Book of Short Papers SIS 2018

Editors: Antonino Abbruzzo - Eugenio Brentari
Marcello Chiodi - Davide Piacentino

SIS 2018 PALERMO

49 TH SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY

20-22 JUNE

PUBLISHED BY PEARSON

WWW.PEARSON.COM

First printing, November 2018, ISBN-9788891910233

SIS 2018
49 TH SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY
PALERMO 20-22 JUNE

Contents

Classification of the Aneurisk65 dataset using PCA for partially observed functional data

Classificazione del dataset Aneurisk65 utilizzando la PCA per dati funzionali parzialmente osservati

Marco Stefanucci, Laura Sangalli and Pierpaolo Brutti

Abstract When functional data are observed over a domain that is subject-specific, most of the techniques for functional data analysis are invalidated. Recently, new methods able to handle this situation were developed and in particular we focus on well-known functional PCA. With the aim of classifying the Aneurisk65 dataset, we apply a few possible methods and we show that carrying out the analysis over the full domain, where at least one of the functional data is observed, may not be the optimal choice. This is also confirmed in a simulation study, where the best interval for classification lies between the common domain and the full domain.

Abstract *Ogniqualvolta dei dati funzionali vengono osservati su un dominio dipendente dal soggetto considerato, non è più possibile utilizzare la gran parte delle tecniche per l'analisi di dati funzionali. Recentemente sono stati sviluppati nuovi metodi in grado di affrontare questa situazione e noi tratteremo la ben nota ACP funzionale. Con lo scopo di classificare il dataset Aneurisk65, abbiamo applicato diversi metodi e mostreremo che eseguire l'analisi sul dominio completo, dove è osservato almeno uno dei dati funzionali, può non essere la scelta ottimale. Ciò è confermato anche da uno studio di simulazione dove il migliore intervallo per la classificazione giace tra il dominio comune e il dominio completo.*

Key words: Functional Data, Partially Observed Data, Classification, Functional PCA

Marco Stefanucci
Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma e-mail: marco.stefanucci@uniroma1.it

Laura Sangalli
Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano e-mail: laura.sangalli@polimi.it

Pierpaolo Brutti
Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma e-mail: pierpaolo.brutti@uniroma1.it

In the last 20 years, functional data analysis emerged as one of the fastest growing fields of modern statistics [1]. Functional data can be viewed as realizations of a functional objects, and in general are collected as discrete and noisy observations. The domain where the data are observed is usually common to all the statistical units, while the observation grid may vary across units. When the domain is unit-dependent, standard procedures are invalidated and a practical solution is to restrict the analysis to the intersection of the domains, where all the curves are observed. However, including intervals where the data are only partially observed may result in a more powerful statistical analysis and in the last decade some authors developed methodology handling this particular situation.

The Aneurisk65 dataset is an important example of partially observed functional data. The data consist in the profiles of radius and curvature of the internal carotid artery of 65 subjects suspected to be affected by cerebral aneurysms [2] and are displayed in figure 1. The domain where all the curves are available is highlighted in light-gray. Outside this domain individual observations are progressively lost when moving towards the full domain. In this application it is relevant to investigate whether the morphology of the internal carotid artery influences aneurysms pathogenesis and this can be done considering that the data can be divided in two groups - displayed in orange and blue in the figure - depending on the presence and location of the cerebral aneurysms.

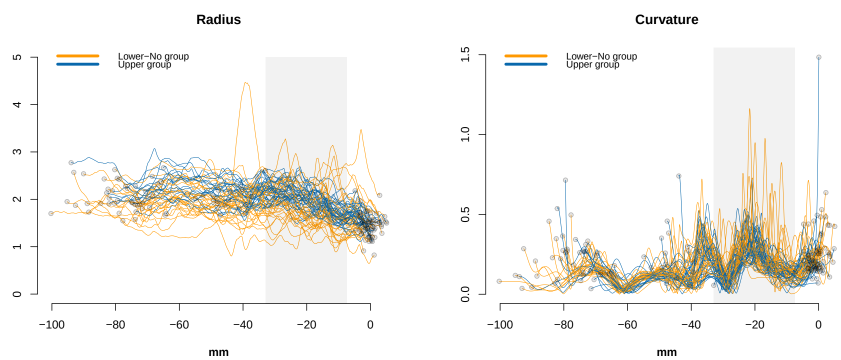

Fig. 1 The Aneurisk dataset.

In [2] a discriminant analysis based on the scores of the principal components of the radius and curvature profiles is presented, restricting the attention to the portion of the domain common across subjects. Our contribution is to improve the discrimination results by considering also portions of the domain where not all data are observed. In doing this, we need a methodology for functional principal components analysis in the case of partially observed data. Here we consider some proposals: the first one [3] is based on a spline reconstruction of the eigenfunctions, the second [4] is a method that estimates the principal component scores using conditional expectations, the third [5] is a generalization of the power method for extracting eigenvectors of a given matrix and the last one [6] is a fully functional approach in which the missing part of each score is predicted via best linear approximation of its conditional expectation.

We then use these methods to perform a discrimination based on the scores of the first K principal components. We show that carry out the analysis on the largest possible domain may not be the best choice for classification purposes. In fact, we suggest to explore different intervals, ranging from the common domain to the full domain. More specifically, we divide the domain where the data are partially observed in L portions, and we consider a collection of progressively larger domains I_l for $l \in \{0,\dots,L\}$, with $I_{l-1} \subset I_l$, where I_0 is the common domain and I_L is the full domain. We select the optimal number of principal components and the optimal domain extension I_l via cross-validation. For all the methods, the best interval for classification is between the common domain and the full domain and such an interval is the same for all the approaches, see figure 2. In this application, the approach [5] achieves the best result, corresponding to a 9 misclassified subjects, thus outperforming the result in [2]. Note that the application of the methodologies for partially observed data on the full domain does not lead to any improvement in the discrimination.

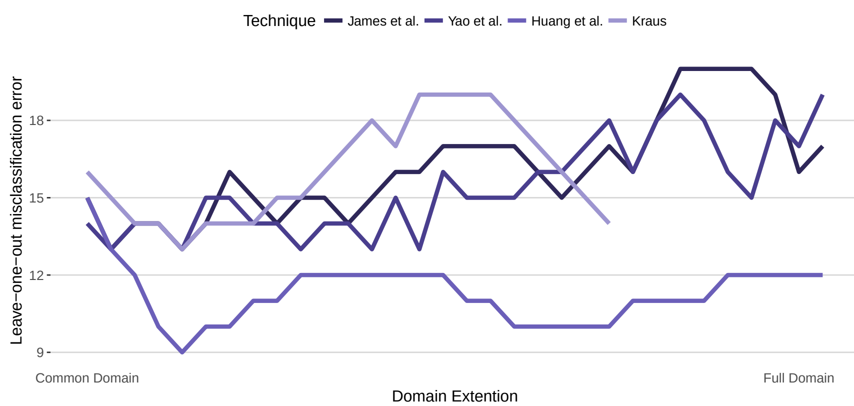

Fig. 2 Performances on the Aneurisk dataset.

In order to understand better the behaviour of the domain extension technique, we carry out simulations. We generate functional data over the interval $I_L = [0,1]$. We then completely retain the data generated over the interval $I_0 = [1/3, 2/3]$, while we censor them over the intervals $I_{\texttt{left}} = [0,1/3]$ and $I_{\texttt{right}} = [2/3,1]$, by sampling the starting point of each functional datum uniformly over $I_{\texttt{left}}$, and its ending point uniformly over $I_{\texttt{right}}$. The data are generated from a cubic B-splines basis with 16 internal knots, corresponding to a total of 20 bases. We generate two groups of 50 units each with the only difference being the mean function, see figure 3. The 100

Fig. 3 Simulated data.

generated curves are evaluated on a regular grid of 150 points in $[0,1]$, contaminated with gaussian noise and finally classified over a set of progressively larger domains. This simulation is repeated 50 times.

In order to compare the performances of the four methods, we also apply standard PCA to the fully observed data. The results are presented in figure 4. The figure displays the boxplots of the leave-one-out misclassification error, for various tech-

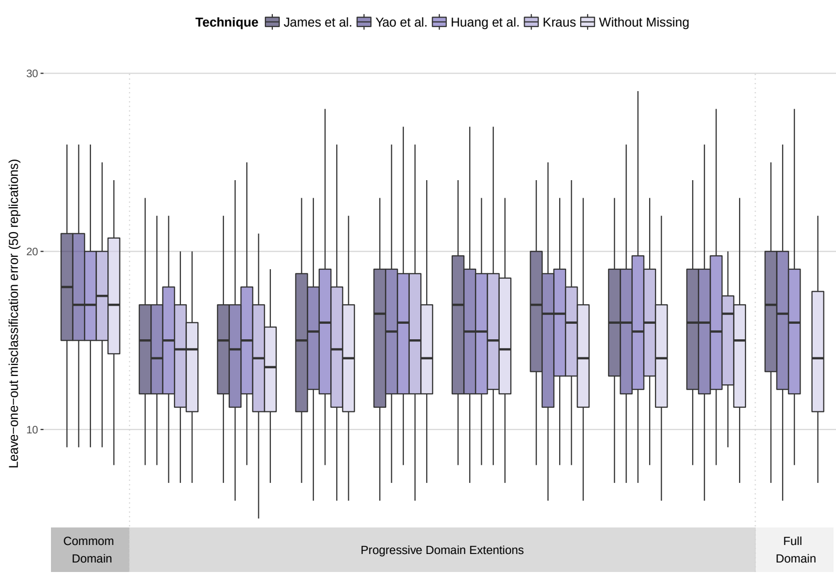

Fig. 4 Simulation results.

niques, for various domain extensions. For all methods, the misclassification error decreases when we start extending the domain with respect to the common domain, but then progressively increases as we approach the full domain. None of the methods outperforms the other. In conclusion, extending the domain with respect to the common domain improves the discrimination between the two groups; on the other hand, larger domain extensions, and in particular the full domain, do not lead to the best discrimination results.

References

1. Ramsay, J. and Silverman, B.W.: Functional Data Analysis. Springer (2005)
2. Sangalli, L.M., Secchi, P., Vantini, S. and Veneziani, A.: A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. Journal of the American Statistical Association **104**, 37-48 (2009)
3. James, G.M. and Hastie, T.J.: Principal component models for sparse functional data. Biometrika **87**, 587-602 (2000)
4. Yao, F., Müller,H.G. and Wang, J.L.:Functional Data Analysis for Sparse Longitudinal Data. Journal of the American Statistical Association **100**, 577-590 (2005)
5. Huang, J.Z., Shen, H., and Buja, A. : Functional principal components analysis via penalized rank one approximation. Electronic Journal of Statistics **2**, 678–695 (2008)
6. Kraus, D.: Components and completion of partially observed functional data. Journal of the Royal Statistical Society: Series B **77**, 777–801 (2015)