The Application of Bayesian Hierarchical Models to Heterogeneous DNA Profiling Data

Thesis submitted to the University of London for the degree of Doctor of Philosophy in the Faculty of Science

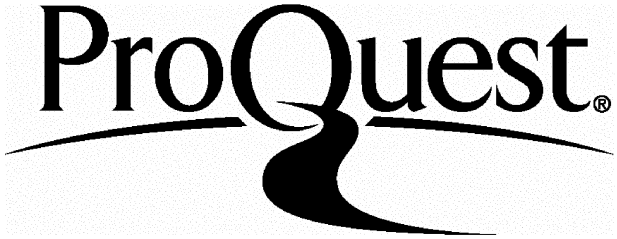
by

John Pueschel

Department of Statistical Science

University College London

January 2002



Abstract

A situation is considered in which a suspect has been found whose DNA profile matches that of a sample, assumed to originate from the offender, found at the scene of a crime being investigated.

The way in which this evidence should be used is reviewed, highlighting the role of the match probability, the probability of a particular individual having the profile in question given the suspect's possession of the profile, and a database of individual profiles. The value of this probability is affected by the heterogeneity of the population, and failure to take account of this could result in a false conviction.

A Bayesian hierarchical model designed to represent population substructure is presented. Parameters are clearly defined at each level within a model displaying justifiable conditional independence properties. This model is then used as a basis for inference about the required match probabilities, highlighting errors in previous approaches.

As the match probability calculations described are impossible analytically, we use MCMC methods to analyse a UK database of DNA profiles. A comparison of results with those of previous methods highlights the practical importance of a clearly defined hierarchy and conducting the correct conditioning upon the database.

Acknowledgements

I would especially like to express my gratitude to Professor Phil Dawid for his excellent guidance, encouragement and patience.

The financial support provided by EPSRC is also much appreciated.

I would also like to thank many other members of the Department for their help and encouragement, in particular fellow members of the research student group, and the undergraduates who 'dared' to enter our room. I would like to mention you all by name, but fear that I would miss someone. You know who you are, and the friendships developed are truly valued.

Various other people require thanking for keeping me (relatively) sane, including the Marshalls crew, the football lads, Ryz, Andy and Dave.

Most of all I would like to thank my family, and particularly my parents, for their support and encouragement. Thanks!

Contents

List of Tables

List of Figures

1 Introduction

1.1 Background

Over the past ten years, 'DNA fingerprinting' has become an important tool in the legal world. High profile cases such as the O.J. Simpson trial [Weir, 1995] have increased public awareness of the technique.

At any criminal trial, the ultimate aim is to reach a decision regarding the guilt of the defendant. It has been argued [Fienberg and Finkelstein, 1996] that Bayesian methods provide a sensible framework for reaching this decision. The application of such a method would result in a posterior probability of guilt of the suspect given the evidence. The suspect would then be convicted if this probability is above a certain value.

A DNA profile consists of pairs of observations, one at each of a small number of well defined sites, or *loci*. One member of each pair is inherited from the individual's father, and one from the mother. The possible observations at each locus are referred to as *alleles*, each effectively representing a number of repeats of a sequence of base pairs (see Appendix A for further detail).

For our purposes, a DNA profile takes the form of a series of pairs of integers, one pair for each locus considered, each number representing a different allele. The empirical frequencies of various alleles at four loci within a Caucasian database can be seen in Figure 1. Within this population, the profile $\{(16,17)(7,10)(5,7)(11,11)\}$ would be considered a relatively common profile, while $\{(13,20)(5,8)(8,8)(8,9)\}$ is a rare profile.

In this thesis we consider a situation in which a crime has been committed and a tissue sample (possibly blood, hair, semen, etc.) has been recovered from the scene. From this sample we have a DNA profile $\boldsymbol{X}_C$ assumed to be that of the culprit C.

In addition we have a suspect s whose DNA profile $\boldsymbol{X}_s$ is found to match $\boldsymbol{X}_C$. We denote the known matching profile by $\boldsymbol{y}$.

Further profiles are available in a database of individuals δ. This could

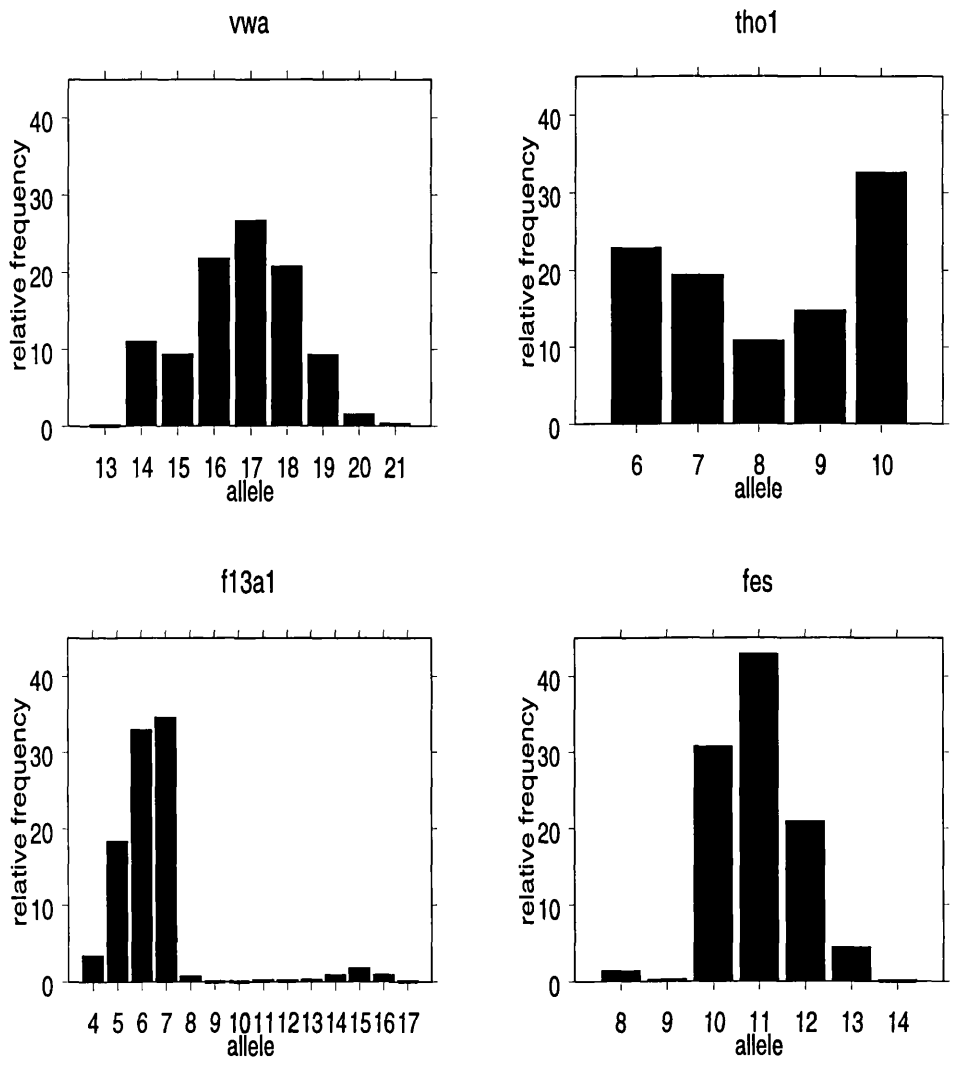


Figure 1: Relative frequency graphs of alleles in a Caucasian database.

comprise a number of suspects, some of their close relatives, a database of known criminals and a "statistical" database drawn from the general population. We restrict ourselves to the instance in which δ represents a statistical database. The collection of DNA profiles from these individuals is labelled χ_δ. Other (non-DNA) evidence, such as eye witness accounts or alibis, which may appear to be for or against the suspect, is labelled ε.

Our ultimate aim is to calculate a probability that the suspect is guilty given

all the evidence and data available

$$p_{guilt} := \Pr(C = s | \boldsymbol{X}_C = \boldsymbol{y}, \boldsymbol{X}_s = \boldsymbol{y}, \chi_\delta = \xi_\delta, \varepsilon).$$

Application of Bayes' Theorem allows this probability to be expressed [Dawid and Mortera, 1996, Weir, 1994] as

$$p_{guilt} = \frac{\pi_s m_s}{\sum_{i \in \mathcal{P}} \pi_i m_i}, \tag{1}$$

where

$$\begin{aligned} \pi_i &= \Pr(C = i | \boldsymbol{X}_s = \boldsymbol{y}, \chi_\delta = \xi_\delta, \varepsilon), \\ m_i &= \Pr(\boldsymbol{X}_C = \boldsymbol{y} | C = i, \boldsymbol{X}_s = \boldsymbol{y}, \chi_\delta = \xi_\delta, \varepsilon), \end{aligned}$$

i labels individuals, and $\mathcal{P}$ represents the population of possible culprits.

Without knowledge of the culprit profile $\boldsymbol{X}_C$, it is reasonable to assume that C is independent of $(\boldsymbol{X}_s, \xi_\delta)$. In this thesis we assume that the sample $\boldsymbol{X}_C$ is definitely that of the culprit, and that there are no errors in the process establishing a profile from a sample. In reality this is not necessarily so, but these assumptions allow us to take m_s to be 1, and simplify equation (1) to give

$$p_{guilt} = \frac{\pi_s}{\sum_{i \in \mathcal{P}} \pi_i m_i}, \tag{2}$$

where

$$\begin{aligned} \pi_i &= \Pr(C = i | \varepsilon), \\ m_i &= \Pr(\boldsymbol{X}_C = \boldsymbol{y} | C = i, \boldsymbol{X}_s = \boldsymbol{y}, \chi_\delta = \xi_\delta, \varepsilon). \end{aligned}$$

A priori the members of the database are potential perpetrators. It is assumed that the DNA profile measurements contain no error, and it is therefore possible to eliminate any database individual whose profile does not match $\boldsymbol{y}$. Thus, defining $\alpha = \{s\} \cup \delta$ as the collection of individuals upon whom we have measurements,

$$p_{guilt} = \frac{\pi_s}{\sum_{i \in \beta} \pi_i + \sum_{i \notin \alpha} \pi_i m_i} \tag{3}$$

where β is the set of individuals in the complete database α whose profiles match the profile $\boldsymbol{y}$ of the culprit.

To evaluate (3), prior probabilities of guilt must be specified for each individual. In particular, members of the jury use the non-DNA evidence to arrive at a subjective prior probability π_s of guilt of the suspect. Match probabilities can then be used to update these prior probabilities, leading to a posterior probability of guilt for the suspect.

A major problem faced is that of evaluating a match probability m_i for each of a potentially large number of individuals in the suspect population $\mathcal{P}$. The calculation of a match probability for each individual in the population is impracticable.

Some methods previously employed have calculated a single match probability for all the individuals. These methods, described in Chapter 2, do not take reasonable account of genetic relationships between individuals. In Chapter 2 we consider why it it is important to take account of these relationships, and consider how this might be achieved. Chapter 3 introduces a hierarchical model designed to incorporate population substructure. In Chapter 4 we consider how this hierarchical model can be used to define two statistical models, and how these can be used to calculate the required match probabilities. The hierarchical model provides a framework which is used in Chapter 5 to clearly define the roles of various genetic parameters.

As the calculations required to evaluate match probabilities are impossible analytically, Markov Chain Monte Carlo methods are introduced in Chapter 6 as a precursor to their application in Chapter 7.

Other authors, in particular Foreman, Evett and Smith [Foreman, Evett and Smith, 1997] and Roeder, Escobar, Kadane and Balazs [Roeder, Escobar, Kadane and Balazs, 1998], have published papers working under a similar principle of calculating match probabilities taking account of population substructure. These papers are reviewed in Chapter 9 and areas in which it is felt omissions have been made outlined. To demonstrate why it is felt

that the approach of this thesis represents a step forward, match probabilities and posterior probabilities of guilt are calculated under the various methods described, and compared.

Chapter 10 considers future work, in particular the generalization of our method to an unknown number of subpopulations within the population.

2 Methods of match probability calculation

To calculate a posterior probability of guilt for the suspect we require a match probability m_i for each individual i in the population $\mathcal{P}$:

$$m_i = \Pr(\boldsymbol{X}_i = \boldsymbol{y} | \boldsymbol{X}_s = \boldsymbol{y}, \chi_\delta = \xi_\delta, \varepsilon).$$

Frequently close relatives of the suspect can be eliminated from the enquiry. If they cannot be eliminated, it is important that they be included in the calculation. If it is impossible to obtain a DNA profile from a close relative, then the match probability of that individual should be calculated separately. Two individuals having recent ancestors in common have a greatly increased probability of inheriting common genes. This means that, for a close relative of the suspect, the probability that they have the profile $\boldsymbol{y}$ given that $\boldsymbol{X}_s = \boldsymbol{y}$ will generally be much higher than for unrelated members of the population, reducing the weight of evidence against the suspect. Derivation of close family match probabilities is described by Donnelly [Donnelly, 1995].

For the remainder of the thesis we concentrate upon methods for match probability calculation for members of the general population.

In this chapter we consider methods of match probability calculation previously presented in court. The errors in these methods are highlighted, leading to an outline of the philosophy employed in this thesis.

2.1 Product rule

The initial method used to calculate the match probability of the offender employed the product rule [NRC, 1996].

We define $g_j(k)$ as the relative frequency of allele k at locus $(j = 1, \ldots, M)$ within the population under consideration. Under the product rule, the match probability is then defined as

$$m_i = \prod_{j=1}^{M} 2^{h(y_{j1}, y_{j2})} g_j(y_{j1}) g_j(y_{j2}), \tag{4}$$

where y_{jb} refers to allele b of the pair observed at locus j of the suspect profile, and

$$h(r,s) = \begin{cases} 0 & \text{if } r = s; \\ 1 & \text{if } r \neq s. \end{cases}$$

for all individuals i.

This is estimated by substituting empirical relative frequency estimates from a suitable database into the product in equation (4).

Derivation of equation (4) requires the assumption of independence of profile alleles across loci. To accept this method, one must also consider mating to be random across the whole population. This is *not* a valid assumption, and its use can result in prejudice against the suspect.

2.2 Taking account of population substructure

Within a large population, mating cannot be considered random. Individuals within sections of the population are more likely to mate with another individual in that section than with someone outside it.

We approximate this situation with a model which divides the population into discrete subpopulations within which individuals are considered randomly mating. These subpopulations are considered to have evolved from a large ancestral population, as outlined in Figure 2.

One interpretation of these subpopulations within, say, an American Caucasian population would be as groups of people whose ancestors came from various countries of Europe. More usually, however, these subpopulations are mere artifacts of the model, designed to reflect the fact that individuals within a large population do not mate at random. It is thus difficult to clearly define these subpopulations, and to allocate individuals to them in the way that we could allocate, say, by nationality.

The ceiling principle and interim ceiling principle were recommended by the first NRC report [NRC, 1992] as methods of match probability calculation accounting for population substructure.

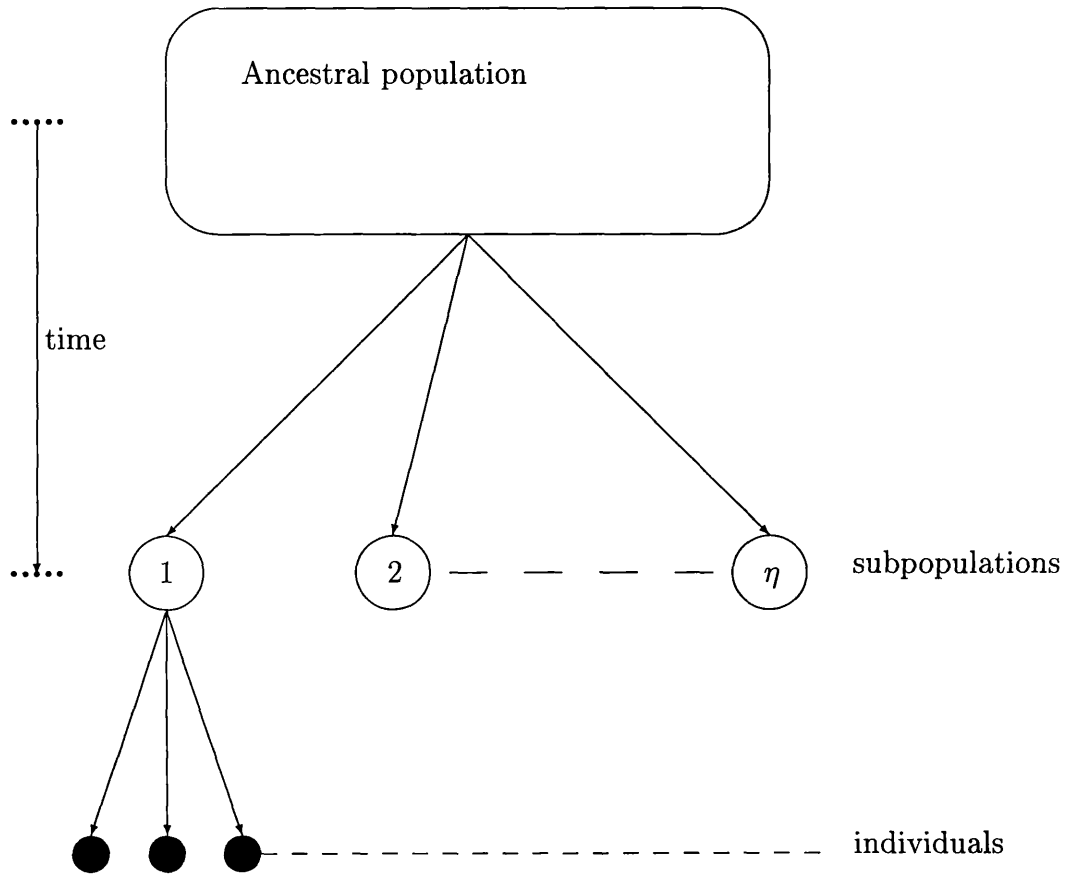


Figure 2: Population substructure model.

The ceiling principle was designed to give a deliberately conservative estimate of the match probability, i.e. greater than the true value. Its execution involves the use of samples from a number of subpopulations and for each allele taking the highest frequency among the groups sampled, or 5%, whichever is larger. The bound on the profile frequency is then obtained by multiplying together these individual allele limits. As the subpopulations are not usually clearly defined in terms of observable characteristics, it is difficult to obtain the samples required.

As a result of this difficulty, the interim ceiling principle was developed. This has been widely used. The rule states that "In applying the multiplication [of frequencies across loci] rule, the 95% upper confidence limit of the frequency of each allele should be calculated for separate 'racial' groups and the highest of these values or 10% (whichever is larger) should be used. Data on at least three major 'races' should be analyzed." The ceiling principle has been heavily criticized [NRC, 1996]. A clear flaw in the interim ceiling principle is that it will give the same limit regardless of the racial group of the individual concerned.

Weir [Weir, 1994] introduced a method far more satisfactory than the ceiling principle. Relating posterior odds to prior odds in the manner outlined in Chapter 1, he too stressed the need to take into account the dependence between the profiles of culprit and suspect when calculating match probabilities.

Wright [Wright, 1951] introduced parameters F_{ST}, F_{IT} and F_{IS} to summarise substructure within a population. They are defined as the correlations of genes of different individuals in the same subpopulation, of genes within individuals and of genes within individuals within populations respectively. This definition of "correlations" is unclear as when defining correlations it is necessary to be clear upon what we are conditioning. Population structure can be considered to be of a hierarchical nature (Figure 2). Assuming such a structure, statements of relative correlation should be made with reference to the characteristics observed at a specified level of the hierarchy. Later studies have introduced quantities supposedly equivalently to those of Wright. However, consideration of these parameters with the conditioning suggested by their definitions shows that they

are not equivalent. This further clouds the study of population substructure and is discussed in more detail in Chapter 5.

However the principle of introducing a parameter to govern subpopulation differentiation within the above model is a very important one. Weir [Weir, 1994] considered the effects of inbreeding on forensic calculations using this parameter based method.

Two genes at a locus are defined as being identical by descent (ibd) if they have the same ancestral gene. Weir defines a number of measures of descent affecting the probabilities of observing a particular set of four genes within two pairs. These include

- θ the probability that any two genes at a locus are ibd;
- γ the probability that any three genes are ibd;
- δ the probability that any four genes are ibd;
- Δ the probability that any two pairs of genes are ibd.

Single locus match probabilities

$$m_i = \Pr(\boldsymbol{X}_i = (y_1, y_2) | \boldsymbol{X}_s = (y_1, y_2), \theta, \gamma, \delta, \Delta)$$

when the individuals i and s are assumed to belong to the same subpopulation are derived to be

$$m_i = \begin{cases} [(1-6\theta+8\gamma+3\Delta-6\delta)p_i^3 & \\ +6(\theta-2\gamma-\Delta+2\delta)p_i^2 & \text{if } y_1 = y_2 \\ +(4\gamma+3\Delta-7\delta)p_i+\delta]\,/\,[p_i+(1-p_i)\theta]\,; & \\ & \\ [(1-6\theta+8\gamma+3\Delta-6\delta)p_ip_j & \\ +2(\theta-2\gamma-\Delta+2\delta)(p_i+p_j) & \text{if } y_{j1} \neq y_{j2} \\ +2(\Delta-\delta)]\,/(1-\theta). & \end{cases} \tag{5}$$

If the population under consideration is in evolutionary equilibrium, the quantities θ, γ, δ and Δ are not changing over the time. Li [Li, 1996] showed that in this instance γ, δ, and Δ can be expressed in terms of θ:

$$\gamma = \frac{2\theta^2}{1+\theta}$$

$$\delta = \frac{6\theta^3}{(1+\theta)(1+2\theta)}$$
$$\Delta = \frac{\theta^2(1+5\theta)}{(1+\theta)(1+2\theta)}$$

Substituting these expressions into the conditional probabilities (5) gives expressions in terms of θ and population-wide allele frequencies γ. These are similar in form to those determined by Balding and Nichols [Balding and Nichols, 1995]. If, at a particular locus, the matching gene pair is homozygous (i.e. the two alleles displayed are similar),

$$m_i = \frac{(2\theta + (1-\theta)\gamma(y_1))(3\theta + (1-\theta)\gamma(y_1))}{(1+\theta)(1+2\theta)}, \tag{6}$$

and if it is heterozygous,

$$m_i = 2\frac{(\theta + (1-\theta)\gamma(y_1))(\theta + (1-\theta)\gamma(y_2))}{(1+\theta)(1+2\theta)} \tag{7}$$

Weir and Evett [Weir and Evett, 1998] discuss methods of estimating the coancestry coefficient θ given data from a number of subpopulations. However, because of the difficulty in allocating individuals to subpopulations, they suggest two possible alternatives to the problem of estimating match probabilities taking into account population substructure.

The first is to refer to previous studies of human population structure such as that by Cavalli-Sforza *et al* [Cavalli-Sforza *et al*, 1994]. Assuming that populations similar to that under consideration have been studied, appropriate estimates can be substituted into equations 6 and 7.

The other alternative covered was proposed in the second NRC report [NRC, 1996] which was produced following heavy criticism of the ceiling principle and other aspects of the initial report. The NRC recommended the substitution of $\theta = 0.03$ into equations (6) and (7) when suspect and culprit are considered to come from the same subpopulation. This is larger than most empirical estimates meaning that the resultant match probability estimates should be conservative.

Foreman *et al* [Foreman, Evett and Smith, 1997] introduced a Bayesian model to represent population substructure. This is used to make inference

about the posterior distribution of θ conditional upon a database of individual profiles. The problem of a lack of subpopulation data is overcome by allowing the individuals to be partitiioned into a specified number of groups, their likelihood assigning most weight to those partitions grouping individuals with the most similar profiles. The results are then combined with the theory of Balding and Nichols to produce match probability estimates in the presence of population substructure. The paper of Foreman *et al* is discussed in more detail in Chapter 9. It proposes the calculation of two match probabilities, one conditional on the suspect being in the same subpopulation as the culprit, and one conditional on the two being in different subpopulations. In this thesis we calculate η match probabilities, each conditional on the suspect being in a particular subpopulation $(\mathcal{P}_l; l = 1, \ldots, \eta)$.

As random mating is assumed within each subpopulation, all individuals within a particular subpopulation are considered to be exchangeable [Dawid and Mortera, 1996]. This means that

$$\Pr(\boldsymbol{X}_i = \boldsymbol{y} | \chi_\alpha = \xi_\alpha) \tag{8}$$

is constant for all individuals i (outside the database α) in subpopulation $\mathcal{P}_l$. This probability will be referred to as the subpopulation match probability m_l. Assuming that there are η subpopulations, the denominator of the posterior probability of guilt of the suspect (equation (3)) becomes

$$\sum_{i \in \beta} \pi_i + \Pr(C \notin \alpha) \sum_{l=1}^{\eta} \lambda_l m_l, \tag{9}$$

where $\lambda_l = \Pr(C \in \mathcal{P}_l | C \notin \alpha)$.

Calculation of the subpopulation match probabilities $(m_l; l = 1, \ldots, \eta)$ is based upon the Bayesian hierarchical model outlined in Chapter 3 and is described in detail in Chapter 4.

US courtrooms currently follow the methods of the second NRC report when dealing with the problem of population substructure in the context of the presentation of DNA profiling evidence. Databases are now available for a number of racial groups, and there is an argument [People vs. Soto, 1999] that mating

within racial groups is of a sufficiently random nature as to justify the use of the product rule in the calcualation of match probabities. This is not so, and this thesis is particularly concerned with presenting the correct method of match probability calculation. In doing this we are careful to apply the correct conditioning throughout to ensure that the information supplied by the database is properly utilised.

3 The Hierarchical Model

3.1 Introduction

The population model described in Chapter 2 displays a clear hierarchical stucture with the following three levels:

(i) ancestral population;

(ii) subpopulations, descended from the ancestral population;

(iii) individuals, within the subpopulations.

3.1.1 Exchangeability

Assumptions of exchangeability feature heavily in this model. Variables $x_1, \ldots, x_n$ are considered exchangeable if their joint probability distribution $f(x_1, \ldots, x_n)$ is invariant to permutations of the indices, i.e. there is no information conveyed by the unit indices themselves.

We assume that, before observing any data, we have no way of distinguishing subpopulations, or individuals within subpopulations. We therefore apply the assumption of exchangeability to the units at stages (ii) and (iii) of the above model.

It is generally appropriate to model exchangeable variables as independently and identically distributed (iid) given some unknown parameter θ, say, with some distribution $p(\theta)$. This arises from *de Finetti's theorem* which states that in the limit as $n \to \infty$, any suitably well-behaved exchangeable distribution on $(x_1, \ldots, x_n)$ can be written in the iid mixture form

$$p(\boldsymbol{x}) = \int \left[\prod_{i=1}^{n} p(x_i|\theta) \right] p(\theta) d\theta.$$

This property is particularly useful in the field of hierarchical modelling in which it is often desirable to model the variables at a particular level to be independent and identically distributed given those at the level above.

The structure and exchangeability properties assumed mean that the population model described is readily translated to a Bayesian hierarchical model involving parameters that can be interpreted in terms of genetic characteristics.

3.2 Outline of the model

The profile $\boldsymbol{X}_i$ of an individual i is a collection of genotypes, one at each of M loci. The genotype at locus j consists of a pair of values from a finite set $\mathcal{R}_j$ of alleles. The set $\mathcal{R}_j$ consists of r_j consecutive integers, and at this point we map this series onto the sequence $(1, \ldots, r_j)$. One member of the pair (the paternal *band*) is inherited from the father, the other (maternal band) from the mother. Generally we would not know which of the pair was the paternal band. However, as it simplifies the notation while making no difference with regard to inference from the database profiles, we consider the paternal and maternal bands to be distinguishable. These bands are labelled by $b = 1$ (paternal), 2 (maternal). It is important to note that, when considering a match between the crime sample and the suspect s or a 'random' individual i, we must take account of the fact that the bands are not truly distinguishable when calculating match probabilities.

The full DNA profile for individual i is

$$\boldsymbol{X}_i \equiv \{X_{ijb}; j = 1, \ldots, M; b = 1, 2\}.$$

However, during this chapter, when describing the model we assume that the subpopulation of each individual is known. This allows us to define an alternative notation, replacing X_{ijb} by X_{lhjb} representing band b at locus j of individual h within subpopulation l.

The model outlined can be represented by the directed acyclic graph of Figure 3. Circular nodes represent parameters which are unknown, while rectangular nodes represent known quantities. An arrow between two variables represents a direct influence of one upon another. Once all direct influences upon a node are known, all other potential influences are considered irrelevant. A double line represents a deterministic link.

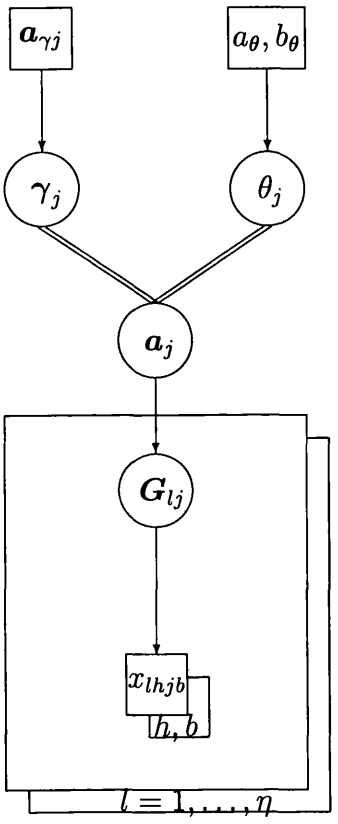


Figure 3: DAG for the case of complete information (at a fixed locus j), introducing the parameters of the hierarchical model.

We consider a heterogeneous population, consisting of (large) subpopulations labelled by $l = 1, \ldots, \eta$, and propose the following 3-stage hierarchical model:

Stage 1 It is assumed that, conditional on some collection $\{\boldsymbol{G} = (G_{lj}(1), \ldots, G_{lj}(r_j)); l = 1, \ldots, \eta; j = 1, \ldots, M\}$ of "within-population" parameter vectors, there is independence of the values (X_{ijb}) across bands within a pair, across loci, and across individuals. We then have, for each individual $i \in \mathcal{P}_l$

$$\Pr(X_{lhjb} = k|\boldsymbol{G}) = G_{lj}(k), \text{ independently for all } (h, j, b, l). \tag{10}$$

With reference to the "genetic model" described earlier, $G_{lj}(k)$ could be interpreted as a limiting relative frequency of allele k at locus j in subpopulation $\mathcal{P}_l$, if the size of that subpopulation tends to infinity.

Stage 2 As we consider the subpopulations exchangeable, it is appropriate, at a fixed locus j, to treat $(\boldsymbol{G}_{lj}; l = 1, \ldots, \eta)$ as independently and identically distributed given some distribution Π.

The distributon Π can be chosen based upon a genetic model, such as that described by Balding and Nichols [Balding and Nichols, 1994, Balding and Nichols, 1995]. At this point we introduce parameters $(\theta_j, \boldsymbol{\gamma}_j)$, both with a genetic interpretation at the ancestral population level. We consider $\boldsymbol{\gamma}_j$ to be a probability vector representing the mean, at locus j, of the process generating subpopulation allele probabilities. To govern the variance of this process, we define θ_j. Such subpopulation differentiation parameters are mentioned frequently in population genetics literature and are discussed further in Chapter 5. Assuming that the model of Balding and Nichols is reasonable, there is some justification for the use, for Π, of the Dirichlet distribution:

$$\boldsymbol{G}_{lj} \sim \text{Dirichlet}(a_j(1), a_j(2), \ldots, a_j(r_j)), \text{ independently for all } l, j, \tag{11}$$

conditionally on hyperparameters $(a_j(1), a_j(2), \ldots, a_j(r_j))$, where $\boldsymbol{a}_j = \frac{1-\theta_j}{\theta_j}\boldsymbol{\gamma}_j$.

Under this distribution,

$$\mathrm{E}[G_{lj}(k)|\boldsymbol{\gamma}, \boldsymbol{\theta}] = \frac{a_j(k)}{a_j(+)} = \boldsymbol{\gamma}_j(k), \tag{12}$$

where $a_j(+) = \sum_{k=1}^{r_j} a_j(k)$. The variance of the subpopulation probabilities generated is then given by

$$\begin{aligned} \mathrm{Var}(G_{lj}(k)|\boldsymbol{\gamma}, \boldsymbol{\theta}) &= \frac{a_j(k)(a_j(+) - a_j(k))}{a_j(+)^2(a_j(+)+1)} \\ &= \theta_j \gamma_j(k)(1 - \gamma_j(k)). \end{aligned} \tag{13}$$

Stage 3 Finally, we need to give a distribution to Π (or, equivalently, to $(\boldsymbol{\gamma}, \boldsymbol{\theta})$). Assuming $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ to be independent, a reasonable prior for $\boldsymbol{a}$ would be based upon

$$\begin{aligned} \boldsymbol{\gamma}_j &\sim \mathrm{Dirichlet}(a_{\gamma j}(1), \ldots, a_{\gamma j}(r_j)); \\ \theta_j &\sim \mathrm{Beta}(a_\theta, b_\theta); \end{aligned}$$

where $(\boldsymbol{a}_\gamma, a_\theta, b_\theta)$ are hyperparameters to be chosen.

This model can be used as a reference point for a discussion of the published methods of tackling the problem of DNA identification using heterogenous databases. Of particular interest in a number of studies is the parameter describing subpopulation differentiation. In this instance, $\boldsymbol{\theta}$ is a parameter controlling the variance of allele frequencies generated. An alternative parameter could be defined to summarise the variation actually observed in the subpopulations 1 to η.

In many cases such parameters are confused and interchanged. Chapter 5 compares these parameters and clearly distinguishes them using the hierarchical framework outlined in this chapter.

3.3 Introducing extra levels to hierarchical models

It has been suggested that subpopulation specific differentiation parameters $(\theta_l; l = 1, \ldots, \eta)$ should be used. This adds an extra level of complexity to the model, but could be justified if it gives greater flexibility.

To investigate whether or not this is indeed the case, we must compare the models with and without subpopulation specific differentiation parameters. To

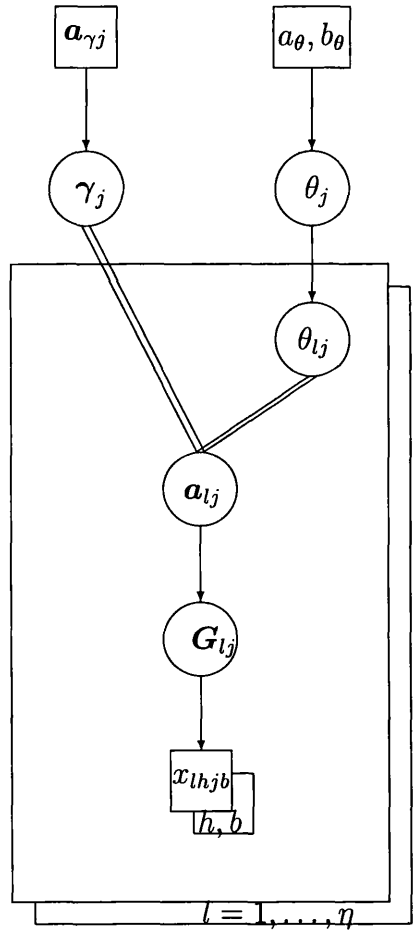


Figure 4: DAG for the case of complete information (at a fixed locus j), introducing subpopulation specific differentiation parameters.

outline the method we initially consider a similar model with Normal distributions, and then go on to consider the model described in this chapter.

It should be noted that adding extra levels of complexity to a hierarchical model does not necessarily result in a more flexible final model. An example is provided by the following model:

Stage 1

$$X_l|(m_i; i = 1, \ldots, \eta), \sigma^2 \sim \text{Normal}(m_l, \sigma^2),$$

independently across l.

Stage 2

$$m_l|\mu, \tau^2 \sim \text{Normal}(\mu, \tau^2),$$

independently across l.

Assuming σ, μ and τ to be known, this 2 stage model can be expressed as follows:

$$X_l = m_l + \sigma Z_l$$

where (Z_l) is defined to be a collection of independent standard Normal random variables, also independent of (m_l);

$$m_l = \mu + \tau W_l$$

where (W_l) is also a collection of independent standard Normal random variables. Then

$$\begin{aligned} X_l &= \mu + (\sigma Z_l + \tau W_l), \\ &\Rightarrow X_l \sim \text{Normal}(\mu, \tau^2 + \sigma^2) \end{aligned}$$

independently across l, i.e. this two stage model can be expressed as a one stage model in which (X_l) is a collection of independent and identically distributed variables with specified mean and variance.

This is an important result which should be borne in mind when working with hierarchical models. Adding extra levels in a bid to extend the model does not necessarily alter its form.

If we now consider the following 'extra' levels as a potential extension to the hierarchical model introduced in this chapter (at this point simplifying by considering a univariate G_l for each subpopulation $(\mathcal{P}_l; l = 1, \ldots, \eta)$ and a single locus):

Stage 1

$$G_l|(\theta_i; i = 1, \ldots, \eta, \gamma) \sim \text{Beta}\left(\frac{1-\theta_l}{\theta_l}\gamma, \frac{1-\theta_l}{\theta_l}(1-\gamma)\right), \tag{14}$$

independently across $(l; l = 1, \ldots, \eta)$.

Stage 2

$$\theta_l|\theta \sim \text{Beta}(k\theta, k(1-\theta)), \tag{15}$$

independently across l, where k is a fixed parameter. The density of $\boldsymbol{G}(= (G_1, \ldots, G_\eta))$ is given by

$$\begin{aligned} f(\boldsymbol{G}|\theta,\gamma) &= \int f(\boldsymbol{G}|\boldsymbol{\theta},\theta,\gamma).\pi(\boldsymbol{\theta}|\theta)d\theta_1 \ldots d\theta_\eta \\ &= \int \prod_{l=1}^{\eta} f(G_l|\theta_l,\gamma).\prod_{l=1}^{\eta}\pi(\theta_l|\theta)d\theta_1 \ldots d\theta_\eta \\ &= \prod_{l=1}^{\eta}\left[\int_0^1 \frac{\Gamma\left(\frac{1-\theta_l}{\theta_l}\right)}{\Gamma\left(\frac{1-\theta_l}{\theta_l}\gamma\right)\Gamma\left(\frac{1-\theta_l}{\theta_l}(1-\gamma)\right)} G_l^{\frac{1-\theta_l}{\theta_l}\gamma-1}(1-G_l)^{\frac{1-\theta_l}{\theta_l}(1-\gamma)-1}\right. \\ &\qquad \left.\times\frac{\Gamma(k)}{\Gamma(k\theta)\Gamma(k(1-\theta))}\theta_l^{k\theta-1}(1-\theta_l)^{k(1-\theta)-1}d\theta_l\right]. \end{aligned}$$

The probability density function of $\boldsymbol{G}$ conditional upon θ is a product of identical terms, one for each subpopulation l, i.e. the subpopulation allele frequencies $(G_1, \ldots, G_\eta)$ are independent and identically distributed.

This is to be compared with the simpler model originally introduced:

$$G_l|\theta,\gamma \sim \text{Beta}\left(\frac{1-\theta}{\theta}\gamma, \frac{1-\theta}{\theta}(1-\gamma)\right)$$

independent across l, identically distributed with density

$$f(G_l|\gamma,\theta) = \frac{\Gamma(\frac{1-\theta}{\theta})}{\Gamma(\frac{1-\theta}{\theta}\gamma)\Gamma(\frac{1-\theta}{\theta}(1-\gamma))} G_l^{\frac{1-\theta}{\theta}\gamma-1}(1-G_l)^{\frac{1-\theta}{\theta}(1-\gamma)-1}.$$

The thinking behind the introduction of the extra level in the hierarchy is that it increases the flexibility of the model, and that it reflects more closely the population substructure exhibited in the 'real world'. However, it can be seen that the form of the model is essentially unchanged. In both cases, conditional upon (θ, γ), the random variables (G_l) are independent and identically distributed.

Adding the extra level will lead to a greater degree of complexity in the model, increasing the difficulty in calculating results of interest. It is therefore appropriate to fully establish its necessity before including it in the model to be used as a basis for calculations within this thesis.

The conditional expectation and variance of G_l given (θ, γ) are calculated as follows:

$$\begin{aligned} \mathrm{E}[G_l|\gamma,\theta] &= \mathrm{E}[\mathrm{E}[G_l|\theta_l,\gamma,\theta]|\gamma,\theta] \\ &= \mathrm{E}[\gamma|\gamma,\theta] \\ &= \gamma. \end{aligned} \tag{16}$$

$$\begin{aligned} \mathrm{Var}(G_l|\gamma,\theta) &= \mathrm{E}[\mathrm{var}(G_l|\theta_l,\gamma,\theta)|\gamma,\theta] \\ &\qquad +\mathrm{var}(\mathrm{E}[G_l|\theta_l,\gamma,\theta]|\gamma,\theta] \\ &= \mathrm{E}[\theta_l\gamma(1-\gamma)|\gamma,\theta] + \mathrm{var}(\gamma|\gamma,\theta) \\ &= \theta\gamma(1-\gamma). \end{aligned} \tag{17}$$

This means that the parameters of the simpler model can be selected to give a conditional mean and variance of G_l equal to the those of the 'extra level' model. This is also true of the multivariate case in which the Beta distribution of equation (14) is replaced by

$$\boldsymbol{G}_l \sim \mathrm{Dirichlet}\left(\frac{1-\theta_l}{\theta_l}\gamma(1), \frac{1-\theta_l}{\theta_l}\gamma(2), \ldots, \frac{1-\theta_l}{\theta_l}\gamma(r)\right)$$

If we decide to choose parameters in such a way, differences in the characteristics of the models will be reflected in the higher moments.

The third and fourth central moments of the more complicated model are as follows:

$$\mathrm{E}[(G_l-\gamma)^3|\gamma,\theta] = \mathrm{E}[G_l^3-3\gamma G_l^2+3\gamma^2 G_l-\gamma^3|\gamma,\theta] \quad (18)$$

$$\mathrm{E}[(G_l-\gamma)^4|\gamma,\theta] = \mathrm{E}[G_l^4-4\gamma G_l^3+6\gamma^2 G_l^2-4\gamma^3 G_l+\gamma^4|\gamma,\theta] \quad (19)$$

The conditional expectation of G_l given (γ,θ) is known to be equal under both models (16), as is the conditional expectation of G_l^2 (17).

$\mathrm{E}[G_l^3|\gamma,\theta]$ and $\mathrm{E}[G_l^4|\gamma,\theta]$ can be evaluated exactly under the single level model and approximated under the two level model:

$$\begin{aligned}\mathrm{E}[G_l^3|\gamma,\theta] &= \mathrm{E}[\mathrm{E}[G_l^3|\theta_l,\gamma,\theta,k]|\gamma,\theta] \\ &= \mathrm{E}\left[\frac{\gamma((1-\theta_l)\gamma+\theta_l)((1-\theta_l)\gamma+2\theta_l)}{(1+\theta_l)}|\gamma,\theta\right] \\ &= \mathrm{E}[\gamma((1-\theta_l)\gamma+\theta_l)((1-\theta_l)\gamma+2\theta_l) \\ &\qquad \times(1-\theta_l+\theta_l^2-\theta_l^3+\ldots)|\gamma,\theta].\end{aligned}$$

Similarly,

$$\begin{aligned}\mathrm{E}[G_l^4|\gamma,\theta] &= \mathrm{E}[E[G_l^3|\theta_l,\gamma,\theta_l] \\ &\qquad \times((1-\theta_l)\gamma+3\theta_l)(1-2\theta_l+4\theta_l^2-8\theta_l^3+\ldots|\gamma,\theta].\end{aligned}$$

Non-centralised moments of θ_l are known for given (θ,γ) under the Beta distribution of equation 15. Expanding these expressions up to θ_l^n (for a chosen value of n), the moments can be approximated for chosen values of θ, γ. The use of such approximations is justified by the fact that θ_l is known to be close to zero.

Similarly to Foreman *at al* [Foreman, Evett and Smith, 1997], we have used values of k and θ which give a conservative prior distribution for θ_l (i.e. giving a higher probability density to values of θ_l greater than we would expect based on previous studies).

The results of tables 1 and 2 are based upon a chosen value of $n=16$ in the above approximation. It can be seen that under the 2 stage model, the third

k	θ	γ	$\mathrm{E}[(G_l-\gamma)^3\|1$ stage model]	$\mathrm{E}[(G_l-\gamma)^3\|2$ stage model]
50	0.025	0.4	5.85×10^{-5}	9.97×10^{-5}
100	0.025	0.4	5.85×10^{-5}	7.97×10^{-5}
50	0.025	0.1	8.78×10^{-5}	1.50×10^{-4}
50	0.04	0.1	2.22×10^{-4}	3.15×10^{-4}

Table 1: Comparison of third centralised moments of 1 and 2 stage models.

k	θ	γ	$\mathrm{E}[(G_l-\gamma)^4\|1$ stage model]	$\mathrm{E}[(G_l-\gamma)^4\|2$ stage model]
50	0.025	0.4	1.04×10^{-4}	1.73×10^{-4}
100	0.025	0.4	1.04×10^{-4}	1.40×10^{-4}
50	0.025	0.1	1.95×10^{-5}	4.30×10^{-5}
50	0.04	0.1	5.57×10^{-5}	9.82×10^{-5}

Table 2: Comparison of fourth centralised moments of 1 and 2 stage models.

and fourth moments of G_l conditional upon (θ,γ) are higher than those under the 1 stage model. This suggests that under the two stage model, there is more weight given to higher values of θ, and the distribution is more skewed. However, the most important consideration is not if there is a difference between the two distributions, but what effect this difference has upon the results of ultimate interest, i.e. the match probabilities.

One aim of this thesis is to consider subpopulation differentiation parameters in the context of the hierarchy introduced in this chapter. By doing this we are looking to clarify the differences between some of the parameters previously introduced, and present a framework within which they can be more easily compared.

Weir and Cockerham [Weir and Cockerham, 1984] use a model which employs a subpopulation differentiation parameter at a similar level to our θ. This is discussed further in Chapter 5. They propose that subpopulation specific

parameters are used when subpopulations develop in differing proportions. The suggestion is that as subpopulations develop, the variance of allele probabilities ($\boldsymbol{G}_l$) is greater for smaller subpopulations. This means that larger values of θ_l would be required to generate $\boldsymbol{G}_l$ for these smaller subpopulations.

However, the major consideration must be the effect of any extension to the model upon our final result. As acknowledged by Devlin *et al*, estimating a number of subpopulation differentiation parameters can be inefficient particularly when, as Foreman *et al* show, results are unlikely to be greatly affected by using a single summary measure θ.

In this work we retain the simpler single differentiation parameter model as the choice between these two models does not affect the message given by this thesis or the methods followed.

The hierarchical model outlined in this chapter forms the basis of inference (Chapter 4), MCMC simulation (Chapter 6) and parameter comparison (Chapter 5) in later chapters.

4 Inference

4.1 Introduction

For each subpopulation $\mathcal{P}_l$, we wish to calculate the subpopulation match probability

$$m_l = \Pr(\boldsymbol{X}_i = \boldsymbol{y} | i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha)$$

for individuals i outside the database α.

To calculate this quantity we shall employ the hierarchical model described in Chapter 3. This requires the specification of a model generating the data, and a prior.

At this point we introduce an additional variable (I_i) to represent the subpopulation label of each individual i. Assuming there to be η subpopulations, this variable can take values $(1, \ldots, \eta)$. It is initially assumed that (I_i) is known for all individuals (including the suspect) in the database α.

As Dawid [Dawid, 1986] describes, in a general case there is no reason that a particular combination of model density $f(\boldsymbol{x}|\boldsymbol{\theta})$ for observables $\boldsymbol{X}$ given parameters Θ and prior $\pi(\boldsymbol{\theta})$ should be regarded as the only option. Any combination of model and prior which implies the same marginal distribution upon $\boldsymbol{X}$ is equally valid. When a choice exists between such combinations of prior and model, selection will depend upon the particular application being considered. A hierarchical model such as that employed in this thesis provides a clear example of this principle.

By 'splitting' the hierarchy at a particular level we define our parameters as the variables at that level. The model density is then defined as the distribution of the observables given these parameters, and the prior is given by combining the higher levels. We have a choice of which level to split the hierarchy at, each option defining a different combination of model density and prior. It must be emphasized that, at whichever level the split is made, the resultant model density and prior will imply the same marginal distribution for our data, and hence the same values for match probabilities.

The specific hierarchical model considered here provides a choice of two levels at which the described split can made.

Figure 5 shows a directed acyclic graph (DAG) representing the Bayesian hierarchical model of Chapter 3.

When such a structure is used for diagnostic purposes, it is referred to as a probabilistic expert system [Cowell, Dawid, Lauritzen and Spiegelhalter, 1999]. The system provides a tool for specifying the joint distribution of all variables. For our purposes, the 'irrelevancies' summarised in the DAG describe conditional independence properties. For example, $\boldsymbol{G}$ can be seen to be independent of $\boldsymbol{a}_\gamma$ conditional upon $\boldsymbol{a}$.

The additional labels on the DAG highlight the two possible statistical models.

Sections 4.1.1 and 4.1.2 consider these models in turn, before Section 4.2 considers inference about the match probability m_l.

4.1.1 Model I

Under model I, the likelihood is defined by the distribution of the data given the "parameter" $\boldsymbol{G}$, the collection of subpopulation allele probability vectors $(\boldsymbol{G}_l)$. Calculation of the prior involves collapsing two levels to give a conditional distribution for $\boldsymbol{G}$ given the collection of hyperprior parameters $(\boldsymbol{a}_{\gamma j}, a_\theta, b_\theta)$ across the loci $j = 1, \ldots, M$. Thus, the posterior distribution of our "parameter" $\boldsymbol{G}$ is given by

$$f(\boldsymbol{G}|\chi_\alpha = \xi_\alpha, \boldsymbol{I}) \propto \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}, \boldsymbol{I}).\pi(\boldsymbol{G}|\boldsymbol{a}_\gamma, a_\theta, b_\theta).$$

Under this model, the individuals are independent given the parameter, making the likelihood a straightforward expression (still considering maternal and paternal bands to be distinguishable),

$$\Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}, \boldsymbol{I}) = \prod_{i \in \alpha} \prod_{j=1}^{M} G_{I_i j}(x_{ij1}) G_{I_i j}(x_{ij2}),$$

where $\alpha = \{s\} \cup \delta$ is the database expanded to include the suspect, and x_{ijb} is the allele observed at (the distinguishable) band b of locus j in individual i.

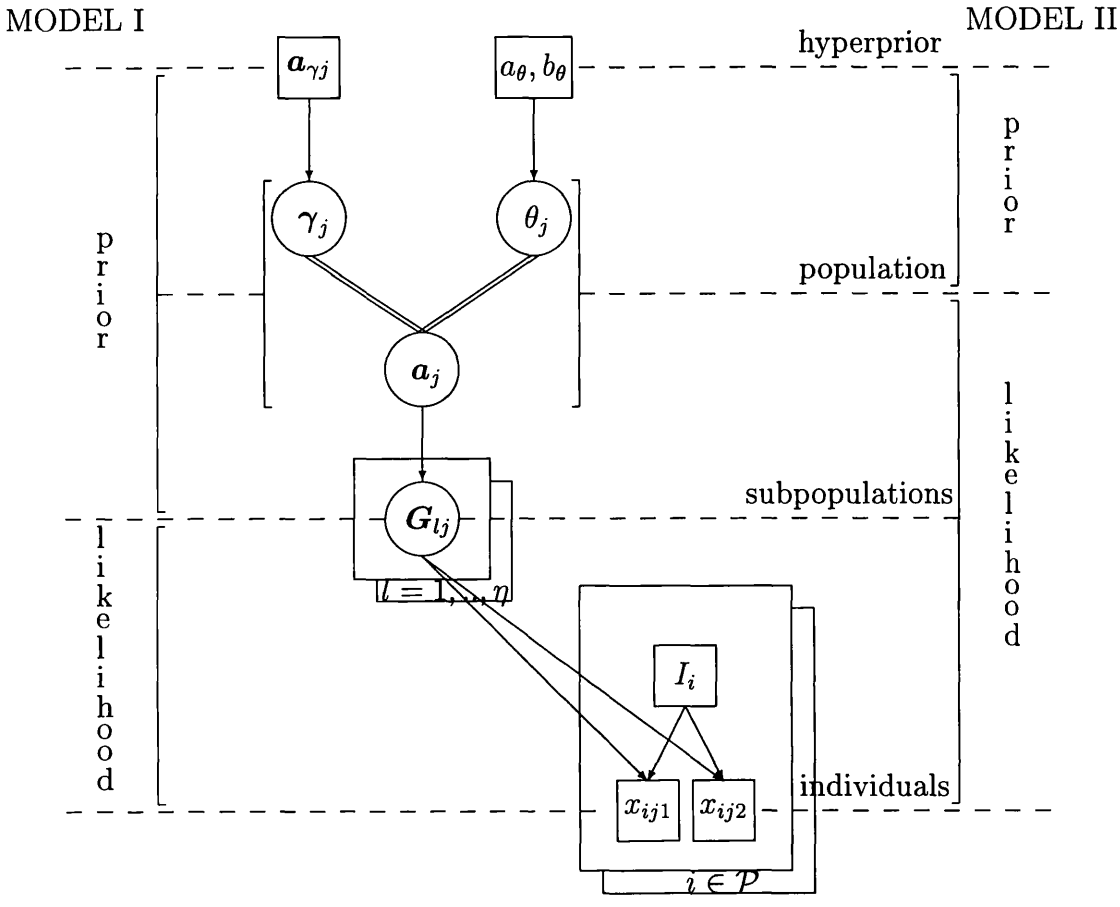


Figure 5: DAG for the case of complete information (at a fixed locus j), highlighting the two possible models. At this point we are assuming subpopulation identifiers (I_i) known for each database individual.

The subpopulation allele probabilities ($\boldsymbol{G}_l$) are not independent given the hyperparameters ($\boldsymbol{a}_{\gamma j}, a_\theta, b_\theta$), but exchangeable. This makes the prior more difficult to derive:

$$\begin{aligned}
\pi(\boldsymbol{G}|\boldsymbol{a}_\gamma, a_\theta, b_\theta) &= \mathrm{E}[\pi(\boldsymbol{G}|\boldsymbol{a}, \boldsymbol{a}_\gamma, a_\theta, b_\theta)|\boldsymbol{a}_\gamma, a_\theta, b_\theta] \\
&= \int_{\boldsymbol{a}} \pi(\boldsymbol{G}|\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{a}_\gamma, a_\theta, b_\theta)d\boldsymbol{a} \qquad (20) \\
&= \int_{\boldsymbol{a}} \prod_{l=1}^{\eta}\prod_{j=1}^{M} \frac{\Gamma(a_j(+))}{\prod_{k=1}^{r_j}\Gamma(a_j(k))} \prod_{k=1}^{r_j} G_{lj}(k)^{a_j(k)-1} \\
&\quad \times \left(\prod_{k=1}^{r_j}\left(\frac{a_j(k)}{a_j(+)}\right)^{a_{\gamma j}(k)-1}\right)\left(\frac{1}{a_j(+)+1}\right)^{a_\theta+r_j-4} \qquad (21) \\
&\quad \times \left(\frac{a_j(+)}{a_j(+)+1}\right)^{b_\theta-r_j} d\boldsymbol{a}.
\end{aligned}$$

This expression for $\pi(\boldsymbol{a}|\boldsymbol{a}_\gamma, a_\theta, b_\theta)$ is explained in more detail in Appendix B.

Equation (21) involves an integral across all $(a_j(k), j = 1, \ldots, M; k = 1, \ldots, r_j)$, where $a_j(k) > 0$. The form of this integral makes its calculation impossible using algebraic methods.

4.1.2 Model II

Under model II the "parameter" is now $\boldsymbol{a}$, and the prior is simply the prior of $\boldsymbol{a}_j$ $(= \frac{1-\theta_j}{\theta_j}\boldsymbol{\gamma}_j)$ derived in Appendix B. This prior assumes, at each locus j, independence between the ancestral population allele probabilities $\boldsymbol{\gamma}_j$ and the subpopulation differentiation parameters θ_j:

$$\begin{aligned}
\boldsymbol{\gamma}_j &\sim \text{Dirichlet}(a_{\gamma j}(1), \ldots, a_{\gamma j}(m_j)); \\
\theta_j &\sim \text{Beta}(a_\theta, b_\theta).
\end{aligned}$$

The likelihood involves collapsing the lower levels of the hierarchy to give the distribution of the data conditional upon the parameter $\boldsymbol{a}$. Conditional upon $\boldsymbol{a}$, knowledge of an individual's profile increases the information available upon his subpopulation's allele frequencies, and thus affects the probability of a fellow subpopulation member having a particular profile. We can therefore see that the individuals are not independent given the "parameter", but as there is no

reason to distinguish individuals *a priori*, they are exchangeable. The likelihood is

$$\begin{aligned}
\Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{a}, \boldsymbol{I}) &= \mathrm{E}[\Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}, \boldsymbol{a}, \boldsymbol{I})|\boldsymbol{a}, \boldsymbol{I}] \\
&= \int_{\boldsymbol{G}} \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}, \boldsymbol{I}).\pi(\boldsymbol{G}|\boldsymbol{a})d\boldsymbol{G} \\
&= \int_{\boldsymbol{G}} \prod_{i\in\alpha}\prod_{j=1}^{M} G_{I_ij}(x_{ij1})G_{I_ij}(x_{ij2}) \\
&\quad \times \prod_{l=1}^{\eta} \frac{\Gamma(a_j(+))}{\prod_{k=1}^{m_j}\Gamma(a_j(k))} \prod_{k=1}^{r_j} G_{lj}(k)^{a_j(k)-1} d\boldsymbol{G} \\
&= \int_{\boldsymbol{G}} \prod_{l=1}^{\eta}\prod_{j=1}^{m} \frac{\Gamma(a_j(+))}{\prod_{k=1}^{r_j}\Gamma(a_j(k))} \prod_{k=1}^{r_j} G_{lj}(k)^{a_j(k)+n_{lj}(k)-1} d\boldsymbol{G} \\
&= \prod_{l=1}^{\eta}\prod_{j=1}^{m} \left[\frac{\Gamma(a_j(+))}{\prod_{k=1}^{r_j}\Gamma(a_j(k))} \frac{\prod_{k=1}^{r_j}\Gamma(a_j(k)+n_{lj}(k))}{\Gamma(a_j(+)+n_{lj}(+))}\right],
\end{aligned}$$

where $n_{lj}(k)$ is the number of alleles of type k at locus j in subpopulation l observed in the database. Given the subpopulation allele probability vectors $(\boldsymbol{G}_{lj})$, these numbers $(\boldsymbol{n}_{lj})$ have a multimomial distribution,

$$(n_lj(1), n_{lj}(2), \ldots, n_{lj}(r_j)) \sim \text{Multinomial}(G_{lj}(1), G_{lj}(2), \ldots, G_{lj}(r_j)).$$

Thus, under model II, the posterior

$$f(\boldsymbol{a}|\chi_\alpha = \xi_\alpha) \propto \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{a}).\pi(\boldsymbol{a}|\boldsymbol{a}_\gamma, a_\theta, b_\theta)$$

for the chosen parameter is available up to a constant of proportionality.

4.2 Inference under the two models

The subpopulation match probability

$$m_l = \Pr(\boldsymbol{X}_i = \boldsymbol{y}|i \in \mathcal{P}_l, i \notin \alpha, \chi_\alpha = \xi_\alpha, \boldsymbol{I})$$

can be expressed as a posterior expectation of a function of either the model I parameters $(\boldsymbol{G}_{lj})$ (see (24) below) or the model II parameters $(\boldsymbol{a}_j)$ (see (25) below),

$$m_l = \int_{\boldsymbol{\phi}} \Pr(\boldsymbol{X}_i = \boldsymbol{y}|\boldsymbol{\phi}, i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha) f(\boldsymbol{\phi}|i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha) d\boldsymbol{\phi}, \qquad (22)$$

where ϕ represents the parameters under the chosen model.

Dawid [Dawid, 1986] notes that if we have data $\boldsymbol{X}_0$ and wish to predict further observables $\boldsymbol{X}_1$ it is often desirable to use a model in which, given parameters $\boldsymbol{\theta}$, $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ are independent. This means that the general predictive density of $\boldsymbol{X}_1$ given $\boldsymbol{X}_0$,

$$f(\boldsymbol{x}_1|\boldsymbol{x}_0) = \int f(\boldsymbol{x}_1|\boldsymbol{\theta}, \boldsymbol{x}_0).f(\boldsymbol{\theta}|\boldsymbol{x}_0)d\boldsymbol{\theta}$$

simplifies to

$$\int f(\boldsymbol{x}_1|\boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{x}_0)d\boldsymbol{\theta},$$

the expectation of the density $f(\boldsymbol{x}_1|\boldsymbol{\theta})$ with respect to the posterior distribution of $\boldsymbol{\theta}$ given the observed data $\boldsymbol{x}_0$.

This property is achieved by model I, under which the culprit's profile is independent of the observed data given the parameters ($\boldsymbol{G}_l$), implying that

$$\begin{aligned}\Pr(\boldsymbol{X}_i = \boldsymbol{x}|i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha) \\ = \int_{\boldsymbol{G}} \Pr(\boldsymbol{X}_i = \boldsymbol{y}|i \in \mathcal{P}_l, \boldsymbol{G}) f(\boldsymbol{G}|i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha) d\boldsymbol{G}. \quad (23)\end{aligned}$$

In this instance, however, there is no great advantage provided by this property, as under model II the probability of the culprit's profile given the parameter and data is reasonably straightforward to calculate (see (25) below).

As an aside, it is interesting to note that in their consideration of this problem, Foreman *et al* [Foreman, Evett and Smith, 1997] make a similar simplification to that above, calculating the expectation of the probability of the culprit's profile given their chosen parameters and the suspect's profile only. However, as is described in Chapter 9, the model specification made by Foreman *et al* does not justify the assumption of independence between the culprit's profile and the data conditional upon the parameters.

The match probability can be expressed as the following expectations:

$$\begin{aligned} m_l &= \mathrm{E}[\Pr(\boldsymbol{X}_i = \boldsymbol{y}|\boldsymbol{G}, i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha)|i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha] \\ &= \mathrm{E}[\prod_{j=1}^{M} 2^{h(y_{j1}, y_{j2})} G_{lj}(y_{j1}).G_{lj}(y_{j2})|i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha] \quad (24) \\ &= \mathrm{E}[m_l^{(1)}|i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha], \end{aligned}$$

where $m_l^{(1)} := \prod_{j=1}^{M} 2^{h(y_{j1},y_{j2})} G_{lj}(y_{j1}).G_{lj}(y_{j2})$ and

$$\begin{aligned} h(r,s) &= 0 \text{ if } r = s; \\ &= 1 \text{ if } r \neq s. \end{aligned}$$

Also,

$$\begin{aligned} m_l &= \mathrm{E}[\mathrm{E}[\prod_{j=1}^{M} 2^{h(y_{j1},y_{j2})} G_{lj}(y_{j1}).G_{lj}(y_{j2}) | \boldsymbol{a}, i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha] \\ &\qquad\qquad | i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha] \\ &= \mathrm{E}\left[\prod_{j=1}^{M} 2^{h(y_{j1},y_{j2})} \frac{(a_j(y_{j1}) + n_{lj}(y_{j1}))(a_j(y_{j2}) + n_{lj}(y_{j2}) + \delta_j)}{(a_j(+) + n_{lj}(+))(a_j(+) + n_{lj}(+) + 1)} \right. \\ &\qquad\qquad \left. | i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha \right] \quad (25) \\ &= \mathrm{E}[m_l^{(2)} | i \in \mathcal{P}_l, \boldsymbol{I}, \chi_\alpha = \xi_\alpha], \quad (26) \end{aligned}$$

where $m_l^{(2)} := \prod_{j=1}^{M} 2^{h(y_{j1},y_{j2})} \frac{(a_j(x_{j1})+n_{lj}(y_{j1}))(a_j(y_{j2})+n_{lj}(y_{j2})+\delta_j)}{(a_j(+)+n_{lj}(+))(a_j(+)+n_{lj}(+)+1)}$

$$\delta_j = \begin{cases} 0 & \text{if } y_{j1} \neq y_{j2} \\ 1 & \text{if } y_{j1} = y_{j2} \end{cases}$$

Ideally we would calculate one or both of the expectations in equations (24) and (25) by integrating over the posterior distribution of the appropriate parameters.

The posterior distribution of $\boldsymbol{G}$ cannot be calculated even to proportionality. The posterior of $\boldsymbol{a}$ can be calculated to proportionality, but the integration required to evaluate the normalizing constant is not feasible using non-numerical methods. The desire for an inexpensive, easily applied method would then dictate that a reasonable approximation to the match probability m_l be considered. This is where model I provides a more satisfactory answer, assuming that we know the subpopulation labels (I_i) of each individual.

If the database α contains extensive data from subpopulation $\mathcal{P}_l$, the expectation of (24) can be approximated by a consistent estimate, for example

$$\hat{m}_l^e = \prod_{j=1}^{M} c_j(\boldsymbol{y}) \hat{G}_{lj}(y_{j1}) \hat{G}_{lj}(y_{j2}), \quad (27)$$

where $c_j(\boldsymbol{y}) = 2^{h(y_{j1},y_{j2})}$ and $\hat{\boldsymbol{G}}_{lj} = \frac{\boldsymbol{n}_{lj}}{n_{lj(+)}}$, $n_{lj}(k)$ representing the number of alleles of type k observed at locus j within individuals of subpopulation $\mathcal{P}_l$ in the database. It should be noted that $(\boldsymbol{n}_{lj})$ is only known if the individual subpopulation labels (I_i) are known.

While data from a large number of individuals will provide a consistent estimate of $\boldsymbol{G}_l$ which can be used to estimate the match probability, it is not generally possible to employ a similar approximation if the statistical model specified defines $\boldsymbol{a}$ as the parameter. As noted, a very large database will provide a large amount of information upon each set of subpopulation allele frequencies $\boldsymbol{G}_l$. The collection of subpopulation frequencies at locus j $(\boldsymbol{G}_{lj}; l = 1, \ldots, \eta)$ can be considered a random sample generated from the Dirichlet$(a_j(1), \ldots, a_j(r_j))$ distribution, with

$$\boldsymbol{a}_j = \frac{1-\theta_j}{\theta_j}(\gamma_j(1), \ldots, \gamma_j(r_j)),$$

where (θ_j) is a collection of subpopulation differentiation parameters and $\boldsymbol{\gamma}_j$ is a vector of 'ancestral population' allele frequencies. This means that, however extensive the database, the 'sample size' from which we may make estimates of $\boldsymbol{\gamma}$ is limited to the number of subpopulations η. Therefore, under model II a consistent empirical estimator of $\boldsymbol{a}$, which might be substituted in (25) instead of calculating the expectation, is not generally available.

Analagous to this situation is the example of a mint producing coins with bias varying about a mean, a. Tossing each of a sample of η coins a large number of times will provide consistent estimates of the bias of each coin $(G_l; l = 1, \ldots, \eta)$. However even absolute knowledge of the bias of each of a small number of coins will not allow consistent estimation of the mean a of the process generating these biases.

A number of simplifications occur if the limit as the number η of subpopulations tends to infinity is considered. Roeder *et al* [Roeder, Escobar, Kadane and Balazs, 1998] simplify the likelihood for the population parameters $(\boldsymbol{\gamma}_j)$ and (θ_j) by assuming that the probability of more than one database individual belonging to the same subpopulation is negligible. Un-

der this assumption, all database individuals are independent given these parameters. The approach of Roeder *et al* is discussed further in Chapter 9.

4.3 Inference under unknown individual subpopulation labels

The assumption that information is available identifying each individual's subpopulation is generally unrealistic. Therefore, we can no longer condition upon $\boldsymbol{I}$ in the match probability.

Under model I,

$$\begin{aligned} m_l &= \Pr(\boldsymbol{X}_i = \boldsymbol{y} | i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha) \\ &= \mathrm{E}[\Pr(\boldsymbol{X}_i = \boldsymbol{y} | \boldsymbol{G}, \boldsymbol{I}, i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha) | i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha] \\ &= \mathrm{E}[\prod_{j=1}^{r_j} 2^{h(y_{j1}, y_{j2})} G_{lj}(y_{j1}).G_{lj}(y_{j2}) | i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha]. \end{aligned} \tag{28}$$

This is identical to equation (24), but excluding the conditioning upon $\boldsymbol{I}$. Derivation of the likelihood of $\boldsymbol{G}$ now requires a summation over all possible $\boldsymbol{I}$. It is assumed *a priori* that subpopulation identifiers (I_i) are independent across individuals. Also assuming that the subpopulation identifiers are independent of the subpopulation allele frequencies ($\boldsymbol{G}_l$):

$$\begin{aligned} l(\boldsymbol{G}) &= \Pr(\chi_\alpha = \xi_\alpha | \boldsymbol{G}) \\ &= \mathrm{E}[\Pr(\chi_\alpha = \xi_\alpha | \boldsymbol{G}, \boldsymbol{I}) | \boldsymbol{G})] \\ &= \sum_{\boldsymbol{I}} \left[\prod_{i \in \alpha} \prod_{j=1}^{M} G_{I_{ij}}(x_{ij1}).G_{I_{ij}}(x_{ij2}).\kappa(I_i) \right], \end{aligned}$$

where $\kappa(l)$ is the prior probability that an individual chosen at random is a member of subpopulation $\mathcal{P}_l$.

This is a sum over η^{n_α} terms, where n_α is the number of individuals in the database α. This number of terms is huge for any reasonably sized database.

However, the more important result of losing the subpopulation identifiers is that the consistent match probability estimator $\hat{m}_l^e$ is no longer available. Calculation of an empirical estimator would now require the above averaging

across all possible combinations of subpopulation identifiers. If at all possible, this would be very computationally expensive, defeating the object of using such an estimator.

Under model II,

$$
\begin{aligned}
m_l &= \Pr(\boldsymbol{X}_i = \boldsymbol{y} | i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha) && (29)\\
&= \mathrm{E}[\mathrm{E}[\Pr(\boldsymbol{X}_i = \boldsymbol{y} | \boldsymbol{G}, \boldsymbol{a}, \boldsymbol{I}, i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha) | \boldsymbol{a}, \boldsymbol{I}, i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha] \\
&\qquad\qquad | i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha] && (30)\\
&= \mathrm{E}\left[\prod_{j=1}^{M} 2^{h(y_{j1}, y_{j2})} \frac{(a_j(y_{j1}) + n_{lj}(y_{j1}))(a_j(y_{j2}) + n_{lj}(y_{j2}) + \delta_j)}{(a_j(+) + n_{lj}(+))(a_j(+) + n_{lj}(+) + 1)}\right. \\
&\qquad\qquad \left. | i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha, \right. && (31)\\
& && (32)
\end{aligned}
$$

where δ_j indicates $y_{j1} = y_{j2}$. Again, the data are no longer enough to evaluate $\boldsymbol{n}$.

In the absence of an empirical estimator, it would appear necessary to evaluate one of the posterior expectations described in (28) and (30). As the integration required is impossible analytically, an alternative method is required.

4.4 Markov chain Monte Carlo

When direct solutions are unavailable due to the complex nature of the required integration, we resort to Markov chain Monte Carlo (MCMC) methods. These methods are discussed generally in Chapter 6, and specific schemes outlined in Chapter 7.

It is important to relate 'real world' knowledge to the mathematical problem at hand. There is a variety of possibilities regarding the state of our knowledge of subpopulation identifiers and related quantities such as subpopulation proportions $\boldsymbol{\alpha}$. Chapter 8 seeks to describe the effect of varying this knowledge upon match probability calculations and consequently upon posterior probabilities of guilt.

5 Measures of subpopulation differentiation

Throughout the extensive literature concerning population substructure, parameters which it is claimed quantify the differentiation between subpopulations are defined by a number of authors.

Indeed the parameter $\boldsymbol{\theta} = (\theta_j; j = 1, \ldots, M)$ of the model defined in Chapter 3 can be seen as a subpopulation differentiation parameter. If reasonable models are to be developed, it is very useful to be able to relate the parameters of such models to meaningful 'real world' quantities. What does $\boldsymbol{\theta}$ really mean? Which, if any, of the population genetics differentiation parameters is it equivalent to? These are interesting and important questions which are not always answered satisfactorily.

As will be seen in this chapter, the degree of difference between parameters which have in some instances been regarded as equivalent is as great as that between sample and population variances.

To demonstrate this, the subpopulation differentiation parameters of Nei [Nei,1987] and Weir and Cockerham [Weir and Cockerham, 1984] are compared in the context of the hierarchical model described in Chapter 3. Papers employing models of population substructure often make no distinction between the two, classing them as equivalent.

To simplify matters, the two allele (0 and 1) single binary locus case is considered here, where $\Pr(X_{ib} = 1|\boldsymbol{G}) = G_l$ for each band $(b = 1, 2)$ within each individual i in subpopulation $\mathcal{P}_l$.

We define $\boldsymbol{\kappa}$ $(= \kappa(l); l = 1, \ldots, \eta)$ as the collection of prior probabilities of a randomly selected individual i being in subpopulation $\mathcal{P}_l$:

$$\kappa(l) = \Pr(I_i = l),$$

where (I_i) is the collection of individual subpopulation identifiers.

5.1 Nei

Nei considers the fixation index F to be a function of the parameters defining the allele probabilities (G_l) of the η actual subpopulations. Using only these present generation parameters, no assumption is required about pedigrees of individuals, selection and migration in the past.

We define

$$g = \sum_{l=1}^{\eta} \kappa(l) G_l,$$

the average of the subpopulation allele '1' probabilities, weighted by $\boldsymbol{\kappa}$.

If the current population is in Hardy-Weinberg equilibrium we have the case where $\eta = 1$ and $g = G_1$:

$$\begin{aligned}
\Pr(\boldsymbol{X}_i = (0,0)|g, \boldsymbol{G}) &= (1-g)^2; \\
\Pr(\boldsymbol{X}_i = (0,1)|g, \boldsymbol{G}) &= 2g(1-g); \\
\Pr(\boldsymbol{X}_i = (1,1)|g, \boldsymbol{G}) &= g^2.
\end{aligned}$$

Following the thinking of Wright [Wright, 1951], any departure from these Hardy-Weinberg probabilities can be measured by the fixation index F so that

$$\Pr(\boldsymbol{X}_i = (0,0)|g, \boldsymbol{G}) = (1-F)(1-g)^2 + F(1-g); \tag{33}$$

$$\Pr(\boldsymbol{X}_i = (0,1)|g, \boldsymbol{G}) = 2(1-F)g(1-g); \tag{34}$$

$$\Pr(\boldsymbol{X}_i = (1,1)|g, \boldsymbol{G}) = (1-F)g^2 + Fg. \tag{35}$$

There are a number of possible causes of departure from Hardy-Weinberg proportions, including inbreeding, assortative mating and selection. At this point, however, we consider only the effect upon overall proportions of a population being split into η randomly mating subpopulations. The overall pair probabilities are then given by

$$\begin{aligned}
\Pr(\boldsymbol{X}_i = (0,0)|g, \boldsymbol{G}) &= \sum_{l=1}^{\eta} \kappa(l)(1-G_l)^2 \\
&= (1-g)^2 + \sum_{l=1}^{\eta} \kappa(l) \left[(1-G_l) - (1-g)\right]^2 \\
&= (1-g)^2 + \sigma^2; \qquad (36)
\end{aligned}$$

$$\Pr(\boldsymbol{X}_i = (0,1)|g, \boldsymbol{G}) = 2\sum_{l=1}^{\eta} \kappa(l) G_l(1 - G_l) = 2g(1-g) - 2\sigma^2; \quad (37)$$

$$\Pr(\boldsymbol{X}_i = (1,1)|g, \boldsymbol{G}) = \sum_{l=1}^{\eta} \kappa(l) G_l^2 = g^2 + \sigma^2; \quad (38)$$

where $\sigma^2 = \sum_{l=1}^{\eta} \kappa(l)(G_l - g)^2$, the 'sample' variance of the gene '1' proportion across subpopulations. The homozygotic frequencies are increased above the Hardy-Weinberg level by an amount σ^2 with an appropriate reduction in the heterozygotic frequency. Comparison of equations (33)-(35) to (36)-(38) shows that the fixation index in the case of 2 alleles is then given by

$$F = \frac{\sigma^2}{g(1-g)}.$$

5.2 Weir and Cockerham

In terms of the hierarchical model, Weir and Cockerham work at the level above Nei. Weir and Cockerham define a *coancestry coefficient* θ that does not depend upon the number of subpopulations observed and governs the level of variability of allele probability (G_l) across subpopulations. It is a parameter related to the ancestral population from which the currently observed subpopulations have developed. To Weir and Cockerham, the observed subpopulations are just a sample of those that could have evolved from the ancestral population under similar conditions. This is directly comparable to the hierarchical model in which the η subpopulations are a sample generated given the parameters of the level above.

The subpopulations ($\mathcal{P}_l; l = 1, \ldots, \eta$) are assumed to have descended separately from the single ancestral population.

Random mating is assumed within subpopulations, and we define γ as the mean of the process generating subpopulation probabilities (G_l), i.e.

$$\mathrm{E}[G_l|\gamma] = \gamma,$$

independently for all l. Weir and Cockerham then define their subpopulation differentiation parameter θ by

$$\Pr(\boldsymbol{X}_i = (1,1)|\gamma, \theta) = \gamma^2 + \gamma(1-\gamma)\theta. \quad (39)$$

Referring to the DAG of Figure 5, the probabilities (G_l) are defined at the level of the subpopulations. We assume that these probabilities are generated with a mean γ and variance V. The important point here is that γ and V are parameters at the level above (G_l) in the hierarchy. In the previous section, g and σ^2 are defined at the same level as (G_l).

$$\begin{aligned}
\Pr(\boldsymbol{X}_i = (1,1)|\gamma, V) &= \mathrm{E}[\Pr(\boldsymbol{X}_i = (1,1)|\gamma, V) \\
&= \mathrm{E}[\Pr(\boldsymbol{X}_i = (1,1)|\boldsymbol{G}, \gamma, V) \\
&= \mathrm{E}[\kappa(l)G_l^2|\gamma, V) \\
&= \sum_{l=1}^{\eta} \kappa(l)(V + \gamma^2) = V + \gamma^2 \qquad (40)
\end{aligned}$$

By comparison of equations (39) and (40), we see that $\theta = \frac{V}{\gamma(1-\gamma)}$, a measure of subpopulation differentiation at the level above Nei's parameter F.

5.3 How are the measures of subpopulation differentiation related?

Firstly it is interesting to compare directly the definitions of Nei's parameter F, and Weir and Cockerham's θ.

$$\begin{aligned}
F &= \frac{\sigma^2}{g(1-g)}; \\
\theta &= \frac{V}{\gamma(1-\gamma)}.
\end{aligned}$$

Presented in this way, one can see the justification behind the earlier statement that the comparison is similar to that between a sample variance and a population variance. The differences in definition between $g = \sum_{l=1}^{\eta} \kappa(l)G_l$ and $\gamma = E[G_l|\gamma]$ should also be noted.

It is also helpful to consider the expectation of the 'sample variance' σ^2 $(= \sum_{l=1}^{\eta} \kappa(l)(G_l - g)^2)$, given γ and the prior variance V and γ.

$$\begin{aligned}
\mathrm{E}[\sigma^2|V, \gamma] &= \mathrm{E}[\sum_l \kappa(l)(G_l - g)^2|V, \gamma] \\
&= \mathrm{E}[\sum_l \kappa(l)G_l^2 - g^2|V, \gamma]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_l \kappa(l)\mathrm{E}[G_l^2|V,\gamma] - \mathrm{Var}(g|V,\gamma) - \mathrm{E}[g|V,\gamma]^2 \\
&= \mathrm{Var}(g|V,\gamma) + E[g|V,\gamma]^2 - \mathrm{Var}(\sum_l \kappa(l)G_l|V,\gamma) - \gamma^2 \\
&= V + \gamma^2 - \sum_l \kappa(l)^2 V - \gamma^2 \\
&= V(1 - \sum_l \kappa(l)^2) \\
\Rightarrow\ & \mathrm{E}[\sigma^2|V,\gamma] = V\sum_l \kappa(l)(1-\kappa(l))
\end{aligned}
$$

In the special case where $\kappa(l) = \frac{1}{\eta}$ for $l = 1, \ldots, \eta$, this simplifies to

$$\mathrm{E}[\sigma^2|V,\gamma] = \frac{\eta - 1}{\eta}V,$$

meaning that

$$\mathrm{E}[Fg(1-g)|\theta,\gamma] = \frac{\eta - 1}{\eta}\theta\gamma(1-\gamma). \tag{41}$$

Equation (41) clarifies the role of the subpopulation differentiation parameters, θ combining with γ to define the process generating $\boldsymbol{G}$, and consequently F.

6 Markov chain Monte Carlo methods

6.1 Introduction

Generally the integration required to calculate match probabilities (equations (24) and (25)) is impossible analytically. For this reason numerical methods are required to obtain estimates of match probabilities. This chapter introduces these methods generally before describing some specific techniques used in later chapters.

Often when applying Bayesian methods we can obtain the posterior distribution only up to a constant of proportionality. Using an illustrative parameter θ,

$$f(\theta|data) = \frac{f(data|\theta).\pi(\theta)}{\int f(data|\theta).\pi(\theta)d\theta} \tag{42}$$

The integral in the denominator of this expression is often impossible to evaluate analytically. The evaluation of posterior expectations and other moments of interest is similarly difficult. Possible approaches include numerical approximation, analytic approximation, and Monte Carlo integration, including Markov chain Monte Carlo (MCMC).

Monte Carlo integration [Gilks, Richardson and Spiegelhalter, 1996] approximates the expectation of a function $f(X)$, where X has a density $\pi(.)$, by

$$\hat{f}_n = \frac{1}{n}\sum_{t=1}^{n} f(X_t), \tag{43}$$

the mean of a sample $(X_t; t = 1, \ldots, n)$ drawn from $\pi(.)$.

Ideally, we would draw an independent sample from $\pi(.)$. However, in many cases this is not feasible, and not absolutely necessary as long as we have some process which, "loosely speaking, draws samples throughout the support of $\pi(.)$ in the correct proportions" [Gilks, Richardson and Spiegelhalter, 1996a]. One way of drawing such samples is through construction of a Markov chain with stationary distribution $\pi(.)$. Substitution of such a sample into equation (43) represents one form of conducting a Markov chain Monte Carlo calculation.

6.2 Markov chains

We consider a sequence of variables $X_0, X_1, X_2, \ldots$. This is a Markov chain if the distribution of X_t given all previous values $(X_0, \ldots, X_{t-1})$ depends only on the most recent X_{t-1}. This means that

$$P(X_t \in A | X_0, \ldots, X_t) = P(X_t \in A | X_{t-1}) \tag{44}$$

for any set A, where $P(.|.)$ denotes a conditional probability. To avoid excessive technical detail in the following descriptions, in this section we restrict attention to Markov chains of discrete state-space. Thus, transition probabilities take the form $P_{ij}(t) = P(X_t = j | X_0 = i)$. Extending the theory to more general state-spaces does not require any major new concepts.

If the following three properties are satisfied, the distribution of X_t can be shown [Meyn and Tweedie, 1993] to converge to a stationary distribution with distribution $\phi(.)$, as $t \to \infty$. We define τ_{ii} as the time of first return to state i, $(\tau_{ii} = \min\{t > 0 : X_t = i | X_0 = i\})$. The chain must be

(i) *irreducible*, i.e. for all i, j, there exists a $t > 0$ such that $P_{ij}(t) > 0$;

(ii) *aperiodic*, i.e.

$$\text{greatest common divider } \{t > 0 : P_{ii}(t) > 0\} = 1;$$

(iii) *positive recurrent.* An irreducible chain $\boldsymbol{X}$ is recurrent if $P(\tau_{ii} < \infty) = 1$ for some (and hence for all) i. The chain $\boldsymbol{X}$ is positive recurrent if it is recurrent and $\mathrm{E}[\tau_{ii}] < \infty$ for some (and hence for all) i.

Under the above conditions,

$$P_{ij}(t) \to \phi(j) \text{ as } t \to \infty \text{ for all } i, j.$$

The Ergodic theorem states that "if X is positive recurrent and aperiodic, then

$$\text{if } \mathrm{E}_\phi[|f(X)|] < \infty, \text{ then } \Pr\left(\hat{f}_n \to \mathrm{E}_\phi[f(X)]\right) = 1,$$

where $\mathrm{E}_\phi[f(X)] = \int_y f(y)\phi(y)dy$, the expectation of $f(X)$ with respect to $\phi(.)$."

One application of MCMC would first conduct a 'burn-in' of m iterations, long enough for the chain (of stationary distribution $\pi(.)$) to 'forget' its starting point, and then use the following $n-m$ points. We would thus use

$$\hat{f}_{n-m} = \frac{1}{n-m} \sum_{t=m+1}^{n} f(x^{(t)})$$

to estimate $\mathrm{E}[f(X)]$.

The Ergodic theorem ensures that this is a consistent estimator. This is fine in theory, but in practice there are a number of issues which must be considered.

(i) How long should the burn-in m be? It should be long enough for the chain to 'forget' its starting position.

- The most common method of burn-in determination is simply visual inspection of trace plots of the chain. There are also a number of more formal tools [Cowles and Carlin, 1994], all of which require a more technical analysis of the output.

(ii) Are X_t and X_{t+1} highly correlated?

- To reduce the correlation between consecutive values of the chain used for estimation, the original chain can be thinned [Gilks, Richardson and Spiegelhalter, 1996]. This involves using every r^{th} iteration.

(iii) For how many iterations should the chain be run? It should be long enough to provide estimators of an acceptable precision, but, on a practical note, it is not desirable to run the MCMC scheme for a far longer time than is necessary.

- An informal way to tackle this problem is to run a number of parallel chains from different starting points, and compare the estimates that they give. The run length should be increased until the estimates agree adequately. More formal methods of run length determination

have been described by Roberts [Roberts, 1992] and Raftery and Lewis [Raftery and Lewis, 1995].

- One method, proposed by Gelman and Rubin [Gelman and Rubin, 1992], is based upon the analysis of variance of a series of parallel runs. The basic assumption is that before the chains have converged, the variability across all sequences combined will be much greater than that within each individual sequence. Assuming that we are interested in a scalar quantity θ, and we have m chains $(\theta_{ij}; i = 1, \ldots, m, j = 1, \ldots, n)$ of length n, the method proceeds by calculating the between-sequence variance B and the within-sequence variance W:

$$B = \frac{n}{m-1} \sum_{i=1}^{m} (\bar{\theta}_{i.} - \bar{\theta}_{..})^2,$$

where $\bar{\theta}_{i.} = \frac{1}{n} \sum_{j=1}^{n} \theta_{ij}, \bar{\theta}_{..} = \frac{1}{m} \sum_{i=1}^{m} \bar{\theta}_{i.}$;

$$W = \frac{1}{m} \sum_{i=1}^{m} s_i^2,$$

where $s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (\theta_{ij} - \bar{\theta}_{i.})^2$.

These quantities are used to construct two estimates of the variance of θ in the target distibution, $\hat{\text{var}}_1(\theta) = \frac{n-1}{n} W + \frac{1}{n} B$ and $\hat{\text{var}}_2(\theta) = W$. As n tends to ∞, these estimates should tend to $\text{var}(\theta)$ from opposite directions. Convergence is monitored by considering the ratio $\hat{R}$ where $\sqrt{\hat{R}} = \sqrt{\frac{\hat{\text{var}}_1(\theta)}{\hat{\text{var}}_2(\theta)}}$. As the chain converges, $\hat{R}$ declines towards 1. Gelman and Rubin consider that chains should be run until the values of $\hat{R}$ for the quantities of interest are less than 1.2.

6.3 Metropolis-Hastings algorithm

We now revert to the continuous state-space case. A Markov chain whose stationary distribution possesses the target density $\pi(.)$ may be constructed using the Metropolis-Hastings algorithm [Hastings, 1970].

Hastings demonstrates how to design a Markov chain $(X_0, X_1, \ldots)$ satisfying

the detailed balance equation,

$$\pi(x_t)f(x_{t+1}|x_t) = \pi(x_{t+1})f(x_t|x_{t+1}). \tag{45}$$

This ensures that the Markov chain has a stationary distribution with the target density $\pi(.)$. Integrating both sides with respect to x_t,

$$\int \pi(x_t)f(x_{t+1}|x_t)\mathrm{d}x_t = \pi(x_{t+1}).$$

This shows that if X_t has been sampled from $\pi(.)$, X_{t+1} will also have density $\pi(.)$. Thus, if we reach an iteration t at which a value is sampled from the stationary distribution, all subsequent values in the chain will be sampled from this distribution also.

Hastings shows that such a chain can be constructed by employing the following algorithm. At iteration $t+1$,

(i) sample a value y from a proposal density $q(y|x_t)$;

(ii) calculate the acceptance probability

$$\alpha_{acc} = \min\left\{1, \frac{\pi(y).q(x_t|y)}{\pi(x_t).q(y|x_t)}\right\}; \tag{46}$$

(iii) generate a Uniform$(0,1)$ random variable U;

$$\text{if } U < \alpha_{acc}, \text{ accept y: set } x_{t+1} = y;$$
$$\text{if } U > \alpha_{acc}, \text{ reject y: set } x_{t+1} = x_t.$$

The proposal distribution must be such that the conditions of Section 6.2 are satisfied. Ideally we would like the proposal density to be as close as possible to the target density. If the variance of proposed values is too great, a large proportion of the proposed moves will be to points of small target density and will be rejected. Conversely, if the variability in the proposal is too small there will be a high acceptance rate, but the chain will take a long time to move about the full support of the density.

It is often convenient to update components of a multivariate variable $\boldsymbol{x}$ individually. Assuming that $\boldsymbol{x}$ can be split into components of varying dimension

$(x_1, x_2, \ldots, x_r)$, iteration t of the sampler would involve r steps, updating each of the r components in turn. Denoting the value of X_i at the end of iteration t by $X_{t.i}$, we define

$$\boldsymbol{x}_{t.-i} = x_{t+1.1}, \ldots, x_{t+1.i-1}, x_{t.i+1}, \ldots, x_{t.r}.$$

At iteration $t+1$, step i would proceed in the following way:

Generate a candidate value y_i from the proposal distribution $q_i(y_i|x_{t.i}, \boldsymbol{x}_{t.-i})$. This value is accepted as $x_{t+1.i}$ with probability

$$\alpha_i = \min\left(1, \frac{\pi(y_i|\boldsymbol{x}_{t.-i}).q_i(x_{t.i}|y_i, \boldsymbol{x}_{t.-i})}{\pi(x_{t.i}|\boldsymbol{x}_{t.-i}).q_i(y_i|x_{t.i}, \boldsymbol{x}_{t.-i})}\right).$$

Hierarchical models are natural candidates for single component Metropolis updating as the complete vector of parameters can be split into components by the levels of the model.

6.3.1 Random walk Metropolis algorithm

The random walk Metropolis algorithm is a special case in which the proposal distribution is a function of the distance from the previous iteration, i.e. $q(x|y) = q(|x-y|)$.

Application of this method often involves the use of a Normal proposal distribution with mean given by the value of the previous iteration.

6.4 Gibbs sampling

Many practical applications of MCMC involve Gibbs sampling [Geman and Geman, 1984]. At each iteration, each component of a vector of variables is updated in turn, conditional on the current values of the other components. It is assumed that we can sample from the full conditional distributions. This is a special case of single-component Metropolis-Hastings in which the proposal distribution for each component is its full conditional. Substitution of $q(y|x) = \pi(y)$ into (46) shows that the acceptance probability is always 1.

6.5 Hybrid MCMC schemes

In some cases, a hybrid MCMC scheme is employed. This involves the combination of more than one MCMC technique into a single MCMC scheme. An example of this is employed by Foreman *et al.* This involves a Gibbs sampling set up in which the full conditional of one of the parameters (θ_j) cannot be sampled from. For this variable, a Metropolis-Hastings step is introduced. Within one iteration of each Gibbs sampler, a number of these Metropolis-Hastings iterations is carried out. This should ensure that there is an acceptable rate of Gibbs sampler iterations in which $(\theta_j^{(t+1)}) \neq (\theta_j^{(t)})$.

6.6 Techniques to improve mixing

Metropolis-Hastings, and in particular Gibbs sampling, is the most commonly used MCMC technique. In some cases however, straightforward application of a Gibbs sampling scheme does not provide a chain which can be used as a sample from the target distribution.

While the chain of a correctly designed MCMC scheme theoretically converges to the target distribution, in practice a multimodal target distribution will often cause problems. This is because in many cases the chain will remain in a single mode for an extended period of time rather than mixing across the full support of the target density. In such instances, methods are required which ensure movement between the modes. Examples of these methods include importance sampling and simulated tempering, both of which are described here.

6.7 Simulated tempering

Simulated tempering [Geyer and Thompon, 1993], and the closely related Metropolis-coupled MCMC [Geyer, 1994], are two of the strategies commonly used if an MCMC scheme is not mixing across a multimodal target distribution.

The chain generated in a simulated tempering scheme randomly moves across $(v+1)$ MCMC samplers, each with a different stationary distribution specified

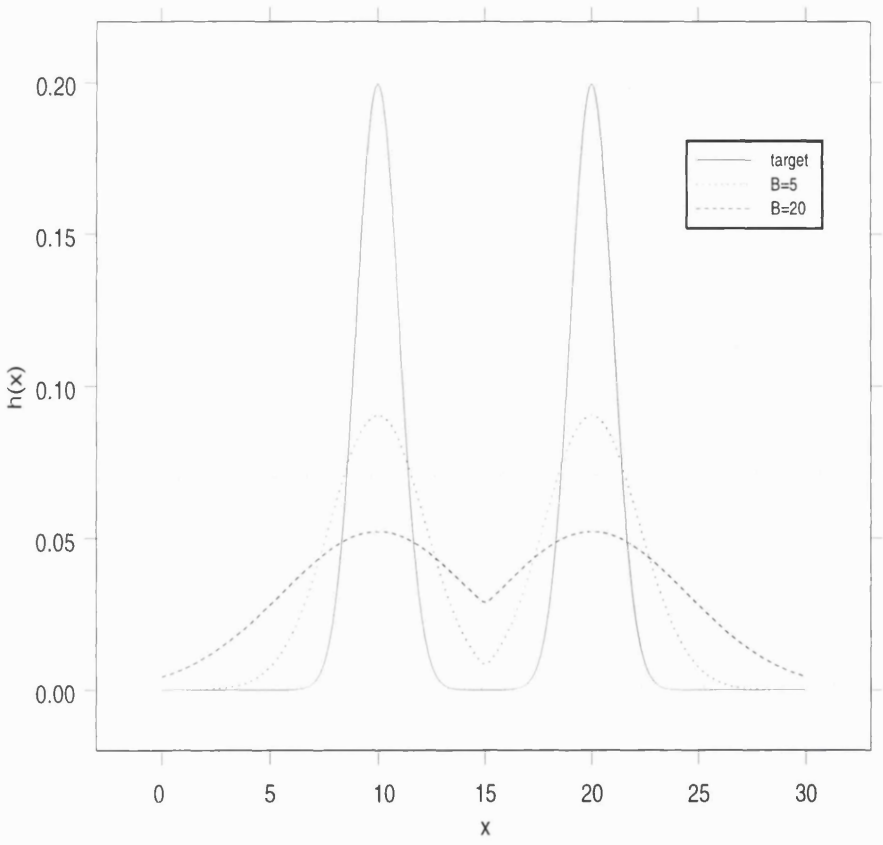


Figure 6: Normalized densities at various temperatures. This demonstrates the effect of "heating" the target density according to the formula of equation 47

by the unnormalized density $h_i(.), i = 0, \ldots, v$. We will generally define $h_0(.)$ as the density of the target distribution of interest and, at the end of the simulated tempering run, only those iterations generated at level 0 will be retained to create the final chain from which estimates are calculated.

The different samplers used can be thought of as corresponding to different "temperatures", heating up the target distribution $\pi_0(.)$ to improve mixing. This heating is often of the form

$$h_i(x) = h(x)^{\frac{1}{B_i}}, \qquad B_i > 0 \text{ for all } i. \tag{47}$$

Figure 6 shows the effect of such heating.

In our example the 'valley' between the two modes of the target density (Figure 6) is relatively deep, so that the original MCMC scheme rarely moves into it, making transitions between the two modes very unlikely. The heating works to flatten the overall distribution, as seen in figure 6. This increases the probability of movement into the 'valley', and thus between the modes.

The state of the chain at time t is represented by the pair (x_t, l_t), where l_t is the sampler used at iteration t. At each iteration t, in addition to the usual MCMC sampling, a move from the current sampler i to an alternative sampler j is proposed. This move is accepted with probability

$$\alpha_{\text{acc}} = \min\left\{1, \frac{c_j \pi_j(x^{(t)}) q_{j,i}}{c_i \pi_i(x^{(t)}) q_{i,j}}\right\}, \tag{48}$$

where $(c_i; i = 0, \ldots, v)$ are conveniently chosen constants,

$q_{i,j}$ is the probability that sampler j is proposed given that sampler i is currently being used.

The constants $(c_i, i = 0, \ldots, v)$ are chosen so that the chain spends an approximately equal amount of time in each sampler. Various techniques to achieve this have been suggested [Geyer and Thompon, 1993].

For the method to work properly, it is important that the chain move about the various samplers. It has been suggested that an appropriate acceptance rate for moves between samplers is between 20% and 40%. This can be achieved by appropriate adjustment of the temperature changes between the levels.

When deciding upon the number of levels, one must make v large enough to improve mixing. However, it must be remembered that only about $\frac{1}{v+1}$ of the iterations will be kept for analysis. Thus, too many levels will necessitate the running of the scheme for a long time to provide a sample large enough to make estimates of acceptable accuracy.

6.8 Importance sampling

Importance sampling [Hammersley and Handscomb, 1964, Geweke, 1989] is another technique aimed at improving the mixing of a Markov chain with stationary density $\pi(.)$. If the quantity of interest is $\mathrm{E}[h(\theta)|y]$ and we cannot sample directly from the posterior density $\pi(\theta|y)$, but can sample from an alternative distribution defined by the density $\pi^*(.)$,

$$\mathrm{E}[h(\theta)|y] = \int \frac{h(\theta)\pi(\theta|y)}{\pi^*(\theta)} \pi^*(\theta) d\theta.$$

This can be estimated by the weighted sum $\frac{1}{n}\sum_{t=1}^{L} h(\theta^{(t)})w(\theta^{(t)})$, where $w(\theta^{(t)}) = \frac{\pi(\theta^{(t)}|y)}{g(\theta^{(t)})}$. The approximating density should be chosen such that the ratio $\frac{h\pi}{g}$ is roughly constant.

If it is not possible to calculate $\pi(\theta|y)$, but an unnormalised density $q(\theta|y)$ can be calculated, the desired posterior expectation can be estimated by

$$\frac{\frac{1}{n}\sum_{t=1}^{n} h(\theta^{(t)})w(\theta^{(t)})}{\frac{1}{n}\sum_{t=1}^{n} w(\theta^{(t)})}$$

where $w(\theta^{(t)}) = \frac{\pi(\theta^{(t)}|y)}{\pi^*(\theta^{(t)})}$.

7 Application of Markov chain Monte Carlo

7.1 Introduction

The hierarchical model described in Chapter 3 involves the following levels:

(i) inheritance of a particular profile $\boldsymbol{x}_i$ by each individual i within a subpopulation $(\mathcal{P}_l, l = 1, \ldots, \eta)$,

$$\Pr(X_{ijb} = k|\boldsymbol{G}) = G_{lj}(k),$$

independently across i, l, band $(b = 1, 2)$ and locus $(j = 1, \ldots, M)$, the collection of observable alleles at a particular locus being denoted by $k = 1, \ldots, r_j$.

(ii) Generation of the allele probabilities $\boldsymbol{G}$ in each subpopulation,

$$G_{lj} \sim \text{Dirichlet}(a_j(1), a_j(2), \ldots, a_j(r_j)), \text{ independently for all } l, j,$$

where $\boldsymbol{a}_j = \frac{1-\theta_j}{\theta_j}\boldsymbol{\gamma}_j$.

(iii) The generation of the ancestral population parameters, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, from a 'hyperprior' distribution,

$$\boldsymbol{\gamma}_j \sim \text{Dirichlet}(a_\gamma(1), \ldots, a_\gamma(r_j)); \quad (49)$$

$$\theta_j \sim \text{Beta}(a_\theta, b_\theta). \quad (50)$$

It is described in Chapter 4 how two alternative statistical models featuring either $\boldsymbol{G} = (\boldsymbol{G}_{lj})$ or $\boldsymbol{a} = (\boldsymbol{a}_j)$ as the "parameter" can be defined.

The subpopulation match probability m_l for an arbitrary individual i (outside the database α) in subpopulation $\mathcal{P}_l$ is defined as

$$m_l = \Pr(\boldsymbol{X}_i = \boldsymbol{y}|i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha),$$

where χ_α represents the collection of profiles in the database α. As shown in Chapter 4 it can be expressed as the posterior expectation of a function of the "parameters" (and possibly the data) dictated by the chosen model,

$$m_l = \mathrm{E}[m_l^{(1)} | \chi_\alpha = \xi_\alpha] \tag{51}$$

where $m_l^{(1)} = c \prod_{j=1}^{M} G_{lj}(y_{j1}) G_{lj}(y_{j2})$, or

$$m_l = \mathrm{E}[m_l^{(2)} | \chi_\alpha = \xi_\alpha] \tag{52}$$

where $m_l^{(2)} = c \prod_{j=1}^{M} \frac{(a_j(y_{j1})+n_{lj}(y_{j1}))(a_j(y_{j2})+n_{lj}(y_{j2})+\delta_j}{(a_j(+)+n_j(+))(a_j(+)+n_j(+)+1)}$ where $c = \prod_{j=1}^{M} 2^{h(y_{j1},y_{j2})}$ and δ_j indicates if $y_{j1} = y_{j2}$.

To find an answer analytically would require the calculation of the posterior density of $\boldsymbol{a}$, or of $\boldsymbol{G}$, complete with normalizing constant, and then integrating to find the expectation of the function in equation (52), or in equation (51). This is not a feasible calculation in this instance, not an unusual situation in Bayesian applications in which the required integration is often impracticable. Thus we resort to MCMC methods, as discussed in Chapter 6.

Initially considered is the simplest situation, that in which the subpopulation membership (I_i) of each individual i is assumed known. We then go on to consider the adjustments required in the absence of such knowledge.

7.2 Subpopulation labels known

When subpopulation labels for the database individuals are known, the match probability can be easily estimated by taking a product across the suspect profile of empirical allele frequencies

$$\hat{G}_{lj}(k) = \frac{n_{lj}(k)}{n_{lj}(+)},$$

where $n_{lj}(k)$ is the number of alleles of type k observed at locus j in database individuals of subpopulation $\mathcal{P}_l$. This gives the match probability estimate

$$\hat{m}_l^e = c \prod_{j=1}^{M} \hat{G}_{lj}(y_{j1}) \hat{G}_{lj}(y_{j2}). \tag{53}$$

Such an estimate takes no account of any prior information available (stage (iii) of the hierarchical model). This omission could be of particular significance if there is a large amount of information in the prior or if data in one or more subpopulations is not extensive.

First, the prior provides knowledge about each subpopulation's distribution to be updated by the data. The prior parameters assumed known are $(\boldsymbol{a}_{\gamma j})$ defining the density of the ancestral population allele probabilities (γ_j) at each locus j, and (a_θ, b_θ), the parameters of the beta prior placed upon the subpopulation differentiation parameter (θ_j). Thus the prior provides an underlying mean about which the subpopulation allele probabilities are distributed, and also information upon the degree of variation demonstrated by these probabilities across subpopulation.

Second, the prior provides information upon the likely variance of the subpopulation distributions. This will be updated by the data and, for example, if the variance is small, we can say that distributions in subpopulations with little data are likely to be close to those with extensive data, meaning that we can, in some sense, use the data of other subpopulations to update the estimated allele probabilities of a particular subpopulation. This effect cannot be utilised if the empirical estimators are employed, as they only provide a direct estimate of each set of subpopulation allele probabilities.

MCMC methods provide a chain of values which can be treated as a sample from the posterior distribution, thus making use of both prior and empirical information. The posterior distribution of interest in the case of complete subpopulation information is

$$\pi(\boldsymbol{a}, \boldsymbol{G} | \chi_\alpha = \xi_\alpha).$$

Assuming that we have generated an appropriate chain of length r with burn-in m, one then has a choice between the two ergodic averages,

$$\hat{m}_l^{(1)} = \frac{1}{r-m} \sum_{t=m+1}^{r} c \prod_{j=1}^{M} G_{lj}^{(t)}(y_{j1}) G_{lj}^{(t)}(y_{j2}) \tag{54}$$

under model 1, and

$$\hat{m}_l^{(2)} = \frac{1}{r-m} \sum_{t=m+1}^{r} c \prod_{j=1}^{M} \frac{(a_j^{(t)}(y_{j1}) + n_{lj}(y_{j1}))(a_j^{(t)}(y_{j2}) + n_{lj}(y_{j2}) + \delta_j)}{(a_j^{(t)}(+) + n_{lj}(+))(a_j^{(t)}(+) + n_{lj}(+) + 1)} \tag{55}$$

under model 2, where t labels the iteration of the Markov chain.

If the database is reasonably large, and the profile in question relatively common, the allele frequencies $n_{lj}(y_{jb})$ should be much larger than the prior

parameters $a_j(y_{jb})$ meaning that $\hat{m}_l^{(1)}$ and $\hat{m}_l^{(2)}$ will be close to the empirical estimate. In this case, the huge savings in computer time would advocate the use of the empirical estimator ahead of an MCMC estimator.

However, it is still important to consider which of the two models it is preferable to use when an MCMC estimator is required, not only for the complete information case in which the empirical estimator may not be acceptable, but also for future use in the cases in which an empirical estimator is not available.

The scheme follows a Gibbs sampling format, updating $\boldsymbol{a}_j$ for each locus j, and then updating $\boldsymbol{G}_{lj}$ for each subpopulation $\mathcal{P}_l$ and locus j. Gibbs sampling requires the ability to sample directly from the full conditional (i.e. conditional upon all other quantities) distributions $f(.|...)$ of each of the parameters under consideration. In this case, sampling from the full conditional distribution of $\boldsymbol{G}$ is straightforward as

$$\begin{aligned} f(\boldsymbol{G}|...) &= f(\boldsymbol{G}|\boldsymbol{a}, \chi_\alpha = \xi_\alpha) \\ &\propto \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}, \boldsymbol{a}).\pi(\boldsymbol{G}|\boldsymbol{a}) \\ &\propto \prod_{l=1}^{\eta}\prod_{j=1}^{M}\prod_{k=1}^{r_j} G_{lj}(k)^{n_{lj}(k)}.G_{lj}(k)^{a_j(k)-1} \end{aligned}$$

$$\Rightarrow G_{lj} \sim \text{Dirichlet}(a_j(1) + n_{lj}(1), \ldots, a_j(r_j) + n_{lj}(r_j)),$$

independent across all l and j. Sampling from the Dirichlet distribution is carried out using the algorithm of Devroye [Devroye, 1986].

In previous studies, much attention has been paid to the distribution of the various parameters of subpopulation differentiation. The approaches of Foreman *et al* [Foreman, Evett and Smith, 1997] and Dawid and Pueschel [Dawid and Pueschel, 1999] use an empirical estimate of γ. Although it has been suggested that, in practice, using this empirical estimate of γ rather than including it as a variable in the MCMC scheme makes little difference to the results, it would seem desirable to develop a scheme which does not require the use of this empirical estimate.

As the quantities of real interest to us are the match probabilities, it would

seem reasonable to update $\boldsymbol{a}_j \left(= \frac{1-\theta_j}{\theta_j}\boldsymbol{\gamma}_j\right)$ rather than the two parameters (θ_j, γ_j) separately. This should provide a more efficient sampler.

The allele frequencies $(\boldsymbol{\gamma}_j)$ of the ancestral population are assumed independent of the subpopulation differentiation parameters $(\boldsymbol{\theta}_j)$ at all loci $(j = 1, \ldots, M)$. This is an assumption which could stimulate some debate, but under this model in which $\boldsymbol{\theta}$ is a set of parameters determining the variance of the process generating $\boldsymbol{G}$, it would currently seem an unnecessary complication to build in dependence at this level. It is important that criticism of this assumption should arise from consideration of the role of these parameters in a population genetics context, and not confusion over their role within the hierarchy. In Chapter 5 we defined parameters F and g for the binary allele, single locus case,

$$\begin{aligned} g &= \sum_{l=1}^{\eta} \kappa(l) G_l, \\ F &= \frac{\sigma^2}{g(1-g)}, \end{aligned}$$

where $\sigma^2 = \sum_{l=1}^{\eta} \kappa(l)(G_l - g)^2$. These parameters can also be considered as a subpopulation differentiation parameter (F) and mean allele probability (g), but at a lower level of the hierarchy. They are dependent given a, but it should be clear that these parameters are utterly distinct from $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$.

The prior independence of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ is used in the derivation of the full conditional distribution of $\boldsymbol{a}$ (Appendix B):

$$\begin{aligned} f(\boldsymbol{a}|\ldots) &= f(\boldsymbol{a}|(\boldsymbol{a}_{\gamma j}, a_{\theta j}, b_{\theta j}, \boldsymbol{G}) \\ &\propto \prod_{j=1}^{M} \left(\prod_{k=1}^{r_j} \left(\frac{a_j(k)}{a_j(+)} \right)^{a_{\gamma j}(k)-1} \right) \left(\frac{1}{a_j(+)+1} \right)^{a_{\theta j}+r_j-4} \left(\frac{a_j(+)}{a_j(+)+1} \right)^{b_{\theta j}-r_j} \\ &\quad \times \prod_{l=1}^{\eta} \left(\frac{\Gamma(a_j(+))}{\prod_{k=1}^{r_j} \Gamma(a_j(k))} \right) \left[\prod_{k=1}^{r_j} G_{lj}(k)^{a_j(k)-1} \right]. \end{aligned}$$

Unlike the case for $\boldsymbol{G}$, it is not straightforward to sample from the full conditional distribution of $\boldsymbol{a}$. There are a number of methods designed to sample

from a density known only up to the constant of proportionality. These include rejection sampling [Ripley, 1987], the ratio-of-uniforms method [Ripley, 1987] and adaptive rejection sampling [Gliks, 1992, Gilks and Wild, 1992].

In this case, the Gibbs sampling set-up is adjusted to a hybrid MCMC strategy. As $\boldsymbol{a}$ cannot be sampled directly, within each iteration of the scheme a number h of Metropolis-Hastings steps is used. At each of these steps, a value is sampled from a proposal distribution, and accepted or rejected according to an acceptance probability α_{acc}. This probability is dependent only upon the ratio of the conditional densities of the proposed and current values of $\boldsymbol{a}$, and therefore does not require knowledge of constants of proportionality.

The number h should be chosen to give a high proportion of Gibbs sampler iterations in which there is at least one acceptance between the first Metropolis-Hastings step and the last. The closer the proposal distribution is to the target distribution, i.e. the full conditional of $\boldsymbol{a}$, the more efficient the scheme. This translates to fewer Metropolis-Hastings steps being required within each iteration of the overall scheme.

A simple proposal distribution for $\boldsymbol{a}$ would generate (θ_j) and $(\boldsymbol{\gamma}_j)$ independently,

$$\begin{aligned} \theta_j &\sim \text{Beta}(k_{\theta 1}, k_{\theta 2}), \\ \boldsymbol{\gamma}_j &\sim \text{Dirichlet}(k_{\gamma j 1}, \ldots, k_{\gamma j r_j}), \end{aligned}$$

and use the equation, $\boldsymbol{a}_j = \frac{1-\theta_j}{\theta_j}\boldsymbol{\gamma}_j$, to calculate $\boldsymbol{a}_j$.

Such a proposal distribution is acceptable for $\boldsymbol{\theta}$, but the relatively high dimension of the parameter space of $\boldsymbol{\gamma}$ makes it very difficult to select $(\boldsymbol{k}_{\gamma j})$ to give a reasonable acceptance rate. Indeed choosing a suitable proposal distribution subject to the constraint, $\sum_{k=1}^{r_j} \gamma_j(k) = 1$, is not straightforward. To allow a wider choice of potential distributions, it was decided to work with $\log(\boldsymbol{a}_j)$, the elements of which can have values in the range $(-\infty, \infty)$.

One option now available is the multivariate normal distribution, i.e.

$$\boldsymbol{Y}_j \sim \text{Normal}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

where $\boldsymbol{Y}_j = \log(\boldsymbol{a}_j)$.

Considering the prior distributions placed upon (θ_j) and $(\boldsymbol{\gamma}_j)$ (equations (49) and (50)), a reasonable choice of the proposal mean would be

$$\boldsymbol{\mu}_j = \log\left(k_4.\left[\frac{1-\frac{a_{\theta j}}{a_{\theta j}+b_{\theta j}}}{\frac{a_{\theta j}}{a_{\theta j}+b_{\theta j}}}\right].\boldsymbol{\pi}_{\gamma j}\right)$$

where $\boldsymbol{\pi}_{\gamma j} = \frac{1}{a_j(+)}.\boldsymbol{a}_j$, and $\frac{a_{\theta j}}{a_{\theta j}+b_{\theta j}}$ is the prior expectation of θ_j. Such a proposal distribution would make use of available prior information, with a constant k_4 allowing tuning of the variance to give the desired acceptance rate.

In general, however, the MCMC scheme will be more efficient if it makes use of the information provided as the chain increases in length. A random walk sampler [Roberts, 1995] does this by shifting the mean of the proposal distribution to the value of the previous iteration.

In this case the suggested distribution of the value $\boldsymbol{a}_j^{(p)}$ proposed for $\boldsymbol{a}_j^{(t+1)}$ given $\boldsymbol{a}_j^{(t)}$ is defined by

$$\log(\boldsymbol{a}_j^{(p)}|\boldsymbol{a}_j^{(t)}) \sim \text{Normal}(\log(\boldsymbol{a}_j^{(t)}), \Sigma_j).$$

The variance-covariance matrix Σ_j can be user defined to achieve a desirable acceptance rate, generally around 40%.

The full density for the proposed value $\boldsymbol{a}_j^{(p)}$ given $\boldsymbol{a}_j^{(t)}$ is then given by

$$\begin{aligned} q(\boldsymbol{a}_j^{(p)}|\boldsymbol{a}_j^{(t)}) &= \frac{1}{J(\boldsymbol{a}_j^{(p)}, \boldsymbol{y}_j^{(p)})} f(\boldsymbol{y}_j^{(p)}|\boldsymbol{a}_j^{(t)}) \\ &= \frac{\exp(-\frac{1}{2}\left\{\log(\boldsymbol{a}_j^{(p)}) - \log(\boldsymbol{a}_j^{(t)}))'\Sigma^{-1}(\log(\boldsymbol{a}_j^{(p)}) - \log(\boldsymbol{a}_j^{(t)}))\right\}}{[\prod_{k=1}^{r_j} a_j^{(p)}(k)](2\pi)^{\frac{r_j}{2}}|\Sigma_j|^{\frac{1}{2}}}, \end{aligned}$$

where $\boldsymbol{y}_j = \log(\boldsymbol{a}_j)$ and $J(\boldsymbol{x}, \boldsymbol{y})$ denotes the Jacobian of $\boldsymbol{x}$ with respect to $\boldsymbol{y}$.

The acceptance probability for the vector proposed at locus j for the $(t+1)^{th}$ iteration is then given by

$$\alpha_{acc} = \min\left(1, \frac{\pi(\boldsymbol{a}_j^{(p)}|\ldots).q(\boldsymbol{a}_j^{(t)}|\boldsymbol{a}_j^{(p)})}{\pi(\boldsymbol{a}_j^{(t)}|\ldots).q(\boldsymbol{a}_j^{(p)}|\boldsymbol{a}_j^{(t)})}\right).$$

Using this proposal distribution within the hybrid MCMC structure allows the MCMC scheme for the case of complete subpopulation information to proceed as described in Figure 7.

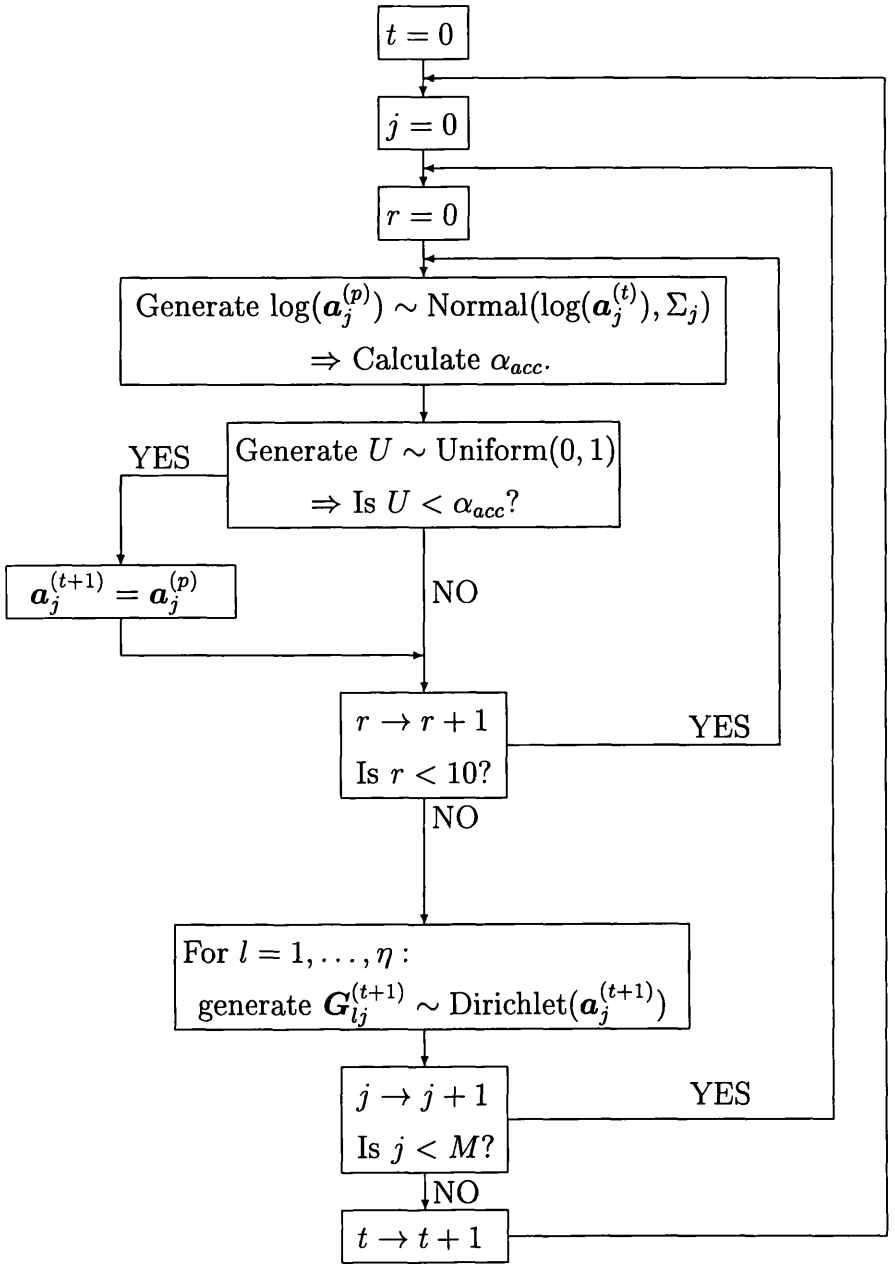


Figure 7: Flow diagram representing MCMC scheme: individual subpopulation labels known. The prior distributions are used to generate values at iteration $t = 0$. The label r tracks the Metropolis-Hastings steps used to generate $\boldsymbol{a}^{(t)}$ within the Gibbs sampling format.

Whichever statistical model is used, the same scheme is applied. The MCMC scheme employed follows a Gibbs sampling format, our target distribution being the joint posterior

$$\pi(\boldsymbol{a}, \boldsymbol{G}|\chi_\alpha = \xi_\alpha).$$

of all variables.

The Markov chain will contain, at each iteration t, values for each of the parameters in the above distribution,

$$(a_1(1)^{(t)}, \ldots, a_j(k)^{(t)}, \ldots, a_M(m_M)^{(t)},$$
$$G_{11}(1)^{(t)}, \ldots, G_{lj}(k)^{(t)}, \ldots, G_{\eta M}(m_M)^{(t)}).$$

Assuming the Ergodic Theorem (see Section 6.2) to hold, the chain of values for each parameter can be treated as a sample from its marginal posterior distribution, and thus a 'sample' mean, $\hat{m}_l^{(1)}$ (54) or $\hat{m}_l^{(2)}$ (55), of an appropriate function of these parameters used to estimate the match probabilities.

The simpler of the two estimators is the ergodic average $\hat{m}_l^{(1)}$ of a product of the subpopulation frequencies across the alleles of the profile $\boldsymbol{y}$. The calculation of this function at each iteration of the chain should afford a small saving in computer time over the more complicated estimator $\hat{m}_l^{(2)}$. However it is assumed that the chain length required is not great enough to cause such a small saving per iteration to be significant overall. One must therefore decide which, if either, of the two estimators is more accurate.

The variances of the two estimators are compared, working under the simplifying assumption that the chain of values represents a random sample from the joint posterior distribution.

$$\begin{aligned} \mathrm{Var}(m_l^{(1)}|\omega) &= \mathrm{Var}(\mathrm{E}[m_l^{(1)}|\boldsymbol{a}, \omega]|\omega) + \mathrm{E}[\mathrm{Var}(m_l^{(1)}|\boldsymbol{a}, \omega)|\omega] \\ &= \mathrm{Var}(m_l^{(2)}|\omega) + \mathrm{E}[\mathrm{Var}(m_l^{(2)}|\boldsymbol{a}, \omega|\omega]. \end{aligned}$$

where $\omega = (i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha)$.

This reveals that $\hat{m}_l^{(2)}$ is a Rao-Blackwell estimator of the match probability m_l, and has a posterior variance which can be no larger than that of $\hat{m}_l^{(1)}$, i.e.

$$\mathrm{Var}(\hat{m}_l^{(2)}|i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha) \leq \mathrm{Var}(\hat{m}_l^{(1)}|i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha).$$

For this reason, it is suggested that when an MCMC estimator is required, that derived under model II is used.

7.3 Subpopulation labels unknown

When subpopulation labels are not available, the vector $\boldsymbol{I}$ containing these labels for the database individuals must be introduced to the MCMC scheme.

With no consistent empirical estimator available, the MCMC estimator $\hat{m}_l^{(2)}$ is used. This expression does not involve $\boldsymbol{I}^{(t)}$ explicitly, but the subpopulation identifiers are required at each iteration to evaluate $\boldsymbol{n}^{(t)}$. Similarly, the full conditional distributions of $\boldsymbol{G}$ and $\boldsymbol{a}$ now involve $\boldsymbol{I}$ indirectly.

The full conditional density of $\boldsymbol{I}$ is given by

$$\begin{aligned} f(\boldsymbol{I}|\boldsymbol{a},\boldsymbol{G},\chi_\alpha=\xi_\alpha) &\propto \Pr(\chi_\alpha=\xi_\alpha|\boldsymbol{G},\boldsymbol{a},\boldsymbol{I}).\pi(\boldsymbol{I}|\boldsymbol{a},\boldsymbol{G}) \\ &\propto \prod_{i\in n}\left[\prod_{j=1}^{M} G_{I_i j}(x_{ij1})G_{I_i j}(x_{ij2})\right]\kappa(I_i) \end{aligned}$$

To this point, the subpopulation proportions $\boldsymbol{\kappa}$ have been assumed known. Again, this is not necessarily so, a consideration discussed in more detail in Chapter 8.

If $\boldsymbol{\kappa}$ is unknown it must also be introduced as a random variable together with a prior distribution. Following the suggestion of Foreman *et al.*, we use the Dirichlet distribution,

$$\boldsymbol{\kappa} \sim \text{Dirichlet}(\pi_\kappa(1),\ldots,\pi_\kappa(\eta))$$

This completes the DAG for incomplete information shown in Figure 8.

The full conditional distribution of $\boldsymbol{\kappa}$ is also Dirichlet, making it straightforward to add to the Gibbs sampling structure of the MCMC scheme,

$$\begin{aligned} \pi(\boldsymbol{\kappa}|...) &\propto \Pr(\boldsymbol{I}|\boldsymbol{\kappa}).\pi(\boldsymbol{\kappa}) \\ &\propto \left[\prod_{i=1}^{n_a}\kappa(I_i)\right]\prod_{l=1}^{\eta}\kappa(l)^{\pi_\kappa(l)-1} \\ &\propto \prod_{l=1}^{\eta}\kappa(l)^{n_a(l)+\pi_\kappa(l)-1}. \\ &\Rightarrow \boldsymbol{\kappa} \sim \text{Dirichlet}(\boldsymbol{n}_a+\boldsymbol{\pi}_\kappa), \end{aligned}$$

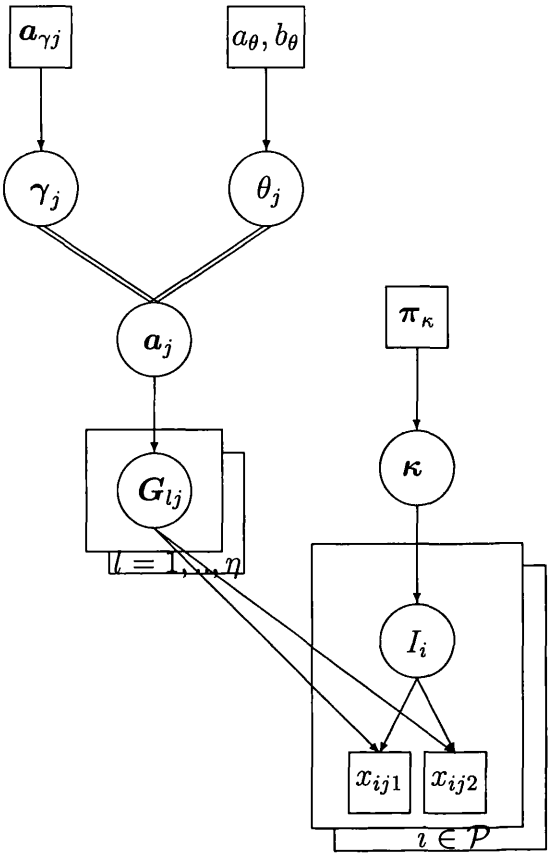


Figure 8: DAG for the case of incomplete information.

where $(n_a(l); l = 1, \ldots, \eta)$ represents the number of database individuals in subpopulation $\mathcal{P}_l$.

When we actually apply the MCMC scheme, we encounter problems, in particularly a lack of mixing across a multimodal posterior. Such practical difficulties and proposed solutions are considered in more detail in Chapter 8.

8 The definition of a subpopulation, and its effects

8.1 Introduction

In this chapter, we consider the effect on the analysis of the way in which the subpopulations are defined. This definition has an effect on both the amount of knowledge we can assume about the subpopulation membership (I_i) of individuals (i), the subpopulation proportions ($\kappa(l); l = 1, \ldots, \eta$) and the level of prior knowledge we have regarding the relationship between suspect and offender.

The definition of "subpopulation" can be considered in two ways:

(i) the subpopulations are tools of the model, used to reflect to some extent the relationships between individuals within the population. In reality, the closeness of the relationship between two distinct individuals can vary from that of siblings, to more distant family, to members of the same racial group and beyond. To build a model involving the possibility of two randomly selected individuals being brothers, cousins, second cousins and so on towards "unrelated" is impractical. The randomly mating subpopulations are incorporated into the model to represent to some extent the fact that some individuals are more closely related than others while ensuring that the model can be employed in practice.

In this instance the subpopulation membership of each individual is generally unknown and further, *a priori* one should consider the subpopulations to fall into equal proportions. One could consider the subpopulations themselves to be subdivided. Indeed the stratification could continue until, in the extreme case, we have subpopulations each consisting of a single individual. However there appears to be little profit in subdividing to this extent.

We consider a single level of stratification with the number of subpopulations within this level assumed known. Relaxation of this assumption is

considered in Chapter 10.

(ii) The subpopulations are clearly defined in terms of ancestries which can be traced. We could for example split the Caucasian population into subpopulations defined by European country of origin. The model is still a clear simplification, as mating between descendents of European countries is known to occur whilst our model would not permit this.

Under this definition, observation of an individual is likely to provide information on his subpopulation membership. It is in this case that we are likely to have information available upon the subpopulation proportions $\{\alpha_l : l = 1, \ldots, \eta\}$. In the extreme case, the subpopulation membership of each database individual is known absolutely. While this is unlikely to occur in practice, the study of such a situation has allowed the development of of much of the theory extended in this thesis. Dawid and Pueschel [Dawid and Pueschel, 1999] employ this assumption when displaying the necessity of conditioning upon the database throughout match probability calculation. In this thesis it is shown how this conditioning is still required when the assumption of known subpopulation membership is relaxed.

The way in which the subpopulations are defined affects what we can reasonably assume about

(i) individual subpopulation labels;

(ii) subpopulation proportions.

In this chapter, we consider the range of information and knowledge which may be at our disposal in the above areas in the context of the possible subpopulation definitions.

The following section introduces the data and analyses used to compare the results of match probability calculations under the various degrees of assumed knowledge.

In Section 8.3 we consider the simplest case in which subpopulation labels (I_i) and proportions $(\kappa(l))$ are assumed known. Comparisons are made between

empirical and MCMC match probability estimates. In Section 8.4 we consider adjustments to the analysis required when subpopulation labels are no longer known, corresponding to the more realistic situation (i) in which the subpopulations are tools of the model. In this instance, it is likely that we must also relax the assumption that the subpopulation proportions ($\kappa(l)$) are known, meaning that these must also be estimated from the data.

We recall (Chapter 2) that match probabilities (m_l) for each subpopulation ($\mathcal{P}_l, l = 1, \ldots, \eta$) are required, where

$$m_l = \Pr(\boldsymbol{X}_i = \boldsymbol{y} | \chi_\alpha = \xi_\alpha)$$

for an individual i (outside the database α) in subpopulation $\mathcal{P}_l$. The collection of individual profiles χ_α includes the suspect's profile $\boldsymbol{X}_s$ which is known to be $\boldsymbol{y}$.

When calculating this probability we employ the hierarchical model defined by the following levels:

(i) inheritance of a particular profile $\boldsymbol{x}_i$ by each individual i within a subpopulation ($\mathcal{P}_l, l = 1, \ldots, \eta$),

$$\Pr(X_{ijb} = k | \boldsymbol{G}) = G_{lj}(k),$$

independently across i, l, band ($b = 1, 2$) and locus ($j = 1, \ldots, M$), the collection of observable alleles at a particular locus being denoted by $k = 1, \ldots, r_j$.

(ii) Generation of the allele probabilities $\boldsymbol{G}$ in each subpopulation,

$$G_{lj} \sim \text{Dirichlet}(a_j(1), a_j(2), \ldots, a_j(r_j)), \text{ independently for all } l, j, \quad (56)$$

where $\boldsymbol{a}_j = \frac{1-\theta_j}{\theta_j}\boldsymbol{\gamma}_j$.

(iii) The generation of the ancestral population parameters ($\boldsymbol{\gamma}_j, \boldsymbol{\theta}_j$) from a 'hyperprior' distribution,

$$\begin{aligned} \boldsymbol{\gamma}_j &\sim \text{Dirichlet}(a_\gamma(1), \ldots, a_\gamma(r_j)); \\ \theta_j &\sim \text{Beta}(a_\theta, b_\theta). \end{aligned}$$

In this thesis we assume that we have no prior knowledge informing us that some alleles have a greater relative frequency than others, and so we take $a_\gamma(k) = 1$ for all k at each locus. The hyperprior parameters for the generation of θ_j are chosen to be $(a_\theta = 1.5, b_\theta = 50)$. These values are similar to those used by Foreman *et al* and, as it is thought that values of θ_j are generally considerably less than the resultant prior mean of 0.291, this prior distribution is considered conservative. This is because a large value of θ_j implies a greater degree of population heterogeneity which will usually decrease the weight of evidence of the profile match.

8.2 Analysis

Estimated match probabilities quoted in this chapter result from analyses carried out using a database consisting of short tandem repeat (Appendix A) profiles typed at up to four loci in 1401 Caucasian individuals, and 558 Afro-Caribbeans. Profiles that had data missing (two of the Caucasian profiles, and 26 Afro-Caribbean profiles) were retained, as the limited information provided is still of use.

It is important to note that the methods of this thesis are only valid when it is legitimate to consider the database as a 'random' sample from the appropriate population. If subpopulation proportions are assumed known, our analysis assumes that these give the prior subopulation membership probabilities of each database individual. In the more general case we use the database to provide information about the subpopulation proportions. It is clearly important that in doing this we are making inference upon the population of interest.

In the following sections, match probabilities are estimated for three different suspect profiles. These profiles are labelled Ca_c, Ca_r, and AC_c and are presented in Table 3. MCMC chains were run under the various degrees of assumed knowledge described, and the appropriate estimators calculated. When subpopulation labels are assumed known, these estimators are compared to the empirical estimators.

	vwa	tho1	f13a1	fes
Ca_c	16, 18	7, 7	5, 7	10, 11
Ca_r	14, 20	5, 8	3, 4	13, 13
AC_c	15, 15	7, 8	3, 5	8, 9

Table 3: Suspect profiles used for analysis.

In later sections, it is assumed that it is no longer known which individuals are Afro-Caribbean, and which Caucasian. While there is greater heterogeneity in this collection of two racial groups than would be expected within a single racial group, analysis of these two 'subpopulations' allows us to witness the extent to which the resultant clusters match the 'true' clusters. Match probability estimates were calculated, having initiated the Markov chain with a series of different random seeds. The grouping observed in these estimates when subpopulation labels are unknown reveals a multimodality in the posterior distribution of the parameters. Estimation problems caused by this multimodality are tackled using the methods of simulated tempering and importance sampling described in Chapter 6. The resultant match probability estimates are then compared to those made when assuming subpopulation information known.

Ca_c and Ca_r are the profiles used in the analysis of Foreman *et al.* These profiles can be seen (Appendix C) to be relatively common and rare respectively within the Caucasian population, while AC_c is a common Afro-Caribbean profile. Whether using the empirical estimators, or conditioning upon the data when running MCMC schemes and calculating ergodic averages, the appropriate profile is added to the database. When there are very few alleles of a suspect profile observed in the original database, it is particularly important that it is included in the database as its addition can have a significant effect upon the empirical relative frequencies of its alleles. As a general principle, it is important to add the suspect's profile to the database anyway to ensure the correct conditioning is executed throughout.

The programs executing all MCMC schemes described were written in the C programming language.

8.2.1 Assumptions regarding subpopulation membership of culprit C and suspect s

The overall match probability $\sum_{l=1}^{\eta} \lambda_l m_l$ requires knowledge of the probabilities $\lambda_l = \Pr(C \in P_l | C \notin \alpha, \varepsilon)$ as well as the match probability m_l for each subpopulation P_l. Here α refers to the database including the suspect s and ε denotes other, 'non-DNA', evidence.

When quoting overall match probabilities for the case of known subpopulation proportions $\boldsymbol{\kappa}$ it is assumed that $\lambda_l = \kappa(l)$ for all l. This might not be appropriate if the subpopulations are defined by known physical characteristics. It is possible that the non-DNA evidence contains eye-witness information which increases the probability of the culprit being in a particular subpopulation, meaning that λ_l is not necessarily equal to $\kappa(l)$ for all l.

Throughout the analyses in this thesis, the suspect is treated similarly to the other members of the database α, in that it is assumed that his subpopulation, when unknown, has a prior distribution defined by the vector of proportions $\boldsymbol{\kappa}$. If relevant eye-witness evidence is available, it would seem reasonable to think that the search resulting in the arrest of the suspect concentrated on members of the population displaying certain characteristics. While this makes it likely that the suspect and culprit originate from the same subpopulation, dependence between the respective subpopulation identifiers I_s and I_C is not required in the model. Any dependence is only apparent before the search for the suspect. Once the suspect is identified, a prior can be placed upon his subpopulation, conditional upon his physical characteristics. Under this conditioning, the suspect's subpopulation is independent of the culprit's. This prior probability upon the suspect's subpopulation should be built into the MCMC scheme, but does not affect (λ_l).

The above adjustments do not significantly complicate the analysis or affect

the general principles employed throughout this thesis.

We later (Section 8.4.2) relax the assumption that the subpopulation proportions ($\kappa(l)$) are known. If there is no physical interpretation to be placed upon the subpopulations, there can be no eye witness evidence regarding the subpopulation of the culprit, and we must therefore assume that $\lambda_l = \kappa(l)$ for all l.

8.3 Individual subpopulation labels known

If the subpopulations are defined in such a way as to allow the membership of individuals to be easily recognized, it is possible that the database includes this information upon the observed individuals, as well as their profiles. This is the situation considered by Dawid and Pueschel [Dawid and Pueschel, 1999].

In this case empirical estimates of the subpopulation allele probabilities ($\boldsymbol{G}_l$) are available. These are given by

$$\hat{G}_{lj}(k) = \frac{n_{lj}(k)}{n_{lj}(+)},$$

where $n_{lj}(k)$ is the number of alleles of type k at locus j, observed in individuals of subpopulation $\mathcal{P}_l$, within the database α. With an extensive database, these subpopulation frequency estimates will be consistent and can be multiplied across the alleles of the suspect profile $\boldsymbol{x}$ to give a match probability estimate

$$\hat{m}_l^e = \prod_{j=1}^{M} c_j \hat{G}_{lj}(y_{j1}) \hat{G}_{lj}(y_{j2}), \tag{57}$$

where $c_j = 2^{h(y_{j1}, y_{j2})}$, $h(r,s)$ indicating if r and s are unequal.

If there is a large amount of prior information available regarding the ancestral population parameters $(\boldsymbol{\gamma}_j, \theta_j)$ or the data is not extensive, these empirical estimates may not be satisfactory (as discussed in Section 7.2). In this instance, the MCMC estimators described in Chapter 7 can be used. The Markov chain generated is designed to have a stationary distribution which matches the target posterior distribution $\pi(\boldsymbol{a}, \boldsymbol{G} | \chi_\alpha = \xi_\alpha)$. Each estimator is an ergodic average along this chain of a function of the parameters defined by the statistical model

	empirical	model I	model II
m_{Ca}	2.94×10^{-5}	2.97×10^{-5}	2.97×10^{-5}
m_{AC}	1.87×10^{-4}	1.87×10^{-4}	1.87×10^{-4}
m	7.43×10^{-5}	7.44×10^{-5}	7.44×10^{-5}

Table 4: Posterior match probabilities for profile Ca_c.

(see Chapter 4) used,

$$\hat{m}_l^{(1)} = \frac{1}{r-m} \sum_{t=m+1}^{r} c \prod_{j=1}^{M} G_{lj}^{(t)}(y_{j1}) G_{lj}^{(t)}(y_{j2}) \tag{58}$$

under model I, where $c = \prod_j c_j$ and t labels the iteration of the Markov chain. Under model II,

$$\hat{m}_l^{(2)} = \frac{1}{r-m} \sum_{t=m+1}^{r} c \prod_{j=1}^{M} \frac{(a_j^{(t)}(y_{j1}) + n_{lj}(y_{j1}))(a_j^{(t)}(y_{j2}) + n_{lj}(y_{j2}) + \delta_j)}{(a_j^{(t)}(+) + n_{lj}(+))(a_j^{(t)}(+) + n_{lj}(+) + 1)} \tag{59}$$

where δ_j indicates if y_{j1} and y_{j2} are equal.

To this point, the subpopulation proportions ($\kappa(l)$) have been assumed known and taken as the empirical proportions in the database, excluding the suspect. In this case, $\kappa(1) = \frac{1401}{1959} = 0.715, \kappa(2) = 0.285$, where the labels 1 and 2 refer to the Caucasian and Afro-Caribbean 'subpopulations' respectively.

Tables 4, 5 and 6 show empirical and MCMC estimates of subpopulation match probabilities (m_1, m_2) and the overall match probability $(m = \kappa(1)m_1 + \kappa(2)m_2)$, for the profiles Ca_c, Ca_r and AC_c respectively.

The MCMC scheme described in Section 7.2 was employed and an analysis of the resultant chains using CODA [BUGS] suggests that the run lengths of 10000 iterations with a burn-in of 3000 are satisfactory. Each estimator is a mean of seven ergodic averages along chains initiated with different random seeds.

A comparison of the results under profiles Ca_c and Ca_r demonstrates the effect of having a small amount of data for the alleles of a particular profile. The empirical and MCMC estimates are similar for the relatively common profile Ca_c, meaning that $\hat{m}_l^e$ is satisfactory in this instance. For the rarer profile Ca_r,

	empirical	model I	model II
m_{Ca}	1.47×10^{-11}	1.69×10^{-11}	1.69×10^{-11}
m_{AC}	3.45×10^{-10}	3.85×10^{-10}	3.86×10^{-10}
m	1.10×10^{-10}	1.22×10^{-10}	1.22×10^{-10}

Table 5: Posterior match probabilities for profile Ca_r.

	empirical	model I	model II
m_{Ca}	3.95×10^{-10}	5.27×10^{-10}	5.27×10^{-10}
m_{AC}	8.01×10^{-6}	6.99×10^{-6}	7.00×10^{-6}
m	2.29×10^{-6}	1.99×10^{-6}	2.00×10^{-6}

Table 6: Posterior match probabilities for profile AC_c.

however, an argument for the use of the MCMC estimate is supported by the much greater difference in the estimates.

A similar comparison for AC_c demonstrates that even in the case of a relatively common overall profile, it may be necessary to use an MCMC estimator. In this case, the profile consists of three common pairs, and a relatively rare pair at locus fes. It is the discrepancy between empirical and MCMC match probability estimates at this locus that is mainly responsible for the difference in overall profile estimates. These individual locus match probabilities can be seen in Table 7.

A comparison of the match probability estimates under models I and II reveals that they are similar. This suggests that, although the estimator under model II is theoretically more accurate, the choice of estimator does not have a significant effect on the results in practice.

	Ca: empirical	Ca: model II	AC: empirical	AC: model II
vwa	8.49×10^{-3}	8.85×10^{-3}	4.82×10^{-2}	4.58×10^{-2}
$tho1$	4.12×10^{-2}	4.17×10^{-2}	1.71×10^{-1}	1.68×10^{-1}
$f13a1$	2.40×10^{-2}	2.41×10^{-2}	8.03×10^{-2}	7.90×10^{-2}
fes	4.73×10^{-5}	5.92×10^{-5}	1.21×10^{-2}	1.12×10^{-2}

Table 7: Individual locus posterior match probabilities for profile AC_c.

8.4 Labels unknown

This is clearly a somewhat more difficult case to handle than that of known subpopulation labels. The clusters are no longer pre-determined, meaning that our population substructure model must be used to cluster individuals according to their profile.

If $\boldsymbol{G}$ is known, then the subpopulation of the individual must also be known to retain independence across loci. If the subpopulation membership of a particular individual is unknown, an allele observed at a particular locus suggests membership of a subpopulation for which the corresponding frequency is relatively high. This gives additional information about the possible allele frequencies at other loci meaning that the alleles at different loci are dependent.

Such dependencies arise from the substructure present within the population. When the assumption of known subpopulation membership of all individuals is relaxed, it is these dependencies which allow us still to cluster the individuals and use the methods of inference described in Chapter 4 for the complete information case.

The subpopulation labels $(I_i; i = 1, \ldots, n)$ must now be included as an unknown in the MCMC scheme, and be updated at each iteration for each individual. This involves the generation, for each database individual, of a uniform random variable and its comparison to the full conditional probability of that individual's subpopulation label,

$$\Pr(I_i|\boldsymbol{G}, \chi_\alpha = \xi_\alpha, \boldsymbol{\kappa}) = \frac{[\prod_{j=1}^{M} G_{I_i j}(x_{ij1}) G_{I_i j}(x_{ij2})]\kappa(I_i)}{\sum_{l=1}^{\eta}[\prod_{j=1}^{M} G_{lj}(x_{ij1}) G_{lj}(x_{ij2})]\kappa(l)},$$

independently across individuals, where $\boldsymbol{x}_i$ represents the profile of the i^{th} individual.

The integration required to calculate the match probabilities is not possible using analytical methods. As the empirical estimators ($\hat{m}_l^e$) are no longer available, the only choice of estimator is between MCMC estimators taking ergodic averages of functions of $\boldsymbol{G}$ and $(\boldsymbol{a}, \boldsymbol{n})$ respectively. It is shown in Chapter 7 that the theoretically more accurate of these is

$$\hat{m}_l^{(2)} = \frac{1}{r-m} \sum_{t=m+1}^{r} c \prod_{j=1}^{r_j} \frac{(a_j^{(t)}(y_{j1}) + n_{lj}^{(t)}(y_{j1}))(a_j^{(t)}(y_{j2}) + n_{lj}^{(t)}(y_{j2}) + \delta_j)}{(a_j^{(t)}(+) + n_{lj}^{(t)}(+))(a_j^{(t)}(+) + n_{lj}^{(t)}(+) + 1)},$$

where $(a_j(1), \ldots, a_j(r_j))$ is the function $\left(\frac{1-\theta_j}{\theta_j}(\gamma_j(1), \ldots, \gamma_j(r_j))\right)$ of the relative allele frequencies $\boldsymbol{\gamma}_j$ and subpopulation differentiation parameter θ_j at locus j in the 'ancestral' population. This is the estimator used for the remainder of this chapter. The number $n_{lj}^{(t)}(k)$ of alleles of type k at locus j within individuals of subpopulation $\mathcal{P}_l$ in the database at iteration t is evaluated by counting the alleles within the profiles of individuals whose subpopulation label $I_i^{(t)} = l$.

If the crime profile $\boldsymbol{x}$ is relatively common, and the database reasonably extensive, the allele frequencies $n_{lj}(y_{jb})$ will generally be much larger than $a_j(y_{jb})$, highlighting the need for accurate clustering of individuals if there is a clear substructure within the overall population.

If the subpopulations do correspond to a split of the population with a physical representation, some information on the subpopulation proportions could be available as a result of previous studies. If such a previous study is extensive enough, it may be adequate to consider $\boldsymbol{\kappa}$ 'known' to be the resultant estimates. If this is not a reasonable assumption, $\boldsymbol{\kappa}$ must be introduced as a random variable, and an appropriate prior placed upon it. If there is no physical interpretation to be placed upon the subpopulations, it is assumed that such information upon $\boldsymbol{\kappa}$ would not be available. In this instance, a vague symmetric prior is used.

Analyses assuming both known and unknown $\boldsymbol{\kappa}$ are presented in the following sections.

8.4.1 Subpopulation proportions assumed known

It is assumed that $\boldsymbol{\kappa}$ agrees with the proportions observed in the Caucasian and Afro-Caribbean populations respectively, i.e. $\kappa(1) = \frac{1401}{1959} = 0.715, \kappa(2) = 0.285$.

The MCMC scheme with stationary distribution $\pi(\boldsymbol{a}, \boldsymbol{G}, \boldsymbol{I}|\chi_\alpha = \xi_\alpha)$ was repeated 7 times for 10000 iterations (including a burn-in of 3000). Initiated using different random seeds, the chains produced the match probability estimates of Table 8 for the profile AC_c.

random seed	m_{Ca}	m_{AC}	m
12	6.13×10^{-10}	6.35×10^{-6}	1.81×10^{-6}
12 000	5.93×10^{-10}	6.43×10^{-6}	1.83×10^{-6}
10^{-6}	5.79×10^{-10}	6.43×10^{-6}	1.83×10^{-6}
Average:	5.95×10^{-10}	6.40×10^{-6}	1.82×10^{-6}
1	3.15×10^{-7}	6.20×10^{-12}	2.26×10^{-7}
125 000	3.23×10^{-7}	1.51×10^{-11}	2.31×10^{-7}
560 000	3.21×10^{-7}	6.80×10^{-12}	2.30×10^{-7}
36×10^6	3.15×10^{-7}	9.26×10^{-12}	2.25×10^{-7}
Average:	3.19×10^{-7}	5.94×10^{-12}	2.28×10^{-7}

Table 8: Posterior match probability estimates when $\boldsymbol{\kappa}$ is known. The table is split to highlight the grouping of results.

As can be seen, the estimated match probabilities are dependent upon the random number seed used in the program. The grouping of the results suggests that the Markov chain is not mixing properly across a multimodal posterior density.

Figure 9 shows the first 500 points of a trace of one particular allele frequency $G_{13}(6)$ for the seven random seeds used. This shows how the chain quickly settles

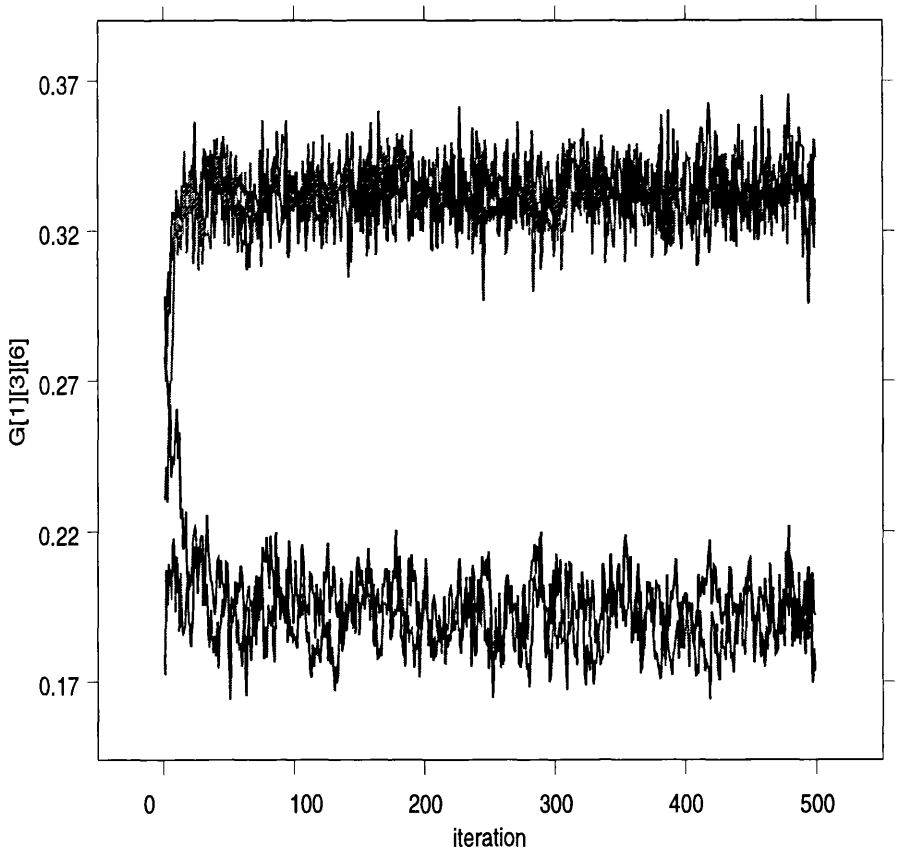


Figure 9: Trace of $G_{13}(6)$ initialised by 7 different random seeds.

into one of the two modes. The evidence of our MCMC results is that there is little, if any, movement between the modes.

Development of an MCMC scheme to cope with such multimodality is aided by a clearer understanding of the reasons behind it. Figures 10 and 11 show the allocation of individuals into subpopulations when the chain remains in each of the two modes.

Figure 10 shows the majority of Caucasians with a very low posterior probability of being in subpopulation 2, and most Afro-Caribbeans having a high probability of being in subpopulation 2, suggesting that this mode corresponds to the 'correct' situation, i.e. recognition of Caucasian and Afro-Caribbean populations. In Figure 11 it appears that the individuals of the minority Afro-Caribbean population are still, in the main, grouped together, but allocated to the larger subpopulation. The diagram of Figure 12 represents a simplified version of the allocation.

A comparison of the average log likelihood of the data across a run of the

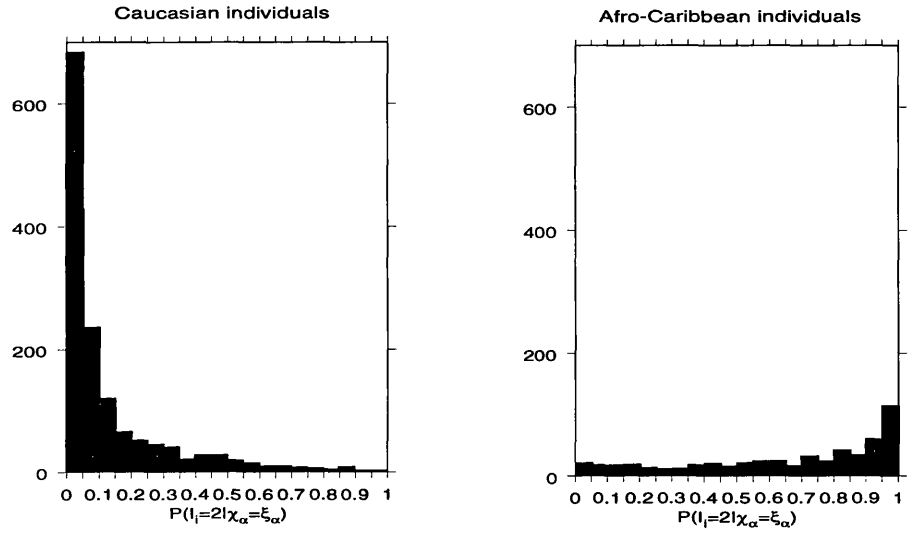


Figure 10: Allocation of individuals: κ known, random seed = 12. For each individual, the posterior probability of him being in a particular subpopulation is estimated by taking the proportion of iterations in which he occupies the particular state. It can be seen that over 850 of the 1401 Caucasian (subpopulation '1') individuals in the database were allocated to the correct subpopulation at least 90% of the time in this instance.

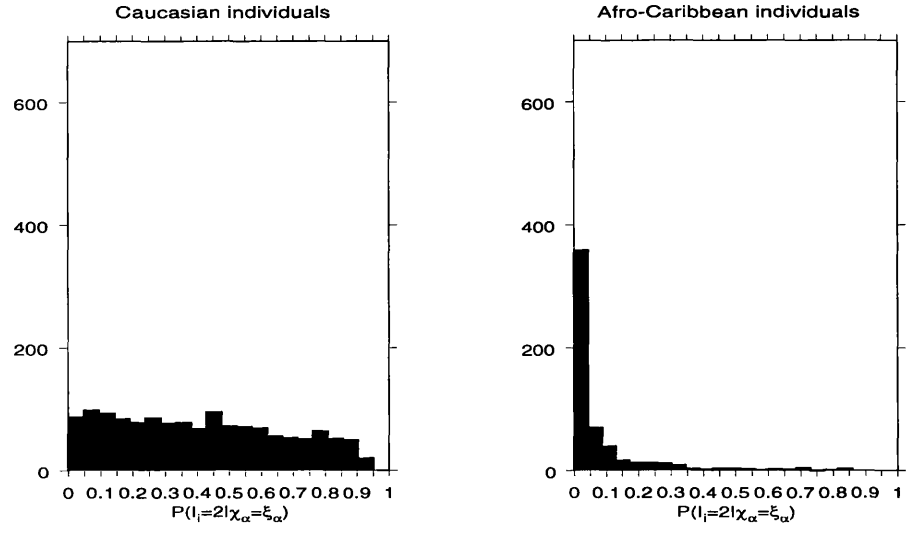


Figure 11: Allocation of individuals: κ known, random seed = 1.

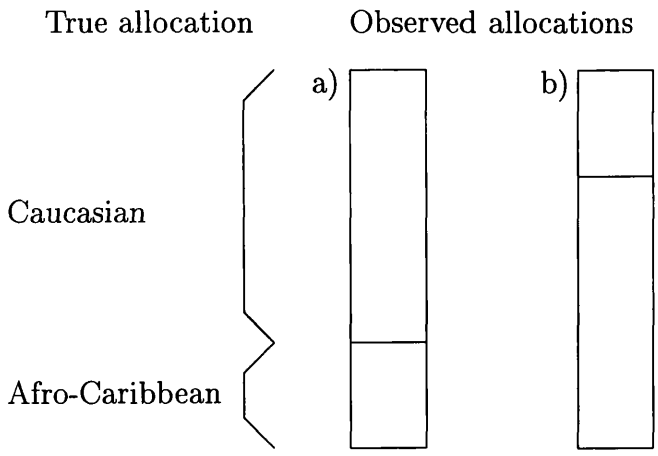


Figure 12: Allocation within the two posterior modes.

MCMC scheme (Figure 13) suggests that the mode corresponding to the incorrect allocation is significantly smaller than that of the 'correct' allocation, as one would generally expect if there is a large difference in the proportions ($\kappa(l)$).

It is certainly reasonable to consider that such a set-up would result in two such modes. All possible vectors $\boldsymbol{I}$ of individual subpopulation allocation have some posterior probability. Those allocations which group together individuals with relatively large numbers of alleles in common will show a larger probability. This means that it is reasonable to think that, in this case in which individuals of two distinct subpopulations are combined to form a single database, those allocations which group together the individuals into the respective subpopulations have the largest posterior probabilities. An allocation in which most Afro-Caribbeans are grouped together in the larger subpopulation will show a smaller posterior probability, but still larger than most arbitrary allocations, thus defining the separate smaller mode.

It could be argued that our interpretation of the results is influenced by the fact that in reality we know there to be two subpopulations present. In an effort to present a balanced judgement, we consider other factors that could conceivably lead to the observation of this apparent multimodality.

It should be noted that a run of 10000 is rather short, particularly considering

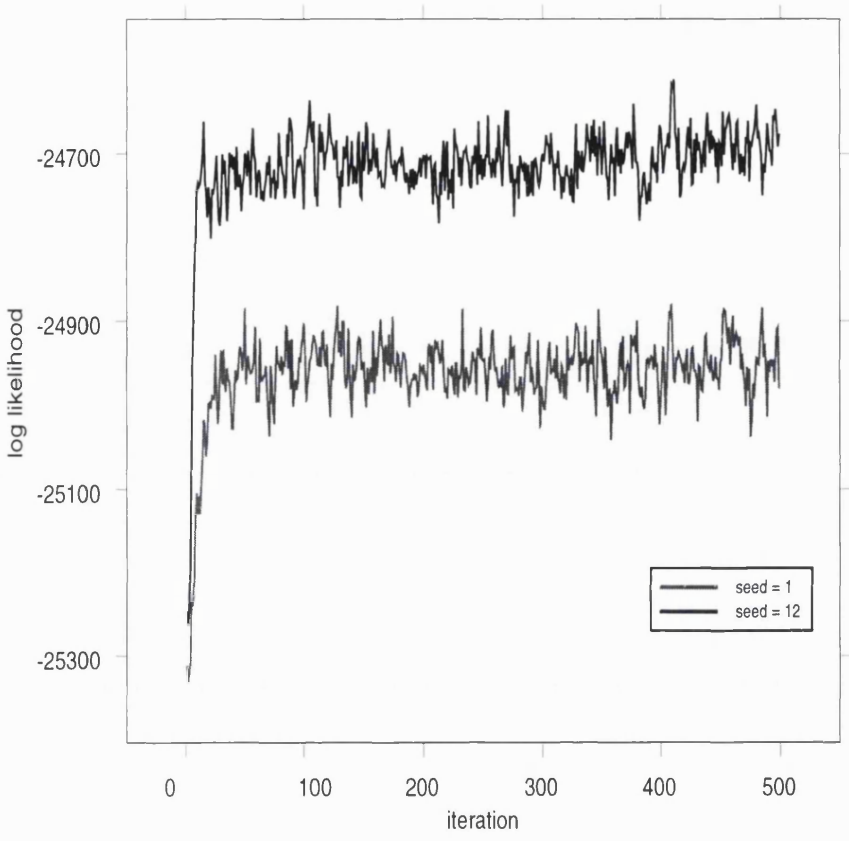


Figure 13: Comparison of log likelihood ($\Pr(\chi_\alpha = \xi_\alpha | \boldsymbol{G}^{(t)}, \boldsymbol{I}^{(t)})$) for random seeds 1 and 12.

the apparently multimodal nature of the posterior, and it could be argued that the multimodality is a convergence problem. With the current program, much longer runs take a prohibitive length of time, but a single run of length 1 million was conducted, initiated by a random seed of 1. The results were very close to those of Table 8 under the same seed, suggesting that the chain remains in this mode throughout the longer run. This result does not prove that the multimodality is not a convergence problem, but does suggest that the chain would have to be run for a very long time to before any other result might be observed. This is not practical, confirming that mixing should be encouraged via the methods of Chapter 6.

There is also the possibility that multimodality is a result of there being more than η subpopulations present.

The nature of these modes is elucidated in Section 8.4.2 in which the subpopulation proportions ($\kappa(l)$) are assumed unknown. This allows us to witness

the effect on the posterior of varying the amount of prior information available upon these proportions.

In the general case of η observed subpopulations (and the number of subpopulations being defined as η), there are $\eta!$ modes, each corresponding to a particular allocation of the clusters in the database to the 'true' subpopulations. Only if the subpopulation proportions ($\kappa(l)$) are equal for all l will the database clusters be similar across all modes.

Should the number of subpopulations present be greater than η some individuals will be forced to be grouped with individuals of another subpopulation. This could result in a greater number of broad groupings as there is a potentially larger number of ways in which the subpopulations could combine.

Thus we would maintain that the multimodality itself is a result of there being a number of subpopulations, and that this multimodality could increase if the number of subpopulations is incorrectly specified. Such reasoning could conceivably be incorprated in a method to establish the ideal number of subpopulations to specify. We consider in Chapter 10 the potential for further study that arises from the restriction we have imposed by assuming the number of subpopulations known. However it should be stressed that the study assuming η known is still of great value, even if it is incorrectly specified. This is because groupings of individuals who are most genetically similar will still have the greatest likelihoods under a given specification of η, and we will still be accounting for a great deal of the substructure present.

Ideally, one would like to design an MCMC scheme which spends an amount of time in each mode proportional to the size of the mode. When one mode is far greater than the others, as in the observed case, the chain should spend the majority of time in this mode. The initial scheme used gives satisfactory estimates if it initially moves to the 'correct' allocation, but has a very low probability of moving to this mode if the chain's starting point leads towards the 'incorrect' allocation. This problem is highlighted by the results of the run of length 1 million along which there was still no 'move' from the smaller mode.

Considering the allocation diagram of Figure 12, movement from one mode to another would be facilitated by the reallocation of a large number of individuals from one subpopulation to another.

Thus an initial attempt to improve the mixing was made involving the introduction of an extra parameter. We now distinguish between the arbitrary subpopulations of the MCMC scheme and the true 'real world' subpopulations, and introduce $\boldsymbol{S}$ as a parameter mapping one onto the other. The r^{th} entry of this vector is an integer representing the true subpopulation corresponding to arbitrary subpopulation r. A symmetric prior π_S is defined such that all $r!$ possible permutations are considered equally likely. In the two subpopulation case that we are currently considering, the subpopulations are coded

$$\begin{aligned} &\text{Caucasian:} && 1 \\ &\text{Afro-Caribbean:} && 2 \end{aligned}$$

and thus, in the prior,

$$\boldsymbol{S} = \begin{cases} (1,2) & \text{with probability } \frac{1}{2}; \\ (2,1) & \text{with probability } \frac{1}{2}. \end{cases}$$

At any iteration, a change in $\boldsymbol{S}$ would immediately reallocate all individuals to a different true subpopulation, corresponding to a change in mode. The updated DAG for this scheme is shown in Figure 14.

The full conditional probability distribution of $\boldsymbol{S}$ is

$$\begin{aligned} \Pr(\boldsymbol{S}|\ldots) &= \Pr(\boldsymbol{S}|\boldsymbol{I}, \pi_S, \boldsymbol{\kappa}) \\ &\propto \Pr(\boldsymbol{I}|\boldsymbol{S}, \boldsymbol{\kappa}).\pi_S(\boldsymbol{S}) \\ &\propto \prod_{i=1}^{n} \kappa(S(I_i)).\pi_S(\boldsymbol{S}) \\ &\propto \prod_{l=1}^{\eta} \kappa(S(l))^{n_a(l)}.\pi_S(\boldsymbol{S}), \end{aligned}$$

where $n_a(l)$ is the number of individuals allocated to subpopulation $\mathcal{P}_l$.

Once the chain has settled into one mode with a particular allocation, the conditional probability of the current $\boldsymbol{S}$ will be much larger than for other possible permutations, unless the subpopulation proportions are close in size. This means that a change of $\boldsymbol{S}$, and therefore a change of mode, is unlikely.

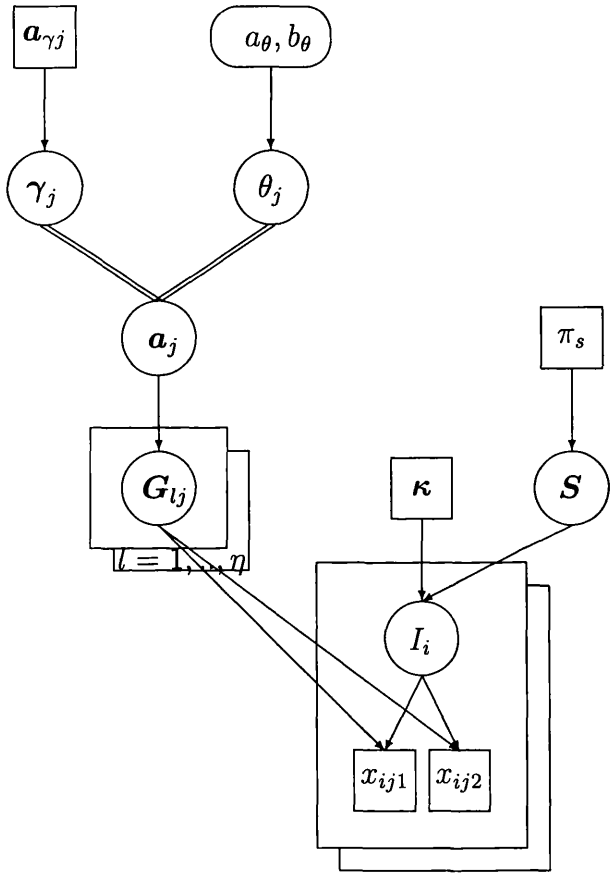


Figure 14: DAG for the case of incomplete information, including 'mixing' variable S.

However, mixing is improved by the introduction of a simulated tempering phase (see Section 6.7) in addition to the 'allocation parameter' $\boldsymbol{S}$. This involves proposing a 'temperature' change at each iteration of the chain. 'Heating' the posterior 'flattens' the density to make a move from one mode to another more likely. The lowest temperature corresponds to the true posterior, and it is only samples made at this level that are retained for estimation purposes.

In the case considered here, 'temperature' changes are represented by changes in the subpopulation proportions. The 'coldest' level 0 corresponds to the true subpopulation proportions. A move from one mode to another is represented by a change in $\boldsymbol{S}$. As this is unlikely to happen unless the proportions $(\kappa(1), \ldots, \kappa(\eta))$ are reasonably similar, at the hottest level v the subpopulation proportions are set equal to $\frac{1}{\eta}$. Setting up the scheme becomes more complicated with a greater number of subpopulations. It has thus far been used successfully in the two subpopulation case described. In this case, the change in 'temperature' between adjacent levels is represented by a change in $\kappa(1)$ of δ, $\kappa(2)$ changing by a similar magnitude in the opposite direction.

Iteration t of the MCMC scheme now involves the sampling of $\boldsymbol{a}^{(t)}$, $\boldsymbol{G}^{(t)}$, $\boldsymbol{I}^{(t)}$ (and consequently the numbers of alleles $(\boldsymbol{n}_{lj}^{(t)})$ of each type at each locus j within each subpopulation $\mathcal{P}_l$), and $\boldsymbol{S}^{(t)}$. In addition we must propose a change in the temperature level. If the process is at level r, we propose a move to level w with probability $q_{r,w}$, where

$$\begin{aligned} q_{0,1} &= 1 \\ q_{v,v-1} &= 1 \\ q_{r,r-1} = q_{r,r+1} &= \frac{1}{2} \qquad \text{if } 0 < r < h \\ q_{r,w} &= 0 \qquad \text{otherwise.} \end{aligned}$$

As the temperature level is characterized by the subpopulation proportions, the acceptance probability α_{acc} involves only terms of the posterior dependent upon $\boldsymbol{\kappa}$.

For a proposed move from level r to level w, the acceptance probability is

given by

$$\begin{aligned}\alpha_{acc} &= \min\left\{1, \frac{c_w \pi(\boldsymbol{a}, \boldsymbol{G}, \boldsymbol{I}, \boldsymbol{S}|\boldsymbol{\kappa}_w \chi_\alpha = \xi_\alpha) q_{w,r}}{c_r \pi(\boldsymbol{a}, \boldsymbol{G}, \boldsymbol{I}, \boldsymbol{S}|\boldsymbol{\kappa}_r, \chi_\alpha = \xi_\alpha) q_{r,w}}\right\} \\ &= \min\left\{1, \frac{c_w [\prod_{l=1}^{\eta} \kappa_w (S_l)^{n_a(l)}] q_{w,r}}{c_r [\prod_{l=1}^{\eta} \kappa_r (S_l)^{n_a(l)}] q_{r,w}}\right\},\end{aligned}$$

where (c_r) is a series of constants chosen to ensure a reasonable acceptance rate and $\boldsymbol{\kappa}_r$ represents the subpopulation proportions characterizing temperature level (r = 0, ..., v).

With $v + 1$ temperature levels overall, the difference between the subpopulation proportions in the two subpopulation case at level r is $2(v - r)\delta$, i.e. it becomes smaller as the temperature increases, and thus a change in $\boldsymbol{S}$ becomes more likely, facilitating a switch from one mode to the other.

The MCMC scheme was set up with 11 levels, and run for 10000 iterations with a burn-in of 3000. The greater the number of levels, the longer the chain must be, as only the iterations at level 0 are used to calculate the final estimator.

Figure 15 shows a section of the sampler initiated by a random seed of 1, which originally led to the chain being unable to move from the smaller mode, together with a trace of $G_{S_{13}}(6)$. This is the allele frequency of the larger subpopulation directly comparable to that traced in Figure 9. This shows movement from the smaller mode to the larger facilitated by 'heating' of the simulated tempering scheme.

Match probability estimates for the profile AC_c under this scheme can be seen in Table 9. The Markov chains used for these results are initiated using the same random seeds as previously.

The lack of variation across random seed suggests that the simulated tempering aids mixing of the process. The similarity of these results to those under the original scheme corresponding to seeds 12, 12000 and 10^{-6} is to be expected if the chain mixes properly according to the relative size of the modes.

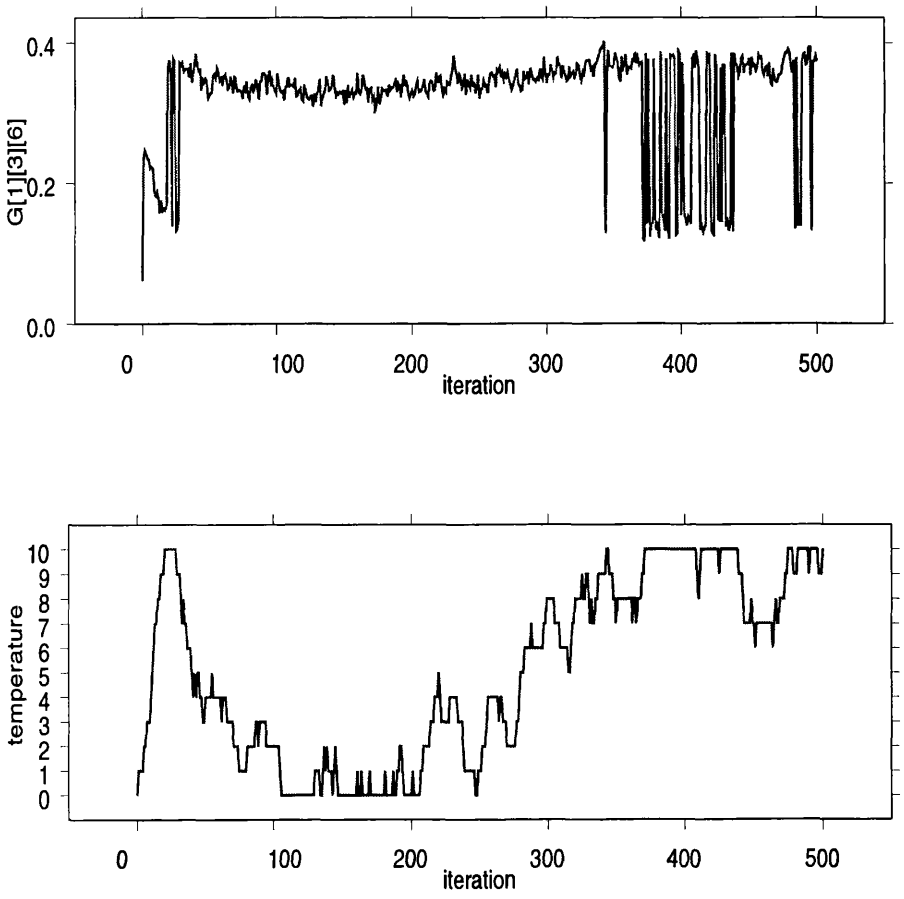


Figure 15: Traces of simulated tempering level and an allele frequency.

random seed	m_{Ca}	m_{AC}	m
1	5.84×10^{-10}	6.39×10^{-6}	1.82×10^{-6}
12	5.90×10^{-10}	6.41×10^{-6}	1.83×10^{-6}
12 000	5.80×10^{-10}	6.46×10^{-6}	1.84×10^{-6}
125 000	6.10×10^{-10}	6.20×10^{-6}	1.77×10^{-6}
560 000	5.36×10^{-10}	6.47×10^{-6}	1.84×10^{-6}
10^{-6}	5.74×10^{-10}	6.45×10^{-6}	1.84×10^{-6}
36×10^{6}	5.96×10^{-10}	6.46×10^{-6}	2.25×10^{-6}
Average:	5.81×10^{-10}	6.41×10^{-12}	1.83×10^{-7}

Table 9: Posterior match probabilities when κ is known, employing simulated tempering.

8.4.2 Subpopulation proportions assumed unknown

The level of information provided by the subpopulation proportions in the absence of individual subpopulation identifiers should not be underestimated.

If subpopulation identifiers are unavailable, one is looking to allocate the individuals of the database according to the most likely genetic clusters. If the size of these clusters can be accurately estimated, a large number of potential allocations will be eliminated. This can be seen mathematically by considering the conditional distributions of $\boldsymbol{I}$ given $\boldsymbol{G}$ in the presence and absence of known subpopulation proportions.

If the subpopulation proportions are considered known,

$$\Pr(\boldsymbol{I}|\boldsymbol{\kappa}, \chi_\alpha = \xi_\alpha, \boldsymbol{G}) \propto \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{I}, \boldsymbol{G}).\Pr(\boldsymbol{I}|\boldsymbol{\kappa}), \tag{60}$$

meaning that the likelihood for each possible allocation is weighted by a prior probability based upon $\boldsymbol{\kappa}$. If the proportions are unknown, these weights are no longer available, and integration across $\boldsymbol{\kappa}$ is required,

$$\begin{aligned}
\Pr(\boldsymbol{I}|\chi_\alpha = \xi_\alpha, \boldsymbol{G}) &= \mathrm{E}[\Pr(\boldsymbol{I}|\chi_\alpha = \xi_\alpha, \boldsymbol{G}, \boldsymbol{\kappa})|\boldsymbol{G}, \chi_\alpha = \xi_\alpha] \\
&\propto \int_{\boldsymbol{\kappa}} \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{I}, \boldsymbol{G}).\Pr(\boldsymbol{I}|\boldsymbol{\kappa}).\pi(\boldsymbol{\kappa})\mathrm{d}\boldsymbol{\kappa} \\
&\propto \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{I}, \boldsymbol{G}) \int_{\boldsymbol{\kappa}} \Pr(\boldsymbol{I}|\boldsymbol{\kappa}).\pi(\boldsymbol{\kappa})\mathrm{d}\boldsymbol{\kappa}.
\end{aligned}$$

This means that the weighting for each allocation's likelihood is an average across all possible $\boldsymbol{\kappa}$. If the prior placed upon $\boldsymbol{\kappa}$ is highly concentrated, these weights should be close to those of (60). If, however, the prior is vague, the weightings are likely to be more uniform meaning that one is relying to a greater extent upon the data to form clusters via the likelihood.

A Dirichlet prior with parameters $(\pi_\kappa(1), \ldots, \pi_\kappa(\eta))$ is placed upon $\boldsymbol{\kappa}$.

To include $\boldsymbol{\kappa}$ in the MCMC scheme requires sampling from its full conditional (now assuming the original model, excluding $\boldsymbol{S}$),

$$\begin{aligned}
\pi(\boldsymbol{\kappa}|\ldots) &= \pi(\boldsymbol{\kappa}|\boldsymbol{I}) \\
&\propto f(\boldsymbol{I}|\boldsymbol{\kappa}).\pi(\boldsymbol{\kappa}) \\
&\propto \prod_{l=1}^{\eta} \kappa(l)^{n_a(l)+\pi_\kappa(l)} \\
&\Rightarrow \boldsymbol{\kappa}|\ldots \sim \text{Dirichlet}(n_a(1)+\pi_\kappa(1),\ldots,n_a(\eta)+\pi_\kappa(\eta)).
\end{aligned}$$

As well as providing information relevant to the clustering process, $\boldsymbol{\kappa}$ has also been used to this point when calculating the weighted sum of the match probabilities $\sum_{l=1}^{\eta} \lambda_l m_l$, where $\lambda_l = \Pr(C \in P_l | C \notin \alpha, \varepsilon)$, α being the database of individuals including the suspect. If it is assumed that there is no additional information upon the culprit's subpopulation available within the non-DNA evidence ε, we may assume that $\lambda_l = \kappa_l$ for all $l = 1,\ldots,\eta$. This suggests that λ_l can be estimated by an ergodic average over (t) of $(\kappa(l)^{(t)})$, and the required weighted sum by an ergodic average of $(\sum_{l=1}^{\eta} \kappa(l)^{(t)} m_l^{(t)})$. The overall match probabilities quoted in this section refer to these ergodic averages.

The match probabilities in Tables 10 - 13 for the suspect profile AC_c result from a series of priors placed upon $\boldsymbol{\kappa}$, from vague (Dirichlet(1,1)) to highly concentrated (Dirichlet(700,300)). These prior densities are plotted in Figure 16.

As the amount of prior information upon $\boldsymbol{\kappa}$ is increased, the distribution tends to a point mass at known proportions. Bearing in mind the results under known $\boldsymbol{\kappa}$, it is reasonable to expect a multimodal posterior once again, and this is reflected in the results of Tables 10 to 13.

Comparison of individual allocations under this series of priors elucidates the effect causing the multimodality described in the previous section.

Figures (17 - 24) show allocations within the two modes as the amount of prior information upon $\boldsymbol{\kappa}$ increases.

The allocations under the Dirichlet(1,1) prior are symmetric in the two modes, individuals being clustered accurately. As there is no information specifying which of the two subpopulations (Caucasian or Afro-Caribbean) is larger, it is to be expected that the larger subpopulation is labelled Caucasian in half the runs.

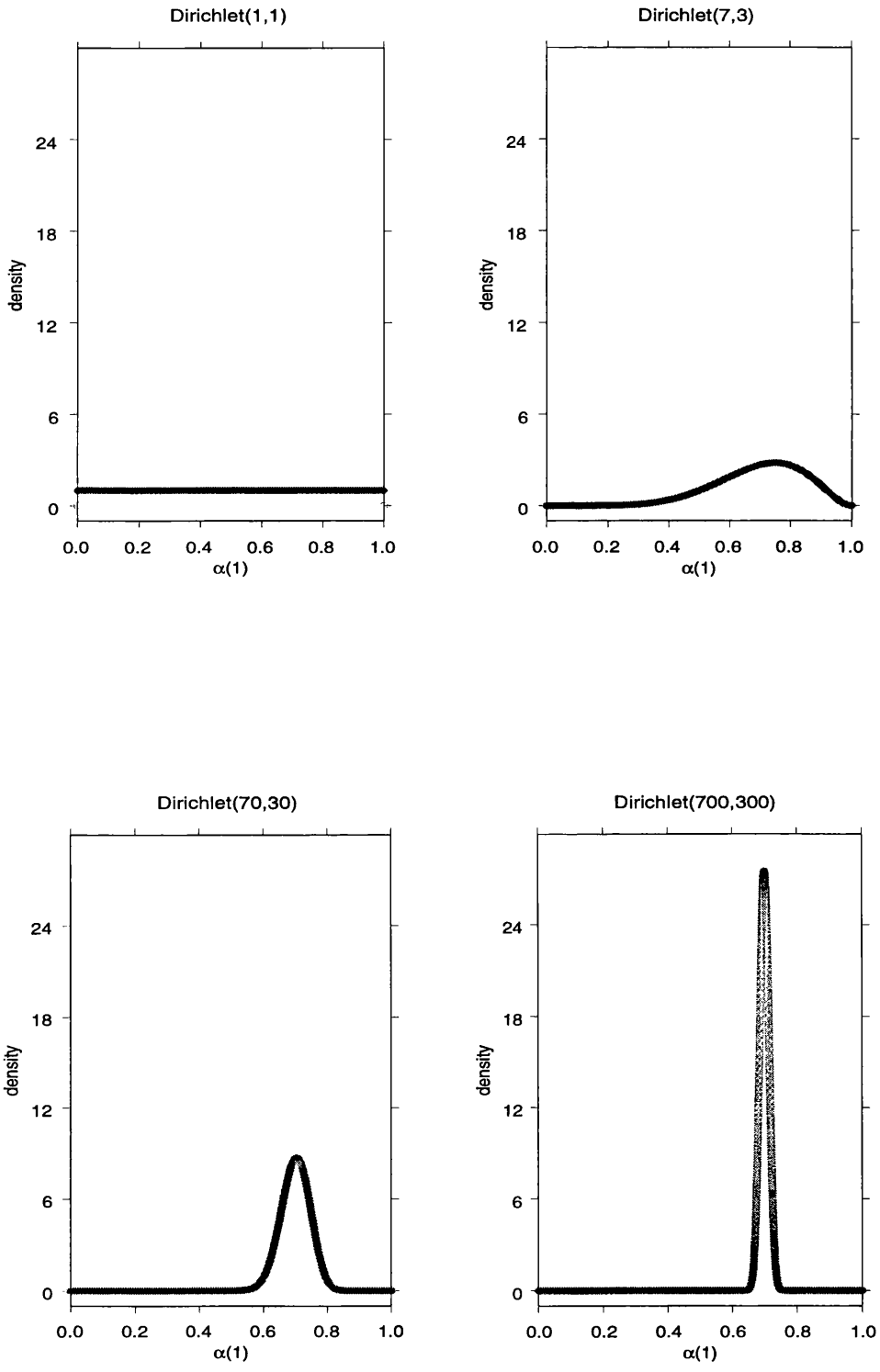


Figure 16: Prior densities for $\kappa(1)$.

random seed	$\hat{\kappa}(1)$	m_{Ca}	m_{AC}	m
12	0.715	10.0×10^{-10}	9.07×10^{-6}	2.24×10^{-6}
12 000	0.783	10.5×10^{-10}	8.56×10^{-6}	2.13×10^{-6}
10^{-6}	0.759	9.46×10^{-10}	8.67×10^{-6}	2.15×10^{-6}
36×10^{6}	0.729	10.7×10^{-10}	8.83×10^{-6}	2.18×10^{-6}
Average:	0.747	10.2×10^{-10}	8.78×10^{-6}	2.17×10^{-6}
1	0.282	8.63×10^{-6}	9.56×10^{-10}	2.14×10^{-6}
125 000	0.244	8.50×10^{-6}	9.86×10^{-10}	2.11×10^{-6}
560 000	0.279	8.43×10^{-6}	10.1×10^{-10}	2.09×10^{-6}
Average:	0.268	8.52×10^{-6}	9.84×10^{-10}	2.11×10^{-6}

Table 10: Posterior match probabilities when $\boldsymbol{\kappa} \sim \text{Dirichlet}(1, 1)$.

random seed	$\hat{\kappa}(1)$	m_{Ca}	m_{AC}	m
12	0.742	10.9×10^{-10}	8.44×10^{-6}	2.08×10^{-6}
12 000	0.756	9.54×10^{-10}	7.99×10^{-6}	2.01×10^{-6}
125 000	0.794	9.55×10^{-10}	8.26×10^{-6}	2.07×10^{-6}
10^{-6}	0.725	9.46×10^{-10}	8.67×10^{-6}	2.15×10^{-6}
36×10^{6}	0.755	10.4×10^{-10}	8.25×10^{-6}	2.05×10^{-6}
Average:	0.754	9.97×10^{-10}	8.32×10^{-6}	2.07×10^{-6}
1	0.308	7.68×10^{-6}	8.05×10^{-10}	2.01×10^{-6}
560 000	0.268	7.32×10^{-6}	8.52×10^{-10}	1.94×10^{-6}
Average:	0.288	7.50×10^{-6}	8.29×10^{-10}	1.98×10^{-6}

Table 11: Posterior match probabilities when $\boldsymbol{\kappa} \sim \text{Dirichlet}(7, 3)$.

random seed	$\hat{\kappa}(1)$	m_{Ca}	m_{AC}	m
12	0.725	8.19×10^{-10}	7.32×10^{-6}	1.94×10^{-6}
12 000	0.715	8.37×10^{-10}	7.57×10^{-6}	1.96×10^{-6}
10^{-6}	0.742	8.85×10^{-10}	7.79×10^{-6}	2.03×10^{-6}
36×10^{6}	0.768	8.19×10^{-10}	7.65×10^{-6}	2.01×10^{-6}
Average:	0.738	8.40×10^{-10}	7.58×10^{-6}	1.99×10^{-6}
1	0.349	3.18×10^{-6}	1.78×10^{-10}	1.18×10^{-6}
125 000	0.369	3.26×10^{-6}	1.41×10^{-10}	1.21×10^{-6}
560 000	0.382	3.19×10^{-6}	1.74×10^{-10}	1.19×10^{-6}
Average:	0.367	3.21×10^{-6}	1.64×10^{-10}	1.19×10^{-6}

Table 12: Posterior match probabilities when $\boldsymbol{\kappa} \sim$ Dirichlet(70, 30).

random seed	$\hat{\kappa}(1)$	m_{Ca}	m_{AC}	m
12	0.719	5.64×10^{-10}	6.31×10^{-6}	1.82×10^{-6}
12 000	0.707	5.58×10^{-10}	6.27×10^{-6}	1.81×10^{-6}
10^{-6}	0.713	5.08×10^{-10}	6.23×10^{-6}	1.80×10^{-6}
Average:	0.713	5.43×10^{-10}	6.27×10^{-6}	1.81×10^{-6}
1	0.661	5.27×10^{-7}	9.53×10^{-12}	3.36×10^{-7}
125 000	0.652	4.93×10^{-7}	1.73×10^{-11}	3.15×10^{-7}
560 000	0.642	5.12×10^{-7}	1.84×10^{-11}	3.28×10^{-7}
36×10^{6}	0.660	5.29×10^{-7}	1.33×10^{-11}	3.38×10^{-7}
Average:	0.654	5.19×10^{-7}	1.46×10^{-11}	3.29×10^{-7}

Table 13: Posterior match probabilities when $\boldsymbol{\kappa} \sim$ Dirichlet(700, 300).

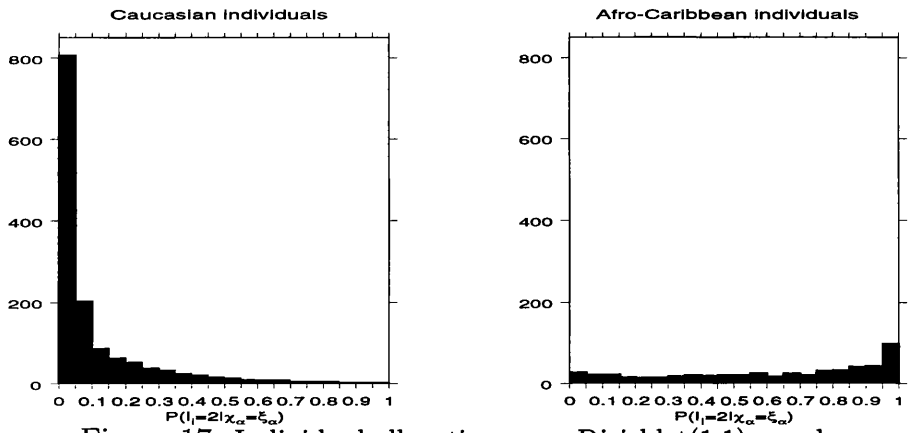


Figure 17: Individual allocation: $\boldsymbol{\kappa} \sim$ Dirichlet(1,1), random seed = 12.

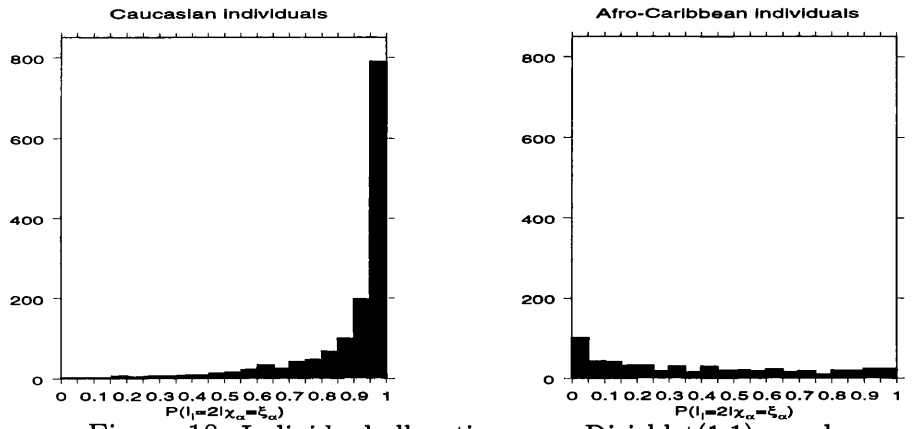


Figure 18: Individual allocation: $\boldsymbol{\kappa} \sim$ Dirichlet(1,1), random seed = 1.

Particular attention should be paid to results under the symmetric prior in which $\pi_{\boldsymbol{\kappa}}(l) = 1$ for all l. It is equivalent to placing one unknown individual in each subpopulation before the analysis is begun.

Such a prior results in a problem of identifiability. In the case of two subpopulations, A and B say, it is not possible to differentiate between the case in which the individuals truly in subpopulation A are generally allocated to arbitrary subpopulation 1, and that in which they are allocated to arbitrary subpopulation 2.

The general problem can be made identifiable by ordering the subpopulation proportions, $\kappa(1) > \kappa(2) > \ldots > \kappa(\eta)$ for example, restricting the chain to a particular mode. In our application however, such a lack of identifiability is unimportant, as the overall match probability $m = \sum_{l=1}^{\eta} \lambda_l m_l$ is itself symmet-

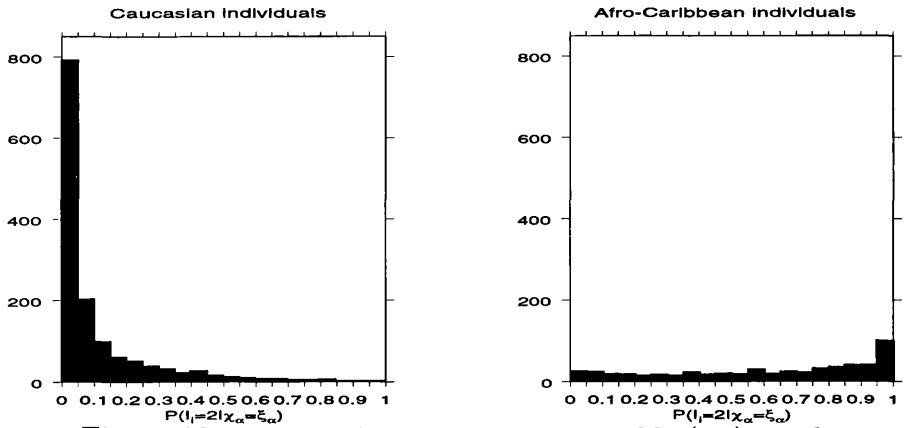


Figure 19: Individual allocation: $\kappa \sim$ Dirichlet(7,3), random seed = 12.

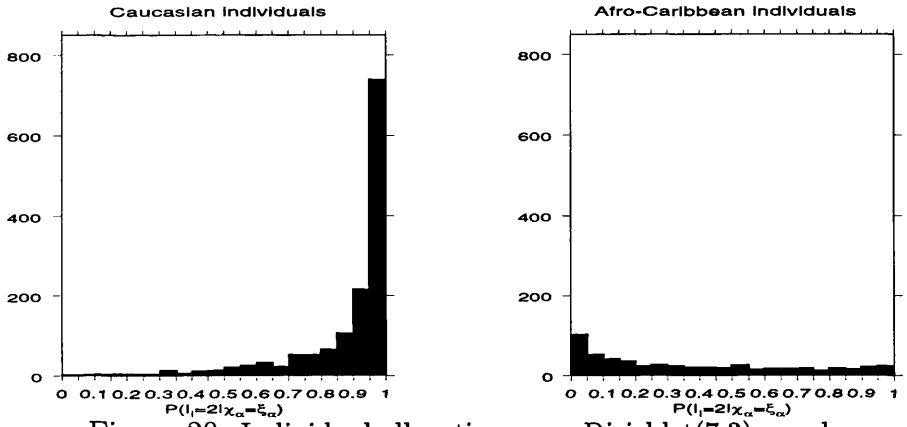


Figure 20: Individual allocation: $\kappa \sim$ Dirichlet(7,3), random seed = 1.

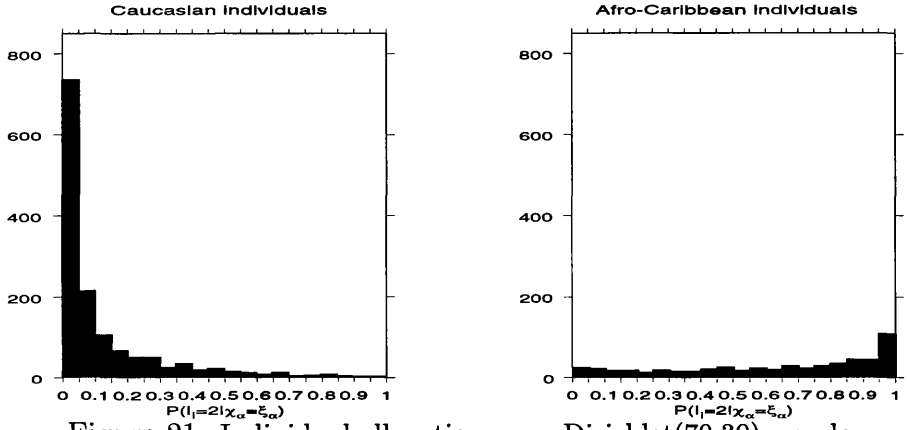


Figure 21: Individual allocation: $\kappa \sim$ Dirichlet(70,30), random seed = 12.

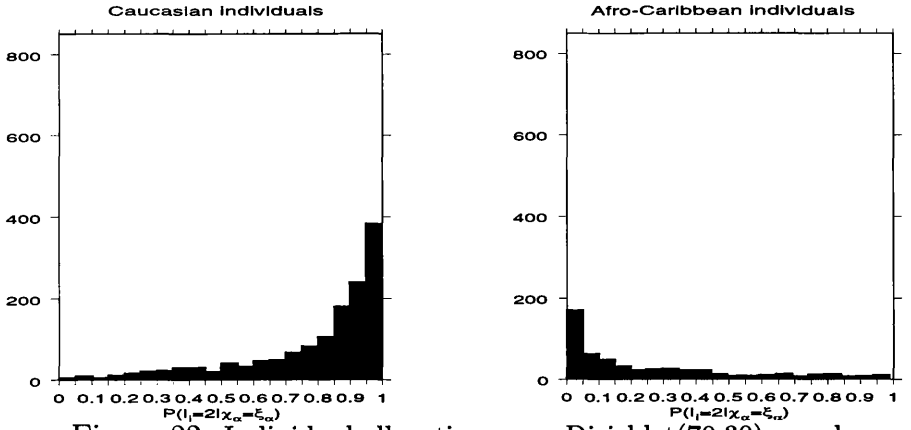


Figure 22: Individual allocation: $\kappa \sim$ Dirichlet(70,30), random seed = 1.

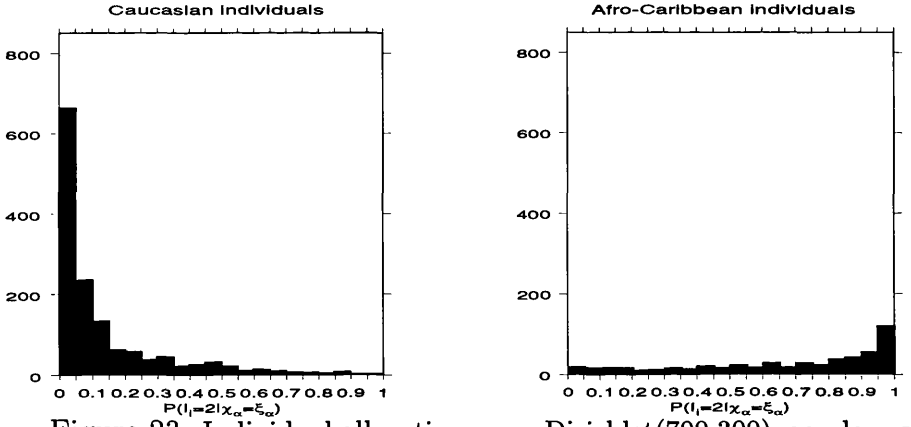


Figure 23: Individual allocation: $\kappa \sim$ Dirichlet(700,300), random seed = 12.

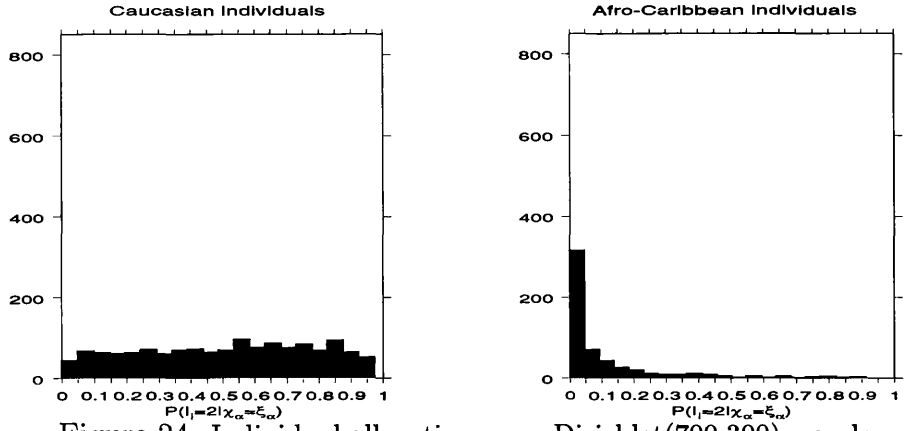


Figure 24: Individual allocation: $\kappa \sim$ Dirichlet(700,300), random seed = 1.

rical if it is assumed that $\lambda_l = \kappa(l)$ where $\lambda_l = \Pr(C \in \mathcal{P}_l | C \notin \alpha)$.

This means that the resultant match probability should be the same however the subpopulations are allocated, a fact substantiated by the overall match probabilities displayed in Table 10.

As the total of the parameters in the Dirichlet distribution increases, the allocations tend to that of known $\boldsymbol{\kappa}$ (Figures 10, 11). The trend of these graphs makes clear how each mode attributes one of the clusters of the data to the Caucasian subpopulation, and one to the Afro-Caribbean. With a small amount of prior information, the relative size of these clusters dictates the posterior estimates of the subpopulation proportions $\boldsymbol{\kappa}$.

The Dirichlet(700, 300) distribution effectively represents the addition of 1000 individuals of unknown profile to the database in the ratio 70:30 of Caucasians to Afro-Caribbeans. With a database of 1960 individuals (including the suspect), this is highly influential in the posterior distribution of $\boldsymbol{\kappa}$, and when the process attempts to allocate the cluster truly corresponding to the Afro-Caribbean subpopulation to the Caucasian subpopulation, the appropriate split in terms of individual numbers is not possible. Thus, at each iteration, a number of true Caucasian individuals are grouped with this cluster to form the larger subpopulation.

If there is some information available facilitating the specification of an asymmetric prior upon $\boldsymbol{\kappa}$, it is important that the chain mixes properly, occupying the modes for the appropriate amount of time. As $\boldsymbol{\kappa}$ is now a variable, it seems possible that its movement could induce switching in the modes without the simulated tempering if the parameter $\boldsymbol{S}$ were again introduced to the scheme. However, if the prior upon $\boldsymbol{\kappa}$ displays a small variance, it is unlikely that $\boldsymbol{\kappa}$ will be allowed to move sufficiently for the desired switching to take place.

In this case, the simulated tempering scheme would be more difficult to employ. This is due to the temperature levels being defined by the subpopulation proportions ($\kappa(l)$). As this is now a random variable, it can no longer be used to define the coldest temperature level. The 'known' constants are now the prior

parameters placed upon this variable, and it is not clear that adjusting these will encourage the chain to change mode in the same way that adjusting $\boldsymbol{\kappa}$ does when the proportions are known. For this reason we opt to use an importance sampling scheme (as described in Chapter 6) in this case. To use this technique one must have a density $\pi^*(\boldsymbol{a}, \boldsymbol{G}, \boldsymbol{I}, \boldsymbol{\kappa}|\chi_\alpha = \xi_\alpha)$ which can be easily sampled from and which mixes well. Although the chain with a stationary distribution of the posterior with a Dirichlet(1,1) prior upon $\boldsymbol{\kappa}$ does not mix well, it would seem reasonable to induce mixing by manually changing the parameter $\boldsymbol{S}$ at set points. It is only allowable to do this with symmetric priors, because it is known that the point in the other mode with the parameters at the same values, but individuals allocated to different subpopulations, will have equal posterior density.

Thus we define

$$\pi^*(\boldsymbol{a}, \boldsymbol{G}, \boldsymbol{I}, \boldsymbol{\kappa}|\chi_\alpha = \xi_\alpha) = \pi(\boldsymbol{a}, \boldsymbol{G}, \boldsymbol{I}, \boldsymbol{\kappa}|\pi_\kappa = (1, \ldots, 1), \chi_\alpha = \xi_\alpha),$$

and parameter values from the resultant chain are weighted at each iteration (t) to give match probability estimates,

$$\hat{m}_l = \frac{\sum_t w_t c \prod_{j=1}^{M} \frac{(a_j^{(t)}(y_{j1}) + n_{lj}^{(t)}(y_{j1}))(a_j^{(t)}(y_{j2}) + n_{lj}^{(t)}(y_{j2}) + \delta_j)}{(a_j^{(t)}(+) + n_{lj}^{(t)}(+))(a_j^{(t)}(+) + n_{lj}^{(t)}(+) + 1)}}{\sum_t w_t},$$

where

$$\begin{aligned} w_t &= \frac{\pi(\boldsymbol{a}^{(t)}, \boldsymbol{G}^{(t)}, \boldsymbol{I}^{(t)}, \boldsymbol{\kappa}^{(t)}|\chi_\alpha = \xi_\alpha)}{\pi^*(\boldsymbol{a}^{(t)}, \boldsymbol{G}^{(t)}, \boldsymbol{I}^{(t)}, \boldsymbol{\kappa}^{(t)}|\chi_\alpha = \xi_\alpha)} \\ &= \frac{\Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}^{(t)}, \boldsymbol{I}^{(t)})\pi(\boldsymbol{a}^{(t)})\pi(\boldsymbol{G}^{(t)}|\boldsymbol{a}^{(t)})\pi(\boldsymbol{I}^{(t)}|\boldsymbol{\kappa}^{(t)})\pi(\boldsymbol{\kappa}^{(t)})}{\Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}^{(t)}, \boldsymbol{I}^{(t)})\pi^*(\boldsymbol{a}^{(t)})\pi^*(\boldsymbol{G}^{(t)}|\boldsymbol{a}^{(t)})\pi^*(\boldsymbol{I}^{(t)}|\boldsymbol{\kappa}^{(t)})\pi^*(\boldsymbol{\kappa}^{(t)})} \\ &= \frac{\pi(\boldsymbol{\kappa}^{(t)})}{\pi^*(\boldsymbol{\kappa}^{(t)})} \\ &= \frac{\Gamma(\pi_\kappa(+))}{(\eta - 1)\prod_{l=1}^{\eta}\Gamma(\pi_\kappa(l))} \prod_{l=1}^{\eta} \kappa(l)^{(t)\pi_\kappa(l)-1}. \end{aligned}$$

Match probability estimates using this importance sampling method under the priors for $\boldsymbol{\kappa}$ used earlier are given in Tables 14 to 16.

It is interesting to note how the subpopulation match probabilities change with the prior subpopulation proportion parameters. Under the symmetric prior,

we always observe similar clusters, but with the larger cluster allocated to subpopulation 1 (Ca) or 2 (AC) at random. In the case of the Dirichlet(700, 300) we have a great deal of prior knowledge regarding the subpopulation proportions. Under this prior we are almost certain that the larger cluster corresponds to subpopulation 1, and therefore the importance sampling weightings are almost negligible when the chain is in the mode in which the subpopulation 2 cluster is larger. However, when the Dirichlet(7,3) prior is assumed, there is a greatly increased probability that '2' is the larger subpopulation. This means that the larger match probability generally attributed to the Afro-Caribbean subpopulation has a contribution to the Caucasian match probability which is no longer negligible.

An important question arises in the case in which there is no apparent physical interpretation of the subpopulations. How many subpopulations are there?

If the subpopulations are purely a tool of the model, it seems unrealistic to consider the answer to this question known before the analysis is started. This is a problem which is considered in Chapter 10.

random seed	m_{Ca}	m_{AC}	m
1	1.15×10^{-7}	8.53×10^{-6}	2.15×10^{-6}
12	1.17×10^{-7}	9.16×10^{-6}	2.26×10^{-6}
12 000	1.06×10^{-7}	8.62×10^{-6}	2.17×10^{-6}
125 000	1.24×10^{-7}	8.36×10^{-6}	2.07×10^{-6}
560 000	1.20×10^{-7}	8.05×10^{-6}	2.03×10^{-6}
10^{-6}	1.17×10^{-7}	8.69×10^{-6}	2.18×10^{-6}
36×10^{6}	1.18×10^{-7}	8.84×10^{-6}	2.19×10^{-6}
Average:	1.17×10^{-7}	8.61×10^{-6}	2.15×10^{-6}

Table 14: Posterior match probabilities when $\boldsymbol{\kappa} \sim$ Dirichlet(7, 3).

Bearing in mind the lack of information generally available regarding the subpopulations present, it is likely that we would generally assume unknown sub-

random seed	m_{Ca}	m_{AC}	m
1	8.35×10^{-10}	7.79×10^{-6}	2.03×10^{-6}
12	9.09×10^{-10}	8.31×10^{-6}	2.13×10^{-6}
12 000	8.05×10^{-10}	7.91×10^{-6}	2.07×10^{-6}
125 000	8.86×10^{-10}	7.53×10^{-6}	1.94×10^{-6}
560 000	8.93×10^{-10}	7.22×10^{-6}	1.89×10^{-6}
10^{-6}	8.20×10^{-10}	8.10×10^{-6}	2.09×10^{-6}
36×10^{6}	9.54×10^{-10}	8.10×10^{-6}	2.08×10^{-6}
Average:	8.72×10^{-10}	7.85×10^{-6}	2.03×10^{-6}

Table 15: Posterior match probabilities when $\boldsymbol{\kappa} \sim$ Dirichlet(70, 30).

random seed	m_{Ca}	m_{AC}	m
1	5.99×10^{-10}	6.24×10^{-6}	1.80×10^{-6}
12	5.33×10^{-10}	6.41×10^{-6}	1.84×10^{-6}
12 000	5.74×10^{-10}	6.40×10^{-6}	1.83×10^{-6}
125 000	6.16×10^{-10}	5.66×10^{-6}	1.63×10^{-6}
560 000	6.49×10^{-10}	5.69×10^{-6}	1.65×10^{-6}
10^{-6}	5.11×10^{-10}	6.59×10^{-6}	1.90×10^{-6}
36×10^{6}	6.65×10^{-10}	6.46×10^{-6}	1.85×10^{-6}
Average:	5.92×10^{-10}	6.21×10^{-6}	1.79×10^{-6}

Table 16: Posterior match probabilities when $\boldsymbol{\kappa} \sim$ Dirichlet(700, 300).

2.5%	1.37×10^{-6}
5%	1.47×10^{-6}
25%	1.82×10^{-6}
Median	2.10×10^{-6}
75%	2.42×10^{-6}
95%	2.97×10^{-6}
97.5%	3.17×10^{-6}

Table 17: Quantiles of the posterior distribution of the overall match probability.

population labels and proportions with a Dirichlet$(1, \ldots, 1)$ prior placed upon $\boldsymbol{\kappa}$. The convergence factor of Gelman and Rubin (see Chapter 6) was estimated via CODA [BUGS] to be 1.01, with a 97.5% quantile of 1.03, for this series of seven runs. This suggests that the sequences have converged, and should be reliable for posterior inference.

The overall mean match probability was estimated to be 2.15×10^{-6} with an estimated standard error of 7.16×10^{-9}. Table 17 shows a summary of the posterior distribution of this overall match probability. In particular, the 95% posterior credible interval for the overall match probability is $(1.37 \times 10^{-6}, 2.42 \times 10^{-6})$. It is interesting to compare this to the model II estimate, assuming subpopulation labels known, of 2.00×10^{-6}.

Having calculated the posterior credible interval, it is necessary to consider how the presence of a credible interval of this magnitude affects the conclusions drawn from our results. Table 18 shows a comparison of posterior probabilities of guilt under a range of prior probabilities. It can be seen that the uncertainty in the match probability does lead to a degree of uncertainty in the posterior probability of guilt. However, in the cases shown, even the discrepancy between values at the extremes of the distribution is unlikely to affect the final decision with regard to the guilt of the suspect.

	Match probability		
Prior probability of guilt	1.37×10^{-6}	2.10×10^{-6}	3.17×10^{-6}
10^{-3}	1.00	1.00	1.00
10^{-4}	0.99	0.98	0.97
10^{-5}	0.88	0.83	0.76
4×10^{-6}	0.74	0.66	0.56
2×10^{-6}	0.59	0.49	0.39
10^{-6}	0.42	0.32	0.24
10^{-7}	0.07	0.05	0.03
10^{-8}	0.01	0.00	0.00

Table 18: Posterior probabilities of guilt for an individual with profile AC_C under a range of prior probabilities of guilt. This shows how the result of interest varies across the posterior distribution of the match probability.

9 Alternative methods

9.1 Introduction

This work is motivated by previously published research in the area. In this chapter we consider two papers which have sought to advance the methods used to present a DNA profile match as evidence within the courtroom.

This chapter describes the methods of papers by Roeder, Escobar, Kadane and Balazs [Roeder, Escobar, Kadane and Balazs, 1998], and Foreman, Smith and Evett [Foreman, Evett and Smith, 1997]. Areas in which it is felt omissions have been made are highlighted and contrasted with the approaches described in this thesis.

Section 9.4 presents comparisons of results under the methods of Foreman *et al.* and of this thesis. While showing that the method of Foreman *et al.* is generally conservative, these comparisons suggest that there are significant gains in accuracy to be made by using the methods of this thesis. It is demonstrated that the two methods could result in different decisions being reached regarding the guilt of the suspect given the profile match.

9.2 Roeder, Escobar, Kadane and Balazs

This analysis [Roeder, Escobar, Kadane and Balazs, 1998] is based upon a model similar to that described in Chapter 3. This means that we have levels describing:

(i) inheritance of a particular profile $\boldsymbol{x}_i$ by each individual i within a subpopulation $(\mathcal{P}_l, l = 1, \ldots, \eta)$,

$$\Pr(X_{ijb} = k|\boldsymbol{G}) = G_{lj}(k),$$

independently across i, l, band $(b = 1, 2)$ and locus $(j = 1, \ldots, M)$, the collection of observable alleles at a particular locus being denoted by $k = 1, \ldots, r_j$.

(ii) Generation of the allele probabilities $\boldsymbol{G}$ in each subpopulation,

$$G_{lj} \sim \text{Dirichlet}(a_j(1), a_j(2), \ldots, a_j(r_j)), \text{ independently for all } l, j, \quad (61)$$

where $\boldsymbol{a}_j = \frac{1-\theta_j}{\theta_j}\boldsymbol{\gamma}_j$.

(iii) The generation of the ancestral population parameters $(\boldsymbol{\gamma}, \boldsymbol{\theta})$ from a 'hyperprior' distribution,

$$\begin{aligned} \boldsymbol{\gamma}_j &\sim \text{Dirichlet}(a_\gamma(1), \ldots, a_\gamma(r_j)); \\ \theta_j &\sim \text{Beta}(a_\theta, b_\theta). \end{aligned}$$

Two extra levels are included by Roeder *et al.* to account for measurement error and coalescence, an effect causing a heterozygote (i.e. a pair in which the alleles are not the same) to appear as a homozygote due to the similarity between the allele lengths of the pair. Advances in technology mean that these levels are no longer necessary.

Roeder *et al.* estimate $\boldsymbol{\gamma}$ empirically, and particular emphasis is placed upon inference about the subpopulation differentiation parameter $\boldsymbol{\theta}$. They employ what has been termed model II in Chapter 4, i.e. the likelihood is defined as

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \Pr(\chi_\alpha = \xi_\alpha | \boldsymbol{\theta}, \boldsymbol{\gamma}), \quad (62)$$

where χ_α is the available data, the collection of profiles from the 'complete database' α comprising the original database δ and the suspect s. It is assumed throughout that the subpopulation membership I_i of each individual i is unknown.

Given $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, the profiles of individuals are independent if they are in different subpopulations. However, given only $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, individuals within a subpopulation are exchangeable, but not independent. This means that terms in which there are at least two individuals allocated to the same subpopulation are more complicated.

Roeder *et al.* consider the case in which the number of subpopulations is much greater than the number of individuals in the database, i.e. $\eta >> n_\alpha$. In

this case, the probability that two individuals belong to the same subpopulation becomes small, and thus the likelihood (62) is dominated by terms involving the product across conditionally independent individuals. In fact Roeder *et al.* use an approximation in which only these terms contribute to the likelihood:

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\gamma}) &\approx \prod_{i\in\alpha} \Pr(\boldsymbol{X}_i|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{I}).\Pr(\boldsymbol{I}|\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &= \prod_{i\in\alpha}\prod_{j=1}^{M} \Pr(\boldsymbol{X}_{ij}|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{I}).\Pr(\boldsymbol{I}|\boldsymbol{\theta},\boldsymbol{\gamma}), \end{aligned}$$

where

$$\begin{aligned} \Pr(\boldsymbol{X}_{ij}|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{I}) &= \mathrm{E}[\Pr(\boldsymbol{X}_{ij}|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{I},\boldsymbol{G})|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{I}] \\ &= \int G_{I_ij}(x_{ij1})G_{I_ij}(x_{ij2})) \\ &\qquad \times \frac{\Gamma(a_j(+))}{\prod_{k=1}^{r_j} a_j(k)} \prod_{k=1}^{r_j} G_{I_ij}(k)^{a_j(k)-1}\mathrm{d}\boldsymbol{G}_j \\ &= \begin{cases} 2(1-\theta_j)\gamma_j(x_{ij1})\gamma_j(x_{ij2}) & \text{if } x_{ij1}\neq x_{ij2} \\ \theta_j\gamma_j(x_{ij1}) + (1-\theta_j)\gamma_j^2(x_{ij1}) & \text{if } x_{ij1}=x_{ij2}, \end{cases} \end{aligned}$$

where $h(r,s)$ indicates inequality between r and s.

As stated by Roeder *et al.* there are a number of advantages to using this likelihood. It is quite simply computed and requires no knowledge of individual subpopulation membership. Furthermore, specification of the number of subpopulations is not required.

The MCMC scheme [Roeder, Escobar, Kadane and Balazs, 1998a] employed by Roeder *et al.* utilises the definition of θ_j as the probability of identity by descent of a pair of alleles [Wright, 1951], augmenting the data set with additional variables indicating this property.

Roeder *et al.* consider match probability calculations under what are termed the 'Affinal' and 'Hardy-Weinberg' models.

The Affinal model refers to the case in which the defendant and perpetrator are assumed to be from the same subpopulation. It is assumed that the culprit is *not* a member of the database α. The match probability for an individual i outside the database α is then given by the product over loci of single locus match probabilities conditional upon the suspect's profile,

$$\Pr(\boldsymbol{X}_i = \boldsymbol{y} | \boldsymbol{X}_s = \boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \prod_{j=1}^{M} \Pr(\boldsymbol{X}_{ij} = \boldsymbol{y}_j | \boldsymbol{X}_{sj} = \boldsymbol{y}_j, \boldsymbol{\gamma}, \boldsymbol{\theta}),$$

where

$$\Pr(\boldsymbol{X}_{ij} = \boldsymbol{y}_j | \boldsymbol{X}_{sj} = \boldsymbol{y}_j, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \begin{cases} 2\frac{(\theta_j+(1-\theta_j)\gamma_j(y_{j1}))(\theta_j+(1-\theta_j)\gamma_j(y_{j2}))}{(1+\theta_j)(1+2\theta_j)} & \text{if } y_{1j} \neq y_{2j} \\ \frac{(2\theta_j+(1-\theta_j)\gamma_j(y_{j1}))(3\theta_j+(1-\theta_j)\gamma_j(y_{j1}))}{(1+\theta_j)(1+2\theta_j)} & \text{if } y_{j1} = y_{j2}. \end{cases} \tag{63}$$

The Hardy-Weinberg model assumes that the suspect and perpetrator come from distinct subpopulations, given that the suspect and perpetrator are not the same person, and gives the single locus match probability of equation (63).

Using the assumption of independence across loci, match probabilities are calculated using a mixed Caucasian/Afro-Caribbean database with population membership assumed unknown. Two methods of match probability calculation are considered, integration of the appropriate match probability across the posterior of $\boldsymbol{\theta}$, and estimation of this posterior expectation by substitution of the posterior median of $\boldsymbol{\theta}$ into the match probability.

If it is truly appropriate to model the current population as split into such a large number of subpopulations, the above approach is correct. However, to use it as an approximation if there is actually a small number of subpopulations can lead to serious error. In this instance, we would expect a number of the members of the database to belong to the same subpopulation as the culprit. By assuming a large number of subpopulations we are nof fully utilising the information provided by the database.

For this reason it is essential that a method is developed which adequately deals with the case of a finite number of subpopulations.

9.3 Foreman, Evett and Smith

Unlike Roeder *et al.*, Foreman *et al.* [Foreman, Evett and Smith, 1997] concentrate mainly on the case involving a finite number η of subpopulations. It is important to do this, although omissions in the analysis mean that the data is not used to its full potential. The hierarchical model of Foreman *et al.* also displays an alternative structure to that employed in this thesis.

When comparing the work of Foreman *et al.* with that described in this thesis, it is important to be aware of differences in the definitions of some of the parameters used.

Throughout population genetics literature there are examples of similarly labelled parameters representing different quantities. Confusion is naturally caused when such parameters are treated as identical. The hierarchical model used in this thesis provides a basis for comparing such parameters, and highlighting their different roles in the model describing the generation of individual DNA profiles. Such comparisons are discussed further in Chapter 5.

Foreman *et al.* define θ_{lj} as a measure of 'genetic distance' at locus j between subpopulation $\mathcal{P}_l$ and the observed population as a whole. They define $\boldsymbol{\gamma}_j$ as $\sum_{l=1}^{\eta} \kappa(l)\boldsymbol{G}_{lj}$, the average of the actual subpopulation allele probabilities, weighted by the probabilities of belonging to each subpopulation, i.e. the vector of allele frequencies at locus j in the present population.

Defined as functions of the subpopulation allele probabilities $\boldsymbol{G}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ describe different quantities to the similarly labelled parameters of this thesis. In our model, (θ_j) and $(\boldsymbol{\gamma}_j)$ are parameters of the process which generates $(\boldsymbol{G}_{lj})$.

Similarly to the approach of this thesis, Foreman *et al.* model the subpopulation allele frequency vectors $(\boldsymbol{G}_{lj})$ as being generated independently with distribution,

$$\boldsymbol{G}_{lj}|\theta_{lj}, \boldsymbol{\gamma}_j \sim \text{Dirichlet}\left(\frac{1-\theta_{lj}}{\theta_{lj}}\gamma_j(1), \ldots, \frac{1-\theta_{lj}}{\theta_{lj}}\gamma_j(r_j)\right).$$

It is acknowledged by Foreman *et al.* that this independence assumption is false for finite η, and that further experimentation is required before the effects of the approximation can be dismissed as negligible. This lack of independence is a result of the alternative definition of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ as parameters at the observed subpopulation level rather than at the ancestral population level. It is clearly demonstrated by considering the case in which there are two observable subpopulations, $\gamma_{jm} = \kappa(1)G_{1j}(k) + \kappa(2)G_{2j}(k)$. Knowledge of $G_{1j}(k)$ in addition to γ_{jk} leads to knowledge of $G_{2j}(k)$ with absolute certainty, given that $\alpha(1)$ and

$\kappa(2)$ are known. Clearly,

$$G_{2j}(k)|\gamma_{jk} \not\sim G_{2j}(k)|G_{1j}(k), \gamma_{jk},$$

meaning that the definition of conditional independence is contravened.

The MCMC scheme of Chapter 6 of this thesis is based upon that of Foreman *et al.*. Using the test combined database of Caucasians and Afro-Caribbeans, despite the differences in the model, the MCMC scheme manages to identify the two 'subpopulations' well. Such a result with a very small number of subpopulations supports the argument that there is not a great effect caused by acceptance of the false independence assumption.

However, an omission that can certainly be seen to be significant is observed in the calculation of match probabilities. Two cases are considered:

(i) accused and offender belong to different subpopulations (either different subpopulations within the same racial group, or from distinct racial groups), i.e. $C \notin \mathcal{P}_s$;

(ii) accused and offender belong to the same subpopulation, i.e. $C \in \mathcal{P}_s$.

These cases are initially considered in isolation before a weighted average across all possible realisations of the offender's subpopulation is carried out to find the overall match probability.

Foreman *et al.* approach the two cases as follows, in each looking to find the match probability

$$\Pr(\boldsymbol{X}_C = \boldsymbol{y}|\boldsymbol{X}_s = \boldsymbol{y}, \chi_\delta = \xi_\delta, C \neq s),$$

where C labels the culprit, s labels the suspect, and χ_δ represents the collection of database profiles, excluding that of the suspect. The basic calculation carried out in both cases involves the following integration across the parameters, adjustments being made according to differing assumptions in cases (i) and (ii).

$$\begin{aligned}&\Pr(\boldsymbol{X}_C = \boldsymbol{y}|\boldsymbol{X}_s = \boldsymbol{y}, \chi_\delta = \xi_\delta, \mathcal{M}, C \neq s)\\ &= \int_{\boldsymbol{\gamma},\boldsymbol{\theta}} \Pr(\boldsymbol{X}_C = \boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\theta}, \chi_\alpha = \xi_\alpha, \mathcal{M}, C \neq s).p(\boldsymbol{\gamma}, \boldsymbol{\theta}|\chi_\alpha = \xi_\alpha, \mathcal{M})\mathrm{d}\boldsymbol{\gamma}\mathrm{d}\boldsymbol{\theta} \quad (64)\end{aligned}$$

where χ_α represents the collection of database profiles, **including** the suspect's profile, Dependent on whether case (i) or (ii) is being considered, $\mathcal{M}$ represents $C \notin \mathcal{P}_s$ or $C \in \mathcal{P}_s$. Given $(\boldsymbol{\gamma}, \boldsymbol{\theta})$, $\boldsymbol{X}_C$ is *not* independent of χ_δ and the database profiles should therefore be conditioned upon in the first term of the above integral. In both cases, the database is omitted from the first term within the integral, meaning that equation (64) becomes

$$\begin{aligned}&\Pr(\boldsymbol{X}_C = \boldsymbol{y} | \boldsymbol{X}_s = \boldsymbol{y}, \chi_\delta = \xi_\delta, C \neq s) \\ &\quad = \int_{\boldsymbol{\gamma},\boldsymbol{\theta}} \Pr(\boldsymbol{X}_C = \boldsymbol{y} | \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{X}_s = \boldsymbol{y}, \mathcal{M}, C \neq s).p(\boldsymbol{\gamma}, \boldsymbol{\theta} | \xi_\alpha, \mathcal{M}) \mathrm{d}\boldsymbol{\gamma} \mathrm{d}\boldsymbol{\theta}, \quad (65)\end{aligned}$$

i.e. the data is used to provide information upon the overall allele frequencies, and the degree of variation observed across subpopulations, but not upon the culprit's match probability.

The effect of this omission is examined within the following sections considering cases (i) and (ii) in turn.

9.3.1 $C \notin \mathcal{P}_s$

As Foreman *et al.* note, "when the accused and offender are taken to belong to different subpopulations, it may be assumed that they are genetically unrelated so that their profiles are independent of each other." It is further assumed that individuals mate at random within the offender's subpopulation $\mathcal{P}_C$. This means that the first term in equation (65) can be further simplified to

$$p(\boldsymbol{X}_C = \boldsymbol{y} | \boldsymbol{\gamma}, \mathcal{M}) = c \prod_{j=1}^{M} \gamma_j(y_{j1}) \gamma_j(y_{j2}),$$

a product of *population* allele frequencies, where $c = \prod_{j=1}^{M} 2^{h(y_{j1}, y_{j2})}$.

It is suggested that this can be calculated in two ways: via plug-in estimates, and by a full Bayesian analysis.

The plug-in estimates $\hat{\boldsymbol{\gamma}}$ are provided by the observed allele frequencies within the database of the racial group of the offender.

In the full analysis, it is assumed that $\boldsymbol{\gamma}$ has a prior distribution given by

$$\boldsymbol{\gamma}_j \sim \text{Dirichlet}(\omega_{j1}, \omega_{j2}, \ldots, \omega_{jr_j}).$$

The need to augment the database δ with the suspect s to form α is recognized, giving a posterior distribution,

$$\gamma_j|\chi_\alpha = \xi_\alpha \sim \text{Dirichlet}(\omega_{j1} + n_j(1), \omega_{j2} + n_j(2), \ldots, \omega_{jr_j} + n_j(r_j)),$$

where $n_j(m)$ is the number of alleles of type m observed at locus j in the database. The integration of equation (65) gives a match probability,

$$\begin{aligned}\Pr(\boldsymbol{X}_C = \boldsymbol{y}|\chi_\alpha = \xi_\alpha, \mathcal{M}, C \neq s) \\ = c\prod_{j=1}^{M} \frac{(\omega_j(y_{j1}) + n_j(y_{j1}))(\omega_j(y_{j2}) + n_j(y_{j2}) + \delta_j)}{(\omega_j(+) + n_j(+))(\omega_j(+) + n_j(+) + 1)},\end{aligned} \tag{66}$$

where δ_j indicates if $y_{j1} = y_{j2}$.

It should be noted that this approach takes no account of subpopulation structure within the racial group being considered. This approach is at odds with that advocated in this thesis. In the general case in which one does not know the subpopulation of the suspect, we consider the suspect as a member of the database with no greater weight than any other individual whose profile is known. Conditioning upon the suspect being from a different subpopulation to the culprit provides no justification for excluding the rest of the database from the conditioning of the match probability. If the number of subpopulations η is small, it is very likely that some of the other database individuals belong to the same subpopulation as the suspect.

9.3.2 $C \in \mathcal{P}_s$

In the case that the culprit and suspect belong to the same subpopulation, it is noted that ideally data from this subpopulation would be available to make inference. Empirical estimators of the subpopulation allele frequencies could then be multiplied across loci to give a match probability estimate similar to $\hat{m}_l^e$ described in Chapter 4.

As such information is not generally available, Foreman *et al.* employ a formula derived by Balding and Nichols [Balding and Nichols, 1994, Balding and Nichols, 1995] to correct for the fact that γ represents the allele

frequency of the racial group rather than the subpopulation. The match probability is then given by

$$\Pr(\boldsymbol{X}_C = \boldsymbol{y} | \boldsymbol{X}_s = \boldsymbol{y}, C \in P_s, \boldsymbol{\gamma}, \boldsymbol{\theta}) = c \prod_{j=1}^{M} \Pr(\boldsymbol{X}_{Cj} = \boldsymbol{y}_j | \boldsymbol{X}_s = \boldsymbol{y}, C \in P_s, \boldsymbol{\gamma}, \boldsymbol{\theta}),$$

where

$$\Pr(\boldsymbol{X}_{Cj} = \boldsymbol{y}_j | \boldsymbol{X}_s = \boldsymbol{y}, C \in P_s, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \begin{cases} \frac{\{2\theta_{Psj}+(1-\theta_{Psj})\gamma_j(y_{j1})\}\{3\theta_{Psj}+(1-\theta_{Psj})\gamma_j(y_{j2})\}}{(1+\theta_{Psj})(1+2\theta_{Psj})} & \text{if } y_{j1} = y_{j2}, \\ \frac{2\{\theta_{Psj}+(1-\theta_{Psj})\gamma_j(y_{j1})\}\{\theta_{Psj}+(1-\theta_{Psj})\gamma_j(y_{j2})\}}{(1+\theta_{Psj})(1+2\theta_{Psj})} & \text{if } y_{j1} \neq y_{j2}. \end{cases} \quad (67)$$

This is a similar match probability equation to that employed by Roeder *et al.* in the Affinal case. In the case of Roeder *et al.* it is the result of modelling the population as consisting of a large number of subpopulations meaning that no two database individuals belong to the same subpopulation. Foreman *et al.* arrive at this equation by not conditioning on the full database, only the suspect.

While the equation is correct under the assumptions of Roeder *et al.*, when there is a finite number of subpopulations there is a significant probability that other members of the database are in the same subpopulation as the culprit. This means that the match probability expression employed by Foreman *et al.* is theoretically incorrect.

9.4 Discussion

The reduction in complexity introduced by assuming a large number of subpopulations, as suggested by Roeder *et al.* is clearly helpful. However, such an assumption would not appear generally applicable, meaning that a method should be devised for the case of a finite number η of subpopulations. Even if this number cannot initially be specified, there are a number of methods suggested (Chapter 10) for introducing η as a variable, each of which relies upon a clearly defined method for dealing with the case of a known finite number of subpopulations. For this reason we concentrate in this section upon a comparison between the methods of Foreman *et al.* and this thesis, both of which look to tackle this problem.

As well as presenting a model with a clearer hierachical structure in which conditional independence of subpopulations given overall allele frequencies and subpopulation differentiation parameters is implicit rather than assumed, the methods of this thesis use the information provided by clustering within the database to provide correct posterior match probabilities.

The following results demonstrate the value of calculating a match probability conditional upon the whole database in the case of the combined Caucasian/Afro-Caribbean database. It is true that this database represents an extreme case, with a greater degree of heterogeneity than would generally be observed within a single racial group. However, the effect upon match probabilities is great enough to suggest that there are cases in which a lack of conditioning upon the database profiles in the match probability could have a significant effect upon the court's decision.

The plots of Figures 25 to 34 are presented to allow comparison of the accuracy of match probability estimates under the various methods discussed. For each individual in the group

The plots of Figures 25 and 26 are presented to allow comparison of the accuracy of match probability estimates under the methods of Foreman *et al.* and this thesis. They show log likelihood ratios, where the likelihood ratio is $\frac{1}{m_l}$, m_l being the match probability for the subpopulation under consideration. This match probability is calculated for each database individual in three ways,

(i) assuming that the culprit belongs to the same subpopulation as the suspect and that this subpopulation $\mathcal{P}_s$ is known, Foreman *et al.* propose the estimator

$$\hat{m}_l = \frac{1}{r-m} \sum_{t=m+1}^{r} c \prod_{j=1}^{r_j} \frac{(a_j^{(t)}(y_{j1}) + 1 + \delta_j)(a_j^{(t)}(y_{j2}) + 1 + 2\delta_j)}{(a_j^{(t)}(+) + 2)(a_j^{(t)}(+) + 3)},$$

where δ_j indicates if $y_{j1} = y_{j2}$, and (t) labels the iteration of the Markov chain.

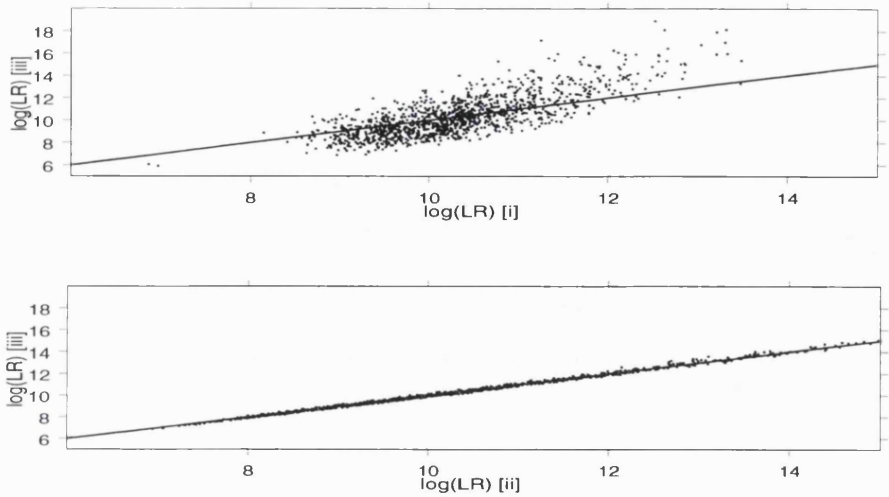


Figure 25: Comparison of log likelihood ratios for Caucasian individuals. These plot log likelihood ratios under method (iii) described in the text against those of methods (i) and (ii) respectively. The calculations are repeated for each profile within the Caucasian database, each point corresponding to one of these profiles. Ideally, the points would lie along the line "$x = y$" indicating that the match probability estimates in the absence of subpopulation information are equal to the accurate match probability estimates when subpopulation information is available. It can be seen that the estimates using the methods of this thesis are far more accurate than those of Foreman *et al.*

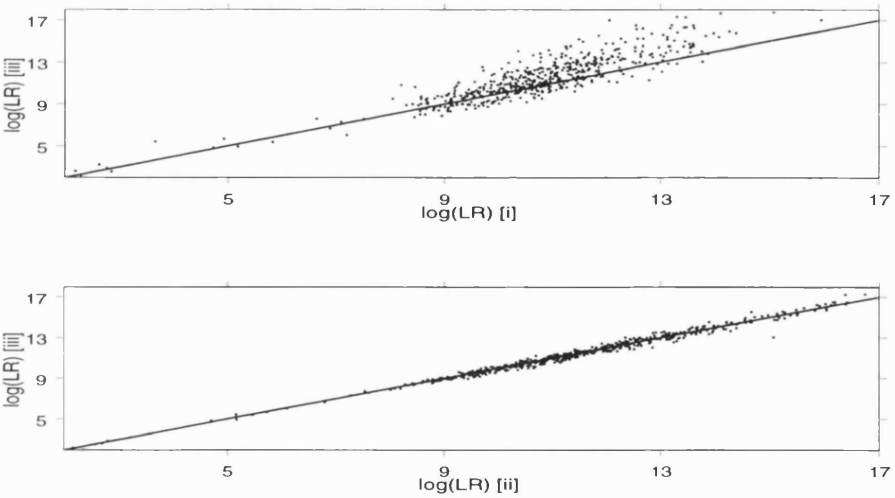


Figure 26: Comparison of log likelihood ratios for Afro-Caribbean individuals.

(ii) In this thesis we suggest the use of the estimator,

$$\hat{m}_l = \frac{1}{r-m} \sum_{t=m+1}^{r} c \prod_{j=1}^{r_j} \frac{(a_j^{(t)}(y_{j1}) + n_{lj}^{(t)}(y_{j1}))(a_j^{(t)}(y_{j2}) + n_{lj}^{(t)}(y_{j2}) + \delta_j)}{(a_j^{(t)}(+) + n_{lj}^{(t)}(+))(a_j^{(t)}(+) + n_{lj}^{(t)}(+) + 1)}$$

(iii) If subpopulation membership of all database individuals can be identified, we advocate the use of a product of empirical subpopulation frequencies,

$$\hat{m}_l^e = c \prod_{j=1}^{M} \hat{G}_{lj}(y_{j1}) \hat{G}_{lj}(y_{j2}), \tag{68}$$

to calculate the match probability for each subpopulation. Using our mixed database, results using this method can be considered a standard against which those of methods (i) and (ii) can be measured.

When calculating the match probabilities under the method of Foreman *et al.*, we have used the values of (θ_j) from our running of the MCMC scheme. Ideally we would use subpopulation-specific values (θ_{lj}) to make the comparison. However the graphs calculated using method (i) are very similar to those displayed by Foreman *et al.* in their response to the discussion of their paper [Foreman, Evett and Smith, 1997].

The comparison of Figures 25 and 26 indicates the greatly increased accuracy in match probability estimates provided by the methods of this thesis. However, to fully appreciate the practical effect of the different methods in the absence of subpopulation information we need to compare the resultant posterior probabilities of guilt.

The posterior probability of guilt of a suspect is calculated by combining the overall match probability m with the prior probability of guilt of the suspect π_s,

$$\Pr(C = s | \boldsymbol{X}_s = \boldsymbol{y}, \boldsymbol{X}_C = \boldsymbol{y}, \chi_\alpha = \xi_\alpha, \varepsilon) = \frac{\pi_s}{\pi_s + (1 - \pi_s)m}.$$

Maintaining the labels (i) - (iii), the overall match probability m is calculated as follows:

(i) Foreman *et al.* specify two match probability formulae, one for the case in which suspect and culprit belong to the same subpopulation, and one for that in which they belong to different subpopulations.

Assuming that we cannot identify the subpopulation to which either the culprit or suspect belong, the overall match probability can be broken down into the following sum across combinations of culprit and suspect subpopulations,

$$\begin{aligned} m &= \Pr(\boldsymbol{X}_C = \boldsymbol{y} | \boldsymbol{X}_s = \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &= \sum_{(l,r)} \Pr(\boldsymbol{X}_C = \boldsymbol{y} | C \in P_l, S \in P_r, \boldsymbol{X}_s = \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\qquad \times \Pr(C \in P_l, S \in P_r | \boldsymbol{X}_s = \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}). \end{aligned}$$

If $l = r$, $\Pr(\boldsymbol{X}_C = \boldsymbol{y} | C \in P_l, S \in P_r, \boldsymbol{X}_s = \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ is given by the expression of equation (67). Otherwise, it is given by the product of allele frequencies given in equation (66). For the purpose of this comparison, we assume that the subpopulations to which culprit and suspect belong are independent, meaning that $\Pr(C \in P_l, S \in P_r | \boldsymbol{X}_s = \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \kappa(l)\kappa(r)$, where $\kappa(l)$ is the prior probability of a randomly chosen individual being a member of subpopulation $\mathcal{P}_l$. As has been shown, the subpopulation differentiation parameters (θ_j) calculated under our model are not the same as those (θ_{lj}) of Foreman *et al.* In this instance, we present results under the above calculation substituting three different values of (θ_j). These range from the prior mean, 0.0291, to $\theta_{lj} = 0$, which is equivalent to ignoring any population substructure.

(ii, iii) In both instances, the overall match probability is calculated by taking the weighted average,

$$m = \sum_{l=1}^{\eta} \kappa(l) m_l.$$

In both cases, the subpopulation match probabilities are calculated as for the previous comparison (of log likelihoods), conditional upon the entire database.

Figures 27 - 34 show comparisons of posterior probabilities of guilt under the methods (i) - (iii). These have been calculated separately assuming two possible prior probabilities of guilt of the suspect, $\pi_s = 10^{-4}$ and $\pi_s = 10^{-6}$.

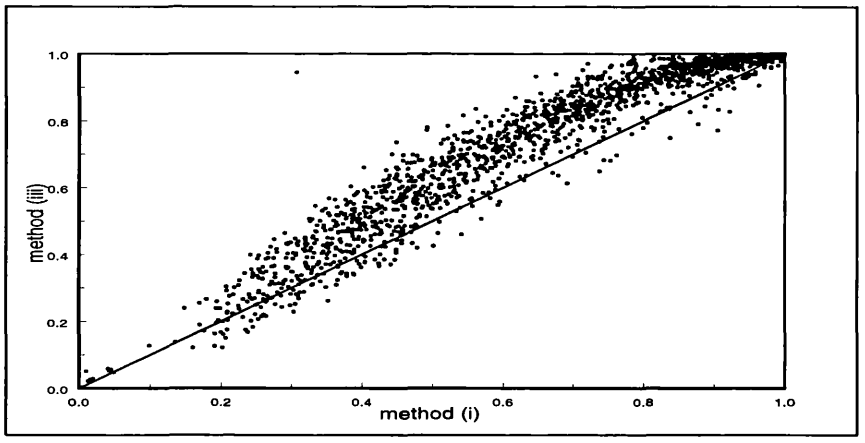


Figure 27: Comparison of posterior probabilities of guilt, $\pi_s = 10^{-4}$, $\theta_j = 0.0291$. This compares the posterior probability of guilt under the method of Foreman *et al.*, substituting the prior mean of (θ_j) into the appropriate match probability equation, to the posterior probability assuming subpopulation information available. A calculation is made assuming a match for each mixed database profile in turn resulting in a point for each database individual. The points displaying posterior probabilities of guilt close to 0 correspond to partial profiles. A match at a smaller number of loci will generally carry a small weight against the suspect.

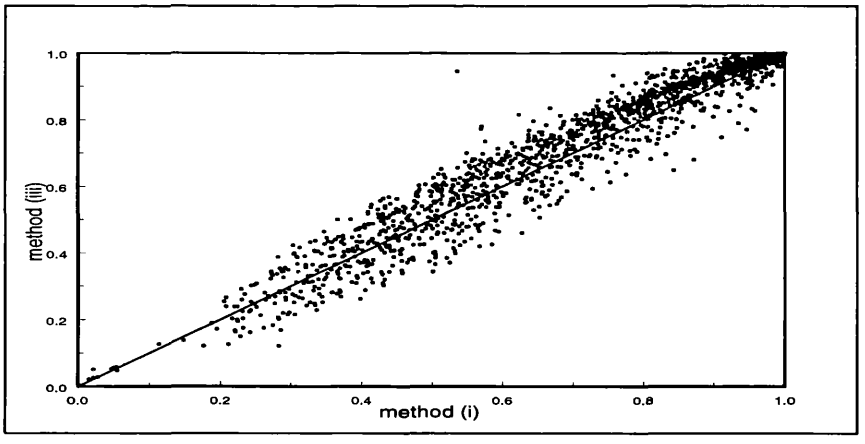


Figure 28: Comparison of posterior probabilities of guilt, $\pi_s = 10^{-4}$, $\theta_j = 0.01$.

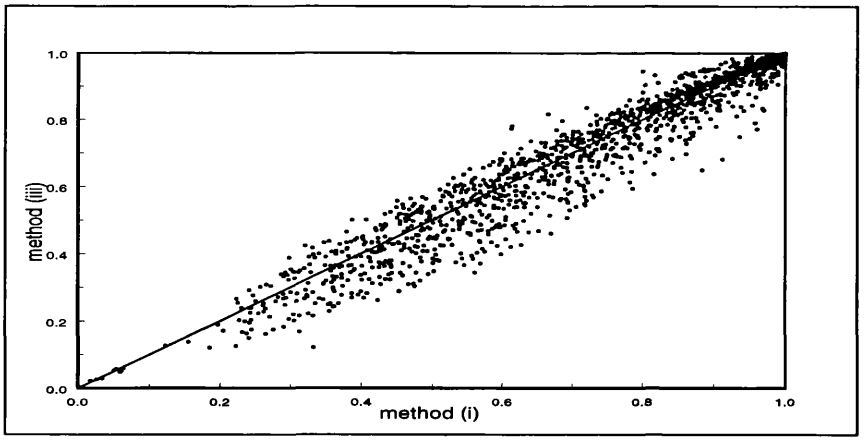


Figure 29: Comparison of posterior probabilities of guilt, $\pi_s = 10^{-4}$, $\theta_j = 0.0$. This compares the posterior probability of guilt assuming the entire population to be in Hardy-Weinberg equilibrium (i.e. under the product rule, as described in Chapter 2) to that assuming subpopulation information. One can see that there are a large number of points under the "$x = y$" line in contrast to Figures 27 and 28. This indicates that the product rule is often anti-conservative, prejudicing against the suspect by providing a posterior probability of guilt which is too high.

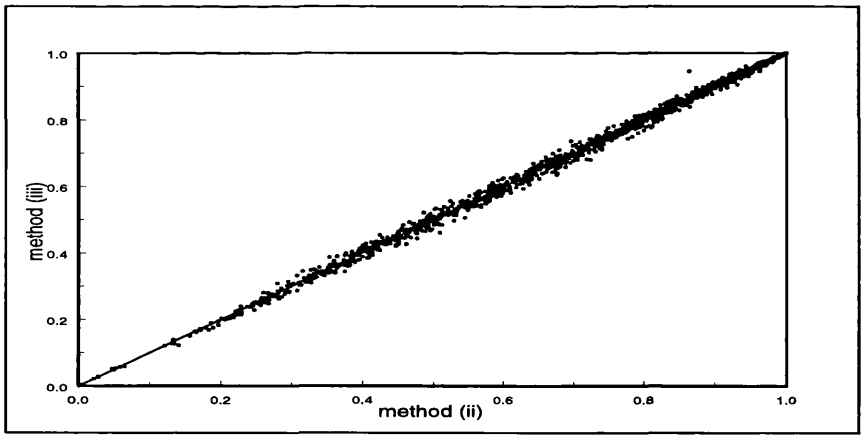


Figure 30: Comparison of posterior probabilities of guilt, $\pi_s = 10^{-4}$. This compares the method of this thesis assuming unknown subpopulation information to the accurate estimates assuming known subpopulation information. It can be seen that the points lie very close to the line "$x = y$", indicating that the posterior probabilities of guilt under this method are more accurate than those under the method of Foreman *et al.*

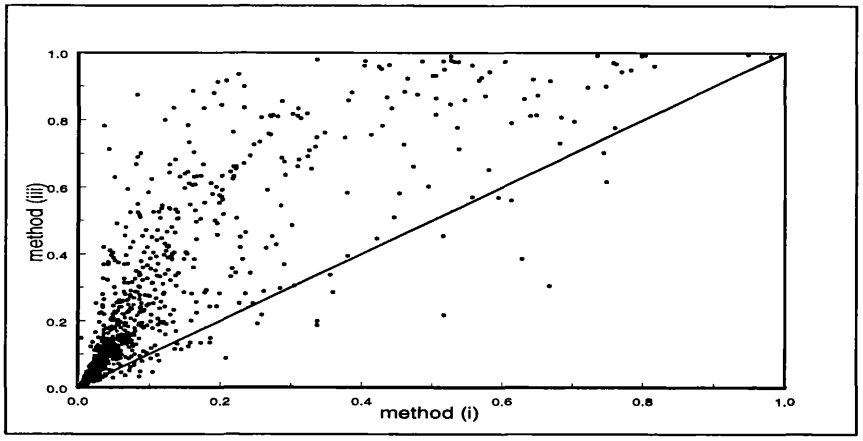


Figure 31: Comparison of posterior probabilities of guilt, $\pi_s = 10^{-6}$, $\theta_j = 0.0291$. This shows a similar comparison to Figure 27, but with a smaller prior probability of guilt of the suspect. This prior probability of 1 in 1 million is still a realistic figure should we be considering, for example, a case in which the crime has been committed in a city and we have very little information to narrow down our possible culprit population. It can be seen that under our "gold standard" method (iii), a number of individuals still have posterior probabilities of guilt close to 1, but that in many cases these are seriously underestimated by method (i).

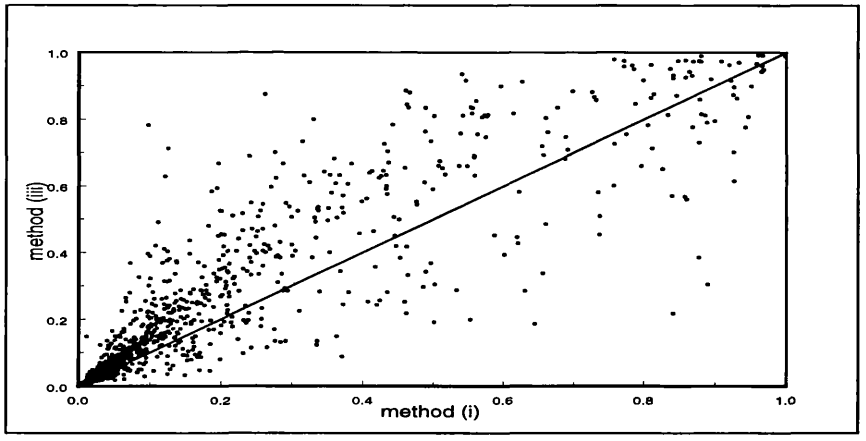


Figure 32: Comparison of posterior probabilities of guilt, $\pi_s = 10^{-6}$, $\theta_j = 0.01$.

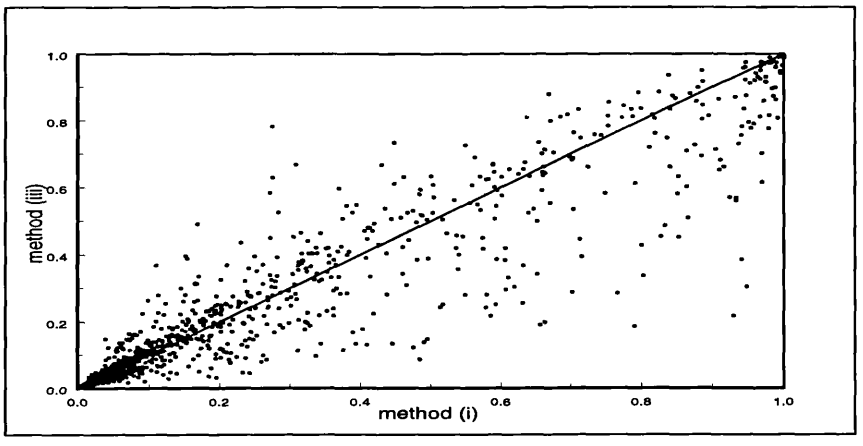


Figure 33: Comparison of posterior probabilities of guilt, $\pi_s = 10^{-6}$, $\theta_j = 0.0$.

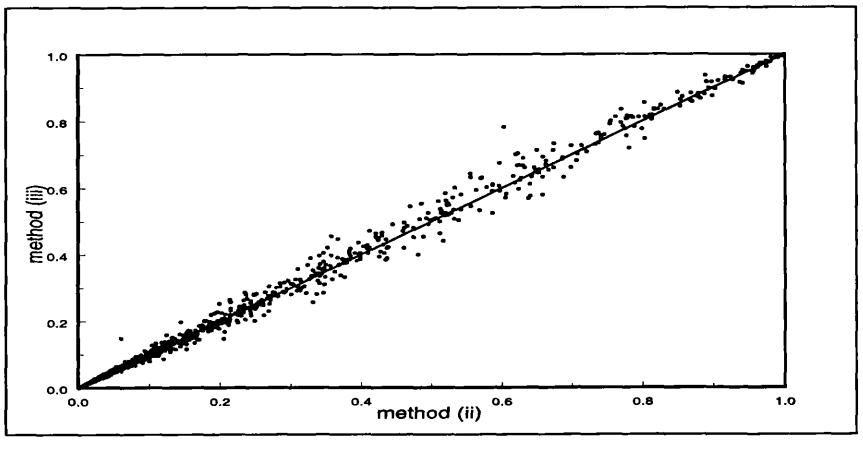


Figure 34: Comparison of posterior probabilities of guilt, $\pi_s = 10^{-6}$. Again we see that the method of this thesis provides a more accurate posterior probability of guilt.

As indicated by Foreman *et al.* [Foreman, Evett and Smith, 1997], match probabilities under their method (i) are generally conservative, giving a posterior probability of guilt less than that in which substructure is ignored ($\theta_{lj} = 0$). However, even if we were to integrate over the posterior densities of (θ_{lj}) presented by Foreman *et al.*, the resultant posterior probabilities cannot be as close to the true probabilities as those of our method (ii).

Although it is very important to avoid prejudicing against the defendant, a large conservative error, possibly resulting in the acquittal of a guilty party, would also seem undesirable. For this reason, we suggest that conditioning upon the full database in the match probability is very important, and that its omission is of practical, and not just theoretical, concern.

10 Future work

In this thesis we have considered how to calculate a posterior probability p_{guilt} of guilt for a suspect whose DNA profile $\boldsymbol{X}_s$ (observed to be $\boldsymbol{y}$) matches that of the culprit C. This requires the calculation of match probabilities for individuals i (outside the database α) in each subpopulation $(\mathcal{P}_l; l = 1, \ldots, \eta)$:

$$m_l = \Pr(\boldsymbol{X}_i = \boldsymbol{y} | \chi_\alpha = \xi_\alpha).$$

These must then be combined with the prior probability π_s of the suspect having committed the crime and $(\lambda_l = \Pr(C \in \mathcal{P}_l | C \notin \alpha))$:

$$p_{guilt} = \frac{\pi_s}{\sum_{i \in \beta} \pi_i + \Pr(C \notin \alpha) \sum_{l=1}^{\eta} \lambda_l m_l},$$

where β is the set of individuals in the database α whose profile matches that of the suspect.

Calculation of these match probabilities is based upon the hierarchical model described in Chapter 3. The parameters of this model have been clearly defined, justifying the conditional independence properties assumed at each level. As the match probability calculations are found to be impossible analytically, we use MCMC methods to obtain estimates. In this thesis we have extended the work of Dawid and Pueschel [Dawid and Pueschel, 1999], in particular showing that it is important to condition upon available data even in the absence of information concerning individual subpopulation membership.

However, there is still scope for future work, particularly in the following areas:

(i) as the subpopulations are not clearly defined, it is not generally reasonable to assume the number η of subpopulations known. This means that we should consider η as a random variable. Two ways of approaching this problem are considered in Section 10.1. Thus far, neither has been found to be satisfactory, meaning that future research is required, either to appropriately adjust these methods, or to find an original method.

(ii) There is much debate as to whether or not Bayesian methods are appropriate for the presentation of DNA profile match evidence in the courtroom. It is our opinion that Bayesian methods represent the best way to combine the scientific evidence of the profile match with other evidence to reach a decision with regard to the conviction/acquittal of the suspect. We discuss this further in Section 10.3.

(iii) While employment of the hierarchical model used in this thesis represents a great advance from the assumption of random mating throughout entire populations, it is still a relatively rough approximation to the true situation. In Section 10.6 we consider the need for adjustments to the current model.

10.1 Variable number of subpopulations

Reversible jump MCMC is a technique first proposed by Green [Green, 1995]. It is designed to tackle problems in which the number v of variables is itself a variable. An application of reversible jump MCMC involves the generation of a Markov chain, as in conventional MCMC, but with an additional step. This step proposes, at each iteration, a change in v that is accepted with some probability dependent upon the current and proposed values.

In this instance, a possible MCMC scheme would proceed in a similar manner to that proposed by Green in his response to Foreman *et al* [Foreman, Evett and Smith, 1997].

At each iteration we consider either a split of an existing subpopulation into two, or the combination of two existing subpopulations into one. Thus, at each iteration, the number η of subpopulations will increase or decrease by one if the proposal is accepted, or remain the same if the proposed move is rejected. A possible method for generating proposed values for the variables corresponding to the newly generated subpopulations is:

(i) split:

Select the subpopulation to be split randomly, and label this subpopulation by s. The new subpopulations are labelled s_- and s_+.

Generate $U \sim \text{Uniform}(0,1)$: $\kappa(s_-) = u \times \kappa(s)$, $\kappa(s_+) = (1-u) \times \kappa(s)$.

Generate $\boldsymbol{G}_{lj} \sim \text{Dirichlet}(C(\boldsymbol{a}_j + \boldsymbol{n}_{sj}))$, where $l = s_-, s_+$ and $n_{sj}(k)$ is the number of alleles of type k at locus ($j = 1, \ldots, M$) in individuals allocated to subpopulation s.

Allocate individuals originally in subpopulation s to s_- and s_+ randomly using the full conditional probability distribution of the subpopulation labels (I_i).

(ii) Combine:

Randomly select two subpopulations, s_- and s_+, to be combined to form a single subpopulation s:

$\kappa(s) = \kappa(s_-) + \kappa(s_+)$.

All individuals originally allocated to the two proposed subpopulations will be allocated to the new subpopulation s.

Generate $\boldsymbol{G}_{sj} \sim \text{Dirichlet}(C(\boldsymbol{a}_j + \boldsymbol{n}_{sj}))$.

The constant C is introduced as a means of adjusting the scheme to achieve a reasonable acceptance rate. In practice it is very difficult to do this due to the high dimension of the parameter space.

10.2 Pritchard *et al* approximation

Pritchard *et al* are faced with a similar clustering problem to that of this thesis.

The "*ad hoc*" approach they follow can be applied to the problem here to estimate

$$\Pr(\eta|\chi_\alpha = \xi_\alpha) \propto \Pr(\chi_\alpha = \xi_\alpha|\eta).\pi(\eta). \tag{69}$$

There are a number of possible choices for the prior distribution of η, for example uniform up to some maximum η_{max}, or a truncated Poisson distribution.

Calculation of the likelihood $\Pr(\chi_\alpha = \xi_\alpha|\eta)$ is where the difficulty lies.

The Pritchard *et al* approach begins by considering the Bayesian deviance,

$$D_\eta(\boldsymbol{G}, \boldsymbol{I}) = -2\log \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}, \boldsymbol{I}, \eta).$$

Assuming that the conditional distribution of D is Normal,

$$\Pr(\chi_\alpha = \xi_\alpha) \approx \exp(-\frac{\mu}{2} - \frac{\sigma^2}{8}),$$

where

$$\mu = \mathrm{E}[D_\eta(\boldsymbol{G}, \boldsymbol{I})|\chi_\alpha = \xi_\alpha, \eta]$$

$$\sigma^2 = \mathrm{Var}(D_\eta(\boldsymbol{G}, \boldsymbol{I})|\chi_\alpha = \xi_\alpha, \eta).$$

The conditional mean μ and variance σ^2 of D can be estimated, using the results of the MCMC scheme, by

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{M} -2\log \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}^{(t)}, \boldsymbol{I}^{(t)}, \eta)$$

and

$$\hat{\sigma^2} = \frac{1}{T}\sum_{t=1}^{M}(-2\log \Pr(\chi_\alpha = \xi_\alpha|\boldsymbol{G}^{(t)}, \boldsymbol{I}^{(t)}, \eta) - \hat{\mu})^2.$$

Using the estimates $\hat{\mu}$ and $\hat{\sigma^2}$, $\Pr(\chi_\alpha = \xi_\alpha|\eta)$ can be estimated leading, by substitution into (69), to an estimate of $\Pr(\eta|\chi_\alpha = \xi_\alpha)$.

Pritchard *et al* show that this method produces acceptable results when assuming a small number of potential subpopulations. If η_{max} is large, we must generate a Markov chain for each possible number η of subpopulations. Thus far we have not managed to write a program that manages this in an acceptable time.

10.3 Application of Bayesian methods in the courtroom

There is much debate as to the role Bayesian methods should play in the courtroom. The argument in favour has not been helped by the frequent use of two major errors:

(i) the prosecutor's fallacy. This is an error of transposing the conditional, confusing Pr(*innocence*|*evidence*) with Pr(*evidence*|*innocence*). This would lead to a match probability of 0.000001 being interpreted as a probability of guilt of 0.999999.

(ii) The defendant's fallacy. This involves using the frequency in the population to argue that the defendant is only one of a large number of possible suspects. The error involves assuming equal prior probabilities of guilt for all individuals. For example, ignoring for a moment the complications of population substructure, if a particular profile has a frequency of 1 in 100,000, one would expect there to be 70 individuals with the profile in a city of 7 million. However, the correct method must use the profile frequency to update the suspect's prior probability of guilt using Bayes' theorem, as described in Chapter 1. If there is a great deal of non-DNA evidence against the suspect his prior probability of guilt will be much greater than 1 in 7 million, and therefore his posterior probability of guilt will be much larger than 1 in 70.

10.4 Presentation of the evidence

The use of methods to correctly utilise DNA evidence requires a greater acceptance of Bayesian methods in the courtroom. To date, in the UK in particular, there has been a marked reluctance to rely on anything other than the common sense of the jurors to evaluate the weight of such scientific evidence. By the mid-90s, some wariness of the use of DNA evidence had developed, a number of convictions having been overturned on appeal as a result of the application of the prosecutor's fallacy (e.g. *R. v. Deen, The Times*, 10 January 1994).

A further blow was dealt to those advocating the use of such methods when the conviction of Dennis Adams was overturned in 1996 (*R. v. D. Adams* [1996] 2 Cr App Rep 467). Adams' DNA profile had been found to match that of a rape sample while most of the other evidence pointed towards Adams' innocence. The defence, concerned that the jury would be overwhelmed by the apparent

strength of the DNA match in suggesting Adams' guilt, introduced a statistician (Peter Donnelly) to explain how all the evidence could be combined using the Bayesian arguments outlined in Chapter sec:intro.

Adams was nevertheless found guilty, but an appeal was successful. Regarding the presentation of Bayes' Theorem as a means of combining different pieces of evidence, the Court of Appeal stated that it had "very grave doubts as to whether that evidence was properly admissible, because it trespasses on an area peculiarly and exclusively within the province of the jury, namely the way in which they evaluate the relationship between one piece of evidence and another...Jurors evaluate evidence and reach a conclusion not by means of a formula, mathematical or otherwise, but by the joint application of their individual common sense and knowledge of the world to the evidence before them." The Appeal Court concluded by stating that "If, as seems entirely possible, the jury abandoned the struggle to understand and apply Bayes, they were left by the summing-up with no other sufficient guidance as to how to evaluate the prosecution case (based as it was entirely on the DNA evidence), in the light of the other non-DNA evidence in the case. This means that their verdict cannot be considered as safe."

A retrial took place in which a questionnaire was prepared for the jury to use. This asked a number of questions to assess the weight placed by the jury upon various aspects of the non-DNA evidence. This was then combined in the appropriate formula to produce a posterior odds ratio. Adams was again convicted and an appeal was dismissed. The Appeal Court ruled that "... expert evidence should not be admitted to induce juries to attach mathematical values to probabilities arising from non-scientific evidence adduced at the trial."

Statements such as those made by the Appeal Court highlight the barriers facing those arguing for the application of Bayesian philosophy in the courtroom.

With regard to the presentation of DNA evidence in court, we concur with the opinions expressed by Evett and Weir [Weir and Evett, 1998]. In order to fully utilise the power of technology such as DNA profiling, it is vital that courtroom

practice also advances.

It is not satisfactory to simply present a match probability, explain its meaning, and then allow the resultant weight of evidence to be used subjectively. Odds are a part of everyday language, and we are used to attaching subjective odds reflecting our feelings regarding the outcome of some event. For this reason it seems reasonable to think that members of the jury would generally be comfortable with the concept of placing odds upon the guilt of a suspect given the evidence presented.

Assuming this to be so, the next step in a presentation to the jury would be a statement of the effect that the match probability in question would have upon these prior odds. With this information it is possible for jurors to update their 'prior' beliefs using Bayesian methods. The Bayesian train of thought which suggests using additional data to update prior beliefs is a logical one confirming to natural intuition. If more concentration is focused upon explaining this logic to jurors rather than upon the underlying mathematics, it may become clearer that Bayesian methods are entirely consistent with intuition.

The actual presentation of the effect of a particular likelihood ratio could be achieved in a number of ways. To avoid the appearance of suggesting a particular prior, the posterior odds ratio resulting from a a range of priors could be demonstrated.

At the moment however, it seems that the greatest challenge is convincing courts steeped in traditional methods of the value in accepting what is to them a new philosophy for combining pieces of evidence.

10.5 A critical analysis

This thesis clearly provides scope for further research, even before considering the problem of an unknown number of subpopulations. A number of further questions can be raised and these are considered in this section.

(i) Should we use subpopulation specific heterogeneity parameters?

- The model of Foreman *et al* [Foreman, Evett and Smith, 1997] orders the subpopulation specific differentiation parameters so that the smallest subpopulations are associated with the largest θ_l values. This is consistent with the thinking of Weir and Cockerham [Weir and Cockerham, 1984] who define subpopulation differentiation parameters at the same level as this thesis. In their paper they concentrate on a situation in which subpopulation sizes are assumed equal, but state that if subpopulation sizes are unequal these parameters will also vary. In retrospect this seems reasonable as the allele probabilities (G_{lj}) of a small isolated subpopulation are more likely to drift further away from the ancestral values than those of a large subpopulation.

Whilst acknowledging that a more accurate model in terms of reflecting the development of subpopulations from a single ancestral population may include subpopulation specific differentiation parameters (as outlined in Chapter 3), their adoption would lead to a more complicated model and a large increase in the running time of any MCMC scheme, particularly when assuming a large number of subpopulations. It is therefore important to establish the necessity of any complication of the model.

It is important to realise that any model is merely an approximation to the true situation. Indeed, it has been suggested [Donnelly, 1997] that the Dirichlet model itself "lacks a sound theoretical justification", and further tests should be made to assess its appropriateness for approximating the population genetics present.

It should be noted that the philosophy of this thesis is applicable however the model is specified. We have carefully defined levels of our model corresponding to those of the population structure. In doing this we present a framework upon which we can derive conditional densities whichever distributions are considered most suitable.

One of the main messages of this thesis is the necessity to condition upon the data at all stages of match probability calculation, even when subpop-

ulation membership is unknown. In doing this, particularly if the database is sizeable, the effect of any inaccuracy in the estimation of θ is greatly reduced when compared to previously published methods. This is particularly relevant when considering the posterior distribution of subpopulation allele probabilities,

$$\boldsymbol{G}_{lj}|\epsilon_\alpha, \theta_j, \boldsymbol{\gamma}_j \sim \text{Dirichlet}\left(\frac{1-\theta_j}{\theta_j}\boldsymbol{\gamma}_j + \boldsymbol{n}_j\right) \tag{70}$$

and also in the evaluation of the match probability,

$$m_l = \mathrm{E}\left[\prod_{j=1}^{M} 2^{h(y_{j1},y_{j2})} \frac{(a_j(y_{j1}) + n_{lj}(y_{j1}))(a_j(y_{j2}) + n_{lj}(y_{j2}) + \delta_j)}{(a_j(+) + n_{lj}(+))(a_j(+) + n_{lj}(+) + 1)} \right. \\ \left. |i \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha\right].$$

It should also be noted that even with θ constant across subpopulations, the posterior distribution 70 displays a greater variance for smaller subpopulations due to the influence of the database numbers. This corresponds to the thinking of Weir and Cockerham.

It should be stressed that it is not our aim to analyse the accuracy of existing models with regard to population genetics, rather to show the way in which we believe existing models, and indeed any newly developed models, should be used when making forensic inference. We have presented a clear framework into which extensions/adjustments to the model can be incorporated. However, it is our opinion that introducing subpopulation specific differentiation parameters will not significantly affect match probability estimates due to the increased influence of the data in the posterior distributions and estimates involved.

(ii) Can we ensure that this method is conservative?

- The more accurate an answer is, the less important is the need to adjust to ensure conservativeness. While it would seem somewhat inefficient, it may be possible to conduct the analysis under the assumption of a number of different values for the number η of subpopulations, selecting the largest (and most conservative) total match probability for use in court.

Another way of ensuring that the calculated match probability is conservative is by building prior dependence between the culprit subpopulation I_C and the suspect subpopulation I_s, or even assigning a prior probability of 1 to the event that culprit and suspect come from the same subpopulation. We introduce a joint distribution between the subpopulations of the culprit and suspect

$$\boldsymbol{Q} = (q_{lm}; l = 1, \ldots, \eta; m = 1, \ldots, \eta) \text{where } q_{lm} = \Pr(C \in \mathcal{P}_l, s \in \mathcal{P}_m).$$

To build in the desired positive dependence, Dawid [Dawid, 1996a] suggests the use of a joint distribution of the form

$$q_{lm} = (1 - \lambda_+)\kappa(l)s(m) + \lambda_l \delta_{lm}$$

where $(s(m))$ is an arbitrary probability distribution placed upon the suspect's subpopulation, and δ_{lm} indicates if $l = m$. The parameters (λ_l) can be adjusted according to the level of dependence required. To ensure that the culprit and suspect come from the same subpopulation, for example, we set $(\lambda_l = \kappa_l; l = 1, \ldots, \eta)$. Under this form of the joint distribution, the conditional distribution of the culprit's subpopulation I_C given that $(I_s = m, I_C \neq m)$ is the same as that if it is assumed that the subpopulations of culprit and suspect are independent.

If we decide to build in this dependence, it is important to remember that the strict definition of the match probability is given by

$$m_l = Pr(\boldsymbol{X}_C = \boldsymbol{y} | C \in \mathcal{P}_l, \chi_\alpha = \xi_\alpha, \varepsilon)$$

We now include in the non-DNA evidence ε the parameters (λ_l) and probabilities $(s(l))$. This means that, if we include the suspect profile in the database when running the (adjusted) MCMC scheme, we should run it η times, as the information upon the suspect's subpopulation will change each time.

10.6 Adjusting the model

.

In this thesis we describe a model based upon those currently in use, clearly defining parameters at all levels. In doing this, we aim to clear confusion caused by the vague definition of parameters in other papers. Before seeking to advance the model, it was felt essential to establish the correct way to make inference using this model.

The hierarchical model used in this thesis is a clear simplification of the true situation, involving as it does discrete subpopulations between which individuals cannot mate, and within which mating is random. However, the major aim when designing any mathematical model is not necessarily to make it as realistic as possible. Rather, we wish to design a model which allows us to easily estimate quantities of interest with acceptable accuracy. This means that a model which reflects the true state of nature more accurately but does not provide any greater accuracy in estimates is not, for our purposes, an improvement if it makes the calculation of these estimates more difficult.

However, it is desirable to gauge the effect of advancing the model. An initial step could involve the introduction of *admixture*. This is incorporated into the model of Pritchard *et al* [Pritchard, Stephens and Donnelly, 2000] and allows an individual's profile to be 'shared' between more than one subpopulation.

10.7 Summary

It is hoped that this thesis contains a number of points of value to those seeking to advance the use of DNA profiling in the courtroom. This thesis builds upon the work of Dawid and Pueschel [Dawid and Pueschel, 1999] which presents what we consider to be the correct basis for using DNA profiling data when estimating a match probability in the presence of subpopuation identifiers. By extending this work to the case in which subpopulation labels are absent, we show how this theory can be applied to a more realistic situation.

It is also felt that there are some points of more general value. Parameters

of subpopulation differentiation have often been confused. The analysis of some of these parameters in the context of the hierarchical model should lend some lucidity.

We have also encountered a number of interesting MCMC mixing problems. It is satisfying to see techniques such as simulated tempering and importance sampling applied, and it is to be hoped that the ways in which they have been used here are also of interest.

It will be interesting to see how the use of DNA profiling data proceeds over the coming years. It is likely that in the future the number of loci over which profiles are defined will increase, increasing the weight of evidence of a profile match to the extent that it will be conclusive proof of guilt. This should not however detract from the value of this work. Bayesian hierarchical models have a wide variety of applications and it is hoped that this example can be of use to those seeking to use similar methods in other areas.

References

Balding, D. J. and Nichols, R. A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* **64**, 125–140.

Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12.

Cavalli-Sforza, L. L., Menozzi, P. and Piazza, A. (1994). *The History and Geography of Human Genes.* Princeton: Princeton University Press.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems.* New York: Springer-Verlag.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo convergence diagnostic: a comparitive review. *J. Am. Statist. Ass.*, **91**, 883–904.

Dawid, A. P. (1985). Probability, symmetry and frequency. *Brit. J. Phil. Sci.* **36**, 107–128.

Dawid, A. P. (1986). A Bayesian view of statistical modelling. In *Bayesian Inference and Decision Techniques* (eds. P. K. Goel and A. Zellner), pp. 391–404. Amsterdam: Elsevier.

Dawid, A. P. (1996). Some thoughts on DNA identification using heterogeneous data-bases. Manuscript dated August 16, 1996. 14pp.

Dawid, A. P. (1997). Modelling issues in forensic inference. In *1997 ASA Proceedings, Section on Bayesian Statistics*, 182–6.

Dawid, A. P. and Mortera, J. (1996). Coherent analysis of forensic identification evidence. *J. Roy. Statist. Soc.* B **58**, 425–443.

Dawid, A. P. and Pueschel, J. (1999). Hierarchical models for DNA profiling using heterogeneous databases. *Bayesian Statistics 6* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 187–212.

Devlin, B., Kadane, J and Roeder, K. (1997). Discussion of the paper by Foreman, Smith and Evett. *J. Roy. Statist. Soc.* A **160**, 464.

Devroye, L. (1986). Non-Uniform Random Variate Generation. New York: Springer-Verlag.

Donnelly, P. The non-independence of matches at different loci in DNA profiles: quantifying the effect of close relatives on the match probability. *Heredity* **72**, 26–34.

Donnelly, P. Discussion of the paper by Foreman, Smith and Evett. *J. Roy. Statist. Soc.* A **160**, 460–461.

Fienberg, S. E. and Finkelstein, M.O. (1996). Bayesian statistics and the law. In *Bayesian Statistics 5* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 129–146. Oxford: Oxford University Press.

Foreman, L. A., Evett, I. W. and Smith, A. F. M. (1997). Bayesian analysis of deoxyribonucleic acid profiling data in forensic identification Applications. *J. Roy. Statist. Soc.* A **160**, 429–459.

Fosdick, L. D. (1963). Monte Carlo calculations on the Ising lattice. *Meth. Comput. Phys.* **1**, 245–280.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis.* Chapman and Hall.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.* **6**, 721–741.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–1339.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (ed. E. M. Keramidas), pp. 156–163. Fairfax Station: Interface Foundation.

Geyer, C. J. and Thompson, E. A. (1993). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Technical Report No. 589* School of Statistics, University of Minnesota.

Gilks, W. R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith), pp. 641-649. Oxford: Oxford University Press.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337-348.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman and Hall.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp 1–19. London: Chapman and Hall.

Green, P. J. (1995). Reversible Jump MCMC Computation and Bayesian Model determination. *Biometrika* **82**, 711–732.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods.* New York: Wiley.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Jones, S. (1996). *In The Blood: God, Genes and Destiny.* Harper Collins.

Li, Y. J. (1996). *Characterizing the Structure of Genetic Populations.* Ph.D Thesis. N.C. State University.

Malecot, G. (1948). *Les Mathematiques de l'Hérédité.* Masson et Cie.

Meyn, S. P. and Tweedie, R. L. (1993) *Markov Chains and Stochastic Stability.* New York: Springer-Verlag.

National Research Council (1992). *DNA Technology in Forensic Science.* Washington, DC: National Academy Press.

National Research Council (1996). *The Evaluation of Forensic DNA Evidence.* Washington, DC: National Academy Press.

Nei, M. (1987). *Molecular Evolutionary Genetics.* Columbia University Press.

Pritchard, J. K., Stephens, M. and Donnelly, P. J. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Raftery, A. E. and Lewis, S. M. (1995). Implementing MCMC. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp 215–239. London: Chapman and Hall.

Ripley, B. D. (1987). *Stochastic Simulation.* New York: Wiley.

Roberts, G. O. (1992). Convergence diagnostics of the Gibbs sampler. In *Bayesian Statistics 4* (eds. J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 775–782. Oxford: Oxford University Press.

Roberts, G. O. (1995). Introduction to Markov chains for simulation. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 45–57. London: Chapman and Hall.

Roeder, K., Escobar, M., Kadane, J. B. and Ballazs, I. (1998). Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika* **85** 269–287.

Roeder, K., Escobar, M., Kadane, J. B. and Balazs, I. (1998). Measuring heterogeneity in forensic databases using hierarchical Bayes models: computational method.
http://lib.stat.cmu.edu/www/cmu-stats/tr/tr662/tr662.html

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc.* B **55**, 3–23.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical Science* **8**, No. 3, 219–283.

Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelationsercheinungen und Korrelationsercheinungen vom Stadtpunkt der Vererbungslehre aus betrachtet. *Hereditas* **11**, 65–106.

Weir, B. S. (1994). The effects of inbreeding on forensic calculations. *Annu. Rev. Genet.* **28**, 597–621.

Weir, B. S. (1995). DNA statistics in the Simpson matter. *Nature Genetics* **11**, 365–368.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Weir, B. S. and Evett, I. W. (1998). Interpreting DNA Evidence. Sunderland, Massachusetts: Sinnauer Associates, Inc.

Wright, S. (1951). The genetic structure of populations. *Ann. Eugenics* **15**, 159–171.

The People v. Frank Lee Soto.
http://freecaselaw.com/ca/S044043.htm

The Bugs Project CODA Read Me.
http://www.mrc-bsu.cam.ac.uk/bugs/classic/coda04/readme.shtml

A Biological background

Each cell in the human body should contain the same 23 pairs of chromosomes. The deoxyribonucleic acid (DNA) is contained within these chromosomes in the form of a double strand twisted to form a helix. Each of these strands consists of a string of bases held together by a sugar-phosphate backbone. There are four bases, A (adenine), T (thymine), G (guanine), and C (cytosine). In the double helix, the bases line up in pairs, an A always opposed by a T and a G always opposite a C. The unit of a base plus a link to the next base is known as a nucleotide. A gene is a stretch of DNA, ranging from a few thousand to tens of thousands of base pairs, that produces a specific product, usually a protein. The position that a gene occupies along the thread is its *locus*. At each locus there are two genes (one maternal and one paternal), and each of these takes one of a number of alternative forms (*alleles*). The DNA profile refers to the combination of alleles observed at a group of analyzed loci.

There are two main 'fingerprinting' techniques currently employed, restriction fragment length polymorphisms (RFLP) and polymerase chain reaction (PCR).

RFLP was developed first and uses regions of DNA known as variable number tandem repeats (VNTRs). These VNTRs are not genes as they have no known product or function, and are referred to as markers. A VNTR consists of a core sequence of bases repeated a number of times. The number of repeats at a particular marker locus varies from person to person, and alleles are defined by this number, usually between 500 and 10000. As a result of the differing number of repeats, the alleles can be identified by their lengths, and this is the basis of the RFLP technique. DNA fragments are placed on a gel in an electric field and migrate at a rate dependent upon their lengths. Thus, alleles can be recognized by the distances travelled in the gel. One disadvantage of this method is that tracks whose lengths differ by small amounts can coalesce. This means that an individual who is heterozygous (has two different alleles) at a particular locus cannot always be separated from a homozygous individual.

PCR works on very small regions of DNA and so cannot, at present, be used for most VNTRs. The technique replicates the DNA sample a number of times and then proceeds in a similar manner to RFLP, identifying alleles by their lengths. Short tandem repeat (STR) regions, which are much smaller than VNTRs can be used for this method. Alleles can be resolved to the scale of single bases, eliminating the problem of coalescence. Due to the smaller range of numbers of base pairs, there is a smaller number of possible alleles at each STR locus, but the number of potentially usable loci is very large. Other advantages of this method are that it is quicker and that it can be used on very small DNA samples such as those found in single hairs or saliva traces on cigarette butts.

B Derivation of the full conditional density of $\boldsymbol{a}$

The parameter $\boldsymbol{a}_j$ is defined at each locus j to be

$$\boldsymbol{a}_j = \frac{1-\theta_j}{\theta_j}(\gamma_j(1), \ldots, \gamma_j(r_j)),$$

where $0 \leq \theta_j \leq 1$ and $\sum_{k=1}^{r_j} \gamma_j(k) = 1$.

At this point we drop the locus label j, considering a parameter $\boldsymbol{a}$ at a single locus.

In deriving the prior density of $\boldsymbol{a}$, we assume the following independent prior distributions upon $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$:

$$\begin{aligned} \boldsymbol{\gamma} &\sim \text{Dirichlet}(a_\gamma(1), \ldots, a_\gamma(r)); \\ \boldsymbol{\theta} &\sim \text{Beta}(a_\theta, b_\theta). \end{aligned}$$

The prior of $\boldsymbol{a}$ is then given by

$$f(\boldsymbol{a}|\boldsymbol{a}_\gamma, a_\theta, b_\theta) = \frac{1}{|\det(M)|} f(\boldsymbol{\gamma}, \theta|\boldsymbol{a}_\gamma, a_\theta, b_\theta), \tag{71}$$

where

$$\begin{aligned} M &= \begin{pmatrix} \frac{\delta a(1)}{\delta g(1)} & \frac{\delta a(1)}{\delta g(2)} & \cdots & \frac{\delta a(1)}{\delta g(r-1)} & \frac{\delta a(1)}{\delta \theta} \\ \frac{\delta a(2)}{\delta g(1)} & \frac{\delta a(2)}{\delta g(2)} & \cdots & \frac{\delta a(2)}{\delta g(r-1)} & \frac{\delta a(2)}{\delta \theta} \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{\delta a(r-1)}{\delta g(1)} & \frac{\delta a(r-1)}{\delta g(2)} & \cdots & \frac{\delta a(r-1)}{\delta g(r-1)} & \frac{\delta a(r-1)}{\delta \theta} \\ \frac{\delta a(r)}{\delta g(1)} & \frac{\delta a(r)}{\delta g(2)} & \cdots & \frac{\delta a(r)}{\delta g(r)} & \frac{\delta a(r)}{\delta \theta} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1-\theta}{\theta} & 0 & \ldots & 0 & -\frac{\gamma(1)}{\theta^2} \\ 0 & \frac{1-\theta}{\theta} & \ldots & 0 & -\frac{\gamma(2)}{\theta^2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & \frac{1-\theta}{\theta} & -\frac{\gamma(r-1)}{\theta^2} \\ -\frac{1-\theta}{\theta} & -\frac{1-\theta}{\theta} & \ldots & -\frac{1-\theta}{\theta} & -\frac{\gamma(r)}{\theta^2} \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
\Rightarrow det(M) &= -\left(\frac{1-\theta}{\theta}\right)^{r-1}\left(\frac{1}{\theta^2}\right) \\
&= -\frac{1}{a(+)^r}(a(+)+1)^2 \\
&= -\left(\frac{a(+)+1}{a(+)}\right)^{r-1}\frac{(a(+)+1)^2}{(a(+)+1)^{r-1}} \\
&= -\left(\frac{a(+)}{a(+)+1}\right)^{-(r-1)}\left(\frac{1}{a(+)+1}\right)^{r-3}.
\end{aligned}
$$

Assuming $\boldsymbol{\gamma}$ is independent of θ,

$$
f(\boldsymbol{\gamma},\theta|\boldsymbol{a}_\gamma,a_\theta,b_\theta) \propto \left(\prod_{k=1}^{r}\gamma(k)^{a_\gamma(k)-1}\right).\theta^{a_\theta-1}(1-\theta)^{b_\theta-1},
$$

and hence

$$
f(\boldsymbol{a}|\boldsymbol{a}_\gamma,a_\theta,b_\theta) \propto \left(\prod_{k=1}^{r}\left(\frac{a(k)}{a(+)}\right)^{a_\gamma(k)-1}\right)\left(\frac{1}{a(+)+1}\right)^{a_\theta+r-4}\left(\frac{a(+)}{a(+)+1}\right)^{b_\theta-r}.
$$

The full conditional density of $\boldsymbol{a}$ is given by

$$
f(\boldsymbol{a}|\boldsymbol{a}_\gamma,a_\theta,b_\theta,\boldsymbol{G}) \propto f(\boldsymbol{a}|\boldsymbol{a}_\gamma,a_\theta,b_\theta).f(\boldsymbol{G}|\boldsymbol{a}).
$$

Conditional upon $\boldsymbol{a}$, the subpopulation allele probabilities $(\boldsymbol{G}_l; l=1,\ldots,\eta)$ follow the following Dirichlet distribution, independently across l:

$$
\boldsymbol{G}_l \sim \text{Dirichlet}(a_1,\ldots,a_r).
$$

The full conditional density of $\boldsymbol{a}$ is therefore given by

$$
\begin{aligned}
f(\boldsymbol{a}|\boldsymbol{a}_\gamma,a_\theta,b_\theta,\boldsymbol{G}) \propto\ & \left(\prod_{k=1}^{r}\left(\frac{a(k)}{a(+)}\right)^{a_\gamma(k)-1}\right)\left(\frac{1}{a(+)+1}\right)^{a_\theta+r-4} \\
& \times\left(\frac{a(+)}{a(+)+1}\right)^{b_\theta-r} \\
& \times\prod_{l=1}^{\eta}\left(\frac{\Gamma(a(+))}{\prod_{k=1}^{r}\Gamma(a(k))}\right)\left[\prod_{k=1}^{r}G_l(k)^{a(k)-1}\right].
\end{aligned}
$$

C Empirical distribution of alleles within databases

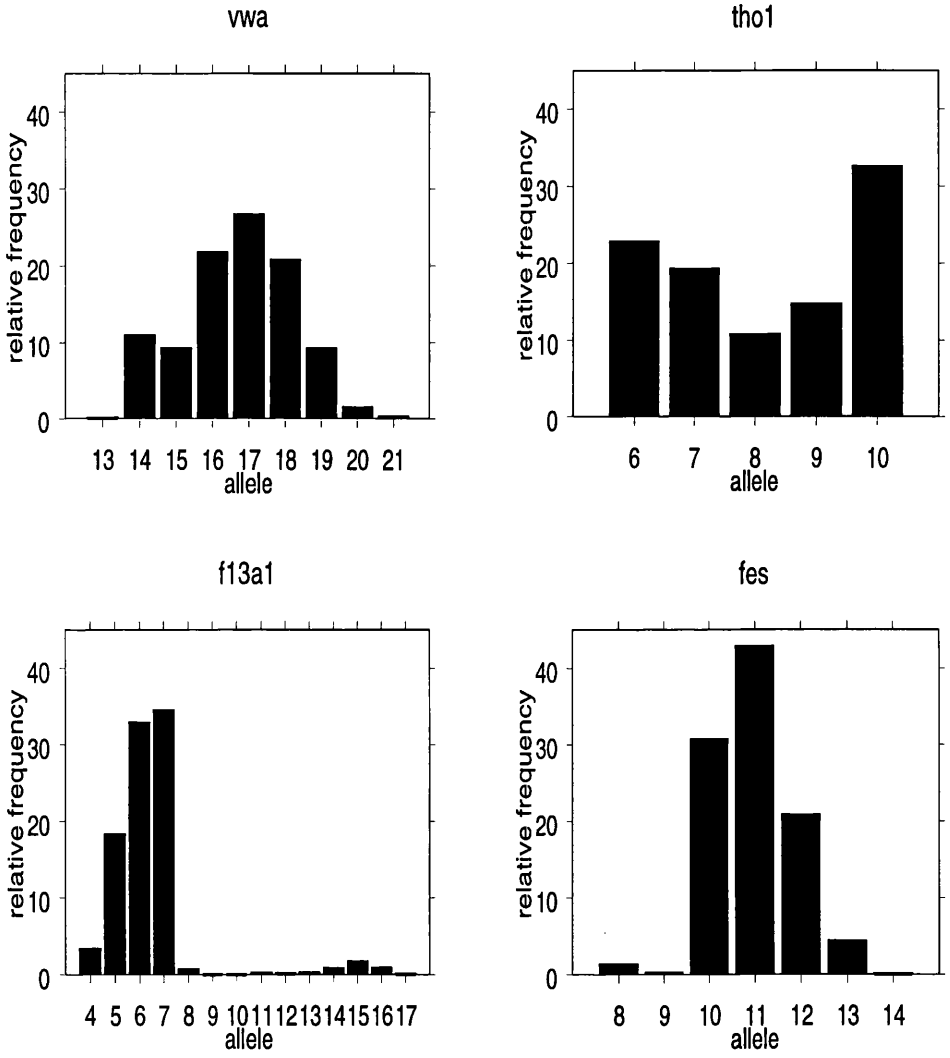


Figure 35: Caucasian database.

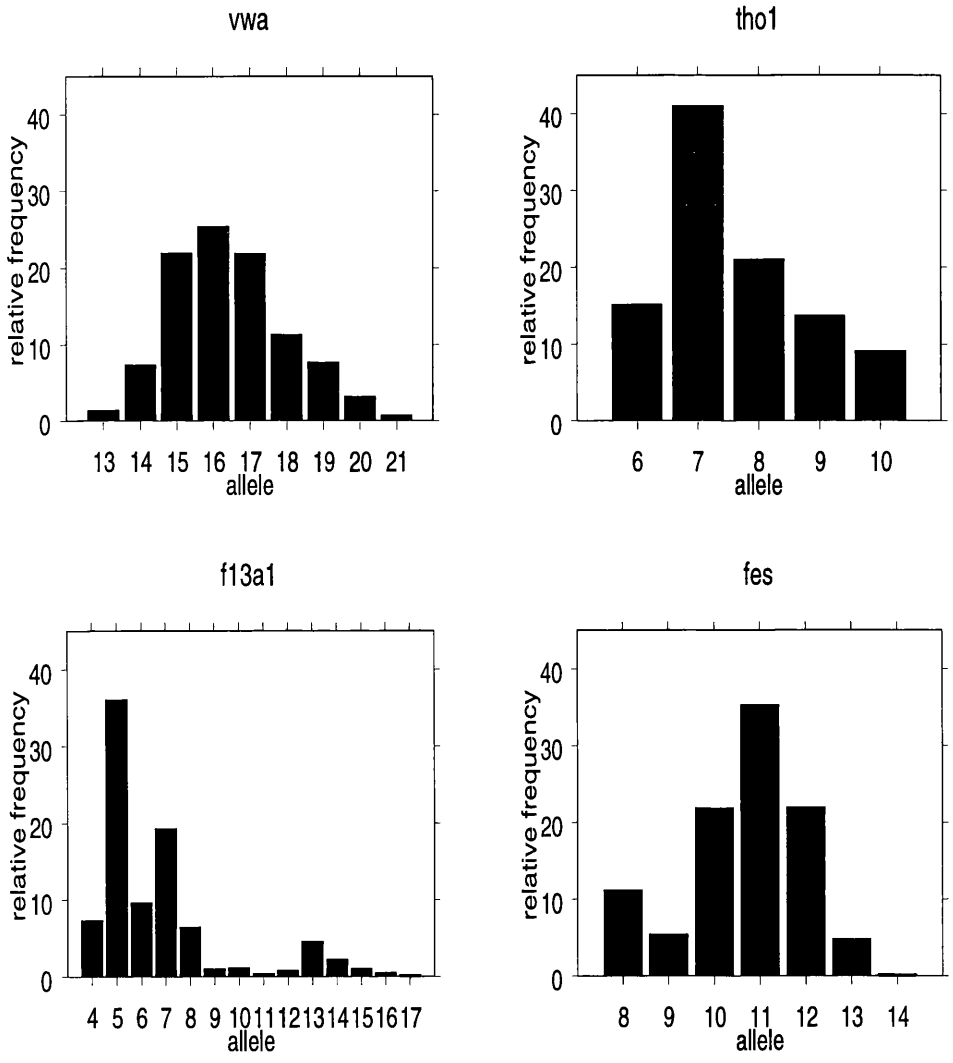


Figure 36: Afro-Caribbean database.

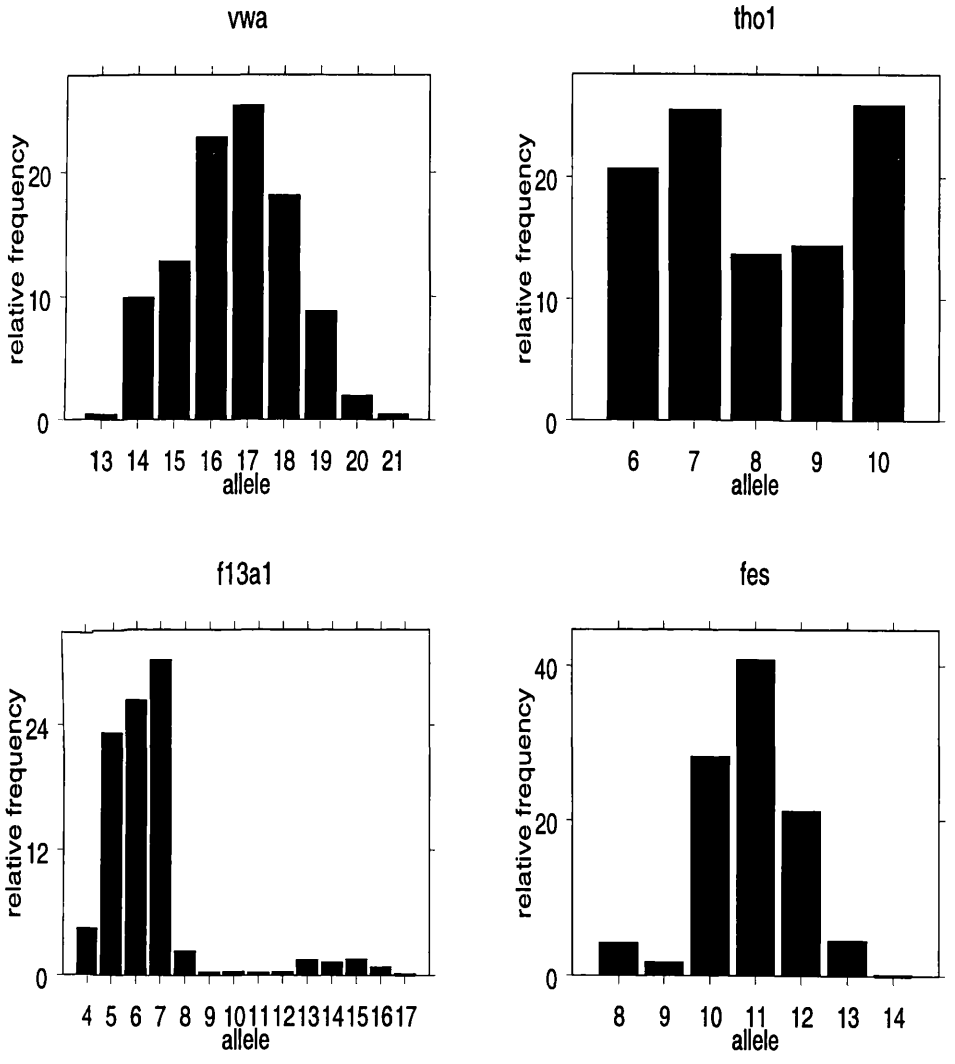


Figure 37: Combined Caucasian/Afro-Caribbean database.