DETECTING REGIME CHANGES IN MARKOV MODELS

Jesús E. García and V. A. González-López

University of Campinas, Brazil.
jg@ime.unicamp.br and veronica@ime.unicamp.br

ABSTRACT

Let C be a data collection, indexed by time. $C = \{D_{t_1}, ..., D_{t_n}\}$, where D_{t_i} was collected at time t_i, $t_i \leq t_j$ if $i \leq j$. Also, each D_{t_i} follows a Markovian model with finite alphabet A, denoted by M_{t_i}. We device a consistent procedure to detect changes in the model at time t_{i_0} that allows to decide if $D_{t_{i_0}}$ and $D_{t_{i_0-1}}$ are coming from the same Markovian source. The procedure is based on the equivalence relationship introduced by the Partition Markov Models, that allows to associate to each Markovian model a minimal number of parameters enough to describe a Markovian source. Under the possibility of regime change, we can have situations in which $D_{t_1}, ..., D_{t_{i_0-1}}$ are coming from a Markovian model, $M_{t_{i_0-1}}$ different to the Markovian model $M_{t_{i_0}}$ appropriated for $D_{t_{i_0}}$. We apply the procedure to detect prosodic changes from classical to modern European Portuguese. Taking in consideration that rhythm is a consequence of several characteristics, like number of syllables in the words, position in the word of the stressed syllable, simple and complex syllabic structure, etc., is possible to look for temporal changes in the rhythm, using written texts. In this context, each D_{t_i} is a written text in European Portuguese and t_i is the author's date of birth from 16th century to the 19th century.

1. INTRODUCTION

Our target is to explore whether rhythm related properties can be computed from European Portuguese written texts, and if by means of Markovian structures those properties can be useful for the study of prosodic changes on datasets of texts ordered by chronological time.

In [1] it is proved that in fact the European Portuguese has undergone a significant alteration, perceived from the 16th century to the 17th. This finding is consistent with the conjecture that the Portuguese is losing some of its features of "romance language" with the passing of the centuries. In [1] significant changes are verified on two phonological features, the size of the words and the position of the accent. Thus, from the 16th century to the centuries 17th, 18th and 19th, [1] shows: (a) a pattern of increase in the proportion of monosyllables (in the universe defined by words of at most two syllables) and (b) a pattern of increase in the proportion of words with stress on the last syllable (in the universe defined by words with accent positioned in the penultimate or in the last syllable). Those are the patterns that need to be further investigated. Also, in this paper we aim to investigate the problem by incorporating the Markovian structure inherent to written texts. So, (i) the number of syllables in each word and (ii) the placement of accent, will be investigated under a richest model that allows to incorporate a dependence structure between words through each text and enables consider jointly, the features (i) and (ii).

To achieve a more comprehensive view of linguistic phenomena studied at present, it should be noted that linguistic structures can be studied in their formats "spoken language" and "written language". The processing and types of statistical models are completely defined by the differentiated nature of the data. For example, for acoustic signal processing see [5] and for recent research about the statistical modeling see [7], [8] and [9].

2. HISTORICAL DATA

Tycho Brahe corpus is an annotated historical corpus, freely accessible at [3]. This corpus uses the chronological criterion of the author's birthdate to assign a time for written text. The subset of historical written texts included in this study, listed in Table 1 is composed by 17 texts from 15 authors, coming from five genres. In Table 1 we report also the number of orthographic words (ow) by text. The data collection $C = \{D_{t_1}, ..., D_{t_n}\}$ is now given by the written texts listed in Table 1.

2.1. Encoding texts

Each written text was processed with a slightly modified version of the perl-code "silaba" that can be freely downloaded for academic purposes at `www.ime.usp.br/~tycho/prosody/vlmc/tools/sil4.pl`. The software was used to extract two components of each orthographic word, denoted by (i, j), where i is the total number of syllables that make up the word, $i = 1, 2, ..., 8$ and j indicates the syllable in which is registered the stress in the word, $j = 0, 1, 2, .., 8$. Where, $j = 0$ means no stress in the word and this just happens in orthographic words with one syllable. The period (final of sentence) was codified as $(0, 0)$.

The alphabet was defined as exposed in Table 2. Note that the set of words represented by $(i, 0), i \geq 2$ corresponds to the empty set.

Table 1. Subset of Tycho Brahe corpus used in this study, coming from five genres: narrative (N), letters (L), philosophical (P), theatre (T) and sermons (S).

D_t	Gândavo	Pinto	Sousa	Brandão
t	1502	1510	1556	1584
Type	N	N	N	N
ow	22850	39941	50218	43192
D_t	Vieira	Vieira	Chagas	Bernardes
t	1608	1608	1631	1644
Type	L	S	P	P
ow	47888	49275	48670	49479
D_t	Oliveira	Aires	Costa	Alorna
t	1702	1705	1714	1750
Type	L	P	L	L
ow	16629	56055	24538	43318
D_t	Garrett	Garrett	Fronteira	Camilo
t	1799	1799	1802	1826
Type	L	N	N	N
ow	30070	45800	54826	20142
D_t	Ortigão			
t	1836			
Type	L			
ow	27420			

Table 2. Definition of the alphabet A.

orthographic word	element alphabet
$(0,0)$	a
$(1,0)$	b
$(1,1)$	c
$(2,1)$	d
$(2,2)$	e
$(i,1),\ i \geq 3$	f
$(i,2),\ i \geq 3$	g
$(i,j),\ i,j \geq 3$	h

3. THE MARKOVIAN MODEL

The Partition Markov Models applied in this paper, were introduced in [**?**]. Those models are generalizations of Variable Length Markov Chains models, used to discover the differences between branches of the Portuguese in [4]. Let (X_t) be a discrete time order M Markov chain on a finite alphabet A. Let us call $\mathcal{S} = A^M$ the state space. Denote the string $a_m a_{m+1} \dots a_n$ by a_m^n, where $a_i \in A$, $m \leq i \leq n$. For each $a \in A$ and $s \in \mathcal{S}$, $P(a|s) = \text{Prob}(X_t = a | X_{t-M}^{t-1} = s)$. Let $\mathcal{L} = \{L_1, L_2, \dots, L_K\}$ be a partition of $\mathcal{S}$, for $a \in A,\ L \in \mathcal{L}$,

$$P(L,a) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s, X_t = a),$$

$$P(L) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s) \text{ and } P(a|L) = \frac{P(L,a)}{P(L)},$$

with $P(L) > 0$.

Definition 1 *Let (X_t) be a discrete time order M Markov chain on a finite alphabet A. We will say that $s, r \in \mathcal{S}$ are equivalent (denoted by $s \sim_p r$) if $P(a|s) = P(a|r)\ \forall a \in A$. For any $s \in \mathcal{S}$, the equivalence class of s is given by $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.*

The previous definition allows to define a Markov chain with a "minimal partition", that is the one which respects the equivalence relationship.

Definition 2 *let (X_t) be a discrete time, order M Markov chain on A and let $\mathcal{L} = \{L_1, L_2, \dots, L_K\}$ be a partition of $\mathcal{S}$. We will say that (X_t) is a Markov chain with partition $\mathcal{L}$, if this partition is the one defined by the equivalence relationship $\sim_p$ introduced by definition 1.*

In a given sample x_1^n, coming from the stochastic process, we denote the number of occurrences of elements into L followed by a for,

$$N_n^{\mathcal{L}}(L,a) = \sum_{s \in L} N_n(s,a),\ L \in \mathcal{L},$$

where the number of occurrences of s in the sample x_1^n is denoted by $N_n(s)$ and the number of occurrences of s followed by a in the sample x_1^n is denoted by $N_n(s,a)$. The accumulated number of $N_n(s)$ for s in L is denoted by,

$$N_n^{\mathcal{L}}(L) = \sum_{s \in L} N_n(s),\ L \in \mathcal{L}.$$

The model, in this context given by the "minimal partition $\mathcal{L}$", can be selected consistently, using the Bayesian Information Criterion. This is, the best partition is the one that maximizes $\text{BIC}(x_1^n, \mathcal{L}) =$
$\sum_{a \in A, L \in \mathcal{L}} N_n^{\mathcal{L}}(L,a) \ln\left(\frac{N_n^{\mathcal{L}}(L,a)}{N_n^{\mathcal{L}}(L)}\right) - \frac{(|A|-1)|\mathcal{L}|}{2} \ln(n)$,
over the space of partitions.

3.1. Criterion of remoteness between the texts

The BIC allows to compare datasets as we will show, previously, we introduce the necessary notation. Given the dataset D_{t_i}, consider the stochastic process X_{t_i} generator of D_{t_i}, with sample $(x_{t_i})_1^{n_i}$ of size n_i. Following the codification given by Table 2, each sample will be composed by the concatenation of symbols from $A = \{a, b, c, d, e, f, g, h\}$. Based on previous works, that investigate similar data (see, for example [4]) the value of M considered here was 4.

Assuming that the data collection is made up of independent texts (which is the case treated here, as each text is a complete work in itself), the BIC under the assumption:

$$X_{t_i} \perp X_{t_j}; X_{t_i} =^d X_{t_j},$$

for an arbitrary partition $\mathcal{L}$ is given by
$\text{BIC}\Big((x_{t_i})_1^{n_i}, (x_{t_j})_1^{n_j}, \mathcal{L}\Big) =$

$$\sum_{a \in A, L \in \mathcal{L}} N_{n_i+n_j}^{\mathcal{L}}(L,a) \ln\left(\frac{N_{n_i+n_j}^{\mathcal{L}}(L,a)}{N_{n_i+n_j}^{\mathcal{L}}(L)}\right) - \frac{(|A|-1)}{2} |\mathcal{L}| \ln(n_i+n_j).$$

If,

$$\text{BIC}\Big((x_{t_i})_1^{n_i}, (x_{t_j})_1^{n_j}, \mathcal{L}\Big) > \sum_{k=i,j} \text{BIC}((x_{t_k})_1^{n_k}, \mathcal{L}^k)$$

so we can consider that both: D_{t_i} and D_{t_j} come from the same model, given by the minimal partition $\mathcal{L}$. In another case D_{t_i} and D_{t_j} come from different models, $\mathcal{L}^i$ and $\mathcal{L}^j$ respectively.
If we define $d_{i,j}$ as

$$\frac{2\sum_{a\in A} B(\mathcal{L}^i, n_i, a) + B(\mathcal{L}^j, n_j, a) - B(\mathcal{L}, n_i + n_2, a)}{(|A| - 1)\left\{|\mathcal{L}^i|\ln(n_i) + |\mathcal{L}^j|\ln(n_j) - |\mathcal{L}|\ln(n_i + n_j)\right\}},$$

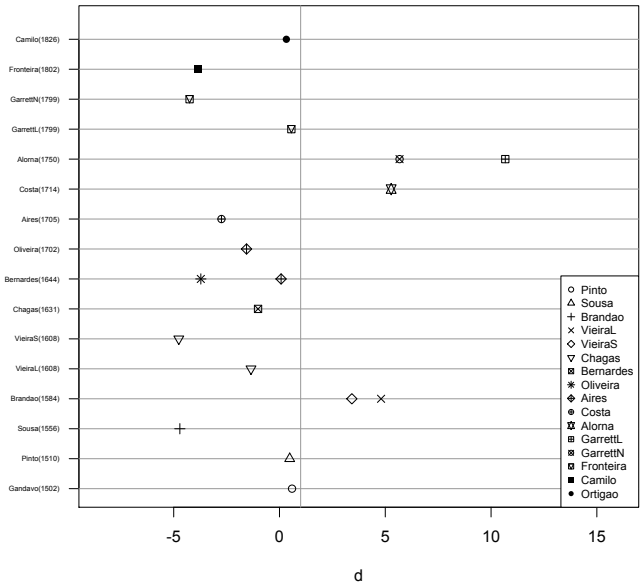
when $d_{i,j} \geq 1$, it indicates that D_{t_i} and D_{t_j} belong from different models, where
$B(\mathcal{L}, n, a) = \sum_{L\in\mathcal{L}} N_n^{\mathcal{L}}(L, a) \ln\left(\frac{N_n^{\mathcal{L}}(L,a)}{N_n^{\mathcal{L}}(L)}\right)$.
In the next section we use the values of $d_{i,j}$ to measure the distance between the models associated with written texts. Thus, texts that are identified with the same model show no change points in the timeline. When $d_{i,j}$ exceeds the value 1, a change point is identified.

4. RESULTS AND CONCLUSIONS

In the Figure 1, each horizontal line represents the text written by a particular author. On the line of each text is shown the value of d computed for two consecutive texts in time. Thus, for example, the text titled as "Gândavo (1502)" was compared with the author's text immediately following, that is the text titled as "Pinto (1510)" and the value of d displayed in Gândavo (1502)'s line. Now, when the line shows two points (two values of d), such as the case of the Brandão (1584)'s line, is because there are two texts in the sample of the same year. For instance, those texts are from Vieira (1608):(a) letters and (b) sermons.

In this analysis we detect two main change points, the first one at the turn of the 16th century to the 17th century. The second one, in the second half of the 18th century that spreads to the end of the century. Our findings complement the results attained in [2], which study the changes of the European Portuguese in the same period of time, through the analysis of clitic placement.

Figure 1. Each horizontal line represents a written text from European Portuguese, those were ordered by time from down to top. The vertical line represents $d = 1$, greater values of d indicates the presence of a change point.



5. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support for this research provided by USP project "Mathematics, computation, language and the brain" and FAPESP's projects (a)"Portuguese in time and space: linguistic contact, grammars in competition and parametric change" 2012/06078-9 and (b)"Research, Innovation and Dissemination Center for Neuromathematics - NeuroMat" 2013/07699-0.

6. REFERENCES

[1] Frota, S., Galves, C., Vigário, M., Gonzalez-Lopez, V. and Abaurre, B. (2012). The phonology of rhythm from Classical to Modern Portuguese. *Journal of Historical Linguistics*, 2(2), 173-207.

[2] Galves, C., Britto, H. and de Sousa, M. C. P. (2005). The Change in Clitic Placement from Classical to Modern European Portuguese. *Journal of Portuguese Linguistics*, 4(1), 39-67.

[3] Galves, C. and Faria, P. (2010). *Tycho Brahe Parsed Corpus of Historical Portuguese*.http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html

[4] Galves, A., Galves, C., Garcia, J. E., Garcia, N. L. and Leonardi, F., "Context tree selection and linguistic rhythm retrieval from written texts", *Annals of Applied Statistics*, **6** 1, 186 – 209, 2012.

[5] Garcia, J., Gut, U., and Galves, A., "Vocalea semi-automatic annotation tool for prosodic research". *Speech Prosody 2002, International Conference*, 2002.

[6] Garcia, J. and Gonzalez-Lopez, V. A., "Minimal Markov Models". arXiv preprint *arXiv:1002.0729*, 2010.

[7] Garcia, J. E., Gonzalez-Lopez, V. A. and Viola, M. L. L., "Robust model selection and the statistical classification of languages", *AIP Conference Proceedings*, vol. 1490, p.160, 2012.

[8] García, Jesús E., V. A. González-López, and R. B. Nelsen (2013). A new index to measure positive de-

pendence in trivariate distributions. *Journal of Multivariate Analysis* **115**, 481-495.

[9] García, Jesús E., González-López, V. A., Viola, M. L. L. (2013). Robust Model Selection for Stochastic Processes. *Communications in Statistics Theory and Methods* (Forthcoming).