Unsupervised Learning with discrete latent variable models

Nicolas Jouvin
nicolas.jouvin@inrae.fr
https://nicolasjouvin.github.io/

M2 Data-Science 2024-2025

Organization

Thursdays 8h30 - 11h45, this room.

$6 \times$ 3h classes

1h30 class + 1h30 practical session (except today)

Important: you need one computer/person for practical sessions

Evaluation

1. CC: assiduity, practical session
2. Final exam on Friday 12th January, 2024
3. "Max exam" : the final not will be $\max(\text{Exam}, \text{mean}(\text{Exam}, \text{CC}))$

Homework: no grade given (but I'll still correct it). The reward is : one exercise of the final exam will be "close".

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

Some relevant lecture/slides on the topic for a different point-of-view (⚠notations)

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

Introduction

Types of statistical learning

Supervised

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 (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...

(Discrete) latent variables models for unsupervised learning

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

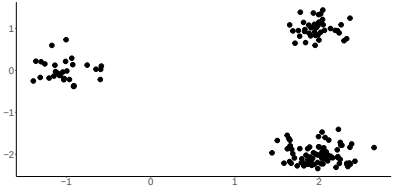
What this course is about...

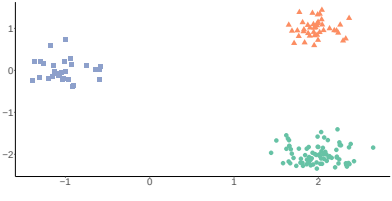
(Discrete) latent variables models for unsupervised learning

$\rightsquigarrow$ 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





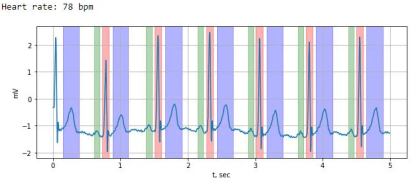
What this course is about...

(Discrete) latent variables models for unsupervised learning

$\rightsquigarrow$ 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



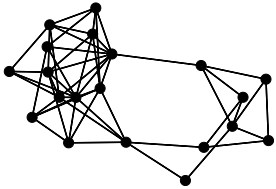
Example of ECG annotation, source: `https://medium.com/data-analysis-center/56f8b9abd83a`

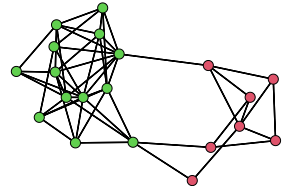
What this course is about...

(Discrete) latent variables models for unsupervised learning

$\rightsquigarrow$ 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)

Course outline

1. Fundamentals of Bayesian statistics
2. Clustering with mixture models
3. Inference in latent variable models: the EM algorithm
4. Hidden Markov Models (HMMs)
5. Stochastic Block Model: an introduction to variational inference
6. Conclusion of the course

Fundamentals of Bayesian statistics

Bayes formula

Frequentist inference

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

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 $\theta^\star$. [1]

Maximum-likelihood estimation

Find the model, hence θ, that maximizes the probability of having seen the data

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

[1] 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)

The Bayesian paradigm

Maximum-likelihood and frequentist statistics produces *point estimates*

Paradigm shift: random parameters

Parameters θ are no longer treated as deterministic but as *random* quantities. The *prior* distribution, denoted as $\pi(\theta)$, encodes knowledge & uncertainty we have on the parameters **before** seeing new data.
$\rightsquigarrow$ 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 $\pi(\theta)$, we posit an observational model on $X \mid \theta \iff$ the likelihood. Bayesian modelization essentially adds one layer to frequentist models : the prior.

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

The posterior

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

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{Z}, \qquad \text{(Bayes formula)}$$

Here, $Z = \int_\Theta p(x \mid \theta)\pi(\theta)\,\mathrm{d}\theta$ is a normalization constant, independent of θ. Thus, it is common to write[a]

$$\pi(\theta \mid x) \propto p(x \mid \theta)\pi(\theta)$$

[a] Although computing this normalization constant is generally a challenging task in Bayesian statistics.

Choosing a prior

Expert knowledge

The prior π may be used to represent any available expert knowledge on θ.

Conjugate priors

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

Conjugate priors are widely used as they greatly simplify computations.

Uninformative prior

When the prior equally charges Θ we say that the prior is *uninformative*, noted $\pi(\theta) \propto 1$. Obviously, $\pi \propto 1$ 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 $\pi(\theta \mid x)$ and the prior is *improper*.

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

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

Observational model: the likelihood
Given a probability of tail θ, we model the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ as *i.i.d.* Bernoulli $X_i \sim Ber(\theta)$ so that

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

Choice of a prior
We use Beta distribution with support $\Theta = [0, 1]$

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

a and b are called *hyper-parameters* 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)

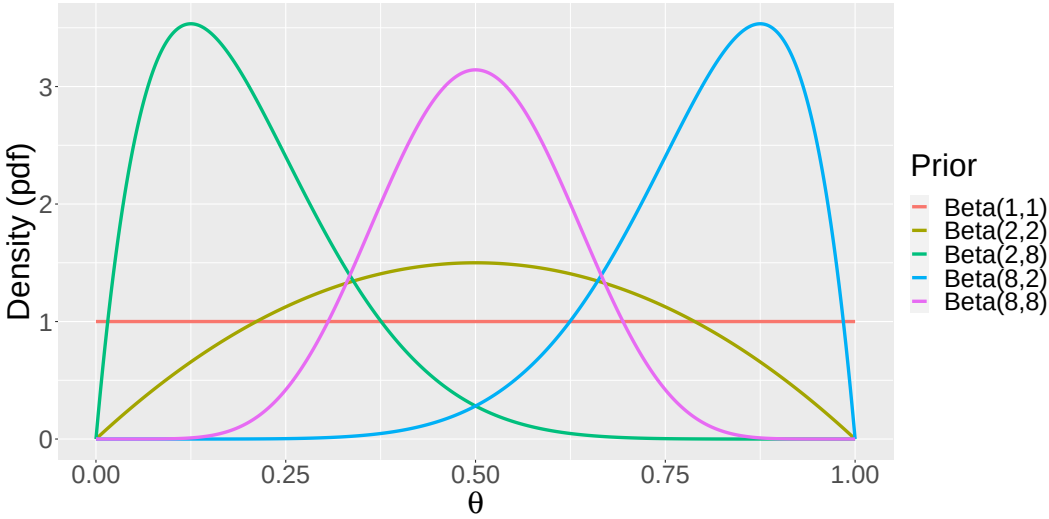


Figure: Graph of the p.d.f. $Beta(\cdot \mid a, b)$ for different values of a and b.

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

We seek to derive the posterior, and we directly have

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

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

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

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.

Bayesian decision theory

Bayesian point estimates

Having derived the posterior: how do we provide point estimates $\hat{\theta}$?

Cost function

A *cost function* is a function $C : \Theta \times \Theta \in \mathbb{R}_+$ where $C(\eta, \theta)$ is the "cost of predicting η for a parameter θ. Some examples

- $C(\eta, \theta) = (\eta - \theta)^p$ (L^p-loss)
- $C(\eta, \theta) = \mathbf{1}_{\eta \neq \theta}$ (0-1 loss)

Bayesian estimator

Remember that θ is random. For a given model and observation x, the Bayesian estimator is the one that minimizes the average cost under the posterior distribution:

$$\hat{\theta} \in \arg\min_{\eta} \left\{ \mathbb{E}_{\theta \sim \pi(\cdot|x)} \left[C(\eta, \theta) \right] = \int_{\Theta} C(\eta, \theta) \pi(\theta \mid x) \, \mathrm{d}\theta \right\}. \qquad \text{(Bayes estimator)}$$

Posterior Mean, Median & Mode

Different cost functions leads to different Bayes estimator among which

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

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

Credibility regions

The posterior may also be used for uncertainty quantification by computing regions $\mathcal{R} \subset \Theta$ s.t. $\pi(\theta \in \mathcal{R} \mid x) = \int_{\mathcal{R}} \pi(\theta \mid x)\, \mathrm{d}\theta = 1 - \alpha$

Latent variable models

Incomplete data models

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

- $\boldsymbol{X}$ observed random variables
- $\boldsymbol{Z}$ unobserved (latent/hidden) variables
- θ unknown parameters

An attempt at defining latent variables (creds. to S. Robin)

- Frequentist setting:

$$\text{latent variables} = \text{random but unobserved}, \quad \text{parameters} = \text{fixed}$$

- Bayesian setting:

$$\text{both latent variables and parameters} = \text{random}$$

but

$$\#\text{ latent variable} \simeq \#\text{ data}, \quad \#\text{ parameters} \ll \#\text{ data}$$

Different types of likelihoods

In this course, we place ourselves in the frequentist setting, using MLE inference. Although Bayesian extension of the proposed models are common.

Complete data likelihood

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

$$p_\theta(\boldsymbol{X}, \boldsymbol{Z}) = p_\theta(\boldsymbol{X} \mid \boldsymbol{Z}) p_\theta(\boldsymbol{Z}).$$

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

Observed data likelihood

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

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

$\rightsquigarrow$ only involves the observed $\boldsymbol{X}$, but not always tractable.

[a] When $\mathcal{Z}$ is discrete, replace $\int$ by $\sum$

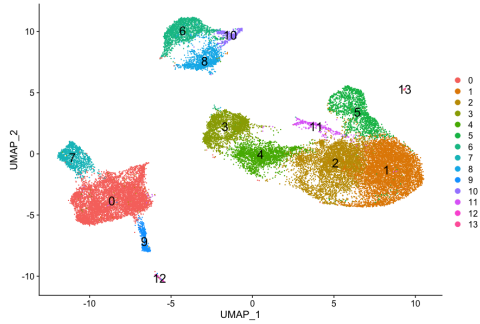
Clustering with mixture models

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*[2]

[2] as opposed to classification in a supervised context

Mathematical context

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

Definition: partition

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

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

Alternative encoding of the partition

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

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

The set $\boldsymbol{Z} = \{z_1, \ldots, z_n\}$ represents a partition of $\{1, \ldots, n\}$. 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 $\boldsymbol{Z}$ as a latent variable and posit a generative model $p_\theta(\boldsymbol{X}, \boldsymbol{Z})$
$\rightsquigarrow$ Clustering becomes an inference problem of finding $\hat{\boldsymbol{Z}}$.

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

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

- Good news: discrete problem $\rightsquigarrow$ there exists an optimum $\boldsymbol{C}^\star$.
- **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 $\mu_1, \ldots, \mu_K$ at random among the sample $\boldsymbol{X}$ 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

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

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)

$\rightsquigarrow$ In practice, we try several restarts with different random inits.

Extensions

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

- Choose μ_1 uniformly among the sample
- then sequentially do for each $k = 2, \ldots, K$
  - compute weight $w_i := \min_{j<k} \|x_i - \mu_j\|_2^2$
  - Choose μ_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 $\phi : \mathcal{X} \to \mathcal{H}$ a *feature map*.

Mixture models

Probabilistic view on clustering

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

Denote $\pi = (\pi_1, \ldots, \pi_K)$ 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 $i = 1, \ldots, n$, 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. $X_i \mid \{z_{ik} = 1\} \sim p_{\gamma_k}$

The model parameters are $\theta = \{\pi_k, \gamma_k\}_{k=1}^K$ and p_γ 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_γ

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

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

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

Complete likelihood

Properties: conditional independence

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

Complete log-likelihood

$$\log p_\theta(\boldsymbol{X}, \boldsymbol{Z}) = \log p_\theta(\boldsymbol{Z}) + \log p_\theta(\boldsymbol{X} \mid \boldsymbol{Z}) = \sum_{i=1}^{n} \log p_\pi(Z_i) + \log p_\gamma(X_i \mid Z_i),$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} Z_{ik} \left[\log \pi_k + \log p_{\gamma_k}(X_i)\right].$$

Posterior distribution of $Z \mid X$

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

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

Notice that τ_i also depends on the parameters θ.

A note on identifiability

Definition: identifiability

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

Intuition: the labels of the clusters $1, \ldots, K$ 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 σ be a permutation of $[\![1, K]\!]$, then for a mixture model with parameters π, γ we have

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

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

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

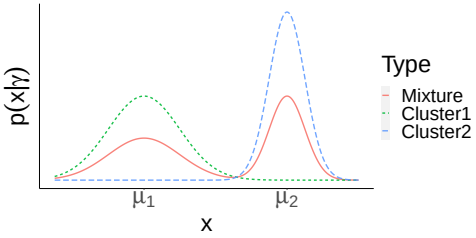
$\rightsquigarrow$ can cause problems in Bayesian inference procedure since the posterior is highly multimodal.

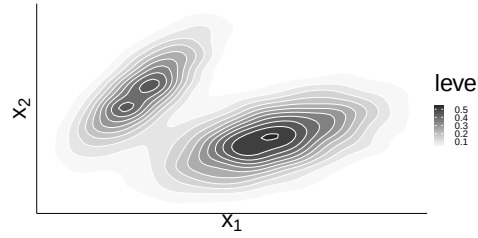
Gaussian Mixture Models (GMM)

Continuous data: $\boldsymbol{x} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$

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

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

$\rightsquigarrow$ **But we do not observe the $\boldsymbol{Z}$!**

Maximum-likelihood estimation (cont'd)

A chicken-and-egg problem

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

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

2. If we knew $\theta^\star$, we could find the best estimate of $\boldsymbol{Z}$ 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

Quizz ! 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

Quizz ! 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 $\phi : \mathbb{R}^d \to \mathbb{R}$ a convex function, then

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

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

Proof:

- ϕ 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 $\mathbb{E}_Z$

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

Entropy of a random variable

Definition: entropy

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

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

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 μ such that $\mathrm{d}Q = q\,\mathrm{d}\mu$ then we can define

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

  Now depends on the base measure μ.

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

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

Properties

- $\mathrm{KL}(p \parallel q) \geq 0$ 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 μ such that $\mathrm{d}P = p\,\mathrm{d}\mu$ and $\mathrm{d}Q = q\,\mathrm{d}\mu$, then

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

  $\rightsquigarrow$ invariant w.r.t. the choice of (p, q, μ) since the ratio $\mathrm{d}P/\,\mathrm{d}Q$ is invariant.

The evidence lower bound (ELBO)

Minorizer of the observed-likelihood

Evidence lower bound

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

$$\log p_\theta(\boldsymbol{X}) \geq \mathcal{L}(q, \theta) \coloneqq \mathbb{E}_q \left[\log p_\theta(X, Z)\right] + \mathcal{H}(q). \qquad \text{(ELBO)}$$

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

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

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

Comments

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

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

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

EM: a universal algorithm for latent variables

Intuition: chicken-and-egg

1. *if we knew* $\boldsymbol{Z}$, we could easily work with $f(\theta) = \log p_\theta(\boldsymbol{X}, \boldsymbol{Z})$
2. *if we knew* θ, the best representation of $\boldsymbol{Z}$ is via its posterior $p_\theta(\boldsymbol{Z} \mid \boldsymbol{X})$

Expectation-Maximization algorithm

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

Expectation step

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

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

It involves (generalized) moments of $\boldsymbol{Z}$ 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 ϵ

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

Rewriting EM: coordinate ascent on the ELBO

EM algorithm (equivalent formulation)

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

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

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

- E-step is equivalent to $\min_q \mathrm{KL}(q \,\|\, p_{\theta^{(t+1)}}(\cdot \mid X)) \implies q^{(t+1)} = p_{\theta^{(t+1)}}(\cdot \mid X)$
- 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

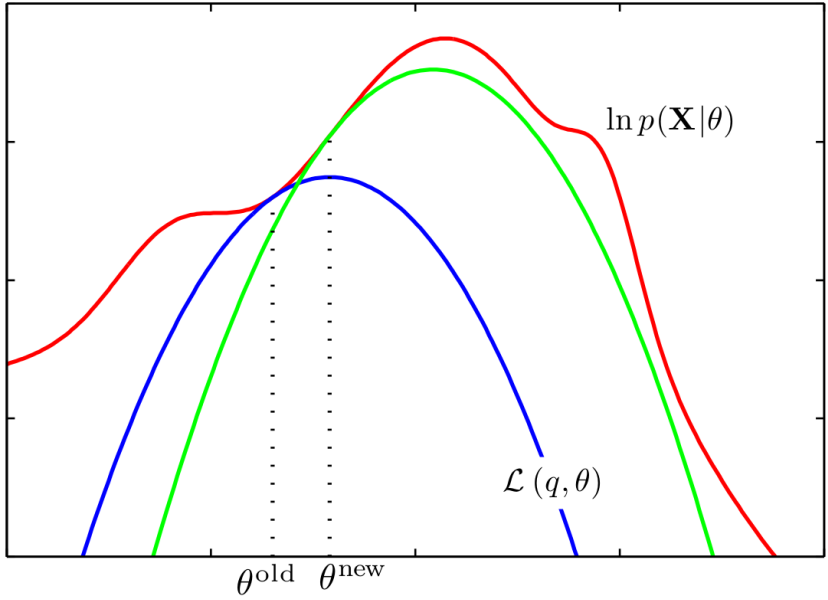
$$\forall t \geq 0, \quad \log p_{\theta^{(t+1)}}(\boldsymbol{X}) \geq \log p_{\theta^{(t)}}(\boldsymbol{X})$$

Proof:

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

- 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.

A graphical illustration of EM algorithm (cred: G. Obozinski)



Source: https://www.iro.umontreal.ca/~slacoste/teaching/MVA_GM/fall2015/cours_MVA_EM.pdf

Illustration with Gaussian mixture

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

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

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

E-step for GMM

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

As seen previously, the posterior for mixture model always writes

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

So that

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

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

M-step for GMM

Solve

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

For GMM, the updates are

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

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

Link 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 $\Sigma_k = \sigma^2 \operatorname{Id}_d$. Dropping the known quantities, the criterion is

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

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 $\boldsymbol{X}$

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.

$\rightsquigarrow$ 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 $\hat{\theta}_K = \arg\max_{\theta \in \Theta_K} \log p_\theta(\boldsymbol{X})$. A *penalized likelihood* estimate of K is given by

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

Different penalties leads to different criterion

Definitions: AIC, BIC, ICL

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

$$\begin{aligned} AIC(K) &\coloneqq \log p_{\hat{\theta}_K}(\boldsymbol{X}) - d_K, \\ BIC(K) &\coloneqq \log p_{\hat{\theta}_K}(\boldsymbol{X}) - \frac{d_K}{2}\log(n), \\ ICL(K) &\coloneqq \mathbb{E}_{Z\sim p_{\hat{\theta}_K}(\cdot|\boldsymbol{X})}\left[\log p_{\hat{\theta}_K}(\boldsymbol{X}, \boldsymbol{Z})\right] - \frac{d_K}{2}\log(n) \end{aligned}$$

Note: the ELBO property gives

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

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 $\boldsymbol{X}$.

Focus on BIC: Bayesian information criterion

Put a prior $p(K)$ on K, and the model: $p(\theta \mid K)$ and $p(\boldsymbol{X} \mid \theta)$. Bayes rule suggests choosing

$$\begin{aligned}\hat{K} &= \arg\max_K \left\{ p(K \mid \boldsymbol{X}) \propto p(K) p(\boldsymbol{X} \mid \theta) \right\}, \\ &= \arg\max_K \log p(K) + \log p(\boldsymbol{X} \mid K), \\ &= \arg\max_K \log p(K) + \log \int p(\boldsymbol{X} \mid \theta, K) p(\theta \mid K) \, \mathrm{d}\theta.\end{aligned}$$

Dropping the prior term $\log p(K)$ which is constant with n, we need to compute the integral in the second term $\rightsquigarrow$ difficult in general !

Under regularity assumptions (see Lebarbier et al. 2004, for details), we have

$$\log p(\boldsymbol{X} \mid K) = \log p_{\hat{\theta}_K}(\boldsymbol{X}) - \frac{d_K}{2} \log(n) + \mathcal{O}_P(1).$$

This justifies the formula of BIC.

Hidden Markov Models (HMMs)

Motivations

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

- time series
- genomic data: observations collected at precise locations in the genome
- etc.

$\rightsquigarrow$ it is likely that "past" influences the "future".

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

Motivations

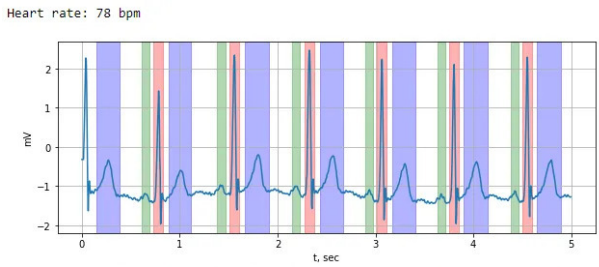
What if observations $\boldsymbol{X} = \{x_i\}_i$ are ordered ? *e.g.*

- time series
- genomic data: observations collected at precise locations in the genome
- etc.

$\rightsquigarrow$ it is likely that "past" influences the "future".

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

Example 1: time series segmentation



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

Motivations

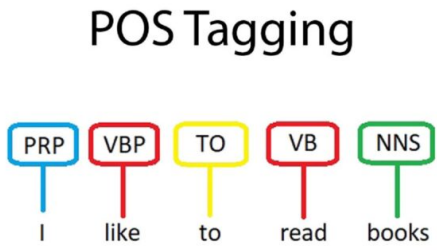
What if observations $\boldsymbol{X} = \{x_i\}_i$ are ordered ? *e.g.*

- time series
- genomic data: observations collected at precise locations in the genome
- etc.

$\rightsquigarrow$ 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



Source: https://byteiota.com/pos-tagging/

Motivations

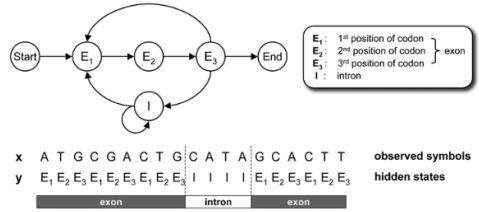
What if observations $\boldsymbol{X} = \{x_i\}_i$ are ordered ? *e.g.*

- time series
- genomic data: observations collected at precise locations in the genome
- etc.

$\rightsquigarrow$ 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[3] 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

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

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.

[3] "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 $p(y_{i+1} \mid y_i)$ is independent time (of i).

Initial distribution

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

Marginal distribution

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

Transition matrix

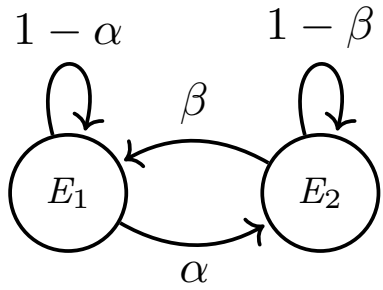
We denote A the $K \times K$ matrix with $A_{kl} = p(z_{i+1} = l \mid z_i = k)$ and properties:

- 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 $m, n \in \mathbb{N}$, $p(y_{n+m} = l \mid y_m = k) = A_{kl}^{(m)}$ (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

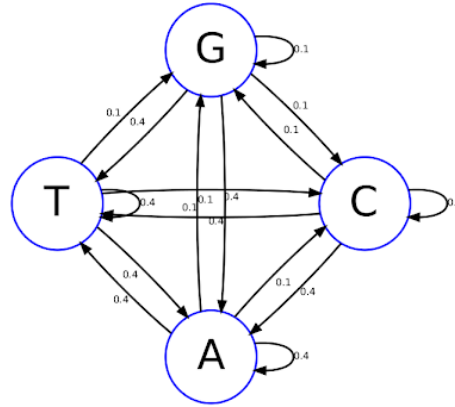
$$A = \begin{array}{c} \\ E_1 \\ E_2 \end{array} \begin{array}{c} \begin{array}{cc} E_1 & E_2 \end{array} \\ \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \end{array}$$



Graphical representation of a 2-state homogeneous Markov chain

A second example: modeling nucleotide transition

$$A = \begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} & & & \\ & & p(y_t{=}G|y_{t-1}{=}T) & \\ & & & \end{pmatrix} \end{array} \rightarrow$$



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

A third example: 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) :
  $\rightarrow$ State $=$ number of balls in this urn

A third example: 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) :
  $\rightarrow$ State $=$ number of balls in this urn

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

Stationary distribution & how to find them

Stationary distribution

Let A be a transition matrix over $[\![1, K]\!]$, we say that a vector π 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. π 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 $q \in \mathbb{N}^\star$ such that $A^{(q)} > 0$ then
   - π is unique and $\pi_k > 0$.
   - $p(y_n = l \mid y_1 = k) = A^{(n)}_{kl} \xrightarrow[n \to +\infty]{} \pi_l$, whatever the initial distribution ν is.

   Such chains "forget their past" after enough steps.

Computing the stationary distribution

First strategy: eigenvector

We know that $A^\top \pi = 1 \cdot \pi$, so that π is an eigenvector associated to the unit[a] 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)

[a] Recall that eigenvalues (but not eigenvectors) of A and $A^\top$ are the same.

Second strategy: linear system

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

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

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

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

Question (at home): 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 `> (Pi - t(Pi)%*%A) < 1e-15`

Hidden Markov Models (HMMs)

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 $\mid (z_i)_i$ and for all $i \in [\![1, n]\!]$, $x_i \mid \{z_{ik} = 1\} \sim p_{\gamma_k}(\cdot)$

The model parameters are $\theta = (\nu, A, \gamma)$ and $p(x_i \mid z_i = k) = p_{\gamma_k}(x_i)$ are called *emission probability*

Marginal likelihood of x_i

Denote $\nu_i = (\nu_{i1}, \ldots, \nu_{iK})$, such that $\nu_{ik} = p_\theta(z_{ik} = 1)$ [a]. Then,

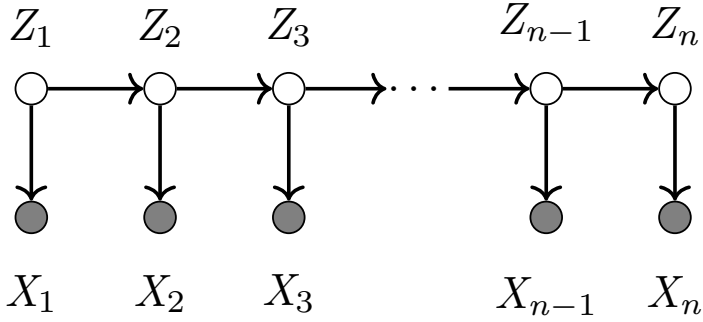
$$p_\theta(x_i) = \sum_k \nu_{ik} p_{\gamma_k}(x_i)$$

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

$\rightsquigarrow$ HMMs can be thought of as a generalization of mixture introducing dependency!

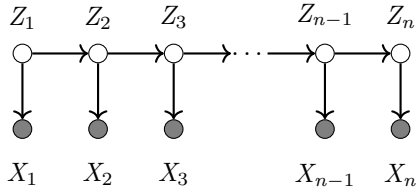
[a] For homogeneous MC we know that $\nu_i = \nu^\top A^{(i-1)}$.

Graphical model representation



- Empty circle ○ represents unobserved random variable
- Gray circles ● 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.

Complete-data likelihood

Complete-data log-likelihood for HMMs

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

Observed-data likelihood

Observed-data log-likelihood for HMMs

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

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

Posterior distribution

Denote

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

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, $p_\theta(\boldsymbol{Z} \mid \boldsymbol{X})$ does not factorizes over i anymore

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

3. $(z_i)_i$ are not independent | $(x_i)_i$ but rather $(z_{1:n})$ | $(x_{1:n})$ is an inhomogeneous Markov C

Posterior distribution

Denote

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

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, $p_\theta(\boldsymbol{Z} \mid \boldsymbol{X})$ does not factorizes over i anymore

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

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

$$\begin{aligned}
p_\theta(z_{i+1} \mid z_{1:i}, x_{1:n}) &= p_\theta(z_{i+1} \mid z_{1:i}, x_{(i+1):n}), && (z_{i+1} \perp\!\!\!\perp x_{1:i} \mid z_i) \\
&\propto p_\theta(x_{(i+1):n} \mid \cancel{z_{1:i}}, z_{i+1}) p_\theta(z_{i+1} \mid \cancel{z_{1:i}}), && \text{(Bayes + HMM)} \\
&\propto p_\theta(x_{(i+1):n}, z_{i+1} \mid z_i), \\
&= p_\theta(z_{i+1} \mid z_i, x_{(i+1):n}), \\
&= p_\theta(z_{i+1} \mid z_i, x_{1:n}). && (z_{i+1} \perp\!\!\!\perp x_{1:i} \mid z_i)
\end{aligned}$$

The "three" HMM problems

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

1. Given θ the model parameters, compute the probability of observing $x_{1:n}$ (*i.e.* the observed likelihood)
$$p_\theta(x_{1:n})$$
2. **Decoding** given θ the model parameters and observations $x_{1:n}$, find the most probable sequence of hidden states
$$\hat{z}_{1:n} = \arg\max_{z_{1:n}} p_\theta(z_{1:n} \mid x_{1:n})$$
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:** $p_\theta(z_{n+m} \mid x_{1:n})$ for $m \geq 1$
- **Filtering:** $p_\theta(z_i \mid x_{1:i})$
- **Smoothing:** $p_\theta(z_i \mid x_{1:n}) \neq$ filtering, notice the conditioning on all the evidence

Inference in HMMs

Reminder on MLE & EM

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

EM algorithm

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

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

$$\mathbb{E}_{\theta(t)}\left[\log p_\theta(\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X}\right] = \mathbb{E}_{\boldsymbol{Z} \sim p_{\theta^{(t)}}(\cdot \mid \boldsymbol{X})}\left[\log p_\theta(\boldsymbol{X}, \boldsymbol{Z})\right].$$

- **M-step:** update the estimate of θ with

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

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

In Slide 76 we derived the expression of $\log p_\theta(\boldsymbol{X}, \boldsymbol{Z})$, hence using linearity of $\mathbb{E}$ we get:

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

Where:

$$\begin{aligned}
\Psi_i(k) &\coloneqq p_{\gamma_k}(x_i), && \text{(Emission probability)} \\
\tau_{ik} &\coloneqq p_{\theta^{(t)}}(z_{ik} = 1 \mid \boldsymbol{X}) = \mathbb{E}\left[z_{ik} \mid \boldsymbol{X}\right], \\
\xi_{i,k,l} &\coloneqq p_{\theta^{(t)}}(z_{(i-1)k} = 1, z_{il} = 1 \mid \boldsymbol{X}) = \mathbb{E}\left[z_{(i-1)k} z_{il} \mid \boldsymbol{X}\right]
\end{aligned}$$

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 $p(z_i \mid \boldsymbol{X}) \neq p(z_i \mid x_i)$ 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[4] into two distinct parts - past and future - conditionally on z_i

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

[4] As opposed to Fleetwood Mac's famous song

The forward-backward algorithm

Proposition

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

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

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

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

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

Then the smoothed posteriors are obtained with

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

Proof of the forward recursion

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

Proof of the backward recursion

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

Proof for the one-slice smoothed marginal τ_{ik}

We previously saw Slide 82 that

$$\begin{aligned}
\tau_{ik} &= p(z_i = k \mid \boldsymbol{X}), \\
&= \frac{p\left(z_i = k, x_{1:n}\right)}{p(x_{1:n})}, \\
&= \frac{\overbrace{p\left(x_{(i+1):n} \mid z_i = k\right)}^{\beta_i(k)} \overbrace{p\left(x_{1:i}, z_i = k\right)}^{\alpha_i(k)}}{p(x_{1:n})}, \\
&\propto \alpha_i(k)\beta_i(k).
\end{aligned} \tag{Slide 82}$$

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

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

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

Using the HMM conditional independencies we can simplify

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

Additional properties of the forward-backward messages

Computational complexity

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

In addition to τ_{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 $cte_i \coloneqq \log \sum_k e^{\log \alpha_i(k)}$. When computing cte_i, we use

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

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

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

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

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

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

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

This leads for $\forall k, l \in [\![1, K]\!]$:

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

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

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

M-step for the emission model part

Obviously dependent on the emission model p_{γ_k}

M-step for the emission model part

Obviously dependent on the emission model p_{γ_k}

Still, there are 2 interesting cases we can think about

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

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

M-step for the emission model part

Obviously dependent on the emission model p_{γ_k}

Still, there are 2 interesting cases we can think about

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

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

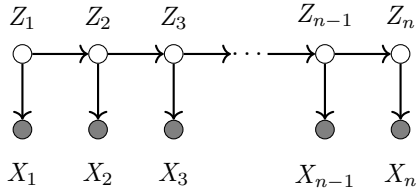
2. *Exponential family* if $\log p_{\eta_k}(x_i) = \eta_k^\top T_k(x_i) - a_k(\eta_k) - b_k(x_i)$, then we seek to solve this implicit equation in η_k

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

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 $\nabla a(\eta) = softmax(\eta) = \gamma$.

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:

$$
\begin{aligned}
p(z_{i+1} = l \mid z_i = k,\, X_{1:(i+1)}) &= p(z_{i+1} = l \mid z_i = k,\, X_{\cancel{1:}(i+1)}), && \text{(HMM)} \\
&\propto p(X_{(i+1)} \mid \cancel{z_i = k},\, z_{i+1} = l) p(z_{i+1} = l \mid z_i = k) && \text{(Bayes)} \\
&\propto p_{\gamma_l}(x_{i+1}) A_{kl}, \\
&= \frac{p_{\gamma_l}(x_{i+1}) A_{kl}}{\sum_l p_{\gamma_l}(x_{i+1}) A_{kl}}.
\end{aligned}
$$

$\rightsquigarrow$ $(Z_{1:n} \mid X_{1:n})$ 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 $\Psi_{i+1}(l)$

Decoding in HMMs: Viterbi algorithm

Reminder: the decoding problem

Context: parameters θ are given and fixed

Decoding is joint MAP estimation of $z_{1:n}$

$$\hat{z}_{1:n} = \arg\max_{z_{1:n}} p_\theta(z_{1:n} \mid X). \qquad \text{(Decoding problem)}$$

For HMM, decoding is solved via the *Viterbi algorithm*.

Warning this is different[5] from classification (marginal MAP)

$$\tilde{z}_i = \arg\max_{k=1,\dots,K} \left\{ \tau_{ik} = p_\theta(z_{ik} = 1 \mid X) \right\}.$$

N.B. classification can be solved via a forward-backward algorithm

[5] Except in mixture model where the two are equivalent since the joint posterior factorizes over i

Intuition for joint MAP

Fix a step i, we can define the quantity

$$V_i(k) \coloneqq \max_{z_1,\dots,z_{i-1}} p_\theta(z_{1:(i-1)}, z_i = k, x_{1:i})$$

Connection to decoding ?

$$p_\theta(Z \mid X) = \frac{p_\theta(Z, X)}{p_\theta(X)} \implies \arg\max_Z p_\theta(Z \mid X) = \arg\max_{k=1,\dots,K} V_{nk}.$$

What do we gain working with V_i ? $\rightsquigarrow$ **recursion**

$$V_i(k) = \Psi_i(k) \max_{l=1,\dots,K} A_{lk}\, V_{i-1}(l)$$

We will prove it later

Viterbi algorithm

Proposition

The most probable hidden state sequence can be computed by the following recursions

Forward $V_1(k) = \nu_{1k} p_{\gamma_k}(x_1)$ and for $i \geq 2$

$$V_i(k) = \Psi_i(k) \cdot \max_{l=1,\dots,K} V_{i-1}(l) A_{lk}, \qquad \text{(Store value)}$$

$$S_i(k) = \arg\max_{l=1,\dots,K} V_{i-1}(l) A_{lk} \Psi_i(k) \qquad \text{(Store best preceding state for going to } k \text{ at step } i\text{)}$$

Backtracking $\hat{z}_n = \arg\max_k V_{nk}$ and for all $2 \leq i \leq n$:

$$\hat{z}_{i-1} = S_i(\hat{z}_i). \qquad \text{(Backtracking)}$$

Proof of the forward recursion

$$
\begin{aligned}
V_i(k) &= \max_{z_{1:(i-1)}} p_\theta(z_{1:(i-1)}, z_i = k, x_{1:i}), \\
&= \max_{z_{1:(i-1)}} p_\theta(z_{1:(i-1)}, z_i = k, x_{1:(i-1)}, x_i), \\
&= \max_{z_{1:(i-1)}} p_\theta(x_i \mid z_i = k, \cancel{z_{1:(i-1)}}, \cancel{x_{1:(i-1)}}) p_\theta(z_i = k, z_{1:(i-1)}, x_{1:(i-1)}), \\
&= \max_{z_{i-1}} \max_{z_{1:(i-2)}} p_\theta(x_i \mid z_i = k)\, p_\theta(z_i = k \mid z_{1:(i-1)}, \cancel{x_{1:(i-1)}})\, p_\theta(z_{1:(i-2)}, z_{i-1}, x_{1:(i-1)}), \\
&= \max_{l=1,\dots,K} \max_{z_{1:(i-2)}} \Psi_i(k)\, p_\theta(z_i = k \mid z_{i-1} = l)\, p_\theta(z_{1:(i-2)}, z_{i-1} = l, x_{1:(i-1)}), \\
&= \max_{l=1,\dots,K} \Psi_i(k)\, A_{lk} \underbrace{\max_{z_{1:(i-2)}} p_\theta(z_{1:(i-2)}, z_{i-1} = l, x_{1:(i-1)})}_{V_{i-1}(l)}, \\
&= \Psi_i(k) \max_{l=1,\dots,K} A_{lk}\, V_{i-1}(l).
\end{aligned}
$$

Where we used the fact that all quantities in the products are ≥ 0 and that for $a \geq 0$, $\max_l a \times f(l) = a \times \max_l f(l)$

Proof of the backtracking step (i)

Computing V_i amounts to assign a score to the succession of optimal choices up to $i-1$ that leads to $z_i = k$ (conditionally on $x_{1:i}$).

The backward recursion traces back the succession of optimal choices to retrieve the whole optimal path.
The justification is based on the fact that, in HMMs:

$$p_\theta(x_{1:n}, z_{1:n}) = p_\theta(x_1, z_1) \prod_{i=2}^{n} p(x_i, z_i \mid z_{i-1}) = f_1(z_1) f_2(z_1, z_2) f_3(z_2, z_3) \dots f_n(z_{n-1}, z_n).$$

So that we can distribute $\max_{z_1,\dots,z_n}$ over the product of these positive functions

Proof of the backtracking step (ii)

$$
\begin{aligned}
\max_{z_{1:n}} p(x_{1:n}, z_{1:n}) &= \max_{z_n} \max_{z_{1:(n-1)}} p(x_{1:n}, z_{1:(n-1)}, z_n), \\
&= \max_{z_n} V_n(z_n) =: V_n(\hat{z}_n), && \text{(def of } \hat{z}_n\text{)} \\
&= \max_{z_{1:(n-1)}} p(z_{1:(n-1)}, \hat{z}_n, x_{1:(n-1)}, x_n), \\
&= \max_{z_{1:(n-1)}} p(\hat{z}_n, x_n \mid z_{n-1})\, p(z_{1:(n-1)}, x_{1:(n-1)}), && \text{(chain rule + HMM)} \\
&= \max_{z_{n-1}} \Psi_n(\hat{z}_n) A_{z_{n-1}\hat{z}_n} \underbrace{\max_{z_{1:(n-2)}} p(z_{1:(n-2)}, z_{n-1}, x_{1:(n-1)})}_{V_{n-1}(z_{n-1})}, \\
&= \underbrace{\Psi_n(\hat{z}_n) A_{\hat{z}_{n-1}\hat{z}_n}}_{p(x_n, \hat{z}_n \mid \hat{z}_{n-1})} V_{n-1}(\hat{z}_{n-1}), && \text{(def of } \hat{z}_{n-1} = S_n(\hat{z}_n)\text{)} \\
&= \dots \\
&= \left[\prod_{i=3}^{n} p(x_i, \hat{z}_i \mid \hat{z}_{i-1}) \right] \times \max_{z_1} p(x_1, z_1) A_{z_1 \hat{z}_2} \Psi_2(\hat{z}_2)
\end{aligned}
$$

Proof of the backtracking step (iii)

$$
\begin{aligned}
\max_{z_{1:n}} p(x_{1:n}, z_{1:n}) &= \dots \\
&= \left[\prod_{i=3}^{n} p(x_i, \hat{z}_i \mid \hat{z}_{i-1})\right] \times \max_{z_1} p(x_1, z_1) A_{z_1 \hat{z}_2} \Psi_2(\hat{z}_2) \\
&= \left[\prod_{i=3}^{n} p(x_i, \hat{z}_i \mid \hat{z}_{i-1})\right] \times p(x_1, \hat{z}_1) A_{\hat{z}_1 \hat{z}_2} \Psi_2(\hat{z}_2), \qquad \text{(def of } \hat{z}_1 = S_2(\hat{z}_2)\text{)} \\
&= p(x_1, \hat{z}_1) \prod_{i=2}^{n} p(x_i, \hat{z}_i \mid \hat{z}_{i-1}), \\
&= p(\hat{z}_{1:n}, x_{1:n}).
\end{aligned}
$$

So the backtracking step captures the posterior mode.

Some notes on Viterbi

At each step we must

- Storage complexity of Viterbi is in $\mathcal{O}(nK)$
- Computational complexity is in $\mathcal{O}(nK^2)$ as each $V_i(k)$ involves a maximum over K values.

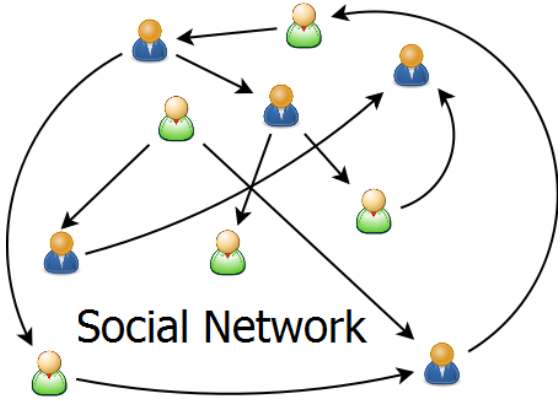
Analogy between $\max$ (Viterbi) and $\sum$ (FB).

The Viterbi algorithm is a particular case of the "max-product algorithm" for computing modes in DAG (see these course notes for details). It is exact when the DAG is a tree, which is the case in HMMs.

Stochastic Block Model: an introduction to variational inference

Qu'est qu'un réseau : quelques exemples concrets

Réseaux sociaux
Individus connectés par des relations de travail, d'amitiés , etc.

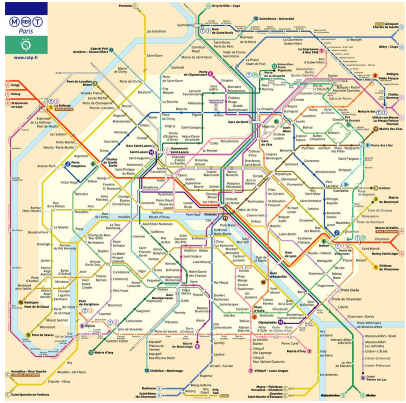


Source : Wikipédia, Zigomitros Athanasios

Qu'est qu'un réseau : quelques exemples concrets

Réseaux de transports

Villes connectées par des routes, gares connectée par des trains, etc.

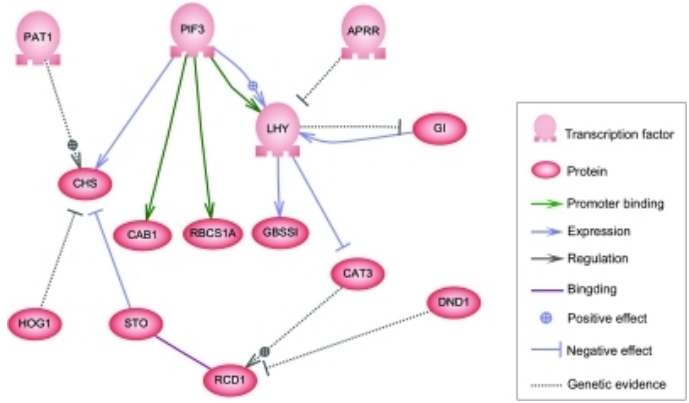


Source : RATP

Qu'est qu'un réseau : quelques exemples concrets

Réseaux de gènes

Gènes ou protéines intéragissant chimiquement entre elles pour réguler l'expression d'autres gènes.



Source : Wikipédia, réseaux de régulation chez le riz

Formalisation mathématique

Un **graphe**[6] est une structure générale qui encode des **liens** entre des **objets**

Vocabulaire
Dans le language de la théorie des graphes

- Objets $\leftrightarrow$ nœuds/sommet/vertex
- Liens $\leftrightarrow$ arrête/arcs/edge
- Le sens de la liaison peut être important (dirigé) ou non (non-dirigé)
- La liaison peut être binaire (présence/absence) ou pondérée

Définition: graphe
Un graphe $G = (V, E)$ est la donnée

- d'un ensemble V de $n = |V|$ noeuds
- d'un ensemble E de $m = |E|$ arrêtes

Définition très générale et flexible $\rightsquigarrow$ un même cadre général pour des problèmes très différents

[6] Ou réseau, l'un est plus commun à la communauté mathématiques, l'autre aux domaines "appliqués".

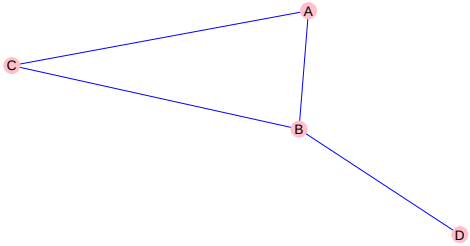
Adjacency matrix

There are several way to represent a graph.

Definition: adjacency matrix

The *adjacency* matrix of $G = (V, E)$ is the $n \times n$ matrix $\boldsymbol{X}$ defined as $X_{ij} = 1$ if edge $e = (i \to j) \in E$ and 0 elsewhere.
Properties:

- $\boldsymbol{X}$ is symmetric if the graph is undirected
- $|E| = m = \sum_{ij} X_{ij}$ (divided by 2 if G undirected).



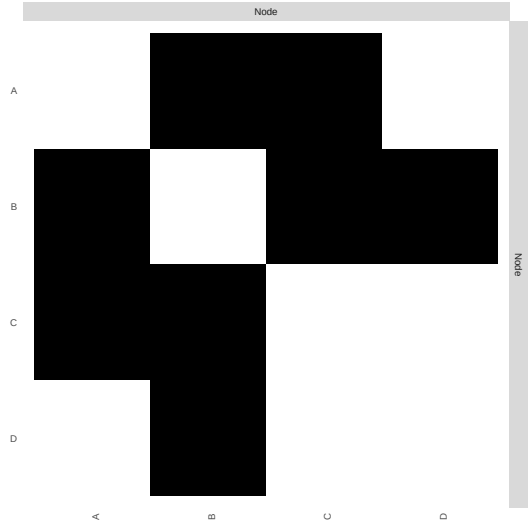
Adjacency matrix

	A	B	C	D
A	0	1	1	0
B	1	0	1	1
C	1	1	0	0
D	0	1	0	0

Dot plot representation

Adjacency matrix

	A	B	C	D
A	0	1	1	0
B	1	0	1	1
C	1	1	0	0
D	0	1	0	0



Random graph models

Random graphs: what for

- Define a probabilistic model of the interactions (i, j) in the graph. Many possibilities
  - statistical model on the adjacency matrix **(this course)**
  - combinatorics, statistical physics

- We want to do **inferential statistics**, answering questions *such as*
  - comparison with a "null model" where edge are uniformly random
  - Explain how local structures may govern the global one
  - Understand the process that generated an observed network

Applications : graph generation, community detection, link prediction, tests, model selection ...

A first random graph model: Erdös-Renyi (ER)

Model for n nodes,

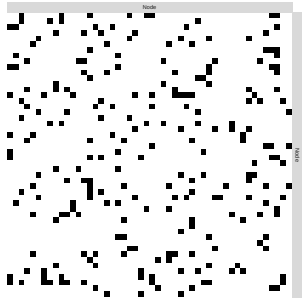
$$X_{ij} \underset{i.i.d.}{\sim} \mathcal{B}(p) \qquad \Longleftrightarrow \qquad \mathbb{P}(X_{ij} = 1) = p \qquad \text{(ER model)}$$

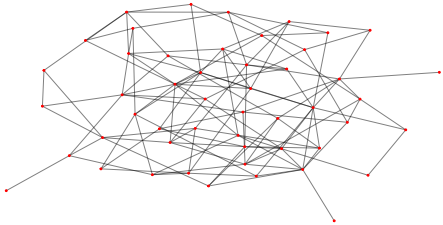
where $\mathcal{B}$ is Bernoulli and $p \in [0, 1]$ is the probability of connection.

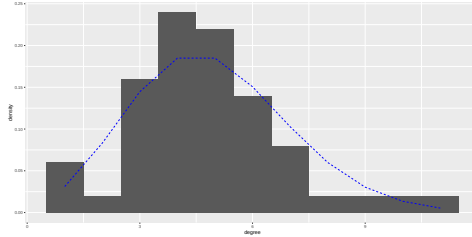
Properties of ER

- Model on the adjacency matrix $\boldsymbol{X}$
- Degree distribution $d_i := \sum_{j \neq i}^{n} X_{ij} \sim Binom(n-1, p)$
- MLE: $\hat{p} = \arg\max_p \log \prod_{i \neq j} \mathcal{B}(X_{ij} \mid p) = \frac{m}{n(n-1)}$ (density of the graph)

Realization of an ER with $n = 50$ nodes and $p = 0.1$







Stochastic block model (SBM)

Hypothesis: nodes belong to groups (clusters) and the probability of connection between node i and j only depends on the pair of cluster z_i and z_j.

Latent variable model of the adjacency matrix

1. $\forall i,\ z_i \overset{i.i.d}{\sim} \mathcal{M}_K(1, \pi)$
2. $\forall i \neq j,\ X_{ij} \mid \{z_{ik} z_{jl} = 1\} \sim \mathcal{B}(\gamma_{kl})$ (no self-loop)

Stochastic block model (SBM)

Hypothesis: nodes belong to groups (clusters) and the probability of connection between node i and j only depends on the pair of cluster z_i and z_j.

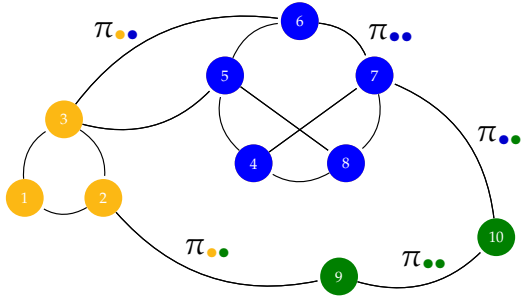
Latent variable model of the adjacency matrix

1. $\forall i,\ z_i \overset{i.i.d}{\sim} \mathcal{M}_K(1, \pi)$
2. $\forall i \neq j,\ X_{ij} \mid \{z_{ik} z_{jl} = 1\} \sim \mathcal{B}(\gamma_{kl})$ (no self-loop)

Remarks

- generative model: easy to sample from SBM
- many interesting real-world structure can emerge from it:
  1. *Communities:* $\gamma_{kk} \gg \gamma_{kl}$
  2. *Nestedness:* hierarchical structure in γ.
  3.
- easily generalize to weighted-graph by replacing Bernoulli with $p_{\gamma_{kl}}$
- *marginal of one edge*: $X_{ij} \sim \sum_{k,l=1}^{K} \pi_k \pi_l p_{\gamma_{kl}} \rightsquigarrow$ "mixture of Erdös-Renyi".

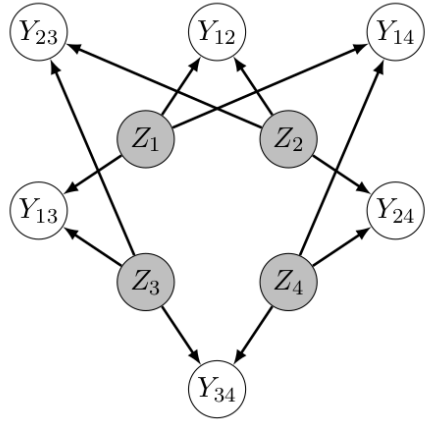
Illustration of SBM



Some example of SBM realizations

DAG representation of SBM

$(Z_i)_i$ are $\perp\!\!\!\perp$ and Y_{ij} only depends on (Z_i, Z_j) which give the DAG (for $n = 4$ nodes)



Source: S. Robin polycopié, notation change: replace Y by X

Remark: n latent variables (node-related) for n^2 observations (edges)

Likelihoods

Complete-data likelihood

$$
\begin{aligned}
p_\theta(\boldsymbol{X}, \boldsymbol{Z}) &= p_\theta(\boldsymbol{X} \mid \boldsymbol{Z}) \times p_\theta(\boldsymbol{Z}), \\
&= \prod_{i \neq j}^{n} p_\gamma(x_{ij} \mid z_i, z_j) \times \prod_{i=1}^{n} p_\pi(z_i), \\
&= \prod_{i \neq j}^{n} \prod_{k,l=1}^{K} p_{\gamma_{kl}}(x_{ij})^{z_{ik} z_{jl}} \times \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_{ik}},
\end{aligned}
$$

Observed-data likelihood (marginal of the whole adjacency matrix)

$$
\begin{aligned}
p_\theta(\boldsymbol{X}) &= \sum_{\boldsymbol{Z}} p_\theta(\boldsymbol{X}, \boldsymbol{Z}), \\
&= \sum_{z_1, \ldots, z_n} \left[\prod_{i \neq j}^{n} \prod_{k,l=1}^{K} p_{\gamma_{kl}}(x_{ij})^{z_{ik} z_{jl}} \times \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_{ik}} \right].
\end{aligned}
$$

Posterior dependencies: DAG moralization

$$p(A, B, C) = p(A)p(B)p(C|A, B) \qquad p(A, B|C) = p(A, B, C)/p(C)$$

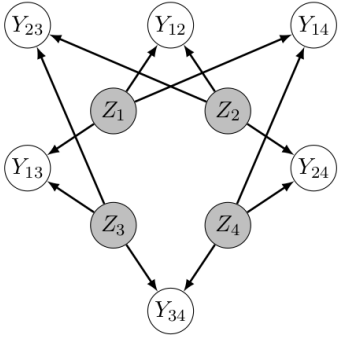
Source: S. Robin polycopié, notation change: replace Y by X

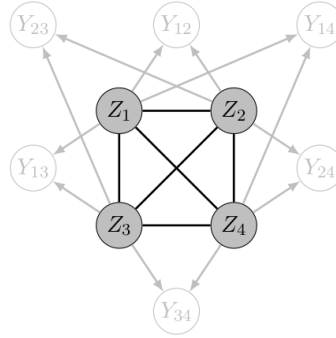
Thus $A, B \mid C$ are not independent since

$$p(A, B \mid C) = \frac{p(A)p(B)p(C \mid A, B)}{p(C)} \neq p(A \mid C)p(B \mid C).$$

does not factorize over A, B

SBM case: intricate posterior dependencies





Source: S. Robin polycopié, notation change: replace Y by X

Problem: computing $p(\boldsymbol{Z} \mid \boldsymbol{X})$ require exploring the K^n configuration $\rightsquigarrow$ no hope of simplifications as in mixture model/HMMs...

Variational inference & illustration for the SBM

Untractable E-step

Until now we always managed to compute the necessary moments of the posterior for E-step, *i.e.* to compute (given $\theta^{(t)}$)

$$f(\theta) = \mathbb{E}_{\boldsymbol{Z} \sim p_{\theta^{(t)}}(\cdot|\boldsymbol{X})} \left[\log p_\theta(\boldsymbol{X}, \boldsymbol{Z}) \right] \tag{1}$$

Untractable E-step

Until now we always managed to compute the necessary moments of the posterior for E-step, *i.e.* to compute (given $\theta^{(t)}$)

$$f(\theta) = \mathbb{E}_{\boldsymbol{Z} \sim p_{\theta^{(t)}}(\cdot|\boldsymbol{X})} \left[\log p_\theta(\boldsymbol{X}, \boldsymbol{Z}) \right] \tag{1}$$

Reminders: computing Equation (1) involve

- *For mixtures:* the marginal $\tau_{ik}^{(t+1)} = p_{\theta^{(t)}}(z_i = k \mid \boldsymbol{X}) = p_{\theta^{(t)}}(z_i = k \mid x_i)$ (posterior independence).
- *For HMMs:* we additionally need $\xi_{i,k,l} = p_{\theta^{(t)}}(z_i = k, z_{i+1} = l \mid \boldsymbol{X})$ + no posterior independence $\rightsquigarrow$ FB procedure in $\mathcal{O}(nK^2)$.

Untractable E-step

Until now we always managed to compute the necessary moments of the posterior for E-step, *i.e.* to compute (given $\theta^{(t)}$)

$$f(\theta) = \mathbb{E}_{\boldsymbol{Z} \sim p_{\theta^{(t)}}(\cdot|\boldsymbol{X})} \left[\log p_\theta(\boldsymbol{X}, \boldsymbol{Z}) \right] \tag{1}$$

Reminders: computing Equation (1) involve

- *For mixtures:* the marginal $\tau_{ik}^{(t+1)} = p_{\theta^{(t)}}(z_i = k \mid \boldsymbol{X}) = p_{\theta^{(t)}}(z_i = k \mid x_i)$ (posterior independence).
- *For HMMs:* we additionally need $\xi_{i,k,l} = p_{\theta^{(t)}}(z_i = k, z_{i+1} = l \mid \boldsymbol{X})$ + no posterior independence $\rightsquigarrow$ FB procedure in $\mathcal{O}(nK^2)$.

Problem: what if there's no hope of reasonable computation time for Equation (1) ? Either because

1. *complicated posterior dependencies:* $\boldsymbol{Z} \mid \boldsymbol{X}$ (as in SBM)
2. *intractable emission model:* even if posterior factorizes, it can be intractable. For example, in mixture $p_\theta(x_i \mid z_i = k) \propto \pi_k p_{\gamma_k}(x_i)$, intractable for a choice of non-analytical p_{γ_k}.

Back to EM: coordinate-ascent on the ELBO

Recall the coordinate-ascent formulation from slide 49

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

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

where $\mathcal{L}(q, \theta)$ is the ELBO:

$$\mathcal{L}(q, \theta) \coloneqq \mathbb{E}_{Z \sim q}[\log p_\theta(\boldsymbol{X}, \boldsymbol{Z})] + H(q) \qquad \text{(ELBO)}$$

Back to EM: coordinate-ascent on the ELBO

Recall the coordinate-ascent formulation from slide 49

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

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

where $\mathcal{L}(q, \theta)$ is the ELBO:

$$\mathcal{L}(q, \theta) \coloneqq \mathbb{E}_{Z \sim q}[\log p_\theta(\boldsymbol{X}, \boldsymbol{Z})] + H(q) \qquad \text{(ELBO)}$$

The E-step in an **unconstrained problem** over distribution $q \in \mathcal{P}(\boldsymbol{Z})$ (proba over $(z_1, \ldots, z_n)$). It can be rewritten as

$$q^{(t+1)} = \arg\min_{q \in \mathcal{P}(\boldsymbol{Z})} \mathrm{KL}(q(\cdot) \parallel p_{\theta^{(t)}}(\cdot \mid \boldsymbol{X})), \qquad \text{(E-step equivalent formulation)}$$

which naturally leads to setting $q^{(t+1)} = p_{\theta^{(t)}}(\cdot \mid \boldsymbol{X})$

Variational inference: constraining the E-step

Since complex models have intractable posteriors, we need to constrain the distribution q to belong in some prescribed family of probability distributions[7] $q \in \mathcal{Q} \subset \mathcal{P}(\boldsymbol{Z})$.

Variational inference: constraining the E-step

Since complex models have intractable posteriors, we need to constrain the distribution q to belong in some prescribed family of probability distributions[7] $q \in \mathcal{Q} \subset \mathcal{P}(\boldsymbol{Z})$.

The *variational, E-step* becomes

$$q^{(t+1)} = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q, \theta^{(t)}). \qquad \text{(VE-step)}$$

Or, equivalently,

$$q^{(t+1)} = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\cdot) \parallel p_{\theta^{(t)}}(\cdot \mid \boldsymbol{X})).$$

[7] The *variational* terminology stems from the fact that we are considering optimization problem over space of functions (probability densities) which is called variational calculus.

Variational inference: constraining the E-step

Since complex models have intractable posteriors, we need to constrain the distribution q to belong in some prescribed family of probability distributions[7] $q \in \mathcal{Q} \subset \mathcal{P}(\boldsymbol{Z})$.

The *variational, E-step* becomes

$$q^{(t+1)} = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q, \theta^{(t)}). \qquad \text{(VE-step)}$$

Or, equivalently,

$$q^{(t+1)} = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\cdot) \parallel p_{\theta^{(t)}}(\cdot \mid \boldsymbol{X})).$$

Key idea: choose $\mathcal{Q}$ such that calculations in (VE-step) are tractable.

[7] The *variational* terminology stems from the fact that we are considering optimization problem over space of functions (probability densities) which is called variational calculus.

A common choice of variational family: mean-field approximation

Natural follow-up question: what choice for q ?

Mean-field family: "forget" conditional dependencies of $\boldsymbol{Z} \mid \boldsymbol{X}$

$$q \in \mathcal{Q} = \left\{ q_\tau \ : \ q_\tau(\boldsymbol{Z}) = \prod_{i=1}^{n} q_{\tau_i}(z_i), \quad \tau_i \in \Psi \right\} \quad \text{so that } \max_{q \in \mathcal{Q}} \mathcal{L}(q) = \max_{\tau \in \Psi^n} \mathcal{L}(q_\tau). \qquad (2)$$

A common choice of variational family: mean-field approximation

Natural follow-up question: what choice for q ?

Mean-field family: "forget" conditional dependencies of $\boldsymbol{Z} \mid \boldsymbol{X}$

$$q \in \mathcal{Q} = \left\{ q_\tau \ : \ q_\tau(\boldsymbol{Z}) = \prod_{i=1}^{n} q_{\tau_i}(z_i), \quad \tau_i \in \Psi \right\} \quad \text{so that } \max_{q \in \mathcal{Q}} \mathcal{L}(q) = \max_{\tau \in \Psi^n} \mathcal{L}(q_\tau). \qquad (2)$$

Important remark: q is not a *model* of the observed data but rather the ELBO (and the KL minimization) connects q to the data & the model (Blei et al. 2017)

A common choice of variational family: mean-field approximation

Natural follow-up question: what choice for q ?

Mean-field family: "forget" conditional dependencies of $\boldsymbol{Z} \mid \boldsymbol{X}$

$$q \in \mathcal{Q} = \left\{ q_\tau \ : \ q_\tau(\boldsymbol{Z}) = \prod_{i=1}^{n} q_{\tau_i}(z_i), \quad \tau_i \in \Psi \right\} \quad \text{so that } \max_{q \in \mathcal{Q}} \mathcal{L}(q) = \max_{\tau \in \Psi^n} \mathcal{L}(q_\tau). \qquad (2)$$

Important remark: q is not a *model* of the observed data but rather the ELBO (and the KL minimization) connects q to the data & the model (Blei et al. 2017)

Property

- Entropy term: by independence $\mathcal{H}(q) = \sum_{i=1}^{n} \mathcal{H}(q_i)$
- when z_i is discrete (this course): we can enforce a parametric form $q_{\tau_i}(z_i) = \mathcal{M}_K(1; \tau_i)$ and $\Psi = \Delta_K$

Variational-EM (VEM) algorithm

VEM algo: coordinate-ascent on the ELBO

Start from initial $\theta^{(0)}$ and set a variational family $\mathcal{Q}$

$$q^{(t+1)} = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q, \theta^{(t)}), \qquad \text{(VE-step)}$$

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

Variational-EM (VEM) algorithm

VEM algo: coordinate-ascent on the ELBO

Start from initial $\theta^{(0)}$ and set a variational family $\mathcal{Q}$

$$q^{(t+1)} = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q, \theta^{(t)}), \qquad \text{(VE-step)}$$

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

In general, maximization over $\tau = (\tau_1, \ldots, \tau_n)$ is done via a coordinate ascent / fixed-point algorithm where we iteratively update q_i keeping q_{-i} fixed, iterating through $i = 1, \ldots, n$:

$$q_i^\star = \arg\max_{q_i} \mathcal{L}(q_i, q_{-i}). \qquad \text{(CAVI)}$$

Variational-EM (VEM) algorithm

VEM algo: coordinate-ascent on the ELBO

Start from initial $\theta^{(0)}$ and set a variational family $\mathcal{Q}$

$$q^{(t+1)} = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q, \theta^{(t)}), \qquad \text{(VE-step)}$$

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

In general, maximization over $\tau = (\tau_1, \ldots, \tau_n)$ is done via a coordinate ascent / fixed-point algorithm where we iteratively update q_i keeping q_{-i} fixed, iterating through $i = 1, \ldots, n$:

$$q_i^\star = \arg\max_{q_i} \mathcal{L}(q_i, q_{-i}). \qquad \text{(CAVI)}$$

Pros & cons of VEM algorithm

- **Pros:**
  1. we choose $\mathcal{Q}$ such that everything is tractable
  2. Approximation of intractable posterior via $q^{(T)}$ in the sense of KL-divergence
- **Cons:** only increase the ELBO, no guarantee to increase the likelihood anymore ! We get an estimator

$$\hat{\theta}_V \in \arg\max_{\theta} \mathcal{L}(q^{(T)}, \theta)$$

The VEM algorithm for SBM: VE-step (i)

Let's write the VE-step for the mean-field approximation

$$q_\tau(\boldsymbol{Z}) = \prod_{i=1}^{n} \mathcal{M}_k(z_i \mid 1, \tau_i) = \prod_{i=1}^{n} \prod_{k=1}^{K} \tau_{ik}^{z_{ik}}, \quad \sum_k \tau_{ik} = 1, \forall i$$

with entropy $\mathcal{H}(q) = -\sum_i \sum_k \tau_{ik} \log \tau_{ik}$.

The ELBO then writes as

$$\begin{aligned}
\mathcal{L}(q_\tau, \theta) =& \mathbb{E}_q \left[\log p_\theta(\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X} \right] + \mathcal{H}(q), \\
=& \sum_{i \neq j} \sum_{k,l=1}^{K} \mathbb{E}_q \left[z_{ik} z_{jl} \right] \log p_{\gamma_{kl}}(x_{ij}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_q \left[z_{ik} \right] \log \pi_k - \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \tau_{ik} \\
=& \sum_{i \neq j} \sum_{k,l=1}^{K} \tau_{ik} \tau_{jl} \log p_{\gamma_{kl}}(x_{ij}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \pi_k - \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \tau_{ik}
\end{aligned}$$

where we used the fact that $z_i \perp\!\!\!\perp z_j$ under q, and $\mathbb{E}_q[z_{ik}] = q_i(z_{ik} = 1) = \tau_{ik}$

The VEM algorithm for SBM: VE-step (ii)

In VE-step, we wish to maximize $\mathcal{L}(q_\tau, \theta)$ with respect to τ under n constraints $\sum_k \tau_{ik} = 1$.

Introducing $(\lambda_i)_{i=1}^n$, we seek stationnary points of the Lagrangian

$$\mathcal{L}(q_\tau, \theta) + \sum_{i=1}^{n} \lambda_i (1 - \sum_{l=1}^{K} \tau_{ik}).$$

Which naturally leads to $n \times K$ equations

$$\sum_{l=1}^{K} \sum_{j \neq i, j=1}^{n} \hat{\tau}_{jl} \log p_{\gamma_{kl}}(x_{ij}) + \log \pi_k - \log \tau_{ik} - 1 - \lambda_i = 0. \tag{3}$$

under the constraints $\sum_{k=1}^{K} \tau_{ik} = 1, \forall i$

The VEM algorithm for SBM: VE-step (iii)

The $n \times K$ equations can be rewritten for fixed i, k

$$\hat{\tau}_{ik} = e^{1+\lambda_i} \pi_k \prod_{j \neq i, j=1}^{n} \prod_{l=1}^{K} p_{\gamma_{kl}}^{\hat{\tau}_{jl}} \quad \propto \pi_k \prod_{j \neq i, j=1}^{n} \prod_{l=1}^{K} p_{\gamma_{kl}}^{\hat{\tau}_{jl}} \pi_k \tag{4}$$

Coordinate-ascent: fixed point algorithm where we iterate through Equation (4) for all $i = 1, \ldots, n$ until a criterion is verified.

Some remarks

1. In fixed point, for each i the coordinate τ_{ik} is normalized to be a probability vector.
2. VE-step is itself iterative: stop after `ve.niter` iterations or increase in $\tau \mapsto \mathcal{L}(\tau, \theta^{(t)})$ below a given threshold.

M-step

Bonus as an homework

Clustering ?

We used to perform MAP estimation of $\boldsymbol{Z}$

$$\hat{\boldsymbol{Z}} \in \arg\max_{\boldsymbol{Z}} p_{\hat{\theta}}(\boldsymbol{Z} \mid \boldsymbol{X}).$$

$\rightsquigarrow$ intractable here.

But VEM also outputs a KL approximation of $q_{\hat{\tau}} \approx p_{\hat{\theta}}(\cdot \mid \boldsymbol{X})$ in the mean-field variational family, so we can use

$$\hat{\boldsymbol{Z}} \in \arg\max q_{\hat{\tau}}(\boldsymbol{Z}) = \arg\max \hat{\tau}$$

Conclusion of the course

What we saw in this course

Three examples of general discrete latent variable models: GMM, HMM, SBM

- incomplete data models: $p_\theta(\boldsymbol{X}) = \sum_{\boldsymbol{Z}} p_\theta(\boldsymbol{X}, \boldsymbol{Z})$
- the complete likelihood is easier to write than the marginal (but we do not observe $\boldsymbol{Z}$)
- Generalizes well to different type of data (discrete, continuous) via the choice of different $\boldsymbol{X} \mid \boldsymbol{Z}$ (*i.e.* p_γ)

Inference procedures for latent variable model

- EM algorithm
- Main difficulties lies in E-step and links to the tractability of the posterior $\boldsymbol{Z} \mid \boldsymbol{X}$
  - tractable for mixture
  - tractable (forward-backward) for HMMs: clever use of the DAG
  - intractable for SBM
- M-step is model dependent, *i.e.* depends on the choice of $\boldsymbol{X} \mid \boldsymbol{Z}$.

$\rightsquigarrow$ this inference part is general and applies to continuous latent variable models as well !

Pratical implementation and caveats

Bibliography I

- Arthur, David and Sergei Vassilvitskii (2007). "K-means++ the advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035.
- Bishop, Christopher M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.
- Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518, pp. 859–877.
- Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1, pp. 1–22.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA.
- Lebarbier, Emilie and Tristan Mary-Huard (2004). "Le critère BIC: fondements théoriques et interprétation". PhD thesis. INRIA.
- MacQueen, James (1967). "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA, pp. 281–297.

Bibliography II

- Murphy, Kevin P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press.
- Peel, DAVID and G MacLahlan (2000). "Finite mixture models". In: *John & Sons*.
- Rabiner, Lawrence R (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2, pp. 257–286.
- Yoon, Byung-Jun (2009). "Hidden Markov Models and their Applications in Biological Sequence Analysis". In: *Current Genomics* 10, pp. 402 –415.