Electronic Journal of Statistics
Vol. 8 (2014) 1842–1847
ISSN: 1935-7524
DOI: 10.1214/14-EJS937D

Analysis of juggling data: Object oriented data analysis of clustering in acceleration functions*

Xiaosun Lu and J. S. Marron

Department of Statistics and Operations Research
University of North Carolina at Chapel Hill NC 27599 USA
e-mail: xiaosun@live.unc.edu; marron@unc.edu

Abstract: This paper describes an analysis of acceleration variability among the juggling cycles. The Fisher Rao curve registration is used for curve alignment. Five different choices of data objects are considered in this paper. We show that one of these choices of data objects leads to a much better clustering into two distinct types of juggling cycles than the other choices.

Keywords and phrases: Curve registration, distance weighted discrimination, clustering, functional data analysis.

Received August 2013.

This paper studies acceleration variability among the juggling cycles from Ramsay et al. (2014). These cycles are treated as being independent of each other. Object Oriented Data Analysis turns out to be very useful terminology throughout the discussion, where *data objects* are understood as the atoms of the statistical analysis. This concept was first brought up by Wang and Marron (2007). Section 1 introduces five potential choices of data objects. Section 2 shows that one of these choices leads to a discovery of two distinct types of juggling cycles, which are hard to detect using the other choices.

1. Data objects

An intuitive choice of data objects is the acceleration curves of each juggling cycle. Figure 1 shows a color coded view of the data preprocessing done to obtain these curves. The top panel shows the acceleration curve of one trial. Each cycle is defined as the period between two neighboring highest peaks indicated by the vertical dashed lines. This is a valid definition, since a careful examination of the other juggling trials confirms that these peaks are always a prominent feature across all cycles. The incomplete cycle fragments at both the beginning and the end are ignored in the following analysis. The resulting acceleration curves of each juggling cycle in this trial are shown in the bottom left panel, using the same colors. Note that these curves are defined on different domains, that is, the

*Main article 10.1214/14-EJS937.

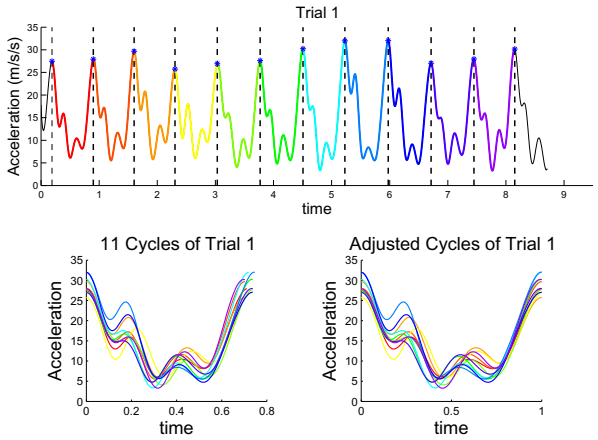

FIG 1. *Trial 1, used to illustrate data preprocessing. Top: Acceleration curve of one trial, where cycles are divided by the vertical dashed lines, color coded by cycle number. Bottom left: The acceleration curves of each cycle, obtained by cutting the above curve at the dashed lines. The incomplete cycles at the two ends are ignored. Bottom right: The adjusted curves defined on the same domain, obtained by linearly warping the domain of each curve in the left to* $[0, 1]$.

duration of each cycle is different. We studied potential relationships between cycle lengths and other aspects of these curves, but do not report details here, because no interesting relationship was found. Next, the domain of each cycle is linearly warped to $[0, 1]$, and a cubic spline interpolation is used to define these curves on a common grid. See the bottom right panel for the resulting curves.

The left panel in Figure 2 shows the adjusted acceleration curves of all 10 trials, which contain both *horizontal* (i.e. phase or tempo) variation and *vertical* (or amplitude) variation. These two types of variation can be separated via the Fisher Rao curve registration proposed by Srivastava et al. (2011), and captured by the resulting domain warping functions and the aligned functions, respectively, shown in the right two panels in Figure 2. Note that the left side of the first bumps in the aligned functions shows much steeper increase than that

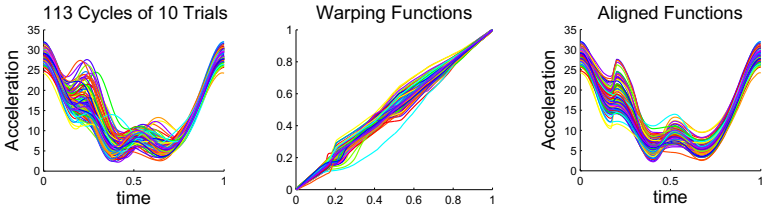

FIG 2. *Fisher Rao alignment of acceleration functions. Left: Unaligned (preprocessed) acceleration functions of all 113 cycles from 10 trials. Middle: Warping functions obtained from Fisher Rao Alignment. Right: Aligned functions. The color indicates the order of cycles, as in Figure 1.*

of the unaligned functions. This is because these cycles consist of two groups with distinct acceleration features at this bump (see the top panels in Figure 4), which is discussed later in Section 2.

This domain warping approach provides two types of data objects: warping functions for studying the horizontal variability, and aligned functions for studying the vertical variability. Meanwhile, the Square-Root Slope Function (SRSF) representation (see Srivastava et al. (2011) for details), i.e. analyzing functions under the Fisher Rao metric, entails considering another two potential choices of data objects: the *horizontal SRSFs*, i.e. SRSFs of warping functions, and *the vertical SRSFs*, i.e. SRSFs of aligned functions. See Lu (2013) and Lu and Marron (2013) for more discussion of these two types of data objects. In summary, the following five different choices of data objects are considered:

(1) Aligned functions;
(2) Warping functions;
(3) Unaligned functions;
(4) Vertical SRSFs;
(5) Horizontal SRSFs.

2. Detection of two types of cycles

Functional Principal Component Analysis (FPCA) was performed on the data objects (1)–(4), respectively. Note that the horizontal SRSFs lie on the surface of a high dimensional sphere (Srivastava et al. (2011)). Two manifold approaches were used separately, Principal Geodesic Analysis (PGA, Fletcher et al. (2004)) and Principal Nested Spheres (PNS, Lu and Marron (2013)), which are two different extensions of FPCA for data lying on curved manifolds.

PGA approximates the spherical surface by a tangent hyperplane centered at the Karcher mean. The data projections on this tangent hyperplane are then analyzed using PCA. In this way, PGA finds the great spheres (i. e. principal geodesics) passing through the mean that best fit the data. In contrast, the PNS method uses a backward approach which starts with the high dimensional sphere and finds the best fitting subsphere of one dimension lower at each step (See Marron et al. (2010)). As a result, unlike PGA, PNS finds the best fitting subspheres regardless of whether they are great spheres or not, with no Karcher mean constraint. It has been shown in a number of cases that PNS can provide more effective analysis of manifold data than many other analogous approaches. See Pizer et al. (2013) for such an example in the study of 3D shapes.

The scores scatterplots of the first two components from each analysis are shown in Figure 3. An interesting data pattern of two clear clusters of cycles is captured by the first principal geodesic of the horizontal SRSFs, shown in Panel $(2, 2)$. These clusters are studied by brushing them with colors and symbols. Clear comparison between choices of data objects comes from applying the same colors and symbols in each panel. These score scatterplots show successive improvement in the detection of the two groups. This visual observation is confirmed by the SWISS scores, i.e. Standardized WithIn class Sum of Squares

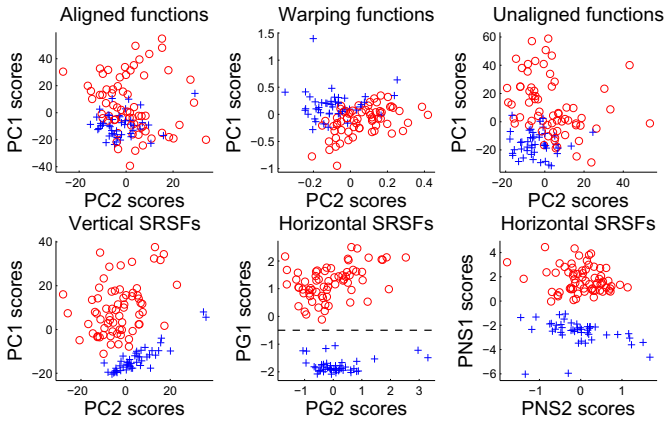

FIG 3. *Score scatterplots of the first two components obtained by performing FPCA on the data objects (1)–(4), and PGA and PNS on (5), respectively. Panel* $(2, 2)$ *indicates two clear clusters, separated by the dashed line. These clusters are highlighted using colors and symbols, which are also used in each other panel. This highlights successive improvement of clustering of the two groups over different data objects.*

(see Cabanski et al. (2010)), which are calculated as the ratio between the total within class sum of squares and the total sum of squares. It reflects the proportion of variation unexplained by clustering. Table 1 shows the corresponding SWISS scores. Smaller values indicate better clustering. On the other hand, DiProPerm (Direction Projection Permutation) t-tests based on the DWD directions (See Wei et al. (2013) for details) were used to test the mean difference between these two groups. The Z-scores of the statistics with respect to the null population (with no group difference) were used to compare the results of these DiProPerm tests. A higher Z-score indicates a bigger difference between two groups. Considering both the clustering and the mean difference, the horizontal SRSFs are shown to be a better choice of data objects for detection of these two groups than the other four, with the lowest SWISS scores and the highest DiProPerm Z-scores (either analyzed by PGA or PNS). All these five choices of data objects lead to significant DiProPerm p-values in both tests.

TABLE 1

Comparison of different choices of data objects with respect to detecting the two groups of cycles, based on the SWISS score and the DiProPerm t-test. Those data objects are in the order of their SWISS scores. It is seen that the horizontal SRSFs lead to the lowest SWISS scores (i.e. the best clustering of the two groups) and the highest DiProPerm Z-scores (i.e. the most significant mean difference)

	Data Objects	Analysis	SWISS	DiProPerm Z-scores
(1)	Aligned functions	PCA	0.89	8.90
(2)	Warping functions	PCA	0.78	28.65
(3)	Unaligned functions	PCA	0.77	13.30
(4)	Vertical SRSFs	PCA	0.67	21.72
(5)	Horizontal SRSFs	PGA	0.52	53.42
		PNS	0.37	34.50

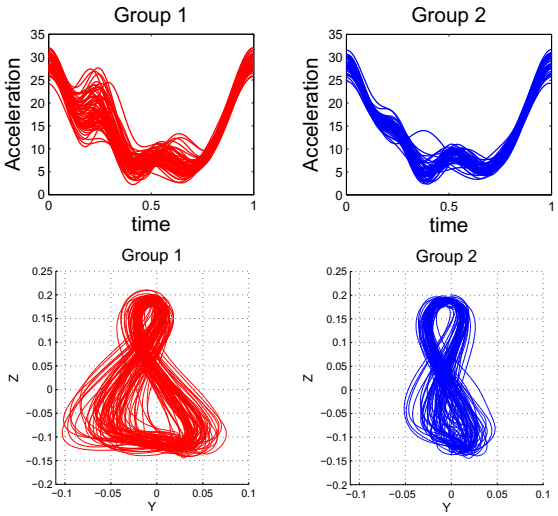

FIG 4. *Top: Adjusted acceleration of the two clusters of cycles, respectively. Bottom: Side view of the juggling trajectories of these two clusters of cycles. The color code is the same as that in Figure 3.*

It is seen from Panel $(2, 2)$ and Panel $(2, 3)$ that both PGA and PNS of the horizontal SRSFs do a good job in separating the two groups. Table 1 shows that the PNS leads to better data clustering (i.e. lower SWISS scores), while the PGA leads to a more significant mean difference (i.e. higher DiProPerm Z-scores). Further analysis shows that the first two components from PGA totally explain 69.8% of data variability (the first component explains 52.7%), while the first two PNS's explain 82.2% of data variability (the first PNS explains 77.4%). That is, the PNS approach is more efficient in capturing data variability and gives better signal compression.

Although no obvious pattern can be found in the distribution of these two types of juggling cycles in the 10 trials, a simple runs test shows that the order of these groups is not random. The acceleration curves of these two types of cycles have distinguishable shapes, shown in the top panels in Figure 4. The main difference is whether the first acceleration bump (at about 0.25 on the time axis) is prominent or not. It is challenging to align a mixture of these two different types of curves simultaneously. In the Fisher Rao alignment shown in Figure 2, the shape of the first bumps in the aligned functions (right) is hard to interpret, compared with the unaligned functions (left). This suggests that it may be better to align these two types of curves separately as proposed by Sangalli et al. (2010). The bottom panels in Figure 4 display a side view of the juggling trajectories of these two types of cycles, respectively. It is seen that the juggler moved his hand back and forth wider in one group of cycles (left) than in the other (right).

In conclusion, we show that among all the five choices of data objects the horizontal SRSFs are the best to study these two different types of juggling cycles,

and the PNS approach leads to the best data clustering. It is seen throughout the paper that the terminology of object oriented data analysis is very helpful in discussion.

Acknowledgements

The authors are grateful to the Mathematical Biosciences Institute for hosting the important meeting which led to this work.

References

Cabanski, C. R., Qi, Y., Yin, X., Bair, E., Hayward, M. C., Fan, C., Li, J., Wilkerson, M. D., Marron, J. S., Perou, C. M., and Hayes, D. N. (2010). Swiss made: Standardized within class sum of squares to evaluate methodologies and dataset elements. *PLoS ONE*, 5:3.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Medical Imaging*, 23:995–1005.

Lu, X. (2013). Object oriented data analysis of cell images and analysis of elastic functions. *Ph.D. Dissertation*, University of North Carolina, Chapel Hill, NC, USA. MR3153472

Lu, X. and Marron, J. S. (2013). Principal nested spheres for time warped functional data analysis. arXiv:1304.6789.

Marron, J. S., Jung, S., and Dryden, I. L. (2010). Speculation on the generality of the backward stepwise view of pca. Proceedings of MIR 2010: 11th ACM SIGMM International Conference on Multimedia Information Retrieval, Association for Computing Machinery, Inc., Danvers, MA, 227–230.

Pizer, S. M., Jung, S., Goswami, D., Vicory, J., Zhao, X., Chaudhuri, R., Damon, J. N., Huckemann, S., and Marron, J. (2013). Nested sphere statistics of skeletal models. In *Innovations for Shape Analysis*, pages 93–115. Springer. MR3075829

Ramsay, J. O., Gribble, P., and Kurtek, S. (2014). Description and processing of functional data arising from juggling trajectories. *Electron. J. Statist.*, 8:1811–1816, 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 & Data Analysis*, 54(5):1219–1233. MR2600827

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011). Statistical analysis and modeling of elastic functions. arXiv:1103.3817.

Wang, H. and Marron, J. S. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873. MR2363955

Wei, S., Lee, C., Wichers, L., Li, G., and Marron, J. S. (2013). Direction-projection-permutation for high dimensional hypothesis tests. arXiv:1304.0796.