Internship topic: Clustering of dependent data.
Advisor: Zacharie Naulet (zacharie.naulet@inrae.fr).
Host laboratory: Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

1 Context

Clustering is the unsupervised tasks aiming at grouping data into homogeneous subgroups. It is used as an exploratory tool to detect structure and uncover latent variables that explain the observed heterogeneity. Clustering is applied in various fields, such as marketing, bioinformatics, psychology and sociology, epidemic surveillance, etc.

Mixture models are a popular statistical framework for clustering. In a mixture model, the data $Y_1, \dots, Y_n$ is assumed to be conditionally independent given an unobserved (*aka.* latent) sequence $X_1, \dots, X_n \in \{1, \dots, K\}$, which represents the subgroup membership of each observation. Specifically, it is postulated that $Y_i \mid X_1, X_2, \dots \overset{law}{=} Y_i \mid X_i \sim F_{X_i}$, where $F_1, \dots, F_K$ denote distinct probability distributions characterizing the subgroups. If the latent variables $X_1, X_2, \dots$ are also independent, then recovering the subgroups without prior assumptions is imposible. This is why most of the scientific literature focuses on mixtures of specific distributions, such as $F_1, \dots, F_K$ being Gaussian distributions. Unfortunately, constraining the form of the distributions composing the mixture can be unrealistic in certain applications, and overcoming this limitation has driven both theoretical and computational research.

A situation where it is possible to recover the subgroups without any prior assumption arises when the latent variables $X_1, X_2, \dots$ are dependent and form a Markov chain. Such mixture model is called a *Hidden Markov Model* (HMM). Theoretical guarantees for the nonparametric learning of HMM have been developed in recent years [5, 4, 8, 1, 2], while clustering has been investigated in [7].

One of the recurrent criticism of HMMs is the lack of realism of the Markovian nature of the latent structure. Yet, having a dependent latent structure is well-accepted in certain applications (for instance, genomics). It has been recently shown that *location* mixtures are nonparametrically identifiable as soon as two consecutive observations are dependent, regardless the type of dependency [6]. This suggests that spectral methods [3] can be used for nonparametric learning and for clustering, beyond Markovian dependencies. The goal of this internship is to study nonparametric estimation and clustering in generally dependent location mixtures.

The internship is motivated by applications in *genomics* where data are inherently dependent because biological sequences, like DNA and RNA, are organized in linear structures where the state or content at one position often influences neighboring positions. These dependencies are critical for understanding biological phenomena, making it essential to use specialized models to analyze and interpret the complex relationships within the data.

2 Objectives and research work

The research work can be organized around the following steps, depending on the student's affinities:

- Develop nonparametric estimators for dependent location mixtures.
- Develop a clustering procedure.
- Study the theoretical properties of the spectral method for estimating and clustering dependent location mixtures.

- Implement the proposed method.
- Compare the performance of the proposed method to misspecified existing methods.
- Extend the framework to general mixtures (non necessarily location mixtures)

Depending on the outcomes of the internship, the work may lead to a PhD with guaranteed funding.

3 Desired profile

Candidates should have a BAC+5-level education (Master's degree or engineering school), skills in theoretical and computational statistics, proficiency in a programming language is welcome, along with scientific rigor, intellectual curiosity, and strong communication skills.

4 Practical details

The internship will take place at the INRAE center in Jouy-en-Josas within the MaIAGE unit. The duration of the internship will be five to six months, between February and September 2025. The monthly stipend is approximately 550 euros (legal rate). The internship will be supervised by Zacharie Naulet (Université Paris-Saclay, INRAE, MaIAGE) and Estelle Kuhn (Université Paris-Saclay, INRAE, MICS). This internship may potentially lead to a funded PhD topic combining mathematical and computational statistics with applications in life sciences.

References

[1] K. Abraham, E. Gassiat, and Z. Naulet. Fundamental limits for learning hidden markov model parameters. *IEEE Transactions on Information Theory*, 69(3):1777–1794, 2022.

[2] K. Abraham, E. Gassiat, and Z. Naulet. Frontiers to the learning of nonparametric hidden markov models. *arXiv preprint arXiv:2306.16293*, 2023.

[3] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In *Conference on learning theory*, pages 33–1. JMLR Workshop and Conference Proceedings, 2012.

[4] Y. De Castro, C. Lacour, et al. Minimax adaptive estimation of nonparametric hidden markov models. *Journal of Machine Learning Research*, 17(111):1–43, 2016.

[5] É. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, 26:61–71, 2016.

[6] E. Gassiat, S. Le Corff, and L. Lehéricy. Identifiability and consistent estimation of nonparametric translation hidden markov models with general state space. *Journal of Machine Learning Research*, 21(115):1–40, 2020.

[7] E. Gassiat, I. Kaddouri, and Z. Naulet. Clustering and classification risks in nonparametric hidden markov and i.i.d. models. *arXiv preprint arXiv:2309.12238*, 2023.

[8] L. Lehéricy. *Estimation adaptative pour les modèles de Markov cachés non paramétriques*. PhD thesis, Université Paris-Saclay, 2018.