Statistical learning

Alain Celisse

SAMM

Paris 1-Panthéon Sorbonne University

alain.celisse@univ-paris1.fr

Lecture 3: Clustering task

—

Master 2 MMMEF – Paris 1 – Fall 2024

Outline of the lectures

Successive topics of the coming lectures:

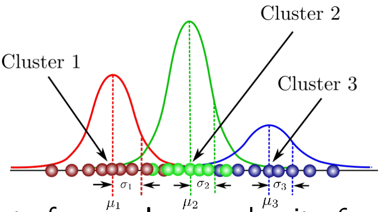
1. Concentration inequalities
2. Linear regression and model selection
3. Clustering task: Mixture models (Today!)
4. Dimension reduction: PCA and Spectral clustering
5. Classification task

Outline of the lecture

- Clustering problem
- Gaussian Mixture Models
- Density estimation/Clustering
- Estimation and EM algorithm
- Model selection

Clustering and GMM

Heterogeneous data within the cloud



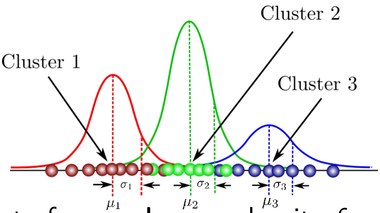
Data

- $X_1, \ldots, X_n \in \mathbb{R}^d$: *i.i.d.* data from **unknown** density f_θ
- θ: parameter vector (to be precised)

Assumptions:

- The data are heterogeneous with G classes
- Each class is spread over a different area
  (can be distinguished from one another)
- Each class has a specific structure (encoded by the parameters)

Heterogeneous data within the cloud



Data

- $X_1, \ldots, X_n \in \mathbb{R}^d$: *i.i.d.* data from **unknown** density f_θ
- θ: parameter vector (to be precised)

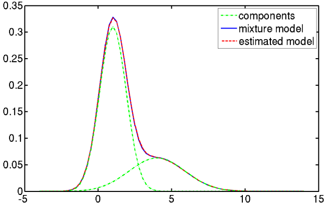
Assumptions:

- The data are heterogeneous with G classes
- Each class is spread over a different area (can be distinguished from one another)
- Each class has a specific structure (encoded by the parameters)

Remark:

Classes can strongly overlap which does not necessarily mean they do not exist!

Mixture Model and GMM



Mixture Model

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_\beta$: density

$$X_i \sim f_\beta(x) = \sum_{g=1}^{G} \pi_g f_{\theta_g}(x)$$

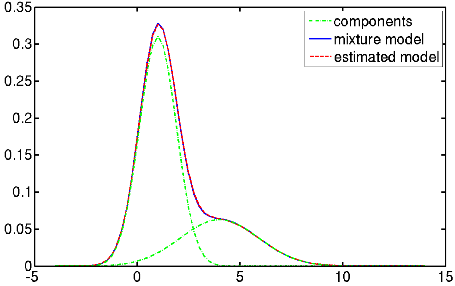
- G classes (clusters)
- π_g: weight of the gth component of the mixture
- f_{θ_g}: density of data within the gth cluster
- θ_g: parameter within the gth cluster
- $\beta = (G, \pi_1, \ldots, \pi_G, \theta_1, \ldots, \theta_G)$

Remark:

- Gaussian Mixture Model (GMM) if all f_{θ_g} are Gaussian

Influence of the proportion

$$X_i \sim f_\beta(x) = \sum_{g=1}^{G} \pi_g f_{\theta_g}(x)$$



- Displayed densities do not (visually) integrate to 1 on the picture!
- "Components" displayed with their proportions $\widehat{\pi}_g$

GMM and Hidden Variables

$$X_i \sim f_\beta(x) = \sum_{g=1}^{G} \pi_g f_{\theta_g}(x)$$

Alternative perspective and model

For each $1 \leq i \leq n$

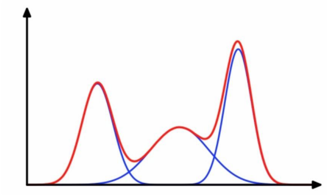
- π_g: Probability belonging to class g $(\pi_1 + \ldots + \pi_G = 1)$
- $H_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_{G-1}, \pi_G)$: Hidden variable (label)
- $X_i \mid H_i = g \quad \sim \quad f_{\theta_g}$: density of data from class g

Remark:

- Clustering individuals means recovering the unknown (hidden) variable H_i for each i
- Gives a strategy for generating data from a mixture!

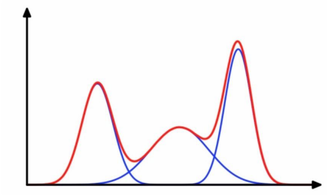
Density estimation versus Clustering

Clustering as an unachievable task

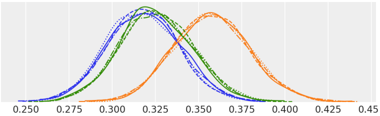


Well separated classes $\longrightarrow$ Clustering is easy

Clustering as an unachievable task



Well separated classes $\rightarrow$ Clustering is easy



Overlapping classes $\rightarrow$ Clustering almost Impossible!

Low SNR $\rightarrow$ Density estimation

Parametric Density estimation (with mixtures)

- f: unknown density of $X_1, \ldots, X_n$
- $\widehat{f} = f_{\widehat{\beta}}$: Parametric estimator of f given by

$$f_{\widehat{\beta}} = \sum_{g=1}^{\widehat{G}} \widehat{\pi}_g f_{\widehat{\theta}_g} \qquad \text{(mixture)}$$

- "Parametric" since β is finite-dimensional

Link with Clustering: Maximum *a posteriori* (MAP) rule

- Outputs the components of the mixture
- Each component corresponds to a cluster

Definition (MAP rule)

The Maximum *a posteriori* (MAP) rule is given by

$$\widehat{g} = Arg \max_{1 \leq g \leq \widehat{G}} \frac{\widehat{\pi}_g f_{\widehat{\theta}_g}(x)}{\sum_{g'=1}^{\widehat{G}} \widehat{\pi}_{g'} f_{\widehat{\theta}_{g'}}(x)} = Arg \max_{1 \leq g \leq \widehat{G}} \left\{ \widehat{\pi}_g f_{\widehat{\theta}_g}(x) \right\}$$

MAP rule justification: Bayes classifier

- $X \in \mathbb{R}^d$: random variable (vector of descriptors)
- H: Hidden label corresponding to location X

Once $X = x$ is observed, what is the label of this point?

Bayes optimal classifier (Reminder)

$$g^\star(x) = Arg \max_{1 \leq g \leq G} \mathbb{P}[\, H = g \mid X = x \,]$$

Justification for the MAP rule.

Bayes' rule yields

$$\mathbb{P}[\, H = g \mid X = x \,] = \frac{\mathbb{P}[\, X = x \mid H = g \,] \cdot \mathbb{P}[\, H = g \,]}{\mathbb{P}[\, X = x \,]} = \frac{\pi_g f_{\theta_g}(x)}{\sum_{g'=1}^{G} \pi_{g'} f_{\theta_{g'}}(x)}$$

Quantifying the clustering uncertainty

Estimated a posteriori probabilities

- Once θ is estimated, we have access to an estimator of the a posteriori probability of each class

$$\widehat{\mathbb{P}}\left[H = g \mid X = x \right] = \frac{\widehat{\pi}_g f_{\widehat{\theta}_g}(x)}{\sum_{g'=1}^{G} \widehat{\pi}_{g'} f_{\widehat{\theta}_{g'}}(x)}$$

- This estimator can serve as a means for quantifying the strength of the overlapping phenomenon

Ex:

- 3 classes exhibit a posteriori probabilities close to $\frac{1}{3}$ at a point x
- *Interpretation:*
  Three overlapping classes in a neighborhood of x
- No strong reasons for choosing one of them...

Parametric Density estimation

Mixture models and key quantities

- Hidden label: $(\delta_g(\cdot)$: Dirac measure)

$$H \sim \mathcal{M}(1; \pi_1, \ldots, \pi_G) \qquad \Leftrightarrow \qquad f_\beta^H(h) = \sum_{g=1}^{G} \pi_g \delta_g(h)$$

Mixture models and key quantities

- Hidden label: ($\delta_g(\cdot)$: Dirac measure)

$$H \sim \mathcal{M}(1; \pi_1, \ldots, \pi_G) \qquad \Leftrightarrow \qquad f_\beta^H(h) = \sum_{g=1}^{G} \pi_g \delta_g(h)$$

- **Conditional density of X given $H = g$:**

$$f_\beta^{X|H=g}(x) = f_{\theta_g}(x)$$

Ex: Gaussian density $\rightarrow$ parametric assumption

Mixture models and key quantities (Cont'd)

$$(f_\beta^{X|H=g}(x) = f_{\theta_g}(x))$$

- Density of X:

$$f_\beta^X(x) = \sum_{g=1}^{G} \pi_g f_{\theta_g}(x)$$

$\rightarrow$ Mixture probability distribution

Mixture models and key quantities (Cont'd)

$$\left(f_\beta^{X|H=g}(x) = f_{\theta_g}(x)\right)$$

- **Density of X:**

$$f_\beta^X(x) = \sum_{g=1}^{G} \pi_g f_{\theta_g}(x)$$

$\rightarrow$ Mixture probability distribution

- **Joint distribution of (X, H):**

$$\begin{aligned} f_\beta^{(X,H)}(x,h) &= f_\beta^{X|H=h}(x) \cdot f_\beta^H(h) \\ &= \sum_{g=1}^{G} (f_{\theta_g}(x) \cdot \pi_g)\delta_g(h) \end{aligned}$$

Density and log-likelihood (Reminder)

- $X_1, \ldots, X_n$: *i.i.d.* data drawn from a density f_β
- **Density of** $X_1^n = (X_1, \ldots, X_n)$:

$$(x_1, \ldots, x_n) \mapsto f_\beta^{X_1^n}(\underbrace{x_1, \ldots, x_n}_{=x_1^n}) = \prod_{i=1}^{n} f_\beta(x_i)$$

- **Likelihood of** β:

$$\beta \mapsto f_\beta^{X_1^n}(\underbrace{x_1, \ldots, x_n}_{=x_1^n}) = \prod_{i=1}^{n} f_\beta(x_i)$$

- log-**likelihood of** β:

$$\beta \mapsto \mathcal{L}_\beta(x_1^n) = \log\left(f_\beta^{X_1^n}(x_1^n)\right) = \sum_{i=1}^{n} \underbrace{\log\left(f_\beta^{X_i}(x_i)\right)}_{=\ell_\beta^{X_i}(x_i)}$$

Maximum likelihood estimator

Maximum Likelihood Estimator of β

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$: unknown density
- $\{f_\beta \mid \beta \in B\}$: parametric model for estimating f

Maximum likelihood estimator

Maximum Likelihood Estimator of β

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$: unknown density
- $\{f_\beta \mid \beta \in B\}$: parametric model for estimating f

Definition (MLE of β)

With $\mathcal{L}_\beta(x_1^n) = \log\left(f_\beta^{X_1^n}(x_1^n)\right)$, the MLE of β is given by

$$\widehat{\beta} \in Arg\max_{\beta} \{\mathcal{L}_\beta(x_1^n)\}$$

Remark: Particular instance of the ERM principle

MLE justification: LLN

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$: unknown density
- $\{f_\beta \mid \beta \in B\}$: parametric model for estimating f

Justification (Heuristic)

$$\frac{\mathcal{L}_\beta(X_1^n)}{n} = \frac{1}{n}\sum_{i=1}^{n} \log\left(f_\beta(X_i)\right) \xrightarrow[n\to+\infty]{P} \int_{\mathbb{R}^d} \log\left(f_\beta(x)\right) \cdot f(x)\, dx$$

Hence,

$$Arg \max_\beta \{\mathcal{L}_\beta(x_1^n)\} \approx Arg \max_\beta \left\{ \int_{\mathbb{R}^d} \log\left(f_\beta(x)\right) \cdot f(x)\, dx \right\}$$
$$= Arg \max_\beta \left\{ \int_{\mathbb{R}^d} \log\left(f_\beta(x)\right) \cdot f(x)\, dx - \int_{\mathbb{R}^d} \log\left(f(x)\right) \cdot f(x)\, dx \right\}$$
$$= Arg \max_\beta \{-KL(f; f_\beta)\} = Arg \min_\beta \{KL(f; f_\beta)\}$$

where $KL(f; g)$: Kullback-Leibler divergence

Kullback-Leibler divergence

Measuring the gap between probability distributions

Definition (KL-Divergence)

f, g: two densities over $\mathbb{R}^d$ w.r.t. λ.
The $KL(f; g)$: Kullback-Leibler divergence between f and g is given by

$$KL(f; g) = \int_{\mathbb{R}^d} \log\left(\frac{f(x)}{g(x)}\right) \cdot f(x)\, d\lambda(x) \geq 0$$

Kullback-Leibler divergence

Measuring the gap between probability distributions

Definition (KL-Divergence)

f, g: two densities over $\mathbb{R}^d$ w.r.t. λ.
The $KL(f; g)$: Kullback-Leibler divergence between f and g is given by

$$KL(f; g) = \int_{\mathbb{R}^d} \log\left(\frac{f(x)}{g(x)}\right) \cdot f(x)\, d\lambda(x) \geq 0$$

- $KL(f; g) \geq 0$

Kullback-Leibler divergence

Measuring the gap between probability distributions

Definition (KL-Divergence)

f, g: two densities over $\mathbb{R}^d$ w.r.t. λ.
The $KL(f;g)$: Kullback-Leibler divergence between f and g is given by

$$KL(f;g) = \int_{\mathbb{R}^d} \log\left(\frac{f(x)}{g(x)}\right) \cdot f(x)\, d\lambda(x) \geq 0$$

- $KL(f;g) \geq 0$ *(Hint: $x \mapsto x\log(x)$ convex)*
- $K(f;f) = 0$
- $KL(f;g)$ measures how much g departs from f

Kullback-Leibler divergence

Measuring the gap between probability distributions

Definition (KL-Divergence)

f, g: two densities over $\mathbb{R}^d$ w.r.t. λ.
The $KL(f;g)$: Kullback-Leibler divergence between f and g is given by

$$KL(f;g) = \int_{\mathbb{R}^d} \log\left(\frac{f(x)}{g(x)}\right) \cdot f(x)\, d\lambda(x) \geq 0$$

- $KL(f;g) \geq 0$ *(Hint: $x \mapsto x\log(x)$ convex)*
- $K(f;f) = 0$
- $KL(f;g)$ measures how much g departs from f
- $KL(f;g) \neq KL(g;f)$: not a distance!
- $KL(f;g) = +\infty$ if
  there exists x s.t. $g(x) = 0$ but $f(x) \neq 0$

Computing the maximum location

Recap

At this point, we aim at computing

$$\widehat{\beta} \in Arg\max_{\beta} \{\mathcal{L}_\beta(x_1^n)\} = Arg\max_{\beta} \left\{ \sum_{i=1}^{n} \log\left(f_\beta^{X_i}(x_i) \right) \right\}$$

Problems

- $f_\beta^{X_i}(x_i) = \sum_{g=1}^{G} \pi_g f_{\theta_g}(x_i)$

  Ex: $f_{\theta_g}(x_i) = \frac{1}{\sqrt{2\pi\sigma_g^2}} e^{-\frac{(x_i-\mu_g)^2}{2\sigma_g^2}}$
- $\log\left(\sum_{g=1}^{G} \pi_g f_{\theta_g}(x_i) \right)$: No closed-form expression for $\widehat{\beta}$

Computing the maximum location

Recap

At this point, we aim at computing

$$\widehat{\beta} \in Arg\max_{\beta} \{\mathcal{L}_{\beta}(x_1^n)\} = Arg\max_{\beta} \left\{ \sum_{i=1}^{n} \log\left(f_{\beta}^{X_i}(x_i) \right) \right\}$$

Problems

- $f_{\beta}^{X_i}(x_i) = \sum_{g=1}^{G} \pi_g f_{\theta_g}(x_i)$

  Ex: $f_{\theta_g}(x_i) = \frac{1}{\sqrt{2\pi\sigma_g^2}} e^{-\frac{(x_i-\mu_g)^2}{2\sigma_g^2}}$
- $\log\left(\sum_{g=1}^{G} \pi_g f_{\theta_g}(x_i)\right)$: No closed-form expression for $\widehat{\beta}$
- Requires an optimization algorithm ($\rightarrow$ see EM-algo.)
- $\beta \mapsto \mathcal{L}_{\beta}(x_1^n)$: Not convex in general

Expectation-Maximization (EM) Algorithm

Optimization strategy

Strategy: Step 1

Goal Find

$$\widehat{\beta} \in Arg \max_{\beta} \{\mathcal{L}_\beta(x_1^n)\} = Arg \max_{\beta} \left\{ \log \left(f_\beta^{X_1^n}(x_1^n) \right) \right\}$$

Key ingredients

First: Bayes' rule:

$$f_\beta^{X_i}(x_i) = \frac{f_\beta^{(X_i,H_i)}(x_i, h)}{f_\beta^{H_i|X_i=x_i}(h)}$$

$$\Rightarrow \quad \log \left(f_\beta^{X_i}(x_i) \right) = \ell_\beta^{X_i}(x_i) = \ell_\beta^{(X_i,H_i)}(x_i, h) - \ell_\beta^{H_i|X_i=x_i}(h)$$

$$\Rightarrow \quad \mathcal{L}_\beta(x_1^n) = \underbrace{\sum_{i=1}^n \ell_\beta^{(X_i,H_i)}(x_i, h_i)}_{=L_\beta^1(x_1^n, h_1^n)} - \underbrace{\sum_{i=1}^n \ell_\beta^{H_i|X_i=x_i}(h_i)}_{=L_\beta^2(h_1^n)}$$

Complete log-likelihood

$$\mathcal{L}_\beta(x_1^n) = \underbrace{\sum_{i=1}^{n} \ell_\beta^{(X_i,H_i)}(x_i, h_i)}_{=L_\beta^1(x_1^n,h_1^n)} - \underbrace{\sum_{i=1}^{n} \ell_\beta^{H_i|X_i=x_i}(h_i)}_{=L_\beta^2(h_1^n)}$$

Vocabulary

- $\beta \mapsto L_\beta^1(x_1^n, h_1^n)$: complete log-likelihood
- $\beta \mapsto L_\beta^2(h_1^n)$: log-likelihood at the latent variables (hidden variables)

Complete log-likelihood

$$\mathcal{L}_\beta(x_1^n) = \underbrace{\sum_{i=1}^{n} \ell_\beta^{(X_i,H_i)}(x_i, h_i)}_{=L_\beta^1(x_1^n, h_1^n)} - \underbrace{\sum_{i=1}^{n} \ell_\beta^{H_i|X_i=x_i}(h_i)}_{=L_\beta^2(h_1^n)}$$

Vocabulary

- $\beta \mapsto L_\beta^1(x_1^n, h_1^n)$: complete log-likelihood
- $\beta \mapsto L_\beta^2(h_1^n)$: log-likelihood at the latent variables (hidden variables)

Remark:

$$\mathcal{L}_\beta(x_1^n) = L_\beta^1(x_1^n, h_1^n) - L_\beta^2(h_1^n)$$

- $\mathcal{L}_\beta(x_1^n)$ does not depend on h_1^n (cancellations...)

Strategy: Step 2

β^t: value of β estimated at iteration t of the optim. algo.

Key ingredients

Second: Cond. expectation of $H_1^n \mid X_1^n = x_1^n$ w.r.t. β^t

$$\mathcal{L}_\beta(X_1^n) = L^1_\beta(X_1^n, H_1^n) - L^2_\beta(H_1^n)$$

$$\Rightarrow \quad \mathbb{E}_{\beta^t}\left[\mathcal{L}_\beta(X_1^n) \mid X_1^n = x_1^n\right] =$$

$$\mathbb{E}_{\beta^t}\left[L^1_\beta(X_1^n, H_1^n) \mid X_1^n = x_1^n\right] - \mathbb{E}_{\beta^t}\left[L^2_\beta(H_1^n) \mid X_1^n = x_1^n\right]$$

$$\Leftrightarrow \quad \mathcal{L}_\beta(x_1^n) =$$

$$\underbrace{\mathbb{E}_{\beta^t}\left[L^1_\beta(X_1^n, H_1^n) \mid X_1^n = x_1^n\right]}_{=Q(\beta|\beta^t)} - \underbrace{\mathbb{E}_{\beta^t}\left[L^2_\beta(H_1^n) \mid X_1^n = x_1^n\right]}_{=R(\beta|\beta^t)}$$

Remark:

- β is to be chosen from the current parameter value β^t
- Maximizing $\mathcal{L}_\beta(x_1^n)$ amounts to maximize $Q(\beta \mid \beta^t) - R(\beta \mid \beta^t)$
- Requires $Q(\beta \mid \beta^t)$ and $R(\beta \mid \beta^t)$ to be computed

Strategy: Step 3

Key ingredients

Third: Computing $Q(\beta \mid \beta^t)$ with Mixture Models

From
$$\log\left(f_\beta^{(X,H)}(x,h)\right) = \sum_{g=1}^{G} \log(f_{\theta_g}(x)\cdot \pi_g)\delta_g(h)$$

we deduce that

$$\begin{aligned}
&Q(\beta \mid \beta^t)\\
&= \mathbb{E}_{\beta^t}\left[\, L_\beta^1(X_1^n, H_1^n) \mid X_1^n = x_1^n \,\right]\\
&= \sum_{i=1}^{n} \mathbb{E}_{\beta^t}\left[\sum_{g=1}^{G} \log(f_{\theta_g}(x_i)\cdot \pi_g)\delta_g(H_i) \mid X_1^n = x_1^n\right]\\
&= \sum_{i=1}^{n}\sum_{g=1}^{G} \log(f_{\theta_g}(x_i)\cdot \pi_g)\underbrace{\mathbb{E}_{\beta^t}\left[\,\delta_g(H_i) \mid X_1^n = x_1^n\,\right]}_{=\mathbb{P}_{\beta^t}(H_i=g|X_i=x_i)}
\end{aligned}$$

Remark: $\mathbb{P}_{\beta^t}(H_i = g \mid X_i = x_i)$: fully known!

Strategy: Step 4

Key ingredients

Fourth: Computing $R(\beta \mid \beta^t)$ with Mixture Models

Using that

$$f_\beta^{H|X=x}(h) = \frac{\pi_h f_{\theta_h}(x)}{\sum_g \pi_g f_{\theta_g}(x)}$$

$$\begin{aligned}
&R(\beta \mid \beta^t) \\
&= \mathbb{E}_{\beta^t}\left[L_\beta^2(H_1^n) \mid X_1^n = x_1^n \right] \\
&= \sum_{i=1}^n \mathbb{E}_{\beta^t}\left[\log\left(f_\beta^{H_i|X_i=x_i}(H_i) \right) \mid X_i = x_i \right] \\
&= \sum_{i=1}^n \int_{\mathbb{R}^d} \log\left(f_\beta^{H_i|X_i=x_i}(h) \right) f_{\beta^t}^{H_i|X_i=x_i}(h)\, dh
\end{aligned}$$

Cannot be easily evaluated in general!

EM-algorithm

EM-algorithm formulation

The goal is (approximately) maximizing $\beta \mapsto \mathcal{L}_\beta(x_1^n)$

Algorithm (EM, general formulation)

1. *Initialize the iterative process with* $\beta = \beta^0$
2. *For* $t = 1, \ldots, T$*:*
   Apply
   - **E-step**: *Compute the Expectation*

     $$Q(\beta \mid \beta^t) = \mathbb{E}_{\beta^t}\left[L_\beta^1(X_1^n, H_1^n) \mid X_1^n = x_1^n \right]$$

   - **M-step**: *Compute the Maximum location*

     $$\beta^{t+1} \in Arg\max_\beta \left\{ Q(\beta \mid \beta^t) \right\}$$

T: defined from a convergence criterion of the difference between $\mathcal{L}_{\beta^t}(x_1^n)$ and $\mathcal{L}_{\beta^{t+1}}(x_1^n)$, e.g.

$$T = \min\left\{ t > 0 \mid \mathcal{L}_{\beta^{t+1}}(x_1^n) - \mathcal{L}_{\beta^t}(x_1^n) \leq 10^{-3} \right\}$$

Specifying the EM algorithm with GMMs

In the case of GMM

Preliminary calculations

$$Q(\beta \mid \beta^t) = \sum_{i=1}^{n}\sum_{g=1}^{G} \log(f_{\theta_g}(x_i)\cdot \pi_g)\underbrace{\mathbb{P}_{\beta^t}\left(H_i = g \mid X_i = x_i\right)}_{=\tau_g^t(x_i)}$$

- From β^0: All $\tau_g^0(x_i)$s are fully computable!
- Using that $\log(f_{\theta_g}(x_i)\cdot\pi_g) = \log(f_{\theta_g}(x_i)) + \log(\pi_g)$

$$\sum_{i=1}^{n}\sum_{g=1}^{G} \log(f_{\theta_g}(x_i)\cdot \pi_g)\tau_g^t(x_i)$$

$$= \sum_{g=1}^{G}\left[\sum_{i=1}^{n}\log(f_{\theta_g}(x_i))\tau_g^t(x_i) + \log(\pi_g)\left(\sum_{i=1}^{n}\tau_g^t(x_i)\right)\right]$$

-

$$\log(f_{\theta_g}(x_i)) = -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(|\Sigma_g|\right) - \frac{1}{2}\left(x_i - \mu_g\right)^\top \Sigma_g^{-1}\left(x_i - \mu_g\right)$$

EM-algorithm for GMM

Algorithm (EM for Gaussian Mixtures)

1. *Initialize the iterative process with* $\beta = \beta^0$
2. *For* $t = 1, \ldots, T$*:*
   Apply
   - **E-step**: *Compute the* $\tau_g^t(x_i)$*s*
     - $Q(\beta \mid \beta^t) = \sum_{i=1}^n \sum_{g=1}^G \log(f_{\theta_g}(x_i) \cdot \pi_g)\tau_g^t(x_i)$
   - **M-step**: *Compute*

$$\pi_g^{t+1} = \frac{1}{n}\sum_{i=1}^n \tau_g^t(x_i), \qquad \mu_g^{t+1} = \sum_{i=1}^n x_i \frac{\tau_g^t(x_i)}{\sum_{j=1}^n \tau_g^t(x_j)}$$

$$\Sigma_g^{t+1} = \sum_{i=1}^n (x_i - \mu_g^{t+1})(x_i - \mu_g^{t+1})^\top \frac{\tau_g^t(x_i)}{\sum_{j=1}^n \tau_g^t(x_j)}$$

By differentiating $Q(\beta \mid \beta^t)$ *w.r.t each coordinate...*

Justifying the EM-algorithm

Why is that meaningful?

First remark

$$(\mathcal{L}_\beta(x_1^n) = Q(\beta \mid \beta^t) - R(\beta \mid \beta^t))$$

$$\mathcal{L}_\beta(x_1^n) - \mathcal{L}_{\beta^t}(x_1^n) = \left[\, Q(\beta \mid \beta^t) - Q(\beta^t \mid \beta^t) \,\right] - \left[\, R(\beta \mid \beta^t) - R(\beta^t \mid \beta^t) \,\right]$$

Kullback-Leibler divergence

$$- \left[\, R(\beta \mid \beta^t) - R(\beta^t \mid \beta^t) \,\right] = - \sum_{i=1}^n \int_{\mathbb{R}^d} \log\left(\frac{f_\beta^{H_i \mid X_i = x_i}(h)}{f_{\beta^t}^{H_i \mid X_i = x_i}(h)} \right) f_{\beta^t}^{H_i \mid X_i = x_i}(h)\, dh = \sum_{i=1}^n KL\left(f_{\beta^t}^{H_i \mid X_i = x_i}(h); f_\beta^{H_i \mid X_i = x_i}(h) \right) \geq 0$$

Maximizing $\beta \mapsto Q(\beta \mid \beta^t)$

$Q(\beta^{t+1} \mid \beta^t) - Q(\beta^t \mid \beta^t) \geq 0$ yields that

$$\mathcal{L}_{\beta^{t+1}}(x_1^n) - \mathcal{L}_{\beta^t}(x_1^n) \geq 0$$

$\Rightarrow \quad \left\{ \mathcal{L}_{\beta^t}(x_1^n) \right\}_{t>0}$ is nondecreasing...

EM algorithm: Pros and cons

Cons

- Results strongly depend on:
  - the initialization β^0 (warm starts)
  - numerous local maxima (no convexity property!)
  - ...
- Some classes can be emptied along the iterations (degeneracy)
- No convergence speed guarantee in general

Pros

- But ... this is the only available optimization algorithm in this context
- Can bare diffficult contexts with missing data...

Model selection

Penalizing candidate models

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f^\star$: unknown density
- $\mathcal{M}_d = \left\{ f_\beta \mid \beta \in \mathbb{R}^d \right\}$: candidate model indexed by d

Assumption

For simplicity, $\{\mathcal{M}_d\}_{1 \leq d \leq D}$ is an ordered family of models

Quality measure of model $\mathcal{M}_d$

Within model $\mathcal{M}_d$:

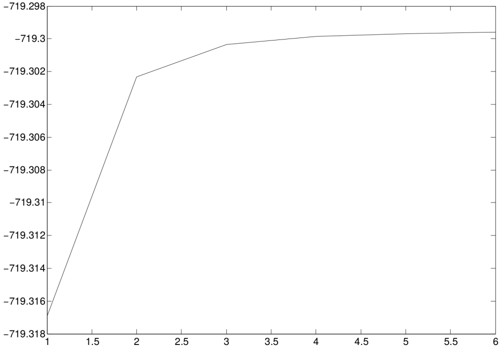
- $\mathcal{L}_\beta(X_1^n)$: likelihood of $\beta \in \mathbb{R}^d$
- $\widehat{\beta}_d^{MLE}$: Maximum Likelihood estimator within $\mathcal{M}_d$

Overfitting

- Never minimize $d \mapsto \mathcal{L}_{\widehat{\beta}_d^{MLE}}(X_1^n)$
- Would lead to overfitting since $-\mathcal{L}_\beta(X_1^n)$ is the empirical risk of β

Overfitting phenomenon

Log-likelihood versus the dimension



Number of parameters (dimension)

Maximizing $d \mapsto \mathcal{L}_{\widehat{\beta}_d^{MLE}}(X_1^n) \Rightarrow$ Choose the largest model

Penalized criteria: AIC

Akaike's Information Criterion (AIC)

- Reaches a good approximation to the density (based on the Kullback-Leibler divergence)
- Does not focus on recovering true mixture components $\to$ Can overestimate the number of components
- Only works with a limited number of models

Definition (AIC penalty)

$$AIC(d) = \mathcal{L}_{\widehat{\beta}_d^{MLE}}(x_1^n) - d, \quad \text{and} \quad \widehat{d} = Arg \max_{1 \leq d \leq D} \{AIC(d)\}$$

Remark:

- Pros: Easy to compute
- Cons: Bad behavior with ill-specified models

Generalizing AIC: TIC

Takeuchi's Information Criterion (TIC)

- Same goal as the one of AIC: Estimation and not identification...
- Only works with a limited number of models

Definition (TIC penalty)

$$TIC(d) = \mathcal{L}_{\widehat{\beta}_d^{MLE}}(x_1^n) - \mathrm{Tr}\left[\widehat{I}(\widehat{\beta}_d^{MLE}) \cdot \widehat{J}^{-1}(\widehat{\beta}_d^{MLE}) \right]$$

with

$$\widehat{I}(\widehat{\beta}_d^{MLE}) = \frac{1}{n}\sum_{i=1}^{n} \left[\partial_\beta \ell_\beta(x_i) \cdot \partial_\beta \ell_\beta(x_i)^\top \right]_{\beta=\widehat{\beta}_d^{MLE}}$$

$$\widehat{J}(\widehat{\beta}_d^{MLE}) = -\frac{1}{n}\sum_{i=1}^{n} \left[\partial_\beta^2 \ell_\beta(x_i) \right]_{\beta=\widehat{\beta}_d^{MLE}}$$

Remark:

- Deals with ill-specified models (more reliable than AIC)
- Cons: Somewhat more complex to calculate

Penalized criteria: BIC for identification

Bayesian Information Criterion (BIC)

- Looks for the best approximation to the truth among candidate models (for density estimation)
- Identification ($\neq$ estimation) purpose
  $\rightarrow$ Reliable estimate of the number of components *if the truth is among the candidate models*

Definition (BIC penalty)

$$BIC(d) = \mathcal{L}_{\widehat{\beta}_d^{MLE}}(x_1^n) - \frac{d}{2}\log(n)$$

Remark:

- Bad behavior with small sample size and/or non-Gaussian distributions (over-estimation of $d^\star$)
- Does not incorporate any structure knowledge regarding potential clusters

Extending BIC to clustering: ICL

The Integrated Complete-data Likelihood (ICL) criterion

Definition (ICL penalty)

$$ICL(d) = \mathcal{L}_{\widehat{\beta}_d^{MLE}}(x_1^n) - \frac{d}{2}\log(n) - ENT(\widehat{\beta}_d^{MLE})$$

where $ENT(\beta) = -\sum_{i=1}^n \sum_{g=1}^G \tau_g^\beta(x_i)\log(\tau_g^\beta(x_i)) \geq 0$
and $\tau_g^\beta(x_i) = \pi_g f_g(x_i)/\sum_h \pi_h f_h(x_i)$.

Remark:
$ENT(\beta) = -\sum_{i=1}^n \sum_{g=1}^G \tau_g^\beta(x_i)\log(\tau_g^\beta(x_i))$ is maximum for $\tau_g^\beta(x_i) = 1/G$ for all $1 \leq g \leq G$

Conclusion

- ICL: dedicated to the clustering task
- Better than BIC for estimating $d^\star$
- ICL looks for "clear clusters": For all is,

$$\tau_g^{\widehat{\beta}_d^{MLE}}(x_i) = \widehat{\mathbb{P}}[\, Y_i \in g \mid X_i = x_i \,] \approx 1 \quad \text{for some } g$$

Illustration: BIC/ICL behaviors