

Statistica Applicata • Italian Journal of Applied Statistics

Volume 27 • Number 2 • 2015

Statistica Applicata

ISSN: 1125-1964

Thematic Issue
Forensic Statistics - I

**Frosini, B.V.,
Taroni, F.**
Co-Editors

Volume 27
Number 2
May – August 2015

**ITALIAN
JOURNAL
OF APPLIED
STATISTICS**

RCE
MULTIMEDIA

UNCERTAINTY IN FORENSIC SCIENCE: EXPERTS, PROBABILITIES AND BAYES' THEOREM

Franco Taroni¹, Alex Biedermann

School of Criminal Justice, The University of Lausanne, Lausanne, Switzerland

Abstract. *Limited and incomplete information represent recurrent constraints in many practical inference problems. Thus, categorical conclusions are unwarranted and assessing the probative strength of scientific results in the light of uncertainty represents the regular case in forensic science, forensic medicine and other disciplines. What is more, the interpretation of scientific results in applied contexts requires the construction of arguments in a balanced, logical, robust and transparent way. It is the duty of scientists to clarify the foundations of such arguments and to handle the possibly multiple sources of uncertainty in a rigorous and coherent way. There now is a widespread agreement among many committed scientists that these requirements are appropriately conceptualized as reasoning in conformity with the laws of probability theory and, derived from this framework, Bayes' theorem, which is central to the understanding of inferential reasoning. This paper is addressed to readers with knowledge of statistical concepts who seek a focused and general overview on how elements of probability and Bayesian inferential statistics can be meaningfully applied to help resolve questions of inference and decision-making at the intersection between forensic science and the law, and support scientists in their interaction with recipients of expert information and decision makers in the legal process.*

Keywords: *Forensic Science, Scientific Evidence, Probabilistic Inference, Bayes' Theorem.*

1. INTRODUCTION

Forensic science and forensic medicine rely on a body of scientific principles and technical methods to help with issues in legal proceedings, such as criminal, civil or administrative investigations. These disciplines seek to help demonstrate the existence or past occurrence of events of legal interest, such as a crime. Forensic science, in particular, assists the various participants in the justice system, such as investigators, public prosecutors and decision-makers at large, in examining events related to persons of interest and recovered traces. This may involve the analysis of the nature of fluids and other materials, such as textile fibres, glass and paint fragments, handwritings, as well as their classification in various categories.

¹ Corresponding author: Franco.Taroni@unil.ch

Forensic medicine, in turn, assists the judicial system by offering information in a variety of domains, such as the cause of death and the estimation of the age of living persons. More generally, forensic disciplines thus take a major interest in aspects such as the investigation of crimes and the direct examination of living or death persons (i.e., victims and suspects) and the vestiges of actions. In other words, forensic disciplines assist in the reconstruction of past events of judicial relevance that are *unknown* to us. Thus, the domain must deal with the fundamental notion of *uncertainty*. The natural response to uncertainty is the search of new data. Naturally, this involves the examination and comparative analysis of so-called ‘evidential material’ (i.e., DNA traces, toxic substances, crime scene findings, data imaging, etc.) followed by an assessment of the evidential strength of these scientific results within the particular context of the relevant event under investigation.

However, throughout the history of forensic science, including more recent periods, challenges arose from the discovery of cases of miscarriage of justice in which scientific findings played a major role. These cases generate a continuing and serious stream of debate about the status of some areas of forensic practice with respect to scientific standards of reliability. At the same time, numerous courts across legal systems have repeatedly emphasized the need for practicing scientists to continually monitor the performance of their domain of specialization. Most importantly, scientists need to scrutinize both the rationale underlying the various areas of practice and the ways in which scientific results are evaluated and presented in context. Today, many of the so-called traditional forensic identification practices (i.e., questioned documents, dentition, X-ray images, etc.) are systematically compared to purportedly better-founded and better-researched fields, in particular forensic DNA analysis, to point out the lack of fundamental research and the predominant reliance upon arbitrary expert opinions. Many legal and scientific researchers and practitioners invoke this observation to call for a revision of research agendas, towards a more systematic generation of data on agreed measurable factors and the development of sound probabilistic methods for evidence evaluation under uncertainty (Saks and Koehler, 2005).

This paper is structured as follows. Section 2 provides further arguments in support of the view that forensic science should openly acknowledge the existence of uncertainty as an inevitable feature of its area of practice and, arguably, that expressing statements of certainty with respect to events of interest should be avoided. Section 3 provides a more formal, but brief outline of the normative probabilistic perspective to dealing with uncertainty while leaving more detailed theoretical considerations to specialized literature on the topic. This discussion will include a general presentation of the Bayesian approach for evaluating forensic

results. Section 4 will rely on the notions introduced in the previous sections to illustrate the proposed probabilistic methodology based on a hypothetical case involving DNA for kinship analysis. Section 5 introduces probabilistic graphical models, that is Bayesian networks, illustrated further in Section 6 through an extended DNA relatedness case. Section 7 concludes the paper by underlying that the Bayesian framework satisfies several desiderata for evidential assessment in forensic science. The paper addresses a statistical readership that seeks a general introduction to elements of *forensic statistics*.

2. THE CORE NEEDS IN FORENSIC SCIENCE: REASONING UNDER UNCERTAINTY

The fundamental constraint in forensic science, in much the same way as in science in general, is that available information is limited and incomplete. This means that categorical conclusions about events of judicial interest are impossible. Reasoning in the light of uncertainty thus represents the regular case. The inevitability of uncertainty implies the necessity to determine the degree of belief that may be assigned to a particular uncertain event or proposition, such as ‘Is the suspect the donor of the recovered trace?’, ‘Is the toxic substance the cause of the patient’s death?’, etc. It is in this context that inferential sciences, including statistics, can offer a valuable and substantial approach. In particular, when the existence of uncertainty is recognized as an inherent aspect of a given inference problem, and a statistical approach is possible, then this approach represents a normative reference in that it captures and indexes uncertainty based upon a precise and logical line of reasoning (Lindley, 2014).

Scientific progress relies on past experience, but how exactly such experience is to be used to inform future directions and decision-making represents a fundamental challenge. On the basis of what one sees, combined with any existing knowledge, one seeks to assess, if possible in a quantitative way, one’s uncertainty about a particular event of interest, yet the reality is that this kind of reasoning for extending knowledge provides only an incomplete basis for a conclusion. It follows naturally from this that scientific discussions and public debates on science should focus on uncertainty explicitly: that is, our perspective should seek to distinguish between what is more likely and what is less likely, rather than attempt to endorse a concept of certainty that cannot be warranted by the limited and imperfect evidence that arises in practical proceedings. With that being said, the very relevant need of all science is a way to deal quantitatively with what is commonly known as the *probabilities of causes*, with the term cause being understood as an uncertain

proposition (e.g., ‘The suspect is the source of the DNA trace found on the victim’). The fundamental task thus consists of discriminating between events of interest, or causes (e.g., in forensic medicine, the patient’s cause of death), in the light of particularly acquired information (i.e., scientific findings) (Redmayne et al., 2011).

From a practical perspective, the ability to deal with reasoning under uncertainty represents a core aspect of operational procedures that seek to qualify as rational. By offering an explicit way for specifying and articulating uncertainties, they help recipients of expert information introduce results of scientific examinations into a coherent whole, along with multiple other items of evidence (Aitken et al., 2010). The latter aspect reveals a further level of challenge. Daily inference tasks encountered by investigators, scientists and other participants in legal proceedings (judges, prosecutors and lawyers) are characterized not only by single and isolated items of evidence, but multiple items associated with a possibly complicated mutual dependency structure. It is therefore natural to enquire about logical procedures that can deal with items of evidence that occur in combination and, in particular, the way in which multiple items of evidence stand in relation to each other. Such analyses reach further levels of complication essentially because they need to be conducted in the light of intricate frameworks of circumstances, that is situations involving many variables.

Probabilistic reasoning alone is not, however, an endpoint of forensic or medical applications in the legal process. Clearly, at the end of the day, *decisions* must be made. Once that uncertainty is recognized and formalized, the combination of uncertainty with the ultimate decision represents the core feature of legal proceedings. For example, a Court of Justice may have to *decide* if it finds a defendant guilty of the offence for which he has been charged (Kaye, 1998). While probabilistic reasoning under uncertainty can be considered a topic to be studied in its own right, systematic research on how probability is coherently applied in the wider context of rational decision-making under uncertainty, in particular with regard to forensic science applications, is still a largely unexplored field. This aspect is beyond the scope of this paper.

3. THE NORMATIVE APPROACH TO SCIENTIFIC INFERENCE: BAYES’ THEOREM

Probability constitutes a reference scheme for measuring uncertainty in any scientific and human endeavour (Oaksford and Chater, 2007). For the purpose of the current discussion, probability theory will be understood in its subjectivistic, also known as epistemic or personalistic, interpretation (Biedermann, 2015). This

view focuses on an individual's personal beliefs about a given event. This view is widely regarded as particularly useful, and by some even as the only meaningful conceptualization of probability, regardless of its application in the field of forensic science or everyday life in general. As noted in the previous sections, however, besides a measure of uncertainty as given through probability theory, it is equally necessary to re-assess probabilities given newly acquired data or, more generally speaking, acquired information. For this purpose, the logic of Bayesian procedures represents a primary choice (Howson and Urbach, 2005).

In essence, scientific reasoning can be considered as an instance of applying the laws of probability theory, with Bayes' theorem providing a solution to the general problem known as *induction*. This type of inference seeks to evaluate which hypothesis is most tenable in the light of what has been seen during examinations. Unlike deduction, which goes from a given postulate to potential observations (which can often be clearly articulated in a straightforward way), induction goes in the reverse direction, which is more challenging, by starting from particular findings to possibly multiple competing hypotheses. How exactly this reverse thinking process ought to be operated in a logically sound way is at the heart of Bayesian inference procedures.

To illustrate and state this more formally for a finite case setting, start by considering a set of mutually exclusive and exhaustive hypotheses or causes H_1, \dots, H_n and a set of experimental (scientific) results or outcomes, say E (short for 'evidence'). Further, let I denote the conditioning information. Bayes' theorem then says that the probability of a hypothesis of interest H_i , given E , is obtained as follows:

$$\Pr(H_i | E, I) = \frac{\Pr(E | H_i, I) \Pr(H_i | I)}{\Pr(E | H_1, I) \Pr(H_1 | I) + \dots + \Pr(E | H_n, I) \Pr(H_n | I)}. \quad (1)$$

In the forensic context, Bayes' theorem thus shows how the beliefs of any person (e.g., a scientist, a judge) required to make an inference based on new data, evolve. This leads one to posterior beliefs, that is a state of belief after that data has been acquired. Such Bayes' oriented reasoning is considered *normative* in the sense that it prescribes a standard that, if followed, allows reasoners to avoid logical fallacies. Stated otherwise, if one is committed to ensure a logically sound way of reasoning, it is in one's interest to conform to the Bayesian norm.

In legal contexts, it is common to consider the hypotheses of interest in pairs. For example, H_1 may denote the hypothesis representing the view of the prosecution, while H_2 denotes the hypothesis proposed by the defense. For such a situation, the

odds form of Bayes' theorem is appropriate. In this formulation of the theorem, H_2 denotes the complement of H_1 so that $\Pr(H_2 | I) = 1 - \Pr(H_1 | I)$. Then, the odds O in favor of H_1 are $\Pr(H_1 | I) / \Pr(H_2 | I)$, denoted $O(H_1 | I)$, and the odds in favor of H_1 given E are denoted $O(H_1 | E, I)$. The ratio form of Bayes' theorem then is:

$$O(H_1 | E, I) = \frac{\Pr(E | H_1, I)}{\Pr(E | H_2, I)} \times O(H_1 | I). \quad (2)$$

The left-hand side of this equation, $O(H_1 | E, I)$, is the posterior odds in favor of the prosecution hypothesis H_1 after the scientific evidence E has been obtained. The odds $O(H_1 | I)$, on the right-hand side, are the prior odds, that is before considering the observation E . The ratio of the two conditional probabilities $\Pr(E | H_1, I) / \Pr(E | H_2, I)$ is known as the *Bayes factor*, or *likelihood ratio*, and converts the prior odds to posterior odds. The Bayes factor can take values between 0 and ∞ and plays a major role in current forensic science thinking. Values greater than 1, for example, support the first proposition, here the prosecution's hypothesis (H_1). Values smaller than 1 support the alternative hypothesis, here the one of the defense (H_2). Scientific results E for which the Bayes factor takes the value 1 are said to be neutral. Such results do not allow one to discriminate between the two competing hypotheses under consideration. Stated otherwise, the scientific observation is equally likely under both hypotheses and, thus, does not allow one to discriminate between the propositions of interest.

In principle, this scheme of reasoning operates analogously in other scientific domains (Aitken and Taroni, 2004). There is no requirement for the hypotheses of interest to relate to events of judicial relevance. In a medical context, for example, H_1 can express the event that a given patient is affected by a given disease 1 and the alternative hypothesis H_2 is that the patient is affected by another disease 2. Similarly, one may be interested in the hypothesis that a victim deceased for reason 1 (H_1) compared to an alternative reason 2 (H_2). Let us exemplify this methodology through a routine case scenario.

Table 1: DNA profiling results (genotype) for the child, the mother and the alleged father at the two genetic markers THO1 and D3S1358. Numbers represent the alleles characterizing the genotype of a given person.

Evidence	Locus	Profiles		
		Child	Mother	Alleged father
E_1	THO1	6 – 6	6 – 7	6 – 9
E_2	D3S1358	18 – 19	16 – 19	18 – 18

4. A TYPICAL CASE IN FORENSIC SCIENCE: DNA PROFILING FOR EXAMINING RELATEDNESS

DNA profiling analyses performed on genetic markers, most often short tandem repeat (STR) markers (i.e. regions on DNA with polymorphisms that can be used to help discriminate between individuals), represents the standard approach to generate information that is relevant for studying various questions of relatedness. For each analyzed marker, the genotype is noted. A genotype is given by two alleles, one being inherited from the mother and the other from the father, although one cannot observe which is which. Suppose the use of such profiling analyses in a scenario involving a child, a mother and an alleged father. Propositions of interest in such scenario generally are:

- H_1 : the alleged father is the true father of the child,
- H_2 : the alleged father is not the true father; an unknown person is the genetical father of the child.

For shortness of notation, we leave aside background information I about the case. Next, consider two items of evidence E_1 and E_2 , representing DNA profiling results for the child, mother and alleged father at two genetic markers (*loci*), THO1 and D3S1358, respectively, as shown in Table 1.

The likelihood ratio, generally referred to as paternity index in the context of kinship analyses, for E_i is $\Pr(E_i | H_1) / \Pr(E_i | H_2)$, with $i = 1, 2$. To pursue this further, let G_C, G_M and G_{AF} denote the genotypes of the child C , the mother M and the alleged father AF , respectively, for evidence E_i . Let A_{Mi} and A_{Pi} denote the maternal and paternal alleles for evidence E_i . Let γ_{ij} be the occurrence (i.e., relevant population proportion) of allele j for evidence E_i .

For E_1 , the numerator of the likelihood ratio equals $\Pr(G_{C1} | G_{M1}, G_{AF1}, H_1) = 1/4$. That is, for parents with genotypes 6 – 7 and 6 – 9, respectively, the probability for their child to have genotype 6 – 6 is 1/4. The denominator equals $\Pr(G_{C1} | G_{M1}, G_{AF1}, H_2) = \Pr(A_{M1} | G_{M1}) \times \Pr(A_{P1} | H_2) = \Pr(A_{M1} = 6 | G_{M1} = 6 - 7) \times \Pr(A_{P1} = 6 | H_2) = (1/2) \times \gamma_{1,6}$. The likelihood ratio for the genetic marker THO1 is then $1/(2\gamma_{1,6})$. For E_2 , the numerator of the likelihood ratio equals $\Pr(G_{C2} | G_{M2}, G_{AF2}, H_1) = 1/2$, that is for parents with genotypes 16 – 19 and 18 – 18, respectively, the probability for their child to have genotype 18 – 19 is 1/2. The denominator equals $\Pr(G_{C2} | G_{M2}, G_{AF2}, H_2) = \Pr(A_{M2} | G_{M2}) \times \Pr(A_{P2} | H_2) = \Pr(A_{M2} = 19 | G_{M2} = 16 - 19) \times \Pr(A_{P2} = 18 | H_2) = (1/2) \times \gamma_{2,18}$. The likelihood ratio for D3S1358 is then $1/\gamma_{2,18}$.

Under the assumption of independence between E_1 and E_2 , the likelihood ratio for the combination of evidence (E_1, E_2) is

$$\frac{\Pr(E_1, E_2 | H_1)}{\Pr(E_1, E_2 | H_2)} = \frac{\Pr(E_1 | H_1)}{\Pr(E_1 | H_2)} \times \frac{\Pr(E_2 | H_1)}{\Pr(E_2 | H_2)} = \frac{1}{2\gamma_{1,6}} \times \frac{1}{\gamma_{2,18}}.$$

Assume further that $\gamma_{1,6} = 0.219$ and $\gamma_{2,18} = 0.1557$. Then $1/(2\gamma_{1,6}\gamma_{2,18}) = 14.7 \approx 15$. The evidence of the two marker systems, that is the DNA profiling results, is about 15 times more probable if the alleged father is the true father than if an unknown man is. Verbally stated, this result can be said to provide moderate support for the proposition that the alleged father is the true father² rather than an unknown and unrelated man.

In cases of alleged paternity, it is appropriate for a judge to consider the posterior probability that the alleged father is the true father, which is the probability that the main hypothesis H_1 is true. This probability is known as the *probability of paternity* and can be shown³ to be equal to:

$$\Pr(H_1 | E_i) = \left\{ 1 + \frac{\Pr(E_i | H_2)}{\Pr(E_i | H_1)} \times \frac{\Pr(H_2)}{\Pr(H_1)} \right\}^{-1}. \quad (3)$$

For illustration, suppose a case in which there is an alleged father and only one other man (of unknown DNA profile) who could be the true father, and that one holds equal probabilities – initially – for each of them being the true father⁴. Then $\Pr(H_1) = \Pr(H_2) = 0.5$ and

$$\Pr(H_1 | E_1) = 1/(1+2\gamma_{1,6}) = 1/(1+0.438) = 0.695.$$

Next, extend the considerations to E_2 . Invoking Equation (3), the posterior odds $\Pr(H_1 | E_1)/\Pr(H_2 | E_1)$ in favour of H_1 , given E_1 , now become the new prior odds so that the posterior probability for H_1 , given E_1 and E_2 , is given by

$$\begin{aligned} \Pr(H_1 | E_1, E_2) &= \left\{ 1 + \frac{\Pr(H_2 | E_1)}{\Pr(H_1 | E_1)} \times \frac{\Pr(E_2 | H_2)}{\Pr(E_2 | H_1)} \right\}^{-1} \\ &= \left(1 + \frac{0.305}{0.695} \times \frac{0.155/2}{1/2} \right)^{-1} \\ &= 0.936. \end{aligned}$$

² Note that, generally, forensic scientists use specific formulae for calculating the probability of DNA profiles for two related individuals under an assumption of independence of genes. Cases when the mother, alleged father and alternative father all belong to the same sub-population require that formulae incorporate additional factors such as the coancestry coefficient, often denoted F_{ST} in scientific literature. To ease notation, this aspect is not included in the development here.

³ For brevity, details of the derivation are left aside (Aitken and Taroni, 2004).

⁴ Notice that the default assumption $\Pr(H_1) = \Pr(H_2) = 0.5$ is unrealistic for many logical and practical reasons. It is generally recommended to assign this initial probability on the basis of the circumstances of the case at hand.

This result assumes again independence between E_1 and E_2 . The probability that the alleged father was the true father, the probability of paternity, was initially 0.5. After presentation of the THO1 evidence (E_1) it became 0.695. After the presentation of the D3S1358 evidence (E_2) it became 0.936. The effect on the posterior probability of altering the prior probability can be determined from equations (3) and (4). Examples of results are given in Table 2.

Table 2: Posterior probabilities of paternity for various prior probabilities for evidence for alleged father $E_1 = 6-9, E_2 = 18-18$.

$\Pr(H_1)$	0.5	0.25	0.1	0.01
$\Pr(H_1 E_1)$	0.695	0.432	0.202	0.023
$\Pr(H_1 E_1, E_2)$	0.936	0.831	0.620	0.129

This example of a paternity scenario portrays the general idea of how to reason coherently under situations of uncertainty. Many real case situations, however, can be much harder to solve, with no simple equations being available. This may be so because more information is available and so the scientist needs to know how the various components of this information interact with and relate to each other. But there may also be more uncertainty due to a lack of information. To help apply probability theory in such contexts, in particular assessing scientific findings and quantifying the probability of hypotheses of interest, graphical frameworks have been developed. So-called probabilistic graphical models, such as Bayesian networks (Bayes nets or BNs, for short).

The level additional complication that may readily be dealt with by Bayesian networks is best illustrated through an example. Suppose that there are two individuals (offspring, denoted here child c1 and child c2, respectively) who share the same two parents (mother m1 and father f). A third individual, say c3, known to have a mother m2 different from m1, is interested in examining the degree of relatedness with respect to c1 and c2 (e.g., half-sibship versus unrelated). Thus, f is considered as a putative father of c3. A particular complication of the scenario consists in the fact that f is deceased and unavailable for DNA profiling analyses. Such a questioned kinship case looks very challenging at first sight, but it can be studied through probabilistic graphical models which are ideally suited to manage the multiple items information. To point this out, we first introduce Bayesian networks in Section 5, and then apply them to this case in Section 6.

5. A PROBABILISTIC GRAPHICAL ENVIRONMENT: BAYESIAN NETWORKS

Bayesian networks combine elements from graph and probability theory. They pictorially represent the assumed dependencies and influences among the variables considered to be relevant for a particular inference problem. Variables of interest are represented by nodes while arcs are used to express assumed dependencies between variables. As a main feature, Bayesian networks allow their users to coordinate probabilistic inferences in different directions, that is, very generally speaking, from causes to effects *and* in the reverse direction, from effects to causes. In the context, this is also called bidirectional inference. This means that one can reason in the direction of a network's arc, which leads to an evaluation of the probability of a particular observational variable given the truth of certain conditioning propositions of interest. On the other hand, one can reason against the direction of an arc, which amounts to inductive inference about propositions of interest, based on particular evidence. These inferential properties find widespread interest in many areas where the study of deduction and induction through probability plays an important role. Typical examples include medical diagnosis but also forensic science (Taroni et al., 2014).

In a Bayesian network, nodes and edges are combined in order to form a directed acyclic graph (or DAG). To express the strength of the relationships between the variables, probability distributions are associated with each node. For discrete variables, this means that, for example, a variable B which has entering arcs from parents A_1, \dots, A_n , will have a conditional node probability table $\Pr(B | A_1, \dots, A_n)$. When a variable has no parents, a table containing unconditional probabilities is assigned. For example, probabilities $\Pr(A)$ will be assigned for a variable A that has no entering arcs from other nodes. In a Bayesian network with variables A_1, \dots, A_n , the joint probability distribution $\Pr(A_1, \dots, A_n)$ is given by the product of all specified conditional probabilities:

$$\Pr = (A_1, \dots, A_n) = \prod_i \Pr(A_i | par(A_i)) \quad (5)$$

where $par(A_i)$ represents the set of parental variables of A_i . Equation (5) is called the chain rule for Bayesian networks. It formally defines the meaning of a Bayesian network: the representation of the joint probability distribution for all the variables. Consider this rule in each of the three cases of basic sequential connections that can be made up with Bayesian networks (see Figure 1). For a path from A to C via B , as shown in Figure 1(i), $\Pr(A, B, C) = \Pr(A)\Pr(B | A)\Pr(C | A, B)$ can be reduced to $\Pr(A, B, C) = \Pr(A)\Pr(B | A)\Pr(C | B)$.

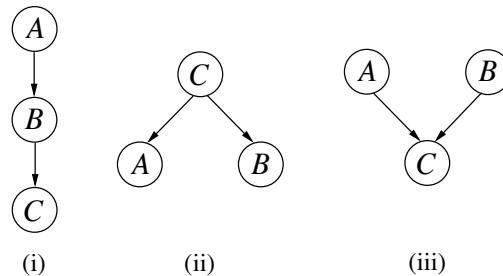


Figure 1: Serial (i), diverging (ii) and converging (iii) connections in Bayesian networks.

For a diverging connection, the joint probability can be written as $\Pr(C,A,B)=\Pr(C)\Pr(A|C)\Pr(B|C)$, whereas in a converging connection it would be $\Pr(A,B,C)=\Pr(A)\Pr(B)\Pr(C|A,B)$.

For building Bayesian networks, the analyst must keep in mind that the purpose is to represent features of a real-world problem. That is, through the use of a Bayesian network, one can graphically and numerically express one's understanding and perception of a real world system. To illustrate this, consider the network fragments shown in Figure 1, representing basic building blocs for encoding dependence and (conditional) independence assumptions that we may have with respect to particular features of an inference problem. Start by considering serial and diverging connections. In these types of connections, a path is said to be 'blocked' if the middle variable is instantiated⁵. To make this explicit, consider a serial connection where *A* represents the proposition 'suspect is the offender', *B* the proposition 'the blood stain found on the crime scene comes from the suspect', and *C* 'the suspect's blood sample and the blood stain from the crime scene share the same DNA profile'. In such a network fragment, the proposition *A* is relevant for *B* and so is *B* for *C*. However, given *B*, the cause of the presence of blood could be different from *A*. Next, consider a diverging connection. This type of connection is appropriate when one judges that knowledge about an event *A* is relevant for another event *B* and that knowledge about a third event *C*, which conditions both *A* and *B*, separates *A* from *B*. This holds, for example, in a case where *C* represents the proposition 'the suspect has assaulted the victim', *A* 'the bloodstain on the suspect's clothes comes from the victim' and *B* 'the bloodstain on the victim comes from the suspect'. In converging connections, a path is blocked for the flow of information as long as the intermediate variable, or one of its descendants, has *not* received evidence. For example, in a medical context, let *A* denote the proposition

⁵ A variable is called instantiated if its state is changed from 'unknown' to 'known'.

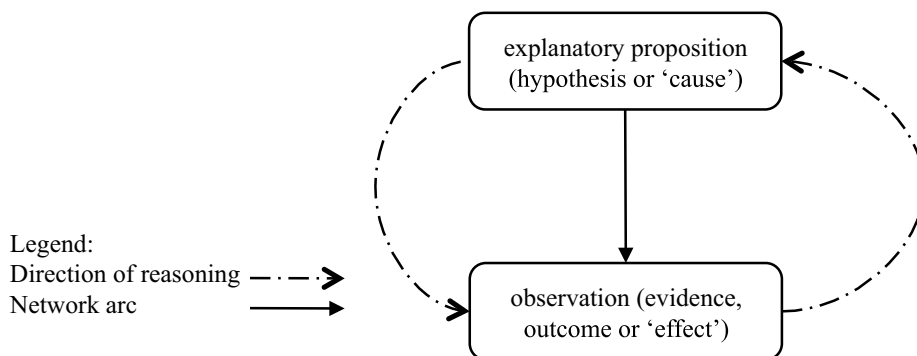


Figure 2: Different modes reasoning in Bayesian networks (prevision and diagnosis).

‘the patient has disease A ’ and B ‘the patient has disease B ’. Then, knowledge that one of these events occurred would *not* provide information about the occurrence of the other. However, if it becomes known that the intermediate variable C holds (e.g., the proposition ‘the patient shows scientific evidence C , such as fever’), then A and B become related.

Figure 2 illustrates the principal modes of probabilistic reasoning in Bayesian networks. For the ease of argument, interpret the network as a representation of relationships of cause and effect. Thus, let us say that causes ‘produce’ the effects, that is, knowing that a cause occurred, it can be foreseen that the effect will occur or might probably occur, too. This kind of reasoning is also known as prevision. For example, if a person of interest (e.g., a suspect) is the source of a crime stain (represented, for instance, in terms of a variable S), then we might expect that laboratory analyses find corresponding DNA profiles between reference material from the suspect and the crime stain. Such a scientific result can be represented, for example, in terms of a second variable E . In this network, with structure $S \rightarrow E$, if S is known, then the probability of the child variable E will be given by the probability $\Pr(E \mid S)$ as specified in the conditional node table of E . Note that, usually, this probability is taken to be considerably larger than the probability for a correspondence if the suspect were not the source of the crime stain: $\Pr(E \mid \bar{S}) < \Pr(E \mid S)$.

Note, however, that the effect does not ‘produce’ the cause. Instead, knowing that the effect occurred, one may *infer* that the cause probably occurred. This is a line of reasoning against the causal direction (i.e., a network’s arc), referred to as ‘diagnostic’. For example, on the basis of analytical results that reveal the same DNA profile for the stain and the reference material from the suspect (represented by variable E), one’s belief in the proposition that the suspect is the source of the

stain (variable S) should increase. This is so whenever $\Pr(E | \bar{S}) < \Pr(E | S)$, that is the finding of corresponding DNA profiles is more probable under the proposition S (i.e., the suspect is the source of the crime stain) rather than under the alternative proposition \bar{S} (i.e., an unknown person is the source of the crime stain). In a medical context, an example of using the network structure $S \rightarrow E$ could be analyses (i.e., diagnostic testing) showing that the patient's blood contains a certain quantity of a given target substance (represented by variable E), which is used as a basis to revise one's belief in the proposition that the patient has a particular disease (conditioning variable S). Within Bayesian networks, such a revision of belief is operated according to Bayes' theorem (Equation (1)). Note that such computations are possible over much larger network structures that considered in the general presentation given in this Section. An example of an extended network is pursued in Section 6.

6. BAYESIAN NETWORKS FOR KINSHIP ANALYSES USING DNA PROFILING RESULTS

Consider again the scenario introduced at the end of Section 4. There are two individuals child $c1$ and child $c2$ who share the same two parents (mother $m1$ and father f). A third individual, child $c3$, known to have a mother $m2$ different from mother $m1$, seeks to investigate the degree of relatedness with respect to individuals $c1$ and $c2$ (i.e., half-sibship versus unrelated). Father f is considered as a putative father of $c3$, but, unfortunately, f is deceased and unavailable for DNA profiling analyses. To approach this case through a Bayesian network, start by considering basic sub-models to deal with genetical characteristics, as shown in Figures 3 (i) and (ii).

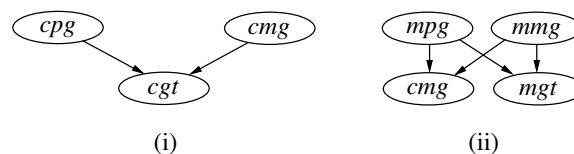


Figure 3: Bayesian network fragments, from methodology described in Dawid et al. (2002), representing (i) a child's genotype, *cgt*, with *cpg* and *cmg* denoting, respectively, the child's paternally and maternally inherited genes, and (ii), a child's maternal gene, *cmg*, reconstructed as a function of the mother's paternal and maternal genes, *mpg* and *mmg*, respectively. The states of gene nodes represent the different forms (i.e., alleles) that a genetic marker can assume whereas the states of the genotype nodes regroup pairs of alleles.

For sake of illustration, consider the network fragment (ii) displayed in Figure 3. Let us suppose that the node *mpg* (short for ‘mother paternal gene’) covers the states 6, 7 and x , representing the number of short tandem repeats (STR) at the locus THO1, with x summarising all alleles other than 6 and 7. The unconditional probabilities required for the various states of the node *mpg* are assigned on the basis of the relevant allelic population proportions, obtained from databases or scientific literature. The same definition applies for the node *mmg* (short for ‘mother maternal gene’). Next, for each marker included in the analysis of our complex scenario, the genotype is recorded. The latter consists of two genes, one being inherited from the mother and the other from the father (although one cannot observe which is which). The Bayesian network fragment in Figure 3 (ii) captures an individual’s genotype for a given marker and the transmission of alleles to a descendant (child).

Based on these considerations, the network in Figure 4 can be constructed to describe the scenario under investigation.

This network can accommodate DNA profiling results for a single marker. The structure of this model is the result of a logical combination of submodels that themselves may be a composition of model fragments. Examples of submodels are shown in Figure 4 using rounded boxes with dotted lines (other submodels may be chosen). The submodel (a) represents the genotypes of the individuals *c1* and *c2*, conditioned on the genotypes of the undisputed parents *m1* and *f*. This submodel is itself a composition of the repeatedly used network fragment described in Figure 3 (i). The same network fragment is invoked to implement the genotype of the

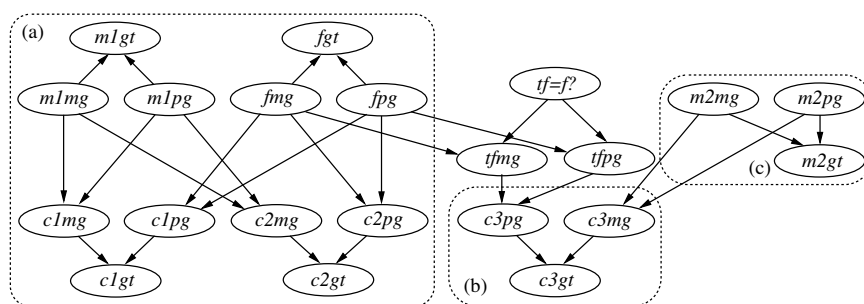


Figure 4: Bayesian network for evaluating DNA profiling results in a case of questioned kinship. Nodes c, f, m (in the first place) and $t f$ denote, respectively, child, father, mother and true father. Nodes with names ‘...mg’ and ‘...pg’ denote, respectively, an individual’s maternally and paternally inherited genes. Nodes with names ‘...gt’ represent an individual’s genotype. The node $t f = f?$ is binary with values ‘yes’ and ‘no’ in answer to the question whether the undisputed father f of the children $c1$ and $c2$ is the true father of the child $c3$.

individuals $c3$ (submodel (b)) and $m2$ (submodel (c)). A subtle constructional detail concerns the connection between the two submodels (a) and (b). As there is uncertainty about whether f is the true father of $c3$, the paternal gene of $c3$, $c3pg$, is not directly conditioned on f 's parental genes (that is, nodes fmg and fpg). Such uncertainty is accounted for through a distinct node $tf=f?$ that regulates the degree to which f 's allelic configuration is allowed to determine c 's true father's parental genes, represented here by nodes $tfmg$ and $tfpg$.

The above considerations clearly illustrate that Bayesian network models are highly versatile and can deal with a variety of aspects that affect the coherent evaluation of scientific results. This includes partial evidence and additional complications such as genetic mutation (Dawid et al., 2002). This explains why the use of Bayesian networks for studying the assessment of the weight of scientific evidence in forensic science is a lively area of research, in particular for DNA profiling results (Biedermann and Taroni, 2012).

7. CONCLUSIONS

We have emphasized in this paper that limited and incomplete information represent recurrent constraints in many practical inference problems. Forensic science and legal medicine provide a strong case for this because of the paucity and limitations of the trace material, that is potential evidence, arising in real cases. Thus, the interpretation of scientific results in context must deal with uncertainty and requires the construction of arguments in a balanced, logical, robust and transparent way (Jackson, 2000). Inferential disciplines, in particular statistics, offer sound frameworks – in particular the Bayesian programme – that provide scientists with a proper approach to this challenge. Although the practical implementation of this perspective may not be straightforward in some instances, there now exist sophisticated frameworks, such as Bayesian networks, available also in both commercially and academically distributed software environments. These support the transition from theoretical analyses to operational applications. The advent of such computational support for implementing probabilistic reasoning in practice has opened many new areas of fundamental research in forensic science. Today, scientists have never been in a better position to invoke the logical framework of probabilistic reasoning when they are required to explain in a clear and explicit way how they have proceeded in solving intricate inferential problems and arrived at their conclusions. This represents an important argument in favour of the requirement of disclosing the rationale behind the work of forensic experts.

ACKNOWLEDGEMENTS

Alex Biedermann gratefully acknowledges the support of the Swiss National Science Foundation through grant No. BSSGI0_155809.

REFERENCES

- Aitken, C.G.G., Roberts, P. and Jackson, G. (2010). *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings (Practitioner Guide No. 1), Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*. Royal Statistical Society's Working Group on Statistics and the Law, London.
- Aitken, C.G.G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley and Sons, Chichester.
- Biedermann, A. (2015). The role of the subjectivist position in the probabilization of forensic science. In *Journal of Forensic Science and Medicine*. **1**: 140–148.
- Biedermann, A. and Taroni, F. (2012). Bayesian networks for evaluating forensic DNA profiling evidence: A review and guide to literature. In *Forensic Science International: Genetics*. **6**: 147–157.
- Dawid, A.P., Mortera, J., Pascali, V.L. and van Boxel, D. (2002). Probabilistic expert systems for forensic inference from genetic markers. In *Scandinavian Journal of Statistics*. **29**: 577–595.
- Howson, C. and Urbach, P. (2005). *Scientific Reasoning: The Bayesian Approach*. Open Court, Chicago.
- Jackson, G. (2000). The scientist and the scales of justice. In *Science and Justice*. **40**: 81–85.
- Kaye, D.H. (1998). What is Bayesianism? In P. Tillers and E.D. Green, eds., *Probability and Inference in the Law of Evidence, The Uses and Limits of Bayesianism (Boston Studies in the Philosophy of Science)*, 1–19. Springer, Dordrecht.
- Lindley, D.V. (2014). *Understanding Uncertainty*. John Wiley and Sons, Hoboken.
- Oaksford, M. and Chater, N. (2007). *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford University Press, Oxford.
- Redmayne, M., Roberts, P., Aitken, C.G.G. and Jackson, G. (2011). Forensic science evidence in question. In *Criminal Law Review*, 347–356.
- Saks, M.J. and Koehler, J.J. (2005). The coming paradigm shift in forensic identification science. In *Science*. **309**: 892–895.
- Taroni, F., Biedermann, A., Bozza, S., Garbolino, P. and Aitken, C.G.G. (2014). *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science*. John Wiley and Sons, Chichester.