A Novel 2-step Iterative Approach for Clustering Functional Data

ZUZANA ROŠŤÁKOVÁ, ROMAN ROSIPAL
INSTITUTE OF MEASUREMENT SCIENCE, SLOVAK ACADEMY OF SCIENCES, SLOVAKIA
e–mail: zuzana.rostakova@savba.sk

INTRODUCTION

A frequent task in functional data analysis is to divide a set of curves $X_1, \dots, X_N$ defined over a common time interval T into K clusters. Here, we address a problem in which classical functional data clustering techniques may fail because curves misalignment is present. We propose and validate a novel 2-step approach, which iteratively combines clustering using the Dynamic Time Warping algorithm [1] with the registration (or time alignment) step applied separately to curves within estimated clusters.

CURVE ALIGNMENT PROBLEM

Let X and Y represent curves defined over the time interval T. To register curves X, Y means to find a strictly increasing warping function $h : T \to \mathbb{R}$ which minimizes a chosen similarity criterion, for example

$$\int_T (X(t) - (Y \circ h)(t))^2 \, dt$$

In this work we use the Self–Modelling Time Warping (SMTW) [7] and Elastic Warping (EW) [3] curve alignment algorithms, which good performance has been proved in practice.

CLUSTERING OF MISALIGNED DATA

There are 3 possible approaches:

1. **raw data clustering**
   → possible incorrect clustering because of curves misalignment (Figure 2)
2. **alignment of the whole dataset followed by clustering of registered curves**
   → possible distortion in the curves shapes, caused by synchronization of curves with different profiles (Figure 3), or the overall poor alignment performance
3. **methods which combine clustering and curve registration**
   - ***k*–means alignment (KMA)**, [4]
     → only linear time transformations
     → insufficient synchronization when the misalignment is of non-linear character (Figure 4)
   - **Joined Probabilistic Curve Clustering and Alignment (JPCCA)**, [5]
     → only linear time transformations
     → EM algorithm within method may fail when curve profiles are complex
   - **truncated Pairwise Curve Synchronization followed by *k*–means clustering (tPCS)**, [6]
     → nonlinear warping functions
     → not always leads to correct clustering which is caused by insufficient curve alignment (Figure 5)
   - a new method ?

DYNAMIC TIME WARPING

Dynamic Time Warping (DTW) is a member of a wider area of registration methods. For discrete observations of two curves $X_1 = (X_1(t_1), \dots, X_1(t_{n_1}))$ and $X_2 = (X_2(s_1), \dots, X_2(s_{n_2}))$ the goal of DTW is to find the optimal warping path

$$w = \{(i_l, j_l), i_l \in \{1, \dots, n_1\}, j_l \in \{1, \dots, n_2\}, l = 1, \dots W\}$$

which minimizes the sum of distances between matched points

$$Q_{X_1,X_2}(w) = \sum_{(i_l, j_l) \in w} d\left(X_1(t_{i_l}), X_2(t_{j_l})\right),$$

where W is the length of the warping path w and d is a chosen distance, for example Euclidean. Instead of curve alignment we use DTW as a similarity measure of misaligned curves

$$dtw(X_1, X_2) = \min_w Q_{X_1,X_2}(w).$$

An overview of DTW and its details can be found in [1] or [2].

REFERENCES

[1] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. In *IEEE Trans. Acoust. Speech Signal Process*, volume 26, pages 43–49, 1978.

[2] K. Wang and T. Gasser. Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25(3):1251–1276, 1997.

[3] J. D. Tucker, W. Wu, and A. Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics and Data Analysis*, 61:50–60, 2013.

[4] L.M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. *k*–mean alignment for curve clustering. *Computational statistics and data Analysis*, 54:1219 – 1233, 2010.

[5] S. Gaffney and P. Smyth. Joint probabilistic curve clustering and alignment. In *In Advances in Neural Information Processing Systems 17*, pages 473–480. MIT Press, 2005.

[6] R. Tang and H. G. Müller. Time-synchronized clustering of gene expression trajectories. *Biostatistics*, 10(1):32–45, 2009.

[7] D. Gervini and T. Gasser. Self-modeling warping functions. *J. R. Statist. Soc. B*, 2004.

[8] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 1987.

2–STEP APPROACH

Let $\{X_1, \dots, X_N\}$ be a set of misaligned curves defined over the time interval T and represents the set which we would like to divide into K clusters.

1. Assign the curves into K clusters using the distance matrix $M_{dtw} = \{dtw\,(X_i, X_j)\}_{i,j=1,\dots,N}$ as an input for the k–medoids algorithm.
2. Separately align curves in each cluster and denote the aligned curves as $X_1^\star, \dots, X_N^\star$. For curves alignment we use the SMTW method, although an arbitrary registration algorithm could be chosen and this can be done according to a given structure of curves. To guarantee that the warping function h is strictly increasing and to avoid registration of distant time segments the following restriction
$$\lambda \int_T \left(\frac{1}{h'(t)} - 1 \right)^2 dt \qquad (1)$$
is added to the chosen registration method. The scaling factor λ influences the weight of the restriction.
3. Compute the average similarity within the formed clusters $C_1, \dots, C_K$
$$L = \sum_{i=1}^{K} \frac{1}{|C_i|} \sum_{j: X_j^\star \in C_i} \int_T \left(X_j^\star(t) - \mu_i(t) \right)^2 dt, \qquad \mu_i(t) = \frac{1}{|C_i|} \sum_{j: X_j^\star \in C_i} X_j^\star(t), \qquad t \in T \qquad (2)$$
4. If the number of iterations exceeds 100 or $L < \varepsilon$, where ε is a small given constant, stop. Otherwise repeat the algorithm with the registered curves $X_1^\star, \dots, X_N^\star$.

COMPARISON WITH OTHER METHODS

In order to compare the quality of clustering and curves alignment of the proposed **2–step approach** with the other existing methods we simulated 100 artificial datasets each consisting of 70 curves $X_1, \dots, X_{70}$. These curves were generated using five different template curves $\nu_1, \dots, \nu_5$, defined over the time interval $[0, 1]$ (Figure 1). Each curve X_i was obtained by warping a chosen template curve by a random time function g_i

$$\begin{aligned} g_i(t) &= c(t + b_2 e^{t-b_1} + a), && b_1 \sim N(0,4), b_2 \sim N(1, 0.01), t \in [0,1], i = 1, \dots, 70, \\ & && a \text{ and } c \text{ are normalizing constants which guarantee } g_i(0) = 0 \text{ and } g_i(1) = 1, \\ X_i(t) &= (\nu_j \circ g_i)(t), && \nu_j \text{ is a chosen template curve} \end{aligned}$$

	k-means without alignemnt	Elastic Warping *k*–means clustering	KMA	JPCCA	tPCS *k*–means clustering	2DTW-SMTW DTW clust.
RCC	49%	100%	9%	3%	61%	100%
AS	0.74	0.9991	-	-	0.79	0.9958
L–crit	0.21	0.0009	-	-	0.15	0.0054

Table 1: Comparison of clustering and alignment quality performance for the set of validated methods. The quality of clustering is expressed as the percentage of 100 generated cases where the whole set of 70 curves is correctly clustered (RCC). The average silhouette (AS), [8] and the L–criterion (2) measure the quality of curve alignment. Because of the low RCC values, the AS and L–criterion were not computed for JPCCA and KMA.

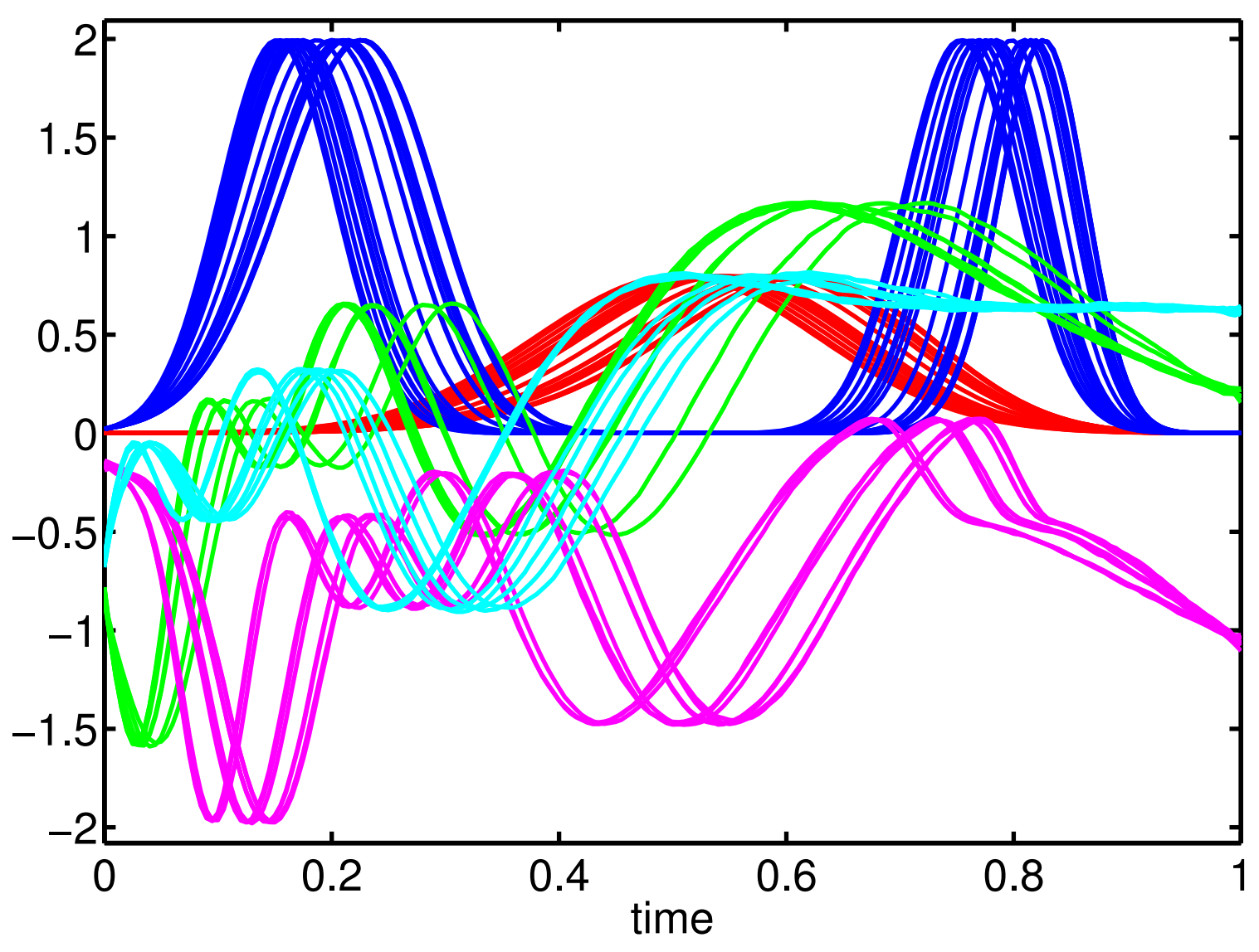

Figure 1: Original simulated data. The true cluster membership is depicted by different colors.

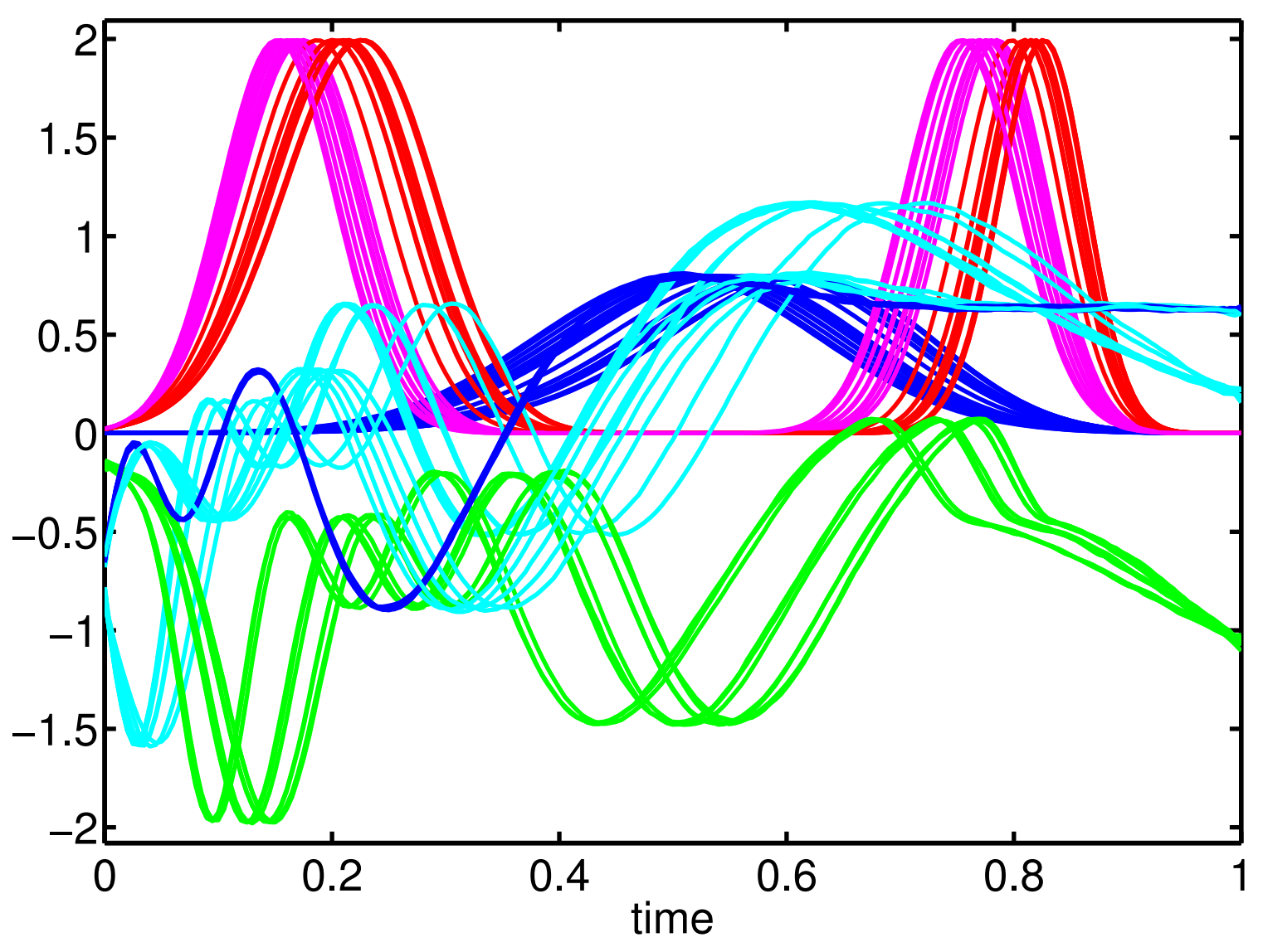

Figure 2: Simulated data clustered by the k–means algorithm without previous registration. Many curves are assigned into incorrect clusters.

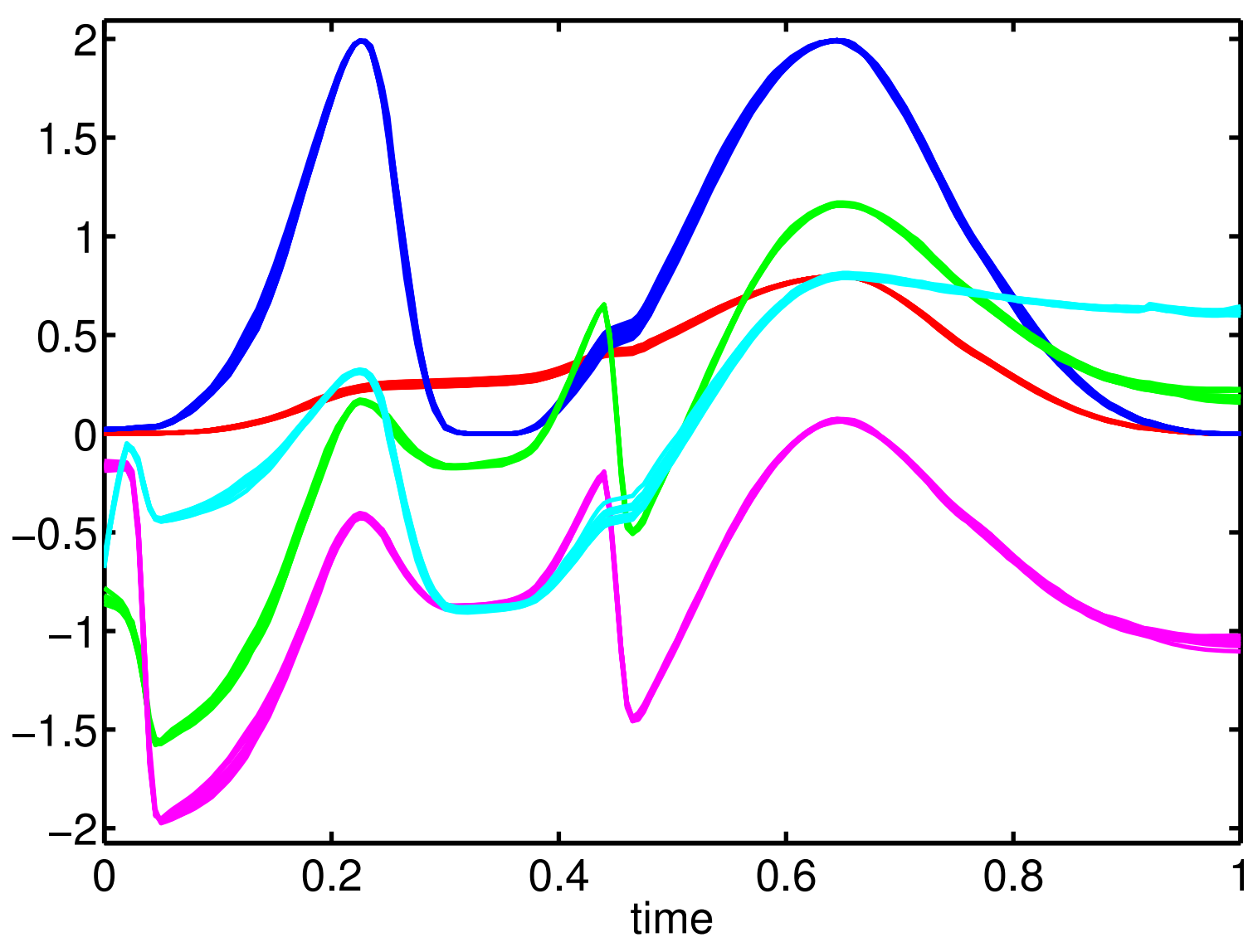

Figure 3: Simulated data registered as a whole dataset by Elastic Warping, [3] and then clustered by k–means. An evident distortion in the shapes of aligned curves is caused by the registration of the whole dataset to one target curve.

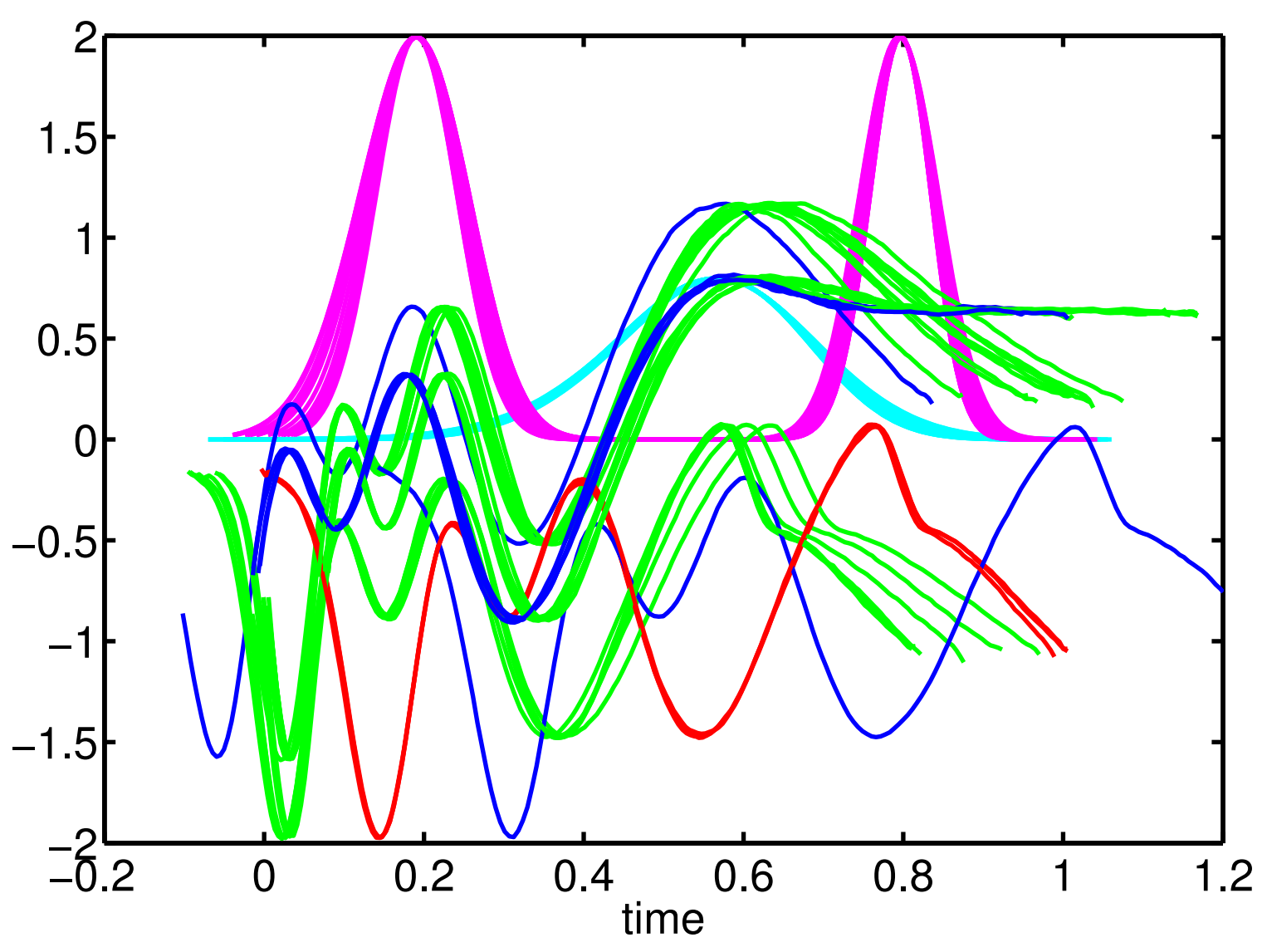

Figure 4: Simulated data registered and clustered by the k–means alignment clustering (KMA). The correct cluster membership is not reached.

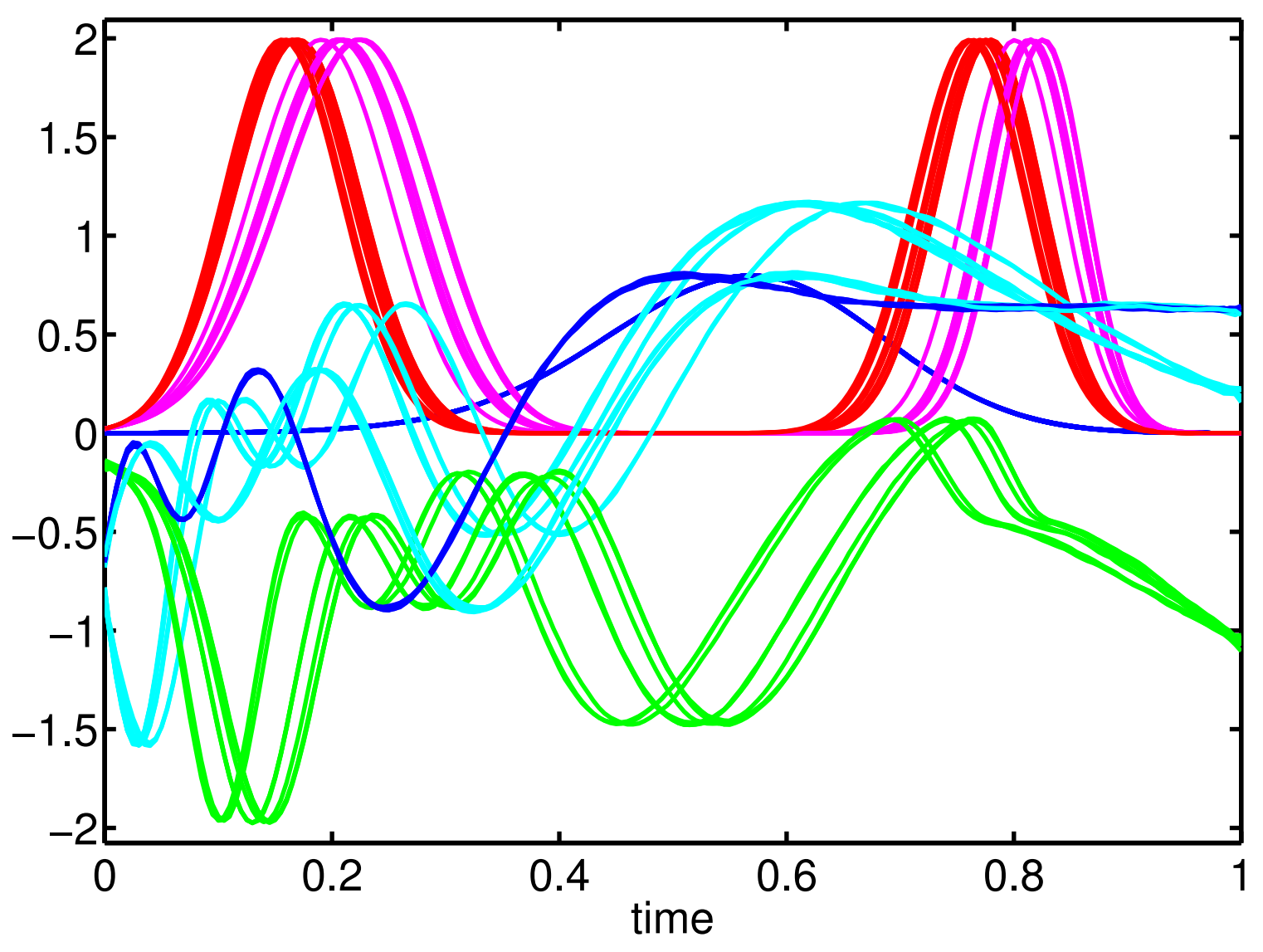

Figure 5: Simulated data registered as a whole set by truncated Pairwise Curve Synchronization (tPCS) followed by the k–means clustering. Because of an improper alignment, k–means assigns some curves into an incorrect cluster.

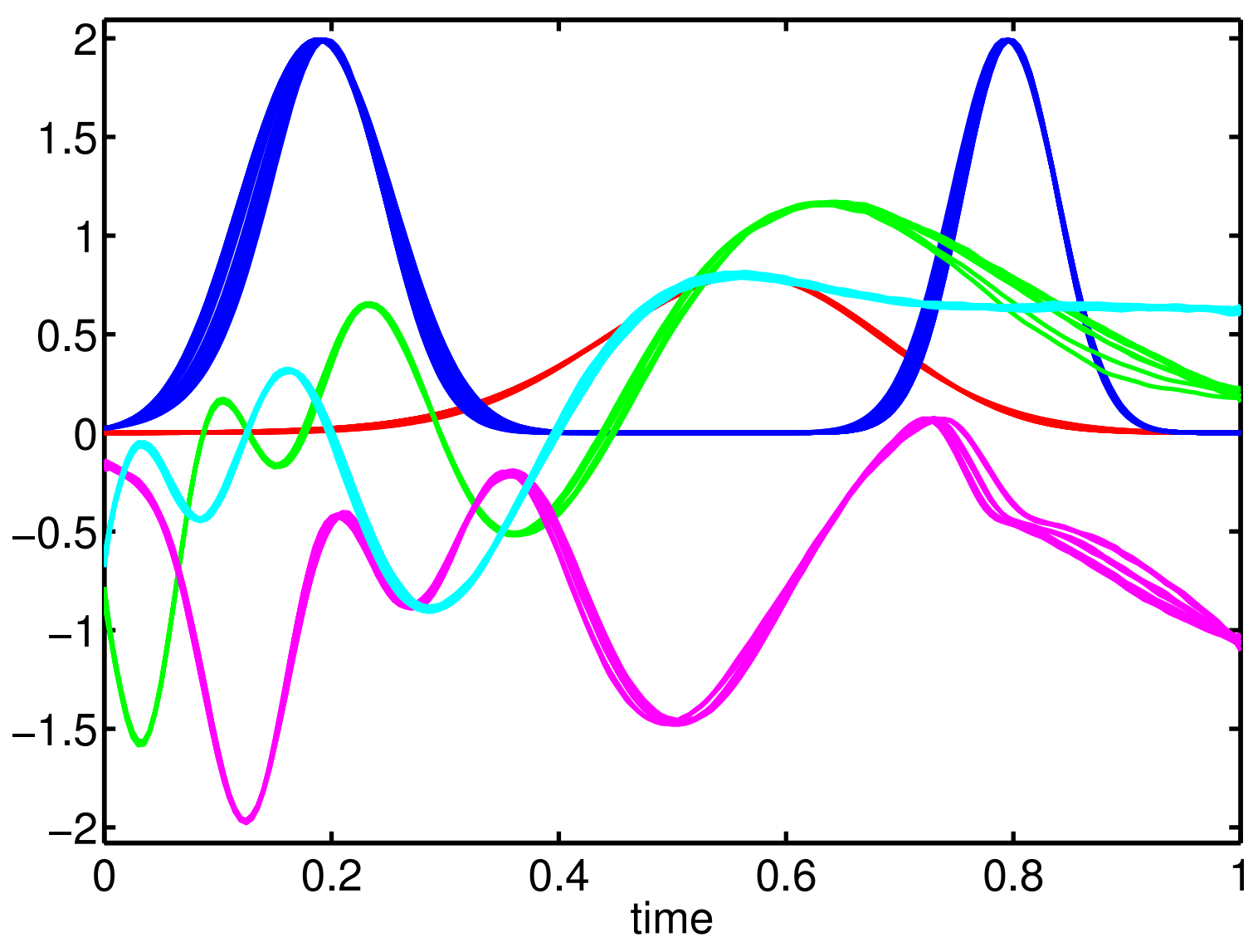

Figure 6: Simulated data registered and clustered by the 2–step approach. For the curve alignment SMTW with the restriction to the warping time defined in eq. (1) was used.

CONCLUSION

The 2–step approach combining the DTW distance matrix based clustering and curve alignment applied to curves within each cluster separately is an intuitive way for clustering misaligned functional data. Applying methods where the alignment of the whole set of investigated curves precedes the clustering step may fail i) due to the insufficient alignment leading to incorrect clusters assignment, or ii) due to the curve shapes distortions as a consequence of an "ideal" curve registration effort. Our 2–step approach successfully overcome both problems. In comparison to the other iterative methods (JPCCA, KMA, tPCS) the approach performs well i) when a non–linear warping functions character is present, and also ii) when many different curve profiles are present in the dataset.

ACKNOWLEDGEMENT

This research was supported by the Ministry of Health of the Slovak Republic, grant number MZ 2012/56-SAV-6 and by the Slovak Research and Development Agency, grant number APVV-0668-12.