

Joint Clustering and Alignment of Functional Data: an Application to Vascular Geometries ¹

Simone Vantini ²

MOX - Dept. of Math. F. Brioschi, Politecnico di Milano, e-mail:
simone.vantini@polimi.it

Abstract: We propose a novel method for jointly clustering and aligning functional data: *k-mean alignment*. 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. This analysis originates from a joint work with P. Secchi, L. M. Sangalli, and V. Vitelli.

Keywords: Functional Data Analysis, Clustering, Alignment, Aneurysm.

1. Introduction

The onset and the rupture of cerebral aneurysms is still a debated matter among neurosurgeons. This disease consists in 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 an event: about 1 event every 10,000 adults per year with a mortality rate exceeding 50%.

Aim of the Aneurisk Project is to provide evidence of an existing relation between this pathology, and the geometry and the hemodynamics of brain vessels. In particular, the present analysis considers the centerlines of 65 internal carotid arteries (ICA) along the paradigm of a two-stage functional data analysis as advocated by Ramsay and Silverman (2005); this work focuses on the second stage of the analysis neglecting any details related to the first stage in which functional data are obtained from discrete observations. In detail, we will present a novel algorithm (*k-mean alignment*) that jointly aligns and clusters the centerlines pointing out interesting issues amenable of a biological interpretation.

2. The *k*-mean Alignment Algorithm

Let $\{c_1, \dots, c_N\}$ be the N curves to be clustered and aligned, $\mathcal{C}$ the set which the curves belong to, W the group of the warping functions, and ρ a index measuring the similarity between any two curves in $\mathcal{C}$. Given some technical properties detailed in Sangalli *et al.* (2008), inspired by functional *k*-mean clustering algorithms and by Procruste methods for functional data alignment, we set the problem of jointly aligning and clustering the

²Address of correspondence: MOX - Dept. of Mathematics “Francesco Brioschi”, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133, Milan, Italy

N curves like the problem of solving the following optimization problem: find $\underline{\varphi} = \{\varphi_1, \dots, \varphi_k\} \subset \mathcal{C}$ and $\underline{h} = \{h_1, \dots, h_N\} \subset W$ such that

$$\frac{1}{N} \sum_{i=1}^N \rho(\varphi_{\lambda(\underline{\varphi}, \mathbf{c}_i)}, \mathbf{c}_i \circ h_i) \geq \frac{1}{N} \sum_{i=1}^N \rho(\psi_{\lambda(\underline{\psi}, \mathbf{c}_i)}, \mathbf{c}_i \circ g_i) \quad (1)$$

for any other set of k templates $\underline{\psi} = \{\psi_1, \dots, \psi_k\} \subset \mathcal{C}$ and any other set of N warping functions $\underline{g} = \{g_1, \dots, g_N\} \subset W$. The labelling function $\lambda(\underline{\psi}, \mathbf{c}_i)$ identifies the element of the set $\underline{\psi}$ which the curve $\mathbf{c}_i$ can be better aligned to. Finding the solution of the optimization problem (1) is generally non trivial. Thus, we face this problem by means of an iterative algorithm. Three different steps are cyclically performed till convergence:

Template identification step. The k templates are estimated using the k clusters and the warped curves from the previous iteration;

Cluster assignment and alignment step. Curves are assigned to the k clusters and aligned according to the k templates computed at the previous step;

Normalization step. The warping functions detected at the previous step are normalized such that the curves are on average not warped.

Note that the k -mean alignment algorithm degenerates to a Procruste method for aligning curves if $k = 1$ (i.e. no clusters assumed) and to a functional k -mean clustering algorithm if $W = \{1\}$ (i.e. no phase variability assumed). Both theoretical details and technical details for the practical implementation of the algorithm can be found in Sangalli *et al.* (2008).

3. Analysis of Internal Carotid Artery Centerlines

In this section we discuss a real application of the k -mean alignment procedure: 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 alignment) to choose a reasonable value for k , and then (iii) find out possible relationships 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.

Figure 1 (top-right) shows the means of similarity indexes obtained by k -mean alignment (orange) and by k -mean without alignment (black) for $k = 1, 2, 3$ using the same ρ and W introduced in Sangalli *et al.* (2009a) are used in the analysis. Note the vertical shift between the orange and the black lines: this points out the presence of a non-negligible phase variability within the original data. In addition, focusing on the orange line, note the absence 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 2-mean alignment algorithm has been used to jointly cluster and align the 65 centerlines. The 2-mean alignment algorithm pointed out some interesting issues amenable of

a biological interpretation that neither a simple k -mean analysis without alignment nor a simple 1-mean alignment have been able to disclose.

Clustering. Focussing on the clusters, we can identify a cluster associated to S -shaped ICAs (the 30 orange curves in the bottom right of Figure 1) and a cluster associated to Ω -shaped ICAs (the 35 green curves in the bottom right of the same figure). To our knowledge, it is the first time that this categorization, proposed in Kraysenbuehl *et al.* (1982), is statistically detected. Moreover, there is statistical evidence of a dependence between cluster membership, and aneurysm presence and location (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 (group N), the presence of an aneurysm along the ICA (group YICA), and the presence of an aneurysm downstream of the ICA (group YDS), we obtain the following conditional contingency table:

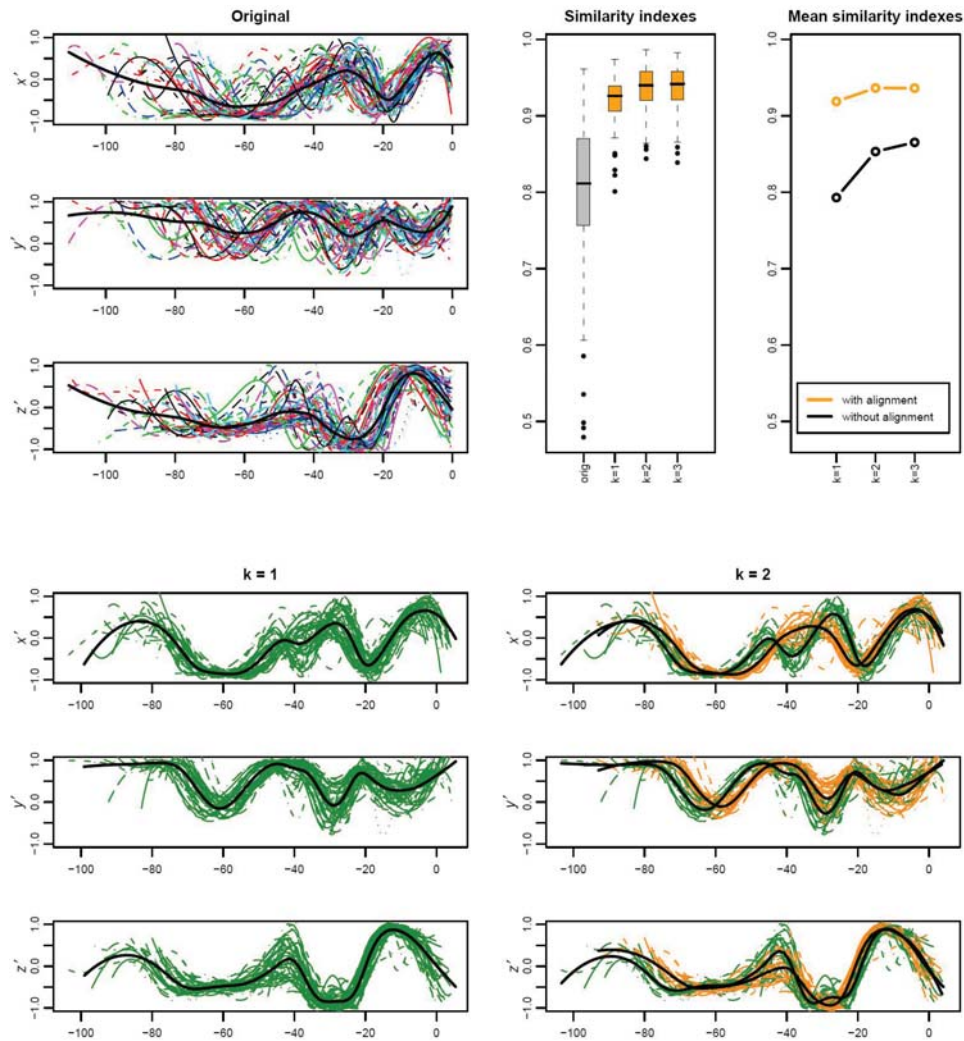
	N	YICA	YDS
S	23.3%	43.3%	33.3%
Ω	00.0%	34.3%	65.7%

It is apparent that a “drift” from class N to class YICA and then YDS occurs while passing from cluster S to cluster Ω . Wall shear stress is suspected to be associated to aneurysm onset and rupture and thus hemodynamics and vessel mechanics could possibly explain this drift. 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 being not surrounded by any muscle tissues. While S -shaped ICAs (two syphons) are expected to be very effective in making the flow steadier (that could be the reason why in this cluster we find people without aneurysm), Ω -shaped ICAs (one syphon) are instead expected to be less effective (that could be the reason why there are no patients without aneurysm within this cluster). 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.

Alignment. Let us now focus on the aligned curves in order to find out possible relationships between centerline geometry and aneurysms. We performed a three-dimensional functional principal component analysis (FPCA) of the aligned centerlines for values of the registered abscissa between -34.0 and -6.9 mm. The fifth principal component (explained total variance 7%, cumulative 93%) appears to be surprisingly easy to interpret: it expresses the prominence of the distal syphon, i.e., lower scores are associated to marked bends while higher to less marked bends. 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 found statistical evidence (Mann-Whitney test p -value 0.0072) that patients with a ruptured aneurysm present more marked syphons with respect to patients with an unruptured aneurysm or without aneurysm.

All these fluid dynamical hypotheses are going to be supported by fluid dynamical simulation 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 highlighted.

Figure 1: k -mean alignment of the AneuRisk dataset.



References

- Krayenbuehl H., Huber P. and Yasargil M.G. (1982) *Krayenbuehl/Yasargil Cerebral Angiography*, Thieme Medical Publishers, 2 edition.
- Ramsay J.O. and Silverman B.W. (2005) *Functional Data Analysis*, Springer New York NY, 2nd edition.
- 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, 485, 37–48.
- Sangalli L.M., Secchi P., Vantini S. and Veneziani A. (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. (2008) K-mean alignment for curve clustering, Technical Report 13/2008, MOX, Dip. di Matematica, Politecnico di Milano.