Model-based Clustering of Non-parametric Hidden Markov Models

Elisabeth Gassiat, Ibrahim Kaddouri, Zacharie Naulet
Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay

Problem

Clustering observations of a Hidden Markov Model where the emission distributions are non-parametric. The purpose is to understand the dependence of the Bayes risk of clustering with respect to the separation between the emission distributions, the transition matrix of the hidden states being fixed.

Goal:

1. Identify the appropriate notion of separation between the emission distributions which measures the difficulty of the problem of clustering.
2. Given a specified level on the risk of clustering, what is the required separation to ensure the Bayes risk of clustering is smaller than the pre-specified threshold?
3. Construct an optimal clustering procedure

Motivation and prior work

Clustering is usually applied to heterogeneous data coming from different populations. Usually, they are modeled by a **mixture model**. However, without any additional assumption, this model is not identifiable. Inference of the HMM parameters and clustering are not possible. However, when the data are derived from a HMM, the model becomes **identifiable** [3] and estimation of the mixture components at **optimal rates** is possible [1].

Mathematical setting

- The hidden states $(X_k)_{k\in\mathbb{N}}$ are assumed to form a **Markov chain** with J hidden states, initial distribution ν and transition matrix Q.
- The observations $(Y_k)_{k\in\mathbb{N}}$ are **independent conditionally to the hidden states** and $Y_k \mid X_k = j \sim F_j$. $(F_j)_{1\le j\le J}$ are called emission distributions.
- $\theta = (\nu, Q, (F_j)_{1\le j\le J})$ will denote the model parameters and $\mathcal{S}_J$ the set of permutations of $\{1, \dots, J\}$.

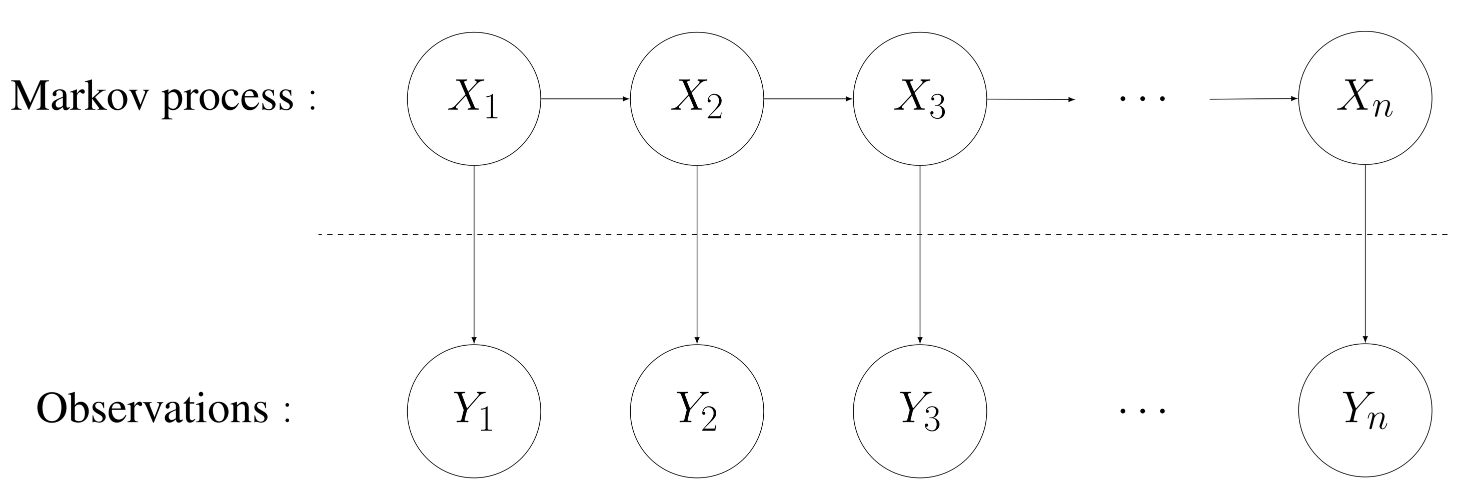

Figure 1: A hidden Markov model.

A classifier is of the form: $h(Y_{1:n}) = (h_i(Y_{1:n}))_{1\le i\le n}$. The associated risk of classification is:

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

Unlike classification, clustering seeks only to identify the clusters or the partitions, but not the associated labels. To each classifier h is associated a clustering procedure g which seeks to infer only the partition, not the labels of the classes themselves:

$$g(Y_{1:n}) = \{\{i \ : \ h_i(Y_{1:n}) = x\} \ : \ x \in \mathbb{X}\}\backslash\{\varnothing\}$$

The associated risk of clustering can then be defined by:

$$\mathcal{R}_n^{\text{clust}}(\theta, g) := \mathbb{E}_\theta\left[\inf_{\tau\in\mathcal{S}_J}\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\tau(X_i)\neq h_i(Y_{1:n})}\right]$$

Quantifying the Bayes risk of clustering

Theorem 1 *Assume $\delta = \min_{i,j} Q_{i,j} > 0$. Then:*

- If $J = 2$, then for all $n \ge 1$ and all θ:

$$\inf_g \mathcal{R}_n^{\text{clust}}(\theta, g) \ge \left[1 - c_1\sqrt{\frac{\log(J)}{2n}}\right]\inf_h \mathcal{R}_n^{\text{class}}(\theta, h)$$

- If $J \ge 3$, then for all $n \ge 1$ and all θ:

$$\inf_g \mathcal{R}_n^{\text{clust}}(\theta, g) \ge \left[1 - c_2\sqrt{\frac{\log(J!)}{2n}}\right]\inf_h \mathcal{R}_n^{\text{class}}(\theta, h) - (J^2+1)e^{-c_3 n}$$

where c_1, c_2 and c_3 are positive constants depending only on ν and Q.

- Complete equivalence between the two risks for n large enough when the HMM is comprised on only two hidden states.
- One can show that the equivalence does not hold in all generality.
- Since the Bayes risk of classification (smallest risk) has a closed formula:

$$\inf_h \mathcal{R}_n^{\text{class}}(\theta, h) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta\left[\min_{1\le j\le J}\mathbb{P}_\theta\left(X_i \neq j \mid Y_{1:n}\right)\right]$$

it is easier to study than the Bayes risk of clustering. Thanks to some recursive formulas [2] ensured by the distributions $(\mathbb{P}_\theta\left(X_i \in . \mid Y_{1:n}\right))_{1\le i\le n}$, we prove that the difficulty of the problem of clustering HMM observations is driven by:

$$\Lambda = \int \min_{1\le j\le J}\left(\sum_{i\neq j} F_i\right)$$

Theorem 2 *If $\delta = \min_{i,j} Q_{i,j} > 0$ and $\Lambda \ge \beta e^{-c_3 n}$:*

$$\alpha\Lambda \le \inf_g \mathcal{R}_n^{\text{clust}}(\theta, g) \le \Lambda$$

where α and β are positive constants depending only on ν, Q and J.

Optimality of plugin procedure

We consider now the plugin procedure defined by:

$$h_{\hat\theta}(Y_{1:n}) = \left(\underset{1\le j\le J}{\arg\max}\,\mathbb{P}_{\hat\theta(Y_{1:n})}\left(X_i = j \mid Y_{1:n}\right)\right)_{1\le i\le n}$$

Theorem 3 *Assume the transition matrix is ergodic, the emission densities are linearly independent and Hölder smooth. Then, the plugin procedure ensures:*

$$\mathcal{R}_n^{\text{clust}}(\theta, g_{\hat\theta}) - \inf_g \mathcal{R}_n^{\text{clust}}(\theta, g) = \mathcal{O}\left(\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}\right)$$

Numerical simulations

Consider the following 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 $\frac{1}{2}\left(\mathcal{N}(3.5, 0.2) + \mathcal{N}(5, 0.4)\right)$ with a stationary hidden chain generated by: $Q = \begin{pmatrix}0.8 & 0.2\\ 0.3 & 0.7\end{pmatrix}$.

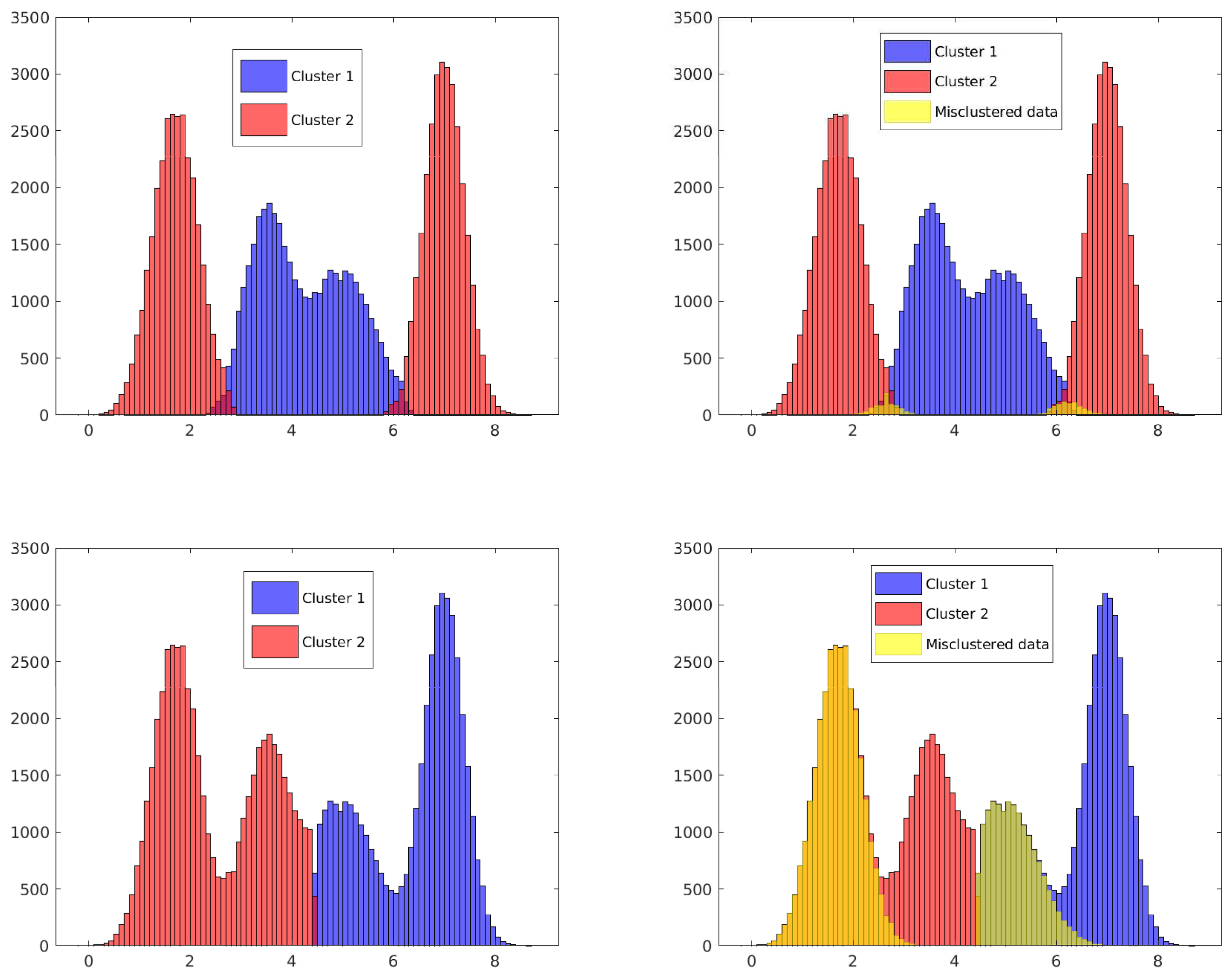

Figure 2: Histograms of clustering results for plugin (top) and k-means (bottom)

Bayes classifier	Plug-in classifier	k-means algorithm	Λ
1.56%	1.61%	46.7%	0.046

Table 1: Errors of clustering for three clustering procedures

Conclusion

1. Non-parametric Hidden Markov Models are capable of handling model-based clustering without any specification of clusters distributions.
2. Unlike clustering algorithms which are purely geometric (such as k-means and its variants), plugin procedures exploit the distribution of the observations and are particularly well-suited to model-based clustering.
3. The measure of separation Λ allows a clear understanding of the difficulty of the task of clustering with respect to the emission distributions.

References

[1] Kweku Abraham, Ismaël Castillo, and Elisabeth Gassiat. Multiple testing in nonparametric hidden markov models: An empirical bayes approach. *Journal of Machine Learning Research*, 23(1):4061–4117, 2022.

[2] Olivier Cappé, Eric Moulines, and Tobias Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag, 2005.

[3] E. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden markov models and applications. *Stat. Comput.*, 26(1-2):61–71, 2016.