Unsupervised Learning with discrete latent variable models

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

M2 Data-Science 2024-2025

Organization

Tuesdays 9h45 - 13h (see agenda)

$3 \times$ 3h classes

Come see me today for the reading report

See course's website

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...

Latent variables models for unsupervised learning

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

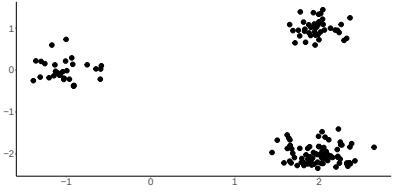
What this course is about...

Latent variables models for unsupervised learning

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

Example: Clustering

X is an unlabeled observation and Z its group membership



- K-means
- Mixture Models
- (hierarchical clustering, HMMs, graphs)

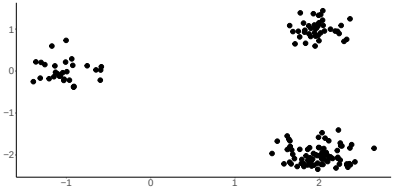
What this course is about...

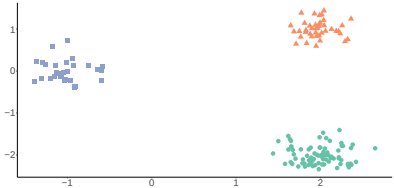
Latent variables models for unsupervised learning

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

Example: Clustering

X is an unlabeled observation and Z its group membership





- K-means
- Mixture Models
- (hierarchical clustering, HMMs, graphs)

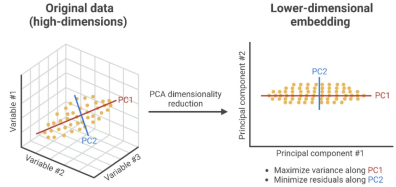
What this course is about...

Latent variables models for unsupervised learning

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

Example: Dimension reduction

X in dimension $p >> 1$ and Z its low-dimensional representation



Source: `https://aiml.com/what-is-dimensionality-reduction-2/`

- Principal Component Analysis (PCA)
- Variational Auto-encoder (VAE)
- (UMAP, t-SNE)

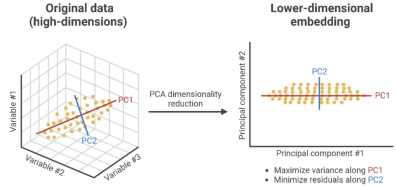
What this course is about...

Latent variables models for unsupervised learning

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

Example: Dimension reduction

X in dimension $p >> 1$ and Z its low-dimensional representation



Source: `https://aiml.com/what-is-dimensionality-reduction-2/`

- Principal Component Analysis (PCA)
- Variational Auto-encoder (VAE)
- (UMAP, t-SNE)

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 $\sum$

Course outline

1. Clustering with mixture models
2. Inference in latent variable models: the EM algorithm
3. Probabilistic dimension reduction
4. Conclusion of the course

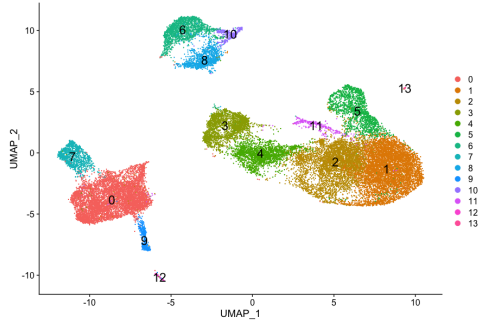
Clustering with mixture models

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*[1]

[1] 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 $\mu_k \coloneqq \frac{1}{|C_k|} \sum_{i \in C_k} x_i$ 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 (Voronoï cells)

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

2. recompute centroids as the barycenter of each center

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

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

Properties of the algorithm

K-means is a greedy algorithm which

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

$\rightsquigarrow$ 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 $\boldsymbol{Z} \mid \boldsymbol{X}$

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 $C = \{C_1, \ldots, C_K\}$.

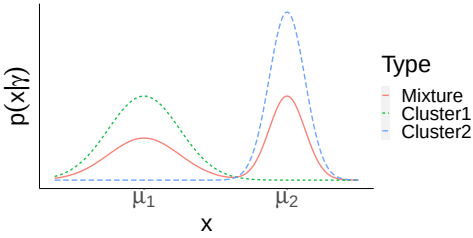
$\rightsquigarrow$ can cause problems in Bayesian inference procedure since the posterior is highly multimodal.

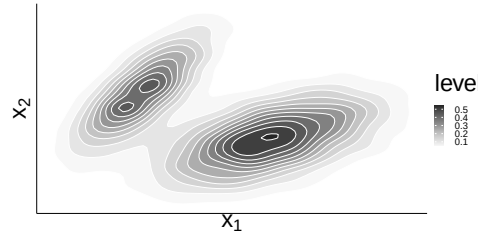
Gaussian Mixture Models (GMM)

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 discrete random variable

Definition: discrete 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. Can be negative.

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

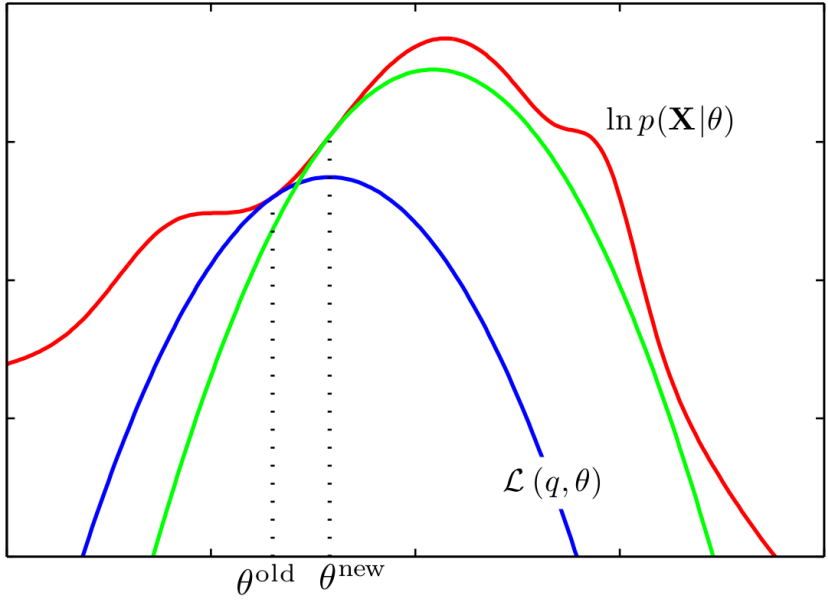
$$\forall t \geq 0, \quad \log p_{\theta^{(t+1)}}(\boldsymbol{X}) \geq \log p_{\theta^{(t)}}(\boldsymbol{X})$$

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

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

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

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

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.

Probabilistic dimension reduction

Principal component analysis (PCA)

Probabilistic PCA (pPCA)

EM algorithm for pPCA

Non-linear extension: variational auto-encoders (VAEs)

Variational inference & illustration for the VAEs

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 pPCA:* Gaussian world, the conditional is also Gaussian

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 pPCA:* Gaussian world, the conditional is also Gaussian

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

1. *complicated posterior dependencies:* $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \mid \boldsymbol{X}$ (DAG moralization)
2. *intractable normalization constant* $p_\theta(x_i) = \int p_\theta(x_i \mid \boldsymbol{Z}) \, \mathrm{d}\boldsymbol{Z}$

Back to EM: coordinate-ascent on the ELBO

Recall the coordinate-ascent formulation from slide 35

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

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

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 35

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

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

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[2] $q \in \mathcal{Q} \subset \mathcal{P}(\boldsymbol{Z})$.

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[2] $q \in \mathcal{Q} \subset \mathcal{P}(\boldsymbol{Z})$.

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,

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

[2] 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[2] $q \in \mathcal{Q} \subset \mathcal{P}(\boldsymbol{Z})$.

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.

[2] 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 $\boldsymbol{Z} \mid \boldsymbol{X}$

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

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

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

Conclusion of the course

What we saw in this course

Three examples of general discrete latent variable models: GMM, pPCA, VAEs

- 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 $\boldsymbol{X} \mid \boldsymbol{Z}$.

What we did not cover: ratical 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.