Norwegian University
of Life Sciences

Master's Thesis 2023 60 ECTS
Chemistry, Biotechnology and Food Science

Combining genetic and non-genetic information to solve forensic identification problems

Alf Erik Borgundvåg Berg
Bioinformatics & Applied Statistics

Combining genetic and non-genetic information to solve forensic identification problems

Alf Erik Borgundvåg Berg

Preface

This is my master thesis in biostatistics at the Norwegian University of Life Sciences supervised by Professor Thore Egeland. In this thesis I wish to explore the possibility of combining different kinds of forensic data to solve forensic identification cases. Forensic cases often rely solely on DNA analysis to identify, I wish to consider other options. I am grateful for the guidance I have received on this 60 credit master thesis.

Abstract

Forensic identification is the process of identifying for judicial purposes using scientific methods. Such techniques may be applied to humans, animals or objects. In this thesis the focus is on identifying humans. Forensic identification problems involving humans range from standard paternity cases to complex identification problems involving a large number of victims. In a murder case there may be trace evidence at the crime scene which helps identify the perpetrator. In a paternity case DNA analysis can determine whether or not a man fathered a child.

In identification cases investigators will make multiple hypotheses in the form of pedigrees. Out of these at most one can be true. The most likely hypothesis may be found through statistical analysis. Several forensic methods exist and may be applied for identification. When multiple data sources are available, it would be favourable for researchers to be able to combine the results based on all available data. The goal of this thesis is to provide a framework for solving identification problems using multiple forms of data for the same hypothesis tests. Specifically, DNA data will be paired with other non-genetic data in the data analysis.

Combining DNA with other data from the forensic case is desired because it allows researchers more material to draw conclusions from. This is particularly useful in cases where no conclusion may be drawn from DNA analysis alone. One such case involves two full siblings of the same sex who have gone missing, where neither sibling has descendants. If DNA is found from one of them, it is impossible to determine which of the siblings it belongs to even if DNA data from their family is accessible. Another way other data may assist in a forensic case is that information like age may be used to limit the hypotheses space, thus simplifying the forensic case. This emphasizes why multiple types of forensic data should be used in forensic analysis. The concept of combining different kinds of information for identification is the core of this thesis.

Sammendrag

Forensisk identifisering er prosesser der en identifiserer ved hjelp av vitenskapelige metoder. Slike teknikker kan bli anvendt på mennesker, dyr eller objekter. I denne avhandlingen er fokuset på å identifisere mennesker, og formålene varierer fra farskapssaker til komplekse problemer med et stort antall offere. I en mordsak kan det finnes bevis på åstedet som kan benyttes til å identifisere gjerningsmannen. I en farskapssak kan DNA-analyse avsløre om en mann er far til et barn.

I identifikasjonssaker vil etterforskere lage flere hypoteser i form av pedigreer. Bare en av disse kan være sann. Den mest sannsynlige hypotesen kan finnes gjennom statistisk analyse. Flere forensiske metoder finnes og kan benyttes for identifikasjon. Når flere datakilder er tilgjengelige, vil det være gunstig for forskere å kunne kombinere resultatene basert på all tilgjengelig data. Målet med denne avhandlingen er å gi et utgangspunkt for å løse identifikasjonsproblemer der flere former av forensisk data kombineres. DNA-data vil pares med ikke-genetiske data i analysen.

Kombinasjon av DNA-data med andre data er ønsket fordi det lar forskere bruke mer materiale til å trekke konklusjoner. Dette er spesielt gunstig i tilfeller der ingen konklusjon kan trekkes fra DNA-analyse alene. Et eksempel på en slik sak involverer to helsøsken av samme kjønn uten etterkommere der begge søsknene er savnet. Om DNA er funnet av den ene, er det umulig å bestemme hvilken av søsknene det kommer fra selv om DNA fra familiemedlemmer ville være tilgjengelig. En annen måte andre data kan være til nytte i en forensisk undersøkelse er ved at informasjon som alder kan brukes til å begrense hypoteserommet, og dermed forenkle undersøkelsen. Dette fremhever hvorfor flere typer data burde brukes i forensisk analyse. Konseptet med å kombinere forskjellige typer informasjon for identifikasjon er kjernen i denne avhandlingen.

Contents

1 Introduction

1.1 Background for the thesis

Within forensic identification, the aim is to find the truth behind forensic cases, with what little information one may have. Traditionally, forensic identification involving genetics has been divided into criminal cases and kinship cases. Cases of the latter form are addressed by this thesis. The simplest kinship cases involve determining paternity, more complex cases involve discovering relationships between distant family members. Such a relationship case may be deciding whether two people are first cousins through forensic genetics. Distant relatives will have less DNA in common than close relatives, and kinship will be harder to prove for distant relatives than for close ones with forensic evidence. The type of kinship case this thesis focuses on, disaster victim identification, may be viewed as multiple kinship cases treated as one.

A *disaster victim identification* (DVI) case is shown in Figure 1. The victims are plotted on the left, the pedigree containing missing people on the right. The victims are referred to as $V1$ and $V2$, the missing people are referred to as $M1$ and $M2$. The sex of the people in the plots is indicated with shapes. A circle means the person is female, a square means the person is male. The plots tell that $M1$, $M2$, $V1$ and $V2$ are all males. The norm when it comes to ordering siblings is to order them from left to right, oldest to youngest. Thus the plot implies that $M1$ is older than $M2$.

Some of the people in the pedigree plots in Figure 1 have been assigned numbers which show their reported genotype. The genotype is the alleles an individual has for a gene. These numbers 1/2, 1/2 and 2/3 reference the alleles the victims and relative have been shown to have in their genome through genotyping.

Table 1 is output from a computation-backed investigation of a simulated DVI case. Each row contains the description and evaluation of one DVI hypothesis. The five columns have the names *"V1"*, *"V2"*, *"loglik"*, *"LR"* and

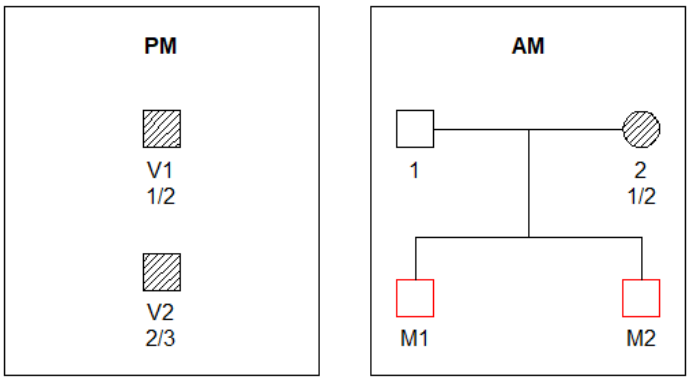


Figure 1: A simple DVI case. There are two genotyped male victims, $V1$ and $V2$, and two missing brothers, $M1$ and $M2$. The mother of the missing brothers is the genotyped reference.

"posterior". The two leftmost columns correspond to the victims, $V1$ and $V2$. These columns contain the assigned missing people. In the first mentioned hypothesis in the upper row, $V1$ was assigned $M1$ while $V2$ was assigned $M2$. This means that the hypothesis claims $V1$ and $M1$ are the same and that $V2$ and $M2$ are the same. The stars (*) visible in some of the rows are empty spaces. An empty space simply means no known missing person was assigned to that victim, and that the victim thus is someone who is not among the reported missing people. For instance, no known missing person was assigned to a victim in the bottom hypothesis, hypothesis #7. This means hypothesis #7 should be interpreted as $V1$ and $V2$ being neither of the two reported missing people, but rather two other missing people not yet considered in the disaster victim investigation.

The three rightmost columns in Table 1 contain the calculated log-likelihood, the likelihood ratio and the posterior probability for each hypothesis. These columns are named "loglik", "LR" and "posterior", respectively. They reflect on the likelihoods of the various hypotheses considering the forensic data available. These columns will be explained later on in Section 3.3. Though, for now,

	V1	V2	loglik	LR	posterior
1	M1	M2	-9.2876	321.9697	0.4313
2	M2	M1	-9.2876	321.9697	0.4313
3	M1	*	-11.8133	25.7576	0.0345
4	M2	*	-11.8133	25.7576	0.0345
5	*	M1	-11.8432	25.0000	0.0335
6	*	M2	-11.8432	25.0000	0.0335
7	*	*	-15.0621	1.0000	0.0013

Table 1: Summary of motivational example discussed in Section 1.1 and in greater detail in Section 3.3. The only data used here was a single genetic marker with alleles denoted 1, 2, 3, 4 with allele frequencies $p_1 = 0.33$, $p_2 = 0.01$, $p_3 = 0.33$, $p_4 = 0.33$. See also Figure 1.

consider the posterior probabilities. The higher the posterior probability, the more likely the hypothesis is to be true. Note that the two top hypotheses in Table 1 have the same posterior probability, and thus one may not be considered more plausible than the other. An explanation of this phenomenon is in Section 2.2.3.

1.2 A brief review of the literature

The book 'Mass Identifications: Statistical Methods in Forensic Genetics' [1] provides the required background for this thesis. The genetical and forensic background are explained in addition to case studies based on realistic cases. The paper 'Joint DNA-based disaster victim identification' [2] provides details on recent methods and also presents the R library dvir that will be used for examples in this thesis. The books [3, 4] give further information on statistical methods and software, particularly on forensic genetics in R. The report 'Making sense of forensic genetics' [5] is intended for the general audience and introduces forensic genetics in a simpler way. However, models for non-genetic evidence and how to combine genetic (DNA) data and non-genetic data, are not covered

in the previously mentioned references. Some models that include non-genetic evidence are contained in the article 'Incorporating non-genetic evidence in large scale missing person searches: A general approach beyond filtering' [6]. There are, however, some problems with these methods and we will investigate some alternatives.

1.3 Aims of the thesis

DNA analysis has historically been invaluable for forensic research, with its history spanning decades. However, data other than DNA data may provide stronger proof in favour of the true hypothesis in cases where such data is available. Suppose that the collected DNA data supports a hypothesis. Then, if the collected non-genetic forensic data also supports this hypothesis, combining the two types of forensic evidence will create stronger conviction in favour of this hypothesis. On the other hand, if DNA data supports a hypothesis which does not fit the other forensic data, this points in the direction that one might have made errors in the investigation.

Methods designed to solve disaster victim identification problems as the one described in Sections 1.1 above, are reviewed. Combining DNA data with other forensic data may result in stronger likelihoods for hypotheses. Because situations exist where DNA readings fail to identify, like the symmetric case presented above, more complex methods are highly desirable. The goal of this thesis is to provide a framework for solving identification problems using multiple forms of data for the same hypothesis tests. Specifically, DNA data will be paired with other, non-genetic data, in the analyses.

A conventional approach to DVI problems is to first list all possible solutions (assignments). However, this may not be feasible for large problems. Therefore we have explored how sex and age information can be used to restrict the number of assignments.

1.4 Organisation of the thesis

This thesis is constructed with focus on explanation of material (data) and methods first, results and discussion thereafter. The section on material and methods (see Section 2) starts with a brief review of DNA data and forensic statistics. The remaining parts of Section 2, starting with Section 2.3 present methods not well covered in existing literature. The theory is exemplified in Section 3 and discussed in 4. The appendices A provide some further details on implementation.

2 Material and methods

2.1 DNA and DNA analysis

DNA is found, with some exceptions, in all the cells in a body. Thus, at any disaster scene or crime scene forensic researchers have a high chance of finding small amounts of DNA, even in cases of severe body degradation. DNA can be found in all sorts of body remains, including trace amounts of spit, blood, hair, semen or dead skin cells. This, combined with the fact that no two humans have identical DNA, makes DNA analysis an invaluable tool in forensic identification.

All humans share approximately 99.9% of their DNA with other humans [5]. Despite this, due to many small differences, everyone has a unique genome which differs from that of everyone else. Even monozygotic twins are expected to have some very small differences in their respective genomes when they are compared [7]. Thus, DNA analysis can, in theory, identify anyone. The segments of the human genome with the most variability may be sampled from a person to create a *DNA profile* for them.

Authorities may store the genetic material of citizens in DNA databases. The DNA which is stored can be that of convicted criminals or from plausible suspects in crime cases. Usually authorities do not have DNA from every citizen on record, as there are ethical dilemmas with storing the genetic material in terms of privacy.

DNA analysis for forensic sciences produces powerful, but not infallible evidence. Forensic investigators must recognize that there are situations where DNA analysis produces misleading results, and hence DNA should never cause a verdict in a criminal investigation as the only piece of evidence. There are a few ways in which gene reading can go wrong and result in misleading conclusions. The DNA found at a crime scene does not need to be that of someone involved in that crime. One way this may happen is that humans spit when they talk, and traces of spit from someone not present during a crime may still be found at the crime scene. Sometimes investigators do not even know how the

DNA showed up in the investigation, as by itself a DNA profile cannot provide information about the body fluid it came from. Further, forensic investigators are prone to making human error, and mistakes done by researchers during the analysis may produce misleading results. One possible error is that researchers accidentally mix samples from different forensic cases [5].

A consequence of the improvements in DNA reading techniques over the last few decades is how increasingly smaller amounts of cells are sufficient to accurately read the DNA of someone. Because low amounts of organic material are required, sometimes investigators detect DNA which they cannot determine the source of. While they may still want to know which individual the DNA belongs to, drawing conclusions on if and how this individual was involved in the case will be difficult, if not impossible, without context.

2.1.1 Genetic markers and genome locations

When we use DNA information to connect disaster victims to missing people, we want to compare the base pairs at specific locations in their genomes to each other. Not all the DNA of a victim or a suspect is important for analytic purposes. Most interesting for analysts is the 0.1% of the human genome which differs the most between people. These 0.1% may be used to create a DNA profile for a human. If DNA profiles from two samples are mostly or entirely identical, investigators may theorize that the two samples come from the same person.

It is possible to use less genetic information than a DNA profile contains to identify humans. A possible narrower data set than a complete DNA profile is a set of genetic markers. A *genetic marker* is a DNA sequence on a specific location in a genome. Genetic markers differ from DNA profiles, not only because the length of the DNA data for markers is much shorter, but also because in a marker the data is limited to a single, continuous segment of the genome. DNA profiles contain multiple segments from multiple chromosomes. Reading a single genetic marker gives little information, therefore reading multiple genetic

markers is recommended for statistical analysis if possible.

Investigators need to decide which genetic markers should be read in a DNA analysis. In disaster victim identification the selected markers are read for both disaster victims and relatives of the missing people, and the results are compared. In a crime investigation the pertinent DNA segments are read from both samples collected from suspects and from DNA strains acquired at the crime scene. DNA strains may be small or damaged and therefore difficult to read, thus investigators may choose which genetic markers to use for the analysis based on which segments they are able to read. In a paternity case the DNA of both the father and the alleged child is read.

For a genetic marker to be useful in the context of forensic analysis, several variations of the gene it tracks should exist within a population. Naturally occurring variations in the DNA sequence of a gene are called *alleles*. Often each allele of a gene is given a unique number or letter. If two alleles of a gene are known to exist within a population, these may be numbered as allele 1 and allele 2.

As a demonstration, an example *autosomal SNP* marker in the human genome has only two alleles, 1 and 2. A SNP (single nucleotide polymorphism) is the genetic variation at a single DNA base pair within a population. An autosomal SNP is an SNP located on a chromosome which is not a sex chromosome.

Any individual will have two copies of an autosomal SNP in their genome, except in cases of polysomy where there are more than two instances of non-sex chromosomes. For the example marker, a human can either have two alleles of type 1 (genotype 1/1), two alleles of type 2 (genotype 2/2) or one allele of each type (genotype 1/2). This is because, with some exceptions, any chosen individual will have two versions of any genetic marker in their genome, one version being maternal and the other being paternal. Exceptions include genetic markers present on sex chromosomes and markers in the mitochondrial DNA. Only autosomal markers will be used in the examples in this thesis.

Out of all versions of this SNP marker within all people in a population, a percentage is of allele 1 and another percentage is of allele 2. These percentages,

or frequencies, will be referred to as p_1 and p_2. Then the probability that a randomly selected chromosome with this markers happens to have allele 1 is p_1. Equivalently, the probability for allele 2 is p_2. Because type 1 and 2 are the only alleles for this gene, p_2 must be $1 - p_1$ as the allele frequencies should sum to 1. Formulas for the probabilities of specific genotypes may be derived from the allele frequencies. This will be done in the section covering the Hardy-Weinberg equilibrium (HWE), Section 2.1.2.

For the genotype to be 2/2 for a person, both chromosomes the person has which contain the genetic marker must have the second variation, allele 2. For this combination to occur, the individual must have inherited this variation both from their mother and from their father. The probability of this happening is p_2^2 assuming HWE. Generally speaking, the probability of an individual having a specific allele i on both strands of the chromosome is p_i^2 if HWE is true. In such cases the individual is homozygote for the marker. It follows that the probability for genotype 1/1 in an individual is p_1^2.

The probability of an individual being heterozygote with alleles 1/2 is $2p_1p_2$, not p_1p_2. This is because the probability of 1 coming from the mother and 2 coming from the father is p_1p_2, while the probability of 1 coming from the father and 2 coming from the mother is also p_1p_2. Both possibilities must be taken into count, as they both produce the genotype 1/2 in the offspring.

The genetic marker which was discussed here is an SNP marker. Another type of genetic marker is the *single tandem repeat* (STR) type marker. This marker marks a piece of the genome which contains a repeating DNA pattern, the number of repeats varies within a population. Single tandem repeats are also called microsatellites.

The database NorwegianFrequencies [8] in the R library forrel contains frequency data for 35 STR markers. Table 2 shows one of these markers, CSF1PO. The allele designations, the numbers 7, 8,...,16, refer to the number of times a specific sequence, **ATCT** in this case, is repeated [1]. As an example, the allele denoted 10 has the sequence repeated 10 times. In relatively rare cases, there may be a partial repeat and that is why the allele called 10.3 may exist. After

ten repeats follows only **ATC**, only three out of four base pairs were present in the eleventh repeat.

allele	frequency
7	0.0006
8	0.0031
9	0.0221
10	0.2472
10.3	0.0001
11	0.2989
12	0.3299
13	0.0812
14	0.0133
15	0.0036
16	0.0001

Table 2: The marker CSF1PO from the database NorwegianFrequencies in the R library forrel [8].

2.1.2 The Hardy-Weinberg principle

The Hardy-Weinberg principle, also called Hardy-Weinberg equilibrium (HWE), is a principle which states that the allele frequencies for a gene within a large population will stay the same over time. Each new generation will then inherit the allele frequencies from the previous. If HWE is assumed true for a population, then it is assumed that for all genes in the genome of the species, genotype frequencies will remain constant over time from generation to generation.

An example gene has alleles A and a, with allele frequencies p and q, respectively. If the Hardy-Weinberg equilibrium holds for this gene, then within a population:

- Genotype AA occurs with frequency p^2
- Genotype Aa occurs with frequency $2pq$

- Genotype aa occurs with frequency q^2

HWE is important for statistical calculations since probabilities of genotypes can then be calculated using no other information than allele frequencies. There are reasons why the Hardy-Weinberg equilibrium may not be true for a population. Some assumptions are therefore necessary when HWE is assumed true. A necessary assumption is that mating is random. The individuals within the population may be objected to some sort of selection, in which case the probability of mating may depend on genotype. This is true if a specific genotype provides a natural advantage. When mating probabilities in a population depend on a gene, then the proportions of the alleles for this gene within this population would increase or decrease in the long run.

HWE states that genotype frequencies will remain constant from generation to generation. Under this assumption one will be able to compute the genotype probabilities for individuals of the current generation as well as for individuals of all future generations with no more information than the current proportions of each allele. Then, the probabilities are as simple as

$$P(AA) = p^2,\ P(Aa) = 2pq,\ P(aa) = q^2$$

as mentioned previously. If there is Hardy-Weinberg *disequilibrium*, then these genotype frequency formulas will not hold. The change in the probability formulas can be explained and modelled through a parameter, θ. The formulas for genotype frequencies will then instead be these;

$$\begin{aligned} P(AA) &= \theta p + (1-\theta)p^2, \\ P(Aa) &= 2pq(1-\theta), \\ P(aa) &= \theta q + (1-\theta)q^2. \end{aligned}$$

Here θ is a measurement of general background relationship between individuals not connected by a pedigree. Typical values for θ are in the interval $[0, 0.03]$. Note that in the case of HWE, $\theta = 0$, and the formulas are simplified. The general case, with more than two alleles, is explained in [9].

The Hardy-Weinberg equilibrium is assumed to be **true** in the analysis done in this thesis.

2.2 Forensic statistics and identification

We first briefly explain forensic statistics generally. Consider two hypotheses

- $H1$: AF is the biological father of C.
- $H2$: AF and C are unrelated.

The objective is to summarise the statistical evidence to help decision makers reach a conclusion. In a case with two hypotheses, two possible errors can be made. These are concluding with H_2 when H_1 is true and concluding with H_1 when H_2 is true. Conventionally, hypotheses are tested using p-values in statistics. This approach assumes that it is most important to avoid the first error. This assumption cannot be made in forensics and p-value based testing is not used. Usually more than two hypotheses are made in a forensic case.

Through statistical analysis investigators want to determine which hypothesis is most likely. This is done by calculating how realistic each possible hypothesis is and comparing them against each other using this information. The recommended approach is to report the likelihood ratio (LR) defined as

$$LR = \frac{P(data \mid H_1)}{P(data \mid H_2)}.$$

A value of 1 is neutral whereas large or small values favour H_1 or H_2, respectively. There are several tables published giving thresholds for the LR, one is reproduced in Figure 2 from [5]. Likelihood ratios are produced using forensic data. An example of a likelihood ratio being calculated in a forensic case (paternity case) is in Section 3.1.

It is possible to make multiple LR-s in a forensic case, using different kinds of data to calculate each LR. If $LR_1, LR_2, \ldots LR_n$ have been calculated from independent data, we can multiply these LR-s to get the overall likelihood ratio.

DNA readings and other forensic evidence do not provide an indisputable answer to forensic cases. However, forensic evidence may provide insight in

Ratio	Expert guidance: the forensic findings…
1	… do not support one proposition over the other
2-10	… provide weak support
10-100	… provide moderate support
100-1,000	… provide moderately strong support
1,000-10,000	… provide strong support
10,000-1 million	… provide very strong support
Over 1 million	… provide extremely strong support[27]

Figure 2: Threshold values for LR. The table is copied from [5]

terms of probability. Through forensic analysis researchers may find that one hypothesis is more plausible than others, the difference in probability between hypotheses may be expressed by magnitude. If the hypothesis with the highest probability is less than ten times as likely than at least one of the others, the forensic data provides weak support in favour of this hypothesis. On the other hand, if the highest probability hypothesis is shown to be millions of times more likely than any other hypothesis, this hypothesis has particularly strong support as shown in Figure 2.

2.2.1 Bayesian approach

Alternatively, a Bayesian approach can be used. Then prior probabilities need to be specified.

The assigned probability of the hypothesis being true before taking the forensic data into count is the *prior probability*. All valid hypotheses need to be assigned a prior probability by the people doing the statistical analysis. The term prior probability is not precisely defined, and it is up to the researcher how much information they will include to form these probabilities. The *posterior probability* of a hypothesis is the probability of the hypothesis being true when the prior probabilities and forensic data are taken into count. Below the Bayesian approach is detailed.

Let $P(H_j)$ denote the prior probability of hypothesis H_j. Furthermore, $LR_{j,1}$ is the likelihood ratio when hypothesis j is compared to hypothesis 1. Note that $LR_{1,1} = 1$. Then, if there are k hypotheses, Bayes theorem gives

$$P(H_i \mid \text{data}) = \frac{LR_{i,1}P(H_i)}{\sum_{j=1}^{k}(LR_{j,1}P(H_j))}$$

which simplifies to

$$P(H_i \mid \text{data}) = \frac{LR_{i,1}}{\sum_{j=1}^{k} LR_{j,1}}. \tag{1}$$

for *flat priors*, i.e. when all prior probabilities are equal, i.e.

$P(H_1) = ... = P(H_k) = 1/k$.

Bayes theorem may alternatively be formulated on odds form

$$\frac{P(H_1 \mid \text{data})}{P(H_2 \mid \text{data})} = \frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_2)} \times \frac{P(H_1)}{P(H_2)} \tag{2}$$

which may be formulated verbally as

$$\text{posterior odds} = LR \cdot \text{prior odds}.$$

In the above example with flat priors, the prior odds is 1 and the posterior odds equals LR, but the interpretation differs. The posterior odds refers to the probability of the hypotheses given the data.

2.2.2 Forensic identification

Forensic identification is the process of identifying someone or something using forensic sciences. Techniques within forensic sciences may be applied to identify humans, animals or objects. The techniques used for identifying humans are relevant for this thesis. With these techniques, forensic researchers analyse various attributes of human subjects such as DNA, size, dental structure, injuries, fingerprints and tattoos. Data collected from evidence in a forensic investigation is referred to as forensic data. Analysts try to find links between the forensic data and suspects.

2.2.3 Kinship cases

Not all forensic cases can be solved using DNA analysis alone. For instance, two close relatives may be impossible to separate based on their post-mortem DNA readings and DNA readings from their extended family. For instance, in a case involving two brothers as the victims, the DNA data may not suffice to solve the case by itself. If the victims have no descendants, the family relation between each of them and each known relative is exactly the same, because their genetic relations go through their parents. The genetic overlap with each parent will be 50% for both of them, so there are no differences in genetic relations. Thus, the DNA of these relatives may not be used to separate the brothers from one another. If one or both the brothers have one or more biological children, then these children could be DNA tested. These tests could give information on how the brothers should be separated as will be exemplified in Section 3.5. If neither of the brothers have children, then all the DNA information to go by is the information from the parents or from someone related to the parents, and they will be impossible to separate.

Another unsolvable case involves two victims. The missing people are a mother and a daughter. No relatives of these missing people are available for genotyping. Separating the two will be impossible, because they share 50% of their genome with one another and there is no information in the DNA which can reveal which victim the mother is and which victim the daughter is.

2.2.4 Disaster victim identification

Within forensic sciences, researchers are sometimes required to identify victims of a disaster. This field is known as disaster victim identification, *DVI*. Determining who a dead victim is may be done with forensic techniques. In an investigation of a DVI case, the data collected by investigators is sorted into two categories; post-mortem (PM) and ante-mortem (AM). PM data is data collected from unidentified disaster victims. AM data is data collected either from family members of missing individuals, or from known information about

the missing individuals themselves. Figure 1 in Section 1.1 provides an example. Here the PM data is data collected from the victims, while the AM data is data collected from the mother of two missing brothers.

In a hypothesis, each disaster victim is paired with either one of the reported missing people, or no individual at all. Thus, a disaster victim is never paired with more than one missing person. Simultaneously, none of the missing people are paired with more than one disaster victim. These restrictions provide grounds for rejecting hypotheses which break them. Some pairings between a victim and a missing person may be deemed impossible prior to the statistical analysis, for instance if the sex of the victim is identified and is not that of the missing person. Hypotheses containing impossible pairings are impossible and may be ignored in the statistical analysis. If no victim-missing pairing is known to be impossible, then all combinations of pairings which can be made need to be considered as valid hypotheses. The total number of valid hypotheses in a forensic case can be in the millions if many victims and missing people are involved. Generation and rejection of DVI hypotheses for a forensic case is one of the goals of this thesis, and general solutions have been implemented as functions in R.

2.2.5 Kinship blind search

Blind search is a DNA analysis method where no AM data is used. When this method is used for kinship analysis, the goal is to find patterns in the data to try to extract information about possible family bonds between victims and similar. The term 'blind' indicates that we are searching for certain specified close relationships among PM samples, like parent - offspring, sibs and half-sibs.

2.3 Models for non-genetic data

In this section, forensic data other than DNA data is used. DNA data can be combined with more DNA data in the same way DNA data can be combined with other forensic data. If there is reason to believe that the DNA data will

be correlated with other data types, then this may in principle be taken into count by including the correlation between the data types into the calculations, similar to how correlated markers may be handled. Ultimately, combining DNA data with non-DNA data in forensic analysis is at the core of this thesis.

For instance, a missing person may be known to have a bone fracture in their body. If this is the case, then one would assume to find the same bone fracture in the victim, if they are the same person. Through knowledge about the injury and statistics, one can calculate posterior probabilities for different DVI hypotheses based on data from victims and missing people about such an injury.

2.3.1 Review of DNA based kinship testing and DVI problems

We briefly review the essential parts of [2]. Assume a DVI case has victims $V_1, \ldots, V_{nV}$ and missing individuals $M_1, \ldots, M_{nM}$. DNA data from the victims will be the PM data in the investigation while DNA data from reference families will be the AM data.

Each family includes at least one genotyped reference individual R_i. A possible solution, referred to as an *assignment*, to the DVI problem we are addressing, is a one-to-one correspondence, denoted a, between a subset of $\mathcal{V} = \{\mathrm{V}_1, \ldots, \mathrm{V}_{nV}\}$ and a subset of $\mathcal{M} = \{\mathrm{M}_1, \ldots, \mathrm{M}_{nM}\}$. Note that the empty assignment $(*, \ldots, *)$ is a valid solution, referred to as the *null model* below and denoted a_0. Essentially all results rely on the assignment likelihood

$$L(a) = P(\text{PM and AM data} \mid a, \Phi) \tag{3}$$

where Φ includes the fixed parameters, i.e., reference pedigree, marker allele frequencies and mutation models.

2.3.2 Statistical model

While the likelihood for the genetic data in equation (3) is based on well established methods, there is little information published on how statistical models

for non-genetic evidence should be formed, and how likelihoods should be calculated. Some statistical models will be formed and used in this thesis, they will differ from those presented in the paper by Franco Marsico and Inés Caridi: "Incorporating Non-Genetic Evidence in Large Scale Missing Person Searches: A General Approach Beyond Filtering" [6].

Considering a statistical model for features, when a forensic feature is accounted for in a DVI case, it will be assumed that one value to describe this feature is observed in each victim and in each missing person. This value will be a whole number between 1 and c, where c is the number of possibilities. Exceptions might be made for binary cases where one feature may be considered a non-feature and given value 0. Such an example may be in the case of observing whether or not a person has an injury, a feature which can be denoted by numbers 1 or 2 for having the injury and not having the injury. Instead, this feature may be labelled with 1 and 0 for having the injury and not having the injury, respectively. This latter approach might seem natural as 0 now means "no injury" or "none". A features with no more than two possible states is called a *binary feature*.

Another example of a feature which may be denoted by numbers is age. This feature could simply note the age of a person in years, but could also denote age categorizations, ranging from 1, the youngest category to c, the oldest category. Let $x = (x_1, \ldots, x_{nV})$ and $y = (y_1, \ldots, y_{nM})$ denote the values for the victims and the missing people, respectively. Here nV denotes the number of victims while nM denotes the number of missing people. We assume independence, i.e.,

$$p(x) = \Pi_{i=1}^{nV} P(x_i) \text{ and } p(y) = \Pi_{j=1}^{nM} P(y_j) \tag{4}$$

if not stated otherwise. The conditional distribution of $x_i \mid (y_j, V_i = M_j)$ is modelled by a $c \times c$ matrix $M = (m_{st})$ where all elements have non-negative values and the sum of each row is 1. M is a transition matrix, used e.g. to model mutations in forensic genetics as explained in [10]. In other words,

$$P(x_i = s \mid y_j = t, V_i = M_j) = m_{st}$$

$$P(x_i = s \mid y_j = t, V_i \text{ and } M_j \text{ are unrelated}) = P(x_i = s) = p_s \tag{5}$$

The simplification in equation (5) is possible because if V_i and M_j are unrelated, then the attributes of V_i and M_j are independent of another. Because y_j and x_i are independent variables the value of y_j does not affect the probability distribution of x_i, hence the expression may be simplified like there was no known prior data. It is reasonable that x and y should have the same marginal distribution. If we assume that the model (M, p) is stationary, i.e., $pM = p$, then $P(y_j = s) = p_s$. A simple sufficient condition for stationarity is that the detailed balance holds [10], i.e.,

$$p_s m_{st} = p_t m_{ts} \text{ if } s \neq t, \tag{6}$$

and then (M, p) is a reversible model. In the binary case, $c = 2$, reversibility and stationarity are equivalent. However, stationarity is a quite strong assumption imposing restrictions on the transition matrix as the following section shows.

2.4 Forming the statistical model

This case only involves one victim, $V1$, and one missing individual, $M1$. x and y are the variables which describe their features, respectively. The data collected on the missing person $M1$ is assumed to be correct, hence, **y is assumed observed without uncertainty**. If $V1$ and $M1$ are the same person, then x is expected to have the same value as y if no error was done in the identification process, and the forensic material was in good shape. Because identification will not always be accurate, the possibility of a misidentification should be taken into count.

A matrix expressing the probability of matches or mismatches in the case of a binary feature is shown in equation (7). The same matrix is shown with parameters instead of probability expressions in equation (8)

$$M = \begin{pmatrix} P(x_i = 1 \mid y_j = 1, V_i = M_j) & P(x_i = 0 \mid y_j = 1, V_i = M_j) \\ P(x_i = 1 \mid y_j = 0, V_i = M_j) & P(x_i = 0 \mid y_j = 0, V_i = M_j) \end{pmatrix} \tag{7}$$

$$M = \left(\begin{array}{c|cc} y\backslash x & 1 & 2 \\ \hline 1 & 1-\mu_1 & \mu_1 \\ 2 & \mu_2 & 1-\mu_2 \end{array}\right) \tag{8}$$

Here, μ_1 is the probability of a victim having feature 1 is misclassified and reported as $x = 2$. Similarly, μ_2 is the probability that feature 2 is misclassified as feature 1. It would be possible that the chance of misclassification was the same for both features, denoted as μ. However, that would lead to some issues with calculations, as explained in this section.

A victim and a missing person are one and the same, and their observed features are x (PM data) and y (AM data), respectively. When a value of y is observed, it is assumed observed without uncertainty, hence a y value of 1 means we assume that the missing person has feature 1 with 100% certainty. The matrix says that if the probability of y being feature 1 is 100% or 1, then the probability of x being feature 1 is $1 - \mu_1$ and the probability of x being feature 2 is μ_1. Hence the matrix is used to calculate probabilities for what an observed value of x will be using values of y. This is shown in equation (9).

$$\begin{aligned} y &= (100\%, 0\%) = (1, 0) \\ x &= yM = (1 - \mu_1, \mu_1) \end{aligned} \tag{9}$$

Example: An unreasonable model

Consider the binary case with only one victim and one missing person. The two features are 1 and 2. Assume that the probability of misclassification, μ, is the same for both features. Here, $H1$ is the hypothesis stating that the victim and

the missing person are one and the same.

$$P(x = 1 \mid y = 2, H_1) = P(x = 2 \mid y = 1, H_1) = \mu.$$

Assume also that the probability of being reported to have feature 1 is the same for both.

$$P(X = 1) = P(Y = 1) = \alpha$$

$$P(X = 2) = P(Y = 2) = 1 - \alpha$$

This is a seemingly reasonable and intuitive approach but it is arithmetically problematic. Problems arise when it is considered that the observed feature of the victim is subject to error, while the feature of a missing person is not. Calculating $P(X = 1)$ given that $H1$ is true gives equation (10).

$$\begin{aligned} P(X = 1|V = M) = {} & P(X = 1|Y = 2, V = M)P(Y = 2) \\ & + P(X = 1|Y = 1, V = M)P(Y = 1) \end{aligned} \quad (10)$$

Then the result in equation (11) follows.

$$P(X = 1|V = M) = \mu * (1 - \alpha) + (1 - \mu) * \alpha = \mu + \alpha - 2\mu\alpha \quad (11)$$

$P(X = 1)$ should be independent of the underlying hypothesis if y is unknown, and should thus be equal to α. According to equation (11), then $\alpha = \mu + \alpha - 2\mu\alpha$. Therefore one must either have the restriction $\mu = 0$, or the restriction $p_1 = p_2 = 0.5$, where $p_1 = \alpha$ is the frequency of feature 1 and $p_2 = (1 - \alpha)$ is the frequency of feature 2. The same conclusions are reached when considering the stationary requirement, as then the detailed balance (6) implies

$$p_1\mu = (1 - p_1)\mu$$

Again one must set μ to 0 or accept $p_1 = p_2 = 0.5$.

The model exemplified above is not reasonable, as one would either need to assume that the frequencies of the two features are equal, or that there is no

possibility of misclassification. Otherwise one would break mathematical rules. Typically the information on p is more precise than on M. Hence, p would be specified. Then M could be chosen based on the information available. One could accept that the model is non-stationary, as is typically the case in forensic genetics. Alternatively, a stationary model could be chosen directly or (M, p) could be transformed to a stationary model [11] or a reversible model [12].

Example: Stationary model

The model (M, p) where

$$M = \begin{pmatrix} 1-\mu & \mu \\ \frac{p_1}{1-p_1}\mu & 1-\frac{p_1}{1-p_1}\mu \end{pmatrix} \tag{12}$$

is well defined and stationary if $0 \leq 1 - \frac{p_1}{1-p_1}\mu \leq 1$.

The model (M, p) is bounded if

$$m_{st} \leq p_t \text{ if } s \neq t. \tag{13}$$

In other words, the probability of observing the value t in the victim as a result of misclassification, is bounded from above by the prevalence of t.

The model (M, p) where M is the matrix and p is the vector is stationary if $pM = p$, as mentioned previously. This means vector p may be multiplied by M as many times as one may want, and it will remain the same.

This matrix corresponds to a case of two possible features, feature 1 with frequency p_1 and feature 2 with frequency p_2. Because the sum of frequencies must be 1, $p_2 = 1 - p_1$. Calculating the product between the vector of feature probabilities and the matrix M will result in the feature probability vector, proving the matrix to be stationary.

One may solve the issue of conditional probabilities by introducing one more parameter. The probabilities of the features 1 and 2 are p_1 and p_2, respectively. They must add up to 1, so p_1 uniquely defines p_2 and vice versa. However, the parameter μ, which denotes the probability of a misdiagnosis, does not need to be the same for the two features. It is possible that that the chance of mistaking

feature 2 for feature 1 is different from the probability of mistaking feature 1 for feature 2. The parameter μ will then be replaced by parameters μ_1 for feature 1 and μ_2 for feature 2. If one is also willing to accept that μ_2 must be given a specific value based on μ_1, p_1 and p_2, then it will be possible to create a statistical model which resolves the conditional probability issue, such that $P(X = 1) = P(Y = 1) = p_1$ and $P(X = 2) = P(Y = 2) = p_2$.

This derivation determines the value μ_2 must take to make the conditional probability matrix stationary.

$$\begin{bmatrix} p_1 & p_2 \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \end{bmatrix} \begin{bmatrix} 1-\mu_1 & \mu_1 \\ \mu_2 & 1-\mu_2 \end{bmatrix}$$

$$\begin{bmatrix} p_1 & 1-p_1 \end{bmatrix} = \begin{bmatrix} p_1 & 1-p_1 \end{bmatrix} \begin{bmatrix} 1-\mu_1 & \mu_1 \\ \mu_2 & 1-\mu_2 \end{bmatrix}$$

$$\begin{bmatrix} p_1 & 1-p_1 \end{bmatrix} = \begin{bmatrix} (1-\mu_1)p_1 & \mu_1 p_1 \\ + & + \\ \mu_2(1-p_1) & (1-\mu_2)(1-p_1) \end{bmatrix}$$

$$\begin{bmatrix} p_1 & 1-p_1 \end{bmatrix} = \begin{bmatrix} p_1+\mu_2-\mu_1 p_1-\mu_2 p_1 & 1-p_1-\mu_2+\mu_1 p_1+\mu_2 p_1 \end{bmatrix}$$

It is clear that $\mu_2 - \mu_1 p_1 - \mu_2 p_1 = 0$. Thus $\mu_2 - \mu_2 p_1 = \mu_1 p_1$, and

$\mu_2 = \mu_1 \frac{p_1}{1-p_1}$

Following is the same calculation done for the sake of verifying that the substitutions $\mu_1 = \mu$ and $\mu_2 = \frac{p_1}{1-p_1}\mu$ do make a stationary model.

$$M = \begin{pmatrix} y\backslash x & 1 & 2 & 3 \\ 1 & \lambda_1 & \lambda_1 * \epsilon_{12} & \lambda_1 * \epsilon_{13} \\ 2 & \lambda_2 * \epsilon_{21} & \lambda_2 & \lambda_2 * \epsilon_{23} \\ 3 & \lambda_3 * \epsilon_{31} & \lambda_3 * \epsilon_{32} & \lambda_3 \end{pmatrix}$$

Figure 3: A probability matrix M in the case of three features. Here, λ_i is the probability of feature i being classified correctly in a victim. The probability of feature i being misclassified as feature j is denoted as $\lambda_i * \epsilon_{ij}$.

$$pM = \begin{bmatrix} p_1 & p_2 \end{bmatrix} \begin{bmatrix} 1-\mu & \mu \\ \frac{p_1}{1-p_1}\mu & 1-\frac{p_1}{1-p_1}\mu \end{bmatrix}$$

$$pM = \begin{bmatrix} p_1 & 1-p_1 \end{bmatrix} \begin{bmatrix} 1-\mu & \mu \\ \frac{p_1}{1-p_1}\mu & 1-\frac{p_1}{1-p_1}\mu \end{bmatrix}$$

$$pM = \begin{bmatrix} (1-\mu)p_1 & \mu p_1 \\ + & + \\ \frac{p_1}{1-p_1}\mu(1-p_1) & (1-\frac{p_1}{1-p_1}\mu)(1-p_1) \end{bmatrix}$$

$$pM = \begin{bmatrix} p_1 - \mu p_1 + \mu p_1 & \mu p_1 + 1 - p_1 - \mu p_1 \end{bmatrix}$$

$$pM = \begin{bmatrix} p_1 & 1-p_1 \end{bmatrix}$$

$$pM = \begin{bmatrix} p_1 & p_2 \end{bmatrix}$$

$$pM = p$$

In the case considered, x and y take on one of two possible values. If there are more than two possibilities for x and y, a larger transition matrix, also denoted by M, is required to contain all the possible combinations of feature values and misdiagnoses. As an example, consider a feature with three possible states, denoted 1, 2 and 3. The former matrix of conditional probabilities may be expanded to include all three states both in victims and in missing people. This is done in Figure 3. This matrix form is proposed in [6].

$$M = \begin{pmatrix} y\backslash x & 1 & 2 & 3 & 4 & 5 \\ 1 & \lambda_1 & \lambda_1 * \epsilon_{12} & \lambda_1 * \epsilon_{13} & \lambda_1 * \epsilon_{14} & \lambda_1 * \epsilon_{15} \\ 2 & \lambda_2 * \epsilon_{21} & \lambda_2 & \lambda_2 * \epsilon_{23} & \lambda_2 * \epsilon_{24} & \lambda_2 * \epsilon_{25} \\ 3 & \lambda_3 * \epsilon_{31} & \lambda_3 * \epsilon_{32} & \lambda_3 & \lambda_3 * \epsilon_{34} & \lambda_3 * \epsilon_{35} \\ 4 & \lambda_4 * \epsilon_{41} & \lambda_4 * \epsilon_{42} & \lambda_4 * \epsilon_{43} & \lambda_4 & \lambda_4 * \epsilon_{45} \\ 5 & \lambda_5 * \epsilon_{51} & \lambda_5 * \epsilon_{52} & \lambda_5 * \epsilon_{53} & \lambda_5 * \epsilon_{54} & \lambda_5 \end{pmatrix}$$

Figure 4: A probability matrix M in the case of five features. Parameters λ_i and ϵ_{ij} have the same meaning as in Figure 3.

One may expand the matrix further, like in Figure 4 where the number of feature values is five. Regardless of how many different values a feature is represented by, the stationary property is desired. Thus, we want $p = pM$ for feature frequency vector p and probability matrix M.

The general formula for the LR for hypothesis H_a against hypothesis H_0 is derived and the result is equation (14). Here H_0 is the hypothesis which contains no matches between victims and missing people.

$$LR_a = \frac{P(XI_1 = xI_1, ..., XI_v = xI_v \mid YI_1 = yI_1, ..., YI_m = yI_m, H_a)}{P(XI_1 = xI_1, ..., XI_v = xI_v \mid YI_1 = yI_1, ..., YI_m = yI_m, H_0)}$$

$$LR_a = \frac{P(XI_1 = xI_1 \mid YI_1 = yI_1, ..., YI_m = yI_m, H_a)}{P(XI_1 = xI_1)} \times ... \\ ... \times \frac{P(XI_v = xI_v \mid YI_1 = yI_1, ..., YI_m = yI_m, H_a)}{P(XI_v = xI_v)}$$

$$LR_a = \prod_{k=1}^{v} \frac{P(XI_k = xI_k | YI_1 = yI_1, ..., YI_m = yI_m, H_a)}{P(XI_k = xI_k)}$$

$$LR_a = \prod_{k \in H_a} \frac{P(XI_k = xI_k | YI_k = yI_k, VI_k = MI_k)}{P(XI_k = xI_k)} \prod_{k \notin H_a} \frac{P(XI_k = xI_k)}{P(XI_k = xI_k)}$$

$$LR_a = \prod_{k \in H_a} \frac{P(XI_k = xI_k | YI_k = yI_k, VI_k = MI_k)}{P(XI_k = xI_k)}$$

$$LR_a = \prod_{k \in H_a} \prod_{f \in features} \frac{P(XI_k f = xI_k f | YI_k f = yI_k f, VI_k = MI_k)}{P(XI_k f = xI_k f)}$$

$$LR_a = \prod_{k \in H_a} \prod_{f \in features} \frac{Mf_{yx}}{P(XI_k f = xI_k f)} \tag{14}$$

Here v is the number of victims and m is the number of missing people. Victims are denoted with V-s, while missing people are denoted with M-s. The X-s and Y-s denote stochastic variables for the reported features for victims and missing people, respectively. Similarly, the x-s and y-s denote the values observed for these stochastic variables in the forensic investigation. All V-s, M-s, X-s, Y-s, x-s and y-s are indexed with $I_1 ... I_v$ for victims and $I_1 ... I_m$ for missing people. For any k such that $1 \leq k \leq v$, VI_k, XI_k and xI_k reference the same victim. Similarly, for any k such that $1 \leq k \leq m$, MI_k, YI_k and yI_k reference the same missing person.

Indices $I_1 ... I_v$ and $I_1 ... I_m$ are arranged such that within any victim-missing pair in the hypothesis H_a, the victim and the missing person are given the same index. In other words, they are victim VI_k and missing person MI_k for some k. The notation $k \in H_a$ means that the victim VI_k with feature XI_k and observed feature value xI_k is in a victim-missing pair in H_a. The X-s and Y-s may contain data for multiple statistically independent features, the individual observed features are denoted $XI_k f$ and $YI_k f$ for some feature f. Mf is here the feature classification probability matrix M for feature f. The probability matrices are on the form explained in this section, for example, with five possible values for a feature, the matrix will be on the form in Figure 4. Mf_{yx} denotes the number in row y and column x in this matrix.

Note that in this result the assumption that features are independent of each other is made. This may not be a particularly realistic assumption.

The following will be a discussion of the paper by Franco Marsico and Inés Caridi [6]. This paper suggests using non-genetic forensic data alongside forensic data to identify. Hypotheses tested in the paper are constructed the same way as the hypotheses in this one, with a set of missing people and a set of victims (called unidentified persons) where victims and missing people are paired against each other. The paper mentions that in the past, when doing forensic analysis,

one would filter out impossible victim-missing pairs from the analysis prior to running a hypothesis search. The paper argues against this approach because of the possibility of errors in the forensic investigation with regards to identifying features in the victims.

The paper argues for splitting the statistical analysis into multiple steps. Bayes theorem is used for this. Assuming the forensic post-mortem data is split into k segments, and all these are statistically independent of each other, the posterior probability computation for a hypothesis H_i may be done like this:

$$P(H_i|D_{j+1}) = \frac{P(D_{j+1}|H_i)P(H_i)}{P(D_{j+1})}$$
$$P(H_i|D_{j+1}) = \frac{P(D_j|H_i)P(d_{j+1}|H_i)P(H_i)}{P(D_j)P(d_{j+1})}$$
$$P(H_i|D_{j+1}) = \frac{P(d_{j+1}|H_i)}{P(d_{j+1})}\frac{P(D_j|H_i)P(H_i)}{P(D_j)}$$
$$P(H_i|D_{j+1}) = \frac{P(d_{j+1}|H_i)P(H_i|D_j)}{P(d_{j+1})}$$

The equation shows how different forms of forensic data, assumed independent, may be combined iteratively. The lower-case d_j denotes one kind of forensic data, the different kinds are called $d_1, d_2, \ldots$. The upper case D_j denotes the combination of all data types up to d_j, i.e., $d_1, \ldots, d_j$.

In this paper, there are five possible values for hair colour and two possible values for sex. However, we do not have a set number of age groups. Age is taken into count through the use of the floating bin approach, which separates people of different ages into different age range groups. This model does to some extent account for the variability in the data. As per the feature model described above, age groups close to one another time-wise may be assigned higher probability of being warped to a close group than to a far away group. The uncertainty in this variable is associated with inaccurate testimony or laboratory estimations. Instead of designing age bins a priori, each bin is designed based on the victim that is being identified. A match between ages happens when there is overlap between the age intervals of the victim and that of the missing person. The age of the missing person may be assumed without uncertainty, though this is not what is done in Marsico/Caridi, rather something done in this thesis.

With sex, hair colour and age as the non-genetic data, the formula for com-

bined likelihood is defined as in equation (15). The likelihood ratios for all non-genetic data combined is here noted as LR_{NG}. The likelihood ratio when only looking at age is LR_A, for sex it is LR_S and for hair colour it is LR_C.

$$LR_{NG} = LR_S * LR_C * LR_A \tag{15}$$

This does assume independence between the attributes, in practice this is not actually the case. There is dependence between age and sex, because women live longer on average. Proportions of hair colour in populations of different ages is also not the same, for example, grey hair is a lot more common among the older generation.

2.4.1 Likelihood ratios

Consider first $H_1 : V_i = M_j$ and $H_2 : V_i$ and M_j are unrelated. Then

$$\mathrm{LR}_{ij}^{\mathrm{NG}} = \frac{P(X_i = x_i, Y_j = y_j \mid H_1)}{P(X_i = x_i, Y_j = y_j \mid H_2)} \tag{16}$$

$$\mathrm{LR}_{ij}^{\mathrm{NG}} = \frac{P(X_i = x_i \mid Y_j = y_j, H_1)P(Y_j = y_j)}{P(X_i = x_i)P(Y_j = y_j)} = \frac{m_{ij}}{p_i} \tag{17}$$

where NG abbreviates 'Non Genetic' evidence. The likelihood ratio LR_{ij}^{NG} is the ratio of the likelihood of the hypothesis $V_i = M_j$ and the likelihood of the hypothesis $V_i \neq M_j$. It was assumed that $P(y_j \mid H_1) = P(y_j \mid H_2) = P(y_j)$. This assumption is justified by noting that while x_i and y_j are dependent stochastic variables if H_1 is true, if y_j is not known then x_i will have marginal probabilities $P(X_i = 1) = p_1$ and $P(X_i = 2) = p_2$, independent on which hypothesis is actually true. The fact that M_j is identical to V_i is irrelevant when we do not condition on data for V_i. If y_j is known but not x_i, the same is true for $P(Y_j = 1) = p_1$ and $P(Y_j = 2) = p_2$.

For a bounded model $\mathrm{LR}_{ij} \leq 1$ if the victim and the missing person do not share a feature.

Consider next a specific assignment a against a_0. Then the likelihood ratio based on non-genetic data is found with equation (18).

$$\mathrm{LR}_a^{\mathrm{NG}} = \frac{P(x, y \mid a)}{P(x, y \mid a_0)} = \Pi_{\{(i,j):\mathrm{V}_i=\mathrm{M}_j\}}\mathrm{LR}_{ij}^{NG}. \tag{18}$$

This is the product of the likelihood ratios for each pair in the hypothesis. For every victim-missing pair in the hypothesis a, the likelihood ratio as defined in equations (16) and (17) is included in the product. The combined likelihood ratio for the hypothesis a is LR_a^{NG}. Statistical independence is assumed between all factors in the products because in a valid hypothesis no victim or missing person will be included in more than one pair.

2.5 Combining genetic and non-genetic evidence

We assume independence between genetic evidence (G) and non-genetic evidence (NG). Hence, for a specific assignment a compared to the null model, the combined likelihood ratio becomes

$$\mathrm{LR}_a = \mathrm{LR}_a^{\mathrm{G}} \cdot \mathrm{LR}_a^{\mathrm{NG}}. \tag{19}$$

The *posterior pairing probabilities* $q_{i,j} = P(\mathrm{V}_i = \mathrm{M}_j \mid D)$ for $i = 1 \ldots, v$ and $j = 1, \ldots, m$, and the *posterior non-pairing probability*, $q_{i,*} = P(\mathrm{V}_i = * \mid D)$, where D denotes all evidence, genetic and non genetic, are computed as explained in [2]. Here v is the number of victims while m is the number of missing people. V_i and M_j are the ith victim and the jth missing person, respectively.

2.5.1 Example: Stationary binary model

Consider the binary case with only one victim and one missing person. Hypotheses H_1 and H_2 are as defined in Section 2.4.1. The transition matrix is given in (12). The likelihoods and likelihood ratios are given in Table 3. A few comments are in order

- If $\mu = 0$, the LR-s vanish if the missing person and victim do not share a feature.

x	y	$L(H_1)$	$L(H_2)$	LR	$p_2 = 1 - p_1$
1	1	$(1-\mu)p_1$	p_1p_1	$\frac{1-\mu}{p_1}$	$\frac{1-\mu}{p_1}$
2	1	μp_1	p_2p_1	$\frac{\mu}{p_2}$	$\frac{\mu}{p_2}$
1	2	$\frac{p_1}{1-p_1}\mu p_2$	p_1p_2	$\frac{\mu}{1-p_1}$	$\frac{\mu}{p_2}$
2	2	$(1-\frac{p_1}{1-p_1}\mu)p_2$	p_2p_2	$\frac{1-\frac{p_1}{1-p_1}\mu}{p_2}$	$\frac{p_2-p_1\mu}{p_2^2}$

Table 3: Distribution of LR for various values of x and y. The rightmost column simplifies the expressions for LR using $p_2 = 1 - p_1$.

- If $\mu = 0$, the LR-s are $1/p_1$ and $1/p_2$ for $(x = 1, y = 1)$ and $(x = 2, y = 2)$. These are intuitive values and also these values will decrease when μ increases.
- The expected likelihood ratio for $H1$ against $H2$ is;
  - $E(LR \mid H_1) = (\frac{\mu}{p_2} - 1)^2 + 1$, here $H1$ is assumed true.
  - $E(LR \mid H_2) = 1$, here $H2$ is assumed true.

The expected likelihood ratio for $H1$ against $H2$ depends on which hypothesis is true. Thus, $E(LR|H_1)$ and $E(LR|H_2)$ have different values.

$$E(LR \mid H_1) = \sum_{x=1}^{2}\sum_{y=1}^{2}(\frac{P(x,y|H_1)}{P(x,y|H_2)}P(x,y|H_1))$$

$$E(LR \mid H_1) = \frac{(1-\mu)^2p_1^2}{p_1p_1} + \frac{\mu^2p_1^2}{p_2p_1} + \frac{\frac{p_1^2}{(1-p_1)^2}\mu^2p_2^2}{p_1p_2} + \frac{(1-\frac{p_1}{(1-p_1)}\mu)^2p_2^2}{p_2p_2}$$

Assuming $p_2 = 1 - p_1$, this expression can be simplified to $(\frac{\mu}{p_2} - 1)^2 + 1$.

$$E(LR \mid H_2) = \sum_{x=1}^{2}\sum_{y=1}^{2} \frac{P(x,y|H_1)}{P(x,y|H_2)} P(x,y|H_2)$$
$$E(LR \mid H_2) = \sum_{x=1}^{2}\sum_{y=1}^{2} P(x,y|H_1)$$
$$E(LR \mid H_2) = 1$$

$E(LR \mid H_2)$ could be found easily because it is just the sum of the probabilities of all possible x, y pairs under a hypothesis. All possible outcomes must sum to 1.

2.6 Generating hypotheses

The solution to DVI problems implemented in the R library dvir may be roughly summarized with two steps.

1. Generate all a priori possible assignments.
2. Calculate all likelihoods and sort them to obtain the best solutions.

In this section we first review a formula counting the number of a priori possible solutions. This number may be prohibitively large. The number may be reduced by taking into account e.g. age information, this will be discussed.

2.6.1 Number of assignments without gender

The number of possible solutions, assignments to a DVI problem is important as it indicates the complexity of the problem. Below we present a formula giving the required number based on [2]. Let $\mathcal{A}$ denote the sex-consistent assignments for a given DVI problem and $n = |\mathcal{A}|$, the number of elements. Assume first that sex is not known, neither for victims nor missing people. Then the total number of assignments is

$$\sum_{k=0}^{\min(s,m)} \binom{s}{k}\binom{m}{k} k!. \tag{20}$$

where s is the number of victims and m the number of missing people. The argument is as follows: For each k, there are $\binom{s}{k}$ different subsets of k victims. Each can be assigned to $\binom{m}{k}$ different subsets of the m missing people. In the end, each assignment can be shuffled in $k!$ ways.

When sex is known, the formula (20) applies to females and males independently, and the total number is

$$\begin{aligned} n &= n(s_F, s_M, m_F, m_M) \\ &= \left[\sum_{k=0}^{\min(s_F, m_F)} \binom{s_F}{k}\binom{m_F}{k}k!\right]\left[\sum_{k=0}^{\min(s_M, m_M)} \binom{s_M}{k}\binom{m_M}{k}k!\right], \end{aligned} \tag{21}$$

where s_F (s_M) is the number of female (male) victims and m_F (m_M) the number of female (male) missing individuals.

2.6.2 Sorting by age

The R-program *expandgridnodup2* includes an R function with the same name which identifies all the possible hypotheses for a forensic identification case. This function has been developed during the writing of this thesis. This kind of program is needed because it is very difficult to manually list and keep track of all hypotheses within a DVI case, as well as to recognize the hypotheses with high likelihood. An example using this code can be found in the appendix, see A.1.

A DVI hypothesis consists of a list of victim/missing person pairs. Not all victims need to be paired with a missing person and vice versa, the hypothesis will still be valid. Then, we wish to screen these hypotheses for maximum likelihood. The hypothesis with the highest likelihood may then be selected as the 'correct' one.

Several parameters need to be taken into count when forming these hypotheses, as some may be impossible. First of all, the hypotheses where one victim or missing person has two or more matches need to be eliminated from the analysis.

Often, the ages of victims are not provided as numbers, but rather as inequalities. Knowledge about the ages of the victims relative to one another may be beneficial for investigation and hypothesis testing. An example DVI case is provided here to demonstrate age restrictions. In this case there are three missing people; $M1$, $M2$ and $M3$, as well as three victims; $V1$, $V2$ and $V3$. The age of all three missing people is known, and from these ages it is deducted that $M1$ is the oldest of the three, while $M2$ is the youngest. Forensic methods performed on the victims concluded that $V1$ was younger than $V2$. One victim is deemed younger than another victim.

A matrix containing information about age is exemplified in Figure 5. In order to account for age in the calculations, the following steps were taken.

The first step was to convert age restrictions from string form to a matrix. For computer algorithms an age restriction matrix is desired, but age restrictions represented as strings are more readable for humans. In R code these strings may be, for example, "$V1 > V2$". The function *expandgridnodup*2 takes age restrictions on the string form as input and the age restriction matrix was constructed within the function. In the example in appendix A.2, the age restriction matrix was instead created manually. In the age restriction matrices, the number $+1$ means the victim corresponding to the row is older than the victim corresponding to the column, while -1 means the opposite. An example matrix is shown in Figure 5.

From the matrix it is clear that all age relations are known except that of $V4$ and $V1$. Information on whether $V1$ is older than $V4$ or vice versa was not provided.

2.6.3 How hypotheses are stored

Like *expandgridnodup*2 described in Section 2.6.2, the R functions included in appendix A.2 generate DVI hypotheses. These hypotheses will be generated in an R matrix and stored in an R data frame. While the idea behind these

String form: "$V1 > V2 > V3$", "$V2 < V4$"

Matrix form:

_	$V1$	$V2$	$V3$	$V4$
$V1$	0	1	1	0
$V2$	-1	0	1	-1
$V3$	-1	-1	0	-1
$V4$	0	1	1	0

Figure 5: Age restrictions represented on matrix form. This is how they are stored in the program *expandgridnodup*2 in appendix A.1. Note that the restrictions on the string form also imply that "$V3 < V4$".

functions is the same as that of *expandgridnodup*2, some changes are done in the implementation for the sake of speed. The functions are stored on Google Drive, and these are the functions which will be discussed the the following sections. The link is in appendix A.2. Like *expandgridnodup*2, these functions were also created during the writing of this thesis.

Among other data, these R functions need to be provided with an age matrix and a pair restriction matrix. This age matrix is on the same form as the one in Figure 5. The pair restriction matrix contains the possible pairings of victims and missing people, pairings not contained in the matrix are assumed to be impossible following results from forensic analysis. These pairings may have been rejected since the sexes of the victim and missing person do not match, or due to other factors.

The restrictions in these two matrices, the age restriction matrix and the pair restriction matrix, are processed by the R function *createrestrictionmatrix* in appendix A.2. This function will output new restriction matrices on a different form which may be difficult to read for a human, but is desirable for further computations. The form of this output is explained later in Section 2.6.5. This output is sent as an argument to the function *hypsolverrestrictions*, which is also in appendix A.2. This function generates the hypotheses and outputs an R data frame containing them.

A DVI hypothesis is defined as a set of pairs containing one victim and one missing person. One way to store them would be to represent them as a list of pairs, but this approach could quickly become messy. Instead, the hypotheses will be stored as lists of people, where the position of a person in the list will tell who the person is paired with. The lists will be made up of missing people, and each position in the list will correspond to a victim. One wants the possibility of a victim to be unassigned, hence empty slots in the lists will be allowed.

The hypotheses search is done in a way such that each of the victims is assigned one column in the hypothesis data frame. The data frame entries will consist of missing people. Each row of the data frame then corresponds to a hypothesis.

In the first step the first victim, or the leftmost column, will be assigned missing people. The missing people which are assigned to this victim are the missing people which may be paired with the victim in a hypothesis according to the restriction matrix.

After this first step, the next step is to fill out the column corresponding to the second victim with all the missing people it may be paired with. Using a Kronecker product, the hypothesis matrix is expanded to contain all possible combinations of pairs involving the first victim and pairs involving the second victim.

Figure 6 demonstrates how hypotheses are stored in the R matrix. The example case in the figure involves five victims, two of the victims have so far been assigned in this instance. Those victims are the two first columns. The first victim has been assigned no one, hence a zero. The second victim has been paired with the third missing person. The next three columns contain restrictions on the remaining victims. They contain the number 4 ($= 2^3$) because the third missing person has already been paired with the second victim. This will be further explained in Section 2.6.5. The last column contains a count of victims so far paired with an empty missing person, 1 in this case.

0	3	4	4	4	1

Figure 6: This figure shows an example of how a hypothesis is stored as a row in a matrix in R.

2.6.4 How hypotheses are generated

These hypotheses are generated by creating permutations of the list containing the missing people. The length of each permutation should be the number of victims. Each hypothesis must take into count which missing people can match with which victims. One hypothesis is a set of pairs consisting of one missing person and one victim, just like the DVI hypotheses discussed earlier.

First find all missing people which may be paired with the first victim. One may also include the possibility of an unassigned victim. Numbers referring to the selected missing people, along with a zero if the first victim may be left unassigned, are then placed in the leftmost column in the hypothesis matrix, starting at the top.

Expand the hypothesis list by Kronecker multiplication, such that each existing hypothesis can be expanded by a new victim-missing pair. All possible missing person matches for the victim corresponding to the next column are identified. If the victim not being assigned a missing person is possible, then this possibility must be taken into count, and "empty" is added to the list of possible matches. The total number of matches, including the empty missing person, is called $n.match$ in the function. The current number of generated hypotheses is called $n.hyp$. With Kronecker multiplication the hypothesis list is repeated $n.match$ times. Then, the next column in the hypothesis table is filled with missing people such that each hypothesis is paired with each of the $n.match$ missing people to form $n.hyp \times n.match$ total hypotheses.

Next, all invalid hypotheses are cut out, before the next victim column in filled. The process of adding missing people to the columns of the hypothesis matrix and removing invalid hypotheses is repeated until all victim columns have been filled. There are four reasons why a hypothesis could be invalid, these

are;

1. The victim and missing person in a pair do not match, due to not being of the same sex or due to other reasons.

2. The age restrictions on the victims and the age restrictions on the missing people contradict each other.

3. One missing person has been paired with two or more different victims. In practice there will at most be two because such hypotheses will be removed before the missing person is added to a third victim.

4. A hypothesis contains more victims assigned an "empty" missing person than what has been allowed.

Impossible pairs between victims and missing people are already accounted for when the columns are expanded. The age restrictions are kept track of in the rightmost victim columns, along with information on already used missing people. The age restriction matrices exemplified in Figure 8 are used to gradually add to the restrictions in the hypothesis matrix. These age restriction matrices will be explained in detail in Section 2.6.6.

2.6.5 How hypotheses are checked for validity

The hypothesis matrix is filled from left to right. When the first k victims have been assigned a missing person or been filled by an empty slot, only the numbers in the k leftmost columns refer to missing people and make up the hypotheses. The other columns may be used to store other information in the meantime.

Because each element in the matrix contains a number, the goal is to store as much information as possible in each. This may be done using bitwise operations. A 32-bit integer can then be used to store 32 boolean statements

Figure 7: The bit lookup table for restrictions for one matrix entry. Here, the missing people not allowed are #02 and #05. The integer value is $2^1 + 2^4 = 18$. Note that the last bit is 2^0, hence it is the "ones" bit.

which are either true or false. For simplicity, restrictions for one matrix cell are stored in that same cell. Hence, the restrictions on the third victim for the fifth hypothesis are stored in the cell in the fifth row and the third column. The matrix cells yet unused could have stored all the necessary information to validate hypotheses, but for simplicity one more column is added to the right end of the matrix. This column will store the number of victims not matched with a missing person but left empty for each row or hypothesis. Only victims representing already completed columns are counted here, not victims whose columns are yet to be filled with missing people.

If each bit corresponds to a missing person, one will be able to keep track of which out of (up to) 32 missing people may be placed in one cell in the matrix. Initially, each cell is assumed to be able to store any given missing person without breaking any of the limiting rules. All bits in all these cells will be 0, hence they all contain the number 0. As restrictions are identified, the corresponding bits will be changed to 1 with a bitwise OR operation.

To reduce computation, all age restrictions already have a set of bits calculated. These restrictions will then be applied continuously as the hypotheses are generated. One small matrix of age restrictions will be made for each victim prior to the generation. The information contained in these matrices will be applied once the victim has been assigned missing people.

This bit representation of hypothesis restrictions is shown in Figure 7. Bits set to 1 denote missing people who may not be eligible for the corresponding matrix cell.

This bit representation of data, with integers being used as lists of boolean values (true or false), is called a *bit field*. The main argument for storing data this way is that it cuts down the amount of computer memory necessary. By storing the restriction information in the same array as the hypotheses, less memory is utilized, and usage of memory is more efficient. More importantly, being able to represent entire lists of missing people as single integers greatly cuts down on required memory. Further, computers are designed to execute bitwise operations very quickly. Thus, the more efficient usage of memory which bit fields allow does not come at the cost of computation time.

2.6.6 Another way to store age information

In Section 2.6.2, age information was stored in a matrix as in Figure 5. With the function *createrestrictionmatrix*, the age information in that kind of matrix is processed and new restriction matrices are made, one matrix per victim. In fact, in these new matrices both age restrictions and restrictions on no overlapping pairs are handled. These matrices will then influence the hypothesis restrictions by bitwise OR operations. To illustrate what this matrix looks like, a DVI case is provided. This case has the following information;

- 5 victims given: $V1$,$V2$,$V3$,$V4$,$V5$
- 5 missing people given: $M1$,$M2$,$M3$,$M4$,$M5$
- Everyone is of the same sex, hence no victim-missing pairs are excluded based on that feature.
- Among ages of the missing people it is known that $M4$ is older than $M1$, that is, $M4 > M1$.
- Among ages of the victims it is known that $V1 > V2 > V3$, $V1 > V4$ and $V5 > V3$.

The restriction matrix has been made for $V1$ in Figure 8. The figure contains two versions of this matrix, one which is on the bit representation form explained

	$V2$	$V3$	$V4$	$V5$
$M1$	9	9	9	1
$M2$	2	2	2	2
$M3$	4	4	4	4
$M4$	8	8	8	8
$M5$	16	16	16	16

	$V2$	$V3$	$V4$	$V5$
$M1$	$M1, M4$	$M1, M4$	$M1, M4$	$M1$
$M2$	$M2$	$M2$	$M2$	$M2$
$M3$	$M3$	$M3$	$M3$	$M3$
$M4$	$M4$	$M4$	$M4$	$M4$
$M5$	$M5$	$M5$	$M5$	$M5$

Figure 8: Restriction matrix for $V1$. The matrix on the left is the numerical bit representation matrix as stored in memory. The matrix on the right contains the same information in the form of lists of missing people to be ruled out.

in Figure 7, and one which shows the missing people being ruled out. In a hypothesis, $V1$ will either be assigned one of the missing people or no one. If $V1$ is assigned a missing person, restrictions need to be implemented for this hypothesis. The matrix row corresponding to the chosen missing person contains these restrictions. If $V1$ is instead assigned no missing person, then the number of unassigned victims is incremented by one for this hypothesis.

Restrictions which matter for $V1$ are that if $V1$ is $M1$, then neither $V2$, $V3$ or $V4$ can be $M4$. Hence, if $V1$ is assigned $M1$ in a hypothesis, $M4$ must be ruled out for $V2$, $V3$ and $V4$ in that same hypothesis. At the same time, the function is not allowed to repeat the same missing person multiple times in one hypothesis. If $V1$ is assigned a missing person (non-empty), this missing person may not be assigned to another victim. Thus, this missing person must be ruled out so that no other victim will be paired with them. All victims except the last ($V5$ in this case) will have a matrix like this. The last victim does not need a restriction matrix, simply because when the column corresponding to this victim is filled out, the hypothesis generation is finished.

In Figure 8 one can see that the missing people $M1$, $M2$, ... , $M5$ all rule themselves out for all other victims if they are paired with $V1$. It is also clear that the age restrictions involving $V1$ are implemented in this matrix, as $V1 = M1$ rules out the possibility that $M4$ is paired with $V2$, $V3$ or $V4$.

2.6.7 Handling hypothesis limit

The function *hypsolverrestrictions* takes a maximum hypothesis number as one of its inputs. If this limit is exceeded, the hypothesis generation needs to be segmented. Instead of trying to create all possible hypotheses simultaneously, the search is divided. When the program registers that the hypothesis limit is about to be passed, the set of missing people matching the current victim will be split in two parts, where only the first part is being done in the current run. The other missing people need to wait for a future run. Information is stored on how the hypothesis search should be continued.

This DVI case is implemented and run with the functions *hypsolverrestrictions* and *createrestrictionmatrix* as an example in appendix A.2. The functions are ran with restrictions as described earlier.

3 Results

3.1 Paternity example

In this example the basic concepts of forensic genetics and paternity testing are illustrated in a simple paternity case. While a paternity case is a kinship case and not a DVI case, many of the same principles apply. Like when solving DVI cases, when solving kinship cases one will make a set of hypotheses and calculate their likelihood ratios. This paternity case involves three people: A male child, the undisputed mother of the child, and the alleged father of the child. The alleged father, the child and the mother have been given the abbreviations AF, C and M, respectively. The investigators will form two hypotheses; one which states that AF is the biological father of the child and one which states that AF and the child are unrelated. The two hypotheses will be compared to each other in a hypothesis test by calculating the likelihood ratio between them. If the hypothesis gives an LR above a specified threshold, see Figure 2, the conclusion is that AF is the biological father.

Formally, the two hypotheses are defined as previously in Section 2.2 as follows:

- H_1: The alleged father (AF) is the biological father.
- H_2: The alleged father and the child are unrelated.

The alleged father, the child and the mother are all genotyped for a set of genetic markers.

Because the biological mother of the child is undisputed in this forensic case, neither of the two hypotheses considered claim otherwise. The fact that the child is male has been implemented in the pedigree plots in Figure 9, as conventionally, squares represent males. A fourth person, the true father (TF), has been included in the plot of $H2$ in Figure 9. $H2$ states that AF and TF are not the same individual.

The code which creates the pedigree plots of the two hypotheses is here.

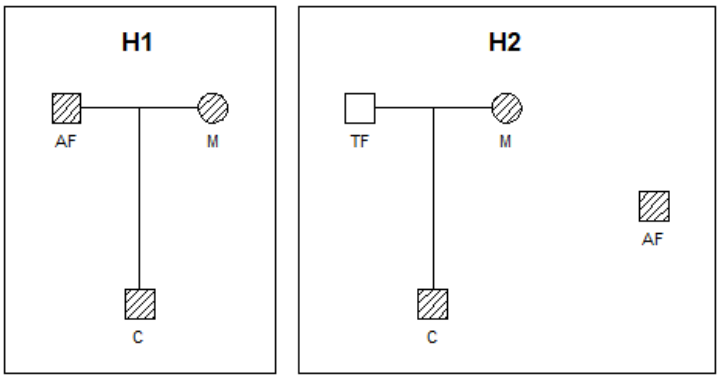


Figure 9: The paternity case. $H1$ and $H2$ denote the two cases.

```
library(dvir)
H1 = nuclearPed(mother = "M", father = "AF", children = "C")
H2 = list(nuclearPed(mother = "M", father = "TF",
children = "C"), singleton("AF"))

plotPedList(list(H1, H2),
hatched = c("M", "C", "AF"),
titles = c("H1", "H2"))
```

Consider only the first marker. This marker has at least two alleles; A and B. The allele frequencies for these alleles are p_A and p_B, respectively. Gene mutations are not considered a possibility in this example.

The likelihood ratio when the genotypes are $g_{AF} = A/A$, $g_M = B/B$ and $g_C = A/B$ for AF, M and C respectively is

$$\frac{P(data \mid H_1)}{P(data \mid H_2)} = \frac{P(g_{AF}, g_M, g_C \mid H_1)}{P(g_{AF}, g_M, g_C \mid H_2)} = \frac{P(g_c \mid g_{AF}, g_M, H_1)P(g_{AF}, g_M \mid H_1)}{P(g_c \mid g_{AF}, g_M, H_2)P(g_{AF}, g_M \mid H_2)} \tag{22}$$

$$= \frac{P(g_c \mid g_{AF}, g_M, H_1)}{P(g_c \mid g_{AF,} g_M, H_2)} = \frac{1}{p_A} \tag{23}$$

When AF and M have genotypes A/A and B/B, respectively, and AF and M

are the biological parents of C, then the probability of C having genotype A/B is 1, or 100%. This is because C must have inherited allele A from AF and allele B from M, if gene mutations are assumed impossible. If M has genotype B/B and is the biological mother of C, and no assumptions about the father are made, the probability that C has genotype A/B is p_A. C must have inherited the gene of allele B from M with probability 1. The probability that the paternal gene is of allele A is p_A.

Equation (23) shows that likelihood ratio is the inverse of p_A. Thus, the LR increases as p_A decreases. This is because the allele is less likely to come from a random person if the frequency of allele A in the population, which p_A denotes, is very small.

3.2 The number of assignments with age restrictions

Here follows an explanation of the number of hypotheses which are valid for the DVI case explained in Section 2.6.6, with two additional restrictions added. The hypothesis list for the case is generated by the code in appendix A.2. The list of restrictions is repeated here, with the additional restrictions added at the bottom.

- 5 victims given: $V1$,$V2$,$V3$,$V4$,$V5$
- 5 missing people given: $M1$,$M2$,$M3$,$M4$,$M5$
- Everyone is of the same sex, hence no victim-missing pairs are excluded based on that feature.
- Among ages of the missing people it is known that $M4$ is older than $M1$, that is, $M4 > M1$.
- Among ages of the victims it is known that $V1 > V2 > V3$, $V1 > V4$ and $V5 > V3$.
- Hypotheses may not contain any unmatched victims. The possibility of unmatched victims is omitted to reduce the number of hypotheses which

will be generated.

- It is known that all victim-missing pairs are possible, except that $V3$ can't be $M2$ or $M3$.

Under these restrictions, the number of possible hypotheses is 50. To arrive at this conclusion we start by finding out how many different ways the 5 missing people can be arranged. The answer to this is 120, because $5 * 4 * 3 * 2 * 1 = 5! = 120$. Two fifths of these 120 hypotheses are removed because of restrictions on $V3$ stating it cannot match with $M2$ or $M3$. That leaves 72, but some of these are invalidated by the age restrictions. These restrictions are shown in Figure 10.

Thus, there are only 50 valid hypotheses. Not more than one out of the five age restrictions on the victims may be broken by a single hypothesis. Thus, there is no overlap between the hypotheses each age restriction invalidates. This is because the one age restriction on missing people, $M4 > M1$, may not be broken more than once in a single hypothesis. If $M4 > M1$ is broken, then the victim assigned $M4$ must be younger than the victim assigned $M1$. These two victims may only correspond to one of the restrictions laid out in Figure 10.

The functions in appendix A.2 will be used to find out what would happen if up to one unmatched victim was allowed per hypothesis. Then, for each of the 50 hypotheses in the previous case, one of the five missing people may be replaced by an empty slot to form a new hypothesis. Thus one may effectively form five new hypotheses from each previous hypothesis. Because all the valid hypotheses under the stricter restrictions are still valid now, this should leave $50 * 6 = 300$ possible hypotheses.

However, the program output 384 hypotheses instead. The actual number is higher than 300 because some of the hypotheses that were removed in the case where no empty victims were allowed, become valid if one removes a victim-missing pair.

List of age restrictions

- $V1 > V2$ invalidates 2 of the remaining hypotheses
- $V1 > V3$ invalidates 6 of the remaining hypotheses
- $V2 > V3$ invalidates 6 of the remaining hypotheses
- $V1 > V4$ invalidates 2 of the remaining hypotheses
- $V5 > V3$ invalidates 6 of the remaining hypotheses

Figure 10: List of age restrictions. To only have two victims in each restriction, $V1 > V2 > V3$ is split into three parts; $V1 > V2$, $V1 > V3$ and $V2 > V3$.

3.3 Calculations for motivational example

We explain the computational details for the example in Figure 1. It shows two male victims and a reference family with two missing brothers and a genotyped mother. Obviously, the second solution in Table 1 cannot be distinguished from the first. We also see from Table 1 that the top two solutions have posteriors 0.43. The number of assignments follows from equation (20):

$$\sum_{k=0}^{\min(s,m)} \binom{s}{k}\binom{m}{k}k! = \binom{2}{0}\binom{2}{0}0! + \binom{2}{1}\binom{2}{1}1! + \binom{2}{2}\binom{2}{2}2! = 1 + 4 + 2 = 7.$$

There are four alleles, denoted 1, 2, 3, 4, with frequencies $p_1 = 0.33$, $p_2 = 0.01$, $p_3 = 0.33$, $p_4 = 0.33$. The likelihood of the null assignment, corresponding to no identifications, the null solution, is obtained my multiplying the genotype probabilities of all genotyped individuals, i.e.,

$$L_7 = 2p_1p_2 \cdot 2p_2p_3 \cdot 2p_1p_2 = 2.87496 \times 10^{-07}.$$

Hence the log likelihood is $l_f = \log(2.87496 \times 10^{-7}) = -15.062057$ as shown in Table 1. The LR comparing the assignment $(V_1 = M1, V_2 = M_2)$ to the null solution is

$$LR_1 = \frac{\exp(-9.287599)}{\exp(-15.062057)} = 321.97.$$

```
library(dvir, quietly =T)
pm = list(singleton("V1"), singleton("V2"))
p = c(0.33, 0.01, 0.33, 0.33)
m = marker(pm[[1]], alleles = 1:4, V1 = 1:2, afreq = p,
           name = "M")
pm[[1]] = setMarkers(pm[[1]], m)
m = marker(pm[[2]], alleles = 1:4, V2 = 2:3, afreq = p, name = "M")
pm[[2]] = setMarkers(pm[[2]], m)
missing = c("M1", "M2")
am = nuclearPed(children = missing)
m = marker(am, alleles = 1:4, "2" = 1:2, name = "M", afreq = p)
am = setMarkers(am, m)
ex1 = dviData(pm = pm, am = am, missing = missing)
res = jointDVI(ex1, verbose = F)
```

Figure 11: R code for Table 1 explained in Section 3.3 .

The prior is $1/7$ for each assignment, a flat prior. Hence, the posterior for the assignment $(V_1 = M1, V_2 = M_2)$ becomes, according to equation (1)

$$\frac{LR_1}{LR_1 + \cdots + LR_6 + 1} = 0.43.$$

Figure 11 shows the R-code used to generate the table.

3.4 Resolving symmetry by restricting hypotheses

This example continues on the previous one. Assume that it is known that $V1$ is older than $V2$ and that $M1$ is older than $M2$. Then the symmetry problem is resolved and the results are summarised in Table 4. In such a case, $V1$ cannot be $M2$ if $V2$ is $M1$. A hypothesis where $V1$ is $M2$ and $V2$ is $M1$ implies that $V1$ is younger than $V2$ because $M2$ is younger than $M1$. However, this contradicts the information that $V1$ is older than $V2$.

	V1	V2	loglik	LR	posterior
1	M1	M2	-9.2876	321.9697	0.8615
3	M1	*	-11.8133	25.7576	0.0689
4	M2	*	-11.8133	25.7576	0.0689
5	*	M1	-11.8432	25.0000	0.0669
6	*	M2	-11.8432	25.0000	0.0669
7	*	*	-15.0621	1.0000	0.0027

Table 4: Age information is used to a priori exclude the assignment where both $V1 = M2$ and $V2 = M1$.

.

3.5 Resolving symmetry by typing more references

The two leftmost panels in Figure 12 are similar to Figure 1. It shows two male victims and a reference family with two missing brothers and a genotyped mother R1. In this example we will consider three alternative pedigrees of relatives to identify $V1$ and $V2$, denoted $AM1$, $AM2$ and $AM3$ in Figure 12. The simulation involved the 13 CODIS markers, allele frequencies were taken from the database NorwegianFrequencies in the R library forrel [8].

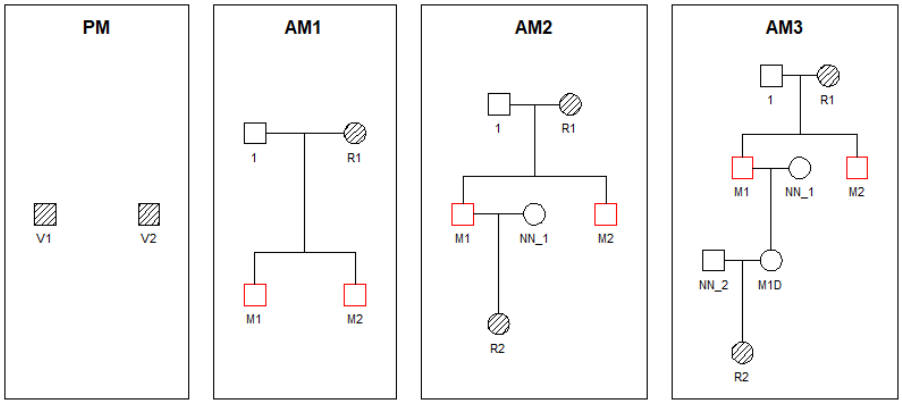


Figure 12: Two missing males are shown to the left followed by three versions of AM data.

3.5.1 Case AM1

The two top solutions in this are ($V1 = M1$, $V2 = M2$) and ($V2 = M1$, $V1 = M2$). They both have posteriors close to 0.5 as shown in Table 5. With the mother R1 being the only reference, no information which may distinguish the brothers from one another is provided. The brothers both have the same familial relationship with the mother.

	V1	V2	loglik	LR	posterior
1	M1	M2	-100.01	4.94e+09	0.50
2	M2	M1	-100.01	4.94e+09	0.50
3	M1	*	-108.17	1.42e+06	0.00
4	M2	*	-108.17	1.42e+06	0.00
5	*	M1	-115.12	1.36e+03	0.00
6	*	M2	-115.12	1.36e+03	0.00
7	*	*	-122.33	1.00e+00	0.00

Table 5: CASE $AM1$: The top two solutions are symmetric.

3.5.2 Case AM2

In this case a daughter of $M1$ is added to the pedigree, see Figure 12. This daughter will be genotyped for the same markers that the two victims were genotyped for. By typing a relative closer to $M1$ than $M2$, one will have a chance of distinguishing these two missing people. The correct underlying hypothesis will likely be identified when a child of a brother can be identified, as this child in expectation will have half their fathers DNA IBD and only a quarter of their uncles DNA IBD. "IBD" means "identical by descent", and DNA identical by descent is DNA which is inherited from a shared ancestor. IBD genes are different from IBS ("Identical by state") genes in that IBS genes may not be traced back to a shared common ancestor from recent time. In this case the posterior probability for hypotheses where $M1$ is $V2$ are not included in the table, because these cases are impossible according to the generated DNA data.

If $M1$ was $V2$, then one of the alleles $V2$ has for each gene must have been inherited by $R2$. The generated DNA data contained genes for which $V2$ and $R2$ did not share a single allele, hence if mutations are assumed impossible, $M1$ being $V2$ causes a contradiction. The expected findings are confirmed in Table 6.

	V1	V2	loglik	LR	posterior
1	M1	M2	-112.86	9.22e+14	1.00
2	M1	*	-126.12	1.61e+09	0.00
3	M2	*	-129.46	5.72e+07	0.00
4	*	M2	-136.94	3.21e+04	0.00
5	*	*	-147.32	1.00e+00	0.00

Table 6: CASE $AM2$: The two solutions given the same probability in Case $AM1$ are now distinguished.

3.5.3 Case AM3

In this last example case, $M1$ has a granddaughter who was genotyped for the markers, though his daughter was not. This approach is less likely to separate $M1$ from $M2$ than when the daughter of $M1$ was genotyped in case $AM2$. This illustrates how a more distant relative implies weaker evidence. The more distant a relative is, the less genes the relative will share with the missing person, hence kinship is harder to prove. The correct solution now has posterior probability 0.911 as shown in Table 7.

All of these three cases were simulated 10000 times in R. These results show the calculated posterior probability of the correct hypothesis. The results can be seen in Figure 13. Here, case $AM1$ is on the left, case $AM2$ is in the middle and case $AM3$ is on the right. In the first case, the brothers will be inseparable every time. This is because DNA information from the mother may only be used to eliminate hypotheses where one or both victims remain unidentified. This information may not be used to separate the hypothesis stating that $V1 = M1$

	V1	V2	loglik	LR	posterior
1	M1	M2	-123.56	2.81e+14	0.91
2	M2	M1	-125.88	2.76e+13	0.09
3	M1	*	-143.23	8.07e+05	0.00
4	M2	*	-143.91	4.08e+05	0.00
5	*	M2	-144.07	3.48e+05	0.00
6	*	M1	-145.56	7.90e+04	0.00
7	*	*	-156.83	1.00e+00	0.00

Table 7: CASE $AM3$: The top two solutions are now distinguished, but not as clearly as in Table 6.

and $V2 = M2$ from the hypothesis stating $V1 = M2$ and $V2 = M1$. In the case $AM2$, the brothers will be separated most of the time. The posterior probability for the correct hypothesis was higher than 0.95 in nearly all the simulations. Case $AM3$ is interesting. Here, conviction was higher than in case $AM1$, but still in a lot of cases not so high that it could be used as undisputable proof. Access to more forensic data would be helpful here.

Table 8 contains information on how many simulations resulted in posterior probabilities above certain thresholds. It is clear from this table that in cases of the form $AM2$ the genetic evidence is quite convincing. The results show an estimated 96% chance of genetic data providing a posterior probability above 0.99 for the correct underlying hypothesis. It is also clear that in cases of the form $AM1$ genetic data will never be able to separate the brothers, as the posterior probability never goes above 0.5 for the correct hypothesis. The simulation data used to make this table is the same data used to generate the boxplots in Figure 13.

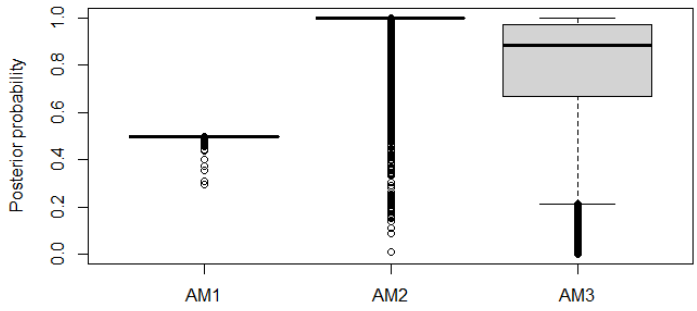


Figure 13: The cases $AM1$, $AM2$ and $AM3$ were simulated 10000 times each. There are two genotyped male victims, $V1$ and $V2$, and two missing brothers, $M1$ and $M2$. The mother of the missing brothers is genotyped in all three cases. In $AM2$ the daughter of $M1$ is also genotyped. In $AM3$ the granddaughter of $M1$ is genotyped instead of the daughter.

3.6 LR for discrete data

A missing person M_i has a property y_i, which describes a feature or attribute which this individual may or may not have. If the individual has this feature, $y_i = 1$, otherwise $y_i = 0$.

The case involves two brothers, $M1$ and $M2$, who have gone missing. Only one victim has been found, $V1$. $V1$ has an attribute, $x_1 = 1$. Among the brothers only $M1$ has the attribute. In other words; $y_1 = 1$ and $y_2 = 0$.

The three DVI hypotheses in this case are;

- $H1 : V1 = *$
- $H2 : V1 = M1$
- $H3 : V1 = M2$

A flat prior is assumed, so all hypotheses are assigned a probability of $1/3$ before data is looked at. By knowing that $M1$ has a feature which $M2$ does

Threshold	AM1	AM2	AM3
> 0.499	9809	9921	8473
> 0.5	0	9918	8468
> 0.75	0	9831	6843
> 0.90	0	9730	4706
> 0.95	0	9679	3403
> 0.975	0	9642	2330
> 0.99	0	9616	1289
> 0.999	0	9463	231
> 0.9999	0	9042	30
> 0.99999	0	8022	5
> 0.999999	0	6411	1

Table 8: A table containing results from simulated DVI cases of the types $AM1$, $AM2$ and $AM3$. The numbers tell how many out of 10000 simulations resulted in a posterior probability above various thresholds.

not, one may conclude who the victim $V1$ really was; $M1$, $M2$, or neither.

$$P(x_i \neq y_j \mid V_i = M_j) = \mu$$

$$P(x_i = 1) = P(y_j = 1) = \alpha$$

These are the definitions of μ and α for an observed binary feature in this DVI case. Here, the probability of a feature being observed in a victim is the same as the frequency of the feature within a population. If only $P(y_j = 1) = \alpha$ is assumed, the probability $P(x_i = 1)$ may be expressed as a function of μ and α.

$$P(x_i \neq y_j \mid V_i = M_j) = \mu$$
$$P(y_j = 1) = \alpha$$
$$P(x_i = 1) = P(y_i = 1) * (1 - \mu) + P(y_i = 0) * \mu$$
$$P(x_i = 1) = \alpha * (1 - \mu) + (1 - \alpha) * \mu$$
$$P(x_i = 1) = \mu + \alpha - 2\alpha\mu$$

For $P(x_i = 1) = \alpha$, the probability of a victim being observed with a feature, to be the same as $P(y_j = 1)$, then μ has to be the same as $2\alpha\mu$, which means either $\mu = 0$ or $\alpha = 1/2$ (or both).

To put this in relation to stationary matrices, the stationary matrix for two states of a feature is as in equation (12) in Section 2.4.

The model (M, p) where

$$M = \begin{pmatrix} P(x_i = 1 \mid y_j = 1, V_i = M_j) & P(x_i = 0 \mid y_j = 1, V_i = M_j) \\ P(x_i = 1 \mid y_j = 0, V_i = M_j) & P(x_i = 0 \mid y_j = 0, V_i = M_j) \end{pmatrix} \tag{24}$$

can be written as

$$M = \begin{pmatrix} 1 - \mu & \mu \\ \frac{p_1}{1-p_1}\mu & 1 - \frac{p_1}{1-p_1}\mu \end{pmatrix} \tag{25}$$

is well defined and stationary if $0 \leq 1 - \frac{p_1}{1-p_1}\mu \leq 1$.

Here, p_1 is the probability for feature 1, i.e. $p_1 = \alpha$. The stationary model allows for other values of α by allowing the two features to have different probability of being misdiagnosed. The probability $P(x_i \neq y_j \mid V_i = M_j)$ is different for $y_j = 1$ and $y_j = 0$ in this case. If $P(x_i \neq y_j \mid V_i = M_j, y_j = 1) = \mu$, then $P(x_i \neq y_j \mid V_i = M_j, y_j = 0) = \frac{\alpha}{1-\alpha}\mu$.

3.6.1 Trying to solve the motivational example

An attempt was made at solving the forensic case in Section 1.1 using genetic evidence. This did not work out, but depending on the situation, the case might be solvable if feature data is provided instead.

This case involved two victims, $V1$ and $V2$, and two missing people, $M1$ and $M2$. If there is one feature recorded for each victim and each missing person, then this feature could be used to draw conclusions within the forensic investigation.

Assume that the hypothesis which states that $V1 = M1$ and $V2 = M2$ is the true underlying hypothesis. If $M1$ and $M2$ have features 1 and 2, respectively, and $V1$ and $V2$ were correctly diagnosed with these features, then the likelihood ratios for the seven possible hypotheses will be those in Table 9. Here the parameters μ_{12} and μ_{21} are introduced. These are the probabilities of feature 1 being misdiagnosed as feature 2 and feature 2 being misdiagnosed as feature 1, respectively. The old parameters μ_1 and μ_2 have the same meaning as before, and are related to the new parameters as they are the sum of the probabilities of all possible misdiagnoses for one feature. For example, if there are three features 1, 2 and 3, then $\mu_1 = \mu_{12} + \mu_{13}$.

	V1	V2	LR
1	M1	M2	$(1\text{-}\mu_1)(1\text{-}\mu_2)/p_1p_2$
2	M2	M1	$\mu_{12}\mu_{21}/p_1p_2$
3	M1	*	$(1\text{-}\mu_1)/p_1$
4	M2	*	μ_{21}/p_1
5	*	M1	μ_{12}/p_2
6	*	M2	$(1\text{-}\mu_2)/p_2$
7	*	*	1

Table 9: Likelihood ratios for the possible hypotheses if $V1$ and $M1$ have been assigned feature 1 and $V2$ and $M2$ have been assigned feature 2.

For small values of p_1, p_2, μ_1 and μ_2 the first hypothesis will correctly be chosen as the most likely when the features are correctly identified. However, in the case of one misdiagnosis (i.e. $V1$ was instead assigned feature 3) the likelihood ratios will be as in Table 10.

If a flat prior is used, and p_3 is greater than μ_{13}, then hypothesis #6 will be

	V1	V2	LR
1	M1	M2	$\mu_{13}(1\text{-}\mu_2)/p_2p_3$
2	M2	M1	$\mu_{12}\mu_{23}/p_2p_3$
3	M1	*	μ_{13}/p_3
4	M2	*	μ_{23}/p_3
5	*	M1	μ_{12}/p_2
6	*	M2	$(1\text{-}\mu_2)/p_2$
7	*	*	1

Table 10: Likelihood ratios for the possible hypotheses if $V1$ was assigned feature 3, $M1$ was assigned feature 1 and both $V2$ and $M2$ were assigned feature 2.

selected over hypothesis #1. For the record, p_3 should be greater than μ_{13} in the large majority of probability models. $V1$ would then remain unidentified.

With DNA data added to the analysis, that could have assisted in identifying the two victims, **if they were unrelated**. However, in this DVI case it would not help because the brothers may not be separated with DNA data from their relatives. DNA data from close relatives like their parents could potentially confirm that the victims were the missing brothers, but it could not assist in telling them apart. For that purpose one would have needed DNA data from a descendant of one of the brothers.

3.7 More likelihood ratios when using the stationary model

The parameters μ and $p_1 = \alpha$ are here as defined in the stationary model in equation (25) in Section 3.6. The two features are denoted as 1 and 0. This DVI case involves one victim $V1$, and two missing people $M1$ and $M2$. $V1$ is reported to have feature 1. $M1$ is known to have feature 1, and $M2$ is known to have feature 0. The hypotheses are defined in the following list.

- $H1$: $V1$ is neither
- $H2$: $V1$ is $M1$

- $H3 : V1$ is $M2$

The conditional probabilities of the observed data are calculated below.

$$
\begin{aligned}
P(data|H1) &= P(X_1 = 1 \mid Y_1 = 1, Y_2 = 0, H1) \times ... \\
&\qquad ... \times P(Y_1 = 1 \mid H1) \times P(Y_2 = 0 \mid H1) \\
P(data|H1) &= P(X_1 = 1)P(Y_1 = 1)P(Y_2 = 0) \\
P(data|H1) &= \alpha^2(1 - \alpha)
\end{aligned}
$$

$$
\begin{aligned}
P(data|H2) &= P(X_1 = 1 \mid Y_1 = 1, Y_2 = 0, H2) \times ... \\
&\qquad ... \times P(Y_1 = 1 \mid H2) \times P(Y_2 = 0 \mid H2) \\
P(data|H2) &= P(X_1 = 1 \mid Y_1 = 1, V_1 = M_1)P(Y_1 = 1)P(Y_2 = 0) \\
P(data|H2) &= (1 - \mu)\alpha(1 - \alpha)
\end{aligned}
$$

$$
\begin{aligned}
P(data|H3) &= P(X_1 = 1 \mid Y_1 = 1, Y_2 = 0, H3) \times ... \\
&\qquad ... \times P(Y_1 = 1 \mid H3) \times P(Y_2 = 0 \mid H3) \\
P(data|H3) &= P(X_1 = 1 \mid Y_2 = 0, V_1 = M_2)P(Y_1 = 1)P(Y_2 = 0) \\
P(data|H3) &= \frac{\alpha}{1 - \alpha}\mu\alpha(1 - \alpha) = \mu\alpha^2
\end{aligned}
$$

From this the likelihood ratio for each hypothesis follows.

$LR(H1) = \frac{P(data|H1)}{P(data|H1)} = 1$

$LR(H2) = \frac{P(data|H2)}{P(data|H1)} = \frac{1-\mu}{\alpha}$

$LR(H3) = \frac{P(data|H3)}{P(data|H1)} = \frac{\mu}{1-\alpha}$

For small μ, the likelihood ratio will be large for $H2$ and small for $H3$. The smaller α is, the more $H2$ is favoured over $H1$ and $H3$.

Next, likelihoods are calculated for the same case, but with $V1$ being reported with feature 0 instead.

$$P(data|H1) = P(X_1 = 0 \mid Y_1 = 1, Y_2 = 0, H1) \times ...$$
$$... \times P(Y_1 = 1 \mid H1) \times P(Y_2 = 0 \mid H1)$$
$$P(data|H1) = P(X_1 = 0)P(Y_1 = 1)P(Y_2 = 0)$$
$$P(data|H1) = \alpha(1-\alpha)^2$$

$$P(data|H2) = P(X_1 = 0 \mid Y_1 = 1, Y_2 = 0, H2) \times ...$$
$$... \times P(Y_1 = 1 \mid H2) \times P(Y_2 = 0 \mid H2)$$
$$P(data|H2) = P(X_1 = 0 \mid Y_1 = 1, V_1 = M_1)P(Y_1 = 1)P(Y_2 = 0)$$
$$P(data|H2) = \mu\alpha(1-\alpha)$$

$$P(data|H3) = P(X_1 = 0 \mid Y_1 = 1, Y_2 = 0, H3) \times ...$$
$$... \times P(Y_1 = 1 \mid H3) \times P(Y_2 = 0 \mid H3)$$
$$P(data|H3) = P(X_1 = 0 \mid Y_2 = 0, V_1 = M_2)P(Y_1 = 1)P(Y_2 = 0)$$
$$P(data|H3) = (1 - \frac{\alpha}{1-\alpha}\mu)\alpha(1-\alpha) = \alpha - \alpha^2(1+\mu)$$

$$LR(H1) = \frac{P(data|H1)}{P(data|H1)} = 1$$
$$LR(H2) = \frac{P(data|H2)}{P(data|H1)} = \frac{\mu}{1-\alpha}$$
$$LR(H3) = \frac{P(data|H3)}{P(data|H1)} = \frac{1-\alpha(1+\mu)}{(1-\alpha)^2}$$

These latter likelihood ratios are very similar to those in the previous example, but with $LR(H2)$ and $LR(H3)$ swapped. This makes sense because the victim $V1$ was now reported with the same feature as $M2$ instead of $M1$. Therefore it is only expected that this time $H3$ is considered more plausible than $H2$ for small μ. If parameters μ and α in the LR expressions in the previous case are substituted with $\frac{\alpha}{1-\alpha}\mu$ and $1-\alpha$, respectively, then one will arrive at the LR expressions for this case. Note that $\frac{\mu}{1-\alpha}$ still is $\frac{\mu}{1-\alpha}$ after this substitution.

3.8 Return to earlier example

This will be another throwback to the example in Section 1.1. Can this case be solved with the use of non-genetic data? The results in this section may be seen as an extension to those in Section 3.6.1.

The DVI case involves two victims and two missing people. The missing people are brothers. One brother has a non-genetic binary feature which the other brother does not have. We denote these as $M1$ has feature 1 and $M2$ has feature 2. The likelihood ratios for the possible hypotheses are in Table 11.

With $p_1 = p_2 = 0.5$ and $\mu_1 = \mu_2 = 0.05$, the likelihood ratios for the various hypotheses are those noted in Table 12. The stationary requirement for models with binary features is fulfilled as $\mu_2 = \frac{p_1}{1-p_1}\mu_1$.

	V1	V2	$x_1 = 1$,$x_2 = 2$	$x_1 = 1$,$x_2 = 1$	$x_1 = 2$,$x_2 = 1$	$x_1 = 2$,$x_2 = 2$
1	M1	M2	$\frac{(1-\mu_1)(1-\mu_2)}{p_1p_2}$	$\frac{(1-\mu_1)\mu_{21}}{p_1p_1}$	$\frac{\mu_{12}\mu_{21}}{p_2p_1}$	$\frac{\mu_{12}(1-\mu_2)}{p_2p_2}$
2	M2	M1	$\frac{\mu_{21}\mu_{12}}{p_1p_2}$	$\frac{\mu_{21}(1-\mu_1)}{p_1p_1}$	$\frac{(1-\mu_2)(1-\mu_1)}{p_2p_1}$	$\frac{(1-\mu_2)\mu_{12}}{p_2p_2}$
3	M1	*	(1-μ_1)/p_1	(1-μ_1)/p_1	μ_{12}/p_2	μ_{12}/p_2
4	M2	*	μ_{21}/p_1	μ_{21}/p_1	(1-μ_2)/p_2	(1-μ_2)/p_2
5	*	M1	μ_{12}/p_2	(1-μ_1)/p_1	(1-μ_1)/p_1	μ_{12}/p_2
6	*	M2	(1-μ_2)/p_2	μ_{21}/p_1	μ_{21}/p_1	(1-μ_2)/p_2
7	*	*	1	1	1	1

Table 11: Likelihood ratios for the seven possible hypotheses.

From Table 12, one can see that in the first column, hypothesis #1 is correctly assigned the highest likelihood ratio. It is also clear that the likelihood ratios for the hypotheses #3 and #6 are quite large. With a flat prior, the posterior probability for hypothesis #1 is no higher than 0.419. From this result we conclude that one fairly common feature is not highly convincing evidence in itself.

Table 13 combines the likelihood ratios from Table 1 in Section 1.1 with those in Table 12 for $x_1 = 1$, $x_2 = 2$. This is assumed to be mathematically possible due to equation (19) in Section 2.5. This equation states that likelihood

	V1	V2	$x_1 = 1$,$x_2 = 2$	$x_1 = 1$,$x_2 = 1$	$x_1 = 2$,$x_2 = 1$	$x_1 = 2$,$x_2 = 2$
1	M1	M2	3.61	0.19	0.01	0.19
2	M2	M1	0.01	0.19	3.61	0.19
3	M1	*	1.9	1.9	0.1	0.1
4	M2	*	0.1	0.1	1.9	1.9
5	*	M1	0.1	1.9	1.9	0.1
6	*	M2	1.9	0.1	0.1	1.9
7	*	*	1	1	1	1

Table 12: Likelihood ratios for the hypotheses in Table 11, with $p_1 = p_2 = 0.5$ and $\mu_1 = \mu_2 = \mu_{12} = \mu_{21} = 0.05$

ratios achieved through DNA analysis may be combined with likelihood ratios achieved through other forensic analysis through simple multiplication. This does of course assume that the data used in the analysis not involving DNA is statistically independent from the DNA data. This assumption is reasonable if the feature is non-genetic, as was assumed in this case. The posterior probability for the correct hypothesis is here 0.917, a clear improvement over both the posterior obtained by only using DNA data (0.431) and the posterior obtained using only feature data (0.419).

	V1	V2	LR^{NG}	LR^{G}	LR	*posterior*
1	M1	M2	3.61	321.9697	1162.3	0.91662
2	M2	M1	0.01	321.9697	3.2197	0.00254
3	M1	*	1.9	25.7576	48.939	0.03859
4	M2	*	0.1	25.7576	2.5758	0.00203
5	*	M1	0.1	25.0000	2.5000	0.00197
6	*	M2	1.9	25.0000	47.500	0.03746
7	*	*	1	1.0000	1.0000	0.00079

Table 13: The results achieved when combining the LR-s in Table 12 for $x_1 = 1$ and $x_2 = 2$ with the LR-s in Table 1 in Section 1.1.

4 Discussion

In this thesis we have presented forensic kinship cases and disaster victim identification problems.

Because of the strength of DNA analysis, other forensic data will usually be redundant, though with notable exceptions. There may be cases where DNA evidence is close to useless due to degradation. In such cases the amount of information researchers are able to collect from DNA readings may be very small. Another situation where DNA evidence may not prove a hypothesis is a case where one is unable to separate two close family members, as demonstrated with cases involving two missing brothers. It was shown that in such a case acquiring other forensic data than DNA data could resolve this issue. This was possible because it was argued both in this thesis and in [6] that likelihood ratios calculated with different kinds of forensic data could be combined through multiplication, though this assumes independence between the different kinds of data.

The main novelty of this thesis is the introduction of models for non-genetic data. Such methods were also addressed in [6]. However, unlike this thesis, the mentioned paper did not assume stationary probability models and did not discuss the problem of non-stationarity. In other words, the distribution

of a discrete characteristic in [6] could differ in the AM and PM samples, i.e. the probabilities of observed features in missing people could be different from those of victims. This may appear unreasonable as normally there is no reason to believe that a feature like a tattoo should appear with different frequencies in victims and missing people. Assuming a stationary model as the underlying probability model avoids this issue.

Whether someone considers the assumption of a stationary model reasonable should depend on what other assumptions that someone wants to make. If a researcher assumes that the frequencies of features in AM and PM samples are the same, and also assumes that misclassification probabilities may be modelled by a probability matrix, we recommend that the researcher assumes that the probability matrix is stationary. This is because if the probability matrix is not assumed stationary, then mathematical rules will be broken by the assumption that observed feature frequencies are the same for victims and missing people.

For DNA based problems there are lots of examples and data available. Also, databases of allele frequencies like the one in Table 2, are published. The statistical models for likelihood calculations have a sound biological basis. For non-genetic data, much less is done and there is not much available of relevance for DVI problems. Hence the specification of the probability distribution for a discrete characteristic becomes more speculative.

Posterior probabilities are often reported for DVI problems as exemplified in Table 1. While posteriors are more easily interpreted than LR-s, they are based on priors. Such priors may be speculative. However, this problem is not unique to non-genetic data.

During the writing of this thesis, functions for generating DVI hypotheses were created. They include functionality not present in the *dvir* library in R. Old methods were only able to reject hypotheses with impossible victim-missing pairs or overlapping victim-missing pairs. The new functionality also allows for rejection based on age restrictions and allows setting a limit on the number of unassigned victims. More work could be done on this functionality, though one may want to focus on one of the implementations in the appendix and not

both, as they solve very similar problems. Currently, the implementation in appendix A.2 is faster, but the implementation in appendix A.1 is more user friendly.

For other possible future work, it would be interesting to perform more simulations and analyse large, realistic cases using the techniques and models presented in this thesis. More simulations would not assist in proving the models in this thesis reasonable, as the simulated data is based on these models. However, simulations could give an idea on how much non-genetic data would be required for solving forensic cases which could not be solved by genetic data alone, and vice versa. By analysing large cases with these models and techniques one would achieve more insight on how accurate or inaccurate the models really are.

5 References

References

[1] Daniel Kling, Thore Egeland, Andreas Tillmar, and Lourdes Prieto. *Mass Identifications: Statistical Methods in Forensic Genetics*. Academic Press, 2021.

[2] Magnus D Vigeland and Thore Egeland. Joint DNA-based disaster victim identification. *Scientific Reports*, 11(1):13661, 2021.

[3] Thore Egeland, Daniel Kling, and Petter Mostad. *Relationship Inference with Familias and R: Statistical methods in Forensic Genetics*. Academic Press, 2015.

[4] Magnus Dehli Vigeland. *Pedigree analysis in R*. Academic Press, 2021.

[5] Making sense of forensic genetics. https://senseaboutscience.org/wp-content/uploads/2017/01/making-sense-of-forensic-genetics.pdf, 2017. Accessed: 2023-05-28.

[6] Franco Marsico and Inés Caridi. Incorporating non-genetic evidence in large scale missing person searches: A general approach beyond filtering. *Preprint available at https://ssrn.com/abstract=4331033*, 2023.

[7] Deping Meng, Peng Zhou, Min Li, Jie Xu, Linchao Lu, Yilin Guo, Chunjiang Yu, Yuliu Xu, Xiaoqun Xu, Chen Fang, et al. Distinguishing between monozygotic twins' blood samples through immune repertoire sequencing. *Forensic Science International: Genetics*, page 102828, 2023.

[8] Norwegianfrequencies database. `https://rdrr.io/cran/Familias/man/NorwegianFrequencies.html`. Accessed: 2023-06-25.

[9] Balding DJ and Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96:3–12, 1995.

[10] A P Dawid, J Mortera, and V L Pascali. Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing. *Forensic Science International*, 124:55–61, 2001.

[11] Ivar Simonsson and Petter Mostad. Stationary mutation models. *Forensic Science International: Genetics*, 23:217–225, 2016.

[12] T Egeland and MD Vigeland. Properties of mutation models with applications in forensic genetics. Manuscript, 2023.

A Implementation

A.1 A simple solution

The function *expandgridnodup2* which was developed during this thesis is part of the R library *dvicomb*. This library is loaded by typing the following in R:

```
install.packages(https://familias.name/alf/dvicomb_0.1.0.zip)
```

The example based on Figure 14 is reproduced below:

```
# We consider the case shown in the below plot:

plotDVI(example2)

# The below code generates the possible assignments

# accounting for sex and that V2 is older than V1.

# Hence we cannot have V1 = M1 and V2 = M2

library(stringr)

library(dvir)

pm = example2$pm

am = example2$am
```

```
pairss = list(V1 = c("*","M1","M2"),

V2 = c("*", "M1", "M2"),V3 =c("M3","*"))

expand.grid.nodup2(pairss, pm, am, age = "V2>V1")
```

```
V1 V2 V3
1   *  * M3
2   *  *  *
3  M1  * M3
4  M1  *  *
5  M2  * M3
6  M2  *  *
7   * M1 M3
8   * M1  *
9  M2 M1 M3
10 M2 M1  *
11  * M2 M3
12  * M2  *
```

A.2 Main Code

The functions *createrestictionmatrix* and *hypsolverrestrictions* developed during this thesis are included in the following link;

`https://drive.google.com/file/d/1V97JqWq6fhZQXGgarM4rG2U9IZfWiPRf/view?usp=sharing`

An example which utilizes these functions is found with this link;

`https://drive.google.com/file/d/1BC2-nAu4lsQv_rNNKF8sa01y1L3uH2b9/`

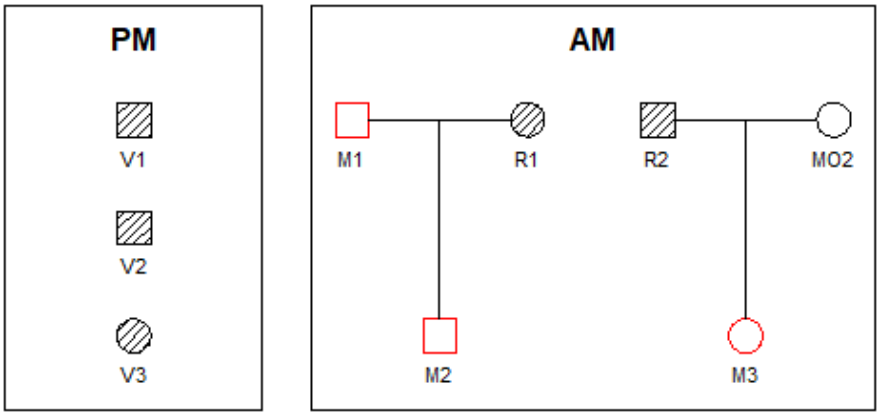
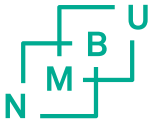


Figure 14: Pedigree plots complementing the example R code in Section A.1 .

`view?usp=sharing`



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway