Clustering categorical functional data Application to medical discharge letters Medical discharge letters

Vincent Vandewalle, Cristian Preda

Clustering categorical functional data
Application to medical discharge letters

Vincent Vandewalle[a,c] & Cristian Preda [b,c]

[a]**Équipe d'Accueil 2694** (Université Lille 2)

[b]**Laboratoire Paul Painlevé** (Université Lille 1 - CNRS)

[c]**Équipe MODAL** (Inria Lille Nord Europe)

Work conducted under the ClinMine ANR project

Medical discharge letters

The data are provided from the GHICL in the scope of the ClinMine ANR project. The goal is to cluster patient stays according to the evolution of the status of their discharge letters over the time. Clustering such type of data could help to improve the patient management by detecting some atypical clusters.

Definition of the states

The discharge could pass by 8 different states:

1. the doctor is dictating the letter.
2. the letter is "waiting" to be type-writing by an assistant
3. the letter is type-writing by the assistant
4. the letter is "waiting" for doctor validation
5. the letter is in validation process by the doctor
6. the letter is "waiting" to be affected to an assistant
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 the data

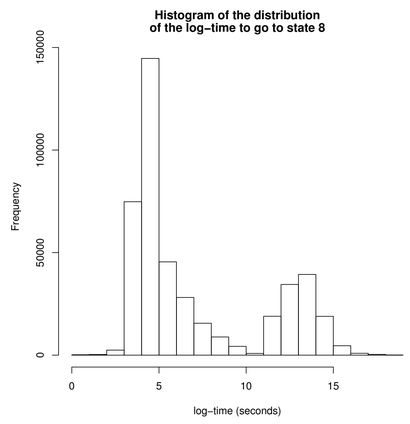
- 443 325 letters

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



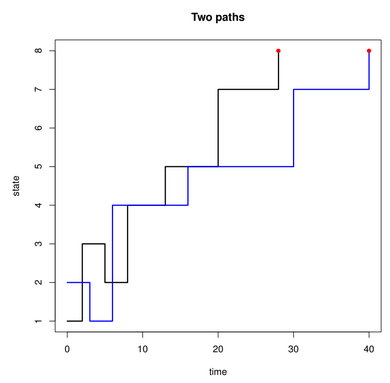
Categorical functional data

The data

- n sample paths : $\mathbf{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_n\}$.
- The ith path:

$$\boldsymbol{x}_i = (\mathbf{s}_{i0}, t_{i0}, \mathbf{s}_{i1}, t_{i1} \dots, t_{i(d_i-1)}, \mathbf{s}_{id_i})$$

  - m the number of states
  - d_i = the number of jumps of the path i,
  - t_{ij} = the length of time spent in the j visited state of path i
  - $\mathbf{s}_{ij} = (s_{ijh}, \dots, s_{ijm})$ the binary coding of the j^{th} state from the 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 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_n\}$.

Mixture of Markov processes

Likelihood: general form

- The n paths come from K different processes characterized by parameters $\boldsymbol{\theta}_k$ ($k \in \{1, \dots, K\}$).
- The likelihood function for the path i coming from cluster k is:

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

Markovian assumptions

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 s_{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

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

Parameters of cluster k: $\boldsymbol{\theta}_k = (\boldsymbol{\alpha}_k, \boldsymbol{\lambda}_k)$

Transition probability matrix $\boldsymbol{\alpha}_k$

- $\alpha_{khh'}$: the probability to move from state h to state h',
- $\boldsymbol{\alpha}_k = (\alpha_{khh'})_{1 \le h \le m-1, 1 \le h' \le m}$

Time distribution $\boldsymbol{\lambda}_k$

- λ_{kh} : parameter of the time distribution in state h,
- $\boldsymbol{\lambda}_k = (\lambda_{k1}, \dots, \lambda_{k(m-1)})$

Parameters estimation and model choice

Parameters

The parameters to be estimated are the cluster specific parameters $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$ and the prior weights $\pi_1, \dots, \pi_K$.

$$\boldsymbol{\theta} = (\pi_1, \boldsymbol{\theta}_1, \pi_2, \boldsymbol{\theta}_2, \dots, \pi_K, \boldsymbol{\theta}_K)$$

Likelihood

The likelihood for the path i is modeled as a mixture

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

The log-likelihood of the data $\mathbf{x}$ for the parameter $\boldsymbol{\theta}$ is

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

EM algorithm

The parameters are estimated by maximum likelihood using the EM algorithm.
The completed log-likelihood is:

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

where z_{ik} equals 1 if the 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

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

Model choice

Since in practice the number of cluster is unknown the choice of the number of cluster can be simply performed by the BIC criterion

$$\text{BIC(K)} = \ell(\hat{\boldsymbol{\theta}}; \mathbf{x}, K) - \frac{\nu_K}{2} \log(n).$$

where ν_K is the number of free parameter of the model with K clusters.

Clustering of discharge letters

Results of the clustering in 2 clusters

- Estimated prior weights: $\hat{\pi}_1 = 0.897$, $\hat{\pi}_2 = 0.103$
- Mean sojourn time (in seconds): $1/\hat{\lambda}_{kj}$

state	1	2	3	4	5	6	7
cluster 1	288.66	290460.96	1136.54	373567.50	569.75	131702.82	712.76
cluster 2	863390.53	268556.31	215645.10	408716.40	380294.60	217268.28	48815.76

- Class conditional transition probabilities: $\alpha_{khh'}$

Cluster 1

from \ to	2	3	4	5	6	7	8
1	0.20						0.80
2		0.96	0.03				
3			0.99		0.01		
4				0.87	0.06		0.07
5					0.83	0.01	0.16
6						0.96	0.04
7							1.00

Cluster 2

from \ to	2	3	4	5	6	7	8
1	0.38						0.62
2		0.98	0.02				
3			0.99				
4				0.88	0.08		0.04
5					0.80	0.03	0.17
6						0.97	0.03
7							1.00

- The main difference between the clusters is the sojourn time distribution. The second cluster is characterized by long sojourn times.
- The transition probabilities are roughly the same except for transition from state 1 to state 8.

Discussion

- Due to the very large number of data ($n = 443\,325$), the model choice issue is difficult to solve since standard model choice criteria leading to large number of clusters making the interpretation difficult.
- The large number of direct transitions from state 1 to state 8 in the data make the interpretation difficult from the practical point of view, further investigations on the data are in progress.

Conclusion and perspectives

Conclusion

- Transition probabilities and sojourn times naturally taken into account
- Model parameters easy to interpret

Perspectives

- A package R will be available soon including parsimonious models on transition probabilities and time distribution parameters
- Adapt the model when the number of states is large (estimation of the transition matrix difficult)
- Apply the model on other clinical data such as data on diabetes

Acknowledgements

We would like to thank Doctor Arnaud Hansske from the GHICL, Julie Jacques, Julien Taillard and David Delerue from the *alicante* society for providing us the data.