

Functional clustering and alignment

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

Abstract The problem of clustering functional data is considered, taking into account possible misalignment of the data. In the proposed setting, joint alignment and clustering of curves is defined via the solution of an optimization problem. This is solved through an iterative procedure, which is detailed in two alternative specifications: a k -mean version and a k -medoid version. The robustness of the alignment and clustering procedure under the different specifications is illustrated via applications to cerebral vascular geometries.

1 Introduction

A frequently encountered problem in functional data analysis is unsupervised classification of a set of curves. Various authors solve the problem of curve clustering developing functional k -mean inspired procedures (see, e.g., [1, 11]). However, when dealing with clustering of curves, we need also to consider a problem which is peculiar to functional data, namely the possible *misalignment* of the data. An instance of this issue is given by the data in Figure 1, which are related to the three spatial coordinates of 65 Internal Carotid Artery (ICA) centerlines (the picture in particular shows the estimated first derivatives of the three-dimensional centerlines). Clustering of these data is of interest for the identification of ICA's with different morphological shapes; however, the evident misalignment of the data acts as a confounding factor when trying to cluster the data, and a plain functional k -mean algorithm would fail to give efficient results (see [10]). This highlights the need for a clustering procedure which is able to jointly deal with data alignment.

The problem of curve alignment (or curve registration) has been considered by a number of authors. Following a line of research advocated by J. O. Ramsay (see, e.g., [5, 6, 7]), alignment can be performed defining a suitable similarity index and

Laura M. Sangalli, Piercesare Secchi, Simone Vantini and Valeria Vitelli
MOX, Dipartimento di Matematica, Politecnico di Milano, via Bonardi 9, 20133, Milano.
e-mail: laura.sangalli@polimi.it, piercesare.secchi@polimi.it, simone.vantini@polimi.it, valeria.-vitelli@mail.polimi.it

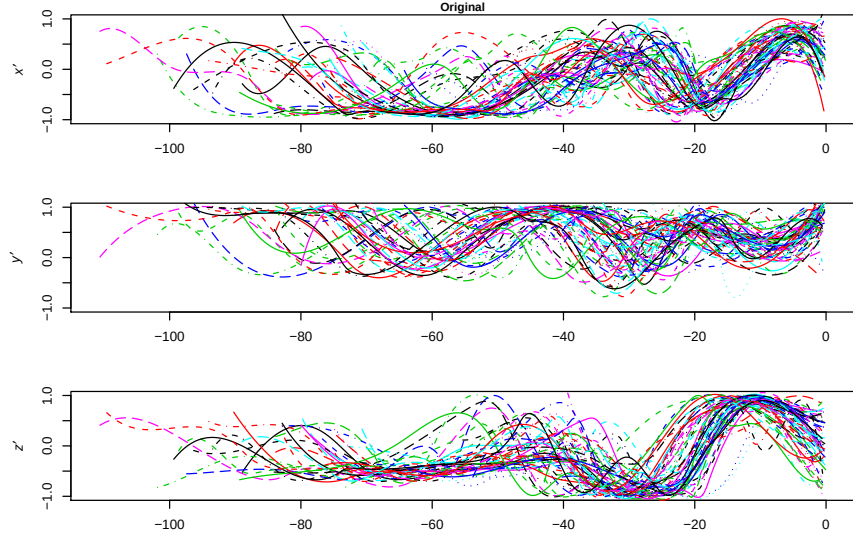


Fig. 1 Estimated first derivatives, x' , y' , z' , of the three spatial coordinates of the 65 ICA centerlines.

a class of transformations of the abscissa, and then finding a reference curve maximizing the total similarity.

In [10] we proposed a procedure which is able to *jointly* cluster and align a set of functional data. We stated the problem of joint clustering and alignment of functional data as an optimization problem, and we proposed an iterative procedure for its solution. We developed two alternative specifications of this same procedure, a *k-mean algorithm* version and a *k-medoid algorithm* version, which only differ in the chosen approach to the template identification. In the former, fully described in [10], the template is estimated by means of a loess, in analogy with the multivariate and functional *k-mean* algorithms. In the latter, introduced in [9], the template is chosen among the curves belonging to the cluster as the one that maximizes the total similarity, thus reducing the potential sensitivity to the presence of anomalous data.

2 Curve clustering when curves are misaligned

The variability among two or more curves can be thought of as having two components: *phase variability* and *amplitude variability*. Heuristically, the aim of alignment procedures is to fully capture phase variability, while the focus of clustering procedures is on amplitude variability. However, in order to avoid that one of the two variabilities acts as a confounding factor while the analysis is trying to capture the other, we shall decouple them jointly through a clustering and alignment procedure.

Two basic elements have to be defined in order to fully describe an alignment Procrustes procedure:

- a *similarity index*, $\rho(\cdot, \cdot)$, which measures the similarity between two curves;
- a *class W of admissible warping functions*, indicating the allowed transformations for the abscissa.

Then, aligning $\mathbf{c}_1$ to $\mathbf{c}_2$, according to (ρ, W) , means finding $h^* \in W$ that maximizes $\rho(\mathbf{c}_1 \circ h, \mathbf{c}_2)$.

Here we shall consider the following bounded similarity index, introduced in [7], between two curves $\mathbf{c}_1, \mathbf{c}_2 \in H^1(\mathbb{R}; \mathbb{R}^d)$, not constant:

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = \frac{1}{d} \sum_{p=1}^d \frac{\int_{\mathbb{R}} c'_{1p}(s) c'_{2p}(s) ds}{\sqrt{\int_{\mathbb{R}} c'^2_{1p}(s) ds} \sqrt{\int_{\mathbb{R}} c'^2_{2p}(s) ds}}, \quad (1)$$

with c_{ip} indicating the p th component of $\mathbf{c}_i$. Geometrically, $\rho(\mathbf{c}_1, \mathbf{c}_2)$ is the average of the cosines of the angles between the derivatives of homologous components of $\mathbf{c}_1$ and $\mathbf{c}_2$. The two curves are said to be *similar* when the index assumes its maximal value 1. The choice of the similarity index $\rho(\cdot, \cdot)$ expressed in (1) comes along with the following definition for the class W

$$W = \{h : h(s) = ms + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\}, \quad (2)$$

the group of strictly increasing affine transformations. The couple (ρ, W) satisfies some minimal requirements for coherence, necessary to ensure the well-posedness of the clustering and alignment optimization problem, and hence the possibility to obtain a coherent procedure; they are described in [10].

2.1 Clustering and alignment iterative procedure

Consider the problem of clustering and aligning a set of n curves $\{\mathbf{c}_1, \dots, \mathbf{c}_n\} \subset H^1(\mathbb{R}; \mathbb{R}^d)$. Ideally, in order to solve the problem of clustering and alignment, we should find the k template curves (one for each of the k clusters) and the n warping functions that maximize the total similarity between each warped curve and the most similar template. Since this problem is highly non-linear and computationally intensive, we choose to deal with it through an iterative procedure.

A schematic flowchart of the proposed clustering and alignment procedure is sketched in Figure 2: given a set of n curves, the procedure iteratively alternates *template identification steps* and *assignment and alignment steps*. Indeed, within the template identification step, k template curves are estimated on the basis of the cluster assignments and of the warped curves detected at the previous iteration; in the alignment and assignment step, the n curves are respectively aligned to one of the k templates and assigned to the corresponding cluster. More precisely, the latter is split in two sub-steps, *alignment* and *assignment*: in the first sub-step, for

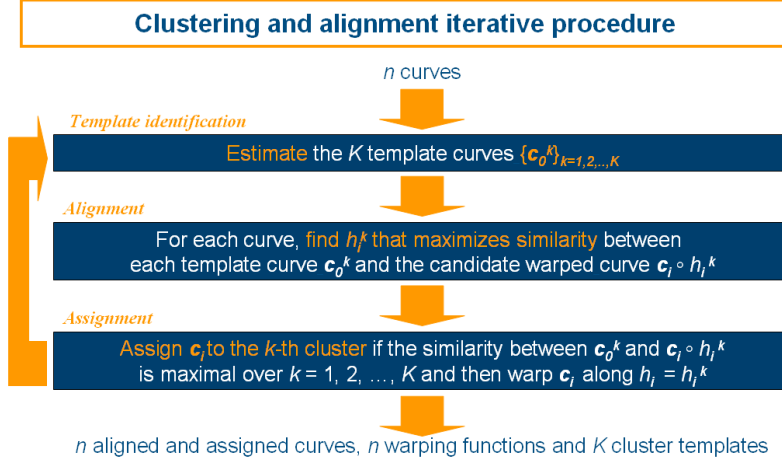


Fig. 2 Schematic flowchart of the clustering and alignment general iterative procedure. Index i refers to the sample unit while index k to the cluster.

each curve, a warping function is found which maximizes the similarity between the warped curve and each of the templates; in the second sub-step, each curve is assigned to the cluster to whose corresponding template is most similar, and warped along the associated warping function.

2.2 Template identification

While the assignment and alignment steps constitute a classical maximization problem, the template identification step can be further discussed from a statistical perspective. In particular, we should estimate the k templates as the curves that maximize within-cluster similarity among all curves belonging to $H^1(\mathbb{R}; \mathbb{R}^d)$. The approximation of this infinite dimensional identification problem has resulted in two different specifications of the clustering and alignment procedure, and hence in two different algorithms:

k-mean alignment algorithm. At each iteration the template is estimated by means of a loess, with Gaussian kernel and appropriate smoothness parameter, of the curves assigned to the cluster at the previous iteration (see [10] for details on the implementation). The algorithm obtained with this specification for the template identification step is named k -mean alignment, in analogy with the k -mean clustering algorithms for multivariate and functional data.

k-medoid alignment algorithm. A more direct approximation to the maximization of the total similarity, consists in restricting the set over which the maximization is carried out, limiting the search to functions belonging to the cluster (see [9] for details on the implementation). The algorithm obtained with this specification is named *k-medoid alignment*, in analogy with the *k-medoid clustering* algorithms described for instance in [2, 3].

3 Analysis of cerebral vascular geometries

The described clustering and alignment procedure has been successfully applied in the analysis of simulated data, and of benchmark datasets for functional data analysis, like the growth curves from Berkeley Growth Study (further details can be found in [10, 12]).

In this Section, we describe the results obtained by *k-mean* and *k-medoid alignment* on the AneuRisk Project¹ dataset (previous exploratory statistical analyses of this dataset are detailed in [7, 10]). Aim of the AneuRisk Project is the evaluation of the possible effects of vascular geometry and hemodynamics on the disrupting pathology of cerebral aneurysm. The data considered in the analysis here presented are the three spatial coordinates (in mm) of 65 Internal Carotid Artery (ICA) centerlines, measured on a fine grid of points along a curvilinear abscissa (in mm), decreasing from the terminal bifurcation of the ICA towards the heart. Estimates of these three-dimensional curves are obtained by means of three-dimensional free-knot regression splines, as described in [8]. Figure 1 displays the first derivatives, x' , y' and z' , of the estimated centerlines. We are interested in clustering the three-dimensional centerlines, with the aim of identifying ICA's with different morphological shapes. Since the shape of the ICA influences the pathogenesis of cerebral aneurysms through its effects on the hemodynamics, such a classification could in fact be helpful in the determination of the risk level of a given patient.

In the medical literature (see, e.g., [4]) ICA's are classified in Γ -shaped, Ω -shaped, and S -shaped ICA's, according to the form of their distal part, which may resemble the letters Γ , Ω or S , in presence of zero, one, or two siphons, respectively. We want here to discuss, in the light of this classification, the clustering results obtained by *k-mean* and *k-medoid alignment*. Details on these analyses can be found in [9, 10].

¹ The project is a joint research program which 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.

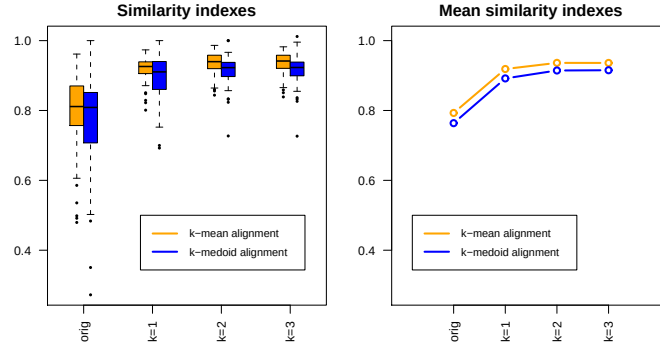


Fig. 3 Left: in orange (blue), boxplots of similarity indexes between the original centerlines and their mean (medoid) curve, “orig”, and boxplots of similarity indexes between the k -mean aligned (k -medoid aligned) centerlines and their estimated templates, for $k = 1, 2, 3$. Right: corresponding means of similarity indexes.

Figure 3, left, shows in orange the boxplots of the similarity indexes between the original centerlines and their estimated loess (indicated in the figure as “orig”), and the boxplots of the similarity indexes between the k -mean aligned centerlines and the associated estimated templates, for $k = 1, 2, 3$. In the same panel are also displayed in blue the boxplots of the similarity indexes obtained by k -medoid alignment. Figure 3, right, displays the corresponding means of the similarity indexes, in orange for k -mean alignment and in blue for k -medoid alignment. Note that both 1-mean alignment and 1-medoid alignment lead to a large increase in the similarities, with respect to the similarities of the original curves, but a further non-neglectable gain can be obtained in both cases by setting $k = 2$; indeed, an eventual choice of $k = 3$ is not justified by an additional increase in the similarities. Thus, the clustering and aligning procedure suggests the presence of $k = 2$ clusters within the analyzed centerlines.

Figure 4 compares the first derivatives of 1-mean and 1-medoid aligned centerlines (top left and right, respectively), and the first derivatives of 2-mean and 2-medoid aligned centerlines (bottom left and right, respectively). Both 1-mean and 1-medoid alignment final templates are not able to fully capture variability in the data, in particular for values of the abscissa between -20 and -50. Instead, the bottom panels of Figure 4 show better results. Moreover, the two clusters identified by both 2-mean and 2-medoid alignment can be described as the Ω -shaped ICA’s cluster (orange) and S -shaped ICA’s cluster (green).

This can be better appreciated in Figures 5 and 6, that give three-dimensional visualizations of the two estimated template curves, obtained by 2-mean and 2-medoid alignment, respectively. In both figures, in fact, the template of the orange cluster is a prototype of Ω -shaped ICA (one siphon in the distal part, red circle), whereas the template of the green cluster is a prototype of S -shaped ICA (two siphons in the distal part, red circle).

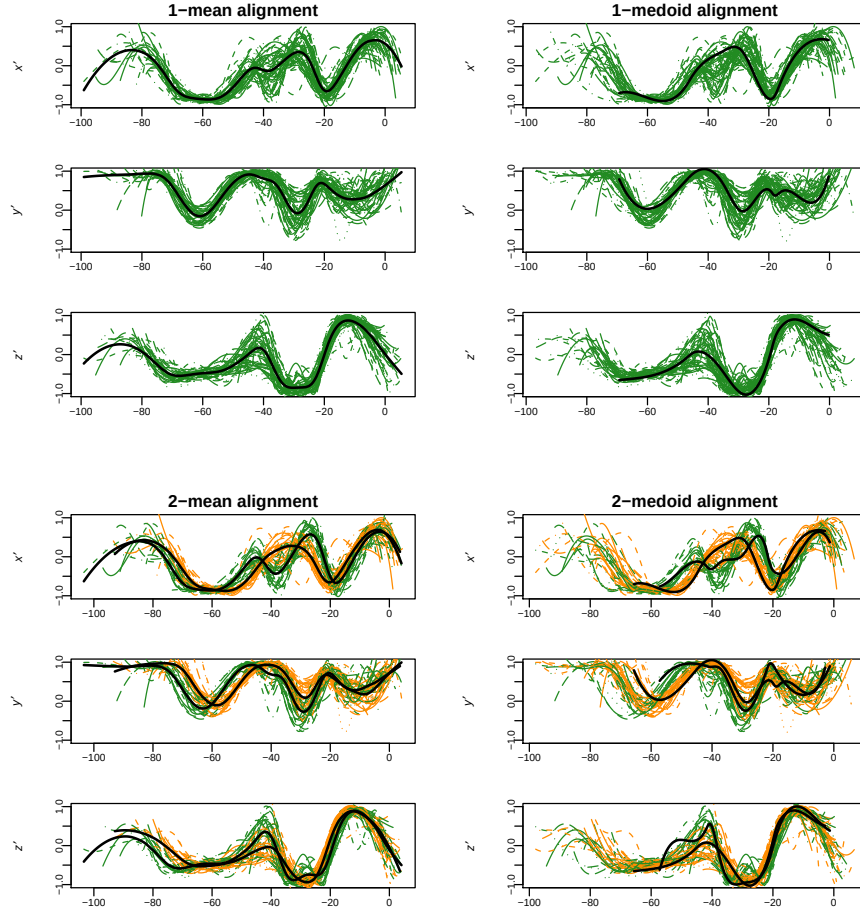


Fig. 4 First derivatives of 1-mean (top left) and 1-medoid (top right) aligned centerlines, and of 2-mean (bottom left) and 2-medoid (bottom right) aligned centerlines, with superimposed first derivatives of estimated templates (black lines); the curve color depends on cluster assignment.

Table 1 Comparison between cluster assignments obtained by 2-mean and 2-medoid alignment algorithms (“ Ω ” stands for the Ω -shaped cluster and “S” stands for the S-shaped one).

		2-medoid	
		Ω	S
2-mean	Ω	34	1
	S	3	27

Finally, Table 1 compares the cluster assignments obtained by 2-mean and 2-medoid alignment (“ Ω ” stands for the cluster of Ω -shaped ICA’s, and “S” for the one of S-shaped ICA’s). This table shows that only 4 out of the 65 ICA’s are differently clustered by the two algorithms, thus confirming the robustness of our clustering and alignment procedure under different specifications of the template identifi-

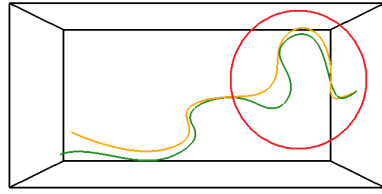


Fig. 5 Three-dimensional image of the estimated templates found by 2-mean alignment of ICA centerlines. The template of the orange cluster is a prototype of Ω -shaped ICA (one siphon), whereas the one of the green cluster is a prototype of S-shaped ICA (two siphons).

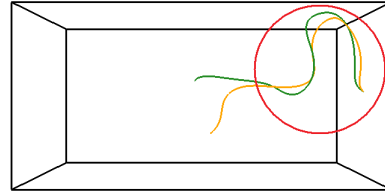


Fig. 6 Three-dimensional image of the estimated templates found by 2-medoid alignment of ICA centerlines. The template of the orange cluster is a prototype of Ω -shaped ICA (one siphon in the distal part), whereas the one of the green cluster is a prototype of S-shaped ICA (two siphons in the distal part).

cation step. Two of these four data might in fact be interpreted as anomalous data, one because of the shortness of the observed centerline, and the other because of the morphological shape, that resembles a Γ -shaped ICA rather than a Ω or a S-shaped one.

References

1. Chiou, J.M., Li, P.L.: Functional clustering and identifying substructures of longitudinal data. *J. R. Statist. Soc. Series B* (69), 679–699 (2007)
2. Hastie, T., Tibshirani, R., J.Friedman: *The Elements of Statistical Learning*. Springer (2001)
3. Kaufman, L., Rousseeuw, P.: *Finding Groups in Data*. Wiley Series in Probability and Mathematical Statistics (1990)
4. Krabenbuehl, H., Huber, P., Yasargil, M.G.: *Krabenbuehl/yasargil cerebral angiography*. Thieme Medical Publishers, 2nd ed. (1982)
5. Ramsay, J.O., Li, X.: Curve registration. *J. R. Stat. Soc. Ser. B Stat. Methodol.* (60), 351–363 (1998)
6. Ramsay, J.O., Silverman, B.W.: *Functional Data Analysis*. Springer (2005)
7. Sangalli, L.M., Secchi, P., Vantini, S., A.Veneziani: A case study in exploratory functional data analysis: Geometrical features of the internal carotid artery. *J. Amer. Statist. Assoc.* **104**, 37–48 (2009)
8. Sangalli, L.M., Secchi, P., Vantini, S., A.Veneziani: 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 (2009)
9. Sangalli, L.M., Secchi, P., Vantini, S., Vitelli, V.: Functional clustering and alignment methods with applications. Tech. Rep. 5, MOX, Dipartimento di Matematica, Politecnico di Milano (2010)
10. Sangalli, L.M., Secchi, P., Vantini, S., Vitelli, V.: K-mean alignment for curve clustering. *Computational Statistics and Data Analysis* (2010, doi:10.1016/j.csda.2009.12.008)
11. Tarpey, T., Kinateder, K.K.J.: Clustering functional data. *Journal of Classification* (20), 93–114 (2003)
12. Tuddenham, R.D., Snyder, M.M.: Physical growth of california boys and girls from birth to age 18. Tech. Rep. 1, University of California Publications in Child Development (1954)