (or *SLR*) compare the probability of observing the score between control and recovered samples given each proposition [6]. The score $s_{r,c}$ is a measure that takes higher values the more the evidence supports H_p and, on the contrary, it takes lower values the more the evidence supports H_d . This score is computed from the multivariate features evidence. The computation of the LR from that score simplifies the LR model due to its low dimensionality (generally univariate), in contrast with the high-dimensional models of the feature space. As a matter of example, the Euclidean distance, the Pearson correlation distance, or uncalibrated LR from a feature-based LR model are already used as scores in different forensic disciplines. The good calibration properties of the SLRs make them used in different branches of forensic science such as glass comparison [6], MDMA comparison [9], forensic speaker recognition [10,11], or biometric systems [12,13]. However, in practice, their main disadvantage is that their performance relies on the selection of the score. Simplistic scores usually lead to a high loss of information, hence degradation of the discrimination power and the calibration of the computed LR. In contrast, more complex score schemes require additional data to be trained, which aggravates the problem of data scarcity inherent in the forensic glass comparison [14]. Despite the good performance of the SLR models in some scenarios, we consider it crucial to delve into feature-based LR models, since modeling complex feature spaces with good calibration might lead to elegant solutions that can be easily extrapolated to other forensic problems and features beyond glass comparison. For this reason, in this work, we propose an approach for the computation of well-calibrated LR with feature-based models. It is worth noting that a feature-based LR model can be also combined with a score-based LR model in order to recalibrate the LR values (as it has been done in the past, see e.g. [6]). However, in this work, we will focus in obtaining the best possible calibration without the use of a score-based stage, and therefore without the need of the necessary data for score-based models.

As it was mentioned before, one of the limitations of the MVN and MVK feature-based models is the definition of Gaussian assumptions of the between-source and within-source variability. These assumptions are far from being fulfilled in real LA-ICP-MS data, which may lead to the computation of bad-calibrated LR. To generate well-calibrated models which fit better to the raw LA-ICP-

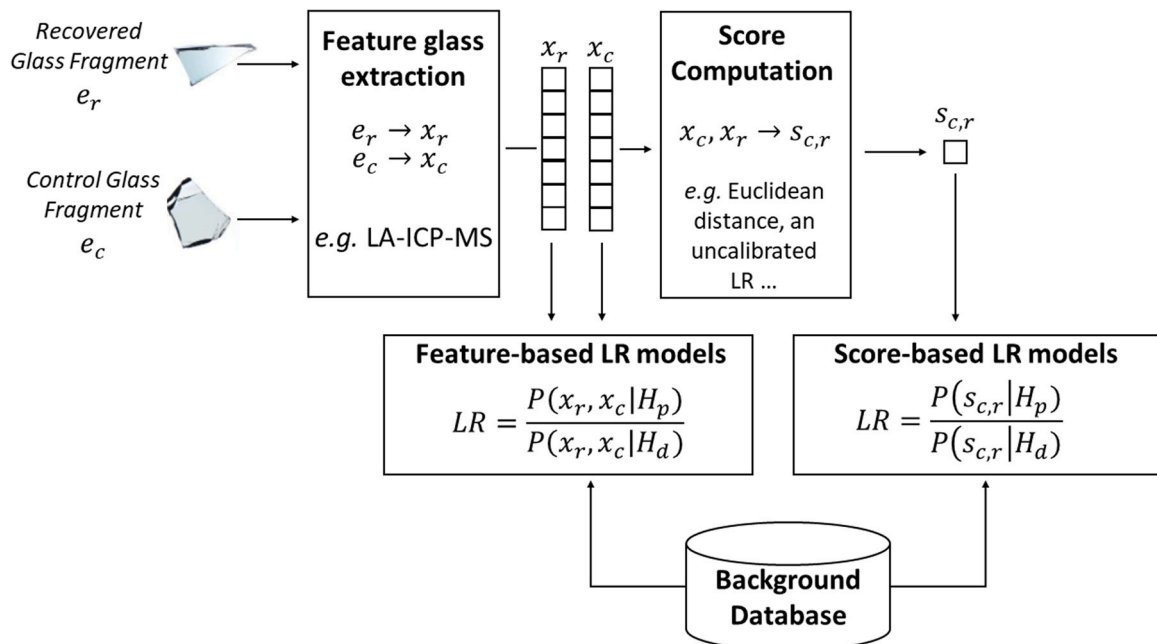


Fig. 1. Likelihood Ratios computation schemes.

MS data, other approaches propose modeling the between-source variability with more complex distributions estimated with Kernel density functions, Variational Autoencoders, or warped Gaussian mixtures [14]. In contrast, we propose the introduction of a pre-processing step to transform the LA-ICP-MS features in such a way that the resulting data conform to Gaussian considerations of the within-source and between-source variabilities. The application of this Gaussianization transformation and the computation of the Likelihood Ratios with simpler models (MVN or MVK) represents an elegant and very intuitive alternative, more robust to data scarcity and that does not require additional data, since the Gaussianization can be trained directly with the *Background* database. Thus, in this work an exploration of different Gaussianization transformations has been performed. Specifically, results have been obtained using a parametric Yeo-Johnson marginal Gaussianization, a non-parametric marginal Gaussianization based on histogram matching, and a more complex joint Gaussianization based on normalizing flows.

For the evaluation of the models and Gaussianization transformations proposed in this work, we have performed a rigorous validation process using different LA-ICP-MS databases. Specifically, we have used the Florida International University (FIU) vehicle dataset and the Bundeskriminalamt (BKA) casework database. These Gaussianization transformations can be performed over different features and tasks. Specifically, in the forensic glass comparison discipline, we report a clear improvement in the calibration of the Likelihood Ratios due to the transformation of the data. Thus, we consider this work may have an impact not only in the forensic glass comparison discipline, but also in other branches of forensic science.

This work is organized as follows. First, in sections 2 and 3, a description of the feature-based LR models and the Gaussianization techniques is provided. Then, in section 4, we present the experimental framework, consisting of the LA-ICP-MS databases, the evaluation metrics, and the evaluation protocol. Next, the results are graphically represented in section 5. Finally, conclusions are drawn in section 6.

2. Likelihood ratio models

In this section, we describe the feature-based models used to compute the Likelihood Ratios. Specifically, these models are based on the classical Bayesian approaches for trace evidence, firstly introduced by [15] for univariate data, and later extended for multivariate features in [8]. Both models were introduced for glass data, and they are still in use today in the forensic glass comparison problem.

To introduce the mathematical notation needed for the description of the models, let us consider a forensic case example, such as a hit and run case, where a recovered glass sample is found at a crime scene, and control glass fragments are extracted from the windshield of the suspect's car.

For each glass sample, some d -dimensional feature vectors are extracted by an analytical chemistry method (in our case LA-ICP-MS). Vectors extracted from the same fragment are called *replicates* or *replicated measurements*. We will refer to the set of n_1 replicates of the control glass as $\mathbf{Y}_1 = (y_1^{(1)}, \dots, y_1^{(n_1)})$. Analogously, we will refer to the set of n_2 replicates extracted from the recovered glass as $\mathbf{Y}_2 = (y_2^{(1)}, \dots, y_2^{(n_2)})$. Thus, the set $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ is the feature-based representation of the evidence of the forensic case. Given, \mathbf{Y} we can compute the mean feature vector of each sample as:

$$\hat{y}_l = \frac{1}{n_l} \sum_{i=1}^{n_l} y_l^{(i)}, \quad l = 1, 2 \quad (1)$$

Moreover, let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_m)$ be the available Background database of m glass objects. Then, we will refer to the set of n_b d -dimensional replicates extracted from glass object i as $\mathbf{X}_i = (x_i^{(1)}, \dots, x_i^{(n_b)})$. For the sake of simplicity, we will assume that the number of replicates are

the same in any case, $n_1 = n_2 = n_b = n$. Then, some statistics can be extracted from the Background database:

- Per-object mean on the background database:

$$\hat{x}_i = \frac{1}{n} \sum_{j=1}^n x_i^{(j)} \quad (2)$$

- Total mean of the background database:

$$\hat{x} = \frac{1}{m} \sum_{i=1}^m \hat{x}_i \quad (3)$$

- Within-source covariance matrix:

$$\hat{\Sigma}_w = \frac{1}{mn - m} \sum_{i=1}^m \sum_{j=1}^n (x_i^{(j)} - \hat{x}_i)(x_i^{(j)} - \hat{x}_i)^T \quad (4)$$

- Between-source covariance matrix:

$$\hat{\Sigma}_b = \frac{1}{m - 1} \sum_{i=1}^m (\hat{x}_i - \hat{x})(\hat{x}_i - \hat{x})^T \quad (5)$$

The classical approach [8,15] for the computation of the Likelihood Ratio is defined by the following expression:

$$\begin{aligned} LR &= \frac{P(E|H_p, \mathbf{X})}{P(E|H_d, \mathbf{X})} = \frac{P(\mathbf{Y}_1, \mathbf{Y}_2|H_p, \mathbf{X})}{P(\mathbf{Y}_1, \mathbf{Y}_2|H_d, \mathbf{X})} \\ &= \frac{\int P(\hat{y}_1|\theta)P(\hat{y}_2|\theta)P(\theta|\mathbf{X})d\theta}{\int P(\hat{y}_1|\theta)P(\theta|\mathbf{X})d\theta \int P(\hat{y}_2|\theta)P(\theta|\mathbf{X})d\theta} \end{aligned} \quad (6)$$

where θ are the parameters of the model. In the previous expression, the *likelihood* $P(y|\theta)$ determines the variation of the features given a fixed set of parameters θ , hence it models the within-source variability. In contrast, the *parameter distribution*, $P(\theta|\mathbf{X})$ determines how the parameters of the model vary given a Background database, hence it models the between-source variability. Note that, as the values are continuous, each P is assumed to be a probability density function.

The models used in this work, the MVN and MVK models, define certain assumptions regarding both distributions (likelihood and parameter distribution) in order to have closed-form solutions of integrals in equation (6). These closed-form equations are detailed in [8]. Despite their high discriminating power, it has been proven in previous works that these models are poorly calibrated. In fact, it has been observed that, when used with LA-ICP-MS data, these models often output infinite and zero LR values, evidencing this miscalibration problem [6,16]. Thus, to avoid numerical and computational pitfalls, we have limited the LR values to lay in the range between the computer minimum and maximum floating-point precision. The calibration problem of these models may be caused either by the presence of strongly misleading evidences, or the presence of cases where the LR cannot be computed due to numerical problems. In those cases where the LR cannot be computed due to numerical issues (i.e. when both the denominator and numerator are null), an LR equal to 1 is generated. These values of LR=1 may cause miscalibration problems, because they increase the value of Cllr.

2.1. MVN model

The model referred to as MVN in [8] defines the following assumptions:

- The likelihood is a multivariate normal distribution with an unknown vector mean and a maximum likelihood point-estimated covariance matrix, whose value is equal to the within-source covariance matrix.

$$P(y_l|\theta) \sim \mathcal{N}(y_l|\theta, \Sigma_w)$$

- The parameter distribution is also a multivariate normal distribution defined by the point-estimated statistics total mean vector and between-source covariance matrix.

$$P(\theta|\mathbf{X}) \sim \mathcal{N}(\theta|\mu, \Sigma_b)$$

2.2. MVK model

A multivariate normal distribution for θ may not always be a reasonable assumption. Thus, the model referred to as MVK in [8] defines the following assumptions:

- As in the MVN model, the likelihood is assumed to be a multivariate normal distribution with an unknown vector mean and a maximum likelihood point-estimated covariance matrix, whose value is equal to the within-source covariance matrix.

$$P(y_i|\theta) \sim \mathcal{N}(y_i|\theta, \Sigma_w)$$

- In contrast, the parameter distribution is assumed to be a non-parametric kernel density function with multivariate Gaussian Kernel functions having a maximum likelihood estimated covariance matrix equal to $h^2\hat{\Sigma}_b$.

$$P(\theta|\mathbf{X}) \sim \frac{1}{m} \sum_{i=1}^m K(\theta|\hat{x}_i, \Sigma_b, h)$$

where the smoothing parameter is computed as in [17]:

$$h = \left(\frac{4}{2D+1} \right)^{\frac{1}{D+4}} m^{-\frac{1}{D+4}}$$

It is important to remark that, in both the MVK and MVN models, the parameters μ , Σ_w , and Σ_b are estimated using their corresponding sample estimations seen in equations (2) to (5).

3. Gaussianization techniques of the feature space

The computation of a feature-level likelihood ratio requires a probabilistic model of the within-source and between-source variabilities of the glass features vectors. Instead of modeling these distributions with complex densities (such as Variational Autoencoders, Gaussian processes or Warped Gaussian mixtures), we propose the introduction of a pre-processing step to obtain a better fit of the features with the Gaussian model assumptions defined in section 2.

Specifically, the proposed models establish Gaussian assumptions over the within-source and/or between-source variability distributions. However, these assumptions are far from being fulfilled in raw data, such as LA-ICP-MS features. As a matter of example, Fig. 2 shows the empirical probability density function of element ^{146}Nd , and its empirical between-source and within-source variabilities. Moreover, these distributions are graphically compared with a Gaussian distribution by Quantile - Quantile representations (or QQplots). As it can be seen, none of the distributions is in fact Gaussian. This analysis has been performed for each LA-ICP-MS element, obtaining similar results that the one represented below. These wrong assumptions over data may cause a degradation of the performance of the calculated LR, especially in terms of calibration. Aiming to minimize the impact of these assumptions on the final performance of the LR-models, we propose a pre-processing step of the LA-ICP-MS features to make more Gaussian-like (or *Gaussianize*) the distributions mentioned above. Following the scheme in Fig. 3 scheme, this data transformation is performed before the computation of the LR with the feature-based models. Thus, the LR is

calculated with the *transformed*, or *gaussianized*, feature vectors of the recovered and control fragments, as well as the background data.

In subsections 3.1 and 3.2, we introduce the different Gaussianization transformations explored. Then, in subsection 3.3 we discuss the application of these transformations to the problem at hand, and define the three different ways, or *schemes*, analyzed to transform the features.

3.1. Gaussianization transformations

Let $\mathcal{X} \in \mathbb{R}^D$ be the corresponding glass *feature space*, for example the LA-ICP-MS feature space (with $D = 17$). Additionally, let $\mathcal{G} \in \mathbb{R}^D$ be a so-called *Gaussianized* or *transformed space* where the variables are intended to be jointly Gaussian distributed. Then, the bijective function g that transforms the feature space \mathcal{X} into the transformed space \mathcal{G} is called a *Gaussianization transformation*. Fig. 4 represents an example where a Gaussianization g is applied over a bivariate feature space formed by elements ^{57}Fe and ^{85}Rb , extracted by the LA-ICP-MS method.

It is important to remark that the joint Gaussianization of high-dimensional multivariate data is a non-trivial problem. The most straightforward approach consists on performing a univariate Gaussianization (also called marginal Gaussianization) of each feature space dimension independently. The univariate Gaussianization techniques found in the literature can be roughly divided into two categories: parametric power transformations, such as Box-Cox or Yeo-Johnson transformations [18,19], whose parameters are learned by maximum likelihood estimation; and non-parametric transformations, such as so-called histogram matching Gaussianization. Nevertheless, the marginal Gaussianization of each dimension may result in a transformed space with a non-Gaussian multivariate distribution.² On the other hand, a much more complex approach is the multivariate Gaussianization of the feature space, which is generally performed with iterative algorithms. As a matter of example we find in literature the *Rotation-based iterative Gaussianization* (RBIG) [20] and Gaussianization based on *normalizing flows* [21–23].

This section describes the different transformations explored to Gaussianize the glass measurements. Specifically, we propose and compare two marginal Gaussianizations transformations: the *Yeo-Johnson Gaussianization* and the *Histogram matching Gaussianization*; and a multivariate Gaussianization using *Gaussianization Flows*.

3.2. Gaussianization techniques

3.2.1. Histogram matching Gaussianization

Let $x \in \mathbb{R}^1$ be a univariate variable with strictly positive and differentiable probability density function. Let $F(\cdot)$ be the cumulative distribution function (CDF) of x . Let also define $\phi(\cdot)$ as a univariate standard Gaussian CDF, and $\phi^{-1}(\cdot)$ its quantile function (inverse of the CDF), namely the *probit* function. We can define the transformation g_{HM} as follows.

$$g_{HM}(x) = \phi^{-1}(F(x)) \quad (7)$$

It can be proven that $y = g_{HM}(x)$ is a univariate standard Gaussian distributed random variable. In practice, the CDF $F(\cdot)$ is not available; it has to be learned from the training data (e.g., with kernel density functions). Thus, the empirical cumulative distribution function $\hat{F}(\cdot)$ is computed from the training database. The estimation of $\hat{F}(\cdot)$ is sensitive to the training data available, leading to a non-perfect Gaussianization transformation. This error in the Gaussianization is increased if there is some dataset shift between the training and the testing databases.

² It is true that a Gaussian multivariate distribution has marginal Gaussian distributions, but not necessarily the other way round.

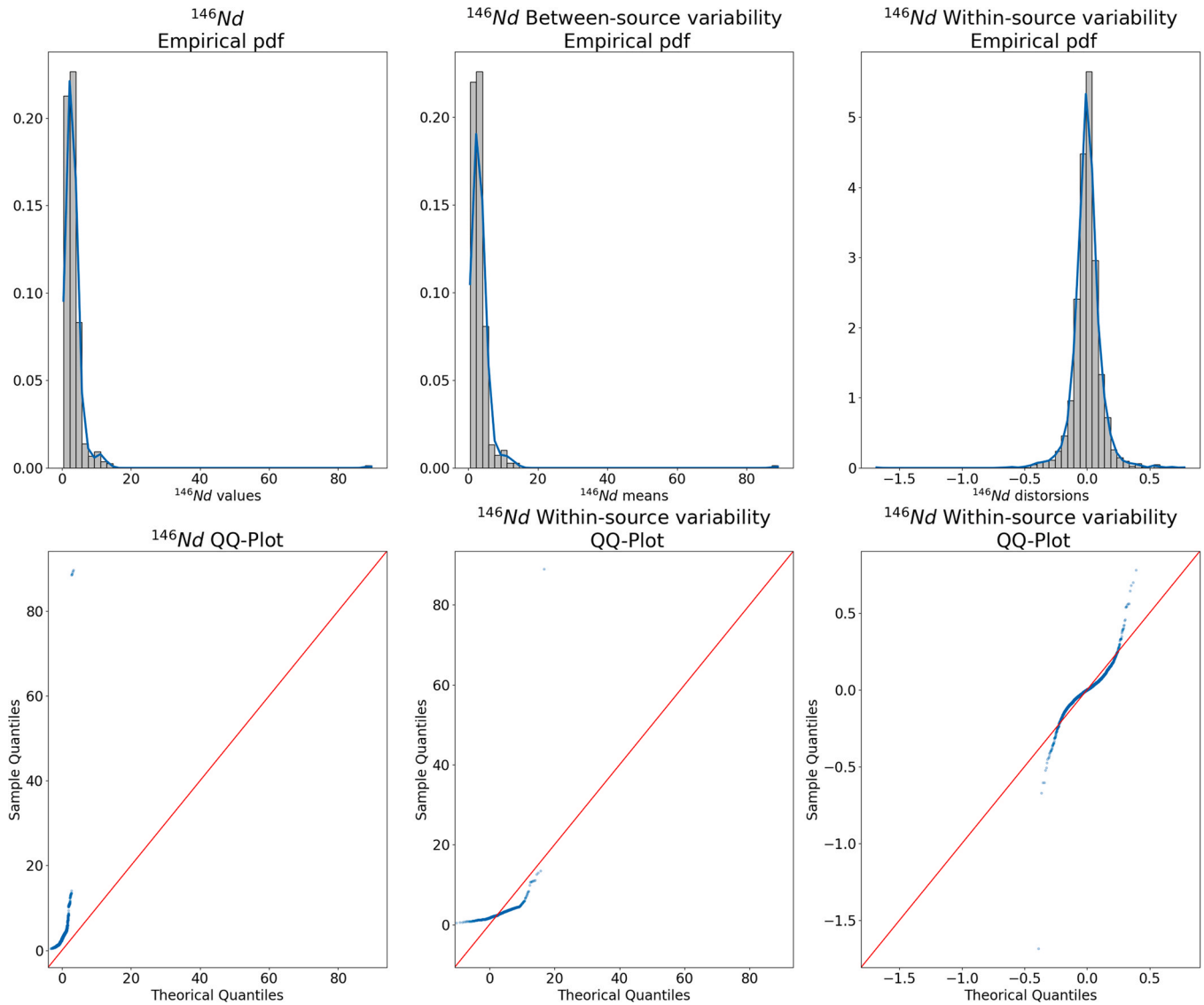


Fig. 2. Above: empirical pdf, between-source variability and within-source variability of element ^{146}Nd . Below: QQplots with the theoretical Quantiles of a Gaussian pdf.

Figure 5 represents this Gaussianization transformation over the element ^{88}Sr . The first column represents its empirical probability density function estimated by KDF and the empirical CDF $\hat{F}(\cdot)$ of the variable. The application of the transformation $y_1 = \hat{F}(x)$ results in a variable uniformly distributed. Then the application of the transformation defined by the probit function $y = \phi^{-1}(y_1)$ results in a Gaussianized variable.

3.2.2. Yeo-Johnson Gaussianization

Unlike the histogram-based Gaussianization, Yeo and Johnson [19] propose a power transform that presents many properties of Box-Cox power transformations [18] without being limited by any input variable restriction. Power transforms are a family of parametric and monotonic functions that aim to map data from any distribution to as close to a Gaussian distribution as possible. Let $x \in \mathbb{R}^1$, Yeo-Johnson transform is given by:

$$g_{\lambda}(x) = \begin{cases} ((x+1)^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0, \quad x \geq 0 \\ \log(x+1) & \text{if } \lambda = 0, \quad x \geq 0 \\ -[(-x+1)^{2-\lambda} - 1]/(2-\lambda) & \text{if } \lambda \neq 2, \quad x < 0 \\ -\log(-x+1) & \text{if } \lambda = 2, \quad x < 0 \end{cases} \quad (8)$$

where the unique parameter of the transformation λ is determined through maximum likelihood. Fig. 6 shows the empirical probability density function of the ^{88}Sr before and after applying the Yeo-Johnson transformation with different values of λ .

As it can be seen, the empirical distribution of the transformed variable is more Gaussian-like than the initial distribution. However, in contrast with the histogram-based Gaussianization, the low complexity of the Yeo-Johnson transformation may yield the generation of variables with distributions that are still far from being Gaussian.

3.2.3. Gaussianization flows

In order to introduce the Gaussianization Flows, let's first define the concept of Normalizing Flows. A Normalizing Flow can be defined as a transformation of a complex probability distribution into a simple distribution (e.g. a standard normal) by a sequence of invertible and differentiable mappings [21]. The main idea behind the Normalizing Flows is the change of variable theorem [24], and we describe the process as follows.

Let $X \in \mathbb{R}^D$ be a multivariate random variable with a known and tractable probability density function p_X . Let f be an invertible function in such a way that $g = f^{-1}$. The probability density function p_Y of $Y = f(X)$ can be computed as follows:

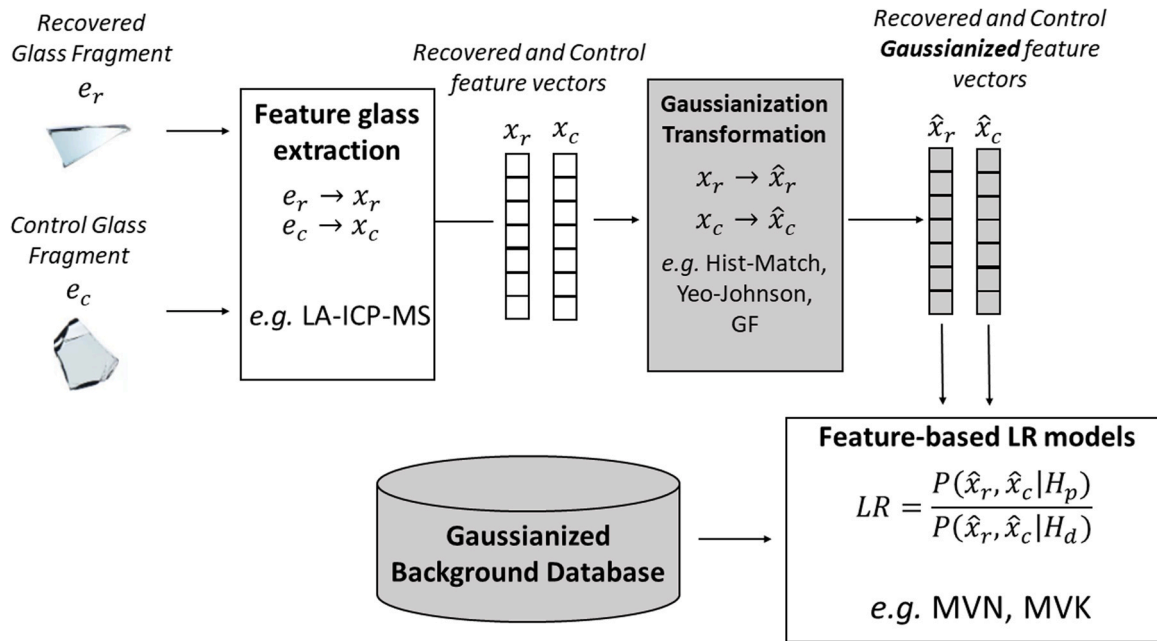


Fig. 3. Proposed pipeline for the computation of Likelihood Ratios combining a Gaussianization transformation and feature-based LR models.

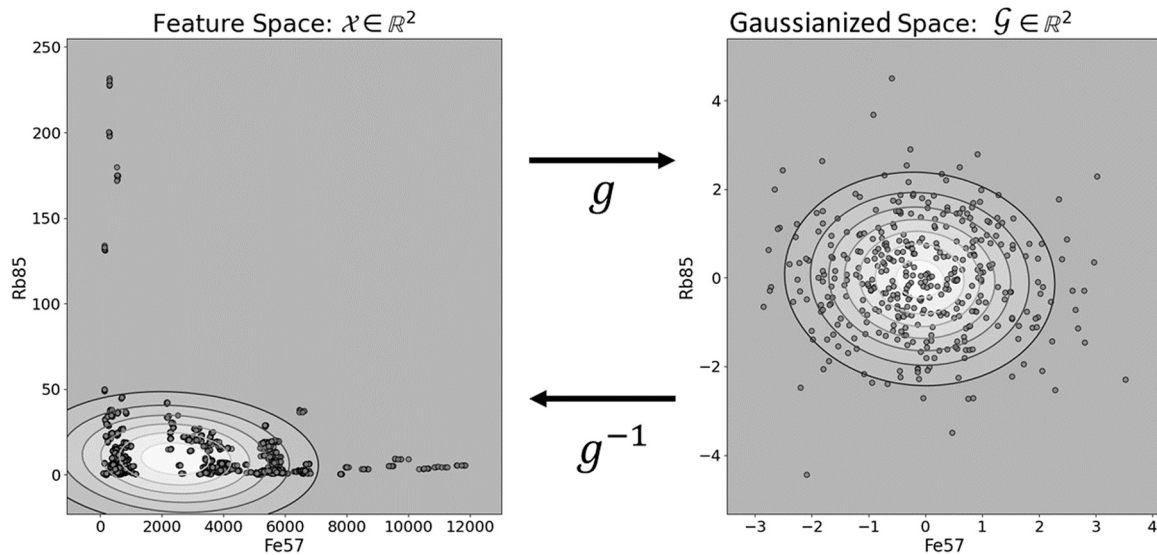


Fig. 4. Gaussianization transformation of a 2-dimensional feature space.

$$\begin{aligned} p_Y(y) &= p_X(g(y)) |\det(Dg(y))| \\ &= p_X(g(y)) |\det(Df(g(y)))|^{-1} \end{aligned} \quad (9)$$

where $Df(y)$ and $Dg(x)$ are the Jacobian matrices of the transformations f and g respectively, and $\det(\cdot)$ denotes the determinant of a matrix. While function f transforms the base density p_X to a more complex density, the inverse function g normalizes the data distribution to the base probability density function. It is worth noting that p_X can be any known distribution. In this work, we consider a D -dimensional standard normal $p_X = \mathcal{N}(0, I)$.

In practice, it is fairly complex to obtain the definition of a transformation g sufficient expressive, convenient to compute, invertible, and with a tractable determinant. Thus, the normalizing flows approach consists of deriving an expressive transformation by sequentially applying multiple simpler functions, with an easy to compute Jacobian. In such a way that the resulting transformation

$g_{NF} = g_1 \circ \dots \circ g_{N-1} \circ g_N$ is also an invertible transformation, whose inverse can be defined as $f_{NF} = f_1 \circ \dots \circ f_{N-1} \circ f_N$, and whose determinant of the Jacobian can be computed as follows:

$$\det Dg_{NF}(x) = \prod_{i=1}^N \det Dg_i(x_i) \quad (10)$$

where $x_i = g_1 \circ \dots \circ g_i$ is the value of the i -th intermediate flow, and $Dg_i(x)$ is the Jacobian of g_i . Fig. 7 schematized a Normalizing flow model.

Gaussianization flows are a type of normalizing flows based on Rotation-Based Iterative Gaussianization (RBIG) [20]. This method iterates between the following steps:

- A one-dimensional Gaussianization of all the marginal distributions of the input space. This Gaussianization is performed by histogram matching (see section 3.2.1).
- The application of a rotation transformation, defined by a fixed rotation matrix T .

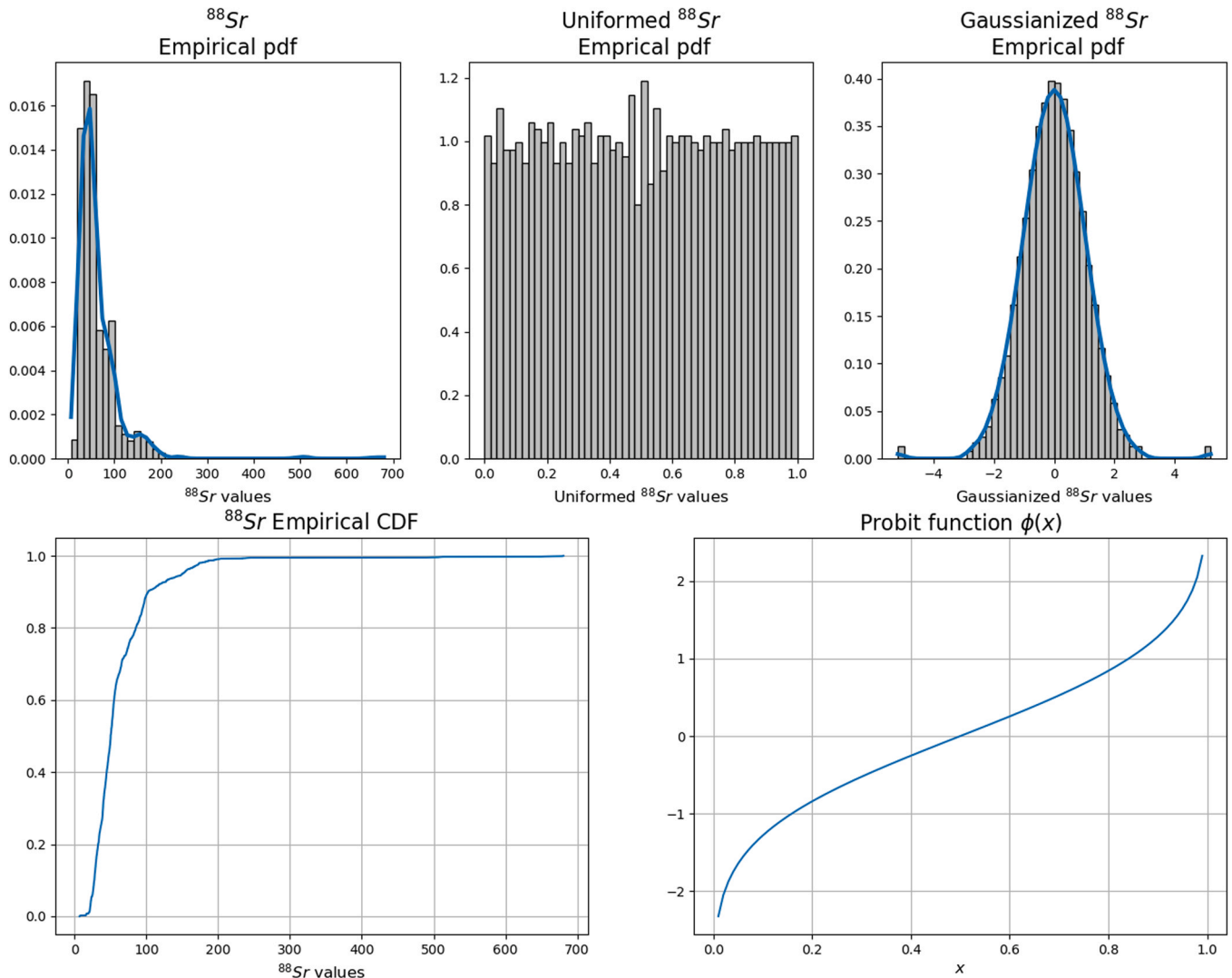


Fig. 5. Histogram - matching Gaussianization process.

Gaussianization flows improve on RBIG by parametrizing these steps. First, it replaces one-dimensional non-parametric Gaussianization with a so-called *trainable kernel layer*. In such a way that, the empirical cumulative density function is no longer trained by choosing a suitable kernel, but optimizing its parameters. Thus, a set of anchor points and bandwidth parameters are learned to estimate the CDF of each variable. Second, the fixed rotation matrix is replaced by a so-called *orthogonal kernel layer*, which consists on a parameterized rotation matrix using Householder reflections [26].

Formally, a Gaussianization flow with L kernel (Ψ) and orthogonal (R) layers can be defined as:

$$g_{GF}(x) = \Psi_L \circ R_L \circ \Psi_{L-1} \circ R_{L-1} \circ \dots \circ \Psi_1 \circ R_1 x \quad (11)$$

Different values of L have been explored. Nevertheless, for the sake of simplicity, we will only consider the best configuration in terms of C_{thr} . Thus, from now on, all the Gaussianization flows mentioned are formed by $L = 6$ kernel and orthogonal trainable layers.

3.3. Gaussianization strategies for forensic glass comparison

In this section, the application of the proposed Gaussianization transformations in the forensic glass comparison task is discussed. In the problem at hand, we aim to transform the data in order to make

it more consistent with a Gaussian distribution, to obtain a better fit with the feature-based models. To do that, we perform Gaussianization transformations in the data. Three Gaussianization strategies or schemes have been explored: the *Gaussianization of the within-source and between-source variabilities*, the *Gaussianization of the between-source variability*, and the *Gaussianization of the feature space*.

The description of all the approaches is done following the notation defined in section 2.

3.3.1. Gaussianization of the within-source and between-source variabilities

According to models assumptions, the between-source and within-source variabilities of the glass fragments should be Gaussian-distributed. However, the available measurements of the same source are too scarce (as a matter of example, BKA dataset has only six replicate measurements per glass object). This precludes the correct estimation and Gaussianization of the within-source variability. Our approach to Gaussianize the within-source variability consists of pooling the within-source variabilities of all the glass objects and Gaussianize the overall.

Given the set of replicates for the recovered and control fragments $\mathbf{Y}_i = (y_i^{(1)}, \dots, y_i^{(n)})$, the transformed replicates $\mathbf{Z}_i = (z_i^{(1)}, \dots, z_i^{(n)})$ can be computed following expression (12).

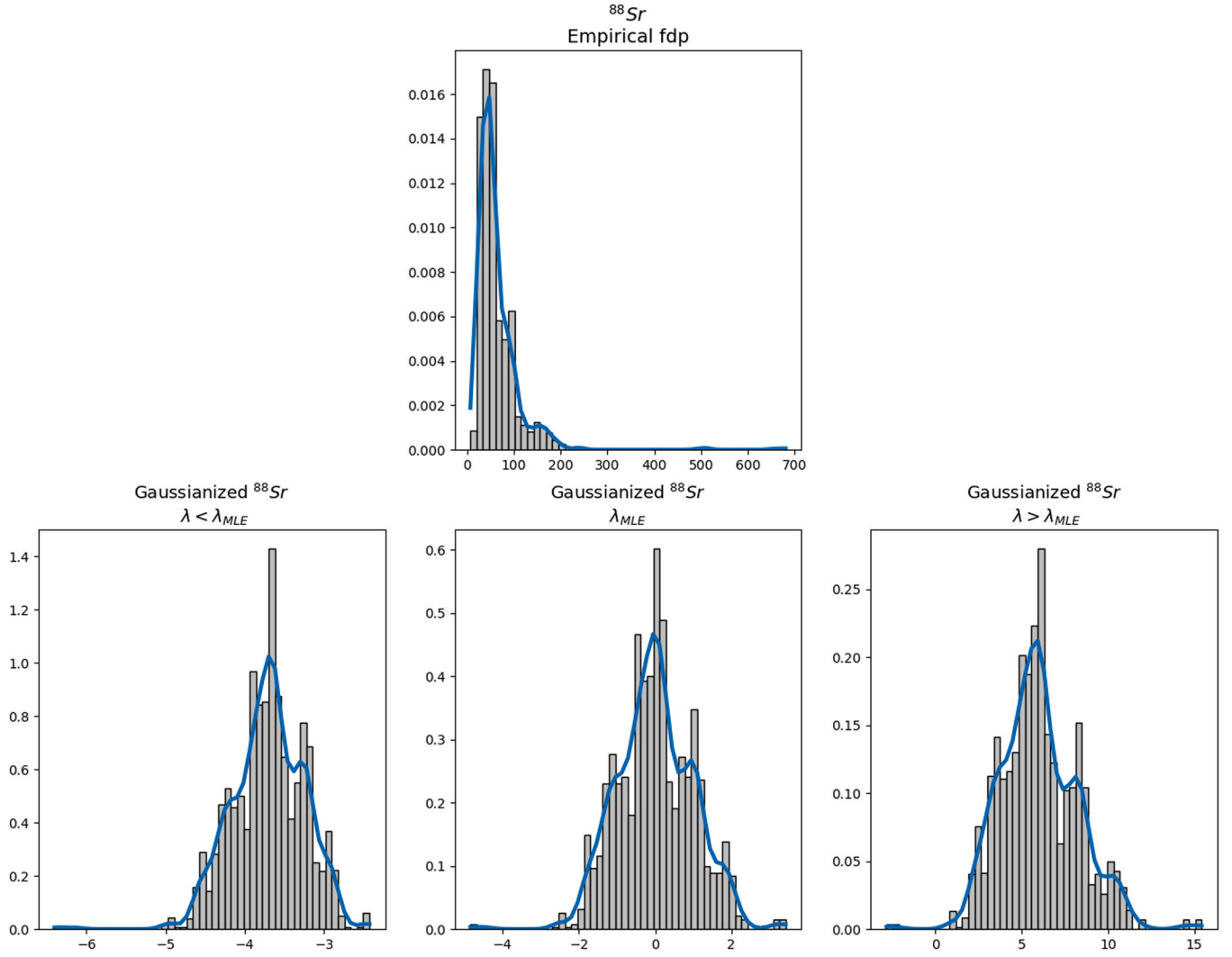


Fig. 6. Gaussianization with Yeo-Johnson transformation using different values of λ .

$$z_l^{(i)} = g^1(y_l^{(i)}) + g^2(g^1(y_l^{(i)}) - g^1(\hat{y}_l)) \quad (12)$$

This transformation is a two-step Gaussianization, where g^1 corresponds to the Gaussianization of the between source-variability, and g^2 corresponds to the Gaussianization of the within-source variability. Specifically, g^1 and g^2 are the same Gaussianization technique, that will be selected from the ones described in section 3.2 (g_{HM} , g_{YJ} and g_{CF}). However, both transforms

differ in their training data. On the one hand, g^1 is trained with the set of per-object vector means of the Background dataset $\hat{\mathbf{X}} = (\hat{x}_1, \dots, \hat{x}_m)$. On the other hand, g^2 is trained with the set of within-source differences of the transformed features $z_{ij} = g^1(x_i^{(j)}) - g^1(\hat{x}_i)$ where $i = (1, \dots, m)$ and $j = (1, \dots, n_r)$.

As a matter of example, Fig. 8 shows the Q-Q representations of the probability density function, the between-source variability and the pool of within-source variabilities for the transformed ^{146}Nd . As

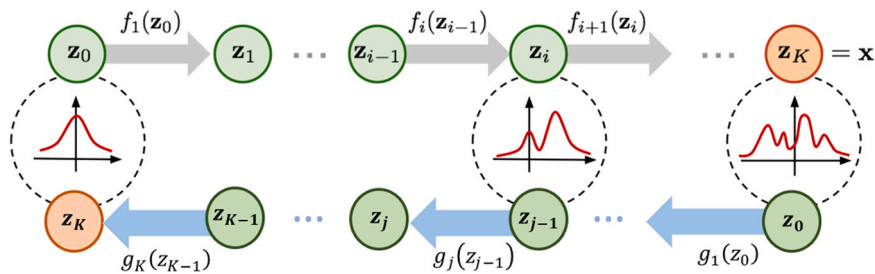


Fig. 7. Normalizing Flow scheme inspired by [25]. On the top, the generative direction: from a base distribution to a complex distribution. On the bottom, the normalizing direction: from a complex distribution to a base distribution.

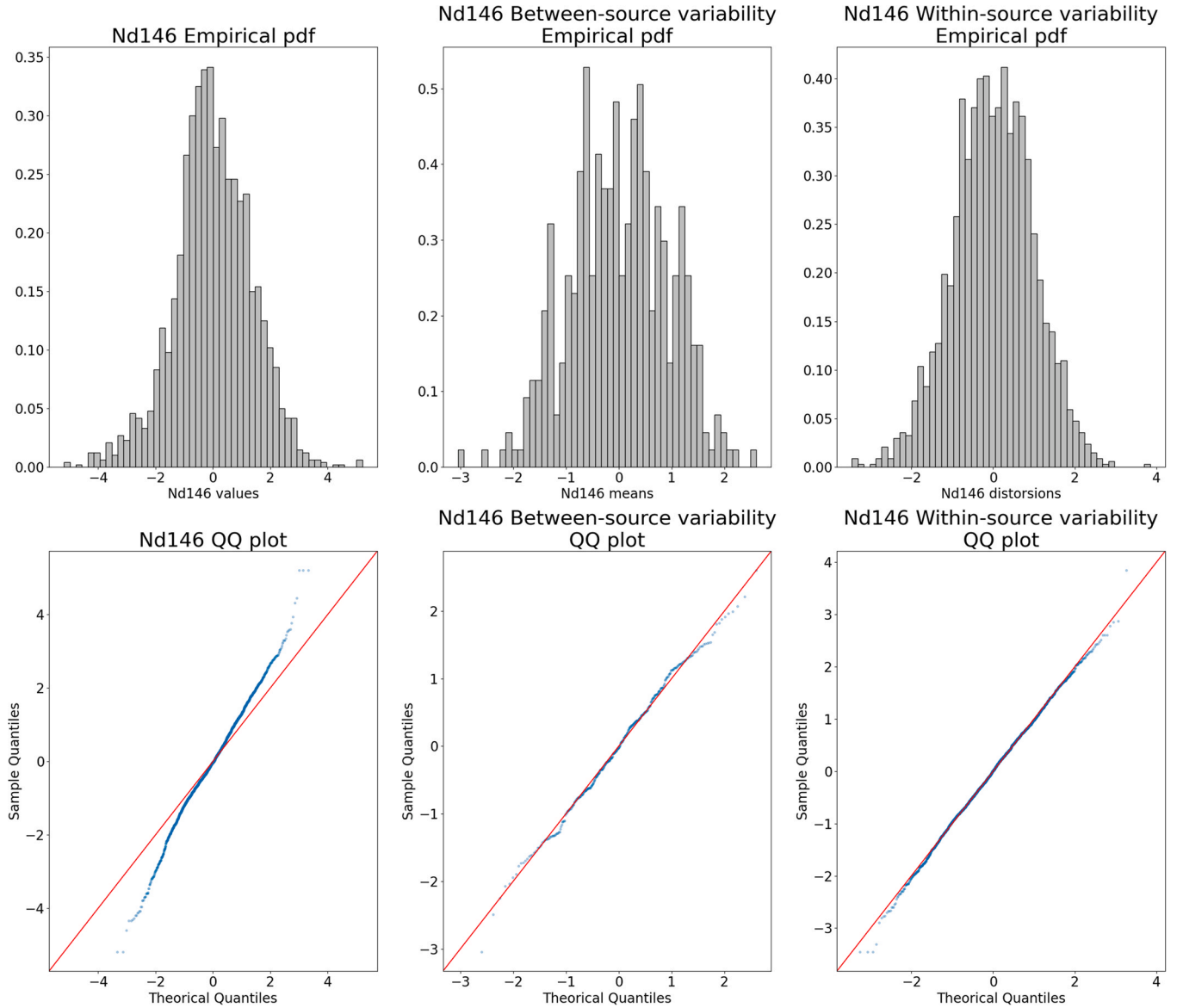


Fig. 8. (a) Quantile-Quantile and histogram representations for the probability density function, the between-source variability and the pool of within-source variabilities for the transformed LA-ICP-MS element 146Nd.

it can be seen and compared with Fig. 2, the transformation defined in equation (12) Gaussianizes the between-source variability and the pool of within-source variabilities.

3.3.2. Gaussianization of the between-source variability

Aiming to avoid the issue concerning the Gaussianization of the within-source variability, our second approach consists on the unique Gaussianization of the between-source variability. This transformation is defined in equation (13).

$$z_l^{(i)} = g(y_l^{(i)}) \quad (13)$$

As described in the previous case, g is one of the Gaussianization transformations in section 3, and it is trained with the set of per-object mean vectors of the Background dataset.

It can be seen in Fig. 9 that this transformation performs a better Gaussianization of the Between-source variability, without Gaussianizing the within-source variability. Moreover, this transformation also

results in a Gaussianization of the probability density function of the variable. It is worth noting that this transformation is trained using only one feature vector per object in the Background database. Nevertheless, the number of glass fragments in the available databases may be too scarce. This may cause overfitting of the Gaussianization transformation to the Background data.

3.3.3. Gaussianization of the LA-ICP-MS feature space

In order to maximize the number of measurements used for training the transformation, our third approach consists on the Gaussianization of the LA-ICP-MS feature space. This approach is also defined by equation (13). Nevertheless, in this case, g is trained with the entire Background dataset \mathbf{X} independently of the source of the fragments. In practice, we noticed that the Gaussianization of the feature space generally leads to a Gaussianization of the between-source variability (as a matter of example see Fig. 10). However, using all the replicated measurements for training may also lead to a transformation more sensitive to possible outliers.

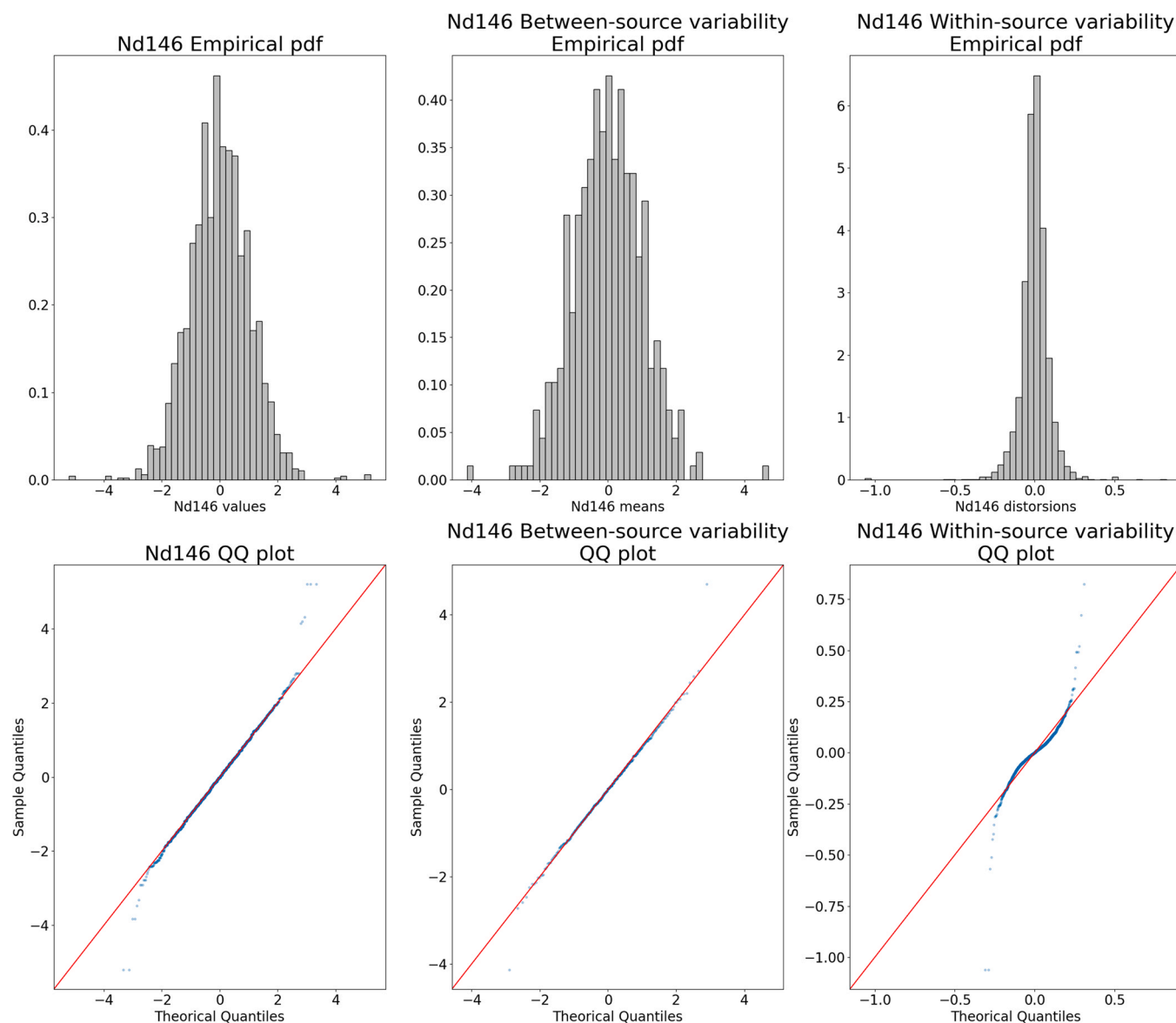


Fig. 9. Quantile-Quantile representations for the probability density function, the between-source variability and the pool of within-source variabilities for the transformed LA-ICP-MS element ^{146}Nd .

4. Experimental framework

4.1. Databases

To evaluate the performance of the feature-based LR models and the Gaussianization techniques, different LA-ICP-MS databases have been used. The LA-ICP-MS analytical technique implies that, following the element menu suggested in the United States standard ASTM E2927, each measurement of glass generates a vector of $D = 17$ values related to chemical isotopes found in the glass object. These chemical isotopes are ^7Li , ^{25}Mg , ^{27}Al , ^{39}K , ^{42}Ca , ^{49}Ti , ^{55}Mn , ^{57}Fe , ^{85}Rb , ^{88}Sr , ^{90}Zr , ^{137}Ba , ^{139}La , ^{140}Ce , ^{146}Nd , ^{178}Hf and ^{208}Pb .

Two LA-ICP-MS glass databases have been used in this work:

1. The Bundeskriminalamt (BKA) casework database is formed by measurements of 385 casework glass objects extracted from windows analyzed in real forensic cases. For each source, a total of six replicated measurements were taken. The database includes several types of glass relevant in forensic practice such as float glass, container glass, etc.

2. The Florida International University (FIU) vehicle dataset considers measurements of 210 windshields from vehicles at the M & M Service and Salvage Yard in Ruckersville, Virginia. Each windshield is formed by an inner and an outer pane, joined together with some fixing technique. The glass in the inner and outer pane can be considered different objects. Therefore we can account for a total of 420 glass objects. However, some similarities or dependencies between the inner and outer panes of the same windshield are expected to be found. Thus, the FIU database has been divided into two subsets, namely 210 *inner* panes and 210 *outer* panes.

A detailed description of the LA-ICP-MS instrumental parameters can be found in [6,14]. Both datasets have been divided in groups of $n_r = 3$ replicates for every single source. Thus, the BKA database has two replicate groups and the FIU database has five. In this way, the number of replicated measurements for the control sample (n_c) and the recovered sample (n_r) is the same in any comparison performed. It is worth noting that, this practice of equalizing the number of replicates is used by many forensic laboratories in order to isolate data variability issues related to the number of replicates.

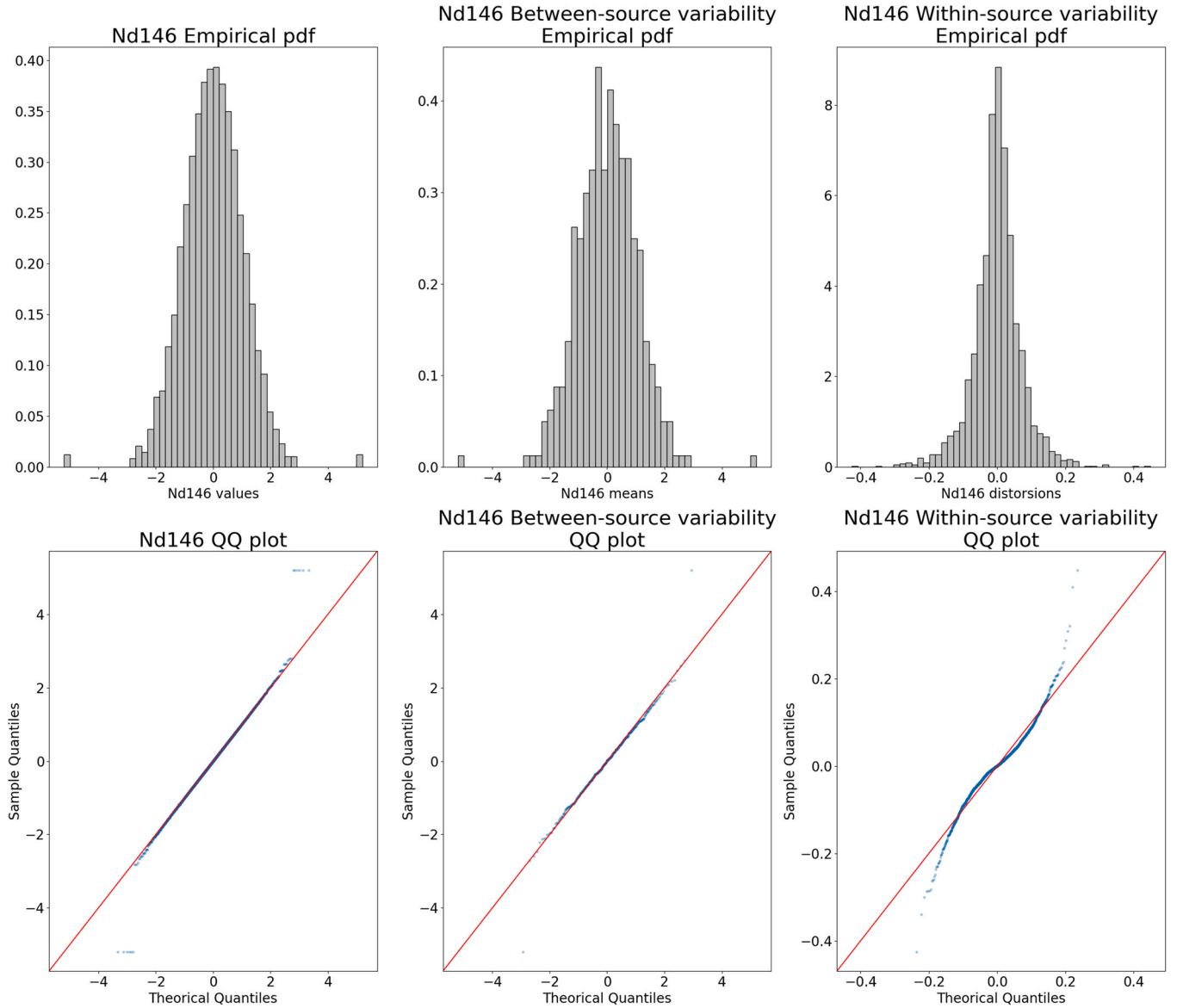


Fig. 10. (a) Quantile-Quantile representations for the probability density function, the between-source variability and the pool of within-source variabilities for the transformed LA-ICP-MS element Nd146.

4.2. Evaluation metric

The performance of the Likelihood Ratios computed by the feature-based LR models with the pre-processed LA-ICP-MS features is evaluated with the *log-likelihood ratio cost* or C_{llr} . Given several same-source comparisons (n_p) and some different-source comparisons (n_d), the C_{llr} can be computed as follows:

$$C_{llr} = \frac{1}{2 \cdot n_p} \sum_{i_p} \log_2 \left(1 + \frac{1}{LR_{i_p}} \right) + \frac{1}{2 \cdot n_d} \sum_{j_d} \log_2 (1 + LR_{j_d}) \quad (14)$$

where indices i_p and j_d correspond to the same-source comparisons and different-source comparisons respectively. As a matter of analysis, this metric can be decomposed into the discrimination cost (C_{llr}^{disc}) and the calibration cost (C_{llr}^{cal}) using the Pool Adjacent Violators (PAV) algorithm [27].

Despite the decomposition of the cost in these two components, we emphasize that the significant metric is the C_{llr} . Approaches with

a C_{llr} value greater than one does not provide any useful information, hence they should be discarded as approaches to use in real casework.

4.3. Evaluation protocol

The flow graph in Fig. 11 schematizes the evaluation procedure followed for a given Background dataset and a Validation dataset. As it can be seen, the Gaussianization transformation is trained with the Background dataset, while the feature-based LR model is trained with the transformed Background data. Then, an LR is computed for each pair of glass fragments of the validation dataset, using the Gaussianization transformation and the model.

A rigorous validation procedure is performed to address data scarcity and analyze the impact of dataset shift. This procedure consists of four experiments, in which we alternate the *Background database* and the *Validation database*. These experiments are summarized in Table 1. Some considerations have to be taken into

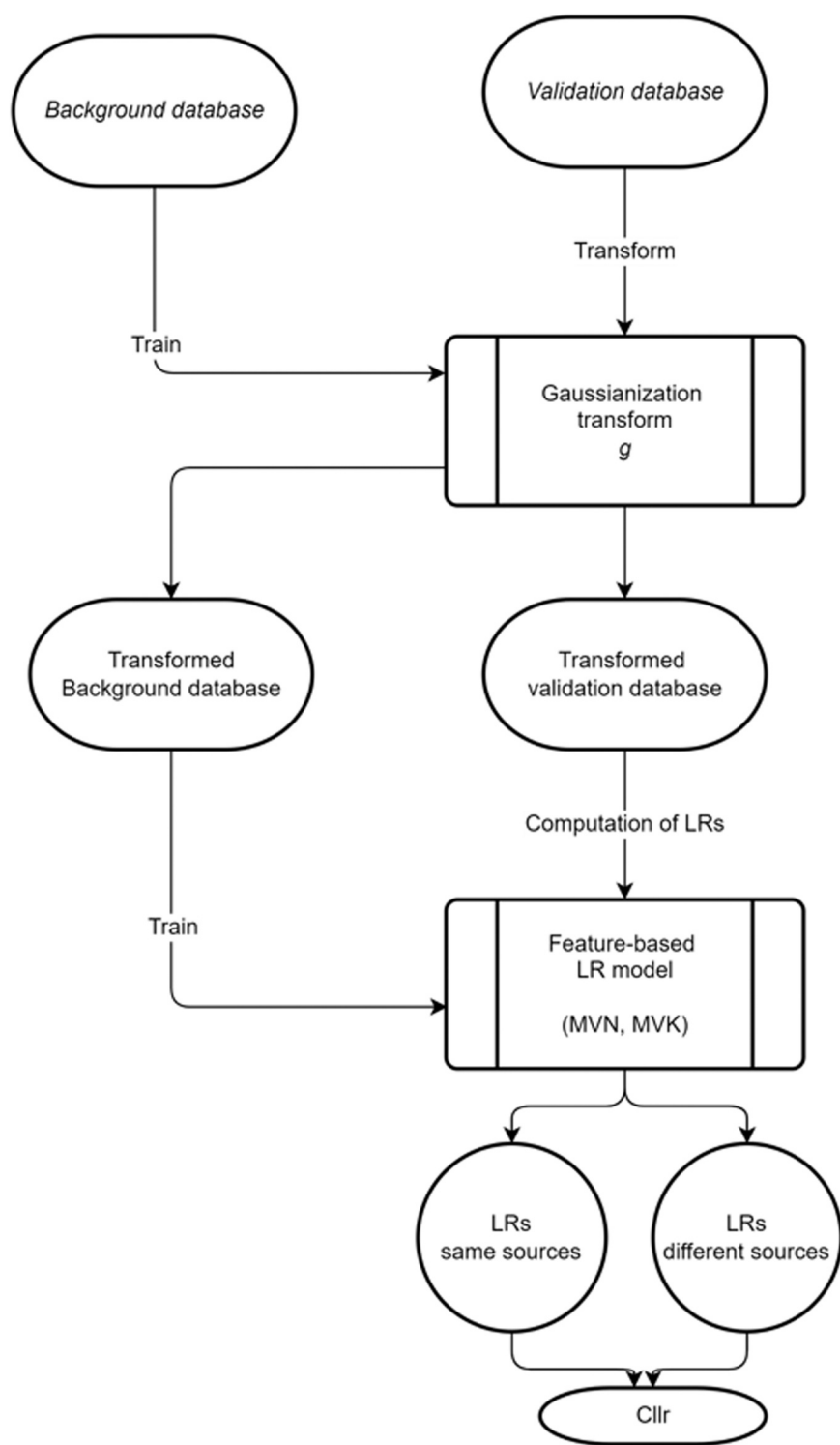


Fig. 11. Evaluation procedure given a Background dataset and a Validation dataset.

Table 1
Summary of the experiments performed.

Experiment name	Background Database	Validation Database	n_p	n_d
Experiment 1	BJA	FIU	4.200	1.097.250
Experiment 2	FIU	BJA	385	295.680
Experiment 3	BJA	BJA	4.200	105.000
Experiment 4	FIU	FIU	385	28.880

account. First, a likelihood ratio is computed for each unique pair of glass fragments available in the respective validation dataset. This gives rise to the number of same-source (n_p) and different-source (n_d) comparisons specified in the table. In the case of validation with the FIU database, these comparisons have been performed for inner and outer panes independently. Finally, aiming at achieving an honest evaluation, experiments 3 and 4 have been done with a 10-fold cross validation procedure. In such a way, no testing glass

Table 2

Transformations evaluated defined by their Gaussianization technique and strategy.

Name	Gaussianization scheme	Gaussianization technique
Hist-match	Feature space	Histogram match
Hist-match _{BTW}	Between-source variability	Histogram match
Hist-match _{BTW W--WTH T H}	Between-source and within-source variabilities	Histogram match
yeo-johnson	Feature space	yeo-johnson
yeo-johnson _{BTW}	Between-source variability	yeo-johnson
yeo-johnson _{BTW W--WTH T H}	Between-source and within-source variabilities	yeo-johnson
GF	Feature space	Gaussianization Flows
GF _{BTW}	Between-source variability	Gaussianization Flows
GF _{BTW W--WTH T H}	Between-source and within-source variabilities	Gaussianization Flows

fragment has been used for training, neither the models nor the Gaussianization transformations.

In any case, the performance of the feature-based LR models with the pre-processed LA-ICP-MS features is measured. Regarding this pre-processing step, we have evaluated the impact in the final C_{lr} of applying the transformations summarized in Table 2.

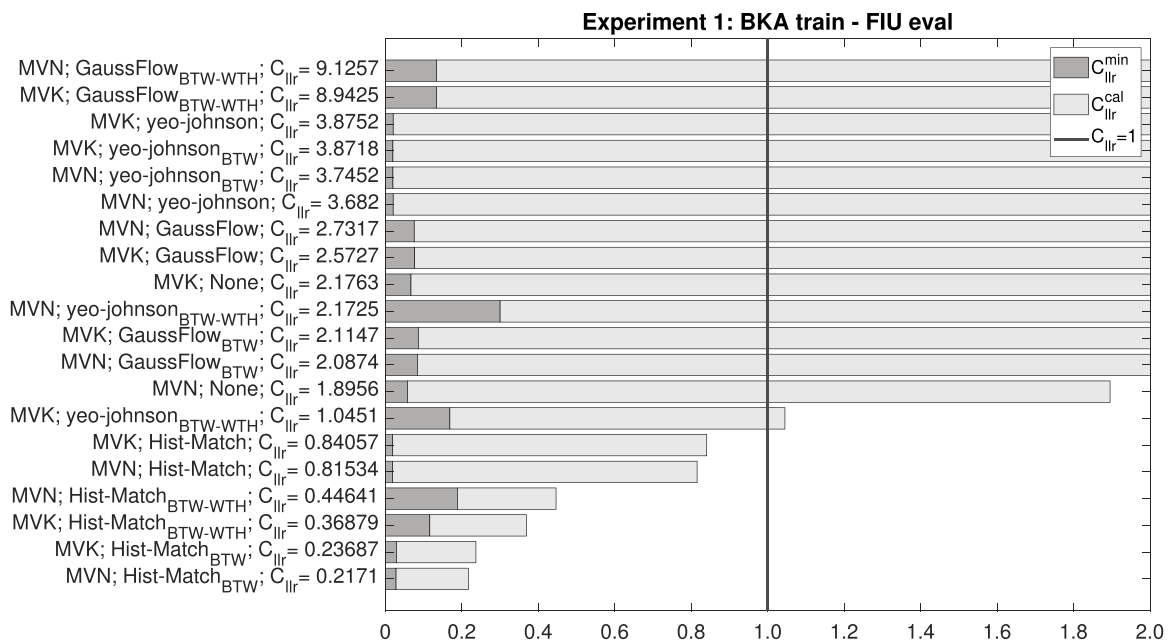
These transformations are defined by the application of the different Gaussianization techniques (described in section 3.2) following all the Gaussianization schemes (see section 3.3).

5. Results

Figures 12, 13, 14 to 15 represent the C_{lr} of the experiments described in section 4.3. To discuss the calibration and discriminating power of the different approaches, the C_{lr} is decomposed into C_{lr}^{cal} and C_{lr}^{min} . The different methods of LR computation are labeled by specifying: first, the name of the LR model; and second, the name of the transformation applied to the LA-ICP-MS features. The transformation names are defined in Table 2. Additionally, results obtained without any pre-processing step are also included (labeled as *None* in the graphs). These approaches are considered the *Baseline* to be improved.

By the analysis of these representations, some conclusions can be extracted. First, the computation of LR directly from the raw LA-ICP-MS features generally gets C_{lr} values greater than 1.0. Thus, Baseline approaches should not be used in real casework as they do not provide any useful information.

Results show that the application of some of the Gaussianization-based pre-processing techniques causes a degradation of the baseline results. For example, in Fig. 12, the Gaussianization of the between-source and within-source variabilities with Gaussianization flows results in a C_{lr} almost five times higher than the one obtained by the same model with the raw features. Additionally, we note that there are transformations that, although they improve the baseline results, do not imply a significant enough enhancement to be used in practice. As a matter of example, in Fig. 12, the Yeo-Johnson Gaussianization of the between-source and within-source variabilities results in C_{lr} values greater than 1.0. Within the transformations that give rise to approaches not usable in real casework, we generally find those based on Yeo-Johnson Gaussianization and Gaussianization Flows. This can be attributed to the existing trade-off between complexity and generalization power of the transformations. As it was mentioned in section 3.2.2, the Yeo-Johnson function is not complex or expressive enough to transform the data distribution into a Gaussian. In contrast, the high number of trainable parameters

**Fig. 12.** Results experiment BKA train, FIU eval.

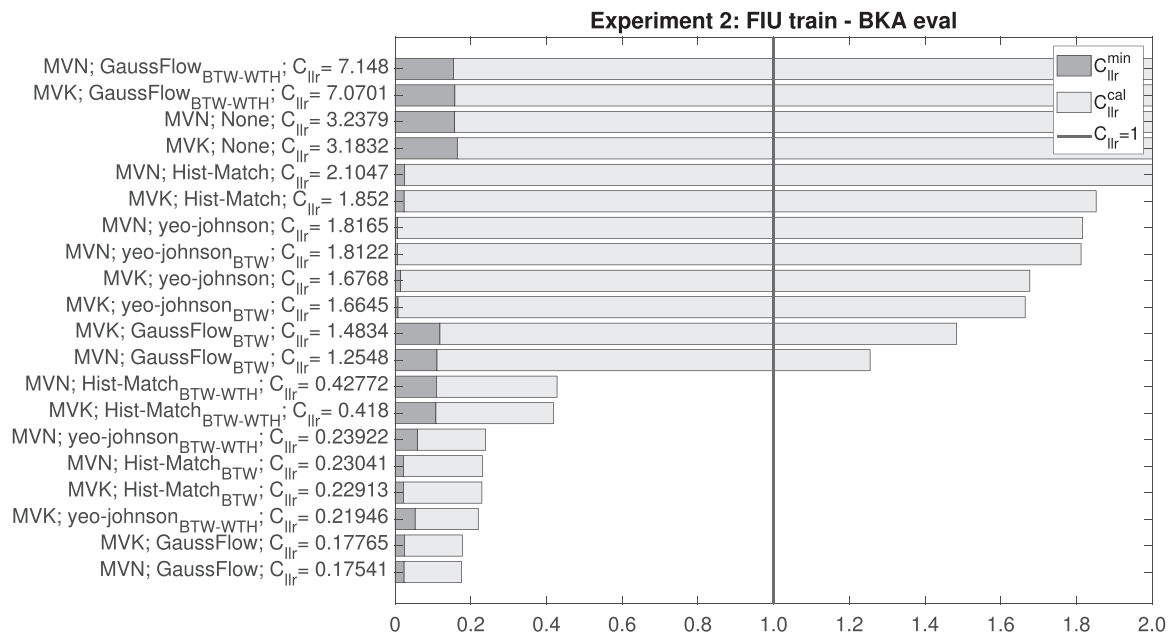


Fig. 13. Results experiment FIU train, BKA eval.

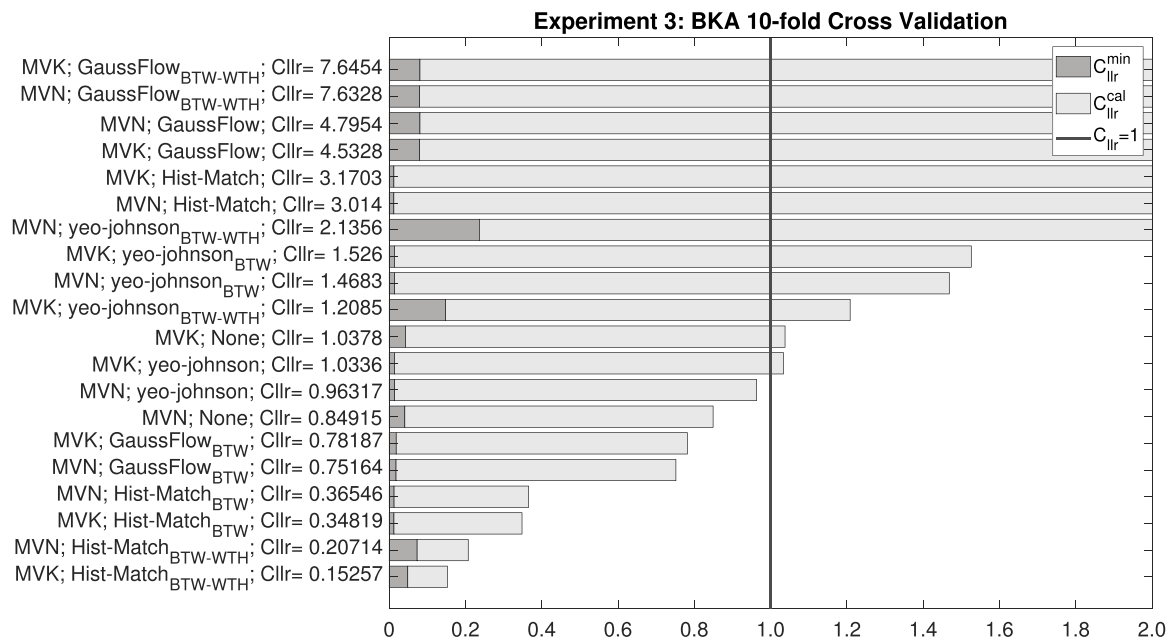


Fig. 14. Results experiment BKA train, 10-fold Cross Validation with BKA.

of the Gaussianization Flows make them a very expressive transformation, which may overfit to the Background data. This limits the generalization power of the transform, hence the capability of correctly Gaussianize new data from the Validation set.

On the other hand, we also report an outstanding improvement in the performance of the feature-based LR models by applying certain transformations over the data. The most consistent results are obtained with a Histogram matching Gaussianization of the Between-source variability (Hist-match_{BTW}); and a Histogram matching Gaussianization of the Between-source and Within-source variabilities (Hist-match_{BTW-WTH}). The computation of the LR models with these

pre-processed features improves the performance of the feature-based LR models in all the experiments, and obtains, in any case, C_{llr} values lower than 1.0. We can therefore conclude that approaches based on these transformations are more robust to dataset shifts.

As it can be seen, this improvement in the C_{llr} is mainly caused by an improvement in the calibration of the final LR models. However, in the case of approaches with a Histogram matching Gaussianization of the between-source variability, we also report a slight improvement in the discrimination power in almost all the experiments. For this reason, we consider this transformation the most optimal within the ones explored all along this work.

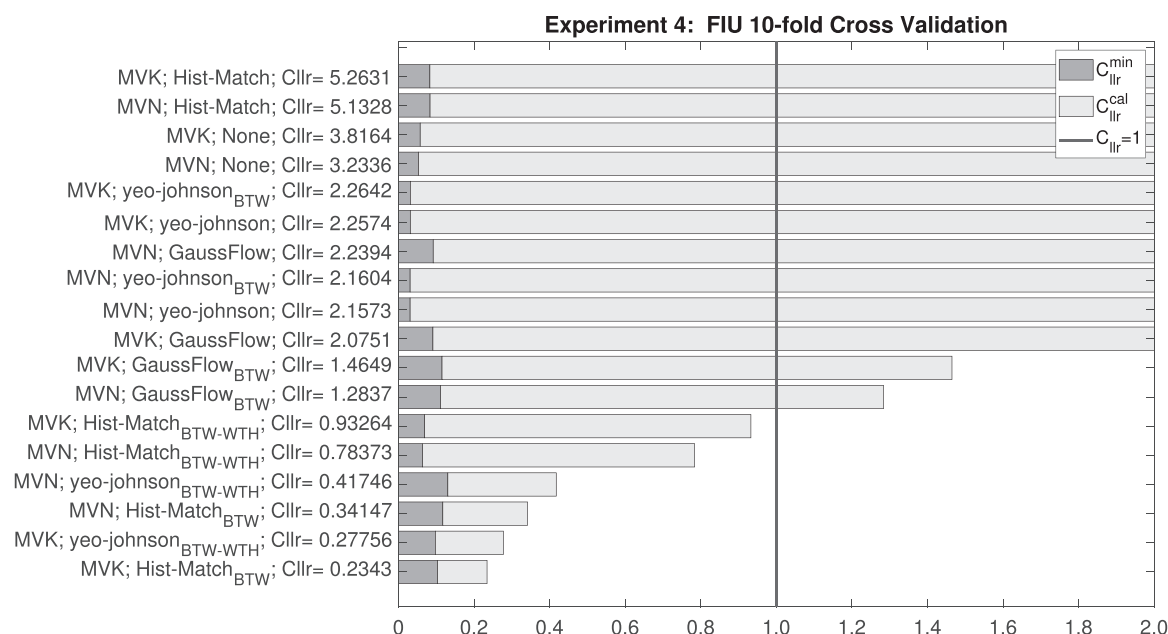


Fig. 15. Results experiment 10-fold Cross Validation with FIU.

Table 3

Computation time for training and evaluating the Gaussianization techniques and the feature-based Likelihood Ratio models. For the Gaussianization techniques, both the training time and the evaluating time has been measured with a dataset of 100 glass fragments with 2 replicates. For the LR models, the training time has been measured with a dataset of 100 glass fragments with 2 replicates, and the evaluating time corresponds to the time needed for the computation of 100 glass comparisons.

Process	Technique	Train time (s)	Eval time (s)
Gaussianization	Hist-match	4.41	0.013
Gaussianization	Hist-match _{BTW}	3.75	0.015
Gaussianization	Hist-match _{BTW W--WTH}	8.37	0.040
Gaussianization	yeo-johnson	5.18	0.004
Gaussianization	yeo-johnson _{BTW}	3.62	0.005
Gaussianization	yeo-johnson _{BTW W--WTH}	8.45	0.014
Gaussianization	GF	1794.60	2.187
Gaussianization	GF _{BTW}	1525.41	2.136
Gaussianization	GF _{BTW W--WTH}	3194.38	6.295
LR model	MVN	0.006	0.178
LR model	MVK	0.006	0.325

5.1. Implementation details

Both Gaussianization techniques and models in this paper have been implemented in Python 3.9.1, using Pythorch 1.7.1, and Sklearn (version 1.0.2). All the experiments have been tested with an Intel Core i7-8750 H (2.20–2.21 GHz) CPU, with 16 GB RAM, and an NVIDIA GEFORCE GTX 1050 Ti GPU. Table 3 shows the average computation time required for training and evaluating the Gaussianization techniques and the LR models of this work. The Gaussianization techniques are referred using the notation described in Table 2.

6. Conclusions

In this work, we have explored different Gaussianization transformations and strategies for feature-based forensic glass comparison using LRs. In this way, it was expected that the Gaussianized measurements would better fit the Gaussian assumptions of these models, hence fostering an improvement of the reliability of the computed LRs.

In light of the results, a Histogram matching-based Gaussianization of the features between-source variability gives rise to an outstanding improvement of the calibration and a slight enhancement of the discrimination power of the models. Thus, we consider that the computation of LRs with these classical models, from the transformed features, presents several advantages. First, this approach produces well-calibrated Likelihood Ratios without employing a further score-based calibration stage, which solves the calibration problem of the feature-based models. Second, it represents a very intuitive, simple, and interpretable solution, since it does not require modeling with very complex distributions of the between-source and within-source variabilities of the features. Third, the low complexity of this approach makes it computationally efficient, since the transformation does not need additional data to be trained, and the models are described with closed-form solutions. And finally, this approach can be easily extrapolated to other forensic disciplines.

It is worth noting that, as it was proven in [14], an efficient incorporation of the uncertainty into the probabilistic models is needed to improve the calibration and avoid unstable Likelihood Ratios. Despite the Bayesian perspective of these models, some of their parameters are point-estimated. The idea of extending these models to *Fully-Bayesian* approaches will be explored in future work.

CRedit authorship contribution statement

Pablo Ramirez-Hereza: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Visualization, Investigation. **Daniel Ramos:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, supervision. **Juan Maroñas:** Formal analysis, Writing – review & editing. **Sergio A. Balanya:** Software, Formal analysis, Writing – review & editing. **Jose Almirall:** Data curation, Resources, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported by the Spanish *Ministerio de Ciencia e Innovación* through grant PID2021-125943OB-I00.

References

- [1] C.G.G. Aitken, F. Taroni, *Statistics and the Evaluation of the Evidence for Forensic Scientists*, third ed., John Wiley and Sons, 2020.
- [2] G. Zadora, A. Martyna, D. Ramos, C. Aitken, *Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data*, John Wiley and Sons., 2014.
- [3] T. Trejos, R. Koons, P. Weis, S. Becker, T. Berman, C. Dalpe, M. Duecking, J. Buscaglia, T. Eckert-Lumsdon, T. Ernst, et al., Forensic analysis of glass by μ -XRF, SN-ICP-MS, LA-ICP-MS and LA-ICP-OES: evaluation of the performance of different criteria for comparing elemental composition, *J. Anal. At. Spectrom.* 28 (8) (2013) 1270–1282, <https://doi.org/10.1039/C3JA50128K>
- [4] P. Weis, M. Dücking, P. Watzke, S. Menges, S. Becker, Establishing a match criterion in forensic comparison analysis of float glass using laser ablation inductively coupled plasma mass spectrometry, *J. Anal. At. Spectrom.* 26 (6) (2011) 1273–1284, <https://doi.org/10.1039/C0JA00168F>
- [5] T. Trejos, S. Montero, J.R. Almirall, Analysis and comparison of glass fragments by laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) and ICP-MS, *Anal. Bioanal. Chem.* 376 (8) (2003) 1255–1264, <https://doi.org/10.1007/s00216-003-1968-0>
- [6] R. Corzo, T. Hoffman, P. Weis, J. Franco-Pedroso, D. Ramos, J. Almirall, The use of LA-ICP-MS databases to calculate likelihood ratios for the forensic analysis of glass evidence, *Talanta* 186 (2018) 655–661, <https://doi.org/10.1016/j.talanta.2018.02.027>
- [7] D. Ramos, D. Meuwly, R. Haraksim, C.E. Berger, Validation of forensic automatic likelihood ratio methods, *Handbook of Forensic Statistics*, Chapman and Hall/CRC, 2020, pp. 143–162.
- [8] C.G.G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, With corrigendum 665–666, *Appl. Stat.* 53 (2004) 109–122, <https://doi.org/10.1046/j.0035-9254.2003.05271.x>
- [9] A. Bolck, H. Ni, M. Lopatka, Evaluating score-and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison, *Law, Probab. Risk* 14 (3) (2015) 243–266, <https://doi.org/10.1093/lpr/mgv009>
- [10] G.S. Morrison, E. Enzinger, D. Ramos, J. Gonzalez-Rodriguez, A. Lozano-Diez, *Handbook of Forensic Statistics*, Chapman and Hall/CRC, 2020, pp. 113–120 (Ch. *Statistical models in forensic voice comparison*).
- [11] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-Garcia, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, *IEEE Trans. Audio, Speech Lang. Process.* 15 (7) (2007) 2072–2084, <https://doi.org/10.1109/TASL.2007.902747>
- [12] R. Haraksim, D. Ramos, D. Meuwly, Validation of likelihood ratio methods for forensic evidence evaluation handling multimodal score distributions, *IET Biom.* 6 (2) (2017) 61–69, <https://doi.org/10.1049/iet-bmt.2015.0059>
- [13] D. Ramos, R.P. Krish, J. Fierrez, D. Meuwly, From Biometric Scores to Forensic Likelihood Ratios, book Edition, Springer, 2017, pp. 305–327, https://doi.org/10.1007/978-3-319-50673-9_14 Ch. 14.
- [14] D. Ramos, J. Maroñas, J. Almirall, Improving calibration of forensic glass comparisons by considering uncertainty in feature-based elemental data, *Chemom. Intell. Lab. Syst.* 217 (2021) 104399, <https://doi.org/10.1016/j.chemolab.2021.104399>
- [15] D.V. Lindley, A problem in forensic science, *Biometrika* 64 (2) (1977) 207–213, <https://doi.org/10.1093/biomet/64.2.207>
- [16] A. van Es, W. Wiarda, M. Hordijk, I. Alberink, P. Vergeer, Implementation and assessment of a likelihood ratio approach for the evaluation of LA-ICP-MS evidence in forensic glass analysis, *Sci. Justice* 57 (3) (2017) 181–192, <https://doi.org/10.1016/j.scijus.2017.03.002>
- [17] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.
- [18] R.M. Sakia, The Box-Cox transformation technique: a review, *J. R. Stat. Soc.: Ser. D.* 41 (2) (1992) 169–178, <https://doi.org/10.2307/2348250>
- [19] S. Weisberg, Yeo-Johnson power transformations, Department of Applied Statistics, University of Minnesota. Retrieved June 1, 2001, 2003.
- [20] V. Laparra, G. Camps-Valls, J. Malo, Iterative gaussianization: from ICA to random rotations, *IEEE Trans. Neural Netw.* 22 (4) (2011) 537–549, <https://doi.org/10.1109/TNN.2011.2106511>
- [21] G. Papamakarios, E.T. Nalisnick, D.J. Rezende, S. Mohamed, B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, *J. Mach. Learn. Res.* 22 (57) (2021) 1–64.
- [22] L. Dinh, J. Sohl-Dickstein, S. Bengio, Density estimation using Real NVP, arXiv preprint arXiv:1605.08803 (2016). [10.48550/arXiv.1605.08803](https://arxiv.org/abs/1605.08803).
- [23] C. Meng, Y. Song, J. Song, S. Ermon, Gaussianization flows, in: International Conference on Artificial Intelligence and Statistics, AISTATS, 2020, 4336–4345.
- [24] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- [25] L. Weng, Flow-based deep generative models, (<https://lilianweng.github.io/posts/2018-10-13-flow-models/>), 2019.
- [26] J.M. Tomczak, M. Welling, Improving variational autoencoders using householder flow, 2016, in: Advances in Neural Information Processing Systems, Workshop on Bayesian Deep Learning, Vol. 1, 2017, 5.
- [27] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, J. Gonzalez-Rodriguez, Deconstructing cross-entropy for probabilistic binary classifiers, *Entropy* 20 (3) (2018) 208, <https://doi.org/10.3390/e20030208>