# SPARSE CLUSTERING FOR FUNCTIONAL DATA

Fabio Centofanti [1], Antonio Lepore[1] Biagio Palumbo[1]

[1] Department of Industrial Engineering, University of Naples Federico II, Piazzale Tecchio 80, Napoli (e-mail: fabio.centofanti@unina.it, antonio.lepore@unina.it, biagio.palumbo@unina.it)

**ABSTRACT**: The sparse and smooth functional clustering (SaS-Funclust) method is presented for sparse clustering of functional data, i.e., to split a sample of curves into homogeneous groups while jointly detecting the most informative portions of the domain. SaS-Funclust relies on a functional adaptive pairwise fusion penalty and a roughness penalty. The former allows identifying the noninformative portion of the domain, whereas the latter improves the interpretability by imposing some degree of smoothing to the cluster means. The practical advantages of the SaS-Funclust method are illustrated through a real-data example in the analysis of the Berkeley growth study dataset. The SaS-Funclust method is implemented in the R package sasfunclust, available on CRAN.

**KEYWORDS**: functional data analysis, functional clustering, model-based clustering, penalized likelihood, sparse clustering

## 1 Introduction

In the last years, due to recent developments in technology and computational power, the majority of the data gathered by practitioners and scientists in many fields contain information about curves or surfaces that are apt to be modelled as functional data, i.e., continuous random functions defined on a compact domain (Ramsay & Silverman, 2005). Cluster analysis is a key tool in functional data analysis, just as it is in the multivariate (non-functional) statistical literature, with applications in several fields. Functional clustering main goal is to classify a sample of functional data into homogenous groups of curves with no explicit information on the actual underlying clustering structure (Capezza *et al.*, 2021). However, as stated in many multivariate data applications, some characteristics could be entirely unhelpful in revealing the desired clustering structure. In this setting, to achieve more accurate group identification, it is important to determine the features in which respect true clusters differ the most, or equivalently noninformative features that may conceal the true clustering structure. More in general, the methods capable of selecting informative

features and eliminating noninformative ones are referred to as *sparse* (Witten & Tibshirani, 2010; Pan & Shen, 2007; Guo *et al.*, 2010). Recently, the notion of sparseness has been translated into a functional data clustering framework. Sparse functional clustering methods have appeared in literature with the aim of clustering functional data while jointly detecting the most informative portion of the domain and improving both the accuracy and the interpretability of the analysis (Floriello & Vitelli, 2017; Vitelli, 2023). In this article, we present the model-based procedure for the sparse clustering of functional data, which has been recently proposed by Centofanti *et al.*, 2023, and referred to as sparse and smooth functional clustering (SaS-Funclust). The SaS-Funclust procedure is implemented in the R package sasfunclust and is openly available on CRAN.

## 2 The SaS-Funclust method

Suppose that $N$ vectors $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$, of size $n_i$, $i = 1, \ldots, N$, of observed values of a function $f_i$ over the time points $t_{i1}, \ldots, t_{in_i}$ are spread among $g = 1, \ldots, G$ unknown clusters and the probability of each observation to belong to the $g$th cluster is $\pi_g$. The function $f_i$ is assumed a Gaussian random process with mean $\mu_g$, covariance $\omega_g$, and values in $L^2(\mathcal{T})$, which denotes the separable Hilbert space of square-integrable functions defined on the compact domain $\mathcal{T}$. We assume that, conditionally on the cluster membership, $\boldsymbol{Y}_i$ is modelled as

$$\boldsymbol{Y}_i = \boldsymbol{f}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N,$$

where $\boldsymbol{f}_i = (f_i(t_{i1}), \ldots, f_i(t_{in_i}))^T$ contains the values of the function $f_i$ at $t_{i1}$, $\ldots, t_{in_i}$ and $\boldsymbol{\varepsilon}_i$ is a vector of random errors zero mean and constant variance $\sigma_e^2$. In this setting, the SaS-Funclust solution (Centofanti *et al.*, 2023) is obtained by maximizing the following penalized log-likelihood

$$L_p(\boldsymbol{\Theta} | \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N) = \sum_{g=1}^{G} \pi_g \psi\left(\boldsymbol{Y}_i; \boldsymbol{\mu}_{gi}, \boldsymbol{\Omega}_{gi} + \boldsymbol{I}\sigma_e^2\right) - \mathcal{P}(\mu_1, \ldots, \mu_G), \quad (1)$$

where $\boldsymbol{\Theta} = \{\pi_g, \mu_g, \omega_g, \sigma_e^2\}_{g=1,\ldots,G}$ is the parameter set of interest, $\boldsymbol{\mu}_{gi} = (\mu_g(t_{i1}), \ldots, \mu_g(t_{in_i}))^T$, $\boldsymbol{\Omega}_{gi} = \{\omega_g(t_{ki}, t_{li})\}_{k,l=1\ldots,n_i}$, $\psi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate Gaussian density distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $\mathcal{P}(\cdot)$ is a penalty

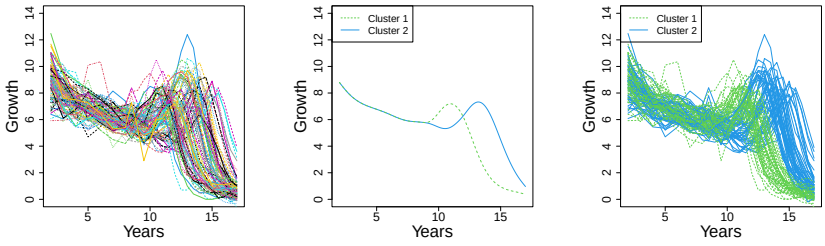
function defined as

$$\mathcal{P}(\mu_1,\ldots,\mu_G)=\lambda_L\sum_{1\le g\le g'\le G}\int_{\mathcal{T}}\tau_{g,g'}(t)\,|\mu_g(t)-\mu_{g'}(t)\,|dt+\lambda_s\sum_{g=1}^{G}\int_{\mathcal{T}}\left(\mu_g^{(s)}(t)\right)^2dt, \tag{2}$$

where $\lambda_L,\lambda_s\ge 0$ are tuning parameters, $\tau_{g,g'}$ are prespecified weight functions, and $\mu_g^{(s)}(\cdot)$ denotes the $s$th-order derivative of $\mu_g$. The first element of the right-hand side of Equation (2) is the functional adaptive pairwise fusion penalty (FAPFP). It allows the pair of cluster means to be equal over a specific portion of the domain that is considered noninformative for separating the cluster means. Thus, the SaS-Funclust method is able to detect, for each cluster pair, the portion of the domain that is noninformative for the cluster analysis, i.e., the portion of the domain where the corresponding cluster means are not fused. The last term in Equation (2) is a roughness penalty, applied on the cluster means to further improve the interpretability of the analysis by constraining, with a magnitude quantified by $\lambda_s$, the cluster means to own a certain degree of smoothness, measured by the derivative order $s$. A specific expectation-conditional maximization (ECM) algorithm is used to maximize the objective function in Equation (1), after some structure is imposed on $f_i$. Then a cross-validation procedure is proposed to select the appropriate model parameters. Further details are in Centofanti *et al.*, 2023.

## 3 A Real-data Example: Berkeley Growth Study Data

In this section, the SaS-Funclust method is applied to the growth dataset from the Berkeley growth study. In this study, 31 height measurements of 54 girls and 39 boys are available from ages 1 through 18. The aim of the analysis is to cluster growth velocities from age 2 to 17. Figure 1 shows (a) the interpolating growth velocity curves for all the individuals, (b) the estimated cluster means, and (c) the clustered growth curves for the SaS-Funclust method. The estimated cluster means are fused over the first portion of the domain, whereas they are separated over the remaining portion. This implies that on average, the two identified clusters do not differ over the first portion of the domain, which can be, thus, regarded as noninformative. The separation between the two groups arises over the remaining informative portion of the domain, where two sharp peaks of growth velocity arise, instead. The latter peaks are referred to as pubertal spurts in the medical literature and in this regard, the obtained results highlight two primary timing/duration groupings. The male pubertal spurt occurs later and lasts longer than the female one. The estimated cluster




(a) (b) (c)

Figure 1: (a) Growth velocities, (b) estimated cluster curve means, and (c) curve clusters for the SaS-Funclust in the Berkeley growth study dataset.

means from some competing methods do not allow for a similar straightforward interpretation.

## References


CAPEZZA, C., CENTOFANTI, F., LEPORE, A., & PALUMBO, B. 2021. Functional clustering methods for resistance spot welding process data in the automotive industry. *Applied Stochastic Models in Business and Industry*, **37**(5), 908–925.

CENTOFANTI, F., LEPORE, A., & PALUMBO, B. 2023. Sparse and smooth functional data clustering. *Statistical Papers*, 1–31.

FLORIELLO, D., & VITELLI, V. 2017. Sparse clustering of functional data. *Journal of Multivariate Analysis*, **154**, 1–18.

GUO, J., LEVINA, E., MICHAILIDIS, G., & ZHU, J. 2010. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, **66**(3), 793–804.

PAN, W:, & SHEN, X. 2007. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, **8**(May), 1145–1164.

RAMSAY, J. O., & SILVERMAN, B. W. 2005. *Functional data analysis*. Wiley Online Library.

VITELLI, V. 2023. A novel framework for joint sparse clustering and alignment of functional data. *Journal of Nonparametric Statistics*, 1–30.

WITTEN, D. M., & TIBSHIRANI, R. 2010. A framework for feature selection in clustering. *Journal of the American Statistical Association*, **105**(490), 713–726.