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

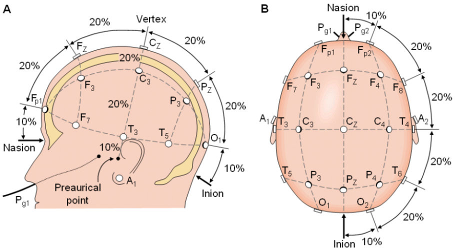
Table of contents

Data description

Data

The sample consists of data coming from 26 patients who suffered from cerebral lesions. For each subject we are provided with:

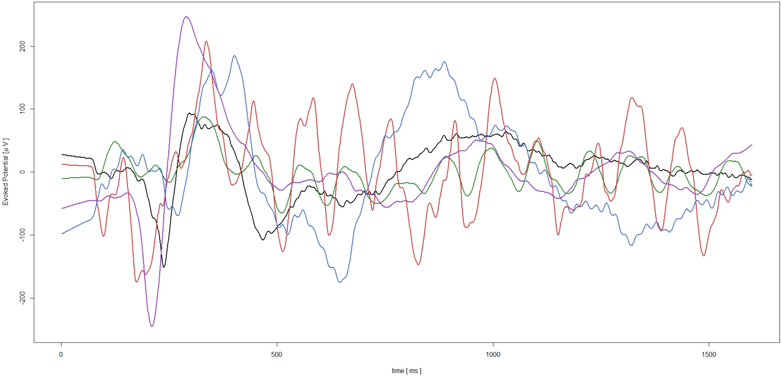
- **Functional data** representing the capability to receive an electric stimulus
- **Covariates** describing the duration of the coma state and the rehabilitation characteristics
- **Outcome surveys** assessing the functional outcome and the level of responsiveness after rehabilitation



Functional data

Each functional datum consists of 1600 measurements of the evolution over time of the sensory evoked potential.

For every patient, **four electrodes** placed in four different areas of the cortex **measure the potential** simultaneously.



Goal

Goal

Research goal: **Cluster** our functional data in K different groups, only considering the response functions of patients.

How?

- By the use of a generalized Bayes approach, bridging loss and model based clustering techniques. Namely, substituting the *likelihood* in a generalized Bayes model:

$$\mathcal{L}(\boldsymbol{c}; \boldsymbol{X}; \lambda) = e^{-\lambda \ell(\boldsymbol{c}; \boldsymbol{X})}$$

- Exploiting an extension of the *Mahalanobis distance* to the functional setting

Model explanation

Setting

Define:

- $\boldsymbol{x_i} = (x_{i1}, ..., x_{id})^T$: vector of observations ($\forall i = 1, ..., n$)
- $\boldsymbol{X}$: collection of all data points
- $\boldsymbol{K}$: number of clusters (fixed a priori)
- $\boldsymbol{C} = (C_1, ..., C_K)$: partition of {1,...,n} into $\boldsymbol{K}$ sets
- $\boldsymbol{X_k} = \{\boldsymbol{x_i} : i \in C_k\}$: k-th cluster
- $\boldsymbol{c} = (c_1, ..., c_n)$: cluster labels: $c_i = k$ iff $i \in C_k$ ($\forall i = 1, ..., n$)

Gibbs posterior and GB-PPM loss

Introduce the **Gibbs posterior**:

$$\pi(\boldsymbol{c}|\lambda, \boldsymbol{X}) \propto \pi(\boldsymbol{c}) exp\left\{-\lambda \ell(\boldsymbol{c}; \boldsymbol{X})\right\} \tag{1}$$

with parameter $\lambda > 0$ and loss function $\ell(\boldsymbol{c}; \boldsymbol{X}) > 0$

The class of **Generalized Bayes product partition models** (GB-PPM) for clustering is characterized by a factorized loss function of the form:

$$\ell(\boldsymbol{c}; \boldsymbol{X}) = \sum_{k=1}^{K} \sum_{i \in C_K} \mathcal{D}(\boldsymbol{x_i}; \boldsymbol{X_k}) \tag{2}$$

where $\mathcal{D}(\boldsymbol{x_i}; \boldsymbol{X_k}) \geq 0$ is some distance function.

Prior and final posterior distribution

Substituting the loss (2) into the Gibbs posterior (1), the **generalized Bayes posterior under a GB-PPM** has the form:

$$\pi(\boldsymbol{c}|\lambda, \boldsymbol{X}) \propto \boldsymbol{\pi}(\boldsymbol{c}) \prod_{k=1}^{K} \rho(C_k; \lambda, X_k) \propto \prod_{k=1}^{K} exp\{-\lambda \sum_{i \in C_K} \mathcal{D}(\boldsymbol{x_i}; \boldsymbol{X_k})\} \quad (3)$$

Having assumed a uniform clustering prior: $\boldsymbol{\pi}(\boldsymbol{c}) = \frac{1}{\mathcal{S}(n,K)}$, where $\mathcal{S}(n, K)$ is the Stirling number of the second kind.

Question: what is a good distance $\mathcal{D}$ to choose?

Mahalanobis distance for functional data

Mahalanobis distance: finite-dimensional set

Given two points $x, y \in \mathbb{R}^d$ with non-singular covariance matrix Σ, the **Mahalanobis distance** is defined as:

$$M(x, y) = ((x - y)'\Sigma^{-1}(x - y))^{1/2}$$

With functional data, given a stochastic process $X(t) \in \mathbb{L}^2[0, 1]$, $t \in [0, 1]$, define:

- The **covariance function** K = K(s,t) = Cov(X(s),X(t))
- The **covariance operator** $\mathcal{K}$: $\mathcal{K}f(t) = \int_0^1 K(t, s)f(s)ds$

Problem: $\mathcal{K}$ is typically not invertible

Mahalanobis distance: functional data context

The α-**Mahalanobis distance** is defined as ($\forall x, y \in \mathbb{L}^2[0,1]$):

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

where:

- $\alpha > 0$ penalization parameter
- $x_\alpha = \underset{f \in \mathcal{H}(K)}{\arg\min} \|x - f\|^2 + \alpha \|f\|_K^2 = (\mathcal{K} + \alpha \mathbb{I})^{-1} \mathcal{K} x =$
  $= \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)} \langle x, e_j \rangle e_j$ approximation of x in $\mathcal{H}(K)$
- $\|f\|_K^2 = \sum_{j=1}^{\infty} \frac{\langle f, e_j \rangle^2}{\lambda_j}$ with $(e_j, \lambda_j > 0)$ eigenfunctions and eigenvalues of $\mathcal{K}$

α-Mahalanobis distance: properties

- Defines a metric in $\mathbb{L}^2[0,1]$
- Is continuous and decreasing wrt the tuning parameter α
- Is invariant wrt isometries
- Given:
  - ◇ a process $X(t)$ with mean m
  - ◇ a sample of observations $X_1(t), ..., X_n(t)$ with sample mean $\bar{X}$, covariance function $\hat{K}(s,t) = \frac{1}{n}\sum_{i=1}^{n}(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))$ and covariance operator $\hat{\mathcal{K}}$

  an estimator for $M_\alpha(X, m)$ is:

$$\widehat{M}_{\alpha,n}(X, \bar{X}) := \|\widehat{X}_\alpha - \bar{X}_\alpha\|_{\hat{K}}$$

  where $\widehat{X}_\alpha$ and $\bar{X}_\alpha$ are sample approximations of X and $\bar{X}$ in $\mathcal{H}(\hat{K})$.
  It converges as $n \uparrow \infty$:

$$\widehat{M}_{\alpha,n}(f, \bar{X}) \xrightarrow{n} M_\alpha(f, m) \qquad \forall f \in \mathbb{L}^2[0,1]$$

Next step

Next step

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

Two possible ways:

1. **MAP**: maximum a posteriori estimation (does not depend on λ): $c_{opt} = \underset{c \,:\, |C|=K}{\arg\min} \sum_{k=1}^{K} \sum_{i \in C_K} \mathcal{D}(\boldsymbol{x_i}; \boldsymbol{X_k})$
2. **NON-MAP**: use Gibbs sampling, i.e. re-allocate the indicators c_i by sampling from their full conditionals.

Further developments:

- Relax the uniform distribution $\boldsymbol{\pi(c)}$ to a more general one
- Perform sensitivity analysis wrt the cluster number K

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).