Current challenges in statistical DNA evidence evaluation

Cereda, G.

Author: Cereda, G.
Title: Current challenges in statistical DNA evidence evaluation
Issue Date: 2017-01-12

Chapter 7

Nonparametric Bayesian approach to LR assessment in case of rare type match

This chapter is based on:
Cereda, G. Nonparametric Bayesian approach to LR assessment in case of rare type match. *arXiv:1506.08444*. Submitted to: *Annals of Applied Statistics*.

Abstract

The evaluation of a match between the DNA profile of a stain found on a crime scene and that of a suspect (previously identified) involves the use of the unknown parameter $\mathbf{p} = (p_1, p_2, ...)$, (the ordered vector which represents the frequencies of the different DNA profiles in the population of potential donors) and the names of the different DNA profiles. We propose a Bayesian nonparametric method which models $\mathbf{p}$ through a random variable $\mathbf{P}$ distributed according to the two-parameter Poisson Dirichlet distribution, and discards the information about the names of the different DNA profiles. The ultimate goal of this model is to evaluate the so-called 'probative value' of DNA matches in the rare type case, that is the situation in which the suspect's profile, matching the crime stain profile, is not in the database of reference.

7.1 Introduction

The largely accepted method for evaluating how much some available data $\mathcal{D}$ (typically forensic evidence) is helpful in discriminating between two hypotheses of interest (the prosecution hypothesis H_p and the defense hypothesis H_d), is the calculation of the *likelihood ratio* (LR), a statistic that expresses the relative plausibility of the data under these hypotheses, defined as

$$\mathrm{LR} = \frac{\Pr(\mathcal{D}|H_p)}{\Pr(\mathcal{D}|H_d)}. \tag{7.1}$$

Widely considered the most appropriate framework to report a measure of the 'probative value' of the evidence regarding the two hypotheses (Robertson and Vignaux, 1995; Evett and Weir, 1998; Aitken and Taroni, 2004; Balding, 2005), it indicates the extent to which data is in favor of one hypothesis over the other. Forensic literature presents many approaches to calculate the LR, mostly divided into Bayesian and frequentist methods (see Cereda (2016a,b) for a careful differentiation between these two approaches).

This paper proposes a Bayesian nonparametric method for the LR assessment in the rare type match case, the challenging situation in which there is a match between some characteristic of the recovered material and of the control material, but this characteristic has not been observed before in previously collected samples (i.e. database of reference). This constitutes a problem because the value of the likelihood ratio depends on the unknown proportion of the matching characteristic in a reference population, and the uncertainty over this proportion is, in standard practice, dealt with using the relative frequency of the characteristic in the available database. In particular, we will focus on Y-STR data, for which the rare type match problem is often recurring (Cereda, 2016b).

To use a Bayesian nonparametric method we assume that there are infinitely many Y-STR profiles: the parameter of the model is the infinite dimensional vector $\mathbf{p}$, made of the (unknown) sorted population proportions of all possible Y-STR profiles. As prior over $\mathbf{p}$ we choose the two-parameter Poisson Dirichlet distribution, and we model the uncertainty over its own parameters α and θ through the use of an hyperprior. The information contained in the names of the profiles is discarded: this means to reduce the data $\mathcal{D}$ to a smaller amount of information D.

The paper is structured in the following way: Section 7.2 introduces the notation, the assumptions of our model and the prior distribution chosen for parameter $\mathbf{p}$. Section 7.3 presents the model, along with some theory on random partitions useful to provide a convenient and compact representation of the reduced data D. An alternative representation of the same model via the two-parameter Chinese restaurant process is also described. Section 7.4 introduces relevant known results regarding the two-parameter Poisson Dirichlet distribution, along with a new lemma that will allow to derive the likelihood ratio in a very elegant way (Section 7.5).

Section 7.6 proposes the application of this model to a real database of Y-STR profiles. We will discuss data driven choices for the hyperpriors, and comparison with the frequentist likelihood ratio values obtained both reducing and not the data in the ideal situation in which vector $\mathbf{p}$ is known.

7.2 A Bayesian nonparametric model for the rare type match

7.2.1 The rare type match problem

The evaluation of a match between the profile of a particular piece of evidence and a suspect's profile depends on the proportion of that profile in the population of potential perpetrators. Indeed, it is intuitive that the rarer the matching profile, the more the suspect is in trouble.

Problems arise when the observed frequency of the profile in a sample from the population of interest (i.e., in a reference database) is 0. Such characteristic is likely to be rare, but it is challenging to quantify how rare it is. The rare type match problem is particularly important in case a new kind of forensic evidence, such as results from DIP-STR markers (see for instance Cereda et al. (2014a)) is involved, and for which the available database size is still limited. The same happens when Y-chromosome (or mitochondrial) DNA profiles are used, since the set of possible Y-STR profiles is extremely large. As a consequence, most of the Y-STR haplotypes are not represented in the database. The Y-STR marker system will thus be retained here as an extreme but in practice common and important way in which the problem of assessing evidential value of rare type match can arise. This problem is so substantial that it has been defined "the fundamental problem of forensic mathematics" (Brenner, 2010).

The *empirical frequency estimator*, also called *naive estimator*, that uses the frequency of the characteristic in the database, puts unit probability mass on the set of already observed characteristics, and it is thus unprepared for the observation of a new type. A solution could be the *add-constant* estimators (in particular the well-known *add-one* estimator, due to Laplace (1814), and the *add-half* estimator of Krichevsky and Trofimov (1981)), which add a constant to the count of each type, included the unseen ones. However, this method requires to know the number of possible unseen types, and it performs badly when this number is large compared to the sample size (see Gale and Church (1994) for an additional discussion). Alternatively, Good (1953), based on an intuition on A.M. Turing, proposed the *Good Turing estimator* for the total unobserved probability mass, based on the proportion of singleton observations in the sample. An extension of this estimator is applied to the frequentist LR assessment in the rare type match case in Cereda (2016b). For a comparison between *add one* and *Good-Turing* estimator, see Orlitsky et al. (2003). As pointed out in Anevski et al. (2013), the *naive estimator*, and the *Good Turing estimator* are in some sense complementary: the first gives a good estimate for the observed types, and the second for the probability mass of the unobserved ones. More recently, Orlitsky et al. (2004) have introduced the *high profile estimator*, which extends the tail of the *naive estimator* to the region of unobserved types. Anevski et al. (2013) improved this estimator and provided the consistency proof. Papers that address the rare Y-STR haplotype problem in forensic context are for instance Egeland and Salas (2008), Brenner (2010), and Cereda (2016a). The latter applies the classical Bayesian approach (the beta binomial and the Dirichlet multinomial problem) to the LR assessment in the rare haplotype case. Moreover, the Discrete Laplace method presented in Andersen et al. (2013b), even though not specifically designed for the rare type match case, can be successfully applied to that purpose (Cereda, 2016b).

Bayesian nonparametric estimators for the probability of observing a new type have been proposed by Tiwari and Tripathi (1989) using Dirichlet process, by Lijoi et al. (2007) using general Gibbs prior, and by Favaro et al. (2009) with specific interest to the two-parameter Poisson Dirichlet prior. However, the LR assessment requires not only the probability of observing a new species but also the probability of observing this same species twice (according to the defense the crime stain profile and the suspect profile are two independent observations): to our knowledge, the present paper is the first one to address the problem of LR assessment in the rare haplotype case using Bayesian nonparametric models. As prior for $\mathbf{p}$ we will use the two-parameter Poisson Dirichlet distribution, which is proving useful in many discrete domains, in particular language modelling (Teh et al., 2006). In addition, it shows a power-law behaviour which describes an incredible variety of phenomena (Newman, 2005). Indeed it can be proved that

7.2.2 Notation

Throughout the paper the following notation is chosen: random variables and their values are denoted, respectively, with uppercase and lowercase characters: x is a realization of X. Random vectors and their values are denoted, respectively, by uppercase and lowercase bold characters: $\mathbf{p}$ is a realization of the random vector $\mathbf{P}$. Probability is denoted with $\Pr(\cdot)$, while density of a continuous random variable X is denoted alternatively by $p_X(x)$ or by $p(x)$ when the subset is clear from the context. For a discrete random variable Y, the density notation $p_Y(y)$ and the discrete one $\Pr(Y = y)$ will be alternately used. Moreover, we will use shorthand notation like $p(y \mid x)$ to stand for the probability density of Y with respect to the conditional distribution of Y given $X = x$.

Notice that in Formula (7.1), $\mathcal{D}$ was regarded as the event corresponding to the observation of the available data. However, later in the paper, $\mathcal{D}$ will be regarded as a random variable generically representing the data. The particular data at hand will correspond to the value d. In that case, the following notation will thus be preferred:

$$\mathrm{LR} = \frac{\Pr(\mathcal{D} = d|H = h_p)}{\Pr(\mathcal{D} = d|H = h_d)} \quad \text{or} \quad \frac{p(d|h_p)}{p(d|h_d)}. \tag{7.2}$$

Lastly, notice that "DNA types" is used throughout the paper as a general term to indicate Y-STR profiles.

7.2.3 Model assumptions

Our model is based on the two following assumptions:

Assumption 1 There are infinitely many DNA types in Nature.

The reason for this assumption is that there are so many possible DNA types that they can be considered infinite. This assumption, already used by e.g. Kimura (1964) in the 'infinite alleles model', allows to use Bayesian nonparametric methods and avoids the problem of specifying how many different types there are in Nature.

Assumption 2 The names of the different DNA types do not contain information.

Actually, the specific sequence of numbers that forms a DNA profile carries information: if two profiles show few differences this means that they are separated by few mutation drifts, hence the profiles share a relatively recent common ancestor. However, this information is difficult to exploit and may be not so relevant for the LR assessment. This is the reason why we will treat DNA types as "colors", and only consider the partition into different categories. Stated otherwise, we put no topological structure on the space of the DNA types.

Notice that this assumption makes the model a priori suitable for any characteristic which shows many different possible types, thus what written still holds, in principle, also replacing 'DNA types' with any other category. However, in this paper we will only test the model with Y-STR profiles as categories.

7.2.4 Prior

In Bayesian statistics, parameter of interest are modeled through random variables. The (prior) distribution over a parameter should represent the uncertainty about its value.

LR assessment for the rare type match involves two unknown parameters of interest: one is $h \in \{h_p, h_d\}$, representing the unknown true hypothesis, the other is $\mathbf{p}$, the vector of the unknown population frequencies of all DNA profiles in the population of potential perpetrators. The dichotomous random variable H is used to model parameter h, and the posterior distribution of this random variable, given the data, is the ultimate aim of the forensic inquiry. In a similar way, random variable $\mathbf{P}$ is used to model the uncertainty over $\mathbf{p}$. Because of Assumption 1, $\mathbf{p}$ is an infinite dimensional parameter, hence the need of Bayesian nonparametric methods (Hjort et al., 2010). In particular, $\mathbf{p} = (p_t | t \in T)$, with T a countable set of indexes, $p_t > 0$, and $\sum_t p_t = 1$. Moreover, because of Assumption 2, data can be reduced to random partitions, as explained in Section 7.3.1, and it will turn out that the distribution of these partitions does not depend on the order of the p_i. Hence, we can force the parameter $\mathbf{p}$ to have values in $\nabla_\infty = \{(p_1, p_2, ...)| p_1 \geq p_2 \geq ..., \sum p_i = 1, p_i > 0\}$, the ordered infinite dimensional simplex. The uncertainty about its value is expressed by the prior distribution over $\mathbf{p}$, for which we choose the two-parameter Poisson Dirichlet distribution (Pitman and Yor, 1997; Feng, 2010; Buntine and Hutter, 2010; Carlton, 1999; Pitman and Picard, 2006), defined in the following way:

Definition 1 (two-parameter GEM distribution)**.** *Given α and θ satisfying the following conditions:*

$$0 \leq \alpha < 1, \text{ and } \theta > -\alpha. \tag{7.3}$$

the vector $\mathbf{W} = (W_1, W_2, ...)$ is said to be distributed according to the GEM(α, θ)*, if*

$$\forall i \quad W_i = V_i \prod_{j=1}^{i-1} (1 - V_j),$$

where V_1, V_2,... are independent random variables distributed according to

$$V_i \sim B(1 - \alpha, \theta + i\alpha).$$

It holds that $W_i > 0$, and $\sum_i W_i = 1$.

The GEM distribution (short for Griffin - Engen - McCloskey distribution') is well known in literature as the "stick breaking prior", since it measures the random sizes in which a stick is broken iteratively. This distribution is invariant under size-biased permutations (Engen, 1975), that is the random permutation defined by sampling from the population and assigning to each type a label, based on the order in which the types are first sampled.

Definition 2 (Two-parameter Poisson Dirichlet distribution)**.** *Given α and θ satisfying condition* (7.3)*, and a vector* $\mathbf{W} = (W_1, W_2, ...) \sim$ *GEM*(α, θ)*, the random vector* $\mathbf{P} = (P_1, P_2, ...)$ *obtained by ordering* $\mathbf{W}$*, such that $P_i \geq P_{i+1}$, is said to be* Poisson Dirichlet distributed *PD*(α, θ)*. Parameter α is called* discount parameter*, while θ is the* concentration parameter.

Notice that the vector $\mathbf{P}$ is obtained by sorting the vector $\mathbf{W}$ in nonincreasing order, while the vector $\mathbf{W}$ can be obtained (in distribution) by the so-called *size-biased permutation* of the indexes of $\mathbf{P}$ (Perman et al., 1992; Pitman and Yor, 1997).

The two-parameter Poisson Dirichlet distribution PD(α, θ) is the generalization of the well-known Poisson Dirichlet distribution with a single parameter θ introduced by Kingman (1975), which is the representation measure (Kingman, 1977, 1978) of the celebrated *Ewens sampling formula* (Ewens, 1972), widely applied in genetics (Karlin and McGregor, 1972; Kingman, 1980). For our model we will not allow $\alpha = 0$, hence we will assume $0 < \alpha < 1$.

It is worth mentioning that an alternative choice for the parameters space is $\alpha < 0, \theta = -m\alpha$ for some $m \in \mathbb{N}$ (Pitman, 1996; Gnedin and Pitman, 2006; Gnedin, 2009; Cerquetti, 2010). It corresponds to a model with finitely many (m) DNA types, where $\mathbf{P} = (P_1, ..., P_m)$ is Dirichlet distributed with m parameters equal to $-\alpha$. We will not consider this case.

Lastly, we point out that, in practice, we cannot assume to know parameters α and θ: we will model the uncertainty about them using an hyperprior.

7.3 The model

The typical data to evaluate in case of a match is $\mathcal{D} = (E, B)$, where $E = (E_s, E_t)$, and

- E_s = suspect's DNA type,
- E_t = crime stain's DNA type (matching with the suspect's type),
- B = a reference database of size n, which contains a sample of DNA types, indexed by $i = 1, ..., n$, from the population of possible perpetrators.

The hypotheses of interest for the case are:

- h_p = The crime stain was left by the suspect,
- h_d = The crime stain was left by someone else.

In agreement with Assumption 2, the model will ignore information about the names of the DNA types: data $\mathcal{D} = (E, B)$ will be reduced to D accordingly. The Bayesian network of Figure 7.1 encapsulates the conditional dependencies of the random variables of the proposed model:

- H is a dichotomous random variable that represents the hypotheses of interest and can take values $h \in \{h_p, h_d\}$, according to the prosecution or the defense, respectively. A uniform prior on the hypotheses is chosen:

$$\Pr(H = h) \propto 1 \quad \text{for } h \in \{h_p, h_d\}.$$

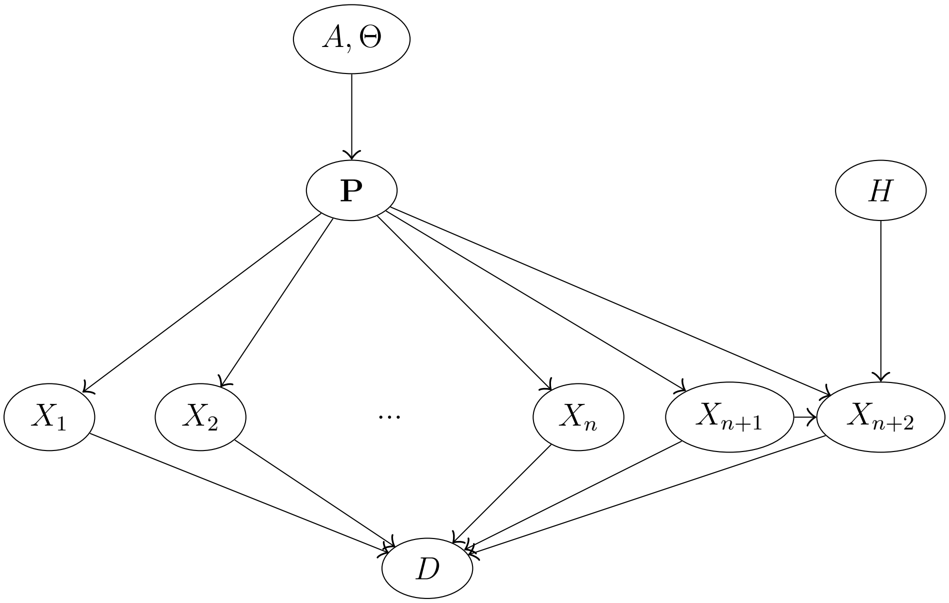


Figure 7.1: Bayesian network to show the conditional dependencies of the relevant random variables in our model.

Notice that this choice is made for mathematical convenience, since it will not affect the likelihood ratio.

- (A, Θ) is the random vector that represents the hyperparameters α and θ, satisfying condition (7.3). The joint prior density of these two parameters (hyperprior) will be generically denoted as $p(\alpha, \theta)$:
$$(A, \Theta) \sim p(\alpha, \theta).$$

- The random vector $\mathbf{P}$ with values in ∇_∞, represents the ranked population frequencies. $\mathbf{P} = (p_1, p_2, ...)$ means that p_1 is the frequency of the most common DNA type in the population, p_2 is the frequency of the second most common DNA type, and so on. As a prior for $\mathbf{P}$ we use the two-parameter Poisson Dirichlet distribution (see Definition 2):
$$\mathbf{P}|A = \alpha, \Theta = \theta \sim PD(\alpha, \theta).$$

- The database is assumed to be a random sample from the population. Integer valued random variables X_1, ..., X_n are here used to represent the ranks of the population proportions of the DNA types in the database. For instance, $X_3 = 5$ means that the third individual in the database has the fifth most common DNA type in the population. Given $\mathbf{p}$ they are an i.i.d. sample from $\mathbf{p}$:
$$X_1, X_2, ..., X_n|\mathbf{P} = \mathbf{p} \sim_{i.i.d.} \mathbf{p}. \tag{7.4}$$
To observe X_1, ..., X_n, one would need to know the rank, in terms of population proportion, of the frequency of each DNA types in the database. This is not known, hence we don't observe $X_1, ..., X_n$.

- X_{n+1} represents the rank of the suspect's DNA type. It is again an indipendent draw from $\mathbf{p}$.

$$X_{n+1}|\mathbf{P} = \mathbf{p} \sim \mathbf{p}.$$

- X_{n+2} represents the rank of the crime stain's DNA type. According to the prosecution, given $X_{n+1} = x_{n+1}$, this random variable is deterministic (it is equal to x_{n+1} with probability 1). According to the defense it is another sample from $\mathbf{p}$, independent of the previous ones:

$$X_{n+2}|\mathbf{P} = \mathbf{p}, X_{n+1} = x_{n+1}, H = h \sim \begin{cases} \delta_{x_{n+1}} & \text{if } h = h_p \\ \mathbf{p} & \text{if } h = h_d \end{cases}.$$

As already mentioned, $X_1, ..., X_{n+2}$ cannot be observed. They represent the database, where the names of the DNA types have been replaced by their (unknown) ranks in $\mathbf{p}$, and constitute an intermediate layer.

Section 7.3.1 recalls some notions about random partitions, useful before defining node D, the 'reduced' data that we want to evaluate.

7.3.1 Random partitions

A *partition of a set* A is an unordered collection of nonempty and disjoint subsets of A the union of which forms A. Particularly interesting for our model are partitions of the set $A = [n] = \{1, ..., n\}$, denoted as $\pi_{[n]}$. The set of all partitions of $[n]$ will be denoted as $\mathcal{P}_{[n]}$. Random partitions of $[n]$ will be denoted as $\Pi_{[n]}$. In addition, a *partition of* n is a finite nonincreasing sequence of positive integers that sum up to n. Partitions of n will be denoted as π_n, random partitions as Π_n.

Given a sequence of integer valued random variables $X_1, ..., X_n$, let $\Pi_{[n]}(X_1, X_2, ..., X_n)$ be the random partition defined by the equivalence classes of their indexes using the random equivalence relation $i \sim j$ if and only if $X_i = X_j$. This construction allows to build a map from the set of values of $X_1, ..., X_n$ to the set of the partitions of $[n]$ as in the following example ($n = 10$):

$$\begin{aligned} \mathbb{N}^{10} &\to \mathcal{P}_{[10]} \\ X_1, ..., X_{10} &\longmapsto \Pi_{[10]}(X_1, X_2, ..., X_{10}) \\ (2, 4, 2, 4, 3, 3, 10, 13, 5, 4) &\longmapsto \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}\} \end{aligned}$$

In agreement with Assumption 2, in our model we can consider the reduction of data which ignores information about the names of the DNA types: this is achieved, for instance, by retaining from the database only the equivalence classes of the indexes of the individuals, according to the equivalence relation "to have the same DNA type". Stated otherwise, the database is reduced to the partition $\pi_{[n]}^{\mathrm{Db}}$, obtained using these equivalence classes. However, data is not only made of the database B. There are also two new DNA profiles which are equal one another and different from the already observed ones. When the suspect's profile is considered we obtain the partition $\pi_{[n+1]}^{\mathrm{Db+}}$, where the first n integers are partitioned as in

$\pi^{\text{Db}}_{[n]}$, and $n+1$ constitutes a class by itself (at least in the rare type match case). When the crime stain profile is considered we obtain the partition $\pi^{\text{Db++}}_{[n+2]}$ where the first n integers are partitioned as in $\pi^{\text{Db}}_{[n]}$, and $n+1$ and $n+2$ belongs to the same (new) class.

Random variables $\Pi^{\text{Db}}_{[n]}$, $\Pi^{\text{Db+}}_{[n+1]}$, and $\Pi^{\text{Db++}}_{[n+2]}$ are used to model $\pi^{\text{Db}}_{[n]}$, $\pi^{\text{Db+}}_{[n+1]}$, and $\pi^{\text{Db++}}_{[n+2]}$, respectively.

Since prosecution and defense agree on the distribution of $X_1, ..., X_{n+1}$, but not on the distribution of X_{n+2}, they also agree on the distribution of $\Pi^{\text{Db+}}_{[n+1]}$ but disagree on the distribution of $\Pi^{\text{Db++}}_{[n+2]}$.

The crucial point of the model is that, by construction, the same random partitions can be defined through random variables X_1, ..., X_{n+2}. Indeed, it holds that:

$$\begin{aligned}\Pi^{\text{Db}}_{[n]} &= \Pi_{[n]}(X_1, ..., X_n),\\ \Pi^{\text{Db+}}_{[n+1]} &= \Pi_{[n+1]}(X_1, ..., X_{n+1}),\\ \Pi^{\text{Db++}}_{[n+2]} &= \Pi_{[n+2]}(X_1, ..., X_{n+2}).\end{aligned}$$

Moreover, although X_1, ..., X_{n+2} were not observable, the random partitions $\Pi^{\text{Db}}_{[n]}, \Pi^{\text{Db+}}_{[n+1]}$, and $\Pi^{\text{Db++}}_{[n+2]}$ are observable.

To clarify, consider the following example of a database (B) with $k = 6$ different DNA types, from $n = 10$ individuals:

$$B = (h_1, h_2, h_1, h_2, h_3, h_3, h_4, h_5, h_6, h_2),$$

where h_i is the name of the ith DNA type according to the order chosen for the database. This can be reduced to the partition of [10]:

$$\pi^{\text{Db}}_{[10]} = \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}\}.$$

Then, the part of data whose distribution is agreed on by prosecution and defense is

$$\pi^{\text{Db+}}_{[11]} = \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}, \{11\}\},$$

while the entire (reduced) data D can be represented as

$$\pi^{\text{Db++}}_{[12]} = \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}, \{11, 12\}\}.$$

Now, assume that we know the rank in the population of each of the DNA types in the database: we know that h_1 is, for instance, the second most frequent type, h_2 is the fourth most frequent type, and so on. Stated otherwise, we are now assuming that we observe the variables X_1, ..., X_{n+2}: for instance, $X_1 = 2$, $X_2 = 4$, $X_3 = 2$, $X_4 = 4$, $X_5 = 3$, $X_6 = 3$, $X_7 = 10$, $X_8 = 13$, $X_9 = 5$, $X_{10} = 4$, $X_{11} = 9$, $X_{12} = 9$. It is easy to check that $\Pi_{[10]}(X_1, ..., X_{10}) = \pi^{\text{Db}}_{[10]}$, $\Pi_{[11]}(X_1, ..., X_{11}) = \pi^{\text{Db+}}_{[11]}$, and $\Pi_{[12]}(X_1, ..., X_{12}) = \pi^{\text{Db++}}_{[12]}$.

As already mentioned, data D is defined as:

- $D = \pi^{\text{Db}++}_{[n+2]}$, obtained partitioning the database enlarged with the two new observations (or partitioning $X_1, ..., X_{n+2}$).

Node D of Figure 7.1 is defined accordingly. Notice that, given $X_1, ..., X_{n+2}$, D is deterministic. An important result is that, according to Proposition 4 in Pitman (1992) it is possible to derive directly the distribution of $D \mid \alpha, \theta, H$. In particular, it holds that if

$$\mathbf{P} \mid \alpha, \theta \sim PD(\alpha, \theta),$$

and

$$X_1, X_2, ... \mid \mathbf{P} = \mathbf{p} \sim_{\text{i.i.d}} \mathbf{p},$$

then, for all $n \in \mathbb{N}$, the random partition $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the following distribution:

$$\mathbb{P}^{\alpha,\theta}_{n}(\pi_{[n]}) := \Pr(\Pi_{[n]} = \pi_{[n]}|\alpha, \theta) = \frac{[\theta+\alpha]_{k-1;\alpha}}{[\theta+1]_{n-1;1}} \prod_{i=1}^{k} [1-\alpha]_{n_i-1;1}, \tag{7.5}$$

where n_i is the size of the ith block of $\pi_{[n]}$ (the blocks are here ordered according to the least element), and $\forall x, b \in \mathbb{R}, a \in \mathbb{N}$, $[x]_{a,b} := \begin{cases} \prod_{i=1}^{a-1}(x+ib) & \text{if } a \in \mathbb{N}\backslash\{0\} \\ 1 & \text{if } a = 0 \end{cases}$. This formula is also known as the *Pitman sampling formula*, further studied in Pitman (1995). Notice that for $\alpha = 0$ we obtain the Ewens's sampling formula.

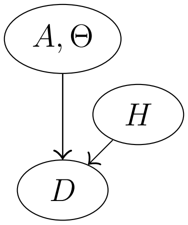


Figure 7.2: Simplified version of the Bayesian network in Figure 7.1

It follows that we can get rid of the intermediate layer of nodes $X_1, ..., X_{n+2}$, and $\Pr(D|\alpha, \theta, h_p) = \mathbb{P}^{\alpha,\theta}_{n+1}(\pi^{\text{Db}+}_{[n+1]})$, while $\Pr(D|\alpha, \theta, h_d) = \mathbb{P}^{\alpha,\theta}_{n+2}(\pi^{\text{Db}++}_{[n+2]})$. The model of Figure 7.1 can thus be simplified to the one in Figure 7.2.

7.3.2 Chinese Restaurant representation

There is an alternative characterization of this model, called "Chinese restaurant process", due to Aldous (1985) for the one parameter case, and studied in details for the two-parameter version in Pitman and Picard (2006). It is defined as follows: consider a restaurant with infinite many tables, each one infinitely large. Let $Y_1, Y_2, ...$ be integer valued random variables that represent the seating plan: tables are ranked in order of occupancy, and $Y_i = j$ means that the ith customer seats at the jth table to be created. The process is described by the following transition matrix:

$$Y_1 = 1,$$

$$\Pr(Y_{n+1} = i | Y_1, ..., Y_n) = \begin{cases} \dfrac{\theta + k\alpha}{n + \theta} & \text{if } i = k + 1 \\ \\ \dfrac{n_i - \alpha}{n + \theta} & \text{if } 1 \leq i \leq k \end{cases} \tag{7.6}$$

where k is the number of tables occupied by the first n customers, and n_i is the number of customers that occupy table i. The process depends on two parameters α and θ with the same conditions (7.3).

$Y_1, ..., Y_n$ are not i.i.d., nor exchangeable, but it holds that $\Pi_{[n]}(Y_1, ..., Y_n)$ is distributed as $\Pi_{[n]}(X_1, ..., X_n)$, with $X_1, ..., X_n$ defined as in (7.4) (in particular they are both distributed according to the Pitman sampling formula (7.5)).

Stated otherwise, we can use the seating plan of n customers $Y_1, ..., Y_n$, or $X_1, ..., X_n$ (the database) and we obtain the same partition $\pi^{\text{Db}}_{[n]}$. Similarly $\pi^{\text{Db+}}_{[n+1]}$ is obtained when a new customer has chosen an unoccupied table (remember we are in the rare type match case), and $\pi^{\text{Db++}}_{[n+2]}$ is obtained when the $n + 2$nd customer goes to the table already chosen by the $n + 1$st customer(suspect and crime stain have the same DNA type). In particular, thanks to (7.6), we can write

$$p(\pi^{\text{Db++}}_{[n+2]} \mid h_p, \pi^{\text{Db+}}_{[n+1]}, \alpha, \theta) = 1, \tag{7.7}$$

and

$$p(\pi^{\text{Db++}}_{[n+2]} \mid h_d, \pi^{\text{Db+}}_{[n+1]}, \alpha, \theta) = \frac{1 - \alpha}{n + 1 + \theta}, \tag{7.8}$$

since the $n+2$nd customer goes to the same table as the $n+1$st (who was sitting alone).

7.4 Some results

This section presents some useful results that will be used in the forthcoming sections. In particular, Lemma 2, suitable to broader applications, is here applied to simplify the likelihood ratio development. Then, some results from Pitman and Picard (2006) regarding the two-parameter Poisson Dirichlet distribution, are listed.

7.4.1 A useful Lemma

The following lemma is a result regarding four general random variables A, X, Y, H whose conditional dependencies are described by the Bayesian network of Figure 8.4. The importance of this result is due to the possibility of applying it to a very common forensic situation: the prosecution and the defense disagree on the distribution of the entirety of data (Y) but agree on the distribution of a part it (X), and these distributions depend on parameters (A).

Lemma 2. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the Bayesian network of Figure 7.3, the likelihood function for h, given $X = x$ and $Y = y$ satisfies*

$$\text{lik}(h \mid x, y) \;\propto \mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

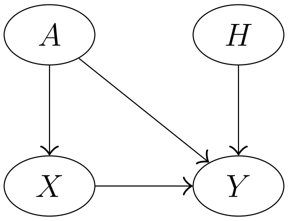
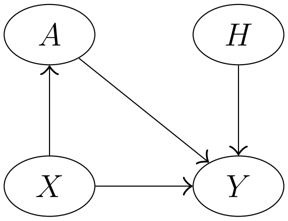


Figure 7.3: Conditional dependencies of the random variables of Lemma 2

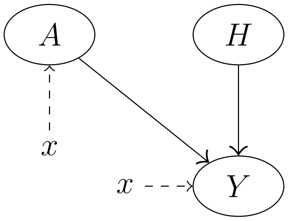
Proof. The model of Figure 7.3 represents four variables A, H, X and Y whose joint probabilty density can be factored as

$$p(a, h, x, y) \;=\; p(a)\,p(x \mid a)\,p(h)\,p(y \mid x, a, h).$$

By Bayes formula, $p(a)\,p(x \mid a) = p(x)\,p(a \mid x)$. This rewriting corresponds to reversing the direction of the arrow between A and X:



The random variable X is now a root node. This means that when we probabilistically condition on $X = x$, the graphical model changes in a simple way: we can delete the node X, but just insert the value x as a parameter in the conditional probability tables of the variables A and Y which formerly had an arrow from node X. The next graph represents this model:



This tells us, that conditional on $X = x$, the joint density of A, Y and H is equal to

$$p(a \mid x)p(h)p(y \mid x, a, h).$$

The joint density of H and Y is obtained by integrating out the variable a. It can be expressed as a conditional expectation value, since $p(a \mid x)$ is the density of A given $X = x$. We find:

$$p(h)\mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

Recall that this is the joint density of two of our variables, H and Y, after conditioning on the value $X = x$. Let us now also condition on $Y = y$. It follows that the density of H given $X = x$ and $Y = y$ is proportional (as function of H, for fixed x and y) to the same expression, $p(h)\mathbb{E}(p(y \mid x, A, h) \mid X = x)$.

This is a product of the prior for h with some function of x and y. Since posterior odds equals prior odds times likelihood ratio, it follows that the likelihood function for h, given $X = x$ and $Y = y$ satisfies

$$\mathrm{lik}(h \mid x, y) \ \propto \mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

□

Corollary 3. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the network of Figure 8.4, the likelihood ratio for $H = h_1$ against $H = h_2$ given $X = x$ and $Y = y$ satisfies*

$$\mathrm{LR} = \frac{\mathbb{E}(p(y|x, A, h_1)|X = x)}{\mathbb{E}(p(y|x, A, h_2)|X = x)}. \tag{7.9}$$

The importance of Lemma 2, and Corollary 3 is due to the possibility of applying it to our model. Indeed, as already noticed, since defense and prosecution agree on the distribution of $\pi^{\mathrm{Db}+}_{[n+1]}$, but not on the distribution of $\pi^{Db++}_{[n+2]}$, and data depends on parameters α and θ.

7.4.2 Known results about the two-parameter Poisson Dirichlet distribution

We will now list some theoretical results which will be useful in the forthcoming analysis. Most of these results can be found in Pitman and Picard (2006).

Denote as K_n the random number of blocks of a partition $\Pi_{[n]}$ distributed according to the Pitman sampling formula with parameters α and θ.

- There exists a positive random variable S_α such that

$$\lim_{n\to+\infty} \frac{K_n}{n^\alpha} = S_\alpha \quad \text{a.s.} \tag{7.10}$$

  the distribution of S_α is a generalization of the Mittag-Leffler distribution (Gorenflo et al., 2014).

- If $\mathbf{P} \sim \mathrm{PD}(\alpha, \theta)$, then

$$\frac{P_i}{Zi^{-1/\alpha}} \to 1, \quad \text{a.s., when } i \to +\infty \tag{7.11}$$

  for a random variable Z such that $Z^{-\alpha} = \Gamma(1-\alpha)/S_\alpha$.

- For a fixed $\alpha \in (0, 1)$, the $\mathrm{PD}(\alpha, \theta)$ (for different θ) are all mutually absolutely continuous. This means that θ cannot be consistently estimated for α in the range of interest. On the other hand, the power-baw behavior described above tells us that α can be consistently estimated.

- Studying (7.6) one can see that when n increases, the parameter θ becomes less and less important. However, it describes how much "social" are the customers: the smaller θ the more the customers tend to seat to already occupied tables. Thus, it determines the sizes of the big tables, but it won't be much important for our application (the more rare DNA types correspond to small tables).
- Given Π_n distributed according to Pitman sampling formula (7.5), it holds that

$$\lim_{n\to+\infty} \frac{m_j(n)}{n^\alpha} = \frac{\alpha\Gamma(j-\alpha)}{\Gamma(1-\alpha)j!} S_\alpha \quad \text{a.s.}\, \forall j \tag{7.12}$$

  where $m_j(n)$, $j = 1, ..., n$ the random number of blocks of the partition $\Pi_{[n]}$ of size j. This result is presented in Gnedin et al. (2007), based on Karlin (1967).

7.5 The likelihood ratio

Using the hypotheses and the reduction of data D defined in Section 7.3, the likelihood ratio will be defined as

$$\text{LR} = \frac{p(\pi^{\text{Db}++}_{[n+2]}|h_p)}{p(\pi^{\text{Db}++}_{[n+2]}|h_d)} = \frac{p(\pi^{\text{Db}+}_{[n+1]}, \pi^{\text{Db}++}_{[n+2]}|h_p)}{p(\pi^{\text{Db}+}_{[n+1]}, \pi^{\text{Db}++}_{[n+2]}|h_d)}.$$

The last equality holds due to the fact that $\Pi^{\text{Db}+}_{[n+1]}$ is a deterministic function of $\Pi^{\text{Db}++}_{[n+2]}$.

Now, we can apply Corollary 3 with (A, Θ) playing the role of A, $X = \Pi^{\text{Db}+}_{[n+1]}$, and $Y = \Pi^{\text{Db}++}_{[n+2]}$ to obtain:

$$\begin{aligned} \text{LR} &= \frac{\mathbb{E}(p(\pi^{\text{Db}++}_{[n+2]} \mid \pi^{\text{Db}+}_{[n+1]}, A, \Theta, h_p) \mid \Pi^{\text{Db}+}_{[n+1]} = \pi^{\text{Db}+}_{[n+1]})}{\mathbb{E}(p(\pi^{\text{Db}++}_{[n+2]} \mid \pi^{\text{Db}+}_{[n+1]}, A, \Theta, h_d) \mid \Pi^{\text{Db}+}_{[n+1]} = \pi^{\text{Db}+}_{[n+1]})} \\ &= \frac{1}{\mathbb{E}\Big(\frac{1-A}{n+1+\Theta} \mid \Pi^{\text{Db}+}_{[n+1]} = \pi^{\text{Db}+}_{[n+1]}\Big)}. \end{aligned}$$

where the last equality is due to (7.7) and (7.8). By defining the random variable $\Phi = n\dfrac{1-A}{n+1+\Theta}$ we can write the LR as

$$\text{LR} = \frac{n}{\mathbb{E}(\Phi \mid \Pi^{\text{Db}+}_{[n+1]} = \pi^{\text{Db}+}_{[n+1]})}. \tag{7.13}$$

7.5.1 True LR

It is now interesting to study the frequentist likelihood ratio values obtained with (7.13), and to compare it with the 'true' ones, meaning the LR values obtained when vector $\mathbf{p}$ is known. This corresponds to having the list of the frequencies of all the DNA types in the population of interest. Then, the model can be represented by the Bayesian network of Figure 7.4.

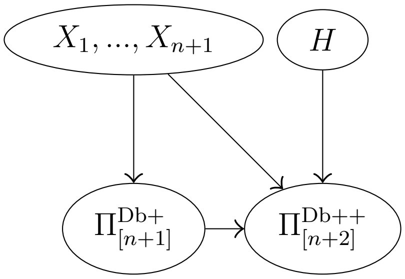


Figure 7.4: Bayesian network for the case in which $\mathbf{p}$ is known.

The LR in this case can be obtained using again Corollary 3, where now X_1, ..., X_{n+1} play the role of A.

$$\text{LR}_{|\mathbf{p}} = \frac{p(\pi^{\text{Db++}}_{[n+2]}, \pi^{\text{Db+}}_{[n+1]} \mid h_p, \mathbf{p})}{p(\pi^{\text{Db++}}_{[n+2]}, \pi^{\text{Db+}}_{[n+1]} \mid h_d, \mathbf{p})} \tag{7.14}$$

$$= \frac{\mathbb{E}(p(\pi^{\text{Db++}}_{[n+2]} \mid \pi^{\text{Db+}}_{[n+1]}, X_1, ..., X_{n+1}, h_p, \mathbf{p}) \mid \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]}, \mathbf{p})}{\mathbb{E}(p(\pi^{\text{Db++}}_{[n+2]} \mid \pi^{\text{Db+}}_{[n+1]}, X_1, ..., X_{n+1}, h_d, \mathbf{p}) \mid \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]}, \mathbf{p})} \tag{7.15}$$

$$= \frac{1}{\mathbb{E}(p_{X_{n+1}}|\Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]}, \mathbf{p})}. \tag{7.16}$$

Notice that, in the rare type case, X_{n+1} is observed only once among the X_1, ..., X_{n+1}. Hence, we call it a singleton. Let s_1 denote the number of singletons, and $\mathcal{S}$ the set of indexes of singletons observations in the database. Notice also that the knowledge of $\mathbf{p}$ and $\pi^{\text{Db+}}_{[n+1]}$, is not enough to observe $X_1, ..., , X_{N+1}$. On the other hand, given $\pi^{\text{Db+}}_{[n+1]}$, both s_1 and $\mathcal{S}$ are fixed and known. Given $\mathbf{p}$ and $\pi^{\text{Db+}}_{[n+1]}$, it holds that the distribution of X_{n+1} is the same as the distribution of all other singletons. This implies that:

$$s_1\mathbb{E}(p_{X_{n+1}}|\pi^{\text{Db+}}_{[n+1]}, \mathbf{p}) = \mathbb{E}(\sum_{i\in\mathcal{S}} p_{X_i}|\pi^{\text{Db+}}_{[n+1]}, \mathbf{p}).$$

Let us denote as X_1^*, .., X_K^* the K different values taken by X_1, ..., X_{n+1}, ordered according to the frequency of their values. Stated otherwise, if n_i is the frequency of x_i^* among $x_1, ..., x_{n+1}$, then $n_1 \geq n_2 \geq ... \geq n_K$. Moreover, in case X_i^* and X_j^* have the same frequency ($n_i = n_j$), then they are ordered according to their values. For instance, if $X_1 = 2$, $X_2 = 4$, $X_3 = 2$, $X_4 = 4$, $X_5 = 3$, $X_6 = 3$, $X_7 = 10$, $X_8 = 13$, $X_9 = 5$, $X_{10} = 4$, $X_{11} = 9$, then $X_1^* = 4, X_2^* = 2, X_3^* = 3, X_4^* = 5, X_5^* = 10, X_6^* = 13$.

By definition, it holds that

$$\mathbb{E}(\sum_{i\in\mathcal{S}} p_{X_i}|\pi^{\text{Db+}}_{[n+1]}, \mathbf{p}) = \mathbb{E}(\sum_{j:\, n_j=1} p_{X_j^*}|\pi^{\text{Db+}}_{[n+1]}, \mathbf{p}).$$

Notice that $(n_1, n_2, ..., n_K)$ is a partition of $n+1$, which will be denoted as $\pi^{\text{Db+}}_{n+1}$. In the example, $\pi^{\text{Db+}}_{n+1} = (3, 2, 2, 1, 1, 1, 1)$. Since the distribution of $\sum_{j:\, n_j=1} p_{x_j^*}$ only depends on $\pi^{\text{Db+}}_{n+1}$, the latter can replace $\pi^{\mathrm{Db}+}_{[n+1]}$. Thus, it holds that

$$\mathrm{LR}_{|\mathbf{p}} = \frac{s_1}{\mathbb{E}(\sum_{j:\, n_j=1} p_{X^*_j} | \pi^{\mathrm{Db}+}_{n+1}, \mathbf{p})}. \tag{7.17}$$

For the same reason explained above, the knowledge of $\mathbf{p}$ and $\pi^{\mathrm{Db}+}_{n+1}$ is not enough to observe $X^*_1, ..., X^*_K$. A more compact representation for $\pi^{\mathrm{Db}+}_{n+1}$ can be obtained by using two vectors $\mathbf{a}$ and $\mathbf{r}$ where a_j are the distinct numbers occurring in the partition, ordered, and each r_j is the number of repetitions of a_j. J is the length of these two vectors, and it holds that $n+1 = \sum_{j=1}^{J} a_j r_j$. In the example above we have that $\pi^{\mathrm{Db}+}_{n+1}$ can be represented by $(\mathbf{a}, \mathbf{r})$ with $\mathbf{a} = (1, 2, 3)$ and $\mathbf{r} = (4, 2, 1)$.

There is a function, χ, treated here as latent variable, which assigns all DNA types, ordered according to their frequency in Nature, to one of the number $\{1, 2, ..., J\}$ corresponding to the position in $\mathbf{a}$ of its frequency in the sample, or to 0 if the type if not observed. Stated otherwise,

$$\chi : \{1, 2, ...\} \longrightarrow \{1, 2, ..., J\}.$$

$$\chi(i) = \begin{cases} 0 & \text{if the } i\text{th most common species is not observed in the sample,} \\ j & \text{if the } i\text{th most common species is one of the } r_j \text{ observed } a_j \text{ times in the sample.} \end{cases}$$

Given $\pi^{\mathrm{Db}+}_{n+1} = (\mathbf{a}, \mathbf{r})$, χ must satisfy the following conditions:

$$\sum_{i=1}^{\infty} \mathbf{1}_{\chi(i)=j} = r_j, \qquad \forall j. \tag{7.18}$$

The map χ can be represented by a vector $\boldsymbol{\chi} = (\chi_1, \chi_2, ...)$ such that $\chi_i = \chi(i)$. In the example above we have that $\boldsymbol{\chi} = (0, 2, 2, 3, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, ..., 0)$.

Notice that, given $\pi^{\mathrm{Db}+}_{n+1} = (\mathbf{a}, \mathbf{r})$, the knowledge of $\boldsymbol{\chi}$ implies the knowledge of X^*_1, ..., X^*_K: indeed it is enough to sort the positive values among the χ_i and take their positions in $\boldsymbol{\chi}$, and solving ties by considering the positions themselves (if $\chi_i = \chi_j$, than the order is given by i and j). For instance, in the example, if we sort the values of $\boldsymbol{\chi}$ and we collect their positions we get $(4, 2, 3, 5, 10, 13)$: the reader can notice that we got back to $X^*_1, ..., X^*_6$.

This means that to obtain the distribution of $X^*_1, ..., X^*_K | \pi^{\mathrm{Db}+}_{n+1}, \mathbf{p}$, which appears in (7.17), it is enough to obtain the distribution of $\boldsymbol{\chi} | \pi^{\mathrm{Db}+}_{n+1}, \mathbf{p}$, and since we are only interested in the mean of the sum of singletons in samples of size $n+1$ from the distribution of $X^*_1, ..., X^*_K | \pi^{\mathrm{Db}+}_{n+1}, \mathbf{p}$, we can just simulate samples from the distribution of $\boldsymbol{\chi} | \pi_{n+1}, \mathbf{p}$ and sum the p_a such that $\chi_a = 1$.

To simulate samples from the distribution of $\boldsymbol{\chi} | \pi_{n+1}, \mathbf{p}$ we use a Metropolis-Hastings algorithm, on the space of the vectors $\boldsymbol{\chi}$ satisfying condition (7.18). Notice that for the model we assumed $\mathbf{p}$ to be infinitely long, but for simulations we will use a finite $\bar{\mathbf{p}}$, of length m. This is equivalent to assume that only m elements in the infinite $\mathbf{p}$ are positive, and the remaining infinite tail is made of zeros. Then the state space of the Metropolis-Hastings Markov chain is made of all vectors of length m whose elements belong to $\{0, 1, ..., J\}$, and satisfy the condition (7.18). If we start with a initial point $\boldsymbol{\chi}_0$ which satisfies (7.18) and, at each allowed move of the Metropolis-Hastings, we swap two different values χ_a and χ_b inside the vector, condition (7.18) remains satisfied. The algorithm is based on a similar one proposed in Anevski et al. (2013).

This method allows us to obtain the 'true' LR when the vector $\mathbf{p}$ is known. This is rarely the case, but we can put ourselves in a fictitious world where we know $\mathbf{p}$, and compare the true values for the LR with the one obtained by applying our model when $\mathbf{p}$ is unknown. This will be done in the forthcoming section.

7.6 Analysis on a real database

In this section we present the study we made on a database of 18,925 Y-STR 23-loci profiles from 129 different locations in 51 countries in Europe (Purps et al., 2014)[1]. Different analyses are performed by considering only 7 Y-STR loci (DYS19, DYS389 I, DYS389 II, DYS3904, DYS3915, DY3926,DY3937) but similar results have been observed with the use of 10 loci.

First, we calculated the maximum likelihood estimators α_{MLE} and θ_{MLE} using the entire database. Their values are $\alpha_{\text{MLE}} = 0.5$ and $\theta_{\text{MLE}} = 216$.

In order to check if the two-parameter Poisson Dirichlet prior is a sensible choice we first compare the ranked frequencies from the database with the relative frequencies of several samples of size n obtained from realisations of $\text{PD}(\alpha_{MLE}, \theta_{MLE})$. The asymptotic behaviour described in (7.11) is also discussed. Lastly, we will analyse the loglikelihood function for the hyperparameters, given the data $\pi^{\text{Db+}}_{[n+1]}$, in order to a perform a data driven choice for the hyperprior.

7.6.1 Model fitting

In Figure 7.5, the ranked frequencies from the database are compared to the relative frequencies of samples of size n obtained from several realizations of $\text{PD}(\alpha_{MLE}, \theta_{MLE})$. To do so we run several times the Chinese Restaurant seating plan (up to $n = 18,925$ customers): each run is equivalent to generate a new realization $\mathbf{p}$ from the $\text{PD}(\alpha_{MLE}, \theta_{MLE})$. The partition of the customers into tables is the same as the partition obtained from an i.i.d. sample of size n from $\mathbf{p}$. The ranked relative sizes of each table (thin lines) are compared to the ranked frequencies of our database (thick line). One can see that for the most common haplotypes (left part of the plot) there is some discrepancy. However, we are interested in rare haplotypes, which typically have a frequency belonging to the right part of the plot. In that region the two-parameter Poisson Dirichlet follows the distribution of the data quite well.

[1] The database has previously been cleaned by Mikkel Meyer Andersen (`http://people.math.aau.dk/~mikl/?p=y23`).

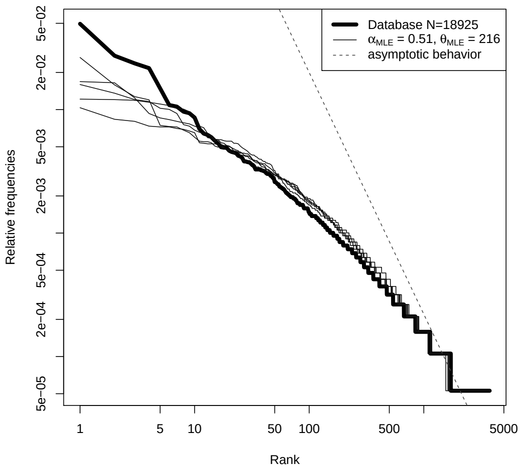


Figure 7.5: Log scale ranked frequencies from the database (thick line) are compared to the relative frequencies of samples of size n obtained from several realizations of PD$(\alpha_{MLE}, \theta_{MLE})$ (thin lines). Asymptotic power-law behavior is also displayed (dotted line).

The asymptotic behavior described in (7.11) is shown in Figure 7.5 with the dotted line. In the limit over i the thin curves are expected to bend to follow that line. This is not what we observe, but we saw from further simulation studies that we should not expect this power-law behaviour to hold at this sample size for such a big value of θ.

7.6.2 Loglikelihood

It is also interesting to investigate the shape of the loglikelihood function for α and θ given $\pi^{\mathrm{Db++}}_{[n+1]}$. It is defined as

$$l_{n+1}(\alpha, \theta) := \log p(\pi^{\mathrm{Db++}}_{[n+1]} | \alpha, \theta).$$

In Figure 7.6 the loglikelihood reparametrized using $\phi = n\dfrac{1-\alpha}{n+1+\theta}$, and θ instead of α and θ, is displayed. The Gaussian distribution is also displayed (in dashed lines). This is not done to show an asymptotic property, but to show the simmetry of the loglikelihood, which allows to approximate $\mathbb{E}(\Phi \mid \Pi^{\mathrm{Db+}}_{[n+1]} = \pi^{\mathrm{Db+}}_{[n+1]})$ with the marginal mode Φ_{MLE}, if the prior $p(\phi, \theta)$ is flat around $(\phi_{MLE}, \theta_{MLE})$, since it holds that $p(\phi, \theta \mid \pi^{\mathrm{Db+}}_{[n+1]}) \propto l_{n+1}(\phi, \theta) \times p(\phi, \theta)$.

Hence, one can approximate the LR itself in the following way:

$$\mathrm{LR} \approx \frac{n+1+\theta_{MLE}}{1-\alpha_{MLE}}. \tag{7.19}$$

Notice that this is equivalent to an hybrid approach, in which the parameters are estimated through the MLE (frequentist) and their values are plugged into the Bayesian LR.

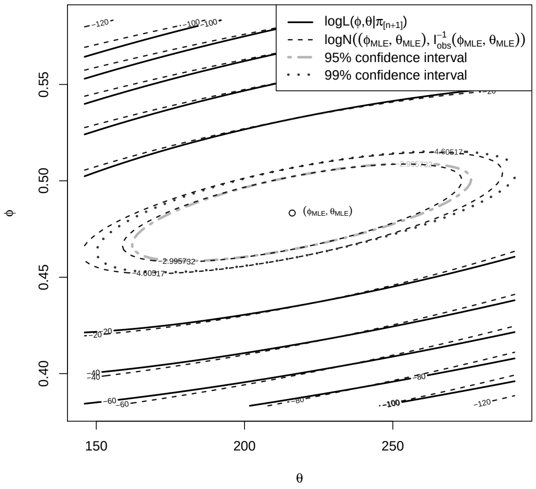


Figure 7.6: Relative loglikelihood for $\phi = n\frac{1-\alpha}{n+1+\theta}$ and θ compared to a Gaussian distribution displayed with 95% and 99% confidence intervals

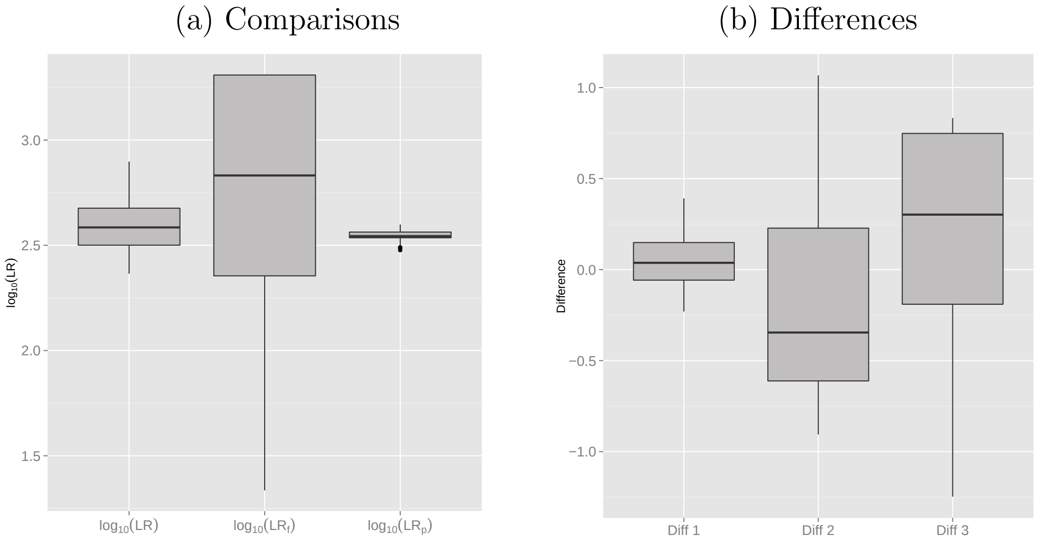
The Gaussian behavior of Figure 7.6 was unexpected. We expect that increasing n, α and θ would become independent, thus the ellipses will rotate.

7.6.3 Analyzing the error

A real Bayesian statistician chooses the prior and hyperprior according to his beliefs. Depending on the choice of the hyperprior over α and θ he may or may not believe in the approximation (7.19), but he does not really talk of ‘error’. However, hardliner Bayesian statisticians are a rare species, and most of the time the Bayesian procedure consists in choosing priors (and hyperpriors) which are a compromise between personal beliefs and mathematical convenience. It is thus interesting to investigate how good it is the choice of such priors. This can be done by comparing the Bayesian likelihood ratio with the likelihood ratio a frequentist would obtain if the vector $\mathbf{p}$ was known, and for the same reduction of data. This is what we call ‘error’: in other words, at the moment we are considering the Bayesian nonparametric method proposed in this paper as a way to estimate (notice the frequentist terminology) the true $\mathrm{LR}_{|\mathbf{p}}$. If we denote by p_x the population proportion of the matching profile, another interesting comparison is the one between the Bayesian likelihood ratio and the frequentist likelihood ratio $1/p_x$ (here denoted as LR_f) that one would obtain knowing $\mathbf{p}$, but not reducing the data to partition. This is a sort of benchmark comparison, and tells us how much we lose by using the Bayesian nonparametric methodology, and by reducing data. In order to evaluate how much we lose due to the sole reduction of the data, one can compare $\mathrm{LR}_{|\mathbf{p}}$ with LR_f. In total there are three quantities of interest ($\log_{10}\mathrm{LR}$, $\log_{10}\mathrm{LR}_{|\mathbf{p}}$, and $\log_{10}\mathrm{LR}_f$), and three differences of interest, which will be denoted as

- $\mathrm{Diff}_1 = \log_{10}\mathrm{LR} - \log_{10}\mathrm{LR}_{|\mathbf{p}}$
- $\mathrm{Diff}_2 = \log_{10}\mathrm{LR} - \log_{10}\mathrm{LR}_f$

- $\mathrm{Diff}_3 = \log_{10} \mathrm{LR}_f - \log_{10} \mathrm{LR}_{|\mathbf{p}}$


(a) Comparisons

(b) Differences

Figure 7.7: (a) comparison between the distribution of $\log_{10} \mathrm{LR}$ and $\mathrm{LR}_{|\mathbf{p}}$. (b) the error $\log_{10} \mathrm{LR}_{|\mathbf{p}} - \log_{10} \mathrm{LR}$.

In order to make the computational effort feasible, instead of using the big database of Purps et al. (2014), we consider the hapolotype frequencies for the sole Dutch population (of size 2037), and we pretend that they are the frequencies from the entire population of possible perpetrators, and we simulate the distribution of the three likelihood ratios of interest.

In Table 7.1 and Figure 7.7 (left part) we compare the distribution of $\log_{10}(\mathrm{LR}_{|\mathbf{p}})$, $\log_{10} \mathrm{LR}$, and $\mathrm{LR}_{|\mathbf{f}}$ obtained by 100 samples of size 100 from this population. The Metropolis-Hastings algorithm explained in Section 7.5.1 can be used to obtain $\mathrm{LR}_{|\mathbf{p}}$.

The distribution of the benchmark likelihood ratio ($\log_{10}(\mathrm{LR}_f)$ has more variation than the distribution of the Bayesian likelihood ratio, while $\log_{10}(\mathrm{LR}_{|\mathbf{p}})$ appears to be the most concentrated around its mean. This is probably due to the small size of the population and of the sampled databases.

In Table 7.2 and Figure 7.7 (right part) we consider the distribution of the three differences, as defined above. Diff_1 is the smallest and the most concentrated: it ranges between -0.4 and 0.23 and has a small standard deviation. It means that the nonparametric Bayesian likelihood ratio obtained as in (7.19) can be thought of as a good approximation of the frequentist likelihood ratio for the same reduction of data ($\log_{10} \mathrm{LR}_{|\mathbf{p}}$). This difference has three components: the approximation (7.19), the MLE estimation of the hyperparameters, and the choice of a prior distribution (two-parameter Poisson Dirichlet) which is quite realistic, as shown in Figure 7.5, but not perfectly fitting the actual population. Moreover, $\log_{10}(\mathrm{LR}_{|\mathbf{p}})$ is not derived exactly, but is obtained using the Metropolis Hasting approximation.

Notice that the difference increases if the Bayesian nonparametric likelihood is compared to the benchmark likelihood ratio (Diff_2). However, it ranges between -1 and +1, but most of the time the difference is between about -0.6 and 0.2, thus small.

The analysis of the distribution Diff_3 tells us that reducing data to the partitions implies a

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd
$\log_{10} \mathrm{LR}$	2.365	2.501	2.585	2.596	2.676	2.897	0.116
$\log_{10} \mathrm{LR}_{\vert\mathbf{p}}$	2.476	2.536	2.543	2.547	2.563	2.599	0.024
$\log_{10} \mathrm{LR}_f$	1.336	2.355	2.832	2.794	3.309	3.309	0.481

Table 7.1: Summaries of the distribution of $\log_{10} \mathrm{LR}$, $\log_{10}(\mathrm{LR}_{|\mathbf{p}})$, and $\log_{10} \mathrm{LR}_f$.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd
Diff_1	-0.23	-0.058	0.037	0.049	0.149	0.391	0.134
Diff_2	-0.905	-0.611	-0.346	-0.198	0.228	1.067	0.492
Diff_3	-1.247	-0.19	0.302	0.247	0.748	0.833	0.484

Table 7.2: Summaries of the distribution of Diff_1, Diff_2, and Diff_3.

loss in the capability to discriminate between the competing hypotheses of at most one order of magnitude, thus not terribly bad.

7.7 Conclusion

This paper discusses the first application of a Bayesian nonparametric method to likelihood ratio assessment in forensic science, in particular to the challenging situation of the rare type match. If compared to traditional Bayesian methods such as those described in Cereda (2016a), it presents many advantages. First of all, the prior chosen for the parameter $\mathbf{p}$ is more realistic for the population whose frequencies we want to model. Moreover, although the theoretical background on which it lies may seem very technical and difficult, the method is extremely simple to apply for practical use, thanks to the discussed approximation: indeed, simulation experiments show that an hybrid empirical approach is justified, at least using Y-STR data from European populations. The likelihood ratio obtained with this method is also compared to the frequentist likelihood ratio obtained, knowing the population frequencies of each type, both reducing and not reducing the data. The differences are quite small, reaching at most 1 order of magnitude. More could be done in the future: for instance, investigate other nonparametric priors, and use more realistic populations.

Acknowledgment

I am indebted to Jim Pitman and Alexander Gnedin for their help in understanding their important theoretical results, and to Mikkel Meyer Andersen for providing a cleaned version of the database of Purps et al. (2014). This research was supported by the Swiss National Science Foundation, through grants no. 105311-144557 and 10531A-156146, and carried out in the context of a joint research project, supervised by Franco Taroni (University of Lausanne, Ecole des sciences criminelles), and Richard Gill (Mathematical Institute, Leiden University).