EMERGING ISSUES
IN THE NATURAL AND
APPLIED SCIENCES

Academic book

Emerging issues in the natural and applied sciences. "Progress" LLC, Baku, 2011, 308 p.

This research book can be used for teaching modern science, including for teaching undergraduates in their final undergraduate course in all fields of natural and applied sciences. More generally, this book should serve as a useful reference for academics, sciences researchers.

TABLE OF CONTENTS

FORENSIC IDENTIFICATION
WITH BAYES' LAW

Dr. **Marina Andrade,** Prof. Dr. **Manuel Alberto M. Ferreira**

Instituto Universitário de Lisboa (ISCTE-IUL),
UNIDE - IUL, Lisboa **(PORTUGAL)**
E-mails: marina.andrade@iscte.pt, manuel.ferreira@iscte.pt

ABSTRACT

The probabilistic fundaments for the evaluation of DNA evidences, based mainly on Bayes' Law, are presented emphasizing the famous island problem. Identification problems are considered, in particular the paternity dispute problem and the crime scene analysis.

Key words: DNA evidence, island problem, paternity dispute, crime scene analysis.

1. INTRODUCTION

It is intended in this section to consider a forensic identification problem in generic terms. Having been committed a crime, the accused identification may be based in accordance with the characteristics that are known to belong to the potential (or true) criminal. The identification may be done through fingerprints, by witnesses, for the type of the used weapon or by the DNA profile (Dawid and Mortera, 1996). For the determination of the DNA fingerprint or of the DNA profiling, since 1984, the forensic scientists used the genetic material (DNA) information as evidence in murder and/or rape crimes, as well as in paternity identification cases.

When the equality of profiles is measured, it is also measured indirectly the prosecuted culpability, since that calculation assumes that the accused is not guilty (Berry, 1991). If the two profiles are not alike it is possible to state that the accused is not the criminal. But,

when the similarity is declared it is not possible to state that the accused is the criminal.

In this context, a natural way to quantify the evidence weight in favour of the accused guiltiness, is to determine the conditional probability of the guiltiness given the evidence, that implies necessarily the application of Bayes' Law (Dawid and Mortera, 1996).

So two hypotheses must be evaluated, one of the defence and the other of the prosecution, i. e.:

H_D: The accused and the criminal are distinct persons

versus

H_P: The accused and the criminal are one and the same person.

In order to use the Bayes' Law suppose that the evidence (called *E*), in the crime scene, states that the criminal has a determined characteristic *C*, that is not for itself incriminating, being *p* the probability of any individual from the population to possess that characteristic. It is known that the accused individual possesses the characteristic (evidence), and be π the probability of any individual from the population, randomly chosen, to be guilty, having in mind the whole remaining evidence, for which it is admitted to be conditionally independent from the identification evidence, given the innocence or the guiltiness.

Be $C_a = 1$ and G_a the events: "The accused possesses the characteristic *C*" and "The accused is guilty", respectively, being C_a a binary random variable that indicates the characteristic *C* presence or absence in the accused individual. By the evidence *E* the criminal possesses *C*. If the accused is the criminal then $P(C_a = 1|G_a) = 1$ and $P(C_a = 1|\overline{G_a}) = p$, the likelihood ratio favouring the guiltiness is p^{-1} and the individual, that is under accusation, guiltiness *posterior* probability is

$$P(G_a|C_a = 1, E) = \frac{P(C_a = 1|G_a)P(G_a|E)}{P(C_a = 1|G_a)P(G_a|E) + P(C_a = 1|\overline{G_a})P(\overline{G_a}|E)}$$

and so

$$P(G_a|C_a = 1, E) = \frac{\pi}{\pi + p - \pi p} \quad (1.1).$$

2. ALGEBRAIC APPROACH

The Island Problem

The forensic scientists, to quantify the evidence weight in favour of the accused guiltiness, developed a theoretical formulation in order to quantify the H_P probability, given the presented evidence.

The island problem was extensively studied by Lindley (1987), Yellin (1979), Eggleston (1983), Dawid (1984) and Balding and Donnelly (1995) and is as follows:

Consider a isolated island, from the rest of the World, at which a murder was performed. Suppose that in the island there are $N + 1$ inhabitants, from which one is the criminal and the remaining ones are not guilty. Each one of the inhabitants has equal probability of being the criminal. Let us still admit that each individual has a probability p of possessing a certain characteristic $C,$ that is independent for two any distinct individuals.

The evidence found at the crime scene may be synthesized in the allegation of that the criminal possesses the characteristic C. So designate by E the evidence that represents the event "the criminal possesses the characteristic C, i. e., C_c = 1". An individual is arrested and presented to the court. The only evidence against him is that he possesses C, event already designated by C_a = 1. The interesting question is to determine the guiltiness probability, i. e., of the accused to be guilty, given the available evidence.

Following the Bayesian approach it is admitted *a priori* that any inhabitant has equal probability, $\pi = \frac{1}{N+1}$, of being the criminal. Substituting in the equation (1.1) it is obtained:

$$P(G_a|C_a = 1, E) = \frac{1}{1 + Np} \quad (2.1).$$

This result is easy to explain in court. In the island there is a criminal that possesses the characteristic *C* and *Np* not guilty persons that also possess *C*. Based on the available information the accused is among the *N* + 1 individuals, among which is also the criminal.

The Search Strategies

In the former analysis of the island problem it was assumed that the accused was identified in a random way among the population, and it is verified that possesses the characteristic *C*. Such a supposition is not always very real. In fact, sometimes is more easily accepted that it is identified through a search proceeding in the population. Being known that the criminal possesses that characteristic *C* it is natural that the police proceeds to a random search in the *N* + 1 individuals of the island, and arrest the first that has *C* presented.

So it is important to evaluate the effect caused by the search proceeding in the probability of guiltiness of the individual that is going to be accused. It is still admitted that to possess the characteristic *C* it is no enough to be considered guiltiness and that distinct individuals possess *C* independently of each other, with probability *p*. Suppose that is performed a search in the population, that may be completely random, either in some way deterministic or in other way stochastic, that is independent of the *C* values and of the criminal identity.

The way the search is performed is not relevant if it is assumed the permutation of the random variables C_i, that is, if it is admitted that the join distribution of any vector of *n* individuals, $\{C_1, ..., C_n\}$ is the same for any set with *n* elements, i.e., $P(C_1, ..., C_n) = P(C_{\psi(1)}, ..., C_{\psi(n)})$ for any permutation ψ of *n* individuals. In this way, the whole strategies are equivalent. Even, to simplify may be admitted that the population is numbered from 1 till *N* + 1, and the search is made by order.

Keeping the notation be C_i = 1 and G_i the event "the individual *i* possesses the characteristic C and the individual *i* is guilty", respectively. In this way, the event "the accused is guilty" will be

designated by $G_{a,k}$. Note that with this strategy $C_a = C_k = C_{a,k} = 1$ is redundant, because the accused was chosen for having characteristic *C*. So the events "the accused has the order *k* and is guilty" and "the accused has the order *k* and the characteristic *C* are $G_{a,k}$ and $C_{a,k} = 1$, respectively.

Two Possibilities

In the exposed context two perspectives are posed. First, given the evidence collected in the crime scene, *E*, that is the guilty person has characteristic *C*, if it is known *K = k* it is concluded that the *k* – 1 first individuals do not possess *C*, and it is possible to think that the individual *k* was randomly chosen among the remaining *N* – *k* + 2 inhabitants of the island, form which one is the criminal and *N* – *k* + 1 are not guilty and may exhibit the characteristic.

So, following the Bayesian argument:

$$P\big(G_{a,k}|E, C_{a,k} = 1, K = k\big) = \frac{1}{1 + (N - k + 1)p} \qquad (2.2).$$

It corresponds to the assumption that the *a priori* guiltiness probability is $\pi = \frac{1}{N-k+2}$. For $k > 1$ the value of this probability is greater than the one given by (1.1).

Consider now that *K* is unknown, that is, it is done a search in the population till finding the first individual that exhibits *C*, that so it is accused. But it is not known how many individuals were observed till finding the accused.

To calculate the accused guiltiness probability conditional to that he is the kth individual to be analyzed and possesses *C*, the whole possible values that *K* may assume must be considered. Then the whole disjoint partitions operated by *k* of the event "G_k and *K = k*" must be taken in account and to sum in its domain, that is the Total Probability Law must be applied. In consequence:

$$P(G_a|C_a = 1, E) = \sum_{k=1}^{N+1} P(G_{a,k}, K = k|C_{a,k} = 1, E)$$

$$= \frac{1 - (1 - p)^{N+1}}{(N + 1)p} \qquad (2.3).$$

This result obtained by Dawid (1994) shows that the Bayesian approach, given by equation (1.1), is not always valid. Indeed, if the accused is found through a search proceeding in the population, the individual guiltiness probability must not be calculated using Bayes' Law; this is appropriated when the suspect is determined in a random way. But the Bayes' Law fails by precaution, what in general is recommended following the innocence presumption.

The interest in the determination of the guiltiness probability of an accused individual is in general caught in the legal context having, because of it, the preoccupation that that result is always advantageous to the accused, that is so conservative.

Note that the former result had been already presented by Yellin (1979) as solution for the island problem. In the context followed by Yellin the result fails for not considering the additional information that the criminal possesses the characteristic *C*.

3. DNA EVIDENCE ALGEBRAIC APPROACH

In a DNA forensic identification simple problem one is in front of a biological trace, of unknown origin, and still biological "information" of unknown individuals based on the collected information it is intended to investigate which is the origin of that trace.

3.1. Paternity Dispute

It matters to think that in any paternity investigation action what firstly is in cause is the official recognizing of the individual parentage whose paternity is being searched, respecting his/her elementary rights.

In a simple way the information that generally is presented in a court decision, that may be called value of the biological evidence *W,* to be evaluated demands the knowledge of a classification table

– the Hummel table[1]. When there is not paternity exclusion the laboratories calculate the value of the paternity probability – *W*.

The laboratories proceeding results from the admission that the paternity and non paternity *a priori* probabilities are equal, $P(H_P) = P(H_D) = 0.5,$ and so the expression of the paternity *a posteriori* probability is $W = \frac{X}{X+Y}$, where *W* is the paternity probability, *X* is the probability that the biological father is the putative father and *Y* is the probability that the biological father is an individual picked up randomly from the population. This formula is also known as Essen-Möller equation[2].

Now the hypotheses outlined in the Introduction may be mathematically detailed. Becoming the assumption of paternity litigious, in formal terms are constituted the prosecution and the defence hypotheses, i.e.,

H_P: The putative father is the true father.

versus

H_D: The true father is another individual, chosen randomly among the population, not related with either the mother or the putative father.

Indeed the court has to decide about the child true paternity in relation with the putative father. In fact, it has to evaluate the disputing hypotheses ratio:

$$\frac{P(H_P|E)}{P(H_D|E)} = \frac{P(E|H_P)}{P(E|H_D)} \times \frac{P(H_P)}{P(H_D)}$$

being *E* the available evidence vector. Admitting that $P(H_P) = P(H_D),$

---

[1] Table of verbal predicates used and indicated by the laboratories to the court to classify the value obtained in laboratorial analysis. Nowadays it is often questioned.
[2] Obtained by Essen-Möller in 1938.

$$\frac{P(H_P|E)}{P(H_D|E)} = \frac{P(E|H_P)}{P(E|H_D)},$$

and in any paternity case at which it is possible to get mother genetic information (*mgt*), child genetic information (*cgt*) and putative father genetic information (*pfgt*), the forensic biology laboratories may compute the likelihood ratio (in the behind mentioned notation $\frac{X}{Y}$).

Likelihood Ratio Calculus in Paternity Dispute

In the hypotheses above considered, and disposing of the triplet $E = (mgt, cgt, pfgt)$, assuming the independence between and across markers it is possible to calculate the *LR* value for each one of the markers separately. And then to multiply those values to obtain the value related to the whole observed markers.

So consider that for a certain marker there is the triplet formula $E = \{(A, B); (B, B); (A, B)\}$, and p_A and p_B are the *A* and *B* allele frequencies, respectively, in the population for the considered marker. It is easily acceptable that, before knowing the child genetic information, it may be assumed that the true father identity is independent from the mother and from the putative father.

So it must be determined the conditional probability of the child genotype given the other two available genotypes. Consequently, to determine $P(E|H_P)$ it is only necessary to apply the Mendel's Laws. To calculate $P(E|H_D)$ the markers allelic frequencies must be known. So

$$P(E|H_P) = P[(mgt, cgt, pfgt)|(mgt, pfgt)]$$
$$= P[cgt = (B, B)|mgt = (A, B), pfgt = (A, B)]$$
$$= 0.5 \times 0.5$$

and

$$P(E|H_D) = P[(mgt, cgt, pfgt)|(mgt, rgt)]$$
$$= P[cgt = (B, B)|mgt = (A, B), rgt] = 0.5 \times p_B$$

where *rgt* is related to the genotype of an individual chosen randomly in the population not related either with the mother or with the putative father. The genotype *rgt* is not known. It is only known that the child inherits the allele *B* from the father, and so inherits with the frequency if this allele in the population. In consequence

$$LR = \frac{P(E|H_P)}{P(E|H_D)} = \frac{0.5}{p_B}.$$

It is important to emphasise that the decision on the child paternity, for any case presented in court, is taken by the judge. Its decision results from the knowledge taken from the whole relevant evidences and facts for each case in appreciation, each one with its own particularities.

This kind of evidence initiated a revolution in the form of appreciating the attribution or not of the paternity. But today it is accepted without any hesitation in court.

3.2. A Complex Mixture Case

The use DNA profiles and quantitative evidences in the appreciation of criminal cases has become more and more trivial, after 1985. But the knowledge of the cases ate which this tools are misunderstood and generate difficulties during the process are still common, contrarily to the desired. The case considered reports to a crime that was committed, in which two persons were violently murdered (Andrade and Faria, 2006). In what follows they will be designated by V_1 (female) and V_2 (male). V_1 and V_2, married, were at home and were murdered in unknown circumstances. During the investigations the police identified an individual, S_1, son of the victims that confessed to the crime, having been signed another individual as being involved in the crime, S_2.

About the Non-DNA Evidence

Of course it is not possible to specify the whole reasons, connected to non-DNA evidences, that driven the police to the sus-

pects, even because it overcomes this work scope. About the non-DNA evidence the interesting elements may be synthesized as follows:

    - There were found some inconsistencies and contradictions during the interrogatories;
    - The testimony of someone that said to have seen two human figures going away from the victims' car, in the morning at which the bodies were found;
    - The coroner stated, after the autopsy, that the wounds in the body of V_2 almost surely had not been inflicted for only one person.

The DNA Evidence

In what concerns DNA evidences, it was found a mixture in the crime scene. The analysis of haematic evidences allowed to identify genetic material matching to the victims', to S_1 the victims' son[3] and to S_2.

The Hypotheses to Consider

The interpretation of a mixture depends very much on the circumstances surrounding the crime, (Evett and Weir, 1998). This is a complex case, but the confession of S_1 allowed some simplification in the case, and consequently in the hypotheses to consider. After the confession of S_1 there is only a suspect: S_2. So, the testing hypotheses are:

H_P: The crime sample contains DNA from both victims (the S_1 profile results from a combination of the victims profiles and, in consequence, it is necessarily in the mixture) and from the suspect S_2.

---

[3]The victims' son could not be excluded, because any individual has in its genetic heritage, for each gene, a part inherited from the father and another inherited from the mother. So if the parents are in the mixture composition their descendents cannot be excluded.

versus

H_D: The crime sample contains DNA from both victims (the S_1 profile results from a combination of the victims profiles and, in consequence, it is necessarily in the mixture) and from an unknown individual.

Model and Analysis

As it was already referred, in front of each case the judge or the jury has to decide. To do so it matters to evaluate the whole evidences (now designated as DNA evidence – *E* and non-DNA evidence – *I*) and then to determine the *a posteriori* guiltiness probability given the evidences, i.e., has to answer the question:

$$P(H_P|E,I)?$$

Before going on with the analysis some notation must be introduced. So

- E_C is the sample collected in the crime scene,
- G_{V_1} is the genotype of the victim V_1,
- G_{V_2} is the genotype of the victim V_2,
- G_{S_1} is the genotype of S_1,
- G_{S_2} is the genotype of S_2.

The likelihood ratio for the mixture may be written in the form:

$$LR = \frac{P\left(E_C|G_{V_1},G_{V_2},\left(G_{S_1}\right),G_{S_2},H_P,I\right)}{P\left(E_C|G_{V_1},G_{V_2},\left(G_{S_1}\right),H_D,I\right)}.$$

As S_1 is necessarily in the mixture it was placed, in the above formula, between parenthesis.

The numerator has the value 1 since under the prosecution hypothesis that DNA mixture comes from the victims and the S_2, and so the presence of S_1 cannot be excluded. In the denominator the

probability of that collected DNA mixture to have contributions from an unknown individual, not related with the victims, must be calculated not excluding the possibility of having the contribution of S_1. So

$$LR = \frac{1}{P(E_C|G_{V_1}, G_{V_2}, (G_{S_1}), H_D, I)}.$$

Dealing any practical case the appropriate should be to present the synthesis of the impact of the evidence, that is the *LR* value and to let the judge to determine its *odds a posteriori*, combining the *odds a priori* with the supplied *LR*. But, not being common the Bayesian methodology use among those who deal with Law and the courts, in general, that does not happens.

4. FINAL REMARK

The proceedings described in this work are often very complicated to implement analytically. A way to overcome these difficulties is to use the so called Probabilistic Expert Systems (PES) to model the DNA interconnections, genetic and others, allowing the application of Bayes' Law in very complicate contexts.

To perform the final calculations it must be used computational software. One of the most used is Hugin[4].

To see practical applications see, for instance, references 8-30.

REFERENCES

1. A. P. Dawid and J. Mortera. Coherent use of identification evidence, Journal of the Royal Statistical Society, series B, 58, 425-443 (1996).
2. D. A. Berry. Inferences using DNA profilling in forensic identification and paternity cases, Statistical Science, 6, 175-205 (1991).

---

[4] www.hugin.com.

3. D. V. Lindley. The probability approach to the treatment of uncertainty in artificial intelligence and expert systems, Statistical Science, 2, 17-24 (1987).

4. J. Yellin. Review of Evidence, Proof and Probability (by R. Eggleston), Journal of Economic Literature, 17, 583-584 (1979).

5. R. Eggleston. Evidence, Proof and Probability, Weidenfeld & Nicholson, London (1983).

6. A. P. Dawid. The island problem: coherent use of identification evidence. Aspects of Uncertainty: a Tribute to D. V. Lindley. eds P. R. Freeman and A. F. M. Smith, ch 11, 159-170. Chichester:Wiley (1994).

7. D. Balding and P. Donnelly. Inference in forensic identification (with discussion), Journal of the Royal Statistical Society, series A, 158, part 1, 21-53 (1995).

8. M. Andrade and J. P. Faria. DNA evidence: a mixture sample. In Maria de Fátima Salgueiro, Maria João Lopes e António Teixeira, eds.: Temas em Métodos Quantitativos 5. 133-139, Edições Sílabo, (2006).

9. M. Andrade and M. A. M. Ferreira. Paternities Search in a very uncommon situation through object-oriented Bayesian networks, Journal of Mathematics and Technology, (2011) Forthcoming.

10. M. Andrade and M. A. M. Ferreira. Some considerations about forensic DNA evidences, International Journal of Academic Research, 3 (1,Part I), 7-10, (2011).

11. M. Andrade and M. A. M. Ferreira. Paternities Search with Object-Oriented Bayesian Networks, Aplimat-Journal of Applied Mathematics, (2011). Forthcoming.

12. M. Andrade and M. A. M. Ferreira. Paternities Search with Object-Oriented Bayesian Networks", *10th International Conference-APLIMAT 2011.Proceedings*, 1 (1), 1409-1418, (2011).

13. M. Andrade and M. A. M. Ferreira. Solving civil identification cases with DNA profiles databases using Bayesian networks, Journal of Mathematics and Technology, 1 (2), 37-40, (2010).

14. M. Andrade and M. A. M. Ferreira. Civil Identification Problems with Bayesian Networks using Official DNA Databases, Aplimat- Journal of Applied Mathematics, 3 (3), 155-162, (2010).

15. M. Andrade and M. A. M. Ferreira. Janus Probability two faces in court, Statistical Review-Journal of the Greek Statistical Association, (2010) Forthcoming.

16. M. Andrade and M. A. M. Ferreira. Evaluation of Paternities with less usual Data using Bayesian Networks, *IEEE Xplore (BMEI 2010 IEEE Catalog Number CFP1093D-PRT)*, 10 (93), 2475-2477, (2010).

17. M. Andrade, M. A. M. Ferreira, D. Abrantes, M. L. Pontes, and M. F. Pinheiro. Object-oriented Bayesian Networks in the evaluation of paternities in less usual environments, Journal of Mathematics and Technology, 1 (1), 161-164, (2010).

18. M. Andrade and M. A. M. Ferreira. Civil Identification Problems with Bayesian Networks using Official DNA Databases, 9th International Conference-APLIMAT 2010. Proceedings, 1 (1), 813-820, (2010).

19. M. Andrade. A Note on Foundations of Probability, Journal of Mathematics and Technology, vol.. 1 (1), pp 96-98, (2010).

20. M. Andrade and M. A. M. Ferreira. Bayesian Networks in Forensic Identification Problems, Aplimat- Journal of Applied Mathematics, 2 (3), 13-30, (2009).

21. M. A. M. Ferreira and M. Andrade. A note on Dawnie Wolfe Steadman, Bradley J. Adams and Lyle W. Konigsberg, "Statistical Basis for Positive Identification in Forensic Anthropology". American Journal of Physical Anthropology 131:15-26 (2006), *I*nternational Journal of Academic Research, 1 (2), 23-26, (2009).

22. M. Andrade and M. A. M. Ferreira. Criminal and Civil Identification with DNA Databases Using Bayesian Networks, International Journal of Security-CSC Journals, 3 (4), 65-74, (2009).

23. M. Andrade and M. A. M. Ferreira. Bayesian Networks in Forensic Identification Problems, 8th international conference-APLIMAT 2009. Proceedings, 1 (1), 699-716, (2009).

24. M. Andrade, M. A. M. Ferreira. J. A. Filipe and M. Coelho. Paternity Dispute: is it important to be conservative?, Aplimat- Journal of Applied Mathematics, 1 (2), (2008).

25. M. Andrade, M. A. M. Ferreira and J. A. Filipe. Evidence Evaluation in DNA Mixture Traces, Journal of Mathematics Statistics and Allied Fields, 2 (2), (2008).

26. D. Abrantes, M. L. Pontes, M. F. Pinheiro, M. Andrade and M. A. M. Ferreira. Towards systematic probabilistic evaluation of parentage casework in forensic genetics: A modest attempt to define a general standardized approach to simple and complex cases, Forensic Science International : Genetics Supplement Series.Elsevier, 1 (1), 351-354, (2008).

27. M. Andrade, M. A. M. Ferreira, J. A. Filipe and M. P. Coelho. Paternity Dispute: Is it important to be conservative?, 7th International Conference-APLIMAT 2008. Proceedings, 1 (1), 1031-1036, (2008).

28. M. Andrade, M. A. M. Ferreira, D. Abrantes, M. L. Pontes and M. F. Pinheiro. Redes bayesianas object-oriented na apreciação de paternidades em contextos menos usuais, XV Congresso Anual da SPE, Lisboa. Portugal, (2008).

29. M. Andrade and M. A. M. Ferreira. Analysis of a DNA Mixture Sample using Object-Oriented Bayesian Networks, 6th International Conference-APLIMAT 2007. Proceedings, 1 (1), 295-306, (2007).

30. M. Andrade and M. A. M. Ferreira. Mixture Traces: comparison of several hypotesis, 56th Session of the ISI-ISI 2007, Lisboa. Portugal, (2007).