



A fast epigraph and hypograph-based approach for clustering functional data

Belén Pulido¹ · Alba M. Franco-Pereira^{1,2} · Rosa E. Lillo^{1,3}

Received: 28 January 2022 / Accepted: 11 January 2023 / Published online: 4 February 2023
© The Author(s) 2023

Abstract

Clustering techniques for multivariate data are useful tools in Statistics that have been fully studied in the literature. However, there is limited literature on clustering methodologies for functional data. Our proposal consists of a clustering procedure for functional data using techniques for clustering multivariate data. The idea is to reduce a functional data problem into a multivariate one by applying the epigraph and hypograph indexes to the original curves and to their first and/or second derivatives. All the information given by the functional data is therefore transformed to the multivariate context, being informative enough for the usual multivariate clustering techniques to be efficient. The performance of this new methodology is evaluated through a simulation study and is also illustrated through real data sets. The results are compared to some other clustering procedures for functional data.

Keywords Epigraph · Hypograph · Functional data · B-spline basis · Cluster analysis

1 Introduction

Nowadays, in several fields of study, much of the data collected and analyzed can be considered as functions $x_i(t)$, $i = 1, \dots, n$, $t \in \mathcal{I}$, where \mathcal{I} is an interval in \mathbb{R} . For example, growth, weather variables, the evolution of the market, ... This has been triggered by recent technological developments that enable a large volume of data to be analyzed in a short period of time. Functional Data Analysis (FDA) arises when this information is studied through the analysis of curves or functions. A complete overview of FDA can be found in the monographs of Ramsay and Silverman (2005), and Ferraty and Vieu (2006), while some interesting reviews

of functional data can be found in Horváth and Kokoszka (2012), Hsing and Eubank (2015), and Wang et al. (2016).

The main drawback when working with functional and multivariate data, unlike in one dimension, is the lack of a total order. Thus, a traditional challenge in FDA and in multivariate analysis is to provide an ordering within a sample of curves that enables the definition of order statistics such as ranks and L -statistics. In this sense, (Tukey 1975) introduced the concept of statistical depth that provided a center-outward ordering for multivariate data. Some other definitions can be found in Oja (1983), Liu (1990), and Zuo (2003). This concept was extended to functional data, leading to different definitions of functional depth. See, for example, (Vardi and Zhang 2000; Fraiman and Muniz 2001; Cuevas et al. 2006; Cuesta-Albertos and Nieto-Reyes 2008; López-Pintado and Romo 2009, 2011), and (Sguera et al. 2014).

More recently, (Franco-Pereira et al. 2011) proposed the epigraph and the hypograph indexes in order to measure the “extremality” of a curve with respect to a bunch of curves, and to provide an alternative order to the one given by the statistical depth.

The combination of these two indexes has already been exploited: (Arribas-Gil and Romo 2014) proposed the outliergram for outliers detection, (Martín-Barragán et al. 2018) defined a functional boxplot, and (Franco-Pereira and Lillo 2020) contributed with a homogeneity test for functional

✉ Belén Pulido
belen.pulido@uc3m.es

Alba M. Franco-Pereira
albfranc@ucm.es

Rosa E. Lillo
rosaelvira.lillo@uc3m.es

¹ uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Getafe, Spain

² Department of Statistics and O.R., Universidad Complutense de Madrid, Madrid, Spain

³ Department of Statistics, Universidad Carlos III de Madrid, Getafe, Spain

data. These works show how the epigraph index, the hypograph index and the band depth provide useful information about both the shape and the magnitude of the curves. The main idea of this work is to use the epigraph and the hypograph indexes to reduce a problem in an infinite dimension into a multivariate one in which multivariate clustering techniques can be applied. The joint use of these indexes in the original curves and in their derivatives largely characterizes aspects of the curves in the sample, providing an ordering of the curves from top to bottom or vice versa.

When studying a high volume of data, there is an increased need to classify the data into groups without any extra information, since this classification makes them easier to manipulate. Clustering is one of the most widely used techniques within unsupervised learning, and has been fully studied for multivariate data. Some of the most frequently used procedures are distance-based techniques such as hierarchical clustering (see (Sibson 1973; Defays 1977; Sokal and Michener 1958; Lance and Williams 1967), and (Ward 1963) for different hierarchical clustering procedures) and k-means clustering (introduced by MacQueen (1967)). Taking into account that k-means is probably the most frequently used clustering method in the literature, different variations have been introduced. See (Ben-Hur et al. 2001), and (Dhillon et al. 2004).

Clustering functional data is a challenging problem since it involves working in an infinite dimensional space. Different approaches have been considered in the literature. In Jacques and Preda (2014), the functional clustering techniques are classified into four categories: (1) the raw data methods, which consist of considering the functional data set as a multivariate one and applying clustering techniques for multivariate data (Boullé 2012); (2) the filtering methods, which first apply a basis to the functional data after applying clustering techniques to the obtained data (Abraham et al. 2003; Rossi et al. 2004; Peng et al. 2008), and (Kayano et al. 2010); (3) the adaptive methods, where dimensionality reduction and clustering are performed at the same time (James and Sugar 2003; Jacques and Preda 2013), and (Giacofci et al. 2013), and (Traore et al. 2019); and (4) the distance-based methods, which apply a clustering technique based on distances considering a specific distance for functional data (Tarpey and Kinateder 2003; Ieva et al. 2013), and (Martino et al. 2019). Recent works that cannot be easily classified into any of these categories for clustering functional data are (Romano et al. 2017), which introduced a method for clustering spatially dependent functional data, (Zambom et al. 2019), which proposed a new method by applying k-means, assigning each element to a cluster or another based on a combination of an hypothesis test of parallelism and a test for equality of means, and (Schmutz et al. 2020), which presented a new strategy for clustering functional data based on applying model-based techniques after

a principal component analysis. Based on the previous classification, the methodology proposed in this paper could be considered as both filtering and adaptive, since dimensionality reduction is performed by using the epigraph and the hypograph indexes after applying a basis to the data.

The paper is organized as follows. In Sect. 2, the epigraph and the hypograph indexes are introduced, as well as their relation with the band depth. The methodology for clustering functional data sets based on these indexes is explained in Sect. 3. In Sect. 4, this methodology is examined through an extensive simulation study, and the results are compared to those obtained with some existing procedures for clustering functional data. In Sect. 5, the applicability of our procedure is illustrated through some real data sets. A discussion and some concluding remarks are finally presented in Sect. 6.

2 Preliminaries: The epigraph, the hypograph and the band depth

Let $C(\mathcal{I})$ be the space of continuous functions defined on a compact interval \mathcal{I} . Consider a stochastic process X with sample paths in $C(\mathcal{I})$ and distribution F_X . The graph of a function x in $C(\mathcal{I})$ is $G(x) = \{(t, x(t)), \text{ for all } t \in \mathcal{I}\}$. Then, the epigraph (epi) and the hypograph (hyp) of x are defined as follows:

$$\begin{aligned} \text{epi}(x) &= \{(t, y) \in \mathcal{I} \times \mathbb{R} : y \geq x(t)\}, \\ \text{hyp}(x) &= \{(t, y) \in \mathcal{I} \times \mathbb{R} : y \leq x(t)\}. \end{aligned}$$

Franco-Pereira et al. (2011) defined two indexes based on these two concepts. Given a sample of curves $\{x_1(t), \dots, x_n(t)\}$, the epigraph index of a curve x ($\text{EI}_n(x)$) is defined as one minus the proportion of curves in the sample that are totally included in its epigraph. Analogously, the hypograph index of x ($\text{HI}_n(x)$) is the proportion of curves totally included in the hypograph of x .

$$\begin{aligned} \text{EI}_n(x) &= 1 - \frac{\sum_{i=1}^n I(\{G(x_i) \subseteq \text{epi}(x)\})}{n} = \\ &= 1 - \frac{\sum_{i=1}^n I\{E_{i,x}\}}{n}, \\ \text{HI}_n(x) &= \frac{\sum_{i=1}^n I(\{G(x_i) \subseteq \text{hyp}(x)\})}{n} = \\ &= \frac{\sum_{i=1}^n I\{H_{i,x}\}}{n}, \end{aligned}$$

where $E_{i,x} = \{x_i(t) \geq x(t), \text{ for all } t \in \mathcal{I}\}$, $H_{i,x} = \{x_i(t) \leq x(t), \text{ for all } t \in \mathcal{I}\}$ and $I\{A\}$ is 1 if A is true and 0 otherwise.

Their population versions are given by:

$$\text{EI}(x, F_X) \equiv \text{EI}(x) = 1 - P(G(X) \subseteq \text{epi}(x))$$

$$\begin{aligned}
 &= 1 - P(X(t) \geq x(t), \text{ for all } t \in \mathcal{I}), \\
 \text{HI}(x, F_X) &\equiv \text{HI}(x) = P(G(X) \subseteq \text{hyp}(x)) \\
 &= P(X(t) \leq x(t), \text{ for all } t \in \mathcal{I}).
 \end{aligned}$$

Franco-Pereira et al. (2011) argued that when the curves in the sample are extremely irregular, with many intersections, the modified versions of these indexes are recommended. If \mathcal{I} is considered as a time interval, the modified epigraph index of x ($\text{MEI}_n(x)$) can be defined as one minus the proportion of time the curves are in the epigraph of x , i.e., the proportion of time the curves of the sample are above x . Analogously, the generalized hypograph index of x ($\text{MHI}_n(x)$) can be considered as the proportion of time the curves in the sample are below x .

$$\text{MEI}_n(x) = 1 - \sum_{i=1}^n \frac{\lambda(t \in \mathcal{I} : x_i(t) \geq x(t))}{n\lambda(\mathcal{I})}, \tag{1}$$

$$\text{MHI}_n(x) = \sum_{i=1}^n \frac{\lambda(t \in \mathcal{I} : x_i(t) \leq x(t))}{n\lambda(\mathcal{I})}, \tag{2}$$

where λ stands for Lebesgue’s measure on \mathbb{R} .

Although these definitions are applicable to an arbitrary curve, from now on the curve x will be considered as a curve of the sample, since the methodology proposed here is based on the computation of these indexes on the sample curves. Note that, since the graph of any curve x is contained in its epigraph and its hypograph, this relation holds when $x(t) = x_i(t)$:

$$\begin{aligned}
 \lambda(t \in \mathcal{I} : x_i(t) \geq x(t)) &= \lambda(\mathcal{I}) = \\
 &\lambda(t \in \mathcal{I} : x_i(t) \leq x(t)).
 \end{aligned}$$

Applying this condition to (1) and (2), we obtain

$$\text{MEI}_n(x) = 1 - \left(\sum_{\substack{i=1 \\ x_i \neq x}}^n \frac{\lambda(t \in \mathcal{I} : x_i(t) \geq x(t))}{n\lambda(\mathcal{I})} + \frac{1}{n} \right), \tag{3}$$

and

$$\text{MHI}_n(x) = \sum_{\substack{i=1 \\ x_i \neq x}}^n \frac{\lambda(t \in \mathcal{I} : x_i(t) \leq x(t))}{n\lambda(\mathcal{I})} + \frac{1}{n}. \tag{4}$$

Moreover, if $x(t) \neq x_i(t)$, then

$$\begin{aligned}
 &\lambda(t \in \mathcal{I} : x_i(t) \leq x(t)) + \\
 &\lambda(t \in \mathcal{I} : x_i(t) \geq x(t)) = \lambda(\mathcal{I}).
 \end{aligned}$$

Now, applying this into (4) we can write:

$$\begin{aligned}
 \text{MHI}_n(x) &= \\
 &1 - \frac{1}{n} - \sum_{\substack{i=1 \\ x_i \neq x}}^n \frac{\lambda(t \in \mathcal{I} : x_i(t) \geq x(t))}{n\lambda(\mathcal{I})} + \frac{1}{n} \tag{3} \\
 &\text{MEI}_n(x) + \frac{1}{n}.
 \end{aligned}$$

Finally, the following relation between the two modified versions of the epigraph and the hypograph indexes is obtained, concluding that they are linearly dependent:

$$\text{MHI}_n(x) - \text{MEI}_n(x) = \frac{1}{n}.$$

Note that this equality does not hold in Franco-Pereira and Lillo (2020) because the way in which the data is considered in the homogeneity test differs from the perspective given in this paper.

One may wonder why the epigraph and the hypograph indexes are considered for summarizing the information of a functional sample instead of the band depth (López-Pintado and Romo 2009). In the following, the band depth will be obtained as a combination of the epigraph and the hypograph indexes. Therefore, these indexes are able to summarize the information provided by this depth.

First of all, some definitions are recalled. Consider the band in \mathbb{R}^2 delimited by two curves x_i and x_j as

$$\begin{aligned}
 b(x_i, x_j) &= \{(t, y) \in \mathcal{I} \times \mathbb{R} : \\
 &\min\{x_i(t), x_j(t)\} \leq y \leq \max\{x_i(t), x_j(t)\}\}.
 \end{aligned}$$

Then, the band depth of x of López-Pintado and Romo (2009) ($\text{BD}_n(x)$) is the proportion of bands $b(x_i, x_j)$ determined by two curves x_i, x_j in the sample, containing the whole graph of x .

$$\begin{aligned}
 \text{BD}_n(x) &= \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n I\{G(x) \subset b(x_i, x_j)\}}{\binom{n}{2}} = \\
 &\frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n I\{B_{i,j,x}\}}{\binom{n}{2}}.
 \end{aligned}$$

where

$$\begin{aligned}
 B_{i,j,x} &= \{\min\{x_i(t), x_j(t)\} \leq x(t) \leq \max\{x_i(t), x_j(t)\}, \\
 &\text{for all } t \in \mathcal{I}\}.
 \end{aligned}$$

The Lebesgue measure can also be used instead of the indicator function, obtaining a more flexible definition of the band

depth. The modified band depth of x ($MBD(x)$) is given by:

$$MBD_n(x) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{\lambda(MB_{i,j,x})}{\lambda(\mathcal{I})}}{\binom{n}{2}},$$

where

$$MB_{i,j,x} = \{t \in \mathcal{I} : \min\{x_i(t), x_j(t)\} \leq x(t) \leq \max\{x_i(t), x_j(t)\}\}.$$

Now, the relation between band depth and the indexes is given by

$$BD_n(x) = 2HI_n(x) + \frac{2}{n-1}EI_n(x) - \frac{2n}{n-1}EI_n(x)HI_n(x),$$

and the relation between the modified band depth and the modified epigraph index is given by

$$MBD_n(x) = \frac{1}{n} + 2MEI_n(x) - \frac{2n}{n-1}MEI_n(x)^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=1}^n \left(\frac{\lambda(ME_{i,x} \cap ME_{j,x})}{\lambda(\mathcal{I})} - \frac{\lambda(ME_{i,x})\lambda(ME_{j,x})}{\lambda(\mathcal{I})} \right),$$

where $ME_{i,x} = \{t \in \mathcal{I} : x_i(t) \geq x(t)\}$. The proof of the first equality is below. The proof of the second one can be found in Arribas-Gil and Romo (2014), but note that they omit “1-” in the definition of the MEI. Also note that, in order to obtain these equations, x is considered to belong to the sample of curves.

As stated before, the band depth, the epigraph and the hypograph indexes can be written as:

$$BD_n(x) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n I\{B_{i,j,x}\}}{\binom{n}{2}},$$

$$EI_n(x) = 1 - \frac{1}{n} \sum_{i=1}^n I\{E_{i,x}\}, \text{ and}$$

$$HI_n(x) = \frac{1}{n} \sum_{i=1}^n I\{H_{i,x}\}.$$

Note that $\sum_{i=1}^{n-1} \sum_{j=i+1}^n I\{B_{i,j,x}\}$ represents the number of bands between two curves that can be obtained taking any two curves in the sample different from x , plus the number of bands that include x . $\sum_{i=1}^n I\{E_{i,x}\}$ represents the number of curves that lie above x plus one, and $\sum_{i=1}^n I\{H_{i,x}\}$ stands for the number of curves that lie below x plus one.



Fig. 1 Scheme of the EHyClus method

It holds that

$$I\{B_{i,j,x}\} = I\{E_{i,x}\}I\{H_{j,x}\} + I\{H_{i,x}\}I\{E_{j,x}\},$$

if $x \neq x_i, x \neq x_j$ and $I\{B_{i,j,x}\} = 1$ otherwise.

Thus,

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^n I\{B_{i,j,x}\} = \sum_{i=1}^n I\{E_{i,x}\} \sum_{i=1}^n I\{H_{i,x}\} - \sum_{i=1}^n I\{E_{i,x}\} - \sum_{i=1}^n I\{H_{i,x}\} + n.$$

Now, since $\sum_{i=1}^n I\{E_{i,x}\} = n(1 - EI_n(x))$, and $\sum_{i=1}^n I\{H_{i,x}\} = nHI_n(x)$, we have

$$BD_n(x) = 2HI_n(x) + \frac{2}{n-1}EI_n(x) - \frac{2n}{n-1}EI_n(x)HI_n(x).$$

3 Clustering functional data through the epigraph and the hypograph indexes

The proposed methodology for clustering functional data is a four-step method, as illustrated in Fig. 1. In what follows, we will refer to this method as EHyClus.

Step 1 (S1) consists of smoothing the data. This is recommended since the amount of data upon which the process is based precludes abrupt changes in value. For this reason, it is common to smooth the data when working with curves. Cubic B-spline bases have been used, but any other functional basis could have been applied. In this case, since the first and the second derivatives of the data are considered, taking a cubic B-spline basis is the most natural option. In order to choose the optimal number of bases, a sensitivity study was carried out. The corresponding results are shown in the Supplementary Material, Section 1, and depend on the data sets one may consider, as happens in almost all the studies in the literature. Here, all the data sets explained in Sect. 4 have been taken. They show that particular changes in the number of bases do not play a crucial role in the results. Nonetheless, this analysis highlights that the best ones can be obtained with a number of bases between 30 and 40. After the

data set is transformed, the second step (S2) is to apply the epigraph and the hypograph indexes (and their generalized versions) to the basis transformed data, as well as to their derivatives, obtaining a multivariate data set. As explained in Sect. 2, the modified epigraph and hypograph indexes are linearly dependent. Because of that, MHI will be discarded since it will not provide “extra information”.

From now on, the term *curves* will refer to the smoothed ones, and *data* will refer to the complete data set with curves and the first and second derivatives. Then, different subsets of the data will be taken to apply multivariate clustering techniques (S3). Finally, the fourth step (S4) consists of obtaining a final clustering partition in a previously fixed number of groups. In general, as explained in Rendón et al. (2011), clustering validity approaches can be divided into two categories: external and internal criteria. The first type of methods requires the ground truth to obtain a final result, and the second one uses some other intrinsic information of the data to achieve a solution. For evaluating the goodness of the classification, in this work, three different external validation strategies will be applied: Purity, F-measure and Rand Index (RI), which are fully explained in Manning et al. (2009), and Rendón et al. (2011).

Purity is the proportion of elements that were classified correctly. The F-measure is the harmonic mean of the precision and the recall values for each cluster. The precision of a cluster is the same as its purity coefficient. The recall of a cluster is the proportion of observations classified as a given class in a correct way. The Rand Index can be viewed as a measure of the percentage of correct decisions made by the algorithm. All these indexes provide values in [0, 1] and verify that the higher the value, the better the classification. The Adjusted Rand Index (ARI) could be considered instead of the ARI. The ARI is a corrected version of the RI that rectifies the fact that good results were obtained by chance. In this work we have considered RI instead of ARI because ARI is not always a number between 0 and 1 and thus, it has a different scale than Purity and F-measure, which are the other validity measures considered here.

In steps S1 to S3, the following procedure is carried out to obtain the clustering partitions to which the external validation criteria explained above may be applied.

A functional data problem is converted into a multivariate one by applying the indexes. Thus, some “information” is lost. Applying these indexes, not only to the original curves but also to their first and second derivatives, allows one to take advantage of the shape/magnitude/amplitude of the curves in the sample. It seems clear that these three attributes play an important role in functional clustering, and these indexes can provide a great deal of information in this regard.

Thereafter, data and indexes are combined to obtain a data set where a multivariate clustering technique is later applied.

Table 1 Representation and description of the combinations of data and indexes

Representation	Description
_.EIHI = (EI, HI)	The epigraph and the hypograph index on the original curves.
dd2.MEI = (dMEI, d2MEI)	The generalized epigraph index applied to the first and second derivatives.
_.dd2.EIHIMEI = (EI, HI, MEI, dEI, dHI, dMEI, d2EI, d2HI, d2MEI)	The epigraph, the hypograph and the generalized version of the epigraph applied to all the data.

Since considering all the possible combinations between data and indexes without fixing any condition leads to a vast number of options, 18 different combinations of data and indexes are contemplated.

Now, the notation used for presenting the results in every table is explained. The combinations sets are represented as (b).(c) where (b) stands for the data, with ‘_’ representing the curves, ‘d’ first derivatives and ‘d2’ second derivatives, and (c) represents the indexes. The 18 different combinations come from applying all the indexes in (c) to all the data in (b). All these possible combinations are listed in the Supplementary Material, but some examples are shown in Table 1.

When the curves are extremely irregular, the epigraph and the hypograph indexes may take values very close to 1 and 0, respectively. This fact causes the indexes to lose discriminatory capacity to differentiate between clusters and also induces computational problems and “ill-conditioned” problems in which singular or near-singular matrices are involved. Combinations leading to these kinds of problems have been avoided in our study.

Finally, the multivariate clustering technique to be applied has to be chosen from among the following: hierarchical clustering using different criteria for calculating similarity between clusters (single linkage, complete linkage, average linkage, centroid linkage), and Ward’s method; k-means and its different versions using a feature space induced by a kernel, such as kernel k-means (kkmeans), and some other approaches with fewer restrictions in the structure of the data, as are spectral clustering (spc) and support vector clustering (svc).

For hierarchical methods, the Euclidean distance has been considered. On the other hand, when implementing k-means, Euclidean and Mahalanobis distances have been used. In order to apply the Mahalanobis distance, data is rescaled using the Cholesky decomposition of the variance matrix before running k-means with the Euclidean distance (see Redko et al. 2019). Moreover, when the method uses a kernel space, three different kernels have been applied: Gaussian, polynomial and linear.

4 Simulation study

With the current EHyClus version, all the cluster partitions derived from applying a clustering technique to the 18 combinations are computed. Then, the external validation criteria explained in Sect. 3 are employed to choose the best one.

A simulation study has been carried out in order to evaluate the performance of EHyClus and to compare it to some other existing methodologies in the literature. Specifically, EHyClus has been compared to five different methodologies fully explained in Yassouridis and Leisch (2017), and it has also been contrasted with two recent ones: the distance-based k-means procedure (functional k-means) introduced by Martino et al. (2019) and the test-based k-means from Zambom et al. (2019).

The first methodology in Yassouridis and Leisch (2017), baseclust, consists in smoothing the data and applying k-means to the resultant data set. It can be easily computed in R by fitting a B-spline basis and then applying k-means. The second one, fitclust, also smooths the data but, instead of applying k-means, it assumes data coming from Gaussian distributions and applies the EM algorithm to the data. The third one, distclust, is a distance approach based on Karhunen-Loève (K-L) expansions. Method four, modelcf, is again based on expansions, but assigns a curve to its closest subspace. These three methods are provided in the R-package ‘fancy’. Finally, the fifth method, curvclust, applies the EM algorithm to a wavelet basis, and it is provided in the R-package ‘curvclust’.

Martino et al. (2019) proposes a k-means algorithm with a generalized Mahalanobis distance for functional data, d_ρ , which was previously defined in Ghiglietti and Paganoni (2017), and where the value of ρ has to be set in advance.

Zambom et al. (2019) propose a methodology based on a hypothesis test applying k-means, where they consider four different possibilities for initializing the clusters: at random, choosing one iteration of k-means, choosing one iteration of a hierarchical method (Ward’s method with Euclidean distance) or taking one iteration of k-means++ (Vassilvitskii and Arthur 2006).

Each simulated scenario is composed of previously known groups proceeding from different processes. Each scenario is simulated 100 times. The eight methodologies previously described are applied each time. In each iteration, the average of each validation criterion: Purity, F-measure, Rand Index, and the mean execution time of all the methods are calculated. Thus, one result from each of these four values is obtained for each evaluated clustering partition. This section has been divided into three parts: The first two deal with the number of clusters and the last one is a summary of the results obtained in all the simulated scenarios.

4.1 Simulation study A: Two clusters

Three different simulation groups of scenarios will be studied in this section. The first one consists of eight different scenarios previously considered in Flores et al. (2018), and Franco-Pereira and Lillo (2020). The second one consists of two scenarios introduced in Martino et al. (2019), and the third one is based on the data presented in Tucker et al. (2013).

First, data simulated in the first group of scenarios is described: Consider eight functional samples defined in $[0, 1]$, which have continuous trajectories in such interval and which are the realizations of a stochastic process X . Each curve has 30 equidistant observations in the interval $[0, 1]$. We generate 100 functions: 50 from Model 1 and 50 from Model i , $i = 2, \dots, 9$, obtaining eight different functional data sets.

Model 1. This is the reference group for all the scenarios. It is generated by a Gaussian process

$$X_1(t) = E_1(t) + e(t),$$

where $E_1(t) = 30t^{\frac{3}{2}}(1-t)$ is the mean function and $e(t)$ is a centered Gaussian process with covariance matrix

$$\text{Cov}(e(t_i), e(t_j)) = 0.3 \exp\left(-\frac{|t_i - t_j|}{0.3}\right).$$

The rest of the models are obtained from the first one by perturbing the generation process.

The first three models contain changes in the mean, while the covariance matrix does not change. Changes in the mean are presented in increasing order from Model 2 to Model 4.

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. $X_4(t) = 30t^{\frac{3}{2}}(1-t) + 1 + e(t)$. The next two samples are obtained by multiplying the covariance matrix by a constant.

Model 5. $X_5(t) = 30t^{\frac{3}{2}}(1-t) + 2e(t)$.

Model 6. $X_6(t) = 30t^{\frac{3}{2}}(1-t) + 0.25e(t)$.

Model 7. This set is obtained by adding to $E_1(t)$ a centered Gaussian process $h(t)$ whose covariance matrix is given by $\text{Cov}(h(t_i), h(t_j)) = 0.5 \exp\left(-\frac{|t_i - t_j|}{0.2}\right)$. In this case $X_7(t) = 30t^{\frac{3}{2}}(1-t) + h(t)$.

The next two samples are obtained by a different mean function.

Model 8. $X_8(t) = 30t(1-t)^2 + h(t)$.

Model 9. $X_9(t) = 30t(1-t)^2 + e(t)$.

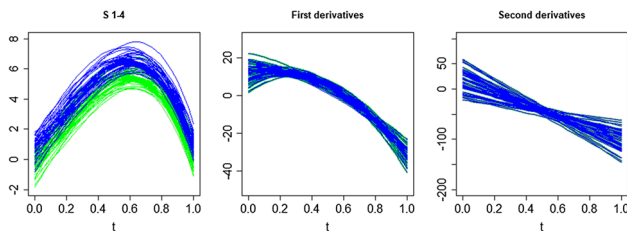


Fig. 2 A sample generated from *S 1-4*. Original data (left panel), first and second derivatives curves (center and right panels respectively)

From now on, the eight resulting data sets will be referred to as scenarios.

The data is smoothed by using a cubic B-spline basis in order to remove noise and to use data derivatives (S1 in Fig. 1). Then, each scenario is simulated 100 times and, each time, EHyClus is applied (S1-S4 in Fig. 1). The mean Purity, F-measure, Rand index (RI) and execution time (ET) are used as criteria to choose the best model. Table 2 presents these results for the top 10 combinations obtained for the scenario *S 1-4*. The rest of the tables are deferred to the Supplementary Material. In these tables, each row represents a description of the process carried out for the 100 realizations, denoted by (a).(b).(c) where (a) represents the name of the considered strategy: a hierarchical method, k-means, support vector clustering, kernel k-means or spectral clustering, and (b).(c) represents the elections of the data and indexes, as represented in Table 1, where (b) is the name of the employed data and (c) the indexes applied on that data.

The smoothed functions and the first and second derivatives of curves generated from *S 1-4* are shown in Fig. 2. These figures confirm that the original curves discriminate between the two clusters better. This can also be noticed when applying the indexes (see Figs. 3 and 4). Regarding the formulation of Model 1 and Model 4, their derivatives are the same. Thus, in this scenario, considering the derivatives does not provide any extra information. Nevertheless, as any previous information is considered, they are included in the process. At first, one could think this is misleading the model, but the combination to be used will be the one with the highest validation criterion. In this case, the top 10 combinations when applying EHyClus are shown in Table 2, and all of them contain the original curves. Hence, the combinations that do not provide information to distinguish between groups are discarded in the top partitions.

In Figs. 3 and 4 the represented data sets have two columns obtained from combining data and indexes, but when applying EHyClus, the higher data set can have up to nine of these combinations.

Finally, the best combination here turns out to be the one obtained when applying k-means on the epigraph, the hypograph and the modified epigraph indexes of the original data (kmeans._.EIHIMEI).

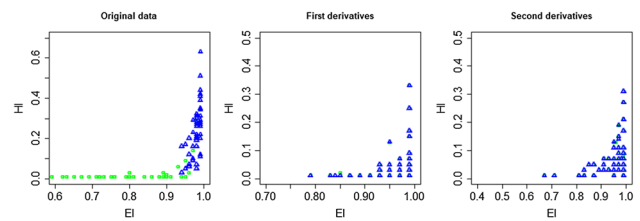


Fig. 3 Scatter plots of the epigraph index (EI) and the hypograph index (HI) of the original data simulated from Model 1 and 4 (left panel), first derivatives (center panel) and second derivatives (right panel)

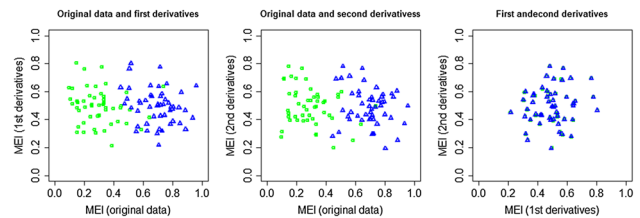


Fig. 4 A sample generated from *S 1-4*. Scatter plots of different combinations of MEI. Original data and first derivatives (left panel), original data and second derivatives (center panel) and first and second derivatives (right panel)

Table 2 Top 10 mean results for *S 1-4* considering Euclidean distance (gray), Mahalanobis distance (pink), a polynomial kernel (blue), kernel k-means for initialization (green) and k-means for initialization (orange)

	Purity	Fmeasure	RI	Time
kmeans._.EIHIMEI	0.929	0.867	0.868	0.00213
kmeans._.EIHIMEI	0.929	0.867	0.868	0.00256
svc._.EIHIMEI	0.929	0.866	0.867	0.00424
kmeans._d.EIHIMEI	0.929	0.866	0.867	0.00252
kmeans._d.EIHIMEI	0.929	0.866	0.867	0.00282
kkmeans._d.EIHIMEI	0.929	0.866	0.867	0.01086
svc._d.EIHIMEI	0.928	0.865	0.866	0.00565
svc._d.EIHIMEI	0.928	0.864	0.866	0.00535
svc._.EIHIMEI	0.928	0.865	0.866	0.00421
kkmeans._.EIHIMEI	0.926	0.861	0.863	0.00934

These results are first compared to those obtained from applying functional k-means and test-based k-means techniques, which are shown in Tables 3 and 4. In the case of functional k-means, each row represents a different distance between the generalized Mahalanobis distance ($d\rho$), the truncated Mahalanobis distance (dk) and the Euclidean distance (L^2). For test-based k-means, each row stands for a different initialization. When considering the first procedure, L^2 distance provides the best RI, 0.847, which is close to those results obtained using methods with a small value of ρ . Nevertheless, when considering ρ equal to 0.02 the execution time is double the one for L^2 distance.

When applying test-based k-means, the method is not able to distinguish between the two groups, since all the metrics lead to values close to 0.5 in all cases.

Table 3 Mean values of Purity, F-measure, Rand Index and execution time for the functional k -means procedure (Martino et al. 2019) with truncated Mahalanobis distance, generalized Mahalanobis distance and L^2 distance to simulated data from S 1-4

	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
$d\rho, \rho = 100$	0.832	0.777	0.773	2.15567
$d\rho, \rho = 1e + 08$	0.863	0.772	0.772	1.87334
$dk, k = 2$	0.791	0.713	0.713	0.84618
$dk, k = 3$	0.724	0.646	0.643	0.71322

Table 4 Mean values of Purity, F-measure, Rand Index and execution time for the test-based k -means procedure (Zamboni et al. 2019) with four different initializations to simulated data from S 1-4

	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

Table 5 Mean values of Purity, F-measure, Rand Index and execution time for the five different procedures described in Yassouridis and Leisch (2017) to simulated data from S 1-4

	Purity	Fmeasure	RI	Time
baseclust	0.923	0.856	0.857	0.00625
fitfclust	0.691	0.565	0.569	0.22153
distclust	0.908	0.833	0.833	4.21785
modelcf	0.901	0.818	0.819	0.04989
curvclust	0.914	0.847	0.847	0.55814

Moreover, results from the five different methodologies considered in Yassouridis and Leisch (2017) appear in Table 5.

The best partition in terms of metrics and ET is the one given by EHyClus (RI=0.868, ET=0.00213), followed then by baseclust (RI=0.857, ET=0.00625). The rest of methodologies do not provide competitive results compared to EHyClus. Thus, the good results achieved in terms of the three different metrics and execution time is key to stating that the proposed methodology is a very good alternative to the existing ones for clustering functional data.

For the other seven scenarios, whose results appear in the Supplementary Material, EHyClus obtains good results in terms of metrics and execution times. These results in terms of RI and ET are commented in Sect. 4.3.

On the other hand, in order to extend the simulation study, the functional data in Martino et al. (2019) is considered. These data sets were specially created for testing clustering techniques for functional data, and consist of two functional samples defined in $[0, 1]$, with continuous trajectories that are generated by independent stochastic processes in $L^2(I)$. Each curve has 150 equidistant observations in the interval $[0, 1]$. We generate 100 functions, 50 from Model 10 and 50 from Model $i, i = 11, 12$, obtaining two different functional samples. These two scenarios will be referred to as $S 10-11$ and $S 10-12$, respectively.

The three different models defined for these simulations are specified below.

Model 10. The first 50 functions are generated as follows:

$$X_{10}(t) = E_2(t) + \sum_{k=1}^{100} Z_k \sqrt{\rho_k} \theta_k(t),$$

where $E_2(t) = t(1 - t)$ is the mean function, $\{Z_k, k = 1, \dots, 100\}$ are independent standard normal variables, and $\{\rho_k, k \geq 1\}$ is a positive real numbers sequence defined as

$$\rho_k = \begin{cases} \frac{1}{k+1} & \text{if } k \in \{1, 2, 3\}, \\ \frac{1}{(k+1)^2} & \text{if } k \geq 4, \end{cases}$$

in such a way that the values of ρ_k are chosen to decrease faster when $k \geq 4$ in order to have most of the variance explained by the first three principal components. The sequence $\{\theta_k, k \geq 1\}$ is an orthonormal basis of $L^2(I)$ defined as

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

where $I_A(t)$ stands for the indicator function of set A . The next two models are defined in the same way, but in each case changing the term which is added to $E_2(t)$ in Model 10. Moreover, the standard normal variables generated for the following two models differ from those of the last model.

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

$$E_4(t) = E_2(t) + \sum_{k=4}^{100} \sqrt{\rho_k} \theta_k(t).$$

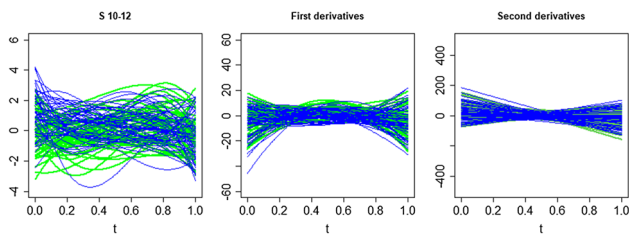


Fig. 5 A sample generated from *S 10-12*. Original data (left panel), first and second derivatives curves (center and right panels respectively)

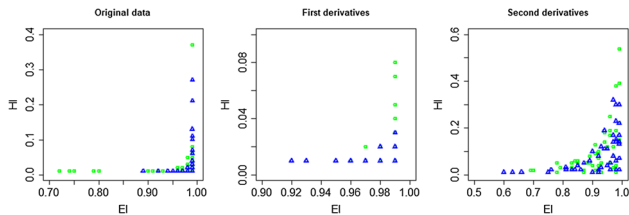


Fig. 6 Scatter plots of the epigraph index (EI) and the hypograph index (HI) of the original data simulated from Model 10 and 12 (left panel), first derivatives (second panel) and second derivatives (right panel)

As before, data is considered after being smoothed with a cubic B-spline basis in order to remove noise and to be able to use its first and second derivatives.

The corresponding smoothed curves simulated from *S 10-12* often cross, as do their derivatives (Fig. 5). When applying the epigraph and the hypograph indexes to these sets of curves, again, the difference between groups is negligible (Fig. 6). Nevertheless, looking at Fig. 7, when applying the MEI, the difference between groups is now much clearer. The best result from EHyClus (see Table 6) is achieved by applying kernel k-means with a polynomial kernel on the generalized epigraph index of the first and second derivatives, obtaining a RI of 0.919. Moreover, when applying the same technique with the same set of data (first and second derivatives) but adding the epigraph and hypograph indexes, the same RI is obtained. This combination is not graphically shown since six different variables are involved. This particular case exemplifies that when considering EI, HI and MEI at the same time, in general one gets equal or better results than when working with any of their subsets. Thus, it would not be necessary to try all the combinations of indexes, but only the union of the three. Nevertheless, in order to see the differences, all the possibilities previously explained have been considered.

These results are compared to those obtained by applying the functional k-means procedure (Table 7). In this case, the best distance is the Mahalanobis distance with a big value of ρ , $\rho = 1e + 08$, getting a RI of 0.718, which is small compared to the value of 0.919 obtained with EHyClus. In addition, EHyClus spends 0.00423s for its best combination, while functional k-means spends 7.9055s for the best election of ρ .

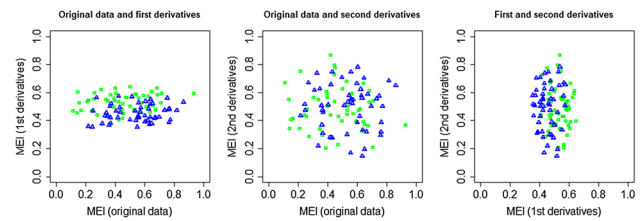


Fig. 7 A sample generated from *S 10-12*. Scatter plots of different combinations of MEI. Original data and first derivatives (left panel), original data and second derivatives (center panel) and first and second derivatives (right panel)

Table 6 Top 10 mean results for S 10-12 considering Euclidean distance (gray), a polynomial kernel (blue), kernel k-means for initialization (green) and k-means for initialization (orange)

	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
kmeans.dd2.EIHIMEI	0.956	0.916	0.917	0.00100
kmeans.dd2.MEI	0.956	0.916	0.917	0.00082
svc.dd2.EIHIMEI	0.956	0.916	0.914	0.00273
ward.D2.dd2.EIHIMEI	0.939	0.889	0.888	0.00015
ward.D2.dd2.MEI	0.939	0.889	0.888	0.00009
average.dd2.MEI	0.921	0.882	0.874	0.00009

Table 7 Mean values of Purity, F-measure, Rand Index and execution time for the functional k-means procedure (Martino et al. 2019) with truncated Mahalanobis distance, generalized Mahalanobis distance and L^2 distance to simulated data from S 10-12

	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
$d\rho, \rho = 0.001$	0.548	0.503	0.502	9.9013
L^2	0.547	0.502	0.502	1.3529
$dk, k = 3$	0.543	0.503	0.501	1.1880
$dk, k = 2$	0.541	0.501	0.500	0.53262

When applying test-based k-means to this type of simulated data (see Table 8) and the five methods described in Yassouridis and Leisch (2017) (Table 9), none of them are able to distinguish between groups, obtaining values close to 0.5 for all metrics.

To sum up, EHyClus leads to the best results in terms of metrics and execution time, being 0.2 points better regarding RI, and being faster compared to functional k-means, which is second in terms of the considered metrics.

Finally, when dealing with functional data, it is important not only to investigate magnitude and amplitude changes in the generated curves, but also those changes in phase due to possible misalignment of data. For that, the models in

Table 8 Mean values of Purity, F-measure, Rand Index and execution time for the test-based k -means procedure (Zambom et al. 2019) with four different initializations to simulated data from S 10-12

	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

Table 9 Mean values of Purity, F-measure, Rand Index and execution time for the five different procedures described in Yassouridis and Leisch (2017) to simulated data from S 10-12

	Purity	Fmeasure	RI	Time
baseclust	0.539	0.501	0.499	0.02385
fitfclust	0.520	0.492	0.496	3.23131
distclust	0.544	0.508	0.501	32.29883
modelcf	0.530	0.492	0.497	0.05702
curvclust	0.545	0.515	0.502	1.24491

Tucker et al. (2013) are taken. Consider a functional sample of continuous trajectories defined in the interval $[-6, 6]$, and which are the realizations of a stochastic process. Each curve has 150 equidistant observations in the interval $[-6, 6]$. We generate 42 functions: 21 for Model 13 and 21 for Model 14.

Model 13. This model has the variation in phase. It is generated as $X_{13}(t) = z_1 \exp(-\frac{(t-a)^2}{2})$ where z_1 and a are independent normal variables $\mathcal{N}(1, 0.05^2)$ and $\mathcal{N}(0, 1.25^2)$, respectively.

Model 14. This model does not present variation in phase and it is combined with the previous one in order to check whether the new methodology is able to distinguish between the two populations. This data is generated considering the equation $X_{14}(t) = z_2 \exp(-\frac{(t-2.5)^2}{2})$ where z_2 follows a $\mathcal{N}(1.5, 0.1^2)$.

The combination of these two models will be referred to as *S 13-14*. As before, a cubic B-spline basis is considered for smoothing the data and to calculate its derivatives. Figures 8 and 9 represent the curves and the groups when applying the modified epigraph index to the different data available. In this case, as the curves present phase variation, there are many intersections between them. This is why the modified versions of the indexes are a must in order to summarise the information.

Top 10 combinations of EHyClus appear in Table 10. There, all the combinations use the modified epigraph index, because it is the one providing the most informative insights. These results are compared to those obtained from the other seven approaches and appear in Tables 11, 12 and 13.

EHyClus and fitfclust are the methodologies with higher RI. Nevertheless, the first one obtains an ET for the best com-

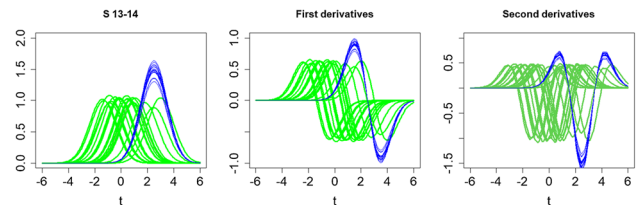


Fig. 8 A sample generated from *S 13-14*. Original data (left panel), first and second derivatives curves (center and right panels respectively)

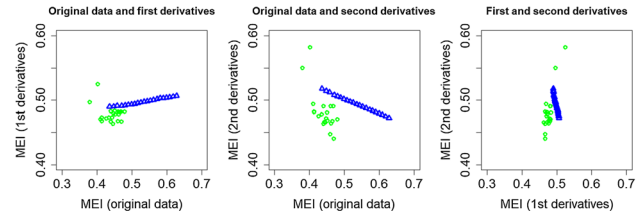


Fig. 9 A sample generated from *S 13-14*. Scatter plots of different combinations of MEI. Original data and first derivatives (left panel), original data and second derivatives (center panel) and first and second derivatives (right panel)

Table 10 Top 10 mean results for S 13-14 considering Euclidean distance (gray), a polynomial kernel (blue), kernel k -means for initialization (green), and k -means for initialization (orange)

	Purity	Fmeasure	RI	Time
complete.d.EIHIMEI	0.952	0.905	0.907	0.00033
kkmeans.d.EIHIMEI	0.952	0.905	0.907	0.00754
kmeans.d.EIHIMEI	0.952	0.905	0.907	0.00365
spc.d.EIHIMEI	0.952	0.905	0.907	0.02359
spc.dd2.MEI	0.952	0.905	0.907	0.03001
svc.d.EIHIMEI	0.952	0.905	0.907	0.00644
svc.d.EIHIMEI	0.952	0.905	0.907	0.00643
ward.D2.d.EIHIMEI	0.952	0.905	0.907	0.00035
ward.D2.dd2.EIHIMEI	0.952	0.905	0.907	0.00037
ward.D2.dd2.MEI	0.952	0.905	0.907	0.00035

bination equal to 0.00033, while the second spends 1.1612 s. Thus, even though there could be some misalignment in one of the classes, EHyClus is able to distinguish between the two populations, and it improves the computational cost with respect to the existing approaches.

4.2 Simulation Study B: More than two clusters

In this case, four data sets will be considered. Three of them with three groups and one with six.

First, three different scenarios coming from three different groups are taken into account. This simulation study previously appeared in Zambom et al. (2019). Each data set is composed of 150 curves, 50 of them belonging to each of the three clusters. Each curve has 100 equidistant points defined in the interval $[0, \frac{\pi}{3}]$.

Table 11 Mean values of Purity, F-measure, Rand Index and execution time for the functional k -means procedure (Martino et al. 2019) with truncated Mahalanobis distance, generalized Mahalanobis distance and L^2 distance to simulated data from S 13-14

	Purity	Fmeasure	RI	Time
$d\rho, \rho = 0.001$	0.929	0.862	0.864	1.46303
$d\rho, \rho = 0.02$	0.9289	0.862	0.864	0.80187
$d\rho, \rho = 1$	0.929	0.862	0.864	1.45995
$d\rho, \rho = 1e + 08$	0.929	0.862	0.864	1.44883
L^2	0.929	0.862	0.864	0.17995
$dk, k = 3$	0.905	0.822	0.824	0.19996
$d\rho, \rho = 100$	0.714	0.634	0.582	1.44405

Table 12 Mean values of Purity, F-measure, Rand Index and execution time for the test-based k -means procedure (Zamboni et al. 2019) with four different initializations to simulated data from S 13-14

	Purity	Fmeasure	RI	Time
random	0.905	0.822	0.824	0.34584
kmeans	0.905	0.822	0.824	0.17640
hclust	0.905	0.822	0.824	0.25854
kmeans ++	0.833	0.724	0.715	0.46918

Table 13 Mean values of Purity, F-measure, Rand Index and execution time for the five different procedures described in Yassouridis and Leisch (2017) to simulated data from S 13-14

	Purity	Fmeasure	RI	Time
baseclust	0.929	0.862	0.864	0.00938
fitfclust	0.952	0.905	0.907	1.1612
distclust	0.928	0.862	0.864	12.1398
modelcf	0.929	0.862	0.864	0.03553
curvclust	0.929	0.862	0.864	0.3678

Each scenario is composed of 50 functions from three different models, each of them of the form:

$$X(t) = Y(t) + \epsilon$$

The nine different models are defined as follows:

Model 15. $X_{13}(t) = \frac{1}{1.3} \sin(1.3t) + t^3 + a + 0.3 + \epsilon_1$

Model 16. $X_{14}(t) = \frac{1}{1.2} \sin(1.3t) + t^3 + a + 1 + \epsilon_1$

Model 17. $X_{15}(t) = \frac{1}{4} \sin(1.3t) + t^3 + a + 0.2 + \epsilon_1$

Model 18. $X_{16}(t) = \sin(1.5\pi t) + \cos(\pi t^2) + b + 1.1 + \epsilon_1$

Model 19. $X_{17}(t) = \sin(1.7\pi t) + \cos(\pi t^2) + b + 1.5 + \epsilon_1$

Model 20. $X_{18}(t) = \sin(1.9\pi t) + \cos(\pi t^2) + b + 2.2 + \epsilon_1$

Model 21. $X_{19}(t) = \frac{1}{1.8} \exp(1.1t) - t^3 + a + \epsilon_2$

Model 22. $X_{20}(t) = \frac{1}{1.7} \exp(1.4t) - t^3 + a + \epsilon_2$

Model 23. $X_{21}(t) = \frac{1}{1.5} \exp(1.5t) - t^3 + a + \epsilon_2$

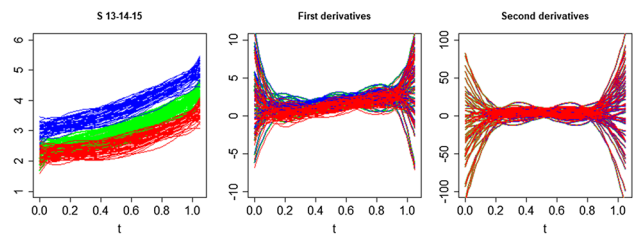


Fig. 10 A sample generated from S 15-16-17. Original data (left panel), first and second derivatives curves (center and right panels respectively)

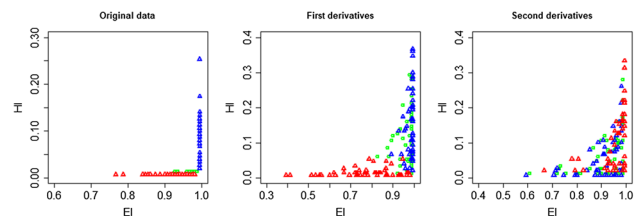


Fig. 11 Scatter plots of the epigraph index (EI) and the hypograph index (HI) of the original data simulated from Model 15, 16 and 17 (left panel), first derivatives (center panel) and second derivatives (right panel)

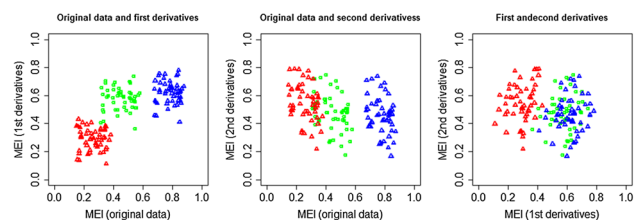


Fig. 12 A sample generated from S 15-16-17. Scatter plots of different combinations of MEI. Original data and first derivatives (left panel), original data and second derivatives (center panel) and first and second derivatives (right panel)

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

Hence, $S 15-16-17$ is composed of 50 functions from Model 15, 50 functions from Model 16 and 50 from Model 17. $S 18-19-20$ are created in an analogous way.

Data considered from $S 15-16-17$ is shown in Fig. 10, where it is clear that the functions in green and red intertwine a lot. When applying the epigraph and the hypograph indexes (Fig. 11) it is possible to identify two well distinguished bunches of curves, because the green curves overlap the red and blue ones. Nevertheless, when considering the modified epigraph index (Fig. 12), it seems that the differences between the three groups are much more evident.

EHyClus with k -means and both Euclidean and Mahalanobis distances for the original data, first and second derivatives with the generalized epigraph index leads to the same and best result in Table 14 (RI=0.977 and ET almost 0.003 s). When considering functional k -means, the best method in Table 15 is the one with a small value of ρ ,

Table 14 Top 10 mean results for S 15-16-17 considering Euclidean distance (gray), Mahalanobis distance (pink), kernel k-means for initialization (green) and k-means for initialization (orange)

	Purity	Fmeasure	RI	Time
kmeans_dd2.MEI	0.988	0.967	0.985	0.00274
kmeans_dd2.MEI	0.988	0.977	0.985	0.00205
svc_dd2.MEI	0.988	0.977	0.985	0.00449
kmeans..d.MEI	0.988	0.976	0.984	0.00175
kmeans..d.MEI	0.988	0.976	0.984	0.00179
svc..d.MEI	0.988	0.976	0.984	0.00433
svc..d.MEI	0.988	0.975	0.984	0.00430
svc_dd2.MEI	0.982	0.969	0.979	0.00429
ward.D2..d.MEI	0.979	0.961	0.974	0.00022
average..d.MEI	0.975	0.959	0.972	0.00027

Table 15 Mean values of Purity, F-measure, Rand Index and execution time for the functional *k*-means procedure (Martino et al. 2019) with truncated Mahalanobis distance, generalized Mahalanobis distance and L^2 distance to simulated data from S 15-16-17

	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
$d\rho, \rho = 100$	0.682	0.662	0.754	5.67288
$dk, k = 2$	0.659	0.606	0.735	1.54468
$d\rho, \rho = 1e + 08$	0.719	0.590	0.725	7.95828
$dk, k = 3$	0.605	0.548	0.695	1.59957

Table 16 Mean values of Purity, F-measure, Rand Index and execution time for the test-based *k*-means procedure (Zambom et al. 2019) with four different initializations to simulated data from S 15-16-17

	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

$\rho = 0.001$, (RI=0.928 and ET=6.50802 s). And when applying test-based *k*-means, the best result in Table 16 is obtained when initializing the process with k-means++ (RI=0.944 and ET=0.99653). Results from the other five methodologies appear in Table 17, with baseclust being the only one obtaining competitive results compared to the previously mentioned ones (RI=0.945 and ET=0.02291).

In summary, EHyClus leads to the best result in terms of RI and ET.

Results for *S 18-19-20* and *S 21-22-23* are shown in the Supplementary Material.

One last scenario from Yassouridis and Leisch (2017) with six different groups is studied. Consider a functional sample

Table 17 Mean values of Purity, F-measure, Rand Index and execution time for the five different procedures described in Yassouridis and Leisch (2017) to simulated data from S 15-16-17

	Purity	Fmeasure	RI	Time
baseclust	0.955	0.917	0.945	0.02291
fitfclust	0.667	0.723	0.760	1.81523
distclust	0.667	0.737	0.770	45.79747
modelcf	0.667	0.676	0.730	0.14528
curvclust	0.667	0.741	0.773	1.05079

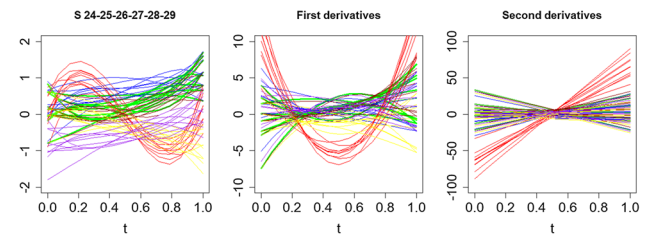


Fig. 13 A sample generated from S 24-25-26-27-28-29. Original data (left panel), first and second derivatives curves (center and right panels respectively)

defined in $[0, 1]$, which have continuous trajectories in such interval and which are the realizations of a stochastic process. Each curve is measured at 15 equidistant observations in the interval $[0, 1]$. We generate 60 functions: 10 for each model from Model 24 to Model 29.

- Model 24.** $X_{24}(t) = x^2(t) + e_2(t)$ where $e_2(t)$ is a centered Gaussian process with standard deviation 0.3.
- Model 25.** $X_{25}(t) = x^2(t) + e_2(t)$
- Model 26.** $X_{26}(t) = \sqrt{x(t)} + e_2(t)$
- Model 27.** $X_{27}(t) = \sin(2\pi x(t)) + e_2(t)$
- Model 28.** $X_{28}(t) = -x^2(t) + e_2(t)$
- Model 29.** $X_{29}(t) = x(t) - 1 + e_2(t)$

Data generated from Models 24 to 29 (S 24-25-26-27-28-29) is shown in Fig. 13. As the curves intertwine a lot, it seems natural that the best result when applying EHyClus will include the modified epigraph index. The results obtained with this methodology, as well as the ones obtained when applying functional *k*-means, hypothesis *k*-means, and the five methodologies of Yassouridis and Leisch (2017) for benchmarking are presented in Tables 18, 19, 20 and 21, respectively. The best result is obtained with baseclust, having a RI equal to 0.909 and ET equal to 0.00519. This methodology obtains small differences compared to EHyClus results, despite its RI being slightly higher.

Table 18 Top 10 mean results for S 24-25-26-27-28-29 considering Euclidean distance (gray), Mahalanobis distance (pink), kernel k-means for initialization (green) and k-means for initialization (orange)

	Purity	Fmeasure	RI	Time
kmeans_dd2.MEI	0.753	0.655	0.889	0.00343
kmeans_dd2.MEI	0.753	0.654	0.889	0.00289
kmeans_dd2.EIHIMEI	0.744	0.643	0.884	0.00450
kmeans_dd2.EIHIMEI	0.744	0.643	0.884	0.00339
ward.D2_dd2.EIHIMEI	0.736	0.646	0.881	0.00023
kmeans_d.MEI	0.732	0.618	0.878	0.00306
kmeans_d.MEI	0.732	0.618	0.878	0.00291
svc_dd2.MEI	0.736	0.625	0.877	0.00969
ward.D2_dd2.MEI	0.727	0.621	0.875	0.00033
svc_dd2.EIHIMEI	0.722	0.620	0.873	0.01540

Table 19 Mean values of Purity, F-measure, Rand Index and execution time for the functional k-means procedure (Martino et al. 2019) with truncated Mahalanobis distance, generalized Mahalanobis distance and L^2 distance to simulated data from S 24-25-26-27-28-29

	Purity	Fmeasure	RI	Time
$d\rho, \rho = 0.001$	0.587	0.495	0.835	0.38078
$d\rho, \rho = 0.02$	0.560	0.477	0.827	0.39563
L^2	0.530	0.458	0.807	0.13257
$dk, k = 3$	0.520	0.405	0.806	0.16657
$d\rho, \rho = 1$	0.550	0.423	0.802	0.28823
$d\rho, \rho = 100$	0.433	0.344	0.785	0.32555
$d\rho, \rho = 1e + 08$	0.400	0.286	0.762	0.41861

Table 20 Mean values of Purity, F-measure, Rand Index and execution time for the test-based k-means procedure (Zambom et al. 2019) with four different initializations to simulated data from S 15-16-17

	Purity	Fmeasure	RI	Time
random	0.645	0.494	0.831	5.41103
kmeans ++	0.604	0.483	0.828	10.81740
hclust	0.608	0.467	0.825	13.34848
kmeans	0.513	0.353	0.779	7.09966

Table 21 Mean values of Purity, F-measure, Rand Index and execution time for the five different procedures described in Yassouridis and Leisch (2017) to simulated data from S 15-16-17

	Purity	Fmeasure	RI	Time
baseclust	0.776	0.7114	0.909	0.00519
fitfclust	0.333	0.474	0.661	0.10180
distclust	0.333	0.461	0.652	0.43145
modelcf	0.333	0.474	0.661	0.02781
curvclust	0.317	0.447	0.606	0.47361

4.3 Simulation summary

Table 22 summarizes the results obtained in all the simulated scenarios, including those deferred to the Supplementary Material. This table presents the RI and ET obtained with EHyClus, with the two methodologies of Martino et al. (2019) and (Zambom et al. 2019), and also with the five different approaches explained in Yassouridis and Leisch (2017). Each row represents one scenario, with the highest RI and the smallest ET shown in bold. Note that the ET for EHyClus in Table 22 corresponds to the one obtained using the best combination of indexes, data and method. The reason is that EHyClus is supposed to be used with one combination fixed in advance. However, in this stage, all the possibilities are calculated and the best one is obtained by comparing all the results.

If the aim of this methodology were to compute all the possibilities and to choose between them with an internal criterion, a global time would be required. In that case, the number of combinations to try could be reduced with no major difference in the results. Nevertheless, even considering all the possibilities and a global ET, this approach is competitive with the others. To illustrate this, the global ET for $S\ 1-4$ when applying EHyClus, functional k-means and hypothesis k-means have been computed, obtaining 2.4991 s, 2.5069 s and 1.1545 s respectively. It is important to note that EHyClus calculates more than 200 combinations, while the other two methodologies compute fewer than 10 different alternatives. Overall, the differences in time are small, but the number of trials is completely different. If the same number of possibilities were computed with the three approaches, the differences in time would be much more noticeable.

EHyClus reaches the best RI in nine out of the fifteen scenarios, and when concerning partial ET, it obtains the best results in all but one. After analysing the difference in the results between EHyClus and the other approaches, one can notice that the RI in the scenarios where EHyClus is not the best strategy have very small differences with the best one. On the other hand, when it gets the best RI it is much higher, and the difference in ET is also high. Finally, baseclust can be considered as the approach with the most competitive results in terms of RI and ET compared to the ones given by EHyClus. Nevertheless, in some cases, when it does not achieve the best result, the differences are high, see the cases S 10-12 and S 21-22-23. In conclusion, EHyClus can be considered a good clustering alternative in terms of metrics and outperforming the existing approaches in terms of ET.

Table 22 Best RI and execution time in seconds (in brackets) for each simulated data set considering EHyClus and seven more different approaches

	EHyClus	Functional k-means	Hypothesis k-means	Baseclust	Fifclust	Distclust	Modelcf	Curvclust
S 1-2	0.644 (0.00944)	0.621 (0.75023)	0.495 (0.34436)	0.629 (0.00657)	0.495 (0.56397)	0.618 (2.11664)	0.594 (0.04352)	0.616 (0.63135)
S 1-3	0.763 (0.00071)	0.751 (0.73622)	0.495 (0.33267)	0.755 (0.00502)	0.506 (0.44235)	0.734 (2.60677)	0.679 (0.05478)	0.749 (0.55363)
S 1-4	0.868 (0.00213)	0.847 (0.72276)	0.495 (0.32627)	0.857 (0.00625 s)	0.569 (0.22153 s)	0.833 (4.21785)	0.819 (0.04989)	0.847 (0.55814)
S 1-5	0.523 (0.01229)	0.496 (1.28681)	0.497 (0.38773)	0.497 (0.00659)	0.495 (0.26672)	0.497 (6.05742)	0.495 (0.05026)	0.503 (0.61242)
S 1-6	0.551 (0.01063)	0.532 (0.23964)	0.529 (0.42781)	0.529 (0.00531)	0.547 (0.24453)	0.525 (2.36152)	0.520 (0.13664)	0.510 (0.57248)
S 1-7	0.501 (0.01748)	0.500 (0.81032)	0.501 (0.24218)	0.499 (0.00557)	0.497 (0.84945)	0.501 (4.72970)	0.508 (0.08188)	0.500 (0.56348)
S 1-8	0.904 (0.00998)	0.993 (0.50043)	0.877 (0.30939)	0.992 (0.00432)	0.716 (1.00264)	0.993 (2.68013)	0.961 (0.04205)	0.986 (0.55182)
S 1-9	0.901 (0.00231)	0.992 (0.50788)	0.878 (0.26316)	0.993 (0.00444)	0.922 (0.85555)	0.989 (4.89978)	1.000 (0.03851)	0.989 (0.54713)
S 10-11	0.603 (0.00483)	0.637 (0.9562)	0.612 (0.61413)	0.652 (0.02038)	0.631 (7.32939)	0.633 (26.62483)	0.632 (0.05706)	0.628 (1.2138)
S 10-12	0.919 (0.00423)	0.718 (7.9055)	0.501 (0.30259)	0.499 (0.02385)	0.496 (3.23131)	0.501 (32.29883)	0.497 (0.05702)	0.502 (1.24491)
S 13-14	0.907 (0.00023)	0.864 (0.65448)	0.824 (0.09722)	0.864 (0.00938)	0.907 (1.16120)	0.864 (12.1398)	0.864 (0.035528)	0.864 (0.36780)
S 15-16-17	0.985 (0.00274)	0.928 (6.50802)	0.944 (0.99653)	0.944 (0.02291)	0.760 (1.81523)	0.770 (45.79747)	0.730 (0.14528)	0.773 (1.05079)
S 18-19-20	0.719 (0.00221)	0.754 (7.76938)	0.821 (1.33809)	0.702 (0.02681)	0.749 (1.80754)	0.724 (41.89710)	0.703 (0.12154)	0.716 (1.17536)
S 21-22-23	0.998 (0.00143)	0.800 (1.54671)	0.937 (0.83632)	0.821 (0.02719)	0.721 (1.80793)	0.723 (47.61030)	0.724 (0.14483)	0.722 (1.14607)
S 24-...-29	0.889 (0.00330)	0.834 (0.38078)	0.831 (5.41103)	0.909 (0.00519)	0.661 (0.10180)	0.652 (0.43145)	0.661 (0.02781)	0.606 (0.47361)

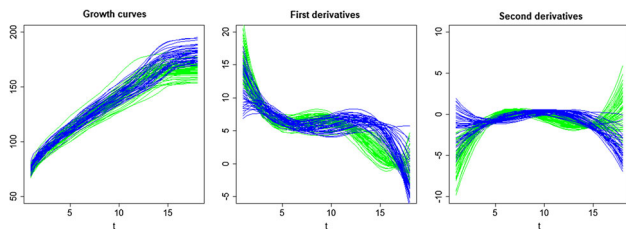


Fig. 14 Growth curves (girls in green and boys in blue) for the original data (left panel) the first derivatives (center panel) and the second derivatives (right panel)

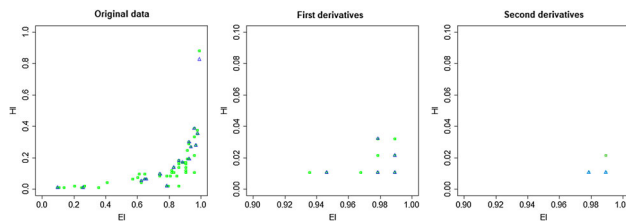


Fig. 15 Scatter plots of the epigraph index (EI) and the hypograph index (HI) of the growth curves original data (left panel), first derivatives (second panel) and second derivatives (right panel)

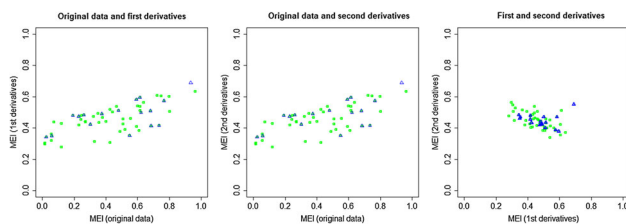


Fig. 16 Growth curves. Scatter plots of different combinations of MEI. Original data and first derivatives (left panel), original data and second derivatives (center panel) and first and second derivatives (right panel)

5 Application to real data

In this section, EHyClus is applied to two different real data sets that have been fully studied in the literature: The Berkeley Growth Study data set and the Canadian Weather data set.

5.1 Case study: Berkeley Growth Study data set

EHyClus has been applied to a popular real data set in the FDA literature: the Berkeley Growth study. This is a classical data set included in Ramsay and Silverman (2005) and available in the ‘fda’ R-package. It contains the heights of 93 children aged from 1 to 18 (54 girls and 39 boys).

A cubic B-spline basis has been fitted to the curves. In Fig. 14, there are differences in shape between the two groups. Thus, when applying the hypograph, epigraph (Fig. 15) and their modified versions (Fig. 16), the two groups have different behaviours despite there being some overlap.

Table 23 Top 10 mean results for growth data set considering Euclidean distance (gray), Mahalanobis distance (pink), a Gaussian kernel (yellow), a polynomial kernel (blue), kernel k-means for initialization (green) and k-means for initialization (orange)

	Purity	Fmeasure	RI	Time
kkmeans.dd2.EIHIMEI	0.968	0.937	0.937	0.02580
kmeans.dd2.EIHIMEI	0.968	0.937	0.937	0.00408
kmeans.dd2.EIHIMEI	0.968	0.937	0.937	0.00434
kmeans.dd2.MEI	0.968	0.937	0.937	0.00399
kmeans.dd2.MEI	0.968	0.937	0.937	0.00000
svc.dd2.EIHIMEI	0.968	0.937	0.937	0.01704
svc.dd2.EIHIMEI	0.968	0.937	0.937	0.02070
svc.dd2.MEI	0.967	0.937	0.937	0.00404
svc.dd2.MEI	0.968	0.937	0.937	0.00407
average.dd2.MEI	0.957	0.917	0.917	0.0001

The best result in Table 23 is obtained when applying different clustering methods: kernel k-means with a polynomial kernel and k-means with Euclidean and Mahalanobis distances, using the three indexes (EI, HI, MEI) applied to first and second derivatives. The resultant clustering partition correctly classifies all boys, but fails to classify 3 girls as boys. The partition is very accurate.

When applying the functional k-means procedure (Table 24), the larger Purity coefficient is equal to 0.850, and it is obtained when applying a big value of ρ , $\rho = 1e + 08$. In addition, the differences in ET compared to EHyClus are evident.

Furthermore, the test-based k-means technique (Table 25) leads to the best result with k-means initialization, obtaining a Purity coefficient of 0.817. EHyClus has also been compared to five different approaches explained in Yassouridis and Leisch (2017) in Table 26, obtaining worse results than the previous ones.

In summary, EHyClus obtains the best result in terms of the three different metrics and in terms of execution time.

5.2 Case study: Canadian weather data set

Another popular real data set in the FDA literature, also included in Ramsay and Silverman (2005) and in the ‘fda’ R-package, is the Canadian weather data set. This data set contains the daily temperature from 1960 to 1994 at 35 different Canadian weather stations grouped into 4 different regions: Artic (3), Atlantic (15), Continental (12) and Pacific (5).

EHyClus has been applied to this data set after smoothing with a cubic B-spline basis. The first and second derivatives by themselves do not give much more information, as shown in Fig. 17. Nevertheless, when applying the indexes and considering them together (Fig. 18), they are able to distinguish between the groups in a better way.

The best configuration between the clustering method, indexes and data is support vector clustering initialized with

Table 24 Mean values of Purity, F-measure, Rand Index and execution time for the functional k -means procedure with truncated Mahalanobis distance, generalized Mahalanobis distance and L^2 distance to data from growth data set

	Purity	Fmeasure	RI	Time
$d\rho, \rho = 1e + 08$	0.850	0.747	0.742	1.93641
$d\rho, \rho = 100$	0.753	0.634	0.624	1.74462
L^2	0.656	0.551	0.544	0.50927
$d\rho, \rho = 0.001$	0.624	0.529	0.525	1.67513
$dk, k = 2$	0.591	0.513	0.511	0.68651
$d\rho, \rho = 0.02$	0.581	0.517	0.508	2.10050
$dk, k = 3$	0.5806	0.4988	0.4974	0.4418
$d\rho, \rho = 1$	0.581	0.496	0.495	2.18262

Table 25 Mean values of Purity, F-measure, Rand Index and execution time for the test-based k -means procedure with four different initializations to data from growth data set

	Purity	Fmeasure	RI	Time
kmeans	0.817	0.702	0.698	0.09893
kmeans ++	0.666	0.552	0.5526	0.34084
hclust	0.666	0.552	0.525	0.09557
random	0.666	0.552	0.525	0.16951

Table 26 Mean values of Purity, F-measure, Rand Index and execution time for the five different procedures described in Yassouridis and Leisch (2017) to data from growth data set

	Purity	Fmeasure	RI	Time
baseclust	0.656	0.551	0.544	0.01212
fitfclust	0.581	0.504	0.499	0.28018
distclust	0.667	0.559	0.550	1.6616
modelcf	0.581	0.542	0.504	0.10094
curvclust	0.677	0.617	0.558	0.71444

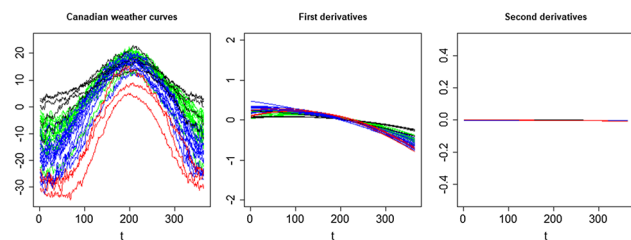


Fig. 17 Canadian weather curves. Original data, first and second derivatives curves

k -means on the three indexes (EI, HI, MEI) of the original data and its second derivatives (see Table 27). In this case, EHyClus provides the following values: a purity value equal to 0.714, F-measure of 0.604 and a Rand index equal to 0.729. These results mean that the final configuration of the groups seems to be accurate, in the sense that the groups obtained by EHyClus have a similar number of elements as the true ones.

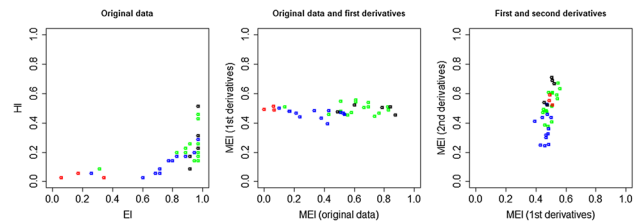


Fig. 18 Canadian weather curves. Epigraph and hypograph index on the original data (left panel), the generalized epigraph index on the original data and first derivatives (center panel) and the generalized epigraph index on the first and second derivatives (right panel)

Table 27 Top 10 mean results for the Canadian weather data set considering Euclidean distance (gray), Mahalanobis distance (pink), a Gaussian kernel (yellow), a polynomial kernel (blue), kernel k -means for initialization (green) and k -means for initialization (orange)

	Purity	Fmeasure	RI	Time
svc._d2.EIHIMEI	0.714	0.604	0.729	0.00801
kkmeans._d2.EIHIMEI	0.685	0.557	0.728	0.00807
kmeans._d2.EIHIMEI-	0.714	0.518	0.728	0.00400
kmeans._d2.EIHIMEI	0.714	0.518	0.728	0.00399
kmeans._dd2.EIHIMEI	0.714	0.518	0.728	0.00464
svc._dd2.EIHIMEI	0.714	0.518	0.728	0.02424
complete._d2.EIHI	0.714	0.517	0.718	0.00401
complete._dd2.EIHI	0.714	0.517	0.718	0.00000
ward.D2._d2.EIHIMEI	0.714	0.520	0.718	0.00000
kkmeans._d2.EIHIMEI	0.714	0.555	0.706	0.01244

Nevertheless, the value of F-measure is smaller because when focusing on the configuration of a concrete group, although the number of observations in the group seems to be correct, the observations inside that group are misclassified. For example, when looking at the Pacific group in Table 28, 6 observations are considered instead of 5, but only 2 out of the 6 are real Pacific observations.

For the functional k -means procedure (Table 29), the best result is obtained when considering the truncated Mahalanobis distance (RI=0.784, F-measure=0.613). Among test-based k -means (Table 30), the best result is obtained with a hierarchical clustering initialization (RI=0.764, F-measure=0.613). Baseclust, fitfclust, distclust and curvclust do not provide competitive results in terms of RI compared to the other three approaches. Modelcf, with RI equal to 0.745, is the only approach from the five in Yassouridis and Leisch (2017) obtaining a closer value of RI, although it is not the best approach (see Table 31).

Regarding the execution time, the only alternative to EHyClus is baseclust, but it obtains the worst results in all the classification metrics. Thus, EHyClus appears as the best alternative.

Table 28 Confusion matrix obtained from comparing real classification to that obtained with our proposal

	Artic	Atlantic	Continental	Pacific	Total
Artic	3	0	0	0	3
Atlantic	1	8	2	4	15
Continental	1	1	10	0	12
Pacific	0	3	0	2	5
Total	5	12	12	6	35

Table 29 Mean values of Purity, F-measure, Rand Index and execution time for the functional *k*-means procedure with truncated Mahalanobis distance, generalized Mahalanobis distance and L^2 distance to data from the Canadian weather data set

	Purity	Fmeasure	RI	Time
<i>dk, k = 3</i>	0.771	0.634	0.784	0.7937
<i>dρ, ρ = 1</i>	0.743	0.598	0.770	0.7697
<i>dρ, ρ = 100</i>	0.743	0.552	0.746	0.7462
<i>dk, k = 2</i>	0.686	0.503	0.694	0.6941
<i>dρ, ρ = 0.001</i>	0.686	0.489	0.681	0.6807
<i>dρ, ρ = 1e + 08</i>	0.657	0.424	0.681	0.6807
L^2	0.686	0.489	0.681	0.6807
<i>dρ, ρ = 0.02</i>	0.657	0.473	0.671	0.6706

Table 30 Mean values of Purity, F-measure, Rand Index and execution time for the test-based *k*-means procedure with four different initializations to simulated data from the Canadian weather data set

	Purity	Fmeasure	RI	Time
hclust	0.771	0.613	0.764	0.12433
kmeans	0.714	0.532	0.731	0.08789
kmeans ++	0.685	0.508	0.717	0.36513
random	0.600	0.427	0.657	0.12229

Table 31 Mean values of Purity, F-measure, Rand Index and execution time for the five different procedures described in Yassouridis and Leisch (2017) to data from the Canadian weather data set

	Purity	Fmeasure	RI	Time
baseclust	0.629	0.468	0.661	0.02593
fitfclust	0.657	0.534	0.704	12.6607
distclust	0.571	0.437	0.640	1.54100
modelcf	0.743	0.548	0.745	0.05882
curvclust	0.657	0.492	0.681	1.1376

6 Discussion

In summary, this paper proposes EHyClus, a new methodology for clustering functional data that is competitive with

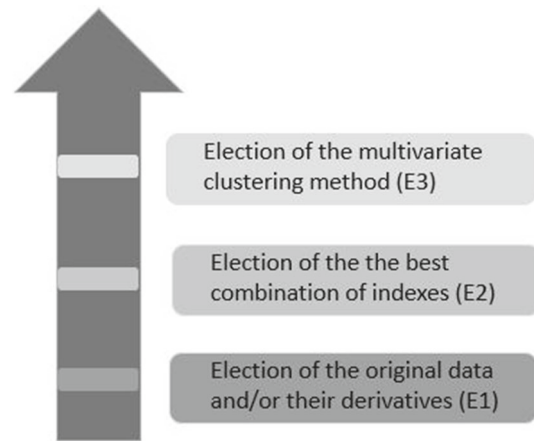


Fig. 19 The three elections to be made during the proposed procedure

respect to the existing ones. EHyClus is based on transforming a functional problem into a multivariate one through the use of the epigraph, the hypograph indexes, their generalized versions and multivariate clustering techniques. It has been compared to seven different clustering procedures, outperforming them in most of the cases in terms of classification metrics and execution time. Finally, the code needed to carry out this analysis and to apply EHyClus is available in the GitHub repository: <https://github.com/bpulidob/EHyClus>.

In order to automatically implement a data-driven procedure able to choose the best combination based on the intrinsic characteristics of the data, further research is needed. Currently, the best combination is obtained based on external validation methods. That is, all the combinations are created, all the clustering methods are applied, and finally, one of them is chosen based on these metrics. To improve this procedure, these elections may be carried out as independent processes, as illustrated in Fig. 19.

The election of the best combination for a given data set without knowing the ground truth is a question for future research. Two possibilities arise in this line. One consists of using internal validation indexes, based on the intrinsic information of the data. (Manning et al. 2009). The Calinski-Harabasz index (Calinski and Harabasz 1974), the Silhouette index (Rousseeuw 1987), and the Davies-Bouldin index (Davies and Bouldin 1979) are examples of these validation measures. The other one consists of studying in depth the a priori potential information given by each combination. For this purpose, different statistical techniques such as the generalized variance proposed in Wilks (1932) could be considered. This methodology leads to a vast reduction in the number of combinations to process. The execution time needed to complete the process would also be reduced. Moreover, these two approaches could be combined if the clustering method is not fixed in advance.

Table 32 Distribution of the number of clusters suggested when applying Silhouette for each Scenario simulated 100 times

	2	3	4	5	6
<i>S 1-2</i>	99	0	0	0	1
<i>S 1-3</i>	100	0	0	0	0
<i>S 1-4</i>	100	0	0	0	0
<i>S 1-5</i>	62	5	5	7	21
<i>S 1-6</i>	72	5	4	6	13
<i>S 1-7</i>	89	2	1	3	5
<i>S 1-8</i>	100	0	0	0	0
<i>S 1-9</i>	100	0	0	0	0
<i>S 10-11</i>	71	15	8	4	2
<i>S 10-12</i>	73	15	5	3	4
<i>S 15-16-17</i>	84	16	0	0	0
<i>S 18-19-20</i>	100	0	0	0	0
<i>S 21-22-23</i>	100	0	0	0	0

It is now generally accepted to fix the number of clusters in advance. This paper has dealt with this issue in that way. The same approach was followed by Martino et al. (2019), Zambom et al. (2019), and (Yassouridis and Leisch 2017). The election of the number of groups may constitute the object of future studies. See for example (Akhanli and Hennig 2020).

Some tests have been carried out, such as the Silhouette index, in order to determine the number of clusters. Each scenario has been simulated 100 times, and the results are given in Table 32. In view of the results, this methodology seems to work well when the number of clusters is two. However, for three groups, the results were unacceptable. As an alternative, the R package ‘NbClust’ has also been considered (see Charrad et al. 2012). However, the results were not as good as expected, concluding that further research is still needed.

Supplementary information

This article is accompanied by a supplementary file.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11222-023-10213-7>.

Acknowledgements This research has been partially supported by Ministerio de Ciencia e Innovación, Gobierno de España, grant numbers PID2019-104901RB-I00, PID2019-104681RB-I00 and PTA2020-018802-I.

Funding Funding for APC: Universidad Carlos III de Madrid (Read & Publish Agreement CRUE-CSIC 2023).

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abraham, C., Cornillon, P.A., Matzner-Løber, E., Molinari, N.: Unsupervised curve clustering using b-splines. *Scand. J. Stat.* **30**(3), 581–595 (2003)
- Akhanli, S.E., Hennig, C.: Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Stat. Comput.* **30**(5), 1523–1544 (2020)
- Arribas-Gil, A., Romo, J.: Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics* **15**(4), 603–619 (2014)
- Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.: Support vector clustering. *J. Mach. Learn. Res.* **2**, 125–137 (2001)
- Boullé, M.: Functional data clustering via piecewise constant nonparametric density estimation. *Pattern Recogn.* **45**(12), 4389–4401 (2012)
- Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. *Commun. Stat. Theory Methods* **3**(1), 1–27 (1974)
- Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A.: Nbclust package. An examination of indices for determining the number of clusters (2012)
- Cuesta-Albertos, J.A., Nieto-Reyes, A.: The random Tukey depth. *Comput. Stat. Data Anal.* **52**(11), 4979–4988 (2008)
- Cuevas, A., Febrero, M., Fraiman, R.: On the use of the bootstrap for estimating functions with functional data. *Comput. Stat. Data Anal.* **51**(2), 1063–1074 (2006)
- Davies, D.L., Bouldin, D.W.: A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **2**, 224–227 (1979)
- Defays, D.: An efficient algorithm for a complete link method. *Comput. J.* **20**(4), 364–366 (1977)
- Dhillon, I.S., Guan, Y., Kulis, B.: Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 551–556 (2004)
- Ferraty, F., Vieu, P.: *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media (2006)
- Flores, R., Lillo, R., Romo, J.: Homogeneity test for functional data. *J. Appl. Stat.* **45**(5), 868–883 (2018)
- Fraiman, R., Muniz, G.: Trimmed means for functional data. *TEST* **10**(2), 419–440 (2001)
- Franco-Pereira, A.M., Lillo, R.E.: Rank tests for functional data based on the epigraph, the hypograph and associated graphical representations. *Adv. Data Anal. Classif.* **14**(3), 651–676 (2020). <https://doi.org/10.1007/s11634-019-00380-9>
- Franco-Pereira, A.M., Lillo, R.E., Romo, J.: Extremality for functional data. In Ferraty, F. (Ed.), *Recent Advances in Functional Data Analysis and Related Topics*, Vol. 14, pp. 651–676. Springer, New York (2011)

- Ghiglietti, A., Paganoni, A.M.: Exact tests for the means of gaussian stochastic processes. *Stat. Probab. Lett.* **131**, 102–107 (2017)
- Giacofci, M., Lambert-Lacroix, S., Marot, G., Picard, F.: Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics* **69**(1), 31–40 (2013)
- Horváth, L., Kokoszka, P.: *Inference for Functional Data with Applications*, Vol. 200. Springer Science & Business Media (2012)
- Hsing, T., Eubank, R.: *Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators*, vol. 997. Wiley, New York (2015)
- Ieva, F., Paganoni, A.M., Pigoli, D., Vitelli, V.: Multivariate functional clustering for the analysis of ecg curves morphology. *J. R. Stat. Soc. Ser. C* **62**(3), 401–418 (2013)
- Jacques, J., Preda, C.: Funclust: a curves clustering method using functional random variables density approximation. *Neurocomputing* **112**, 164–171 (2013)
- Jacques, J., Preda, C.: Functional data clustering: a survey. *Adv. Data Anal. Classif.* **8**(3), 231–255 (2014)
- James, G.M., Sugar, C.A.: Clustering for sparsely sampled functional data. *J. Am. Stat. Assoc.* **98**(462), 397–408 (2003)
- Kayano, M., Dozono, K., Konishi, S.: Functional cluster analysis via orthonormalized gaussian basis expansions and its application. *J. Classif.* **27**(2), 211–230 (2010)
- Lance, G.N., Williams, W.T.: A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput. J.* **9**(4), 373–380 (1967)
- Liu, R.: On a notion of data depth based upon random simplices. *Ann. Stat.* **18**, 405–414 (1990)
- López-Pintado, S., Romo, J.: On the concept of depth for functional data. *Am. Stat. Assoc.* **104**, 327–332 (2009)
- López-Pintado, S., Romo, J.: A half-region depth for functional data. *Comput. Stat. Data Anal.* **55**, 1679–1695 (2011)
- MacQueen, J. 1967. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297
- Manning, C.D., P. Raghavan, and H. Schüte. 2009. *Introduction to information retrieval*. Cambridge, UP
- Martino, A., Ghiglietti, A., Ieva, F., Paganoni, A.M.: A k-means procedure based on a mahalanobis type distance for clustering multivariate functional data. *Stat. Methods Appl.* **28**(2), 301–322 (2019)
- Martín-Barragán, B., Lillo, R.E., Romo, J.: Functional boxplots based on half-regions (2018)
- Oja, H.: Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1**, 327–332 (1983)
- Peng, J., Müller, H.G., et al.: Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Stat.* **2**(3), 1056–1077 (2008)
- Ramsay, J.O., Silverman, B.W.: *Functional Data Analysis* (2 ed.). Springer (2005.)
- Redko, I., Habrard, A., Morvant, E., Sebban, M., Bennani, Y.: *Advances in Domain Adaptation Theory*. Elsevier, New York (2019)
- Rendón, E., Abundez, I., Arizmendi, A., Quiroz, E.M.: Internal versus external cluster validation indexes. *Int. J. Comput. Commun.* **5**(1), 27–34 (2011)
- Romano, E., Balzanella, A., Verde, R.: Spatial variability clustering for spatially dependent functional data. *Stat. Comput.* **27**(3), 645–658 (2017)
- Rossi, F., Conan-Guez, B., El Golli, A.: Clustering functional data with the som algorithm. In *ESANN*, pp. 305–312 (2004)
- Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
- Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., Martin, P.: Clustering multivariate functional data in group-specific functional subspaces. *Comput. Stat.* **35**(3), 1–31 (2020)
- Sguera, C., Galeano, P., Lillo, R.: Spatial depth-based classification for functional data. *TEST* **23**(4), 725–750 (2014)
- Sibson, R.: Slink: an optimally efficient algorithm for the single-link cluster method. *Comput. J.* **16**(1), 30–34 (1973)
- Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationship. *Univ. Kansas Sci. Bull.* **28**, 1409–1438 (1958)
- Tarpey, T., Kinader, K.K.: Clustering functional data. *J. Classif.* **20**(1), 22–93 (2003)
- Traore, O., Cristini, P., Favretto-Cristini, N., Pantera, L., Vieu, P., Viguier-Pla, S.: Clustering acoustic emission signals by mixing two stages dimension reduction and nonparametric approaches. *Comput. Stat.* **34**(2), 631–652 (2019)
- Tucker, J.D., Wu, W., Srivastava, A.: Generative models for functional data using phase and amplitude separation. *Comput. Stat. Data Anal.* **61**, 50–66 (2013)
- Tukey, J.: Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematics (Vancouver, 1974)* vol 2, pp. 523–531 (1975)
- Vardi, Y., Zhang, C.H.: The multivariate l1-median and associated data depth. *Proc. Natl. Acad. Sci.* **97**(4), 1423–1426 (2000)
- Vassilvitskii, S., Arthur, D.: k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035 (2006)
- Wang, J.L., Chiou, J.M., Müller, H.G.: Functional data analysis. *Ann. Rev. Stat. Appl.* **3**, 257–295 (2016)
- Ward, J.H.: Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**(301), 236–244 (1963)
- Wilks, S.S.: Certain generalizations in the analysis of variance. *Biometrika* **1**, 471–494 (1932)
- Yassouridis, C., Leisch, F.: Benchmarking different clustering algorithms on functional data. *Adv. Data Anal. Classif.* **11**(3), 467–492 (2017)
- Zambom, A.Z., Collazos, J.A., Dias, R.: Functional data clustering via hypothesis testing k-means. *Comput. Stat.* **34**(2), 527–549 (2019)
- Zuo, Y.: Projection-based depth functions and associated medians. *Inst. Math. Stat.* **31**, 1460–1490 (2003)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.