Forensic Science International: Genetics 6 (2012) 689–696

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Inference about the number of contributors to a DNA mixture: Comparative analyses of a Bayesian network approach and the maximum allele count method

A. Biedermann [a,*], S. Bozza [b], K. Konis [c], F. Taroni [a]

[a] University of Lausanne, School of Criminal Justice, Lausanne, Switzerland
[b] University 'Ca' Foscari' of Venice, Department of Economics, Venice, Italy
[c] École Polytechnique Fédérale de Lausanne, Chair of Mathematical Statistics, Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 10 September 2011
Received in revised form 21 March 2012
Accepted 26 March 2012

Keywords:
DNA mixture profiling results
Number of contributors
Bayesian networks
Simulation
Probability and decision theory

ABSTRACT

In the forensic examination of DNA mixtures, the question of how to set the total number of contributors (*N*) presents a topic of ongoing interest. Part of the discussion gravitates around issues of bias, in particular when assessments of the number of contributors are not made prior to considering the genotypic configuration of potential donors. Further complication may stem from the observation that, in some cases, there may be numbers of contributors that are incompatible with the set of alleles seen in the profile of a mixed crime stain, given the genotype of a potential contributor. In such situations, procedures that take a single and fixed number contributors as their output can lead to inferential impasses. Assessing the number of contributors within a probabilistic framework can help avoiding such complication. Using elements of decision theory, this paper analyses two strategies for inference on the number of contributors. One procedure is deterministic and focuses on the minimum number of contributors required to 'explain' an observed set of alleles. The other procedure is probabilistic using Bayes' theorem and provides a probability distribution for a set of numbers of contributors, based on the set of observed alleles as well as their respective rates of occurrence. The discussion concentrates on mixed stains of varying quality (i.e., different numbers of loci for which genotyping information is available). A so-called qualitative interpretation is pursued since quantitative information such as peak area and height data are not taken into account. The competing procedures are compared using a standard scoring rule that penalizes the degree of divergence between a given agreed value for *N*, that is the number of contributors, and the actual value taken by *N*. Using only modest assumptions and a discussion with reference to a casework example, this paper reports on analyses using simulation techniques and graphical models (i.e., Bayesian networks) to point out that setting the number of contributors to a mixed crime stain in probabilistic terms is, for the conditions assumed in this study, preferable to a decision policy that uses categoric assumptions about *N*.

1. Introduction

"Given typing results on one or several loci, what assumptions should be made about the number of contributors?" This is a recurrent question that arises in the context of DNA typing of biological staining, in particular when allelic configurations are observed that cannot be explained by a single contributor. In a strict sense, the true number of contributors to a given sample cannot – in view of the currently used STR polymorphisms – be known with certainty. Because of possible effects of masking [1], the fact that no more than two alleles are observed at any locus in a profile does not imply that a stain could not be a mixture.

Notwithstanding, a widely followed conventional approach to mixture assessment considers explicit assumptions about the unknown (untyped) contributors to an evidential mixture under each of the competing propositions. Most often, assessing a mixture consists of a comparison between the probabilities of obtaining the typing results given that a specified individual – that is a named suspect – is (is not) the source of the crime stain, along with a certain number of additional untyped individuals e.g., [2]. The result of such a comparison is a likelihood ratio that provides an expression of the degree of discrimination among the competing propositions.

In the context, literature also reported on procedures that allow one to obtain an upper bound for the number of unknown persons that need to be considered as contributors to a mixed stain [3]. Much of current discussions on the 'number of contributors' issue involves, however, a breach of argument. Terms such as 'determining' the

* Corresponding author.
E-mail address: alex.biedermann@unil.ch (A. Biedermann).

1872-4973/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2012.03.006

numbers of contributors are regularly encountered but this opposes to the scientist's practical impossibility to set numbers for contributors accurately. Besides, this also suggests a deterministic view about the unknown number of contributors. It will be part of this paper to take some closer look at the relative performance of a typical categoric perspective of this kind, that is a method known as the maximum allele count. In a strict sense, this method is not a proper inference procedure, but merely a *rule* that sets the lower bound on the number of contributors to the minimum required to explain the set of alleles observed in a mixture.

Yet other contributions emphasise the idea of 'estimating' the number of contributors. There are frequentist procedures, for instance, that have a given number of contributors as their output. An example for this is the recent report on the maximum likelihood estimator [4,5]. This procedure selects that number of contributors for which the probability of the observed allelic configuration is maximal. Here, this approach is not pursued because it does not lead to a genuine expression of uncertainty about the number of contributors to a mixed stain (i.e., in terms of a probability distribution).

Approaches that have a fixed number of contributors as their output have some appeal because of their ease of application. In particular, they allow scientists to calculate the probability of an allelic configuration in a single step. This makes it unnecessary to account for several different numbers of contributors along with their respective probability, as required, for example, by the likelihood ratio approach of Brenner et al. [6]. It is questionable, however, whether the sole argument of ease of application should be considered as sufficient to justify a practical application. In fact, forensic scientists may be required to inform recipients of expert evidence about how well a chosen procedure performs. But this, in order to be of some value, asks for a comparison with the performance of an alternative procedure.

This paper intends to approach this aspect by an investigation and comparison of the potential of two procedures, based on simulated mixtures with varying numbers of contributors. One approach is the maximum allele count method, chosen as an example for a deterministic procedure. As a second approach, Bayesian inference is chosen as a probabilistic alternative. Bayesian inference is retained here as a method for belief revision about different numbers of contributors based on a mixture's allelic configuration as well as the relative rarity of the various observed alleles. It is such beliefs that are needed to weight the probability of observing a mixture's allelic configuration according to various numbers of contributors [6]. As an aside, this will also serve as an argument in support of the feasibility of an informed specification of a probability distribution for the number of contributors. This is of interest because practitioners sometimes criticise or do not use the approach of Brenner et al. [6] because of its involvement of probabilities (for numbers of contributors) that are claimed to be difficult to find.

This paper is structured as follows. Section 2 presents the general methodology and (computational) procedures (based, in part, on graphical models) used for (i) the simulation of DNA mixture profiles, (ii) the revision of beliefs about various numbers of contributors (for each sampled mixture profile) and (iii) the scoring of these inferences. The results of these analyses are presented in Section 3. A discussion with reference to a case example and conclusions are given in Section 4.

2. Methods and methodology

2.1. Software

The general computational environment chosen for this study is R, a widely used free software for statistical computing and graphics [7]. R was used for simulating STR profiling data and combining these in order to produce DNA mixture profiles. The program R was also used to write and apply a routine for processing the simulated DNA mixture data for finding the maximum allele count and, thus, the minimum number of contributors that, in combination, could have produced each conceptual DNA mixture profile.

Bayesian inference for the number of contributors was performed by using a Bayesian network. The formal network structure was set up in Hugin Researcher (Version 7.4). Further details on the definition of the Bayesian network is given below in Section 2.3. Bayes' theorem for inference about the number of contributors, given information about observed allelic configurations, is given hereafter in Section 2.2. Because the Bayesian network was needed to process a large number of simulated DNA mixture profiles, the RHugin package [8] was used as an interface for the Hugin Decision Engine (HDE).[1]

2.2. Bayesian inference for the number of contributors

For situations in which the number of contributors to a mixed DNA stain cannot be agreed, Brenner et al. [6] described an approach that allows for uncertainty about the total number of contributors. This approach combines the probabilities of typing results E given different numbers of contributors N. Each of these conditional probabilities is weighted by multiplying with the probability that the stain actually contains the respective number of contributors. Here, the discussion is confined only to inference about the total number of contributors. Thus, a proposition H, specifying a suspect as contributor, or some appropriate alternative, will not be included in the development.[2] Further, there will be no attempt to 'infer' the genotypes possessed by individual contributors.

To complete notation for generality, let I denote the framework of circumstances. Then, the probability of the typing results with uncertainty about the number of contributors can be written as follows:

$$Pr(E|I) = \sum_{j=1}^{n} Pr(E|N = j, I)Pr(N = j|I), \quad \text{for } j = 1, \ldots, n. \tag{1}$$

For the purpose of this paper, an exhaustive expansion $j = 1, \ldots, \infty$ to all relevant propositions N is avoided. Following argument by Buckleton et al. [1,9], attention can be confined to those hypotheses that imply a relevant contribution to the prior and the likelihood.

The probability defined in Eq. (1) is needed in Bayes' theorem for calculating the probability of a given number of contributors N, conditional on the typing results E. For example, the probability that $N = 3$ individuals contributed to a mixed DNA stain, given the typing results E, is obtained as follows:

$$Pr(N = 3|E, I) = \frac{Pr(E|N = 3, I)Pr(N = 3|I)}{Pr(E|I)}. \tag{2}$$

It remains to be defined how to calculate the probability of an allelic configuration E given a fixed number of contributors N.

[1] The HDE is part of Hugin Development Environment for performing operations on Bayesian networks (in particular probabilistic calculations). A main component, it contains a compiler that transforms networks into junction trees (i.e., structures that make it possible to perform inference in Bayesian networks efficiently).

[2] A development and use of the approach presented in this paper as part of a likelihood ratio based procedure for inference of source is feasible, but beyond the scope of this paper (see also Section 4.2 for discussion on this point and further references).

Summary formulae are available for this in literature (e.g., [10]), but these may become rather complex with increasing numbers of contributors and increasing number of distinct alleles in a mixture profile. Following Weir [10], for example, one can calculate the probability that x unknown contributors to a mixed profile have a set of alleles $\{u\}$ between them, but do not have any alleles that are not part of the profile E. More formally this is written as $Pr_x(\{u\} \mid E)$. As may be seen, one would need several of such equations, and combine them with the prior probabilities for each number of contributors, in order to calculate the posterior probability of a given number of contributors as defined by Eq. (2). This would then need to be repeated for each number of contributors N of interest.

In the study pursed here, it was decided to support this task by conducting these calculations within a Bayesian network (see also Section 2.3). The main advantage of using a Bayesian network is that one only needs to define a single model in order to account for several possible numbers of contributors and allelic configurations (of the mixed stain). Moreover, the calculation of the posterior probability distribution for a specified set of numbers of contributors is obtained in a single propagation. It is sufficient to specify the typing results for a given locus of the mixed stain. As explained later on, 'specifying the typing results' means here that, broadly speaking, there is a particular layer of nodes in the Bayesian network that take the role of 'capturing' the alleles that make up a mixture profile at a given locus (see, for example, nodes at the bottom in Fig. 2). This information is then processed in the Bayesian network and the posterior probability distribution for each value N is retrieved from the node that represents the variable N.

2.3. Bayesian networks

2.3.1. Representation of a mixture contributor

In this study, analyses were confined to mixtures that may contain up to four contributors. This is thought to cover the bulk part of situations encountered in practice where mixture profiling data would be retained for evaluation. In fact, as noted by Kelly et al. [11], it makes sense to limit the efforts to what is biologically feasible and in relation to the limitations of current typing technologies, but not to push, at the same time, modelling approaches for the number of contributors too far.

The study here captures situations involving several contributors to a DNA mixture in terms of a Bayesian network and by pursuing a staged modelling approach. That is, first, a local network fragment that models the alleles that a single individual may contribute to a DNA mixture was constructed. This network fragment was then duplicated repeatedly to represent additional potential contributors. The distinct network fragments were then combined in order establish a connection to the variable representing the number of contributors and the variables representing the observed set of alleles in the DNA mixture. Further details on this modelling approach are given here below.

Start by considering Fig. 1. It represents the alleles that an individual may contribute to a DNA stain. The nodes '*pg*(1)' and '*mg*(1)' represent the paternal and maternal alleles of individual 1. These are numerical, each with a set of nine states {1, 2, 3, 4, 5, 6, 7, 8, 99}. Each number represents an allele of some unspecified locus. Notice that neither the locus, nor alleles are specified in further detail at this point because the intention is to set up a generic Bayesian network that can be completed and instantiated according to particular typing results at hand. A total of eight alleles is specified because this corresponds to the maximum number of distinct alleles that may be seen in four heterozygous contributors if they share no alleles between them. A state arbitrarily named '99' is included in order to account for all alleles other than those labelled '1'–'8'. The nodes '*pg*(1)' and '*mg*(1)' are so-called founder gene nodes. They contain (unconditional) probability distributions that reflect the occurrence of the specified alleles in the relevant population. The probability assigned to the state '99' will thus be the cumulative probability of all alleles other than '1'–'8'. In turn, the probabilities assigned to the states '1'–'8' reflect the probability with which the respective alleles occur in the population of interest.

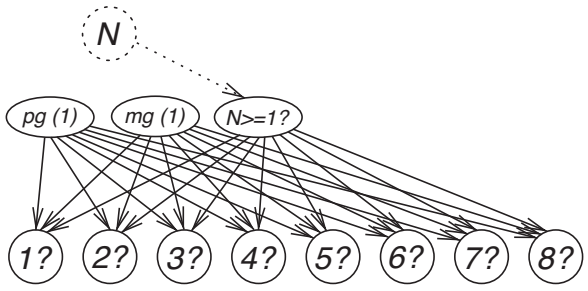

Fig. 1. Bayesian network for representing the allelic configuration of an individual (at a given locus) as well as the alleles that this individual may contribute to a DNA stain. Nodes '*pg*(1)' and '*mg*(1)' represent, respectively, the paternal and maternal allele (of a generic individual numbered '1'). The nodes '*i* ?' are Boolean and state whether an allele *i* is contributed to the DNA stain. The node '$N >= 1$?' is Boolean and defines whether there is at least one contributor to the DNA stain. The node N represents the possible total numbers of contributors.

The node '$N >= 1$?' is Boolean and assumes the state '*true*' whenever the node N, a numerical node with states {0, 1, 2, 3, 4} representing the possible number of contributors,[3] is in a state greater or equal to 1. If N is in the state '0', then '$N >= 1$?' is false and the submodel shown in Fig. 1 will be prevented from transmitting information about the allelic constitution of the individual at hand further on (i.e., the individual will not be considered as a contributor to the DNA mixture). The node N is shown here with a dashed line because it is part of the global network wherein several network fragments for distinct individuals will be combined.

The nodes '*i* ?', for $i = 1, 2, \ldots, 8$, are Boolean. They assume the state '*true*' whenever the node '$N >= 1$?' is true and at least one of the nodes '*pg*(1)' and '*mg*(1)' is in state *i*. Stated otherwise, the nodes '*i* ?' reflect – in combination – the allele(s) that individual one may contribute to the mixture. As already mentioned, this translates a purely qualitative approach throughout the whole model in the sense that peak area/intensities are not taken into account.

2.3.2. Combination of local network fragments for distinct contributors

The Bayesian network fragment described in the previous section can be reused[4] for representing the allelic constitution of a further individual as well as the alleles that such an individual may contribute to a mixed stain. This is illustrated in Fig. 2, that shows a Bayesian network for modelling the alleles that two individuals may contribute to a mixed DNA stain. Here, '*pg*(2)' and '*mg*(2)'

[3] The state '0' is included here for the sake of completeness. Although, later on, simulations will always suppose at least one contributor, one could also think of generating DNA 'samples' that do not contain material from any contributors. In such a case, the model would need to be able to infer $N = 0$ contributors. In turn, one may think of extensions of the model to situations in which allele drop-out occurs. In such a setting, detecting no alleles at a given locus will oppose propositions of no contributors and contributors whose alleles have undergone drop-out. More generally, beliefs about the numbers of contributors may even be specified prior to analysing a sample and this may require that the possibility of 0 contributors could have a probability different from zero.

[4] Notice that the node N modelling the number of contributors is not duplicated. It is a node that the various sub-models for distinct stain contributors have in common.

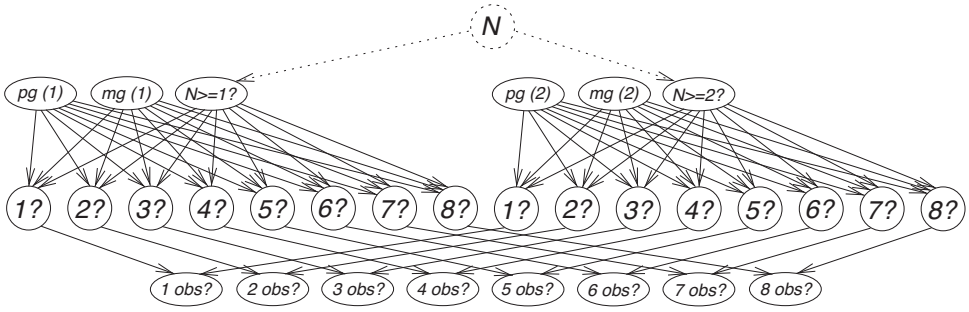

Fig. 2. Bayesian network for representing the allelic configuration of two individuals (at a given locus) as well as the alleles that these individuals may contribute to a DNA stain. Nodes '*pg*' and '*mg*' represent, respectively, the paternal and maternal allele of individual one (1) and two (2). The nodes '*i* ?' are Boolean and state whether an allele *i* is contributed to the DNA stain. The nodes '$N >= 1$?' and '$N >= 2$?' are Boolean and define whether there are, respectively, at least one and two contributors to the DNA stain. The node *N* represents the possible total numbers of contributors. The nodes '*i obs*?' model the alleles in the DNA mixture, depending on the allelic contributions from individual one and two.

represent, respectively, the paternal and maternal allele of the second potential contributor. The node '$N >= 2$?' is Boolean and assumes the state '*true*' whenever the node *N* is in a state greater or equal to 2 (i.e., when there are at least two contributors). The set of Boolean nodes '*i* ?' (for $i = 1, 2, \ldots, 8$) of each individual feed a set of nodes termed '*i obs* ?'. This bottom layer of nodes models the alleles that may be found in a mixture to which individual one and two have contributed. The nodes '*i obs* ?' are Boolean and assume the state '*true*' whenever at least one of its parental node is true (i.e., either individual one or two has allele *i*). They thus serve the purpose of 'communicating' to the model the alleles that are observed in a mixture. Note again that this node layer accounts only for the presence or absence of particular contributed alleles. This qualitative approach does not take into account the possibility that the contributions from different individuals may involve DNA in different quantities.

Using analogous argument, two further network fragments of the kind shown in Fig. 1 have been added so as to obtain a Bayesian network that models the possible contribution of up to four individuals. This amounts to a duplication of the structure shown in Fig. 2, except for the node *N* as well as the layer of nodes '*i obs*'?. On the whole, this results in a rather densely connected network structure. Clearly, the staged modelling approach built on distinct sub-models would have been ideally suited for an object-oriented network construction, but this could not be retained here because the file format for object-orientation is currently not supported by RHugin [8].

2.4. Simulations and analyses

The simulations and analyses in this study involved several distinct steps. One of these is concerned with the generation of sets of DNA mixture profiles. The DNA profiles for individual mixture contributors were simulated by drawing alleles independently at their relative rate of occurrence. Whilst simulating a profile for a mixture contributor, possible correlations between the two alleles of one contributor or between the alleles of different individuals were not taken into account. Allele probabilities were those for U.S. African Americans published in Butler et al. [12]. Each mixture contributor was simulated independently of any other mixture contributor. When generating DNA mixture profiles, genotypes from up to four individuals were combined. The following 10 loci were considered: D13S1358, VWA, FGA, D8S1179, D21S11, D5S818, D13S317, TH01, TPOX, D7S820. Sets of DNA mixtures were also set up for lower numbers of loci, that is 7 and 5, chosen randomly among these 10 loci (for each mixture). Any given DNA mixture is thus summarized by an allele vector. As no quantitative information is considered, this involves no such step as a conversion of observed peaks to alleles which, otherwise, would require additional discussion.

The number of contributors was not kept fixed when simulating sets of DNA mixture profiles. Instead, that number was sampled in each trial with rates thought to represent the proportions of practical cases involving, for example, two, three or four contributors. The reason for this is that – from a practical point of view – it is of interest to detain information about the performance of inference approaches for the number of contributors in repeated trials where each trial may involve one of several distinct numbers of contributors, rather than a given constant number of contributors. Although in practical work it may not precisely be known at which proportion the various possible numbers of contributors appear, these proportions can nonetheless be framed in reasonable orders of magnitude (based, for example, on empirical studies (e.g., [13]). Here, the analyses reported several such proportions.

A second main step focused on applying the two competing inference methods. For each conceptual mixture profile, the maximum allele count was determined along with the minimum number of distinct contributors necessary to 'explain' the mixture. In a subsequent step, the same mixture profiles were processed using Bayesian inference (Section 2.2). This latter step led to (posterior) probability distributions for the number of contributors, calculated according to Eq. (2). Practically, these posterior probability distributions were found by loading the Bayesian network described in Section 2.3.2 in R and communicating to the model the set of alleles observed for each locus (for a given mixture profile). Prior to using the Bayesian network for inference about *N* at a given locus, it was necessary to assign the allele proportions of the set of alleles under investigation (i.e., the set of alleles at the locus at hand) to the generic states of all founder gene nodes *pg* and *mg*. After processing the data for each locus of a profile, the overall posterior probability distribution for the profile under investigation is obtained in the node *N*. Before using the Bayesian network for processing a new mixture profile, the distribution for the node *N* was set back to an initial distribution. In order to allow for a common starting point for each profile across the various sets of simulated DNA mixture profiles (i.e., sets with distinct proportions for the different number of contributors), a default initial distribution {0.25, 0.25, 0.25, 0.25} was chosen for $N = \{1, 2, 3, 4\}$. This avoids the choice of prior distributions for the Bayesian inference method that might initially approximate or otherwise (and thus, provide this method an advantage or handicap) the proportions with which two, three and four contributors are simulated within each set of DNA mixture profiles.

The last step of the analyses focused on scoring and comparing the results of the two inference methods. Let us recall that the

method of maximum allele count provides a deterministic conclusion that can readily be compared with the actual (and known) number of contributors that was used when constructing the mixture profile of interest. It is thus straightforward to determine whether a given 'conclusion' based on the maximum allele count correctly determines the number of contributors of a given mixed DNA profile.

The Bayesian inference method (Section 2.2) does not provide such a categorical statement and this raises the question of how to assess its output with respect to the true state of affairs (i.e., the actual number of contributors of a given conceptual DNA mixture). In the study here, a Brier score [14] is used. It penalizes the divergence between the inferred and true state by a sum of squared errors.[5] An advantage of this scoring rule is that it can also accommodate the output of the maximum allele count method. Suppose, for example, that this method would indicate the correct result (i.e., two contributors). The resulting score would thus be zero because the true state of affaires coincides with the maximum allele count result. However, notice that the deterministic conclusion of two contributors, that is {0, 1, 0, 0}, when in fact there are three contributors, that is {0, 0, 1, 0}, would lead to a total score of 2. In turn, for a probabilistic conclusion of the kind {0, 0.9, 0.06, 0.04}, for example, the reader can verify that the quadratic score would be given by 0.01 + 0.0036 + 0.0016 = 0.0152 .

3. Results

3.1. Ten loci DNA mixture profiles

Four different sets of 100 DNA mixture profiles at 10 loci were simulated. For the first set of 100 profiles, the number of contributors for each mixture, that is two, three or four, was sampled with equal probability. For the three other sets of 100 DNA mixtures, the number of contributors was sampled with, respectively, the vectors of probabilities {0.4, 0.4, 0.2}, {0.5, 0.4, 0.1} and {0.1, 0.45, 0.45}. Applying the maximum allele count and Bayesian inference to each mixture profile, and cumulating the scores, led to the overall penalties summarised in Table 1. Fig. 3 plots the evolution of the cumulated score for both methods when processing the mixture profiles one at the time – in each of the four sets of 100 DNA mixture profiles with 10 loci.

The results show that the performance of the maximum allele count depends considerably on the proportion of mixture profiles with low (i.e., two) contributors. As shown by the scores in Table 1, the maximum allele count performs best – but still not as good as Bayesian inference – for the set of profiles with the highest proportion of profiles with two contributors (i.e., {0.5, 0.4, 0.1}). The score increases for sets of profiles with lower proportions of two contributor mixtures. It achieves 90 for the setting in which two contributor mixtures were simulated with a rather low probability of 0.1. This score of 90 in an evaluation of 100 profiles means that the maximum allele count led to 45 categorically wrong conclusions. This variation in performance is in agreement with the fact that, with low numbers of contributors, 'overlapping' of alleles from different contributors is less probable than for higher numbers of contributors. Notice that when contributors have alleles in common, then this may lead to the impression that there are less contributors than that there actually are. In turn, Bayesian inference appears to be rather insensitive to the apportionment of numbers of contributors. In fact, the cumulated score for each set of simulation remained in a rather narrow range below 20.

[5] Other scoring rules may be chosen, but one should be aware that there are rules that are not proper in the sense defined in decision theory [15].

Table 1
Cumulated scores for the maximum allele count (column two) and Bayesian inference (column three, rounded to two decimals) for four sets of 100 DNA mixture profiles. The first column gives the rates at which two, three and four contributors were sampled within each of the four sets of profiles.

Proportions for number of contributors	Maximum allele count	Bayesian inference
{1/3, 1/3, 1/3}	54	19.48
{0.4, 0.4, 0.2}	54	13.41
{0.5, 0.4, 0.1}	20	15.68
{0.1, 0.45, 0.45}	90	18.21

3.2. Partial DNA mixture profiles (with 7 and 5 loci)

Retaining the same total number of contributor apportionments as in the previous section, that is {1/3, 1/3, 1/3}, {0.4, 0.4, 0.2}, {0.5, 0.4, 0.1} and {0.1, 0.45, 0.45}, further sets of DNA mixture profiles with, respectively, 7 and 5 loci were generated. Within each of these sets, the loci were selected at random among the loci used for simulating the 10 loci mixture profiles (see also Section 2.4). Each locus had the same probability to be selected.

Table 2 summarises the overall scores for both the maximum allele count and Bayesian inference. Graphical illustrations of the evolution of the cumulated scores within each set of profiles are shown in Fig. 4. Throughout the simulations here, Bayesian inference has again turned out to provide better scores than the maximum allele count. Consideration of profiles with less loci tends to increase the scores for both methods. Again, the maximum allele count shows its best scores for sets of mixtures with high proportions of low numbers of contributors, such as the apportionment {0.5, 0.4, 0.1}. But even for this setting, Bayesian inference has a score almost half as great. More generally, the score for Bayesian inference is, on average, approximately only 1/3 of that for the maximum allele count. With a high apportionment of more than two contributors, such as {0.1, 0.45, 0.45}, and a low number of loci for which genotypic information is available (e.g., setting with 5 loci), the maximum allele count was found to give a score as high as 110. This particular result means that for more than half of the 100 mixture profiles in this setting (actually, for 55 profiles), the number of contributors was falsely 'determined'. In contrast to this, Bayesian inference led to a score of about 36, that is about three times lower.

4. Discussion and conclusions

4.1. Need for a balanced approach

The widely used maximum allele count method owes much of its popularity to its ease and rapidity of application. Although there is evidence that suggests that this procedure has some appealing performance when the mixture involves a low number of contributors (e.g., [4]), this cannot be put forward as a justification

Table 2
Cumulated scores for the maximum allele count (MAC) and Bayesian inference (BI; rounded to two decimals) for eight sets of 100 DNA mixture profiles. The first column gives the apportionment of two, three and four contributors in each mixture set. For each of these apportionments, two distinct sets of 100 mixture profiles were generated, one with 7 loci profiles and one with 5 loci profiles.

Proportions for number of contributors	7 loci		5 loci	
	MAC	BI	MAC	BI
{1/3, 1/3, 1/3}	84	25.62	86	29.55
{0.4, 0.4, 0.2}	52	19.72	66	29.63
{0.5, 0.4, 0.1}	50	28.21	48	28.84
{0.1, 0.45, 0.45}	88	24.07	110	36.28

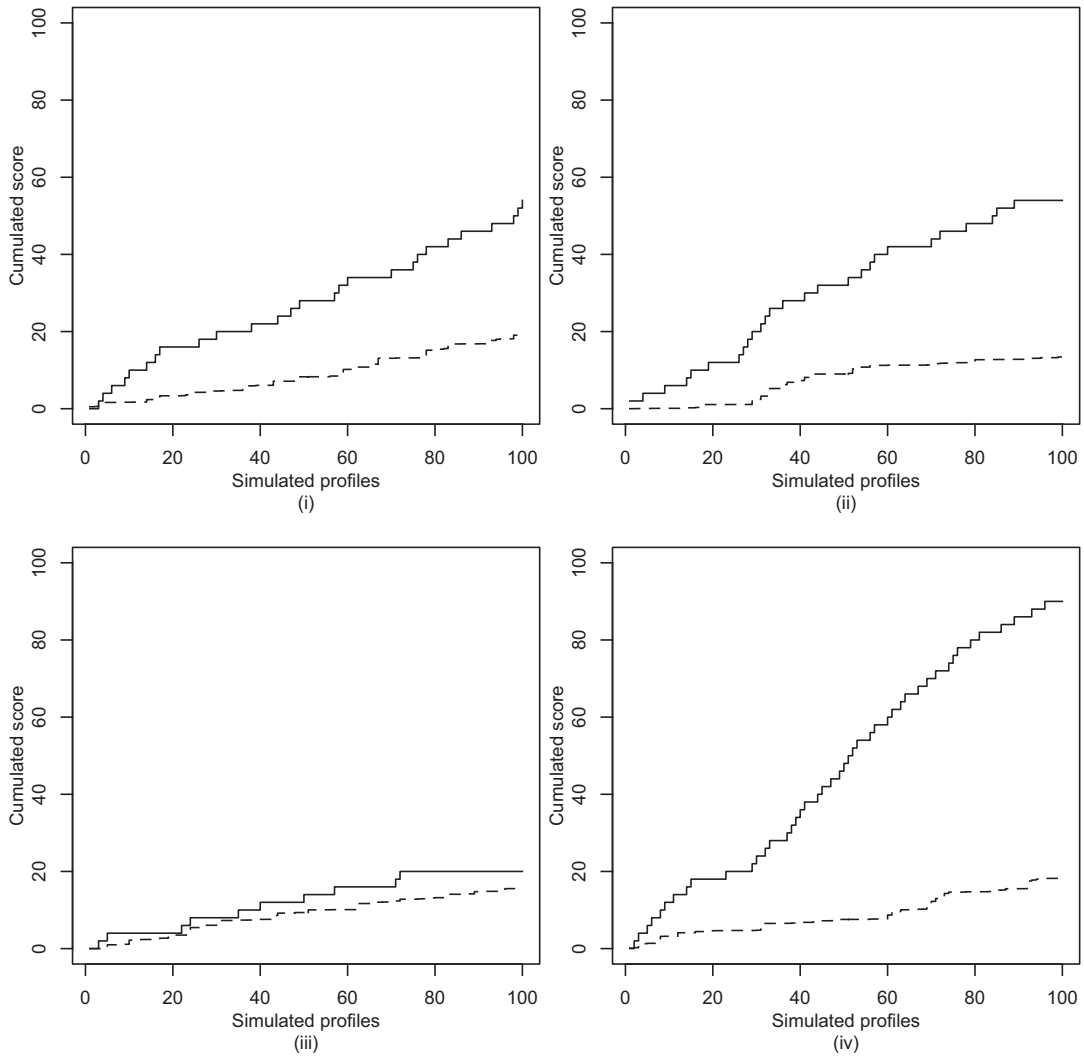

Fig. 3. Cumulated scores for the maximum allele count (solid line) and Bayesian inference (dashed line) for four sets of 100 DNA mixture profiles with 10 loci. Figure (i) illustrates the evolution of the score when processing the DNA profiles of the set for which two, three and four contributors were sampled at equal rates (i.e., {1/3, 1/3, 1/3}). Figures (ii), (iii) and (iv) show plots for the sets with contributor proportions of, respectively, {0.4, 0.4, 0.2}, {0.5, 0.4, 0.1} and {0.1, 0.45, 0.45}.

for using this method in practice because in actual casework, the true number of contributors is typically unknown. According to the observations in this study, it may be generally unsafe to use the maximum allele count because this rule can have a high number of false determinations. This is the case, for example, when working on kinds of crime stains that possibly involve high rates of multiple contributors (i.e., more than two) and/or when profiles are incomplete (i.e., not all loci provide typing results).

Besides, a categorical view with respect to the number of contributors to a DNA mixture can lead to conflicting standpoints. The partial mixed crime stain summarised in Table 3, drawn from a real case, illustrates this point. An intriguing aspect of this case is that, judged solely on the basis of the alleles present in the mixture profile, a total number of two contributors, as suggested by the method of maximum allele count, is sufficient to 'explain' the mixture. The prosecution, however, retained a number of at least three contributors. As a consequence, a suspect whose profile is shown in column three in Table 3, could be considered as a potential contributor. Due to homozygosity at loci D21S11 and D5S818, this suspect could not be a contributor to the crime stain under the assumption of only two contributors.

Contradicting and extreme conclusions such as 'exclusion' and 'non-exclusion' of a particular individual as a potential donor to a mixed crime stain can be avoided when focusing on the probability of the typing results given different numbers of contributors, weighted by the respective probability of each number of contributors (e.g., [16]). This is expressed, for instance, by Eq. (1). Such an approach is more prudent because it does not involve the claim of a categorical statement about a proposition – the number of contributors – that is one that cannot actually be known with certainty. Any categorical statement about such a proposition would involve a suppression of uncertainty that is not warranted by available evidence.

In view of this, it thus appears natural to argue that uncertainty about the number of contributors ought to be expressed by probability. This implies the need for Bayesian inference. For the purpose of illustration, Table 4 summarises some posterior distributions for two, three and four contributors (to the mixed stain described in Table 3), based on different prior distributions and using Bayesian inference as described earlier in Section 2. An additional advantage of a Bayesian procedure is that it is capable of incorporating circumstantial information (e.g., when specifying a prior distribution). This is not the case for the method of maximum allele count.

More generally, it may be argued that the analyses pursued here are limited in the sense that potentially helpful information on peak area/height is not taken into account. However, practitioners often seek a quick method to provisionally inform their beliefs

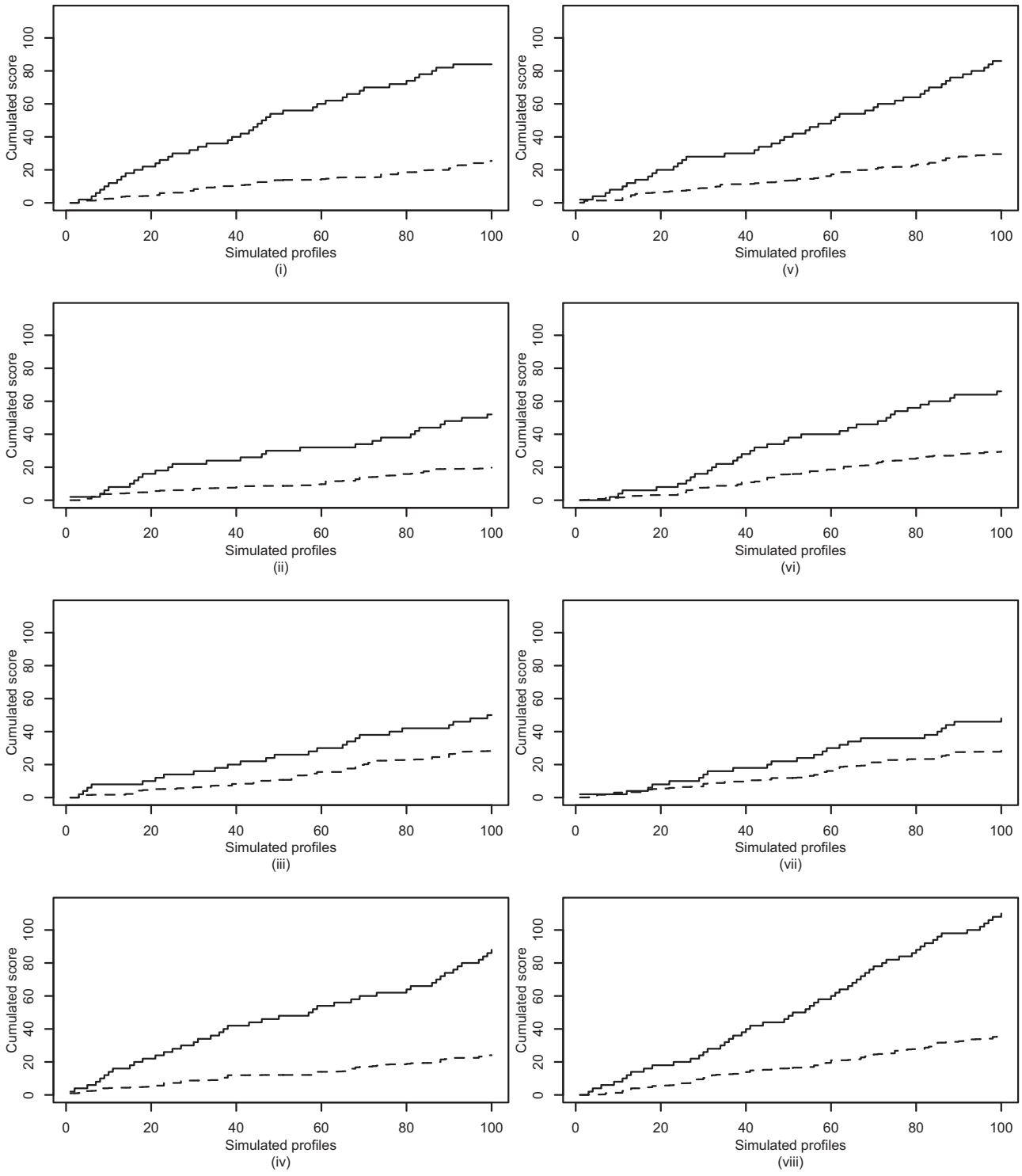

Fig. 4. Cumulated scores for the maximum allele count (solid line) and Bayesian inference (dashed line) for eight sets of 100 DNA mixture profiles. The plots in the column on the left illustrate the evolution of the cumulated scores for profiles with 7 loci. The plots in the column on the right show the results for sets of profiles with 5 loci. The plots (i) and (v) represent the evolution of the cumulated scores for equal apportionments of two, three and four contributors (i.e., {1/3, 1/3, 1/3}). Plots (ii) and (vi) refer to the set {0.4, 0.4, 0.2}, plots (iii) and (vii) to the set {0.5, 0.4, 0.1} and plots (iv) and (viii) to the set {0.1, 0.45, 0.45}.

about potential numbers of contributors. In such contexts, the mere counting of alleles represents an attractive method. It is for this reason that this study focused on assessing that particular method, by comparing it with a Bayesian procedure. Notwithstanding, it is acknowledged at this point that, potentially, Bayesian networks can deal with quantitative profiling data (e.g., [17,18]). The particular topic of inference about the number of contributors within a more quantitative perspective of using Bayesian networks thus represents a potential area of further research. Although a graphical probabilistic approach represents a clearly versatile method that offers room for dealing with further aspects, such as typing artifacts and multiple stains, their development and implementation may be computationally demanding [19] and, thus, represent a practical challenge.

Table 3
Typing results for a crime stain and a suspect drawn from a real case.

Locus	Crime stain	Suspect
D3S1358	14, 15, 16	14, 15
VWA	15, 16, 17	16, 16
FGA	19.2, 23, 24, 25	19.2, 23
D8S1179	12, 14, 15	14, 15
D21S11	28, 29, 30, 32.2	30, 30
D5S818	8, 11, 12, 13	13, 13
D13S317	9, 11, 12	9, 11

Table 4
Bayesian inference for the number of contributors to a mixed DNA stain as defined in Table 3, based on four different prior distributions and using the procedure described in Section 2.

Probability:	Prior			Posterior		
Number of contributors:	2	3	4	2	3	4
	1/3	1/3	1/3	0.293	0.674	0.033
	0.4	0.4	0.2	0.298	0.686	0.016
	0.5	0.4	0.1	0.349	0.643	0.008
	0.1	0.45	0.45	0.084	0.874	0.042

4.2. The role of Bayesian networks

Widespread interest in categorical approaches, such as the maximum allele count rule, can in part be explained by the fact that the application of alternative approaches may be computationally more demanding. Bayesian inference is sometimes put forward as an example for this essentially because expressions like Eq. (2) involve several components that need to be calculated separately and combined with assumptions about prior probabilities. In fact, the extent of computations that is necessary to find a posterior distribution may be discouraging. However, methods currently exist that allow one to translate the logic of Bayesian inference into an abstract graphical representation (i.e., a Bayesian network) that can be instantiated as required [20–23]. This allows genotypic information about a particular mixed crime stain to be incorporated on a case specific level along with relevant population data. The feasibility of such a procedure in a simulation study was illustrated in this paper by analysing several sets of simulated DNA mixture profiles and using using current Bayesian network software.

In the study presented here, Bayesian inference was confined solely to reasoning about the number of contributors to given mixed DNA stains and comparing that approach to a commonly used deterministic procedure. It is worth noting, however, that a graphical approach for inference about the number of contributors can readily be developed as part of a larger network that seeks to discriminate between the propositions according to which the suspect is (is not) the contributor to a given (mixed) crime stain. This allows multiple sources of uncertainty to be approached within a coherent whole and supports the implementation of existing formal likelihood ratio developments (e.g., Eq. (2)) on a case-specific level. This also points out the versatility of Bayesian networks, that can be constructed in a way that allows one to cope with multiple inference tasks. That is, they can be used to discriminate between target propositions regarding the source of a crime stain and tackle inferences about the number of contributors (e.g., [16]).

Conflict of interest

None of the authors A. Biedermann, S. Bozza, K. Konis, F. Taroni has a financial or personal relationship with other people or organisations that could inappropriately influence or bias the paper entitled "Inference about the number of contributors to a DNA mixture: Comparative analyses of a Bayesian network approach and the maximum allele count method".

References

[1] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci. Int.: Genet. 1 (2007) 20–28.
[2] I.W. Evett, B.S. Weir, Interpreting DNA Evidence, Sinauer Associates Inc., Sunderland, 1998.
[3] S.L. Lauritzen, J. Mortera, Bounding the number of contributors to mixed DNA stains, Forensic Sci. Int. 130 (2002) 125–126.
[4] H. Haned, L. Pène, F. Sauvage, D. Pontier, The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci. Int.: Genet. 5 (2011) 281–284.
[5] H. Haned, L. Pène, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci. 56 (1) (2011) 23–28.
[6] C.H. Brenner, R. Fimmers, M.P. Baur, Likelihood ratios for mixed stains when the number of donors cannot be agreed, Int. J. Legal Med. 109 (1996) 218–219.
[7] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2009). http://www.R-project.org.
[8] K. Konis, RHugin: RHugin, R package version 7.4/r247 (2010). http://rhugin.r-forge.r-project.org/.
[9] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, FL, 2005.
[10] B.S. Weir, Genetic Data Analysis II, Sinauer Associates, Sunderland, MA, 1996.
[11] H. Kelly, J.-A. Bright, J. Curran, J. Buckleton, The interpretation of low level DNA mixtures, Forensic Sci. Int. 6 (2012) 191–197.
[12] J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, M.C. Kline, Allele frequencies of 15 autosomal STR loci on U.S. Caucasian, African American and Hispanic populations, J. Forensic Sci. 48 (2003) 908–911.
[13] Y. Torres, I. Flores, V. Prieto, M. López-Soto, M. José Farfán, A. Carracedo, P. Sanz, DNA mixtures in forensic casework: a 4-year retrospective study, Forensic Sci. Int. 134 (2003) 180–186.
[14] G.W. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Rev. 78 (1950) 1–3.
[15] G. Parmigiani, L. Inoue, Decision Theory: Principles and Approaches, John Wiley & Sons, Chichester, 2009.
[16] A. Biedermann, F. Taroni, W.C. Thompson, Using graphical probability analysis (Bayes nets) to evaluate a conditional DNA inclusion, Law, Probability Risk 10 (2011) 89–121.
[17] R.G. Cowell, S.L. Lauritzen, J. Mortera, Identification and separation of DNA mixtures using peak area information, Forensic Sci. Int. 166 (2007) 28–34.
[18] R.G. Cowell, S.L. Lauritzen, J. Mortera, A Gamma model for DNA mixture analyses, Bayesian Anal. 2 (2007) 333–348.
[19] R. G. Cowell, S. L. Lauritzen, J. Mortera, Probabilistic modelling for DNA mixture analysis, Forensic Science International: Genetics Supplement Series 1 (2008) 640–642, Progress in Forensic Genetics 12 – Proceedings of the 22nd International ISFG Congress.
[20] A.P. Dawid, J. Mortera, V.L. Pascali, D. van Boxel, Probabilistic expert systems for forensic inference from genetic markers, Scand. J. Stat. 29 (2002) 577–595.
[21] F. Taroni, C.G.G. Aitken, G. Garbolino, A. Biedermann, Bayesian Networks and Probabilistic Inference in Forensic Science, John Wiley & Sons, Chichester, 2006.
[22] A.P. Dawid, J. Mortera, P. Vicard, Object-oriented Bayesian networks for complex forensic DNA profiling problems, Forensic Sci. Int. 169 (2007) 195–205.
[23] A. Biedermann, F. Taroni, Bayesian networks for evaluating forensic DNA profiling evidence: a review and guide to literature, Forensic Sci. Int.: Genet. 6 (2012) 147–157.