A novel regularized approach for functional data clustering: an application to milking kinetics in dairy goats

Christophe Denis

Joint work with:
E. Lebarbier, C. Lévy-Leduc, O. Martin, and L. Sansonnet

LAMA, Université Gustave Eiffel

26/03/2021

MIA-Toulouse, INRAE

Motivation

Objective

- valuable information on biological processes
- understanding of the variability in milk flow kinetics
- understanding of the lactation process
- controlling udder health

Milking kinetics

- classically described and classified through synthetic parameters
- consider milking kinetics as a whole function

Data description

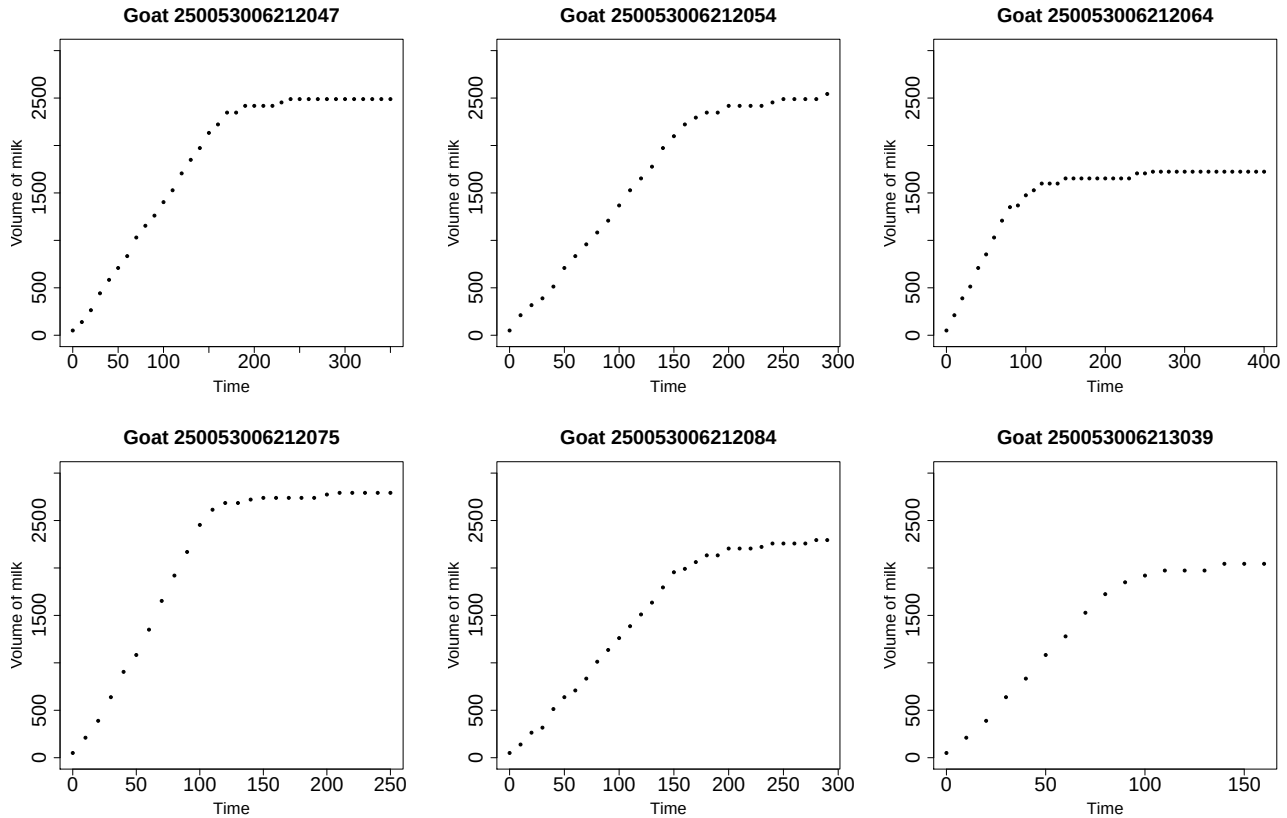
100470 milking kinetics of goats

- two breeds: Alpine and Saanen
- morning kinetics
- several kinetics by goat
- two different parity

487 goats

- 296 in parity 1 and 191 in parity 2
- kinetics observed each day around 5 month

Illustration



Statistical goal

Clustering of the milking kinetics

- for each parity provide a clustering of the kinetics
- non-decreasing curves
- heterogeneity of the milking kinetics
- change points detection

Clustering of goats according to the parity

- based on the resulting clustering of the trajectories
- provide a clustering of the goats

Functional data clustering (Jacques and Biernacki (2014))

Two-step approach

- dimension reduction step
- build a vector of summary measures for each trajectory
- apply unsupervised procedure to this vector

Nonparametric approach

- define suitable distance between trajectories
- perform `kmeans` type algorithm

Model-based approach

- consider Λ the vector of the expansion coefficients of each curve into an adapted basis B-spline basis)
- assume that Λ is distributed according to a Gaussian mixture
- perform EM type algorithm

Overview of the procedure: two-step approach

First step: for each curve

- estimate the change-points
- B-spline basis of order 2 defined by the change-points
- estimate the expansion coefficients of the curve into this basis

Second step

- summary measures: change points and coefficient estimates
- perform `kmeans` algorithm

Contributions

- new method to estimate the change points in the slop
- include the change points estimates in the clustering procedure

Change points estimation: *trend filtering*

Trend filtering (Tibshirani (2014))

- Observation of a curve $\mathbf{Y} = (Y_1, \ldots, Y_n)$ at $(x_1, \ldots, x_n)$.
- For $\lambda > 0$, consider

$$\widehat{\boldsymbol{\beta}}(\lambda) = \text{Argmin}_{\boldsymbol{\beta} \in \mathbb{R}^n} \left\{ \|\mathbf{Y} - \boldsymbol{\beta}\|_2^2 + \lambda \|D^{(2)} \boldsymbol{\beta}\|_1 \right\},$$

- $D^{(2)} = D^{(1)}.D^{(1)}$, with $D^{(1)}\beta = (\beta_{i+1} - \beta_i)_{i=1,\ldots,n-1}$.

Drawback of this approach

- tuning parameter λ, usually chosen by cross-validation
- omitted change points due to resampling
- leads to oversegmentation phenomena

Change points estimation: *find the most relevant*

To avoid the use of resampling methods

- choose $K_{\max}$ and consider all λ leading to $K_{\max}$ change points
- over this set compute $\hat{\lambda}$ the minimizer of $\|\mathbf{Y} - \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\|_2^2$
- change points indices $(\hat{n}_1, \ldots, \hat{n}_{K_{\max}})$

Find the relevant change points

- apply DP algorithm to the restricted set $(Y_{\hat{n}_1}, \ldots, Y_{\hat{n}_{K_{\max}}})$
- modification to make the piecewise linear fit to data continuous
- number of change points $\hat{K}$ is then chosen by using the criterion proposed by Lavielle (2005)
- resulting change points $(\hat{t}_1, \ldots, \hat{t}_{\hat{K}})$

Projection onto the B-spline basis

B-spline basis

- knots sequence $(\tau_1, \ldots, \tau_{\hat{K}+4})$
- $\tau_1 = \tau_2 = \hat{t}_0 = x_1$, $\tau_{\hat{K}+3} = \tau_{\hat{K}+4} = \hat{t}_n = x_n$
- $\tau_{j+2} = \hat{t}_j$, $j = 1, \ldots \hat{K}$

B-spline function

- $B_{i,2} = \frac{u-\tau_i}{\tau_{i+1}} B_{i,1}(u) + \frac{\tau_{i+2}-u}{\tau_{i+2}-\tau_{i+1}} B_{i+1,1}(u)$, $i = 1, \ldots, \hat{K}+2$
- $B_{i,1}(u) = \mathbf{1}_{[\tau_i, \tau_{i+1})}(u)$

Piecewise linear estimate

- $\hat{f}_{\boldsymbol{\tau}, \hat{\boldsymbol{\theta}}}(u) = \sum_{i=1}^{\hat{K}+2} \hat{\theta}_i B_{i,2}(u)$
- $\hat{\theta}_i$ obtained from a least square criterion
- $\hat{\theta}_i$ values of the estimated piecewise linear curves at the change points

Clustering of curves

For each curve $\mathbf{c}$

- $X_c = (\hat{\theta}_1, \ldots, \hat{\theta}_{K_c+2}, \hat{t}_1, \ldots, \hat{t}_{\hat{K}_c})$
- $K_M = \max_c K_c$

Define $\tilde{\boldsymbol{X}}_{\boldsymbol{c}} \in \mathbf{R}^{\boldsymbol{2K_M+2}}$

- $\tilde{X}_c = X_c$ if $\hat{K}_c = K_M$
- else missing t_k and θ_k replace by 0

$$\hookrightarrow \quad \tilde{X}_c = (\hat{\theta}_1, \ldots, \hat{\theta}_{K_c+2}, 0, ., 0, \hat{t}_1, \ldots, \hat{t}_{\hat{K}_c}, 0, ., 0)$$

Clustering part

- centered and normalize $(\tilde{X}_c)_{c \in \mathcal{C}}$
- apply K-means to this dataset

k-means algorithm

k-means objective

- find a partition S of the data into k clusters
- minimize $\arg\min_S \sum_{i=1}^{K} |S_i| \mathrm{Var}(S_i)$

Lloyd heuristic

- start with a set of k random points (centers)
- repeat until convergence
  $\hookrightarrow$ assign each data point to its nearest center (k clusters)
  $\hookrightarrow$ for each cluster, update its center as the average of its points.

Number of cluster k

- Different criterion used to chose k, Charrad *et al* (2014)
- k is chosen by using majority voting rule

Simulation study: *simulation scheme*

Model

- complete observed data $(\mathbf{Y}, Z)$
- $\mathbf{Y}$ observations at $x_i = 10(i-1)$, $i = 1, \dots 51$
- $Z \in \{1, 2, 3, 4\}$ label associated to $\mathbf{Y}$
- cluster $\mathcal{C}_Z$ defined by $\{K_Z, \mathbf{t}^Z, \theta^Z\}$

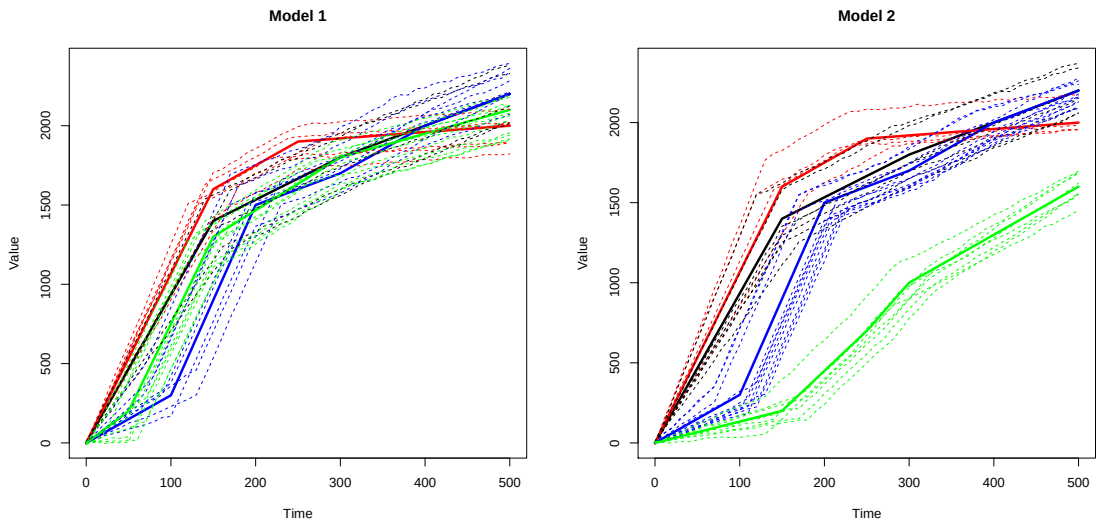
Simulation scheme

- label z is drawn from $\mathcal{U}(\{1, 2, 3, 4\})$
- generate $\tilde{\mathbf{t}}^z = \mathbf{t}^z + U$ and $\tilde{\boldsymbol{\theta}}^z = \boldsymbol{\theta}^z + V$ (U and V uniform)
- based on $(0, \tilde{\mathbf{t}}^z, 500)$ and $(0, \tilde{\boldsymbol{\theta}}^z)$ compute $f_{\tilde{\mathbf{t}}^z, \tilde{\boldsymbol{\theta}}^z}(.)$
- define $Y_i = f_{\tilde{\mathbf{t}}^z, \tilde{\boldsymbol{\theta}}^z}(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Simulation study: *model*

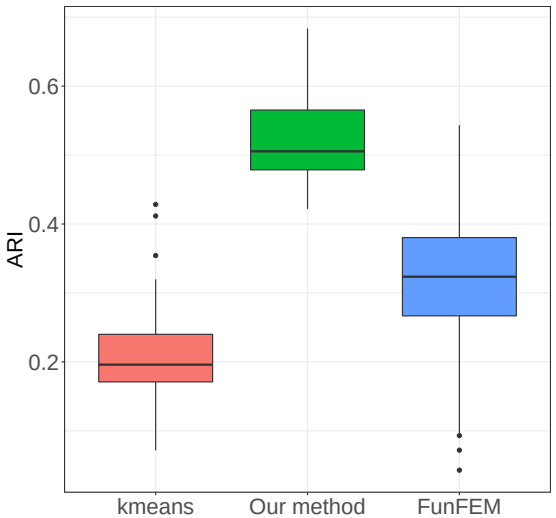
Model 1			
z	K_z	t^z	θ^z
1	2	$(150, 250)$	$(1600, 1900, 2000)$
2	2	$(150, 300)$	$(1400, 1800, 2200)$
3	4	$(100, 200, 300, 400)$	$(300, 1500, 1700, 2000, 2200)$
4	3	$(50, 150, 300)$	$(200, 1300, 1800, 2100)$

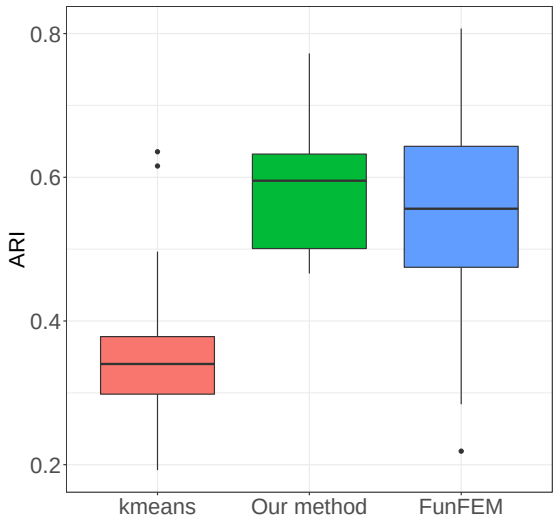
Model 2			
z	K_z	t^z	θ^z
1	2	$(150, 250)$	$(1600, 1900, 2000)$
2	2	$(150, 300)$	$(1400, 1800, 2200)$
3	4	$(100, 200, 300, 400)$	$(300, 1500, 1700, 2000, 2200)$
4	3	$(150, 250, 300)$	$(200, 700, 1000, 1600)$

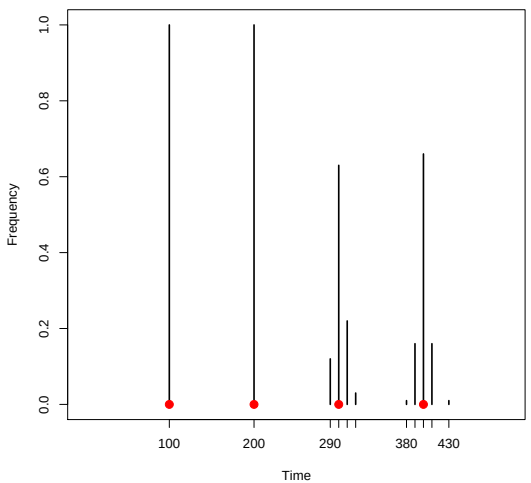


Simulation study: *results*

- Comparison with `FunFEM` (Bouveyron *et al* (2015))
- $K_{max} = 10$, $\sigma = 5$



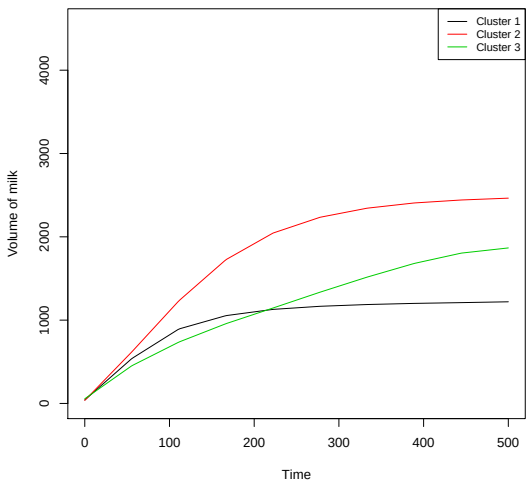




Application to real data: *kinetics clustering (1/2)*

Description

- $K_{\max} = 2$
- 100470 kinetics
- 36757 in cluster 1 , 57498 in cluster 2 , 6215 in cluster 3

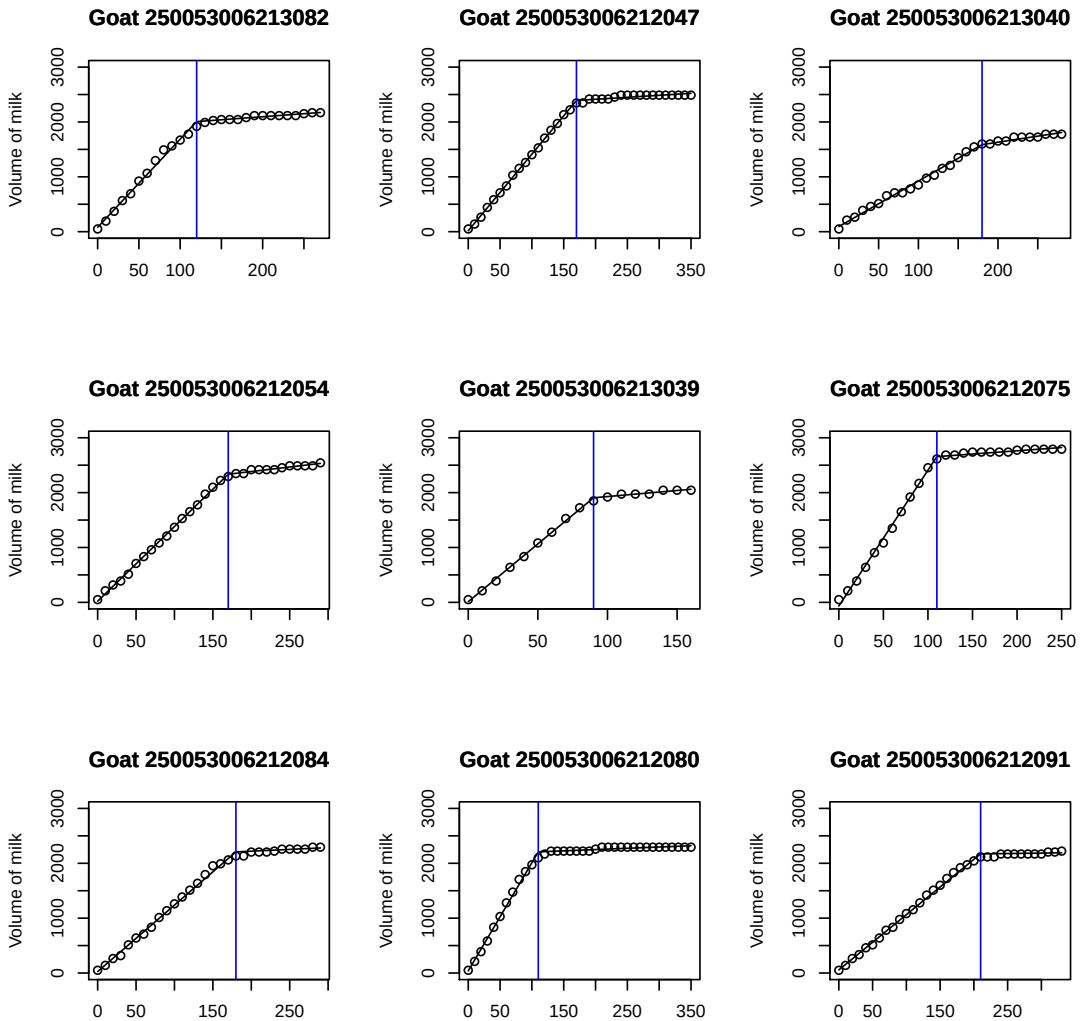


Interpretation

- number of change points 1 for cluster 1 and 2, 2 for cluster 3
- clusters can be discriminated according to the milk production
- cluster 1 and 2 differ from the change points location

Application to real data: *kinetics clustering (2/2)*

- Example of trajectories belonging to cluster 2



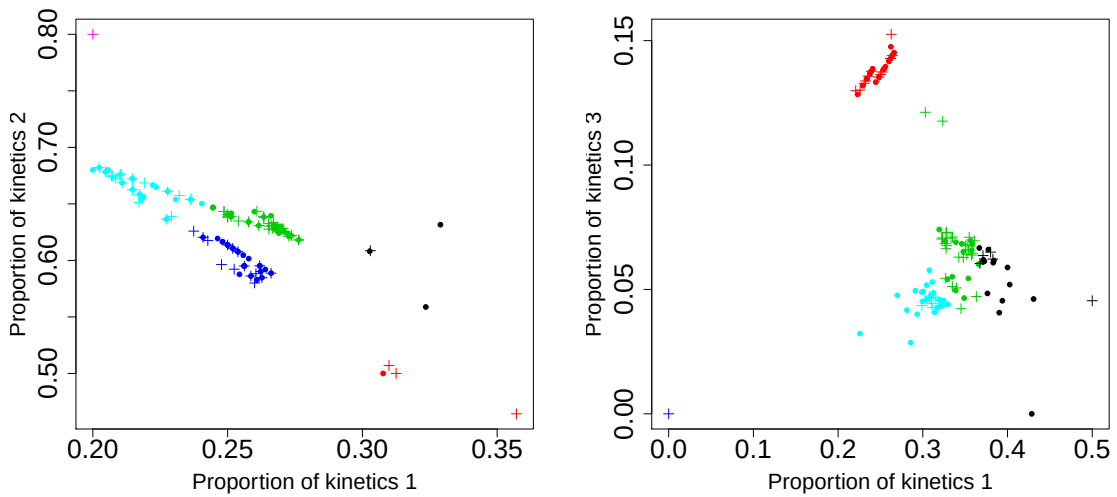
Application to real data: *goats clustering*

Method

- for a given parity
- for each goat, we compute a vector of proportions of its kinetics belonging to clusters 1,2 and 3
- apply k-means algorithm to the vectors of proportions.

Results

- 6 clusters for parity 1 (left) and 5 clusters for parity 2 (right)



Conclusion

Proposed procedure

- functional data clustering
- two-step procedure which involves change points estimation
- good performance and low computational burden

Analysis of the milking kinetics

- clustering of the kinetics can provide a relevant characterization of their shape
- Further analysis to propose options for individual milking management