
# Penalized Model-based Functional Clustering: a Regularization Approach via Shrinkage Methods

Nicola Pronello, Rosaria Ignaccolo, Luigi Ippoliti, and Sara Fontanella

**Abstract** With the advance of modern technology, and with data being recorded continuously, functional data analysis has gained a lot of popularity in recent years. Working in a mixture model-based framework, we develop a flexible functional clustering technique achieving dimensionality reduction schemes through a $L_1$ penalization. The proposed procedure results in an integrated modelling approach where shrinkage techniques are applied to enable sparse solutions in both the means and the covariance matrices of the mixture components, while preserving the underlying clustering structure. This leads to an entirely data-driven methodology suitable for simultaneous dimensionality reduction and clustering. Preliminary experimental results, both from simulation and real data, show that the proposed methodology is worth considering within the framework of functional clustering.

**Keywords:** functional data analysis, $L_1$-penalty, silhouette width, graphical LASSO, mixture model

Nicola Pronello (✉)
Department of Neurosciences, Imaging and Clinical Sciences, University of Chieti-Pescara, Chieti, Italy, e-mail: nicola.pronello@unich.it

Rosaria Ignaccolo
Department of Economics and Statistics "Cognetti de Martiis", University of Torino, Torino, Italy, e-mail: rosaria.ignaccolo@unito.it

Luigi Ippoliti
Department of Economics, University of Chieti-Pescara, Pescara, Italy,
e-mail: luigi.ippoliti@unich.it

Sara Fontanella
National Heart and Lung Institute, Imperial College London, London, United Kingdom,
e-mail: s.fontanella@imperial.ac.uk



P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_34

# 1 Introduction

In recent decades, technological innovations have produced data that are increasingly complex, high dimensional, and structured. A large amount of these data can be characterized as functions defined on some continuous domain and their statistical analysis has attracted the interest of many researchers. This surge of interests is explained by the ubiquitous examples of functional data that can be found in different application fields (see for example [2], and references therein for specific examples). With functions as the basic units of observation, the analysis of functional data poses significant theoretical and practical challenges to statisticians. Despite these difficulties, methodology for clustering functional data has advanced rapidly during the past years; recent surveys of functional data clustering are presented in [7] and [2]. Popular approaches have extended classical clustering concepts for vector-valued multivariate data to functional data.

In this paper, we consider a finite mixture as a flexible model for clustering. In particular, applying a functional model-based clustering algorithm with an $L_1$-penalty function on a set of projection coefficients, we extend the results of [8] and [9] for vector-valued multivariate data to a functional data framework. This approach appears particularly appealing in all cases in which the functions are spatially heterogeneous, meaning that some parts of the function can be smoother than in other parts, or that there may be distant parts of the function that are correlated with each other. Furthermore, the introduction of a shrinkage penalty allows to look for directions in the feature space (that is now the space of expansion/projection coefficients) that are the most useful in separating the underlying groups without first applying dimensionality reduction techniques.

In Section 2 we present at first the methodology along with some details on model estimation (subsection 2.2). Secondly, in Section 3, we perform a validation study with simulated and real data for which the classes are known a-priori.

# 2 Shrinkage Method for Model-based Clustering for Functional Data

Here we consider the problem of clustering a set of $n$ observed curves into $K$ homogeneous groups (or clusters). To this end, we propose a flexible model based on a finite mixture of Gaussian distributions, with a $L_1$ penalized likelihood, which we name *Penalized model-based Functional Clustering* (PFC-$L_1$).

## 2.1 Model Definition

We consider a set of $n$ observed curves, $x_1, \ldots, x_n$, that are independent realizations of a continuous stochastic process $X = \{X(t)\}_{t \in [0,T]}$ taking values in $L_2[0,T]$. In

practice, such curves/trajectories are available only at a discrete set of the domain points $\{t_{is} : i = 1, \ldots, n,\ s = 1, \ldots, m_i\}$ and the $n$ curves need to be reconstructed. To this goal, it is common to assume that the curves belong to a finite dimensional space spanned by a basis of functions, so that given a basis of functions $\boldsymbol{\Phi} = \{\psi_1, ..., \psi_p\}$ each curve $x_i(t)$ admits the following decomposition:

$$x_i(t) = \sum_{j=1}^{p} \beta_{j,i}\psi_j(t), \qquad i = 1, \ldots, n; \tag{2.1}$$

that is the stochastic process $X$ admits a corresponding truncated basis expansion

$$X(t) = \sum_{j=1}^{p} \beta_j(X)\psi_j(t),$$

where $\boldsymbol{\beta} = \{\beta_1(X), \ldots, \beta_p(X)\}$ is a random vector in $\mathbb{R}^p$. By considering observations with a sampling error, such that

$$x_i^{obs}(t) = x_i(t) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{2.2}$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, the realizations of the random coefficients $\beta_{j,i}$ for $j = 1, \ldots, p$ describing each curve can be obtained via least squares as $\hat{\boldsymbol{\beta}}_i = (\boldsymbol{\Theta}_i'\boldsymbol{\Theta}_i)^{-1}\boldsymbol{\Theta}_i'\mathbf{X}_i^{obs}$ where $\boldsymbol{\Theta}_i = (\psi_j(t_{is}))$, $1 \le j \le p$, $1 \le s \le m_i$ contains the basis functions evaluated at the fixed domain points and $\mathbf{X}_i^{obs} = (x_i^{obs}(t_{i1}), \ldots, x_i^{obs}(t_{im_i}))'$ is the vector of observed values of the $i$-th curve.

With the goal of dividing into $K$ homogeneous groups the observed curves $x_1, \ldots, x_n$, let us assume that it exists an unobservable grouping variable $\mathbf{Z} = (Z_1, ..., Z_K) \in [0, 1]^K$ indicating the cluster membership: $z_{i,k} = 1$ if $x_i$ belongs to cluster $k$, 0 otherwise (and $z_{i,k}$ is indeed what we want to predict for each curve).

In adopting a model-based clustering approach, we denote with $\pi_k$ the (a-priori) probabilities of belonging to a group:

$$\pi_k = \mathbb{P}(Z_k = 1), \qquad k = 1, \ldots, K,$$

such that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$ for each $k$, and we assume that, conditionally on $Z$, the random vector $\boldsymbol{\beta}$ follows a multivariate Gaussian distribution, that is for each cluster

$$\boldsymbol{\beta}|(Z_k = 1) = \boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{\mu}_k = (\mu_{1,k}, \ldots, \mu_{p,k})^T$ and $\boldsymbol{\Sigma}_k$ are respectively the mean vector and the covariance matrix of the $k$-th group. Then the marginal distribution of $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}$ can be written as a finite mixture with mixing proportions $\pi_k$ as

$$p(\boldsymbol{\beta}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{\beta}_k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $f$ is the multivariate Gaussian density function. The log-likelihood function can then be written as

$$l(\boldsymbol{\theta};\boldsymbol{\beta}) = \sum_{i=1}^{n} log \sum_{k=1}^{K} \pi_k f(\boldsymbol{\beta}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

where $\boldsymbol{\theta} = \{\pi_1,\ldots,\pi_K;\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K;\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K\}$ is the vector of parameters to be estimated and $\boldsymbol{\beta}_i = (\beta_{1,i},\ldots,\beta_{p,i})^T$ is the vector of projection coefficients of the $i$-th curve.

In this modeling framework, we consider a very general situation without introducing any kind of constraints neither for cluster means nor for covariance matrices, that can be different in each cluster. This flexibility, however, leads to overparameterization and, as an alternative to any kind of constraints, we consider a penalty that allows regularized parameters' estimation.

To define a suitable penalty term, we follow the penalized approach introduced by Zhou et al. [8] in the high-dimensional setting, and so we consider a penalty composed by two terms: the first one on the mean vector of each cluster $\boldsymbol{\mu}_k$, and the second one on the inverse of the covariance matrix in each group $\mathbf{W}_k = \boldsymbol{\Sigma}_k^{-1}$, otherwise said "precision" matrix, with elements $W_{k;j,l}$. The proposed penalized log-likelihood function, given the projection coefficients $\boldsymbol{\beta}_i$, is

$$l_P(\boldsymbol{\theta};\boldsymbol{\beta}) = \sum_{i=1}^{n} log \sum_{k=1}^{K} \pi_k f(\boldsymbol{\beta}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) - \lambda_1 \sum_{k=1}^{K} ||\boldsymbol{\mu}_k||_1 - \lambda_2 \sum_{k=1}^{K} \sum_{j,l}^{p} |W_{k;j,l}|,$$

where $||\boldsymbol{\mu}_k||_1 = \sum_{j=1}^{p} |\mu_{k,j}|$, $\lambda_1 > 0$ and $\lambda_2 > 0$ are penalty parameters to be suitably chosen.

The penalty term on the cluster mean vectors allow for component selection in the functional data framework (whereas it would be variable selection in the multivariate case), considering that when the $j$-th component in the basis expansion is not useful in separating groups it has a common mean across groups, that is $\mu_{1,j} = \ldots = \mu_{K,j} = 0$. Then to realize component selection the considered term is $\sum_{k=1}^{K} ||\boldsymbol{\mu}_k||_1$.

The second part of the penalty, namely $\sum_{k=1}^{K} \sum_{j,l}^{p} |W_{k;j,l}|$, imposes a shrinkage on the elements of the precision matrices, thus avoiding possible singularity problems and facilitating the estimation of large and sparse covariance matrices.

## 2.2 Model Estimation via E-M Algorithm

Since the membership of each observation to a cluster is unobservable, data related to the grouping variable $\mathbf{Z}$ is inevitably missing and the maximum penalized log-likelihood estimator can be obtained by means of the E-M algorithm [4], that iterates over two steps: expectation (E) of the complete data (penalized) log-likelihood by considering the unknown parameters equal to those obtained at the previous iteration

(with initialization values), and maximization (M) of a lower bound of the obtained expected value with respect to the unknown parameters.

In particular, at the $d$-th iteration, given a current estimate $\boldsymbol{\theta}^{(d)}$, the lower bound after the E-step assumes the following form:

$$Q_P(\boldsymbol{\theta};\boldsymbol{\theta}^{(d)}) = \sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{k,i}^{(d)}[\log\pi_k + \log f(\boldsymbol{\beta}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)] - \lambda_1\sum_{k=1}^{K}||\boldsymbol{\mu}_k||_1 - \lambda_2\sum_{k=1}^{K}\sum_{j,l}^{p}|W_{k;j,l}|,$$

where $\tau_{k,i} = \mathbb{P}(Z_k = 1|X = x_i)$ is the posterior probability of observation $i$ to belong to group $k$. The M-step maximizes the function $Q_P$ in order to update the estimate of $\boldsymbol{\theta}$.

As suggested by [9], it is possible to maximize each of the $K$ term using a "graphical lasso" (GLASSO) algorithm (first proposed by [5]), thanks to the close connection between fitting Gaussian mixture models and Gaussian graphical models. Indeed, in GLASSO the objective function looks like $\log\det(\mathbf{W}) - \text{tr}(\mathbf{SW}) - \lambda\sum_{j,l}^{p}|W_{j,l}|$ so that the algorithm implemented in the R package "`glasso`" can be used with $\mathbf{W} = \mathbf{W}_k$, $S = \tilde{\mathbf{S}}_k$ and $\lambda = \frac{2\lambda_2}{\sum_{i=1}^{n}\tau_{k,i}^{(d)}}$ for each $k$ to obtain the elements $\widehat{W}_{k;j,l}^{(d+1)}$ of the precision matrices.

## 2.3 Model Selection via Silhouette Profile

A fundamental, and probably unsolved, problem in cluster analysis is determining the "true" number of groups in a dataset. To this purpose, for simplicity, here we approach the problem choosing the number of groups as cluster validation problem and use the *average silhouette width* index as a model selection heuristic. The silhouette value for curve $i$ is given by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average distance of curve $i$ to all other curves $h$ assigned to the same cluster (if $i$ is the only observation in its cluster, then $s(i) = 0$), and $b(i)$ is the minimum average distance of curve $i$ to observations $h$ which are assigned to a different cluster. This definition ensures that $s(i)$ takes values in $[-1, 1]$, where values close to one indicate "better"clustering solutions. Conditional on $K$ and a pair of values $(\lambda_1, \lambda_2)$, we thus assess the overall cluster solution using the total average of silhouette values

$$S(K, \lambda_1, \lambda_2) = \frac{1}{n}\sum_{i=1}^{n}s(i).$$

In particular, by doing a grid search for the triple $(K, \lambda_1, \lambda_2)$, the best cluster solution is obtained by looking for the largest value of the *average silhouette width* (*ASW*) index. Note that, to evaluate $s(i), i = 1, \ldots, n$, and then the objective function $S(K, \lambda_1, \lambda_2)$, we need to compute a distance between pairs of curves $X_i$ and $X_h$. One

possibility is to compute the euclidean distance

$$d_E^2(i,h) = \int \|X_i(t) - X_h(t)\|^2 dt.$$

# 3 Experimental Results

## 3.1 Simulation

We present here a simulated scenario in order to investigate the effectiveness of the $L_1$ regularization in removing noise while preserving dominant local features, accommodating for spatial heterogeneity of the curves.

The statistical analysis is illustrated for data simulated by means of a finite mixture of multivariate Gaussian distributions. In particular, based on equation (2.1) and (2.2), the curves are simulated using a combination of $p = 25$ Fourier basis functions defined over a one-dimensional regular grid with 100 observations. We consider a mixture of four ($K = 4$) multivariate Gaussian distributions with isotropic covariance matrices, i.e.

$$\boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\mu}_k; \mathcal{I}_k) \text{ where } \epsilon_i \sim \mathcal{N}(0; 0.5), \quad k = 1, \ldots, 4.$$

With the exclusion of 3 entries per group, the means $\boldsymbol{\mu}_k$ are all zero mean vectors. Under this scenario, the simulated curves (25 per group) and the non-zero group expansion coefficients are represented in Figure 1. For this simple simulation setting, estimation results suggest that, using euclidean distance to computed the *ASW*, the grid search procedure is always able to correctly select the cluster-relevant basis functions. This is confirmed by Figure 2 which shows both the distribution (over 100 replications) of the selected basis functions and the data projected on these bases that clearly highlight the identification of 4 clusters. Under this scenario, the quality of the estimated clusters thus appears very good as the analysis of the misclassification rate suggests an 100% of accuracy in all the replicated datasets.

Similar results hold for more complex simulation designs, where we consider different structure of the covariance matrices in the data generating process.

## 3.2 Performance on Real Data Sets

We evaluate the PFC-$L_1$ model on a well-known benchmark data set, namely the electrocardiogram (ECG) data set (data can be found at the UCR Time Series Classification Archive [3]).

The ECG data set comprises a set of 200 electrocardiograms from 2 groups of patients, myocardial infarction and healthy, sampled at 96 time instants in time.

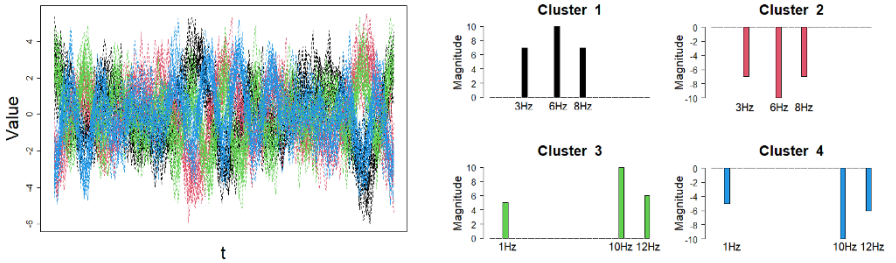


**Fig. 1** Left: 25 simulated curves for each group. Right: Vector of expansion coefficients for each group, with only three non-zero coefficients corresponding to basis functions with specific periodicities (Hertz values).

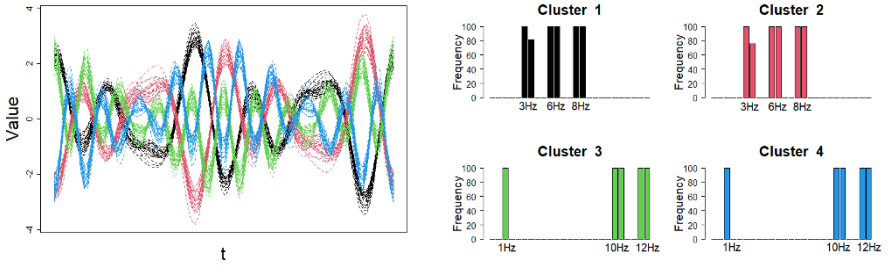


**Fig. 2** Left: Data projected on cluster specific functional subspace generated by the selected basis functions. Right: Distribution (over 100 replications) of the selected basis functions shown for pairs of sine and cosine basis functions, according to the Hertz values.

This data set were previously used to compare the performance of several functional clustering models in [1]. The results in Table 5 of [1] show that the FunFEM models, compared to other state of the art methodologies, achieved the best performances in terms of accuracy. Hence, here, we limit the comparison to the results obtained with the PFC-$L_1$ and the FunFEM models. Although FunFEM models relay on a mixture of Gaussian distributions describing the likelihood of the data similarly to our proposal, they differ on facing the intrinsic high dimension of the problem by estimating a latent discriminant subspace in parallel with the steps of an EM algorithm.

For all the data, we reconstruct the functional form from the sampled curves choosing arbitrarily 20 cubic spline basis of functions. We tested the PFC-$L_1$ models considering five different values for the number of clusters, $K = \{2, 3, 4, 5, 6\}$, and six values for $\lambda_1 = \{0.5, 1, 5, 10, 15, 20\}$.

Considering that the GLASSO penalty parameter $\lambda$ depends linearly from $\lambda_2$, the choice of $\lambda_2$ has to provide suitable values for $\lambda$. A practical approach is to choose values avoiding convergence problems with GLASSO. Here $\lambda_2$ was set to $\{5, 7.5, 10, 12, 15, 20\}$ for the ECG data. Both PFC-$L_1$ and FunFEM algorithms were initialized using a $K$-means procedure.

The clustering accuracies, computed with respect to the known labels, are 69% for FunFEM DFM$_{[\alpha_{kj}\beta_k]}$ (choosing among 12 different model parameterizations with BIC index), and 75% for PFC-L$_1$ [$\lambda_1 = 0.5$ , $\lambda_2 = 5$] (values of tuning parameters chose by ASW index) . Thus PFC-$L_1$ achieves good performance, with an increase in the accuracy about 9%.

## 4 Discussion

In this paper we tried to investigate the potential of shrinkage methods for clustering functional data. Our numerical examples show the advantages of performing clustering with features selection, such as uncover interesting structures underlying the data while preserving good clustering accuracy. To the best of our knowledge, this is the first proposal that considers a penalty for both means and covariances of mixture components in functional model-based clustering. In the model selection section we defined an heuristic criterion to choose among different model parameterizations based on average silhouette index. It may be interesting to evaluate different distances (i.e. not euclidean) to compute this index in future research. Moreover, we will consider more complex simulation designs to investigate the robustness of the proposal and extend the comparison with the state of the art methodologies on more benchmark datasets.

## References

1. Bouveyron, C., Come, E., Jacques, E.: The discriminative functional mixture model for a comparative analysis of bike sharing systems. Ann. Appl. Stat. **9**, 1726–1760 (2015)
2. Chamroukhi, F., Nguyen, H.: Model-based clustering and classification of functional data. Wiley Interdiscip. Rev.: Data Min. and Knowl. Discov. **9**, e1298, 1–36 (2019)
3. Dau, H. A., Keogh, E., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G., Hexagon-ML: The UCR Time Series Classification Archive (October 2018)
`https://www.cs.ucr.edu/$\sim$eamonn/time\_series\_data\_2018/`
4. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.). **39**, 1–38 (1977)
5. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostat. **9**, 432–41 (2008)
6. Friedman, J., Hastie, T., Tibshirani, R.: glasso: Graphical Lasso: Estimation of Gaussian Graphical Models, R package version 1.11 (2019).
`https://CRAN.R-project.org/package=glasso`
7. Jacques, J., Preda, C.: Functional data clustering: A survey. Adv. Data Anal. Classif. **8**, 231–255 (2013)
8. Pan, W., Shen, X.: Penalized model-based clustering with application to variable selection. J. Mach. Learn. Res. **8**, 1145–1164 (2007)
9. Zhou, H., Pan, W., Shen, X.: Penalized model-based clustering with unconstrained covariance matrices. Electron. J. Stat. **3**, 1473–1496 (2009)