Year : 2023

Bayesian models in questioned handwriting and signatures

Gaborini Lorenzo

Gaborini Lorenzo, 2023, Bayesian models in questioned handwriting and signatures

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_38564422C6201

UNIL | Université de Lausanne
Faculté de droit,
des sciences criminelles
et d'administration publique

Faculté de droit, des sciences criminelles et d'administration publique
École des Sciences Criminelles

Bayesian models in questioned handwriting and signatures

PhD THESIS of

Lorenzo GABORINI

Thesis supervisor: **Professor Franco TARONI**

Lausanne, 2023

UNIL | Université de Lausanne
Ecole des sciences criminelles
bâtiment Batochime
CH-1015 Lausanne

IMPRIMATUR

A l'issue de la soutenance de thèse, le Jury autorise l'impression de la thèse de Monsieur Lorenzo Gaborini, candidat au doctorat en science forensique, intitulée

« Bayesian model in questioned handwriting and signatures»

Le Président du Jury

Professeur Christophe Champod

Lausanne, le 7 décembre 2021

Contents

Abstract

Forensic document examination is one of the oldest areas of forensic science. Despite the advent of personal computers and portable digital tools, the discipline has enjoyed relatively few methodological advances compared to other forensic areas. Moreover, the use of handwritten evidence in court has historically faced many issues, particularly in the US-centric system.

Among the specificities of this field, one can identify the lack of physical laws governing the fundamental principles of handwriting, the difficulty in assessing the general validity of these principles, and the reliance on human experts' judgement to provide an opinion to the stakeholders. The reporting of the evidential value also lags behind other forensic areas, such as DNA interpretation, where the modern Bayesian approach of the ENFSI Guideline for Evaluative Reporting in Forensic Science is fully implemented.

Starting from the fundamental principles governing handwriting, necessarily qualitative in their nature, this thesis first considers several well-defined forensic scenarios in which such principles can be translated to a statistical description of a series of measurements. The necessary evidence is collected either on request via a panel of writers, or from real casework. We considered scenarios where authorship is discussed, either of signatures or naturally handwritten content. Next, the Bayesian approach is introduced, from the theoretical notions to the computational requirements. As no universal technique to compute the Bayes Factor is available (the only coherent measure for evaluative purpose), every scenario requires a distinct path and a tailored approach. The stability and the validity of each developed model is also approached, for instance by performing sensitivity analyses on its parameters or on the data.

Bayesian reasoning can be easily generalized, and allows one to approach the issue of combining multiple kinds of evidence. As an example, we consider a hypothetical scenario where an anonymous letter is found jointly with salivary evidence, under the hypothesis that both came from the same person of interest. The person of interest declares that his twin brother was the source of both traces. In the last Chapter we first show how the developed models can be adapted for each type of evidence, then how they can be combined together to produce an evaluative report that is coherent and justified.

Résumé

L'expertise en documents est l'une des plus anciennes branches des sciences forensiques. Malgré l'époque numérique et l'apparition de l'ordinateur et de nombreux outils numériques portables, la discipline n'a eu le même taux de développement méthodologique par rapport à d'autres branches forensiques. De plus, l'utilisation d'évidences manuscrites dans les tribunaux a historiquement été confrontée à de nombreux problèmes, notamment dans le cadre du système judiciaire des États-Unis.

Parmi les spécificités (et les lacunes) du domaine, on rencontre l'absence de lois physiques qui règlent les principes fondamentaux de l'écriture manuscrite, la difficulté d'évaluation de la validité de ces principes, et l'approximation logique de l'expert dans la formulation de ses conclusions. La manière dont la valeur de l'indice est reportée est aussi en retard comparé à d'autres branches forensiques, notamment l'analyse ADN, où l'approche dite "Bayésienne" moderne de l'ENFSI Guideline for Evaluative Reporting in Forensic Science est pleinement adoptée.

Partant des principes fondamentaux régissant l'écriture manuscrite, nécessairement qualitatifs par nature, cette thèse considère d'abord plusieurs scénarios forensiques bien définis où ces principes peuvent être traduits dans un descriptif statistique d'une série de mesures. Le matériel nécessaire (corps d'écriture, soit de signatures, soit d'écriture naturelle) est récolté soit à travers un panel de scripteurs, soit provenant de cas réels. Nous avons traité des scénarios où l'identité du scripteur est disputée. Ensuite, l'approche Bayésienne est adoptée, suivie d'une exposition des notions théoriques et computationnelles. Puisqu'il n'existe aucune technique universelle de calcul du Facteur de Bayes (seule méthode cohérente pour l'évaluation des observations forensiques), tout scénario nécessite une démarche adaptée et ciblée à la problématique d'intérêt. La stabilité et la validité de chaque modèle développé sont également approchées, par exemple à travers d'une analyse de sensitivité aux paramètres ou aux données.

Le raisonnement Bayésien peut être facilement généralisé, permettant l'inclusion de plusieurs types d'observations forensiques différentes. Par exemple, nous considérons un scénario hypothétique dans lequel une lettre anonyme manuscrite est saisie ainsi que du matériel salivaire, sous l'hypothèse que les deux proviennent d'une même personne d'intérêt. Cette dernière déclare que son jumeau est à l'origine des traces. Dans le dernier chapitre, nous montrons d'abord comment les modèles ici développés peuvent être adaptés à ces types de preuves, puis comment ils peuvent être combinés pour produire un rapport d'évaluation de type évaluatif cohérent et justifié.

Chapter 1

Introduction

In this Chapter we give an overview on handwriting evidence and examination in forensic science, from the first historical cases to contemporary times. We present the current state of the art of research on handwriting examination, the relationships it has with other fields such as machine learning, and reactions from the forensic community.

The work of forensic handwriting examiners rests on several fundamental principles that govern the act of handwriting production, as well as a specific vocabulary. These aspects are briefly exposed in the Chapter.

We present the structured process that examiners follow to reach their conclusions, beginning with the search of specimens and ending with the statement of the expert opinion to the stakeholders. We also focus on the way their conclusions are stated, as it is the subject of a lively transversal debate across all areas of forensic science.

As forensic handwriting examination is an extremely complex process, we present the most important issues that characterize the field. Some of these have been successfully solved by past research, for instance the issue of admissibility of handwritten evidence in court trials (see Section 1.5.1 for a US-centric overview). Other issues are still open for research and discussion.

It is upon both the fundamental principles and current research that this thesis is based. Our contributions to the solution of the open issues in determined cases are detailed. First, we adopt a wide perspective, showing we address these issues within the generic scope of forensic handwriting examination. Then, we expose how we intend to solve these problems on a case-by-case basis.

Finally, the end of the Chapter contains a summary of the structure of the thesis.

1.1 A brief history of handwriting in forensic science

Forensic document examination is one of the oldest areas of forensic science. The Roman magistrate Gaius Verres was condemned in 70 B.C., on the basis of alleged forgery of official documents (Locard, 1959; Stone, 2018). In the third century, Titus was known to be involved in forgeries. Handwritten evidence was frequently reported in court during history: Justinian I (539), the "La Roncière" case (1835), the "La Bussinière" case (1891), the Dreyfus case (1894) (Münch, 2000).

Forensic document examiners (FDE) and forensic handwriting examiners (FHE) acted as experts in many recent highly mediatized cases (Koppenhaver, 2007): we can recall the Howard Hughes fake autobiography (1971), the false Hitler Diaries (1983), the Gregory case (1983), the Omar case (1991) and the Seznec case (1923).

Nowadays, FHE — despite the electronic era and the diffusion of electronic devices — are still being frequently confronted with cases involving handwriting or signatures, for example in wills, contracts or authentication of paintings (Montani, 2015). However, the discipline has enjoyed relatively few methodological advances compared to other forensic areas.

In 1910, Albert Osborn set the bases for modern forensic document examination, establishing its scope, its tools, the techniques and the principles that experts still use in current practice (Koppenhaver, 2007, p. 49; Osborn, 1910). Many scholars have further developed the discipline during the last century. We cite the works of Harrison (1958), Hilton (1992), Huber & Headrick (1999), Morris & Morris (2000), Lewis (2014) and Koppenhaver (2007).

Efforts were mainly directed at understanding the intrinsic characteristics of handwriting, how it is learned and taught, how it varies with time, how it is affected by illnesses and medical conditions, and how it differs across writers, populations, cultures and writing systems. The area was also concerned with examining printed or typed documents, exploiting ink characteristics and chemical composition, paper properties and latent marks on the document surface. Forensic document examination has also recently been influenced by advances in neurosciences (Caligiuri & Mohammed, 2012), which provide a description of how the writing process is dissected by the brain.

The advent of the digital era introduced new automatic tools to the arsenal available to help FDEs in their work. Among these, we cite the systems Script (1986), Forensic Information System on Handwriting (FISH) (1990), CEDAR-FOX (1996), WANDA (2003), TRIGRAPH (2005), Graphlog (2006), Masquerade (2012), Biografo (2015) and GRAPHJ (2018) (Fabiańska et al., 2006; Galbally et al., 2015;

Guarnera et al., 2018; Leedham & Srihari, 2003; Marquis, 2007; Natural Intelligent Technologies Srl, 2012; Niels et al., 2005). These systems automatically extract a set of features (assumed relevant) given sufficient handwritten material, and produce either a list of similar specimens in the system, a list of selected measurements extracted from the images, or a binary conclusion on the authenticity (i.e. whether the questioned specimen(s) are attributed to the same writer). These automated systems share some major weaknesses (Marquis, 2007). A generally shared trait between those systems is the assumption of independence between the set of features which are extracted, but which may not hold. For example, it is known that the writing system shown in the acquired material has an effect on multiple features (Huber & Headrick, 1999). Another issue is that comparisons made by automated systems are intrinsically different to comparisons made by experts. These systems consider only the features that have been programmed to be considered as relevant, while FDEs are able to reach conclusions exploiting subjective characterizations and context-related information. Also, FDE are allowed to state a "no conclusion", for example when the acquired material is in insufficient quantity (Saunders et al., 2011). For these reasons, automated systems are rarely used by forensic document examiners (Found, 2012), although they can be employed as a means to process large quantities of documents to create a database of items of handwriting (for example Graphlog in Maciaszek (2011) and Dziedzic (2016)). Moreover, the criteria for inferential and decisional processes are questionable from a logical and statistical point of view.

Contemporary research has also been influenced by the pattern recognition and machine learning communities. A series of deep reviews of their research output was conducted, showing the great interest by scholars in these fields (Impedovo & Pirlo, 2008; Leclerc & Plamondon, 1994; Plamondon & Lorette, 1989). A number of competitions were also held, with the goal of collecting and obtaining the best performing algorithm for typical FDE tasks on a given database, such as writer identification (detecting a writer given a set of samples from several writers) and signature verification (given a written specimen, determine whether it could have been written by a given writer) (Hassaine et al., 2013; Liwicki et al., 2012; Louloudis et al., 2011, 2013; Malik et al., 2015).

The recent appearance and enormous success of deep learning techniques in many computationally intensive fields (from image processing to autonomous driving) inspired new applications by the machine learning community. The main difference with the pattern recognition algorithms is that deep learning approaches heavily exploit computational power in order to learn the *best* discriminatory features from data, rather than relying on characteristics defined as relevant by humans. Some of these techniques have recently been applied to FDE tasks such as writer identification

(Christlein et al., 2015; Christlein et al., 2017) and signature verification (Hafemann et al., 2016), beating in both cases the state of the art according to machine learning metrics. Recent reviews on deep learning research have been conducted by (Rehman et al., 2019) and (Hafemann et al., 2017).

Concerning these modern techniques, we will avoid citing any article in particular, and rather referring the interested reader to the resourceful reviews cited above. One common factor of the pattern recognition literature applied to FDE is that the metrics used to report the conclusions are not well suited for forensic purposes. The goal of most of these works is the creation of an automated system which decides whether a set of specimens is consistent with the hypothesis of source (e.g. a given writer) or authenticity. As it will be explained later in the Chapter, this approach is not consistent with the Bayesian evaluative approach of forensic evidence, as the decision ultimately bears on the judge, not on the expert (or in this case, the automated system). The complete adoption of the evaluative standard should also impact the entire process, for instance the choice of an appropriate reference database to elicit prior knowledge about the evidence (Champod et al., 2004). However, these works have the merit of exploring and suggesting novel handwriting features that might guide FDEs in their decision.

1.2 State of the art of handwriting examination in forensic science

It is important to note that the fundamental principles governing handwriting, as established by Osborn and his scholars, are mostly qualitative in their nature. Some of the discriminating elements that are used by FDE can be measured (e.g. slants), but a standard approach on how the measurements are performed, compared and catalogued, is lacking. One's handwriting fluency and intricacy are often qualitatively described, yet they are important factors that may alter a judgement of genuineness (Huber & Headrick, 1999, sec. 55).

With the adoption of computers and the appearance of more sophisticated measurement tools, a large amount of empirical quantitative studies on handwriting appeared in recent literature.

Among these, we should cite the works of Srihari et al. (Srihari et al., 2008; Srihari et al., 2002; Srihari et al., 2005). In these studies, the authors analyze several large datasets with the main goal of verifying the principles behind forensic document examinations, in particular the "uniqueness" of handwriting. Through their proposed methods, they claim to be able to discriminate writers with a very low error rate.

Despite attracting vivid criticisms by FDE practitioners (Marquis et al., 2005; Saks, 2003; Thiéry, 2014), they are frequently used in court to motivate the admissibility of handwritten evidence in court, as well as raising important questions, perhaps involuntarily, that the discipline needs to answer.

The simple presence or absence of certain characteristics in specimens may be exploited by FDEs, provided that their rarity in a given relevant population is known. Two studies were conducted in order to provide these relative rarities for specified characteristics in natural handwriting (Johnson et al., 2017) and written numbers (Vastrick et al., 2018). The authors warn that these results do not provide sufficient information to correctly evaluate the value of evidence. For example, these studies consider only the presence or absence of selected features, but not their range of variation. Some of the most important characteristics (such as line quality) were not analyzed. The independence between features was only partially characterized, so the product rule of probability theory is not directly applicable to evaluate the joint value for the occurrence of certain characteristics. Also, they consider a demographic which allegedly represents the US population. An extrapolation of these results to other populations might not be possible.

Other studies provided a quantitative description of various features that FDE can exploit in their work. Starting from a large scale, a signature may be composed by two elements, for example the name and the surname. It has been shown on 60 subjects that the horizontal length of these components is stable on a window of 15 weeks (Matuszewski & Maciaszek, 2008). In another study, experts considered the variability of the total height and width of a signature across time (days or months), grouping all components together. The study compared 2320 signatures from 8 individuals collected over 3 years, showing that in most individuals the dimensions were significantly different across time. Length and height for most individuals were lowly correlated when considering signatures which were written close in time (Evett & Totty, 1985). These results were extended by considering various absolute and relative measurements from the initial letters of signatures from 30 individuals. This study showed the lack of correlation of absolute height and width of initials, as well as showing that these are stable over time (Maciaszek, 2011). It is important to note that these studies are not directly comparable, as they differ in the reference population that supplied the specimens, the time window of collection, the criteria used to measure absolute and relative lengths, and the statistical quantification of the variability.

On a smaller scale, Lizega Rika (2018) described the variability of letter proportions (height, width and their ratio) in natural handwriting, showing that these measurements are highly characteristic across 21 writers (Lizega Rika, 2018). Muehlberger et

al. (1977) considered the letter combination "th", common in English texts, and gave a description of the variability of several geometrical relations that can be measured on the letter pair. Ling (2002) built upon this article, adding other measurements (both lengths and spacings) from more letter combinations.

Marquis et al. (2005) investigated the shape of handwritten character loops, such as those found in letter "o". This work has been extended to letters "a", "d", "q" in later works, and the effect of forced enlargement on letter shape was discussed (Marquis et al., 2006, 2007). Thiéry (2014) attempted to generalize the previous descriptors to characters of a generic shape.

The direction of strokes in selected characters as an indicator of authorship was studied by Franks et al. (1985). FDE can also investigate cases where the questioned material is handwritten, but does not represent any letter nor signature. For example, Marquis, Mazzella, et al. (2019) considered the variability of "x"-shaped marks, such as ones that can be requested by checkboxes in forms. Such marks appear simple to the untrained eye, yet it has been shown that even simple marks contain enough information to form a opinion on authorship (Marquis, Mazzella, et al., 2019; Marquis, Hicks, et al., 2019). It is naturally possible to consider multiple features from the same document. When the handwritten evidence is composed of words, several stroke-based measurements can be taken, and the distributions compared (Marcelli et al., 2015). Again, the features were supposed to be independent.

Despite the recent evolution of the discipline, very few studies venture into providing an evaluation of handwritten evidence consistent with the principles of forensic interpretation that will be explained later in the Chapter. Among these we cite (Bozza et al., 2008; Marquis et al., 2011; Marquis, Hicks, et al., 2019; Srihari et al., 2008; Taroni, Marquis, et al., 2014). It is upon these studies that this thesis is built, to further provide support to the usage of the Bayesian framework in forensic document examinations.

1.3 Principles and elements characterizing handwriting

Post-Osborn research manifested in the development of a structured body of knowledge of writing characteristics that FDEs routinely exploit in their work. Many scholars identified a set of fundamental principles that appear to arise when dealing with handwriting. Here we report these principles as defined by Huber & Headrick (1999). Alternative formulations exist, such as in Hilton (1963) or Morris & Morris (2000).

First, handwriting is based on *habituation*. Everyone learns first to write using a

common model (a copybook, or letters shown by the teacher on a blackboard) (Kelly & Lindblom, 2006). As one grows, experience and repetition create a set of habits which contribute in differentiating handwriting across individuals. These habits may involve a number of characteristics such as word choice and letter formation. The "imprecision with which the habits of the writer are executed on repeated occasions" constitute the *natural variation* (Huber & Headrick, 1999, sec. 28).

The second principle is that handwriting is claimed to be unique to the individual. In the words of Huber and Headrick, two writings of the same material by two different persons are different (Huber & Headrick, 1999, sec. 28, as principle of heterogeneity). This principle is generally accepted by the forensic community, and is analogous to the claim of unicity of fingerprints. However, some care is needed to interpret it correctly. As Huber & Headrick (1999) say, fingerprints have a system of classification which allows one to verify this claim on very large databases. Despite this fact, no one will ever be able to *prove* unicity by sampling the entire population of the world, yet unicity of fingerprints is generally accepted. Actually, it is not even necessary to prove this principle, but only to be able to assess the ability of FDEs in differentiating writers (Champod, 2009). In fact, the notion of unicity is to be intended as conditioned on the characteristics exhibited by the handwritten material under evaluation, not on the process of handwriting in general. For example, available material might be of insufficient quantity, or it may only show a limited set of features which provide little discrimination. Equivalently, a fingermark might exhibit a very poor quality. Hence, this principle relates only to the ability of a trained FDE to correctly attribute (or exclude) the fingermark to a specific person.

It is then both natural and necessary to express the principle of heterogeneity in a weaker form, by considering the writer's natural variation. In general, one's handwriting evolves around a *master pattern* (Kelly & Lindblom, 2006; Morris & Morris, 2000). afterward, FHEs in their expertise must first recognize the writer's master pattern, then assess the degree of consistency of the acquired material. The master pattern usually evolves slowly with time, while larger variations are due to illness, aging or abnormal psychomotor states. The very slowness of this variation allows FHEs to perform the required comparisons.

Quantitative studies describe this principle with the term *intra-variability*.

Definition 1.1 (Intra-variability)
The normal range of a writer around a master pattern, characteristic of the writer.

The principle of unicity can be analogously translated with the term *inter-variability*:

Definition 1.2 (Inter-variability)
The variation of master patterns across different writers.

Other fundamental principles stemming from modern research concern the elements that FDEs exploit in order to discriminate between writers. These are directly related with the first principle, as they result from the acquired habit of handwriting.

Huber and Headrick classified them mostly under two large categories (Huber & Headrick, 1999, sec. 30):

- elements of style: "aspects of writing that play a significant role in creating a pictorial, or general or overall effect" (Huber & Headrick, 1999, sec. 30);
- elements of execution: aspects of writing that relate to less visible changes in handwriting, as opposed to those implied by the elements of style.

To show some examples, among the elements of style one can find the arrangement on a page, the characteristics of allographs, slant, slope and spacing. The interplay of these features produces the general aspect of the written material. Among the elements of execution one can find the choice of abbreviations, item alignments, line production attributes (endings, continuity, quality) and pen control attributes.

To the untrained eye, the elements of style appear more markedly than elements of execution, as the latter might imply the usage of an instrument to be assessed. The elements of execution are "the personal idiosyncrasies of writing in which we find the subtle dissimilarities between the writing of one individual and the next" (Huber & Headrick, 1999, sec. 30).

The variation of the elements of style and execution, as well as their mutual interactions, is itself discriminative. Huber and Headrick further consider these aspects as two additional separate categories, one concerning the "consistency, the natural variation and the persistency", the other describing lateral expansion and word proportions (Huber & Headrick, 1999, sec. 30).

Other fundamental principles concern the act of forgery. Handwriting is a conscious process, thus it can be voluntarily modified in order to reproduce a signature, or disguise one own's writing. Recalling the first principle, notice that to successfully produce a signature different that one own's, one needs to modify one's own habits and reproduce the victim's in a plausible and uniform way. If the forger has not been trained in forensic handwriting examination, it is reasonable to assume that he will fail to imitate the elements of execution, since by definition they are much more subtle than the elements of style. As a result, an expert handwriting examiner will detect the emerging inconsistencies between the reference and the questioned material. These reasonings are stated by Huber & Headrick (1999) under the names

of "principle of Exclusion and Inclusion" and "principle of Interference". It is also expected that the difficulty of simulating a signature increases with its length, its complexity and the similarity between the writer's and the victim's writing habits. This has been postulated and quantitatively verified in Brault & Plamondon (1993), Found & Rogers (1996) and Dewhurst et al. (2007).

1.3.1 Types of handwritten evidence

As said before, handwritten evidence can appear under many forms. In this section we give a partial classification, to delineate and restrict the scope of this thesis. It is by no means an attempt to produce a complete classification; nor is it in full agreement with the rest of the literature.

Definition 1.3 (Medium)
The physical support containing the written evidence.

Definition 1.4 (Writing instrument)
The physical tool used to produce the written evidence.

We will consider only handwritten evidence on a flat physical surface, such as a sheet of paper, with ballpoint pens (the most common tool encountered in practice).

Definition 1.5 (Natural writing)
The result of execution of an individual's writing habits, when performing a familiar task that does not significantly impact the fluency (e.g. spontaneously taking notes, or copying short familiar words).

Definition 1.6 (Signature)
"A way for a person to endorse the content of a document (. . .), coming in a wide variety of forms ranging from simple to complex, from legible to stylised." (Allen, 2015).

Notice that the differences between signatures and natural writing amount not only to their content (in terms of legibility) and their purpose (e.g. authentication), but also to the way they are compared and evaluated by FDEs. FDEs must make comparisons on a "like-by-like" basis (Allen, 2015, p. 63). For example, it is reasonable to assume that the purpose of natural writing is to communicate a message. Letter forms will be present and recognizable by readers. An expertise will require reference written material that shows the same letter forms appearing in the questioned writing, thus enabling the FDE to evaluate the writer's natural variation. On the contrary, when dealing with questioned signatures, FDEs must acquire specimens of signatures from the putative source, that might not show any specific letter form.

Definition 1.7 (Forged signature)
A signature which has been written by someone other than the claimant, by imitating the shape of an authentic specimen without any external aid and with unlimited practice, with "an intent to deceive" (from Ellen, 2005).

To distinguish against other kinds of forgeries in terms of instruments and way of production, Huber and Headrick give this definition as "freehand simulated signature" (Huber & Headrick, 1999, sec. 54). We will not consider traced signatures, nor disguised signatures, nor signatures obtained by external aids (such as guided or assisted signatures). Notice that this definition implies the existence of a forger, a person whose own signature is, by definition, different from the victim's.

Definition 1.8 (Off-line writing)
The image of a writing, obtained by digitising the medium.

Definition 1.9 (On-line writing)
The recording of the process of production of a writing, as a set of space-time positions of the writing instrument across the medium.

Concerning on-line writings, notice that the amount of information available to FDEs is much greater than off-line writings. Digital tablets also provide information on pressure and tip-to-surface distance. The time ordinate is important to establish a conclusion: it is not only intimately tied with many of the elements of execution (in particular line quality, line continuity and fluidity of the writing movement), but it is also an important feature to consider when the contemporaneity of writing is contested, or when dealing non-occidental scripts (see for example Li (2019) for Chinese handwriting). In these cases the stroke order needs also to be recovered from the image. In off-line writings, in some cases one can infer the stroke direction (Marquis, Hicks, et al., 2019; Snape, 1980) and the stroke order on line crossings, for example with an optical microscope or mechanical methods (Brito et al., 2017; Shiver, 2009). A multi-stage process has recently been proposed to recover velocity and pressure information from an image. Results are highly dependent on the accuracy of each stage, in particular the extraction of the strokes from the image and the determination of the strike order (Diaz et al., 2017). In general, a complete recovery of the dynamical parameters of the stroke is unachievable. An extended description of the dynamical parameters in forensic sciences is given by Linden et al. (2018).

In this thesis we consider handwritten evidence in the forms of off-line natural writing, off-line signatures, and off-line forged signatures.

1.3.2 Factors influencing handwriting

Handwriting is the result of a very complex process, thus it may be influenced by a multitude of factors.

Huber and Headrick (2010) divide them into two large categories: *intrinsical* and *extrinsical* factors.

Intrinsical factors can be at least partially controlled by the writer. Among these, we cite the influence of the physical environment with which the writer interacts. As an example, the effect of body posture on writing under normal (sitting at a table) and unusual conditions (kneeling on the floor) was investigated in Sciacca et al. (2008). The authors show that the natural variation does not depend strongly on the analyzed body postures. However, the actual shape of the written material was not analyzed. A successive study confirmed the findings on two new writing postures (standing and lying down) and considering a vertical orientation of the writing surface (Sciacca et al., 2011). The authors exercised care in interpreting these findings, as the studies were limited in their experimental settings. The change of aspect ratio of signatures as a function of posture was investigated by Thiéry et al. (2013). Results show that some of the writers were sensitive to the adopted posture, while others were not. Notice that the aspect ratio is not robust to accidental variations such as longer ending strokes (Thiéry et al., 2013).

The change of shape when artificially forcing a writing size was taken into account in Marquis et al. (2007). The authors analyzed the shape of closed character loops, first written in writers' normal size of writing, then three times larger. The same Fourier methodology as in Marquis et al. (2005) was applied to parametrize the shape, and a discriminant analysis classifier was trained with the goal of discriminating writers. Results show that loop shapes were modified in the same way by most (but not all) writers during the enlargement process, by increasing the roundness and decreasing the slant. However, the classifier performed badly when trained only on one version of the set of characters. In other words, the authors underline the necessity of acquiring reference specimens with the same size of questioned material, to eliminate the possibility of an artificial enlargement of the handwriting which could lead to misleading evidence.

A factor which is relevant to signatures is the influence of the space that surrounds them. It is frequent in forms and contracts that signatures have to be written on an horizontal guide line, or in a box. A few studies are available in literature. Morton (1980) show that an external constraint usually results either in a reduction of the horizontal dimension, or a miniaturization of the whole signature, with little alterations of the remaining writing aspects. Fazio (2015) studied the effect of external constraints, both in on-line and off-line signatures. Results support the

hypothesis that the constraints have an effect on a signature, increasing in strength as the constraints become stricter. Concerning off-line signatures, the effect is not as pronounced as on-line signatures, but the authors advise caution, and recommend searching for reference specimens that exhibit the same limitations.

The second category of factors are those which are not controlled by the writer, the so-called extrinsical factors. Among them, one extrinsical factor which is relevant to the scope of this thesis is the genetic influence on handwriting.

The best way to study its effects is to consider handwriting in twins, who share most of their genetic material. It is also reasonable to suppose that most twins share the same environment and same education, at least in the first years of life. Available studies which involve handwriting of twins conclude that twins can be discriminated from unrelated persons, albeit at a higher error rate (Ahuja et al., 2018; Dziedzic et al., 2007; Gamble, 1980; Srihari et al., 2008; Thorndike, 1915). A famous case involving twins is the so-called Dionne quintuplets of Canada, whose handwriting has been reported to be differentiated even during the twins' formative years (Huber & Headrick, 1999, sec. 25).

Another extrinsical factor which could potentially play a major confounding role during an expertise is the influence of the writing system. In older times, handwriting was taught using one of the standard models, the so-called copybooks. The influence of the copybooks persisted as writers grew, notwithstanding the acquisition of a set of personalized habits (according to the principle of habituation). Groups of writers could have been differentiated by recognizing the taught copybook (*class characteristics*) (Huber & Headrick, 1999). Nowadays, however, handwriting is taught differently: pupils learn to write by copying the teacher's handwriting on the blackboard, rather than a prescribed copybook, reaching for legibility rather than the fidelity in shape and good penmanship. As a consequence, the influence of the writing system is generally greatly diminished, if not absent in signatures (Kelly & Lindblom, 2006, pp. 60–61). A counterexample has recently been observed, showing that Polish writers may still possess a stronger copybook influence than English writers. However, the study included participants with a very wide age range (Turnbull et al., 2010).

Class characteristics commonly also include habits that are shared by groups of writers. For example, it has been reported a greater similarity of the handwriting between members of a family (Kiran & Sridhar, 2017). Notice that this effect might be a confounder of the aforementioned studies conducted on handwriting similarity across twins. The same phenomenon may happen during adolescence, when writers may be exposed through their peers to a particular style of writing. This was the case for the so-called bubble writing, a particularly round style diffused among female adolescents in the North-Eastern United States (Totty, 1991, p. 120).

Nowadays, most features which are interesting to the purpose of writing identification spur from the individual habits acquired during the formative years. However, FDEs still need to be aware of class characteristics, for instance when the case involves older people, foreign alphabets, or the design of particular letters.

It is important to remind that the reference material should have been produced, as far as possible, in the same period of the questioned material (Sulner, 2018, pp. 648–649). This is due to the fact that the writer's habits may change over time. If the reference specimens are not sufficiently representative, differences in writing characteristics may wrongly weigh towards the hypothesis of non-authenticity, rather than the modification of a writer's habits. As the writer reaches old age, the effect of time may also be associated with other factors, such as a decline in health. In that case, writing may manifest increased tremor, loss of design quality and, in conclusion, will have a greater probability to err, due to the absence of lines of reasoning to deal with uncertainty (Kelly & Lindblom, 2006, p. 85).

1.4 The forensic document examination process

An important outcome of the post-Osborn research is the rationalization of the process of forensic handwriting examination. FDEs adopted the so-called "ACE-V" process, which is also shared by other forensic fields such as fingermark analysis (Huber & Headrick, 1999).

In short, evidence is examined by experts in three separate phases:

- *Analysis*: the principles governing the specific forensic discipline are exploited in order to extract elements to be analyzed. In this case, writer's habits are identified from the writings, and the most discriminating elements extracted. This is done first on the questioned material, then on the reference material.
- *Comparison*: the questioned material is compared with the reference material, to collect similarities and differences between the discriminating elements identified in the *Analysis* step.
- *Evaluation*: the outputs of the previous steps are considered in the light of the hypotheses of interest, and the evidential value of the findings is assessed. The experts' report is finally stated in a verbal or numeric form. Commonly (and unfortunately), experts state their conclusion as an identification, an elimination or declare evidence to be inconclusive.

A fourth step, *Validation*, is typically added to the chain, and involves the blind re-examination of the evidence by one or more experts in order to confirm or contradict the conclusions.

The third step, *Evaluation*, produces what is ultimately communicated in court or to the stakeholders: to correctly conduct an evaluation is of utter importance. Many miscarriages of justice have occurred due to fallacies in evaluative reporting. DNA evidence is particularly sensitive to these issues, but an improperly conducted evaluation can also occur in other domains (Gill, 2012). Forensic document examination is not free from these issues either. One of the most widely known examples is the Dreyfus case, where a group of mathematicians led by Poincaré pointed out that the forensic expert Bertillon had drawn fallacious conclusions in his report (Bertillon, 1901; Mansuy & Mazliak, 2008). For an historical review of the case, see (Taroni et al., 1998), (Champod et al., 2000) and (Kaye, 2007).

To dispel the possibility of any miscarriage of justice, the evaluative step is the subject of several guidelines recently established by various forensic institutes, calling for a trans-disciplinary standardization of evaluative reports. In the UK the AFSP guidelines were published in 2009 (Association of Forensic Science Providers, 2009), while European forensic institutions (ENFSI) published their own version in 2015 (Biedermann et al., 2017; Willis et al., 2015).

These guidelines require forensic experts to acknowledge the usage of probability into uncertainty assessment for evaluative reporting. The whole reasoning behind the formulation of the report should follow the probability axioms. The result of the evaluation (i.e. the strength of the evidence) is to be stated using an expression called Likelihood ratio (or, more properly, Bayes factor). Such an approach has also recently been adopted by the Document Section of the Canadian Society of Forensic Science (CSFS) that feels the necessity to express a formal position about the use of an alternative evaluation and reporting scheme, often referred to as "the logical approach to evidence evaluation" (Ostrum, 2019).

Interestingly, it has been observed that the adoption of these guidelines should affect the whole ACE process. In fact it is helpful to consider which results are more discriminative according to the evaluative hypotheses of interest, before proceeding with the physical analyses on the collected evidence (Biedermann et al., 2017; Cook et al., 1998b). This has also been observed for handwritten evidence (Stockton & Day, 2001).

1.5 Issues in forensic document examination

The usage of handwritten evidence in court has historically faced many issues, and still does nowadays.

1.5.1 What has been solved

Handwritten evidence is very different from evidence analyzed in other forensic fields, such as physics, chemistry, biology and genetics (Huber & Headrick, 1999). One of the differences lies in the fact that handwriting is the result of a chain of interactions that start in the human brain and involves the neuromuscular system. The conscious act is a fundamental component of the writing process, along with the precision of the repetition of the neuromotor task. Moreover, there is no physical law that dictates how handwriting is produced and evolves over time. This markedly contrasts with other kinds of evidence, for example DNA. We know that DNA in an individual is mostly immutable, and the way that DNA propagates across generations is described by the laws of genetics. Further theoretical research provided a set of theoretical models that describe how populations can be characterized by DNA. Major technological improvements in terms of DNA laboratory analyses have finally given origin to DNA profiling, the usage of DNA for identification (Jeffreys et al., 1985). In other terms, the existence of a strong theoretical, empirical and scientific background is grounds for the successful admissibility of DNA evidence in court (Buckleton et al., 2018; Champod et al., 2017).

On the other hand, handwritten evidence did not enjoy the same luxuries. As DNA evidence gradually was affirmed as *the* model for forensic identification science (Saks & Koehler, 2005), the admissibility of handwritten evidence was the subject of strong criticism, most notably in the United States of America.

In 1989 Risinger strongly affirmed that forensic handwriting examination lacked any empirical bases, and many of the existing studies were either flawed, anecdotal, speculative or non-rigorous (Risinger et al., 1989). This article was itself severely criticized in FDE community (see for example Moenssens (1997–1998), p. 300), but it started to raise several issues that grew in importance in the following years.

Until 1993, evidence resulting from forensic handwriting examinations was admitted in the courts of the United States on the basis of *Frye v. United States*[1]. Under Frye, expert opinion evidence is accepted as long as it is based on principles generally accepted by the scientific or technical community where it belongs (Frye's *general acceptance* test) (Jamieson & Moenssens, 2009, p. 1331).

In 1993, in the case *Daubert v. Merrell Dow Pharmaceuticals, Inc.*[2], the US Supreme Court established four criteria that expert evidence should satisfy in order to be admissible for testimony. First, *Frye*'s general acceptance test was demoted to be a necessary condition. The other criteria concerned the scientific foundations

[1]Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).
[2]Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

of the methodology used by the expert. Admissible methodologies would have been subjected to peer review, would be testable through falsification, and error rates would have been available (Jamieson & Moenssens, 2009, pp. 693–694). Moreover, the judge was allowed to exclude any evidence not meeting all the aforementioned criteria (the "gatekeeper role"). A later decision (case *Kumho Tire*[3]) further expanded the scope of the *Daubert* criteria to all kinds of evidence, scientific and not. In 1995, in the US District Case *United States v. Starzecpyzel*[4], following the *Daubert* requirements, the judge affirmed that handwriting examination "has never been validated as credible scientific or technical knowledge", but is only a "technical skill". The courts' decisions, at their face value, would have excluded from the courtroom many professional figures due to the lack of published error rates, such as members from the social sciences (psychiatrists and psychologists), forensic pathologists and fingerprint examiners (Moenssens, 1997–1998).

This chain of events severely impacted the whole forensic area, and stimulated a major rethinking of the forensic handwriting discipline in the subsequent years. The *Daubert* and *Kumho Tire* trials brought to attention how the scientific bases of forensic handwriting expertise could have been justified and improved, and raised the relevant question of how performant examiners are in what their profession claims to be able to do (Berger, 2011, p. 32).

Concerning the scientific foundations, it has been recognized that empirical research on handwriting is important, as it leads to fundamental and factual knowledge about the phenomenon of handwriting. Such information is valuable. First, it adds to the theoretical background that supports operational techniques for comparisons. Secondly, it allows one to formulate recommendations on which handwriting features can be exploited by FHE according to their discrimination power. Thirdly, once these features are determined, one can evaluate the performance of FHE by establishing a set of standardized proficiency tests, providing an unambiguous measure of reliability of the examiners, for example by assessing their error rate. Nevertheless, some of the FHE practitioners have themselves been pushing against the recognition of forensic handwriting analysis among more "scientific" disciplines, as most of the founding principles (such as the principle of uniqueness) are not experimentally verifiable. The same practitioners, instead, advocate embracing uncertainty and focusing on testing the reliability of experts (Sulner, 2018).

The issue of expert performance on FHE tasks was investigated by several scholars after the *Daubert* trial. Recent studies were conducted, comparing trained forensic document examiners against laypersons, both on handwriting and on signatures.

[3]Kumho Tire Co., Ltd v. Carmichael, 522 U.S. 136 (1997).
[4]United States v. Starzecpyzel, 880 F. Supp. 1027 (S.D.N.Y. 1995).

(Found et al., 1999; Found et al., 2002; Found & Rogers, 2003; Kam et al., 1994, 1997, 1998, 2001; Kam & Lin, 2003; Risinger, 2007) In general, it was found that FHE perform their tasks at a lower error rate than laypersons, supporting the actual expertise of handwriting examiners.

Another result of research on FHE performance is the evaluation of the intrinsic difficulty of the tasks. For example, Found & Rogers (2008) have shown that FHE are more confident in expressing an authorship opinion on genuine signatures rather than simulated or disguised ones, and do so with an error rate as low as 2.5%. On the other hand, if the signature is simulated or disguised, expressing an authorship opinion is a more difficult task, and the error rate can rise up to, respectively, 6.9% and 40.1%. Also, FHE tend to favor the inconclusive opinion rather than call for an authorship or an elimination (Found & Rogers, 2008). Laypersons, instead, are less conservative than FHE, calling more often for the wrong decision (Sita et al., 2002). It has been observed that the worsening of the performance on disguised signatures might be due to the production of the specimens, not to the examiner's skill. In fact, it appears that some people may fail to disguise their signatures, producing specimens whose natural variation is consistent with their own master pattern (Michel, 1978).

Two of the largest FDE organizations, namely the American Society of Questioned Document Examiners (ASQDE) and the American Board of Forensic Document Examiners (ABFDE), created several certifications to train and demonstrate the reliability of practitioners (Day, 2009). Several other certifications exist (Huber & Headrick, 1999, sec. 67).

The post-*Daubert* evolution of the forensic handwriting examination discipline, such as the aforementioned quantitative studies by Srihari et al. (2002), and the establishment of reliability assessments for the forensic profession, finally resulted in the successful readmission of handwritten evidence in court in the case *United States v. Prime*[5] (Harralson, 2014; Kelly & Lindblom, 2006). However, the many critics raised in the *Daubert*-era still remain.

It is important to note that these issues affect also other judicial systems, albeit to a lesser extent than the US. Swiss courts, for example, allow the usage of handwritten evidence as long as the expert is able to well justify its usage. The many critics, thus, could potentially be used to invalidate experts' justifications (Marquis, 2007).

1.5.2 What has not been solved

Despite the many post-*Daubert* efforts, the forensic handwriting discipline is still facing a number of open issues.

[5]United States v. Prime, 220 F. Supp. 2d 1203 (W.D. Wash. 2002).

One is related to the push for a stronger evaluative reporting across forensic science, for instance through the adoption of the ENFSI Guideline for Evaluative Reporting in Forensic Science (Willis et al., 2015). The new standard was modeled on the Bayes factor approach to forensic interpretation of DNA profiles, calling for a reporting that is *balanced, robust, transparent and logical* (Aitken et al., 2021; Willis et al., 2015). Since its conception, it has been successfully applied to many areas (Meuwly et al., 2017) including, but not limited to: glass evidence (Curran et al., 2000; Curran, 2003), fingermarks (Neumann et al., 2006; Neumann et al., 2007), speaker recognition (Gonzalez-Rodriguez et al., 2007), questioned documents (Biedermann et al., 2011), fibers (Grieve et al., 2017), transfer evidence (Samie et al., 2016), footwear marks (Evett et al., 1998), fire investigation (Biedermann et al., 2005), drug seizure classification (Bozza et al., 2014).

However, the thorough application of the Bayesian forensic approach to handwritten evidence is lacking, as its implementation is not trivial. The first difficulty is the need for a definition of a set of features to be extracted from the handwritten evidence. Due to the complexity of handwriting, this task cannot be standardized but must be conducted on a case-by-case basis, possibly under the guidance of a FHE. Consequently, machine learning-based approaches are already barred out since the features are often chosen to be highly discriminative for a large number of cases, irrespective of their similarities/differences as evaluated by a FHE.

Once the features are defined, the second difficulty is the assignment of probabilities to their relative rarities. A naïve adoption of the Bayesian forensic approach requires the direct elicitation of probabilities by experts, a task that may be revealed to be counter-intuitive and non-trivial for multiple reasons. For example, there exist multiple definitions of the term *probability* that conflict in fundamental ways, for instance in the attribution of a numerical value to an uncertain event (see for example Köller et al. (2004), Cosmides & Tooby (1996) and Taroni et al. (2018) for a discussion). It has recently been shown by Martire et al. (2018) that FDE are not yet fully comfortable with the assignment of the probability of occurrence of certain written characteristics from one own's experience. The resulting likelihood ratio values might therefore be flawed (Martire et al., 2018). It is important to note that the study does not disprove the existence of the skills that FDE claim to have. It focuses, instead, on the reliability of conclusions as given by FDEs under the very specific form of the likelihood ratio value.

We conjecture that this issue might be resolved with the adoption of appropriate evaluative procedures that are both grounded in computational methods and consistent with the forensic Bayesian framework. Probabilities would be elicited with the help of casework data, and the resulting computations would automatically be developed

by a thoroughly tested computer program. All possible detrimental biases would therefore be restricted to the choice of the data and the statistical model at hand. Robustness checks are available, to further verify the dependence of the computed quantities to the assumptions of the methods, even before any evidence is collected (Cook et al., 2006; Gelman et al., 2009; Vehtari & Ojanen, 2012). To our knowledge, to date only a few works consistent with this approach are available (Bozza et al., 2008; Marquis et al., 2011; Marquis, Hicks, et al., 2019; Taroni, Marquis, et al., 2014).

Notice, however, that the complete procedure is sensitive to the creation of the statistical model, a task that does not typically pertain to the educative path of a forensic document examiner. A successful application of the Bayesian approach would require the contribution of forensic scientists from other areas, and specialists familiar with it from other scientific disciplines. We dare to say that all areas of forensic science would benefit by this collaborative work to find a common language for handling the problem in question.

The push for an increase in interdisciplinarity would contribute to the solution of a second major issue that affects forensic handwritten evidence examination as it is taught today, that is the inability to evaluate the value of evidence spanning multiple domains (see (Taroni, 2005) for an example involving handwriting and fingerprints). To this purpose, the Bayesian evaluative framework proposes the usage of probabilistic graphical models (Bayesian networks) as basic tools to formalize the rational thinking (Biedermann, 2007). These have been successfully applied to many problems in forensic science (see for example Taroni, Biedermann, et al. (2014)), but the integration of handwritten evidence is a problem that has not yet been explored.

Another issue concerns the tools that the forensic document examiners use in their profession. Recalling the fundamental principles governing handwriting, it would be pretentious to replace the human skill of an FDE with the output of a computer program at this stage of research. Past attempts such as the aforementioned systems FISH, CEDAR-FOX and WANDA, encountered limited enthusiasm among FDE practitioners, although some of them were explicitly developed for police forces (for instance GRAPHJ (Guarnera et al., 2018)). A Bayesian approach, however, is not limited to *hard data*, but allows scientists to include *any* kind of variables into the model that formalizes the evaluative reasoning. A properly constructed model would allow the FDE not only to supervise the procedure, but also to react to the change of the evidential scenario. For example, scientists may be asked whether their conclusion would change (and if so, in what way) if further information would become available, such as additional reference specimens from the suspect. We conjecture that tools delivered under the Bayesian approach could be revealed to actually be useful to experts, as they are designed to *assist*, not replace them.

1.6 Originality of the current research

On June 4-5, 2013, the National Institute of Standards and Technology (NIST) hosted the Measurement Science and Standards in Forensic Handwriting Analysis Conference in Gaithersburg, Maryland. NIST planned and organized this event in collaboration with the American Academy of Forensic Sciences (Questioned Document Section), the American Board of Forensic Document Examiners (ABFDE), the American Society of Questioned Document Examiners (ASQDE), the Federal Bureau of Investigation Laboratory, the National Institute of Justice, and the Scientific Working Group for Forensic Document Examination. The general discussion focused on the future of forensic handwriting analysis. Three major points were emphasized. First, the future of the discipline will incorporate the use of more quantitative analysis tools to handle the handwriting examination process. Secondly, forensic document examiners would like to use soundly based statistical models to explain the significance of results. Thirdly, researchers should publish more studies involving the use of quantitative methods for examinations, which will both improve the understanding of these advancements and validate examination methods by converting research into the best practice that examiners can incorporate into their standard operating procedures.

This thesis aims to be a major research contribution towards the points highlighted in the NIST conference and the ENFSI guidelines for evaluative reporting. In particular, this thesis will provide four types of contributions:

1. *Empirical*: our results are based on sound research as working hypotheses, the fundamental principles governing handwriting, and state-of-the-art results such as Marquis' shape descriptors (Marquis et al., 2005).
   Firstly, we investigate the usage of these descriptors on the same dataset as in the original article, in the context of natural handwriting. Secondly, we apply these descriptors in a novel context with a purposely collected dataset, comprising forged and genuine signatures that contain character loops. Thirdly, we use these descriptors to investigate the quantitative differences in handwriting in twins, adding to the limited existing research available in literature. The dataset has been collected on purpose. Finally, we provide an entirely novel descriptor that can be used to quantify a class of particular signatures (*simple signatures*) that do not show sufficient complexity to allow for a clear evaluative examination by FHEs using the traditional elements of style and execution.
   In this case, the dataset consists of real casework data. The latter descriptor is extensible to other kinds of research questions spanning other forensic domains, such as the comparisons of two sets of proportions of multiple items.

2. *Statistical*: we adopt a rigorous Bayesian approach, from data collection to communication.
   We mainly focus on the so-called evaluative scenario: under this setting, two items of evidence are compared, the source of one item is known (for example, a suspect), the other is not (e.g. a questioned document). The forensic scientist must, then, evaluate and quantify the value of the evidence at hand against (at least) a pair of hypotheses on the source of the questioned material. Results will be expressed in the form of Bayes factors, as requested by the ENFSI guideline[6]. As the computation of Bayes factors is not a trivial task, we propose several computational techniques with the aid of specialized libraries for R, the platform for statistical computing (R Core Team, 2019). In particular, we first exploit the model used to quantify the probative value of loop shapes, which has been already validated on natural handwriting in current literature (Marquis et al., 2005). The model is described in (Bozza et al., 2008; Bozza et al., 2014). Its operative implementation has been revised and improved: this enabled us to verify its performance on novel datasets as well as assessing its intrinsic sensitivity to assumptions. Secondly, we introduce a novel statistical model for the descriptor used in the context of simple signatures. This model allows us to compute Bayes factors for novel scenarios, as well as to assess the sensitivity of this metric to assumptions on the prior parameters. This thesis would, therefore, contribute to the almost non-existent body of research concerning the usage of the Bayesian framework in forensic handwriting examinations.

3. *Transdisciplinary*: we show how the descriptors for simple signatures can be used to evaluate the value of evidence in a completely unrelated domain (i.e. microbial composition of human saliva), and how to jointly evaluate it along with handwritten evidence. The Swiss National Science Foundation has supported a joint research between the École des Sciences Criminelles (ESC) of the University of Lausanne (UNIL) and the Institute of Microbiology of the Centre Hospitalier Universitaire Vaudois (CHUV) to deal with the collection, analysis and joint evaluation of handwritten characters and salivary microbiota of twins. With the proposed approach we will we able to tackle, for the first time, casework that involves, for example, a handwritten letter and a trace in the form of a salivary stain, in which classic DNA analyses cannot be applied (due to degradation or the claim of a twin brother as the alternative source).

4. *Methodological*: the programs that have been developed to perform the necessary computations are packaged into a set of fully documented open-source packages

[6]The ENFSI guideline uses only the term “Likelihood ratio”, not “Bayes factor”.

for R, ready for use by the entire forensic community: `bayessource` (Gaborini, 2019), `rstanBF` (Gaborini, 2020a), `rdirdirgamma` (Gaborini, 2020b). Moreover, the procedure of extraction of numerical data from the acquired images required the development of several graphical interfaces for the programs R and MATLAB (The Mathworks, 2019). These interfaces are easy to use, and can be adapted to specific demands of interested users.

This thesis has been supported and financed by the grants no. 10001A_156290 (concerning goals 1, 2 and 4) and no. 10531C_170280 (concerning the transdisciplinary objectives).

1.7 Research questions

This thesis rests on several hypotheses that are commonly stated in FHE literature. We assume that writers have a master pattern which can differ significantly across persons. A writer's natural variations manifest themselves in terms of departures from a writer's handwriting "master pattern". Moreover, the master pattern is assumed to be reasonably stable, as long as the time window is limited. In other terms, the features in a writers' handwriting are characterized by an intra-variability and an inter-variability, and are quantified according to the descriptors that have been developed in this thesis. The statistical models that we developed translate the interaction between intra- and inter-variability into a Bayesian model separating the two components. It is important to note that the usage of the Bayesian framework as well as the adopted design of the experiences allow us to operatively verify whether and when this hypothesis does not hold, namely by obtaining a numerical expression for the value of evidence that points to the "wrong" hypothesis (e.g. there is evidence that the questioned material comes from the reference writer, whereas the questioned material has been taken from another person).

In this thesis we consider a set of scenarios that differ in the type of evidence under consideration, as well as which statistical models are used to quantify its value. The goal of each scenario is to establish a complete procedure to evaluate the value of evidence in a particular context. Particularly, each scenario consists in a step of data acquisition, a step of development of an appropriate statistical model, a step of validation of the statistical properties of the model (e.g. sensitivity), and the establishment of a procedure to compute Bayes factor values.

a. When handwritten material contains character loops, we exploit Marquis' Fourier descriptors as well as the Bayesian model for evidence evaluation. The method

has already been validated on natural writing in past works (Bozza et al., 2008; Marquis et al., 2005; Marquis et al., 2011). In this thesis we apply the descriptors to other kinds of handwriting, namely questioned signatures, and natural handwriting coming from related persons (i.e. twins). The first goal of the thesis is to establish whether this method can be successfully applied to cases that were not considered during their developments.

b. If the handwritten material does not contain character loops, we introduce a novel descriptor based on the variability of specified proportions inside a particular signature. By imitating the Bayesian model of the previous case, we introduce an analogous model to assess the value of evidence in this case. The second goal of this thesis is to assess whether the new descriptors and the new model are able to quantify the value of evidence under an alternative context.

It is important to note that these methods seek to describe the variability of specific features (i.e. loop shapes, and proportional distances), not the handwriting as a whole. As a consequence, our conclusions cannot be taken as a substitute for an expert analysis, but can quantitatively support FHE in their work (Thiéry, 2014, p. 13).

c. During the development of the proportional descriptors, it occurred to us that it might have a much broader scope of application, in particular to describe all kinds of data related to proportions. To provide an example, we apply these descriptors and the proportional model to data consisting in the composition of the microbiota in adult twins. The third goal of this thesis is to quantitatively evaluate whether microbiota differ in twins rather than unrelated persons.

d. The combination of the three goals allows us to consider cases of forensic interest, namely the recovery of a salivary trace and a handwritten item from the crime scene, and a putative source related to the suspect, such as a sibling or a twin.

An element transversal to these goals is the usage of a Bayesian perspective to assess the value of evidence, in particular through the Bayes factor. This raises a number of significant technical and operative difficulties, for instance when the (a) evidence consists of multivariate data, (b) when the amount of recovered data is insufficient, or (c) the background information might be lacking. Also, Bayes factors themselves attracted significant criticisms, both from forensic scientists (on the appropriate definition, see for example (Hepler et al., 2012; van den Hout & Alberink, 2016)) and statisticians (on their sensitivity as well as their operative relevance) (Kamary et al., 2014, p. 3; Morey et al., 2016). These issues, where relevant, will be isolated and discussed.

1.8 Thesis structure

The thesis is organized as follows. Chapter 2 gives an introduction to the use of Bayesian statistics in forensic science. Chapters 3 to 5 are dedicated to the analyses of the collected datasets. Each chapter is independent of each other, and builds on the generic framework introduced in Chapter 2. In general, each chapter is dedicated to a particular type of evidence: character loops and simple signatures. Chapter 5 is an exception, as we will exploit the results of the previous chapters in a completely different context, the combined evaluation of handwritten and microbial evidence in twins. Each chapter contains a discussion on the improvements, future extensions and open problems that resulted from the analysis of each case. Chapter 6 is a review of Chapters 3 to 5, showing problems and characteristics shared by the different contexts. The structure of the thesis ,detailing datasets and statistical models, is resumed in Table 1.1.

In more detail, in Chapter 2 we briefly review the basic concepts behind the theory of Bayesian statistics and the usage of Bayes theorem in forensic science. Afterward, we show how a typical evaluative forensic scenario can be modeled with this framework. We explain the Bayes factor, the main quantitative tool that is used to report evaluative conclusions in court (Willis et al., 2015). As the calculation of the Bayes factor is mathematically challenging outside very simple cases, we show how it can be performed with the help of numerical methods, introducing them in a symbolic form. These methods will be concretely used in the subsequent chapters, given the appropriate models, data and hypotheses for the considered cases.

In Chapter 3 we briefly recall Marquis' shape descriptors (Marquis et al., 2005) as well as the Bayesian two-level model from the state-of-the-art literature (Bozza et al., 2008). First, we introduce our implementation of the model, showing how it agrees and differs with the original article, and test its performance on two datasets, a simulated dataset and the one used in the original articles. Next, once the code is verified on natural handwriting, we apply the method to the context of forged and genuine signatures containing character loops. A second dataset is introduced and described, consisting of a set of forged and genuine signatures containing several loops. The method is then applied, and the results discussed.

In Chapter 4 we consider an actual casework consisting of a set of reference and questioned signatures. From now on, we call them *simple signatures*: by the term *simple* we refer to the apparent lack of strongly discriminating features. As these signatures contain no loops, we introduce a descriptor tuned to the case at hand. The general framework of Chapter 3 is adapted to the specificities of this Chapter by stating a Bayesian model for the newly introduced descriptor. Its implementation is

discussed, and the Bayes factors calculated for the considered casework.

In Chapter 5 we consider a context in which available evidence consists of samples of natural handwriting (in the form of character loops, as in Chapter 3), and the composition of salivary microbiota. This context stems from a novel dataset, collected in a collaboration with the Institute of Microbiology of the University of Lausanne. The donors of the samples are adult monozygotic twins. Evidence is analyzed, first separately using the results developed in Chapters 3 and 4, respectively. Afterward, we approach the problem of combination of items of evidence, to reach a single evaluative conclusion.

Table 1.1: list of datasets and descriptors.

Name of dataset	Data source	Chapter	Descriptor
Natural handwriting	(Bozza et al., 2008)	3	Fourier
Questioned signatures with loops	Ad-hoc	3	Fourier
Questioned simple signatures	Casework	4	Proportional
Natural handwriting in twins	CHUV	5	Fourier
Salivary microbiome	CHUV	5	Proportional

Chapter 2

The Bayesian framework

2.1 On uncertainty

Who is the likely source of this blood trace? How similar is the smudged fingermark with the suspect's prints? How often does one encounter a particular kind of paint chip among all cars that were present in a given region? Forensic science is intrinsically concerned with uncertainty, spanning almost every aspect of it.

Data is collected to gain insight into the unobserved. Laboratories are often calibrated to deliver measurements within some established error bounds. Traces may be of poor quality compared to standards. Standards are just a simplified representation of the physical object, limited by the properties of the support or the instrument that created them. Samples can be destructively collected, altering subsequent analyses on the same material. Samples may have been improperly stored, or the chain of custody may have been breached.

How do these findings relate to the material in possession of the investigators?

Context is uncertain. We might not know the identity of a victim, or the source of a particular trace. We might consider the suspect among a panel of potential offenders. A suspect might disprove his own association with a particular crime scene, yet evidence is pointing at him. Or we might know the source of the trace, but not its actual involvement with the offender or the chain of actions. Has this trace been deposited for innocent reasons? Is the offender the same person that delivered the mortal blow to the victim? Traces are *not* present, whereas they should have been in a hypothetical situation. How do these findings change our belief about this particular hypothesized chain of events?

Events are uncertain. Witness reports are not always reliable. Forensic experts may state their qualified opinion on the basis of wrong assumptions (hence the need

for the validation step according to the "ACE-V" process). Even the apparently simple affirmation "The suspect is the source of the trace" has no truth or falsity in general[1]. More complex affirmations such as "This blood trace contains DNA from at least two persons" are associated with a degree of belief that must consider multiple issues related to the laboratory results, such as the possibility of a contamination.

Clearly, all those questions cannot be answered with a clear "yes" or "no", but a more nuanced answer is required. This answer must connect data, context and events in the most reasonable way, taking into account the intrinsic uncertainty characterizing the three aspects.

The appropriate language of choice is probability theory (Taroni et al., 1998). Historically, it has been introduced to forensic science already in 1894 during the Dreyfus case. However, some major flaws in the probabilistic reasoning were highlighted in the second appeal (1904) by a group of mathematicians led by Poincaré.

Probabilistic reasoning was used in more subsequent cases, not without controversies. In particular, in *People v. Collins*[2] (1968) an argument involving probability was fallaciously used to convict a couple based on a witness' description, and back-of-the-hand statistics on the occurrence of the witnessed attributes. Another major error was the so-called *fallacy of the transposed conditional*, that manifests itself in statements such as(Aitken et al., 2021, p. 189):

> There is a 10% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 90% chance that he is guilty.

In fact, one may prove that it is not possible to obtain information on the probability of the defendant being the offender without considering other elements, such as the circumstantial evidence and the prior belief about the statement (i.e. the belief about the truth of the statement before having collected *any* evidence: in this case, the witness report).

One can identify many other related fallacies, commonly collected under the name of *fallacies in interpretation* (Aitken et al., 2021). As the name implies, these fallacies escape intuition, and can cast confusion during a fair court trial process. The appearance of even more sensitive forensic techniques, such as Low Copy Number (LCN) DNA profiling, is a double-edged sword if the conveyed information is misused by an incoherent reasoning scheme.

The possibility of making such errors can be ruled out once the reasoning is formalized into an appropriate probabilistic language, in particular by adopting the

[1] Unless the trace has been deposited under controlled conditions, or there is a reliable declaration by the suspect.

[2] People v. Collins, 68 Cal.2d 319.

Bayesian model as the main reasoning scheme. This was already observed initially during the second appeal of the Dreyfus case, in 1904 (Taroni et al., 1998, p. 190). The mathematicians called for the usage of Bayes' theorem, the pillar of Bayesian probability, as *the* only way to correctly apply a probabilistic reasoning to a forensic case, if a coherent quantification is to be made[3]. However, the Bayesian framework was not popularized until the aftermath of *People v. Collins*, more than sixty years later.

2.2 Probability theory and forensic science

As we said before, forensic science is fundamentally concerned with uncertainty. Probability theory is a branch of mathematics that is concerned with quantifying uncertainty, and does so by using numbers.

There exist multiple branches of probability theory that differ mostly in the way the uncertainty is considered (e.g. how it relates to the outcome of future events or whether one can perform repeated experimental trials to change uncertainty). However, these branches agree on the basic rules governing the mathematical objects that translate uncertainty into formulas, the so-called *Kolmogorov axioms* (1933). Common to all approaches is the possiblity to associate a numerical value (in the $[0, 1]$ interval), called *probability*, that quantifies the uncertainty of an event.

We share the Bayesian view: this definition always applies on the outcome of all events, past, present and future, without referring to their repeatability. It is also allowed to assign a probabilistic value to events that are entirely hypothetical, such as the affirmation "The sun will not rise tomorrow". We accept that this value might differ depending on the person who assesses it, as it translates *his own* belief about the uncertainty to a single number. Due to this aspect, Bayesian probability is said to be *subjective*. However, the combination of beliefs is mathematically sound, and does not actually depend on the adopted probabilistic branch, but only to the Kolmogorov axioms. For instance, De Finetti proved that it is mathematically correct to attribute a numerical value to one's beliefs on uncertain quantities under mild conditions, and showed how to do it by means of betting strategies (De Finetti, 1930; De Finetti et al., 1964).

We will skip most details related to the construction of this set of axioms and the

[3]Actually, the group of mathematicians contested the application of mathematics to moral matters due to its dangerousness (Taroni et al., 1998). However, a properly constructed Bayesian approach limits itself to the quantification of the value of the collected evidence, whereas any decision upon guilt or innocence is left to the judge. Thus, forensic scientists need not to be concerned with moral matters, but only by measurable facts.

related theorems. In this section we recall what is needed to show how the Bayesian framework is used in a forensic context. We refer the interested reader to dedicated texts, such as the classic tome from Jaynes (2003), Berger (1985) and Lindley (1971) for an introduction to Bayesian probability and its links with decision theory, and Aitken et al. (2021) for the utilization of Bayesian probability in a forensic context.

In forensic science one reasons about the relationship between items of evidence, noted with the generic letter E, to the hypotheses of interest, generically represented with the random variable H. Usually, H will assume the values h_p (for the prosecution hypothesis) and h_d (for the defense hypothesis). More than two states are allowed. Context-related information can be exploited, for example the frequency of occurrence of a certain characteristic in a given reference population. This can be indicated with the letter I, but it is usually omitted in subsequent equations for simplicity of notation.

For example, E might be the observation of a correspondence between the genotypes of a blood trace found on a crime scene and that of a person of interest (suspect or victim). $H = h_p$ might indicate the statement, "The blood trace has been left by Mr. X.". $H = h_d$ might indicate the statement, "The blood trace has been left by somebody else.". I might contain the information on the occurrence of the blood type found on a crime scene.

In real cases, the source of the trace is typically unknown, therefore we recur to the usage of probability to assign a value value to our belief about the relationships between E, h_p and h_d. For instance, $\Pr(E = e \mid h_p)$ represents the probability that the blood has a particular genotype e, assuming that it came from Mr. X (h_p). The operator $\Pr(\cdot)$ translates the probability of the event in the argument to a number in the closed interval $[0, 1]$. When not ambiguous, we will use it also to indicate the probability that a discrete random variable X takes on a particular value x. We also consider continuous random variables: in this case $\Pr(\cdot)$ will indicate the probability density function of a random variable evaluated in the argument. Deviations from this rule will be made explicit when they occur. For brevity, we will often avoid the distinction between a random variable and its realization. When the distinction is needed, we usually employ upper-case letters for the former, and lower-case letters for the latter.

The vertical bar appearing in $\Pr(\cdot \mid \cdot)$ statements expresses a conditioning, i.e. we assume complete knowledge of the variable(s) on the right-hand side of the vertical bar.

2.2.1 The Bayes' theorem and the Bayes factor

A fundamental consequence of the Kolmogorov axioms is the so-called Bayes' theorem. In its simplest form it states:

Theorem 2.1 (Bayes' theorem)
Let A, B be two events, with $\Pr(B) > 0$. Then:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}\,.$$

An equivalent expression can be obtained if one considers the complementary event $\bar{A}$ that happens if and only if A does not hold. By dividing the Bayes' theorem with the equivalent form, one can restate the Bayes' theorem as follows:

Theorem 2.2 (Bayes' theorem in odds form)
Let A, B be two events, with $\Pr(B) > 0$. Then:

$$\frac{\Pr(A \mid B)}{\Pr(\bar{A} \mid B)} = \frac{\Pr(B \mid A)}{\Pr(B \mid \bar{A})}\frac{\Pr(A)}{\Pr(\bar{A})}\,.$$

Considering a generic forensic case, assume that the evidence E has been observed (e.g. the genotypes of a recovered stain and of a person of interest), assuming a particular value e (e.g. the genotype AB).

Our view of subjective (i.e. Bayesian) probability allows us to associate a corresponding number to hypotheses. This represents our personal belief about their truth or falsity. Namely, $\Pr(H = h_p)$ and $\Pr(H = h_d)$ indicate one's prior belief about the prosecution and defense hypotheses, respectively. The ratio of those two numbers is called *prior odds* in favor of h_p.

Notice that in this case the events $H = h_p$ and $H = h_d$ are complementary, so the term *odds* is fully justified. If the considered hypotheses are not exhaustive (i.e. the probability of their union is 1), the odds ratio more properly describes the *relative odds* in favor of h_p against h_d (Aitken et al., 2021, p. 108).

In the forensic context we usually make the following substitutions in Theorem 2.1[4]: A is the event $H = h_p$, $\bar{A}$ is the event $H = h_d$, B is the event $E = e$. Then one obtains:

$$\frac{\Pr(H = h_p \mid E = e)}{\Pr(H = h_d \mid E = e)} = \frac{\Pr(E = e \mid H = h_p)}{\Pr(E = e \mid H = h_d)}\frac{\Pr(H = h_p)}{\Pr(H = h_d)}\,. \tag{2.1}$$

[4]Here we highlight explicitly the role of H. For the sake of brevity, H is often left out in favor of its states h_p and h_d. Analogously, a similar reasoning is done for the evidence E: in simple forensic models, evidence is always observed, therefore e can be substituted to E without ambiguity.

On the right side of Equation (2.1), one immediately recognizes the prior odds term $\Pr(H = h_p)/\Pr(H = h_d)$. A corresponding term can be located on the left side, $\Pr(H = h_p \mid E = e)/\Pr(H = h_d \mid E = e)$. This is called *posterior odds* in favor of h_p.

The remaining term is the so-called *Bayes factor* (BF) (Kass & Raftery, 1995). It is improperly known in forensic science as *Likelihood ratio*, as in this specific case (for simple vs simple hypotheses) it is constituted by a ratio of likelihoods. In general, it might be different: e.g. see (Taroni et al., 2010, p. 53) for a discussion.

In other terms, Equation (2.1) may be rewritten as:

$$\text{posterior odds} = \text{BF} \times \text{prior odds}\,. \tag{2.2}$$

From this, we *define* the Bayes factor to be the ratio of the posterior and the prior odds:

Definition 2.1 (Bayes factor)
The Bayes factor of h_p against h_d for the evidence e is:

$$\text{BF} := \frac{\Pr(H = h_p \mid E = e)}{\Pr(H = h_d \mid E = e)} \Big/ \frac{\Pr(H = h_p)}{\Pr(H = h_d)}\,.$$

Forensically speaking, the relative belief about the hypotheses of interest can be assessed before any evidence is collected. For example, one may establish *a priori* the probability that the suspect is the source of the blood trace. This is usually done by the mandating authority (Willis et al., 2015).

Collected evidence may change the relative belief about the propositions of interest. We can also say that the evidence has a certain *value* for/against the considered hypotheses. The way it occurs is described by Equations (2.1) and (2.2), namely through the action of the Bayes factor. There is *no* other logical way to update our beliefs using evidence, provided that we accept the Kolmogorov axioms, the postulates of Bayesian probability and the generic probabilistic model for a forensic case[5] (Good, 1991). Notice also that the Bayes factor is a single number, since both the prior odds and the posterior odds are numbers (for a discussion on this aspect see (Taroni, Bozza, et al., 2016)). We believe that using the above definition of the Bayes factor encourages practitioners to develop the statistical model around the question of interest to the court ("What are the posterior odds in favor of h_p ?"), necessarily introducing the connection of the hypotheses to the observed evidence.

[5]Actually, it is not even necessary to recur to Bayesian probability to state that the Bayes factor is the only way to measure the value of evidence in favor of an hypothesis. See Aitken et al. (2018). It becomes necessary once one uses Bayesian theory to measure his belief about such hypothesis.

2.3 Bayesian networks

The definition of conditional probability and Bayes' theorem have a graphical representation that can be used to explain and present statistical models and probabilistic reasoning in general. This allows the adoption of a graphical representation and reasoning process, for instance, by immediately showing which are the interconnections and the dependence structure between random variables of interest. Graphical models are also easier to present to laypersons, without introducing heavy mathematical jargon that would risk being incomprehensible.

Consider three random variables X, Y and Z along with their joint probability distribution $\Pr(X, Y, Z)$. It can be shown that the joint probability distribution can be factorized into a product of conditional probability distributions. Since the decomposition is not unique, modeling effort is put forward to identify which factors are representative of the statistical problem at hand. For example, suppose that we judge $\Pr(X|Y, Z)$, $\Pr(Y)$ and $\Pr(Z)$ (these two do not depend on any other) as relevant factors. Then, the joint probability distribution is: $\Pr(X, Y, Z) = \Pr(X|Y, Z)\Pr(Y)\Pr(Z)$.

This assumption can be represented graphically with a particular kind of graph called *Bayesian network*. Bayesian networks are directed acyclic graphs (DAG for short), where each node represents a random variable, and arcs represent the conditional probabilistic relationship from the source nodes (on the right of the conditioning symbol) to the target node. In this specific case, one may represent $\Pr(X, Y, Z)$ with the following Bayesian network:

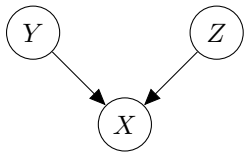


Figure 2.1: A simple Bayesian network.

All nodes (i.e. random variables) have a probability distribution on their *states* (i.e. the values of the random variables). When the node has no parents, its distribution does not depend on other variables, and is assumed to be its prior distribution. Nodes can be observed: the information carried by this operation is propagated according to Bayes' theorem to the rest of the network, in what is commonly called *Bayesian updating*. Notice that information can flow in both directions across an arc. In the forward direction, the information is propagated through the definition of the conditional probability of the node. Backwards, it is propagated using Bayes' theorem. Once one or more nodes have been observed and the information distributed to

the whole network, all node distributions are the respective posterior distributions, conditioned on the observed node(s).

A major trait of Bayesian statistical models is that they both allow inference from an observed dataset, as well as generate datasets from the described joint distribution. Due to this property, they are also called *generative*.

Another desirable property of Bayesian networks stems from the interaction between probability theory and graph theory. The interplay between those two aspects results in a set of criteria (known as d-separation (Pearl, 1988)) that allows for the isolation of specific sets of nodes by considering only the topology of the network (i.e. interconnections). Across these sets, information may or may not flow depending on the state of the nodes separating these sets. This results in a concrete functional separation between network components, that translates to an encoding of the dependence structure to a macroscopic scale.

The usage of Bayesian networks in forensic and legal applications enables practitioners to model complex scenarios that involve multiple types of evidence and multiple hypotheses in an unified framework. The graphical aspect, such as the intrinsic directionality of the propagation process, helps in eliminating reasoning errors and fallacies that could have easily been made otherwise. A deeper look into Bayesian networks is available in (Nielsen & Jensen, 2009) and (Kjaerulff & Madsen, 2008). An introduction to Bayesian networks in forensic science is available in (Taroni, Biedermann, et al., 2014) as well as in (Biedermann, 2007) and (Gittelson, 2013).

There exist also multiple software programs that allow practitioners to define their desired Bayesian network, assess their prior knowledge and obtain the posterior distributions for any node of choice. These programs are, however, often limited to the subset of networks where all nodes are discrete or, at best, continuous Gaussian random variables. In these cases, the posterior distributions can be computed in an exact form. If the network contains nodes with more general distributions, an automatic computational theory is no longer available, and the inference process needs to be carried out using some approximate inference procedures. Often, these solutions are developed on an ad-hoc basis, or are confined to research applications (Korb & Nicholson, 2010, ch. 3).

2.3.1 Marginalization

A fundamental operation that must be performed to propagate belief between two distant nodes in a Bayesian network is the so-called *marginalization*.

Suppose that a three-node Bayesian network is available:

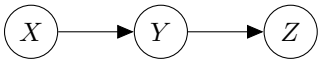


Figure 2.2: A three-node Bayesian network.

This Bayesian network can be used in forensic scenarios to model the uncertainty tying the hypotheses (X) to the collected evidence (Z) through a distinct node Y.

Example 2.1 (A glass fragment)**.** For example, suppose that a fragment of glass is recovered from a crime scene. Two glass objects, a broken bottle and a window fragment, are seized: their refractive index Y is known only with a large uncertainty (for instance, no laboratory measurements are available). The hypothesis node is used to discriminate whether the glass pieces came from the bottle or the window item (states $X = x_1$ and $X = x_2$). Then, Z can be modeled as a continuous random variable, for instance Gaussian, with the mean parameter corresponding to the refractive index of the true source, and the standard deviation corresponding to the laboratory measurement error. However, the mean of Z is not known, and depends on the state of X. Through the operation of marginalization, all uncertainty on the intermediary node Y is exploited to connect nodes X and Z. For example, if it is certain that object 1 has a refractive index greater than 1.6, any observation of Z lower than 1.6 would heavily support the hypothesis that the fragment came from object 2.

Suppose that one is interested in evaluating the change of belief in Z after observing a realization of X. To do so, the law of total probability[6] and the definitions of conditional probability can be applied:

$$\Pr(Z = z \mid X = x) = \int \Pr(Z = z \mid Y = y) \Pr(Y = y \mid X = x) \, \mathrm{d}y \,, \qquad (2.3)$$

where the symbol of integration in Equation (2.3) becomes a simple sum if the evidence assumes discrete values.

If Z is a continuous random variable and the function $\Pr(Z = z \mid Y = y)$ represents the density function for the data z with parameter y, the term $\Pr(Z = z \mid X = x)$ is also called the *marginal likelihood* for the data z under x. If X is discrete and acts as a model indicator, the marginal likelihood can also be seen as the *prior predictive density* for Z, evaluated in z, under the model $X = x$. One can also define the *posterior predictive density*, considering a situation where the belief about

[6] In the forensic context it is also called "*extension of the conversation*" (Lindley, 1991).

Y has been already updated through a past observation of z_b, another realization of Z. The argument of the integral would change from $\Pr(Y = y \mid X = x)$ to $\Pr(Y = y \mid X = x, Z_b = z_b)$, to indicate the updated belief.

To reconnect with the forensic context, notice that the left-hand side of Equation (2.3) is the numerator of the Bayes factor, where the evidence is $Z = z$ and the prosecution hypothesis is $X = x$. Y is a node which is never observed, but models latent properties that are relevant to describing the dependence of the evidence Z from the hypotheses X. An analogous term appears in the denominator, with the respective defense hypothesis in the place of x. This motivates the paramount importance of being able to compute Equation (2.3) in actual cases.

Notation

Here we introduce a concise notation for the marginal likelihood in Equation (2.3). For clarity, this notation is shared with the article that suggested the hierarchical structure of the models (Bozza et al., 2008). The marginal likelihood m for the evidence z given the hypothesis x over variable Y is:

$$m(e \mid h) := \Pr(Z = z \mid X = x) = \int \Pr(Z = z \mid Y = y) \Pr(Y = y \mid X = x) \, \mathrm{d}y \,. \quad (2.4)$$

2.3.2 On causality

As the factorization of the joint probability distribution is not unique, the Bayesian network is not unique (Dawid, 2008). This means that Bayesian networks do not necessarily encode a causality relation, but simply describe the structure of the data. Moreover, forensic science is not concerned with issues of causality, but only with describing which hypothesis is more supported by the collected data.

For the sake of completeness, causality structures *can* be represented by a DAG. This way of thinking was recently popularized by Pearl et al. (2016), and is currently encountering a large support in many scientific communities. Particularly, it helps in addressing issues that, up to now, were often improperly modeled in experimental studies, such as the "correlation $\neq$ causation" debate, determination of counterfactual effects, or deciding which variables need to be measured to account for confounders (i.e. unobserved common causes or associations). However, not all scientists agree with this approach to causality modeling: see for example (Gelman, 2011; Krieger & Davey Smith, 2016) for a discussion.

DAGs developed for causal inference share many properties with Bayesian networks, yet it can be shown that causal interventions, called *do*-operators, involve the modification of the structure of the graph (i.e. cutting incoming arcs) instead of a

simple instantiation of the affected node. This operation is not typically done in Bayesian networks, as forensic scientists are not concerned with issues of causality but only a probabilistic description of the relation between hypotheses and data (Taroni, Biedermann, et al., 2014, sec. 2.1.11).

2.4 Forensic scenarios

In forensic science, one can distinguish two common situations, depending upon if a putative source of a questioned item is available (Taroni et al., 2012).

2.4.1 Investigative scenario

The *investigative scenario* (or investigative setting) occurs when evidence (a questioned item) has been collected, but no person of interest (control item) is available to provide reference material. Equivalently, the pair of hypotheses may no longer refer to a single putative source, but to broader statements that concern a population of potential sources (e.g. "the recovered sample comes from a Caucasian male population or an alternative population").

The forensic procedure under an investigative scenario is aimed at providing support to investigating authorities, for instance by narrowing the search for a person of interest to a particular set of candidates (de Zoete et al., 2017; Jackson et al., 2010), or by helping to establish connections between cases providing that the evidence is relevant to the case at hand (Taroni et al., 2006).

It is also the scenario under which searches for matching evidence (e.g. DNA) against a database are conducted. In this case, if the procedure results in a correspondence between objects, investigation may proceed to an "evaluative" scenario once the person of interest is apprehended. Other items of evidence from the alleged source can then be exploited for comparative purposes. Hypotheses of interest also shift into an evaluative mode, for example by considering the activity of the person of interest during the offense.

Probability theory can be applied under an investigative scenario, and a Bayes factor can be computed (Taroni et al., 2012). Considering the illustrative Example 2.1 for the glass fragment, we suppose that no laboratory measurements on the bottle and window fragments are available. The Bayes factor may be used to decide which items should be sent to the forensic laboratory, and which kinds of objects can be further seized. The hypotheses are only concerned with the generic source of Z: Z is a fragment of a bottle glass (under $X = x_1$), or is made up of window glass (under $X = x_2$). This results in the investigative Bayes factor:

$$\mathrm{BF} = \frac{\Pr(Z = z \mid X = x_1)}{\Pr(Z = z \mid X = x_2)}, \tag{2.5}$$

where the numerator and the denominator must be computed using the marginalization procedure.

During the early phase of this thesis, an application of probability theory to handwritten evidence has been developed and published (Gaborini et al., 2017). We admittedly used the plug-in approximation to the Bayes factor for illustrative purposes (see Section 2.6.2), as more sophisticated methods (e.g. the implementation of the bridge sampler used in Chapter 4 (Gronau et al., 2017)) were not available at the time. A discussion of the investigative Bayes factor on anonymous handwritten documents can be found in (Bozza, 2015, sec. 3).

2.4.2 Evaluative scenario

Notice that the investigative scenario involves only one type of evidence, whose source is unknown. In the glass example (Example 2.1), only Z was available. Imagine that one of the glass objects was broken apart under laboratory-controlled conditions, and the refractive index of one of its fragments (the "control fragment") was measured. Let us represent the measured refractive index on the control fragment with the random variable C. The hypotheses X reported in the investigative scenario (i.e. the recovered fragments come from object 1, or from object 2) would also change to X':

$X' = x'_1$: "the glass fragment Z and the control fragment C come from the same glass object",
$X' = x'_2$: "the glass fragment Z and the control fragment C come from (two) different glass objects".

This pair of hypotheses characterizes the so-called *evaluative* scenario. Notice that now the evidence is the recovered item and the control material; the evidence contributes to the update of beliefs on the states of X'.
In this case, the Bayes factor is the following:

$$\mathrm{BF} = \frac{\Pr(Z = z, C = c \mid X' = x'_1)}{\Pr(Z = z, C = c \mid X' = x'_2)}. \tag{2.6}$$

An example for the calculation of a Bayes factor on glass fragment data can be found in (Aitken & Lucy, 2004). This methodology was further extended in an application to handwritten data in (Bozza et al., 2008) and (Bozza, 2015). In this thesis we will consider only evaluative scenarios, adapting the methodology developed

in (Bozza et al., 2008) to account for characteristics such as the intra- and the inter-variability of source parameters.

2.5 Hierarchical models

In this section, the general form of the probabilistic models that we developed for this thesis is described. The notation is purposely generic, to give a unified description of the procedures that will be adopted in the subsequent chapters, and to determine an expression for the computation of the value of the evidence through the Bayes factor. For instance, all probability distributions will be left unspecified, and no details will be given on individual evidence items as well as their types. The form of the model follows the one developed in (Bozza et al., 2008) for handwriting comparison. In particular, one can distinguish a dependence structure that spans multiple levels, describing the evidence and the sources in turn. In statistical literature, this model structure is called *hierarchical*.

First, we consider only evaluative scenarios, where the source of N_{q} questioned items is disputed. N_{r} reference items from the putative source are available. We indicate these sets of items as e_{q} and e_{r}, respectively. Each set is a collection of random variables that represent the measurements, possibly multivariate, on the single item. For the sake of clarity we could expand $e_{\mathrm{r}} = \left\{e_{\mathrm{r},i}\right\}_{i=1}^{N_{\mathrm{r}}}$ and $e_{\mathrm{q}} = \left\{e_{\mathrm{q},i}\right\}_{i=1}^{N_{\mathrm{q}}}$, but we will often drop the index subscript unless necessary.

The evaluative hypotheses are:

$$\begin{aligned} H = h_p &: \quad \text{“the items } e_{\mathrm{r}} \text{ and } e_{\mathrm{q}} \text{ come from the same source”}, \\ H = h_d &: \quad \text{“the items } e_{\mathrm{r}} \text{ and } e_{\mathrm{q}} \text{ come from two different sources”}. \end{aligned}$$

Notice that these hypotheses are not exhaustive. For example, we do not consider cases where one of the e_{q} came from the putative source while others did not. In other words, all the e_{q} either came from the putative source, or another unseen source, belonging to the reference population. The exhaustive model for two total traces ($N_{\mathrm{r}} + N_{\mathrm{q}} = 2$) can be found in (Gittelson et al., 2013): in that case, inference is noticeably more complicated even if inference is restricted to binary-valued random variables.

We assume that the properties of the questioned and reference sources are described by the latent source parameters θ_{q} and θ_{r}, respectively. For instance, in the glass example (Example 2.1), θ_{q} indicates the refractive index of the entire glass object, which is never measured but only inferred through the observations e_{q}. In handwriting, it indicates the intra-variability component, or *within*-source variation (Bozza et al.,

2008). We suppose that e_{r} and e_{q} are observations from the respective source, conditionally i.i.d. on the respective latent source parameter θ, with density function $f(\cdot;\theta)$.
The evaluative hypotheses can be translated on θ to a pair of events H':

$$\begin{aligned} H' = h'_p &: \quad \text{“}\theta_{\mathrm{r}} = \theta_{\mathrm{q}}\text{”}\,, \\ H' = h'_d &: \quad \text{“}\theta_{\mathrm{r}} \text{ and } \theta_{\mathrm{q}} \text{ are independent”}\,. \end{aligned}$$

Remark. In principle, H' could be viewed as a new hypothesis pair with a different nature from H. For instance, H are discrete and mutually exclusive hypotheses, H' involve statements on continuous variables that are no longer mutually exclusive. It is also clear that h_p implies h'_p, and h_d implies h'_d, so H is more restrictive than H' (more details are given in Sections 2.6.3 and 2.6.4). However, the “hypothesis pair” H' is free from forensic interest, as the connection with the alleged source of the questioned material is no longer explicit. Also, we must remember that we are testing H, not H'. Pragmatically, we conflate H and H', e.g. assuming h'_p whenever h_p is valid, and we avoid indicating H' in the formulae. In other words, we are saying that sources are completely characterized by their latent parameters θ. Two different sources have independent parameters θ_1 and θ_2.

The model can be resumed using the common Bayesian model notation:

$$\begin{aligned} e_{\mathrm{r}} \mid \theta_{\mathrm{r}} &\overset{iid}{\sim} f(e_r;\theta_{\mathrm{r}})\,, \\ e_{\mathrm{q}} \mid \theta_{\mathrm{q}} &\overset{iid}{\sim} f(e_q;\theta_{\mathrm{q}})\,. \end{aligned} \tag{2.7}$$

The upper level of the hierarchical model is now introduced, to model the uncertainty and the dependence between the within-source parameters. We assume that θ_{q} and θ_{r} are i.i.d. observations from the *between*-source prior distribution g, itself characterized by a latent parameter ψ, with density distribution $g(\cdot;\psi)$.
Using the common Bayesian model notation:

$$\begin{aligned} \theta_{\mathrm{r}} \mid \psi &\sim g(\theta_{\mathrm{r}};\psi)\,, \\ \theta_{\mathrm{q}} \mid \theta_{\mathrm{r}}, \psi, H = h_p &\sim \mathbb{1}_{\{\theta_{\mathrm{r}}\}}(\theta_{\mathrm{q}})\,, \\ \theta_{\mathrm{q}} \mid \psi, H = h_d &\sim g(\theta_{\mathrm{q}};\psi)\,, \end{aligned} \tag{2.8}$$

where $\mathbb{1}_{\{\theta_{\mathrm{r}}\}}(\theta_{\mathrm{q}})$ means that θ_{q} can assume only the value θ_{r}.

There are other possible parameterizations that try to disambiguate the interactions between e_{q}, θ_{q} and H. For instance, we could introduce a latent node θ_{t} that models the source properties of the *true* source for e_{q}. θ_{t} becomes θ_{r} under h_p, and

θ_{q} under h_d. Under every hypothesis, e_{q} are assumed to be i.i.d. samples from the true source with density $f(e_q; \theta_{\mathrm{t}})$.

Notice that another layer in the hierarchy might be added on the bottom, for example to take into account three sources of variation such as the error introduced by multiple measurements of the same item (Bozza, 2015, p. 190).

All models in this thesis will differ in the choice of f, g and the distributions for θ and ψ.

2.5.1 Background observations

Information on ψ might not be readily available. However, a set of background observations can be exploited to elicit a distribution for ψ[7]. The way the observations are exploited is discussed in Section 2.5.3.

We assume that the background dataset comprises M_{b} sources (unrelated given ψ), where the i-th source is characterized by $N_{\mathrm{b}i}$ observations. For clarity, we avoid the double subscript for background observations, indicating them collectively as e_{b}. We assume that the model describes all observations coming from every considered source, be it in the background, reference or questioned set. As a consequence, background sources will be characterized by their own latent source parameter θ_{b}. Moreover, we assume that the between-source variability is described by the random variable ψ. ψ becomes the only random variable that conditions the background dataset, the reference dataset and the questioned dataset.

2.5.2 Bayesian network

The model can be represented with a Bayesian network, as in Figure 2.3. The root node ψ conditions all parameters θ, be they reference, questioned or background sources. The plate notation indicates repeated nodes, as many times as the indicated number of repetitions. The leftmost part models the background dataset, with M_{b} sources. The middle part models the reference items, coming from a single source (i.e. the single θ_{r}), that does not depend on the hypothesis H. The right part models the questioned items, coming from a single source. According to the evaluative hypotheses (Section 2.5), the single source is θ_{r} under $H = h_p$, and θ_{q} under $H = h_d$.

[7]Notice that we could put a prior distribution on ψ, but there would be no apparent way to elicit its hyperparameters in this simplified model.

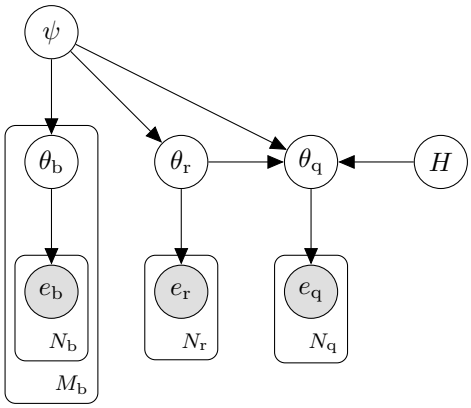


Figure 2.3: The generic two-level hierarchical model. The plate notation indicates repeated elements, whose subscript has been omitted. Shaded nodes are observed.

2.5.3 Evidence propagation

The graphical model may also illustrate how the beliefs are updated using the available data. First, the background information e_b is collected, thus the node is instantiated. Taroni, Biedermann, et al. (2014) collapse e_b into the context information I, to stress that it is already known before e_r and e_q are observed: instead, we keep indicating e_b for clarity. Next, as ψ separates[8] the left (background) and the right (casework) part of the Bayesian network, this procedure can be split into three phases:

1. the background data e_b updates the first level of the network, θ_b, then the root node ψ. Now the prior $\Pr(\psi)$ has become the posterior $\Pr(\psi \mid e_\mathrm{b})$.
2. The updated knowledge on ψ is exploited to update θ_r and θ_q (generically indicated with $\theta_\bullet$). Now we are able to evaluate $\Pr(\theta_\bullet \mid \psi, e_\mathrm{b})$[9].
3. The densities for the casework evidence e_r, e_q can be evaluated by marginalizing over $\theta_\bullet \mid \psi, e_\mathrm{b}$. The Bayes factor is a ratio of these values evaluated in the hypotheses states.

In most cases, the addition of the second level of the Bayesian network results in the general impossibility of either obtaining an analytical expression for $\Pr(\psi \mid e_\mathrm{b})$, or the posterior distribution of ψ itself, requiring the usage of a numerical method.

[8] Once ψ is observed, information cannot flow from the left to the right part of the network and vice versa. This is a consequence of the d-separation criterion. See (Pearl, 1988) for details.

[9] In (Cereda, 2017), they equivalently rewrite the probability $\Pr(\theta_\bullet \mid \psi, e_\mathrm{b})$ to a new probability $\widehat{\Pr}(\theta_\bullet \mid \psi) := \Pr(\theta_\bullet \mid \psi, e_\mathrm{b})$.

The posterior for ψ is in general known only through a set of samples from a Monte Carlo method, possibly autocorrelated. This fact represents a further step in the quantification of the Bayes factor.

The methods exploited in this thesis require the knowledge of the distribution of $\psi \mid e_\mathrm{b}$. To prove this, we remember that the Bayes factor can be computed as a ratio of marginal likelihoods. From Equation (2.6), Z is e_r, C is e_q, X' is H. Numerator and denominator can computed by marginalizing the posteriors on θ_r, θ_q and ψ (Equation (2.3)). The integrand would, therefore, involve $\Pr(\psi \mid e_\mathrm{b})$ that is not known.

The plug-in approximation

The solution we adopted is to replace the previous Step 1 with a two-part procedure:

1a) the background data e_b is used to obtain $\widehat{\psi}$, a point estimate of ψ, for instance using its maximum likelihood estimator,

1b) the propagation continues, collapsing $\psi \mid e_\mathrm{b}$ to a point mass located in $\widehat{\psi}$.

Step 2 is therefore replaced with the evaluation of $\Pr(\theta_\bullet \mid \psi, e_\mathrm{b})$ from the model, using the point estimate for ψ. We are approximating $\Pr(\theta_\bullet \mid \psi, e_\mathrm{b}) \approx \Pr(\theta_\bullet \mid \widehat{\psi})$ $(= g(\theta_\bullet; \widehat{\psi})$ within the hierarchical model). After part 1b., the procedure continues to Step 3, where the marginalization on $\theta_\bullet$ can be done.

This procedure is similar to the "plug-in" estimation (see Equation (2.12)), and can be related to the so-called *empirical Bayes* method (Casella, 1985), and the "Bayesian plug-in method" described in (Cereda, 2017).

The *d*-separation criterium makes clear the dependence between casework evidence (e_r and e_q) and background information (e_b), namely through the knowledge of ψ. To approximate ψ with $\widehat{\psi}$ means that the Bayes factor estimate, which is obtained by considering the right part of the Bayesian network (as in Figure 2.3), depends on the background only through the point estimate $\widehat{\psi}$.

Compared to the many criticisms reported in Section 2.6.2 and in literature such as (Cereda, 2017), however, a "proper" marginalization step is retained, and the approximation is performed in the between-writer layer of the Bayesian network, far from the evidence layer. As long as a well-justified point estimate for ψ is available (for instance by estimating ψ on a large background dataset, or eliciting ψ from another expert's knowledge), we suppose that the obtained Bayes factor is less suffering from the issues that affect the "plug-in LR". We also perform extensive sensitive analysis to evaluate the sensitivity of the Bayes factor to the assumed $\widehat{\psi}$. Predictive

checking procedures (see Section 2.6.5) represent another tool that can be used to this purpose. More pragmatically, this approximation allows us to tackle the computation of the Bayes factor, that otherwise would not have been possible with the considered methods.

2.5.4 The Bayes factor

Once the model is defined, the Bayes factor follows by performing the marginalization in Equation (2.6) on the writer's parameters θ. Using the notation introduced in Equation (2.4), the marginal likelihood for the evidence e given the hypothesis h is:

$$\begin{aligned} m(e \mid h) &= \Pr(E = e \mid H = h) = \\ &= \int \Pr(E = e \mid \Theta = \theta, H = h) \Pr(\Theta = \theta \mid H = h) \, \mathrm{d}\theta \,, \end{aligned} \tag{2.9}$$

where Θ denotes the writer's parameters seen as a random variable.

Consider the generic two-level hierarchical model (Section 2.5), with the plug-in approximation for the background parameter (Section 2.5.3).

We are interested in quantifying the Bayes factor for the hypothesis pair h_p against h_d, with evidence e_r and e_q. We suppose that ψ is known, approximated with the point estimate $\widehat{\psi}$ according to Section 2.5.3. It can be proven that the Bayes factor in Equation (2.6) can be written as:

$$\mathrm{BF} = \frac{m(e_\mathrm{r}, e_\mathrm{q} \mid h_p)}{m(e_\mathrm{r} \mid h_d) \, m(e_\mathrm{q} \mid h_d)} \,. \tag{2.10}$$

Explicitly, all marginal likelihoods with the model notation are:

$$\begin{aligned} m(e_\mathrm{r}, e_\mathrm{q} \mid h_p) &= \int f(e_\mathrm{r}; \theta) \, f(e_\mathrm{q}; \theta) \, g(\theta; \widehat{\psi}) \, \mathrm{d}\theta \,, \\ m(e_\mathrm{r} \mid h_d) &= \int f(e_\mathrm{r}; \theta) \, g(\theta; \widehat{\psi}) \, \mathrm{d}\theta \,, \\ m(e_\mathrm{q} \mid h_d) &= \int f(e_\mathrm{q}; \theta) \, g(\theta; \widehat{\psi}) \, \mathrm{d}\theta \,. \end{aligned} \tag{2.11}$$

All evaluative situations can, therefore, be reduced to the problem of calculating three marginal likelihood values.

2.5.5 Scenario simulations

During the thesis two kinds of operative conditions may be encountered:

1. *background-dominant*: e_b is provided, but not necessarily e_r and e_q,
2. *evidence-only*: e_r and e_q are provided, but not e_b.

The *background-dominant* situation is important for the development of the methodology, as it enables us to explore the structure of the data, and choose a statistical model that allows for the calculation of the Bayes factor. Usually, we assume that e_b is large, i.e. is constituted by many observations coming from many different sources.

In this thesis, the first situation is particularly important, as it is most often encountered when dealing with simulated datasets (i.e. data generated from a model which is completely known, including all latent parameters). Their analysis allows us to understand whether the implemented models behave correctly, for instance by obtaining correct estimates of the true parameter values (when known, such as in a artificially generated dataset), and examining the general properties of the computed Bayes factors such as their asymptotic behavior, or the influence of prior parameters. This phase is called *model validation*.

The background-dominant situation is also encountered when e_b comes from a situation where obtaining a large background dataset is relatively easy. This was the case for the dataset of character loops in natural writing.

The *evidence-only* situation typically occurs during casework where no background data is initially available. As we have seen before, the computation of a Bayes factor requires a validated statistical model, and informed knowledge of the between-source parameter ψ. These prerequisites can be satisfied if one is able to collect information on ψ by searching in available literature, or by *creating* background data (under the hypothesis that the generated data can be modeled by the same model that describes e_r and e_q). For instance, if two naturally written manuscripts are compared, and the authenticity of one is disputed, a forensic scientist can request a background corpus e_b from a large relevant population, to inform his knowledge on ψ. If either the literature search or the simulation succeed, we fall back to a background-dominant situation, where knowledge of ψ is available.

The evidence-only situation was encountered twice in this thesis. When dealing with forged signatures with character loops (Chapter 3), a background dataset comprising specimens coming from other, unseen, forgers, was collected. In that case, the model had already been validated on simulated data as well as in Bozza et al. (2008).

When dealing with simple signatures in a real casework, a model validated on simulated data was proposed, under the assumption that the simulated data contains relevant information on ψ. In this case, request forgeries could not have been obtained due to privacy concerns.

Bayes factors distribution

It is easy to pass from a background-dominant to an evidence-only situation. This procedure will be named *scenario simulation*, as it allows one to simulate the performance of our model as if one were confronted with real casework data. Moreover, one is no longer restricted to a fixed pair e_r and e_q, being free instead to exploit the entire background dataset, leveraging on pure computational power. This procedure is similar to the so-called *cross-validation*, typically used in the machine learning context to train robust classifiers and statistical models.

To simulate a casework scenario, one must consider a pair of mutually exclusive hypotheses h_p and h_d. The output of a scenario simulation is a list of Bayes factor values, computed when one of the hypotheses is true. Their distribution[10] allows one to verify, for example, whether the model is able to obtain Bayes factors that point to the correct hypothesis. Whenever this result does not hold, it means that the model is unable to describe the data under the assumed hypotheses. Ruling out programming errors, it implies that one or more model assumptions are wrong.

Scenario simulation: the procedure

The fundamental prerequisite for scenario simulation is the set of sources M in the background dataset, supposed large. We also suppose to know the source of the i-th item in the background dataset, indicated with $m_i \in M$.

Firstly, we choose the size of e_r and e_q, i.e. how many evidence items are collected in the simulated casework. These are indicated with k_r and k_q, respectively.

Then, at the g-th step:

1. A random source for the reference items, ${m_\mathrm{r}}^{(g)} \in M$ is picked. The eligible sources will, in general, coincide with M.
2. A random source for the questioned items, ${m_\mathrm{q}}^{(g)} \in M$ is picked. Here one chooses the hypothesis to be tested: for instance, under h_p, ${m_\mathrm{q}}^{(g)} = {m_\mathrm{r}}^{(g)}$.
3. One randomly picks k_r items from the background dataset, to form ${e_\mathrm{r}}^{(g)}$.
4. One randomly picks k_q items from the background dataset, to form ${e_\mathrm{q}}^{(g)}$.
5. The casework background ${e_\mathrm{b}}^{(g)}$ is formed by items that do not appear in any of the ${e_\mathrm{r}}^{(g)}$ and ${e_\mathrm{q}}^{(g)}$.
6. The tuple $e^{(g)} := \left({e_\mathrm{b}}^{(g)}, {e_\mathrm{r}}^{(g)}, {e_\mathrm{q}}^{(g)}\right)$ forms a hypothetical dataset that is used to obtain a *single* value of the Bayes factor, $\mathrm{BF}^{(g)}$, using the procedure described

[10] With "distribution" we mean their descriptive statistics, such as their histogram. We do not assume that Bayes factors are random variables. For a critical discussion of the misunderstandings surrounding this issue, the interested reader may refer to (Taroni, Bozza, et al., 2016).

in Section 2.5.3.

The collected Bayes factor values $\left\{\mathrm{BF}^{(g)}\right\}_{g=1}^{G}$ form the desired distribution, which is analyzed to the purpose of the simulation.

There are many choices that can be tuned during this procedure. For instance, the samples ${e_\mathrm{r}}^{(g)}$ and ${e_\mathrm{q}}^{(g)}$ can be picked either with or without replacement. One may also want to make sure that any item appears at most once across all $e^{(g)}$. The background ${e_\mathrm{b}}^{(g)}$ may be restricted to sources that *never* appear in any of the casework items ${e_\mathrm{r}}^{(g)}, {e_\mathrm{q}}^{(g)}$. Depending on the structure of H, step 2 can be modified to allow picking multiple sources for the questioned material. In this case, one may also want to observe each questioned source at least once in ${m_\mathrm{q}}^{(g)}$.

In this thesis we assume that sampling is always done without replacement, and each item appears at most once in $e^{(g)}$. Further assumptions and details will be specified when needed.

2.6 Issues and criticisms to the Bayes factor

2.6.1 Bayes factor for model choice

The Bayes factor is often cited in introductory texts to Bayesian statistics as a first approach to the problem of hypothesis testing. In frequentist statistics, this is usually done by measures such as the p-value (a value p such that $p := \Pr(E \geq e|H_0)$, where the evidence E is observed to be e, and H_0 is the null hypothesis; if p is lower than a fixed threshold, then H_0 is rejected in favor of another hypothesis[11]).

This approach is known as *null hypothesis significance testing* (NHST). p-values are particularly easy to compute, as one needs to assume a null hypothesis H_0 (which is in general uninteresting) and identify a distribution of the evidence (more generally, a test statistic) under H_0. However, this recipe is treacherous for a number of reasons that will be explained later in this Chapter.

The widespread use (or, more properly, misuse) of p-values in multiple scientific branches, in particular life and social sciences, was one of the factors that resulted in widespread replication failures of published studies, in what has been known as the "replication crisis". As a consequence, p-values used as a measure of evidence attracted strong criticisms in recent years, culminating in suggestions such as lowering the minimum p-value thresholds for publication (Benjamin et al., 2018), replacing

[11] Notice here the fallacy of the transposed conditional, and how the alternative hypothesis does not play any role in the definition of the p-value (although it does when the sample size is decided). A discussion can be found in (Taroni, Biedermann, et al., 2016).

p-values with more appropriate measures (Benjamin et al., 2018), rethinking the approach to hypothesis testing (Haaf et al., 2019), or abandoning completely *any* procedure that results in a dichotomization of the result (Johnson, 1999; McShane et al., 2019). Similar objections were raised in matters of justice and forensic science many years ago, anticipating the current replication crisis (Kaye, 1986).

The Bayes factor is often chosen as one of the replacements for p-values, as it *is* the formal definition of the value of evidence (Good, 1991). Also, it is less treacherous than p-values: for instance, it avoids the fallacy of the transposed conditional.

However, the use of Bayes factors as a replacement for the old experimental designs, such as NHST, is not problem-free. For instance, the structure of the design phase (i.e. the Bayesian equivalent of the notion of "effect size", which prior distributions should be chosen, and how many items of evidence must be collected to obtain compelling Bayes factor values) is still under discussion (Schönbrodt & Wagenmakers, 2018). A recent statistical discussion on the specific form of the competing hypotheses under consideration can be found in Etz et al. (2018). Note that this does not represent a theoretical criticism, just an operational one.

The recent increase in the utilization of Bayes factors resulted also in an increase of discussion and criticisms, notably from statistic branches, but also from forensic scientists.

Criticisms coming from statisticians mostly address issues related to the general properties of Bayes factors, for instance their sensitivity to parameters or their statistical well posedness and behavior. These aspects are generally avoided if a subjectivist point of view is assumed and accepted.

Forensic scientists have debated (and still do), instead, on issues such as the proper way to define the forensic scenario to the purpose of computing a Bayes factor, and how the Bayes factor value can be effectively communicated to the court. This aspect mostly relates to psychological arguments of communication (see, for example, (Martire et al., 2014) and (Thompson et al., 2018)).

We report some of these issues, showing which ones are relevant to this context, and what is our approach to solving them in our work. Another major issue of Bayes factors is the well-known general difficulty of their computation outside "toy cases".

We believe that these issues are not detrimental to the usage of Bayes factors in forensic science. One cannot simply refuse the utilization of Bayes factors simply because they are difficult to compute (Nordgaard & Rasmusson, 2012, p. 309). Even a Bayes factor that has been computed using a simplified approach could be helpful to communicate the value of evidence in court, provided that the whole procedure has been guaranteed to be transparent and honest in its assumptions and limitations (Nordgaard & Rasmusson, 2012).

2.6.2 Criticisms from forensic scientists

Another common approximation to Equation (2.3) arises in forensic literature, known as "plug-in" (as it is used in the so-called "plug-in LR") (Cereda, 2017). Bernardo & Smith (1994) called it "estimated likelihood" (Bernardo & Smith, 1994, pp. 480, 483).

Firstly, one first supposes to have a background dataset of observations of Z. This dataset is then used to infer some knowledge about Y. In the context of the glass example (Example 2.1), one could collect information on glass objects of the same kind as the one that has been recovered, and analyze similarly obtained fragments.

Secondly, a maximum likelihood (ML) point estimate for Y replaces the full probability distribution of Y, and the integrand is evaluated using the point estimate. In other terms, one assumes that the density of Y is concentrated to a point mass located in the ML estimate, throwing away most of the information regarding Y's incertitude. In formulæ, Equation (2.3) can be approximated with Equation (2.12):

$$\begin{aligned}\Pr(Z = z \mid X = x) = \int \Pr(Z = z \mid Y = y) \Pr(Y = y \mid X = x) \, \mathrm{d}y \simeq \\ \simeq \Pr(Z = z \mid Y = \hat{y}^{\mathrm{ML}}) \, . \end{aligned} \tag{2.12}$$

This approximation is clearly very crude, and it has been shown to provide a misleading estimation[12] of the Bayes factor value in certain cases (Bernardo & Smith, 1994, p. 483; Cereda, 2017; Dawid, 2017).

Some forensic scientists suggest that Bayes factors computed in this way do not "consume" all uncertainty, but should still retain a dependence on the spread of Y, since it has been approximated with a point value. Therefore, these "Bayes factors" would awkwardly still be functions of Y (or some statistic on its distribution). Remember also that Bayes factors can be computed as the ratio between the posterior odds (where Y has been observed to assume the value of y, thus being fixed) and the prior odds (where Y plays no role). These are both scalar numbers, and they bear no dependence on Y, as it has been already observed, thus the proper Bayes factor does not depend on Y.

We strongly disagree with this view, as the Bayes factor value is a) a scalar number, b) computed by integration (or, equivalently, by dividing the posterior odds by the prior odds).

Other forensic scientists try to circumvent the necessity of eliciting a full probabilistic model for the evidence, relying instead on projecting the dissimilarities between

[12] We use the term "estimation" to mean the numerical procedure of obtaining an approximation to the true Bayes factor value that can be defined using Equation (2.2).

measurements on evidence items to an unidimensional space (the "score metric"), then eliciting a distribution on the obtained unidimensional distribution. The Bayes factor is then approximated by a ratio of probabilities of the score values. In literature this approach is called "score-based LR". It is often appealing for its apparent simplicity, as the definition of a distance measure is often straightforward compared to what is typically needed to compute a full Bayes factor. Also, the projection of complex data to an unidimensional space is a procedure that can be visualized, helping thus intuition and communication in the courtroom.

However, it has been proven that this approach is, at best, formally unjustified. Even in simplified scenarios, the score-based LR might be an unacceptable approximation to the full Bayes factor (Hepler et al., 2012). Also, it brings forward many issues that are not encountered with the usage of "proper" Bayes factors. For instance, the choice of the probability distributions for the score metrics is known as "calibration" (Robertson et al., 2016, sec. 7.2). These distributions characterized by parameters that are unknown, need to be estimated, but their connection to the "physical" world of the evidence space is lost. Hepler et al. (2012) and Neumann & Ausdemore (2019) give a good review of these issues. Notice also that the score-based LR must involve a comparison between two items, potentially coming from two different sources, whereas the full Bayes factor can be computed even in an investigative scenario.

2.6.3 Criticisms from statisticians

The Bayes factor is often stated in statistical literature as a means to choose, among two competing models, the one that better explains the observed data. For example, suppose that one wants to study the effect of a new medical treatment to affect the blood pressure. First, a dose X of medicament is administered, then one measures the increase (or decrease) of blood pressure Y between two reference time points. A control group of subjects does not receive any medicament, and their increase of blood pressure Y is taken in the same conditions. It might appear to be reasonable to investigate a pair of hypotheses such as $h_p : Y = 0$ ("the treatment has no effect") against $h_d : Y \neq 0$ ("the treatment has an effect"). Notice that the form of these hypotheses replicates those typically assumed under the NHST framework. However, this pair of hypotheses (where h_p is known in this case as a "sharp null hypothesis") cannot be appropriately investigated with the Bayes factor, as they suffer from the same issues as those under NHST. In fact, it is reasonable to assume that Y will *never* be exactly 0, but will depend on many unobserved confounding factors. Moreover, it was also suggested that a naïve Bayesian approach would introduce issues in the choice of the priors for the parameters (Gelman et al., 2009, p. 190 for a similarly

posed problem; also Kass, 1993).

Luckily, there are instances where the usage of Bayes factors is considered to be justified and meaningful. Gelman et al. (2009) gives an example, where the Bayes factor can be computed using the prior and the posterior odds, both hypotheses are scientifically sound, and there are "no obvious scientific models in between" (Gelman et al., 2009, p. 190). We believe that Bayesian models for forensic science, as we stated them, account for one of these cases, provided that the set of hypotheses is well-posed (i.e. hypotheses must be discrete, mutually exclusive and exhaustive). Notice that exhaustivity is not required if we accept that the Bayes factor is a *relative* measure of evidence (Robertson et al., 2016, p. 34).

Another issue which is frequently reported by statisticians is that Bayes factors are very sensitive to the priors (see for example (Yuan & Johnson, 2008)). This is in marked contrast with other statistical estimators, such as the maximum likelihood method, which converge to the *true* value as the sample size grows. However, in our opinion, one must always consider that Bayesian probabilities are subjective (Gelman et al., 2009). Different users might assume different values for the values of prior parameters, for example by considering different relevant populations. The reported sensitivity of Bayes factors to prior assumptions is thus a desirable trait that should not constitute an element of surprise. A more in-depth discussion can be found in (Morey et al., 2016, sec. 4.1).

2.6.4 Computation of the Bayes factor

Through marginalization

We have shown that the Bayes factor can often be written as a ratio of two integrals, such as the one appearing in Equation (2.3). If the Bayes factor is computed by separately computing the numerator and the denominator, we call this the "marginalization approach". When all random variables are discrete (or finite), the integration symbol in Equation (2.3) becomes a simple sum, thus all computations may be exhaustively carried out. With continuous or vector-valued random variables, the integration is often notoriously difficult or impossible to compute in a closed form.

Notice also that if the evidence Z consists of multiple observations that are conditionally i.i.d. given Y, the first integrand term in Equation (2.3) becomes a product of likelihood functions, that can rarely be analytically simplified. In the left hand side of the Equation, one finds the predictive distribution of a random vector. Even if one had a closed form for the predictive distribution of the scalar Z, it would not generally be possible to exploit it in the vector-valued case. Moreover, these likelihood values could easily result to be extremely small, moreso if the dimension of

the space increases, contributing to the potential propagation of numerical errors.

Luckily, the problem of computing the marginalization integrals is shared by many scientific fields. For instance, the marginalization integral is known as the *partition function* in statistical mechanics (Jaynes, 2003, pp. 281–282, 364). Shannon's information theory has deep connections with the partition function (Jaynes, 2003, p. 631). More recently, the Nobel Prize laureate Richard Feynman and others invented the so-called *path integrals* to compute the partition function in Quantum Electrodynamics (QED) (Zinn-Justin, 2002).

A number of methods have appeared in statistic and forensic literature to approximate the marginal likelihood value. We have already seen an example, the plug-in method (Equation (2.12)). An extensive comparison of major methods is available in (Bos, 2002). Most methods are either numerical (the integral is computed using numerical methods, such as quadrature), analytical (where an analytical approximation to the probability terms is sought) or stochastical (where integration is performed by random sampling). They differ also in how much information they require: some of them require that all elements appearing in the integrands must be in a closed form, others need only random samples from their distributions. However, no method is clearly superior, as some of them might be well suited for particular problems but fail to reach acceptable approximations in others. Moreover, the implementation difficulties greatly vary as well as the computation times.

The choice of an appropriate computational method strongly depends on the problem at hand. For instance, it has been shown that some of the most attractive methods (e.g. simple Monte Carlo sampling, and the harmonic mean estimator), extremely easy to derive and implement, often provide unacceptable approximations to the marginal likelihood value (Gamerman & Lopes, 2006; Lartillot & Philippe, 2006).

In this thesis we will use two methods to compute the marginal likelihood value:

- Gibbs sampling (Chib, 1995) for the Fourier loop shape descriptors,
- bridge sampling (Gronau et al., 2017; Meng & Wong, 1996) for the proportional descriptors.

The first method is operatively explained in (Bozza et al., 2008); bridge sampling is far too technical to be explained in this thesis, operative details can be found in (Gronau et al., 2017).

Other approaches

Notice that there is a class of methods that target directly the computation of a ratio of marginal likelihood values, that corresponds to the Bayes factor in non-trivial

cases. Some of these methods are closely related to those that attempt to compute a single marginal likelihood value, for instance bridge sampling (Meng & Wong, 1996).

Others are entirely novel, and they rather exploit other characteristics of the model. Lodewyckx et al. (2011) propose to compute the Bayes factor by introducing a prior on the model indicator (i.e. the hypothesis), creating a "supermodel" that encompasses both h_p and h_d, run a Markov Chain Monte Carlo, and count how many times each hypothesis is more supported by data (Lodewyckx et al., 2011). The method is very general but its implementation is delicate, as one needs to make sure that the chain explores both hypotheses, even if one is very unlikely. Also, it requires the awkward introduction of prior distributions (the so-called *pseudopriors*) on parameters that might not appear under one of the hypotheses under consideration.

Other approaches exploit the structure of the hypotheses under evaluation. If these hypotheses are nested (i.e. the model specified under h_d is contained into h_p), the Savage-Dickey ratio provides an easy way to compute the Bayes factor value (Wagenmakers et al., 2010). It is not without its own difficulties and paradoxes, however (Consonni & Veronese, 2008; Heck, 2019; Marin & Robert, 2010). Moreover, it has recently been shown that this method cannot be used to compute a Bayes factor in a nested model where one of the hypotheses has an exact equality constraint between two continuous random variables (Wetzels et al., 2010). Notice that this case contains all models presented in this thesis if H' (the statements relating θ_{r} and θ_{q} in Section 2.5) were considered as an hypothesis pair. In that case, the authors recommend the adoption of other techniques such as the marginalization approach, the one that has been followed.

2.6.5 Open problems

Besides these issues, it is important to remember that the Bayes factor is a *relative* measure for the value of evidence. Namely, it shows which one among the compared hypothesis better explains the observed data. It does not guarantee that the most supported explication is *true* (Morey et al., 2016). Bernardo & Smith (1994) introduced the notation $\mathcal{M}$-closed for the case where one of the compared models is assumed to be true. If this does not happen, the case is said to be $\mathcal{M}$-open (Bernardo & Smith, 1994, sec. 6.1).

The burden of determining whether the chosen hypothesis is plausible lies on the ability of the forensic scientist to choose a statistical model that is both defensible and close to the phenomenon that one wants to describe.

To this purpose, several solutions have been proposed to verify whether the models are close to the data.

The easiest method exploits the fact that the form of the marginal likelihood (Equation (2.3)) is the same as the form as the predictive distribution[13]. If one uses the posterior distribution for the latent parameter (Y in the Example), the marginal likelihood equation can be used to sample new observations from the same model that is assumed to be generating the data. Instead of checking whether the obtained parameters are reasonable (e.g. their credibility intervals for Y overlap scientifically sound values), a visual comparison can be made directly in the space of the original data. Moreover, sampling from the marginal likelihood is a much easier operation than computing its value in a given point. Notice that the same procedure can be repeated using the prior distribution for Y instead of the posterior. Notably, this allows practitioners to check whether their prior assumptions on the latent parameters generate plausible data, even before considering actual evidence into the case. This approach has been described in multiple texts under the name of *prior (posterior) predictive checking*, such as (Gelman, Meng, et al., 1996; Lynch & Western, 2004). In the forensic world, this procedure is known as "pre-assessment" (Taroni, Biedermann, et al., 2014, ch. 10).

Another possible extension relates to the fact that the choice of the probability distributions (for example those in Equation (2.3)) is fixed by the forensic scientists. One could in principle allow the form to vary among a specified set, for example to allow for uncertainty *on the model*. In forensic literature a similar procedure has been proposed for glass evidence, replacing the probability distributions with kernel density estimates (Aitken & Lucy, 2004).

The Bayesian way of choice to allow for the variability of probability distributions, is to put a prior on them, resulting in a so-called *Bayesian nonparametric* model. We hypothesize that this procedure can be applied to the numerator and denominator of the Bayes factor. It could also potentially lead to specific advantages. For instance, if the Bayesian nonparametric models encompass all *possible* distributions for a set of random variables, one could transition from an $\mathcal{M}$-open to an $\mathcal{M}$-closed case. Current literature on a nonparametric Bayes factor is scarce. Among available works we cite (Holmes et al., 2015), which could be of relevant forensic interest. To our knowledge, this approach has been applied only once in a forensic context (Cereda, 2015).

We also emphasize that the statistical procedure of model choice can be performed using approaches other than the Bayes factor. An example is the so-called *Bayesian model averaging* (BMA), where multiple models are combined to obtain a single

[13]The predictive density is the LHS of (Equation (2.3)), seen as a function of z (Equation (2.3)). The marginal likelihood value, or equivalently the denominator of the Bayes theorem, is the predictive density evaluated in the observed z: it is therefore a scalar value.

expression for a quantity of interest that appears in all considered models, accounting for their respective epistemological uncertainty (Hoeting et al., 1999). This does not apply to forensic purposes since we are solely interested in obtaining the Bayes factor value, as it is by definition the number that allows recipients of expert evidence to compute the posterior odds ratio. Note also that the Bayes factor cannot be considered among the quantities of interest for the BMA procedure, since it results from the comparison of models, rather than living inside them.

Chapter 3

Quantifying loops

This Chapter is dedicated to handwritten evidence that includes characters featuring closed loops, such as those in lower-case letters “a”, “d” and “o”.

We first introduce two datasets that may fall under this definition.

One has been the subject of several past forensic literature works, in particular (Marquis et al., 2005) and (Bozza et al., 2008). It is upon these works that this Chapter is built, in particular by suggesting the Bayesian model that we use to quantify the value of evidence. This dataset comprises measurements taken on a set of naturally written corpora.

The second dataset has been collected and introduced in a publication within the scope of this thesis (Gaborini et al., 2017). It consists of a set of genuine signatures of a single person, and a set of specimens created by several forgers, who had been exposed to the genuine set, and were allowed to practice imitation at will. All genuine and forged signatures contain character loops, that can be analyzed under the same statistical framework that has been employed to analyze the first dataset.

Secondly, we briefly recall the Fourier descriptors that we adopt to describe this kind of evidence. These descriptors are used by both datasets.

Thirdly, we detail the Bayesian model used in this Chapter. We will refer to the original article (Bozza et al., 2008), relating it to the more general inferential scheme introduced in Chapter 2 (Section 2.5.2).

Afterward, we show our optimized implementation of the Bayesian model, enabling us to discuss its sensitivity, strengths and weaknesses in greater detail than the original article. We will do so also by leveraging several simulated datasets, following the procedure described in Section 2.5.5.

Lastly, we will exploit our implementation to discuss results in the form of Bayes factors for each of the two datasets.

3.1 The datasets

3.1.1 Natural handwriting dataset

This dataset was collected among a sample of 42 French native speakers from the School of Criminal Justice, University of Lausanne, Switzerland. The dataset comprises 6868 specimens of the character loops "a" and "d", with a median of 177.5 loops per writer. All specimens were written in the writers' natural handwriting. The shape of each loop was described by their Fourier descriptors as presented in (Bozza et al., 2008; Marquis et al., 2005). 5 harmonics were retained (from order 0 to 4) as well as the loop surface: details on the meaning of Fourier coefficients are given in Section 3.2. Note that no image of the specimens nor the specimens themselves were available. The numerical values related to the 5 harmonics plus the surfaces have been used as in (Marquis, 2007).

3.1.2 Forged signatures dataset

A subset of this new dataset was described in (Gaborini et al., 2017). However, note that the article mostly considered a different set of features (absolute signature dimensions), and character loops were only briefly described.

This dataset was built starting from 143 specimens of genuine signatures of a single person, collected over the period of one month, writing in small batches to avoid nuisance factors such as fatigue. 6 persons were recruited in the project, with no experience in signature forgery and document examination. Each person was provided the full set of genuine signatures and was told to produce at least 20 freehand forgeries over a week period, practicing the forgeries at will. Tracing was forbidden. All specimens were produced on white unruled paper, writing with a ballpoint pen. In total, 444 specimens were collected.

All genuine signatures could present at most 5 character loops, corresponding to characters "a", "b", "o" in different positions. The same loops were extracted in the forged specimens, if present.

A partial reproduction of a signature is shown in Figure 3.1.

Digitalization and loop extraction

All specimens were digitalized at 600 dpi, then converted to a black and white uncompressed format. Every loop is extracted from the digitalized images of samples through various image processing steps, that isolate the *skeleton* of each character loop from the ink trace. The procedure is rather manual, as some of these steps

Figure 3.1: A partial reproduction of a signature in the dataset. Two loops ("o"s) are visible.

require minor image retouches (e.g. noise removal, or loop closure) and calibration of parameters (e.g. choosing the appropriate morphological filters, or binary thresholding). The main reasoning we adopted is that loops were manually closed if this operation did not significantly alter their shape (i.e. very small opening, or the closing is strongly suggested by the surrounding character traits) (Gaborini et al., 2017). Given a single character loop, the output of this step is a list of planar coordinates to each pixel, as well as additional information that links the loop to the original specimen (e.g. its relative position in the signature). The planar coordinates were afterward converted to polar coordinates, using the barycenter of the pixel sets as the center of the representation, and discretizing the angle over 128 equispaced nodes.

In total, the dataset comprises 1336 loops, whose 420 appear in genuine specimens. Table 3.1 summarizes the composition of the signature dataset.

The superposition of all extracted character loops is shown in Figure 3.2, distinguishing by writer and letter. It can be visually appreciated that some forgers never reproduced (or failed to close) some character loops, present in most genuine specimens. Also, the shape of genuine loops does not significantly vary across letters, except for the "b" loops, appearing slightly more elongated than "o"s and "a"s.

Table 3.1: Number of specimens and loops per writer in the forged signature dataset. "A" indicates the genuine author, "F1" to "F6" indicate the forgers.

Writer	Total specimens	Total loops
A	143	420
F1	35	98
F2	20	49
F3	16	23
F4	20	56
F5	153	538
F6	57	152

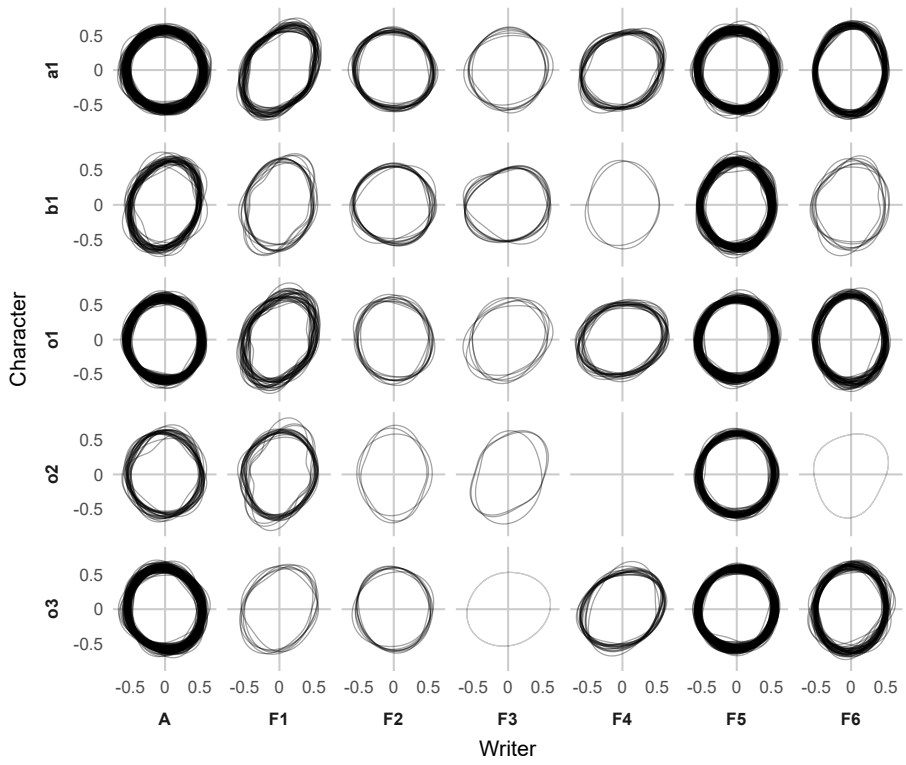


Figure 3.2: All characters in the forged signature dataset, superposed, distinguishing by writer (columns) and letter (rows). All character loops have been rescaled to have the same surface. Only closed loops were shown.

3.2 Features

The method introduced in this Chapter is based on the representation of closed shapes (e.g. character loops) as a sum of harmonic components, through the Fourier series on a polar parametrization of the shape contour. The main idea was introduced in this forensic context in several articles (Marquis et al., 2005, 2006; Marquis et al., 2011), although the extraction of said features was originally described for anthropological applications in (Schmittbuhl et al., 1998). We briefly recall the construction of the Fourier features, referring the reader to the cited articles to obtain further details.

Let us consider a single character loop. We initially assume to have a list of $x - y$ coordinates to the pixels that belong to the loop, possibly those close to the center of the stroke (*skeleton*). Following (Marquis et al., 2005), each loop was converted in polar coordinates $(\theta, R(\theta))$ with respect to its barycenter, then the Fourier descriptors are extracted according to the truncated Fourier series in Equation (3.1). Figure 3.3 shows the representation of a loop in polar coordinates.

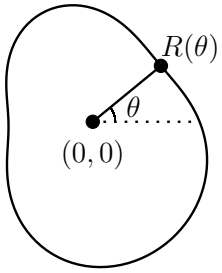


Figure 3.3: A contour in polar coordinates.

$$R(\theta) = A_0 + \sum_{k=1}^{n} A_k \cos(k\theta + \theta_k) \qquad \text{for } \theta \in [0, 2\pi)\,. \tag{3.1}$$

Each harmonic contribution is characterized by the harmonic index k (an integer), an amplitude A_k (a non-negative real number) and a phase θ_k (typically in radians or degrees).
The meaning of all harmonics contributions can be represented in Figure 3.4, where each harmonic contribution of amplitude $A_k = 0.5$ is added, in turn, to the unit circle. The reference article considered as features the sets $\{A_k\}$ and $\{\theta_k\}$ for $k > 0$, and the loop surface.

It is also known (Zahn & Roskies, 1972) that the series in Equation (3.1) can be rewritten as Equation (3.2):

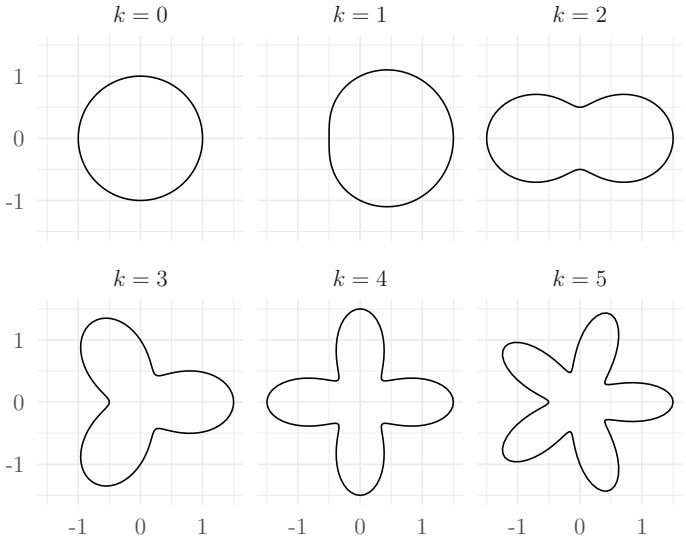


Figure 3.4: The unit circle ($k = 0$) along with some shapes obtained by summing a Fourier contribution of amplitude $A_k = 0.5$ and phase $\theta = 0$ to the unit circle.

$$R(\theta) = a_0 + \sum_{k=1}^{n} \left(a_k \cos(k\theta) + b_k \sin(k\theta)\right) \qquad \text{for } \theta \in [0, 2\pi)\,. \tag{3.2}$$

Each harmonic contribution can also be described by the harmonic index k, and the terms a_k and b_k (real numbers). It can be proven that $a_k = A_k \cos\theta_k$ and $b_k = A_k \sin\theta_k$ for $k > 0$.

In this Chapter, the features are constituted by the sets $\{a_k\}$ and $\{b_k\}$ for $k > 0$, and a_0.

The number of retained harmonics n determines the degree of approximation of the Fourier representation. In both datasets as well as the reference article n did never exceed the value of 4, as it achieves a good balance between the graphical approximation of the loops, and the reduced number of parameters to fit.

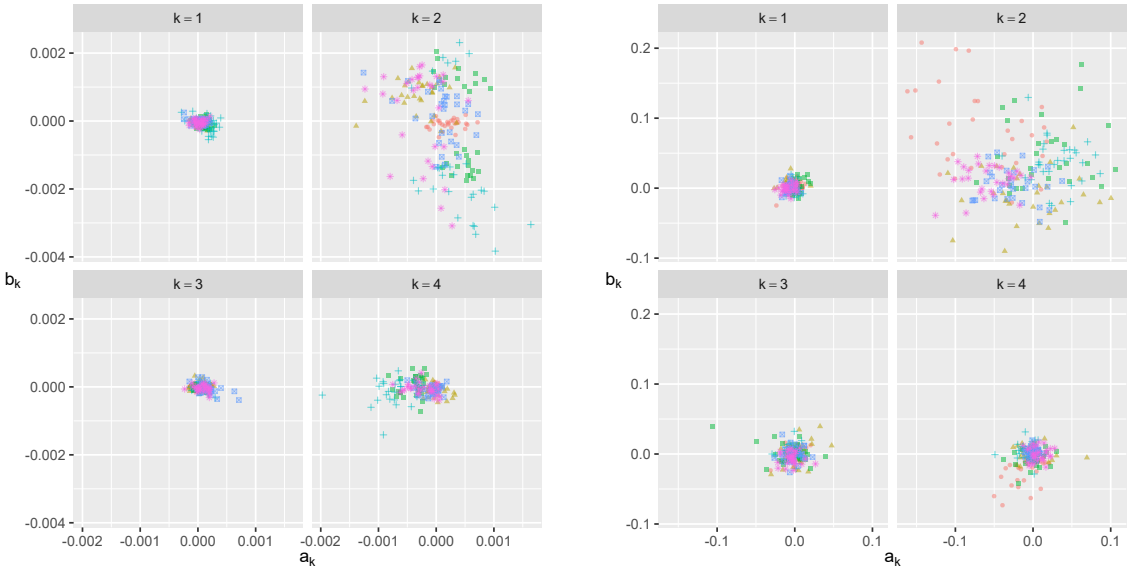
Notice that n is fixed, as it has been decided during the digitalization procedure. However, the number of Fourier coefficients that take part in the statistical model may vary, depending for instance on their discrimination power. As a consequence, this choice has been also briefly investigated in the next sections.

In both datasets, it can be shown that authors tend to differentiate on components $k = 2$, which determine the "elongation" of shapes. This behavior was already observed in the reference articles, and we observe the same phenomenon in the forged

signature dataset. Moreover, the amplitudes of harmonic components $k = 2$ are much greater than other harmonic orders (except $k = 0$, which mostly determines the average radius of the loop). A comparison of the harmonic structure across the two datasets is shown in Figure 3.5.

3.2.1 Parameters and choices

This procedure requires the establishment of a number of choices for various parameters. The first choice is the number of Fourier components n used to represent character loops (more generally, which k were considered to form the feature vectors). In general, it is advisable to choose an n that guarantees a reconstruction that is sufficiently close to the actual loops, whilst yielding a number of parameters to fit ($2n + 1$ coefficients: $2n$ for the Fourier coefficients, 1 for a_0) that is compatible with the available number of observations. In the forged signature dataset we considered $n = 4$, as 4 harmonics were sufficient to characterize loop shapes, thus describing each loop with a vector of maximal length 9.



(a) Natural writing dataset

(b) Forged signature dataset

Figure 3.5: Distribution of pairs of harmonic coefficients of order k. Each shape and color is associated with a writer. Each point represents a loop. Only a subset of loops and writers is shown. Notice how writers tend to separate on $k = 2$.

A second choice that needs to be made is whether loops should be rescaled in order to have a unitary area. In the reference articles, the authors used the rescaled versions. In the forged signature dataset, we briefly investigate this assumption through the developed Bayesian model. If Bayes factors are not sensitive to the loop surface, the choice of whether the rescaling is needed is not relevant. From the forensic handwriting examination perspective, the usage of rescaled loops attempts to reduce the influence of external constraints that are known to affect absolute size, such as the presence of boxes or a horizontal line (Ellen, 2005; Fazio, 2015).

3.3 Statistical model

The statistical model developed in (Bozza et al., 2008) was adopted to treat the new set of data. We recall it briefly here.

3.3.1 Background observations

We initially suppose to have a background dataset on m writers. n_i character loops are available coming from the i-th writer, $i = 1, \ldots, m$. Each loop is characterized by p features, the number of Fourier coefficients $\{a_k, b_k\}$ (possibly including a_0) forming the feature vectors.

The set of features for the i-th writer and the j-th loop coming from the i-th writer is a random vector, indicated with $X_{ij} \in \mathbb{R}^p$, where $j = 1, \ldots, n_i$.
A hierarchical model is stated for X_{ij}:

$$\begin{aligned} X_{ij} &\sim \mathcal{N}_p(\theta_i, W_i) \\ \theta_i &\sim \mathcal{N}_p(\mu, B) \\ W_i &\sim IW(U, \nu)\,, \end{aligned} \tag{3.3}$$

where θ_i is the mean vector of the i-th writer, W_i is the non-constant within-writer covariance matrix, μ is the mean vector between writers and B is the between-writers covariance matrix.

To reconnect to the evaluative inferential scheme described in Section 2.5, evidence e from the i-th writer and j-th loop is represented by the random vector X_{ij}. f is the density of a multivariate Normal, parametrized by (θ_i, W_i) in the i-th writer (previously indicated with θ). The between-writer hierarchical level models the distribution of the within-writer parameters (θ_i, W_i). In this case, θ_i and W_i are

modeled separately, with a multivariate Normal and an inverse Wishart distribution. Consequently, g is the density of the product of those two random variables.

Some statistical differences were introduced. As said in Section 3.2, we note that we describe Fourier contributions using the coefficients $\{a_k, b_k\}$ rather than $\{A_k, \theta_k\}$[1]. This no longer imposes sign and domain constraints on the variables, and they both share the same order of magnitude for any harmonic index k. In this way, the hypothesis of multivariate normality is supported.

Background parameter elicitation

We now apply the plug-in approximation (Section 2.5.3) to elicit values for μ and B. The hyperparameters μ, B and W_i are estimated using maximum likelihood on the background dataset, as in (Bozza et al., 2008, sec. 4). In particular, we obtain an estimate of W, the within-writer covariance matrix pooled over all W_i. U and ν are elicited using the mean of the inverse Wishart distribution (Press, 2012):

$$\mathbb{E}[W_i] = \frac{U}{\nu - 2(p+1)}\,. \tag{3.4}$$

We initially set the number of degrees of freedom ν to be as low as possible, in order to let W_i have a finite expectation in Equation (3.4).
The smallest possible value for ν is indicated by $\nu_{\min}$:

$$\nu_{\min} := \nu = 2(p+1) + 1\,. \tag{3.5}$$

The inverse Wishart scale matrix U is elicited by substituting the pooled maximum likelihood estimate for $\mathbb{E}[W_i]$ in Equation (3.4), the chosen value for ν, and solving for U.

3.3.2 Evaluative scenario

We now introduce the evaluative scenario and the evaluative hypotheses that we wish to use for evaluating evidence, following the structure introduced in Sections 2.4.2 and 2.5.

We suppose that the questioned material shows multiple character loops, indicated with e_{q}. The reference material, provided by the putative writer, shows multiple character loops, indicated with e_{r}. All loops are described using the aforementioned Fourier descriptors, keeping the same fixed number of values.

[1] There is no relation between θ_k and any θ_i in Equation (3.3).

The evaluative hypotheses of interest are:

$H = h_p$: "the character loops e_r and e_q come from the same writer"
$H = h_d$: "the character loops e_r and e_q come from two different writers"

Under h_p, e_r and e_q are samples from the same source, with parameters θ_rq and W_rq. Under h_d, e_r and e_q are two independent random vectors: the source for e_r is parametrized by the parameters θ_r and W_r, the source for e_q is parametrized by the parameters θ_q and W_q.

The Bayes factor value can be computed using Equation (2.10). In particular, it requires the computation of three marginal likelihood values that share the same structure (Equation (2.11)). The generic marginal likelihood is:

$$m(e \mid h) = \int f(e; \theta, W)\, \pi_1(\theta; \widehat{\mu}, \widehat{B}) \pi_2(W; \widehat{U}, \widehat{\nu})\, \mathrm{d}\theta\, \mathrm{d}W\,,$$

where $\widehat{\mu}$, $\widehat{B}$, $\widehat{U}$ and $\widehat{\nu}$ are the plug-in estimates on the background dataset, and π_1 and π_2 are the densities for the prior parameters θ and W, respectively.

3.3.3 Bayes factor computation

The Bayes factor value is obtained by computing the three marginal likelihoods, separately. The procedure follows the one described in (Bozza et al., 2008), and is done in two steps.

First, we use the fact that $m(e \mid h)$ is the normalizing constant of the posterior for (θ, W), indicated with π.

$$m(e \mid h) = \frac{f(e; \theta, W)\, \pi_1(\theta; \widehat{\mu}, \widehat{B}) \pi_2(W; \widehat{U}, \widehat{\nu})}{\pi(\theta, W; e, \widehat{\mu}, \widehat{B}, \widehat{U}, \widehat{\nu}, h)}\,. \tag{3.6}$$

Therefore, the marginal likelihood can be estimated from a set of samples for the posterior of (θ, W) (Chib, 1995). Equation (3.6) is used to deliver an estimate of $m(e \mid h)$ in a given point of (θ, W), indicated with (θ^*, W^*). The Equation holds for any choice of (θ^*, W^*), but it is desirable to choose a point where many samples are available (Chib, 1995, sec. 2.1). For (θ^*, W^*), we chose the point where the likelihood f assumes the maximum value.

In the second step, we sample from the posterior distribution of (θ, W). The posterior is not available in closed form, but the full conditional distributions $\Pr(\theta \mid W, e)$ and $\Pr(W \mid \theta, e)$ are known, allowing for the usage of a Gibbs sampler. More details are available in (Bozza et al., 2008) and (Aitken et al., 2021).

3.3.4 Implementation

This Chapter was first approached by using the code that was developed for the reference article. The original implementation was written using the R language (R Core Team, 2019). Given the need for a detailed description of the sensitivity of this method to new datasets and prior parameters, notwithstanding the required heavy computational loads, it was necessary to improve the user-facing components and the general user experience, particularly the speed of computation of each marginal likelihood value.

A new code implementation has been afterward packaged to the R package `bayessource` (Gaborini, 2019), providing a set of functions to obtain the background plug-in estimates, marginal likelihoods and the Bayes factor values for the proposed model.

The core functions were translated from pure R routines to C++, a high-performance language that directly generates machine code. This was done using the Rcpp R package, a tool that enables the creation of a transparent interface between the two languages (Eddelbuettel & François, 2011).

Moving core computational procedures to a C++-based implementation allowed for extremely large speedups for the computation of Bayes factors. In Table 3.2 we compare the computation times for a single Bayes factor across three different implementations of the model:

a. the reference pure R implementation,
b. an improved pure R implementation, exploiting vectorization and matrix factorizations,
c. our R/C++ package.

One can see that the mean time for computation of a single Bayes factor value reduces from around 3031.924 milliseconds to 13.47519 milliseconds, with an average 225× speedup. Even faster speedups are achievable by parallelizing the algorithm and exploiting various matrix factorizations for the covariance and scale matrices.

Further benefits for wrapping the article code to an R package include the increased reproducibility of the analyses, the facilitated creation of package documentation, and the establishment of an automated suite of tests. These tests make sure that most components of the package behave correctly in scenarios where results are known. For example, we check that the implemented definitions for the Wishart densities and samplers reduce to known distributions for known particular parametrizations, and that all covariance matrices can be provided either in their full form or through their Cholesky factors. We also verify that the package computes Bayes factor values that

Table 3.2: Comparison of Bayes factor computing times (in milliseconds) across the three implementations. All statistics were computed over 10 replicas. 100 burn-in iterations, 1000 sampling iterations, $p = 2$ variables, $k_\mathrm{r} = k_\mathrm{q} = 30$ samples, 510 background samples.

Implementation	min [ms]	mean [ms]	median [ms]	max [ms]
Reference (pure R)	2719.91	3031.92	2859.97	3882.45
Improved (pure R)	1274.06	1369.07	1347.24	1515.70
Ours (Rcpp)	12.67	13.48	13.18	14.96

support the true hypothesis in a situation where the generating model is fully known. The suite of tests is supported and ran by the `testthat` R package (Wickham, 2011).

The package has not been released to CRAN (the official R package repository), but its source is open and available on request (Gaborini, 2019). In the repository, one also finds abundant documentation on how to use it, including some worked-out simplified scenarios.

3.4 Model validation

With the available package, we first proceeded to verify its behavior in a situation where the generating model is known, and a large set of background samples are available (a background-dominant situation: see Section 2.5.5).

The background data consists of 200 samples from $m = 3$ bivariate ($p = 2$) Gaussians, representing three different sources, $i \in \{1, 2, 3\}$. In total, 600 samples are available. The model parameters $\theta_i, W_i, \mu, B, U, \nu$ were chosen to produce slightly overlapping distributions, shown in Figure 3.6.

We consider $i = 1$ to mark the reference source. We want to evaluate the hypothesis pair:

$H = h_p$: "the samples e_r and e_q come from the source 1",
$H = h_d$: "the samples e_r and e_q come from sources 1 and 2, respectively".

The scenario simulation procedure (Section 2.5.5) is applied twice to simulate two situations:

① under h_p, we deal $k_\mathrm{r} = 30$ samples from the reference source ($i = 1$) and $k_\mathrm{q} = 30$ samples from the reference source ($i = 1$)

② under h_d, we deal $k_\mathrm{r} = 30$ samples from the reference source ($i = 1$) and $k_\mathrm{q} = 30$ samples from the questioned source (we pick $i = 2$)

The remaining 510 samples constitute the background dataset, which is used to elicit the hyperparameters using the plug-in estimation procedure.

The generated dataset is shown in Figure 3.6. The point shapes mark in which sample set they belong.

3.4.1 Priors and computational parameters

Using the plug-in estimation procedure, one can obtain the point estimates for the between- writer parameters $\widehat{\mu}, \widehat{B}, \widehat{U}, \widehat{\nu}$ using the background observations e_{b}, as detailed in Section 3.3.1. Particularly, $\widehat{\nu}$ was set to $\nu_{\min} = 7$ (Equation (3.5)).

For the Gibbs sampling procedure, 10000 Gibbs samples are obtained, whose 1000 are dedicated to the burn-in process. The number of samples was decided to provide a good mixing of the Gibbs chain. Only one Gibbs chain is run. Trace plots did not show any anomalies. Also, the reference article used only 1000 Gibbs samples in total.

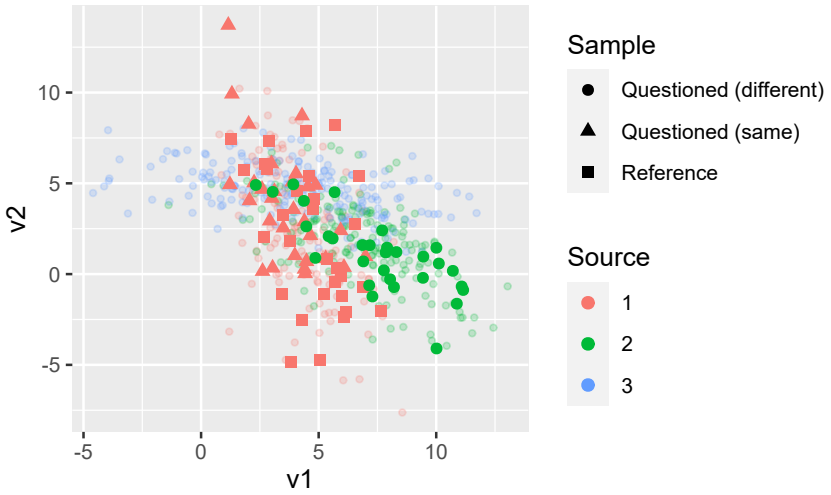


Figure 3.6: A generated bivariate dataset for model verification. Each point represents a bivariate observation. The small circles constitute the background samples. Colors represent the true source. Shapes indicate whether the sample belongs to the reference or the questioned set.

3.4.2 Bayes factors

By plugging in our point estimates for the prior hyperparameters in Equations (2.11) and (3.6), one can compute the Bayes factor values under the scenarios ① and ② of Section 3.4.

In the situation ① (h_p is true), e_r and e_q have been sampled from the same source. The log-Bayes factor is:

$$\log_{10} \mathrm{BF} = 1.1989623\,,$$

which is greater than 0, as expected.

In the situation ② (h_d is true), e_r and e_q have been sampled from two different writers. The log-Bayes factor is:

$$\log_{10} \mathrm{BF} = -9.0001065\,,$$

which is lower than 0, as expected.

As further verification, we compare the results of the Bayes factors across the three implementations (Section 3.3.4), and an informed prior selection (where we chose the true generating parameters as priors, being known). The comparison, shown in Table 3.3, proves that our implementation produces Bayes factor values that are coherent with the established hypotheses and the generated data. Moreover, we also show that the maximum likelihood (ML) estimators constitute a good choice in a background-dominant situation.

Table 3.3: Comparison of the value of evidence quantified through the $\log_{10}$ Bayes factor computed on the same dataset using different implementations and prior elicitation methods.

Implementation	$\log_{10}\mathrm{BF}$ (h_p)	$\log_{10}\mathrm{BF}$ (h_d)	Prior choice
Reference (pure R)	1.1719	-9.0332	ML
Improved (pure R)	1.1923	-9.0048	ML
Ours (Rcpp)	1.1990	-9.0001	ML
Reference (pure R)	0.9902	-9.1955	Informed
Improved (pure R)	1.0265	-9.1439	Informed
Ours (Rcpp)	1.0256	-9.1634	Informed

3.4.3 Diagnostics and sensitivity

The major computational speedups enable us to evaluate the sensitivity of the obtained Bayes factor values to various assumptions, such as the prior parameters and the dataset properties.

Notice that Bayes factors are difficult to "debug", as they are rooted in a combination of three quantities (the three marginal likelihoods) that may wildly vary across orders of magnitude. In turn, each quantity depends on the plug-in estimates for the hyperparameters, and how the respective posteriors for the latent source parameters relate to the chosen hyperparameters.

In a simulated scenario, one could compare how close they are to the generating values. For example, when data is generated under h_p from the source $i = 1$, the posteriors for θ_{r} and θ_{q}, calculated using e_{r} and e_{q} respectively, should be concentrated around θ_1. Analogously, the posterior for θ_{rq}, calculated using all the available e_{r} and e_{q}, should also be concentrated around θ_1.

In some situations, however, this characterization is not possible. For example, when data is generated under h_d, θ_{rq} tries to capture the mean vector of *the* source that is generating the reference and questioned samples. However, we know that this source does not exist, as e_{r} and e_{q} truly come from two different sources. The computed marginal likelihood depends, therefore, on how distant these sources are in terms of their respective mean vectors θ.

A more exhaustive sensitivity study should take into account the interplay of all these aspects. For instance, one could develop inequalities to bound the Bayes factor, as done in (de Zoete & Sjerps, 2018), or to bound the individual marginal likelihoods using a Variational Bayes approach (Blei et al., 2017). In this Section we provide only some elements of sensitivity analysis, to gain insight into the basic properties of the developed method.

3.4.4 Convergence

Since the model parameters are known, one can check whether the Gibbs chains are sampling near the correct values. Notice that since the computation of a Bayes factor requires three marginal likelihoods, there will be three sets of Gibbs samples. To simplify, we consider only the chain for the reference samples that appear in the denominator of the Bayes factor, i.e. $m(e_{\mathrm{r}} \mid h_d)$. Figure 3.7 displays the comparison between the posterior distributions of θ and W^{-1} and the true values for the reference source (respectively θ_{r} and ${W_{\mathrm{r}}}^{-1}$), showing a good agreement between the MCMC behavior and the model generating values.

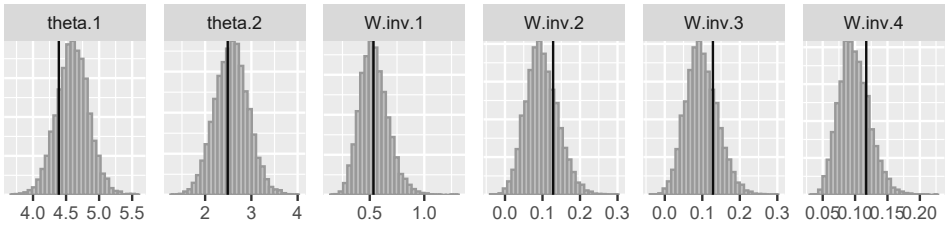


Figure 3.7: Histogram of the obtained posterior samples for θ_{r} and ${W_{\mathrm{r}}}^{-1}$, along with their true values (vertical bars). Each histogram represents the sampling distribution of the corresponding entry.

Degrees of freedom

One open question that was highlighted in the reference article was the sensitivity to ν, the number of degrees of freedom of the inverse Wishart distribution for the between-writer variability.

The sensitivity to ν was investigated by choosing values of $\widehat{\nu}$ in the range $[\nu_{\min}, \nu_{\max}]$. $\nu_{\min}$ was set to be 7 (the smallest value such that the scale matrix is invertible, see Section 3.3.1 and Equation (3.5)), and $\nu_{\max}$ was taken to be large.

Following the previous section, the Bayes factor values were recalculated, obtaining Figure 3.8. Results show that Bayes factor values tend to become more extreme (i.e. $|\log \mathrm{BF}|$ increases) as ν is taken to assume larger values. This behavior is expected, since the prior distribution for the within-writer covariance matrix W concentrates in a smaller region of the $\mathbb{R}^p$ space, attributing greater weight to the (pooled) maximum likelihood estimates.

One can see that, concerning h_d, the performance reverses as ν increases, resulting in log-Bayes factor values greater than 0 from a certain value of ν. These two considerations suggest choosing small values for ν (e.g. $\nu_{\min}$), as they are more conservative (under h_p) and avoid the production of contradicting conclusions (under h_d).

Sensitivity to the dataset

One important question to answer is the sensitivity of the Bayes factor to the collected observations. As an example, one might be interested in examining at which point e_{r} and e_{q} become "different enough" to obtain a Bayes factor that crosses the neutral value of 1. Another question of interest would be to assess the dependence of the Bayes factor (thus, the decision whether to prosecute or not) to uncertainties in laboratory analyses. For instance, suppose that in the glass fragment example (Example 2.1),

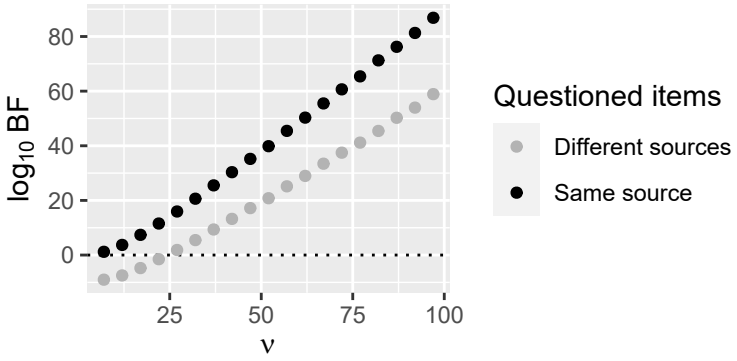


Figure 3.8: Sensitivity of the Bayes factor to ν. The black series is obtained when h_p is true. The gray series is obtained when h_d is true.

the laboratory reports an interval on the refractive index of the fragment instead of a point value (representing a confidence interval, a credibility interval, or any other way to indicate values that are highly plausible according to the laboratory). If the Bayes factor varies significantly by considering observations inside the given interval, then the strength of the evidence e_{q} may be weaker than expected, the model might be wrong or incomplete[2], or more observations are needed. Another output of the sensitivity analysis to the dataset is the rate of change of the Bayes factor to small perturbations of the casework evidence e_{r} and e_{q}.

To continue the analysis, it would be necessary to precisely state how observations (for instance, e_{q}) could vary, and what is the magnitude of the variation. If evidence is described by a number (e.g. the refractive index or a fiber count), it could be sufficient to let it increase or decrease. However, for multivariate observations there are infinitely many possible directions of increase and decrease.
To avoid introducing unnecessary complexity, this issue can be generically solved by defining a notion of distance between probability distributions, in particular the ones generating the sets e_{r} and e_{q}: the first distribution will be indicated with p, the second with q. Observations can be transformed by modifying the generating distribution, for instance by translating the mean vector or rotating the covariance matrix. One suitable notion is the Bhattacharyya distance (Kailath, 1967):

[2]In this specific case, a properly specified model would also include the uncertainty on the reported e_{q} among the random variables.

Definition 3.1 (Bhattacharyya distance)
Let p, q be two continuous probability distributions on Ω, with densities $p(x)$ and $q(x)$. The Bhattacharyya distance between distributions p and q is:

$$d(p,q) = -\log \int_{x\in\Omega} \sqrt{p(x)\,q(x)}\,dx\,.$$

In our case, e_r has distribution $p = N(\theta_\mathrm{r}, W_\mathrm{r})$, e_q has distribution $q = N(\theta_\mathrm{q}, W_\mathrm{q})$. The reference and questioned sources are known, so the distance can be exactly computed. Substituting the definitions, one obtains:

$$d(p,q) = \frac{1}{8}(\theta_\mathrm{r}-\theta_\mathrm{q})^T W^{-1}(\theta_\mathrm{r}-\theta_\mathrm{q}) + \frac{1}{2}\log\left(\frac{\det W}{\sqrt{\det W_\mathrm{r}\,\det W_\mathrm{q}}}\right),$$

with $W = (W_\mathrm{r}+W_\mathrm{q})/2$. Notice that one obtains the well-known Mahalanobis distance if the distributions have the same covariance matrices.

Next, one can apply a transformation to the distributions that generate e_r and e_q. For e_q, we consider the case where h_d is true: the sources are different.

To simplify, e_r is kept fixed, setting instead the questioned mean vector θ_q to a new position $\tilde{\theta}_\mathrm{q}$, leaving the covariance matrices unchanged. Moreover, a linear shift is applied, parametrized by the coordinate $\alpha \in \mathbb{R}$, such that one interpolates between θ_r and θ_q. One can thus write:

$$\tilde{\theta}_\mathrm{q}(\alpha) = \theta_\mathrm{q} + (\alpha-1)(\theta_\mathrm{q}-\theta_\mathrm{r})\,.$$

By applying the same shift to the questioned samples e_q, one obtains a shifted version $\tilde{e}_\mathrm{q} = \tilde{e}_\mathrm{q}(\alpha)$. The effect of the geometrical transformation of e_r and e_q is shown in Figure 3.9.

Finally, one can compute the Bayes factor using e_r and $\tilde{e}_\mathrm{q}$. Notice that the Bayes factor is now a function of α as well as the Bhattacharyya distance $d(e_\mathrm{r}, \tilde{e}_\mathrm{q})$. The behavior of the Bayes factor is shown in Figure 3.10. One can see, for instance, that the log-Bayes factor in this case is always lower than 0: the questioned samples e_q are always sufficiently "different" than e_r to support h_d rather than h_p. Notice this is true also when $\alpha = 0$ (i.e. when their means overlap: see the first panel in Figure 3.9), since the Bhattacharyya distance also considers the covariances of the distributions.

The first-order approximation to the curve could be used as a measure of sensitivity of the Bayes factor to small perturbations around the observed casework data e_q.

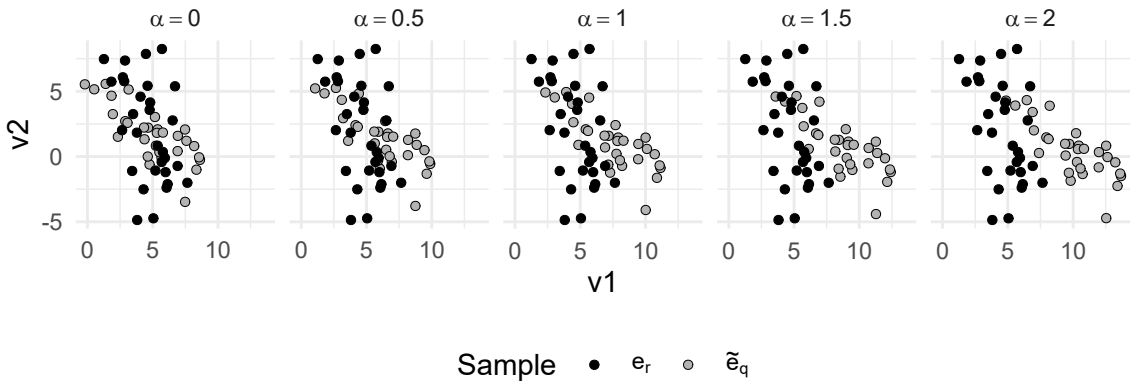


Figure 3.9: The datasets obtained by transforming the bivariate validation dataset (Figure 3.6) at various values of α. In particular, the mean of the questioned samples θ_q from the second source was linearly shifted to the value $\tilde{\theta}_q(\alpha)$. Choosing $\alpha = 0$ sets $\tilde{\theta}_q = \theta_r$. Choosing $\alpha = 1$ sets $\tilde{\theta}_q = \theta_q$. Choosing $\alpha > 1$ sets the reference and the questioned samples farther apart than their original positions, inflating their distance.

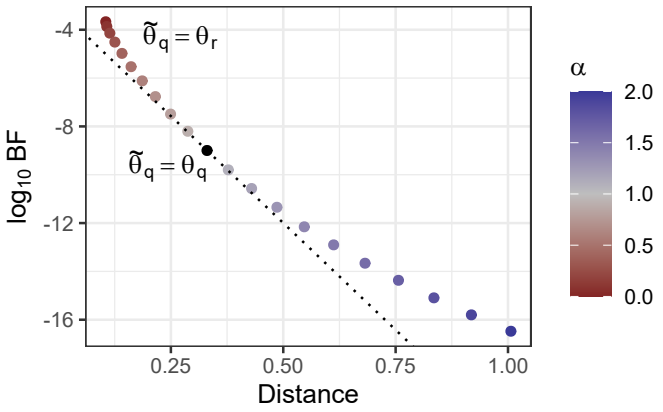


Figure 3.10: The log-Bayes factor as a function of the Bhattacharyya distance and the shift α. The dotted line is the tangent to the observed casework data. Notice that the Bayes factor always points to h_d even when the mean vectors overlap ($\alpha = 0$): this should be due to the different covariances.

Gibbs iterations

Another analysis that can be performed is the sensitivity of the Bayes factor to the number of Gibbs samples, and how many of those are discarded. In Figure 3.11 we recompute the Bayes factor by exploring various combinations of the number of burn-in samples and the number of Gibbs samples after the burn-in (from 100 up to 50000). One can see that convergence (in terms of BF value) is easily reached even with few Gibbs samples.

3.5 Results on natural handwriting

We can now proceed to analyze the same natural handwriting dataset appearing in the reference article (Bozza et al., 2008). Again, some differences concerning the form of the extracted features are introduced, so the results are not directly comparable. Nevertheless, we exploit the decreased computational load of the algorithm to provide new insights on the same dataset.

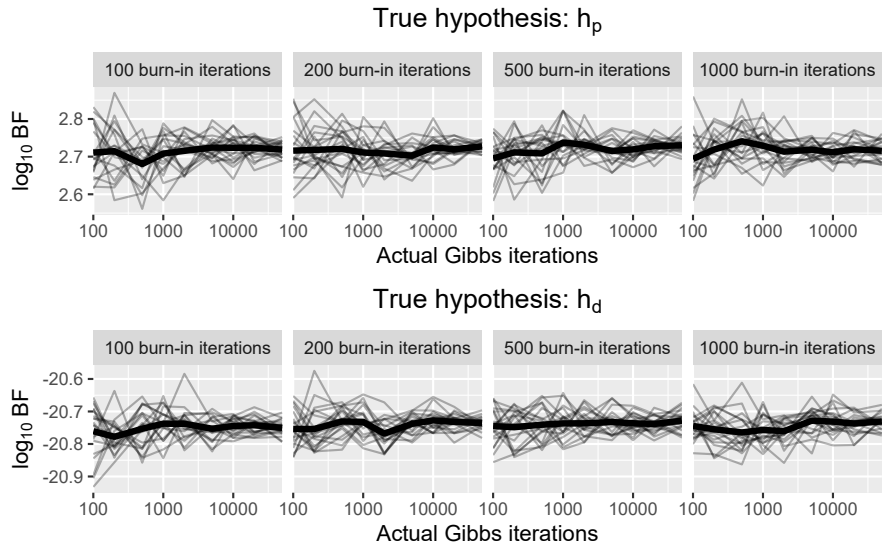


Figure 3.11: log-Bayes factor values as a function of the number of Gibbs samples, burn-in iterations, and true hypothesis. Each parameter combination was repeated 20 times (the thin lines). The thick lines show the BF averaged over 20 trials. One can see that the BF values are stable over a wide range of values for the MCMC parameters.

3.5.1 Evaluative scenario

Let e_{r} be a set of character loops, parametrized by their Fourier coefficients, from the reference writer. Let e_{q} be a set of character loops, parametrized by their Fourier coefficients, from the questioned writer (i.e. the writer of the questioned material, who may or may not correspond to the reference writer).

The evaluative scenario described in Section 3.3.2 is adopted, in particular with the same hypothesis pair:

$$\begin{aligned} H = h_p &: \quad \text{“the loops } e_{\mathrm{r}} \text{ and } e_{\mathrm{q}} \text{ come from the same writer”} \\ H = h_d &: \quad \text{“the loops } e_{\mathrm{r}} \text{ and } e_{\mathrm{q}} \text{ come from two different writers”} \end{aligned}$$

As the natural handwriting dataset constitutes a background-dominant situation, one can proceed with the scenario simulation procedure that was described in Section 2.5.5. In particular, every possible pairing between a reference and a questioned writer is explored. Once the pair is chosen, the procedure is repeated $G = 100$ times, to account for the writer's intra-variability. Each time new e_{r}, e_{q} and e_{b} are sampled, re-eliciting the priors with the new background dataset. In the end, since the dataset contains 42 different writers, $100(\binom{42}{2} + 42) = 90300$ Bayes factor values are computed. To decrease the computational cost, the symmetric comparisons (e.g. writer 2 against writer 1, as opposed to writer 1 against writer 2) are afterward added to the list of Bayes factor values, duplicating the opposite result.

3.5.2 Parameters and choices

The previous procedure can be modified in multiple ways. With respect to the reference article (Bozza et al., 2008), we initially chose not to separate loops by characters (e.g. sample only “a”s). Also, the number of loops in the reference and questioned sets, respectively k_{r} and k_{q}, is no longer fixed. Instead, a wide range of choices for such values is explored, to approach forensic cases where the questioned material varies in length, ranging from just a few words containing loops to a full-length text.

One can also investigate whether the choice of using unit area loops has an impact on the value of the evidence, by comparing the Bayes factors with and without rescaling. Notice that this choice can always be made since it refers to the way evidence is processed.

Finally, one can evaluate the impact on the Bayes factor of the choice of the harmonic components forming the feature vectors. This choice entails a number of effects. The usage of more harmonics (i.e. longer feature vectors, thus higher dimensional features) should provide stronger evidence weights, as more information

is used during evaluation. However, more background samples are required to keep up with the increased number of parameters to fit. For instance, feature vectors of length p require p free parameters for the within-writer mean vector θ, and $p(p+1)/2$ free parameters for the within-writer covariance matrix W. The between-writer parameters μ, B, U, also increase in dimension accordingly.

All Bayes factor computations were performed by obtaining 3000 Gibbs samples, and 1000 burn-in iterations.

3.5.3 Results

Due to the extremely large number of Bayes factor values computed in this section, and the large amount of explored choices, we start first by showing Bayes factors values aggregating all writer combinations together. We name this view *writer-independent*, as it provides initial insight into the model performance and the dataset characteristics, without focusing on the individual writer combinations.

This view is also useful to approach an $\mathcal{M}$-open forensic casework, where none of the writers belongs to this dataset, yet one supposes that their writing behavior can be captured by this model. In this case, one is not able to refer to a specific reference-questioned writer combination, but must accept the generic answer that this view is providing.

We take the opportunity also to evaluate the impact of several choices on the discriminative properties of our method, for instance:

1. rescaling character loops to have a unitary surface,
2. how many character loops are considered as reference and questioned sets,
3. consider only a subset of all available harmonics.

Next, the performance across the reference writers is detailed, in order to learn if (and which) writer combinations provide contrasting evidence. We call this view *writer-dependent*. This view is useful in cases where one of the writers (or somebody whose writing is very close to his/her) belongs to this dataset, as it provides an answer that accounts for his specific variability.

Writer-independent analysis

One can start by comparing the computed "distributions" of the Bayes factors, aggregating on the *true* hypothesis, without distinguishing writer pairs. Instead, we compare the effect of using loops with unit area against using loops with the original scale, considering $k_{\mathrm{r}} = k_{\mathrm{q}} = 20$ questioned and reference loops. All harmonics

($k = 1, 2, 3, 4$) are considered along with a_0 (corresponding to $p = 9$ variables). Following the Validation section (Section 3.4.4), the lowest possible value for the number of degrees of freedom of the inverse Wishart distribution is chosen, setting $\nu = \nu_{\min} = 27$ (Equation (3.5)).

Figure 3.12 shows the comparison between these distributions using boxplots. Statistics in tabular form are also available in Table 3.4. One first notices that the range of the computed log-Bayes factors is extremely large, whose 5% and 95% quantiles are approximately, respectively, 10^{-36} and 10^{16}. This is expected since we are dealing with multivariate data ($p = 9$), coming from continuous distributions with support $\mathbb{R}^9$, with multiple observations per sample ($k_{\mathrm{r}} = 20$ for instance) (Taroni et al., 2012).

Concerning the decision outputs, one notices that the model supports a true h_p during most replications (sensitivity of 97.76%). Under h_d, however, the model outputs a Bayes factor greater than 0 more than one trial over four (specificity of 67.29%). Another issue is that, under h_d, the spread of the Bayes factor values is much larger than those computed under h_p. The key to solving these issues lies in taking into account the variability between writers, as it will be shown later in this Section.

Concerning the usage of loops with unitary surface, the normalization operation has a negligible effect on the statistics, aggregated across all writers. In particular, it tends to improve the less-than-optimal performance under h_d. From now on, we decide to keep using loops with unitary surface, as originally done in the reference

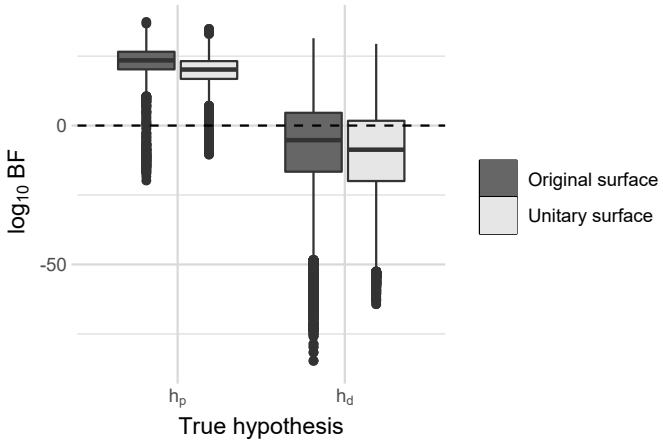


Figure 3.12: Computed log-Bayes factors according to the true hypotheses and the usage of unitary area loops. The dashed line marks BF = 1.

Table 3.4: Comparisons between the direction of the Bayes factor and the true hypothesis. In parentheses, the total number of trials.

True hypothesis	BF > 1	BF < 1
h_p	97.76% (8,212)	2.24% (188)
h_d	32.71% (112,664)	67.29% (231,736)

article.

One may wish to investigate the impact of considering a different number of reference and questioned samples, i.e. k_{r} and k_{q} respectively. To do so, the same procedure is repeated first by considering the same number of reference and questioned samples (i.e. $k_{\mathrm{r}} = k_{\mathrm{q}} = k_{\mathrm{rq}}$), then by varying k_{rq} over a wide range. All loops have unitary surface. Figure 3.13 shows again the Bayes factors distributions, according to k_{rq}. One can appreciate that the weight of evidence increases towards the respective direction as one takes into account more samples into the evaluative procedure. The method increases also in discriminative capacity, as for large k_{rq} most of the log-Bayes factors distributions stray away from the neutral value of 0. Notice again the much larger spread under h_d.

The last experiment performed within the writer-independent view is to form the feature vectors by considering a subset of all available harmonic contributions. In particular, one can explore the case where only one harmonic is considered, i.e. the terms a_k and b_k for a given k. This produces feature vectors of length $p = 2$, and

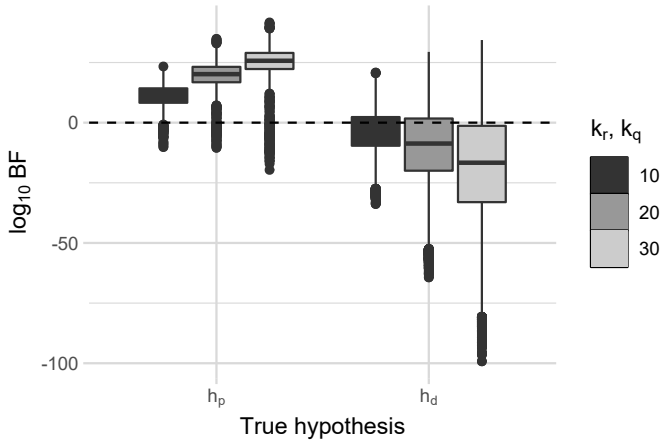


Figure 3.13: Computed log-Bayes factors according to the true hypotheses and the number of reference and questioned samples.

the degrees of freedom are adjusted accordingly to the minimum value ($\nu = 7$). We take h ranging from 1 to 4; the case $k = 0$ is always discarded since the code is not currently able to consider unidimensional distributions. $k_{\mathrm{r}} = k_{\mathrm{q}} = 30$ samples for the reference and questioned sets are considered, using only loops with unitary surface. Figure 3.14 shows the Bayes factors distributions, distinguishing by the considered harmonic indexes. The model using the full feature vector (using harmonics 1 to 4 plus a_0) is also juxtaposed.

As expected, feature vectors of higher dimensionality are much more discriminative than bivariate feature vectors, particularly under h_p. Also, as previously observed, writers tend to differentiate on $h = 2$, the harmonic that determine the loop elongations. The Bayes factor partially reflects this behavior, showing slightly more discriminative values for $h = 2$ as opposed to other harmonic contributions, such as $h = 1$ and $h = 3$.

The analysis of the distributions under h_d also shows that it is necessary to consider sufficiently complex evidence (e.g. with higher dimensionality) in order to avoid a high rate of false decisions. In fact, one can see that most of the single-harmonic distributions at best overlap the neutral log-Bayes factor value of 0, and in some cases the medians lie towards the "wrong" side. Under h_p, however, this phenomenon looks to be limited in magnitude.

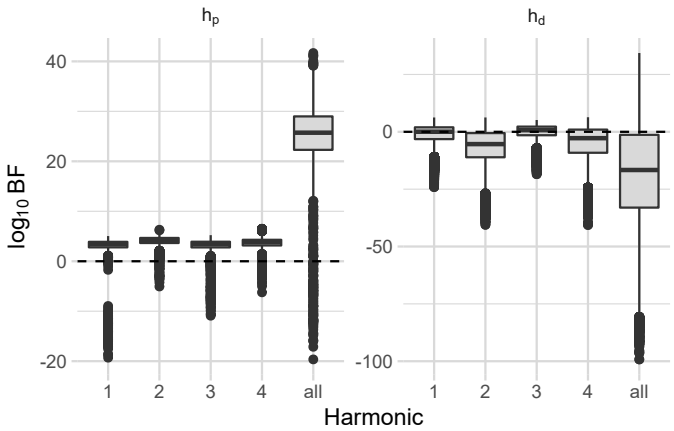


Figure 3.14: Computed log-Bayes factors according to the true hypotheses, isolating a specific harmonic contribution. The case "all" marks the results obtained with the complete feature vector. $k_{\mathrm{r}} = k_{\mathrm{q}} = 30$.

Writer-dependent analysis

One can now "zoom" on the obtained results, taking into account the individual writers that took place in the comparisons. For ease of visualization, we will consider every combination of *reference* writer and true hypothesis, thus obtaining $2 \times 42 = 84$ Bayes factors distributions. Considering the same parameters as Figure 3.12 (i.e. $k_{\mathrm{r}} = k_{\mathrm{q}} = 20$, all available harmonics), and imposing the usage of loops with unitary surface, one obtains Figure 3.15.

This view clears up some of the ambiguities shown in writer-independent analyses, such as the large spread under h_d and the median log-BF shifted towards 0. In particular, one notices that most writers produce loops that favor h_p when compared against themselves. Writers 9, 20 and 22 are exceptions to this rule, as log-Bayes factor values lower than 0 under h_p were obtained in a significant number of trials.

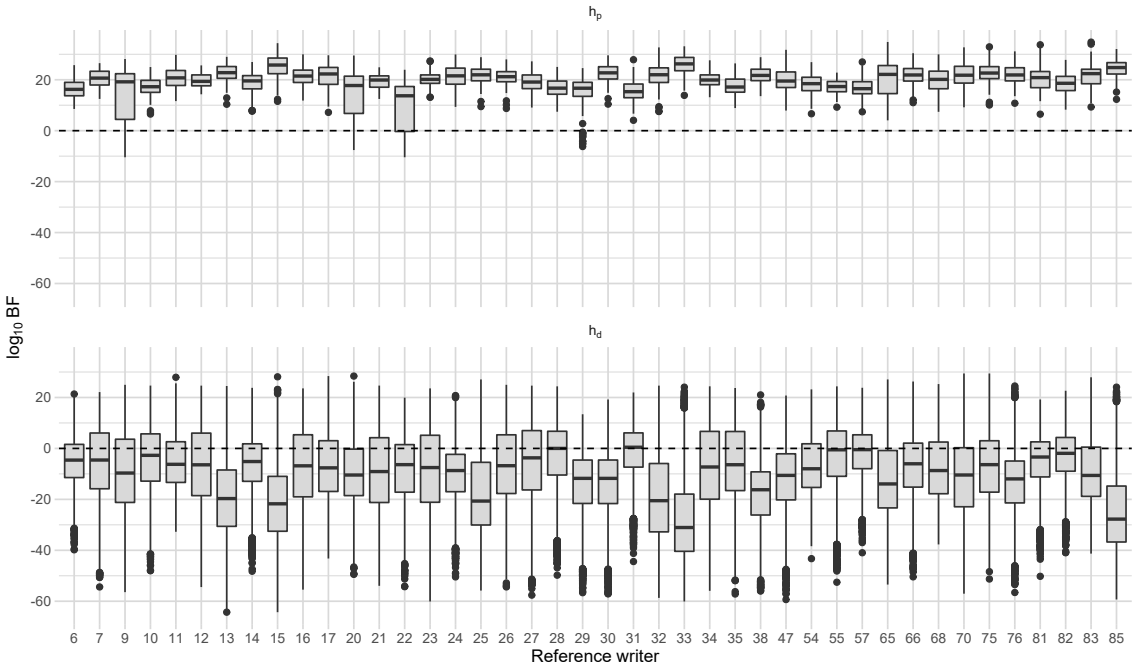


Figure 3.15: Computed log-Bayes factors according to the true hypotheses and the reference writer. In the upper section h_p is true; in the bottom part, h_d is true. Each boxplot is the distribution of the log-Bayes factors obtained by comparing the reference writer on the x-axis with the reference writer (in the top panel), or all other writers (in the bottom panel).

Under h_d, instead, performance varies greatly according to the writer who is providing the reference samples. Some of the writers, including (but not limited to) 13, 15, 33 and 85, write dissimilarly (according to the model) when compared to any other writer in the dataset: this results in log-Bayes factors under h_d that stray away from 0. Most of the other writers, instead, contribute to the high spread of the distributions: some of the reference-questioned writer pairings produce values that are closer to 1 compared to other writer combinations. To discover which ones, one should proceed with a deeper analysis, separately considering each writer-writer pairing.

In this case, a graphical presentation of the results is particularly complicated due to the large amount of distributions to analyze, each one with 100 replications of the Bayes factor: with 42 writers, one has $42^2 = 1764$ pairs. A simplified view of the results is given in Figure 3.16, where only the median of these distributions is shown. One can see that on the diagonal (i.e. when h_p is true), the median Bayes factor always points to h_p. The converse is often true under h_d, particularly with writers whose writing is dissimilar than most of the other writers (e.g. writers 32 and 33), and with selected pairs (e.g. writer 13 against 15). In cases where the Bayes factor wrongly supports h_p, the evidence is weaker compared to cases where h_p is true.

These considerations confirm that loop shapes can potentially constitute an interesting feature to discriminate writers, according to the examined dataset, as previously shown by (Marquis et al., 2005) and later works. In particular, the hypotheses (see Section 1.3) that writers possess a *master pattern* (the first level of the hierarchical model) and exhibit some degree of *natural variation* around it (the second level of the hierarchical model) seem to be well-captured by the specified model. The computed Bayes factors serve as an additional confirmation of this fact for most of the considered writers. However, it also results that, in some cases, some writer pairs seem to contradict these hypotheses, by providing samples that are similar (as per their master pattern) and within the range of natural variation of the reference writer. When it happens, however, the strength of the (contradicting) evidence is generally weaker than the situation where only one writer provides both the reference and the questioned samples.

The contradicting evidence can, probably, be accounted for by considering more details into the model, such as a third level, a character-level dependence or the position of the letter inside the word (Bozza et al., 2008).

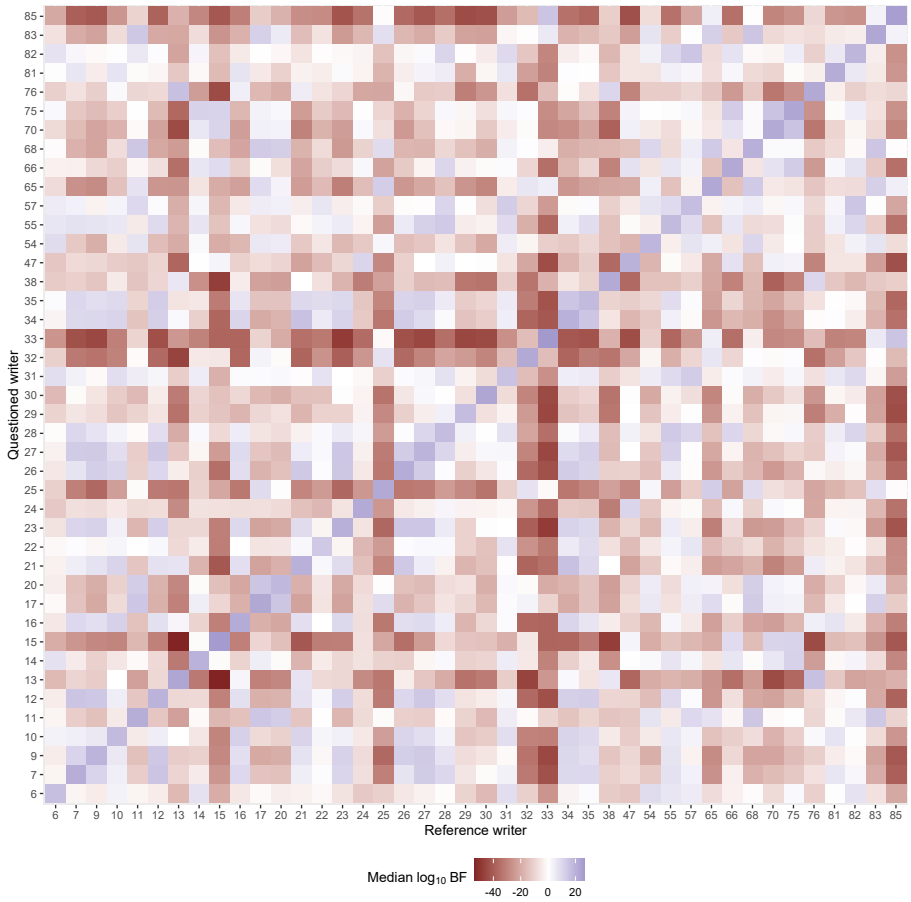


Figure 3.16: Median of 100 log-Bayes factors obtained when comparing one writer against another. Colors mark the median of the log-Bayes factors. On the diagonal, h_p is true. Blue marks evidence towards h_p, red marks evidence towards h_d.

3.6 Results on forged signatures

In this Section we analyze the loop shapes in the forged signatures dataset (Section 3.1.2).

3.6.1 Evaluative scenario

Let e_{r} be a set of character loops, parametrized by their Fourier coefficients, from the reference writer. Let e_{q} be a set of character loops, parametrized by their Fourier coefficients, from the questioned writer (i.e. the writer of the questioned material, who may or may not correspond to the reference writer).

The evaluative scenario described in Section 3.3.2 is adopted, in particular with the same hypothesis pair:

$$\begin{aligned} H = h_p &: \quad \text{“the loops } e_{\mathrm{r}} \text{ and } e_{\mathrm{q}} \text{ come from the same writer”,} \\ H = h_d &: \quad \text{“the loops } e_{\mathrm{r}} \text{ and } e_{\mathrm{q}} \text{ come from two different writers”.} \end{aligned}$$

As opposed to the simulated dataset and the natural handwriting situation, the interpretation of the evaluative scenario is delicate. In particular, we assume that there exists a relevant population, i.e. there exists a set of writers that produce character loops using the same “mechanism” followed by the reference writer. One possible way of how this could happen is to consider the case where a set of writers, the reference population, attempt to reproduce the reference signature as if it were their own. Thus, the reference population is constituted by a set of forgers performing freehand simulations (see Section 1.3.1) of the victim's signature. Notice that, by doing so, we disregard the differences between handwriting (where writers follow their own master pattern) and signatures (where writers try to imitate others' master patterns). Thus, the applicability of the model itself may also be discussed: this is detailed in Section 3.7.1.

Typically, as the size of the relevant population is usually very small (corresponding to the number of alleged forgers), the scenario simulation procedure (Section 2.5.5) is not directly applicable. As the Bayesian model for character loops (Section 3.3.1) requires some prior knowledge on the model parameters (in particular writer's inter-variability parameters μ, B, U and ν), the additional samples provided by the various forgers were used as background observations.

From this point onwards, the procedure is identical to the one followed in the natural handwriting situation.

In this case, the total number of writers in the dataset is 7. Every possible pairing between a reference and a questioned writer is explored. Once the pair is chosen, the

procedure is repeated $G = 100$ times, to account for the writer's intra-variability. Each time new e_{r}, e_{q} and e_{b} are sampled, re-eliciting the priors with the new background dataset. In the end, since the dataset contains 7 different writers, $100(\binom{7}{2} + 7) = 2800$ Bayes factor values are computed. To decrease the computational cost, the symmetric comparisons (e.g. writer 2 against writer 1, as opposed to writer 1 against writer 2) are afterward added to the list of Bayes factor values, duplicating the opposite result.

3.6.2 Parameters and choices

In the light of the results on the natural handwriting dataset, loops are always normalized to have unitary surface. Following the Validation section (Section 3.4.4), the lowest possible value for the number of degrees of freedom of the inverse Wishart distribution is chosen, depending on the number of considered harmonic contributions.

The number of loops in the reference and questioned sets, respectively k_{r} and k_{q}, is no longer fixed.

The number of harmonic contributions in the feature vector is allowed to vary, from the full vector (considering harmonic terms with $k = 1, 2, 3, 4$ plus a_0) to single harmonic contributions (where $k \in \{1, 2, 3, 4\}$). The length of the feature vector p ranges from $p = 2$ (when one harmonic contribution is considered) to $p = 9$ (when one considers 4 harmonic contributions along with a_0).

The settings for the Gibbs sampler were unchanged with respect to the natural handwriting case: all Bayes factor computations were performed by obtaining 3000 Gibbs samples, and 1000 burn-in iterations.

3.6.3 Operative limitations

Let us indicate with m the number of writers in the dataset. It is easy to see that m and p, the length of the feature vector, are related. This constitutes a significant obstacle to the procedure described above when one wants to obtain an estimate of B, the between-writer covariance matrix, by using the sample covariance matrix $\widehat{B}$ (its maximum likelihood estimator). It can be proven that $\widehat{B}$ is not invertible if $m \leq p$ (Johnson & Wichern, 2007, sec. 3.4).

This situation is encountered when considering all harmonic contributions together ($p = 9$) with the $m = 7$ writers in the dataset. As the Gibbs sampler iterations require the calculation of the inverse of $\widehat{B}$, the calculation of any Bayes factor was not possible.

There are several ways to circumvent this problem, including, but not limited to:

1. recruit more forgers in order to increase m,

2. use another estimator for B from the background observations, such as a regularized or shrinkage estimator (Ayyıldız et al., 2012; Bai & Shi, 2011),
3. change the elicitation method for B, for example by considering information coming from a related study, or
4. consider less Fourier coefficients in order to decrease p.

Single harmonic contributions (i.e. $p = 2$) did not pose any problem during the estimation of the Bayes factor. As a consequence, all the following results are presented by considering all harmonics separately ($k \in \{1, 2, 3, 4\}$). The problem of achieving a combined measure of evidence is later discussed in Section 3.7.

3.6.4 Results

As in the natural handwriting dataset, one can approach the problem first in the writer-independent view, by distinguishing upon which hypothesis is true. Also, the influence of the choice of k_{r} and k_{q} is evaluated.

Afterward, the performance across the reference writers is detailed in the writer-dependent view, in order to learn if (and which) writer combinations provide contrasting evidence. Following the natural handwriting procedure, this also would include comparing the handwriting coming from two different forgers.

In the context of forged handwriting, one could evaluate the capacity of each person of reproducing the character loops in a consistent way. Replications of a single person's handwriting should not stray far from the victim's master pattern, while replications coming from different persons could be different. Notice, however, that the forgers attempt to reproduce a master pattern that might be different from their own. This could lead to inconsistencies in the reproduction, a variability that is much larger than the true writer's, or even completely disregard the fidelity of replication of shape. As a consequence, one should not be surprised by any contradictory results in terms of the value of the evidence, rather interpret them in the light of this scenario. An alternative approach to the questioned signature problem is discussed in Section 3.7.1.

Victim-based view

A sub-problem of interest when dealing with forged handwriting consists in considering the legitimate owner of the signature as the reference writer, and one person among the forgers as the questioned writer. This is the *victim-based* view.

In this way, one could evaluate the skill of each forger in reproducing the victim's handwriting characteristics. Particularly, character loops coming from a skillful forger

(h_d) would be close to the victim's master pattern and within his range of natural variation, incorrectly supporting h_p.

Notice also that if the victim's handwriting is inconsistent (e.g. the victim's writing habits evolved significantly over the time of collection of the reference material, so to disrupt the victim's master pattern), it would be possible to obtain Bayes factor values that incorrectly support h_d, as the compared character loops coming from the victim would be "too different" from each other. For instance, such inconsistency could be revealed by comparing character loops written at very distant times. Notice that this phenomenon cannot be directly revealed in this Thesis, as the e_{r}s are randomly sampled from the available reference material. However, the resulting log-Bayes factors distributions under h_p would shift towards 0 as the number of compared loops increases, since the distributions would include samples taken from periods where the victim's habits were different.

Writer-independent analysis

As opposed to the natural handwriting case, the reduced number of writers in the dataset forces us to compute the Bayes factors for each harmonic contribution. It is nevertheless interesting to compare these resulting values with the natural handwriting results.

To this purpose, one might want to start by evaluating the sensitivity of the Bayes factors to the number of considered loops, k_{r} and k_{q}. Figure 3.17 shows that the second harmonic is more discriminating, as already previously noted. Results under h_p are still in accordance with the postulated hypothesis, i.e. character loops drawn by the same person are more similar than character loops drawn by different persons. However, the general performance of the results under h_d is worse than the one obtained in the natural handwriting case (compare against Figure 3.13 and Figure 3.14). In particular, the median of many of the log-Bayes factors distributions is greater than the neutral value of 0, pointing thus towards the wrong hypothesis. This anomaly persists even if the number of considered samples k_{r} and k_{q} increases up to 50. It can also be shown that the usage of loops on the original scale does not offer any improvement.

As previously reported in Section 3.6.4, one might now consider the victim-based view, i.e. the distributions obtained when the reference writer is the legitimate owner of the signature. The distributions, represented in Figure 3.18, show that the victim's signature is consistent according to the model, resulting in log-Bayes factor values greater than 0.

Under h_d, a forger is providing the questioned character loops. The distributions

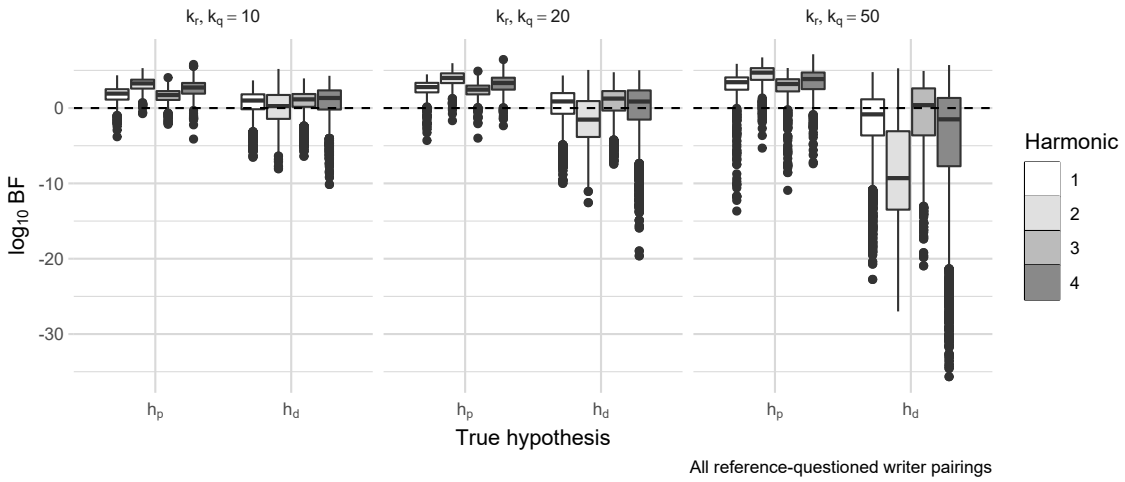


Figure 3.17: Computed log-Bayes factors according to the true hypotheses, the number of reference and questioned samples k_{r} and k_{q}, and the considered harmonic contribution k. All possible writer pairings are considered.

at a moderate amount of recovered material (e.g. $k_{\mathrm{q}} \leq 20$) support h_p most of the times. This shows that, in general, it is not possible to formulate a judgment of forgery using 20 reference and questioned loops under this specific statistical model.

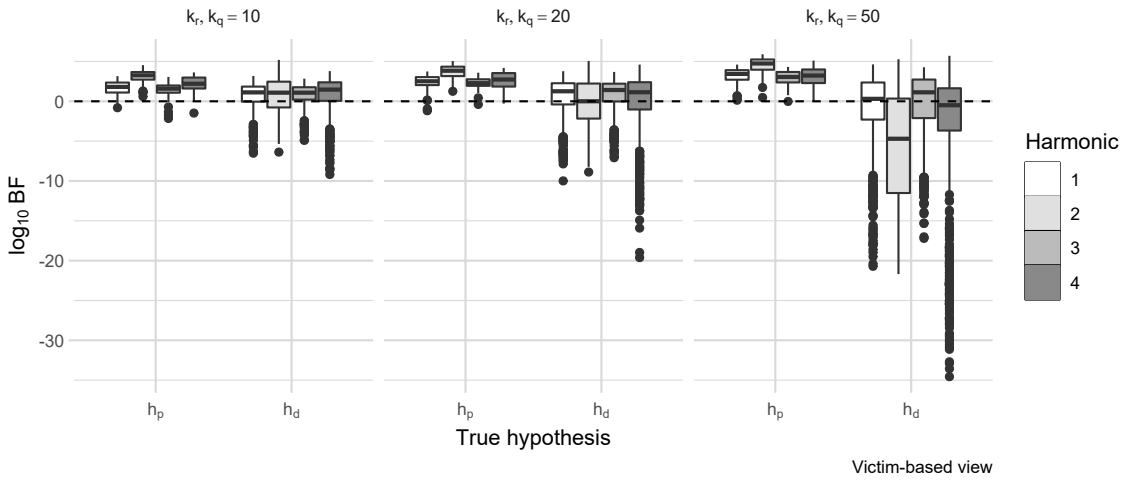


Figure 3.18: Computed log-Bayes factors according to the true hypotheses, the number of reference and questioned samples k_r and k_q, and the considered harmonic contribution k. The reference writer is the victim.

Writer-dependent analysis

Similarly to the writer-dependent analysis in the natural handwriting case (Section 3.5.3), one can detail the previous analysis by distinguishing across the *reference* writer and the individual harmonic contributions. For ease of visualization, we will consider every combination of reference writer, harmonic contribution and true hypothesis, thus obtaining $2 \times 4 \times 7 = 56$ Bayes factors distributions. Fixing $k_\mathrm{r} = k_\mathrm{q} = 20$, one obtains the Bayes factors distributions shown in Figure 3.19.

The method performs as expected under h_p, as most of the Bayes factors distributions support h_p most of the time, and across most harmonic coefficients. This shows that every writer is consistent in his reproduction of the character loop shape. It can be noted that the writer "F3" has a rather large spread of the Bayes factor values concerning the first harmonic contribution: also, approximately one-fourth of the times a log-Bayes factor lower than 0 was obtained, wrongly supporting h_d. This means that the writer was less able to reproduce the ovate contribution to the shape in a consistent way.

Under h_d, however, most distributions have a median value of the log-Bayes factor that is greater than 0, supporting the wrong hypothesis. As happened in the natural handwriting case, most distributions have a large spread, calling for more detailed analyses across harmonics and writer pairings.

The victim-based view is forensically interesting, as it allows us to concretely

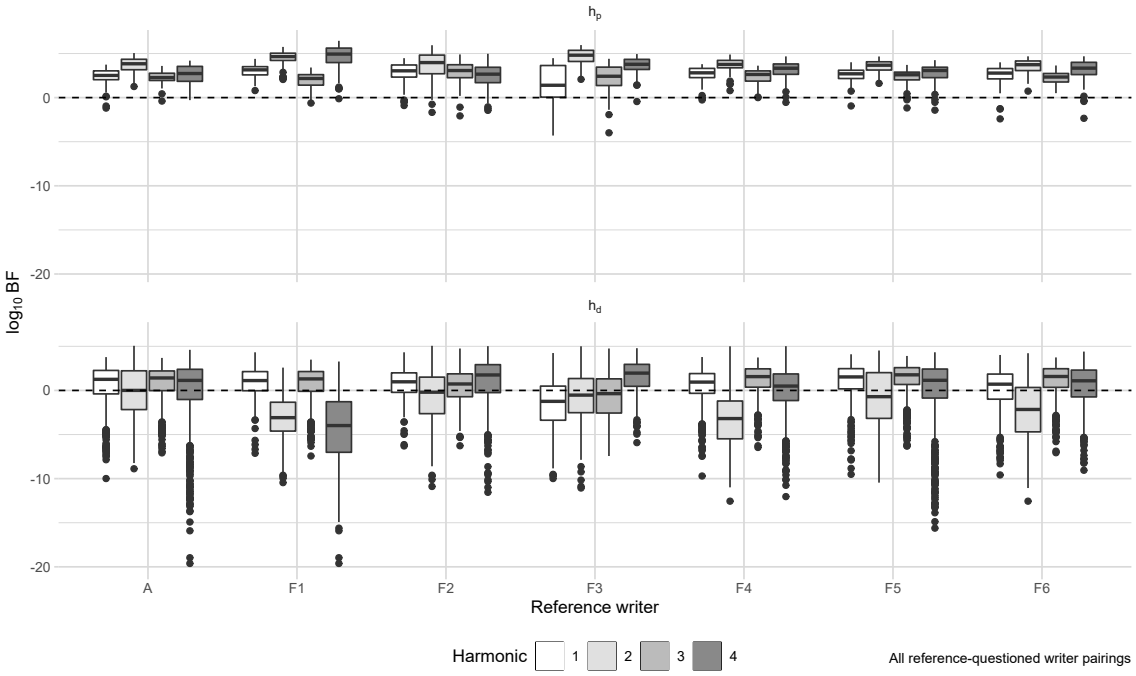


Figure 3.19: Computed log-Bayes factors according to the true hypotheses and the reference writer. All possible writer pairings are considered.

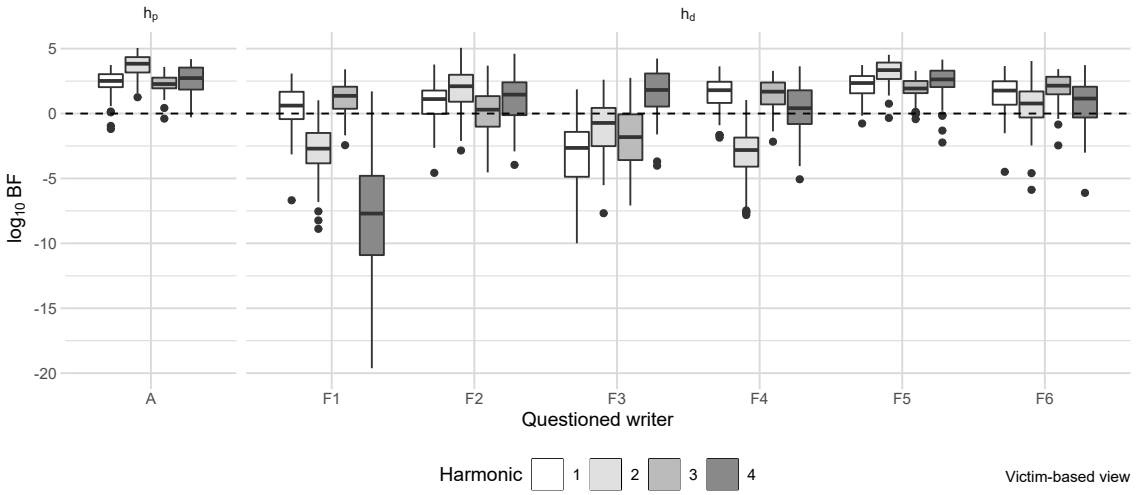


Figure 3.20: Computed log-Bayes factors according to the true hypotheses and the writer of the questioned material. The reference writer is the victim.

evaluate the skill of each forger according to this model. By considering the legitimate owner of the signature as the reference writer, one obtains the distributions shown in Figure 3.20. As previously said, a skillful forger would successfully be able to reproduce the victim's habits, expecting a log-Bayes factor close to 0. This is particularly true for forgers "F5" and "F6", who were able to achieve the same order of magnitude of the log-Bayes factor as the reference writer across all shape contributions. Other forgers are less successful in their task, in particular forger "F1" who was not able to properly imitate the loop shapes according to the model.

The writer's consistency can be also evaluated by allowing the number of considered samples to vary, i.e. k_r and k_q, over the set $\{10, 20, 50\}$. The distributions, represented in Figure 3.21, confirm firstly that the author is consistent in reproducing his loop shape (under h_p), as the Bayes factor does not decrease if the amount of considered material increases. Secondly, the victim-based view confirms that there are less skilled forgers, particularly "F1". Even at moderate sample sizes (e.g. 20 loops), the model is discriminating their replications from the ones coming from the victim. Thirdly, as the sample size increases, most forgers tend to show some inconsistencies in their reproductions, which contribute to decreasing the Bayes factor across all harmonic contributions. Even a moderately skilled forger, such as "F6", fails to reproduce some aspects of the shape at high sample sizes, for instance the triangularity (the second harmonic). Finally, the victim-based view confirms that forger "F5" is particularly skilled, reaching the same performance as the reference writer under h_p across all shape contributions. This achievement is also stable at all sample sizes, showing that "F5" has been able to achieve a consistent reproduction of the character loops.

3.7 Extensions

In the case of forged signatures, all harmonics were treated separately, as it was not possible to compute the Bayes factor using all harmonic coefficients at the same time (see Section 3.6.3).

It is a well-known fact that, given two independent evidence items e_1 and e_2, the Bayes factor obtained by jointly considering e_1 and e_2 is the product of two Bayes factors, one obtained when considering e_1, and the other obtained when considering e_2 respectively (Taroni, Biedermann, et al., 2014, ch. 8).

Let us fix a set of reference and questioned character loops, e_r and e_q respectively. Let us indicate with BF_k the Bayes factor calculated using the harmonic coefficients of the k-th harmonic. By extension, BF_0 is the Bayes factor calculated considering solely a_0. Let us also indicate with $\mathrm{BF}_\mathrm{full}$ the Bayes factor calculated using all available harmonic coefficients: $\{a_0\} \cup \{a_k, b_k\}_{k=1}^{4}$. By supposing that harmonic contributions

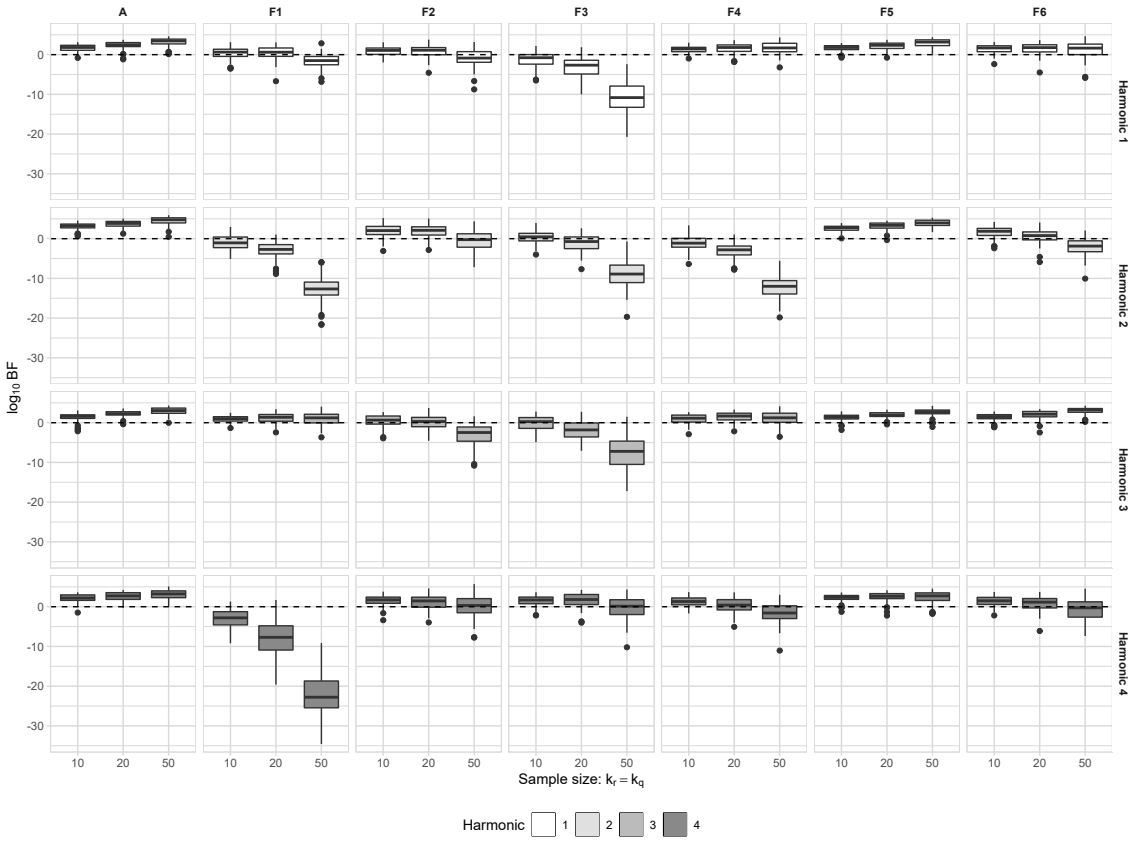


Figure 3.21: Computed log-Bayes factors according to the questioned writer (a column), the harmonic contribution (a row) and number of considered samples $k_{\mathrm{r}} = k_{\mathrm{q}}$ (the x-axis). The reference writer is the victim. In the first column, h_p is true. In the other columns, h_d is true.

are independent, one could exploit the above fact to compute $\mathrm{BF}_{\mathrm{full}}$ from the set of BF_k. In that case, one should have that

$$\log \mathrm{BF}_{\mathrm{full}} = \sum_{k=0}^{4} \log \mathrm{BF}_k \ . \tag{3.7}$$

The independence assumption, however, needs to be verified using the full model. For instance, if harmonic contributions are pairwise-independent, all within-writer covariance matrices W_i can be partitioned to have blocks of 0s on rows that belong to different harmonic contributions. Notice that W_i are latent, so an estimator is needed. A similar reasoning should apply for all other covariance matrices such as B, and the inverted Wishart scale matrix U.

An alternative approach can be attempted when one is able to compute the Bayes factors using the full vector, for example in the natural handwriting case. Then, it is possible to compute and verify Equation (3.7).

As an example, Figure 3.22 shows a graphical representation of Equation (3.7) with the natural handwriting dataset, under the writer-independent view. In particular, we first compute each BF_k for all k, then represent the partial sums in Equation (3.7) for k from 1 to 4, until all harmonics are considered. The value of $\mathrm{BF}_{\mathrm{full}}$ is also juxtaposed for comparison. Notice that it was not possible to compute BF_0 since our model implementation does not allow unidimensional vectors (i.e. those involving only a_0).

If all harmonic contributes are independent, the Bayes factor obtained by summing all individual contributes should be equal to the Bayes factor calculated by considering the full vector from the beginning. In the natural handwriting case, the Bayes factors distributions overlap under h_d, but not under h_p. This means that the harmonics are not independent under h_p: the equality in Equation (3.7) does not hold.

If harmonic independence holds, one could exploit Equation (3.7) to calculate the left-hand side (i.e. $\mathrm{BF}_{\mathrm{full}}$) by computing only the right-hand side (i.e. BF_k for every k). This could reveal useful in cases when one cannot compute the left-hand side, for instance when the number of writers in the dataset is too small (see Section 3.6.3). However, the graphical procedure shown in this Section cannot substitute a proper evaluation of the independence hypothesis, which should call for stronger assumptions involving the model structure and expert knowledge.

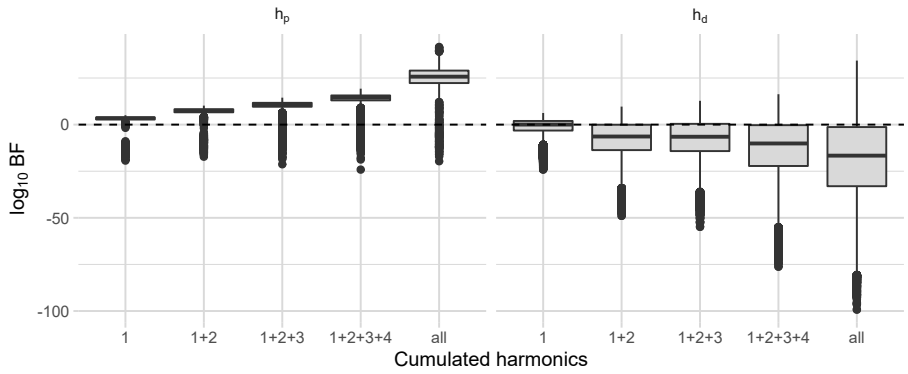


Figure 3.22: Computed log-Bayes factors according to the true hypotheses, obtained by calculating the Bayes factors on single harmonics, then by cumulating across k according to the x-axis labels. The case "all" marks the results obtained with the complete feature vector (including the contribution of $k = 0$).

3.7.1 Alternative models for questioned signatures

As previously shown, natural handwriting and questioned signatures are intrinsically different situations, therefore should require different statistical approaches. Natural handwriting assumes that writers follow their own master pattern, neglecting psychological factors and disguise attempts. From the statistical point of view, it is reasonable to assume that under h_d (samples coming from different writers), the writers are independent, as their master patterns are not related. However, signature forgers attempt to reproduce the victim's master pattern, with varying degrees of success depending on the forger's skill, the amount of reference material available to the forgers, and the victim's own variability. In other terms, since forgers usually have some knowledge of the victim's signature, it is therefore unreasonable to assume that the samples are independent from the victim's under h_d.

3.7.2 Non-independence

The effect of this assumption on the computed Bayes factor can be briefly investigated. As known, the generic Bayes factor for an evaluative scenario (Equation (2.6)) can be written as follows:

$$\mathrm{BF} = \frac{\Pr(e_\mathrm{r}, e_\mathrm{q} \mid h_p)}{\Pr(e_\mathrm{r}, e_\mathrm{q} \mid h_d)}, \tag{3.8}$$

where $\Pr(\cdot)$ indicates a probability density function or a marginal likelihood as in the hierarchical model (Equation (2.10)). By assuming that the sources are independent under h_d and applying the methods shown in this Chapter, one computes the following Bayes factor (indicated with BF_0):

$$\mathrm{BF}_0 := \frac{\Pr(e_\mathrm{r}, e_\mathrm{q} \mid h_p)}{\Pr(e_\mathrm{r} \mid h_d)\Pr(e_\mathrm{q} \mid h_d)} \,. \tag{3.9}$$

To quantify the degree of dependence between the items e_r and e_q, it is helpful to recur to the conflict measure introduced by Chamberlain et al. (2013):

Definition 3.2 (Conflict measure)
The conflict C between items of evidence e_r and e_q under the hypothesis h_d is:

$$C = \mathrm{conf}(e_r, e_q \mid h_d) := \log \frac{\Pr(e_r \mid h_d)\ \Pr(e_q \mid h_d)}{\Pr(e_r, e_q \mid h_d)}$$

The conflict C has the following properties:

- $C = C(e_\mathrm{r}, e_\mathrm{q}) \in \mathbb{R}$: C is itself a random variable, and depends on the observed evidence.
- $C > 0$ implies **conflict**: evidence is less likely to be jointly seen,
  - $\Pr(e_\mathrm{r}, e_\mathrm{q} \mid h_d) < \Pr(e_\mathrm{r} \mid h_d)\Pr(e_\mathrm{q} \mid h_d)$,
  - $\Pr(e_\mathrm{q} \mid h_d, e_\mathrm{r}) < \Pr(e_\mathrm{q} \mid h_d)$,
  - $C \to +\infty$ the more e_r and e_q are incompatible.
- $C < 0$ implies **synergy**: evidence is more likely to be jointly seen,
  - $\Pr(e_\mathrm{r}, e_\mathrm{q} \mid h_d) > \Pr(e_\mathrm{r} \mid h_d)\Pr(e_\mathrm{q} \mid h_d)$,
  - $\Pr(e_\mathrm{q} \mid h_d, e_\mathrm{r}) > \Pr(e_\mathrm{q} \mid h_d)$,
  - $C \to -\infty$ the more e_r and e_q are dependent.
- $C = 0$ implies the **independence**,
  - $\Pr(e_\mathrm{r}, e_\mathrm{q} \mid h_d) = \Pr(e_\mathrm{r} \mid h_d)\Pr(e_\mathrm{q} \mid h_d)$,
  - $\Pr(e_\mathrm{q} \mid h_d, e_\mathrm{r}) = \Pr(e_\mathrm{q} \mid h_d)$.

Finally, the generic Bayes factor in Equation (3.8) can be rewritten as follows:

$$\mathrm{BF} = \exp(C)\frac{\Pr(e_\mathrm{r}, e_\mathrm{q} \mid h_p)}{\Pr(e_\mathrm{r} \mid h_d)\Pr(e_\mathrm{q} \mid h_d)} = \exp(C)\,\mathrm{BF}_0 \,. \tag{3.10}$$

Considering the situation of interest where e_{r} and e_{q} represent handwritten signatures, it is reasonable to assume that some information is transferred between the random variables, for example the generic aspect of the signature. As a consequence, it is likely that $C < 0$ (synergy), giving $\mathrm{BF} < \mathrm{BF}_0$.

Under h_p one has that $\mathrm{BF} > 1$, so the independence assumption is less conservative and over-evaluates the evidence (BF_0 is stronger than BF). However, this line of reasoning does not allow to conclude anything on the properties of BF_0 under h_d.

Even though the specific value of C is not easy to determine, we postulate that C becomes larger (in absolute value) the more information can be extracted by the questioned writer from the reference signatures. For instance if the forger has access to good quality reference material, C should be much lower than if the forger were only aware of the victim's name. In the latter case it would be safe to assume the independence between sources, thus $\mathrm{BF} = \mathrm{BF}_0$.

For the same principle, if evidence is constituted by character traits that are less apparent to the untrained eye, C should be closer to 0 than the more informative situation when the victim's signature is constituted by a single letter with a closed loop: in the latter case, the forger would probably focus on imitating the loop shapes, thereby linking e_{r} and e_{q}.

3.7.3 Literature

Linden et al. (2021) recently raised the same considerations and developed an alternative approach to the questioned signature problem. In particular, they proposed to collect two kinds of background signature datasets. The first is constituted by the study participants' genuine signatures, each one replicated multiple times: this is the genuine signature dataset. For the second kind, one of the participants is chosen as the "victim". Forgers were recruited, and were instructed to produce a set of replications of the victim's signature: this is the forged signature dataset. For the sake of completeness of the article, three victims were chosen in order to evaluate the proposed method on varying types of signatures.

As opposed to our approach (Section 3.6.1), these datasets are used separately to elicit the background source parameters θ_i and W_i as well as the hyperparameters. Under h_p, the priors are updated using the genuine dataset, while under h_d the forged signature dataset is used instead.

A second major difference between the two approaches is that Linden et al. (2021) avoid modeling the between-writer variability (B in our Equation (3.3)), putting instead a Normal-Wishart prior on the within-writer parameter vector (θ_i, W_i). This choice produces a Bayesian model that is conjugated, and the Bayes factor is available

in closed form.

The advantages of their proposed approach are multiple: the calculation of the Bayes factor value is much faster, and the model is far simpler to implement and verify. However, it is necessary to collect two datasets instead of one, and their approach could not be valid in cases where the sources are independent under h_d.

Concerning the performance of decision a procedure that chooses h_p if BF > 1 and h_d if BF < 1, our approach appears to perform better with two victims' signatures out of three, irrespective of the available number of reference signatures k_{r}. Notice also that the articles are not directly comparable, as our dataset is constituted by 6 forgers, while they recruited at least 16 participants. Also, the feature vectors and operative conditions are very different: this Chapter considers off-line signatures described with a 9-dimensional vector based on the shape of closed character loops, while Linden et al. (2021) examined on-line signatures represented with a bivariate descriptor that exploits simple dynamical properties of the signature (including speed, duration, pressure and pen lift timings).

3.8 Discussion

The significant computational improvements over the original model, introduced by Bozza et al. (2008) in the non-constant within-writer covariance form, allowed us to apply their technique to cases that were not considered by the original authors, such as questioned signatures. Despite the relative simplicity of the procedure, the method offers a large number of choices that need to be made, such as the loop area normalization or the methods for elicitation of priors. It has been shown that their impact strongly depends on the specific case. The increased computational speed of our implementation allows scientists to conveniently evaluate these choices, comparing the obtained Bayes factors distributions before and after any choice is made.

Concerning the results from a forensic perspective, we confirm that the proposed method is capable of evaluating handwritten evidence showing closed character loops. If evidence consists of naturally handwritten material, the method shows a performance that is comparable to results in past literature, allowing for the objective evaluation of characteristics such as writers' consistency and adherence to their master pattern. If evidence involves questioned signatures, the technique allows scientists to concretely evaluate the forgers' skill in replicating the victim's character loop shapes.

It is nevertheless important to note that the method may appear to show several limitations and requirements. In this Chapter we have shown how an interested user can possibly deal with some of them. For instance, the number of degrees of freedom of the inverse Wishart distribution, ν, has to be set as low as possible, to

avoid obtaining contradicting Bayes factor values.

Some limitations, such as the requirement of a background dataset, relate rather to the general Bayesian setting of the evaluative scenario ("Is it reasonable to assume that a reference population exists when considering questioned signatures?").

Other limitations are intrinsic to any procedure that exploits any statistical model, such as the appropriateness of the multivariate Gaussian model to the considered evidence. Typically, this assumption must be motivated by a statistical test, a graphical method, or the existence of a physical law. In the FHE context, and given the high-dimensionality of the treated data, the verification of this assumption is a challenging task. Future works could address and weaken the assumption on the structure of the model, to introduce distributions that are data-driven, ranging from the usage of kernel density estimators (Aitken & Lucy, 2004) to Bayesian nonparametric techniques (Ghosh & Ramamoorthi, 2003).

It is promising to note that this model can easily extend to other types of evidence, as the multivariate Normal distribution is often used to describe correlated variables that cluster around a particular point in space (the mean).

Chapter 4

Quantifying simple signatures

This Chapter is dedicated to the analysis of handwritten evidence that does not contain any character loop. In particular, this Chapter is inspired by real casework, where an expert was faced with a set of questioned signatures that did not appear to show any strongly individualizing features. The specific casework data composes the dataset analyzed in this Chapter.

A descriptor is introduced, tuned to the specific case at hand. Following the generic Bayesian framework of Chapters 2 and 3, a statistical model is introduced for this specific descriptor.

The model sensitivity, behavior and specificities are first verified on fake (generated) data, then the actual casework dataset is approached. Since this specific case does not provide any background dataset, as opposed to Chapter 3, a novel data-driven mathematical procedure is introduced to elicit the hyperparameters.

The Chapter ends by a discussion of the generalizability of this descriptor to other kinds of data, such as repeated measurements of the composition of a certain substance. This generalization will be further exploited in Chapter 5, to analyze bacterial populations in human saliva in conjunction with handwritten characters. The Chapter will focus on the joint evaluation of two sets of evidence.

4.1 The dataset

The dataset comprises 24 signatures of a single person, whose authenticity was disputed. Accordingly, 44 reference signatures were collected, appearing in the same type of document as the questioned material. Due to the confidentiality agreement, more details on their origin cannot be disclosed. In total, the dataset comprises 68 signatures.

An element of novelty characterizing this specific case lies in the flowing nature of all signatures in the dataset. A questioned specimen is shown in Figure 4.1. No letters can be identified, all signatures (reference and questioned) are constituted by a single pen stroke, and a paraph at the end. One of the dominant features of these signatures is the presence of a number of "peaks" (and corresponding "valleys"), whose numbers and shapes may vary across the seized material. For instance, the signature shown in Figure 4.1 has four peaks and four valleys. The stroke can be self-intersecting, eventually drawing small loops at some of the upper peaks. The pen motion is always continuous and smooth, except for the peaks.

The FHE postulated that the peaks can be exploited for the purpose of conducting a forensic handwriting examination under the Bayesian framework introduced in Chapter 2. The challenges are manifold. Firstly, one needs to translate these features to numerical vectors, for quantitative comparison. Secondly, one needs a suitable statistical model to conduct this analysis. As it has already been seen in Chapter 3, this model must deliver a statistical description of the evidence that is defensible, for instance by appealing to a physical law or to expert knowledge. Lastly, the model must allow for a feasible computation of the Bayes factor.

Another feature that could be exploited is the position of the small paraph at the end of the signature. It was not considered in this analysis, as it belongs to a different stroke than the main one, and it has a less pronounced appearance than the signature peaks.

4.2 Features

To begin the extraction of the peak-based features, all signatures were digitalized at 600 dpi, then converted to a black and white uncompressed format.

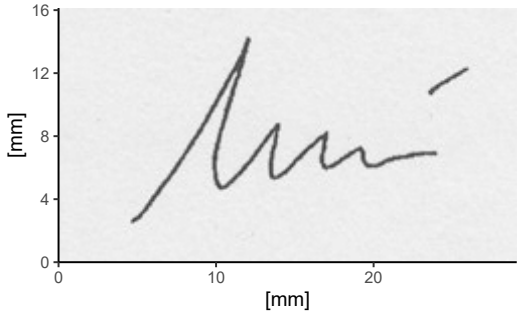


Figure 4.1: A questioned signature.

The main stroke was manually traced with a piecewise spline curve, then parametrized with the curve representation $\gamma(t)$:

$$\gamma : t \mapsto (x(t), y(t)) \quad t \in [0, 1]\,, \tag{4.1}$$

where t is the "time" coordinate, running from the starting point ($t = 0$) to the end ($t = 1$).

Notice that the term "time" is conventionally used to describe curve representations, even if it bears no relation with the concept of physical time. In our setting, the coordinate t simply indicates one mapping from $[0, 1]$ to the position of the pen on the paper.

In principle, it is possible to parametrize a stroke with a parametric curve $\gamma(t)$ such that the pen stops for a certain amount of time, yet the same stroke is traced as t spans the interval $[0, 1]$. Since the Chapter concerns off-line signatures, those possibilities are excluded to avoid introducing ambiguities in the definitions.

4.2.1 Curvature representation

One wishes to locate the positions of the "peaks" and the "valleys" along the signature stroke. A useful tool is the (signed) curvature:

Definition 4.1 (Curvature)
*Given a curve $\gamma(t) = (x(t), y(t))$, the **signed curvature** $\kappa(t) \in \mathbb{R}$ is:*

$$\kappa(t) := \frac{x'(t)\, y''(t) - y'(t)\, x''(t)}{(x'(t)^2 + y'(t)^2)^{\frac{3}{2}}}\,,$$

where $x'(t) = \mathrm{d}x/\,\mathrm{d}t$ and $x''(t) = \mathrm{d}^2x/\,\mathrm{d}t^2$.

$\kappa(t)$ has a precise geometrical definition: the quantity $R(t) = 1/\kappa(t)$ is the (signed) radius of the osculating circle at time t. Intuitively, the circle has the smallest radius when the stroke abruptly changes direction. The objective definition of "peak" and "valley" used in the rest of the Chapter involves $R(t)$ through its reciprocal $\kappa(t)$:

Definition 4.2 (Peak and valley)
When $|\kappa(t)|$ has a local maximum in $t \in (0, 1)$, t is either a peak or a valley.

As $\kappa(t)$ is signed, one has $\kappa(t) > 0$ (equivalently, $R(t) > 0$) if the tangent vector to the curve rotates counterclockwise when t increases.
In stroke terms (for non-intersecting strokes going from the left to the right), one has a peak if $\kappa(t) < 0$, or a valley if $\kappa(t) > 0$. Moreover, the behavior of $R(t)$ near any

minimum point enables us to characterize peaks and valleys according to their shape. The definition of the curvature is graphically clarified in Figure 4.2.

If $\kappa(t) = 0$, the stroke is straight at time t, and a corresponding definition could be given to identify these points. However, experiments show that it is difficult to precisely locate the straight points, as they are sensitive to small perturbations of the curve (such as those stemming from the spline fitting procedure). Also, the peaks and valley appear to be more prominent to the eye rather than the location of straight sections. For these reasons, we will only consider peaks and valleys in this Chapter.

For technical reasons that will be explained in the next Section, it is forbidden to have peaks and valleys at the beginning ($t = 0$) or at the end ($t = 1$) of the stroke. Intuitively, the curvature is not defined at the stroke endpoints. This is not a significant limitation, as all signature peaks and valleys of interest occur inside the stroke.

Curvature maximization

According to the definition of peak and valley (Definition 4.2), we are interested in locating the points t where $|\kappa(t)|$ reaches a local maximum. The definition of $\kappa(t)$ requires the computation of the first and second derivatives of the quantities $x(t)$ and

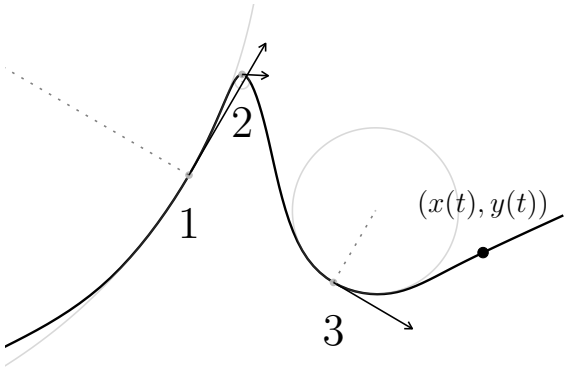


Figure 4.2: A 2D curve (black, thick) with three example points, along with the tangent vectors (i.e. $(x'(t), y'(t))$), the tangent circles (gray), and the radiuses of those circles (gray, dotted). The curvature is the reciprocal of their lengths. It is evident how broader curves imply a longer radius of curvature, thus a smaller curvature $\kappa(t)$.

$y(t)$.

$x(t)$ and $y(t)$ are obtained by fitting a piecewise spline curve over a relatively small number of pixels. This process might introduce spurious turns at the interconnections between these splines. Another source of error might be the slight inconsistencies brought in by the manual process of spline fitting, for instance if the source image provides a very low number of data points. For convenience, we refer to these sets of perturbations of $x(t)$ and $y(t)$ with the term "noise".

This noise is almost invisible to the naked eye, but its presence is greatly amplified in a naively computed $\kappa(t)$, severely affecting the identification process of peaks and valleys.

A very useful technique to compute $\kappa(t)$ whilst being robust to the noise involves a Gaussian smoothing of $x(t)$ and $y(t)$ prior to the curvature computation (Lowe, 1989):

$$\begin{aligned}\tilde{x}(t) &:= (G_\sigma \circledast x)(t)\\ \tilde{y}(t) &:= (G_\sigma \circledast y)(t)\,,\end{aligned}$$

where $\circledast$ is the convolution operator:

$$(f \circledast g)(t) := \int_{-\infty}^{\infty} f(\tau) g(t-\tau)\,\mathrm{d}\tau\,,$$

and G_σ is the probability density function of a Gaussian random variable with 0 mean and standard deviation σ.

Then, one computes $\kappa(t)$ using the smoothed $\tilde{x}(t)$ and $\tilde{y}(t)$. It can be shown that the local minimum and maximum points of the smoothed $\kappa(t)$ match those in the original $\kappa(t)$. In stroke terms, peaks (valleys) appearing in the smoothed stroke will also appear in the original stroke, and at the same "time" value t (provided that σ is small). The converse is not true, as the spurious peaks (valleys) are removed in the smoothed stroke.

The smoothing of $(x(t), y(t))$, however, alters slightly the shape of the curve as well as the values for $\kappa(t)$ and $R(t)$. For instance, if a curve $\gamma : t \mapsto (x(t), y(t))$ is a circle, the smoothed curve $\tilde{\gamma} : t \mapsto (\tilde{x}(t), \tilde{y}(t))$ is a circle with a smaller radius. The distortion effect can be corrected, but the locations of the minima and maxima are not affected (Lowe, 1989). Figure 4.3 shows a plane curve along with its smoothed version: the distortion effect "shrinks" the curve towards the centers of the turns, but the maximal (minimal) curvature points are reached at the same "time" as the original curve.

The standard deviation of the Gaussian filter σ is chosen by a grid-search and a visual comparison between the original stroke and its smoothed version, seeking to

Figure 4.3: A 2D curve (in black) along with its smoothed versions (in gray) for increasing values of the smoothing factor σ. It is appreciable how the smoothing affects the shape of the signature, in particular by reducing the total length.

identify the most prominent peaks and valleys. It can be shown that the procedure is not particularly sensitive to the choice of σ, once all spurious peaks are removed.

Once $\kappa(t)$ is computed, one can then easily proceed to find its local extrema points, i.e. the positions of the peaks and valleys. The process, applied to the signature in Figure 4.1, is represented in Figure 4.4.

Curvature absolute value

Although peaks and valleys can be precisely located with this method, Figure 4.4 also suggests that the absolute value of $\kappa(t)$, mathematically related to the radius of curvature, seems to be weakly correlated with the shape of the peak (or valley). In fact, the curvature $\kappa(t)$ does not correctly behave near turn points that become sharper and sharper since the radius of curvature tends to 0, or, equivalently, $|\kappa(t)|$ tends to $+\infty$. Rigorously, the curvature $\kappa(t)$ is not defined in corners, as the coordinate functions $x(t)$ and $y(t)$ are not differentiable.

Operatively, those functions are approximated by discretizing over an equispaced grid of "time" points (e.g. $(x_i, y_i) = (x(t_i), y(t_i))$ for $i = 1, \ldots, n$), so that the derivatives of $x(t)$ and $y(t)$ can be numerically computed. However, this approximation greatly amplifies the noise contributions introduced by the spline fitting process. If the approximation improves (e.g. the number of points n is increased), it is expected that $\kappa(t)$ explodes in sharp turns.

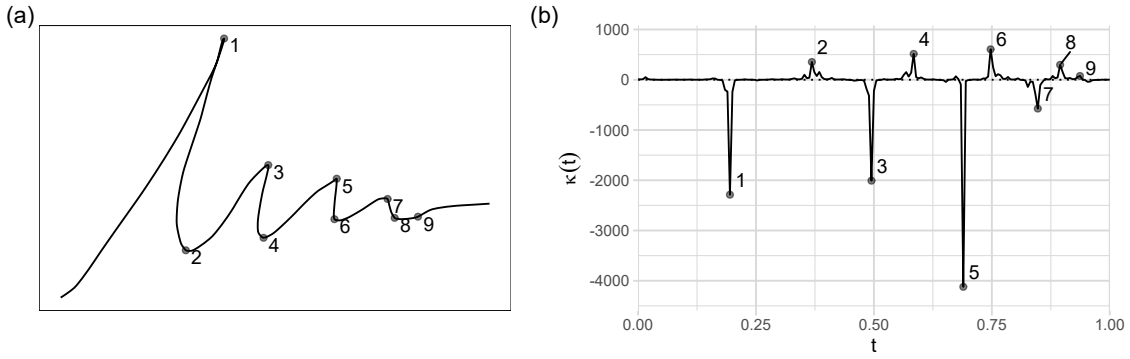


Figure 4.4: (a) The main stroke of the questioned signature of Figure 4.1; (b) Values of the curvature as a function of "time" t. 9 local maxima and minima are highlighted, corresponding to the peaks and valleys.

Notice that the locations of the peaks and valleys (the quantities of interest in this Section) are unaffected, as the curvature $\kappa(t)$ is well-defined outside of the critical points.

4.2.2 Arc-length parametrization

It is possible to compute the stroke length $L(t)$ from the start to time t by integration:

$$L(t) = \int_0^t \sqrt{x'(s)^2 + y'(s)^2}\, \mathrm{d}s\,. \tag{4.2}$$

$L(1)$, thus, is the total length of the main stroke.

Note that to compute $L(t)$ it is required to use the unsmoothed $x(t)$ and $y(t)$, since the total length of the curve may significantly change, as already shown in Figure 4.3. Figure 4.5 illustrates how $L(t)$ is measured on a signature, from the beginning to a generic point at "time" $\tilde{t} = 0.25$.

4.2.3 Peak and valley dataset

One non-trivial feature that can be compared between signatures in the dataset is the rectified distance between the start of the main stroke, and the point where a specific peak or a valley is located. Without loss of generality, such points will be named **points of interest**.

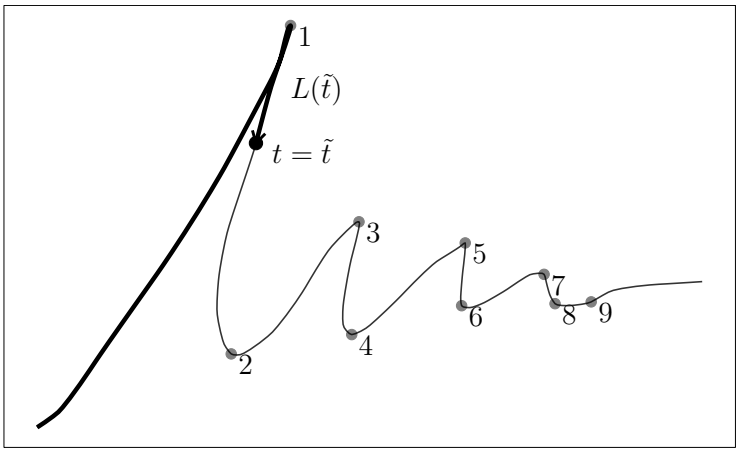


Figure 4.5: How the signature length $L(\tilde{t})$ is measured by tracing the curve from the start ($t = 0$) to the point $\tilde{t} = 0.25$.

To be able to compare these quantities across signatures with varying lengths, these distances need to be normalized by the total length of the signature. The previous sections introduced the tools to accomplish this procedure.

Let us consider the main stroke of the i-th signature, with total length L_i (given by Equation (4.2)).

Let t_0 be the "time" coordinate of a point of interest, for instance where $|\kappa(t_0)|$ reaches a local maximum. According to Definition 4.2, t_0 is either a peak or a valley. The i-th signature will have n_i of such points: let us t_{ik} indicate the k-th point, with $k \in \{1, \ldots, n_i\}$.

From the Equation (4.2), the distance from the beginning of the stroke to t_{ik} is $l_{ik} = L(t_{ik})$.

One may want to compare l_{ik} across all signatures in the dataset. As signatures will have different total lengths, we define the normalized distance of the k-th point in the i-th signature:

$$s_{ik} = \frac{l_{ik}}{L_i} \in (0, 1)\,.$$

The i-th signature will be described and quantified by the set of features $\{s_{ik}\}_{k=1}^{n_i}$. Notice that the length of this vector varies with i (the signature).

4.2.4 Peak and valley count

The number of detected peaks and valleys may alone be considered as a feature. In Figure 4.6 we show how their count varies in the dataset. In particular, the signatures in the reference dataset appear to have a greater number of peaks and valleys, on average. Also, both distributions appear to be bimodal.

To evaluate this information according to the Bayesian framework (Chapter 2), one would require to establish two statistical models for the evidence, one per hypothesis. As previously said, these models must be defensible, for instance by appealing to a physical law or to expert knowledge. In this case, the bar plots of Figure 4.6 do not suggest any known statistical distribution for the number of peaks and valleys, preventing us to define the Bayes factor.

4.2.5 Delay parametrization

Let us consider the i-th signature, with n_i peaks and valleys. The k-th point of interest of the i-th signature occurs at position s_{ik}, where $k \in \{1, \ldots, n_i\}$. It holds that $s_{ik} \in (0, 1)$, but it is difficult to describe the relations between two points of the same curve, i.e. s_{ik} and another s_{ij}.

Instead of measuring its position from the beginning of the stroke, we measure the rectified distance from the previous point. In other terms, the (distance) **delay** between one point of interest, and the previous. We also add one entry to the vector, the distance from the last point to the end.

The new measure will be indicated as t_{ik}. It can be defined as follows:

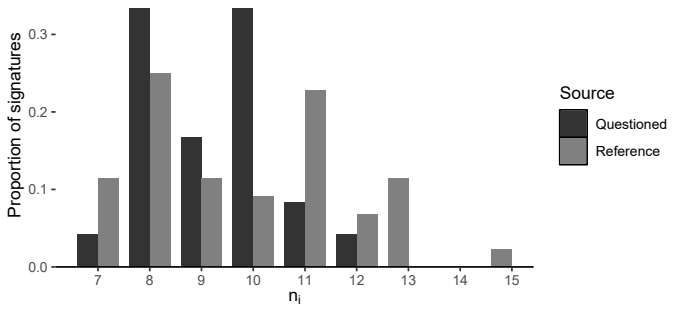


Figure 4.6: Distributions of the number of peaks and valleys in the dataset. On the x-axis, the number of peaks and valleys of a signature. On the y-axis, the proportion of signatures with a given number of peaks and valleys, respective to the source.

$$t_{ik} = \begin{cases} s_{ik} & \text{if } k = 1 \\ s_{ik} - s_{i,k-1} & \text{if } k = 2, \dots, n_i \\ 1 - s_{i,n_i} & \text{if } k = n_i + 1 \end{cases} \tag{4.3}$$

Notice that now k spans from 1 to $n_i + 1$ (although the last term is determined by the previous n_i terms).
The i-th signature is described by the feature vector $(t_{ik})_{k=1}^{n_i+1}$.

The following properties hold:

1. $t_{ik} \in (0, 1) \quad \forall i, \forall k$

2. $\sum_{k=1}^{n_i+1} t_{ik} = 1 \quad \forall i, \forall k$

The set of $(n_i + 1)$-dimensional vectors satisfying these properties is called a (n_i)-**simplex**. As their entries must sum to 1 (property 2), the (n_i)-simplex is contained into $\mathbb{R}^{n_i}$.

Graphical display

The previous Figure 4.5 can be modified to illustrate the current geometric parametrization. This is shown in Figure 4.7: only the first two points of interest are shown. Notice also that, compared to the previous Figure, all measurements are now relative to the length of the full stroke.

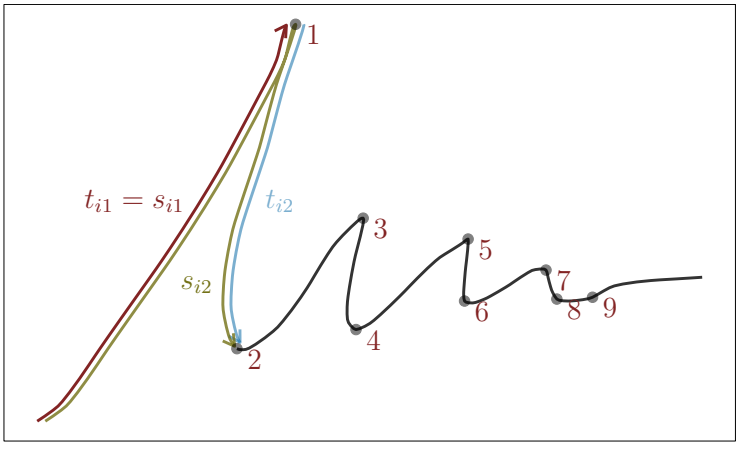


Figure 4.7: The delay parametrization applied to the Figure 4.5.

4.3 Statistical model

In this section we introduce a model for the peak and valley positions starting from the peak and valley dataset (Section 4.2.3). It is first necessary to introduce some notation and definitions.

4.3.1 The Dirichlet distribution

First, the peak and valley positions are represented using the delay parametrization (Section 4.2.5): t_{ik} is the normalized distance in the i-th signature between the k-th point of interest and the previous one. This characterization has multiple consequences:

1. t_{ik} indicates how much "space" is taken up by the k-component compared to all $n_i + 1$ entries.
2. $(t_{ik})_{k=1}^{n_i+1}$ is a partitioning of the segment $(0, 1)$ into $n_i + 1$ non-overlapping intervals.
3. $(t_{ik})_{k=1}^{n_i+1}$ are probabilities: t_{ik} is the probability that a discrete random variable with $n_i + 1$ possible values assumes the k-th value.

If the signature had only one peak ($n_i = 1$), the resulting feature vector would be entirely characterized by the single parameter (t_{i1}, t_{i2}). A common distribution that is often chosen to model random variables with the same constraints as t_{i1} is the Beta distribution. Its two parameters allow for a wide range of shapes as well as bimodality whilst being mathematically tractable.

The natural generalization of the Beta distribution to any number of points n_i is the Dirichlet distribution. Firstly, it reduces to the Beta distribution when $n_i = 1$. Secondly, its allows to model a wide range of distributions on the (n_i)-simplex, from the uniform one (where no t_{ik} has more weight than the other components) to ones which are heavily concentrated (where one of the t_{ik} dominates the others). Lastly, it is mathematically tractable for Bayesian purposes, as it is the conjugate prior of the multinomial distribution (although this fact is not exploited in this Chapter).

The Dirichlet distribution is formally defined in Definition 4.3:

Definition 4.3 (Dirichlet distribution)

Let $\alpha = (\alpha_k)_{k=1}^{n+1} \in \mathbb{R}^{n+1}$, with $\alpha_k > 0$. Let $X = (X_k)_{k=1}^{n+1}$ be a random variable in the n-simplex. X has the Dirichlet distribution ($X \sim Dir(\alpha)$) if it has density:

$$p(X = x) = \frac{1}{\beta(\alpha)} \prod_{k=1}^{n+1} x_k^{\alpha_k - 1} ,$$

where $\beta(\cdot)$ is the Beta function (Jackman, 2009).

It is now useful to introduce a few mathematical properties of the Dirichlet distribution, as they will be exploited later within the casework model in Section 4.3.2.

Properties

The vector parameter α can be split in a scalar component $\alpha^{(0)} > 0$ (the **concentration parameter**) and a vector parameter $\nu = (\nu_k)_{k=1}^{n+1}$ (the **base measure**):

$$\alpha = \alpha^{(0)} \nu ,$$

where $\alpha^{(0)} = \sum_{k=1}^{n+1} \alpha_k$. As $\nu = \frac{\alpha}{\alpha^{(0)}}$, ν belongs to the same n-simplex as X.
Two useful properties are:

$$\begin{aligned} \mathbb{E}[X_j] &= \frac{\alpha_j}{\sum_{k=1}^{n+1} \alpha_k} = \nu_j \\ \mathrm{Var}[X_j] &= \frac{1}{\alpha^{(0)} + 1} \nu_j (1 - \nu_j) . \end{aligned} \tag{4.4}$$

ν_k is thus the expected value for the k-th Dirichlet component. The variance of each component can be adjusted according to $\alpha^{(0)}$, but "certain" or "impossible" components (i.e. those whose ν_k is close to 1 or 0, respectively) are unchanged. Moreover, if all α_k are equal to 1 (i.e. ν is uniform and $\alpha^{(0)} = n + 1$), the Dirichlet distribution is uniform on the n-simplex. From the modeling point of view, $\alpha^{(0)}$ has an attractive or repulsive effect on the samples toward ν, representing the average composition of X. These mechanisms are shown in Figure 4.8 when $n = 2$: in that case, the 2-simplex is the filled triangle with vertices $(0, 0, 1)$, $(0, 1, 0)$, $(1, 0, 0)$.

These facts as well as the above decomposition allow us to consider any available information on peak and valley distances during the prior elicitation procedure.

4.3.2 The Dirichlet model

Given the definition of the Dirichlet distribution (Definition 4.3), it is possible to specify a model for the peak and valley distances. The approach follows the one

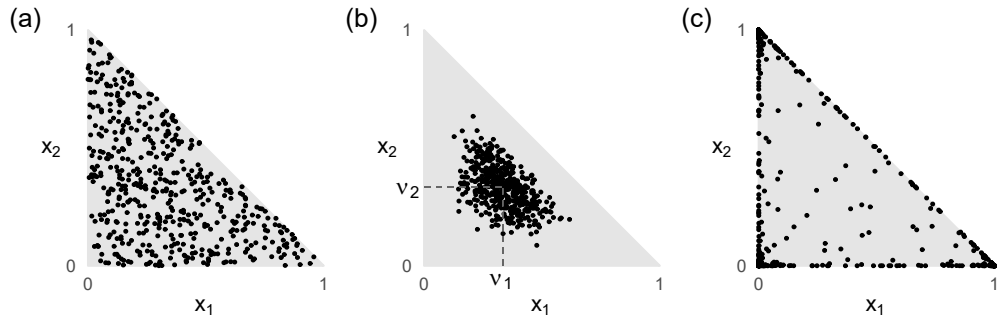


Figure 4.8: Samples from the Dirichlet distribution over its support, the 2-simplex (the shaded area), for various values of the parameters $\alpha^{(0)}$ and $\nu = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Each point is the realization of a possible repartition of the interval $(0, 1)$ in three parts of lengths x_1, x_2 and $x_3 = 1 - x_1 - x_2$ (in our case, the distances between points of interest in the signatures). As x_3 is known given x_1 and x_2, only the first two components x_1 and x_2 are shown. (a) $\alpha^{(0)} = 3 = n + 1$: uniform sampling, all partitions are equally likely; (b) $\alpha^{(0)} = 30$: sampling concentrates around ν; (c) $\alpha^{(0)} = 0.3$: sampling concentrates on the borders, where one of the components vanishes.

theoretically introduced in Section 2.5, further adapted in Section 3.3.1 to the Fourier loop shape descriptors.

Consider the j-th signature with n_j points of interest, described as before (Section 4.2.5) using the elapsed distances t_{jk}, with $k \in \{1, \dots, n_j + 1\}$. To shorten the notation, all points for the j-th signature are collected in the vector $d_j = (t_{jk})_{k=1}^{n_j+1}$.

By construction, the vectors d_j share the same properties as the realizations from the Dirichlet distribution (4.3.1). Similarly to the model for character loops, the distribution of the d_j will depend on the writer through the Dirichlet parameter α. By indicating with d_{ij} the feature vector of the j-th signature written by the i-th writer, one would have:

$$d_{ij} \sim \text{Dirichlet}(\alpha_i)\,.$$

Each α_i might be assumed known for each writer, for instance from expert knowledge.

Alternatively, we suppose to have a background dataset, constituted by m writers, to obtain information on the α_is. Assuming that the peak and valley features exhibit variability between writers, a between-writer level could be added:

$$\alpha_i \sim g(\psi)\,,$$

where g is a density function that describes the between-writer variability, parametrized by ψ.

The Dirichlet-Dirichlet-Gamma model

Choices for g and ψ can take advantage of the decomposition of α_i in a measure component ν_i and a scalar concentration parameter $\alpha_i^{(0)}$ (see Section 4.3.1).
In this thesis we propose to model the components separately, with a Dirichlet and a Gamma distribution respectively:

$$\begin{aligned} \alpha_i &= \alpha_i^{(0)} \nu_i \\ \alpha_i^{(0)} &\sim \text{Gamma}(\alpha_0, \beta_0) \\ \nu_i &\sim \text{Dirichlet}(\nu_0)\,. \end{aligned}$$

The Dirichlet distribution for the within-writer parameters ν_i sets the "average" proportion of peak and valley distances around ν_0, and the Gamma distribution for the within-writer concentration parameters $\alpha_i^{(0)}$ allows us to tune the between-writer variability around the common mean ν_0. The Gamma distribution has been chosen since it has a positive support, and can represent a wide variety of prior beliefs using only two scalar parameters. With this parametrization, ψ is the vector $(\nu_0, \alpha_0, \beta_0)$, and g is its joint probability density function.

The **Dirichlet-Dirichlet-Gamma** model is the following:

$$\begin{aligned} d_{ij} &\sim \text{Dirichlet}(\alpha_i) \\ \alpha_i &= \alpha_i^{(0)} \nu_i \\ \alpha_i^{(0)} &\sim \text{Gamma}(\alpha_0, \beta_0) \\ \nu_i &\sim \text{Dirichlet}(\nu_0)\,. \end{aligned} \tag{4.5}$$

To resume the evaluative inferential scheme described in Section 2.5, evidence e from the i-th writer and j-th signature is represented by the random vector d_{ij}. f is the density of a Dirichlet distribution, parametrized by α_i in the i-th writer (previously indicated with θ in Section 2.5). The between-writer hierarchical level models the distribution of the within-writer parameters α_i through the density function g, parametrized by ψ.

Variability in number of points

One element that further complicates this model with respect to the one introduced, e.g. for the Fourier descriptors in Section 3.3.1, lies in the fact that signatures differ in

the number of points of interest n_j, both within and across the writers. Accordingly, the feature vectors d_{ij} would vary in dimension (length) according to i and j. Since the parameter of the Dirichlet distribution has the same dimension as the data vector, for a given writer i one would have multiple within-writer parameters, say α_{il}, describing the distributions of all signatures from writer i with l points of interest.

This element of variability could raise issues concerning the treatment of "missing data": for instance, in cases when a certain writer never provided signatures with a certain number of points $\tilde{l}$, yet the questioned material contains signatures with $\tilde{l}$ points.

Concerning the modeling approach, one may wish to describe the dependence of all α_{il} by referring to a common α_i. However, the task is not trivial as the addition of a new point of interest could influence the distribution of the remaining points.

Some situations could support to pad all α_{il} with zeroes, up to a common length. In this case, this would be consistent with the idea of adding overlapping points of interest at the end of signature, so that their rectified distances are 0. We believe that this is an inappropriate extension, as it is not possible to have overlapping points according to our approach (see Section 4.2).

The simplest workaround is to consider only signatures with the same number of points. This reduces the amount of data that is used to update the beliefs on the random variables, but also decreases the number of latent random variables, and the number of equations to describe their distributional links.

This situation does not happen in the general compositional case when the length of the feature vector is constant for all observations and all sources. For instance, assume that the vectors d_{ij} represent the repeated measurements of the composition of a given substance across sources i in terms of a set of elements. Their lengths would be constant if the set of searched elements does not change across all observations, and the instrument is always reporting their measured quantity, even if under the limit of detection.

Background parameter elicitation

If a background dataset is available, one can apply the plug-in approximation (Section 2.5.3) to elicit values for ψ, as done in Section 3.3.1.

The simple signature dataset does not contain a sufficient number of writers. We fix $\psi = \widehat{\psi}$ using a data-driven approach, based on stochastic simulation. More details are given in Section 4.5.1 when the practical implementation is discussed.

4.3.3 Evaluative scenario

The same evaluative scenario and the evaluative hypotheses of Chapter 3 can be introduced. The reference material, provided by the putative writer, is represented by a feature vector indicated with e_{r}. In particular, e_{r} is a collection of d_j coming from the reference writer. Without loss of generality, we consider signatures with the same number of points of interest.

The evaluative hypotheses of interest are:

$H = h_p$: "the signatures e_{r} and e_{q} come from the same writer"
$H = h_d$: "the signatures e_{r} and e_{q} come from two different writers"

Under h_p, e_{r} and e_{q} are samples from the same source, supposed parametrized with α_{rq}.
Under h_d, e_{r} and e_{q} are two independent random vectors: the source for e_{r} is parametrized by the α_{r}, the Dirichlet parameter for the reference source. The source for e_{q} is parametrized by the α_{q}, the Dirichlet parameter for the questioned source.

Source parameters are further modeled as being sampled from the between-writer distribution g, parametrized with ψ. The Dirichlet-Dirichlet-Gamma model provides definitions for g and ψ (Equation (4.5)).

The Bayes factor value can be computed using Equation (2.10) as a ratio of marginal likelihoods:

$$\mathrm{BF} = \frac{m(e_{\mathrm{r}}, e_{\mathrm{q}} \mid h_p)}{m(e_{\mathrm{r}} \mid h_d)\, m(e_{\mathrm{q}} \mid h_d)}\,, \tag{4.6}$$

where each $m(e_\bullet \mid h_\bullet)$ involves the integration over the latent parameters $\alpha_\bullet$ as in (2.11), conditioned on the between-writer parameters ψ, supposed known and assuming value $\widehat{\psi}$:

$$m(e_\bullet \mid h_\bullet) = \int f(e_\bullet; \alpha_\bullet)\, g(\alpha_\bullet; \widehat{\psi})\, \mathrm{d}\alpha_\bullet\,. \tag{4.7}$$

4.3.4 Bayes factor computation

The Bayes factor value can be obtained by computing the required marginal likelihoods, as done in Chapter 3 (see Section 3.3.3).

Similarly to the model for the character loops, the marginal likelihood distributions cannot be computed in closed form, so a numerical method is required.

In this Chapter, however, setting up a Gibbs sampler is unfeasible, as the required full conditionals are difficult or impossible to isolate. Instead, we first describe the model in the probabilistic language Stan (Carpenter et al., 2017) to obtain

posterior samples using a particular Markov Chain Monte Carlo (MCMC) method, the Hamiltonian Monte Carlo (HMC).

Secondly, we use the so-called *bridge sampler* on its output to estimate the value of the marginal likelihood (Gronau et al., 2017). The bridge sampler is based on an identity obtained by rewriting Equation (4.7), introducing a bridge function and a proposal distribution. More information can be found in (Gronau et al., 2017).

4.3.5 Implementation

Stan allows to specify models using a language that is very close to the probabilistic description, reducing the need of doing calculations at hand. Among the advantages of Stan compared to other BUGS languages, Stan enforces automated checks on the mathematical constraints on the sampled parameters (e.g. bounded parameters or simplex constraints) and facilitates the specification of prior and posterior predictive quantities, useful to perform model diagnostics. Moreover, the particular sampler used by Stan is reportedly much more efficient than traditional BUGS languages at exploring the sampling space, requiring less iterations, thus less time to fit (McElreath, 2015, sec. 8.2). The sampler fails destructively if the model is ill-specified, producing the so-called *divergent iterations*: in these cases the user is encouraged to re-parametrize the model. Classical MCMC samplers, instead, keep drawing samples that are, possibly, correlated, and the user is not made aware of the phenomenon unless diagnostic procedures are run.

Among the disadvantages, Stan is unable to sample discrete distributions: this does not pose any problem within this Chapter. Also, the sampling speed is much slower than the optimized Gibbs sampler implemented in Chapter 3, specifically tailored to the problem at hand, representing the trade-off between sampling speed and flexibility in terms of modeling specifications.

Concerning the bridge sampling, it is performed in R by the package `bridgesampling` (Gronau et al., 2017), itself implementing the method as formulated in (Meng & Wong, 1996). Particularly, the package takes an output of any Markov Chain Monte Carlo method (such as Stan) and computes the marginal likelihood value for the supplied data. Very few parameters are needed: the proposal function has been set to the multivariate Normal distribution, and the bridge function is obtained by an iterative method. Furthermore, the package also provide an error estimate for the marginal likelihood value, provided that the posterior samples represent well the posterior distribution of the parameters.

To resume, the computation of one Bayes factor value requires the specification of two statistical models, one for each competing hypothesis. Stan runs the MCMC chains

in the numerator and in the denominator of the Bayes factor, then `bridgesampling` independently computes the respective marginal likelihood values. Finally, the Bayes factor value is obtained by computing their ratio.

As this procedure introduces significant overhead in the model fitting workflow, the R package `rstanBF` (Gaborini, 2020a) has been created within the scope of this thesis to wrap the workflow steps to a set of functions. As in Chapter 3, further benefits include the increased reproducibility of the analyses, the facilitated creation of package documentation, and the establishment of an automated suite of tests.

4.4 Model validation

With the available package, we first proceeded to verify its behavior in a situation where the generating model is known, and a large set of background samples are available (a background-dominant situation: see Section 2.5.5).

For the sake of simplicity, we consider the case where the true generating model is a Dirichlet-Dirichlet:

$$\begin{aligned} d_{ij} &\sim \text{Dirichlet}(\nu_i) \\ \nu_i &\sim \text{Dirichlet}(\nu_0)\,. \end{aligned}$$

The Dirichlet-Dirichlet-Gamma model is more general, as it encompasses the Dirichlet-Dirichlet model when:

$$\alpha_i^{(0)} = 1 \qquad \forall i = 1, \ldots, n\,.$$

This is true when the Gamma distribution is degenerate, i.e. $\alpha_0 = \beta_0$ and $\alpha_0 \to +\infty$, resulting in a distribution with unitary mean and vanishing variance. While recovering the hyperparameters from sample data, we expect to obtain a well-behaved estimate of ν_0, and large estimates of α_0 and β_0.

The simulated background data consists of $n = 100$ samples from $m = 10$ 4-dimensional Dirichlet distributions, representing three different sources, $i \in \{1, \ldots, m\}$. In total, 1000 samples are available. The true ν_0 was set to $(1, 1, 1, 1)$.

We consider $i = 1$ to mark the reference source. We want to evaluate the hypothesis pair:

$H = h_p$: "the samples e_r and e_q come from the source 1",
$H = h_d$: "the samples e_r and e_q come from sources 1 and 2, respectively".

The scenario simulation procedure (Section 2.5.5) is applied twice, to evaluate the behavior of the model in two different scenarios:

1. under h_p, we deal $k_{\mathrm{r}} = 20$ samples from the reference source ($i = 1$) and $k_{\mathrm{q}} = 20$ samples from the reference source ($i = 1$),
2. under h_d, we deal $k_{\mathrm{r}} = 20$ samples from the reference source ($i = 1$) and $k_{\mathrm{q}} = 20$ samples from the questioned source (we pick $i = 2$)

The remaining 960 samples constitute the background dataset, which is used to elicit the hyperparameters using the plug-in estimation procedure.

4.4.1 Hyperparameter elicitation

This is done in three steps:

1. for each source i we estimate the Dirichlet parameters $\alpha_i = \alpha_i^0 \nu_i$ using all observed d_{ij}. We indicate the point estimate with $\hat{\alpha}_i$:

$$\hat{\alpha}_i = f_{\mathrm{MLE}}(d_{ij}),$$

2. for each source i we normalize the obtained $\hat{\alpha}_i$, allowing the estimation of the base measure ν_i and the concentration parameter α_i^0:

$$\hat{\alpha}_i = \hat{\alpha_i^0} \hat{\nu}_i,$$

3. finally, the hyperparameters ν_0, α_0 and β_0 are estimated by applying the respective MLE:

$$\hat{\nu_0} = f_{\mathrm{MLE}}(\hat{\nu_i})$$
$$(\hat{\alpha_0}, \hat{\beta}_0) = g_{\mathrm{MLE}}(\hat{\alpha_i^0}),$$

where $f_{\mathrm{MLE}}(x)$ gives the MLE for the parameter of the Dirichlet distribution given observations x, and $g_{\mathrm{MLE}}(x)$ gives the MLE for both parameters of a Gamma distribution given observations x. These are commonly available in the literature as iterative methods, for instance implemented in R by the package `sirt` (Minka, 2000; Robitzsch, 2020) for the Dirichlet MLE, the package `MASS` (Venables & Ripley, 2002) for the Gamma MLE.

In this case we obtain:

$$\hat{\nu_0} = (1.1605139, 1.772515, 1.8833803, 1.9907211)\,.$$

The Gamma hyperparameters are estimated to be very large, as predicted:

$$(\hat{\alpha_0}, \hat{\beta}_0) = (157.4391182, 149.2058713)\,.$$

Notice that this procedure requires a background dataset.

4.4.2 Bayes factors

By plugging in our point estimates for the prior hyperparameters, one can compute the two Bayes factor values for the two scenarios using the `rstanBF` package.

For the Hamiltonian Monte Carlo (HMC) procedure, 10000 iterations are performed, whose 1000 are dedicated to the burn-in process. Six HMC chains are ran in parallel, giving a total of 54000 samples.

For the first scenario (h_p is true), e_r and e_q have been sampled from the same source. The obtained Bayes factor is:

$$\mathrm{BF} = 3.6636522\,,$$

which is greater than 1, as expected.

For the second scenario (h_d is true), e_r and e_q have been sampled from two different sources. The obtained Bayes factor is:

$$\mathrm{BF} = 1.8775933 \times 10^{-8}\,,$$

which is lower than 1, as expected.

Concerning the computational requirements for the computation of a single Bayes factor value, the HMC sampling requires 13 seconds, not including the bridge sampling iterations. This is noticeably longer than the ad-hoc Gibbs sampler shown in Chapter 3.

4.4.3 Convergence

Since the model parameters are known, one can check if the posterior distributions are concentrated near the known values, to make sure that the HMC chains are correctly exploring the sample space. In particular, under the first scenario (h_p is true), the posterior for α_1 should capture the Dirichlet parameter of the reference source $i = 1$. Likewise, under the second scenario (h_d is true), the posterior for α_2 should capture the Dirichlet parameter of the questioned source $i = 2$. As the Figure 4.9 shows, the model seems to be able to capture the generating values.

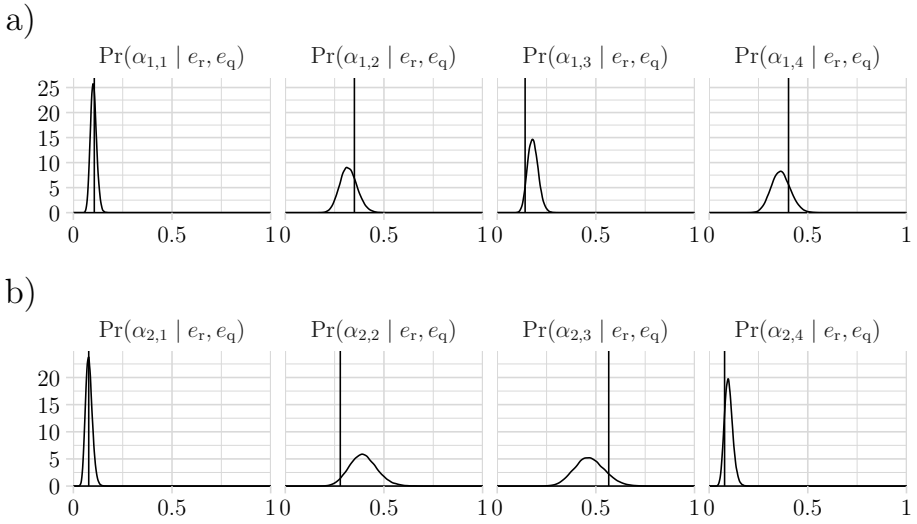


Figure 4.9: Posterior density distributions for the components of α along with their true values (the vertical line): a) h_p is true and the reference source is shown; b) h_d is true and the questioned (different) source is shown. See Section 4.3.3 for details on the notation.

4.5 Results

4.5.1 Evidence-only situation

We can now proceed to analyze the simple signature dataset with the developed package. With respect to the validation scenario, the casework offers no background dataset that could help eliciting the required hyperparameters $(\nu_0, \alpha_0, \beta_0)$. This is an evidence-only operative condition (2.5.5).

As the hyperparameters try to capture the variability of selected handwriting features in a reference population, one could approach this problem by creating a forgery dataset. Multiple forgers should be recruited, then instructed to deliver simulated signatures following the same psychomotor patterns and variability shown by the casework material. For instance, if the questioned signatures have been written in an unusual standing position, the adjoint forgers should try to reproduce the same circumstances, to enable a comparison on a "like-by-like" basis. The validity of this approach ultimately rests on the hypothesis that the recruited forgers are well representing the reference population: this could require a large number of participants in the study. Also, each participant should be made aware of the full-range of reference

signatures, potentially disclosing sensitive information.

Hyperparameters can also be elicited by appealing to forensic literature, related studies or some degree of technical expertise on the collected data. For instance, forensic scientists working with glass fragments may exploit physical knowledge on the range of variation of glass used in common household objects. This is infeasible when one deals with evidence that does not possess a uniform quantitative body of knowledge that is easily transferrable between cases, such as signatures.

A weaker form of elicitation is to exploit the fact that Bayesian models are generative in nature: given any hyperparameter, it is easy to generate synthetic data since the distributions are defined by the probabilistic model. One could identify a range of hyperparameter values that produce data that "look plausible", possibly encompassing the casework material (and more).

Clearly, this task is easier if the statistical model is very close to the casework data, introducing as less transformations as possible. For instance, in the questioned handwriting context, it would be desirable to have a model capable of reconstructing the ink path traced by the signature: the "acceptable" hyperparameters would produce signatures with a behavior comparable to the casework material, that a FHE could examine.

As one goes towards a more elaborated representation to find a suitable statistical description of the evidence, one typically focuses on some aspects of it, discarding content not relevant to the statistical model. For instance, in this Chapter we first extract the main stroke of the signature, discarding the paraph, then the distance between peaks and valleys is considered. Since the shape of the line is no longer predictable, it becomes difficult to assess whether a set of generated distances could appear in the written material. Even if the model were able to generate main strokes, their relative position with the paraph would not be considered, potentially producing strokes that would overlap the paraph.

The decision to "accept" or "reject" a hyperparameter, thus, may ultimately rest on the ability of the forensic scientist to interpret this latter elaborate representation, disregarding most aspects of the evidence that lie on a larger scale. Despite this limitation, this paradigm is rather easy to implement, and could offer a good starting point to begin formulating and discussing a Bayes factor.

4.5.2 Approximate Bayesian Computation

The notion of "acceptable" (hyper)parameters generating "plausible" data can be made more precise, resulting in the family of statistical methods called "Approximate Bayesian Computation" (ABC for short) (Tavare et al., 1997). Particularly, these

methods are used when it is easier to generate data rather than evaluate the likelihood function, for instance by repeating an experiment, or performing simulations.

First, ABC methods sample (hyper)parameters from a distribution or a grid, then generate a dataset conditioning on the sampled (hyper)parameter. Generated datasets are "plausible" when they are *close* to the observed dataset according to some criterium, usually by computing a distance metric between two summary statistics. Those (hyper)parameters that generate "plausible" data are **accepted** and form the posterior samples, as if they were sampled with a traditional MCMC method.

Under appropriate assumptions, ABC methods can deliver good approximations to the posterior distributions (Robert et al., 2011), but their usage for model selection (in particular, the computation of Bayes factors) is controversial (Marin et al., 2014; Robert et al., 2011). ABC methods are commonly used in genetics and population dynamics (Leuenberger & Wegmann, 2010; Pritchard et al., 1999; Tavare et al., 1997; Weiss & Haeseler, 1998), and some early results in forensic science are available (Hendricks et al., 2020).

In the next Section we introduce a method strongly inspired by the ABC approach to elicit the hyperparameters, required to compute the Bayes factor.

4.5.3 ABC in practice

In short, ABC works by generating a dataset (indicated with x^{gen}), then comparing it with the observed one (indicated with $x^{\text{obs}} = (d_{ij})_{ij}$, the set of all observed feature vectors, where i is the index of the writer, j is the signature index for the writer i), and deciding whether x^{gen} is close to x^{obs}.

We assume that there are two independent writers that generate the dataset, consistently with h_d. The hyperparameters $(\nu_0, \alpha_0, \beta_0)$ are set to a value we wish to evaluate its "plausibility". Generation of x^{gen} follows the Dirichlet-Dirichlet-Gamma model (Section 4.3.2), first by choosing the hyperparameters, then by sampling the conditional distributions. Notice that the identifier assigned to the writer (e.g. whether the sample has been provided by the reference writer, the alleged questioned writer, or some other synthetic source) is not considered, as we are interested in evaluating whether the generated dataset could represent *any* background dataset, regardless of the truth of the evaluated hypothesis.

The datasets x^{gen} and x^{obs} can be organized as matrices with the same number of rows (the number of observed feature vectors, here indicated with n) and the same number of columns (here indicated with $p = n_i + 1$, where n_i is the number of points of interests contained in the signature).
By considering only signatures with the same number of points (following the as-

sumption introduced in Section 4.3.2, in order to have feature vectors of the same length p), we have that $n = 28$ and $p = 5$.

According to ABC, all datasets are resumed through a set of summary statistics. The model and the framework do not suggest any summary statistic that is relevant to the task, so we considered a set of measures of central tendency, dispersion and shape:

- s_{mean}: mean
- s_{sd}: standard deviation
- s_{kurtosis}: kurtosis
- s_{skewness}: skewness

In particular, all summary statistics are computed for each column, and are therefore 5-dimensional.

Afterwards, ABC computes the distance between the summary statistics of x^{gen} and x^{obs}: the distance is taken to be the L_1 norm of their difference vectors:

$$\|x - y\|_1 := \sum_{k=1}^{p} |x_k - y_k| \,.$$

Finally, if this distance is lower than a given threshold, the hyperparameters $(\nu_0, \alpha_0, \beta_0)$ are accepted.

To resume, the ABC algorithm proceeds as follows:

1. choose the (hyper)parameters $(\nu_0, \alpha_0, \beta_0)$,
2. compute the observed summary statistics $s^{\text{obs}}_{\text{mean}}, s^{\text{obs}}_{\text{sd}}, s^{\text{obs}}_{\text{kurtosis}}$, and $s^{\text{obs}}_{\text{skewness}}$,
3. generate one dataset x^{gen},
4. compute its summary statistics $s^{\text{gen}}_{\text{mean}}, s^{\text{gen}}_{\text{sd}}, s^{\text{gen}}_{\text{kurtosis}}$, and $s^{\text{gen}}_{\text{skewness}}$,
5. compute the distances between the generated and the observed summary statistics:
   - $d^{\text{gen}}_{\text{mean}} = \|s^{\text{gen}}_{\text{mean}} - s^{\text{obs}}_{\text{mean}}\|_1$
   - $d^{\text{gen}}_{\text{sd}} = \|s^{\text{gen}}_{\text{sd}} - s^{\text{obs}}_{\text{sd}}\|_1$
   - $d^{\text{gen}}_{\text{kurtosis}} = \|s^{\text{gen}}_{\text{kurtosis}} - s^{\text{obs}}_{\text{kurtosis}}\|_1$
   - $d^{\text{gen}}_{\text{skewness}} = \|s^{\text{gen}}_{\text{skewness}} - s^{\text{obs}}_{\text{skewness}}\|_1$,
6. accept the ABC sample if all distances are under the acceptance threshold ϵ:

$$(d^{\text{gen}}_{\text{mean}} < \epsilon) \wedge (d^{\text{gen}}_{\text{sd}} < \epsilon) \wedge (d^{\text{gen}}_{\text{kurtosis}} < \epsilon) \wedge (d^{\text{gen}}_{\text{skewness}} < \epsilon)$$

7. go to step 3 and repeat n_{ABC} times.

The number of times an ABC sample is retained indicates how often the chosen hyperparameters $(\nu_0, \alpha_0, \beta_0)$ generate "plausible" data: in particular, the ratio between this number and n_{ABC} is the so-called **acceptance ratio**. To this purpose, the ABC procedure is repeated by varying $(\nu_0, \alpha_0, \beta_0)$ over a very large space. Particularly, α_0 varies in $[0.01, 10^5]$, β_0 varies in $[0.01, 10^5]$, while ν_0 is set to the value recovered by the ML estimators (as during the model validation, Section 4.4.1).

The acceptance threshold has been set to $\epsilon = (0.1, 0.4, 10, 6)$ by visual inspection: this choice selects datasets whose marginal distributions seem to be similar to the observed ones, yet without being too restrictive. The number of ABC repetitions has been set to $n_{\text{ABC}} = 1000000$.

The "plausible" choice for $(\nu_0, \alpha_0, \beta_0)$, indicated with $(\nu_0^\star, \alpha_0^\star, \beta_0^\star)$, is the one that maximizes the acceptance ratio over the explored combinations of hyperparameters.

Figure 4.10 shows the acceptance ratios obtained over a regular grid of choices for $(\alpha_0^\star, \beta_0^\star)$. In principle, ν_0 should also be allowed to change. However, this revealed to be difficult, as ν_0 is multivariate. Also, when the dimensionality becomes high, ABC might suffer from the so-called "curse of dimensionality", resulting in an extremely low acceptance ratio and in an inefficient sampling (Blum & François, 2010; Hendricks et al., 2020). For instance, the best solution in this case has an acceptance ratio of approximately 2%.

Figure 4.11 shows an example of a synthetic dataset, generated using the hyperparameters that maximize the acceptance ratio. The visual inspection of the shape of the generated distributions was particularly important to define the summary statistics leading to a "correct" acceptance: otherwise it would have been possible to select datasets that are close to x^{obs} according to the summary statistics, yet completely different when observed in their full complexity[1]. For instance, by considering only the mean and the standard deviation, it was possible to generate datasets that would have been accepted under strict values for the tolerance ϵ, but degenerate in their distributions (i.e. where the distribution masses were concentrated in very small regions, unlike the observed data). The solution was found by introducing more summary statistics that would account for the shape of the datasets.

Remark. Compared to the other approaches that can be found in literature, ABC is only used in this Chapter to elicit an informed data-driven choice for the values of the hyperparameters $(\nu_0, \alpha_0, \beta_0)$. Once the choice is made, the computation of the Bayes factor is performed using more traditional methods such as the bridge sampler. ABC

[1]Another famous example of this phenomenon is the Anscombe's quartet, where four datasets are visually very distinct but share the same descriptive statistics (Anscombe, 1973).

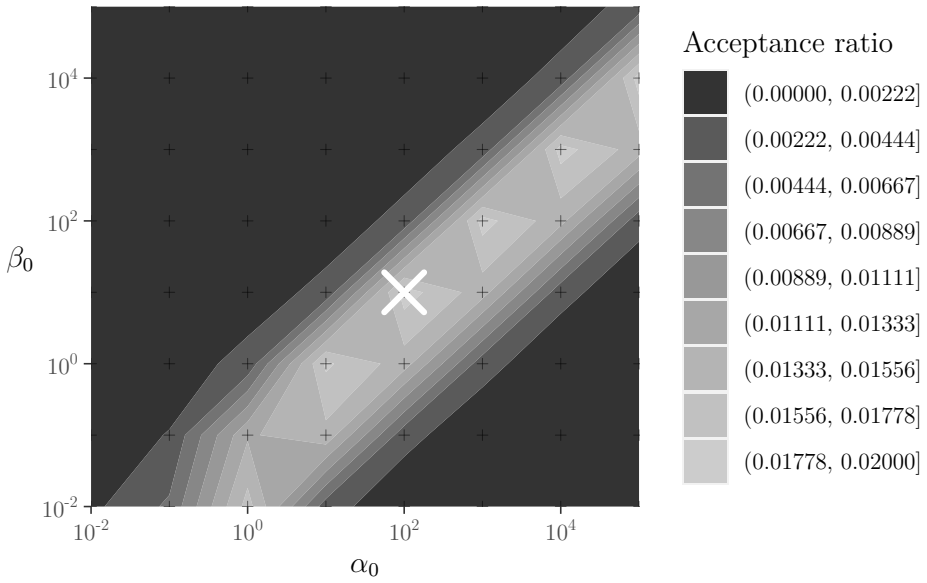


Figure 4.10: The ABC acceptance ratio for different choices (the plusses) of the hyperparameters α_0 and β_0. The white cross marks the accepted combination of $(\alpha_0^\star, \beta_0^\star)$.

methods, instead, would retain all samples of the latent writer parameters (e.g. all α_is) that would generate acceptable datasets. However, given the very low acceptance ratio, this approach would be extremely inefficient, possibly incurring also limitations raised by Robert et al. (2011).

Implementation

The R package `rdirdirgamma` has been created to perform the whole ABC procedure for the specific Dirichlet-Dirichlet-Gamma model, from the dataset generation to the acceptance step. As in Chapter 3.3.4, all code has been optimized for speed using Rcpp (Eddelbuettel & François, 2011). This package is open source and available on request (Gaborini, 2020b).

4.5.4 Bayes factor

Now that the ABC method provides the values for the hyperparameters $(\nu_0^\star, \alpha_0^\star, \beta_0^\star)$, the single Bayes factor for the evaluative scenario (Section 4.3.3) can be computed. This can be performed with a bridge sampler, as in the Model validation procedure (Section 4.4.2).

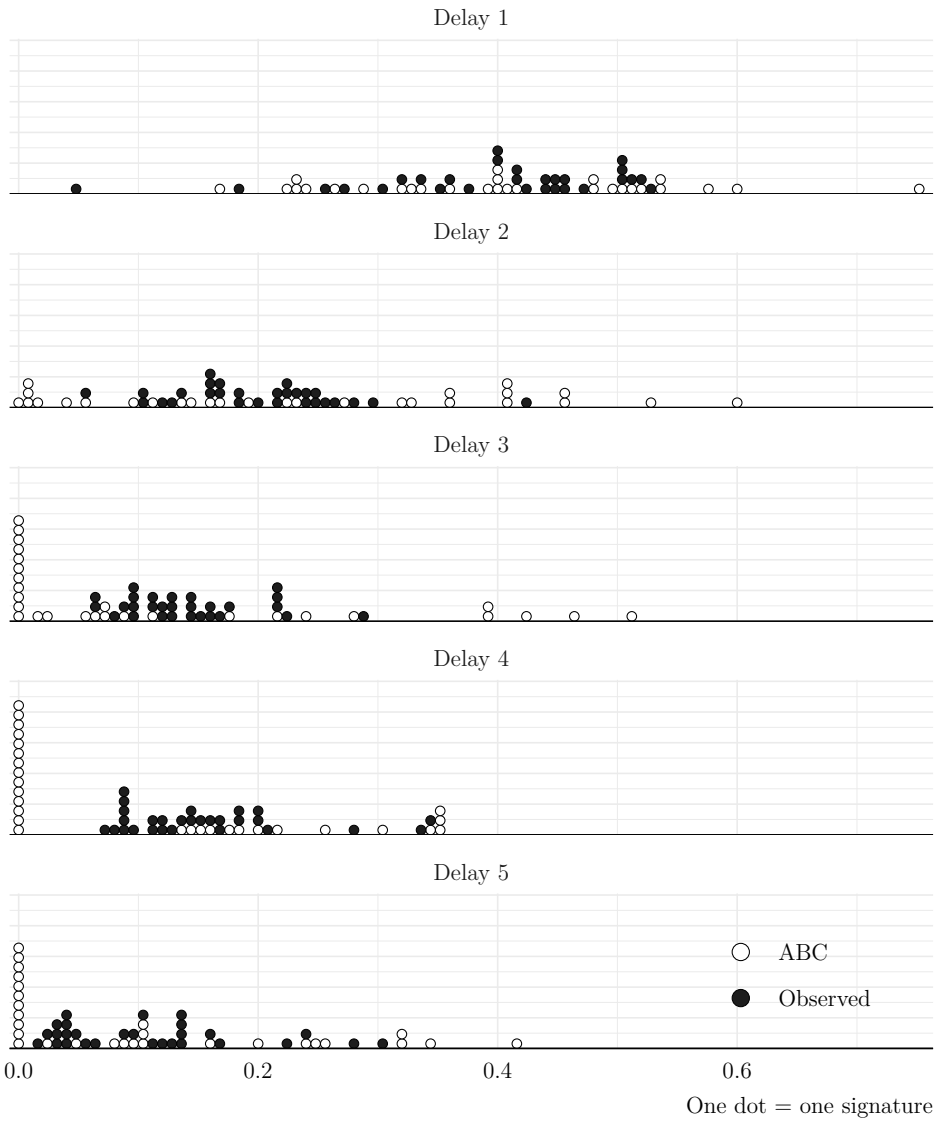


Figure 4.11: Comparison between the observed data x^{obs} and one of the generated datasets x^{gen} that is accepted by the ABC criterium using the best hyperparameter combination $(\nu_0^\star, \alpha_0^\star, \beta_0^\star)$. Each panel is one component of the distribution (a specific point of interest). As each dot represents one feature vector (a signature), the y-axis is not relevant. In this case, the distances between the summary statistics are: $(d^{\text{gen}}_{\text{mean}}, d^{\text{gen}}_{\text{sd}}, d^{\text{gen}}_{\text{kurtosis}}, d^{\text{gen}}_{\text{skewness}}) = (0.092, 0.356, 8.040, 3.022)$.

For the Hamiltonian Monte Carlo (HMC) procedure, 10000 iterations are performed, whose 2000 are dedicated to the burn-in process. Six chains are ran in parallel, giving a total of 48000 HMC samples.

The obtained Bayes factor is:

$$\mathrm{BF} = 1.4454577 \times 10^4 \,,$$

which gives a strong support for the prosecution hypothesis.

4.6 Extensions

This model can be noticeably improved under many aspects.

Firstly, the Dirichlet likelihood can be substituted with any likelihood that has support on the simplexes. In fact, the Dirichlet distribution, despite its nice mathematical properties, has been reportedly been difficult to use to describe compositional data, most notably since it contains a number of independence assumptions between the components (Aitchison, 1982). An alternative modeling has been proposed by Gelman, Bois, et al. (1996), where each component is first separately modeled as a log-Normal distribution, then the whole vector is normalized (Gelman, Bois, et al., 1996).

Secondly, the need to consider signatures with the same number of points of interest is particularly restricting (Section 4.3.2). A prior distribution on the number of points could be introduced. However, the dimensions of the feature vectors and the latent writer parameters would change across MCMC iterations. Notice that this issue is similar to the computation of Bayes factors using a model indicator (see Section 2.6.4), therefore it could share the same solutions such as the introduction of pseudopriors, and the usage of a sampling algorithm that can sample from spaces of varying dimensions, such as the Reversible-Jump Markov chain Monte Carlo sampler (Green, 1995).

Once the model is determined, the method to elicit the hyperparameters could also be improved. As previously said, this could be done by creating a background dataset constituted by request forgeries (see Section 4.5.1), or by refining the ABC approach (see Section 4.5.2).

4.6.1 Non-independence

Like any scenario involving questioned signatures, the assumption of independence between sources under h_d could be undermined. This has already been encountered

and discussed in the Section 3.7.1, and the same conclusions apply to the case at hand, including the over-evaluation of evidence under h_p. However, the approach shown in this Chapter does not attempt to solve the problem.

4.7 Epilogue

The case has been further blindly evaluated by a Forensic Document Examiner using the traditional appproach. The FDE ultimately concluded that the evidence supports the defense hypothesis, as opposed to the Bayes factor computed in Section 4.5. However, the conclusion was based on the relative position of the main stroke and the paraph, whereas in this Chapter only the main stroke was studied.

This fact further stresses the importance of properly assessing evidence using the general-to-specific pattern, and the difficulty of combining evidence coming from multiple sources, items, or levels of detail of the same item.

Chapter 5

Combining evidence

Let us focus on the impact of relatives in evidence evaluation. The most extreme case where two persons are related is when they are mono-zygotic twins. As twins share most of their genetic material, such situation bears great interest in forensic disciplines other than DNA, for instance handwriting (Srihari et al., 2008), speaker recognition (Loakes, 2008), fingerprints (Jain et al., 2002), bitemarks (Sognnaes et al., 1982) and hair analysis (Bisbing & Wolner, 1984).

In this Chapter we consider a hypothetical situation where handwritten and biological evidence (a salivary stain) is recovered from the crime scene or a relevant object (e.g. an envelope), and the defense hypothesis involves a mono-zygotic twin of the person of interest (POI). This case could arise when a POI provides a handwritten ransom note on a sheet of paper, showing many characters with closed loops, and an item with salivary traces of the suspect is seized, constituted for instance by a glass item, a bottle or the envelope that contained the letter. As the DNA is not easily exploitable due to the genetic constraint (Gringras & Chen, 2001), the microbiome (the composition of the salivary bacteria) can be analyzed instead (Leake, 2014; Leake et al., 2016).

The scenario is approached first by treating evidence separately: handwriting evidence is discussed by applying the character loop model from Chapter 3, while biological evidence exploits the same model introduced in Chapter 4 to evaluate compositional data. The results of these separate evaluations are then combined to a single Bayes factor using a Bayesian network.

This Chapter involves experimental data collected in collaboration with the Institute of Microbiology (CHUV) of the University of Lausanne. The analytical methodology of the laboratory has already been satisfactorily conducted and validated in past experimental research (Leake, 2014; Leake et al., 2016). In this Chapter we

propose an evaluative framework that could potentially integrate these results with a Bayesian (forensic) approach.

Remark. Due to major time constraints and the sudden appearance of the SARS-CoV-2 pandemic, only handwritten evidence has been considered in full detail, although on a reduced set of participants and visits. The microbiome data (not available at the time of writing) and the problem of combination of evidence is only introduced as a theoretical device, without relying on any experimental data. Future works could address this part by building over the proof of concept given in this Chapter.

5.1 Research design

The research involved 8 pairs of twins at the time of writing. All twins were proven to be mono-zygotic through DNA analyses. To reduce age and learning factors influence, all twins pairs were at least 18 years old, learned to write in French as their first language, grew up and were living in Switzerland.

All research participants were required to visit the collection infrastructures at times t_0, $t_1 = t_0 + 1$ month, $t_2 = t_0 + 12$ months, $t_3 = t_0 + 13$ months. Data from visits t_2 and t_3 was not available at the time of writing.

Participants were given a form containing 3 lower-case single letters (“a”, “d”, “o”) and 11 lower-case words, containing said letters in various positions inside the words. In each form, every letter was replicated 10 times in isolated form, and every word was replicated 2 times. On the page only a horizontal-printed guide was provided. Participants were told to write in their usual sitting position, and the writing conditions were standardized across all sessions, with a unique blue ball-point pen on a standardized surface.

Concerning the biological evidence, during each visit participants provided salivary samples for microbiotic analyses. The collection procedure and the subsequent analyses were conducted by the Institute of Microbiology (CHUV) of the University of Lausanne. All twins also filled a form containing health-related questions to control for changes in environmental factors, antibiotics, medications and diseases, that were postulated by Leake (2014) to be potential confounders.

5.2 Handwritten evidence

An aspect of importance to consider in large-scale studies on handwriting — as underlined by (Huber & Headrick, 1999) — is the characterization of the relevant population. Among the elements that could contribute to the variability of handwrit-

ing across persons of interest, the genetic factor has been frequently discussed in past literature.

Available studies which involve the handwriting of twins (Ahuja et al., 2018; Dziedzic et al., 2007; Gamble, 1980; Srihari et al., 2008; Thorndike, 1915) conclude that twins can be discriminated from unrelated persons, albeit at a higher error rate. Of these studies, only (Ahuja et al., 2018; Srihari et al., 2008) used a quantitative characterization of handwriting. Moreover, (Srihari et al., 2008) involves a global description of handwriting, rather than the variability of specific features.

5.2.1 Digitalization

All acquired handwritten forms were digitalized at 600 dpi resolution, then pre-processed in MATLAB to separate written content from the paper. After binarization, the extraction of loops from the word skeletons was performed using Topological Data Analysis. This is a mathematical framework that aims to provide a robust description of the structure of a point cloud (represented by the pixels where ink was detected), in particular where "holes" are located (Chazal & Michel, 2017). In our application, holes in the topology represent closed loops in characters. This enabled us to segment multiple loops at once, still under human supervision, but greatly decreasing the need for manual intervention. Results are comparable to those that could have been obtained by manually tracing the loop contours. This was carried out using R package `TDA` (Fasy et al., 2015).

Only closed loops were retained. Also, strokes crossing the loop but not belonging to it were eliminated, to comply with our goal of targeting the analysis on the main loop of the characters only. Table 5.1 reports the material for analysis used for this study.

Next, the character loops were parametrized and modeled using the Fourier descriptors as done in Chapter 3. In particular, only the harmonic contributes from order $k = 1$ to order $k = 3$ were kept, as they were sufficient to characterize loop shapes[1], thus describing each loop with a vector of length $p = 7$. Consequently, Fourier coefficients were rescaled in order to describe loops with unit area.

Notice that additional metadata are available for each character loop, such as its containing character ("a", "d", "o") and its position inside the word (beginning, middle, end, or isolated letter). In this Chapter only the character information has been used. Other characteristics could be exploited in future works to investigate forensically interesting questions such as the relation between the shape of the loop

[1]The number of retained contributes differs from Chapter 3 due to the different handwriting size (not controlled in the printed forms) and the usage of a different flatbed scanner.

Table 5.1: Number of closed character loops across characters and writers.

	Character			
Writer code	a	d	o	Total
1A	41	23	47	111
1B	45	13	60	118
2A	46	32	57	135
2B	41	12	48	101
3A	58	62	59	179
3B	58	56	59	173
4A	59	42	58	159
4B	53	62	57	172
5A	56	53	55	164
5B	57	62	53	172
6A	59	63	58	180
6B	60	63	59	182
7A	59	55	54	168
7B	33	35	58	126
8A	43	48	51	142
8B	56	59	51	166
Total	824	740	884	2448

and its position inside the word.

5.2.2 Dimensional reduction

The Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) technique (McInnes et al., 2018) was first applied to the dataset, to evaluate whether the Fourier descriptors offered information to discriminate writers based on their shape, as supposed in Marquis et al. (2005) and Chapter 3. The same UMAP Python implementation was used as described by the above article. The Euclidean distance was chosen as a distance metric, while the UMAP parameters were set to `min_dist = 0.8` and `n_neighbors = 5` to visually obtain a good trade-off between writer discrimination and preservation of the local structure; this choice bears no relation to the experimental design of the study, such as the number of twins.

The results of applying the UMAP technique to the dataset are reported in Figure

5.1. It can be seen that writers tend to cluster together. In Figure 5.2, the features learned by UMAP discriminate loops based on their shape: however, the loops are not well discriminable across the character they belong to, as Figure 5.3 shows.

To further investigate the genetic influence on handwriting, one can highlight the UMAP representations for each pair of twins in the study, as shown in Figure 5.4. It can be seen that some twins tend to write similarly, while others are set apart by UMAP.

5.2.3 Statistical model

The same general statistical model developed in Section 3.3 was applied to the set of Fourier coefficients. In particular, the collected dataset can be considered to be the set of background observations.

Maintaining the same notation, let $X_{ij} \in \mathbb{R}^p$ denote the set of $p = 7$ Fourier coefficients of the j-th character loop written by the i-th participant. The model for

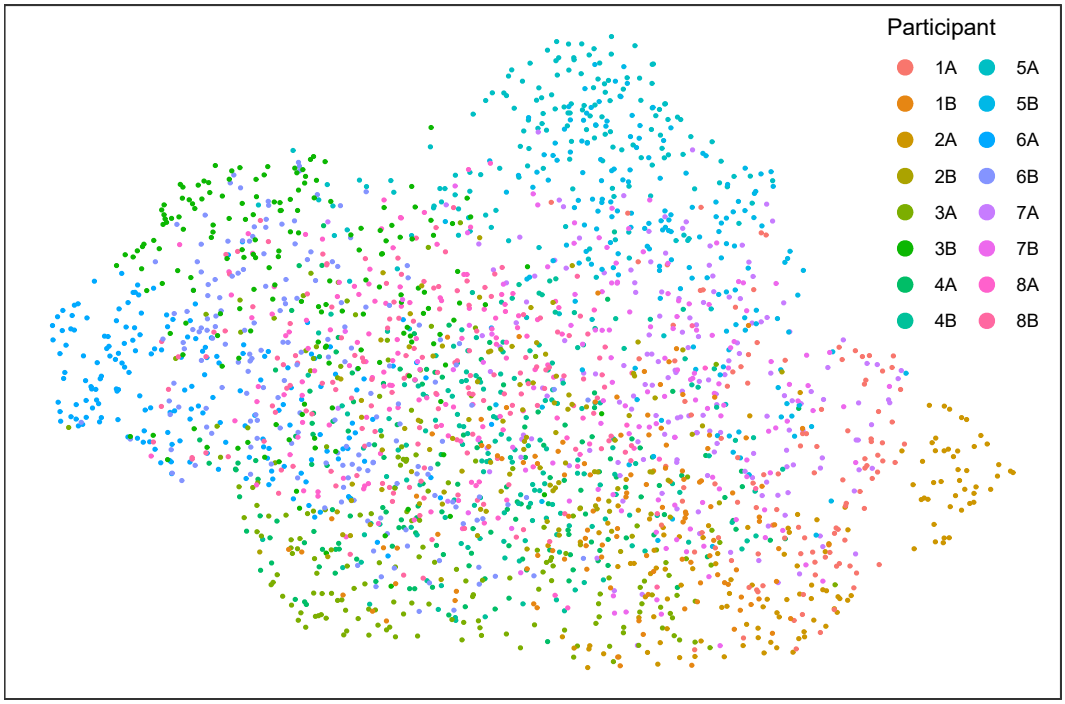


Figure 5.1: UMAP 2D representation of the Fourier twins dataset. Each character is represented as a single point in the 2D plane, colored by participant.

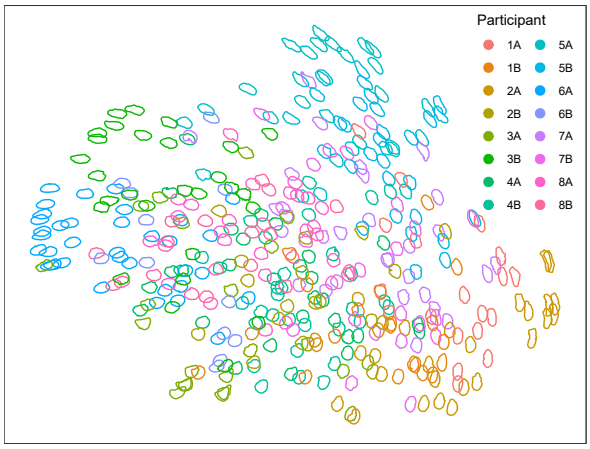


Figure 5.2: Projected dataset on the learned UMAP representation. The original loop shapes are superimposed onto the 2D plane, with the barycenters in their UMAP coordinates, colored by participant. To avoid overplotting, only 400 loops are represented.

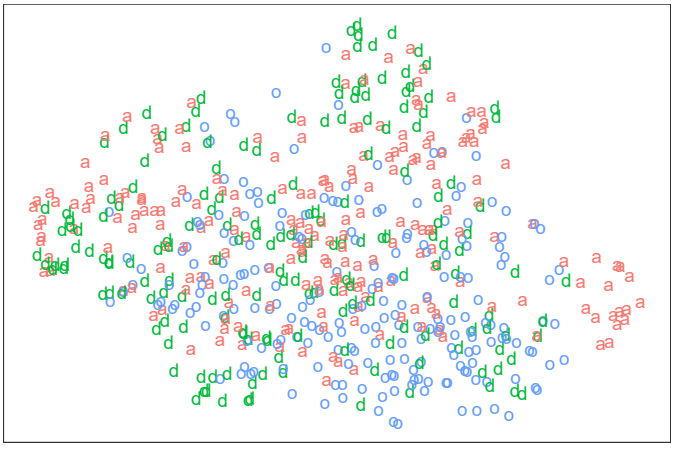


Figure 5.3: UMAP 2D representation of the Fourier twins dataset. Each character is represented as a single letter in the 2D plane, also colored by letter. To avoid overplotting, only 600 loops are represented.

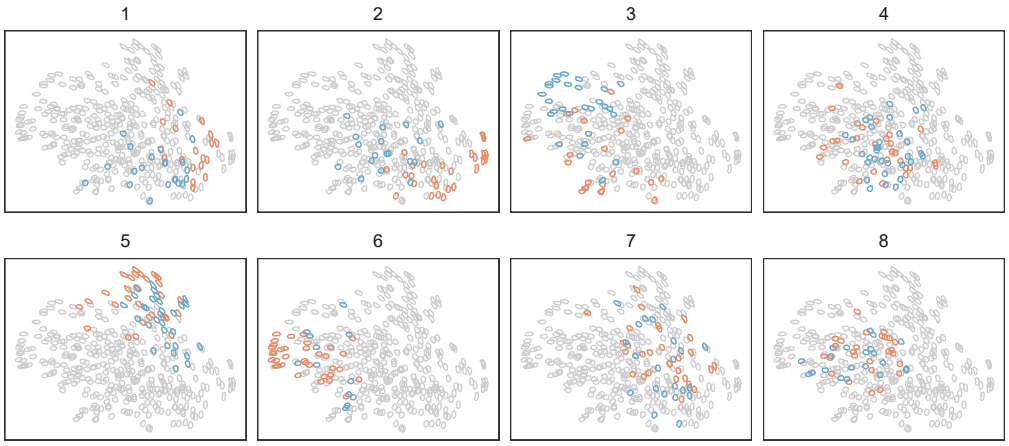


Figure 5.4: Projected dataset on the learned UMAP representation. The original loop shapes are superimposed onto the 2D plane, with the barycenters in their UMAP coordinates. The UMAP representation is replicated for each pair of twins in the dataset, highlighting both participants. To avoid overplotting, only 400 loops are represented.

the background is given by Equation (3.3), repeated here for clarity:

$$\begin{aligned} X_{ij} &\sim \mathcal{N}_p(\theta_i, W_i) \\ \theta_i &\sim \mathcal{N}_p(\mu, B) \\ W_i &\sim IW(U, \nu)\,, \end{aligned}$$

where θ_i is the mean vector of the i-th writer, W_i is the non-constant within-writer covariance matrix, μ is the mean vector between writers and B is the between-writers covariance matrix.

Background parameter elicitation

Since the dataset contains a sufficient number of participants, it is possible to use the plug-in approximation as in Section 3.3.1 to obtain point estimates for μ, B and W_i using the background dataset. Known p, ν is set to be as small as possible as in Section 3.3.1.

5.2.4 Evaluative scenario

As this is a background-dominant operative condition (see Section 2.5.5), the background dataset can be exploited to simulate a real casework. The character loops found in the questioned handwritten material are indicated with e_{q}. The character loops found in the putative writer's corpus are indicated with e_{r}. Notice also that the truth of the hypothesis is known since it is known which participants are twins.

The evaluative scenario introduced in Section 3.3.2 can be applied. However, in this case it is interesting to consider two defense hypotheses, depending on the relation between the suspect and the true writer of the questioned material.

The hypotheses of interest are:

$H = h_p$: "the character loops e_{r} and e_{q} come from the same writer"
$H = h_u$: "the character loops e_{r} and e_{q} come from two unrelated writers"
$H = h_t$: "the character loops e_{r} and e_{q} come from two twins[2]".

As the evaluative framework requires two hypotheses, we will discuss two scenarios:

1. h_p against h_u
2. h_p against h_t

Once the scenario is defined, for convenience, we indicate with h_d the defense hypothesis, chosen among h_u or h_t. Finally, one can compute the respective Bayes factor value for the chosen scenario by applying the scenario simulation procedure, theoretically defined in Section 2.5.5 and implemented on character loop data in Section 3.4.

The procedure was repeated 100 times for each unique combination of reference and questioned writers, thus producing $100 \times 16^2 = 25600$ Bayes factors. The effect of the choice of the sample sizes k_{ref} and k_{quest} was also investigated by sweeping over the range $\{5, 10, 20, 30, 50\}$. The analysis was limited to cases involving $k_{\mathrm{ref}} = k_{\mathrm{quest}} = k$.

5.2.5 Results

Given the large number of available character loops, results are discussed under two perspectives:

- character-independent: all characters ("a", "d", "o") are pooled together,

[2]This abuse of notation includes all situations where the suspect and the true writer share the same DNA: twins, triplets, quadruplets and so on. Notice that among the analyzed participants there were only twin couples.

- character-dependent: all characters ("a", "d", "o") are considered separately during the scenario simulation procedure.

In addition, each perspective can be discussed in turn under the writer-independent and writer-dependent views, defined in Section 3.5.3. These are, respectively, by aggregating the distributions of the Bayes factors over all possible reference writers, or by detailing each reference writer separately.

Character-independent results

In the character-independent view, the obtained Bayes factors are summarized by the hypothesis that has generated the data.

Figure 5.5 reports the "distributions" of the log-Bayes factors across combinations of reference and questioned writers, as the number of samples k grows. At first, the writers relationship is considered. It is expected that the strength of evidence increases with the number of samples under comparison, as the model better approximates writers' patterns. In particular:

1. when comparing material coming from the same writer, it is expected that evidence correctly supports the hypothesis h_p.
2. when comparing material coming from unrelated writers, it is expected that evidence correctly supports the hypothesis h_u.

Notice that these results are in accordance with the ones obtained using an independent corpus of unrelated writers (Section 3.5.3), thus providing a further verification of the model.

The comparison of material from twin pairs results in log-Bayes factor values which are distributed around the neutral value of 0, with a spread increasing with k. This result supports the importance of writer choice: results seem to be writer-dependent. This is consistent with what has been reported in (Bozza et al., 2008).

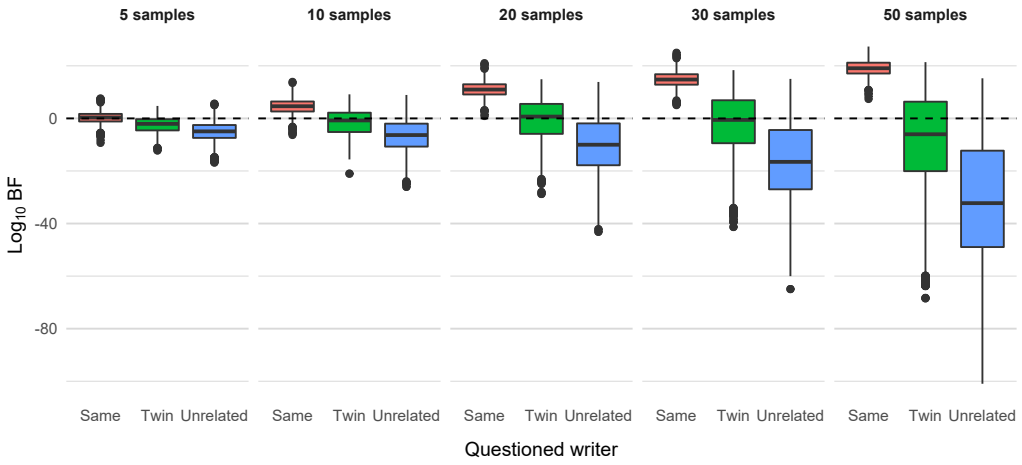


Figure 5.5: log-Bayes factor values across hypotheses and the number of samples k.

Table 5.2: Bayes factor values across all scenarios as a function of the relationship between writers and the number of samples k.

k	Questioned writer	BF < 1	BF > 1	Total
5	Same	694	906	1600
	Twin	1274	326	1600
	Unrelated	20900	1500	22400
10	Same	108	1492	1600
	Twin	886	714	1600
	Unrelated	19050	3350	22400
20	Same	0	1600	1600
	Twin	752	848	1600
	Unrelated	18084	4316	22400
30	Same	0	1600	1600
	Twin	836	764	1600
	Unrelated	18932	3468	22400
50	Same	0	1600	1600
	Twin	980	620	1600
	Unrelated	20662	1738	22400

Table 5.3: Number of times a contradictory Bayes factor value has been obtained across 100 trials, with k samples from each writer combination. In bold, the reference writer and the questioned writers are the same person (h_p is true), "T" marks pairs of twins (h_t is true), in the other entries the participants are unrelated (h_u is true).

k	Reference writer	Questioned writer															
		1A	1B	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B	7A	7B	8A	8B
5	1A	**57**	7^{T}	6	3	0	0	1	1	0	3	0	0	20	10	0	0
	1B	7^{T}	**38**	10	28	22	0	31	30	0	3	0	0	15	12	1	2
	2A	6	10	**43**	1^{T}	0	0	1	1	0	0	0	0	3	2	0	0
	2B	3	28	1^{T}	**33**	8	1	45	47	0	0	1	1	12	10	27	15
	3A	0	22	0	8	**43**	4^{T}	31	20	1	0	1	6	2	2	8	7
	3B	0	0	0	1	4^{T}	**47**	4	1	1	0	17	21	0	0	14	9
	4A	1	31	1	45	31	4	**36**	38^{T}	0	0	8	6	5	5	16	24
	4B	1	30	1	47	20	1	38^{T}	**40**	0	0	1	2	2	15	21	21
	5A	0	0	0	0	1	1	0	0	**40**	15^{T}	0	0	1	3	0	0
	5B	3	3	0	0	0	0	0	0	15^{T}	**45**	0	0	9	10	0	0
	6A	0	0	0	1	1	17	8	1	0	0	**14**	29^{T}	0	0	9	10
	6B	0	0	0	1	6	21	6	2	0	0	29^{T}	**34**	0	0	17	33
	7A	20	15	3	12	2	0	5	2	1	9	0	0	**61**	28^{T}	2	1
	7B	10	12	2	10	2	0	5	15	3	10	0	0	28^{T}	**68**	5	7
	8A	0	1	0	27	8	14	16	21	0	0	9	17	2	5	**53**	41^{T}
	8B	0	2	0	15	7	9	24	21	0	0	10	33	1	7	41^{T}	**42**
10	1A	**5**	13^{T}	28	0	1	0	0	2	0	8	0	0	59	42	0	0
	1B	13^{T}	**4**	4	64	48	0	50	56	0	0	0	0	38	39	4	5
	2A	28	4	**11**	0^{T}	0	0	0	0	0	0	0	0	3	1	0	0
	2B	0	64	0^{T}	**3**	28	0	78	85	0	0	0	1	19	25	36	45
	3A	1	48	0	28	**5**	7^{T}	57	46	0	0	3	11	4	16	23	36
	3B	0	0	0	0	7^{T}	**5**	1	4	3	0	30	46	0	1	46	25
	4A	0	50	0	78	57	1	**5**	78^{T}	0	1	2	4	5	8	42	37
	4B	2	56	0	85	46	4	78^{T}	**6**	0	0	1	3	20	50	59	57
	5A	0	0	0	0	0	3	0	0	**5**	52^{T}	0	0	2	3	1	0
	5B	8	0	0	0	0	0	1	0	52^{T}	**11**	0	0	29	28	0	1
	6A	0	0	0	0	3	30	2	1	0	0	**3**	50^{T}	0	0	20	24
	6B	0	0	0	1	11	46	4	3	0	0	50^{T}	**7**	0	0	53	69
	7A	59	38	3	19	4	0	5	20	2	29	0	0	**12**	68^{T}	1	0
	7B	42	39	1	25	16	1	8	50	3	28	0	0	68^{T}	**11**	16	18
	8A	0	4	0	36	23	46	42	59	1	0	20	53	1	16	**10**	89^{T}
	8B	0	5	0	45	36	25	37	57	0	1	24	69	0	18	89^{T}	**5**

Table 5.3: Number of times a contradictory Bayes factor value has been obtained across 100 trials, with k samples from each writer combination. In bold, the reference writer and the questioned writers are the same person (h_p is true), "T" marks pairs of twins (h_t is true), in the other entries the participants are unrelated (h_u is true).

		Questioned writer															
k	Reference writer	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B	7A	7B	8A	8B
20	1A	**0**	17^{T}	36	0	1	0	0	0	0	5	0	0	90	65	0	0
	1B	17^{T}	**0**	0	81	70	0	65	89	0	0	0	0	49	52	3	2
	2A	36	0	**0**	0^{T}	0	0	0	0	0	0	0	0	3	1	0	0
	2B	0	81	0^{T}	**0**	28	0	99	100	0	0	0	0	18	39	55	51
	3A	1	70	0	28	**0**	5^{T}	91	70	0	0	0	2	1	17	33	49
	3B	0	0	0	0	5^{T}	**0**	2	0	0	0	29	50	0	0	47	42
	4A	0	65	0	99	91	2	**0**	92^{T}	0	0	0	2	2	9	55	67
	4B	0	89	0	100	70	0	92^{T}	**0**	0	0	0	0	24	61	78	81
	5A	0	0	0	0	0	0	0	0	**0**	59^{T}	0	0	4	4	0	0
	5B	5	0	0	0	0	0	0	0	59^{T}	**0**	0	0	62	45	0	0
	6A	0	0	0	0	0	29	0	0	0	0	**0**	54^{T}	0	0	13	21
	6B	0	0	0	0	2	50	2	0	0	0	54^{T}	**0**	0	0	84	86
	7A	90	49	3	18	1	0	2	24	4	62	0	0	**0**	97^{T}	0	0
	7B	65	52	1	39	17	0	9	61	4	45	0	0	97^{T}	**0**	9	16
	8A	0	3	0	55	33	47	55	78	0	0	13	84	0	9	**0**	100^{T}
	8B	0	2	0	51	49	42	67	81	0	0	21	86	0	16	100^{T}	**0**
30	1A	**0**	8^{T}	21	0	0	0	0	0	0	0	0	0	90	61	0	0
	1B	8^{T}	**0**	0	80	62	0	62	80	0	0	0	0	30	43	0	0
	2A	21	0	**0**	0^{T}	0	0	0	0	0	0	0	0	0	0	0	0
	2B	0	80	0^{T}	**0**	9	0	100	99	0	0	0	0	4	12	46	43
	3A	0	62	0	9	**0**	0^{T}	83	64	0	0	0	0	1	1	11	20
	3B	0	0	0	0	0^{T}	**0**	0	0	0	0	10	37	0	0	34	21
	4A	0	62	0	100	83	0	**0**	98^{T}	0	0	0	0	1	3	59	57
	4B	0	80	0	99	64	0	98^{T}	**0**	0	0	0	0	8	61	64	71
	5A	0	0	0	0	0	0	0	0	**0**	35^{T}	0	0	1	0	0	0
	5B	0	0	0	0	0	0	0	0	35^{T}	**0**	0	0	45	52	0	0
	6A	0	0	0	0	0	10	0	0	0	0	**0**	41^{T}	0	0	0	3
	6B	0	0	0	0	0	37	0	0	0	0	41^{T}	**0**	0	0	80	94
	7A	90	30	0	4	1	0	1	8	1	45	0	0	**0**	100^{T}	0	0
	7B	61	43	0	12	1	0	3	61	0	52	0	0	100^{T}	**0**	6	5
	8A	0	0	0	46	11	34	59	64	0	0	0	80	0	6	**0**	100^{T}
	8B	0	0	0	43	20	21	57	71	0	0	3	94	0	5	100^{T}	**0**

Table 5.3: Number of times a contradictory Bayes factor value has been obtained across 100 trials, with k samples from each writer combination. In bold, the reference writer and the questioned writers are the same person (h_p is true), "T" marks pairs of twins (h_t is true), in the other entries the participants are unrelated (h_u is true).

		Questioned writer															
k	Reference writer	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B	7A	7B	8A	8B
50	1A	**0**	0^{T}	0	0	0	0	0	0	0	0	0	0	83	24	0	0
	1B	0^{T}	**0**	0	31	22	0	35	69	0	0	0	0	3	4	0	0
	2A	0	0	**0**	0^{T}	0	0	0	0	0	0	0	0	0	0	0	0
	2B	0	31	0^{T}	**0**	0	0	100	99	0	0	0	0	0	0	2	5
	3A	0	22	0	0	**0**	0^{T}	79	29	0	0	0	0	0	0	0	0
	3B	0	0	0	0	0^{T}	**0**	0	0	0	0	0	5	0	0	5	0
	4A	0	35	0	100	79	0	**0**	89^{T}	0	0	0	0	0	0	11	18
	4B	0	69	0	99	29	0	89^{T}	**0**	0	0	0	0	0	18	28	34
	5A	0	0	0	0	0	0	0	0	**0**	6^{T}	0	0	0	0	0	0
	5B	0	0	0	0	0	0	0	0	6^{T}	**0**	0	0	17	5	0	0
	6A	0	0	0	0	0	0	0	0	0	0	**0**	16^{T}	0	0	0	0
	6B	0	0	0	0	0	5	0	0	0	0	16^{T}	**0**	0	0	60	83
	7A	83	3	0	0	0	0	0	0	0	17	0	0	**0**	99^{T}	0	0
	7B	24	4	0	0	0	0	0	18	0	5	0	0	99^{T}	**0**	0	0
	8A	0	0	0	2	0	5	11	28	0	0	0	60	0	0	**0**	100^{T}
	8B	0	0	0	5	0	0	18	34	0	0	0	83	0	0	100^{T}	**0**

^{T} Twins

Table 5.2 reports the number of replications supporting the hypotheses under the three above scenarios. The errors across 100 comparisons and k samples are detailed in Table 5.3 in the writer-dependent view.

The direction of the log-Bayes factor with respect to the neutral value of 0 could, in principle, be used in a decision theory framework, to "choose" which one of the hypotheses h_p and h_d is more supported by the evidence. The resulting decisions taken based on the obtained Bayes factors can be, then, analyzed with the usual metrics (that is, sensitivity and specificity) from a classification perspective:

Definition 5.1 (Sensitivity)
Sensitivity is the relative frequency across 100 comparisons in which a Bayes factor value is greater than 1 *when the prosecution hypothesis h_p is true.*

Definition 5.2 (Specificity)
Specificity is the relative frequency across 100 comparisons in which a Bayes factor value is lower than 1 *when the defense hypothesis h_d is true.*

Table 5.4: Bayes factor performance: h_p vs. h_d (h_t or h_u), according to the reference writer and the number of samples k. Notice the reduced specificity for twins even when large sample sizes are considered.

k	h_d	Sensitivity	Specificity
5	Twin	0.57	0.80
	Unrelated	0.57	0.93
10	Twin	0.93	0.55
	Unrelated	0.93	0.85
20	Twin	1.00	0.47
	Unrelated	1.00	0.81
30	Twin	1.00	0.52
	Unrelated	1.00	0.85
50	Twin	1.00	0.61
	Unrelated	1.00	0.92

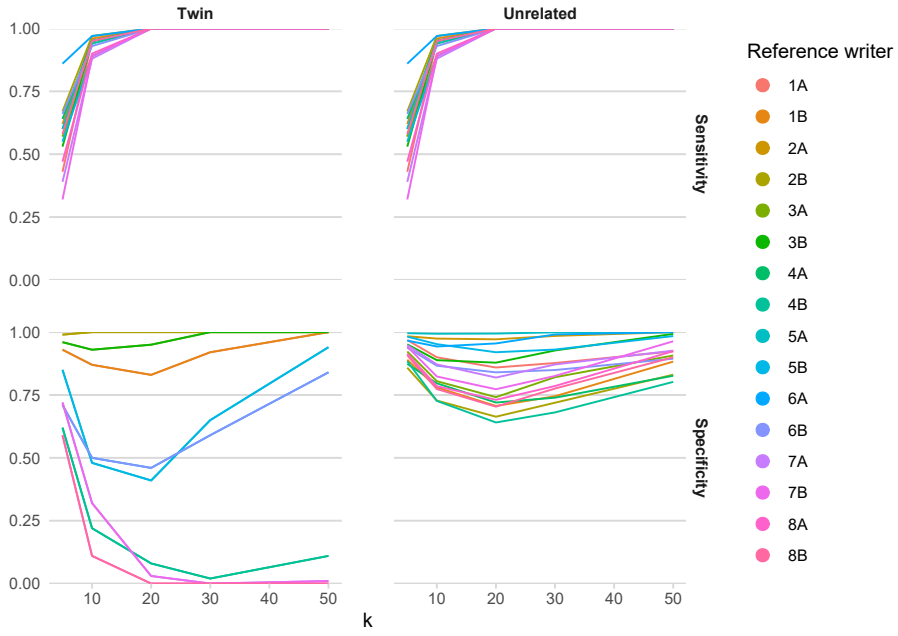


Figure 5.6: Bayes factor performance: h_p vs. h_d (h_t or h_u), according to the reference writer and the number of samples k.

The performance of the decision criterium across hypothesis pairs and the number of samples k is reported in Table 5.4 and in Figure 5.6 (distinguishing by reference writer). This confirms what has been seen in the past figures: results are dependent on which writers are considered. However, the Bayes factor supports the true hypothesis as k increases.

Note that we are not attempting to classify the items of evidence e_{r} and e_{q}, but only assessing the performance of this method, to understand how often a misleading Bayes factor is obtained. A classification using the Bayes factor would imply choosing between h_p and h_d, disregarding their prior probabilities, the utility/loss function and violating our understanding of Bayesian reasoning, expressing our conclusions in absolute terms. (Morey et al., 2016; Taroni et al., 2010)

The full distributions of the Bayes factor values across reference writers are shown in Figure 5.7. Concerning the hypotheses h_p and h_u, one can observe the same conclusions already shown in Figure 5.5 (not distinguishing writers). The support for/against h_t strongly depends on the selected twin pairs. In particular, some of the writers produce false identifications when compared with their twin (ex. writers 4A, 7A, 8A). Other twins write differently, as if they were unrelated (ex. writers 2A, 3A). Across k we observe the same trend as in the preceding figure: evidence strength increases (in absolute value) as more data is compared.

Character-dependent results

For a character-dependent analysis, collected data was first split across the characters "a", "d", "o". Then, the same protocol as in the previous section was followed inside each split.

A major side effect is that it is no longer possible to investigate large sample sizes, as the sampling with replacement was forbidden by the scenario simulation protocol. Available character counts, thus, greatly reduce the amount of obtainable results, to only those where writers provided a sufficient number of closed loops. The range of the sample sizes to consider has also been reduced.

The same statistical analysis as the one performed in the previous character-independent case was repeated. In particular, the values of the Bayes factors are shown in Figures 5.8 and 5.9, conditioning by the given character of interest.

These figures did not show any significant difference between the previously obtained results, confirming what UMAP suggested in the exploratory data analysis section.

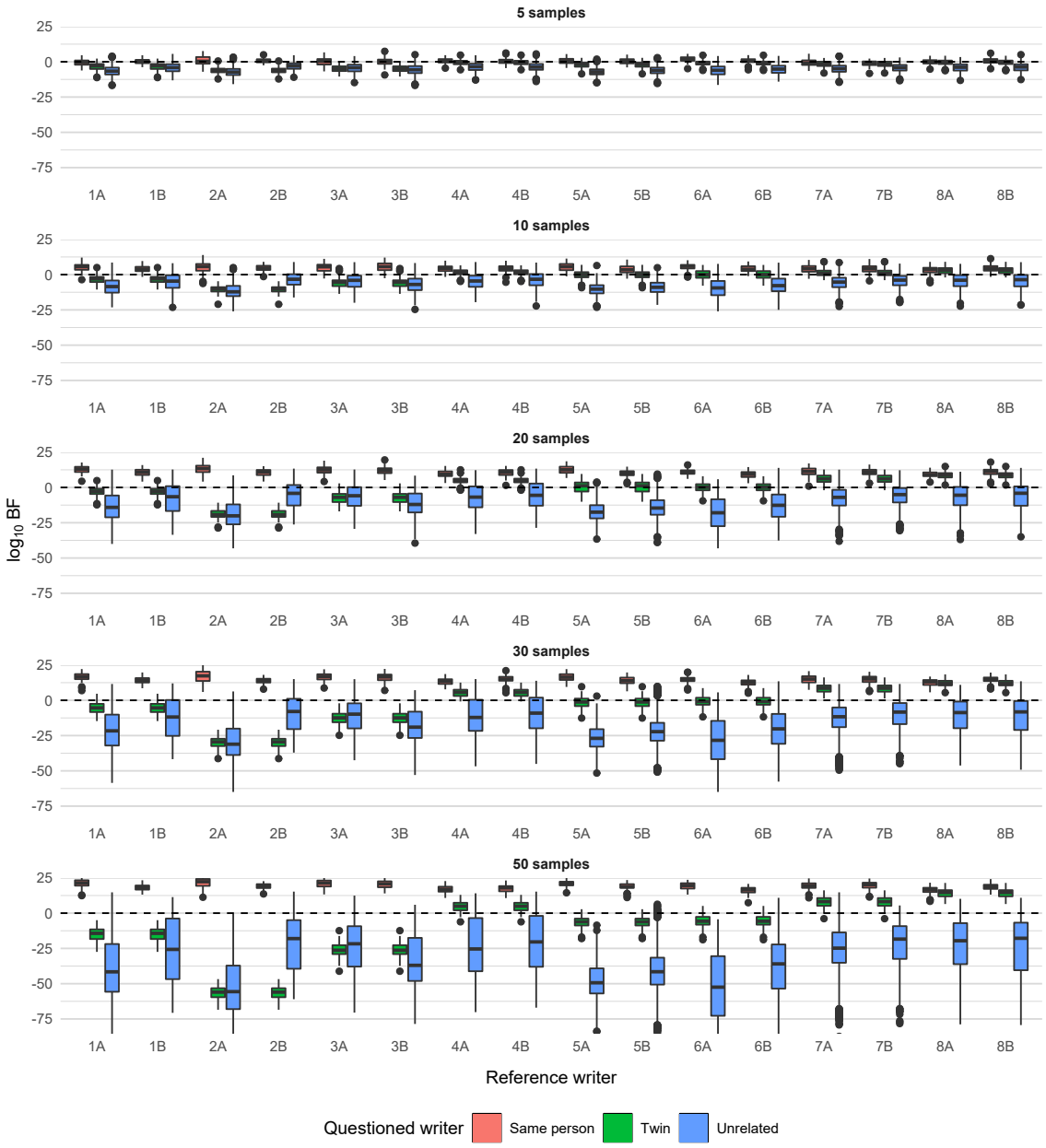


Figure 5.7: log-Bayes factor distributions across reference writers and the number of samples k. Notice that each writer appears at least thrice in comparisons: as the reference writer, as an unrelated writer or as a twin.

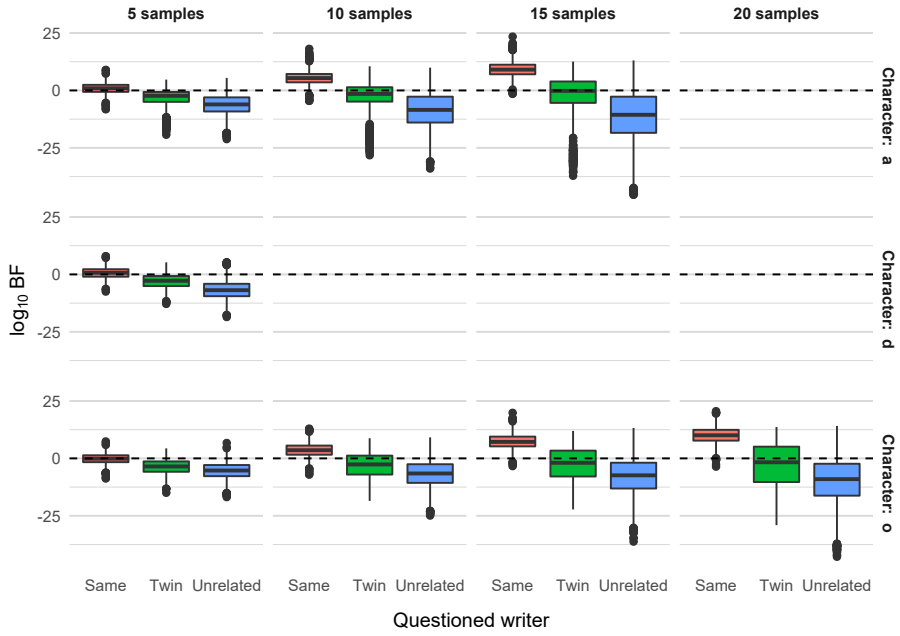


Figure 5.8: Distributions for the log-Bayes factor values across reference writers and number of samples, character-dependent. Notice that it is not possible to compare performances in many of the investigated cases due to the insufficient number of samples.

5.2.6 Discussion and conclusion

Evaluative findings are consistent with previous results presented in (Bozza et al., 2008) and Chapter 3, and support the hypothesis that different persons write differently. In particular, Tables 5.2 and 5.4 show that one needs at least 20 closed loops in evidence sets to be able to correctly support the source of a written material coming from a reference writer.

Concerning material coming from different writers, it is interesting to note that the error rate may not improve by collecting more samples if questioned writers are aggregated. However, Table 5.3 and Figure 5.6 show that this phenomenon can be mostly attributed to the performance of particular pairs of twins who write similarly. The phenomenon, however, does not appear when comparing unrelated persons, despite the number of comparisons being much greater. This implies that care must be taken when comparing handwriting coming from different writers, as some of them may show similar behaviors in terms of the analyzed features.

The limited number of twins in this study is a serious limiting factor, thus it

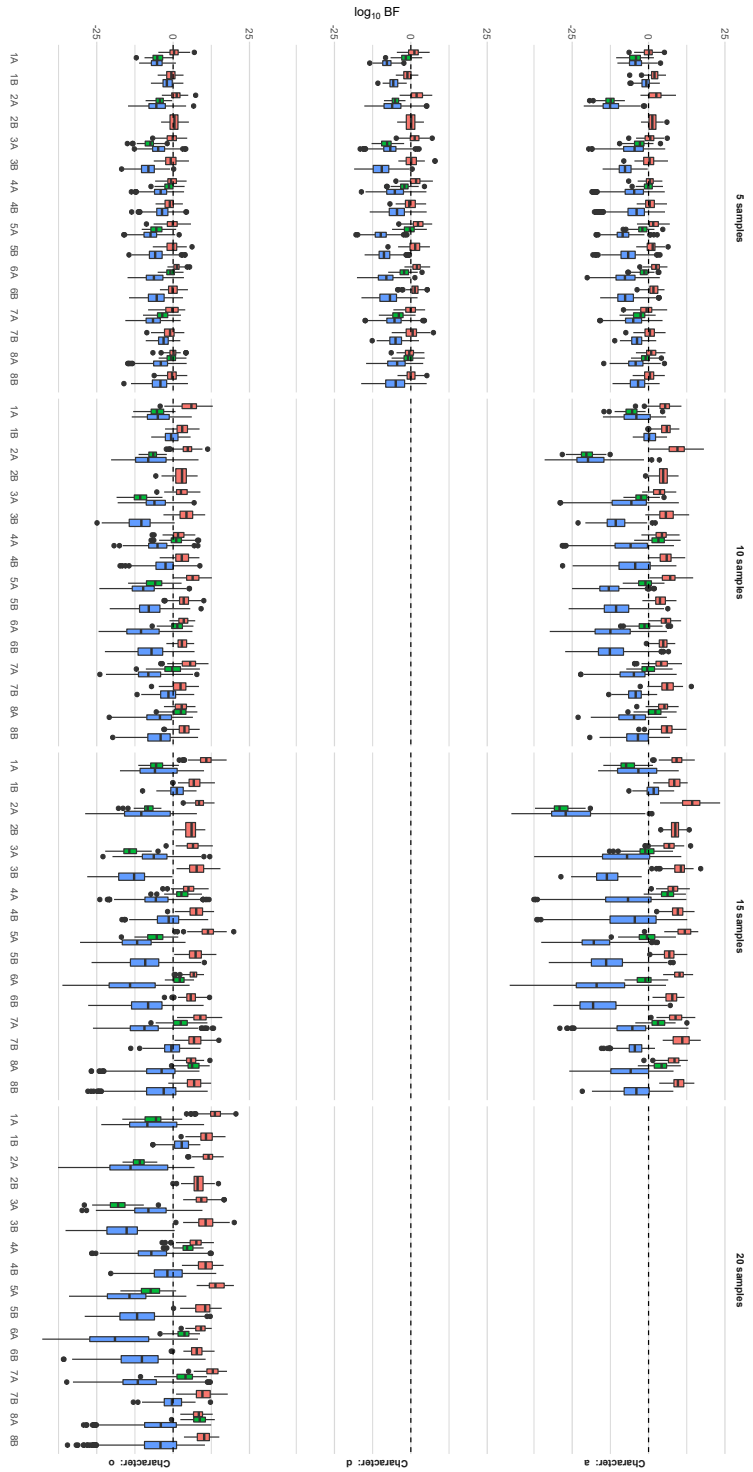


Figure 5.9: Distributions for the Bayes factor values across reference writers and number of samples k, writer- and character-dependent. Notice that it is not possible to compare performances in many of the investigated cases due to the insufficient number of samples.

is not possible to make any strong claim on the similarity of handwriting across twins as a general principle. However, Figure 5.7 shows that many log-Bayes factor values obtained considering twin pairs tend to lie closer to the neutral value of 0 than those across unrelated writers. Also, the decisionalization performance (in terms of "error rate") is bad when comparing twins instead of unrelated persons. These phenomena suggest that twins may reproduce more similar character loops than unrelated persons.

5.3 Microbiome evidence

In addition to the handwritten specimens, study participants also provided salivary samples that were afterwards analyzed by the Institute of Microbiology (CHUV), in order to characterize the salivary microbiome. Among the various possibilities for giving a quantitative description of the bacterial population in the saliva, the model described in Chapter 4 could be exploited if the laboratory reports the relative abundance of each taxon, as done in (Leake et al., 2016). We suppose that the laboratory is consistently searching for the same taxa across all samples, and no result is censored (e.g. under the limit of detection, therefore reported as "0").

5.3.1 Statistical model

Let us adopt the notation of the model for simple signatures in Section 4.3.2. Let p be the total number of taxa that the laboratory is able to detect ($p = 31$ according to the study design), and let t_k be the relative abundance of the k-th taxon, where $t_k \in [0, 1]$. The vector of relative abundances can be written as $d = (t_k)_{k=1}^{p}$, with $\sum_{k=1}^{p} t_k = 1$.

Additionally, d will generally differ across the participants and the visits: let us indicate with d_{ij} the relative abundance vector of the i-th participant obtained at the j-th visit. By noticing that d_{ij} belongs to the $(p-1)$-simplex, the model in Section 4.3.2 is directly applicable.

Particularly, one may introduce the parameters α_i to govern the bacterial variability across visits, and the parameter ψ to model the between-participants variability. The Dirichlet-Dirichlet-Gamma model (Section 4.3.2) could be further exploited to associate α_i and ψ to tractable distributions.

Background parameter elicitation

Depending on the structure of the between-participant level, one could obtain point estimates for the hyperparameters α_i and ψ either using the plug-in approximation (Section 4.4.1) or the ABC-like algorithm (Section 4.5.1). The choice between the methods will also depend on the number of observations (i.e. visits) for each participant: it might be possible that the plug-in estimators require a large number of visits to deliver stable estimates.

However, the structure of the microbiome dataset should be similar to the one used to validate the model for the simple signatures (Section 4.4). If the Dirichlet-Dirichlet-Gamma model is chosen, the plug-in estimators performed well.

In particular, the collected dataset can be considered to be the set of background observations.

5.3.2 Evaluative scenario

As this is a background-dominant operative condition (see Section 2.5.5), the background dataset can be exploited to simulate a real case. The evaluative setting is identical to the one already set for the handwritten data (Section 5.2.4).

The salivary traces (in terms of relative abundances) found on the crime scene are indicated with e_{q}. The reference salivary samples are indicated with e_{r}.

The evaluative hypotheses of interest are:

$H = h_p$: "the salivary traces e_{q} and e_{r} come from the same person"
$H = h_u$: "the salivary traces e_{q} and e_{r} come from two unrelated persons"
$H = h_t$: "the salivary traces e_{q} and e_{r} come from two twins".

Once one of the two defense hypotheses is considered (indicated with h_d), one can compute the Bayes factor value for h_p against h_d by applying the scenario simulation procedure, theoretically defined in Section 2.5.5 and implemented for the Dirichlet-Dirichlet-Gamma model in Section 4.4.2. Notice that the usage of a bridge sampler allows a much larger flexibility in the choice for the between-participant model rather than an ad-hoc Gibbs sampler.

5.3.3 What to expect

The same limitations discussed in Section 4.3.2 and in Section 4.3.2 could also potentially apply to this situation.

Particularly, it is desirable that the feature vectors d_{ij} do not change their length p. In other terms, the laboratory must search for the same taxa at all visits, and

report their abundances in the same order. We believe that this is not a significant limitation, as the laboratory processes must have been calibrated and standardized prior to the data collection.

The intra-variability of the taxa for a given participant must be estimated using at most 4 observations (the number of visits). If the number of taxa is large and if their distributions are uncorrelated, the problem is underdetermined, and the amount of observations will deliver broad credibility intervals unless one adopts very strong priors. This problem could be reduced if one reduces the dimensionality of the feature vector space, for instance by applying a suitable dimensionality reduction algorithm. However, the compositional model could lose its applicability unless one guarantees that the reduced vectors still belong to a simplex.

Notice that the relatively low number of visits might not be an issue if the intra-variability is lower than the inter-variability. To extremes, if the microbiota were as stable as the DNA, only one visit would suffice to completely define the participants' microbiota "profile". However, the stability can only be evaluated after multiple sample collection sessions.

The Dirichlet distribution could also be a problematic choice. Firstly, one should verify that the Dirichlet model is a good description of the taxa distributions, both theoretically and by validating it against the laboratory results. Secondly, it has been noticed that the Dirichlet-Dirichlet-Gamma model seems to be numerically sensitive when dealing with very small parameter values. This reflected not only during the sampling and background estimation procedures, but also to the computational stability of the Bayes factor value calculation. Alternative and less opinionated models are available (Section 4.6).

5.4 Combining evidence

In the first part of this Chapter, two separate Bayesian models are introduced for the handwritten and the salivary evidence, respectively. Once the available prior information is integrated, each model allows expressing the value of "its" evidence in the form of a Bayes factor. However, no connection between these models was made, and the combined evaluation of the evidence is not possible without introducing two additional assumptions.

Consider the hypothetical crime scene depicted at the beginning of this Chapter, where two kinds of evidence (a salivary stain and a ransom note) are recovered from the crime scene. Depending on the particular situation, one could consider several potential donors of the recovered traces. For instance, it might be reasonable to suppose that the glass item may have been stained by a third person that was able

to access the crime scene.

The **first assumption** is that all crime scene traces have been left by the same donor, whose identity (i.e. the suspect, a twin of the suspect, or another unrelated person) is disputed. This greatly simplifies the mathematical model for the combination of evidence, as it is not necessary to describe two putative sources (one for the handwritten evidence, the other for the salivary trace). By doing so, the hypotheses of the joint model can be easily composed by the hypotheses of the two sub-models.

The **second assumption** is the independence between the handwritten characters and the salivary stain, conditioned on the fact that they came from the same donor. In other terms, knowing the microbiome of a certain person's saliva does not inform how that person writes.

5.4.1 Evaluative scenario

Let us introduce some notation. Every variable appearing in the handwriting model (Section 5.2.4) and in the microbiome model (Section 5.3.2) will be indicated as $\bullet^{\mathcal{H}}$ and $\bullet^{\mathcal{B}}$, respectively. The crime scene traces are indicated with $e_{\mathrm{q}} = \left(e_{\mathrm{q}}^{\mathcal{H}}, e_{\mathrm{q}}^{\mathcal{B}}\right)$, while the reference material is indicated with $e_{\mathrm{r}} = \left(e_{\mathrm{r}}^{\mathcal{H}}, e_{\mathrm{r}}^{\mathcal{B}}\right)$. The same will also apply to the evaluative hypotheses of each model.

The first assumption will influence how the evaluative hypotheses are specified.

In this case, the hypotheses of interest are:

$$
\begin{aligned}
H = h_p &: \quad \text{“the traces } e_{\mathrm{q}} \text{ and } e_{\mathrm{r}} \text{ come from the same person”,} \\
H = h_u &: \quad \text{“the traces } e_{\mathrm{q}} \text{ and } e_{\mathrm{r}} \text{ come from two unrelated persons”,} \\
H = h_t &: \quad \text{“the traces } e_{\mathrm{q}} \text{ and } e_{\mathrm{r}} \text{ come from two twins”.}
\end{aligned}
$$

Notice that h_u is strictly contained into the intersection of $h_u^{\mathcal{H}}$ and $h_u^{\mathcal{B}}$: to be more precise, h_u implies that e_{r} will all come from donor D_{r}, while all crime scene traces $e_{\mathrm{q}}^{\mathcal{H}}$ and $e_{\mathrm{q}}^{\mathcal{B}}$ will all come from D_{q}, unrelated to the donor D_{r}. The same reasoning also applies to h_t: e_{r} will all come from donor D_{r}, while all crime scene traces e_{q} will all come from donor D_{q}, a twin of D_{r}.

The second assumption, instead, allows one to compute the value of the combined Bayes factor for a hypothesis pair (say, h_p versus h_d) by exploiting the available methods for the separate models.

Recall from Equation (2.10) that the Bayes factor for a hierarchical model can be written as a ratio of marginal likelihoods:

$$\mathrm{BF} = \frac{m(e_\mathrm{r}, e_\mathrm{q} \mid h_p)}{m(e_\mathrm{r} \mid h_d)\, m(e_\mathrm{q} \mid h_d)} \,.$$

Notice that the second assumption allows splitting each conditional probability as a product:

$$m(e_\mathrm{r}, e_\mathrm{q} \mid h_p) = m(e_\mathrm{r}^{\mathcal{H}}, e_\mathrm{r}^{\mathcal{B}}, e_\mathrm{q}^{\mathcal{H}}, e_\mathrm{q}^{\mathcal{B}} \mid h_p) = m(e_\mathrm{r}^{\mathcal{H}}, e_\mathrm{q}^{\mathcal{H}} \mid h_p)\, m(e_\mathrm{r}^{\mathcal{B}}, e_\mathrm{q}^{\mathcal{B}} \mid h_p) \,.$$

Also, since h_p implies both $h_p^{\mathcal{H}}$ and $h_p^{\mathcal{B}}$, one has that

$$m(e_\mathrm{r}^{\mathcal{H}}, e_\mathrm{q}^{\mathcal{H}} \mid h_p) = m(e_\mathrm{r}^{\mathcal{H}}, e_\mathrm{q}^{\mathcal{H}} \mid h_p^{\mathcal{H}}) \,.$$

This is also true for any other hypothesis, as traces coming from different sources are always conditionally independent. As a consequence, the Bayes factor can be written as the product of the Bayes factor values of the separated models:

$$\mathrm{BF} = \frac{m(e_\mathrm{r}^{\mathcal{H}}, e_\mathrm{q}^{\mathcal{H}} \mid h_p^{\mathcal{H}})}{m(e_\mathrm{r}^{\mathcal{H}} \mid h_d^{\mathcal{H}})\, m(e_\mathrm{q}^{\mathcal{H}} \mid h_d^{\mathcal{H}})} \, \frac{m(e_\mathrm{r}^{\mathcal{B}}, e_\mathrm{q}^{\mathcal{B}} \mid h_p^{\mathcal{B}})}{m(e_\mathrm{r}^{\mathcal{B}} \mid h_d^{\mathcal{B}})\, m(e_\mathrm{q}^{\mathcal{B}} \mid h_d^{\mathcal{B}})} = \mathrm{BF}^{\mathcal{H}}\, \mathrm{BF}^{\mathcal{B}} \,.$$

Finally, the values $\mathrm{BF}^{\mathcal{H}}$ and $\mathrm{BF}^{\mathcal{B}}$ can be computed by following the algorithms presented in the related Chapters.

5.4.2 A Bayesian network for the combination of evidence

The joint evaluation of items of evidence can be graphically represented with a Bayesian network, starting from the one introduced for the generic two-level hierarchical model (Section 2.5.2). A similar model is also given in (Taroni, Biedermann, et al., 2014, ch. 4).

To simplify the diagrams that will be shortly shown, assume that all available prior information (e.g. background observations, literature search, expert advice) is already integrated into the model. For instance, if the background observations e_b are available, the between-source parameter ψ should be distributed as $\psi \mid e_b$. By doing so, it is safe to omit the nodes corresponding to the background e_b and θ_b.

Consider the generic hypothesis $H = h_i$, where $i \in \{p, d\}$. Figure 5.10 shows a simple Bayesian network for the problem where a trace e_q is recovered from the crime scene, coming from the true source with parameters[3] θ_q. The reference material

[3]From now on, we identify a source with its parameters whenever no ambiguity is introduced: for instance, we will indicate "the source θ_q" instead of "the source parametrized by θ_q".

e_r is available, coming from the known donor θ_r. The node θ_q represents the true source of the questioned material: as it is never observed, θ_q becomes θ_r under h_p, and the defense source θ_d under h_d (e.g. a twin or an unknown person). Notice the reduced graph contains enough elements to calculate the Bayes factor by applying the previously described procedures.

This Bayesian network is the basic building block for the one that could describe the problem of combination of evidence. By considering evidence of two different kinds $\{\mathcal{H}, \mathcal{B}\}$, the network in Figure 5.10 can be duplicated and juxtaposed, adding the evidence superscripts where necessary. The hypothesis pair for the combined evidence H can be introduced as an additional node: $H = h_p$ implies $H^\mathcal{H} = h_p^\mathcal{H}$ and $H^\mathcal{B} = h_p^\mathcal{B}$, and $H = h_d$ implies $H^\mathcal{H} = h_d^\mathcal{H}$ and $H^\mathcal{B} = h_d^\mathcal{B}$. Graphically, this is represented with a divergent connection $H^\mathcal{H} \leftarrow H \rightarrow H^\mathcal{B}$. The obtained Bayesian network is shown in Figure 5.11.

Notice that the two assumptions are incorporated into the graph. By the first assumption, the handwritten and biological evidence have been deposited by the same donor. Graphically, the nodes $\theta_\mathrm{q}^\mathcal{H}$ and $\theta_\mathrm{q}^\mathcal{B}$ both refer to the donor d. If the true source of the handwritten material differed from the one of the biological evidence, two nodes $\theta_{\mathrm{d}^\mathcal{H}}^\mathcal{H}$ and $\theta_{\mathrm{d}^\mathcal{B}}^\mathcal{B}$ should have been respectively introduced to allow this possibility. The second assumption, instead, states that different types of evidence are conditionally independent. Graphically, the hypothesis nodes d-separate the left and the right part of the network: in other terms, whenever any hypothesis is instantiated (for instance when evaluating a Bayes factor), information cannot flow from one type of evidence to the other.

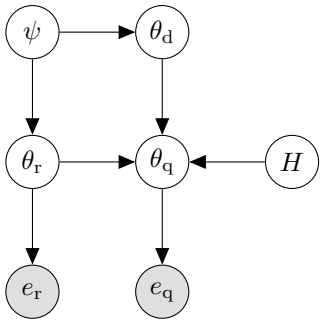


Figure 5.10: The basic Bayesian network for the problem of combination of items of evidence, described with a two-level hierarchical model. e_r is the reference material, e_q is the questioned material coming from the true source θ_q (latent). Shaded nodes are observed.

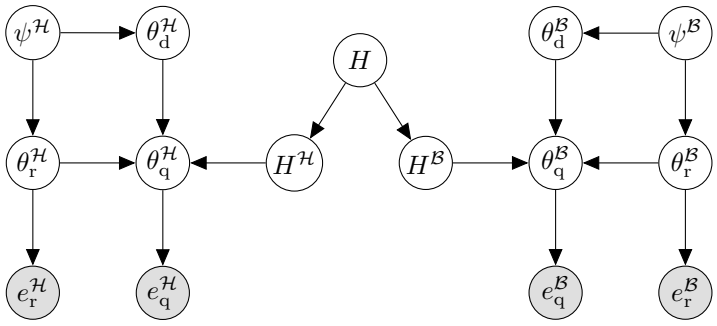


Figure 5.11: A Bayesian network for the problem of combination of evidence of types $\mathcal{H}$ and $\mathcal{B}$. The node H is the hypothesis pair for the combined model. Shaded nodes are observed.

5.4.3 What to expect

As frequently happening when combining information from multiple items of evidence, one can expect to encounter several issues that are not present when analyzing the items separately.

One of these issues is the need for a probabilistic description of the relation between the items (in this case, the handwritten material and the biological trace). The model shown in this Chapter is fully specified by introducing the aforementioned first and second assumptions, thus rendering the items of evidence independent, conditionally on the donor. We believe that these assumptions are fit for the purposes of this Chapter. If these assumptions do not hold, one may encounter situations where the observations may lead to different conclusions if they are considered jointly rather than separately: this fact is generally known as "the problem of conjunction". The interested reader may refer to (Taroni, Biedermann, et al., 2014, ch. 8) for more details.

Another issue is the fact that the items of evidence may convey strongly different information in terms of the order of magnitude of the respective Bayes factor. This can be problematic if the model is not sufficiently complex, as the "stronger" item may overcome the "weakest" observations. For instance, assume that a ransom note and a blood stain (both supposed independent, but coming from the same source) are recovered from a crime scene, and a suspect is available. The separate analyses both support the hypothesis that the suspect is the source of the trace: the DNA profile matches the suspect's DNA profile with a Bayes factor of 10^9, while a forensic document examiner reports a Bayes factor value of 100. It is evident how the handwritten evidence vanishes compared to the DNA analyses, since the FDE's conclusion has a limited value on the joint model. However, it is known that DNA

evidence frequently leads to large values of the Bayes factor, potentially shadowing problems (e.g. contaminations, traces left for innocent reasons) that might not be relevant at the source level, or may not appear in other "weaker" items of evidence. A properly stated Bayesian model should account for these possibilities, for instance by addressing multiple propositions in a hierarchical form (Cook et al., 1998a). Even if the model is well-posed and the results properly interpreted, the shadowing effect may still incur if a verbal scale for the Bayes factor is used.

Concerning this Chapter, we expect this second issue to be particularly concerning. All evidence items are mathematically translated to a high-dimensional representation that escapes intuition (the "curse of dimensionality"). As a consequence, the Bayes factor is computed only through numerical simulations, with no analytical insights available. Its range and its behavior are potentially sensitive to a number of parameters that must be set each time new casework is considered (e.g. the number of recovered items), and other parameters that relate to the computational procedures. The respective Bayes factors will, thus, strongly depend on how these parameters are set, and computational[4] and epistemological[5] inaccuracies may result in Bayes factor values that do not correctly represent the reality (see i.e. (Hopwood et al., 2012; Kaye, 2009)).

As further complication, the handwritten material and salivary traces have clearly different characteristics in terms of observations and number of variables. For instance, it may be reasonable to suppose that the handwritten material might show a large number of closed character loops (e.g. a ransom note), but only one salivary stain is available, providing no information on the salivary intra-variability. Moreover, each character loop is described by a vector of length $p^{\mathcal{H}} = 7$ (the number of retained Fourier harmonics, and the mean radius), while the salivary profile has a length that depends on the test kit used ($p^{\mathcal{B}} = 31$ according to the study design). A more familiar although unrealistic[6] situation would be the comparison of two DNA profiles with $p^{\mathcal{H}}$ and $p^{\mathcal{B}}$ matching loci. Consequently, it might be possible that the handwritten evidence may overcome the salivary observations or vice versa, depending on the complete model obtained once the laboratory results will be made available.

[4]E.g. how well the calculated posteriors converge, or the sensitivity of the model on a particular (hyper)parameter.

[5]E.g. how different is the model from the reality, or whether the model is capable of correctly evaluating the evidence under both hypotheses.

[6]DNA evidence has zero intra-variability, neglecting laboratory errors, drop-outs, spurious contaminations and significant degradations.

5.5 Epilogue

The situation encountered in this Chapter offers many opportunities for reflection, further improvements, as well as novel discoveries.

The handwriting part shares the same approach as Chapter 3 (natural handwriting) and is therefore subject to the same limitations and observations. Some of them concern the intrinsic characteristics of the model, such as its sensitivity on various parameters (e.g. the degrees of freedom of the inverse Wishart distribution) and the dependence on a background dataset. Others relate to the issue of combining evidence, the central theme of this Chapter.

In Chapter 3, it has been shown how a model can assess multiple items of evidence of the same kind (Section 3.7) whenever a joint description is available (i.e. a model encompassing multiple harmonic contributes). However, Chapter 4 has stressed that a proper assessment is required, and it would be desirable to consider evidence as a whole. For example, one of the handwritten features (the position of the paraph) was not quantified by our model, yet it revealed to bear a large evidential value by an FDE (Section 4.7).

The present Chapter, instead, considers a situation where multiple kinds of evidence are jointly evaluated. As this task is significantly different than the previous instances, the solution should also be built using a different strategy. In this case, by pairing two additional model assumptions with a Bayesian network (Section 5.4), we have been able to consider each kind of evidence separately, falling back to the previous situation. However, domain expertise is required in order to criticize and verify the additional assumptions: failure to address this aspect would lead to an incorrect joint evaluation of the evidence, such as in *People v. Collins*[7], no matter how sophisticated is the statistical model.

[7]People v. Collins, 68 Cal.2d 319.

Chapter 6

Conclusion

6.1 Outcomes

The primary goal of this thesis was to strengthen the way handwritten evidence is quantified and reported in an evaluative setting. We focused on giving quantitative descriptions of handwritten material consisting of natural handwriting samples and a particular instance of a forged signature by considering two types of quantitative descriptors: one exploits the presence of characters with closed loops, the other seeks to quantify certain proportions between line lengths. Evidence was evaluated following the ENFSI guidelines for evaluative reporting (Willis et al., 2015), notably by adopting a Bayes factor approach to translate the observations to a measure of support towards contrasting hypotheses of interest.

The first concrete situation, encountered in Chapter 3, involved questioned handwritten material exhibiting characters with closed loops, for instance a ransom letter, along with material coming from a reference writer. The approach was based on the Fourier descriptors first introduced by Marquis et al. (2005), and the Bayesian hierarchical model introduced by Bozza et al. (2008). An optimized ad-hoc implementation was developed, allowing for a much deeper exploration of the properties of the model, in particular of its Bayes factor. The implementation was first validated using fake (generated) data, as recommended by recent Bayesian practices (Gabry et al., 2019; Gelman et al., 2009). Results in past literature have been successfully reproduced by considering data collected under similar circumstances. Also, it has been possible to obtain several bounds on quantities of forensic interest, for instance the minimum number of samples needed in order to obtain a Bayes factor value with a given order of magnitude. Additionally, the Fourier descriptors were extended to a novel situation, notably the comparative analysis of a particular forged signature exhibiting closed

loops. In this case, the model appears to be able to correctly support the "same writer" hypothesis; when the questioned material has been written by someone else than the reference writer, results are mixed, strongly depending on the writer.

The second situation of forensic interest, encountered in Chapter 4, considered another signature that did not appear to show strongly discriminating features except for the presence of "peaks" and "valleys". The subjectively flowing character of the signature's master pattern has been translated to a novel quantitative descriptor, more properly characterizing the ratio between various distances between these peaks. A Bayesian hierarchical model is introduced, simultaneously leveraging on the framework of Chapter 3 and adapting it to the alleged quantitative properties of the observations. The model has been implemented using the Stan modeling language, and the Bayes factor is computed with a bridge sampling approach, trading off computational speed for flexibility. As before, the implementation has been validated against fake data, to make sure that the model is able to recover the generating parameters. Since the casework lacked a background dataset and quantitative expert knowledge, a strategy to elicit the hyperparameters has been devised inspired by ABC methods. To our knowledge, this result, although purely data-driven, is an entirely novel development in Bayesian methods applied to forensic science. Also, it should not suffer from the limitations raised by Robert et al. (2011) when ABC is used to compute a Bayes factor, since the Bayes factor is computed with a bridge sampler.

As the Bayesian methodology is not tied to any specific situation, any development could potentially have a transdisciplinary scope, bridging more complex scenarios involving various kinds of evidence. In this thesis, the model that described the forged signatures in the latter Chapter revealed to be applicable to compositional data.

A challenging and forensically interesting situation was considered in Chapter 5: a case where handwritten material and a salivary trace are recovered on a crime scene, a suspect is available for comparison, and the suspect claims that a twin was implied (to preclude the exploitation of DNA).

First, evidence was separately considered: Chapter 3 has been used to describe the handwritten material, while the model for compositional data developed in Chapter 4 has been proposed to be adapted to the salivary stain. Then, a third model for the combination of evidence has been introduced. Operatively, Chapter 5 involved experimental data collected in collaboration with the Institute of Microbiology (CHUV) of the University of Lausanne, in the form of handwritten specimens as well as salivary samples from twin pairs. However, the sample collection and analysis have been disrupted by the COVID-19 pandemic.

The handwritten material has been analyzed in full detail, although on a limited cohort. Results support the same conclusions as Chapter 3, namely that a writer's

intra-variability is generally smaller than inter-variability, so that the developed model can discriminate between the competing hypotheses. Interestingly, the performance is mixed when twins are considered: this supports the hypothesis that twins tend to write more similar than unrelated persons. However, a follow-up study is required on a much larger cohort, to further confirm or disprove these results.

The salivary samples have been collected and sequenced by the partner institution, but were not further statistically analyzed. Instead, Chapter 5 discussed how the previously developed methods could be exploited for the purpose of the study. To end the Chapter, a simple model for the combination of the two types of evidence is devised based on two assumptions that should hold on the complete dataset. Since a number of problems were possibly expected to arise once the analyses resume, the corresponding arguments and solutions were put forward as an effort of anticipating and leading future works.

6.2 Major issues and their solutions

Bayesian statistics is the common thread that underpins our understanding of uncertainty in forensic science, thus linking together all premises and results in this thesis. As a consequence, many difficulties of different nature have arisen, and the corresponding solutions were sought.

6.2.1 Theoretical issues

The first class of issues concerns the more theoretical aspects of the Bayesian framework. It is always assumed that two (or more) competing models are available, both explaining the evidence according to the prosecution and the defense hypothesis. The Bayes factor is a measure of the value of evidence, restricted to those models. However, to our knowledge, only few practitioners investigated the possible implications if one of these models is "incorrect" (meaning that it misrepresents the probability of a hypothetical set in the evidence space), thus also raising the question if a "correct" model even exists. Bernardo & Smith (1994) introduced the terms $\mathcal{M}$-closed and $\mathcal{M}$-open, respectively whether one of the evaluated models is "correct" or whether no attempt at searching for a correct model is made. The extent of these consequences on probability assignments, for instance, the numerator and the denominator of the Bayes factor, is currently an open problem (see Section 2.6.5), and the issue worsens once their ratio is computed. A related phenomenon can happen due to measure-theoretical reasons when the model and the hypothesis pair are defined such that the Bayes factor is not mathematically well-posed (Wetzels et al., 2010).

Notice that this situation might not appear pathological at first glance, but can arise when one of the hypotheses involves an exact equality constraint between continuous variables (see Section 2.6.4).

Since realistic data never perfectly "fit" any statistical model, very recent Bayesian developments consider the model specification issue to be unavoidable, advocating instead for a simulation-based approach, including the choice of the prior among its steps (Gelman et al., 2017; Vanpaemel & Lee, 2012). However, this results in considerable friction between Bayesian statisticians and forensic scientists, notably concerning the usage of the Bayes factor. The formers see the Bayes factor as a way to choose one model out of a set of alternative explanations: to that purpose, there are far better tools, such as the aforementioned simulation-based approach (Gelman et al., 2017; Vanpaemel & Lee, 2012). Forensic scientists, instead, consider the Bayes factor as *the* measure of choice, since it is the value that translates prior odds to posterior odds. To this purpose, it could be interesting to explore new ways to obtain theoretical bounds on the Bayes factor, for instance whether developing a more complete model of the casework is too costly (whilst being, perhaps, practically useless) (de Zoete & Sjerps, 2018).

To the more theoretical issues, one should also include the debate on which information is used to infer the values of the (hyper)parameters, and how the elicitation is performed. As already seen in Section 2.5.3, casework data can be used together with the background observations, or the process could be split into two distinct steps, first by eliciting the priors using background data alone, then by evaluating the casework evidence. In both cases, Bayesian reasoning should translate the updated belief to a probability distribution on the (hyper)parameters, to feed the algorithm that computes the Bayes factor. This step can be further simplified or approximated, depending on the statistical model or the chosen algorithm. In this thesis the aforementioned probability distribution on the (hyper)parameters (namely, the generic inter-variability parameter ψ) was collapsed to a single point estimate, as designed by the hierarchical model. Alternatively, one could have added another level to the hierarchical model, assigning a probability distribution to ψ.

6.2.2 Operative issues

Even if the theoretical underpinnings were completely understood, most situations considered in this thesis involve many assumptions that are difficult to verify. As it is known, the Bayesian framework uses probability distributions to represent the uncertainty on available measurements. The lack of empirical research and a strong reason (e.g. a physical law) to select the model's probability distributions implied that

the chosen model should be verified using data-oriented procedures ("goodness-of-fit"). However, most of these measurements are highly dimensional, so classical tools to check model assumptions (e.g. Q-Q plots) are either not applicable or are currently under active research. Additionally, analogously to the previous section, the key question should not be whether the specified model is "correct", rather whether the unavoidable deviations of the data from the model have a strong impact on the resulting inferences, the Bayes factor among them. In this thesis, these issues have been investigated by conducting sensitivity analyses of the Bayes factor to selected model or casework parameters, for instance the number of recovered specimens or the number of degrees of freedom. Notice that a more thorough analysis should also evaluate the dependence of the inferences of interest to the model itself, for example by supplying data generated by a process that is markedly different ("wrong") from the fitted one.

Another instance of an operative issue consisted in the situation where a large background dataset was not available at the time of analysis for the reasons given in Section 4.5.1. A data-driven algorithm was developed by exploiting the generative properties of Bayesian models, analogously to the commonly known Approximate Bayesian Computation methods: instead of relying on the background observations, parameter beliefs were updated by comparing the casework data to the generated one. We believe that this procedure is novel in this context, and could be used as a starting point during the case pre-assessment, as well as offering a number of future research opportunities.

6.2.3 Computational issues

The third class of issues relates to the notoriously difficult computation of the Bayes factors. A large number of methods are available in the statistical literature, each one with its own strengths and weaknesses (Section 2.6.4). However, no method is clearly superior to all the others, but they dramatically differ in terms of ease of implementation, assumptions, operative conditions, stability and flexibility. To give an example, in this thesis we used an ad-hoc Gibbs sampler for the model in Chapter 3 and a bridge sampler in Chapter 4. The Gibbs sampler has been thoroughly optimized for the problem at hand to allow extremely fast calculations: however, this step required months of work, and the obtained algorithm cannot be easily adapted to another model, for instance by introducing another parameter. The bridge sampler, instead, rests on the Stan modeling language: inference is slower, sampling from discrete variables is forbidden, but model specification is far more flexible, allowing parameters to be introduced or removed at will. Also, one does not

need to worry about implementation details, such as checking that all distributions have been correctly parametrized.

Notice that having a fast computational algorithm is not useless: on the contrary, it allows conducting much deeper sensitivity analyses to assess the performance of the model, that the Bayes factor is well-behaved, and that the evidence is robustly assessed. Moreover, the Bayesian workflow should be a closed loop, where a model is evaluated against its predictions, then eventually refined (Gabry et al., 2019).

6.3 Future research directions

As discussed in this Chapter, this thesis has explored various forensic scenarios, each one coupled with its own issues, challenges and the proposed approach to the interpretation of evidence. Notice that the Bayesian models hereby developed are by no means restricted to handwritten material, as demonstrated by the proposed application of Chapter 4 to the study of microbiological populations, or the features in Chapter 3 originating from the field of anthropology (Schmittbuhl et al., 1998). At the same time, Bayesian statistics has been enjoying lively and vivid developments while this thesis was being written, from its theoretical foundations (see, for instance, Section 2.6.5) to the available computational tools (e.g. the recent diffusion of the Stan modeling language (Carpenter et al., 2017)). Forensic science did not enjoy the same luxury: future research could address the integration of the novel Bayesian practices into all forensic disciplines. On the other hand, many of the open problems are still standing, such as the understanding and the communication of Bayes factors in complex scenarios. We believe that the adoption of these new tools and paradigms could bring novel interest to the matter, stimulating new developments in all forensic disciplines.

References

Ahuja, P., Chaudhary, P. K., ... Dahiya, M. (2018). Study of Genetic and Environmental Influence on Handwriting of Monozygotic Twins. *Indian Journal of Scientific Research*, 17–22.

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139–177.

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

Aitken, C., Nordgaard, A., ... Biedermann, A. (2018). Commentary: Likelihood Ratio as Weight of Forensic Evidence: A Closer Look. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00224

Aitken, C., Taroni, F., & Bozza, S. (2021). *Statistics and the evaluation of evidence for forensic scientists* (3rd ed). John Wiley & Sons. Retrieved from https://serval.unil.ch/notice/serval:BIB_A1A3D9997C08

Allen, M. (2015). *Foundations of forensic document analysis: Theory and practice*. Chichester, West Sussex ; Hoboken, NJ: John Wiley & Sons, Ltd.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, *49*(3), 161–164. https://doi.org/10.1016/j.scijus.2009.07.004

Ayyıldız, E., Gazi, V., & Wit, E. (2012). A Short Note on Resolving Singularity Problems in Covariance Matrices. *International Journal of Statistics and Probability*, *1*. https://doi.org/10.5539/ijsp.v1n2p113

Bai, J., & Shi, S. (2011). Estimating high dimensional covariance matrices and its applications. *Annals of Economics and Finance*, *112*(2), 199–215.

Benjamin, D. J., Berger, J. O., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-4286-2

Berger, M. A. (2011). The Supreme Court's Trilogy on the Admissibility of Expert Testimony. In *Reference manual on scientific evidence* (Vol. 9, p. 30). Federal Judicial Center Washington, DC.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. Chichester, Eng. ; New York: Wiley.

Bertillon, A. (1901). *La comparaison des écritures et l'identification graphique*. Paris: F.C.W. Vogel.

Biedermann, A. (2007). *Bayesian networks and the evaluation of scientific evidence in forensic science*. Université de Lausanne, Faculté de droit et des sciences criminelles. Retrieved from https://serval.unil.ch/notice/serval:BIB_E79B4396F90D

Biedermann, A., Champod, C., & Willis, S. (2017). Development of European standards for evaluative reporting in forensic science: The gap between intentions and perceptions. *The International Journal of Evidence & Proof*, *21*(1-2), 14–29. https://doi.org/10.1177/1365712716674796

Biedermann, A., Taroni, F., ... Davison, A. C. (2005). The evaluation of evidence in the forensic investigation of fire incidents. Part II. Practical examples of the use of Bayesian networks. *Forensic Science International*, *147*(1), 59–69. https://doi.org/10.1016/j.forsciint.2004.04.015

Biedermann, A., Taroni, F., ... Mazzella, W. D. (2011). 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), 58–66. https://doi.org/10.1016/j.forsciint.2010.05.001

Bisbing, R. E., & Wolner, M. F. (1984). Microscopical Discrimination of Twins' Head Hair. *Journal of Forensic Science*, *29*(3), 780–786. https://doi.org/

10.1520/JFS11736J

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Blum, M. G. B., & François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, *20*(1), 63–73. https://doi.org/10.1007/s11222-009-9116-0

Bos, C. S. (2002). A comparison of marginal likelihood computation methods. In *Compstat* (pp. 111–116). Springer.

Bozza, S. (2015). The value of scientific evidence for forensic multivariate data. *Statistica Applicata - Italian Journal of Applied Statistics*, *27*(2), 187–202.

Bozza, S., Broséus, J., ... Taroni, F. (2014). Bayesian classification criterion for forensic multivariate data. *Forensic Science International*, *244*, 295–301. https://doi.org/10.1016/j.forsciint.2014.09.017

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

Brault, J.-J., & Plamondon, R. (1993). A complexity measure of handwritten curves: Modeling of dynamic signature forgery. *Systems, Man and Cybernetics, IEEE Transactions on*, *23*(2), 400–413.

Brito, L. R. e, Martins, A. R., ... Pimentel, M. F. (2017). Critical review and trends in forensic investigations of crossing ink lines. *TrAC Trends in Analytical Chemistry*, *94*, 54–69. https://doi.org/10.1016/j.trac.2017.07.005

Buckleton, J. S., Bright, J.-A., & Taylor, D. (2018). *Forensic DNA evidence interpretation*. CRC press.

Caligiuri, M. P., & Mohammed, L. A. (2012). *The neuroscience of handwriting: Applications for forensic document examination*. CRC Press.

Carpenter, B., Gelman, A., ... Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1, 1), 1–32. https://doi.org/10.18637/jss.v076.i01

Casella, G. (1985). An Introduction to Empirical Bayes Data Analysis. *The American Statistician*, *39*(2), 83–87. https://doi.org/10.2307/2682801

Cereda, G. (2015). Non parametric Bayesian approach to LR assessment in case of rare haplotype match. *ArXiv e-Prints*. Retrieved from https://arxiv.org/abs/1506.08444

Cereda, G. (2017). Bayesian approach to LR assessment in case of rare type match. *Statistica Neerlandica*, *71*(2), 141–164. https://doi.org/10.1111/stan.12104

Chamberlain, B., Jensen, F. V., ... Nordahl, T. (2013, March 27). *Analysis in HUGIN of Data Conflict*. Retrieved from http://arxiv.org/abs/1304.1146

Champod, C. (2009). Identification and individualization. *Wiley Encyclopedia of Forensic Science*.

Champod, C., Evett, I. W., & Jackson, G. (2004). Establishing the most appropriate databases for addressing source level propositions. *Science & Justice*, *44*(3), 153–164.

Champod, C., Lennard, C. J., ... Stoilovic, M. (2017). *Fingerprints and Other Ridge Skin Impressions*. CRC Press. https://doi.org/10.1201/b20423

Champod, C., Taroni, F., & Margot, P. (2000). The Dreyfus Case-An Early Debate on Expert's Conclusions. *International Journal of Forensic Document Examiners*, *5*, 446–459.

Chazal, F., & Michel, B. (2017, October 11). *An introduction to Topological Data Analysis: Fundamental and practical aspects for data scientists*. Retrieved from http://arxiv.org/abs/1710.04019

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313–1321. Retrieved from http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476635

Christlein, V., Bernecker, D., ... Angelopoulou, E. (2015). Offline Writer Identification Using Convolutional Neural Network Activation Features. In J. Gall, P. Gehler, & B. Leibe (Eds.), *Pattern Recognition* (pp. 540–552). Springer International Publishing. https://doi.org/10.1007/978-3-319-24947-6_45

Christlein, V., Gropp, M., ... Maier, A. (2017). Unsupervised Feature Learning for Writer Identification and Writer Retrieval. In *2017 14th IAPR International*

Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 991–997). https://doi.org/10.1109/ICDAR.2017.165

Consonni, G., & Veronese, P. (2008). Compatibility of Prior Specifications Across Linear Models. *Statistical Science*, *23*(3), 332–353. https://doi.org/10.1214/08-STS258

Cook, R., Evett, I. W., ... Lambert, J. A. (1998a). A hierarchy of propositions: Deciding which level to address in casework. *Science & Justice*, *38*(4), 231–239.

Cook, R., Evett, I. W., ... Lambert, J. A. (1998b). A model for case assessment and interpretation. *Science & Justice*, *38*(3), 151–156. https://doi.org/10.1016/S1355-0306(98)72099-4

Cook, S. R., Gelman, A., & Rubin, D. B. (2006). Validation of Software for Bayesian Models Using Posterior Quantiles. *Journal of Computational and Graphical Statistics*, *15*(3), 675–692. https://doi.org/10.1198/106186006X136976

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1–73. https://doi.org/10.1016/0010-0277(95)00664-8

Curran, J. M. (2003). The Statistical Interpretation of Forensic Glass Evidence. *International Statistical Review*, *71*(3), 497–520. https://doi.org/10.1111/j.1751-5823.2003.tb00208.x

Curran, J. M., Champod, T. N. H., & Buckleton, J. S. (2000). *Forensic interpretation of glass evidence*. CRC Press.

Dawid, A. P. (2008). Beware of the DAG! In *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment - Volume 6* (pp. 59–86). Whistler, Canada: JMLR.org.

Dawid, A. P. (2017). Forensic likelihood ratio: Statistical problems and pitfalls. *Science & Justice*, *57*(1), 73–75. https://doi.org/10.1016/j.scijus.2016.09.002

Day, S. P. (2009). Handwriting and Signatures, Interpretation of Comparison Results. In *Wiley Encyclopedia of Forensic Science*. American Cancer Society. https://doi.org/10.1002/9780470061589.fsa138

De Finetti, B. (1930). Fondamenti logici del ragionamento probabilistico. *Bollettino Dell'Unione Matematica Italiana*, (5), 261–263.

De Finetti, B., Kyburg, jr H. E., & Smokler, H. E. (1964). Foresight: Its logical laws, its subjective sources. In *Studies in Subjective Probability* (pp. 93–158). New York: Wiley.

de Zoete, J., & Sjerps, M. (2018). Combining multiple pieces of evidence using a lower bound for the LR. *Law, Probability and Risk*, *17*(2), 163–178. https://doi.org/10.1093/lpr/mgy006

de Zoete, J., Sjerps, M., & Meester, R. (2017). Evaluating evidence in linked crimes with multiple offenders. *Science & Justice*, *57*(3), 228–238. https://doi.org/10.1016/j.scijus.2017.01.003

Dewhurst, T., Found, B., & Rogers, D. (2007). The relationship between quantitatively modelled signature complexity levels and forensic document examiners' qualitative opinions on casework. *Journal of Forensic Document Examination*, *18*, 21–40.

Diaz, M., Ferrer, M. A., ... Marcelli, A. (2017). Recovering Western On-Line Signatures from Image-Based Specimens. In (pp. 1204–1209). Presented at the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE. https://doi.org/10.1109/ICDAR.2017.199

Dziedzic, T. (2016). The influence of lying body position on handwriting. *Journal of Forensic Sciences*, *61*, S177–S183.

Dziedzic, T., Fabianska, E., & Toeplitz, Z. (2007). Handwriting of Monozygotic and Dizygotic Twins. *Problems of Forensic Sciences*, *69*, 30–36.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Ellen, D. (2005). *Scientific examination of documents: Methods and techniques*. Boca Raton, Florida: CRC Press.

Etz, A., Haaf, J. M., ... Vandekerckhove, J. (2018). Bayesian Inference and Testing Any Hypothesis You Can Specify. *Advances in Methods and Practices in Psychological Science*, *1*(2), 281–295. https://doi.org/10.1177/2515245918773087

Evett, I. W., Lambert, J. A., & Buckleton, J. S. (1998). A Bayesian approach to interpreting footwear marks in forensic casework. *Science & Justice*, *38*(4), 241–247. https://doi.org/10.1016/S1355-0306(98)72118-5

Evett, I. W., & Totty, R. N. (1985). A study of the variation in the dimensions of genuine signatures. *Journal of the Forensic Science Society*, *25*(3), 207–215.

Fabiańska, E., Kunicki, M., ... Bułka, D. (2006). Graphlog - Computer system supporting handwriting analysis. *Problems of Forensic Sciences*, *68*, 394–408.

Fasy, B. T., Kim, J., ... Maria, C. (2015, January 29). *Introduction to the R package TDA*. Retrieved from `http://arxiv.org/abs/1411.1830`

Fazio, K. (2015). The effects of constraint on a signature's static and dynamic features. *Journal of the American Society of Questioned Document Examiners*, *18*(1), 41–55.

Found, B. (2012). Handwriting and Signatures, Comparison of. In *Wiley Encyclopedia of Forensic Science*. American Cancer Society. `https://doi.org/10.1002/9780470061589.fsa331.pub2`

Found, B., & Rogers, D. (1996). The forensic investigation of signature complexity. In M. Simner, G. Leedham, & A. Thomassen (Eds.), *Handwriting and Drawing Research: Basic and Applied Issues* (pp. 483–492). IOS Press.

Found, B., & Rogers, D. (2008). The probative character of Forensic Handwriting Examiners' identification and elimination opinions on questioned signatures. *Forensic Science International*, *178*(1), 54–60. `https://doi.org/10.1016/j.forsciint.2008.02.001`

Found, B., & Rogers, D. K. (2003). The Initial Profiling Trial of a Program to Characterize Forensic Handwriting Examiners' Skill. *Journal of the American Society of Questioned Document Examiners*, *6*(2), 71–81.

Found, B., Rogers, D., & Herkt, A. (2002). The skill of a group of forensic document examiners in expressing handwriting and signature authorship and production process opinions. *J. Forensic Document Examination*, *14*, 15–30.

Found, B., Sita, J., & Rogers, D. (1999). The development of a program for characterizing forensic handwriting examiners' expertise: Signature examination pilot study. *Journal of Forensic Document Examination*, *12*, 69–80.

Franks, J., Davis, T., ... Grove, D. (1985). Variability of stroke direction between left-and right-handed writers. *Journal of the Forensic Science Society*, *25*(5), 353–370.

Gaborini, L. (2019, May 31). R package bayessource. https://doi.org/10.5281/zenodo.3570578 (Original work published 3 January 2019)

Gaborini, L. (2020a, May 15). R package rstanBF. https://doi.org/10.5281/zenodo.4404588 (Original work published 10 January 2019)

Gaborini, L. (2020b, September 13). R package rdirdirgamma. https://doi.org/DOI:%2010.5281/zenodo.4404592 (Original work published 28 August 2020)

Gaborini, L., Biedermann, A., & Taroni, F. (2017). Towards a Bayesian evaluation of features in questioned handwritten signatures. *Science & Justice*, *57*(3), 209–220. https://doi.org/10.1016/j.scijus.2017.01.004

Gabry, J., Simpson, D., ... Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Galbally, J., Gonzalez-Dominguez, S., ... Ortega-Garcia, J. (2015). Biografo: An Integrated Tool for Forensic Writer Identification. In U. Garain & F. Shafait (Eds.), *Computational Forensics* (pp. 200–211). Springer International Publishing.

Gamble, D. J. (1980). The Handwriting of Identical Twins. *Canadian Society of Forensic Science Journal*, *13*(1), 11–30. https://doi.org/10.1080/00085030.1980.10757337

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference* (2nd ed). Chapman and Hall/CRC.

Gelman, A. (2011). Causality and statistical learning. *American Journal of Sociology*, *117*(3), 955–966.

Gelman, A., Bois, F., & Jiang, J. (1996). Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions. *Journal of the American Statistical Association*, *91*(436), 1400–1412.

Gelman, A., Carlin, J. B., ... Rubin, D. B. (2009). *Bayesian data analysis* (2nd ed.). Chapman and Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 733–760.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, *19*(10), 555. https://doi.org/10.3390/e19100555

Ghosh, J. K., & Ramamoorthi, R. (2003). *Bayesian nonparametrics*. Springer Science & Business Media.

Gill, P. (2012). Misleading DNA Evidence: Reasons for Miscarriages of Justice. *International Commentary on Evidence*, *10*(1), 55–71. https://doi.org/10.1515/ice-2014-0010

Gittelson, S. (2013). *Evolving from inferences to decisions in the interpretation of scientific evidence*. Université de Lausanne, Faculté de droit et des sciences criminelles. Retrieved from https://serval.unil.ch/notice/serval:BIB_620A73F01CCC

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

Gonzalez-Rodriguez, J., Rose, P., ... Ortega-Garcia, J. (2007). Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2104–2115. https://doi.org/10.1109/TASL.2007.902747

Good, I. (1991). Weight of evidence and the Bayesian likelihood ratio. In *The Use Of Statistics In Forensic Science* (pp. 85–106). CRC Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732.

Grieve, M., Roux, C., ... Taroni, F. (2017). Interpretation of fibre evidence. In *Forensic examination of fibres* (pp. 345–426). CRC Press.

Gringras, P., & Chen, W. (2001). Mechanisms for differences in monozygous twins. *Early Human Development*, *64*(2), 105–117. https://doi.org/10.1016/S0378-3782(01)00171-2

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). *Bridgesampling: An R Package for Estimating Normalizing Constants*. Retrieved from https://arxiv.org/abs/1710.08162

Guarnera, L., Farinella, G. M., ... Battiato, S. (2018). Forensic analysis of handwritten documents with GRAPHJ. *Journal of Electronic Imaging*, *27*(5), 051230.

Haaf, J. M., Ly, A., & Wagenmakers, E.-J. (2019). Retire significance, but still test hypotheses. *Nature*, *567*, 461–461. https://doi.org/10.1038/d41586-019-00972-7

Hafemann, L. G., Sabourin, R., & Oliveira, L. S. (2016). Analyzing features learned for offline signature verification using Deep CNNs. In *Pattern Recognition (ICPR), 2016 23rd International Conference on* (pp. 2989–2994). IEEE.

Hafemann, L. G., Sabourin, R., & Oliveira, L. S. (2017). Offline handwritten signature verification - Literature review. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1–8). https://doi.org/10.1109/IPTA.2017.8310112

Harralson, H. (2014). *Developments in handwriting and signature identification in the digital age*. Routledge.

Harrison, W. R. (1958). *Suspect Documents, Their Scientific Examination*. Sweet & Maxwell.

Hassaine, A., Al Maadeed, S., & Bouridane, A. (2013). ICDAR 2013 Competition on Handwriting Stroke Recovery from Offline Data. In (pp. 1412–1416). IEEE. https://doi.org/10.1109/ICDAR.2013.285

Heck, D. (2019). A Caveat on the Savage-Dickey Density Ratio: The Case of Computing Bayes Factors for Regression Parameters. *British Journal of Mathematical and Statistical Psychology*, *72*, 316–333. https://doi.org/10.1111/bmsp.12150

Hendricks, J. H., Neumann, C., & Saunders, C. P. (2020, January 13). *Quantification of the weight of fingerprint evidence using a ROC-based Approximate Bayesian Computation algorithm for model selection*. Retrieved from http://arxiv.org/abs/1803.10121

Hepler, A. B., Saunders, C. P., ... Buscaglia, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Science International*, *219*(1), 129–140. https://doi.org/10.1016/j.forsciint.2011.12.009

Hilton, O. (1963). Some basic rules for the identification of handwriting. *Medicine, Science and the Law*, *3*(1), 107–117.

Hilton, O. (1992). *Scientific examination of questioned documents*. CRC press.

Hoeting, J. A., Madigan, D., ... Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, *14*(4), 382–401.

Holmes, C. C., Caron, F., ... Stephens, D. A. (2015). Two-sample Bayesian Nonparametric Hypothesis Testing. *Bayesian Analysis*, *10*(2), 297–320. https://doi.org/10.1214/14-BA914

Hopwood, A. J., Puch-Solis, R., ... Tully, G. (2012). Consideration of the probative value of single donor 15-plex STR profiles in UK populations and its presentation in UK courts. *Science & Justice*, *52*(3), 185–190. https://doi.org/10.1016/j.scijus.2012.05.005

Huber, R. A., & Headrick, A. M. (1999). *Handwriting identification: Facts and fundamentals*. CRC press.

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

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester, U.K: Wiley.

Jackson, G., Aitken, C., & Roberts, P. (2010). Case assessment and interpretation of expert evidence. *Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses. Practitioner Guide*, (4).

Jain, A. K., Prabhakar, S., & Pankanti, S. (2002). On the similarity of identical twin fingerprints. *Pattern Recognition*, *35*(11), 2653–2663. https://doi.org/10.1016/S0031-3203(01)00218-7

Jamieson, A., & Moenssens, A. A. (Eds.). (2009). *Wiley encyclopedia of forensic science*. Chichester, West Sussex, U.K. ; [Hoboken, N.J.]: John Wiley & Sons.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Individual-specific 'fingerprints' of human DNA. *Nature*, *316*(6023), 76–79. https://doi.org/10.1038/316076a0

Johnson, D. H. (1999). The Insignificance of Statistical Significance Testing. *The Journal of Wildlife Management*, *63*(3), 763–772. https://doi.org/10.2307/3802789

Johnson, M. E., Vastrick, T. W., ... Schuetzner, E. (2017). Measuring the Frequency Occurrence of Handwriting and Handprinting Characteristics. *Journal of Forensic Sciences*, *62*(1), 142–163. https://doi.org/10.1111/1556-4029.13248

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed). Upper Saddle River, N.J: Pearson Prentice Hall.

Kailath, T. (1967). The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communication Technology*, *15*(1), 52–60. https://doi.org/10.1109/TCOM.1967.1089532

Kam, M., Fielding, G., & Conn, R. (1997). Writer Identification by Professional Document Examiners. *Journal of Forensic Sciences*, *42*(5), 14207J. https://doi.org/10.1520/JFS14207J

Kam, M., Fielding, G., & Conn, R. (1998). Effects of Monetary Incentives on Performance of Nonprofessionals in Document-Examination Proficiency Tests. *Journal of Forensic Sciences*, *43*(5), 14348J. https://doi.org/10.1520/JFS14348J

Kam, M., Gummadidala, K., ... Conn, R. (2001). Signature Authentication by Forensic Document Examiners. *Journal of Forensic Sciences*, *46*(4), 15062J. https://doi.org/10.1520/JFS15062J

Kam, M., & Lin, E. (2003). Writer Identification Using Hand-Printed and Non-Hand-Printed Questioned Documents. *Journal of Forensic Sciences*, *48*(6), 2002321. https://doi.org/10.1520/JFS2002321

Kam, M., Wetstein, J., & Conn, R. (1994). Proficiency of Professional Document Examiners in Writer Identification. *Journal of Forensic Sciences*, *39*(1), 13565J. https://doi.org/10.1520/JFS13565J

Kamary, K., Mengersen, K., ... Rousseau, J. (2014). *Testing hypotheses via a mixture estimation model*. Retrieved from https://arxiv.org/abs/1412.2044

Kass, R. E. (1993). Bayes factors in practice. *The Statistician*, *42*(5), 551–560.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kaye, D. (2007). Revisiting 'Dreyfus': A More Complete Account of a Trial by Mathematics. *Minnesota Law Review*, *91*, 825.

Kaye, D. H. (1986). Is Proof of Statistical Significance Relevant? *Washington Law Review*, *61*, 35.

Kaye, D. H. (2009). Trawling DNA databases for partial matches: What is the FBI afraid of. *Cornell JL & Pub. Pol'y*, *19*, 145.

Kelly, J. S., & Lindblom, B. S. (Eds.). (2006). *Scientific examination of questioned documents* (2nd ed.). Boca Raton, Florida: CRC/Taylor & Francis.

Kiran, P., & Sridhar, D. (2017). A Study to Determine the Inheritance Pattern of Characteristics of Handwriting between Parents and Off-springs. *Journal of Forensic Science & Criminology*, *5*(3), 1.

Kjaerulff, U. B., & Madsen, A. L. (2008). Bayesian networks and influence diagrams. *Springer Science+ Business Media*, *200*, 114.

Köller, N., Niessen, K., ... Sadorf, E. (2004). Probability Conclusions in Expert Opinions on Handwriting: Substantiation and Standardization of Probability Statements in Expert Opinions. *Substantiation and Standardization of Probability Statements in Expert Opinions. Luchterhand, München.*

Koppenhaver, K. (2007). *Forensic document examination: Principles and practice.* Totowa, N.J: Humana Press.

Korb, K. B., & Nicholson, A. E. (2010). *Bayesian artificial intelligence* (2nd ed). CRC press.

Krieger, N., & Davey Smith, G. (2016). The tale wagged by the DAG: Broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology*, *45*(6), 1787–1808. https://doi.org/10.1093/ije/dyw114

Lartillot, N., & Philippe, H. (2006). Computing Bayes Factors Using Thermodynamic Integration. *Systematic Biology*, *55*(2), 195–207. https://doi.org/10.1080/10635150500433722

Leake, S. (2014). *Human identification through analysis of the salivary microbiome: Proof of principle*. Université de Lausanne, Faculté de droit et des sciences criminelles. Retrieved from https://serval.unil.ch/notice/serval:BIB_6FBD46E02D68

Leake, S. L., Pagni, M., ... Greub, G. (2016). The salivary microbiome for differentiating individuals: Proof of principle. *Microbes and Infection*, *18*(6), 399–405.

Leclerc, F., & Plamondon, R. (1994). Automatic signature verification: The state of the art—1989–1993. *International Journal of Pattern Recognition and Artificial Intelligence*, *8*(03), 643–660.

Leedham, G., & Srihari, S. (2003). A survey of computer methods in forensic handwritten document examination. In (pp. 278–281). Presented at the Proceeding the Eleventh International Graphonomics Society Conference, Scottsdale Arazona.

Leuenberger, C., & Wegmann, D. (2010). Bayesian Computation and Model Selection Without Likelihoods. *Genetics*, *184*(1), 243–252. https://doi.org/10.1534/genetics.109.109058

Lewis, J. A. (2014). *Forensic document examination: Fundamentals and current trends*. Oxford ; San Diego: Academic Press.

Li, C. (2019). Competency for Chinese Handwriting and Signature Examination. *Journal of Forensic Sciences*, *64*(2), 607–615. https://doi.org/10.1111/1556-4029.13895

Linden, J., Marquis, R., ... Taroni, F. (2018). Dynamic signatures: A review of dynamic feature variation and forensic methodology. *Forensic Science International*, *291*, 216–229. https://doi.org/10.1016/j.forsciint.2018.08.021

Linden, J., Taroni, F., ... Bozza, S. (2021). Bayesian multivariate models for case assessment in dynamic signature cases. *Forensic Science International*, *318*, 110611. https://doi.org/10.1016/j.forsciint.2020.110611

Lindley, D. V. (1971). *Making decisions*. Wiley Interscience.

Lindley, D. V. (1991). Probability. In C. G. G. Aitken & D. A. Stoney (Eds.), *The use of statistics in forensic science*. CRC Press.

Ling, S. (2002). A preliminary investigation into handwriting examination by multiple measurements of letters and spacing. *Forensic Science International*, *126*(2), 145–149.

Liwicki, M., Malik, M. I., ... Found, B. (2012). ICFHR 2012 Competition on Automatic Forensic Signature Verification (4NsigComp 2012). In *2012 International Conference on Frontiers in Handwriting Recognition* (pp. 823–828). https://doi.org/10.1109/ICFHR.2012.217

Lizega Rika, J. (2018). Relative Width and Height of Handwritten Letter. *Journal of Forensic Sciences*, *63*(1), 178–190. https://doi.org/10.1111/1556-4029.13483

Loakes, D. (2008). A forensic phonetic investigation into the speech patterns of identical and non-identical twins. *International Journal of Speech Language and the Law*, *15*(1), 97–100.

Locard, E. (1959). Les faux en écriture et leur expertise.

Lodewyckx, T., Kim, W., ... Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*(5), 331–347. https://doi.org/10.1016/j.jmp.2011.06.001

Louloudis, G., Gatos, B., ... Papandreou, A. (2013). ICDAR 2013 Competition on Writer Identification. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 1397–1401). https://doi.org/10.1109/ICDAR.2013.282

Louloudis, G., Stamatopoulos, N., & Gatos, B. (2011). ICDAR 2011 Writer Identification Contest. In *2011 International Conference on Document Analysis and Recognition* (pp. 1475–1479). https://doi.org/10.1109/ICDAR.2011.293

Lowe, D. G. (1989). Organization of smooth image curves at multiple scales. *International Journal of Computer Vision*, *3*(2), 119–130.

Lynch, S. M., & Western, B. (2004). Bayesian Posterior Predictive Checks for Complex Models. *Sociological Methods & Research*, *32*(3), 301–335. https://doi.org/10.1177/0049124103257303

Maciaszek, J. (2011). Natural variation in measurable features of initials. *Z Zagadnien Nauk Sadowych*, *85*, 25–39.

Malik, M. I., Ahmed, S., ... Liwicki, M. (2015). ICDAR2015 competition on signature verification and writer identification for on-and off-line skilled forgeries (SigWIcomp2015). In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on* (pp. 1186–1190). IEEE.

Mansuy, R., & Mazliak, L. (2008). 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), 60–63.

Marcelli, A., Parziale, A., & Stefano, C. D. (2015). Quantitative evaluation of features for Forensic Handwriting Examination. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1266–1271). https://doi.org/10.1109/ICDAR.2015.7333952

Marin, J.-M., Pillai, N. S., ... Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(5), 833–859.

Marin, J.-M., & Robert, C. P. (2010). On resolving the Savage–Dickey paradox. *Electronic Journal of Statistics*, *4*, 643–654. https://doi.org/10.1214/10-EJS564

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

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

Marquis, R., Hicks, T., & Mazzella, W. (2019). Forensic Investigative and Evaluative Assessment of Handwritten X-Marks. *Journal of the American Society of Questioned Document Examiners*, *22*(1), 11.

Marquis, R., Mazzella, W., & Hicks, T. (2019). X-marks: Too simple to be useful? *Nowa Kodyfikacja Prawa Karnego*, *49*, 103–110. https://doi.org/10.19195/2084-5065.49.8

Marquis, R., Schmittbuhl, M., ... Taroni, F. (2005). 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), 23–32. https://doi.org/10.1016/j.forsciint.2004.06.028

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

Marquis, R., Taroni, F., ... Schmittbuhl, M. (2007). Size influence on shape of handwritten characters loops. *Forensic Science International*, *172*(1), 10–16. https://doi.org/10.1016/j.forsciint.2006.11.005

Martire, K. A., Growns, B., & Navarro, D. J. (2018). What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts. *Psychonomic Bulletin & Review*, *25*(6), 2346–2355. https://doi.org/10.3758/s13423-018-1448-3

Martire, K. A., Kemp, R. I., ... Newell, B. R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International*, *240*, 61–68. https://doi.org/10.1016/j.forsciint.2014.04.005

Matuszewski, S., & Maciaszek, J. (2008). Natural variation in length of signature components. *Z Zagadnien Nauk Sadowych*, *74*, 182–189.

McElreath, R. (2015). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Vol. 179). Chapman and Hall/CRC. Retrieved from http://dx.doi.org/10.1111/rssa.12221

McInnes, L., Healy, J., & Melville, J. (2018, February 9). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Retrieved from http://arxiv.org/abs/1802.03426

McShane, B. B., Gal, D., ... Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, *73*, 235–245. https://doi.org/10.1080/00031305.2018.1527253

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*(4), 831–860.

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

Michel, L. (1978). Disguised Signatures. *Journal of the Forensic Science Society*, *18*(1), 25–29. https://doi.org/10.1016/S0015-7368(78)71179-5

Minka, T. P. (2000). *Estimating a Dirichlet distribution*. Retrieved from https://tminka.github.io/papers/dirichlet/

Moenssens, A. A. (1997–1998). Handwriting Identification Evidence in the Post-Daubert World. *UMKC Law Review*, *66*, 251.

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

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Morris, R., & Morris, R. N. (2000). *Forensic handwriting identification: Fundamental concepts and principles*. San Diego, California: Academic Press.

Morton, S. E. (1980). How Does Crowding Affect Signatures? *Journal of Forensic Science*, *25*(1), 141–145. https://doi.org/10.1520/JFS10948J

Muehlberger, R., Newman, K., ... Wichmann, J. (1977). A Statistical Examination of Selected Handwriting Characteristics. *Journal of Forensic Sciences*, *22*(1), 206–215. https://doi.org/10.1520/JFS10388J

Münch, A. (2000). *L'expertise en écritures et en signatures*. Les éditions du Septentrion.

Natural Intelligent Technologies Srl. (2012). Masquerade. Retrieved 18 September 2019, from https://www.nitesrl.com

Neumann, C., & Ausdemore, M. A. (2019, October 11). *Defence Against the Modern Arts: The Curse of Statistics "Score-based likelihood ratios"*. Retrieved from http://arxiv.org/abs/1910.05240

Neumann, C., Champod, C., ... Bromage-Griffiths, A. (2006). Computation of likelihood ratios in fingerprint identification for configurations of three minutiae. *Journal of Forensic Sciences*, *51*(6), 1255–1266. https://doi.org/10.1111/j.1556-4029.2006.00266.x

Neumann, C., Champod, C., ... Bromage-Griffiths, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences*, *52*(1), 54–64. https://doi.org/10.1111/j.1556-4029.2006.00327.x

Niels, R., Vuurpijl, L., & Schomaker, L. (2005). Introducing TRIGRAPH - trimodal writer identification. *Proc. European Network of Forensic Handwr. Experts*.

Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.

Nordgaard, A., & Rasmusson, B. (2012). The likelihood ratio as value of evidence—more than a question of numbers. *Law, Probability and Risk*, *11*(4), 303–315. https://doi.org/10.1093/lpr/mgs019

Osborn, A. S. (1910). *Questioned documents: A study of questioned documents with an outline of methods by which the facts may be discovered and shown*. New York, NY: Lawyers' Co-operative Pub. Company.

Ostrum, R. B. (2019). CSFS Document Section Position on the Logical Approach to Evidence Evaluation and Corresponding Wording of Conclusions. *Canadian Society of Forensic Science Journal*, *52*(3), 129–138. https://doi.org/10.1080/00085030.2019.1635736

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan kaufmann.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

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

Press, S. J. (2012). *Applied multivariate analysis: Using Bayesian and frequentist methods of inference* (2nd ed). Courier Corporation.

Pritchard, J. K., Seielstad, M. T., ... Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, *16*(12), 1791–1798. https://doi.org/10.1093/oxfordjournals.molbev.a026091

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from

https://www.R-project.org/

Rehman, A., Naz, S., & Razzak, M. I. (2019). Writer identification using machine learning approaches: A comprehensive review. *Multimedia Tools and Applications*, *78*(8), 10889–10931. https://doi.org/10.1007/s11042-018-6577-1

Risinger, D. (2007). Cases Involving the Reliability of Handwriting Identification Expertise Since the Decision in Daubert. *Tulsa L Rev*, *43*.

Risinger, D. M., Denbeaux, M. P., & Saks, M. J. (1989). Exorcism of ignorance as a proxy for rational knowledge: The lessons of handwriting identification "expertise". *University of Pennsylvania Law Review*, *137*(3), 731–792.

Robert, C. P., Cornuet, J.-M., ... Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, *108*(37), 15112–15117.

Robertson, B., Vignaux, G. A., & Berger, C. E. (2016). *Interpreting evidence: Evaluating forensic science in the courtroom*. John Wiley & Sons.

Robitzsch, A. (2020). *Sirt: Supplementary item response theory models*. manual. Retrieved from https://CRAN.R-project.org/package=sirt

Saks, M. J. (2003). Commentary on: Srihari SN, Cha S-H, Arora H, Lee S. Individuality of handwriting. J Forensic Sci 2002; 47(4):856–72. *Journal of Forensic Sciences*, *48*(4), 2002428. https://doi.org/10.1520/JFS2002428

Saks, M. J., & Koehler, J. J. (2005). The Coming Paradigm Shift in Forensic Identification Science. *Science*, *309*(5736), 892–895. https://doi.org/10.1126/science.1111565

Samie, L., Hicks, T., ... Taroni, F. (2016). Stabbing simulations and DNA transfer. *Forensic Science International: Genetics*, *22*, 73–80. https://doi.org/10.1016/j.fsigen.2016.02.001

Saunders, C. P., Davis, L. J., & Buscaglia, J. (2011). Using Automated Comparisons to Quantify Handwriting Individuality. *Journal of Forensic Sciences*, *56*(3), 683–689. https://doi.org/10.1111/j.1556-4029.2011.01713.x

Schmittbuhl, M., Le Minor, J.-M., ... Schaaf, A. (1998). Shape of the piriform aperture in Gorilla gorilla, Pan troglodytes, and modern Homo sapiens: Characterization and polymorphism analysis. *American Journal of Physical Anthropology*, *106*(3), 297–310.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Sciacca, E., Langlois-Peter, B.-B., ... Velay, J.-L. (2011). Effects of different postural conditions on handwriting variability. *Journal of Forensic Document Examination*, *21*, 51–60.

Sciacca, E., Langlois-Peter, M.-B., ... Velay, J.-L. (2008). The range of handwriting variability under different writing conditions. *Journal of Forensic Document Examination*, *19*, 5–13.

Shiver, F. C. (2009). Intersecting Lines: Documents. In *Wiley Encyclopedia of Forensic Science*. American Cancer Society. https://doi.org/10.1002/9780470061589.fsa328

Sita, J., Found, B., & Rogers, D. K. (2002). Forensic handwriting examiners' expertise for signature comparison. *Journal of Forensic Sciences*, *47*(5), 1117–1124.

Snape, K. W. (1980). Determination of the Direction of Ball-Point Pen Motion from the Orientations of Burr Striations in Curved Pen Strokes. *Journal of Forensic Sciences*, *25*(2), 12142J. https://doi.org/10.1520/JFS12142J

Sognnaes, R. F., Rawson, R. D., ... Nguyen, N. B. (1982). Computer comparison of bitemark patterns in identical twins. *Journal of the American Dental Association (1939)*, *105*(3), 449–451. https://doi.org/10.14219/jada.archive.1982.0338

Srihari, S. N., Beal, M. J., ... Krishnamurthy, P. (2005). A statistical model for writer verification. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)* (pp. 1105–1109 Vol. 2). Seoul, South Korea: IEEE. https://doi.org/10.1109/ICDAR.2005.33

Srihari, S. N., Cha, S.-H., ... Lee, S. (2002). Individuality of handwriting. *Journal of Forensic Sciences*, *47*(4), 856–872.

Srihari, S., Huang, C., & Srinivasan, H. (2008). On the Discriminability of the Handwriting of Twins. *Journal of Forensic Sciences*, *53*(2), 430–446. https://doi.org/10.1111/j.1556-4029.2008.00682.x

Stockton, A., & Day, S. (2001). Bayes, handwriting and science. *Proceedings of the 59th Annual ASQDE Meeting - Handwriting Er Technology: At the Crossroads*,

1–10.

Stone, M. (2018). Gaius Verres Troubleshooter. In H. van der Blom, C. Gray, & C. Steel (Eds.), *Institutions and Ideology in Republican Rome* (1st ed., pp. 299–313). Cambridge University Press. https://doi.org/10.1017/9781108681476.017

Sulner, A. (2018). Critical Issues Affecting the Reliability and Admissibility of Handwriting Identification Opinion Evidence How They Have Been Addressed (or Not) Since the 2009 NAS Report, and How They Should Be Addressed Going Forward: A Document Examiner Tells All. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3062250

Taroni, F. (2005). Inadequacies of posterior probabilities for the assessment of scientific evidence. *Law, Probability and Risk*, *4*(1-2), 89–114. https://doi.org/10.1093/lpr/mgi008

Taroni, F., Biedermann, A., ... Aitken, C. (2014). *Bayesian networks for probabilistic inference and decision analysis in forensic science*. Chichester, United Kingdom; Hoboken, NJ: John Wiley and Sons.

Taroni, F., Biedermann, A., & Bozza, S. (2016). Statistical hypothesis testing and common misinterpretations: Should we abandon p-value in forensic science applications? *Forensic Science International*, *259*, e32–e36. https://doi.org/10.1016/j.forsciint.2015.11.013

Taroni, F., Bozza, S., ... Aitken, C. (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, *15*(1), 1–16. https://doi.org/10.1093/lpr/mgv008

Taroni, F., Bozza, S., & Biedermann, A. (2006). Two items of evidence, no putative source: An inference problem in forensic intelligence. *Journal of Forensic Sciences*, *51*(6), 1350–1361. https://doi.org/10.1111/j.1556-4029.2006.00272.x

Taroni, F., Bozza, S., ... Garbolino, P. (2010). *Data analysis in forensic science: A Bayesian decision perspective*. Chichester, United Kingdom; Hoboken, New York: John Wiley and Sons.

Taroni, F., Champod, C., & Margot, P. (1998). Forerunners of Bayesianism in Early Forensic Science. *Jurimetrics Journal*, *38*, 183–200. Retrieved from https://serval.unil.ch/notice/serval:BIB_8195

Taroni, F., Garbolino, P., ... Bozza, S. (2018). Reconciliation of subjective probabilities and frequencies in forensic science. *Law, Probability and Risk*, *17*(3), 243–262. https://doi.org/10.1093/lpr/mgy014

Taroni, F., Marquis, R., ... Bozza, S. (2012). The use of the likelihood ratio for evaluative and investigative purposes in comparative forensic handwriting examination. *Forensic Science International*, *214*(1–3), 189–194. https://doi.org/10.1016/j.forsciint.2011.08.007

Taroni, F., Marquis, R., ... Bozza, S. (2014). Bayes factor for investigative assessment of selected handwriting features. *Forensic Science International*, *242*, 266–273. https://doi.org/10.1016/j.forsciint.2014.07.012

Tavare, S., Balding, D. J., ... Donnelly, P. (1997). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1207814. *Genetics*, *145*(2), 505–518.

The Mathworks, I. (2019). MATLAB version 9.6.0.1072779 (R2019a). Natick, Massachusetts: The Mathworks, Inc.

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

Thiéry, A., Marquis, R., & Montani, I. (2013). Statistical evaluation of the influence of writing postures on on-line signatures. Study of the impact of time. *Forensic Science International*, *230*(1), 107–116. https://doi.org/10.1016/j.forsciint.2012.10.033

Thompson, W. C., Grady, R. H., ... Stern, H. S. (2018). Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk*, *17*(2), 133–155. https://doi.org/10.1093/lpr/mgy012

Thorndike, E. L. (1915). The Resemblance of Young Twins in Handwriting. *The American Naturalist*, *49*(582), 377–379. https://doi.org/10.1086/279488

Totty, R. N. (1991). Recent Developments in Handwriting Examination. In A. Maehly & R. L. Williams (Eds.), *Forensic Science Progress* (Vol. 5, pp. 91–128). Springer Berlin Heidelberg.

Turnbull, S. J., Jones, A. E., & Allen, M. (2010). Identification of the Class Characteristics in the Handwriting of Polish People Writing in English. *Journal of Forensic Sciences*, *55*(5), 1296–1303. https://doi.org/10.1111/j.1556-4029.2010.01449.x

van den Hout, A., & Alberink, I. (2016). Posterior distributions for likelihood ratios in forensic science. *Science & Justice*, *56*(5), 397–401. https://doi.org/10.1016/j.scijus.2016.06.011

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*(6), 1047–1056. https://doi.org/10.3758/s13423-012-0300-4

Vastrick, T. W., Schuetzner, E., & Osborn, K. (2018). Measuring the Frequency Occurrence of Handwritten Numeral Characteristics. *Journal of Forensic Sciences*, *63*(4), 1215–1220. https://doi.org/10.1111/1556-4029.13678

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. https://doi.org/10.1214/12-SS102

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Wagenmakers, E.-J., Lodewyckx, T., ... Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Weiss, G., & Haeseler, A. von. (1998). Inference of Population History Using a Likelihood Approach. *Genetics*, *149*(3), 1539–1546.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio. *Computational Statistics & Data Analysis*, *54*(9), 2094–2102. https://doi.org/10.1016/j.csda.2010.03.016

Wickham, H. (2011). Testthat: Get started with testing. *The R Journal*, *3*, 5–10. Retrieved from https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf

Willis, S., McKenna, L., ... Taroni, F. (2015). ENFSI guideline for evaluative reporting in forensic science. *European Network of Forensic Science Institutes*.

Yuan, Y., & Johnson, V. E. (2008). Bayesian hypothesis tests using nonparametric statistics. *Statistica Sinica*, *18*, 1185–1200.

Zahn, C. T., & Roskies, R. Z. (1972). Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, *C-21*(3), 269–281.

Zinn-Justin, J. (2002). *Quantum Field Theory and Critical Phenomena*. Oxford University Press.