Clustering and classification risks in non-parametric Hidden Markov and I.I.D models

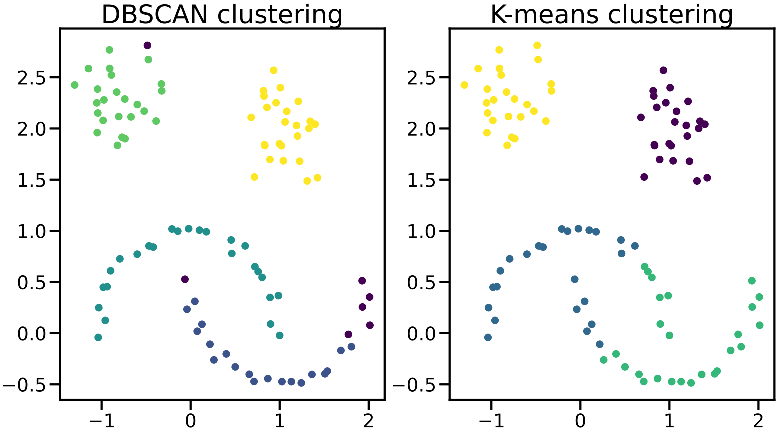
Ibrahim Kaddouri

Joint work with Elisabeth Gassiat and Zacharie Naulet

AHIDI2024 Workshop

Clustering

Clustering is an ill-posed problem which aims to find out interesting structures in the data or to derive a useful grouping of the observations.



Model-based clustering: Mixture models

Observations $Y = (Y_k)_{1 \leq k \leq n}$ coming from J populations.
Define latent variables $X = (X_k)_{1 \leq k \leq n}$ such that: for each k,

$$Y_k \mid X_k = j \sim f_j$$

Model-based clustering: Mixture models

Observations $Y = (Y_k)_{1 \leq k \leq n}$ coming from J populations.
Define latent variables $X = (X_k)_{1 \leq k \leq n}$ such that: for each k,

$$Y_k \mid X_k = j \sim f_j$$

Then Y_k has distribution

$$\sum_{j=1}^{J} \pi_j f_j$$

π_j: Probability to come from population j

Useful to model data coming from heterogeneous populations.

Mixture models: Identifiability

Mixture models are not identifiable :

$$\sum_{j=1}^{J} \pi_j f_j = \frac{\pi_1}{2} f_1 + \left(\frac{\pi_1}{2} + \pi_2\right) \left(\frac{\frac{\pi_1}{2} f_1 + \pi_2 f_2}{\frac{\pi_1}{2} + \pi_2}\right) + \sum_{j=3}^{J} \pi_j f_j$$

Mixture models: Identifiability

Mixture models are not identifiable :

$$\sum_{j=1}^{J} \pi_j f_j = \frac{\pi_1}{2} f_1 + \left(\frac{\pi_1}{2} + \pi_2\right) \left(\frac{\frac{\pi_1}{2} f_1 + \pi_2 f_2}{\frac{\pi_1}{2} + \pi_2}\right) + \sum_{j=3}^{J} \pi_j f_j$$

Learning of population components possible only under additional structural assumptions such as:

- Parametric mixtures
- Shape restrictions (gaussian, multinomial, ...)

$\rightarrow$ Might lead to poor results in practice

Hidden Markov Models and why they are useful

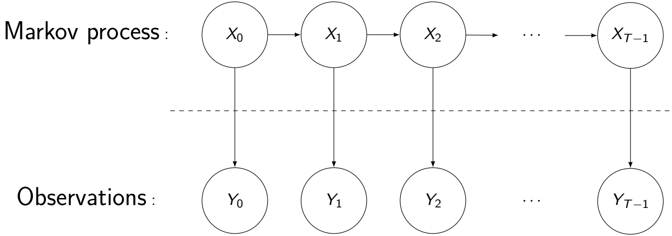


Figure: A Hidden Markov Model.

Latent (unobserved) variables $(X_k)_k$ form a Markov chain.
Observations $(Y_k)_k$ are independent conditionnally to $(X_k)_k$.

Hidden Markov Models and why they are useful

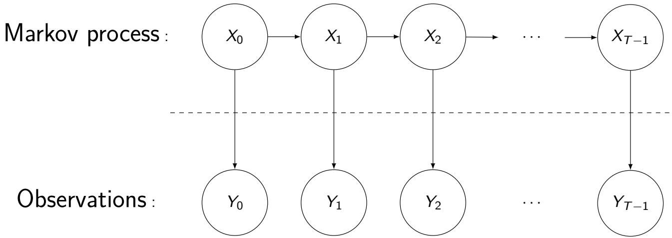


Figure: A Hidden Markov Model.

Latent (unobserved) variables $(X_k)_k$ form a Markov chain.
Observations $(Y_k)_k$ are independent conditionnally to $(X_k)_k$.

HMMs are identifiable without any shape restriction!

Outline

Risk of classification

Consider the classification loss function:

$$L_1(x'_{1:n}, x_{1:n}) = \frac{1}{n}\sum_{k=1}^{n} 1_{x'_k \neq x_k}$$

Risk of classification

Consider the classification loss function:

$$L_1(x'_{1:n}, x_{1:n}) = \frac{1}{n} \sum_{k=1}^{n} 1_{x'_k \neq x_k}$$

Let $\theta = \left(\nu, Q, (f_x)_{1 \leq x \leq J}\right)$ denote the model parameters.
The risk associated to a classifier $h = (h_i)_{1 \leq i \leq n}$ is:

$$\mathcal{R}_n^{class}(\theta, h) = \mathbb{E}_\theta[L_1(h(Y_{1:n}), X_{1:n})] = \mathbb{E}_\theta \left[\frac{1}{n} \sum_{i=1}^{n} 1_{h_i(Y_{1:n}) \neq X_i} \right]$$

Risk of classification

Consider the classification loss function:

$$L_1(x'_{1:n}, x_{1:n}) = \frac{1}{n} \sum_{k=1}^{n} 1_{x'_k \neq x_k}$$

Let $\theta = \left(\nu, Q, (f_x)_{1 \leq x \leq J}\right)$ denote the model parameters.
The risk associated to a classifier $h = (h_i)_{1 \leq i \leq n}$ is:

$$\mathcal{R}_n^{class}(\theta, h) = \mathbb{E}_\theta[L_1(h(Y_{1:n}), X_{1:n})] = \mathbb{E}_\theta \left[\frac{1}{n} \sum_{i=1}^{n} 1_{h_i(Y_{1:n}) \neq X_i} \right]$$

The Bayes risk of classification corresponds to $\inf_h \mathcal{R}_n^{class}(\theta, h)$ and the Bayes classifier has a closed formula:

$$h_\theta^\star = \left(\mathbb{P}_\theta \left(X_i = . \mid Y_{1:n}\right)\right)_{1 \leq i \leq n}$$

Risk of clustering

To measure the loss between two partitions A and B of $\{1, .., n\}$, we use the loss

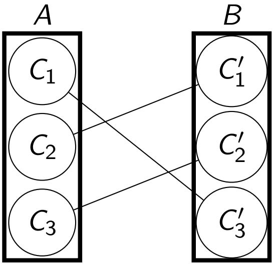
$$L_2(A, B) = 1 - \frac{1}{n} \sup_{\substack{M \subseteq \mathcal{E}(A,B) \\ M \text{ is a matching} \\ \text{between A and B}}} \sum_{\{C, C'\} \in M} \mathrm{Card}(C \cap C')$$

Risk of clustering

To measure the loss between two partitions A and B of $\{1, .., n\}$, we use the loss

$$L_2(A, B) = 1 - \frac{1}{n} \sup_{\substack{M \subseteq \mathcal{E}(A,B) \\ M \text{ is a matching} \\ \text{between A and B}}} \sum_{\{C, C'\} \in M} \mathrm{Card}(C \cap C')$$

where the supremum is over the set of matchings which are subsets of the edge set $\mathcal{E}(A, B) := \{\{C, C'\} \ : \ C \in A, \ C' \in B\}$.



Risk of clustering

We first define the map π_n by:

$$\pi_n(x_{1:n}) = \{\{i \ : \ x_i = a\} \ : \ a \in \{1, .., J\}\} \backslash \{\varnothing\}$$

Risk of clustering

We first define the map π_n by:

$$\pi_n(x_{1:n}) = \{\{i \ : \ x_i = a\} \ : \ a \in \{1, .., J\}\}\backslash\{\varnothing\}$$

The risk of a clusterer g can be defined as:

$$\mathcal{R}_n^{clust}(\theta, g) := \mathbb{E}_\theta \left[L_2(g(Y_{1:n}), \pi_n(X_{1:n}))\right]$$

where

- $\pi_n(X_{1:n})$ is the partition induced by the labels $X_{1:n}$
- $g(Y_{1:n})$ is the partition generated by the clusterer g

The Bayes risk of clustering corresponds to $\inf_g \mathcal{R}_n^{clust}(\theta, g)$.

Risk of clustering

We first define the map π_n by:

$$\pi_n(x_{1:n}) = \{\{i \ : \ x_i = a\} \ : \ a \in \{1, .., J\}\} \backslash \{\varnothing\}$$

The risk of a clusterer g can be defined as:

$$\mathcal{R}_n^{clust}(\theta, g) := \mathbb{E}_\theta \left[L_2(g(Y_{1:n}), \pi_n(X_{1:n})) \right]$$

where

- $\pi_n(X_{1:n})$ is the partition induced by the labels $X_{1:n}$
- $g(Y_{1:n})$ is the partition generated by the clusterer g

The Bayes risk of clustering corresponds to $\inf_g \mathcal{R}_n^{clust}(\theta, g)$.

Questions:

Risk of clustering

We first define the map π_n by:

$$\pi_n(x_{1:n}) = \{\{i \ : \ x_i = a\} \ : \ a \in \{1, .., J\}\} \backslash \{\varnothing\}$$

The risk of a clusterer g can be defined as:

$$\mathcal{R}_n^{clust}(\theta, g) := \mathbb{E}_\theta \left[L_2(g(Y_{1:n}), \pi_n(X_{1:n})) \right]$$

where

- $\pi_n(X_{1:n})$ is the partition induced by the labels $X_{1:n}$
- $g(Y_{1:n})$ is the partition generated by the clusterer g

The Bayes risk of clustering corresponds to $\inf_g \mathcal{R}_n^{clust}(\theta, g)$.

Questions:

- Is there any relationship between the Bayes classifier and the Bayes clusterer? If so, under what condition?

Risk of clustering

We first define the map π_n by:

$$\pi_n(x_{1:n}) = \{\{i \ : \ x_i = a\} \ : \ a \in \{1, .., J\}\} \backslash \{\varnothing\}$$

The risk of a clusterer g can be defined as:

$$\mathcal{R}_n^{clust}(\theta, g) := \mathbb{E}_\theta \left[L_2(g(Y_{1:n}), \pi_n(X_{1:n})) \right]$$

where

- $\pi_n(X_{1:n})$ is the partition induced by the labels $X_{1:n}$
- $g(Y_{1:n})$ is the partition generated by the clusterer g

The Bayes risk of clustering corresponds to $\inf_g \mathcal{R}_n^{clust}(\theta, g)$.

Questions:

- Is there any relationship between the Bayes classifier and the Bayes clusterer? If so, under what condition?
- Under what condition do the Bayes risk of classification and the Bayes risk of clustering have the same magnitude? In what sense?

Relationship between the minimizers

Let J the number of hidden states. Let Θ^{ind} the set of parameters for which observations are independent (all the lines of the transition matrix Q are equal,...) and let Θ^{dep} be the set of the remaining parameters. We recall that $g^{\star}_{\theta}$ is the Bayes clusterer and $h^{\star}_{\theta}$ the Bayes classifier.

Relationship between the minimizers

Let J the number of hidden states. Let Θ^{ind} the set of parameters for which observations are independent (all the lines of the transition matrix Q are equal,...) and let Θ^{dep} be the set of the remaining parameters. We recall that $g^\star_\theta$ is the Bayes clusterer and $h^\star_\theta$ the Bayes classifier.

Theorem

If $J = 2$, then for all $\theta \in \Theta^{\mathrm{ind}}$ and all $n \geq 2$.

$$g^\star_\theta(Y_{1:n}) = \pi_n \circ h^\star_\theta(Y_{1:n}) \quad \mathbb{P}_\theta\text{-}a.s.$$

Relationship between the minimizers

Let J the number of hidden states. Let Θ^{ind} the set of parameters for which observations are independent (all the lines of the transition matrix Q are equal,...) and let Θ^{dep} be the set of the remaining parameters. We recall that $g^\star_\theta$ is the Bayes clusterer and $h^\star_\theta$ the Bayes classifier.

Theorem

If $J = 2$, then for all $\theta \in \Theta^{\mathrm{ind}}$ and all $n \geq 2$.

$$g^\star_\theta(Y_{1:n}) = \pi_n \circ h^\star_\theta(Y_{1:n}) \quad \mathbb{P}_\theta\text{-}a.s.$$

Theorem

If $J > 2$ or $\theta \in \Theta^{\mathrm{dep}}$, then for all $n \geq 2$.

$$\mathbb{P}_\theta\left(g^\star_\theta(Y_{1:n}) \neq \pi_n \circ h^\star_\theta(Y_{1:n})\right) > 0.$$

Relationship between the Bayes risks

Theorem

Assume $\delta = \min_{i,j} Q_{i,j} > 0$. *For* $J = 2$ *and* $\theta \in \Theta^{\mathrm{ind}} \cup \Theta^{\mathrm{dep}}$, *there exist* $c, c', \beta > 0$ *depending only on* δ *such that*

$$\left(1 - \frac{c}{\sqrt{n}}\right) \inf_h \mathcal{R}_n^{class}(\theta, h) \leq \inf_g \mathcal{R}_n^{clust}(\theta, g) \leq \inf_h \mathcal{R}_n^{class}(\theta, h)$$

Relationship between the Bayes risks

Theorem

Assume $\delta = \min_{i,j} Q_{i,j} > 0$. *For* $J = 2$ *and* $\theta \in \Theta^{\mathrm{ind}} \cup \Theta^{\mathrm{dep}}$, *there exist* $c, c', \beta > 0$ *depending only on* δ *such that*

$$\left(1 - \frac{c}{\sqrt{n}}\right) \inf_h \mathcal{R}_n^{class}(\theta, h) \leq \inf_g \mathcal{R}_n^{clust}(\theta, g) \leq \inf_h \mathcal{R}_n^{class}(\theta, h)$$

For $J > 2$ *and* $\theta \in \Theta^{\mathrm{ind}} \cup \Theta^{\mathrm{dep}}$ *and all* $n \geq 1$

$$\left(1 - \frac{c'}{\sqrt{n}}\right) \inf_h \mathcal{R}_n^{class}(\theta, h) - J^2 e^{-n\beta} \leq \inf_g \mathcal{R}_n^{clust}(\theta, g) \leq \inf_h \mathcal{R}_n^{class}(\theta, h)$$

Analyzing the Bayes risk of clustering

Theorem

Assume $\delta = \min_{i,j} Q_{i,j} > 0$. *Then,*

- *When* $J = 2$

$$\delta(1-\alpha_n)\int f_0 \wedge f_1 \leq \inf_g \mathcal{R}_n^{clust}(\theta, g) \leq (1-\delta)\int f_0 \wedge f_1$$

- *When* $J > 2$

$$\delta(1-\alpha_n)\Lambda - J^2 e^{-n\beta} \leq \inf_g \mathcal{R}_n^{clust}(\theta, g) \leq (1-\delta)\,\Lambda$$

where α_n *decays to* 0 *and* β *depends on* δ *and* J *and*

$$\Lambda = \int \min_{1\leq x_0 \leq J} \sum_{x \neq x_0} f_x(y)dy$$

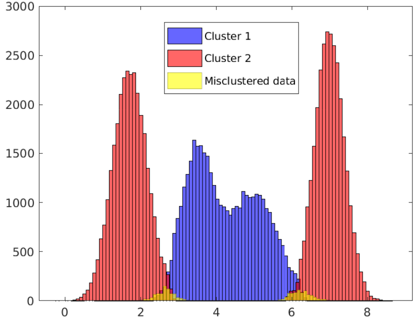
Examples where HMMs are useful

Data are generated through the same transition matrix $Q = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}$.

- **First example:** A sample of size $n = 5.10^4$ is generated from two gaussian mixtures : $\frac{1}{2}\left(\mathcal{N}(1.7, 0.2) + \mathcal{N}(7, 0.15)\right)$ and $\frac{1}{2}\left(\mathcal{N}(3.5, 0.2) + \mathcal{N}(5, 0.4)\right)$.
- **Second example:** A sample of size $n = 10^5$ is generated from two gaussian mixtures : $\frac{1}{2}\left(\mathcal{N}(3, 0.6) + \mathcal{N}(7, 0.4)\right)$ and $\frac{1}{2}\left(\mathcal{N}(5, 0.3) + \mathcal{N}(9, 0.4)\right)$.

Purpose: Retrieve the sequence of hidden states using only the observations.

Example 1



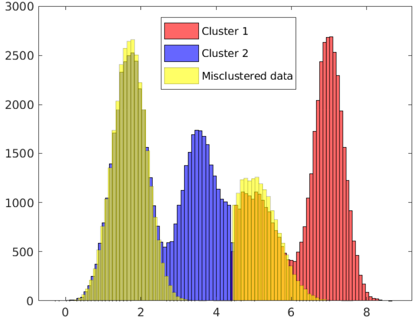
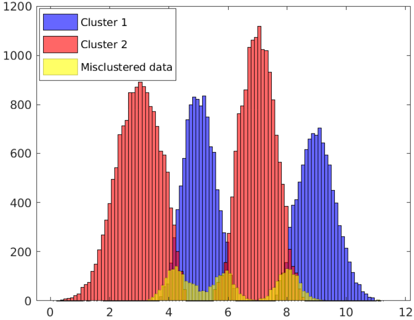


Figure: Histograms of the clusters. Left: clustering using plug-in classifier. Right: K-means clustering

Example 2



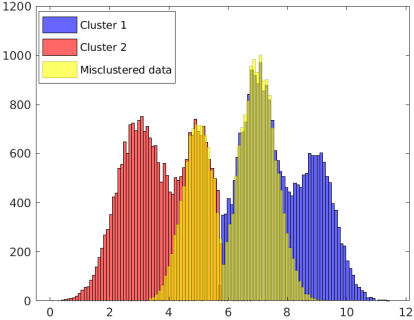


Figure: Histograms of the clusters. Left: clustering using plug-in classifier. Right: K-means clustering

Clustering errors

	Bayes classifier	Plug-in classifier	K-means algorithm
Example 1	1.56%	1.61%	46.7%
Example 2	6.42%	6.51%	47.3%

Table: Errors of clustering using 3 algorithms: the Bayes classifier (using the true model parameters), the plug-in classifier (using the estimated parameters) and the K-means algorithm.