LR assessment for the rare type match using a Bayesian nonparametric method

Giulia Cereda

November 11, 2016

LR assessment

The rare type match case

A Bayesian nonparametric method

LR assessment

The rare type match case

A Bayesian nonparametric method

LR assessment

The rare type match case

A Bayesian nonparametric method

LR assessment

The rare type match case

A Bayesian nonparametric method

Likelihood ratio assessment

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

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

Likelihood ratio assessment

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

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

Likelihood ratio assessment

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

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

Likelihood ratio assessment

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

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

Likelihood ratio assessment

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

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

Likelihood ratio assessment

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

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

Likelihood ratio assessment

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

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

The rare type match case

Y-STR profiles.

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

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

The rare type match case

Y-STR profiles.

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

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

The rare type match case

Y-STR profiles.

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

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

The rare type match case

Y-STR profiles.

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

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

How rare?

The rare type match case

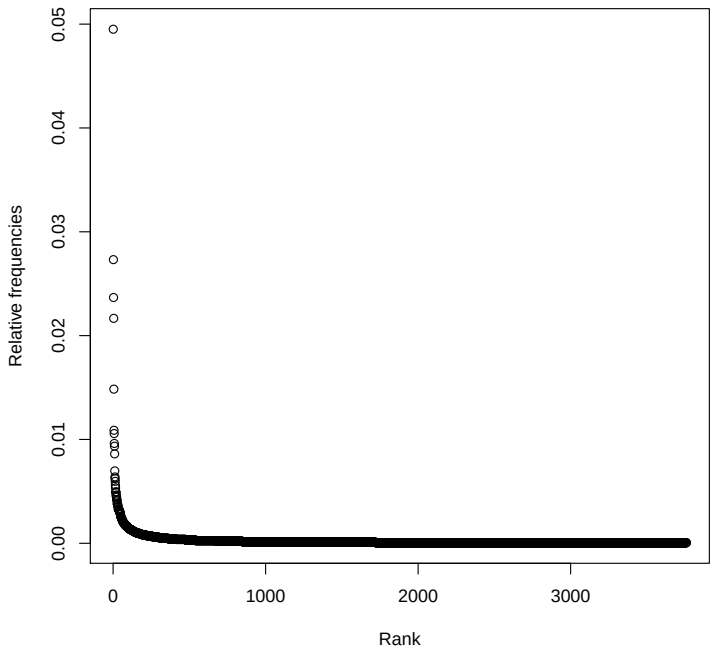
Y-STR profiles.

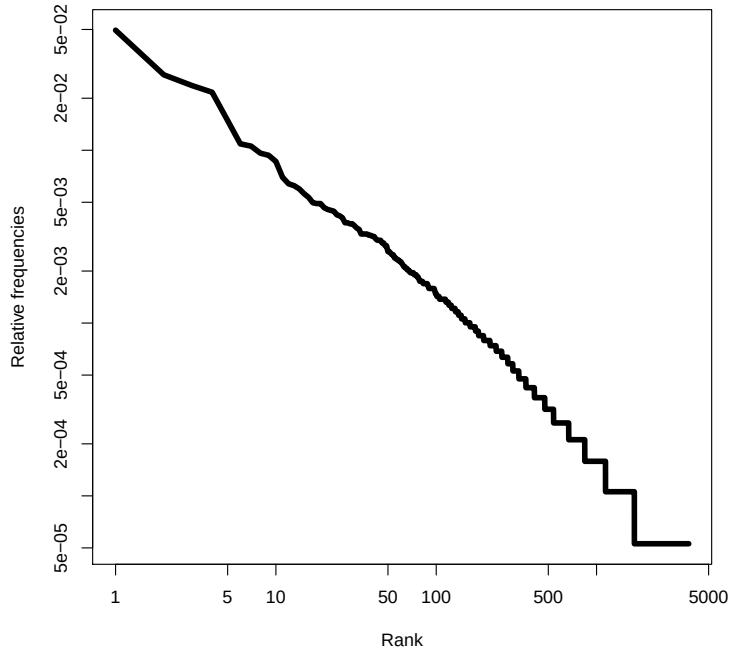
- A match between the suspect's DNA profile and the crime stain's DNA profile.
- This profile is not contained in the database.

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

Y-STR data

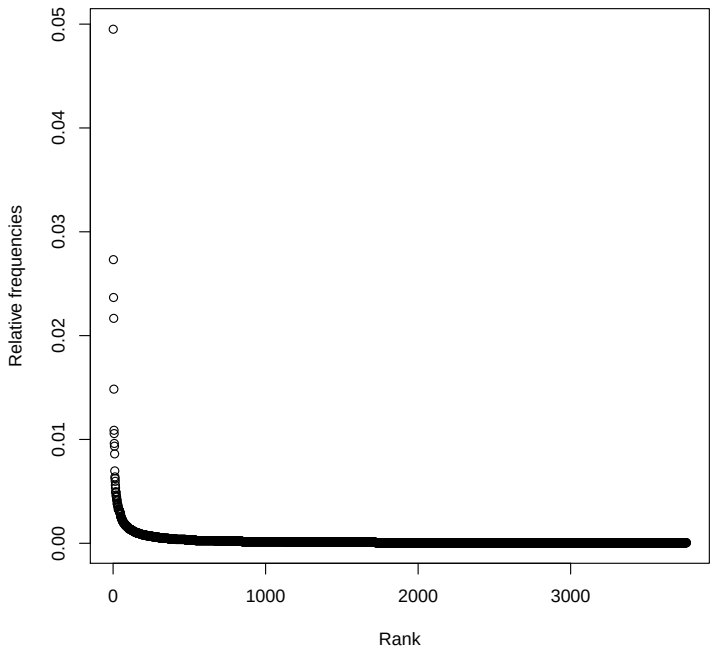
Y-STR database N=18925,
distinct profiles= 3759,
singletons=2038,
duplets= 584.

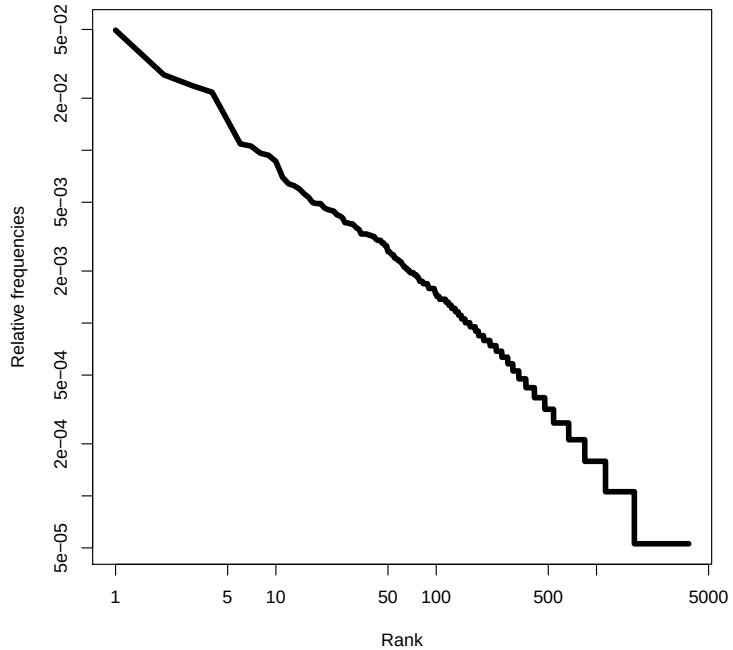




Y-STR data

Y-STR database N=18925,
distinct profiles= 3759,
singletons=2038,
duplets= 584.

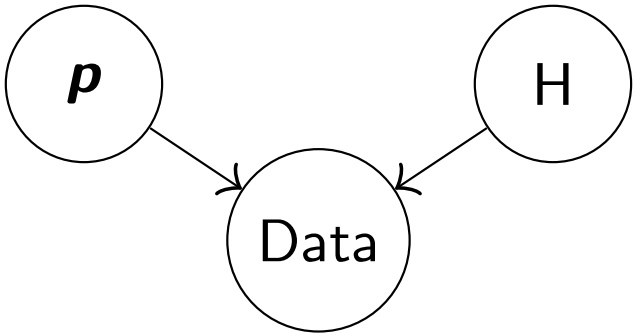




Bayesian nonparametric model

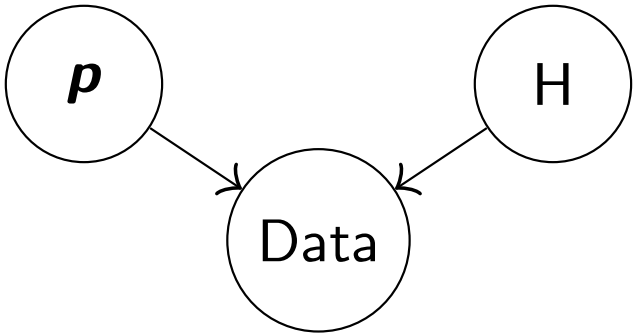
- a prior distribution for p,
- p is infinite dimensional parameter.

Bayesian nonparametric model



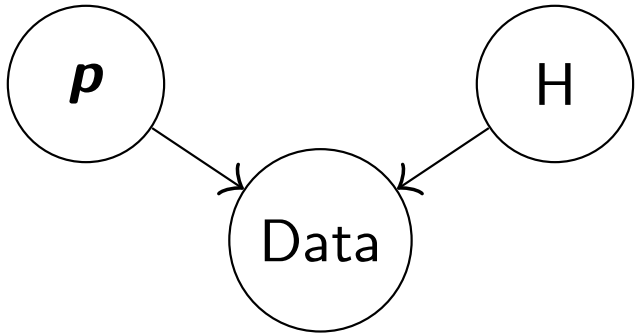
- a prior distribution for p,
- p is infinite dimensional parameter.

Bayesian nonparametric model



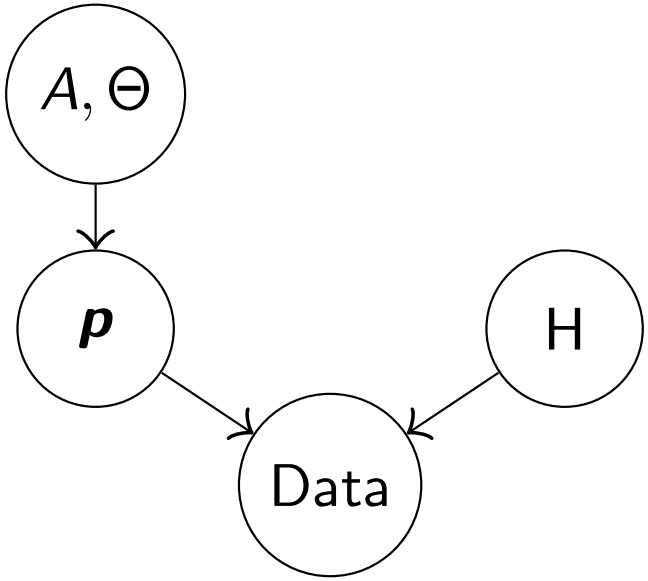
- a prior distribution for $\boldsymbol{p}$,
- $\boldsymbol{p}$ is infinite dimensional parameter.

Bayesian nonparametric model



- a prior distribution for $\boldsymbol{p}$,
- $\boldsymbol{p}$ is infinite dimensional parameter.

Final result



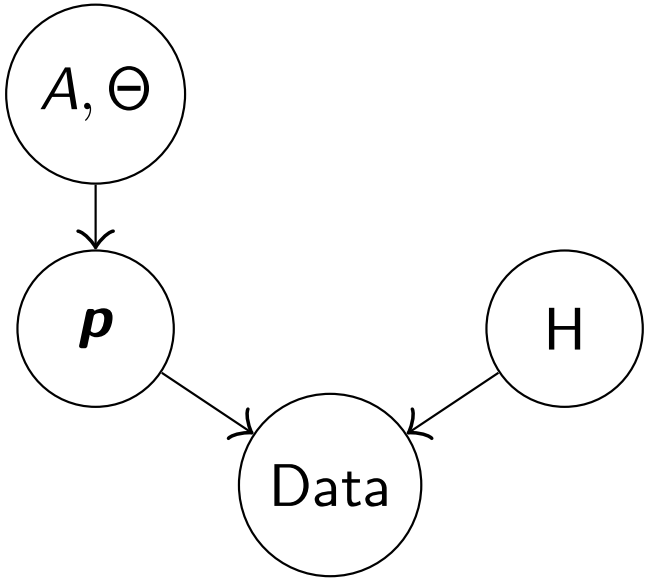
Based on our prior choice and assumptions, and data reduction, we obtain

$$\text{LR} = \frac{1}{\mathbb{E}(\frac{1-\alpha}{n+1+\theta} \mid \text{reduced data})}$$

If the hyperprior over α and θ is noninformative,

$$\text{LR} \approx \frac{1}{\frac{1-\widehat{\alpha}_{MLE}}{n+1+\widehat{\theta}_{MLE}}}.$$

Final result



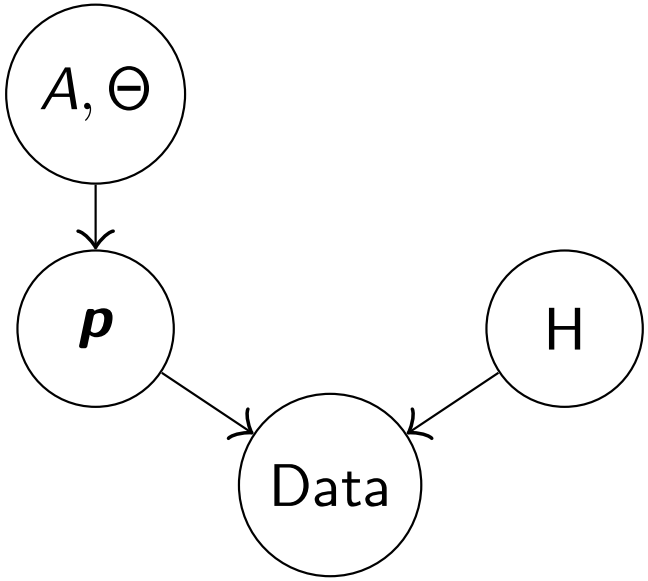
Based on our prior choice and assumptions, and data reduction, we obtain

$$\text{LR} = \frac{1}{\mathbb{E}(\frac{1-\alpha}{n+1+\theta} \mid \text{reduced data})}$$

If the hyperprior over α and θ is noninformative,

$$\text{LR} \approx \frac{1}{\frac{1-\widehat{\alpha}_{MLE}}{n+1+\widehat{\theta}_{MLE}}}.$$

Final result



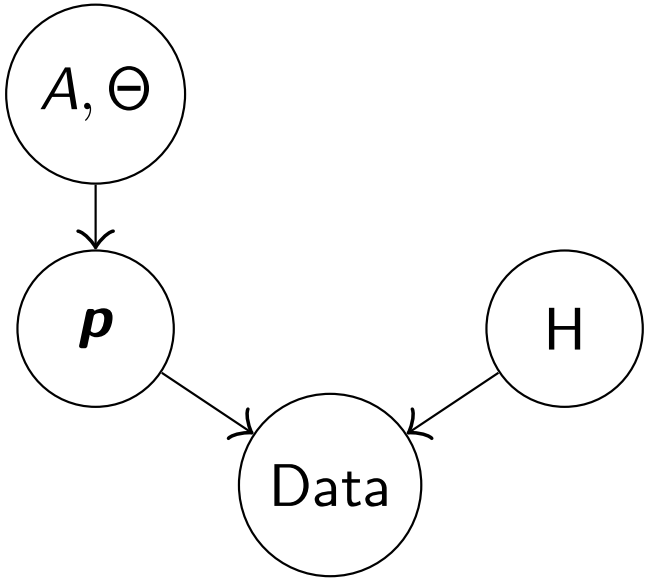
Based on our prior choice and assumptions, and data reduction, we obtain

$$\text{LR} = \frac{1}{\mathbb{E}(\frac{1-\alpha}{n+1+\theta} \mid \text{reduced data})}$$

If the hyperprior over α and θ is noninformative,

$$\text{LR} \approx \frac{1}{\frac{1-\widehat{\alpha}_{MLE}}{n+1+\widehat{\theta}_{MLE}}}.$$

Final result



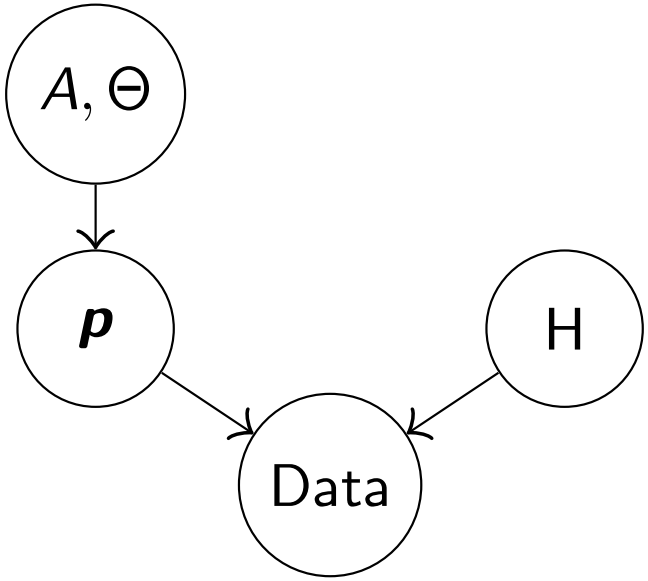
Based on our prior choice and assumptions, and data reduction, we obtain

$$\text{LR} = \frac{1}{\mathbb{E}(\frac{1-\alpha}{n+1+\theta} \mid \text{reduced data})}$$

If the hyperprior over α and θ is noninformative,

$$\text{LR} \approx \frac{1}{\frac{1-\widehat{\alpha}_{MLE}}{n+1+\widehat{\theta}_{MLE}}}.$$

Final result



Based on our prior choice and assumptions, and data reduction, we obtain

$$\text{LR} = \frac{1}{\mathbb{E}(\frac{1-\alpha}{n+1+\theta} \mid \text{reduced data})}$$

If the hyperprior over α and θ is noninformative,

$$\text{LR} \approx \frac{1}{\frac{1-\widehat{\alpha}_{MLE}}{n+1+\widehat{\theta}_{MLE}}}.$$

two parameter Chinese Restaurant process

Depends on two parameters, α and θ such that $0 \leq \alpha < 1$, $\theta > -\alpha$
A restaurant with infinite tables, each infinitely large.
$Y_1, Y_2, ...$ is the seating plan:

- First customer always seats at the first table: $Y_1 = 1$
- Second customer seats at
  - the first table: $Y_2 = 1$ with probability $\frac{1-\alpha}{1+\theta}$
  - new unoccupied table: $Y_2 = 2$ with probability $\frac{\theta+\alpha}{1+\theta}$
- ...
- Given the first n customers have seated according to $Y_1, ..., Y_n$ to k different tables, it holds that

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

where n_i is the number of customers that occupy table i

two parameter Chinese Restaurant process

Depends on two parameters, α and θ such that $0 \leq \alpha < 1$, $\theta > -\alpha$

A restaurant with infinite tables, each infinitely large.

$Y_1, Y_2, ...$ is the seating plan:

- First customer always seats at the first table: $Y_1 = 1$
- Second customer seats at
  - the first table: $Y_2 = 1$ with probability $\frac{1-\alpha}{1+\theta}$
  - new unoccupied table: $Y_2 = 2$ with probability $\frac{\theta+\alpha}{1+\theta}$
- ...
- Given the first n customers have seated according to $Y_1, ..., Y_n$ to k different tables, it holds that

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

where n_i is the number of customers that occupy table i

two parameter Chinese Restaurant process

Depends on two parameters, α and θ such that $0 \leq \alpha < 1$, $\theta > -\alpha$

A restaurant with infinite tables, each infinitely large.

$Y_1, Y_2, ...$ is the seating plan:

- First customer always seats at the first table: $Y_1 = 1$
- Second customer seats at
  - the first table: $Y_2 = 1$ with probability $\frac{1-\alpha}{1+\theta}$
  - new unoccupied table: $Y_2 = 2$ with probability $\frac{\theta+\alpha}{1+\theta}$
- ...
- Given the first n customers have seated according to $Y_1, ..., Y_n$ to k different tables, it holds that

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

where n_i is the number of customers that occupy table i

two parameter Chinese Restaurant process

Depends on two parameters, α and θ such that $0 \leq \alpha < 1$, $\theta > -\alpha$
A restaurant with infinite tables, each infinitely large.
$Y_1, Y_2, ...$ is the seating plan:

- First customer always seats at the first table: $Y_1 = 1$
- Second customer seats at
  - the first table: $Y_2 = 1$ with probability $\frac{1-\alpha}{1+\theta}$
  - new unoccupied table: $Y_2 = 2$ with probability $\frac{\theta+\alpha}{1+\theta}$
- ...
- Given the first n customers have seated according to $Y_1, ..., Y_n$ to k different tables, it holds that

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

where n_i is the number of customers that occupy table i

two parameter Chinese Restaurant process

Depends on two parameters, α and θ such that $0 \leq \alpha < 1$, $\theta > -\alpha$
A restaurant with infinite tables, each infinitely large.
$Y_1, Y_2, ...$ is the seating plan:

- First customer always seats at the first table: $Y_1 = 1$
- Second customer seats at
  - the first table: $Y_2 = 1$ with probability $\frac{1-\alpha}{1+\theta}$
  - new unoccupied table: $Y_2 = 2$ with probability $\frac{\theta+\alpha}{1+\theta}$
- ...
- Given the first n customers have seated according to $Y_1, ..., Y_n$ to k different tables, it holds that

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

where n_i is the number of customers that occupy table i

two parameter Chinese Restaurant process

Depends on two parameters, α and θ such that $0 \leq \alpha < 1$, $\theta > -\alpha$
A restaurant with infinite tables, each infinitely large.
$Y_1, Y_2, ...$ is the seating plan:

- First customer always seats at the first table: $Y_1 = 1$
- Second customer seats at
  - the first table: $Y_2 = 1$ with probability $\frac{1-\alpha}{1+\theta}$
  - new unoccupied table: $Y_2 = 2$ with probability $\frac{\theta+\alpha}{1+\theta}$
- ...
- Given the first n customers have seated according to $Y_1, ..., Y_n$ to k different tables, it holds that

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

where n_i is the number of customers that occupy table i

two parameter Chinese Restaurant process

Depends on two parameters, α and θ such that $0 \leq \alpha < 1$, $\theta > -\alpha$
A restaurant with infinite tables, each infinitely large.
$Y_1, Y_2, \ldots$ is the seating plan:

- First customer always seats at the first table: $Y_1 = 1$
- Second customer seats at
  - the first table: $Y_2 = 1$ with probability $\frac{1-\alpha}{1+\theta}$
  - new unoccupied table: $Y_2 = 2$ with probability $\frac{\theta+\alpha}{1+\theta}$
- ...
- Given the first n customers have seated according to $Y_1, \ldots, Y_n$ to k different tables, it holds that

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

where n_i is the number of customers that occupy table i

two parameter Chinese Restaurant process

Depends on two parameters, α and θ such that $0 \leq \alpha < 1$, $\theta > -\alpha$
A restaurant with infinite tables, each infinitely large.
$Y_1, Y_2, \ldots$ is the seating plan:

- First customer always seats at the first table: $Y_1 = 1$
- Second customer seats at
  - the first table: $Y_2 = 1$ with probability $\frac{1-\alpha}{1+\theta}$
  - new unoccupied table: $Y_2 = 2$ with probability $\frac{\theta+\alpha}{1+\theta}$
- ...
- Given the first n customers have seated according to $Y_1, \ldots, Y_n$ to k different tables, it holds that

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

where n_i is the number of customers that occupy table i

two parameter Chinese Restaurant process

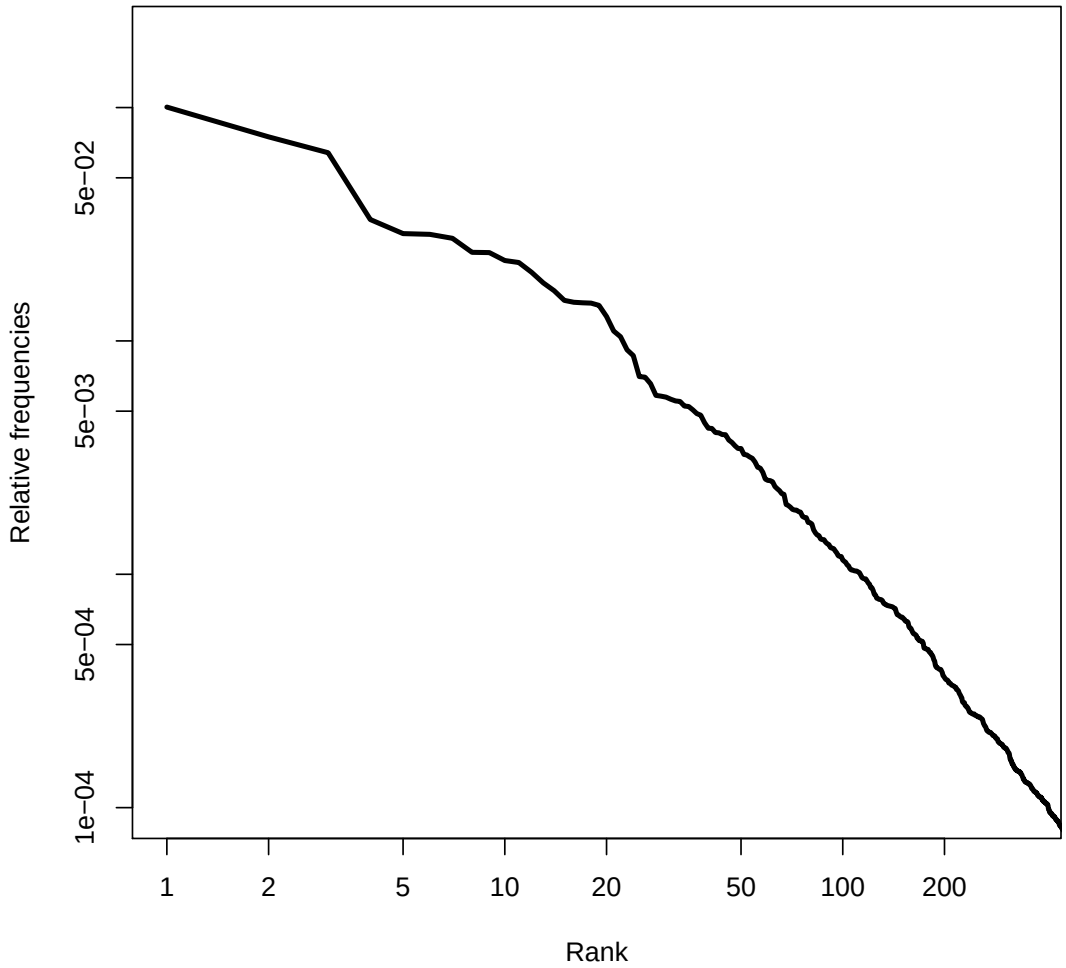
Depends on two parameters, α and θ such that $0 \leq \alpha < 1$, $\theta > -\alpha$
A restaurant with infinite tables, each infinitely large.
$Y_1, Y_2, \ldots$ is the seating plan:

- First customer always seats at the first table: $Y_1 = 1$
- Second customer seats at
  - the first table: $Y_2 = 1$ with probability $\frac{1-\alpha}{1+\theta}$
  - new unoccupied table: $Y_2 = 2$ with probability $\frac{\theta+\alpha}{1+\theta}$
- ...
- Given the first n customers have seated according to $Y_1, \ldots, Y_n$ to k different tables, it holds that

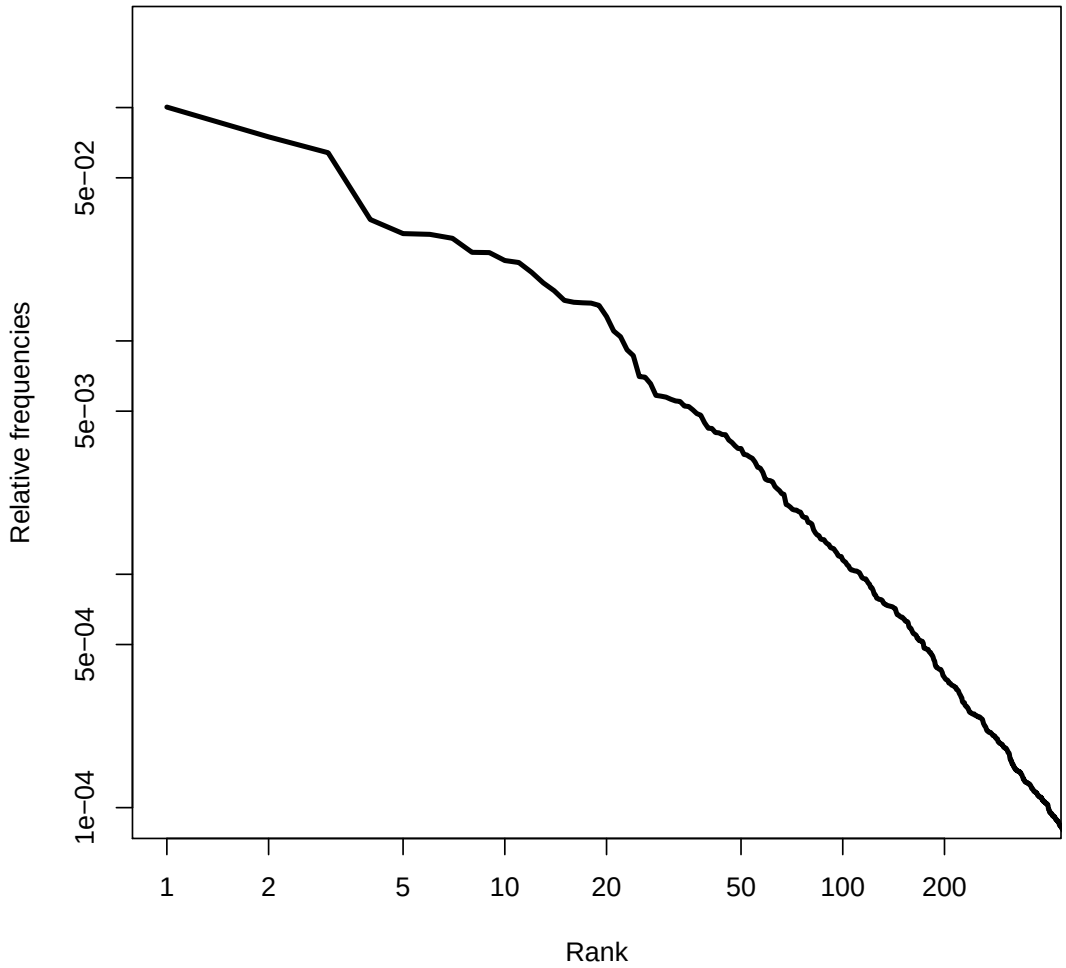
$$\Pr(Y_{n+1} = i | Y_1, \ldots, Y_n) = \begin{cases} \frac{\theta+k\alpha}{n+\theta} & \text{if } i = k+1, \\ \frac{n_i-\alpha}{n+\theta} & \text{if } 1 \leq i \leq k \end{cases}$$

where n_i is the number of customers that occupy table i

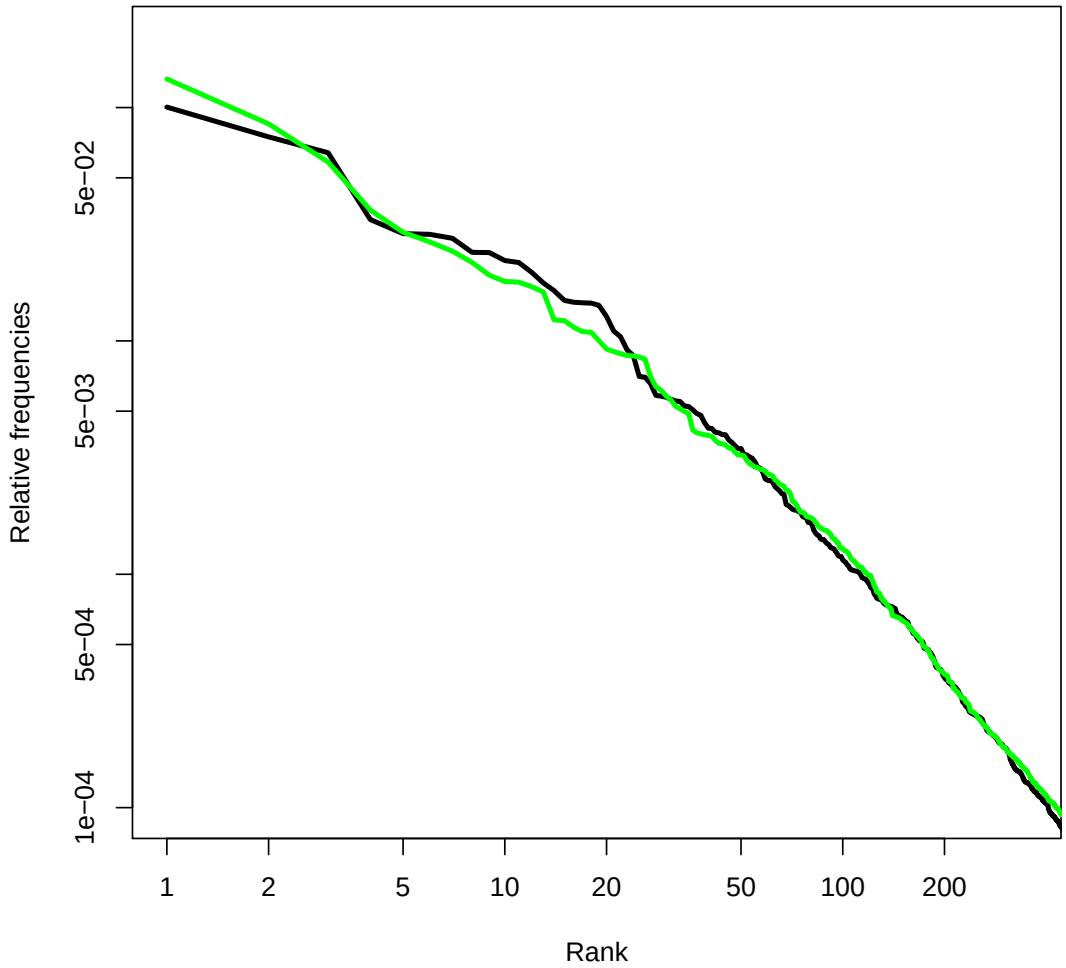
Let the customers enter for an infinite long time. Let $\boldsymbol{p}$ be the infinite vector containing the limit ordered relative size of each table.



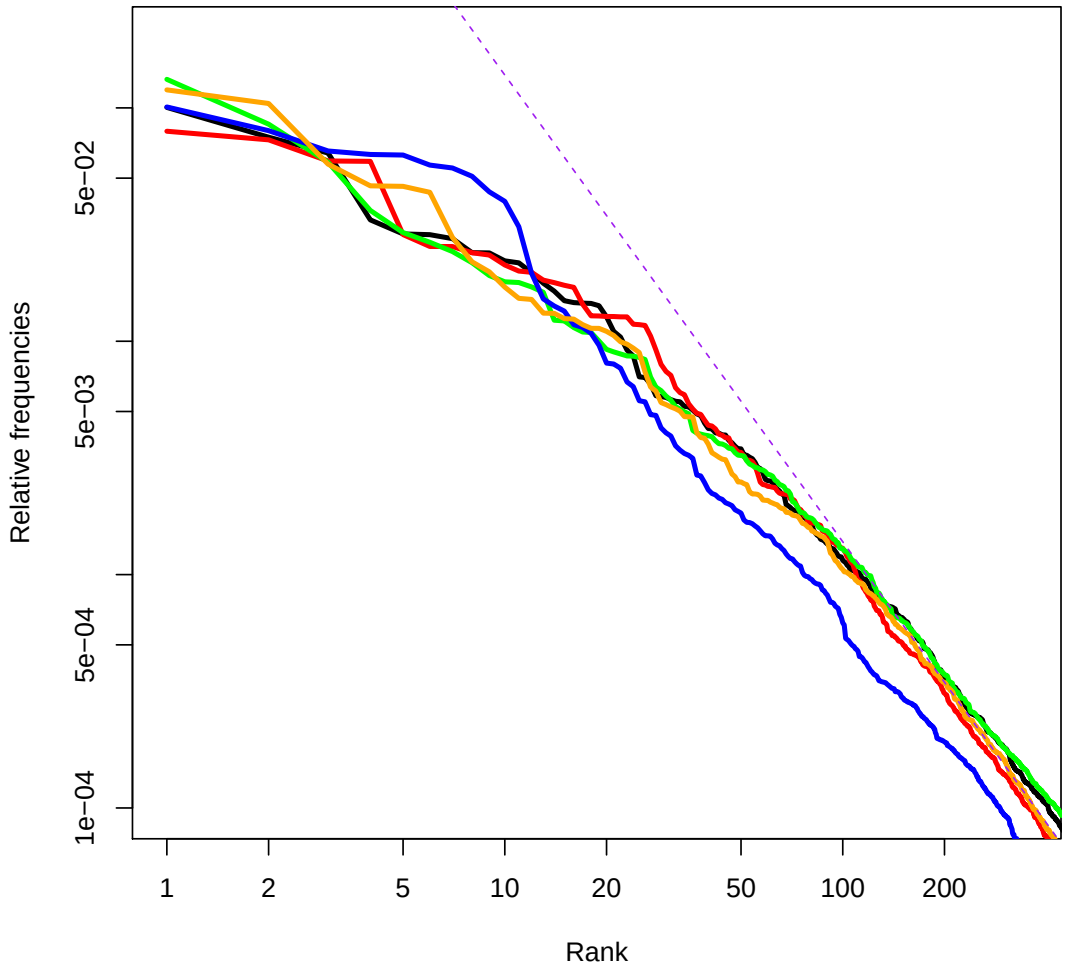
Let the customers enter for an infinite long time. Let $\boldsymbol{p}$ be the infinite vector containing the limit ordered relative size of each table.



Let the customers enter for an infinite long time. Let $\boldsymbol{p}$ be the infinite vector containing the limit ordered relative size of each table.

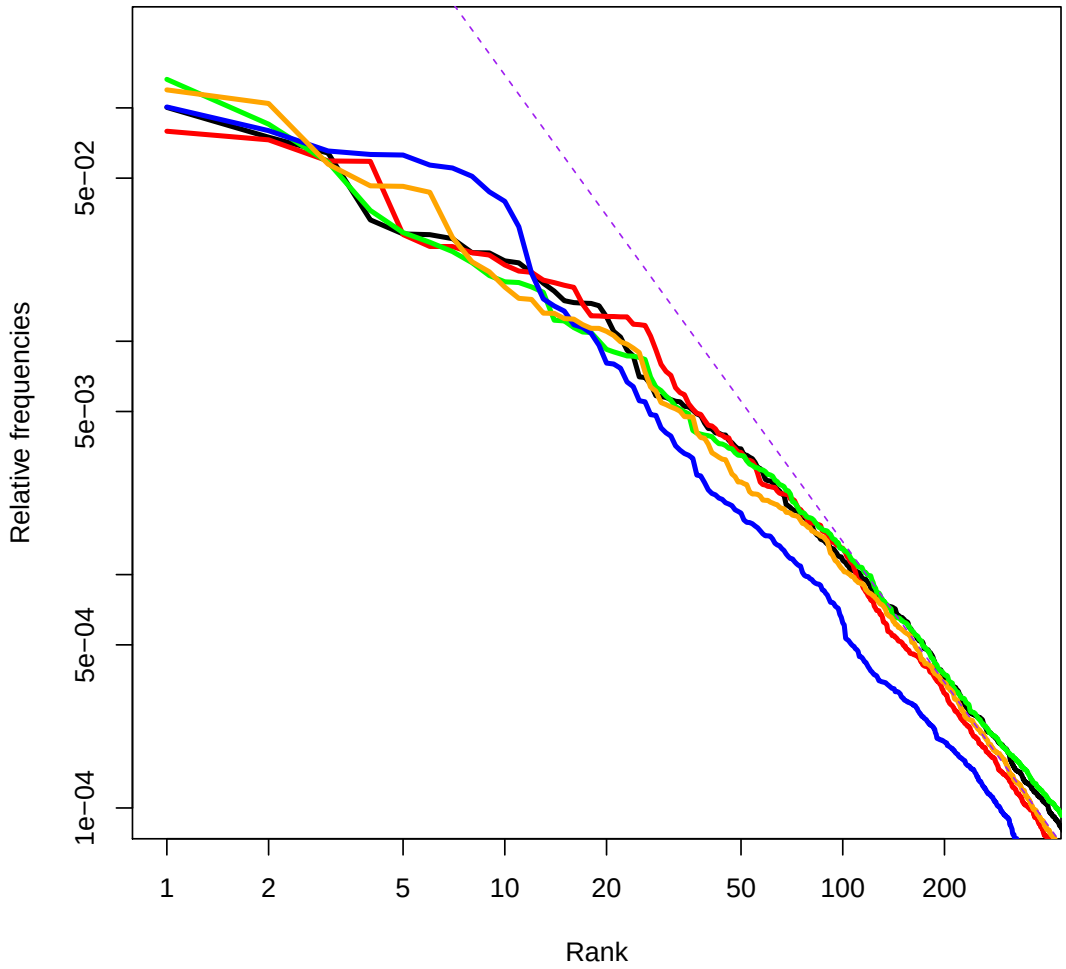


Let the customers enter for an infinite long time. Let $\boldsymbol{p}$ be the infinite vector containing the limit ordered relative size of each table.



$\boldsymbol{p}$ follows the two parameter Poisson Dirichlet distribution $PD(\alpha, \theta)$

Let the customers enter for an infinite long time. Let $\boldsymbol{p}$ be the infinite vector containing the limit ordered relative size of each table.



$\boldsymbol{p}$ follows the two parameter Poisson Dirichlet distribution $PD(\alpha, \theta)$

The distribution of the data given $\mathbf{p}$ and H

Some notation...

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

A partition of the set $[n]$ will be denoted as $\pi_{[n]}$.

For instance if $n = 10$, $\pi_{[n]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}\}$

Random partitions on the set $[n]$ will be denoted as $\Pi_{[n]}$.

The distribution of the data given $\mathbf{p}$ and H

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

A partition of the set $[n]$ will be denoted as $\pi_{[n]}$.

For instance if $n = 10$, $\pi_{[n]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}\}$

Random partitions on the set $[n]$ will be denoted as $\Pi_{[n]}$.

The distribution of the data given $\mathbf{p}$ and H

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

A partition of the set $[n]$ will be denoted as $\pi_{[n]}$.

For instance if $n = 10$, $\pi_{[n]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}\}$

Random partitions on the set $[n]$ will be denoted as $\Pi_{[n]}$.

The distribution of the data given $\mathbf{p}$ and H

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

A partition of the set $[n]$ will be denoted as $\pi_{[n]}$.

For instance if $n = 10$, $\pi_{[n]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}\}$

Random partitions on the set $[n]$ will be denoted as $\Pi_{[n]}$.

The distribution of the data given $\mathbf{p}$ and H

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

A partition of the set $[n]$ will be denoted as $\pi_{[n]}$.

For instance if $n = 10$, $\pi_{[n]} = \{\{1\}, \{2, 9\}, \{3, 5, 6, 8, 10\}, \{4\}, \{7\}\}$

Random partitions on the set $[n]$ will be denoted as $\Pi_{[n]}$.

Assumptions

Assumption 1

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

Parameter: $\mathbf{p} \in \nabla_\infty = \{(p_1, p_2, ...), p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$.
p_3 is the frequency of the third most frequent type in the population. We chose the two parameter Poisson Dirichlet distribution as prior for $\mathbf{p}$.

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.

Assumptions

Assumption 1

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

Parameter: $\mathbf{p} \in \nabla_\infty = \{(p_1, p_2, ...), p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$.
p_3 is the frequency of the third most frequent type in the population. We chose the two parameter Poisson Dirichlet distribution as prior for $\mathbf{p}$.

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.

Assumptions

Assumption 1

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

Parameter: $\mathbf{p} \in \nabla_\infty = \{(p_1, p_2, ...), p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$.
p_3 is the frequency of the third most frequent type in the population. We chose the two parameter Poisson Dirichlet distribution as prior for $\mathbf{p}$.

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.

Assumptions

Assumption 1

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

Parameter: $\mathbf{p} \in \nabla_\infty = \{(p_1, p_2, ...), p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$.
p_3 is the frequency of the third most frequent type in the population. We chose the two parameter Poisson Dirichlet distribution as prior for $\mathbf{p}$.

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.

Assumptions

Assumption 1

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

Parameter: $\mathbf{p} \in \nabla_\infty = \{(p_1, p_2, ...), p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$.
p_3 is the frequency of the third most frequent type in the population. We chose the two parameter Poisson Dirichlet distribution as prior for $\mathbf{p}$.

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.

Assumptions

Assumption 1

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

Parameter: $\mathbf{p} \in \nabla_\infty = \{(p_1, p_2, ...), p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$.
p_3 is the frequency of the third most frequent type in the population. We chose the two parameter Poisson Dirichlet distribution as prior for $\mathbf{p}$.

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.

DNA database can be reduced

DATABASE of size 10
Person 1 $(16 - 12 - 17 - 23 - 11 - 13 - 12)$
Person 2 $(15 - 14 - 17 - 24 - 10 - 13 - 14)$
Person 3 $(13 - 13 - 18 - 24 - 10 - 11 - 13)$
Person 4 $(14 - 14 - 16 - 24 - 11 - 11 - 10)$
Person 5 $(13 - 13 - 18 - 24 - 10 - 11 - 13)$
Person 6 $(13 - 13 - 18 - 24 - 10 - 11 - 13)$
Person 7 $(13 - 12 - 16 - 25 - 10 - 13 - 12$
Person 8 $(13 - 13 - 18 - 24 - 10 - 11 - 13)$
Person 9 $(15 - 14 - 17 - 24 - 10 - 13 - 14)$
Person 10 $(13 - 13 - 18 - 24 - 10 - 11 - 13)$

DNA database can be reduced

DATABASE of size 10

Person 1 $(16-12-17-23-11-13-12)$

Person 2 $(15-14-17-24-10-13-14)$

Person 3 $(13-13-18-24-10-11-13)$

Person 4 $(14-14-16-24-11-11-10)$

Person 5 $(13-13-18-24-10-11-13)$

Person 6 $(13-13-18-24-10-11-13)$

Person 7 $(13-12-16-25-10-13-12)$

Person 8 $(13-13-18-24-10-11-13)$

Person 9 $(15-14-17-24-10-13-14)$

Person 10 $(13-13-18-24-10-11-13)$

Assumption 2 $\rightarrow$ data can be replaced by the equivalence classes on the indices of the relation "to have the same DNA type".

This is a partition of the set $[n]$: $\{\{1\},\{2,9\},\{3,5,6,8,10\},\{4\},\{7\}\}$

DNA database can be reduced

DATABASE of size 10
Person 1 $(16-12-17-23-11-13-12)$
Person 2 $(15-14-17-24-10-13-14)$
Person 3 $(13-13-18-24-10-11-13)$
Person 4 $(14-14-16-24-11-11-10)$
Person 5 $(13-13-18-24-10-11-13)$
Person 6 $(13-13-18-24-10-11-13)$
Person 7 $(13-12-16-25-10-13-12)$
Person 8 $(13-13-18-24-10-11-13)$
Person 9 $(15-14-17-24-10-13-14)$
Person 10 $(13-13-18-24-10-11-13)$
Assumption 2 $\rightarrow$ data can be replaced 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

The database of size n is reduced to the partition $\pi_{[n]}^{\text{Db}}$.

Data $\mathcal{D}$ is made of the database, and of 2 new observations: the suspect's DNA type and the crime stain's DNA type (the same type).

$\mathcal{D} = \pi_{[n+2]}^{\text{Db++}}$, 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 partition

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

Reduced data

The database of size n is reduced to the partition $\pi_{[n]}^{\text{Db}}$.

Data $\mathcal{D}$ is made of the database, and of 2 new observations: the suspect's DNA type and the crime stain's DNA type (the same type).

$\mathcal{D} = \pi_{[n+2]}^{\text{Db++}}$, 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 partition

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

Reduced data

The database of size n is reduced to the partition $\pi^{\text{Db}}_{[n]}$.

Data $\mathcal{D}$ is made of the database, and of 2 new observations: the suspect's DNA type and the crime stain's DNA type (the same type).

$\mathcal{D} = \pi^{\text{Db++}}_{[n+2]}$, 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 partition

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

Reduced data

The database of size n is reduced to the partition $\pi^{\text{Db}}_{[n]}$.

Data $\mathcal{D}$ is made of the database, and of 2 new observations: the suspect's DNA type and the crime stain's DNA type (the same type).

$\mathcal{D} = \pi^{\text{Db++}}_{[n+2]}$, 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 partition

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

Reduced data

The database of size n is reduced to the partition $\pi^{\text{Db}}_{[n]}$.

Data $\mathcal{D}$ is made of the database, and of 2 new observations: the suspect's DNA type and the crime stain's DNA type (the same type).

$\mathcal{D} = \pi^{\text{Db++}}_{[n+2]}$, 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 partition

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

Reduced data

The database of size n is reduced to the partition $\pi^{\text{Db}}_{[n]}$.

Data $\mathcal{D}$ is made of the database, and of 2 new observations: the suspect's DNA type and the crime stain's DNA type (the same type).

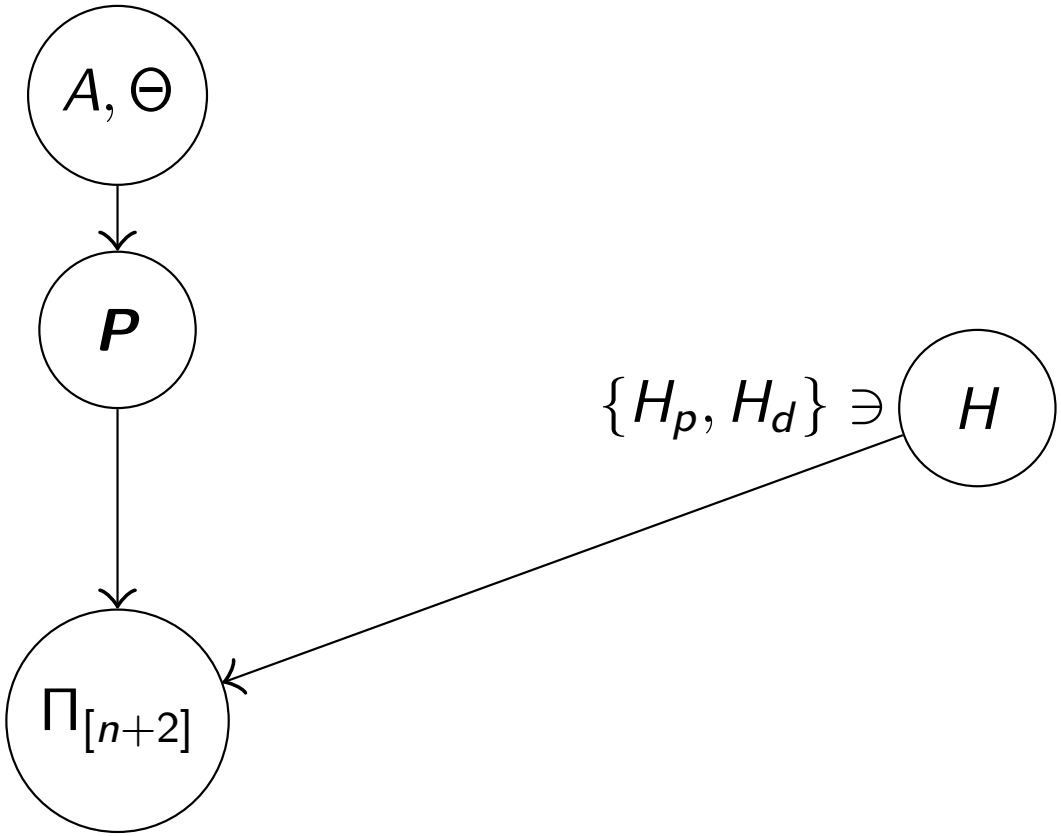
$\mathcal{D} = \pi^{\text{Db++}}_{[n+2]}$, partition of the set $\{1, 2, ..., n+2\}$.

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 partition

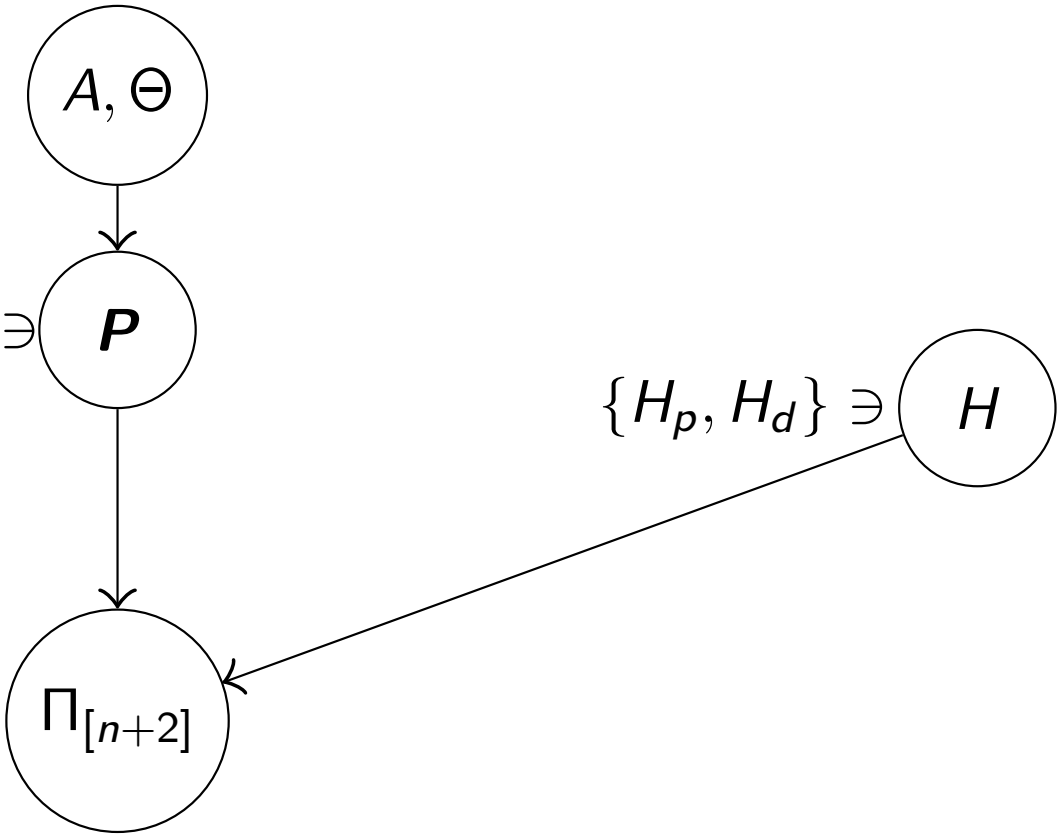
$$\mathcal{D} = \Pi_{[n+2]}$$

The model



The model

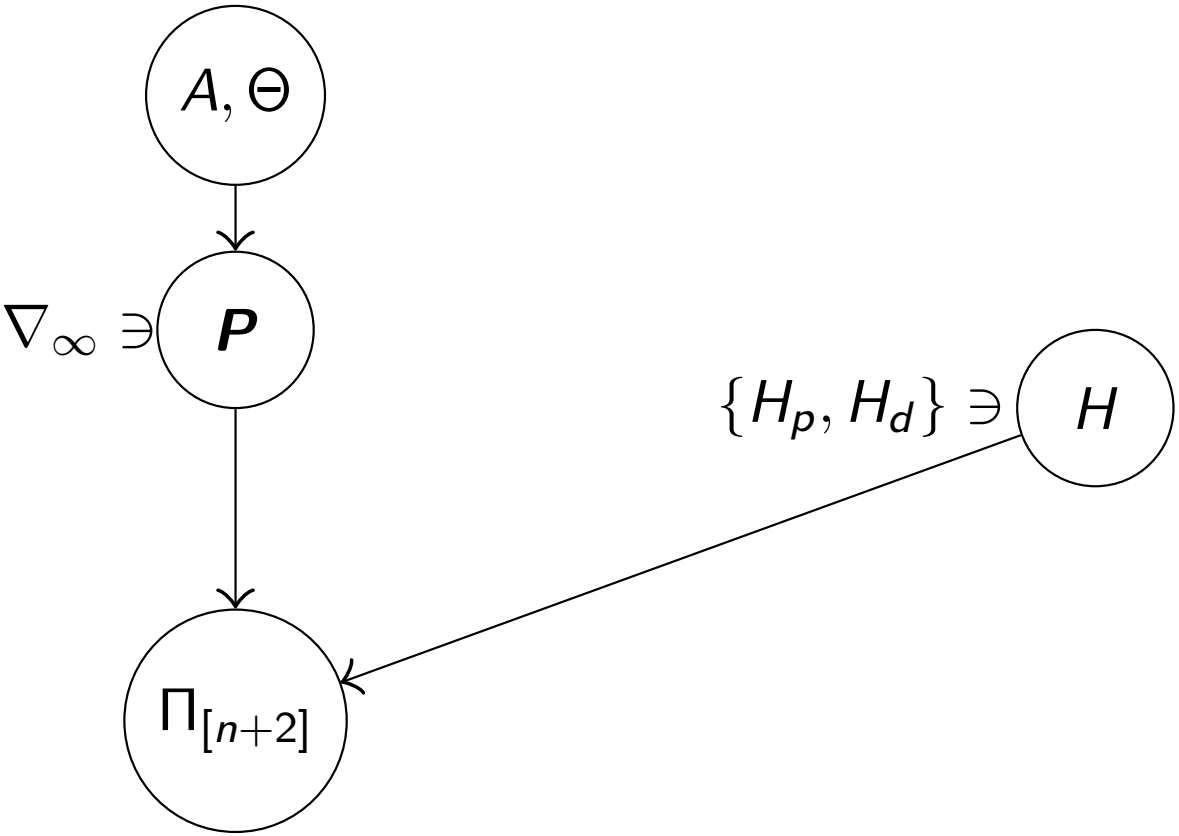
$$\boldsymbol{P}|\alpha,\theta \sim PD(\alpha,\theta) \qquad \nabla_\infty \ni \boldsymbol{P} \qquad \{H_p, H_d\} \ni H$$



The model

$$(A, \Theta) \sim f$$

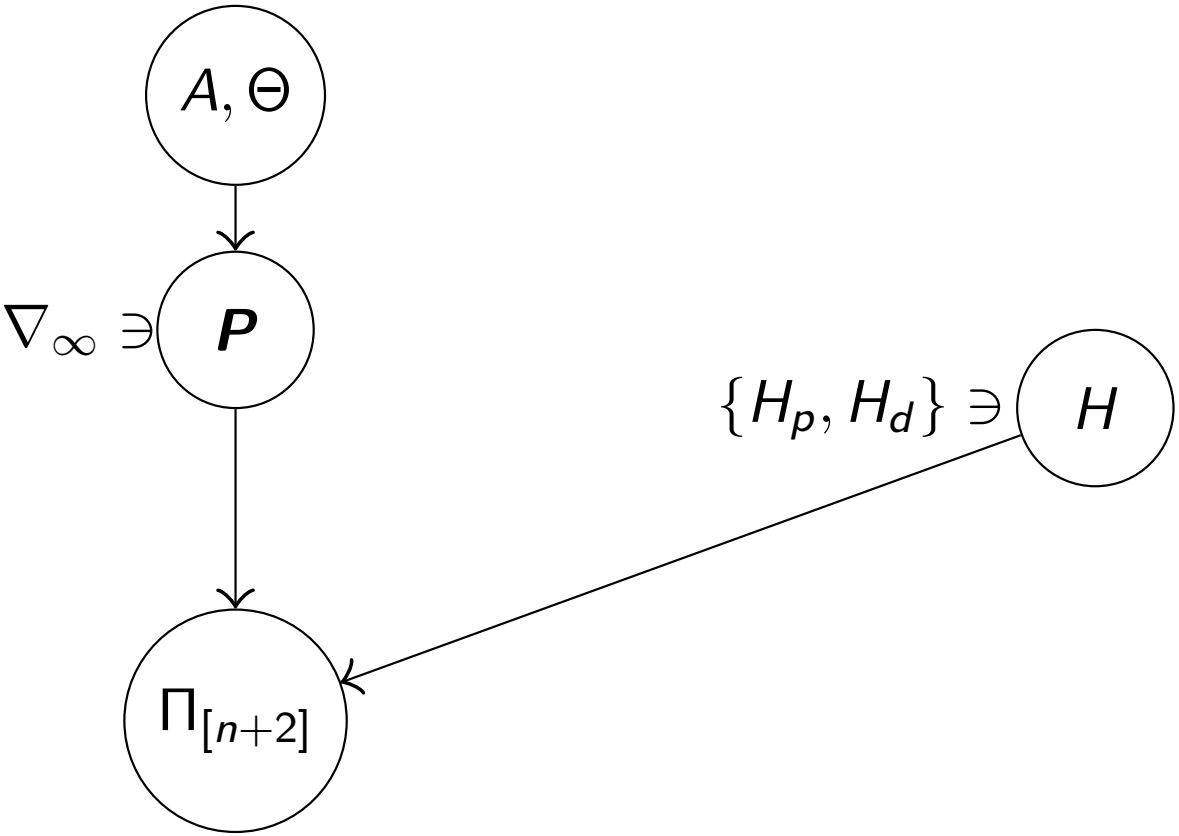
$$\boldsymbol{P}|\alpha, \theta \sim PD(\alpha, \theta)$$



The model

$$(A, \Theta) \sim f$$

$$\boldsymbol{P}|\alpha, \theta \sim PD(\alpha, \theta)$$



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

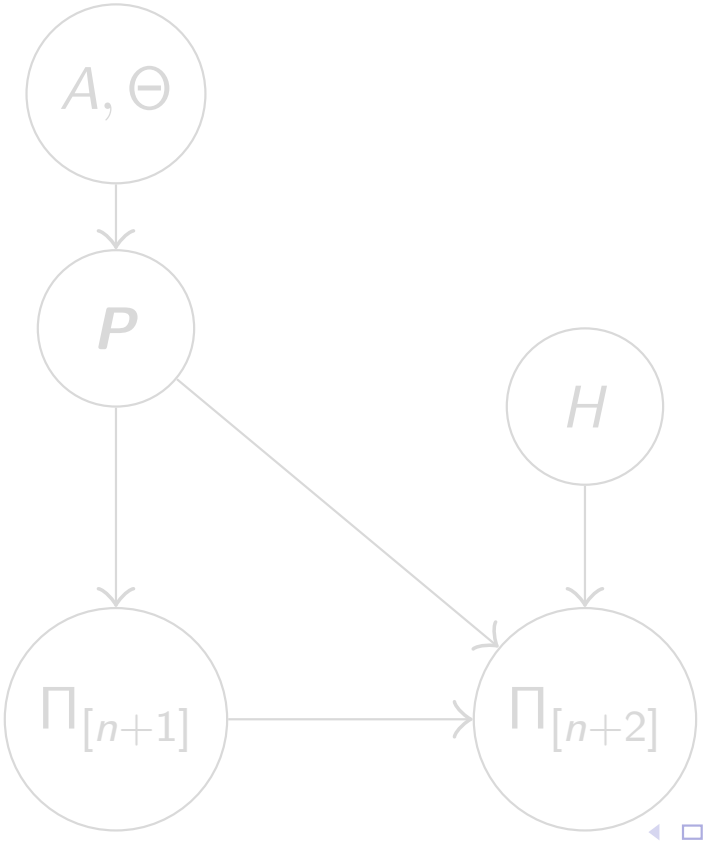
The model

$\Pi_{[n]}$ obtained from the database

$\Pi_{[n+1]} = \{\Pi_{[n]}, \{n+1\}\}$

$\Pi_{[n+2]} = \{\Pi_{[n+1]}, \{n+1, n+2\}\}$

The defence and prosecution disagree on the probability that the $n+2$nd observation (crime stain) goes in the same class of $n+1$st.



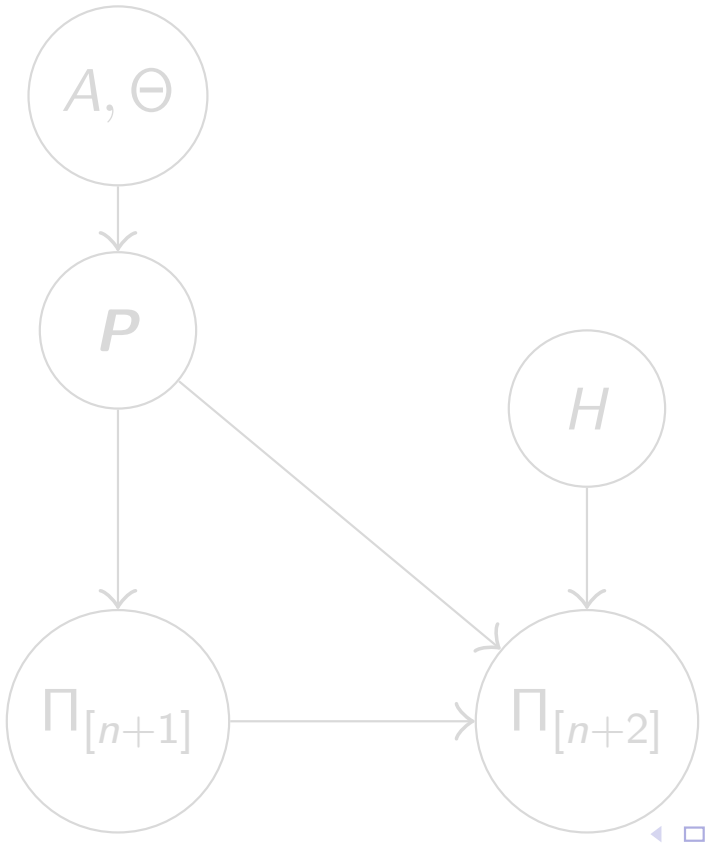
The model

$\Pi_{[n]}$ obtained from the database

$\Pi_{[n+1]} = \{\Pi_{[n]}, \{n+1\}\}$

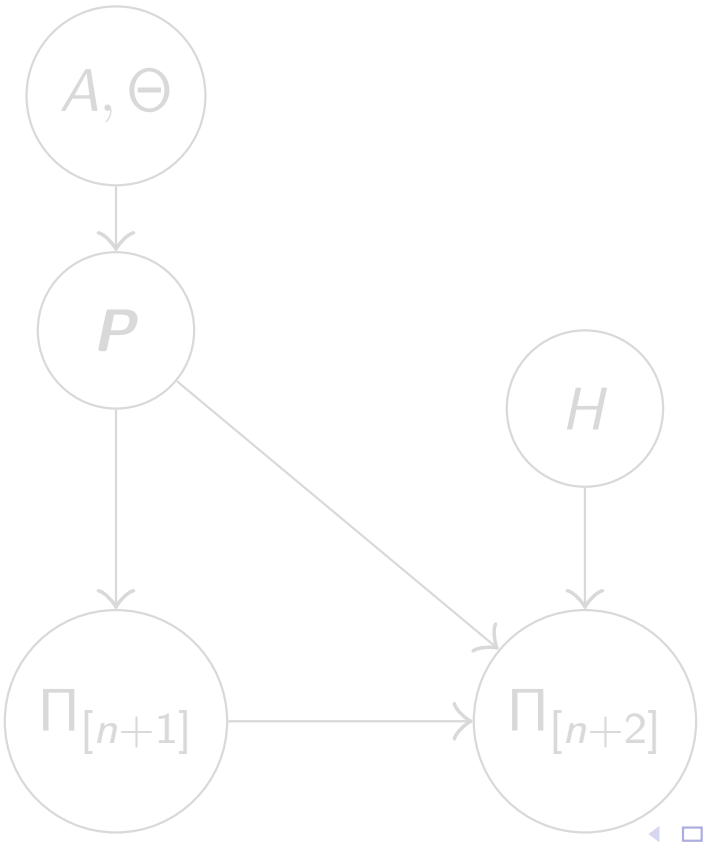
$\Pi_{[n+2]} = \{\Pi_{[n+1]}, \{n+1, n+2\}\}$

The defence and prosecution disagree on the probability that the $n+2$nd observation (crime stain) goes in the same class of $n+1$st.



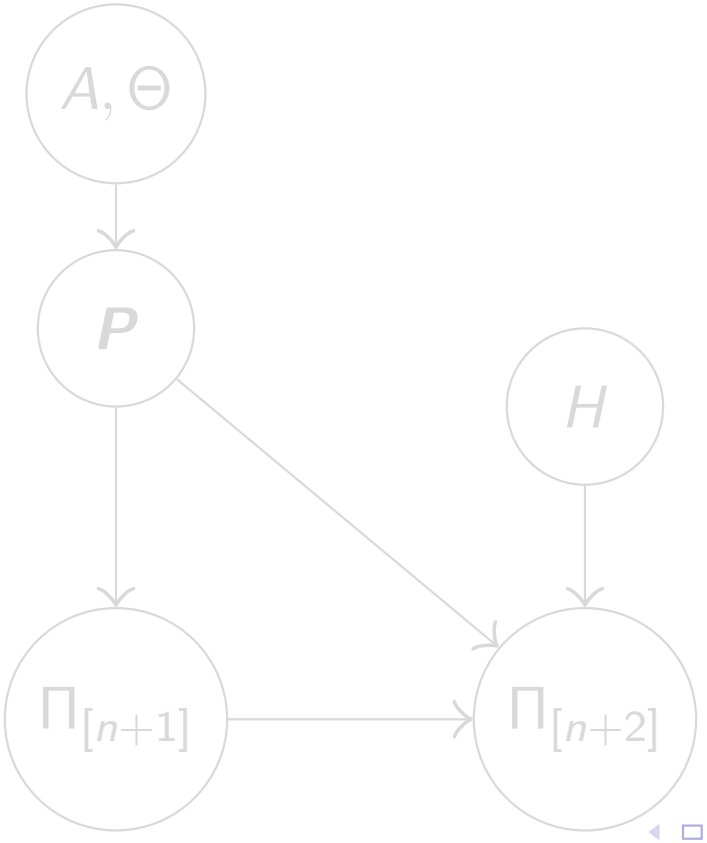
The model

$\Pi_{[n]}$ obtained from the database
$\Pi_{[n+1]} = \{\Pi_{[n]}, \{n+1\}\}$
$\Pi_{[n+2]} = \{\Pi_{[n+1]}, \{n+1, n+2\}\}$
The defence and prosecution disagree on the probability that the $n+2$nd observation (crime stain) goes in the same class of $n+1$st.



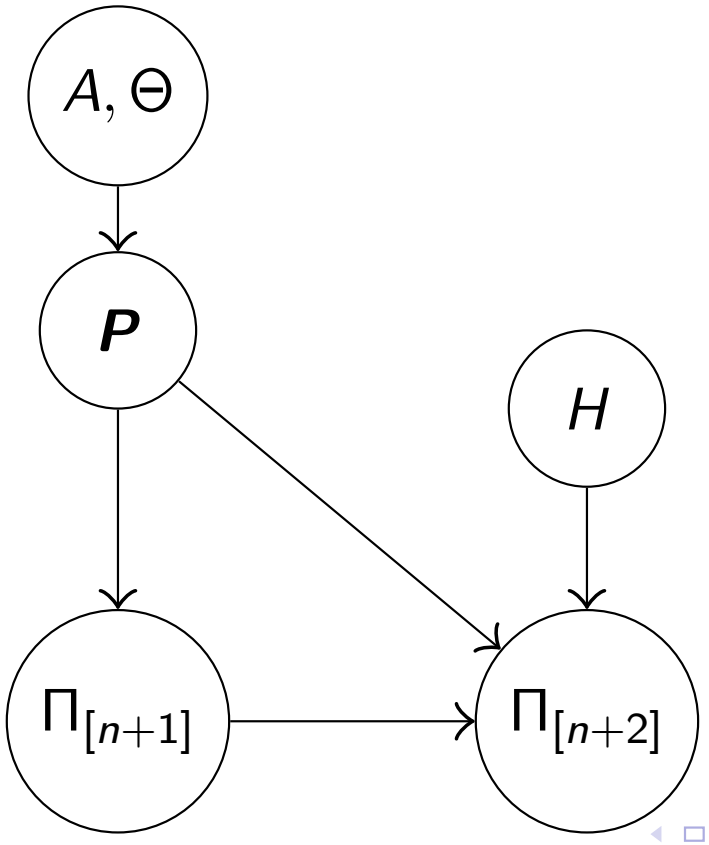
The model

$\Pi_{[n]}$ obtained from the database
$\Pi_{[n+1]} = \{\Pi_{[n]}, \{n+1\}\}$
$\Pi_{[n+2]} = \{\Pi_{[n+1]}, \{n+1, n+2\}\}$
The defence and prosecution disagree on the probability that the $n+2$nd observation (crime stain) goes in the same class of $n+1$st.



The model

$\Pi_{[n]}$ obtained from the database
$\Pi_{[n+1]} = \{\Pi_{[n]}, \{n+1\}\}$
$\Pi_{[n+2]} = \{\Pi_{[n+1]}, \{n+1, n+2\}\}$
The defence and prosecution disagree on the probability that the $n+2$nd observation (crime stain) goes in the same class of $n+1$st.

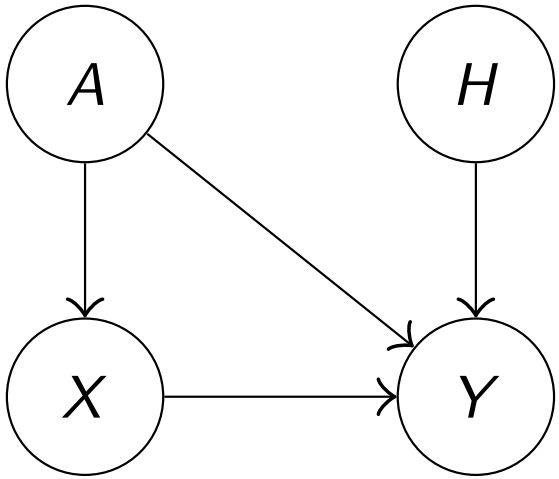


Lemma

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

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

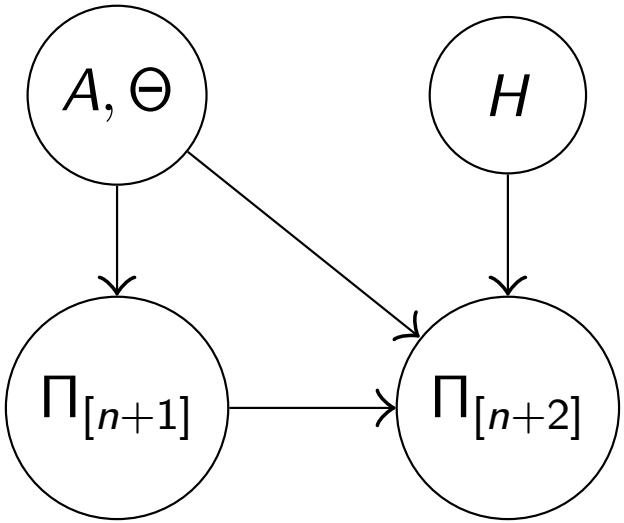
Lemma



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



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

Likelihood ratio

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

Lemma allows to write

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

Result:

For any n, the random partition $\Pi_{[n]}$ obtained form an i.i.d sample from $\mathbf{P}$ ($\sim PD(\alpha, \theta)$), is distributed as the random partition generated by the seating plan of n customers according the the Chinese Restaurant process with parameters α and θ.

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

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

Result:

For any n, the random partition $\Pi_{[n]}$ obtained form an i.i.d sample from $\mathbf{P}$ ($\sim PD(\alpha, \theta)$), is distributed as the random partition generated by the seating plan of n customers according the the Chinese Restaurant process with parameters α and θ.

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

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

Result:

For any n, the random partition $\Pi_{[n]}$ obtained form an i.i.d sample from $\mathbf{P}$ ($\sim PD(\alpha, \theta)$), is distributed as the random partition generated by the seating plan of n customers according the the Chinese Restaurant process with parameters α and θ.

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

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

Result:

For any n, the random partition $\Pi_{[n]}$ obtained form an i.i.d sample from $\mathbf{P}$ ($\sim PD(\alpha, \theta)$), is distributed as the random partition generated by the seating plan of n customers according the the Chinese Restaurant process with parameters α and θ.

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

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

Result:

For any n, the random partition $\Pi_{[n]}$ obtained form an i.i.d sample from $\mathbf{P}$ ($\sim PD(\alpha, \theta)$), is distributed as the random partition generated by the seating plan of n customers according the the Chinese Restaurant process with parameters α and θ.

Likelihood ratio

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

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

Likelihood ratio

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

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

Likelihood ratio

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

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

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

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 reparametrize and assume a noninformative prior

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

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

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 reparametrize and assume a noninformative prior

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

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

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

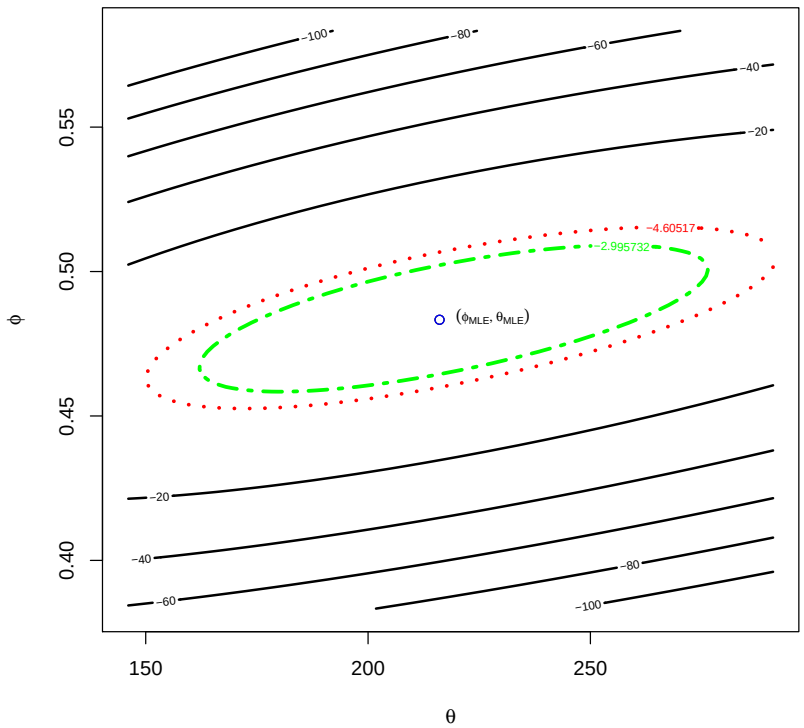
$$\mathrm{LR} = \frac{n}{\mathbb{E}(\Phi \mid \Pi_{[n+1]} = \pi_{[n+1]})}.$$

We reparametrize and assume a noninformative prior

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

Log likelihood with ϕ and θ

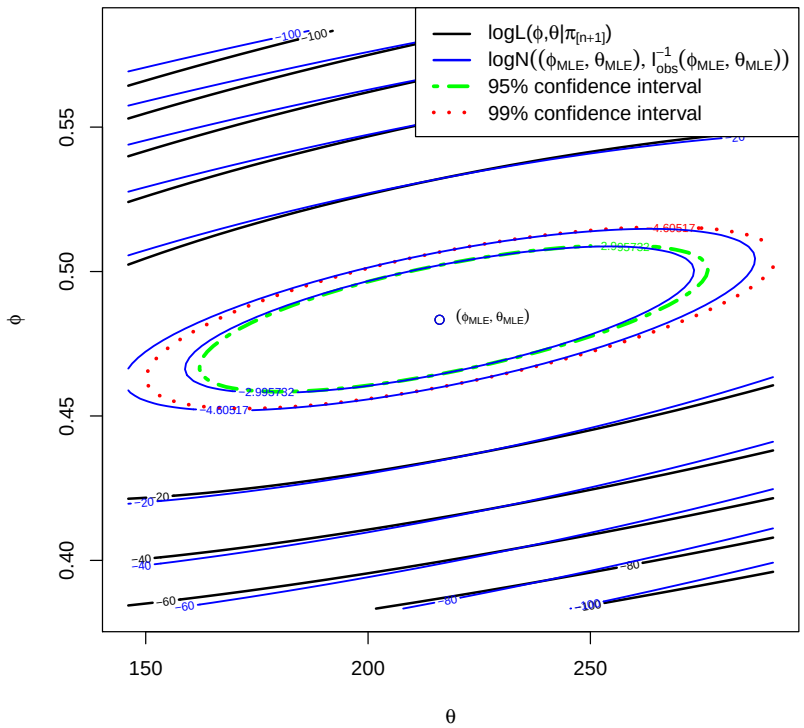
$\log_{10} p(\pi_{[n+1]} \mid \phi, \theta)$



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

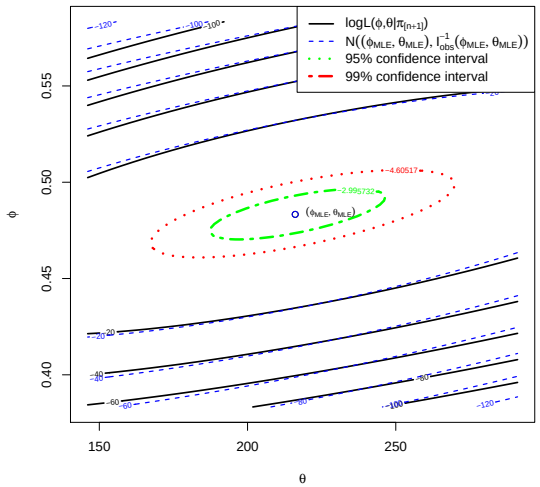
Log likelihood with ϕ and θ

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



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

Log likelihood as a function of ϕ and θ

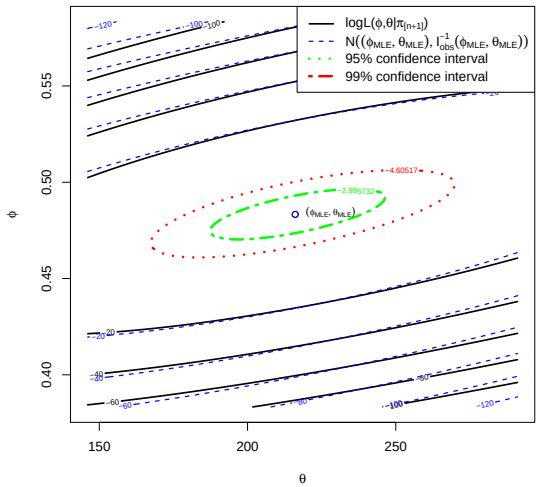


$$p(\phi, \theta \mid \pi_{[n+1]}) \approx N((\phi_{MLE}, \theta_{MLE}), I_{MLE}^{-1}).$$

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

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

Log likelihood as a function of ϕ and θ

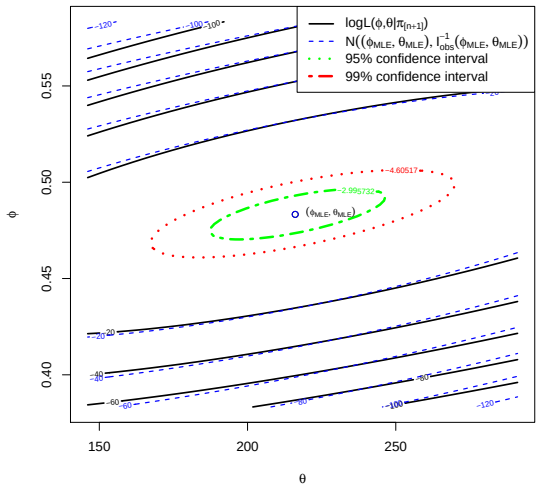


$$p(\phi, \theta \mid \pi_{[n+1]}) \approx N((\phi_{MLE}, \theta_{MLE}), I_{MLE}^{-1}).$$

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

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

Log likelihood as a function of ϕ and θ



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

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

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

An example

Dutch database $N = 2038$

Chinese Restaurant method	LR = 5423
Generalized Good (Cereda, 2017)	LR = 4547
Brenner's k method (Brenner, 2010)	LR = 2472
Augmented database count method	LR = 2039

(Generalized Good method is based on a similar reduction of data. $LR_{GG} = \frac{NN_1}{2N_2}$)

An example

Dutch database $N = 2038$

Chinese Restaurant method	LR = 5423
Generalized Good (Cereda, 2017)	LR = 4547
Brenner's k method (Brenner, 2010)	LR = 2472
Augmented database count method	LR = 2039

(Generalized Good method is based on a similar reduction of data. $LR_{GG} = \frac{NN_1}{2N_2}$)

An example

Dutch database $N = 2038$

Chinese Restaurant method	LR = 5423
Generalized Good (Cereda, 2017)	LR = 4547
Brenner's k method (Brenner, 2010)	LR = 2472
Augmented database count method	LR = 2039

(Generalized Good method is based on a similar reduction of data. $LR_{GG} = \frac{NN_1}{2N_2}$)

An example

Dutch database $N = 2038$

Chinese Restaurant method	LR = 5423
Generalized Good (Cereda, 2017)	LR = 4547
Brenner's k method (Brenner, 2010)	LR = 2472
Augmented database count method	LR = 2039

(Generalized Good method is based on a similar reduction of data. $LR_{GG} = \frac{NN_1}{2N_2}$)