Post-print (Final draft post-refereeing)

Final version published in Science & Justice 57 (2017) 209-220

Accepted Manuscript

Towards a Bayesian evaluation of features in questioned handwritten signatures

Lorenzo Gaborini, Alex Biedermann, Franco Taroni

PII:	S1355-0306(17)30004-7
DOI:	doi:10.1016/j.scijus.2017.01.004
Reference:	SCIJUS 643
To appear in:	*Science & Justice*
Received date:	20 June 2016
Revised date:	10 January 2017
Accepted date:	22 January 2017

Please cite this article as: Lorenzo Gaborini, Alex Biedermann, Franco Taroni, Towards a Bayesian evaluation of features in questioned handwritten signatures, *Science & Justice* (2017), doi:10.1016/j.scijus.2017.01.004

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Towards a Bayesian evaluation of features in questioned handwritten signatures

Abstract

In this work we propose the construction of a evaluative framework for supporting experts in questioned signature examinations. Through the use of Bayesian networks, we envision to quantify the probative value of well defined measurements performed on questioned signatures, in a way that is both formalised and part of a coherent approach to evaluation.

At the current stage, our project is explorative, focusing on the broad range of aspects that relate to comparative signature examinations. The goal is to identify writing features which are both highly discriminant, and easy for forensic examiners to detect. We also seek for a balance between case-specific features and characteristics which can be measured in the vast majority of signatures. Care is also taken at preserving the interpretability at every step of the reasoning process.

This paves the way for future work, which will aim at merging the different contributions to a single probabilistic measure of strength of evidence using Bayesian networks.

Keywords: Bayesian networks, signature evidence, Fourier descriptors, multivariate likelihood ratio

1. Introduction

Handwritten signatures have been employed since centuries as a means of authenticating one's identity on official documents. Their study has been one of the oldest disciplines in forensic science, yet its evaluative part did not achieve the same level of refinement as others.

Several professional groups, such as questioned document examiners, are trained to testify in courts by following established examination protocols for handwritten signatures. However, the usage of handwritten evidence in courts raises a number of issues. The scientific foundations of forensic handwriting comparisons are regularly doubted, in particular the mechanism by which forensic examiners arrive at and state their conclusions is often questioned. Specifically, the evaluation process is highly expert dependent and does not rely on standardized measurements and lines of reasoning, being thus highly dependent on the skill and proficiency of each examiner.

The use of forensic science in legal proceedings is based on the so-called "evaluative" framework: instead of stating a probability for a hypothesis, forensic experts report an expression of strength of support against two competing hypotheses, of forensic and legal interest[1]. To help evaluate the strength of support, a likelihood ratio is used in or-

ACCEPTED MANUSCRIPT

der to formalise the reasoning of the expert with respect to the relevant scientific findings.

The advantages of this evaluative framework are multiple: while formalised reasoning is much less liable to logical fallacies, experts will not express their beliefs on matters for which a court is responsible, notably on the hypotheses of interest. Further, the approach clarifies that the probability of hypotheses of interest also depends on information other than the scientific findings, allowing thus legal decision makers to incorporate in their reasoning a broad range of collateral case information.

1.1. The defence hypothesis

In forensic science, case-based evidence is collected and assessed under at least two competing hypotheses, those of the prosecution and the defence. In the domain of comparative forensic document examination, evidence takes the form of observed similarities and differences between questioned and reference ("known") items. To assess its value with respect to the competing hypotheses, the forensic scientist needs to evaluate the rarity of such similarities and differences in a given population of potential writers.

The choice and the size of the relevant population is of utmost importance, as it is very easy to overestimate the relevance of a character trait if it is shared by many or all the users of a determinate writing system [2]. For example, one may compare the writing features of a questioned item against those of two individuals, a number of suspects, or any other set of potential writers. If the relevant population spans a restricted number of individuals, the comparison is said to be a "closed-set": on the other hand, if the population at large is considered, the situation is labelled "open-set" [3].

As a result, the value of the scientific findings strongly depends on their rarity in the reference population, though some traits might be more discriminating between two individuals rather than among a broader group of writers.

In this article we mostly focus on closed-set circumstances, leaving the possibility to extend to open-set situations in future works.

1.2. Elements of Bayesian networks

To depict the reasoning using the previously illustrated intepretative framework, consider a single variable H which can assume two mutually exclusive states H_p and H_d, respectively the prosecution and the defence hypothesis. In a questioned signature examination scenario, we may associate e.g., H_p = "Person A has written the questioned signature" and H_d ="An unknown person has written the questioned signature". Let E be the set of findings, as detected by the expert (e.g., similarities and differences between the questioned signature and the reference specimens). We denote with I the background information on the case, available to the expert.

Relevant to the recipient of expert information are the prior beliefs on H_p and H_d, conditioned by the background information: these are the probabilities $\Pr(H = h_p \mid I)$ and $\Pr(H = h_d \mid I)$, respectively[1]. More precisely, their ratio (called *prior odds*) is the relative strength of belief in H *a priori*.

[1] The reason for which h_p and h_d are written in lower-case letters is explained in Section 3.

ACCEPTED MANUSCRIPT

The role of the expert is to evaluate the probability of having observed E under h_p and h_d: these terms are $\Pr(E \mid H = h_p, I)$ and $\Pr(E \mid H = h_p, I)$, respectively.

Bayes' theorem then states that the relative strength of belief in H *a posteriori* is proportional to the prior odds. In formulae:

$$\frac{\Pr(H = h_p \mid E, I)}{\Pr(H = h_p \mid E, I)} = \mathrm{LR}\,\frac{\Pr(H = h_p \mid I)}{\Pr(H = h_d \mid I)}$$

where

$$\mathrm{LR} = \frac{\Pr(E \mid H = h_p, I)}{\Pr(E \mid H = h_p, I)}$$

is called *likelihood ratio*. We observe that LR provides the expression for the strength of support of E versus the considered H: if $\mathrm{LR} > 1$, E provides more support to h_p rather than h_d, conditioned on the background information, and vice versa. Notice that it does *not* imply that h_p is more probable than h_d. To dissect the definition of the LR, the numerator reads as the probability of having observed E under h_p: referring to the previous example, it amounts to asking "what is the probability of observing the set of concordances and discordances in genuine signatures of Person A?". The denominator, instead, is the probability of observing the same set of findings in signatures that appear to belong to Person A, but instead have been forged by someone else: this is assessed using the relevant population, defined in Section 1.1. In other terms, the LR is the ratio of two probabilities that account for, respectively, the *intra-* and *inter-* variability of findings.

Note that to apply the evaluative framework, one needs to specify not only the numerical values for the beliefs, but also the dependences between the variables in terms of conditional probabilities. This can be intuitively represented in a graphical notation, which enables the forensic scientist to consider cases with multiple variables with differing interdependence. The obtained graphs are named *Bayesian networks* [4]: for instance, the previous example can be represented in Figure 1. Note that the background information I is used to condition all relevant probabilities associated to H and E. For the sake of simplicity, such dependence is generally just assumed without a clear representation in the network. As a consequence, information I is usually omitted from explicit graphical representation in Bayesian networks.

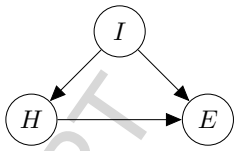


Figure 1: The example in Section 1.2 as a Bayesian network. Note that the node I is usually omitted.

Bayesian networks are very flexible, and have been used to support evaluative reasoning in very different forensic branches such as firearms [5], printed documents [6], signatures [7], forensic medicine [8] and DNA [9]. A review on the usage of Bayesian networks in forensic science can be found in [10].

1.3. Hierarchical evidence evaluation

A Bayesian network can be built for a rather generic evaluative procedure (e.g., the two-trace problem in [11]), but its structure can also be modified in order to accommodate for missing evidence

ACCEPTED MANUSCRIPT

[10]. Specifically, a binary-valued node M can be added to the network, encoding the fact that some evidence E is expected, but has not been retrieved (i.e., is missing). The usual assessments on the probability of E are now conditional on M.

The flexibility of Bayesian networks allows us to assess the probative value in complex cases, using evidence which is more and more case-based. In the works described later on, we envision to build a useful model to capture the essential features of the process of signature comparison, to help evaluating evidence which can always be measured (e.g., physical dimensions of the signatures) up to specific traits of one's signature (e.g., inner angles), which might not always be recognisable.

2. Related work

2.1. Pattern Recognition literature

A large amount of work has been recently done in building automatic classifiers for signatures. Most often, they exploit features which are not easy to describe (either empirically or mathematically), or have limited forensic interest. Such systems can be designed to work in a multi-writer environment, with known or unknown sources. A number of literature review papers are available, such as [12, 13, 14].

Several international competitions of automated forensic handwriting analysis (AFHA) have also been organised, mostly related to ICDAR conferences: typically, they required participants to test their working systems on a common signature database, built specifically for the contest.

The literature distinguishes between studying *online* signatures (i.e., whose speed, acceleration and stroke-order data is recorded in real-time) and *offline* signatures (i.e., where only the finished written signature is studied) [14], but also classifies the signatures according to their "grade" of forgery. Since an automatic approach is sought, researchers frequently aim at distinguishing one's signature from another's, with no attempt at forgery (i.e., *random forgery*) [12]. If the forger is aware of the victim's name but signed in his own style, we talk about *blind forgery*; if the forger has obtained some genuine specimens and has actively tried to reproduce the victim's signature, one talks about *skilled forgery*.

However, our goal is not to build an entirely automatic system to classify signatures (which is the goal of AFHA discipline), but to find a means to support the forensic document examiner in his evaluation. A fully automatic approach would be useless to him, since, for example, by his expertise he would also consider a number of characteristics which are related to the context of the contested specimens (e.g., type of instrument, writing position, nature of document, secondary traces such as blood or fingermarks, and reasons behind a forgery): an automatic approach cannot deal with these characteristics, since they are necessarily tied to a specific forensic case. Moreover, works in Pattern Recognition literature achieve very high classification rates (see Table V in [14]), but do so in a not forensically interesting environment, both due to the usage of different hypotheses (e.g., writing position, type of instrument and psychological conditions) and to the way in which conclusions are

ACCEPTED MANUSCRIPT

reported. They also typically make no reference to the chosen population under the defence hypothesis (see Section 1.1).

To reconnect with the Pattern Recognition nomenclature, the aim of this work is to study skilled forgeries in an off-line setting. However, the choice of features which will be examined and their integration into the evaluative framework is very different.

2.2. Forensic handwriting examination literature

As stated before, forensic handwriting examination is one of the oldest disciplines of forensic science. A great deal of literature is available on the subject, suggesting a number of traits of signatures and handwriting to be considered for evaluation: for example see [2, 15, 16, 17] and [18]. However, few of these works actually report quantitative means for obtaining measurements, and those who do have been often targets of criticism (see [19] and [20]). Scarce progress was made in the quest for the formalisation of handwriting examinations until very recent years, with the advent of increasing computational power, refined mathematical techniques and sophisticated forensic tools. Among other works, we highlight [21], [22] and [23].

Here, we chose to follow the initial part of [21] on signatures on paintings, for multiple reasons. Her work strives to achieve the same goals as ours, albeit in different contexts. The analysis of painted signatures deeply draws from the classical expertise in forensic document examination, on which our work will be based: as a consequence, the adopted features are related only to distances and angles inside each signature. The scope of [21] is mainly quantitative, and is aimed at building an evaluative procedure under the Bayesian framework.

In [23], the authors illustrated a method to discriminate writers based on the shape of the capital letter "O", further extended to other characters in later works [24, 25, 26]. Specifically, loops of closed characters are described in terms of harmonic content of the shape contour, being thus easily classifiable by a set of numbers separating, e.g., "small, shaky and elongated loops" from "round and smooth circles". This method is striking in its match between simplicity and visual effectiveness, therefore we decided to adopt it as a part of our framework.

3. Notation

In this section we define the notation that will be used in this work.

Where possible, upper-case letters will denote random variables (e.g., H), whose realisations are indicated with lower-case letters (e.g., $H = h$). Vectors will be underlined (e.g., $\underline{X}$). We may also obtain a number of observations from a single random variable: the i-th realisation will be indicated with a parenthesized superscript. For example, $\underline{s} = (s^{(i)})_{i=1}^{n}$ is a collection of n realisations of the random variable S.

In forensic literature it is customary to specify the competing hypotheses by means of the variable H taking values in $\{H_p, H_d\}$, respectively for the prosecution and the defence. In light of previous remarks, H_p and H_d will then be indicated in lower-case notation (i.e., h_p and h_d).

In this work several signatures of a single per-

ACCEPTED MANUSCRIPT

son, denoted A, were analysed. In addition, a number of forged specimens were provided by several forgers. The latter are anonymously denoted by F_1, F_2, F_3, F_4. The set of all forgers is denoted with F_{All}. We indicate the writer of the i-th specimen with the variable $w^{(i)}$, taking values in $W_{\text{All}} = \{A\} \cup F_{\text{All}}$. Notice that w is supposed to be known, but it will be used only for ease of notation rather than inference.

Further definitions about the acquired corpora are given in Table 1 and Table 2.

Writer	N°	Description
A	$n_A = 143$	Authentic corpus
F_1	$n_{F_1} = 35$	
F_2	$n_{F_2} = 20$	
F_3	$n_{F_3} = 21$	
F_4	$n_{F_4} = 20$	
$F_{\text{All}} = \{F_1, F_2, F_3, F_4\}$	$n_F = 96$	Forged corpora
$W_{\text{All}} = \{A, F_1, F_2, F_3, F_4\}$	$n = 239$	Full corpus

Table 1: Composition and notation of the corpora.

4. Methodology

4.1. Signature acquisition: the corpora

143 signatures of a single person A were collected over the period of a month, writing in small batches to avoid hand fatigue and adaptation. All signatures have been written with a black ball-point pen on unruled white paper in a normal sitting position. As [27] reports, the absolute size of writing is deeply affected by available space, hence enough space has been left on the sides to prevent any restraining effect.

Forged samples were provided by 4 other persons, with no past experience in signature forgery and document examination. Each forger received the authentic corpus, and was instructed to practice forgeries to one's liking. Each forger reproduced at least 20 forgeries each over a week period, with a black pen in a normal sitting position. Tracing has been forbidden.

All collected signatures have been digitalized at 600 dpi and saved in an uncompressed grayscale format. The composition of the corpora is summarised in Table 1.

4.2. Signature processing

As some features require the extraction of the contour and pixel values, each image needs to be further processed. Specifically, each signature is isolated from the background by means of a combination of morphological operators [28]. Consequently, each pixel of the image is either part of the background or of the signature.

As no guideline is provided onto the writing paper, signatures are rotated in order to compensate for the baseline inclination, evaluated by inspection. However, there were no cases in the corpora with extreme slant. This step is not necessary if only distance-based features (e.g., length of a particular ascender, or width of a word) are retained.

The implementation of [23] necessitates to locate closed loops in signatures. For each signature in the corpora, at most 5 loops were identified, each one attributable to the shape of single characters ("a", "b", "o"). Loops were closed by hand if a closure did not significantly alter their shape (i.e., very small opening, or the closing is strongly suggested by the surrounding character traits), while missing

ACCEPTED MANUSCRIPT

loops (i.e., either open or not present at all) were labelled as such. After normalizing the area of each loop, the contour was separated using morphological operators, and a vector of Fourier descriptors was extracted. 4 harmonics were retained, thereby producing 8 pairs of values (amplitude and phase) for each loop. Harmonics greater than the fifth order were discarded, as their shape contribution was visually negligible.

4.3. Keypoints and measures

Following [21], 19 keypoints were hand picked in order to identify features which are both relevant to forensic examination, and easy to locate in the authentic signatures. The keypoints were afterwards matched in all corpora by human inspection. It is worthwhile noting that a given keypoint can fail to be present on a specimen: this information has been retained rather than discarded, and provides a further novelty in future works with respect to [21].

The list of chosen keypoints is graphically represented in Figure 2, along with two samples from the corpora.

The set of keypoints serves as a basis to obtain measurements such as distances, angles, and ratios of distances. Such a list of measures is reported in Table 2. To reduce the huge amount of measurements which can be extracted, we decided to initially focus only on those which can be reliably transposed to other kinds of signatures. Nevertheless, as mentioned in Section 1.3, features which are highly signature-dependent (such as $\Theta_1, \Theta_2, \Theta_3, \Phi_1, \Phi_2$) can be integrated in the Bayesian network to provide further support to (or against) the hypotheses.

5. Results

5.1. Absolute dimensions

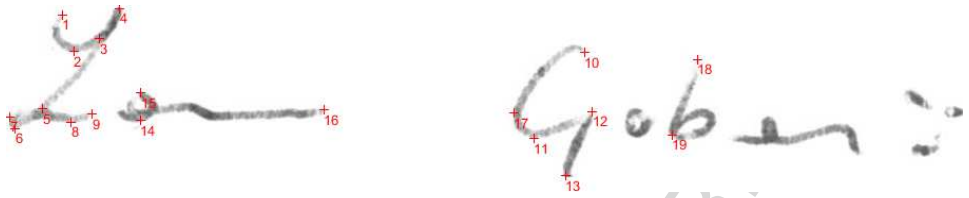
The simplest analysis which can be done is to study the absolute dimensions of signature samples, S_1 and S_2. Collected data is represented in Figure 3. It is evident that layman forgers were disregarding the absolute dimensions of the specimens, thereby producing forgeries with high probative value in favour of the hypothesis of forgery (rather than authenticity). The reasons are multiple: first of all, specimens were provided in an electronic format rather than in a printed form, impairing absolute visual comparisons between one's forgery and a genuine sample. Also, there were no visual guides on the document, to reduce the aforementioned restraining effect. Hence, the detected differences are tied to the mode of presentation of specimens to forgers.

We nevertheless chose to report this analysis, as the method is applicable to any set of quantitative measurements performed on every specimen. As an example, we may repeat the same analyses on relative measurements (i.e., measurements which are independent of absolute sizes).

From Figure 3, the visual clustering of signature dimensions suggests us to treat them as samples from k bivariate Gaussians, where $k = 5$ (i.e., number of writers in the corpora).

According to Bayes' theorem (see Section 1.2), in the numerator of the LR, the variability of the signature dimensions is assessed in the genuine signatures. In a real-world forensic context, as stated in Section 1.1, one needs to specify the defence hypothesis, hence the population of interest (denoted

ACCEPTED MANUSCRIPT



(a) An authentic signature.

(b) A forged signature. Notice that keypoint #9 is missing.

Figure 2: Two signatures in the corpora, with marked keypoints.

Symbol	Name	Domain	Formula
S_1	Signature width	$[0, \infty)$	-
S_2	Signature height	$[0, \infty)$	-
D_1	Absolute word spacing	$[0, \infty]$	$\lvert H_{17} - H_{16} \rvert$
D_2	Relative word spacing	$[0, 1]$	D_1/S_1
D_3	Initial height	$[0, \infty)$	$\lvert \min(V_1, V_3) - \max(V_7, V_8, V_9) \rvert$
D_4	Small caps height	$[0, \infty)$	$\lvert H_{14} - H_{15} \rvert$
D_5	Caps height ratio	$[0, 1]$	D_4/D_3
D_6	Absolute first name length	$[0, \infty)$	$\lvert H_{16} - H_7 \rvert$
D_7	Relative first name length	$[0, 1]$	D_6/S_1
Θ_1	Angle inside upper loop of L	$[0°, 360°)$	Angle between keypoints 1, 2, 4
Θ_2	Angle between ascender of L and keypoint 9	$[0°, 360°)$	Angle between keypoints 3, 5, 9
Θ_3	Angle inside G	$[0°, 360°)$	Angle between keypoints 11, 12, 13
Φ_1	Slant of ascender of L	$[0°, 360°)$	Slope of line between keypoints 3, 5
Φ_2	Slant of descender of G	$[0°, 360°)$	Slope of line between keypoints 12, 13

Table 2: List of keypoint-based measures. Formulae are omitted with angles and slants. H_i (V_i) denotes the horizontal (vertical) distance in mm between the i-th keypoint and the top-left corner of the signature. Please note that H_i is not related to the hypothesis H.

ACCEPTED MANUSCRIPT

with $\mathcal{F}$). The probability of the evidence under the defence hypothesis appears as the denominator of the LR.

We consider the following hypotheses: $H = h_p =$ "The questioned signature has been written by A" vs $H = h_d =$ "The questioned signature has been written by someone in $\mathcal{F}$".

If two writers need to be compared, our corpora are able to produce 4 kinds of comparisons with the reference material, one for each forger (i.e. $\{A \text{ vs } F_i\}_{i=1}^4$). In that case, $\mathcal{F}$ is represented by the corpus produced by the considered forger.

If the number of potential writers remains unknown, however, the 5-class comparison problem becomes a two-class comparison problem, i.e., "authentic" ($H = h_p$) vs "forged" ($H = h_d$). In the latter case, $\mathcal{F} = \{F_1, F_2, F_3, F_4\} = F_{\text{All}}$: i.e., under h_d we state that a signature has been produced by one of the $\mathcal{F}$. Notice that we consider writers in $\mathcal{F}$ as being "indistinguishable" from each other, as they are grouped together under a single distribution, the one under h_d. In other words, we consider forged specimens as having been written by a single "virtual" writer F_{All}. We also observe that, under the latter, the Gaussian structure under h_d cannot be assumed in our corpora.

To simulate a real-world comparison procedure, we performed a *leave-one out* cross-validation. From the corpora we extracted a questioned signature; the remaining part will serve as the training data, which is used to infer the various distributional parameters. The questioned signature is finally evaluated against the updated distributions, thereby producing a LR value. This procedure is repeated until the corpora are exhausted, thereby

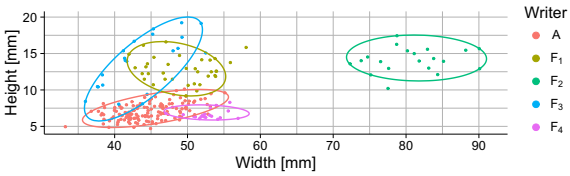


Figure 3: Plot of signature dimensions. Colours represent the identity of the forger, where "A" marks the author of genuine specimens. To aid visual separation and to suggest distributional hypotheses, we superimposed the smallest ellipses covering 95% of the points [29].

producing $n = 239$ LR values. We expect to obtain LRs greater than 1 in the authentic corpus, and LRs smaller than 1 otherwise.

In the light of previous remarks, we collect S_1 and S_2 in a vector, and we model the vector $\underline{S} = (S_1, S_2)$ as follows:

$$\underline{S} \,|\, H = h, \underline{\mu_h}, \Sigma_h \sim N_2(\underline{\mu_h}, \Sigma_h), \tag{1}$$

where $h \in \{h_p, h_d\}$ is value of the considered hypothesis ($H = h_p$ in authentic signatures, $H = h_d$ otherwise), and $(\underline{\mu_h}, \Sigma_h)$ are respectively the maximum likelihood estimates of the mean and the covariance matrix under h. By abuse of notation, we will write, e.g., $\underline{\mu_p}$ instead of $\underline{\mu_{h_p}}$. In short, the model is the following:

$$\left.\underline{s}^{(i)}\right| H^{(i)} = h, \underline{\mu_h}, \Sigma_h \overset{iid}{\sim} N_2(\underline{\mu_h}, \Sigma_h) \quad \forall i : w_i \in \{A\} \cup \mathcal{F}. \tag{2}$$

We are assuming that signature dimensions are temporally indistinguishable (i.e., there is no dependence on i), and their distribution corpora can be fully described with a mean vector, and a covariance matrix. The model (1) is represented as a

ACCEPTED MANUSCRIPT

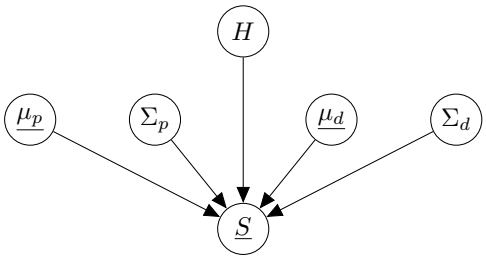


Figure 4: The model (1) *represented as a Bayesian network.*

Bayesian network in Figure 4.

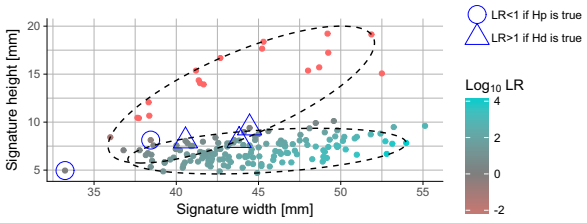
The LR value is computed by evaluating the probability density function of the questioned pair (S_1, S_2) under the competing hypotheses. In formulae:

$$\mathrm{LR}^{(i)} = \frac{f(\underline{s}^{(i)}; \underline{\mu}_p, \Sigma_p)}{f(\underline{s}^{(i)}; \underline{\mu}_d, \Sigma_d)} \quad \forall i : w_i \in \{A\} \cup \mathcal{F},$$

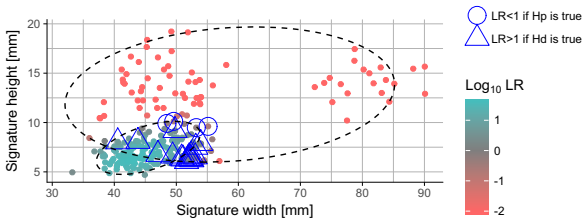
where $\underline{\mu}_p$, Σ_p, $\underline{\mu}_d$, Σ_d have been estimated using the remaining parts of the considered corpora (i.e., for the i-th LR, all specimens with indexes $\{j : w_j = A \wedge j \neq i\}$ under h_p, and all specimens with indexes $\{j : w_j \in \mathcal{F} \wedge j \neq i\}$ under h_d).

In Figure 5 we show the LR obtained for each specimen under two comparison scenarios. Notice the very large LR values: this may be due to the fact that in our model there is no uncertainty on the parameters of the Gaussian distributions, as we substituted their respective maximum likelihood estimates. Tippett plots for all scenarios are reported in Figure 6, while corresponding confusion matrices are reported in Table 3.

Notice that in order to produce a fully Bayesian model, in (1) we should specify a prior on $\underline{\mu}_h$, Σ_h and the hypothesis H. [30] For this study, however, we substituted the maximum likelihood es-



(a) $\mathcal{F} = \{F_3\}$*, i.e., comparison between two writers,* h_d *="The signature has been written by* F_3*".*



(b) $\mathcal{F} = \{F_1, F_2, F_3, F_4\}$*, i.e., comparison between 5 writers,* h_d *="The signature has been written by any of the* $\mathcal{F}$*".*

Figure 5: LR*s obtained through cross-validation in two different scenarios. Each point represents a questioned signature, evaluated against the rest of the corpora. The resulting* Log LR *value is represented by its colour. Ellipses containing 95% of the competing distributions of the* $\underline{S}$ *are added, while triangles and circles are superimposed on cases where the LR supports the wrong hypothesis. Note the large difference in evidence strength between Figure 5a and Figure 5b.*

ACCEPTED MANUSCRIPT

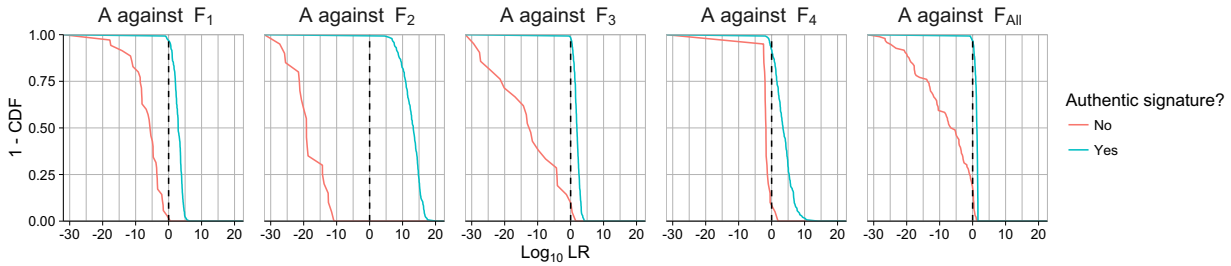


Figure 6: Tippett plots of signature sizes under 5 different comparison scenarios. The dashed line marks $\mathrm{LR} = 1$.

	$\mathcal{F} = F_{\mathrm{All}}$		$\mathcal{F} = F_1$		$\mathcal{F} = F_2$		$\mathcal{F} = F_3$		$\mathcal{F} = F_4$	
Ground truth	$\mathrm{LR} < 1$	$\mathrm{LR} \geq 1$	$\mathrm{LR} < 1$	$\mathrm{LR} \geq 1$	$\mathrm{LR} < 1$	$\mathrm{LR} \geq 1$	$\mathrm{LR} < 1$	$\mathrm{LR} \geq 1$	$\mathrm{LR} < 1$	$\mathrm{LR} \geq 1$
H_p	4	139	4	139	0	143	2	141	11	132
H_d	79	17	34	1	20	0	18	3	18	2

Table 3: Confusion matrices across 5 different comparison scenarios.

timates in their corpora, as performed in [31]: this both simplifies the reasoning, and greatly reduces the amount of calculations which have to be computed. A fully Bayesian modelling can be attempted to account for parameter uncertainty, as done in [30, 32, 33, 34].

A further possible extension is to also infer $|\mathcal{F}|$ (i.e., the number of forgers in the corpora), and the individual characteristics of each writer. The fully Bayesian tools of choice are Dirichlet Process Mixture Models [35].

5.1.1. Sensitivity to dataset size

To investigate on the sensitivity of the method introduced in 5.1, we conducted a sensitivity analysis of the LR on the particular choice of the dataset. Specifically, instead of using the entire authentic corpus to estimate the population parameters ($\underline{\mu_p}$ and Σ_p), we considered smaller random subsets of specimens of increasing size (8 sizes being $\{5, 10, 12, 15, 20, 30, 50, 90\}$, while the authentic corpus has size 143). For each subset size, we repeated the analysis 10 times, each time choosing a new subset, thus obtaining $10 \times 8 = 80$ likelihood ratios for each specimen, for a fixed $\mathcal{F}$ (the set of possible forgers). An averaged LR can be formed by combining the 10 repetitions for each image and for each subset size. The arithmetic mean of such repetitions has been indicated with the symbol $\overline{\mathrm{LR}}$. It is worthwhile noting that it is possible to perform other kinds of aggregations, such as majority voting. Large sample properties, however, do not change.

Only the authentic corpus has been subsampled for two reasons. First, the forged corpora are already very small with respect to the authentic corpus (see Table 1): a further resampling could intro-

ACCEPTED MANUSCRIPT

duce artifacts in the obtained distributions. Secondly, this enables us to approach dataset sizes which are more commonly encountered in practical cases: e.g., we computed the likelihood ratios by comparing 20 authentic specimens and 20 forgeries.

To summarise the sensitivity results, we use Tippett plots shown in Figure 7. To ease interpretation, we have only shown the averaged likelihood ratio $\overline{\mathrm{LR}}$: also, the tails of the distributions have been cut off to highlight the increasing spread for smaller dataset sizes. Notice how the likelihood ratio is more sensitive for forgers who produced forgeries with more variable sizes (e.g., F_2). Also, it reveals how important it is to consider writers separately instead of grouping them together under the virtual writer F_{All}.

It can be shown that the size of the subsets heavily impacts the range of the likelihood ratios obtained, but does not significantly affect whether a likelihood ratio obtained for a given specimen supports the correct hypothesis. This last effect is shown in the confusion matrices in Table 4. For small dataset sizes, a signature can obtain contrasting LR values across repetitions. As more data is considered for estimation of population parameters, the LR converges to the one obtained using the entire dataset, as in the previous section. This is mostly due to the fact that the maximum likelihood estimators are unreliable unless the size of the dataset is sufficiently large.

A fully Bayesian analysis would heavily reduce this problem by leveraging on past data (represented by priors) as well as hypotheses on population parameters. However, the development of a fully Bayesian method is not straight-forward, and is out of scope of the current article.

5.2. Angles and slants

As stated in Section 4.3, angles and slants are highly dependent on the specific case. Moreover, within the collected authentic corpus it is not apparent how to define measures which are both angle-based and easy to extend to other kinds of signatures.

Another difficulty lies in the properties of the numerical domain where angles and slants lie. Common classifiers do not reliably work, and particular care is needed even to define the sample mean [36]. Circular statistics, and in particular Bayesian extensions, still pose many open research questions.

We nevertheless decided to briefly report our collected data, leaving the possibility to expand its analysis in future works, as suggested in Section 1.3.

Summary statistics are reported in Figures 8 and 9. Notice that angles and slants are much less clearly distributed than the absolute dimensions of signatures. Also, some keypoints are often missing in the corpora (e.g., keypoint 9), which leads to the impossibility of measuring an angle or a slant. As a consequence of this interplay, in some cases we speculate that evidence provided by absolute angular measurements can be dominated by other characteristics, such as presence or absence of keypoints, or the impossibility to obtain angular measurements.

5.3. Proportions and distances

In this subsection we explore a joint representation for the measures reported in Table 2. Specifically, we focus on the measures $\underline{D} = (D_j)_{j=1}^{7}$, as

ACCEPTED MANUSCRIPT

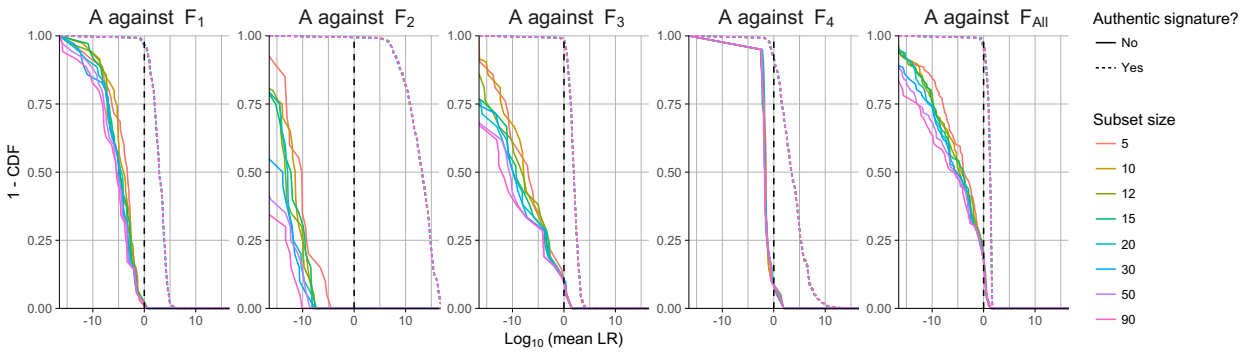
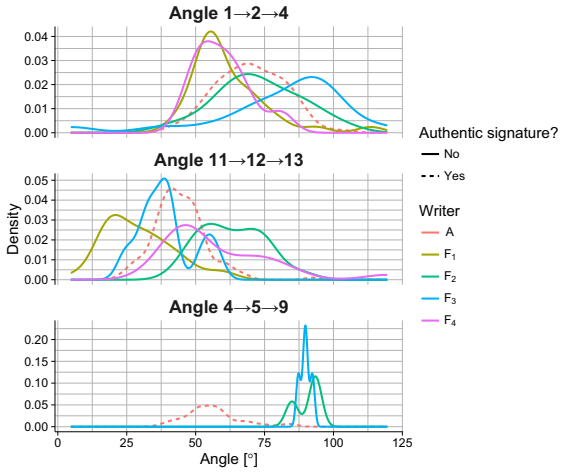


Figure 7: Tippett plots of sensitivity on dataset size under 5 different comparison scenarios. The LR *is averaged over* 10 *subsamplings for a fixed subset size. The dashed vertical line marks* $\mathrm{LR} = 1$.

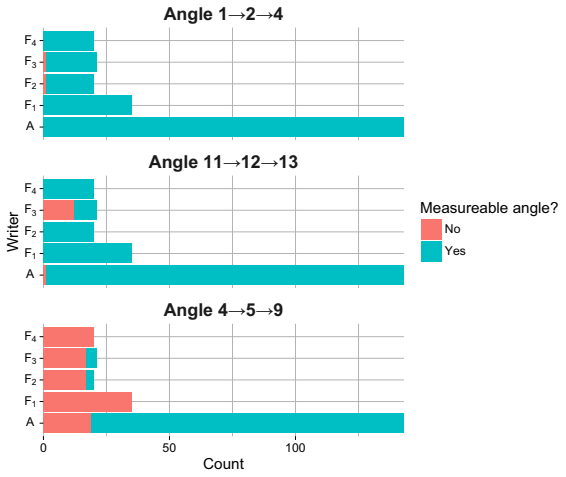
		$\mathcal{F} = F_{\mathrm{All}}$		$\mathcal{F} = F_1$		$\mathcal{F} = F_2$		$\mathcal{F} = F_3$		$\mathcal{F} = F_4$	
Ground truth	Subset size	$\overline{\mathrm{LR}} < 1$	$\overline{\mathrm{LR}} \geq 1$	$\overline{\mathrm{LR}} < 1$	$\overline{\mathrm{LR}} \geq 1$	$\overline{\mathrm{LR}} < 1$	$\overline{\mathrm{LR}} \geq 1$	$\overline{\mathrm{LR}} < 1$	$\overline{\mathrm{LR}} \geq 1$	$\overline{\mathrm{LR}} < 1$	$\overline{\mathrm{LR}} \geq 1$
H_p	5	7	136	4	139	0	143	2	141	11	132
	10	5	138	4	139	0	143	3	140	11	132
	12	6	137	4	139	0	143	1	142	12	131
	15	4	139	4	139	0	143	2	141	11	132
	20	4	139	4	139	0	143	1	142	11	132
	30	4	139	4	139	0	143	1	142	11	132
	50	4	139	4	139	0	143	1	142	11	132
	90	4	139	4	139	0	143	2	141	11	132
H_d	5	75	21	34	1	20	0	18	3	18	2
	10	78	18	34	1	20	0	18	3	18	2
	12	78	18	34	1	20	0	18	3	18	2
	15	78	18	34	1	20	0	18	3	18	2
	20	78	18	34	1	20	0	18	3	18	2
	30	79	17	34	1	20	0	18	3	18	2
	50	78	18	34	1	20	0	18	3	18	2
	90	79	17	34	1	20	0	18	3	18	2

Table 4: Sensitivity of the confusion matrices in Table 3 to subset size across 5 different comparison scenarios. $\overline{\mathrm{LR}}$ *denotes the likelihood ratio averaged over* 10 *trials for each chosen subset size.*

ACCEPTED MANUSCRIPT

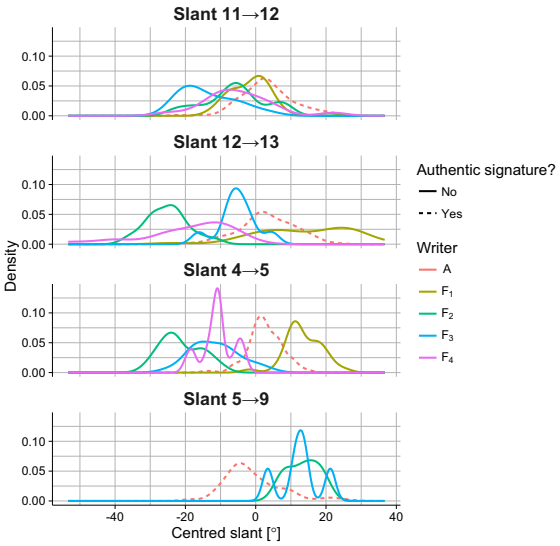


(a) Kernel density estimates of angle densities across writers. Notice that densities of angle $4 \to 5 \to 9$ are extremely noisy due to the limited number of samples.

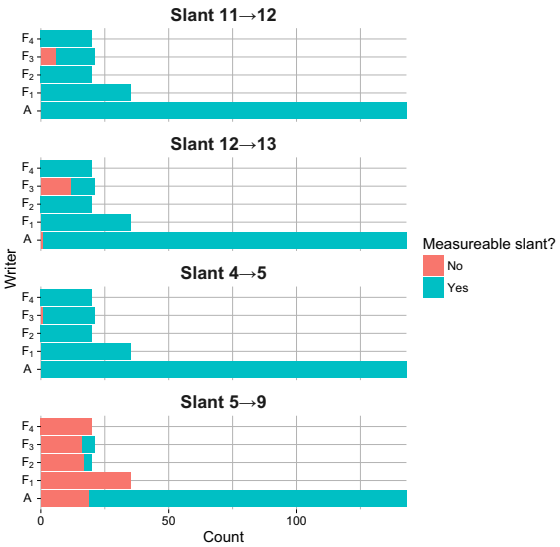


(b) Proportion of measurable angles in a given corpus. An angle is not measurable if any of its keypoints is missing.

Figure 8: Summary of angle-based measures.



(a) Kernel density estimates of slant densities across writers. Slants have been beforehand centered around their circular mean. Notice that densities of slant $5 \to 9$ are extremely noisy due to the limited number of samples.



(b) Proportion of measurable slants in a given corpus. A slant is not measurable if any of its keypoints is missing.

Figure 9: Summary of slant-based measures.

ACCEPTED MANUSCRIPT

they do not suffer issues that characterise angles and slants.

Their distributions have been summarised in Figure 10. To isolate the most discriminant features, the article [21] proceeds with Boruta feature selection. On our corpora, however, all features are labelled as being important, thus defeating the goal of the feature selection step.

Nevertheless, to gain insight into the structure of the corpora with a joint visualisation, we explored a representation of the observed $\underline{d} = \{d_i\}_{i=1}^{7}$ into a lesser-dimensional space. To this purpose, there are a host of techniques which fall under the category of *Multidimensional Scaling* (`MDS`): as an example, `PCA` is strongly related to the simplest `MDS` techniques. All `MDS` techniques map points to the lesser-dimensional space such that those who are most similar, they are represented as being closer, while points that are very dissimilar get spread far apart. Notice that `MDS` techniques work with unlabelled ("unsupervised") data.

In particular, we applied *t-distributed Stochastic Neighbor Embedding* (`t-SNE`)[37]: this technique is very powerful, well suited for high-dimensional databases, and it has been specifically conceived to map vectors from any space to a plane ($\mathbb{R}^2$). However, it has some drawbacks: the convergence to a global optimum is not guaranteed, it is a stochastic method (hence results are not easily reproducible), and it does not provide a means to map a new point on a past representation. Furthermore, it is often difficult to interpret the learned characteristics, as they are non-linear and do not have a specific geometrical meaning (unlike other dimensional reduction techniques such as principal components).

The results of `t-SNE` have been represented in Figure 11: notice the intrinsic similarity between specimens produced by each forger. This gives us hope for obtaining a statistical model capable of distinguishing evidence from different forgers in future works.

5.4. Marquis' Fourier descriptors

The idea behind this approach is to decompose each closed loop as a Fourier series of harmonic components. Following [23], from each closed loop we first extract the contour, we normalize its area[2], we describe it in polar coordinates as $\rho = f(\theta)$ (see Figure 12) and we represent f with a Fourier series[3]:

$$f(\theta) = A_1 + \sum_{k=2}^{\infty} A_k \cos\left(\theta\,(k-1) + \tau_k\right)$$

The set of amplitudes and phases $(A_k, \tau_k)_{k=k_0}^{K}$ forms the Fourier descriptors, which characterize the shape of the loop.

We chose to consider only harmonics with $k \in \{2, \ldots, 5\}$, as A_1 is tied to the loop radius, and harmonics above $k = 5$ have a negligible effect on the shape. The distribution of amplitude coefficients across letters, writers and harmonic index is reported in Figure 13.

A number of observations can be made on Fourier descriptors.

[2]The motivation behind area normalization is to be able to measure deviations from circular shapes rather than variations in scale. As a consequence, all loops have approximately the same mean radius, represented by the term A_1.

[3]To conform to the literature, we use a upper-case notation: the distinction between random variables and their realisations is left to the context.

ACCEPTED MANUSCRIPT

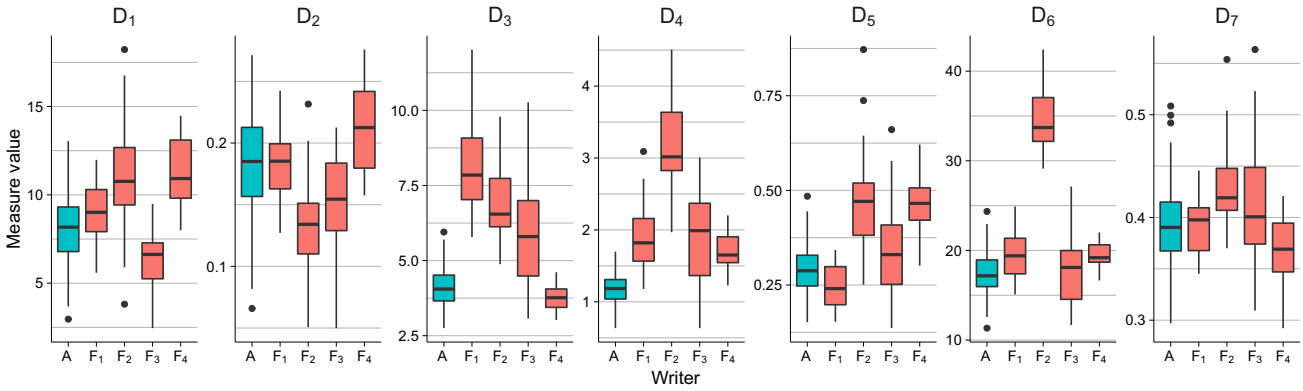


Figure 10: Boxplots for all measures in Table 2.

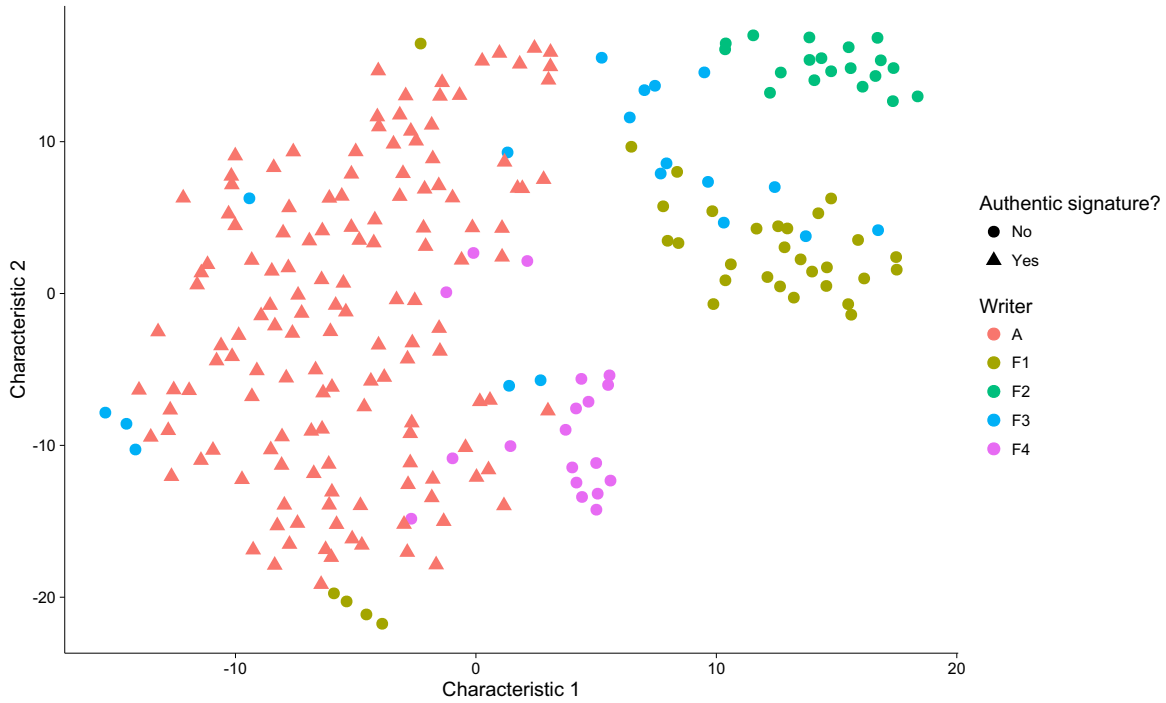


Figure 11: `t-SNE` representation of the set of measures d. Colors (forgers) have been added afterwards. The obtained characteristics have no specific meaning, yet they form clusters which correspond to individual writers: specimens which are represented as being closer, are "more similar".

ACCEPTED MANUSCRIPT

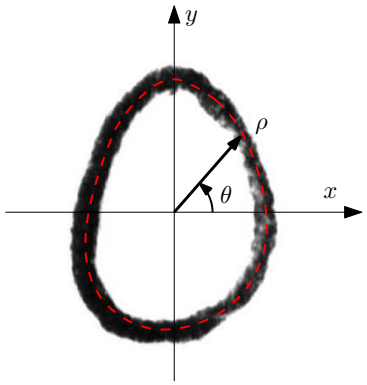


Figure 12: A closed loop in polar coordinates. Figure adapted from [23].

First of all, shape information can be properly described only by considering pairs (A_k, τ_k). All pairs are shown in Figure 14, distinguishing by letter and writer. Notice the discriminative power of the third harmonic.

It is necessary to note that all τ_ks are affected by the same problems as the angles in Section 5.2. In fact, the τ_ks are not necessarily restricted to a small neighbourhood of some τ_0, but can easily assume any angle in the unit circle: this defeats methods based on small deviations from a common mean, or methods that do not wrap distributions around the unit circle. Consequently, rather than building a classifier as in [23], in future works we will seek a representation of Fourier coefficients which is free from these issues. In particular, the Cartesian representation of Fourier descriptors lies in the Euclidean plane. We hence conjecture that $(A_k \cos \tau_k, A_k \sin \tau_k)$ can be modelled as samples from Gaussian bivariate distributions, whose parameters mainly depend on writers, letters and harmonic index. Based on these hypotheses, one can then apply the model described in Section 5.1 for each harmonic.

Similarly to keypoints, the presence or the absence of a loop can be a strong indicator for, or against, the authenticity of a signature. This is not directly expressed through Fourier coefficients, but will be nevertheless integrated into the Bayesian network by means of a binary node M, as stated in Section 1.3.

6. Discussion

So far we presented brief insights on the capabilities of the corpora we collected. It is clear that future works will heavily exploit Bayesian multivariate statistical models. We expect that significant challenges will be posed by the study of individual characteristics.

In particular, the Bayesian approach on Fourier descriptors illustrated in Section 5.4 opens a number of research questions concerning the description of the joint distribution of harmonic coefficients. It is noteworthy that Fourier analysis appears in many different technical disciplines. Any progress achieved in this field could provide insight on the usage of Bayesian methods in other forensic domains such as spectroscopy or chemistry.

Another facet to be explored is the denominator of the likelihood ratio. As detailed in Section 5.1, its definition changes according to the relevant population that is accounted for (previously noted as $\mathcal{F}$). As a consequence, one may introduce Bayesian mixture models and non-parametric methods to automatically evaluate the presence of several classes in the data.

The next challenge is posed by the construction of

ACCEPTED MANUSCRIPT

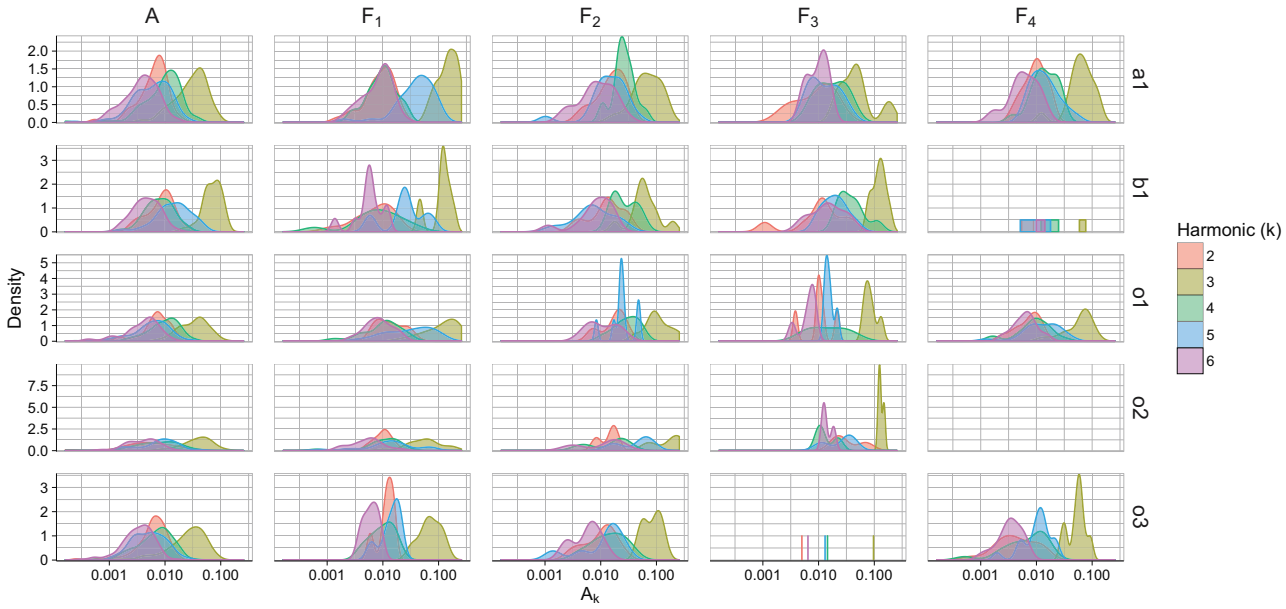


Figure 13: Amplitude density across harmonic indexes, writers and letters. Notice the logarithmic x axis. The third harmonic is always the strongest.

the Bayesian network. This phase will significantly exploit domain knowledge from forensic experts, in order to account for interdependence between features.

In general, there are no guidelines on how to construct a sensible network. However, there are a number of attempts in literature detailing individual aspects. As an example, a very recent study [34] related to Marquis' Fourier descriptors is available. Such study will serve us as a basis to integrate them into a more general framework. Moreover, in Section 1.3 we described the motivation behind the evaluation of missing evidence: this section will certainly benefit from techniques addressed to treat missing data.

A fundamental problem underlying the project is the very high data dimensionality. Forensic document examiners commonly work with very few specimens, so a good characterization of the distributions is often challenging. It is then imperative to integrate techniques to perform feature selection or dimensionality reduction. To this purpose we can also exploit domain knowledge from fingerprint comparison: specifically, one may aim at introducing a biometric score between individual characteristics in signatures. Such measure is one-dimensional, therefore drastically reducing data dimensionality and easing inferences through the Bayesian network.

A further extension of the work is the addition of new measures and new features. In particular, so far we only studied simple geometrical properties of signatures, while forensic experts commonly rely also on other types of evidence such as pres-

ACCEPTED MANUSCRIPT

sure variations, line quality and background information. Bayesian networks are easily extensible, requiring the specification of the conditional distributions for the desired feature, and the addition the node(s) to the network.

We also plan to collect new forgeries of the authentic corpus, as well as new corpora based on other signatures of different complexity. Specifically, this enables us to verify assumptions stated in Section 4.3 on universality of measures, and in Section 1.3 on case-based evidence. Moreover, with respect to Section 1.1, it also will help us to establish the feasibility of transposing the results from closed to open-set situations, where the relevant population is much larger.

Acknowlegements

We are deeply thankful to the reviewers for their helpful comments.
This research was supported by the Swiss National Science Foundation, through grant no. 10001A_156290.

References

[1] European Network of Forensic Science Institutes (ENFSI), ENFSI guideline for evaluative reporting in forensic science: Strengthening The Evaluation Of Forensic Results Across Europe (STEOFRAE)., http://enfsi.eu/sites/default/files/documents/external_publications/m1_guideline.pdf (2015).

[2] R. A. Huber, A. M. Headrick, Handwriting Identification: Facts and Fundamentals, CRC Press, Boca Raton, Florida, 2010.

[3] C. Champod, Identification / individualization: Overview and meaning of ID, in: J. A. Siegel (Ed.), Encyclopedia of Forensic Sciences, Elsevier, Oxford, 2000, pp. 1077 – 1084.

[4] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter, Probabilistic Networks and Expert Systems, Springer, New York, 1999.

[5] A. Biedermann, F. Taroni, A probabilistic approach to the joint evaluation of firearm evidence and gunshot residues, Forensic Science International 163 (1-2) (2006) 18–33. doi:10.1016/j.forsciint.2005.11.001.

[6] A. Biedermann, F. Taroni, S. Bozza, W. D. Mazzella, Implementing statistical learning methods through Bayesian networks (Part 2): Bayesian evaluations for results of black toner analyses in forensic document examination, Forensic Science International 204 (1-3) (2011) 58–66. doi:10.1016/j.forsciint.2010.05.001.

[7] A. Biedermann, R. Voisard, F. Taroni, Learning about Bayesian networks for forensic interpretation: An example based on the 'the problem of multiple propositions', Science & Justice 52 (3) (2012) 191–198. doi:10.1016/j.scijus.2012.05.004.

[8] E. Sironi, V. Pinchi, F. Taroni, Probabilistic age classification with Bayesian networks: A study on the ossification status of the medial clavicular epiphysis, Forensic Science International 258 (2016) 81–87. doi:10.1016/j.forsciint.2015.11.010.

[9] G. Cereda, A. Biedermann, D. Hall, F. Taroni, Object-oriented Bayesian networks for evaluating DIP–STR profiling results from unbalanced DNA mixtures, Forensic Science International: Genetics 8 (1) (2014) 159–169. doi:10.1016/j.fsigen.2013.09.001.

[10] F. Taroni, A. Biedermann, S. Bozza, P. Garbolino, C. Aitken, Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science, John Wiley and Sons, Chichester, United Kingdom; Hoboken, NJ, 2014.

[11] S. Gittelson, A. Biedermann, S. Bozza, F. Taroni, Modeling the forensic two-trace problem with Bayesian networks, Artificial Intelligence and Law 21 (2) (2013) 221–252. doi:10.1007/s10506-012-9136-5.

[12] R. Plamondon, G. Lorette, Automatic signature verification and writer identification - the state of the art, Pattern Recognition 22 (2) (1989) 107–131.

[13] F. Leclerc, R. Plamondon, Automatic signature verification: The state of the art—1989–1993, International Journal of Pattern Recognition and Artificial Intelli-

Acknowledgements (Final version)
This research was supported by the Swiss National Science Foundation (grants no. 10001A_156290 and BSSGI0_155809) and the University of Lausanne.

ACCEPTED MANUSCRIPT

gence 8 (03) (1994) 643–660.

[14] D. Impedovo, G. Pirlo, Automatic signature verification: The state of the art, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 38 (5) (2008) 609–635.

[15] P. Frazer, Des Faux En Écriture et de L'écriture, Méthode Scientifique Nouvelle D'analyse et D'examen, Guillaumin, Paris, 1899.

[16] A. Bertillon, La Comparaison Des Écritures et L'identification Graphique, F.C.W. Vogel, Paris, 1901.

[17] E. Locard, Traité de Criminalistique: L'expertise Des Documents Écrits. Les Correspondances Secrètes. Les Falsifications, J. Desvigne, Lyon, 1936.

[18] R. Morris, R. N. Morris, Forensic Handwriting Identification: Fundamental Concepts and Principles, Academic Press, San Diego, California, 2000.

[19] R. Mansuy, L. Mazliak, Introduction au rapport de Poincaré pour le proces en cassation de Dreyfus en 1904, Bulletin de la Sabix. Société des amis de la Bibliothèque et de l'Histoire de l'École polytechnique (42) (2008) 60–63.

[20] P. L. Kirk, Crime Investigation: Physical Evidence and the Police Laboratory Interscience, Vol. 75, Interscience Publishers, New York, 1953.

[21] I. Montani, Exploring transparent approaches to the authentication of signatures on artwork, Ph.D. thesis, Université de Lausanne, Faculté de droit, des sciences criminelles et d'administration publique (2015).

[22] A. Thiéry, Développement d'un processus de quantification et d'évaluation de caractères manuscrits : théorie et applications, Ph.D. thesis, Université de Lausanne, Faculté de droit, des sciences criminelles et d'administration publique (2014).

[23] R. Marquis, M. Schmittbuhl, W. D. Mazzella, F. Taroni, Quantification of the shape of handwritten characters: A step to objective discrimination between writers based on the study of the capital character O, Forensic Science International 150 (1) (2005) 23–32. doi:10.1016/j.forsciint.2004.06.028.

[24] R. Marquis, F. Taroni, S. Bozza, M. Schmittbuhl, Quantitative characterization of morphological polymorphism of handwritten characters loops, Forensic Science International 164 (2–3) (2006) 211–220. doi:10.1016/j.forsciint.2006.02.008.

[25] R. Marquis, Etude de caractères manuscrits: de la caractérisation morphologique à l'individualisation du scripteur, Ph.D. thesis, Université de Lausanne, Faculté de droit, des sciences criminelles et d'administration publique (2007).

[26] R. Marquis, S. Bozza, M. Schmittbuhl, F. Taroni, Handwriting evidence evaluation based on the shape of characters: Application of multivariate likelihood ratios, Journal of Forensic Sciences 56 (2011) S238–S242. doi:10.1111/j.1556-4029.2010.01602.x.

[27] A. J. Quirke, Forged, Anonymous and Suspect Documents, George Routledge, London, 1930.

[28] R. Gonzalez, R. Woods, Digital Image Processing, Pearson Education, Upper Saddle River, NJ, USA, 2009.

[29] S. Van Aelst, P. Rousseeuw, Minimum volume ellipsoid, Wiley Interdisciplinary Reviews: Computational Statistics 1 (1) (2009) 71–82. doi:10.1002/wics.19.

[30] S. Bozza, F. Taroni, R. Marquis, M. Schmittbuhl, Probabilistic evaluation of handwriting evidence: Likelihood ratio for authorship, Journal of the Royal Statistical Society: Series C (Applied Statistics) 57 (3) (2008) 329–341. doi:10.1111/j.1467-9876.2007.00616.x.

[31] F. Taroni, S. Bozza, A. Biedermann, P. Garbolino, Data Analysis in Forensic Science: A Bayesian Decision Perspective, Statistics in Practice, John Wiley and Sons, Chichester, United Kingdom; Hoboken, New York, 2010.

[32] C. G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, Journal of the Royal Statistical Society: Series C (Applied Statistics) 53 (1) (2004) 109–122.

[33] F. Taroni, R. Marquis, M. Schmittbuhl, A. Biedermann, A. Thiéry, S. Bozza, The use of the likelihood ratio for evaluative and investigative purposes in comparative forensic handwriting examination, Forensic Science International 214 (1–3) (2012) 189–194. doi:10.1016/j.forsciint.2011.08.007.

[34] F. Taroni, R. Marquis, M. Schmittbuhl, A. Biedermann, A. Thiéry, S. Bozza, Bayes factor for investigative assessment of selected handwriting features, Forensic Science International 242 (2014) 266–273. doi:10.1016/j.forsciint.2014.07.012.

ACCEPTED MANUSCRIPT

[35] M. D. Escobar, M. West, Bayesian density estimation and inference using mixtures, Journal of the American Statistical Association 90 (1994) 577–588.

[36] S. R. Jammalamadaka, A. Sengupta, Topics in Circular Statistics, Vol. 5, World Scientific, 2001.

[37] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2579-2605) (2008) 85.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

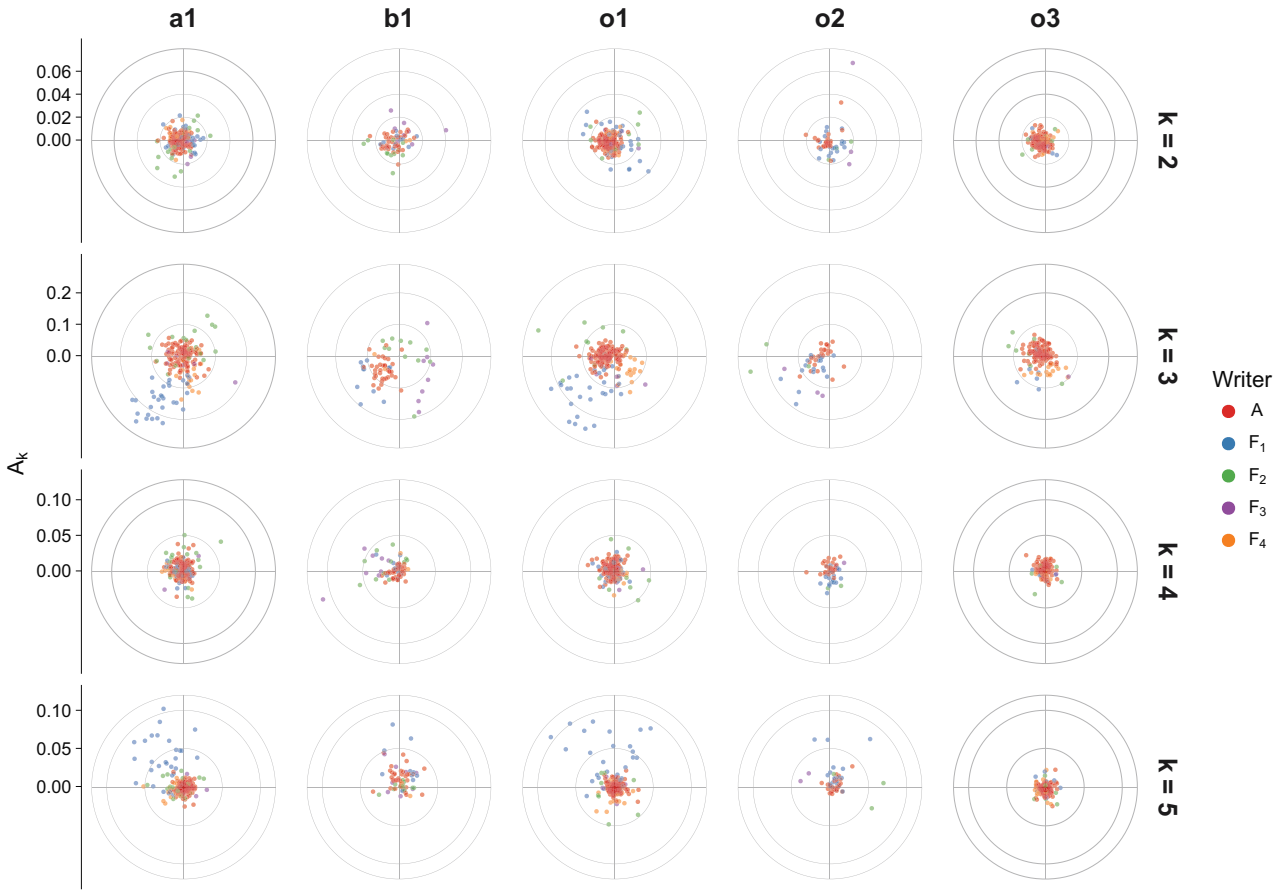


Figure 14: Polar plot of pairs (A_k, τ_k) across harmonic indexes, writers and letters. Notice the different amplitude scales.

ACCEPTED MANUSCRIPT

Highlights

- The problem of evidence evaluation in comparative signature examinations is explored.
- We propose a set of measurements which can be performed on signatures, along with the corresponding statistical models.
- The aim is to build a Bayesian network for evaluating evidence of different nature.

ACCEPTED MANUSCRIPT