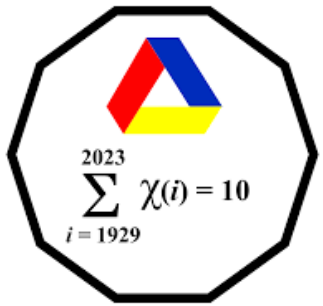
Learning with categorical functional data

Cristian Preda[1,2]

[1]Université de Lille, MODAL/Inria Lille Nord-Europe
[2] ISMMA Romanian Academy

The Tenth Congress of Romanian Mathematicians



June 30 - July 5, 2023 Pitești, Romania

Functional data

Definition [Ferraty and Vieu (2006)] A random variable X is called *functional* if it takes values in some infinite dimensional space. An observation of X is called a *functional data*.

Functional data

Model : Stochastic process,

$$X = \{X_t, \quad t \in \mathcal{T}\},$$

for some (continuous) index set $\mathcal{T}$

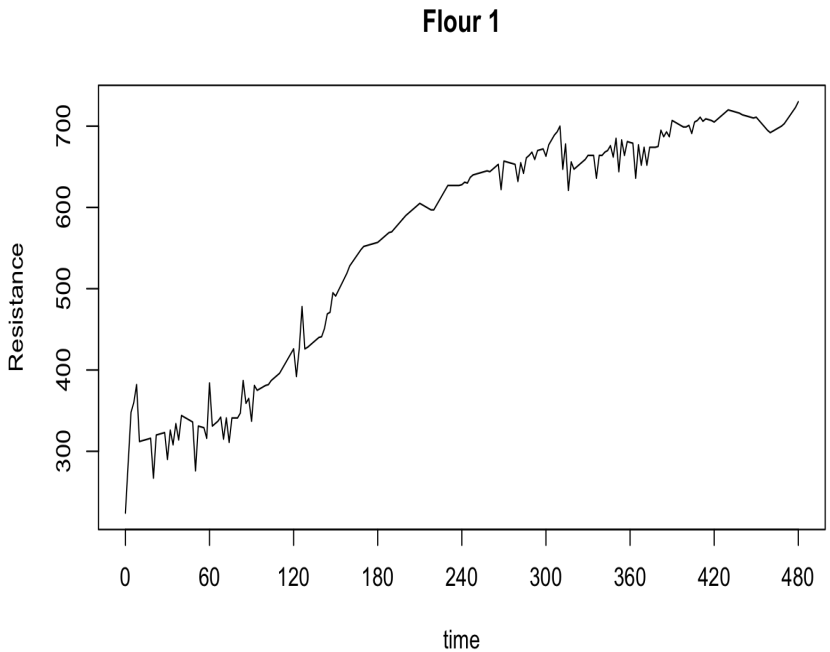
Common examples.

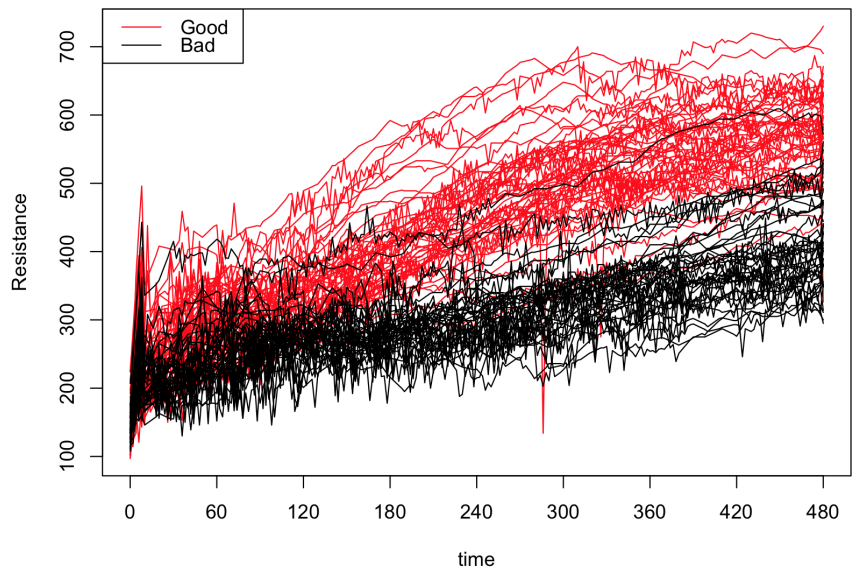
- $\mathcal{T} = [0, T],\ T > 0.$
  $X_t \in \mathbb{R}$: *univariate* functional data
  $X_t \in \mathbb{R}^p,\ p \geq 2$: *multivariate* functional data
  $X_t \in \{s_1, s_2, \ldots\}$: *categorical* functional data

- $\mathcal{T} = [0, T_1] \times [0, T_2],\ T_1 > 0, T_2 > 0.$
  $X_{t,s} \in \mathbb{R}$: image data
- ...

Functional data and the K-L expansion

$X_t \in \mathbb{R}, t \in [0, T]$





Hypothesis : L_2–continuity.

$$X_t = \mu(t) + \sum_{i \geq 1} z_i u_i(t), \quad \forall t \in [0, T].$$

Order q approximation : $X_t^{(q)} = \mu(t) + \sum_{i=1}^{q} z_i u_i(t), \quad \forall t \in [0, T].$

Functional data and the K-L expansion

Example

X is the Brownian motion on $[0, T]$:

for all $k \geq 1$,

$$u_k(t) = \sqrt{\frac{2}{T}} \sin(\frac{k-1}{2}\pi\frac{t}{T}),$$

$$z_k \sim \mathcal{N}(0, \lambda_k), \quad \lambda_k = \frac{4T^2}{(2k-1)\pi^2}$$

Remark : A homogeneous Poisson process has the same weight functions u_k.

Functional data and the K-L expansion

Example (cont.)

The OLS criteria for functional linear regression model : an ill posed problem.

X is the Brownian motion on $[0, T]$, $Y = X_{T+h}$, $h > 0$.
Find $\beta \in L_2([0, T])$:

$$Y = \int_0^T \beta(t) X_t dt + \varepsilon$$

OLS :

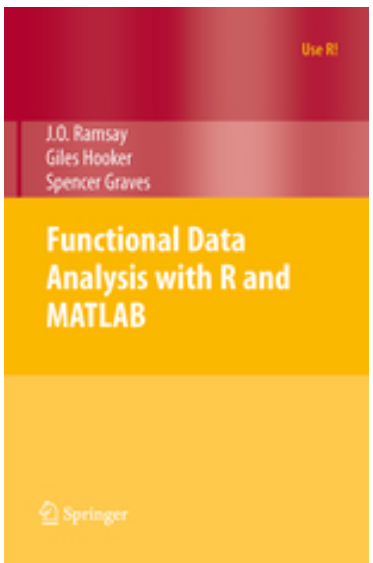
$$\beta = \underset{\alpha \in L_2([0,T])}{\arg\min}\, \mathbb{E}\left(Y - \int_0^T X_t \alpha(t) dt\right)^2$$

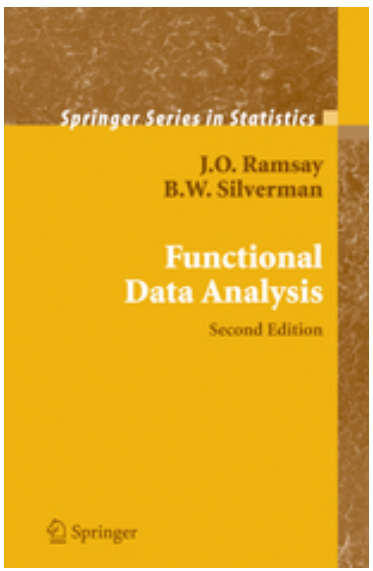
has no solution in $L_2[0, T]$.

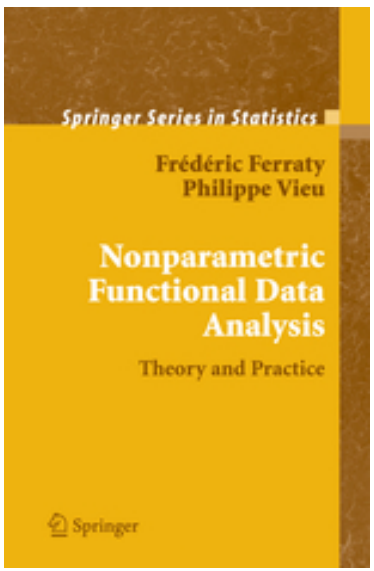
$$\beta = \delta_T, \quad \hat{Y} = \langle X, \beta \rangle = X_T.$$

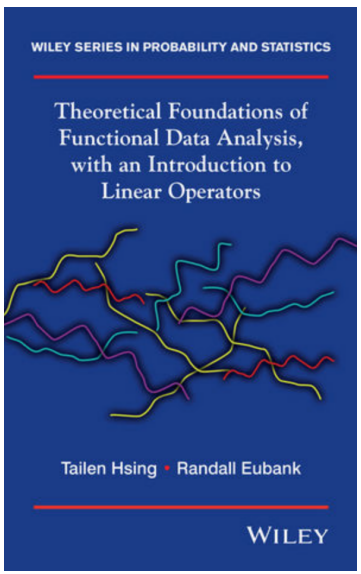
Functional data and the K-L expansion

Some references :
Deville (1974), Saporta (1981), Ramsay and Silverman (1997, 2002, 2005, 2012), Ferraty and Vieu (2006), Hsing and Eubank (2015) :









Functional data and images : double K-L expansion

$(t, s) \in [0, T_1] \times [0, T_2]$,
$X_{t,s} \in \mathbb{R}$:

Thoracic scanner

Chen and Müller (JASA, 2012) : *Modelling repeated functional observations*
Idea : double K-L expansion

Functional data and images : double K-L expansion :

$(t, s) \in [0, T_1] \times [0, T_2]$,
$X_{t,s} \in \mathbb{R}$:

- For each $s \in [0, T_2]$:

$$X(t|s) = \mu(t|s) + \sum_{k \geq 1} z_k(s) u_k(t|s), \forall t \in [0, T_1]$$

$$z_k(s) = \sum_{i \geq 1} Z_{k,i} v_{k,i}(s).$$

Reconstruction formula :

$$X(t|s) = \mu(t|s) + \sum_{k \geq 1} \sum_{i \geq 1} Z_{k,i} v_{k,i}(s) u_k(t|s).$$

Principal components : $Z_{k,i} \Longrightarrow$ visualisation, clustering, regression.

Multivariate functional data

$\vec{X}_t \in \mathbb{R}^p, \ \ \vec{X}_t = (X_{1,t}, \ldots, X_{p,t}) \ , \ t \in [0, T]$

K-L expansion and the $[C(t,t)]^{-1}$ metric (canonical analysis of $\vec{X}$) :

$$\lambda \vec{u}(t) = [C(t,t)]^{-1} \int_0^T [C(t,s)] \vec{u}(s) ds,$$

$$z_k = \int_0^T \langle \vec{X}_t - \vec{\mu}(t), \vec{u}_k(t) \rangle_{\mathbb{R}^p} dt$$

$$\vec{X}_t = \vec{\mu}(t) + \sum_{k \geq 1} z_k [C(t,t)] \vec{u}_k(t).$$

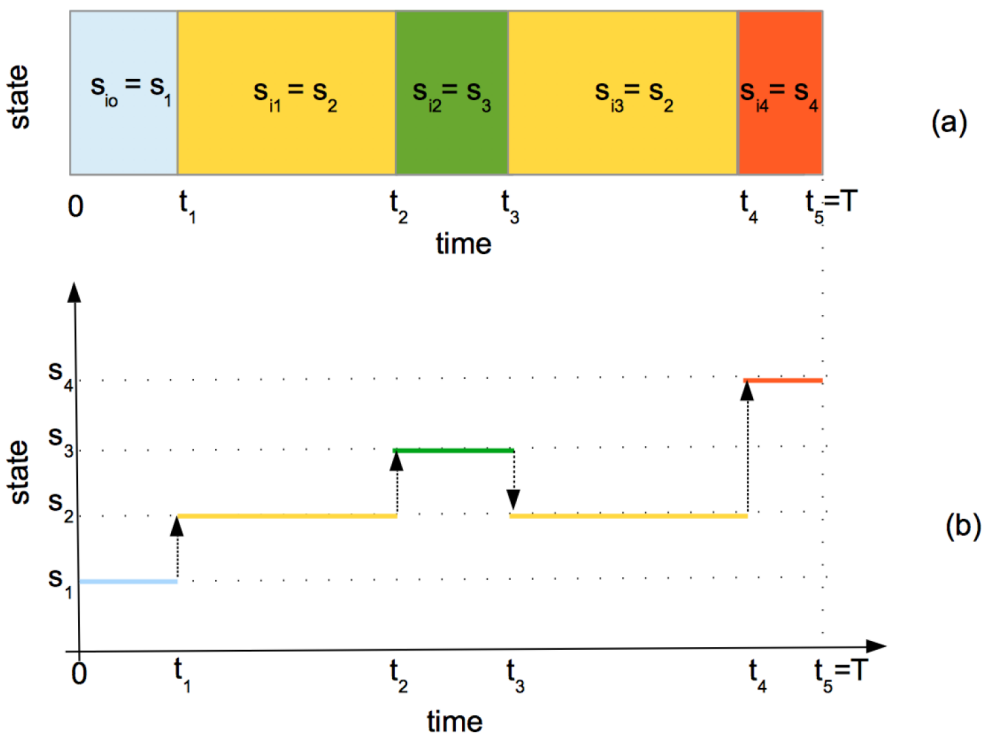
Saporta (1981).

Categorical functional data

Set of states : $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$, $K \geq 2$

$$X_t : \Omega \to \mathcal{S}.$$

A path of X on $[0, T]$ is a sequence of states $\mathbf{s_{i_j}}$ and times points $\mathbf{t_j}$:
$\{(0 = t_0, s_{i_0}), (t_1, s_{i_1}), (t_2, s_{i_2}), \ldots, (t_p = T, s_{i_p})\}$,



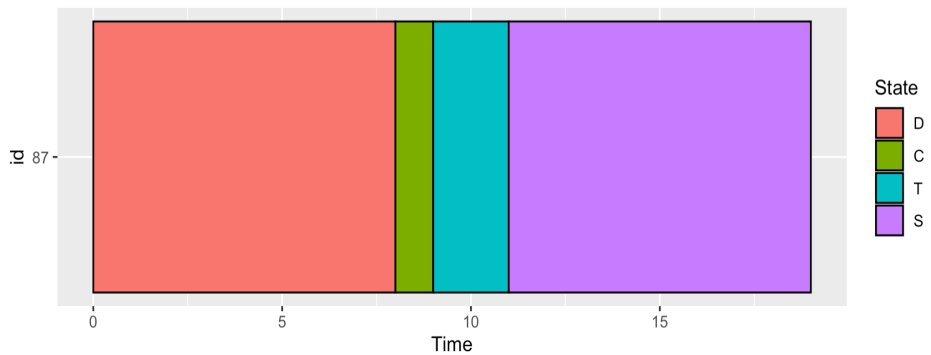
Categorical functional data

An example : Paths of infected patients.

$$\mathcal{S} = \{D, C, T, S\},$$

A path on $[0, T = 19]$:

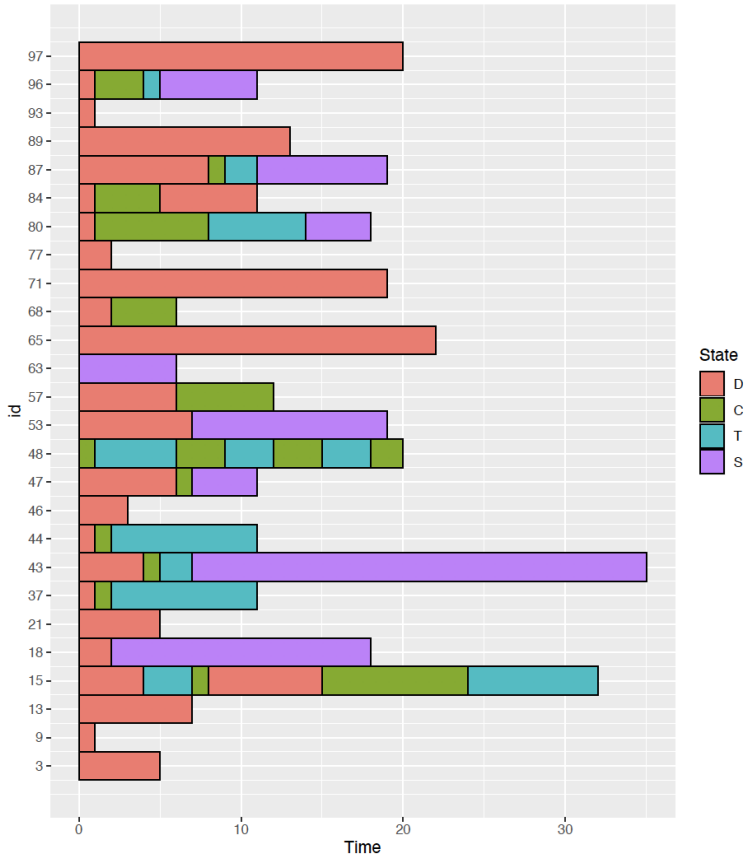
time	state
0	D
8	C
9	T
11	S
19	S



Categorical functional data

An example : Paths of infected patients.

Several paths :



Develop tools for visualisation, clustering and regression (`cfda R package`, 2021).

Categorical functional data analysis

The principal components

Hypothesis : X is continuous in probability.

Deville (1978), Saporta (1981) : find $z \in L_2(\Omega)$ that maximizes

$$\int_0^T \eta^2(z; X_t)dt,$$

with $\eta^2(z; X_t) = \frac{var(\mathbb{E}_t(z))}{var(z)}$ and $\mathbb{E}_t(z) = \mathbb{E}(z|X_t)$.

Solution :

$$\int_0^T \mathbb{E}_t(z)dt = \lambda z.$$

$\{(\lambda_i, z_i)\}_{i\geq 1}$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and

$$\int_0^T \mathbb{E}_t(z_i)dt = \lambda_i z_i.$$

Categorical functional data analysis

Optimal encodings of states and principal components

Optimal encoding functions : a^x, $x \in \mathcal{S}$:

$$z = \int_0^T \sum_{x \in \mathcal{S}} a^x(t) \mathbf{1}_t^x dt,$$

with

$$\int_0^T \sum_{y \in \mathcal{S}} p^{x,y}(t,s) a^y(s) ds = \lambda a^x(t) p^x(t), \quad \forall t \in [0,T], \forall x \in \mathcal{S},$$

where $p^x(t) = \mathbb{P}(X_t = x)$ and $p^{x,y}(t,s) = \mathbb{P}(X_t = x, X_s = y)$.

Categorical functional data analysis

Two expansion formulas :

$$\mathbf{1}_t^x = p^x(t) + \sum_{i \geq 1} z_i a_i^x(t) p^x(t), \quad \forall x \in \mathcal{S}.$$

and

$$p^{x,y}(t,s) = p^x(t)p^y(s) \sum_{i \geq 1} \lambda_i a_i^x(t) a_i^y(s),$$

$\forall t, s \in [0, T], \forall x, y \in \mathcal{S}.$

Categorical functional data : the Markov model example

$\{X_t\}_{t\in[0,T]}, \quad X_t : \Omega \to \mathcal{S} = \{1, \ldots, K\},$

- $\{\alpha_x \in \mathbb{R}_+\}_{x\in\mathcal{S}}$: parameters of sojourn time in each state
- $\mathbf{A}$ the infinitesimal generator of X.

Proposition. *If $(X_t)_{t\in[0,T]}$ is stationary with reversible distribution then the coding functions of its states are solutions of the system :*

$$\lambda \vec{a}''(t) = \mathbf{A}(\lambda\mathbf{A} + 2\mathbf{I}_m)\vec{a}(t), \quad \forall t \in [0, T].$$

If $\mathbf{A} = \mathbf{B}\Delta\mathbf{B}^{-1}$, $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_m)$, then

$$a_i^x(t) = C_{1,i}\cos(\omega_i t) + C_{2,i}\sin(\omega_i t), \quad i \geq 1, x \in \mathcal{S},$$

with $C_{1,i}, C_{2,i} \in \mathbb{R}$ and

$$\omega_i \quad : \quad \frac{\delta_i^2 - \omega_i^2}{2\delta_i\omega_i} = \mathrm{cotg}(\omega_i T).$$

Categorical functional data : the Markov model example

Two state Markov process

$$\alpha_1 = \alpha,\ \alpha_2 = \beta,\ \mathbf{A} = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}.$$

$$\pi = (\tfrac{\beta}{\alpha+\beta}, \tfrac{\alpha}{\alpha+\beta}$$

$$\lambda = \frac{\sin(\omega T)}{\omega}, \quad \frac{\omega^2 - (\alpha+\beta)^2}{2\omega(\alpha+\beta)} = \operatorname{cotg}(\omega T),$$

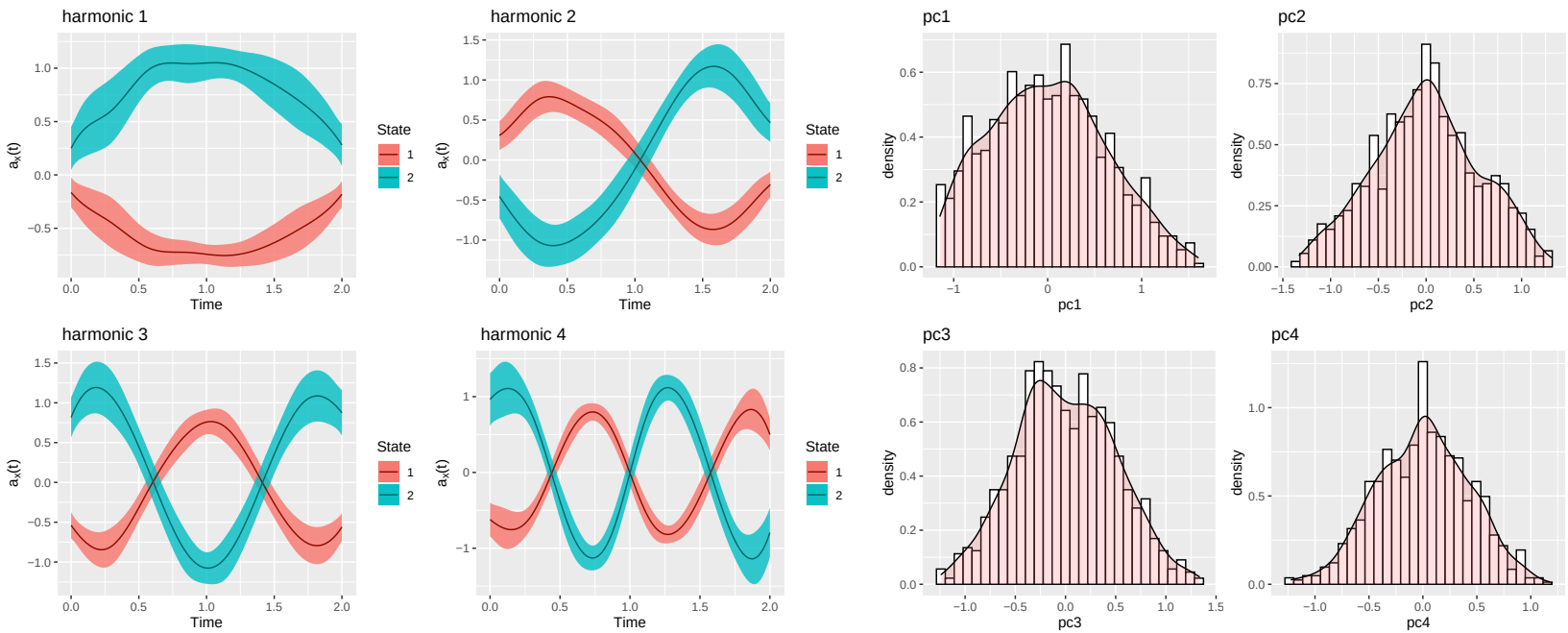
$$\begin{cases} a^1(t) = \dfrac{C\alpha}{\cos(\varphi)} \cos(\omega t + \varphi), \\ \\ a^2(t) = -\dfrac{C\beta}{\cos(\varphi)} \cos(\omega t + \varphi), \end{cases}$$

$$\varphi \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[: tg(\varphi) = -\frac{\alpha+\beta}{\omega},\ C \in \mathbb{R}.$$

Categorical functional data : the Markov model example

Two state Markov process : simulation

n= 1000, $K = 2$ $\alpha_1 = 2$, $\alpha_2 = 3$, $\pi = (0.6, 0.4)$:

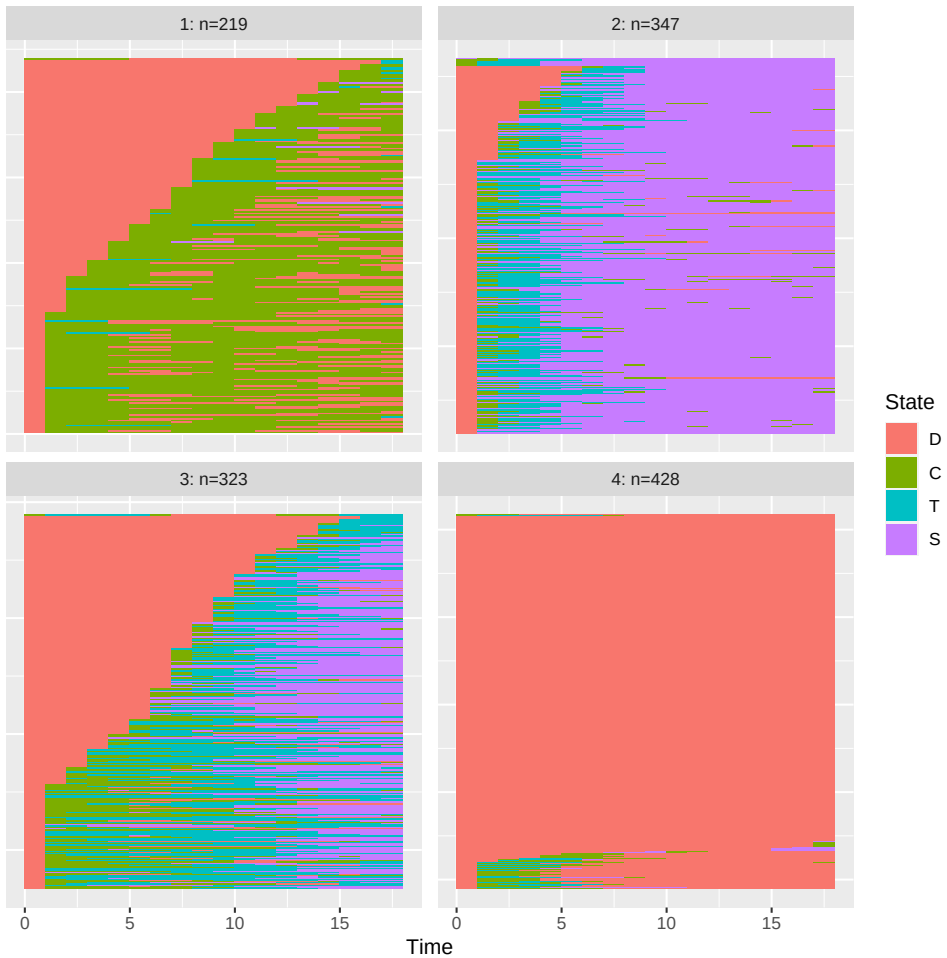


encodings

principal components

Clustering categorical functional data

- Finite mixtures of semi-Markov processes (Cardot et al. (2019)
- Hierarchical clustering on principal components (`cfda` *R package, 2021*).



Regression with categorical functional data

The linear model

$$\mathbf{1}_t^x = p^x(t) + \sum_{i \geq 1} z_i a_i^x(t) p^x(t), \quad \forall x \in \mathcal{S}.$$

Models for $f(X) = \mathbb{E}(Y|X)$:

- Regression on principal components :

$$\mathbb{E}(Y|X) \approx \mathbb{E}(Y|z_1, \ldots, z_k).$$

The linear model :

$$\mathbb{E}(Y|z_1, \ldots, z_k) = b_0 + b_1 z_1 + \ldots + b_k z_k$$

$$= b_0 + \int_0^T \sum_{x=1}^K \beta_x(t) \mathbf{1}_t^x dt.$$

where $\beta_x(t) = \sum_{j=1}^k b_j a_j^x(t)$.

Regression with categorical functional data : kernel methods

Let H be the space whose elements are the paths of X.

Let $x, y \in H$, $x = \{x_t, t \in [0, T]\}$, $y = \{y_t, t \in [0, T]\}$.

$$x(t) = \left(\mathbf{1}_x^1(t), \ldots, \mathbf{1}_x^K(t)\right)$$

$$y(t) = \left(\mathbf{1}_y^1(t), \ldots, \mathbf{1}_y^K(t)\right)$$

The inner product :

$$\langle x, y \rangle_H = \int_0^T \sum_{i=1}^K \mathbf{1}_x^i(t)\mathbf{1}_y^i(t)dt.$$

$\langle x, y \rangle_H$ = the length of time within $[0, T]$ where x and y are in the same state.

Notice that $\|x\|^2 = T$.

Regression with categorical functional data : kernel methods

$\mathrm{f}(X) = \mathbb{E}(Y|X)$:

- Non-parametric regression and RKHS methods.

$f \in \mathcal{F} = \mathcal{H}_{\mathbf{K}}$, $\mathcal{H}_{\mathbf{K}}$ is an RKHS with kernel $\mathbf{K} : H \times H \to \mathbb{R}$.

Representer Theorem (Kimeldorf and Wahba (1970), Scölkhopf et al (2001)). *Let* $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in H$, $y_i \in \mathbb{R}$, $n > 0$, $n \in \mathbb{N}$, $\lambda > 0$ *and* $\mathcal{C} : H \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ *be a convex loss function of third argument. Then, the solution to the problem : find* $\hat{f} \in \mathcal{H}_{\mathbf{K}}$ *to minimize*

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{C}\left(x_i, y_i, \hat{f}(x_i)\right) + \lambda \|\hat{f}\|^2_{\mathcal{H}_{\mathbf{K}}}$$

exists, is unique and admits a representation of the form

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i \mathbf{K}(x, x_i), \quad \forall x \in H,$$

where $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, n$.

Regression with categorical functional data : Kernel methods

Example of kernels

Gaussian :

$$\mathbf{K}(x,y) = e^{-\frac{\|x-y\|_H^2}{2\sigma^2}}, \quad \sigma > 0,$$

Polynomial :

$$\mathbf{K}(x,y) = (c + \langle x, y \rangle_H)^d, \quad c > 0, d > 0.$$

and many others !

Categorical functional data : some references

1. Saporta G. (1981) *Méthodes exploratoires d'analyse de données temporelles*, Cahiers du B.U.R.O, Université Pierre et Marie Curie,37-38, Paris.

2. Deville J.C. (1982) *Analyse de données chronologiques qualitatives : comment analyser des calendriers ?*, Annales de l'INSEE, No 45, p. 45-104.

3. Cardot H., Lecuelle G., Schlich P. ,Visalli M. (2019), Estimating Finite Mixtures of Semi-Markov Chains : An Application to the Segmentation of Temporal Sensory Data, JRSS Series C : Applied Statistics, Volume 68, Issue 5, p.1281–1303.

4. Preda C., Grimonprez Q., Vandewalle V., *Categorical Functional Data Analysis. The cfda R Package*. Mathematics. 2021 ; 9(23) :3074. https ://doi.org/10.3390/math9233074