Chronological analysis of textual data and curve clustering: preliminary results based on wavelets

Matilde Trevisani and Arjuna Tuzzi

Abstract In textual analysis, many corpora include texts which have a chronological order. The temporal evolution of (key) words is relevant in order to highlight the distinctive features of the chronological corpus. In a typical bag-of-words approach data are organized in word-type x time-point contingency tables. Such discrete data can be thought of as continuous objects represented by functional relationships. The aims of this study are identifying a specific sequential pattern for each word as a functional object, and determining prototype patterns representing clusters of words portraying a similar evolution. We propose the application of a flexible wavelet-based model for curve clustering to a corpus of end-of-year addresses delivered by the ten Presidents of Italian Republic in the period 1949-2011.

Key words: textual analysis, word patterns, curve clustering, wavelets

1 Introduction

In textual analysis many corpora include texts which have a chronological order and this temporal connotation is crucial to understand their structure. Common examples are: addresses delivered by institutional representatives in different years, articles retrieved from newspaper archives, literary works written by an author during his/her active life, essays written by a student in different steps of his/her educational experience, etc.

In a typical bag-of-words approach data are organized in word-type x time-point contingency tables. A very large number of word-types are often characterized by

Matilde Trevisani
Department of Economics, Business, Mathematics and Statistics, University of Trieste, 34100 Trieste, e-mail: matildet@econ.units.it

Arjuna Tuzzi
Department of FISPPA, University of Padova, 35123 Padova, e-mail: arjuna.tuzzi@unipd.it

a relatively low number of time occurrences, and, moreover, are sparsely spaced over time. In the context of chronological corpora sparsity is represented by a large number of zeros for most of the time interval length; these zero-cells are due to the large number of word-types with low number of corresponding word-tokens (intrinsic feature of textual data known as large p small n problem) as well as to the size of time-point subcorpora (the richness of information is highly variable across time in terms of different number of documents and of word-tokens therein).

Such discrete data can be thought of as continuous objects represented by functional observations [6]. Object of this study is twofold: to identify a specific sequential pattern for each word as a functional object; to partition these curves (word patterns) into homogeneous clusters [3] . To this aim we illustrate a flexible wavelet-based model with specific reference to political and institutional texts.

2 Corpus

The chronological corpus includes 63 end-of-year messages (1949-2011) of ten Presidents of the Italian Republic. The messages are available on the institutional website of the presidential office but a manual correction of the official versions was necessary to obtain the texts actually delivered by the Presidents and improve corpus homogeneity. Moreover, the lemmatization of the corpus allowed a part-of-speech analysis. The number of word-tokens (the size of the corpus in terms of occurrences, i.e. observations) is 104152; the number of lemma-types (the size of the corpus in terms of different words, i.e. categories of the observations) is 6352.

The size of the addresses in terms of word-tokens (Figure 1) shows an increasing trend over time between two extremes: Einaudi for his conciseness and Scalfaro for his loquacity. In order to enhance the comparability of the shorter addresses we aggregated consecutive addresses when their length was lower than a threshold set to 600 word-tokens (Einaudi 1949-1951, 1952-1954; Gronchi 1955-1956; Saragat 1964-1965; Leone 1971-1972; Cossiga 1990-1991). In this application only the 535 names (lemma-types with frequency higher than 10 in the corpus) have been considered and the frequencies in the addresses are expressed as rates per 1000 word-tokens.

3 Curve clustering and wavelets

In this study we focus on methods for model-based curve clustering in presence of individual variability. Curve clustering has longly been studied using splines, however they are not appropriate when dealing with high-dimensional data and cannot be used to model irregular curves such as peak-like data. On the contrary, wavelet representation can accomodate a wider range of functional shapes and proves more flexible than splines. In our case the temporal pattern of the frequencies of names shows peaks and irregularities (spot curve). For instance, the absence of a name in a

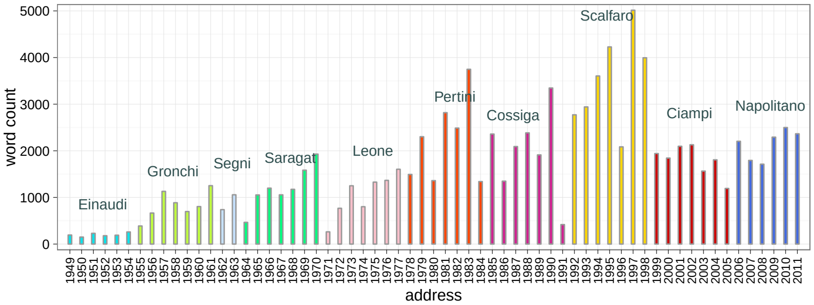

Fig. 1 Length of end-of-year addresses: number of word-tokens.

message (time-point) originates a zero-peak (sparsity) as well as its high frequency in another message produces a high-peak. Moreover, data are high-dimensional which is a common feature of chronological corpora. Hence, a method based on a wavelet decomposition seems to be more suitable [5].

In our work we obtain preliminary results by suitably fitting a functional clustering mixed model (FCMM) in the wavelet domain. The model resumes to a linear mixed-effects model that can be used for a model-based clustering algorithm. More formally, let $Y_i(t)$ be the individual curve of name i, from 1 to n, observed at equally spaced time points $t_1, \ldots, t_M$, and suppose that individuals are spread among L unknown clusters of prior size π_l, with l from 1 to L. Then, a FCMM takes the form

$$Y_i(t) = \mu_l(t) + U_i(t) + E_i(t) \tag{1}$$

where $\mu_l(t)$ is the principal functional fixed effect that characterizes cluster l and $U_i(t)$ is a random function that is introduced for handling inter-individual random deviation and modelled as centered Gaussian process independent from $E_i(t)$, a random measurement error. Once defined in the functional domain, a classical approach for dealing with infinite-dimensional clustering problems is the filtering method which consists in projecting each curve onto a finite-dimensional functional basis. For the foregoing arguments, we adopt a wavelet representation which, in practice, is obtained by a discrete wavelet transform that converts the linear mixed effect model into the wavelet coefficients domain. Furthermore, this class of FCMMs allows random effect variance to vary over wavelet positions and/or groups. For estimation we resort to the EM-algorithm for maximum likelihood estimation provided by the recently developed R package curvclust [4] which is dedicated to model-based curve clustering and was originally thought for microarray-type data.

Two criteria for model selection have been used: the Bayesian Information Criterion (BIC) and an Integrated Classification Likelihood criterion (ICL) [1], which has proved to perform better than BIC both for choosing a mixture model and a relevant number of clusters. Both criteria select a FCMM with group-specific random effect variance. Vice versa, the directions for selecting the optimal number of clusters seem less clear-cut. In this work, for ease of interpretation, we illustrate the results of the selected model according to the BIC criterion (see left panel of Figure 2; right panel shows individual Y_i curves as well as group mean μ_l curves as inverse wavelet transform of estimated fixed effects for the identified clusters).

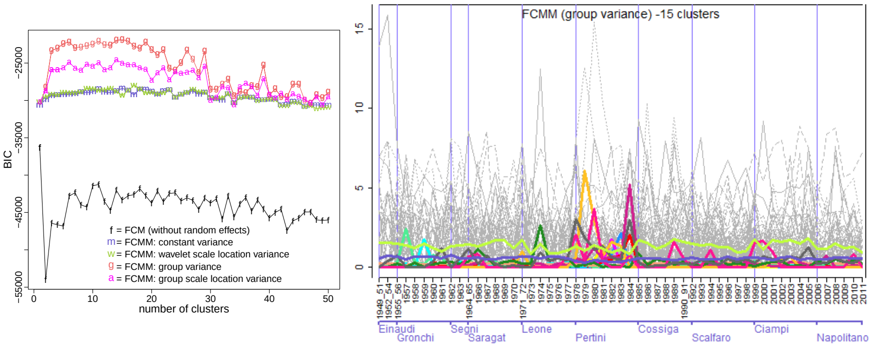

Fig. 2 Model selection according to the BIC criterion (left); individual and group mean curves for the selected FCMM with group-specific random effect variance and 15 clusters.

The found pattern in many clusters seem to prove that the “President effect” is predominant in shaping the principal trajectories. Previous research has already shown that the individual choices of each President makes the end-of-year messages less stereotyped than expected [2]. Moreover, with the group variance model, many clusters appear associated with Pertini who in fact distinguishes himself from all the others for giving an impromptu speech instead of sticking to the written text. An interesting issue then consists in disentangling lower-scale patterns from the higher-scale ones in order to detect the importance of a possible ”regime” factor (e.g. the President’s term of office) relatively to the temporal evolution of a chronological corpus. Investigation into wavelet coefficients domain turns out to be useful to inspect on different scales of the process.

References

1. Biernacki, C. , Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE PAMI, **22**(7), 719–725 (2000)
2. Cortelazzo, M. A., Tuzzi, A. (eds): Messaggi dal Colle. I discorsi di fine anno dei presidenti della Repubblica. Marsilio, Venezia (2007)
3. Gareth, M. J., Sugar, C. A.: Clustering for Sparsely Sampled Functional Data. J. Am. Stat. Ass. **98**, 397–408 (2003)
4. Giacofci, M., Lambert-Lacroix, S., Marot, G., Picard, F.: curvclust. R package (2012-01-07) http://cran.r-project.org/web/packages/curvclust/index.html.
5. Morris, J. S., Carroll, R. J.: Wavelet-based functional mixed models. J. Roy. Stat. Soc. B Met. **199**, 68–179 (2006)
6. Reithinger, F., Jank, W., Tutz, G., Shmuell, G.: Modelling price paths in on-line auctions: smoothing sparse and unevenly sampled curves by using semiparametric mixed models. Appl. Statist. **57**, 127–148 (2008)