Comparison of Stochastic Processes 17th ASMDA Conference

Jesús E. García V.A. González-López

June 9, 2017

Introduction

- A distance which allows to compare Markovian processes is introduced.
- It is shown the relationship of this distance to the divergence of Kullback Leibler and revealed its stochastic behavior in terms of the Chi-squared distribution.
- The distance allows to decide if there is any discrepancy between two samples and to find the strings where the discrepancy is manifested.
- We apply the distance to written texts of European Portuguese coming from two authors.
- In the application the distance reveals the linguistic configurations that expose discrepancies between written texts of different genres from the same author.

Notation

For (X_t) a discrete time, order $o < \infty$ Markov chain on a finite alphabet A,

- $\mathcal{S} = A^o$ is the state space,
- $a_m^n = a_m a_{m+1} \dots a_n$ where $a_l \in A$, $m \leq l \leq n$,
- for each $a \in A$ and $s \in \mathcal{S}$, $P_X(a|s) = \text{Prob}(X_t = a | X_{t-M}^{t-1} = s)$.
- $x_1^{n_x}$, a size n_x sample of process (X_t).
- $N_x(s)$ is the number of occurrences of s in the sample x_1^n.
- $N_x(s, a)$ is he number of occurrences of s followed by a in the sample x_1^n,

For (X_t) and (Y_t) two discrete time, order $o < \infty$ Markov chains on the same finite alphabet A, consider $x_1^{n_x}$, a size n_x sample of process (X_t) and $y_1^{n_y}$, a size n_y sample of process (Y_t).

- $N_{x,y}(s) = N_x(s) + N_y(s)$,
- $N_{x,y}(s, a) = N_x(s, a) + N_y(s, a)$,

Distance

Definition

For two such samples $x_1^{n_x}$, $y_1^{n_y}$ and $s \in \mathcal{S}$,

$$d_s(x_1^{n_x}, y_1^{n_y}) = \frac{1}{b(x_1^{n_x}, y_1^{n_y}} \sum_{a \in A} \left\{ N_x(s,a) \ln\left(\frac{N_x(s,a)}{N_x(s)}\right) \right.$$

$$\left. + N_y(s,a) \ln\left(\frac{N_y(s,a)}{N_y(s)}\right) - N_{x,y}(s,a) \ln\left(\frac{N_{x,y}(s,a)}{N_{x,y}(s)}\right) \right\}$$

where $b(x_1^{n_x}, y_1^{n_y}) = |A| - 1) \ln(n_x + n_y)$

Properties

i. *For a fixed string $s \in \mathcal{S}$, the function $d_s(x_1^{n_x}, y_1^{n_y})$ is a distance.* Consider $x_1^{n_x}, y_1^{n_y}, z_1^{n_z}$ three samples of order o Markov Chains on the alphabeth A,

$$d_s(x_1^{n_x}, y_1^{n_y}) \geq 0 \text{ with equality } \Leftrightarrow \frac{N_x(s,a)}{N_x(s)} = \frac{N_y(s,a)}{N_y(s)} \ \forall a \in A,$$

$$d_s(x_1^{n_x}, y_1^{n_y}) = d_s(y_1^{n_y}, x_1^{n_x}),$$

$$d_s(x_1^{n_x}, y_1^{n_y}) \leq d_s(x_1^{n_x}, z_1^{n_z}) + d_s(z_1^{n_z}, y_1^{n_y}).$$

ii. *Local consistence.* If the stochastic laws of (X_t) and (Y_t) are the same in s, then $d_s(x_1^{n_x}, y_1^{n_y}) \underset{\min(n_x,n_y)\to\infty}{\longrightarrow} 0$.

Otherwise $d_s(x_1^{n_x}, y_1^{n_y}) \underset{\min(n_x,n_y)\to\infty}{\longrightarrow} \infty$.

Properties

Generalizing (for partition Markov models) ressults in [1] we obtain the following theorem.

Theorem

Let (X_t) be an order o Markov chain with finite alphabet A and $x_1^{n_x}$, $y_1^{n_y}$ two independent samples of (X_t). Consider also $s \in \mathcal{S}$.
a)

$$\begin{aligned} d_s(x_1^{n_x}, y_1^{n_y}) &= \frac{N_x(s)}{b(x_1^{n_x}, y_1^{n_y})} D\Big(\frac{N_x(s,\cdot)}{N_{n_k}(s)} \parallel \frac{N_{x,y}(s,\cdot)}{N_{x,y}(s)}\Big) \\ &+ \frac{N_x(s)}{b(x_1^{n_x}, y_1^{n_y})} D\Big(\frac{N_y(s,\cdot)}{N_y(s)} \parallel \frac{N_{x,y}(s,\cdot)}{N_{x,y}(s)}\Big) \end{aligned}$$

and

Theorem

b)

$$
\begin{aligned}
d_s(x_1^{n_x}, y_1^{n_y}) \quad \sim_d \quad & \frac{1}{2b(x_1^{n_x}, y_1^{n_y})} \chi^2\left(\frac{N_x(s,\cdot)}{N_x(s)}, P_X(.|s)\right) \\
+ \quad & \frac{1}{2b(x_1^{n_x}, y_1^{n_y})} \chi^2\left(\frac{N_y(s,\cdot)}{N_y(s)}, P_X(.|s)\right) \\
+ \quad & \frac{1}{2b(x_1^{n_x}, y_1^{n_y})} \chi^2\left(\frac{N_{x,y}(s,\cdot)}{N_{x,y}(s)}, P_X(.|s)\right)
\end{aligned}
$$

where $\sim_d$ means similarity in distribution.

Application to Linguistic Data

Tycho Brahe corpus is an annotated historical corpus, freely accessible at Galves and Faria (2010) [3]. This corpus uses the chronological criterion of the author's birthdate to assign a time for written texts. The subset of written texts included in this study, listed in table 9 is composed by six texts from two authors.

Author	Vieira	Vieira	Vieira
Date	1608	1608	1608
Type	dissertation	letters	sermons
Notation	1608d	1608c	1608s
Author	Garrett	Garrett	Garrett
Date	1799	1799	1799
Type	letters	narrative	theater
Notation	1799c	1799n	1799t

Application to Linguistic Data

Linguistic studies show that the variability observed in different written texts of European Portuguese involves, among other aspects,

- changes in the proportion of occurrence of the placement of the stress in the last or in the penultimate syllable of the word and
- alterations in the use of monosyllables, with or without stress, see for instance Frota et al. (2012)[2].

For this reason we guide our inspection to the position in the word occupied by the stress and the size of the word (number of syllables).

Application to Linguistic Data

Each written text was processed with a slightly modified version of the perl-code "silaba" by Miguel Galves.
`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),
- i is the total number of syllables which compound the word, $i = 1, 2, ..., 8$ and
- j indicates the syllable (from left to right) in which is registered the stress in the word. Where, $j = 0$ means no stress in the word.
- The period (final of sentence) was codified as $(0, 0)$.

Application to Linguistic Data

Word code	Letter in A	Meaning
$(0,0)$	0	final of sentence
$(1,1)$	1	monosyllable with stress
$(1,0)$	2	monosyllable without stress
$(2,2)$	3	dissyllable - stress in the last syllable
$(2,1)$	4	dissyllable - stress in the first syllable
(i,i), $i \geq 3$	6	*oxytone* word
$(i,i-1)$, $i \geq 3$	7	*paroxytone* word
$(i,i-2)$, $i \geq 3$	8	*proparoxytone* word

Table: Definition of the alphabet A.

Application to Linguistic Data

We can define

$$dmax = \max\{d_s(x_1^{n_x}, y_1^{n_y}), s \in \mathcal{S}\} \quad (1)$$

and

$$smax = \arg\max\{d_s(x_1^{n_x}, y_1^{n_y}), s \in \mathcal{S}\}. \quad (2)$$

Observe that $dmax < \epsilon$ if and only if $d_s(x_1^{n_x}, y_1^{n_y}) < \epsilon, \forall s \in \mathcal{S}$. That is, a small value of $dmax$ indicates the stochastic laws on s are similar for all $s \in \mathcal{S}$. In other words the distributions of the processes are similar.

Application to Linguistic Data

Let $d_s^n(1,2)$ be the distance for $s \in \mathcal{S}$ when the size of the sample for both process is n. If the two processes have the same law, then for each $s \in \mathcal{S}$,

$$d_s^n(1,2) \underset{n\to\infty}{\longrightarrow} 0.$$

If the local laws for s are different then,

$$d_s^n(1,2) \underset{n\to\infty}{\longrightarrow} \infty.$$

We can see that if *dmax* is large, *smax* is exactly the string we want to recognize, as being relevant in terms of discrepancy but all the strings with a large relative value of d will reveal changes on the local laws of the processes relative to the string.
We note that the comparison is made between the different texts of the same author. The memory o used in this application is equal to 2.

$d_s(1608c, 1608d)$	s	$d_{1608c,1608s}(s)$	s	$d_s(1608d, 1608s)$	s
1.02591	7-6	1.18101	4-7	1.07770	1-7
1.11191	1-6	1.28567	2-3	1.07883	4-4
1.13048	3-6	1.98674	2-7	1.33124	4-7
2.14046	**7-2**	**3.86756**	**2-4**	1.67395	2-4
				1.74245	**2-7**

$d_s(1799c, 1799t)$	s	$d_s(1799t, 1799n)$	s
1.13432	1-7	1.01398	1-7
1.20717	4-4	1.07517	6-2
1.29197	7-0	1.24806	1-2
2.15512	4-2	1.34589	3-2
2.35864	4-7	2.56588	2-4
2.84146	2-7	2.57690	4-7
3.40959	7-2	3.56924	4-2
3.46598	**2-4**	3.74332	2-7
		4.49460	**7-2**

Table: Cases with values of $d > 1$;1608c 1608d

Application to Linguistic Data

Meaning of each *smax* detected by *dmax*.

- 2-4 a monosyllable without stress followed by a dissyllable with stress in the first syllable
- 2-7 a monosyllable without stress followed by a paroxytone word
- 7-2 a paroxytone word followed by a monosyllable without stress

Other studies in the area show that the strings 2-4, 7-2 and 2-7 are volatile configurations of the European Portuguese (from the 16th century to the 19th century) see García et al. (2017) [2]. We can see that this characteristic persists when analyzing the variability of different written texts of the same author, being that author: Vieira or Garrett.

Application to Linguistic Data

$a \in A$	*smax*: 7-2 1608c	(2.14046) 1608d	*smax*: 2-4 1608c	(3.86756) 1608s	*smax*: 2-7 1608d	(1.74245) 1608s
0	0.00000	0.00000	0.02277	0.06825	0.05221	0.11444
1	0.09461	0.09927	0.06624	0.07620	0.05200	0.05488
2	0.20037	0.15025	0.31714	0.35255	0.50989	0.47124
3	0.07616	0.04350	0.04251	0.04473	0.02926	0.03284
4	0.31379	0.29417	0.20402	0.22949	0.16547	0.16809
6	0.03763	0.03986	0.02638	0.02175	0.02337	0.02203
7	0.26082	0.34266	0.31144	0.19307	0.15053	0.12464
8	0.01662	0.03028	0.00949	0.01397	0.01726	0.01183

$a \in A$	*smax*: 2-4 1799c	(3.46598) 1799t	*smax*: 7-2 1799t	(4.49460) 1799n
0	0.05469	0.13364	0.01100	0.00183
1	0.06448	0.10649	0.13265	0.10353
2	0.28798	0.27959	0.34777	0.15735
3	0.04389	0.04200	0.06598	0.05153
4	0.19514	0.24608	0.23505	0.29684
6	0.02971	0.01273	0.02749	0.02863
7	0.30959	0.17522	0.17113	0.32410
8	0.01452	0.00424	0.00893	0.03619

Table: Conditional probabilities $P(a|smax)$, $\forall a \in A$ computed from each written text: 1608c, 1608d; 1608c, 1608s;1608d, 1608s; 1799c, 1799t; 1799t, 1799n

Application to Linguistic Data

Table 4 shows the transition probabilities $P(a|smax)\ \forall a \in A$, for each pair of compared texts. With this information we can check the differences between the written texts in relation to the prosodic construction, for example $P(2|7-2)$ is 0.34777 in the text 1799t (theater) and it goes to 0.15735 in the written text 1799n (narrative) both texts from Garrett. Moreover, the most probable choice for the second text, since the string 7-2 has been observed is 7 ($P(7|7-2) = 0.3241$).

Application to Linguistic Data

We can define 3 groups of strings:

- (i) strings that show discrepancies between Vieira's texts but not in the case of Garrett's texts,
- (ii) strings that show discrepancies between Garrett's texts and not in the case of Vieira's texts and
- (iii)strings that show discrepancies between texts for each of these authors.

Author	String	Meaning
Vieira	1-6	monosyllable with stress followed by an *oxytone* word
	2-3	monosyllable without stress followed by a dissyllable with stress in the last syllable
	3-6	dissyllable with stress in the last syllable followed by an *oxytone* word
	7-6	*paroxytone* word followed by an *oxytone* word
Garrett	1-2	monosyllable with stress followed by a monosyllable without stress
	3-2	a dissyllable with stress in the last syllable followed by a monosyllable without stress
	4-2	a dissyllable with stress in the first syllable followed by a monosyllable without stress
	6-2	an *oxytone* word followed by a monosyllable without stress
	7-0	a *paroxytone* word followed by final of sentence
Both	1-7	a monosyllable with stress followed by a *paroxytone*
	4-4	a dissyllable with stress in the first syllable followed by a dissyllable with stress in the first sylla
	4-7	a dissyllable with stress in the first syllable followed by a *paroxytone* word

Strings (see table 3) and meaning of the linguistic compositions that characterize the variability between the texts of the same author. We also list the strings (with $d > 1$) that are common among the authors, $2-4$, $2-7$ and $7-2$ are excluded.

Application to Linguistic Data

Values of d greater than 1 have not been detected in the comparison between the texts: 1799c (letters) and 1799n (narrative). Thus, these texts can be considered as coming from the same Markovian process.

Conclusions

- The distance proposed in this paper has a clear relation to the divergence of Kullback Leibler, we show this in the theorem.
- In addition, the adequately scaled distance has its stochastic behavior described by a sum of Chi-squared dependent random variables, also seen in the theorem
- In relation to the application, note that the distance introduced here makes it possible to decide whether two Markovian stochastic processes follow the same law or not. And it also allows to identify discrepancies pointing out the strings responsible for them.

Bibliography


I. Csiszar and P. Shields. The consistency of the bic markov order estimator. *The Annals of Statistics* 28, (6), 1601-1619, 2000


S. Frota, C. Galves, M. Vigário, V.A. González-López and B. Abaurre. The phonology of rhythm from Classical to Modern Portuguese. *Journal of Historical Linguistics* (2.2) 173-207, 2012.


C. Galves and P. Faria. Tycho Brahe Parsed Corpus of Historical Portuguese
`http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html`


A. Galves, C. Galves, J.E. Garcia, N.L. Garcia and F. Leonardi, F. Context tree selection and linguistic rhythm retrieval from written texts. *The Annals of Applied Statistics* 6(1), 186-209, 2012.


J.E. García, R. Gholizadeh and V.A. González-López. Linguistic Compositions Highly Volatile in Portuguese. *Submitted*.

Bibliography

J.E. García and M. Fernández. Copula based model correction for bivariate Bernoulli financial series. In *11TH INTERNATIONAL CONFERENCE OF NUMERICAL ANALYSIS AND APPLIED MATHEMATICS 2013: ICNAAM 2013,* vol. 1558, no. 1, pp. 1487-1490. AIP Publishing, 2013.

García, J.E.; González-López, V.A. Consistent Estimation of Partition Markov Models. Entropy 2017, 19, 160.

J.E. García and V.A. González-López. Detecting regime changes in Markov models. In SMTDA2014 Book.