Improving Calibration of Forensic Glass Comparisons by Considering Uncertainty in Feature-Based Elemental Data

Daniel Ramos[a,1,*], Juan Maroñas[b,a,1], Jose Almirall[c]

[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] *Pattern Recognition and Human Language Tecnology (PRHLT) Research Center. Escuela Técnica Superior de Ingeniería Informática. Universitat Politècnica de València. Camino de Vera, s/n. 46022 Valencia. Spain.*
[c] *Center for Advanced Research in Forensic Science, and Department of Chemistry and Biochemistry, Florida International University.*

Abstract

The computation of likelihood ratios (LR) to measure the weight of forensic glass evidence with LA-ICP-MS data directly in the feature space without computing any kind of score as an intermediate step is a complex problem. A probabilistic two-level modeling of the within-source and between-source variability of the glass samples is needed in order to compare the elemental profiles measured from glass recovered from a suspect or a crime scene and compared to glass samples of a known source of origin. Calibration of the likelihood ratios generated using previously reported models is essential to the realistic reporting of the value of the glass evidence comparisons.

We propose models that outperform previously proposed feature-based LR models, in particular by improving the calibration of the computed LRs. We assume that the within-source variability is heavy-tailed, in order to incorporate uncertainty when the available data is scarce, as it typically happens in forensic glass comparison. Moreover, we address the complexity of the between-source

[*] Corresponding author
Email address: daniel.ramos@uam.es (Daniel Ramos)
URL: http://audias.ii.uam.es (Daniel Ramos)
[1] Equal contribution.

variability by the use of probabilistic machine learning algorithms, namely a variational autoencoder and a warped Gaussian mixture. Our results show that the overall performance of the likelihood ratios generated by our model is superior to classical approaches, and that this improvement is due to a dramatic improvement in the calibration despite some loss in discriminating power. Moreover, the robustness of the calibration of our proposal is remarkable.

Keywords: Likelihood ratio, forensic glass comparison, LA-ICP-MS, variational autoencoder, warped Gaussian mixture, heavy-tailed

1. Introduction

The aim of a comparison between items within the forensic context is to obtain information about the linkage between a suspect and a crime event by the use of scientific procedures to analyze and evaluate the evidence. In this context, a fact finder (e.g., a judge or a jury) must make the final decision on this issue, but the responsibility of the examination of the physical evidence and subsequent interpretation of the data generated falls on a forensic examiner. As an example, if the forensic case is a hit-and-run car accident, the glass in the clothes of the victim can be compared to the glass that originated from the windshield of a car suspected as involved in the accident.

A likelihood ratio (LR) approach has been proposed as the logical way of evaluating forensic evidence [1, 2, 3, 4, 5]. The LR is the ratio of the probability of the evidence given each of two mutually exclusive propositions in the case. According to [6, 7], a validation process is recommended prior to the use of a LR model in casework. In this work, we will follow this recommendation, and we measure performance with a proper scoring rule (in our case, by the use of the so-called likelihood ratio cost C_{llr})[8], that can be further analyzed as the additive contribution of the discriminating power and the calibration of LR values. The former measures the ability of the compared features (elemental profiles) to distinguish between cases where both sets of features belong to the same source, and cases where they belong to different sources. The latter measures how reliable are the conclusions that a decision maker can make with those LR values [9, 10].

The incorporation of good analytical data derived from laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) analyses have enabled the use of statistical tools for the comparison and objective evaluation of glass evidence. Score-based LR approaches have been specially successful [11, 12, 13, 14, 15, 16]. A *score* is firstly computed from the features [17], and a calibration step can transform it in a LR based on the size and composition of a background population of glass samples. The score-based approaches typically present good calibration properties despite the scarcity of the training data [18, 19]. These score-based LR methods have been successfully applied in many other forensic disciplines [20, 21, 22, 23, 24, 25, 15, 14].

One problem of score-based LR models relies upon the need of several databases. The probabilsitic model to transform the score into a LR requires a database to be trained. Moreover, it has been shown in [16] that the score computation process must take into account the typicality of the features, otherwise important information may be lost. As a matter of an example, scores based only on a distance between the features are shown to be inadequate. The reality of operational forensic laboratories is that most background databases are relatively small in size ranging from 200-500 distinct glass samples representing the local population of glass sources relevant to the particular jurisdiction or population of sources. Data scarcity not only impacts the performance of models, but also the evaluation protocol. Typically, the available data must be split into training, validation, and test datasets. If the data are scarce, that split will yield too small datasets, unable to guarantee generalization to other datasets. Notice that here we split the data into training, validation, and testing datasets as it is typically done in chemometrics, pattern recognition and machine learning [26]; instead of two sets of training and forensic validation as in previous work in forensic science [6]. This is a matter of terminology[2]. Cross-validation techniques

[2] In forensic science, the training and validation datasets are referred to as *training* or

[26] help on the optimal usage of data, but its intensive use might lead to an unrealistic match between data splits, leading to overoptimistic results. The need of more databases to train the models, as it happens in score-based approaches, exacerbates the problem. These data requirements also appear when using principal component analysis (PCA) [27] or other dimensionality reduction techniques.

Another option to score-based LR computation is the use of so-called feature-based models [1, 17, 13], which directly assigns probabilities to the features and not to a pre-computed score. Unfortunately, to date there has not been feature-based LR models that present reasonable calibration for the problem of forensic glass comparison using LA-ICP-MS features. Therefore, research in these models is relevant and important.

Moreover, there is additional interest to conduct research in feature-based LR computation: if good calibration is obtained with LA-ICP-MS features using some probabilistic models, these models could be potentially usable in other forensic problems and features beyond glass comparison. In this sense, probabilistic machine learning may offer powerful algorithms to help with this objective [28].

Despite the potential advantages of feature-based models, their complexity with respect to score-based LR methods is significantly higher. The problem of higher dimension increases the need of additional data. Moreover, the feature space is fairly complex, requiring models with a high number of free parameters, and even though non-parametric approaches. All these alternatives are known to be also highly sensitive to the lack of data, which aggravates the data scarcity problem. As a consequence, previously proposed feature-based models present bad calibration in high dimensions [17, 29], leading to unreliable models that cannot be used in practice for LA-ICP-MS features. In this sense, other authors

development dataset; and the testing dataset is referred to as the *validation* dataset [6]. As this might create confusion to readers more used to the chemometrics field, we decided to stick to the more accepted machine learning terminology.

have advocated to compute the weight of the evidence with alternatives to the likelihood ratio with good results, as it is proposed in [30], but we believe that research must still be conducted to reliably compute feature-based likelihood ratios.

In this work, we propose the use of two feature-based models that dramatically improve the performance of likelihood ratios as compared to previous feature-based approaches. This performance boost is mainly based on the improvement on the calibration. Our proposal is based on the classical two-level model presented in [17], with two main important differences. First, the within-source variability of the features is not modeled by a multivariate Gaussian, but by a Student's t-distribution. This allows the incorporation of uncertainty in the model by the selection of the number of degrees of freedom of the Student's t-distribution. Second, the between-source variability is modeled by two different probabilistic machine learning approaches: On the one hand, a variational autoencoder [31]; and on the other hand, a warped Gaussian mixture [32]. These approaches enable flexible densities, and direct Monte Carlo sampling to numerically approximate likelihood ratio. Our results show that the proposed solution outperforms the baseline model and is better calibrated. Furthermore, we will use relevant LA-ICP-MS databases from the Florida International University, and from the German Federal Police (*Bundeskriminalamt*) as background population.

This article is organized as follows. First, the likelihood ratio framework is described in Section 2. Next, Section 3 describes the LR models proposed in this work, including the baseline used. The experimental section is described in Section 4. Finally, conclusions are drawn in Section 5.

2. The likelihood ratio framework in forensic science

In this work, we follow the likelihood ratio framework, which has been proposed as a logical way of evaluating the evidence in a forensic case [5]. In forensic glass comparison using LA-ICP-MS, the evidence are the glass samples to be compared. For instance, in a hit-and-run car accident, one of the samples would be recovered from the crime scene or from the clothes of a victim, and the other sample would be known to originate from the windshield of a suspect car.

Some feature vectors are measured from each of the samples. From the crime-scene sample, we obtain a total of n_1 replicates, namely $\mathbf{Y}_1 = \left(\mathbf{y}_1^{(1)}, \cdots, \mathbf{y}_1^{(n_1)}\right)$, where each measurement is denoted as $\mathbf{y}_1^{(i)} = \left(y_{1,1}^{(i)}, \cdots, y_{1,D}^{(i)}\right)^T$. For LA-ICP-MS, the dimension of the feature vectors is $D = 17$. Analogously, for the suspected sample, some feature vectors are measured, for a total of n_2 replicates, namely $\mathbf{Y}_2 = \left(\mathbf{y}_2^{(1)}, \cdots, \mathbf{y}_2^{(n_2)}\right)$, where each measurement is denoted as $\mathbf{y}_2^{(i)} = \left(y_{2,1}^{(i)}, \cdots, y_{2,D}^{(i)}\right)^T$.

Both sets of vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are the feature-based representation of the evidence. Within the likelihood ratio framework, we evaluate these features in the context of two competing *propositions* or *hypotheses*. Let H be the so-called *proposition* variable, that represents what is the true proposition that happens in the case (also known as *the ground-truth* proposition, or *the true label*). This is a categorical variable with alphabet $\{h_{\text{ss}}, h_{\text{ds}}\}$, with possible values as follows:

- h_{ss} is the so-called *prosecution proposition*, or in our case, the *same-source* proposition, because a prosecutor typically contributes to relate the suspect to the crime. For simplicity, we will define h_{ss} as follows: *The crime-scene sample and the suspect samples come from the same source.*

- h_{ds} is the so-called *defense proposition*, or in our case, the *different-source* proposition, because the defense typically contributes to demonstrate that the suspect is unrelated to the crime. For simplicity, h_{ds} will be: *The crime-scene sample and the suspect samples come from different sources.*

Other definitions of the propositions are possible. In particular, we defined the propositions for a so-called common-source scenario, but a definition of the propositions for a specific-source scenario would also be possible. See [33] for a discussion on this issue.

To compute the LR, the forensic scientist should have a so-called *background* database, equivalent to the pooling of the training and validation databases if we use the typical terminology in pattern recognition and machine learning (as we do in this work). This background database contains glass features of many other glass objects (sources). We will assume that those sources are *representative* of the population from which the findings in the case are drawn. This means that if that background database was drawn from a population, the forensic findings were also drawn from the same population.

For the sake of simplicity, we will assume that the number of replicates of each object in the background database is the same, namely n_b. Let m be the number of objects in the dataset. Then, $\mathbf{X}_i = \left(\mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(n_b)}\right)$ will be the matrix with all the replicated measurements from object i in the background database, with $\mathbf{x}_i^{(j)} = \left(x_{i,1}^{(j)}, \cdots, x_{i,D}^{(j)}\right)^T$. Thus, $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_m)$ will be the complete background dataset of features. Note that we have organized the notation in a way in which it is implicitly assumed that the true labels of each object (source) in the background database are known. Finally, we organize the whole set of data observations as $\mathcal{D} = (\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{X})$.

The LR framework is summarized with the odds-form of the Bayes theorem:

$$\frac{P\left(H = h_{\mathrm{ss}} \middle| \mathcal{D}\right)}{P\left(H = h_{\mathrm{ds}} \middle| \mathcal{D}\right)} = \frac{P\left(\mathcal{D} \middle| H = h_{\mathrm{ss}}\right)}{P\left(\mathcal{D} \middle| H = h_{\mathrm{ds}}\right)} \times \frac{P\left(H = h_{\mathrm{ss}}\right)}{P\left(H = h_{\mathrm{ds}}\right)}, \tag{1}$$

Thus, the posterior probabilities $P\left(H = h \middle| \mathcal{D}\right)$ with $h \in \{h_{\mathrm{ss}}, h_{\mathrm{ds}}\}$ are used to make decisions. The prior probabilities $P\left(H = h\right)$, with $h \in \{h_{\mathrm{ss}}, h_{\mathrm{ds}}\}$, summarize all the knowledge about the propositions in the case *before* the analysis of the evidence has been conducted. This might include information such as police investigative reports, witnesses, past criminal records of the suspect, other evidence (e.g., a recorded conversation, a psychological report), and so on. The prior probabilities are not the province of the forensic examiner.

These prior probabilities are updated by the evidence evaluation, the province of the forensic examiner. The factor that transforms the prior odds into the posterior odds is the likelihood ratio (LR):

$$LR = \frac{P\left(\mathcal{D}\middle|\, H = h_{\text{ss}}\right)}{P\left(\mathcal{D}\middle|\, H = h_{\text{ds}}\right)}, \tag{2}$$

If we simplify notation $H = h_{\text{ss}}$ into h_{ss} inside probabilities, we have:

$$LR = \frac{P\left(\mathcal{D}\middle|\, h_{\text{ss}}\right)}{P\left(\mathcal{D}\middle|\, h_{\text{ds}}\right)} = \frac{P\left(\mathbf{Y}_1, \mathbf{Y}_2\middle|\, h_{\text{ss}}, \mathbf{X}\right)}{P\left(\mathbf{Y}_1, \mathbf{Y}_2\middle|\, h_{\text{ds}}, \mathbf{X}\right)}, \tag{3}$$

because $P\left(\mathbf{X}\middle|\, h_{\text{ss}}\right) = P\left(\mathbf{X}\middle|\, h_{\text{ds}}\right)$, i.e., the background data are independent of the relationship between the crime-scene sample and the suspect sample.

Thus, in this work, we will explore models to compute the LR as in Eq. 3.

3. Likelihood Ratio models

In this section, we describe the models that will be used to compute a feature-based LR. We start with the description of the classical Bayesian inference approach to compute likelihood ratios in forensic science with continuous data. We then present our baseline system, widely known and proposed in [17]. Finally, our two proposed methods are presented.

3.1. Classical Bayesian model for LR computation

The classical Bayesian model for trace evidence evaluation, firstly applied in univariate glass data in [1] and extended to multivariate glass data in [17], considers the following general expression for the LR:

$$\begin{aligned} &\tfrac{P(\mathbf{Y}_1,\mathbf{Y}_2|h_{\text{ss}},\mathbf{X})}{P(\mathbf{Y}_1,\mathbf{Y}_2|h_{\text{ds}},\mathbf{X})} \\ &= \tfrac{\int P(\mathbf{Y}_1|\theta)P(\mathbf{Y}_2|\theta)P(\theta|\mathbf{X})d\theta}{\int P(\mathbf{Y}_1|\theta)P(\theta|\mathbf{X})d\theta \int P(\mathbf{Y}_2|\theta)P(\theta|\mathbf{X})d\theta}, \end{aligned} \tag{4}$$

where θ are the parameters of the selected model. We clearly identify the following elements:

- Likelihood $P\left(\mathbf{y}\middle|\,\theta\right)$.
- Parameter distribution $P\left(\theta\middle|\,\mathbf{X}\right)$.

This model is not necessarily *fully-Bayesian*, because sometimes θ contains an incomplete set of the total number of parameters in the model, and the rest of parameters are assigned using point-estimate approaches (*e.g.*, maximum likelihood). In that sense, $P\left(\theta|\, \mathbf{X}\right)$ is not precisely a *posterior* parameter distribution, but some density assigned to the parameters by some optimization method (e.g., maximum likelihood). That is why we will call it *parameter distribution* and not *parameter posterior* as typically happens in Bayesian statistics literature. Moreover, in particular, to date and to our knowledge, the classical models only consider that θ is the mean vector $\boldsymbol{\mu}$ of each glass object feature distribution, whereas the between-source and within-source covariance matrices are assigned with maximum likelihood, as in the cited earliest works [1, 17].

All the models presented in this work (baseline model and proposed models) are based on this classical model. The main difference between them relies upon the selection of the likelihood, the selection of the parameter distribution, and the computation of the integrals in Equation 4 (analytically or by Monte-Carlo sampling).

3.2. Baseline model: ALK model

The model ALK (after *Aitken and Lucy with Kernels*) was proposed in [17], and follows the model in Equation 4, considering the following:

- Likelihood $P\left(\mathbf{y}|\, \boldsymbol{\mu}\right) \sim \mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_w\right)$, where $\hat{\boldsymbol{\Sigma}}_w$, the within-source variability covariance matrix, is computed by maximum likelihood.
- Parameter distribution $P\left(\boldsymbol{\mu}|\, \mathbf{X}, k\hat{\boldsymbol{\Sigma}}_b\right)$, a non-parametric kernel density function (KDF) with Gaussian kernels having covariance $k\hat{\boldsymbol{\Sigma}}_b$, where $\hat{\boldsymbol{\Sigma}}_b$ is obtained by maximum likelihood.

The maximum-likelihood statistics involved in this computation are:

- Per-object mean on the background set:

$$\hat{\bar{\mathbf{x}}}_i = \frac{1}{n_b}\sum_{j=1}^{n_b}\mathbf{x}_i^{(j)} \tag{5}$$

- Total mean of the background set:

$$\hat{\hat{\mathbf{x}}} = \frac{1}{m}\sum_{i=1}^{m}\hat{\hat{\mathbf{x}}}_i \tag{6}$$

- Within-object covariance matrix:

$$\hat{\mathbf{\Sigma}}_w = \frac{1}{mn_b - m}\sum_{i=1}^{m}\sum_{j=1}^{n_b}\left(\mathbf{x}_i^{(j)} - \hat{\hat{\mathbf{x}}}_i\right)\left(\mathbf{x}_i^{(j)} - \hat{\hat{\mathbf{x}}}_i\right)^T \tag{7}$$

- Between-object covariance matrix:

$$\hat{\mathbf{\Sigma}}_b = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\hat{\mathbf{x}}}_i - \hat{\hat{\mathbf{x}}}\right)\left(\hat{\hat{\mathbf{x}}}_i - \hat{\hat{\mathbf{x}}}\right)^T - \frac{\hat{\mathbf{\Sigma}}_w}{n_b} \tag{8}$$

Also, the so-called kernel smoothing parameter was computed as in [34]:

$$k = \left(\frac{4}{2D+1}\right)^{\frac{1}{D+4}} m^{-\frac{1}{D+4}} \tag{9}$$

This model assumes that the features of objects are represented by their mean vectors. Thus, the glass samples replicated from one object are generated according to a Gaussian distribution centered in the object mean vector $\mathbf{x}_i$, that is computed from the background set as $\hat{\hat{\mathbf{x}}}_i$. Then, the distribution of object means is represented by the parameter distribution, which is modeled according to a KDF.

The main advantage of this model is that it has a closed-form solution (see [17]). The main disadvantage is that it does not consider much of the uncertainty in the parameters, since it is not fully-Bayesian and assigns maximum-likelihood statistics to several important parameters. Therefore, it is expected that the model will be poorly calibrated. Moreover, in our preliminary experiments it has been observed that, when used with LA-ICP-MS data, the model will often output infinite and zero LR values, which is and indication of overstatement in the weight of the evidence. Thus, to preserve Cromwell's rule, stating that no value of 0 or 1 must be assigned to any probability [35], we will limit the LR values to lay in the range between the maximum and minimum floating-point precision in the computed where the model is used. This way, we avoid numerical and computational pitfalls on the baseline results. Strategies to limit the value of the LR are not new, and have also been previously proposed in [19, 18, 36].

3.3. Proposed LR models

In this section, we briefly propose two alternate models that aim at better incorporating the uncertainty in the glass comparison problem. It is worth noting that in these models we use a multivariate Student's-t distribution to incorporate uncertainty in the Bayesian model, instead of using it for a hypothesis test like in [17].

3.3.1. HW model: Heavy-tailed with Warped Gaussian mixtures

This model assumes the following:

- Likelihood $P\left(\mathbf{y}|\,\boldsymbol{\mu}\right) \sim \mathcal{T}\left(\mathbf{y}|\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_w, \nu\right)$. Here, $\mathcal{T}$ denotes a multivariate Student's t-distribution with within-source variability scale matrix $\hat{\boldsymbol{\Sigma}}_w$ and degrees of freedom ν. Here, $\hat{\boldsymbol{\Sigma}}_w$ is computed with maximum likelihood in the same way as in Section 3.2, and ν is heuristically obtained from a validation set[3].

- Parameter distribution $P\left(\boldsymbol{\mu}|\,\mathbf{X}, \mathcal{W}\right)$, the predictive distribution of a warped Gaussian mixture [32] representing the distribution of the mean vectors of the glass features. The Warped Gaussian Mixture is a fully Bayesian generative model where an infinite Gaussian Mixture model [37] is warped using a Gaussian Process Latent Variable Model [38]. Inference in this model relies on a Hamiltonian within Gibbs Markov Chain Monte Carlo sampler. In this way, the warped Gaussian mixture can adapt to complex parameter distributions while incorporating the uncertainty in $\boldsymbol{\mu}$. Details

[3] As mentioned before, here we use the term *validation dataset* to refer to the dataset used for model selection, in contrast to the *forensic validation dataset* typical in forensic science, whose analogous in machine learning would be the *testing dataset*. The reasons of this change in the terminology have been explained before.

about this model can be found in [32], with indications of a Matlab^{TM} code with an implementation of the model.

As there is no closed-form of the integrals in Equation 4 with the warped Gaussian mixture as the parameter distribution, we compute the LR by Monte Carlo sampling using the following expression:

$$\frac{P\left(\mathbf{Y}_1, \mathbf{Y}_2 \middle| h_{\mathrm{ss}}, \mathbf{X}\right)}{P\left(\mathbf{Y}_1, \mathbf{Y}_2 \middle| h_{\mathrm{ds}}, \mathbf{X}\right)} = \frac{\frac{1}{M}\sum_{i=1}^{M} P\left(\mathbf{Y}_1 \middle| \boldsymbol{\mu}_i\right) P\left(\mathbf{Y}_2 \middle| \boldsymbol{\mu}_i\right)}{\left[\frac{1}{M}\sum_{i=1}^{M} P\left(\mathbf{Y}_1 \middle| \boldsymbol{\mu}_i\right)\right]\left[\frac{1}{M}\sum_{i=1}^{M} P\left(\mathbf{Y}_2 \middle| \boldsymbol{\mu}_i\right)\right]} \tag{10}$$

where M mean vectors are sampled from the model:

$$\boldsymbol{\mu}_i \sim P\left(\boldsymbol{\mu}_i \middle| \mathbf{X}, \mathcal{W}\right). \tag{11}$$

3.3.2. HVAE model: Heavy-tailed with Variational Autoencoder

This model is fairly similar to the HW model, but replacing the warped Gaussian mixture with a variational autoencoder, which also allows sampling from the parameter distribution. In particular:

- Likelihood $P\left(\mathbf{y} \middle| \boldsymbol{\mu}\right) \sim \mathcal{T}\left(\mathbf{y}|\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_w, \nu\right)$. Here, $\mathcal{T}$ denotes a multivariate Student's t-distribution with within-source variability scale matrix $\hat{\boldsymbol{\Sigma}}_w$ and degrees of freedom ν. Here, $\hat{\boldsymbol{\Sigma}}_w$ is computed with maximum likelihood in the same way as in Section 3.2, and ν is heuristically obtained from a validation set.

- Parameter distribution $P\left(\boldsymbol{\mu} \middle| \mathbf{X}, \psi\right)$, a variational autoencoder [31], representing the mean vectors of the glass objects. The variational autoencoder has generative parameters ψ that parameterize a Gaussian likelihood modeling both the mean and the variance (aleatoric uncertainty) of our observations. The generative parameters ψ are learned following the procedure described by [31], i.e. by using amortized inference and stochastic variational inference and using a pathwise gradient estimator. To sample from the variational autoencoder we construct a Markov Chain using the encoder and decoder to form the transition operator, see e.g. [39].

Again, the model uses Equation 10 to approximate the LR by Monte-Carlo using samples from the Markov Chain generated using the variational autoencoder.

4. Experiments

4.1. Databases

Here we describe the two databases used in this work, that together allow a realistic experimental set-up. The databases contain vectors of glass elemental compositions coming from LA-ICP-MS analysis derived from quantitative analysis using a standard LA-ICP-MS method (ASTM E2927-16e1). Two LA-ICP-MS glass databases were used in this study: the Bundeskriminalamt (BKA) casework database (385 glass objects) and the Florida International University (FIU) vehicle database (420 glass objects). The LA-ICP-MS analytical technique implies that each measurement generates $D = 17$ chemical isotopes found in the glass, which each forms a feature vector. Each of the values of each element of the vector represents a number associated with that particular isotope. Although each measurement generates a single vector, several measurements can be performed in the same object, leading to so-called *replicates*, or *replicated measurements* in one single object. The meaning of the numbers in the vectors are chemically relevant, but irrelevant for the purpose of statistical forensic evaluation, and, of course, data transformation or dimensionality reduction techniques are welcome if convenient. What matters is that all vectors will tend to present high mean separation if they come from different sources, and that the within-source variability will be low enough with respect to between-source variability to be able to reach an important discriminating power. Moreover, the LA-ICP-MS technique is quite robust, and a low database shift is expected between different LA-ICP-MS databases, even from different laboratories. These hypothesis have been supported by past studies [18, 40, 41].

The available databases are described as follows:

- **BKA (background database)**. This database from the German Federal Police (Bundeskriminalamt) contains features from float glass in windows analyzed in real forensic cases. The database contains 385 different sources (glass objects). For each source, a total of 6 replicated measurements were taken. The database includes several different types of glass relevant in forensic practice: float glass, container glass, pre-float window glass, etc. The BKA typically analyzes the following 18 elements: ^{7}Li, ^{23}Na, ^{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. However, ^{23}Na was omitted in this study since it is not included in the element menu suggested in the United States standard, ASTM E2927 [41]. LA-ICP-MS analysis was followed by signal processing using the Glitter^{TM} software (MacQuarie University, Australia). The LA-ICP-MS instrumental parameters were set following the relevant literature [42, 43, 41].

- **FIU (testing database)**. This database from the Florida International University considers measurements from car windshields. The database consists of 210 windshields from vehicles at the M & M Service and Salvage Yard in Ruckersville, Virginia. Each windshield consists of inner and outer panes, joint together with some fixing technique. The dataset includes vehicles from 26 different automotive manufacturers, and they were produced between the years 2004 and 2017. LA-ICP-MS was used to analyze all objects and the 17 isotopes listed in ASTM E2927 were quantified using Glitter^{TM}, with the Float Glass Standard 2 (FGS 2) used as the calibrator [44]. The analysis was conducted using a ns-213 nm Nd: YAG laser (ESI New Wave Research, Portland, OR, USA) coupled to a quadrupole ELAN DRC II (Perkin Elmer LAS, Shelton, CT, USA), with the following parameters: 100% laser energy ($\simeq$ 0.65 mJ), 10 Hz, 90 μm spot size, and 60-second dwell. For the purposes of the statistical evaluation, the glass in the inner and in the outer pane can be considered as different objects (because they behave differently from the chemical point of view), and therefore we can account for a total number of 420 objects (sources). However, it is expected that some similarities will be found between inner and outer panes of each windshield, even though are supposed to be different glass objects. Thus, the testing FIU database has been divided into two different subsets, namely 210 *inner* panes and 210 *outer* panes of the considered windshields. Both subsets will be used in an isolated way in the experimental protocol, and the LR values from each of the subsets will be pooled to generate the final testing results. For each of the 420 objects (sources), a total of 15 replicated measurements were taken.

For the purposes of model comparison and benchmarking typical in pattern recognition and machine learning, we will consistently use the BKA database for training and validation and the FIU database for testing. This is because the BKA database, from realistic forensic cases, is more adequate to represent a real population than the FIU database, created in laboratory conditions. This allows the comparison with other works in the literature, where the same experimental set-up has been used, such as [18].

Moreover, in order to isolate data variability issues related to the number of replicates, we will divide the training (BKA) and testing (FIU) sets into groups of $n_r = 3$ replicates for every single object. Thus, the BKA databases will have 5 of these n_r-replicate groups, and the BKA testing database will have 2 of these n_r-replicate groups. This situation of equalizing the number of replicates to compare is realistic, as many forensic laboratories use standardized numbers for n_r in casework comparisons. This can be evidenced from standards for glass comparison using LA-ICP-MS features (see *e.g.* [41]).

4.2. Forensic relevance of databases

A legitimate question is whether the databases presented in this work are representing a realistic and relevant population. We aim at addressing this issue here.

The authors have previously shown [45, 46, 40] that there are observable differences in the LRs calculated based on the background database used for the LR calculation. The differences observed are based on 1) the size of the background database and 2) the composition of the background database. According to this previous work, the differences calculated are small with models with BKA data and with FIU data. Therefore, the main conclusion is that the differences in the type of the data in these two databases are not particularly relevant for LA-ICP-MS features when they are used as population samples.

According to this previous observations, the relevant population in casework is typically defined locally, i.e., from German cases, in the case of the BKA and from US vehicles, in the case of the FIU. Therefore, the population depends on what type of case. What we have found, however, is that any differences found in the LLR calculation will be affected by the size of the database, not necessarily the geographic origin of the glass given that glass is a global commodity and especially, vehicle glass moves around the world freely.

In casework, the recommendation is to use a local *relevant* background database in the laboratory that serves as a sample of that *population.* For example, a locally collected database in the US should be used for US cases and a locally collected database in Australia should be used for Australia-originated casework, etc. There are now 6-7 locally produced databases in The Netherlands, Germany (two available), the US (two available), Singapore and Australia. Those laboratories that have access to a local database are recommended to use it. What we have found, however, is that the size of the database plays a more important role in the Log-LR calculation than where it was generated [45, 46, 40].

According to this discussion, we can define database selection in a typical forensic case. The most likely type of glass forensic case in the US is a hit-and-run accident where the broken windshield is involved in the glass collected from the crime event. The windshield glass transfers to both any victim of a hit-and-run and, we have found also to any driver or passenger in the vehicle. These types of transfers often answer the forensic questions of *activity* along with source comparisons. While other types of vehicle glass (rear window and side windows) also can break, of course, they are less likely to break and, would provide similar information, were they to make up the background database samples, we have found.

4.3. Data pre-processing

LA-ICP-MS data present a very different behavior for each of the dimensions of the feature vector, since each of them represents the concentrations of different chemical elements. Moreover, the data presents some outliers. Therefore, to facilitate the modeling, we have normalized the data using a two-step process. First, each of the marginal histograms of the logarithm of each variable has been normalized to their interquartile distance, and their median has been subtracted. This will transform all the marginal distributions of the variables ideally to zero-median and one interquartile distance. Then, a standard sigmoid has been applied to each of the variables, to reduce the influence of the outliers. As a result, the data to be used in these experiments will consist of so-called normalized feature vectors, which in previous experiments have demonstrated to be much more adequate to the proposed models. A possible drawback of this normalization strategy is that the rarity of the features can be reduced, which might result on a discrimination loss, but it is expected to favor a calibration improvement. As the main problem of feature-based approaches is the calibration, we decided to follow this normalization procedure as a proof-of-concept that the calibration can be dramatically improved at the feature level. Moreover, in our previous experiments, other normalization schemes seemed to yield worse calibration results precisely due to the outliers. The exploration of those effects in the normalization of glass features is one of our future avenues to improve feature-based models.

The normalization process of the training split of the BKA database has been performed via a leave-one-object-out cross-validation procedure. Then, the statistics on the raw, unnormalized training BKA data (interquartile distance and median) are used to normalize the validation BKA data and the test FIU data. Thus, the data usage is honest and does not contaminate results, as all vectors are normalized with a dataset that does not contain themselves.

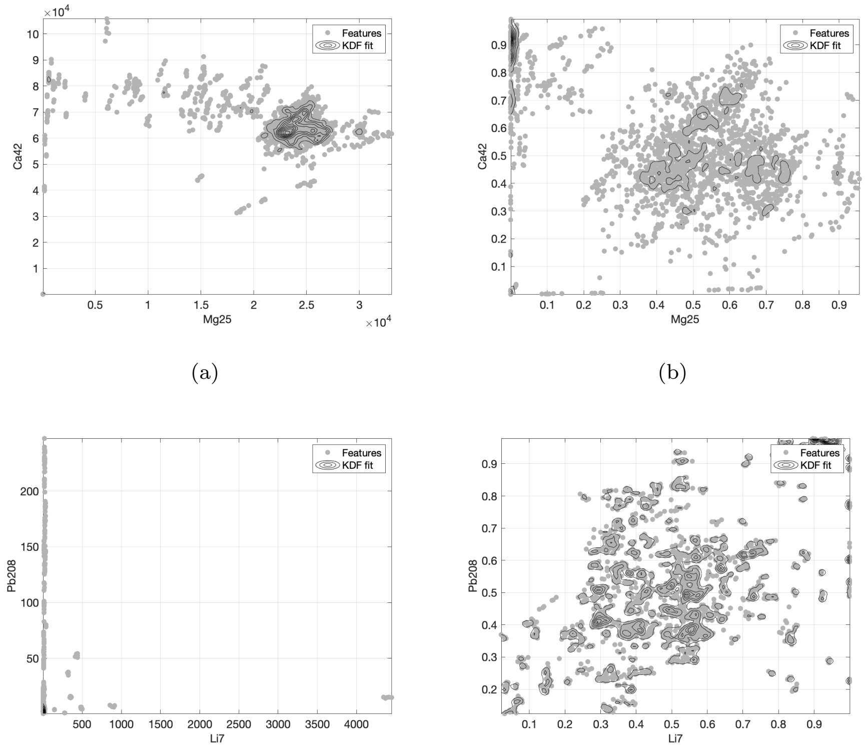
For the sake of illustration, Figures 1 and 2 show two examples of two-dimensional scatter plots, each of them considering two of the 17 variables in the LA-ICP-MS feature vector. A bidimensional kernel density fitting [34] has been represented for more clarity. Also, plots for the LA-ICP-MS data in the FIU database and the BKA database before and after the proposed normalization are performed. It can be seen that the normalization process is fit for purpose, as the features are normalized to zero median and one interquartile distance, and because the effect of the outliers is reduced.

More generally, Figures 3 and 4 show the correlation diagrams of the BKA and FIU databases before and after the normalization process. It can be seen that the normalization process sets the range of values in more sensible and similar ranges across variables, and that the outliers are severely reduced. Finally, the correlations of the variables are in some cases very clearly kept after the normalization process.

4.4. Comparison protocol

In our problem, we measure performance empirically by repeating comparisons of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ features from the FIU database, whereas the model is trained on the background BKA database $\mathbf{X}$ described before. For each comparison, a LR is computed as in Equation 4. In that sense, sets $\mathbf{Y}_1$ and $\mathbf{Y}_2$ will be assumed to represent respectively the crime-scene and suspect glass samples in each comparison. Thus, each comparison can be viewed as a simulated forensic case.

The experimental protocol involves the splitting of all the data into training, validation, and test sets. The training set consists of the first 300 glass objects extracted from the BKA database, whereas the rest of 85 glass objects of the BKA database are used for validation. Training data have been used to train model parameters, and validation data are used for model selection described below. The testing FIU database has been used to measure the performance of models. This protocol resembles the one used in previous work with these



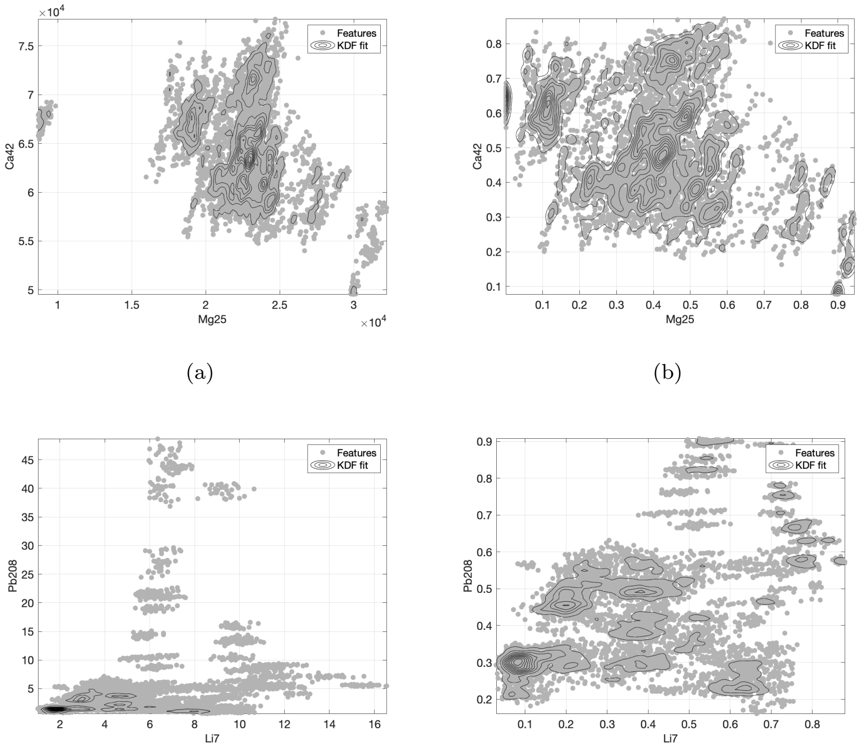
(a)

(b)

(c)

(d)

Figure 1: Bidimensional scatter plots for two examples of sample variables for the BKA database, before normalization (a and c on the left) and after normalization (b and d on the right). A kernel density function (KDF) fitting is shown as contour lines.



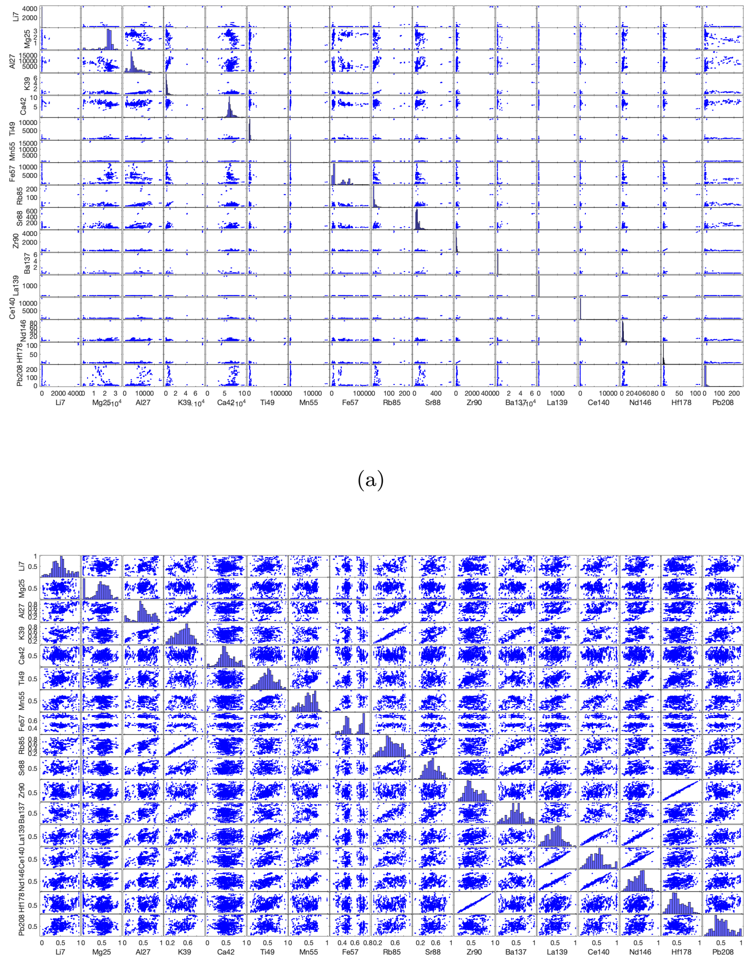
(a)

(b)

(c)

(d)

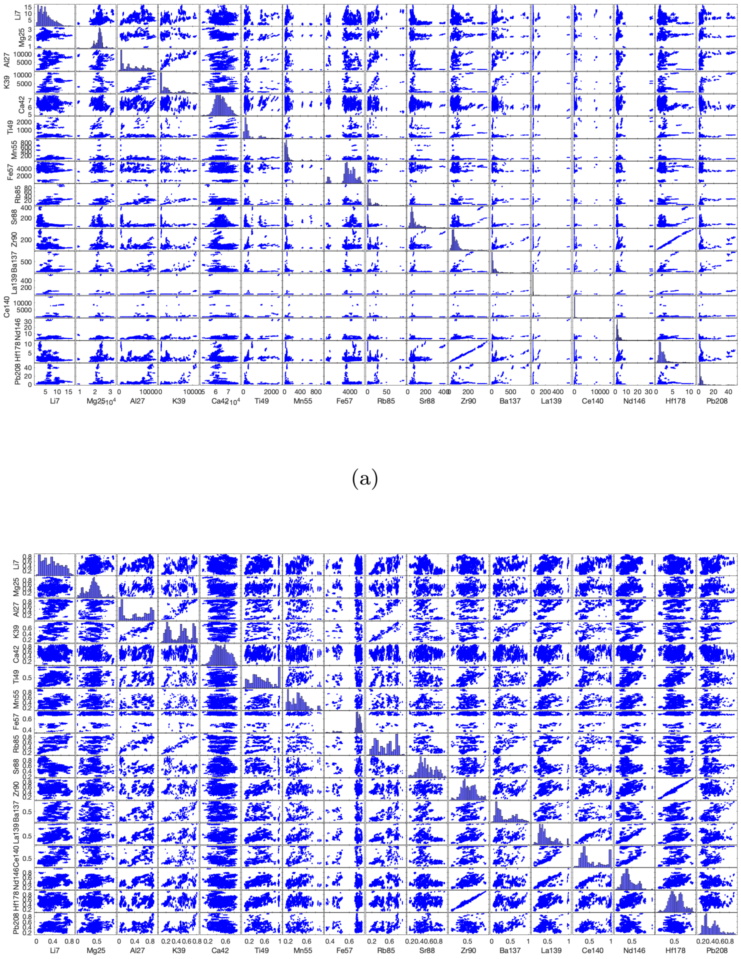
Figure 2: Bidimensional scatter plots for two examples of sample variables for the FIU database, before normalization (a and c on the left) and after normalization (b and d on the right). A kernel density function (KDF) fitting is shown as contour lines.



(a)

(b)

Figure 3: Correlation diagram for the BKA database, before normalization (a) and after normalization (b). The main diagonal in the diagram shows marginal histograms of each variable. Scatter plots are then represented in the intersections of variables.



(a)

(b)

Figure 4: Correlation diagram for the FIU database, before normalization (a) and after normalization (b). The main diagonal in the diagram shows marginal histograms of each variable. Scatter plots are then represented in the intersections of variables.

databases, see [18]. Following the methodology described in [6, 7], the BKA database would be the development set, used to train and select the best model; and the FIU database would be the forensic validation database, used to decide whether the method is suitable to be used in casework according to criteria. Recall that we used here the terminology *training*, *validation* and *testing* for it is more widely known from machine learning, but the analogy with development database and forensic validation databases as in [6, 7] is clear as described above.

We measure performance by comparing simulated crime-scene glass features and suspected glass features for which we know if h_{ss} is true (both features sets come from the same glass object), or conversely if h_{ds} is true (both features sets come from different glass objects). Thus, the testing FIU database has been divided into two different subsets, namely 210 *inner* panes and 210 *outer* panes of the considered windshields. Both subsets will be treated as different testing datasets for the purposes of simulating LR values. As each pane has a total of 15 replicate features, the comparison protocol is as follows. For same-source comparisons (where h_{ss} is true), the 3 first replicates are compared to the rest of non-overlapping groups of 3 replicates in the same object, to a total of 4 same-source LR values per each of the panes. This results in 840 LR values in the *inner* subset and 840 LR values in the *outer* subset, to a total of $1,680$ same-source testing LR values. The different-source comparisons (for which h_{ds} is true) are computed comparing the 3 first replicates of each pane with the rest of 5 non-overlapping groups of 3 replicates of the rest of objects in each subset separately (*inner* or *outer*). This leads to a total number of different-source testing LR values of $1,097,250$.

The model selection experiments using the BKA validation dataset were conducted using the same protocol as for the FIU testing database, considering the following. First, only 6 replicates were available for BKA objects, and therefore only 1 same-source comparison was possible for each BKA object in the validation dataset, to a total of 85 same-source LR values for model selection. On the other hand, following the same protocol as the one in the FIU test database, we obtained a total of 14280 different-source LR values for model selection.

4.5. Performance metrics

A solution to measure the performance of LR values has been proposed in [8] for speaker recognition and has been extended and adapted to forensic comparison problems [10, 6]. The main metric for performance measurement has been dubbed *log-likelihood-ratio cost* (C_{llr}), a scalar defined as follows:

$$C_{\text{llr}} = \frac{1}{2 \cdot n_{\text{ss}}} \sum_{i_{\text{ss}}} \log_2 \left(1 + \frac{1}{LR_{i_{\text{ss}}}}\right) + \frac{1}{2 \cdot n_{\text{ds}}} \sum_{j_{\text{ds}}} \log_2 \left(1 + LR_{j_{\text{ds}}}\right).$$

The indices i_{ss} and j_{ds} respectively refer to h_{ss} and h_{ds} comparisons, and n_{ss} and n_{ds} are respectively the total number of h_{ss} (i.e., same-source) comparisons and h_{ds} (i.e., different-source) comparisons in the testing LR set. It is easily seen that C_{llr} is the well-known binary cross-entropy function when prior probabilities $P(H = h_{\text{ss}}) = P(H = h_{\text{ds}}) = 0.5$ [21].

An important result is derived in [8], where it is proven that minimizing the value of C_{llr} also encourages to obtain better decisions. This property has been highlighted as extremely important in forensic science [9]. Moreover, in [8], the Pool Adjacent Violators (PAV) algorithm is used in order to decompose C_{llr} as follows:

$$C_{\text{llr}} = C_{\text{llr}}^{\text{min}} + C_{\text{llr}}^{\text{cal}} \tag{12}$$

where:

- $C_{\text{llr}}^{\text{min}}$ represents the *discrimination cost* of the LR method, and it is due to non-perfect discriminating power. $C_{\text{llr}}^{\text{min}}$ can be viewed as the same kind of measure as the well-known Area Under ROC curve (AUC) [47, 8].
- $C_{\text{llr}}^{\text{cal}}$ represents the *calibration cost* of the LR values.

A detailed revision of this decomposition of cross-entropy can be found in [21]. Moreover, a *neutral* value of $C_{\text{llr}} = 1$ is obtained when all the LR values computed by the system are equal to 1, representing the performance of a *useless* LR model that does not gives any information to the trier of fact (i.e., if LR is

1, the evidence is said to have *no value*). Thus, this value of $C_{\mathrm{llr}} = 1$ represents an ultimate performance limit, and if $C_{\mathrm{llr}} > 1$ for a LR model, it should be discarded as a possible model in casework.

4.6. Results

Here we describe the model selection experiments using the BKA validation dataset, also referred to as *validation experiments*. Next, we will present the test experiments in the FIU database.

4.6.1. Training the models with BKA train database

We trained the parameter distributions of HW and HVAE models with the training split of the BKA database. For the warped Gaussian mixture, we trained it by varying many of the hyperparameters of the model, as described in [32]. We briefly describe the main ones here. The mixture was initialized with 5 components. The step size of the leap-frog algorithm used to perform the integrator was 0.01. A Gaussian process latent variable model was used firstly to initialize the model. The rest of the parameters were varied according to Table 1, see the caption for the definition of the parameters of the model, and see [32] for further details.

As the number of hyperparameters of the VAE models was higher, we select the best VAE models by their marginal log-likelihood with the training split of the BKA data, which represents the consistency of the trained models with the training data. In Table 2 we show the configurations of the VAE selected in this way. For all the VAE models, the aleatoric uncertainty is modeled. Details can be found in [31].

4.6.2. Model selection results using BKA validation database

The aim of these experiments is two-fold. On the one hand, we want to find what model configuration performs best in the validation dataset, in order to select it for the testing experiments. On the other hand, we want to show the robustness of the approach to the selection of parameters.

Model ID	M_c	WU	LatDim	τ
HW1	10000	4000	5	60
HW2	10000	4000	2	100
HW3	10000	4000	5	20
HW4	10000	4000	10	100
HW5	4000	1000	10	100
HW6	4000	1000	5	60
HW7	10000	4000	10	60
HW8	4000	1000	10	60
HW9	4000	1000	2	60
HW10	10000	4000	10	20
HW11	4000	1000	10	20
HW12	10000	4000	2	20
HW13	10000	4000	2	60
HW14	4000	1000	5	20
HW15	10000	4000	5	100
HW16	4000	1000	5	100
HW17	4000	1000	2	100
HW18	4000	1000	2	20

Table 1: Configurations of the warped Gaussian mixtures used for the HW model that performed better than the baseline in the validation dataset (split of the BKA database). The fields are defined as follows: M_c is the number of Monte-Carlo samples inferred from the parameter posterior of the mixture. *WU* are the warm-up samples for Markov-Chain Monte-Carlo. *LatDim* is the dimension of the latent space of the mixture. Finally, τ is the number of steps used for the integrator based on Hamiltonian Monte Carlo. See [32] for details.

Model ID	Layers	Neurons	LatDim	β	$\mathcal{L}$
VAE1 no Dropout	1	50	10	1	14.3
VAE2 no Dropout	2	50	2	1	12.0
VAE3 no Dropout	2	25	2	1	11.9
VAE1 Dropout	1	50	5	0.1	9.4
VAE2 Dropout	1	50	10	0.1	9.3
VAE3 Dropout	1	50	2	1	9.2

Table 2: Configurations of the VAEs used for the HVAE model that performed better than the baseline in the validation dataset (split of the BKA database). The fields are defined as follows: *Layers* is the number of hidden layers of the neural network of the autoencoder. *Neurons* is the number of neurons of the hidden layers. *LatDim* is the dimension of the latent space. β is a power applied to the Kullback-Leibler divergence when computing the Evidence Lower Bound (ELBO) of the variational inference. Finally, $\mathcal{L}$ is the marginal likelihood of the data for each of the models. See [31] for details.

For all the model selection experiments using the BKA validation dataset, the degrees of freedom of the heavy-tailed within-source distribution, namely ν, have been varied across the following values: $\{1, 2, 3, 5, 10, 20, 50\}$. As an intuition, the value $\nu = 1$ means that the distribution is more "spread" in the feature space, indicating more uncertainty about the features in the model. On the other hand, $\nu \geq 50$ greatly approximates the Gaussian within-source variability, indicating that the model is more similar to the Gaussian, maximum likelihood distribution, and therefore incorporating less uncertainty to the model. Thus, finding the best value for ν can be seen as finding the best amount of uncertainty to be considered by the model.

Figures 5 and 6 show the model selection experiments using the BKA validation dataset with HW and HVAE models, respectively. Some observations can be made about these models. Firstly, it can be seen that the optimal degrees of freedom of the within-source Student's t-distribution are around $\nu = 10$ in all the cases. This means that the best models (in terms of C_{llr}) are different from a maximum-likelihood Gaussian within-source variability, meaning that incorporating uncertainty in the model improves the performance by obtaining better calibration. However, the incorporation of the uncertainty has an optimal value, and if a too low ν is used, then the discriminating power is degraded.

Also, it can be seen that the baseline presents better discriminating power than the proposed models. This may be explained as follows: as the uncertainty is incorporated poorly by the baseline ALK model, as long as two LA-ICP-MS sets of features are slightly different, the ALK model distinguished between them extremely well. However, this distinction is heavily overstated, and the likelihood ratio for such a comparison turns out to be zero. As this happens to almost all different-source comparisons, there will be many LRs providing very strong evidence (in fact, infinitely strong). This presents the risk that future LR values will be also misleadingly strong, as there is not a guarantee that models yielding overstating LR values will not generate a misleading, overstated LR value in the future. On the other hand, for same-source comparisons, the analogous situation happens, where the baseline ALK model assigns a very high number to the numerator of the LR, due to more concentrated distributions (i.e., Gaussian instead of heavy-tailed). This leads to an overstatement of the weight of the evidence for many of the same-source comparisons. These observations are consistent with [18] and [19], where the baseline ALK model provides these overstated LR values, which on itself are used as scores that are further calibrated and limited. However, here, as described before, we compute better calibrated LR values in a single, feature-based step.

Secondly, we observe that the model is quite robust to the selection of parameters. We see that many models are better than the baseline in terms of C_{llr} for a variety of parameter configurations and degrees of freedom ν. This is specially apparent for the HW model (Figure 6). This evidences that the behavior of the model is robust to small variations of the hyperparameters of the VAE and the warped Gaussian mixture, and also to the degrees of freedom. Therefore, we can expect robustness in real casework.

Although the improvement on the calibration of the proposed models with respect to the baseline is clear, it is true that some calibration loss is presented by the models. This is evidenced by the remaining values of $C_{\text{llr}}^{\text{cal}}$.

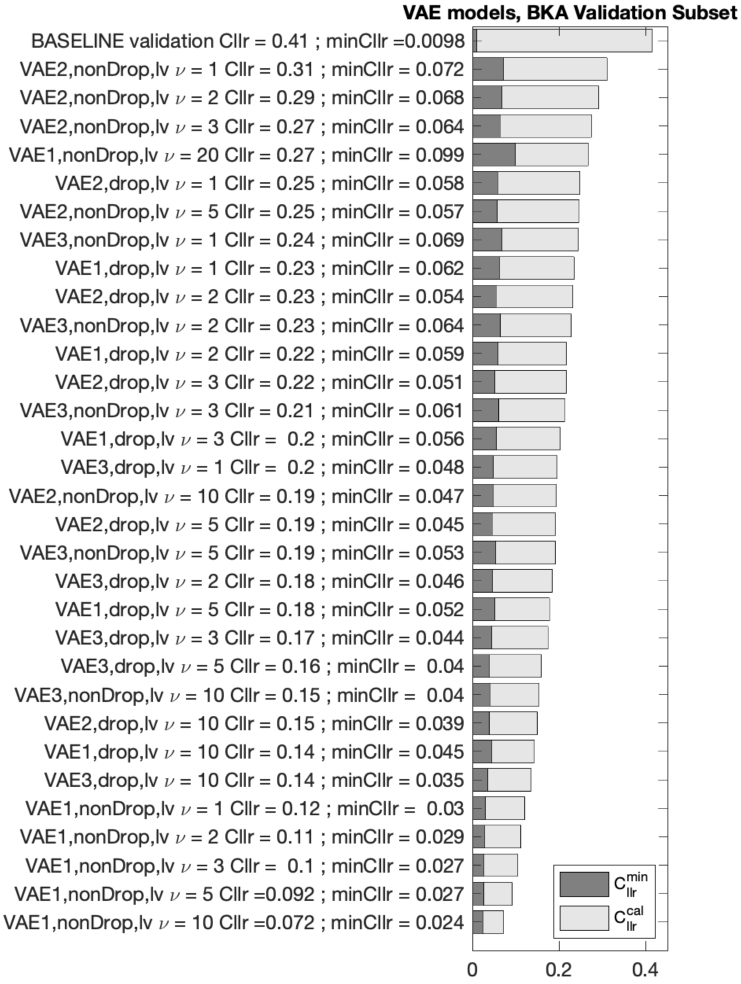


Figure 5: Performance of the HVAE models in the validation split of the BKA database, compared to the baseline. Only models that are better than the baseline are represented. The VAEs are as defined in Table 2, ν are the degrees of freedom of the Student's t-distribution for the within-source variability, and *lv* indicates that the aleatoric uncertainty was learned by the VAEs.

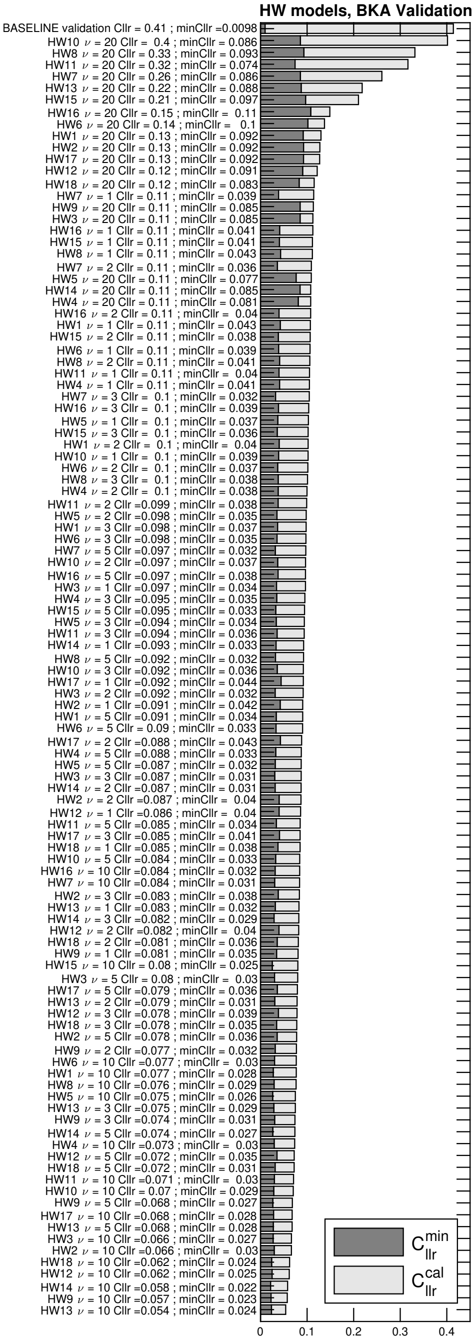


Figure 6: Performance of the HW models in the validation split of the BKA database, compared to the baseline. Only models that are better than the baseline are represented. The warped Gaussian mixture models are described in Table 1, and ν are the degrees of freedom of the Student's t-distribution for the within-source variability.

Following model selection experiments using the BKA validation dataset, we selected the best-performing VAE and warped Gaussian mixture (Figures 5−6) for the testing results using the FIU database.

4.6.3. Test results using the FIU database as training with the BKA database

Figure 7 shows the comparative performance of the baseline model and the best HVAE and HW models. Some observations are in order. First, it can be seen that both HW and HVAE models dramatically outperform the baseline in terms of the primary performance measure, namely C_{llr}. However, and secondly, we also observe some loss in the discriminating power of the HW and HVAE models with respect to the baseline. This is possibly because the effect of incorporating uncertainty by the Student's t-distribution might be leading to more *confusable* objects in the testing database. Therefore, we observe some trade-off within these models: if we want to incorporate uncertainty by the use of more heavy-tailed distributions, the discriminating power may be affected, but the calibration will dramatically improve. It might be the case that a more consistent model, such as a fully-Bayesian model, would help to improve the discriminating power while retaining the uncertainty necessary for a good calibration. This will be explored in future work.

As it was seen in the model selection experiments using the BKA validation dataset, we still observe considerable values of $C_{\text{llr}}^{\text{cal}}$ for the proposed models. This suggest that the calibration could still be improved, for instance by the use of score-based stages like in [15, 19, 18]. However, although we have not considered score-based approaches in this work, it also seems evident that the proposed models are still far from the state-of-the-art score-based approaches for LA-ICP-MS data (see e.g. [19, 18]). Nevertheless, these results are reported as the experimental demonstration that it is possible to achieve a value of $C_{\text{llr}} < 1$ just with a feature-based LR model. To our knowledge, it is the first time that this is achieved for LA-ICP-MS data.

To gain further insight into the rationale of the proposed models, we attempt to analyze two simultaneous effects: the influence of the use of a more complex

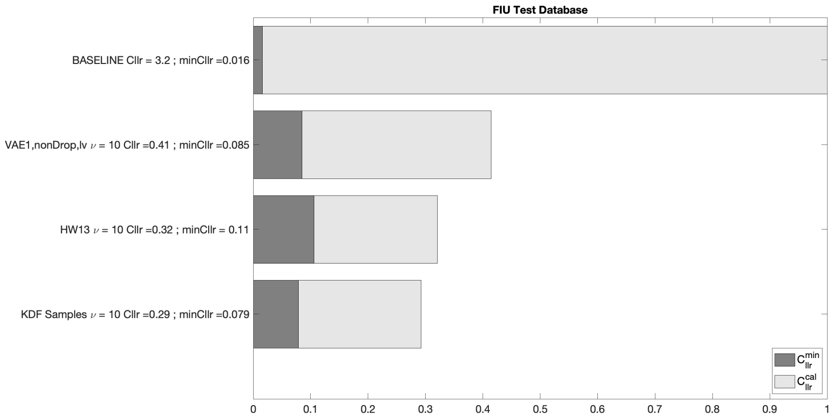


Figure 7: Performance of the selected HVAE and HW models in the FIU testing database, compared to the baseline and a heavy-tailed version of the baseline (*KDF Samples*).

between-source distribution for the parameter distribution; and the incorporation of uncertainty in the within-source variability by the use of a heavy-tailed Student's t distribution. In order to do this, we tested a Monte-Carlo sampled version of the ALK model, namely *KDF Samples* model. Here, we have sampled from the KDF defined by the parameter distribution of the ALK baseline (see Section 3.2), and then we used those samples to compute Equation 4 with a heavy-tailed within-source distribution, using $\nu = 10$ as in the HW and HVAE models. More specifically, the likelihood of the KDF samples model would be $P\left(\mathbf{y}|\,\boldsymbol{\mu}\right) \sim \mathcal{T}\left(\mathbf{y}|\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_w, \nu\right)$ as in the HW and HVAE models, and the LR would be computed by Monte Carlo sampling as:

$$\frac{P\left(\mathbf{Y}_1, \mathbf{Y}_2|\, h_{\text{ss}}, \mathbf{X}\right)}{P\left(\mathbf{Y}_1, \mathbf{Y}_2|\, h_{\text{ds}}, \mathbf{X}\right)} = \frac{\frac{1}{M}\sum_{i=1}^{M} P\left(\mathbf{Y}_1|\,\boldsymbol{\mu}_i\right) P\left(\mathbf{Y}_2|\,\boldsymbol{\mu}_i\right)}{\left[\frac{1}{M}\sum_{i=1}^{M} P\left(\mathbf{Y}_1|\,\boldsymbol{\mu}_i\right)\right]\left[\frac{1}{M}\sum_{i=1}^{M} P\left(\mathbf{Y}_2|\,\boldsymbol{\mu}_i\right)\right]} \tag{13}$$

where M mean vectors are sampled from the parameter distribution of the ALK model (Section 3.2), as:

$$\boldsymbol{\mu}_i \sim P\left(\boldsymbol{\mu}|\,\mathbf{X}, k\hat{\boldsymbol{\Sigma}}_b\right). \tag{14}$$

Figure 7 shows the comparative results of this model with the proposed

HW and HVAE models in the FIU database. It is clearly seen that the *KDF Samples* model has comparable performance as the HW model, and better than the HVAE model in terms of C_{llr}. This is a surprising result since we hypothesized that a better model for the parameter distribution would help to obtain better calibration. However, it is seen that the important contribution of the model relies upon the use of the heavy-tailed distribution for the within-source variability, and apparently not based upon the proposed, more complex parameter distributions.

A further analysis can be performed from Figure 8, that shows the histograms of the different models proposed, as compared to the baseline. It turns out that the baseline presents overstating LR values, that in fact, if they are not artificially limited, would be zero in the majority of the different-source comparisons. However, this effect is damped by the incorporation of the uncertainty via the Student's t-distribution within-source variability, leading to much more sensible LR values in the proposed models, and much better calibration. However, this is at the cost of some loss in the discriminating power, as observed in the model selection experiments using the BKA validation dataset. As the histograms suggest in Figure 8, the discriminating power of the baseline model is still much better than for the proposed models, and therefore the gain on calibration for the proposed models is still at the cost of a considerable loss of discriminating power.

Some strange behavior of different-source scores needs an explanation. First, for Figures 8(a), (b) and (d), the density of larger different-source LR values increases. This is not a quite typical behavior in different-source scores as previously seen in the literature. Our explanation is that, as the tails of the within-source variability are heavy, there is some kind of limiting effect on the ability of the model to discriminate between two feature vectors. This might also apply to very similar different-source feature vectors, explaining that those pairs could not yield very high LR values for the heavy-tailed model. Thus, the model might be limiting the LR values for different-source scores coming from these comparisons, leading to an increase of the density of different-source log-LRs when approaching the limit. This does not happen in the baseline model in Figure 8(a), where different-source comparisons are extremely well discriminating, but at the cost of LR overstatement. This effect again evidences the tradeoff between calibration and discriminating power of the proposed models.

We also see that, although $C_{\text{llr}} < 1$ for the proposed models, there is still some calibration loss that can be evidenced in some of the histograms. For instance, if we consider that calibrated LR values must hold that *the LR of the LR is the LR*, in Figures 8(a), (b) and (c), this is not totally observed for most ranges. This is in accordance with the still relevant values of $C_{\text{llr}}^{\text{cal}}$ previously reported. Thus, although we achieved a dramatic improvement in C_{llr} with respect to the baseline, further research efforts should be made to improve both the discriminating power and the calibration of the proposed models at the feature level.

Finally, in Figure 9 we show the plot of the LR values obtained for the FIU testing dataset, both before and after applying Pool Adjacent Violators (PAV) over the same set of LR values. This plot can be interpreted as follows: the black dashed line is the $x = y$ line, indicating perfect calibration. The red solid line indicates the PAV transformation to the LR values, when the same LR values are used to train PAV. If the red line is equal to the black dashed line, the calibration is perfect. A larger deviation from the $x = y$ by the red line indicates worse calibration. It can be seen that the calibration of the baseline model strongly deviates from the $x = y$ line, indicating very bad calibration. On the other hand, the red solid line is much closer to the $x = y$ line for the proposed models, indicating a significant calibration improvement. However, the calibration of the proposed models is still far from perfect, as the red solid line is still far from the $x = y$ black dashed line. This is coherent with our previous observations.

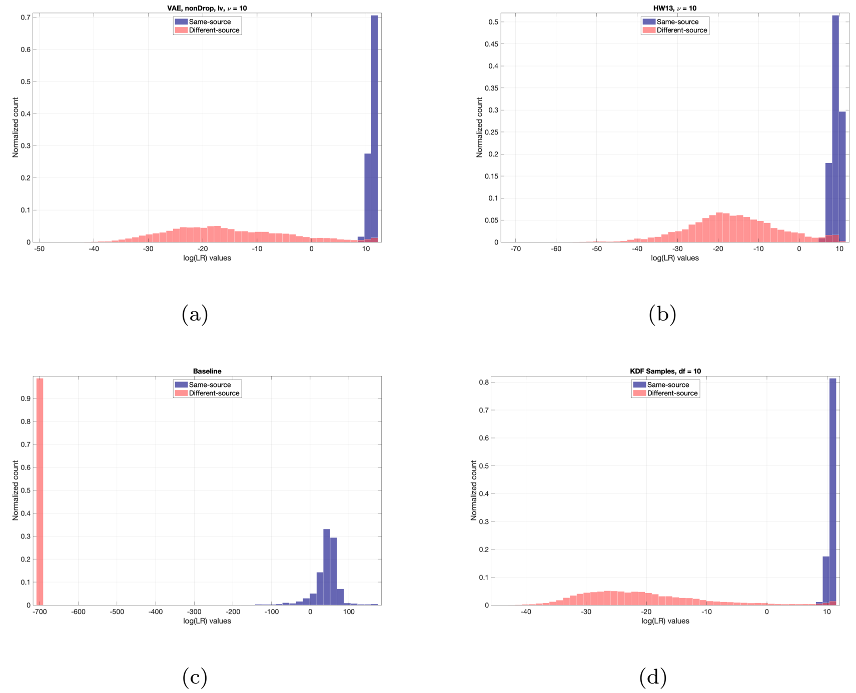
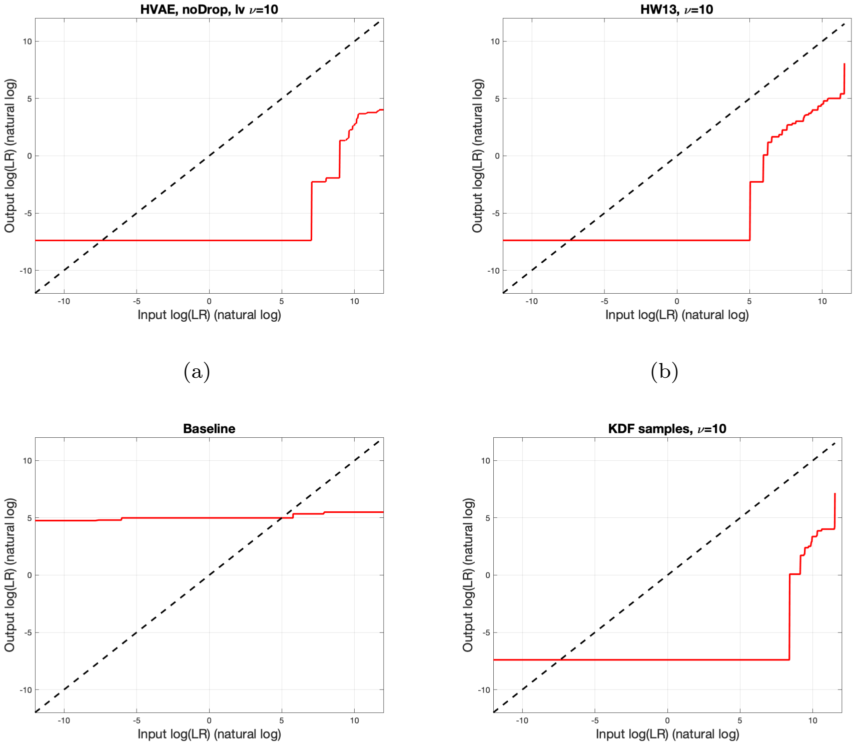


Figure 8: Histograms of the best HVAE (a) and HW (b) models selected in validation, compared to the ALK baseline (c) and a heavy-tailed version of the baseline named *KDF Samples* (d). Log-LRs are computed with natural logarithms.

5. Conclusions

We have proposed several models for feature-level forensic glass comparison, namely HW and HVAE. These models clearly outperform classical feature-based models as presented in [17] in terms of C_{llr}, a popular performance measure that takes into account both discriminating power and calibration [8, 6, 7, 21]. Remarkably, our proposal dramatically improves the calibration of the likelihood ratios obtained, at the cost of a considerable loss in the discriminating power. Moreover, although the proposed models are complex in terms of the number of parameters needed, the robustness is demonstrably acceptable. The models present a value of $C_{\text{llr}} < 1$, and therefore they could be adopted for use in a forensic comparison case.



(a)

(b)

(c)

(d)

Figure 9: PAV plots of the best HVAE (a) and HW (b) models, compared to the ALK baseline (c) and the heavy-tailed version of the baseline named *KDF Samples* (d). Log-LRs are computed with natural logarithms.

However, it seems reasonable that a forensic laboratory or practitioner will always want to use the best model available, and currently the proposed models are outperformed by score-based approaches such as [18, 19]. Therefore, it seems unreasonable to use our proposal instead of those score-based approaches. On the other hand, $C_{\text{llr}} < 1$ is an evaluation criterion with a theoretical basis, but the validation of a model for forensic practice must consider many other validation criteria as defined by the forensic laboratory or practitioner [6, 7]. The value of this contribution relies upon models presenting $C_{\text{llr}} < 1$ for this task with only a feature-based approach, but more research must be conducted in order to make these models more reliable, more discriminating, and therefore, more amenable for use in forensic practice.

To our knowledge, this is the first proposal of a feature-based likelihood ratio model for LA-ICP-MS-based glass comparison that presents $C_{\mathrm{llr}} < 1$ without the use of a further score-based calibration step. In this sense, we believe that this work will help with the use of probabilistic models for LA-ICP-MS data directly in the feature space. Also, we believe that our conclusions will shed some light in the rationale of comparison models in complex feature spaces, that may enable further applications in other areas of forensic science.

The main conclusion of our work is that there is a strong need for probabilistic models to efficiently incorporate uncertainty. Our observations are in accordance with the idea that, if the models do not consider the uncertainty, the likelihood ratios computed in the feature space become unstable, yielding unreasonable strength of evidence. The main harmful effect of overstated evidence is that sometimes it will support the incorrect proposition, a situation to be avoided in critical applications such as forensic science. However, by the use of a heavy-tailed distribution for the within-source variability of the glass features, we force this uncertainty to be considered by the model, dramatically improving the calibration and preventing evidence over-statement.

We cannot forget the loss of discrimination that is observed in the proposed models in a tradeoff with good calibration. This issue may be related to the fact that the uncertainty is incorporated in a forced way, directly affecting only one of the levels of the model (the within-source variability) and only one of its parameters (the mean vectors). This suggests the use of fully-Bayesian models in order to incorporate all the uncertainty present in the whole set of model parameters, not only in some of them, an idea that will drive our future research.

Another important observation is that the consistent modeling of the parameter distribution remains an issue. We have tested two complex and flexible models in order to achieve this goal, but they did not improve over the baseline, kernel density model. Moreover, we believe that the inadequate definition of the parameter distribution may affect the influence of the incorporation of the uncertainty by the heavy-tailed distribution. All this could be accommodated with extensions of the proposed models, such as fully-Bayesian versions of a variational autoencoder. This idea will be further explored in future work.

6. Acknowledgements

This work has been partially developed during a research stay of D. R. at the Machine Learning Group of the Department of Engineering, University of Cambridge, UK; funded by the Spanish *Ministerio de Educación, Cultura y Deporte* under the program for mobility of professors and researchers in higher education and research centers. We strongly thank Dr. José Miguel Hernández-Lobato for all the ideas and discussions about the algorithmic developments in this article, and for hosting the aforementioned research stay. We also thank Peter Weis and Sonja Menges from the German Federal Police (Bundeskriminalamt) for their support, their revision of the manuscript, and for providing the BKA database. This development of the Florida International University (FIU) background database of vehicle glass samples was supported by Award No. 2018-DU-BX-0194 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice to Florida International University. The opinions, findings, and conclusions or recommendations expressed in this manuscript are those of the authors and do not necessarily reflect those of the National Institute of Justice or the U.S. Department of Justice. D. R and J. M. were supported by the Spanish *Ministerio de Educación, Cultura y Deporte* through grant RTI2018-098091-B-I00. J.M is also supported by grant FPI-UPV associated to the DeepHealthProject, grant agreement No 825111.

References

[1] D. V. Lindley, A problem in forensic science, Biometrika 64 (2) (1977) 207–213.

[2] I. W. Evett, Towards a uniform framework for reporting opinions in forensic science casework, Science and Justice 38 (3) (1998) 198–202.

[3] C. G. G. Aitken, F. Taroni, Statistics and the evaluation of the evidence for forensic scientists, John Wiley and Sons, 2004.

[4] I. Evett, Expressing evaluative opinions: A position statement, Science and Justice 51 (2011) 1–2, several signatories.

[5] S. Willis, ENFSI Guideline for the Formulation of Evaluative Reports in Forensic Science. Monopoly Project MP2010: The development and implementation of an ENFSI standard for reporting evaluative forensic evidence, Tech. rep., European Network of Forensic Science Institutes, Wiesbaden, Germany (2015).

[6] D. Ramos, D. Meuwly, R. Haraksim, C. E. H. Berger, Validation of Forensic Automatic Likelihood Ratio Methods, CRC Press, 2021, pp. 143–164. doi:https://doi.org/10.1201/9780367527709.

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

[8] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Computer Speech and Language 20 (2-3) (2006) 230–275.

[9] D. Ramos, J. Gonzalez-Rodriguez, Reliable support: measuring calibration of likelihood ratios, Forensic Science International 230 (2013) 156–169.

[10] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, Information-theoretical assessment of the performance of likelihood ratio models, Journal of Forensic Sciences 58 (6) (2013) 1503–1518. doi:http://dx.doi.org/10.1111/1556-4029.12233.

[11] D. Meuwly, Reconaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approache Automatique, Ph.D. thesis, IPSC-Université de Lausanne, 2001.

[12] D. Ramos, Forensic evaluation of the evidence using automatic speaker recognition systems, Ph.D. thesis, Depto. de Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain (2007).

[13] A. Bolck, H. Ni, M. Lopatka, Evaluating score-and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison, Law, Probability and Risk 14 (3) (2015) 243–266. doi:10.1093/lpr/mgv009.

[14] A. B. Hepler, C. P. Saunders, L. J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Foresic Science International 219 (1-3) (2011) 129–140.

[15] G. S. Morrison, Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio, Australian Journal of Forensic Sciences 45 (2013) 173—-197. doi:10.1080/00450618.2012.733025.

[16] G. S. Morrison, E. Enzinger, Score based procedures for the calculation of forensic likelihood ratios - scores should take account of both similarity and typicality, Science & Justice 58 (1) (2018) 47–58. doi:https://doi.org/10.1016/j.scijus.2017.06.005.

[17] C. G. G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, Applied Statistics 53 (2004) 109–122, With corrigendum 665-666.

[18] R. Corzo, T. Hoffman, P. Weis, J. Franco-Pedroso, D. Ramos, J. Almirall, The use of LA-ICP-MS databases to estimate likelihood ratios for the forensic analysis of glass evidence, Talanta 186 (1) (2018) 655–661. doi:https://doi.org/10.1016/j.talanta.2018.02.027.

[19] 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, Science & Justice 57 (3) (2017) 181–192. doi:https://doi.org/10.1016/j.scijus.2017.03.002.

[20] G. S. Morrison, E. Enzinger, D. Ramos, J. Gonzalez-Rodriguez, A. Lozano-Diez, Statistical Models for Forensic Voice Comparison, CRC Press, 2021, pp. 451–498. doi:https://doi.org/10.1201/9780367527709.

[21] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, J. Gonzalez-Rodriguez, Deconstructing cross-entropy for probabilistic binary classifiers, Entropy 20 (3) (2018) 208. doi:10.3390/e20030208.

[22] D. Ramos, J. Maroñas-Molano, A. Lozano-Diez, Bayesian strategies for likelihood ratio computation in forensic voice comparison with automatic systems, in: Subsidia: Tools and Resources fot the Speech Sciences, 2017.

[23] M. E.Sigman, M. R. Williams, Assessing evidentiary value in fire debris analysis by chemometric and likelihood ratio approaches, Forensic Science International 264 (2016) 113–121. doi:https://doi.org/10.1016/j.forsciint.2016.03.051.

[24] D.-M. K. Dennis, M. R. Williams, M. E. Sigman, Assessing the evidentiary value of smokeless powder comparisons, Forensic Science International 259 (2016) 179–187. doi:https://doi.org/10.1016/j.forsciint.2016.03.051.

[25] D. Ramos, J. Franco-Pedroso, J. Gonzalez-Rodriguez, Calibration and weight of the evidence by human listeners: the ATVS-UAM submission to NIST human-aided speaker recognition 2010, in: Proc. of ICASSP 2011, Pague, Czeck Republic, 2011, pp. 5908 – 5911.

[26] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[27] A. Gupta, R. Corzo, A. Akmeemana, K. Lambert, K. Jimenez, J. M. Curran, J. R. Almirall, Dimensionality reduction of multielement glass evidence to calculate likelihood ratios, Journal of Chemometrics 35 (1) (2021) e3298. doi:https://doi.org/10.1002/cem.3298.

[28] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature 521 (2015) 452–459. doi:https://doi.org/10.1038/nature14541.

[29] J. Franco-Pedroso, D. Ramos, J. Gonzalez-Rodriguez, Gaussian mixture models of between-source variation for likelihood ratio computation from multivariate data, PLoS ONE 11 (2) (2016) 1–25. doi:http://dx.doi.org/10.1371/journal.pone.0149958.

[30] J. P. Williams, D. M. Ommen, J. Hannig, Generalized fiducial factor: an alternative to the bayes factor for forensic identification of source problems (2020). arXiv:2012.05936.

[31] D. P. Kingma, M. Welling, Autoencoding variational Bayes, in: Proceedings of the International Conference on Learning Represesntations (ICLR), 2014. arXiv:1312.6114v10.

[32] T. Iwata, D. Duvenaud, Z. Ghahramani, Warped mixtures for nonparametric cluster shapes, in: Uncertainty in Artificial Intelligence, 2013. arXiv:1206.1846v2.

[33] D. M. Ommen, C. P. Saunders, Building a unified statistical framework for the forensic identification of source problems, Law. Probability and Risk 17 (2) (2018) 179–197. doi:https://doi.org/10.1093/lpr/mgy008.

[34] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.

[35] D. V. Lindley, Understanding Uncertainty, Wiley, 2006.

[36] G. S. Morrison, N. Poh, Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/bayes factors, Science & Justice 58 (3) (2018) 200–218. doi:https://doi.org/10.1016/j.scijus.2017.12.005.

[37] C. Rasmussen, The infinite gaussian mixture model, in: S. Solla, T. Leen, K. Müller (Eds.), Advances in Neural Information Processing Systems, Vol. 12, MIT Press, 2000, pp. 554–560.

URL https://proceedings.neurips.cc/paper/1999/file/97d98119037c5b8a9663cb21fb8ebf47-Paper.pdf

[38] N. D. Lawrence, Learning for larger datasets with the gaussian process latent variable model, in: M. Meila, X. Shen (Eds.), Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, Vol. 2 of Proceedings of Machine Learning Research, PMLR, San Juan, Puerto Rico, 2007, pp. 243–250.
URL http://proceedings.mlr.press/v2/lawrence07a.html

[39] D. J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, in: E. P. Xing, T. Jebara (Eds.), Proceedings of the 31st International Conference on Machine Learning, Vol. 32 of Proceedings of Machine Learning Research, PMLR, Bejing, China, 2014, pp. 1278–1286.
URL http://proceedings.mlr.press/v32/rezende14.html

[40] A. Akmeemana, P. Weis, R. Corzo, D. Ramos, P. Zoon, T. Trejos, T. Ernst, C. Pollock, E. Bakowska, C. Neumann, J. Almirall, Interpretation of chemical data from glass analysis for forensic purposes, Journal of Chemometrics 35 (1) (2021) e3267. doi:https://doi.org/10.1002/cem.3267.

[41] E2927-16e1 In Standard Test Method for the Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Ablation Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons, Standard, ASTM International, West Conshohocken, PA (2015).

[42] T. Trejos, J. Almirall, Sampling strategies for the analysis of glass fragments by LA-ICP-MS part ii: sample size and sample shape considerations, Talanta 67 (2) (2005) 396–401.

[43] 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, Journal of Analytical Atomic Spectrometry 26 (6) (2011) 1273–1284.

[44] C. Latkoczy, S. Becker, M. Ducking, D. Gunther, J. Hoogewerff, J. Almirall, J. Buscaglia, A. Dobney, R. Koons, S. Montero, G. Q. Peijl, W. Stoecklein, T. Trejos, J. Watling, V. Zdanowicz, Development and evaluation of a standard method for the quantitative determination of elements in float glass samples by LA-ICP-MS, Journal of Forensic Sciences 50 (6) (2005) 1327–1341.

[45] R. Corzo, T. Hoffman, T. Ernst, T. Trejos, T. Berman, S. Coulson, P. Weis, A. Stryjnik, H. Dorn, E. Pollock, M. Workman, P. Jones, B. Nytes, T. Scholz, H. Xie, K. Igowsky, R. Nelson, K. Gates, J. Gonzalez, L. Voss, E. Steel, J. Almirall, An interlaboratory study evaluating the interpretation of forensic glass evidence using refractive index measurements and elemental composition, Forensic Chemistry 22 (2021) 100307. doi:doi.org/10.1016/j.forc.2021.100307.

[46] T. Hoffman, R. Corzo, P. Weis, E. Pollock, A. van Es, W. Wiarda, A. Stryjnike, H. Dorne, A. Heydon, E. Hoise, S. L. Franc, X. Huifang, B. Pena, T. Scholz, J. Gonzalez, J. R. Almirall, An inter-laboratory evaluation of la-icp-ms analysis of glass and the use of a database for the interpretation of glass evidence, Forensic Chemistry 11 (2018) 65–76. doi:https://doi.org/10.1016/j.forc.2018.10.001.

[47] T. Fawcett, A. Niculescu-Mizil, PAV and the ROC convex hull, Machine Learning 68 (1) (2007) 97–106.