Uncertainty Quantification and Multivariate Functional Data: an Application to the Supervised Classification of Augmented Plane Trajectories

Rémi PERRICHON (ENAC)

Advisors: Thierry KLEIN (ENAC) & Xavier GENDRE (ISAE Supaero)

MASCOT-NUM 2022
Clermont Ferrand, France
7-9 June 2022

Context

To date, the statistical analysis of aircraft trajectories has been under-exploited in the Airspace Traffic Management (ATM) literature. This assessment was early made in [1], promoting the use of Functional Data Analysis (FDA) to study aircraft trajectories. Concurrently, FDA has experienced substantial growth and development in recent years. From the early work of Ramsay and Silverman [2], some topics have gained visibility such as inference procedures or curve registration. Dealing with model-based discriminant analysis and clustering for functional data, preprocessing steps are often neglected as they are suspected to remove clusters. This work aims at reintroducing the interest of constraint smoothing and curve registration for the supervised classification of aircraft trajectories.

Data

In its raw form, a so-called augmented trajectory is a vector of observation times associated to values referring to the usual components of an aircraft trajectory (longitude, latitude, altitude) and a weather dimension (experienced wind speed). Matching trajectories and weather is done using:

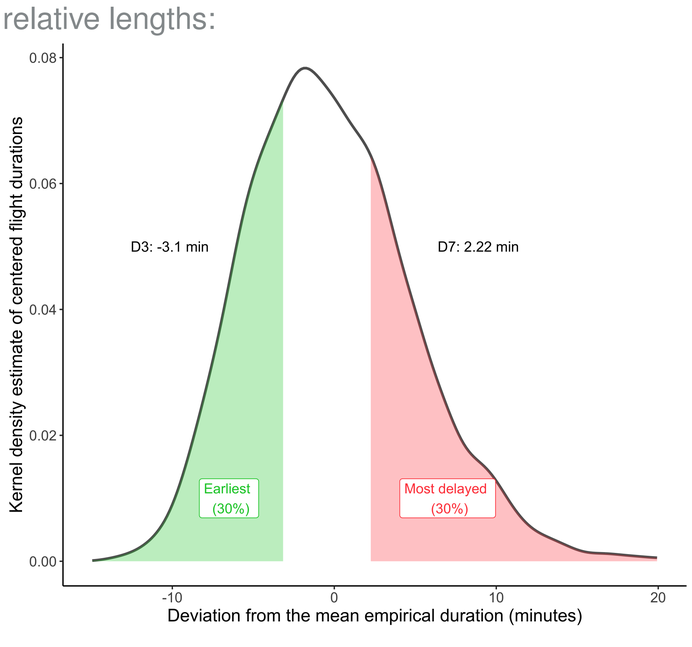
- **(Trajectory data) R&D data from Eurocontrol**

  The R&D data archive contains more than 14 million flights as of April 2021. The data are collected from all commercial flights operating in and over Europe. Data are available for 4 months each year: March, June, September and December. This work is about the 3,000+ flights that departed from Toulouse-Blagnac and landed at Paris-Orly in 2015.

- **(Weather data) ERA 5 data**

  ERA5 is the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis for the global climate and weather.

Two groups of trajectories (earliest and most delayed) can be made according to the empirical deciles of the relative lengths:



References

[1] Puechmorel, S. and Delahaye, D. (2007) 4D trajectories: a functional data perspective. pp 1.C.6. IEEE.

[2] Ramsay, J. O. and Silverman, B. W. (2005) Functional data analysis. Springer series in statistics. Springer-Verlag New York Inc., 2nd edn.

[3] Jacques, J. and Preda, C. (2014) Functional data clustering: a survey. Advances in Data Analysis and Classification.

[4] Jacques, J. and Preda, C. (2014) Model-based clustering for multivariate functional data. Computational Statistics Data Analysis, 71, 92–106.

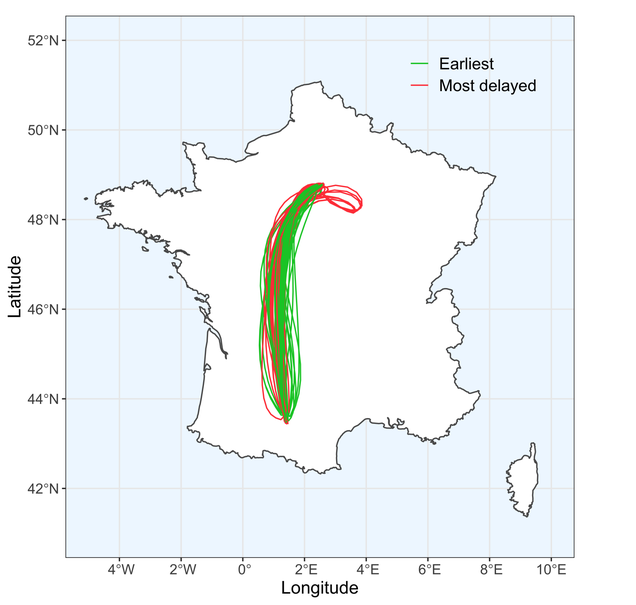
[5] Srivastava, A. and Klassen, E. P. (2016) Functional and Shape Data Analysis. Springer-Verlag New York Inc., 1st edn.

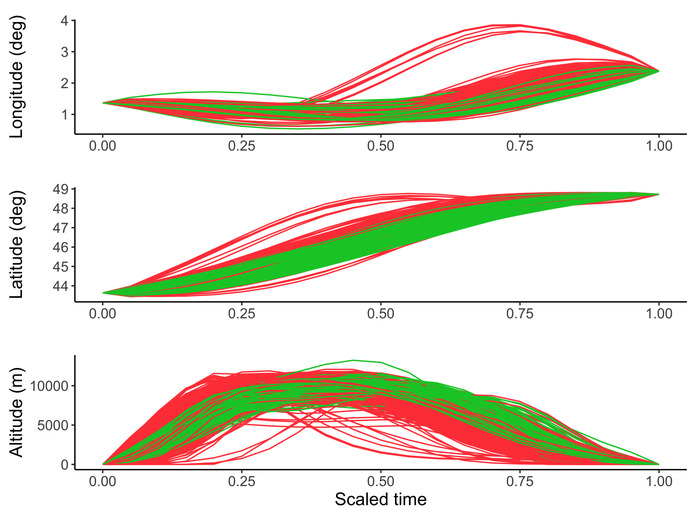
Preprocessing

By nature, an aircraft trajectory is a multivariate object. Let N_{Total} be the total number of flights that can be divided into N earliest trajectories and M delayed trajectories such that $N_{\text{Total}} = N + M$.
For each flight i in group g an augmented trajectory is

$$\{(\boldsymbol{y_{g,i,j}}, t_{g,i,j}), j = \{1, ..., m_{g,i}\}\}$$

where $\boldsymbol{y_{g,i,j}}$ is a four-dimensional vector, $t_{g,i,j}$ a timestamp and $m_{g,i}$ the number of points in trajectory i in group g. The components of $\boldsymbol{y_{g,i,j}}$ are denoted $y_{g,i,j,x}$, $y_{g,i,j,y}$, $y_{g,i,j,z}$, $y_{g,i,j,w}$, respectively referring to the longitude value for point j in trajectory i from group g, the latitude, the altitude and the wind speed. Note that the sampling is not regular within a trajectory, nor between two trajectories. FDA offers a rigorous framework because it directly deals with underlying curves. For each dimension (longitude, latitude, altitude, wind speed), observed augmented trajectories are modeled as independent realizations of an underlying stochastic process taking values in $[0, 1] \rightarrow [-180, 180] \times [-90, 90] \times \mathbb{R}^+$ and given by $\boldsymbol{Y_{g,i}} : t \mapsto (Y_{g,i,x}(t), Y_{g,i,y}(t), Y_{g,i,z}(t), Y_{g,i,w}(t))'$. Reconstructing individual curves from discrete values is a traditional preprocessing step. This smoothing step is particularly challenging for trajectories as some dimensions are naturally constrained.





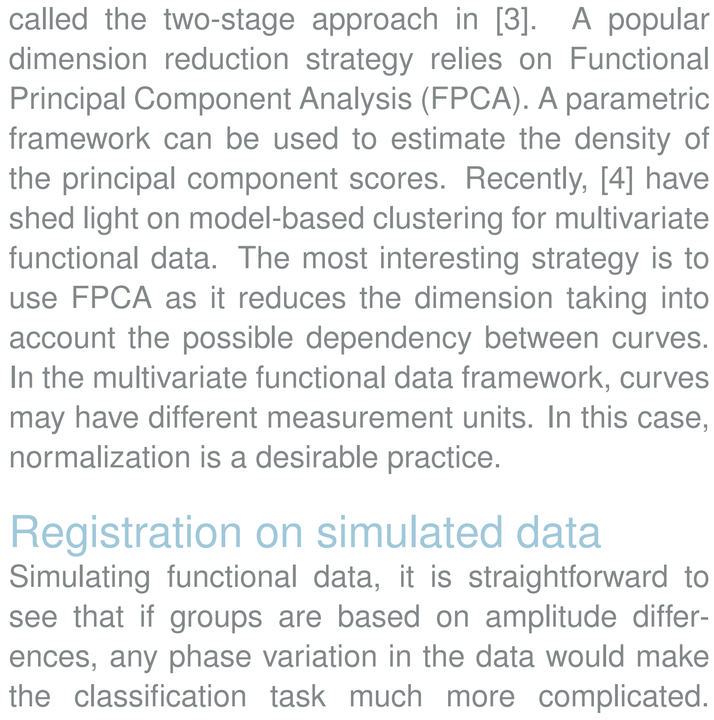
Supervised classification

In Multivariate Data Analysis (MDA), discriminant analysis (also known as pattern recognition or supervised classification) aims at classifying observations in known groups (or classes). The classification task revolves around establishing a rule that can allocate a new observation into one of the groups. In practice, as group-conditional densities are unknown, the optimal decision is unknown and a sample-based rule is estimated from a training data set. A classical method to estimate the group-conditional densities is to rely on a parametric framework. It is a model-based supervised classification.

When dealing with functional data, a dimensionality problem arises immediately. Sampling smoothed curves on a regular grid does not solve the problem as it leads to temporal features that are highly correlated and possibly uninformative. Any model-based approach usually assumes a probability density function on a finite number of parameters describing the curves. In this setting, most of existing approaches for functional data consist of two steps. The first step is to transform the infinite dimensional problem into a finite dimensional one. The second step is to use a MDA technique to perform the supervised classification. This approach has been called the two-stage approach in [3]. A popular dimension reduction strategy relies on Functional Principal Component Analysis (FPCA). A parametric framework can be used to estimate the density of the principal component scores. Recently, [4] have shed light on model-based clustering for multivariate functional data. The most interesting strategy is to use FPCA as it reduces the dimension taking into account the possible dependency between curves. In the multivariate functional data framework, curves may have different measurement units. In this case, normalization is a desirable practice.

Registration on simulated data

Simulating functional data, it is straightforward to see that if groups are based on amplitude differences, any phase variation in the data would make the classification task much more complicated. One may use the Square-Root Slope Function (SRSF) representation of functions developed in [5] to register functions and remove phase variability.



Scaling, registration and supervised classification for trajectories

Empirically, two flights never have the exact same duration. To compare two trajectories, a scaling is made to the unified time interval [0,1]. This transformation is popular in FDA but modifies phases. Using the SRSF framework to perform registration, a FPCA is done to reduce the dimension. The chosen normalization is the one using the maximum of each functional random variable. Four principal components are enough to keep 90% of the variance. A Gaussian-mixture supervised classification is done on the FPCA scores. The density for class g follows a Gaussian mixture distribution given by

$$f_g(\boldsymbol{x}) = \sum_{k=1}^{K_g} \pi_{g,k} \phi(\boldsymbol{x}; \boldsymbol{\mu}_{g,k}, \boldsymbol{\Sigma}_{g,k})$$

where $\pi_{g,k}$ are the mixing probabilities for class g ($\pi_{g,k} > 0, \sum_{k=1}^{K_g} \pi_{g,k} = 1$),$\boldsymbol{\mu}_{g,k}$ the means for component k within class g, $\boldsymbol{\Sigma}_{g,k}$ the covariance matrix of component k within class g, and ϕ the probability density function of the multivariate normal distribution. The number of clusters in the Gaussian mixture is chosen according to the Bayesian Information Criterion (BIC) as well as the covariance structure for each class. The 10-fold cross-validation procedure gives a 7.5 % error rate.

Contact: remi.perrichon@enac.fr