Clustering multivariate functional data using unsupervised binary trees

Steven Golovkine[1] **and Nicolas Klutchnikoff**[2] **and Valentin Patilea**[3]

[1]*Groupe Renault & CREST - UMR 9194, Rennes, France*

[2]*Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France*

[3]*Ensai, CREST - UMR 9194, Rennes, France*

Abstract

We propose a model-based clustering algorithm for a general class of functional data for which the components could be curves or images. The random functional data realizations could be measured with error at discrete, and possibly random, points in the definition domain. Based on [1], the idea is to build a set of binary trees by recursive splitting of the observations. At each node of the tree, a model selection test is performed, after expanding the multivariate functional data into a well chosen basis. We consider the Multivariate Functional Principal Component basis, developed in [2]. Similarly to [3], using the Bayesian Information Criterion, we test whether there is evidence that the data structure is a mixture model or not at each node of the tree. The number of groups are determined in a data-driven way and does not have to be pre-specified before the construction of the tree. Moreover, the tree structure allows us to consider only a small number of basis functions at each node. The new algorithm provides easily interpretable results and fast predictions for online data sets. Results on simulated datasets reveal good performance in various complex settings. The methodology is applied to the analysis of vehicle trajectories on a German roundabout. The open-source implementation of the algorithm can be accessed at `https://github.com/StevenGolovkine/FDApy`. Complete version of the work is available at arxiv:2012.05973.

Keywords: Gaussian mixtures, Model-based clustering, Multivariate Functional Principal Components

AMS subject classifications: 62R10

Acknowledgements: The authors wish to thank Groupe Renault and the ANRT for their financial support via the CIFRE convention no. 2017/1116.

References

[1] Fraiman, R., Ghattas, B. and Svarc, M. (2013). Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification*, 7.

[2] Happ, C. and Greven, S. (2018). Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association* 113 649–659.

[3] Pelleg, D. and Moore, A. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *In Proceedings of the 17th International Conf. on Machine Learning* 727–734.