MODELLISTICA E CALCOLO SCIENTIFICO

MOX

MODELING AND SCIENTIFIC COMPUTING

MOX–Report No. 09/2010

Joint Clustering and Alignment of Functional Data: an Application to Vascular Geometries

Laura M. Sangalli, Piercesare Secchi,
Simone Vantini, Valeria Vitelli

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

Joint Clustering and Alignment of Functional Data: an Application to Vascular Geometries

Laura M. Sangalli, Piercesare Secchi, Simone Vantini, Valeria Vitelli

MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
laura.sangalli@polimi.it, piercesare.secchi@polimi.it,
simone.vantini@polimi.it, valeria.vitelli@mail.polimi.it

Keywords: Functional data analysis, clustering, alignment, registration.

Abstract

We show an application of the *k-mean alignment* method presented in Sangalli et al. (2010b). This is a method for jointly clustering and aligning functions that puts in a unique framework two widely used methods of functional data analysis: Procrustes continuous alignment and functional k-mean clustering. These two methods turn out to be two special cases of the new method. In detail we use this algorithm to analyze 65 internal carotid arteries (ICA) in relation to the presence and rupture of cerebral aneurysms. Some interesting issues, amenable of a biological interpretation and pointed out by the analysis, are briefly discussed.

1 Introduction

Both the onset and the rupture of cerebral aneurysms are still matters of research among neuro-surgeons. A cerebral aneurysm is essentially a bulge in the wall of a brain vessel, it is generally not disrupting, and it is not rare among adult population: epidemiological studies suggest that between 1% and 6% of adults develop a cerebral aneurysm during their lives. On the contrary, the rupture of a cerebral aneurysm is quite uncommon but very severe event: about 1 event every 10,000 adults per year with a mortality rate exceeding 50%.

Aim of the Aneurisk Project[1] is to provide evidence of an existing relation between this pathology and the geometry and hemodynamics of brain vessels. In particular, the present analysis considers the centerlines of 65 internal carotid arteries (ICA) whose functional form is obtained from discrete observations by means of free-knot regression splines as shown in Sangalli et al. (2009b). Details about the elicitation of discrete observations from row data can be found in Antiga et al. (2008). Before the analysis, the 65 centerlines are jointly aligned and clustered by means of the *k-mean alignment* method proposed in Sangalli et al. (2010a,b) (in the same works, the properties of the method are widely discussed both from a theoretical point of view and by means of the analysis of synthetic and real data). The aligned and clustered centerlines are then here analyzed along the paradigm of functional data analysis as advocated by Ramsay and Silverman (2005). In the end, some interesting issues amenable of a biological interpretation are discussed.

2 The k-mean Alignment Algorithm

The k-mean alignment algorithm - whose technical and numerical details for practical implementation can be found in Sangalli et al. (2010a,b) - originates from the need of consistently aligning and clustering a set of functional data. This algorithm can be seen as the result of an integration of two algorithms that are currently widely used in functional data analysis: the Procrustes continuous registration algorithm (e.g Sangalli et al. 2009a) and the functional k-mean clustering algorithm (e.g Tarpey and Kinateder 2003). With these two mother algorithms, the new algorithm shares both aims and basic operations. Schematic flowcharts of both algorithms are sketched in Figure 1. Alternative approaches to the joint clustering and alignment of curves can be found in Liu and Muller (2003), Liu and Yang (2009), and Boudaoud et al. (2010).

The aim of the Procrustes continuous alignment algorithm is to decouple phase and amplitude variability; this task is essentially achieved by iteratively performing an *identification step* and an *alignment step.* The former step consists in the identification of a template function on the basis of the n functions as aligned at the previous iteration; the latter step consists instead in the maximization of the similarity between each function and the template, as identified at the previous identification step, by means of subject-by-subject warping of

[1] The project involves MOX Laboratory for Modeling and Scientific Computing (Dip. di Matematica, Politecnico di Milano), Laboratory of Biological Structure Mechanics (Dip. di Ingegneria Strutturale, Politecnico di Milano), Istituto Mario Negri (Ranica), Ospedale Niguarda Ca' Granda (Milano), and Ospedale Maggiore Policlinico (Milano), and is supported by Fondazione Politecnico di Milano and Siemens Medical Solutions Italia.

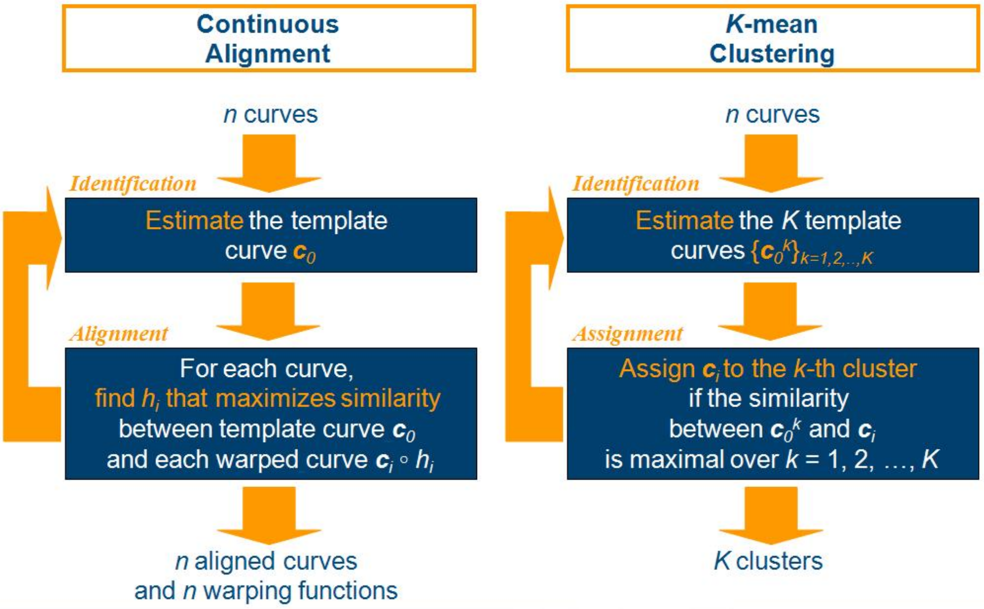
Continuous Alignment
n curves
Identification
Estimate the template curve c_0
Alignment
For each curve, find h_i that maximizes similarity between template curve c_0 and each warped curve $c_i \circ h_i$
n aligned curves and n warping functions
K-mean Clustering
n curves
Identification
Estimate the K template curves $\{c_0^k\}_{k=1,2,..,K}$
Assignment
Assign c_i to the k-th cluster if the similarity between c_0^k and c_i is maximal over k = 1, 2, …, K
K clusters

Figure 1: Schematic flowcharts of the Procrustes continuous registration algorithm (left) and the functional k-mean clustering algorithm (right). Index i refers to the sample unit while index k to the cluster.

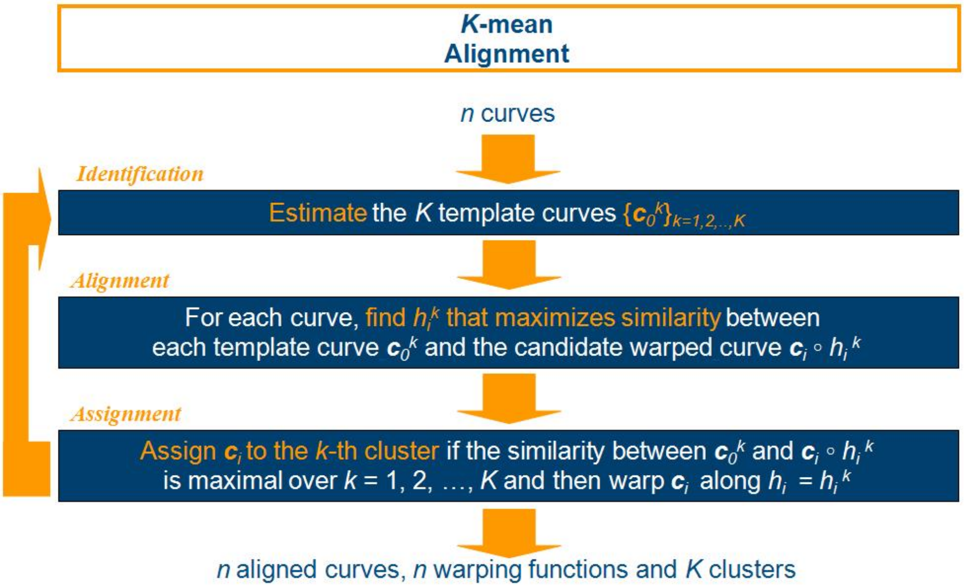
K-mean Alignment
n curves
Identification
Estimate the K template curves $\{c_0^k\}_{k=1,2,..,K}$
Alignment
For each curve, find h_i^k that maximizes similarity between each template curve c_0^k and the candidate warped curve $c_i \circ h_i^k$
Assignment
Assign c_i to the k-th cluster if the similarity between c_0^k and $c_i \circ h_i^k$ is maximal over k = 1, 2, …, K and then warp c_i along $h_i = h_i^k$
n aligned curves, n warping functions and K clusters

Figure 2: Schematic flowchart of the k-mean alignment algorithm. Index i refers to the sample unit while index k to the cluster.

the abscissa. The problem of curve alignment is theoretically well set when a similarity index ρ between two functions and a set W of admissible warping functions of the abscissa are chosen.

The aim of the k-mean clustering algorithm is instead to decouple within and between-cluster variability (in this context within and between-cluster amplitude variability); this task is here achieved by iteratively performing an *identification step* and an *assignment step*. In this algorithm, the identification step consists in the identification of k cluster template functions on the basis of the k clusters detected at the previous iteration; the assignment step consists in the assignment of each function to one of the k clusters, this assignment is achieved by maximizing the similarity between each function and the k templates, as identified at the previous identification step. The problem of clustering curves is theoretically well set when a similarity index ρ between two functions and a number of cluster k to be detected are chosen.

The k-mean alignment algorithm, as a fall out of the two previous algorithms, aims at jointly decoupling phase variability, within-cluster amplitude variability, and between-cluster amplitude variability. It reaches this task by putting together the basic operations of the two mother algorithms (a schematic flowchart of the k-mean alignment algorithm is sketched in Figure 2) and thus iteratively performing an *identification step*, an *alignment step*, and an *assignment step*. Indeed, within the identification step, k template functions are identified on the basis of the k clusters and of the n aligned functions detected at the previous iteration. Within the alignment step the n functions are aligned to the k templates detected at the previous iteration and k candidate aligned versions of each curve are obtained; within the assignment step, each curve is then assigned to the cluster which the curve can be best aligned to, i.e. the cluster for which the similarity among its template and the corresponding candidate aligned curve is maximized.

On the whole, the k-mean alignment algorithm takes as input a set of n functions $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ (like both mother algorithms do) and gives as output k clusters (like the k-mean clustering algorithm does) and n aligned functions together with the corresponding n warping functions $\{h_1, \ldots, h_n\}$ (like the continuous alignment algorithm does).

From a theoretical point of view, the problem of jointly aligning and clustering curves is soundly posed when the number of cluster k, the similarity index ρ between two functions, and the set W of warping functions are chosen. Let us mention two special choices that make the k-mean alignment algorithm degenerate to the two mother algorithms, respectively: $k = 1$ and $W = \{\mathbf{1}\}$. Indeed, if just one single cluster is assumed (i.e. no clustering within the data), the k-mean alignment algorithm turns out to be the continuous alignment algorithm, while if the group of warping functions is assumed to be made just by the identity function (i.e. no phase variability within the data), it turns out to be the k-mean clustering algorithm.

From a practical point of view, different procedures can be used for the implementation of the identification step, of the alignment step, and of the assignment step. In particular, the procedure used within the identification step appears to be very sensitive a point to the good outcome of the k-mean alignment algorithm. To this purpose, in Sangalli et al. (2010a) two different procedures are extensively compared: identification by local regression and identification by means of medoids.

3 Analysis of Internal Carotid Artery Centerlines

In this section we discuss a real application of the k-mean alignment procedure that is also the one that urged us to develop such method: the analysis of the AneuRisk dataset. In detail, we deal with 65 three-dimensional curves, each one representing the centerline of an ICA of a person hospitalized at the Neuro-radiology Department of Ca' Granda Hospital - Niguarda Milan. Details about the elicitation of a discrete representation of the centerline from the three-dimensional angiography can be found in Sangalli et al. (2009a), while the consequent elicitation of the curve - by means of three-dimensional free-knot splines - from the discrete data is detailed in Sangalli et al. (2009b). The idea of the analysis is (i) to perform a k-mean alignment algorithm for different values of k, (ii) compare the performances (measured by means of the mean similarity achieved after the k-mean alignment) to choose a reasonable value for k, and then (iii) find out possible relations between both geometry and cluster membership of the aligned curves on the one hand, and presence, rupture, and location of cerebral aneurysms on the other.

Consistently with Sangalli et al. (2009a), where another analysis of the AneuRisk dataset is presented, we use, as similarity index ρ between two curves $\mathbf{c}_1$ and $\mathbf{c}_2$, the average of the cosine of the angle between the first derivatives of homologous components of the two curves:

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = \frac{1}{3} \sum_{p \in \{x,y,z\}} \frac{\int c'_{1p}(s) c'_{2p}(s) ds}{\sqrt{\int c'_{1p}(s)^2 ds} \sqrt{\int c'_{2p}(s)^2 ds}} \ , \tag{1}$$

and, as the set of warping functions W, we use the group of the affine transformation with positive slope. This joint choice for ρ and W descends from both theoretical and medical issues that are detailed in Sangalli et al. (2009a).

Figure 3 graphically reports the main results of the application of the k-mean alignment algorithm to the analysis of the 65 ICA centerlines. In the

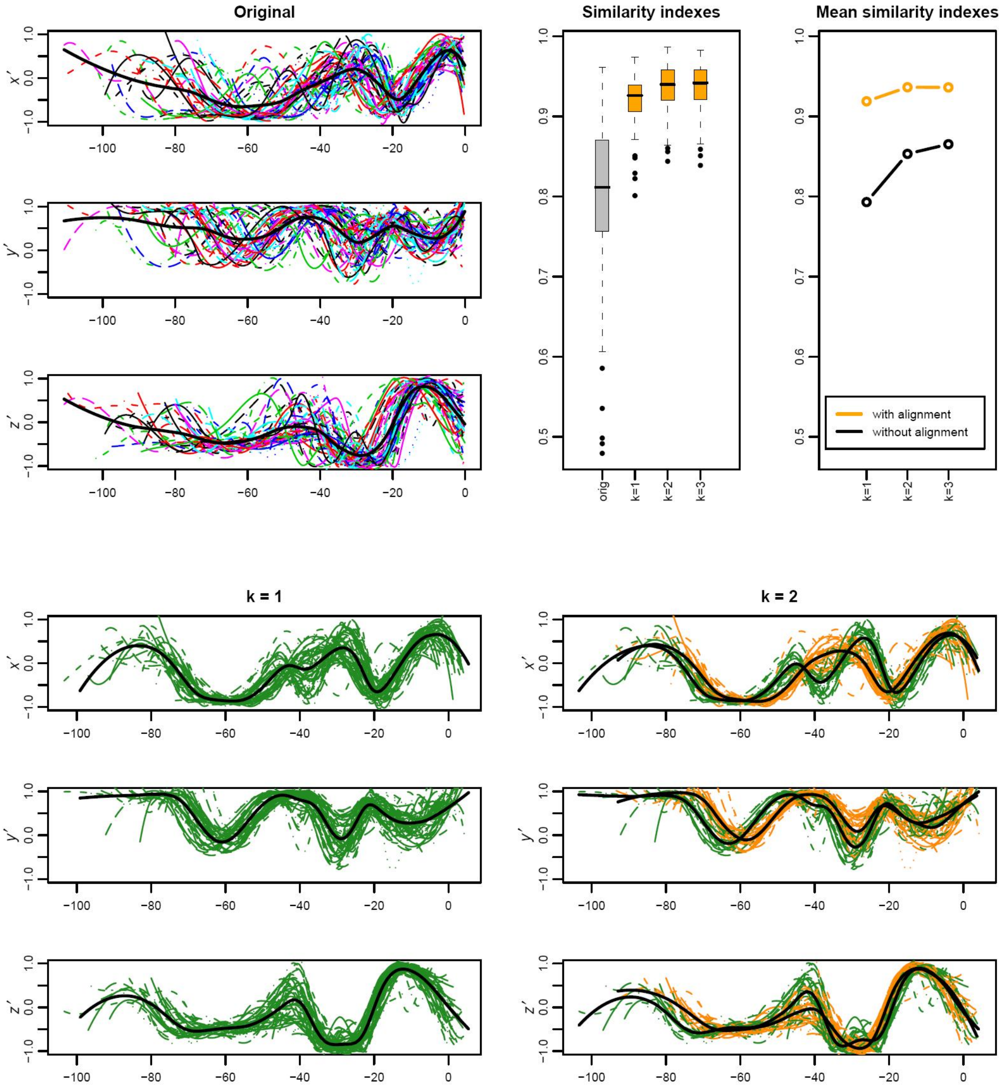

Figure 3: Top left: first derivative of the three spatial coordinates x', y', z' of ICA centerlines. Bottom: first derivatives of one-mean and two-mean aligned curves (left and right respectively); first derivatives of templates always in black. Top center: boxplots of similarity indexes between each curve and the corresponding template for original curves and for k-mean aligned curves, $k = 1, 2, 3$. Top right: means of similarity indexes between each curve and the corresponding template obtained by k-mean alignment and by k-mean without alignment.

top-left plots, the original data are plotted (i.e. first derivative of the three spatial coordinates x', y', z' of 65 ICA centerlines). In the bottom-left plots, the output provided by the one-mean alignment (i.e. first derivative of the three spatial coordinates x', y', z' of 65 ICA centerlines aligned with respect to a single template) is reported. Similarly, in the bottom-right plots, the output provided by the two-mean alignment (i.e. first derivative of the three spatial coordinates x', y', z' of 65 ICA centerlines aligned with respect to two templates) is reported; the two detected clusters are identified by different colors. In the top-center plot: boxplots of similarity indexes between each curve and the corresponding template are reported for original curves and for k-mean aligned curves, $k = 1, 2, 3$. Finally, in the top-right plot, the performances of the algorithm are shown: the orange line reports, as a function of the number of clusters k, the mean of similarity indexes (between curves and the corresponding template) obtained by k-mean alignment; the black line reports the mean of similarity indexes (between curves and the corresponding template) obtained by k-mean clustering without alignment.

Focussing on the last plot, at least two features need to be discussed. Firstly, note the clear vertical shift between the orange and the black line: this points out the presence of a non-negligible phase variability within the original data and thus the necessity of aligning the data before undertaking any further analysis. The important contribution of phase variability to the variability that characterizes these data is also confirmed in Vantini (2009); indeed in the latter work - where consistent notions of total, amplitude, and phase variability are introduced - the phase variability is estimated to contribute to nearly 2/3 of the total variability.

Secondly, once decided that alignment is needed, note the absence in the orange line of an evident improvement in the performance when three clusters are used in place of two: this suggests $k = 2$ to be the correct number of clusters. Consequently, the two-mean alignment algorithm will be used to jointly cluster and align the 65 ICA centerlines.

In the next two subsections, we will discuss some interesting issues amenable of a biological interpretation that the two-mean alignment algorithm points out while neither the simple two-mean clustering without alignment nor the simple one-mean alignment (i.e. continuous alignment) have been able to disclose. In particular, the most interesting finds relative to the association between cluster membership and the aneurysmal pathologies are tackled in Subsection 3.1; the ones relative to the association between the shape of the aligned centerlines and the aneurysmal pathologies are instead shown in Subsection 3.2; the analysis of the warping functions is omitted since no interesting associations have been found.

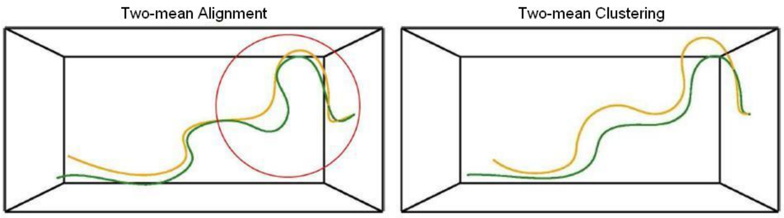

Figure 4: Left: cluster template curves detected by two-mean alignment (S group in green and Ω group in orange). Right: cluster template curves detected by the simple two-mean clustering.

3.1 Centerline Clusters *vs* Cerebral Aneurysms.

Focussing on the two clusters detected by the two-mean alignment algorithm (bottom-right plots of Figure 3), it is noticeable that the two clusters essentially differ within the region between 20 and 50 mm from the end of the ICA, that is also the region where the amplitude variability is maximal within the one-mean aligned data (bottom-left plots of Figure 3). In particular (left plot of Figure 4 where the two cluster templates are reported), we can identify a cluster associated to S-shaped ICAs (two siphons in the distal part of the ICA), i.e. the 30 green curves, and a cluster associated to Ω-shaped ICAs (just one siphon in the distal part of the ICA) i.e. the 35 orange curves. To our knowledge, it is the first time that this categorization, proposed in Krayenbuehl et al. (1982), is statistically detected. To show the primacy of the two-mean alignment, not only over the one-mean alignment but also over the simple two-mean clustering, in Figure 4 the cluster templates detected by two-mean alignment (top) and by the simple two-mean clustering (bottom) are compared. It is evident that while the former algorithm detects two morphologically different templates (the S and the Ω are clearly visible within the red circle), the latter detects two templates that are essentially equal in shape but just shifted. This is not surprising since the two-mean clustering algorithm (that is optimal if no phase variability is present within the data) is completely driven in this case by phase variability, providing fictitious and uninteresting amplitude clusters.

Moreover, the two clusters detected by the two-mean alignment turn out to be associated to the aneurysmal pathology, since there is statistical evidence of a dependence between cluster membership, and aneurysm presence and location (MC simulated p-value of Pearson's χ^2 test for independence equal to 0.0013): indeed, if we label the 65 patients according to the absence of an aneurysm (NO group), the presence of an aneurysm along the ICA (YES-ICA group), and the presence of an aneurysm downstream of the ICA (YES-DS group), we obtain the following conditional contingency table:

	NO	YES-ICA	YES-DS
S	100.0%	52.0%	30.3%
Ω	00.0%	48.0%	69.7%

A close look at the previous table makes evident that: (*i*) within this data set, there are no healthy subjects within the Ω cluster and all healthy subjects belong to the S cluster; (*ii*) within the YES-DS group the number of Ω patients is nearly twice the number of S patients, while within the YES-ICA group the two proportions are nearly equal. Wall shear stress is suspected to be associated to aneurysm onset and rupture and thus vessel geometry and hemodynamics could possibly explain this dependence.

Indeed, both ICA and arteries downstream of the ICA are very stressed vessels from a mechanical point of view: the former because its syphons are expected to act like a fluid dynamical shock-absorber for the brain; the latter because they are floating in the brain humor without being surrounded by any muscle tissues. In particular, while S-shaped ICAs (two syphons) are expected to be very effective in making the flow steadier, Ω-shaped ICAs (one syphon) are instead expected to be less effective (this could be a possible explanation to (*i*)). Moreover for this same reason, in Ω-shaped ICAs, the blood flow downstream of the ICA is expected to be less steady, providing an overloaded mechanical stress to downstream arteries (this could be an explanation to (*ii*)).

3.2 Centerline Shapes *vs* Cerebral Aneurysms.

Let us now focus on the two-mean aligned curves in order to find out possible relations between centerline geometry and aneurysms. In order to reduce data dimensionality, we perform a three-dimensional functional principal component analysis (e.g. Ramsay and Silverman 2005) of the aligned centerlines for values of the registered abscissa between -34.0 and -6.9 mm, i.e. the abscissa interval where all records are available. In the right plot of Figure 5 the fractions and the cumulative fractions of explained total variance are displayed, it is evident that one, three, or five principal components can be used to represent the centerlines. We decide to summarize the data by means of the first five principal components comforted by the fact that they provide a visually good representation of the data, by the fact that they explain more than the 90% of the total variance, and by the fact that all remaining principal components seem not to be related to any structural mode of variability but just noise.

In the left plot of Figure 5 the projections of the 65 ICA centerlines along the first principal component are reported (orange for the Ω cluster centerlines and green for the S cluster ones). Nearly 42% of the total variability is explained by this component. It is evident that the variability associated to the first compo-

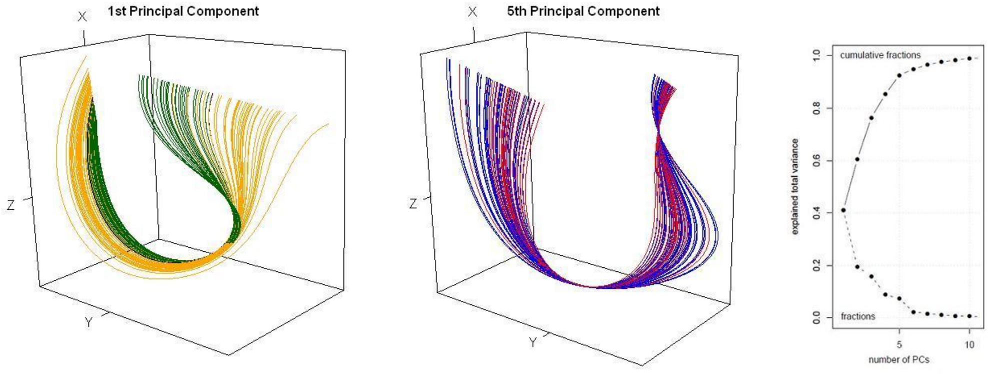

Figure 5: Left: the projections of the 65 ICA centerlines along the first principal component (in orange centerlines belonging to the Ω cluster and in green centerlines belonging to the S cluster). Center: the projections of the 65 ICA centerlines along the fifth principal component (in red centerlines associated to patients with a ruptured aneurysm and in blue patients without aneurysm or with unruptured aneurysm). Right: fractions of explained total variance.

nent is mostly concentrated at the top-right extreme (i.e. the proximal part of the portion of centerline under investigation), and moreover it is indicating the presence and magnitude of a second syphon before the distal one (in this picture blood flows from right to left). The Mann-Whitney test for the first principal component scores of the S and the Ω cluster centerline projections presents a p-value equal to 10^{-14}. This result strongly supports the identification of the two clusters - detected by the two-mean alignment - with the S and Ω shaped ICAs proposed by Krayenbuehl et al. (1982).

The second, third, and fourth principal components result difficult to interpret and moreover no associations have been found between the latter ones and the aneurysmal pathologies. For this reason they will not be discussed in this work.

The fifth principal component (explained total variance 7%, cumulative 93%) appears instead to be surprisingly easy to interpret (in the center plot of Figure 5 the projections of the 65 ICA centerlines along the fifth principal component are reported: in red the centerlines associated to patients with a ruptured aneurysm and in blue the ones associated to patients without aneurysm or with unruptured aneurysm). Indeed, it expresses the prominence of the distal syphon, i.e., along the fifth principal component, ICA centerlines progressively evolve from having a very sharped distal syphon (lower scores) toward smoother distal syphons (higher scores). It is known that the more curved the vessel is, the higher the vorticity in the fluid and the shear stress on the wall are. Analyzing the scores relevant to the fifth components, we find that patients with a ruptured aneurysm present significant lower scores than patients with an unruptured aneurysm or without aneurysm (Mann-Whitney test p-value 0.0072), i.e. the former ones present more marked syphons than the latter ones. These results could support the idea of a fluid dynamical origin of the onset and/or rupture of cerebral aneurysms.

All these fluid dynamical hypotheses are going to be supported, in the future, by fluid dynamical simulations in order to provide a mechanical interpretation of the relation between geometry and hemodynamics on one side, and aneurysm onset and rupture on the other, that this analysis partially already highlights.

4 Conclusions

We showed in this work an application of the k-mean alignment method proposed in Sangalli et al. (2010b) that jointly clusters and aligns curves. This method puts in a unique framework two widely used methods of functional data analysis: functional k-mean clustering and Procrustes continuous alignment. Indeed, these latter two methods turn out to be two special cases of the new one.

In particular, we used this method to perform a functional data analysis of 65 three-dimensional curves representing 65 internal carotid artery centerlines. In this application the k-mean alignment algorithm outdoes both functional k-mean clustering and Procrustes continuous alignment by pointing out interesting features from a medical and fluid dynamical point of view that former methods were not able to point out.

References

Antiga, L., Piccinelli, M., Botti, L., Ene-Iordache, B., Remuzzi, A., and Steinman, D. (2008), "An image-based modeling framework for patient-specific computational hemodynamics," *Medical and Biological Engineering and Computing*, 1097–112.

Boudaoud, S., Rix, H., and Meste, O. (2010), "Core Shape modelling of a set of curves," *Computational Statistics and Data Analysis*, 308–325.

Krayenbuehl, H., Huber, P., and Yasargil, M. G. (1982), *Krayenbuhl/Yasargil Cerebral Angiography*, Thieme Medical Publishers, 2nd ed.

Liu, X. and Muller, H. G. (2003), "Modes and clustering for time-warped gene expression profile data," *Bioinformatics*, 19, 1937–1944.

Liu, X. and Yang, M. C. K. (2009), "Simultaneous curve registration and clustering for functional data," *Computational Statistics and Data Analysis*, 1361–1376.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer New York NY, 2nd ed.

Sangalli, L. M., Secchi, P., Vantini, S., and Veneziani, A. (2009a), "A case study in exploratory functional data analysis: geometrical features of the internal carotid artery," *Journal of the American Statistical Association*, 104, 37–48.

— (2009b), "Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines applied to the analysis of inner carotid artery centrelines," *Journal of the Royal Statistical Society, Ser. C, Applied Statistics*, 58, 285–306.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010a), "Functional clustering and alignment methods with applications," Tech. Rep. 5/2010, MOX, Dip. di Matematica, Politecnico di Milano.

— (2010b), "K-mean alignment for curve clustering," *Computational Statistics and Data Analysis*, available at http://dx.doi.org/10.1016/j.csda.2009.12.008.

Tarpey, T. and Kinateder, K. K. J. (2003), "Clustering functional data," *Journal of Classification*, 20, 93–114.

Vantini, S. (2009), "On the definition of Phase and Amplitude Variability in Functional Data Analysis," Tech. Rep. 33/2009, MOX, Dip. di Matematica, Politecnico di Milano.

MOX Technical Reports, last issues

Dipartimento di Matematica “F. Brioschi”,
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

09/2010 Laura M. Sangalli, Piercesare Secchi, Simone Vantini, Valeria Vitelli:
Joint Clustering and Alignment of Functional Data: an Application to Vascular Geometries

08/2010 Francesca Ieva, Anna Maria Paganoni:
Multilevel models for clinical registers concerning STEMI patients in a complex urban reality: a statistical analysis of MOMI^2 survey

07/2010 Laura M. Sangalli, Piercesare Secchi, Simone Vantini, Valeria Vitelli:
Functional clustering and alignment methods with applications

06/2010 Jordi Alastruey, Tiziano Passerini, Luca Formaggia, Joaquim Peiró:
The effect of visco-elasticity and other physical properties on aortic and cerebral pulse waveforms: an analytical and numerical study

05/2010 Matteo Longoni, A.C.I. Malossi, Alfio Quarteroni, Andrea Villa:
A complete model for non-Newtonian sedimentary basins in presence of faults and compaction phenomena

04/2010 Marco Discacciati, Paola Gervasio, Alfio Quarteroni:
Heterogeneous mathematical models in fluid dynamics and associated solution algorithms

03/2010 P.E. Farrell, Stefano Micheletti, Simona Perotto:
A recovery-based error estimator for anisotropic mesh adaptation in CFD

02/2010 Pietro Barbieri, Niccolo’ Grieco, Francesca Ieva, Anna Maria Paganoni and Piercesare Secchi:
Exploitation, integration and statistical analysis of Public Health Database and STEMI archive in Lombardia Region

01/2010 G.M. Porta, S. Perotto, F. Ballio:
Anisotropic Mesh Adaptation Driven by a Recovery-Based Error Estimator for Shallow Water Flow Modeling

35/2009 C. D'Angelo, P. Zunino:
Robust numerical approximation of coupled Stokes and Darcy flows applied to vascular hemodynamics and biochemical transport