Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2001/06

The evaluation of DNA evidence in pedigrees requiring population inference

Fabio Corradi, Giampietro Lago
and Federico M. Stefanini

Università degli Studi
di Firenze

Statistics, Biostatistics

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

The evaluation of DNA evidence in pedigrees requiring population inference

Fabio Corradi[1], Giampietro Lago and Federico M. Stefanini

[1]Department of Statistics, University of Florence,

Viale Morgagni 59, 50134, Florence, Italy,

e.mail: corradi@ds.unifi.it

Tel. +39-055-4237 209 Fax +39-055-4223560,

Keywords:Weight of evidence, pedigrees, graphs, Bayesian Inference

12 March 2001

Abstract

The evaluation of nuclear DNA evidences for identification purposes is here performed taking account of the uncertainty about population parameters. Graphical models are used to detail out the hypotheses under forensic debate, those that determine the pedigree structure. Graphs clarify the set of evidences that contribute to population inferences and they also describe the conditional independence structure of DNA evidences. Numerical illustrations are provided by reexamining three case studies taken from the literature. Our calculations of the weight of evidence differ from those given by the authors of case studies in the direction of more conservative values.

1 Introduction

The use of nuclear DNA in paternity test, identification of missing persons, siblings recognition is gaining increasing attention in courts. A careful statistical analysis of DNA fingerprints in forensic science is an issue for the consequences it embraces, from considerations about the cost of assessment to ethical concerns.

The evaluation of the weight of evidence (WE) is the core of forensic identification. Here the joint probability of observing the genotypes that constitute observed evidences is quantified in turn by conditioning on pairs of hypotheses which are of forensic interest. Each hypothesis specifies a kinship scheme among individuals so that the two hypotheses define different conditional independence structures among genotypes. The ratio of the two probability values is called WE and it is indicated as W_E in formulas.

A comprehensive discussion of methods and pitfalls related to the evaluation of DNA evidences in forensic science is given by Evett and Weir [5]. Essen-Möller [4] deals with paternity tests and Ihm and Hummel [7] calculate the WE for common kinship schemes. Brenner [2] has developed an original symbolic program to obtain formulas of the WE. Mostad and Egeland follow the Ihm and Hummel track and numerically evaluate the WE in a large variety of kinship schemes, including those involving family reconstruction.

Entry points in the literature dealing with probability calculus for pedigrees are Lange [9, chapter 7] and Ott [12], the last one focused on linkage analysis.

In this field, Bayesian methods are often invoked since the WE may be interpreted as the updating factor that map hypotheses prior odds to posterior odds. Nevertheless, the Bayesian paradigm is not strictly followed because the uncertainty on population parameters is typically neglected while calculating the WE. Point estimates based on a reference population sample are often plugged into the model instead.

We believe that population parameters may be coherently included among the unknowns, and this implies that the reference population sample is considered as part of the available evidences, together with case evidences. The WE arises from the joint probability associated to case and population sample evidences when they are calculated under different forensic hypotheses. An example of joint use of sample and case evidence is given in Dawid and Mortera [3] that deal with a suspect-crime identification problem.

The probabilistic analysis is here achieved making use of graphs to express the substantive (research) hypotheses Wermuth and Lauritzen [14] under forensic debate and a related conditional independence structure.

The consequences of coherently considering all the unknowns in the forensic debate are described also providing illustrative examples. In all the case studies we found that our method gives a smaller WE then the 'plug-in' approach. The issues considered in this paper have important consequences in the forensic use of nuclear DNA evidences. Our software implementation is freely available (WEB site http://bayes1.ds.unifi.it/RaCis/ini.html).

2 Preliminaries on the nature of evidences

Sometimes, in forensic identification based on nuclear DNA, the crime and/or the suspect evidences are missing. A group of individuals genetically related to the missing evidence is also typed at several genetic loci: these are called case evidences.

The kinship among some members of this group is assumed to be known, because this information does not enter the forensic dispute as a questionable fact. Nevertheless, the full specification of kinship among group members is only achieved making at least two alternative hypotheses, and these are to be debated in the court.

Besides the case evidences, a sample (of evidences) is taken from a reference population defined by auxiliary information strictly related to the case (spatial location, race, etc.). The sample is often not random (convenience samples, [5] pp 44). Nevertheless, the observations are considered to be exchangeable if weak information on the genetic structure of the population makes hard further distinctions. Anyway, cares are taken to avoid the selection of strictly related individuals in the sample, e.g. siblings are non included in the sample to avoid the over-representation of highly correlated genotypes.

2.1 Notation

In forensic studies based on nuclear DNA, several polymorphic genetic loci are considered, i.e. those that have two or more alleles maintained in the population.

Let $\mathcal{L} = \{1, 2, \ldots\}$ be the set of natural numbers associated to the genetic loci that are included in a study, and $n(\mathcal{L})$ its cardinality (from now on $n(.)$ stands for cardinality of). Let $\mathcal{A}_l = \{a_{l,1}, \ldots, a_{l,i}, \ldots\}$ be the set of alleles at locus $l \in \mathcal{L}$. An individual carries two alleles at one locus, and this unordered pair is called genotype. Let $\mathcal{G}_l = \{g : (g \subset \mathcal{A}_l) \cap (1 \leq n(g) \leq 2)\}$ be the sample space of genotypes at locus l. Then, the sample space of genotypes at $n(\mathcal{L})$ loci is $\mathcal{G}_{\mathcal{L}} = \{(g_1, \ldots, g_l, \ldots) : g_l \in \mathcal{G}_l, l \in \mathcal{L}\}$, and its cardinality is $n(\mathcal{G}_{\mathcal{L}})$.

We define X_l to be the random vector that associate an ordered pair of integers to a generic genotype g:

$$X_l\,(g) = (i, j), i \leq j, \text{ whenever } g = \{a_{l,i}, a_{l,j}\} \tag{1}$$

By definition, the random vector X_l is unaffected by order in the subset of alleles, namely $X_l\,(\{a_{l,i}, a_{l,j}\}) = X_l\,(\{a_{l,j}, a_{l,i}\})$. The range of X_l is $\mathcal{X}_l = \{(i, j) : 1 \leq i \leq j \leq n(\mathcal{A}_l)\}$. A realization of the random vector X_l is $x_l \in \mathcal{X}_l$. If $n(\mathcal{L}) > 1$ than several loci are considered and the range is given by the Cartesian product $\mathcal{X}_{\mathcal{L}} = \times_{l \in \mathcal{L}} \mathcal{X}_l$.

A (reference) population is biologically defined as a finite collection of genotypes, e.g. $\{g_j : g_j = (g_{j,1}, \ldots, g_{j,l}, \ldots) \in \mathcal{G}_{\mathcal{L}}\}$.

The probability of sampling an allele $a_{l,i}$ from a reference population at locus l is given by its relative frequency in the reference population. Let $\theta_l = \{\theta_{l,i} : i = 1, 2, \ldots, n(\mathcal{A}_l)\}$ be the vector of allele frequencies at locus l. If one locus is considered, we will neglect the locus index l. If several reference populations would be considered, then $\theta_{l,p}$ would indicate the vector of parameters at locus l in the reference population p. If indices l, p are neglected in the formulas then one locus and one reference population is considered.

2.2 The kinship of individuals

The relatedness of genotypes may be described by a graph. Although graphs called pedigrees are well known and widely used in genetics [12], here we will use the definition commonly adopted in statistics under the label of graphical models [10].

By a graph we mean a pair $\mathcal{G} = (V, E)$ in which V is a finite set of vertices and $E \subset V \times V$ is the set of edges. The set V may be considered as a list of labels referred to random vectors associated to the genotype of individuals and to all the other unobserved quantities, like the parameter vector θ, required to perform the calculation of the WE. The set E is a subset of the Cartesian product $V \times V$, in which $(a, b) \in E$ implies $a \neq b$. If $(a, b) \in E$ and $(b, a) \notin E$ than X_b is a descendant of X_a, thus the parent-offspring relationship $a \rightarrow b$ holds. An undirected association between X_a and X_b is represented by the two pairs (a, b) and (b, a) both in E, and in this case an undirected edge joins a and b.

A genetic sub-system made by individuals, and possibly other unobserved components, is indexed by the subset of vertices $S \subset V$. The subgraph induced by S is $\mathcal{G}_S = (S, E_S)$, where $E_S = E \cap (S \times S)$. The subset S also induces the random vector $X_S = \{X_i : i \in S \subset V\}$.

Given a set of vertices S, the boundary $bd(S)$ of S is defined as $bd(S) = \{a \in V \setminus S \mid$

$\exists b \in S, (a,b) \in E\}$. In a special case S contains just one vertex s and edges of the boundary are all directed to s: they are called parents of s, $pa(s)$.

A graph is called chain graph if the set of vertices can be partitioned in a ordered collection of subsets $(V_1, \ldots, V_k)$ called chain components. Edges within a chain component are undirected and edges joining vertices located in different chain components are directed.

It will be always assumed that an observed genotype is not affected by measurement errors and that it may be directly assessed (codominant markers). This is typically the case for microsatellites genetic loci, widely used in forensic applications.

In graphs, we distinguish the observed from the unobserved random variables, the former indicated by a black dot vertex, the latter by an empty dot vertex.

3 The weight of evidence for competing hypotheses

The information used to build a graph $\mathcal{G}$ is called informative set, indicated as $\mathcal{I}$.

At the most basic level, the informative set $\mathcal{I}_0$ collects all the information related to the forensic case that is not matter of debate. Given $\mathcal{I}_0$, the set V_0 is defined by the collection of vertices representing observed DNA evidences (case and reference population sample evidences). The set $\mathcal{I}_0$ is also used to define the relationships among vertices in V_0 and these constitute the set E_0.

The graph $\mathcal{G}_0 = (V_0, E_0)$ based on $\mathcal{I}_0$ is a chain graph. Undirected edges join genotypes that are related weather they came from the same family or they belong to the same reference population. Directed edges relate genotypes that belong to the same lineage by the parents-to-offspring Mendelian law. It must be remarked that conditionally on both parents' genotypes be known, their offsprings' genotypes are independent, thus directed edges suffice to account for stochastic dependence.

The graph $\mathcal{G}_0$ is not directly suited to calculate the probability of observing the collected evidences because $\mathcal{I}_0$ does not include those kinship relations that are debated and needed to fully specify the conditional independence structure. The working hypothesis $\mathcal{I}_j$ extends $\mathcal{I}_0$ through additional information.

If the working hypothesis $\mathcal{I}_j$ deals with samples taken from several reference populations, then it introduces at least one parametric model. A first expansion of the set of vertices V_0 is performed by adding vertices $\{r_1, r_2, \dots\}$ to label the vectors of population parameters. Since we will focus on just one reference population, the special role played by the only population parameter vector is highlighted in formulas by using θ instead of X_{r_1} and using the vertex label r in graphs. An undirected edge joining two genotypes that do not share recent common ancestors is substituted by directed edges from the population parameter to each genotype.

The formal distinction among recently related and recently unrelated genotypes is obtained by introducing a partition of the set of vertices V_0: founder vertices, F_0, and descendant vertices, $\overline{F}_0$ (the symbol o stands for 'observed evidence'). We remark that members of the reference population sample are always considered as founders. We simplify the graph representation by introducing $D = (d_1, d_2, \dots)$ as the set of founders taken from the reference population sample (for example Figure 1). The subgraph $r \rightarrow D$ may be expanded into a detailed representation of vertices in D, that is $r \rightarrow d_1$, $r \rightarrow d_2$, etc.. Moreover, if both parents of a vertex v are unobserved and no information is provided about their kinship (e.g. brothers) then v is a founder, $v \in F_0$. Otherwise as many vertices (and arrows) are introduced as required to have the oldest individuals of each lineage in the set of founders. Within $\overline{F}_0$, a generic vertex v might refer to a genotype that has just one parent observed. Then, one vertex is introduced for the unobserved parent to fill-in the pair of parents, according to a Mendeleian representation.

The fill-in procedure is repeated for each vertex in $\overline{F}_0$. If some vertices are introduced during the fill-in, than the procedure is iterated until no more vertices are added and all the vertices either have both parents specified (observed or unobserved) or they are founders.

The result of the fill-in step is that vertices connected by undirected edges in $\mathcal{G}_0$ are substituted by arrows from observed/unobserved parents to their children.

The whole procedure given above produces a special kind of graph given a working hypothesis: a directed acyclic graph (DAG). It is characterized by the lack of undirected edges, that is $(a, b) \in E$ implies $(b, a) \notin E$.

Given two working hypotheses, $\mathcal{I}_1, \mathcal{I}_2$, the joint probability of observing a realization of the random vector labeled by the vertices in $\mathcal{G}_j$ may be calculated for $j = 1, 2$, namely $p(X_{V_j} \mid \mathcal{G}_j)$.

The marginal conditional probability on the observed vertices is obtained by (Lebesgue) integration with respect to the appropriate vertices, as specified by the working hypothesis $\mathcal{I}_j$, that is

$$p(X_{V_0} \mid \mathcal{I}_j) = \int_{\mathcal{X}_{V_j \setminus V_0}} dF(X_{V_j}), \tag{2}$$

where $V_j \setminus V_0$ is the set of labels for unobserved genotypes, and $\mathcal{X}_{V_j \setminus V_0}$ the sample space.

The WE of the working hypothesis $\mathcal{I}_1$ against the working hypothesis $\mathcal{I}_2$ is calculated as the ratio among the marginal conditional probability values:

$$W_E = \frac{p(X_{V_0} \mid \mathcal{I}_1)}{p(X_{V_0} \mid \mathcal{I}_2)} \tag{3}$$

Actual calculations of WE require the choice of specific probability models, as described in the next section.

3.1 Probability models for DAG subgraphs

The joint probability of observing genotypes labelled by vertices of a DAG is [6, 10]:

$$p(X_{V_j} \mid \mathcal{I}_j) = \prod_{v \in V_j} p(X_v \mid X_{pa(v)}, \mathcal{I}_j). \tag{4}$$

We first consider just one locus, i.e. $\mathcal{L} = \{1\}$, while specifying conditional distributions in equation (4), because calculations for several loci, i.e. $n(\mathcal{L}) > \{1\}$, may be performed by exploiting the conditional independence structure in equations (8) and (9):

$$p(X_{V_j} \mid \mathcal{I}_j) = \prod_{l \in \mathcal{L}} \prod_{v \in V_j} p(X_{l,v} \mid X_{l,pa(v)}, \mathcal{I}_j). \tag{5}$$

If the population is at Hardy-Weinberg equilibrium [9, pp2], then the genotipic relative frequency may be calculated from the alleles relative frequency in the population:

$$\mathbf{Pr}[X_l = (i,j) | \theta_l] = \theta_{l,i} \cdot \theta_{l,j} + \theta_{l,i} \cdot \theta_{l,j} \cdot \mathbf{I}_{\{i \neq j\}}(i,j). \tag{6}$$

The probability mass function defined in equation (6) will be indicated as $Gen(X_l \mid \theta_l)$, where θ_l,$\forall l \in \mathcal{L}$, are typically unknown. It defines the probability of observing the genotype (i,j) in a individual of the founders. The equation (6) is a compact notation for the multinomial sampling model from an urn containing alleles where the number of sampled alleles is 2.

As regards the elements of $\overline{F}$, each vertex has two parents specified by the graph $\mathcal{G}_j$. For example, $X_p \to X_o$ and $X_q \to X_o$ are the two directed edges that indicate parents (p and q) to offspring (o) relationship. The Mendel law on alleles segregation at one locus, specifies the probability of the event $X_o = (x_{o,a}, x_{o,b})$ given the values $X_p = (x_{p,i}, x_{p,j})$ and $X_q = (x_{q,r}, x_{q,s})$ of parents:

$$\begin{aligned} \mathbf{Pr}[X_o \mid X_p, X_q] = \big\{ & \mathbf{I}_{\{x_{p,i}, x_{p,j}\}}(x_{o,a}) \cdot \mathbf{I}_{\{x_{q,r}, x_{q,s}\}}(x_{o,b}) + \mathbf{I}_{\{x_{p,i}, x_{p,j}\}}(x_{o,b}) \cdot \\ & \mathbf{I}_{\{x_{q,r}, x_{q,s}\}}(x_{o,a}) \big\} \cdot \frac{1}{2^{1 - \mathbf{I}_{(u,u)}(x_{p,i}, x_{p,j})}} \cdot \frac{1}{2^{1 - \mathbf{I}_{(z,z)}(x_{q,r}, x_{q,s})}} \end{aligned} \tag{7}$$

where the terms in the curly brackets account for the combinatorial aspect of the Mendel law, and the last two factors determine the probability value.

The probability mass function defined in equation (7) will be indicated as $Des(X_{l,o} \mid X_{l,a}, X_{l,b})$ to remark that it deals with probability values of offspring's genotype.

The extension of equation (7) to several loci, $n(\mathcal{L}) > 1$, is based on the conditional independence of alleles at loci that are located on different chromosomes.

Let $X_o = (x_{o,1,a}, x_{o,1,b}, \ldots, x_{o,l,a}, x_{o,l,b}, \ldots)$ be the offspring multilocus genotype, and $X_p = (x_{p,1,i},\, x_{p,1,j}, \ldots,\, x_{p,l,i}, x_{p,l,j}, \ldots)$, and $X_q = (x_{q,1,r}, x_{q,1,s}, \ldots, x_{q,l,r}, x_{q,l,s}, \ldots)$ be the parents genotypes. According to the Mendel law of independent assortment, the conditional probability of obtaining the genotype x_o is:

$$\begin{aligned}&\mathbf{Pr}[X_o \mid X_p, X_q] = \qquad (8)\\ &\quad = \prod_{l \in \mathcal{L}} \mathbf{Pr}[X_{o,l} \quad = \quad (x_{o,l,a}, x_{o,l,b}) \mid X_{p,l} = (x_{p,l,i}, x_{p,l,j}), X_{q,l} = (x_{q,l,r}, x_{q,l,s})].\end{aligned}$$

Equation (8) approximately holds if genetic loci are located on the same chromosome but their genetic distance is large (say about 45 centiMorgan or more).

The probability of sampling a given genotype from the reference population can be factored in a simple expression under the assumption of genetic equilibrium in the population. Let $\theta = \{\theta_l : l = 1, 2, \ldots, n(\mathcal{L})\}$ be the vector of parameters for $n(\mathcal{L})$ loci. The probability of sampling a given genotype at one of the founder nodes v is:

$$\mathbf{Pr}[X_v = (x_{v,1,a}, x_{v,1,b}, , \ldots, x_{v,l,a}, x_{v,l,b}, \ldots)|\theta] = \prod_{l \in \mathcal{L}} \mathbf{Pr}[X_{v,l} = (x_{v,l,a}, x_{v,l,b}) \mid \theta_l]. \quad (9)$$

We will always assume that the population is at equilibrium, but we mention in the discussion section the path to be followed if the assumption is relaxed.

3.2 The probability of observed evidence

The evaluation of equation (3) is performed taking account of nodes features: founders (F) or not founders ($\overline{F}$) individuals, observed (o) or unobserved (u) genotypes.

We give now details on the computation of the conditional distribution in (2) by distinguishing nodes according to the partition F,$\overline{F}$:

$$p(X_{V_j} \mid \mathcal{I}_j) = p(\theta \mid \mathcal{I}_j) \cdot \prod_{k \in \overline{F}} Des(X_k \mid X_{pa(k)}, \mathcal{I}_j) \cdot \prod_{h \in F} Gen(X_h \mid \theta, \mathcal{I}_j), \tag{10}$$

where again $X_r \equiv \theta, r \in V_j$.

We choose a Dirichlet prior distribution for θ, with parameter vector α, as a reasonable compromise between simple and realistic models. Closed-form calculations depends on our choice of a Dirichlet as family of prior distributions but any other reasonable choice might be performed switching to Monte Carlo computation.

Values of vector α are selected to have an expected value of each population allele frequency equal to $\frac{1}{n(\mathcal{A})}$, the number of locus alleles, and a strength of prior belief equal to 1 (interpreted as 'prior sample size').

From equation (10), by keeping the distinction between observed (o) and unobserved (u) random variables, we rewrite the joint probability associated to graph $\mathcal{G}_j$ as:

$$\begin{aligned} p(X_{V_j} \mid \mathcal{I}_j) \quad = \quad & \prod_{k \in \overline{F}_o} Des(x_k \mid X_{pa(k)}, \mathcal{I}_j) && (11) \\ \cdot \quad & \prod_{w \in \overline{F}_u} Des(X_w = x_w \mid X_{pa(w)}, \mathcal{I}_j) && (12) \\ \cdot \quad & \prod_{h \in F_u} Gen(X_h = x_h \mid \theta, \mathcal{I}_j) && (13) \\ \cdot \quad & \prod_{z \in F_o} Gen(x_z \mid \theta, \mathcal{I}_j) \cdot p(\theta \mid \mathcal{I}_j), && (14) \end{aligned}$$

where capital letters are used for unobserved random variables (notation given in section 2.1).

The (Lebesgue) integration specified in equation (2) involves variables representing unobserved nodes: summation for genotypes and integration for the population parameters.

The calculation conveniently starts with X_w for $w \in \overline{F}_u$, thus summations involve all unobserved nodes among the set of founders' descendants.

The integration of θ concerns factors displayed in lines (13) and (14). The calculation is clarified by rewriting factors in line (14) as a product of the posterior distribution of θ (conditionally to founder nodes) times the marginal probability of the observed founders:

$$\prod_{z \in F_o} Gen(x_z \mid \theta, \mathcal{I}_j) \cdot p(\theta \mid \mathcal{I}_j) = Dirichlet(\theta \mid \alpha^*) \cdot \prod_{z \in F_o} Md(x_z \mid 2, \widetilde{\alpha}_z), \tag{15}$$

where $Md(\cdot)$ stands for the multinomial-Dirichlet distribution; α^* is a vector of elements $\alpha_i^* = \alpha_i + n_i$, and n_i is the number of alleles of type a_i in the set of founders; the vector $\widetilde{\alpha}_z =$ has elements $\widetilde{\alpha}_{i,z} = \alpha_i + n(i,z)$, where $n(i,z)$ is the number of alleles of type a_i observed up to the evidence labeled by $z-1$ in any given ordered sequence $\mathcal{Z}$ of vertices in F_o. Note that the posterior distribution of θ is obtained by sequential application of the Bayes theorem. This is required since the observational model is parameterized in terms of population allele frequencies: in order to obtain the closed-form posterior distribution of θ, genotypes must be considered one at a time.

The proposition clarify how to perform the integration step.

PROPOSITION: Let $\mathcal{Z}$ be any ordered sequence labeling the elements of F_u. Let $n_{i,u}$ be the number of observed alleles of type a_i in the set of evidences pertaining F_u, and $n_u = \sum_i^{n_{\mathcal{A}}} n_{i,u}$, where $n(\mathcal{A})$ is the number of different alleles at the considered locus. Let $n_{i,h} \in \{0,1,2\}$ be the number of alleles of type a_i scored on the individual labeled by $h \in F_u$, and $n_h = \sum_{i=1}^{n(\mathcal{A})} n_{i,h} = 2$ the sum calculated over $h \in F_u$. Then the identity *below holds:*

$$\int \prod_{h\in F_u} Gen(X_h = x_h \mid \theta, \mathcal{I}_j) \cdot \prod_{z\in F_o} Gen(x_z \mid \theta, \mathcal{I}_j) \cdot p(\theta \mid \mathcal{I}_j) \cdot d\theta \quad = \quad C \cdot Md(n_{1,u}, \cdots, n_{n(\mathcal{A}),u} \mid n_u, \alpha^*), \tag{16}$$

that is the value given by a Multinomial-Dirichlet distribution, where C is a constant that may be calculated (proof in Appendix).

By applying the Proposition, we finally obtain the conditional probability of the observed evidence as:

$$p(X_{V_0} \mid \mathcal{I}_j) = C \cdot \sum_{x_w \in \mathcal{X}_w} \prod_{k\in \overline{F}_o} Des(x_k \mid X_{pa(k)}, \mathcal{I}_j) \cdot \prod_{w\in \overline{F}_u} Des(X_w = x_w \mid X_{pa(w)}, \mathcal{I}_j) \cdot Md(n_{1,u}, \cdots, n_{n(\mathcal{A}),u} \mid n_u, \alpha^*). \tag{17}$$

3.3 Computational remarks

The calculation of equation (3) is simplified if all individuals belong to the same reference population. In this case many factors hidden into the C term of equation (17) are simplified in the WE without being numerically evaluated. If the simplification is unfeasible then a database of individual genotypes must be provided to calculate those constants, that is allele frequencies do not suffice.

Further computational savings are realized by detailing out only those alleles that are effectively involved in the forensic case under evaluation. This goal is achieved by exploiting the well known properties of the marginal distributions of the Dirichlet family of distributions [1, pp 135].

4 Case studies

In the literature, probability values on a pedigree have been calculated in several forensic debates involving nuclear DNA. We compared the results from three already published studies with the WE obtained following our approach.

In this section, the notation strictly follows that given in the above sections. Nevertheless, we simplify the notation when feasible, e.g. $\underline{n}(x_a)$ is the vector of allele frequencies observed in the individual labeled by a.

4.1 The paternity test

Let us consider the paternity identification problem as discussed in Brenner [2]. The list of alleles at the considered locus is r, p, q, s. Mother's (m), alleged father's (a) and child's (c) genotypes are the case evidences respectively equal to $\{r, p\}$, $\{p, q\}$ and $\{q, s\}$.

A sample of genotypes is also collected from the reference population and the allele frequencies in the reference population sample are, respectively $\widehat{\theta}_r, \widehat{\theta}_p, \widehat{\theta}_q, \widehat{\theta}_s$. These are point estimates that in the 'plug-in' approach substitute the unknown population parameters.

By casting the author's solution into our notation, after simplifications that depend on this specific case study, we have:

$$W_E = \frac{Des(x_c \mid x_m, x_a)}{\sum_{x_f \in \{(q,q),(q,\overline{q})\}} Des(x_c \mid x_m, x_t)) \cdot Gen(X_t \mid \underline{\widehat{\theta}})} = \frac{1}{2\widehat{\theta}_q}$$

We instead account for the uncertainty of population parameters by using case evidences as well as the reference population sample to infer about θ. To provide details, let V_0 be the set of nodes $V_0 = \{m, c, a\} \bigcup D$, where $D = (d_1, d_2, \dots)$ stands for the sample of size $n(D)$ taken from the reference population. The set $\mathcal{I}_0$ is defined by the collection of sentences:

$\mathcal{I}_0 = \{$ 'V_0 above is the set of nodes labeling genotypes',

'm label the mother and comes from the reference population',

'c is the child of m ',

'D is a set of labels for genotypes taken from the reference population',

' the alleged father a comes from the reference population ' $\}$

The first working hypothesis $\mathcal{I}_1$ adds to $\mathcal{I}_0$ the sentence 'The alleged father is the true father'. The graph $\mathcal{G}_1$, shown in Figure 1, is obtained using the working hypothesis $\mathcal{I}_1$.

The second working hypothesis $\mathcal{I}_2$ adds to $\mathcal{I}_0$ the sentence 'the alleged father is not the true father', therefore a new vertex t is introduced to mark the genotype of the true father taken from the reference population. The graph $\mathcal{G}_2$ shown in Figure 1 is defined using $\mathcal{I}_2$.

Conditionally to the hypothesis of paternity, all the genotypes in the graph are observed. Let $\underline{n}(x_a)$ be the vector of allele frequencies assessed on individual a, and $\underline{n}(D)$ be the vector of allele frequencies assessed on the reference population sample. Then, the marginal probability with respect to θ is:

$$p(X_{V_0}|\mathcal{I}_1) \quad = \quad Des(x_c|x_m,x_a) \cdot Md(x_m|2,\widetilde{\alpha}_{a,d}) \cdot Md(x_a|2,\widetilde{\alpha}_d) \cdot \prod_{i=1}^{n(D)} Md(x_{d_i}|2,\widetilde{\alpha}_{i-1}),$$

where $\widetilde{\alpha}_{a,d} = \alpha + \underline{n}(x_a) + \underline{n}(D)$, $\widetilde{\alpha}_d = \alpha + \underline{n}(D)$ and $\widetilde{\alpha}_i = \alpha + \sum_{j=1}^{i} \underline{n}(x_{d_j})$ (where $\widetilde{\alpha}_0 = \alpha$).

Similarly, the marginal probability for the non-paternity hypothesis is:

$$\begin{aligned} p(X_{V_0}|\mathcal{I}_2) \quad = \quad & \sum_{x_t \in \mathcal{X}_t} (Des(x_c|x_m, x_t) \cdot Md(X_t = x_t \mid 2, \widetilde{\alpha}_{a,m,d}) \\ & \cdot Md(x_m \mid 2, \widetilde{\alpha}_{a,d}) \cdot Md(x_a \mid 2, \widetilde{\alpha}_d) \cdot \prod_{j=1}^{n(D)} Md(x_{d_i} \mid 2, \widetilde{\alpha}_{i-1}), \end{aligned}$$

where $\widetilde{\alpha}_{a,m,d} = \alpha + \underline{n}(x_a) + \underline{n}(x_m) + \underline{n}(D)$.

The WE is therefore

$$W_E = \frac{Des(x_c|x_m,x_a)}{\sum_{x_t \in \mathcal{X}_t} \cdot Des(x_c|x_m,x_t) \cdot Md(X_t = x_t \mid 2, \widetilde{\alpha}_{a,m,D})}.$$

Several (hypothetical) samples are here considered to assess the sensitivity of W_E to the reference population sample size: all of them provide the same point estimate of $\widehat{\theta}_q = 0.01$, but the size is in the range 50 to 1000.

Results are shown in Figure 2. Values of WE span from 25 to nearly 50, therefore our coherent analysis always gives a more conservative evaluation of WE.

4.2 The missing father case study

Let's consider now a less usual paternity case as in Jourqueira [8]. Here we have two children (c_1, c_2) of the same mother (m). The father of the first child is well identified (f_1) but missing. The debate is about the possibility that he is also the father of the second child $(\mathcal{I}_1)$. If children have two different fathers then another (unknown) man (f_2) is introduced in the graph $(\mathcal{I}_2)$. Additional data are the parents (p_1, p_2) of f_1 and a sample from the reference population labeled as $D = (d_1, d_2, \cdots)$. In Figure 3, the graphs obtained under the two hypotheses are shown. The original data from [8] are summarized in Table 1.

First, we detail the solution obtained using our approach. If $\mathcal{I}_1$ holds we have:

$$\begin{aligned} p(X_{V_0}|\mathcal{I}_1) \quad = \quad & \sum_{x_{f_1} \in \mathcal{X}_{f_1}} Des(x_{c_1}|x_m, X_{f_1}) \cdot Des(x_{c_2}|x_m, X_{f_1}) \cdot Des(X_{f_1}|x_{p_1}, x_{p_2}) \cdot \\ & Md(x_{p_1} \mid 2, \widetilde{\alpha}_{p_2,D}) \cdot Md(x_m \mid 2, \widetilde{\alpha}_{a,p_1,p_2,D}) \cdot Md(x_{p_2} \mid 2, \widetilde{\alpha}_D) \\ & \cdot \prod_{i=1}^{n(D)} Md(x_{d_i} \mid 2, \widetilde{\alpha}_{i-1}), \end{aligned}$$

where symbols have been already introduced.

Similar calculation may be performed by conditioning on the two-fathers hypothesis ($\mathcal{I}_2$), thus:

$$\begin{aligned}
p(X_{V_0}|\mathcal{I}_2) \quad = \quad & \sum_{\mathcal{X}_{f_1}} Des(x_{c_1}|x_{m,}X_{f_1}) \cdot Des(X_{f_1}|x_{p_1,}x_{p_2}) \\
& \cdot \sum_{\mathcal{X}_{f_2}} Des(x_{c_2}|x_{m,}X_{f_2}) \cdot Md(X_{f_2} \mid 2, \widetilde{\alpha}_{m,p_1,p_2,D}) \\
& \cdot Md(x_m \mid 2, \widetilde{\alpha}_{p_1,p_2,D}) \cdot Md(x_{p_1} \mid 2, \widetilde{\alpha}_{p_2,D}) \cdot Md(x_{p_2} \mid 2, \widetilde{\alpha}_{D}) \\
& \cdot \prod_{j=1}^{n(D)} Md(x_{d_i} \mid 2, \widetilde{\alpha}_{i-1}).
\end{aligned}$$

The WE, after simplification, becomes:

$$W_E = \frac{\sum_{x_{f_1} \in \mathcal{X}_{f_1}} Des(x_{c_1} \mid x_m, x_{f_1}) \cdot Des(x_{c_2} \mid x_m, x_{f_1}) \cdot Des(X_{f_1} = x_{f_1} \mid x_{p_1,}x_{p_2})}{\begin{array}{r} \sum_{x_{f_1} \in \mathcal{X}_{f_1}} Des(x_{c_1} \mid x_m, x_{f_1}) \cdot Des(X_{f_1} = x_{f_1} \mid x_{p_1}, x_{p_2}) \cdot \\ \cdot \displaystyle\sum_{x_{f_2} \in \mathcal{X}_{f_2}} Des(x_{c_2} \mid x_m, x_{f_2}) \cdot Md(X_{f_2} = x_{f_2} \mid 2, \widetilde{\alpha}_{m,p_1,p_2,D}). \end{array}}$$

Jourqueira's solution differs from our equation because the equation term $Md(X_{f_2} = x_{f_2} \mid 2, \widetilde{\alpha}_{m,p_1,p_2,D})$ is substituted by $Gen(X_{f_2} = x_{f_2} \mid \widehat{\underline{\theta}})$, i.e. the plug-in method is invoked.

In Figure 4, the numerical consequences on the WE are evaluated at four loci (it seems that original Jorqueira's calculations are based on the allele frequency 0.021 instead of 0.012 for allele 29 at locus D12S1090). The change of overall WE is appreciable and it depends on the sample size. If the the sample size is 100 then the author's WE is reduced from 234.38 to 150.21. The Figure highlights numerical differences between Egeland-Mostad's results and our findings, especially if a rare allele is found in the alleged person (e.g. on top right, the locus vd12S1090 versus others).

4.3 A missing person case study

Evett and Weir [5] have illustrated a missing person problem in which a corpse, labelled by x, is found. The genotypes of the mother (m), of four siblings ($s_1, \cdots s_4$), of the spouse (sp) of the missing person and their child (c) are assessed in a family. A sample $D = (d_1, d_2, \cdots)$ is taken from the reference population.

Once more, we start showing our solution. In Figure 5 (left), $\mathcal{G}_1$ summarizes the conditional independence structure obtained if the corpse belongs to the missing person. In Figure 5 (right), the graph $\mathcal{G}_2$ is defined by assuming that the corpse comes from a generic individual tf of the reference population.

The probability of observed evidences given $\mathcal{I}_1$ is:

$$\begin{aligned} p(X_{V_0}|\mathcal{I}_1) \quad &= \quad Des(x_c \mid x_x, x_{sp}) \cdot \sum_{x_f \in \mathcal{X}_f} \cdot Des(x_x \mid x_m, x_f) \prod_{i=1}^{4} \cdot Des(x_{s_i} \mid x_m, x_f) \\ & \quad \cdot Md(X_f = x_f \mid 2, \widetilde{\alpha}_{m,sp,D}) \cdot Md(x_m \mid 2, \widetilde{\alpha}_{sp,D}) \cdot Md(x_{sp} \mid 2, \widetilde{\alpha}_D) \\ & \quad \prod_{i=1}^{n(D)} Md(x_{d_i} \mid 2, \widetilde{\alpha}_{i-1}). \end{aligned}$$

Given the alternative hypothesis $\mathcal{I}_2$, the marginal probability is:

$$\begin{aligned} p(X_{V_0} \mid \mathcal{I}_2) \quad &= \quad \sum_{x_{tf} \in \mathcal{X}_{tf}} \sum_{x_f \in \mathcal{X}_f} Des(x_c \mid x_{sp}, x_{tf}) \cdot \prod_{i=1}^{4} Des(x_{s_i} \mid x_m, x_f) \\ & \quad \cdot Des(X_{tf} = x_{tf} \mid x_m, x_f) \cdot Md(X_f = x_f \mid 2, \widetilde{\alpha}_{x,m,sp,D}) \\ & \quad \cdot Md(x_x \mid 2, \widetilde{\alpha}_{m,sp,D}) \cdot Md(x_m \mid 2, \widetilde{\alpha}_{sp,D}) \cdot Md(x_{sp} \mid 2, \widetilde{\alpha}_D) \\ & \quad \cdot \prod_{i=1}^{n(D)} Md(x_{d_i} \mid 2, \widetilde{\alpha}_{i-1}) \end{aligned}$$

The WE, after simplifications, becomes:

$$\frac{Des(x_c \mid x_x, x_{sp}) \cdot \sum_{x_f \in \mathcal{X}_f} Des(x_x|x_m, x_f) \cdot \prod_{i=1}^{4} Des(x_{s_i} \mid x_m, x_f) \cdot Md(X_f = x_f \mid 2, \widetilde{\alpha}_{m,sp,D})}{\begin{array}{r} \sum_{x_{tf} \in \mathcal{X}_{tf}, x_f \in \mathcal{X}_f} Des(x_c \mid x_{sp}, x_{tf}) \cdot \prod_{i=1}^{4} Des(x_{s_i} \mid x_m, x_f) \cdot Des(X_{tf} = x_{tf} \mid x_m, x_f) \cdot \\ \cdot Md(X_f = x_f \mid 2, \widetilde{\alpha}_{x,m,sp,D}) \cdot Md(x_x \mid 2, \widetilde{\alpha}_{m,sp,D}) \end{array}}$$

Evett and Weir follow the symbolic approach and calculate the WE through the appealing expression $W_E = \frac{2}{2\widehat{\theta}_3 + \widehat{\theta}_4}$, where $\widehat{\theta}_i$ is the relative frequency of allele i.

Unfortunately, under the informative set $\mathcal{I}_2$, they did not consider the stochastic vertex tf thus their W_E is flawed. Fixing the cited flaw, we obtained $W_E = \frac{2}{\widehat{\theta}_3 \cdot (2\widehat{\theta}_3 + \widehat{\theta}_4)}$, and the corrected expression numerically matches the numerical solution given by Mostad and Egeland [11]. Nevertheless, the uncertainty about the population parameters is not accounted.

The sensitivity of the WE to the sample size for fixed allele relative frequencies is shown in Figure (6). In Table 2 the artificial reference population sample used to build the graph is listed.

5 Discussion

The use of graphs in which a vertex represents population parameters characterizes our evaluation of the probability of observing the collected DNA evidences given two or more forensic hypotheses.

In our approach, the set of genotypes that shall be used to make inference on population parameters is made explicit. We called these contributing genotypes 'founders', whether they belong to the reference population sample or they pertain case evidences.

The coherent treatment of unknowns entails appreciable changes in the evaluation of WE, as we described in the re-analysis of case studies from the literature. A common feature of all the case studies is that greater differences occur if rare alleles in the population are found in the alleged person(s). The explanation follows from the inclusion of the alleged person's relatives (or some of them) among the founders, so that inference on population parameters is strongly affected by the augmentation of the reference population sample, making rare alleles 'less rare'.

The results of our analysis are always conservative, that is favorable to the non-identification. This is an important feature, especially if the reference population sample is in the order of few hundreds, as it very often happens in reference populations that are not widely typed (e.g. Eastern Europe).

Our approach asymptotically gives the same value of WE obtained using the "plug-in" with an infinite sample size, that is, if the sample size diverges then the contribution of the case-related founders becomes negligible.

Closed-form expressions has been obtained but this achievement depends on the presence of Hardy-Weinberg equilibrium in the reference population. The generalization of the approach to reference populations far from equilibrium could require a more complex model addressing features like the structure of subpopulations. Nevertheless, the conditional independence structure expressed by a DAG still holds and it might be exploited by the BUGS simulation software [13] that accepts a graph as description of a probabilistic model.

The calculation of the WE by simulation should not be particularly intensive following our specifications. The marginalization over unobserved nodes does not entail additional computer simulations after the updating of population parameters. This property follows from the rules we gave to build the graph.

References

[1] J. Bernardo and A. F. M. Smith. *Bayesian Theory.* Wiley, New York, 1994.

[2] C. H. Brenner. Symbolic kinship program. *Genetics*, 145:535–542, 1997.

[3] A. P. Dawid and J. Mortera. Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society ser. B*, 58:425–443, 1996.

[4] E. Essen-Möller. Die beweiskraft der hnlichkeit im vatershaftsnachweis teoretische graundlagen. *Mitteilung Antropologishe Graundlagen*, 68:9–53, 1938.

[5] I. W. Evett and B. S. Weir. *Interpreting DNA evidence.* Sinauer, Sunderland, 1998.

[6] M. Frydenberg. The chain graph markov property. *Scandinavian Journal of Statistic*, 17:333–353, 1990.

[7] P. Ihm and K. Hummel. A method to calculate the plausability of paternity using blood groups results of any relatives. *Zeitschrift fur immunitatsforschung experimentelle und klinische immunologie*, 149:405–416, 1975.

[8] H. Jorqueira, L. Cifuentes, F. Moreno, and E. Aguirre. Paternity analysis when the putative father is missing: first case in chile. *Journal of Forensic Science*, 44:627–629, 1999.

[9] K. Lange. *Mathematical and statistical methods for genetic analysis.* Springer-Verlag, New York, 1997.

[10] S. Lauritzen. *Graphical models.* Oxford University Press, Oxford, 1996.

[11] P. E. Mostad and T. Egeland. Probability assessment of family relations using the program "pater". Technical report, Norvegian Computing Center, 1998.

[12] J. Ott. *Analysis of Human Linkage 3rd edition.* Johns Hopkins University Press Baltimore, 1999.

[13] D. J. Spiegelhalter, A. Thomas, and N. G. Best. Winbugs version 1.2 user manual. Technical report, MRC Biostatistics Unit, 1999.

[14] N. Wermuth and S. L. Lauritzen. On substantive research hypotheses, conditional

independence graphs and graphical chain models. *Journal of the Royal Statistical Society series B*, 52:21–50, 1990.

APPENDIX: Proof of the proposition.

Let's first consider the joint probability of observing individuals in F_u, using the multinomial representation:

$$\begin{aligned}\prod_{h\in F_u} Gen(X_h = x_h \mid \theta, \mathcal{I}_j) = \quad &= \textstyle\prod_{h\in F_u}\left[\left(\frac{n_h!}{\prod_{i=1}^{n(\mathcal{A})} n_{i,h}!}\right)\cdot \prod_{i=1}^{n(\mathcal{A})} \theta_i^{n_{i,h}}\right]\\ &= \textstyle\prod_{h\in F_u}\left(\frac{n_h!}{\prod_{i=1}^{n(\mathcal{A})} n_{i,h}!}\right)\cdot \prod_{i=1}^{n(\mathcal{A})} \theta_i^{n_{i,u}},\end{aligned}$$

where $n_{i,u} = \sum_{h\in F_u} n_{i,h}$, $n_{i,h} \in \{0,1,2\}$ and $n_h = \sum_{i=1}^{n(\mathcal{A})} n_{i,h} = 2$ for each $h \in F_u$. Conditionally on θ, the above probability value is proportional to the probability of observing n_u alleles without regards for the assortment of alleles into genotypes.

By exploiting the conjugate structure of the Dirichlet model for θ, the integration proceeds as:

$$\int \prod_{h\in F_u} Gen(X_h = x_h \mid \theta, \mathcal{I}_j) \cdot \prod_{z\in F_o} Gen(x_z \mid \theta, \mathcal{I}_j) \cdot p(\theta \mid \mathcal{I}_j) \cdot d\theta =$$

$$= \int \prod_{h\in F_u}\left(\frac{n_h!}{\prod_{i=1}^{n(\mathcal{A})} n_{i,h}!}\right) \cdot \prod_{i=1}^{n(\mathcal{A})} \theta_i^{n_{i,u}} \cdot Dirichlet(\theta \mid \alpha^*) \cdot \prod_{z\in F_o} Md(x_z \mid 2, \widetilde{\alpha}_z) \cdot d\theta =$$

$$= \int \frac{\prod_{h\in F_u} \frac{n_h!}{\prod_{i=1}^{n(\mathcal{A})} n_{i,h}!}}{\frac{n_u!}{\prod_{i=1}^{n(\mathcal{A})} n_{i,u}!}} \cdot \frac{n_u!}{\prod_{i=1}^{n(\mathcal{A})} n_{i,u}!} \cdot \prod_{i=1}^{n(\mathcal{A})} \theta_i^{n_{i,u}} \cdot Dirichlet(\theta \mid \alpha^*) \cdot \prod_{z\in F_o} Md(x_z \mid 2, \widetilde{\alpha}_z) \cdot d\theta =$$

$$= C \cdot Md(n_{1,u}, \cdots, n_{n(\mathcal{A}),u} \mid n_u, \alpha^*),$$

where C is given by:

$$C = \frac{\prod_{h\in F_u} \frac{n_h!}{\prod_{i=1}^{n(\mathcal{A})} n_{i,h}!}}{\frac{n_u!}{\prod_{n_{1,u}}^{n(\mathcal{A})} n_{i,u}!}} \prod_{z\in F_o} Md(x_z \mid 2, \widetilde{\alpha}_z) \tag{18}$$

Table 1: Allele relative frequencies and evidences for Jorqueira case study (section 4.2).

	Case Evidences				
Loci	p_1	p_2	m	c_1	c_2
D1S80	$(31, 33)$	$(24, 31)$	$(22, 31)$	$(22, 33)$	$(31, 31)$
D12S1090	$(20, 22)$	$(22, 29)$	$(12, 25)$	$(12, 29)$	$(12, 29)$
D3S1744	$(18, 21)$	$(18, 21)$	$(18, 21)$	$(18, 21)$	$(18, 18)$
D18S849	$(16, 16)$	$(16, 16)$	$(16, 17)$	$(16, 16)$	$(16, 16)$

Relevant allele frequencies (locus, allele)			
D1S80, 31	D12S1090, 29	D3S1744, 18	D18S849, 16
0.110	0.021	0.341	0.339

Table 2: Reference population sample and evidences for Evett and Weir case study (section 4.3).

Alleles	1	2	3	4	5	6	7	8
$\hat{\theta}_3$	.004	.008	.0444	.240	.32	.32	.052	.012
n	1	2	11	60	80	80	13	3

Genotypes	m	s_1	s_2	s_3	s_4	x	sp	c
	$(3,4)$	$(2,4)$	$(2,4)$	$(2,4)$	$(3,4)$	$(3,3)$	$(5,6)$	$(3,5)$

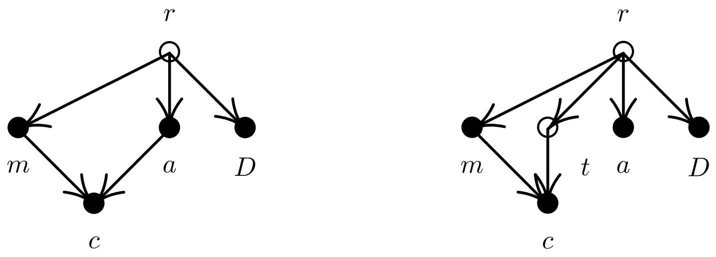

Figure 1: Brenner case study. Graph $\mathcal{G}_1$ (left) refers to the paternity hypothesis, while graph $\mathcal{G}_2$ deals with not paternity (section 4.1).

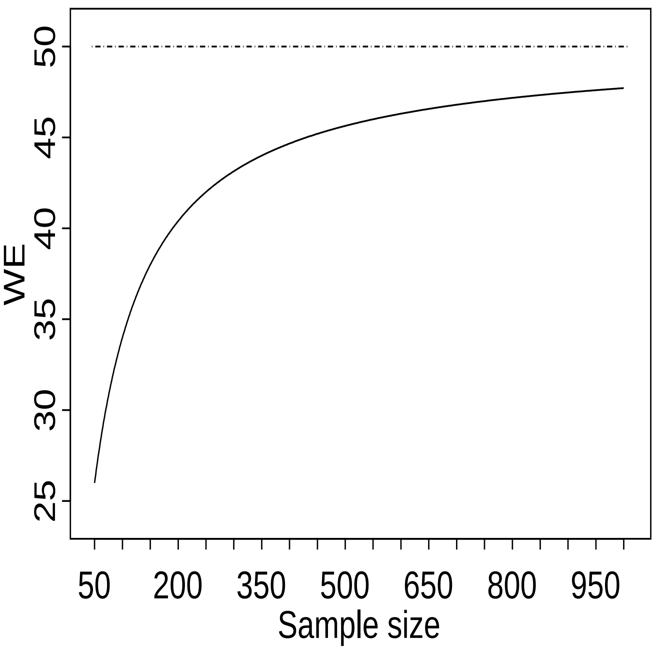

Figure 2: Brenner case study. Solid lines represent the weight of evidences for different sample sizes (full Bayesian approch). Dotted lines represent the weight of evidences obtained by the plug-in approach (section 4.1).

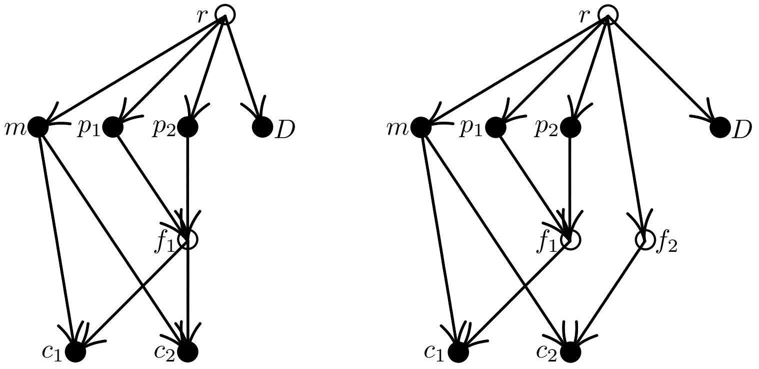

Figure 3 : Jourqueira case study. Graph $\mathcal{G}_1$ (left) represents the "one missing father" hypothesis, while $\mathcal{G}_2$ (right) refers to the "two missing fathers" hypothesis (section 4.2).

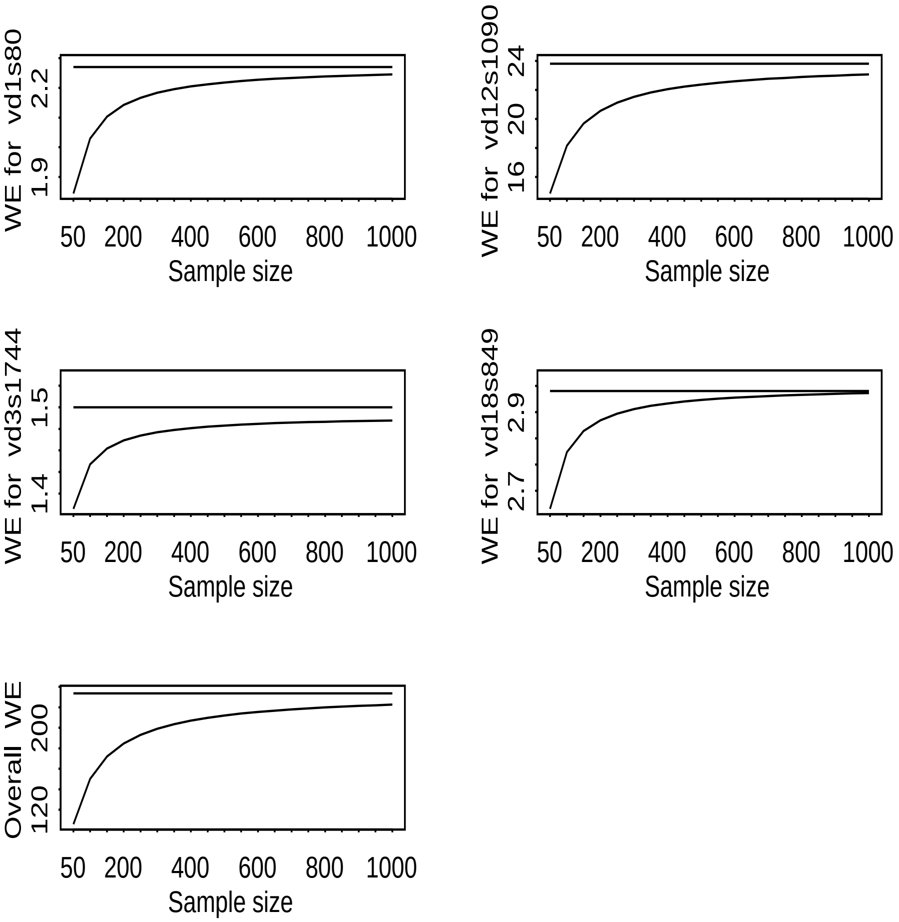

Figure 4: Jourqueira case study. Solid lines represent the weight of evidences (four loci) for different sample sizes (full Bayesian approch). Dotted lines represent the weight of evidences obtained by the plug-in approach (section 4.2).

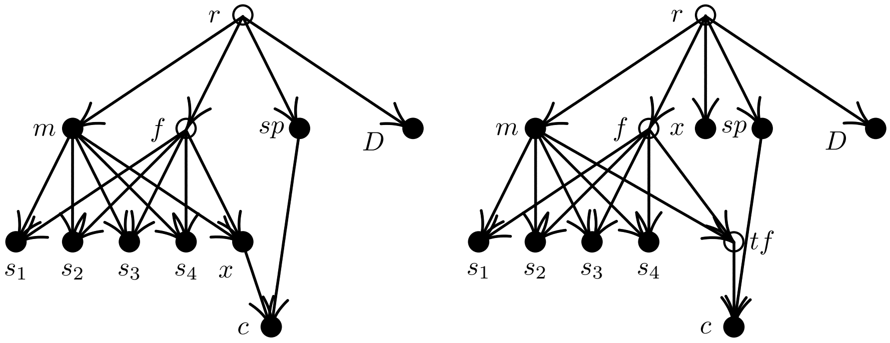

Figure 5: Evett and Weir case study. Graph $\mathcal{G}_1$ (left) refers to the missing person identification hypothesis, while $\mathcal{G}_2$ (right) deals with the no identification hypothesis (section 4.3).

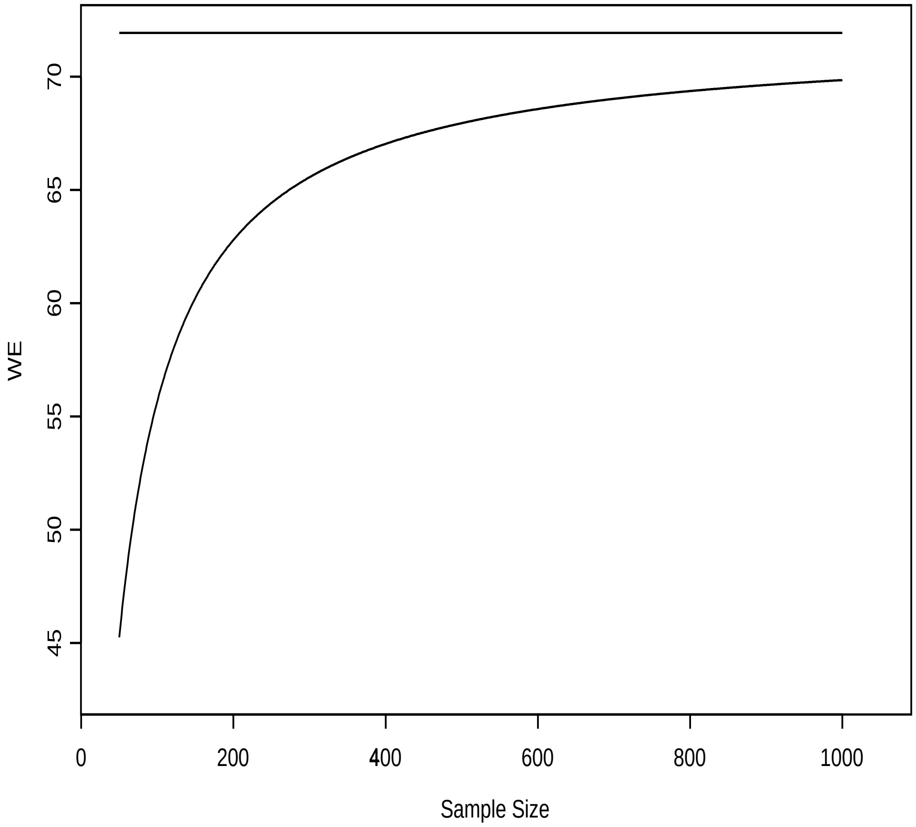

Figure 6: Evett and Weir case study. Solid lines refers to the weight of evidences for different sample sizes (full Bayesian approch). Dotted lines deals with the weight of evidences obtained by the plug-in approach (section 4.3).