

Computational Issues in the Application of Functional Data Analysis to Imaging Data

Juan A. Arias-López[1(✉)], Carmen Cadarso-Suárez[1], and Pablo Aguiar-Fernández[2,3]

[1] Biostatistics and Biomedical Data Science Unit. Department of Statistics, Mathematical Analysis, and Operational Research, Universidade de Santiago de Compostela, Santiago, Spain
juanantonio.arias.lopez@usc.es

[2] Nuclear Medicine Department and Molecular Imaging Group, University Clinical Hospital (CHUS) and Health Research Institute of Santiago de Compostela (IDIS), Santiago, Spain

[3] Molecular Imaging Group, Department of Psychiatry, Radiology and Public Health, Faculty of Medicine, Universidade de Santiago de Compostela, Santiago, Spain

Abstract. Functional Data Analysis (FDA) is the field of statistics which deals with the analysis of data expressed in the form of functions, which is extensible to data in the form of images. In a recent publication, Wang et al. [1] settled the mathematical groundwork for the application of FDA to the estimation of mean function and simultaneous confidence corridors (SCC) for a group of images and for the difference between two groups of images. This approach presents at least two advantages compared to previous methodologies: it avoids loss of information in complex data structures and also avoids the multiple comparison problem which arises from pixel-to-pixel comparison techniques. However, the computational costs of applying these procedures are yet to be fully explored and could outweigh the benefits resulting from the use of an FDA approach. In the present study, we aim to apply these novel procedures to simulated data and measure computing times both for the estimation of mean function and SCC for a one-group approach, for the comparison between two groups of images, and for the construction of Delaunay triangulations necessary for the implementation of the methodology. We also provide the computational tools to ensure replicability of the results herein presented.

Keywords: Functional data analysis · Image processing · Computational biology

1 Introduction

1.1 Functional Data Analysis

Functional Data Analysis (FDA) can be defined as the field of statistics concerned with the theoretical basis and the development of analytical tools

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12950, pp. 630–638, 2021.
https://doi.org/10.1007/978-3-030-86960-1_46

oriented towards the study of data in the form of functions. From this scope of view, a function is the minimum unit of the analysed data and usually every subject or participant in a study has one or more functions associated. In the last decades, FDA has experimented a rise in popularity in several research areas and different publications - including monographs [2] and review articles [3] - are now available to the public explaining its theoretical basis and applications.

However, FDA is still in its infancy and thus a strict definition of the field is difficult and not recommended. Instead, there are a few characteristics generally assumed to be inherent to functional data. First, functional data are continuously defined and single instances of this data are only considered as realizations of the underlying function, a necessary constraint in order to use this data with the computational means available to us. Second, the whole function - rather than individual points - is the basic element of the analytical process carried out in FDA. Additionally, functional data is usually associated with time as a variable and often this data is smoothed or at least some regularity conditions are imposed on it [2]. Functional data usually consist of a sample of independent functions whose values are located in a compact interval (I) and are usually assumed to be in a Hilbert space (L^2):

$$X_1(t), X_2(t), ..., X_n(t); I = [0, T] \in L^2 \tag{1}$$

1.2 FDA for Imaging Data

FDA techniques can easily be extended to work with images, and especial attention has been paid to its use with biomedical imaging data such as brain scanner data and images of tumor tissue [2]. However, smoothing methods proposed to date in the scientific literature for imaging analysis (e.g. kernel smoothing, tensor product smoothing...) suffer from a problem of *leakage* when the data structure is complex, showing difficulties in boundary regions' estimation resulting in inappropriate smoothing.

Besides, when analysing medical imaging data, problems arise not only for the estimation of the mean value for a given point, but also for the estimation of the associated uncertainty (i.e. confidence band). This problem becomes even more complicated when considering that the spatial correlation has to be taken in account. To date, the predominant analytic techniques usually rely on *mass univariate approaches*, which consider pixels in the image as independent units and then perform comparisons between them with classical methods such as a simple *T-test*. This brings up the problem of multiple comparisons, which is usually addressed with popular approaches such as the Bonferroni correction or extensions of random field theory [4], which are *ad hoc* corrections heavily dependent on the appropriate choice of a threshold.

In a recent publication, Wang et al. [1] proposed a way to avoid these problems using FDA techniques. First, the challenge of *leakage* on complex domains is addressed by using bivariate splines over Delaunay triangulations (see Sect. 3.1), thus preserving important features of the imaging data. Second, the proposed

methodology treats imaging data as an instance of functional data continuously defined and only observed on a regular grid. If the image is considered as functional, attention moves from the pixel as an individual unit towards the analysis of images as a whole, allowing for calculations such as the simultaneous confidence corridors (SCC; also known as *simultaneous confidence bands*), an approach which has been proven superior to conventional multiple comparison approaches [5].

In their article, Wang and colleagues [1] describe the proposed bivariate spline estimators, test their asymptotic properties, describe the attributes of SCC based on these estimators, and extract coverage probability for the obtained mean function, concluding that the proposed SCC account for the correct probability coverage both in one-group and two-group setups. However, although the proposed methodology accounts for the correct probability coverage, the computational resources necessary for its application are not addressed by the authors and thus its utility for a practical case is yet to be tested, as computing times, which include among others the calculation of Delaunay triangulation parameters, could outweigh the benefits of choosing an FDA approach.

2 Objectives

Given the gap in knowledge mentioned in the previous section, we set the goal of testing the practical utility of this novel methodology by studying the computational efforts necessary to implement it. In order to do that, we evaluate computing times for the calculation of the polygonal domain of Delaunay triangulations for imaging data (necessary for the technique's implementation), evaluate computing times for the calculation of mean function and SCC for a sample of simulated imaging data, and also for the calculation of mean function and SCC for the difference between two samples of simulated imaging data.

3 Methods

3.1 Delaunay Triangulations

In mathematical terms, Delaunay triangulations consist on a series of triangles created by the union of different vertices in which no vertices falls inside the circumcircle of a given triangle. The examined FDA approach relies on applying bivariate splines over these triangulations to preserve imaging data features with the minimum loss of information. These Delaunay triangulations are essential for the construction of SCC and thus its computing times also need to be tested. In order to do this, we will be using the *Triangulation* R package [6] and test its performance for growing values of the triangulation fineness degree. Computing times and examples of the results are analysed in Sect. 4.1.

3.2 One-Group Mean Function and SCC

The methodology herein tested allows for the estimation of the mean function for a series of images and also for the calculation of its associated SCC. With the aim of evaluating performance in terms of computing times we simulate a growing number of data using functions implemented in *ImageSCC* R package [7] with what would be the boundaries of a brain used in a hypothetical neuroimaging study. We then evaluate computing times for the estimation of mean function and SCC for this data using different triangulation parameters. Computing times and examples of the results are analysed in Sect. 4.2.

3.3 Two-Group Mean Function and SCC

This technique is also extensible to a two-sample setup. Following this approach, estimated mean functions are calculated for the two different groups separately and then SCC are estimated for the difference between the two groups' mean functions. The results display areas in the analysed images whose difference fall outside estimated SCC for a given α level, thus suggesting a significant difference in that region in one group compared to another. With the aim of testing this methodology's performance, we simulate a growing number of neuroimaging data for two groups and then compute the estimated mean functions for each group and SCC for the difference between these groups' mean functions, visualizing areas falling outside estimated SCC. Computing times and examples of the results are analysed in Sect. 4.3.

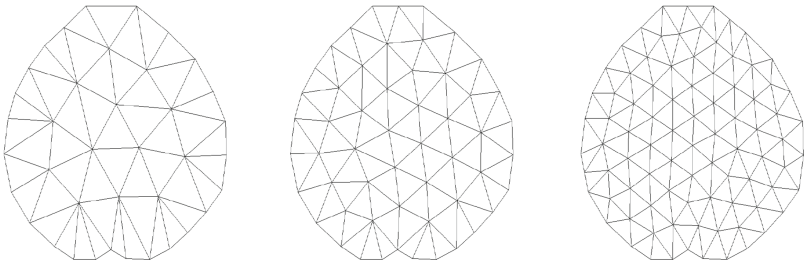
4 Results

In the following sections we summarize results for the analysis proposed in our Objectives (Sect. 2), including computing times and sample graphics of the obtained results.

4.1 Delaunay Triangulations

In Fig. 1 we visually represent three output examples of Delaunay triangulations for an hypothetical brain imaging study. The brain is naturally tri-dimensional, however, we will be working with a two-dimensional axial cut of the brain for practical reasons. These triangulations are the basis for imaging mean function and SCC estimation and will be used in the following sections in order to evaluate computing times of this methodology.

In Fig. 2 we visually assess computing times for these triangulations depending on the selected degree of fineness. We can see a pattern of growing computing times with the increase of n value. According to the authors [1] a value of $n = 8$ is sufficient, although higher values such as $n = 15$ and $n = 25$ still take reasonable amounts of time and will be considered in this study.



(a) N=8 (b) N=15 (c) N=25

Fig. 1. Delaunay triangulations with increasing N values representing increases in triangulation's degree of fineness.

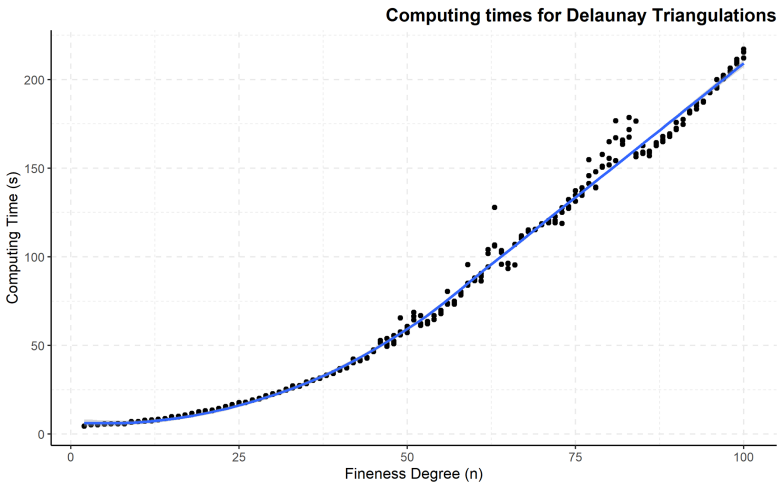
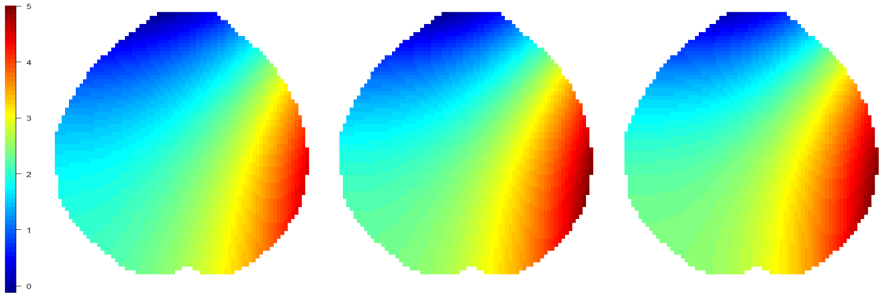


Fig. 2. Computing times for Delaunay triangulations with growing fineness degree values. Curve fitted with LOESS regression (interactive version).

4.2 One-Group Mean Function and SCC Estimation

In Fig. 3 we present an example of the images obtained using this methodology with one-group imaging data, displaying estimated mean function, lower, and upper confidence intervals in the form of images.

Simulation of the data presented in Fig. 3 was carried out with ImageSCC [7] package using a $n = 50$, with cubic function for data generation and eigenvalues to adjust subject-level variation in the simulated data of $\lambda_1 = 0.5$ and $\lambda_2 = 0.2$. Besides for triangulation degree of fineness ($n = 25$), estimated mean function and SCC were calculated using parameters recommended by



(a) Lower SCCs (b) Mean Function (c) Upper SCCs

Fig. 3. Results for one-sample SCC estimation with estimated mean function (center), lower SCC (left), and upper SCC (right) calculated for $\alpha = 0.05$ using triangulation's fineness degree $N = 25$.

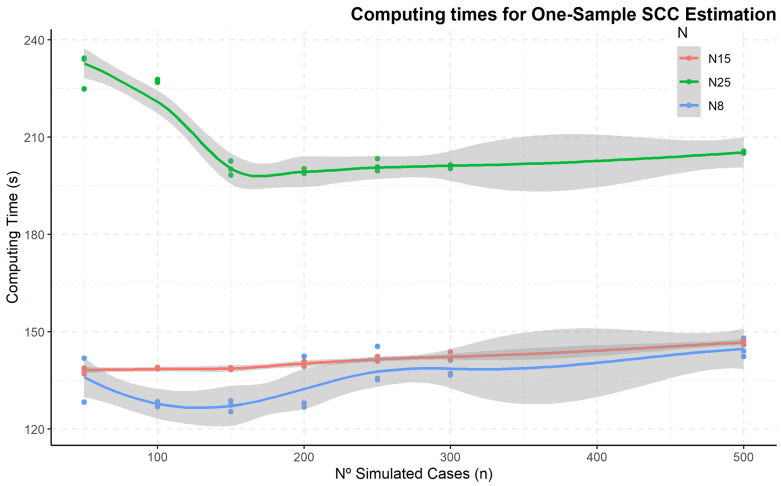


Fig. 4. Computing times for one-group mean function and SCC estimation for imaging data with growing number of simulated cases and three Triangulation fineness degrees (N = 8, N = 15, N = 25). Curves fitted using local (LOESS) regression (interactive version).

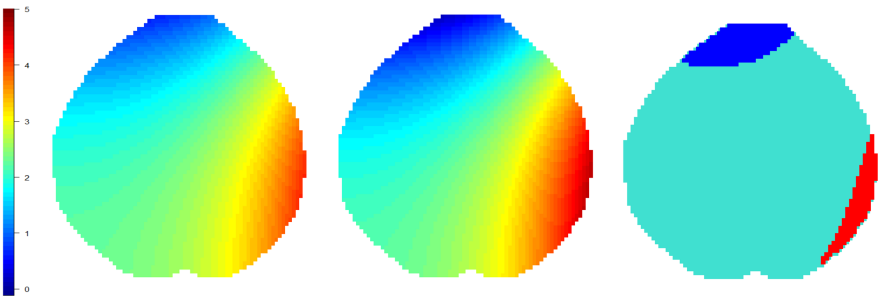
Wang et al. [1] including: degree of bivariate spline for mean estimation $d.est = 5$, degree of bivariate spline for construction of SCC $d.band = 2$, smoothness parameter $r = 1$, and a vector of candidates for penalty parameter with values ranging from 10^{-6} to 10^{3}. In addition, SCC were calculated for three different α levels.

In Fig. 4 we visually assess computing times for mean function and SCC estimation for a one-group setup with growing number of simulated subjects and three different triangulation parameters. We see that these times are similar for fineness degrees of $N = 8$ and $N = 15$, and much larger for $N = 25$. We

can also observe that these times are stable and tend to orbit around a certain value with slow increases as the number of simulated subjects grows. It is also worth mentioning that the results for $N = 25$ show high computing times for low number of simulated cases which tends to then be reduced and stabilize.

4.3 Two-Group Mean Function and SCC Estimation

In Fig. 5 we represent an example of the images obtained using this methodology with two-group imaging data, displaying estimated mean function for both groups and also areas whose values fall outside estimated SCC for the difference between groups of images.



(a) Mean Function G1 (b) Mean Function G2 (c) Areas outside SCC

Fig. 5. Results for a two-sample setup showing estimated mean function for Group 1 (left) and Group 2 (center), together with estimated areas with significant ($\alpha = 0.05$) inter-group differences (blue: hypo-activity; red: hyper-activity). Triangulation's fineness degree $N = 25$. (Color figure online)

Simulation of data for the creation of estimated mean functions and SCC was carried out following the same parameters described in Sect. 4.2. In addition, the additional parameter $\delta = 0.8$ was used, indicating the scale of difference between the two simulated groups' mean functions. In Fig. 6 we visually assess computing times for obtaining the two simulated groups' mean function and SCC for the difference between them. We carry out this calculation for a growing number of simulated subjects and three triangulation setups. The obtained timings are similar for $N = 8$ and $N = 15$, and much larger for $N = 25$. We can also see that computing times are extremely stable around a certain value and present very slow growth with regards to growing number of simulated subjects. Compared to Fig. 4, the variability of computing times is almost null.

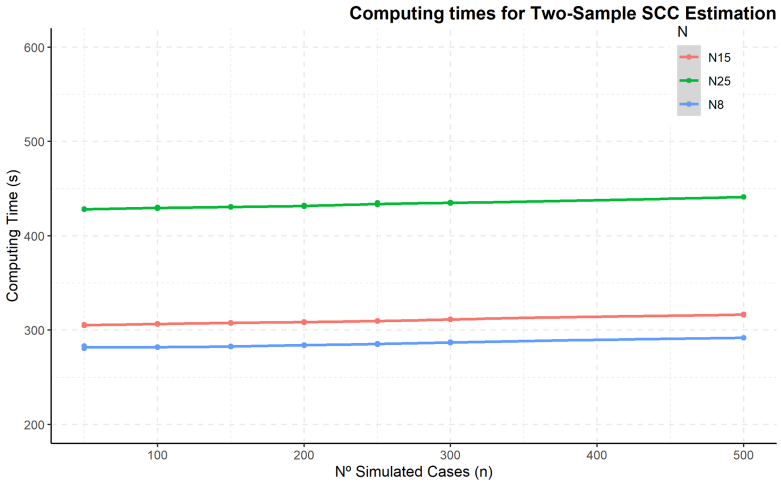


Fig. 6. Computing times for two-group mean function and SCC estimation for the differences between groups with growing number of simulated cases and three Triangulation parameters. Curves fitted using LOESS regression (interactive version).

5 Discussion

In this article we aimed to assess computational costs for Wang and colleagues' [1] proposal of an FDA methodology in order to obtain mean functions and SCC for imaging data. Our appraisal consisted on the calculation of computing times for Delaunay triangulations necessary for this method, simulation of data, and evaluation of time necessary for the estimation of mean function and SCC (one-group approach) and estimation of mean functions and SCC for the difference between groups (two-groups approach).

The obtained results suggest the following. First, computation times for Delaunay triangulations grow with the increase of selected fineness degree, although these computing times are low, especially regarding the fineness degree values recommended by the authors (see Fig. 2). Second, computing times for a one-group approach are much more dependent on the selected triangulation parameters than the number of simulated cases, as shown in Fig. 4; besides, it seems sensible to use higher values of triangulation's degree of fineness than the recommended by the authors, as computing times for $N = 8$ and $N = 15$ are almost identical. Third, computing times for the two-group approach are also heavily dependent on the selected triangulation parameters rather than the variation in number of simulated subjects, for which the results remain stable as shown in Fig. 6.

The strong dependence of computing times on the triangulation parameters, rather than the number of studied cases, goes in line with expected outcomes for functional data methodologies, which are meant to be applied to a high number of cases, whereas increases in the intricacy of the triangulation meshes tend to

produce cumulative effects deriving in increased computation times due to the higher complexity of the calculations involved.

In conclusion, due to the utility of this methodology to avoid other recurrent problems in imaging analysis and also due to the relatively low computing times of this methodology, it seems reasonable to recommend further application of this technique to practical cases, especially bearing in mind the great time stability displayed in the two-sample case for big numbers of simulated cases, a property which is valuable in medical research dealing with a great number of patients. Nevertheless, this is a simulation study and there is still a need for testing this methodology and its results in a complete setup, with real data over more realistic image structures. In short, these preliminary results suggest that functional data analysis techniques can be useful for imaging analysis, displaying desirable properties such as stability and low computing times.

Computer Specifications

This study was carried out using a personal computer with the following specifications and R version. Computer model: MSI Modern 14 A10M; operating system: Windows 10 Enterprise LTSC 64-bit; CPU: Intel(R) Core(TM) i5-10210U CPU @ 1.60 GHz (8 CPUs); RAM memory: 16384 Mb; R version: 4.0.3 (2020-10-10).

Complementary Materials

Complementary material are stored in our GitHub code repository.

References

1. Wang, Y., Wang, G., Wang, L., Ogden, R.T.: Simultaneous confidence corridors for mean functions in functional data analysis of imaging data. Biometrics **76**(2), 427–437 (2020). https://doi.org/10.1111/biom.13156
2. Ramsay, J.O.: Functional data analysis. In: Encyclopedia of Statistical Sciences, vol. 4 (2004). https://doi.org/10.1002/0471667196.ess3138
3. Wang, J.-L., Chiou, J.-M., Müller, H.-G. : Functional data analysis. Ann. Rev. Stat. Appl. **3**, 257–295 (2016). https://doi.org/10.1146/annurev-statistics-041715-033624
4. Worsley, K.J., Taylor, J.E., Tomaiuolo, F., Lerch, J.: Unfied univariate and multivariate random field theory. NeuroImage **23**, S189–S195 (2004). https://doi.org/10.1016/j.neuroimage.2004.07.026
5. Degras, D.A.: Simultaneous confidence bands for nonparametric regression with functional data. Statistica Sinica, 1735–1765 (2011). https://doi.org/10.5705/ss.2009.207
6. Lai, M.J., Wang, L.: Triangulation: Triangulation in 2D domain. R package version 0.1.0 (2020)
7. Wang, Y., Wang, G., Wang, L.: ImageSCC: SCC for Mean Function of Imaging Data. R package version 0.1.0 (2020)