Electronic Journal of Statistics
Vol. 8 (2014) 1891–1904
ISSN: 1935-7524
DOI: 10.1214/14-EJS938A

Analysis of AneuRisk65 data: k-mean alignment*

Laura M. Sangalli†, Piercesare Secchi and Simone Vantini

MOX – Department of Mathematics, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133, Milano, Italy
e-mail: laura.sangalli@polimi.it; piercesare.secchi@polimi.it; simone.vantini@polimi.it

Abstract: We describe the k-mean alignment procedure, for the joint alignment and clustering of functional data and we apply it to the analysis of the AneuRisk65 data. Thanks to the efficient separation of the variability in phase variability and within/between clusters amplitude variability, we are able to discriminate subjects having aneurysms in different cerebral districts and identifying different morphological shapes of Inner Carotid Arteries, unveiling a strong association between arteries morphologies and the aneurysmal pathology.

Keywords and phrases: k-mean alignment, registration, functional clustering, AneuRisk65 data.

Received August 2013.

1. k-mean alignment

We here summarize the k-mean alignment procedure that we shall use in Section 2 to analyze the AneuRisk65 data described in Sangalli, Secchi and Vantini (2013). This procedure, introduced in Sangalli et al. (2010), is able to efficiently align and cluster in k groups a set of curves. The procedure can be seen as a continuous alignment with $k \geq 1$ templates, or equivalently as a k-mean clustering of curves with warping allowed. In fact, if the number of clusters k is set equal to 1, the algorithm implements the Procrustes aligning procedure described in Sangalli et al. (2009), whereas, if no alignment is allowed, it implements a functional k-mean clustering of curves (see, e.g., Tarpey and Kinateder (2003)).

The described procedure merges the goal of alignment, i.e., decoupling phase and amplitude variability, with the goal of k-mean clustering, i.e., decoupling within and between-cluster amplitude variability. K-mean alignment is also able to disclose clustering structures in the phase, even though this is not one of the stated goals of the procedure. Overall, the technique has the capacity to point out features that can neither be captured by simple k-mean clustering without alignment nor by simple curve alignment without clustering.

*Main article 10.1214/14-EJS938.
†Corresponding author.

1.1. A problem-specific definition of phase and amplitude variabilities

Consider a set $\mathcal{F}$ of (possibly multidimensional) curves $f(s) : \mathbb{R} \to \mathbb{R}^d$. The choice of the considered space $\mathcal{F}$ depends on the data features and on the regulari ty required by the data analysis; typical examples are L^2 spaces or a subset of L^2 composed of curves in L^2 having derivatives up to a given order also in L^2. Continuously aligning $f_1 \in \mathcal{F}$ to $f_2 \in \mathcal{F}$ means finding a warping function $h(s)$: $\mathbb{R} \to \mathbb{R}$, of the abscissa parameter s, such that the two curves $f_1 \circ h$ and f_2 are the most similar or equivalently the least dissimilar (with $(f \circ h)(s) := f(h(s))$); see, e.g., Ramsay and Silverman (2005). It is thus necessary to specify a similarity index $\rho(\cdot,\cdot) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ that translates the concept of similarity between two functional data for the problem under study (or equivalently a dissimilarity index $\mathcal{E}(\cdot,\cdot)$: $\mathcal{F} \times \mathcal{F} \to \mathbb{R}$ that measures the dissimilarity between two functional data), and a class $\mathcal{H}$ of warping functions h (such that $f \circ h \in \mathcal{F}$, for all $f \in \mathcal{F}$ and $h \in \mathcal{H}$) that indicates which transformations for the abscissa are admissible for the considered problem. Aligning f_1 to f_2, according to $(\rho, \mathcal{H})$, means finding $h^* \in \mathcal{H}$ that maximizes $\rho(f_1 \circ h, f_2)$ (or equivalently that minimizes $\mathcal{E}(f_1 \circ h, f_2)$). The chosen index of similarity/dissimilarity and class of warping functions define univocally what are phase and amplitude variabilities for the problem being analyzed. The choice of the couple $(\rho, \mathcal{H})$, or equivalently $(\mathcal{E}, \mathcal{H})$, must hence be problem-specific.

The couple similarity/dissimilarity index and class of warping functions must satisfy the following properties, that we deem minimal requirements for coherence:

(a) The similarity index ρ is bounded from above, with maximum value equal to 1. Moreover, ρ is

reflexive: $\rho(f, f) = 1, \ \forall\ f \in \mathcal{F}$;

symmetric: $\rho(f_1, f_2) = \rho(f_2, f_1), \ \forall\ f_1, f_2 \in \mathcal{F}$;

transitive: $\left[\rho(f_1, f_2) = 1 \wedge \rho(f_2, f_3) = 1\right] \Rightarrow \rho(f_1, f_3) = 1$ $\forall\ f_1, f_2, f_3 \in \mathcal{F}$.

Equivalently, the dissimilarity index $\mathcal{E}$ is bounded from below, with minimal value equal to 0, and the properties above are suitably reformulated.

(b) The class of warping functions $\mathcal{H}$ is a convex vector space and has a group structure with respect to function composition $\circ$.

(c) The index ρ and the class $\mathcal{H}$ are coherent in the sense that, if two curves f_1 and f_2 are simultaneously warped by the same warping function $h \in \mathcal{H}$, their similarity does not change

$$\rho\left(f_1, f_2\right) = \rho\left(f_1 \circ h, f_2 \circ h\right), \quad \forall\ h \in \mathcal{H},$$

and the same property can of course be formulated for the dissimilarity $\mathcal{E}$. This guarantees that it is not possible to obtain a fictitious increment of the similarity between two curves f_1 and f_2 by simply warping them simultaneously to $f_1 \circ h$ and $f_2 \circ h$. This property is also referred as the

parallel-orbit property or *isometry of the action of* $\mathcal{H}$ (e.g., Wu and Srivastava, 2014).

Together, (b) and (c) imply the following property

(d) For all h_1 and $h_2 \in \mathcal{H}$,

$$\rho\left(f_1 \circ h_1, f_2 \circ h_2\right) = \rho\left(f_1 \circ h_1 \circ h_2^{-1}, f_2\right) = \rho\left(f_1, f_2 \circ h_2 \circ h_1^{-1}\right).$$

This means that a change in similarity between f_1 and f_2 obtained by warping simultaneously f_1 and f_2, can also be obtained by warping the sole f_1 or the sole f_2.

Many indices for measuring similarity (dissimilarity) between functions have been considered in the literature on functional data analysis (see, e.g., Ferraty and Vieu, 2006, for a proficient mathematical introduction to the issue). Table 1 reports some possible choices of couples (dissimilarity index, class of warping functions) that satisfy properties (a)–(d). Equivalent similarity indices can of course be considered. For instance, the normalized Pearson correlation $\frac{\langle f_1, f_2\rangle}{||f_1||^2||f_2||^2}$ is a similarity index naturally induced by the dissimilarity (semi-metric) $||\frac{f_1}{||f_1||} - \frac{f_2}{||f_2||}||$, where $\langle\cdot,\cdot\rangle$ denotes the inner product in L^2 and $||\cdot||$ the corresponding norm. A multivariate extension of the similarity $\frac{\langle f_1', f_2'\rangle}{||f_1'||^2||f_2'||^2}$, i.e., a normalized Pearson correlation of first derivatives, induced by the dissimilarity (semi-metric) $||\frac{f_1'}{||f_1'||} - \frac{f_2'}{||f_2'||}||$, is introduced in Section 2 to analyze the AneuRisk65 data set; see, eq. (1.3). The class of warping functions indicated in

TABLE 1

Examples of coherent couples (dissimilarity index, class of warping functions) satisfying property (c). $\bar{f}$ denotes the spatial mean of the curve f; f' denotes the first derivative of f; sign denotes the sign function, i.e., $sign(f(t)) = -1$ if $f(t) < 0$, $sign(f(t)) = 0$ if $f(t) = 0$, and $sign(f(t)) = 1$ if $f(t) > 0$

dissimilarity $\mathcal{E}$	class $\mathcal{H}$
$\|\|f_1 - f_2\|\|$	$\mathcal{H}_{shift}$
$\|\|f_1' - f_2'\|\|$	$\mathcal{H}_{shift}$
$\|\|(f_1 - \bar{f}_1) - (f_2 - \bar{f}_2)\|\|$	$\mathcal{H}_{shift}$
$\|\|(f_1' - \bar{f}_1') - (f_2' - \bar{f}_2')\|\|$	$\mathcal{H}_{shift}$
$\left\|\left\|\frac{f_1}{\|\|f_1\|\|} - \frac{f_2}{\|\|f_2\|\|}\right\|\right\|$	$\mathcal{H}_{affinity}$
$\left\|\left\|\frac{f_1'}{\|\|f_1'\|\|} - \frac{f_2'}{\|\|f_2'\|\|}\right\|\right\|$	$\mathcal{H}_{affinity}$
$\left\|\left\|\text{sign}(f_1')\sqrt{\|f_1'\|} - \text{sign}(f_2')\sqrt{\|f_2'\|}\right\|\right\|$	$\mathcal{H}_{diffeomorphism}$

the table are

$$\begin{aligned} \mathcal{H}_{\text{shift}} &= \{h : h(t) = t + q \text{ with } q \in \mathbb{R}\}, \\ \mathcal{H}_{\text{dilation}} &= \{h : h(t) = mt \text{ with } m \in \mathbb{R}^+\}, \\ \mathcal{H}_{\text{affine}} &= \{h : h(t) = mt + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\}, \end{aligned} \tag{1.1}$$

and the more general class of diffeomorphisms $\mathcal{H}_{diffeomorphism}$, composed by increasing functions that are smooth and have a smooth inverse. The case where no alignment is performed corresponds to the special case $\mathcal{H}_{\text{identity}} = \{h : h(t) = t\}$.

1.2. Aligning to k templates

Consider the problem of aligning and clustering a set of n curves $\{f_1, \ldots, f_n\} \subset \mathcal{F}$ with respect to a set of k template curves $\underline{\mu} = \{\mu_1, \ldots, \mu_k\} \subset \mathcal{F}$. For each template curve μ_j, define its domain of attraction

$$\Delta_j(\underline{\mu}) = \{f \in \mathcal{F} : \sup_{h \in \mathcal{H}} \rho(\mu_j, f \circ h) \geq \sup_{h \in \mathcal{H}} \rho(\mu_r, f \circ h), \forall\, r \neq j\}, \quad j = 1, \ldots, k,$$

and the labeling function $\lambda(\underline{\mu}, f) = \min\{r : f \in \Delta_r(\underline{\mu})\}$. Thus $\mu_{\lambda(\underline{\mu}, f)}$ indicates the template whose f can be best aligned to and hence $\lambda(\underline{\mu}, f)$ indicates the cluster that f should be assigned to.

If the k templates $\underline{\mu} = \{\mu_1, \ldots, \mu_k\}$ were known, then aligning and clustering the set of n curves $\{f_1, \ldots, f_n\}$ with respect to $\underline{\mu}$ would simply mean assigning f_i to the cluster $\lambda(\underline{\mu}, f_i)$ and aligning it to the corresponding template $\mu_{\lambda(\underline{\mu}, f_i)}$, for $i = 1, \ldots, n$. Here we are interested in the more complex case when the k templates are unknown. Ideally, if our aim is aligning and clustering the set of n curves $\{f_1, \ldots, f_n\}$ with respect to k unknown templates, we should first solve the following optimization problem

(i) find $\underline{\mu} = \{\mu_1, \ldots, \mu_k\} \subset \mathcal{F}$ and $\underline{\mathbf{h}} = \{h_1, \ldots, h_n\} \subset \mathcal{H}$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\mu_{\lambda(\underline{\mu}, f_i)}, f_i \circ h_i) \geq \frac{1}{n} \sum_{i=1}^{n} \rho(\psi_{\lambda(\underline{\psi}, f_i)}, f_i \circ g_i),$$

for any other set of k templates $\underline{\psi} = \{\psi_1, \ldots, \psi_k\} \subset \mathcal{F}$ and any other set of n warping functions $\underline{\mathbf{g}} = \{g_1, \ldots, g_n\} \subset \mathcal{H}$,

and then, for $i = 1, \ldots, n$,

(ii) assign f_i to the cluster $\lambda(\underline{\mu}, f_i)$ and warp f_i along h_i.

The optimization problem (i) describes a search both for the set of optimal k templates, and for the set of optimal n warping functions. Note that the solution $(\underline{\mu}, \underline{\mathbf{h}})$ to (i) has mean similarity $\frac{1}{n} \sum_{i=1}^{n} \rho(\mu_{\lambda(\underline{\mu}, f_i)}, f_i \circ h_i)$ equal to 1 if and only if it is possible to perfectly align and cluster in k groups the set of n curves, i.e., if and only if there exists $\underline{\mathbf{h}} = \{h_1, \ldots, h_n\} \subset \mathcal{H}$ and a partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ of $\{1, \ldots, n\}$ in k elements, such that $\rho(f_i \circ h_i, f_j \circ h_j) = 1$ for all i and j belonging to the same element of $\mathcal{P}$.

It should also be noted that, thanks to property (c), if $\{\mu_1, \ldots, \mu_k\}$ and $\{h_1, \ldots, h_n\}$ provide a solution to (i), then also $\{\mu_1 \circ g_1, \ldots, \mu_k \circ g_k\}$ and $\{h_1 \circ g_{\lambda(\underline{\mu}, f_1)}, \ldots, h_n \circ g_{\lambda(\underline{\mu}, f_n)}\}$ are a solution to (i), for any $\{g_1, \ldots, g_k\} \subset \mathcal{H}$. Moreover, this solution identifies the same clusters (i.e., is associated to the same partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ of $\{1, \ldots, n\}$).

The non-linear optimization problem (i) is not analytically solvable in its complete generality. For this reason, in Sangalli et al. (2010) we proposed to simultaneously deal with (i) and (ii) via a k-mean alignment algorithm that iteratively alternates *template identification* steps, *assignment and alignment* steps and *normalization* steps. In the *template identification* step, we estimate the set of k templates associated to the k clusters identified at the previous *assignment and alignment* step. The jth template can be identified as the curve μ_j, in some set of curves $\mathcal{C}$, that maximizes the total similarity within the jth cluster:

$$argmax_{\mu_j \in \mathcal{C}} \sum_{i:\ f_i \ \in\ j\text{th cluster}} \rho(\mu_j, f_i). \tag{1.2}$$

Two choices for the set of curves $\mathcal{C}$ are particularly natural: $\mathcal{C}$ may coincide with the entire considered functional space $\mathcal{F}$, in which case the templates are the within cluster Frechet templates, or $\mathcal{C}$ may coincide with the sample of curves $\{f_1, \ldots, f_n\}$, in which case the templates coincide with the within cluster medoids, or Karcher templates. In the *assignment and alignment* step, we align the n curves to the set of the k templates obtained in the previous template identification step, as described above. The k-mean alignment algorithm also considers the problem of non-uniqueness of the solution, by targeting a specific solution via a *normalization step*. The algorithm is initialized with a set of initial templates, and stopped when, in the assignment and alignment step, the increments of the similarity indices are all lower than a fixed threshold.

In many practical situations, as for instance in the case of the AneuRisk65 dataset, the functional data are not available on the entire real axis but observed on arbitrary intervals and thus both the template identification step and the assignment and alignment step have to be carried out in an approximated way. In this case, the similarities between two functions are computed over the intersection of the domains of the two functions; moreover, if the template identification is carried out on the entire space $\mathcal{F}$, the Frechet mean (no longer analytically available) is approximated by a local mean.

Details for the practical implementation of a k-mean alignment procedure are given in Sangalli et al. (2010). The procedure is coded in the `fdakma` R package downloadable from CRAN Parodi et al. (see 2014).

1.3. Shape invariant models

When analyzing the AneuRisk65 data, it makes sense to consider two vessel centrelines to be perfectly aligned if they are identical except for a shift and/or a dilation along the three space coordinates. Because location of the scanned volume and proportions of the skull change across patients (see the data presentation),

different shift and/or dilation for each space coordinate must be permitted. For this reason, Sangalli et al. (2009, 2010), used the following bounded similarity index between two curves $f_1, f_2 \in \mathcal{F}$, where $\mathcal{F} = \{f : f \in L^2(\mathbb{R};\mathbb{R}^d), f' \in L^2(\mathbb{R};\mathbb{R}^d), f' \neq \mathbf{0}\}$:

$$\rho(f_1, f_2) = \frac{1}{d}\sum_{p=1}^{d} \frac{\int_{\mathbb{R}} c'_{1p}(t)c'_{2p}(t)dt}{\sqrt{\int_{\mathbb{R}} c'_{1p}(t)^2 dt}\sqrt{\int_{\mathbb{R}} c'_{2p}(t)^2 dt}}, \tag{1.3}$$

with c_{ip} indicating the pth component of f_i, $f_i = \{c_{i1}, \ldots, c_{id}\}$. Geometrically, (1.3) represents the average of the cosines of the angles between the derivatives of homologous components of f_1 and f_2, which is a possible multidimensional extension of the univariate similarity index reported at the sixth row of Table 1. The index 1.3 assumes its maximal value 1 when the two curves are identical except for shifts and dilations of homologous components, i.e.,

$$\rho(f_1, f_2) = 1 \quad \Leftrightarrow \quad \begin{array}{l} \text{for } p = 1, \ldots, d,\ \exists a_p \in \mathbb{R}^+, b_p \in \mathbb{R}: \\ c_{1p}(t) = a_p c_{2p}(t) + b_p. \end{array} \tag{1.4}$$

This similarity index is coherent, in the sense of property (c), with the classes of warping functions of the abscissa given in eq. (1.1). Here, in particular we shall consider the largest of these classes,

$$\mathcal{H}_{\text{affine}} = \{h : h(t) = mt + q \ \text{ with } \ m \in \mathbb{R}^+, q \in \mathbb{R}\} \tag{1.5}$$

i.e., the group of strictly increasing affine transformations.

For $f = \{c_1, \ldots, c_d\} \in \mathcal{F}$, where c_p indicates the pth component of f, for $p = 1, \ldots, d$, assume the existence of $\mu = \{\mu_1, \ldots, \mu_d\} \in \mathcal{F}$ and of a parameter vector $(a_1, \ldots, a_d, b_1, \ldots, b_d, m, q)$, with $a_p \in \mathbb{R}^+$ and $b_p \in \mathbb{R}$ for $p = 1, \ldots, d$, $m \in \mathbb{R}^+$, $q \in \mathbb{R}$, such that

$$c_p(t) = a_p \mu_p(mt + q) + b_p \qquad \text{for } p = 1, \ldots, d. \tag{1.6}$$

We shall write $f \in \text{SIM}(\mu)$, since the condition (1.6) means that f falls within a *shape invariant model* (SIM), with *characteristic shape curve* μ. For $d = 1$, SIM models were introduced by Lawton, Sylvestre and Maggio (1972). For further details, see Kneip and Gasser (1988). SIM models are strongly connected with the couple $(\rho, \mathcal{H})$ defined in (1.3) and (1.5). Indeed,

$$\exists h \in \mathcal{H} : \rho(f \circ h, \mu) = 1 \quad \Leftrightarrow \quad f \in \text{SIM}(\mu). \tag{1.7}$$

Note that, thanks to property (d), the roles of f and μ can be swapped. Now, consider a set of n curves $\{f_1, \ldots, f_n\} \subset \mathcal{F}$, such that $f_i \in \text{SIM}(\mu)$ for all $i = 1, \ldots, n$; then, the following property follows immediately:

(f) For all f_i, f_j, with $i, j = 1, \ldots, n$, there exist $h_i \in \mathcal{H}$, $h_j \in \mathcal{H}$ such that

$$\rho\left(f_i \circ h_i, f_j \circ h_j\right) = \rho\left(f_i \circ h_i, \mu\right) = \rho\left(f_j \circ h_j, \mu\right) = 1 \quad \forall\ i, j = 1, \ldots, n.$$

Because of (1.7), when using the couple $(\rho, \mathcal{H})$ defined in (1.3) and (1.5), it is possible to perfectly align and cluster, in k groups, a set of n curves if there

exist k characteristic shape curves, $\mu_1, \ldots, \mu_k$, such that

$$\forall\ i = 1, \ldots, n, \quad \exists\ l_i \in \{1, \ldots, k\} :\ \ f_i \in \mathrm{SIM}(\mu_{l_i}).$$

In this case the optimization problem (i) is solved by setting $\mu_{\lambda(\underline{\mu}, f_i)} \equiv \mu_{l_i}$, and its objective function achieves the maximum total similarity 1.

1.4. Theoretical framework

The introduction, in a functional data analysis, of a similarity (derived by a metric $\mathcal{E}$ satisfying some specific properties) and of a group $\mathcal{H}$ of warping functions, with respect to which the similarity is invariant (i.e., property (d) in Section 1.1), provides a mathematical framework where a sound and not ambiguous definition of phase and amplitude variability can be given. Indeed in this framework, we can prove that the analysis of a continuously-registered functional data set can be re-interpreted as the analysis of a set of suitable equivalence classes associated to unaligned functions and induced by the group of the warping functions. The theoretical investigation for a coherent formalization of the problem of registration, in relation to properties required to the metric $\mathcal{E}$ and to the group $\mathcal{H}$ of warping functions, is deepened and detailed in Vantini (2012).

The most important required property is the $\mathcal{H}$-*invariance* of the metric $\mathcal{E}$ (Sangalli et al. (2009, 2010)) which indeed induces property (d) of Section 1.1. $\mathcal{H}$-invariance provides the quotient set made by the orbits induced by the action of $\mathcal{H}$ over $\mathcal{F}$ with a natural metric δ (dependent on the joint choice of $\mathcal{E}$ and $\mathcal{H}$) defined as follows: $\delta([f_i], [f_j]) := \min_{h_i, h_j \in \mathcal{H}} \mathcal{E}(f_i \circ h_i, f_j \circ h_j)$.

The introduction of a quotient set provided with a natural metric jointly induced by the original metric $\mathcal{E}$ and by the group $\mathcal{H}$ enables a not ambiguous definition of *Phase Variability* and *Amplitude Variability*. Phase variability is defined as that occuring between functions belonging to the same equivalence class, i.e. the variability within equivalence classes, and the amplitude variability is the one between functions not belonging to the same equivalence class and not imputable to phase variability, i.e. the variability between equivalence classes. Moreover, within this mathematical framework the k-mean alignment algorithm is an implementation of a k-mean clustering algorithm on the n equivalence classes $\{[f_1],[f_2], \ldots, [f_n]\}$ induced by the action of $\mathcal{H}$ over the functional data set $\{ f_1, f_2, \ldots, f_n\}$ performed by using the naturally defined metric δ.

2. *k*-mean alignment of the AneuRisk65 data

2.1. Decoupling phase and amplitude variability in the AneuRisk65 data by 1-mean alignment

To enable meaningful comparisons across subjects in the AneuRisk study, it is necessary to first efficiently decouple the phase and the amplitude variability. As described in the data presentation, in this application phase variation is mainly due to differences in the dimension of skulls among subjects, whereas the amplitude variation is mainly due to differences in the carotid morphological

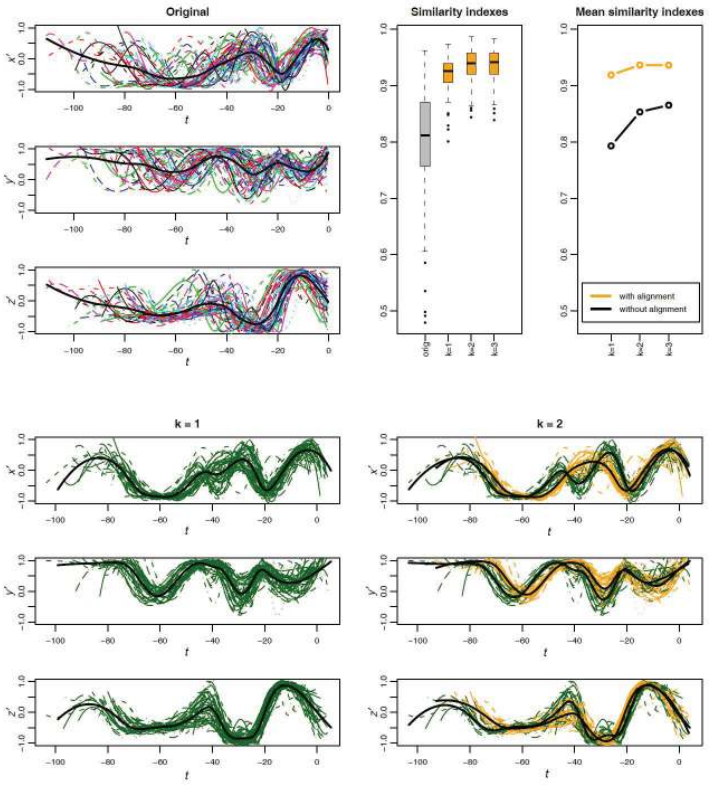

FIG 1. *Top-left: First derivatives of the original ICA centrelines. The thick black lines indicate the first derivatives of the estimated template. Top-center: Boxplots of the similarity indices for unaligned curves and the k-mean aligned curves, for* $k = 1, 2, 3$. *Top-right: Means of the similarity indices obtained by k-mean alignment (orange) and by k-mean without alignment (black), for* $k = 1, 2, 3$. *Bottom: First derivatives of the 1-mean (left) and 2-mean (right) aligned ICA centrelines; the thick black lines indicate the first derivatives of the estimated templates.*

shapes. We shall in particular explore if the morphological features of these vessels relates with aneurysms presence and location, contrasting subjects in the Upper group and subjects in the Lower and No-aneurysm groups, often joined in a unique "Lower-No" group, as indicated in the data description.

First derivatives of the vessel centrelines have been aligned by the 1-mean alignment, using the similarity index (1.3) and the class of warping functions (1.5). The variability captured by the optimal warping functions found during this alignment process was analyzed in Sangalli et al. (2009) and was not found to be associated to the aneurysmal pathology. In particular, no significant difference exists between the warping functions of subjects in the Upper and Lower-No groups. Subsequent analysis may hence focus on the aligned data. The optimal warping functions can be used to correspondingly align the radius and curvature profiles; see Figure 3. After alignment it is possible to start appre-

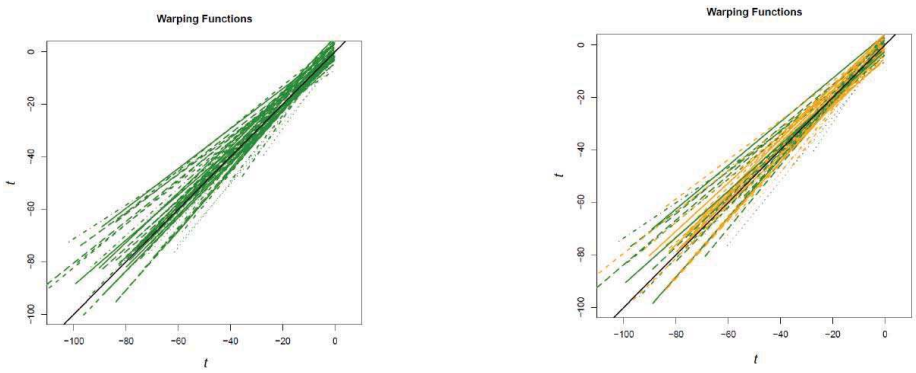

FIG 2. *Warping functions corresponding to 1-mean (left) and 2-mean (right) alignment of ICA centerlines. First derivatives of aligned curves are displayed in the bottom panel of Figure 1. The identity function is plotted in black.*

ciating a common pattern for the curvature profiles (bottom right panel) that was not visible before alignment (bottom left panel). The registered radius and curvature profiles highlight many interesting aspects. Figure 3 shows that the vessel gets progressively narrower toward the terminal bifurcation of the ICA; this is the so-called tapering effect. Tapering concerns all arteries, but it is particularly apparent close to the terminal bifurcation of the ICA, where the artery has to enter the skull. The figure also shows that most ICAs display two peaks of curvature in the terminal part of the vessel; these peaks of curvature are in correspondence with the carotid syphon. The same figure displays also Gaussian kernel density estimates of the aneurysm location along the ICA, before

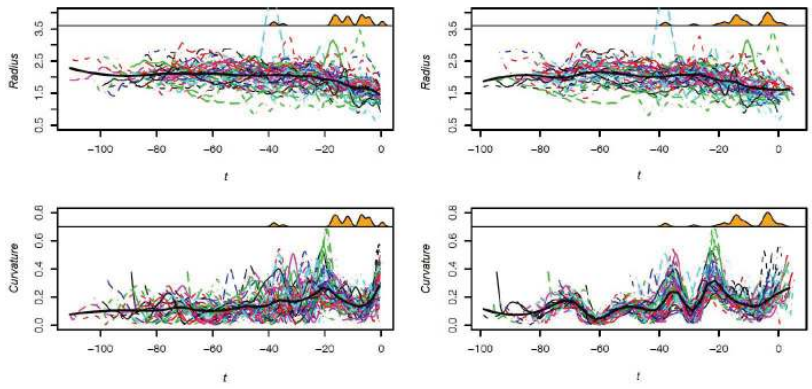

FIG 3. *Radius (top) and curvature (bottom) profiles of the 65 patients respectively before (left) and after (right) alignment using the optimal warping functions displayed in the left panel of Figure 2. Solid black lines show mean curves. The upper part of each picture also displays the estimated probability density function of the location of aneurysms along the ICA.*

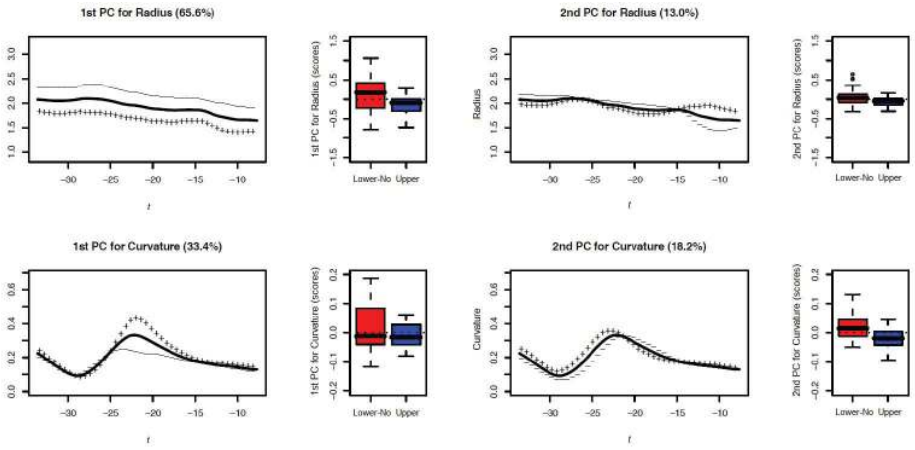

FIG 4. *Estimates of the first (left) and the second (right) eigenfunctions for radius (top) and curvature (bottom), and boxplots of the corresponding scores for the two groups (red for the Lower-No group and blue for the Upper group).*

and after alignment (left and right, respectively), based on data from patients having an aneurysm along the ICA or at its terminal bifurcation. The majority of the ICA aneurysms are located in the terminal part of the vessel, where tapering is stronger, and after the main curvature peak. These results support the conjecture concerning the influence of the vessel morphology and the aneurysm onset, via the hemodynamics. In fact, the tapering of the vessel and the peak in its curvature determine hemodynamic regimes that may facilitate aneurysm formation and development. The density estimates of the aneurysm location show that, after alignment, the locations of the ICA aneurysms cluster in two neatly separated groups, before and after -13 mm from the vessel terminal bifurcation. This fact suggests that this is the average position of the dural ring, i.e., the hole in the skull bone the ICA goes through to enter inside the skull. Notice that this ring cannot be detected directly through angiographies, but indications of the location of the aneurysm relative to the dural ring may be of great importance, since aneurysms within the skull are more dangerous, as explained in the data description.

The autocovariance structures of aligned radius and curvature profiles are thus explored separately by means of Functional Principal Component Aanalysis (FPCA), in order to estimate their main uncorrelated modes of variability. Since the 65 curves are observed on different abscissa intervals, these analyses focus on the interval where all curves are available, i.e., for values of the abscissa between -3.37 cm and -0.78 cm. Figure 4 shows the first two eigenfunctions of the autocovariance function for radius and for curvature, respectively, and the percentage of total variance explained is printed over each graph. Figure 4 also reports, for each considered principal component, the distributions of the corresponding scores for subjects in the Upper group and subjects in the Lower-No

group. These scores may be used to discriminate the two groups of patients. In fact, the distribution of FPCA scores have significantly different means and/or variances for the two groups, as confirmed by appropriate t-tests and F-tests for equality of means and variances. According to these differences, Upper group patients tend to have wider, more tapered and less curved ICA's compared to patients belonging to the Lower-No group. Moreover the variance of these geometrical features is significantly smaller in the Upper group than in the Lower-No group. The Upper group is indeed very well characterized in terms of the geometrical features represented by the first two principal components of aligned radius and curvature profiles. Sangalli et al. (2009) shows in fact that a quadratic discriminant analysis of the scores of the first two principal components of aligned radius and curvature profiles correctly identifies 31 out of the 33 patients in the Upper group. This gives strong statistical evidence in favor of the conjecture explored within the project.

2.2. Identifying ICA's with different morphological shapes by k-mean alignment

The problem at hand might suggest the presence of more than one prototype of morphological shape of ICA. To evaluate this possibility we use k-mean alignment.

The top-center and top-right panels of Figure 1 give an indication of how many clusters k should be considered, i.e., how many morphological shapes of ICA are present in the datasest. The first panel reports the boxplot of the similarity indices between the unaligned centrelines and their estimated mean curve ("unaligned"), and the boxplots of the similarity indices between the k-mean aligned centrelines and the associated k estimated templates, for $k = 1, 2, 3$. The second panel displays the corresponding means of the similarity indices (orange circles linked by orange lines). The same plot also shows the means of the similarity indices that would be obtained by the simple k-mean algorithm without alignment (black circles linked by black lines). As highlighted by this figure, 1-mean alignment leads to a large increase in the similarities, with respect to the similarities of the unaligned centrelines, but a further significant gain can be obtained by setting $k=2$ in the aligning and clustering procedure; instead, a choice of $k=3$ would not be justified by any additional increase in the similarities. Thus the k-mean alignment algorithm suggests the presence of $k = 2$ amplitude clusters within the analyzed centrelines. The bottom-right panels of Figure 1 show the results of 2-mean alignment of these three-dimensional curves. Figure 5 gives a three-dimensional visualization of the estimated templates of the two amplitude clusters. These two templates indeed identify two prototypal shapes of ICA commonly used in the medical literature, see e.g. Krayenbuehl, Huber and Yasargil (1982), and the two clusters can be described as the Ω-shaped ICA's cluster (35 curves in orange), whose siphons feature just one bend, and the S-shaped ICA's cluster (30 curves in green), whose siphons are characterized by two bends. As displayed in the right panel of Figure 6, the simple k-mean clus-

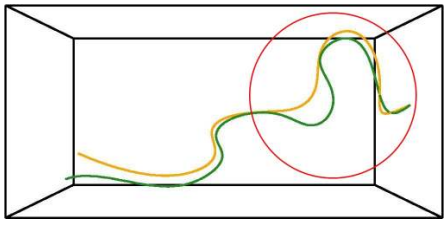

FIG 5. *3D image of the estimated templates of the 2 amplitude clusters, found by 2-mean alignment of the ICA centrelines. The template of the orange cluster is the prototype of an Ω-shaped ICA (single-bend siphon), whereas the one of the green cluster is the prototype of an S-shaped ICA (double-bend siphon).*

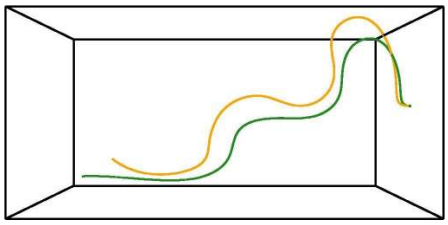

FIG 6. *3D image of the estimated templates of the 2 clusters found by simple 2-mean clustering without alignment of the ICA centrelines. The two templates appear to have almost the same morphological shape, and seem to differ mainly in their phase. Clustering, without joint alignment, is driven by the predominant phase variability.*

tering of these data, without joint alignment, leads instead to uninteresting classification results, failing to identify ICA's with different morphological shapes.

It is very interesting to note that the clustering found by 2-mean alignment seems indeed to be relevant for the aneurysmal pathology, in the sense that there is statistical evidence of a dependence between cluster membership and aneurysm presence and location. Looking at how subjects in the Upper, Lower and No-aneurysm group have been allocated respectively to the Ω-shaped ICA's and S-shaped ICA's clusters, we obtain the following conditional contingency table.

	Upper group (33)	Lower group (25)	No-aneurysm (7)
Ω	70%	48%	0%
S	30%	52%	100%

Note that the 7 subjects in the No-aneurysm group all display S-shaped ICA's and that among the 33 patients in the Upper group (those having the most dangerous aneurysms) only a minority has an S-shaped ICA, whilst the majority (70%) has an Ω-shaped one. This prompted us to conjecture that the ICA syphons act as flow energy dissipators. While S-shaped (double bend syphon) ICAs are expected to be very effective in dissipating the flow energy, Ω-shaped (single-bend syphon) ICA's would not be as efficient. A higher flow energy downstream of Ω-shaped ICA's would result in an overloaded mechanical stress for downstream arteries, creating more stimuli for aneurysm onset and development. More results in support of this conjecture are reported in Passerini et al. (2012), where the relationship between morphological and hemodynamical features, and their impact on aneurysm pathology, is further explored.

3. Discussion

In this work, after the decoupling of phase and amplitude variabilities by k-mean alignment, we focussed on amplitude variability and on clustering in the am-

plitude. This is because, in the specific application here considered, the phase variability, which is mostly due to the different dimensions of patients skulls, seems not to be relevant in the study, as it does not influences the considered pathology. In other applications, though, the clustering might be in the phase, rather than in the amplitude, or even in both phase and amplitude. It is thus important to mention that the alignment and clustering procedure here described, beside correctly detecting true amplitude clusters, is also able to simultaneously disclose clustering structures present in the phase, as illustrated for instance in Sangalli et al. (2010) and in Bernardi et al. (2014a,b); Patriarca et al. (2014), via both simulations and applications to real data.

Acknowledgements

We thank Alessandro Veneziani, P.I. of the AneuRisk Project. All authors are grateful to the MBI Mathematical Biosciences Institute http://mbi.osu.edu/, The Ohio State University, for support. L. M. Sangalli acknowledges funding by the research program Dote Ricercatore Politecnico di Milano – Regione Lombardia, project "Functional data analysis for life sciences", and by MIUR Ministero dell'Istruzione dell'Università e della Ricerca, *FIRB Futuro in Ricerca* starting grant project "Advanced statistical and numerical methods for the analysis of high dimensional functional data in life sciences and engineering" http://mox.polimi.it/users/sangalli/firbSNAPLE.html.

References

Bernardi, M., Sangalli, L. M., Secchi, P. and Vantini, S. (2014a). Analysis of proteomics data: Block k-mean alignment. *Electronic Journal of Statistics* **8** 1714–1723, Special Section on Statistics of Time Warpings and Phase Variations.

Bernardi, M., Sangalli, L. M., Secchi, P. and Vantini, S. (2014b). Analysis of juggling data: An application of k-mean alignment. *Electronic Journal of Statistics* **8** 1817–1824, Special Section on Statistics of Time Warpings and Phase Variations.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer. MR2229687

Kneip, A. and Gasser, T. (1988). Convergence and consistency results for self-modeling nonlinear regression. *The Annals of Statistics* **16** 82–112. MR0924858

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

Lawton, W. H., Sylvestre, E. A. and Maggio, M. S. (1972). Self modeling nonlinear regression. *Technometrics* **14** 513–532.

Parodi, A., Patriarca, M., Sangalli, L., Secchi, P., Vantini, S. and Vitelli, V. (2014). fdakma: Clustering and alignment of a functional dataset, R package version 1.1.

Passerini, T., Sangalli, L. M., Vantini, S., Piccinelli, M., Bacigaluppi, S., Antiga, L., Boccardi, E., Secchi, P. and Veneziani, A. (2012). An integrated CFD-statistical investigation of parent vasculature of cerebral aneurysms. *Cardio. Eng. and Tech.* **3** 26–40.

Patriarca, M., Sangalli, L. M., Secchi, P. and Vantini, S. (2014). Analysis of spike train data: An application of k-mean alignment. *Electronic Journal of Statistics* **8** 1769–1775, Special Section on Statistics of Time Warpings and Phase Variations.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*, second ed. *Springer Series in Statistics*. Springer, New York. MR2168993 MR2168993

Sangalli, L. M., Secchi, P. and Vantini, S. (2013). AneuRisk65: A dataset of three-dimensional cerebral vascular geometries. *Electronic Journal of Statistics* **8** 1879–1890, Special Section on Statistics of Time Warpings and Phase Variations.

Sangalli, L. M., Secchi, P., Vantini, S. and Veneziani, A. (2009). A case study in exploratory functional data analysis: Geometrical Features of the Internal Carotid Artery. *J. Amer. Statist. Assoc.* **104** 37–48. MR2663032

Sangalli, L. M., Secchi, P., Vantini, S. and Vitelli, V. (2010). K-mean alignment for curve clustering. *Computational Statistics and Data Analysis* **54** 1219–1233. MR2600827

Tarpey, T. and Kinateder, K. K. J. (2003). Clustering functional data. *Journal of Classification* **20** 93–114. MR1983123

Vantini, S. (2012). On the definition of phase and amplitude variability in functional data analysis. *TEST* **21** 676–696. MR2992088

Wu, W. and Srivastava, A. (2014). Analysis of spike train data: Alignment and comparisons using extended Fisher-Rao metric. *Electronic Journal of Statistics* **8** 1776–1785, Special Section on Statistics of Time Warpings and Phase Variations.