A new methodology for classification of partially observed curves: an application to aneurysm patients

Pavel Hernández-Amaro[1], *Maria Durban*[2], *M. Carmen Aguilera-Morillo*[3]

[1]pahernan@est-econ.uc3m.es, Department of Statistics, University Carlos III de Madrid
[2]mdurban@est-econ.uc3m.es, Department of Statistics, University Carlos III de Madrid
[3]mdagumor@eio.upv.es, Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València

Functional data analysis is one of the fastest growing fields in statistical analysis. Modern data sets often consist of complex objects, such as functions. Functional data is usually found as discrete and often noisy observations of the true underlying function, measured at different locations in time, space, or other continuum. In most cases it is assumed that all functions are observed over the full extension of their domain. However, in many real data sets, each curve is observed in a subset of the domain, which may even be different for each curve. This type of data is know as partially observed functional data.

In this work we present a new methodology to fit a generalized scalar-on-function regression model to deal with partially observed functional data. The proposed functional model considers each curve only within its observed subset of the domain; also a penalty is added to the estimation of the functional coefficient in order to control its smoothness trough the smoothing parameter. Additionally a basis representation of the functional data and the functional coefficient of the model is made. This representation allows us to transform our functional model into a mixed effect model and then estimate directly all the model coefficients including the smoothing parameter. We use B-spline basis for our representations but other suitable basis can be chosen.

The performance of the proposed model is tested on a real classification problem from the AneuRisk65 data set (https://statistics.mox.polimi.it/aneurisk/). The goal is to classify each patient into one of two groups depending on the presence and location of the aneurysms, for this classification two functional variables are taken into consideration: the radius and the curvature of the internal carotid artery. Additionally we also test our new methodology in a simulation study and compared its performance with other classification techniques for partially observed functional data.

Keywords: Partially observed functional data, generalized scalar-on-function regression model, B-splines.