Credit: © F. Cormon/Auvergne-Rhône-Alpes Tourisme

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

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

Rémi Perrichon
Ecole Nationale de l'Aviation Civile (ENAC)

Supervisor(s): Thierry Klein (ENAC) and Xavier Gendre (Isae-Supaéro)

PhD expected duration: Oct. 2021 - Sep. 2024

Address:
Ecole Nationale de l'Aviation Civile
7 Avenue Edouard Belin
31400 Toulouse, France
Office Z.133

E-mail: remi.perrichon@enac.fr

Abstract:

A common way to assess the reliability of a supervised classification procedure is to evaluate its performance on a test set, using a suitable evaluation metric. As an example, the well-known F-score is a good candidate to take accuracy and precision into account. Yet, this strategy is not giving any insight on the model's efficiency when classifying a particular sample. In the uncertainty quantification framework, we are interested in giving a degree of confidence in a prediction, leading to the use of dedicated probability distributions.

The model-based discriminant analysis task has been gaining visibility thanks to some influential papers such as [3]. To perform a model-based discriminant analysis, one has to assume that observations in a given class are generated by a probability distribution specific to that class. In the so-called mixture discriminant analysis setup, the density associated to a class is assumed to be a mixture of normal distributions. This approach is particularly interesting when boundaries are nonmonotonic. The parameters of the probability density function are estimated by the Expectation-Maximization algorithm (EM). To select the number of clusters in the Gaussian mixture among a given set as well as the covariance structure of each class, the Bayesian Information Criterion (BIC) is widely used.

When dealing with functional data, a dimensionality problem arises immediately, and, as a consequence, a probability density function generally does not exist. In this setting, most of clustering algorithms for functional data consists of a first step of transforming the infinite dimensional problem into a finite dimensional one and a second step using a method designed for finite dimensional data. This approach is called the two-stage approach in [4]. In most two-stage approaches, a parametric distribution on some finite set of coefficients characterizing the curves is assumed. For instance, if the density of the basis coefficients are assumed to be a mixture of normal distributions, the above procedure can be performed. Another interesting finite set of coefficients is given by the principal component scores.

The two-stage approach is not entirely satisfactory as the two steps are not jointly considered. In this spirit, a sophisticated model-based technique is early proposed in [6] for sparsely sampled functional data: the spline decomposition is class-dependent. Likewise, relying on the pseudo-density for functional data introduced in [2], some authors put effort in representing functional data in group-specific functional subspaces with an extension to the multivariate case [5]. Recently, [7] has proposed a promising approach. In short, the idea is to model the functional data to retain both their most important characteristics and the ones that are the most associated to a covariate. This method is used to perform some density estimation, among other things.

To our knowledge, all the above methods have not been compared for a model-based discriminant analysis task on the same simulated data. This paper aims at filling this gap.

All approaches are then compared on an original data-set made of so-called augmented trajectories. This data-set is born from a simple assessment: no data-set is currently available to perform a statistical study of the weather impact on plane trajectories despite manifest interest for the airspace management. Indeed, weather delays are a key component of overall delays for given periods of the year. As an example, weather-related delays accounted as much as governmental causes (security or immigration, customs and COVID-19 related delays) to the average delay per flight in the first quarter of 2021 [1].

Raw data for trajectories (longitude, latitude, altitude, time) are taken from the R&D database made available by the European Organisation for the Safety of Air Navigation, commonly known as Eurocontrol. Data are in open access for academics upon request. Trajectory points are associated to weather data coming from ERA5 hourly data on pressure levels (open data). ERA5 is the the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis for the global climate and weather. Adding weather dimensions to a trajectory gives a so-called augmented trajectory. The latter have been constructed for eight departure-arrival airport couples located in Europe, for flights between 2015 and 2018. The data sum up to tens of thousands flights in total.

The focus is made on en-route delays and the role played by wind speed and direction on these delays. As the raw data sampling is very irregular, a suitable framework is naturally the one of Functional Data Analysis (FDA) that allows to evaluate variables' values at the same time for all flights. Each augmented trajectory is viewed as a multivariate functional object, consisting in at least five stochastic processes that are dependent. For a given air link, say flights from Toulouse-Blagnac to Paris-Orly, each trajectory belongs to a class: on-time or delayed. Performances of the model-based methods to predict the correct class of a new trajectory as well as interpretability are compared.

References

[1] All-Causes Delays to Air Transport in Europe, Quarter 1, 2021. Technical report, Eurocontrol, June 2021.

[2] Aurore Delaigle and Peter Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193, April 2010. Publisher: Institute of Mathematical Statistics.

[3] Chris Fraley and Adrian E Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, June 2002.

[4] Julien Jacques and Cristian Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

[5] Julien Jacques and Cristian Preda. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106, March 2014.

[6] Gareth M James and Catherine A Sugar. Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association*, 98(462):397–408, June 2003.

[7] Simon Nanty, Céline Helbert, Amandine Marrel, Nadia Pérot, and Clémentine Prieur. Uncertainty quantification for functional dependent random variables. *Computational Statistics*, 32(2):559–583, June 2017.

Short biography – My background revolves around Statistics and Econometrics (Master's degree at Toulouse School of Economics, Magistère diploma at Université Paul Sabatier (Toulouse), apprenticeship at Airbus Helicopters). My PhD is funded by the ENAC-Isae-Supaéro-ONERA foundation and is entitled Statistical Modeling of Plane Trajectories for Classification and Prediction.