Electronic Journal of Statistics
Vol. 8 (2014) 1817–1824
ISSN: 1935-7524
DOI: 10.1214/14-EJS937A

Analysis of juggling data: An application of k-mean alignment*

Mara Bernardi, 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: marasabina.bernardi@mail.polimi.it; laura.sangalli@polimi.it; piercesare.secchi@polimi.it; simone.vantini@polimi.it

Abstract: We analyze the juggling data by means of the k-mean alignment algorithm using cycles as the experimental units of the analysis. Allowing for affine warping, we detect two clusters distinguishing between mainly-planar trajectories and trajectories tilted toward the body of the juggler in the lower part of the cycle. In particular we detect an anomalous presence of tilted trajectories among the trial third cycles. We also find warping functions to be clustered according to trials suggesting that each trial is performed at a different pace and thus associated to a different typical cycle-duration.

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

Received August 2013.

1. Pre-processing

We analyzed the juggling data described in Ramsay, Gribble and Kurtek (2014a) from two perspectives: trials as experimental units of the analysis and cycles – composing each trial – as experimental units of the analysis. In this manuscript we focus on the second approach and thus we deal with 113 three-dimensional curves indicating different trajectories of the juggler's right hand forefinger. We arbitrarily indicate as *cycle* the period between two subsequent releases of a ball and, for biological reasons, we identify the moment when a ball is released with the moment when the tangential acceleration is maximal. These moments identify the end of a cycle and the beginning of the following one. As an example, in Figure 1 we report the evolution across time of the tangential acceleration for the first trial. Colored stars represent stationary points. In particular, blue stars represent the ones used to cut the trial into cycles.

This procedure was applied to all ten trials thus obtaining a set of 113 cycles. As suggested in Ramsay, Gribble and Kurtek (2014a), the origin of each cycle has been set at the time when the tangential acceleration is maximal, i.e., the starting points of the 113 curves. Note that all cycles present different durations

*Main article 10.1214/14-EJS937.
†Corresponding author.

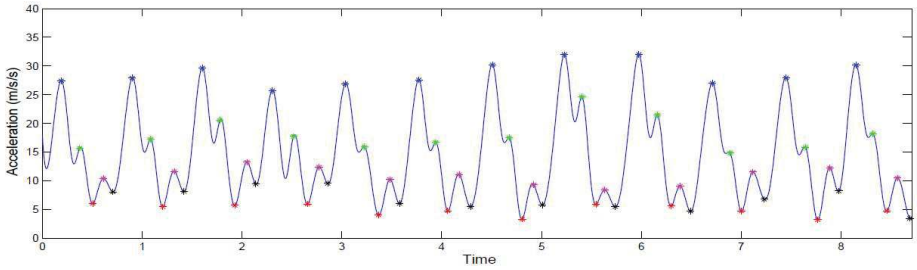

FIG 1. *Tangential acceleration of the first trial.*

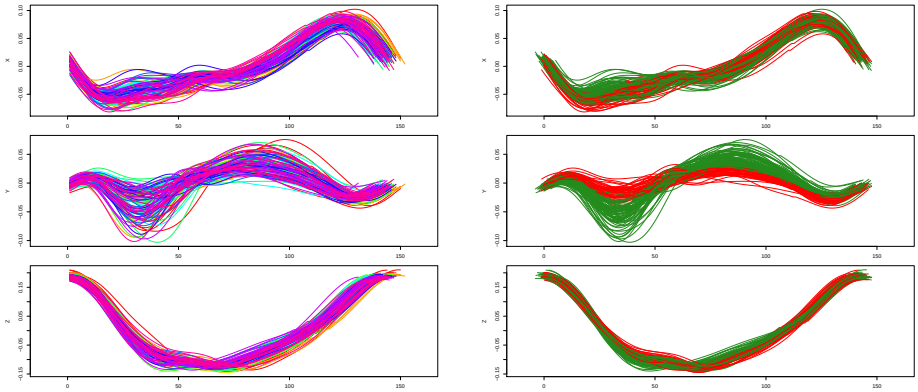

FIG 2. *Left panels: the 113 cycles. The three panels show the X, Y, and Z components of the curves, respectively. Right panels: Aligned and clustered cycles provided by the k-mean alignment algorithm performed with affine warping functions and* $k = 2$. *The three panels show the X, Y, and Z components of the aligned cycles. The cycles of the first cluster are in green, while those of the second cluster are in red.*

and thus the final time instants differ across cycles. In the left panels of Figure 2 the X, Y, and Z components of the 113 cycles are reported.

2. *k*-mean alignment

To look for clusters of trajectories in presence of phase variability we applied the k-mean alignment algorithm, detailed in Sangalli et al. (2010) and summarized in Sangalli, Secchi and Vantini (2014), to the 113 trajectories. Since data pre-processing centered and partially rotated data but did not rescaled them, we will consider two cycles as similar if they are identical up to a multiplicative factor along each component. Therefore, we shall use the following similarity index:

$$\rho(f_i, f_j) = \frac{1}{3}\sum_{p=1}^{3} \frac{\int f_{ip}(t) f_{jp}(t) dt}{\sqrt{\int f_{ip}(t)^2 dt}\sqrt{\int f_{jp}(t)^2 dt}}, \tag{2.1}$$

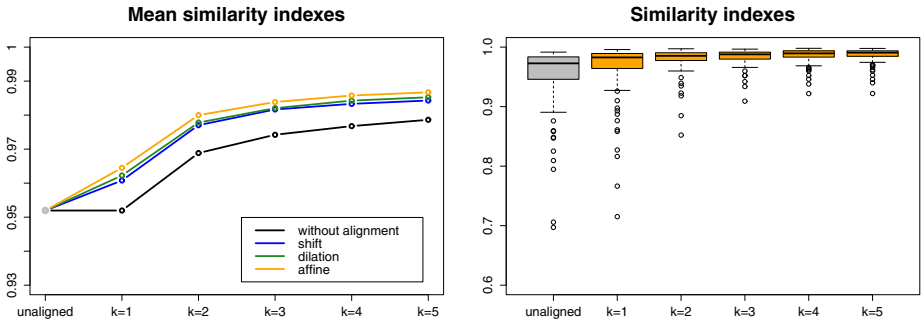

FIG 3. *Performance of the k-mean alignment algorithm. The left panel shows the mean similarity between the aligned curves and their respective templates obtained with different values of k and different classes of warping functions. The right panel displays the boxplots of the similarity indexes between the aligned curves and their respective templates with the group of warping functions* $\mathcal{H}_{affine}$ *and different values of k.*

where f_i and f_j are cycles. Indeed, this similarity index assigns similarity equal to 1 (its maximal value) to couples of curves that differ only for a positive multiplying factor along each component:

$$\rho(f_i, f_j) = 1 \;\Leftrightarrow\; \forall p \in \{1,2,3\}\; \exists a_p \in \mathbb{R}^+ : f_{ip}(t) = a_p f_{jp}(t). \tag{2.2}$$

Since the physical phenomenon does not suggest any particular group of warping functions to be the best suited to align the cycles, we run the analysis using different groups of warping functions coherent with the previous similarity index:

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

The analysis here presented has been performed using `fdakma` R package downloadable from CRAN (Parodi et al. (2014)).

The left panel of Figure 3 shows the results of the k-mean alignment algorithm applied with different choices for the number k of clusters and the group $\mathcal{H}$ of warping functions. For each couple $(k, \mathcal{H})$ the mean similarity between the aligned curves and their respective templates is reported. The gray dot on the left represents the mean similarity between the unaligned curves and their mean which acts as a lower bound for the algorithm performance. The mean similarities achieved by using $\mathcal{H}_{\text{affine}}$, $\mathcal{H}_{\text{shift}}$, $\mathcal{H}_{\text{dilation}}$, and $\mathcal{H}_{\text{identity}}$ are reported in orange, blue, green, and black, respectively. Note that, as already pointed out in Sangalli et al. (2010) and in Sangalli, Secchi and Vantini (2014), running the k-mean alignment without allowing for warping (i.e, choosing $\mathcal{H}_{\text{identity}}$) is equivalent to perform a simple functional k-mean clustering, while setting $k = 1$ is equivalent to perform a simple continuous alignment with just one template.

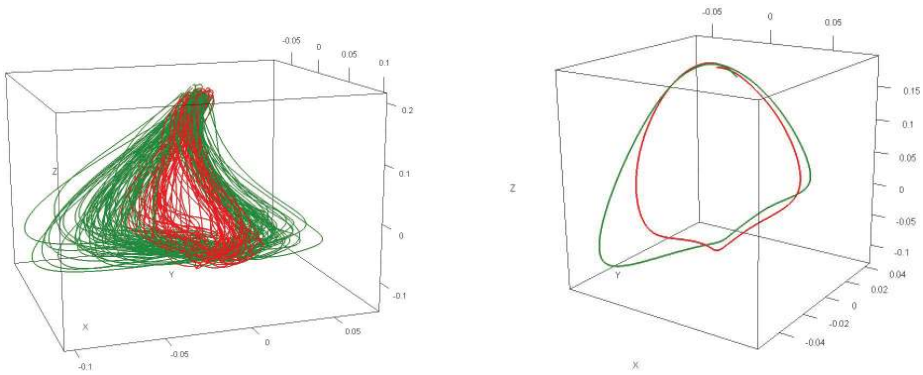

FIG 4. *Left panel: 3D plot of the cycles colored according to the clusters provided by the k-mean alignment algorithm performed with affine warping functions and* $k = 2$. *The cycles of the first cluster are in green, while those of the second cluster are in red. Right panel: the corresponding template curves.*

As described in Sangalli et al. (2010) and in Sangalli, Secchi and Vantini (2014), being the curves not defined on the entire real axis, the integrals in (2.1) are computed over the intersection of the domains of f_i and f_j, and the cluster templates are estimated by means of local polynomial regression.

The vertical displacement between the black curve and the others suggests the presence of phase variability. In detail, the slightly higher performances achieved using the group of affinity (i.e., orange curve) seem to suggest this latter group as the most suitable to align cycles. This choice is consistent with a possible wrong detection of the starting point of the cycle and a possible different velocity across cycles. The right panel of Figure 3 details the orange curve by showing the boxplots of the similarities between each aligned curve and its respective template. Focussing on the orange curve, the elbow observed for $k = 2$ suggests the existence of two clusters. No significant gain in the mean similarity is indeed obtained by introducing an extra cluster (i.e., $k = 3$). We thus now focus on the clustering and alignment obtained by setting the number of clusters equal to 2 (i.e., $k = 2$) and by considering positive affinities (i.e., $\mathcal{H}_{\text{affine}}$) as the group of warping functions. The three right panels of Figure 2 show the X, Y, and Z components of the aligned cycles. The cycles of the first cluster are colored in green, while those of the second cluster in red.

3. Analysis of the clusters

Differences between the two clusters are observed in all three components, though the major ones pertain to the front-back component (i.e., Y-axis) where the trajectories of the first cluster (i.e, green) present a more oscillating behavior. As it is even clearer in the left panel of Figure 4, the trajectories of the second cluster (i.e, red) are mainly planar curves in the X-Z plane, while the trajectories of the first cluster are tilted toward the body of the juggler in the

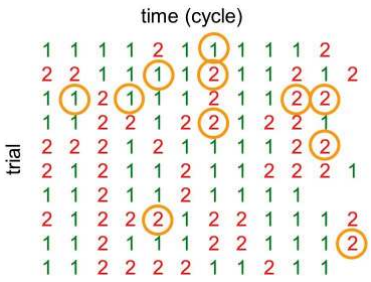

FIG 5. *Clustering assignments across trials (i.e., rows) and positions in the sequence (i.e., columns). Mismatches between our classification and the one obtained by Lu and Marron (2014) are pointed out in orange.*

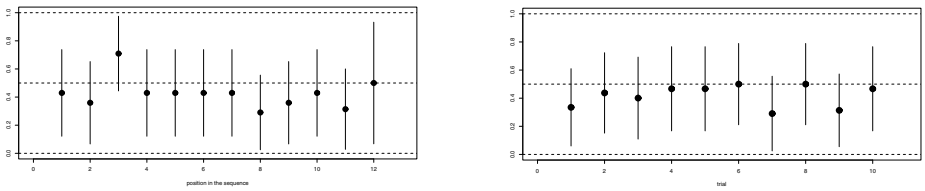

FIG 6. *Confidence (95%) intervals for the proportion of cycles belonging to the second clusters across trials (right panel) and across position in the sequence (left panel).*

lower part of the cycle. The same differences are captured by the two template curves reported in the right panel of Figure 4.

We now want to test if there is any relation between the cluster each cycle belongs to and the trial it belongs to and/or its position in the cycle sequence. Figure 5 shows cluster assignment for each cycle. Rows are associated to trials and columns to positions in the sequence. In Figure 6 we report confidence intervals for the proportion of cycles belonging to the second cluster across trials (right panel) and across positions in the sequence (left panel). We found no statistical differences among the proportions across trials (i.e., curves of each clusters seem randomly spread across trials), while we found a higher proportion of curves belonging to the second cluster among the third cycles of each trial (i.e., third column in Figure 5). Note that in the third cycle for the first time the juggler needs to catch and then throw a ball with his right hand. Indeed, the first two cycles are warm-up cycles in which no ball is caught, and only balls previously handled by the juggler are thrown.

4. Analysis of the warping functions

We now focus on the analysis of warping functions and in particular on possible associations between warping functions and trials or warping functions and position in the sequence. Analyzing the mean shift and dilation across trials and positions in the sequence (Figure 7) we found an opposite scenario with respect to the one illustrated in the previous section. Indeed we found no association

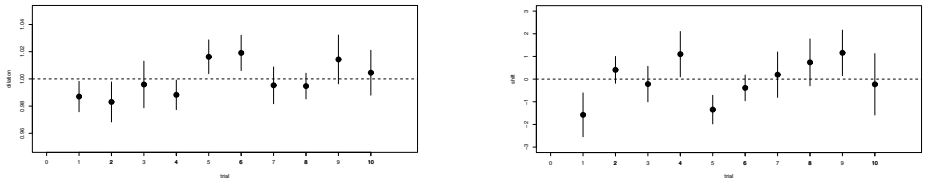

FIG 7. *Confidence (95%) intervals for the mean of the dilations (left panel) and the shifts (right panel) of the warping functions across trials.*

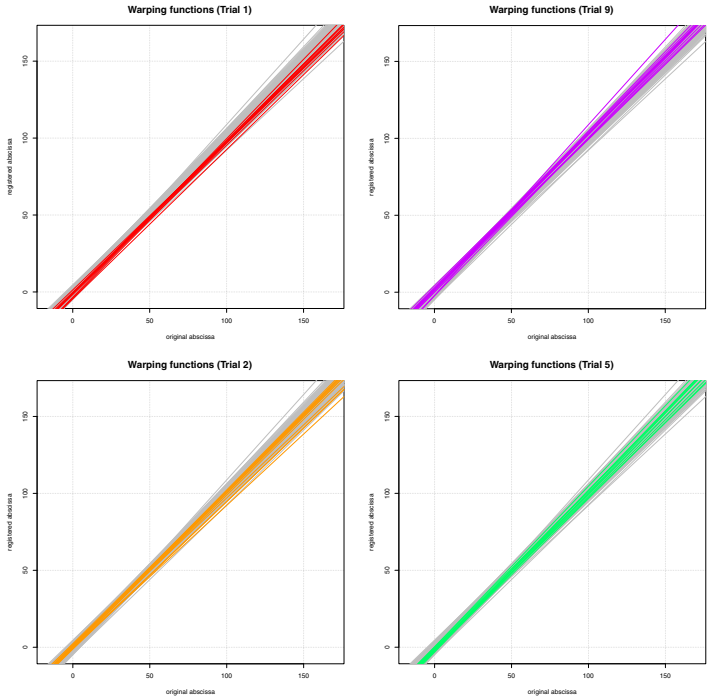

FIG 8. *Warping function associated to cycles belonging to trials 1, 9, 2, and 5, respectively.*

between warping functions and position in the sequence, while we observed significant differences across trials. As an example, in Figure 8 we point out with different colors over the gray background the warping function associated to cycles belonging to trials 1, 9, 2, and 5, respectively. Indeed, all warping functions of trial 1 have negative intercepts, while those of trial 9 have positive intercepts: the cycles of trial 1 have been anticipated to be aligned to the other ones, while those of trial 9 have been delayed. All warping functions of trial 2 have slopes lower than 1, while those of trial 5 have slopes greater than one: the cycles of trial 2 have been sped up to be aligned, while those of trial 5 have been slowed down. Our analysis suggests that each trial has been performed at a different pace (i.e., cycle durations are homogeneous within trials but not across trials).

5. Discussion

Our findings pertaining to the amplitude variability agree with Lu and Marron (2014), Kurtek, Xie and Srivastava (2014), and Poss and Wagner (2014). Our clustering structure is indeed very similar to the one detected in Lu and Marron (2014): only 20 cycles among the 113 have been assigned to different clusters by the two analyses. In Figure 5 the mismatches are pointed out by means of orange circles. The differences between the two clusters shown in the right panels of Figures 2 and 4 are also the same gathered by the first functional principal component detected in both Kurtek, Xie and Srivastava (2014) and Poss and Wagner (2014). Moreover, Poss and Wagner (2014) look for possible association between scores and the position in the sequence. Since their first principal component seems to be related to our clustering structure, it would be interesting to know if they find an anomaly in the first principal component score distribution in correspondence of third cycles analogous to the one we find for the cluster assignment.

On the contrary, with respect to phase variability, our results disagree with the ones reported in Kurtek, Xie and Srivastava (2014) and Ramsay, Gribble and Kurtek (2014b). In details, Kurtek, Xie and Srivastava (2014) point out a "compensation" effect within each cycle, suggesting that cycles that started faster then slowed down or cycles that started slower then sped up. Possibly, because of the affine warping we used, we did not detect any similar effect. Finally, our analysis of the warping functions suggests that cycle durations are homogeneous within trials but not across trials. This contrasts with the "return to the base frequency" effect reported in Ramsay, Gribble and Kurtek (2014b) where early cycles at the beginning of each trial are found to be followed by late cycles at the end of the trial.

Acknowledgements

All authors are grateful to MBI Mathematical Biosciences Institute http://mbi.osu.edu/ 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 SNAPLE: Statistical and Numerical methods for the Analysis of Problems in Life sciences and Engineering http://mox.polimi.it/users/sangalli/firbSNAPLE.html.

References

Kurtek, S., Xie, Q. and Srivastava, A. (2014). Analysis of juggling data: Alignment, extraction, and modeling of juggling cycles. *Electronic Journal of Statistics* **8** 1865–1873, Special Section on Statistics of Time Warpings and Phase Variations.

Lu, X. and Marron, J. S. (2014). Analysis of juggling data: Object oriented data analysis of clustering in acceleration functions. *Electronic Journal of Statistics* **8** 1842–1847, Special Section on Statistics of Time Warpings and Phase Variations.

Parodi, A., Patriarca, M., Sangalli, L., Secchi, P., Vantini, S. and Vitelli, V. (2014). fdakma: Functional data analysis: k-mean alignment, R package version 1.1.1.

Poss, D. and Wagner, H. (2014). Analysis of juggling data: Registering data to principal components to explain amplitude variation. *Electronic Journal of Statistics* **8** 1825–1834, Special Section on Statistics of Time Warpings and Phase Variations.

Ramsay, J. O., Gribble, P. and Kurtek, S. (2014a). Description and processing of functional data arising from juggling trajectories. *Electronic Journal of Statistics* **8** 1811–1816, Special Section on Statistics of Time Warpings and Phase Variations.

Ramsay, J., Gribble, P. and Kurtek, S. (2014b). Analysis of juggling data: Landmark and continuous registration of juggling trajectories. *Electronic Journal of Statistics* **8** 1835–1841, Special Section on Statistics of Time Warpings and Phase Variations.

Sangalli, L. M., Secchi, P. and Vantini, S. (2014). Analysis of AneuRisk65 data: k-mean alignment. *Electronic Journal of Statistics* **8** 1891–1904, Special Section on Statistics of Time Warpings and Phase Variations.

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