Aalborg University

Department of Mathematical Sciences

MSc Thesis

Amplification of DNA mixtures

Missing data approach

June 2007

Torben Tvedebrink

Department of Mathematical Sciences, Aalborg University,
Fredrik Bajers Vej 7 G, 9220 Aalborg East, Denmark

AALBORG UNIVERSITET

Department of Mathematical Sciences
Fredrik Bajers Vej 7G
9220 Aalborg
Telephone: +45 96 35 88 02
Fax: +45 98 15 81 29
Web: http://www.math.aau.dk

Topic:
DNA mixtures in forensic science

Title:
Amplification behaviour of mixed DNA samples
- Missing data approach

Project period:
February 1st - June 21th, 2007

Author:
Torben Tvedebrink

Supervisor:
Poul Svante Eriksen

Semester:
MAT6, Spring 2007

Number of copies: 13
Report – number of pages: 64
Deadline: 8 June 2006.

Preface

Abstract

This thesis presents a model for the interpretation of results of STR typing of DNA mixtures based on a multivariate normal distribution of peak areas. From previous analyses of controlled experiments with mixed DNA samples, we exploit the linear relationship between peak heights and peak areas, and the linear relations of the means and variances of the measurements. Furthermore the contribution from one individual allele to the mean area of this allele, is assumed proportional to the average of height measurements on alleles where the individual is the only contributor.

For shared alleles in mixed DNA samples, it is only possible to observe the cumulative peak heights and areas. Complying with this latent structure, we use the EM-algorithm to impute the missing variables based on a compound symmetry model. This allows intra- and intersystem correlations on the measurements and does not depend on the alleles of the DNA profiles. Due to factorization of the likelihood and properties of the normal distribution, an ordinary implementation of the EM-algorithm solves the missing data problem.

We estimate the parameters in the model based on a training data set. In order to asses the weight of evidence provided by the model, we use the model with the estimated parameters on STR data from real crime cases with DNA mixtures.

The model work under certain limitations. In the estimation phase we exclude cases with drop-outs. These limitations are important and must be solved before the model can be used for real crime case work and the limitations are therefore subject to further investigation.

Outline

Chapter 1 contains an introduction and description of the problem of DNA STR mixtures. We discuss the problem related to DNA mixtures with a focus on the matters addressed in this thesis, but also a detailed overview of additional aspects is presented.

Chapter 2 gives a summary of the data analysis performed on the preceding semester. The relevant conclusions for the modelling phase are included with plots as supportive argumentation.

Chapter 3 presents the model of this thesis and introduces the notation used in the rest of the report. The assumptions of compound symmetry and discussions of the missing data nature of DNA mixtures are given. In the last section we derive the estimators used in the EM-algorithm.

Chapter 4 is about the EM-algorithm and its use in this present project. The useful properties of the EM-algorithm applied to our problem and a schematic pseudo code for its implementation is provided. Also a description of how to execute the R-scripts for estimation is included.

Chapter 5 presents the parameter estimates from the EM-algorithm from several initial value sets. Simplifications of the model are also discussed in relation to the estimated values. The expected Fisher information is also derived and is used for computing the asymptotic covariance matrix of the parameters.

Chapter 6 analyzes data from some real crime cases. Different test based on Mahalanobis distances is used for assessing the goodness of fit of the model.

Chapter 7 rounds off the thesis with a discussion and conclusion. Furthermore some considerations are made on possible future work with a link to the additional aspects mentioned under Chapter 1 which is not included in this project.

Appendix A contains some more work from the preceding semester. A summary of the biology of DNA and populations genetics is included for quick references to the technical terms used in the body text.

Acknowledgements

I would like to thank my supervisor Poul Svante Eriksen, Department of Mathematical Sciences Aalborg University, who have helped me set up the model framework, debugging my R-implementation of the EM-algorithm and gave me critical comments during the process.

I would also like to thank Professor Niels Morling, Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, for his and the department's commitment and enthusiasm in my project and for providing me with data.

Torben Tvedebrink
Aalborg 2007

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The issue of DNA STR mixtures are of great importance since DNA mixtures arise from various contexts and are complex to interpret. In crime cases there are several examples from which DNA mixtures can occur:

- Rapes with one or more rapists. In these cases the profile of the victim is always available this is useful in determining possible suspects. In gang rapes where there is more than one rapist it is more complex to separate the mixed profiles.
- Burglaries where the burglars leave a stain behind with more than one burglar contributing. This could be a blood stain where they cut themselves on the same object.
- A cigarette butt where more than one person has placed some saliva.

It is only possible to observe the mixed DNA profile and not the relevant single profiles. Using STR DNA there is for each locus a finite number of alleles and therefore it is (almost) unavoidable to have some shared alleles in a mixture. Using the quantitative information from peak heights and peak areas it is possible for trained forensic geneticists to come up with possible combinations. For complicated cases with more than two contributors it are however often difficult to resolve the donor profiles and therefore some more objective methods needs to be developed. In the next section we list some of the related problems to DNA mixtures and DNA evidence in general.

1.1 DNA mixtures

In real crime cases we seldom have access to information on how the mixed stain was provided. Therefore several important informations are attached with uncertainty which increases the complexity. Below we discuss each of the most common problems.

Unknown number of contributors Information on how many individuals have contributed to the sample is important in order to determine the number of possible peaks for each loci. The fewer contributors the less complicated the separation of DNA mixtures gets and thus is the evidence more conclusive.

Degraded DNA When DNA is exposed to heat, direct sunlight, moisture and humid environment, acids and other inhibitors the sequences of deoxyribonucleic acid (DNA) are likely to break into smaller fractions. This can cause the peaks of alleles to amplify less than if the DNA material were kept under optimal conditions. A common assumption is that longer DNA sequences have a higher probability of breakage than shorter sequences. Also degraded DNA can cause entire systems to drop-out under amplification.

Drop-outs As mentioned above drop-outs can be caused by degradation of the DNA but also from malfunctioning machinery. Under the controlled experiments analyzed on the preceding semester the most likely reason for drop-outs were observed to be low DNA concentration. This implies that in mixtures with a major and minor contributor with respect to DNA concentration it is likely that alleles of the minor contributor are subject to drop-outs. When analyzing DNA samples laboratories often set a lower limit on the peak heights. This threshold is however a trade-off between drop-outs and drop-ins as a lower threshold reduces the number of drop-outs but might also introduce more drop-ins and stutters.

Stutters, drop-ins and pull-ups Artificial alleles are often observed in DNA samples even though none of the contributor profiles contain these alleles. This kind of contamination of a DNA profile is known as stutters, drop-ins and pull-ups each having their own meaning. Stutters are a product of the PCR procedure run before amplification of a DNA sample to increase the amount of DNA. A stutter for allele n is a peak observed at the position of allele $n - 1$ but with reduced size. In non-mixture samples the ratio of stutters and the real peaks are 5%-15%. The stutter effect may also be observed for alleles $n - k$ for $k > 1$ but with even lower magnitude. Drop-ins are artificial alleles observed outside the stutter range. DNA material from plastic and other material are likely reasons for drop-ins. An other explanation for some drop-in peaks may be pull-ups which is caused by the spectral overlap of the allelic ladder across dye bands. This is caused by the overlap of the colours used in the fluorescent reaction.

Amplification variation For the model to be useful across laboratories and machines it is important to know the variation induced to these factors. The controlled experiments showed that personal effects should be seen as an important source of variation. This applies also to the different loci and alleles where the amplification behaviour varies with both system and allele within systems.

Null alleles The model introduced in Chapter 3 does not allow the so-called null-alleles. Null-alleles are unobservable alleles and is a competing event with homozygosity. Null-alleles is by definition off-ladder alleles that are undetectable by any kit used in DNA profiling. If for a system a profile only amplify at one allele, this person can either by homozygote or have a null-allele for this system.

Due to both the importance of crime detection of the above mentioned examples and the challenges outlined above, DNA mixtures have received extensive focus from forensic geneticists and statisticians. In the bibliography there is a list of references to some of the most important of the published literature on the subject. The approach discussed in this present thesis where we allow intersystem covariance is however not seen elsewhere.

CHAPTER 2

Summary of MAT5 project

This chapter is a summary of the authors own work (Tvedebrink, 2006) from the preceding semester (MAT5) at Aalborg University. This semester was dedicated to obtain an understanding of relevant issues related to mixtures of STR DNA and an analyzes of a data set from controlled experiments conducted at the Section of Forensic Genetics, University of Copenhagen. For a summary on the biology of DNA we refer to Appendix A on page 57.

The aim of the data analysis was to get insight on the amplification behaviour of mixed DNA samples. The DNA mixtures of this data set where sampled in a controlled environment where the laboratory pursued to keep temperature, humidity, exposure to sunlight and UV-light constant. There were however some variation in the data which could not be explained by known covariates which indicated that this was not possible.

2.1 Data exploration

Included in the analysis were the occurrence of stutters and their relative size to the real peak and the occurrence of drop-ins and drop-outs for different DNA ratios and systems. These two types of contamination play a central role in real world cases as their presence changes the evidence based on a DNA stain related to a crime scene. This implies that the determination of the number of contributors to a stain gets more complicated and also that probabilities involved in the inclusion and exclusion of possible suspects alters. Since the profiles of the contributors to the DNA mixtures in our data set were known these quantities could be estimated. The profiles are given in Table 2.1.

Table 2.1: DNA profiles of the four individuals in the experiment. Allele numbers in *italic* are reported as null alleles.

	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
A	(14;18)	(17;19)	(12;14)	(20;24)	(10;13)	(30.2;32.2)	(13;*13*)	(12;13)	(8;9)	(20;22)
B	(15;16)	(14;16)	(10;12)	(17;25)	(13;16)	(30;*30*)	(13;*13*)	(14;15)	(6;9)	(19;23)
C	(15;16)	(15;17)	(11;*11*)	(19;25)	(8;12)	(29;31)	(15;17)	(13;*13*)	(6;8)	(23;24)
D	(16;19)	(15;17)	(10;12)	(23;25)	(13;*13*)	(28;30)	(12;16)	(13;15)	(6;7)	(20;23)

The findings confirmed the common assumption of increased drop-out frequency as the DNA concentration decreases. We have summarized the drop-out frequencies in Figure 2.1. We see that there is no drop-out when the amount of DNA is above 150 pg and except for two observations this threshold can be lowered to 75 pg.

The main focus of the investigation was to reveal patterns in the data which could be used in the later modeling phase. For each observed peak we have information on the

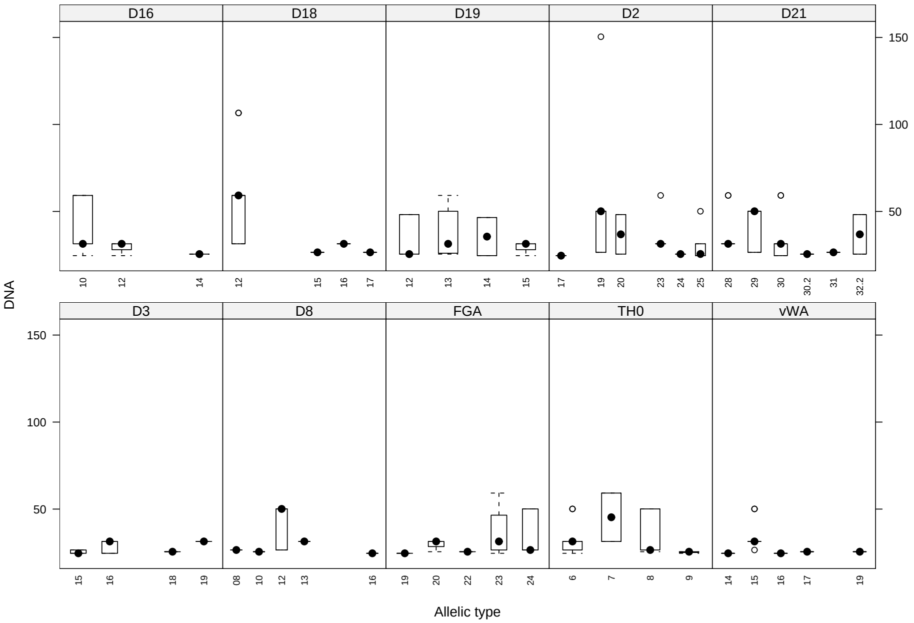

Figure 2.1: Box plot of amount of DNA contributed by the donor when a drop-out is observed. There is stratification on systems and alleles.

STR system, dye band, fragment length in base pairs, allelic number, peak height and peak area as well as the possible donor(s). The latter refers to the possibility of shared alleles between the two donors, and that hidden drop-outs may occur if a stutter of another peak were observed instead of the true peak.

The main conclusions based on the data analysis were that,

- there is a strict linear relation between peak heights and peak areas,
- the amplification properties across dye bands vary,
- the variance of the measurements is proportional to the mean of the measurements,
- there is an approximately linear increase in amplification as a function of the amount of DNA,
- the ratio of the mean peak areas of the two donors are proportional to the ratio of the amount of DNA from the donors.

Below we will justify these conclusions by graphical plots indicating these properties.

The linearity between peak height and area depends on the system as shown in Figure

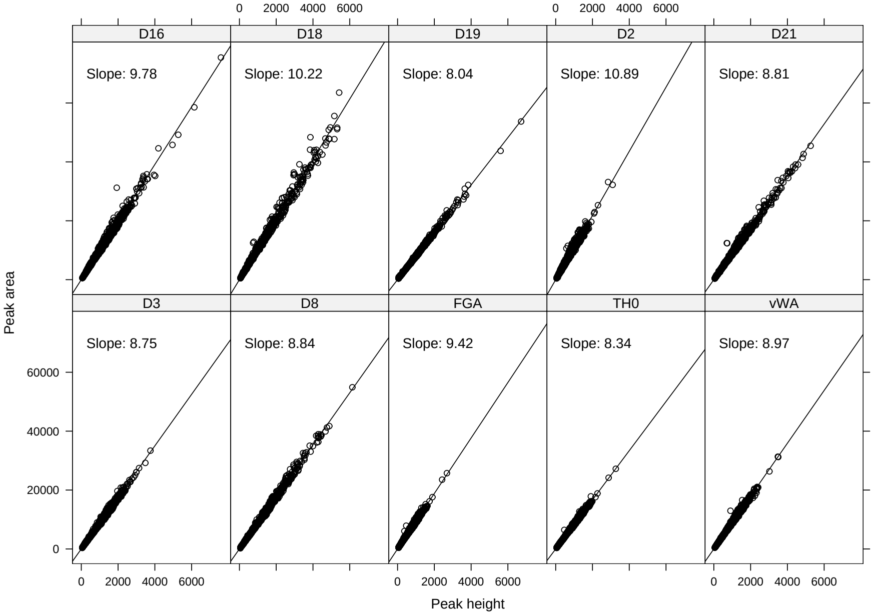

Figure 2.2: Linear relationship between peak area and peak height.

2.2. This indicate that our mean of the peak areas must be parametrized by at least ten parameters, but also that the different alleles within the same STR system behave similarly in terms of this relation.

In Figure 2.3 the box plots show the aggregated sum over each locus in the different cases. The letters below the locus names indicate which dye band the system belongs to. Note the pattern of the yellow band being the less amplified band and also green tends to amplify more than blue.

In Figure 2.4 the numbers reflect the contributors to the observed peak area. “One heterozygote” means that the observed allele can only originate from one person which is heterozygote and for “Two heterozygote” we have both persons are heterozygote and share the observed allele. Similar for the homozygotes, whereas the fifth category is one homozygote together with a heterozygote who share the allele of the homozygote donor. We see that the deviation of the measurements increases with the mean which again increases with the amount of DNA. The approximate linearity mentioned above referrers to the bend-off observed on the curve (second order polynomial fit) superimposed.

The final item refers to the ratio

$$R = \frac{H^{(1)}/H^{(2)}}{DNA_1/DNA_2}$$

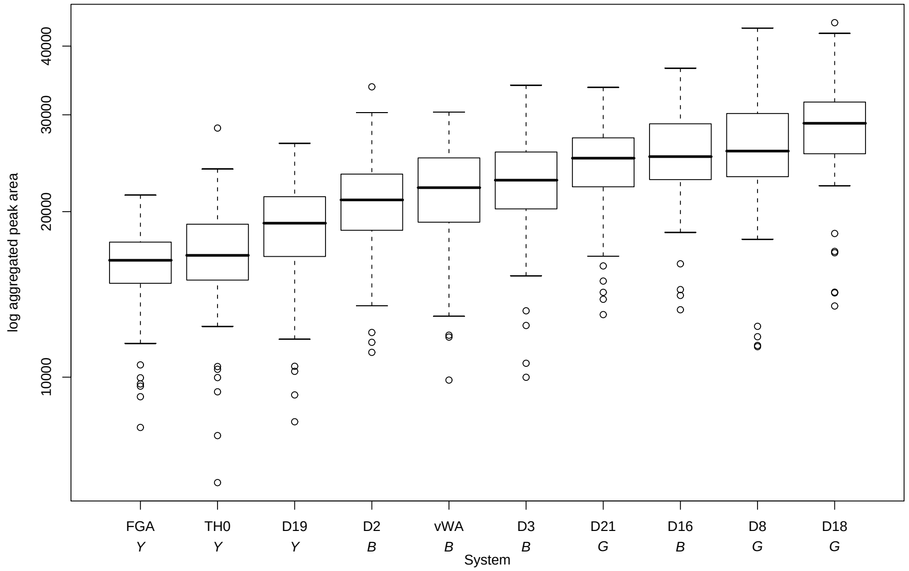

Figure 2.3: Box plot of aggregated peak areas within each locus. Plotted on log scale to reduce the variability. The letters *Y*, *G* and *B* indicate the dye of the STR system.

where $H^{(k)}$ is the mean of all heights where only person k have contributed and DNA_k is the amount of DNA contributed by person k. A mathematical definition of $H^{(k)}$ is given in (3.1). In Figure 2.5 we have plotted $\log(H^{(1)}/H^{(2)})$ against $\log(DNA_2/DNA_1)$ to reduce the variance. We see that except for two observations all the points lie on the identity line. The outliers can be explained by a typing error as commented by the laboratory as person-water mixture (left) and wrongly registered DNA concentrations (right).

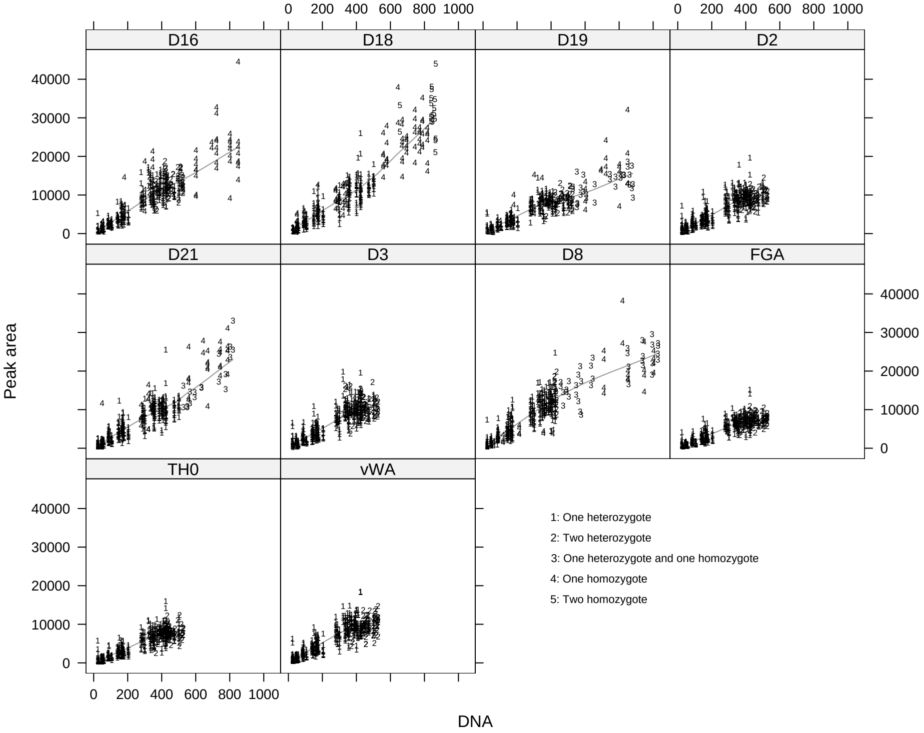

Figure 2.4: Scatter plot of peak area and amount of DNA stratified on STR systems.

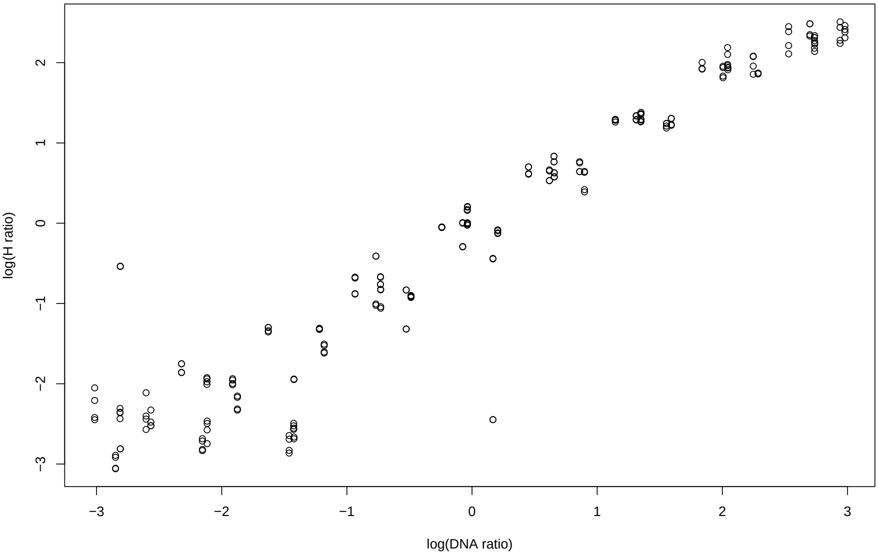

Figure 2.5: Log of DNA ratio against log of mean height ratio.

CHAPTER 3

Missing data model

In the following we assume that a DNA typing kit testing on S systems is used. We denote the set of systems sys, i.e. $|\text{sys}| = S$. This gives reason for the matrix and vector dimensions stated below. In this chapter we exploit the relationships observed in the data analysis to form a model based on a multivariate normal distribution of the unobservable peak areas.

3.1 Model for the unobservable peak areas

Let the two profiles from person 1 and person 2 be denoted as $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. Let $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_S$ be the set of possible alleles. The term "possible" is interpreted as detectable and excludes therefore the so-called null-alleles. The observed alleles are given as $B_{i,s}^{(k)} \in \mathcal{B}_s$ which refers to the ith allele of system s for person k. Hence we have that $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ can be written as

$$\boldsymbol{P}_1 = \left(B_{1,1}^{(1)}, B_{2,1}^{(1)}, B_{1,2}^{(1)}, B_{2,2}^{(1)}, \ldots, B_{1,S}^{(1)}, B_{2,S}^{(1)}\right) \quad \boldsymbol{P}_2 = \left(B_{1,1}^{(2)}, B_{2,1}^{(2)}, B_{1,2}^{(2)}, B_{2,2}^{(2)}, \ldots, B_{1,S}^{(2)}, B_{2,S}^{(2)}\right),$$

and then the mixed sample is given as $\boldsymbol{P}{=}(P_1, \ldots, P_S)$ where $P_s{=}\left(B_{1,s}^{(1)}, B_{2,s}^{(1)}, B_{1,s}^{(2)}, B_{2,s}^{(2)}\right)$ for $s = 1, \ldots, S$.

Defining the area function as $a : \mathcal{B}_s \to \mathbb{R}_+$, then we denote the area $a\left(B_{i,s}^{(k)}\right)$ of $B_{i,s}^{(k)}$ as $A_{i,s}^{(k)}$ and similarly for the peak height function h. Further we define

$$\boldsymbol{A} = a(\boldsymbol{P}) = \left(A_{1,1}^{(1)}, A_{2,1}^{(1)}, A_{1,1}^{(2)}, A_{2,1}^{(2)}, \ldots, A_{1,S}^{(1)}, A_{2,S}^{(1)}, A_{1,S}^{(2)}, A_{2,S}^{(2)}\right)$$

which is called the area vector. We assume that $\boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_S)$ has a $(4S)$-dimensional normal distribution, $\boldsymbol{A} \sim \mathcal{N}_{(4S)}(\boldsymbol{\mu}, D)$ where D is a diagonal matrix. $\boldsymbol{A}$ is however not observable in DNA STR mixtures since when the contributors to a sample share one or several alleles only the cumulative peak areas (and heights) are observable. Information on which alleles that are shared is not available hence $\boldsymbol{A}$ is unobservable by nature. The peak areas in $\boldsymbol{A}$ do however contain relevant information in relation to separate the mixed profiles. This is based on an assumption of similar amplification behaviour across systems for each contributor. That is if a mixture consists of DNA from two individuals we assume that $(A_{1,s}^{(1)} + A_{2,s}^{(1)})/(A_{1,s}^{(2)} + A_{2,s}^{(2)})$ is rather constant across systems.

Let $H^{(k)}$ be the mean of all peak heights where only person k have contributed,

$$H^{(k)} = |\mathcal{I}^{(k)}|^{-1} \sum_{s \in \text{sys}} \sum_{i=1}^{2} h\left(B_{i,s}^{(k)}\right) \mathbb{I}_{\mathcal{I}_s^{(k)}}\left(B_{i,s}^{(k)}\right) \tag{3.1}$$

where $\mathcal{I}_s^{(k)} = \left\{B_{i,s}^{(k)} : B_{i,s}^{(k)} \neq B_{i',s}^{(k')}, \text{ for all } i' \text{ and } k \neq k'\right\}$, $\mathcal{I}^{(k)} = \bigcup_{s\in\text{sys}} \mathcal{I}_s^{(k)}$ and $\mathbb{I}$ is the indicator function. In order to link elements in the area vector $\boldsymbol{A}$ with the correct mean peak height observations, we define $\boldsymbol{H}$=$(\boldsymbol{H}_1, \ldots, \boldsymbol{H}_S)$ with $\boldsymbol{H}_s$=$(H^{(1)}, H^{(1)}, H^{(2)}, H^{(2)})$. The mean $\boldsymbol{\mu}$ is a linear function of the mean peak heights $\boldsymbol{H}$. We have from Figure 2.2 that the parameters differ among systems hence

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_S)^\top, \quad \text{where } \boldsymbol{\mu}_s = \left(\mu_s^{(1)}, \mu_s^{(1)}, \mu_s^{(2)}, \mu_s^{(2)}\right)^\top \text{ for all } s \in \text{sys}.$$

That is for each $\mu_s^{(k)} = H^{(k)}\alpha_s$ with α_s being a common parameter for system s. Figure 2.5 indicates that $H^{(k)}$ is proportional to the DNA concentration of person k. This implies that the mean $\boldsymbol{\mu}$ is modelled proportional to the amount of DNA which is supported by Figure 2.4. Furthermore the mean within a system is the same for the two alleles of person k. Previous studies suggest that the ratio of the two peaks often is above 90% (Applied Biosystems, 2006, p.9-29) for non-mixtures. Since we assume that $\boldsymbol{A}$ behave as a non-mixture sample this applies to observations in $\boldsymbol{A}$ and hence also to its mean.

Also the variance is proportional to the DNA concentration and therefore proportional to H (see Figure 2.4). Incorporating this into the structure on D imply that $D_s^{(k)} = \sigma_s^2 H^{(k)}$. Let $A_{i,s,c}^{(k)}$ be $A_{i,s}^{(k)}$ in case c and similarly for $H_c^{(k)}$, $c = 1, \ldots, C$ with C being the number of cases. Since the elements of $\boldsymbol{A}$ are independent their marginal distributions follow a univariate normal distribution,

$$A_{i,s,c}^{(k)} \sim \mathcal{N}\left(\alpha_s H_c^{(k)}, \sigma_s^2 H_c^{(k)}\right).$$

emphasizing the proportionality of $H^{(k)}$ in both mean and variance and that the distribution does not depend on i, i.e. we have the same distribution of all alleles of system s for person k. Furthermore the parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_S)$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_S^2)$ are common for all cases.

The observed peak area values are denoted $\boldsymbol{M}$ and are determined by a transformation T and an error term ε,

$$\boldsymbol{M} = T\boldsymbol{A} + \varepsilon. \tag{3.2}$$

We partition $\boldsymbol{M}$ as $(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_S)$ where each $\boldsymbol{M}_s$ is the observations for system s and have dimension n_s with $n = \sum_s n_s$. The transformation T is given as an $n \times (4S)$-matrix with elements 0 and 1 based on $\boldsymbol{P}$ and adds together peak areas from the same alleles within each system. T is a diagonal block matrix with diagonal elements T_s,

$$T = \begin{bmatrix} T_1 & O & \ldots & O \\ 0 & T_2 & \ldots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \ldots & T_S \end{bmatrix},$$

where T_s are of dimension $n_s \times 4$. Given that for system s person 1 and person 2 have profiles $\left(B_{1,s}^{(1)}, B_{2,s}^{(1)}\right)$ and $\left(B_{1,s}^{(2)}, B_{2,s}^{(2)}\right)$ respectively where only $B_{1,s}^{(1)} = B_{1,s}^{(2)}$ otherwise different, then T_s will be

$$T_s = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

yielding $T_s \boldsymbol{A}_s = \left(A^{(1)}_{1,s} + A^{(2)}_{1,s}, A^{(1)}_{2,s}, A^{(2)}_{2,s}\right)$. I.e. for system s in this example we would observe three alleles where the first is a sum of two contributing peak areas. Below we give an example of the quantities defined above.

In Table 3.1 the available data from a case is given. The data are from the same data set that was used for the analysis in Chapter 2. The column names $B^{(k)}_{i,s}$ refers to the ith allele for person k for system s where s is determined by the system column.

Table 3.1: Data from a controlled experiment. Here a 8:1 mixture of person D and B.

$B^{(1)}_{1,s}$	$B^{(1)}_{2,s}$	$B^{(2)}_{1,s}$	$B^{(2)}_{2,s}$	System	Allele	Height	Area
10	12	10	12	D16	10	1261	12381
10	12	10	12	D16	12	1249	12475
13	13	12	16	D18	13	3141	32097
13	13	12	16	D18	12	274	2833
13	13	12	16	D18	16	146	1545
14	15	13	15	D19	14	1097	8799
14	15	13	15	D19	15	1045	8334
14	15	13	15	D19	13	222	1795
17	25	23	25	D2	17	929	10089
17	25	23	25	D2	25	889	10031
17	25	23	25	D2	23	125	1354
30	30	28	30	D21	30	2654	23601
30	30	28	30	D21	28	224	2038
15	16	16	19	D3	15	959	8614
15	16	16	19	D3	16	1284	11296
15	16	16	19	D3	19	154	1289
13	16	13	13	D8	13	1722	15242
13	16	13	13	D8	16	1226	10943
19	23	20	23	FGA	19	862	8184
19	23	20	23	FGA	23	656	6280
19	23	20	23	FGA	20	111	1046
6	9	6	7	TH0	6	919	7570
6	9	6	7	TH0	9	865	7300
6	9	6	7	TH0	7	97	780
14	16	15	17	vWA	14	1019	9298
14	16	15	17	vWA	16	1019	9315
14	16	15	17	vWA	15	174	1770
14	16	15	17	vWA	17	103	973

From these information we can compute $H = (H^{(1)}, H^{(2)})$ using the definition from (3.1). Here

$$
\begin{aligned}
H^{(1)} &= \frac{3141 + 1097 + 929 + 959 + 1226 + 862 + 865 + 1019 + 1019}{2 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1} = 1111.7 \\
H^{(2)} &= \frac{274 + 146 + 222 + 125 + 224 + 154 + 111 + 97 + 174 + 103}{1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1} = 163
\end{aligned}
$$

The theoretical mixture ratio of the sample was 8:1 but since the different test persons vary in DNA concentrations the actual ratio was 6.28 based on the DNA concentration measurements. However several studies (e.g. Nielsen et al., 2007) suggests that the methods for determining the amount of DNA are inaccurate and subject to large variability. The ratio $H^{(1)}/H^{(2)}$ yields 6.82 which is close to the ratio based on the DNA concentrations.

The $\boldsymbol{M}$ vector is simply the cell values from the Area column in Table 3.1 and similarly for $\boldsymbol{h}$ being the observations in the Height column. The shadings separate the systems and the number of rows within each block we denote n_s. I.e. here we have $\boldsymbol{n} = (2, 3, 3, 3, 2, 3, 2, 3, 3, 4)$ with $n = \sum_s n_s = 28$. Below we have formed the T matrix of this case which is of dimension 28×40 since we have 28 observations and ten systems. Owing to lack of space we have only shown the first six T_s (shaded) across the diagonal.

$$
T = \begin{bmatrix}
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \\
 & & & & & & & & & & & \vdots & & & & & & & & & & & & & \ddots
\end{bmatrix}
$$

The first block matrix T_1 refers to system D16 and since the two individuals have the same genotype $(10, 12)$ we add $A^{(1)}_{1,1} + A^{(2)}_{1,1}$ and $A^{(1)}_{2,1} + A^{(2)}_{2,1}$. The next block refers to D18, the third D19 and so on using the same ordering of the systems as in Table 3.1.

3.2 Compound symmetry of the error term

The observed values $\boldsymbol{M}$ also follow a normal distribution with dimension n with mean $T\boldsymbol{\mu}$ and covariance matrix $TDT^\top + \Sigma$, where Σ is the covariance matrix of ε. The covariance matrix of $(\boldsymbol{A}, \boldsymbol{M})$ is given as

$$\Sigma(\boldsymbol{A}, \boldsymbol{M}) = \begin{bmatrix} D & DT^\top \\ TD & TDT^\top + \Sigma \end{bmatrix},$$

since we have that

$$\mathrm{cov}(\boldsymbol{M}, \boldsymbol{A}) = \mathrm{cov}(T\boldsymbol{A} + \varepsilon, \boldsymbol{A}) = T\mathrm{cov}(\boldsymbol{A}) = TD.$$

The mutual distribution of $\boldsymbol{A}$ and $\boldsymbol{M}$ is again normal with

$$\begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{M} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ T\boldsymbol{\mu} \end{pmatrix}, \begin{bmatrix} D & DT^\top \\ TD & TDT^\top + \Sigma \end{bmatrix}\right)$$

and the conditional distribution of $\boldsymbol{A}|\boldsymbol{M}$ is therefore (Lauritzen, 1996, Proposition C.5)

$$\boldsymbol{A}|\boldsymbol{M} \sim \mathcal{N}\left(\boldsymbol{\mu} + DT^\top \left\{TDT^\top + \Sigma\right\}^{-1} (\boldsymbol{M} - T\boldsymbol{\mu}), D - DT^\top \left\{TDT^\top + \Sigma\right\}^{-1} TD\right). \quad (3.3)$$

From previous analyses we might expect the different alleles of a locus to have different amplification behaviour. However due to the nature of DNA mixtures we will most likely have different alleles present from case to case which makes it difficult to incorporate a covariance structure covering all such combinations. A possible way to go about this is to have a compound symmetry structure on the residuals with equal within and between covariance. This does not satisfy the allelic variability, but is operational feasible. If we let $\varepsilon = \boldsymbol{M} - T\boldsymbol{A}$ be the residuals then we assume it to have zero mean and covariance Σ. A way to let the structure of Σ be affected by the present alleles is to scale the residuals by the associated heights $\boldsymbol{h}$ which are observable. This case specific scaling allows the different alleles within systems to be scaled according to their amplification behaviour. That is if for a system shorter alleles amplify more easily than longer this is taken into account by this scaling. Let $\tilde{\varepsilon} = \mathrm{diag}(\boldsymbol{h})^{-1/2}\varepsilon$, where $\mathrm{diag}(\boldsymbol{v})$ forms a diagonal matrix with the elements of $\boldsymbol{v}$. Then we assume $\tilde{\varepsilon}$ to have a compound symmetry structure specified by,

$$\mathrm{cov}(\tilde{\varepsilon}_s, \tilde{\varepsilon}_t) = \tilde{\Sigma}_{st} = \begin{cases} \nu_{st}\mathbf{1}_{n_s}\mathbf{1}_{n_t}^\top, & s \neq t \\ \tau_s I_{n_s} + \nu_{ss}\mathbf{1}_{n_s}\mathbf{1}_{n_s}^\top, & s = t. \end{cases}, \quad (3.4)$$

where we have $\boldsymbol{M}_s = T_s\boldsymbol{A}_s + \varepsilon_s$ with $\tilde{\varepsilon}_s = \mathrm{diag}(\boldsymbol{h}_s)^{-1/2}\varepsilon_s$. This implies that

$$\tilde{\Sigma} = \mathrm{cov}(\tilde{\varepsilon}) = \mathrm{cov}(\mathrm{diag}(\boldsymbol{h})^{-1/2}\varepsilon) = \mathrm{diag}(\boldsymbol{h})^{-1/2}\Sigma\mathrm{diag}(\boldsymbol{h})^{-1/2},$$

hence $\Sigma = \mathrm{diag}(\boldsymbol{h})^{1/2}\tilde{\Sigma}\mathrm{diag}(\boldsymbol{h})^{1/2}$. Specifying the structure as above we have that

$$\mathrm{cov}(\varepsilon_s, \varepsilon_t) = \Sigma_{st} = \begin{cases} \nu_{st}\mathrm{diag}(\boldsymbol{h}_s)^{1/2}\mathbf{1}_{n_s}\mathbf{1}_{n_t}^\top\mathrm{diag}(\boldsymbol{h}_t)^{1/2}, & s \neq t \\ \mathrm{diag}(\boldsymbol{h}_s)^{1/2}(\tau_s I_{n_s} + \nu_{ss}\mathbf{1}_{n_s}\mathbf{1}_{n_s}^\top)\mathrm{diag}(\boldsymbol{h}_s)^{1/2}, & s = t. \end{cases},$$

i.e. the covariance structure on the residuals takes the different amplification behaviour over the alleles into account.

Below $\tilde{\Sigma}$ is given in matrix notation. The blocks on the diagonal are of dimensions $n_s \times n_s$ determining the dimensions of the off-diagonal blocks.

$$\tilde{\Sigma} = \left[\begin{array}{cccc:cccc:c:cccc}
\gamma_1^2 & \nu_{11} & \cdots & \nu_{11} & \nu_{1.2} & \nu_{1.2} & \cdots & \nu_{1.2} & \cdots & \nu_{1.S} & \nu_{1.S} & \cdots & \nu_{1.S} \\
\nu_{11} & \gamma_1^2 & \cdots & \nu_{11} & \nu_{1.2} & \nu_{1.2} & \cdots & \nu_{1.2} & \cdots & \nu_{1.S} & \nu_{1.S} & \cdots & \nu_{1.S} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\
\nu_{11} & \nu_{11} & \cdots & \gamma_1^2 & \nu_{1.2} & \nu_{1.2} & \cdots & \nu_{1.2} & \cdots & \nu_{1.S} & \nu_{1.S} & \cdots & \nu_{1.S} \\
\hdashline
\nu_{2.1} & \nu_{2.1} & \cdots & \nu_{2.1} & \gamma_2^2 & \nu_{22} & \cdots & \nu_{22} & \cdots & \nu_{2.S} & \nu_{2.S} & \cdots & \nu_{2.S} \\
\nu_{2.1} & \nu_{2.1} & \cdots & \nu_{2.1} & \nu_{22} & \gamma_2^2 & \cdots & \nu_{22} & \cdots & \nu_{2.S} & \nu_{2.S} & \cdots & \nu_{2.S} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\
\nu_{2.1} & \nu_{2.1} & \cdots & \nu_{2.1} & \nu_{22} & \cdots & \nu_{22} & \gamma_2^2 & \cdots & \nu_{2.S} & \nu_{2.S} & \cdots & \nu_{2.S} \\
\hdashline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
\hdashline
\nu_{S.1} & \nu_{S.1} & \cdots & \nu_{S.1} & \nu_{S.2} & \nu_{S.2} & \cdots & \nu_{S.2} & \cdots & \gamma_S^2 & \nu_{SS} & \cdots & \nu_{SS} \\
\nu_{S.1} & \nu_{S.1} & \cdots & \nu_{S.1} & \nu_{S.2} & \nu_{S.2} & \cdots & \nu_{S.2} & \cdots & \nu_{SS} & \gamma_S^2 & \cdots & \nu_{SS} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\
\nu_{S.1} & \nu_{S.1} & \cdots & \nu_{S.1} & \nu_{S.2} & \nu_{S.2} & \cdots & \nu_{S.2} & \cdots & \nu_{SS} & \nu_{SS} & \cdots & \gamma_S^2
\end{array}\right]$$

where $\gamma_j^2 = \tau_j + \nu_{jj}$ is the variance parameter for system j.

3.3 Derivation of EM estimators

In this section we derive estimators for the parameters in the model outlined above to be used in the EM-algorithm. Since we assume normality of our observations, missing as observable, we can use standard results about the normal distribution. From Section 4.2.1 on the EM-algorithm for exponential families, we need only to determine the sufficient statistics in the E-step.

First we show a general result which we are going to use multiple times in the following. Let $\boldsymbol{X}$ be a p-dimensional stochastic vector with mean $\mathrm{E}(\boldsymbol{X})$ and covariance $V = \mathrm{cov}(\boldsymbol{X})$. Then for an arbitrary $p \times p$-matrix A we have

$$\mathrm{E}(\boldsymbol{X}^\top A\boldsymbol{X}) = \mathrm{E}(\mathrm{tr}[\boldsymbol{X}^\top A\boldsymbol{X}]) = \mathrm{tr}[A(\mathrm{E}\boldsymbol{X}\mathrm{E}\boldsymbol{X}^\top + V)] = \mathrm{E}\boldsymbol{X}^\top A\mathrm{E}\boldsymbol{X} + \mathrm{tr}[AV], \quad (3.5)$$

where we used that expectation is linear and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ if both products exists.

Next we introduce two frequently used matrices Q and Q_c which we define as

$$Q = \begin{bmatrix} \mathbf{1}_4 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_4 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}_4 \end{bmatrix} \quad \text{and} \quad Q_c = \begin{bmatrix} \mathbf{1}_{n_{1c}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_{2c}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}_{n_{Sc}} \end{bmatrix},$$

where $\mathbf{1}_n$ is a n-dimensional vector of ones and $\mathbf{0}$ are zero vectors of suitable dimensions. The dimensions of Q and Q_c are $(4S)\times S$ and $(n_c)\times S$, respectively. The subscript c indicates that Q_c is case dependent where as Q is fixed across cases. Let $n_c = \sum_s n_{sc}$ be the number of observations in case c, then we see that $Q_c^\top Q_c = \mathrm{diag}(n_{1c}, \ldots, n_{Sc})$ and therefore $I_{n_c} - Q_c(Q_c^\top Q_c)^{-1}Q_c^\top$ and $Q_c(Q_c^\top Q_c)^{-1}Q_c^\top$ are idempotent which will be useful later.

Now from (3.3) we have that the conditional expectation of $\boldsymbol{A}$ given $\boldsymbol{M}$ can be found from

$$\mathrm{E}(\boldsymbol{A}_c|\boldsymbol{M}_c) = \boldsymbol{\mu}_c + D_c T_c^\top (T_c D_c^\top + \mathrm{diag}(\boldsymbol{h}_c)^{1/2}\tilde{\Sigma}\mathrm{diag}(\boldsymbol{h}_c)^{1/2})^{-1}(\boldsymbol{M}_c - T_c\boldsymbol{\mu}_c).$$

In the expression of the mean of $\boldsymbol{A}|\boldsymbol{M}$ we have $\boldsymbol{\mu}$ which is just a linear function of $\boldsymbol{H}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_S)$. Using Q we find that

$$\boldsymbol{\mu} = \mathrm{diag}(Q\boldsymbol{\alpha})\boldsymbol{H} = \mathrm{diag}(\underbrace{\alpha_1, \ldots, \alpha_1}_{4}, \ldots, \underbrace{\alpha_S, \ldots, \alpha_S}_{4})\boldsymbol{H}.$$

Since we have the same relationship on the variance with $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_S^2)$ replaced for $\boldsymbol{\alpha}$ we need to divide $\boldsymbol{A}$ by $\boldsymbol{H}^{1/2}$ in order to satisfy the assumptions of homoskedasticity of linear models. That is

$$\mathrm{diag}(\boldsymbol{H})^{-1/2}\boldsymbol{A} \sim \mathcal{N}(\mathrm{diag}(Q\boldsymbol{\alpha})\boldsymbol{H}^{1/2}, \mathrm{diag}(Q\boldsymbol{\sigma}^2)).$$

Since the different $A_{i,s,c}^{(k)}$'s are independent maximization with respect to $\boldsymbol{\alpha}$ can be done within each case and system. That is we need to minimize

$$\sum_{i,k,c}\left(\frac{A_{i,s,c}^{(k)}}{\sqrt{H_c^{(k)}}} - \sqrt{H_c^{(k)}}\alpha_s\right)^2$$

with respect to α_s. The MLE is found to be $\hat{\alpha}_s = \sum_{i,k,c} A_{i,s,c}^{(k)} / \sum_{i,k,c} H_c^{(k)}$. Multiplying a vector with $Q^\top$ from the left adds together elements of the same system, therefore this is also $\hat{\boldsymbol{\alpha}} = \sum_c Q^\top \boldsymbol{A}_c / \sum_c Q^\top \boldsymbol{H}_c$ where the division is done component-wise.

As usual $\hat{\sigma}_s^2$ is estimated by evaluating $(n-1)^{-1}(\boldsymbol{y} - \hat{\boldsymbol{y}})^2$. Instead of $\mathrm{diag}(Q\boldsymbol{\alpha})\boldsymbol{H}^{1/2}$ we write $\mathrm{diag}(\boldsymbol{H})^{-1/2}\boldsymbol{\mu}$ since $\boldsymbol{\mu} = \mathrm{diag}(Q\boldsymbol{\alpha})\boldsymbol{H}$. Then using (3.5) we have

$$\begin{aligned}
\hat{\sigma}_s^2 &= (4C-1)^{-1}\sum_c \mathrm{E}\left([\mathrm{diag}(\boldsymbol{H}_{sc})^{-1/2}(\boldsymbol{A}_{sc} - \boldsymbol{\mu}_{sc})]^\top[\mathrm{diag}(\boldsymbol{H}_{sc})^{-1/2}(\boldsymbol{A}_{sc} - \boldsymbol{\mu}_{sc})]\middle|\boldsymbol{M}_c\right) \\
&= (4C-1)^{-1}\sum_c \Big(\mathrm{E}\left\{\mathrm{diag}(\boldsymbol{H}_{sc})^{-1/2}(\boldsymbol{A}_{sc} - \boldsymbol{\mu}_{sc})\middle|\boldsymbol{M}_c\right\}^\top \mathrm{E}\left\{D(H_c)^{-1/2}(\boldsymbol{A}_{sc} - \boldsymbol{\mu}_{sc})\middle|\boldsymbol{M}_c\right\} \\
&\qquad + \mathrm{tr}\left\{D(H_c)^{-1}\mathrm{cov}(\boldsymbol{A}_{sc}|\boldsymbol{M}_c)\right\}\Big) \\
&= (4C-1)^{-1}\sum_c \Big(\mathrm{E}(\boldsymbol{A}_{sc} - \boldsymbol{\mu}_{sc}|\boldsymbol{M}_c)^\top \mathrm{diag}(\boldsymbol{H}_{sc})^{-1}\mathrm{E}(\boldsymbol{A}_{sc} - \boldsymbol{\mu}_{sc}|\boldsymbol{M}_c) \\
&\qquad + \mathrm{tr}\left\{D(H_c)^{-1}\mathrm{cov}(\boldsymbol{A}_{sc}|\boldsymbol{M}_c)\right\}\Big)
\end{aligned}$$

We divide by $4C-1$ since for each case we have two persons and two observations from each and only one parameter for the mean. Define the vector operator $\boldsymbol{x}^2 = \mathrm{diag}(\boldsymbol{x}\boldsymbol{x}^\top)$. This implies that $\mathrm{E}(\boldsymbol{x}^2) = \mathrm{E}(\mathrm{diag}(\boldsymbol{x}\boldsymbol{x}^\top)) = \mathrm{diag}\left\{\mathrm{E}(\boldsymbol{x})\mathrm{E}(\boldsymbol{x})^\top + \mathrm{cov}(\boldsymbol{x})\right\}$ since the diag() operator is linear. Now we can compute $\hat{\sigma}_s^2$ for all systems by using Q,

$$\hat{\boldsymbol{\sigma}}^2 = (4C-1)^{-1}\left\{\sum_c Q^\top \left([\mathrm{E}(A_c|M_c) - \boldsymbol{\mu}_c]^2/\boldsymbol{H}_c + \mathrm{diag}[\mathrm{cov}(\boldsymbol{A}_c|M_c)]/\boldsymbol{H}_c\right)\right\},$$

where the divisions are component-wise and $\mathrm{diag}(V)$ is the vector formed by the diagonal elements of the matrix V. From (3.3) we have $\mathrm{cov}(\boldsymbol{A}_c|\boldsymbol{M}_c) = D_c - D_c T_c^\top[T_c D_c T_c^\top + \mathrm{diag}(\boldsymbol{h}_c)^{1/2}\tilde{\Sigma}\mathrm{diag}(\boldsymbol{h}_c)^{1/2}]^{-1}T_c D_c$.

In order to derive estimators for the parameters in Σ without solving the likelihood equations, we define for each system s a rotation given by an orthogonal matrix O_s. The first row of O_s is given by $\boldsymbol{e}_{n_s}(1)^\top O_s = n_s^{-1/2}\mathbf{1}_{n_s}^\top$, where $\boldsymbol{e}_n(i)$ is the i'th canonical unit vector in $\mathbb{R}^n$. Define $\boldsymbol{\xi}_s$ as the rotated residual, $\boldsymbol{\xi}_s = O_s\tilde{\varepsilon}_s$ which has a zero-mean normal distribution with covariance matrix $O_s(\tau_s I_{n_s} + \nu_{ss}\mathbf{1}_{n_s}\mathbf{1}_{n_s}^\top)O_s = \tau_s I_{n_s} + \nu_{ss}n_s\boldsymbol{e}_{n_s}(1)\boldsymbol{e}_{n_s}(1)^\top$. Let $\overline{\tilde{\varepsilon}}_s = n_s^{-1}\sum_{i=1}^{n_s}\tilde{\varepsilon}_{s_i}$ with $\tilde{\varepsilon}_s = (\tilde{\varepsilon}_{s_1},\ldots,\tilde{\varepsilon}_{s_{n_s}})$. Since $\tilde{\varepsilon}_s = O_s^\top\boldsymbol{\xi}_s$ we find that

$$\begin{aligned}
||\tilde{\varepsilon}_s - \overline{\tilde{\varepsilon}}_s\mathbf{1}_{n_s}||^2 &= (\tilde{\varepsilon}_s - \overline{\tilde{\varepsilon}}_s\mathbf{1}_{n_s})^\top(\tilde{\varepsilon}_s - \overline{\tilde{\varepsilon}}_s\mathbf{1}_{n_s}) \\
&= \tilde{\varepsilon}_s^\top\tilde{\varepsilon}_s - \overline{\tilde{\varepsilon}}_s^2 n_s \\
&= \boldsymbol{\xi}_s^\top O_s O_s^\top\boldsymbol{\xi}_s - n_s(n_s^{-1}\mathbf{1}_{n_s}^\top O_s^\top\boldsymbol{\xi}_s)^\top(n_s^{-1}\mathbf{1}_{n_s}^\top O_s^\top\boldsymbol{\xi}_s) \\
&= \boldsymbol{\xi}_s^\top\boldsymbol{\xi}_s - \boldsymbol{\xi}_s^\top\boldsymbol{e}_{n_s}(1)\boldsymbol{e}_{n_s}(1)^\top\boldsymbol{\xi}_s \\
&= \sum_{i=1}^{n_s}\xi_{s_i}^2 - \xi_{s_1}^2 = \sum_{i=2}^{n_s}\xi_{s_i}^2.
\end{aligned}$$

But from the choice of O_s we have that $\xi_{s_1} = \sqrt{n_s\overline{\tilde{\varepsilon}}_s}$ implying that $\xi_{s_1}n_s^{-1/2} = \overline{\tilde{\varepsilon}}_s$. Combining these two expressions yield that $\overline{\tilde{\varepsilon}}_s \perp\!\!\!\perp ||\tilde{\varepsilon}_s - \overline{\tilde{\varepsilon}}_s\mathbf{1}_{n_s}||^2$. The compound symmetry on $\tilde{\Sigma}$ also imply that $\overline{\tilde{\varepsilon}}_t \perp\!\!\!\perp \tilde{\varepsilon}_s - \overline{\tilde{\varepsilon}}_s\mathbf{1}_{n_s}$ which is shown below,

$$\begin{aligned}
\mathrm{cov}(\tilde{\varepsilon}_s - \overline{\tilde{\varepsilon}}_s\mathbf{1}_{n_s}, \overline{\tilde{\varepsilon}}_t) &= [I_{n_s} - n_s^{-1}\mathbf{1}_{n_s}\mathbf{1}_{n_s}^\top]\mathrm{cov}(\tilde{\varepsilon}_s, \tilde{\varepsilon}_t)n_t^{-1}\mathbf{1}_{n_t} \\
&= [I_{n_s} - n_s^{-1}\mathbf{1}_{n_s}\mathbf{1}_{n_s}^\top]\left\{\delta_{st}\tau_s I_{n_s} + \nu_{st}\mathbf{1}_{n_s}\mathbf{1}_{n_t}\right\}n_t^{-1}\mathbf{1}_{n_t} \\
&= \mathbf{0}, \quad \text{for all } s \text{ and } t, \qquad (3.6)
\end{aligned}$$

where δ_{st} is Kronecker's delta.

From the length of $\tilde{\varepsilon}_s - \overline{\tilde{\varepsilon}}_s\mathbf{1}_{ns}$ we also have that

$$\mathrm{E}||\tilde{\varepsilon}_s - \mathbf{1}_{n_s}\overline{\tilde{\varepsilon}}_s||^2 = \sum_{i=2}^{n_s}\mathrm{E}\xi_{s_i}^2 = (n_s - 1)\tau_s$$

An unbiased estimator for τ_s is therefore $(n_s-1)^{-1}||\tilde{\varepsilon}_s - \mathbf{1}_{n_s}\overline{\tilde{\varepsilon}}_s||^2$ in accordance with standard results for linear models. Let $SSD_{sc} = ||\tilde{\varepsilon}_{sc} - \mathbf{1}_{n_{sc}}\overline{\tilde{\varepsilon}}_{sc}||^2$ which is distributed as $SSD_{sc} \sim \tau_s \chi^2_{(n_{sc}-1)}$. Since we have independence across cases the sum of SSD_{sc} is again chi-squared, $\sum_c SSD_{sc} \sim \tau_s \chi^2_{(n_{s+}-C)}$ where $n_{s+} - C = \sum_c (n_{sc} - 1)$.

When determining the covariance of $\overline{\tilde{\varepsilon}_s}$ and $\overline{\tilde{\varepsilon}_t}$ we use that

$$\begin{aligned}
\operatorname{cov}(\overline{\tilde{\varepsilon}_s}, \overline{\tilde{\varepsilon}_t}) &= n_s^{-1} \mathbf{1}_{n_s}^\top \operatorname{cov}(\tilde{\varepsilon}_s, \tilde{\varepsilon}_t) n_t^{-1} \mathbf{1}_{n_t} \\
&= n_s^{-1} \mathbf{1}_{n_s}^\top \left\{ \delta_{st} \tau_s I_{n_s} + \nu_{st} \mathbf{1}_{n_s} \mathbf{1}_{n_t}^\top \right\} n_t^{-1} \mathbf{1}_{n_t} \\
&= \delta_{st} \tau_s / n_s + \nu_{st}
\end{aligned}$$

Now we let $\overline{\tilde{\varepsilon}} = (\overline{\tilde{\varepsilon}_s})_{s \in \text{sys}}$. When we summarize our findings in

$$\operatorname{cov}(\overline{\tilde{\varepsilon}}) = \operatorname{diag}\left(\frac{\tau_s}{n_s} \right)_{s \in \text{sys}} + \Lambda, \quad \text{where } \Lambda = \{\nu_{st}\}_{s,t \in \text{sys}}. \tag{3.7}$$

From the above expressions a straight forward approach to obtain the estimates would be to take average over the cases in our data. However the dimension of $\boldsymbol{M}$ and thus n_s for each case vary according to the mixed profiles. Due to the covariance structure specified in (3.7) we need to include auxiliary variables in order to handle this. These variables are unobservable and thus we need to impute them in our EM-algorithm.

Now we write $\overline{\tilde{\varepsilon}_c}$ as a linear combination of two independent variables, $\overline{\tilde{\varepsilon}_c} = \boldsymbol{u}_c + \boldsymbol{v}_c$. Both $\boldsymbol{u}_c$ and $\boldsymbol{v}_c$ follow a zero mean normal distribution with variances $\operatorname{diag}(\tau_s / n_{sc})_{s \in \text{sys}}$ and Λ, respectively and they are assumed independent of SSD_{sc}. Now let $\boldsymbol{x}_c$ be either of $\boldsymbol{u}_c$ and $\boldsymbol{v}_c$, then since $\operatorname{cov}(\boldsymbol{x}_c, \overline{\tilde{\varepsilon}_c}) = \operatorname{cov}(\boldsymbol{x}_c)$ and $\varepsilon_c = \operatorname{diag}(\boldsymbol{h}_c)^{1/2} \tilde{\varepsilon}_c$, we have that

$$\begin{aligned}
\operatorname{cov}(\boldsymbol{x}_c, \boldsymbol{M}_c) &= \operatorname{cov}(\boldsymbol{x}_c, \tilde{\varepsilon}_c - Q_c \overline{\tilde{\varepsilon}_c} + Q_c \overline{\tilde{\varepsilon}_c}) \operatorname{diag}(\boldsymbol{h}_c)^{1/2} \\
&= \operatorname{cov}(\boldsymbol{x}_c, Q_c \overline{\tilde{\varepsilon}_c}) \operatorname{diag}(\boldsymbol{h}_c)^{1/2} \\
&= \operatorname{cov}(\boldsymbol{x}_c) Q_c^\top \operatorname{diag}(\boldsymbol{h}_c)^{1/2},
\end{aligned}$$

where the second equality holds due to the independence shown in (3.6). We denote $\operatorname{diag}(\boldsymbol{h}) = \mathrm{d}(\boldsymbol{h})$ and state the covariance matrices of $(\boldsymbol{u}_c, \boldsymbol{M}_c)$ and $(\boldsymbol{v}_c, \boldsymbol{M}_c)$,

$$\begin{aligned}
\Sigma(\boldsymbol{u}_c, \boldsymbol{M}_c) &= \begin{bmatrix} \mathrm{d}(\boldsymbol{\tau}_c) & \mathrm{d}(\boldsymbol{\tau}_c) Q_c^\top \mathrm{d}(\boldsymbol{h}_c)^{1/2} \\ \mathrm{d}(\boldsymbol{h}_c)^{1/2} Q_c \mathrm{d}(\boldsymbol{\tau}_c) & T_c D_c T_c + \Sigma_c \end{bmatrix} \\
\Sigma(\boldsymbol{v}_c, \boldsymbol{M}_c) &= \begin{bmatrix} \Lambda & \Lambda Q_c^\top \mathrm{d}(\boldsymbol{h}_c)^{1/2} \\ \mathrm{d}(\boldsymbol{h}_c)^{1/2} Q_c \Lambda & T_c D_c T_c + \Sigma_c \end{bmatrix},
\end{aligned}$$

where $\mathrm{d}(\boldsymbol{\tau}_c) = \operatorname{diag}(\tau_s / n_{sc})_{s \in \text{sys}}$ and $\Sigma_c = \mathrm{d}(\boldsymbol{h}_c)^{1/2} \tilde{\Sigma} \mathrm{d}(\boldsymbol{h}_c)^{1/2}$.

By assumption u_{sc} is independent of $||\tilde{\varepsilon}_{sc} - \overline{\tilde{\varepsilon}_{sc}} \mathbf{1}_{n_{sc}}||^2$ which imply $u_{sc}^2 \perp\!\!\!\perp ||\tilde{\varepsilon}_{sc} - \overline{\tilde{\varepsilon}_{sc}} \mathbf{1}_{n_{sc}}||^2$, furthermore we have that $u_{sc}^2 \sim \tau_s / n_{sc} \chi_1$ implying that $\sum_c n_{sc} u_{sc}^2 \sim \tau_s \chi_C$. Since the two contributors in estimating τ_s are independent they can be computed separately in the M-step of the EM-algorithm. First we note that $\mathrm{E}(\boldsymbol{u}_c \boldsymbol{u}_c^\top | \boldsymbol{M}_c) = \mathrm{E}(\boldsymbol{u}_c | \boldsymbol{M}_c) \mathrm{E}(\boldsymbol{u}_c | \boldsymbol{M}_c)^\top +$

$\mathrm{cov}(\boldsymbol{u}_c|\boldsymbol{M}_c)$ where we use (3.3) to determine the right hand side,

$$\begin{aligned} \mathrm{E}(\boldsymbol{u}_c|M_c) &= \mathrm{d}(\boldsymbol{\tau}_c)Q_c^\top \mathrm{d}(\boldsymbol{h}_c)^{1/2}(T_cD_cT_c^\top + \Sigma_c)^{-1}(\boldsymbol{M} - T_c\boldsymbol{\mu}_c) \\ \mathrm{cov}(\boldsymbol{u}_c|M_c) &= \mathrm{cov}(\boldsymbol{u}_c) - \mathrm{cov}(\boldsymbol{u}_c, \boldsymbol{M}_c)\mathrm{var}(\boldsymbol{M}_c)^{-1}\mathrm{cov}(\boldsymbol{M}_c, \boldsymbol{u}_c) \\ &= \mathrm{d}(\boldsymbol{\tau}_c) - \mathrm{d}(\boldsymbol{\tau}_c)Q_c^\top \mathrm{diag}(\boldsymbol{h}_c)^{1/2}(T_cD_cT_c^\top + \Sigma_c)^{-1}\mathrm{diag}(\boldsymbol{h}_c)^{1/2}Q_c\mathrm{d}(\boldsymbol{\tau}_c). \end{aligned}$$

Since we only need $\mathrm{E}(u_{sc}^2|\boldsymbol{M})$ the components for each system of $\boldsymbol{\tau}$ is just the diagonal $\mathrm{diag}(\mathrm{E}(\boldsymbol{u}_c\boldsymbol{u}_c^\top|\boldsymbol{M}_c))$ which we multiply by $\boldsymbol{n}_c$ to have a central estimate. Also the SSD_{sc} can be computed simultaneously for all systems as,

$$Q_c^\top \mathrm{E}\left\{(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c})^2\middle|\boldsymbol{M}_c\right\} = Q_c^\top \mathrm{E}(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c|}\boldsymbol{M}_c)^2 + Q_c^\top \mathrm{diag}\left\{\mathrm{cov}(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c|}\boldsymbol{M}_c)\right\}$$

Since $\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c} = (I_{n_c} - Q_c[Q_c^\top Q_c]^{-1}Q_c^\top)\tilde{\varepsilon}_c$ and the covariance of $(\varepsilon, \boldsymbol{M})$ is

$$\Sigma(\varepsilon_c, \boldsymbol{M}_c) = \begin{bmatrix} \Sigma_c & \Sigma_c \\ \Sigma_c & T_cD_cT_c^\top + \Sigma_c \end{bmatrix},$$

we have that $\mathrm{E}(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c|}\boldsymbol{M}_c)$ is just

$$\begin{aligned} \mathrm{E}(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c|}\boldsymbol{M}_c) &= (I_{n_c} - Q_c[Q_c^\top Q_c]^{-1}Q_c^\top)\mathrm{E}(\tilde{\varepsilon}_c|\boldsymbol{M}_c) \\ &= (I_{n_c} - Q_c[Q_c^\top Q_c]^{-1}Q_c^\top)\mathrm{d}(\boldsymbol{h}_c)^{-1/2}\Sigma_c(T_cD_cT_c^\top + \Sigma_c)^{-1}(\boldsymbol{M}_c - T_c\boldsymbol{\mu}_c) \\ &= (I_{n_c} - Q_c[Q_c^\top Q_c]^{-1}Q_c^\top)\tilde{\Sigma}_c\mathrm{d}(\boldsymbol{h}_c)^{1/2}(T_cD_cT_c^\top + \Sigma_c)^{-1}(\boldsymbol{M}_c - T_c\boldsymbol{\mu}_c). \end{aligned}$$

It is easy to verify that $\tilde{\Sigma}_c = \mathrm{diag}(Q_c\boldsymbol{\tau}) + Q_c\Lambda Q_c^\top$ gives the correct dimensions of $\tilde{\Sigma}_c$ and the sub-matrices hereof. Now,

$$\begin{aligned} (I_{n_c} - Q_c(Q_c^\top Q_c)^{-1}Q_c^\top)\tilde{\Sigma}_c &= (I_{n_c} - Q_c(Q_c^\top Q_c)^{-1}Q_c^\top)(\mathrm{diag}(Q_c\boldsymbol{\tau}) + Q_c\Lambda Q_c^\top) \\ &= (I_{n_c} - Q_c(Q_c^\top Q_c)^{-1}Q_c^\top)\mathrm{diag}(Q_c\boldsymbol{\tau}). \end{aligned}$$

Therefore $\mathrm{E}(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c|}\boldsymbol{M}_c)$ is expressed as

$$\mathrm{E}(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c|}\boldsymbol{M}_c) = (I_{n_c} - Q_c(Q_c^\top Q_c)^{-1}Q_c^\top)\mathrm{diag}(Q_c\boldsymbol{\tau})\mathrm{d}(\boldsymbol{h})^{1/2}\mathrm{var}(\boldsymbol{M}_c)^{-1}(\boldsymbol{M}_c - T_c\boldsymbol{\mu}_c),$$

with $\mathrm{var}(\boldsymbol{M}_c) = T_cD_cT_c^\top + \Sigma_c$. Let $K_c = I_{n_c} - Q_c(Q_c^\top Q_c)^{-1}Q_c^\top$, then the covariance is found by similar arguments,

$$\begin{aligned} \mathrm{cov}(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c|}\boldsymbol{M}_c) &= K_c\mathrm{d}(\boldsymbol{h}_c)^{-1/2}\left\{\Sigma_c - \Sigma_c\mathrm{var}(\boldsymbol{M}_c)^{-1}\Sigma_c\right\}\mathrm{d}(\boldsymbol{h}_c)^{-1/2}K_c \\ &= K_c\mathrm{diag}(Q_c\boldsymbol{\tau}) - K_c\mathrm{diag}(Q_c\boldsymbol{\tau})\mathrm{d}(\boldsymbol{h}_c)^{1/2}\mathrm{var}(\boldsymbol{M})^{-1}\mathrm{d}(\boldsymbol{h}_c)^{1/2}K_c\mathrm{diag}(Q_c\boldsymbol{\tau}), \end{aligned}$$

where we used $K_c\tilde{\Sigma}_cK_c = K_c\mathrm{diag}(Q_c\boldsymbol{\tau})$. Since $\sum_c u_{sc}^2 n_{sc} \sim \tau_s\chi_C^2$ and $\sum_c SSD_{sc} \sim \tau_s\chi_{n_{s+}-C}^2$ are independent their sum is distributed as $\tau_s\chi_{n_{s+}}^2$, i.e.

$$\hat{\tau}_s = n_{s+}^{-1}\sum_c\left\{\mathrm{E}(u_{sc}^2|\boldsymbol{M}_c)n_{sc} + SSD_{sc}\right\}$$

Letting $\boldsymbol{n}_+ = (n_{1+}, \dots, n_{S+})$, we find that

$$\hat{\boldsymbol{\tau}} = \mathrm{diag}(\boldsymbol{n}_+)^{-1} \sum_c \left\{ (Q_c^\top Q_c)\mathrm{E}(u_c^2|\boldsymbol{M}_c) + Q_c^\top \mathrm{E}([K_c\tilde{\boldsymbol{\varepsilon}}_c]^2|\boldsymbol{M}_c) \right\},$$

where both terms in the sum expands as the expectation squared plus the diagonal of the covariance matrix.

By construction of $\boldsymbol{v}_c$ we have $\mathrm{E}(\boldsymbol{v}_c\boldsymbol{v}_c^\top) = \mathrm{E}(\boldsymbol{v}_c)\mathrm{E}(\boldsymbol{v}_c)^\top + \mathrm{cov}(\boldsymbol{v}_c) = \Lambda$. Now as for $\boldsymbol{u}$ we need to calculate $\mathrm{E}(\boldsymbol{v}_c\boldsymbol{v}_c^\top|\boldsymbol{M}_c)$. The same considerations apply hence we have $\mathrm{E}(\boldsymbol{v}_c\boldsymbol{v}_c^\top|\boldsymbol{M}_c) = \mathrm{E}(\boldsymbol{v}_c|\boldsymbol{M}_c)\mathrm{E}(\boldsymbol{v}_c|\boldsymbol{M}_c)^\top + \mathrm{cov}(\boldsymbol{v}_c|\boldsymbol{M}_c)$ where

$$\begin{aligned} \mathrm{E}(\boldsymbol{v}_c|\boldsymbol{M}_c) &= \Lambda Q_c^\top \mathrm{d}(\boldsymbol{h}_c)^{1/2}\mathrm{var}(\boldsymbol{M}_c)^{-1}(\boldsymbol{M}_c - T_c\boldsymbol{\mu}_c) \\ \mathrm{cov}(\boldsymbol{v}_c|\boldsymbol{M}_c) &= \Lambda - \Lambda Q_c^\top \mathrm{diag}(\boldsymbol{h}_c)^{1/2}(T_cD_cT_c^\top + \Sigma_c)^{-1}\mathrm{diag}(\boldsymbol{h}_c)^{1/2}Q_c\Lambda. \end{aligned}$$

Hence Λ is just estimated as the mean over all cases

$$\hat{\Lambda} = C^{-1} \sum_c \mathrm{E}(\boldsymbol{v}_c|\boldsymbol{M}_c)\mathrm{E}(\boldsymbol{v}_c|\boldsymbol{M}_c)^\top + \mathrm{cov}(\boldsymbol{v}_c|\boldsymbol{M}_c).$$

Note that by decomposing $\overline{\tilde{\boldsymbol{\varepsilon}}_c}$ as two independent Gaussian variables $\boldsymbol{u}_c$ and $\boldsymbol{v}_c$ we force the covariance ν_{ss} to be positive since Λ is an ordinary positive definite covariance matrix for $\boldsymbol{v}_c$, i.e. the diagonal elements are positive. This implies that $\mathrm{cov}(\tilde{\boldsymbol{\varepsilon}}_s) = \tilde{\Sigma}_{ss}$,

$$\tilde{\Sigma}_{ss} = \begin{bmatrix} \tau_s + \nu_{ss} & \nu_{ss} & \dots & \nu_{ss} \\ \nu_{ss} & \tau_s + \nu_{ss} & \dots & \nu_{ss} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{ss} & \tau_s + \nu_{ss} & \dots & \tau_s + \nu_{ss} \end{bmatrix},$$

are all non-negative entrances.

CHAPTER 4

EM algorithm

In this chapter we formulate the EM algorithm in broad terms and demonstrate how we use it in our missing data approach to the DNA mixture problem. The EM algorithm consist of two general steps, the E-step where the missing observations are replaced by the expected values under the current estimates of the parameters $\boldsymbol{\theta}$, say, which are found by maximization in the M-step assuming full observations.

4.1 Missing data mechanisms

Following the terminology of Little and Rubin (2002) presented by Lauritzen (2006) the missing data can be generated by various mechanisms. In our concrete example the area observations $\boldsymbol{A}$ are missing completely at random (MCAR). That is the missing observations are independent from the observed values given the parameters.

Let $Y = (Y_{obs}, Y_{mis})$ and introduce the missing data matrix M which is 1 if Y is missing and 0 otherwise. Then we have that

$$f(M|Y,\boldsymbol{\theta}) = f(M|\boldsymbol{\theta}), \qquad \text{i.e.} \quad M \perp\!\!\!\perp Y|\boldsymbol{\theta} \qquad \text{(MCAR)}$$

$$f(M|Y,\boldsymbol{\theta}) = f(M|Y_{obs},\boldsymbol{\theta}) \qquad \text{i.e.} \quad M \perp\!\!\!\perp Y_{mis}|(Y_{obs},\boldsymbol{\theta}) \qquad \text{(MAR)}$$

yielding that MCAR imply MAR.

Below we factorize the likelihood

$$L(\boldsymbol{\theta}|M, y_{obs}) \propto \int L_{mis}(\boldsymbol{\theta}) f(y_{obs}, y_{mis}|\boldsymbol{\theta}) \mathrm{d}y_{mis},$$

where $L_{mis}(\boldsymbol{\theta}) \propto f(M|y_{obs}, y_{mis}, \boldsymbol{\theta})$ is based on an explicit model for the missing data mechanism. The likelihood function ignoring the missing mechanism is

$$L(\boldsymbol{\theta}|y_{obs}) \propto f(y_{obs}|\boldsymbol{\theta}) = \int f(y_{obs}, y_{mis}|\boldsymbol{\theta}) \mathrm{d}y_{mis}.$$

Assuming that $\boldsymbol{\theta} = (\phi, \psi)$ with ψ governing the missingness are separate from the parameter of interest ϕ (i.e. the parameters vary in a product region) and the data are MAR we have that $L_{mis}(\boldsymbol{\theta}) = L_{mis}(\psi) \propto f(M|y_{obs}, y_{mis}, \psi) = f(M|y_{obs}, \psi)$. Using this we get our final result

$$\begin{aligned} L(\boldsymbol{\theta}|M, y_{obs}) &\propto \int L_{mis}(\boldsymbol{\theta}) f(y_{obs}, y_{mis}|\boldsymbol{\theta}) \mathrm{d}y_{mis} = L_{mis}(\psi) \int f(y_{obs}, y_{mis}|\phi) \mathrm{d}y_{mis} \\ &\propto L_{mis}(\psi) L(\phi|y_{obs}). \end{aligned}$$

This shows that the missingness mechanism can be ignored when concerned with likelihood inference about ϕ.

From the present data we only have access to the peak area observations $\boldsymbol{M}$ for each mixture. In the controlled experiments discussed and analyzed in Chapter 2, we also have the single DNA profiles given. Hence we can construct the mapping T since we know which alleles the two donors share. The task is to recover the $\boldsymbol{A}$ vector since this contain information of the DNA profiles. In the terminology of Little and Rubin (2002) this implies the missing data mechanism of $\boldsymbol{A}$ is missing completely at random (MCAR). This implies we can ignore the missing data mechanism in the steps of the EM algorithm.

4.2 Theory of the EM-algorithm

In this section we show that after a complete cycle of the EM-algorithm the incomplete data log likelihood has never decreased. The proof is based on Lauritzen (2006). The EM-algorithm converges to either a saddlepoint, local or global maxima. For practical purposes it is often possible to avoid convergence to a saddlepoint by performing several independent runs with small perturbations of the initial parameter values.

In the E-step of the EM-algorithm we take the expectation of the log-likelihood ratio with respect to Y_{mis} given y_{obs} and current $\boldsymbol{\theta}_{(n)}$ estimates of $\boldsymbol{\theta}$,

$$\begin{aligned} q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(n)}) &= \mathrm{E}\left(\log \frac{f(Y_{mis}, y_{obs}; \boldsymbol{\theta})}{f(Y_{mis}, y_{obs}; \boldsymbol{\theta}_{(n)})} \,\Big|\, y_{obs}, \boldsymbol{\theta}_{(n)}\right) \\ &= \int \log \frac{f(y_{mis}, y_{obs}; \boldsymbol{\theta})}{f(y_{mis}, y_{obs}; \boldsymbol{\theta}_{(n)})} f(y_{mis}|y_{obs}, \boldsymbol{\theta}_{(n)}) \mathrm{d}y_{mis} \end{aligned}$$

Since $f(z|x;\phi) = f(z,x;\phi)/f(x;\phi)$ we have that

$$\begin{aligned} &= \int \log \frac{f(y_{obs}; \boldsymbol{\theta}) f(y_{mis}|y_{obs}; \boldsymbol{\theta})}{f(y_{obs}; \boldsymbol{\theta}_{(n)}) f(y_{mis}|y_{obs}; \boldsymbol{\theta}_{(n)})} f(y_{mis}|y_{obs}, \boldsymbol{\theta}_{(n)}) \mathrm{d}y_{mis} \\ &= \log f(y_{obs}; \boldsymbol{\theta}) - \log f(y_{obs}; \boldsymbol{\theta}_{(n)}) + \int \log \frac{f(y_{mis}|y_{obs}; \boldsymbol{\theta})}{f(y_{mis}|y_{obs}; \boldsymbol{\theta}_{(n)})} f(y_{mis}|y_{obs}, \boldsymbol{\theta}_{(n)}) \mathrm{d}y_{mis} \\ &= \ell_{y_{obs}}(\boldsymbol{\theta}) - \ell_{y_{obs}}(\boldsymbol{\theta}_{(n)}) - KL\left(f^{y_{obs}}_{\boldsymbol{\theta}_{(n)}}; f^{y_{obs}}_{\boldsymbol{\theta}}\right), \end{aligned} \tag{4.1}$$

where $\ell_{y_{obs}}(\boldsymbol{\theta}) = \log \int f(y_{mis}, y_{obs}; \boldsymbol{\theta}) \mathrm{d}y_{mis}$ is the incomplete data log-likelihood and KL is the Kullback-Leibler divergence defined by

$$KL(f,g) = \int f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x.$$

Let f and g be densities then we have that

$$KL(f,g) = -\int f(x) \log \frac{g(x)}{f(x)} \mathrm{d}x \geq -\log \int f(x) \frac{g(x)}{f(x)} \mathrm{d}x = 0,$$

due to Jensen's inequality for the concave function $-\log(x)$. We see that $KL(f, g) = 0$ for $f = g$ hence $KL\left(f^{y_{obs}}_{\boldsymbol{\theta}_{(n)}}; f^{y_{obs}}_{\boldsymbol{\theta}}\right)$ is minimized for $\boldsymbol{\theta} = \boldsymbol{\theta}_{(n)}$. Using (4.1) this furthermore imply that

$$\frac{\partial}{\partial\boldsymbol{\theta}} q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(n)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{(n)}} = \frac{\partial}{\partial\boldsymbol{\theta}} \ell_{y_{obs}}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{(n)}}. \tag{4.2}$$

In the M-step we set $\boldsymbol{\theta}_{(n+1)} = \arg\max_{\boldsymbol{\theta}} q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(n)})$, which yields that $q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(n)}) \geq 0$ since

$$\arg\max_{\boldsymbol{\theta}} q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(n)}) = \arg\max_{\boldsymbol{\theta}} \int (\log f(y; \boldsymbol{\theta}) - \log f(y; \boldsymbol{\theta}_{(n)}))\, f(y_{mis}|y_{obs}, \boldsymbol{\theta}_{(n)}) \mathrm{d}y_{mis},$$

where $y = (y_{mis}, y_{obs})$ and the latter integral is constant in $\boldsymbol{\theta}$ and hence

$$\int \log f(y; \boldsymbol{\theta}_{(n+1)}) f(y_{mis}|y_{obs}, \boldsymbol{\theta}_{(n)}) \mathrm{d}y_{mis} \geq \int \log f(y; \boldsymbol{\theta}_{(n)}) f(y_{mis}|y_{obs}, \boldsymbol{\theta}_{(n)}) \mathrm{d}y_{mis}.$$

From (4.1) we get

$$\begin{aligned}\ell_{y_{obs}}(\boldsymbol{\theta}^{(n+1)}) &= q(\boldsymbol{\theta}|\boldsymbol{\theta}_{(n)}) + \ell_{y_{obs}}(\boldsymbol{\theta}_{(n)}) + KL\left(f^{y_{obs}}_{\boldsymbol{\theta}_{(n)}}; f^{y_{obs}}_{\boldsymbol{\theta}}\right)\\ &\geq \ell_{y_{obs}}(\boldsymbol{\theta}_{(n)}).\end{aligned} \tag{4.3}$$

Hence after a complete step in the EM-algorithm the incomplete data log-likelihood has never decreased.

4.2.1 EM-algorithm for exponential families

As in many other areas of statistics distributions from a regular exponential family have useful properties in the settings of the EM-algorithm. The common functional form of the exponential family is

$$f(\boldsymbol{Y}; \boldsymbol{\theta}) = b(\boldsymbol{Y}) \exp(s(\boldsymbol{Y})\boldsymbol{\theta}/a(\boldsymbol{\theta})), \tag{4.4}$$

where the expressions indicate that the only data terms need to be incorporated in the log likelihood is $s(\boldsymbol{Y})$ which is a d-dimensional vector of sufficient statistics. When the complete data Y follow a distribution of the regular exponential family the E- and M-step of the EM-algorithm reduces to, $s_{(n+1)} = \mathrm{E}(s(\boldsymbol{Y})|\boldsymbol{y}_{obs}, \boldsymbol{\theta})$ and due to (4.2) the solution to the likelihood equations $\mathrm{E}(s(\boldsymbol{Y})|\boldsymbol{\theta}) = s_{(n+1)}$, respectively.

4.3 Application of the EM-algorithm

A useful property of the EM-algorithm used in models with additional constraints on the parameters, is that the E-step in unaffected by these constraints. This however adds no further complications to situations with missing data since the calculations in the M-step is as if we had access to the complete data. By additional constraints we refer to restrictions induced by some model assumptions and not by the mathematical assumptions such as semi-definite covariance matrices.

The model described in Section 3.1 and Section 3.2 indicate an two-step structure where the first part involves $\boldsymbol{A}$ and $\boldsymbol{M}$ together with the parameters from their distributions D and $\boldsymbol{\mu}$. The other part models the error structure on ε with its covariance structure $\Sigma(\boldsymbol{\tau}, \Lambda)$. This is also obvious from the factorization of the likelihood shown below,

$$f(D, \boldsymbol{\mu}, \Sigma(\boldsymbol{\tau}, \Lambda); \boldsymbol{A}, \boldsymbol{M}, \boldsymbol{H}, \tilde{\varepsilon}, \boldsymbol{u}, \boldsymbol{v}) = f(D, \boldsymbol{\mu}; \boldsymbol{A}, \boldsymbol{M}, \boldsymbol{H}) f(\Sigma(\boldsymbol{\tau}, \Lambda); \tilde{\varepsilon}, \boldsymbol{u}, \boldsymbol{v}).$$

Hence when maximizing the likelihood with respect to $(D, \boldsymbol{\mu})$ and $\Sigma(\boldsymbol{\tau}, \Lambda)$, respectively, we get

$$\max_{D,\boldsymbol{\mu}} f(D, \boldsymbol{\mu}, \Sigma(\boldsymbol{\tau}, \Lambda) | \boldsymbol{A}, \boldsymbol{M}, \boldsymbol{H}, \tilde{\varepsilon}, \boldsymbol{u}, \boldsymbol{v}) = \max_{D,\boldsymbol{\mu}} f(D, \boldsymbol{\mu} | \boldsymbol{A}, \boldsymbol{M}, \boldsymbol{H})$$
$$\max_{\Sigma(\boldsymbol{\tau},\Lambda)} f(D, \boldsymbol{\mu}, \Sigma(\boldsymbol{\tau}, \Lambda) | \boldsymbol{A}, \boldsymbol{M}, \boldsymbol{H}, \tilde{\varepsilon}, \boldsymbol{u}, \boldsymbol{v}) = \max_{\Sigma(\boldsymbol{\tau},\Lambda)} f(\Sigma(\boldsymbol{\tau}, \Lambda) | \tilde{\varepsilon}, \boldsymbol{u}, \boldsymbol{v})$$

This ensures that we can perform calculations for both parts of our model within the same EM-algorithm (Little and Rubin, 2002, Section 7.1). Without this factorization the EM-procedure may have been split into more than one EM-procedure.

We can summarize our findings from Section 3.3 in Figure 4.1 and in terms of the estimators of the two steps in the EM-algorithm below.

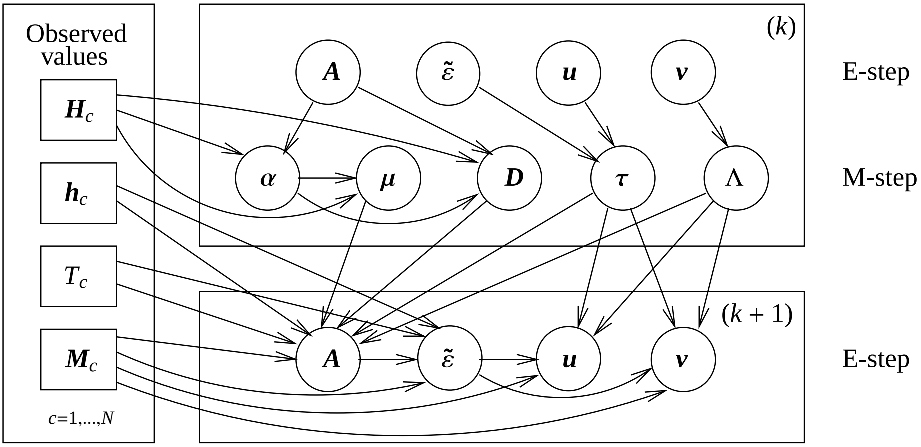

Figure 4.1: Graphical representation of the steps in our implementation of the EM-algorithm for fixed k. We have dropped the subscripts on the parameters and imputed values to keep the picture simple.

E-step In the E-step we need to compute $\mathrm{E}(\boldsymbol{A}|\boldsymbol{M})$ and $\mathrm{cov}(\boldsymbol{A}|\boldsymbol{M})$ for the outer procedure, and $\mathrm{E}(\boldsymbol{u}|\boldsymbol{M})$, $\mathrm{cov}(\boldsymbol{u}|\boldsymbol{M})$, $\mathrm{E}(\boldsymbol{v}|\boldsymbol{M})$, $\mathrm{cov}(\boldsymbol{v}|\boldsymbol{M})$, $\mathrm{E}(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c}|\boldsymbol{M}_c)$ and $\mathrm{cov}(\tilde{\varepsilon}_c -$

$Q_c\overline{\tilde{\varepsilon}_c}|\boldsymbol{M}_c)$ for the inner procedure. For both $\mathrm{cov}(\boldsymbol{A}|\boldsymbol{M})$, $\mathrm{cov}(\boldsymbol{u}|\boldsymbol{M})$ and $\mathrm{cov}(\tilde{\varepsilon}_c - Q_c\overline{\tilde{\varepsilon}_c}|\boldsymbol{M}_c)$ we need only the diagonal elements, i.e. when implementing in computer software we only have to store the diagonal elements. Below we list the estimators of the moments with appropriate indexing,

$$
\begin{aligned}
\mathrm{E}(\{\boldsymbol{A}_{(n+1)}\}_c|\boldsymbol{M}_c) &= \{\boldsymbol{\mu}_{(n)}\}_c + \{D_{(n)}\}_c T_c^\top \{\boldsymbol{V}_{(n)}\}_c^{-1}(\boldsymbol{M}_c - T_c\{\boldsymbol{\mu}_{(n)}\}_c),\\
\mathrm{cov}(\{\boldsymbol{A}_{(n+1)}\}_c|\boldsymbol{M}_c) &= \{D_{(n)}\}_c - \{D_{(n)}\}_c T_c^\top \{\boldsymbol{V}_{(n)}\}_c^{-1} T_c\{D_{(n)}\}_c\\
\mathrm{E}(\{\boldsymbol{u}_{(n+1)}\}_c|M_c) &= \mathrm{d}(\{\boldsymbol{\tau}_{(n)}\}/\boldsymbol{n}_c)Q_c^\top \mathrm{d}(\boldsymbol{h}_c)^{1/2}\{\boldsymbol{V}_{(n)}\}_c^{-1}(\boldsymbol{M}_c - T_c\{\boldsymbol{\mu}_{(n)}\}_c)\\
\mathrm{cov}(\{\boldsymbol{u}_{(n)}\}_c|M_c) &= \mathrm{d}(\{\boldsymbol{\tau}_{(n)}\}/\boldsymbol{n}_c)\\
&\quad - \mathrm{d}(\{\boldsymbol{\tau}_{(n)}\}/\boldsymbol{n}_c)Q_c^\top \mathrm{diag}(\boldsymbol{h}_c)^{1/2}\{\boldsymbol{V}_{(n)}\}_c^{-1}\mathrm{diag}(\boldsymbol{h}_c)^{1/2}Q_c\mathrm{d}(\{\boldsymbol{\tau}_{(n)}\}/\boldsymbol{n}_c)\\
\mathrm{E}(\{\boldsymbol{v}_{(n+1)}\}_c|M_c) &= \{\Lambda_{(n)}\}Q_c^\top \mathrm{d}(\boldsymbol{h}_c)^{1/2}\{\boldsymbol{V}_{(n)}\}_c^{-1}(\boldsymbol{M}_c - T_c\{\boldsymbol{\mu}_{(n)}\}_c)\\
\mathrm{cov}(\{\boldsymbol{v}_{(n)}\}_c|M_c) &= \{\Lambda_{(n)}\} - \{\Lambda_{(n)}\}Q_c^\top \mathrm{diag}(\boldsymbol{h}_c)^{1/2}\{\boldsymbol{V}_{(n)}\}_c^{-1}\mathrm{diag}(\boldsymbol{h}_c)^{1/2}Q_c\{\Lambda_{(n)}\}\\
\mathrm{E}(K_c\{\tilde{\varepsilon}_{(n)}\}_c|\boldsymbol{M}_c) &= K_c\{\tilde{\Sigma}_{(n)}\}_c\mathrm{d}(\boldsymbol{h}_c)^{1/2}(T_c\{D_{(n)}\}_c T_c^\top + \{\Sigma_{(n)}\}_c)^{-1}(\boldsymbol{M}_c - T_c\{\boldsymbol{\mu}_{(n)}\}_c)\\
\mathrm{cov}(K_c\{\tilde{\varepsilon}_{(n)}\}_c|\boldsymbol{M}_c) &= K_c\mathrm{diag}(Q_c\{\boldsymbol{\tau}_{(n)}\})\\
&\quad - K_c\mathrm{diag}(Q_c\{\boldsymbol{\tau}_{(n)}\})\mathrm{d}(\boldsymbol{h}_c)^{1/2}\{\boldsymbol{V}_{(n)}\}_c^{-1}\mathrm{d}(\boldsymbol{h}_c)^{1/2}K_c\mathrm{diag}(Q_c\{\boldsymbol{\tau}_{(n)}\}),
\end{aligned}
$$

where $\{\boldsymbol{V}_{(n)}\}_c = T_c\{D_{(n)}\}_c T_c^\top + \{\Sigma_{(n)}\}_c$, $K_c = I_{n_c} - Q_c[Q_c^\top Q_c]^{-1}Q_c^\top$ and the subscript (n) indicates the current estimates of the parameters.

M-step We use the expressions where we estimate the parameters for all systems at a time.

$$
\begin{aligned}
\boldsymbol{\alpha}_{(n+1)} &= \sum_c Q^\top \mathrm{E}(\{\boldsymbol{A}_{(n+1)}\}_c|\boldsymbol{M}_c) / \sum_c Q^\top \boldsymbol{H}_c\\
\{\boldsymbol{\mu}_{(n+1)}\}_c &= \mathrm{diag}(Q\boldsymbol{\alpha}_{(n+1)})\boldsymbol{H}_c\\
\boldsymbol{\sigma}^2_{(n+1)} &= (4C-1)^{-1}\sum_c Q^\top[\mathrm{E}(\{A_{(n+1)}\}_c|M_c) - \{\boldsymbol{\mu}_{(n+1)}\}_c]^2/\boldsymbol{H}_c\\
&\quad + (4C-1)^{-1}\sum_c Q^\top \mathrm{diag}[\mathrm{cov}(\{\boldsymbol{A}_{(n+1)}\}_c|M_c)]/\boldsymbol{H}_c\\
\boldsymbol{\tau}_{(n+1)} &= \mathrm{diag}(\boldsymbol{n}_+)^{-1}\sum_c (Q_c^\top Q_c)\mathrm{diag}\Big[\mathrm{E}(\{\boldsymbol{u}_{(n+1)}\}_c|M_c)\mathrm{E}(\{\boldsymbol{u}_{(n+1)}\}_c|M_c)^\top + \mathrm{cov}(\{\boldsymbol{u}_{(n)}\}_c|M_c)\Big]\\
&\quad + \mathrm{diag}(\boldsymbol{n}_+)^{-1}\sum_c Q_c^\top \mathrm{E}(\{\tilde{\varepsilon}_{(n)}\}_c - Q_c\{\overline{\tilde{\varepsilon}_{(n)}}\}_c|\boldsymbol{M}_c)^2 + Q_c^\top \mathrm{diag}\,\{\mathrm{cov}(\{\tilde{\varepsilon}_{(n)}\}_c - Q_c\{\overline{\tilde{\varepsilon}_{(n)}}\}_c|\boldsymbol{M}_c)\}\\
\Lambda_{(n+1)} &= C^{-1}\sum_c \mathrm{E}(\{\boldsymbol{v}_{(n+1)}\}_c|\boldsymbol{M}_c)\mathrm{E}(\{\boldsymbol{v}_{(n+1)}\}_c|\boldsymbol{M}_c)^\top + \mathrm{cov}(\{\boldsymbol{v}_{(n+1)}\}_c|\boldsymbol{M}_c)
\end{aligned}
$$

where all vector divisions are done component-wise.

4.4 Implementation of EM-algorithm

The expression from Section 4.3 will be used to estimate the parameters in the model from cases with *full* information. That is cases where no drop-outs have been observed. Each case in the data were amplified using six and 12 seconds injection time, respectively, during the PCR reaction. This doubling of the injection time resulted in higher peaks and larger peak areas, hence the two subsets represents different populations of samples. We therefore restrict the data only to contain mixtures with an injection time of six seconds.

In our data we have also access to cases where only one person has contributed to the sample. These cases will however not be used in the parameter estimation phase as they are subject to an even higher degree of uncertainty compared to real mixtures.

From the expressions specified under the E-step and M-step in the previous section it is possible to compute case-wise contributions to the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\sigma}^2$, $\boldsymbol{\tau}$ and Λ. The pseudo code in Figure 4.2 emphasizes this.

```
**EM-algorithm**(init = list{$\boldsymbol{\alpha}$, $\boldsymbol{\sigma}^2$, $\boldsymbol{\tau}$, diag($\Lambda$)},data)
    for $n \in \{1, \ldots, N\}$
        for $c \in \{1, \ldots, C\}$
            Update moments:
                $\mathtt{A}_c \leftarrow \mathrm{E}(\{\boldsymbol{A}_{(n)}\}_c | \boldsymbol{M}_c)$
                $\mathtt{cA}_c \leftarrow \mathrm{diag}(\mathrm{cov}(\{\boldsymbol{A}_{(n)}\}_c | \boldsymbol{M}_c))$
                $\mathtt{u}_c \leftarrow \mathrm{E}(\{\boldsymbol{u}_{(n)}\}_c | \boldsymbol{M}_c)$
                $\mathtt{cu}_c \leftarrow \mathrm{diag}(\mathrm{cov}(\{\boldsymbol{u}_{(n)}\}_c | \boldsymbol{M}_c))$
                $\mathtt{v}_c \leftarrow \mathrm{E}(\{\boldsymbol{v}_{(n)}\}_c | \boldsymbol{M}_c)$
                $\mathtt{cv}_c \leftarrow \mathrm{cov}(\{\boldsymbol{v}_{(n)}\}_c | \boldsymbol{M}_c)$
                $\mathtt{ep}_c \leftarrow \mathrm{E}(\{\tilde{\varepsilon}_{(n)}\}_c - Q_c \{\overline{\tilde{\varepsilon}_{(n)}}\}_c | \boldsymbol{M}_c)$
                $\mathtt{cep}_c \leftarrow \mathrm{diag}(\mathrm{cov}(\{\tilde{\varepsilon}_{(n)}\}_c - Q_c \{\overline{\tilde{\varepsilon}_{(n)}}\}_c | \boldsymbol{M}_c))$
            Compute parameter contributions:
                $\boldsymbol{\alpha}_c^{(1)} \leftarrow Q^\top \mathtt{A}_c$
                $\boldsymbol{\alpha}_c^{(2)} \leftarrow Q^\top \boldsymbol{H}_c$
                $\boldsymbol{\sigma}_c^2 \leftarrow Q^\top \left( \left\{ \mathtt{cA}_c + (\mathtt{A}_c - \{\boldsymbol{\mu}_{(n)}\}_c)^2 \right\} / \boldsymbol{H}_c \right)$
                $\boldsymbol{\tau}_c \leftarrow Q_c^\top \left( \mathtt{cep}_c + \mathtt{ep}_c^2 \right) + \mathrm{diag}(\boldsymbol{n}_c)(\mathtt{cu}_c + \mathtt{u}_c^2)$
                $\Lambda_c \leftarrow \mathtt{v}_c \mathtt{v}_c^\top + \mathtt{cv}_c$
        Compute parameters:
        $\boldsymbol{\alpha} \leftarrow \sum_c \boldsymbol{\alpha}_c^{(1)} / \sum_c \boldsymbol{\alpha}_c^{(2)}$
        $\boldsymbol{\sigma}^2 \leftarrow (4C - 1)^{-1} \sum_c \boldsymbol{\sigma}_c^2$
        $\boldsymbol{\tau} \leftarrow \boldsymbol{n}_+^{-1} \sum_c \boldsymbol{\tau}_c$
        $\Lambda \leftarrow \boldsymbol{C}^{-1} \sum_c \Lambda_c$
**Return**(list{$\boldsymbol{\alpha}$, $\boldsymbol{\sigma}^2$, $\boldsymbol{\tau}$, $\Lambda$},data)
```

Figure 4.2: Pseudo code for the EM-algorithm in this application.

To perform and compute the iterations of the EM-algorithm we will use the open source S-language R as we will make extensive use of matrix algebra and manipulation with matrices. The pseudo code of Figure 4.2 also works as indirect comments for the R-source code available at http://www.math.aau.dk/~tvede. To be able to use the R script the data must be given in a data frame with the same structure as in Table 3.1. Below we have stated the data frame column names and lines needed for a successful execution of the script.

```
> names(DATA)
 "case" "p1c1" "p1c2" "p2c1" "p2c2" "sys" "allele" "height" "area"
> source(file="datahandle.R")
> source(file="engine.R")
> source(file="em.R")
> em.output <- EM(n=N,x=DATA,inu=NU,itau=TAU,iD=D)
```

The three files loaded before the run of the `EM`-function computes $\boldsymbol{H}_c$, T_c, and $\boldsymbol{n}_c$ for each case; contains the core functions of the EM-algorithm based on the estimators in Chapter 3; and the managing functions as described by Figure 4.2, respectively. The `EM`-function takes apart from the arguments specified, which is the number of iterations (`n`), data frame (`x`) and initial values (`inu`, `itau` and `iD`), also a Boolean `print` argument. If true (default) the current parameter estimates are printed to the screen after each iteration. The traces of parameters and deviance are stored by the script and returned at termination. The output of the `EM`-function (here named `em.output`) will contain the data and estimates of i.a. $\boldsymbol{A}_c$ and $\boldsymbol{\mu}_c$, together with traces and the final parameter estimates.

CHAPTER 5

Results

In this chapter we present the parameter estimates computed using the EM-algorithm of the previous chapter. We also investigate insignificance and asympotic variance of the parameters in the model.

5.1 Parameter estimates

Since our likelihood may be multimodal the EM-algorithm is sensitive to initial values. In order to verify the convergence of the algorithm we have used several different sets of initial values for $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$, Λ and $\boldsymbol{\sigma}^2$. In Table 5.1 we have listed the nine sets of initial values which will be used to analyze the convergence properties in this chapter.

Table 5.1: Initial values used for the EM-algorithm in the left table. For all sets we made 30,000 iterations. The two columns in the right table contain the deviance after 1100 and 30000 iterations, respectively.

	$\boldsymbol{\tau}$	$\boldsymbol{\sigma}^2$	diag(Λ)	D_{1100}	D_{30000}
Run 1	300	1000	1000	33015.03	33013.97
Run 2	0	1000	1000	33014.97	33014.77
Run 3	500	200	10	33016.66	33014.00
Run 4	0	1000	0	33342.88	33342.88
Run 5	100	100	100	33015.12	33013.97
Run 6	0	100	100	33014.97	33014.77
Run 7	40	1000	0	33302.96	33301.69
Run 8	1000	1000	1000	33015.07	33013.97
Run 9	1000	100	300	33015.43	33013.98

From Table 5.1 we see that for *Run 2, 4* and *6* we have initialized $\boldsymbol{\tau} = \mathbf{0}$ and for *Run 4* and *7* the Λ matrix were initialized as a zero matrix. The estimates for Λ and $\boldsymbol{\tau}$ found in Section 3.3 are both evaluated using the previous parameter estimate for Λ and $\boldsymbol{\tau}$, respectively. It is therefore not possible for these two parameters to attain other values when they at some point are zero. This is in particular true when they are initialized to zero.

A measure of the convergence is the deviance defined as $-2\log L(\boldsymbol{\theta}_{(n)})$, where $L(\boldsymbol{\theta}_{(n)})$ is the likelihood evaluated with the current parameter estimates $\boldsymbol{\theta}_{(n)}$. Since the likelihood of $\boldsymbol{M}$ is given as

$$L(\boldsymbol{M}; \Sigma, \boldsymbol{\mu}) = (2\pi|\Sigma|)^{-1/2} \exp\left[-\tfrac{1}{2}(\boldsymbol{M} - T\boldsymbol{\mu})^\top \mathrm{var}(\boldsymbol{M})^{-1}(\boldsymbol{M} - T\boldsymbol{\mu})\right],$$

the deviance D is just $\log(|\text{var}(\boldsymbol{M})|)+(\boldsymbol{M}-\boldsymbol{T}\boldsymbol{\mu})^\top\text{var}(\boldsymbol{M})^{-1}(\boldsymbol{M}-\boldsymbol{T}\boldsymbol{\mu})$ where $\text{var}(\boldsymbol{M}) = \boldsymbol{T}\boldsymbol{D}\boldsymbol{T}^\top + \boldsymbol{\Sigma}$. The total deviance is just the sum of case-wise deviances since we assume independence between cases. Since the deviance is just a monotone function of the likelihood the deviance is a measure of the goodness of fit implying that for different models the one with the lowest deviance has the largest likelihood. In Figure 5.1 we have plotted the trace of the deviance for the different initial values.

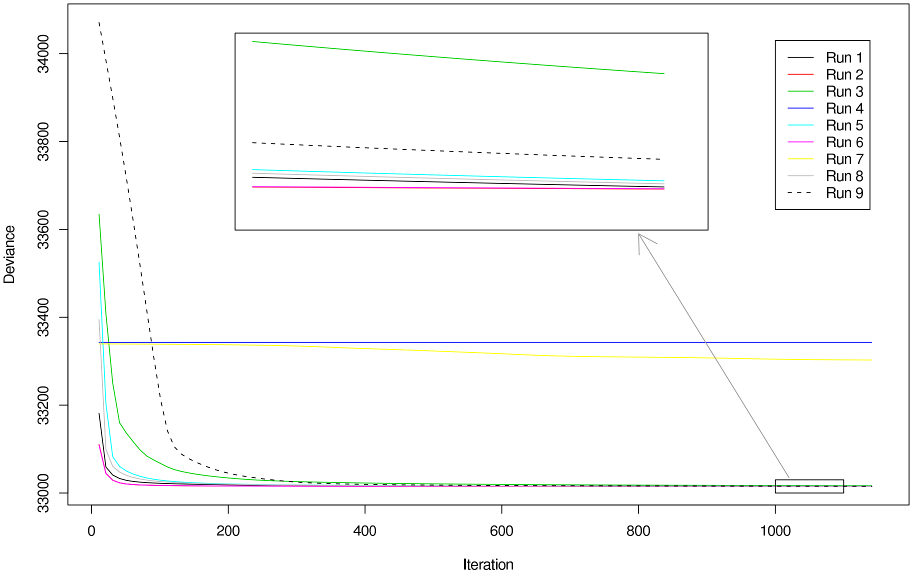

Figure 5.1: Traces of the deviance from iterations with different initial values.

The cut point of 1100 iterations was chosen since after this point none of the deviances improved more than 0.01 per iteration. In the right table of Table 5.1 we have listed the deviance for the runs after 1100 and 30000 iterations. We see that there is only marginal improvement of the fit with an additional 28900 iterations. However it is worth notice that the deviances of *Run 2* and *Run 6*, where $\boldsymbol{\tau} = \boldsymbol{0}$, are smaller than runs with $\boldsymbol{\tau} \neq \boldsymbol{0}$ after 1100 iterations, but that after 30000 iterations it is reversed. This implies that the model with $\boldsymbol{\tau} \neq \boldsymbol{0}$ fits marginally better, but as we see below this improvement is statistical insignificant. Since $\boldsymbol{\tau}$ is of dimension $S = 10$ the test of $\boldsymbol{\tau} = \boldsymbol{0}$ is approximately asymptotic chi-square distributed with ten degrees of freedom (Cox and Hinkley, 1974, Section 9.3),

$$D_{Run\ 5} - D_{Run\ 6} \sim \chi^2_{10},$$

where *Run 5* and *Run 6* are chosen since they differ only in initial values on $\boldsymbol{\tau}$. Using the deviance after 30000 iterations yields a test statistic of 0.7979 and a p-value of 0.9999 implying that we reject the hypothesis of $\boldsymbol{\tau} \neq \boldsymbol{0}$.

From the analysis of the deviance it is clear that cases where $\text{diag}(\boldsymbol{\Lambda})$ is initialized as or close to $\boldsymbol{0}$ has the worst fit. This implies that the covariance within and between systems are important to incorporate in the model. For further investigations of the parameters we chose *Run 2* since it has $\boldsymbol{\tau}$ initialized as $\mathbf{0}$ and small deviance.

The estimated parameters after 30000 iterations of *Run 2* are given in Table 5.2. For Λ the intrasystem covariances are displayed on the diagonal, the upper triangle the intersystem covariances and intersystem correlations in the lower triangle of the matrix (shaded). For $\boldsymbol{\tau}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\alpha}$ the estimates are listed below. By construction only the diagonal of Λ were forced to be positive but all system correlations are positive with values larger than 0.36 for all but one. This indicates that analyzing DNA profiles with the assumption of independence between systems is an extensive simplification. Furthermore we see that the correlations between systems on the same dye band are not necessarily larger than the others (Blue: D3, vWA, D16 and D2; Green: D8, D21 and D18; Yellow: D19, TH0 and FGA).

Since these estimates are based on pairwise mixtures of only four individuals and one machine the parameters are likely to be very data set specific. Note for example the large variation on system D18 where the variance pair $(\nu, \sigma^2)_{D18} = (2197, 3209)$. From Table 2.1 we see that there are two homozygous with shared allele for this system and in Figure 2.4 we recognize the large spread for this system. Conversely the systems where only a few alleles are shared between the individuals the overall variance is comparatively small, e.g. FGA and vWA.

An interesting observation about the intrasystem covariances ν_{ss}, $s \in \text{sys}$ is that their magnitude tends to follow the unbalances pictured in Figure A.1 with respect to the alleles included in our data set. That is system D8 with the lowest intrasystem covariance also tends to have the most homogeneous amplification behaviour across alleles by visual inspection of the allelic ladder. The pattern can only be examined with respect to the included alleles in the four profiles, however for the present alleles the intrasystem covariances seem to reflect the variance in allele amplification.

In Table 5.3 we have listed the parameters from the six runs with diag(Λ) not initialized to zero. The concordance of the parameters indicates that the EM-algorithm has converges to a maxima and not a saddlepoint. However there is no guarantee for this to be a global maximum.

An indication of the goodness of fit can also be assessed by linking the parameter estimates of $\boldsymbol{\alpha}$ with the box plots in Figure 2.3. Sorting α_s in the same order as in the box plots gives an increasing sequence except for D8 and D18. This reflects that system on the green dye band has more narrow peaks than systems on the yellow band. We find, except for D2 and vWA, a similar pattern for the variance component $\boldsymbol{\sigma}^2$ with the same ordering. This supports the assumption of proportionality of the mean and variance of the peak areas. Except for system D18 there is a linear relation between α_s and σ_s^2. A least square fit yields a coefficient of approximately 190 implying that we may write $\sigma_s^2 = 190\alpha_s$ for $s \in \text{sys}\backslash\{D18\}$, such that $\mathrm{E}(\boldsymbol{A}) = \boldsymbol{\alpha H}$ and $\mathrm{var}(\boldsymbol{A}) = K\boldsymbol{\alpha H}$ with $K = 190$. In Figure 5.2 we have plotted the estimates from the nine runs of Table 5.1 against each other. We see the obvious linear relationship but also that an affine mapping (dashed line) might be more appropriate than just a scaling (solid line).

Table 5.2: The estimated parameters after 30000 iterations of *Run 2*, i.e. $\tau = 0$.

		D16	*D18*	*D19*	*D2*	*D21*	*D3*	*D8*	*FGA*	*TH0*	*vWA*
	D16	1915.776	1653.911	1716.853	1848.140	1201.832	954.082	246.497	1305.358	527.211	1339.643
	D18	0.806	2196.651	1964.678	2077.812	1353.252	1142.766	750.020	1461.591	1085.372	1449.906
	D19	0.866	0.925	2052.831	2151.759	1279.925	1050.949	619.224	1441.000	1042.031	1319.491
	D2	0.848	0.890	0.953	2481.701	1354.935	1237.071	765.346	1492.006	1090.164	1438.362
$\Lambda =$	*D21*	0.890	0.936	0.915	0.881	952.266	776.642	380.722	1033.495	654.255	975.914
	D3	0.742	0.829	0.789	0.845	0.856	864.066	536.557	857.456	664.553	821.782
	D8	0.197	0.558	0.477	0.536	0.431	0.637	820.995	397.344	582.881	340.353
	FGA	0.879	0.919	0.937	0.883	0.987	0.860	0.409	1151.536	773.909	1044.479
	TH0	0.396	0.761	0.756	0.719	0.697	0.743	0.669	0.750	925.693	587.580
	vWA	0.930	0.940	0.885	0.878	0.961	0.850	0.361	0.936	0.587	1082.097
	$\sigma^2 =$	1821.04	3208.726	1002.926	1236.31	1797.387	1331.789	1854.943	596.532	730.285	1146.784
	$\alpha =$	9.102	10.175	6.149	7.014	8.918	8.248	10.189	5.529	5.992	7.642

Table 5.3: Parameter values after 30000 iterations for *Run 1, 2, 3, 5, 6, 8* and *9*.

	D16	D18	D19	D2	D21	D3	D8	FGA	TH0	vWA	*Run*
α	9.12	10.21	6.16	7.03	8.94	8.26	10.20	5.53	6.00	7.66	*1*
	9.10	10.17	6.15	7.01	8.92	8.25	10.19	5.53	5.99	7.64	*2*
	9.11	10.23	6.16	7.03	8.95	8.27	10.19	5.53	6.00	7.64	*3*
	9.12	10.21	6.16	7.03	8.94	8.26	10.20	5.53	6.00	7.66	*5*
	9.10	10.17	6.15	7.01	8.92	8.25	10.19	5.53	5.99	7.64	*6*
	9.12	10.21	6.16	7.03	8.94	8.26	10.20	5.53	6.00	7.66	*8*
	9.12	10.21	6.16	7.03	8.94	8.26	10.20	5.53	6.00	7.66	*9*
τ	158.53	443.76	1.04	1.88	4.46	1.27	1.48	92.32	3.04	1.56	*1*
	0	0	0	0	0	0	0	0	0	0	*2*
	151.93	478.79	1.05	1.88	4.93	1.29	1.48	90.13	2.96	1.51	*3*
	158.47	443.81	1.04	1.88	4.54	1.27	1.49	91.82	3.05	1.57	*5*
	0	0	0	0	0	0	0	0	0	0	*6*
	158.32	444.76	1.04	1.88	4.51	1.27	1.48	92.10	3.04	1.56	*8*
	158.62	441.59	1.05	1.88	4.69	1.28	1.49	90.62	3.09	1.58	*9*
diag(Λ)	1840.03	1929.18	2026.49	2447.12	922.66	849.22	813.05	1115.74	913.04	1060.55	*1*
	1915.78	2196.65	2052.83	2481.70	952.27	864.07	821.00	1151.54	925.69	1082.10	*2*
	1852.61	1837.99	2021.97	2448.81	898.54	835.87	820.18	1125.75	920.02	1086.10	*3*
	1840.08	1929.15	2025.25	2447.58	919.51	848.30	813.34	1117.49	912.91	1060.84	*5*
	1915.79	2197.09	2052.50	2481.80	951.52	863.88	821.04	1151.92	925.63	1082.03	*6*
	1840.43	1926.56	2025.96	2447.33	920.96	848.55	813.35	1116.59	913.20	1061.38	*8*
	1839.59	1935.37	2022.13	2448.67	912.61	846.64	813.65	1121.53	912.12	1059.87	*9*
σ^2	1656.76	2746.88	1002.00	1234.82	1795.95	1331.81	1853.55	533.57	728.08	1148.60	*1*
	1821.04	3208.73	1002.93	1236.31	1797.39	1331.79	1854.94	596.53	730.29	1146.78	*2*
	1663.31	2717.22	1001.70	1234.87	1795.38	1332.11	1853.92	534.98	728.24	1148.95	*3*
	1656.84	2746.84	1001.98	1234.83	1795.87	1331.84	1853.56	533.88	728.07	1148.60	*5*
	1821.04	3208.69	1002.93	1236.31	1797.39	1331.80	1854.95	596.52	730.28	1146.79	*6*
	1656.98	2746.03	1001.98	1234.82	1795.90	1331.83	1853.56	533.71	728.08	1148.60	*8*
	1656.73	2748.75	1001.98	1234.84	1795.73	1331.90	1853.57	534.63	728.04	1148.58	*9*

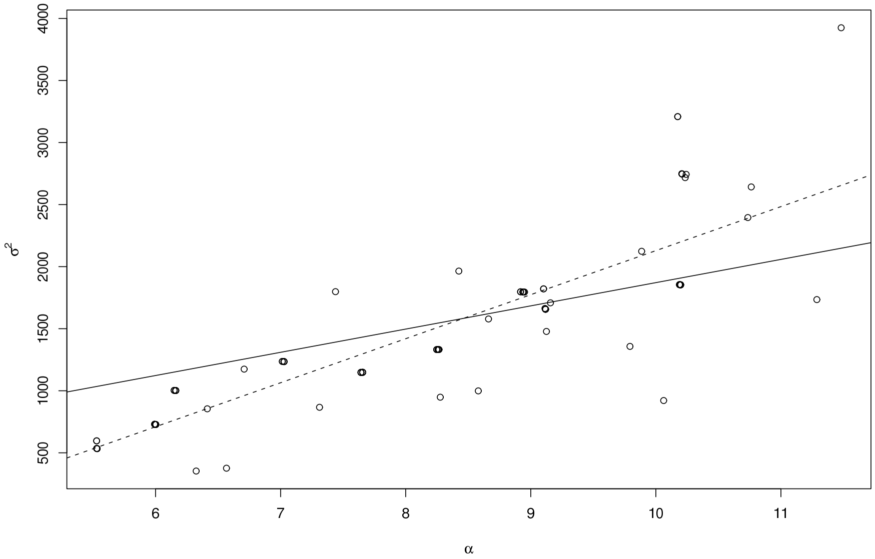

Figure 5.2: Scatter plot of the estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}^2$ from the nine runs of Table 5.1 with least squares linear fits superimposed.

The Mahalanobis distance $(\boldsymbol{M} - T\boldsymbol{\mu})^\top \text{var}(M)^{-1}(\boldsymbol{M} - T\boldsymbol{\mu})$ with $\boldsymbol{\mu}$ and $\text{var}(M) = TDT^\top + \Sigma$ known, is chi-square distributed with $n = \sum_s n_s$ degrees of freedom. Hence using the estimate of $\boldsymbol{\sigma}^2$, $\boldsymbol{\tau}$ and Λ of Table 5.2, together with $\boldsymbol{\alpha}$ for the mean $\boldsymbol{\mu}$, we can calculate a p-value for each case to assess if the case is indeed a mixture in terms of our model. A p-value less than 0.05 indicate that the model does not describe the observed data for that case significantly well. In Chapter 6 we introduce two different Mahalonobis distances to assess the fit of two proposed profiles to a mixed sample. The aim of this test is however to assess information on the fit of the model to data from a DNA STR mixture and not whether the proposed profiles match observed mixture.

In Figure 5.3 we have plotted a histogram for the p-values assuming a χ^2_n-distribution of $(\boldsymbol{M} - T\boldsymbol{\mu})^\top \text{var}(\boldsymbol{M})^{-1}(\boldsymbol{M} - T\boldsymbol{\mu})$. For the model to be supported by data the p-values must be uniform distributed which seems satisfied from inspection of Figure 5.3 and the p-value using Fisher's omnibus test, $-2\sum_{i=1}^n \log(p_i) \sim \chi^2_{2n}$, yields 0.6257 supporting uniformity of the p-values. Also the three cases with a p-value less than 0.05 matches the expected 3.55 which is 5% of 71 cases.

5.2 Variance of parameter estimates

The precision by which the parameters are estimated is a natural measure to include in the evaluation of the model fit. The asymptotic distribution of $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ is a zero-mean normal distribution with the inverse of the expected Fisher Information $\mathcal{I}(\boldsymbol{\theta})$ as covariance matrix. Let $\Delta = \Sigma^{-1}$ and let $\dot{\Delta}_k = \partial\Delta/\partial\boldsymbol{\theta}_k$ with a similar definition for $\dot{\boldsymbol{\mu}}_k$.

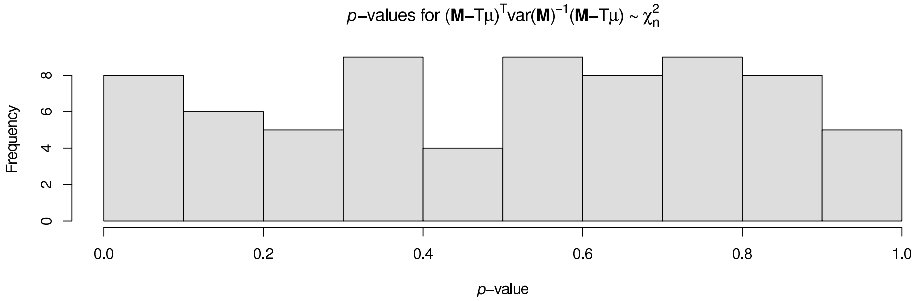

Figure 5.3: Histogram of the p-values computed from $(\boldsymbol{M}-T\boldsymbol{\mu})^\top \mathrm{var}(\boldsymbol{M})^{-1}(\boldsymbol{M}-T\boldsymbol{\mu})$. The distribution looks reasonable uniform supporting the model.

Since $\Delta\Sigma = I$ we have that $\dot{\Delta}_k\Sigma = -\Delta\dot{\Sigma}_k$ and thus $\dot{\Delta}_k = -\Sigma^{-1}\dot{\Sigma}_k\Sigma^{-1}$. Furthermore we have $-\Delta^{-1}\dot{\Delta}_k = \dot{\Sigma}_k\Sigma^{-1}$ but also

$$-\frac{\partial}{\partial\boldsymbol{\theta}_k}\log|\Delta| = \frac{\partial}{\partial\boldsymbol{\theta}_k}\log|\Sigma| = \mathrm{tr}(\Sigma^{-1}\dot{\Sigma}_k) \quad \text{hence} \quad \frac{\partial}{\partial\boldsymbol{\theta}_k}\log|\Delta| = \mathrm{tr}(\Delta^{-1}\dot{\Delta}_k).$$

For a normal distributed $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$ the Fisher information $\mathcal{I}(\boldsymbol{\theta})_{ij}$ can be found by the following arguments. Using the precision Δ the log-likelihood is

$$-2\ell(\boldsymbol{\theta}) = \log(2\pi) - \log|\Delta| + \mathrm{tr}(\Delta(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top).$$

Then differentiation with respect to $\boldsymbol{\theta}_i$ yields

$$\begin{aligned}
-2\dot{\ell}(\boldsymbol{\theta})_i &= -\mathrm{tr}(\Delta^{-1}\dot{\Delta}_i) + \mathrm{tr}(\dot{\Delta}_i[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top] - \Delta[\dot{\boldsymbol{\mu}}_i(\boldsymbol{X}-\boldsymbol{\mu})^\top + (\boldsymbol{X}-\boldsymbol{\mu})\dot{\boldsymbol{\mu}}_i^\top]) \\
-2\ddot{\ell}(\boldsymbol{\theta})_{ij} &= \mathrm{tr}(\Delta^{-1}\dot{\Delta}_j\Delta^{-1}\dot{\Delta}_i) - \mathrm{tr}(\Delta^{-1}\ddot{\Delta}_{ij}) + \mathrm{tr}(\ddot{\Delta}_{ij}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top]) - \\
&\quad \mathrm{tr}(\dot{\Delta}_i[\dot{\boldsymbol{\mu}}_j(\boldsymbol{X}-\boldsymbol{\mu})^\top + (\boldsymbol{X}-\boldsymbol{\mu})\dot{\boldsymbol{\mu}}_j^\top]) - \mathrm{tr}(\dot{\Delta}_j[\dot{\boldsymbol{\mu}}_i(\boldsymbol{X}-\boldsymbol{\mu})^\top + (\boldsymbol{X}-\boldsymbol{\mu})\dot{\boldsymbol{\mu}}_i^\top]) - \\
&\quad \mathrm{tr}(\Delta[\ddot{\boldsymbol{\mu}}_{ij}(\boldsymbol{X}-\boldsymbol{\mu})^\top - \dot{\boldsymbol{\mu}}_i\dot{\boldsymbol{\mu}}_j^\top - \dot{\boldsymbol{\mu}}_j\dot{\boldsymbol{\mu}}_i^\top + (\boldsymbol{X}-\boldsymbol{\mu})\ddot{\boldsymbol{\mu}}_{ij}^\top]).
\end{aligned}$$

Taking expectation on both sides yields,

$$\begin{aligned}
2\mathcal{I}(\boldsymbol{\theta})_{ij} &= \mathrm{tr}(\Delta^{-1}\dot{\Delta}_j\Delta^{-1}\dot{\Delta}_i) - \mathrm{tr}(\Delta^{-1}\ddot{\Delta}_{ij}) + \mathrm{tr}(\Delta^{-1}\ddot{\Delta}_{ij}) + \mathrm{tr}(\Delta[\dot{\boldsymbol{\mu}}_i\dot{\boldsymbol{\mu}}_j^\top + \dot{\boldsymbol{\mu}}_j\dot{\boldsymbol{\mu}}_i^\top]) \\
\mathcal{I}(\boldsymbol{\theta})_{ij} &= \tfrac{1}{2}\mathrm{tr}(\Delta^{-1}\dot{\Delta}_j\Delta^{-1}\dot{\Delta}_i) + \dot{\boldsymbol{\mu}}_i\Delta\dot{\boldsymbol{\mu}}_j^\top
\end{aligned}$$

Substituting Σ^{-1} for Δ using the relations from above gives the final expression

$$\mathcal{I}(\boldsymbol{\theta})_{ij} = \frac{\partial\mu(\boldsymbol{\theta})^\top}{\partial\boldsymbol{\theta}_i}\Sigma(\boldsymbol{\theta})^{-1}\frac{\partial\mu(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j} + \frac{1}{2}\mathrm{tr}\left(\Sigma(\boldsymbol{\theta})^{-1}\frac{\partial\Sigma(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i}\Sigma(\boldsymbol{\theta})^{-1}\frac{\partial\Sigma(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_j}\right), \tag{5.1}$$

where

$$\frac{\partial\mu(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_k} = \left(\frac{\partial\mu_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_k}\right)_{i=1}^{p} \quad \text{and} \quad \frac{\partial\Sigma(\boldsymbol{\theta})}{\partial\theta_k} = \left(\frac{\partial\Sigma_{i,j}(\boldsymbol{\theta})}{\partial\theta_k}\right)_{i,j=1}^{p}.$$

In our setting we have $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda)$ where $\boldsymbol{M}_c \sim \mathcal{N}(\mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$, with

$$\begin{aligned}
\mu(\boldsymbol{\theta}) &= T_c \mathrm{diag}(\boldsymbol{H}_c) Q \boldsymbol{\alpha} \\
\Sigma(\boldsymbol{\theta}) &= T_c \mathrm{diag}\left\{\mathrm{diag}(\boldsymbol{H}_c) Q \boldsymbol{\sigma}^2\right\} T_c^\top + \mathrm{diag}(Q_c \boldsymbol{\tau}) + Q_c \Lambda Q_c^\top = \mathrm{var}(\boldsymbol{M}_c).
\end{aligned}$$

This implies that all differentiations of $\mu(\boldsymbol{\theta})$ with respect to $(\boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda)$ is zero and also $\partial\Sigma(\boldsymbol{\theta})/\partial\boldsymbol{\alpha} = 0$. Since our data contains 10 systems, we have $10_{\boldsymbol{\alpha}} + 10_{\boldsymbol{\sigma}^2} + 10_{\boldsymbol{\tau}} + 55_\Lambda = 85$ parameters and therefore the Fisher Information is a 85×85 matrix.

We have that

$$\frac{\partial \mu(\boldsymbol{\theta})}{\partial \alpha_i} = \frac{\partial}{\partial \alpha_i} T_c \mathrm{diag}(\boldsymbol{H}) Q \boldsymbol{\alpha} = T_c \mathrm{diag}(\boldsymbol{H}) Q \boldsymbol{e}_S(i),$$

and as mentioned $\mathbf{0}$ for all other differentiations. Next we see that,

$$\begin{aligned}
\frac{\partial \Sigma(\boldsymbol{\theta})}{\partial \sigma_i^2} &= \frac{\partial}{\partial \sigma_i^2} T_c \mathrm{diag}\left\{\mathrm{diag}(\boldsymbol{H}) Q \boldsymbol{\sigma}^2\right\} T_c^\top = T_c \mathrm{diag}\left\{\mathrm{diag}(\boldsymbol{H}) Q \boldsymbol{e}_S(i)\right\} T_c^\top \\
\frac{\partial \Sigma(\boldsymbol{\theta})}{\partial \tau_i} &= \frac{\partial}{\partial \tau_i} \mathrm{diag}(Q_c \boldsymbol{\tau}) = \mathrm{diag}(Q_c \boldsymbol{e}_S(i)) \\
\frac{\partial \Sigma(\boldsymbol{\theta})}{\partial \nu_{ij}} &= \frac{\partial}{\partial \nu_{ij}} Q_c \Lambda Q_c^\top = Q_c \left[\left\{\boldsymbol{e}_S(i)\boldsymbol{e}_S(j)^\top + \boldsymbol{e}_S(j)\boldsymbol{e}_S(i)^\top\right\}/(1+\delta_{ij})\right] Q_c^\top,
\end{aligned}$$

where $i, j = 1, \ldots, S$ and δ_{ij} is Kronecker's delta. Since the Fisher information with respect to $\boldsymbol{\alpha}$ is independent of $\boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda$ and vice versa we can invert these blocks separately for determine the asymptotic variance of the parameters,

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{bmatrix} \mathcal{I}(\boldsymbol{\alpha}) & O \\ O & \mathcal{I}(\boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda) \end{bmatrix} \quad \text{and} \quad \mathcal{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} \mathcal{I}(\boldsymbol{\alpha})^{-1} & O \\ O & \mathcal{I}(\boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda)^{-1} \end{bmatrix}.$$

Since we assume independence across cases the total Fisher information is just the sum over the Fisher information for each case, $\mathcal{I}(\boldsymbol{\theta}) = \sum_c \mathcal{I}_c(\boldsymbol{\theta})$. For a fixed case the $\mathcal{I}_c(\boldsymbol{\alpha})$ can now be found as,

$$\mathcal{I}_c(\boldsymbol{\alpha})_{ij} = \boldsymbol{e}_S(i)^\top Q^\top \mathrm{diag}(\boldsymbol{H}_c) T_c^\top \mathrm{var}(M_c)^{-1} T_c \mathrm{diag}(\boldsymbol{H}_c) Q \boldsymbol{e}_S(j).$$

Note the simple form of $\frac{\partial}{\partial \alpha_j} \boldsymbol{\mu}(\boldsymbol{\theta})$

$$\begin{aligned}
T_c \mathrm{diag}(\boldsymbol{H}_c) Q \boldsymbol{e}_S(j) &= T_c \mathrm{diag}(\boldsymbol{H})[\underbrace{0, \ldots, 0}_{4(j-1)}, \mathbf{1}_4^\top, \underbrace{0, \ldots, 0}_{4(S-j)}]^\top \\
&= [\underbrace{0, \ldots, 0}_{n_1 + \cdots + n_{j-1}}, \underbrace{(T_j \boldsymbol{H}_j)^\top}_{n_j}, \underbrace{0, \ldots, 0}_{n_{j+1} + \cdots + n_S}]^\top,
\end{aligned}$$

where $T_j \boldsymbol{H}_j$ adds together the $H_c^{(k)}$s contributing to the same alleles. Hence $\mathcal{I}_c(\boldsymbol{\alpha})_{ij}$ simplifies to

$$\mathcal{I}_c(\boldsymbol{\alpha})_{ij} = (T_i \boldsymbol{H}_i)^\top \mathrm{var}(M)^{-1}_{ij} T_j \boldsymbol{H}_j$$

We can partition $\mathcal{I}_c(\boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda)$ as

$$\mathcal{I}_c(\boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda) = \begin{bmatrix} \mathcal{I}_c(\boldsymbol{\sigma}^2) & \mathcal{I}_c(\boldsymbol{\sigma}^2, \boldsymbol{\tau}) & \mathcal{I}_c(\boldsymbol{\sigma}^2, \Lambda) \\ \mathcal{I}_c(\boldsymbol{\sigma}^2, \boldsymbol{\tau}) & \mathcal{I}_c(\boldsymbol{\tau}) & \mathcal{I}_c(\boldsymbol{\tau}, \Lambda) \\ \mathcal{I}_c(\boldsymbol{\sigma}^2, \Lambda) & \mathcal{I}_c(\boldsymbol{\tau}, \Lambda) & \mathcal{I}_c(\Lambda) \end{bmatrix}$$

where the symmetry is secured since the trace operator is symmetric and the first term of (5.1) is zero. Using the above expressions we have,

$$\begin{aligned} \mathcal{I}_c(\boldsymbol{\sigma}^2)_{ij} &= \operatorname{tr}\left(\operatorname{var}(\boldsymbol{M}_c)^{-1}\frac{\partial\Sigma(\boldsymbol{\theta})}{\partial\sigma_i^2}\operatorname{var}(\boldsymbol{M}_c)^{-1}\frac{\partial\Sigma(\boldsymbol{\theta})}{\partial\sigma_j^2}\right) \quad \text{where,} \\ \frac{\partial\Sigma(\boldsymbol{\theta})}{\partial\sigma_i^2} &= T_c\operatorname{diag}\left\{\operatorname{diag}(\boldsymbol{H}_c)Q\boldsymbol{e}_S(i)\right\}T_c^\top = T_c\operatorname{diag}\left\{[\mathbf{0}_{4(i-1)}, \boldsymbol{H}_{ic}, \mathbf{0}_{4(s-i)}]^\top\right\}T_c^\top. \end{aligned}$$

The latter expression is just a matrix product of two block diagonal matrices and a diagonal. From hereon we drop the case subscript c to keep the expressions simpler except for Q_c to distinguish it from Q. Now let $Z^i = \partial\Sigma(\boldsymbol{\theta})/\partial\sigma_i^2$, then we have,

$$Z^i = \operatorname{diag}(O, \ldots, O, Z_i, O, \ldots, O) \quad \text{with } Z_i = T_i\operatorname{diag}(\boldsymbol{H}_i)T_i^\top \text{ in the } i\text{'th block.}$$

Since $\operatorname{var}(\boldsymbol{M})^{-1}$ also have a block structure, $\operatorname{var}(M)_{ij}^{-1}$, the trace therefore becomes a sum expressed as,

$$\mathcal{I}(\boldsymbol{\sigma}^2)_{ij} = \operatorname{tr}\left(\operatorname{var}(\boldsymbol{M})^{-1}Z^i\operatorname{var}(M)^{-1}Z^j\right) = \sum_{k,l,m,n}\operatorname{tr}\left(\operatorname{var}(\boldsymbol{M})_{km}^{-1}Z_{ml}^i\operatorname{var}(\boldsymbol{M})_{ln}^{-1}Z_{nk}^j\right).$$

Since $Z_{ml}^i = 0$ unless $m = l = i$ with $Z_{ii}^i = Z_i$ all terms in the sum is zero but for $m = l = i$ and $k = n = j$. Thus we have

$$\mathcal{I}(\boldsymbol{\sigma}^2)_{ij} = \operatorname{tr}\left(\operatorname{var}(\boldsymbol{M})_{ji}^{-1}Z_i\operatorname{var}(\boldsymbol{M})_{ij}^{-1}Z_j\right).$$

When determining the next diagonal element of $\mathcal{I}(\boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda)$ we can use the same reasoning as above. Hence $\mathcal{I}(\boldsymbol{\tau})_{ij}$ yield,

$$\mathcal{I}(\boldsymbol{\tau})_{ij} = \operatorname{tr}\left(\operatorname{var}(M)^{-1}\operatorname{diag}(Q_c\boldsymbol{e}_S(i))\operatorname{var}(M)^{-1}\operatorname{diag}(Q_c\boldsymbol{e}_S(j))\right).$$

Here $\operatorname{diag}(Q_c\boldsymbol{e}_S(j))$ is simply a zero matrix with I_{n_i} on the i'th block of the diagonal. Hence using the trace expansion again we have,

$$\mathcal{I}(\boldsymbol{\tau})_{ij} = \operatorname{tr}\left(\operatorname{var}(\boldsymbol{M})_{ji}^{-1}I_{n_i}\operatorname{var}(\boldsymbol{M})_{ij}^{-1}I_{n_j}\right) = \operatorname{tr}\left(\operatorname{var}(\boldsymbol{M})_{ji}^{-1}\operatorname{var}(\boldsymbol{M})_{ij}^{-1}\right).$$

Since Λ is symmetric we need only ν_{st} with $1 \leq s \leq t \leq S$ of Λ. That is

$$\begin{bmatrix} \nu_{1.1} & \nu_{1.2} & \cdots & \nu_{1.S} \\ \nu_{2.1} & \nu_{2.2} & \cdots & \nu_{2.S} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_{S.1} & \nu_{S.2} & \cdots & \nu_{S.S} \end{bmatrix} \mapsto [\nu_{1.1}, \ldots, \nu_{1.S}, \nu_{2.2}, \ldots, \nu_{2.S}, \ldots, \nu_{S-1.S-1}, \nu_{S-1.S}, \nu_{S.S}]^\top = \boldsymbol{\nu},$$

hence $\mathcal{I}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda) = \mathcal{I}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}, \boldsymbol{\nu})$. In row k of Λ we use $S - (k-1)$ elements. Hence the index number of $\nu_{k,S}$ in $\boldsymbol{\nu}$, $\boldsymbol{i}(k,S)$, is determined as, $\boldsymbol{i}(k,S) = \sum_{i=1}^{k} S - (i-1)$ with $k \leq S$. Hence $\boldsymbol{i}(k+1,k+1) = \boldsymbol{i}(k,S) + 1$ from which we have $\boldsymbol{i}(k+1,l) = \boldsymbol{i}(k+1,k+1) + (l-(k+1))$ where $k \leq l \leq S$. Hence we have the general expression,

$$\boldsymbol{i}(k,l) = S(k-1) - k(k-1)/2 + l.$$

This implies that $\boldsymbol{\nu}_{\boldsymbol{i}(k,l)} = \nu_{k,l}$. The final diagonal term $\mathcal{I}_c(\Lambda)$ is found by

$$\mathcal{I}(\Lambda)_{\boldsymbol{i}(k,l)\boldsymbol{i}(m,n)} = \operatorname{tr}\left(\operatorname{var}(\boldsymbol{M})^{-1}\frac{\partial}{\partial \nu_{kl}} Q_c \Lambda Q_c^\top \operatorname{var}(\boldsymbol{M})^{-1} \frac{\partial}{\partial \nu_{mn}} Q_c \Lambda Q_c^\top\right).$$

Now evaluating the right hand side using the expression for $\partial\Sigma(\boldsymbol{\theta})/\partial\nu_{kl}$ we have for $k \neq l$,

$$\begin{aligned}
\frac{\partial}{\partial \nu_{kl}} Q_c \Lambda Q_c^\top &= (\underbrace{0,\ldots,0}_{n_1+\cdots+n_{k-1}}, \mathbf{1}_{n_k}^\top, \underbrace{0,\ldots,0}_{n_{k+1}+\cdots+n_S})^\top (\underbrace{0,\ldots,0}_{n_1+\cdots+n_{l-1}}, \mathbf{1}_{n_l}^\top, \underbrace{0,\ldots,0}_{n_{l+1}+\cdots+n_S}) + \\
&\quad (\underbrace{0,\ldots,0}_{n_1+\cdots+n_{l-1}}, \mathbf{1}_{n_l}^\top, \underbrace{0,\ldots,0}_{n_{l+1}+\cdots+n_S})^\top (\underbrace{0,\ldots,0}_{n_1+\cdots+n_{k-1}}, \mathbf{1}_{n_k}^\top, \underbrace{0,\ldots,0}_{n_{k+1}+\cdots+n_S}) \\
&= C^{kl} = C^{kl}_{qr} = \begin{cases} \mathbf{1}_{n_k}\mathbf{1}_{n_l}^\top, & q = k \text{ and } r = l \\ \mathbf{1}_{n_l}\mathbf{1}_{n_k}^\top, & q = l \text{ and } r = k \\ O, & \text{otherwise.} \end{cases}
\end{aligned}$$

Inserting this in $\mathcal{I}_c(\Lambda)_{\boldsymbol{i}(k,l)\boldsymbol{i}(m,n)}$ yields,

$$\mathcal{I}_c(\Lambda)_{\boldsymbol{i}(k,l)\boldsymbol{i}(m,n)} = \sum_{p,q,r,s} \operatorname{tr}\left(\operatorname{var}(\boldsymbol{M})^{-1}_{pq} C^{kl}_{qr} \operatorname{var}(\boldsymbol{M})^{-1}_{rs} C^{mn}_{sp}\right) = \frac{2w_{nk}w_{lm} + 2w_{nl}w_{km}}{(1+\delta_{kl})(1+\delta_{mn})},$$

where $w_{rs} = \mathbf{1}_{n_r}^\top \operatorname{var}(\boldsymbol{M})^{-1}_{rs}\mathbf{1}_{n_s}$ with $w_{rs} = w_{sr}$.

For the off-diagonal terms in $\mathcal{I}_c(\boldsymbol{\sigma}^2, \boldsymbol{\tau}, \Lambda)$ we can use the expressions derived above to get the following entrances,

$$\begin{aligned}
\mathcal{I}_c(\boldsymbol{\sigma}^2, \boldsymbol{\tau})_{ij} &= \operatorname{tr}\left(\operatorname{var}(\boldsymbol{M})^{-1}_{ji} Z_i \operatorname{var}(\boldsymbol{M})^{-1}_{ij}\right) \\
\mathcal{I}_c(\boldsymbol{\sigma}^2, \Lambda)_{i\boldsymbol{i}(k,l)} &= (\mathbf{1}_{n_l}^\top \operatorname{var}(\boldsymbol{M})^{-1}_{li} Z_i \operatorname{var}(\boldsymbol{M})^{-1}_{ik}\mathbf{1}_{n_k} + \mathbf{1}_{n_k}^\top \operatorname{var}(\boldsymbol{M})^{-1}_{ki} Z_i \operatorname{var}(\boldsymbol{M})^{-1}_{il}\mathbf{1}_{n_l})/(1+\delta_{kl}) \\
\mathcal{I}_c(\boldsymbol{\tau}, \Lambda)_{i\boldsymbol{i}(k,l)} &= (\mathbf{1}_{n_l}^\top \operatorname{var}(\boldsymbol{M})^{-1}_{li} \operatorname{var}(\boldsymbol{M})^{-1}_{ik}\mathbf{1}_{n_k} + \mathbf{1}_{n_k}^\top \operatorname{var}(\boldsymbol{M})^{-1}_{ki} \operatorname{var}(\boldsymbol{M})^{-1}_{il}\mathbf{1}_{n_l})/(1+\delta_{kl})
\end{aligned}$$

which completes the Fisher information matrix.

The asymptotic variances and correlations estimated by the Fisher information are given in Table 5.4 and Table 5.5 for the full and restricted ($\boldsymbol{\tau} = \mathbf{0}$) models, respectively.

For the full model (with $\boldsymbol{\tau}$ not set to zero) the correlations seems constant across systems. Hence we may summarize this correlation structure in a matrix form,

$$\begin{array}{c} \boldsymbol{\sigma}^2 \\ \boldsymbol{\tau} \\ \Lambda_{ss} \end{array} \begin{bmatrix} 1 & -0.7 & 0.15 \\ -0.7 & 1 & -0.5 \\ 0.15 & -0.5 & 1 \end{bmatrix}$$

Table 5.4: Standard deviation and correlation of parameters fitted for the full model. We denote diag(Λ) as $\boldsymbol{\nu}_{ss}$ due to lack of space.

	Standard deviations				Correlation		
	α	σ^2	τ	ν_{ss}	σ^2, τ	σ^2, ν_{ss}	τ, ν_{ss}
D16	0.147	223.608	108771.53	83843.23	–0.711	0.159	–0.498
D18	0.168	387.382	155198.38	92326.20	–0.721	0.080	–0.425
D19	0.119	126.066	61039.40	55008.97	–0.725	0.216	–0.545
D2	0.135	146.348	58264.26	46547.08	–0.769	0.180	–0.447
D21	0.125	201.491	76094.18	42136.96	–0.736	0.076	–0.404
D3	0.118	152.801	60907.14	39453.75	–0.745	0.130	–0.433
D8	0.144	236.997	101856.93	75112.52	–0.747	0.161	–0.475
FGA	0.087	68.134	28800.48	21892.66	–0.734	0.158	–0.474
TH0	0.101	91.615	40686.60	34805.48	–0.758	0.232	–0.540
vWA	0.113	128.337	52310.19	35328.37	–0.734	0.125	–0.422

Table 5.5: Standard deviation and correlation of parameters fitted for the restricted model with $\boldsymbol{\tau} = \mathbf{0}$.

	Standard deviations			Correlation
	α	σ^2	ν_{ss}	σ^2, ν_{ss}
D16	0.148	157.194	73453.16	–0.316
D18	0.171	267.388	86516.20	–0.345
D19	0.119	86.779	45871.55	–0.312
D2	0.135	93.454	41423.60	–0.288
D21	0.125	136.206	38584.17	–0.356
D3	0.118	101.766	35518.81	–0.321
D8	0.144	157.494	66067.68	–0.330
FGA	0.087	46.237	19334.22	–0.315
TH0	0.101	59.743	29252.41	–0.323
vWA	0.112	86.949	31957.80	–0.299

Note that the strongest correlations are between $\boldsymbol{\tau}$ and the two other variance parameters, $\boldsymbol{\sigma}^2$ and Λ.

Apart from the insignificant p-value from the approximate χ^2-test these large correlations and the smaller standard deviations on the parameters in the reduced model support eliminating $\boldsymbol{\tau}$. Since the parameters are based on a limited training set, $\boldsymbol{\tau}$ might be estimated significant different from $\mathbf{0}$ using a more representative data set.

CHAPTER 6

Evidence

Section of Forensic Genetics, University of Copenhagen, has provided data from 74 real crime cases. Under real circumstances the contributing profiles are not known in advance and therefore there are some additional uncertainty attached to these cases. In eight of these cases we have observed one or several drop-outs and they are therefore excluded from this analysis. The 66 remaining cases will be analyzed in the following.

6.1 Real crime cases

When drawing conclusions based on the p-value derived from the Mahalanobis distance with $\boldsymbol{\mu}$ and var$(\boldsymbol{M})$ assumed known, we have to bare in mind that the estimated values of these two quantities is based on pairwise mixtures of only four DNA profiles. This causes bias towards the represented alleles in the parameter estimates. Also the analyzed mixtures from the controlled experiments are performed on the same machinery and thus the amplification results reflect this machine's specific behaviour.

Let $\overline{\boldsymbol{M}}_c$ be the system-wise sum over alleles such that $\overline{\boldsymbol{M}}_c = (\sum_{i=1}^{n_1} \boldsymbol{M}_{1_i}, \dots, \sum_{i=1}^{n_S} \boldsymbol{M}_{S_i})$. Using the matrix notation from before this can be expressed as $\overline{\boldsymbol{M}}_c = Q_c^\top \boldsymbol{M}_c$, hence the distribution of $\overline{\boldsymbol{M}}_c$ is

$$\overline{\boldsymbol{M}}_c \sim \mathcal{N}\left(Q^\top \boldsymbol{\mu}_c, Q^\top DQ + \text{diag}(\boldsymbol{n}_c\boldsymbol{\tau}) + \text{diag}(\boldsymbol{n}_c)\Lambda\text{diag}(\boldsymbol{n}_c)\right),$$

where we used $Q_c^\top T_c = Q^\top$, $Q_c^\top \text{diag}(Q_c\boldsymbol{\tau})Q_c = Q_c^\top Q_c \text{diag}(\boldsymbol{\tau})$ and $Q_c^\top Q_c = \text{diag}(\boldsymbol{n}_c)$. Since the Mahalanobis distance considered in the previous chapter takes both the mixture and system balances into account it is likely to yield low p-values for real crime cases. System unbalances caused by degraded DNA and other sources of contamination are likely to affect the amplification of the different systems and therefore reject the hypothesis of a mixture of the two proposed profiles. However conditioning on the system sums $\overline{\boldsymbol{M}}_c$ we can evaluate the match of the two profiles within each system.

In order to find the distribution of $\boldsymbol{M}|\overline{\boldsymbol{M}}$ we need to specify the covariance of $(\boldsymbol{M}, \overline{\boldsymbol{M}})$,

$$\Sigma(\boldsymbol{M}, \overline{\boldsymbol{M}}) = \begin{bmatrix} TDT^\top + \Sigma & (TDT^\top + \Sigma)Q_c \\ Q_C^\top(TDT^\top + \Sigma) & Q_c^\top(TDT^\top + \Sigma)Q_c \end{bmatrix},$$

since $\text{cov}(\boldsymbol{M}, \overline{\boldsymbol{M}}) = \text{cov}(\boldsymbol{M}, Q_c^\top \boldsymbol{M}) = \text{var}(\boldsymbol{M})Q_c$. Now the conditional distribution of $\boldsymbol{M}$ given $\overline{\boldsymbol{M}}$ is degenerated since conditioning imply that the covariance matrix is not of full rank,

$$\begin{aligned} \text{E}(\boldsymbol{M}|\overline{\boldsymbol{M}}) &= T\boldsymbol{\mu} + (TDT^\top + \Sigma)Q_c\left[Q_c^\top(TDT^\top + \Sigma)Q_c\right]^{-1}(\overline{\boldsymbol{M}} - Q^\top\boldsymbol{\mu}) \\ \text{cov}(\boldsymbol{M}|\overline{\boldsymbol{M}}) &= TDT^\top + \Sigma - (TDT^\top + \Sigma)Q_c\left[Q_c^\top(TDT^\top + \Sigma)Q_c\right]^{-1}Q_c^\top(TDT^\top + \Sigma). \end{aligned}$$

However using the generalized inverse $\mathrm{cov}(\boldsymbol{M}|\overline{\boldsymbol{M}})^-$ we can determine the Mahalanobis distances,

$$\left(\boldsymbol{M} - \mathrm{E}(\boldsymbol{M}|\overline{\boldsymbol{M}})\right)^\top \mathrm{cov}(\boldsymbol{M}|\overline{\boldsymbol{M}})^- \left(\boldsymbol{M} - \mathrm{E}(\boldsymbol{M}|\overline{\boldsymbol{M}})\right) \sim \chi^2_{n_c-S} \tag{6.1}$$

$$(\overline{\boldsymbol{M}} - Q^\top \boldsymbol{\mu})^\top \left[Q_c^\top (TDT^\top + \Sigma) Q_c\right]^- (\overline{\boldsymbol{M}} - Q^\top \boldsymbol{\mu}) \sim \chi^2_S. \tag{6.2}$$

The p-values determined from these Mahalanobis distances can be interpreted as a two-step evaluation of the DNA mixture. The interpretation of p-values from (6.1) is whether the two profiles are the likely contributors to the mixture. Since we condition on the possible unbalances between systems low p-values is an indication of a low agreement of the observed relation, and the one explained by the pair of proposed profiles. However if this p-value is above some fixed value, e.g. 0.01 say, we have no evidence for excluding the mixture as a possible explanation.

Assuming that the p-value from (6.1) is significantly different from zero, we assume the proposed profiles are a possible explanation of the observed mixture. Therefore we assume that $\overline{\boldsymbol{M}}$ is the system sums over a mixture of the two proposed profiles and the assumptions that imply the validity of (6.2) are met. Under this assumption the interpretation of the p-value from (6.2) is the quality of the available sample. A low p-value indicates unbalances between systems and conclusions should therefore be made with extra caution. Since our model does not incorporate the possibility of DNA to be degraded and other unbalances between systems the evidence with a low p-value is of limited use.

As expected did all cases from the controlled experiments have high p-values based on both (6.1) and (6.2). This indicates that for all cases the correct profiles were mixed and also that there were no abnormal system unbalances.

In the real crime cases only three cases indicated a poor fit between the observed and proposed mixture. In Figure 6.1 we have plotted the observed peak areas for the three

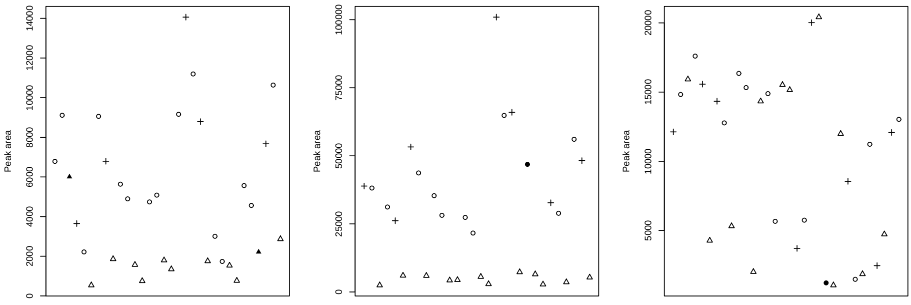

Figure 6.1: The three cases with p-values less than 0.01. The point characters indicates the donors, $\circ$ and $\triangle$, with $+$ being a shared allele and the filled versions being non-shared homozygote alleles. The ordering along the first axis is random.

cases respectively. The panels indicate large within person variability for the peak areas not supporting the proposed profiles as donors to the mixture.

In Figure 6.2 we have plotted histograms for the p-values from (6.1) and (6.2) evaluated for the real crime cases. We have only included the (6.2) p-values from cases where the (6.1) p-values were greater than 0.01.

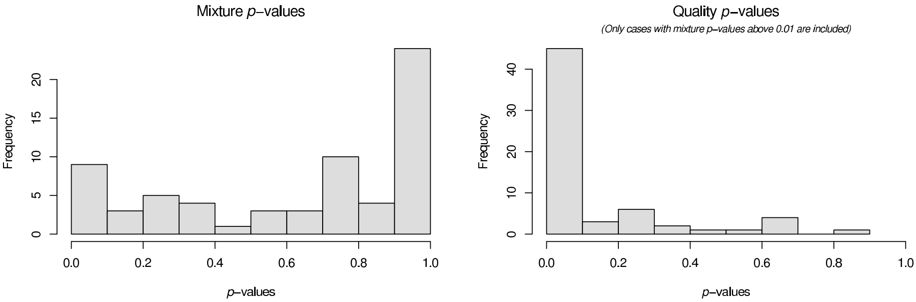

Figure 6.2: Histograms of the p-values from (6.1) and (6.2) for the real crime cases. In the right panel only cases with mixture p-values above 0.01 are included.

The non-uniformity of the p-values in the two histograms indicates that the model does not fit perfectly to the data. For the right panel of Figure 6.2 the histogram shows that a majority of cases have low p-value with 35 cases having a p-value less than 0.01. This may indicate that several of the cases have a low copy number or degraded DNA or simply that the model is too simple to cope with real world data.

The person-water mixtures included in the controlled experiments can be used as additional data for verification of the model. These cases were excluded before parameter estimation due to the variability observed for person-water mixtures during the data analysis summarized in Chapter 2. Computing both the Mahalanobis distances of Chapter 5 and those of (6.2) and (6.1) indicates that the model also has a reasonable fit to single contributor cases. The histograms of p-values from the three Mahalanobis distances in Figure 6.3 show uniformity supporting the goodness of fit.

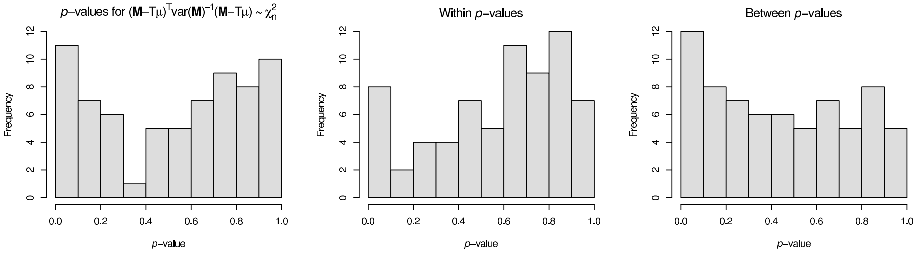

Figure 6.3: p-values from Mahalanobis distances for single contributor cases. The uniformity indicates the model fits well for one-person samples.

CHAPTER 7

Epilogue

In this chapter we summarize, discuss and conclude on the work presented in the present thesis. Furthermore we set out some possible problems for future work within this framework for mixed DNA samples.

7.1 Discussion

As mentioned several times throughout the present thesis the interpretations made about model structure and parameter estimates are biased towards the four profiles included in our data set. For the model to be sufficiently supported by data a vaster data set with several profiles must be used for estimation of parameters. This is both due to the allelic variability but also the general variable nature of DNA STR amplifications as seen from the present data set. Another source of variation not represented by the available data is machine effect as all samples are analyzed using the same machinery. This is an important factor to include in order having confidence in the estimates of the parameters. It would be interesting to perform a meta analysis comparing estimates based on data from different machines in order to evaluate the robustness of the estimates.

The data used for the parameter estimates were processed by the forensic laboratory using no stutter filter but still a fixed threshold for the peak heights. This threshold excludes all peaks with lower peak heights than 50 RFU and introduces a kind of censoring depending on the (un-)observed peak height. The problem with such a threshold is instead of having a peak area observation of e.g. 500 we may instead register a drop-out. In Gilder et al. (2007) they discuss the use of run specific threshold determined by a model based on the white noise of the machinery. This seems like a more reasonable approach since cases with low contributions from a donor may result in many drop-outs using a fixed threshold.

On the preceding semester we examined the empirical inter-correlation structure for the systems. By stratifying on the trial number (each mixture were separately analyzed two times) we found that some correlations were significantly different in both magnitude and sign. This may explain some of the uncertainty of the estimates of the covariance structure Λ.

In addition to the runs with initial values specified in Table 5.1 we also performed runs with $\boldsymbol{\sigma}^2_{(0)} = \mathbf{0}$. However since the estimator of $\boldsymbol{\sigma}^2_{(n+1)}$ is not dependent of the previous values it may eventually converge towards the $\boldsymbol{\sigma}$ estimates found from the other initial values. In Figure 7.1 we have plotted the traces for $\boldsymbol{\sigma}^2$ and $\boldsymbol{\tau}$ with $\boldsymbol{\sigma}^2$ initialized to $\mathbf{0}$. After approximately 17000 iterations the parameters changes and together with them the deviance drops (not included in the plot). The plot supports the insignificance of $\boldsymbol{\tau}$ as it converges to relative small values compared to both $\boldsymbol{\sigma}^2$ and diag(Λ).

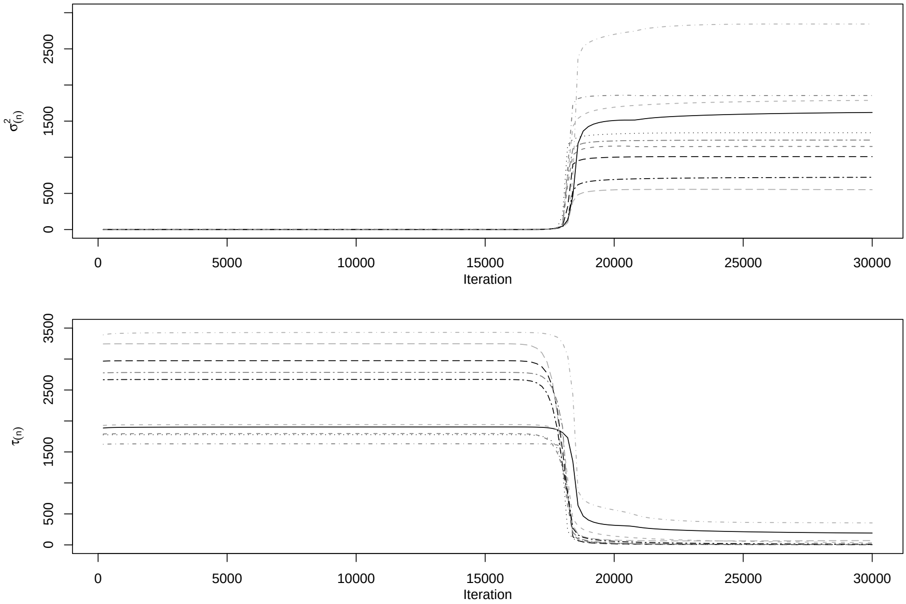

Figure 7.1: Traces for $\boldsymbol{\sigma}^2_{(n)}$ and $\boldsymbol{\tau}_{(n)}$ with $\boldsymbol{\sigma}^2$ initialized as $\mathbf{0}$ with a line for each system.

As mentioned in Section 4.4 we only used data with a PCR injection time of six seconds. On the preceding semester we used a simple linear regression model to predict the peak areas by the amount of DNA. By stratisfying on systems and dropping insignificant terms, such as laboratorian-effect, the final model reduced to $\text{Area}_s = (\alpha_s + \mathbb{I}_{12}(t)\beta_s)\text{DNA}$, where $\mathbb{I}_{12}$ is the indicator function for the injection time being 12 seconds. In addition we found an approximate linear relation of β_s and α_s as $\beta_s = 0.146\alpha_s$ for all systems. Hence the ratio of samples with different injection time is 1.146 or equivalent $\text{Area}_{12} = 1.146\text{Area}_6$ not depending on the system. However since our model includes $\boldsymbol{H}$ in both mean and variance we expect the estimates of both $\boldsymbol{\alpha}$ and the variance components $\boldsymbol{\sigma}^2$ and Λ to be identical for the different injection times.

In Figure 7.2 we have plotted the estimates based on six and 12 seconds data. The panels indicate that $\boldsymbol{\alpha}$ is independent of the injection time in contrast to the variance components. We find that the variance components for 12 seconds are approximately 1.5 times larger than for the six seconds data. This indicates that the variance is not linear in injection time. This non-linearity may be due possible saturation of the machinery occurring more frequently for longer injection tim es relative to shorter injection times. This shows that the parameters used in assessing the weight of evidence depend on the injection time.

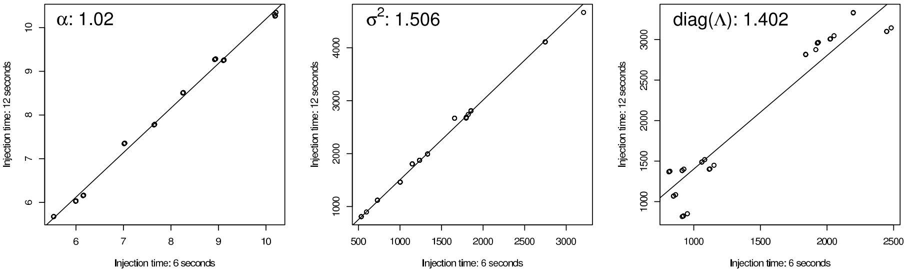

Figure 7.2: Parameters determined for data with six and 12 seconds injection time. We see the variance components have increased by a half whereas α seems constant.

7.2 Future work

In this section we discuss issues and problems subject to future work since the current model is not applicable in all real mixture cases. First we evaluate the current model and the parameters reported in Chapter 5.

7.2.1 Model reductions

In Chapter 5 we showed that $\boldsymbol{\tau}$ was insignificant with respect to the present data. By doing so we placed more of the variance on the v_{ss} components and $\boldsymbol{\sigma}^2$ but also reduced the variance on these parameter estimates. However the standard deviations reported in Table 5.5 indicates that data does not support having a covariance for each inter- and intrasystem combination and calls for further model reductions.

Since the components of $\boldsymbol{\sigma}^2$ seems be ordered by dye band colour this may also apply to Λ, hence a possible covariance structure could be,

$$\begin{bmatrix} v_{YY} & v_{YB} & v_{YG} \\ v_{YB} & v_{BB} & v_{BG} \\ v_{YG} & v_{BG} & v_{GG} \end{bmatrix}. \tag{7.1}$$

This allows systems on the same dye band to share covariance parameters. The parameters fitted for the model with Λ specified as in (3.7) does not indicate that covariance between systems on the same dye band should be similar. However with the uncertainty attached to the parameters we can not reject this covariance structure. The reduction to (7.1) imply that instead of estimating $S(S+1)/2$ parameters for the covariance matrix we need only to estimate 6 parameters. Hence with 10 systems as in our data set we drop 49 parameters.

An even simpler covariance structure assumes equal within and between covariance for all systems,

$$\begin{bmatrix} v & \gamma \\ \gamma & v \end{bmatrix}.$$

This two parameter covariance structure may cause $\boldsymbol{\tau}$ to be significantly different from $\mathbf{0}$ in order to have a reasonable fit. However compared to the model with $\Lambda = \nu_{st}$ and $\boldsymbol{\tau} = \mathbf{0}$ we still reduce the number of parameters by 43.

7.2.2 Model limitations

The model described in Chapter 3 works under several limitations which makes it limited for immediate use in real world crime cases.

Degraded DNA

Since we assume the mean (and variance) of $\boldsymbol{A}$ to be proportional to the DNA concentration mimicked by $\boldsymbol{H}$ the only system-wise differences possible of the mean is captured in $\boldsymbol{\alpha}$. For non-degraded DNA, as in the controlled experiments, we found that this structure is sufficient for describing the different amplification properties of systems. Exposing DNA to different kinds of inhibitors increases the probability of the DNA sequences to break into shorter structures. This damage imposed to the DNA may cause alleles to drop-out or in milder degree imply lower amplification than expected. Experts in forensic DNA expect longer sequences to have a higher probability of breakage than shorter sequences and by the nature of STR DNA therefore systems to have different risks due to the allelic ladder in Figure A.1. For inclusion of DNA degradation in the model we need to have a separate model of the degradation behaviour for the different systems and alleles. In Figure 2.3 we have indications of the expected levels of amplification for each system. Dividing $\overline{\boldsymbol{M}}$ by $\boldsymbol{\alpha}$ imply that the expectation of all systems to be the same constant, $2(H^{(1)} + H^{(2)})$. Now an analysis of variance can be used as an approximation to test for differences across systems and thereby indicate if the DNA has degraded. The dependence between and within systems causes the method to be approximate. Note that the within system variation is of less importance in this analysis since we expect a significant difference between individual peaks for mixtures with $H^{(1)}/H^{(2)}$ different from one.

Applying this approach to the data from the controlled experiments imply a few cases to have p-values less than 0.05. Further investigation show that all these cases is mixtures of person A and B, both homozygote on system D18 sharing allele 13. Removing D18 from all cases and reanalyzing by the same means none of the controlled experiments had p-values less than 0.10. This indicates that the amplification behaviour of systems where two homozygous share an allele is not captured sufficiently by the model.

Degradation of DNA may be modeled using the methodology of survival data analysis as lower amplification than expected is due to failure in amplification of some the DNA material. One can interpret the observed peak areas as the proportion of material *surviving* the degradation. In Figure 7.3 we have plotted the fragment length for some real crime cases against the aggregated sums of peak areas weight by the reciprocal $\boldsymbol{\alpha}$ estimates.

In Figure 7.3 we see the reasonable good fit to a linear curve for the log-transformed

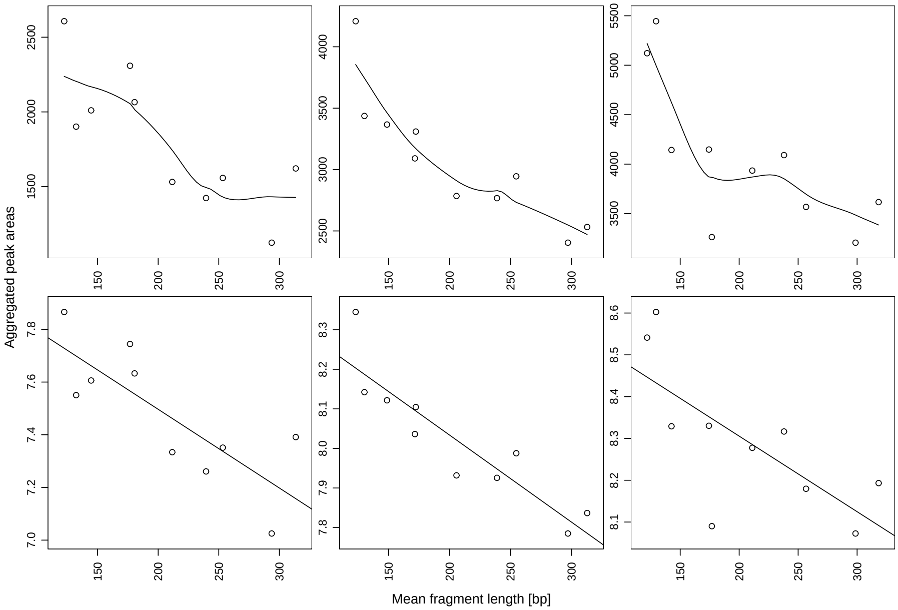

Figure 7.3: Aggregated peak areas for each system weighted by the associated reciprocal α_s for some real crime cases. In the top panels we have superimposed a loess curve, and in the bottom panels a linear fit to the log-transformed data.

data. The points in the top panels can be interpreted as estimates of the survival function $S(l)$ where l is the fragment length. The bottom panel indicates that the cumulative hazard is linear in the fragment lengths such that an exponential survival model is sufficient in order to describe the decay in amplification. However further experiments and analyses of the behaviour of degraded DNA needs to be performed for a proper inclusion of the problem in the model.

Drop-ins, stutters, drop-outs and pull-up effects

The contamination of the observations from stutters, drop-ins, drop-outs and pull-up effects increases the complexity of the analysis of DNA samples and in particular mixed DNA samples. As for degraded DNA separate models for these issues needs to be constructed for proper understanding of their behaviour. If we consider the four issues as separate problems we may assume they can be modelled independently given the quality of the DNA sample. However as these four events may occur at the same time, it can be difficult to assess which are present. For example may the presence of stutters, drop-ins and/or pull-up effect hide a drop-out. This may in particular be true for mixtures with a minor and major contributor where the ratio of their DNA concentrations is close to the stutter percentages of 5%-17% (Applied Biosystems, 2006, p. 9-22).

Furthermore the three *adding* events are difficult to distinguish from one another. This is due to their overlapping definitions where stutter and pull-up effects are the most restrictive and drop-ins are the artificial peaks not classifiable by the two. If neglecting back-stutters then stutters are only a possible explanation if the position of the artificial peak is before the true peaks. Pull-up effects can also be restricted to certain intervals on the allelic ladder. For a pull-up to occur there has to be an overlap on the allelic ladder between the dye bands. In Figure A.1 we see that for example D8 on the green band is in the span of both D3 and vWA on the blue band.

7.2.3 Finding possible matches

In Chapter 6 we were only able to assess whether a mixture with the proposed profiles fits our model assumptions for DNA STR mixtures. It gives however no indications in direction of which other two profiles that might be more likely under the observed alleles and peak areas. In the search over possible pairs of profiles we need only to search over one system at a time under assumption of Hardy-Weinberg equilibrium implying independence over systems with respect to the presence of alleles. However the results of this present thesis clearly indicates that when evaluating the evidence of each pair of profiles we need to include all systems. If we assume that no drop-outs have occurred, an algorithm for finding the best fitting pair could be as follows:

(1) Find all possible sets of pairs that match the observed alleles for each system s (see Table 7.1).

(2) Construct the associated T_s matrices.

(3) Construct all T matrices from these sub-matrices and determine H.

(4) Chose the configuration with the largest likelihood.

The likelihood comparisons mentioned in (4) needs to be evaluated with respect to some confidence limits.

In the worst case there will for each system be three observations implying there is 12 possible combinations (see Table 7.1). This induce that for each system we construct 12 sub-matrices T_s. The total number of possible T matrices is therefore in the worst case 12^S which even for moderate S is an intractable number of combinations.

The deviance involve determining the Mahalanobis distance and computing the determinant of $\mathrm{var}(M) = T_c D_c T_c + \mathrm{diag}(Q_c \boldsymbol{\tau}) + Q_c \Lambda Q_c$. Now in this expression only $T_c D_c T_c$ is affected by the current configuration of profiles, the latter terms do only depend on the profiles though $\boldsymbol{n}_c = (n_{1_c}, \ldots, n_{S_c})$ which is constant given the observations $\boldsymbol{M}$.

In order to make the search for possible profiles manageable we need to consider some more sophisticated optimization techniques. For our concrete setup with ten systems a worst case scenario is $12^{10} = 61917364224$ possible configurations. If assuming more than two persons contribute to a mixture the number of possible combinations increases dramatically hence for future use this problem needs careful consideration.

Table 7.1: Possible combinations of profiles for system s.

Obs. alleles	Possible combinations			
B_1	$(B_1, B_1)\times(B_1, B_1)$			1
B_1, B_2	$(B_1, B_1)\times(B_1, B_2)$	$(B_1, B_1)\times(B_2, B_2)$	$(B_1, B_2)\times(B_1, B_1)$	
	$(B_1, B_2)\times(B_1, B_2)$	$(B_1, B_2)\times(B_2, B_2)$	$(B_2, B_2)\times(B_1, B_1)$	
	$(B_2, B_2)\times(B_1, B_2)$			7
B_1, B_2, B_3	$(B_1, B_2)\times(B_3, B_3)$	$(B_1, B_2)\times(B_2, B_3)$	$(B_1, B_2)\times(B_1, B_3)$	
	$(B_1, B_3)\times(B_2, B_2)$	$(B_1, B_3)\times(B_1, B_2)$	$(B_1, B_3)\times(B_2, B_3)$	
	$(B_2, B_3)\times(B_1, B_1)$	$(B_2, B_3)\times(B_1, B_2)$	$(B_2, B_3)\times(B_1, B_3)$	
	$(B_1, B_1)\times(B_2, B_3)$	$(B_2, B_2)\times(B_1, B_3)$	$(B_3, B_3)\times(B_1, B_2)$	12
B_1, B_2, B_3, B_4	$(B_1, B_2)\times(B_3, B_4)$	$(B_1, B_3)\times(B_2, B_4)$	$(B_1, B_4)\times(B_2, B_3)$	
	$(B_2, B_3)\times(B_1, B_4)$	$(B_2, B_4)\times(B_1, B_3)$	$(B_3, B_4)\times(B_1, B_4)$	6

The approach mentioned above makes no use of the quantitative information available from the peak areas. Doing so may exclude some of possible combinations in Table 7.1 and hence reduce the number of total possible profiles. If the mixture ratio is not close to one then for systems with four observations most probable only one configuration of the alleles will be likely under the model assumptions. Hence this fixes the profiles for these systems and reduces the total number of possible profiles by a factor six.

Next a partial estimate of H can be made based on systems with four observations and then used when considering the remaining systems. This method tends to have a recursive structure and reduces the number of possible profiles. This approach needs further investigation but seems intuitively feasible from both a theoretical and practical point of view.

7.3 Conclusion

The experiences from the data analysis summarized in Chapter 2 were used though out this present thesis. The relations found in the data from extensive exploration are together with expert knowledge important tools for setting up a model that fits a complex structure as DNA mixtures. The main results from the data analysis discussed here are most likely independent of the number of contributors to a DNA mixture. Hence the basic assumptions of the model are still valid with more than two contributors. However the limitation of the model discussed in Section 7.2.2 may well be easier to incorporate before extending the model to multiple contributors.

Using the normal distribution for the model simplifies the estimation phase as many standard results were drawn upon and also closed form solutions of the estimators were guaranteed. The non-zero probability of having negative peak areas is of less importance and future work is therefore on the issues mentioned in Section 7.2 and not on implementing positive distributions such as the gamma distribution. By introducing a compound symmetry structure on the covariance of the error term we solved the problem of having varying number of observations for each case. The multiplication of $\text{diag}(\boldsymbol{h})^{-1/2}$ were incorporated to increase the fit of the model as we observed some alleles amplified differently across systems which were not captured by the original model.

Another advantage of the normal distribution is that the conditional distributions of the variables in the model are easily determined using standard formulae. This was extensively used when deriving the estimators for the EM-algorithm and also in the direct implementation of the EM-algorithm. When estimating the conditional means in the *inner* EM-procedure we first made use of the properties of the normal distribution also used in the Kalman filter, $\text{E}(\boldsymbol{Z}|\boldsymbol{X}) = \text{E}(\boldsymbol{Z}|\boldsymbol{Y}{=}\text{E}(\boldsymbol{Y}|\boldsymbol{X}))$. However it was not directly clear that a similar result was valid for the conditional mean of $\boldsymbol{Z}\boldsymbol{Z}^\top$.

The choice of R as the language for the implementation of the EM-algorithm was based on previous experiences and in comparison to e.g. C matrix multiplication is well implemented in R. The implementation is easily altered in order to cope with the covariance structures discussed in Section 7.2.1.

Assuming independence of the different systems in a DNA sample is a simplification which can not be supported by any work done in this present thesis. Modelling each system separately introduces considerable bias to the result since the approach makes use of the same information about the mixture for each system. Hence intersystem correlations need to be considered when assessing the weight of evidence in forensic DNA STR settings.

We found that Λ is significantly different from νI_S but also that the estimates of ν_{st} are subject to large variability. This indicates possibilities of model reductions and maybe incorporating a covariance structure based on dye bands and fragments lengths. This could for instance be done by including the difference in fragment length into the covariance function, $\text{cov}\left(B^{(k)}_{i,s}, B^{(k')}_{i',s'}\right) = \delta^t_{s,s'}$, where t is the difference in fragment length of the two alleles.

The insignificance of $\boldsymbol{\tau}$ in the model indicates that $\text{cov}(\tilde{\varepsilon}_{sj}, \tilde{\varepsilon}_{si})$ is constant for all $i, j = 1, \ldots, n_s$. This imply that the correlation matrix of $\tilde{\varepsilon}$ will have $\mathbf{1}_{n_s}\mathbf{1}^\top_{n_s}$-blocks down the diagonal. We may interpret this as the weighted errors within each systems is a linear function of each other.

The estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}^2$ supports the assumption of proportionality of the mean $\boldsymbol{\mu}$ and variance D of the unobservable peak areas $\boldsymbol{A}$. Hence this part of the model can be retained even if the covariance structure of $\tilde{\varepsilon}$ is changed to incorporate different symmetries.

The model's goodness of fit to the training data was assess by evaluating the Mahalanobis distance for each case given the estimated values of $\boldsymbol{\alpha}$, $\boldsymbol{\sigma}^2$ and Λ. The uniformity of the p-values from the χ^2-test showed a reasonable fit of the controlled experiments. A similar approach was applied to the real crime cases by splitting the evidence into two parts; mixture match within each system and the quality of the mixture. Only three cases had a poor fit of the mixture, whereas several mixtures showed unbalances across systems. The latter may be due to degraded DNA or contamination of the DNA. As a final assessment of the model we analyzed the single person-water “mixtures”. The p-values from the three tests mentioned above indicated a reasonable fit to the model. When only one person contributes the interpretations of (6.1) and (6.2) are merely balances within and between systems.

APPENDIX A

Biology of DNA

This chapter is taken from the authors own work written on the preceding semester.

In this chapter we will present the basic terms of the human DNA system and related biological topics with relevance for understanding of the overall subject of DNA typing. The chapter is meant as an introduction for statisticians and other interested with no further biological knowledge of chromosomes, alleles, DNA typing or forensic science as such. Therefore the depth and accuracy might not be of the same scientific standard compared to textbooks dedicated to these topics, but merely an overview and definition of the words used though out this present report. The chapter is based on Butler (2005) and Evett and Weir (1998).

A.1 Deoxyribonucleic acid

Deoxyribonucleic acid also known as DNA is the building blocks of all life on Earth. DNA is a double helix structure found in every nuclear cell in living organism. DNA is inherited from parent to offspring during reproduction. During reproduction the maternal chromosome pairs are separated into single chromosomes. A similar process applies to the paternal chromosomes and these single chromosomes then recombines by random with each other for the respective chromosomes - one from each parent. That is if the maternal chromosome pair is mM and the paternal is pP, the combinations mp, mP, Mp and MP all happens with equal probability. For humans each nuclear cell contains 23 chromosome pairs which constitute the human genome. Thus every human carry multiple copies of our DNA sequence. The characteristic double helix form of the DNA consists of two single stranded DNA sequences of four bases adenine (A), thymine (T), cytosine (C) and guanine (G). These bases form the structure of DNA by their repeat patterns and unique combinations. Due to the unique combinations of the four bases in AT/TA and CG/GC the double helix is kept together. An example of a DNA sequence is $TCTA$ which is a tetra nucleotide repeat pattern due to the four bases in the pattern. Throughout the rest of the present report the data used is of tetra nucleotide form.

In this paragraph we define some often used phrases in DNA and thus also in forensic DNA science. The genome is the entire DNA of the human body contained in the nuclear of the cells. That is the DNA in the mitochondria genome is considered separate. The chromosomes are the level just below the genome. The human DNA is made up of the 22 chromosome pairs and the sex chromosome pair. A gene is the part on the chromosomes where the DNA codes for some biological properties. The remaining part of the chromosome is called junk DNA. Loci which is plural for locus, is the term used for the position on the gene. That is in order to locate a position on the genome we must include information on chromosome numbers and loci, e.g. D16S539 is position 539 on chromosome 16. An allele corresponds to the length of the tetra nucleotide repeats on the particular loci. The length of the DNA fragments are measured on continuous scale but are discretised into a so called allele ladder, i.e. the length of DNA fragments are binned into intervals coding for the alleles. For a person with the same allele on a locus for both chromosomes in a chromosome pair we say they are homozygote otherwise heterozygote. There are a variable number of alleles for the different loci on the human genome. This variability is crucial for discriminating individuals. The kit used for the DNA typing in the data this project is Applied Biosystems Amplifiler STR SGM Plus®. The electropherogram (EPG) in Figure A.1 show the allelic ladder for this kit.

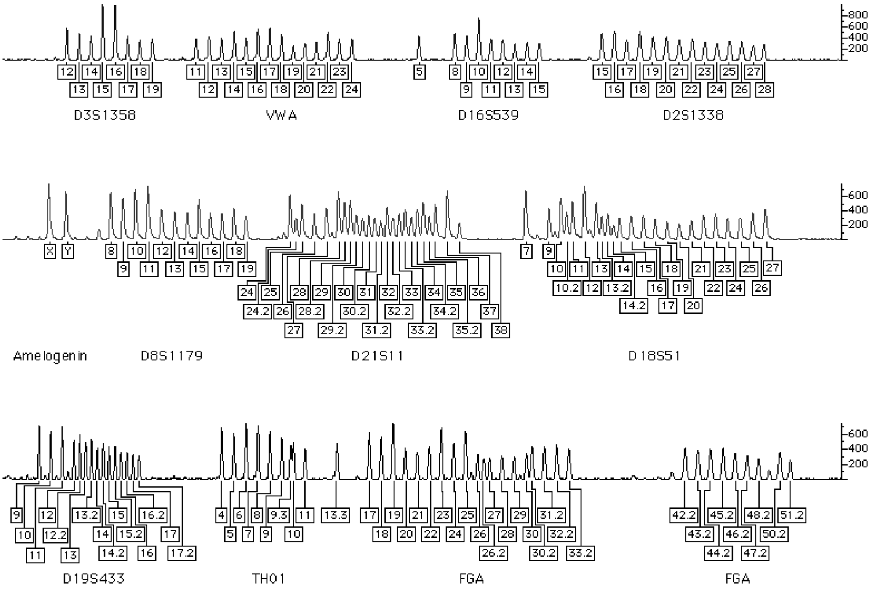

Figure A.1: Electropherogram (EPG) showing the allelic ladder for the Applied Biosystems Amplifiler STR SGM Plus®kit. The dyes used from top to bottom are blue, green and yellow. (Applied Biosystems, 2006, p. 9-7).

When DNA material is used in crime cases it is often found after some time in non-optimal environment. It may have been exposed to direct sun light, water, bacteria, heat, etc. The chemical reactions which degrade the DNA may break the DNA string into shorter pieces causing some methods of DNA typing to fail. The most commonly used methods for forensic DNA typing is called Short tandem repeat (STR) and is a methods which is fairly applicable for typing degraded DNA. Prior to the STR typing the DNA is amplified using Polymerase chain reaction (PCR). The method involves breaking the double stranded helix into single strands by the use of heat and enzymes after which specialized primers binds to single strands making new double helices.

By repeating this process millions of DNA copies can be made in a spell. A side-effect from PCR amplifying is the so called stutters. That is small peaks in the EPG located a few base pairs (bp) before the actual allele peak. A proposed explanation for the mechanism causing stutters is when the primers experience a mis-pairing during the amplification and hence producing allele markings for the amplified loci to be a few repeats shorter. In single contributor DNA samples it is often possible to detect stutters by eye. This is due to the significant difference in peak area of stutters relative to the real allele peak. Experimental data suggests that the peak areas of stutters in most cases are less than 15% of the area of the true allele peak areas. Other issues which blur the picture when analyzing DNA samples are drop-ins and drop-outs. As the phrases indicate it is when foreign allele peaks are observed or when some peaks are missing. Together with stutters and the fact that these three issues may arise on the same time gives reason for some concern. That is estimating the profiles of unknown contributors, exclusion of suspects and determining the number of contributors to a mixture gets complicated by these possible sources of error.

A.2 Population genetics

The uniqueness of the genotype for each individual is crucial in order to use DNA typing as discrimination tool for identification use. In court rooms where the forensic expert evaluate the strength of the evidence it is important to be able to determine how likely the present DNA profile could have originated from a random selected person in the reference population rather than the suspect. To assess estimates of the occurrence of genotypes in the reference population, e.g. individuals with the same ethnicity, nationality or cultural background as the suspect, we need to make some assumptions.

The simplest models in population genetics theory define the concept of an ideal population. The validity of the model is based on some assumptions where the two most fundamental are an infinite reference population from which the present population have descended, and the assumption of random mating. That is given an individual in the reference population then any other individual in the reference population is eligible for mating independent of sex, age, etc.

From the reference population we imagine a series of populations of size N are descending as shown in Figure A.2. Every one such population is origin for a chain of populations all of same size, which does not coexist with each other. The populations (both across chains and generations) differ from each other in allelic diversity due to random mating. This random mating hence implies the variability in present alleles in every population.

Assume now there exists for a gene $\boldsymbol{A}$ two alleles A_1 and A_2, then the allele proportion of A_i in the population is called p_i. From Section A.1 we can deduce that we are only able to observe pairs of alleles, e.g. observing A_1A_2 for $\boldsymbol{A}$ in this case. Now let P_{ij} be the proportion of A_iA_j and similar P_{ii} for homozygote A_iA_i. Then we can calculate p_i

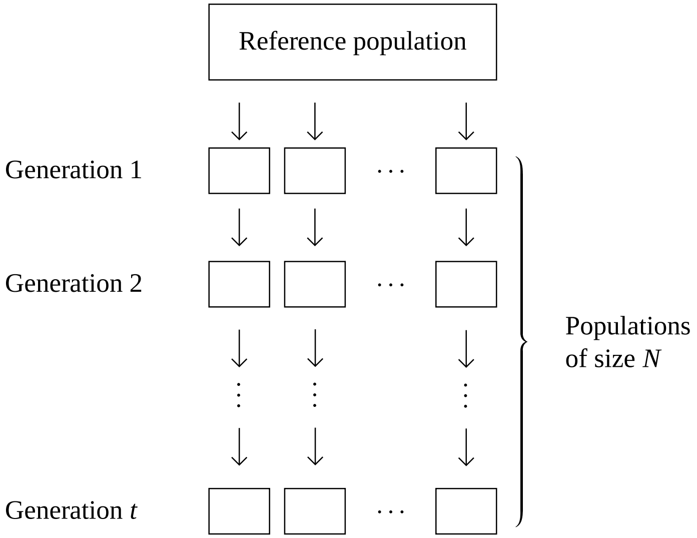

Figure A.2: Diagram of the infinite reference population and descending chains of sub populations of size N.

as

$$p_i = P_{ii} + \tfrac{1}{2} \sum_{j \neq i} P_{ij}, \tag{A.1}$$

where we by convention only consider P_{ij} for which $i < j$ and similar for A_iA_j. This relation between the genotype proportions and allele proportions does not rely on any of the assumptions mentioned above.

From the assumptions of random mating and infinite reference population we have that the genotype of one individual does not provide any information of the genotypes of others, hence a independence property among the genotypes of individuals. Returning to $\boldsymbol{A}$ with the two alleles A_1 and A_2, we now look at the proportion of P'_{ij}, which is the proportion of A_iA_j in generation $t+1$, where P_{ij} relates to generation t. From the independence between genotypes we have that we can multiply the probabilities of the genotypes P_{ij} together to have the probabilities of the offspring genotypes. Due to the low number of possible alleles of $\boldsymbol{A}$ we can summaries this in Table A.1.

From the information in Table A.1 it is rather simple to find that

$$P'_{11} = P_{11}^2 + \tfrac{1}{2}\left(P_{11}P_{12} + P_{12}P_{11}\right) + \tfrac{1}{4}P_{12}^2 = \left(P_{11} + \tfrac{1}{2}P_{12}\right)^2 = p_1^2,$$

and similar for $P'_{12} = 2p_1p_2$ and $P'_{22} = p_2^2$ where we used (A.1).

We can generalize these equations for a gene with an arbitrary number of alleles to yield $P'_{ii} = p_i^2$ and $P'_{ij} = 2p_ip_j$. In population genetics these equations is referred to as the Hardy-Weinberg law.

Table A.1: Probabilities of genotypes among offspring.

			Offspring		
Mother	Father	Probability	A_1A_1	A_1A_2	A_2A_2
A_1A_1	A_1A_1	$P_{11}P_{11}$	1	0	0
	A_1A_2	$P_{11}P_{12}$	1/2	1/2	0
	A_2A_2	$P_{11}P_{22}$	0	1	0
A_1A_2	A_1A_1	$P_{12}P_{11}$	1/2	1/2	0
	A_1A_2	$P_{12}P_{12}$	1/4	1/2	1/4
	A_2A_2	$P_{12}P_{22}$	0	1/2	1/2
A_2A_2	A_1A_1	$P_{22}P_{11}$	0	1	0
	A_1A_2	$P_{22}P_{12}$	0	1/2	1/2
	A_2A_2	$P_{22}P_{22}$	0	0	1

In the discussion above we assumed infinite population size and random mating, but also that selection, mutation and migration were not present. These three concepts covers the issue where some combinations of alleles are preferable compared to other combinations, where alleles is present in generation $t + 1$ but not in generation t and were the populations from different chains in Figure A.2 interact, respectively. To model these issues one must include some further notation, but for discussion of these issues we refer to Evett and Weir (1998). A final term involved in population genetics is an equilibrium situation which was demonstrated above when selection, mutation and migration were not allowed. Equilibrium can also occur when the three disturbing forces are included in the model.

When one consider real world data it is necessary to evaluate if the data satisfy Hardy-Weinberg equilibrium in order to verify that the probabilities of the genotypes can be estimated by multiplying the allelic proportions.

Bibliography

Applied Biosystems (2006). *AmpFℓSTR®SGM Plus®PCR Amplification Kit User's Manual.* Applied Biosystems. Retrieved June 4, 2007, from http://docs.appliedbiosystems.com/pebiodocs/04309589.pdf.

Balding, J. M. (2005). *Forensic DNA Typing* (2nd ed.). Elsevier Science & Technology.

Bøttcher, S. G., Christensen, E. S., Lauritzen, S. L., Mogensen, H. S., and Morling, N. (2007). Investigation of a Gamma model for mixture STR samples. Unpublished.

Butler, J. M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2nd ed.). Academic Press Inc.,U.S.

Clayton, T. M., Whitaker, J. P., Sparkes, R., and Gill, P. D. (1998). Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Science International*,* 91(1): 55–70.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2004). Identification and seperation of DNA mixtures using peak area information. Cass Statistics Research Paper No. 25, London: Cass Business School, City of London.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2007). Identification and seperation of DNA mixtures using peak area information. Forensic Science International*,* 166(1): 28–34.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall Ltd.

Evett, I. W., Gill, P. D., and Lambert, J. A. (1998). Taking account of peak areas when interpreting mixed DNA profiles. Journal of Forensic Sciences*,* 43(1): 62–69.

Evett, I. W. and Weir, B. S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists.* Sinauer Associates Inc.

Gilder, J., Doom, T., Inman, K., and Krane, D. (2007). Run-Specific Limits of Detection and Quantitation for STR-based DNA Testing. Journal of Forensic Science,*,* 52(1): 97–101.

Gill, P., Brenner, C. H., Buckleton, J. S., Carracedo, A., Krawczak, M., Mayr, W. R., Morling, N., Prinz, M., Schneider, P. M., and Weir, B. S. (2006). DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Science International*,* 160(2-3): 90–101.

Gill, P. D., Sparkes, R., Pinchin, R., Clayton, T. M., Whitaker, J. P., and Buckleton, J. (1998). Interpreting simple STR mixtures using allele peak areas. Forensic Science International*,* 91(1): 41–53.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press.

Lauritzen, S. L. (2006, Hillary Term). Further Statistical Methods. Department of Statistics, University of Oxford. Lecture notes.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons, Ltd.

McLachlan, G. and Krishnan, T. (1996). *The EM-algorithm and Extensions.* John Wiley & Sons, Ltd.

Mortera, J., Dawid, A. P., and Lauritzen, S. L. (2003). Probabilistic expert systems for DNA mixture profiling. Theoretical Population Biology, 63: 191–205.

Nielsen, K., Mogensen, H. S., Eriksen, B., Hedman, J., Niedersttter, H., Parson, W., and Morling, N. (2007). Comparison of six DNA quantification methods. Work in progress.

Rudin, N. and Inman, K. (2001). *An Introduction to Forensic DNA Analysis* (2nd ed.). CRC Press.

Taroni, F., Aitken, C., Garbolino, P., and Biedermann, A. (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science.* John Wiley & Sons, Ltd.

Tvedebrink, T. (2006). Amplification of DNA mixtures. MAT5 project report, Department of Mathematical Sciences, Aalborg University.

Votaw, D. F. (1948). Testing Compound Symmetry in a Normal Multivariate Distribution. The Annals of Mathematical Statistics, 19(4): 447–473.