THE VALUE OF SCIENTIFIC EVIDENCE FOR FORENSIC MULTIVARIATE DATA

Silvia Bozza[1]
Department of Economics, Ca' Foscari University of Venice, Venice, Italy
School of Criminal Justice, The University of Lausanne, Lausanne, Switzerland

Abstract *Multivariate continuous data are becoming more prevalent in forensic science. Available databases may present a complex dependence structure, with several variables and several levels of variation. The assessment of the value of evidence can be performed by the derivation of a likelihood ratio, a rigorous concept that measures the change produced by a given item of information in the odds in favour of a proposition as opposed to another, when going from the prior to the posterior distribution. The derivation of a likelihood ratio may be a demanding task, essentially because of the complexity of the scenario at hand and the possible poor information at the forensic examiner's disposal. This opened the door in the forensic community to a large debate about what should be the most appropriate way to take charge of uncertainty while presenting expressions of evidential value at trial. These ideas will be illustrated with reference to a comparative handwriting scenario.*

Keywords: *Likelihood ratio assessment; Evaluation of scientific evidence; Multivariate data; Handwriting.*

1. INTRODUCTION

In forensic science, statistical methods are currently largely used for assessing the probative value of criminal traces, such as DNA or other recovered materials. Decades ago, discussions about this topic were less structured and formalized than they are today, and the diversity of opinion could be substantial. Today, the evaluation of measurements on characteristics associated to trace evidence when a recovered item of unknown origin is compared with a control item whose origin is known is generally performed through the derivation of a Bayes factor (in the forensic context often referred as a likelihood ratio), a rigorous concept that provides a balanced measure of the degree to which the evidence is capable of discriminating among competing propositions that are suggested by opposing parties at trial (Lindley, 1991). The use of this metric of probative value is largely supported by operational standards and recommendations in different forensic disciplines (ENFSI, 2015). A Bayes factor can also be assessed with reference to

[1] Corrisponding author: silvia.bozza@unil.ch

investigative settings, when no trace is available for comparative purposes. The recovered evidence may be valuable to generate hypotheses and suggestions for explanations in order to give assistance to investigative authorities.

Undoubtedly, the assessment of a Bayes factor may be a demanding task, essentially because of the likely complexity of the scenario at hand and the possible poor information at the forensic scientist's disposal. Moreover, forensic laboratories have frequently access to equipment (e.g., scanning electronic microscope) which can readily provide continuous multivariate data, and scientific evidence is often presented in this form. Glass fragments that are searched and recovered at a crime scene, or drug samples that are seized since under suspicious of containing illicit substances may be analyzed and compared on the basis of a profile of chemical compounds as well as physical characteristics. Multivariate data also arise in other domains of forensic science, such as handwriting examination. A handwritten character can in fact be described by means of several variables, such as the width, the height, the surface, the orientation of the strokes, or by Fourier descriptors as it will be outlined later. This has originated an abundance of databases that often present a complex dependence structure with a large number of variables and multiple sources of variation.

It must be emphasized that the use of multivariate statistical techniques in forensic science applications has been often criticized because of the lack of background data from which to estimate parameters (e.g., the first and second order moments within- and between-sources) and several attempts have been proposed to lead a dimensionality reduction. For example, score-based models have been proposed with the aim of reducing multivariate information to a univariate distance or similarity score between items (see e.g. Bolck et al. (2015) for forensic MDMA comparison). Alternatively, the multivariate likelihood ratio can be simplified to a product of univariate likelihood ratios whenever variables can be taken as independent. However, this hypothesis is seldom warranted, and a likelihood ratio for multivariate data accounting for correlation between variables and possibly several levels of variation must be provided.

A graphical probability environment was proposed by Aitken et al. (2006) to reduce a dataset with a considerable number of variables to a product of mutually independent sets of reduced dimensions. In this way, the number of parameters are considerably reduced whilst retaining the dependence structure, not recognized in a model which assumes full independence. Clearly, any statistical methodology which leads to reduction of the multivariate structure to fewer or even only one dimension will need careful justification in order to avoid the challenge of suppression of evidence. Bayesian multilevel models for the evaluation of multivariate measurements on characteristics associated to questioned material that are capable to deal with such constraints have been proposed, among others, by Aitken and Lucy (2004) for the evaluation of scientific evidence that consists in glass fragments, by Bozza et al. (2008) for handwriting examination, and by Alberink et al. (2013) for comparison of ecstasy tablets on MDMA content.

Numerical procedures are often implemented to handle the complexity and to compute the marginal likelihoods under competing propositions. This, along with the acknowledgement of subjective evaluations that are unavoidably involved in the Bayes factor assessment process, and sensitivity upon available measurements or observations, has given rise to a large debate in the forensic community about what should be the most appropriate ways to take charge of uncertainty while presenting expressions of evidential value to a court of justice. These ideas will be illustrated with reference to handwriting examination, a forensic discipline that attracts nowadays considerable attention due to its uncertain status under new admissibility standards.

2. BACKGROUND DATA AND MODELS

Available background data may present increasing levels of complexity: observations can be collected in several groups, take for example single individuals in a population of writers, glass fragments collected in a population of glass windows or different plants in a population of Cannabis plants and so on. Consider the case where observations are divided into m groups with several members for each group. The data structure may suggest a two-level hierarchy, where the hierarchical ordering takes into account two sources of variation: that between measurements within the same source (the *within-source* variation), and that between sources (the *between-source* variation). For sake of illustration, imagine a database collecting the handwriting features of a population of m writers with several observations for each writer. The background data can be denoted as $\mathbf{x}_{ij} = (x_{ij1},\ldots,x_{ijp})$, where $i = 1,\ldots,m$ denotes the number of groups (i.e., writers), $j = 1,\ldots,n_i$ denotes the available measurements for each writer and p is the number of variables. A Bayesian statistical model can be introduced, with $f(\mathbf{x}_{ij} \mid \psi_i)$ measuring personal degree of belief in the data taking certain values given the hypothetical information that ψ_i takes certain values, and $\pi(\psi_i \mid H_k)$ measuring the personal belief about ψ_i prior to observing the data given the proposition at hand.

The data structure may be far more complex, requiring an additional level of variability alongside the levels above to take into account, for example, the measurements error (i.e., the error related to the precision of the instrument). In this way the hierarchical ordering can take into account three sources of variation: that between replicate measurements on the same item, that between items within the same group, and that between groups. Looking back at the previous example, one may consider for each writer several measurements of characters of different type.

A Bayesian multilevel model for the evaluation of transfer evidence for three-level multivariate data has been proposed by Aitken et al. (2006) in a different forensic domain, where the available database encompasses replicated measurements of the elemental composition of several glass fragments originating from a population of glass windows.

3. PROBABILISTIC MODELS FOR EVALUATIVE AND INVESTIGATIVE PURPOSES

Suppose that evidentiary samples are collected by investigative authorities, and that control samples are taken for comparative purposes. Let us denote the recovered and the control measurements by, respectively, $\mathbf{y}_1 = (\mathbf{y}_{11}, \dots, \mathbf{y}_{1n_1})$ and $\mathbf{y}_2 = (\mathbf{y}_{21}, \dots, \mathbf{y}_{2n_2})$, where $\mathbf{y}_{ij} = (y_{ij1}, \dots, y_{ijp})$, and $n_{1(2)}$, is the number of measurements on the recovered(control) material. The distribution of the measurements $\mathbf{y}_1$ and $\mathbf{y}_2$ on the recovered and the control items can be denoted by $f(\mathbf{y}_i \mid \psi_i)$, $\psi_i = \{\boldsymbol{\theta}_i, W_i\}$, where $\boldsymbol{\theta}_i$ represents the mean vector within source i and W_i the matrix of variances and covariances within source i.

The propositions of interest to the court may be the following:

H_p: The recovered (i.e., questioned) sample is from the same source as the control sample;

H_d: The recovered (i.e., questioned) sample is from a source that is different from that of the control sample,

where the subscript p stands, usually, for the prosecution's proposition, and the subscript d stands for the defence proposition. Statistical methods are often used to infer identity of a common source. Two fundamental ingredients are necessary for the evaluation of findings in forensic science: the probability distribution of the forensic results if proposition H_p is true, and the probability distribution of those results if proposition H_d is true. The question of interest is: given which of the competing propositions is the forensic result more reliable? The value of the evidence $\mathbf{y}_1$ and $\mathbf{y}_2$ is the ratio of two probability distributions under the two competing propositions:

$$LR = \frac{f(\mathbf{y}_1, \mathbf{y}_2 \mid H_p)}{f(\mathbf{y}_1, \mathbf{y}_2 \mid H_d)}. \tag{1}$$

The assessment of the value of evidence is typically considered to be in the domain of the forensic scientist's duties and the reported value of a likelihood ratio[2] implies either an increase or a decrease in the prior odds once forensic findings are taken into account. The likelihood ratio considers a particular case and answers the post-data question about how the evidence in the particular case alters the odds in favour of a particular proposition. In the numerator, where proposition H_p is assumed to be true, the model's parameters are assumed to be equal, say $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}$ and $W_1 = W_2 = W$ and the parameter vector takes the form $\psi = \{\boldsymbol{\theta}, W\}$. The marginal likelihood under proposition H_p can therefore be computed as:

$$f(\mathbf{y}_1, \mathbf{y}_2 \mid H_p) = \int f(\mathbf{y}_1, \mathbf{y}_2 \mid \psi, H_p)\pi(\psi \mid H_p)d\psi. \tag{2}$$

In the denominator, where proposition H_d is assumed to be true, the model's parameters are not equal and the marginal likelihood can therefore be computed as:

$$f(\mathbf{y}_1, \mathbf{y}_2 \mid H_d) = \int f(\mathbf{y}_1 \mid \psi_1, H_d)\pi(\psi_1 \mid H_d)d\psi_1 \int f(\mathbf{y}_2 \mid \psi_2, H_d)\pi(\psi_2 \mid H_d)d\psi_2. \tag{3}$$

The integrations above in (2) and (3) do not always have an analytical solution. Whenever the latter is not available, numerical procedures must be implemented to handle the complexity and to compute the marginal likelihood under the competing propositions.

Consider the two-level model that was proposed in Section 2. In some cases data present regular characteristics (e.g., symmetry or unimodality) that can reasonably be described using standard parametric models. If this happens, the variation at the two levels can be approximated by a normal distribution, that is one can take $\mathbf{X}_{ij} \sim \mathscr{N}(\boldsymbol{\theta}_i, W_i)$, and $\boldsymbol{\theta}_i \sim \mathscr{N}(\boldsymbol{\mu}, B)$ for the between-source variation, where $\boldsymbol{\mu}$ denotes the mean vector between sources and B denotes the matrix of variances and covariances between sources. If the variation within sources W_i can be reasonably assumed to be constant, that is $W_1 = W_2 = \cdots = W_m = W$ and is

[2] The term likelihood ratio will be used as a synonym of Bayes factor, to include the wider use of the first in forensic science applications, though it must be underlined that the Bayes factor does not always simplify to a ratio of likelihoods.

estimated from the available background data along with the mean vector $\boldsymbol{\mu}$ and the covariance matrix B between sources[3], the integrations above have an analytical solution[4] and it can be shown with some effort that the value of the evidence becomes the ratio of

$$f(\mathbf{y}_1,\mathbf{y}_2 \mid H_p) \propto \mid 2\pi \left[(n_1+n_2)W^{-1}+B^{-1}\right]^{-1} \mid^{\frac{1}{2}}$$
$$\exp\left\{-\frac{1}{2}\left[(\bar{\mathbf{y}}-\boldsymbol{\mu})'\left(\frac{W}{n_1+n_2}+B\right)^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu})+(\bar{\mathbf{y}}_1-\bar{\mathbf{y}}_2)'\left(\frac{W}{n_1}+\frac{W}{n_2}\right)^{-1}(\bar{\mathbf{y}}_1-\bar{\mathbf{y}}_2)\right]\right\} \tag{4}$$

against

$$f(\mathbf{y}_1,\mathbf{y}_2 \mid H_d) \propto \mid 2\pi B \mid^{-1/2} \prod_{i=1}^{2} \mid 2\pi(n_i W^{-1}+B^{-1})^{-1} \mid^{1/2}$$
$$\exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}_i-\boldsymbol{\mu})'(n_i^{-1}W+B)^{-1}(\bar{\mathbf{y}}_i-\boldsymbol{\mu})\right\}, \tag{5}$$

where $\bar{\mathbf{y}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{y}_{ij}$ and $\bar{\mathbf{y}} = \frac{1}{n_1+n_2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\mathbf{y}_{ij}$ (see Aitken and Lucy (2004)).

Nevertheless, there may often be practical situations where observations or measurements do not have such regular characteristics that make it suitable to use standard parametric models. The probability distribution can be estimated for each of the competing propositions by means of kernel density estimation, which is sensitive to multimodality and skewness and may provide a better representation of the available data. This approach is not novel in forensic science, and several applications can be found. It was proposed, for example, by Aitken and

[3] The within-source covariance matrix W, the mean vector between sources $\mu$and the between-source covariance matrix B can be estimated using the available background data, as

$$\begin{aligned}
\hat{W} &= \frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)', \\
\hat{\boldsymbol{\mu}} &= \sum_{i=1}^{m}\bar{\mathbf{x}}_i\frac{n_i}{n}, \\
\hat{B} &= \frac{1}{m-1}\sum_{i=1}^{m}n_i(\bar{\mathbf{x}}_i-\hat{\boldsymbol{\mu}})(\bar{\mathbf{x}}_i-\hat{\boldsymbol{\mu}})',
\end{aligned}$$

with $\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$.

[4] An analytical solution is also available whenever an additional level of variation alongside the levels above is considered (e.g., to take into account the measurement error), and it is supposed normally distributed. See Aitken et al. (2006) for details.

Lucy (2004) in the context of elemental composition of glass fragments, where the assumption of normality between sources was relaxed by introducing a kernel density estimate. Note that whenever the kernel density function is assumed to be normally distributed the integrations in (2) and (3) still have an analytical solution (as it is given in Aitken and Lucy (2004)). In a different context, it was proposed by Aitken and Taroni (2004) to classify banknotes as coming from drug trafficking rather than from general circulation on the basis of detected traces of cocaine.

The assumption of constant variation within sources makes the task of computing the marginal likelihoods in (2) and (3) much more feasible. However, while for some kind of trace evidence this assumption is sound (e.g., when available measurements consist of the elemental composition of glass fragments), there may be found forensic domains where a constant variability can be hardly justified, as it is the case for comparative handwriting to infer authorship in presence of questioned documents, since each writer can be characterized by a peculiar variation. To model prior uncertainty about the within-source parameters, a semi-conjugate model can be specified by choosing statistically independent prior distributions for the mean vector $\boldsymbol{\theta}$, and for the matrix of variances and covariances W, that is

$$\pi(\boldsymbol{\theta}, W \mid H_k) = \pi(\boldsymbol{\theta} \mid H_k)\pi(W \mid H_k), \tag{6}$$

where $\pi(\boldsymbol{\theta} \mid H_k)$ is taken to be normal as before, and $\pi(W \mid H_k)$ is taken of type Wishart-inverse distribution (Gelman et al., 2014). This model was proposed by Bozza et al. (2008) for a comparative handwriting scenario, with a Wishart-inverse distribution centred at the common within-source covariance matrix that was estimated from the available population database. Note that in this case, the assessment of the likelihood ratio for a given case is slightly more problematic, since the marginal likelihoods under the competing propositions can not be computed analytically. One may refer to numerical integration methods to provide an approximation of the marginal likelihood by means of numerical procedures. The marginal likelihood can however be approximated by a direct application of Bayes theorem (Chib and Jeliazkov, 2001), since it can be seen as the normalizing constant of the posterior density $\pi(\boldsymbol{\psi} \mid \mathbf{y}_1, \mathbf{y}_2, H_k)$, that is

$$f(\mathbf{y}_1, \mathbf{y}_2 \mid H_k) = \frac{f(\mathbf{y}_1, \mathbf{y}_2 \mid \boldsymbol{\psi}, H_k)\pi(\boldsymbol{\psi} \mid H_k)}{\pi(\boldsymbol{\psi} \mid \mathbf{y}_1, \mathbf{y}_2, H_k)}, \qquad k = \{p, d\}. \tag{7}$$

This is valid for any parameter point $\boldsymbol{\psi}$. So, if an opportune value $\boldsymbol{\psi}^*$ of $\boldsymbol{\psi}$ is chosen (e.g., the maximum likelihood estimate), the marginal likelihood can be approximated as

$$\hat{f}(\mathbf{y}_1, \mathbf{y}_2 \mid H_k) = \frac{f(\mathbf{y}_1, \mathbf{y}_2 \mid \psi^*)\pi(\psi^* \mid H_k)}{\hat{\pi}(\psi^* \mid \mathbf{y}_1, \mathbf{y}_2, H_k)}, \tag{8}$$

where the estimate of the posterior ordinate $\hat{\pi}(\psi^* \mid \mathbf{y}_1, \mathbf{y}_2, H_k)$ can be obtained by inspecting the output of a MCMC algorithm. This means that, starting from the proposed semi-conjugate model, under the assumption of normality alongside the levels above and non constant variability within sources, the posterior distribution

$$\pi(\boldsymbol{\theta}, W \mid \mathbf{y}) = \pi(\boldsymbol{\theta} \mid \mathbf{y})\pi(W \mid \boldsymbol{\theta}, \mathbf{y}) \tag{9}$$

can be estimated at the parameter point $\psi^* = (\boldsymbol{\theta}^*, W^*)$ by multiplying the estimates of the posterior densities $\hat{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ and $\hat{\pi}(W \mid \boldsymbol{\theta}, \mathbf{y})$ at the selected points $\boldsymbol{\theta}^*$ and W^*. The conditional densities being known (i.e., a normal distribution and a Wishart-inverse distribution, respectively), a Gibbs-sampling algorithm can be implemented and a Monte Carlo estimate of the posterior ordinates can be obtained in a straightforward manner from the Gibbs output (Bozza et al., 2008).

While the use of likelihood ratios (or, Bayes factors) for evaluative purposes is rather well established, presented and discussed in both theory and practice (Aitken and Taroni, 2004), focus on investigative settings still remains rather beyond considerations. The likelihood ratio represents a coherent metric for evidence assessment in general, but it can also be developed for investigative purposes, that is when no immediate suspect is available for comparison purposes. Investigative authorities (and ongoing investigations) may profit of valuable informations coming from the sole recovered items. For sake of illustration, consider the forensic examination of anonymous handwritten documents, that regularly arises in contexts where no suspect is available, and there will be no possibility for evaluating characteristics observed in a questioned document and those in reference (or control) material as it would be the case in a conventional evaluative scenario. Knowledge about the influence of demographic parameters, such as gender or handedness, may provide useful assistance in reducing the population of putative writers. Two propositions[5] may be considered:

H_1: the recovered item comes from population 1 (e.g., the population of left-handed writers);

H_2: the recovered item comes from population 2 (e.g., the population of right-handed writers).

[5] Note that these propositions need to be mutually exclusive, but not necessarily exhaustive.

Imagine a database is available with the handwriting features of individuals originating from population 1 and from population 2. A Bayesian statistical model can be introduced as before, with $f(\mathbf{x}_{ij} \mid \psi_i^k, H_k)$ representing the distribution of the available measurements given the hypothetical information that the item belongs to population k, and $\pi(\psi_i^k \mid H_k)$ measuring prior beliefs about population parameters ψ_i^k prior to observing data. Denote the available measurements on the recovered item to be classified by $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, where $\mathbf{y}_j = (y_{j1}, \ldots, y_{jp})$ and $\mathbf{y} \sim f(\mathbf{y} \mid \psi^k, H_k)$. The posterior probability of the competing propositions can be straightforwardly obtained by a direct application of Bayes theorem

$$\Pr(H_k \mid \mathbf{y}) = \frac{\Pr(H_k) f(\mathbf{y} \mid H_k)}{\sum_{k=1}^{2} \Pr(H_k) f(\mathbf{y} \mid H_k)},$$

where

$$f(\mathbf{y} \mid H_k) = \int f(\mathbf{y} \mid \psi^k, H_k)\pi(\psi^k \mid H_k) d\psi^k \tag{10}$$

is the predictive distribution.

The classification problem can be seen as a special case of decision making, where decision d_i can be formalized as: *the recovered item is to be classified in population i.* A coherent classification procedure would suggest to make the decision d_i that allows to minimize the posterior expected loss, that is

$$\min_i EL(d_i) = \sum_k L(d_i, H_k) \Pr(H_k \mid \mathbf{y}), \tag{11}$$

where $\Pr(H_k \mid \mathbf{y})$ is the posterior probability of proposition H_k and $L(d_i, H_k)$ represents the loss of classifying in population i (d_i) an item belonging to population k (H_k), $i \neq k$. The loss is zero whenever a correct decision is taken (i.e., $i = k$). The optimal decision is therefore to classify a recovered item in population 1 whenever the expected loss of decision d_1 is smaller than the expected loss of decision d_2:

$$L(d_1, H_2) \Pr(H_2 \mid \mathbf{y}) < L(d_2, H_1) \Pr(H_1 \mid \mathbf{y}). \tag{12}$$

Rearranging terms, and dividing both sides by the prior odds,

$$\frac{\Pr(H_1 \mid \mathbf{y})}{\Pr(H_2 \mid \mathbf{y})} \Big/ \frac{\Pr(H_1)}{\Pr(H_2)} > \frac{L(d_1, H_2)}{L(d_2, H_1)} \Big/ \frac{\Pr(H_1)}{\Pr(H_2)}, \tag{13}$$

one obtains a threshold for the interpretation of the Bayes factor: a recovered item is classified in population 1(2) whenever the Bayes factor is larger(smaller) than the quantity at the right-hand side in (13). The advantage is that there is not only a decision as to which population the recovered item can be classified, but also through the Bayes factor there is a measure of the strength of the conclusion. This classification criteria was proposed by Bozza et al. (2014) to classify two-class Cannabis seedlings. The necessity to choose a prior probability for competing propositions and a loss function to assess the undesirability of misclassification may be felt as a struggling issue as there is not an ad-hoc recipe. Subjectivity do often have a connotation of arbitrariness, and forensic science is not an exception. Probabilities depend on one's extent of knowledge, may change as the informations change and may vary amongst individuals, as well as there is not a 'correct' loss function, since each individual will have his own system of preferences. Personal degrees of belief also enter the enumeration of the Bayes factor through the elicitation of the prior probability distributions. Finally, as it was highlighted for the evaluative scenario, the assessment of a Bayes factor can be more or less feasible depending on the specific setting of interest, as the integral in (10) may be analytically intractable.

There is actually an ongoing discussion in the forensic community whether a forensic scientist should report to the court a single value of the Bayes factor or a range of values to acknowledge for uncertainty in its ratio assessment (Taroni et al., 2015). It is not uncommon in fact to encounter forensic scientists who argue the need to determine the probability distribution of a given expression of evidential value, or to fit an interval on such an expression. These ideas will be addressed in Section 4 with reference to a comparative handwriting scenario.

4. A CASE STUDY: LIKELIHOOD RATIO FOR ASSESSING HANDWRITING EVIDENCE

Handwriting examination involving questioned documents consists in describing handwriting features, such as elements of style or elements of execution, and studying their range of variation. Characterization of writing habits is largely dependent on the experience of the document examiner, who usually evaluate the handwriting features in a qualitative or subjective way. Various studies have already been undertaken to partially automate the analysis process and support the examiner. Among these, an image analysis procedure has been developed and tested by Marquis et al. (2005) to quantify and provide a global description of handwriting features. According to the proposed technique, each contour loop can be expressed by a set of p variables representing the first four pairs of Fourier coefficients.

Suppose that an anonymous document is available for comparative purposes.

A suspect is apprehended, and written material from this suspect is selected and will be analyzed to infer common identity. A number n_1 of measurements are performed on the anonymous manuscript (these will be referred as recovered data). A number n_2 of measurements are performed on a manuscript originating from the suspect (these will be referred as control data). For illustration, consider the following two propositions of interest:

H_p: the suspect is the author of the manuscript;

H_d: the suspect is not the author of the manuscript, an unknown person is the author of the manuscript.

Starting from the evidence, and assuming a two-level model as the one defined above in Section 3 with the assumption of normality for both levels of variation and a Wishart-inverse distribution for the within-source variability, the likelihood ratio is computed as a ratio of two estimated marginal density ordinates as in (8): one for the numerator, where proposition H_p is supposed to be true, and one for the denominator, where proposition H_d is supposed to be true. Imagine a value equal to, say, 125 is obtained, supporting hypothesis H_p. Such a value is not substantially different from 120 or 130! What is the information that the assessed number does really convey? Recalling one of the fundamental laws of handwriting according to which no one writer writes the same word exactly the same way twice (the so called *within-writer variability*), it may be observed that the reported likelihood ratio is sensitive to the shape's variability of handwritten characters. So, to what extent can a forensic examiner rely upon a case-specific likelihood ratio? To approach this question, one may consider the related question: "How often may the document examiner obtain a likelihood ratio larger or smaller than 1 for handwriting evidence originating from the same source?". In the same way, "How often may the document examiner obtain a likelihood ratio smaller or larger than 1 whenever the questioned documents do not origin from the alleged writer?". To provide an answer, several pairs of observations for each setting of interest (i.e., the competing propositions H_p and H_d) may be selected from the background population[6], and for every pair the likelihood ratio can be

[6] The handwriting of 100 writers from the School of Criminal Justice of the University of Lausanne was collected, and the contour shape of several characters was analyzed and described by Fourier coefficients according to the methodology that was described in Marquis et al. (2005). In other forensic domains, data can be compiled through practical experiments because target materials and substances are easily available (Aitken and Lucy, 2004). Pairs of observations can also be generated, e.g. in applications where the distribution of the source features is known as for kinship identification scenario (Corradi and Ricciardi, 2013).

assessed.

To test hypothesis H_p: *the suspect is the author of the manuscript'*, several groups of measurements can be randomly selected from the same writer to act as recovered and control data, and for each draw the likelihood ratio can be computed. So, imagine for sake of illustration that a writer is to be selected from the available database, denote it as *writer 1* ($w1$), and that several draws of characters of type *a* are extracted as it was previously described. Accordingly, the selected measurements are randomly divided into two groups: values coming from the first group are denoted as measurements from a recovered manuscript, whereas values coming from the second group are denoted as measurements from a control sample. In the same way, a second writer is selected, denote it as *writer 2* ($w2$), and again groups of measurements are randomly selected to act as recovered and control data. Results are summarized in Figure 1, where the logarithm of the likelihood ratios obtained in correspondence of one thousand of draws of measurements of character *a* originating from *w1* and *w2* is displayed. The variability for each writer can be less or more pronounced, as one might reasonably expect because of the non constant within-writers variability.

In the same way, to test hypothesis H_d: *the suspect is not the author of the manuscript, an unknown person is the author of the manuscript'*, several groups of measurements originating from different writers can be selected to act as recovered and control data, respectively. For sake of illustration, suppose that *w1* is selected to act as the author of the questioned anonymous manuscript, while *w2* is selected to act as the author of the control document: several draws are extracted from each of them, and for each draw a likelihood ratio is assessed. Results are summarized in Figure 2, where it is displayed the logarithm of the likelihood ratios obtained in correspondence of one thousand of draws of characters *a*. Clearly, one may expect a negative log-ikelihood ratio, though with a variable magnitude according to the selected writers.

In Figure 2 there are also summarized the log likelihood ratios that can be assessed whenever *w2* is substituted by a third writer, *w3*, to act as the author of the control document. The triplet was chosen for illustrative purposes because of shape's similarities between *w1* and *w3* (i.e., both presenting rounded characters), and because of the pronounced dissimilarities between *w1* and *w2*, the last one being characterized by handwritten characters with substantial elongation toward the right. Though this is only an example, it is interesting to observe that as the separation between control and recovered measurements increases, the value of the evidence decreases.

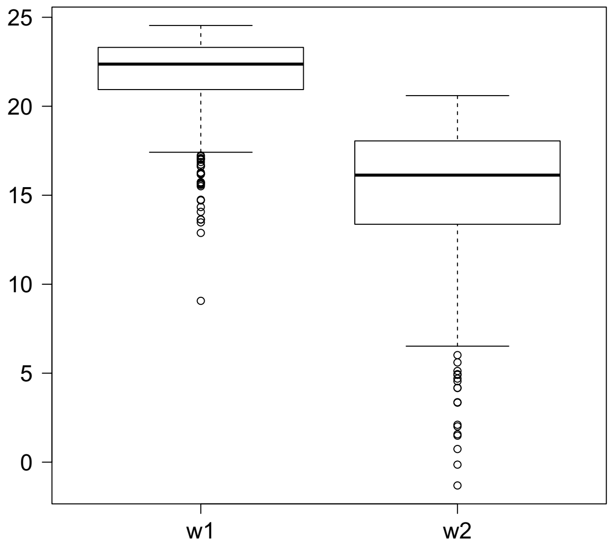

Figure 1: Log-likelihood ratios for several draws of measurements of character *a* extracted from *writer 1* to act as control and recovered data (left); log-likelihood ratios for several draws of measurements of character *a* extracted from *writer 2* to act as control and recovered data (right).

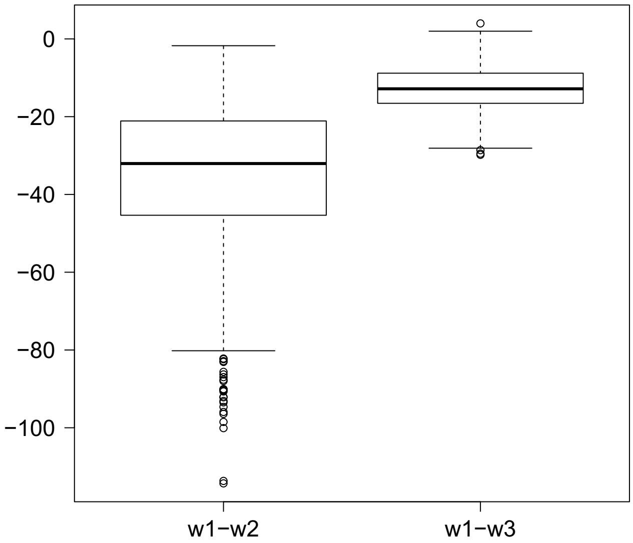

Figure 2: Log-likelihood ratios for several draws of measurements of character *a* extracted from *writer 1* to act as recovered data, and from *writer 2* to act as control data (left); log-likelihood ratios for several draws of measurements of character *a* extracted from *writer 1* to act as recovered data, and from *writer 3* to act as control data (right).

A recovered anonymous document may also be used for investigative purposes to reduce the pool of potential writers. Several draws may be taken from each setting of interest (i.e., propositions H_1 and H_2) to infer the discriminative capacity of the proposed approach. It must be said that the contour shape alone appeared in general unable to infer the gender or the handedness of an anonymous writer, with a non negligible percentage of misclassified that do not actually allows to transpose the current methodology at an operational level (Taroni et al., 2014). Still, the Bayesian approach offers a logical framework also for investigative settings, where no suspect is available for comparative purposes, though other types of reference material may be needed to discriminate between populations of interest.

5. DISCUSSION AND CONCLUSIONS

Continuing developments in science and technology provide an increasing amount of information at the forensic scientist's disposal during criminal investigations and it is fundamental that the evidential strength of available results is derived reliably, so that the justice system can take advantage of this. The assessment of value of scientific evidence requires probabilistic and statistical reasoning and improved methods are necessary to deal with all sort of uncertainties and complexities that are inevitably associated with forensic scenarios.

Simulation studies may be extremely valuable to inform the court about the robustness of the proposed statistical methodologies. A 'likelihood ratio distribution' can be obtained when the analyzed findings come from the same source or from different sources to quantify how often a likelihood ratio taking values in a given range can be obtained. This would be valuable to quantify how often a likelihood ratio points in the wrong direction (i.e., giving rise to false negatives or false positives).

The assessment of the magnitude of false positives and false negatives for a given setting may therefore be informative about the potential of misleading evidence to investigate the discriminative capacity of the proposed methods with respect to propositions at hand. Such information is obtainable before findings are made and is independent of the observations made in a given case. The answers to these questions are not of particular relevance for the evaluation of evidence in a particular case. The fact that if one were to take another sample one's resulting marginal likelihood would be different, and the likelihood ratio too, is uncontroversial. However, this is of no detriment to the assessment of one single value to report for the given scenario of interest. This is an objective assignment: two forensic examiners with the same statistical model, the same prior probability distributions, and the same measurements, will provide the same value. Arguing otherwise would bear the risk of embracing ideas that care about possible values that have not actually been observed, which is in contradiction with the likelihood principle according to which once data are observed, no other values matter, and the hypothetical extreme values that might have been observed are irrelevant. Undoubtedly, the reported value is based on all available knowledge at a given instant of time: there may be uncertainty because it may change, and not least because of the complexity of the model that could make it necessary numerical procedures (Alberink et al., 2013). The exploration of a MCMC output, for example, does not provide a sample from a hypothetical likelihood ratio distribution. By treating the resulting draws as single scores for the questions of interest would amount equating knowledge about the posterior distribution to knowledge about the marginal likelihood the examiner is required to report.

To conclude, it may be felt that a likelihood ratio approach to measure the value of evidence should only be restricted to forensic domains where a large background information is available, as in DNA evidence. There are cases where a limited knowledge about the type of evidence is available (Nordgaard and Rasmusson, 2012), as in the handwriting scenario where a document examiner observes similarities and dissimilarities between a questioned document and reference material. The likelihood ratio may be hard to specify, and it will vary according to available evidential value. This does not prevent scientists from mentioning that their beliefs may be based on a limited amount of data. The scale of the likelihood ratio may be inevitably rough and not sufficient to take a decision, nevertheless the evidence may be quite compelling in support of one of the alternative propositions.

6. ACKNOWLEDGEMENTS

The author gratefully acknowledges the support of the Swiss National Science Foundation through grant No. BSSGI0_155809.

The author gratefully acknowledges Dr Raymond Marquis from the School of Criminal Justice of the University of Lausanne for the data that has been used in Section 4.

REFERENCES

Aitken, C.G.G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. In *Applied Statistics*, 53: 109–122.

Aitken, C.G.G., Lucy, D., Zadora, G. and Curran, J.M. (2006). Evaluation of transfer evidence for three-level multivariate data with the use of graphical models. In *Computational Statistics & Data Analysis*, 50: 2571–2588.

Aitken, C.G.G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley and Sons, Inc., Chichester, UK, 2 edn.

Alberink, I., Bolck, A. and Menges, S. (2013). Posterior likelihood ratios for evaluation of forensic trace evidence given a two-level model on the data. In *Journal of Applied Statistics*, 40: 2579–2600.

Bolck, A., Ni, H. and Lopatka, M. (2015). Evaluating score- and feature-based likelihood ratio models for multivariate continuos data: Applied to forensic MDMA comparison. In *Law, Probabiltiy & Risk*, 14: 243–266.

Bozza, S., Broséus, J., Esseiva, P. and Taroni, F. (2014). Bayesian classification criterion for forensic multivariate data. In *Forensic Science International*, 244: 295–301.

Bozza, S., Taroni, F., Marquis, R. and Schmittbuhl, M. (2008). Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. In *Applied Statistics*, 57: 329–341.

Chib, S. and Jeliazkov, S. (2001). Marginal likelihood from the MetropolisHastings output. In *Journal of the American Statistical Association*, 96 (453): 270–281.

Corradi, F. and Ricciardi, F. (2013). Evaluation of kinship identification systems based on short tandem repeat DNA profiles. In *Applied Statistics*, 62: 649–668.

ENFSI (2015). Guidelines for evaluative reporting in forensic science.'*Tech. rep.*, European Network'of Forensic Science Institutes, (www.enfsi.eu).

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2014). *Bayesian Data Analysis*. CRC Press, Boca Raton, 3 edn.

Lindley, D.V. (1991). Probability. In *The Use of Statistics in Forensic Science*, Aitken, C.G.G. and Stoney, D.A. (Eds), Ellis Horwood, New York, pp. 27–50

Marquis, R., Schmittbuhl, M., Mazzella, W. and Taroni, F. (2005). Quantification of the shape of handwritten characters loops. In *Forensic Science International*, 164: 211–220.

Nordgaard, A. and Rasmusson, B. (2012). Likelihood ratio as value of evidence: more than a question of numbers. In *Law, Probability & Risk*, 11: 303–315.

Taroni, F., Bozza, S., Biedermann, A. and Aitken, A. (2015). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. In *Law, Probability & Risk*, 15: 1–16.

Taroni, F., Marquis, R., Schmittbuhl, M., Biedermann, A. and Bozza, S. (2014). Bayes factor for investigative assessment of selected handwriting features. In *Forensic Science International*, 242: 266–273.