Machine learning and statistical methods for clustering in FDA

Belén Pulido [1] Alba M. Franco-Pereira [1 2] Rosa E. Lillo [1 3]

[1]uc3m-Santander Big Data Institute (IBiDat), Madrid, Spain

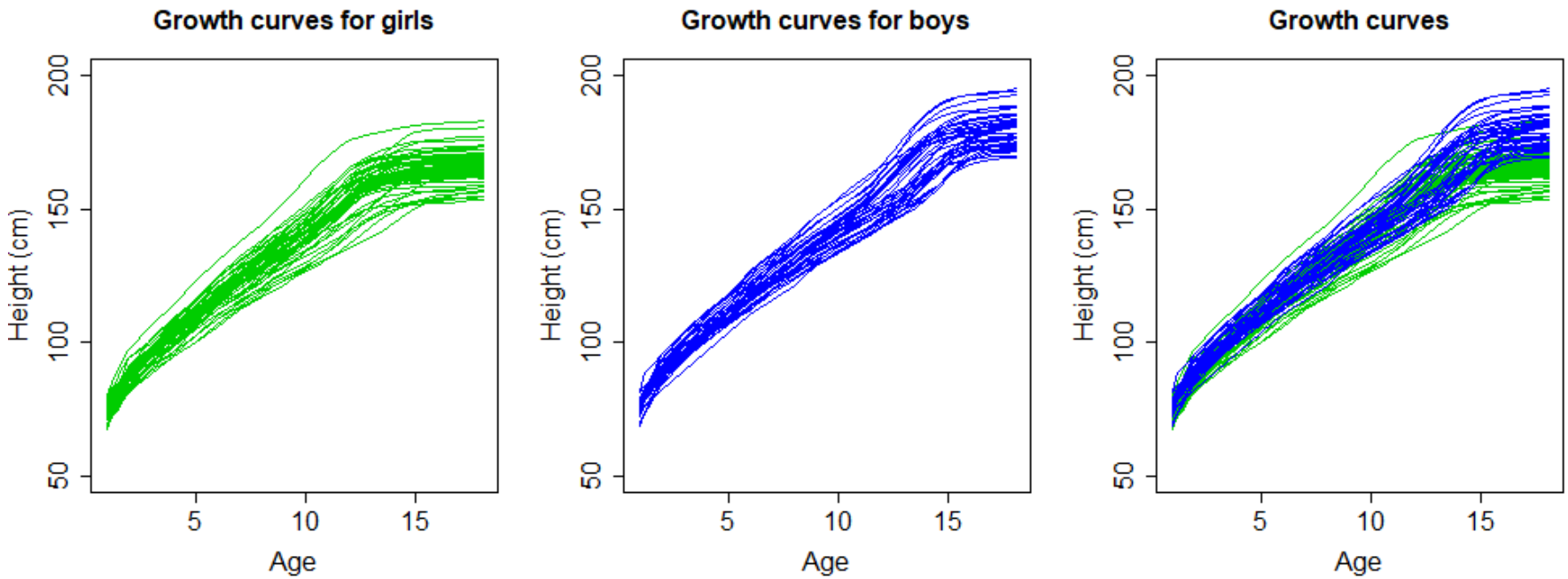
[2]Universidad Complutense de Madrid, Spain

[3]Universidad Carlos III de Madrid, Spain

New Bridges between Mathematics and Data Science,
9th November 2021

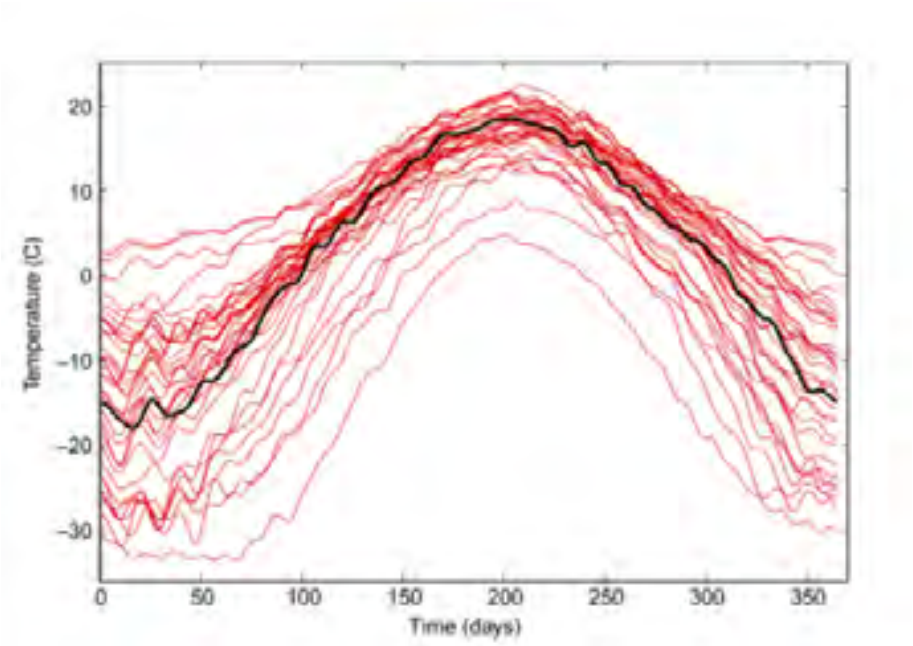
What is Functional Data Analysis?

FDA refers to the study of data where each observation is a real function $x_i(t)$, $i = 1, ..., n$, $t \in I$, where I is an interval in $\mathbb{R}$.



How to order functions?

- A total order does not exist for continuous functions, $C(I)$.
- Statistical depth provides a criterion for ordering the sample of curves from center-outward. López-Pintado and Romo (2009).



- Is there any partial order "equivalent" to the natural order of $\mathbb{R}$?
- **Our option**: Epigraph and hypograph indexes.

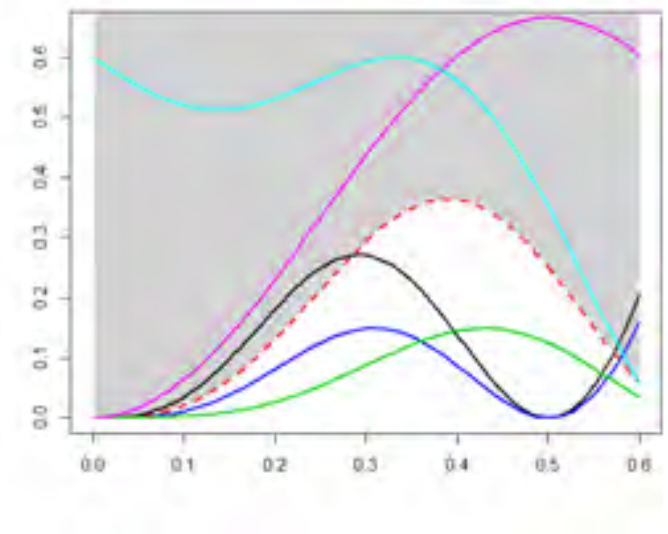
Epigraph and hypograph indexes

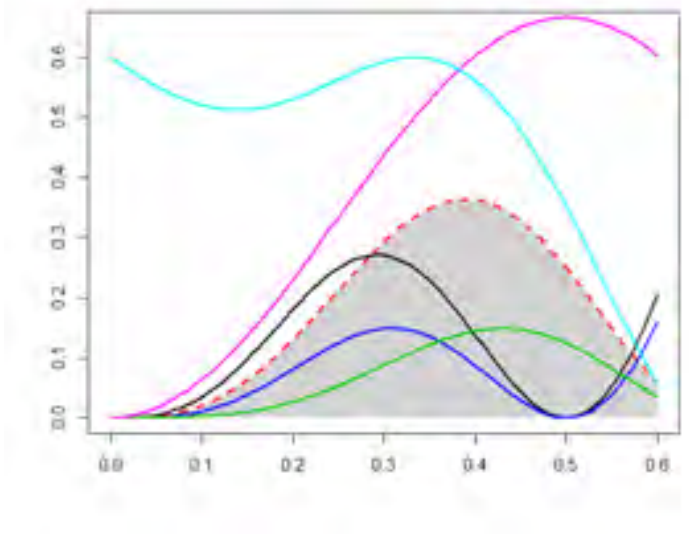
Let consider a stochastic process X with sample paths in $C(I)$ and distribution F_X. The graph of a function x in $C(I)$ is $G(x) = \{(t, x(t)), t \in I\}$.

The epigraph (epi) and the hypograph (hyp) of x are defined:

$$epi(x) = \{(t, y) \in I \times \mathbb{R} : y \geq x(t)\},$$

$$hyp(x) = \{(t, y) \in I \times \mathbb{R} : y \leq x(t)\}.$$





Epigraph and hypograph indexes

Given a sample of curves $\{x_1(t), ..., x_n(t)\}$

Sample versions:

$$EI_n(x) = 1 - \frac{\sum_{i=1}^{n} I(\{G(x_i) \subseteq epi(x)\})}{n} = 1 - \frac{\sum_{i=1}^{n} I(\{x_i(t) \geq x(t), t \in I\})}{n},$$

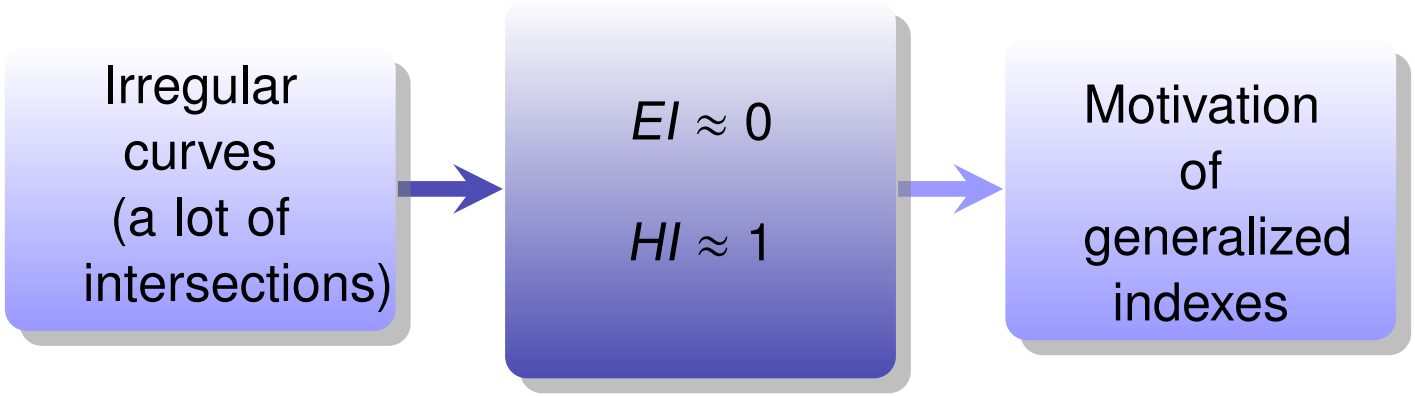
$$HI_n(x) = \frac{\sum_{i=1}^{n} I(\{G(x_i) \subseteq hyp(x)\})}{n} = \frac{\sum_{i=1}^{n} I(\{x_i(t) \leq x(t), t \in I\})}{n}.$$

Population versions:

$$EI(x, F_X) \equiv EI(x) = P(G(X) \subseteq epi(x)) = 1 - P(X(t) \geq x(t), t \in I),$$

$$HI(x, F_X) \equiv EI(x) = P(G(X) \subseteq hyp(x)) = 1 - P(X(t) \leq x(t), t \in I).$$

Generalized epigraph and hypograph indexes



Given a sample of curves $\{x_1(t), ..., x_n(t)\}$

$$MEI_n(x) = 1 - \sum_{i=1}^{n} \frac{\lambda(\{G(x_i) \subseteq epi(x)\})}{n\lambda(I)},$$

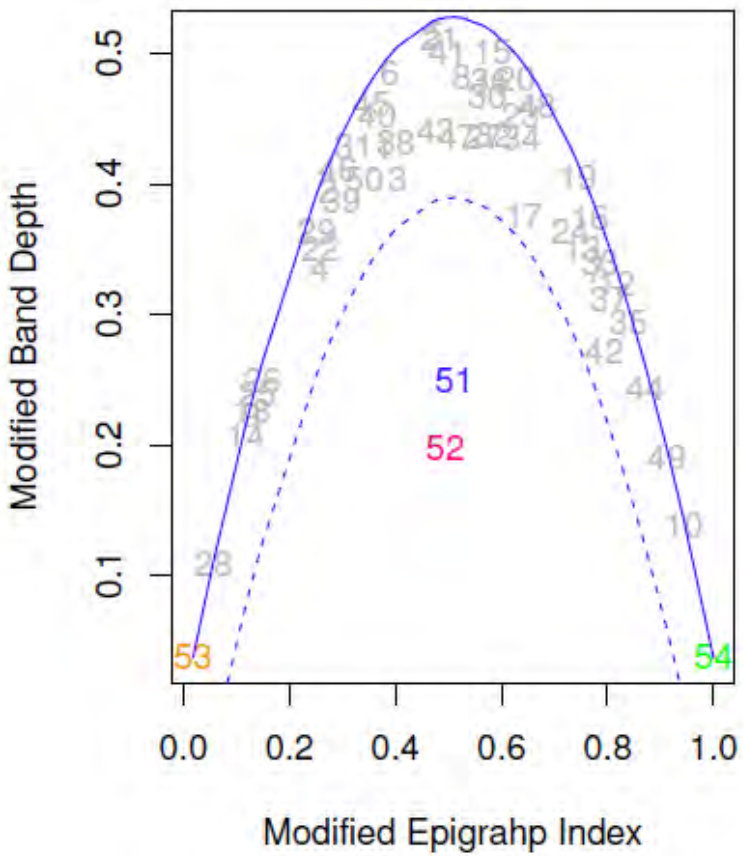
$$MHI_n(x) = \sum_{i=1}^{n} \frac{\lambda(\{G(x_i) \subseteq hyp(x)\})}{n\lambda(I)},$$

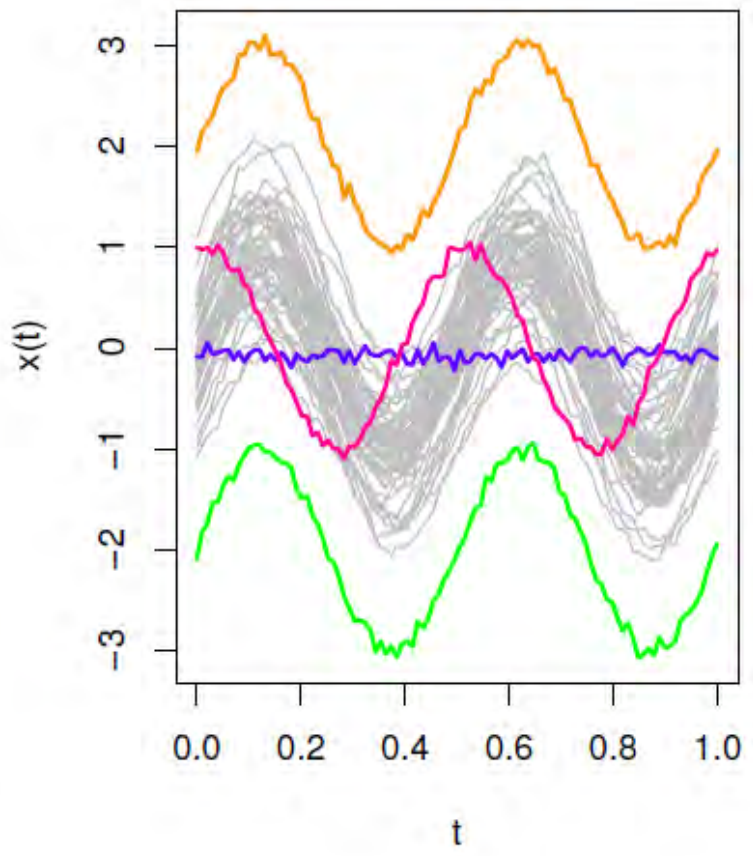
where λ stands for the Lebesgue's measure on $\mathbb{R}$.

Combining epigraph and hypograph indexes

The joint use of two indexes and their generalized versions have been exploited in recent papers:

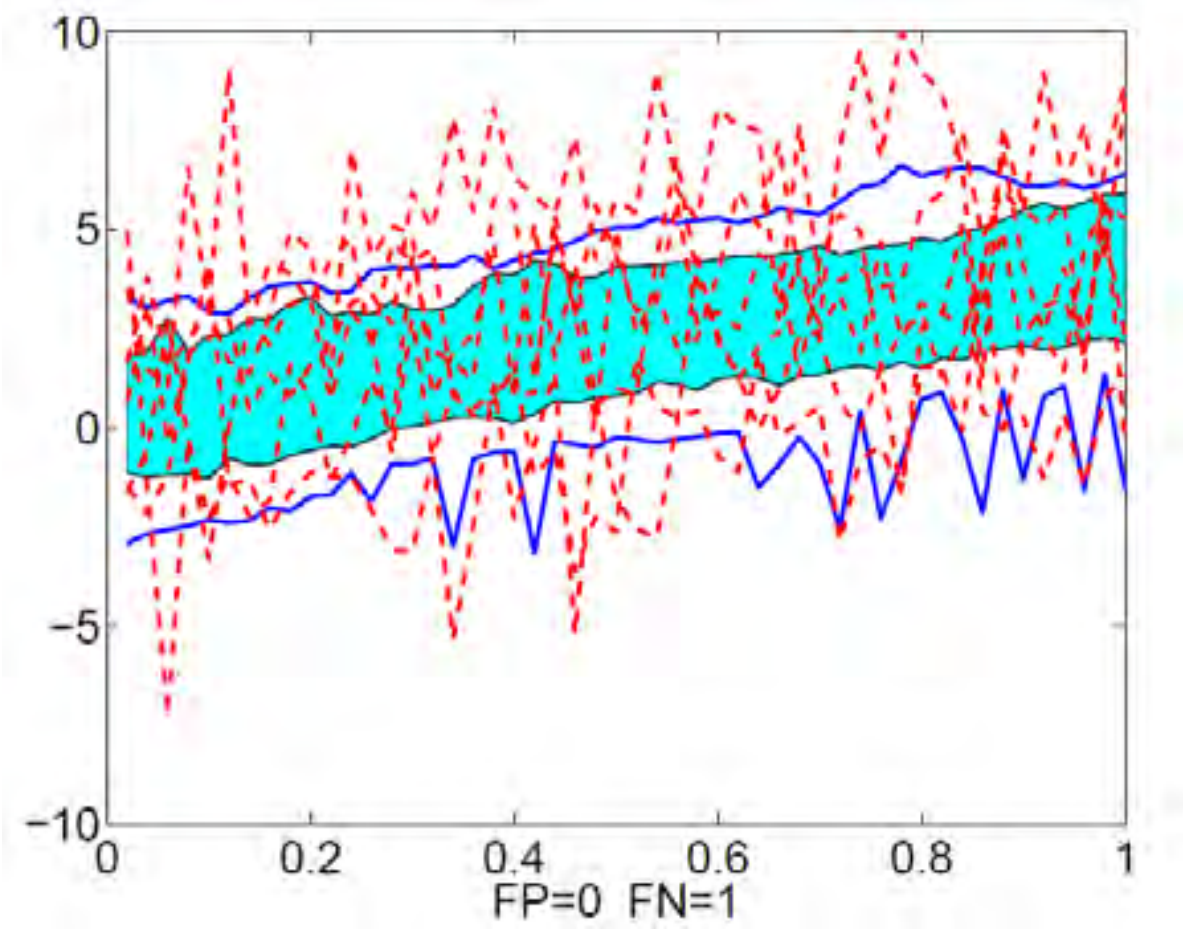
- Arribas-Gil and Romo (2014). Outlier detection.





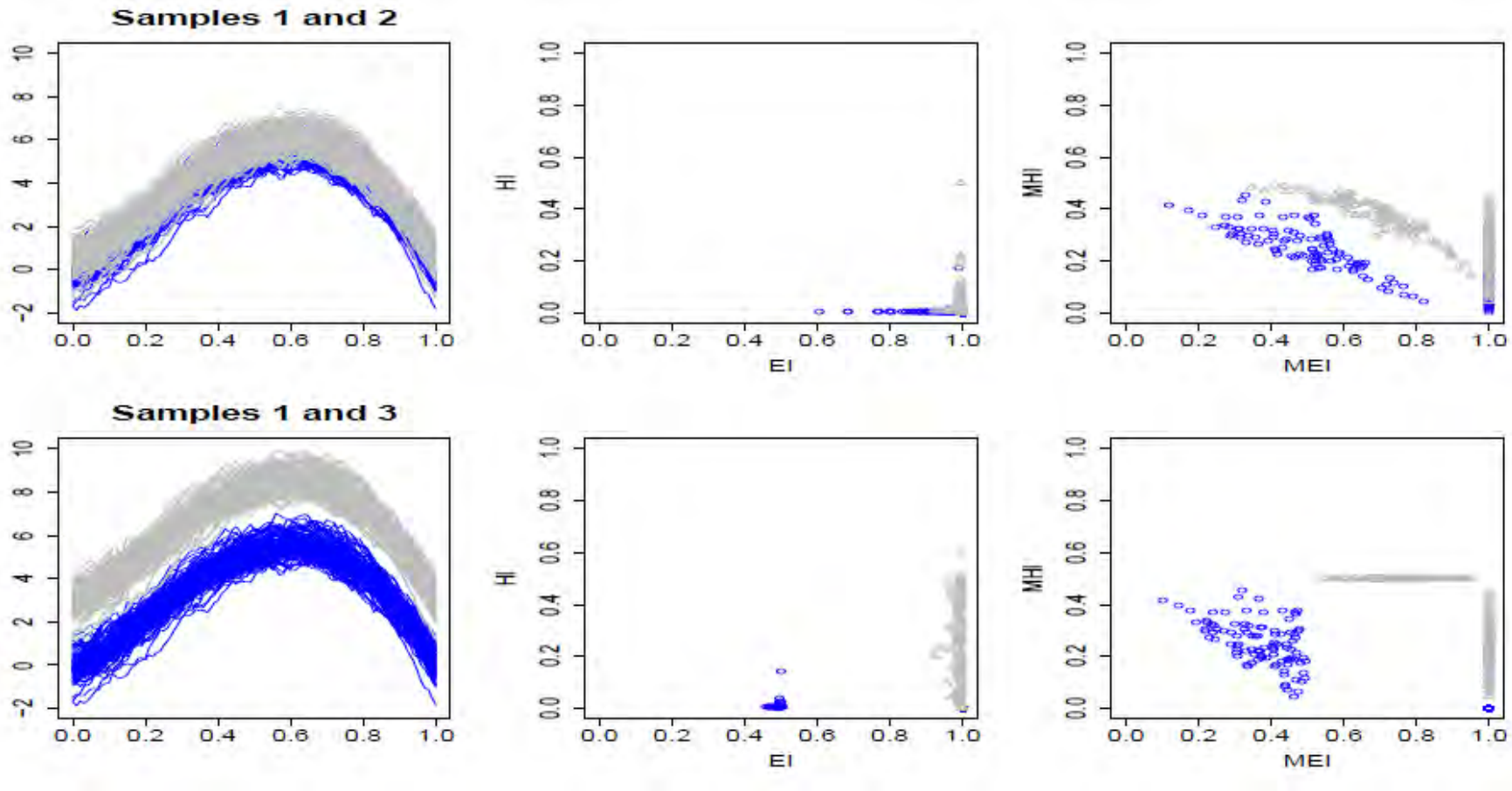
Combining epigraph and hypograph indexes

- Martín-Barragán et al. (2018). Functional boxplot.

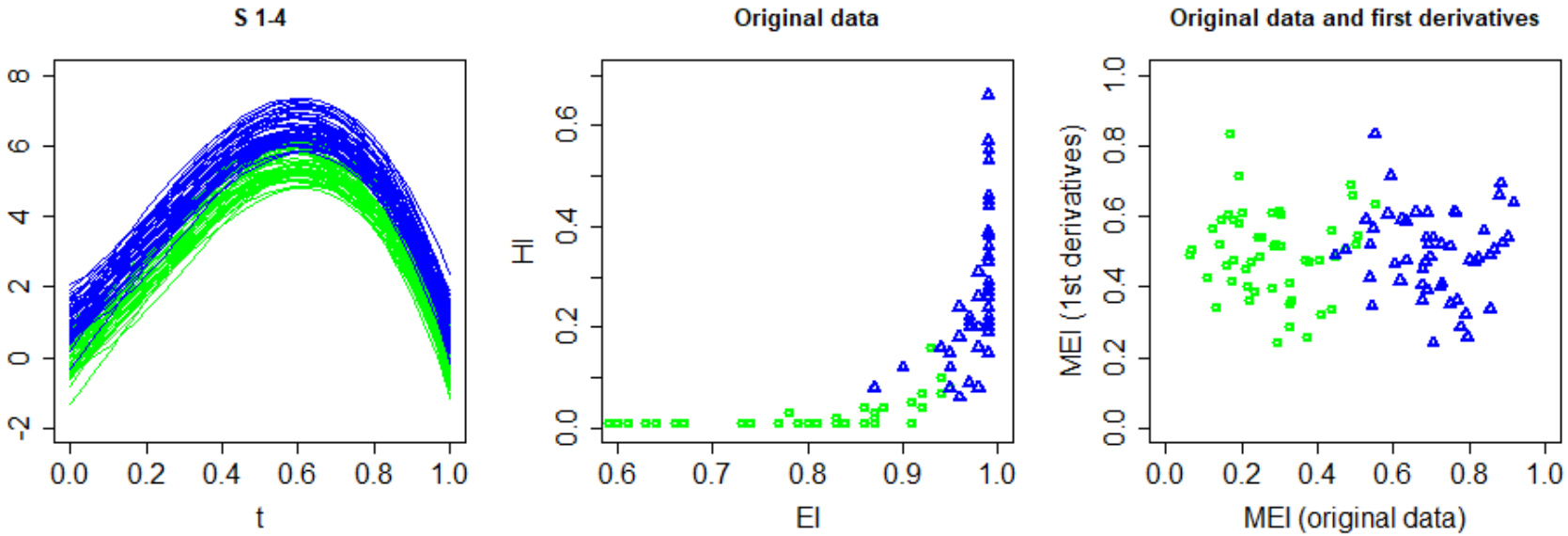


Combining epigraph and hypograph indexes

- Franco-Pereira and Lillo (2019). Homogeneity test.



A motivating example



Main idea for clustering functional data

Functional data set (S1) → Apply indexes to the data (S2) → Apply multivariate clustering techniques (S3) → Obtain a clustering partition into k groups (S4)

Cluster Analysis directly applied to functions

Martino et al. (2019). 'gmfd' R package.
Functional k-means

Distances

Let X_i, X_j, $i, j \in \mathbb{N}$ be two realizations of a stochastic process X.

- **Generalized Mahalanobis distance,** d_ρ

$$d_\rho(X_i, X_j) = \sqrt{\sum_{l=1}^{\infty} d^2_{M,l}(X_i, X_j) h_l(\rho)}$$

where $d_{M,l}(X_i, X_j) = \frac{\langle X_i - X_j, \varphi_l \rangle)^2}{\lambda_l}$, being φ_l and λ_l the eigenfunctions and eigenvalues of the covariance kernel, and $h_l(\rho) = \int_0^\infty \lambda_l \, exp(-\lambda_l c) \, g_\rho(c) dc$.

Values of ρ: $\rho_1 = 0.001$, $\rho_2 = 0.02$, $\rho_3 = 1$, $\rho_4 = 100$ and $\rho_5 = 10^8$

Some more distances

- **Truncated Mahalanobis distance,** d_k

$$d_K(X_i, X_j) = \sqrt{\sum_{l=1}^{K} \hat{d}^2_{M,l}(X_i, X_j)},$$

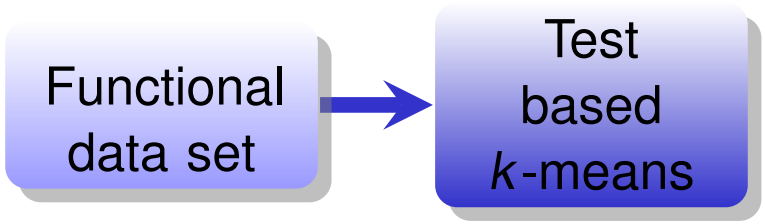
where $\hat{d}^2_{M,l}(X_i, X_j)$ stands for the empirical version of $d^2_{M,l}(X_i, X_j)$.

- L^2 **distance**

$$d_{L^2}(X_i, X_j) = \|X_i - X_j\|$$

Cluster Analysis directly applied to functions

Zambom et al. (2019)
Test based k-means



- Initialization
  - Random
  - k-means
  - hierarchical method
  - k-means ++
- Test based clustering
  - ANOVA test (parallelism)
  - t-test (equallity of means)

Multivariate Cluster Analysis

Methods used in this project (S3). Distance based clustering.

- Hierarchical clustering
  - Single, complete, average and centroid linkage
  - Ward's method

  Distance: Euclidean

- Partitional clustering
  - *k*-means

  Distances: Euclidean and Mahalanobis

- Other methods based on k-means
  - Kernel k-means (kkmeans)
  - Spectral Clustering (spc)
  - Support Vector Clustering (svc)

Our proposal

Pulido et al. (2021)

1. **Preparing functional data (S1):**
   - Fit a B-spline basis
   - Calculate the first and the second derivatives of the data

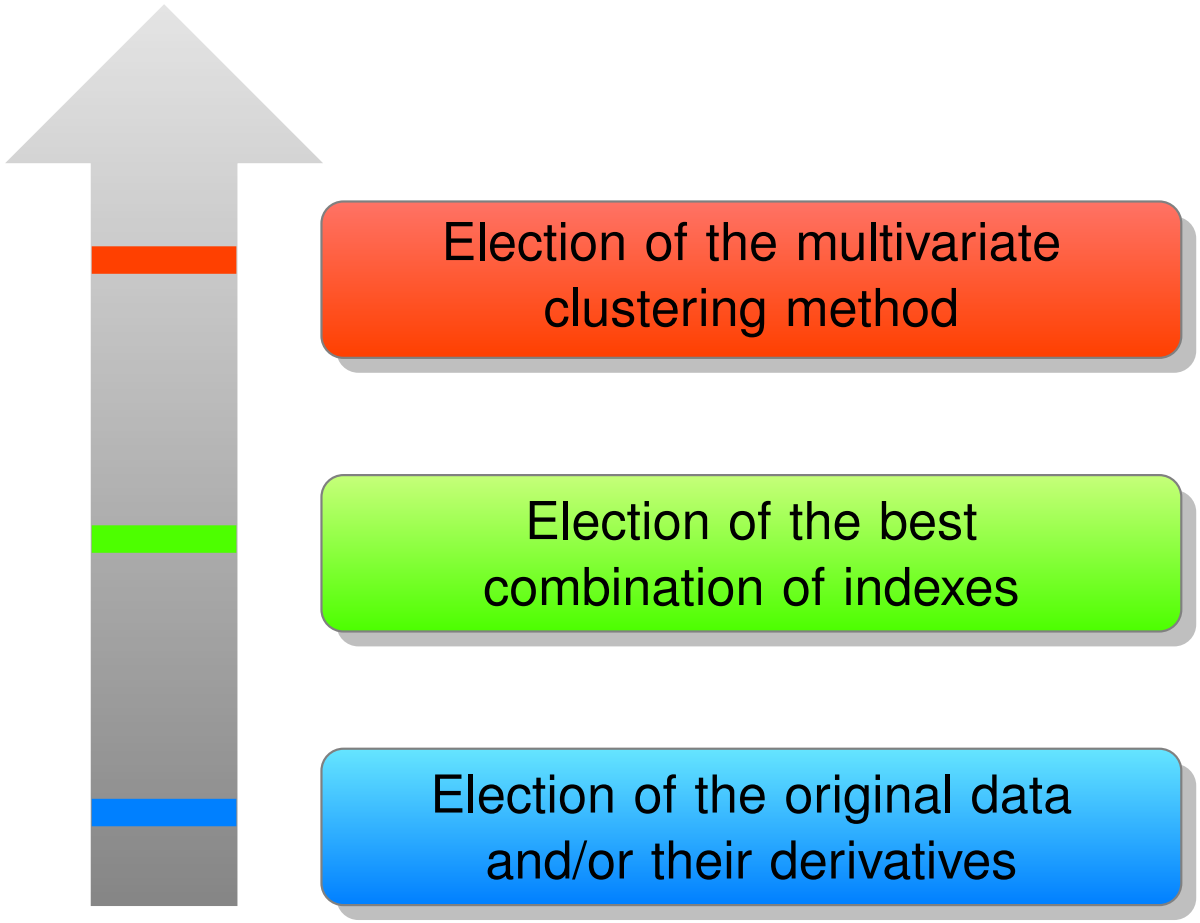
2. **Applying the indexes (S2):**

$$\_.EIHI = (EI, HI)$$

$$dd2.MEI = (dMEI, d2MEI)$$

$$\_dd2.EIHIMEI = (EI, HI, MEI, dEI, dHI, dMEI, d2EI, d2HI, d2MEI)$$

Three elections



Simulations scheme

- Two and three clusters
- Previously defined number of functions
- Previously defined interval I
- Previously defined number of equidistant points
- 100 simulations for each scheme

External validation criteria

Purity | **F-measure** | **Rand Index (RI)**

First simulation study

First simulation type. Flores et al. (2018), Franco-Pereira and Lillo (2019)

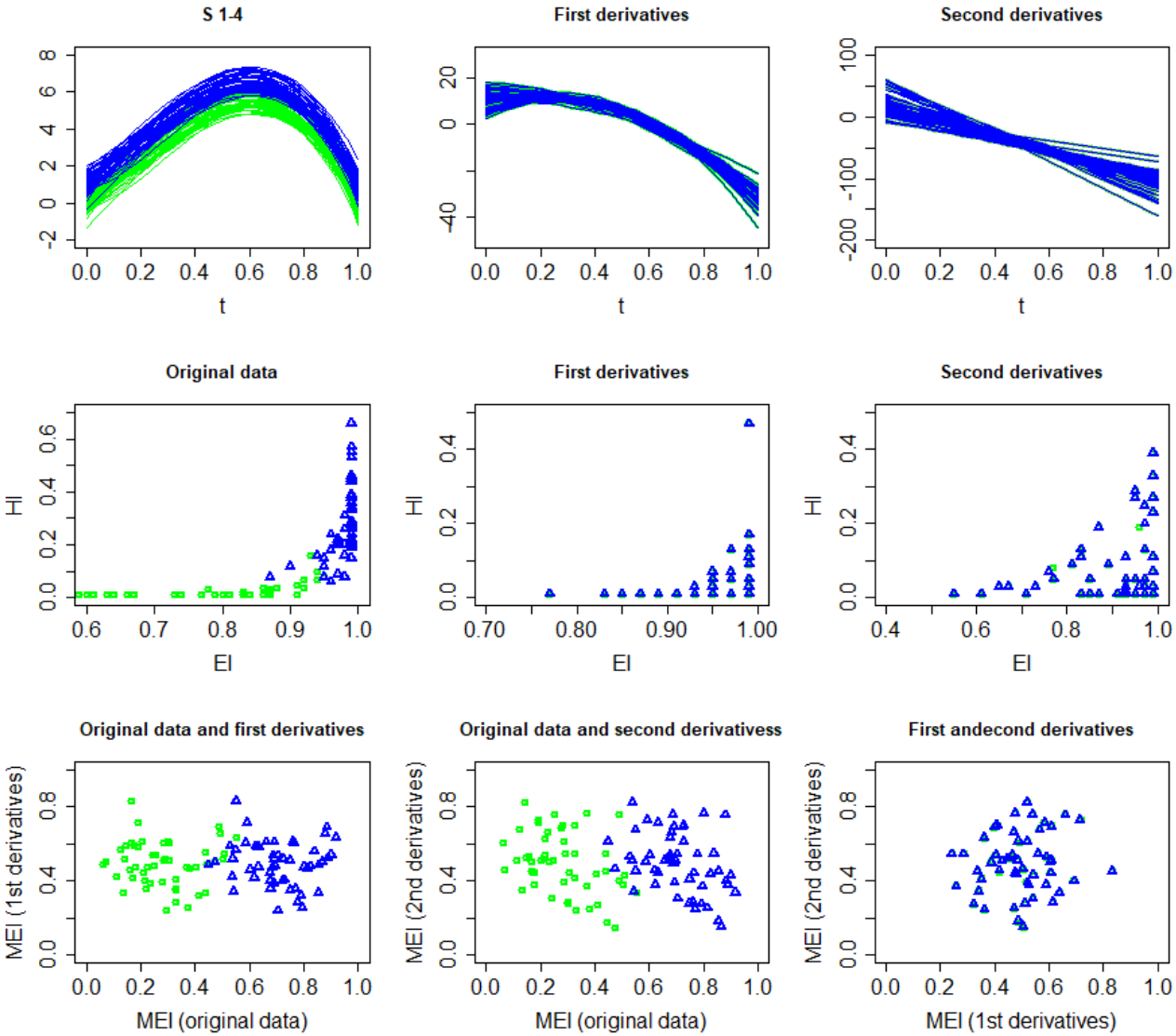
- Two clusters
- 100 functions defined in $[0, 1]$
- 30 equidistant points
- 50 functions for each model
- 100 simulations for each scheme

First simulation study

- **Model 1.** $X_1(t) = E_1(t) + e(t)$, where $E_1(t) = E_1(X(t)) = 30t^{\frac{3}{2}}(1-t)$ is the mean function and $e(t)$ is a centered Gaussian process with covariance matrix $Cov(e_i, e_j) = 0.3\ exp(-\frac{|t_i - t_j|}{0.3})$.
- **Model 2.** $X_2(t) = 30t^{\frac{3}{2}}(1-t) + 0.5 + e(t)$.
- **Model 3.** $X_3(t) = 30t^{\frac{3}{2}}(1-t) + 0.75 + e(t)$.
- **Model 4.** $X_4(t) = 30t^{\frac{3}{2}}(1-t) + 1 + e(t)$.
- **Model 5.** $X_5(t) = 30t^{\frac{3}{2}}(1-t) + 2\ e(t)$.
- **Model 6.** $X_6(t) = 30t^{\frac{3}{2}}(1-t) + 0.25\ e(t)$.
- **Model 7.** $X_7(t) = 30t^{\frac{3}{2}}(1-t) + \ h(t)$, where $h(t)$ is a centered Gaussian process with covariance matrix $Cov(e_i, e_j) = 0.5\ exp(-\frac{|t_i - t_j|}{0.2})$.
- **Model 8.** $X_8(t) = 30t(1-t)^2 + h(t)$.
- **Model 9.** $X_9(t) = 30t(1-t)^2 + e(t)$.

First simulation study

Samples 1 and 4



First simulation study

Samples 1 and 4

S 1-4 Top results				
	Purity	Fmeasure	RI	Time
svc._.EIHI	0.924	0.859	0.860	0.00237
svc._.EIHI	0.924	0.859	0.860	0.00260
kkmeans._.EIHI	0.924	0.859	0.860	0.00600
kmeans._.EIHI	0.923	0.857	0.858	0.00121

Initialization: kkmeans

Initialization: kmeans

Polynomial kernel

Euclidean distance

First simulation study

Samples 1 and 4

Functional k-means Top results				
	Purity	Fmeasure	RI	Time
L^2	0.917	0.846	0.847	0.72276
$d\rho$, $\rho = 0.02$	0.916	0.845	0.846	1.69700
$d\rho$, $\rho = 1$	0.916	0.844	0.846	1.73552
$d\rho$, $\rho = 0.001$	0.914	0.842	0.843	1.73665

Test based k-means Top results				
	Purity	Fmeasure	RI	Time
kmeans ++	0.500	0.507	0.495	0.32627
hclust	0.500	0.498	0.495	0.20851
kmeans	0.500	0.498	0.495	0.20502
random	0.500	0.502	0.495	0.27125

Second simulation study

Second simulation type. Martino et al. (2019)

- Two clusters
- 100 functions defined in $[0, 1]$
- 150 equidistant points
- 50 functions for each model
- 100 simulations for each scheme

Second simulation study

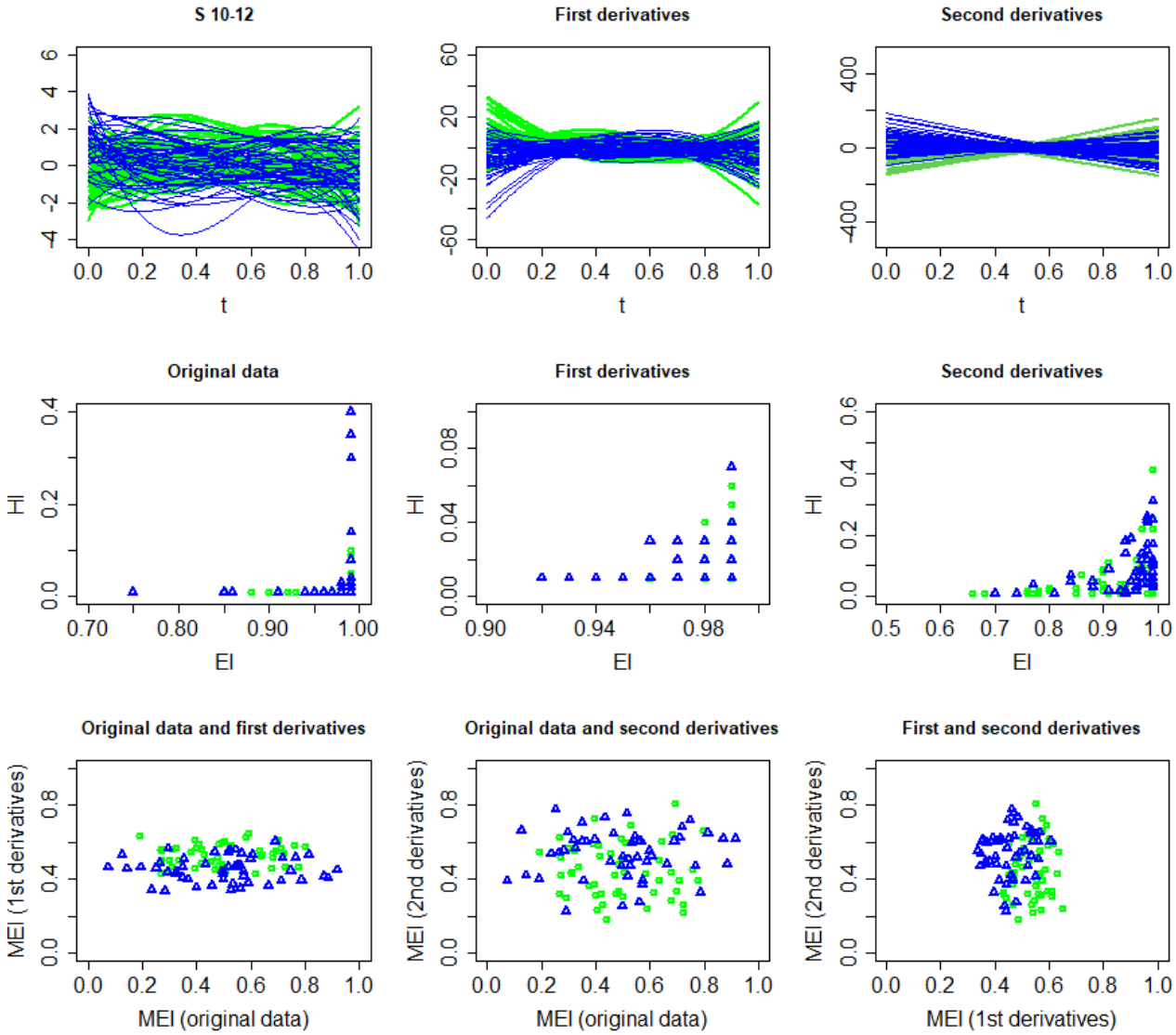
- **Model 10.** $X_{10}(t) = E_2(t) + \sum_{k=1}^{100} Z_k \sqrt{\rho_k}\theta_k(t)$, where $E_2(t) = t(1-t)$ is the mean function, $\rho_k = \begin{cases} \frac{1}{k+1} & \text{if} \quad k \in \{1,2,3\}, \\ \frac{1}{(k+1)^2} & \text{if} \quad k \geq 4, \end{cases}$ and $\{\theta_k, k \geq 1\}$ is an orthonormal basis of $L^2(I)$ defined as

$$\theta_k(t) = \begin{cases} \mathbb{1}_{[0,1]}(t) & \text{if} \quad k = 1, \\ \sqrt{2}\,sin\,(k\pi t)\mathbb{1}_{[0,1]}(t) & \text{if} \quad k \geq 2, \quad k \text{ even}, \\ \sqrt{2}\,cos\,((k-1)\pi t)\mathbb{1}_{[0,1]}(t) & \text{if} \quad k \geq 3, \quad k \text{ odd}. \end{cases}$$

- **Model 11.** $X_{11}(t) = E_3(t) + \sum_{k=1}^{100} Z_k \sqrt{\rho_k}\theta_k(t)$, where $E_3(t) = E_2(t) + \sum_{k=1}^{3} \sqrt{\rho_k}\theta_k(t)$
- **Model 12.** $X_{12}(t) = E_4(t) + \sum_{k=1}^{100} Z_k \sqrt{\rho_k}\theta_k(t)$, where $E_4(t) = E_2(t) + \sum_{k=4}^{100} \sqrt{\rho_k}\theta_k(t)$

Second simulation study

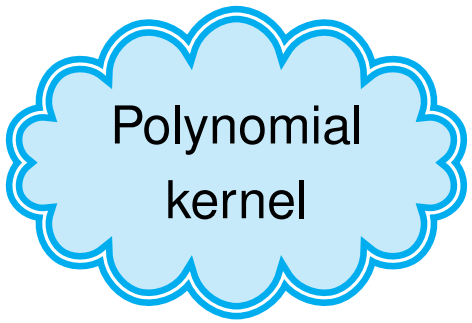
Samples 10 and 12

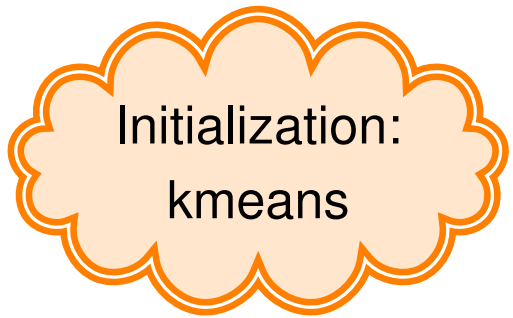


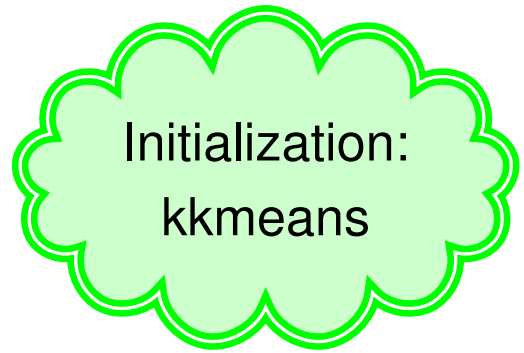
Second simulation study

Samples 10 and 12

S 10-12 Top results				
	Purity	Fmeasure	RI	Time
kkmeans.dd2.MEI	0.957	0.918	0.919	0.00423
kkmeans.dd2.EIHIMEI	0.957	0.918	0.918	0.00466
svc.dd2.EIHIMEI	0.956	0.916	0.917	0.00276
svc.dd2.MEI	0.956	0.916	0.917	0.00185







First simulation study

Samples 10 and 12

Functional k-means Top results				
	Purity	Fmeasure	RI	Time
$d\rho, \rho = 1e+08$	0.831	0.718	0.718	7.9055
$d\rho, \rho = 100$	0.637	0.554	0.548	10.1940
$d\rho, \rho = 0.02$	0.551	0.504	0.502	9.2982
$d\rho, \rho = 1$	0.549	0.503	0.502	9.8511

Test based k-means Top results				
	Purity	Fmeasure	RI	Time
kmeans	0.545	0.502	0.501	0.30259
hclust	0.542	0.502	0.500	0.39506
random	0.539	0.501	0.499	0.53736
kmeans ++	0.536	0.503	0.499	0.99704

Third simulation study

Third simulation type. Zambom et al. (2019)

- Three clusters
- 150 functions defined in $[0, \frac{\pi}{3}]$
- 100 equidistant points
- 50 functions for each model
- 100 simulations for each scheme

Third simulation type

Model 13. $X_{13}(t) = \frac{1}{1.3}\, sin(1.3t) + t^3 + a + 0.3 + \epsilon_1$

Model 14. $X_{14}(t) = \frac{1}{1.2}\, sin(1.3t) + t^3 + a + 1 + \epsilon_1$

Model 15. $X_{15}(t) = \frac{1}{4}\, sin(1.3t) + t^3 + a + 0.2 + \epsilon_1$

Model 16. $X_{16}(t) = sin(1.5\pi t) + cos(\pi t^2) + b + 1.1 + \epsilon_1$

Model 17. $X_{17}(t) = sin(1.7\pi t) + cos(\pi t^2) + b + 1.5 + \epsilon_1$

Model 18. $X_{18}(t) = sin(1.9\pi t) + cos(\pi t^2) + b + 2.2 + \epsilon_1$

Model 19. $X_{19}(t) = \frac{1}{1.8}\, exp(1.1t) - t^3 + a + \epsilon_2$

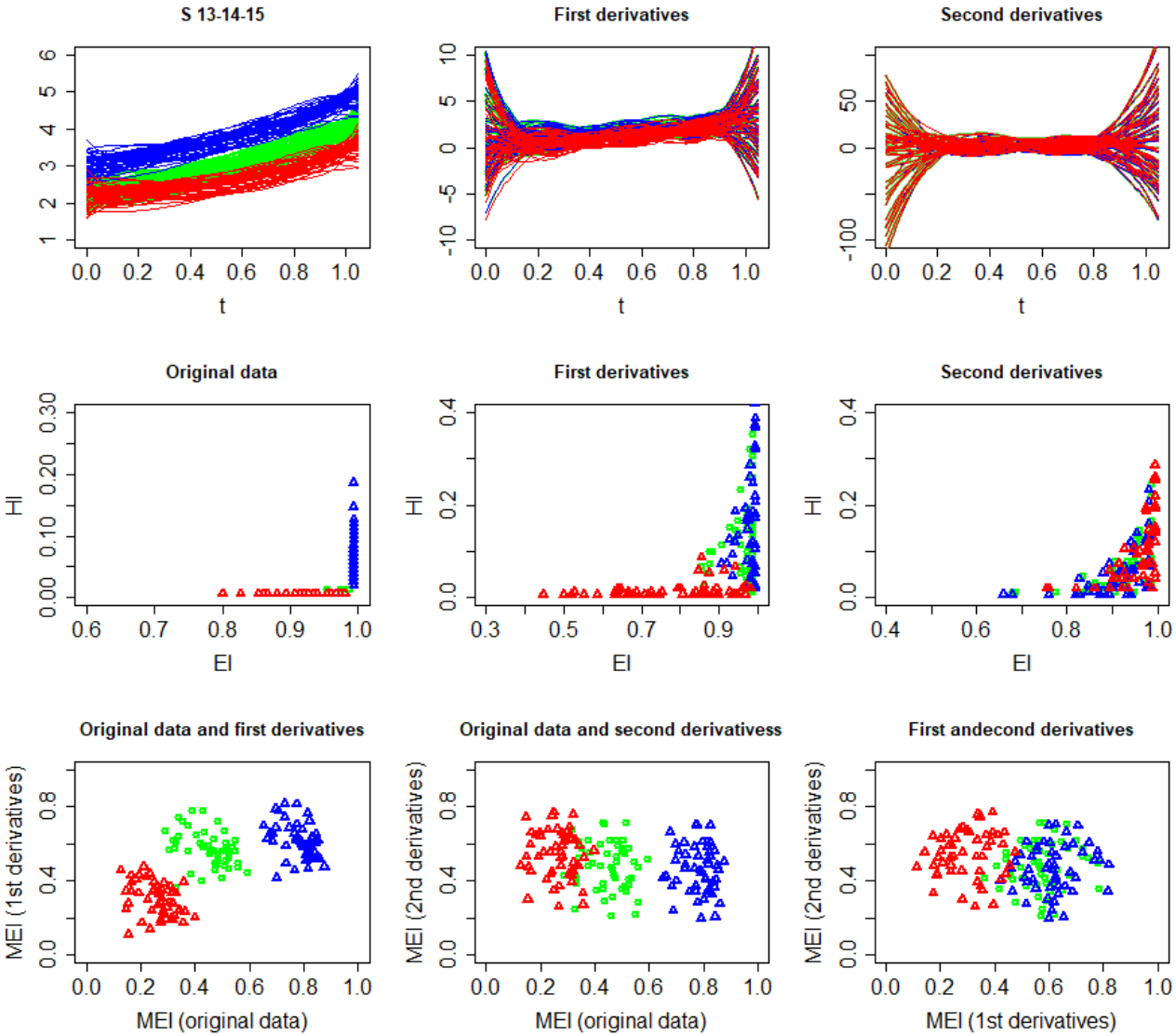
Model 20. $X_{20}(t) = \frac{1}{1.7}\, exp(1.4t) - t^3 + a + \epsilon_2$

Model 21. $X_{21}(t) = \frac{1}{1.5}\, exp(1.5t) - t^3 + a + \epsilon_2$

$a \sim U(\frac{-1}{4}, \frac{1}{4})$ $\quad b \sim U(\frac{-1}{2}, \frac{1}{2})$ $\quad \epsilon_1 \sim N(2, 0.4^2)$ $\quad \epsilon_2 \sim N(2, 0.4^2)$

Third simulation study

Samples 13, 14 and 15



Second simulation study

Samples 13, 14 and 15

S 10-12 Top results				
	Purity	Fmeasure	RI	Time
kmeans._dd2.MEI	0.987	0.974	0.983	0.00257
kmeans._dd2.MEI	0.987	0.974	0.983	0.00267
svc._d.MEI	0.986	0.974	0.982	0.00694
kmeans._d.MEI	0.986	0.973	0.982	0.00216

Euclidean distance

Mahalanobis: distance

Initialization: kkmeans

First simulation study

Samples 13, 14 and 15

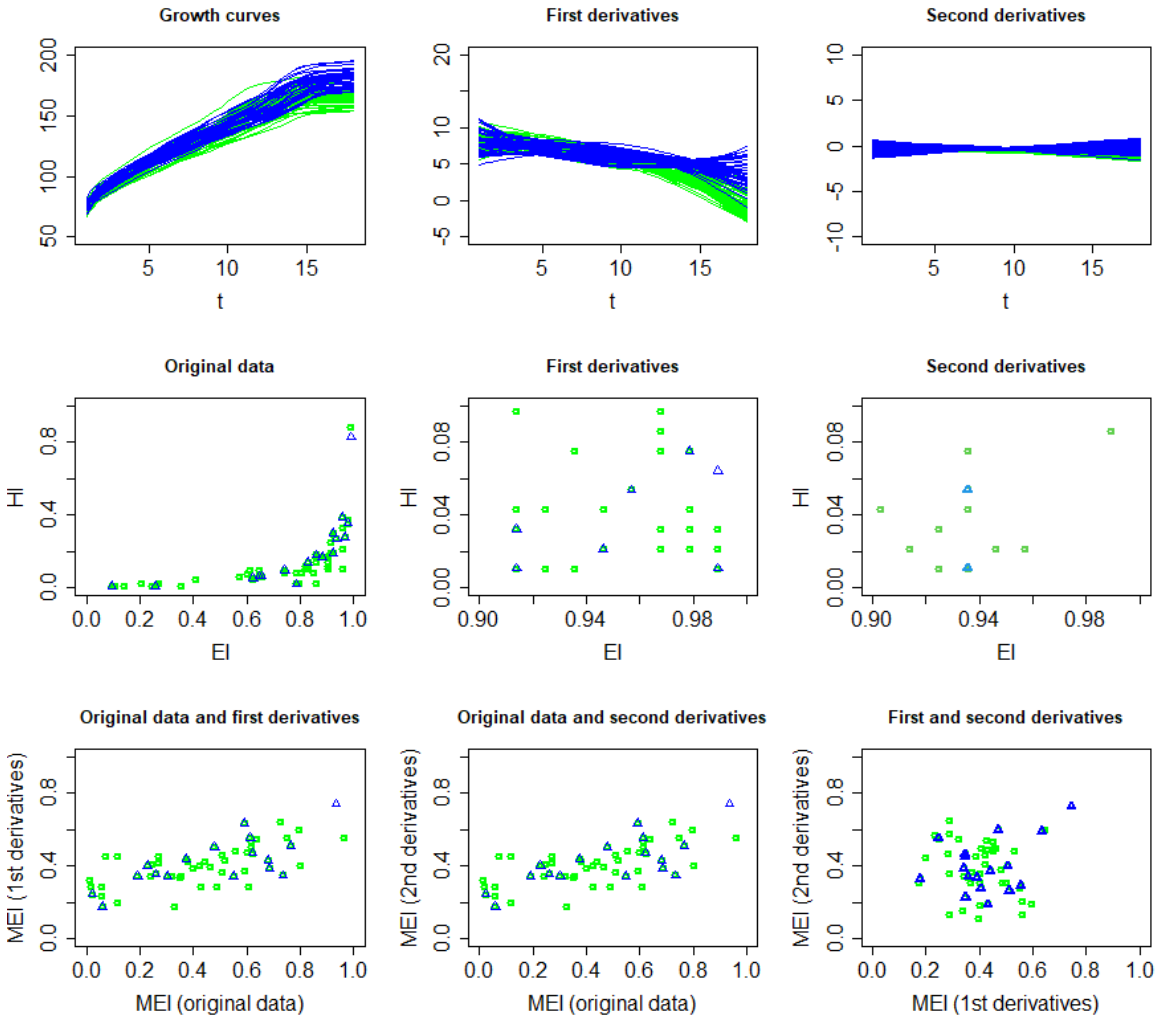
Functional k-means Top results				
	Purity	Fmeasure	RI	Time
$d\rho$, $\rho = 0.001$	0.936	0.894	0.928	6.50802
$d\rho$, $\rho = 0.02$	0.934	0.890	0.925	6.28237
L^2	0.934	0.887	0.923	1.38996
$d\rho$, $\rho = 1$	0.927	0.885	0.921	6.39378

Test based k-means Top results				
	Purity	Fmeasure	RI	Time
kmeans $++$	0.955	0.915	0.944	0.99653
kmeans	0.953	0.912	0.942	4.96029
random	0.952	0.910	0.940	1.05298
hclust	0.947	0.903	0.936	4.98203

Application to a real data set

Berkeley Growth Study

'fda' R-package



Application to a real data set

Our best result.
kmeans.dd2.MEI (euclidean distance)
RI: 0.936
Time: 0.00498s

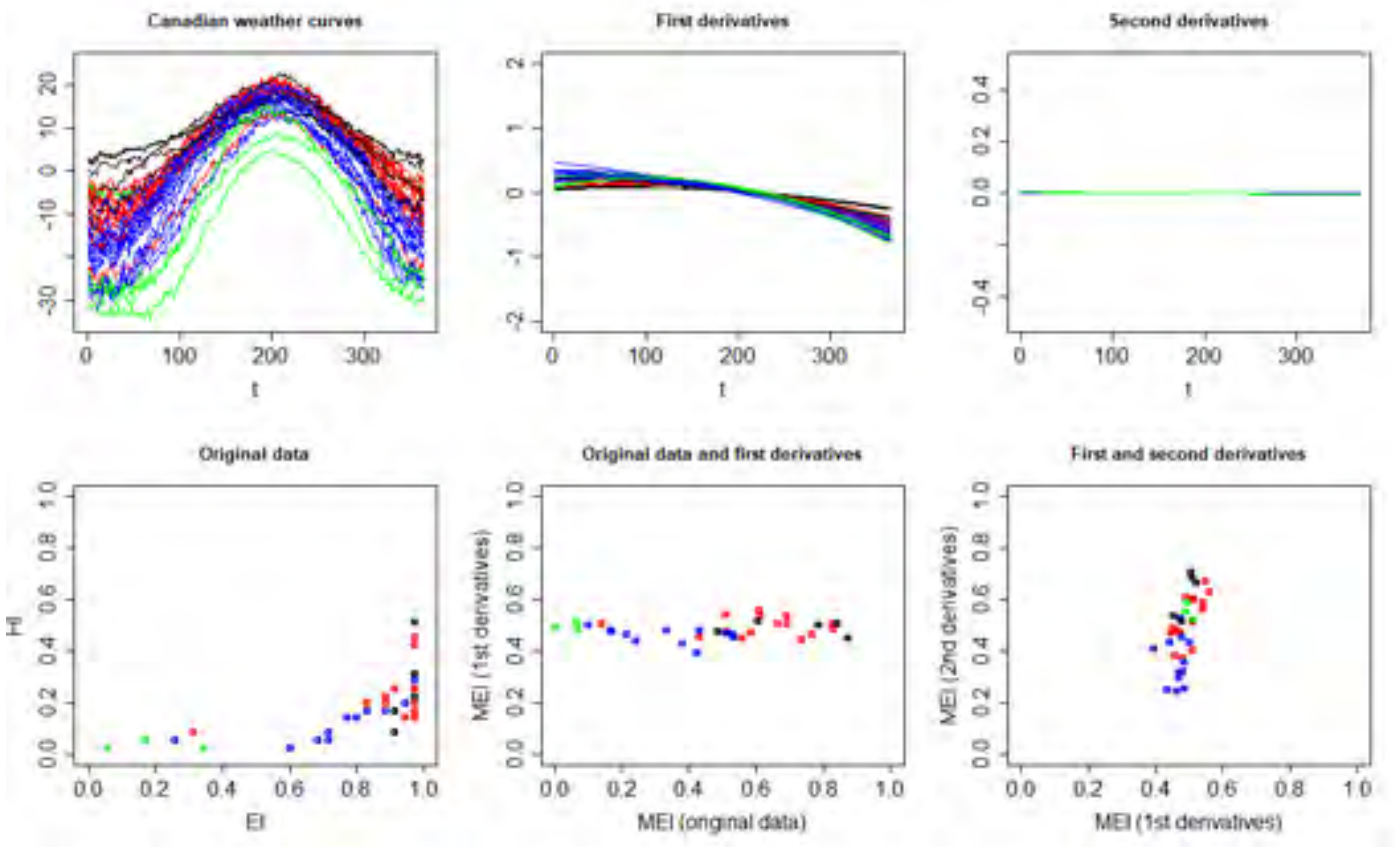
Functional k-means best result. $d\rho$, $\rho = 1e + 08$
RI: 0.742
Time: 1.93641s

Test based k-means best result. Initialization: k-means
RI: 0.698
Time: 0.09893s

Application to a real data set

Canadian Weather Study

'fda' R-package



Application to a real data set

Our best result.

svc.dd2.MEI (Initialization: kernel k-means)

RI: 0.722

Time: 0.05883s

Functional k-means best result. dk, $k = 3$

RI: 0.784

Time: 0.7937s

Test based k-means best result. Initialization: hierarchical clustering

RI: 0.764

Time: 0.12433s

Conclusions

We propose a new procedure for clustering functional data based on multivariate clustering techniques that is:

- **Competitive in terms of performance**
- **Extremely fast**

Future work

- Choose the **number of clusters** with a criterion (for example Silhouette).
- Obtain an **authomatic criterion for the three elections that are the base of our methodology** (like a pre-test).
- Apply the indexes to different contexts.

Main references

Arribas-Gil, A., & Romo, J. (2014). Shape outlier detection and visualization for functional data: The outliergram. *Biostatistics*, *15*(4), 603–619.

Flores, R., Lillo, R., & Romo, J. (2018). Homogeneity test for functional data. *Journal of Applied Statistics*, *45*(5), 868–883.

Franco-Pereira, A. M., & Lillo, R. E. (2019). Rank tests for functional data based on the epigraph, the hypograph and associated graphical representations. *Advances in Data Analysis and Classification*.

López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. *American Statistical Association*, *104*, 327–332.

Martín-Barragán, B., Lillo, R. E., & Romo, J. (2018). Functional boxplots based on half-regions. *Journal of Applied Statistics*, 1088–1103.

Martino, A., Ghiglietti, A., Ieva, F., & Paganoni, A. M. (2019). A k-means procedure based on a mahalanobis type distance for clustering multivariate functional data. *Statistical Methods & Applications*, *28*(2), 301–322.

Pulido, B., Franco-Pereira, A. M., & Lillo, R. E. (2021). Functional clustering via multivariate clustering.

Zambom, A. Z., Collazos, J. A., & Dias, R. (2019). Functional data clustering via hypothesis testing k-means. *Computational Statistics*, *34*(2), 527–549.

Thank You!

belenpulidobravo@gmail.com

rosaelvira.lillo@uc3m.es
albfranc@ucm.es