Analysis of spike train data:
An application of k-mean alignment[*]

Mirco Patriarca, 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: mirco.patriarca@mail.polimi.it*;* laura.sangalli@polimi.it*;*
piercesare.secchi@polimi.it*;* simone.vantini@polimi.it

Abstract: We analyze the spike train data by means of the k-mean alignment algorithm in a double perspective: data as non periodic and data as periodic. In the first analysis, we show that alignment is not needed to identify paths. Indeed, without allowing for warping, we detect four clusters strongly associated to the four possible paths. In the second analysis, by exploiting the circular nature of data and allowing for shifts, we detect two clusters distinguishing between spike trains presenting higher or lower neuronal activity during the bottom-left/bottom-right movement respectively. In this latter case, the alignment procedure is able to match the four movements across paths.

Keywords and phrases: k-mean alignment, registration, functional clustering, spike trains.

Received August 2013.

1. Introduction

We here analyze the spike train data presented in Wu, Hatsopoulos and Srivastava (2014) with the aim of detecting spike trains associated to different paths or movements. This manuscript is divided in two sections: in the first one we analyze the 240 spike trains as functions defined on a common domain along the real axis (i.e., the interval [0,5]); in the second section, given the circularity of the four possible paths, we analyze the 240 spike trains as periodic functions. All analyses have been performed using the `fdakma` R package downloadable from CRAN (Parodi et al. (2014)).

2. Non-periodic data analysis

To look for clusters among spike trains we applied the k-mean alignment algorithm, detailed in Sangalli et al. (2010) and summarized in Sangalli, Secchi and Vantini (2014), to the 240 spike trains. Since a null value of intensity in a

---

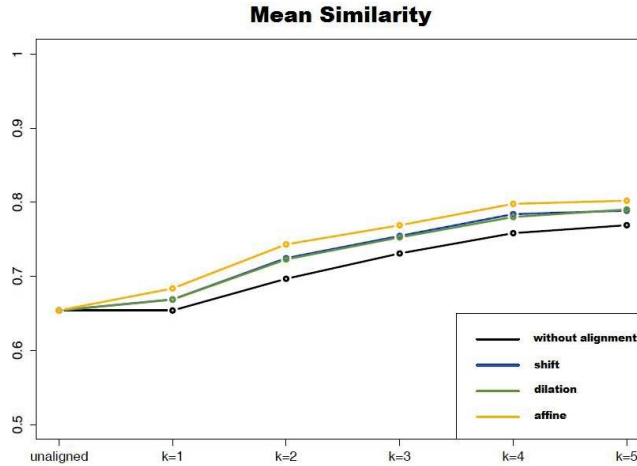[*]Main article 10.1214/14-EJS865.

[†]Corresponding author.

FIG 1. *Performance of the k-mean alignment algorithm when data are not assumed as periodic. The plot shows the mean similarity obtained with different values of k and different classes of warping functions.*

spike train means no neuronal activity, we used a similarity index that considered vertical shifts of the function as informative of a higher or lower neuronal activity. Therefore, we shall use the following similarity index:

$$\rho(f_i, f_j) = \frac{\int f_i(t) f_j(t) dt}{\sqrt{\int f_i(t)^2 dt} \sqrt{\int f_j(t)^2 dt}}. \tag{2.1}$$

Note that this similarity index assigns similarity equal to 1 (its maximal value) to couples of curves that differ only for a positive multiplying factor:

$$\rho(f_i, f_j) = 1 \iff \exists a \in \mathbb{R}^+ : f_i(t) = a f_j(t). \tag{2.2}$$

We performed the k-mean alignment algorithm allowing for affine warping functions, being the group of affinity the maximal group compatible with the index. We tested also the subgroups of shifts, dilations, and the degenerative identity subgroup:

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

Figure 1 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 first dot on the left represents
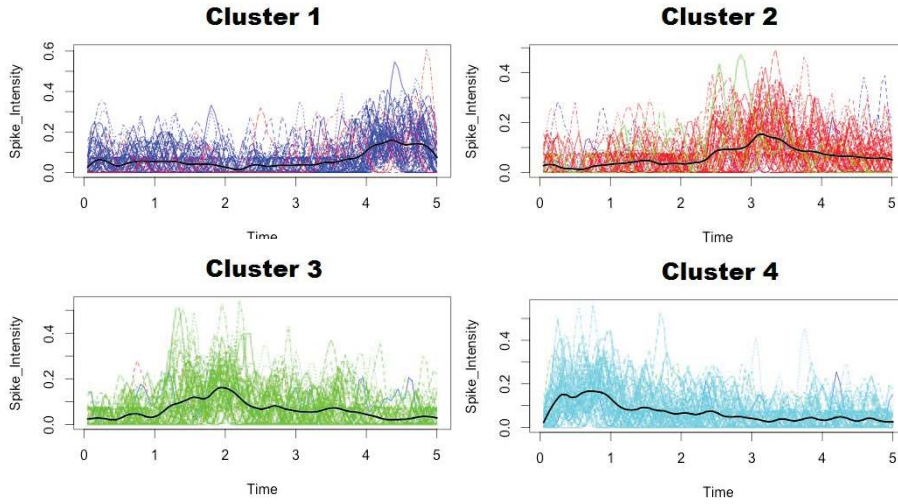
FIG 2. *Clusters and templates identified by setting $k = 4$ and the group of warping functions equal to $\mathcal{H}_{identity}$ when data are not assumed as periodic. Each panel refer to a different cluster. Spike trains are colored according to their respective path. Templates are black colored.*

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. 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 (2.2) 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 similar values and patterns of the four curves suggest the absence of phase variability. The low mean similarities achieved are instead evidence of an important residual amplitude variability in the data set that is not captured by the templates. All four curves present an elbow for $k = 4$ suggesting the presence of four clusters. In Figure 2 the four clusters obtained when no warping is allowed are reported. Almost the same clusters are obtained if groups $\mathcal{H}_{\text{affine}}$, $\mathcal{H}_{\text{shift}}$, or $\mathcal{H}_{\text{dilation}}$ are used instead. The four clusters turn out to be strongly associated to paths. In the left table of Figure 3 we classify indeed the 240 spike trains according to both clusters and paths showing a 92.5% agreement between the two classifications. This analysis shows that to assign each spike train to the correct path no alignment is needed. If the target of the analysis were instead to detect the four movements in each spike train, alignment would be of course needed. This latter issue is explored in the next section.

FIG 3. *The left panel displays the confusion matrix between paths and clusters identified by setting $k = 4$ and the group of warping functions equal to $\mathcal{H}_{identity}$ when data are not assumed as periodic. The right panel displays the confusion matrix between paths and clusters identified by setting $k = 2$ and the group of warping functions equal to $\mathcal{H}_{shift}$ when data are assumed as periodic.*

3. Periodic data analysis

Since the trajectories of the monkey right hand should be ideally close curves, always the same across paths, and with just the starting points differing across paths, we here analyze the 240 spike trains as periodic functions and apply the k-mean alignment with a similarity measure similar to that used in the previous section:

$$\rho(f_i, f_j) = \frac{\int_0^5 f_i(s)f_j(s)ds}{\sqrt{\int_0^5 f_i(s)^2 ds}\sqrt{\int_0^5 f_j(s)^2 ds}}. \tag{3.1}$$

The only difference is that integrals are not defined over the real axis but just on a single period (i.e., $[0,5]$) and that the functions are here assumed to be periodic. Then we tested $\mathcal{H}_{shift}$ and $\mathcal{H}_{identity}$ as possible groups of warping functions being dilations non-coherent with the similarity measure in the case of periodic function. From an algorithmic point of view, the only difference, with respect to the analysis presented in the previous section, is that the similarity between the template and the candidate warped functions is not computed on the common domain but always on the interval $[0,5]$. Indeed, being the functions periodic, what exceeds one interval extreme is considered at the other interval extreme.

In Figure 4 the mean similarities between the aligned curves and their relative templates are reported as functions of k (i.e., the number of clusters). The plot clearly suggests the use of shifts and two templates to align and cluster data. Thus we chose to set $k = 2$ and the group of warping functions equal to $\mathcal{H}_{shift}$.

In Figure 5 the two obtained clusters are reported. The first cluster (left panel) is made of 181 spike trains presenting a symmetric activity pattern around the higher activity peak. The second cluster (right panel) is instead made of 59 left-skewed spike trains characterized by a certain rate of activity also before the higher activity peak. This classification, as shown by the confusion matrix reported in the right panel of Figure 3, is not simply related to paths and thus it is worth further investigation in a biological perspective.
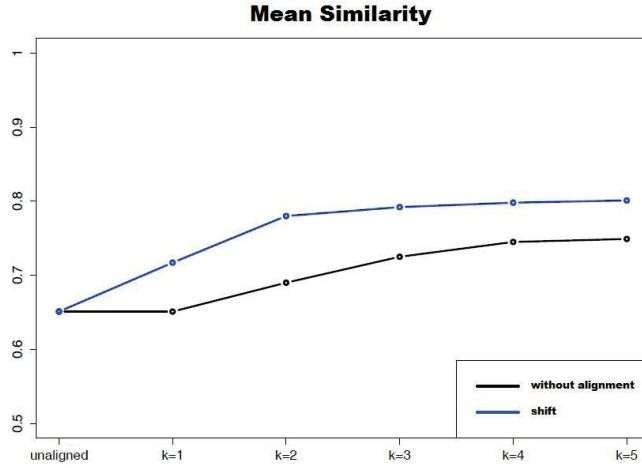
Mean Similarity

FIG 4. *Performance of the k-mean alignment algorithm when data are assumed periodic. The left panel shows the mean similarity obtained with different values of k and different classes of warping functions.*
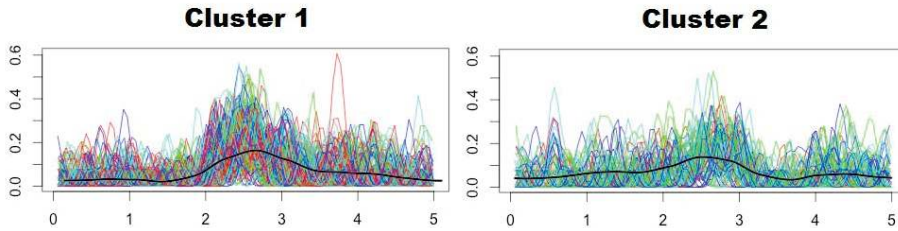

FIG 5. *Clusters and templates identified by setting $k = 2$ and the group of warping functions equal to $\mathcal{H}_{shift}$ when data are assumed periodic. Each panel refer to a different cluster. Spike trains are colored according to their respective path. Templates are black colored.*

On the contrary, the warping functions (i.e., shifts) result to be strongly associated to paths. Indeed, to effectively visualize the warping performed by the k-mean alignment algorithm on the 240 periodic curves, in Figure 6, we report, separately for each path, the 240 corresponding shifts as planar rotations. With the exception of a reduced number of spike trains, warping functions are clustered according to paths: all spike trains associated to the same path are shifted nearly the same way as if the algorithm were trying to match the movements across curves. If this is so, the spike trains assigned to the first cluster will be the ones characterized by a very high neuronal activity during the movement from button four to button one (i.e., bottom-right/top-right), while those assigned to the second cluster will be the ones characterized by high neuronal activity also during the movement from button three to button four (i.e., bottom-left/bottom-right).

Finally, in order to check for any relation between the two clusters and the warping functions, in Figure 7 we report the warping shifts separated according
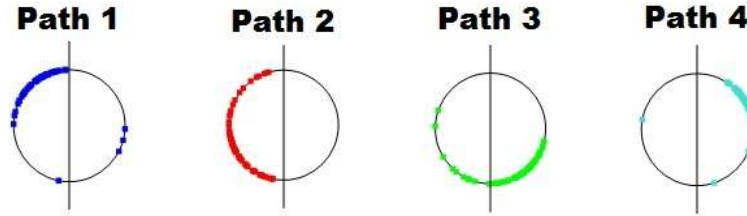
FIG 6. *Warping shifts by setting $k = 2$ and the group of warping functions equal to $\mathcal{H}_{shift}$. Each panel is associated to one of the four possible paths.*
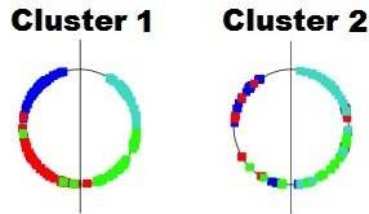

FIG 7. *Warping shifts by setting $k = 2$ and the group of warping functions equal to $\mathcal{H}_{shift}$. Each panel is associated to one of the two clusters.*

to cluster assignment. The picture clearly shows that there is no relation between warping functions and clusters. As a final comment, it is important to note that the choice of a proper similarity measure (i.e., eq. (3.1) in a periodic setting) and a proper group of warping functions (i.e., shifts) has been the key to unveil an hidden clustering structure in the signal shape that was completely masked by clusters in the phase directly related to the four path types and which are here captured by the warping functions.

4. Discussion

In the non-periodic data analysis we pointed out that alignment was not needed for the identification of clusters associated to paths. This finding is supported by the analysis presented in Lu and Marron (2014). Indeed, they show that the first two principal components of the unaligned data clearly point out four groups of data associated to paths while these groups are confounded if data are aligned. Instead, in the periodic data analysis we found a strong association between phase variability (i.e., periodic shifts) and paths. This kind of association has also been pointed out by Wu and Srivastava (2014).

Acknowledgements

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

Lu, X. and Marron, J. S. (2014). Analysis of spike train data: Comparison between the real and the simulated data. *Electronic Journal of Statistics* **8** 1793–1796, 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.

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

Wu, W., Hatsopoulos, N. and Srivastava, A. (2014). Introduction to neural spike train data for phase-amplitude analysis. *Electronic Journal of Statistics* **8** 1759–1768, Special Section on Statistics of Time Warpings and Phase Variations.

Wu, W. and Srivastava, A. (2014). Analysis of spike train data: Alignment and comparisons using the extended Fisher-Rao metric. *Electronic Journal of Statistics* **8** 1776–1785, Special Section on Statistics of Time Warpings and Phase Variations.