Bayesian Inference and Hypothesis Testing

Christos Dimitrakakis

October 23, 2024

Conditional Probability and the Theorem of Bayes

Simple Bayesian hypothesis testing

Bayesian Inference

Bayes theorem

- Recall the definition of Conditional probability:

$$P(A|B) = P(A \cap B)/P(B)$$

i.e. the probability of A happening if B happens.

Bayes theorem

- Recall the definition of Conditional probability:

$$P(A|B) = P(A \cap B)/P(B)$$

  i.e. the probability of A happening if B happens.
- It is also true that:

$$P(B|A) = P(A \cap B)/P(A)$$

Bayes theorem

- Recall the definition of Conditional probability:

$$P(A|B) = P(A \cap B)/P(B)$$

  i.e. the probability of A happening if B happens.
- It is also true that:

$$P(B|A) = P(A \cap B)/P(A)$$

- Combining the two equations, reverse the conditioning:

$$P(A|B) = P(B|A)P(A)/P(B)$$

Bayes theorem

- Recall the definition of Conditional probability:

$$P(A|B) = P(A \cap B)/P(B)$$

  i.e. the probability of A happening if B happens.
- It is also true that:
$$P(B|A) = P(A \cap B)/P(A)$$
- Combining the two equations, reverse the conditioning:

$$P(A|B) = P(B|A)P(A)/P(B)$$

- So we can reverse the order of conditioning, i.e. relate to the probability of A given B to that of B given A.

The cards problem

1. Print out a number of cards, with either [A|A], [A|B] or [B|B] on their sides.
2. If you have an A, what is the probability of an A on the other side?
3. Have the students perform the experiment with:
   3.1 Draw a random card.
   3.2 Count the number of people with A.
   3.3 What is the probability that somebody with an A on one side will have an A on the other?
   3.4 Half of the people should have an A?

The cards problem

1. Print out a number of cards, with either [A|A], [A|B] or [B|B] on their sides.
2. If you have an A, what is the probability of an A on the other side?
3. Have the students perform the experiment with:
   3.1 Draw a random card.
   3.2 Count the number of people with A.
   3.3 What is the probability that somebody with an A on one side will have an A on the other?
   3.4 Half of the people should have an A?

The prior and posterior probabilities

A	A	2/6	A observed	2/3
A	B	1/6	A observed	1/3
B	A	1/6		
B	B	2/6		

Bayesian Inference

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

Prior elicitation

- All those that think the accused is guilty, raise their hand.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

Prior elicitation

- All those that think the accused is guilty, raise their hand.
- Divide by the number of people in class

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

Prior elicitation

- All those that think the accused is guilty, raise their hand.
- Divide by the number of people in class
- Let us call this $P(H_1)$.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

Prior elicitation

- All those that think the accused is guilty, raise their hand.
- Divide by the number of people in class
- Let us call this $P(H_1)$.
- This is a purely subjective measure!

DNA test

- Let us now do a DNA test on the suspect

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 1\%$: False positive rate

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 1\%$: False positive rate
- $P(D|H_1) = 100\%$: True positive rate

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 1\%$: False positive rate
- $P(D|H_1) = 100\%$: True positive rate

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 1\%$: False positive rate
- $P(D|H_1) = 100\%$: True positive rate

Run the test

- The result is either positive or negative ($\neg D$).

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 1\%$: False positive rate
- $P(D|H_1) = 100\%$: True positive rate

Run the test

- The result is either positive or negative ($\neg D$).
- What is your belief now that the suspect is guilty?

Everybody is a suspect

- Run a DNA test on everybody in the house.

Everybody is a suspect

- Run a DNA test on everybody in the house.
- What is different from before?

Everybody is a suspect

- Run a DNA test on everybody in the house.
- What is different from before?
- Who has a positive test?

Everybody is a suspect

- Run a DNA test on everybody in the house.
- What is different from before?
- Who has a positive test?
- What is your belief that the people with the positive test are guilty?

Explanation

- Prior: $P(H_i)$.

Explanation

- Prior: $P(H_i)$.
- Likelihood $P(D|H_i)$.

Explanation

- Prior: $P(H_i)$.
- Likelihood $P(D|H_i)$.
- Posterior: $P(H_i|D) = P(D \cap H_i)/P(D) = P(D|H_i)P(H_i)/P(D)$

Explanation

- Prior: $P(H_i)$.
- Likelihood $P(D|H_i)$.
- Posterior: $P(H_i|D) = P(D \cap H_i)/P(D) = P(D|H_i)P(H_i)/P(D)$
- Marginal probability: $P(D) = P(D|H_0)P(H_0) + P(D|H_1)P(H_1)$

Explanation

- Prior: $P(H_i)$.
- Likelihood $P(D|H_i)$.
- Posterior: $P(H_i|D) = P(D \cap H_i)/P(D) = P(D|H_i)P(H_i)/P(D)$
- Marginal probability: $P(D) = P(D|H_0)P(H_0) + P(D|H_1)P(H_1)$
- Posterior: $P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0)+P(D|H_1)P(H_1)}$

Explanation

- Prior: $P(H_i)$.
- Likelihood $P(D|H_i)$.
- Posterior: $P(H_i|D) = P(D \cap H_i)/P(D) = P(D|H_i)P(H_i)/P(D)$
- Marginal probability: $P(D) = P(D|H_0)P(H_0) + P(D|H_1)P(H_1)$
- Posterior: $P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0)+P(D|H_1)P(H_1)}$
- Assuming $P(D|H_1) = 1$, and setting $P(H_0) = q$, this gives

$$P(H_0|D) = \frac{0.1q}{0.1q + 1 - q} = \frac{q}{10 - 9q}$$

Explanation

- Prior: $P(H_i)$.
- Likelihood $P(D|H_i)$.
- Posterior: $P(H_i|D) = P(D \cap H_i)/P(D) = P(D|H_i)P(H_i)/P(D)$
- Marginal probability: $P(D) = P(D|H_0)P(H_0) + P(D|H_1)P(H_1)$
- Posterior: $P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0)+P(D|H_1)P(H_1)}$
- Assuming $P(D|H_1) = 1$, and setting $P(H_0) = q$, this gives

$$P(H_0|D) = \frac{0.1q}{0.1q + 1 - q} = \frac{q}{10 - 9q}$$

- The posterior can always be updated with more data!

Python example

```
# the input to the function is the prior, the likelihood function, a
# Input:
# - prior for hypothesis 0 (scalar)
# - data (single data point)
# - likelihood[data][hypothesis] array unction
# Returns:
# - posterior for the data point (if multiple points are given, the
def get_posterior(prior, data, likelihood):
    marginal = prior * likelihood[data][0] + (1 - prior) * likelihoo
    posterior = prior * likelihood[data][0] / marginal
    return posterior


import numpy as np
prior = 0.9
likelihood = np.zeros([2, 2])
# pr of negative test if not a match
likelihood[0][0] = 0.9
# pr of positive test if not a match
likelihood[1][0] = 0.1
# pr of negative test if a match
```

Types of hypothesis testing problems

Simple Hypothesis Test

Example: DNA evidence, Covid tests

- Two hypothesese H_0, H_1
- $P(D|H_i)$ is defined for all i

Multiple Hypotheses Test

Example: Model selection

- H_i: One of many mutually exclusive models
- $P(D|H_i)$ is defined for all i

Null Hypothesis Test

Example: Are men's and women's heights the same?

- H_0: The 'null' hypothesis
- $P(D|H_0)$ is defined
- The alternative is undefined

Pitfalls

Problem definition

- Defining the models $P(D|H_i)$ incorrectly.

The garden of many paths

Pitfalls

Problem definition

- Defining the models $P(D|H_i)$ incorrectly.
- Using an "unreasonable" prior $P(H_i)$

The garden of many paths

Pitfalls

Problem definition

- Defining the models $P(D|H_i)$ incorrectly.
- Using an "unreasonable" prior $P(H_i)$

The garden of many paths

- Having a huge hypothesis space

Pitfalls

Problem definition

- Defining the models $P(D|H_i)$ incorrectly.
- Using an "unreasonable" prior $P(H_i)$

The garden of many paths

- Having a huge hypothesis space
- Selecting the relevant hypothesis after seeing the data

Bayesian Inference

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$

Bayesian Inference

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$
- Each model β assigns probabilities $P_\beta(x)$ to possible $x \in X$.

Bayesian Inference

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$
- Each model β assigns probabilities $P_\beta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution P over the parameters.

Bayesian Inference

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$
- Each model β assigns probabilities $P_\beta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution P over the parameters.
- Given x, we calculate the posterior distribution

Bayesian Inference

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$
- Each model β assigns probabilities $P_\beta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution P over the parameters.
- Given x, we calculate the posterior distribution

Bayesian Inference

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$
- Each model β assigns probabilities $P_\beta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution P over the parameters.
- Given x, we calculate the posterior distribution

$$P(\beta|x) = \frac{P_\beta(x)P(\beta)}{\sum_{\beta' \in \mathcal{B}} P_{\beta'}(x)P(\beta')}, \qquad \text{(finite } \mathcal{B}, P \text{ is a probability)}$$

$$p(\beta|x) = \frac{P_\beta(x)p(\beta)}{\int_{\mathcal{B}} P_{\beta'}(x)p(\beta')d\beta'}, \qquad \text{(continuous } \mathcal{B}, p \text{ is a density)}$$

$$P(B|x) = \frac{\int_B P_{\beta'}(x)dP(\beta)}{\int_{\mathcal{B}} P_{\beta'}(x)dP(\beta)}, \qquad B \subset \mathcal{B} \qquad \text{(arbitrary } \mathcal{B}, P \text{ is a measure)}$$

Bayesian Inference

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$
- Each model β assigns probabilities $P_\beta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution P over the parameters.
- Given x, we calculate the posterior distribution

$$P(\beta|x) = \frac{P_\beta(x)P(\beta)}{\sum_{\beta' \in \mathcal{B}} P_{\beta'}(x)P(\beta')}, \qquad \text{(finite } \mathcal{B}, P \text{ is a probability)}$$

$$p(\beta|x) = \frac{P_\beta(x)p(\beta)}{\int_{\mathcal{B}} P_{\beta'}(x)p(\beta')d\beta'}, \qquad \text{(continuous } \mathcal{B}, p \text{ is a density)}$$

$$P(B|x) = \frac{\int_B P_{\beta'}(x)dP(\beta)}{\int_{\mathcal{B}} P_{\beta'}(x)dP(\beta)}, \qquad B \subset \mathcal{B} \qquad \text{(arbitrary } \mathcal{B}, P \text{ is a measure)}$$

Alternative notation for different probability spaces

- The prior $P(\beta) = \mathbb{P}(\beta)$ and posterior $P(\beta \mid x) = \mathbb{P}(\beta \mid x)$ belief.
- The likelihood $P_\beta(x) = \mathbb{P}(x \mid \beta)$
- The marginal $\mathbb{P}_P(x) = \sum_\beta P_\beta(x)P(\beta)$.

Probabilistic machine learning

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$

Maximum likelihood approach

- Model selection: $\beta^*_{ML}(x) = \arg\max_\beta P_\beta(x)$.
- Model prediction: $P_{\beta^*_{ML}(x)}(x_{t+1})$

Maximum a posteriori approach

- Model selection: $\beta^*_{MAP}(x) = \arg\max_\beta P_\beta(x) P(\beta)$.
- Model prediction: $P_{\beta^*_{MAP}(x)}(x_{t+1})$

Bayesian approach

- Posterior calculation: $P(\beta|x) = P_\beta(x) P(\beta) / \mathbb{P}_P(x)$
- Model prediction: $\mathbb{P}_P(x_{t+1}|x) = \sum_\beta P_\beta(x_{t+1}) P(\beta|x)$

Probabilistic machine learning

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$
- Prior P on $\mathcal{B}$

Maximum likelihood approach

- Model selection: $\beta^*_{ML}(x) = \arg\max_\beta P_\beta(x)$.
- Model prediction: $P_{\beta^*_{ML}(x)}(x_{t+1})$

Maximum a posteriori approach

- Model selection: $\beta^*_{MAP}(x) = \arg\max_\beta P_\beta(x) P(\beta)$.
- Model prediction: $P_{\beta^*_{MAP}(x)}(x_{t+1})$

Bayesian approach

- Posterior calculation: $P(\beta|x) = P_\beta(x) P(\beta) / \mathbb{P}_P(x)$
- Model prediction: $\mathbb{P}_P(x_{t+1}|x) = \sum_\beta P_\beta(x_{t+1}) P(\beta|x)$

Probabilistic machine learning

- Model family $\{P_\beta | \beta \in \mathcal{B}\}$
- Prior P on $\mathcal{B}$
- Observations $x = x_1, \ldots, x_t$.

Maximum likelihood approach

- Model selection: $\beta^*_{ML}(x) = \arg\max_\beta P_\beta(x)$.
- Model prediction: $P_{\beta^*_{ML}(x)}(x_{t+1})$

Maximum a posteriori approach

- Model selection: $\beta^*_{MAP}(x) = \arg\max_\beta P_\beta(x) P(\beta)$.
- Model prediction: $P_{\beta^*_{MAP}(x)}(x_{t+1})$

Bayesian approach

- Posterior calculation: $P(\beta | x) = P_\beta(x) P(\beta) / \mathbb{P}_P(x)$
- Model prediction: $\mathbb{P}_P(x_{t+1} | x) = \sum_\beta P_\beta(x_{t+1}) P(\beta | x)$

Differences between approaches

Maximum likelihood approach

- Ignores model complexity
- Is an optimisation problem

Maximum a posteriori approach

- Regularises model selection using the prior
- Can be seen as solving the optimisation problem

$$\max_{\beta} \ln P_{\beta}(x) + \ln P(\beta),$$

where the prior term $\ln P(\beta)$ acts as a regulariser.

Bayesian approach

- Does not select a single model
- Averages over all models according to their fit and the prior
- Does not result in an optimisation problem.

The n-meteorologists problem

- Consider n meteorological stations $\{\mu\}$ predicting rainfall.
- $x_t \in \{0, 1\}$ with $x_t = 1$ if it rains on day t.
- We have a prior distribution $P(\mu)$ for each station.
- At time t, station μ makes as a prediction $P_\mu(x_{t+1}|x_1, \ldots, x_t)$
- We observe x_{t+1} and calculate the posterior $P(\mu|x_1, \ldots, x_t, x_{t+1})$.

The marginal distribution

To take into account all stations, we can marginalise:

$$\mathbb{P}_P(x_{t+1} \mid x_1, \ldots x_t) = \sum_\mu P_\mu(x_{t+1}|x_t)P(\mu)$$

The posterior

- Show that

$$P(\mu \mid x_1, \ldots, x_{t+1}) = \frac{P_\mu(x_t \mid x_1, \ldots, x_t)P(\mu|x_1, \ldots, x_t)}{\sum_{\mu'} P_{\mu'}(x_t \mid x_1, \ldots, x_t)P(\mu'|x_1, \ldots, x_t)}$$

- How would you implement an ML or a MAP solution to this problem?

Sufficient statistics

A statistic f

This is any function $f : X \to S$ where

- X is the data space
- S is an arbitrary space

Example statistics for $X = \mathbb{R}^*$ (the set of all real-valued sequences)

- The sample mean of a sequence $1/T \sum_{t=1}^{T} x_t$
- The total number of samples T

Sufficient statistic

f is sufficient for a family $\{P_\beta : \beta \in \mathcal{B}\}$ when

$$f(x) = f(x') \Rightarrow P_\beta(x) = P_\beta(x') \forall \beta \in \mathcal{B}.$$

If there exists a finite-dimensional sufficient statistic, Bayesian and ML learning can be done in closed form within the family.

Conjugate priors

Consider a parametrised family of priors $\mathcal{P}$ on $\mathcal{B}$ and a distribution family $\{P_\beta\}$
The pair is conjugate if, for any prior $P \in \mathcal{P}$, and any observation x, there exists $P' \in \mathcal{P}$ such that $P'(\beta) = P(\beta|x)$

Standard Parametric conjugate families

Prior	Likelihood	Parameters β	Observations x
Beta	Bernoulli	$[0, 1]$	$\{0, 1\}^T$
Multinomial	Dirichlet	^{n}	$\{1, \ldots, n\}^T$
Gamma	Normal	$\mathbb{R}, \mathbb{R}$	$\mathbb{R}^T$
Wishart	Normal	$\mathbb{R}^n, \mathbb{R}^{n \times n}$	$\mathbb{R}^{n \times T}$

The Simplex $^n = \{\boldsymbol{\beta} \in [0, 1]^n : \|\boldsymbol{\beta}\|_1\}$ is the set of all n-dimensional probability vectors.

Extensions

- Discrete Bayesian Networks.
- Linear-Gaussian Models (i.e. Bayesian linear regression)
- Gaussian Processes.

Beta-Bernoulli

Definition of the Bernoulli distribution

If $x_t \mid \beta \sim \mathrm{Bernoulli}(\beta)$. $\beta \in [0, 1]$, $x_t \in \{0, 1\}$ and:

$$P_\beta(x_t = 1) = \beta$$

Definition of the Beta density

If $\beta \sim \mathrm{Beta}(\alpha_1, \alpha_0)$, $\alpha_0, \alpha_1 > 0$ and

$$p(\beta | \alpha_1, \alpha_0) \propto \beta^{\alpha_1 - 1}(1 - \beta)^{\alpha_0 - 1}$$

Beta-Bernoulli conjugate pair

- $\beta \sim \mathrm{Beta}(\alpha_1, \alpha_0)$.
- $x_t \mid \beta \sim \mathrm{Bernoulli}(\beta)$.

Then, for any $x = x_1, \ldots, x_T$, the posterior distribution is

- $\beta \mid x \sim \mathrm{Beta}(\alpha_1 + \sum_t x_t, \alpha_0 + T - \sum_t x_t)$.

Dirichlet-Multinomial

$$\boldsymbol{\alpha} \longrightarrow \boldsymbol{\beta} \longrightarrow x_t$$

Definition of the Multinomial distribution

If $x_t \mid \boldsymbol{\beta} \sim (\boldsymbol{\beta})$, with $\beta \in^n$ and $x_t \in \{1, \ldots, n\}$ and:

$$P_{\boldsymbol{\beta}}(x_t = i) = \beta_i$$

Definition of the Dirichlet density

If $\boldsymbol{\beta} \sim (\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} \in \mathbb{R}^n_+$ then

$$p(\beta|\boldsymbol{\alpha}) \propto \prod_i \beta_i^{\alpha_i - 1}$$

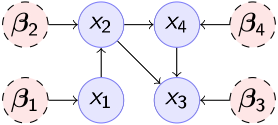
Dirichlet-Multinomial conjugate pair

- $\beta \sim (\boldsymbol{\alpha})$.
- $x_t \mid \beta \sim \mathrm{Bernoulli}(\boldsymbol{\beta})$.

Then, for any $x = x_1, \ldots, x_T$, the posterior distribution is

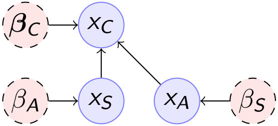
- $\beta \mid x \sim (\boldsymbol{\alpha} + s_T)$, where $s_{T,i} = \sum_{t=1}^{T} \mathbb{I}\{x_t = i\}$.

Discrete Bayesian Networks



- A directed acyclic graph (DAG) defined on variables $x_1, \ldots, x_n$ with each x_n taking a finite number of values,
- Let S_i be the indices corresponding to parent variables of x_i.
- $x_i \mid \boldsymbol{\beta}_i, x_{S_i} = k \sim (\boldsymbol{\beta}_{i,k})$.

Example: Lung cancer, smoking and asbestos

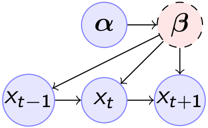


$$P_{\beta_A}(x_A = 1) = \beta_A \tag{1}$$

$$P_{\beta_S}(x_S = 1) = \beta_S \tag{2}$$

$$P_{\boldsymbol{\beta}_C}(x_C = 1 \mid X_A = j, X_S = k) = \beta_{C,j,k} \tag{3}$$

Markov model



A Markov model obeys

$$\mathbb{P}_{\boldsymbol{\beta}}(x_{k+1}|x_k, \ldots, x_1) = \mathbb{P}_{\boldsymbol{\beta}}(x_{k+1}|x_k)$$

i.e. the graphical model is a chain. We are usually interested in homogeneous models, where

$$\mathbb{P}_{\boldsymbol{\beta}}(x_{k+1} = i \mid x_k = j) = \beta_{i,j} \qquad \forall k$$

Inference for finite Markov models

- If $x_t \in [n]$ then $x_{t+1} \mid \boldsymbol{\beta}, x_t = i \sim (\boldsymbol{\beta}_i)$, $\boldsymbol{\beta}_i \in^n$
- Prior $\boldsymbol{\beta}_i \mid \boldsymbol{\alpha} \sim (\boldsymbol{\alpha})$ for all $i \in [n]$.
- Posterior $\boldsymbol{\beta}_i \mid x_1, \ldots, x_t, \boldsymbol{\alpha} \sim (\boldsymbol{\alpha}^{(t)})$ with

$$\alpha^t_{i,j} = \alpha_{i,j} + \sum_{k=1}^{t} \mathbb{I}\left\{x_k = i \wedge x_{k+1} = j\right\}, \qquad \boldsymbol{\alpha}^0 = \boldsymbol{\alpha}.$$