

# Bayesian statistics and latent variable models

Arthur Leroy

`arthur.leroy@inrae.fr`

<https://arthur-leroy.netlify.app/>

M2 Data-Science 2025-2026



# Organization

Wednesdays 8h30 - 11h45, this room.

9 × 3h classes

1h30 class + 1h30 practical session

**Important:** you need one computer/person for practical sessions

# Evaluation

- 1 CC: assiduity, practical session
- 2 Final exam on Wednesday 4th February, 2026
- 3 "Max exam" : the final grade will be  $\max(\text{Exam}, \text{mean}(\text{Exam}, \text{CC}))$

# Bibliography & relevant sources

## General ML / Stats books

- Kevin P. Murphy (2022). *Probabilistic Machine Learning: An introduction*. MIT Press
- Trevor Hastie et al. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA
- Christopher M. Bishop (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer
- Martin J Wainwright et al. (2008). *Graphical models, exponential families, and variational inference*. Vol. 1. 1–2. Now Publishers, Inc., pp. 1–305

## Some relevant lecture/slides on the topic with a different point-of-view ( $\triangle$ notations)

- S. Robin lectures
- Some lectures of this course on Graphical Models

# Introduction

# Types of statistical learning

## Supervised (this course)

Data  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$  with  $y_i$  an output (*response*) and  $x_i$  some features (*covariates*).

The goal is to learn a good predictor  $\hat{f}$  such that  $y_i \approx \hat{f}(x_i)$  that generalizes well on new data.

## Unsupervised (also this course)

The data  $\mathcal{D} = \{x_i\}_{i=1}^n$  The goal is to learn "interesting" and hidden structure in the data to

- partition the data, aka clustering
- visualize/compress the data, aka dimension reduction

**Generative models:** posit a statistical model on the distribution of  $(X_i)$

## Many flavors in modern ML

semi-supervised, self-supervised, reinforcement learning, multi-task, etc.

# What this course is about...

Bayesian statistics, uncertainty quantification and probabilistic modelling

↪ reasoning in terms of prior and posterior probability distributions

## **Parameter estimation**

Learn from iterative data collection the value of a parameter  $\mathbf{p}$

# What this course is about...

(Discrete) latent variables models for unsupervised learning

~> we will assume the generative process of  $X$  involves an unobserved (latent) variable  $Z$

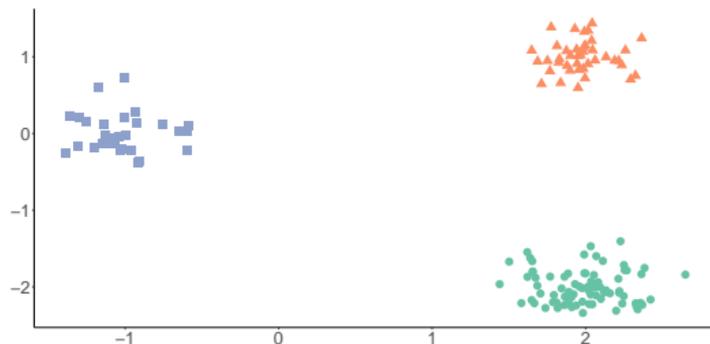
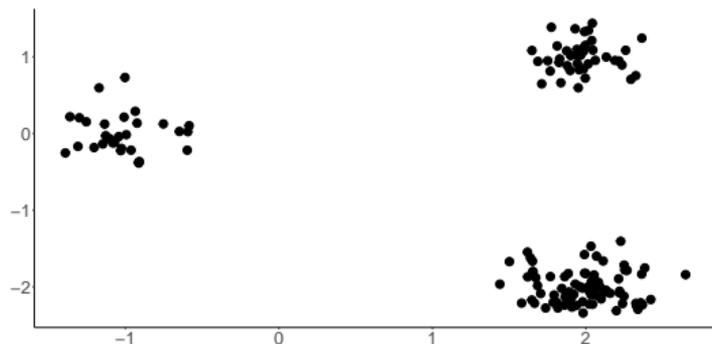
# What this course is about...

(Discrete) latent variables models for unsupervised learning

↪ we will assume the generative process of  $X$  involves an unobserved (latent) variable  $Z$

## Clustering

$X$  is an unlabeled observation and  $Z$  its group membership



# What this course is about...

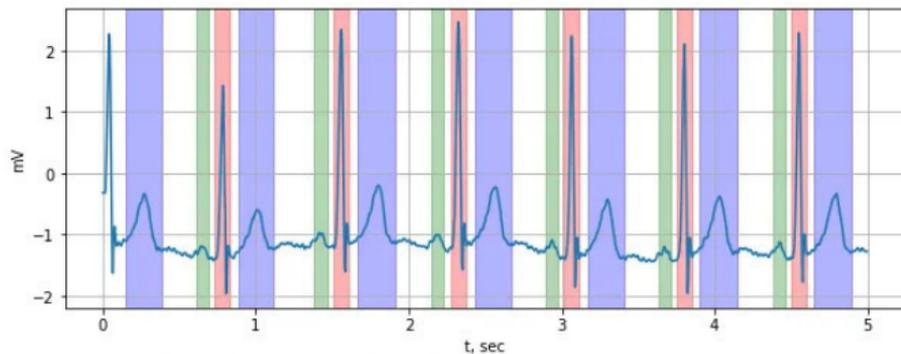
(Discrete) latent variables models for unsupervised learning

↪ we will assume the generative process of  $X$  involves an unobserved (latent) variable  $Z$

## Time series segmentation

$X$  is the temporal signal and  $Z$  the cardiac phase

Heart rate: 78 bpm



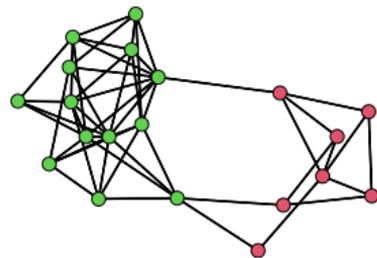
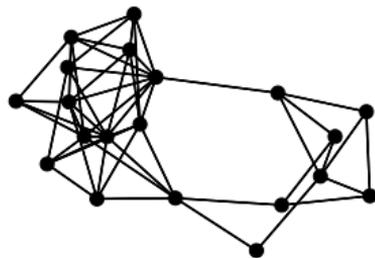
Example of ECG annotation, source: <https://medium.com/data-analysis-center/56f8b9abd83a>

# What this course is about...

(Discrete) latent variables models for unsupervised learning

↪ we will assume the generative process of  $X$  involves an unobserved (latent) variable  $Z$

## Node clustering in a network



$X$  is the graph (connection between node) and  $Z$  the group of the node (community)

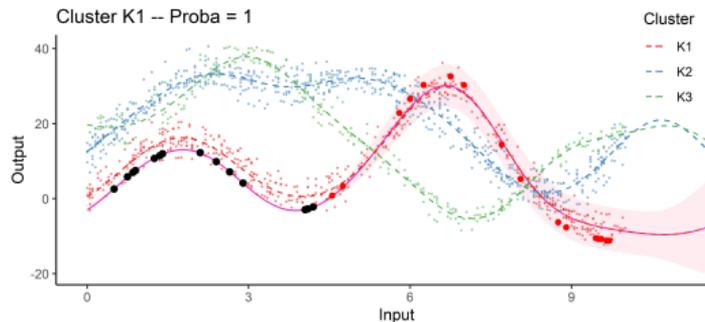
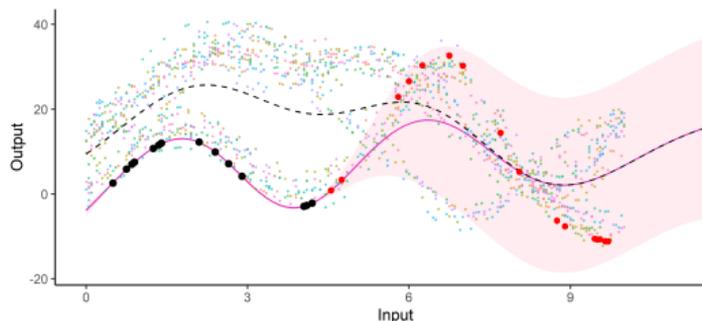
# What this course is about...

(Continuous) latent variables models for both supervised and unsupervised learning

↪ generative processes involving both discrete and continuous (latent) variables  $Z$

## Modelling, predicting and clustering functional data

$X$  is a temporal signal,  $Z$  group memberships and  $\mu$  latent mean processes



# Course outline

- 1** Fundamentals of Bayesian statistics
- 2** Hierarchical modelling and sampling methods
- 3** Clustering and mixture models
- 4** Inference in latent variable models: the EM algorithm
- 5** Hidden Markov Models (HMMs)
- 6** Probabilistic learning and Gaussian processes
- 7** Approximation schemes for probabilistic inference
- 8** Multi-task Gaussian processes and clustering
- 9** Stochastic Block Model or project on real data (your choice)

# Fundamentals of Bayesian statistics

# Let's start with a guessing game

Assume there exists some disease such that:

- 1 person out of 1000 contracted the disease on average,
- a detection test is 99% reliable if you have the disease,
- this same detection test is 99% reliable if you **don't** have the disease,

What is the probability to have contracted the disease if the result of your detection test was positive?

# Frequentist inference

**Assumption:** the observation  $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$  is a realization of a random vector  $\mathbf{X} = \{X_1, \dots, X_n\}$  with distribution  $p_{\theta^*}$ .

**Posit:** a statistical model  $\{p_{\theta}, \theta \in \Theta\}$ , i.e. a family of parametric distribution on  $\mathcal{X}^n$

**Goal:** Provide an estimate  $\hat{\theta}$  of  $\theta^*$ .<sup>1</sup>

## Maximum-likelihood estimation

Find the model, hence  $\theta$ , that maximizes the probability of having seen the data

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} \log p_{\theta}(x_1, \dots, x_n) \quad (\text{MLE})$$

---

<sup>1</sup>and eventually derive theoretical guarantees such as convergence and confidence intervals on  $\hat{\theta}_n(X_1, \dots, X_n)$  (e.g. via central limit theorem)

# What is Uncertainty?

**Uncertainty** represents our lack of knowledge or inherent randomness in data and models.

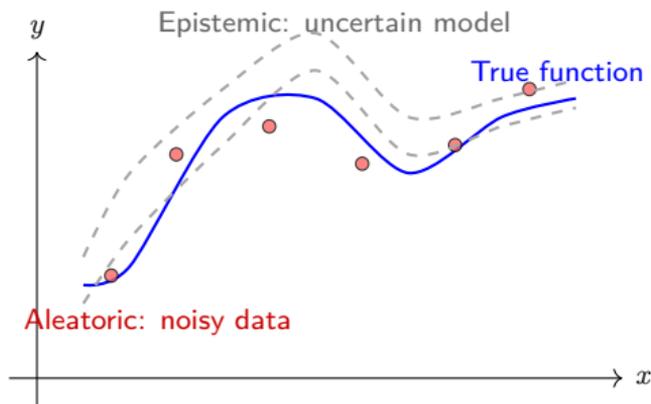
**Two main types of uncertainty:**

## Aleatoric (Data) Uncertainty

- Inherent randomness in the process.
- Cannot be reduced by more data.
- *Example:* Sensor noise, dice roll.

## Epistemic (Model) Uncertainty

- Due to lack of knowledge or limited data.
- Can be reduced with more observations.
- *Example:* Unknown model parameters.



**Bayesian Statistics** quantifies both types via probabilities — anything unobserved is treated as random.

# The Bayesian paradigm

Maximum-likelihood and frequentist statistics produces *point estimates*

## Paradigm shift: random parameters

Parameters  $\theta$  are no longer treated as deterministic but as *random* quantities. The *prior* distribution, denoted as  $p(\theta)$ , encodes knowledge & uncertainty we have on the parameters **before** seeing new data.

↪ the goal is to update this a priori knowledge when new data comes: this is the essence of Bayes formula.

## A bit of history...

The terminology *Bayesian* has been coined that way thanks to the work of Reverend Thomas Bayes (1701-1761) and his posthumous *essay in view of solving the doctrine of chance*. Pierre-Simon Laplace independently proposed a version in 1774.

N.B. : this course will not settle the somewhat sterile debate "Bayesian VS Frequentist".

# Bayes formula

Equipped with a prior  $p(\theta)$ , we posit an observational model on  $X | \theta \iff$  the likelihood. Bayesian modelization essentially adds one layer to frequentist models : the prior.

1.  $\theta \sim p(\cdot)$ , (prior)
2.  $X | \theta \sim p(\cdot | \theta) = p_\theta$  (likelihood).

## The posterior

Given a realization  $x$ , we update our prior via a new distribution called the *posterior*:

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}, \quad (\text{Bayes formula})$$

Where,  $p(x) = \int_{\Theta} p(x | \theta)p(\theta) d\theta$  is a normalization constant, independent of  $\theta$ . Thus, it is common to write<sup>a</sup>

$$p(\theta | x) \propto p(x | \theta)p(\theta)$$

---

<sup>a</sup>Although computing this normalization constant is generally a challenging task in Bayesian statistics.

## Think as proper Bayesian about our previous enigma

- $\mathbb{P}(D) = 0.001$ , 1 person out of 1000 contracted the disease on average,
- $\mathbb{P}(T | D) = 0.99$ , a detection test is 99% reliable if you have the disease,
- $\mathbb{P}(\bar{T} | \bar{D}) = 0.99$ , this same detection test is 99% reliable if you **don't** have the disease,

## Think as proper Bayesian about our previous enigma

- $\mathbb{P}(D) = 0.001$ , 1 person out of 1000 contracted the disease on average,
- $\mathbb{P}(T | D) = 0.99$ , a detection test is 99% reliable if you have the disease,
- $\mathbb{P}(\bar{T} | \bar{D}) = 0.99$ , this same detection test is 99% reliable if you **don't** have the disease,

$$\begin{aligned}\mathbb{P}(T) &= \mathbb{P}(T \cap D) + \mathbb{P}(T \cap \bar{D}) \\ &= \mathbb{P}(T|D) \times \mathbb{P}(D) + \mathbb{P}(T|\bar{D}) \times \mathbb{P}(\bar{D}) \\ &= \mathbb{P}(T|D) \times \mathbb{P}(D) + (1 - \mathbb{P}(\bar{T}|\bar{D})) \times (1 - \mathbb{P}(D)) \\ &= 0.99 \times 0.001 + (1 - 0.99) \times (1 - 0.001) \\ &= 0.00099 + 0.00999 \\ &= 0.01098\end{aligned}$$

## Think as proper Bayesian about our previous enigma

- $\mathbb{P}(D) = 0.001$ , 1 person out of 1000 contracted the disease on average,
- $\mathbb{P}(T | D) = 0.99$ , a detection test is 99% reliable if you have the disease,
- $\mathbb{P}(\bar{T} | \bar{D}) = 0.99$ , this same detection test is 99% reliable if you **don't** have the disease,

$$\begin{aligned}\mathbb{P}(T) &= \mathbb{P}(T \cap D) + \mathbb{P}(T \cap \bar{D}) \\ &= \mathbb{P}(T|D) \times \mathbb{P}(D) + \mathbb{P}(T|\bar{D}) \times \mathbb{P}(\bar{D}) \\ &= \mathbb{P}(T|D) \times \mathbb{P}(D) + (1 - \mathbb{P}(\bar{T}|\bar{D})) \times (1 - \mathbb{P}(D)) \\ &= 0.99 \times 0.001 + (1 - 0.99) \times (1 - 0.001) \\ &= 0.00099 + 0.00999 \\ &= 0.01098\end{aligned}$$

$$\mathbb{P}(D|T) = \frac{\mathbb{P}(T|D) \times \mathbb{P}(D)}{\mathbb{P}(T)} = \frac{0.99 \times 0.001}{0.01098} \approx 0.09016$$

Hence, we only have 9.016% chance to actually be sick despite a positive result to the detection test.

## Think as proper Bayesian: sequential updates

After the first test, we updated our beliefs about chances of being sick, thus:

- $\mathbb{P}(D | T) = 0.09016$  (current probability of the disease),

Imagine we take the same kind of test again, independently, and get a positive results.

## Think as proper Bayesian: sequential updates

After the first test, we updated our beliefs about chances of being sick, thus:

- $\mathbb{P}(D | T) = 0.09016$  (current probability of the disease),

Imagine we take the same kind of test again, independently, and get a positive results.

**We now treat the previous posterior as a new prior for a second independent test:**

$$\begin{aligned}\mathbb{P}(T_2) &= \mathbb{P}(T_2 | D) \mathbb{P}(D | T) + \mathbb{P}(T_2 | \bar{D}) \mathbb{P}(\bar{D} | T) \\ &= 0.99 \times 0.09016 + 0.01 \times (1 - 0.09016) \\ &\approx 0.08926 + 0.00909 \\ &\approx 0.09836\end{aligned}$$

## Think as proper Bayesian: sequential updates

After the first test, we updated our beliefs about chances of being sick, thus:

- $\mathbb{P}(D | T) = 0.09016$  (current probability of the disease),

Imagine we take the same kind of test again, independently, and get a positive results.

**We now treat the previous posterior as a new prior for a second independent test:**

$$\begin{aligned}\mathbb{P}(T_2) &= \mathbb{P}(T_2 | D) \mathbb{P}(D | T) + \mathbb{P}(T_2 | \bar{D}) \mathbb{P}(\bar{D} | T) \\ &= 0.99 \times 0.09016 + 0.01 \times (1 - 0.09016) \\ &\approx 0.08926 + 0.00909 \\ &\approx 0.09836\end{aligned}$$

$$\mathbb{P}(D | T, T_2) = \frac{\mathbb{P}(T_2 | D) \mathbb{P}(D | T)}{\mathbb{P}(T_2)} = \frac{0.99 \times 0.09016}{0.09836} \approx \boxed{0.9075}$$

- After one positive test:  $\approx 9.0\%$  chance of disease.
- After a *second independent* positive test:  $\approx 90.8\%$  chance.

## Odd ratios: comparing hypotheses without computing evidence

Two possible diseases can explain  $T$ :  $H_1$  : Disease A vs.  $H_0$  : Disease B.

- Prior belief (before seeing  $T$ ):

$$P(H_1) = 0.2, \quad P(H_0) = 0.8$$

- Suppose the test result  $T$  is much more typical of Disease A:

$$P(T | H_1) = 0.6, \quad P(T | H_0) = 0.1$$

**Bayes' rule in odds form: Posterior odds = Prior odds  $\times$  Likelihood ratio (Bayes factor)**

$$\frac{P(H_1 | T)}{P(H_0 | T)} = \frac{P(H_1)}{P(H_0)} \times \frac{P(T | H_1)}{P(T | H_0)} \times \frac{\cancel{P(T)}}{\cancel{P(T)}}$$

so the **Bayes factor** is

$$\text{BF} = \frac{0.6}{0.1} = 6.$$

**Posterior odds:**

$$\frac{P(H_1 | T)}{P(H_0 | T)} = \frac{0.2}{0.8} \times \frac{0.6}{0.1} = 0.25 \times 6 = 1.5$$

# Clarifications: Distributions, Densities, Likelihood

## 1. Probability Distribution

- Describes how a random variable  $X$  behaves.
- For discrete  $X$ : probability mass function (PMF)  $p(x) = \mathbb{P}(X = x)$ .
- For continuous  $X$ : probability density function (PDF)  $p(x)$  such that

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x) dx$$

## 2. Probability Density / Mass Function

- PMF or PDF is a *function* that assigns probabilities (discrete) or densities (continuous) to outcomes.
- **Notation:** often  $p(x)$  in both cases.

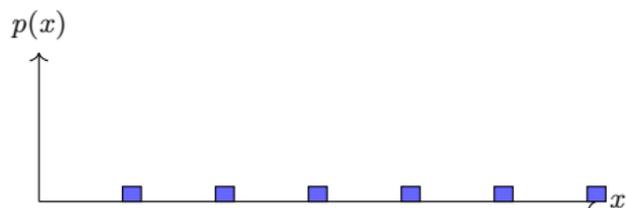
## 3. Likelihood Function

- Treats data as fixed and parameters  $\theta$  as variable.
- Measures “how likely” the observed data is for different parameter values.
- **Notation:**  $\mathcal{L}(\theta | x) = p(x | \theta)$
- Key point: Likelihood is a function of  $\theta$ , not  $x$ !

# Distributions, Densities, and Likelihood: Visual Overview

## Discrete Random Variable

- PMF:  $p(x) = \mathbb{P}(X = x)$
- Example: Dice roll



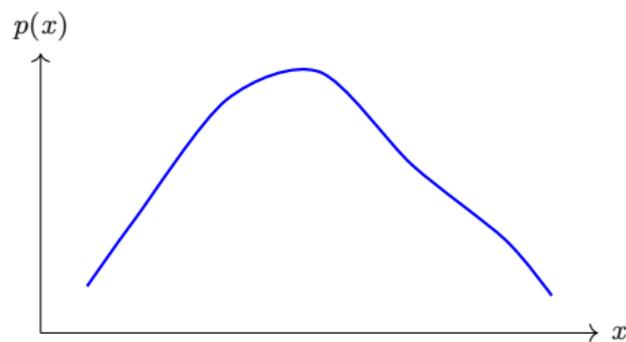
**Likelihood for parameter  $\theta$  (e.g., biased die):**

$$\mathcal{L}(\theta | x) = p_{\theta}(x)$$

Data fixed,  $\theta$  variable.

## Continuous Random Variable

- PDF:  $p(x)$ ,  $\mathbb{P}(a \leq X \leq b) = \int_a^b p(x) dx$
- Example: Heights of students



**Likelihood for parameter  $\theta$  (e.g., mean  $\mu$ ):**

$$\mathcal{L}(\theta | x) = p(x | \theta)$$

Data fixed,  $\theta$  variable.

# Choosing a prior

## Expert knowledge

The prior  $p(\theta)$  may be used to represent any available expert knowledge on  $\theta$ .

## Conjugate priors

When the prior  $p(\theta)$  and the posterior  $p(\theta | x)$  belong to the same family of distributions (e.g. Gaussian, Beta, etc.), then we say that the prior is *conjugate* to the observational model  $p(x | \theta)$ .

Conjugate priors are widely used as they greatly simplify computations.

## Uninformative prior

When the prior equally charges  $\Theta$  we say that the prior is *uninformative*, noted  $p(\theta) \propto 1$ . Obviously,  $p(\theta) \propto 1$  does not always define a proper p.d.f. (consider  $\Theta = \mathbb{R}$ ). Still, as long as the posterior is well defined (i.e. the normalization constant  $Z$  exists and is finite) then we can still use the posterior  $p(\theta | x)$  and the prior is *improper*.

## Example of conjugacy: the Beta-Binomial model (1)

**Experiment & question** Given a sequence of independent coin flips  $\mathbf{x} = \{x_1, \dots, x_n\}$ , determine the probability of getting tail.

### Observational model: the likelihood

Given a probability of tail  $\theta$ , we model the random vector  $\mathbf{X} = (X_1, \dots, X_n)$  as *i.i.d.* Bernoulli  $X_i \sim \text{Ber}(\theta)$  so that

$$p(\mathbf{X} \mid \theta) = \prod_{i=1}^n \text{Ber}(x_i \mid \theta) = \theta^{\sum_i x_i} (1 - \theta)^{\sum_i 1 - x_i}.$$

### Choice of a prior

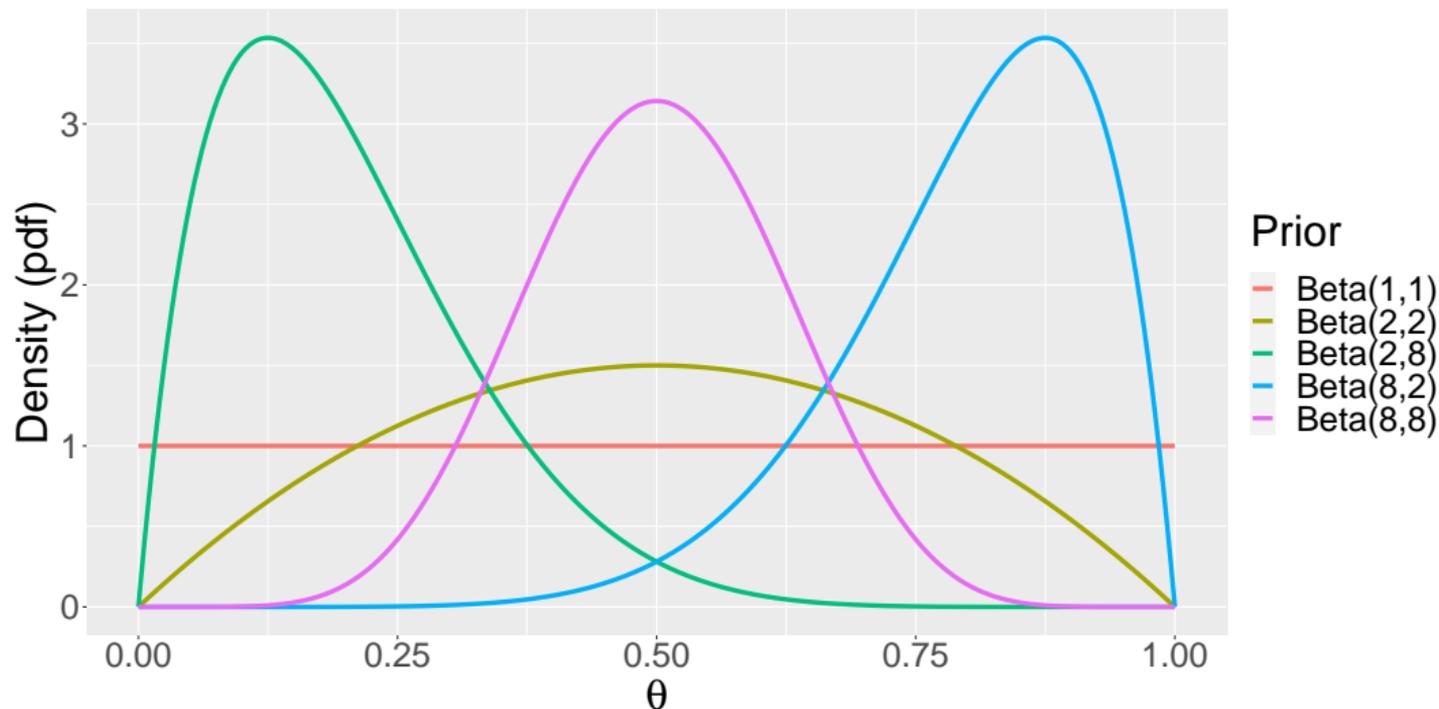
We use Beta distribution with support  $\Theta = [0, 1]$

$$p(\theta) = \text{Beta}(a, b) \propto \mathbf{1}_{[0,1]}(\theta) \theta^{a-1} (1 - \theta)^{b-1}.$$

$a$  and  $b$  are called *hyperparameters* and they control our level of a priori

- $a = b = 1$  : uniform on  $[0, 1]$  (uninformative)
- $a = b > 1$  : in favor of a balanced coin, the greater  $a$ , the stronger the prior
- $a > b$  (resp.  $a < b$ ): in favor of tail (resp. head).

## Example of conjugacy: the Beta-Binomial model (2)



**Figure:** Graph of the p.d.f.  $Beta(\cdot | a, b)$  for different values of  $a$  and  $b$ .

## Example of conjugacy: the Beta-Binomial model (3)

We seek to derive the posterior, and we directly have

$$\begin{aligned} p(\theta | \mathbf{X}) &\propto p(\mathbf{X} | \theta)p(\theta), \\ &\propto \theta^{\sum_i x_i} (1 - \theta)^{\sum_i 1 - x_i} \theta^{a-1} (1 - \theta)^{b-1} \mathbf{1}_{[0,1]}(\theta), \\ &\propto \theta^{a + \sum_i x_i - 1} (1 - \theta)^{b + n - \sum_i x_i - 1} \mathbf{1}_{[0,1]}(\theta). \end{aligned}$$

We recognize the p.d.f of a Beta distribution

$$\theta | \mathbf{X} \sim \text{Beta} \left( a + \sum_i X_i, b + n - \sum_i X_i \right)$$

Remarks :

- 1  $a$  and  $b$  act as *pseudo-counts* for head and tails, smoothing the estimates when  $n$  is small.
- 2 This conjugacy between the Beta prior and the binomial model always hold : property of the model (prior + likelihood) and not our specific experiment.

# Beta-Binomial conjugacy: a visual illustration

# Classical Conjugacy: Gaussian-Gaussian

We observe  $x_1, \dots, x_n$  assumed i.i.d. from a Gaussian likelihood with known variance  $\sigma^2$ :

$$x_i \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$$

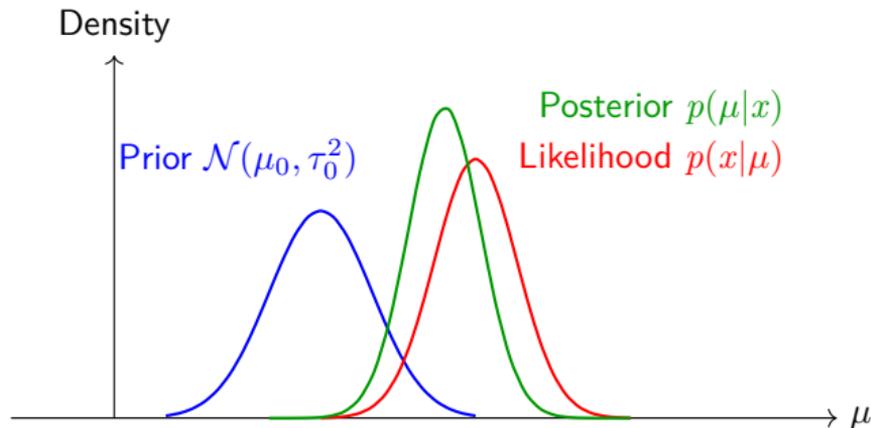
**Prior:** Gaussian on the mean:

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$

**Posterior:** Also Gaussian (conjugacy):  $\mu \mid x_{1:n} \sim \mathcal{N}(\mu_n, \tau_n^2)$ ,

with

$$\tau_n^2 = \left( \frac{n}{\sigma^2} + \frac{1}{\tau_0^2} \right)^{-1}, \quad \mu_n = \tau_n^2 \left( \frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right)$$



# Classical Conjugacy: Multinomial–Dirichlet ( $K = 3$ )

**Setup:** We model categorical data  $x_1, \dots, x_n$  with three possible outcomes:

$$(x_1, \dots, x_n) \mid \boldsymbol{\theta} \sim \text{Multinomial}(n, \boldsymbol{\theta}), \quad \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3), \quad \sum_k \theta_k = 1$$

**Prior:**

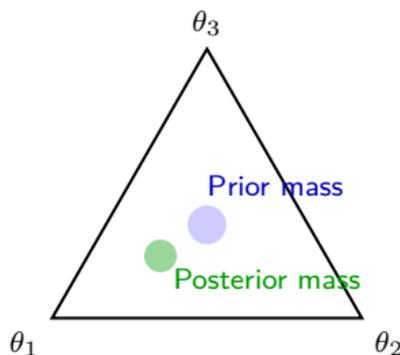
$$\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$$

**Posterior:**

$$\boldsymbol{\theta} \mid \mathbf{x} \sim \text{Dirichlet}(\alpha_1 + n_1, \alpha_2 + n_2, \alpha_3 + n_3)$$

**Example:**  $\alpha = (2, 2, 2)$ ,  $\mathbf{n} = (10, 4, 6)$

$$\Rightarrow \boldsymbol{\theta} \mid \mathbf{x} \sim \text{Dirichlet}(12, 6, 8)$$



# Posterior Mean, Median & Mode

Different cost functions leads to different Bayes estimator among which

- 1 posterior mean  $\hat{\theta} = \mathbb{E}[\theta | x]$  corresponds to the  $L^2$ -loss
- 2 posterior median  $\hat{\theta}$  such that  $p(\theta \geq \hat{\theta} | x) = p(\theta \leq \hat{\theta} | x) = 0.5$  ( $L^1$ -loss)
- 3 **posterior mode (aka MAP)**:  $\hat{\theta} \in \arg \max_{\theta} p(\theta | x)$  (0-1 loss)

Maximum a posteriori is one of the most popular

- reduces to an optimization problem
- log-prior can be interpreted in a frequentist setting as a regularizer for MLE

$$\log p(\theta | x) = cte + \underbrace{\log p_{\theta}(x)}_{\text{likelihood}} + \underbrace{\log p(\theta)}_{\text{regularizer}}$$

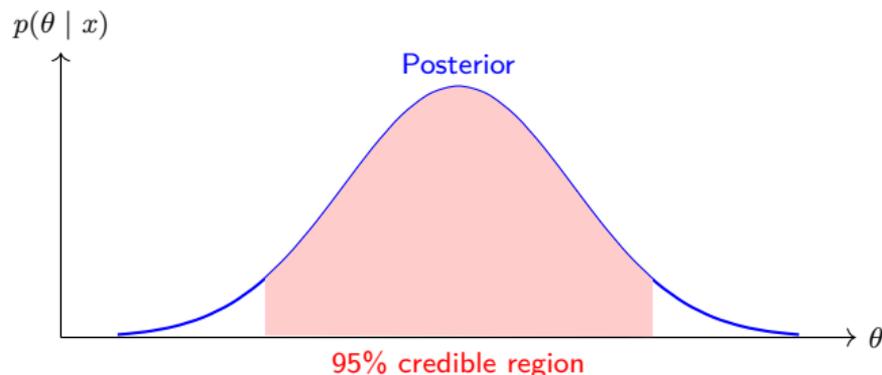
# The main benefit of Bayes: Uncertainty Quantification

The posterior distribution provides a full description of uncertainty about parameters. A **credible region**  $\mathcal{R} \subset \Theta$  is defined by:

$$p(\theta \in \mathcal{R} | x) = \int_{\mathcal{R}} p(\theta | x) d\theta = 1 - \alpha$$

## Interpretation:

- Unlike a *confidence interval*, a credible region means:  $P(\theta \in \mathcal{R} | x) = 0.95$  i.e. “Given the data, there is a 95% probability that  $\theta$  lies in  $\mathcal{R}$ .”
- All uncertainty is captured by the posterior but can be summarized (mean, credible interval, ...).
- If more data are observed, the posterior becomes narrower  $\rightarrow$  less uncertainty.



**Uncertainty instead of binary decisions, posteriors instead of tests**

# Uncertainty instead of binary decisions, posteriors instead of tests

## Frequentist view:

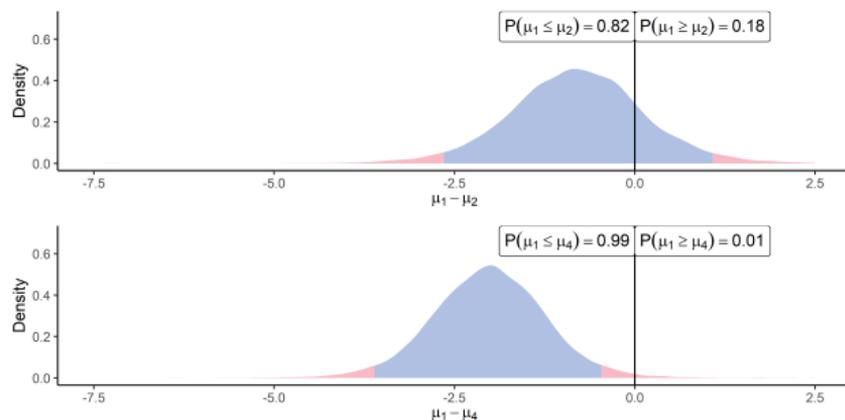
- Decision is **binary**: accept or reject  $H_0$  based on a threshold (e.g.  $p < 0.05$ ).
- Output: a yes/no answer — *no direct measure of uncertainty about hypotheses*.

## Bayesian view:

- Compute the **posterior probability** of each hypothesis or model, e.g.

$$P(H_1 \mid \text{data}) = 0.82, \quad P(H_0 \mid \text{data}) = 0.18$$

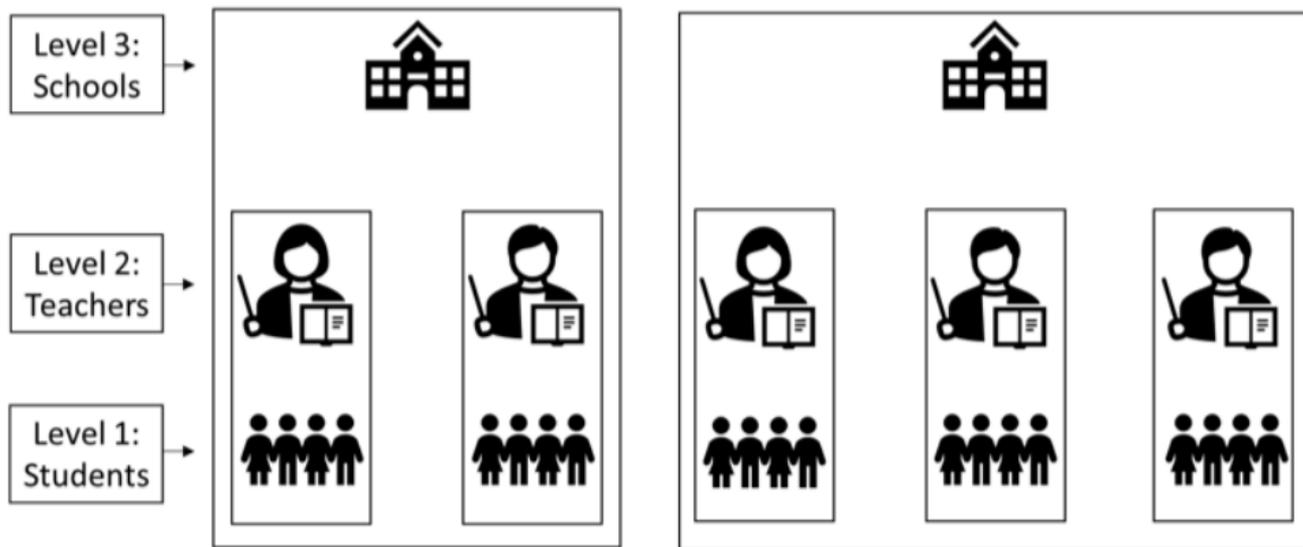
- This expresses **degrees of belief** rather than a binary verdict.
- Richer information: you can compare models, quantify uncertainty, and propagate it.



# Hierarchical modelling and sampling methods

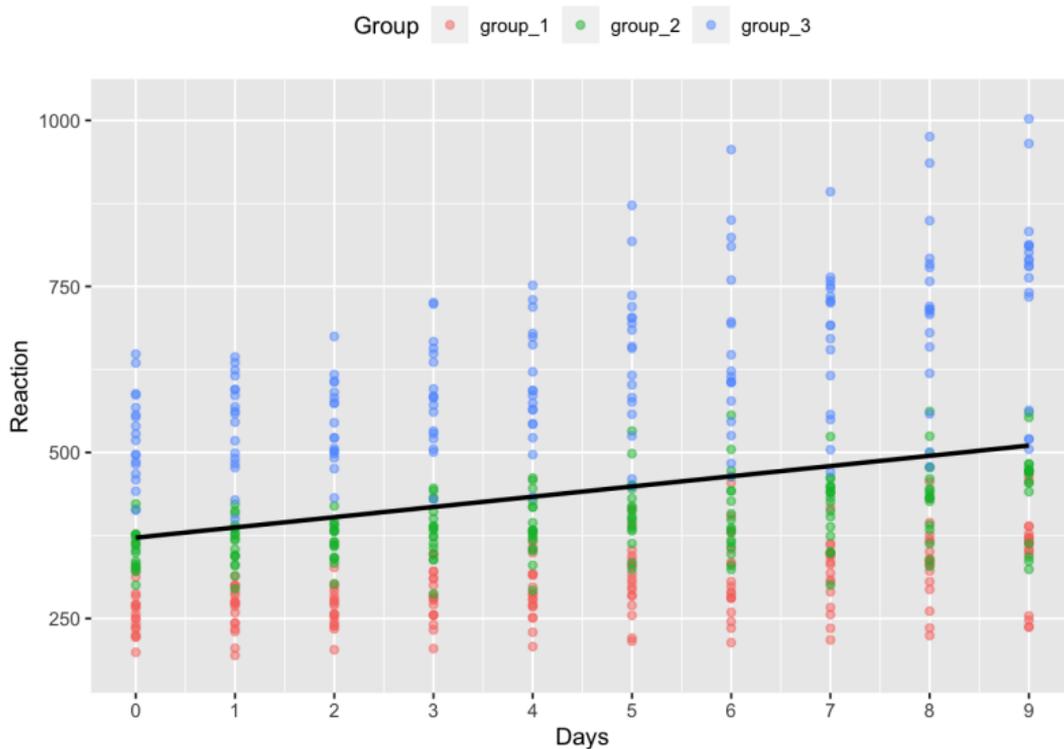
# Motivation: why hierarchical models?

- Data are often grouped (schools, hospitals, customers, sensors)
- Need to share statistical strength across groups (partial pooling)
- Hierarchical models capture multi-level variation and can propagate uncertainty



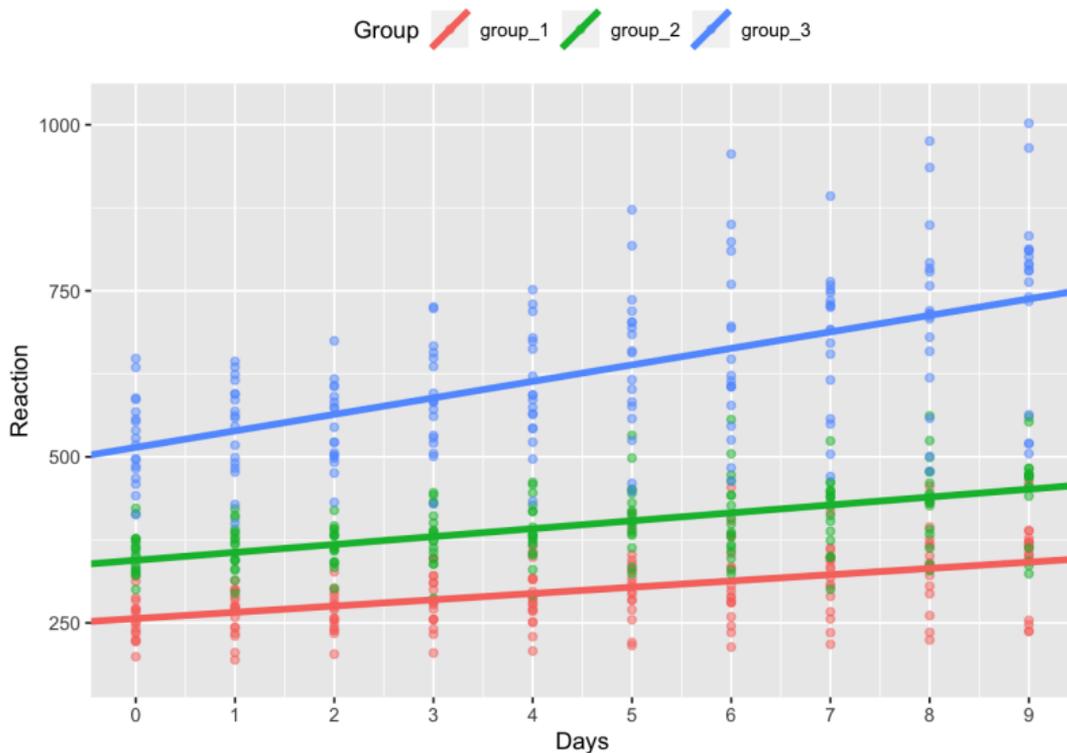
# Illustrative example: linear regression

Disregarding group structure in data may yield overly simple (though robust) models



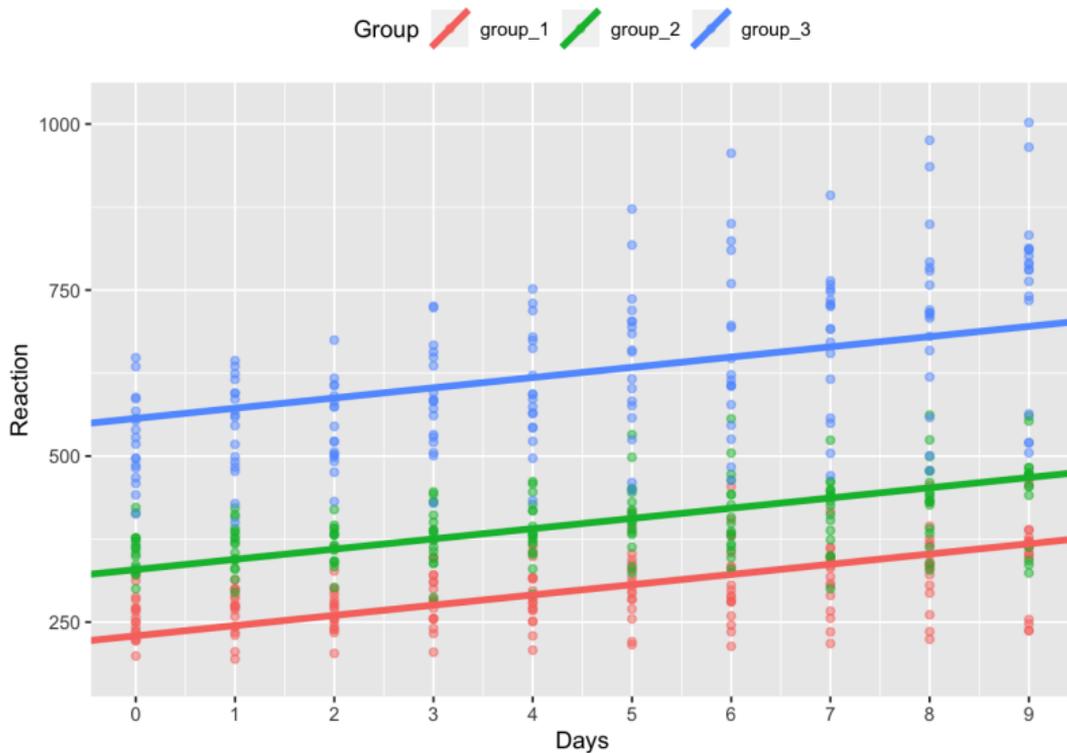
## Illustrative example: modeling independent groups

Conversely, considering each group independently provides tailored models, but ignores possibly *shared* information, and reduces statistical power (more parameters to estimate with less data).



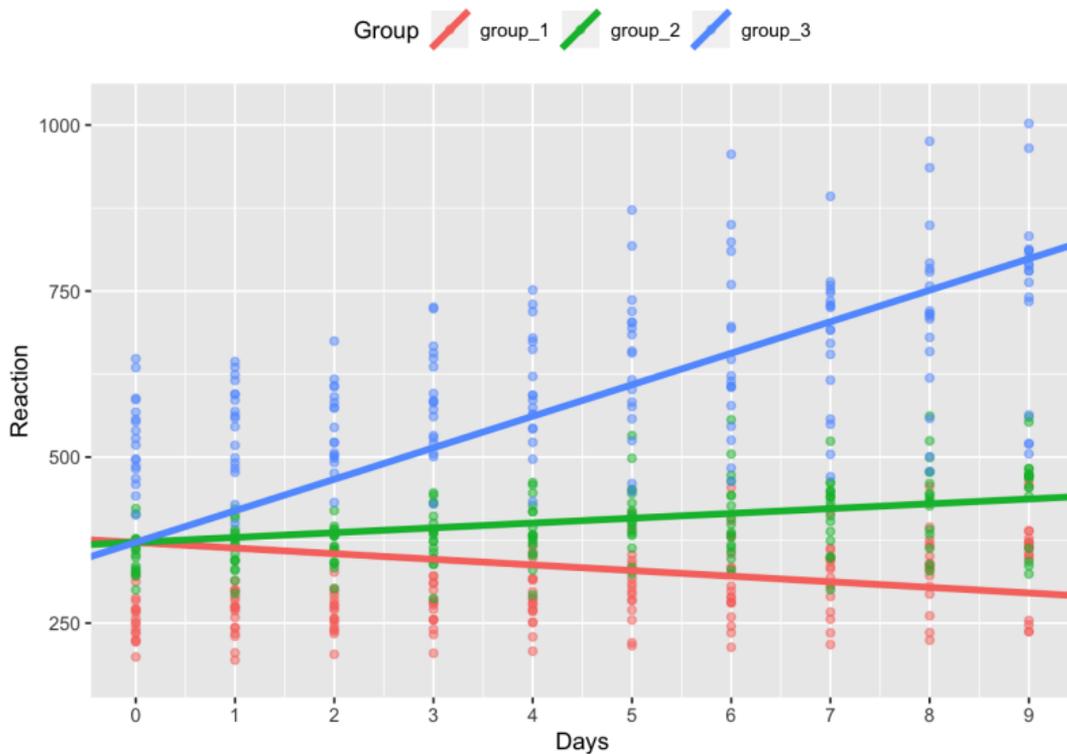
# Illustrative example: pooling the slope

If all groups share the same slope, we assume that the underlying mechanism is identical.



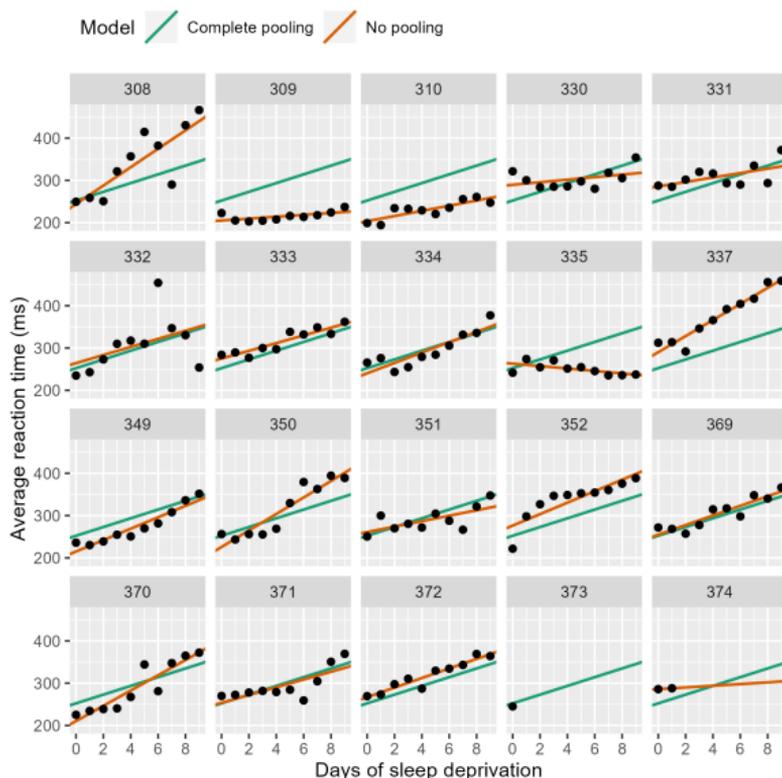
# Illustrative example: pooling the intercept

If all groups share the same intercept, it is the baseline value that is assumed common.



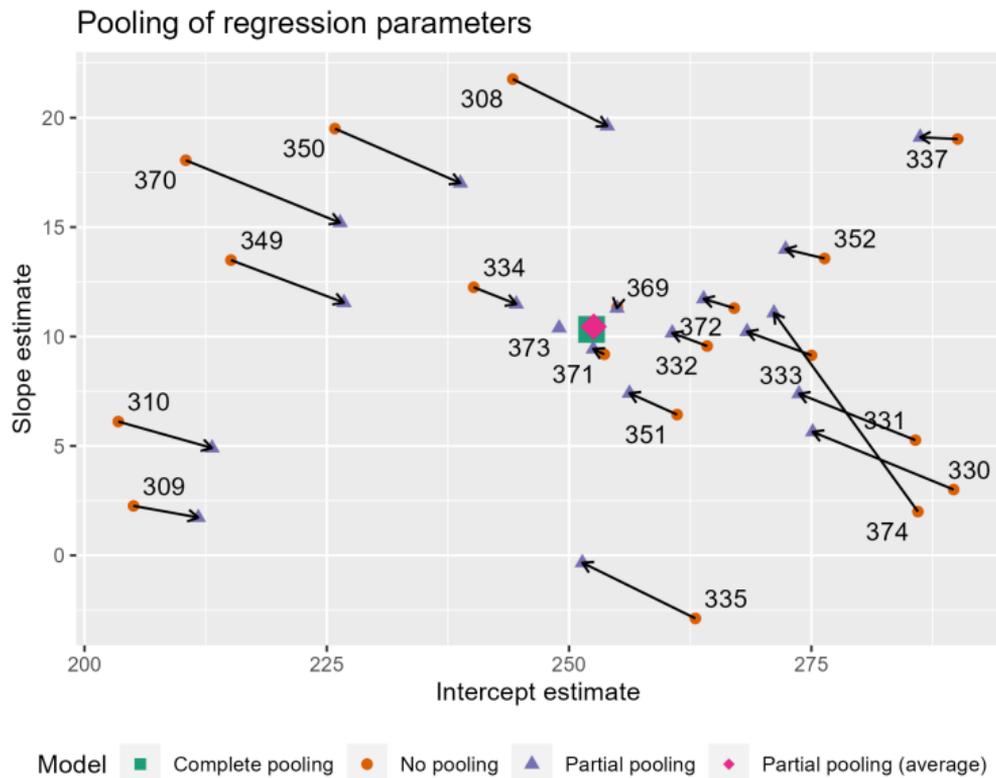
# Illustrative example: layers of hierarchy

Hierarchical models can model several layers (e.g. several individuals within groups).



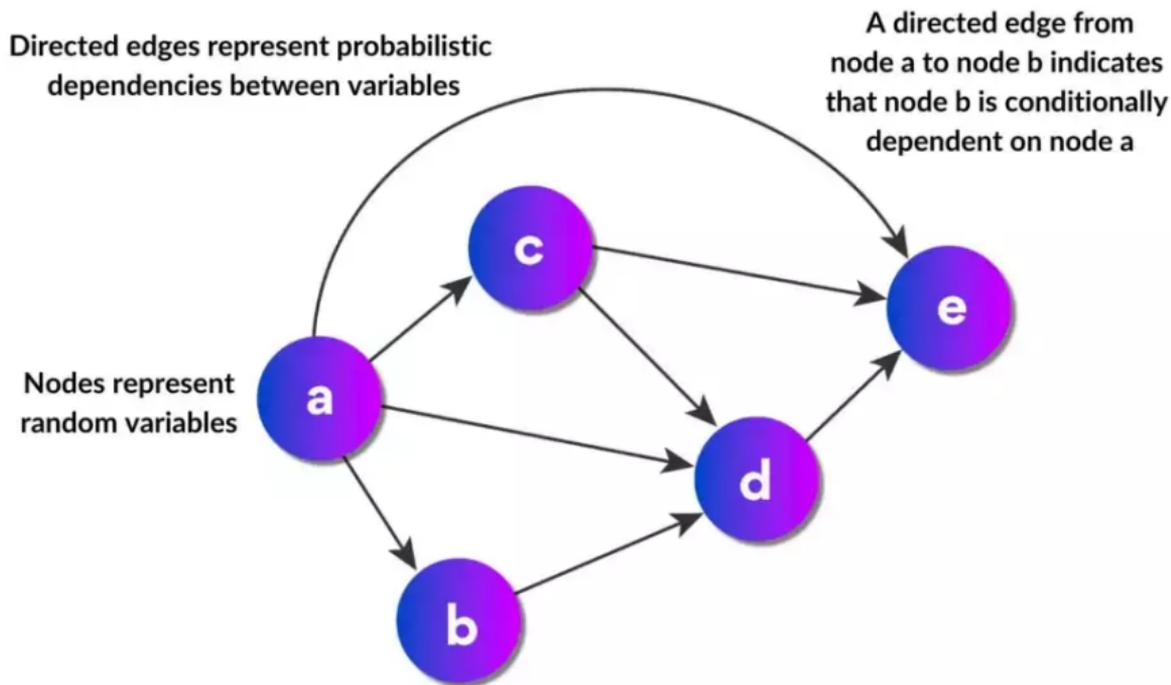
# Illustrative example: shrinkage effect

A typical phenomenon, called *shrinkage*, tends to "pull" estimated parameters towards the global ones as you share information.



# Directed Acyclic Graphs (DAGs)

Often, Bayesian hierarchical models use DAGs to visually represent the dependencies structure between variables.



## Exercise: recovering dependencies and expressing probabilities

Road Condition

Driver's fatigue

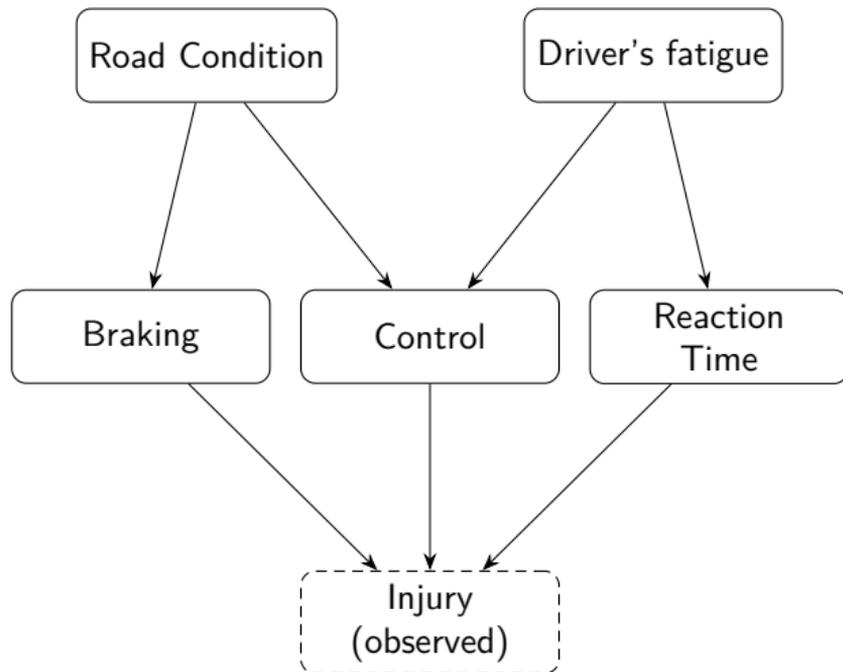
Braking

Control

Reaction  
Time

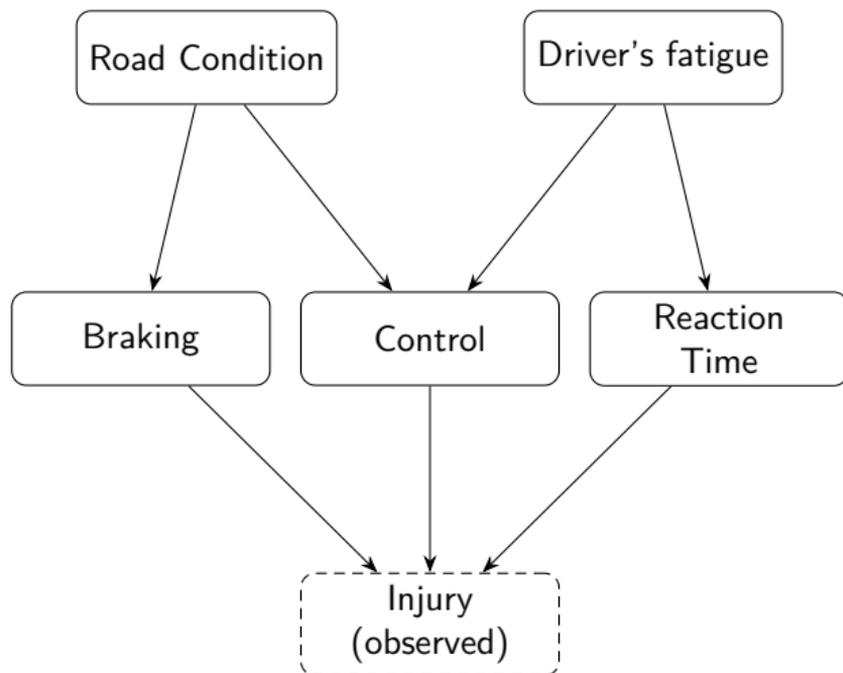
Injury  
(observed)

## Exercise: recovering dependencies and expressing probabilities



Corresponding expression of the joint probability distribution:

## Exercise: recovering dependencies and expressing probabilities



Corresponding expression of the joint probability distribution:

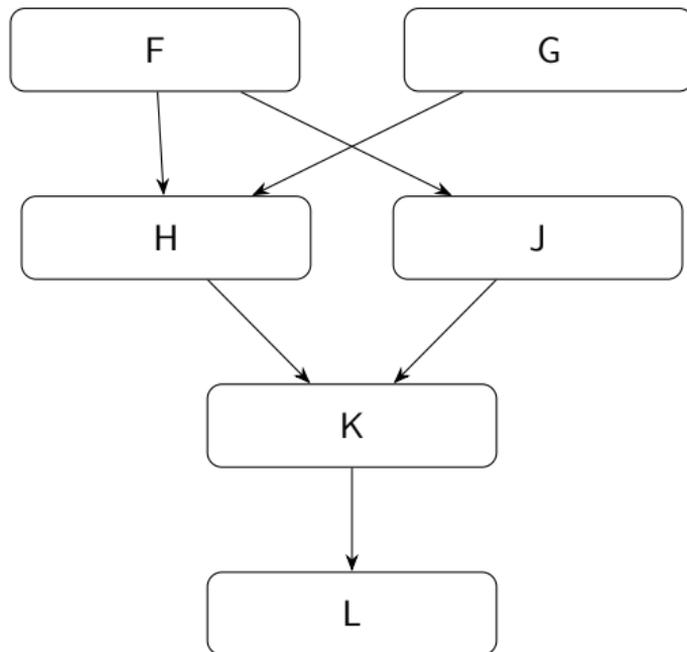
$$p(I, B, C, Re, Ro, D) = p(Ro)p(D)p(B | Ro)p(C | Ro, D)p(Re | D)p(I | B, C, Re)$$

## Exercise: Draw the generative graph from the joint distribution

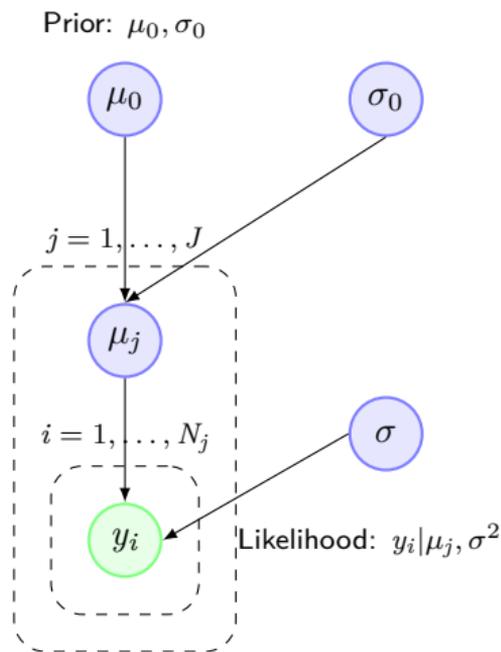
$$p(F, G, H, J, K, L) = p(F) p(G) p(H | F, G) p(J | F) p(K | H, J) p(L | K)$$

## Exercise: Draw the generative graph from the joint distribution

$$p(F, G, H, J, K, L) = p(F) p(G) p(H | F, G) p(J | F) p(K | H, J) p(L | K)$$



# Bayesian Hierarchical Model: Gaussian Example



## Model equations:

$$y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$$

$$\mu_j \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

$\mu_0$ : hyperparameter

$\sigma_0$ : hyperparameter

$\sigma$ : likelihood variance

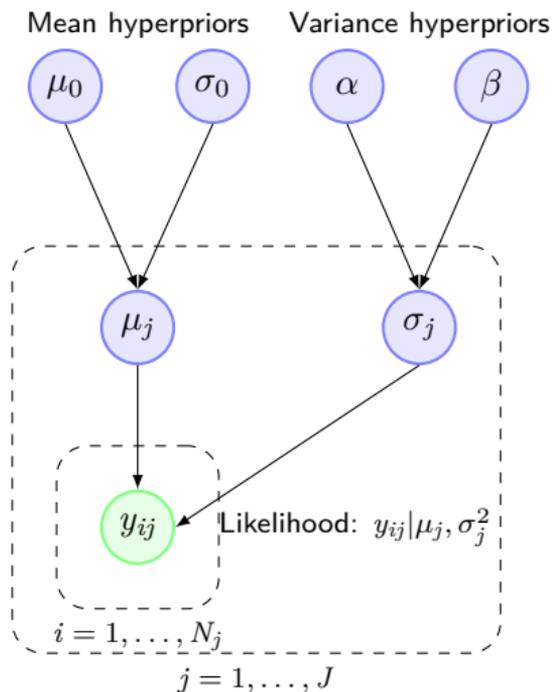
## Key points:

- Observations  $y_{ij}$  are grouped by  $j$ .
- Group means  $\mu_j$  share information via the population-level mean  $\mu_0$ .
- Shrinkage: estimates for small groups are pulled toward  $\mu_0$ .

Hierarchical representation clarifies dependencies and factorization:

$$p(\{y_{ij}\}, \{\mu_j\} | \mu_0, \sigma_0^2, \sigma^2) = \prod_{j=1}^J p(\mu_j | \mu_0, \sigma_0^2) \prod_{i=1}^{N_j} p(y_{ij} | \mu_j, \sigma^2)$$

# Bayesian Hierarchical Model: modularity of the models



## Model equations:

$$y_{ij} \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

$$\mu_j \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

$$\sigma_j^2 \sim \Gamma^{-1}(\alpha, \beta)$$

$\mu_0, \sigma_0, \alpha, \beta$ : hyperparameters

## Key points:

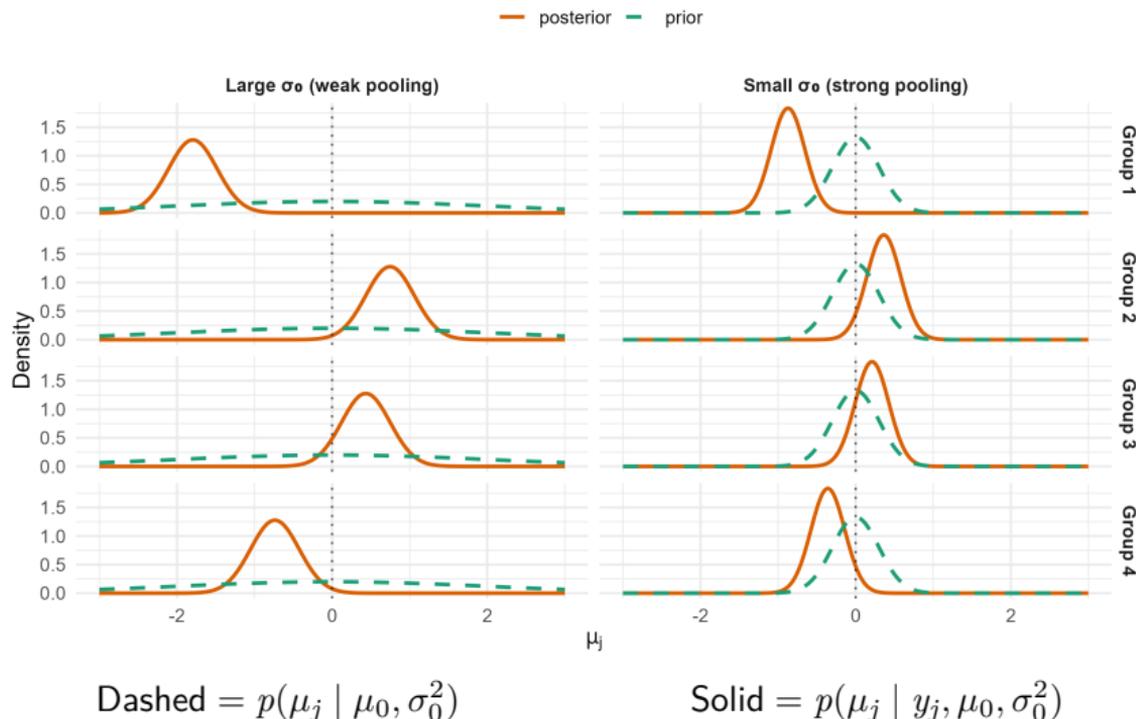
- Both means  $\mu_j$  and variances  $\sigma_j^2$  are group-specific.
- Hierarchical models are highly modular. Each sub-part can be adapted or extended with no change on the others.

Factorization of joint distribution naturally adapts:

$$p(\{y_{ij}\}, \{\mu_j\}, \{\sigma_j^2\} | \mu_0, \sigma_0, \alpha, \beta) = \prod_{j=1}^J p(\mu_j | \mu_0, \sigma_0) p(\sigma_j^2 | \alpha, \beta) \prod_{i=1}^{N_j} p(y_{ij} | \mu_j, \sigma_j^2)$$

# Hyperparameters, hyperpriors, and Empirical Bayes

- Hyperprior choices encode prior belief about population-level variation.
- Empirical Bayes: estimate hyperparameters from data (e.g., marginal MLE) — pragmatic but less fully Bayesian.



## Limits of the conjugate framework

Conjugate priors/likelihoods are specific cases of 2-layer hierarchical models, where the posterior is analytically tractable.

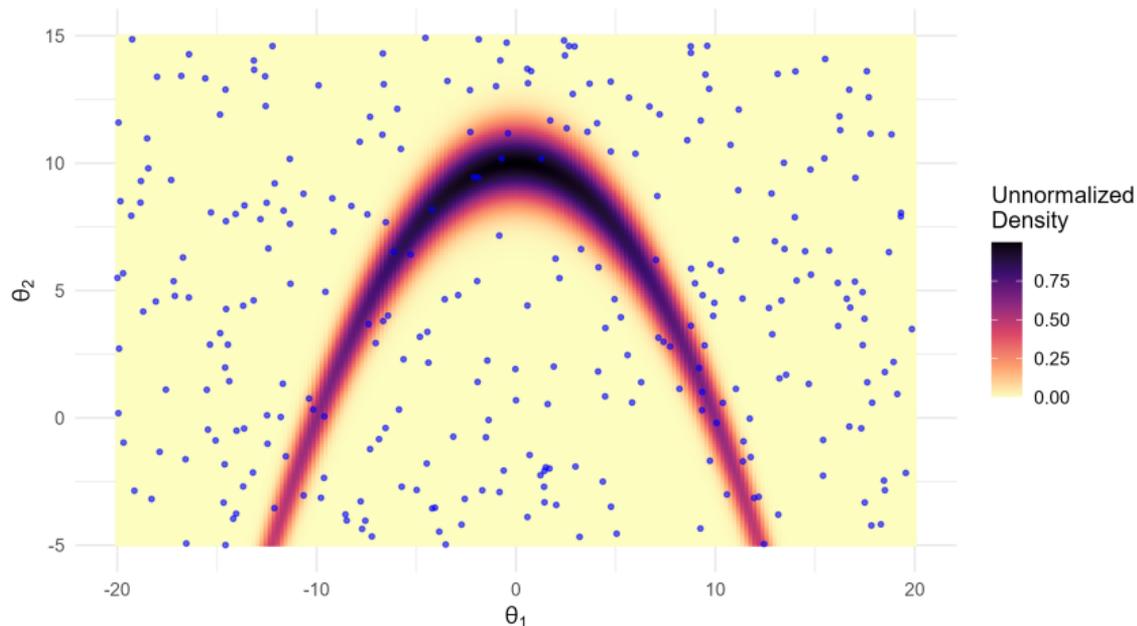
Likelihood	Prior	Posterior
Binomial	Beta	Beta
Negative Binomial	Beta	Beta
Poisson	Gamma	Gamma
Geometric	Beta	Beta
Exponential	Gamma	Gamma
Normal (mean unknown)	Normal	Normal
Normal (variance unknown)	Inverse Gamma	Inverse Gamma
Normal (mean and variance unknown)	Normal / Gamma	Normal / Gamma
Multinomial	Dirichlet	Dirichlet

- Conjugacy exists only for specific likelihood-prior pairs.
- Many models of practical interest (logistic, complex hierarchies, ...) are non-conjugate.

We therefore need **approximate** or **sampling-based** methods to compute general posteriors.

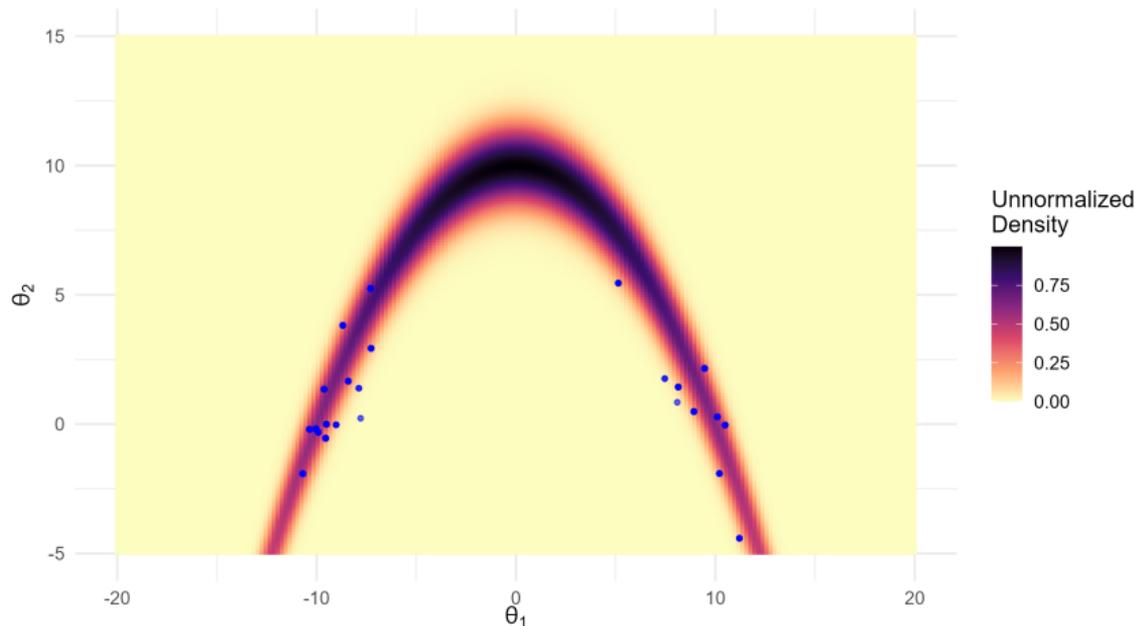
# Unnormalized posteriors and sampling intuition

- Posterior often known up to normalization:  $p(\theta|y) \propto p(y|\theta)p(\theta)$ .
- Goal: compute expectations  $\mathbb{E}[f(\theta)] = \int f(\theta)p(\theta|y)d\theta$ .
- Monte Carlo: approximate integral with samples  $\{\theta_i\}_{i=1}^N$ .



# Unnormalized posteriors and sampling intuition

- Posterior often known up to normalization:  $p(\theta|y) \propto p(y|\theta)p(\theta)$ .
- Goal: compute expectations  $\mathbb{E}[f(\theta)] = \int f(\theta)p(\theta|y)d\theta$ .
- Monte Carlo: approximate integral with samples  $\{\theta_i\}_{i=1}^N$ .



# Intuition behind Markov Chains

- A **Markov chain** is a sequence  $\{\theta^{(t)}\}$  depending **only on the previous one**:

$$p(\theta^{(t+1)}|\theta^{(1)}, \dots, \theta^{(t)}) = p(\theta^{(t+1)}|\theta^{(t)}).$$

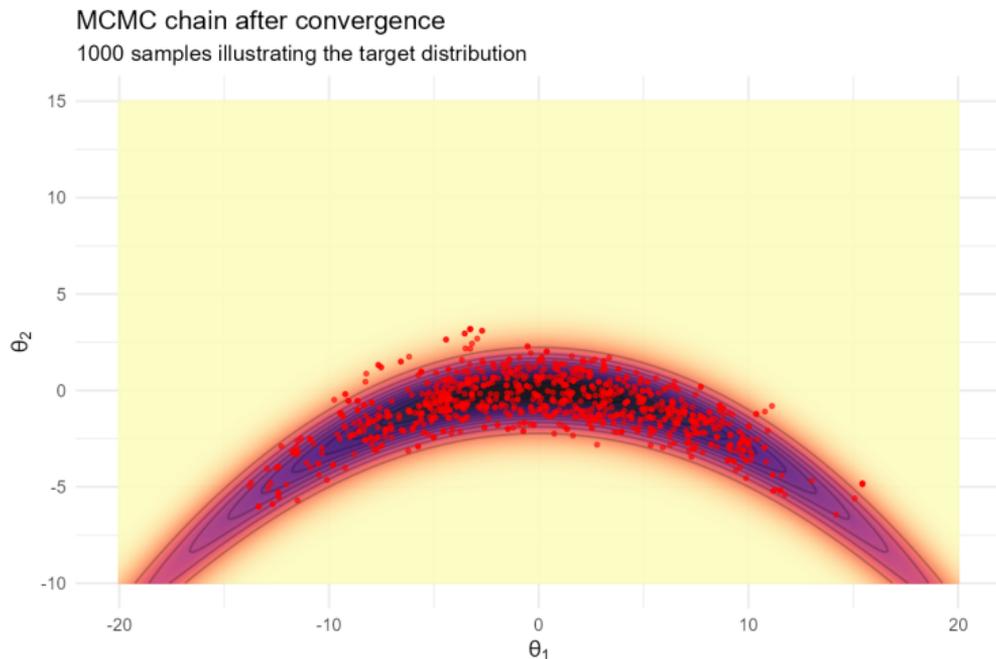
- Designed so that, as  $t \rightarrow \infty$ , the chain's **stationary distribution** is the target.
- Exploring **the parameter space**, spending more time where  $p(\theta|y)$  is large.

# Intuition behind Markov Chains

- A **Markov chain** is a sequence  $\{\theta^{(t)}\}$  depending **only on the previous one**:

$$p(\theta^{(t+1)}|\theta^{(1)}, \dots, \theta^{(t)}) = p(\theta^{(t+1)}|\theta^{(t)}).$$

- Designed so that, as  $t \rightarrow \infty$ , the chain's **stationary distribution** is the target.
- Exploring **the parameter space**, spending more time where  $p(\theta|y)$  is large.



# Rejection Sampling

■ **Goal:** sample from a target density  $p(\theta)$  when direct sampling is difficult.

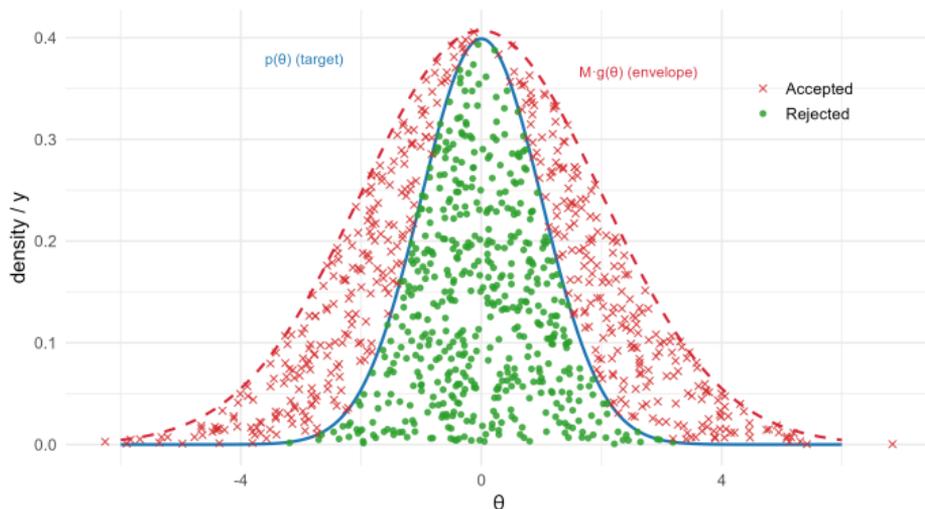
■ **Idea:**

- 1 Draw  $\theta \sim g(\theta)$  (a proposal distribution covering  $p$ ).
- 2 Accept  $\theta$  with probability

$$\alpha(\theta) = \frac{p(\theta)}{M \times g(\theta)},$$

where  $M$  is a constant such that  $p(\theta) \leq M \times g(\theta)$  for all  $\theta$ .

- **Intuition:** Points under the scaled proposal envelope are “kept”; others are rejected.
- **Efficiency:** Low-dimensional problems only; acceptance rate  $\approx 1/M$ .



# Gibbs Sampling: Sampling Along Conditional Distributions

- **Goal:** Sample from a joint distribution  $p(x, y)$ .
- **Idea:** Update **one variable** using its *conditional distribution* given current values of others:

$$x^{(t)} \sim p(x | y^{(t-1)}), \quad y^{(t)} \sim p(y | x^{(t)})$$

- **Intuition:** The chain moves in a *zig-zag* path along the axes of the variables.
- Each update is *always accepted* (no rejection like importance, Metropolis-Hastings).
- Works well when conditional distributions are easy to sample.

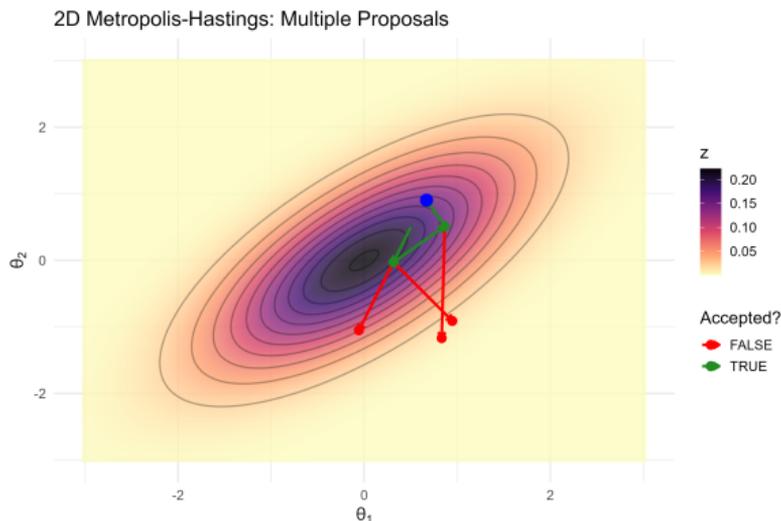
# Metropolis-Hastings: Proposal and Acceptance Mechanism

## ■ Mechanism:

- 1 Start at current point  $\theta^{\text{curr}}$  (blue dot).
- 2 Propose a new point  $\theta^{\text{prop}} \sim q(\theta^{\text{curr}}, \cdot)$  (random nearby).
- 3 Compute acceptance probability:

$$\alpha = \min \left( 1, \frac{p(\theta^{\text{prop}})}{p(\theta^{\text{curr}})} \right)$$

- 4 Accept the proposal with probability  $\alpha$  (green = accepted, red = rejected).
- 5 If rejected, stay at current point.



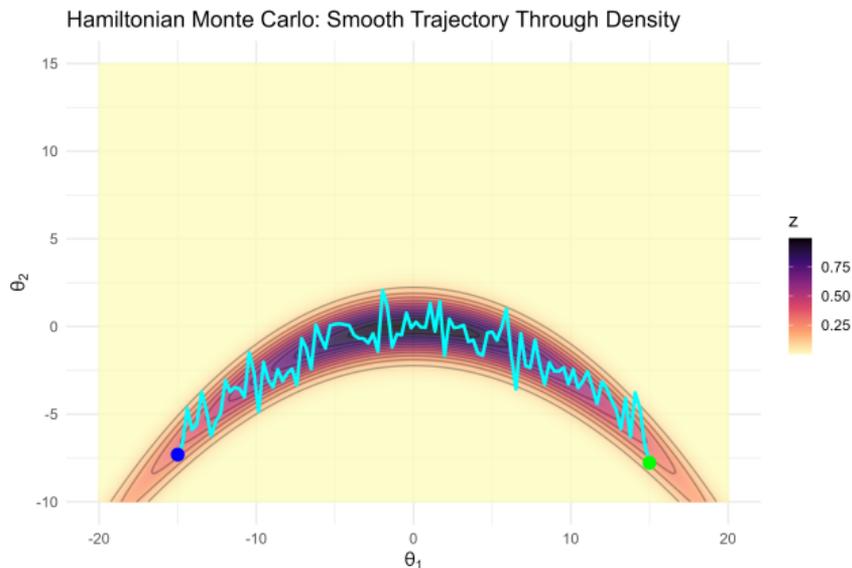
# Hamiltonian Monte Carlo: Efficient Exploration via Trajectories

- **Key Idea:** Introduce a *momentum variable*  $p$  and simulate Hamiltonian dynamics:

$$H(\theta, p) = -\log p(\theta) + \frac{1}{2}p^T M^{-1}p$$

to move along contours of high probability.

- The particle “slides” through the probability landscape along smooth trajectories.
- Momentum prevents random-walk behavior and allows long-distance moves.
- Each trajectory is approximately energy-conserving, leading to high acceptance rates.



# MCMC Methods: Comparison and Use Cases

Method	Pros / Benefits	Limits / When to Use
Metropolis-Hastings	<ul style="list-style-type: none"><li>- Simple, general-purpose</li><li>- Can target any distribution</li><li>- Easy to implement in 1–2 dimensions</li></ul>	<ul style="list-style-type: none"><li>- Random walk <math>\rightarrow</math> slow mixing in high dimensions</li><li>- Acceptance rate depends on proposal tuning</li><li>- Inefficient in complex correlated distributions</li></ul>
Gibbs Sampling	<ul style="list-style-type: none"><li>- Always accepted (no rejection)</li><li>- Efficient if conditional distributions are easy to sample</li><li>- Good for structured models</li></ul>	<ul style="list-style-type: none"><li>- Requires analytically tractable conditionals</li><li>- Can be slow if variables are highly correlated</li><li>- Not general-purpose for arbitrary distributions</li></ul>
Hamiltonian MC	<ul style="list-style-type: none"><li>- Efficient exploration of high-dimensional, correlated spaces</li><li>- Smooth, long-distance moves <math>\rightarrow</math> fast convergence</li><li>- High acceptance rate</li></ul>	<ul style="list-style-type: none"><li>- Requires gradients of log-density</li><li>- Implementation more complex</li><li>- Step-size and trajectory length need tuning</li><li>- State-of-the-art for continuous, differentiable targets</li></ul>

# MCMC Mechanics: Ensuring Reliable Sampling

## ■ Burn-in:

- Initial iterations may be far from the target distribution.
- Discard a fraction of early samples to allow the chain to “warm up”.

## ■ Convergence Assessment:

- Run multiple chains from different starting points.
- Check if chains converge to the same region/distribution.
- Visual tools: trace plots, density overlays.

## ■ Mixing:

- Chains should explore the target efficiently, not get “stuck”.
- Poor mixing → high autocorrelation between successive samples.

## ■ Sanity Checks / Diagnostics:

- *Trace plots*: visualize parameter values over iterations.
- *Autocorrelation plots*: quantify dependence between samples.
- *Gelman-Rubin  $\hat{R}$* : compare variance between and within chains.
- *Effective Sample Size (ESS)*: number of independent samples.

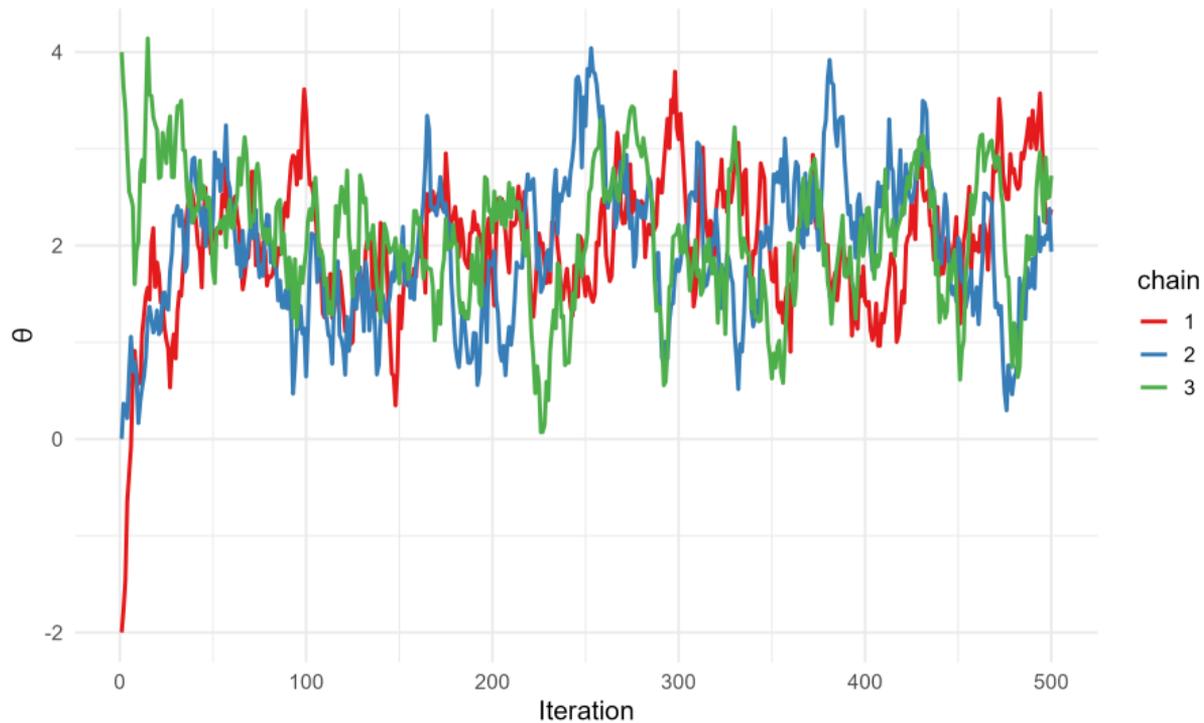
## ■ Key Takeaway:

- MCMC samples are only useful if the chain has converged and is well-mixed.
- Diagnostics prevent misleading inferences from biased or correlated samples.

# MCMC traceplots

## Multiple Chains Converging

Chains start far apart and converge to stationary region



# Hierarchical Model Syntax with nimble

```
library(nimble)

# --- 1) Define the model ---
code <- nimbleCode({
  for (i in 1:N) {
    y[i] ~ dnorm(mu[i], sd = sigma)
    mu[i] <- alpha + beta * x[i]
  }
  alpha ~ dnorm(0, sd = 10)
  beta ~ dnorm(0, sd = 10)
  sigma ~ dunif(0, 10)
})

# --- 2) Provide data ---
data <- list(y = c(2.3, 1.9, 3.1, 2.7))
constants <- list(N = length(data$y), x = c(0, 1, 2, 3))

# --- 3) Build model and compile ---
model <- nimbleModel(code, data = data, constants = constants)
cmodel <- compileNimble(model)

# --- 4) Configure MCMC ---
conf <- configureMCMC(model)
Rmcmc <- buildMCMC(conf)
cmcmc <- compileNimble(Rmcmc, project = model)

# --- 5) Run MCMC ---
samples <- runMCMC(cmcmc, niter = 1000)
head(samples)
```

# Incomplete data models and latent variables

Most often, the observations are involved in complicated (biological, ecological, physical) processes, with many unobserved variables and complex dependency structure.

- $X$  observed random variables
- $Z$  unobserved (latent/hidden) variables
- $\theta$  unknown parameters

## An attempt at defining latent variables

- Frequentist setting:

latent variables = random but unobserved, parameters = fixed

- Bayesian setting:

both latent variables and parameters = random

but

$\#$  latent variable  $\simeq$   $\#$  data,  $\#$  parameters  $\ll$   $\#$  data

# Different types of likelihoods

In this course, we mainly adopt a *empirical Bayes* (also called *type II*) viewpoint, using MLE inference for latent variables, although Bayesian counterparts of proposed models also exist.

## Complete data likelihood

Joint likelihood of the whole random process  $(\mathbf{X}, \mathbf{Z})$  with given parameters  $\theta$ .

$$p_{\theta}(\mathbf{X}, \mathbf{Z}) = p_{\theta}(\mathbf{X} | \mathbf{Z})p_{\theta}(\mathbf{Z}).$$

↪ tractable in many models, but we do not observe  $\mathbf{Z}$  !

## Observed data likelihood

Marginal likelihood of the observed random variables  $\mathbf{X}$

$$p_{\theta}(\mathbf{X}) = \int_{\mathcal{Z}} p_{\theta}(\mathbf{X}, \mathbf{z}) d\mathbf{z}^a$$

↪ only involves the observed  $\mathbf{X}$ , but not always tractable.

---

<sup>a</sup>When  $\mathcal{Z}$  is discrete, replace  $\int$  by  $\sum$

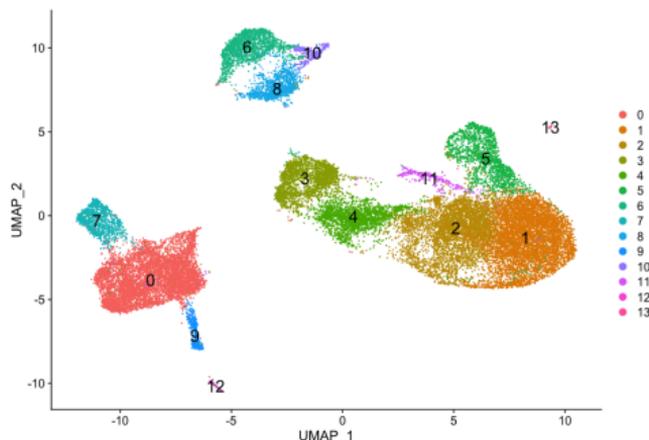
# Clustering and mixture models

# Motivation

Sometimes our data is organized in sub-population: groups of individuals we call *clusters*.

## Example

In modern biology, discovering cell-types via their gene expression profile is an important task.



When the groups are unknown, we call the task of discovering them *clustering*<sup>2</sup>

---

<sup>2</sup>as opposed to classification in a supervised context

# Mathematical context

We search for an optimal partition of  $x = \{x_1, \dots, x_n\}$  into  $K$  groups.

## Definition: partition

A partition  $C = \{C_1, \dots, C_K\}$  of  $\{1, \dots, n\}$  is a set of sets s.t.

$$\bigcup_k C_k = \{1, \dots, n\}, \quad \forall k \neq l, \quad C_k \cap C_l = \emptyset.$$

## Alternative encoding of the partition

For each individual  $i = 1, \dots, n$ , we define its *cluster membership*  $z_i \in \{0, 1\}^K$

$$k = 1, \dots, K, \quad z_{ik} = \begin{cases} 1 & \text{if } i \text{ belongs to cluster } k, \\ 0 & \text{otherwise} \end{cases}.$$

The set  $Z = \{z_1, \dots, z_n\}$  represents a partition of  $\{1, \dots, n\}$ . This particular encoding is sometimes referred to as one-hot encoding.

# Clustering criteria

"Optimality" implies the definition of some criterion  $L \iff$  assumptions on the nature of clusters. Methods can be roughly split in two

## Similarity-based methods

Design  $L$  via geometric notions of similarity between  $x_i$ 's, favoring e.g.

- elliptic clusters
- convex clusters
- connected clusters

## Statistical methods

Consider the partition  $Z$  as a latent variable and posit a generative model  $p_\theta(\mathbf{X}, \mathbf{Z})$

$\rightsquigarrow$  Clustering becomes an inference problem of finding  $\hat{Z}$ .

There are connections between both !

## K-means

# The K-means problem

K-means seeks clusters well concentrated around their centroids  $\mu_k := \frac{1}{|C_k|} \sum_{i \in C_k} x_i$  by minimizing

$$\arg \min_{\mathcal{C}} \left\{ L(\mathcal{C}, \mathbf{X}) = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \mu_k\|_2^2 \right\} \quad (\text{K-means problem})$$

- Good news: discrete problem  $\rightsquigarrow$  there exists an optimum  $\mathcal{C}^*$ .
- **Bad news:** there are  $K^n$  possible partitions  $\rightsquigarrow$  enumeration is not an option.

In fact, K-means problem is a **nonconvex NP-hard** problem and one need to resort to fast heuristics.

⚠ With a slight abuse, we drop distinction between K-means problem and heuristics to solve it.

# The K-means algorithm (MacQueen 1967)

Draw centroids  $\mu_1, \dots, \mu_K$  at random among the sample  $X$  and

- 1 Assign each point to its closest centroid

$$C_k \leftarrow \left\{ i : \|x_i - \mu_k\|_2^2 = \min_l \|x_i - \mu_l\|_2^2 \right\}$$

- 2 recompute centroids as the barycenter of each center

$$\mu_k := \frac{1}{|C_k|} \sum_{i \in C_k} y_i$$

- 3 Go to 1 until clusters (hence barycenters) are unchanged

## Properties of the algorithm

K-means is a greedy algorithm which

- monotonically decreases the criterion
- converges in a finite number of iterations
- will get stuck in local minima of  $L$  (non-convex)

↪ In practice, we try several restarts with different random inits.

# Extensions

**Kmeans++** initialization matter !  $\rightsquigarrow$  stop drawing centroids at random

- Choose  $\mu_1$  uniformly among the sample
- then sequentially do for each  $k = 2, \dots, K$ 
  - compute weight  $w_i := \min_{j < k} \|x_i - \mu_j\|_2^2$
  - Choose  $\mu_k$  among the sample with proba  $\propto w_i$

Optimality bounds can be obtained (Arthur et al. 2007)

**Sparse K-means** include variable selection, useful when  $x_i$  in dimension  $d \gg n$

**Kernel K-means** compute distance between  $\phi(x_i)$  with  $\phi : \mathcal{X} \rightarrow \mathcal{H}$  a *feature map*.

## Mixture models

# Probabilistic view on clustering

The partition is now seen as a set of discrete latent variables  $\mathbf{Z} = \{z_1, \dots, z_n\}$

Denote  $\pi = (\pi_1, \dots, \pi_K)$  the (unknown) cluster proportions, we have

$$p_\pi(z_{ik} = 1) = \pi_k \iff z_i \sim \mathcal{M}(1, \pi)$$

## Mixture models

For all  $i = 1, \dots, n$ , mixture models suppose that  $(z_i, x_i)$  are drawn *i.i.d.* according to the two-stage hierarchical model

- 1  $Z_i \sim \mathcal{M}_K(1, \pi)$
- 2  $X_i | \{z_{ik} = 1\} \sim p_{\gamma_k}$

The model parameters are  $\theta = \{\pi_k, \gamma_k\}_{k=1}^K$  and  $p_\gamma$  can be any parametric distribution over  $X_i$ .

Clusters are sometimes called *components*

$\rightsquigarrow$  **general** and **flexible** framework, adapt to nature of the data (discrete, continuous, mixed-type) via  $p_\gamma$

# Observed (marginal) likelihood

## Properties: independence

In a mixture model,  $(Z_i)_i$  are *i.i.d.* and  $(X_i)_i$  also are *i.i.d.*

### Observed likelihood

$$\begin{aligned} p_{\theta}(\mathbf{X}) &= \sum_{z_1, \dots, z_n} p_{\theta}(\mathbf{Z}, \mathbf{X}) = \sum_{z_1, \dots, z_n} \prod_{i=1}^n p_{\theta}(X_i | z_i) p_{\theta}(z_i), \\ &= \prod_{i=1}^n \sum_{z_i} p_{\gamma}(X_i | z_i) p_{\theta}(z_i), \\ &= \prod_{i=1}^n \left( \sum_{k=1}^K \pi_k p_{\gamma_k}(X_i) \right). \end{aligned}$$

$\rightsquigarrow$  the marginal distribution of  $X_i$  is a convex combination (*mixture*) of the  $K$  base distributions  $(p_{\gamma_k})_k$ , with weights  $\pi_k$ .

# Complete likelihood

## Properties: conditional independence

In a mixture model,  $(X_i)_i \perp\!\!\!\perp \mathbf{Z}$  and  $(Z_i)_i \perp\!\!\!\perp \mathbf{X}$ , **but not** identically distributed

## Complete log-likelihood

$$\begin{aligned}\log p_{\theta}(\mathbf{X}, \mathbf{Z}) &= \log p_{\theta}(\mathbf{Z}) + \log p_{\theta}(\mathbf{X} \mid \mathbf{Z}) = \sum_{i=1}^n \log p_{\pi}(Z_i) + \log p_{\gamma}(X_i \mid Z_i), \\ &= \sum_{k=1}^K \sum_{i=1}^n Z_{ik} [\log \pi_k + \log p_{\gamma_k}(X_i)].\end{aligned}$$

## Posterior distribution of $\mathbf{Z} \mid \mathbf{X}$

For  $i = 1, \dots, n$ ,  $Z_i \mid X_i \sim \mathcal{M}_K(1, \tau_i)$  with

$$\tau_{ik} := p_{\theta}(z_{ik} = 1 \mid X_i) \propto \pi_k p_{\gamma_k}(X_i)$$

Notice that  $\tau_i$  also depends on the parameters  $\theta$ .

# A note on identifiability

## Definition: identifiability

A statistical model  $p_\theta$  is said to be identifiable iff the mapping  $\theta \mapsto p_\theta$  is injective.

**Intuition:** the labels of the clusters  $1, \dots, K$  should have no impact on the marginal likelihood

$$\pi_1 p_{\gamma_1}(x) + \pi_2 p_{\gamma_2}(x) = \pi_2 p_{\gamma_2}(x) + \pi_1 p_{\gamma_1}(x)$$

## Label switching

Let  $\sigma$  be a permutation of  $\llbracket 1, K \rrbracket$ , then for a mixture model with parameters  $\pi, \gamma$  we have

$$p(\mathbf{X} \mid \pi, \gamma) = p(\mathbf{X} \mid \sigma(\pi), \sigma(\gamma))$$

Hence, there are  $K!$  equivalent formulations of a mixture model.

↪ conceptually not a problem, it simply states that there are  $K!$  different encoding  $\mathbf{Z}$  of a given partition  $C = \{C_1, \dots, C_K\}$ .

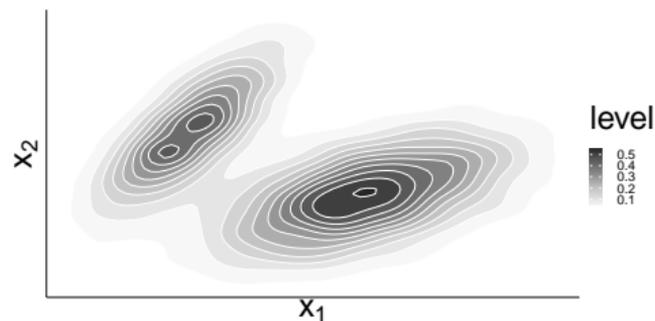
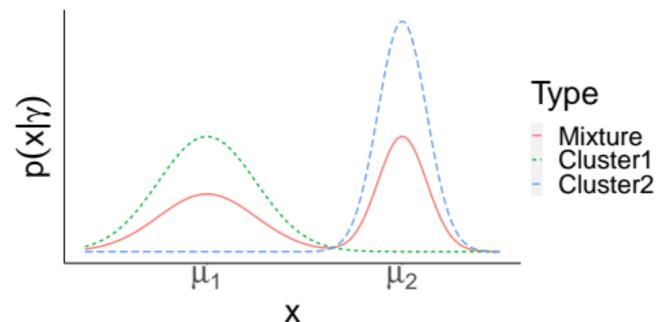
↪ can cause problems in Bayesian inference procedure since the posterior is highly multimodal.

# Gaussian Mixture Models (GMM)

**Continuous data:**  $\mathbf{x} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$

**Model:** Mixture of Gaussians  $p_{\gamma_k}(x) = \mathcal{N}(x \mid \mu_k, \Sigma_k)$ , with  $\gamma_k = (\mu_k, \Sigma_k)$

Multimodal marginal density around the  $(\mu_k)_k$ 's



**Number of free parameters:**  $K - 1 + Kd + K \frac{d(d+1)}{2} = \mathcal{O}(Kd^2)$  to estimate

# Maximum-likelihood estimation

Non-convex MLE problem

$$\arg \max_{\pi_k, \mu_k, \Sigma_k} \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \right).$$

- Much more complex to maximize than in standard Gaussian models ( $K = 1$ )
- No closed-form solution, gradients can be derived but
  - 1 they are not cheap to compute at each iteration (although one could resort to stochastic optimization to leverage this issue).
  - 2 Requires re-projecting on the cone of p.d. matrices  $\Sigma_k \succ 0$ .

By contrast, the complete log-likelihood is much simpler to handle

$$\log p_{\theta}(\mathbf{x}, \mathbf{Z}) = \sum_{k=1}^K \sum_{i=1}^n Z_{ik} [\log \pi_k + \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)].$$

↪ **But we do not observe the  $Z$  !**

# Maximum-likelihood estimation

## A chicken-and-egg problem

- 1 If we knew  $\mathbf{Z}$  we could maximize  $p_{\theta}(\mathbf{X}, \mathbf{Z}) \rightsquigarrow$  amount to compute MLE  $\hat{\gamma}_k$  in each cluster. In the Gaussian case we'd have cluster's empirical means and covariance

$$n_k = \sum_i z_{ik}, \quad \hat{\mu}_k = \sum_i z_{ik} x_i / n_k, \quad \hat{\Sigma}_k = \sum_i z_{ik} \frac{(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^{\top}}{n_k}$$

- 2 If we knew  $\theta^*$ , we could find the best estimate of  $\mathbf{Z}$  via the posterior distribution

$$\tau_{ik}(\theta) = p_{\theta}(z_{ik} = 1 \mid x_i) = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_l \pi_l \mathcal{N}(x_i \mid \mu_l, \Sigma_l)}$$

$\rightsquigarrow$  this suggest an iterative scheme between 1) & 2) to solve MLE.

# Inference in latent variable models: the EM algorithm

## Some tools from information theory

# Jensen's inequality

**Quizz !** Which is larger:  $\mathbb{E}[Z^2]$  or  $\mathbb{E}[Z]^2$  ?

## Jensen's inequality

**Quiz !** Which is larger:  $\mathbb{E}[Z^2]$  or  $\mathbb{E}[Z]^2$  ?

**Answer:**  $\mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \mathbb{V}(Z) \geq 0$

# Jensen's inequality

**Quiz !** Which is larger:  $\mathbb{E}[Z^2]$  or  $\mathbb{E}[Z]^2$  ?

**Answer:**  $\mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \mathbb{V}(Z) \geq 0$

## General result: Jensen's inequality

Let  $Z$  be a random vector in  $\mathcal{Z} \subset \mathbb{R}^d$  and  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$  a convex function, then

$$\mathbb{E}_Z [\phi(Z)] \geq \phi(\mathbb{E}_Z[Z]). \quad (\text{Jensen})$$

$\rightsquigarrow$  the inequality is reversed with  $\phi$  concave ( $\phi \leftarrow -\phi$ )

*Proof:*

- $\phi$  convex  $\implies$  it is above its tangents, hence at any point  $z_0 \in \mathbb{R}^d, \exists a$  s.t.

$$\forall z \in \mathbb{R}^d, \quad \phi(z) \geq \phi(z_0) + a(z - z_0).$$

- Take  $z_0 = \mathbb{E}_Z[Z]$ , since the above inequality is true for all  $z$ , it generalizes to  $\mathbb{E}_Z$

$$\mathbb{E}_Z [\phi(Z)] \geq \phi(z_0) + \underbrace{a(\mathbb{E}_Z[Z] - z_0)}_{=0} = \phi(z_0) = \phi(\mathbb{E}_Z[Z])$$

# Entropy of a random variable

## Definition: entropy

For a discrete random variable  $Z$  with distribution  $q(Z = z)$  we define its entropy as

$$\mathcal{H}(Z) = \mathcal{H}(q) = -\mathbb{E} [\log q(Z)] = - \sum_{z \in \mathcal{Z}} q(z) \log q(z)$$

with the convention that  $0 \times \log 0 = 0$

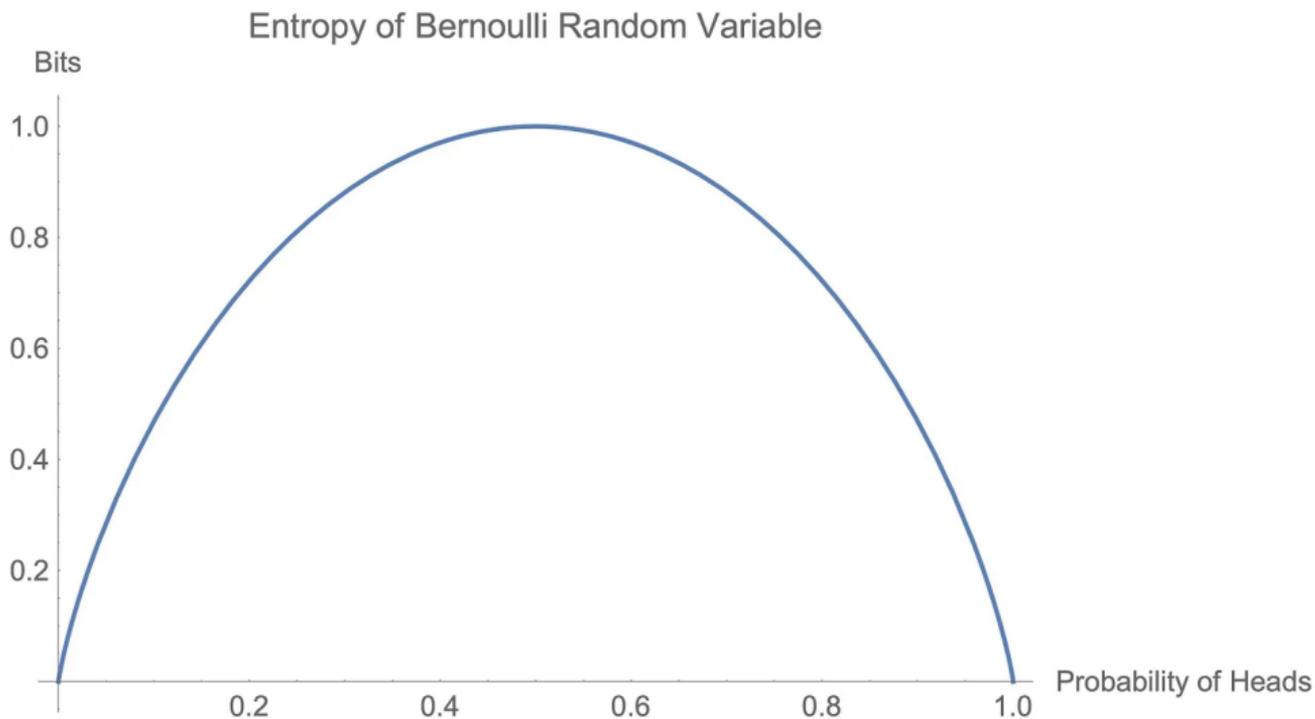
## Properties

- $\mathcal{H}(q) \geq 0$
- **Continuous formulation:** Let  $Z$  be a r.v. with distribution  $Q$ . If there exist a measure  $\mu$  such that  $dQ = q d\mu$  then we can define

$$\mathcal{H}(Q) = \mathcal{H}_\mu(q) = - \int \log q(z) q(z) d\mu(z)$$

Now depends on the base measure  $\mu$ .

# Example: entropy of a binomial random variable



Source: <https://mlberkeley.substack.com/p/max-ent>

# Kullback-Leibler (KL) divergence

## Definition: KL divergence (discrete case)

Let  $p$  and  $q$  be two distribution over discrete set  $\mathcal{Z}$ , we define the KL-divergence as

$$\text{KL}(p \parallel q) := \mathbb{E}_{Z \sim p} \left[ \log \frac{p(Z)}{q(Z)} \right] = \sum_{z \in \mathcal{Z}} p(z) \log \frac{p(z)}{q(z)}$$

## Properties

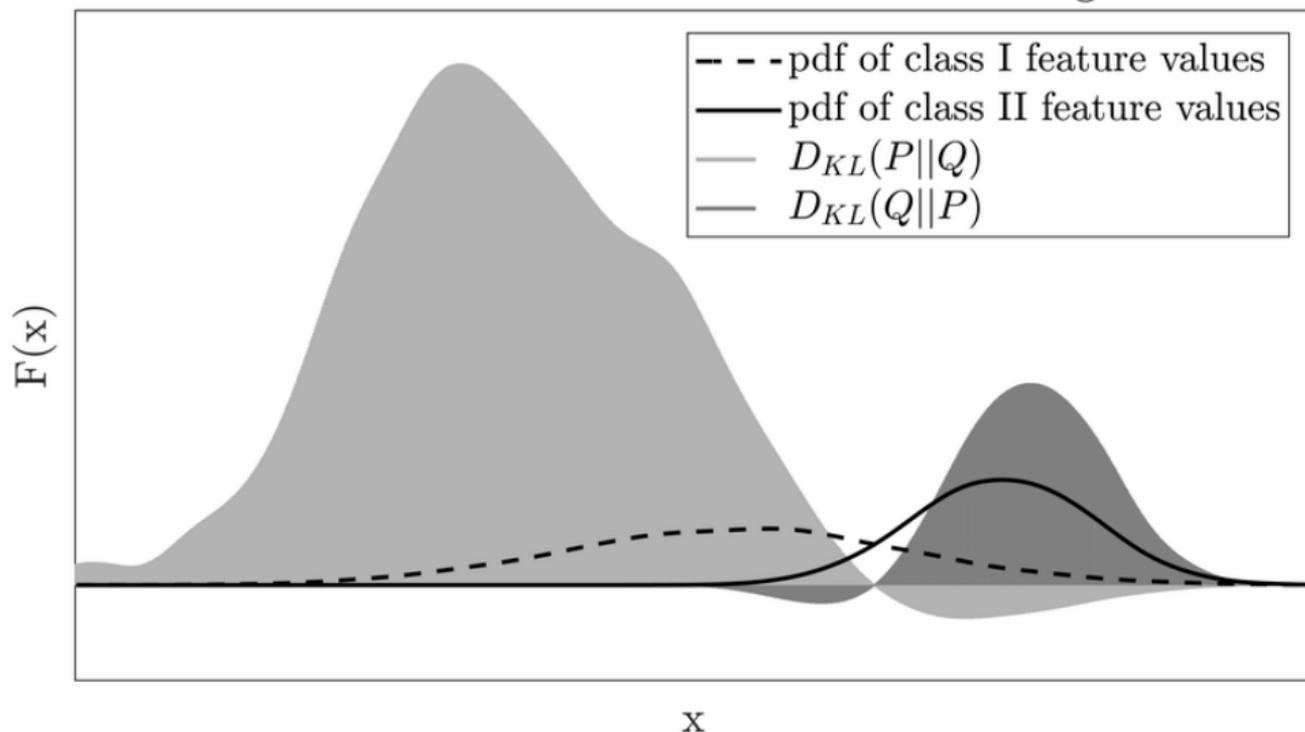
- $\text{KL}(p \parallel q) \geq 0$  with equality iff  $p = q$  (*proof*: Jensen on  $\frac{q}{p}(Z)$  with convex  $\phi(x) = -\log x$ )
- Diverges if  $\exists z_0$  such that  $q(z_0) = 0$  when  $p(z_0) \neq 0$
- Not a distance (not symmetric)
- **Continuous formulation:** For two distribution  $P$  and  $Q$ , if there exists a measure  $\mu$  such that  $dP = p d\mu$  and  $dQ = q d\mu$ , then

$$\text{KL}(P \parallel Q) = \int \log \frac{dP}{dQ} dP = \int \log \frac{p(z)}{q(z)} p(z) d\mu(z).$$

$\rightsquigarrow$  invariant w.r.t. the choice of  $(p, q, \mu)$  since the ratio  $dP/dQ$  is invariant.

# KL divergence: what does it look like?

Visualization of the Kullback–Leibler divergence



Source: <https://ieeexplore.ieee.org/document/9353046>

# The evidence lower bound (ELBO)

## Evidence lower bound

Let  $q$  be a distribution over  $\mathcal{Z}$  absolutely continuous with respect to  $p_\theta(X, Z)$ . Then,

$$\log p_\theta(\mathbf{X}) \geq \mathcal{L}(q, \theta) := \mathbb{E}_q [\log p_\theta(X, Z)] + \mathcal{H}(q). \quad (\text{ELBO})$$

The quantity  $\mathcal{L}$  is called *the evidence lower-bound*, moreover the gap is expressed as

$$\log p_\theta(X) - \mathcal{L}(q, \theta) = \text{KL}(q \parallel p_\theta(\cdot | X)).$$

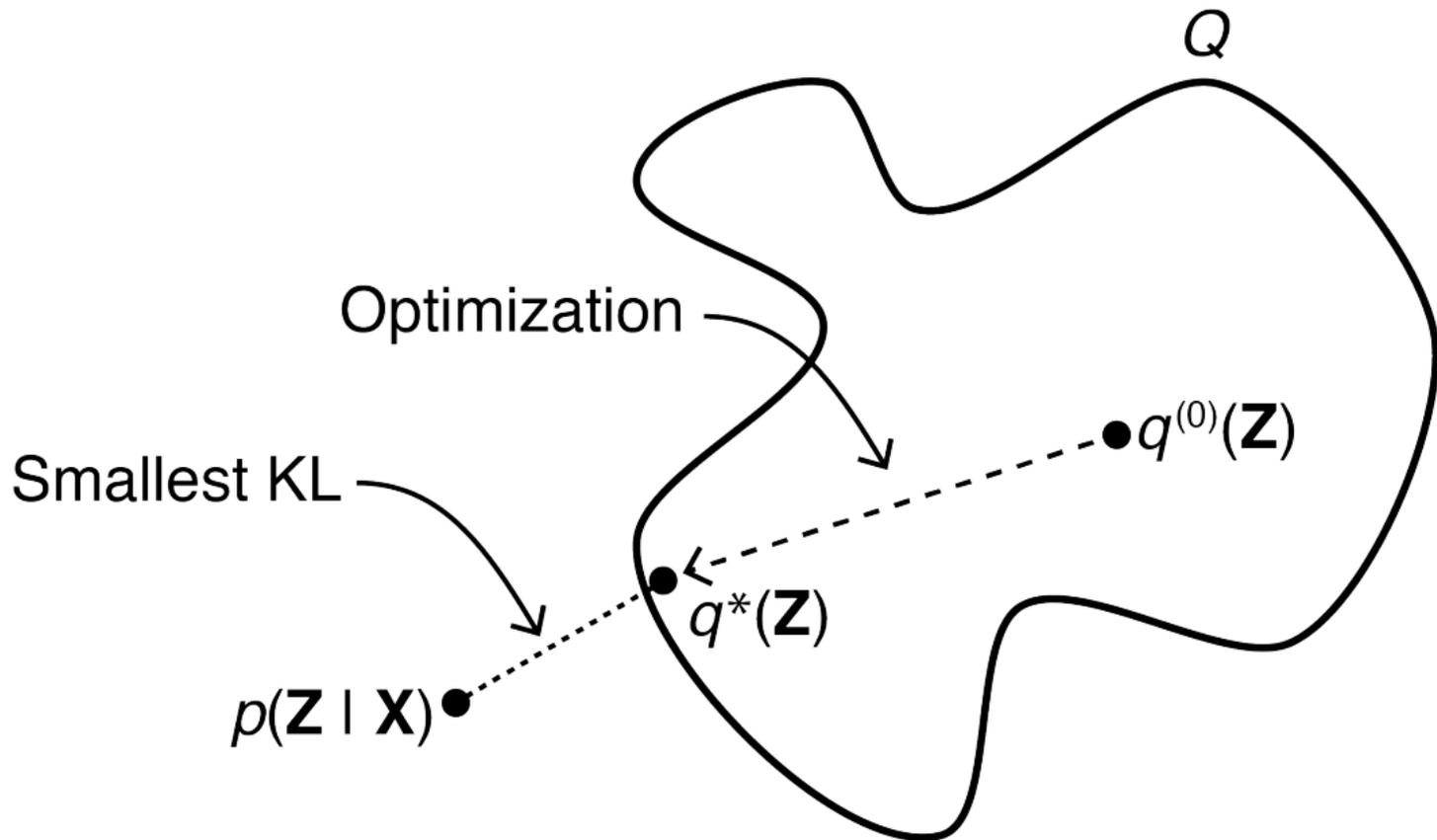
*Proof:*  $\log p_\theta(X) = \log \int p_\theta(X, z) dz = \log \mathbb{E}_q \left[ \frac{p_\theta(X, Z)}{q(Z)} \right] \stackrel{\text{Jensen}}{\geq} \mathbb{E}_q \left[ \log \frac{p_\theta(X, Z)}{q(Z)} \right] = \mathcal{L}(q, \theta)$

## Comments

- The ELBO holds for any distribution  $q$  on  $\mathcal{Z}$
- For a given  $\theta$ , the gap is 0 iff

$$q(z) = p_\theta(z | X)$$

## Visual intuition of the ELBO optimisation



**Expectation-maximization (EM, Dempster et al. 1977)**

# EM: a universal algorithm for latent variables

**Intuition:** chicken-and-egg

- 1 if we knew  $\mathbf{Z}$ , we could easily work with  $f(\theta) = \log p_{\theta}(\mathbf{X}, \mathbf{Z})$
- 2 if we knew  $\theta$ , the best representation of  $\mathbf{Z}$  is via its posterior  $p_{\theta}(\mathbf{Z} | \mathbf{X})$

## Expectation-Maximization algorithm

Starting from  $\theta^{(0)}$ , iterate between

### Expectation step

Use  $q^{(t+1)}(\mathbf{Z}) = p_{\theta^{(t)}}(\mathbf{Z} | \mathbf{X})$  to form the objective function

$$f(\theta) = Q(\theta, \theta^{(t)}) = \mathbb{E}_{\mathbf{Z} \sim q^{(t+1)}} [\log p_{\theta}(\mathbf{X}, \mathbf{Z})].$$

It involves (generalized) moments of  $\mathbf{Z}$  under  $q^{(t+1)}$ .

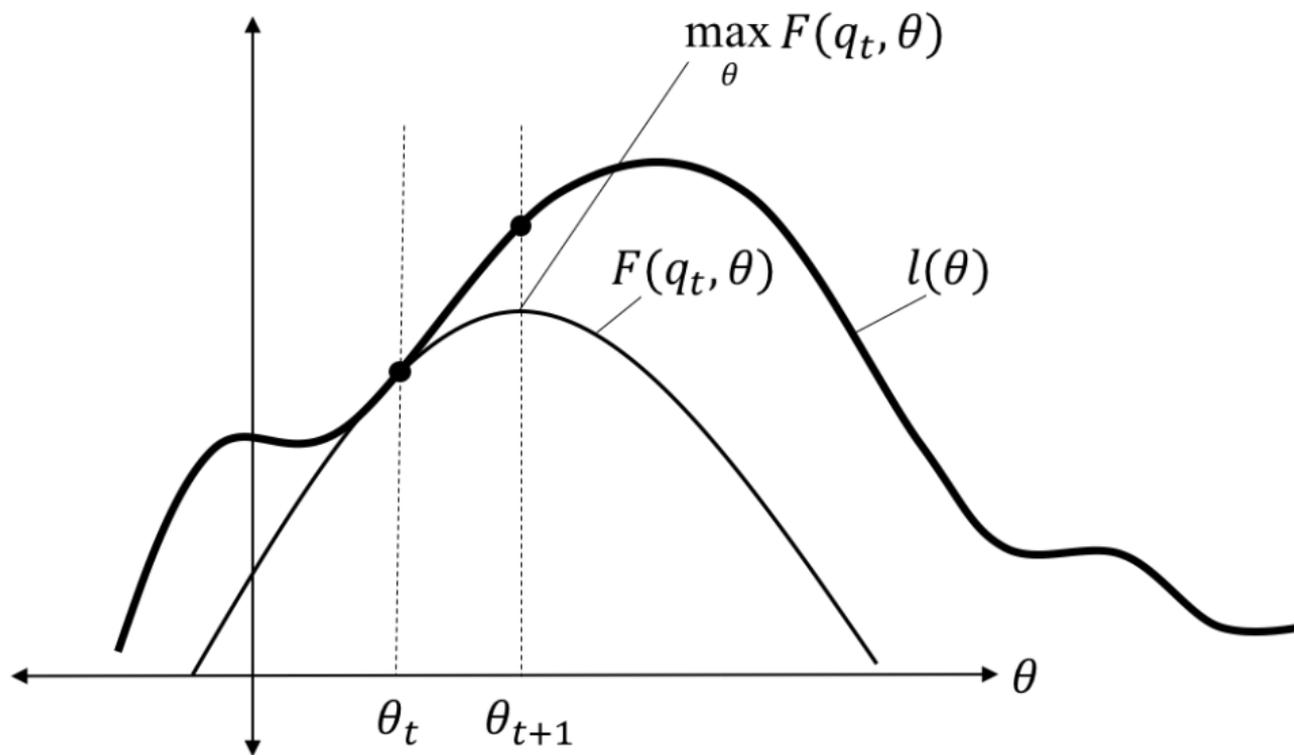
### Maximization step

Solve  $\theta^{(t+1)} \in \arg \max_{\theta} Q(\theta, \theta^{(t)})$

In practice, EM stop after likelihood gaps fall below a given threshold  $\epsilon$

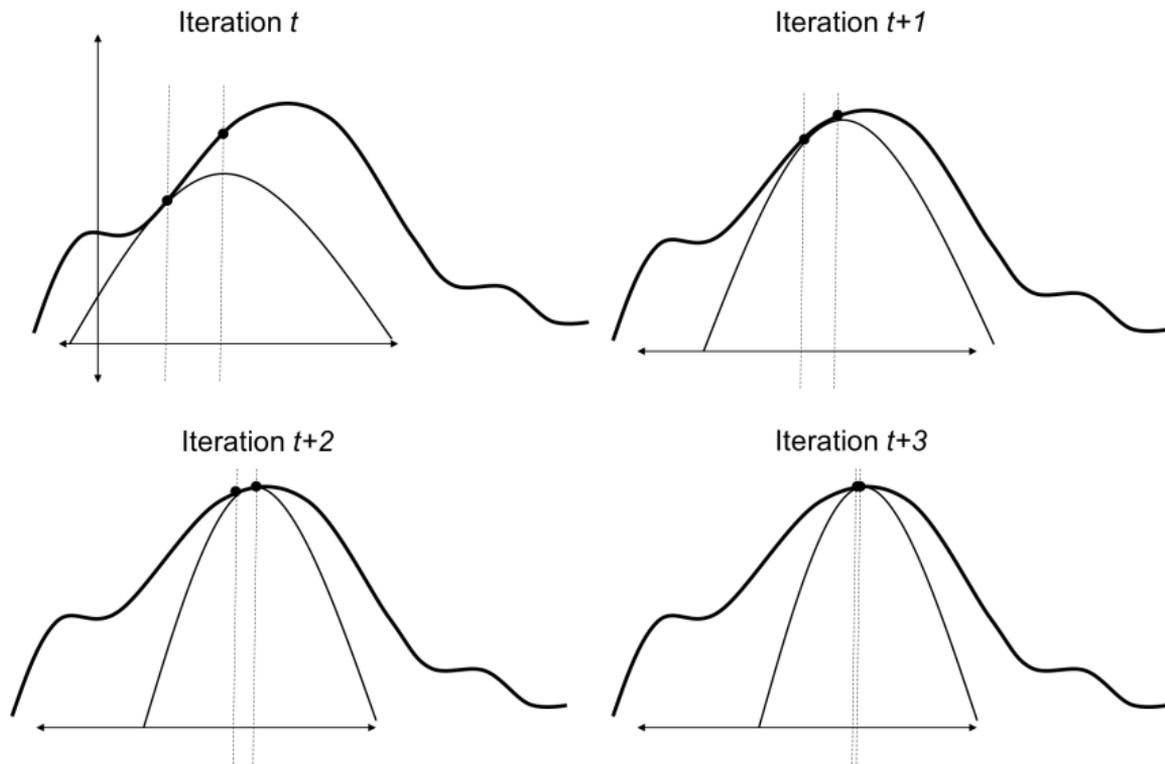
$$|\mathcal{L}(q^{(t+1)}, \theta^{(t)}) - \mathcal{L}(q^{(t)}, \theta^{(t-1)})| = |\log p_{\theta^{(t)}}(\mathbf{X}) - \log p_{\theta^{(t-1)}}(\mathbf{X})| < \epsilon$$

# A graphical illustration of EM algorithm



Source: <https://mbernste.github.io/posts/em/>

# A graphical illustration of EM algorithm



Source: <https://mbernste.github.io/posts/em/>

# Rewriting EM: coordinate ascent on the ELBO

## EM algorithm (equivalent formulation)

Starting from  $\theta^{(0)}$ , iterate between

$$q^{(t+1)} = \arg \max_q \mathcal{L}(q, \theta^{(t)}), \quad (\text{E-step})$$

$$\theta^{(t+1)} = \arg \max_{\theta} \mathcal{L}(q^{(t+1)}, \theta). \quad (\text{M-step})$$

- E-step is equivalent to  $\min_q \text{KL}(q \parallel p_{\theta^{(t+1)}}(\cdot \mid X)) \implies q^{(t+1)} = p_{\theta^{(t+1)}}(\cdot \mid X)$
- basis of inference in latent variable models, many extensions: see e.g. [Peel et al. \(2000\)](#) for mixture models

# Monotonic increase of the observed likelihood

## Property of EM algorithm

The sequence of iterates  $\{\theta^{(t)}\}_t$  returned by EM verifies

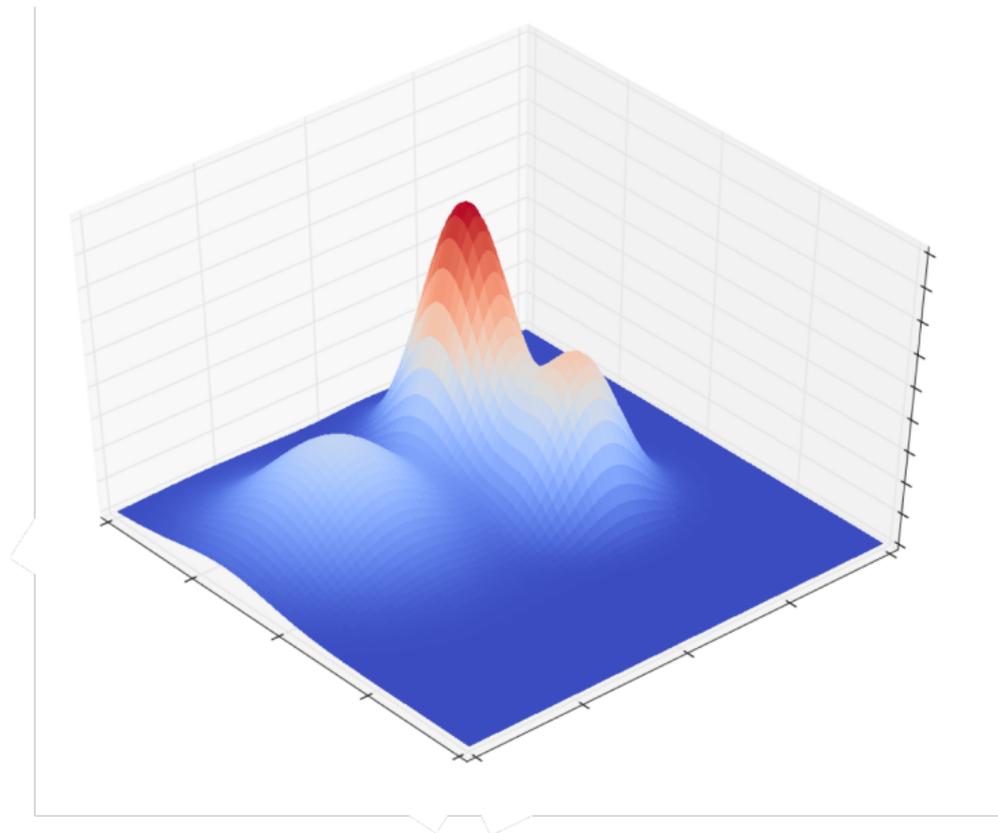
$$\forall t \geq 0, \quad \log p_{\theta^{(t+1)}}(\mathbf{X}) \geq \log p_{\theta^{(t)}}(\mathbf{X})$$

*Proof:*

$$\log p_{\theta^{(t+1)}}(\mathbf{X}) \underbrace{\geq}_{\text{ELBO}} \mathcal{L}(q^{(t+1)}, \theta^{(t+1)}) \underbrace{\geq}_{\text{M-step}(t+1)} \mathcal{L}(q^{(t+1)}, \theta^{(t)}) \underbrace{=}_{\text{E-step}(t)} \log p_{\theta^{(t)}}(\mathbf{X})$$

- Guarantees EM converges with the likelihood gaps criterion
- In general, only converges to local maxima of the likelihood
- Does not guarantee convergence of the sequence of parameters  $\{\theta^{(t)}\}_t$  itself.

# Gaussian mixture inference with the EM algorithm



Source: [https://mbernste.github.io/posts/gmm\\_em/](https://mbernste.github.io/posts/gmm_em/)

## Expected complete log-likelihood

Denote  $\tau_{ik}^{(t)} := p_{\theta^{(t-1)}}(Z_{ik} = 1 \mid x_i) \underset{\text{Multinomial}}{=} \mathbb{E}_{q^{(t)}}[Z_{ik}]$ , then

$$\begin{aligned} f(\theta) &= \mathbb{E}_{q^{(t)}} [\log p_{\theta}(\mathbf{X}, \mathbf{Z})], \\ &= \mathbb{E}_{q^{(t)}} \left[ \sum_{i=1}^n \log p_{\theta}(x_i, Z_i) \right], \\ &= \mathbb{E}_{q^{(t)}} \left[ \sum_{k=1}^K \sum_{i=1}^n Z_{ik} [\log \pi_k + \log \mathcal{N}_d(x_i \mid \mu_k, \Sigma_k)] \right], \\ &= \sum_{k=1}^K \sum_{i=1}^n \mathbb{E}_{q_i^{(t)}} [Z_{ik}] [\log \pi_k + \log \mathcal{N}_d(x_i \mid \mu_k, \Sigma_k)], \\ &= \sum_{k=1}^K \sum_{i=1}^n \tau_{ik}^{(t)} [\log \pi_k + \log \mathcal{N}_d(x_i \mid \mu_k, \Sigma_k)], \end{aligned}$$

It involves  $\tau_{ik}^{(t)}$  : (first) moments of  $Z$  under  $q^{(t)}$ .

## E-step for GMM

Compute the posterior given  $\theta^{(t-1)}$ ,  $q^{(t)} = p_{\theta^{(t-1)}}(\mathbf{Z} | \mathbf{X})$

As seen previously, the posterior for mixture model always writes

$$p_{\theta}(\mathbf{Z}) = \prod_{i=1}^n \mathcal{M}_K(1, \tau_i(\theta)), \quad \text{with: } \tau_{ik}(\theta) \propto \pi_k p_{\gamma_k}(x_i).$$

So that

$$\tau_{ik}^{(t)} = \tau_{ik}(\theta^{(t-1)}) = \frac{\pi_k \mathcal{N}_d(x_i | \mu_k^{(t-1)}, \Sigma_k^{(t-1)})}{\sum_{l=1}^K \pi_l \mathcal{N}_d(x_i | \mu_l^{(t-1)}, \Sigma_l^{(t-1)})}.$$

**Careful** with numerical underflow  $\rightsquigarrow$  better to work with in log-space with  $\log \tau$ .

# M-step for GMM

Solve

$$(\pi_k^{(t)}, \mu_k^{(t)}, \Sigma_k^{(t)})_{k=1}^K \in \arg \max_{\theta} \left\{ f(\theta) = \mathbb{E}_{q^{(t)}} [\log p_{\theta}(\mathbf{X}, \mathbf{Z})] \right\}$$

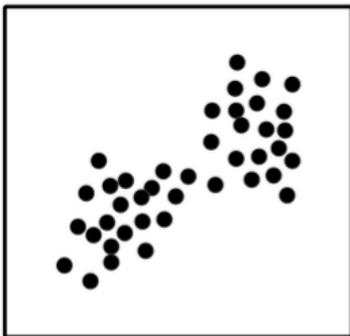
For GMM, the updates are

$$\begin{cases} \tilde{n}_k^{(t)} = \sum_{i=1}^n \tau_{ik}^{(t)}, \\ \pi_k^{(t)} = \frac{\tilde{n}_k^{(t)}}{n}, \\ \mu_k^{(t)} = \frac{1}{\tilde{n}_k^{(t)}} \sum_{i=1}^n \tau_{ik}^{(t)} x_i, \\ \Sigma_k^{(t)} = \frac{1}{\tilde{n}_k^{(t)}} \sum_{i=1}^n \tau_{ik}^{(t)} (x_i - \mu_k^{(t)})(x_i - \mu_k^{(t)})^{\top} \end{cases}$$

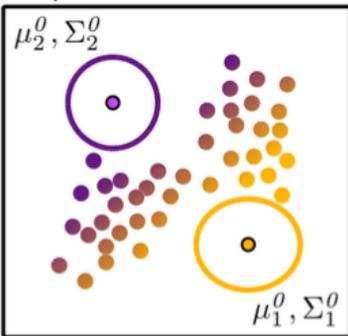
We recognize standard Gaussian MLE in each cluster, using soft probability memberships  $\tau$  in place of unknown  $\mathbf{Z}$ .

# A graphical illustration of the EM algorithm

Unknown distribution



Step 0



Step 1



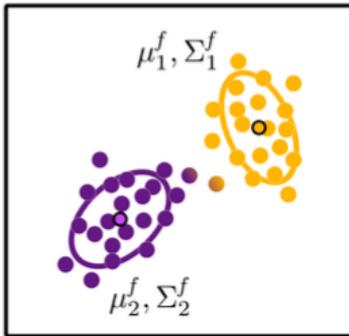
Step 2



Step 10



Final GMM



# Links with K-means algorithm

The K-means algorithm can be interpreted as an EM algorithm for a constrained GMM with equal proportions  $\pi_k = 1/K$ , known isotropic covariance  $\Sigma_k = \sigma^2 \text{Id}_d$ . Dropping the known quantities, the criterion is

$$\arg \min_{\mu_1, \dots, \mu_K, \mathbf{Z}} -\log p_{\mu}(\mathbf{X}, \mathbf{Z}) = cte + \sum_k \sum_{i \in C_k} \|x_i - \mu_k\|_2^2.$$

## Rewriting K-means (Classification-EM for GMM)

- 1 *Hard E-step*: set partition  $C^{(t+1)}$  via MAP  $\arg \max_l \tau_{il}^{(t+1)} = \arg \min_l \|x_i - \mu_l^{(t)}\|_2^2$
- 2 *M-step*: update the centroids  $\mu_k^{(t+1)} \leftarrow (1/n_k) \sum_{i \in C_k^{(t+1)}} x_i$

## Comments

- highlight connections between similarity-based and probabilistic methods
- unveil hypothesis behind K-means criterion: spherical, equal-volume and equal-size clusters.

# Choosing the number of components $K$

**Challenge:** how to choose the number of clusters  $K$  ?

**Intuition:** the larger the likelihood, the better our model fits the data  $\mathbf{X}$

**Caveat:** complex models tend to provide larger likelihood, for example

- mixture models with  $K - 1$  components are nested in models with  $K$  components.
- models with constraints (diagonal, spherical) are nested in unconstrained ones.

↪ we need to account for "model complexity"

**Definition: dimension/size of a model**

Let  $\mathcal{M} = \{p_\theta, \theta \in \Theta_{\mathcal{M}}\}$ , we denote  $d_{\mathcal{M}}$  the number of free parameters in the model. For unconstrained mixtures, it is  $d_K = K - 1 + Kd_\Gamma$ ,  $\gamma_k \in \Gamma$ .

## Penalized likelihood criterion

For a mixture model with  $K$  components, denote  $\hat{\theta}_K = \arg \max_{\theta \in \Theta_K} \log p_\theta(\mathbf{X})$ . A *penalized likelihood* estimate of  $K$  is given by

$$\hat{K} = \arg \max_K \left\{ \log p_{\hat{\theta}_K} - \text{pen}(K) \right\}.$$

# Different penalties leads to different criterion

## Definitions: AIC, BIC, ICL

For a model  $\mathcal{M}$  and observations  $\mathbf{X}$ , we have several choice of penalize likelihood criteria

$$AIC(K) := \log p_{\hat{\theta}_K}(\mathbf{X}) - d_K,$$

$$BIC(K) := \log p_{\hat{\theta}_K}(\mathbf{X}) - \frac{d_K}{2} \log(n),$$

$$ICL(K) := \mathbb{E}_{Z \sim p_{\hat{\theta}_K}(\cdot | \mathbf{X})} \left[ \log p_{\hat{\theta}_K}(\mathbf{X}, \mathbf{Z}) \right] - \frac{d_K}{2} \log(n)$$

**Note:** the ELBO property gives

$$ICL(K) = BIC(K) - \mathcal{H}(p_{\hat{\theta}_K}(\cdot | \mathbf{X})).$$

Hence, ICL is more focused on models with strongly separable clusters (peaked posterior  $\implies$  low entropy), while BIC is more focused on fitting the marginal density of  $\mathbf{X}$ .

# Hidden Markov Models (HMMs)

# Motivations

**What if** observations  $\mathbf{X} = \{x_i\}_i$  are ordered ? e.g.

- time series (occurring in weather forecast, finance, biology, ...)
- genomic data: observations collected at precise locations on the genome
- etc.

↪ it is likely that "past" influences the "future".

Need to introduce **dependence** between observations/latent variables in the model

# Motivations

**What if** observations  $\mathbf{X} = \{x_i\}_i$  are ordered ? e.g.

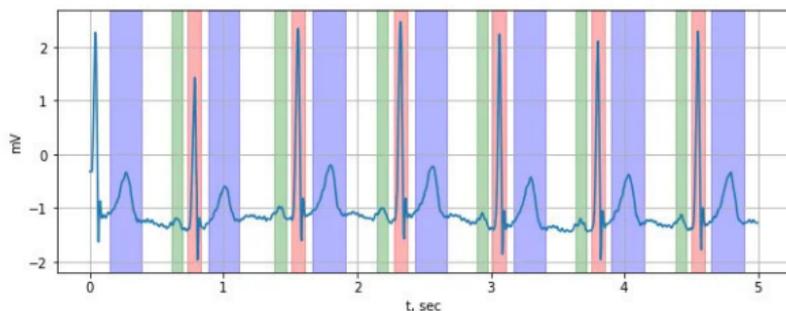
- time series (occurring in weather forecast, finance, biology, ...)
- genomic data: observations collected at precise locations on the genome
- etc.

↪ it is likely that "past" influences the "future".

Need to introduce **dependence** between observations/latent variables in the model

## Example 1: time series segmentation

Heart rate: 78 bpm



Source: <https://medium.com/data-analysis-center/56f8b9abd83a>

# Motivations

**What if** observations  $\mathbf{X} = \{x_i\}_i$  are ordered ? e.g.

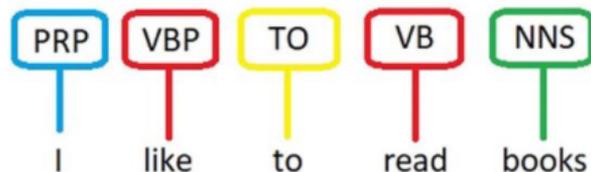
- time series (occurring in weather forecast, finance, biology, ...)
- genomic data: observations collected at precise locations on the genome
- etc.

↪ it is likely that "past" influences the "future".

Need to introduce **dependence** between observations/latent variables in the model

**Example 2: part-of-speech tagging**

## POS Tagging



# Motivations

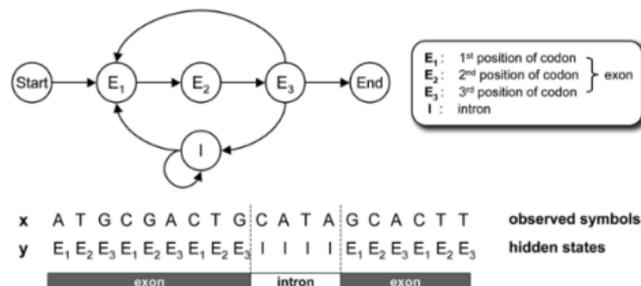
What if observations  $\mathbf{X} = \{x_i\}_i$  are ordered? e.g.

- time series (occurring in weather forecast, finance, biology, ...)
- genomic data: observations collected at precise locations on the genome
- etc.

↪ it is likely that "past" influences the "future".

Need to introduce **dependence** between observations/latent variables in the model

## Example 3: protein coding



From Yoon (2009)

## Reminder on discrete Markov chains

# Markov Chains (discrete)

Suppose we observe a sequence  $y_{1:n} := \{y_1, \dots, y_n\}$  at *discrete* time<sup>3</sup> steps  $1, \dots, n$ , with discrete outcomes  $y_i \in \{1, \dots, K\}$

## Markov chain (MC)

We say that the sequence  $y_{1:n}$  is a Markov Chain if for all  $i = 1, \dots, n$ ,

$$p(y_{i+1} \mid y_{1:i}) = p(y_{i+1} \mid y_i)$$

*"The future is independent from the past knowing the present."*

### Joint distribution of the sequence

$$p(y_{1:n}) = p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_2) \cdots p(y_n \mid y_{n-1}) = p(y_1) \prod_{i=2}^n p(y_i \mid y_{i-1}).$$

Proof of all the statements made about Markov Chains can be found e.g. in Sophie Lemaire's course : <https://www.imo.universite-paris-saclay.fr/~sophie.lemaire/coursCM13.pdf>.

---

<sup>3</sup>"Time" may also refer to locations within a sequence of words/genes/etc.

# Vocabulary around MC

## Homogeneous Markov chain

We say that a markov chain is homogeneous (or time invariant) if the transition probability  $p(y_{i+1} | y_i)$  is independent time (of  $i$ ).

### Initial distribution

We denote as  $\nu = (\nu_1, \dots, \nu_K)$  the vector  $\nu_k := p(y_1 = k)$

### Marginal distribution

We denote as  $\nu_i = (\nu_{i1}, \dots, \nu_{iK})$  the vector  $\nu_{ik} := p(y_i = k)$

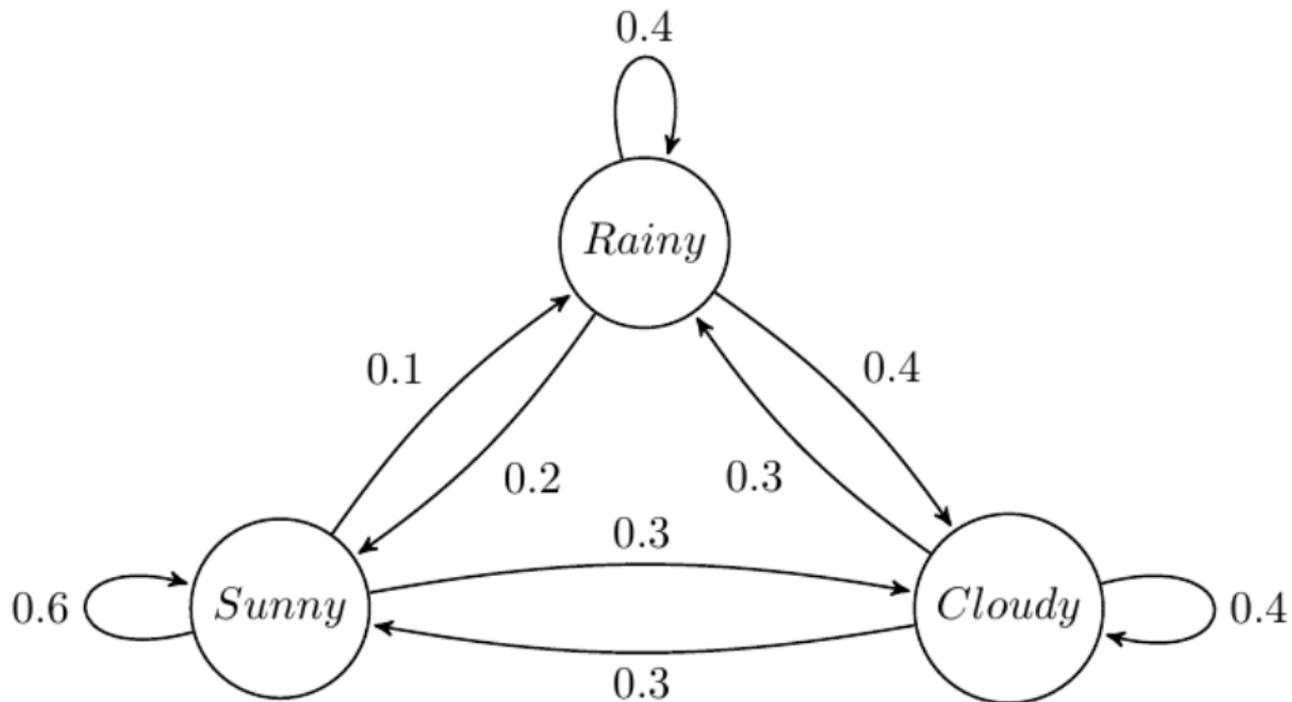
### Transition matrix

We denote  $A$  the  $K \times K$  matrix with  $A_{kl} = p(z_{i+1} = l | z_i = k)$  and properties:

- stochastic matrix: each row sum to 1 -  $\sum_{l=1}^K A_{kl} = 1$
- eigenvalue 1 associated to the column vector  $e = (1, \dots, 1)^\top$ :  $Ae = 1 \cdot e$
- For any  $m, n \in \mathbb{N}$ ,  $p(y_{n+m} = l | y_n = k) = A_{kl}^{(m)}$  ( $m$ -th matrix power)
- Moreover  $\nu_i = \nu_1 A^{(i-1)}$

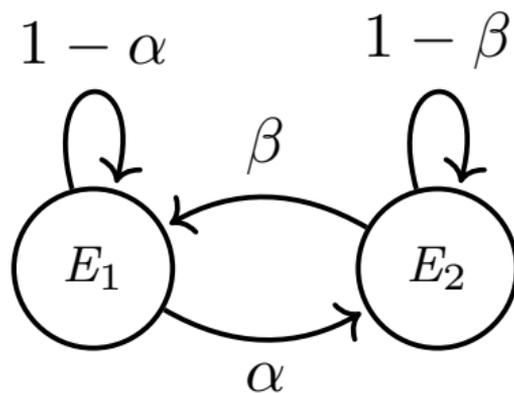
**Notation:**  $y_{1:n} \sim MC(\nu, A)$

## Diagram representation: a toy example



## Link between transition matrix and diagram representation

$$A = \begin{matrix} & \begin{matrix} E_1 & E_2 \end{matrix} \\ \begin{matrix} E_1 \\ E_2 \end{matrix} & \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix} \end{matrix}$$



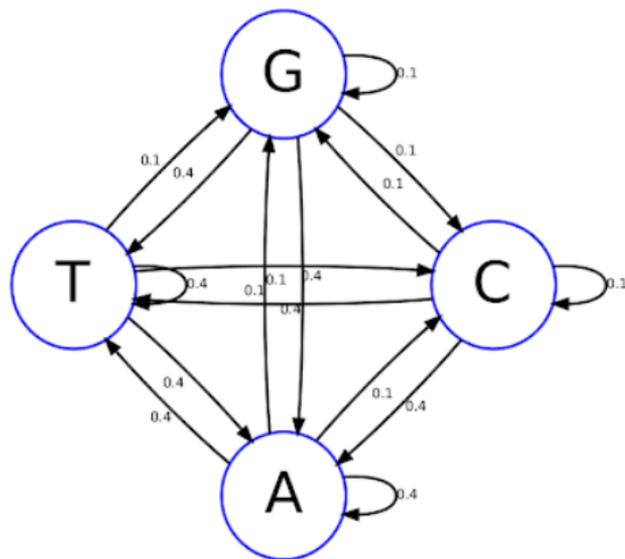
*Graphical representation of a 2-state homogeneous Markov chain*

## A second example: modeling nucleotide transition

$$A = \begin{matrix} & A & T & G & C \\ \begin{matrix} A \\ T \\ G \\ C \end{matrix} & \left( \begin{array}{cccc} & & & \\ & & & \\ & & p(y_t=G|y_{t-1}=T) & \\ & & & \end{array} \right) \end{matrix}$$

$$= \begin{matrix} & A & T & G & C \\ \begin{matrix} A \\ T \\ G \\ C \end{matrix} & \left( \begin{array}{cccc} 0.4 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.1 \end{array} \right) \end{matrix}$$

→



Source: <https://www.r-bloggers.com/2012/04/introduction-to-markov-chains-and-modeling-dna-sequences-in-r/>

# Let's practice: Ehrenfest's urn model

- 4 balls distributed across 2 urns
- Each turn, we pick a ball and change its urn
- Let  $A$  be the transition of one urn (symmetric problem) :

→ State = number of balls in this urn

Can you draw the diagram representation and the associated transition matrix?

## Let's practice: Ehrenfest's urn model

- 4 balls distributed across 2 urns
- Each turn, we pick a ball and change its urn
- Let  $A$  be the transition of one urn (symmetric problem) :  
→ State = number of balls in this urn

$$A = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix}$$

# Stationary distribution & how to find them

## Stationary distribution

Let  $A$  be a transition matrix over  $[[1, K]]$ , we say that a vector  $\pi$  such that

$$\pi^\top A = \pi^\top \qquad \sum_{k=1}^K \pi_k = 1, \quad \pi_k \geq 0$$

is a **stationary** (or **invariant**) distribution for the homogeneous chain  $MC(\nu, A)$ .

## Properties

- 1  $\pi$  is a discrete probability vector & eigenvector of  $A^\top$  associated to the eigenvalue  $\lambda = 1$
- 2 if  $y_1 \sim \pi$ , then  $\forall n \in \mathbb{N}$ ,  $y_n \sim \pi$  (hence the name *stationary*)
- 3 **Existence:** for discrete MC it is an application of Perron-Frobenius theorem to  $A$
- 4 **Uniqueness & convergence:** if there exists some power  $q \in \mathbb{N}^*$  such that  $A^{(q)} > 0$  then
  - $\pi$  is unique and  $\pi_k > 0$ .
  - $p(y_n = l \mid y_1 = k) = A_{kl}^{(n)} \xrightarrow[n \rightarrow +\infty]{} \pi_l$ , whatever the initial distribution  $\nu$  is.

Such chains "forget their past" after enough steps.

# Computing the stationary distribution

## First strategy: eigenvector

We know that  $A^T \pi = 1 \cdot \pi$ , so that  $\pi$  is an eigenvector associated to the unit<sup>a</sup> eigenvalue.

**Careful**, most scientific softwares give eigenvector such that  $\|v\|_2 = 1$ , so we need to post process  $\pi := v / (\sum_k v_k)$ .

When  $K$  is big, there are efficient algorithms to find only largest eigenvector under conditions on  $A$  (e.g. Lanczos algorithm for symmetric matrices)

---

<sup>a</sup>Recall that eigenvalues (but not eigenvectors) of  $A$  and  $A^T$  are the same.

## Second strategy: linear system

We have  $K$  unknown  $\pi_1, \dots, \pi_K$  and  $K + 1$  equations  $\pi^T (A - I) = \mathbf{0}_{1 \times K}$  &  $\sum_k \pi_k = 1$   
 $\rightsquigarrow$  over-determined linear system.

Thus, we can create a new matrix  $M$  by arbitrarily replace a column (say last one) in  $(A - I)$  by  $\mathbf{1}_{K \times 1}$  and solve for  $\pi^T M = (0, \dots, 0, 1)$ .

## 2-state example

Compute the stationary distribution of

$$A = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}$$

## 2-state example

Compute the stationary distribution of

$$A = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}$$

$$A - I = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$$

Replacing last column by  $(1, 1)^\top$  and solving the linear system when

$$(\pi_1 \quad \pi_2) \begin{pmatrix} -\alpha & 1 \\ \beta & 1 \end{pmatrix} = (0 \quad 1)$$

leads to  $\pi = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$  provided  $\alpha + \beta \neq 0$ .

**Bonus question:** when do we have convergence of  $A^n$ ? (Consider the matrix  $A$  on limit cases  $\alpha = \beta = r$ ,  $r \in \{0, 1\}$ )

## Numerical example

Find the stationary distribution of  $A = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}$

```
A <- matrix(c(0, 1/2,1,1,0,0,0,1/2,0),3,3)
```

### Eigenvector

```
eigen.res <- eigen(t(A))  
Pi <- eigen.res$vectors[,1]  
Pi/sum(Pi)
```

```
 [,1]  [,2]  [,3]  
[1,] 0.4+0i 0.4+0i 0.2+0i
```

### Linear system

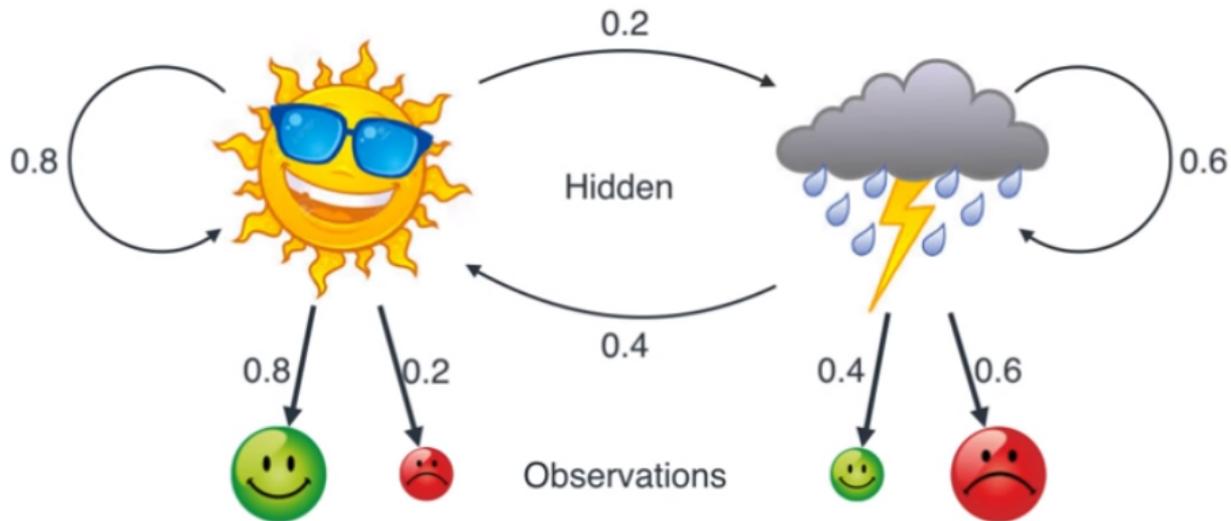
```
M<-diag(1, 3, 3) - A  
M[,3] <- rep(1,3)  
Pi <- solve(t(M),b=c(0,0,1))  
Pi
```

```
[1] 0.4 0.4 0.2
```

**Sanity check**  $(\text{Pi} - \text{t}(\text{Pi})\%*\%A) < 1\text{e-}15$

# Hidden Markov Models (HMMs)

## Hidden Markov Model



Source: <https://medium.com/analytics-vidhya/hidden-markov-models-for-time-series-in-r-studio-5ae2b9fb0701>

# HMM: the model

## Generative model

A general (discrete) hidden Markov model is defined as

1  $z_{1:n} \sim MC(\nu, A)$

2  $(x_i)_i$  independent |  $(z_i)_i$  and for all  $i \in \llbracket 1, n \rrbracket$ ,  $x_i | \{z_{ik} = 1\} \sim p_{\gamma_k}(\cdot)$

The model parameters are  $\theta = (\nu, A, \gamma)$  and  $p(x_i | z_{ik} = 1) = p_{\gamma_k}(x_i)$  are called *emission probability*

### Marginal likelihood of $x_i$

Denote  $\nu_i = (\nu_{i1}, \dots, \nu_{iK})$ , such that  $\nu_{ik} = p_{\theta}(z_{ik} = 1)$ <sup>a</sup>. Then,

$$p_{\theta}(x_i) = \sum_k p(x_i | z_{ik} = 1) p_{\theta}(z_{ik} = 1) = \sum_k \nu_{ik} p_{\gamma_k}(x_i)$$

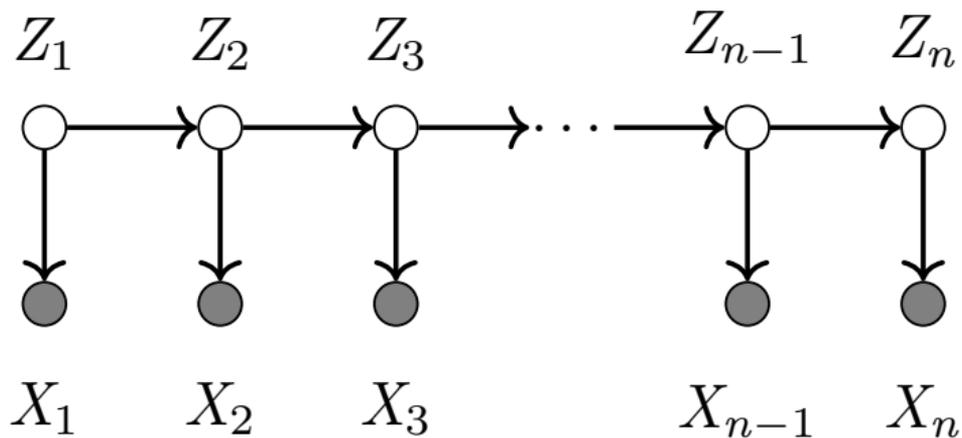
Moreover, if  $\nu_1 = \pi$  (the chain's stationary distribution) then  $p_{\theta}(x_i) = \sum_k \pi_k p_{\gamma_k}(x_i)$

↪ HMMs can be thought of as a generalization of mixture introducing dependency!

---

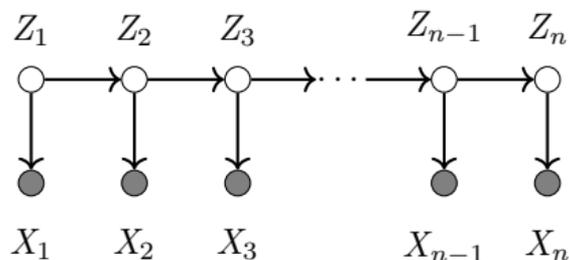
<sup>a</sup>For homogeneous MC we know that  $\nu_i = \nu^{\top} A^{(i-1)}$ .

## Graphical model representation



- Empty circle  $\circ$  represents unobserved random variable
- Gray circles  $\bullet$  represents observed random variables

# Conditional independence



Looking at the DAG, we have the three fundamental properties of HMM

- 1  $Z_{i+1} \perp\!\!\!\perp Z_{1:(i-1)} \mid Z_i$  (i.e.  $Z_{1:n}$  is a MC)
- 2  $Z_{i+1} \perp\!\!\!\perp X_{1:i} \mid Z_i$
- 3  $X_{i+1} \perp\!\!\!\perp X_{1:i} \mid Z_{i+1}$  (and also  $\mid Z_i$ )

This basically states that knowing the hidden state at step  $i$  captures all relevant information about the past.

# Observed-data likelihood

## Observed-data log-likelihood for HMMs

$$\begin{aligned}\log p_{\theta}(\mathbf{X}) &= \log \sum_{\mathbf{Z}} p_{\theta}(\mathbf{X} \mid \mathbf{Z}) \times p_{\theta}(\mathbf{Z}), \\ &= \log \left[ \sum_{z_1, \dots, z_n} \prod_{i=1}^n \prod_{k=1}^K p_{\gamma_k}(x_i)^{z_{ik}} \times \prod_{k=1}^K \nu_k^{z_{1k}} \prod_{k,l=1}^K \prod_{i=2}^n A_{k,l}^{z_{(i-1)k} z_{il}} \right].\end{aligned}$$

Brute force computation involves  $\mathcal{O}(K^n)$  operations !

# Complete-data likelihood

## Complete-data log-likelihood for HMMs

$$\begin{aligned}\log p_{\theta}(\mathbf{X}, \mathbf{Z}) &= \log p_{\theta}(\mathbf{X} \mid \mathbf{Z}) \times p_{\theta}(\mathbf{Z}), \\ &= \log \left[ \prod_{k=1}^K \prod_{i=1}^n p_{\gamma_k}(x_i)^{z_{ik}} \times \prod_{k=1}^K \nu_k^{z_{1k}} \prod_{k,l=1}^K \prod_{i=2}^n A_{k,l}^{z_{(i-1)k} z_{il}} \right], \\ &= \sum_{k=1}^K \sum_{i=1}^n z_{ik} \log p_{\gamma_k}(x_i) + \sum_{k=1}^K z_{1k} \log \nu_k + \sum_{k,l=1}^K \sum_{i=2}^n z_{(i-1)k} z_{il} \log A_{k,l}.\end{aligned}$$

# Posterior distribution

Denote

$$\tau_{ik} := p_{\theta}(z_{ik} = 1 \mid X)$$

## Important: posterior dependencies

Contrary to mixture models

- 1  $\tau_{ik} \neq p(z_{ik} = 1 \mid x_i) \rightsquigarrow$  we need the whole set of observations
- 2 More generally,  $p_{\theta}(\mathbf{Z} \mid \mathbf{X})$  does not factorizes over  $i$  anymore

$$p_{\theta}(\mathbf{Z} \mid \mathbf{X}) \neq \prod_i \prod_k \tau_{ik}^{z_{ik}}$$

- 3  $(z_i)_i$  are not independent  $\mid (x_i)_i$  but rather  $(z_{1:n}) \mid (x_{1:n})$  is an inhomogeneous MC

# Posterior distribution

Denote

$$\tau_{ik} := p_{\theta}(z_{ik} = 1 \mid X)$$

## Important: posterior dependencies

Contrary to mixture models

- 1  $\tau_{ik} \neq p(z_{ik} = 1 \mid x_i) \rightsquigarrow$  we need the whole set of observations
- 2 More generally,  $p_{\theta}(\mathbf{Z} \mid \mathbf{X})$  does not factorizes over  $i$  anymore

$$p_{\theta}(\mathbf{Z} \mid \mathbf{X}) \neq \prod_i \prod_k \tau_{ik}^{z_{ik}}$$

- 3  $(z_i)_i$  are not independent  $\mid (x_i)_i$  but rather  $(z_{1:n}) \mid (x_{1:n})$  is an inhomogeneous MC

**Exercise: Prove that the following Markov property indeed holds**

$$p_{\theta}(z_{i+1} \mid z_{1:i}, x_{1:n}) = p_{\theta}(z_{i+1} \mid z_i, x_{1:n})$$

## Solution for the Markov property

$$\begin{aligned} p_{\theta}(z_{i+1} \mid z_{1:i}, x_{1:n}) &= p_{\theta}(z_{i+1} \mid z_{1:i}, x_{(i+1):n}) \\ &\stackrel{\text{Bayes}}{\propto} p_{\theta}(x_{(i+1):n} \mid z_{1:i}, z_{i+1}) p_{\theta}(z_{i+1} \mid z_{1:i}) \\ &\stackrel{p(a,b)=p(a|b)p(b)}{=} p_{\theta}(x_{(i+1):n}, z_{i+1} \mid z_{1:i}) \\ &\stackrel{\text{Markov}}{=} p_{\theta}(x_{(i+1):n}, z_{i+1} \mid z_i) \\ &\stackrel{p(a,b)=p(a|b)p(b)}{=} p_{\theta}(x_{(i+1):n} \mid z_{i+1}, z_i) p_{\theta}(z_{i+1} \mid z_i) \\ &\stackrel{\text{Bayes}}{\propto} p_{\theta}(z_{i+1} \mid z_i, x_{(i+1):n}) \\ &= p_{\theta}(z_{i+1} \mid z_i, x_{1:n}). \end{aligned}$$

## Solution for the Markov property

$$\begin{aligned} p_{\theta}(z_{i+1} \mid z_{1:i}, x_{1:n}) &= p_{\theta}(z_{i+1} \mid z_{1:i}, x_{(i+1):n}) \\ &\stackrel{\text{Bayes}}{\propto} p_{\theta}(x_{(i+1):n} \mid z_{1:i}, z_{i+1}) p_{\theta}(z_{i+1} \mid z_{1:i}) \\ &\stackrel{p(a,b)=p(a|b)p(b)}{=} p_{\theta}(x_{(i+1):n}, z_{i+1} \mid z_{1:i}) \\ &\stackrel{\text{Markov}}{=} p_{\theta}(x_{(i+1):n}, z_{i+1} \mid z_i) \\ &\stackrel{p(a,b)=p(a|b)p(b)}{=} p_{\theta}(x_{(i+1):n} \mid z_{i+1}, z_i) p_{\theta}(z_{i+1} \mid z_i) \\ &\stackrel{\text{Bayes}}{\propto} p_{\theta}(z_{i+1} \mid z_i, x_{(i+1):n}) \\ &= p_{\theta}(z_{i+1} \mid z_i, x_{1:n}). \end{aligned}$$

# The "three" HMM problems

Following Rabiner (1989), there are three problems related to HMMs:

- 1 Given  $\theta$  the model parameters, compute the probability of observing  $x_{1:n}$  (i.e. the observed likelihood)

$$p_{\theta}(x_{1:n})$$

- 2 **Decoding** given  $\theta$  the model parameters and observations  $x_{1:n}$ , find the most probable sequence of hidden states

$$\hat{z}_{1:n} = \arg \max_{z_{1:n}} p_{\theta}(z_{1:n} | x_{1:n})$$

- 3 **Inference**: estimate the model parameters, e.g. by MLE

$$\hat{\nu}, \hat{A}, \hat{\gamma} \in \arg \max_{\theta} p_{\theta}(x_{1:n})$$

Actually, many others linked problems...

- **Prediction**:  $p_{\theta}(z_{n+m} | x_{1:n})$  for  $m \geq 1$
- **Filtering**:  $p_{\theta}(z_i | x_{1:i})$
- **Smoothing**:  $p_{\theta}(z_i | x_{1:n}) \neq$  filtering, notice the conditioning on all the evidence

## Inference in HMMs

# Reminder on MLE & EM

$$\hat{\theta} \in \arg \max_{\theta} p_{\theta}(\mathbf{X})$$

## EM algorithm

Start with  $\theta^{(0)}$  and repeat until convergence

- **E-step:** given the current estimate  $\theta^{(t)}$ , compute the posterior  $p_{\theta^{(t)}}(\mathbf{Z} | \mathbf{X})$ , or at least all its necessary moments to compute

$$\mathbb{E}_{\mathbf{Z} \sim p_{\theta^{(t)}}(\cdot | \mathbf{X})} [\log p_{\theta}(\mathbf{X}, \mathbf{Z})] .$$

- **M-step:** update the estimate of  $\theta$  with

$$\theta^{(t+1)} \in \arg \max_{\theta} \mathbb{E}_{\theta^{(t)}} [\log p_{\theta}(\mathbf{X}, \mathbf{Z}) | \mathbf{X}] .$$

## E-step: compute $\mathbb{E}_{\mathbf{Z} \sim p_{\theta^{(t)}}(\cdot | \mathbf{X})} [\log p_{\theta}(\mathbf{X}, \mathbf{Z})]$

In Slide 123 we derived the expression of  $\log p_{\theta}(\mathbf{X}, \mathbf{Z})$ , hence using linearity of  $\mathbb{E}$  we get:

$$\begin{aligned}\mathbb{E} [\log p_{\theta}(\mathbf{X}, \mathbf{Z}) | \mathbf{X}] &= \mathbb{E} \left[ \sum_{k=1}^K z_{1k} \log \nu_k + \sum_{k,l=1}^K \sum_{i=2}^n z_{(i-1)k} z_{il} \log A_{k,l} | \mathbf{X} \right] \\ &\quad + \mathbb{E} \left[ \sum_{k=1}^K \sum_{i=1}^n z_{ik} \log p_{\gamma_k}(x_i) | \mathbf{X} \right], \\ &= \sum_{k=1}^K \tau_{1k} \log \nu_k + \sum_{k,l=1}^K \sum_{i=2}^n \xi_{i,k,l} \log A_{k,l} + \sum_{k=1}^K \sum_{i=1}^n \tau_{ik} \log \Psi_i(k).\end{aligned}$$

Where:

$$\Psi_i(k) := p_{\gamma_k}(x_i), \quad (\text{Emission probability})$$

$$\tau_{ik} := p_{\theta^{(t)}}(z_{ik} = 1 | \mathbf{X}) = \mathbb{E} [z_{ik} | \mathbf{X}],$$

$$\xi_{i,k,l} := p_{\theta^{(t)}}(z_{(i-1)k} = 1, z_{il} = 1 | \mathbf{X}) = \mathbb{E} [z_{(i-1)k} z_{il} | \mathbf{X}]$$

Hence, we need to compute "smoothed" posterior of all *unigrams*  $z_i$  and *bi-grams*  $(z_{i-1}, z_i)$   
 $\rightsquigarrow$  no straight-forward closed form as in mixture since  $p(z_i | \mathbf{X}) \neq p(z_i | x_i)$  anymore

## Intuition: "breaking" the chain

The smoothed posteriors can be computed thanks to a recursion called forward-backward.

The key decomposition lies with the fact that the chain can be split into two distinct parts - past and future - conditionally on  $z_i$

$$\begin{aligned} p(z_i = k, x_{1:n}) &= p\left(z_i = k, x_{1:i}, x_{(i+1):n}\right), \\ &= p\left(x_{(i+1):n} \mid z_i = k, \cancel{x_{1:i}}\right) p(x_{1:i}, z_i = k). \end{aligned}$$

# The forward-backward algorithm

## Proposition

For a given parameter  $\theta$ , the posterior probabilities  $\tau_{ik}$  and  $\xi_{i,k,l}$  can be computed by the two following recursions (we drop the  $\theta$  dependencies for readability,  $p = p_\theta$ )

**Forward-step** filtering step  $\alpha_i = (\alpha_i(1), \dots, \alpha_i(K))$  with

$$\alpha_i(k) = p(z_i = k, x_{1:i}) \longrightarrow \begin{cases} \alpha_1 = \nu \odot \Psi_1, \\ \alpha_i = \Psi_i \odot (A^\top \alpha_{i-1}). \end{cases} \quad (\text{Forward recursion})$$

**Backward** compute likelihood of future evidence given that  $z_i = k$

$$\beta_i(k) = p(x_{(i+1):n} \mid z_i = k) \longrightarrow \begin{cases} \beta_n = 1, \\ \beta_{i-1} = A(\Psi_i \odot \beta_i). \end{cases} \quad (\text{Backward recursion})$$

**Then** the smoothed posteriors are obtained with

$$\begin{aligned} \tau_{ik} &= p(z_i = k \mid \mathbf{X}) \propto \alpha_i(k) \beta_i(k), \\ \xi_{i,k,l} &= p(z_i = k, z_{i+1} = l \mid \mathbf{X}) \propto \alpha_i(k) \Psi_{i+1}(l) \beta_{i+1}(l) A_{kl} \end{aligned}$$

## Proof of the forward recursion

$$\begin{aligned}\alpha_i(k) &= p(x_{1:i}, z_i = k) = \sum_{l=1}^K p(x_{1:i}, z_{i-1} = l, z_i = k), \\ &= \sum_{l=1}^K p(x_{1:i-1}, x_i, z_{i-1} = l, z_i = k), \\ &= \sum_{l=1}^K p(x_i, z_i = k \mid x_{1:i-1}, z_{i-1} = l) p(x_{1:i-1}, z_{i-1} = l), \\ &= \sum_{l=1}^K p(x_i \mid z_i = k, \cancel{x_{1:i-1}, z_{i-1} = l}) p(z_i = k \mid \cancel{x_{1:i-1}, z_{i-1} = l}) p(x_{1:i-1}, z_{i-1} = l), \\ &= p(x_i \mid z_i = k) \sum_{l=1}^K p(z_i = k \mid z_{i-1} = l) p(x_{1:i-1}, z_{i-1} = l), \quad (\text{HMM model}) \\ &= \Psi_i(k) \sum_{l=1}^K A_{lk} \alpha_{i-1}(l).\end{aligned}$$

$$\implies \alpha_i = \Psi_i \odot (A^\top \alpha_{i-1}).$$

## Proof of the backward recursion

$$\begin{aligned}\beta_{i-1}(k) &= p(x_{i:n} \mid z_{i-1} = k) = \sum_{l=1}^K p(x_{i:n}, z_i = l \mid z_{i-1} = k), \\ &= \sum_{l=1}^K p(x_i, x_{(i+1):n}, z_i = l \mid z_{i-1} = k), \\ &= \sum_{l=1}^K p(x_{(i+1):n} \mid z_i = l, \cancel{z_{i-1} = k}, \cancel{x_i}) p(z_i = l, x_i \mid z_{i-1} = k), \\ &= \sum_{l=1}^K p(x_{(i+1):n} \mid z_i = l) p(x_i \mid z_i = l, \cancel{z_{i-1} = k}) p(z_i = l \mid z_{i-1} = k), \\ &= \sum_{l=1}^K \beta_i(l) \Psi_i(k) A_{kl}, \\ &= \sum_{l=1}^K \beta_i(l) A_{kl} \Psi_i(k). \\ \implies \beta_{i-1} &= A(\Psi_i \odot \beta_i).\end{aligned}$$

## Proof for the one-slice smoothed marginal $\tau_{ik}$

We previously saw Slide 129 that

$$\begin{aligned}\tau_{ik} &= p(z_i = k \mid \mathbf{X}), \\ &= \frac{p(z_i = k, x_{1:n})}{p(x_{1:n})}, \\ &= \frac{\overbrace{p(x_{(i+1):n} \mid z_i = k)}^{\beta_i(k)} \overbrace{p(x_{1:i}, z_i = k)}^{\alpha_i(k)}}{p(x_{1:n})}, \\ &\propto \alpha_i(k)\beta_i(k).\end{aligned}\tag{Slide 129}$$

In addition, we get that the normalization factor (*i.e.* the observed likelihood) is

$$p(x_{1:n}) = \sum_l \alpha_i(l)\beta_i(l), \quad \text{at any time step } i = 1, \dots, n$$

## Proof for the two-slice smoothed marginal $\xi_{i,k,l}$

Using the HMM conditional independencies we can simplify

$$\begin{aligned}\xi_{i,k,l} &= p(z_i = k, z_{i+1} = l \mid x_{1:n}) = \frac{p(x_{1:n}, z_i = k, z_{i+1} = l)}{p(x_{1:n})}, \\ &\propto p(x_{1:n}, z_i = k, z_{i+1} = l), \\ &\propto p(x_{1:i} \mid z_i = k, \cancel{z_{i+1} = l}, \cancel{x_{(i+1):n}}) p(z_i = k, z_{i+1} = l, x_{(i+1):n}), \\ &\propto p(x_{1:i} \mid z_i = k) p(z_i = k, z_{i+1} = l, x_{i+1}, x_{(i+2):n}), \\ &\propto p(x_{1:i} \mid z_i = k) p(x_{(i+2):n} \mid z_{i+1} = l, \cancel{x_{i+1}}, \cancel{z_i = k}) p(z_i = k, z_{i+1} = l, x_{i+1}), \\ &\propto p(x_{1:i} \mid z_i = k) p(x_{(i+2):n} \mid z_{i+1} = l) p(x_{i+1} \mid z_{i+1} = l, \cancel{z_i = k}) p(z_{i+1} = l \mid z_i = k) p(z_i = k), \\ &\propto p(x_{1:i} \mid z_i = k) \beta_{i+1}(l) \Psi_{i+1}(l) A_{kl} p(z_i = k), \\ &\propto p(x_{1:i}, z_i = k) \beta_{i+1}(l) \Psi_{i+1}(l) A_{kl}, \\ &\propto \alpha_i(k) \beta_{i+1}(l) \Psi_{i+1}(l) A_{kl}.\end{aligned}$$

# Additional properties of the forward-backward messages

## Computational complexity

The FB procedure is in  $\mathcal{O}(nK^2)$

In addition to  $\tau_{ik}$  and  $\xi_{i,k,l}$

- The observed likelihood can be computed in two equivalent ways:

- 1 with a single forward pass as  $p_{\theta}(x_{1:n}) = \sum_l \alpha_n(k)$
- 2 at any step  $i$ :  $p_{\theta}(x_{1:n}) = \sum_k \alpha_i(k)\beta_i(k)$

Using **1** is called a *forward* algorithm.

- The *filtered* marginal at step  $i$  is

$$p(z_i = j \mid x_{1:i}) = \alpha_i(k) / \sum_l \alpha_i(l)$$

## Some remarks on forward-backward

Not complicated to implement but

- 1 **Careful with indices**, notations easily get mixed up
- 2 **Numerical error**: code in log-space  $\log \alpha$ ,  $\log \tau$  and  $\log \xi$  with the "log-sum-exp trick" for computing the normalizing constant. An example

$$\log \alpha_i = \log \Psi_i + \log A^\top \alpha_{i-1} - cte_i$$

with  $cte_i := \log \sum_k e^{\log \alpha_i(k)}$ . When computing  $cte_i$ , we use

$$\log \sum_k e^{y_k} = m^* + \log \underbrace{\sum_k e^{y_k - m^*}}_{\geq 1},$$

with  $y_k = \log \alpha_i(k)$  to ensure there is at least one  $e^0 = 1$  in the sum for numerical stability.

# M-step

Assume  $\tau_{ik}^{(t)}$  and  $\xi_{i,k,l}^{(t)}$  have been computed by FB recursion (E-step). We need to solve

$$\theta^{(t)} \in \arg \max_{\theta=(\nu, A, \gamma)} \left\{ f_t(\theta) := \mathbb{E}_{\mathbf{Z} \sim p_{\theta^{(t-1)}}} [\log p_{\theta}(\mathbf{X}, \mathbf{Z})] \right\}.$$

With

$$f_t(\theta) = \underbrace{\sum_{k=1}^K \tau_{1k}^{(t)} \log \nu_k + \sum_{k,l=1}^K \sum_{i=2}^n \xi_{i,k,l}^{(t)} \log A_{k,l}}_{\text{Markov part}} + \underbrace{\sum_{k=1}^K \sum_{i=1}^n \tau_{ik}^{(t)} \log p_{\gamma_k}(x_i)}_{\text{Emission part}},$$

and constraints

$$\sum_{k=1}^K \nu_k = 1 \quad \text{and} \quad \sum_{l=1}^K A_{kl} = 1, \quad \forall l = 1, \dots, K \quad \text{and} \quad \gamma_k \in \Gamma$$

## M-step for the Markov Chain part

Introducing Lagrange multipliers  $\lambda_0, \dots, \lambda_K$  associated to the  $K + 1$  equality constraints we seek stationary points of

$$\begin{aligned}\mathcal{L}(\nu, A; \lambda) &= \sum_{k=1}^K \tau_{1k}^{(t)} \log \nu_k + \sum_{k,l=1}^K \sum_{i=2}^n \xi_{i,k,l}^{(t)} \log A_{k,l} \\ &+ \lambda_0 \left(1 - \sum_k \nu_k\right) + \sum_k \lambda_k \left(1 - \sum_l A_{kl}\right).\end{aligned}$$

This leads for  $\forall k, l \in \llbracket 1, K \rrbracket$ :

$$\hat{\nu}_k^{(t)} = \frac{\tau_{1k}^{(t)}}{\lambda_0}, \quad \hat{A}_{kl} = \frac{\sum_{i=1}^{n-1} \xi_{i,k,l}^{(t)}}{\lambda_k}.$$

Injecting into the  $K + 1$  constraints we get the Lagrange multipliers

- $\lambda_0 = \sum_k \tau_{1k}^{(t)} = 1$
- $\forall k = 1, \dots, K, \quad \lambda_k = \sum_{i=1}^{n-1} \sum_{l=1}^K \xi_{i,k,l}^{(t)} = \sum_{i=2}^n \tau_{ik}^{(t)}.$

## M-step for the emission model part

Obviously dependent on the emission model  $p_{\gamma_k}$

# M-step for the emission model part

Obviously dependent on the emission model  $p_{\gamma_k}$

Still, there are 2 interesting cases we can think about

- 1 *Discrete emissions*  $x_i \in \{1, \dots, V\}$  and  $x_i | \{z_{ik} = 1\} \sim \mathcal{M}_V(1, \gamma_k)$  with each  $\gamma_k$  a probability vector over  $V$  outcomes. Minimizing the Lagrangian accounting for  $\sum_v \gamma_v = 1$ , we then have

$$\hat{\gamma}_{kv} = \sum_{i=1}^n \tau_{ik} x_{iv} / \tilde{n}_k, \quad \text{with: } \tilde{n}_k = \sum_{i=1}^n \tau_{ik}.$$

# M-step for the emission model part

Obviously dependent on the emission model  $p_{\gamma_k}$

Still, there are 2 interesting cases we can think about

- 1 *Discrete emissions*  $x_i \in \{1, \dots, V\}$  and  $x_i | \{z_{ik} = 1\} \sim \mathcal{M}_V(1, \gamma_k)$  with each  $\gamma_k$  a probability vector over  $V$  outcomes. Minimizing the Lagrangian accounting for  $\sum_v \gamma_v = 1$ , we then have

$$\hat{\gamma}_{kv} = \sum_{i=1}^n \tau_{ik} x_{iv} / \tilde{n}_k, \quad \text{with: } \tilde{n}_k = \sum_{i=1}^n \tau_{ik}.$$

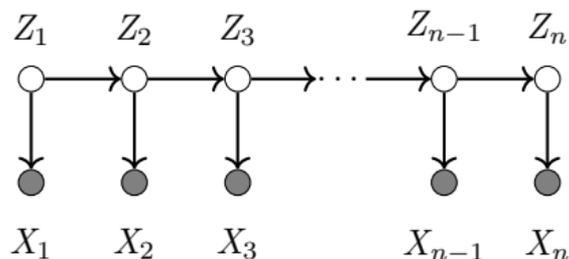
- 2 *Exponential family* if  $\log p_{\eta_k}(x_i) = \eta_k^\top T_k(x_i) - a_k(\eta_k) - b_k(x_i)$ , then we seek to solve this implicit equation in  $\eta_k$

$$\nabla a(\eta_k) = \frac{\sum_{i=1}^n \tau_{ik} T_k(x_i)}{\tilde{n}_k}$$

1 is a particular case since  $\mathcal{M}_V(1, \gamma)$  can be cast in the exponential family. Its minimal form involves  $\eta = \log \gamma / \gamma_V$ ,  $a(\eta) = \log \sum_v e^{\eta v}$  and  $T(x) = x$ . Notice that  $\nabla a(\eta) = \text{softmax}(\eta) = \gamma$ .

## Final comment: prediction of $Z_{i+1} \mid Z_i, X_{1:(i+1)}$

Recall the DAG



Hence, in a HMM we have that for all  $k$ :

$$p(z_{i+1} = l \mid z_i = k, X_{1:(i+1)}) = p(z_{i+1} = l \mid z_i = k, X_{\cancel{1:(i+1)}}), \quad (\text{HMM})$$

$$\propto p(X_{(i+1)} \mid \cancel{z_i = k}, z_{i+1} = l) p(z_{i+1} = l \mid z_i = k) \quad (\text{Bayes})$$

$$\propto p_{\gamma_l}(x_{i+1}) A_{kl},$$

$$= \frac{p_{\gamma_l}(x_{i+1}) A_{kl}}{\sum_l p_{\gamma_l}(x_{i+1}) A_{kl}}.$$

$\rightsquigarrow (Z_{1:n} \mid X_{1:n})$  is an inhomogeneous MC with the transition probability at step  $i$  that are biased according to the likelihood of the data under the arrival state  $\Psi_{i+1}(l)$

# Probabilistic learning and Gaussian processes

# Aproximation schemes for probabilistic inference

# Multi-task Gaussian processes and clustering

**Stochastic Block Model or project on  
real data (your choice)**