Clustering categorical functional data

Cristian Preda [1] & Vincent Vandewalle [2]

[1] *MODAL - Inria Lille Nord Europe, Laboratoire Paul Painlevé, Université Lille 1, Lille, France, et cristian.preda@polytech-lille.fr*
[2] *MODAL - Inria Lille Nord Europe, EA 2694, Université Lille 2, Lille, France, et vincent.vandewalle@univ-lille2.fr*

Résumé. La classification non-supervisée des données fonctionnelle qualitatives représentées par des trajectoires d'un processus de sauts est considérée. Nous proposons un algorithme EM pour estimer un mélange de processus de Markov. Une étude de simulation et une application sur des données hospitalières sont présentées.

Mots-clés. Données fonctionelles qualitatives, classification, algorithme EM.

Abstract. Categorical functional data represented by paths of a stochastic jump process with continuous time are considered for clustering. For Markov models we propose an EM algorithm to estimate a mixture of Markov processes. A simulation study as well as a real application on hospital stays will be presented.

Keywords. Categorical functional data, clustering, EM algorithm.

1 Introduction

Most literature devoted to functional data considers data as sample paths of a real-valued stochastic process, $X = \{X_t, t \in \mathcal{T}\}$, $X_t \in \mathbb{R}^p$, $p \geq 1$ where $\mathcal{T}$ is some continuous set. Among a considerable record of papers on the subject, the monographs of Ramsay and Silverman (2002, 2005) and Ferraty and Vieu (2006) still remain references presenting the main methodologies for visualisation, denoising, classification and regression when dealing with functional data represented by real-valued functions.

We consider the case where the underlying stochastic model generating the data is a continuous-time stochastic process $X = \{X_t, t \in \mathcal{T}\}$ such that for all $t \in \mathcal{T}$, X_t is a categorical random variable rather than a real-valued one.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{S} = \{s_1, \ldots, s_m\}$, $m \geq 2$, be a set of m states and

$$X = \{X_t \ ; \ X_t : \Omega \longrightarrow \mathcal{S}, \quad t \in \mathcal{T}\} \tag{1}$$

be a family of categorical random variables indexed by $\mathcal{T}$. Thus, a path of X is a sequence of states s_{i_j} and times points t_i of transitions from one state to another one : $\{(s_{i_1}, t_1), (s_{i_2}, t_2), \ldots\}$, with $s_{i_j} \in \mathcal{S}$ and $t_i \in \mathcal{T}$.

We call the sample paths of the process (1) *categorical functional data*. The Figure 1 presents graphically scalar and categorical functional data.

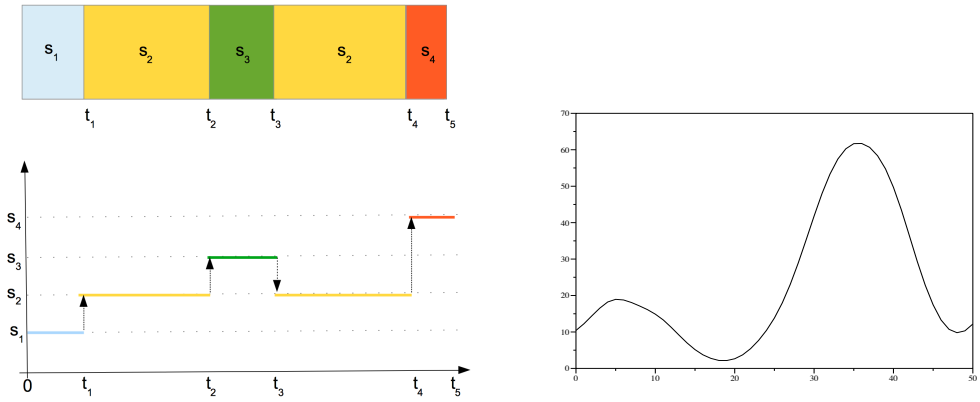

Figure 1: Examples of categorical (left) and scalar (right) functional data.

To the best of our knowledge, and quite surprisingly, there is no recent researches devoted to this type of functional data despite its ability to model real situations in different fields of applications: health and medicine (status of a patient over time), economy (status of the market), sociology (evolution of social status), and so on. As a start point in research on this topic we consider the works of Boumaza(1980), Deville (1982), Deville and Saporta (1983), Saporta (1981). These works are devoted to the extension of factorial techniques (canonic and multiple correspondances analysis) towards functional data. Applications of these techniques are presented in Heijden (1997) for analysing career data and in Preda (1998) for studying spectral properties of the transition probability matrix of a the stationary Markovian jump process with continuous time.

In this work we present a model-based methodology of clustering categorical functional data. Instead of the classical setting considering a fixed length of the paths of X, i.e. the process is observed over a fixed length of time $\mathcal{T} = [0, T]$, $T > 0$, we consider that the process X has an absorbing state and thus, we allow sample paths of different lengths. In the Markovian framework, based on the likelihood function, we derive an EM algorithme for clustering categorical functional data. A simulation study and an application on clustering medical discharge letters according to their status of dictating, type-writing and delivery to the end-user (patient or medicine) are presented.

Bibliographie

[1] Boumaza R. (1980) *Contribution à l'étude descriptive d'une fonction aléatoire qualitative*, Thèse de 3ème cycle, Université Paul Sabatier, Toulouse, France.

[2] Deville J.C. (1982) *Analyse de données chronologiques qualitatives : comment analyser des calendriers ?*, Annales de l'INSEE, No. 45, 45–104.
[3] Deville J. C., Saporta G. (1983) *Correspondence analysis with an extension towards nominal time series*, Journal of Econometrics, 22, 169–189.
[4] Ferraty F., Vieu P. (2006) *Nonparameric Functional Data Analysis. Theory and Practice*, Second Edition, Springer Series in Statistics.
[5] Heijden P.G.M. , Teunissen J., van Orlé C. (1997) *Multiple correspondence analysis as a tool for quantification or classification of career data*, Journal of Educational and Behavioral Statistics, 22, 447–477.
[6] Preda C. (1998), *Analyse harmonique qualitative des processus markoviens de sauts stationnaires*, Scientific Annals of Alexandru Ioan Cuza University of Iasi (Romania), Computer Science Section, Tome VII, 5–18.
[7] Ramsay J.O., Silverman B.W. (2002) *Applied functional data analysis. Methods and case studies*. Springer Series in Statistics, Springer-Verlag, New York.
[8] Ramsay J.O., Silverman B.W. (2005) *Functional Data Analysis*, Springer Series in Statistics, Springer-Verlag, New York.
[9] Saporta G. (1981) *Méthodes exploratoires d'analyse de données temporelles*, Cahiers du B.U.R.O., No. 37-38, Université Pierre et Marie Curie, Paris.