

Clustering multivariate functional data defined on random domains: an application to vehicle trajectories analysis

S. GOLOVKINE

National School for Statistics and Information Analysis (ENSAI)

Supervisor(s): Prof. Patilea (ENSAI), Prof. Klutchnikoff (University Rennes 2) and Dr. Cembrzynski (Renault)

Ph.D. expected duration: Jan. 2018 - Dec. 2020

Address: 3, rue Philibert Delorme, 78280 Guyancourt, FRANCE

Email: steven.s.golovkine@renault.com

Abstract:

With the recent development of sensing devices, more and more data are recorded continuously (or at least at high frequency) through time and space. These measures lead large to amounts of data under the form of curves. In the statistical modeling this type of data is usually called *functional data*. We retrieve functional data in large variety of domains. For example, in biology, growth curves have probably been the first dataset considered as functional data [3]. But, one can find application in physics (spectroscopy), economics (index evolution), musics (sounds recognition), medicine (electroencephalography comparison), and so on. Lately, multivariate functional data have been considered. For instance, one can cite two famous examples from Ramsey and Silverman [2], gait cycle data and Canadian weather data. Moreover, the automotive industry also generates large volume of multivariate functional data. In particular, vehicle trajectories, which are our case of interest, could be naturally represented as multivariate functional data. Functional data analysis (FDA) develops the theory and statistical methodology for studying such data. By analogy with multivariate data analysis where an observation is represented by a random vector of scalars, in FDA, an observation is a random vector of functions. Hence, functional data are intrinsically infinite dimensional. However, we can not generally observed directly the functions but only a discretization of the functions over a fixed or random grid of points.

Now, recall our case of interest: vehicle trajectories. Nowadays, a vehicle records a lot of information about its environment through his different sensors (camera, radar, lidar). More particularly, it registers some characteristics about vehicles around him at high frequency. These characteristics can be the longitudinal and lateral position, the acceleration, the size, the type of the vehicle for instance. All the information are recorded relatively to the considered vehicle (also known as EGO car). Define a driving scene as a small period of time, say \mathcal{T} , during which we record the environment of the EGO car. This environment is constituted by a certain number of vehicles, say P , whose one records a certain number of characteristics for each vehicle, say D . However, we do not assume that all of the P vehicles are recorded on the complete interval \mathcal{T} , but only on a random compact subset of \mathcal{T} . So, an observation of a scene can be represented as a random vector of functions :

$$\mathbf{Z} = \left(Z^{(1)}, \dots, Z^{(P)} \right), \quad \text{where } \forall i \in \llbracket 1, P \rrbracket, Z^{(i)} : \mathcal{T}^{(i)} \subset \mathcal{T} \longrightarrow \mathbb{R}^D.$$

Moreover, $Z^{(i)}$ is assumed to be in $L^2(\mathcal{T}^{(i)})$ for all $i \in \llbracket 1, P \rrbracket$. So, realizations of \mathbf{Z} are multivariate functional data which are defined on different domains. The analysis of such data is performed in three major steps: smoothing, dimension reduction and then clustering.

The smoothing step has two major goals. The first one is to remove the eventual noise in the measurements because the sensors are not perfect and they can not retrieve exactly the reality.

Secondly, as the functions are defined on different domains, we use change-of-time methods to put them on a common interval, for instance $[0, 1]$. Here, consider that the functions are already defined on a same domain and we focus on retrieving the signals from the noisy curves.

In a second time, dimension reduction is done using multivariate functional principal components analysis, as proposed by [1]. The idea is to write all the observations of the scenes into a common multivariate basis of functions. In fact, by the multivariate version of the Karhunen-Loève decomposition:

$$\mathbf{Z}(t) = \boldsymbol{\mu}(t) + \sum_{j=1}^{\infty} c_j \boldsymbol{\Phi}_j(t),$$

where $\boldsymbol{\mu} = (\mathbb{E}(Z^{(1)}), \dots, \mathbb{E}(Z^{(P)}))$ is the mean vector of each function, $\{\boldsymbol{\Phi}_j\}_{j \geq 1}$ are the multivariate eigenfunctions found by an eigenanalysis of the covariance operator of \mathbf{Z} and the c_j are the projection of \mathbf{Z} onto $\boldsymbol{\Phi}_j$. In practice, we truncate the KL expansion at M terms. However, it is known that in general the KL truncated decomposition is optimal in term of expected quadratic error given the dimension M . Usually, M is chosen to explain a certain percentage of variance (95% or 99% generally) of the data.

So, our multivariate functions \mathbf{Z} are summarized by M coefficients which capture both the information on the curves of each vehicle in the scene, as well as the correlation between the vehicles' curves. Next the clustering step follows directly from that. Classical clustering algorithms are used with the set of coefficient. We consider some appropriate metric in order to take into account the variability of the data in the coefficients.

An algorithm is proposed to analyze vehicle trajectories data using such a methodology.

References

- [1] C. Happ and S. Greven. Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, 113(522):649–659, April 2018. arXiv: 1509.02029.
- [2] James Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition.
- [3] R.D. Tuddenham and M.M. Snyder. *Physical Growth of California Boys and Girls from Birth to Eighteen Years*. Publications in child development. University of California Press, 1954.

Short biography – I have a MSc in Big Data and an engineering degree in statistics from ENSAI. After my end-of-study internship at Renault, the company proposed me to continue as a PhD student working on the data analysis on the autonomous vehicle. I started my thesis at the beginning of January 2018 thanks to the CIFRE plan.