Distance-based probabilistic clustering for functional data

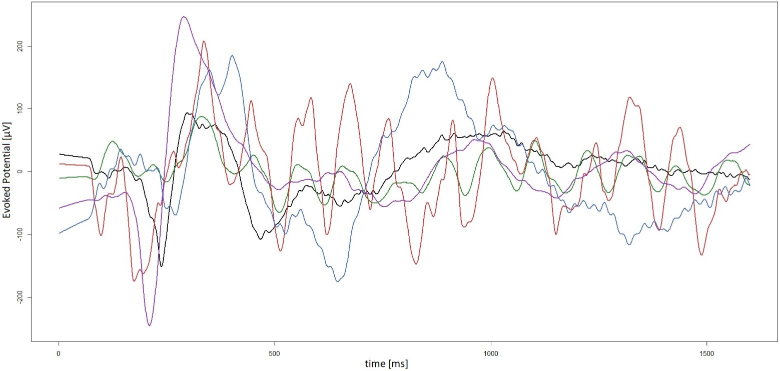
Group members: Giulia Caruso, Alessio Facincani, Giulia Romani, Pietro Spina, Matteo Vescovi
Tutors: Mario Beraha, Riccardo Corradin

POLITECNICO
MILANO 1863

Framework of the project

26 **multivariate functional observations**, each component is a function observed at 1600 time points.

RESEARCH GOAL: **Cluster** observations in different groups, first considering only one functional component.



Model

Prior and posterior

- *Gibbs posterior*:

$$\pi(\boldsymbol{c}|\lambda, \boldsymbol{X}) \propto \pi(\boldsymbol{c}) \exp\{-\lambda \ell(\boldsymbol{c}; \boldsymbol{X})\}$$

- *Loss function* under a GB-PPM (Generalized Bayes Product Partition Model) with *Mahalanobis distance* $M_\alpha(\boldsymbol{x_i}; \boldsymbol{X_k}) \geq 0$:

$$\ell(\boldsymbol{c}; \boldsymbol{X}) = \sum_{k=1}^{K} \sum_{i \in C_K} \mathcal{D}(\boldsymbol{x_i}; \boldsymbol{X_k}) = \sum_{k=1}^{K} \sum_{i \in C_K} M_\alpha(\boldsymbol{x_i}; \boldsymbol{X_k})$$

- *Generalized Bayes posterior under a GB-PPM*:

$$\pi(\boldsymbol{c}|\lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} exp\{-\lambda \sum_{i \in C_K} M_\alpha(\boldsymbol{x_i}; \boldsymbol{X_k})\}$$

- *Uniform prior (Stirling number of the second kind)*:

$$\boldsymbol{\pi}(\boldsymbol{c}) = \frac{1}{\mathcal{S}(n, K)}$$

Mahalanobis distance in a functional context

Given a stochastic process $X(t) \in \mathbb{L}^2[0,1],\ t \in [0,1]$:

- *Covariance function K and operator $\mathcal{K}$*:

$$K(s,t) = Cov(X(s), X(t)) \qquad \mathcal{K}f(t) = \int_0^1 K(t,s)f(s)ds$$

- *α-Mahalanobis distance* with *smoothing parameter $\alpha > 0$*: $(\forall x, y \in \mathbb{L}^2[0,1])$

$$M_\alpha(x,y)^2 = \|x_\alpha - y_\alpha\|_K^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)^2} \langle x - y, e_j \rangle^2$$

- *α-approximation* of a function $x \in \mathbb{L}^2[0,1]$:

$$x_\alpha = \underset{f \in \mathcal{H}(K)}{\arg\min} \|x - f\|^2 + \alpha \|f\|_K^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)} \langle x, e_j \rangle e_j$$

Maximum a posteriori estimation (MAP)

Target of inference: optimal and unknown partition c_{opt}

$$
\begin{aligned}
c_{opt} &= \underset{c\ :\ |C|=K}{\arg\max}\, \pi(\boldsymbol{c}|\lambda, \boldsymbol{X}) \\
&= \underset{c\ :\ |C|=K}{\arg\min}\, \ell(\boldsymbol{c}; \boldsymbol{X}) \\
&= \underset{c\ :\ |C|=K}{\arg\min} \sum_{k=1}^{K} \sum_{i \in C_K} M_\alpha(\boldsymbol{x_i}; \boldsymbol{X_k})
\end{aligned}
$$

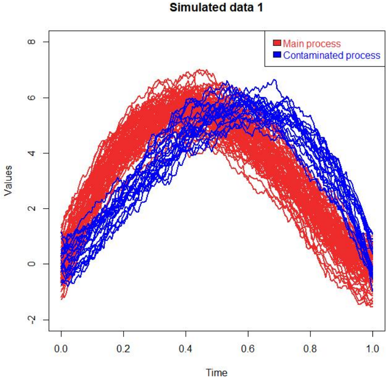
Where the second equality is justified by the uniform prior assumption.

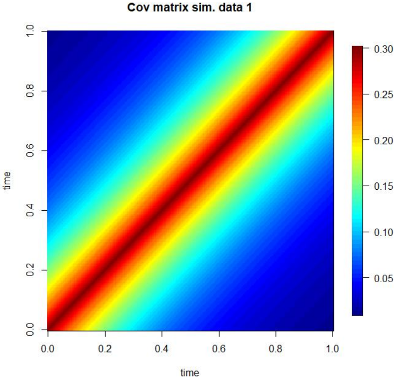
Posterior inference

Simulated data 1

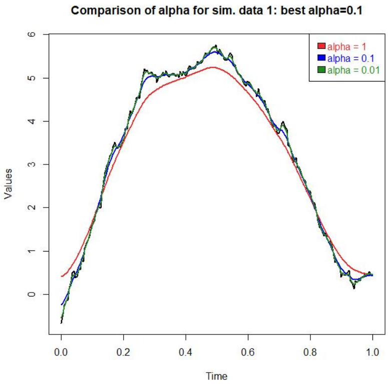
- Main process: $X(t) = 30t(1-t)^{3/2} + \epsilon(t)$
- Contaminated process: $X(t) = 30t^{3/2}(1-t) + \epsilon(t) \quad t \in [0,1]$

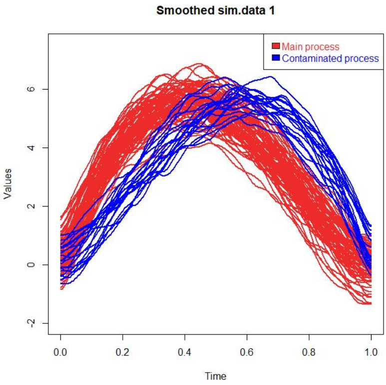
$\epsilon(t) \sim \mathrm{GP}(0, \mathcal{C})$ where $\mathcal{C}(s,t) = 0.3 \cdot exp(-|s-t|/0.3)$





Simulated data 1: best $\alpha = 10^{-1}$





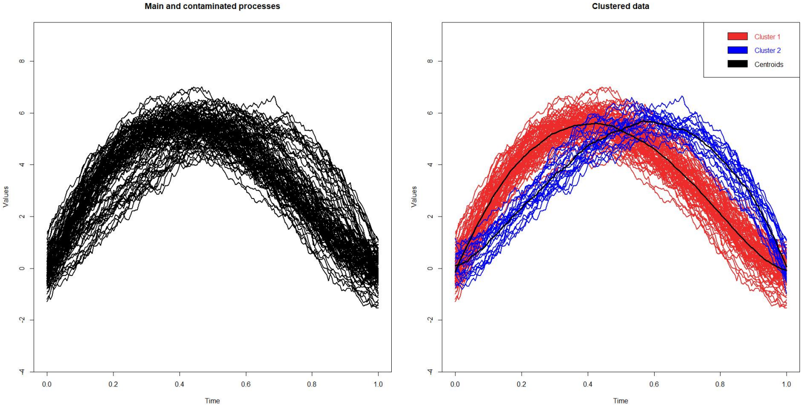
Algorithm with "fixed" covariance

Algorithm 0: Mahalanobis dist. clustering with fixed covariance

Input : n° clusters, covariance matrix, α, toll, data
Output: Optimal partition (labels, centroids and loss)

Sample random observations as initial centroids $\mathbf{m}_1, \ldots, \mathbf{m}_K$
Compute eigenvalues and eigenvectors of the covariance matrix
Initialize $\ell_1(\boldsymbol{c}; \boldsymbol{X})$ and $\ell_2(\boldsymbol{c}; \boldsymbol{X})$
while $(\ell_1(\boldsymbol{c}; \boldsymbol{X}) - \ell_2(\boldsymbol{c}; \boldsymbol{X})) > \mathit{toll}$ **do**
    $\ell_1(\boldsymbol{c}; \boldsymbol{X}) = \ell_2(\boldsymbol{c}; \boldsymbol{X})$
    for *i=1,...,n* **do**
        Set the cluster indicator c_i equal to k, so that $M_\alpha(\mathbf{x}_i, \mathbf{m}_k)$ is minimum
    end
    for *k=1,...,K* **do**
        Set $\mathbf{m}_k$ as the functional mean of the observations belonging to cluster k
    end
    Update $\ell_2(\boldsymbol{c}; \boldsymbol{X})$
end

Application on simulated data 1



Algorithm with "updated" covariance

Algorithm 1: Mahalanobis dist. clustering with covariance updating

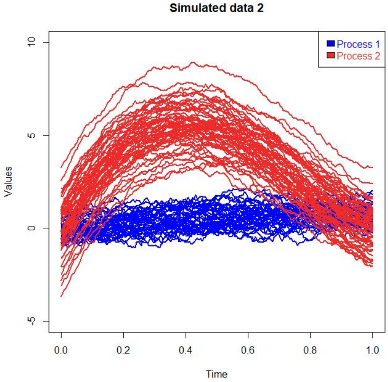
Input : n° clusters, covariance matrix, α, toll, data
Output: Optimal partition (labels, centroids and loss)

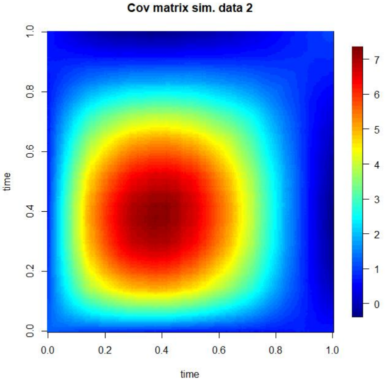
```
Sample random observations as initial centroids m_1, ..., m_K;
Compute eigenvalues and eigenvectors of the covariance matrix;
Initialize ℓ_1(c; X) and ℓ_2(c; X);
while (ℓ_1(c; X) − ℓ_2(c; X)) > toll do
    ℓ_1(c; X) = ℓ_2(c; X) ;
    for i=1,...,n do
        Set the cluster indicator c_i equal to k, so that M_α(x_i, m_k) is minimum
    end
    for k=1,...,K do
        Set m_k as the functional mean of the observations belonging to cluster k;
        Set cov_k as the covariance matrix of the k-th cluster and compute its eigenvalues
          and eigenvectors;
    end
    Update ℓ_2(c; X)
end
```

Simulated data 2

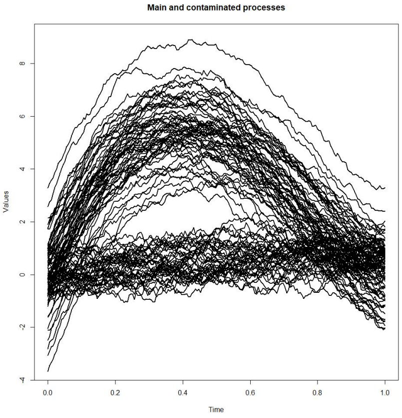
Data clustered into two groups:

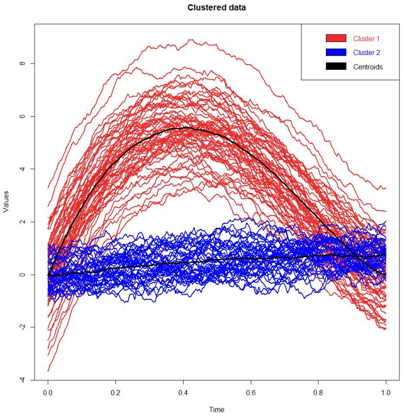
- First process: $X(t) = sin(t) + \epsilon_1(t) \quad \epsilon_1(t) \sim \text{GP }(0, \mathcal{C}_1)$
  $\mathcal{C}_1(s,t) = 0.3 \cdot exp(-|s-t|/0.3)$
- Second process: $X(t) = 30t(1-t)^{3/2} + \epsilon_2(t) \quad \epsilon_2(t) \sim \text{GP }(0, \mathcal{C}_2)$
  $\mathcal{C}_2(s,t) = 1.5 \cdot exp(-|s-t|/3) \qquad (t \in [0,1])$



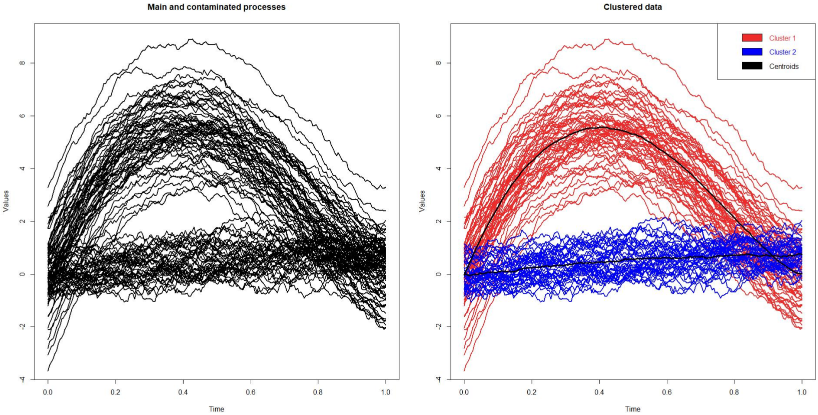


Application on simulated data 2





Application on simulated data 2



Problem: the number of clusters has to be given a priori

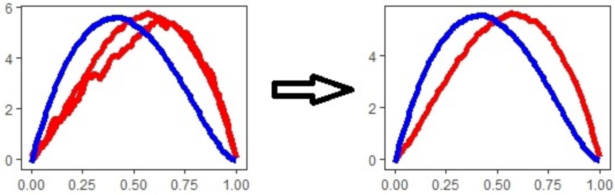
Union of similar clusters

Algorithm 2: Union of similar clusters

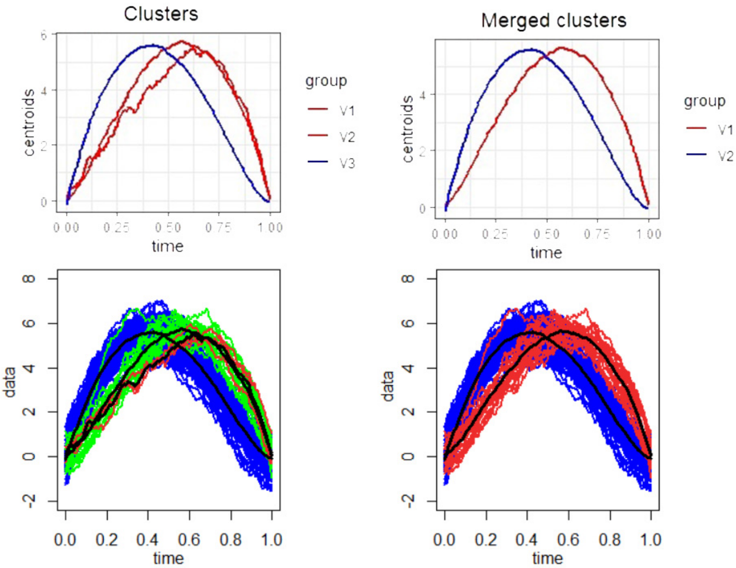
Input : optimal centroids $\mathbf{m}_1, ..., \mathbf{m}_k$

Output: united centroids where needed

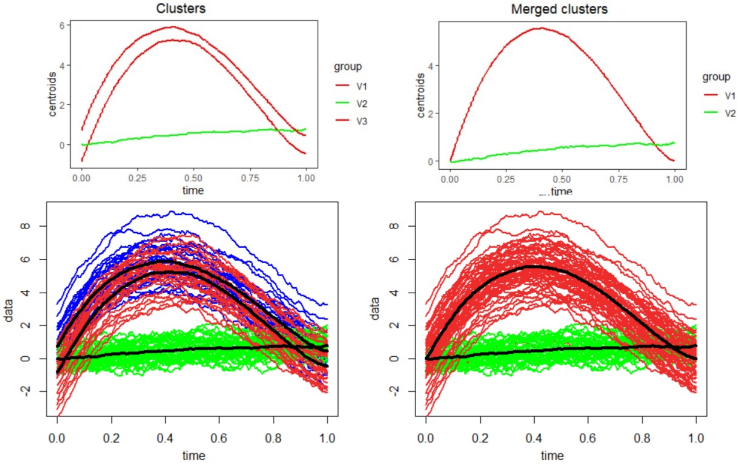
Compute distances matrix d : $d_{ij} = \|\mathbf{m}_i - \mathbf{m}_j\|_2^2 \quad \forall i, j$;
Set $\varepsilon = 0.5 \cdot (median(d) + mean(d))$;
if $d_{ij} < \varepsilon$ **then**
  Merge clusters i, j
end
Recompute the centroids and repeat until $d_{i_{new} j_{new}} > \varepsilon_{new} \quad \forall i_{new}, j_{new}$



Union of similar clusters: Simulated data 1



Union of similar clusters: Simulated data 2



Next steps

Next steps

- Relax the uniform prior distribution $\pi(\boldsymbol{c})$ to a prior that allows "small cluster" penalization
- Application on real data:
    - Tuning of the smoothing parameter α
- Further development: Gibbs sampling strategy for uncertainty quantification

References

References

[1] José R. Berrendero, Beatriz Bueno-Larraz, and Antonio Cuevas. "On Mahalanobis Distance in Functional Settings". In: *Journal of Machine Learning Research* 21 (2020), pp. 1–33.

[2] Tommaso Rigon, Amy H. Herring, and David B. Dunson. "A generalized Bayes framework for probabilistic clustering". In: *arXiv:2006.05451* (2020).

[3] Sara Wade and Zoubin Ghahramani. "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)". In: *Bayesian Anal.* 13 (2) (2018), pp. 559–626.