Mixtures with relatives: a pedigree perspective

Thore Egeland^{a,b *}

Guro Dørum^a

Magnus Dehli Vigeland^c

Nuala A. Sheehan^d

^a*IKBM, Norwegian University of Life Sciences, Ås, Norway*

^b*Norwegian Institute of Public Health, Oslo, Norway*

^c*Department of Medical Genetics, Oslo University Hospital, Oslo, Norway*

^d*Departments of Health Sciences and Genetics, University of Leicester, UK*

June 22, 2013

Abstract

DNA mixture evidence pertains to cases where several individuals may have contributed to a biological stain. Statistical methods and software for such problems are available and a large number of cases can be handled adequately. However, one class of mixture problems remains untreated in full generality in the literature, namely when the contributors may be related. Disregarding a plausible close relative of the perpetrator as an alternative contributor (identical twin is the most extreme case) may lead to overestimating the evidence against a suspect. Existing methods only accommodate pairwise relationships such as the case where the suspect and the victim are siblings, for example. In this paper we consider relationships in full generality, conveniently represented by pedigrees. In particular, these pedigrees may involve inbreeding, for instance when the parents of an individual of interest are first cousins. Furthermore our framework handles situations where the opposing parties in a court case (prosecution and defence) propose different family relationships. Consequently, our approach combines classical mixture and kinship problems.

The basic idea of this paper is to formulate the problem in a way that allows for the exploitation of currently available methods and software designed originally for linkage applications. We have developed a freely available R package, **euroMix** based on another package, **paramlink**, and we illustrate the ideas and methods on real and simulated data.

Keywords: Likelihoods; Pedigrees; Forensics; DNA mixtures; Related Contributors.

*Corresponding author at: Norwegian University of Life Sciences, Norway. Phone +4764966391.
E-mail address: Thore.Egeland@umb.no (T. Egeland)

1 Introduction

The broad motivation for this paper is the evaluation of DNA evidence in criminal cases. We deal specifically with *mixture evidence* which refers to cases where several individuals may have contributed to a biological stain recovered from a crime scene. Rape cases present an important example and occur frequently in forensic case work. Typically, the DNA profile based on a vaginal swab will indicate the presence of the victim and one or more men. The forensic scientist can evaluate the likelihood of the DNA data assuming the prosecutor and defence hypotheses, respectively, and report a *likelihood ratio* (LR).

Our calculations are based on qualitative data obtained from discretising the continuous measurements delivered in electrophorograms prior to calculation. This is in line with traditional approaches and implementations. However, there have been attempts to use the continuous measurements directly (see [1] for a recent paper analysing DNA mixture evidence using Bayesian Networks). Our reasons for taking the qualitative data as a starting point are mainly practical. It seems reasonable to first provide the extensions to standard implementations described below in the conventional setting before considering more ambitious models that utilise more of the information in the data. However, it should be noted that modelling the continuous data comes at a price in that more assumptions are required and implementation becomes less straightforward.

Statistical methods for mixtures have been developed when all contributors are assumed to be unrelated [2]. Most cases can therefore be handled adequately by existing software [3]. The motivation for this paper is cases involving related contributors. Typically, the evidence against a suspect with matching DNA profile will be strong if the alternative is that somebody unrelated to the suspect contributed: Intuitively it is very unlikely that a randomly selected person will fit the evidence. However, it is much more likely that a brother of the suspect matches and this must be accounted for as explained in [4]. Fung and Hu have written widely on the situation where some of the contributors to the biological stain could be related (references include [5, 6, 7, 8]). However, their approach is limited to specific relationships between two contributors. They argue that consideration of more than two relatives is not very relevant practically, and becomes difficult computationally as kinship coefficients involving more than two individuals would have to be derived. However, there are important cases involving more complicated family relationships and there are also, as we will see, alternatives to calculations based on kinship coefficients.

In this paper we propose a general method for handling mixture problems with related contributors. The relationships may involve any number of individuals and families with inbreeding. Furthermore, contrary to previous methods [7, 2] the opposing hypotheses need not agree on the relationships. Hence, the approach presented here is also applicable to forensic identification problems where both mixture data and reference data for certain individuals are available. Combining genotype data at the individual level from reference samples with a trace DNA profile assumed to originate from a single individual was discussed in [9]. Here we discuss the much more general

case where the trace could be a mixture from several individuals, some of whom may be related.

We formulate the problem in a way that allows for methods and software designed originally for linkage applications, to be applicable. Thus, well established implementations that are freely available can be used for these applications. General overviews of algorithms for pedigree likelihood calculations such as the Elston-Stewart algorithm (or more generally *peeling* algorithms) [10, 11] used in our implementation together with further references, can be found in [12, 13]. However, it should be noted that the traditional approach based on the general formulae presented in [14] is likely to be more efficient when contributors to the mixture are unrelated or simply pairwise related as summarised in [2, 7].

We will commence with a motivating example to provide the essential background and notation after which we will formulate the problem more generally. Three examples are presented in Section 3, the last one based on a real case. In Section 4 we discuss the assumptions underlying our approach, along with possible extensions and suggestions for future work. Some further examples are presented in the supplementary material and the online documentation.

2 Methods

2.1 Motivating example

It is convenient to introduce the context and essential concepts via an example. The example is specifically constructed to allow for simple calculations. A DNA mixture from three individuals is available and consists of *alleles* denoted 1 and 2 from a *genetic marker*. In addition, reference samples are available from two undisputed contributors to the mixture: the typed individuals labelled 3 and 4 in Figure 1.

The genotype and identity of the third contributor is unknown and disputed. For this paper we are assuming *unlinked markers* and *linkage equilibrium* [12] and therefore we only need one marker to explain the concepts since likelihood ratios are multiplied across independent markers. Obviously a reasonable number of markers, such as the database with 17 EXS17 markers and 6 additional markers that comes with our software **euroMix** described below, is required to obtain sufficient statistical power to discriminate between opposing hypotheses.

Data like this could occur in a crime context where individuals 3 and 4 could be victims in a murder investigation, for instance, and where the third contributor could be the presumed perpetrator, or murderer. The identity of the perpetrator is disputed, and the suspects are individuals 5 and 6 in the two suggested pedigrees, respectively, shown in Figure 1. Neither of them have been genotyped. On the other hand, this could be an identification problem as noted in the introduction where the task is to determine the pedigree connecting the typed individuals.

The assessment of the evidence starts by formulating the prosecution (P) and defence

Figure 1: The pedigrees representing the prosecution (left) and defence (right) hypotheses of the motivating example. The victims 3 and 4 are known contributors to the mixture $R = \{1, 2\}$. The opposing hypotheses disagree on whether the victims are sisters or half sisters, and whether the third, untyped, contributor is their brother (5) or the father (6) of one of them).

(D) hypotheses which in this case are

- H_P : The contributors are the individuals 3, 4 and 5 related as shown in the left part of Figure 1.
- H_D : The contributors are the individuals 3, 4 and 6 related as shown in the right part of Figure 1.

The strength of the evidence is normally summarised by the likelihood ratio

$$LR = \frac{P(R, T|H_P)}{P(R, T|H_D)} = \frac{P(R|T, H_P) P(T|H_P)}{P(R|T, H_D) P(T|H_D)}, \quad (1)$$

where $R = \{1, 2\}$ denotes the set of alleles found in the mixture evidence and $T = \{(3, g_3), (4, g_4)\}$, with $g_3 = 1/1$ and $g_4 = 2/2$ being the known genotypes of individuals 3 and 4. We have tried to stay close to the notation of [15], but some modification is required as genotypes need to be allied with individual members of a pedigree. Thus, each element of T above is a pair (i, g_i) , where g_i is the genotype of a contributing individual i . It is commonly assumed [7, 15] that $P(T|H_P) = P(T|H_D)$ using the argument that information on typed individuals will not vary between the prosecution and defense hypotheses. While this seems reasonable, we note that it is only true if:

1. The genetic marker is not linked or associated with an individual being a contributor to the mixture and so the genetic marker is not a ‘crime’ gene.
2. The two hypotheses dictate the same family relationships between the typed individuals.

Assumption 1 is reasonable and will be adopted throughout. Assumption 2 holds in many applications and is always true if the contributors are assumed to be unrelated.

When assumptions 1 and 2 both hold, we get the following simplification of the likelihood ratio in (1):

$$LR = \frac{P(R|T, H_P)}{P(R|T, H_D)}. \quad (2)$$

In our simple example, it can be seen from the two pedigrees in Figure 1 that Assumption 2 does not hold. This means that the simplification in (2) does not apply and we must work with the general expression for the likelihood ratio in (1). In [7] pairwise family relationships are handled starting from (2) by conditioning on the number of alleles shared identical by descent (IBD). However, in this example there are three related individuals and it is not practical to generalise this approach, as noted by the authors themselves [5]. Rather, we suggest calculating the required likelihoods numerically as described below and illustrated in Section 3 where we continue this example.

2.2 General problem formulation

Any hypothesis H may be formulated by specifying the individuals who have contributed to the mixture, and a pedigree describing how these individuals are related. It should be noted that the term *pedigree* is used throughout the paper in a broad sense, i.e., possibly disconnected, and possibly including *singletons* (individuals not related to anyone else). The pedigree typically contains untyped individuals who are not explicitly central to the problem, but are required to specify the relationship. For example, if we suspect that two individuals are siblings, we need to introduce their untyped parents to represent this relationship.

As noted earlier, we use a modified version of the notation in [15]:

- R : The set of distinct marker alleles in the mixture.
- T : Contributors to the mixture with known genotypes.
- V : Non-contributors to the mixture with known genotypes.
- U : The alleles of the unknown contributors to the mixture.

Individual genotypes are referenced by the individual's position in the pedigree as was seen in the motivating example above. These pedigree positions can differ under the prosecution and defense hypotheses. Specifically, the elements of the sets T and V are written as $(i, g_{i1}/g_{i2})$, where i corresponds to the label of a pedigree member and the observed alleles are given by the unordered pair g_{i1}/g_{i2} . The likelihood of interest is

$$P(R, T, V|H) = \sum_u P(U = u, T, V|H) \quad (3)$$

where the sum extends over all genotype combinations of the unknown contributors such that the result is consistent with the mixture R .

2.3 Implementation

Each term of the sum on the RHS of (3) is calculated using the Elston-Stewart algorithm, or a generalisation, for pedigree likelihood calculation. (See [12] for an overview.) We have implemented this calculation tool in the R package `euroMix` (http://arken.umb.no/~theg/euroMix_1.1.zip), which is a wrapper for another R package, `paramlink` (<http://cran.r-project.org/web/packages/paramlink>).

For some particular calculations involving the ‘theta-correction’ [16] that are not emphasised in this paper but illustrated briefly in Example 1 below, the R version of the package `Familias` (<http://cran.r-project.org/web/packages/Familias>) can be used. Existing approaches and implementations that we are aware of are based on the simplified likelihood ratio of Equation (2). By using Equation (3), we do not have to assume that different hypotheses specify identical family relationships between the typed individuals. The core of our code is to generate all sets and terms of the summation in Equation (3). This part may actually be of some interest in its own right [15]. Note that the number of possible genotypes for the unknown contributors will also affect implementations that are not concerned with general pedigrees, such as `Forensim` [3], for example.

3 Results

We present three worked examples in this section. The first one continues with the motivating example of Section 2.1 by deriving an exact expression for the *LR*. Example 2 is based on simulated data from a complex pedigree where a large number of family members, or an unrelated individual, are considered as potential contributors to the mixture. The example indicates how likely it is to identify the true perpetrator. Finally, Example 3 is based on data from a real case and shows that it is crucial to account for the family relationship.

Example 1 (Continuation of the motivating example). Recall that the problem is summarised in Figure 1 which shows the two hypothesised pedigrees and the genotypes for the two known contributors. The calculations are based on Equation (1). For each hypothesis $H \in \{H_P, H_D\}$ the probability $P(T|H)$ is computed by conditioning on the number Z of alleles that individuals 3 and 4 share identically by descent. Then

$$P(T|H) = \sum_{z=0}^2 P(Z = z|H)P(T|Z = z, H) = P(Z = 0|H)P(T|Z = 0)$$

since 3 and 4 share no alleles identical by state (IBS) and hence no alleles identical by descent. IBD probabilities for simple relationships including those for siblings (corresponding to H_P) and half-siblings (H_D) can be deduced directly or found in standard references such as [12]. In particular, $P(Z = 0|H_P) = 0.25$ and $P(Z = 0|H_D) = 0.5$. Therefore

$$\frac{P(T|H_P)}{P(T|H_D)} = \frac{0.25P(T|Z = 0)}{0.5P(T|Z = 0)} = \frac{1}{2}.$$

Figure 2: The consanguineous pedigree used in Example 2. Individual 9 is the victim, and the true perpetrator is the brother (10). After simulation of their DNA mixture, the posterior probability of contribution is computed for all pedigree members, given the mixture and the victim’s genotypes.

Furthermore, $P(R|T, H_P) = 1$ since the data on siblings 3 and 4 implies that both parents are heterozygous $1/2$ and hence the sibling 5 has to be compatible with the observed mixture alleles.

Next we consider $P(R|T, H_D)$. From the observed data on the pedigree on the right hand side of Figure 1, individual 6 must carry one 2 allele since he is the parent of 4. His second allele has to be either 1 or 2 in order to be compatible with the mixture and the pedigree under the defense hypothesis. Therefore $P(R|T, H_D) = (p_1 + p_2)$ and so

$$LR = \frac{1}{(p_1 + p_2)} \frac{1}{2}. \quad (4)$$

The factor $\frac{1}{2}$ accounts for different relationships between the typed individuals. The general case with the theta correction introduced to model deviation from Hardy-Weinberg equilibrium and, more generally, population substructure [16] leads to (details omitted):

$$LR = \frac{1 + 3\theta}{4\theta + (p_1 + p_2)(1 - \theta)} \frac{1}{2}, \quad 0 \leq \theta \leq 1. \quad (5)$$

There are numerical examples in the documentation of **euroMix** confirming Equations (4) and (5).

Example 2 (A complex pedigree). This example is based on simulated data, the advantage being that the ‘truth’ is known and it is therefore possible to assess how well the method actually performs given the ‘available’ data. There is a victim, labelled 9 in Figure 2, known to have contributed to the mixture. Based on a reference sample from the victim a DNA profile consisting of 23 markers is obtained. Allele frequencies for the markers are taken from the database available in **euroMix**. In each simulation, the

following procedure is performed: First, genotypes for 9 and 10 are simulated jointly by gene dropping. (Alleles are assigned randomly to the founders according to frequencies and Hardy-Weinberg equilibrium, and then segregated down the pedigree.) The mixture R is computed as the union of the brothers' alleles at each marker. We then pretend that we don't know the genotype of individual 10 (the true perpetrator), and instead consider every pedigree member as a second contributor to the mixture. We thus want to consider all prosecution hypotheses of the form:

- H_i : The contributors to the mixture are 9 and the relative i .

In order to take account for several competing hypotheses, it is practical to use a Bayesian framework initially [17]. A traditional frequentist approach is considered subsequently and then the defence hypothesis is

- H_{11} : The contributors to the mixture are 9 and 11.

We note that individual 11 in Figure 2 is actually unrelated to 9 and H_{11} thus corresponds to the standard hypothesis whereby the unexplained mixture alleles are assumed to be randomly drawn from the population.

Assuming a priori that all individuals are equally likely contributors, the posterior probability for each hypothesis H_i is given by:

$$P(H_i|R, T, V) = \frac{P(R, T, V|H_i)}{\sum_i P(R, T, V|H_i)} \quad (6)$$

Since the posterior probability ratio equals the likelihood ratio, the likelihood ratio for each member of the pedigree compared to the alternative of unrelated, H_{11} , is given by

$$LR_i = \frac{P(H_i|R, T, V)}{P(H_{11}|R, T, V)}. \quad (7)$$

The results of 1000 simulations are given in Table 1. For the true contributor, 10, the mean posterior probability is 0.66 and the mean $LR = 1.01e+11$ — far higher than corresponding values for any other relative. This is also reflected in that individual 10 has the highest posterior probability in 74 % of the simulations.

Figure 3 shows boxplots of the posterior probabilities from the 1000 simulations for each individual. The true contributor, 10, clearly has the overall highest posterior probability, but also the largest variation. The probability varies from values near 0 to values close to 1. The only individuals that can possibly have posterior probabilities of 0 are the parents of 9 and 10: If 9 and 10 are heterozygote for different alleles, a parent cannot possibly fit the mixture.

Example 3 (Case with real data). This example is based on a real case. The alleles have been relabelled to preserve anonymity. A rape is suspected and a DNA profile is obtained from the penis of the man. The critical issue here is whether the victim's DNA is in the profile. The mixture profile and observed genotypes for the suspect and

Individual	LR vs. unrelated	Posterior	Power
1, 2	1.27e+06	4.50e-02	0.01
3, 4	6.44e+07	1.47e-01	0.12
5, 6	1.14e+05	4.11e-02	0.02
7, 8	1.32e+09	1.04e-01	0.11
10	1.01e+11	6.61e-01	0.74
11	1.00e+00	6.25e-04	0.00

Table 1: Results of 1000 simulation experiments for Example 2. Note that some pairs of individuals, like 1 and 2, must have the same LR and posterior probability by symmetry. The LR for the stated relative versus an unrelated person is followed by the average posterior probability (flat prior) under the assumption that a given family member is the perpetrator. The average posterior probability for the true perpetrator, 10, is 0.661. The last column which we loosely refer to as ‘power’, gives the proportion of simulations for which the individual had the largest posterior probability.

victim are given in Figure 4 for 10 markers. The allele frequencies are again taken from the database used above. When the case was first analysed, the following hypotheses were suggested for consideration:

- H_P : The contributors are the suspect, victim and some unknown.
- H_{D1} : The contributors are the suspect and some unknown person.

All individuals are assumed unrelated. Having looked at the evidence for all 10 markers, it transpired that the victim and suspect shared at least one allele for all markers and it seemed obvious that the suspect was the father of the victim. This was confirmed by the party requesting the analysis. It was decided to assume this family relationship. This does not affect the prosecution hypothesis. However, the defence hypothesis can be revised to

- H_{D2} : The contributors are the suspect (father of the victim) and mother of the victim.

The results of the calculations are given in Table 2. Observe that the evidence as measured by the LR of 15385 strongly favours the prosecution if the family relationship is ignored whereas proposing that the contributors are the victim, suspect (father of victim) and mother (untyped) changes the result dramatically, yielding a LR close to 0 (0.00002). Generally, the opposing parties are free to formulate their own hypotheses. In this case, the prosecution hypothesis initially involved three unrelated individuals and this may be unlikely as no more than four alleles have been observed.

Figure 3: Boxplot of the posterior probabilities from the 1000 simulations for Example 2. The whiskers mark the minimum and maximum probabilities. The true contributor, 10, has the overall highest posterior probability, but also the largest variation. The posterior probability of 7 and 8, who are the parents of 9 and 10, can vary from 0 to almost 1.

4 Discussion

This paper provides an extension to existing methods and implementations by enabling consideration of general relationships between contributors to a DNA mixture in forensic likelihood ratio calculations. The pedigrees describing relationships between the typed individuals are allowed to differ between hypotheses. This is important as opposing parties are not obliged to agree on family relationships between contributors to a mixture. Freely available open source software has been presented and described. We have not yet encountered cases that we consider to be of potential practical interest that become computationally prohibitive. For unrelated individuals, we note that our approach is likely to be less efficient than standard methods designed specifically to deal with this case. This is mostly due to the fact that the sum in Equation (3) could potentially extend over a very large number of terms.

Our approach here has centred on calculating a likelihood ratio corresponding to specific hypotheses. In case there is uncertainty as to the number of contributors, the likelihood in Equation (3) can be viewed as a function of the number of contributors and maximised numerically to obtain the maximum likelihood estimate, see e.g. [18, 19].

We have not considered artefacts like drop-in, drop-out or stutter. Obviously, the

Figure 4: Mixtures and genotypes for suspect and victim for 10 markers (alleles have been relabelled) used in Example 3.

	LR(unrelated)	LR(related)
D3S1358	3.60	0.50
VWA	1.20	0.20
D16S539	6.60	0.20
D2S1338	1.30	0.10
D8S1179	1.70	0.50
D21S11	1.80	0.20
D18S51	2.60	0.30
D19S434	9.20	1.10
TH01	1.60	0.30
FGA	3.30	0.70
LR.all	15385.40	< 0.01

Table 2: Results of LR calculations in Example 3 corresponding to two different formulations of the defense hypotheses.

continuous data from electrophorograms can be processed prior to LR calculations as is commonly done in other implementations. Certain genotypes, deemed impossible can be excluded. For our approach this corresponds simply to omitting the corresponding terms from (3). A general approach based directly on continuous data would involve modifying the likelihood calculation. This is beyond the scope of the current paper and will be considered as future work. Implementations based on Bayesian networks provide an attractive alternative (see [1] for a recent paper). However, incorporating general relationships into a continuous model would seem to be non-trivial.

Our example based on real data shows that ignoring relationships can completely change the interpretation of the evidence. This should come as no surprise: intuitively it is much more likely that a relative of the perpetrator matches the evidence than an unrelated person.

4.1 Acknowledgement

TE and GD have received funding support from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 285487. NS acknowledges funding from the Medical Research Council (Project Grant G1002312).

References

- [1] R.G. Cowell, T. Graversen, S. Lauritzen, and J. Mortera. Analysis of DNA mixtures with artefacts. *arXiv preprint arXiv:1302.4404*, 2013.
- [2] J. Buckleton, C.M. Triggs C.M., and S.J. Walsh, editors. *Forensic DNA Evidence Interpretation*. CRC Press, Florida, USA, 2005.
- [3] H. Haned. *Forensim*: An open-source initiative for the evaluation of statistical methods in forensic genetics. *Forensic Science International: Genetics*, 5(4):265–268, 2011.
- [4] I.W. Evett. Evaluating DNA profiles in a case where the defence is "it was my brother". *Journal of the Forensic Science Society*, 32(1):5–14, 1992.
- [5] Y.Q. Hu and W.K. Fung. Interpreting DNA mixtures with the presence of relatives. *Int. J. Legal Med.*, 117:39–45, Feb 2003.
- [6] W.K. Fung and Y.Q. Hu. Interpreting DNA mixtures with related contributors in subdivided populations. *Scandinavian Journal of Statistics*, 31(1):115–130, 2004.
- [7] W.K. Fung and Hu YQ. *Statistical DNA Forensics: Theory, Methods and Computation*. Wiley, England, 2008.
- [8] Y.Q. Hu, W.K. Fung, and Y.T. Choy. Interpreting DNA mixtures with relatives of a missing suspect. In *Remote Sensing, Environment and Transportation Engineering (RSETE), 2011 International Conference on*, pages 7649–7652. IEEE, 2011.
- [9] A.P. Dawid, J. Mortera, V.L. Pascali, and D. van Boxel. Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, 29:577–595, 2002.
- [10] R.C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Human heredity*, 21(6):523–542, 1971.
- [11] C. Cannings, E.A. Thompson, and M.H. Skolnick. Probability functions on complex pedigrees. *Advances in Applied Probability*, 10:26–61, 1978.
- [12] E.A. Thompson. Statistical inference from genetic data on pedigrees. In *NSF-CBMS regional conference series in probability and statistics*. JSTOR, 2000.

- [13] A. Ziegler, I.R. König, and F. Pahlke. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an e-learning platform*. Wiley-VCh, 2010.
- [14] B.S. Weir, C.M. Triggs, L. Starling, L.I. Stowell, K.A. Walsh, and J. Buckleton. Interpreting DNA mixtures. *J. Forensic Sci.*, 42:213–222, Mar 1997.
- [15] J.M. Curran, C.M. Triggs, J. Buckleton, and B.S. Weir. Interpreting DNA mixtures in structured populations. *Journal of Forensic Sciences*, 44:987–995, 1999.
- [16] D.J. Balding and R.A. Nichols. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, 64(2):125–140, 1994.
- [17] N.A. Sheehan and T. Egeland. Structured incorporation of prior information in relationship identification problems. *Annals of Human Genetics*, 71:501–518, 2007.
- [18] T. Egeland, I. Dalen, and P.F. Mostad. Estimating the number of contributors to a DNA profile. *Int. J. Legal Med.*, 117:271–275, Oct 2003.
- [19] H. Haned, L. Pène, J.R. Lobry, A.B. Dufour, and D. Pontier. Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count? *Journal of Forensic Sciences*, 56:23–28, 2009.

5 Supplementary material. Worked examples in R/euroMix

In the following two examples we provide details about the implementation of the software. The first example deals with pairwise kinship and is based on material presented in [7, Chapter 7], in which pairwise IBD probabilities are used to compute likelihood ratios. We reproduce some results reported in that chapter with our software and give the relevant R commands needed for the computations. The second example involves kinship between three individuals and therefore cannot be handled with kinship coefficients. Further examples with code follow thereafter.

Example S-1 (Relationship between two individuals). The data are taken from a Hong Kong rape case. As noted previously, it is sufficient to consider a single genetic marker to convey the basic ideas, and so we will only use the FGA marker. The genetic data comprise a mixture and individual DNA profiles from a victim and a suspect. For simplicity, we have renumbered the FGA alleles 20, 24, 25 and a 'pooled allele' as 1, 2, 3 and 4, respectively, and the corresponding allele frequencies are $(p_1, p_2, p_3, p_4) = (0.044, 0.166, 0.110, 0.68)$. Note that lumping all unobserved alleles into one pooled allele does not affect the calculations in this case, but this always has to be checked for every application.

The prosecution and defense hypotheses are

- H_P : The victim and the suspect contributed to the mixture.
- H_D : The victim and a relative of the suspect contributed to the mixture.

Several alternatives for the suspect's relative are considered in [7]. Here we consider

- H_D : The victim and the father of the suspect contributed to the mixture.

The victim is assumed to be unrelated to any other contributor. Figure 5 shows the pedigree that is assumed under both hypotheses. The mixture evidence consists of the allele set $R = \{1, 2, 3\}$. The victim's genotype is 1/2 and the suspect's genotype is 3/3.

Figure 5: The pedigree in Example S-1.

```

require(euroMix) 1
alleles <- 1:4 2
afreq <- c(0.044, 0.166, 0.11, 0.68) 3
R <- 1:3 4
x <- nuclearPed(1) 5
y <- singleton(4, sex=2) 6
known <- list(c(3,3,3), c(4,1,2)) 7
lp <- paraMix(list(x,y), R, id.U=NULL, id.V=NULL, alleles=alleles, 8
              afreq=afreq, known=known)$likelihood 9
ld <- paraMix(list(x,y), R, id.U=1, id.V=3, alleles=alleles, 10
              afreq=afreq, known=known)$likelihood 11
lp/ld 12

```

Table 3: R commands for Example S-1.

H_P assumes individuals 3 and 4 to be the contributors, while H_D assumes 1 and 4 to be the contributors. Table 3 shows how to calculate the LR in R.

We go through the commands one by one. The first line loads the packages `euroMix` (which loads `paramlink`) assuming that this has already been installed as explained in Section 2.3. In lines 2-4 the marker alleles, the allele frequencies and the mixture R are defined. Next, lines 5 and 6 create the pedigree as shown in Figure 5: First the trio (a nuclear pedigree with 1 child), and then the victim (she is assumed unrelated to everyone else, hence a singleton). The known genotypes are specified in line 7 as a list of triples. For instance, the triple (4, 1, 2) means that individual 4 has genotype 1/2. In lines 8-9, the likelihood for the prosecution hypothesis is calculated using the core function `paraMix`. The argument `id.U=NULL` signifies that the set U is empty, i.e., that there are no unknown contributors under H_P . Similarly we set `id.V=NULL` since there are no genotyped non-contributors under this hypothesis (i.e. $V = \emptyset$). For the defense hypothesis, in lines 10-11, the argument `id.U=1` says that individual 1 (the suspect's father) is an unknown contributor. Furthermore, individual 3 (the suspect) is a non-contributor under this hypothesis. He belongs in the set V since he has known genotype, thus we set `id.V=3`. The final line produces the LR .

Example S-2 (Kinship between three individuals). In this example we consider the same structure of the pedigree and the same hypotheses as in the previous example, but with an extra genotyped individual in the pedigree. The mixture evidence alleles and the victim's genotype is the same as before, however the suspect's genotype is now 1/3, and in addition we know that the mother of the suspect has genotype 2/3. Figure 6 shows the pedigree with the known markers. The prosecution assumes individuals 3 and 4 to be the contributors to the mixture, while the defense assumes individuals 1 and 4 to be the contributors.

The R commands for producing the LR are as follows. Lines 1-6 are identical to Table 3 and are not shown here.

Figure 6: The pedigree in Example S-2. Individual 2 is genotyped, but assumed to be a non-contributor under both hypotheses.

```
known <- list(c(2,2,3), c(3,1,3), c(4,1,2))          7
lp <- paraMix(list(x,y), R, id.U=NULL, id.V=2, alleles=alleles,  8
               afreq=afreq, known=known)$likelihood          9
ld <- paraMix(list(x,y), R, id.U=1, id.V=c(2,3), alleles=alleles, 10
               afreq=afreq, known=known)$likelihood          11
lp/ld  # = 9.091                                           12
```

The only differences from the previous example are the known genotypes, and the inclusion of the mother (individual 2) in the `id.V` argument in both likelihood calculations. The final line results in a LR of 9.091, which agrees with the exact formula $1/p_3$ (details omitted).

To see the effect the mother's genotype has on the likelihood ratio, we recalculate the LR , but this time ignoring the mother's genotype. Without this additional genotype information, the evidence in support of the prosecution hypothesis is weakened, producing a likelihood ratio of 4.651. The commands are as follows:

```
known <- list(c(3,1,3), c(4,1,2))                    7
lp <- paraMix(list(x,y), R, id.U=NULL, alleles=alleles,  8
               afreq=afreq, known=known)$likelihood          9
ld <- paraMix(list(x,y), R, id.U=1, id.V=3, alleles=alleles, 10
               afreq=afreq, known=known)$likelihood          11
lp/ld  # = 4.651                                           12
```

5.1 Additional examples

In the next sections we explore variations on the rape case considered above. In particular the same marker is used throughout, the mixture is $R = \{1, 2, 3\}$, the victim is a singleton with genotype $1/2$, and the suspect has genotype $3/3$. We will let the victim have ID label 99 to avoid conflicting labels in the suspects pedigree. The ID of the suspect will vary from case to case. but the following setup is the same for all :


```

require(euroMix)
alleles <- 1:4
afreq <- c(0.044, 0.166, 0.11, 0.68)
R <- 1:3
vic <- singleton(99, sex=2)
vic.geno <- c(99, 1, 2)

```

In the last example we also show a computation with theta correction.

5.1.1 Relative of the suspect

We first consider variations of the defense hypothesis proposed in Example S-1, where the relationship between the suspect and the relative is varied. The hypotheses proposed are

- H_P : The victim and the suspect contributed to the mixture.
- H_D : The victim and a relative of the suspect contributed to the mixture.

Example S-3 (Full sibling). The defence claims that a sibling of the suspect contributed to the mixture. This gives $LR = 2.36$, computed as follows:

```

x <- nuclearPed(2)          # Suspect is 3, sib is 4
known <- list(c(3,3,3), vic.geno)
lp <- paraMix(list(x,vic), R, id.U=NULL, id.V=NULL, alleles=alleles,
              afreq=afreq, known=known)$likelihood
ld <- paraMix(list(x,vic), R, id.U=4, id.V=3, alleles=alleles,
              afreq=afreq, known=known)$likelihood
lp/ld

```

Example S-4 (Half sibling). The defence claims that a half sibling of the suspect contributed to the mixture. This gives $LR = 5.29$, computed as follows:

```

x <- halfCousinPed(0)       # Suspect is 4, half sib is 5
known <- list(c(4,3,3), vic.geno)
lp <- paraMix(list(x,vic), R, id.U=NULL, id.V=NULL, alleles=alleles,
              afreq=afreq, known=known)$likelihood
ld <- paraMix(list(x,vic), R, id.U=5, id.V=4, alleles=alleles,
              afreq=afreq, known=known)$likelihood
lp/ld

```

Example S-5 (First cousin). The defence claims that a first cousin of the suspect contributed to the mixture. This gives $LR = 8.08$, computed as follows:

```

x <- cousinPed(1)           # Suspect is 7, cousin is 8      7
known <- list(c(7,3,3), vic.geno)                             8
lp <- paraMix(list(x,vic), R, id.U=NULL, id.V=NULL, alleles=alleles, 9
              afreq=afreq, known=known)$likelihood           10
ld <- paraMix(list(x,vic), R, id.U=8, id.V=7, alleles=alleles,    11
              afreq=afreq, known=known)$likelihood             12
lp/ld                                                                13

```

Example S-6 (Second cousin). The defence claims that a second cousin of the suspect contributed to the mixture. This gives $LR = 13.39$, computed as follows:

```

x <- cousinPed(2)           # Suspect is 11, 2nd cousin is 12  7
known <- list(c(11,3,3), vic.geno)                             8
lp <- paraMix(list(x,vic), R, id.U=NULL, id.V=NULL, alleles=alleles, 9
              afreq=afreq, known=known)$likelihood           10
ld <- paraMix(list(x,vic), R, id.U=12, id.V=11, alleles=alleles,    11
              afreq=afreq, known=known)$likelihood             12
lp/ld                                                                13

```

Example S-7 (Unrelated). The defence claims that a person unrelated to the suspect contributed to the mixture. This gives $LR = 17.15$, computed as follows:

```

x <- singleton(2)           # Suspect                          7
y <- singleton(3)           # The unrelated person              8
known <- list(c(2,3,3), vic.geno)                             9
lp <- paraMix(list(x,y,vic), R, id.U=NULL, id.V=NULL, alleles=alleles, 10
              afreq=afreq, known=known)$likelihood           11
ld <- paraMix(list(x,y,vic), R, id.U=3, id.V=2, alleles=alleles,    12
              afreq=afreq, known=known)$likelihood             13
lp/ld                                                                14

```

5.1.2 Unknown and relative of the suspect

Next, we consider hypotheses where the defense proposes that the contributors are the relative of the suspect and one unrelated unknown.

- H_P : The victim and the suspect contributed to the mixture.
- H_D : A relative of the suspect and an unrelated unknown contributed to the mixture.

The relationship between the suspect and the relative is varied. In the R scripts, the unknown individual proposed by the defense is created as a female singleton with ID label 10.

Example S-8 (Parent). The defence claims that the father of the suspect, and an unrelated unknown, contributed to the mixture. This gives $LR = 106.13$, computed as follows:

```
x <- nuclearPed(1)           # Suspect is 3, father is 1      7
y <- singleton(10, sex=2)     # Unrelated unknown            8
known <- list(c(3,3,3), vic.geno)                               9
lp <- paraMix(list(x,y,vic), R, id.U=NULL, alleles=alleles,    10
               afreq=afreq, known=known)$likelihood           11
ld <- paraMix(list(x,y,vic), R, id.U=c(2,10), id.V=c(3,99),    12
               alleles=alleles, afreq=afreq, known=known)$likelihood 13
lp/ld                                                                14
```

Example S-9 (Full sibling). The defence claims that a sibling of the suspect, and an unrelated unknown, contributed to the mixture. This gives $LR = 109.48$, computed as follows:

```
x <- nuclearPed(2)           # Suspect is 3, sibling is 4      7
y <- singleton(10, sex=2)     # Unrelated unknown            8
known <- list(c(3,3,3), vic.geno)                               9
lp <- paraMix(list(x,y,vic), R, id.U=NULL, alleles=alleles,    10
               afreq=afreq, known=known)$likelihood           11
ld <- paraMix(list(x,y,vic), R, id.U=c(4,10), id.V=c(3,99),    12
               alleles=alleles, afreq=afreq, known=known)$likelihood 13
lp/ld                                                                14
```

5.1.3 Two related unknowns

Finally, we consider hypotheses where the defense proposes that the contributors are two untyped related persons, who are unrelated to both the suspect and the victim.

- H_P : The victim and the suspect contributed to the mixture.
- H_D : Two unknown relatives contributed to the mixture.

The relationship between the relatives in the defense hypothesis is varied. Note that the suspect is a singleton in these cases.

Example S-10 (Parent-child). The defence claims that an unknown male and his son contributed to the mixture. This gives $LR = 207.44$, computed as follows:

```
x <- nuclearPed(1)           # Unknown father and son are 1 and 3  7
y <- singleton(5)            # The suspect                        8
known <- list(c(5,3,3), vic.geno)                               9
lp <- paraMix(list(x,y,vic), R, id.U=NULL, alleles=alleles,    10
               afreq=afreq, known=known)$likelihood           11
ld <- paraMix(list(x,y,vic), R, id.U=c(1,3), id.V=c(5,99),    12
               alleles=alleles, afreq=afreq, known=known)$likelihood 13
lp/ld                                                                14
```

Example S-11 (Full siblings). The defence claims that two unknown siblings contributed to the mixture. This gives $LR = 314.31$, computed as follows:

```
x <- nuclearPed(2)           # Unknown siblings are 3 and 4      7
y <- singleton(5)           # The suspect                        8
known <- list(c(5,3,3), vic.geno)                               9
lp <- paraMix(list(x,y,vic), R, id.U=NULL, alleles=alleles,    10
              afreq=afreq, known=known)$likelihood            11
ld <- paraMix(list(x,y,vic), R, id.U=c(3,4), id.V=c(5,99),    12
              alleles=alleles, afreq=afreq, known=known)$likelihood 13
lp/ld                                                                14
```

5.1.4 Parent-child with theta correction

Example S-12. We consider the same father-child relationship as in Example S-1, but with $\theta = 0.01$. To apply theta correction, we use the `euroMix` function `famMix` which relies on the R version of `Familias`. Unlike `paraMix`, `famMix` can not handle singletons, and unrelated individuals must thus be specified as part of the pedigree. The following commands were used to produce the likelihood ratio of 2.89:

```
require(Familias)                                                5
theta <- 0.01                                                    6
x <- nuclearPed(1)                                               7
x <- addOffspring(x,father=1,noff=1)                             8
m1 <- marker(x, 4, c(1,2), 3, c(3,3), alleles=alleles, afreq=afreq) 9
lp <- famMix(x, R, id.U=NULL, partialmarker=m1,                 10
            theta=theta)$likelihood                             11
ld <- famMix(x, R, id.U=1, id.V=3, partialmarker=m1,            12
            theta=theta)$likelihood                             13
lp/ld                                                            14
```