MOX–Report No. 13/2008

K-means alignment for curve clustering

Laura Maria Sangalli, Piecesare Secchi,
Simone Vantini, Valeria Vitelli

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 29 - 20133 Milano (Italy)

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

K-means alignment for curve clustering *

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

June 3, 2008

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

Keywords: Functional data analysis, curve alignment, curve clustering, k-means algorithm

AMS Subject Classification: 62H30, 68T10

Abstract

We deal with the problem of curve clustering when curves are misaligned. We propose a k-means alignment algorithm which jointly cluster and align the curves. We illustrate the procedure via simulation studies and applications to real data.

A problem, often encountered in functional data analysis, is misalignment of the data. A typical example, considered by a number of authors, is given by the kids growth curves (see for example Ramsay and Li [13], Sheehy et al. [18, 19], Ramsay and Silverman [14], and James [6]). Figure 1 shows the growth curves of 93 kids (39 boys and 54 girls) from Berkeley Growth Study data (see Tuddenham and Snyder [22]). Looking at the corresponding growth velocities, also displayed in Figure 1, it is apparent that all growth curves follow a similar course, characterized by a sharp peak of growth velocity around 12 years, the pubertal spurt, and a minor velocity peak around 4 years, the mid-spurt; but different kids have their growth spurts at different times, some take more time in their spurts, others less, each kid following his/her personal biological clock. Thus, to learn something about the common growth path, it is first necessary

*This work has been supported by Ministero dell'Istruzione dell'Università e della Ricerca (research project "Metodi numerici avanzati per il calcolo scientifico" PRIN2006). The dataset analized in Section 5 is provided by the Aneurisk Project.

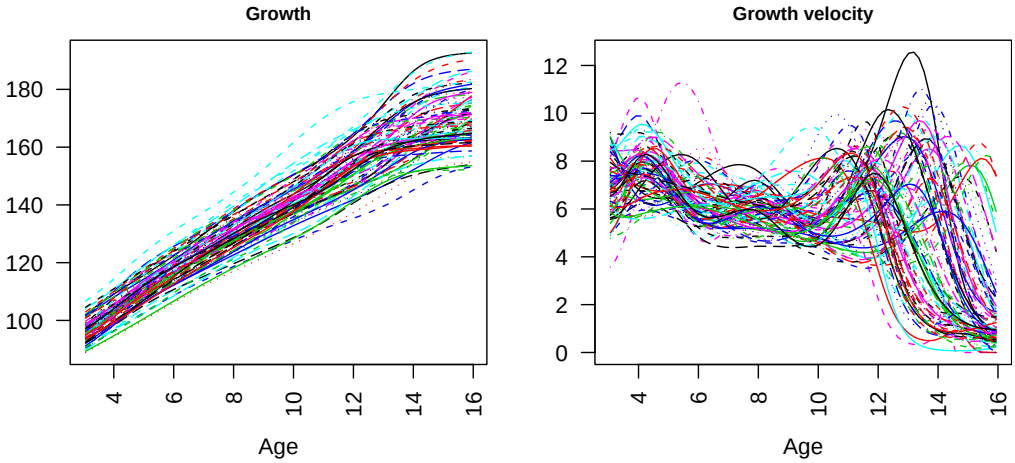

Figure 1: Growth curves of 93 kids from Berkeley Growth Study data (left) and corresponding growth velocity (right).

to register the biological clocks of the kids, separating the variability due to the different timings.

Many methods for curve alignment (or curve registration) have been proposed in the literature. For example, Lawton et al. [10] and Altman and Villarreal [1] deal with this problem using self-modelling non-linear regression methods, Lindstrom and Bates [11] instead develop non-linear mixed-effects models, and Ke and Wang [8] 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 Ramsay and Li [13], Ramsay and Silverman [14], James [6], Kaziska and Srivastava [7] and Sangalli et al. [17], who define suitable similarity indexes between curves and thus align the curves maximizing their similarities, by means of a Procustes procedure.

The present paper is in the latter line of research, and moves forward from the problem of curve alignment, per se, focussing on the more complex problem of curve clustering when curves are misaligned. Look at Figure 2. Do the two clusters of curves in case B and case C represent two sets of curves with distinct shapes? Or do they reflect a clustering in the phase, that could be eliminated if the curves were suitably aligned? How many set of curves with distinct shape are present in case D?

We describe a procedure that is able to efficiently cluster and align in k groups a set of curves. If the number of clusters k is set equal to 1, the algorithm implements a Procustes aligning procedure as the ones mentioned above, whereas, if no alignment is allowed, it implements a k-means clustering of curves (see Heckman and Zamar [5], Tarpey and Kinateder [21] and Shimizu and Mizuta [20] for other implementations of k-means algorithms for curve clustering). For this reason we well call it a k-means alignment algorithm.

The paper is organized as follows. In Section 1 we formally describe the prob-

lem of curve alignment. In Section 2 we consider the problem of curve clustering when curves are misaligned and describe the k-means alignment algorithm. Section 3 illustrates the efficiency of the algorithm via simulated studies. Section 4 shows the application to growth curves data, whilst Section 5 is devoted to the application to another real dataset, concerning three-dimensional vascular geometries. Finally, some conclusive considerations are drawn in Section 6.

All simulations and analysis of real datasets are performed in R©.

1 Defining phase and amplitude variabilities

The variability among two or more curves can be though of as having two components: phase variability and amplitude variability. Heuristically, phase variability is the variability that can be eliminated by suitably aligning the curves, and amplitude variability is the remaining variability among the curves once they have been aligned. Consider a space $\mathcal{C}$ of curves $\mathbf{c}(s) : \mathbb{R} \to \mathbb{R}^d$. Aligning $\mathbf{c}_1(s) \in \mathcal{C}$ to $\mathbf{c}_2(s) \in \mathcal{C}$ means finding a warping function $h(s) : \mathbb{R} \to \mathbb{R}$, of the abscissa parameter s, such that the two curves $\mathbf{c}_1(h(s))$ and $\mathbf{c}_2(s)$ are the most similar. It is thus necessary to choose a class W of admissible warping functions h (with $(\mathbf{c} \circ h)(s) := \mathbf{c}(h(s)) \in \mathcal{C}$, for any $\mathbf{c} \in \mathcal{C}$ and $h \in W$), and a similarity index $\rho(\cdot, \cdot) : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ that measures the similarity between two curves. Aligning $\mathbf{c}_1$ to $\mathbf{c}_2$, according to (ρ, W), means finding $h^* \in W$ that maximizes $\rho(\mathbf{c}_1 \circ h, \mathbf{c}_2)$. This procedure decouples phase and amplitude variability without loss of information: the phase variability is captured by the optimal warping function h^*, whilst the amplitude variability is the remaining variability between $\mathbf{c}_1 \circ h^*$ and $\mathbf{c}_2$. Note that the choice of the couple (ρ, W) defines what is meant by phase variability and by amplitude variability.

Many similarity indexes for measuring closeness 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 [3]. Sangalli et al. [17] proposed the following similarity index between two curves $\mathbf{c}_1 \in L^2(S_1 \subset \mathbb{R}; \mathbb{R}^d)$ and $\mathbf{c}_2 \in L^2(S_2 \subset \mathbb{R}; \mathbb{R}^d)$, where $\mathbf{c}_1' \in L^2(S_1 \subset \mathbb{R}; \mathbb{R}^d)$, $\mathbf{c}_2' \in L^2(S_2 \subset \mathbb{R}; \mathbb{R}^d)$ and $S_{12} = S_1 \cap S_2$ has positive Lebesgue measure:

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = \frac{1}{d} \sum_{p=1}^{d} \frac{\int_{S_{12}} c_{1p}'(s) c_{2p}'(s) ds}{\sqrt{\int_{S_{12}} c_{1p}'(s)^2 ds} \sqrt{\int_{S_{12}} c_{2p}'(s)^2 ds}}.$$

with c_{ip} indicating the pth component of $\mathbf{c}_i$, $\mathbf{c}_i = \{c_{i1}, \ldots, c_{id}\}$. According to this index, two curves are completely similar when they are identical except for shifts and dilations of the components, i.e.

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = 1 \Leftrightarrow \text{for } p = 1, \ldots, d, \ \exists A_p \in \mathbb{R}^+, B_p \in \mathbb{R} : c_{1p} = A_p c_{2p} + B_p. \quad (1)$$

The choice of this similarity index comes along with the following choice for the class W of warping functions:

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

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

The similarity index ρ and the class of warping function W in (1) and (2) satisfy some minimal requirements, that we deem necessary for the well posedness of the alignment problem:

- The similarity index ρ is bounded, with maximum value equal to 1, so that two curves $\mathbf{c}_1$ and $\mathbf{c}_2$ are completely similar when $\rho(\mathbf{c}_1, \mathbf{c}_2) = 1$; moreover, ρ is reflexive (i.e., $\rho(\mathbf{c}, \mathbf{c}) = 1, \ \forall\ \mathbf{c} \in \mathcal{C}$), symmetric (i.e., $\rho(\mathbf{c}_1, \mathbf{c}_2) = \rho(\mathbf{c}_2, \mathbf{c}_1), \ \forall\ \mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$) and transitive (i.e., $\left[\rho(\mathbf{c}_1, \mathbf{c}_2) = 1, \ \rho(\mathbf{c}_2, \mathbf{c}_3) = 1\right] \Rightarrow \rho(\mathbf{c}_1, \mathbf{c}_3) = 1, \ \forall\ \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3 \in \mathcal{C}$).

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

- The choices of the similarity index ρ and the class of warping functions W are consistent in the sense that, if two curves $\mathbf{c}_1$ and $\mathbf{c}_2$ are simultaneously warped along the same warping function $h \in W$, their similarity does not change:

$$\rho\left(\mathbf{c}_i, \mathbf{c}_j\right) = \rho\left(\mathbf{c}_i \circ h, \mathbf{c}_j \circ h\right) \qquad \forall\ h \in W. \tag{3}$$

  This guarantees that it is not possible to obtain a fictitious increment of the similarity between two curves $\mathbf{c}_i$ and $\mathbf{c}_j$ by simply moving them simultaneously to $\mathbf{c}_i \circ h$ and $\mathbf{c}_j \circ h$.

2 Curve clustering when curves are misaligned

Consider the problem of clustering and aligning a set of N curves $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$ with respect to a set of k template curves $\underline{\varphi} = \{\varphi_1, \ldots, \varphi_k\}$ (with $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\} \in \mathcal{C}^N$ and $\underline{\varphi} \in \mathcal{C}^k$). For each template curve φ_j in $\underline{\varphi}$, define the domain of attraction

$$\Delta_j(\underline{\varphi}) = \{\mathbf{c} \in \mathcal{C} : \sup_{h \in W} \rho(\varphi_j, \mathbf{c} \circ h) \geq \sup_{h \in W} \rho(\varphi_r, \mathbf{c} \circ h), \forall\ r \neq j\}, \quad j = 1, \ldots, k\,;$$

moreover define the labeling function

$$\lambda(\underline{\varphi}, \mathbf{c}) = \min\{r : \mathbf{c} \in \Delta_r(\underline{\varphi})\}.$$

Note that $\lambda(\underline{\varphi}, \mathbf{c}) = j$ means that the similarity index obtained by aligning $\mathbf{c}$ to φ_j is at least as big as the similarity index obtained by aligning $\mathbf{c}$ to any other template φ_r, with $r \neq j$. Thus $\lambda(\underline{\varphi}, \mathbf{c})$ indicates a template the curve $\mathbf{c}$ can be best aligned to and hence a cluster it should be assigned to.

Now, if the k templates $\underline{\varphi} = \{\varphi_1, \ldots, \varphi_k\}$ were known, then clustering and aligning the set of N curves $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$ with respect to $\underline{\varphi}$ would simply mean to assign $\mathbf{c}_i$ to the cluster $\lambda(\underline{\varphi}, \mathbf{c}_i)$ and align it to the corresponding template $\varphi_{\lambda(\underline{\varphi}, \mathbf{c}_i)}$, for $i = 1, \ldots, N$.

Here we are interested in the more complex case were the k templates are unknown. Ideally, in order to cluster and align the set of N curves $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$ with respect to k unknown templates we should solve the following optimization problem:

1. find $\underline{\varphi} = \{\varphi_1, \ldots, \varphi_k\} \in \mathcal{C}^k$ such that

$$\sum_{i=1}^{N} \sup_{h \in W} \rho(\varphi_{\lambda(\underline{\varphi}, \mathbf{c}_i)}, \mathbf{c_i} \circ h) \geq \sum_{i=1}^{N} \sup_{h \in W} \rho(\psi_{\lambda(\underline{\psi}, \mathbf{c}_i)}, \mathbf{c_i} \circ h)$$

   for any other set of k templates $\underline{\psi} = \{\psi_1, \ldots, \psi_k\} \in \mathcal{C}^k$;

2. cluster and align the N curves to $\underline{\varphi} = \{\varphi_1, \ldots, \varphi_k\}$.

Note that if $\{\varphi_1, \ldots, \varphi_k\}$ is a solution to the optimization problem, then also $\{\varphi_1 \circ h_1, \ldots, \varphi_k \circ h_k\}$ is a solution, for any $h_1, \ldots, h_k \in W$.

Unfortunately, point 1 of the optimization problem is not easily solvable. We thus deal with the optimization problem through a $k-$means alignment algorithm that iteratively alternates expectation and maximization steps. In the expectation steps we estimate the set of k candidate templates, identifying the k clusters; in the maximization steps we assign each of the N curves to one of the k clusters and align it to the corresponding template, maximizing the similarity index. Moreover, after each maximization step, we perform a normalization step in which we select, among all candidate solutions to the optimization problem, the one that leaves the locations of the clusters unchanged, as will be clarified in Remark 1.

2.1 k-means alignment algorithm

We describe here a k-means aligning algorithm suitable for the choice of (ρ, W) described in Section 1; by adapting the technical details, the algorithm can be applied for different choices of (ρ, W).

k-means alignment algorithm. Let $\underline{\varphi}^{[q-1]} = \{\varphi_{1\,[q-1]}, \ldots, \varphi_{k\,[q-1]}\}$ be the set of templates after iteration $q-1$, and $\{\mathbf{c}_{1\,[q-1]}, \ldots, \mathbf{c}_{N\,[q-1]}\}$ be the N curves aligned and clustered to $\underline{\varphi}^{[q-1]}$. At the qth iteration the algorithm performs the following steps.

Expectation step. For $j = 1, \ldots, k$, the template of the jth cluster, $\varphi_{j\,[q]}$, is estimated using all curves assigned to cluster j at iteration $q-1$, i.e. all curves $\mathbf{c}_{i\,[q-1]}$ such that $\lambda(\underline{\varphi}^{[q-1]}, \mathbf{c}_{i\,[q-1]}) = j$. Ideally the template should be estimated as the curve $\varphi \in \mathcal{C}$ that maximizes the similarity:

$$\sum_{i:\lambda(\underline{\varphi}^{[q-1]}, \mathbf{c}_{i\,[q-1]})=j} \rho(\varphi, \mathbf{c}_{i\,[q-1]}).$$

The implementation of the algorithm that we illustrate here computes the template $\varphi_{j\,[q]}$ by means of Loess, as detailed in Remark 2.

Maximization step. The set of curves $\{\mathbf{c}_{1\,[q-1]}, \ldots, \mathbf{c}_{N\,[q-1]}\}$ is clustered and aligned to the set of templates $\underline{\varphi}_{[q]} = \{\varphi_{1\,[q]}, \ldots, \varphi_{k\,[q]}\}$: for $i = 1, \ldots, N$, the i-th curve $\mathbf{c}_{i\,[q-1]}$ is aligned to $\varphi_{\lambda(\underline{\varphi}_{[q]}, \mathbf{c}_{i\,[q-1]})}$ and the aligned curve $\tilde{\mathbf{c}}_{i\,[q]} = \mathbf{c}_{i\,[q-1]} \circ h_{i\,[q]}$ is assigned to cluster $\lambda(\underline{\varphi}_{[q]}, \mathbf{c}_{i\,[q-1]}) \equiv \lambda(\underline{\varphi}_{[q]}, \tilde{\mathbf{c}}_{i\,[q]})$.

Normalization step. For $j = 1, \ldots, k$, all the $N_{j\,[q]}$ curves $\tilde{\mathbf{c}}_{i\,[q]}$ assigned to cluster j are warped along the warping function $(\bar{h}_{j\,[q]})^{-1}$, where

$$\bar{h}_{j\,[q]} = \frac{1}{N_{j\,[q]}} \sum_{i:\lambda(\underline{\varphi}_{[q]}, \tilde{\mathbf{c}}_{i\,[q]})=j} h_{i\,[q]}$$

obtaining $\mathbf{c}_{i\,[q]} = \tilde{\mathbf{c}}_{i\,[q]} \circ (\bar{h}_{j\,[q]})^{-1} = \mathbf{c}_{i\,[q-1]} \circ h_{i\,[q]} \circ (\bar{h}_{j\,[q]})^{-1}$.

The algorithm is initialized with k distinct templates, $\underline{\varphi}_{[0]} = \{\varphi_{1\,[0]}, \ldots, \varphi_{k\,[0]}\}$, chosen at random among the N curves available, and stopped when, in the maximization step, the increments of the similarity indexes are all lower than 0.01 (i.e., 1% of the achievable maximum).

Remark 1. At the qth iteration, the average warping underwent by curves assigned to cluster j is the identity transformation $h(s) = s$. Indeed:

$$\frac{1}{N_{j\,[q]}} \sum_{i:\lambda(\underline{\varphi}_{[q]}, \mathbf{c}_{i\,[q]})=j} \big(h_{i\,[q]} \circ (\bar{h}_{j\,[q]})^{-1}\big)(s) = s, \qquad j = 1, \ldots, n.$$

Hence, the normalization step is 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 the 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., $\lambda(\underline{\varphi}_{[q]}, \tilde{\mathbf{c}}_{i\,[q]}) = \lambda(\underline{\varphi}_{[q]}, \mathbf{c}_{i\,[q]})$ for all i.

Remark 2. When performing the k-means alignment algorithm with the choices of ρ and W described in Section 1, we implement the maximization step as follows. We estimate the first derivative of the template of each cluster, starting from the first derivatives of the curves assigned to that cluster at the previous iteration, by means of Loess with gaussian kernel and an appropriate smoothness parameter α (see for example Cleveland and Grosse [2]). We use this adaptive fitting method in order to keep the variance of the estimator of the template as constant as possible along the abscissa (see for example Hastie and Tibshirani [4]), since the domains of the curves are no longer the same due to curve alignment. Note that, for the implementation of the algorithm, it is sufficient to estimate the first derivatives of the templates, rather than the templates themselves, thanks to the specific choice of (ρ, W).

3 Simulation studies

In this section we illustrate the potential of our k-means alignment algorithm through a 4-case simulation study.

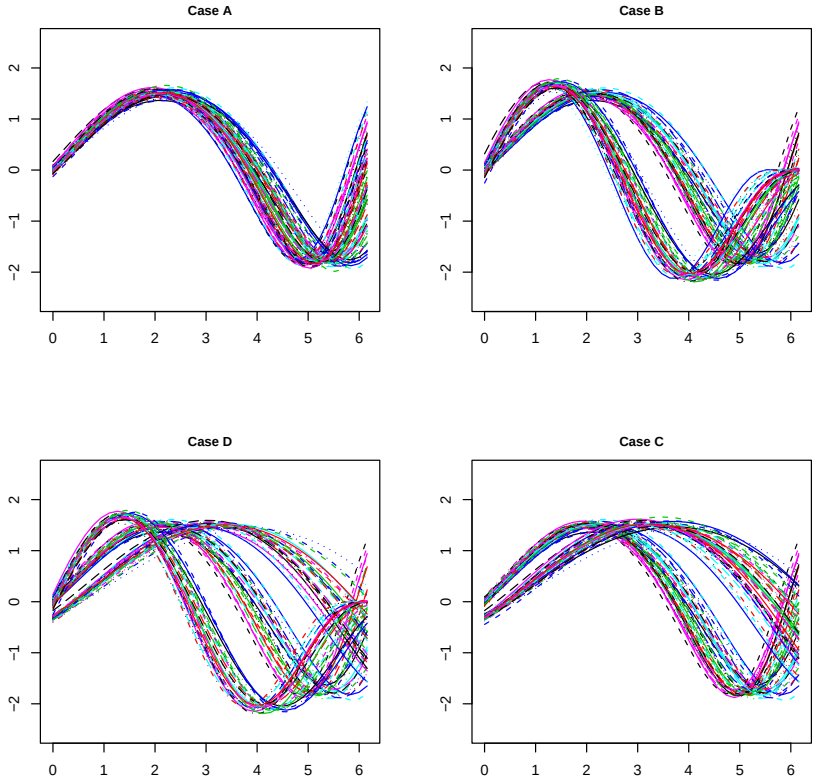

Figure 2: Starting from top left, clockwise, curves simulated in case A, case B, case C and case D.

3.1 Data generation

Consider the template:

$$c(s) = 1 * \sin(s) + 1 * \sin\left(\frac{s^2}{2\pi}\right) \qquad 0 \leq s \leq 2\pi \tag{4}$$

CASE A. We simulate 90 curves from template (4), with small errors in amplitude and phase, i.e. for $i = 1, \ldots, 90$ we generate

$$c_i^{[A]}(s) = (1 + \varepsilon_{1i}) * \sin(\varepsilon_{3i} + \varepsilon_{4i} s) + (1 + \varepsilon_{2i}) * \sin\left(\frac{(\varepsilon_{3i} + \varepsilon_{4i} s)^2}{2\pi}\right) \qquad 0 \leq s \leq 2\pi$$

where the errors ε are all independent and normally distributed with mean 0 and standard deviation 0.05. The simulated 90 curves are displayed in case A of Figure 2.

CASE B. The 90 curves displayed in case B of Figure 2, $c_1^{[B]}, \ldots, c_{90}^{[B]}$, are obtained as follows:
- for $i = 1, \ldots, 45$, $c_i^{[B]} = c_i^{[A]}$;
- for $i = 46, \ldots, 90$, $c_i^{[B]}$ is obtained from $c_i^{[A]}$ by modifying its amplitude: instead

of considering as template the curve in (4) we take as template

$$2 * \sin(s) - 1 * \sin(\frac{s^2}{2\pi})$$

using the same amplitude and phase errors that were sampled for $c_{46}, \ldots, c_{90}$.

CASE C. The 90 curves displayed in case C of Figure 2, $c_1^{[C]}, \ldots, c_{90}^{[C]}$, are obtained as follows:
- for $i = 1, \ldots, 45$, $c_i^{[C]} = c_i^{[A]}$;
- for $i = 46, \ldots, 90$, $c_i^{[C]}$ is obtained from $c_i^{[A]}$ by modifying its phase: instead of considering phase s we take as phase

$$-\frac{1}{3} + \frac{3}{4}s$$

using the same amplitude and phase errors that were sampled for $c_{46}, \ldots, c_{90}$.

CASE D. The 90 curves displayed in case C of Figure 2, $c_1^{[D]}, \ldots, c_{90}^{[D]}$, are obtained as follows:
- for $i = 1, \ldots, 30$, $c_i^{[D]} = c_i^{[A]}$;
- for $i = 31, \ldots, 60$, $c_i^{[D]}$ is obtained as in case B;
- for $i = 61, \ldots, 90$, $c_i^{[D]}$ is obtained as in case C.

3.2 Data analysis with k-means alignment

Cases B and C in Figure 2 display two clusters of curves each. We know that in case B, the two clusters are due to clustering in the amplitude, i.e. to the presence of two groups of curves with distinct shapes; whereas in case C the two clusters are due to clustering in the phase, but all curves, once suitably aligned, belong to the same amplitude cluster. Again, case D in the same figure displays three clusters of curves, but we know that only two amplitude clusters are present, and one of the two has associated a further clustering in the phase. We want our procedure to be able to correctly identify when the clustering is in the amplitude and when it is in the phase, and more generally to be able to efficiently separate amplitude variability and phase variability.

Figure 4, case A, shows the aligned curves and warping functions resulting from 1-mean alignment of curves A. Figure 3 shows the boxplot of the similarity indexes between the original A curves and their mean estimated by Loess ("A, orig"), and the boxplots of the similarity indexes between the k-means aligned curves and their estimated templates, for $k = 1, 2, 3$ ("$A, k = 1$", "$A, k = 2$" and "$A, k = 3$" respectively). Note that the 1-mean alignment procedure leads to a significant increase of the similarity indexes, with respect to the similarities of the original curves, leaving not much scope for further improvement when k is set equal to 2 or 3. Thus, the procedure correctly suggest to use $k = 1$ amplitude cluster.

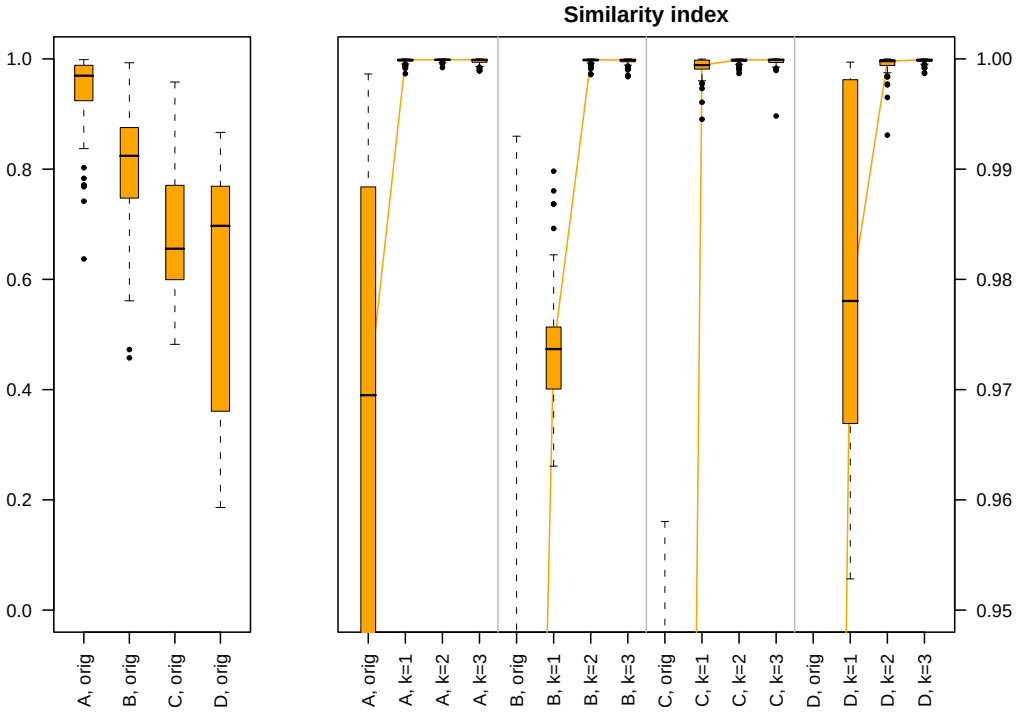

Figure 3: Left: boxplots of similarity indexes between the original curves and their mean curve estimated by Loess (the scale of the plot is from 0 to 1). Right: boxplots of similarity indexes of the original curves (which do not appear in full since the scale of the plot is now from 0.95 to 1) and boxplots of the similarity indexes between the k-means aligned curves and their estimated templates, for $k = 1, 2, 3$ (cases A, B, C and D, respectively); for each case, orange lines link the median of the boxplots corresponding to original curves, 1-mean, 2-means and 3-means aligned curves.

Figure 4, case B, shows the aligned curves and warping functions resulting from 1-mean alignment and 2-means alignment of curves B ("$B, k = 1$" and "$B, k = 2$" respectively). The 1-mean alignment seems to find two clusters in phase, but fails to give a clear picture of the single amplitude cluster that is looked for, since the aligned curves still appear to be separated in two groups. A better picture is instead given by the 2-means alignment, with the 2 amplitude clusters neatly separated and no clustering in phase. Figure 3 shows the similarity indexes of the original curves B and of the k-means aligned curves, for $k = 1, 2, 3$ ("B, orig", "$B, k = 1$", "$B, k = 2$" and "$B, k = 3$" respectively). Note that 1-mean alignment leads to an high increase in the similarities, but a further significant gain can be obtained by setting $k = 2$, whereas an eventual choice of $k = 3$ is not justified by an additional increase in the similarities. Thus, the procedure correctly suggests to use 2 amplitude clusters. We can compare the similarities attained by 2-means alignment of curves B to the ones attained by 1-

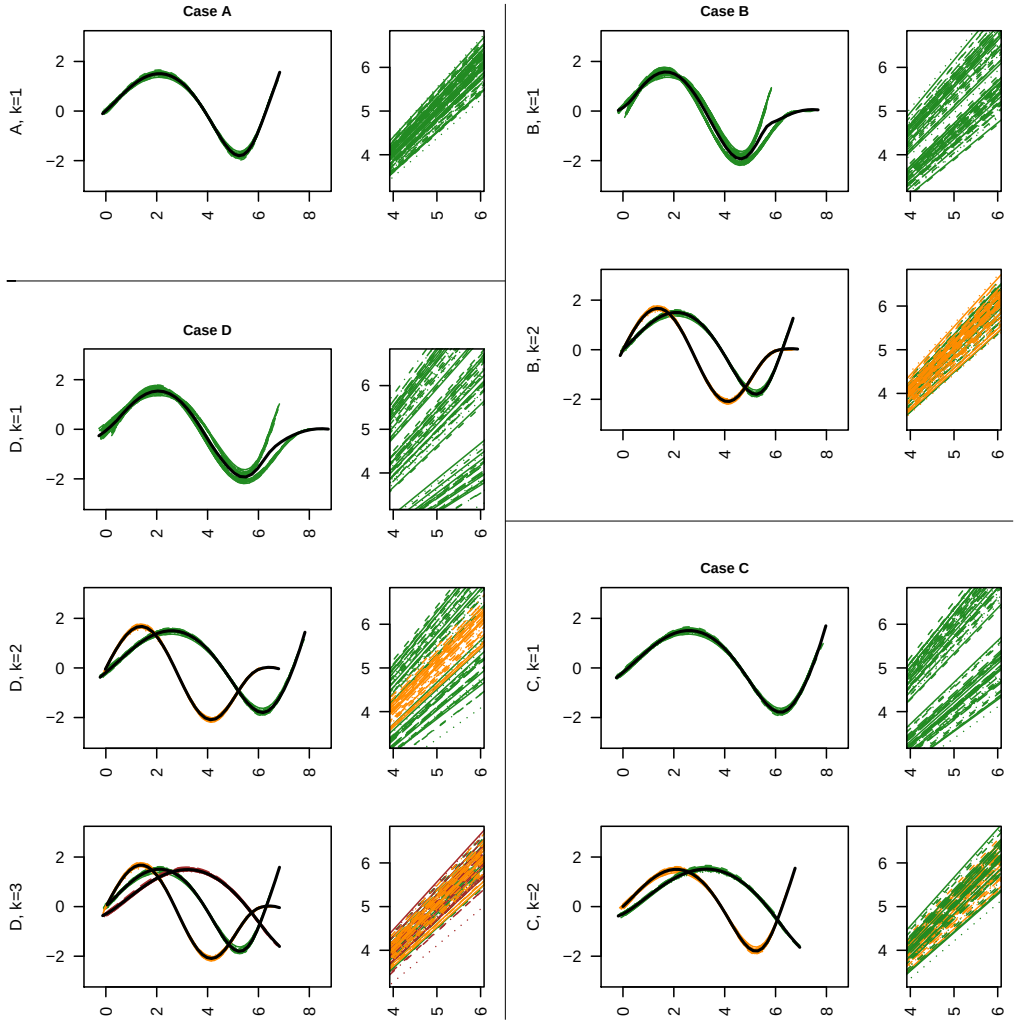

Figure 4: Starting from top left, clockwise, k-means aligned curves and detail of corresponding warping functions for case A ($k=1$), case B ($k=1,2$), case C ($k=1,2$) and case D ($k=1,2,3$); the colors of aligned curves and corresponding warping functions depend on the cluster; superimposed to the aligned curves are the estimated templates (black lines).

mean alignment of curves A. In fact, since half of the curves B coincide with the corresponding curves A, and the other half is obtained from the corresponding curves A by a common modification of their amplitude, one expects that 2-means alignment of curves B should lead to a comparable result, in term of similarities, with respect to 1-mean alignment of curves A. This is confirmed by inspection of the boxplots.

Figure 4, case C, shows the aligned curves and warping functions resulting from 1-mean alignment and 2-means alignment of curves C ("$C, k = 1$" and "$C, k = 2$" respectively). In this case, the 1-mean alignment seems already to give good results, with the curves nicely aligned in one single group, and two clusters evidenced in phase. Also the 2-means alignment gives visually good results, with two neatly separated amplitude clusters and no clustering in phase. But, 2 amplitude clusters can really better capture the similarity of the 90 curves, with respect to just 1? Figure 3 shows the similarity indexes of the original curves C and of the k-means aligned curves, for $k = 1, 2, 3$ ("C, orig", "$C, k = 1$", "$C, k = 2$" and "$C, k = 3$" respectively). Note that the similarities attained with $k = 1$ amplitude cluster are already very high and the use of $k = 2$ amplitude clusters is not paid off by a further reasonable gain in the similarities. Thus, the procedure correctly suggests that $k = 1$ amplitude cluster is sufficient to capture the similarity of the curves; hence, the clustering observed in Figure 3, case C, is due to clustering in the phase, and is captured by the clustering of the warping functions relative to 1-mean alignment of the curves. Note that when we set $k = 2$, the procedure uses the unnecessary second amplitude cluster to explain a clustering that is instead present in the phase space.

Finally, Figure 4, case D, shows the aligned curves and warping functions resulting from 1-mean, 2-means and 3-means alignment of curves D ("$D, k = 1$", "$D, k = 2$" and "$D, k = 3$"). The boxplots of the similarity indexes, shown in Figure 3 ("D, orig", "$D, k = 1$", "$D, k = 2$" and "$D, k = 3$"), correctly suggest to use 2 amplitude clusters. The 2-means alignment procedure efficiently identifies the 2 amplitude clusters and evidences that one of the two clusters (the green one in the picture) has associated a further clustering in the phase. Note that when we set $k = 1$, the procedure tries to explain the clustering of curves D by imputing it to the phase, where it finds three clusters of warping functions, but the procedure fails to give a clear picture of the single amplitude cluster. Whereas, when we set $k = 3$, the procedure uses the unnecessary third amplitude cluster to explain a clustering that is instead present in the phase space, as noticed for case C.

4 An application to the analysis of growth data

In this section we present the results obtained by applying the k-means alignment algorithm to Berkeley Growth Study data, which include the heights (in cm) of 39 boys and 54 girls, measured quarterly from 1 to 2 years, annually from 2 to

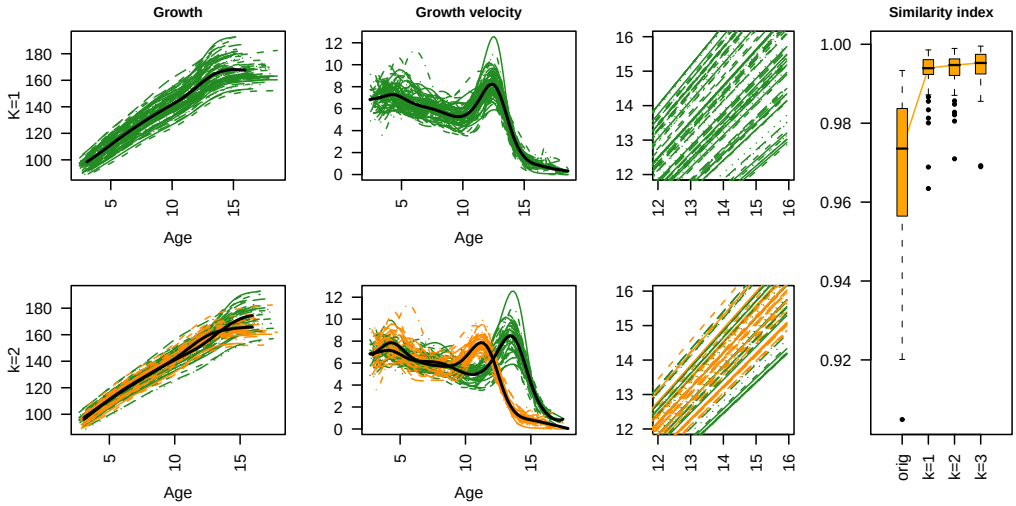

Figure 5: Results of k-means alignment of growth curves, for $k=1,2$: aligned growth curves (with superimposed estimated templates, in black) and corresponding growth velocities (with superimposed first derivatives of estimated templates, in black), together with detail of warping functions and boxplots of similarity indexes (for original curves and k-means aligned curves, $k=1,2,3$).

8 years and biannually from 8 to 18 years. We estimate the growth curves by means of monotonic cubic regression splines (see Ramsay and Silverman [14]), implemented using the R© function `smooth.monotone` available in *fda* package. Figure 5 shows the results obtained by 1-mean and 2-means alignment of these curves. If 2-means alignment is performed, the 2 amplitude clusters discriminate boys from girls, with a misclassification error of only 14%. This fact can be appreciated in Figure 6, which displays growth velocities and warping functions corresponding to $k=2$ (right), colored in blue for boys and pink for girls. But the boxplots of the similarity indexes attained by k-means alignment, for $k=1,2,3$, suggest that the correct number of amplitude clusters to be used is just 1, not 2, since the choice of $k=2$ is not payed off by a reasonable further gain in the similarities (See Figure 5).Thus, the clustering of boys and girls must instead be looked for in the phase space. This clustering is evidenced in Figure 6, were the warping functions corresponding to $k=1$ (left) are displayed in blue for boys and pink for girls, evidencing that boys grow later and more slowly than girls.

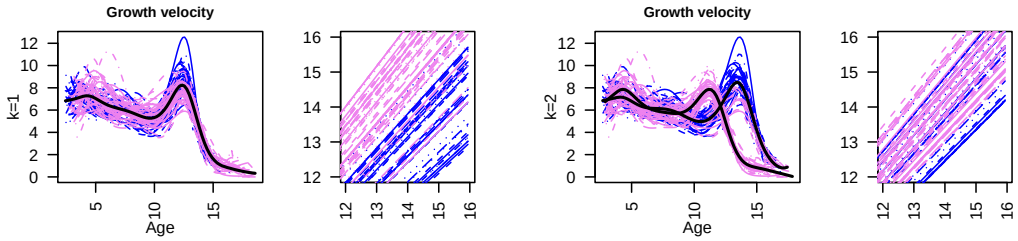

Figure 6: Growth velocities and warping functions corresponding to $k=1$ and $k=2$ (left and right respectively), displayed in blue for boys and pink for girls. The warping functions corresponding to $k=1$ evidence the clustering of boys and girls in the phase space.

5 An application to the analysis of 3D cerebral vascular geometries

Finally, we show the results obtained by applying the k-means alignment algorithm to the AneuRisk dataset[1]. This dataset includes the three spatial coordinates (in mm) of 65 Internal Carotid Artery (ICA) centerlines, measured on a fine grid of points along a curvilinear abscissa, that goes from the terminal bifurcation of the ICA towards the heart. The estimates of these three-dimensional curves are obtained by means of 3D free knot regression splines, described in Sangalli et al. [16]. The first derivatives, x', y', z', of estimated ICA centerlines are displayed in Figure 7, top left. Sangalli et al. [17, 15] presents 1-mean alignment of these curves. Figure 7 shows the first derivatives of 1-mean aligned curves and also the first derivatives of 2-means aligned curves (bottom left and bottom right respectively). Looking at the boxplots of the similarity indexes attained by k-means alignment, it could be argued that the use of $k=2$ amplitude clusters leads in fact to a reasonable further gain in the similarities, whilst no additional improvement is obtained for $k=3$. If $k=2$ is used, the 2 amplitude clusters discriminate between Ω-shaped ICA (green cluster) and S-shaped ICA (orange cluster). This can be appreciated in Figure 8 that gives a 3D image of the estimated templates of the 2 amplitude clusters. The classification of ICA in Ω-shaped and S-shaped, introduced in the medical field and used among others by Krayenbuehl [9], is based on the shape of the terminal part of the ICA (the part inside the red circle in Figure 8): in some cases this has the form of the letter Ω, in others the form of the letter S. Since the shape of the ICA influences

[1] 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 project involves MOX Laboratory for Modeling and Scientific Computing (Dip. di Matematica, Politecnico di Milano), Laboratory of Biological Structures (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.

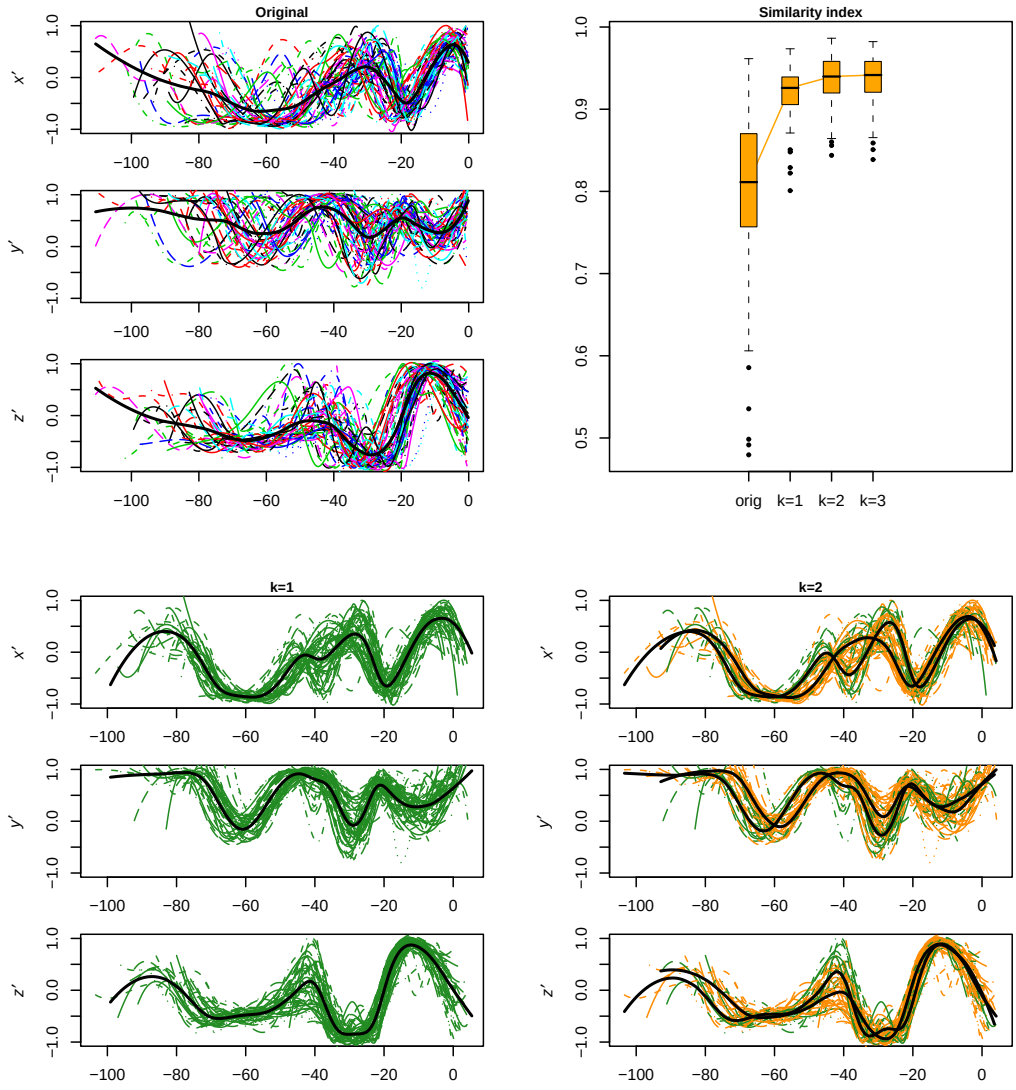

Figure 7: Top left: first derivative of the three estimated spatial coordinates x', y', z' of ICA centerlines. Top right: boxplots of similarity indexes (for original curves and k-means aligned curves, $k = 1, 2, 3$). Bottom: first derivatives of 1-mean and 2-means aligned curves (left and right respectively).

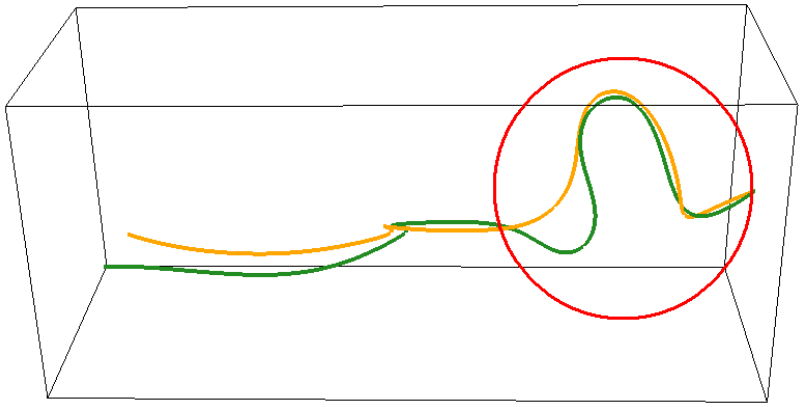

Figure 8: 3D image of the estimated templates of the 2 amplitude clusters, found by 2-means alignment of ICA centerlines. The template of the green cluster is to prototype of an Ω-shaped ICA, whereas the template of the orange cluster is to prototype of an S-shaped ICA.

the pathogenesis of cerebral aneurysms through its effects on the hemodynamics (as discussed in Piccinelli et al. [12], Sangalli et al. [16, 17, 15]) the classification provided by the 2-means alignment of the ICA centerlines could be helpful in the determination of the risk level of a given patient.

6 Discussion

We described the problem of curve clustering when curves are misaligned and proposed a k-means alignment algorithm that jointly clusters and aligns the curves with respect to k unknown templates. We illustrated the power of this procedure via simulation studies and applications to real data.

References

[1] Altman, N. S. and Villarreal, J. C. (2004), "Self-modelling regression for longitudinal data with time-invariant covariates," *Canad. J. Statist.*, 32, 251–268.

[2] Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992), *Local regression models. Chapter 8 of Statistical Models in S*, eds Chambers, J. M. and Hastie, T. J., Wadsworth & Brooks, Pacific Grove, California.

[3] Ferraty, F. and Vieu, P. (2006), "Functional nonparametric statistics in action," in *The art of semiparametrics*, Physica-Verlag/Springer, Heidelberg, Contrib. Statist., pp. 112–129.

[4] Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized additive models*, vol. 43 of *Monographs on Statistics and Applied Probability*, London: Chapman and Hall Ltd.

[5] Heckman, N. E. and Zamar, R. H. (2000), "Comparing the shapes of regression functions," *Biometrika*, 87, 135–144.

[6] James, G. M. (2007), "Curve alignment by moments," *The Annals of Applied Statistics*, 1, 480–501.

[7] Kaziska, D. and Srivastava, A. (2007), "Gait-Based Human Recognition by Classification of Cyclostationary Processes on Nonlinear Shape Manifolds," *Journal of the American Statistical Association*, 102, 1114–1128.

[8] Ke, C. and Wang, Y. (2001), "Semiparametric nonlinear mixed-effects models and their applications," *J. Amer. Statist. Assoc.*, 96, 1272–1298, with comments and a rejoinder by the authors.

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

[10] Lawton, W. H., Sylvestre, E. A., and Maggio, M. S. (1972), "Self Modeling Nonlinear Regression," *Technometrics*, 14, 513–532.

[11] Lindstrom, M. J. and Bates, D. M. (1990), "Nonlinear mixed effects models for repeated measures data," *Biometrics*, 46, 673–687.

[12] Piccinelli, M., Bacigaluppi, S., Boccardi, E., Ene-Iordache, B., Remuzzi, E., Veneziani, A., and Antiga, L. (2007), "Influence of internal carotid artery geometry on aneurism location and orientation: a computational geometry study," Available at www.mathcs.emory.edu.

[13] Ramsay, J. O. and Li, X. (1998), "Curve registration," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60, 351–363.

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

[15] Sangalli, L. M., Secchi, P., and Vantini, S. (2008a), "Explorative Functional data analisys for 3D-geometries of the Inner Carotid Artery," in *Functional and Operatorial Statistics*, ed. Dabo-Niang, Sophie; Ferraty, F., Springer, Contributions to Statistics, pp. 289–296.

[16] Sangalli, L. M., Secchi, P., Vantini, S., and Veneziani, A. (2007), "Efficient estimation of 3-dimensional centerlines of inner carotid arteries and their curvature functions by free knot regression splines," Tech. Rep. 23/2007, MOX, Dipartimento di Matematica, Politecnico di Milano.

[17] — (2008b), "A Case Study in Explorative Functional Data Analysis: Geometrical Features of the Internal Carotid Artery," *J. Amer. Statist. Assoc.*, to appear.

[18] Sheehy, A., Gasser, T., Molinari, L., and Largo, R. H. (2000a), "An analysis of variance of the pubertal and midgrowth spurts for length and width," *Annals of Human Biology*, 26, 309–331.

[19] — (2000b), "Contribution of growth phases to adult size," *Annals of Human Biology*, 27, 281–298.

[20] Shimizu, N. and Mizuta, M. (2007), "Functional Clustering and Functional Principal Points," in *Knowledge-Based Intelligent Information and Engineering Systems 11th International Conference, KES 2007, XVII Italian Workshop on Neural Networks, Vietri sul Mare, Italy, September 12-14, 2007. Proceedings, Part III*, eds. Apolloni, B., Howlett, R. J., and Jain, L. C., Berlin Heidelberg: Springer-Verlag, Lecture Notes in Artificial Intelligence 4693, pp. 501–508.

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

[22] Tuddenham, R. D. and Snyder, M. M. (1954), "Physical growth of California boys and girls from birth to age 18," Tech. Rep. 1, University of California Publications in Child Development.

MOX Technical Reports, last issues

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

13/2008 L.M. Sangalli, P. Secchi, S. Vantini, V. Vitelli:
K-means alignment for curve clustering

12/2008 T. Passerini, M.R. de Luca, L. Formaggia, A. Quarteroni, A. Veneziani:
A 3D/1D geometrical multiscale model of cerebral vasculature

11/2008 L. Gerardo Giorda, L. Mirabella, F. Nobile, M. Perego, A. Veneziani:
A model preconditioner for the Bidomain problem in electrocardiology

10/2008 N. Grieco, E. Corrada, G. Sesana, G. Fontana, F. Lombardi, F. Ieva, A.M. Paganoni, M. Marzegalli:
Predictors of the reduction of treatment time for ST-segment elevation myocardial infarction in a complex urban reality. The MoMi^2 survey

9/2008 P. Secchi, E. Zio, F. Di Maio:
Quantifying Uncertainties in the Estimation of Safety Parameters by Using Bootstrapped Artificial Neural Networks

8/2008 S. Micheletti, S. Perotto:
Space-time adaptation for purely diffusion problems in an anisotropic framework

7/2008 C. Vergara, R. Ponzini, A. Veneziani, A. Redaelli, D. Neglia, O. Parodi:
Reliable CFD-based Estimation of Flow Rate in Hemodynamics Measures. Part II: Sensitivity Analysis and First Clinical Application

6/2008 E. Fumagalli, L. Lo Schiavo, A.M. Paganoni, P. Secchi:
Statistical analyses of exceptional events: the Italian experience

5/2008 S. Badia, A. Quaini, A. Quarteroni:
Modular vs. non-modular preconditioners for fluid-structure systems with large added-mass effect

4/2008 R. Milani, A. Quarteroni, G. Rozza:
Reduced basis method for linear elasticity problems with many parameters