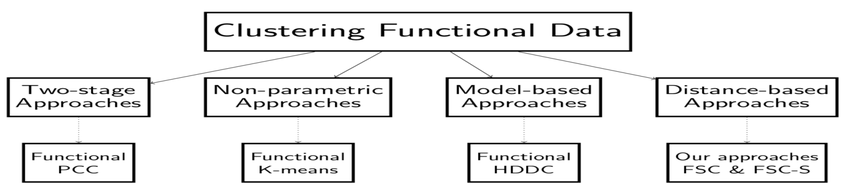
# FSC-S technique for clustering functional data

Maryam Al Alawi[1], Surajit Ray[1], and Mayetri Gupta[1]
[1] School of Mathematics and Statistics, University of Glasgow, UK
E-mail: m.al-alawi.1@research.gla.ac.uk

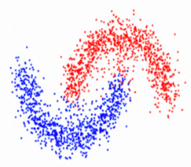
## Introduction

Clustering functional data (CFD) has been an active area of research in recent years. CFD aims to group curves with similar features in one cluster, and the cluster is usually represented by the mean of all the curves in the cluster. It is well documented in the clustering literature that clustering is an ill-defined problem, and the challenges such as finding the right number of clusters and proposals for appropriate measures of cluster accuracy is still an active area of research for multivariate clustering. Additionally, the high dimensionality of functional data and the lack of clear distributional theory for functional data makes CFD even more challenging.

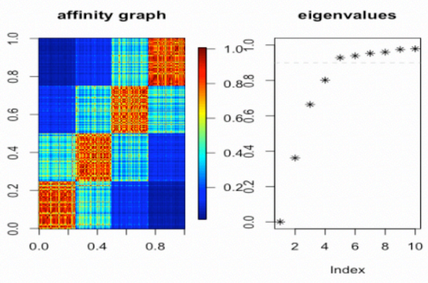
## Overview




## Methodology



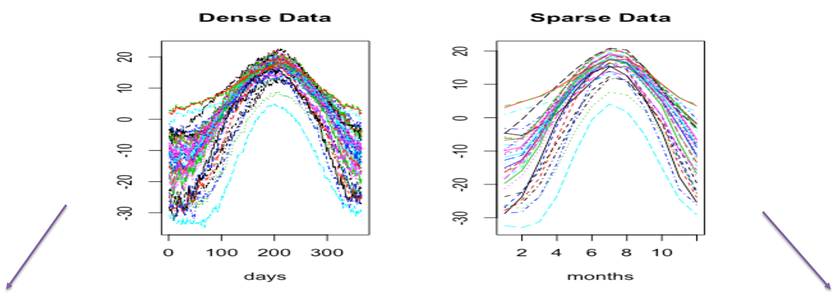
In a multivariate setting, spectral clustering has been successfully applied to cluster high-dimensional data embedded in nonlinear manifolds. The swiss roll example is commonly used to illustrate the power of the spectral method which outperforms most other standard clustering methods such as k-means or model-based clustering. Moreover, this clustering method can be easily implemented and most importantly does not require strong assumptions about the data. Taking into account the challenges for CFD as mentioned above, we develop a flexible framework to implement a spectral clustering method for functional data and through extensive simulations we have seen clear evidence of its superior performance in different functional data contexts.

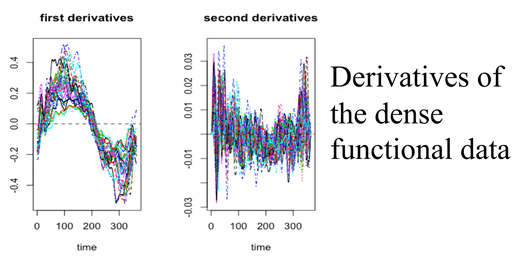



1. Smooth the data
2. Measure the Distance
3. Create the Similarity matrix
4. Create the Laplician Graph matrix
5. Cluster by K-means of the k eigenvectors from step 4

## Simulations, Applications, and Results

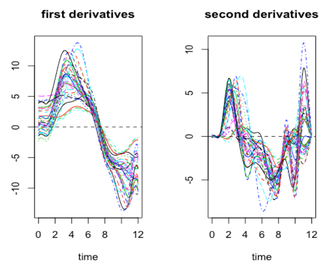
The Canadian Weather data consists of temperature measures for 35 selected cities distributed across Canada (available in FDA package). It has the daily temperature (365 time-points) and the monthly temperature (12 time-points) measures, which we consider as the dense and the sparse data respectively. We selected this data set for simulation because, it has been widely employed in FDA researches, the true clusters can be easily found, and it can be used as sparse or dense data.




Derivatives of the dense functional data

Derivatives of the sparse functional data



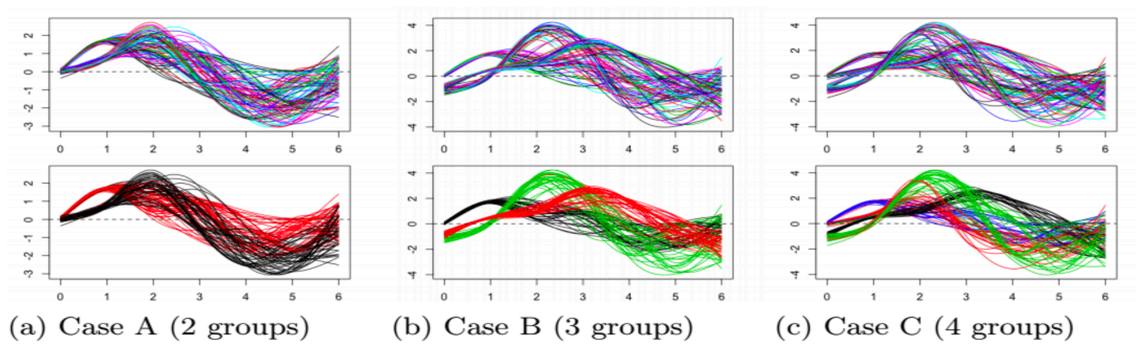

The smoothing technique used for the Canadian temperature functional data comes from the B-splines basis of order 4, with a penalized fit. Using this smoothing technique on the data, the error vectors for each curve are evaluated by $\varepsilon_i = Y_i - X(t_i)$. These error matrices are then perturbated by changing their distributions parameters, mixing and relocating the error values along the curves. Thus, we end up with 500 simulated data sets as dense data, and another 500 as sparse data $Y_1^*, Y_2^*, \ldots, Y^*_{1000}$.

These simulated data go through the clustering analysis using the above methods. Then the average classification correct rate is computed, the true clusters are available in (Ramsay and Silverman, 2005). It is important to know that for FSC and FSC-S, the distances between curves were measured by using the rates of changes in temperature (first derivatives), as they contain more discriminative information.

**Classification Correct Rate**

| Methods | Nonparametric perturbation | Parametric N(0, 0.5σ) | Parametric N(0, 1 σ) | Parametric N(0, 2 σ) | Parametric N(0, 5 σ) |
|---|---|---|---|---|---|
| **Dense Case** | | | | | |
| **funHDDC** | 72% | 71% | 72% | 72% | 70% |
| **Fd K-means** | 68% | 68% | 67% | 68% | 69% |
| **FPCC** | 51% | 51% | 51% | 52% | 53% |
| **FSC-S** | 82% | 83% | 83% | 83% | 79% |
| **Sparse Case** | | | | | |
| **funHDDC** | 71% | 74% | 74% | 70% | 61% |
| **Fd K-means** | 73% | 76% | 75% | 73% | 68% |
| **FPCC** | 75% | 77% | 76% | 73% | 64% |
| **FSC-S** | 72% | 72% | 72% | 72% | 72% |

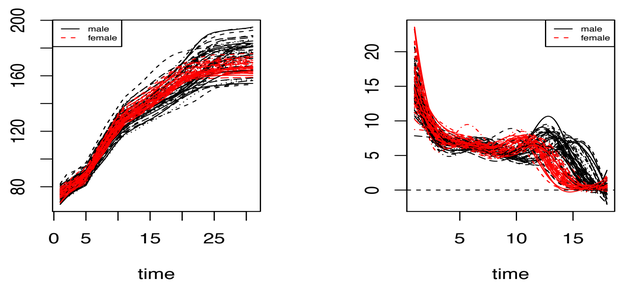
To evaluate the performance of our approach, we set up a simulation study to show the performance of the functional spectral clustering approaches on functional data that involves shifts in either phase, or amplitude or both. That simulation scheme was initially introduced by Sangalli et al. (2010), but we have expanded and made some modifications to the scheme. Through the general template $f(t) = \sin(t) + \sin\left(\frac{t^2}{2\pi}\right)$ we added shifts to some curves to create a new group. We started with creating 90 curves over the period from 0 to $2\pi$. Then we created another data set that consists of 90 curves over the period from 0 to $10\pi$ to examine the performance of our method on periodic functional data with phase/amplitude variations. According to Chen et al. (2012), this type of functional data can be assumed to lie on a nonlinear functional manifold.



(a) Case A (2 groups) (b) Case B (3 groups) (c) Case C (4 groups)

The top row shows the smoothed version of the generated data, the bottom row shows the clustering results after applying FSC-S($D_o$). In (A) the red curves make up the first group, while the black curves represent the amplitude shift. (B) shows 3 groups, 2 of them, black and green curves, similar to case (A), along with the third group (red) with a phase shift added to the data. In (C), a more difficult scenario is represented with 4 groups, three of which come from case (B). Our algorithm was successful in all three cases.

We applied FSC-S to the Berkeley Growth Study data (Ramsay and Silverman, 2005) which includes the heights of 39 boys and 54 girls. Some individuals reach puberty earlier than others, which is reflected clearly in the first derivatives of the data, that shows the rate of change in heights over the years for the two genders. We applied both FSC-S($D_o$) and FSC-S($D_1$), to compare the results. Once more, both these methods outperform the other clustering methods in the growth data.




The figure shows the resulting clusters of the Berkeley Growth data using FSC-S($D_o$) gives accuracy rates 86% (left) and using FSC-S($D_1$) gives accuracy rate 90% (right).

## Conclusions

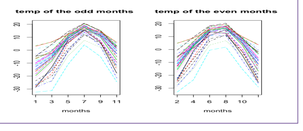
In summary, our distance based functional clustering approach is flexible, can accommodate different forms of functional data, and is easy to implement. Besides, it shows high accuracy rates and outperformed other methods in a variety of scenarios. Also, it is computationally faster than the other algorithms (like funHDDC, and FPCC) and always converges, whilst other methods might not like fd K-means.

## Future Work

Extending the algorithm:

- For the number of clusters k: we are currently using the eigengap criteria with some restrictions.
- For the scaling parameter σ: we are using a range of values of σ automatically.

We are examining the above through the **Down-Sampling**. This criterion refers to splitting the curves into two copies and applying the algorithm in each copy to validate the results.




## References

Bouveyron, C., and Jacques, J. (2011). Model-based clustering of time se- ries in group-specific functional subspaces. Advances in Data Analysis and Classification 5(4), 281– 300.
Chen, D., and Muller, H.-G. (2012). Nonlinear manifold representations for functional data. The Annals of Statistics 40(1), 1 – 29.
Febrero-Bande, M., and de la Fuente, M. O. (2012). Statistical computing in functional data analysis: The R package fda. usc. Journal of Sta- tistical Software 51(4), 1 – 28.
Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. Ad- vances in Data Analysis and Classification 8(3), 231 – 255.
Ng, A.Y., Jordan, M. I. and Weiss, Y. (2002). On spectral clustering: anal- ysis and an algorithm. In: Dietterich, T., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems 14, pp. 849–856. MIT Press, Cambridge.
Ramsay, J.O. and Silverman, B.W. (2005). Functional data analysis. Springer- Verlag New York.
Sangalli, L. M., Secchi, P., Vantini, S. and Valeria, V. (2010). K-mean align- ment for curve clustering. Computational Statistics & Data Analysis 54(5), 1219 – 1233.
Tzeng, S., Hennig, C., Li, Y.F. and Lin, C.J. (2018). Dissimilarity for Func- tional Data Clustering Based on Smoothing Parameter Commutation. Statistical Methods in Medical Research 27(11), 3492 – 3504.