

Multivariate functional data clustering using unsupervised binary trees

Steven Golovkine^{1,2}, Nicolas Klutchnikoff³, Valentin Patilea²

¹Groupe Renault

²CREST, Ensai

³IRMAR, Université Rennes 2

EYSM 2021

Virtual Conference, 6-10 September 2021



Functional Data Analysis

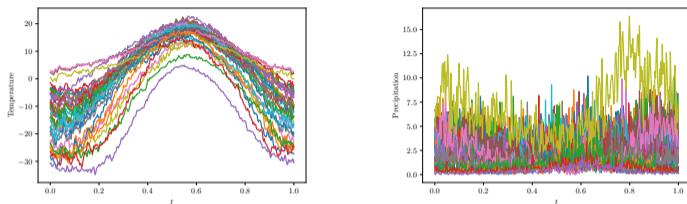


Figure 1: Canadian weather dataset (Ramsay and Silverman, 2005)

Examples

- ▶ Spectroscopy;
- ▶ Sounds recognition;
- ▶ Electroencephalography comparison;
- ▶ Various sensors.

Model

- ▶ Let

$$\mathcal{T} := [0, 1] \quad \text{and} \quad \mathcal{H} := L^2(\mathcal{T}) \times \cdots \times L^2(\mathcal{T}).$$

- ▶ We are interested by independent realizations of the P -dimensional stochastic process

$$X = \left\{ (X^{(1)}(t_1), \dots, X^{(P)}(t_P)) : t_1, \dots, t_P \in \mathcal{T} \right\}$$

taking values in \mathcal{H} .

- ▶ Note $\langle\langle \cdot, \cdot \rangle\rangle$ the inner product in \mathcal{H} .
- ▶ We aim to develop a clustering procedure to find some meaningful partition of realizations of the process X .

A mixture model for curves

- ▶ Let K be a positive integer, and let Z be a discrete random variable taking values in $\{1, \dots, K\}$ such that

$$\mathbb{P}(Z = k) = p_k \quad \text{with} \quad p_k > 0 \quad \text{and} \quad \sum_{k=1}^K p_k = 1.$$

- ▶ We consider that the stochastic process X admits the following decomposition:

$$X(\mathbf{t}) = \sum_{k=1}^K \mu_k(\mathbf{t}) \mathbf{1}_{\{Z=k\}} + \sum_{j \geq 1} \xi_j \phi_j(\mathbf{t}), \quad \mathbf{t} \in \mathcal{T},$$

where

- ▶ $\mu_1, \dots, \mu_K \in \mathcal{H}$ are the mean curves per cluster.
- ▶ $\{\phi_j\}_{j \geq 1}$ in an orthonormal basis of \mathcal{H} .
- ▶ For each $1 \leq k \leq K$, $\xi_j | Z = k \sim \mathcal{N}(0, \sigma_{kj}^2)$, for all $j \geq 1$.

Lemma

Assume X admits the previous decomposition. Let $\{\psi_j\}_{j \geq 1}$ be another orthonormal basis in \mathcal{H} and consider

$$c_j = \langle\langle X - \mu, \psi_j \rangle\rangle, \quad j \geq 1 \quad \text{where} \quad \mu(\cdot) = \sum_{k=1}^K p_k \mu_k(\cdot).$$

Then,

$$c_j | Z = k \sim \mathcal{N}(m_{kj}, \tau_{kj}^2),$$

where

$$m_{kj} = \langle\langle \mu_k - \mu, \psi_j \rangle\rangle \quad \text{and} \quad \tau_{kj}^2 = \sum_{l \geq 1} \langle\langle \phi_l, \psi_j \rangle\rangle^2 \sigma_{kl}^2.$$

► The clusters will be preserved after expressing the realizations of the process into an orthonormal basis.

The data

- ▶ Let $X_n, n \in \{1, \dots, N\}$ be independent trajectories of X .
- ▶ In practice, such trajectories cannot be observed at any \mathbf{t} .
- ▶ Moreover, only noisy data are available:
 - ▶ the observed values on the trajectory $X_n(\cdot)$ are contaminated with additive errors.
- ▶ For any $1 \leq n \leq N, 1 \leq p \leq P$, we observe $M_n^{(p)} \geq 2$ random pairs $(T_{n,m}^{(p)}, Y_{n,m}^{(p)})$ which are defined as:

$$Y_{n,m}^{(p)} = X_n^{(p)}(T_{n,m}^{(p)}) + \epsilon_{n,m}^{(p)}, \quad m = 1, \dots, M_n^{(p)}$$

where

- ▶ $(T_{n,1}^{(p)}, \dots, T_{n,M_n}^{(p)})$ are i.i.d. random sampling points in \mathcal{T} ;
- ▶ $\epsilon_{n,m}^{(p)}$ are i.i.d. random errors.

Example of such data

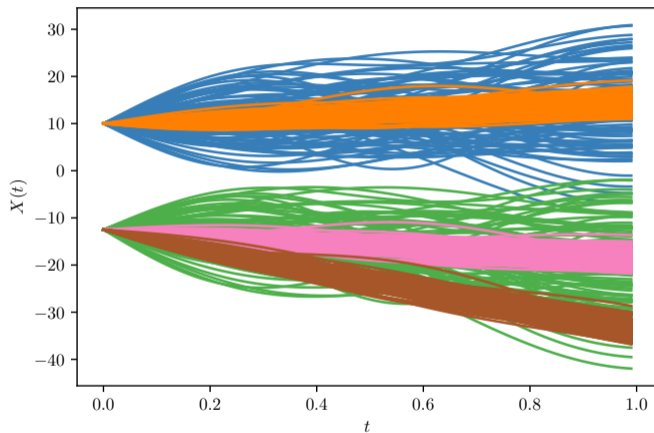


Figure 2: Example of data.

- ▶ Let $\mathcal{S} = \{X_1, \dots, X_N\}$ be a sample of realizations of the process X .
- ▶ We consider the problem of learning a meaningful partition \mathcal{U} of \mathcal{S} .
- ▶ For that, the idea is to build a full binary tree using a topdown procedure by recursive splitting.
- ▶ The procedure is based on Fraiman et al. (2010), adapted to functional data.
- ▶ The splitting criterion is similar to the one from Pelleg and Moore (2000).

How to split a node?

Given a training sample \mathcal{S} of realizations of X .

1. Perform a MFPCA with n_{comp} components and get the associated eigenvalues and eigenfunctions Φ .
2. Build the matrix C of the projection of the element of \mathcal{S} onto the elements Φ .
3. For each $k = 1, \dots, K_{\text{max}}$, fit a k -components GMM using an EM algorithm on the columns of C . The models are denoted by $\{\mathcal{M}_1, \dots, \mathcal{M}_{K_{\text{max}}}\}$.
4. Estimate the number of mixture components \hat{K} as

$$\hat{K} = \arg \max_{k=1, \dots, K_{\text{max}}} \text{BIC}(\mathcal{M}_k).$$

5. If $\hat{K} > 1$, we split the node in two using the model \mathcal{M}_2 .

- ▶ The construction of a branch of the tree is stopped if one of the following criterion is true:
 - ▶ The estimation of K is equal to 1.
 - ▶ There are less than `minsize` elements in the node.
- ▶ Three hyperparameters have to be set by the user:
 - ▶ `ncomp` – The number of components to keep for the MFPCA.
 - ▶ `Kmax` – The maximum number of components to consider for the mixture model.
 - ▶ `minsize` – The minimal number of elements in a node to be considered to be split.

Example of a tree

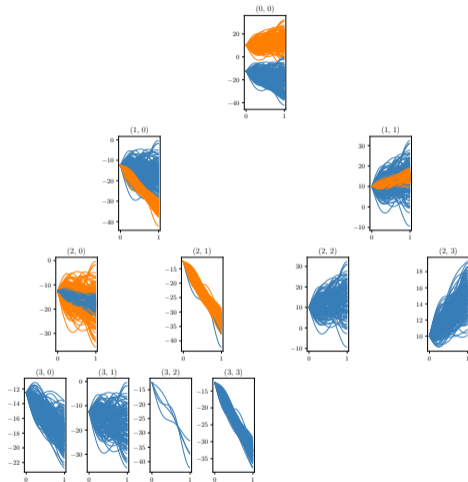


Figure 3: Example of a grown tree.

How to join nodes?

Given a set of terminal nodes V from the construction of the tree.

1. Build the graph $\mathcal{G} = (V, E)$ such that

$$E = \{(A, B) | A, B \in V, A \neq B \text{ and } \hat{K}_{A \cup B} = 1\}.$$

2. Associate to each element of E the value of the BIC that corresponds to $\hat{K}_{A \cup B}$.
3. Remove the edge with the maximum BIC value and replace the associated vertices by their union.
4. Continue the procedure by applying **1.** with

$$V = \{V \setminus \{A, B\}\} \cup \{A \cup B\}$$

until E is empty or V is reduced to a unique element.

Example of Canadian weather dataset

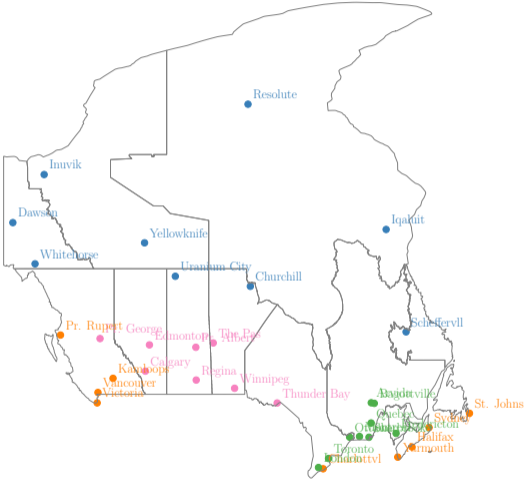


Figure 4: Clustering results using fCUBT

Takeaway ideas

- ▶ Model-based clustering of functional data:
 - ▶ multivariate functional data in both input and output dimension;
 - ▶ noisy data;
 - ▶ random discrete measurement points;
 - ▶ unknown number of groups.
- ▶ Prediction for new observation is easy.
- ▶ A preprint of the paper is available at
<https://arxiv.com/abs/2012.05973>
- ▶ An implementation of the fCUBT procedure is available at
<https://github.com/StevenGolovkine/FDApy>

References I

- Fraiman, R., Ghattas, B., and Svarc, M. (2010). Clustering using Unsupervised Binary Trees: CUBT. *Computing Research Repository - CORR*.
- Pelleg, D. and Moore, A. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *In Proceedings of the 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann.
- Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition.