A Bayesian nonparametric method for the LR assessment in case of rare type match

Giulia Cereda

October 8, 2015

Forensic Statistics

Ingredients:

Forensic Statistics

Ingredients:

- Crime case

Forensic Statistics

Ingredients:

- Crime case
- Evidence (E)

Forensic Statistics

Ingredients:

- Crime case
- Evidence (E)
- 2 Hypotheses of Interest: H_p vs H_d

Forensic Statistics

Ingredients:

- Crime case
- Evidence (E)
- 2 Hypotheses of Interest: H_p vs H_d
- Background (B)

Forensic Statistics

Ingredients:

- Crime case
- Evidence (E)
- 2 Hypotheses of Interest: H_p vs H_d
- Background (B)

D = (E,B).

Forensic Statistics

Ingredients:

- Crime case
- Evidence (E)
- 2 Hypotheses of Interest: H_p vs H_d
- Background (B)

D = (E,B).
The court asks for the likelihood ratio

Forensic Statistics

Ingredients:

- Crime case
- Evidence (E)
- 2 Hypotheses of Interest: H_p vs H_d
- Background (B)

D = (E,B).
The court asks for the likelihood ratio

$$\frac{\Pr(H_p \mid D)}{\Pr(H_d \mid D)} = \underbrace{\frac{\Pr(D \mid H_p)}{\Pr(D \mid H_d)}}_{LR} \frac{\Pr(H_p)}{\Pr(H_d)}$$

Example

Ingredients:

Example

Ingredients:

- Crime case: murder

Example

Ingredients:

- Crime case: murder
- Evidence (E): profile of the DNA trace found on the crime scene matches the suspect's DNA profile.

Example

Ingredients:

- Crime case: murder
- Evidence (E): profile of the DNA trace found on the crime scene matches the suspect's DNA profile.
- 2 Hypotheses of Interest:

Example

Ingredients:

- Crime case: murder
- Evidence (E): profile of the DNA trace found on the crime scene matches the suspect's DNA profile.
- 2 Hypotheses of Interest:
  - H_p: The suspect left the stain

Example

Ingredients:

- Crime case: murder
- Evidence (E): profile of the DNA trace found on the crime scene matches the suspect's DNA profile.
- 2 Hypotheses of Interest:
  - H_p: The suspect left the stain
  - H_d: Someone else left the stain

Example

Ingredients:

- Crime case: murder
- Evidence (E): profile of the DNA trace found on the crime scene matches the suspect's DNA profile.
- 2 Hypotheses of Interest:
  - H_p: The suspect left the stain
  - H_d: Someone else left the stain
- Background (B): database of DNA profiles from the population of possible perpetrators

DNA profiles

DNA profiles

A DNA profile is a list of integers $h = (4 - 5 - 2 - 10)$ that code some characteristics in some portions of the DNA sequence of an individual: different persons can share the same profile.

DNA profiles

A DNA profile is a list of integers $h = (4 - 5 - 2 - 10)$ that code some characteristics in some portions of the DNA sequence of an individual: different persons can share the same profile.

For H_p the match is a sure event,

DNA profiles

A DNA profile is a list of integers $h = (4 - 5 - 2 - 10)$ that code some characteristics in some portions of the DNA sequence of an individual: different persons can share the same profile.

For H_p the match is a sure event,

For H_d the match is a random event with probability p_h = frequency of the profile h of the suspect in the population of possible perpetrators.

DNA database

Database: a list of DNA profiles from a sample from the population of possible perpetrators

DNA database

Database: a list of DNA profiles from a sample from the population of possible perpetrators

DATABASE of size 10
Person 1 $(4-10-6-7)$
Person 2 $(3-5-6-8)$
Person 3 $(3-7-8-10)$
Person 4 $(10-1-4-5)$
Person 5 $(3-7-8-10)$
Person 6 $(3-7-8-10)$
Person 7 $(1-5-7-2)$
Person 8 $(3-7-8-10)$
Person 9 $(3-5-6-8)$
Person 10 $(3-7-8-10)$

DNA database

Database: a list of DNA profiles from a sample from the population of possible perpetrators

DATABASE of size 10
Person 1 $(4-10-6-7)$
Person 2 $(3-5-6-8)$
Person 3 $(3-7-8-10)$
Person 4 $(10-1-4-5)$
Person 5 $(3-7-8-10)$
Person 6 $(3-7-8-10)$
Person 7 $(1-5-7-2)$
Person 8 $(3-7-8-10)$
Person 9 $(3-5-6-8)$
Person 10 $(3-7-8-10)$

The database is used to find out the rarity of the matching profile.

LR assessment in the rare type match case

My research focuses on the LR assessment in the rare type match case, that is:

LR assessment in the rare type match case

My research focuses on the LR assessment in the rare type match case, that is:

- A match between the suspect's DNA profile and the crime stain's DNA profile.

LR assessment in the rare type match case

My research focuses on the LR assessment in the rare type match case, that is:

- A match between the suspect's DNA profile and the crime stain's DNA profile.
- This profile is not contained in the database B.

LR assessment in the rare type match case

My research focuses on the LR assessment in the rare type match case, that is:

- A match between the suspect's DNA profile and the crime stain's DNA profile.
- This profile is not contained in the database B.

Especially if the database is big, the profile seems to be rare.

LR assessment in the rare type match case

My research focuses on the LR assessment in the rare type match case, that is:

- A match between the suspect's DNA profile and the crime stain's DNA profile.
- This profile is not contained in the database B.

Especially if the database is big, the profile seems to be rare.
How rare?

Previous models

Previous models

- Frequentist model:
  (Cereda 2015) Frequentist approach to LR assessment in case of rare haplotype match
  arXiv:1502.04083

Previous models

- Frequentist model:
  (Cereda 2015) Frequentist approach to LR assessment in case of rare haplotype match
  arXiv:1502.04083
- Bayesian model:
  (Cereda 2015) Full Bayesian approach to LR assessment in case of rare haplotype match
  arXiv:1502.02406

Previous models

- Frequentist model:
  (Cereda 2015) Frequentist approach to LR assessment in case of rare haplotype match
  arXiv:1502.04083
- Bayesian model:
  (Cereda 2015) Full Bayesian approach to LR assessment in case of rare haplotype match
  arXiv:1502.02406
- (Cereda 2015) Nonparametric Bayesian approach to LR assessment in case of rare haplotype match
  arXiv:1506.08444

Assumptions

Assumption 1

There are so many different DNA types that they may be considered infinite.

Assumptions

Assumption 1

There are so many different DNA types that they may be considered infinite.

Parameter: $\mathbf{p} = (p_t | t \in T)$, T an infinite countable set, $p_t > 0$, $\sum p_t = 1$, to represent the (unknown) frequencies of all DNA types in Nature.

Assumptions

Assumption 1

There are so many different DNA types that they may be considered infinite.

Parameter: $\mathbf{p} = (p_t | t \in T)$, T an infinite countable set, $p_t > 0$, $\sum p_t = 1$, to represent the (unknown) frequencies of all DNA types in Nature.

Assumption 2

The particular list of integers that forms a DNA type is just a category: no structure assumed.

Assumptions

Assumption 1

There are so many different DNA types that they may be considered infinite.

Parameter: $\mathbf{p} = (p_t | t \in T)$, T an infinite countable set, $p_t > 0$, $\sum p_t = 1$, to represent the (unknown) frequencies of all DNA types in Nature.

Assumption 2

The particular list of integers that forms a DNA type is just a category: no structure assumed.

"DNA types" or "colors" is now the same.

Random partitions of $[n]$

Let $[n]$ denote the set $[n] = \{1, 2, ..., n\}$.

Random partitions of $[n]$

Let $[n]$ denote the set $[n] = \{1, 2, ..., n\}$.
A partition of the set $[n]$ will be denoted as $\pi_{[n]}$.

Random partitions of $[n]$

Let $[n]$ denote the set $[n] = \{1, 2, ..., n\}$.
A partition of the set $[n]$ will be denoted as $\pi_{[n]}$.
Random partitions on the set $[n]$ will be denoted as $\Pi_{[n]}$.

DNA database can be reduced

DATABASE of size 10
Person 1 $(4-10-6-7)$
Person 2 $(3-5-6-8)$
Person 3 $(3-7-8-10)$
Person 4 $(10-1-4-5)$
Person 5 $(3-7-8-10)$
Person 6 $(3-7-8-10)$
Person 7 $(1-5-7-2)$
Person 8 $(3-7-8-10)$
Person 9 $(3-5-6-8)$
Person 10 $(3-7-8-10)$

DNA database can be reduced

DATABASE of size 10
Person 1 $(4-10-6-7)$
Person 2 $(3-5-6-8)$
Person 3 $(3-7-8-10)$
Person 4 $(10-1-4-5)$
Person 5 $(3-7-8-10)$
Person 6 $(3-7-8-10)$
Person 7 $(1-5-7-2)$
Person 8 $(3-7-8-10)$
Person 9 $(3-5-6-8)$
Person 10 $(3-7-8-10)$

DNA database can be reduced

DATABASE of size 10
Person 1 $(4-10-6-7)$
Person 2 $(3-5-6-8)$
Person 3 $(3-7-8-10)$
Person 4 $(10-1-4-5)$
Person 5 $(3-7-8-10)$
Person 6 $(3-7-8-10)$
Person 7 $(1-5-7-2)$
Person 8 $(3-7-8-10)$
Person 9 $(3-5-6-8)$
Person 10 $(3-7-8-10)$
Assumption 2 $\rightarrow$ data can be replaces by the equivalence classes on the indices of the relation "to have the same DNA type".
This is a partition of the set $[n]$: $\{\{1\},\{2,9\},\{3,5,6,8,10\},\{4\},\{7\}\}$

Reduced data

Reduced data

Data $\mathcal{D}$ is made of the database + 2 new observations

Reduced data

Data $\mathcal{D}$ is made of the database + 2 new observations

$\mathcal{D} = \pi_{[n+2]}$ partition of the set $\{1, 2, ..., n+2\}$

Reduced data

Data $\mathcal{D}$ is made of the database + 2 new observations

$\mathcal{D} = \pi_{[n+2]}$ partition of the set $\{1, 2, ..., n+2\}$

Example:
Database $\rightarrow \pi_{[10]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}\}$

Reduced data

Data $\mathcal{D}$ is made of the database + 2 new observations

$\mathcal{D} = \pi_{[n+2]}$ partition of the set $\{1, 2, ..., n+2\}$

Example:
Database $\rightarrow \pi_{[10]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}\}$
$\mathcal{D} \rightarrow \pi_{[12]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}, \{11, 12\}\}$

Reduced data

Data $\mathcal{D}$ is made of the database + 2 new observations

$\mathcal{D} = \pi_{[n+2]}$ partition of the set $\{1, 2, ..., n+2\}$

Example:
Database $\rightarrow \pi_{[10]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}\}$
$\mathcal{D} \rightarrow \pi_{[12]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}, \{11, 12\}\}$

We can see the data as a random variable. In that case,

$$\mathcal{D} = \Pi_{[n+2]}.$$

The distribution of $\mathcal{D} = \Pi_{[n+2]}$ depends on $\mathbf{p}$. However, it does not depend on the order of the p_i.

The distribution of $\mathcal{D} = \Pi_{[n+2]}$ depends on $\mathbf{p}$. However, it does not depend on the order of the p_i.

$$\downarrow$$

We can consider directly the ordered vector
$\mathbf{p} \in \nabla_\infty = \{(p_1, p_2, ...), p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$.

The distribution of $\mathcal{D} = \Pi_{[n+2]}$ depends on $\mathbf{p}$. However, it does not depend on the order of the p_i.

$$\downarrow$$

We can consider directly the ordered vector
$\mathbf{p} \in \nabla_\infty = \{(p_1, p_2, ...), p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$.

For instance, p_3= the frequency of the third most frequent DNA type in Nature.

Prior distribution on $\mathbf{p} \in \nabla_\infty$

Bayesian nonparametrics: we need a prior for the parameter $\mathbf{p}$.

Prior distribution on $\mathbf{p} \in \nabla_\infty$

Bayesian nonparametrics: we need a prior for the parameter $\mathbf{p}$.
Two parameter Poisson Dirichlet distribution.

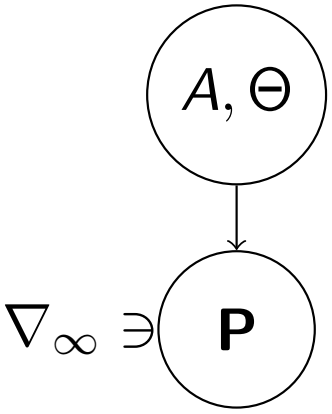
Prior distribution on $\mathbf{p} \in \nabla_\infty$

Bayesian nonparametrics: we need a prior for the parameter **p**.
Two parameter Poisson Dirichlet distribution.

Parameters:
$0 < \alpha < 1$, $\theta > -\alpha$

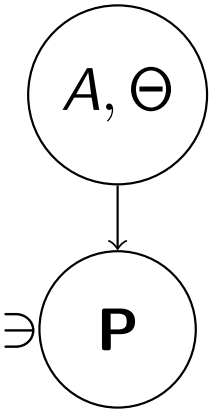
The model (first part)

The model (first part)

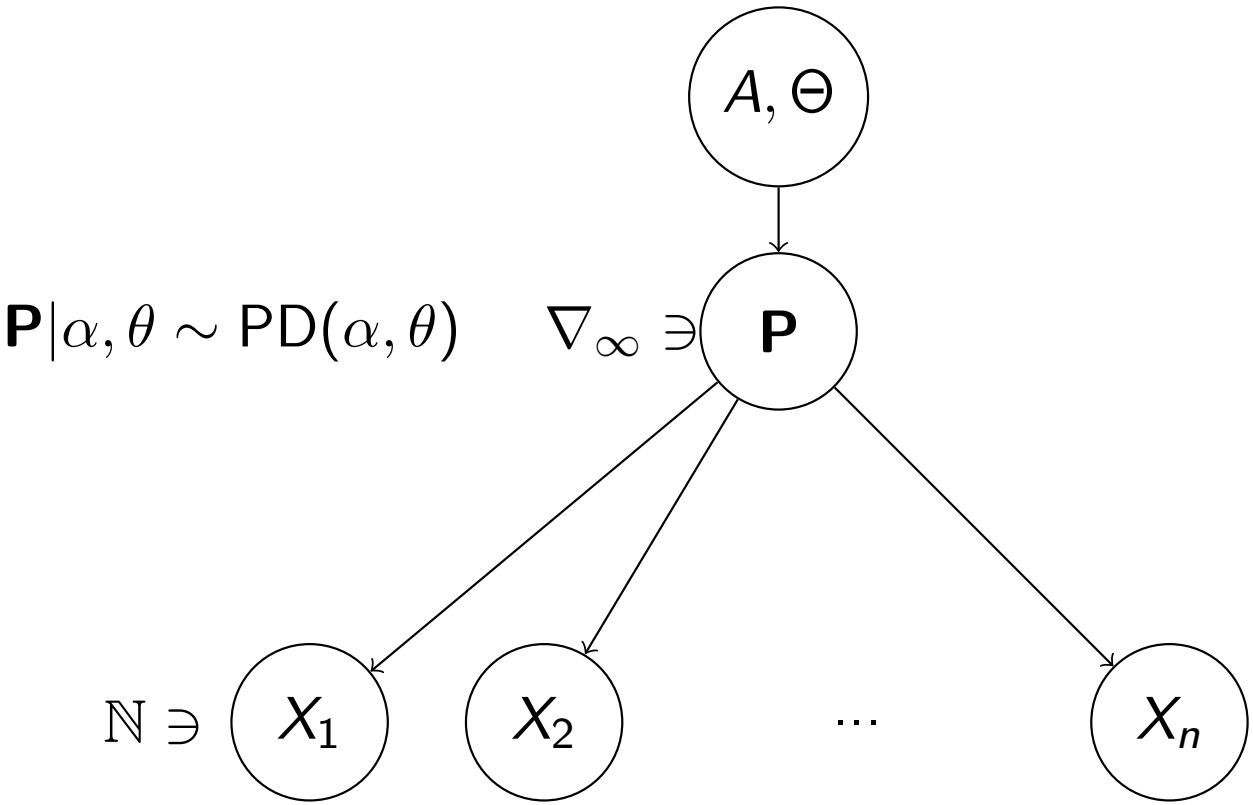


The model (first part)

$$\mathbf{P}|\alpha,\theta \sim \mathrm{PD}(\alpha,\theta) \qquad \nabla_\infty \ni$$

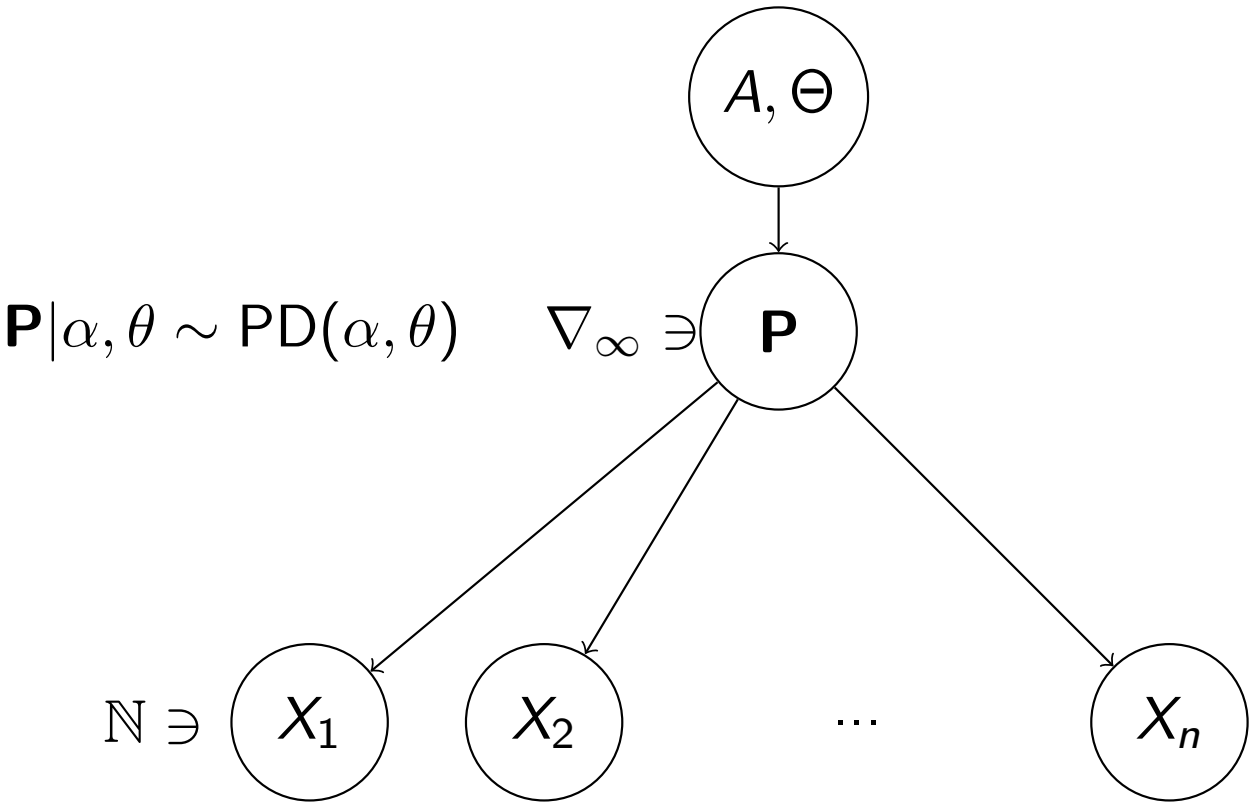


The model (first part)



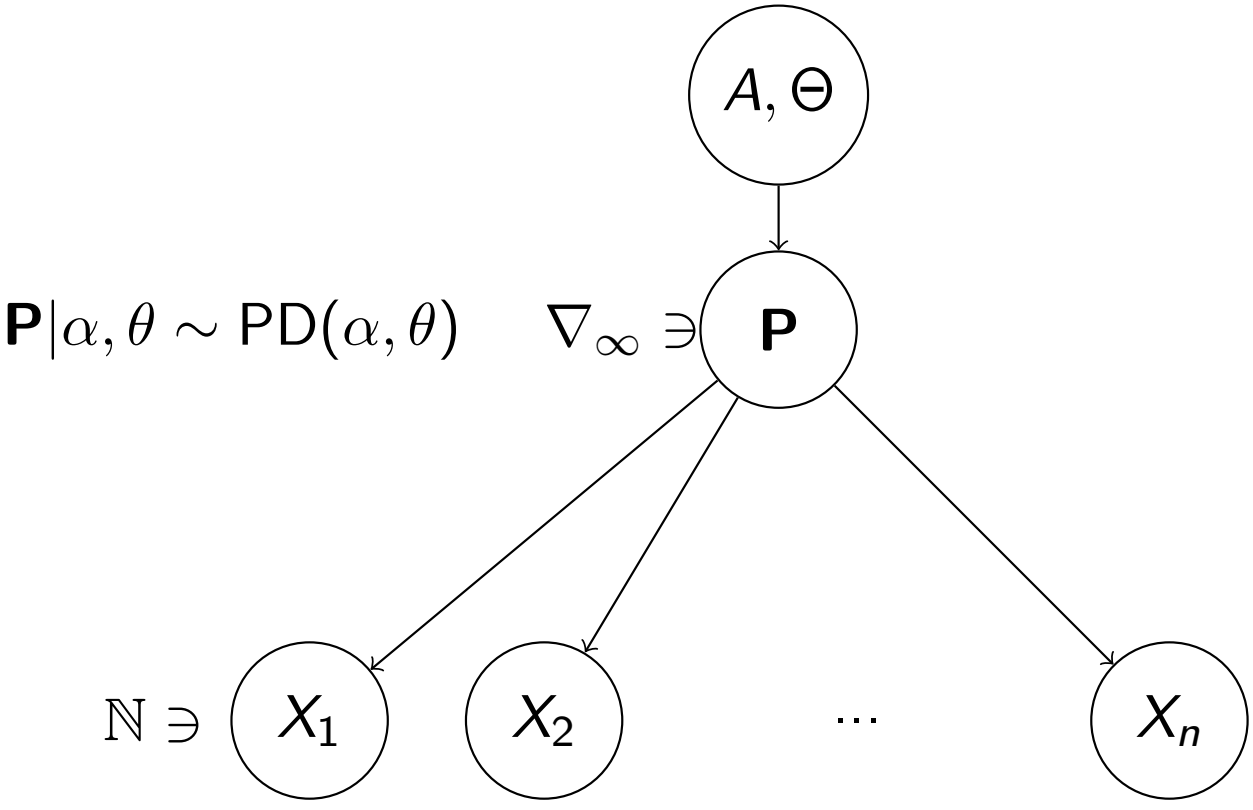
$X_i = j \rightarrow$ the i-th observation has the jth most common type in Nature.

The model (first part)

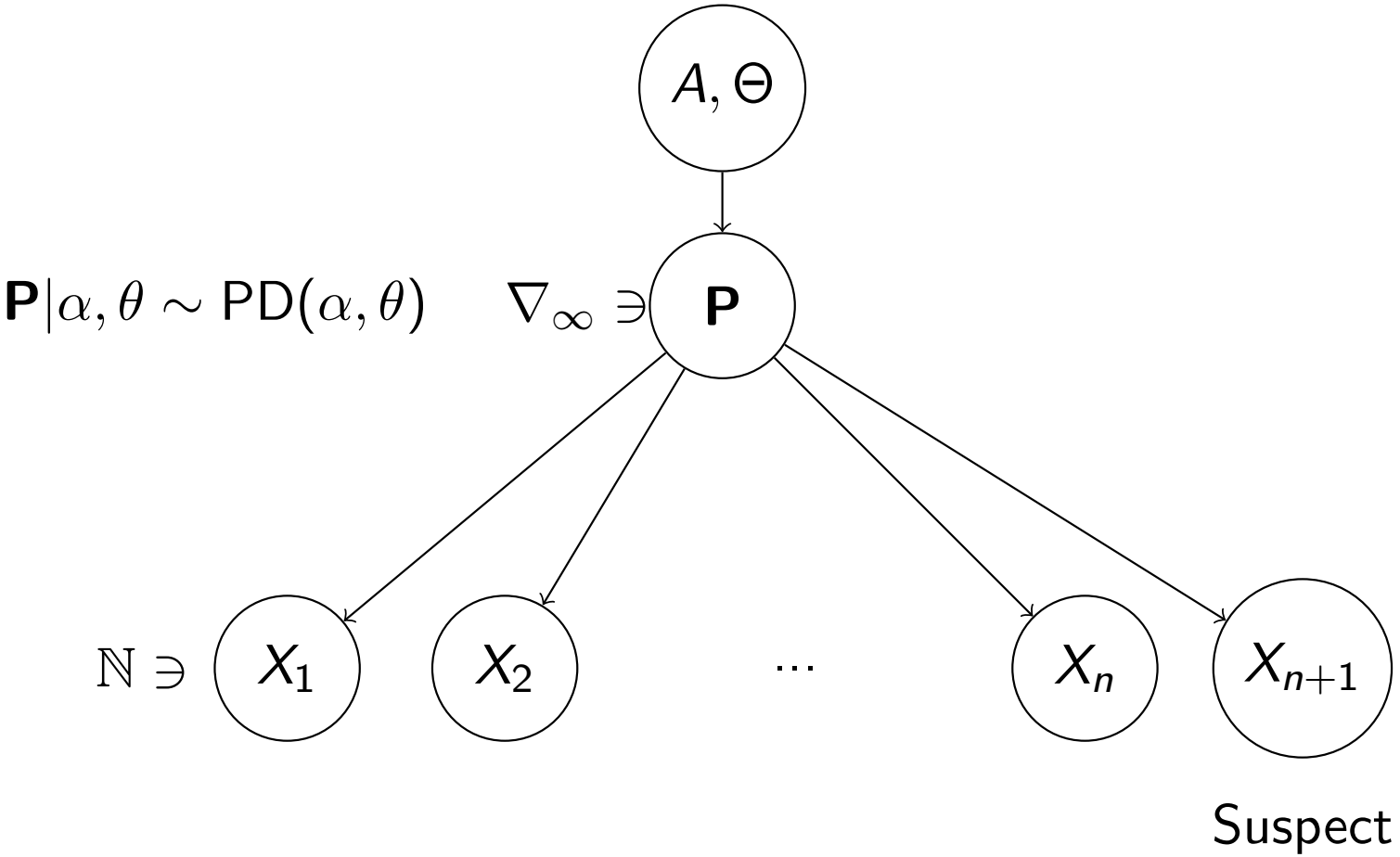


$X_i = j \rightarrow$ the i-th observation has the jth most common type in Nature.

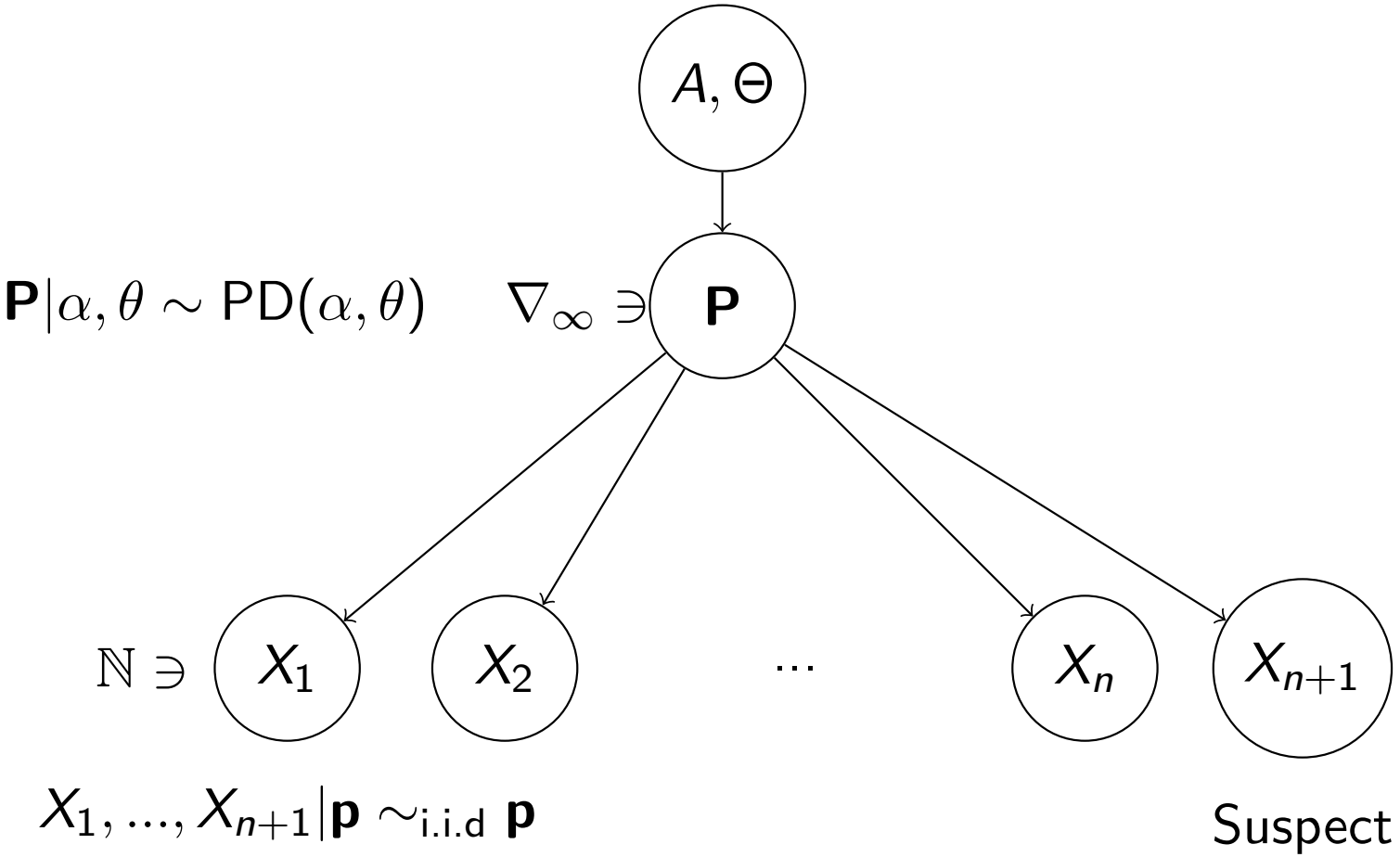
The model (first part)



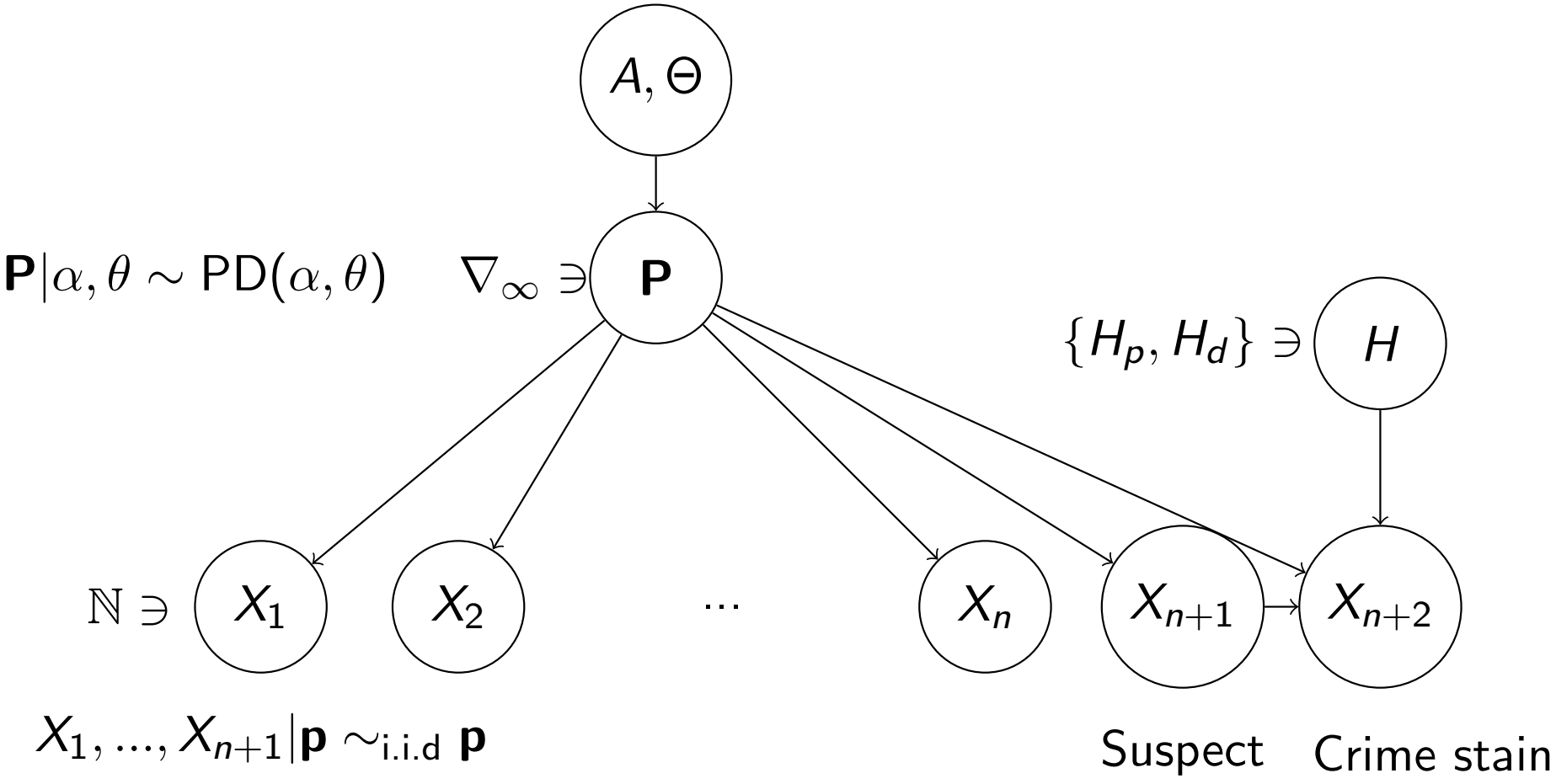
The model (first part)



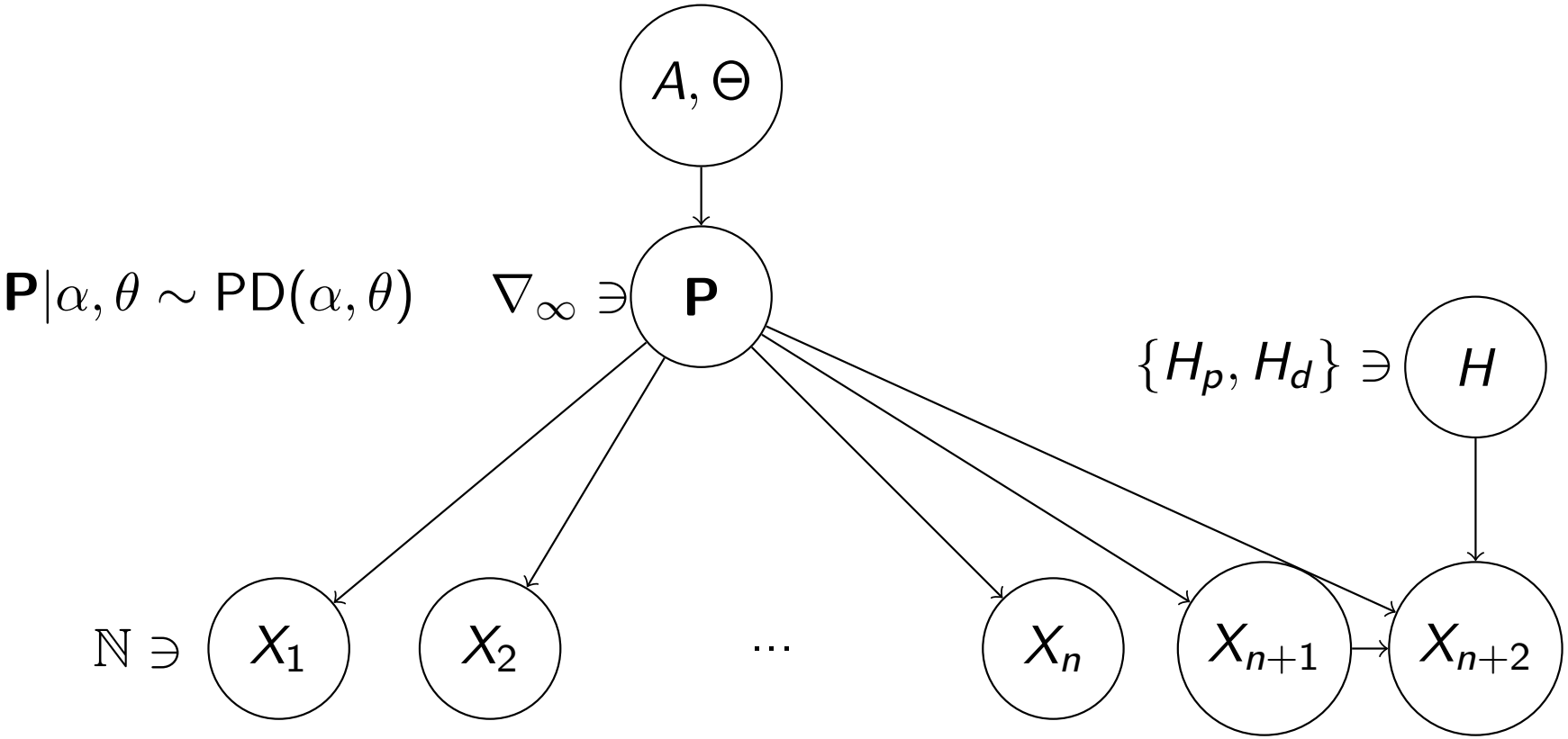
The model (first part)



The model (first part)



The model (first part)

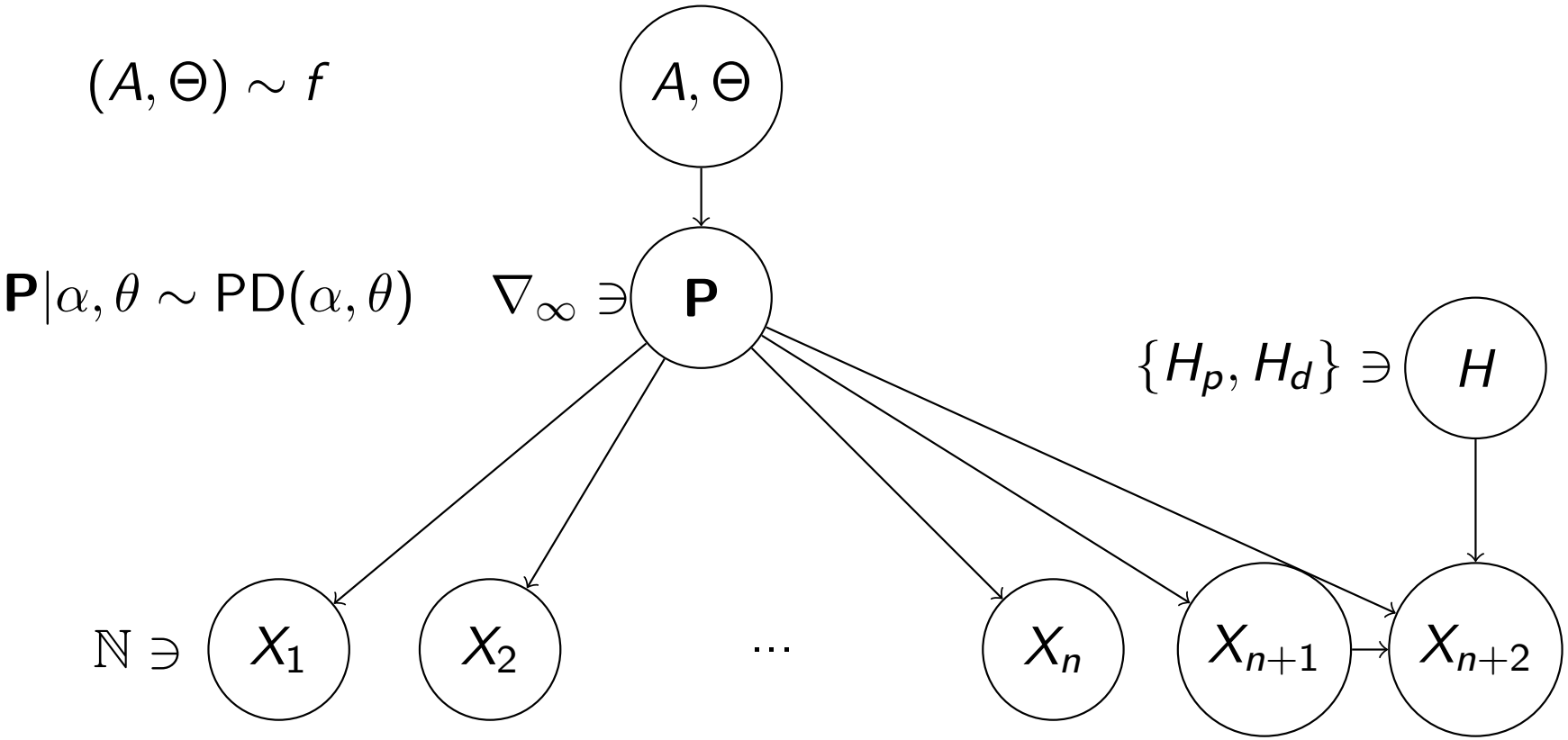


$X_1, ..., X_{n+1}|\mathbf{p} \sim_{\text{i.i.d}} \mathbf{p}$

Suspect Crime stain

$$X_{n+2}|\mathbf{p}, H, x_{n+1} \sim \begin{cases} \delta_{x_{n+1}} & \text{if } H = H_p \\ \mathbf{p} & \text{if } H = H_d \end{cases}$$

The model (first part)

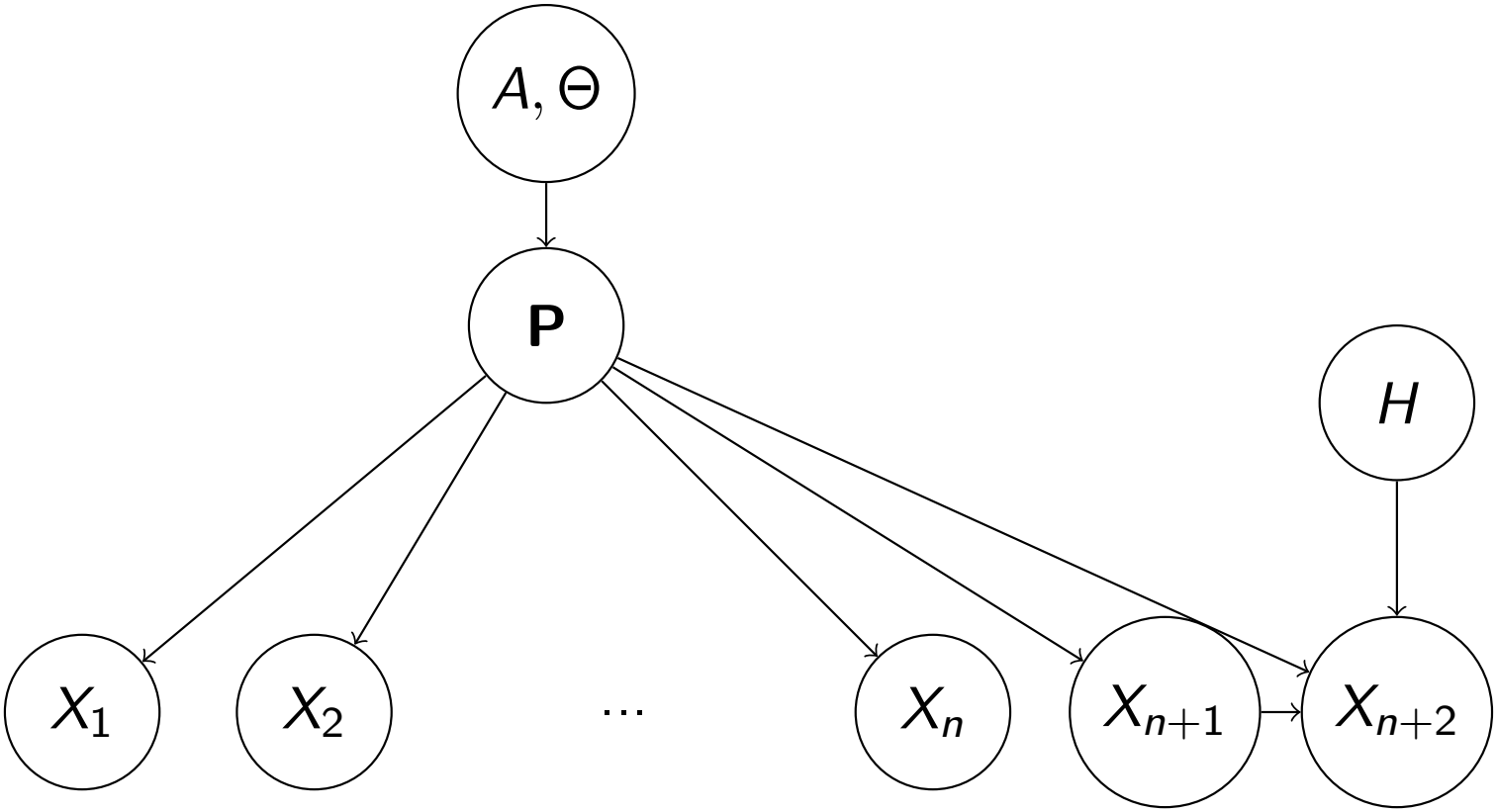


$X_1, ..., X_{n+1}|\mathbf{p} \sim_{\text{i.i.d}} \mathbf{p}$

Suspect Crime stain

$$X_{n+2}|\mathbf{p}, H, x_{n+1} \sim \begin{cases} \delta_{x_{n+1}} & \text{if } H = H_p \\ \mathbf{p} & \text{if } H = H_d \end{cases}$$

The model (first part)



Random partitions

Some notation:
Given $X_1, ..., X_n \in \mathbb{N}$, random variables, $\Pi_{[n]}(X_1, X_2, ..., X_n)$ is the random partition defined by the equivalence classes of $i \sim j$ iff $X_i = X_j$.

Random partitions

Some notation:
Given $X_1, ..., X_n \in \mathbb{N}$, random variables, $\Pi_{[n]}(X_1, X_2, ..., X_n)$ is the random partition defined by the equivalence classes of $i \sim j$ iff $X_i = X_j$.

$$
\begin{aligned}
X_1, ..., X_n &\longrightarrow \Pi_{[n]} = \pi^{\mathrm{Db}}_{[n]} \\
X_1, ..., X_n, X_{n+1} &\longrightarrow \Pi_{[n+1]} = \pi^{\mathrm{Db}+}_{[n+1]} \\
X_1, ..., X_n, X_{n+1}, X_{n+2} &\longrightarrow \Pi_{[n+2]} = \pi^{\mathrm{Db}++}_{[n+2]}
\end{aligned}
$$

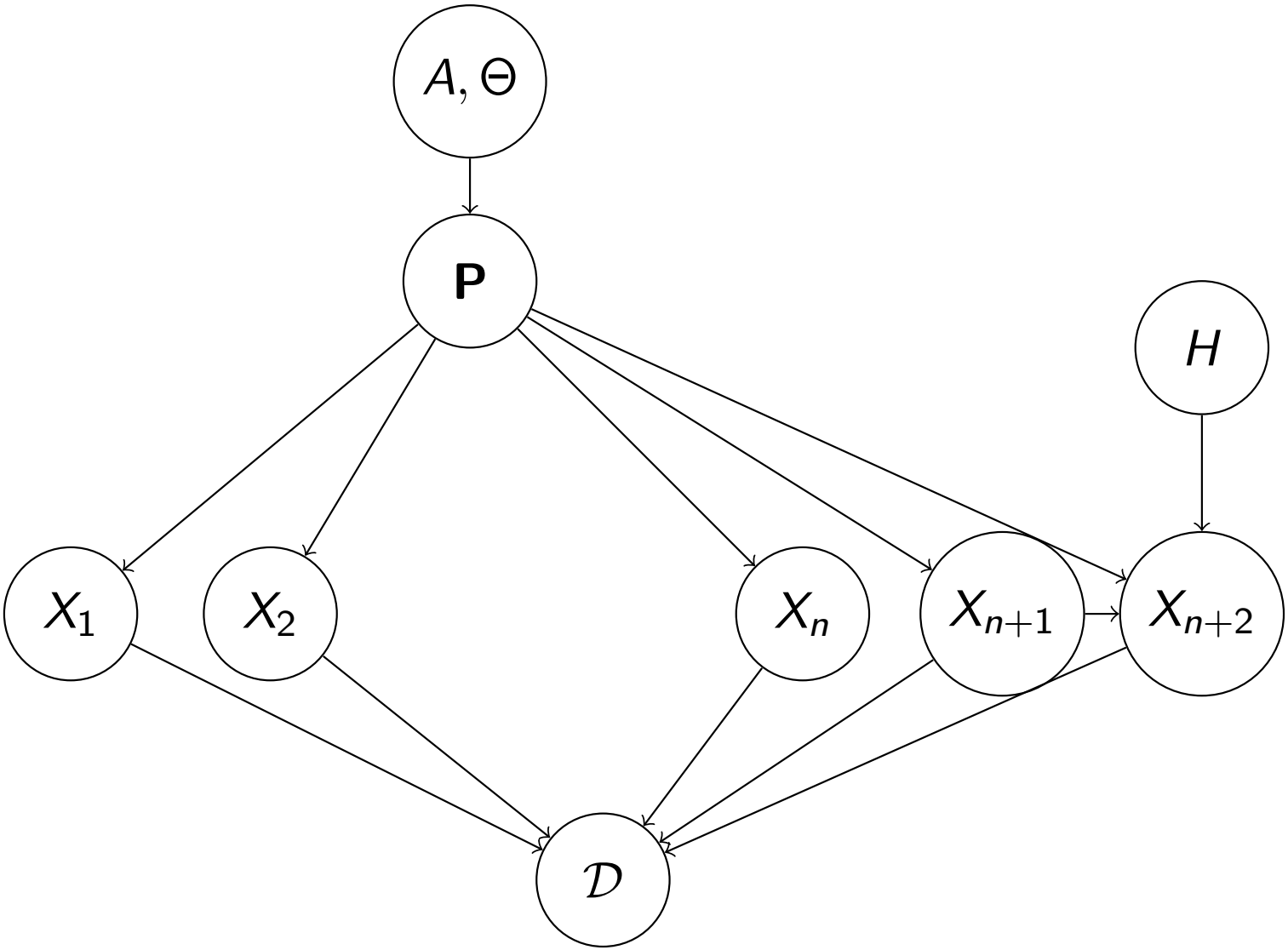
Random partitions

Some notation:
Given $X_1, ..., X_n \in \mathbb{N}$, random variables, $\Pi_{[n]}(X_1, X_2, ..., X_n)$ is the random partition defined by the equivalence classes of $i \sim j$ iff $X_i = X_j$.

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

$X_1, ..., X_n$ are not observed, but generates the same partition as the original database.
Data can be defined as $\mathcal{D} = \Pi_{[n+2]}$.

The complete model



Pitman sampling formula

Pitman sampling formula

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

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

Pitman sampling formula

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

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

then $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the following distribution:

Pitman sampling formula

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

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

then $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the following distribution:

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

Pitman sampling formula

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

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

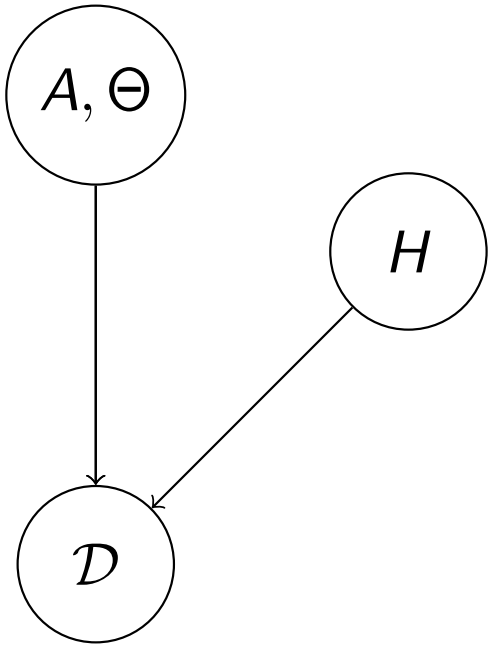
then $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the following distribution:

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

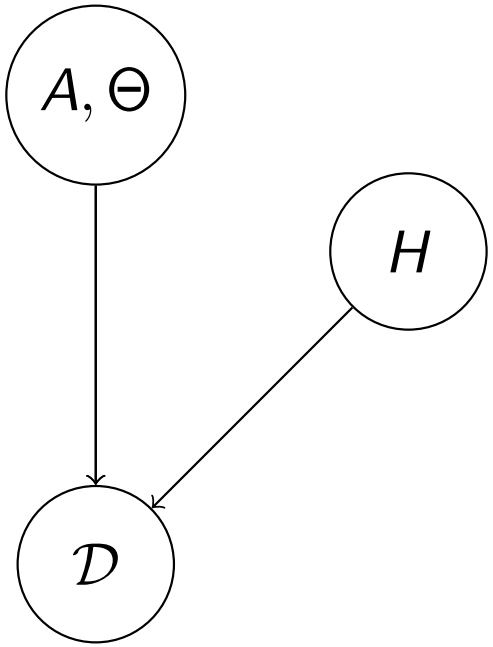
In our model

$$\Pr(D|\alpha, \theta, h) = \Pr(\Pi_{[n+2]} = \pi^{Db++}_{[n+2]} | \alpha, \theta, h) = \begin{cases} \mathbb{P}^{n+2}_{\alpha,\theta}(\pi^{\text{Db++}}_{[n+2]}) & \text{if } h = H_d \\ \mathbb{P}^{n+1}_{\alpha,\theta}(\pi^{\text{Db+}}_{[n+1]}) & \text{if } h = H_p \end{cases}$$

The model, simplified



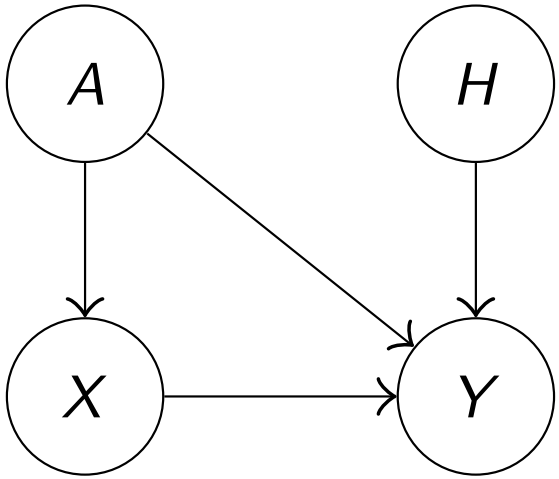
The model, simplified



$\mathcal{D} = \Pi_{[n+2]}$.

Lemma

Lemma

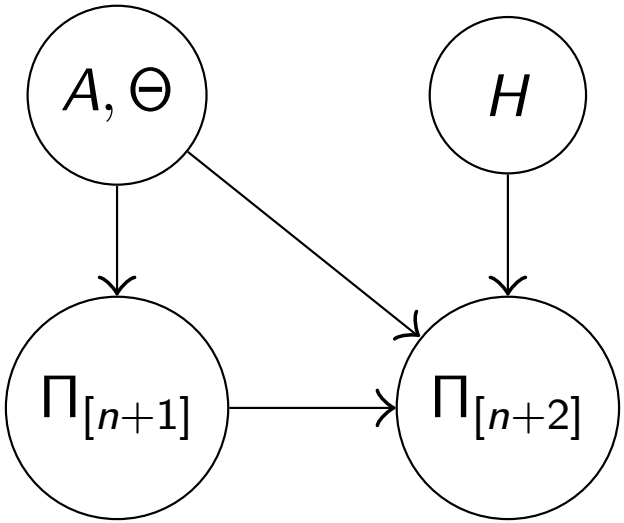


Lemma

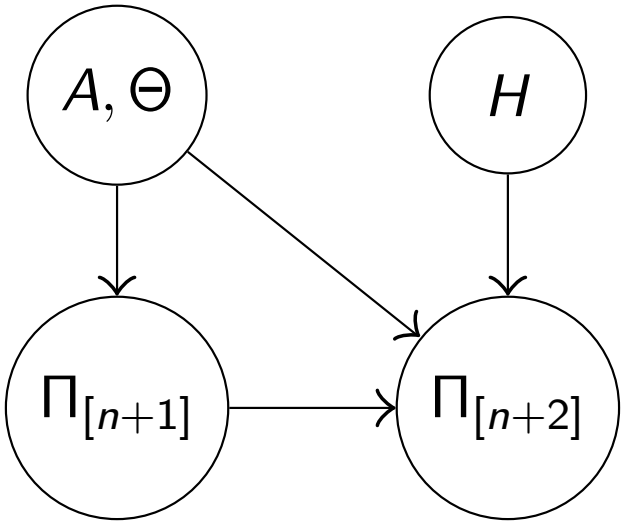
Given four random variables A, H, X and Y, as above, the likelihood function for h, given $X = x$ and $Y = y$, satisfies

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

Lemma



Lemma



$$\text{lik}(h \mid \pi_{[n+1]}, \pi_{[n+2]}) \propto \mathbb{E}(p(\pi_{[n+2]} \mid \pi_{[n+1]}, A, \Theta, h) \mid \Pi_{[n+1]} = \pi_{[n+1]}).$$

Likelihood ratio

$$\text{LR} = \frac{p(\pi_{[n+2]}|H_p)}{p(\pi_{[n+2]}|H_d)} = \frac{p(\pi_{[n+1]}, \pi_{[n+2]}|H_p)}{p(\pi_{[n+1]}, \pi_{[n+2]}|H_d)} = \frac{\text{lik}(H_p|\pi_{[n+1]}, \pi_{[n+2]})}{\text{lik}(H_d|\pi_{[n+1]}, \pi_{[n+2]})}$$

Likelihood ratio

$$\text{LR} = \frac{p(\pi_{[n+2]}|H_p)}{p(\pi_{[n+2]}|H_d)} = \frac{p(\pi_{[n+1]}, \pi_{[n+2]}|H_p)}{p(\pi_{[n+1]}, \pi_{[n+2]}|H_d)} = \frac{\text{lik}(H_p|\pi_{[n+1]}, \pi_{[n+2]})}{\text{lik}(H_d|\pi_{[n+1]}, \pi_{[n+2]})}$$

Lemma allows to write

$$\text{LR} = \frac{\mathbb{E}(\overbrace{p(\pi_{[n+2]} \mid \pi_{[n+1]}, A, \Theta, H_p)}^{1} \mid \Pi_{[n+1]} = \pi_{[n+1]})}{\mathbb{E}(\underbrace{p(\pi_{[n+2]} \mid \pi_{[n+1]}, A, \Theta, H_d)}_{\frac{1-A}{n+1+\Theta}} \mid \Pi_{[n+1]} = \pi_{[n+1]})}$$

Likelihood ratio

$$\text{LR} = \frac{p(\pi_{[n+2]}|H_p)}{p(\pi_{[n+2]}|H_d)} = \frac{p(\pi_{[n+1]}, \pi_{[n+2]}|H_p)}{p(\pi_{[n+1]}, \pi_{[n+2]}|H_d)} = \frac{\text{lik}(H_p|\pi_{[n+1]}, \pi_{[n+2]})}{\text{lik}(H_d|\pi_{[n+1]}, \pi_{[n+2]})}$$

Lemma allows to write

$$\text{LR} = \frac{\mathbb{E}(\overbrace{p(\pi_{[n+2]} \mid \pi_{[n+1]}, A, \Theta, H_p)}^{1} \mid \Pi_{[n+1]} = \pi_{[n+1]})}{\mathbb{E}(\underbrace{p(\pi_{[n+2]} \mid \pi_{[n+1]}, A, \Theta, H_d)}_{\frac{1-A}{n+1+\Theta}} \mid \Pi_{[n+1]} = \pi_{[n+1]})}$$

$$= \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]}\right)}.$$

Likelihood ratio

$$\text{LR} = \frac{p(\pi_{[n+2]}|H_p)}{p(\pi_{[n+2]}|H_d)} = \frac{p(\pi_{[n+1]}, \pi_{[n+2]}|H_p)}{p(\pi_{[n+1]}, \pi_{[n+2]}|H_d)} = \frac{\text{lik}(H_p|\pi_{[n+1]}, \pi_{[n+2]})}{\text{lik}(H_d|\pi_{[n+1]}, \pi_{[n+2]})}$$

Lemma allows to write

$$\text{LR} = \frac{\mathbb{E}(\overbrace{p(\pi_{[n+2]} \mid \pi_{[n+1]}, A, \Theta, H_p)}^{1} \mid \Pi_{[n+1]} = \pi_{[n+1]})}{\mathbb{E}(\underbrace{p(\pi_{[n+2]} \mid \pi_{[n+1]}, A, \Theta, H_d)}_{\frac{1-A}{n+1+\Theta}} \mid \Pi_{[n+1]} = \pi_{[n+1]})}$$

$$= \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]}\right)}.$$

$$\mathrm{LR} = \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]}\right)}$$

$$\mathrm{LR} = \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]}\right)}$$

By defining the random variable $\Phi = n\frac{1-A}{n+1+\Theta}$ we can write the LR as

$$\mathrm{LR} = \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]}\right)}$$

By defining the random variable $\Phi = n\frac{1-A}{n+1+\Theta}$ we can write the LR as

$$\mathrm{LR} = \frac{n}{\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]})}.$$

$$\mathrm{LR} = \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]}\right)}$$

By defining the random variable $\Phi = n\frac{1-A}{n+1+\Theta}$ we can write the LR as

$$\mathrm{LR} = \frac{n}{\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]})}.$$

We are interested in the distribution of $\Phi, \Theta | \Pi_{[n+1]}$

$$\mathrm{LR} = \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]}\right)}$$

By defining the random variable $\Phi = n\frac{1-A}{n+1+\Theta}$ we can write the LR as

$$\mathrm{LR} = \frac{n}{\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]})}.$$

We are interested in the distribution of $\Phi, \Theta | \Pi_{[n+1]}$

$$p(\phi, \theta \mid \pi_{[n+1]}) \propto p(\pi_{[n+1]} \mid \phi, \theta) f(\phi, \theta)$$

$$\mathrm{LR} = \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]}\right)}$$

By defining the random variable $\Phi = n\frac{1-A}{n+1+\Theta}$ we can write the LR as

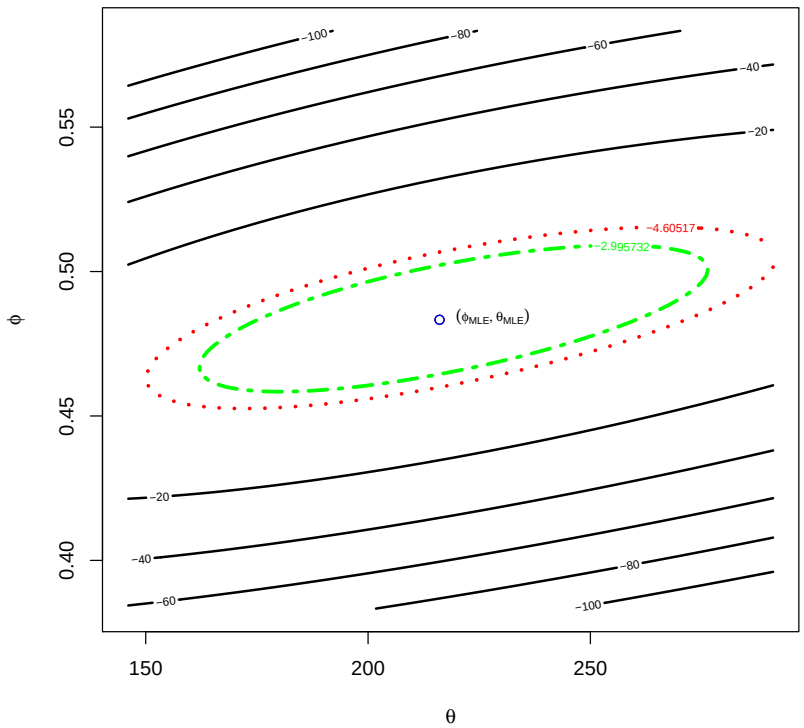
$$\mathrm{LR} = \frac{n}{\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]})}.$$

We are interested in the distribution of $\Phi, \Theta | \Pi_{[n+1]}$

$$p(\phi, \theta \mid \pi_{[n+1]}) \propto p(\pi_{[n+1]} \mid \phi, \theta) f(\phi, \theta)$$

Log likelihood with ϕ and θ

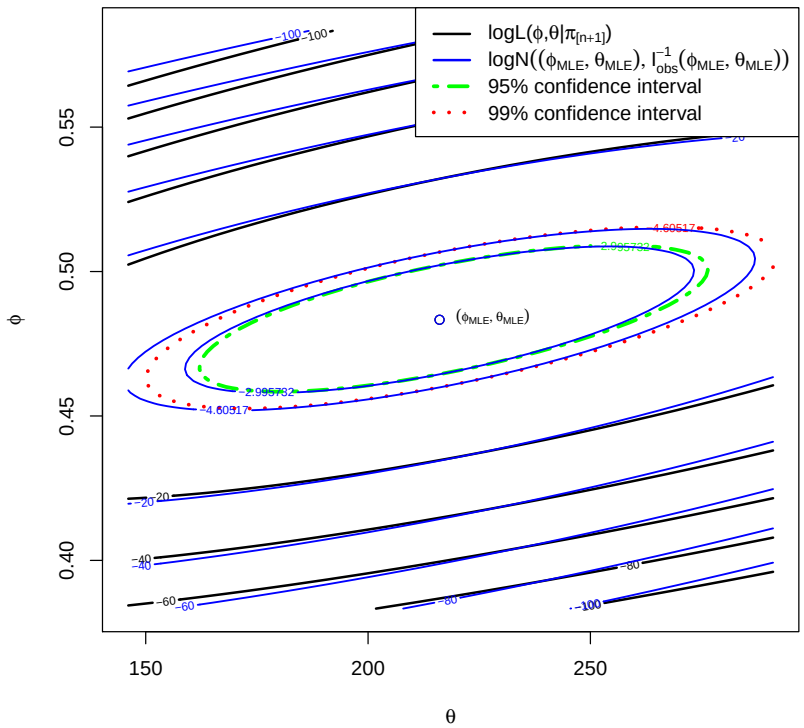
$\log_{10} p(\pi_{[n+1]} \mid \phi, \theta)$



Dutch Y-STR database, 7 loci, N=18,925

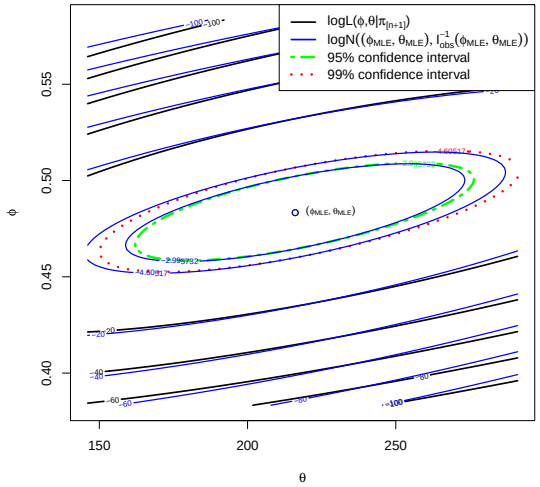
Log likelihood with ϕ and θ

$\log_{10} p(\pi_{[n+1]} \mid \phi, \theta)$



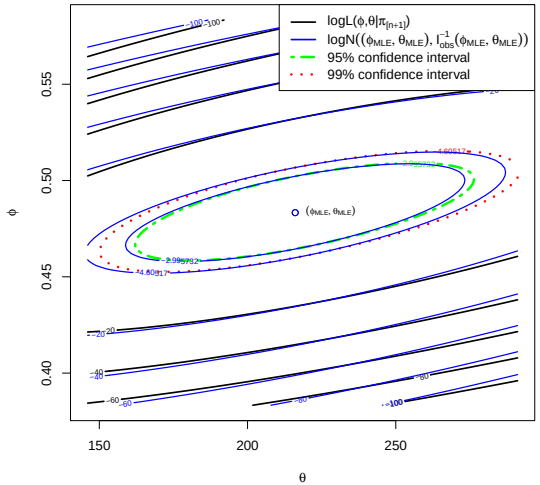
Dutch Y-STR database, 7 loci, N=18,925

Log likelihood as a function of ϕ and θ



$$p(\pi_{[n+1]} \mid \phi, \theta) \approx N((\phi_{MLE}, \theta_{MLE}), I^{-1}_{MLE})$$

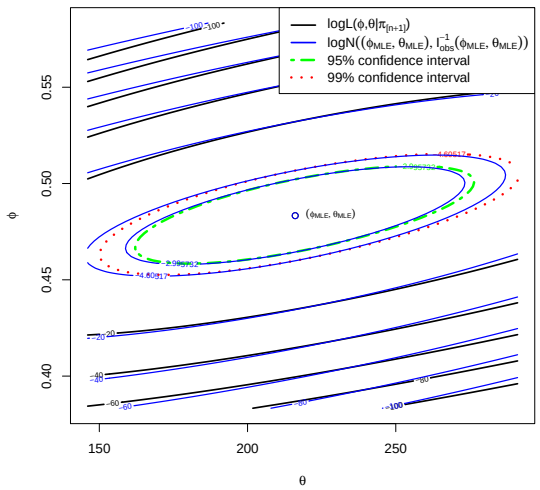
Log likelihood as a function of ϕ and θ



$$p(\pi_{[n+1]} \mid \phi, \theta) \approx N((\phi_{MLE}, \theta_{MLE}), I^{-1}_{MLE})$$

$$p(\phi, \theta \mid \pi_{[n+1]}) \propto p(\pi_{[n+1]} \mid \phi, \theta) f(\phi, \theta)$$

Log likelihood as a function of ϕ and θ

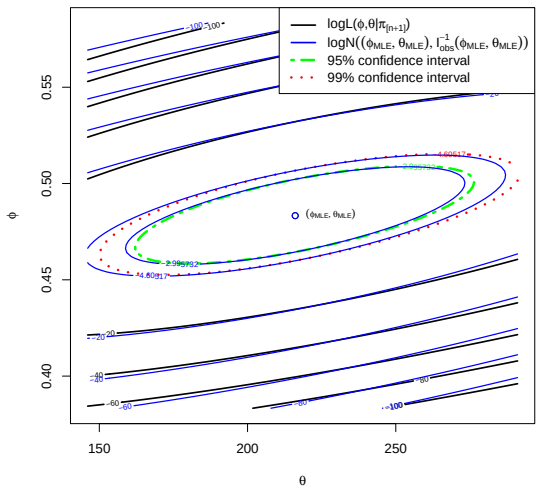


$$p(\pi_{[n+1]} \mid \phi, \theta) \approx N((\phi_{MLE}, \theta_{MLE}), I_{MLE}^{-1})$$

$$p(\phi, \theta \mid \pi_{[n+1]}) \propto p(\pi_{[n+1]} \mid \phi, \theta) f(\phi, \theta)$$

If the prior is smooth around the MLE then

Log likelihood as a function of ϕ and θ



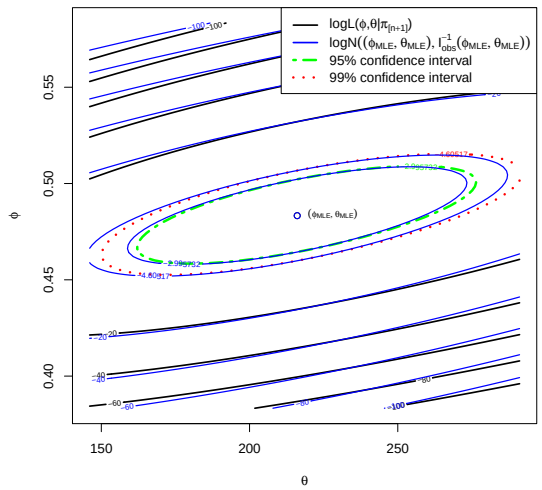
$$p(\pi_{[n+1]} \mid \phi, \theta) \approx N((\phi_{MLE}, \theta_{MLE}), I_{MLE}^{-1})$$

$$p(\phi, \theta \mid \pi_{[n+1]}) \propto p(\pi_{[n+1]} \mid \phi, \theta) f(\phi, \theta)$$

If the prior is smooth around the MLE then

$$p(\phi, \theta \mid \pi_{[n+1]}) \approx N((\phi_{MLE}, \theta_{MLE}), I_{MLE}^{-1}).$$

Log likelihood as a function of ϕ and θ



$$p(\pi_{[n+1]} \mid \phi, \theta) \approx N((\phi_{MLE}, \theta_{MLE}), I_{MLE}^{-1})$$

$$p(\phi, \theta \mid \pi_{[n+1]}) \propto p(\pi_{[n+1]} \mid \phi, \theta) f(\phi, \theta)$$

If the prior is smooth around the MLE then

$$p(\phi, \theta \mid \pi_{[n+1]}) \approx N((\phi_{MLE}, \theta_{MLE}), I_{MLE}^{-1}).$$

$$p(\phi, \theta \mid \pi_{[n+1]}) \approx N((\phi_{MLE}, \theta_{MLE}), I^{-1}_{MLE}).$$

$$p(\phi, \theta \mid \pi_{[n+1]}) \approx N((\phi_{MLE}, \theta_{MLE}), I^{-1}_{MLE}).$$

It follows that $\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]}) \approx \phi_{MLE}$.

$$p(\phi, \theta \mid \pi_{[n+1]}) \approx N((\phi_{MLE}, \theta_{MLE}), I^{-1}_{MLE}).$$

It follows that $\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]}) \approx \phi_{MLE}$.

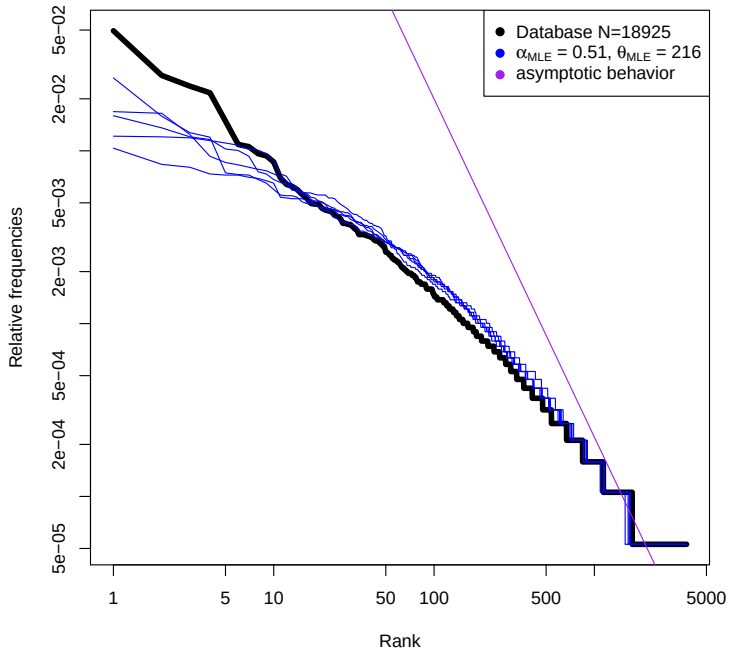
$$\text{LR} = \frac{n}{\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]})} \approx \frac{n + 1 + \theta_{MLE}}{1 - \alpha_{MLE}}$$

Sorted relative frequencies: how good is our prior?

Comparison between the spectrum from a big database, and simulations from PD(α, θ) using MLE estimators of the parameters.

Sorted relative frequencies: how good is our prior?

Comparison between the spectrum from a big database, and simulations from PD(α, θ) using MLE estimators of the parameters.



Thick black line: ranked relative frequencies in the database.
Thin black lines: simulations from the PD($\alpha_{MLE}, \theta_{MLE}$).
Dotted line: asymptotics.

The LR when $\mathbf{p}$ is known

Imagine we know $\mathbf{p}$.

The LR when $\mathbf{p}$ is known

Imagine we know $\mathbf{p}$.

$$\mathrm{LR}_{|\mathbf{p}} = \frac{p(\pi^{\mathrm{Db}++}_{[n+2]}|H_p, \mathbf{p})}{p(\pi^{\mathrm{Db}++}_{[n+2]}|H_d, \mathbf{p})} =$$

The LR when $\mathbf{p}$ is known

Imagine we know $\mathbf{p}$.

$$\text{LR}_{|\mathbf{p}} = \frac{p(\pi^{\text{Db++}}_{[n+2]}|H_p, \mathbf{p})}{p(\pi^{\text{Db++}}_{[n+2]}|H_d, \mathbf{p})} = \text{Applying Lemma} =$$

The LR when $\mathbf{p}$ is known

Imagine we know $\mathbf{p}$.

$$\text{LR}_{|\mathbf{p}} = \frac{p(\pi^{\text{Db++}}_{[n+2]}|H_p, \mathbf{p})}{p(\pi^{\text{Db++}}_{[n+2]}|H_d, \mathbf{p})} = \text{Applying Lemma} = \frac{1}{\mathbb{E}(p_{x_{n+1}}|\pi^{\text{Db+}}_{[n+1]}, \mathbf{p})}.$$

The LR when $\mathbf{p}$ is known

Imagine we know $\mathbf{p}$.

$$\text{LR}_{|\mathbf{p}} = \frac{p(\pi^{\text{Db++}}_{[n+2]}|H_p, \mathbf{p})}{p(\pi^{\text{Db++}}_{[n+2]}|H_d, \mathbf{p})} = \text{Applying Lemma} = \frac{1}{\mathbb{E}(p_{x_{n+1}}|\pi^{\text{Db+}}_{[n+1]}, \mathbf{p})}.$$

How is this compared to the one we get with our method when $\mathbf{p}$ is unknown?

Test Dutch database (N=2085, 7 loci)

Database of 2085 Y-STR profiles form Dutch men.

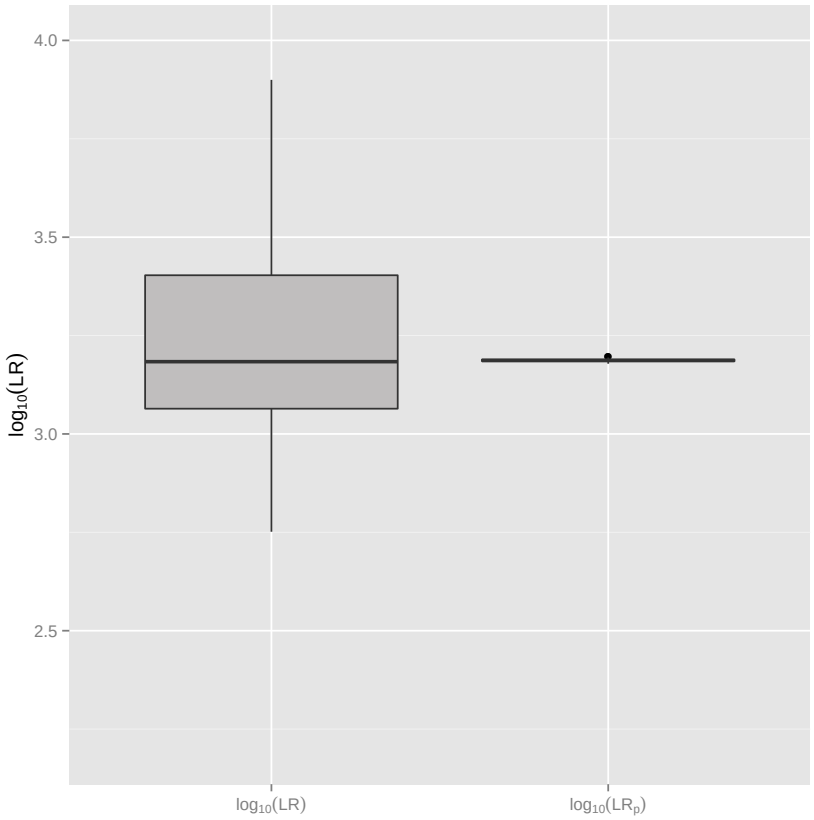
Test Dutch database (N=2085, 7 loci)

Database of 2085 Y-STR profiles form Dutch men.

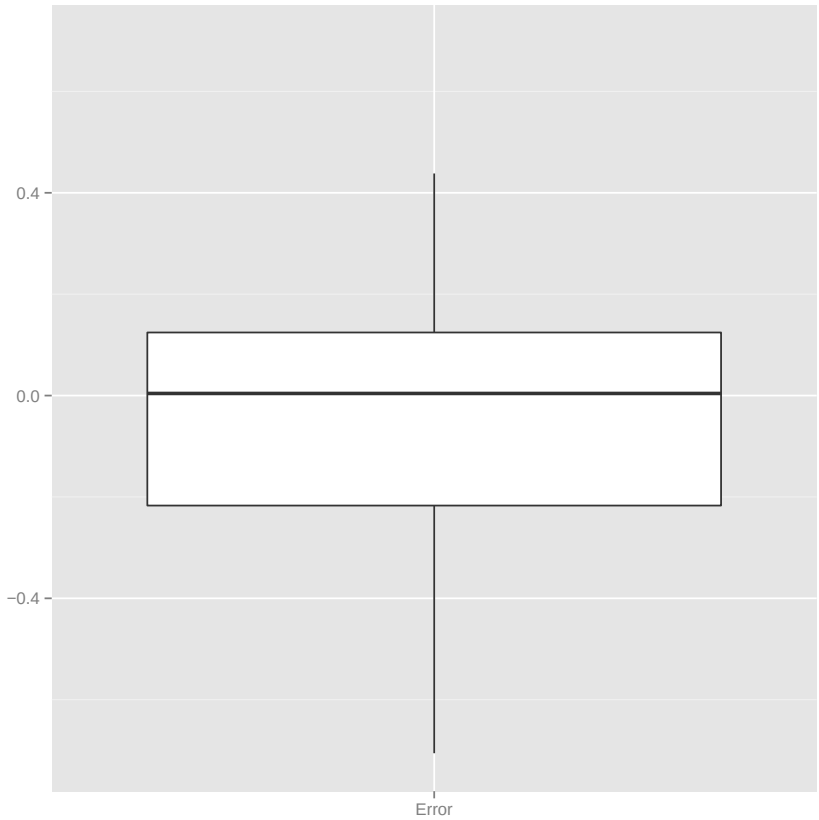
Test: Compare the distribution of $\log_{10}(\mathrm{LR}_{|\mathbf{p}})$ and $\log_{10}\mathrm{LR}$ obtained by 100 samples of size 100 from this population.

Results

Compare the distribution of $\log_{10}(\mathrm{LR}_{|\mathbf{p}})$ and $\log_{10}\mathrm{LR}$ obtained by 100 samples of size 100 from this population



(a) Comparison



(b) Error