

### UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

RAMIN GHOLI ZADEH

### A BIC-Based Distance Between Stochastic Processes

# Uma distância baseada no BIC entre processos estocásticos

Campinas 2018 Ramin Gholi Zadeh

### A BIC-Based Distance Between Stochastic Processes

# Uma distância baseada no BIC entre processos estocásticos

Tese apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Estatística.

Thesis presented to the Institute of Mathematics, Statistics and Scientific Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Statistics.

Supervisor: Verónica Andrea González-López

Este exemplar corresponde à versão final da Tese defendida pelo aluno Ramin Gholi Zadeh e orientada pela Profa. Dra. Verónica Andrea González-López.

> Campinas 2018

Ficha catalográfica Universidade Estadual de Campinas Biblioteca do Instituto de Matemática, Estatística e Computação Científica Ana Regina Machado - CRB 8/5467

Zadeh, Ramin Gholi, 1982-A BIC-based distance between stochastic processes / Ramin Gholi Zadeh. – Campinas, SP : [s.n.], 2018.
Orientador: Verónica Andrea González-López. Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.
1. Markov, Processos de. 2. Estatística robusta. 3. Processo estocástico. 4. Métodos estatísticos robustos. I. González-López, Verónica Andrea, 1970-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

#### Informações para Biblioteca Digital

Título em outro idioma: Uma distância baseada no BIC entre processos estocásticos Palavras-chave em inglês: Markov processes **Robust statistics** Stochastic processes Robust statistical methods Área de concentração: Estatística Titulação: Doutor em Estatística Banca examinadora: Verónica Andrea González-López Pedro Jose Catuogno Adriano Polpo de Campos Márcio Luis Lanfredi Viola Luís Gustavo Esteves Data de defesa: 23-02-2018 Programa de Pós-Graduação: Estatística

Tese de Doutorado defendida em 23 de fevereiro de 2018 e aprovada

pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). VERÓNICA ANDREA GONZÁLEZ LÓPEZ

Prof(a). Dr(a). PEDRO JOSE CATUOGNO

Prof(a). Dr(a). ADRIANO POLPO DE CAMPOS

Prof(a). Dr(a). MÁRCIO LUIS LANFREDI VIOLA

**Prof(a). Dr(a). LUIS GUSTAVO ESTEVES** 

As respectivas assinaturas dos membros encontram-se na Ata de defesa

This thesis is dedicated to my wife, who has been a constant source of encouragement throughout all my life and has made countless sacrifices to help me get to this point. This work is also dedicated to my parents, who have always loved me unconditionally and whose good example have taught me to work hard for the things that I aspire to achieve.

### Acknowledgements

First at all, I am extremely grateful to my supervisor, *Prof. Dra. Verónica* Andrea González-López, for the patient guidance, encouragement and advice she has provided throughout my PhD program. Her deep insights helped me at various stages of my research. I have been extremely lucky to have a supervisor who cared so much about my thesis, and who responded to my questions and queries so promptly.

I would also like to take this opportunity to thank *Prof. Dr. Jesus Enrique Garcia* for his very helpful comments and suggestions.

Very special thanks to the Coordination of Improvement of Higher Level Personnel (CAPES) and to the National Council for Scientific and Technological Development (CNPq) for giving me the opportunity to carry out my doctoral research and for their financial support. Finally, I want to give a special thanks to the Jury Committee include of Professors Pedro José Catuogno, Adriano Polpo de Campos, Marcio Luis Lanfredi Viola and Luis Gustavo Esteves for corrections and suggestions.

"There are three kinds lies: lies, damned lies and statistics, but remember: statistics do not lie, liars fake statistic.

Benjamin Disraeli

# Resumo

Nesta tese lidamos com dois problemas estatísticos. Investigamos as propriedades de uma métrica entre amostras de processos estocásticos e usamos ela para desenvolver um procedimento robusto de seleção de amostras. Abordamos o problema de decidir se duas amostras independentes provenientes de processos Markovianos discretos são regidas pela mesma lei estocástica. No caso em que as leis que geram os processos não são as mesmas, a metodologia apresentada neste documento permite detectar os elementos específicos do espaço de estados em que as discrepâncias se manifestam. Nós estabelecemos uma métrica local entre amostras com base no critério de Informação Bayesiano (Bayesian Information Criterion), derivamos a fronteira que deve ser usada nesta métrica para tomar a decisão. Mostramos que a distância é estatisticamente consistente para detectar se as amostras seguem a mesma lei, tendendo a zero quando os tamanhos das amostras aumentam. Mostramos que a métrica assume valores arbitrariamente grandes quando os tamanhos de amostra aumentam e as leis estocásticas são diferentes. Também mostramos a relação dessa distância com a divergência de Kullback Leibler e revelamos seu comportamento estocástico em termos de distribuição Qui-quadrado. O objetivo de segundo problema é o de postular um método de seleção de amostras de um conjunto de amostras provenientes de processos Markovianos de ordem finita e alfabeto finito. Sob a suposição da existência de uma lei que prevalece em pelo menos 50% das amostras da coleção, mostramos que o procedimento permite identificar amostras regidas pela lei predominante. A abordagem é baseada na distância local entre amostras, que tende a zero quando comparamos amostras de lei idêntica e tende ao infinito ao comparar amostras com diferentes leis. A distância local permite definir um critério que registra valores arbitrariamente grandes quando a suposição anterior sobre a existência de uma lei predominante não é válida. Aplicamos os conceitos e resultados aqui introduzidos em bases de dados de diversas áreas, como linguística, genêtica, indústria e finanças.

**Palavras-chave**: Critério de informação Bayesiano, Entropia relativa, Processo de Markov, Robustez.

## Abstract

In this thesis we deal with two statistical issues. We investigate the properties of a metric between samples from stochastic processes and we use it to develop a robust sample selection procedure. We address the problem of deciding if two independent samples, coming from discrete Markovian processes, are governed by the same stochastic law. In the case on which is decided that the laws generating the processes are not the same, the methodology presented in this research allows to detect the specific elements of the state space where the discrepancies are manifested. We establish a local metric between samples based on the Bayesian Information Criterion, we derive the bound that must be used in this metric to take the decision. We show that the distance is statistically consistent to detect if the samples follow the same law, tending to zero when the sample sizes increase. we show that the metric assumes arbitrarily large values when the sample sizes increase and the stochastic laws are different. Also we show the relationship of this distance to the divergence of Kullback Leibler which reveals its stochastic behavior in terms of the Chi-squared distribution. We introduce a method of selecting samples from a set of samples coming from Markovian processes of finite order and finite alphabet. Under the assumption of the existence of a law that prevails in at least 50% of the samples of the collection, we show the procedure which allows to identify samples governed by the predominant law. The approach is based on the local metric between samples, which tends to zero when we compare samples of identical law and tends to infinity when comparing samples with different laws. The local distance allows to define a criterion which takes arbitrarily large values when the previous assumption about the existence of a predominant law does not hold. We illustrate both, the use of this metric and the procedure of selecting samples through applications in real data.

**Keywords**: Bayesian information criterion, Relative entropy, Markov processes, Robustness.

# List of Figures

Figure 1 –	Graphical representation: on the left the tree $\mathcal{T}_1$ and on the right the tree $\mathcal{T}_2$ . In both cases, the contexts correspond to the sequences obtained	
	by concatenating the symbols from the leaves to the root of the trees	25
Figure 2 –	Left: context tree $\mathcal{T}_P$ . Middle: context tree $\mathcal{T}_Q$ . Right: context tree $\mathcal{T}_{PQ}$ .	30
Figure 3 –	Graphical representation: context tree in the example 2.3.3	31
Figure 4 –	<i>dmax</i> values (on the vertical axis) denoted by the year of the second written text (column 2 of the table 4). In the case of a year with several texts, a symbol was attached to the year, which indicates the type of written text: narrative (n), letters (c), sermons (s), theater (t), dissertation (d)	51
Figure 5 –	Dendrograms build through the <i>dmax</i> values (tables 3.13-14), agglomeration method: <i>Average</i> , on the left and <i>Complete</i> , on the right	61
Figure 6 –	Representative fifth period, 5 September 2016 - 31 October 2017, for BDDC4 using of equation (4.4).	86
Figure 7 –	Fourth period, 7 July 2015 - 2 September 2016, for BDDC4 using of equation (4.4).	86
Figure 8 –	First period, 3 January 2012 - 7 March 2013, for BDDC4 using of time series.	102
Figure 9 –	Second period, 3 January 8 March 2013 - 7 May 2014, for BDDC4 using of time series.	102
Figure 10 –	Third period, 8 May 2014 - 6 July 2015, for BDDC4 using of time series	.103
Figure 11 –	First period, 3 January 2012 - 7 March 2013, for BVMF3 using of time series.	103
Figure 12 –	Second period, 3 January 8 March 2013 - 7 May 2014, for BVMF3 using of time series.	104
Figure 13 –	Representative third period, 8 May 2014 - 6 July 2015, for BVMF3 using of time series.	104
Figure 14 –	Fourth period, 7 July 2015 - 2 September 2016, for BVMF3 using of time series.	105
Figure 15 –	Fifth period, 5 September 2016 - 31 October 2017, for BVMF3 using of time series.	105
Figure 16 –	First period, 3 January 2012 - 7 March 2013, for BBAS3 using of time series.	106
Figure 17 –	Second period, 3 January 8 March 2013 - 7 May 2014, for BBAS3 using of time series	106
Figure 18 –	Third period, 8 May 2014 - 6 July 2015, for BBAS3 using of time series	. 107

Figure 19 – Representative fourth $\pm$	period, 7 July 2015 - 2 September 2016, for
BBAS3 using of time se	ries
Figure 20 – Fifth period, 5 Septemb	er 2016 - 31 October 2017, for BBAS3 using of
time series.	
Figure 21 – First period, 3 January	2012 - $7$ March $2013$ , for ITUB4 using of time
series.	
Figure 22 – Third period, 8 May 20 $$	$14$ - $6~{\rm July}~2015$ , for ITUB4 using of time series. 109
Figure 23 $-$ Representative second p	period, 8 March 2013 - 7 May 2014, for ITUB4
using of time series	
Figure 24 – Fourth period, 8 May 20	) 14 - 6 July 2015, for BBAS3 using of time series. 110 $$
Figure 25 – Fifth period, 5 Septemb	er 2016 - 31 October 2017, for ITUB4 using of
time series.	

# List of Tables

Table 1 –	Table of transition probabilities with alphabet $\mathcal{A} = \{0, 1\}$	26
Table 2 –	The set of the Tycho Brahe corpus.	49
Table 3 –	Definition and meaning of each element $a \in \mathcal{A}$	49
Table 4 –	Values of $dmax$ between a written text and the written text, dated	
	immediately after the previous one	50
Table 5 –	Conditional probabilities from $smax$ to each element $a$ of the alphabet	
	$\mathcal{A}$ . Texts: 1502, 1510, 1556, 1584, 1608c, 1608d, 1608s, 1631	52
Table 6 –	Conditional probabilities from $smax$ to each element $a$ of the alphabet	
	$\mathcal{A}$ . Texts: 1631, 1644, 1702, 1705, 1714, 1750, 1799c, 1799t, 1799n	53
Table 7 $-$	Conditional probabilities from $smax$ to each element $a$ of the alphabet	
	$\mathcal{A}$ . Texts: 1799c, 1799t, 1799n, 1802, 1826	54
Table 8 –	Cases with bigger values of $d_s$ and different $smax$ : 1750-1799t, 1799n-	
	1802,1799t-1802. In bold the bigrams that most often produce the highest	
	values of $d_s$ .	55
Table 9 $-$	Values of $d_s$ and bigrams such that $d_s > 1$ , between texts of the 16th	
	century and Vieira's texts: 1608c, 1608d and 1608s. In bold letter the $\hfill \hfill \hfill$	
	most frequent bigrams, according to the previous section. $\ldots$ $\ldots$	56
Table 10 –	Values of $d_s$ and bigrams such that $d_s > 1$ , between texts of the 16th	
	century and Vieira's texts: 1608c, 1608d and 1608s. In bold letter the	
	most frequent bigrams, according to the previous section	56
Table 11 –	Values of $d_s$ and bigrams such that $d_s > 1$ , between texts of the 16th	
	century and Vieira's texts: 1608c, 1608d and 1608s. In bold letter the	
	most frequent bigrams, according to the previous section	57
Table 12 –	Bigrams that announce changes between texts of the 16th century when	
	compared to texts of beginning of the 17th century: 1608c, 1608d, 1608s.	
	In the third column are indicated the cases covered by the configuration,	
	see tables 3.8, 3.9, 3.10	57
Table 13 –	Sample sizes $n$ of DNA sequence coming from 15 patients with Burkitt	
	lymphoma/leukemia, AM2871z.1, where $z = 39, 40, 41, 46, 50, 52, 57,$	
<b>T</b> 11 4 (	58, 59, 61, 62, 65, 76, 81, 87	59
Table 14 –	$dmax(i, j)$ values, $i \neq j, i, j = AM2871z.1$ , where $z = 39, 40, 41, 46, 50, 52, 57$ ,	<u>a</u> 0
	58, 59, 61, 62, 65, 76, 81, 87.	<u>50</u>
Table 15 –	$dmax(i, j)$ values, $i \neq j, i, j = AM28712.1$ , where $z = 39, 40, 41, 46, 50, 52, 57, 50, 50, 50, 61, 62, 65, 76, 61, 62, 41, 46, 50, 52, 57, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50$	
	$58, 59, 61, 62, 65, 76, 81, 87$ . In bold the lowest value of $\mathcal{S}$ , associated to	
	the sequence with $z = 816$	50

Table 16 –	Relation between the sequences used in the estimation and number of	
	parts of the estimated partition, for AM2871z.1, where $z=39, 40, 41, 46,$	
	$50, 52, 57, 58, 59, 61, 62, 65, 76, 81, 87. \dots \dots \dots \dots \dots \dots \dots \dots$	62
Table 17 –	Parts of the partition selected through the Bayesian Information Criterion,	
	using AM2871z.1, where $z=39, 46, 52, 57, 61, 76, 81. \dots$	63
Table 18 –	Transition probabilities $P(\cdot L_i)$ with $\cdot \in \{a,c,g,t\}$ and $i = 1, \dots, 27$ . For	
	each part $i$ , listed on the left column (see table 17), we indicate in bold	
	the highest transition probability to the elements of the alphabet	64
Table 19 –	Selected parts, from table 17, which have the greater (on top)/null (on	
	bottom) transition probabilities to each element of the alphabet {a,c,g,t}.	65
Table 20 –	Parts of the partition selected through the Bayesian Information Criterion,	
	using all the sequences AM2871z.1, where $z = 39, 40, 41, 46, 50, 52, 57, 58$ ,	
	59, 61, 62, 65, 76, 81, 87	66
Table 21 –	Transition probabilities $P(. L_i)$ with $. \in \{a, c, g, t\}$ and $i = 1, \dots, 42$ . For	
	each part $i$ , listed on the left column (see table 20), we indicate in bold	
	the highest transition probability to the elements of the alphabet.	67
Table 22 –	Selected parts, from table 20 and 21, which have the greater transition	
	probabilities to each element of the alphabet $\{a,c,g,t\}$ .	68
Table 23 –	Relation between the parts listed in table 19 and 20. On left we display	
	the parts coming from the model using only $50\%$ of the DNA sequences,	
	on right the parts coming from the model using all the DNA strains. In	
	the same line, on the right we list the parts in which are identified the	
	elements into the part on the left.	68
Table 24 –	For each order $M, M = 1, 2$ , on the right we list the strings $(s)$ in which	
	the columns $i, i = 1, 2$ are considered as being different. On the left	
	we inform the value of $d_s$ . In bold, we highlight the cases with greater	
	distances.	70
Table 25 –	Transition probabilities from the string $s$ to each element of the alphabet	
	$\mathcal{A} = \{0,1\}^5$ (listed on the left), for each process (column 1 and column	
	2) where $s = s_0, s_1, s_2; s_0 = (1, 1, 1, 1, 1), s_1 = (1, 0, 1, 1, 1)$ and $s_2 = $	
	$(1, 1, 0, 1, 1)$ with $d_s > 1$ , for the case $M = 1$ - see table 24	71
Table 26 –	On the right we list the strings $(s)$ in which the columns $i, i = 1, 2$ are	
	considered as being different, according to the variable $X_t^i(1)$ built from	
	the <i>alcoholic content</i> . On the left we inform the intervals in which the	
	values of $d_s$ are included. The strings are listed according to the $d_s$ values,	
	in an increasing order from left to right and from top to bottom	72

Table 27 –	On the right we list the strings $(s)$ in which the columns $i, i = 1, 2$ are considered as being different, according to the variable $X_t^i(2)$ built from the <i>fill level</i> . On the left we inform the intervals in which the values of $d_s$ are included. The strings are listed according to the $d_s$ values, in an increasing order from left to right and from top to bottom	73
Table 28 –	On the right we list the strings $(s)$ in which the columns $i, i = 1, 2$ are considered as being different, according to the variable $X_t^i(j)$ , built from the <i>temperatures</i> recorded at the entrance $(j = 3)$ and at the exit $(j = 4)$ . On the left we inform the intervals in which the values of $d_s$ are included. The strings are listed according to the $d_s$ values, in an increasing order from left to right and from top to bottom	74
Table 29 –	Transition probabilities from the string $s$ to 0 and 1, assigned by the two processes, given by columns 1 and 2, respectively, $j = 1$ records the	14
Table 30 –	results for the <i>alcoholic content</i> , $j = 2$ indicates the results for the <i>fill level</i> . Transition probabilities from the string s to 0 and 1, assigned by the two processes, given by columns 1 and 2, respectively, $j = 3$ records the results for the <i>temperature of entrance</i> , $j = 4$ indicates the results for	75
	the temperature of exit.	76
Table 31 –	Four financial series with sample size $n=1446$ . Period: 3 January 2012- 3	
	November 2017	81
Table 32 –	$dmax_{j,k}$ (up) and $dmean_{j,k}$ (down) values for $j \neq k, j, k \in \{1, 2, 3, 4\}$ .	0.0
<b>T</b> .11. 99	Order $M = 5$ . See equations 4.1 and 4.2.	82
Table $33 -$	For the processes 1, 2 and 3, from left to right we list the transition probabilities from $a^*$ = 00100 to a 5.4 Order M = 5	09
Table 34 –	probabilities from $s^{*} = 00100$ to $a \in \mathcal{A}$ . Order $M = 5$	83
	In bold letter the values registered in table 32	83
Table 35 –	Temporal period associated with the subsample for $j = 1, 2, 3, 4$	84
Table 36 –	Results of equation (4.4), in bold letter the minimum values which select $x_{(1),1}^{j,n^*}$ , for $j = 1, 2, 3, 4, \ldots, \ldots$	84
Table 37 –	Sample selection following procedure 4.2.5, for series $j = 1, 2, 3, 4, \ldots$	85
Table 38 –	$dmax_{j,k}$ (up) and $dmean_{j,k}$ (down) between the samples $x_{(1),1}^{j,n^*}$ and $x_{(1),1}^{k,n^*}$	
	for $j \neq k; j, k \in \{1, 2, 3, 4\}$ . Order $M = 5$	87
Table 39 –	$dmax_{j,k}$ (up) and $dmean_{j,k}$ (down) between the samples $x_{(1),1}^{j,n^*}, x_{(2),1}^{j,n^*}$ and	
	$x_{(1),1}^{\kappa,n}, x_{(2),1}^{\kappa,n}$ for $j \neq k; j, k \in \{1, 2, 3, 4\}$ . Order $M = 5$	87
'Table 40 –	$dmax_{j,k}$ (up) and $dmean_{j,k}$ (down) between the samples $x_{(1),1}^{j,n}, x_{(2),1}^{j,n}, x_{(3),1}^{j,n}$ and $x_{(1),1}^{k,n^*}, x_{(2),1}^{k,n^*}, x_{(3),1}^{k,n^*}$ for $j \neq k; j, k \in \{1, 2, 3, 4\}$ . Order $M = 5$	88

- Table 42 Details about the discrepancy related in table 39, between the processes 3 and 4,  $d_s = 1.10101, s = 01100...$  89
- Table 44 Details about the discrepancies related in table 41,  $s^* = 00100$ . . . . . 90

# List of abbreviations and acronyms

- VLMC Variable Length Markov Chain
- PMM Partition Markov models
- BIC Bayesian Information Criterion
- ML Maximum Likelihood
- i.i.d Independent and identically distributed

# List of symbols

- $\mathcal{A}$  Finite set of alphabet
- $|\mathcal{A}|$  Cardinal of the alphabet
- $\mathcal{T}$  The set of context Tree
- $\mathcal{I}$  The family of irreducible trees

 $N_n(s, a)$  The number of occurrences of the string s followed by the symbol a

 $N_n(s)$  The number of occurrences of s

- D(PQ) The relative entropy between two laws Pand Q
- $d_s(x_{1,1}^{n_1},x_{2,1}^{n_2})$   $\ \ \,$  The distance between two stochastic processes

# Contents

1	INTRODUCTION	20
2	MARKOVIAN MODELS	23
2.1	Introduction	23
2.2	Context Tree	24
2.2.1	Consistent Estimation of a Context Tree via BIC	27
2.3	Partition Markov Model	30
2.3.1	Partition Markov Model Estimation via BIC	32
2.4	Measures between Markovian Processes and Related Structures	34
2.4.1	Relative Entropy	34
2.4.2	Measure for Partition Markov Model	37
3	NEW MEASURE BETWEEN STOCHASTIC PROCESSES	39
3.1	Comparison of Stochastic Processes	39
3.2	Results: The Measure and its Properties	40
3.3	First Case Study: Linguistic Data	48
3.3.1	The Most Variable Configurations	50
3.3.2	From the 16th Century to the Beginning of the 17th Century	55
3.3.3	Conclusion	58
3.4	Second Case Study: Comparison between DNA Strains	58
3.4.1	DNA Data	59
3.4.2	Results	59
3.4.3	Conclusion	68
3.5	Third Case Study: Production of Alcohol Fuel	69
3.5.1	Conclusion	77
4	ROBUST SAMPLE SELECTION STRATEGY	78
4.1	Sample Selection	78
4.2	Results	78
4.3	Case Study: Daily Trading Volume	81
4.3.1	Proximity between the Full Series	82
4.3.2	Selecting Subsamples from the Robust Procedure	83
4.3.3	Re-evaluating the Proximity between the Series	87
4.3.4	Conclusion	90
5	FINAL CONCLUSION	92

	BIBLIOGRAPHY	. 95
	APPENDIX	97
A.1	APPENDIXA – PROOFSProofs of Theorems of Chapter 1	. 98 . 98
B.1	APPENDIX       B – GRAPHS         Graphs of time series for financial study in chapter 4	. 102 . 102

## 1 Introduction

We begin this thesis by introducing the problems. Given two samples of Markov processes, we want to decide if the two samples follow the same stochastic law. If we arrive to the conclusion that the stochastic laws are different, it will be relevant to know in which way those samples are different. That is, it will be necessary to identify the states in the space state in which the transition probabilities estimated through the samples are different or maybe it will be necessary to know for which states the discrepancy between the samples is larger. This issue can be formulated as a model selection problem. The decision about if the two samples are coming from the same stochastic law can be formulated as follows: For a fixed  $s \in \mathcal{S}$ , where  $\mathcal{S}$  is the state space and s is a string, when we estimate the transition probabilities (conditioned to s) by mean of the two samples, should we estimate only one set of transition probabilities for s which will be used for both samples? or should we estimate two different sets of transition probabilities, one set for each sample? We will formulate those questions from the Bayesian Information Criterion point of view. The Bayesian Information Criterion consists of two terms. The first one is the logarithm of the maximum likelihood, which measures the goodness of fit of the sample to the model class. The second term is the half number of free parameters in the model class times the logarithm of the sample size. This second term penalizes too complex models.

In several works in the area, criteria have been proposed to achieve this goal, see for example Garcia e González-López (2013), García, González-López e Andrade (2017) among others. In this thesis, we study the theoretical properties of one of these criteria and show how it operates. This criterion was experimentally used in García, Gholizadeh e González-López (2017b) and it is built through the Bayesian Information Criterion (BIC), introduced in Schwarz et al. (1978). The BIC was used to obtain consistent estimates in several Markov models, as is the case of Context Tree estimation (CSISZÁR; TALATA, 2006) and Partition Markov model estimation (GARCÍA; GONZÁLEZ-LÓPEZ, 2017).

The measure was used to investigate several real problems, by means of those situations we investigate the performance of the distance  $d_s$  that, when evaluated in a given string s, allows us to define how far or near the processes are.

In the first application we inspect written texts of Portuguese dated between 16th century and 18th century using this tool  $(d_s)$ . We identify the most voluble structures throughout the period and also we identify robust linguistic compositions that should be considered when studying the linguistic changes from Classical Portuguese to Modern Portuguese.

In the second application we explore the performance of  $d_s$  in order to establish a

notion of natural proximity between DNA sequences from patients with identical diagnosis, which is: Burkitt lymphoma/leukemia. And we present a robust strategy of estimation to identify the law that governs most of the sequences considered, thus mapping out a common profile to all these patients, via their DNA sequences.

In the third application, we compare lines of production of alcohol fuel. The database considers several variables arising from the production of fuel, based on sugar cane. Fuel based on sugar cane, corresponds to 18% of the Brazilian energy consumption. For a complete review of the historical and political context governing these productions in the Brazilian market, we cite Goldemberg (2008) and Cortez et al. (2014). Regarding the importance of production processes based on sugar cane, we cite two references that show the impacts of these productions on the health of the population: Arbex et al. (2007) and Cançado et al. (2006).

Chapter 4 is devoted to propose and explore a selection procedure which applied in a collection of samples from Markov processes over a finite alphabet, with finite order, selects the more representative samples of the full collection.

Collet et al. (2008) showed that a small random Bernoulli perturbation in the sample from a Markovian variable memory process will effectively transform the process to an infinity memory process. They also show a variation of the original context algorithm given by Rissanen (1983) which can recover the context tree of the original chain, provided that the noise is small enough. García, González-López e Viola (2014) considered a different type of contamination model considering a set of m independent samples, with most of them coming from the same stochastic process with law Q but the remained portion of the sample following different laws, and also they defined he asymptotic breakdown point  $\gamma$  for a model selection procedure under the assumption of context tree structures.

We applied the robust selection procedure in 4 series related to brazilian stocks traded at B3 S.A. – Brasil, Bolsa, Balcão. We identify groups of series with similar traded volume patterns which support for determining potential profits or risk scenarios. Also the method let us to map similarities and dissimilarities of the daily trading volume dynamics between the series.

Finally, we present the organization of this thesis:

- In Chapter 2, we define the Markovian structures which are consistently estimated through the Bayesian Information Criterion, see Csiszár e Talata (2006) and García e González-López (2017) for the estimation of context trees and partition Markov model, respectively, via the information criterion BIC, see Csiszár, Shields et al. (2000) and Schwarz et al. (1978).
- In Chapter 3, we propose a consistent measure for the comparison between Markovian

processes, we also explore its theoretical properties and we apply this concept to several real problems as historical linguistics, Burkitt lymphoma/leukemia DNA and alcohol fuel production.

- In Chapter 4, we introduce a robust sample selection procedure, using the measure presented in the chapter 3. Also in this chapter we investigate its theoretical properties. To make this procedure more clear, we use this concept for the selection of subsamples of series coming from the financial sector of the Brazilian market.
- In Chapter 5, we report the final conclusions obtained in chapters 3 and 4.

# 2 Markovian Models

### 2.1 Introduction

Context tree models describe processes where each "past" has a suffix which is enough to determine the transition probability to the next symbol in the alphabet. Context trees are generalizations of Markov chains and were introduced by Rissanen (1983) as an efficient tool for data compression. The context trees have been studied and used in the modeling of several practical problems, such as the analysis of linguistic data Galves et al. (2012), protein classification Leonardi (2007) and identification of genes Bühlmann, Wyner et al. (1999).

Several researchers have studied various aspects related to context trees. Bühlmann, Wyner et al. (1999) proved properties of the Context estimator allowing the model to grow with the sample size. They also studied a bootstrap scheme based on fitted Variable Length Markov Chains. Ferrari e Wyner (2003) consider processes with infinite dependence for which there exist "good" context tree approximations. They established new results on a sieve methodology based on an adaptation of the context algorithm. Galves e Leonardi (2008) investigated an exponential upper limit for the convergence rate of the Context algorithm when the tree is not necessarily limited. Collet et al. (2008) studied the retrieval of the tree from contexts of an unlimited variable memory Markov chain from a sample with the noise is small enough.

Recently, García e González-López (2017) introduce a new class of finite order Markov chain models, called Partition Markov models (PMM). It characterizes the process by a partition  $\mathcal{L}$ , of the state space, where the elements in each part of  $\mathcal{L}$  share the same transition probability to an arbitrary element in the alphabet. This class of models includes the full Markov chains and the Variable Length Markov Chain models because a context tree can be seen as a particular case of a PMM. Under the assumption of this new family, they addressed the problem of model selection, showing that the model can be selected consistently using the Bayesian Information Criterion (BIC).

The BIC proposed by Schwarz et al. (1978) was used to obtain consistent estimates in several Markov models, as is the case of Context Tree estimation Csiszár e Talata (2006) and Partition Markov model estimation García e González-López (2017). Therefor, in this chapter we will cover the main results for the estimation of context trees and partition Markov models via the BIC criterion, see Csiszár, Shields et al. (2000) and Schwarz et al. (1978).

The process of estimation of full Markov chains is hard because it involves a

number of parameters  $|\mathcal{A}|^{M}(|\mathcal{A}| - 1)$ , which grows exponentially with the order M, where  $|\mathcal{A}|$  denotes the cardinal of the alphabet  $\mathcal{A}$ . For example, if  $|\mathcal{A}| = 5$  and M = 5, the number of parameters is 12500, which is prohibitive. Another limitation is that the class of full Markov chains is not very rich, since, fixed the alphabet with cardinal  $|\mathcal{A}|$  there is just one model for each order M and in practical situations it could be necessary a more flexible structure. A richer class of finite order Markov models introduced by Rissanen (1983) and Bühlmann, Wyner et al. (1999) is composed by the Variable Length Markov Chains (VLMC). In the VLMC class, each model is identified by a suffix tree  $\mathcal{T}$  called context tree. For a given model with a context tree  $\mathcal{T}$ , the total number of parameters is  $|\mathcal{T}|(|\mathcal{A}| - 1)$ .

Here is a description of the issues addressed in this chapter, section by section. In Section 2.2, we describe some concepts related to the suffix tree Csiszár e Talata (2006) and its association with a variable memory Markov process Bühlmann, Wyner et al. (1999), Csiszár e Talata (2006) and also we present the consistent estimation of a context tree proposed by Csiszár e Talata (2006). In Section 2.3, we introduce the concept of Markov chain with partition  $\mathcal{L}$  and we show that the optimal partition can be obtained through the BIC criterion, eventually almost surely, when the sample size tends to infinity. In Section 2.4, we explore some notions of proximity between Markovian Processes. First the concept of relative entropy between two stochastic processes is defined and second we introduce a measure to quantify the distance between the parts of a partition García e González-López (2017).

### 2.2 Context Tree

Let  $\mathcal{A}$  be a finite set called alphabet and  $|\mathcal{A}|$  its cardinality, the set  $\mathcal{A}^*$  is formed by all the finite sequences of elements in  $\mathcal{A}$ .

A string  $s = a_m a_{m+1} \dots a_n$  of symbols in  $\mathcal{A}$  will be denoted by  $a_m^n$  with length l(s) = n - m + 1. This notation is also valid for  $m = -\infty$  in which case we obtain a left-infinite sequence  $a_{-\infty}^n$ . The empty string is denoted by  $\phi$  with  $l(\phi) = 0$ .

A string v is a suffix of a string s, denoted by  $s \ge v$ , when there exists a string u such that s = uv. For a proper suffix, that is, when  $s \ne v$ , we write s > v.

**Remark 2.2.1.** It should be noted that the employed structure  $a_m^n$  in the current thesis is exactly the same used by Csiszár e Talata (2006), while in Bühlmann, Wyner et al. (1999) the notation is partially different, for instance  $a_m^n$  is built as  $a_n a_{n-1} \dots a_m$ .

**Definition 2.2.2.** A set  $\mathcal{T} \subset \mathcal{A}^*$  is called an irreducible tree if it satisfies the following rules:

- i). Suffix property: No sequence  $s_1 \in \mathcal{T}$  is a suffix of any other sequence  $s_2 \in \mathcal{T}$ ;
- ii). Irreducibility: No sequence  $s_1 \in \mathcal{T}$  can be replaced by a proper suffix without violating rule *i*.

The family of irreducible trees will be denoted by  $\mathcal{I}$ .

A tree  $\mathcal{T}$  is any set of strings, called leaves, such that no  $w \in \mathcal{T}$  is a proper suffix of any other  $s \in \mathcal{T}$ . This property enables us to represent the set  $\mathcal{T}$  as a graphical rooted tree by identifying the elements in  $\mathcal{T}$  with paths from the terminal nodes of the tree to the root. As an example of finite tree, consider the set  $\mathcal{T}_1 = \{00, 010, 110, 1\}$  over the alphabet  $\mathcal{A} = \{0, 1\}$ . On the other hand, an example of an infinite tree over  $\mathcal{A}$  is given by  $\mathcal{T}_2 = \{10_1^i : i = 0, 1...\} \cup \{0^\infty\}$  which has a unique infinite element, the left-infinite  $0^\infty$ . The graphical representation of these trees can be found in Fig 1, both irreducible trees. An example of a non-irreducible tree is  $\mathcal{T}_3 = \{000, 010, 110, 1\}$ , because substituting 000 by 00 leads to  $\mathcal{T}_1$  that satisfies the property i) of definition 2.2.2



Figure 1 – Graphical representation: on the left the tree  $\mathcal{T}_1$  and on the right the tree  $\mathcal{T}_2$ . In both cases, the contexts correspond to the sequences obtained by concatenating the symbols from the leaves to the root of the trees.

The depth of the tree  $\mathcal{T}$  denoted by  $d(\mathcal{T})$ , is given by

$$d(\mathcal{T}) = \max\left\{l(s), \ s \in \mathcal{T}\right\}$$

and the tree  $\mathcal{T}$  truncated at level K will be denoted by  $\mathcal{T}|_{K}$ , it satisfies

$$\mathcal{T}|_{K} = \{s' : s' \in \mathcal{T} \text{ with } l(s') \leq K \text{ or } s' \text{ is a suffix of length } \lfloor K \rfloor \text{ of some } s \in \mathcal{T} \}.$$

Let  $(X_t)_{t\in Z}$  be a stationary and ergodic process assuming values in the finite alphabet  $\mathcal{A}$ . We denote by  $P(a_m^n)$  the stationary probability of the string  $a_m^n$ , that is

$$P(a_m^n) = Prob(X_m^n = a_m^n).$$

If  $s \in \mathcal{A}^*$  is such that P(s) > 0 we write

$$P(a|s) = \operatorname{Prob}(X_0 = a|X_{-k}^{-1} = s)$$

**Definition 2.2.3.** A string  $s \in \mathcal{A}^*$  is a context for a process P if P(s) > 0 and

$$Prob(X_0 = a | X_{-\infty}^{-1} = x_{-\infty}^{-1}) = P(a|s), \ \forall a \in \mathcal{A}$$

whenever s is a suffix of the semi-infinite sequence  $x_{-\infty}^{-1}$ , and no proper suffix of s has this property.

By this definition, the set of contexts of a process  $(X_t)_{t \in Z}$  with measure P is an irreducible tree, it will be denoted by  $\mathcal{T}_P$ . The set of contexts associated with the process  $(X_t)_{t \in Z}$  is called the probabilistic tree of contexts.

**Definition 2.2.4.** A probabilistic context tree is a pair  $(\mathcal{T}, p)$  such that:

- i).  $\mathcal{T}$  is an irreducible tree;
- *ii).*  $p = \{P(\cdot|s) : s \in \mathcal{T}\}$  is a family of transition probabilities over  $\mathcal{A}$

**Example 2.2.5.** Consider the stationary Markov chain of order 3 over the alphabet  $\mathcal{A} = \{0, 1\}$  defined by the transition probabilities

s	P(0 s)	P(1 s)
ab1	0.2	0.8
a00	0.5	0.5
010	0.3	0.7
110	0.7	0.3

Table 1 – Table of transition probabilities with alphabet  $\mathcal{A} = \{0, 1\}$ .

where  $a, b \in \mathcal{A}$  are arbitrary. By Definition 2.2.3, the only contexts of this process are the strings 1, 00, 010 and 110. The context tree  $\mathcal{T}_P$  is the tree  $\mathcal{T}_1$  represented in Fig. 1

In a context tree, the probability of a sequence can be obtained by the probability of contexts and transition probability. For example, in example 2.2.5, the contexts are 1, 00, 010 and 110, then the probability  $\operatorname{Prob}(X_{-4}^0 = 01000)$  is given by

$$Prob(X_0 = 0 | X_{-4} = 0, X_{-3} = 1, X_{-2} = 0, X_{-1} = 0) Prob(X_{-4} = 0, X_{-3} = 1, X_{-2} = 0, X_{-1} = 0)$$

On the other hand  $Prob(X_{-4} = 0, X_{-3} = 1, X_{-2} = 0, X_{-1} = 0)$  is equals

$$Prob\left(X_{-1}=0|X_{-4}=0,X_{-3}=1,X_{-2}=0\right)Prob\left(X_{-4}=0,X_{-3}=1,X_{-2}=0\right)$$

Since a sequence 00 is one of the contexts, we have

$$Prob\left(X_{0}=0|X_{-4}=0,X_{-3}=1,X_{-2}=0,X_{-1}=0\right)=Prob\left(X_{0}=0|X_{-2}=0,X_{-1}=0\right)$$

therefor the probability  $Prob(X_{-4} = 0, X_{-3} = 1, X_{-2} = 0, X_{-1} = 0, X_0 = 0)$  is given by

$$Prob (X_0 = 0 | X_{-2} = 0, X_{-1} = 0) \times Prob (X_{-1} = 0 | X_{-4} = 0, X_{-3} = 1, X_{-2} = 0)$$
$$\times Prob (X_{-4} = 0, X_{-3} = 1, X_{-2} = 0)$$

whereas the sequence 010 is one of the contexts.

#### 2.2.1 Consistent Estimation of a Context Tree via BIC

The purpose of this section is to describe the estimation process of a context tree  $\mathcal{T}_0$  and its transition probabilities from their contexts. Consider a sample  $x_1^n$ , which is a realization of the process  $(X_t)_{t \in \mathbb{Z}}$ .

Let  $N_n(s, a)$  denote the number of occurrences of the string  $s \in \mathcal{A}^{l(s)}$  followed by the symbol  $a \in \mathcal{A}$  in the sample  $x_1^n$ , where s is supposed to be of length at most D(n). As  $N_n(s, a)$  denote the number of occurrences of the string  $s \in \mathcal{A}^{l(s)}$  followed by the symbol  $a \in \mathcal{A}$ , only the cases with i > D(n), are considered, thus  $N_n(s, a)$  is given by

$$N_n(s,a) = \left| \left\{ i : D(n) < i \le n; \ x_{i-l(s)}^{i-1} = s, x_i = a \right\} \right|$$

and the number of occurrences of s is given by

$$N_n(s) = \left| \left\{ i : D(n) < i \le n; \ x_{i-l(s)}^{i-1} = s \right\} \right|$$

Given a sample  $x_1^n$ , a feasible tree is any tree  $\mathcal{T}$  of depth  $d(\mathcal{T}) \leq D(n)$  such that  $N_n(s) \geq 1$ for all  $s \in \mathcal{T}$ , and each string s' with  $N_n(s') \geq 1$  is either a suffix of some  $s \in \mathcal{T}$  or has a suffix  $s \in \mathcal{T}$ . A feasible tree  $\mathcal{T}$  is called *r*-frequent if  $N_n(s) \geq r$  for all  $s \in \mathcal{T}$ . The family of all feasible, *r*-frequent trees is denoted by  $\mathcal{F}_r(x_1^n, D(n))$ , so for r = 1, we have  $\mathcal{F}_1(x_1^n, D(n))$ .

Clearly, 
$$\sum_{a \in \mathcal{A}} N_n(s, a) = N_n(s)$$
 and  $\sum_{s \in \mathcal{T}} N_n(s) = n - D(n)$ , for any feasible tree  $\mathcal{T}$ .

Considering the tree  $\mathcal{T}$  as a context tree associated with a law process P, the probability of the sample  $x_1^n$  can be written as

$$P(x_1^n) = P\left(x_1^{D(n)}\right) \prod_{s \in \mathcal{T}, a \in \mathcal{A}} P\left(a|s\right)^{N_n(s,a)}$$
(2.1)

For context trees in  $\mathcal{F}_1(x_1^n, D(n))$ , we define the maximum likelihood of equation (2.1), denoted by  $ML(x_1^n, \mathcal{T})$  as the maximum of the second factor in equation (2.1), given by  $\prod_{s \in \mathcal{T}, a \in \mathcal{A}} \hat{P}(a|s)^{N_n(s,a)}$ , where  $\hat{P}(a|s)$ ,  $\forall a \in \mathcal{A}, s \in \mathcal{T}$  are the estimators of P(a|s) $\forall a \in \mathcal{A}, s \in \mathcal{T}$ , that maximize the function  $\prod_{s \in \mathcal{T}, a \in \mathcal{A}} P(a|s)^{N_n(s,a)}$  subject to the restriction  $\sum_{a \in \mathcal{A}} P(a|s) = 1$  for each  $s \in \mathcal{T}$ . The maximum likelihood  $ML(x_1^n, \mathcal{T})$  is given by the expression

$$ML(x_1^n, \mathcal{T}) = \prod_{s \in \mathcal{T}, N_n(s) \ge 1} \prod_{a \in \mathcal{A}} \left( \frac{N_n(s, a)}{N_n(s)} \right)^{N_n(s, a)}$$
(2.2)

which can be factored as

$$ML(x_1^n, \mathcal{T}) = \prod_{s \in \mathcal{T}} \widetilde{P}_s(x_1^n)$$

where

$$\widetilde{P}_{s}(x_{1}^{n}) = \begin{cases} \prod_{a \in \mathcal{A}} \left( \frac{N_{n}(s,a)}{N_{n}(s)} \right)^{N_{n}(s,a)} & \text{if } N_{n}(s) \ge 1; \\ \\ 1 & \text{if } N_{n}(s) = 0 \end{cases}$$

For the estimation of the tree  $\mathcal{T}_0$  is used the criterion BIC that assigns a score for each hypothetical model (in this case, the models are context trees), the estimator will be that model with maximal score. One of the most relevant criterion used in model selection of stochastic processes is the Bayesian Information Criterion (BIC). Schwarz et al. (1978) derived an information criterion by means of an asymptotic approximation of the maximum posterior Probability estimator. Csiszár e Talata (2006) have used this criterion to obtain a consistent estimator of a context tree. The BIC consists of two terms:

- 1). The logarithm of the maximum likelihood, which measures the goodness of fit of the sample to the model class,
- 2). the half number of the free parameters in the model class times the logarithm of the sample size. This term penalizes complex models.

Consider a density functions belonging to the exponential family

$$f_{\theta}(x_i) = exp\left(\theta y(x_i) - b(\theta)\right); \ \theta \in \Theta$$

In the natural parametric  $\Theta$  which is a convex subset of the K-dimensional Euclidean space, y is a sufficient K-dimensional statistic and

$$b(\theta) = \log \int exp(\theta y(x_i)) dx_i$$

Consider a prior distribution of the parameter vector written in the form  $\mu = \sum \alpha_j \mu_j$ , where  $\alpha_j$  is the priori probability of the *j*th model being the true one, and  $\mu_j$  is the conditional priori distribution of  $\theta$  given the *j*th model with  $\theta \in \Theta$ . Schwarz et al. (1978) shows that, under regularity conditions, maximum posteriori Probability estimator of the parameter vector  $\theta$  from an i.i.d. sample  $x_1, \ldots, x_n$  asymptotically does not depend on  $\mu$ , and is equivalent to the maximum likelihood estimator  $\hat{\theta} = \arg \max_{\theta \in \Theta} f_{\theta}(x)$ . The BIC chooses the *j* that maximizes

$$S(\mathbf{Y}, n, j) = \log \int_{m_j \cap \Theta} \alpha_j \exp\left(\left(\theta Y - b(\theta)\right)n\right) d\alpha_j(\theta)$$

where  $\mathbf{Y} = \frac{1}{n} \sum_{i} y(x_i)$  and  $m_j$  is a dimension of the *j*-th model.

Considering an asymptotic treatment for  $S(\mathbf{Y}, n, j)$  when n goes to infinity, then the BIC is defined as follows:

**Definition 2.2.6.** Given the maximum likelihood function  $L(x_1^n, K_j)$  corresponding to a model  $K_j$  and a sample  $x_1^n$ , the BIC is defined by

$$BIC(x_1^n, K) = \ln (L(x_1^n, K_j)) - \frac{M_j}{2} \ln(n)$$

where  $K = \{K_j\}_j$  and  $M_j$  is the total number of parameters to be estimated for the model  $K_j$ . In the scope of context trees,  $M_j = (|\mathcal{A}| - 1)|\mathcal{T}|$ , so the BIC definition for context trees estimation is:

**Definition 2.2.7.** Given a sample  $x_1^n$ , the BIC for a feasible tree  $\mathcal{T}$  is

$$BIC_{\mathcal{T}}(x_1^n) = \sum_{s \in \mathcal{T}, a \in \mathcal{A}} N_n(s, a) \ln\left(\frac{N_n(s, a)}{N_n(s)}\right) - \frac{\left(|\mathcal{A}| - 1\right) |\mathcal{T}|}{2} \ln(n).$$

The estimator for  $\mathcal{T}$ , denoted by  $\widehat{\mathcal{T}}_{BIC}(x_1^n)$ , is defined as

$$\widehat{\mathcal{T}}_{BIC}(x_1^n) = \arg\min_{\mathcal{F}\in\mathcal{F}_1(x_1^n, D(n))\cap\mathcal{I}} BIC_{\mathcal{T}}(x_1^n)$$

This estimator is a consistent estimator, that in the case  $D(\mathcal{T}_0) < \infty$ , consistency means that estimated context tree is equal to the tree  $\mathcal{T}_0$ , almost certainly, so  $\exists n_0$  such that the estimated context tree is equal to the tree  $\mathcal{T}_0, \forall n \ge n_0$ .

**Theorem 2.2.8.** Csiszár e Talata (2006) If  $d(\mathcal{T}_0) < \infty$ , the estimator  $\widehat{\mathcal{T}}_{BIC}$ , with  $D(n) = o(\log n)$ , satisfies  $\widehat{\mathcal{T}}_{BIC}(X_1^n) = \mathcal{T}_0$  almost certainly, when  $n \longrightarrow \infty$ .

**Proof.** See Section A.1 of Appendix A.

In order to show that the Kullback Leibler concept can be adapted to the structure of context trees, we introduce the next definition. García, González-López e Viola (2014)

introduced a structure of tree  $\mathcal{T}_{PQ}$  and they demonstrated that the relative entropy between two processes can be expressed, through its conditional relative entropies. For this purpose, first it is necessary to define the structure of tree  $\mathcal{T}_{PQ}$ , which is given as follows.

**Definition 2.2.9.** Let  $\mathcal{T}_P$  and  $\mathcal{T}_Q$  be two context trees with probability law P and Q, respectively. The tree  $\mathcal{T}_{PQ}$ , resulting from the concatenation of  $\mathcal{T}_P$  and  $\mathcal{T}_Q$ , is defined as

$$\mathcal{T}_{PQ} = \{ s \in \mathcal{T}_P \cup \mathcal{T}_Q : \ \nexists s' \in \mathcal{T}_P \cup \mathcal{T}_Q, \ suffix \ of \ s \}$$

In Figure 2 we show an example of the concatenation of  $\mathcal{T}_P$  and  $\mathcal{T}_Q$  trees resulting in the context tree  $\mathcal{T}_{PQ}$ .



Figure 2 – Left: context tree  $\mathcal{T}_P$ . Middle: context tree  $\mathcal{T}_Q$ . Right: context tree  $\mathcal{T}_{PQ}$ .

#### 2.3 Partition Markov Model

In this section, we introduce the concept of Markov chain with partition  $\mathcal{L}$ , which is a partition of the state space defined through a stochastic equivalence relationship between strings of the state space as shown below.

Consider  $(X_t)_{t\in\mathbb{Z}}$  be a discrete time order M stationery Markov chain on a finite alphabet  $\mathcal{A}$  with  $M < \infty$ . Let us call  $\mathcal{S} = \mathcal{A}^M$  the state space. Denote the string  $a_m a_{m+1} \dots a_n$  by  $a_m^n$ , where  $a_i \in \mathcal{A}$ ,  $m \leq i \leq n$ . For each  $a_i \in \mathcal{A}$  and  $s \in \mathcal{S}$ ,  $P(a|s) = Prob(X_t = a|X_{t-M}^{t-1} = s)$ . Let  $\mathcal{L} = \{L_1, L_2, \dots, L_{|\mathcal{L}|}\}$  be a partition of  $\mathcal{S}$ , for  $a_i \in \mathcal{A}$ ,  $L \in \mathcal{L}$ ,  $P(L, a) = \sum_{s\in L} Prob(X_{t-M}^{t-1} = s)$  and  $P(a|L) = \frac{P(L, a)}{P(L)}$  with P(L) > 0.

The following definition, which was proposed by García e González-López (2017), defines a Markov chain with a "minimal partition" based on the equivalent relationship on S.

**Definition 2.3.1.** Let  $(X_t)_{t \in \mathbb{Z}}$  be a discrete time order M stationery Markov chain on a finite alphabet  $\mathcal{A}$ ; with state space  $\mathcal{S} = \mathcal{A}^M$ ,

*i.*  $s, r \in S$  are equivalent (denoted by  $s \sim r$ ) if  $P(a|s) = P(a|r) \forall a \in A$ .

ii.  $(X_t)_{t\in\mathbb{Z}}$  is a Markov chain with partition  $\mathcal{L} = \{L_1, L_2, \cdots, L_{|\mathcal{L}|}\}$  if this partition is the one defined by the equivalence relationship introduced by item *i*.

The set of parameters for a Markov chain over the alphabet  $\mathcal{A} = \{a_1, a_2, \cdots a_{|\mathcal{A}|}\}$  with partition  $\mathcal{L} = \{L_1, L_2, \cdots, L_{|\mathcal{L}|}\}$  can be denoted by  $\{P(a_i|L_j) : 1 \leq i < |\mathcal{A}|, 1 \leq j \leq |\mathcal{L}|\}$ . This means that there are  $(|\mathcal{A}| - 1)$  transition probabilities for each part. Then, the total number of parameters for the model is  $|\mathcal{L}|(|\mathcal{A}| - 1)$ ; where  $|\mathcal{A}|$  and  $|\mathcal{L}|$  denote the cardinal of  $\mathcal{A}$  and  $\mathcal{L}$  respectively.

For a better understanding of this definition, consider the example below.

**Example 2.3.2.** Let  $(X_t)_{t\in \mathbb{Z}}$  be a discrete time order M = 2 stationary Markov chain taking values on  $\mathcal{A} = \{0, 1\}$  with state space  $\mathcal{S} = \mathcal{A}^M = \{00, 01, 10, 11\}$ . Suppose that this chain follows the transition probabilities given by

$$P(0|00) = 0.4 = P(0|01)$$
 and  $P(0|10) = 0.2 = P(0|11)$ 

the process as a full chain have 4 parameters. Then, according the definition 2.3.1-ii, the partition for this Markov chain is  $\mathcal{L} = \{\{00, 01\}, \{10, 11\}\}$ . Note that only 2 parameters are needed to describe the source, since  $P(0|\{00, 01\}) = 0.4$  and  $P(0|\{10, 11\}) = 0.2$ .

There is a link between context trees or VLMC with finite depth M and Partition Markov models, which is illustrated by the next example.

**Example 2.3.3.** Let  $(X_t)_{t\in\mathbb{Z}}$  be a finite order Markov chain taking values on  $\mathcal{A} = \{0, 1\}$ and  $\mathcal{T}$  a set of contexts. Consider  $d(\mathcal{L}) = 3$  and  $\mathcal{T} = \{\{00\}, \{10\}, \{11\}, \{001\}, \{101\}\},$ illustrated by Figure 3. Also suppose that  $P(\cdot|s)$  is the probability of  $s \in \{00, 10, 11\}$  and  $Q(\cdot|s)$  is the probability of  $s \in \{001, 101\}$ , this mean that  $P(\cdot|00) = P(\cdot|10) = P(\cdot|11)$  and the same occurs with Q. This context tree corresponds to the partition  $\mathcal{L} = \{L_1, L_2\}$  where  $L_1 = \{\{000\}, \{100\}, \{010\}, \{110\}, \{011\}, \{111\}\}$  and  $L_2 = \{\{001\}, \{101\}\}.$ 



Figure 3 – Graphical representation: context tree in the example 2.3.3

since

$$Prob(\cdot|000) = Prob(\cdot|100) = Prob(\cdot|\underbrace{00}_{context}) = P(\cdot|s)$$

$$Prob(\cdot|010) = Prob(\cdot|110) = Prob(\cdot|\underbrace{10}_{context}) = P(\cdot|s)$$

this is also true for the strings 011 and 111 with context 11., so  $\forall s, s' \in L_i$ , i = 1, 2 we have  $Prob(\cdot|s) = Prob(\cdot|s')$ .

#### 2.3.1 Partition Markov Model Estimation via BIC

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 as  $N_n(L, a) = \sum_{s \in L} N_n(s, a), \ L \in \mathcal{L}$ . The accumulated number of  $N_n(s)$  for s in L is denoted by  $N_n(L) = \sum_{s \in L} N_n(s), \ L \in \mathcal{L}$ .

Under the assumption of a hypothetical partition  $\mathcal{L}$  of  $\mathcal{S}$ , the Prob $(X_1^n = x_1^n)$  denoted by  $P(x_1^n)$  follows

$$P(x_1^n) = P(x_1^M) \prod_{L \in \mathcal{L}, a \in \mathcal{A}} P(a|L)^{N_n(L,a)}.$$
(2.3)

According to the definition of the Bayesian Information Criterion (BIC) will be necessary to maximize the second term in the equation 2.3 called maximum likelihood for a given observation  $x_1^n$ . Then the BIC definition for partition Markov models is:

**Definition 2.3.4.** Given a sample  $x_1^n$  of the process  $(X_t)_{t\in \mathbb{Z}}$ , a discrete time order M stationary Markov chain on a finite alphabet  $\mathcal{A}$  with state space  $\mathcal{S} = \mathcal{A}^M$  and  $\mathcal{L}$  a partition of  $\mathcal{S}$ . The BIC of the model given by definition 2.3.1-ii is

$$BIC_{\alpha}(\mathcal{L}, x_1^n) = \sum_{a \in \mathcal{A}, L \in \mathcal{L}} N_n(L, a) \ln\left(\frac{N_n(L, a)}{N_n(L)}\right) - \frac{(|\mathcal{A}| - 1)|\mathcal{L}|}{\alpha} \ln(n).$$

for a given positive value  $\alpha$ .

**Remark 2.3.5.** The theoretical results of this thesis show that can be used any positive values of  $\alpha$ . The consistency of our results will be remain valid when n goes to infinity, for any  $\alpha$ . As was done by Schwarz et al. (1978) and Csiszár e Talata (2006), in the applications we use by simplicity  $\alpha = 2$ . In this thesis we didn't explore different values of alpha since we assume that n is big enough. We note that any possible difference can be detected by simulations for moderate sample sizes, but that is not the focus of this thesis.

By means of definition 2.3.4 it is possible to, archive the minimal partition of S, by maximizing the BIC. In order to show how this criterion works, we introduce below some concepts.

**Definition 2.3.6.** Let  $\mathcal{L} = \{L_1, L_2, \cdots, L_{|\mathcal{L}|}\}$  be a partition of  $\mathcal{S}$ .

- (i)  $L \in \mathcal{L}$  is a good part of  $\mathcal{L}$  if  $\forall s, s' \in L$ ,  $P(X_t = .|X_{t-M}^{t-1} = s) = P(X_t = .|X_{t-M}^{t-1} = s');$
- (ii)  $\mathcal{L}$  is a good partition of  $\mathcal{S}$  if for each  $i \in \{1, \ldots, |\mathcal{L}|\}$ ,  $L_i$  verifies item (i).

**Theorem 2.3.7.** (GARCÍA; GHOLIZADEH; GONZÁLEZ-LÓPEZ, 2017a) Given a sample  $x_1^n$  of the process  $(X_t)_{t\in Z}$ , a discrete time order M stationary Markov chain on a finite alphabet  $\mathcal{A}$ ; with state space  $\mathcal{S} = \mathcal{A}^M$ :

*i.* According to definition 2.3.6 (i), suppose that i and j exist, and  $i \neq j$  such that  $L_i$ and  $L_j$  are good parts. Then  $P(a|L_i) = P(a|L_j)$ ,  $\forall a \in \mathcal{A}$  if, and only if, eventually almost surely as  $n \to \infty$ ,

$$BIC_{\alpha}(\mathcal{L}^{ij}, x_1^n) > BIC_{\alpha}(\mathcal{L}, x_1^n)$$

ii. Let  $\mathcal{P}$  be the set of all the partitions of  $\mathcal{S}$ . Define

$$\mathcal{L}_n = argmax_{\mathcal{L}\in\mathcal{P}} \left\{ BIC_\alpha(\mathcal{L}, x_1^n) \right\}$$

Then, eventually almost surely as  $n \to \infty$ ,  $\mathcal{L}^* = \mathcal{L}_n$ , where  $\mathcal{L}^*$  is the minimal partition of S.

where  $\mathcal{L}^{ij} = \{L_1, \dots, L_{i-1}, L_{ij}, L_{i+1}, \dots, L_{j-1}, L_{j+1}, \dots, L_{|\mathcal{L}|}\}$ , with  $L_{ij} = L_i \cup L_j$ . And, for each  $a \in \mathcal{A}$ , we set  $N_n(L_{ij}, a) = N_n(L_i, a) + N_n(L_j, a)$  and  $N_n(L_{ij}) = N_n(L_i) + N_n(L_j)$ .

**Example 2.3.8.** Suppose  $\mathcal{A} = \{0, 1\}$  with M = 3 and consider contexts tree as following

$$\mathcal{T}_1 = \{\{0\}, \{01\}, \{011\}, \{111\}\}$$

The good partition corresponding with the these context tree is given by

$$\mathcal{L} = \{L_1, L_2, L_3, L_4\}$$

where  $L_1 = \{\{000\}, \{100\}, \{010\}, \{110\}\}, L_2 = \{\{001\}, \{101\}\}, L_3 = \{011\} and L_4 = \{111\}.$ 

Suppose  $P(\cdot|s) \neq P(\cdot|s')$ ,  $\forall s, s' \in \mathcal{T} \setminus \{111, 011\}$  and  $P(\cdot|011) = P(\cdot|111)$ . Define  $L' = L_3 \cup L_4$  and  $\mathcal{L}' = \{L_1, L_2, L'\}$  is also good partition and it is also minimal, according to definition 2.3.1.

### 2.4 Measures between Markovian Processes and Related Structures

#### 2.4.1 Relative Entropy

In this section, we introduce the concepts of relative entropy between two random variables of the family of VLMC or Context Trees models, and also we define the relative entropy rate between two stochastic processes.

**Definition 2.4.1.** Let X, Y be two discrete random variables over the finite set (alphabet)  $\mathcal{A}$ , with probability mass functions  $P(\cdot)$  and  $Q(\cdot)$ , respectively. Then the relative entropy or Kullback Leibler divergence between two laws P and Q is

$$D(P||Q) = \sum_{x \in \mathcal{A}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right) = E_P\left(\ln\left(\frac{P(X)}{Q(X)}\right)\right)$$

In the above definition, using the convention (based on continuity) that  $0\ln(\frac{0}{0}) = 0$ ,  $0\ln(\frac{0}{Q}) = 0$  if P(x) = 0 and  $P\ln(\frac{P}{0}) = \infty$  if P(x) > Q(x) = 0.

**Example 2.4.2.** Let P be a Bernoulli distribution with success probability 1/2, and also Q be a Bernoulli distribution with success probability q, then according to definition 2.4.1, we have

$$D(P||Q) = \frac{1}{2}\ln\left(\frac{1}{2q}\right) + \frac{1}{2}\ln\left(\frac{1}{2(1-q)}\right) = -\frac{1}{2}\ln\left(4q(1-q)\right)$$

and when  $q \to 0$ ,  $D(P||Q) \to \infty$ ,

also

$$D(Q||P) = q\ln(2q) + (1-q)\ln(2(1-q))$$

and when  $q \to 0$ ,  $D(Q||P) \to \ln(2)$ .

So as a consequence

$$D(P||Q) \neq D(Q||P).$$

It is worth mentioning that the relative entropy is not a distance because it is not a symmetric function and does not satisfy the triangular inequality. Nonetheless, it quantifies the similarity/divergence between distributions. Now we show some relevant characteristics of the relative entropy.

**Theorem 2.4.3.** (Gibbs Inequality) Consider P(.) and Q(.) as being two probability functions defined in the same finite alphabet  $\mathcal{A}$ . Then,  $D(P||Q) \ge 0$ . The equality occurs if, and only if P(x) = Q(x),  $\forall x \in \mathcal{A}$ .

**Proof.** Consider  $\mathcal{A} = \{x : P(x) > 0\}$ , then

$$-D(P||Q) = -\sum_{x \in \mathcal{A}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right) = \sum_{x \in \mathcal{A}} P(x) \ln\left(\frac{Q(x)}{P(x)}\right)$$

Since  $\ln(t)$  is a strictly concave function at t, using the inequality of Jensen we have

$$\sum_{x \in \mathcal{A}} P(x) \ln\left(\frac{Q(x)}{P(x)}\right) \leq \ln\left(\sum_{x \in \mathcal{A}} P(x)\frac{Q(x)}{P(x)}\right).$$
(2.4)

So,

$$-D(P||Q) = \sum_{x \in \mathcal{A}} P(x) \ln\left(\frac{Q(x)}{P(x)}\right) \leq \ln\left(\sum_{x \in \mathcal{A}} P(x)\frac{Q(x)}{P(x)}\right)$$
$$= \ln\left(\sum_{x \in \mathcal{A}} Q(x)\right) = \ln(1) = 0$$

Therefore  $D(P||Q) \ge 0$ .

The equality in the equation (2.4) occurs if, and only if  $\frac{Q(x)}{P(x)} = 1$ ,  $\forall x \in \mathcal{A}$ . Therefore D(P||Q) = 0 if, and only if  $P(x) = Q(x) \ \forall x \in \mathcal{A}$ .

**Theorem 2.4.4.** (Log-Sum Inequality) Let  $a_1, a_2, \ldots$  and  $b_1, b_2, \ldots$  non-negative numbers so that  $\sum a_i < \infty$  and  $0 < \sum b_i < \infty$ . Then

$$\sum a_i \ln \frac{a_i}{b_i} \ge \left(\sum a_i\right) \ln \frac{\sum a_i}{\sum b_i}$$
(2.5)

The equality occurs if and only if  $\frac{a_i}{b_i} = c, \ \forall i.$ 

**Proof.** Let

$$a'_i = \frac{a_i}{\sum_j a_j}$$
;  $b'_i = \frac{b_i}{\sum_j b_j}$ 

Hence  $(a'_1, a'_2, \ldots)$  and  $(b'_1, b'_2, \ldots)$  are probability measures, from the Gibbs inequality, it follows

$$0 \leq D\left((a_1', a_2', \ldots)\right) ||(b_1', b_2', \ldots)) = \sum a_i' \ln\left(\frac{a_i'}{b_i'}\right) = \sum \left(\frac{a_i}{\sum_j a_j}\right) \ln\left(\frac{\frac{a_i}{\sum_j a_j}}{\frac{\sum_j a_j}{\sum_j b_j}}\right)$$
$$= \frac{1}{\sum_j a_j} \left[\sum a_i \ln\left(\frac{a_i}{b_i}\right) - \left(\sum a_i\right) \ln\left(\frac{\sum_j a_j}{\sum_j b_j}\right)\right]$$
$$\Rightarrow \sum a_i \ln\frac{a_i}{b_i} \geq \left(\sum a_i\right) \ln\frac{\sum a_i}{\sum b_i}$$

The inequality 2.5 follows. We have that  $D((a'_1, a'_2, \ldots)||(b'_1, b'_2, \ldots)) = 0$  iff  $a'_i = b'_i$ . This, implies that  $\frac{a_i}{b_i} = \frac{\sum_j a_j}{\sum_j b_j}$ ,  $\forall i$ .

**Remark 2.4.5.** Note that log-sum inequality and Gibbs inequality are equivalent.

**Theorem 2.4.6.** Consider  $P_1, P_2, Q_1, Q_2$  distributions on  $\mathcal{A}$ . Then convexity of relative entropy is given by:

$$D(\lambda P_1 + (1 - \lambda)P_2 ||\lambda Q_1 + (1 - \lambda)Q_2) \leq \lambda D(P_1 ||Q_1) + (1 - \lambda)D(P_2 ||Q_2)$$

**Proof.** Fix  $x \in \mathcal{A}$ . According to Log-sum inequality, we have:

$$\lambda P_1(x) \ln \frac{\lambda P_1(x)}{\lambda Q_1(x)} + (1-\lambda) P_2(x) \ln \frac{(1-\lambda) P_2(x)}{(1-\lambda) Q_2(x)}$$
$$\geqslant (\lambda P_1(x) + (1-\lambda) P_2(x)) \ln \left(\frac{\lambda P_1(x) + (1-\lambda) P_2(x)}{\lambda Q_1(x) + (1-\lambda) Q_2(x)}\right)$$

Sum over  $\mathcal{A}$ .

**Theorem 2.4.7.** (CSISZÁR; TALATA, 2006) Consider two distributions  $P_1$  and  $P_2$  on  $\mathcal{A}$ , then,

$$D(P_1||P_2) \leq \sum_{x \in \mathcal{A}} \frac{(P_1(x) - P_2(x))^2}{P_2(x)}$$
(2.6)

Proof.

$$D(P_1||P_2) = \sum_{x \in \mathcal{A}} P_1(x) \ln \frac{P_1(x)}{P_2(x)}$$
  
  $\leq \sum_{x \in \mathcal{A}} P_1(x) \left( \frac{P_1(x)}{P_1(x)} - \frac{1}{P_1(x)} \right) \sum_{x \in \mathcal{A}} \frac{(P_1(x) - P_2(x))}{P_1(x)}$ 

$$\leq \sum_{x \in \mathcal{A}} P_1(x) \left( \frac{P_1(x)}{P_2(x)} - 1 \right) = \sum_{x \in \mathcal{A}} \frac{(P_1(x) - P_2(x))^2}{P_2(x)}.$$

Since the relative entropy between two laws P and Q is not symmetric, a complementary concept is defined below.

**Definition 2.4.8.** The symmetrized relative entropy between two laws P and Q is defined by

$$\overline{D}(P,Q) = \frac{D(P||Q) + D(Q||P)}{2}$$

Now the concept of relative entropy is extended to stochastic processes. The rate of relative entropy is calculated for two stationary and Markovian processes defined under the same finite alphabet  $\mathcal{A}$  and with laws P and Q. The next theorem shows how to use the joint context tree structure (definition 2.2.9) to evaluate the relative entropy between the processes.
**Theorem 2.4.9.** (GARCÍA; GONZÁLEZ-LÓPEZ; VIOLA, 2014) Let  $(X_t)_{t\in \mathbb{Z}}$  and  $(Y_t)_{t\in \mathbb{Z}}$ be two stationary, ergodic and Markovian stochastic process on a finite alphabet  $\mathcal{A}$  of finite order, with probability law P and Q, respectively. Then, the relative entropy rate between them is given by

$$D(P||Q) = \sum_{s \in \mathcal{T}_{PQ}} P(s)D(P(\cdot|s)||Q(\cdot|s))$$

**Proof.** See Section A.1 of Appendix A.

**Remark 2.4.10.** For  $s \in \mathcal{T}_{PQ}$ , we observe that  $P(\cdot|s)$  is the usual probability when  $s \in \mathcal{T}_P$ . If  $s \notin \mathcal{T}_P$ ,  $\exists s_1 \in \mathcal{T}_P$  and x some string, such that  $s = xs_1$  and  $P(\cdot|s) = P(\cdot|s_1)$ .

#### 2.4.2 Measure for Partition Markov Model

In this section we introduce a distance between the parts of a partition, and this concept defines a metric on the state space and also allows to build efficient algorithms for estimating the minimal partition see definition 2.3.1 in García e González-López (2017).

**Definition 2.4.11.** Let  $(X_t)_{t\in \mathbb{Z}}$  be a stationary Markov chain of order M, with finite alphabet  $\mathcal{A}$  and state space  $\mathcal{S} = \mathcal{A}^M$ ,  $x_1^n$  a sample of the process and let  $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$  be a good partition of  $\mathcal{S}$ 

$$\begin{aligned} d_{\mathcal{L}}(i,j) &= \frac{1}{\ln(n)} \sum_{a \in \mathcal{A}} \left\{ N_n(L_i,a) \ln\left(\frac{N_n(L_i,a)}{N_n(L_i)}\right) + N_n(L_j,a) \ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right) \right. \\ &\left. - N_n(L_{ij},a) \ln\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right) \right\} \end{aligned}$$

where  $N_n(L_{ij}, a) = N_n(L_i, a) + N_n(L_j, a)$  and  $N_n(L_{ij}) = N_n(L_i) + N_n(L_j)$ . García e González-López (2017) shows that  $d_{\mathcal{L}}$  is a distance in  $\mathcal{L}$ .

**Theorem 2.4.12.** Let  $(X_t)_{t\in\mathbb{Z}}$  be a stationary Markov chain of order M over a finite alphabet  $\mathcal{A}$  and  $\mathcal{S} = \mathcal{A}^M$  the state space and  $x_1^n$  a sample of the Markov process. If  $\mathcal{L} = \{L_1, L_2, \ldots, L_{\mathcal{L}}\}$  is a good partition of  $\mathcal{S}$ , for each n, and for any  $i, j, k \in \{1, 2, \ldots, \mathcal{L}\}$ :

(i)  $d_{\mathcal{L}}(i,j) \ge 0$  with equality if and only if  $\frac{N_n(L_i,a)}{N_n(L_i)} = \frac{N_n(L_j,a)}{N_n(L_j)} \quad \forall a \in \mathcal{A}$ (ii)  $d_{\mathcal{L}}(i,j) = d_{\mathcal{L}}(j,i)$ (iii)  $d_{\mathcal{L}}(i,k) \le d_{\mathcal{L}}(i,j) + d_{\mathcal{L}}(j,k)$ 

**Proof.** See García e González-López (2017).

**Remark 2.4.13.** As a consequence of Theorem 1 proved in García e González-López (2017), if  $(X_t)_{t\in\mathbb{Z}}$  is a discrete time, order M stationary Markov chain on a finite alphabet  $\mathcal{A}$  and  $x_1^n$  is a sample of the process, then for n large enough, for each  $s, r \in \mathcal{S}, d_{\mathcal{L}}(r, s) < 1$  iff s and r belong to the same part of the partition.

# 3 New Measure between Stochastic Processes

We describe below the issues addressed in this chapter. In section 3.2 we expose and prove the main results about the properties of a BIC-based measure  $d_s$ built between samples. In section 3.3 we use the distance  $d_s$  between strings to identify linguistic compositions that show a different performance when comparing written texts of Portuguese. In section 3.4 we describe the strains of 15 patients, using  $d_s$  to establish a notion of natural proximity between DNA sequences from patients with identical diagnosis, which is: Burkitt lymphoma/leukemia. In section 3.5 we apply the measure to the problem of fuel alcohol production, comparing two lines of production.

# 3.1 Comparison of Stochastic Processes

There are several practical situations in which it is necessary to quantify divergences between the laws of samples coming from stochastic processes. Whether to establish that all the samples follow the same stochastic law or in case of being governed by the same law, the interest may lie in establishing a metric that allows to decide which of these samples are closer. For example, the interest may be in inspecting industrial processes, in which the behavior of production lines could point to concrete evidence of the (dis)similarities found in the final products. For instance, industrial plants usually carry out their productions in parallel. That is, these are planned to operate in various production lines. The raw material comes from the same source and is separated into different systems that process them. The lines of production are monitored in order to obtain equivalent products to the end of the production process.

In this section, we introduce a distance which allows to compare Markovian processes. We show 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 of stochastic processes. When a discrepancy exist, the use of this distance allows us to find the strings where the discrepancy is manifested. Also we establish a local metric between samples based on the Bayesian Information Criterion, we derive the bound that must be used in this metric to take the decision. We show that the distance is statistically consistent to detect if the samples follow the same law, tending to zero when the sample sizes increase. We show that the metric assumes arbitrarily large values when the sample sizes increase and the stochastic laws are different. In this chapter, we investigate the performance of  $d_s$  in several real problems, coming from several areas as linguistics, genetics and industry.

## 3.2 Results: The Measure and its Properties

Consider the independent samples  $x_{1,1}^{n_1}$ ,  $x_{2,1}^{n_2}$  of  $(X_{1,t})_{t\in Z}$  with law P and  $(X_{2,t})_{t\in Z}$ with law Q respectively. Define  $N_{n_1+n_2}(s, a) = N_{n_1}(s, a) + N_{n_2}(s, a)$ ,  $N_{n_1+n_2}(s) = N_{n_1}(s) + N_{n_2}(s)$ , where  $N_{n_1}$  and  $N_{n_2}$  are given as usual, computed from the samples  $x_{1,1}^{n_1}$  and  $x_{2,1}^{n_2}$ respectively. We start calculating the BIC value for the joint model on which  $P(\cdot|r) \neq Q(\cdot|r)$ ,  $\forall r \in S$ . If the samples are independent then the likelihood of the two samples is

$$P(x_{1,1}^M)P(x_{2,1}^M)\prod_{a\in\mathcal{A},r\in\mathcal{S}}P(a|r)^{N_{n_1}(r,a)}Q(a|r)^{N_{n_2}(r,a)}$$

and the total number of parameters to be estimated is  $2(|\mathcal{A}|-1)|\mathcal{S}|$ . Under the assumption that the memory of both processes is M, and the state space for both is  $\mathcal{S}$ . Obtaining the log-maximum likelihood for the two samples, as

$$\sum_{a \in \mathcal{A}, r \in \mathcal{S}} \left\{ N_{n_1}(r, a) \ln \left( \frac{N_{n_1}(r, a)}{N_{n_1}(r)} \right) + N_{n_2}(r, a) \ln \left( \frac{N_{n_2}(r, a)}{N_{n_2}(r)} \right) \right\}$$

On the other hand, if we consider that there is an specific  $s \in S$  on which  $P(\cdot|s) = Q(\cdot|s)$ , the number of parameters is now  $(|\mathcal{A}| - 1)(2|\mathcal{S}| - 1)$  and under this constraints the log-maximum likelihood is,

$$\sum_{a \in \mathcal{A}, r \in \mathcal{S} \setminus \{s\}} \left\{ N_{n_1}(r, a) \ln\left(\frac{N_{n_1}(r, a)}{N_{n_1}(r)}\right) + N_{n_2}(r, a) \ln\left(\frac{N_{n_2}(r, a)}{N_{n_2}(r)}\right) \right\} + \sum_{a \in \mathcal{A}} N_{n_1+n_2}(s, a) \ln\left(\frac{N_{n_1+n_2}(s, a)}{N_{n_1+n_2}(s)}\right).$$

In this way we have all the elements to show how the BIC can be used to decide if the two sets of probabilities related to an specific  $r \in S$  should be considered as being the same. We will formulate a distance  $d_s$  that, when evaluated in a given string s, allows us to define how far or near the processes are. All our results are shown to be connected with the relative entropy  $D(P(\cdot)||Q(\cdot))$  between distributions.

**Definition 3.2.1.** Consider two independent and stationary Markov chains  $(X_{1,t})_{t\in \mathbb{Z}}$  and  $(X_{2,t})_{t\in \mathbb{Z}}$ , of order M, with finite alphabet  $\mathcal{A}$ , state space  $\mathcal{S} = \mathcal{A}^M$  and samples  $x_{1,1}^{n_1}$ ,  $x_{2,1}^{n_2}$  respectively. Define for a string  $s \in \mathcal{S}$ ,

$$d_{s}(x_{1,1}^{n_{1}}, x_{2,1}^{n_{2}}) = \frac{\alpha}{(|\mathcal{A}| - 1)\ln(n_{1} + n_{2})} \sum_{a \in \mathcal{A}} \left\{ N_{n_{1}}(s, a) \ln\left(\frac{N_{n_{1}}(s, a)}{N_{n_{1}}(s)}\right) + N_{n_{2}}(s, a) \ln\left(\frac{N_{n_{2}}(s, a)}{N_{n_{2}}(s)}\right) - N_{n_{1}+n_{2}}(s, a) \ln\left(\frac{N_{n_{1}+n_{2}}(s, a)}{N_{n_{1}+n_{2}}(s)}\right) \right\}$$

with  $\alpha$  a real and positive value.

**Remark 3.2.2.** According to definition 3.2.1,  $d_s$  is only the difference between two formulations of the BIC. One of them considering that the samples follow different conditional laws and the other considering that the samples follow the same law. Moreover, since  $d_s$  is a measure any clustering algorithm can be used to deal with the magnitude of  $d_s$ .

The following theory is represented to demonstrate the relationship between the distance  $d_s$  and the Kullback-Leibler divergence  $D(\cdot || \cdot)$  (presented in section 2.4).

**Theorem 3.2.3.** Let  $(X_{k,t})_{t\in\mathbb{Z}}$  be a stationary Markov chain of order M, with finite alphabet  $\mathcal{A}$ , state space  $\mathcal{S} = \mathcal{A}^M$  and  $x_{k,1}^{n_k}$  a sample of the process for k = 1, 2. Consider also  $s \in \mathcal{S}$ . If  $D\left(\frac{N_{n_k}(s,\cdot)}{N_{n_k}(s)} \parallel \frac{N_{n_1+n_2}(s,\cdot)}{N_{n_1+n_2}(s)}\right) < \infty$ , for k = 1, 2, then

$$d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = \frac{\alpha}{(|\mathcal{A}| - 1)\ln(n_1 + n_2)} \sum_{k=1,2} N_{n_k}(s) D\left(\frac{N_{n_k}(s, \cdot)}{N_{n_k}(s)} \parallel \frac{N_{n_1 + n_2}(s, \cdot)}{N_{n_1 + n_2}(s)}\right).$$

**Proof.** Note that  $\ln(n_1 + n_2) \frac{(|\mathcal{A}| - 1)}{\alpha} d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2})$  is

$$= \sum_{a \in \mathcal{A}} \left\{ N_{n_1}(s, a) \ln \left( \frac{N_{n_1}(s, a)}{N_{n_1}(s)} \right) + N_{n_2}(s, a) \ln \left( \frac{N_{n_2}(s, a)}{N_{n_2}(s)} \right) \right. \\ \left. - (N_{n_1}(s, a) + N_{n_2}(s, a)) \ln \left( \frac{N_{n_1+n_2}(s, a)}{N_{n_1+n_2}(s)} \right) \right\} \\ = \sum_{a \in \mathcal{A}} \left\{ N_{n_1}(s, a) \left( \ln \left( \frac{N_{n_1}(s, a)}{N_{n_1}(s)} \right) - \ln \left( \frac{N_{n_1+n_2}(s, a)}{N_{n_1+n_2}(s)} \right) \right) \right\} + \\ \left. \sum_{a \in \mathcal{A}} \left\{ N_{n_2}(s, a) \left( \ln \left( \frac{N_{n_2}(s, a)}{N_{n_2}(s)} \right) - \ln \left( \frac{N_{n_1+n_2}(s, a)}{N_{n_1+n_2}(s)} \right) \right) \right\} \right\} \\ = \sum_{k=1,2} N_{n_k}(s) \sum_{a \in \mathcal{A}} \frac{N_{n_k}(s, a)}{N_{n_k}(s)} \ln \left( \frac{N_{n_k}(s, a)}{N_{n_k}(s)} / \frac{N_{n_1+n_2}(s, a)}{N_{n_1+n_2}(s)} \right) \\ = \sum_{k=1,2} N_{n_k}(s) D\left( \frac{N_{n_k}(s, \cdot)}{N_{n_k}(s)} \parallel \frac{N_{n_1+n_2}(s, \cdot)}{N_{n_1+n_2}(s)} \right).$$

The questions derived from the definition of  $d_s$  are, is  $d_s$  a metric?, how is its relation with the BIC?, what can we say about its behavior in terms of statistical consistency?. In the next theorems we address and respond to all these points. We start by showing that  $d_s$  is a metric.

**Theorem 3.2.4.** Consider three stationary Markov chains  $(X_{i,t})_{t\in\mathbb{Z}}$ , i = 1, 2, 3 of order M, with finite alphabet  $\mathcal{A}$ , state space  $\mathcal{S} = \mathcal{A}^M$  and independent samples  $x_{i,1}^{n_i}$ , i = 1, 2, 3. Consider a string  $s \in \mathcal{S}$ ,

*i.*  $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \ge 0$  with equality  $\Leftrightarrow \frac{N_{n_1}(s, a)}{N_{n_1}(s)} = \frac{N_{n_2}(s, a)}{N_{n_2}(s)} \quad \forall a \in \mathcal{A},$ *ii.*  $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = d_s(x_{2,1}^{n_2}, x_{1,1}^{n_1}),$ 

*iii.* 
$$d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \leq d_s(x_{1,1}^{n_1}, x_{3,1}^{n_3}) + d_s(x_{3,1}^{n_3}, x_{2,1}^{n_2})$$

**Proof.** Following similar calculations of theorem 2 in García e González-López (2017). For (i) define  $a_i = N_{n_i}(s, a)$  and  $b_i = N_{n_i}(s)$ , i = 1, 2. Consider the log-sum inequality:  $a_1 \ln(\frac{a_1}{b_1}) + a_2 \ln(\frac{a_2}{b_2}) \ge \sum_{i=1,2} a_i \ln\left(\frac{\sum_{i=1,2} a_i}{\sum_{i=1,2} b_i}\right)$ , with equality  $\Leftrightarrow \frac{a_1}{b_1} = \frac{a_2}{b_2}$ , and the result follows.

For (iii) assume by simplicity  $n_1 = n_2 = n_3$ , we note that the statement to prove is equivalent to

$$0 \leqslant \sum_{k=1,2} \left\{ N_{n_3}(s) D\left(\frac{N_{n_3}(s,\cdot)}{N_{n_3}(s)} || \frac{N_{n_k+n_3}(s,\cdot)}{N_{n_k+n_3}(s)}\right) + \sum_{a \in \mathcal{A}} \frac{N_{n_k}(s,a) N_{n_1+n_2}(s)}{N_{n_1+n_2}(s,a)} \frac{N_{n_1+n_2}(s,a)}{N_{n_1+n_2}(s)} \ln\left(\frac{N_{n_1+n_2}(s,a)}{N_{n_1+n_2}(s)} / \frac{N_{n_k+n_3}(s,a)}{N_{n_k+n_3}(s)}\right) \right\},$$

and, the right side of the previous inequality is greater to the next positive term

$$N_{n_3}(s) \sum_{k=1,2} D\left(\frac{N_{n_3}(s,\cdot)}{N_{n_3}(s)} || \frac{N_{n_k+n_3}(s,\cdot)}{N_{n_k+n_3}(s)}\right) + \frac{1}{n} \sum_{k=1,2} D\left(\frac{N_{n_1+n_2}(s,\cdot)}{N_{n_1+n_2}(s)} || \frac{N_{n_k+n_3}(s,\cdot)}{N_{n_k+n_3}(s)}\right). \quad \blacksquare$$

The next result shows the connection of  $d_s$  with the BIC.

**Theorem 3.2.5.** Consider two stationary Markov chains  $(X_{1,t})_{t\in\mathbb{Z}}$  and  $(X_{2,t})_{t\in\mathbb{Z}}$ , of order M, with finite alphabet  $\mathcal{A}$ , state space  $\mathcal{S} = \mathcal{A}^M$  and independent samples  $x_{1,1}^{n_1}$ ,  $x_{2,1}^{n_2}$  respectively. Denote by  $\{P(a|r)\}_{a\in\mathcal{A},r\in\mathcal{S}}$  and  $\{Q(a|r)\}_{a\in\mathcal{A},r\in\mathcal{S}}$  the sets of conditional probabilities of  $(X_{1,t})_{t\in\mathbb{Z}}$  and  $(X_{2,t})_{t\in\mathbb{Z}}$  respectively. Given  $s \in \mathcal{S}$ ,

$$BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2}) < BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2}, =_s) \Leftrightarrow d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) < 1,$$

where  $BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2})$  and  $BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2}, =_s)$  are given by definition 2.2.6 and the second is formulated under the assumption:  $P(a|s) = Q(a|s), \forall a \in \mathcal{A}$ .

**Proof.** Since the samples are independent, from definition 2.2.6 we obtain

$$BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = \sum_{a \in \mathcal{A}, r \in \mathcal{S}} N_{n_1}(r, a) \ln\left(\frac{N_{n_1}(r, a)}{N_{n_1}(r)}\right) \\ + \sum_{a \in \mathcal{A}, r \in \mathcal{S}} N_{n_2}(r, a) \ln\left(\frac{N_{n_2}(r, a)}{N_{n_2}(r)}\right) \\ - \frac{(|\mathcal{A}| - 1)}{\alpha} 2|\mathcal{S}| \ln(n_1 + n_2),$$

where the number of parameters to be estimated is twice in comparison with the case of one sample. If we assume: P(a|s) = Q(a|s) where s is a string in S, we obtain

$$BIC(x_{1,1}^{n_1}, x_{2,1}^{n_2}, =_s) = \sum_{a \in \mathcal{A}, r \in \mathcal{S} \setminus \{s\}} N_{n_1}(r, a) \ln\left(\frac{N_{n_1}(r, a)}{N_{n_1}(r)}\right) + \sum_{a \in \mathcal{A}, r \in \mathcal{S} \setminus \{s\}} N_{n_2}(r, a) \ln\left(\frac{N_{n_2}(r, a)}{N_{n_2}(r)}\right) + \sum_{a \in \mathcal{A}} N_{n_1+n_2}(s, a) \ln\left(\frac{N_{n_1+n_2}(s, a)}{N_{n_1+n_2}(s)}\right) - \frac{(|\mathcal{A}| - 1)}{\alpha} \{(|\mathcal{S}| - 1)2 + 1\} \ln(n_1 + n_2).$$

Then, the result follows directly.

As we can see, a consequence of this result is that the BIC indicates the decision to consider the samples as being governed by the same stochastic law, when the value of  $d_s$  is less than 1. Allowing to create an specific decision rule that corresponds directly with the formulation of the BIC. Also, we can assert that the samples are governed by different stochastic laws when  $d_s$  assumes values greater than 1, in this case the BIC recognizes that it is necessary to estimate two groups of conditional probabilities and not more one group as it happens in the previous case, when  $d_s < 1$ .

On the next result we show the statistical consistency of  $d_s$ .

**Theorem 3.2.6.** Consider two Markov chains  $X_{1,t}$  and  $X_{2,t}$  of order M, with finite alphabet A, state space  $S = A^M$  and independent samples  $x_{1,1}^{n_1}$ ,  $x_{2,1}^{n_2}$  respectively. Denote by  $\{P(a|r)\}_{a \in A, r \in S}$  and  $\{Q(a|r)\}_{a \in A, r \in S}$  the sets of conditional probabilities of  $X_{1,t}$  and  $X_{2,t}$ respectively. Consider a string  $s \in S$ ,

$$i. if P(a|s) = Q(a|s) \forall a \in \mathcal{A} then, \ d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \xrightarrow[\min(n_1, n_2) \to \infty]{} 0,$$
$$ii. if there is \ a \in \mathcal{A} : P(a|s) \neq Q(a|s) then, \ d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \xrightarrow[\min(n_1, n_2) \to \infty]{} \infty$$

**Proof.** *i*. : Since  $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \ge 0$  from the previous result, given an arbitrary and positive value  $\delta$ , it is enough to prove that there exists a sample size  $n_0$  such that, if  $\min(n_1, n_2) > n_0$ ,  $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) < c\delta$ , for a constant and positive value *c*. Let's see that this happens.

Since  $\frac{N_{n_1}(s,a) + N_{n_2}(s,a)}{N_{n_1}(s) + N_{n_2}(s)}$  is the maximum likelihood estimator of P(a|s) then,

$$\frac{N_{n_1}(s,a) + N_{n_2}(s,a)}{N_{n_1}(s) + N_{n_2}(s)} \ge P(a|s), \text{ as a consequence } d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \text{ is}$$

$$\leqslant \frac{\alpha}{(|\mathcal{A}| - 1) \ln(n_1 + n_2)} \sum_{i=1,2} \sum_{a \in \mathcal{A}} N_{n_i}(s,a) \ln\left(\frac{N_{n_i}(s,a)}{N_{n_i}(s)}/P(a|s)\right)$$

$$= \frac{\alpha}{(|\mathcal{A}| - 1) \ln(n_1 + n_2)} \sum_{i=1,2} N_{n_i}(s) D\left(\frac{N_{n_i}(s,\cdot)}{N_{n_i}(s)}||P(\cdot|s)\right)$$

$$\leqslant_{(1)} \frac{\alpha}{(|\mathcal{A}| - 1) \ln(n_1 + n_2)} \sum_{i=1,2} N_{n_i}(s) \sum_{a \in \mathcal{A}} \frac{\left(\frac{N_{n_i}(s,a)}{N_{n_i}(s)} - P(a|s)\right)^2}{P(a|s)}$$

$$\leqslant_{(2)} \frac{\alpha}{(|\mathcal{A}| - 1) \ln(n_1 + n_2)} \sum_{i=1,2} N_{n_i}(s) \sum_{a \in \mathcal{A}} \delta \frac{\ln(n_i)}{N_{n_i}(s)P(a|s)}$$

$$\leqslant \frac{\alpha\delta|\mathcal{A}|}{(|\mathcal{A}| - 1) \ln(n_1 + n_2)p^*} \sum_{i=1,2} \ln(n_i)$$

$$\leqslant_{(3)} \frac{\alpha\delta|\mathcal{A}|2}{(|\mathcal{A}| - 1)p^*} \frac{\ln(n^*)}{\ln(n_1 + n_2)} \leqslant_{(4)} \frac{2\alpha|\mathcal{A}|}{(|\mathcal{A}| - 1)p^*}\delta.$$

Where  $p^* = \min\{P(a|s) : a \in \mathcal{A}\}$  and  $n^* = \max\{n_1, n_2\}$ . (1) is a consequence of lemma 6.3 of Csiszár e Talata (2006). (2) is coming from lemma 6.2 of Csiszár e Talata (2006) , since, given an arbitrary  $\delta > 0$ , for each i (i = 1, 2) there exists  $k_i > 0$  such that  $M < k_i \ln(n_i)$ . (3) and (4) are immediate, since  $1 < n^* < n_1 + n_2$ .

According to the proof, the constant c is given by  $\frac{2\alpha|\mathcal{A}|}{(|\mathcal{A}|-1)p^*}$  and  $n_0 = \lceil \max\{e^{\frac{M}{k_1}}, e^{\frac{M}{k_2}}\} \rceil$ . *ii.*: From theorem 1 of García e González-López (2017),

$$d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = \frac{\alpha}{(|\mathcal{A}| - 1)\ln(n_1 + n_2)} \sum_{k=1,2} N_{n_k}(s) D\left(\frac{N_{n_k}(s, \cdot)}{N_{n_k}(s)} \parallel \frac{N_{n_1 + n_2}(s, \cdot)}{N_{n_1 + n_2}(s)}\right)$$

When  $n_1 \to \infty$ ,  $\frac{N_{n_1}(s, a)}{N_{n_1}(s)} \to P(a|s)$ , and when  $n_2 \to \infty$ ,  $\frac{N_{n_2}(s, a)}{N_{n_2}(s)} \to Q(a|s)$ . Since  $P(a|s) \neq Q(a|s)$ , if we denote by  $P_{1,2}$  the law of the mixture, when  $\min\{n_1, n_2\} \to \infty$ ,  $\frac{N_{n_1+n_2}(s, a)}{N_{n_1+n_2}(s)} \to P_{1,2}(a|s)$ . As a consequence when  $\min\{n_1, n_2\} \to \infty$ ,

$$D\left(\frac{N_{n_k}(s,\cdot)}{N_{n_k}(s)} \parallel \frac{N_{n_1+n_2}(s,\cdot)}{N_{n_1+n_2}(s)}\right) \to \begin{cases} D\left(P(\cdot|s) \parallel P_{1,2}(\cdot|s)\right) & \text{if } k=1;\\ D\left(Q(\cdot|s) \parallel P_{1,2}(\cdot|s)\right) & \text{if } k=2, \end{cases}$$

with a positive limit in both cases. In addition  $\frac{N_{n_k}(s)}{\ln(n_1+n_2)} = \frac{n_k}{\ln(n_1+n_2)} \frac{N_{n_k}(s)}{n_k} \to \infty$ , when  $\min\{n_1, n_2\} \to \infty$ , since,

$$\frac{N_{n_k}(s)}{n_k} \to \begin{cases} P(s) & \text{if } k = 1; \\ Q(s) & \text{if } k = 2. \end{cases}$$

Then, the result follows.

The latter theorem shows that increasing the sample sizes increases the ability of  $d_s$  of to detect discrepancies or similarities between the stochastic laws involved. That is, as is to be expected, the statistical consistency of the transition probabilities carry their characteristics for the metric  $d_s$ .

In the following result, we show the asymptotic behavior of the distance using the following notations of Kullback-Leibler divergence D(P||Q) and  $\chi^2(P(\cdot), Q(\cdot)) = \sum_{a \in \mathcal{A}} \frac{(P(a) - Q(a))^2}{Q(a)}$ .

**Theorem 3.2.7.** Considering two distributions P and Q, with empirical distribution  $\hat{P}(a) = \frac{X(a)}{k_1}$  and  $\hat{Q}(a) = \frac{Y(a)}{k_2}$  where two samples of sizes  $k_1$  and  $k_2$  generated from the law W, respectively, defined in the alphabet  $\mathcal{A}$ , for  $Q(a) \neq 0$ ,  $a \in \mathcal{A}$  and k = 1, 2, then

$$\frac{D(P(\cdot)||Q(\cdot))}{\frac{1}{2k_1}\chi^{2,k_1}(\hat{P}(\cdot),W(\cdot)) + \frac{1}{2k_2}\chi^{2,k_2}(\hat{Q}(\cdot),W(\cdot))} \to 1.$$
(3.1)

**Proof.** Consider the function  $f(x) = x \ln(x)$ , near to x = 1, by the Taylor's expansion we have  $f(x) = (x - 1) + \frac{(x - 1)^2}{2} + \delta(x)(x - 1)^2$  where  $\delta(x) = -\frac{(x - 1)}{6t^2}$  for some value  $t \in (x, 1)$  (Lagrange's form). We note that when  $x \to 1, \delta(x) \to 0$ . Thus, for two probability distributions P and Q in  $\mathcal{A}$ ,

$$\begin{split} P(a) \ln \left(\frac{P(a)}{Q(a)}\right) &= Q(a) f\left(\frac{P(a)}{Q(a)}\right) \\ &= P(a) - Q(a) + \frac{1}{2} \frac{(P(a) - Q(a))^2}{Q(a)} + \delta \left(\frac{P(a)}{Q(a)}\right) \frac{(P(a) - Q(a))^2}{Q(a)}, \end{split}$$

for  $a \in \mathcal{A}$ ,

$$D(P(\cdot)||Q(\cdot)) = \frac{1}{2}\chi^{2}(P(\cdot),Q(\cdot)) + \sum_{a\in\mathcal{A}}\delta\Big(\frac{P(a)}{Q(a)}\Big)\frac{(P(a)-Q(a))^{2}}{Q(a)}$$
(3.2)

and

$$\frac{D(P(\cdot)||Q(\cdot))}{\chi^2(P(\cdot),Q(\cdot))} = \frac{1}{2} + \frac{\sum_{a\in\mathcal{A}} \delta\left(\frac{P(a)}{Q(a)}\right) \frac{(P(a)-Q(a))^2}{Q(a)}}{\chi^2(P(\cdot),Q(\cdot))}$$

$$\begin{split} & \text{If } \frac{P(a)}{Q(a)} \to 1, \text{ given } \epsilon \text{ positive and small enough, } |\delta\left(\frac{P(a)}{Q(a)}\right)| < \epsilon \text{ and } \Big|\frac{\sum_{a \in \mathcal{A}} \delta\left(\frac{P(a)}{Q(a)}\right) \frac{(P(a) - Q(a))^2}{Q(a)}}{\chi^2(P(\cdot), Q(\cdot))}\Big| < \epsilon, \text{ so } \frac{D(P(\cdot)||Q(\cdot))}{\chi^2(P(\cdot), Q(\cdot))} \to \frac{1}{2}. \end{split}$$

If one of the probabilities is the empirical distribution, say  $\hat{P}(a) = \frac{X(a)}{k}$ , where the occurrences of a in the sample of size k is denoted by X(a), and the sample is generated from the law Q,  $\chi^2(\hat{P}(\cdot)||Q(\cdot)) = \frac{1}{k} \sum_{a \in \mathcal{A}} \frac{(X(a) - kQ(a))^2}{kQ(a)}$ . Thus, if we introduce the quantity  $\chi^{2,k}(\hat{P}(\cdot),Q(\cdot)) = \sum_{a \in \mathcal{A}} \frac{(X(a) - kQ(a))^2}{kQ(a)}$ , we can recognize the typical Chi-square statistic. From the equation 3.2 we obtain

$$\begin{split} D(\hat{P}(\cdot)||Q(\cdot)) &= \frac{1}{2k} \chi^{2,k}(\hat{P}(\cdot),Q(\cdot)) + \sum_{a \in \mathcal{A}} \frac{1}{k} \delta\Big(\frac{\hat{P}(a)}{Q(a)}\Big) \frac{(X(a) - kQ(a))^2}{kQ(a)} \\ \text{and when } \frac{\hat{P}(a)}{Q(a)} \to 1, \\ &\qquad \frac{D(\hat{P}(\cdot)||Q(\cdot))}{\chi^{2,k}(\hat{P}(\cdot),Q(\cdot))} \to \frac{1}{2k}. \end{split}$$

If we have two samples of sizes  $k_1$  and  $k_2$  generated from the law W, with empirical distribution  $\hat{P}(a) = \frac{X(a)}{k_1}$  and  $\hat{Q}(a) = \frac{Y(a)}{k_2}$  respectively. According to equation 3.2 we obtain

$$D(\hat{P}(\cdot)||\hat{Q}(\cdot)) = \frac{1}{k_1} \sum_{a \in \mathcal{A}} \left(\frac{W(a)}{\hat{Q}(a)}\right) \left(\frac{1}{2} + \delta\left(\frac{\hat{P}(a)}{\hat{Q}(a)}\right)\right) \frac{(X(a) - k_1 W(a))^2}{k_1 W(a)} + \frac{1}{k_2} \sum_{a \in \mathcal{A}} \left(\frac{W(a)}{\hat{Q}(a)}\right) \left(\frac{1}{2} + \delta\left(\frac{\hat{P}(a)}{\hat{Q}(a)}\right)\right) \frac{(Y(a) - k_2 W(a))^2}{k_2 W(a)} + \sum_{a \in \mathcal{A}} \left(1 + 2\delta\left(\frac{\hat{P}(a)}{\hat{Q}(a)}\right)\right) (\hat{P}(a) - W(a)) \left(\frac{W(a)}{\hat{Q}(a)} - 1\right).$$

So, when 
$$\frac{W(a)}{\hat{Q}(a)} \to 1$$
 and  $\frac{\hat{P}(a)}{\hat{Q}(a)} \to 1$ ,  
$$\frac{D(\hat{P}(\cdot)||\hat{Q}(\cdot))}{\frac{1}{2k_1}\chi^{2,k_1}(\hat{P}(\cdot),W(\cdot)) + \frac{1}{2k_2}\chi^{2,k_2}(\hat{Q}(\cdot),W(\cdot))} \to 1.$$

These simple relationships between empirical distributions allows us to delineate the behavior of the distance  $d_s$  (definition 3.2.1).

**Theorem 3.2.8.** Let  $(X_{k,t})_{t\in\mathbb{Z}}$  be a stationary Markov chain of order M, with finite alphabet  $\mathcal{A}$  state space  $\mathcal{S} = \mathcal{A}^M$  and  $x_{k,1}^{n_k}$  a sample of the process for k = 1, 2. Consider also  $s \in \mathcal{S}$ . If  $D\left(\frac{N_{n_k}(s,\cdot)}{N_{n_k}(s)} \parallel \frac{N_{n_1+n_2}(s,\cdot)}{N_{n_1+n_2}(s)}\right) < \infty$ , when  $\frac{N_{n_k}(s,\cdot)/N_{n_k}(s)}{W(\cdot)} \to 1$  for k = 1, 2, then

$$2\ln(n_1+n_2)\frac{(|\mathcal{A}|-1)}{\alpha}d_s(x_{1,1}^{n_1},x_{2,1}^{n_2})\sim_d$$

$$\sum_{k=1,2} \chi^{2,N_{n_k}(s)} \Big( \frac{N_{n_k}(s,\cdot)}{N_{n_k}(s)}, W(\cdot) \Big) + \chi^{2,N_{n_1+n_2}(s)} \Big( \frac{N_{n_1+n_2}(s,\cdot)}{N_{n_1+n_2}(s)}, W(\cdot) \Big),$$

where  $\sim_d$  means similarity in distribution.

**Proof.** According to theorem 3.2.3  $\ln(n_1 + n_2) \frac{(|\mathcal{A}| - 1)}{\alpha} d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2})$  is

$$\sum_{k=1,2} N_{n_k}(s) D\left(\frac{N_{n_k}(s,\cdot)}{N_{n_k}(s)} \parallel \frac{N_{n_1+n_2}(s,\cdot)}{N_{n_1+n_2}(s)}\right)$$

Following the equation 3.1, we have

$$\sum_{k=1,2} N_{n_k}(s) D\left(\frac{N_{n_k}(s,\cdot)}{N_{n_k}(s)} \parallel \frac{N_{n_1+n_2}(s,\cdot)}{N_{n_1+n_2}(s)}\right) \sim_d$$

$$\sum_{k=1,2} \frac{N_{n_k}(s)}{2} \Big\{ \frac{\chi^{2,N_{n_k}(s)}\left(\frac{N_{n_k}(s,\cdot)}{N_{n_k}(s)},W(\cdot)\right)}{N_{n_k}(s)} + \frac{\chi^{2,N_{n_1+n_2}(s)}\left(\frac{N_{n_1+n_2}(s,\cdot)}{N_{n_1+n_2}(s)},W(\cdot)\right)}{N_{n_1+n_2}(s)} \Big\}.$$

Then,

$$2\ln(n_1+n_2)\frac{(|\mathcal{A}|-1)}{\alpha}d_s(x_{1,1}^{n_1},x_{2,1}^{n_2})\sim_d$$

$$\sum_{k=1,2} \chi^{2,N_{n_k}(s)} \Big( \frac{N_{n_k}(s,\cdot)}{N_{n_k}(s)}, W(\cdot) \Big) + \chi^{2,N_{n_1+n_2}(s)} \Big( \frac{N_{n_1+n_2}(s,\cdot)}{N_{n_1+n_2}(s)}, W(\cdot) \Big). \quad \blacksquare$$

We show now a notion of proximity between samples considering all the values of  $d_s$  with  $s \in \mathcal{S}$  (see García, Gholizadeh e González-López (2017b)).

**Definition 3.2.9.** Consider two stationary Markov chains  $(X_{1,t})_{t\in\mathbb{Z}}$  and  $(X_{2,t})_{t\in\mathbb{Z}}$ , of order M, with finite alphabet  $\mathcal{A}$ , state space  $\mathcal{S} = \mathcal{A}^M$  and independent samples  $x_{1,1}^{n_1}, x_{2,1}^{n_2}$  respectively,

$$dmax(x_{1,1}^{n_1}, x_{2,1}^{n_2}) = \max_{s \in \mathcal{S}} \{ d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \}$$

and

$$smax = \arg\max\left\{dmax(x_{1,1}^{n_1}, x_{2,1}^{n_2})\right\}$$

Observe that  $dmax < \epsilon$  if and only if  $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2}) < \epsilon, \forall s \in \mathcal{S}$ . 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.

**Remark 3.2.10.** From the properties observed for  $d_s(x_{1,1}^{n_1}, x_{2,1}^{n_2})$  it follows that,

- $i. \ P(a|s) = Q(a|s) \ \forall a \in \mathcal{A}, s \in \mathcal{S} \ if, and only \ if, \ dmax(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \xrightarrow[\min(n_1, n_2) \to \infty]{} 0,$
- ii.  $P(a|s) \neq Q(a|s)$  for some  $a \in \mathcal{A}$  and  $s \in \mathcal{S}$  if, and only if,  $dmax(x_{1,1}^{n_1}, x_{2,1}^{n_2}) \xrightarrow{\min(n_1, n_2) \to \infty} \infty$ .

Three case studies are presented in the following section applying the concept given by definition 3.2.1, which is already described in section 3.2. The first case study concerns to a linguistic issue that considers written texts dated between 16th century and 18th century. The second case study is about stochastic distance between Burkitt lymphoma/leukemia Strains, while the third case study considers the production of alcohol fuel. A step by step strategy was employed in each case by comparing several processes to tackle real problems accurately. In these applications we consider a method to measure discrepancies between samples avoiding to fit a model.

# 3.3 First Case Study: Linguistic Data

Investigations in the field of historical linguistics record a body of evidence on the changes occurring in written texts, from Classical to Modern Portuguese, including the period from 16th to 19th century. Frota et al. (2012) shows clear evidence of changes occurred from the 16th century to the 17th century, in relation to the prosody of the language. In this case study, our focus is to identify the most relevant linguistic constructions (N-grams) that lead to these changes. That is, by sequentially observing the texts, we want to identify the constructions that lead to relevant changes. One line of research is to treat a written text, or some strategic coding thereof, as a sequence of N-grams. The structure of N-grams plays a very important role in the inspection and modeling of language profiles, see for instance Manning and Manning e Schütze (1999). In some of the investigations, the focus has been on discriminating between languages, for example see García and Garcia e González-López (2016) (under the scope of Partition Markov Models (PMM)), in others, the purpose has been to discriminate between varieties of the same language, for example see Galves et al. (2012) (under the scope of Context Tree Models).

Tycho Brahe corpus is an annotated historical corpus, freely accessible at Galves e Faria (2010). This corpus uses the chronological criterion of the author's birthdate to assign a time for written texts. The subset of historical written texts included in this study, listed in table 2 is composed by 19 texts from 15 authors, coming from five genres.

Author	Gândavo	Pinto	Sousa	Brandão	Vieira
Date	1502	1510	1556	1584	1608
Type	narrative	narrative	narrative	narrative	dissertation
Author	Vieira	Vieira	Chagas	Bernardes	Oliveira
Date	1608	1608	1631	1644	1702
Type	letters	sermons	letters	narrative	letters
Author	Aires	Costa	Alorna	Garrett	Garrett
Date	1705	1714	1750	1799	1799
Type	dissertation	letters	letters	letters	narrative
Author	Garrett	Fronteira	Camilo	Ortigão	
Date	1799	1802	1826	1836	
Type	theater	narrative	narrative	letters	

Table 2 – The set of the Tycho Brahe corpus.

There are previous studies (see Frota et al. (2012)) that show that historical texts such as the listed in table 2 reveal changes in the proportion of occurrence of the placement of the stress in the last or in the penultimate syllable of the word. Also the written texts reveal alterations in the use of monosyllables. These changes are found predominantly from the 16th century to the 17th century. For this reason we guide our inspection to the position in the word occupied by the stress and the word size. Each written text was processed with a slightly modified version of the perl-code "silaba" (by Miguel Galves) 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 which integrate the word, i = 1, 2, ..., 8 and *j* indicates the position of the stressed syllable in the word (from left to right). j = 0 means no stress in the word. The period (final of sentence) was codified as (0, 0). The alphabet  $\mathcal{A}$  used here was defined as exposed in table 3.

Orthographic word code	Element in the alphabet $\mathcal{A}$	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
$(i,i), i \ge 3$	6	oxytone word
$(i, i-1), i \ge 3$	7	paroxytone word
$(i, i-2), i \ge 3$	8	proparoxytone word

Table 3 – Definition and meaning of each element  $a \in \mathcal{A}$ .

In this approach we used linguistic composition of two words (bigrams), for technical reasons: size of the alphabet and size of the available texts. For example, the linguistic

structure 2-7 represents an unstressed monosyllable followed by a *paroxytone* word. The perspective introduced in this study aims to incorporate in the analysis of written texts the dependence between the words that compose them. When considering a bigram s we see that the discrepancies between two written texts will be confirmed, if the next word a to be found in the text 1 and 2, are different. Precisely, given a bigram s, if  $d_s(1,2) > 1$  we will have that  $Prob_1(a|s) \neq Prob_2(a|s)$  with  $Prob_i$  computed from the text i, i = 1, 2 and a a word of the alphabet.

## 3.3.1 The Most Variable Configurations

According to definition 3.2.9 and remark 3.2.10, 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  $d_s$  will reveal changes on the local laws of the processes relative to the string. In this application a value larger than 1 will be considered significant.

Date text 1	Date text 2	dmax	smax
1502	1510	1.08679	2-4
1510	1556	1.64926	2-7
1556	1584	0.71331	1-6
1584	1608c	1.73511	2-4
1584	1608d	1.00874	7-2
1584	1608s	1.35197	2-7
1608c	1631	4.44799	2-4
1608d	1631	3.10919	2-4
1608s	1631	1.78843	2-4
1631	1644	2.00082	2-4
1644	1702	1.03420	4-4
1702	1705	2.04039	7-0
1705	1714	2.59383	2-4
1714	1750	4.46181	2-4
1750	1799c	1.45198	4-7
1750	1799t	6.14052	2-4
1750	1799n	2.29650	2-4
1799c	1802	6.62204	7-2
1799t	1802	6.45512	2-7
1799n	1802	9.04413	7-2
1802	1826	6.39264	7-2
1826	1836	0.52470	2-4

Table 4 – Values of dmax between a written text and the written text, dated immediately after the previous one.



Figure 4 – dmax values (on the vertical axis) denoted by the year of the second written text (column 2 of the table 4). In the case of a year with several texts, a symbol was attached to the year, which indicates the type of written text: narrative (n), letters (c), sermons (s), theater (t), dissertation (d).

Figure 4 shows the values of dmax over the years, recorded on the horizontal axis. We note that over the years, the discrepancies detected by dmax are more pronounced, except at the end of the 19th century (right). It is worth emphasizing that each discrepancy can be produced by bigrams that are not necessarily identical for all the texts, as detailed in the last column of table 4.

	a	<i>smax</i> : 2-4	(1.08679)	smax: 2-7	(1.64926)	
		1502	1510	1510	1556	
	0	0.03251	0.01590	0.01974	0.05232	
	1	0.06572	0.04623	0.04877	0.04559	
	2	0.34383	0.39770	0.52833	0.52970	
	3	0.04220	0.04310	0.05864	0.03270	
	4	0.28087	0.23577	0.19509	0.16385	
	6	0.02421	0.01485	0.01635	0.01944	
	7	0.19301	0.23159	0.12602	0.13896	
	8	0.01764	0.01485	0.00705	0.01744	
a	smax: 2-4	(1.73511)	<i>smax</i> : 7-2	(1.00874)	<i>smax</i> : 2-7	(1.35197)
	1584	1608c	1584	1608d	1584	1608s
0	0.03356	0.02277	0.00022	0.00000	0.05627	0.11444
1	0.07102	0.06624	0.10044	0.09927	0.05447	0.05488
2	0.38474	0.31714	0.13624	0.15025	0.52607	0.47124
3	0.04528	0.04251	0.07707	0.04350	0.03614	0.03284
4	0.20363	0.20402	0.32729	0.29417	0.15643	0.16809
6	0.02528	0.02638	0.03930	0.03986	0.01859	0.02203
7	0.21719	0.31144	0.29127	0.34266	0.13552	0.12464
8	0.01931	0.00949	0.02817	0.03028	0.01652	0.01183
a	smax: 2-4	(4.44799)	smax: 2-4	(3.10919)	smax: 2-4	(1.78843)
	1608c	1631	1608d	1631	1608s	1631
0	0.02277	0.04852	0.03027	0.04852	0.06825	0.04852
1	0.06624	0.08622	0.07731	0.08622	0.07620	0.08622
2	0.31714	0.32035	0.33885	0.32035	0.35255	0.32035
3	0.04251	0.09578	0.03550	0.09578	0.04473	0.09578
4	0.20402	0.23178	0.21733	0.23178	0.22949	0.23178
6	0.02638	0.01768	0.03223	0.01768	0.02175	0.01768
7	0.31144	0.19012	0.24913	0.19012	0.19307	0.19012
8	0.00949	0.00956	0.01938	0.00956	0.01397	0.00956

Table 5 – Conditional probabilities from smax to each element a of the alphabet  $\mathcal{A}$ . Texts: 1502, 1510, 1556, 1584, 1608c, 1608d, 1608s, 1631.

a	<i>smax</i> : 2-4	(2.00082)	<i>smax</i> : 4-4	(1.03420)	<i>smax</i> : 7-0	(2.04039)
	1631	1644	1644	1702	1702	1705
0	0.04852	0.03472	0.03494	0.11220	0.00000	0.00000
1	0.08622	0.07377	0.10519	0.09368	0.06766	0.12903
2	0.32035	0.33500	0.31225	0.30174	0.53498	0.25605
3	0.09578	0.04860	0.05112	0.07190	0.05791	0.05645
4	0.23178	0.23085	0.22030	0.18954	0.22764	0.30847
6	0.01768	0.02864	0.03715	0.02832	0.00516	0.03427
7	0.19012	0.22282	0.21552	0.18736	0.09117	0.15323
8	0.00956	0.02560	0.02354	0.01525	0.01548	0.06250
		1		1		
	a	<i>smax</i> : 2-4	(2.59383)	<i>smax</i> : 2-4	(4.46181)	
		1705	1714	1714	1750	
	0	0.07313	0.04440	0.04440	0.03273	
	1	0.09521	0.07985	0.07985	0.05687	
	2	0.33216	0.33246	0.33246	0.29573	
	3	0.03062	0.10373	0.10373	0.03412	
	4	0.23695	0.23657	0.23657	0.19684	
	6	0.01824	0.01493	0.01493	0.02484	
	7	0.20030	0.17761	0.17761	0.34123	
	8	0.01339	0.01045	0.01045	0.01764	
1		(1 (5100)	2.4	(0.1.050)	2.4	(2,20,450)
a	smax: 4-7	(1.45198)	smax: 2-4	(6.14052)	smax: 2-4	(2.29650)
	1750	1799c	1750	1799t	1750	1799n
0	0.08432	0.10345	0.03273	0.13364	0.03273	0.06361
1	0.05295	0.05314	0.05687	0.10649	0.05687	0.07905
2	0.40000	0.44432	0.29573	0.27959	0.29573	0.30952
3	0.12310	0.03957	0.03412	0.04200	0.03412	0.05047
4	0.17774	0.16789	0.19684	0.24608	0.19684	0.21318
6	0.02226	0.02600	0.02484	0.01273	0.02484	0.02374
7	0.11872	0.14811	0.34123	0.17522	0.34123	0.23070
8	0.02091	0.01752	0.01764	0.00424	0.01764	0.02973

Table 6 – Conditional probabilities from smax to each element a of the alphabet  $\mathcal{A}$ . Texts: 1631, 1644, 1702, 1705, 1714, 1750, 1799c, 1799t, 1799n.

a	<i>smax</i> : 7-2	(6.62204)	$\parallel smax: 2-7$	(6.45512)	<i>smax</i> : 7-2	(9.04413)
	1799c	1802	1799t	1802	1799n	1802
0	0.00064	0.00031	0.25458	0.06325	0.00183	0.00031
1	0.12162	0.06796	0.08215	0.03171	0.10353	0.06796
2	0.16651	0.38500	0.35777	0.51374	0.15735	0.38500
3	0.05890	0.04704	0.04684	0.04228	0.05153	0.04704
4	0.28335	0.22984	0.14053	0.14711	0.29684	0.22984
6	0.04011	0.02673	0.01154	0.02872	0.02863	0.02673
7	0.31137	0.22770	0.10251	0.16138	0.32410	0.22770
8	0.01751	0.01542	0.00407	0.01180	0.03619	0.01542
		a	smax: 7-2	(6.39264)		
			1802	1826		
		0	0.00031	0.00000		

	1802	1826
0	0.00031	0.00000
1	0.06796	0.08152
2	0.38500	0.13136
3	0.04704	0.05919
4	0.22984	0.32347
6	0.02673	0.03998
7	0.22770	0.33022
8	0.01543	0.03427

Table 7 – Conditional probabilities from smax to each element a of the alphabet  $\mathcal{A}$ . Texts: 1799c, 1799t, 1799n, 1802, 1826.

We see that the most frequent bigram which produces a high  $dmax \ (> 1)$  is 2-4. From the meaning of  $d_s$  this means that, the probability conditioned to an unstressed monosyllable followed by a disyllable with stress at the beginning of the word motivates such discrepancies between consecutive texts, i.e. these conditional probabilities are very different from a text to the following text, in the cases: 1502-1510, 1584-1608c, 1608c-1631,1608d-1631, 1608s-1631, 1631-1644, 1705-1714, 1714-1750, 1750-1799t, 1750-1799n. Although this same bigram causes the maximum value of dmax, between the texts of 1826 and 1836, the value of dmax in this case does not indicate a discrepancy between them (because dmax < 1). The bigram 7-2 appears as the next string responsible for discrepancy between texts. The transition probability of a paroxytone word followed by an unstressed monosyllable can be considered different for the cases: 1584-1608d, 1799c-1802, 1799n-1802, 1802-1826. Similarly occurs with 2-7, the transition probability of an unstressed monosyllable followed by a paroxytone word, can be considered different for the cases: 1510-1556, 1584-1608s, 1799t-1802. In tables 5, 6 and 7 we expose the conditional probabilities from the bigram smax, to each value of the alphabet  $\mathcal{A}$  and for each pair of texts. We found cases in which the disparity between the processes is evident, since the conditional probabilities are markedly different, see for instance: 1702-1705, 1714-1750, 1750-1799t, 1750-1799n, 1799c-1802, 1799t-1802, 1799n-1802, 1802-1826.

$d_s(1750, 1799t)$	s	$d_s(1799n, 1802)$	s	$d_s(1799t, 1802)$	s
1.12915	2-2	1.00300	4-7	1.05268	2-6
1.13272	7-4	1.21533	1-2	1.13008	4-3
1.18684	4-1	1.39296	6-2	1.16699	4-1
1.21775	1-4	1.55913	2-2	1.23502	7-0
1.22487	3-2	1.69919	2-7	1.31616	1-2
1.22692	1-7	1.96170	3-2	1.51818	0-2
1.26838	6-2	2.01062	2-4	1.62364	1-4
1.29315	7-7	2.08092	4-4	1.62953	7-7
1.39431	1-2	5.31407	4-2	1.66957	7-2
2.18292	2-1	9.04413	7-2	1.75069	1 - 7
2.35651	4-4			1.77862	7-4
3.37187	4-2			2.56165	2-1
3.38483	7-2			2.58853	2-2
4.00310	4-7			3.62165	4-2
4.66029	2-7			4.05870	4-4
6.14052	2-4			4.39322	4-7
				6.39768	2-4
				6.45512	2-7

Table 8 – Cases with bigger values of  $d_s$  and different smax: 1750-1799t, 1799n-1802,1799t-1802. In bold the bigrams that most often produce the highest values of  $d_s$ .

In table 8 we list all the bigrams that show values of  $d_s > 1$  for the cases: 1750-1799t, 1799n-1802 and 1799t-1802 that are those that show a higher dmax for the 3 most frequent smax, 2-4, 2-7, 7-2. All these cases are in the 18th and early 19th century. We can note that the bigrams (i) an unstressed monosyllable followed by a disyllable with stress at the beginning of the word, code: 2-4; (ii) a paroxytone word followed by an unstressed monosyllable, code: 7-2 and (iii) an unstressed monosyllable followed by a paroxytone word, code: 2-7 detect values of  $d_s$  greater than 1, practically in all the written texts, so they should not necessarily be considered as responsible for the changes from the 16th century to the 17th century. We can argue that those constructions between others (also with  $d_s > 1$ ) are constructions with a tendency to report the particularity of each text.

## 3.3.2 From the 16th Century to the Beginning of the 17th Century

In Frota et al. (2012) significant changes are reported in the language from the 16th century to the 17th century. In the previous section we noticed that some bigrams are intrinsically variable, being characterized by large values of dmax. In this section, we examine the transition from the 16th century to the 17th century, taking into account that the 3 configurations cited in the previous section do not necessarily lead to drastic changes in the language. We record all the bigrams which report changes (i.e. with values

of  $d_s > 1$ ), considering each written text of the 16th century in relation to the written texts dated immediately afterwords, until the beginning of the 17th century. Tables 9, 10 and 11 present the results.

$d_s(1502, 1$	$(510) \mid s$		$d_s(1502, 1556)$	s	$  d_s(1502, 1584)  $	s
1.08679 2-4		1	1.10873	7-2	1.10897	4-2
			1.51328	4-2	1.32348	4-4
	1					
$d_s(1502, 1608c)$	s		$d_s(1502, 1608d)$	s	$d_s(1502, 1608s)$	s
1.15103	4-2		1.85333	4-2	1.07436	1-4 (V)
1.21610	2-3 (II)		1.93021	7-2		
2.16232	2-4					

Table 9 – Values of  $d_s$  and bigrams such that  $d_s > 1$ , between texts of the 16th century and Vieira's texts: 1608c, 1608d and 1608s. In bold letter the most frequent bigrams, according to the previous section.

$d_s(1510, 1556)$	s	$d_s(1510, 1584)$	s
1.39871	4-2	1.02257	4-0
1.64242	7-2	1.04342	2-4
1.64926	2-7	1.06105	1-3
		1.12283	7-2
		1.15678	4-4
		1.19666	3-6
		1.33785	1-6
		1.37050	7-0
		1.51808	2-7

$d_s(1510, 1608c)$	s	$d_s(1510, 1608d)$	s	$d_s(1510, 1608s)$	s
1.20665	2-7	1.04371	1-7 (VI)	1.41470	4-4
2.03579	2-4	1.07434	1-4 (V)	2.06415	4-7 (I)
		1.13177	7-0	3.20374	2-4
		1.31517	1-6	4.43740	2-7
		1.36703	4-2		
		1.55650	2-4		
		1.81723	2-7		
		2.43819	7-2		

Table 10 – Values of  $d_s$  and bigrams such that  $d_s > 1$ , between texts of the 16th century and Vieira's texts: 1608c, 1608d and 1608s. In bold letter the most frequent bigrams, according to the previous section.

	$d_s(1556, 160)$	s	$d_s(15)$	56, 160	(8s)	S			
—	1.32334		2-4	1.	57438		4-7(I)		
	1.66714		7-2	1.	67042		2-4		
				1.	94014		2-7		
$d_s(1584, 1608c)$	s	$d_s($	1584, 1	608d)	s	$d_s(1$	1584, 1608s)	s	
1.07605	3-1 (III)		1.008'	74	7-2		1.07634	<b>2-</b> 4	Ł
1.10511	6-1 (IV)						1.14234	3-1(I	II)
1.10537	7-2						1.35197	2-7	7
1.14962	1-6								
1.23217	2-3 (II)								
1.41250	3-6								
1.73511	2-4								

Table 11 – Values of  $d_s$  and bigrams such that  $d_s > 1$ , between texts of the 16th century and Vieira's texts: 1608c, 1608d and 1608s. In bold letter the most frequent bigrams, according to the previous section.

In table 12 we list the bigrams detected as a change in the comparison between each written text of the 16th century with the first 3 written texts of the 17th century: 1608c, 1608d and 1608s. In that list we exclude configurations that are identified as changes between texts of the 16th century itself.

Code string	Bigram	Reference
4-7	a disyllable with stress on the first syllable	(I)
	followed by a <i>paroxytone</i> word	
2-3	an unstressed monosyllable followed by	(II)
	a disyllable with stress on the last syllable	
3-1	a disyllable with stress on the last syllable	(III)
	followed by a stressed monosyllabic word	
6-1	an <i>oxytone</i> word followed by	(IV)
	a stressed monosyllabic word	
1-4	a stressed monosyllabic word followed by	(V)
	a disyllable with stress on the first syllable	
1-7	a stressed monosyllabic word	(VI)
	followed by a <i>paroxytone</i> word	

Table 12 – Bigrams that announce changes between texts of the 16th century when compared to texts of beginning of the 17th century: 1608c, 1608d, 1608s. In the third column are indicated the cases covered by the configuration, see tables 3.8, 3.9, 3.10

## 3.3.3 Conclusion

In this study we introduce a strategy to identify linguistic structures (bigrams) that generate alterations of the Portuguese. Also it is possible to identify the bigrams more strongly associated with historical changes. Bigrams with large values of  $d_s$  unrelated to temporal changes could possibly be used to discriminate linguistic genres or particular aspects of texts. Moreover, the idea of identifying the language with sequences of N-grams thus adopting the measure  $d_s$  to proceed to the detection of changes, can be applied to other contexts and problems, helping to solve and review linguistic alterations proclaimed in the literature of the area of historical linguistics. In this instance, it is necessary to make some observations. The dmax detects volatile linguistic constructions that expose changes in several moments from Classical Portuguese to Modern Portuguese (period: 16th century to 19th century). Among them, the most outstanding constructions, with maximum  $d_s$  value and most frequent, are: (i) an unstressed monosyllable followed by a paroxytone disyllable word, (ii) a paroxytone word followed by an unstressed monosyllable and (iii) an unstressed monosyllable followed by a paroxytone word. These voluble linguistic constructions allow to delineate the profile of the Portuguese language in the period: 16th century to 19th century, showing in a clear way the constructions more associated to the changes of the period. These results already show that bigrams composed by unstressed monosyllables and paroxytone words (and viceversa) are the most likely to suffer alteration. It should be remembered that in Frota et al. (2012) these two characteristics indicate significant changes in the Portuguese of the period: 16th-17th. In the present work we go further, because bigrams take into account the dependence between both aspects: unstressed monosyllables and paroxytone words. When comparing the texts of the 16th century with the first texts of the 17th century, it is possible to detect a series of bigrams that indicate important differences, since, in these cases, the measure d adopts values greater than 1. These are (a) a disyllable with stress on the first syllable followed by a paroxytone word (indicated by two texts), (b) an unstressed monosyllable followed by a disyllable with stress on the last syllable (indicated by one text), (c) a disyllable with stress on the last syllable followed by a stressed monosyllabic word (indicated by two texts), (d) an oxytone word followed by a stressed monosyllabic word (indicated by two texts), (e) a stressed monosyllabic word followed by a disyllable with stress on the first syllable (indicated by two texts) and (f) a stressed monosyllabic word followed by a paroxytone word (indicated by one text).

# 3.4 Second Case Study: Comparison between DNA Strains

The Burkitt lymphoma occurs when the chromosome 8 (locus of gene MYC) is broken, which produces a change in the cellular proliferation. The data used in this application corresponds to the most frequent variant, produced by the translocation

between chromosomes 8 and 14. It is known, so far, three variants of Burkitt lymphoma, which are (i) endemic, (ii) sporadic, (iii) produced by immunodeficiency. The first case is observed in childs in Equatorial Africa and it is associated with chronic Malaria infections. It does not exist until the moment and according to what we know, a clear notion of the profile of the Burkitt lymphoma's DNA. Considering that it is natural to expect diversity between DNA strains, we will measure the distance between 15 of them. We adopt a distance between the strains which is conditioned to each possible common string s, where s is an element of the state space. That is, suppose that  $x_{1,1}^{n_1}$  and  $x_{2,1}^{n_2}$  are the concatenations of elements a, c, g and t of the DNA of two patients, say 1 and 2,  $d_s(1,2)$  will be the distance between the sequences in relation to s some string of interest, for instance s =aggc. As there are a variety of possible strings, which we should observe to measure the discrepancy between the strains, we will compute the maximum of all:  $\max_{s} \{d_s(1,2)\}$ , so as to focus on the most extreme situation among them. This notion allows to identify which of these strings can be considered more distant of the majority, and allows us to select the strains which will be used to define the profile of the DNA. To strengthen our conclusions, we compared the model constructed with the selected strains with the model constructed using the 15 available strains.

### 3.4.1 DNA Data

The database is composed by 15 DNA sequences, available in the repository: <https://www.ncbi.nlm.nih.gov/nuccore/>, coming from 15 patients with Burkitt lymphoma/leukemia carrying the t(8;14)(q24;q32) with IgH-MYC fusion, breakpoint in the joining region. The registers (genbank numbers) of the sequences are: AM2871z.1, where z=39, 40, 41, 46, 50, 52, 57, 58, 59, 61, 62, 65, 76, 81, 87. For each sequence, the concatenation of bases a,c,g,t observed in the code is the realization denoted by  $x_1^n$ . The size of each sequence is shown in table 13.

z	39	40	41	46	50	52	57	
n	3641	2965	4464	2731	5428	2475	3907	
z	58	59	61	62	65	76	81	87
n	3636	4291	2642	3206	2906	2635	3608	3734

Table 13 – Sample sizes n of DNA sequence coming from 15 patients with Burkitt lymphoma/leukemia, AM2871z.1, where z = 39, 40, 41, 46, 50, 52, 57, 58, 59, 61, 62, 65, 76, 81, 87.

#### 3.4.2 Results

In tables 14 and 15 we expose the *dmax* values between the DNA sequences, where  $dmax(i, j) = \max_{s \in S} \{d_s(i, j)\}, i \neq j, i, j = AM2871z.1$ , with z = 39, 40, 41, 46, 50, 52, 57, 58, 59, 61, 62, 65, 76, 81, 87. At the end of each column we record the sum of the dmax, that is:

$$S(i) = \sum_{j} dmax(i, j)$$
, for each sequence  $i = AM2871z.1$ ,

where z = 39, 40, 41, 46, 50, 52, 57, 58, 59, 61, 62, 65, 76, 81, 87. Through  $d_s$  we have a criterion to rescue the greatest distance between two DNA sequences. From the magnitudes found, we can affirm that the processes can be considered as coming from the same stochastic law, dmax < 1. We also verified the above statement from the dendrograms constructed using the values recorded in tables 14 and 15, see figure 5.

$j \setminus i$	39	40	41	46	50	52	57
40	0.23625						
41	0.16160	0.25648					
46	0.22578	0.24031	0.21857				
50	0.20218	0.25847	0.17855	0.22644			
52	0.19479	0.17870	0.21143	0.16253	0.33231		
57	0.09777	0.24533	0.13885	0.22058	0.12481	0.19363	
58	0.27729	0.21783	0.30105	0.25156	0.28312	0.23041	0.25738
59	0.12485	0.32050	0.09723	0.24232	0.15545	0.21165	0.09821
61	0.20229	0.10170	0.22626	0.20598	0.30120	0.12572	0.25328
62	0.32556	0.34309	0.35858	0.26633	0.47720	0.24569	0.32362
65	0.22234	0.15183	0.26545	0.15812	0.25339	0.29264	0.27469
76	0.19421	0.24629	0.20804	0.12923	0.23960	0.12786	0.19308
81	0.16363	0.17050	0.19272	0.16614	0.22817	0.17392	0.12994
87	0.26047	0.16796	0.24704	0.25130	0.41112	0.26425	0.22481
S(i)	2.8890	3.13523	3.06186	2.96519	3.67203	2.94553	2.77597

Table 14 – dmax(i, j) values,  $i \neq j, i, j$  = AM2871z.1, where z = 39, 40, 41, 46, 50, 52, 57, 58, 59, 61, 62, 65, 76, 81, 87.

$j \setminus i$	58	59	61	62	65	76	81	87
59	0.30177							
61	0.20284	0.32032						
62	0.27748	0.34112	0.25478					
65	0.25707	0.27412	0.21689	0.30528				
76	0.13397	0.21109	0.13318	0.27990	0.21237			
81	0.25801	0.14463	0.11904	0.23329	0.21155	0.15334		
87	0.23089	0.24762	0.20658	0.37764	0.20689	0.23144	0.19405	
S(i)	3.48067	3.09091	2.87007	4.40955	3.30265	2.69363	2.53891	3.52205

Table 15 – dmax(i, j) values,  $i \neq j, i, j$  = AM2871z.1, where z = 39, 40, 41, 46, 50, 52, 57, 58, 59, 61, 62, 65, 76, 81, 87. In bold the lowest value of S, associated to the sequence with z = 81.



Figure 5 – Dendrograms build through the *dmax* values (tables 3.13-14), agglomeration method: *Average*, on the left and *Complete*, on the right.

The model we will apply in the data, is extensively investigated in García e González-López (2017). This is the most general model known to be used in finite order Markov chains on a finite alphabet, since this model includes fixed order Markov chains and the variable lenght Markov chains (VLMC). Essentially what this model proposes as we describe in chapter 2 is to estimate the transition probabilities that describe the process by identifying a partition  $\mathcal{L} = \{L_1, \dots, L_{|\mathcal{L}|}\}$  in the state space  $\mathcal{S}$ . The state space is divided into parts  $L_i, i = 1, \dots, |\mathcal{L}|$  which constitute a partition. The strings of each part have in common the characteristic of sharing the same transition probability to any element of the alphabet. In practice, all strings included in the same part of that partition will be used for the computation of the transition probability that identifies them. As we show in chapter 2, the identification of such partition is done using the Bayesian Information Criterion (BIC), which also is the basis to the concept  $d_s$ , previously introduced.

Table 16 shows some general characteristics that are observed in the adjustment of the model introduced in García e González-López (2017). We include progressively (from top to bottom) the closest sequences, according to the criterion  $\mathcal{S}$ . That is, first using the sequence 81, second, using two sequences: 81 and 76 and so on. In other words, we are increasing the sample size from one stage to the next, following as inclusion criterion the magnitude of  $\mathcal{S}$ .

2	Sample size	$ \mathcal{S} $	$ \mathcal{L} $
81	3604	134	6
81,76	6235	193	13
81,76,57	10138	241	18
81,76,57,61	12776	249	21
81,76,57,61,39	16413	255	27
81,76,57,61,39,52	18884	255	28
81,76,57,61,39,52,46	21611	255	27
81,76,57,61,39,52,46,41	26071	256	31
81, 76, 57, 61, 39, 52, 46, 41, 59	30358	256	33
81,76,57,61,39,52,46,41,59,40	33319	256	31
81,76,57,61,39,52,46	36221	256	34
$41,\!59,\!40,\!65$			
81,76,57,61,39,52,46	39853	256	37
$41,\!59,\!40,\!65,\!58$			
81,76,57,61,39,52,46	43583	256	39
$41,\!59,\!40,\!65,\!58,\!87$			
81,76,57,61,39,52,46	49007	256	40
$41,59,40,\ 65,58,87,50$			
81,76,57,61,39,52,46	52209	256	42
$41,\!59,\!40,\!65,\!58,\!87,\!50,\!62$			

Table 16 – Relation between the sequences used in the estimation and number of parts of the estimated partition, for AM2871z.1, where z=39, 40, 41, 46, 50, 52, 57, 58, 59, 61, 62, 65, 76, 81, 87.

We can not state unequivocally that by increasing the sample sizes we increase the parts of the estimated partition, but it seems to be a trend, as seen in the table 16. But this could also be the result of incorporating in the model gradually the more distant sequences according to criterion S.

We apply in all the adjustments the agglomerative method, whose performance is analyzed in García e González-López (2017), the memory used in all the adjustments is equal to  $M = 4 \leq \lfloor \log_{|\mathcal{A}|}(2475) \rfloor - 1$ , with alphabet  $\mathcal{A} = \{a, c, g, t\}$  where 2475 is the smallest sample size reported in table 13.

We describe in a comparative way the results when applying the model in: (i) the 7 closest sequences according to S, which are: AM2871z.1, where z=39, 46, 52, 57, 61, 76, 81 (see tables 17 and 18) and (ii) the 15 sequences (see tables 20 and 21).

$i$ of part $L_i$	Strings			
1	acgc, accg, ccag, gacg, acac, gcag, caat, atca			
2	ccgc, cggt, cctc, agga, tcac, tagg, acca, gcac			
3	gcgc, cgtg, ctat, tctc, cagg, cacc, taag, cgtt, ttct, cccc, ggct, gtca			
	ctct, agac, tctt, ctta, tgcc, atgc, gttt, tatc, gctt, cttg, agct			
4	tcgc, gaca, tttc, ctgc, ttgt, gata, gtta			
5	agge, aaca, agte, agea, atte, tteg, aagt, taca, ageg, cage, gtee			
6	cggc, cgct, tttg, atac, gccg, caga, ttat, ctaa, tagt, ctca, gaaa			
7	gggc, gtct, aatt, ttgg, cctt, ttgc, tgga, ctgt, taat, tgta, ccat			
	tcct, ttca, ggaa, ctcg, tgtt, agaa, gtga			
8	tggc, atct, gagt, gagc, aatc, tacg, ggcc, ggtt, agtg, cact, ataa			
9	gtgc, gatc, catc, aaga, gctc, aaat, aata			
10	actc, tgaa, acag, gtat			
11	cgtc, tgtc, gtag, aggt, ttag, ttga, gtac, gcct			
12	ggtc, gtaa, agtt, caaa, gttc, gaag, atag			
13	cttc, taaa, ttta, catt, attt, aaag, ttaa, acat, aagc, cttt, aaac			
14	caag, gcca, tgag, gcaa, aacg, acga, ccac			
15	tcag, attg, agag, atcc, aact, cgat, cgta, catg, taga, tccc, ttcc, acaa			
16	cgag, tccg, tcta, ggta, taac, acgg, gaga, cata			
17	ggag, tgtg, tgct, tcca, tctg, tttt, ccca, ccga, ggca, gcgg, gtgg			
	tgat, gggg			
18	ctag, tcgt, ctgg, aaaa, tcgg, gtgt, gggt, tggt, gatt			
19	cacg, aagg, tcaa, cgcg, actg, cgca, tgcg, tcat, ccgt			
20	cccg, ctac, atcg, aggg, aatg, ggcg, cggg			
21	gtcg, atta, ggat, cgaa, gagg, ggac, tact, tgca, tata, agat, acct			
	gcga, tcga			
22	ccgg, gatg, caac, cctg, atgg, tatt, tggg, tatg, accc, ggtg, cgac, atgt			
	gcta, ggga			
23	gctg, gaat, gttg, acgt, tgac, gacc			
24	gcat, gact, ccta, gcgt, caca, acta, gaac, ttac			
25	atat, cgga, actt, atga, ccct, cagt, aacc, ccaa, agta, ctga			
26	tacc			
27	gccc, cgcc, ctcc, agcc			

Table 17 – Parts of the partition selected through the Bayesian Information Criterion, using AM2871z.1, where z=39, 46, 52, 57, 61, 76, 81.

$i$ of part $L_i$	a	с	g	t
1	0.52430	0.24041	0.18670	0.04859
2	0.26705	0.53977	0.13636	0.05682
3	0.31714	0.25073	0.30718	0.12495
4	0.16245	0.27557	0.53791	0.02407
5	0.47689	0.07948	0.27542	0.16821
6	0.20235	0.17204	0.33822	0.28739
7	0.11628	0.31924	0.46564	0.09884
8	0.26162	0.34661	0.36122	0.03054
9	0.14495	0.11376	0.57982	0.16147
10	0.00487	0.42336	0.37226	0.19951
11	0.02754	0.11864	0.61441	0.23941
12	0.26225	0.08357	0.62824	0.02594
13	0.17198	0.23406	0.41527	0.17869
14	0.35484	0.00000	0.18894	0.45622
15	0.27907	0.22161	0.12996	0.36936
16	0.02667	0.72800	0.11200	0.13333
17	0.23476	0.35264	0.24390	0.16870
18	0.13996	0.43050	0.31757	0.11197
19	0.36605	0.38650	0.02658	0.22086
20	0.20380	0.51813	0.09845	0.17962
21	0.08753	0.43885	0.11631	0.35731
22	0.16603	0.35227	0.21136	0.27034
23	0.14469	0.25736	0.23303	0.36492
24	0.01155	0.33949	0.22864	0.42032
25	0.07393	0.47471	0.28664	0.16472
26	0.03922	0.05882	0.03922	0.86275
27	0.26437	0.04310	0.46264	0.22989

Table 18 – Transition probabilities  $P(\cdot|L_i)$  with  $\cdot \in \{a,c,g,t\}$  and  $i = 1, \dots, 27$ . For each part *i*, listed on the left column (see table 17), we indicate in bold the highest transition probability to the elements of the alphabet.

We note (see table 18) that in relation to the transition probabilities from part i to the elements of the alphabet, 3 of these parts show their highest values in the transition to a, 10 parts expose their highest values in the transition to c, 9 of those parts show their highest values in the transition to g and 5 of those parts expose their greater probabilities in the transition to t.

we give their transition probabilities in table 21.

i of	part $L_i$		Strings	Probabilit	у
	1	acgo	e, accg, ccag, gacg, acac, gcag, caat, atca	$P(a L_1) = 0.52$	2430
	16	cgag	g, tccg, tcta, ggta, taac, acgg, gaga, cata	$P(c L_{16}) = 0.7$	2800
	12	g	gtc, gtaa, agtt, caaa, gttc, gaag, atag	$P(g L_{12}) = 0.6$	2824
	26		tacc	$P(t L_{26}) = 0.8$	6275
	i of par	part $L_i$ Strings		Probability	
	14		caag, gcca, tgag, gcaa, aacg, acga, ccac	$P(\mathbf{c} L_{14}) = 0$	

Table 19 – Selected parts, from table 17, which have the greater (on top)/null (on bottom) transition probabilities to each element of the alphabet {a,c,g,t}.

In table 19 we highlight the composition of four parts 1, 16, 12 and 26 that show the highest values of transition probability for a, c, g and t respectively. We also emphasize in table 19, the part 14 that joins all those strings whose transition probability to c is zero. We list in table 20 the elements of the partition obtained using all the strains, and then

65

$i$ of part $L_i$	Strings
1	acgc, accg
2	ccgc, gagt, acca, gagc, ggcg, cggt
3	gcgc, tacg, cgtg, ccca, gccg, ctat, cacc, ttat, ctaa, caga
4	tcgc, ggtt, aatc, ggcc, atca, agtg, ataa, tggc, atct
5	aggc, agtc, aaca, agca
6	cggc, cgct, tttg, atac
7	gggc, ctcg, gtct, cctt, gaca, tagc
8	atgc, gttt, cttg, tatc, ttgg, cttc, catt, taaa, gctt
9	ctgc, ctgt, gttc, tttc
10	gtgc, ggtc, gaag, atag, agtt, caaa, gtaa
11	ttgc, tgga, taat, aaac, ttca, aaag, ttaa, acat, aagc, ttta, cttt
12	catc, ctcc, gctc, aaga, gccc
13	gatc, aata, aaat, gtac, gtag, ttga, ttag, aggt
14	actc, agaa, cgga, atat, atga
15	cctc, tgcc, ggct, tcca, acac, ggag, tgat, tgtg, tgct, gcac, tctg, tttt, gtca
16	tctc, taag, cccc, cagg, ttct, ctct, tctt
17	cgtc, tgtc, gcct
18	attc, aagt, agct, taca, tagt, ctca, gaaa
19	caag, gcca, ccac
20	acag, gtat, tgaa
21	ccag, agac, agga, gcag, gacg, caat, cgtt
22	tcag, attg, aact, agag, catg, atcc, aatg, cgat, cgta
23	cgag, tccg
24	tgag, acga, gcaa, agcc, cgcc
25	ctag, ggga, accc, gctg, atgt, gaat, acgt, gttg, tgac, gacc
26	aacg, taga, acaa, tccc, ttcc
27	cacg, aagg, tcaa, cgca, actg, cgcg, tagg, tcac
28	cccg, ctac, aggg, atcg, gtcg, atta, cggg
29	agcg, cagc, gtcc, ttcg
30	tgcg, tcat, ccgt, acct
31	gagg, ggac, tgca, gcat, ttac, gact, gaac, caca, tata
32	acgg, agat
33	ccgg, gatg, tact, atgg, cctg, gcta, caac, tggg, tatg, ggtg, tatt, cgac, ggca, ccga
34	gcgg, gtgg, cact, ctta, attt, aaaa, gggg, cgaa
35	tcgg, actt, tggt, gggt, ctgg, tcgt, gatt
36	ccat, ggaa, tgta, tcct, aatt, gata, gtta, gtgt, tgtt, gtga
37	ggat, gcga, gaga, tcga
38	ccct, cagt, aacc, ccta, ccaa, agta, ctga
39	gcgt, acta
40	ttgt
41	cata, tcta, ggta, taac
42	tacc

Table 20 – Parts of the partition selected through the Bayesian Information Criterion, using all the sequences AM2871z.1, where z = 39, 40, 41, 46, 50, 52, 57, 58, 59, 61, 62, 65, 76, 81, 87.

$i \text{ of part } L_i$	a	с	g	t
1	0.62162	0.17568	0.11824	0.08446
2	0.20141	0.52669	0.18127	0.09063
3	0.27795	0.22742	0.25900	0.23563
4	0.26518	0.31020	0.37473	0.04989
5	0.49296	0.03873	0.34859	0.11972
6	0.15173	0.17569	0.36646	0.30612
7	0.17982	0.28801	0.44956	0.08260
8	0.26409	0.22741	0.38224	0.12625
9	0.13973	0.20960	0.58923	0.06145
10	0.25732	0.09728	0.57741	0.06799
11	0.15512	0.23870	0.42244	0.18373
12	0.22067	0.08288	0.50377	0.19268
13	0.07543	0.16140	0.58475	0.17843
14	0.07708	0.49605	0.26383	0.16304
15	0.26020	0.34949	0.22874	0.16156
16	0.34054	0.25250	0.28011	0.12685
17	0.05213	0.06398	0.55450	0.32938
18	0.29789	0.14042	0.34416	0.21753
19	0.38671	0.04532	0.09970	0.46828
20	0.01037	0.41014	0.38134	0.19816
21	0.43774	0.26038	0.21384	0.08805
22	0.27390	0.29363	0.11684	0.31563
23	0.08333	0.82222	0.05556	0.03889
24	0.26710	0.05375	0.35668	0.32248
25	0.14725	0.28990	0.24258	0.32027
26	0.28553	0.14211	0.18027	0.39211
27	0.38189	0.38091	0.06102	0.17618
28	0.19352	0.51001	0.09724	0.19924
29	0.48899	0.11006	0.19654	0.20440
30	0.22482	0.37230	0.02698	0.37590
31	0.05018	0.35636	0.17236	0.42109
32	0.02055	0.66438	0.02740	0.28767
33	0.17956	0.35776	0.19200	0.27069
34	0.19475	0.34030	0.30463	0.16031
35	0.15342	0.42826	0.30464	0.11369
36	0.09625	0.36320	0.44644	0.09401
37	0.07349	0.48294	0.14961	0.29396
38	0.09836	0.36339	0.31785	0.22040
39	0.00339	0.35254	0.28136	0.36271
40	0.17865	0.31828	0.49281	0.01027
41	0.05017	0.63378	0.16890	0.14716
42	0.05970	0.10448	0.12687	0.70896

Table 21 – Transition probabilities  $P(.|L_i)$  with  $. \in \{a, c, g, t\}$  and  $i = 1, \dots, 42$ . For each part *i*, listed on the left column (see table 20), we indicate in bold the highest transition probability to the elements of the alphabet.

According to table 21, 7 parts exhibit their highest transition probability values for the element a, 13 for the element c, 14 for the element g, and 8 for the element t.

$i$ of part $L_i$	Strings	Probability
1	acgc, accg	$P(a L_1) = 0.62162$
23	cgag, tccg	$P(c L_{23}) = 0.82222$
9	ctgc, ctgt, gttc, tttc	$P(g L_9) = 0.58923$
42	tacc	$P(t L_{42}) = 0.70896$

Table 22 – Selected parts, from table 20 and 21, which have the greater transition probabilities to each element of the alphabet {a,c,g,t}.

Note that the parts recorded in the selection given in table 19, where we use only 50% of the nearest strains, are combinations of those listed in table 22 with other parts, in the latter case we use all the strains. We detail the connection in the table 23.

Index of part from table 19	Indices of parts - table 20
1	1,4,15,21
16	23,32,37,41
12	9,10
26	42
14	19,24,26

Table 23 – Relation between the parts listed in table 19 and 20. On left we display the parts coming from the model using only 50% of the DNA sequences, on right the parts coming from the model using all the DNA strains. In the same line, on the right we list the parts in which are identified the elements into the part on the left.

We see that the listed parts (to the left of table 23) are dispersed in several parts of the model adjusted with all the sequences. In the case of the last line, the strings listed in part 14 of table 19 occur with nonzero frequencies, when using all the sequences. This last aspect shows evidences of the natural dispersion that is imprinted to the model with only 50% of the sequences more near, when we use all the sequences.

### 3.4.3 Conclusion

In this application we show how to use the measure  $d_s$  to establish a notion of proximity between strains of Burkitt lymphoma/leukemia, over the alphabet  $\mathcal{A} = \{a, c, g, t\}$ , we deal with 15 strains. The state space is formed by strings that are concatenations of size 4 of elements coming from the alphabet, and the DNA sequences are identified with Markov processes of memory 4. From  $d_s$  it is also possible to propose a strategy of selection of strains, for the construction of a model that allows to describe the way the elements of the state space are organized. The measure  $d_s$  allows to select the nearest strains to build the model whose represents the majority of the strains. We estimate the transition probability of each string for any element of the alphabet  $\mathcal{A}$ . By the conception of the model it is possible to classify the strings into 27 categories, where each category contains strings with the same transition probability to elements of the alphabet, i.e. within each category, the strings are stochastically equivalent. Comparing the model constructed from the closest strains to the model with all the strains, we noticed that the categories practically double. An open question is to be able to quantify with some level of significance the impact of the inclusion of each strain on the model, as the quantity  $\mathcal{S}$  increases. An answer in that line would allow to classify the different possible models, given the 15 strains.

# 3.5 Third Case Study: Production of Alcohol Fuel

The case investigated in this section is the process of distillation of sugar cane. for the production of fuel. After the fermentation, the product is heated in the same batch and immediately it is introduced in two different columns, in order to extract the hydrated alcohol. Those columns work under the same specifications. For each column i, i = 1, 2, ...there are 5 variables collected in a period of 1 month, one observation by minute t which means a total of n = 44643 observations. Those variables are (1) alcoholic contents  $M_t^i(1)$ in INPM degrees (the alcohol weight percentage in a hydro-alcoholic solution), (2) fill level  $M_t^i(2)$  in percentage, (3) entrance temperature  $M_t^i(3)$  in degrees Celsius, (4) exit temperature  $M_t^i(4)$  in degrees Celsius, (5) vapor pressure  $M_t^i(5)$  in  $kgf/cm^2$  (kilogramforce per square centimeter). In García, González-López e Andrade (2017) a preliminary study is carried out to determine whether or not the processes are different. The processes are compared by mean of the joint comparison of the 5 variables of column 1 with the 5 variables of column 2, the criterion used is also BIC-based, but it is not a measure. That study already points to divergence between the columns. The purpose of this application is to determine which of these variables most contribute to the divergence. To do that, we use the local measure  $d_s$ , for each pair of variables, identifying precisely which are the configurations (strings) that lead to the discrepancy between the processes. This information points out which are the mayor problems that must be corrected to avoid divergences.

For each time t and each column i = 1, 2 define  $X_t^i(j) = 1$  if the value of  $M_t^i(j) > M_{t-1}^i(j)$  and  $X_t^i(j) = 0$  otherwise (then  $\mathcal{A} = \{0, 1\}$ ) for j = 1, ..., 5. In terms of the joint process, i.e. with alphabet  $\mathcal{A} = \{0, 1\}^5$  and using memory M = 1 (by the rule described above) we obtain the results detailed in table 24. To compare we have included the results with memory M = 2.

	$d_s$		String $(s)$		
M = 1	1.01564		(0,0,0,1,1)		
	1.26571		$(1,\!1,\!0,\!1,\!0)$		
	1.56578		$(1,\!1,\!1,\!1,\!0)$		
	2.08815		$(0,\!1,\!0,\!1,\!1)$		
	2.80355		$(1,\!1,\!0,\!1,\!1)$		
	2.82101		$(1,\!0,\!1,\!1,\!1)$		
	5.29509		$(1,\!1,\!1,\!1,\!1)$		
M = 2	$d_s$		String $(s)$		
	1.27915	(1,1)	,0,1,1)(1,1,0,1,1)		
	1.64902	(1,0)	$,\!1,\!1,\!1)(1,\!0,\!1,\!1,\!1)$		
	2.37569	(1,1)	$,\!1,\!1,\!1)(1,\!1,\!1,\!1,\!1)$		

Table 24 – For each order M, M = 1, 2, on the right we list the strings (s) in which the columns i, i = 1, 2 are considered as being different. On the left we inform the value of  $d_s$ . In bold, we highlight the cases with greater distances.

We see that according to the records in table 24, the 3 extreme cases (for each order) involve 3 elements of  $\mathcal{A}$  those are: (1, 1, 1, 1, 1), (1, 0, 1, 1, 1) and (1, 1, 0, 1, 1).

column	1	2	1	2	1	2
$a \in A$	$P(a s_0)$	$P(a s_0)$	$P(a s_1)$	$P(a s_1)$	$P(a s_2)$	$P(a s_2)$
(0, 0, 0, 0, 0)	0.00012	0.00046	0.00000	0.00000	0.00000	0.00000
(1,0,0,0,0)	0.00000	0.00000	0.00014	0.00122	0.00012	0.00000
(0, 1, 0, 0, 0)	0.00012	0.00034	0.00000	0.00000	0.00037	0.00034
(1, 1, 0, 0, 0)	0.00012	0.00160	0.00000	0.00000	0.00220	0.00271
(0, 0, 1, 0, 0)	0.00000	0.00011	0.00027	0.00052	0.00000	0.00000
(1,0,1,0,0)	0.00048	0.00034	0.00217	0.00226	0.00012	0.00000
(0, 1, 1, 0, 0)	0.00000	0.00011	0.00000	0.00000	0.00012	0.00023
(1, 1, 1, 0, 0)	0.00227	0.00239	0.00054	0.00000	0.00024	0.00136
(0, 0, 0, 1, 0)	0.00084	0.00023	0.00217	0.00417	0.00098	0.00079
(1,0,0,1,0)	0.00358	0.00445	0.00666	0.03056	0.00415	0.00701
(0, 1, 0, 1, 0)	0.00131	0.00274	0.00068	0.00017	0.00598	0.00724
(1, 1, 0, 1, 0)	0.00836	0.05347	0.00462	0.00590	0.06798	0.08492
(0, 0, 1, 1, 0)	0.00239	0.00080	0.00747	0.00625	0.00024	0.00023
(1,0,1,1,0)	0.02233	0.01015	0.05787	0.07380	0.00354	0.00565
(0, 1, 1, 1, 0)	0.00299	0.00559	0.00109	0.00052	0.00073	0.00441
(1, 1, 1, 1, 0)	0.06294	0.08118	0.01562	0.00903	0.01684	0.05552
(0, 0, 0, 0, 1)	0.00024	0.00011	0.00014	0.00052	0.00024	0.00023
(1,0,0,0,1)	0.00107	0.00080	0.00163	0.00556	0.00110	0.00045
(0, 1, 0, 0, 1)	0.00048	0.00068	0.00027	0.00035	0.00171	0.00090
(1, 1, 0, 0, 1)	0.00084	0.00638	0.00041	0.00347	0.01855	0.01029
(0, 0, 1, 0, 1)	0.00072	0.00034	0.00149	0.00174	0.00012	0.00011
(1,0,1,0,1)	0.00561	0.00319	0.01345	0.00799	0.00037	0.00068
(0, 1, 1, 0, 1)	0.00060	0.00125	0.00054	0.00052	0.00012	0.00079
(1, 1, 1, 0, 1)	0.01588	0.00946	0.00367	0.00243	0.00525	0.00746
(0, 0, 0, 1, 1)	0.00478	0.00456	0.01793	0.01858	0.00842	0.00339
(1,0,0,1,1)	0.02723	0.03238	0.07390	0.20455	0.03662	0.03856
(0, 1, 0, 1, 1)	0.00836	0.01767	0.00421	0.00886	0.05029	0.03426
(1, 1, 0, 1, 1)	0.06330	0.22780	0.03152	0.07380	0.61089	0.43250
(0, 0, 1, 1, 1)	0.01696	0.00775	0.05910	0.04150	0.00574	0.00328
(1,0,1,1,1)	0.18094	0.08186	0.51691	0.38999	0.02124	0.03290
(0, 1, 1, 1, 1)	0.02257	0.02816	0.01386	0.00955	0.00513	0.02001
(1, 1, 1, 1, 1)	0.54258	0.41364	0.16166	0.09620	0.13060	0.24378

Table 25 – Transition probabilities from the string s to each element of the alphabet  $\mathcal{A} = \{0, 1\}^5$  (listed on the left), for each process (column 1 and column 2) where  $s = s_0, s_1, s_2; s_0 = (1, 1, 1, 1, 1), s_1 = (1, 0, 1, 1, 1)$  and  $s_2 = (1, 1, 0, 1, 1)$  with  $d_s > 1$ , for the case M = 1 - see table 24.

Given the strings that show greater values of  $d_s$  (being  $d_s > 1$ ), according to table 24 and for the case of order 1, we see in table 25 in a comparative way the transition probabilities of the processes (columns 1 and 2). We note that the difference between the probabilities is perceptible, although it is much more practical to use  $d_s$  to detect such discrepancies.

In this work, when defining the order of the processes we follow a line of reasoning

similar to the introduced and detailed in García, González-López e Hirsh (2016),  $M = \lfloor \frac{(\log(n) - \log(\log(n))}{\log(|\mathcal{A}|)} \rfloor - 1 = 10$ . Instead of using n = 44643, which is the original size of the sample, we use n = 43643, since we discard the 500 initial observations and the 500 final observations.

Interval of $d_s$ values	Strings $(s)$			
[1,2)	1111100010, 0111111001, 1111110011, 1100111000,			
	0111111011, 1101111110, 0011111101, 111001100			
	1110000111, 1100100111, 1110111100, 1000001111,			
	1110011000, 1111001001, 1111011110, 1110010011,			
	1100001111, 1100001110, 1110001101, 1110111110,			
	0111111110, 0011111110, 0011111000, 1111100000,			
	1110011111, 1111100110			
[2,3)	1111111011, 1011111100, 1111100011, 1110001110,			
	1100111110, 1110000110			
[3,4)	0000111111, 1111110010, 1110011110, 0001111110,			
	111111100			
[4,5)	1111001110, 0111111000, 1111111001			
[5, 6)	011111111, 1111000010			
[6,7)	1111000110, 1001111110			
[8,9)	0011111111			
[10, 11)	0001111111			
[19, 20)	111111110			
[41, 42]	111111000			

Table 26 – On the right we list the strings (s) in which the columns i, i = 1, 2 are considered as being different, according to the variable  $X_t^i(1)$  built from the *alcoholic content*. On the left we inform the intervals in which the values of  $d_s$  are included. The strings are listed according to the  $d_s$  values, in an increasing order from left to right and from top to bottom.
Interval of $d_s$ values	Strings $(s)$
[1,2)	1011100000, 0110001000, 1110111100, 0011001000,
	1100010000, 1100100100, 0111110000, 1001100000,
	1011000100, 0000010111, 0111010000, 0000100001,
	0111100000, 0000001100, 0111111110, 0000111000,
	0000111100, 0110110000, 0001010000, 1110100000,
	0110100000, 1101111111, 0101101000, 0000110100,
	0001001000, 1100100000, 1100001000, 1101100000,
	0010111111, 1110000100, 0100100000, 1011000000,
	0110010000, 0000100100, 1100000001, 1010010000,
	1010100000, 0011111101, 0011100000
[2,3)	1000001000, 0101000000, 1111100000, 0001100000
[3,4)	0010100000, 1101000000, 0001000000, 1001000000,
	0011000000
[4,5)	100000000, 0111000000
[5,6)	1010000000, 1111000000, 0000000001
[6,7)	0110000000, 0010000000, 0100000000, 011111111
[8,9)	000000011
[10, 11)	0000111111, 1110000000, 0001111111
[11, 12)	110000000
[12, 13)	0000000111, 001111111
[13, 14]	0000001111, 0000011111

Table 27 – On the right we list the strings (s) in which the columns i, i = 1, 2 are considered as being different, according to the variable  $X_t^i(2)$  built from the *fill level*. On the left we inform the intervals in which the values of  $d_s$  are included. The strings are listed according to the  $d_s$  values, in an increasing order from left to right and from top to bottom.

In tables 26 and 27 we show the strings (s) where the marginal processes 1 (alcohol) and 2 (fill level) are different, when compared the columns 1 and 2. We also expose the intervals of  $d_s$  values where such strings show discrepancies. In table 28 we show the strings where the columns behave differently in relation to the temperatures's records. Analogously to the previous cases, on the left we expose the intervals of values of  $d_s$  where the magnitudes of the discrepancies are recorded. About variable 5 (vapor pressure), no records of discrepancies are detected.

	Interval of $d_s$ values	Strings $(s)$
	[1,2)	1111110100, 1111010011, 0111111100, 0000001111,
		0000111111, 1111100100, 1111110011, 0111111000,
		1111111011, 1111100111, 1111001100, 1111100000,
j = 3		0011111111, 00011111111, 1111101100
0	[2,3)	111111001
	[3, 4)	111111111
	[4, 5)	1111110000, 1111111110
	[7, 8)	1111111000, 1111111100
	Interval of $d_s$ values	Strings $(s)$
	[1, 2)	0111111001, 1100110000, 1110100000, 0000011111,
		0100111100, 1111011111, 0001000001, 1000000100,
		1000100000, 0100001100, 0000000011, 1010000000,
		0110000000, 1111111100, 1000000000, 0100000000
j = 4	[2,3)	0000011000, 0010000000, 0000000001
	[3, 4)	1100000000, 1111100000
	[4, 5)	1111000000, 1110000000
	[5, 6)	111111111
	[6,7)	1111110000
	[8,9)	1111111000

Table 28 – On the right we list the strings (s) in which the columns i, i = 1, 2 are considered as being different, according to the variable  $X_t^i(j)$ , built from the *temperatures* recorded at the entrance (j = 3) and at the exit (j = 4). On the left we inform the intervals in which the values of  $d_s$  are included. The strings are listed according to the  $d_s$  values, in an increasing order from left to right and from top to bottom.

The fill level contributes with 66 strings with  $d_s > 1$ , 39 of them with  $d_s \in [1, 2)$ and 2 strings with  $d_s \in [13, 14)$ ; the alcoholic content contributes with 48 strings with  $d_s > 1$ , 26 of them with  $d_s \in [1, 2)$  and 2 strings with  $d_s \in [19, 42)$ . The temperature of entrance contributes with 21 strings with  $d_s > 1$ , 15 of them with  $d_s \in [1, 2)$  and 2 strings with  $d_s \in [7, 8)$ . The temperature of exit contributes with 26 strings with  $d_s > 1$ , 16 of them with  $d_s \in [1, 2)$  and 2 strings with  $d_s \in [6, 9)$ , see table 28. Tables 29 and 30 show for each variable the 4 strings that expose the highest values of  $d_s$ . Also in these tables we show the transition probabilities that explain the discrepancies detected by  $d_s$ . The measure  $d_s$  already captures that the performance of the processes conducted in the columns are essentially different, in the list of strings detailed in the tables 26, 27 and 28. More precisely, if we focus on the four most extreme cases of each variable, we can see how this measure identifies the differences between the transition probabilities when comparing the values between the columns. See for instance tables 29 and 30. We can note that when comparing the processes marginally, variable to variable, there is a tendency to show strings s with large values of  $d_s$ , which are made up of runs of zeros (or runs of ones) followed by runs

			P(0 s)	P(1 s)
j = 1	s = 1111111000	column 1	0.80362	0.19638
		column 2	0.29775	0.70225
	$(d_s = 41.81420)$		1	1
	1111111110	. 1 1	0.04070	0.05090
	s = 111111110	column 1	0.94970	0.05030 0.01700
	$(d_s = 19.03538)$	column 2	0.78208	0.21792
	c = 0001111111	column 1	0 02068	0 07032
	5 – 0001111111	column 2	0.02908	0.91032
	$(d_s = 10.47809)$		0.10210	0.01102
	s = 0.0111111111	column 1	0.07895	0 92105
	5 - 001111111	column 2	0.20325	0.79675
	$(d_s = 8.95013)$			
			P(0 s)	P(1 s)
j = 2	s = 0000011111	column 1	$P(0 s) \\ 0.01379$	P(1 s) 0.98621
j=2	s = 0000011111	column 1 column 2	$\begin{array}{c} P(0 s) \\ 0.01379 \\ 0.18511 \end{array}$	$ \begin{array}{c} P(1 s) \\ 0.98621 \\ 0.81489 \end{array} $
<i>j</i> = 2	s = 0000011111 $(d_s = 13.81704)$	column 1 column 2	$\begin{array}{c} P(0 s) \\ 0.01379 \\ 0.18511 \end{array}$	$ \begin{array}{c} P(1 s) \\ 0.98621 \\ 0.81489 \end{array} $
j = 2	s = 0000011111 $(d_s = 13.81704)$ s = 0000001111	column 1 column 2	$\begin{array}{c} P(0 s) \\ 0.01379 \\ 0.18511 \end{array}$	$\begin{array}{c} P(1 s) \\ 0.98621 \\ 0.81489 \end{array}$
j = 2	s = 0000011111 $(d_s = 13.81704)$ s = 0000001111	column 1 column 2 column 1 column 2	$\begin{array}{c} P(0 s) \\ 0.01379 \\ 0.18511 \\ 0.00771 \\ 0.16183 \end{array}$	$\begin{array}{c} P(1 s) \\ 0.98621 \\ 0.81489 \\ 0.99229 \\ 0.83817 \end{array}$
j = 2	$s = 0000011111$ $(d_s = 13.81704)$ $s = 0000001111$ $(d_s = 13.77702)$	column 1 column 2 column 1 column 2	$\begin{array}{c} P(0 s) \\ 0.01379 \\ 0.18511 \\ 0.00771 \\ 0.16183 \end{array}$	$\begin{array}{c} P(1 s) \\ 0.98621 \\ 0.81489 \\ 0.99229 \\ 0.83817 \end{array}$
j = 2	$s = 0000011111$ $(d_s = 13.81704)$ $s = 0000001111$ $(d_s = 13.77702)$ $s = 0011111111$	column 1 column 2 column 1 column 2	$\begin{array}{c} P(0 s) \\ 0.01379 \\ 0.18511 \\ 0.00771 \\ 0.16183 \end{array}$	$\begin{array}{c} P(1 s) \\ 0.98621 \\ 0.81489 \\ 0.99229 \\ 0.83817 \\ 0.96186 \end{array}$
j = 2	$s = 0000011111$ $(d_s = 13.81704)$ $s = 0000001111$ $(d_s = 13.77702)$ $s = 001111111$	column 1 column 2 column 1 column 2 column 1 column 2	$\begin{array}{c} P(0 s) \\ 0.01379 \\ 0.18511 \\ 0.00771 \\ 0.16183 \\ 0.03814 \\ 0.24505 \end{array}$	$\begin{array}{c} P(1 s) \\ 0.98621 \\ 0.81489 \\ 0.99229 \\ 0.83817 \\ 0.96186 \\ 0.75495 \end{array}$
j = 2	$s = 0000011111$ $(d_s = 13.81704)$ $s = 0000001111$ $(d_s = 13.77702)$ $s = 001111111$ $(d_s = 12.13914)$	column 1 column 2 column 1 column 2 column 1 column 2	$\begin{array}{c} P(0 s) \\ 0.01379 \\ 0.18511 \\ 0.00771 \\ 0.16183 \\ 0.03814 \\ 0.24505 \end{array}$	$\begin{array}{c} P(1 s)\\ 0.98621\\ 0.81489\\ 0.99229\\ 0.83817\\ 0.96186\\ 0.75495\\ \end{array}$
<i>j</i> = 2	$s = 0000011111$ $(d_s = 13.81704)$ $s = 0000001111$ $(d_s = 13.77702)$ $s = 0011111111$ $(d_s = 12.13914)$ $s = 000000111$	column 1 column 2 column 1 column 2 column 1 column 2	$\begin{array}{c} P(0 s) \\ 0.01379 \\ 0.18511 \\ 0.00771 \\ 0.16183 \\ 0.03814 \\ 0.24505 \\ 0.00312 \end{array}$	$\begin{array}{c} P(1 s) \\ 0.98621 \\ 0.81489 \\ \end{array} \\ \begin{array}{c} 0.99229 \\ 0.83817 \\ \end{array} \\ \begin{array}{c} 0.96186 \\ 0.75495 \\ \end{array} \\ \begin{array}{c} 0.99688 \end{array} \end{array}$
<i>j</i> = 2	$s = 0000011111$ $(d_s = 13.81704)$ $s = 0000001111$ $(d_s = 13.77702)$ $s = 0011111111$ $(d_s = 12.13914)$ $s = 0000000111$	column 1 column 2 column 1 column 2 column 1 column 2 column 1 column 2	$\begin{array}{c} P(0 s)\\ 0.01379\\ 0.18511\\ \end{array}\\ 0.00771\\ 0.16183\\ 0.03814\\ 0.24505\\ \end{array}\\ 0.00312\\ 0.13412 \end{array}$	$\begin{array}{c} P(1 s)\\ 0.98621\\ 0.81489\\ 0.99229\\ 0.83817\\ 0.96186\\ 0.75495\\ 0.99688\\ 0.86588\\ \end{array}$

Table 29 – Transition probabilities from the string s to 0 and 1, assigned by the two processes, given by columns 1 and 2, respectively, j = 1 records the results for the *alcoholic content*, j = 2 indicates the results for the *fill level*.

			P(0 s)	P(1 s)
j = 3	s = 1111111100	$column \ 1$	0.43993	0.56007
		$\operatorname{column} 2$	0.24449	0.75551
	$(d_s = 7.76765)$		1	
	1111111000	1 1	0.0007	0.96779
	s = 111111000	column 1	0.03227	0.30773
	$(d_s = 7.53980)$	column 2	0.28340	0.71034
	1111111110	1 1	0 70074	0.00100
	s = 1111111110	column 1	0.70874	0.29126
	(d - 4.78082)	column 2	0.58032	0.41968
	$(u_s - 4.10302)$			
	s = 1111110000	column 1	0.77159	0.22841
		$\operatorname{column} 2$	0.35135	0.64865
	$(d_s = 4.16771)$		I	
			P(0 s)	P(1 s)
j = 4	s = 1111111000	column 1	0.76239	0.23761
j = 4	s = 1111111000	column 1 column 2	0.76239 0.92345	$0.23761 \\ 0.07655$
j = 4	s = 1111111000 $(d_s = 8.59985)$	column 1 column 2	0.76239 0.92345	0.23761 0.07655
j = 4	s = 1111111000 $(d_s = 8.59985)$	column 1 column 2	0.76239 0.92345	0.23761 0.07655
j = 4	s = 111111000 $(d_s = 8.59985)$ s = 111110000	column 1 column 2 column 1	0.76239 0.92345 0.78345	0.23761 0.07655 0.21655
j = 4	s = 111111000 $(d_s = 8.59985)$ s = 111110000	column 1 column 2 column 1 column 2	0.76239 0.92345 0.78345 0.92347	0.23761 0.07655 0.21655 0.07653
j = 4	s = 111111000 $(d_s = 8.59985)$ s = 111110000 $(d_s = 6.03633)$	column 1 column 2 column 1 column 2	0.76239 0.92345 0.78345 0.92347	0.23761 0.07655 0.21655 0.07653
j = 4	s = 111111000 $(d_s = 8.59985)$ s = 111110000 $(d_s = 6.03633)$ s = 111111111	column 1 column 2 column 1 column 2	0.76239 0.92345 0.78345 0.92347	0.23761 0.07655 0.21655 0.07653
j = 4	s = 111111000 $(d_s = 8.59985)$ s = 111110000 $(d_s = 6.03633)$ s = 111111111	column 1 column 2 column 1 column 2 column 1 column 2	0.76239 0.92345 0.78345 0.92347 0.18722 0.14170	0.23761 0.07655 0.21655 0.07653 0.81278 0.85830
j = 4	s = 111111000 $(d_s = 8.59985)$ s = 111110000 $(d_s = 6.03633)$ s = 111111111 $(d_s = 5.38372)$	column 1 column 2 column 1 column 2 column 1 column 2	0.76239 0.92345 0.92345 0.92347 0.18722 0.14170	0.23761 0.07655 0.21655 0.07653 0.81278 0.85830
j = 4	$s = 111111000$ $(d_s = 8.59985)$ $s = 111110000$ $(d_s = 6.03633)$ $s = 111111111$ $(d_s = 5.38372)$ $s = 1110000000$	column 1 column 2 column 1 column 2 column 1 column 2	0.76239 0.92345 0.92345 0.92347 0.92347 0.18722 0.14170	0.23761 0.07655 0.21655 0.07653 0.81278 0.85830
j = 4	$s = 111111000$ $(d_s = 8.59985)$ $s = 111110000$ $(d_s = 6.03633)$ $s = 111111111$ $(d_s = 5.38372)$ $s = 1110000000$	column 1 column 2 column 1 column 2 column 1 column 2	$\begin{array}{c} 0.76239\\ 0.92345\\ 0.92345\\ 0.92347\\ 0.92347\\ 0.18722\\ 0.14170\\ 0.67971\\ 0.85825\\ \end{array}$	0.23761 0.07655 0.21655 0.07653 0.81278 0.85830 0.32029 0.14175

Table 30 – Transition probabilities from the string s to 0 and 1, assigned by the two processes, given by columns 1 and 2, respectively, j = 3 records the results for the *temperature of entrance*, j = 4 indicates the results for the *temperature of exit*.

of ones (or runs of zeros). This shows that the processes react differently with different transition probabilities to atribute the next symbol a zero (or one). We can visualize these discrepancies in table 29 and 30. For example if we consider the variable j = 2 and the string s = 0000011111, P(1|s) = 0.98621 for column 1 while P(1|s) = 0.81480 for column 2. Similar behaviour is observed on the other cases for j = 2; on a long sequence of ones, the probability of the next symbol be one is always larger for column 1 than for column 2. The interpretation in practice is the following, a long sequence of ones means that the value of the variable keep increasing for long time while a long sequence of zeros means that the variable is in a decreasing mode for a long time. The behaviour observed corresponds with differences on the reaction velocity on the self-regulation mechanism of the variable. The self-regulation mechanism for j = 2 is faster for column 2 than for column 1. A similar behaviour is observed for the variables j = 1 and j = 3. For variable j = 4 we observe an inverse behaviour, column 2 seems to self-regulate faster than column 1.

#### 3.5.1 Conclusion

In the application we address a real problem that is to compare two lines of production of alcohol fuel. Our purpose goes beyond deciding whether or not production lines are equivalent. Through the measure explored in this section, we identify which are the strings of the processes that mark discrepancies between the lines of production. We do it variable to variable and jointly. We were able to identify from a range of 5 variables, which are the ones that contribute most in terms of discrepancies. For this, two situations can be observed: the first one is to determine the number of strings with which each variable contributes and the second is when a variable contributes with very marked discrepancies, in terms of quantity of strings (66) and the *alcoholic content* is the first in terms of magnitude ( $d_s > 41$ ). The variables recording *temperatures* show each of them about 20 strings that expose discrepancies between the lines of production and the magnitudes of the differences are similar between them ( $1 < d_s < 9$ ). On the other hand in relation to the *vapor pressure* the production lines can be considered equivalent.

## 4 Robust Sample Selection Strategy

In section 4.1 we introduce the problem, In section 4.2. we propose a robust procedure that has the highest possible robustness rate (50%). We apply this procedure to the dataset of the financial sector in section 4.3.

## 4.1 Sample Selection

In this chapter we propose a robust strategy for selecting samples which come from stochastic processes in discrete time taking values on a finite alphabet. Suppose that it is suspected that although most of the samples come from the same stochastic law, a minor proportion of them may come from different laws. So, if our purpose is to discover the prevailing stochastic law, we should be able to identify the samples associated with that law. Being able to identify those samples with more discrepant behavior or associated with the alternatives laws.

In this chapter, we consider m samples  $\{x_{i,1}^{n_j}\}_{i=1}^m$  coming from Markov processes of order M, over the finite alphabet  $\mathcal{A}$  and state space  $\mathcal{S}$ . Assume that the samples follow the laws denoted by  $\{P_i\}_{i=1}^m$  where  $k^*$  of these samples follow the same law, say P. Our goal is to establish a sample selection procedure which identifies samples following the same law P. We want to ensure that this procedure is capable of selecting samples with law P when  $m - k^* < k^*$ . Noting that those  $m - k^*$  samples can be generated by laws that are different of P. A procedure with such characteristics will be able to withstand a fraction of alternative laws (different of P) less than q% of the total number of samples, allowing its robustness. A method of selecting samples with this ability should involve some notion of proximity or a distance between the samples that reflects the generating law. It is desirable that, when the sample sizes are large enough such a distance tends to zero if both samples are generated by the same law and also it is desirable that such a distance be arbitrarily large when the laws generating the samples are different.

In the next section we introduce and investigate the robust procedure of sample selection, using the dmax notion.

### 4.2 Results

First, given a collection of samples, we use the notion introduced in the definition 3.2.9 to quantify the proximity between a specific sample in the collection  $x_{i,1}^n$  and the rest of the members of the collection of samples. This is a way to compare a sample with a group

of samples. In conceptual terms, establishing the notion of proximity between samples is equivalent to quantifying the proximity between the stochastic laws that generate such samples.

**Definition 4.2.1.** Given a finite collection  $\{x_{i,1}^{n_j}\}_{i=1}^m$  of samples from the processes  $\{X_{j,t}\}_{j=1}^m$ with probabilities  $\{P_j\}_{j=1}^m$ , over the finite alphabet  $\mathcal{A}$  with state space  $\mathcal{S} = \mathcal{A}^M$   $(M < \infty)$ . For a fixed  $i \in \{1, 2, ..., m\}$  define

$$V(x_{i,1}^{n_i}) = median\{dmax(x_{i,1}^{n_i}, x_{j,1}^{n_j}) : j \neq i, 1 \leq j \leq m\}.$$

Where, given a sequence  $\{z_j\}_{j=1}^l$ ,  $median\{z_j, 1 \leq j \leq l\} = z_{(k+1)}$  if l = 2k + 1 and  $median\{z_j, 1 \leq j \leq l\} = \frac{z_{(k)} + z_{(k+1)}}{2}$  if l = 2k, for k an integer and  $z_{(j)}$  denoting the jth order statistic of the collection  $\{z_j\}_{j=1}^l$ .

We can use  $V(x_{i,1}^{n_i})$  to compare stochastic laws, under the assumption that  $\min\{n_1, \dots, n_m\}$  is large enough. For instance, if all the samples follow the same law, according to remark 3.2.10,

$$V(x_{i,1}^{n_i}) \xrightarrow[\min\{n_1, \cdots, n_m\} \to \infty]{} 0, \ i = 1, \dots, m.$$

In a more realistic case, we may have that not all the samples in the collection have the same law. Even in this case, we will have that, if  $J_i = \{j : 1 \leq j \leq m, P_j = P_i\}$ , then  $V(x_{i,1}^{n_i})$  goes to zero when  $\min\{n_1, \dots, n_m\}$  goes to infinity, if and only if  $|J_i| > \lceil \frac{m}{2} \rceil$ , where  $\lceil r \rceil$  represents the smallest integer which is also larger than r. From now on we will use the following notation. Under the assumptions of definition 4.2.1, for each  $i, 1 \leq i \leq m$ ,  $\xi_i = |J_i|$  (cardinal of the set  $J_i$ ).

The following result formally exposes the conditions for which  $V(x_{i,1}^{n_i})$  takes high values indicating a high proportion of samples from different stochastic laws, in comparison with the law of  $x_{i,1}^{n_i}$ .

**Theorem 4.2.2.** Under the assumptions of definition 4.2.1, for each  $i, 1 \leq i \leq m$ ,  $V(x_{i,1}^{n_i}) \xrightarrow{}_{\min\{n_1, \cdots, n_m\} \to \infty} \infty$ , if, and only if,  $\xi_i \leq \lceil \frac{m}{2} \rceil$ .

**Proof.** To simplify the exposition of the proof we will assume that all the samples have the same size *n*. Denote by  $z_j^n = dmax(x_{i,1}^n, x_{j,1}^n), \ j \neq i, 1 \leq j \leq m$  (i.e. we have m-1 values of  $z_j^n$ ) then  $V(x_{i,1}^n) = z_{(k+1)}^n$  if m-1 = 2k+1 and  $V(x_{i,1}^n) = \frac{z_{(k)}^n + z_{(k+1)}^n}{2}$  if m-1 = 2k. For the first part of the proof, for the case *m* even i.e. m-1 = 2k+1 note that, if  $V(x_{i,1}^n) \to_{n\to\infty} \infty$ , then,  $z_{(k+1)}^n \to_{n\to\infty} \infty$ , and,  $z_{(v)}^n \to_{n\to\infty} \infty$ , for  $v = k+1, k+2, \cdots, m-1$  which correspond to k+1 samples of the total of m = 2k+2 samples. Also, for each  $v = k+1, \cdots, m-1$  if j(v) is the original index of the sample,  $dmax(x_{i,1}^n, x_{j(v),1}^n) \to_{n\to\infty} \infty$ . As seen in remark 3.2.10, this means that  $P_i \neq P_{j(v)}$  and  $j(v) \notin J_i$ . That is  $\xi_i \leq k+1 = \lceil \frac{m}{2} \rceil$ . For the second part of the proof, consider the complementary set of  $J_i$ , say  $\{j(v)\}$ . For each  $j(v) \notin J_i$ ,  $P_i \neq P_{j(v)}$  and according to the remark 3.2.10, this means that  $z_{j(v)}^n \to_{n\to\infty} \infty$ . Then, the condition  $\xi_i \leq \lceil \frac{m}{2} \rceil$  implies that more than 50%  $z_j^n$  values go to infinity when n goes to infinity and so does the median  $V(x_{i,1}^n)$ .

That is,  $V(x_{i,1}^{n_i})$  takes large values if, and only if, more than 50% of the available samples exhibit different laws, compared to the law of  $x_{i,1}^{n_i}$ . In a collection of stochastic processes we will admit the existence of a predominant stochastic law. This supposition is essential to ensure that it is possible to extract from the collection of samples one specific sample which is associated with the predominant law.

**Definition 4.2.3.** Under the assumptions of definition 4.2.1  $P_{j^*}$ , for some  $1 \leq j^* \leq m$ , is denoted as the majority law in  $\{X_{j,t}\}_{j=1}^m$  if  $\xi_{j^*} > \xi_j, \forall j \neq j^*, j \in \{1, \dots, m\}$ .

**Remark 4.2.4.** If  $\xi_{j*} \leq \lceil \frac{m}{2} \rceil$ , then, for every  $i, 1 \leq i \leq m, \xi_i \leq \lceil \frac{m}{2} \rceil$  and from theorem 4.2.2 we will have that  $V(x_{i,1}^{n_i}) \xrightarrow{\longrightarrow} \infty$ . This mean that if the majority law does not correspond to at least half of the samples, then any procedure based on the V statistic will be inconclusive. In this work we will only consider the case on which  $\xi_{j*} > \lceil \frac{m}{2} \rceil$ .

Suppose we have *m* samples  $\{x_{j,1}^{n_j}\}_{j=1}^m$ , and we want to identify one of them which can be considered as being the nearest sample to all the other samples. We will see that finding the nearest sample is equivalent to identifying the majority law. The procedure described below allows the ordering of the samples of the set  $\{x_{j,1}^{n_j}\}_{j=1}^m$ . That is, once this procedure is executed on the set, let's say by doing  $\mathcal{P}(\{x_{j,1}^{n_j}\}_{j=1}^m)$  we will have the ordered set:  $\{x_{(j),1}^{n_{(j)}}\}_{j=1}^m$ , so the sample in the *i* position is  $\mathcal{P}(\{x_{j,1}^{n_j}\}_{j=1}^m)(i) = x_{(i),1}^{n_{(i)}}$ .

**Procedure 4.2.5.**  $(\mathcal{P}(\{x_{j,1}^{n_j}\}_{j=1}^m))$ 

- a. imput (sample set):  $\{x_{i,1}^{n_j}\}_{j=1}^m$ ;
- *i.* for each  $j \in \{1, 2, \dots, m\}$  compute  $V(x_{j,1}^{n_j})$ ;
- ii.  $v_{(i)}$  the ith order statistic of  $\{V(x_{i,1}^{n_j}), 1 \leq j \leq m\}$ , for  $i = 1, \cdots m$ ;
- iii. denote by  $x_{(i),1}^{n_{(i)}}$  the sample related to  $v_{(i)}$ , for  $i = 1, \dots, m$ ;
- b. output (ordered sample set):  $\{x_{(j),1}^{n_{(j)}}\}_{j=1}^m$

The procedure takes a set of samples and return the set of samples ordered by the value of V.

**Theorem 4.2.6.** Consider the samples  $\{x_{j,1}^{n_j}\}_{j=1}^m$  coming from the processes  $\{X_{j,t}\}_{j=1}^m$  under the assumptions of definition 4.2.3, set  $n = \min\{n_1, \cdots, n_m\}$ . Suppose that  $\xi_{j^**} > \lceil \frac{m}{2} \rceil$ 

and define  $x_1^{n_x} = \mathcal{P}(\{x_{j,1}^{n_j}\}_{j=1}^m)(i)$ , where  $\mathcal{P}$  is the ordering procedure 4.2.5. Then, for n large enough and  $i \leq \xi_{j^**}, x_1^{n_x}$  will be a sample of the majority law.

**Proof.** Fix a value M > 0, if  $\xi_{j^**} > \lceil \frac{m}{2} \rceil$  then, for each l such that  $P_l$  is not the majority law,  $\xi_l < \frac{m}{2}$  and from theorem 4.2.2  $V(x_{l,1}^{n_l}) \xrightarrow[\min\{n_1, \cdots, n_m\} \to \infty]{} \infty$ . This implies that there is  $N_1$  such that if  $\min\{n_1, \cdots, n_m\} \ge N_1$  then,  $V(x_{l,1}^{n_l}) > M$  for all l such that  $P_l$  is not the majority law.

On the other hand if  $P_k$  is the majority law, then,  $V(x_{k,1}^{n_k}) \xrightarrow{\min\{n_1, \dots, n_m\} \to \infty} 0$ , and there is  $N_2$  such that, if  $\min\{n_1, \dots, n_m\} \ge N_2$ , then,  $V(x_{k,1}^{n_k}) < M$  for all k such that  $P_k$  is the majority Law. Now, for  $\min\{n_1, \dots, n_m\} \ge \max\{N_1, N_2\}$ ,  $V(x_{k,1}^{n_k}) < M$  if  $P_k$  is the majority law and  $V(x_{l,1}^{n_l}) > M$  if  $P_l$  is not the majority law. Then, for  $n = \min\{n_1, \dots, n_m\}$ large enough on the ordering procedure 4.2.5, the samples generated from the majority law will precede the samples coming from other distributions.

The next section shows a case study of similarity and discrepancy, on financial series. It is worth noting that according to our knowledge with a possible exception see García, González-López e Viola (2014), there is not any other method of selecting samples in stochastic processes in the literature. García, González-López e Viola (2014) expressed a VLMC structure for selecting samples in stochastic processes but in our case only a Markovian behavior is required.

### 4.3 Case Study: Daily Trading Volume

Series	Description
BBDC4	Banco Bradesco SA Preference Shares (private bank)
BVMF3	B3 S.A. – Brasil, Bolsa, Balcão
	(public company of trading and clearing services)
BBAS3	Banco do Brasil (state owned bank)
ITUB4	Itau Unibanco Holding SA Preference Shares (private bank)

In this section we inspect four financial series, in table 31 we give a general description of the series.

Table 31 – Four financial series with sample size n=1446. Period: 3 January 2012- 3 November 2017.

The proposed discretization (just with daily up-down) is a first approach for liquidity modelling. By simplicity BBDC4 is associated with j = 1, BVMF3 (j = 2), BBAS3 (j = 3) and ITUB4 (j = 4) and for each j = 1, 2, 3, 4, we codify the sequences in the following way:  $w_{j,t} = 1$ , if the volume negotiated on day t is greater than the one negotiated on day t-1 and  $w_{j,t} = 0$  otherwise, so the alphabet is  $\mathcal{A} = \{0, 1\}$ . We wish to establish if the behavior of the four series  $W_{1,t}, W_{2,t}, W_{3,t}$  and  $W_{4,t}$  represented by the samples  $\{w_{j,1}^n\}_{j=1}^4$  can be considered similar or not, this is discussed in subsection 4.3.1. In subsection 4.3.2 we apply the robust procedure 4.2.5 for selection of subsamples and in subsection 4.3.3 we re-evaluate the results of subsection 4.3.1 using the subsamples selected by mean of the robust procedure.

#### 4.3.1 Proximity between the Full Series

Usually in Markovian processes the maximum memory to be considered in practical terms is limited by the size of the sample n and the size of the alphabet  $\mathcal{A}$ . Thus, if the alphabet  $\mathcal{A}$  has size  $|\mathcal{A}|$  the memory M is such that  $M \leq \lfloor \log_{|\mathcal{A}|}(n) \rfloor - 1$ , where  $\lfloor z \rfloor$ is the greatest integer less than or equal to z. The series treated here are sequences of size n = 1446 and  $|\mathcal{A}| = 2$ , then the memory is  $M \leq 9$ . Since these series are taken during the workweek days, it is natural to consider M = 5, 10, 15, etc. From the limitation mentioned above we have adopted M = 5.

jackslash k	1	2	3	4
1	-	1.00574	0.55082	0.78985
2	-	-	1.50145	0.55436
3	-	-	-	0.85719
4	-	-	-	-
1	-	0.12342	0.10213	0.10733
2	-	-	0.17352	0.09399
3	-	-	-	0.17949
4	-	-	-	-

Table 32 –  $dmax_{j,k}$  (up) and  $dmean_{j,k}$  (down) values for  $j \neq k, j, k \in \{1, 2, 3, 4\}$ . Order M = 5. See equations 4.1 and 4.2.

Given that the measure  $d_s$  can be computed string by string, table 32 shows two global concepts to compare the series, these are

$$dmax_{j,k} = \max\{d_s(w_{j,1}^n, w_{k,1}^n), s \in \mathcal{S}\}$$
(4.1)

and

$$dmean_{j,k} = mean\{d_s(w_{j,1}^n, w_{k,1}^n), s \in \mathcal{S}\},$$
(4.2)

for  $j \neq k, j, k \in \{1, 2, 3, 4\}$ . We observe that in terms of the average value (*dmean*) all the series can be considered similar, since *dmean* < 1. And this is not the case of *dmax*, for *dmax* we register two discrepancies, one between processes 1 and 2 and another between processes 2 and 3. Since all the stocks belongs to the same sector, the financial services industry, it is expected to have a similar behavior. This is shown in the general result of

table 32 when we look at the average distance computed using the whole samples. But when we look at extreme values, we identify different behaviors or higher dmax values. This indicates that for liquidity risk analysis we should have a spread between them. About a detailed inspection on the performance of  $d_s$  on all the strings of the state space, we note that there is only one string  $s^* = 00100$  that is responsible for the values of dmaxgreater than 1 and there is no other string with  $d_s > 1$ , see table 33 for details.

j	1 (BBDC4)	2 (BVMF3)	3 (BBAS3)
$P(0 s^*)$	0.35088	0.11364	0.44737
$P(1 s^*)$	0.64912	0.88636	0.55263

Table 33 – For the processes 1, 2 and 3, from left to right we list the transition probabilities from  $s^* = 00100$  to  $a \in \mathcal{A}$ . Order M = 5.

As it is perceived from the results reported in table 33, the series 2 (BVMF3) shows the most discrepant probabilities, when we compare the three samples. Now we test what happens with the discrepancies  $d_{s*}(w_{1,1}^n, w_{2,1}^n) = 1.00574$  and  $d_{s*}(w_{2,1}^n, w_{3,1}^n) = 1.50145$  by varying the order, let's say for M = 4 and M = 6, we infer that the discrepancy between the processes 1 (BBDC4) and 2 (BVMF3) is no longer relevant, but the discrepancy between the processes 2 (BVMF3) and 3 (BBAS3) becomes more marked, substituting the string  $s^* = 00100$  by 0100. Already if we move to an order M = 6, and in relation to these two cases also, we have that the discrepancy between the processes 1 and 2 ceases to be relevant and the discrepancy between the processes 2 and 3 continues to be relevant but with a lesser magnitude, substituting the string  $s^*$  by \*00100 for \* = 0 or = 1. The table 34 summarizes the information.

Order	M = 4	M = 5	M = 6
String	0100	$s^* = 00100$	*00100
$d_s(w_{1,1}^n, w_{2,1}^n)$	0.91468	1.00574	0.62200 (* = 0)
$d_s(w_{2,1}^n, w_{3,1}^n)$	1.85619	1.50145	1.05025 (* = 1)

Table 34 – Magnitude of  $d_s$  for three consecutive orders (*M*) from left to right, reporting the discrepancies between processes 1(BBDC4) and 2(BVMF3) and between processes 2(BVMF3) and 3(BBAS3), from top to bottom. In bold letter the values registered in table 32.

#### 4.3.2 Selecting Subsamples from the Robust Procedure

Another big deal when modeling daily financial data is setting the ideal period for calibrating the model. Let us explore different periods by selecting the more representative ones. We apply the robust procedure to each of the series j = 1, 2, 3 and 4 individually, in order to select representative fractions (in the sense indicated by theorem 4.2.6) of each of these financial processes. We divide the sample  $w_{j,1}^n$  into m = 5 disjoint parts

$$\{w_{j,1}^{n^*}, w_{j,n^{*}+1}^{2n^*}, w_{j,2n^{*}+1}^{3n^*}, w_{j,3n^{*}+1}^{4n^*}, w_{j,4n^{*}+1}^{5n^*}\}$$
(4.3)

of size  $n^* = \lfloor \frac{n}{m} \rfloor = 289$  each. By simplicity denote by  $\{x_{i,1}^{j,n^*}\}_{i=1}^m$  at the sequence given by the list 4.3, then  $x_{i,1}^{j,n^*} = w_{j,(i-1)n^*+1}^{in^*}, i = 1, \dots, m$ . Each subsample is associated with a specific temporal period, as described by table 35.

Subsample	Period
$x_{1,1}^{j,n^*}$	3 January 2012 - 7 March 2013
$x_{2,1}^{j,n^*}$	8 March 2013 - 7 May 2014
$x_{3,1}^{j,n^*}$	8 May 2014 - 6 July 2015
$x_{4,1}^{j,n*}$	7 July 2015 - 2 September 2016
$x_{5,1}^{j,n*}$	5 September 2016 - 31 October 2017

Table 35 – Temporal period associated with the subsample for j = 1, 2, 3, 4.

For each subsample  $x_{i,1}^{j,n^*}$  we compute

$$z_{i}^{j} = \text{median}\{\max\{d_{s}(x_{i,1}^{j,n^{*}}, x_{k,1}^{j,n^{*}}), s \in \mathcal{S}\}, k \neq i, 1 \leq k \leq m\},$$
(4.4)

denote by  $\{x_{(i),1}^{j,n^*}\}_{i=1}^m$  the sequence ordered by (4.4) (output of the procedure 4.2.5). Table 36 shows the results of equation (4.4) and table 37 shows the ordered blocks for each series j = 1, 2, 3, 4. By way of illustration fix j = 1, which is the process BBDC4. For this process, the procedure 4.2.5 identifies as the most representative block the fifth block (5 September 2016 - 31 October 2017) because according to the values in table 36, this block shows a value of  $z_1^5 = 0.08347$ . As the second block following this criterion we will have the block 3 (with  $z_1^5 = 0.14360$ ), and so on, as shown by the column associated with j = 1, in table 37. The column j = 1 of table 37, from up to down, arranges the subsamples in increasing order according to the magnitude of  $\{x_i^1\}_{i=1}^5$ .

j	Series	$z_1^j$	$z_2^j$	$z_3^j$	$z_4^j$	$z_5^j$
1	BBDC4	0.19210	0.14954	0.14360	0.26676	0.08347
2	BVMF3	0.24919	0.18972	0.13113	0.21829	0.15970
3	BBAS3	0.15297	0.18481	0.22455	0.13880	0.15351
4	ITUB4	0.14616	0.11654	0.22273	0.15714	0.13980

Table 36 – Results of equation (4.4), in bold letter the minimum values which select  $x_{(1),1}^{j,n^*}$ , for j = 1, 2, 3, 4.

Subsample	j = 1  (BBDC4)	j = 2  (BVMF3)	j = 3 (BBAS3)	j = 4  (ITUB4)
$x_{(1),1}^{j,n^*}$	$x_{5,1}^{1,n^*}$	$x_{3,1}^{2,n^*}$	$x_{4,1}^{3,n^*}$	$x_{2,1}^{4,n^{*}}$
$x_{(2),1}^{j,n^*}$	$x_{3,1}^{1,n*}$	$x_{5,1}^{2,n*}$	$x_{1,1}^{3,n^*}$	$x_{5,1}^{4,n^{*}}$
$x_{(3),1}^{j,n^*}$	$x_{2,1}^{1,n^*}$	$x_{2,1}^{2,n*}$	$x_{5,1}^{3,n*}$	$x_{1,1}^{4,n^{*}}$
$x_{(4),1}^{j,n^*}$	$x_{1,1}^{1,n^*}$	$x_{4,1}^{2,n^*}$	$x_{2,1}^{3,n^*}$	$x_{4,1}^{4,n^{*}}$
$x_{(5),1}^{j,n^*}$	$x_{4,1}^{1,n^{*}}$	$x_{1,1}^{2,n^*}$	$x^{3,n^{*}}_{3,1}$	$x^{4,n^{m{*}}}_{3,1}$

Table 37 – Sample selection following procedure 4.2.5, for series j = 1, 2, 3, 4.

The third period (8 May 2014 - 6 July 2015) is the most discrepant for both series: BBAS3 and ITUB4. The first period (3 January 2012 - 7 March 2013) is the most discrepant for the series BVMF3, while the fourth period (7 July 2015 - 2 September 2016) is the most discrepant for the series BBDC4. This last case reports the greater magnitude of  $z_i^j$ , when compared with all the other cases (0.26676374). The recent period, the fifth period (5 September 2016 - 31 October 2017) is the most representative between the 5 blocks of the series, in the case of BBDC4 and also this case reports the lesser magnitude of  $z_i^j$ , when compared with all the other cases (0.08346967). The third period is the most representative for the series BVMF3, the fourth in the case of BBAS3 and for the series ITUB4, the second period (8 March 2013 - 7 May 2014) is the most representative. Note that in some sense the series BBAS3 behaves temporally different in comparison with the others three series, since the period 4 (considered the most representative for BBAS3) is in all the remaining cases considered the most discrepant or the second most discrepant block. One aspect that is notorious is that series 1 (BBDC4) and 2 (BVMF3) show some coincidence in relation to the blocks of more or less representative samples, we see that blocks 3 and 5 are the most representative and blocks 1 and 4 the most discrepant in both series. About the results of table 36 and in a summarized way, we conclude that the most representative periods for each series can be identified, see also the first line of table 37. For example, thus, any forecast of the series j = 1, should seriously consider the performance of the period: 5 September 2016 - 31 October 2017, since such period has been selected as the most representative, according to the procedure 4.2.5. According to table 37 we investigate all period of series, with time series. Because of the reduced values of the alphabet  $\{0, 1\}$ , it is hard to identify some specific pattern in the graphic, so this task is easy considering the values of table 36. On the other hand, the number of data are large and to illustrate the changes, we divide each period into 4 periods. As we can see the picture of the best period (5 September 2016 - 31 October 2017), picture 6 and the worst period (7 July 2015 - 2 September 2016), picture 7 in the case of BBDC4, it is a difficulty task to identify a clear patter. But we can see in the picture 7 (8 October 2015- 7 January 2016) a very volatile period, maybe this is the reason which case the worst position of the period 7 July 2015 - 2 September 2016 (0.26676). In the picture 6, we can observe a very stable period (7 July 2017- 31 October 2017), we note that the whole period (5 September 2016 - 31 October 2017) is considered the most representative so is natural also to have high variation.



Figure 6 – Representative fifth period, 5 September 2016 - 31 October 2017 , for BDDC4 using of equation (4.4).



Figure 7 – Fourth period, 7 July 2015 - 2 September 2016 , for BDDC4 using of equation (4.4).

#### 4.3.3 Re-evaluating the Proximity between the Series

With the selected subsamples, we will proceed to re-evaluate the comparison between the financial entities in the following way. Table 38 shows the results of dmax(equation 4.1) and dmean (equation 4.2) between  $x_{(1),1}^{j,n^*}$  and  $x_{(1),1}^{k,n^*}$ , for j, k = 1, 2, 3, 4.

$j \backslash k$	1	2	3	4
1	-	0.34081	0.44222	0.67331
2	-	-	0.56085	0.34323
3	-	-	-	0.63592
4	-	-	-	-
1	-	0.09016	0.13409	0.13087
2	-	-	0.13688	0.11386
3	-	-	-	0.19879
4	-	-	-	-

Table 38 –  $dmax_{j,k}$  (up) and  $dmean_{j,k}$  (down) between the samples  $x_{(1),1}^{j,n^*}$  and  $x_{(1),1}^{k,n^*}$  for  $j \neq k; j, k \in \{1, 2, 3, 4\}$ . Order M = 5.

Table 39 shows the results of dmax and dmean between  $x_{(1),1}^{j,n^*}, x_{(2),1}^{j,n^*}$  and  $x_{(1),1}^{k,n^*}, x_{(2),1}^{k,n^*}$  for j, k = 1, 2, 3, 4. Table 40 shows the results of dmax and dmean between  $x_{(1),1}^{j,n^*}, x_{(2),1}^{j,n^*}, x_{(3),1}^{j,n^*}$  and  $x_{(1),1}^{k,n^*}, x_{(2),1}^{k,n^*}$  for j, k = 1, 2, 3, 4.

$j \backslash k$	1	2	3	4
1	-	0.92315	0.54757	0.71179
2	-	-	0.83523	0.33198
3	-	-	-	1.10101
4	-	-	-	-
1	-	0.14222	0.12811	0.18238
2	-	-	0.17223	0.11028
3	-	-	-	0.20563
4	-	-	-	-

Table 39 –  $dmax_{j,k}$  (up) and  $dmean_{j,k}$  (down) between the samples  $x_{(1),1}^{j,n^*}, x_{(2),1}^{j,n^*}$  and  $x_{(1),1}^{k,n^*}, x_{(2),1}^{k,n^*}$  for  $j \neq k; j, k \in \{1, 2, 3, 4\}$ . Order M = 5.

$j \backslash k$	1	2	3	4
1	-	1.30840	0.61490	1.01026
2	-	-	1.25238	0.52223
3	-	-	-	0.82828
4	-	-	-	-
1	-	0.16503	0.12915	0.19999
2	-	-	0.17599	0.13989
3	-	-	-	0.19314
4	-	-	-	-

Table 40 –  $dmax_{j,k}$  (up) and  $dmean_{j,k}$  (down) between the samples  $x_{(1),1}^{j,n^*}, x_{(2),1}^{j,n^*}, x_{(3),1}^{j,n^*}$  and  $x_{(1),1}^{k,n^*}, x_{(2),1}^{k,n^*}, x_{(3),1}^{k,n^*}$  for  $j \neq k; j, k \in \{1, 2, 3, 4\}$ . Order M = 5.

Table 41 shows the results of dmax and dmean between  $x_{(1),1}^{j,n^*}, x_{(2),1}^{j,n^*}, x_{(3),1}^{j,n^*}, x_{(4),1}^{j,n^*}$  and  $x_{(1),1}^{k,n^*}, x_{(2),1}^{k,n^*}, x_{(3),1}^{k,n^*}, x_{(4),1}^{k,n^*}$  for j, k = 1, 2, 3, 4.

$j \backslash k$	1	2	3	4
1	-	1.58548	0.64465	0.64251
2	-	-	1.41443	0.54859
3	-	-	-	0.64070
4	-	-	-	-
1	-	0.15505	0.12432	0.13754
2	-	-	0.16492	0.12252
3	-	-	-	0.17647
4	-	-	-	-

Table 41 –  $dmax_{j,k}$  (up) and  $dmean_{j,k}$  (down) between the samples  $x_{(1),1}^{j,n^*}, x_{(2),1}^{j,n^*}, x_{(3),1}^{j,n^*}, x_{(4),1}^{j,n^*}$ and  $x_{(1),1}^{k,n^*}, x_{(2),1}^{k,n^*}, x_{(3),1}^{k,n^*}$  for  $j \neq k; j, k \in \{1, 2, 3, 4\}$ . Order M = 5.

Table 38 reports no discrepancy between the processes, but by construction the samples with index (1) only capture discrepancies when those are very strong, and therefore the results of table 38 are very reasonable. Despite this, we must take into account the relatively small sample size in that case  $(n^*=289)$ , so it may make more sense to include more samples to support any decision. Table 39 uses for the calculations the two most robust samples, which are those of indices (1) and (2). With this change we detect a discrepancy between processes 3 and 4, although the magnitude of the discrepancy is moderate. Table 42 shows the reasons of this discrepancy.

Original series	Samples used for the calculation	P(0 s)	P(1 s)
BBAS3	$x_{(1),1}^{3,n^{st}}, x_{(2),1}^{3,n^{st}}$	0.48276	0.51724
ITUB4	$x^{4,n^{st}}_{(1),1},x^{4,n^{st}}_{(2),1}$	0.13043	0.86957

Table 42 – Details about the discrepancy related in table 39, between the processes 3 and 4,  $d_s = 1.10101, s = 01100.$ 

Table 40 uses for the calculations the three most robust samples, which are those of indices (1), (2) and (3). With this change we detect three discrepancies, two of them are those detected by the initial inspection in table 32 an extra discrepancy is now detected between the processes 1 (BBDC4) and 4 (ITUB4), although the magnitude of the discrepancy is moderate. Table 43 shows the reasons of this discrepancy, also we explore details about the previous discrepancies.

s	Original series	Samples used for the calculation	P(0 s)	P(1 s)
00100	BBDC4	$\{x_{(k),1}^{1,n^*}\}_{k=1,2,3}$	0.37143	0.62857
$(d_s = 1.30840)$	BVMF3	$\{x_{(k),1}^{2,n^*}\}_{k=1,2,3}$	0.06452	0.93548
10111	BBDC4	$\{x_{(k),1}^{1,n^*}\}_{k=1,2,3}$	0.68750	0.31250
$(d_s = 1.01026)$	ITUB4	$\{x_{(k),1}^{4,n^*}\}_{k=1,2,3}$	1.00000	0.00000
10100	BVMF3	${x^{2,n^*}_{(k),1}}_{k=1,2,3}$	0.15909	0.84091
$(d_s = 1.25238)$	BBAS3	$\{x_{(k),1}^{3,n^*}\}_{k=1,2,3}$	0.45455	0.54545

Table 43 – Details about the discrepancies related in table 40.

Hence, we are able to identify, through the technique proposed here, the subsample or period more representative for studying the daily trading volume dynamic. When there is no economic reasons for selecting a sample period, we can select one with statistical properties such as the robustness. In addition, we got a kind of classification metric according the stocks' liquidity. It can be seen, in table 39 and 40, BVMF3 and ITUB4 are closer than the others and BBAS3 and BBDC4 are next one each other. So we get two groups for the liquidity behavior.

Table 41 uses for the calculations the four most robust samples, which are those of indices (1), (2), (3) and (4). With this change we detect two discrepancies, those detected by the initial inspection in table 32. Table 44 shows the reasons of this discrepancy.

$d_{s}*$	Original series	Samples used for the calculation	$P(0 s^*)$	$P(1 s^*)$
1.58548	BBDC4	$\{x_{(k),1}^{1,n^*}\}_{k=1,2,3,4}$	0.41026	0.58974
	BVMF3	${x_{(k),1}^{2,n^*}}_{k=1,2,3,4}$	0.07895	0.92105
1.41443	BVMF3	${x_{(k),1}^{2,n^*}}_{k=1,2,3,4}$	0.07895	0.92105
	BBAS3	$\{x_{(k),1}^{\dot{3},\dot{n}^*}\}_{k=1,2,3,4}$	0.41379	0.58621

Table 44 – Details about the discrepancies related in table 41,  $s^* = 00100$ .

First of all it should be noted that the latter case offers similar information to those given by the study with all the subsets of the series, pointing the same string as responsible for the discrepancies, i.e.  $s^* = 00100$ , and involving three processes 1 (BBDC4), 2 (BVMF3) and 3 (BBAS3). The table 44 reflects the information of the table 33.

#### 4.3.4 Conclusion

The technique had been applied to real financial data, specifically to four Brazilian stocks' daily trading volume. It had given a clustering of the stocks according their liquidity and had identified the most representative period for describing this dynamic. When quantifying the distances between the 4 complete series we had identified that those that could be considered closest are BBDC4 (code 1) and BBAS3 (code 3) on the one hand, and, BVMF3 (code 2) and ITUB4 (code 4) on the other hand, see table 32. This result had kept when calculating the distances with the subsamples selected by the robust procedure for the cases of tables 39 and 40 and also for the case of the *dmean* criterion in table 41, this means, when considering more than one selected subsample. In relation to the selection of subsamples based on the robust procedure, the four financial series generally had pointed to different periods within the interval: 3 January 2012 to 31 October 2017, but the series BBDC4 and BVMF3 had shown some similarity in this selection, since they had shown as the most representative subsamples the same periods 3 and 5 (8 May 2014 -6 July 2015 and 5 September 2016 - 31 October 2017, respectively) and also those series had shown as less representative subsamples the same periods 1 and 4 (3 January 2012 -7 March 2013 and 7 July 2015 - 2 September 2016, respectively). When the series had been compared restricted to the robust subsamples of index (1), the series had appeared as being similar, revealing their practical nature, since all of them belong to the same sector, see table 38. This comparison had contrasted with the results presented by the comparison between the complete sequences (table 32) which had been affected by periods that insert internal discrepancies, in the robust sense explored in this paper. In relation to the discrepancies detected by dmax, we had highlighted two cases that had been confirmed by several selections of subsamples performed on the full series: the discrepancy between BVMF3 and BBAS3 and the discrepancy between BBDC4 and BVMF3, whose had been

presented in strings ending in 0100, that is to say an increase in the volume traded on day t-2 followed by two decreases experienced on days t-1 and t and also preceded by a decrease in the volume traded on day t-3. The discrepancies had been revealed when comparing the complete series (table 32), and with several selected subsamples (three and four of the five possibles and selected through the procedure 4.2.5) indicated in tables 40 and 41.

## 5 Final Conclusion

In this thesis we discuss two problems related to (i) the comparison of samples coming from Markovian processes by means of a BIC-based metric and (ii) a robust sample selection procedure. Both of which were motivated by several real applications.

The first problem consists of the construction of a distance which allows to compare and decide if there is any discrepancy between two samples of stochastic processes. When a discrepancy exists, the use of this distance allow us to find the strings where the discrepancy is manifested. It is shown the relationship of this distance to the divergence of Kullback Leibler and is also revealed its stochastic behavior in terms of the Chi-squared distribution, as seen in the theorem 3.2.8.

The Bayesian Information Criterion has already been used to obtain consistent methods for the selection of models, in Markovian processes, see for example Csiszár e Talata (2006) and García e González-López (2017). Chapter 3 theoretically contributes to this line of research, showing that the Bayesian Information Criterion can also be used to consistently decide whether two independent samples of stochastic processes can be considered to be coming from the same process (theorem 3.2.6-i), i.e. it allows identifying if such samples are governed by the same law of probability. Moreover, this criterion makes it possible to construct a local measure (in the mathematical sense) between the samples (theorem 3.2.4). We also see that this measure assumes unlimited values when the stochastic laws of the samples are different and the sample sizes grow, which indicates their strong discriminatory ability (theorem 3.2.6-ii). In addition, the measure tends to zero consistently when the samples are governed by the same law. We perceive a natural connection of the distance  $d_s$  with the relative entropy D computed between the empirical laws involved, which strengthens the notion that  $d_s$  is governed by similar theoretical and practical bases supporting the relative entropy D (theorem 3.2.3). The connection between the BIC concept and the measure  $d_s$  is exposed in theorem 3.2.5. This shows that to decide if two samples are coming from the same stochastic law is enough to verify that  $d_s < 1$ . The property described by theorem 3.2.5. is practical and easy to check.

We use  $d_s$  in several real applications for clarifying this topic. First we introduce a strategy to identify linguistic structures (bigrams) that generate alterations of the Portuguese, related to the period from 16th to 19th century. With use of  $d_s$  and dmax(definition 3.2.1 and definition 3.2.9), we inspect written texts of Portuguese dated between 16th century and 18th century. We identify the most voluble structures throughout the period and also we identify robust linguistic compositions that should be considered when studying the linguistic changes from Classical Portuguese to Modern Portuguese. According to remark 3.2.10 the *dmax* detects volatile linguistic constructions that expose changes in several moments from Classical Portuguese to Modern Portuguese (period: 16th century to 19th century). This type of study could motivate others that allow in fact to identify precisely how the prosody concerns intonational and rhythmic patterns involving stress alternation in a language.

In the second application, we use DNA sequences of Burkitt lymphoma. It occurs when the chromosome 8 (locus of gene MYC) is broken, which produces a change in the cellular proliferation. In this application we show how to use the measure  $d_s$  to establish a notion of proximity between strains of Burkitt lymphoma/leukemia, over the alphabet  $\mathcal{A} = \{a, c, g, t\}$ , we deal with 15 strains. The measure  $d_s$  allows to select the nearest strains to build the model whose represents the majority of the strains. Comparing the model constructed from the closest strains to the model with all the strains, we noticed that the categories practically double.

In the third application, we address a real problem that is to compare two lines of production of alcohol fuel. The purpose of this application is to determine which of these variables most contribute to the divergence. To do that, we use the local measure  $d_s$ , for each pair of variables, identifying precisely which are the configurations (strings) that lead to the discrepancy between the processes. This information points out which are the mayor problems that must be corrected to avoid divergences. The strings where the discrepancies occur expose the differences in the mechanisms of self-regulation between the columns of fuel production. We note that the processes differ in the time that they remain in certain states, for that reason long runs showing growth (code=1)/decrease (code=0) are the strings where the discrepancies appear with greater magnitude.

In chapter 4 we propose a method of selecting samples, starting from a collection of samples from Markov processes over a finite alphabet, with finite order. The selection procedure 4.2.5 shows that it is possible to detect samples with the predominate law, under the requirement that the percentage of samples with a predominant law is greater than or equal to 50% of the total of samples in the collection. The theoretical properties of this procedure are investigated in the theorems 4.2.2 and 4.2.6. Theorem 4.2.2 explains how operates the criterion of proximity between the samples, when the previous condition about the predominant law is violated. This criterion is incorporated in the selection procedure. Theorem 4.2.6 shows how the selection procedure allows us to select a quantity of samples (with the predominate law) until the total of samples that experience the predominant law, under the previous condition of to have at least 50% of the samples coming from the same law.

The technique was applied to real financial data, specifically to four Brazilian stocks' daily trading volume. With this study we had given the initial step in the sense of using statistical criteria to determine representative and robust periods of financial series, using tools coming from the stochastic processes field as is the case of the Bayesian Information Criterion. The selection procedure introduced here had lead us to describe the profile of financial series and, in the future, it could be used to determine periods of crisis in blocks of series.

# Bibliography

ARBEX, M. A.; MARTINS, L. C.; OLIVEIRA, R. C. de; PEREIRA, L. A. A.; ARBEX, F. F.; CANÇADO, J. E. D.; SALDIVA, P. H. N.; BRAGA, A. L. F. Air pollution from biomass burning and asthma hospital admissions in a sugar cane plantation area in brazil. *Journal of Epidemiology & Community Health*, BMJ Publishing Group Ltd, v. 61, n. 5, p. 395–400, 2007. Citado na página 21.

BÜHLMANN, P.; WYNER, A. J. et al. Variable length markov chains. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 27, n. 2, p. 480–513, 1999. Citado 2 vezes nas páginas 23 and 24.

CANÇADO, J. E.; SALDIVA, P. H.; PEREIRA, L. A.; LARA, L. B.; ARTAXO, P.; MARTINELLI, L. A.; ARBEX, M. A.; ZANOBETTI, A.; BRAGA, A. L. The impact of sugar cane-burning emissions on the respiratory system of children and the elderly. *Environmental health perspectives*, National Institute of Environmental Health Science, v. 114, n. 5, p. 725, 2006. Citado na página 21.

COLLET, P.; GALVES, A.; LEONARDI, F. et al. Random perturbations of stochastic processes with unbounded variable length memory. *Electronic Journal of Probability*, The Institute of Mathematical Statistics and the Bernoulli Society, v. 13, p. 1345–1361, 2008. Citado 2 vezes nas páginas 21 and 23.

CORTEZ, L. A. B.; SOUZA, G. M.; CRUZ, C. H. de B.; MACIEL, R. An assessment of brazilian government initiatives and policies for the promotion of biofuels through research, commercialization, and private investment support. In: *Biofuels in Brazil.* [S.1.]: Springer, 2014. p. 31–60. Citado na página 21.

CSISZÁR, I.; SHIELDS, P. C. et al. The consistency of the bic markov order estimator. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 28, n. 6, p. 1601–1619, 2000. Citado 2 vezes nas páginas 21 and 23.

CSISZÁR, I.; TALATA, Z. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information theory*, IEEE, v. 52, n. 3, p. 1007–1016, 2006. Citado 11 vezes nas páginas 20, 21, 23, 24, 28, 29, 32, 36, 44, 92, and 98.

FERRARI, F.; WYNER, A. Estimation of general stationary processes by variable length markov chains. *Scandinavian Journal of Statistics*, Wiley Online Library, v. 30, n. 3, p. 459–480, 2003. Citado na página 23.

FROTA, S.; GALVES, C.; VIGÁRIO, M.; GONZALEZ-LOPEZ, V.; ABAURRE, B. The phonology of rhythm from classical to modern portuguese. *Journal of Historical Linguistics*, John Benjamins Publishing Company, v. 2, n. 2, p. 173–207, 2012. Citado 4 vezes nas páginas 48, 49, 55, and 58.

GALVES, A.; GALVES, C.; GARCIA, J. E.; GARCIA, N. L.; LEONARDI, F. et al. Context tree selection and linguistic rhythm retrieval from written texts. *The Annals of Applied Statistics*, Institute of Mathematical Statistics, v. 6, n. 1, p. 186–209, 2012. Citado 2 vezes nas páginas 23 and 48. GALVES, A.; LEONARDI, F. Exponential inequalities for empirical unbounded context trees. *In and Out of Equilibrium 2*, Springer, p. 257–269, 2008. Citado na página 23.

GALVES, C.; FARIA, P. Tycho brahe parsed corpus of historical portuguese. 2010. Citado na página 48.

GARCÍA, J.; GONZÁLEZ-LÓPEZ, V.; HIRSH, I. Copula-based prediction of economic movements. In: AIP PUBLISHING. *AIP Conference Proceedings*. [S.I.], 2016. v. 1738, n. 1, p. 140005. Citado na página 72.

GARCÍA, J. E.; GHOLIZADEH, R.; GONZÁLEZ-LÓPEZ, V. A. Comparison of stochastic processes. In: *In Books of ASMDA*. [S.l.: s.n.], 2017. Citado na página 33.

\_\_\_\_\_. Linguistic compositions highly volatile in portuguese. *Cadernos de Estudos Linguísticos*, v. 59, n. 3, p. 617–630, 2017. Citado 2 vezes nas páginas 20 and 47.

GARCIA, J. E.; GONZÁLEZ-LÓPEZ, V. Detecting regime changes in markov models. In: Proceedings of The Sixth Workshop on Information Theoretic Methods in Science and Engineering. [S.l.: s.n.], 2013. Citado na página 20.

\_\_\_\_\_. Optimal partition of markov models and automatic classification of languages. *ISAST Publishing*, p. 207–215, 2016. Citado na página 48.

GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V.; ANDRADE, F. K. de. Dissimilarity between markovian processes applied to industrial processes. In: AIP PUBLISHING. *AIP Conference Proceedings*. [S.l.], 2017. v. 1863, n. 1, p. 220002. Citado 2 vezes nas páginas 20 and 69.

GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V.; VIOLA, M. Robust model selection for stochastic processes. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 43, n. 10-12, p. 2516–2526, 2014. Citado 4 vezes nas páginas 21, 29, 37, and 81.

GARCÍA, J. E.; GONZÁLEZ-LÓPEZ, V. A. Consistent estimation of partition markov models. *Entropy*, Multidisciplinary Digital Publishing Institute, v. 19, n. 4, p. 160, 2017. Citado 12 vezes nas páginas 20, 21, 23, 24, 30, 37, 38, 42, 44, 61, 62, and 92.

GOLDEMBERG, J. The brazilian biofuels industry. *Biotechnology for biofuels*, BioMed Central, v. 1, n. 1, p. 1–7, 2008. Citado na página 21.

LEONARDI, F. G. Cadeias estocásticas parcimoniosas com aplicações à classificação e filogenia das seqüências de proteínas. Tese (Doutorado) — Universidade de São Paulo, 2007. Citado na página 23.

MANNING, C. D.; SCHÜTZE, H. Foundations of statistical natural language processing. [S.l.]: MIT press, 1999. Citado na página 48.

RISSANEN, J. A universal data compression system. *IEEE Transactions on information theory*, IEEE, v. 29, n. 5, p. 656–664, 1983. Citado 3 vezes nas páginas 21, 23, and 24.

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado 6 vezes nas páginas 20, 21, 23, 28, 29, and 32.

Appendix

## APPENDIX A - Proofs

### A.1 Proofs of Theorems of Chapter 1

To prove the theorem 2.2.8, it is necessary to show the following lemmas introduced by Csiszár e Talata (2006).

**Lemma A.1.1.** For any proper suffix s of some  $s_0 \in \mathcal{T}_0$ , there exists an irreducible tree  $\widetilde{\mathcal{T}}$  with  $d(\widetilde{\mathcal{T}}) < \infty$  such that u > s and Q(u) > 0 for each  $u \in \widetilde{\mathcal{T}}$ , each  $v \geq s$  with Q(v) > 0 has a suffix in  $\widetilde{\mathcal{T}}$ , and

$$\ln \widetilde{P}_{ML,s}(x_1^n) - \sum_{u \in \widetilde{\mathcal{T}}} \ln \widetilde{P}_{ML,u}(x_1^n) < -cn$$
(A.1)

eventually almost surely as  $n \to \infty$ , where c > 0 is a sufficiently small constant.

**Lemma A.1.2.** For any irreducible tree  $\mathcal{T}$  with  $d(\mathcal{T}) \leq D(n)$ ,  $D(n) = o(\ln n)$  and  $s \in \mathcal{T}$  that has a proper suffix  $s_0 \in \mathcal{T}_0$  with  $l(s_0) \leq K$ , there exists w satisfying  $s > w \geq s_0$  such that, for  $\widetilde{\mathcal{T}} = \{u \in \mathcal{T} : u > w\}$  and arbitrary v > 0

$$\sum_{u \in \widetilde{\mathcal{T}}} \ln \widetilde{P}_{ML,u}(x_1^n) - \ln \widetilde{P}_{ML,w}(x_1^n) < v |\widetilde{\mathcal{T}}| \ln n$$
(A.2)

holds simultaneously for all  $\mathcal{T}$  and as above, eventually almost surely as  $n \to \infty$ . Moreover, here  $w = a_{-k}a_{-k+1} \dots a_{-1}$  be chosen such that  $a_{-k+1} \dots a_{-1}$  is a proper suffix of some  $u \in \mathcal{T} \setminus \widetilde{\mathcal{T}}$ .

**Theorem A.1.3.** If  $d(\mathcal{T}_0) < \infty$ , the estimator  $\widehat{\mathcal{T}}_{BIC}$ , with  $D(n) = o(\ln n)$ , satisfies  $\widehat{\mathcal{T}}_{BIC}(X_1^n) = \mathcal{T}_0$  almost certainly, when  $n \longrightarrow \infty$ .

**Proof.** For demonstrate the theorem it is enough to show that if  $\mathcal{T} \neq \mathcal{T}_0$  for some  $\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{I}$  then there exists a modification  $\mathcal{T}'$  of  $\mathcal{T}$  also satisfying  $\mathcal{T}' \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{I}$  such that

$$BIC_{\mathcal{T}}(x_1^n) > BIC'_{\mathcal{T}}(x_1^n) \tag{A.3}$$

simultaneously for all considered trees  $\mathcal{T}$ , eventually almost surely as  $n \to \infty$ .

According to Eq. 2.2 the likelihood can be described by

$$ML_{\mathcal{T}}(x_1^n) = \prod_{s \in \mathcal{T}} \widetilde{P}_{ML,s}(x_1^n),$$

where

$$\widetilde{P}_{ML,s}(x_1^n) = \begin{cases} \prod_{a \in \mathcal{A}} \left[ \frac{N_n(s,a)}{N_n(s)} \right]^{N_n(s,a)} & \text{if } N_n(s) \ge 1; \\ 1 & \text{if } N_n(s) = 0. \end{cases}$$

Using of the definition of BIC, Eq. A.3 is equivalent to

$$\sum_{s \in \mathcal{T}} \ln \widetilde{P}_{ML,s}(x_1^n) - \sum_{s' \in \mathcal{T}'} \widetilde{P}_{ML,s'}(x_1^n) < \frac{(|\mathcal{T}| - 1)}{2} (|\mathcal{T}| - |\mathcal{T}'|) \ln n.$$
(A.4)

If  $\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{T}$  is of such form that  $\mathcal{T} \neq \mathcal{T}'$ , there exist a sequence  $\tilde{s} \in \mathcal{T}$  and  $\tilde{s}_0 \in \mathcal{T}_0$  such that  $\tilde{s} < \tilde{s}_0$  or  $\tilde{s}_0 < \tilde{s}$ . Equivalently, there exist  $s \in \mathcal{T}$  and  $s_0 \in \mathcal{T}_0$  such that one of the following cases is true:

- (a)  $s < s_0$
- (b)  $s_0 < s$

Let us first consider the case (a), then there is a modification  $\mathcal{T}' \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{I}$  of  $\mathcal{T}$  given by

$$\mathcal{T}^{'} = (\mathcal{T} \setminus \{s\}) \cup \widetilde{\mathcal{T}}$$

where  $\tilde{\mathcal{T}}$  given by Lemma A.1.1. that in case (a), we have  $|\mathcal{T}| - |\mathcal{T}'| = 1 - |\tilde{\mathcal{T}}|$  and the left-hand side of Eq. A.4 is equal to that of Eq. A.1.1. By Lemma A.4 the latter is less than -c n, eventually almost surely  $n \to \infty$ , and thus (A.4) certainly holds.

Let us now consider the case (b) Then there exist a modification  $\mathcal{T}' \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{I}$  of  $\mathcal{T}$  given by

$${\mathcal T}^{'}=({\mathcal T}ackslash {\mathcal T}^{'})\cup\{w\}$$
 ,

where  $\widetilde{\mathcal{T}}$  and  $\{w\}$  given by Lemma A.1.2. In case (b), we have  $|\mathcal{T}| - |\mathcal{T}'| = |\widetilde{\mathcal{T}}| - 1$ , and the left-hand side of Eq. A.4 is equal to that of Eq. A.1.2 below. Hence, by Lemma A.1.2, Eq. (A.4) is satisfied also in this case, eventually almost surely as for all considered.

**Theorem A.1.4.** Let  $(X_t)$  and  $(Y_t)$  be two stationary, ergodic and Markovian stochastic process on a finite alphabet  $\mathcal{A}$  of finite order, with probability law P and Q, respectively. Then, the relative entropy rate between them is given by

$$D(P||Q) = \sum_{s \in \mathcal{T}_{PQ}} P(s)D(P(.|s)||Q(.|s))$$

**Proof.** Let  $d = \max \{ d(\mathcal{T}_P), d(\mathcal{T}_Q) \}$  and  $\mathcal{A}^d$  be the set of all sequence of length d, it is denoted by  $P_n$  and  $Q_n$ , respectively the probabilities P(x) and Q(x) for some  $x \in \mathcal{A}^d$ .

Given a sequence  $z \in \mathcal{A}^{n+1}$ , n > d, exist  $x \in \mathcal{A}^n$ ,  $y \in \mathcal{A}$  such that z = xy. Then

$$D(P_{n+1}||Q_{n+1}) = \sum_{z \in \mathcal{A}^{n+1}} P(z) \ln\left(\frac{P(z)}{Q(z)}\right) = \sum_{x \in \mathcal{A}^n} \sum_{y \in \mathcal{A}} P(xy) \ln\left(\frac{P(xy)}{Q(xy)}\right)$$
$$= \sum_{x \in \mathcal{A}^n} \sum_{y \in \mathcal{A}} P(y|x)P(x) \ln\left(\frac{P(y|x)P(x)}{Q(y|x)Q(x)}\right) = \sum_{x \in \mathcal{A}^n} \sum_{y \in \mathcal{A}} P(y|x)P(x) \ln\left(\frac{P(y|x)}{Q(y|x)}\right)$$
$$+ \sum_{x \in \mathcal{A}^n} \sum_{y \in \mathcal{A}} P(y|x)P(x) \ln\left(\frac{P(x)}{Q(x)}\right)$$

Using  $\sum_{y \in \mathcal{A}} P(y|x) = 1$  we get

$$D(P_{n+1}||Q_{n+1}) = \sum_{x \in \mathcal{A}^n} \sum_{y \in \mathcal{A}} P(y|x)P(x)\ln\left(\frac{P(y|x)}{Q(y|x)}\right) + \sum_{x \in \mathcal{A}^n} P(x)\ln\left(\frac{P(x)}{Q(x)}\right)$$

Now, using the definition

$$D(P(Y|x)||Q(Y|x)) = \sum_{y \in \mathcal{A}} P(y|x) \ln\left(\frac{P(y|x)}{Q(y|x)}\right)$$

for some sequence  $x \in \mathcal{A}^n$ , and

$$D(P_n||Q_n) = \sum_{x \in \mathcal{A}^n} P(x) \ln\left(\frac{P(x)}{Q(x)}\right)$$
  
we have 
$$D(P_{n+1}||Q_{n+1}) = \sum_{x \in \mathcal{A}^n} P(x)D(P(.|x)||Q(.|x)) + D(P_n||Q_n).$$

Given  $x \in \mathcal{A}^n$ ,  $\exists$  a sequence  $x_1$  such that  $x = x_1 s$ ,  $s \in \mathcal{S}$  and s is suffix of  $x_1$ . Thus

$$D(P_{n+1}||Q_{n+1}) = \sum_{s \in \mathcal{T}_{PQ}} \sum_{a \in \mathcal{A}^k: s} \sum_{\text{ is suffix of } x_1} P(x)D(P(.|s)||Q(.|s)) + D(P_n||Q_n)$$
$$= \sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s)) \left[ \sum_{a \in \mathcal{A}^k: s} \sum_{\text{ is suffix of } x_1} P(x) \right] + D(P_n||Q_n)$$
$$= \sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s))P(s) + D(P_n||Q_n)$$
(A.5)

as

 $P(s) = \sum_{a \in \mathcal{A}^k: s \text{ is suffix of } x_1} P(x).$ 

Then, using the reasoning above,  $D(P_n||Q_n)$  can be written as

$$D(P_n||Q_n) = \sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s))P(s) + D(P_{n-1}||Q_{n-1})$$
(A.6)

Substituting (A.6) into (A.5) we have

$$\begin{aligned} D(P_{n+1}||Q_{n+1}) &= \sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s))P(s) + D(P_n||Q_n) \\ &= \sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s))P(s) + \sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s))P(s) + D(P_{n-1}||Q_{n-1}) \\ &= 2\sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s))P(s) + D(P_{n-1}||Q_{n-1}) \end{aligned}$$

Developing, the expression  $D(P_{n-1}||Q_{n-1})$  until the sequence  $s \in \mathcal{T}_{PQ}$  of length d we obtain

$$D(P_{n+1}||Q_{n+1}) = (n-d) \sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s))P(s) + D(P_d||Q_d)$$

Thus, the relative entropy rate between the laws processes

$$D(P||Q) = \lim_{n \to \infty} \left[ \frac{n-d}{n} \sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s))P(s) + \frac{1}{n} D(P_d||Q_d) \right]$$
  
= 
$$\lim_{n \to \infty} \left[ \frac{n-d}{n} \sum_{s \in \mathcal{T}_{PQ}} D(P(.|s)||Q(.|s))P(s) \right] + \lim_{n \to \infty} \frac{1}{n} D(P_d||Q_d)$$
  
= 
$$\sum_{s \in \mathcal{S}} D(P(.|s)||Q(.|s))P(s)$$

Therefore,

$$D(P||Q) = \sum_{s \in \mathcal{S}} D(P(.|s)||Q(.|s))P(s).$$

# APPENDIX B - Graphs

## B.1 Graphs of time series for financial study in chapter 4



Figure 8 – First period, 3 January 2012 - 7 March 2013 , for BDDC4 using of time series.



Figure 9 – Second period, 3 January 8 March 2013 - 7 May 2014 , for BDDC4 using of time series.



Figure 10 – Third period, 8 May 2014 - 6 July 2015 , for BDDC4 using of time series.



Figure 11 – First period, 3 January 2012 - 7 March 2013, for BVMF3 using of time series.



Figure 12 – Second period, 3 January 8 March 2013 - 7 May 2014 , for BVMF3 using of time series.



Figure 13 – Representative third period, 8 May 2014 - 6 July 2015, for BVMF3 using of time series.



Figure 14 – Fourth period, 7 July 2015 - 2 September 2016, for BVMF3 using of time series.



Figure 15 – Fifth period, 5 September 2016 - 31 October 2017, for BVMF3 using of time series.



Figure 16 – First period, 3 January 2012 - 7 March 2013, for BBAS3 using of time series.



Figure 17 – Second period, 3 January 8 March 2013 - 7 May 2014 , for BBAS3 using of time series.



Figure 18 – Third period, 8 May 2014 - 6 July 2015, for BBAS3 using of time series.



Figure 19 – Representative fourth period, 7 July 2015 - 2 September 2016, for BBAS3 using of time series.



Figure 20 – Fifth period, 5 September 2016 - 31 October 2017, for BBAS3 using of time series.



Figure 21 – First period, 3 January 2012 - 7 March 2013, for ITUB4 using of time series.


Figure 22 – Third period, 8 May 2014 - 6 July 2015 , for ITUB4 using of time series.



Figure 23 – Representative second period, 8 March 2013 - 7 May 2014, for ITUB4 using of time series.



Figure 24 – Fourth period, 8 May 2014 - 6 July 2015, for BBAS3 using of time series.



Figure 25 – Fifth period, 5 September 2016 - 31 October 2017, for ITUB4 using of time series.