

Model-Based Clustering of Functional Data Via Random Projection Ensembles Supplementary Material

Matteo Mori^{1*} and Laura Anderlucci¹

¹Department of Statistical Sciences, University of Bologna, via Belle Arti 41, Bologna, 40126, BO, Italy.

*Corresponding author(s). E-mail(s): matteo.mori8@unibo.it;
Contributing authors: laura.anderlucci@unibo.it;

Journal: Advances in Data Analysis and Classification

1 Some heuristics on the ordering criteria

The theoretical result that motivates the use of RPs is the Johnson–Lindenstrauss lemma. However, a single projection may not necessarily reveal a clear cluster structure, making it necessary to select a subset of informative projections for the ensemble step. Each of the three proposed criteria, Kullback-Leibler (KL), Wasserstein and entropy, captures a different perspective for ordering the clustering results, and simulation studies showed that they behave differently in practice. In this section, we discuss the KL ordering criterion in detail for the three scenarios considered in the simulation study, providing insights into its behaviour.

In order to summarize the information across replicates, we consider the distribution of the ARI between each partition and the true labels for projections ranked first, second, and so on, over the 100 replicates of each scenario. Specifically, Figure 1 reports the median ARI for each rank k : the first point corresponds to the median ARI of the best projection (i.e., rank 1) across all simulations, the second point to that of the second-best projection, and so forth. The shaded area represents the interquartile range.

Overall, there is good agreement between the ordering criterion and the ARI of the individual projections. In particular, in the first two simulation settings the KL criterion is able to place the projections with the highest ARI within the first 100 positions (see the second row of Figure 1 for the zoomed-in plots). A similar pattern is observed in Scenario 3; however, in this case the decrease in ARI values is more pronounced, reflecting the greater difficulty of the clustering task.

Nevertheless, it would be incorrect to assume that including projections beyond the first few necessarily deteriorates the results. Subsequent projections, although possibly associated with lower ARI values, may produce different yet still meaningful clusterings that add useful information to the ensemble step.

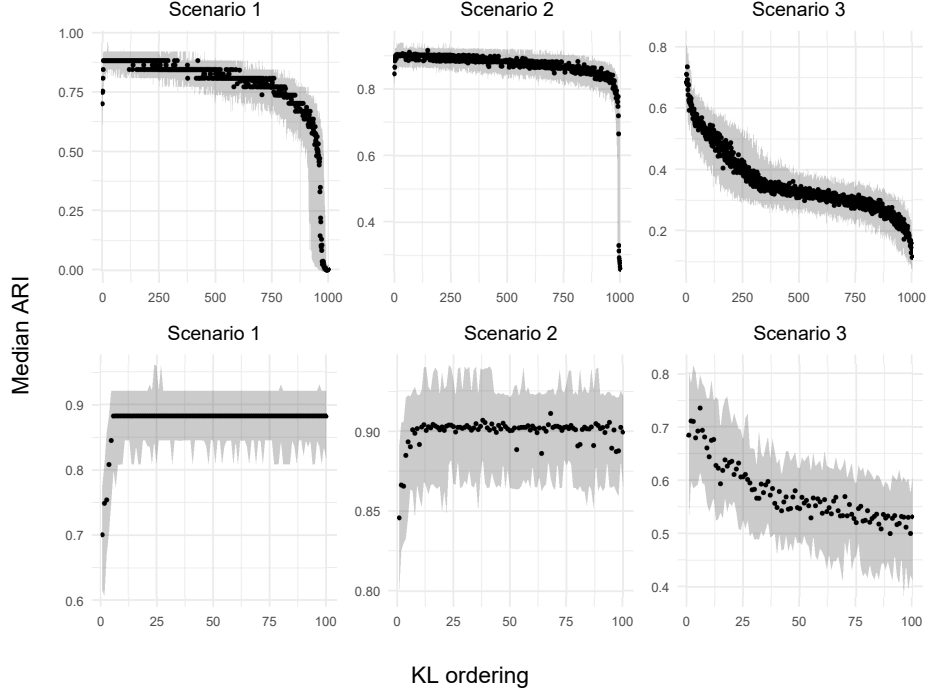


Fig. 1 First row: Median ARI across replications for projections ordered according to the KL criterion in the three simulation scenarios; shaded areas represent the interquartile ranges. Second row: Zoomed-in view of the first 100 projections.

Figures 2 and 3 display similar pattern for the Wasserstein and the Entropy criteria, respectively. Their performance is comparatively less good than that exhibited by KL, as the ARI behaviour seems to be generally less stable in the examined scenarios.

2 Additional simulation study

In the following, we present complementary simulations to assess the robustness of the model, FunCluRPE, under common challenging situations; specifically, we study

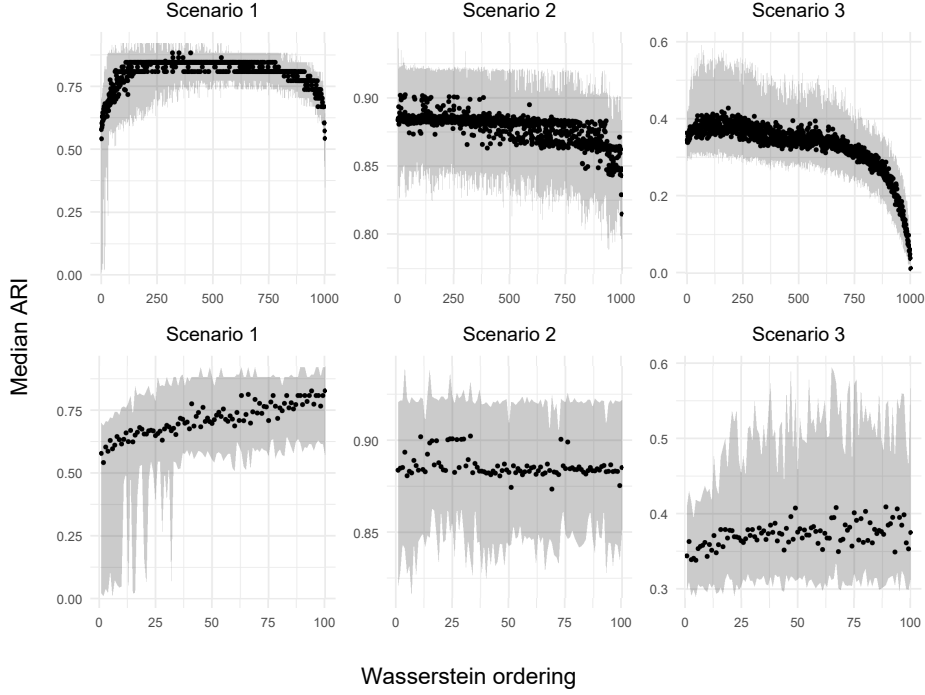


Fig. 2 First row: Median ARI across replications for projections ordered according to the Wasserstein criterion in the three simulation scenarios; shaded areas represent the interquartile ranges. Second row: Zoomed-in view of the first 100 projections.

the behaviour of the proposal in presence of: (i) unbalanced clusters, (ii) non-normal errors and (iii) phase variability.

The first two considered settings are deviations of Scenario 1 presented in the article. Recall that the general model for functional data can be defined as

$$\mathbf{y} = x(\mathbf{t}) + \mathbf{e}, \quad (1)$$

where $\mathbf{y}$ is the vector of observed values, $\mathbf{t}$ is the vector of equi-spaced (time) points, x is the underlying function and $\mathbf{e}$ is the vector of random errors, whose elements are usually modeled as realization of some exogenous random variable with zero mean.

The scenario is built considering $G = 2$ groups, $n = 1001$ equidistant sampling points and $t \in [1, 21]$. The underlying functions are:

$$\begin{aligned} x_1(t) &= U_1 h_1(t) + U_2 h_2(t), \\ x_2(t) &= U_1 h_1(t), \end{aligned}$$

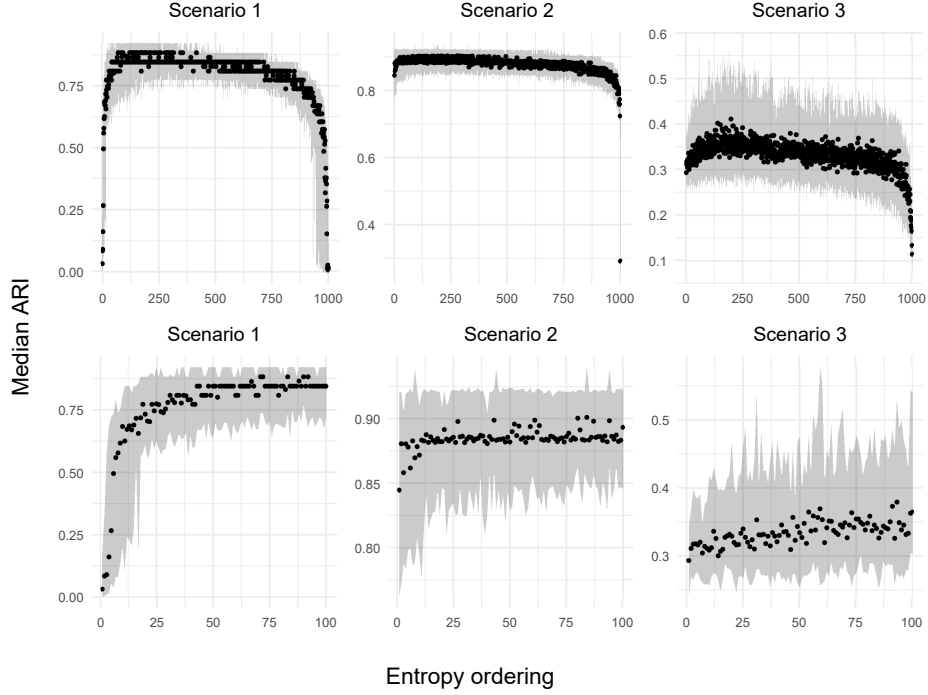


Fig. 3 First row: Median ARI across replications for projections ordered according to the Entropy criterion in the three simulation scenarios; shaded areas represent the interquartile ranges. Second row: Zoomed-in view of the first 100 projections.

where U_1 and U_2 are independent Gaussian variables such that $E[U_1] = E[U_2] = 0$ and $\text{Var}(U_1) = \text{Var}(U_2) = 1/12$, while

$$\begin{aligned} h_1(t) &= \max(6 - |t - 7|, 0), \\ h_2(t) &= \max(6 - |t - 15|, 0). \end{aligned}$$

The first setting considers the case of unbalanced classes: in one simulation the majority class is generated from x_1 , while in the other most observations are generated from x_2 . The assignment probability for the majority class is set to 0.8, and the total sample size N is set to 200. FunCluRPE ran with Haar projection matrices, $d = 5$, $B = 1000$, and $B^* = 100$.

Figure 4 reports the ARI for the balanced setting alongside the two unbalanced scenarios; performance depends on which generating function defines the majority class. Functions x_2 generally form a compact cluster (trajectories concentrate near zero in the second half, see Figure 7 in the article). Therefore, when x_2 is the majority, the partition is more stable and class separability is higher. Conversely, x_1 is more dispersed and overlaps more with x_2 ; when x_1 dominates, the fitted partition is driven

by the heterogeneous majority, and the compact minority x_2 may be harder to detect in some cases. Nevertheless, the ARI exceeds 0.75 in about half of the simulations.

These results suggest that the method performance is not directly affected by unbalanced clusters, but rather by other factors, like the curves dispersion, that can make the clustering problem harder. Indeed, mixture models are generally more flexible than other clustering methods (e.g. k -means) in situations of unbalanced clusters, as G is chosen by the model itself and, on the other hand, if more components are needed to characterise the clustering, this might impact the ensemble step.

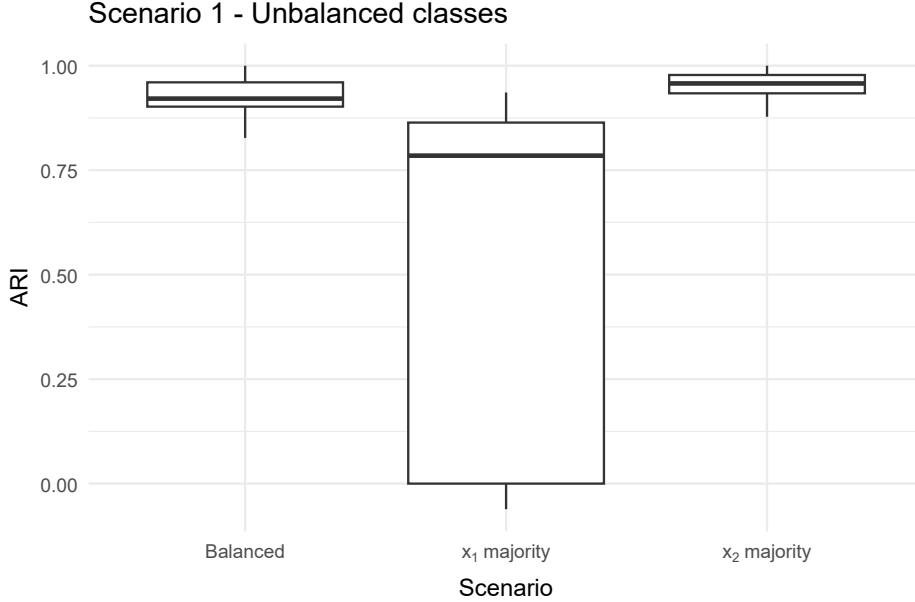


Fig. 4 Comparison of ARI performance for Scenario 1, in case of unbalanced classes.

We further explore the behaviour of the proposed method under non-Gaussian errors. We consider three deviations from Scenario 1. In the first, the error vector $\mathbf{e}$ of equation 1 is generated from independent Student's t distributions with 3 degrees of freedom. In the second case, the elements of $\mathbf{e}$ follow independent Student's t distributions with 5 degrees of freedom. In the third deviation, the errors follow a skew-normal distribution (Azzalini 2013) with shape parameter $\alpha = -5$ and location ξ and scale ω parameters equal to

$$\xi = \frac{5}{\sqrt{6(13\pi - 25)}}, \quad \omega = \sqrt{\frac{13\pi}{6(13\pi - 25)}},$$

respectively. These values are chosen so that $E(\mathbf{e}) = 0$ and $\text{Var}(\mathbf{e}) = 1/6$, that is twice that of the original scenario. In all cases we use Haar projection matrices with $d = 5$, $B = 1000$, and $B^* = 100$.

Figure 5 reports the results based on 100 replicates. Overall, the performance of the proposed method does not deteriorate substantially in the presence of skew-normal errors. The ARI shows a slight decline under Student’s t errors as the degrees of freedom decrease; this can be partly attributed to the much larger error variance, which makes the basis representation less accurate. Indeed, while in the Gaussian scenario the error variance is $\text{Var}(\mathbf{e}) = 1/12$, it increases to $5/3$ and 3 for the Student’s t distributions with 5 and 3 degrees of freedom, respectively.

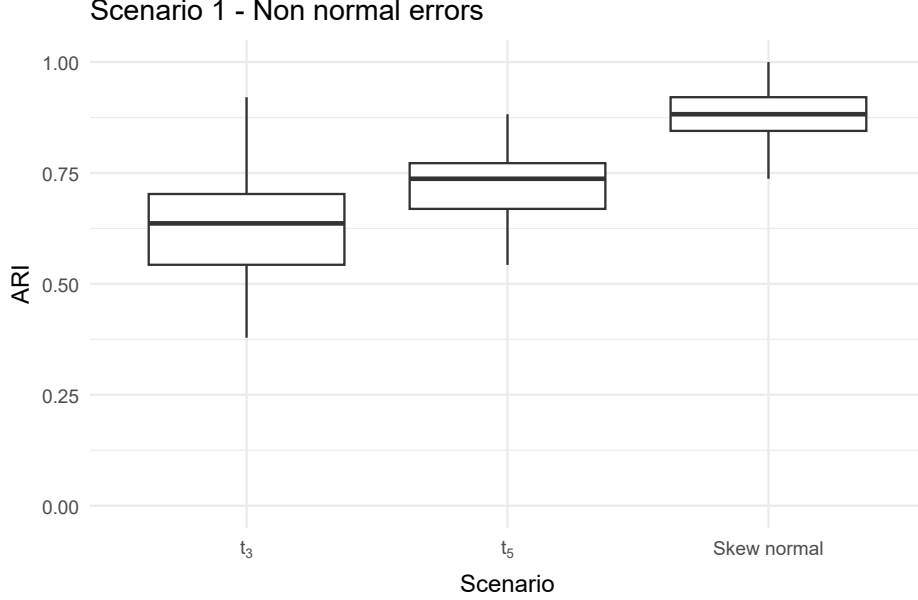


Fig. 5 Comparison of ARI performance for Scenario 1, in case of non normal errors.

Finally, we consider an additional simulation scenario consisting of a two-class problem, with equal assignment probability, where phase variability is introduced through time warping. In this setting, both classes share the same underlying signal function, but their features are temporally shifted by accelerating or delaying the signal.

Specifically, the underlying functions are defined as

$$\begin{aligned} x_1(t) &= U h(\tau_1(t)), \\ x_2(t) &= U h(\tau_2(t)), \end{aligned}$$

where U is a Gaussian random variable such that $E[U] = 0$ and $\text{Var}(U) = 1/12$.

The common signal function $h(t)$ is defined as

$$h(t) = \max\left(5 - \left|\frac{t}{2} - 5\right|, 0\right),$$

while phase variability is introduced through the time-warping functions $\tau_i(t)$, which are constructed from the cumulative densities of Beta distributions and rescaled to the interval $[t_0, t_n]$, with $t_0 = 0$ and $t_n = 20$:

$$\begin{aligned}\tau_1(t) &= \text{Beta}(a_1, b_1) (t_n - t_0) + t_0, \\ \tau_2(t) &= \text{Beta}(a_2, b_2) (t_n - t_0) + t_0.\end{aligned}$$

The parameters of the Beta distributions are chosen for each curve as follows: $a_1 = 1$ and $b_2 = 1$, while a_2 and b_1 are independently drawn from a uniform distribution on $(0, 3)$. This parametrization induces asymmetric time distortions across the two classes: curves in class x_1 tend to exhibit an acceleration of the signal, whereas curves in class x_2 tend to display a delay.

We observe $N = 100$ curves on $n = 1001$ observed time points. Figure 6 illustrates one replicate.

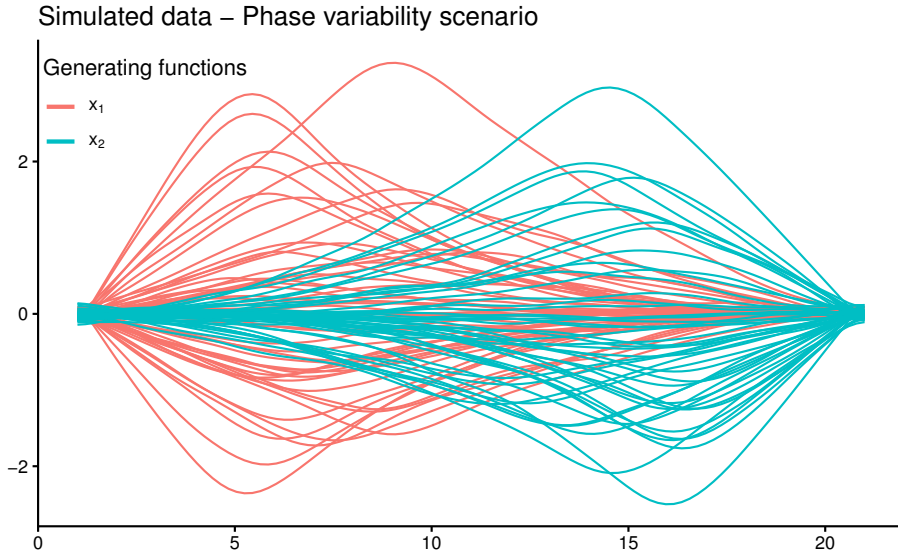


Fig. 6 Example of simulated data according to the time warped model. Different colours reflect different cluster membership. The curves are smoothed using cubic splines penalized via an integrated squared second derivative term.

Figure 7 reports the ARI performance of FunCluRPE alongside competing methods in the literature (see Table 2 in the main article). It can be observed that our method performs generally well in capturing clustering structure.

3 Real data application on periodic data

An additional example, drawn from Ramsay and Silverman (2005) and available in the R package `fda` (Ramsay et al. 2025), consists of daily temperature curves recorded at 35

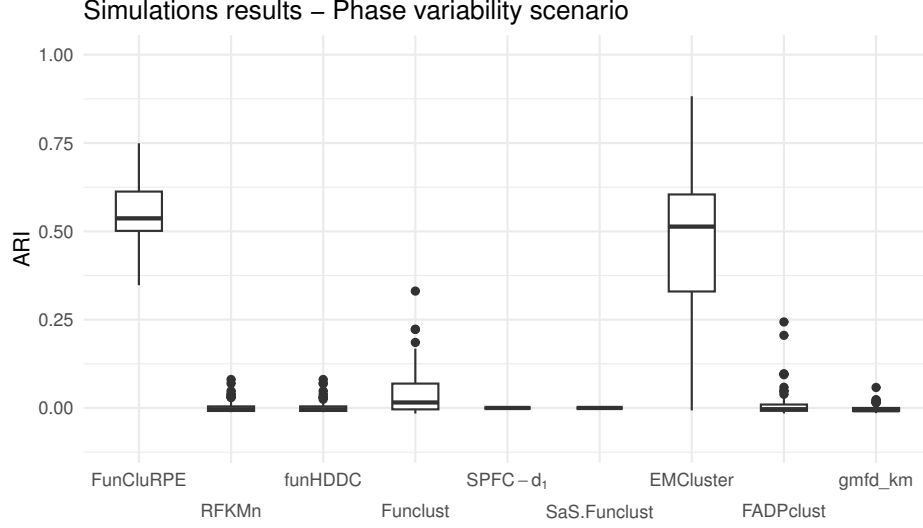


Fig. 7 Comparison of ARI performance for Scenario warped, showing FunCluRPE alongside the competing methods described in the paper.

weather stations across Canada. The stations are associated with $G = 4$ climate regions (Arctic, Atlantic, Continental, and Pacific). Given the periodic nature of the data, a Fourier basis representation was adopted. The GCV procedure selected $K = 328$ basis functions with smoothing parameter $\lambda = 0.01$.

The FunCluRPE model ran using Gaussian projection matrices and with parameters $B = 1000$ and $B^* = 100$. In this case $d = \lceil 5 \log(4) \rceil + 1 = 8$ due to the reduced sample size (if $d = \lceil 10 \log(4) \rceil + 1 = 15$ were chosen, the dimension of the projected space would be almost half of the number of observed curves). The resulting performance, reported in Table 1, is competitive with that of alternative approaches from the literature.

Figure 8 displays a map of Canada where each weather station is coloured according to the resulting clustering. It should be noted that the clustering relies solely on temperature profiles, while climate classification depends on a broader set of environmental factors. Recovering climatic regions from temperature alone is therefore a challenging task. Nevertheless, the spatial pattern observed in the figure suggests that the method is able to identify a meaningful and geographically coherent grouping structure.

Table 1 ARI and computation time (in seconds) for the Canadian temperature data of FunCluRPE, FunCluRPE_G and the alternative methods described in the paper.

Method	ARI	time (s)
FunCluRPE	0.38	73
FunCluRPE_G	0.33	25
RFKMn	0.30	93
funHDDC	0.22	134
Funclust	0.02	1
SPFC-d _{0.25}	0.24	7638
SaS-Funclust	0.22	98
EMCluster	0.13	1718
FADPclust	0.31	1
gmfd_km	0.22	8

Canadian stations - clustering results



Fig. 8 Map of the Canadian stations. Colours indicate the clusters obtained by FunCluRPE. Some station names are omitted to avoid label overlap.

References

- Azzalini, A.: The Skew-normal and Related Families vol. 3. Cambridge University Press, Cambridge, UK (2013)
- Ramsay, J.O., Hooker, G., Graves, S.: fda: Functional Data Analysis. R package version 6.3.0, URL <https://cran.r-project.org/package=fda> (2025)
- Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York (2005)