Clustering categorical functional data
Application to medical discharge letters

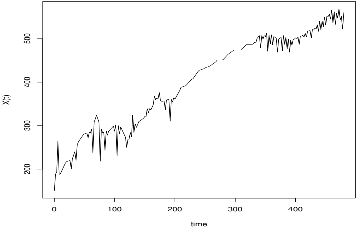
Vincent Vandewalle, Cristina Cozma, Cristian Preda
MODAL Team /Inria Lille Nord Europe

December 14, 2015
ERCIM 2015, London

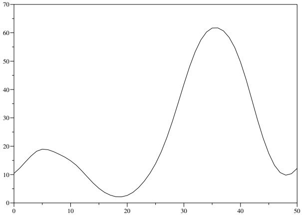
Functional data

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*.

In the literature, most of the fd is scalar (univariate/ multivariate) :



Dough resistance



Gait data

Model : Stochastic process, $X = \{X_t, \quad t \in \mathcal{T}\}$,

$$X_t \in \mathbb{R}^p, p \geq 1,$$

for some index set $\mathcal{T}$

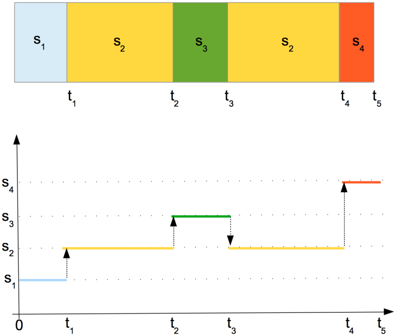
Categorical functional data

Set of states : $\mathbf{S} = \{s_1, s_2, \ldots, s_m\}$, $m \geq 2$

$\mathbf{X} = \{X_t \; : \; t \in [0, T]\}$, $(\Omega, \mathcal{A}, P)$, $\quad X_t : \Omega \to \mathbf{S}$.

A path of $\mathbf{X}$ on $[0, T]$ is a sequence of states $\mathbf{s_i}$ and times points $\mathbf{t_i}$ of transitions from one state to another :

$$\{(s_1, t_1), (s_2, t_2), \ldots, (s_f, t_f)\}$$



Categorical functional data

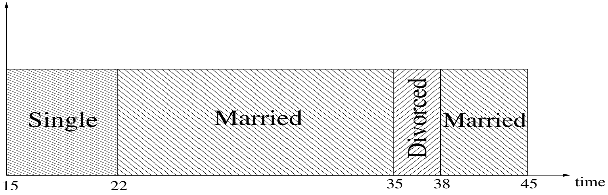
Saporta et Deville (1982,1983) :

- *Analyse de données chronologiques qualitatives : comment analyser des calendriers ?*, Annales de l'INSEE, No. 45, p. 45-104.

- *Correspondence analysis with an extension towards nominal time series*, Journal of Econometrics, 22, p. 169-189.

X = marital status of women from 15 to 45

$$X_t \in \{\text{"Single"}, \text{"Maried"}, \text{"Divorced"}, \text{"Widowed"}\}$$



Categorical functional data

Model :

$\mathbf{S} = \{s_1, s_2, \dots s_m\}$, $(\Omega, \mathcal{A}, P)$,

$\mathbf{X} = \{X_t \; : \; t > 0\}$, $\quad X_t : \Omega \to \mathbf{S}$, a continuous-time jump process

Markov jump process and categorical functional data :

Harmonic qualitative analysis (extension of the multiple correspondance analysis towards categorical fd)

- Richard (1988), Heijden et al (1997), Preda (1998).

The absorbing states

- X can have one or several *absorbing states* :

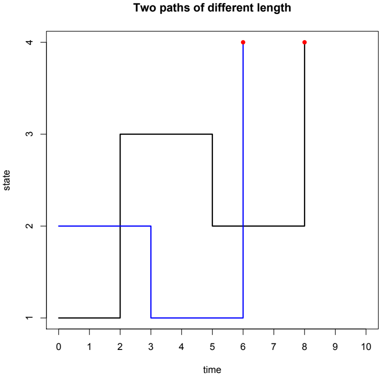
$$\mathbb{P}(X_{t+s} = s \big| X_t = s) = 1, \forall s > 0$$

Two situations :

X is observed :

- until an absorbant state is reached

  Remark :
  functional data with different lengths !



- on a predefined period $[0, T]$
  (classical framework of functional data)

The aim

The data : n sample paths of $\mathbf{X}$: $\{\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_n\}$.
The *ith* path :

$$x_i = (\mathrm{s}_{i0}, t_{i0}, \mathrm{s}_{i1}, t_{i1} \ldots, t_{i(d_i-1)}, \mathrm{s}_{id_i})$$

- d_i = the number of jumps of the path i,
- s_{ij} = the length of time spent in the j visited state of path i

It is supposed that the paths are uncensured, i.e. the paths are observed until they have reached the absorbing state.

Objective : Clustering $\{\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_n\}$.

Clustering categorical functional data

Notations

- m the number of states
- $\mathcal{S} = \{1, \ldots, m\}$ the state space, (m is the absorbing state)
- n is the number of observed paths
- d_i is the length of the i^{th} observed path
- s_{ijh} equals to 1 if the j^{th} state of the path i is h and 0 otherwise.
- $\mathbf{s}_{ij} = (s_{ijh}, \ldots, s_{ijm})$ the binary coding of the j^{th} state from the path i
- t_{ij} the length of time in the j^{th} of the path i,
- $\boldsymbol{x}_i = (\mathbf{s}_{i0}, t_{i0}, \mathbf{s}_{i1}, t_{i1} \ldots, t_{i(d_i-1)}, \mathbf{s}_{id_i})$ is the data from the path i,
- $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ the whole dataset

Mixture of Markov processes

- The n paths come from K different processes (K to be determined) caracterized by parameters $\boldsymbol{\theta}_k$.
- The likelihood function for the path i coming from cluster k,

$$p(\boldsymbol{x}_i; \boldsymbol{\theta}_k) = p(\mathbf{s}_{i0}; \boldsymbol{\theta}_k) p(t_{i0}, \mathbf{s}_{i1} | \mathbf{s}_{i0}; \boldsymbol{\theta}_k) \prod_{j=1}^{d_i - 1} p(t_{ij}, \mathbf{s}_{i(j+1)} | \mathbf{s}_{ij}, t_{i(j-1)}, \ldots, \mathbf{s}_{i1}, t_{i0}, \mathbf{s}_{i0}; \boldsymbol{\theta}_k)$$

- Estimate from the data the parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ and then determine the class membership for each sample paths.

Markovian hypotheses

H1 The distribution of $(t_{ij}, \mathbf{s}_{i(j+1)})$ is independent of the past given $\mathbf{s}_{ij}$

$$p(t_{ij}, \mathbf{s}_{i(j+1)} | \mathbf{s}_{ij}, t_{i(j-1)}, \dots, \mathbf{s}_{i1}, t_{i0}, \mathbf{s}_{i0}; \boldsymbol{\theta}_k) = p(t_{ij}, \mathbf{s}_{i(j+1)} | \mathbf{s}_{ij}; \boldsymbol{\theta}_k)$$

H2 The distributions of t_{ij} and $\mathbf{s}_{i(j+1)}$ are independent given $\mathbf{s}_{ij}$

$$p(t_{ij}, \mathbf{s}_{i(j+1)} | \mathbf{s}_{ij}; \boldsymbol{\theta}_k) = p(t_{ij} | \mathbf{s}_{ij}; \boldsymbol{\theta}_k) p(\mathbf{s}_{i(j+1)} | \mathbf{s}_{ij}; \boldsymbol{\theta}_k)$$

H3 The distribution of t_{ij} given $\mathbf{e}_{ij}$ is an exponential distribution

H4 The distribution of the initial state does not depends on the cluster

$$p(\mathbf{s}_{i0}; \boldsymbol{\theta}_k) = p(\mathbf{s}_{i0})$$

Then,

$$p(\boldsymbol{x}_i; \boldsymbol{\theta}_k) = p(\mathbf{s}_{i0}) \prod_{j=0}^{d_i-1} \underbrace{p(t_{ij} | \mathbf{s}_{ij}; \boldsymbol{\theta}_k)}_{\text{time}} \underbrace{p(\mathbf{s}_{i(j+1)} | \mathbf{s}_{ij}; \boldsymbol{\theta}_k)}_{\text{transition}}$$

Model parameters

Cluster k : $\boldsymbol{\theta}_k$

- transition probability matrix $\boldsymbol{\alpha}_k$
  - $\alpha_{khh'}$: the probability to move from state h to state h',
    $\boldsymbol{\alpha}_k = (\alpha_{khh'})_{1 \leq h \leq m-1, 1 \leq h' \leq m}$
- time distribution $\boldsymbol{\lambda}_k$
  - λ_{kh} : parameter of the time distribution in state h of cluster k
    $\boldsymbol{\lambda}_k = (\lambda_{k1}, \ldots, \lambda_{k(m-1)})$

$$\boldsymbol{\theta}_k = (\boldsymbol{\alpha}_k, \boldsymbol{\lambda}_k)$$

.

All data :

- π_k prior weight for cluster k
  $\boldsymbol{\pi} = (\pi_k, \ldots, \pi_K)$

$$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$$

The likelihood

The pdf for the path i given the parameter $\boldsymbol{\theta}$ is

$$p(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_i;\boldsymbol{\theta}_k).$$

The log-likelihood $\ell(\boldsymbol{\theta};\mathbf{x})$,

$$\ell(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right).$$

Estimation : $\hat{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta};\mathbf{x}).$$

The EM algorithm

The completed log-likelihood :

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log\left(\pi_k p(\boldsymbol{x}_i; \boldsymbol{\theta}_k)\right)$$

where z_{ik} equals 1 if path i comes from the k^{th} Markov process.

- **E step :**

$$t_{ik}^{(r+1)} = E[Z_{ik} | \boldsymbol{x}_i; \boldsymbol{\theta}^{(r)}] = P(Z_{ik} = 1 | \boldsymbol{x}_i; \boldsymbol{\theta}^{(r)}) = \frac{\pi_k^{(r)} p(\boldsymbol{x}_i; \boldsymbol{\theta}_k^{(r)})}{\sum_{k'=1}^{K} \pi_{k'}^{(r)} p(\boldsymbol{x}_i; \boldsymbol{\theta}_{k'}^{(r)})}$$

- **M step :** Let

$$n_{khh'}^{(r+1)} = \sum_{i=1}^{n} \sum_{j=0}^{d_i - 1} t_{ik}^{(r+1)} s_{ijh} s_{i(j+1)h'}, \qquad n_{kh}^{(r+1)} = \sum_{i=1}^{n} \sum_{j=0}^{d_i - 1} t_{ik}^{(r+1)} s_{ijh}, \qquad n_k^{(r+1)} = \sum_{i=1}^{n} t_{ik}^{(r+1)}$$

The update formulas are

$$\pi_k^{(r+1)} = \frac{n_k^{(r+1)}}{n}, \qquad \alpha_{khh'}^{(r+1)} = \frac{n_{khh'}^{(r+1)}}{n_{kh}^{(r+1)}}, \qquad \lambda_{kh}^{(r+1)} = \frac{n_k^{(r+1)}}{\sum_{i=1}^{n} \sum_{j=1}^{d_i} t_{ik}^{(r+1)} s_{ijh} t_{ij}}.$$

Clustering

$$\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\pi}}_k, \hat{\boldsymbol{\lambda}}_k, \hat{\boldsymbol{\alpha}}_k\}_{k=1,\dots,K}.$$

$$\hat{z}_{ik} = \begin{cases} 1 \text{ if } k = \underset{k'}{\operatorname{argmax}} P(Z_{ik'} = 1 | \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}), \\ 0 \quad \text{otherwise.} \end{cases}$$

$$P(Z_{ik} = 1 | \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{\pi_k p(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_k)}{\sum_{k'=1}^{K} \pi_{k'} p(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_{k'})}$$

Choice of the number of clusters K

The BIC criterion for K clusters noted $BIC(K)$ is

$$BIC(K) = L(\hat{\boldsymbol{\theta}}^K; \mathbf{x}) - \frac{\nu_K}{2} \log n$$

where $\hat{\boldsymbol{\theta}}^K$ is the maximum likelihood parameter estimated when considering K clusters, and ν_K are the number of estimated parameters when considering K clusters.

$$\nu_K = K(m-1)^2 + K - 1$$

The chosen number of clusters :

$$\hat{K} = \underset{K \in \{1,\dots,K_{\max}\}}{\operatorname{argmax}} BIC(K)$$

Application to medical discharge letters status.

Definition of the states :

1. the doctor is dictating the letter.
2. the letter is "waiting" to be type-writing by an assistant (queue)
3. the letter is type-writing by the assistant
4. the letter is "waiting" for doctor validation (queue)
5. the letter is in validation process by the doctor
6. the letter is "waiting" to be affected to an assistant (queue)
7. the letter is treated by the assistant
8. the letter is sent to the patient (end).

Data description :

A state is caracterised by 4 values :

1. date of the beginnig of the state
2. the day number (into a week) of the beginning date (1=Monday, 7=Sunday)
3. length of time spent into the state
4. the name of the state (1 to 8)

beginning date	day	length	state
10/01/2012 02 :39 :19	2	0h0m0s	1
10/01/2012 02 :42 :38	2	0h7m20s	2
10/01/2012 02 :49 :58	2	18h29m34s	3
11/01/2012 09 :19 :42	3	4h43m59s	4
11/01/2012 02 :14 :08	3	3h13m13s	6
11/01/2012 05 :27 :21	3	0h0m7s	7
11/01/2012 05 :30 :44	3		8

Summary of data :

- 443 325 letters

Number of jumps (length of the path) :

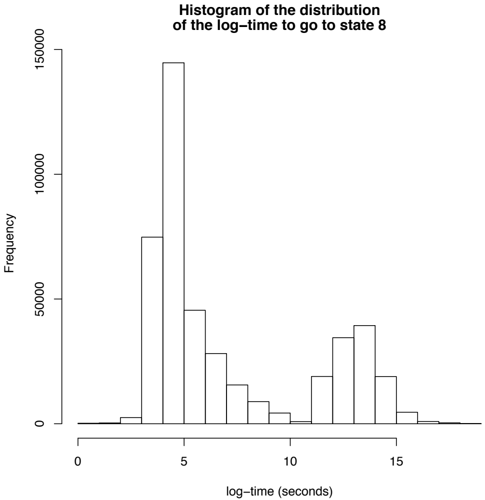
Length	2	3	4	5	6	7	8
Frequence	336181	1118	2752	8157	23688	8541	62888

Number of transitions from one state to another state :

from \ to	1	2	3	4	5	6	7	8
1	0	93042	201	0	0	0	0	335306
2	0	0	90453	2849	32	0	0	317
3	0	0	0	100452	113	974	1	73
4	0	0	0	0	92351	6629	191	6694
5	0	0	0	0	0	76523	887	15353
6	0	0	0	0	0	0	81180	3184
7	0	0	0	0	0	0	0	82398

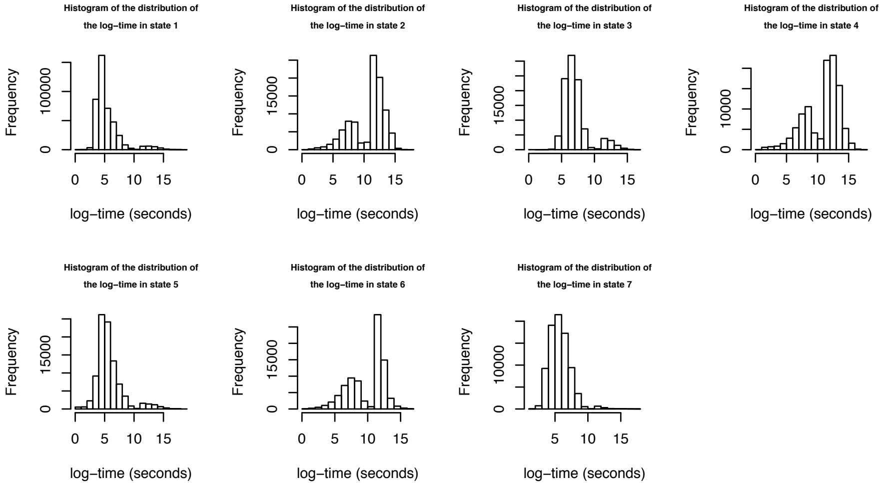
Time length of the paths :

Quantile	Time (s)
0%	1.00
10%	47.00
20%	57.00
30%	69.00
40%	90.00
50%	147.00
60%	383.00
70%	2643.00
80%	231211.40
90%	637633.20
100%	89717350.00



Distribution of log(length), cut-off at exp(10) seconds $\simeq$ 6 hours.

Distribution of the length of time for each state



All distributions are bimodal.

Classification

$K = 4$ clusters :

Cluster	1	2	3	4
Size	37783	23159	358498	23844

Average time by state :