Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica

Márcio Luis Lanfredi Viola

Tópicos em Seleção de Modelos Markovianos

TESE DE DOUTORADO APRESENTADA AO INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA DA UNICAMP PARA OBTENÇÃO DO TÍTULO DE DOUTOR EM ESTATÍSTICA.

Orientador: Prof. Dr. Jesus Enrique Garcia
Coorientadora: Verónica Andrea González-López

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE/DISSERTAÇÃO DEFENDIDA PELO ALUNO MÁRCIO LUIS LANFREDI VIOLA, E ORIENTADA PELO PROF. DR. JESUS ENRIQUE GARCIA

Prof. Dr. Jesus Enrique Garcia
Orientador

Profa. Dra. Verónica Andrea González-López
Coorientadora

Profa. Dra. Verónica Andrea González López
Diretora Associada
IMECC - UNICAMP
Matric. 28485-3

CAMPINAS, 2011

FICHA CATALOGRÁFICA ELABORADA POR
MARIA FABIANA BEZERRA MÜLLER - CRB8/6162
BIBLIOTECA DO INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E
COMPUTAÇÃO CIENTÍFICA - UNICAMP

V811t

Viola, Márcio Luis Lanfredi, 1978-
Tópicos em seleção de modelos markovianos / Márcio Luis Lanfredi Viola. - Campinas, SP : [s.n.], 2011.

Orientador: Jesus Enrique Garcia.
Coorientador: Verónica Andrea González-López.

Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

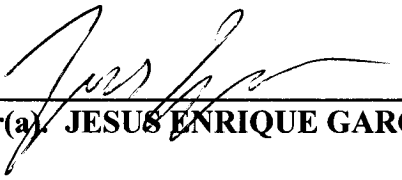
1. Markov, Processos de. 2. Estatística robusta. 3. Comprimento Mínimo de Descrição. 4.Compressão de dados. I. García, Jesus Enrique, 1966-. II.González-López, Verónica Andrea, 1970- . III. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. IV. Título.

Informações para Biblioteca Digital

Título em inglês: Topics in selection of Markov models
Palavras-chave em inglês:
Markov processes
Robust statistica
Iminimum description lengh (Information theory)
Data compression (Computer science)
Área de concentração: Estatística
Titulação: Doutor em Estatística
Banca examinadora:
Jesus Enrique García [Orientador]
Florencia Graciela Leonardi
Miguel Natalio Abadi
Nancy Lopes Garcia
Pedro José Catuogno
Data da defesa: 13-12-2011
Programa de Pós-Graduação: Estatística

Tese de Doutorado defendida em 13 de dezembro de 2011 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

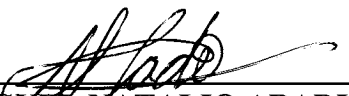


Prof(a). Dr(a). JESUS ENRIQUE GARCIA

Prof(a). Dr(a). NANCY LOPES GARCIA

Prof(a). Dr(a). PEDRO JOSÉ CATUOGNO

Prof(a). Dr(a). FLORENCIA GRACIELA LEONARDI



Prof(a). Dr(a). MIGUEL NATALIO ABADI

Agradecimentos

Agradeço a todos que contribuiram, direta ou indiretamente para a realização deste trabalho.

À minha família.

À todos os meus amigos.

À CAPES pelo suporte financeiro.

Resumo

Nesta tese abordamos o problema estatístico de seleção de um modelo Markoviano de ordem finita que se ajuste bem a um conjunto de dados em duas situações diferentes.

Em relação ao primeiro caso, propomos uma metodologia para a estimação de uma árvore de contextos utilizando-se amostras independentes sendo que a maioria delas são provenientes de um mesmo processo de Markov de memória variável finita e as demais provêm de um outro processo Markoviano de memória variável finita. O método proposto é baseado na taxa de entropia relativa simetrizada como uma medida de similaridade. Além disso, o conceito de ponto de ruptura assintótico foi adaptado ao nosso problema de seleção a fim de mostrar que o procedimento proposto, nesta tese, é robusto.

Em relação ao segundo problema, considerando um processo de Markov multivariado de ordem finita, propomos uma metodologia consistente que fornece a partição mais fina das coordenadas do processo de forma que os seus elementos sejam condionalmente independentes. O método obtido é baseado no BIC (Critério de Informação Bayesiano). Porém, quando o número de coordenadas do processo cresce, o custo computacional do critério BIC torna-se excessivo. Neste caso, propomos um algoritmo mais eficiente do ponto de vista computacional e a sua consistência é demonstrada.

A eficiência das metodologias propostas foi estudada através de simulações e elas foram aplicadas em dados linguísticos.

Palavras-chave: Processo de Markov, Processo de Markov de Memória Variável, Árvore de Contextos, Entropia Relativa, Robustez, Critério de Informação Bayesiano, Independência Condicional.

Abstract

This work related two statistical problems involving the selection of a Markovian model of finite order.

Firstly, we propose a procedure to choose a context tree from independent samples, with more than half of them being realizations of the same finite memory Markovian processes with finite alphabet with law P. Our model selection strategy is based on estimating relative entropies to select a subset of samples that are realizations of the same law. We define the asymptotic breakdown point for a model selection procedure, and show the asymptotic breakdown point for our procedure. Moreover, we study the robust procedure by simulations and it is applied to linguistic data.

The aim of other problem is to develop a consistent methodology for obtain the finner partitions of the coordinates of an multivariate Markovian stationary process such that their elements are conditionally independents. The proposed method is establishment by Bayesian information criterion (BIC). However, when the number of the coordinates of process increases, the computing of criterion BIC becomes excessive. In this case, we propose an algorithm more efficient and the its consistency is demonstrated. It is tested by simulations and applied to linguistic data.

Keywords: Context Tree, Variable Memory Markov Chain, CTM Algorithm, Relative Entropy, Robustness.

Sumário

Lista de Figuras

Introdução

Nesta tese abordamos o problema estatístico de seleção de um modelo Markoviano, discreto, estacionário e de ordem finita sobre um alfabeto finito, que se ajuste bem a um conjunto de dados em duas situações diferentes.

A metodologia desenvolvida, para cada uma das situações, foi motivada por um problema linguístico. A primeira surgiu do estudo de correlação acústica de classes rítmicas [24, 18, 10] cujos dados integram o corpus pertecente à *Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS/CNRS)*. Os dados consistem de sinais acústicos resultante de sentenças lidas por falantes dos idiomas Inglês, Japonês, Espanhol, Francês, Holandês, Italiano, Polonês e Catalão.

Como, recentemente, Galves et al. [13] tem aplicado processos Markovianos de memória variável na caracterização de idiomas, a ideia é obter uma árvore de contexto que caracterize cada um dos idiomas. Porém, podemos nos perguntar se a dicção de um indivíduo pode interferir no sinal acústico de forma que a estrutura da árvore de contextos possa sofrer alguma modificação. Isso nos motiva a desenvolver um procedimento que estime uma árvore de contextos, para um determinado idioma, de forma que a presença de algum tipo de ruído, na minoria das amostras, não afete a sua estrutura.

Mais especificamente, o conjunto de dados é formado por amostras independentes sendo que a maioria das amostras são provenientes de uma mesma cadeia de Markov de memória variável finita e as demais provêm de um outro processo Markoviano de memória variável finita. Estimando-se uma árvore de contextos para cada uma das amostras, criamos um mecanismo robusto, baseado na taxa de entropia relativa simetrizada como uma medida de similaridade, que permite obter o modelo representante da maioria das amostras.

O segundo problema foi motivado pelo artigo escrito por Galves et al. [13] no qual cadeias de símbolos obtidos através da codificação de textos escritos extraídos de um corpus de jornais

impressos, em Português Brasileiro e Português Europeu, foram modeladas através de um processo Markoviano de memória variável a fim de obter evidências de que o Português Europeu e o Português Brasileiro pertencem a diferentes classes rítmicas. Para isto, um processo bivariado foi criado a partir duas características rítmicas básicas: se a sílaba é acentuada ou não; e se a sílaba é o início de uma palavra prosódica ou não.

Podemos nos questionar se os dois processos, que formam o processo bivariado, são condicionalmente independentes ou não dado uma sequência de símbolos do alfabeto. Mais especificamente, considerando um processo de Markov multivariado, estacionário, de ordem finita sobre um alfabeto finito, o objetivo é propor uma metodologia consistente que forneça a partição mais fina das coordenadas do processo multivariado de forma que os seus elementos sejam condionalmente independentes. O método obtido é baseado no BIC (Critério de Informação Bayesiano) [27, 7].

Vale observar que o primeiro problema foi desenvolvido para processos de Markov de memória variável finita [25, 4, 22, 9] sendo, também, válido para para modelos Markovianos de ordem fixa e finita, modelos Markovianos mínimos e [16] e modelos multinomiais. Já o segundo foi desenvolvido para modelos markovianos de ordem fixa e finita sendo válido para modelos Markovianos de memória variável finita e modelos multinomiais multivariados.

Para finalizar, apresentamos a organização da tese ressaltando os principais resultados originais obtidos:

- No Capítulo 1 definimos processos Markovianos e descrevemos o algoritmo proposto por Csiszár e Talata [9] para a estimação de árvores de contextos via o critério de informação BIC [27, 7];

- No Capítulo 2 propomos um procedimento robusto de estimação de árvores de contextos e obtemos o seu ponto de ruptura assintótico. O conceito de ponto de ruptura assintótico, criado por Donoho e Huber [12], foi adaptado para o nosso problema. Por fim, a eficiência do métodos robusto foi estudada através de simulações e ele foi aplicado em dados linguísticos;

- No Capítulo 3 propomos um método consistente com a finalidade de obter a partição mais fina das coordenadas de um processo Markoviano multivariado de modo que seus elementos sejam condicionalmente independentes dada uma sequência de símbolos do alfabeto. Porém, quando o número de coordenadas do processo cresce, seu custo computacional torna-se excessivo.

Neste caso, propomos um algoritmo consistente mais eficiente do ponto de vista computacional. Além disso, a eficiência da metodologia proposta foi estudada através de simulações e ela foi aplicada em dados linguísticos;

- No Capítulo 4 relatamos as conclusões obtidas nos Capítulos 2 e 3;
- No Capítulo 5 apresentamos as demonstrações dos teoremas citados nos nos Capítulos 2 e 3.

Capítulo 1

Árvores de Contextos

1.1 Introdução

As árvores de contextos [25, 4, 22] têm sido estudadas e utilizadas na modelagem de vários problemas práticos como, por exemplo, na análise de dados linguísticos (Galves et al., 2009) [13], na classificação de proteínas (Leonardi, 2007) [21, 22] e na identificação de genes (Bulhmann & Wyner, 1999) [4].

Vários pesquisadores têm estudado diversos aspectos relacionados às árvores de contextos. Galves e Leonardi [15] pesquisaram um limite superior exponencial para a taxa de convergência do algoritmo Contexto quando a árvore não é necessariamente limitada. Collet, Galves e Leonardi [5] estudaram a recuperação da árvore de contextos de uma cadeia de Markov de memória variável ilimitada a partir de uma amostra com ruído desta cadeia.

Na literatura existem vários algoritmos propostos para a estimação de árvores de contextos como, por exemplo, os algoritmos propostos por Rissanen (Algoritmo Contexto) [25], Buhlmann & Wyner [4], Csiszár & Talata [9], Bejerano & Yona [2], Leonardi [21, 22], Galves, Galves, Garcia & Leonardi [13].

O resultado que será apresentado no Teorema 3.3 foi baseado no Critério de Informação de Bayes (BIC) e a sua demonstração utiliza alguns fatos apresentados no artigo mo qual Csiszár e Talata [9] propuseram um algoritmo para a estimação de árvores de contextos. Por isso, neste capítulo, abordaremos os principais resultados deste artigo como, por exemplo, o algoritmo proposto por

Csiszár e Talata [9] para a estimação de árvores de contextos via o critério de informação BIC [27, 7].

Na Seção 1.2 definimos o conceito de processo de Markov de memória variável [4, 9], na Seção 1.3 apresentamos alguns conceitos relacionados à árvore sufixo [9] e a sua associação com um processo de Markov de memória variável [4, 9] e na Seção 1.4 relatamos o algoritmo proposto por Csiszár e Talata [9].

1.2 Cadeias de Markov de Memória Variável

Por simplicidade de notação, a sequência $x_t x_{t-1} \dots x_0$ será denotada por x_0^t.

Um processo estocástico é uma família $(X_t)_{t\in\mathbb{T}}$ de variáveis aleatórias definidas em um mesmo espaço de probabilidade. Para cada $t \in \mathbb{T}$, X_t é uma variável aleatória que assume valores num conjunto $\mathcal{A}$ chamado alfabeto. Usualmente, o índice t é interpretado como sendo o tempo e, informalmente, um processo estocástico descreve uma história que se desenvolve de forma aleatória ao longo de um período representado por $\mathbb{T}$ que, neste caso, será considerado como o conjunto dos números inteiros $\mathbb{Z}$. Uma classe muito importante de processos a tempo discreto são os processos de Markov ou cadeias de Markov.

Em uma cadeia (ou processo) de Markov de ordem um, a previsão do próximo passo, conhecendo toda a história passada do processo desde o seu início é tão boa quanto a previsão feita conhecendo-se apenas o valor do processo no presente.

Definição 1.1. *Um processo estocástico discreto $(X_t)_{t\in\mathbb{Z}}$ é um processo de Markov de ordem um se para todo $t \geq 1$ e para quaisquer elementos $x_0, x_1, \dots, x_n$ de $\mathcal{A}$, tais que $Prob\big(X_0^{t-1} = x_0^{t-1}\big) > 0$, a igualdade (1.2.1) é válida.*

$$Prob\big(X_t = x_t | X_0^{t-1} = x_0^{t-1}\big) = Prob(X_t = x_t | X_{t-1} = x_{t-1}) \tag{1.2.1}$$

Se o processo for estacionário, podemos trocar os índices $-\infty, ..., t-1, t$ pelos índices $-\infty, ..., -1, 0$, para todo $t \in \mathbb{Z}$.

As cadeias de Markov completas de ordem k estendem os processos Markovianos de ordem 1.

Definição 1.2. *Um processo estocástico discreto $(X_t)_{t\in\mathbb{Z}}$ é um processo de Markov de ordem k (ou processo de Markov completo de ordem k) se para todo $t \geq k$ e para toda sequência $x_{-1}x_{-2}\ldots x_{-t} \in \mathcal{A}^t$, tal que $Prob\big(X_{-t}^{-1} = x_{-t}^{-1}\big) > 0$, valer a igualdade (1.2.2).*

$$Prob\big(X_0 = x_0 | X_{-t}^{-1} = x_{-t}^{-1}\big) = Prob\big(X_0 = x_0 | X_{-k}^{-1} = x_{-k}^{-1}\big) \tag{1.2.2}$$

sendo

$$k = \min\big\{l : Prob(X_0 = x_0 | X_{-t}^{-1} = x_{-t}^{-1}) = Prob(X_0 = x_0 | X_{-l}^{-1} = x_{-l}^{-1}), \forall x_{-1}x_{-2}\ldots x_{-t} \in \mathcal{A}^t\big\}$$

Vale observar que o número de parâmetros de uma cadeia de Markov cresce exponencialmente com a sua ordem. Há casos em que a descrição mais eficiente pode ser aquela em que a ordem da cadeia de Markov não é fixa. Este tipo de processo Markoviano, denominado processo de Markov de memória variável, será definido a seguir [21].

Definição 1.3. *Seja $(X_t)_{t\in\mathbb{Z}}$ um processo estacionário com valores em $\mathcal{A}$. Diremos que o processo $(X_t)_{t\in\mathbb{Z}}$ é uma cadeia de Markov de memória variável (VLMC) se para toda sequência x_{-t}^{-1} satisfazendo $Prob\big(X_{-t}^{-1} = x_{-t}^{-1}\big) > 0$, existe um inteiro k, determinado a partir de x_{-t}^{-1}, tal que*

$$Prob\big(X_0 = x_0 | X_{-t}^{-1} = x_{-t}^{-1}\big) = Prob\big(X_0 = x_0 | X_{-k}^{-1} = x_{-k}^{-1}\big) \tag{1.2.3}$$

Assume-se que $k\big(x_{-t}^{-1}\big)$ é o menor inteiro que satisfaz (1.2.3). Assim, k é o comprimento da memória do processo, dado que o passado foi x_{-t}^{-1}, ou seja, cada estado presente da cadeia depende de um passado até encontrar um determinado evento o qual faz com que todo o passado restante seja irrelevante.

As sequências finitas dadas pelos símbolos $(x_{-k}, \ldots, x_{-1})$ são os passados relevantes e são chamadas de contextos.

1.3 Árvore de Contextos

Primeiramente, serão abordados alguns conceitos relativos à árvore sufixo.

Seja $\mathcal{A}$ um alfabeto finito e $|\mathcal{A}|$ a sua cardinalidade. O conjunto $\mathcal{A}^\star$ é formado por todas as sequências finitas sobre $\mathcal{A}$.

Uma sequência $s = a_m a_{m+1} ... a_n$, $a_i \in \mathcal{A}$, $m \leq i \leq n$, será denotada por a_m^n e $l(s) = n - m + 1$ é o comprimento de s sendo que a sequência vazia será denotada por $\emptyset$ a qual $l(\emptyset) = 0$.

Denotando a concatenação de u e v por uv, uma sequência v será um sufixo de uma sequência s quando existir uma sequência u tal que $s = uv$.

Definição 1.4. *Um subconjunto finito $\mathcal{T}$ de $\mathcal{A}^\star$ é uma árvore irredutível se satisfaz as seguintes propriedades:*

1. *Propriedade de sufixo: Nenhuma sequência $s \in \mathcal{T}$ é um sufixo de uma outra sequência $r \in \mathcal{T}$;*
2. *Irredutibilidade: Nenhuma sequência $s \in \mathcal{T}$ pode ser substituída por um sufixo dela sem violar a propriedade do sufixo.*

A família de árvores irredutíveis será denotada por $\mathbb{I}$.

Cada sequência $s = a_k^1 \in \mathcal{T}$ é visualizada como sendo o caminho das folhas até a raiz (topo da árvore), consistindo de k arestas rotuladas pelos símbolos $a_1 \dots a_k$ e uma sequência semi-infinita $a_{-\infty}^{-1} \in \mathcal{T}$ é visualizada como um caminho infinito até a raiz. As sequências $s \in \mathcal{T}$ são, também, identificadas com as folhas da árvore $\mathcal{T}$ sendo que a folha s é aquela conectada com a raiz pelo caminho visualizado s. Similarmente, os nós da árvore $\mathcal{T}$ são identificados com o sufixo finito de todos (finito ou infinito) $s \in \mathcal{T}$ sendo que a raiz é identificada pela sequência vazia $\emptyset$.

A profundidade de uma árvore $\mathcal{T}$, denotada por $d(\mathcal{T})$, é dada por

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

e a árvore $\mathcal{T}$ trucada no nível K será denotada por $\mathcal{T}\big|_K$ sendo

$$\mathcal{T}\big|_K = \{s' : s' \in \mathcal{T} \text{ com } l(s') \leq K \text{ ou } s' \text{ é um sufixo de comprimento } \lfloor K \rfloor \text{ de algum } s \in \mathcal{T}\}$$

Agora, a árvore sufixo será vinculada a um processo de Markov de memória variável (VLMC). Considere um processo estocástico ergódico estacionário $(X_t)_{t \in \mathbb{Z}}$ tomando valores no alfabeto finito $\mathcal{A}$. Denota-se $P(a_m^n) = Prob(X_m^n = a_m^n)$ e $P(a|s) = Prob(X_0 = a | X_{-k}^{-1} = s)$, $s \in \mathcal{A}^k$, se $P(s) > 0$.

Definição 1.5. *Uma sequência $s \in \mathcal{A}^\star$ é um contexto para o processo de lei P se $P(s) > 0$ e para toda sequência $x_{-\infty}^{-1}$ tal que s é sufixo da sequência $x_{-\infty}^{-1}$ tem-se*

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

e nenhum sufixo de s tem essa propriedade.

Com essa definição pode-se ver que o conjunto de contextos de um proceso $(X_t)_{t\in\mathbb{Z}}$ é uma árvore irredutível. O conjunto de contextos associado ao processo $(X_t)_{t\in\mathbb{Z}}$ é chamada de árvore probabilística de contextos.

Definição 1.6. *Uma árvore probabilística de contexto é um par ordenado* $(\mathcal{T}, p)$ *tal que:*

1. $\mathcal{T}$ *é uma árvore irredutível;*

2. $p = \left\{P(\cdot|s) : s \in \mathcal{T}\right\}$ *é uma família de probabilidades de transição sobre* $\mathcal{A}$.

Em um processo de Markov de memória variável, o conjunto de sequências passadas relevantes é o conjunto de contextos. Essa característica permite representar o modelo como uma árvore probabilística de contextos, em que cada contexto é representado por um nó terminal. Chamaremos por árvore de contextos as árvores probabilísticas de contextos.

Em uma árvore de contextos, a probabilidade de uma sequência pode ser obtida através das probabilidades dos contextos e das probabilidades de transição. Por exemplo, na árvore de contextos mostrada na figura 1.3.1, os contextos são 000, 100,10 e 1 e a probabilidade
$Prob(X_0 = 0, X_{-1} = 0, X_{-2} = 0, X_{-3} = 0, X_{-4} = 1)$ é dada por

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

sendo que $Prob(X_{-1} = 0, X_{-2} = 0, X_{-3} = 0, X_{-4} = 1)$ é igual a

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

Como a sequência 000 é um dos contextos, temos que

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

Logo, a probabilidade $Prob(X_0 = 0, X_{-1} = 0, X_{-2} = 0, X_{-3} = 0, X_{-4} = 1)$ é dada por

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

já que a sequência 100 é um dos contextos.

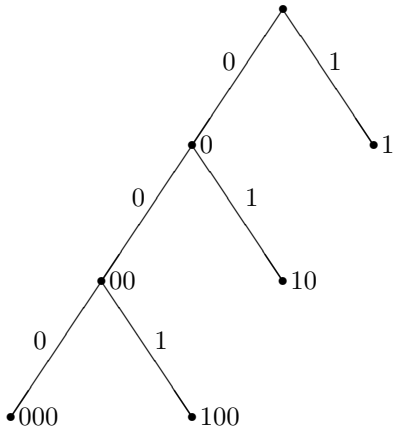


Figura 1.3.1: Exemplo de uma Árvore de Contextos Binária

1.4 Estimação de Árvores de Contextos via BIC

Nesta seção, através de uma amostra x_1^n, que é uma realização de $(X_t)_{t\in\mathbb{Z}}$, o objetivo é estimar uma árvore de contextos $\mathcal{T}_0$ e as probabilidades de transição de seus contextos. Para isto, o algoritmo proposto por Csiszár e Talata [9] é descrito a seguir.

Seja $N_n(s,a)$ o número de ocorrências da sequência $s \in \mathcal{A}^{l(s)}$ seguida do símbolo $a \in \mathcal{A}$ na amostra x_1^n supondo que o comprimento de s é no máximo $D(n)$. Como $N_n(s,a)$ é o número de ocorrências da sequência $s \in \mathcal{A}^{l(s)}$ seguida do símbolo $a \in \mathcal{A}$, deve-se considerar somente símbolos nas posições $i > D(n)$, ou seja, considera-se as $i \leq D(n)$ posições iniciais para o cálculo do número de ocorrências da sequência $s \in \mathcal{A}^{l(s)}$ seguida pelo símbolo $a \in \mathcal{A}$ o qual ocupa a posição $i = D(n)+1$. Desta forma, $N_n(s,a)$ é dado por

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

O número de ocorrências de s é dado por

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

Observação: Quando houver várias amostras independentes, $N_n(s,a)$ e $N_n(s)$, representam, respectivamente, a soma do número de ocorrências da sequência $s \in \mathcal{A}^{l(s)}$ seguida pelo símbolo $a \in \mathcal{A}$ em cada amostra e a soma do número de ocorrências de s em cada amostra.

Dada uma amostra x_1^n, uma árvore factível é qualquer árvore $\mathcal{T}$ de profundidade $d(\mathcal{T}) \leq D(n)$ tal que $N_n(s) \geq 1, \forall s \in \mathcal{T}$, e cada sequência $s^{'}$ com $N_n(s^{'}) \geq 1$ é um sufixo de algum $s \in \mathcal{T}$ ou possui um sufixo $s \in \mathcal{T}$. Uma árvore factível $\mathcal{T}$ é denominada r-frequente se $N_n(s) \geq r, \forall s \in \mathcal{T}$. A família de todas as árvores possíveis e a família das árvores r-frequentes são denotadas por $\mathcal{F}_1(x_1^n, D(n))$ e $\mathcal{F}_r(x_1^n, D(n))$, respectivamente.

Claramente, $\sum_{a \in \mathcal{A}} N_n(s,a) = N_n(s)$ e $\sum_{s \in \mathcal{T}} N_n(s) = n - D(n)$ para qualquer árvore factível $\mathcal{T}$.

Considerando a árvore $\mathcal{T}$ como uma árvore de contextos associada a um processo de lei P, a probabilidade da amostra x_1^n pode ser escrita como

$$P(x_1^n) = P\left(x_1^{D(n)}\right) \prod_{s \in \mathcal{T}, a \in \mathcal{A}} P(a|s)^{N_n(s,a)} \tag{1.4.1}$$

Para árvores de contextos $\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n))$, a verossimilhança máxima de (1.4.1), denotada por $ML(x_1^n, \mathcal{T})$, é aproximada por $\prod_{s \in \mathcal{T}, a \in \mathcal{A}} \hat{P}(a|s)^{N_n(s,a)}$ na qual $\hat{P}(a|s)$, $\forall a \in \mathcal{A}, s \in \mathcal{T}$, são os estimadores de $P(a|s)$, $\forall a \in \mathcal{A}, s \in \mathcal{T}$, que maximizam a função $\prod_{s \in \mathcal{T}, a \in \mathcal{A}} P(a|s)^{N_n(s,a)}$ sujeito à restrição $\sum_{a \in \mathcal{A}} P(a|s) = 1$ para cada $s \in \mathcal{S}$. A verossimilhança máxima $ML(x_1^n, \mathcal{T})$ é dada pela expressão

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

a qual pode ser fatorada como

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

sendo

$$\tilde{P}_s(x_1^n) = \begin{cases} \prod_{a \in \mathcal{A}} \left(\frac{N_n(s,a)}{N_n(s)} \right)^{N_n(s,a)} & se \quad N_n(s) \geq 1; \\ 1 & se \quad N_n(s) = 0. \end{cases}$$

Para a estimação da árvore $\mathcal{T}_0$ é utilizado um critério de informação que designa uma pontuação para cada modelo hipotético (neste caso, os modelos são árvores de contextos) baseada na amostra. O estimador é o modelo cuja pontuação seja a mínima. Um dos critérios de informação utilizado é o BIC [27, 7].

O BIC é um critério de informação que utiliza uma formulação bayesiana. Ele seleciona o modelo que possui a posteriori mais provável.

Considere $f(x,\theta)$ uma função densidade pertencente à família exponencial, ou seja, $f(x,\theta)$ escrita como $f(x,\theta) = \exp(\theta \mathbf{y}(x) - b(\theta))$ onde θ pertence ao espaço paramétrico natural Θ que é um subconjunto convexo do espaço euclideano K-dimensional e $\mathbf{y}$ é uma estatística suficiente K-dimensional e seja α_j a priori associada ao modelo j. O BIC escolhe j que maximiza

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

onde $\mathbf{Y} = \frac{1}{n}\sum_i \mathbf{y}(X_i)$ e m_j é a dimensão do j-ésimo modelo.

Ao longo da tese, log representa o logaritmo na base e.

Considerando um tratamento assintótico para $S(\mathbf{Y}, n, j), n \to \infty$, o BIC torna-se:

Definição 1.7. *Dada a função de verossimilhança L_{M_j} correspondente a um modelo M_j e uma amostra x_1^n, o BIC é definido por*

$$BIC(x_1^n, M) = -\log(L(x_1^n, M_j)) + \frac{k_j}{2}\log(n)$$

No escopo de árvores de contextos, $k_j = (|A| - 1)|\mathcal{T}|$. Então, a definição do BIC para árvores de contextos é:

Definição 1.8. *Dada uma amostra x_1^n, o BIC correspondente à uma árvore factível $\mathcal{T}$ é dado por*

$$BIC(x_1^n, \mathcal{T}) = -\log(ML(x_1^n, \mathcal{T})) + \frac{(|\mathcal{A}| - 1)|\mathcal{T}|}{2}\log(n)$$

O estimador para $\mathcal{T}_0$, denotado por, $\hat{\mathcal{T}}_{BIC}(x_1^n)$ é definido como sendo

$$\hat{\mathcal{T}}_{BIC}(x_1^n) = arg \min_{\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathbb{I}} BIC(x_1^n, \mathcal{T}) \tag{1.4.2}$$

Este estimador obtém árvores de contextos consistentes sendo que, para o caso $D(\mathcal{T}_0) < \infty$, consistência significa que a árvore de contextos estimada é igual à árvore $\mathcal{T}_0$, quase certamente, quando $n \to \infty$, e quase certamente significa, com probabilidade 1, $\exists n_0$ tal que a árvore de contextos estimada é igual à árvore $\mathcal{T}_0, \forall n \geq n_0$.

Teorema 1.1. *Se $d(\mathcal{T}_0) < \infty$, o estimador BIC, $\hat{\mathcal{T}}_{BIC}$, com $D(n) = o(\log n)$, satisfaz $\hat{\mathcal{T}}_{BIC}(X_1^n) = \mathcal{T}_0$, quase certamente, quando $n \to \infty$.*

No caso geral, este estimador satisfaz, para qualquer constante K, $\hat{\mathcal{T}}_{BIC}(X_1^n)\big|_K = \mathcal{T}_0\big|_K$, quase certamente, quando $n \to \infty$.

O estimador (1.4.2) pode ser escrito como

$$\hat{\mathcal{T}}(x_1^n) = arg \max_{\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathbb{I}} \prod_{s \in \mathcal{T}} \tilde{P}_s(x_1^n)$$

sendo $\tilde{P}_s(x_1^n) = n^{-\frac{|A|-1}{2}} \tilde{P}_{ML,s}(x_1^n)$. Este fato admite um tratamento conjunto do cálculo do BIC via uma extensão do algoritmo CTM [29] cuja construção é abordada a seguir.

Considere a árvore completa $\mathcal{A}^D$, $D = D(n) = o(\log(n))$, e $\mathcal{S}_D$ o conjunto de seus nós, isto é, o conjunto de todas as sequências de comprimento máximo D. Baseado na amostra x_1^n, atribui-se, para cada nó, um valor e um indicador binário. Estas atribuições são feitas recursivamente, ou seja, o valor e o indicador atribuídos a um nós são calculados a partir dos valores atribuídos aos filhos destes nós. O estimador será a subárvore determinada pelos indicadores como especificado a seguir.

Definição 1.9. *Dada a amostra x_1^n, para cada sequência $s \in \mathcal{S}_D$ com $N_n(s) \geq 1, D = D(n)$, atribuí-se, recursivamente, começando a partir das folhas da árvore completa $\mathcal{A}^D$, o valor*

$$V_s^D(x_1^n) = \begin{cases} \max \left\{ \tilde{P}_s(x_1^n), \prod_{a \in \mathcal{A}: N_n(as) \geq 1} V_{as}^D(x_1^n) \right\}, & 0 \leq l(s) < D; \\ \tilde{P}_s(x_1^n), & l(s) = D. \end{cases}$$

e o indicador

$$\aleph_s^D(x_1^n) = \begin{cases} 1 \quad se \quad 0 \leq l(s) < D \quad e \quad \prod_{a \in \mathcal{A}: N_n(as) \geq 1} V_{as}^D(x_1^n) > \tilde{P}_s(x_1^n); \\ 0 \quad se \quad 0 \leq l(s) < D \quad e \quad \prod_{a \in \mathcal{A}: N_n(as) \geq 1} V_{as}^D(x_1^n) \leq \tilde{P}_s(x_1^n); \\ 0 \quad se \quad l(s) = D. \end{cases}$$

Vale observar que, na prática, é inviável cálcular os estimadores computando-se o valor do BIC para cada modelo factível dado que o número de árvores de contextos hipotéticas é muito grande. O número de cálculos necessários para a obtenção do estimador $\hat{\mathcal{T}}_{BIC}(x_1^n)$, para uma dada amostra x_1^n, é $O(n)$ e isto pode ser alcançado armazenando-se $O(n^\epsilon)$ dados, $\epsilon > 0$ arbitrário [9].

Para finalizar a seção enunciamos dois lemas apresentados em [9] que serão utilizados no capítulo 3.

Lema 1.1. *Para distribuições de probabilidade P_1 e P_2, cujo alfabeto é $\mathcal{A}$, temos que*

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

sendo $D(\cdot||\cdot)$ a entropia relativa (veja a Seção 2.2 do Capítulo 2).

Lema 1.2. *Dado um processo de lei P, cujos espaço de estados é $\mathcal{A}$, para todo $\delta > 0$, $\exists \kappa > 0$ tal que, eventualmente quase certamente quando $n \to \infty$,*

$$\left| \frac{N_n(s,a)}{N_n(s)} - P(a|s) \right| < \sqrt{\frac{\delta \log(n)}{N_n(s)}}$$

simultaneamente para toda sequência s com $l(s) \leq \kappa \log(n)$ e $N_n(s) \geq 1$ a qual possui um sufixo na árvore contextos correspondente ao processo P.

Capítulo 2

Estimação Robusta de Árvores de Contextos

2.1 Introdução

Neste capítulo propomos uma estratégia robusta para a seleção de modelos a partir de amostras provenientes de um processo estocástico a tempo discreto tomando valores num alfabeto finito porém contaminado. Por simplicidade, consideramos o caso em que todas as contaminações são produzidas por um mesmo processo Markoviano de memória variável finita com lei Q.

Collet, Galves e Leonardi [5] mostraram que uma pequena perturbação aleatória Bernoulli na amostra proveniente de um processo Markoviano de memória variável irá, efetivamente, transformá-lo em um processo Markoviano de memória infinita. Eles, também, mostraram uma variação do algoritmo Contexto original [25] que permite recuperar da árvore de contextos da cadeia de Markov de memória variável original, desde que o ruído seja suficientemente pequeno. Neste capítulo consideramos um tipo diferente de modelo de contaminação considerando um conjunto formado por m amostras independentes das quais mais da metade delas são provenientes de um mesmo processo estocástico com lei P cujo modelo queremos estimar. No estudo por meio simulações, também, consideramos o caso em que temos somente uma amostra produzida pela concatenação de amostras do processo P e do outro processo contaminante.

Primeiramente, na Seção 2.2 são definidos os conceitos de entropia relativa entre duas variáveis

aleatórias [6] e taxa de entropia relativa entre dois processos estocásticos. Nesta tese desenvolvemos o cálculo da taxa de entropia relativa entre dois processos Markovianos de memória variável a qual é utilizada no procedimento proposto na Seção 2.3.

Na Seção 2.3 definimos o conceito de ponto de ruptura assintótico para o nosso problema de seleção de modelos e ele foi obtido.

Na Seção 2.4, a eficiência do método robusto é estudada através de simulações e, na Seção 2.5, eles são aplicados em dados linguísticos.

O capítulo é desenvolvido utilizando-se os modelos Markovianos de memória variável finita. Porém, vale observar que o procedimento proposto é válido para modelos Markovianos de ordem fixa finita, modelos Markovianos mínimos [16] e modelos multinomiais.

2.2 Entropia Relativa ou Distância de Kullback Leibler

Sejam X e Y duas variáveis aleatórias discretas, definidas num mesmo suporte (alfabeto) $\mathcal{A}$ finito, com leis P e Q, respectivamente.

Definição 2.1. *A entropia relativa ou Distância de Kullback Leibler entre duas leis P e Q é definida por*

$$D(P||Q) = \sum_{x\in\mathcal{A}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) = \mathbb{E}_P\left(\log\left(\frac{P(X)}{Q(X)}\right)\right)$$

sendo $\mathbb{E}_P$ a esperança em relação à lei P e será convencionado que $0\log\left(\frac{0}{0}\right) = 0$, $0\log\left(\frac{0}{Q}\right) = 0$ e $P\log\left(\frac{P}{0}\right) = \infty$.

Vale observar que a entropia relativa não é uma distância pois não é uma função simétrica e não satisfaz a desigualdade triangular. Apesar disto, é frequentemente útil pensá-la como uma "distância" entre distribuições.

Teorema 2.1. *Considere as leis de probabilidade P e Q definidas no mesmo alfabeto $\mathcal{A}$ finito. Então, $D(P||Q) \geq 0$ com igualdade se e somente se $P = Q$.*

Demonstração. Seja $\chi = \{x : P(x) > 0\}$. Então,

$$-D(P||Q) = -\sum_{x\in\chi} P(x)\log\left(\frac{P(x)}{Q(x)}\right) = \sum_{x\in\chi} P(x)\log\left(\frac{Q(x)}{P(x)}\right)$$

Como $\log(t)$ é uma função estritamente côncava em t, usando a desigualdade de Jensen temos

$$\sum_{x\in\chi} P(x)\log\left(\frac{Q(x)}{P(x)}\right) \leq \log\left(\sum_{x\in\chi} P(x)\frac{Q(x)}{P(x)}\right) \tag{2.2.1}$$

Logo,

$$\begin{aligned} -D(P||Q) &= \sum_{x\in\chi} P(x)\log\left(\frac{Q(x)}{P(x)}\right) \leq \log\left(\sum_{x\in\chi} P(x)\frac{Q(x)}{P(x)}\right) = \\ & \log\left(\sum_{x\in\chi} Q(x)\right) \leq \log\left(\sum_{x\in\mathcal{A}} Q(x)\right) = \log(1) = 0 \end{aligned}$$

Portanto, $D(P||Q) \geq 0$.

Desde que $\log(t)$ é uma função estritamente côncava em t, a igualdade na expressão (2.2.1) ocorre se e somente se $\frac{Q(x)}{P(x)} = 1, \forall x \in \mathcal{A}$, isto é, $P = Q$. Portanto, $D(P||Q) = 0$ se e somente se $P = Q$. □

Como a entropia relativa entre duas leis P e Q não é simétrica, consideramos a entropia relativa simetrizada definida a seguir.

Definição 2.2. *A entropia relativa simetriza entre duas leis P e Q é definida por*

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

Agora o conceito de entropia relativa é estendido para processos estocásticos.

Considere dois processos estocásticos $(X_t)_{t\in\mathbb{Z}}$ e $(Y_t)_{t\in\mathbb{Z}}$ com leis P e Q, respectivamente. A taxa de entropia relativa (entropia relativa por unidade de tempo) [6] entre $(X_t)_{t\in\mathbb{Z}}$ e $(Y_t)_{t\in\mathbb{Z}}$ é dada pela expressão

$$D\big(P||Q\big) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_P\left(\log\left(\frac{P(X_1^n)}{Q(Y_1^n)}\right)\right) \tag{2.2.2}$$

sendo $\mathbb{E}_P$ a esperança em relação à lei P.

Nesta tese, a taxa de entropia relativa é calculada para dois processos, estacionários, Markovianos de memória variável finita definidos sob um mesmo alfabeto finito $\mathcal{A}$ e com leis P e Q.

Conforme visto no Capítulo 1, um processo Markoviano de memória variável com valores em um alfabeto finito é associado a uma árvore de contextos. Desta forma, sejam $\mathcal{T}_P$ e $\mathcal{T}_Q$ as árvores

de contextos associadas aos processos de leis P e Q, respectivamente. A árvore $\mathcal{T}_{PQ}$, resultante da concatenação de $\mathcal{T}_P$ e $\mathcal{T}_Q$, é definida como

$$\mathcal{T}_{PQ} = \{s \in \mathcal{T}_P \cup \mathcal{T}_Q : \forall s' \in \mathcal{T}_P \cup \mathcal{T}_Q, s = s' \text{se } \ s \ \text{ é sufixo de } \ s'\}$$

Na Figura 2.2.1 mostramos um exemplo da concatenação das árvores $\mathcal{T}_P$ e $\mathcal{T}_Q$ resultando na árvore de contextos $\mathcal{T}_{PQ}$.

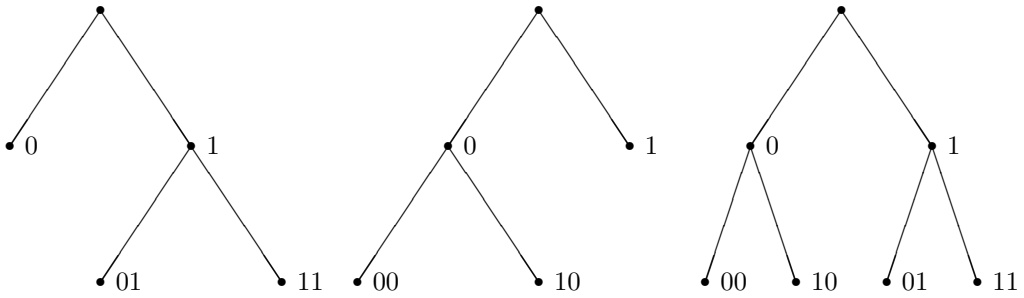


Figura 2.2.1: À esquerda: Árvore de contextos $\mathcal{T}_P$. Ao centro: Árvore de contextos $\mathcal{T}_Q$. À direita: Árvore de contextos $\mathcal{T}_{PQ}$

Teorema 2.2. *Considere dois processos, estacionários, Markovianos de memória variável finita tomando valores no alfabeto finito $\mathcal{A}$ e com leis P e Q, respectivamente. Então, a taxa de entropia relativa entre eles é dada por*

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

Demonstração. Veja Seção 5.1 do Capítulo 5. □

Agora definimos um estimador consistente da taxa de entropia relativa entre dois processos estocásticos, P e Q, estacionários, ergódicos, de ordem finita e tomando valores no alfabeto finito $\mathcal{A}$.

Para qualquer amostra x_1^n, seja $\widehat{CTM}(x_1^n)$ o modelo estimado utilizando-se o algoritmo proposto por Csiszár e Talata [9].

Vale ressaltar que há vários outros algoritmos para a estimação de árvores de contextos como, por exemplo, aqueles propostos por Rissanen [25], Buhlmann & Wyner [4], Bejerano & Yona [2], Leonardi [21, 22], Galves, Galves, Garcia, Garcia & Leonardi [13].

Considere os processos de Markov de memória variável, $(X_{1,t})_{t\in\mathbb{Z}}$ e $(X_{2,t})_{t\in\mathbb{Z}}$, com leis Q e P, respectivamente. Sejam duas amostras $(x_{1,t})_{t=1}^{n_1}$ e $(x_{2,t})_{t=1}^{n_2}$ dos processos $(X_{1,t})_{t\in\mathbb{Z}}$ e $(X_{2,t})_{t\in\mathbb{Z}}$, respectivamente. Denotemos $\hat{Q}_{n_1} = \widehat{CTM}((x_{1,t})_{t=1}^{n_1})$ e $\hat{P}_{n_2} = \widehat{CTM}((x_{2,t})_{t=1}^{n_2})$.

Lema 2.1. *Quando $n_1, n_2 \to \infty$, $D(\hat{P}_{n_2}||\hat{Q}_{n_1}) \to D(P||Q)$ quase certamente.*

Demonstração: Sem perda de generalidade, assuma que $n_1 = n_2 = n$. Existe N_0 tal que se $n > N_0$, então a árvore estimada a partir das duas amostras são as árvores verdadeiras correspondentes a Q and P (veja [9]).

Suponha que existe $s \in \mathcal{T}_{PQ}$ tal que $Q(s) = 0$. Então, $s \in \mathcal{T}_{PQ}$ implica que $P(s) > 0$ e $D(P||Q) = \infty$. Agora, $\hat{Q}_n(s) \to Q(s) = 0$ e $\hat{P}_n(s) \to P(s) > 0$ e $D(\hat{P}_n||\hat{Q}_n) \to \infty$.

Suponha que $Q(s) > 0\ \forall s \in \mathcal{T}_{PQ}$. Quando $n \to \infty$, $\hat{Q}(s) \to Q(s)$ (e $\hat{P}(s) \to P(s)$) quase certamente, $\forall s \in A^{d(\mathcal{T}_{PQ})}$ (veja [6, 8, 9]). Logo, pelo teorema do mapeamento contínuo [28], quando $n \to \infty$, $\frac{\hat{P}(s)}{\hat{Q}(s)} \to \frac{P(s)}{Q(s)}$ quase certamente, $\forall s \in A^{d(\mathcal{T}_{PQ})}$. Portanto, usando novamente o teorema do mapeamento contínuo, quando $n \to \infty$, $D(\hat{P}||\hat{Q}) \to D(P||Q)$ quase certamente. □

2.3 Estimação Robusta de Árvores de Contextos

Considere $\big\{(X_{i,t})_{t\in\mathbb{Z}}, i = 1, \ldots, m\big\}$, m processos, estacionários, Markovianos de memória variável finita M, tomando valores num mesmo alfabeto finito $\mathcal{A}$, dos quais k deles possuem lei Q e os $m - k$ restantes possuem lei P. Para cada i, $(x_{i,t})_{t=1}^{n_i}$ é uma amostra de tamanho n_i do processo $(X_{i,t})_{t\in\mathbb{Z}}$. Sem perda de generalidade, suponha que $n_i = n$ para todo $i = 1, \ldots, m$.

Denotemos por $\mathcal{C}_n^m = \{(x_{1,t})_{t=1}^n, (x_{2,t})_{t=1}^n, \ldots, (x_{m,t})_{t=1}^n\}$ uma coleção formada por m amostras independentes de tamanho n. Além disso, sejam $\hat{\mathcal{T}}$ o estimador da árvore de contextos associada a um processo Markoviano de memória variável e $\hat{\mathcal{T}}(\mathcal{C}_n^m)$ a árvore estimada utilizando-se as amostras $\mathcal{C}_n^m$.

Considerando $\hat{Q}_i(\mathcal{C}_n^m) = \widehat{CTM}((x_{i,t})_{t=1}^n)$, para cada $i \in \{1, 2, ..., m\}$, a taxa de entropia relativa

entre $\hat{Q}_i(\mathcal{C}_n^m)$ e $\hat{Q}_j(\mathcal{C}_n^m)$, denotada por $\hat{d}_{(i||j)}(\mathcal{C}_n^m)$, é

$$\hat{d}_{(i||j)}(\mathcal{C}_n^m) = D(\hat{Q}_i(\mathcal{C}_n^m)||\hat{Q}_j(\mathcal{C}_n^m))$$

e a taxa de entropia relativa simetrizada entre $\hat{Q}_i(\mathcal{C}_n^m)$ e $\hat{Q}_j(\mathcal{C}_n^m)$, denotada por $\overline{d}_{(i,j)}(\mathcal{C}_n^m)$, é

$$\overline{d}_{(i,j)}(\mathcal{C}_n^m) = \frac{\hat{d}_{(i||j)}(\mathcal{C}_n^m) + \hat{d}_{(j||i)}(\mathcal{C}_n^m)}{2}$$

Para cada $j \in \{1, 2, ..., m\}$, seja $\hat{V}_j(\mathcal{C}_n^m)$ a taxa de entropia relativa simetrizada média entre o processo estimado utilizando-se a amostra j e o processo estimado para cada uma das amostras restantes, isto é,

$$\hat{V}_j(\mathcal{C}_n^m) = \frac{1}{m}\sum_{i=1}^{m} \overline{d}_{(j,i)}(\mathcal{C}_n^m)$$

Abusando da notação, $\overline{d}_{(i,j)}(\mathcal{C}_n^m)$ é a taxa de entropia relativa simetrizada entre as amostras i and j pertencentes a $\mathcal{C}_n^m$. Além disso, $\hat{V}_j(\mathcal{C}_n^m)$ é a taxa de entropia relativa simetrizada média entre a amostra j e as demais amostras pertencentes a $\mathcal{C}_n^m$.

Ordenando, em ordem crescente, o conjunto $\{\hat{V}_j(\mathcal{C}_n^m),\ j = 1, \ldots, m\}$, denotamos $j_i^*(\mathcal{C}_n^m)$ o índice da amostra na i-ésima posição do conjunto ordenado. Por exemplo,

$$j_1^*(\mathcal{C}_n^m) = argmin_{j=1,...,m}\left\{\hat{V}_j(\mathcal{C}_n^m)\right\}$$

e, também,

$$j_m^*(\mathcal{C}_n^m) = argmax_{j=1,...,m}\left\{\hat{V}_j(\mathcal{C}_n^m)\right\}$$

A seguir definimos o ponto de ruptura assintótico de um estimador $\hat{\mathcal{T}}$ cuja construção é inspirada na versão amostral do ponto de ruptura por substituição (*finite sample replacement breakdownpoint*) [23, 30] que foi introduzido por Donoho e Huber [12]. Para isto, considere

$$\mathcal{I}_P = \{i \in \{1, ..., m\} \text{ tais que } (X_{it}) \sim P\}$$

e

$$\mathcal{I}_Q = \{i \in \{1, ..., m\} \text{ tais que } (X_{it}) \sim Q\}$$

Definição 2.3. *Dizemos que o estimador $\hat{\mathcal{T}}$ possui um ponto de ruptura assintótico igual a $\gamma \in [0,1]$ se γ é o menor valor pertencente a $[0,1]$ tal que se $\frac{|\mathcal{I}_Q|}{m} < \gamma$ então, quase certamente,*

$$\lim_{n\to\infty} \hat{\mathcal{T}}(\mathcal{C}_n^m) \neq Q$$

Teorema 2.3. *Considere o estimador definido por*

$$\hat{\mathcal{T}}_i(\mathcal{C}_n^m) = \hat{Q}_{j_i^*(\mathcal{C}_n^m)}(\mathcal{C}_n^m)$$

Então, $\hat{\mathcal{T}}_i(\mathcal{C}_n^m)$ possui ponto de ruptura assintótico igual a $\frac{1}{2}$ para todo $i < \frac{m}{2}$.

Demonstração. Veja Seção 5.1 do Capítulo 5. □

Definição 2.4. *Definimos o estimador CTM α-truncado para $\mathcal{C}_n^m$ como sendo*

$$\hat{\mathcal{T}}_\alpha(\mathcal{C}_n^m) = CTM\left(\left\{\left(x_{j_1^*(\mathcal{C}_n^m),t}\right)_{t=1}^n, \ldots, \left(x_{j^*_{[(1-\alpha)m]}(\mathcal{C}_n^m),t}\right)_{t=1}^n\right\}\right) \tag{2.3.1}$$

para α tal que $[(1-\alpha)m] \geq 1$ sendo $[(1-\alpha)m]$ a parte inteira de $(1-\alpha)m$.

Teorema 2.4. *Para $0 < \alpha < \frac{1}{2}$, $\hat{\mathcal{T}}_\alpha(\mathcal{C}_n^m)$ possui ponto de ruptura assintótico $\gamma = \alpha$. Para $\frac{1}{2} \leq \alpha \leq 1$, $\hat{\mathcal{T}}_\alpha(\mathcal{C}_n^m)$ possui ponto de ruptura assintótico $\gamma = \frac{1}{2}$.*

Demonstração. A prova é consequência direta do Teorema 2.3. □

Na sequência do capítulo, o procedimento proposto é aplicado a dados reais e a dados simulados. Para estes casos, usamos os parâmetros $\alpha = (1 - \frac{1}{m})$ e $\alpha = \frac{1}{2}$. O valor $\alpha = (1 - \frac{1}{m})$ corresponde a escolher o modelo estimado para a amostra minimizando-se a taxa de entropia relativa média, isto é, $\hat{Q}_{j_1^*(\mathcal{C}_n^m)}(\mathcal{C}_n^m)$.

2.4 Simulações

Nesta seção, através de simulações, comparamos a eficiência do método proposto por Csiszár & Talata (Método CTM que corresponde ao Método CTM truncado com $\alpha = 1$) [9] com o método proposto (Método CTM α-truncado) com $\alpha = (1 - \frac{1}{m})$ e $\alpha = \frac{1}{2}$ sendo m o número de amostras independentes de tamanho n.

Gerando-se 1.000 vezes m amostras independentes de tamanho n, os métodos são comparados através da porcentagem de acertos, ou seja, a porcentagem de vezes em que o método selecionou o modelo correto.

Nas Subseções 2.4.1 e 2.4.2 são geradas m amostras independentes de tamanho n, das quais k e $m-k$ correspondem, respectivamente, aos processos de Markov de memória variável representados pelas árvores de contextos $\mathcal{T}_Q$ e $\mathcal{T}_P$. Na primeira subseção as árvores não possuem contextos em comum enquanto que, na segunda, as árvores possuem alguns contextos em comum.

Nas Subseções 2.4.1 e 2.4.2, cada amostra é gerada ou do processo de lei P ou do processo de lei Q (amostras contaminadas). Já, na Subseção 2.4.3, considera-se que, em uma mesma amostra, há alguns pedaços do processo de lei P e outros do processo de lei Q.

2.4.1 Simulação 1

São geradas m amostras independentes de tamanhos 25, 50, 100 e 200 das quais k e $m-k$ correspondem, respectivamente, aos processos de Markov de memória variável representados pelas árvores de contextos $\mathcal{T}_Q$ e $\mathcal{T}_P$, ambas definidas sob o mesmo alfabeto $\mathcal{A} = \{0, 1\}$.

Os contextos s e as probabilidades condicionais $\big(Prob(0|s), Prob(1|s)\big)$ das árvores $\mathcal{T}_P$ e $\mathcal{T}_Q$ são mostradas na Figura 2.4.1.

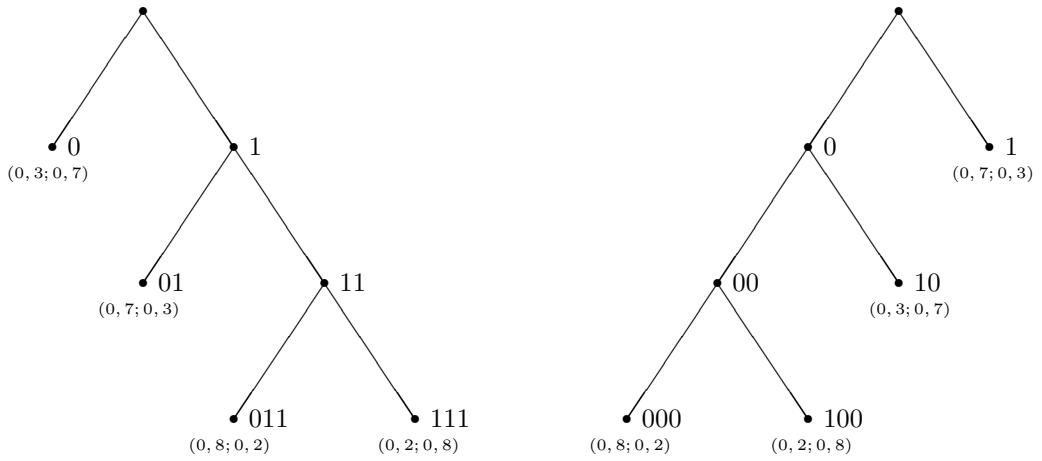


Figura 2.4.1: À esquerda: Árvore de contextos $\mathcal{T}_P$. À direita: Árvore de contextos $\mathcal{T}_Q$

Pelas Tabelas 4.1, 4.2, 4.3 e 4.4 observa-se que, para o caso em que há 9, 15 e 50 amostras independentes, a performance do Método CTM truncado com $\alpha = \frac{1}{2}$ é superior à dos demais para quaisquer valores de k (número de amostras contaminadas) e n (tamanho amostral). Para o caso em que há 5 amostras, o Método CTM apresenta as maiores porcentagens de acertos apenas para $n = 50$.

A performance do Método CTM truncado com $\alpha = \frac{1}{2}$ sempre é superior à do CTM truncado com $\alpha = (1 - \frac{1}{m})$. Comparando o Método CTM truncado com $\alpha = (1 - \frac{1}{m})$ com o CTM, para amostras de tamanho n=200, independentemente dos valores de m e k, a superioridade deste primeiro é nítida.

A melhora do Método CTM truncado com $\alpha = (1 - \frac{1}{m})$, para amostras grandes, é devido ao fato de que, neste método, uma árvore é estimada de cada uma das amostras enquanto que, nos demais, uma árvore de contextos é estimada a partir de uma coleção de amostras independentes, ou seja, neste método, necessita-se de amostras grandes para estimar uma árvore que capte a estrutura do processo de lei P.

Observa-se que, nos três métodos, a porcentagem de acertos diminui conforme aumenta-se o número de amostras contaminadas k.

Tabela 4.1: Porcentagem de acertos para o caso em que são geradas m=5 amostras independentes

			Métodos		
Tamanho Amostral	$m-k$	k	CTM	CTM Truncado com $\alpha = (1 - \frac{1}{m})$	CTM Truncado com $\alpha = \frac{1}{2}$
n=50	4	1	76,4%	26,6%	55,7%
	3	2	55,4%	19,5%	35,9%
n=100	4	1	72,1%	51,0%	77,0%
	3	2	44,8%	33,1%	51,2%
n=200	4	1	65,4%	89,3%	95,2%
	3	2	22,6%	70,8%	76,0%

Tabela 4.2: Porcentagem de acertos para o caso em que são geradas m=9 amostras independentes

Tamanho Amostral	$m-k$	k	Métodos		
			CTM	CTM Truncado com $\alpha = (1 - \frac{1}{m})$	CTM Truncado com $\alpha = \frac{1}{2}$
n=50	8	1	79,1%	32,9%	90,6%
	7	2	63,7%	25,1%	83,8%
	6	3	53,2%	21,9%	73,8%
	5	4	38,7%	15,4%	59,9%
n=100	8	1	80,3%	60,4%	97,1%
	7	2	58,4%	52,7%	93,4%
	6	3	35,5%	41,9%	85,7%
	5	4	17,7%	28,1%	69,4%
n=200	8	1	74,9%	96,2%	98,3%
	7	2	37,1%	92,7%	98,3%
	6	3	11,1%	88,4%	94,7%
	5	4	1,7%	67,2%	75,7%

Tabela 4.3: Porcentagem de acertos para o caso em que são geradas m=15 amostras independentes

Tamanho Amostral	$m-k$	k	Métodos		
			CTM	CTM Truncado com $\alpha = (1 - \frac{1}{m})$	CTM Truncado com $\alpha = \frac{1}{2}$
n=50	14	1	78,9%	27,4%	96,4%
	12	3	55,8%	27,1%	94,5%
	10	5	30,4%	19,6%	86,2%
	8	7	12,3%	13,2%	74,3%
n=100	14	1	81,8%	57,0%	97,5%
	12	3	44,2%	55,4%	97,2%
	10	5	11,5%	42,0%	90,9%
	8	7	1,9%	23,0%	67,9%
n=200	14	1	82,9%	95,6%	98,6%
	12	3	25,2%	94,3%	99,3%
	10	5	1,3%	88,3%	96,2%
	8	7	0,0%	66,4%	51,3%

Tabela 4.4: Porcentagem de acertos para o caso em que são geradas m=50 amostras independentes

Tamanho Amostral	$m-k$	k	Métodos		
			CTM	CTM Truncado com $\alpha = (1 - \frac{1}{m})$	CTM Truncado com $\alpha = \frac{1}{2}$
n=25	49	1	25,2%	7,5%	95,4%
	45	5	9,3%	8,3%	91,7%
	40	10	1,5%	6,7%	79,1%
	35	15	0,1%	7,1%	62,5%
	30	20	0,0%	7,9%	41,3%
	26	24	0,0%	8,0%	26,3%
n=50	49	1	35,0%	7,6%	96,3%
	45	5	13,8%	6,7%	92,8%
	40	10	1,4%	7,5%	84,0%
	35	15	0,1%	6,7%	64,4%
	30	20	0,0%	5,4%	36,0%
	26	24	0,0%	3,7%	17,2%
n=100	49	1	54,4%	24,7%	98,1%
	45	5	13,8%	23,3%	95,8%
	40	10	0,3%	26,2%	91,1%
	35	15	0,0%	21,2%	70,7%
	30	20	0,0%	14,6%	34,9%
	26	24	0,0%	8,3%	70,0%
n=200	49	1	63,7%	81,8%	98,7%
	45	5	7,5%	84,1%	99,3%
	40	10	0,0%	84,8%	98,3%
	35	15	0,0%	84,5%	94,7%
	30	20	0,0%	74,7%	47,6%
	26	24	0,0%	67,9%	2,0%

2.4.2 Simulação 2

São geradas m amostras independentes de tamanhos 100, 250, 500 e 1.000 das quais k e $m-k$ correspondem, respectivamente, aos processos de Markov de memória variável representados pelas árvores $\mathcal{T}_Q$ e $\mathcal{T}_P$, ambas definidas sob o mesmo alfabeto $\mathcal{A} = \{0, 1, 2\}$, as quais possuem alguns contextos em comum.

Os contextos s e as probabilidades condicionais $\big(Prob(0|s), Prob(1|s), Prob(2|s)\big)$ das árvores $\mathcal{T}_P$ e $\mathcal{T}_Q$ são mostradas na Figura 2.4.2.

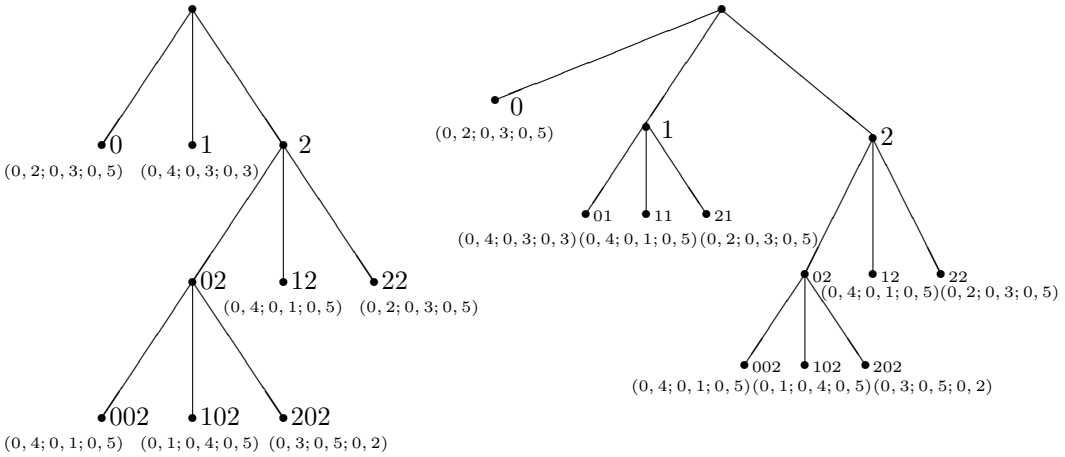


Figura 2.4.2: À esquerda: Árvore de contextos $\mathcal{T}_P$. À direita: Árvore de contextos $\mathcal{T}_Q$

Pelas Tabelas 4.5, 4.6, 4.7 e 4.8 observa-se que, em geral, a performance do Método CTM truncado com $\alpha = \frac{1}{2}$ é superior à dos demais.

A performance do Método CTM truncado com $\alpha = \frac{1}{2}$ sempre é superior à do CTM truncado com $\alpha = (1 - \frac{1}{m})$. Além disso, comparando o Método CTM com o CTM truncado com $\alpha = (1 - \frac{1}{m})$ amostras de tamanho n=1.000, independentemente dos valores de m e k, a superioridade deste último é nítida.

A melhora do Método CTM truncado com $\alpha = (1 - \frac{1}{m})$, para amostras grandes, é devido ao fato de que uma árvore de contextos é estimada de cada uma das amostras enquanto que, nos demais, estima-se uma árvore de contextos a partir de uma coleção de amostras independentes, ou seja, neste método, necessita-se de amostras grandes para estimar uma árvore que capte a estrutura do

processo de lei P.

Observa-se que o Método CTM truncado com $\alpha = \frac{1}{2}$ apresentou uma diminuição abrupta na porcentagem de acertos para valores de m e k tais que $k = \lceil m/2 \rceil$.

Tabela 4.5: Porcentagem de acertos para o caso em que são geradas m=5 amostras independentes

			Métodos		
Tamanho Amostral	$m-k$	k	CTM	CTM Truncado com $\alpha = (1 - \frac{1}{m})$	CTM Truncado com $\alpha = \frac{1}{2}$
n=250	4	1	70,9%	0,0%	8,0%
	3	2	2,1%	0,0%	9,6%
n=500	4	1	55,5%	1,8%	55,6%
	3	2	0,0%	1,5%	57,9%
n=1.000	4	1	6,0%	69,5%	98,7%
	3	2	0,0%	62,8%	99,0%

Tabela 4.6: Porcentagem de acertos para o caso em que são geradas m=9 amostras independentes

Tamanho Amostral	$m-k$	k	Métodos		
			CTM	CTM Truncado com $\alpha = (1 - \frac{1}{m})$	CTM Truncado com $\alpha = \frac{1}{2}$
n=250	8	1	97,6%	0,0%	60,9%
	7	2	34,0%	0,0%	63,1%
	6	3	0,5%	0,0%	65,1%
	5	4	0,0%	0,0%	55,4%
n=500	8	1	88,1%	0,2%	99,4%
	7	2	1,0%	1,0%	99,1%
	6	3	0,0%	0,2%	99,2%
	5	4	0,0%	0,2%	74,9%
n=1.000	8	1	48,7%	73,9%	100,0%
	7	2	0,0%	72,2%	100,0%
	6	3	0,0%	66,5%	100,0%
	5	4	0,0%	65,6%	64,3%

Tabela 4.7: Porcentagem de acertos para o caso em que são geradas m=15 amostras independentes

Tamanho Amostral	$m-k$	k	Métodos		
			CTM	CTM Truncado com $\alpha = (1 - \frac{1}{m})$	CTM Truncado com $\alpha = \frac{1}{2}$
n=250	14	1	97,9%	0,0%	97,5%
	12	3	7,2%	0,0%	98,0%
	10	5	0,0%	0,0%	98,1%
	8	7	0,0%	0,0%	48,6%
n=500	14	1	95,9%	0,0%	100,0%
	12	3	0,0%	0,1%	100,0%
	10	5	0,0%	0,0%	100,0%
	8	7	0,0%	0,2%	39,8%
n=1.000	14	1	80,2%	80,1%	100,0%
	12	3	0,0%	76,8%	100,0%
	10	5	0,0%	72,3%	100,0%
	8	7	0,0	50,4%	18,8%

Tabela 4.8: Porcentagem de acertos para o caso em que são geradas m=50 amostras independentes

Tamanho Amostral	$m-k$	k	Métodos		
			CTM	CTM Truncado com $\alpha = (1 - \frac{1}{m})$	CTM Truncado com $\alpha = \frac{1}{2}$
n=100	49	1	13,7%	0,0%	94,6%
	45	5	1,9%	0,0%	95,6%
	40	10	0,0%	0,0%	92,6%
	35	15	0,0	0,0%	85,3%
	30	20	0,0%	0,0%	32,1%
	26	24	0,0%	0,0%	1,9%
n=250	49	1	33,3%	0,0%	98,2%
	45	5	0,1%	0,0%	98,4%
	40	10	0,0%	0,0%	98,8%
	35	15	0,O%	0,0%	97,9%
	30	20	0,0%	0,0%	68,0%
	26	24	0,0%	0,0%	0,3%
n=500	49	1	47,5%	0,1%	99,4%
	45	5	0,0%	0,0%	99,4%
	40	10	0,0%	0,0%	99,6%
	35	15	0,0%	0,0%	99,1%
	30	20	0,0%	0,0%	79,4%
	26	24	0,0%	0,1%	0,1%
n=1.000	49	1	52,0%	96,2%	99,8%
	45	5	0,0%	94,0%	99,9%
	40	10	0,0%	90,2%	100,0%
	35	15	0,0%	86,9%	99,8%
	30	20	0,0%	78,2%	95,6%
	26	24	0,0%	23,5%	0,0%

2.4.3 Simulação 3

Nesta subseção, cada amostra é formada por alguns pedaços do processo de lei P e por outros do processo de lei Q, ou seja, a contaminação está presente em alguns pedaços de uma mesma amostra. Para isto, são geradas duas amostras de tamanho n sendo que a primeira e a segunda correspondem, respectivamente, aos processos de Markov de memória variável $\{X_t\}_{t\in\mathbb{Z}}$ e $\{Y_t\}_{t\in\mathbb{Z}}$ representados pelas árvores de contextos $\mathcal{T}_P$ e $\mathcal{T}_Q$. Estas árvores são as mesmas utilizadas na Subseção 2.4.1.

Em seguida, é gerada uma amostra de um processo Markoviano, $\{W_t\}_{t\in\mathbb{Z}}$, de ordem 1 tomando valores no alfabeto $\{0,1\}$ e possuindo probabilidades de transição, $Prob(W_t = 0|W_{t-1} = 0)$ e $Prob(W_t = 1|W_{t-1} = 1)$, próximas de um. A partir de uma amostra gerada deste processo, obtém-se uma amostra do processo $\{Z_t\}_{t\in\mathbb{Z}}$ definido, para cada t, por

$$Z_t = \begin{cases} X_t & \text{se} \quad W_t = 0 \\ Y_t & \text{se} \quad W_t = 1 \end{cases}$$

sendo $Z_0 \equiv 0$.

A amostra de $\{Z_t\}_{t\in\mathbb{Z}}$ é formada pela junção de amostras de $\{X_t\}_{t\in\mathbb{Z}}$ e $\{Y_t\}_{t\in\mathbb{Z}}$. Finalmente, a amostra gerada de $\{Z_t\}_{t\in\mathbb{Z}}$ é cortada em m pedaços e, em cada pedaço, são eliminados os três primeiros valores para que a dependência entre eles seja desprezível. Dessa forma, cada um dos pedaços passa a ser uma amostra.

Os Métodos CTM e CTM truncado com $\alpha = (1 - \frac{1}{m})$ e $\alpha = \frac{1}{2}$ são aplicados na coleção formada pelas amostras (pedaços da amostra gerada de $\{Z_t\}_{t\in\mathbb{Z}}$) de tamanho n construídas conforme a descrição apresentada.

Pela tabela 4.9 observa-se que, para esta situação em que a contaminação está presente em alguns pedaços de uma mesma amostra, o Método CTM truncado com $\alpha = (1 - \frac{1}{m})$ foi superior aos demais métodos em todas as situações analisadas.

Tabela 4.9: Porcentagem de acertos

Tamanho da Amostra gerada de $\{Z_t\}_{t\in\mathbb{Z}}$	$P(0\|0)$	$P(1\|1)$	n	Métodos		
				Método CTM	Método CTM Truncado com $\alpha = (1 - \frac{1}{m})$	Método CTM Truncado com $\alpha = \frac{1}{2}$
5.000	0,999	0,993	200	55,2%	82,6%	48,7%
			500	57,9%	97,9%	88,8%
			1.000	57,3%	92,5%	87,9%
	0,999	0,995	200	41,2%	78,2%	43,3%
			500	44,5%	96,0%	85,9%
			1.000	39,9%	84,1%	81,2%
	0,999	0,990	200	72,4%	82,5%	46,4%
			500	70,6%	99,7%	91,6%
			1.000	68,7%	97,4%	93,3%
10.000	0,999	0,993	200	34,6%	70,9%	0,5%
			500	32,0%	99,1%	66,8%
			1.000	30,5%	95,5%	79,8%
	0,995	0,999	200	16,8%	69,0%	0,8%
			500	18,0%	97,4%	60,6%
			1.000	19,0%	88,4%	71,7%
	0,999	0,990	200	53,3%	69,7%	0,8%
			500	55,9%	99,8%	68,5%
			1.000	51,3%	99,2%	88,7%

2.5 Aplicação

Como os idiomas diferem em seus ritmos da fala [1], em Linguística, o estudo de correlação acústica de classes rítmicas é um problema de grande interesse. Originalmente, três classes rítmicas foram propostas:

(i) Acentual (*stress-timed languages);*

(ii) Silábica (syllable-timed languages);

(iii) Moraica (mora-timed languages).

Essas classes são baseadas na ideia de que, em cada classe, elementos de diferenciação causam a organização temporal. De acordo com psicolinguistas, o tipo rítmico pode ser correlacionado com a unidade de segmentação da fala. De fato, falantes de idiomas pertencentes às classes acentual, silábica e moraica podem segmentar a fala, respectivamente, em unidades rítmicas, sílabas e moras.

Dauer [11] e Ramus et al. [24] extraíram duas principais propriedades fonéticas/fonológicas que diferenciam as classes (i) e (ii), citadas anteriormente. Estas características são:

(a) Estrutura Silábica: A classe acentual possui uma maior variedade de tipos silábicos em relação à classe silábica;

(b) Redução Vocálica: Na classe acentual, usualmente, sílabas não acentuadas possuem um sistema de redução vocálica.

De acordo com essa classificação, os idiomas Francês, Italiano e Espanhol pertencem à classe rítmica silábica e o Inglês e o Holandês pertencem à classe rítmica acentual. Já o Japonês pertence à classe moraica

Em 1999, Ramus, Nespor e Mehler [24], obtiveram evidências de que estatísticas baseadas na segmentação manual de vogais e consoantes de um sinal da fala poderia ser utilizado para a diferenciação de classes rítmicas. Porém, o problema desta metodologia é o trabalho manual feito por um foneticista. Por causa disto, eles usaram conjuntos de dados pequenos e, assim, os resultados não são estatisticamente significantes. Galves et al. [14] e García & González-López [17] obtiveram resultados similares à Ramus com o uso de uma metodogia completamente automatizada sem a necessidade de trabalhos manuais.

2.5.1 Os Dados

Nesta aplicação são utilizadas 153, 212, 212, 216, 228, 216, 216, 216 sentenças pertencentes aos idiomas Inglês, Japonês, Espanhol, Francês, Holandês, Italiano, Polonês e Catalão, respectivamente. Cada falante lê cerca de 50 sentenças com duração indo de 2 à 3,5 segundos, digitalizadas em 16.000 amostras por segundo (isto é, a uma taxa amostral de 16 kHz). Estes dados integram o corpus pertecente à *Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS/CNRS)*. O corpus inclui as 160 sentenças analizadas em [24].

2.5.2 Bandas de Energia

Denotando-se por $\vartheta_t(f)$, a densidade espectral no tempo t e frequência f é o quadrado do coeficiente, para a frequência f, da decomposição de Fourier local do sinal acústico. O tempo e a frequência foram discretizados, respectivamente, a cada 25 milisegundos e 20 Hz. Os valores da densidade espectral são estimados utilizando-se uma janela Gaussiana de 25 milisegundos.

Fixado um idioma l, considere a sentença j de tamanho $T_{l,j}$. Dada uma frequência f, denota-se por $\vartheta_t^{l,j}(f)$ a densidade espectral no tempo t, $t = 1, \cdots, T_{l,j}$, para a sentença j. Para cada tempo t, sejam os processos estocásticos

$$\chi_1^{l,j}(t) = \sum_{f=80,100,\ldots,800} \vartheta_t^{l,j}(f)$$

e

$$\chi_2^{l,j}(t) = \sum_{f=1500,1520,\ldots,5000} \vartheta_t^{l,j}(f)$$

os quais são chamados de energia.

As frequências para as bandas foram escolhidas baseadas em trabalhos anteriores [14] sobre segmentação automática em vogais e consoantes.

2.5.3 Codificação dos Dados

Para cada idioma l e sentença j deste idioma, considere uma banda de frequências b ($b = 1$ representa a banda inferior de energia $\chi_1^{l,j}(t)$ e $b = 2$ representa a banda superior de energia $\chi_2^{l,j}(t)$).

Defina

$$Y_t^{l,j,b} = \begin{cases} 1 & \text{se} \quad \chi_b^{l,j}(t+1) \geq \chi_b^{l,j}(t) \\ 0 & \text{se} \qquad \text{caso contrário} \end{cases}$$

Seja $Z_t^{l,j} = 2 * Y_t^{l,j,2} + Y_t^{l,j,1}$. Esta variável pode ser interpretada como:

- O valor $Z_t^{l,j} = 0$ significa que ambas energias decrescem no tempo $t+1$;
- O valor $Z_t^{l,j} = 1$ significa que a energia na banda inferior cresce e a energia na banda superior decresce no tempo $t+1$;
- O valor $Z_t^{l,j} = 2$ significa que a energia na banda superior cresce e a energia na banda inferior decresce no tempo $t+1$;
- O valor $Z_t^{l,j} = 3$ significa que ambas energias crescem no tempo $t+1$.

2.5.4 Resultados

O Método CTM truncado com $\alpha = (1 - \frac{1}{m})$ e $\alpha = \frac{1}{2}$ é utilizado para a estimação de uma árvore de contextos para cada idioma usando-se os valores de $\left\{ Z_t^{l,j} \right\}_{t=1}^{T_{l,j}}$. Em seguida, a taxa de entropia relativa simetrizada é utilizada como uma medida de "distância" entre as árvores selecionadas de cada idioma.

As distâncias obtidas, utilizando-se as árvores estimadas pelo Método CTM truncado com $\alpha = (1 - \frac{1}{m})$, são agrupadas conforme a Figura 2.5.1 na qual observa-se os seguintes grupos:

- Holandês e Inglês;
- Japonês;
- Francês, Italiano, Polonês e Espanhol;
- Catalão.

Além disso, o Método K-means com 4 grupos confirma o agrupamento supra-citado o qual coincide com a conjectura linguística de classes rítmicas em que o Inglês e o Holandês pertencem à

classe rítmica acentual, o Japonês pertence à classe rítmica moraica e os idiomas Francês, Italiano e Espanhol pertencem à classe rítmica silábica. Veja, também, os resultados obtidos por García & González-López [17].

Já o Catalão e o Polonês são considerados idiomas mistos. Conforme relatado por Ramus et al. [24], os idiomas intermediários misturam as seguintes características: estrutura silábica e redução vocálica. O Polonês possui uma grande complexidade silábica mas sem a redução vocálica esperada para a classe rítmica acentual enquanto que o Catalão possui o mesmo sistema silábico do Espanhol mas com a presença da redução vocálica.

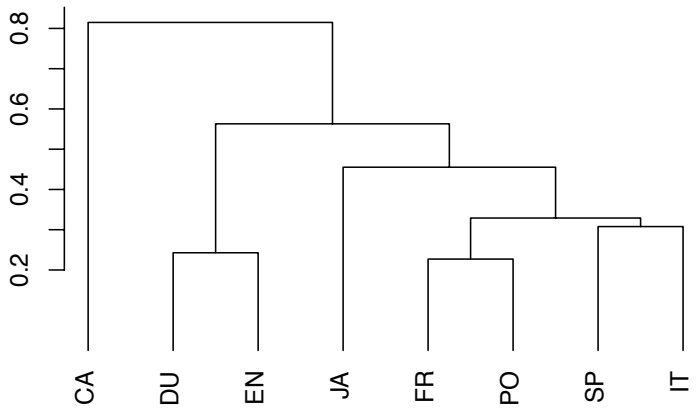


Figura 2.5.1: Cluster obtido a partir das distâncias entre árvores estimadas pelo Método CTM truncado com $\alpha = (1 - \frac{1}{m})$ considerando os idiomas Holandês, Espanhol, Francês, Inglês, Italiano, Japonês, Polonês e Catalão rotulados como DU, SP, FR, EN, IT, JA, PO, e CA, respectivamente

Utilizando-se as árvores estimadas pelo Método CTM truncado com $\alpha = \frac{1}{2}$, as distâncias obtidas são agrupadas conforme a Figura 2.5.2 na qual observa-se os seguintes grupos:

- Catalão e Espanhol;
- Japonês, Francês e Italiano;
- Polonês, Holandês e Inglês;

Além disso, o Método K-means com 3 grupos confirma o agrupamento supra-citado. A classicação do Japonês como pertencente à classe rítmica silábica não coincide com a conjectura linguística já que ele pertence à classe rítmica moraica. Note que os idiomas Francês e Italiano estão mais próximos entre si do que em relação ao Japonês.

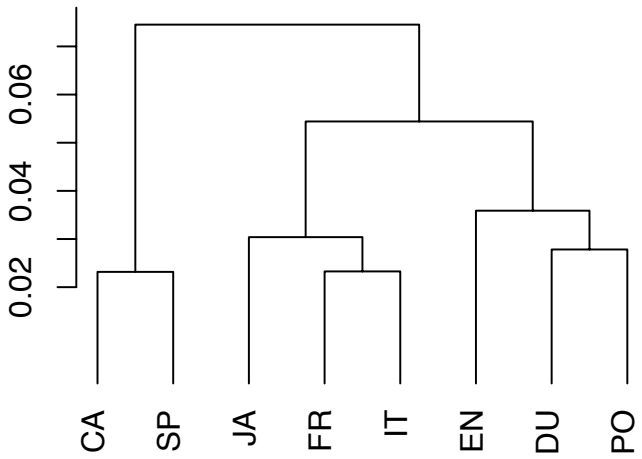


Figura 2.5.2: Dendograma obtido a partir das distâncias entre árvores estimadas pelo Método CTM truncado com $\alpha = \frac{1}{2}$ considerando os idiomas Holandês, Espanhol, Francês, Inglês, Italiano, Japonês, Polonês e Catalão rotulados como DU, SP, FR, EN, IT, JA, PO, e CA, respectivamente

Capítulo 3

Independência Condicional entre as coordenadas de um Processo de Markov l-variado

3.1 Introdução

Considere $(Z_t)_{t\in\mathbb{Z}} = \left(Z_t^{(1)}, \ldots, Z_t^{(l)}\right)_{t\in\mathbb{Z}}$ um processo de Markov estacionário de ordem finita M com alfabeto $\mathcal{A} = \mathcal{B}_1 \times \ldots \times \mathcal{B}_l$ sendo $\mathcal{B}_i$ o alfabeto da i-ésima coordenada de $(Z_t)_{t\in\mathbb{Z}}$, ou seja, o alfabeto de $\left(Z_t^{(i)}\right)_{t\in\mathbb{Z}}$. O modelo associado a este processo possui $(|\mathcal{A}|-1)|\mathcal{A}|^M$ parâmetros a serem determinados sendo que o número de parâmetros de uma cadeia de Markov cresce exponencialmente com a sua ordem. Uma descrição mais eficiente seria aquela em que a ordem da cadeia de Markov não é fixa, ou seja, a memória da cadeia de Markov estacionária é variável, uma função dos valores do passado. Este tipo de modelo Markoviano foi originado através do trabalho de Jorma Rissanen em 1986 [26] e motivou o surgimento dos modelos VLMC (Cadeias de Markov de Memória Variável) (ver Buhlmann e Wyner [4], Csiszár e Talata [9], Galves et al. [13], Leonardi [21, 22]) que requer um número muito menor de parâmetros a serem determinados.

Um problema interessante relativo à estas famílias de modelos, que têm sido estudado originalmente por Rissanen [26] e mais recentemente por Csiszár e Talata [9], Galves [13], Leonardi [21, 22], é a seleção de um modelo, dentro de famílias de modelos, que se ajuste a uma fonte de dados. Nesta

família, os modelos são indexados por um conjunto de árvores prefixo e um método consistente para a escolha do modelo é o BIC (Critério de Informação Bayesiano).

Uma outra família de modelos que inclui a dos VLMC (Cadeias de Markov de Memória Variável) é a dos modelos de Markov de partições. Esta família permite uma economia de parâmetros ainda maior [16].

Neste capítulo, o objetivo é a criação de uma família de modelos Markovianos que inclua as cadeias de Markov de alcance fixo e que permita modelar este tipo de cadeia com o menor número possível de parâmetros aproveitando a estrutura de dependência condicional para diferentes combinações de estados. Primeiramente definimos uma família de modelos que inclua as cadeias de Markov de alcance fixo e que é indexada pela estrutura de dependência condicional entre as cadeias marginais no espaço de estados. Além disso, apresentamos uma metodologia consistente de seleção de modelos baseada no BIC que é implementada por um algoritmo. A eficiência do algoritmo é estudada por simulações e este é aplicado em dados linguísticos.

O capítulo será desenvolvido utilizando-se os modelos Markovianos de ordem fixa finita. Porém, vale observar que o procedimento proposto neste capítulo é válido para modelos Markovianos de memória variável finita e modelos multinomiais multivariados.

Na Seção 3.2 enunciamos algumas definições relacionadas à Teoria de Conjunto e Probabilidade que são utilizadas na tese. Na Seção 3.3 obtemos os estimadores das probabilidades de transição de um processo Markoviano estacionário l-variado e de um processo Markoviano estacionário l-variado com partes condicionalmente independentes. Estes estimadores são utilizados na Seção 3.4 para o estabelecimento de um critério que possibilite a obtenção da estrutura de independência condicional do processo. Tal critério é implementado por um algoritmo cuja eficiência é estudada por simulações na Seção 3.5 e ele é aplicado em dados linguísticos na Seção 3.6.

3.2 Definições Básicas

Nesta seção enunciamos algumas definições relacionadas à Teoria de Conjunto e Probabilidade.

Definição 3.1. *Seja $A_1, A_2, \ldots$ uma sequência de subconjuntos não-vazios de um conjunto A. Esta sequência de subconjuntos é uma partição de A se $\cup_{i=1}^{\infty} A_i = A$ e $A_i \cap A_j = \emptyset$ para todo $i \neq j$ [20, 3].*

Definição 3.2. *Seja Π o conjunto de todas as partições de um conjunto A. Dadas duas partições π*

e σ pertencentes a Π, π é mais fina que σ, denotado por $\pi \preceq \sigma$, se e somente se, $\forall B \in \pi$, $\exists C \in \sigma$ tal que $B \subseteq C$ [20, 3].

O refinamento define uma relação de ordem parcial em Π sendo que uma relação é de ordem parcial se ela é reflexiva, anti-simétrica e transitiva [20, 3].

Os Teoremas 3.3, 3.4 e 3.5 utilizam os conceitos de convergência eventual e quase certa [19].

Definição 3.3. *A sequência $(x_n)_{n\in\mathbb{N}}$ de números reais converge eventualmente para x se $\exists n_o$ tal que $x_n = x$ para todo $n \geq n_o$.*

Definição 3.4. *Sejam $X, X_1, X_2, \ldots$ variáveis aleatórias definidas em um mesmo espaço de probabilidade. A sequencia $(X_n)_{n\in\mathbb{N}}$ converge para X quase certamente se*

$$P\big(X_n \overset{n\to\infty}{\longrightarrow} X\big) = 1$$

isto é, o evento $[w : X_n(w) \to X(w)]$ é de probabilidade 1.

3.3 Estimadores das Probabilidades de Transição de uma Cadeia de Markov l-variada

Considere $(Z_t)_{t\in\mathbb{Z}}$ um processo de Markov estacionário l-variado de ordem finita M tomando valores em $\mathcal{A} = \mathcal{B}_1 \times \ldots \times \mathcal{B}_l$ sendo $\mathcal{B}_i$, $i = 1, \ldots, l$, o alfabeto da i-ésima coordenada de $(Z_t)_{t\in\mathbb{Z}}$. Sejam $\mathcal{S} = \mathcal{A}^M$ e $z_1^n = z_1 \ldots z_n$, $z_i \in \mathcal{A}$, $i = 1, \ldots, n$, respectivamente, o espaço de estados e uma amostra desse processo.

Sem perda de generalidade, no decorrer do capítulo, será considerado $\mathcal{B}_1 = \cdots = \mathcal{B}_l = \mathcal{B}$ e, consequentemente, o alfabeto associado ao processo $(Z_t)_{t\in\mathbb{Z}}$ será $\mathcal{A} = \mathcal{B}^l$.

Seja $N_n(s, a)$ o número de ocorrências da sequência $s \in \mathcal{A}^M$ seguida do símbolo $a \in \mathcal{A}$ na amostra z_1^n, ou seja,

$$N_n(s, a) = \Big|\Big\{u : M < u \leq n, z_{u-M}^{u-1} = s, z_u = a\Big\}\Big|$$

Note que $\sum_{a\in\mathcal{A}} N_n(s, a) = N_n(s)$.

O número de ocorrências da sequência s é dado por

$$N_n(s) = \Big|\Big\{u : M < u \leq n, z_{u-M}^{u-1} = s\Big\}\Big|$$

A função de verossimilhança, denotada por $L\big(z_1^n|\mathcal{P}\big)$, é dada por

$$L\big(z_1^n|\mathcal{P}\big) = Prob\big(Z_1^n = z_1^n\big) = Prob\big(z_1^M\big) \prod_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} P(a|s)^{N_n(s,a)}$$

sendo $P(a|s) = Prob(Z_0 = a|Z_{-M}^{-1} = s)$ e $\mathcal{P}$ o vetor formado pelas probabilidades $P(a|s)$, $\forall a \in \mathcal{A}$, $\forall s \in \mathcal{S}$.

Logo, a função de log-verossimilhança $l\big(z_1^n|\mathcal{P}\big)$ é dada por

$$l\big(z_1^n|\mathcal{P}\big) = \ln\big(L\big(z_1^n|\mathcal{P}\big)\big) = \ln\big(Prob\big(z_1^M\big)\big) + \sum_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} N_n(s,a)\ln(P(a|s))$$

Conforme Csiszár e Talata (2006) [9], a verossimilhança máxima de $L\big(z_1^n|\mathcal{P}\big)$, denotada por $ML(z_1^n, \mathcal{S})$, é aproximada por $\prod_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} \hat{P}(a|s)^{N_n(s,a)}$ na qual $\hat{P}(a|s)$, $\forall a \in \mathcal{A}, s \in \mathcal{S}$, são os estimadores de $P(a|s)$, $\forall a \in \mathcal{A}, s \in \mathcal{S}$, que maximizam a função $\prod_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} P(a|s)^{N_n(s,a)}$ sujeito à restrição $\sum_{a\in\mathcal{A}} P(a|s) = 1$ para cada $s \in \mathcal{S}$.

Teorema 3.1. *O valor de $P(a|s)$ que maximiza $\prod_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} P(a|s)^{N_n(s,a)}$ sujeito à restrição $\sum_{a\in\mathcal{A}} P(a|s) = 1$, para cada $s \in \mathcal{S}$, é $\hat{P}(a|s) = \frac{N_n(s,a)}{\sum_{a\in\mathcal{A}} N_n(s,a)}$.*

Demonstração. Veja Seção 5.2 do Capítulo 5. □

Logo, utilizando o Teorema 3.1, a verossimilhança máxima de $L\big(z_1^n|\mathcal{P}\big)$ é dada por

$$ML(z_1^n, \mathcal{S}) = \prod_{\substack{s\in\mathcal{S}\\ a\in\mathcal{A}}} \left(\frac{N_n(s,a)}{N_n(s)}\right)^{N_n(s,a)}$$

Agora, considere a_i a i-ésima coordenada de $a \in \mathcal{A}$ e, para para cada t, $Z_t^{(i)}$ é a i-ésima coordenada de Z_t. Dado $U = \big\{i_1, \ldots, i_m\big\} \subseteq \{1, \ldots, l\}$, $i_u \neq i_v$, define-se o evento $\big[Z_t^U = a^U\big]$ como sendo

$$\big[Z_t^U = a^U\big] = \Big[\Big(Z_t^{(i_1)}, \ldots, Z_t^{(i_m)}\Big) = \big(a_{i_1}, \ldots, a_{i_m}\big)\Big]$$

e a probabilidade $P(a^U|s)$ como sendo

$$P\big(a^U|s\big) = Prob\Big(Z_t^U = a^U | Z_{t-M}^{t-1} = s\Big), \forall s \in \mathcal{S} \tag{3.3.1}$$

com $P\big(Z_{t-M}^{t-1} = s\big) > 0$.

Além disso, seja

$$N_n\big(s, a^U\big) = \Big|\Big\{u : M < u \leq n, z_{u-M}^{u-1} = s, z_u^U = a_u^U\Big\}\Big|$$

Em palavras, $N_n\big(s, a^U\big)$ é o número de ocorrências da sequência sa tal que nas coordenadas $i_1, \ldots, i_m$ da sequência a aparecem $a_{i_1}, \ldots, a_{i_m}$ e nas demais posições pode aparecer qualquer elemento de $\mathcal{B}$.

A seguir definimos o conceito de independência condicional dada uma sequência $s \in \mathcal{S}$.

Definição 3.5. *Dado $s \in \mathcal{S}$, uma partição $\mathcal{I} = \big\{I_1, \ldots, I_{|\mathcal{I}|}\big\}$ de $\{1, \ldots, l\}$ é compatível com a lei condicional do processo dado s se*

$$Prob\big(Z_t = a | Z_{t-M}^{t-1} = s\big) = \prod_{j=1}^{|\mathcal{I}|} Prob\Big(Z_t^{I_j} = a^{I_j} | Z_{t-M}^{t-1} = s\Big), \forall a \in A, \forall t \in \mathbb{Z} \tag{3.3.2}$$

Usando (3.3.1), a equação (3.3.2) pode ser reescrita como $P\big(a|s\big) = \prod_{j=1}^{|\mathcal{I}|} P\big(a^{I_j}|s\big)$.

Definição 3.6. *O conjunto de partições $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ será chamado compatível com a lei do processo se, para cada $s \in \mathcal{S}$, $\mathcal{I}_s$ é compatível com a lei condicional do processo dado s.*

Considere $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ um conjunto de partições de $\{1, \ldots, l\}$ compatível com a lei do processo sendo $\mathcal{I}_s = \{I_1^s, \ldots, I_{|\mathcal{I}_s|}^s\}$. Logo,

$$L\big(z_1^n, \mathcal{J}|\mathcal{P}\big) = Prob\Big(z_1^M\Big) \prod_{s\in\mathcal{S}} \prod_{a\in\mathcal{A}} \prod_{j=1}^{|\mathcal{I}_s|} P\big(a^{I_j^s}|s\big)^{N_n(s,a)}$$

sendo $\mathcal{P}$ o vetor formado pelas probabilidades $P\big(a^{I_j^s}|s\big)$, para todo $j = 1, \ldots, \big|\mathcal{I}_s\big|$, $\forall a \in \mathcal{A}$, $\forall s \in \mathcal{S}$.

Teorema 3.2. *O valor de* $P(a^{I_j^s}|s)$ *que maximiza* $\prod_{s\in\mathcal{S}}\prod_{a\in\mathcal{A}}\prod_{j=1}^{|\mathcal{I}_s|}P\big(a^{I_j^s}|s\big)^{N_n(s,a)}$ *sujeito à* $\sum_{a^{I_j^s}\in\mathcal{B}^{|I_j^s|}}P^{I_j^s}(a|s)=1$, *para cada* $s\in\mathcal{S}$ *e* $j=1,\ldots,|\mathcal{I}_s|$, *é*

$$\hat{P}\big(a^{I_j^s}|s\big)=\frac{N_n\big(s,a^{I_j^s}\big)}{N_n(s)}$$

Demonstração. Veja Seção 5.2 do Capítulo 5. □

Utilizando o Teorema 3.2, a verossimilhança máxima de $L\big(z_1^n,\mathcal{J}|\mathcal{P}\big)$ é dada por

$$ML(z_1^n,\mathcal{S},\mathcal{J})=\prod_{s\in\mathcal{S}}\prod_{a\in\mathcal{A}}\prod_{j=1}^{|\mathcal{I}_s|}\left(\frac{N_n\left(s,a^{\mathcal{I}_j^s}\right)}{N_n(s)}\right)^{N_n(s,a)}$$

3.4 Partições das Coordenadas de uma Cadeia de Markov l-variada em Partes Condicionalmente Independentes

Considere $(Z_t)_{t\in\mathbb{Z}}$ um processo de Markov estacionário de ordem finita M com valores em $\mathcal{A}=\mathcal{B}^l$. Seja $\mathcal{S}=\mathcal{A}^M$ o seu espaço de estados.

Considere a_i a i-ésima coordenada de $a\in\mathcal{A}$ e, para para cada t, $Z_t^{(i)}$ é a i-ésima coordenada de Z_t. Dado $U=\big\{i_1,\ldots,i_m\big\}\subseteq\{1,\ldots,l\}$, $i_u\neq i_v$, define-se o evento $\big[Z_t^U=a^U\big]$ como sendo

$$\big[Z_t^U=a^U\big]=\Big[\Big(Z_t^{(i_1)},\ldots,Z_t^{(i_m)}\Big)=\big(a_{i_1},\ldots,a_{i_m}\big)\Big]$$

e a probabilidade $P\big(a^U|s\big)$ como sendo

$$P\big(a^U|s\big)=Prob\Big(Z_t^U=a^U|Z_{t-M}^{t-1}=s\Big),\forall s\in\mathcal{S}$$

com $P\big(Z_{t-M}^{t-1}=s\big)>0$.

Na Seção 3.3 foi definido que, dada uma sequência $s\in\mathcal{S}$, uma partição $\mathcal{I}=\big\{I_1,\ldots,I_{|\mathcal{I}|}\big\}$ de $\{1,\ldots,l\}$ é compatível com a lei condicional do processo dado s se

$$P\big(a|s\big)=\prod_{j=1}^{|\mathcal{I}|}P(a^{I_j}|s),\forall a\in\mathcal{A}\tag{3.4.1}$$

Considerando $s \in \mathcal{S}$, vale observar que $\mathcal{I}$ compatível com a lei condicional do processo dado s significa que $Z_t^{I_1}, \ldots, Z_t^{I_{|\mathcal{I}|}}$ são condicionalmente independendentes dado s.

Definição 3.7. *O conjunto de partições $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ será chamado compatível com a lei do processo se, para cada $s \in \mathcal{S}$, $\mathcal{I}_s$ é compatível com a lei condicional do processo dado s.*

Definição 3.8. *Dado $s \in \mathcal{S}$, seja $\mathcal{I}_s = \left\{I_1^s, \ldots, I_{|\mathcal{I}_s|}^s\right\}$ uma partição de $\{1, \ldots, l\}$ compatível com a lei condicional do processo dado s. Se o conjunto $\mathcal{I}_s$ é a partição mais fina de $\{1, \ldots, l\}$ então ele será chamado de partição de estrutura de dependência condicional de s.*

Definição 3.9. *Se $\mathcal{I}_s$ é a partição de estrutura de dependência condicional de s, $\forall s \in \mathcal{S}$, então o conjunto de partições $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ será chamado de estrutura de dependência condicional do processo.*

Supondo que $Z_t^{I_1^s}, \ldots, Z_t^{I_{|\mathcal{I}_s|}^s}$ são condicionalmente independendentes dado $s \in \mathcal{S}$, $\forall s \in \mathcal{S}$, a função de verossimilhança deste modelo é dada por

$$L\big(z_1^n, \mathcal{J}|\mathcal{P}\big) = Prob\Big(z_1^M\Big) \prod_{s\in\mathcal{S}} \prod_{a\in\mathcal{A}} \prod_{j=1}^{|\mathcal{I}_s|} P(a^{I_j^s}|s)^{N_n(s,a)} \tag{3.4.2}$$

sendo $\mathcal{P}$ o vetor formado pela probabilidade $P(a^{I_j^s}|s)$ para todo $j = 1, \ldots, \big|\mathcal{I}_s\big|$, $a \in \mathcal{A}$ e $s \in \mathcal{S}$, $\mathcal{J} = \big\{\mathcal{I}_s\big\}_{s\in\mathcal{S}}$ e

$$N_n(s,a) = \left|\left\{u : D(n) < u \leq n, z_{u-M}^{u-1} = s, z_u = a\right\}\right|$$

Além disso, a verossimilhança máxima de (3.4.2) é dada por

$$ML(z_1^n, \mathcal{S}, \mathcal{J}) = \prod_{s\in\mathcal{S}} \prod_{a\in\mathcal{A}} \prod_{j=1}^{|\mathcal{I}_s|} \left(\frac{N_n\left(s, a^{\mathcal{I}_j^s}\right)}{N_n(s)}\right)^{N_n(s,a)}$$

sendo

$$N_n(s) = \left|\left\{u : D(n) < u \leq n, z_{u-M}^{u-1} = s\right\}\right|$$

e

$$N_n\big(s, a^{\mathcal{I}_j^s}\big) = \left|\left\{u : M < u \leq n, z_{u-M}^{u-1} = s, z_u^{\mathcal{I}_j^s} = a_u^{\mathcal{I}_j^s}\right\}\right|$$

ou seja, $N_n\big(s, a^{\mathcal{I}_j^s}\big)$ é o número de ocorrências de sa tal que nas coordenadas $i_1, \ldots, i_m$ da sequência a aparecem $a_{i_1}, \ldots, a_{i_m}$ e nas demais posições pode aparecer qualquer elemento de $\mathcal{B}$.

Com a suposição de que $Z_t^{I_1^s}, \ldots, Z^{I_{|\mathcal{I}_s|}^s}$ são condicionalmente independendentes dado s, $\forall s \in \mathcal{S}$, o BIC correspondente a este modelo é

$$BIC(z_1^n, \mathcal{S}, \mathcal{J}) = -\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{j=1}^{|\mathcal{I}_s|} N_n(s,a)\ln\left(\frac{N_n\left(s, a^{I_j^s}\right)}{N_n(s)}\right) + \sum_{s\in\mathcal{S}}\sum_{j=1}^{|\mathcal{I}_s|}\frac{|\mathcal{B}|^{|I_j^s|}-1}{2}\ln(n) \quad (3.4.3)$$

O termo $\displaystyle\sum_{s\in\mathcal{S}}\sum_{j=1}^{|\mathcal{I}_s|}\frac{|\mathcal{B}|^{|I_j^s|}-1}{2}\ln(n)$ pode ser escrito como $\frac{\mathcal{C}-\mathcal{D}}{2}\ln(n)$ sendo $\mathcal{C} = \displaystyle\sum_{s\in\mathcal{S}}\sum_{j=1}^{|\mathcal{I}_s|}|\mathcal{B}|^{|I_j^s|}$ e $\mathcal{D} = \displaystyle\sum_{s\in\mathcal{S}}|\mathcal{I}_s|$.

No caso em que, para cada $s \in \mathcal{S}$, a partição de estrutura de dependência condicional de s for o conjunto $\mathcal{L} = \{1, \ldots, l\}$, a expressão (3.4.3) torna-se

$$BIC(z_1^n, \mathcal{S}, \mathcal{R}) = -\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} N_n(s,a)\ln\left(\frac{N_n(s,a)}{N_n(s)}\right) + \frac{(|\mathcal{B}|^l-1)|\mathcal{S}|}{2}\ln(n) \quad (3.4.4)$$

na qual $\mathcal{R} = \mathcal{L}^{\mathcal{S}}$.

O próximo teorema estabelece um critério baseado no BIC que fornece um meio para detectar se um processo não possui grupos de coordenadas condicionalmente independentes ou se possui grupos de coordenadas condicionalmente independentes.

Teorema 3.3. *Considere $(Z_t)_{t\in\mathbb{Z}}$ um processo de Markov estacionário de ordem finita M tomando valores no alfabeto finito $\mathcal{A} = \mathcal{B}^l$ e seja $\mathcal{S} = \mathcal{A}^M$ o seu espaço de estados. Além disso, considere $z_1^n = z_1 \ldots z_n$, $z_i \in \mathcal{A}$, $i = 1, \ldots, n$, uma amostra deste processo, $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ um conjunto de partições de $\mathcal{L} = \{1, \ldots, l\}$ e $\mathcal{R} = \mathcal{L}^{|\mathcal{S}|}$. Então, eventualmente quase certamente, quando $n \to \infty$,*

$$BIC(z_1^n, \mathcal{S}, \mathcal{R}) \geq BIC(z_1^n, \mathcal{S}, \mathcal{J})$$

se, e somente se, $\mathcal{I}_s$ é compatível com a lei condicional do processo dado s, $\forall s \in \mathcal{S}$.

Demonstração. Veja Seção 5.2 do Capítulo 5. □

O seguinte teorema permite decidir se é recomendável ou não ter índices numa mesma partição tais que as coordenadas do processo associadas a estes índices sejam independentes.

Teorema 3.4. *Sob as suposições do Teorema 3.3, seja* $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ *um conjunto de partições compatível com a lei do processo sendo* $\mathcal{I}_s = \left\{I_1^s, \ldots, I_{|\mathcal{I}_s|}^s\right\}$, $\forall s \in \mathcal{S}$. *Suponha que, em relação à sequência* s_m, *para cada* $j = 1, \ldots, |\mathcal{I}_{s_m}|$, *existem subconjuntos* $\tilde{I}_{j_k}^{s_m}$, $k = 1, \ldots, k_j$, *de* $I_j^{s_m}$ *tais que* $\tilde{I}_{j_k}^{s_m} \preceq I_j^{s_m}$ *e* $P\big(a|s\big) = \prod_{k=1}^{k_j} P^{\tilde{I}_{j_k}^{s_m}}(a|s)$, $\forall a \in \mathcal{A}$. *Então, eventualmente quase certamente, quando* $n \to \infty$,

$$BIC\big(z_1^n, \mathcal{S}, \mathcal{J}\big) \geq BIC\big(z_1^n, \mathcal{S}, \tilde{\mathcal{J}}\big)$$

se, e somente se,

$$P(a|s_m) = \prod_{j=1}^{|\mathcal{I}_{s_m}|} \prod_{k=1}^{k_j} P\left(a^{\tilde{I}_{j_k}^{s_m}}|s_m\right), \forall a \in \mathcal{A}$$

sendo $\tilde{\mathcal{J}} = \left\{\mathcal{I}_{s_1}, \ldots, \mathcal{I}_{s_{m-1}}, \tilde{\mathcal{I}}_{s_m}, \mathcal{I}_{s_{m+1}}, \ldots, \mathcal{I}_{s_{|\mathcal{S}|}}\right\}$ *e* $\tilde{\mathcal{I}}_{s_m} = \left\{\tilde{I}_{j_k}^{s_m}\right\}_{\substack{k=1,\ldots,k_j \\ j=1,\ldots,|\tilde{\mathcal{I}}_{s_m}|}}$.

Demonstração. Veja Seção 5.2 do Capítulo 5. □

O próximo teorema mostra que, para n suficientemente grande, pode-se obter a estrutura de dependência condicional do processo pela minimização do BIC.

Teorema 3.5. *Considere* $(Z_t)_{t\in\mathbb{Z}}$ *um processo de Markov estacionário de ordem finita tomando valores no alfabeto finito* $\mathcal{A} = \mathcal{B}^l$ *e seja* $\mathcal{S} = \mathcal{A}^M$ *o seu espaço de estados. Além disso, sejam* $z_1^n = z_1 \ldots z_n$ *uma amostra deste processo,* $\wp = \mathcal{C}^{|\mathcal{S}|}$ *no qual* $\mathcal{C}$ *é o conjunto formado por todas as partições de* $\{1, 2, \ldots, l\}$ *e* $\mathcal{L}^\star = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ *a estrutura de dependência condicional do processo. Seja*

$$\mathcal{L}_n^\star = \arg\min_{\mathcal{J}\in\wp} BIC\big(z_1^n, \mathcal{S}, \mathcal{J}\big)$$

Então, eventualmente quase certamente, quando $n \to \infty$, $\mathcal{L}^\star = \mathcal{L}_n^\star$.

Demonstração. Veja Seção 5.2 do Capítulo 5. □

Quando o número de coordenadas do processo cresce, o custo computacional do critério BIC, apresentado no teorema 3.5, torna-se excessivo. Neste caso propomos um algoritmo mais eficiente do ponto de vista computacional e a sua consistência é demonstrada.

O algoritmo é inicializado com $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ sendo $\mathcal{I}_{s_i} = \{\{1\},\ldots,\{l\}\}$, $\forall i = 1,\ldots,|\mathcal{S}|$. Em cada iteração, para cada estado $s_i \in \mathcal{S}$, o algoritmo gera partições de $\{1,\ldots,l\}$ a fim de obter-se a partição de estrutura de dependência condicional de s_i, ou seja, queremos obter a partição $\mathcal{I}_{s_i}$ que cause o maior descrécimo no valor do BIC.

A ideia é que, em cada iteração, dois elementos de $\mathcal{I}_{s_i}$ sejam unidos para que uma outra partição seja obtida. Por exemplo, unindo os elementos indexados por j e k em $\mathcal{I}_{s_i}$, obtém-se $I^{s_i}_{jk} \leftarrow I^{s_i}_j \cup I^{s_i}_k$. Logo, $\mathcal{I}^{jk}_{s_i} \leftarrow \left(\mathcal{I}^{jk}_{s_i} \setminus \{I^{s_i}_j, I^{s_i}_j\}\right) \cup I^{s_i}_{jk}$ é o conjunto $\mathcal{I}_{s_i}$ no qual foram retirados os elementos $I^{s_i}_j$ e $I^{s_i}_k$ e adicionado o elemento $I^{s_i}_{jk}$.

Considerando $\mathcal{J} \leftarrow \left\{\mathcal{I}_{s_1},\ldots,\mathcal{I}_{s_i},\ldots,\mathcal{I}_{s_{|\mathcal{S}|}}\right\}$ e $\mathcal{J}^{jk} \leftarrow \left\{\mathcal{I}_{s_1},\ldots,\mathcal{I}^{jk}_{s_i},\ldots,\mathcal{I}_{s_{|\mathcal{S}|}}\right\}$ temos que se $BIC\left(z_1^n,\mathcal{S},\mathcal{J}\right) > BIC\left(z_1^n,\mathcal{S},\mathcal{J}^{jk}\right)$ então $\mathcal{J}^{jk}$ é um candidato a estrutura de dependência condicional do processo e, neste caso, $\mathcal{J} \leftarrow \mathcal{J}^{jk}$.

O pseudo-código do algoritmo é mostrado a seguir.

INPUT: l, $\mathcal{S} = \mathcal{A}^M$, $z_1^n = z_1 \ldots z_n$
OUTPUT: $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ tal que $\mathcal{J}$ é a estrutura de dependência condicional do processo
$\mathcal{I}_{s_i} \leftarrow \{\{1\},\{2\},\ldots,\{l\}\}$, $i = 1,\ldots,|\mathcal{S}|$
$\mathcal{J} \leftarrow \left\{\mathcal{I}_{s_1},\ldots,\mathcal{I}_{s_i},\ldots,\mathcal{I}_{s_{|\mathcal{S}|}}\right\}$
for $i = 1$ to $|\mathcal{S}|$ **do**
  $\mathcal{J}^o \leftarrow \mathcal{J}$
  $stop \leftarrow 0$
  while $stop = 0$ **do**
    $j \leftarrow 1$
    while $j \leq l-1$ **do**
      for $k = j+1$ to l **do**
        $I^{s_i}_{jk} \leftarrow I^{s_i}_j \cup I^{s_i}_k$
        for $u = j$ to $l-2$ **do**
          if $k = u+1$ **then**
            $I^{s_i}_j \leftarrow I^{s_i}_{u+2}$
          else
            $I^{s_i}_j \leftarrow I^{s_i}_{u+1}$

end if
end for
$I_{l-1}^{s_i} \leftarrow I_{jk}^{s_i}$
$\mathcal{I}_{s_i}^{jk} \leftarrow \{I_1^{s_i}, \ldots, I_{l-1}^{s_i}\}$
$\mathcal{J}^{jk} \leftarrow \left\{\mathcal{I}_{s_1}, \ldots, \mathcal{I}_{s_i}^{jk}, \ldots, \mathcal{I}_{s_{|\mathcal{S}|}}\right\}$
if $BIC\left(z_1^n, \mathcal{S}, \mathcal{J}\right) > BIC\left(z_1^n, \mathcal{S}, \mathcal{J}^{jk}\right)$ **then**
$\mathcal{J} \leftarrow \mathcal{J}^{jk}$
end if
end for
$j \leftarrow j + 1$
end while
if $\mathcal{J}^o \neq \mathcal{J}$ **then**
$\mathcal{J}^o \leftarrow \mathcal{J}$
$l \leftarrow l - 1$
else
$stop \leftarrow 1$
end if
end while
end for

Corolário 3.1. *Sob as suposições do Teorema 3.5, quando $n \leftarrow \infty$, $\mathcal{L}_n^\star$, dado pelo algoritmo anterior, converge eventualmente quase certamente para $\mathcal{L}^\star = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ que é a estrutura de dependência condicional do processo.*

Demonstração. Como $l < \infty$, para n suficientemente grande, pelo Teorema 3.4, o algoritmo retorna $\mathcal{L}^\star = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ que é a estrutura de dependência condicional do processo. □

3.4.1 Partições das Coordenadas de uma Cadeia de Markov de Memória Variável l-variada em Partes Condicionalmente Independentes

Os Teoremas 3.3, 3.4 e 3.5 podem ser estendidos, de forma natural, para um processo Markoviano de memória variável trocando o espaço de estados pelo conjunto formado pelos contextos do processo.

Desta forma, considere $(Z_t)_{t\in\mathbb{Z}}$ um processo de Markov de memória variável estacionário de ordem finita tomando com valores no alfabeto finito $\mathcal{A} = \mathcal{B}^l$ e seja $\mathcal{T}$ a árvore de contextos associada ao processo.

Teorema 3.6. *Considere $(Z_t)_{t\in\mathbb{Z}}$ um processo de Markov de memória variável estacionário de ordem finita M tomando valores no alfabeto finito $\mathcal{A} = \mathcal{B}^l$ e seja $\mathcal{T}$ a árvore de contextos associada ao processo. Além disso, considere $z_1^n = z_1 \dots z_n$, $z_i \in \mathcal{A}$, $i = 1, \dots, n$, uma amostra deste processo, $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{T}}$ um conjunto de partições de $\mathcal{L} = \{1, \dots, l\}$ e $\mathcal{R} = \mathcal{L}^{|\mathcal{T}|}$. Então, eventualmente quase certamente, quando $n \to \infty$,*

$$BIC(z_1^n, \mathcal{T}, \mathcal{R}) \geq BIC(z_1^n, \mathcal{T}, \mathcal{J})$$

se, e somente se, $\mathcal{I}_s$ é compatível com a lei condicional do processo dado s, $\forall s \in \mathcal{T}$.

Teorema 3.7. *Sob as suposições do Teorema 3.6, seja $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{T}}$ um conjunto de partições compatível com a lei do processo sendo $\mathcal{I}_s = \left\{I_1^s, \dots, I_{|\mathcal{I}_s|}^s\right\}$, $\forall s \in \mathcal{T}$. Suponha que, em relação à sequência s_m, para cada $j = 1, \dots, |\mathcal{I}_{s_m}|$, existem subconjuntos $\tilde{I}_{j_k}^{s_m}$, $k = 1, \dots, k_j$, de $I_j^{s_m}$ tais que $\tilde{I}_{j_k}^{s_m} \preceq I_j^{s_m}$ e $P\big(a|s\big) = \prod_{k=1}^{k_j} P^{\tilde{I}_{j_k}^{s_m}}(a|s)$, $\forall a \in \mathcal{A}$. Então, eventualmente quase certamente, quando $n \to \infty$,*

$$BIC\big(z_1^n, \mathcal{T}, \mathcal{J}\big) \geq BIC\big(z_1^n, \mathcal{T}, \tilde{\mathcal{J}}\big)$$

se, e somente se,

$$P(a|s_m) = \prod_{j=1}^{|\mathcal{I}_{s_m}|} \prod_{k=1}^{k_j} P^{\tilde{I}_{j_k}^{s_m}}(a|s_m), \forall a \in \mathcal{A}$$

sendo $\tilde{\mathcal{J}} = \left\{\mathcal{I}_{s_1}, \dots, \mathcal{I}_{s_{m-1}}, \tilde{\mathcal{I}}_{s_m}, \mathcal{I}_{s_{m+1}}, \dots, \mathcal{I}_{s_{|\mathcal{T}|}}\right\}$ e $\tilde{\mathcal{I}}_{s_m} = \left\{\tilde{I}_{j_k}^{s_m}\right\}_{\substack{k=1,\dots,k_j \\ j=1,\dots,|\tilde{\mathcal{I}}_{s_m}|}}$.

Teorema 3.8. *Considere $(Z_t)_{t\in\mathbb{Z}}$ um processo de Markov de memória variável, estacionário, de ordem finita tomando valores no alfabeto finito $\mathcal{A} = \mathcal{B}^l$ e seja $\mathcal{T}$ a árvore de contextos associada ao processo. Além disso, sejam $z_1^n = z_1 \dots z_n$ uma amostra deste processo, $\wp = \mathcal{C}^{|\mathcal{T}|}$ no qual $\mathcal{C}$ é o conjunto formado por todas as partições de $\{1, 2, \dots, l\}$ e $\mathcal{L}^\star = \{\mathcal{I}_s\}_{s\in\mathcal{T}}$ a estrutura de dependência condicional do processo. Seja*

$$\mathcal{L}_n^\star = \arg\min_{\mathcal{J}\in\wp} BIC\big(z_1^n, \mathcal{T}, \mathcal{J}\big)$$

Então, eventualmente quase certamente, quando $n \to \infty$, $\mathcal{L}^\star = \mathcal{L}_n^\star$.

As demonstrações dos Teoremas 3.6, 3.7 e 3.8 são análogas às demonstrações dos Teoremas 3.3, 3.4 e 3.5 trocando-se o espaço de estados pelo conjunto formado pelos contextos do processo.

3.5 Simulações

3.5.1 Simulação 1

Considere o processo estacionário de Markov de ordem 1, $(Z_t)_{t\in\mathbb{Z}} = \left(Z_t^{(1)}, Z_t^{(2)}\right)_{t\in\mathbb{Z}}$, possuindo alfabeto $\mathcal{A} = \{0,1\} \times \{0,1\}$. As probabilidades de transição $Prob(Z_0 = a|Z_{-1} = s)$, $\forall a \in \mathcal{A}$, $\forall s \in \mathcal{S} = \mathcal{A}$, são dadas pela matriz (3.5.1) sendo que as suas linhas e colunas representam, respectivamente, s e a.

$$T_Z = \begin{pmatrix} 0,25 & 0,25 & 0,25 & 0,25 \\ 0,2 & 0,3 & 0,2 & 0,3 \\ 0,2 & 0,2 & 0,3 & 0,3 \\ 0,16 & 0,24 & 0,24 & 0,36 \end{pmatrix} \tag{3.5.1}$$

Pelas probabilidades de transição mostradas em (3.5.1), os processos $\left(Z_t^{(1)}\right)_{t\in\mathbb{Z}}$ e $\left(Z_t^{(2)}\right)_{t\in\mathbb{Z}}$, possuindo alfabeto $\{0,1\}$, são condicionalmente independentes dadas as sequências (0,0), (1,0), (0,1) e (1,1) pois, dado qualquer $s \in \mathcal{A}$,

$$Prob\big(Z_0 = a|Z_{-1} = s\big) = Prob\Big(Z_0^{(1)} = a^{(1)}|Z_{-1} = s\Big)Prob\Big(Z_0^{(2)} = a^{(2)}|Z_{-1} = s\Big), \forall a \in \mathcal{A}$$

com probabilidades $Prob\Big(Z_0^{(1)} = a^{(1)}|Z_{-1} = s\Big)$ e $Prob\Big(Z_0^{(2)} = a^{(2)}|Z_{-1} = s\Big)$ dadas a seguir.

- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)}|Z_{-1} = (0,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)}|Z_{-1} = (0,0)\Big) = 0,5$;
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)}|Z_{-1} = (0,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)}|Z_{-1} = (0,0)\Big) = 0,5$;
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)}|Z_{-1} = (0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)}|Z_{-1} = (0,0)\Big) = 0,5$;
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)}|Z_{-1} = (0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)}|Z_{-1} = (0,0)\Big) = 0,5$;
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)}|Z_{-1} = (1,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)}|Z_{-1} = (1,0)\Big) = 0,4$;

- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-1} = (1,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-1} = (1,0)\Big) = 0,6;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-1} = (1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-1} = (1,0)\Big) = 0,5;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-1} = (1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-1} = (1,0)\Big) = 0,5;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-1} = (0,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-1} = (0,1)\Big) = 0,5;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-1} = (0,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-1} = (0,1)\Big) = 0,5;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-1} = (0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-1} = (0,1)\Big) = 0,4;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-1} = (0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-1} = (0,1)\Big) = 0,6;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-1} = (1,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-1} = (1,1)\Big) = 0,4;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-1} = (1,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-1} = (1,1)\Big) = 0,6;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-1} = (1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-1} = (1,1)\Big) = 0,4;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-1} = (1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-1} = (1,1)\Big) = 0,6.$

Simulou-se 500 amostras independentes, de tamanhos 3.000, 10.000, 15.000 e 50.000, do processo $(Z_t)_{t\in\mathbb{Z}}$ a fim de obter a proporção de vezes em que o procedimento proposto detecta a estrutura de dependência condicional imposta. Na Tabela 5.1 são mostradas as porcentagens de acertos.

Tabela 5.1: Porcentagem de Acertos por Tamanho Amostral

Tamanho das amostras Simuladas	Porcentagem de Acertos
3.000	93,6%
10.000	95,8%
15.000	97,2%
50.000	99%

Observa-se que o procedimento proposto detectou a estrutura de dependência condicional do processo em mais de 90% das amostras geradas sendo que a sua eficiência foi melhorando conforme aumentou-se o tamanho das amostras simuladas.

3.5.2 Simulação 2

Considere o processo estacionário de Markov de ordem 2, $(Z_t)_{t\in\mathbb{Z}} = \left(Z_t^{(1)}, Z_t^{(2)}\right)_{t\in\mathbb{Z}}$, possuindo alfabeto $\mathcal{A} = \{0,1\} \times \{0,1\}$. As probabilidades de transição $Prob(Z_0 = a|Z_{-2}^{-1} = s)$, $\forall a \in \mathcal{A}$, $\forall s \in \mathcal{S} = \{(0,0)(0,0), (1,0)(0,0), (0,1)(0,0), (1,1)(0,0), (0,0)(1,0), (1,0)(1,0), (0,1)(1,0), (1,1)(1,0), (0,0)(0,1), (1,0)(0,1), (0,1)(0,1), (1,1)(0,1), (0,0)(1,1), (1,0)(1,1), (0,1)(1,1), (1,1)(1,1)\}$, são mostradas pela matriz (3.5.2) sendo que as suas linhas e colunas representam, respectivamente, s e a.

$$T_Z = \begin{pmatrix} 0,36 & 0,24 & 0,24 & 0,16 \\ 0,18 & 0,42 & 0,12 & 0,28 \\ 0,18 & 0,42 & 0,12 & 0,28 \\ 0,42 & 0,18 & 0,28 & 0,12 \\ 0,18 & 0,12 & 0,42 & 0,28 \\ 0,09 & 0,21 & 0,21 & 0,49 \\ 0,09 & 0,21 & 0,21 & 0,49 \\ 0,21 & 0,09 & 0,49 & 0,21 \\ 0,24 & 0,16 & 0,36 & 0,24 \\ 0,12 & 0,28 & 0,18 & 0,42 \\ 0,12 & 0,28 & 0,18 & 0,42 \\ 0,28 & 0,12 & 0,42 & 0,18 \\ 0,48 & 0,32 & 0,12 & 0,08 \\ 0,24 & 0,56 & 0,06 & 0,14 \\ 0,24 & 0,56 & 0,06 & 0,14 \\ 0,56 & 0,24 & 0,14 & 0,06 \end{pmatrix} \tag{3.5.2}$$

Pelas probabilidades de transição mostradas em (3.5.2) os processos $\left(Z_t^{(1)}\right)_{t\in\mathbb{Z}}$ e $\left(Z_t^{(2)}\right)_{t\in\mathbb{Z}}$, possuindo alfabeto $\{0,1\}$, são condicionalmente independentes dado qualquer $s \in \mathcal{S}$ pois, para cada $s \in \mathcal{S}$, $Prob\big(Z_0 = a|Z_{-2}^{-1} = s\big) = Prob\Big(Z_0^{(1)} = a^{(1)}|Z_{-2}^{-1} = s\Big)Prob\Big(Z_0^{(2)} = a^{(2)}|Z_{-2}^{-1} = s\Big)$, $\forall a \in \mathcal{A}$,

com probabilidades $Prob\Big(Z_0^{(1)} = a^{(1)} | Z_{-2}^{-1} = s\Big)$ e $Prob\Big(Z_0^{(2)} = a^{(2)} | Z_{-2}^{-1} = s\Big)$ dadas a seguir.

- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (0,0)(0,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (0,0)(0,0)\Big) = 0,6;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (0,0)(0,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (0,0)(0,0)\Big) = 0,4;$
  $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (0,0)(0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (0,0)(0,0)\Big) = 0,6;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (0,0)(0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (0,0)(0,0)\Big) = 0,4;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (1,0)(0,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (1,0)(0,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (1,0)(0,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (1,0)(0,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (1,0)(0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (1,0)(0,0)\Big) = 0,6;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (1,0)(0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (1,0)(0,0)\Big) = 0,4;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (0,1)(0,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (0,1)(0,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (0,1)(0,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (0,1)(0,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (0,1)(0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (0,1)(0,0)\Big) = 0,6;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (0,1)(0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (0,1)(0,0)\Big) = 0,4;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (1,1)(0,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (1,1)(0,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (1,1)(0,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (1,1)(0,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (1,1)(0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (1,1)(0,0)\Big) = 0,6;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (1,1)(0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (1,1)(0,0)\Big) = 0,4;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (0,0)(1,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (0,0)(1,0)\Big) = 0,6;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (0,0)(1,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (0,0)(1,0)\Big) = 0,4;$

- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (0,0)(1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (0,0)(1,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (0,0)(1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (0,0)(1,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (1,0)(1,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (1,0)(1,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (1,0)(1,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (1,0)(1,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (1,0)(1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (1,0)(1,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (1,0)(1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (1,0)(1,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (0,1)(1,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (0,1)(1,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (0,1)(1,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (0,1)(1,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (0,1)(1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (0,1)(1,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (0,1)(1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (0,1)(1,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (1,1)(1,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (1,1)(1,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (1,1)(1,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (1,1)(1,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (1,1)(1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (1,1)(1,0)\Big) = 0,3;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (1,1)(1,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (1,1)(1,0)\Big) = 0,7;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (0,0)(0,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (0,0)(0,1)\Big) = 0,6;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (0,0)(0,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (0,0)(0,1)\Big) = 0,4;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (0,0)(0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (0,0)(0,1)\Big) = 0,4;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (0,0)(0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (0,0)(0,1)\Big) = 0,6;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (1,0)(0,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (1,0)(0,1)\Big) = 0,3;$

- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (1,0)(0,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (1,0)(0,1)\Big) = 0,7;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (1,0)(0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (1,0)(0,1)\Big) = 0,4;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (1,0)(0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (1,0)(0,1)\Big) = 0,6;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (0,1)(0,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (0,1)(0,1)\Big) = 0,3;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (0,1)(0,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (0,1)(0,1)\Big) = 0,7;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (0,1)(0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (0,1)(0,1)\Big) = 0,4;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (0,1)(0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (0,1)(0,1)\Big) = 0,6;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (1,1)(0,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (1,1)(0,1)\Big) = 0,7;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (1,1)(0,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (1,1)(0,1)\Big) = 0,3;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (1,1)(0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (1,1)(0,1)\Big) = 0,4;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (1,1)(0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (1,1)(0,1)\Big) = 0,6;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (0,0)(1,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (0,0)(1,1)\Big) = 0,6;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (0,0)(1,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (0,0)(1,1)\Big) = 0,4;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (0,0)(1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (0,0)(1,1)\Big) = 0,8;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (0,0)(1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (0,0)(1,1)\Big) = 0,2;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (1,0)(1,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (1,0)(1,1)\Big) = 0,3;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (1,0)(1,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (1,0)(1,1)\Big) = 0,7;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (1,0)(1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (1,0)(1,1)\Big) = 0,8;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (1,0)(1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (1,0)(1,1)\Big) = 0,2;$

- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (0,1)(1,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (0,1)(1,1)\Big) = 0,3;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (0,1)(1,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (0,1)(1,1)\Big) = 0,7;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (0,1)(1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (0,1)(1,1)\Big) = 0,8;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (0,1)(1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (0,1)(1,1)\Big) = 0,2;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)} | Z_{-2}^{-1} = (1,1)(1,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)} | Z_{-2}^{-1} = (1,1)(1,1)\Big) = 0,7;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)} | Z_{-2}^{-1} = (1,1)(1,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)} | Z_{-2}^{-1} = (1,1)(1,1)\Big) = 0,3;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)} | Z_{-2}^{-1} = (1,1)(1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)} | Z_{-2}^{-1} = (1,1)(1,1)\Big) = 0,8;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)} | Z_{-2}^{-1} = (1,1)(1,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)} | Z_{-2}^{-1} = (1,1)(1,1)\Big) = 0,2.$

Simulou-se 500 amostras independentes, de tamanhos 3.000, 10.000, 15.000 e 50.000, do processo $(Z_t)_{t\in\mathbb{Z}}$ a fim de obter a proporção de vezes em que o procedimento proposto detecta a estrutura de dependência condicional imposta. Na Tabela 5.2 são mostradas as porcentagens de acertos.

Tabela 5.2: Porcentagem de Acertos por Tamanho Amostral

Tamanho das amostras Simuladas	Porcentagem de Acertos
3.000	67,6 %
10.000	80,0%
15.000	81,8%
50.000	88,2%

Observa-se que o procedimento proposto detectou a estrutura de independência condicional do processo em mais de 80% das amostras geradas, exceto para amostras de tamanho 3.000, sendo que a sua eficiência foi melhorando conforme aumentou-se o tamanho das amostras simuladas.

3.5.3 Simulação 3

Considere o processo estacionário de Markov de ordem 1, $(Z_t)_{t\in\mathbb{Z}} = \left(Z_t^{(1)}, Z_t^{(2)}\right)_{t\in\mathbb{Z}}$, possuindo alfabeto $\mathcal{A} = \{0,1\} \times \{0,1\}$ e as probabilidades de transição $Prob(a|s)$, $\forall a \in \mathcal{A}$, $\forall s \in \mathcal{A}$ são dadas pela matriz (3.5.3) sendo que as suas linhas e colunas representam, respectivamente, s e a.

$$T_Z = \begin{pmatrix} 0,25 & 0,25 & 0,25 & 0,25 \\ 0,5 & 0 & 0 & 0,5 \\ 0,25 & 0,25 & 0,25 & 0,25 \\ 0,5 & 0 & 0 & 0,5 \end{pmatrix} \tag{3.5.3}$$

De acordo com as probabilidades apresentadas em (3.5.3), os processos $(Z_t^{(1)})_{t\in\mathbb{Z}}$ e $(Z_t^{(2)})_{t\in\mathbb{Z}}$, possuindo alfabeto $\{0,1\}$, são condicionalmente independentes dadas as sequências (0,0) e (0,1) pois, $Prob\big(Z_0 = a|Z_{-1}^{t-1} = s\big) = Prob\Big(Z_0^{(1)} = a^{(1)}|Z_{-1} = s\Big)Prob\Big(Z_0^{(2)} = a^{(2)}|Z_{-1} = s\Big)$ é satisfeita para todo $a \in \mathcal{A}$, somente para $s = (0,0), (0,1)$, com probabilidades:

- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)}|Z_{-1} = (0,0)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)}|Z_{-1} = (0,0)\Big) = 0,5;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)}|Z_{-1} = (0,0)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)}|Z_{-1} = (0,0)\Big) = 0,5;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)}|Z_{-1} = (0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)}|Z_{-1} = (0,0)\Big) = 0,5;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)}|Z_{-1} = (0,0)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)}|Z_{-1} = (0,0)\Big) = 0,5;$
- $Prob\Big(Z_0^{(1)} = (0,0)^{(1)}|Z_{-1} = (0,1)\Big) = Prob\Big(Z_0^{(1)} = (0,1)^{(1)}|Z_{-1} = (0,1)\Big) = 0,5;$
- $Prob\Big(Z_0^{(1)} = (1,0)^{(1)}|Z_{-1} = (0,1)\Big) = Prob\Big(Z_0^{(1)} = (1,1)^{(1)}|Z_{-1} = (0,1)\Big) = 0,5;$
- $Prob\Big(Z_0^{(2)} = (0,0)^{(2)}|Z_{-1} = (0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,0)^{(2)}|Z_{-1} = (0,1)\Big) = 0,5;$
- $Prob\Big(Z_0^{(2)} = (0,1)^{(2)}|Z_{-1} = (0,1)\Big) = Prob\Big(Z_0^{(2)} = (1,1)^{(2)}|Z_{-1} = (0,1)\Big) = 0,5.$

Simulou-se 500 amostras independentes, de tamanhos 3.000, 10.000, 15.000 e 50.000, a fim de obter a proporção de vezes em que o procedimento proposto detecta a estrutura de independência condicional imposta ao processo $(Z_t)_{t\in\mathbb{Z}}$, para cada tamanho da amostra considerado. Na Tabela 5.3 são mostradas as porcentagens de acertos.

Tabela 5.3: Porcentagem de Acertos por Tamanho Amostral

Tamanho das amostras Simuladas	Porcentagem de Acertos
3.000	96,6%
10.000	98,6%
15.000	97,8%
50.000	99,2%

Observa-se que o procedimento proposto detectou a estrutura de independência condicional imposta ao processo em mais de 90% das amostras geradas sendo que a sua eficiência foi melhorando conforme aumentou-se o tamanho das amostras simuladas.

3.5.4 Simulação 4

Nesta subseção geramos 500 amostras independentes, de tamanhos 15, 50, 100, 500 e 5.000, de uma distribuição multinomial bivariada, tomando valores em $\{0,1\} \times \{0,1\}$, os quais são todos igualmente prováveis. Note que as variáveis marginais são independentes.

Desta forma, cada coordenada da distribuição multinomial bivariada é uma variável aleatória cujos possíveis valores são 0 ou 1. Utilizando o procedimento proposto neste capítulo e o Teste Qui-Quadrado de Independência, verificamos se as coordenadas da distribuição multinomial bivariada são independentes.

A Tabela 5.4 mostra, para cada tamanho amostral considerado, a proporção de vezes em que o procedimento proposto e o Teste Qui-Quadrado de Independência detectam a independência entre as coordenadas das distribuição em questão.

Observa-se que, para amostras de tamanho 15, o desempenho do procedimento proposto é inferior ao do Teste Qui-Quadrado de Independência enquanto que, para os demais tamanhos amostrais, o procedimento proposto e o Teste Qui-Quadrado de Independência apresentam resultados semelhantes.

Tabela 5.4: Porcentagem de Acertos por Tamanho Amostral

Tamanho das Amostras Simuladas	Método Proposto	Teste Qui-Quadrado de Independência ($\alpha = 5\%$)
15	87,8%	99,2%
50	95,8%	97,2%
100	96,2%	96,4%
500	99,2%	96,4%
3.000	99,2%	95,2%

3.5.5 Simulação 5

Nesta subseção geramos 500 amostras independentes, de tamanhos 15, 50, 100, 500 e 5.000, de uma distribuição multinomial bivariada cujos possíveis valores (0,0), (1,0), (0,1), (1,1) possuem probabilidade $\frac{1}{16}$, $\frac{1}{4}$, $\frac{1}{2}$ e $\frac{3}{16}$, respectivamente. Observe que as variáveis marginais não são independentes.

Desta forma, cada coordenada da distribuição multinomial bivariada é uma variável aleatória cujos possíveis valores são 0 ou 1. Através do procedimento proposto e do Teste Qui-Quadrado de Independência, verificamos se as coordenadas da distribuição multinomial bivariada são independentes.

A Tabela 5.5 mostra, para cada tamanho amostral considerado, a proporção de vezes em que o procedimento proposto e o Teste Qui-Quadrado de Independência detectam a independência entre as coordenadas das distribuição em questão.

Observe que, para amostras de tamanho 15, o desempenho do procedimento proposto se destaca em relação ao Teste Qui-Quadrado de Independência enquanto que, para os demais tamanhos amostrais, o procedimento proposto e o Teste Qui-Quadrado de Independência apresentam resultados semelhantes.

Tabela 5.5: Porcentagem de Acertos por Tamanho Amostral

Tamanho das Amostras Simuladas	Método Proposto	Teste Qui-Quadrado de Independência ($\alpha = 5\%$)
15	68,2%	23,0%
50	94,2%	90,8%
100	100,0%	100,0%
500	100,0%	100,0%
3.000	100,0%	100,0%

3.6 Aplicação

A pesquisa em assinaturas rítmicas em textos escritos é importante de diferentes pontos de vista científicos. Por exemplo, é uma importante ferramenta para o desenvolvimento de sintetizadores de fala; para a descrição da evolução histórica do ritmo de linguagens naturais; para o estudo de correlação acústica de classes rítmicas entre outros.

Galves et al. [13] aplicou Cadeia de Markov com Memória variável para modelar as cadeias de símbolos obtidos através codificação de textos escritos. Neste artigo, o Português Europeu e o Português Brasileiro foram analisados. Do ponto de vista de linguagem externa, eles produzem um grande número de sentenças superficialmente idênticas mas há argumentos de que eles pertencem a diferentes classes rítmicas. Na literatura linguística há a conjectura de que o Português Europeu e o Português Brasileiro pertencem a diferentes classes rítmicas.

O conjunto de dados linguístico consistiu de textos escritos, escolhidos aleatoriamente, extraídos de um corpus de jornais impressos em português brasileiro e português europeu. Os textos foram codificados utilizando-se um alfabeto finito de símbolos, expressando um número pequeno de características rítmicas básicas as quais podem ser automaticamente resgatadas dos textos escritos.

Foram extraídos, de 1994 a 1995, 40 artigos de cada um dos jornais: "Folha de São Paulo" (jornal brasileiro) e "O Público" (jornal português).

Primeiramente, selecionou-se, aleatoriamente, 20 edições de cada jornal em cada ano. Dentro de cada edição foram descartados todos os textos com menos que 1000 palavras assim como artigos considerados inadequados para o propósito da pesquisa (entrevistas, sinopses, transcrições de leis e

trabalhos coletados). Dos artigos remanescentes, selecionou-se, aleatoriamente, um de cada edição previamente escolhida. Depois, cada um dos textos selecionados foram codificados e submetidos para um procedimento de limpeza linguisticamnete orientada. Palavras compostas por hífen foram reescritas como duas palavras separadas, exceto quando um dos coordenadas é não acentuado. Reticências, pontos de interrogação e exclamação foram substituídos por ponto. Todos os parênteses foram removidos.

As amostras consistiram de 2070 e 2273 sentenças codificadas obtidas, respectivamente, dos jornais "Folha de São Paulo" e "O Público" e a codificação das sentenças foi feita atribuindo-se, para cada sílaba do texto, um dentre quatro símbolos possíveis, observando as seguintes características:

(i) A sílaba é acentuada ou não;

(ii) A sílaba é o início de uma palavra prosódica ou não.

Os símbolos utilizados na codificação foram 0, 1, 2 e 3 os quais representam:

- 0: sílaba inicial de uma palavra não prosódica e não acentuada;
- 1: sílaba inicial de uma palavra não prosódica e acentuada;
- 2: sílaba inicial de uma palavra prosódica e não acentuada;
- 3: sílaba inicial de uma palavra prosódica e acentuada.

Além desses símbolos adicionou-se o símbolo 4 para indicar o final de cada sentença.

Nesta aplicação consideramos o processo $(Z_t)_{t\in\mathbb{Z}} = \left(Z_t^{(1)}, Z_t^{(2)}, Z_t^{(3)}\right)_{t\in\mathbb{Z}}$ no qual a primeira, a segunda e a terceira coordenada indicam, respectivamente, se a sílaba é acentuada (codificada por 1) ou não (codificada por 0), se a sílaba é o início de uma palavra prosódica (codificada por 1) ou não (codificada por 0) e se é final de sentença (situação codificada por 1) ou não (situação codificada por 0).

Para cada idioma (Português brasileiro e Português europeu), o procedimento desenvolvido neste capítulo é utilizado a fim de obter-se quais coordenadas do processo $(Z_t)_{t\in\mathbb{Z}}$ são condicionalmente independentes dados os contextos deste.

Os contextos obtidos por Galves et al. [13] são:

- Português Brasileiro: 000, 100, 200, 300, 0010, 2010, 210, 20, 30, 001, 201, 21, 2, 3 e 4.
- Português Europeu: 000, 100, 200, 300, 010, 210, 20, 30, 001, 201, 21, 02, 012, 212, 32, 42, 3 e 4.

Aplicando-se o procedimento proposto ao Português brasileiro, obteve-se que, para todos os contextos, $\left(Z_t^{(1)}\right)_{t\in\mathbb{Z}}$ e $\left(Z_t^{(2)}\right)_{t\in\mathbb{Z}}$ não são condicionalmente independentes e estes são condicionalmente independentes a $\left(Z_t^{(3)}\right)_{t\in\mathbb{Z}}$. Em outras palavras, os processos que representam o acento de uma palavra e o início de uma palavra prosódica não são condicionalmente independentes e estes são condicionalmente independentes ao processo que indica o final de frase.

Além disso, aplicando-se o procedimento proposto ao Português europeu, obteve-se as mesmas conclusões referentes ao Português brasileiro, isto é, os processos que representam o acento de uma palavra e o início de uma palavra prosódica não são condicionalmente independentes e estes são condicionalmente independentes ao processo que indica o final de frase.

De fato, esperava-se que os processos que representam o acento de uma palavra e o início de uma palavra prosódica fossem condicionalmente independentes ao processo que indica o final de frase.

Capítulo 4

Conclusão

Nesta tese foram abordados dois problemas relacionados à seleção de modelos Markovianos sendo que ambos foram motivados por aplicações em Linguística. Um deles envolve um modelo de mistura entre dois processos Markovianos e o outro envolve um processo Markoviano multivariado.

O primeiro problema consistiu na construção de um procedimento que permita a escolha de uma árvore de contextos quando há árvores provenientes de dois processos distintos. O procedimento proposto (Método CTM α-truncado) foi baseado na taxa de entropia relativa simetrizada entre duas árvores de contextos como uma medida de similaridade e a expressão para o seu cálculo foi desenvolvida nesta tese.

Para o estudo do procedimento proposto, o conceito de ponto de ruptura assintótico foi definido para o nosso problema de seleção de modelos e ele foi obtido.

Através de simulações, comparou-se a eficiência do método proposto por Csiszár & Talata (Método CTM que corresponde ao Método CTM truncado com $\alpha = 1$) [9] com o método proposto (Método CTM α-truncado) com $\alpha = (1 - \frac{1}{m})$ e $\alpha = \frac{1}{2}$ sendo m o número de amostras. Eles foram comparados através da porcentagem de acertos, ou seja, a porcentagem de vezes em que o método selecionou o modelo correto.

Nas subseções 2.4.1 e 2.4.2 foram geradas m amostras independentes de tamanho n das quais k e $m - k$ correspondem, respectivamente, aos processos de Markov de memória variável representados pelas árvores de contextos $\mathcal{T}_Q$ e $\mathcal{T}_P$. Em outras palavras, cada amostra é gerada ou do processo de lei P ou do processo de lei Q (amostras contaminadas). Para este caso, em geral, o Método

CTM truncado com $\alpha = \frac{1}{2}$ apresentou as maiores porcentagens de acertos em relação aos demais e a performance do Método CTM foi superior à do CTM truncado com $\alpha = (1 - \frac{1}{m})$.

O Método CTM truncado com $\alpha = (1 - \frac{1}{m})$ foi superior à do CTM somente para tamanhos amostrais grandes. Esta melhora é devido ao fato de que, no Método CTM truncado com $\alpha = (1 - \frac{1}{m})$, uma árvore de contextos é estimada de cada uma das amostras enquanto que, nos demais, estima-se uma árvore de contextos a partir de uma coleção formada por amostras independentes, ou seja, neste método, necessita-se de amostras maiores para estimar uma árvore que capte a estrutura do processo de lei P.

Observou-se que os métodos podem ser menos eficientes quando a proporção de amostras contaminadas é próxima de 50% do total de amostras.

Na subseção 2.4.3 simulou-se amostras formadas por alguns pedaços do processo de lei P e por outros do processo de lei Q, ou seja, a contaminação está presente em alguns pedaços de uma mesma amostra. Neste caso, observou-se que o Método CTM truncado com $\alpha = (1 - \frac{1}{m})$ foi superior aos demais em todas as situações analisadas.

De forma geral, para os casos simulados, conclui-se que o Método CTM truncado com $\alpha = \frac{1}{2}$ é superior para o caso em que há amostras inteiramente contaminadas, ou seja, algumas amostras são geradas pelo processo verdadeiro e outras geradas pelo processo contaminado. Já o Método CTM truncado com $\alpha = (1 - \frac{1}{m})$ é superior para o caso em que a contaminação está presente em alguns pedaços de uma mesma amostra.

Uma extensão desse problema é o desenvolvimento de uma metodologia que permita a escolha de uma árvore de contextos quando há árvores provenientes de vários processos distintos.

O segundo problema, considerando processos Markovianos multivariados, consistiu na criação de uma nova família de modelos Markovianos de ordem finita utilizando o Critério de Informação Bayesiano (BIC) para a seleção de modelos dentro desta família. Para este caso, a consistência do BIC foi demonstrada.

Em outras palavras, através da máxima verossimilhança de um processo Markoviano estacionário multivariado e da verossimilhança máxima de um processo Markoviano estacionário multivariado com partes condicionalmente independentes, ambas obtidas nesta tese, foi possível o estabelecimento do critério para a obtenção da estrutura de independência condicional do processo. Este critério estabelece que a estrutura de independência condicional é obtida minimizando-se o BIC sob todas

as partições do conjunto $\{1, \ldots, l\}$ e sob todas as sequências pertencentes ao espaço de estado sendo que l é o número de coordenadas do processo multivariado.

Porém, quando o número de coordenadas do processo cresce, o custo computacional do critério BIC torna-se excessivo. Neste caso, foi proposto um algoritmo mais eficiente do ponto de vista computacional e a sua consistência foi demonstrada.

A eficiência do critério proposto foi estudada através de simulações. Foram realizadas algumas simulações de processos Markovianos estacionários
$(Z_t)_{t\in\mathbb{Z}} = \left(Z_t^{(1)}, Z_t^{(2)}\right)_{t\in\mathbb{Z}}$ de ordens 1 e 2 cujas coordenadas são condicionalmente independentes para cada um dos espaços de estados.

Para os processos de ordem 1, o critério proposto detectou corretamente a estrutura de independência condicional do processo $(Z_t)_{t\in\mathbb{Z}}$ em mais de 90% das amostras geradas. Já, para os de ordem 2, o critério detectou a estrutura de independência em mais de 80% das amostras geradas, exceto para amostras de tamanho 3.000. Em ambos os processos, a eficiência do método foi melhorando conforme aumentou-se o tamanho das amostras simuladas

Em um processo Markoviano de ordem 2 cujas coordenadas são condicionalmente independentes para cada estado, é necessário que o método proposto capte a independência condicional em cada um dos 16 possíveis estados do processo enquanto que, para processos de ordem 1, o método terá que captar a independência condicional em cada um dos 4 possíveis estados do processo. Desta forma, é esperado que a eficiência do método para cadeias de Markov de ordem 2 seja inferior em relação às cadeias de ordem 1.

Além disso, simulamos um processo Markoviano estacionário $(Z_t)_{t\in\mathbb{Z}} = \left(Z_t^{(1)}, Z_t^{(2)}\right)_{t\in\mathbb{Z}}$ de ordem 1 cujas coordenadas $(Z_t^{(1)})_{t\in\mathbb{Z}}$ e $(Z_t^{(2)})_{t\in\mathbb{Z}}$ são condicionalmente independentes dadas as sequências (0,0) e (0,1). O procedimento proposto detectou a estrutura de independência condicional do processo em mais de 90% das amostras geradas sendo que a sua eficiência foi melhorando conforme aumentou-se o tamanho das amostras simuladas.

Por fim, foram simuladas amostras independentes de uma distribuição multinomial bivariada com variáveis marginais independentes e dependentes. Em relação ao primeiro caso, foram geradas amostras independentes de uma distribuição multinomial cujos possiveis valores são 0, 1, 2 ou 3 os quais são igualmente prováveis. Observou-se que, para amostras de tamanho 15, o desempenho do procedimento proposto foi inferior ao do Teste Qui-Quadrado de Independência enquanto que, para

os demais tamanhos amostrais, o procedimento proposto e o Teste Qui-Quadrado de Independência apresentaram resultados semelhantes.

Já, em relação ao outro caso, foram geradas amostras independentes de uma distribuição multinomial cujos possiveis valores são 0, 1, 2 ou 3 com probabilidades $\frac{1}{16}$, $\frac{1}{4}$, $\frac{1}{2}$ e $\frac{3}{16}$, respectivamente. Observou-se que, para amostras de tamanho 15, o desempenho do procedimento proposto se destaca em relação à eficiência do Teste Qui-Quadrado de Independência enquanto que, para os demais tamanhos amostrais, o procedimento proposto e o Teste Qui-Quadrado de Independência apresentaram resultados semelhantes.

De modo geral, pelas simulações realizadas, conclui-se que o desempenho do critério proposto foi satisfatório, ou seja, o método é viável para a obtenção da estrutura de dependência condicional de um processo Markoviano estacionário multivariado. Além disso, o seu desempenho foi similar ao do Teste Qui-Quadrado de Independência para testar se as coordenadas de uma distribuição multinomial multivariada são independentes com a vantagem de não necessitar de uma distribuição de probabilidade para a verificação da independência em questão.

Capítulo 5

Demonstrações

5.1 Demonstrações dos Teoremas do Capítulo 2

Teorema 2.2. *Considere dois processos markovianos de memória variável, estacionários de ordem finita tomando valores no alfabeto finito $\mathcal{A}$ e com leis P e Q, respectivamente. Então, a taxa de entropia relativa entre eles é dada por*

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

Demonstração. Sejam $d = \max\{d(\mathcal{T}_P), d(\mathcal{T}_Q)\}$ e $\mathcal{A}^d$ o conjunto de todas as sequências de comprimento d. Denota-se por P_n e Q_n, respectivamente, as probabilidades $P(x)$ e $Q(x)$, para algum $x \in \mathcal{A}^d$.

Dada uma sequência $z \in \mathcal{A}^{n+1}$, $n > d$, existem $x \in \mathcal{A}^n$, $y \in \mathcal{A}$ tais que $z = xy$. Então,

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

$$\sum_{x \in \mathcal{A}^n} \sum_{y \in \mathcal{A}} P(y|x)P(x) \log\left(\frac{P(y|x)P(x)}{Q(y|x)Q(x)}\right) = \sum_{x \in \mathcal{A}^n} \sum_{y \in \mathcal{A}} P(y|x)P(x) \log\left(\frac{P(y|x)}{Q(y|x)}\right) +$$

$$\sum_{x \in \mathcal{A}^n} \sum_{y \in \mathcal{A}} P(y|x)P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Usando $\sum_{y\in\mathcal{A}} P(y|x) = 1$ obtemos

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

Agora, utilizando as definições

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

para alguma sequência $x \in \mathcal{A}^n$, e

$$D(P_n||Q_n) = \sum_{x\in\mathcal{A}^n} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

temos $D(P_{n+1}||Q_{n+1}) = \sum_{x\in\mathcal{A}^n} P(x)D(P(\cdot|x)||Q(\cdot|x)) + D(P_n||Q_n)$.

Dado $x \in \mathcal{A}^n$, $\exists$ uma sequência x_1 tal que $x = x_1 s$, $s \in \mathcal{S}$ e x_1 é sufixo de s. Desta forma,

$$D(P_{n+1}||Q_{n+1}) = \sum_{s\in\mathcal{T}_{PQ}} \sum_{x_1\in\mathcal{A}^k : s \text{ é sufixo de } x_1} P(x)D(P(\cdot|s)||Q(\cdot|s)) + D(P_n||Q_n) =$$

$$\sum_{s\in\mathcal{T}_{PQ}} D(P(\cdot|s)||Q(\cdot|s)) \left[\sum_{x_1\in\mathcal{A}^k : s \text{ é sufixo de } x_1} P(x)\right] + D(P_n||Q_n) =$$

$$\sum_{s\in\mathcal{T}_{PQ}} D(P(\cdot|s)||Q(\cdot|s))P(s) + D(P_n||Q_n) \tag{5.1.1}$$

já que

$$P(s) = \sum_{x_1\in\mathcal{A}^k : s \text{ é sufixo de } x_1} P(x).$$

Então, utilizando o raciocínio acima, $D(P_n||Q_n)$ pode ser escrito como

$$D(P_n||Q_n) = \sum_{s\in\mathcal{T}_{PQ}} P(s)D(P(\cdot|s)||Q(\cdot|s)) + D(P_{n-1}||Q_{n-1}) \tag{5.1.2}$$

Substituindo (5.1.2) em (5.1.1) temos

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

Desenvolvendo, recursivamente, a expressão $D(P_{n-1}||Q_{n-1})$ até a sequência $s \in \mathcal{T}_{PQ}$ de comprimento d obtemos

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

Logo, a taxa de entropia relativa entre os processos de leis P e Q é dada por

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

Portanto,

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

□

Teorema 2.3. *Considere o estimador definido por*

$$\hat{\mathcal{T}}_i(\mathcal{C}_n^m) = \hat{Q}_{j_i^*(\mathcal{C}_n^m)}(\mathcal{C}_n^m)$$

Então, $\hat{\mathcal{T}}_i(\mathcal{C}_n^m)$ possui ponto de ruptura assintótico igual a $\frac{1}{2}$ para todo $i < \frac{m}{2}$.

Demonstração. Agora, suponha que $|\mathcal{I}_P| = m-k$ e $|\mathcal{I}_Q| = k$, isto é, k das m amostras são geradas pelo processo de lei Q e as $m-k$ restantes são geradas pelo processo de lei P.

Para qualquer j,

$$\hat{V}_j(\mathcal{C}_n^m) = \frac{1}{m}\sum_{i\in\mathcal{I}_P}\overline{d}_{(j,i)}(\mathcal{C}_n^m) + \frac{1}{m}\sum_{i\in\mathcal{I}_Q}\overline{d}_{(j,i)}(\mathcal{C}_n^m)$$

A partir do Lema 2.1, para qualquer $\epsilon > 0$, existe $N_0^\epsilon > 0$ tal que, para $j \in \mathcal{I}_Q$ e para qualquer $n > N_0^\epsilon$,

$$\begin{aligned} 0 \leq \ & \tfrac{1}{m}\sum_{i\in\mathcal{I}_Q}\overline{d}_{(j,i)}(\mathcal{C}_n^m) \ \leq \epsilon\frac{k}{m} \ \text{ e} \\ \Big(\overline{D}\Big(Q,P\Big) - \epsilon\Big)\frac{(m-k)}{m} \leq \ & \tfrac{1}{m}\sum_{i\in\mathcal{I}_P}\overline{d}_{(j,i)}(\mathcal{C}_n^m) \ \leq \Big(\overline{D}\Big(Q,P\Big) - \epsilon\Big)\frac{(m-k)}{m}. \end{aligned}$$

Então, $\hat{V}_j(\mathcal{C}_n^m)$ é limitado por,

$$\Big(\overline{D}\Big(Q,P\Big) - \epsilon\Big)\frac{(m-k)}{m} \leq \hat{V}_j(\mathcal{C}_n^m) \leq \epsilon + \overline{D}\Big(Q,P\Big)\frac{(m-k)}{m}, \ \ \forall j \in \mathcal{I}_Q \tag{5.1.3}$$

Além disso, existe $N_1^\epsilon > 0$ tal que, para $j \in \mathcal{I}_P$ e para qualquer $n > N_1^\epsilon$,

$$\Big(\overline{D}\Big(Q||P\Big) - \epsilon\Big)\frac{k}{m} \leq \hat{V}_j(\mathcal{C}_n^m) \leq \overline{D}\Big(Q,P\Big)\frac{k}{m} + \epsilon \tag{5.1.4}$$

Temos que $\frac{(m-k)}{m} < \frac{1}{2} \Rightarrow m-k < k$ e, como consequência, para ϵ suficientemente pequeno e $n > \max\{N_0^\epsilon, N_1^\epsilon\}$,

$$\epsilon + \overline{D}\Big(Q,P\Big)\frac{(m-k)}{m} < \Big(\overline{D}\Big(Q||P\Big) - \epsilon\Big)\frac{k}{m}$$

e $\hat{V}_j(\mathcal{C}_n^m) < \hat{V}_l(\mathcal{C}_n^m)$, para todo $j \in \mathcal{I}_Q$ e $l \in \mathcal{I}_P$.

Então,

$$j_i^*(\mathcal{C}_n^m) \leq k, \ \forall i < \frac{m}{2},$$

e

$$\lim_{n\to\infty}\hat{\mathcal{T}}_i(\mathcal{C}_n^m) = Q, \ \text{ quase certamente } \ \forall i < \frac{m}{2}.$$

Do mesmo modo, $\frac{(m-k)}{m} > \frac{1}{2} \Rightarrow m-k > k$, e

$$\lim_{n\to\infty}\hat{\mathcal{T}}_i(\mathcal{C}_n^m) = P, \ \text{ quase certamente } \ \forall i < \frac{m}{2}.$$

□

5.2 Demonstrações dos Teoremas do Capítulo 3

Teorema 3.1. *O valor de* $P(a|s)$ *que maximiza* $\prod_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}} P(a|s)^{N_n(s,a)}$ *sujeito à restrição* $\sum_{a\in\mathcal{A}} P(a|s) = 1$, *para cada* $s \in \mathcal{S}$, *é* $\hat{P}(a|s) = \frac{N_n(s,a)}{\sum_{a\in\mathcal{A}} N_n(s,a)}$.

Demonstração. Como a função $\prod_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}} P(a|s)^{N_n(s,a)}$ é maximizada em relação a $P(a|s)$, $\forall a \in \mathcal{A}$ e $\forall s \in \mathcal{S}$, sujeito à restrição $\sum_{a\in\mathcal{A}} P(a|s) = 1$, para cada $s \in \mathcal{S}$, considere o Lagrangeano

$$\mathcal{L}\big(z_1^n|\mathcal{P}\big) = \sum_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}} N_n(s,a)\ln(P(a|s)) - \sum_{s\in\mathcal{S}} \lambda_s \left(\sum_{a\in\mathcal{A}} P(a|s) - 1\right)$$

Derivando $\mathcal{L}\big(z_1^n|\mathcal{P}\big)$ em relação à $P(a|s)$ temos

$$\frac{\partial\mathcal{L}\big(z_1^n|\mathcal{P}\big)}{\partial P(a|s)} = \frac{N_n(s,a)}{P(a|s)} - \lambda_s \tag{5.2.1}$$

Igualando a expressão (5.2.1) a zero obtemos $P(a|s) = \frac{N_n(s,a)}{\lambda_s}$. Como $\sum_{a\in\mathcal{A}} P(a|s) = 1$ temos

$$\sum_{a\in\mathcal{A}} \left(\frac{N(s,a)}{\lambda_s}\right) = 1 \Rightarrow \lambda_s = \sum_{a\in\mathcal{A}} N_n(s,a)$$

Portanto, $\hat{P}(a|s) = \frac{N_n(s,a)}{\sum_{a\in\mathcal{A}} N_n(s,a)}$ é o valor de $P(a|s)$ que maximiza $\prod_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}} P(a|s)^{N_n(s,a)}$ sujeito à restrição $\sum_{a\in\mathcal{A}} P(a|s) = 1$ para cada $s \in \mathcal{S}$.

Como $\sum_{a\in\mathcal{A}} N_n(s,a) = N_n(s)$ temos

$$\hat{P}(a|s) = \frac{N_n(s,a)}{N_n(s)}$$

□

Teorema 3.2. *O valor de* $P(a^{I_j^s}|s)$ *que maximiza* $\prod_{s\in\mathcal{S}}\prod_{a\in\mathcal{A}}\prod_{j=1}^{|\mathcal{I}_s|}P\big(a^{I_j^s}|s\big)^{N_n(s,a)}$ *sujeito à* $\sum_{a^{I_j^s}\in\mathcal{B}^{|I_j^s|}}P^{I_j^s}(a|s)=1$, *para cada* $s\in\mathcal{S}$ *e* $j=1,\ldots,|\mathcal{I}_s|$, *é*

$$\hat{P}\big(a^{I_j^s}|s\big)=\frac{N_n\big(s,a^{I_j^s}\big)}{N_n(s)}$$

Demonstração. Considere o Lagrangeano

$$\mathcal{L}\big(z_1^n,\mathcal{J}|\mathcal{P}\big)=\sum_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}}\sum_{j=1}^{|\mathcal{I}_s|}N_n(s,a)\ln\big(P^{I_j^s}(a|s)\big)-\sum_{s\in\mathcal{S}}\sum_{j=1}^{|\mathcal{I}_s|}\lambda_j^s\left(\sum_{a^{I_j^s}\in\mathcal{B}^{|I_j^s|}}P^{I_j^s}(a|s)-1\right)$$

Derivando $\mathcal{L}\big(z_1^n,\mathcal{J}|\mathcal{P}\big)$ em relação à $P^{I_j^s}(a|s)$ temos

$$\frac{\partial\mathcal{L}\big(z_1^n|\mathcal{P}\big)}{\partial P^{I_j^s}(a|s)}=\sum_{a^{\{1,\ldots,l\}\setminus I_j^s}\in\mathcal{B}^{l-|I_j^s|}}\frac{N_n\big(s,a\big)}{P^{I_j^s}(a|s)}-\lambda_j^s=\frac{N_n\big(s,a^{I_j^s}\big)}{P^{I_j^s}(a|s)}-\lambda_j^s \tag{5.2.2}$$

Igualando a expressão (5.2.2) a zero obtemos $P^{I_j^s}(a|s)=\frac{N_n(s,a^{I_j^s})}{\lambda_j^s}$.
Como, para cada $j=1,\ldots,|\mathcal{I}_s|$ e $s\in\mathcal{S}$, $\sum_{a^{I_j^s}\in\mathcal{B}^{|I_j^s|}}P^{I_j^s}(a|s)=1$, então

$$\sum_{a^{I_j^s}\in\mathcal{B}^{|I_j^s|}}\left(\frac{N\big(s,a^{I_j^s}\big)}{\lambda_j^s}\right)=1\Rightarrow\lambda_j^s=\sum_{a^{I_j^s}\in\mathcal{B}^{|I_j^s|}}N_n\big(s,a^{I_j^s}\big)$$

Portanto, $\hat{P}^{I_j^s}(a|s)=\dfrac{N_n\big(s,a^{I_j^s}\big)}{\sum_{a^{I_j^s}\in\mathcal{B}^{|I_j^s|}}N_n\big(s,a^{I_j^s}\big)}$.
Como $\sum_{a^{I_j^s}\in\mathcal{B}^{|I_j^s|}}N_n\big(s,a^{\mathcal{I}_j^s}\big)=N_n(s)$, obtemos

$$\hat{P}^{I_j^s}(a|s)=\frac{N_n\big(s,a^{I_j^s}\big)}{N_n(s)}$$

□

Teorema 3.3. *Considere $(Z_t)_{t\in\mathbb{Z}}$ um processo de Markov estacionário de ordem finita M tomando valores no alfabeto finito $\mathcal{A} = \mathcal{B}^l$ e seja $\mathcal{S} = \mathcal{A}^M$ o seu espaço de estados. Além disso, considere $z_1^n = z_1 \dots z_n$, $z_i \in \mathcal{A}$, $i = 1, \dots, n$, uma amostra deste processo, $\mathcal{J} = \{\mathcal{I}_s\}_{s\in\mathcal{S}}$ um conjunto de partições de $\mathcal{L} = \{1, \dots, l\}$ e $\mathcal{R} = \mathcal{L}^{|\mathcal{S}|}$. Então, eventualmente quase certamente, quando $n \to \infty$,*

$$BIC(z_1^n, \mathcal{S}, \mathcal{R}) \geq BIC(z_1^n, \mathcal{S}, \mathcal{J})$$

se, e somente se, $\mathcal{I}_s$ é compatível com a lei condicional do processo dado s, $\forall s \in \mathcal{S}$.

Demonstração. Inicialmente, será analisada a diferença entre as expressões (3.4.4) e (3.4.3).

$$\begin{aligned} BIC(z_1^n, \mathcal{S}, \mathcal{R}) - BIC(z_1^n, \mathcal{S}, \mathcal{J}) &= -\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} N_n(s,a)\ln\left(\frac{N_n(s,a)}{N_n(s)}\right) + \frac{(|\mathcal{B}|^l - 1)|\mathcal{S}|}{2}\ln(n) \\ + \quad \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{j=1}^{|\mathcal{I}_s|} N_n(s,a)\ln\left(\frac{N_n\left(s,a^{I_j^s}\right)}{N_n(s)}\right) &- \frac{\mathcal{C}-\mathcal{D}}{2}\ln(n) = \\ \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-N_n(s,a)\ln\left(\frac{N_n(s,a)}{N_n(s)}\right) + \sum_{j=1}^{|\mathcal{I}_s|} N_n(s,a)\ln\left(\frac{N_n\left(s,a^{I_j^s}\right)}{N_n(s)}\right)\right] &+ \frac{\mathcal{E}}{2}\ln(n) \end{aligned} \tag{5.2.3}$$

sendo $\mathcal{E} = (|\mathcal{B}|^l - 1)|\mathcal{S}| - \mathcal{C} + \mathcal{D}$, $\mathcal{C} = \sum_{s\in\mathcal{S}}\sum_{j=1}^{|\mathcal{I}_s|} |\mathcal{B}|^{|I_j^s|}$ e $\mathcal{D} = \sum_{s\in\mathcal{S}} |\mathcal{I}_s|$.

Definindo, para cada $a \in \mathcal{A}$ e $s \in \mathcal{S}$, $r_n(s,a) = \frac{N_n(s,a)}{n}$, $r_n(s) = \frac{N_n(s)}{n}$ e $r_n\left(s, a^{I_j^s}\right) = \frac{N_n\left(s,a^{I_j^s}\right)}{n}$, a expressão (5.2.3) pode ser reescrita como

$$\begin{aligned} &BIC(z_1^n, \mathcal{S}, \mathcal{R}) - BIC(z_1^n, \mathcal{S}, \mathcal{J}) = \\ &n\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-\frac{N_n(s,a)}{n}\ln\left(\frac{N_n(s,a)/n}{N_n(s)/n}\right) + \sum_{j=1}^{|\mathcal{I}_s|}\frac{N_n(s,a)}{n}\ln\left(\frac{N_n\left(s,a^{I_j^s}\right)/n}{N_n(s)/n}\right)\right] + \frac{\mathcal{E}}{2}\ln(n) = \\ &n\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-r_n(s,a)\ln\left(\frac{r_n(s,a)}{r_n(s)}\right) + \sum_{j=1}^{|\mathcal{I}_s|} r_n(s,a)\ln\left(\frac{r_n\left(s,a^{I_j^s}\right)}{r_n(s)}\right)\right] + \frac{\mathcal{E}}{2}\ln(n) \end{aligned} \tag{5.2.4}$$

Supondo que $BIC(z_1^n, \mathcal{S}, \mathcal{R}) - BIC(z_1^n, \mathcal{S}, \mathcal{J}) \geq 0$, a relação (5.2.4) fornece

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-r_n(s,a)\ln\left(\frac{r_n(s,a)}{r_n(s)}\right)+\sum_{j=1}^{|\mathcal{I}_s|} r_n(s,a)\ln\left(\frac{r_n\left(s,a^{I_j^s}\right)}{r_n(s)}\right)\right]+\frac{\mathcal{E}}{2}\frac{\ln(n)}{n}\geq 0 \qquad (5.2.5)$$

Como o termo $\frac{\mathcal{E}}{2}\frac{\ln(n)}{n}$ tende a zero, quando $n\to\infty$, e $r_n(s,a)\to Prob(Z_t=a, Z_{t-M}^{t-1}=s)=P(sa)$, $\frac{r_n(s,a)}{r_n(s)}\to P(a|s)$, $\frac{r_n\left(s,a^{I_j^s}\right)}{r_n(s)}\to P^{I_j^s}(a|s)$, quando $n\to\infty$, a desigualdade (5.2.5) fornece

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-P(sa)\ln(P(a|s))+\sum_{j=1}^{|\mathcal{I}_s|}P(sa)\ln\left(P^{I_j^s}(a|s)\right)\right]\geq 0$$

ou seja,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}P(sa)\ln\left(\frac{\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(a|s)}{P(a|s)}\right)\geq 0 \qquad (5.2.6)$$

Por outro lado, utilizando $P(s)=Prob(Z_{t-l(s)}^t=s)>0$ e a relação $P(a|s)=\frac{P(sa)}{P(s)}$, temos

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}P(sa)\ln\left(\frac{\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(a|s)}{P(a|s)}\right)=\sum_{s\in\mathcal{S}}P(s)\sum_{a\in\mathcal{A}}\frac{P(sa)}{P(s)}\ln\left(\frac{\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(a|s)}{P(a|s)}\right)=$$

$$-\sum_{s\in\mathcal{S}}P(s)\sum_{a\in\mathcal{A}}P(a|s)\ln\left(\frac{P(a|s)}{\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(a|s)}\right)=-\sum_{s\in\mathcal{S}}P(s)D\left(P(\cdot|s)||\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(\cdot|s)\right)$$

sendo $D\left(P(\cdot|s)||\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(\cdot|s)\right)$ a entropia relativa entre $P(\cdot|s)$ e $\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(\cdot|s)$.

Pelo Teorema 2.1, $D\left(P(\cdot|s)||\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(\cdot|s)\right)\geq 0$. Logo,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}P(sa)\ln\left(\frac{\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(a|s)}{P(a|s)}\right)\leq 0 \tag{5.2.7}$$

Pelas desigualdades (5.2.6) e (5.2.7) conclui-se que

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}P(sa)\ln\left(\frac{\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(a|s)}{P(a|s)}\right)=-\sum_{s\in\mathcal{S}}P(s)D\left(P(\cdot|s)||\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(\cdot|s)\right)=0 \tag{5.2.8}$$

Pelo Teorema 2.1, a igualdade (5.2.8) ocorre quando $P(a|s)=\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(a|s)$ para qualquer $a\in\mathcal{A}$ e $s\in\mathcal{S}$.

Portanto, para n suficientemente grande, se $BIC(z_1^n,\mathcal{S},\mathcal{R})\geq BIC(z_1^n,\mathcal{S},\mathcal{J})$ então $\mathcal{I}_s$ é compatível com a lei condicional do processo dado s, $\forall s\in\mathcal{S}$.

Agora, será provada a recíproca do teorema. Suponha que $\mathcal{I}_s$ é compatível com a lei condicional do processo dado s, $\forall s\in\mathcal{S}$.

Como $\frac{N_n\left(s,a^{I_j^s}\right)}{N_n(s)}$ é o estimador de máxima verossimilhança de $P^{I_j^s}(a|s)$, então

$$\prod_{s\in\mathcal{S}}\prod_{a\in\mathcal{A}}\prod_{j=1}^{|\mathcal{I}_s|}\left(\frac{N_n\left(s,a^{I_j^s}\right)}{N_n(s)}\right)^{N_n(s,a)}\geq\prod_{s\in\mathcal{S}}\prod_{a\in\mathcal{A}}\prod_{j=1}^{|\mathcal{I}_s|}\left(P^{I_j^s}(a|s)\right)^{N_n(s,a)} \tag{5.2.9}$$

Usando (5.2.9) temos

$$BIC(z_1^n,\mathcal{S},\mathcal{R}) - BIC(z_1^n,\mathcal{S},\mathcal{J}) =$$
$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-N_n(s,a)\ln\left(\frac{N_n(s,a)}{N_n(s)}\right)+\sum_{j=1}^{|\mathcal{I}_s|}N_n(s,a)\ln\left(\frac{N_n\left(s,a^{I_j^s}\right)}{N_n(s)}\right)\right]+\frac{\mathcal{E}}{2}\ln(n)\geq$$
$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-N_n(s,a)\ln\left(\frac{N_n(s,a)}{N_n(s)}\right)+\sum_{j=1}^{|\mathcal{I}_s|}N_n(s,a)\ln\left(P^{I_j^s}(a|s)\right)\right]+\frac{\mathcal{E}}{2}\ln(n) \qquad (5.2.10)$$

Utilizando a hipótese $P(a|s)=\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(a|s)$, para quaisquer $a\in\mathcal{A}$ e $s\in\mathcal{S}$, a relação (5.2.10) torna-se

$$BIC(z_1^n,\mathcal{S},\mathcal{R}) - BIC(z_1^n,\mathcal{S},\mathcal{J}) \geq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-N_n(s,a)\ln\left(\frac{N_n(s,a)}{N_n(s)}\right)+\sum_{j=1}^{|\mathcal{I}_s|}N_n(s,a)\ln\left(P^{I_j^s}(a|s)\right)\right]$$
$$+\quad \frac{\mathcal{E}}{2}\ln(n)=\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-N_n(s,a)\ln\left(\frac{N_n(s,a)}{N_n(s)}\right)+N_n(s,a)\ln\left(\prod_{j=1}^{|\mathcal{I}_s|}P^{I_j^s}(a|s)\right)\right]+\frac{\mathcal{E}}{2}\ln(n)=$$
$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[-N_n(s,a)\ln\left(\frac{N_n(s,a)}{N_n(s)}\right)+N_n(s,a)\ln(P(a|s))\right]+\frac{\mathcal{E}}{2}\ln(n)$$

a qual pode ser reescrita como

$$BIC(z_1^n,\mathcal{S},\mathcal{R}) - BIC(z_1^n,\mathcal{S},\mathcal{J}) \geq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}N_n(s,a)\ln\left(\frac{P(a|s)N_n(s)}{N_n(s,a)}\right)+\frac{\mathcal{E}}{2}\ln(n)=$$
$$-\quad \sum_{s\in\mathcal{S}}N_n(s)\sum_{a\in\mathcal{A}}\frac{N_n(s,a)}{N_n(s)}\ln\left(\frac{N_n(s,a)}{P(a|s)N_n(s)}\right)+\frac{\mathcal{E}}{2}\ln(n) \qquad (5.2.11)$$

Utilizando os Lemas 1.1 e 1.2 temos

$$\sum_{a\in\mathcal{A}}\frac{N_n(s,a)}{N_n(s)}\ln\left(\frac{N_n(s,a)}{P(a|s)N_n(s)}\right)=D\left(\frac{N_n(s,a)}{N_n(s)}\middle\|P(a|s)\right)\leq\sum_{a\in\mathcal{A}}\frac{\left(\frac{N_n(s,a)}{N_n(s)}-P(a|s)\right)^2}{P(a|s)}\leq\sum_{a\in\mathcal{A}}\frac{\frac{\delta\ln(n)}{N_n(s)}}{P(a|s)}$$

para qualquer $\delta>0$ e n suficientemente grande.

Logo, a relação (5.2.11) torna-se

$$\begin{aligned} BIC(z_1^n,\mathcal{S},\mathcal{R}) - BIC(z_1^n,\mathcal{S},\mathcal{J}) \;\; &\geq \;\; -\sum_{s\in\mathcal{S}} N_n(s) \sum_{a\in\mathcal{A}} \frac{\frac{\delta\ln(n)}{N_n(s)}}{P(a|s)} + \frac{\mathcal{E}}{2}\ln(n) = \\ &- \;\; \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{\delta\ln(n)}{P(a|s)} + \frac{\mathcal{E}}{2}\ln(n) \end{aligned}$$

Utilizando $p = \max\Big\{P(a|s) : a \in \mathcal{A} \text{ e } \; s \in \mathcal{S}\Big\}$, temos

$$BIC(z_1^n,\mathcal{S}) - BIC(z_1^n,\mathcal{S},\mathcal{J}) \geq -\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{\delta\ln(n)}{p} + \frac{\mathcal{E}}{2}\ln(n) = -\frac{\delta|\mathcal{S}||\mathcal{A}|}{p}\ln(n) + \frac{\mathcal{E}}{2}\ln(n)$$

Em particular, tomando

$$\delta < \frac{p}{|\mathcal{S}||\mathcal{A}|}\frac{\mathcal{E}}{2}$$

para n suficientemente grande, $BIC(z_1^n,\mathcal{S},\mathcal{R}) \geq BIC(z_1^n,\mathcal{S},\mathcal{I})$.

Logo, para n suficientemente grande, se $P(a|s) = \prod_{j=1}^{|\mathcal{I}_s|} P^{I_j^s}(a|s)$, $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$, então $BIC(z_1^n,\mathcal{S},\mathcal{R}) \geq BIC(z_1^n,\mathcal{S},\mathcal{J})$.

Portanto, demonstrou-se que, eventualmente quase certamente, $BIC(z_1^n,\mathcal{S},\mathcal{R}) \geq BIC(z_1^n,\mathcal{S},\mathcal{J})$ se, e somente se, $P(a|s) = \prod_{j=1}^{|\mathcal{I}_s|} P^{I_j^s}(a|s)$, $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$. □

Teorema 3.4. *Sob as suposições do Teorema 3.3, sejam* $\mathcal{I}_s = \Big\{I_1^s,\ldots,I_{|\mathcal{I}_s|}^s\Big\}$ *uma partição de* $\{1,\ldots,l\}$ *compatível com a lei condicional do processo dado* s, $\forall s \in \mathcal{S}$, *e* $\mathcal{J} = \big\{\mathcal{I}_s\big\}_{s\in\mathcal{S}}$ *um conjunto de partições compatível com a lei do processo. Suponha que, em relação à sequência* s_m, *para cada* $j = 1,\ldots,|\mathcal{I}_{s_m}|$, *existem subconjuntos* $\tilde{I}_{j_k}^{s_m}$, $k = 1,\ldots,k_j$, *de* $I_j^{s_m}$ *tais que* $\tilde{I}_{j_k}^{s_m} \preceq I_j^{s_m}$ *e* $P\big(a|s\big) = \prod_{k=1}^{k_j} P^{\tilde{I}_{j_k}^{sm}}(a|s)$, $\forall a \in \mathcal{A}$. *Então, eventualmente quase certamente, quando* $n \to \infty$,

$$BIC\big(z_1^n,\mathcal{S},\mathcal{J}\big) \geq BIC\big(z_1^n,\mathcal{S},\tilde{\mathcal{J}}\big)$$

se, e somente se,

$$P(a|s_m) = \prod_{j=1}^{|\mathcal{I}_{sm}|}\prod_{k=1}^{k_j} P^{\tilde{I}_{j_k}^{sm}}(a|s_m), \forall a \in \mathcal{A}$$

sendo $\tilde{\mathcal{J}} = \left\{\mathcal{I}_{s_1}, \ldots, \mathcal{I}_{s_{m-1}}, \tilde{\mathcal{I}}_{s_m}, \mathcal{I}_{s_{m+1}}, \ldots, \mathcal{I}_{s_{|\mathcal{S}|}}\right\}$ *e* $\tilde{\mathcal{I}}_{s_m} = \left\{\tilde{I}_{j_k}^{s_m}\right\}_{\substack{k=1,\ldots,k_j \\ j=1,\ldots,|\tilde{\mathcal{I}}_{s_m}|}}$.

Demonstração. A expressão do $BIC(z_1^n, \mathcal{S}, \mathcal{J})$ é (3.4.3), isto é,

$$BIC(z_1^n, \mathcal{S}, \mathcal{J}) = -\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{j=1}^{|\mathcal{I}_s|} N_n(s,a)\ln\left(\frac{N_n\left(s, a^{I_j^s}\right)}{N_n(s)}\right) + \sum_{s\in\mathcal{S}}\sum_{j=1}^{|\mathcal{I}_s|}\frac{|\mathcal{B}|^{|\mathcal{I}_j^s|}-1}{2}\ln(n)$$

enquanto o $BIC\left(z_1^n, \mathcal{S}, \tilde{\mathcal{J}}\right)$ é dado por

$$\begin{aligned} BIC\left(z_1^n, \mathcal{S}, \tilde{\mathcal{J}}\right) = \; & - \sum_{\substack{s\in\mathcal{S} \\ s\neq s_m}}\sum_{a\in\mathcal{A}}\sum_{j=1}^{|\mathcal{I}_s|} N_n(s,a)\ln\left(\frac{N_n\left(s, a^{I_j^s}\right)}{N_n(s)}\right) \\ & - \sum_{j=1}^{|\mathcal{I}_{s_m}|}\sum_{a\in\mathcal{A}}\sum_{k=1}^{k_j} N_n(s_m,a)\ln\left(\frac{N_n\left(s_m, a^{\tilde{I}_{j_k}^{s_m}}\right)}{N_n(s_m)}\right) \\ & + \sum_{\substack{s\in\mathcal{S} \\ s\neq s_m}}\sum_{j=1}^{|\mathcal{I}_s|}\frac{|\mathcal{B}|^{|I_j^s|}-1}{2}\ln(n) + \sum_{j=1}^{|\mathcal{I}_{s_m}|}\sum_{k=1}^{k_j}\frac{|\mathcal{B}|^{\left|\tilde{I}_{j_k}^{s_m}\right|}-1}{2}\ln(n) \end{aligned}$$

Logo,

$$\begin{aligned} & BIC\left(z_1^n, \mathcal{S}, \mathcal{J}\right) - BIC\left(z_1^n, \mathcal{S}, \tilde{\mathcal{J}}\right) = -\sum_{j=1}^{|\mathcal{I}_{s_m}|}\sum_{a\in\mathcal{A}} N_n(s_m,a)\ln\left(\frac{N_n\left(s_m, a^{I_j^{s_m}}\right)}{N_n(s_m)}\right) \\ & + \sum_{j=1}^{|\mathcal{I}_{s_m}|}\frac{|\mathcal{B}|^{|I_j^{s_m}|}-1}{2}\ln(n) + \sum_{j=1}^{|\mathcal{I}_{s_m}|}\sum_{a\in\mathcal{A}}\sum_{k=1}^{k_j} N_n(s_m,a)\ln\left(\frac{N_n\left(s_m, a^{\tilde{I}_{j_k}^{s_m}}\right)}{N_n(s_m)}\right) \\ & - \sum_{j=1}^{|\mathcal{I}_{s_m}|}\sum_{k=1}^{k_j}\frac{|\mathcal{B}|^{\left|\tilde{I}_{j_k}^{s_m}\right|}-1}{2}\ln(n) = -\sum_{j=1}^{|\mathcal{I}_{s_m}|}\sum_{a\in\mathcal{A}} N_n(s_m,a)\ln\left(\frac{N_n\left(s_m, a^{I_j^{s_m}}\right)}{N_n(s_m)}\right) \\ & + \sum_{j=1}^{|\mathcal{I}_{s_m}|}\sum_{a\in\mathcal{A}}\sum_{k=1}^{k_j} N_n(s_m,a)\ln\left(\frac{N_n\left(s_m, a^{\tilde{I}_{j_k}^{s_m}}\right)}{N_n(s_m)}\right) + \frac{\mathcal{F}}{2}\ln(n) \end{aligned} \tag{5.2.12}$$

na qual $\mathcal{F} = \sum_{j=1}^{|\mathcal{I}_{s_m}|}\left(|\mathcal{B}|^{|I_j^{s_m}|} - \sum_{k=1}^{k_j}|\mathcal{B}|^{\left|\tilde{I}_{j_k}^{s_m}\right|} + k_j\right) - |\mathcal{I}_{s_m}|$.

Logo, a expressão (5.2.12) fornece

$$
\begin{aligned}
&BIC\big(z_1^n,\mathcal{S},\mathcal{J}\big)-BIC\big(z_1^n,\mathcal{S},\tilde{\mathcal{J}}\big)=\\
&-\sum_{j=1}^{|\mathcal{I}_{s_m}|}\sum_{a^{I_j^{s_m}}\in\mathcal{B}^{|I_j^{s_m}|}} N_n(s_m,a^{I_j^{s_m}})\ln\left(\frac{N_n\left(s_m,a^{I_j^{s_m}}\right)}{N_n(s_m)}\right)\\
&+\sum_{j=1}^{|\mathcal{I}_{s_m}|}\sum_{a^{I_j^{s_m}}\in\mathcal{B}^{|I_j^{s_m}|}}\sum_{k=1}^{k_j} N_n(s_m,a^{I_j^{s_m}})\ln\left(\frac{N_n\left(s_m,a^{\tilde{I}_{j_k}^{s_m}}\right)}{N_n(s_m)}\right)\\
&+\frac{\mathcal{F}}{2}\ln(n)
\end{aligned}
\tag{5.2.13}
$$

A partir de (5.2.13), a demonstração do teorema segue a mesma idéia utilizada na demonstração do Teorema 3.3. □

Teorema 3.5. *Considere $(Z_t)_{t\in\mathbb{Z}}$ um processo de Markov estacionário de ordem finita tomando valores no alfabeto finito $\mathcal{A}=\mathcal{B}^l$ e seja $\mathcal{S}=\mathcal{A}^M$ o seu espaço de estados. Além disso, sejam $z_1^n=z_1\dots z_n$ uma amostra deste processo, $\wp=\mathcal{C}^{|\mathcal{S}|}$ no qual $\mathcal{C}$ é o conjunto formado por todas as partições de $\{1,2,\dots,l\}$ e $\mathcal{L}^\star=\{\mathcal{I}_s\}_{s\in\mathcal{S}}$ a estrutura de dependência condicional do processo. Seja*

$$\mathcal{L}_n^\star=\arg\min_{\mathcal{J}\in\wp} BIC\big(z_1^n,\mathcal{S},\mathcal{J}\big)$$

Então, eventualmente quase certamente, quando $n\to\infty$, $\mathcal{L}^\star=\mathcal{L}_n^\star$.

Demonstração. Sem perda de generalidade, considere, para todo $s\in\mathcal{S}$, $s\neq s_m$, uma partição $\mathcal{I}_s$ de $\{1,\dots,l\}$ que é compatível com a lei condicional do processo dado s. Além disso, sejam $\mathcal{I}_{s_m}=\big\{I_1^{s_m},\dots,I_{|\mathcal{I}_{s_m}|}^{s_m}\big\}$ uma partição de $\{1,\dots,l\}$ que não é compatível com a lei condicional do processo dado s_m e $\mathcal{J}=\big\{\mathcal{I}_s\big\}_{s\in\mathcal{S}}$. Suponha que existem dois conjuntos $I_u^{s_m}$ e $I_v^{s_m}$, $u\neq v$, pertencentes a $\mathcal{I}_{s_m}$ tais que $\overline{\mathcal{I}}_{s_m}=I_{uv}^{s_m}\cup\big\{I_j^{s_m}\big\}_{\substack{j=1,\dots,|\mathcal{I}_{s_m}|\\ j\neq u,v}}$ é compatível com a lei condicional do processo dado s_m sendo $I_{uv}^{s_m}=I_u^{s_m}\cup I_v^{s_m}$.

Suponha que $BIC\big(z_1^n,\mathcal{S},\mathcal{J}\big)<BIC\big(z_1^n,\mathcal{S},\mathcal{R}\big)$, $\forall\mathcal{R}\in\wp$. Considerando $\mathcal{J}^{uv}=\overline{\mathcal{I}}_{s_m}\cup\big\{\mathcal{I}_s\big\}_{\substack{s\in\mathcal{S}\\ s\neq s_m}}$, conclui-se, pelo Teorema 3.4, que $BIC\big(z_1^n,\mathcal{S},\mathcal{J}\big)\geq BIC\big(z_1^n,\mathcal{S},\mathcal{J}^{uv}\big)$ quando $n\to\infty$, o que é um absurdo pela suposição da existência de $\mathcal{J}=\big\{\mathcal{I}_s\big\}_{s\in\mathcal{S}}$ tal que o BIC é o mínimo em $\wp$.

Agora, considere $\wp^{'}$ um conjunto de partições compatível com a lei do processo, $\mathcal{J}$ um elemento arbitrário de $\wp^{'}$, $\mathcal{R} = \mathcal{L}^{|\mathcal{S}|}$ e $\mathcal{L} = \{1, \ldots, l\}$. Pelo Teorema 3.3, eventualmente quase certamente quando $n \to \infty$, $BIC\big(z_1^n, \mathcal{S}, \mathcal{R}\big) \geq BIC\big(z_1^n, \mathcal{S}, \mathcal{J}\big)$.

Defina, quando $n \to \infty$, $\mathcal{L}^\star = \arg\min_{\mathcal{J} \in \wp^{'}} BIC\big(z_1^n, \mathcal{S}, \mathcal{J}\big)$. Por construção, $\mathcal{L}^\star \in \wp^{'}$.

Suponha que $\mathcal{L}^\star = \{\mathcal{I}_s^\star\}_{s \in \mathcal{S}}$ não é a estrutura de dependência condicional do processo sendo $\mathcal{I}_s^\star = \left\{I_1^{\star^s}, \ldots, I_{|\mathcal{I}_s^\star|}^{\star^s}\right\}$. Então, para cada $s \in \mathcal{S}$ e para cada $j = 1, \ldots, |\mathcal{I}_s^\star|$, existem $\tilde{I}_{j_k}^{\star_s} \preceq I_j^{\star_s}$, $k = 1, \ldots, k_j$, tais que $\mathcal{J}^\star = \{\tilde{\mathcal{I}}_s^\star\}_{s \in \mathcal{S}}$ é a estrutura de dependência condicional do processo sendo $\tilde{\mathcal{I}}_s^\star = \left\{\tilde{I}_{j_k}^{\star_s}\right\}_{\substack{k=1,\ldots,k_j \\ j=1,\ldots,|\mathcal{I}_s^\star|}}$. Logo, $BIC\big(z_1^n, \mathcal{S}, \mathcal{L}^\star\big) \geq BIC\big(z_1^n, \mathcal{S}, \mathcal{J}^\star\big)$ e isto é impossível pois $\mathcal{L}^\star$ foi definido como sendo a estrutura de dependência condicional do processo. $\square$

Referências Bibliográficas

[1] Abercrombie, D. *Elements of General Phonetics*. Chicago, USA: Aldine, 1967.

[2] Bejerano, G., and Yona, G. Variations on probabilistic suffix trees: Statistical modeling and prediction of protein families. *Bioinformatics 17*, 1 (2001), 23–43.

[3] Brualdi, R. A. *Introductory Combinatorics*, 5th ed. Prentice Hall, 2009.

[4] Buhlmann, P., and Wyner, A. J. Variable length markov chains. *The Annals of Statistics 27*, 2 (1999), 480–513.

[5] Collet, P., Galves, A., and Leonardi, F. Random perturbations of stochastic processes with unbounded variable length. *Electronic Journal of Probability 13* (2008), 1345–1361.

[6] Cover, T. M., and Thomas, J. A. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, INC., 1991.

[7] Csiszár, I., and Shields, P. C. The consistency of the bic markov order estimator. *The Annals of Statistics 28*, 6 (2000), 1601–1619.

[8] Csiszár, I., and Shields, P. C. Information theory and statistics: A tutorial. *Foundations and Trends^{TM} in Communications and Information Theory 4*, 1 (2004).

[9] Csiszár, I., and Talata, Z. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information Theory 52*, 3 (2006), 1007–1016.

[10] Cuesta, A., Fraiman, R., Galves, A., García, J. E., and Svarc, M. Identifying rhythmic classes of languages using their sonority: a kolmogorov-smirnov approach. *(to appear in Journal of Applied Statistics)* (2007).

[11] Dauer, R. M. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics 11* (1983), 51–62.

[12] Donoho, D., and Huber, P. J. *The Notion of Breakdown Point.* A Festschrift for Erich Lehmann, 1983.

[13] Galves, A., Galves, C., García, J. E., Garcia, N. L., and Leonardi, F. Context tree selection and linguistic rhythm retrieval from written texts. *Can be retrieved from arXiv:0902.3619v2* (2009).

[14] Galves, A., García, J. E., Duarte, D., and Galves, C. Sonority as a basis for rhythmic class discrimination. *Paper presented at Speech Prosody, Aix-en-Provence (can be downloaded from www.lpl.univ-aix.fr/sp2002/pdf/galves-etal.pdf)* (2002).

[15] Galves, A., and Leonardi, F. Exponential inequalities for empirical unbounded context trees. *Progress in Probability 60* (2008), 257–270.

[16] García, J. E., and González-López, V. A. Minimal markov models. *Can be retrieved from arXiv:1002.0729v1* (2010).

[17] García, J. E., and González-López, V. A. Speech signal processes discriminate between rhythmic classes. *In Proceedings. Fifth Brazilian Conference on Statistical Modelling in Insurance and Finance, 2011, Maresias-SP. Brazil.* (2011).

[18] García, J. E., Gut, U., and Galves, A. Vocale-a semi-automatic annotation tool for prosodic research. *Paper presented at Speech Prosody 2002, Aix-en-Provence (can be download from http://aune.lpl.univ-aix.fr/sp2002/pdf/garcia-gut-galves.pdf)* (2002).

[19] James, B. R. *Probabilidade: um Curso em Nível Intermediário*, 2a ed. Projeto Euclides. Rio de Janeiro, Brasil: IMPA, 2002.

[20] Kurtz, D. C. *Foundations of Abstract Mathematics*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., 1992.

[21] Leonardi, F. A generalization of the pst algorithm: Modeling the sparse nature of protein sequences. *Bioinformatics 22*, 11 (2006), 1302–1307.

[22] Leonardi, F. *Cadeias Estocásticas Parcimoniosas com Aplicações à classificação e filogenia das sequências de proteínas*. Tese de doutorado. Universidade de São Paulo, 2007.

[23] Maronna, R., Martin, D., and Yohai, V. *Robust Statistics: Theory and Methods*. John Wiley & Sons, 2006.

[24] Ramus, F., Nespor, M., and Mehler, J. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73 (1999), 265–292.

[25] Rissanen, J. A universal data compression system. *IEEE Transactions on Information Theory 29*, 5 (1983), 656–664.

[26] Rissanen, J. Complexity of strings in the class of markov sources. *IEEE Transactions on Information Theory 32* (1986), 526–532.

[27] Schwarz, G. Estimation the dimension of a model. *The Annals of Statistics 6*, 2 (1978), 461–464.

[28] Van der Vaart, A. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 1998.

[29] Willems, F. M. J., Shtarkov, Y. M., and Tjalkens, T. J. Context-tree maximizing. *Conference on Information Sciences and Systems, Princeton University* (March 2000), TP6–7 – TP6–12.

[30] Zuo, Y. Some quantitative relationships between two types of finite sample breakdown point. *Statistics & Probability Letters 51* (2001), 369–375.