CLUSTERING MULTIVARIATE FUNCTIONAL DATA DEFINED ON RANDOM DOMAINS

An application to vehicle trajectories analysis

Steven Golovkine

National School for Statistic and Information Analysis

steven.s.golovkine@renault.com

Introduction

Nowadays, a vehicle records a lot of information about its environment through his different sensors (camera, radar and 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 vehicle for instance. All the information are recorded relatively to the considered vehicle. We define a driving scene as a small period, say $\mathcal{T}$, during which we record the environment of the car. This environment is constitued by an certain number of vehicles, say P, whose one records a certain number of characteristics, 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 subset of it. So, an observation of a scene can be represented as a random vector of functions:

$$\mathbf{Z} = \left(Z^{(p,d)}(t) : t \in \mathcal{T}^{(p)}, 1 \leq p \leq P, 1 \leq d \leq D\right) \in \mathbb{R}^{P \times D}, \mathcal{T}^{(p)} \subset \mathcal{T}.$$

Moreover, $Z^{(p,d)}$ is assumed to be in $L^2(\mathcal{T}^{(p)})$ for all $p \in [\![1, P]\!]$. So, realizations of $\mathbf{Z}$ are multivariate functional data which are defined on different domains. We aim to analyze such data.

Methodology

Mathematical model

Consider the following mathematical model for independant realizations $\mathbf{Z}_i = \left(Z_i^{(p,d)}(\cdot) : 1 \leq p \leq P, 1 \leq d \leq D\right), 1 \leq i \leq n$ of $\mathbf{Z}$:

$$Z_i^{(p,d)}(t) = \eta^{(d)}(t) + v_i^{(p,d)}(t) + \epsilon_i^{(p,d)}(t), \quad t \in \mathcal{T}. \tag{1}$$

The $P \times D$-dimensional trajectories, $\left(v_i^{(p,d)}(\cdot)\right)$ and $\left(\epsilon_i^{(p,d)}(\cdot)\right)$, are unobserved independant realizations of zero mean multivariate stochastic processes. The processes $\eta^{(d)}$ are deterministic processes. We denote by $\gamma_{\mathbf{v}}(s,t), s, t \in \mathcal{T}$ the covariance structure of the multivariate process corresponding to $v_i^{(p,d)}$ that we aim to identify.

In practice, the data are not recorded continously. Denote by $\mathcal{G}_i = \{t_{i,j} : 1 \leq j \leq m_i\}$ the design points of $\mathbf{Z}_i$. So, the sampling of the equation (1) is written:

$$Z_i^{(p,d)}(t_{i,j}) = \eta^{(d)}(t_{i,j}) + v_i^{(p,d)}(t_{i,j}) + \epsilon_i^{(p,d)}(t_{i,j}), \quad 1 \leq j \leq m_i, 1 \leq i \leq n. \tag{2}$$

Multivariate functional principal components analysis

Smoothing. A smoothing is performed on all the curves. It has two purposes. The first one is to remove the eventual noise in the measurements as the sensors are assumed to not record 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. Following Zhang and Chen (2007), we use Local Polynomial Kernel smoothing for the reconstruction of the curves and adopt the method of Goldenshluger and Lepski (2011) for the bandwidth selection.

Functional Principal Components Analysis. Dimension reduction is done using multivariate functional principal components analysis, as proposed by Happ and Greven (2018). 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 expansion:

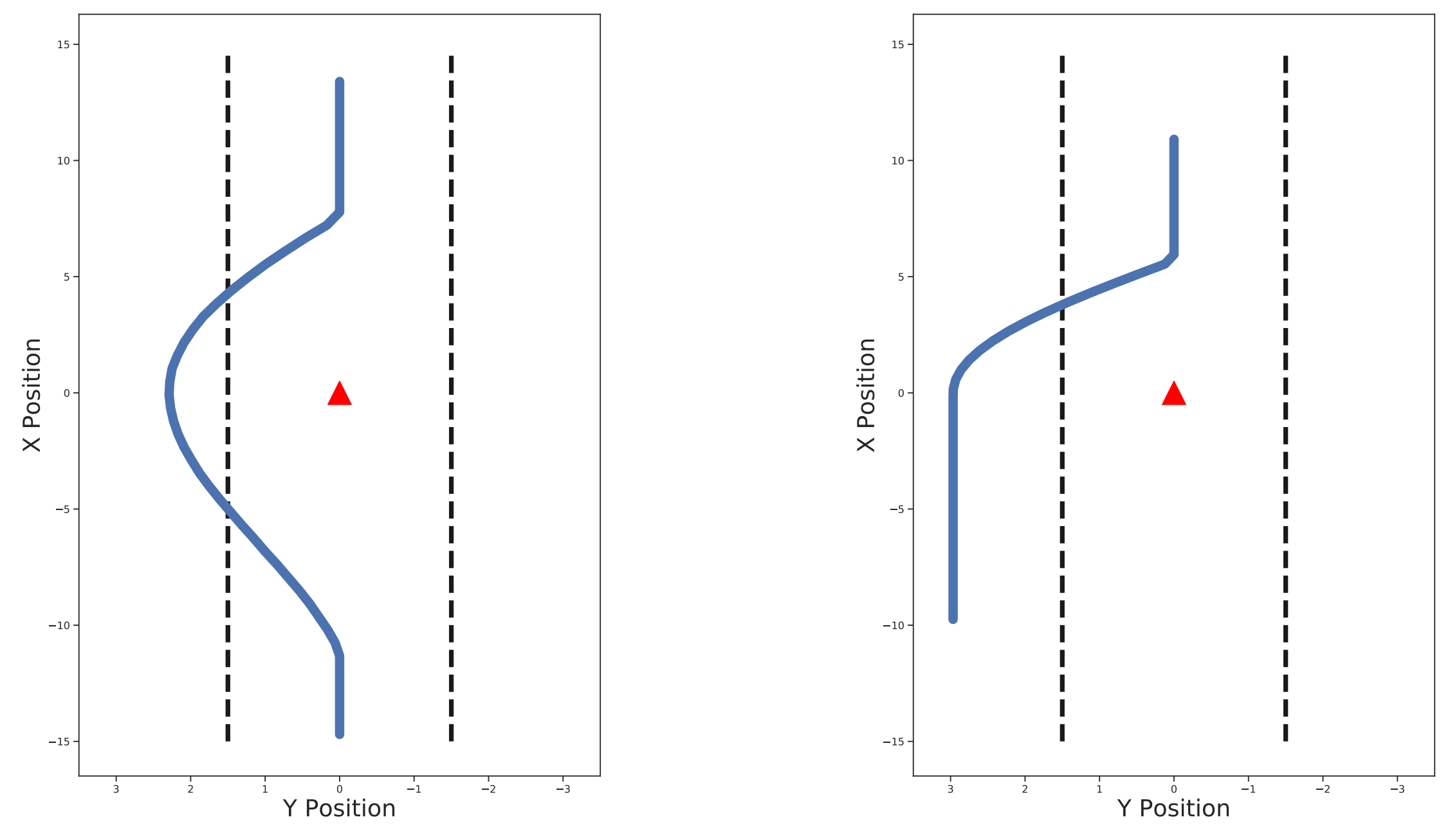
$$\mathbf{Z}(t) = \boldsymbol{\mu}(t) + \sum_{j=1}^{\infty} c_j \mathbf{\Phi}_j(t), \tag{3}$$

where $\boldsymbol{\mu} = \left(\mathbb{E}(Z^{(1)}), \dots, \mathbb{E}(Z^{(P)})\right)$ is the mean vector of each function, $\{\mathbf{\Phi}_j\}_{j \geq 1}$ 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 $\mathbf{\Phi}_j$. 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.

Clustering. Classical clustering algorithms are used with the set of coefficient. In particular, one can cite the k-means algorithm and the spectral clustering algorithm. We consider some appropriate metric in order to take into account the variability of the data in the coefficients.

Tests on simulation

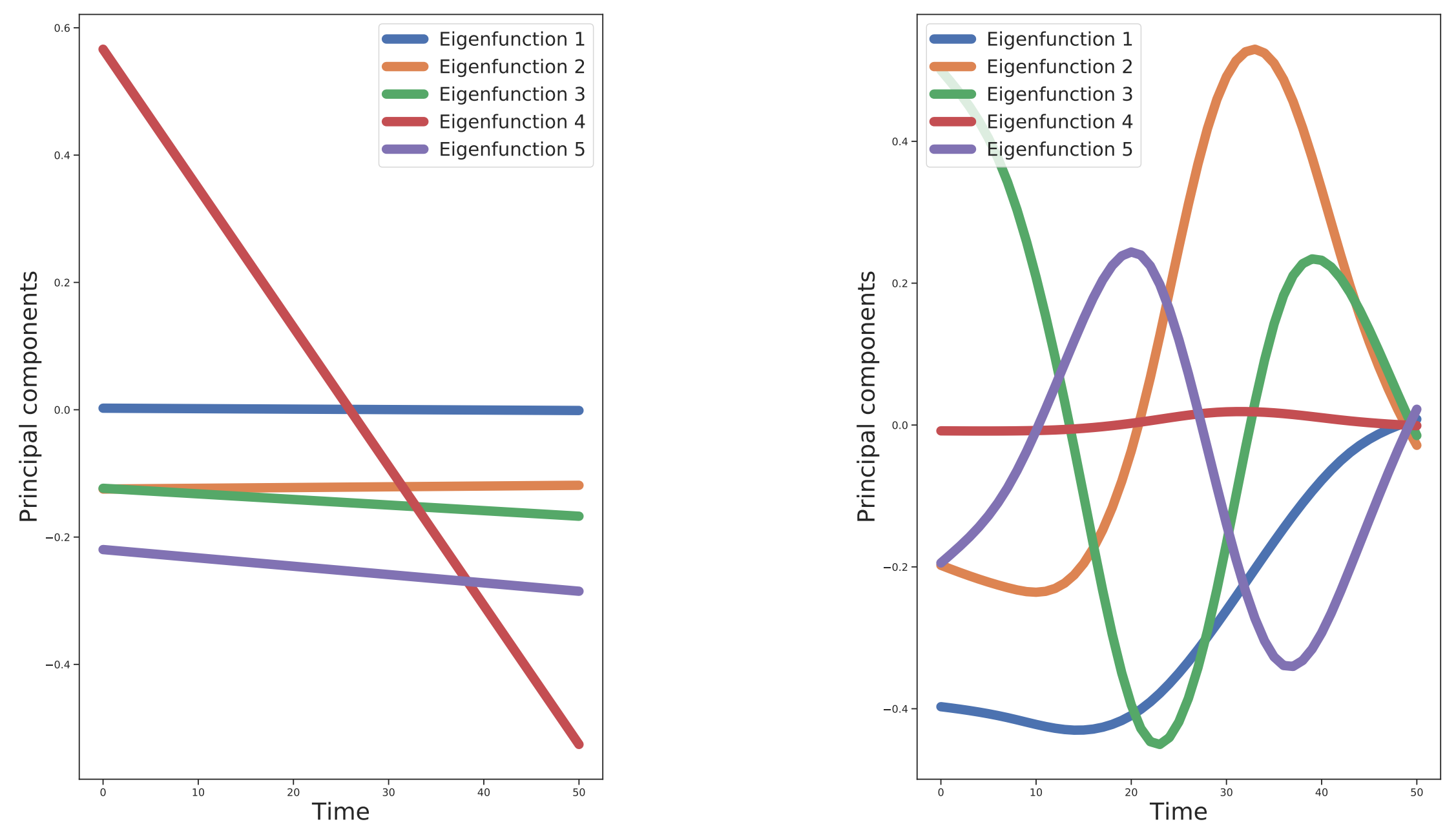
In order to demonstrate the performance of the methodology, we have tried it on simulation. So, we generated different signals that correspond to real driving situations: overtaking (figure 1a) and cut-in (insertion in the vehicle lane, figure 1b). Moreover, the number of overtaking or cut-in by the left is much more important than the ones by the right. This setting corresponds to $P = 1$ and $D = 2$ (one vehicle and two features) in the model 1, and we simulate $n = 500$ independent realizations of the process $\mathbf{Z}$.



(a) Overtaking **(b)** Cut-in

Figure 1: Examples of simulation

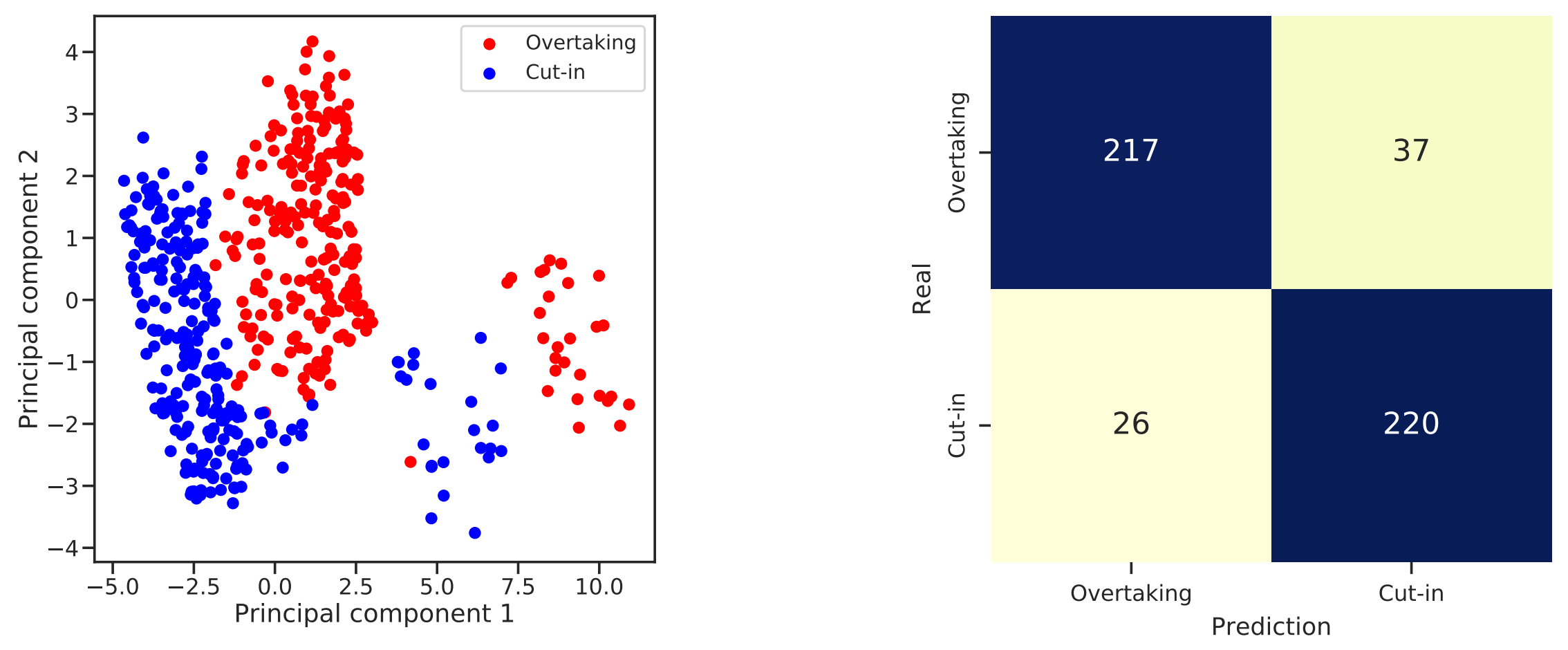
The results of the multivariate functional principal components analysis are shown in the figure 2. This graphs represent the multivariate eigenfunctions of the Karhunen-Loève decomposition. These eigenfunctions explain 95% of the variability in the data.



(a) X Position **(b)** Y Position

Figure 2: Eigenfunctions of the K-L expansion

The figure 3a represents the projection of the multivariate signals into the first principal plan defined by the two first multivariate eigenfunctions. The colors are the predicted class for each of the signal. As we do simulation, we have the true class of each signal, and thus we can compute some metrics to measure method's performance. As we consider non-supervised classification, we can not say which class is overtaking and which is cut-in. So, we manually look at the signals to say that. The algorithm gave the right class in 87% of the time (figure 3b).



(a) Clustering of the coefficients **(b)** Confusion matrix

Figure 3: Results of the clustering

Acknowledgements

This CIFRE thesis n°2017/1116 is supported by Groupe Renault and ENSAI.

References

Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632.

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(522):649–659. arXiv: 1509.02029.

Zhang, J.-T. and Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics*, 35(3):1052–1079.

March 2019 – Groupe Renault