Functional clustering and alignment methods with applications

Laura M. Sangalli\textsuperscript{1}, Piercesare Secchi\textsuperscript{1}, Simone Vantini\textsuperscript{1}, Valeria Vitelli\textsuperscript{1}

\textsuperscript{1} MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy.
laura.sangalli@polimi.it
piercesare.secchi@polimi.it
simone.vantini@polimi.it
valeria.vitelli@mail.polimi.it

Communicated by Giorgio Fotia

Abstract

We consider the issue of classification of functional data and, in particular, we deal with the problem of curve clustering when curves are misaligned. In the proposed setting, we aim at jointly aligning and clustering the curves, via the solution of an optimization problem. We describe an iterative procedure for the solution of the optimization problem, and we detail two alternative specifications of the procedure, a \textit{k}-mean version and a \textit{k}-medoid version. We illustrate via applications to real data the robustness of the alignment and clustering procedure under the different specifications.

\textbf{Keywords:} Functional data analysis, curve alignment, curve clustering, \textit{k}-mean algorithm, \textit{k}-medoid algorithm.

\textbf{AMS Subject Classification:} 62H30, 68T10

1. Introduction

Unsupervised classification (or clustering) methods are an important topic in statistics, with many applications in various fields; the aim of such techniques is to classify a sample of data into homogeneous groups, without having any a priori knowledge about the true underlying clustering structure. In particular, \textit{k}-mean clustering for multivariate data is a well-established algorithm, first proposed in the ‘70 by Hartigan \cite{Hartigan} and widely used in statistical classification problems for its flexibility. The algorithm divides a set of \textit{n} data in \textit{k} subgroups, iteratively alternating the search for the templates of the \textit{k} clusters, that minimize within-cluster variability, and the assignment of each of the \textit{n} data to the cluster whose template is the nearest. Multivariate \textit{k}-medoid is a less computationally intensive modification of the \textit{k}-mean algorithm \cite{Kaufman}. The main difference involves the way each
cluster template is estimated, with the \( k \) templates simply selected among the \( n \) data.

Here, we shall consider the problem of clustering of functional data, and in particular of curves. For an introduction to the statistical analysis of functional data, see the books by Ramsay and Silverman [3] and by Ferraty and Vieu [4]. Many methods for curve clustering have been proposed in the literature on functional data analysis, most of which are based on generalizations of the multivariate \( k \)-mean algorithm. For example, Tokushige et al. [5] proposes a generalization of \( k \)-mean clustering algorithms for functional data, as a way to solve the problem of principal points estimation. In Cuesta-Albertos and Fraiman [6], a robust \( k \)-mean clustering procedure is developed, based on the idea of “impartial trimming”, which proves to be useful for high dimensional data. Another \( k \)-mean algorithm for functional data can be found in Chiou and Li [7], where the efficiency of the clustering procedure is improved thanks to the use of a non-parametric random-effect model.

When dealing with clustering of curves, we need 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. On the other hand, the evident misalignment of the data acts as a confounding factor when trying to cluster the data, and the above cited \( k \)-mean algorithms fail to give efficient results (see [8]). This highlights the need for a clustering procedure which is able to jointly deal with data alignment, decoupling the variability due to data misalignment (phase variability) and the variability due to the shape (amplitude variability).

The problem of curve alignment (or curve registration) has been considered by a number of authors. Altman and Villarreal [9] face this problem using self-modelling non-linear regression methods, Lindstrom and Bates [10] develop non-linear mixed-effects models, and Ke and Wang [11] merge the above approaches in the unifying framework of semiparametric non-linear mixed-effects models. A different line of research, advocated by J. O. Ramsay, is followed by Kaziska and Srivastava [12] and Sangalli et al. [13], who define suitable similarity indexes between curves and thus align the curves, maximizing their similarities by means of a Procrustes procedure.

Following the latter line of research, in Sangalli et al. [8] 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. This procedure was thus specified in a $k$-mean algorithm. Here, we describe an alternative specification of the procedure, in a $k$-medoid algorithm version. This new version approximates more directly the original optimization problem, and is potentially less sensitive to the presence of anomalous data. Moreover, this alternative version gives us the possibility of testing the robustness of our alignment and clustering procedure. In this work, we shall call robust a procedure whose final results are not influenced by different specifications in the algorithm (in particular, different approximations of the optimization problem which identifies cluster ideal templates). See Boudaoud et al. [14] and Liu and Yang [15] for other recent approaches to the problem of clustering of misaligned functional data.

The paper is organized as follows. In Section 2 we introduce a proper framework for the problem of clustering and alignment, defining phase and amplitude variability. In Section 3 we state the problem of curve clustering and alignment as an optimization problem, and we propose an iterative procedure for its solution. In Section 4 we describe two approaches to the template identification step in the procedure proposed in Section 3, obtaining a $k$-mean and a $k$-medoid specification of the iterative procedure. The two subsequent sections are devoted to the application of the two algo-
rithm versions to real data; in particular, Section 5 describes the clustering and alignment analysis of the data shown in Figure 1, concerning three-dimensional vascular geometries, while in Section 6 the procedure is tested on a benchmark dataset, the Berkeley Growth Study dataset. Some concluding considerations are drawn in Section 7. All analyses of real datasets are performed in R [16].

2. Defining phase and amplitude variabilities

The variability among two or more curves can be thought of as having two components: phase variability and amplitude variability. Heuristically, phase variability is the one that can be eliminated by suitably aligning the curves, and amplitude variability is the one that remains among the curves once they have been aligned. Consider a set $C$ of curves $c(s) : \mathbb{R} \to \mathbb{R}^d$. Aligning $c_1 \in C$ to $c_2 \in C$ means finding a warping function $h(s) : \mathbb{R} \to \mathbb{R}$, of the abscissa parameter $s$, such that the two curves $c_1 \circ h$ and $c_2$ are the most similar (with $(c \circ h)(s) := c(h(s))$). It is thus necessary to specify a similarity index $\rho(\cdot, \cdot) : C \times C \to \mathbb{R}$ that measures the similarity between two curves, and a class $W$ of warping functions $h$ (such that $c \circ h \in C$, for all $c \in C$ and $h \in W$) indicating the allowed transformations for the abscissa. Aligning $c_1$ to $c_2$, according to $(\rho, W)$, means finding $h^* \in W$ that maximizes $\rho(c_1 \circ h, c_2)$. This procedure decouples phase and amplitude variability without loss of information: phase variability is captured by the optimal warping function $h^*$, whilst amplitude variability is the remaining variability between $c_1 \circ h^*$ and $c_2$. Note that the choice of the couple $(\rho, W)$ defines what is meant by phase variability and amplitude variability.

Many similarity indexes for measuring similarity between functions have been considered in the literature on functional data analysis; for a proficient mathematical introduction to the issue see the book by Ferraty and Vieu [4]. Sangalli et al. [8,13] proposed the following bounded similarity index between two curves $c_1, c_2 \in C$, where $C = \{c : c \in L^2(\mathbb{R}; \mathbb{R}^d), c' \in L^2(\mathbb{R}; \mathbb{R}^d), c' \neq 0\}$

$$\rho(c_1, c_2) = \frac{1}{d} \sum_{p=1}^{d} \frac{\int_{\mathbb{R}} c_{1p}'(s)c_{2p}'(s)ds}{\sqrt{\int_{\mathbb{R}} c_{1p}'(s)^2ds} \sqrt{\int_{\mathbb{R}} c_{2p}'(s)^2ds}},$$

with $c_{ip}$ indicating the $p$th component of $c_i$, $c_i = \{c_{i1}, \ldots, c_{id}\}$; geometrically, (1) represents the average of the cosines of the angles between the derivatives of homologous components of $c_1$ and $c_2$. The two curves are said to be similar when the index assumes its maximal value 1; for the similarity index defined in (1), this happens when the two curves are identical except
for shifts and dilations of their components

\[(2) \quad \rho(c_1, c_2) = 1 \iff \text{for } p = 1, \ldots, d, \exists \theta_{0p} \in \mathbb{R}, \theta_{1p} \in \mathbb{R}^+: c_{1p}(s) = \theta_{0p} + \theta_{1p}c_{2p}(s).\]

The choice of this similarity index comes along with the following choice for the class \(W\) of warping functions of the abscissa

\[(3) \quad W = \{h : h(s) = ms + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\}\]

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

The couple \((\rho, W)\) defined in (1) and (3) satisfies the following properties (a)-(c) that we deem to be minimal requirements for coherence:

(a) The similarity index \(\rho\) is bounded, with maximum value equal to 1. Moreover, \(\rho\) is

- reflexive: \(\rho(c, c) = 1, \forall c \in C;\)
- symmetric: \(\rho(c_1, c_2) = \rho(c_2, c_1), \forall c_1, c_2 \in C;\)
- transitive: \(\left[\rho(c_1, c_2) = 1 \land \rho(c_2, c_3) = 1\right] \Rightarrow \rho(c_1, c_3) = 1\)
\[\forall c_1, c_2, c_3 \in C.\]

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

(c) The index \(\rho\) and the class \(W\) are consistent in the sense that, if two curves \(c_1\) and \(c_2\) are simultaneously warped along the same warping function \(h \in W\), their similarity does not change

\[(4) \quad \rho\left(c_1, c_2\right) = \rho\left(c_1 \circ h, c_2 \circ h\right), \quad \forall h \in W.\]

This guarantees that it is not possible to obtain a fictitious increment of the similarity between two curves \(c_1\) and \(c_2\) by simply warping them simultaneously to \(c_1 \circ h\) and \(c_2 \circ h\).

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

(d) For all \(h_1\) and \(h_2\) \(\in W,\)

\[\rho\left(c_1 \circ h_1, c_2 \circ h_2\right) = \rho\left(c_1 \circ h_1 \circ h_2^{-1}, c_2\right) = \rho\left(c_1, c_2 \circ h_2 \circ h_1^{-1}\right).\]

This means that a change in similarity between \(c_1\) and \(c_2\) obtained by warping simultaneously \(c_1\) and \(c_2\), can also be obtained by warping the sole \(c_1\) or the sole \(c_2\). Moreover, the couple \((\rho, W)\) defined in (1) and (3) satisfies the additional auxiliary property

209
(e) Let \( W^d \) be the set of all transformations \( r : \mathbb{R}^d \rightarrow \mathbb{R}^d \) such that
\[
x = (x_1, \ldots, x_d) \in \mathbb{R}^d \mapsto r(x) = (r_1(x_1), \ldots, r_d(x_d)) \in \mathbb{R}^d,
\]
with \( r_1, \ldots, r_d \in W \). Then, for all \( r_1 \) and \( r_2 \in W^d \),
\[
\rho(r_1(c_1), r_2(c_2)) = \rho(c_1, c_2).
\]
In words, the similarity index between two curves is unaffected by strictly increasing affine transformations of one or more components of the curves.

For \( c = \{c_1, \ldots, c_d\} \in \mathcal{C} \), assume the existence of \( \varphi = \{\varphi_1, \ldots, \varphi_d\} \in \mathcal{C} \) and of a parameter vector \( \theta = (\theta_{01}, \ldots, \theta_{0d}, \theta_{11}, \ldots, \theta_{1d}, \theta_2, \theta_3) \), with \( \theta_{0p} \in \mathbb{R} \) and \( \theta_{1p} \in \mathbb{R}^+ \) for \( p = 1, \ldots, d \), \( \theta_2 \in \mathbb{R} \), \( \theta_3 \in \mathbb{R}^+ \), such that
\[
(5) \quad c_p(s) = \theta_{0p} + \theta_{1p} \varphi_p(\theta_2 + \theta_3 s) \quad \text{for } p = 1, \ldots, d.
\]
We shall write \( c \in \text{SIM}(\varphi) \), since the condition (5) means that \( c \) falls within a shape invariant model (SIM), with characteristic shape curve \( \varphi \). For \( d = 1 \), SIM models were introduced by Lawton et al. [17]. For further details, see Kneip and Gasser [18]. SIM models are strongly connected with the couple \((\rho, W)\) defined in (1) and (3). Indeed,
\[
(6) \quad \exists h \in W : \rho(c \circ h, \varphi) = 1 \iff c \in \text{SIM}(\varphi);
\]
this follows directly from (2) and (3). Note that, thanks to property (d), the roles of \( c \) and \( \varphi \) can be swapped. Now, consider a set of \( N \) curves \( \{c_1, \ldots, c_N\} \subset \mathcal{C} \), such that \( c_i \in \text{SIM}(\varphi) \) for \( i = 1, \ldots, N \); then, the following property follows immediately:

(f) For all \( c_i, c_j \), with \( i, j = 1, \ldots, N \), \( \exists h_i \in W, h_j \in W \) such that
\[
\rho(c_i \circ h_i, c_j \circ h_j) = \rho(c_i \circ h_i, \varphi) = \rho(c_j \circ h_j, \varphi) = 1 \quad \forall i, j = 1, \ldots, N.
\]

3. Curve clustering when curves are misaligned

Consider the problem of clustering and aligning a set of \( N \) curves \( \{c_1, \ldots, c_N\} \) with respect to a set of \( k \) template curves \( \varphi = \{\varphi_1, \ldots, \varphi_k\} \) (with \( \{c_1, \ldots, c_N\} \subset \mathcal{C} \) and \( \varphi \subset \mathcal{C} \)). For each template curve \( \varphi_j \) in \( \varphi \), define the domain of attraction
\[
(7) \quad \Delta_j(\varphi) = \{c \in \mathcal{C} : \sup_{h \in W} \rho(\varphi_j, c \circ h) \geq \sup_{h \in W} \rho(\varphi_r \circ h), \forall r \neq j\}, \quad j = 1, \ldots, k.
\]
Moreover, define the labeling function

\[
\lambda(\varphi, c) = \min \{ r : c \in \Delta_r(\varphi) \}.
\]

Note that \(\lambda(\varphi, c) = j\) means that the similarity index obtained by aligning \(c\) to \(\varphi_j\) is at least as large as the similarity index obtained by aligning \(c\) to any other template \(\varphi_r\), with \(r \neq j\). Thus \(\varphi_{\lambda(\varphi, c)}\) indicates a template the curve \(c\) can be best aligned to and hence \(\lambda(\varphi, c)\) a cluster which \(c\) should be assigned to.

**Case of known templates.** If the \(k\) templates \(\varphi = \{\varphi_1, \ldots, \varphi_k\}\) were known, then clustering and aligning the set of \(N\) curves \(\{c_1, \ldots, c_N\}\) with respect to \(\varphi\) would simply mean to assign \(c_i\) to the cluster \(\lambda(\varphi, c_i)\) and align it to the corresponding template \(\varphi_{\lambda(\varphi, c_i)}\), for \(i = 1, \ldots, N\).

Here we are interested in the more complex case when the \(k\) templates are unknown.

**Case of unknown templates.** Ideally, if our aim is clustering and aligning the set of \(N\) curves \(\{c_1, \ldots, c_N\}\) with respect to \(k\) unknown templates, we should first solve the following optimization problem

\[(i) \text{ find } \varphi = \{\varphi_1, \ldots, \varphi_k\} \subset \mathcal{C} \text{ and } h = \{h_1, \ldots, h_N\} \subset \mathcal{W} \text{ such that}
\]

\[
\frac{1}{N} \sum_{i=1}^{N} \rho(\varphi_{\lambda(\varphi, c_i)} \circ h_i, c_i) \geq \frac{1}{N} \sum_{i=1}^{N} \rho(\psi_{\lambda(\psi, c_i)} \circ g_i, c_i) \text{ for any other set of } k \text{ templates } \psi = \{\psi_1, \ldots, \psi_k\} \subset \mathcal{C} \text{ and any other set of } N \text{ warping functions } g = \{g_1, \ldots, g_N\} \subset \mathcal{W},
\]

and then, for \(i = 1, \ldots, N\),

\[(ii) \text{ assign } c_i \text{ to the cluster } \lambda(\varphi, c_i) \text{ and warp } c_i \text{ along } h_i.\]

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 \((\varphi, h)\) to (i) has mean similarity \(\frac{1}{N} \sum_{i=1}^{N} \rho(\varphi_{\lambda(\varphi, c_i)} \circ h_i)\) 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 \(h = \{h_1, \ldots, h_N\} \subset \mathcal{W}\) and a partition \(\mathcal{P} = \{P_1, \ldots, P_k\}\) of \(\{1, \ldots, N\}\) in \(k\) elements, such that \(\rho(c_i \circ h_i, c_j \circ h_j) = 1\) for all \(i\) and \(j\) belonging to the same element of \(\mathcal{P}\).
Because of (6), this is equivalent to the existence of \( k \) characteristic shape curves, \( \varphi_1, \ldots, \varphi_k \), such that

\[
\forall \ i = 1, \ldots, N, \quad \exists \ l_i \in \{1, \ldots, k\} : \ c_i \in \text{SIM}(\varphi_{l_i}). \tag{10}
\]

In this case the optimization problem (i) is solved by setting \( \varphi_\lambda(\varphi_1c_i) = \varphi_{l_i} \).

It should also be noted that, thanks to property (c), if \( \{\varphi_1, \ldots, \varphi_k\} \) and \( \{h_1, \ldots, h_N\} \) provide a solution to (i), then also \( \{\varphi_1 \circ g_1, \ldots, \varphi_k \circ g_k\} \) and \( \{h_1 \circ g_\lambda(\varphi_1), \ldots, h_N \circ g_\lambda(\varphi_N)\} \) is a solution to (i), for any \( \{g_1, \ldots, g_k\} \subset W \). 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. Numerical methods can be proposed for its solution, for instance employing mixing of discrete and continuous optimization algorithms. Following a statistical perspective, we here prefer to give an approximate solution to the optimization problem, integrating a classical functional \( k \)-mean or \( k \)-medoid clustering algorithm and a Procrustes procedure for curve alignment; the resulting iterative procedure simultaneously deals with (i) and (ii) alternating template identification steps, clusters assignment steps and alignment steps (see [8] for further details).

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; 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, and we assign each of the curves to one of the \( k \) clusters. As we shall see, the proposed clustering and alignment procedure also considers the problem of non-uniqueness of the solution, by targeting a specific solution via a normalization step.

3.1. Clustering and alignment iterative procedure

Let \( \varphi_{[q-1]} = \{\varphi_1^{[q-1]}, \ldots, \varphi_k^{[q-1]}\} \) be the set of templates after iteration \( q-1 \), and \( \{c_1^{[q-1]}, \ldots, c_N^{[q-1]}\} \) be the \( N \) curves aligned and clustered to \( \varphi_{[q-1]} \). At the \( q \)th iteration the algorithm performs the following steps.

**Template identification step.** For \( j = 1, \ldots, k \), the template of the \( j \)th cluster, \( \varphi_j^{[q]} \), is estimated using all curves assigned to cluster \( j \) at iteration \( q-1 \), i.e. all curves \( c_{i^{[q-1]}} \) such that \( \lambda(\varphi_{[q-1]}, c_i^{[q-1]}) = j \). Ideally, the template \( \varphi_j^{[q]} \) should be estimated as the curve \( \varphi \in \mathcal{C} \) that maximizes the within-cluster total similarity

\[
\sum_{i : \lambda(\varphi_{[q-1]}, c_i^{[q-1]}) = j} \rho(\varphi, c_i^{[q-1]}), \tag{11}
\]
i.e., $\varphi_{j[q]}$ should be the functional median, or Fréchet median, associated to the similarity $\rho$.

Assignment and alignment step. The set of curves $\{c_{1[q−1]}, \ldots, c_{N[q−1]}\}$ is clustered and aligned to the set of templates $\mathcal{C}[q] = \{\varphi_{1[q]}, \ldots, \varphi_{k[q]}\}$: for $i = 1, \ldots, N$, the $i$-th curve $c_{i[q−1]}$ is aligned to $\varphi_{\lambda(\mathcal{C}[q],c_{i[q−1]})}$ and the aligned curve $\tilde{c}_{i[q]} = c_{i[q−1]} \circ h_{i[q]}$ is assigned to cluster $\lambda(\mathcal{C}[q],\tilde{c}_{i[q]}).$

Normalization step. After each assignment and alignment step, we also perform a normalization step. In detail, for $j = 1, \ldots, k$, all the $N_j[q]$ curves $\tilde{c}_{i[q]}$ assigned to cluster $j$ are warped along the warping function $(h_{j[q]})^{-1}$, where

\begin{equation}
\tilde{h}_{j[q]} = \frac{1}{N_j[q]} \sum_{i: \lambda(\mathcal{C}[q],\tilde{c}_{i[q]}) = j} h_{i[q]} \tag{12}
\end{equation}

obtaining $c_{i[q]} = \tilde{c}_{i[q]} \circ (\tilde{h}_{j[q]})^{-1} = c_{i[q−1]} \circ h_{i[q]} \circ (\tilde{h}_{j[q]})^{-1}$. In this way, at each iteration, the average warping undergone by curves assigned to cluster $j$ is the identity transformation $h(s) = s$. Indeed:

\begin{equation}
\frac{1}{N_j[q]} \sum_{i: \lambda(\mathcal{C}[q],c_{i[q]}) = j} (h_{i[q]} \circ (\tilde{h}_{j[q]})^{-1})(s) = s, \quad j = 1, \ldots, k. \tag{13}
\end{equation}

The normalization step is thus used to select, among all candidate solutions to the optimization problem, the one that leaves the average locations of the clusters unchanged, thus avoiding the drifting apart of clusters or the global drifting of the overall set of curves. Note that the normalization step preserves the clustering structure chosen in the maximization step, i.e., $\lambda(\mathcal{C}[q],\tilde{c}_{i[q]}) = \lambda(\mathcal{C}[q],c_{i[q]})$ for all $i$.

The algorithm is initialized with a set of initial templates, $\mathcal{C}[0] = \{\varphi_1[0], \ldots, \varphi_{k[0]}\} \subset \mathcal{C}$, and with $\{c_{1[0]}, \ldots, c_{N[0]}\} = \{c_1, \ldots, c_N\}$, and stopped when, in the assignment and alignment step, the increments of the similarity indexes are all lower than a fixed threshold.

4. Template identification

Whilst the assignment and alignment step and the normalization step are less troublesome, the template identification step raises more issues from a theoretical and computational point of view. In fact, the identification of the template $\varphi_{j[q]}$, as the curve $\varphi \in \mathcal{C}$ that maximizes the total similarity $(11)$, cannot be easily dealt with. For this reason, in [8] we proposed to estimate the template $\varphi_{j[q]}$ as a loess, with Gaussian kernel and appropriate
smoothness parameter, of the curves assigned to cluster $j$ at iteration $q-1$ (i.e. all curves $c_{i[q-1]}$ such that $\lambda(\varphi[q-1], c_{i[q-1]}) = j$). See [8] for details on the implementation. The algorithm obtained with this specification for the template identification step was named $k$-mean alignment, in analogy with the $k$-mean clustering algorithms for multivariate and for functional data.

On the other hand, the fact that the template is estimated by loess of the curves assigned to the cluster, instead of the curve that maximize the total similarity (11), raises doubts about a possible distortion of the algorithm. Moreover, estimating the template by loess of the curves assigned to the cluster might make this step sensitive to the presence of anomalous data.

For this reason, we propose here an alternative specification of the template identification step, that constitutes a direct approximation to the maximization of the total similarity (11). In particular, we restrict the set over which the maximization is carried out, limiting the search to functions of the sample. The template $\varphi[j\|q]$ is thus estimated as the curve $\varphi$, among all curves assigned to cluster $j$ at iteration $q-1$ (i.e. all curves $c_{i[q-1]}$ such that $\lambda(\varphi[q-1], c_{i[q-1]}) = j$), that maximizes the total similarity (11)

$$\varphi[j\|q] = \arg \max_{c_{i[q-1]}: \lambda(\varphi[q-1], c_{i[q-1]}) = j} \sum_{k: \lambda(\varphi[q-1], c_{k[q-1]}) = j} \rho(c_{i[q-1]}, c_{k[q-1]}).$$

Note that the $k$ curves selected as templates, in the template identification step, shall skip the subsequent assignment and alignment step and normalization step.

The algorithm obtained with this specification for the template identification step will be named $k$-medoid alignment, in analogy with the $k$-medoid clustering algorithms described for instance by [2] and [19]. The same analogy suggests that this novel specification of the template identification step is less sensitive to the presence of anomalous data. In particular, the comparison of the clustering results obtained with the two alternative algorithm versions, the $k$-mean and the $k$-medoid version, might indicate accidental anomalous data, as we shall see in the applications to real data described in the following sections. Moreover, thanks to these alternative specifications, we will have the possibility of highlighting the robustness of our alignment and clustering procedure when different algorithm specifications are considered.
5. An application to the analysis of three-dimensional cerebral vascular geometries

In this section, \( k \)-mean and \( k \)-medoid alignment are used to improve upon the exploratory statistical analyses of the AneuRisk Project\(^a\) dataset (previous analyses are detailed in \([8,13,20]\)). The AneuRisk Project is a joint research program that aims at evaluating the role of vascular geometry and hemodynamics in the pathogenesis of cerebral aneurysms. 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 Sangalli et al. \([21]\). Details about the elicitation of discrete observations from row data can be found in Antiga et al. \([22]\). 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., \([23]\)) ICA’s are classified in \( \Gamma \)-shaped, \( \Omega \)-shaped, and \( S \)-shaped ICA’s, according to the form of their distal part, which may resemble the letters \( \Gamma, \Omega \) or \( S \), in presence of zero, one, or two siphons, respectively. In \([8]\), using \( k \)-mean alignment, we were able to identify a cluster of \( \Omega \)-shaped ICA’s and a cluster \( S \)-shaped ICA’s, in the AneuRisk dataset. We want here to verify if this clustering result is confirmed by application of \( k \)-medoid alignment, and recognize accidental anomalous data.

Figure 2, 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 \( k = 1, 2, 3 \); the same panel also displays in blue the boxplots of the similarity indexes between the original centerlines and their estimated medoid, and the boxplots of the similarity indexes obtained by \( k \)-medoid alignment, for \( k = 1, 2, 3 \).

---

\(^a\)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.
Figure 2. 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 2, 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 considerable gain can be obtained by setting $k = 2$ in the clustering and aligning procedure, both $2$-mean and $2$-medoid, whereas an eventual choice of $k = 3$ is not justified by an additional increase in the similarities. Thus, $k$-medoid alignment, likewise $k$-mean alignment, suggests the presence of $k = 2$ shape characteristic curves within the analyzed centerlines.

Figure 3 compares the first derivatives of $2$-mean and $2$-medoid aligned centerlines (left and right, respectively). The two clusters identified by $2$-medoid alignment, similarly to the two clusters identified by $2$-mean alignment, can be described as the $\Omega$-shaped ICA’s cluster (orange) and $S$-shaped ICA’s cluster (green). This can be better appreciated in Figures 4 and 5 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 $\Omega$-shaped ICA (one siphon in the distal part), whereas the template of the green cluster is a prototype of $S$-shaped ICA (two siphons in the distal part).

Finally, Table 1 compares the cluster assignments obtained by $2$-mean and $2$-medoid alignment (“$\Omega$” stands for the cluster of $\Omega$-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. 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
Figure 3. First derivatives of 2-mean (left) and 2-medoid (right) aligned centerlines, with superimposed first derivatives of estimated templates (black lines); the curve color depends on cluster assignment.

Figure 4. 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).

Figure 5. 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).

morphological shape, that resembles a Γ-shaped ICA rather than a Ω or a S-shaped one.

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).

<table>
<thead>
<tr>
<th></th>
<th>2-medoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ω</td>
<td>34</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
</tr>
<tr>
<td>Ω</td>
<td>3</td>
</tr>
<tr>
<td>S</td>
<td>27</td>
</tr>
</tbody>
</table>
A detailed comparison between the results achieved by simple $k$-mean clustering without alignment and $k$-mean alignment is reported in [8]. Similar conclusions can be drawn by comparing simple $k$-medoid clustering without alignment and $k$-medoid alignment. In particular, if alignment is not taken into account, the confounding effect of phase variability drives the iterative procedure. Indeed, the two templates detected by simple $k$-medoid clustering without alignment appear to have almost the same morphological shape, and seem to differ mainly in their phase, failing to identify the $S$-shaped and $\Omega$-shaped clusters.

6. An application to the analysis of growth data

In this section we apply $k$-mean and $k$-medoid alignment for the analysis of a benchmark data set in the functional data analysis literature: the 93 growth curves from Berkeley Growth Study (see Tuddenham and Snyder [24]). These data have been previously considered by a number of authors (see for example Ramsay and Li [25], Ramsay and Silverman [3], James [26], and references therein). In particular, in this paper we will improve upon the analysis illustrated in Sangalli et al. [8], confirming the interesting insight of joint clustering and alignment with respect to sole clustering or sole alignment.

The heights (in cm) of the 93 children in the data set are measured quarterly from 1 to 2 years, annually from 2 to 8 years and biannually from 8 to 18 years. The growth curves are estimated by means of monotonic cubic regression splines (see Ramsay and Silverman [3]), implemented using the R function smooth.monotone available in the fda package [27]. Figure 6 shows the estimated growth curves and their derivatives, the growth velocities. Looking at the growth velocities, it is apparent that the children follow a similar growth course, but that each child has a personal biological clock.
Figure 7. Results of $k$-mean alignment of growth curves, for $k=1; 2$: aligned growth curves (with superimposed estimated templates, in black) and corresponding growth velocities (with superimposed first derivatives of estimated templates, in black), together with warping functions. The colors of aligned curves and warping functions depend on cluster assignment.

Figure 7 shows 1-mean and 2-mean aligned growth curves, the corresponding growth velocities and warping functions; Figure 8 shows the corresponding results obtained by $k$-medoid alignment. Figure 9, left, displays in orange (blue), the boxplots of the similarity indexes between the original growth curves and their mean (medoid) curve, indicated with “orig”, and the boxplots of the similarity indexes between the $k$-mean aligned ($k$-medoid aligned) growth curves and their estimated templates, for $k = 1, 2, 3$. The right panel of Figure 9 displays the corresponding means of similarity indexes. From inspection of the similarity indexes, both $k$-mean and $k$-medoid alignment suggest the presence of just one characteristic curve, since the choice of $k=2$ is not payed off by a reasonable gain in the similarities.

Since, out of the 93 children, 39 are boys and 54 are girls, we might wonder if the analysis points out some differences among them (notice that here we are not performing any supervised classification of boys and girls). Figure 10 is obtained from Figure 8 (top panels) displaying in blue the 1-medoid aligned growth curves of boys, and the corresponding growth velocities and warping functions, and in pink the ones of girls. The warping functions show a pretty neat separation of boys and girls in the phase; this highlights that the biological clocks of boys and girls run at different speeds, and in particular that boys start to grow later, having warping functions with smaller
Figure 8. Results of $k$-medoid alignment of growth curves, for $k=1; 2$: aligned growth curves (with superimposed estimated templates, in black) and corresponding growth velocities (with superimposed first derivatives of estimated templates, in black), together with warping functions. The colors of aligned curves and warping functions depend on cluster assignment.

Figure 9. Left: in orange (blue), boxplots of similarity indexes between the original growth curves and their mean (medoid) curve, “orig”, and boxplots of similarity indexes between the $k$-mean aligned ($k$-medoid aligned) growth curves and their estimated templates, for $k = 1, 2, 3$. Right: corresponding means of similarity indexes.

intercepts, and grow slower, having warping functions with smaller slopes. The left panel of Figure 10 also shows that, once the biological clocks of the children have been aligned, the height of boys stochastically dominates the one of girls for any registered biological age. Finally, boys seem also to have a more pronounced growth, especially during puberty, as highlighted
by their more prominent growth velocity peak. All these features are in complete agreement with the results obtained by 1-mean alignment, and discussed in detail in [8].

Figure 10. Figure obtained from Figure 8 (top panels) displaying in blue the growth curves, growth velocities and warping functions of boys and in pink the ones of girls.

Table 2. Left: results of 2-mean bivariate clustering of slopes and intercepts of the warping functions obtained by 1-mean alignment. Right: results of 2-medoid bivariate clustering of slopes and intercepts of the warping functions obtained by 1-medoid alignment. Cluster assignment vs gender.

<table>
<thead>
<tr>
<th></th>
<th>2-mean clusters</th>
<th>2-medoid clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>clusters</td>
<td>clusters</td>
</tr>
<tr>
<td>F</td>
<td>44 10</td>
<td>43 11</td>
</tr>
<tr>
<td>M</td>
<td>1 38</td>
<td>1 38</td>
</tr>
</tbody>
</table>

The grouping structure of the warping functions obtained by 1-mean and 1-medoid alignment of the growth curves, can be explored by a coherent unsupervised classification of their slopes and intercepts, i.e. by 2-mean and 2-medoid bivariate clustering respectively. The results of these unsupervised classifications are shown in Table 2. Note that both 2-mean clustering of the warping functions obtained by 1-mean alignment of growth curves, and 2-medoid clustering of the warping functions obtained by 1-medoid alignment of growth curves, assign 1 boy to the female prevalent cluster, and respectively 10 and 11 girls to the male prevalent cluster. From inspection of Table 3, which compares the cluster assignments of the two procedures, we conclude that this boy and the 10 girls are exactly the same. Thus, both algorithms agree that these ten girls have biological clocks closer to those of boys, and that the boy has a biological clock closer to those of girls. This fact is evident in Figure 11, which displays the slopes and intercepts of the warping functions (pink for girls and blue for boys), and highlights in red the ones of the ten girls and in green the one of the boy. Figure 11 also displays in black the only mismatch between the two algorithms, which has in fact a biological clock borderline between the two groups.
Table 3. A comparison between cluster assignments obtained by 2-mean and 2-medoid bivariate clustering of slopes and intercepts of the warping functions obtained by 1-mean alignment and 1-medoid alignment of growth curves.

<table>
<thead>
<tr>
<th></th>
<th>2-mean</th>
<th>2-medoid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48</td>
</tr>
</tbody>
</table>

Figure 11. Slopes and intercepts of the warping functions resulting from 1-mean alignment (left) and 1-medoid alignment (right) of growth curves. Blue circles correspond to boys and pink circles to girls. Red (green) circles are girls (boys) which in both cases have been assigned to the male (female) prevalent cluster when the corresponding bivariate clustering procedure is applied. The black circle is the only mismatching case between the two procedures (see Table 3).

7. Discussion

We have considered the issue of classification of functional data and, in particular, we have dealt with the problem of curve clustering when curves are misaligned. We have described a proper mathematical framework where the problem of clustering and alignment is stated as an optimization problem. We have thus illustrated two alternative specifications of an iterative procedure used to solve this optimization problem. The first one, a $k$-mean alignment version, was proposed in [8]. The second one, named $k$-medoid alignment algorithm, is novel; this version approximates more directly the clustering and alignment optimization problem and is potentially less sensitive to the presence of anomalous data. Moreover, this alternative version has given us the possibility of testing the robustness of our alignment and clustering procedure: the agreement in the results obtained with different algorithm specifications supports the soundness of the approximated proce-
dure we use to solve the optimization problem.

Both approaches have confirmed the potentiality of joint clustering and alignment with respect to sole clustering or sole alignment, pointing out features that can neither be captured by simple clustering without alignment nor by simple curve alignment without clustering. Details of the comparison of clustering and alignment with the other methods can be found in [8].

REFERENCES

13. L. M. Sangalli, P. Secchi, S. Vantini, and A. Veneziani, A case study in exploratory functional data analysis: Geometrical features of the internal


