Local Bayesian clustering for functional data

Giovanni Toto, Antonio Canale

Department of Statistical Sciences, University of Padova

UNIVERSITÀ DEGLI STUDI DI PADOVA

Introduction

For standard Euclidean data, clustering is quite straightforward since a point may belong to only a cluster. On the other hand, for functional data the problem becomes more complex since one may perform

- **clustering at global level**: a whole function is assigned to a single cluster,
- **clustering at local level**: the same function may belong to different clusters depending on the point of its domain at which it is evaluated.

Goal: performing local clustering for functional data.

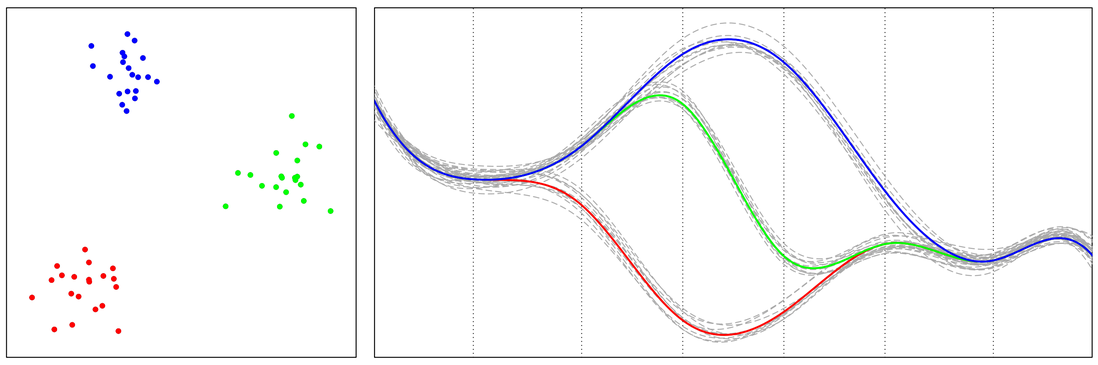


Figure 1: Comparison of a clustering problem for non-functional and functional data.

Contributions

We propose a Bayesian approach for the analysis of functional data that

- exploits the local property of B-spline basis expansion to perform indirect local clustering,
- defines **unit-specific B-spline parameters in terms of unit- and basis-specific cluster assignments**,
- employs an **ad-hoc definition for contiguous cluster-specific parameters ensuring smooth functions**,
- employs a novel **dependent random partition model inducing sequences of random partitions exhibiting semi-Markovian dependence**.

Modeling expected values via B-spline basis expansion

We assume that a random curve $Y_i(x)$, evaluated at the point $x \in \mathbb{R}$, follows

$$Y_i(x) \mid \boldsymbol{\theta}^*, \mathbf{c}_i, \sigma^2 \overset{ind}{\sim} \mathsf{N}\left(\boldsymbol{b}(x)^\top \boldsymbol{\theta}_i, \sigma^2\right),$$

where $\boldsymbol{b}(x)$ is a d-degree B-spline with basis coefficient

$$\boldsymbol{\theta}_i = (\theta^*_{1,c_{i,1}}, \ldots, \theta^*_{K,c_{i,K}}),$$

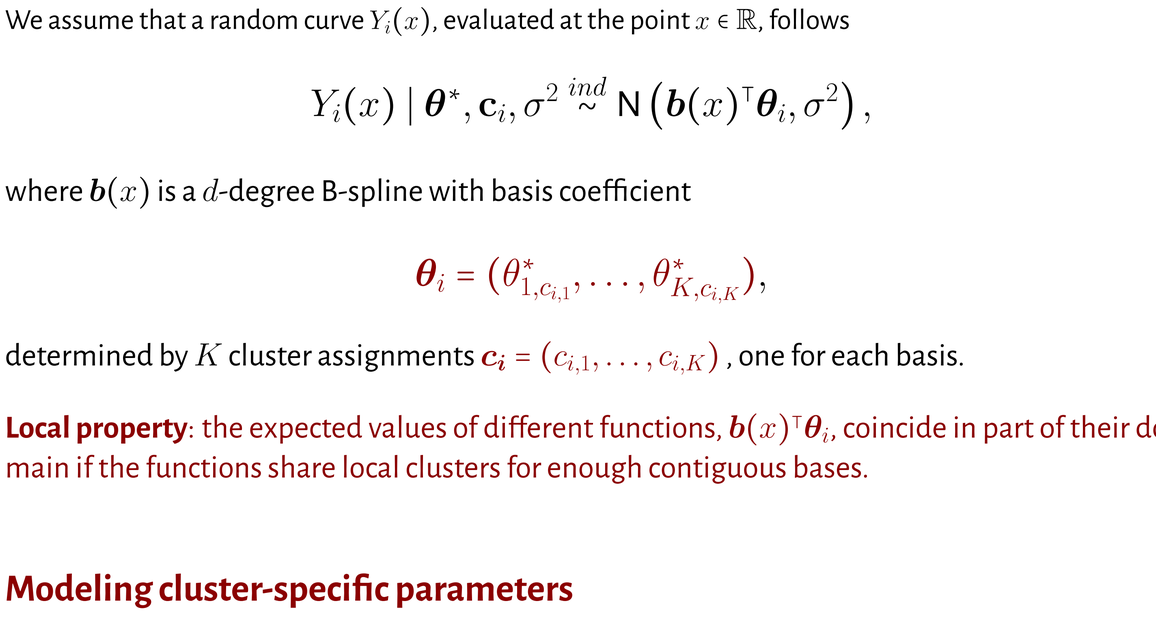
determined by K cluster assignments $\boldsymbol{c_i} = (c_{i,1}, \ldots, c_{i,K})$, one for each basis.

Local property: the expected values of different functions, $\boldsymbol{b}(x)^\top \boldsymbol{\theta}_i$, coincide in part of their domain if the functions share local clusters for enough contiguous bases.

Modeling cluster-specific parameters

We consider a partition of the curves at each basis, ρ_k, and define the cluster-specific parameters is such a way that contiguous parameters in each $\boldsymbol{\theta}_i$ are similar:

$$\theta^*_{kj} \mid \boldsymbol{\theta}^*_{k-1}, \rho_k, \rho_{k-1}, \phi, \tau^2 \overset{ind}{\sim} \mathsf{N}\left(\frac{\phi}{|C^{(\to j)}_{k-1}|} \sum_{l \in C^{(\to j)}_{k-1}} \theta^*_{k-1,l}, \tau^2\right).$$



Semi-Markovian Random Partition Model (smRPM)

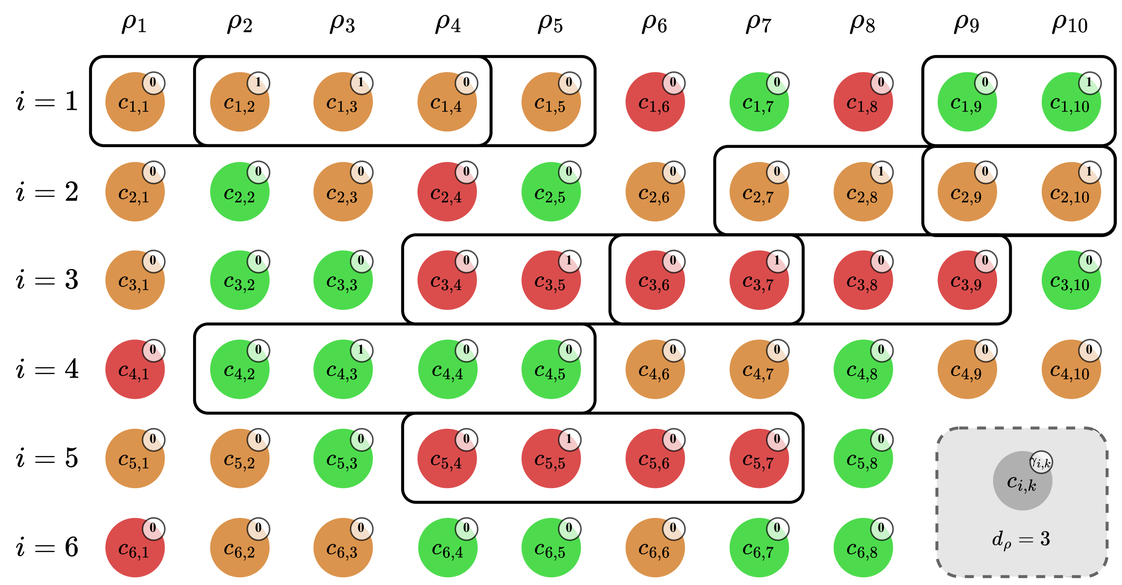
We introduce auxiliary variables explicitly modeling the evolution of the partitions at different basis:

$$\gamma_{ik} = \begin{cases} 1 & \text{if curve } i \textit{ cannot} \text{ be reallocated when moving from basis } k-1 \text{ to } k+d_\rho-1 \\ 0 & \text{otherwise} \end{cases}.$$

The distribution of the partition at basis k is influenced by the auxiliary variables related to bases $k, \ldots, k-d_\rho+1$, collected in $\boldsymbol{\gamma}^{(d_\rho)}_k$:

$$\Pr(\rho_k = \lambda \mid \boldsymbol{\gamma}^{(d_\rho)}_k, \rho_{k-1}) = \frac{\Pr(\rho_k = \lambda)\mathbb{I}\left(\lambda \in P_{\mathcal{R}_k}\right)}{\sum_{\lambda' \in P} \Pr(\rho_k = \lambda')\mathbb{I}\left(\lambda' \in P_{\mathcal{R}_k}\right)},$$

where $P_{\mathcal{R}_k}$ is the set of partitions that are compatible with ρ_{k-1} based on $\boldsymbol{\gamma}^{(d_\rho)}_k$.



The probabilities of success of the auxiliary variables at different basis can be independent,

$$\gamma_{ik} \mid \alpha_k \overset{ind}{\sim} \mathsf{Ber}(\alpha_k),$$
$$\alpha_k \overset{iid}{\sim} \mathsf{Beta}(a_\alpha, b_\alpha),$$

or have d_γ-order dependence,

$$\gamma_{ik} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}^{(d_\gamma)}_{i,k-1} \overset{ind}{\sim} \mathsf{Ber}\left(\pi\left(\alpha_0 + \alpha_1 \sum_{q=1}^{d_\gamma} \gamma_{i,k-q}\right)\right),$$
$$\boldsymbol{\alpha} \sim \mathsf{N}_2(\mathbf{a}, \mathbf{A}), \quad \omega_{ik} \overset{iid}{\sim} \mathsf{PG}(0,1). \qquad \text{(Polson et al. (2013))}$$

Notation: assuming $\rho_1 \sim \mathrm{CRP}(M)$, the model is denoted as $\mathrm{smRPM}_{d_\rho,d_\gamma}(\boldsymbol{\alpha}, M)$.

If $d_\rho = 1, d_\gamma = 0$, the model coincides with temporal Random Partition Model (Page et al., 2022).

Simulations

We define $n^{(ref)} = 5$ reference functional observations with known cluster assignments, and we simulate $R = 50$ datasets containing $n^{(rep)} \in \{10, 30\}$ realizations of each of these reference functional observations.

We simulate the observations and the cluster-specific parameters as

$$Y_i(x) \mid \boldsymbol{\theta}^*, \mathbf{c}_i, \sigma^2 \overset{ind}{\sim} \mathsf{N}\left(\boldsymbol{b}(x)^\top \boldsymbol{\theta}_i, \sigma^2\right),$$
$$\theta^*_{1j} \overset{ind}{\sim} (10j, 5), \quad \theta^*_{kj} \mid \boldsymbol{\theta}^*_{k-1}, \rho_k, \rho_{k-1} \overset{ind}{\sim} \left(\frac{1}{|C^{(\to j)}_{k-1}|} \sum_{l \in C^{(\to j)}_{k-1}} \theta^*_{k-1,l}, 5\right),$$

where $\sigma^2 \in \{1, 4\}$ quantifies the noise added to the reference functional observations.

Evaluation: Adjusted Rand Index (Hubert and Arabie, 1985) computed on each MCMC iteration.

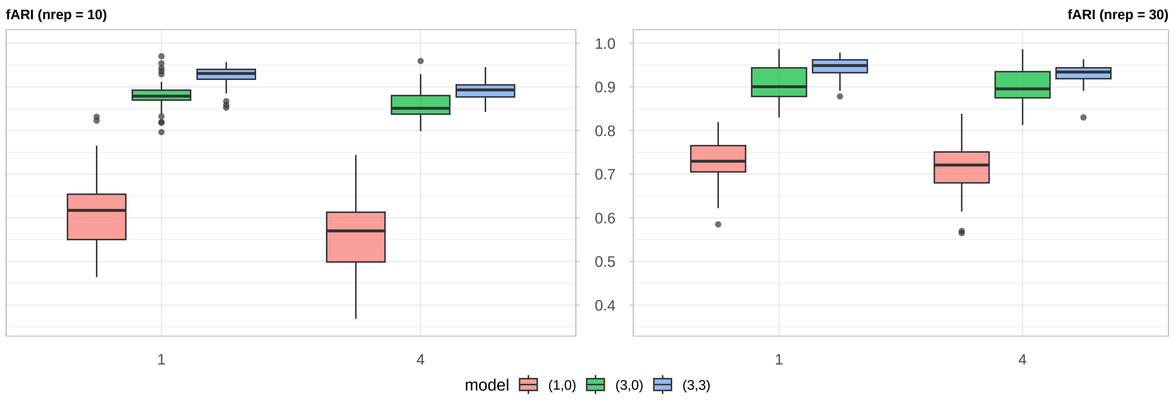


Figure 2: Boxplot of the average posterior ARI across the sequence of partitions computed on $R = 50$ simulated datasets.

References

Page, G. L., Quintana, F. A., Dahl, D. B. (2022). Dependent Modeling of Temporal Sequences of Random Partitions. In: Journal of Computational and Graphical Statistics, **31 (2)**, 614–627.

Hubert, L., and Arabie, P. (1985). Comparing Partitions. In: Journal of Classification, **2**, 193–218.

Polson, N. G., Scott, J. G., Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. In: Journal of the American Statistical Association, **108 (504)**, 1339–1349..