Inferência em Assinaturas de Amostras em Cadeias de Memória de Alcance Variável

Wecsley Otero Prates

Dissertação de Mestrado apresentada ao Departamento de Estatística da Universidade Federal de Minas Gerais para obtenção do título de Mestre

Programa: Pós-Graduação em Estatística
Orientadora: Prof[a]. Dr[a]. Denise Duarte[1]
Co-Orientador: Prof. Dr. Marcos A. da Cunha[1]

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CNPq - Edital MCT CNPq 70/2009 - PGAEST

Belo Horizonte, Setembro de 2011

[1]Universidade Federal de Minas Gerais

Inferência em Assinaturas de Amostras em Cadeias de Memória de Alcance Variável

Este exemplar corresponde a Dissertação de Mestrado

defendida por Wecsley Otero Prates

Banca Examinadora:

- Profª. Drª. Denise Duarte (Orientadora) - Dest-UFMG
- Prof. Dr. Marcos Antonio da Cunha (Co-Orientador) - Dest-UFMG
- Prof. Dr. Enrico Antônio Colosimo - Dest-UFMG
- Prof. Dr. Gustavo Leonel Gilardoni Avalle - Depto Estatística UnB

Agradecimentos

A preocupação, o medo e as incertezas são assuntos completos e estão associados com tantas de nossas ansiedades e problemas que se sobrepõem, com cada um deles tornando-se um fator em relação direta com os outros. Por isso mesmo, vai se tornando impossível separá-los completamente.

Agradeço primeiramente a Deus que me deu força, coragem e persistência para cansar as adversidades que encontrei pelo caminho.

Agradeço a Profª e amiga Denise Duarte pela sua orientação. A sua generosidade e entusiasmo foram um grande incentivo nesses anos.

Agradeço ao Profº e Co-Orientador Marcos Antonio pelo seu apoio e suas idéias, que foram muito importantes para o desenvolvimento deste trabalho. Sou grato ao Profº Enrico Colosimo por me ajudar com seu amplo conhecimento em parte do trabalho. Aos colegas Diego Leal Togni e a Marina Lobato que tiveram partcipações significativas na parte das simulações. Ao colega Rodolfo Lorenzutti por seus conhecimentos computacionais.

Aos meus Professores que dos conhecimentos que foram passados, o mais importante é faríamos muitas coisas se não julgassem tantas coisas impossíveis. Aos amigos que conquistei nessa jornada e que foram importantes para mais essa etapa.

E claro um agradecimento especial aos colegas Gabriel Bahia, Paulo Cerqueira e Rodolfo Lorenzutti, que posso chamá-los de amigos e aprendi de cada um deles uma lições de responsabilidade e companherismo.

E um grande agradecimento á minha família que me deram amor, apoio e me encorajaram a realizar mais um sonho.

Resumo

A análise de um modelo estocástico que descreva, realisticamente, uma situação prática é um grande desafio, em particular porque os fenômenos reais exibem várias dependências. Neste contexto os modelos markovianos desempenham um papel fundamental, uma vez que permitem soluções mais eficientes. Uma cadeia de Markov $\{X_t, t \in \mathbb{Z}\}$ de ordem k assumindo valores em um alfabeto $\mathcal{A}$ finito tem $|\mathcal{A}|^k(|\mathcal{A}| - 1)$ parâmetros a serem estimados. Esse número cresce exponencialmente em k e, portanto, pode tornar-se inviável mesmo para valores não muito grandes de k.

Uma alternativa mais viável do ponto de vista da estimação, é a utilização de Cadeias de Memória de Alcance Variável (VLMC), conhecidas também por Árvores Probabilísticas de Contexto (PCT). O modelo VLMC foi introduzido por Rissanen em 1983. Nesse modelo o tamanho do passado relevante para prever o próximo símbolo muda de uma sequência para outra. Desta forma, o número de parâmetros a serem estimados dimimui muito, uma vez que não precisamos considerar todos os passados de ordem k, mas apenas aqueles relevantes que, em geral, são em número bem menor.

A estimação dos parâmetros do modelo VLMC pode ser feita de maneira consistente através do critério de informação Bayesiano (BIC). O estimador BIC consiste em penalisar a máxima verossimilhança pelo número de parâmetros a serem estimados, estabelecendo um equilíbrio entre a verossimilhança e o número de parâmetros do modelo.

Nesse trabalho foi utilizada a metodologia BIC para estimar as VLMC's baseado no algoritmo proposto no artigo de Csiszar e Talata(2006). Construímos um programa na linguagem R (www.r-project.org) para fazer a estimação das VLMC e utilizamos uma

variante desse algoritmo, proposto em Galves et al (2011), para estimar os valores das constantes de penalização C de cada VLMC candidata dada uma amostra da cadeia. Uma constante ótima de penalização é obtida mudando os valores da constante de penalização e escolhendo aquela que agrega um valor significativo à verossimilhança . Para cada valor de C, temos um valor da penalização da verossimilhança, obtendo assim uma sequência de constantes de penalização $C_n > C_{n-1} > \ldots > C_{opt}$, que chamamos de **Assinatura da Amostra**. E de acordo com a assinatura deixada pela amostra, encontramos um padrão de diferenciação entre textos do Corpus Histórico Tycho Brahe (www.tycho.unicamp.br) através da metodologia das Equações de Estimação Generalizadas (GEE) que vai de encontro à conjectura linguística que diz que houve mudança no ritmo do Português Brasileiro por volta do século 17.

Palavras-chave: Cadeias de Memória de Alcance Variável, Árovre de Contexto, Assinatura da Amostra, Equações de Estimação Generalizada.

Abstract

The analysis of a stochastic model to describe realistically a practical situation is a challenge often insurmountable, especially because the real phenomena exhibit different dependencies. In this context the Markov models play a fundamental role, since they allow more efficient solutions. A Markov chain $\{X_t, t \in \mathbb{Z}\}$ of order k taking values on an alphabet $\mathcal{A}$ finite, has $|\mathcal{A}|^k(|\mathcal{A}| - 1)$ parameter to be estimated. This number grows exponentially in k, and therefore a more viable alternative in terms of estimation, is the use of variable length memory chains (VLMC), also known in literature as Probabilistic Context Tree (PCT), since in this model we have, in general, to estimate fewer parameters.

In this work we introduce the Sample Signature of a Probabilistic Context Tree (PCT) or VLMC, as a way to distinguish samples of discrete random variables coming from different sources. The PCT model is much more interesting than Markov chains of fixed order because it is more parsimonious in the sense that we need fewer parameters to describe it. Moreover, we introduce the Sample Signature of a PCT and show that it can bring more information about the generating source than the model itself. We face in this work the challenge of prosodic patterns detention in the written texts of the Historical Portuguese Corpus Tycho Brahe by using the Sample Signatures of the texts. We also use the Generalized Estimating Equation marginal model as a tool to obtain the results.

Keywords: Variabel Lenght Memory Chains, Sample Signature, Generalized Estimating Equations.

Índice

Lista de Figuras

Lista de Gráficos

Lista de Tabelas

1 *Introdução*

A análise de um modelo estocástico que descreva, realisticamente, uma situação prática é um grande desafio, em particular porque os fenômenos reais exibem várias dependências. Coloca-se então a questão de incorporação de dependências no modelo que o tornem mais adequado à descrição da realidade versus a complexidade daí resultante. Neste contexto os modelos markovianos desempenham um papel fundamental, uma vez que permitem soluções mais eficientes. As cadeias de Markov, apesar de serem uma boa alternativa para modelar muitos problemas não são, em geral, aconselháveis no que se refere à questão prática da estimação. Uma cadeia de Markov $\{X_t, t \in \mathbb{Z}\}$ de ordem k assumindo valores em um alfabeto $\mathcal{A}$ finito tem $|\mathcal{A}|^k(|\mathcal{A}| - 1)$ parâmetro a serem estimados, em que $|\mathcal{A}|$ é a cardinalidade de $\mathcal{A}$. Tem-se que esse número cresce exponencialmente em k e, portanto, para valores não tão grandes de k temos uma quantidade muito grande de parâmetros no modelo. Uma alternativa mais viável, do ponto de vista da estimação, é a utilização de Cadeias de Memória de Alcance Variável (VLMC), conhecidas também por Árvores Probabilísticas de Contexto (PCT).

As cadeias de memória de alcance variável, introduzidas por Rissanem em 1983 são uma classe promissora de modelos que podem auxiliar por exemplo nas áreas da genética, lingüística ou qualquer outra cuja necessidade seja estimar a melhor ordem para uma cadeia com o conjunto de dados disponível, que em contraste com os modelos de cadeia de Markov onde cada variável no tempo t depende de um número fixo de variáveis no passado, os modelos VLMC o tamanho do passado é relevante para prever o próximo símbolo e pode variar com base na realização específica observada.

A idéia é que para cada passado, apenas um sufixo finito do passado (sequência finita

de símbolos), chamada de *contexto* é suficiente para predizer o próximo símbolo. Esses contextos podem ser representados por uma árvore enumerável completa de contextos finitos. Em uma VLMC existe uma probabilidade de transição associada a cada contexto.

2 *Cadeias de Memória de Alcance Variável (VLMC)*

As VLMC´s são cadeias estocásticas de ordem infinita ou finita em um alfabeto finito. A idéia é que para cada passado, apenas um sufixo finito do passado (sequência finita de símbolos), chamada de *contexto*, é suficiente para predizer o próximo símbolo.

As VLMC´s foram introduzidas por Rissanen em 1983 com o nome de Árvores Probabilística de Contexto (Probabilistic Context Trees) para modelar sequências de dados. Ele chamou seu modelo *finitely generated source*. Em seu trabalho ele não apenas introduz o modelo como também propõe um algoritmo para estimar as VLMC´s dada uma amostra. Em seu artigo ele apresenta uma prova da consistência (fraca) do algoritmo no caso de uma VLMC fixa. Mas as Cadeias de Memória de Alcance Variável só se tornaram populares na literatura estatística com o nome de *variable length Markov chains* utilizado por Bühlman e Wyner (1999). Há inclusive um pacote chamado VLMC no software estatístico livre R-Project para fazer estimação das VLMC´s . Eles provaram a consistencia de uma variante do algoritmo de Contexto proposto por Rissanen para cadeias finitas permitindo que a altura da VLMC crescesse com o tamanho da amostra. Apesar de consistente, esse algoritmo demora muito para convergir. Em 2006 Csiszár & Tálata mostraram que um algoritmo baseado no BIC (Bayesian Information Criterion) é eficiente para a estimação de uma VLMC e é com uma versão deste algoritmo, que será apresentada oportunamente, que trabalhamos nesta dissertação.

2.1 *Modelo Probabilístico*

Considera-se um alfabeto $\mathcal{A}$ como sendo qualquer conjunto finito ou infinito formado por símbolos, ou seja, $\mathcal{A}$ pode ser um espaço de estados discreto e finito de um processo estocástico. Um processo estocástico é uma sequência de variáveis aleatórias $\{X_t, t \in \mathbb{Z}\}$, definidas sobre um mesmo espaço de probabilidade $(\Omega, \mathfrak{F}, P)$. Assim, para cada $\omega \in \Omega$ fixo, a função $X_t(\omega)$ na variável t, denotada por $\{X_t(\omega), t \in \mathbb{Z}\}$, é chamada uma realização do processo.$\{X_t\}_{t\in\mathbb{Z}}$ é um processo a tempo discreto se o conjunto de índices $\mathbb{Z}$ for enumerável, e um processo a tempo contínuo, se $\mathbb{Z}$ for não enumerável.

Formalmente definiremos o que vem a ser uma Cadeia de Memória de Alcance Variável, bem como algumas de suas propriedades. Primeramente iremos definir uma Cadeia de Markov de ordem K.

Definição 2.1.1: Cadeia de Markov de Ordem *k: Seja* $\{X_t\}_{t\in\mathbb{Z}}$ *uma cadeia de Markov de ordem k estacionária definida em um alfabeto finito* $\mathcal{A}$*. Denotando* $\left\{\omega \in \Omega \,|X_{-1}(\omega) = x_{-1}, ..., X_{-k}(\omega) = x_{-k}\right\} := x_{-k}^{-1}$ *e deste modo, as probabilidades de transição podem ser escritas da seguinte forma:*

$$\mathbb{P}\left(X_0 = x_0 \,|X_{-1} = x_{-1}, ..., X_{-k} = x_{-k}\right) = p\left(x_0 \left|x_{-k}^{-1}\right.\right), \forall k \in \mathbb{N}.$$

A cadeia será dita estacionária se para todo $t \in \mathbb{Z}$ *tivermos* $\mathbb{P}\{X_t = a\} = \pi(a)$*, onde* $\pi(a)$ *é a medida estacionária da cadeia.*

Para a definição de uma VLMC considere um conjunto finito $\mathcal{A}$ em que $|\mathcal{A}|$ é a cardinalidade de $\mathcal{A}$. Um ramo $s = \omega_m\omega_{m+1}...\omega_n$ (com $\omega_i \in \mathcal{A}, m \leq i \leq n$) é denotado por ω_m^n de tamanho $l(s) = n - m + 1$. O ramo vazio é denotado por $l(\emptyset) = 0$.

A concatenação dos ramos u e v é denotado por uv. Dizemos que um ramo v é um sufixo de uma sequência s, denotado por $s \succeq v$, quando existe um ramo u tal que $s = uv$. Quando $s \neq v$, escreve-se $s \succ v$. Um sufixo de uma sequência semi-infinita

$\omega_{-\infty}^{-1} = ...\omega_{-k}...\omega_{-1}$ é definido da mesma maneira. Onde $\succeq$ significa que o ramo s é maior do que o ramo v.

Propriedade de Sufixo: *Se alguma sequência $\omega_1^k \in \mathcal{T}$ é um sufixo, então para nenhum $\omega_1^k \in \mathcal{T}$ com $k \in \mathbb{N}$ existe $u_1^j \in \mathcal{T}$ com $j < k$ tal que $\omega_i = u_i$ para $i = 1, ..., j$*

Desde modo, temos que;

Definição 2.1.2: *Uma VLMC com finitos galhos será definida como um subconjunto enumerável $\mathcal{T}$ de $\Gamma = \bigcup_{k=0}^{\infty} \mathcal{A}^k$, que satisfaça a propriedade do sufixo.*

O comprimento de uma VLMC será dado por $\mathcal{T} = \max\{|\omega||\omega \in \mathcal{T}\}$, onde $\omega = \omega_1^s : s \in \mathbb{N}$.

Uma árvore $\mathcal{T}$ com número finitos de galhos, será dita completa se $\mathcal{T}$ define uma partição de $\mathcal{A}^{\{1,2,...\}}$. Cada elemento da partição coincide com o conjunto das seqüências em $\mathcal{A}^{\{1,2,...\}}$ tendo ω_1^k como sufixo, para algum $\omega_1^k \in \mathcal{T}$.

O contexto de uma sequência $\omega \in$ a uma Cadeia de Memória de Alcance Variável, é definida como:

Definição 2.1.3: *Considere $\{X_t\}_{t\in\mathbb{Z}}$ uma Cadeia de Memória de Alcance Variável estacionária definidas sobre um mesmo espaço de probabilidade $(\Omega, \mathfrak{F}, P)$ em um alfabeto finito $\mathcal{A} = \{x_0, x_1, ..., x_n\}$. A função;*

$$t : \mathcal{A}^{\infty} \longrightarrow \bigcup_{k=0}^{\infty} \mathcal{A}^k$$

$$x_{-\infty}^{-1} \longrightarrow x_{-l}^{-1} \text{ , em que}$$

$$l = \left|t\left(x_{-\infty}^{-1}\right)\right| := min\left\{k \left| \mathbb{P}(x_0 \left| x_{-\infty}^{-1}\right.) = \mathbb{P}(x_0 \left| x_{-k}^{-1}\right.)\right.\right. \text{ ,para algum } x_0 \in \mathcal{A}, x \in \mathcal{A}^k$$

é denotada função contexto da cadeia, em que x é uma sequência qualquer de símbolos $\in \mathcal{A}^k$.

O comprimento l indica a quantidade de passados relevantes, ou seja, $k < \infty$ o menor inteiro, tal que $\left|t\left(x_{-\infty}^{-1}\right)\right| = l \leq k, \forall x \in \mathcal{A}^{\infty}$.

A forma mais conveniente de representar esta classe de modelos é através da sua árvore de contexto. Uma cadeia $\{X_t\}_{t\in\mathbb{Z}}$ de ordem K com função contexto c, sua árvore é definida com ramos $\left\{s \left| s = t\left(x_{-\infty}^{-1}\right), \forall x \in \mathcal{A}^{\infty}\right.\right\}$

Claramente, uma Cadeia de Memória de Alcance Variável de ordem k é uma cadeia de memória estacionária de ordem k.

Pela exigência de estacionariedade, a distribuição de probabilidade $\mathbb{P}$ de uma cadeia de memória de alcance variável é completamente especificada pelas probabilidades de transição $\mathbb{P}\left\{X_0 = x_0 | t\left(x_{-\infty}^{-1}\right)\right\}$. Desta forma, um modo conveniente de representar estes estados, o espaço de estado minimal, é a representação por árvores (árvores de contexto).

A função contexto pode ser obtida diretamente da VLMC, que é o conjunto de passados relevantes para predição do próximo símbolo, representado como uma árvore. O conjunto $\mathcal{T}$ pode ser identificado por um conjunto de ramos da seguinta forma:

- O primeiro nó é a raiz, que significa o presente;
- Os ramos são os passados relevantes. Quanto mais longe da raiz é o nó, mais passos passados precisam ser observados para predição do próximo símbolo;
- Cada nó tem no máximo $|\mathcal{A}|$ galhos, que é o tamanho do espaço de estados da VLMC;
- O contexto nada mais é que os ramos que ligam o último nó à raiz;
- Cada contexto é representado por um ramo completo;
- O contexto $l = t\left(x_{-\infty}^{-1}\right)$ é representado por um ramo, cujo o sub-ramo do topo é determinado por x_{-1}, o próximo sub-ramo é determinado x_{-2} e assim sucessivamente.

Um exemplo de uma VLMC e contextos usando o espaço de estados igual a 2, onde s=1,2 está descrito a seguir:

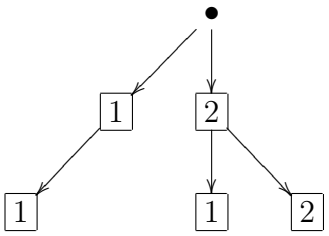


Figura 1 - Exemplo de uma VLMC com espaço binário 1,2

Contextos:

$$l_{-\infty}^{-1} = \begin{cases} 1,1, & \text{se} \quad x_{-1} = 1 \quad e \quad x_{-2} = 1, x_{-\infty}^{-3} = qualquer \\ 2,1, & \text{se} \quad x_{-1} = 2 \quad e \quad x_{-2} = 1, x_{-\infty}^{-3} = qualquer \\ 2,2, & \text{se} \quad x_{-1} = 2 \quad e \quad x_{-2} = 2, x_{-\infty}^{-3} = qualquer \end{cases}$$

Os ramos $s \in \mathcal{T}$ são identificados também com as folhas da árvore, *folha* s é a folha conectada com a raiz pelo caminho s. Similarmente, os nós da árvore $\mathcal{T}$ são identificados com os finitos sub-ramos de todo (finito ou infinito) $s \in \mathcal{T}$, sendo a raiz identificada com o ramo vazio $\emptyset$. Um *filho* de um nó s são aqueles ramos $as, a \in \mathcal{A}$, que são eles próprios os nós, isto é, um sub-ramo de algum $s' \in \mathcal{T}$. Uma VLMC $\mathcal{T}$ é completa se cada nó, exceto as folhas, tem no máximo $|\mathcal{A}|$ filhos.

Cada ramo $s = \omega_1^k \in \mathcal{T}$ é visto como um caminho a partir de uma folha para a raiz (com a raiz no topo), consiste em k simbolos $\omega_1 ... \omega_k$. Uma sequência semi-infinita $\omega_{-\infty}^{-1} \in \mathcal{T}$ é visto como infinitos caminhos para a raiz, (veja Figura 2).

Figura 2 (a)

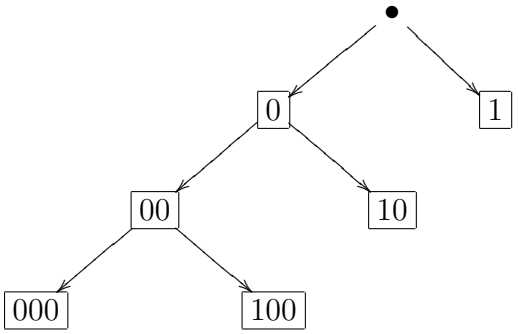


Figura 2 (b)

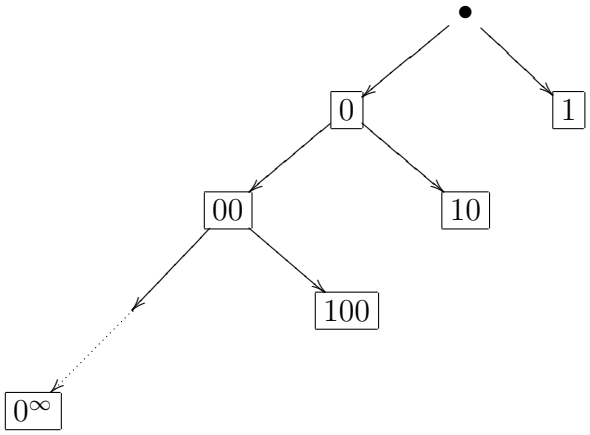


Figura 2 - VLMC de um processo de renovação com espaço binário 0,1, (a) $k_0 = 3$**, (b)** $k_0 = \infty$

Dizemos que uma VLMC é ilimitada se o conjunto $\mathcal{T}$ é enumerável e infinito, portanto teremos que a função contexto é ilimitada.

Oberva-se que uma Cadeia de Memória de Alcance Variável em um alfabeto $\mathcal{A}$ é um par ordenado ($\mathcal{T}$,$\mathbb{P}$) tal que;

- $\mathcal{T}$ *é uma VLMC completa;*
- $\mathbb{P} = \{\mathbb{P}(.|\omega), \omega \in \mathcal{T}\}$ *é uma família de transição de probabilidades em* $\mathcal{A}$.

Mas por uma questão de simplicidade de notação, denotaremos uma VLMC apenas como árvore $\mathcal{T}$, deixando implícita a família de probabilidades de transição.

Definição 2.1.4: *Em uma VLMC $\mathcal{T}$ se nenhuma sequência pertencente a $\mathcal{T}$ pode ser substituída por um sufixo sem violar a propriedade do sufixo, então $\mathcal{T}$ será dita irredutível. Chamamos $\mathcal{I}$ o conjunto das VLMC irredutíveis.*

Quando o comprimento da VLMC $d(\mathcal{T}_0)$=k$_0$ $< \infty$, o processo $\mathbb{P}$ é uma cadeia de memória de alcance variável de ordem k_0. Nesse caso, a VLMC nos leva a uma descrição parcimoniosa do processo, pois a coleção $(|\mathcal{A}|-1)\,|\mathcal{T}_0|$ de probabibilidades de transições são suficientes para descrever o processo, ao invés de $(|\mathcal{A}|-1)\,|\mathcal{A}|^{k_0}$. Observe que a VLMC de um processo independente e identicamente distribuído ($i.i.d$) consiste somente na raiz $\emptyset$, assim $|\mathcal{T}_0| = 1$.

3 *Estimação de uma VLMC via BIC*

O critério de informação Bayesiano (BIC), que apresentamos a seguir, é a principal ferramenta que utilizamos para fornecer estimadores consistentes de VLMC's através da comparação das logverossimilhanças penalizadas de árvores candidatas. Consideraremos árvores com profundidade finita, crescendo com o tamanho da amostra n com ordem $o(logn)$. A literatura fornece alguns algoritmos para estimar VLMC's em tempo $O(n)$, entre eles o VLMC implementado no software R(Buhlman, 1999), e computá-los para todos $i \leq n$ em tempo $o(n\ log\ n)$ (Csiszár & Talata, 2006).

Csiszár & Talata, 2006, provaram que o BIC fornece um estimador consistente da VLMC. Além disso, estes estimadores podem ser implementadas em tempo linear. O conjunto de candidatos das VLMC´s é especificado por um limite no comprimento dos candidatos a contextos, permitindo crescer na ordem $o(logn)$.

Dada uma amostra x_1^n, seja $N_n(s,a)$ o número de ocorrências do ramo $s \in \mathcal{A}^{l(s)}$ seguido pela letra $a \in \mathcal{A}$ na amostra x_1^n. O comprimento máximo de s é $D(n)$, em que $D(n) = (log(n))$- por razões técnicas (Csiszár & Talata, 2006)- somente os símbolos nas posições $i > D(n)$ serão consideradas. Portanto, temos que;

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

E o número dessas ocorrências de s é denotado por $N_n(s)$, onde;

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

Dada uma amostra x_1^n, uma VLMC viável é alguma VLMC de comprimento $d(\mathcal{T}) \leq D(n)$ tais que $N_n(s) \geq 1$ para todo $s \in \mathcal{T}$, e cada ramo s' com $N_n(s') \geq 1$ seja um sub-ramo de algum $s \in \mathcal{T}$ ou tem um sub-ramo $s \in \mathcal{T}$. Uma VLMC viável $\mathcal{T}$ é chamada

de *frequencia* r se $N_n(s) \geq r$ para todo $s \in \mathcal{T}$. A família de todas as possíveis VLMC´s de*frequencia* é denotada por $\mathcal{F}_r(x_r^n, D(n))$.

Temos que;

$$\sum_{a \in A} N_n(s,a) = N_n(s) \text{ e } \sum_{s \in \mathcal{T}} N_n(s) = n - D(n)$$

para alguma VLMC viável. Em relação a VLMC $\mathcal{T}$ como sendo a VLMC de um processo $\mathbb{P}$, a probabilidade da amostra x_1^n pode ser escrita como;

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

para uma VLMC $\mathcal{F}_1(x_1^n, D(n))$, define-se a função de verossimilhança $ML_{\mathcal{T}}(x_1^n)$ como a verossimilhança em $\mathbb{P}(a|s)$ no segundo fator acima, isto é;

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

O critério de informação BIC para estimar $\mathcal{T}_0$ é um critério que atribui uma pontuação para cada modelo candidato (aqui, cadeia de memória de alcance variavel) com base na amostra. O modelo escolhido será aquele cuja pontuação é máxima.

3.1 *Metodologia BIC*

O BIC consiste em penalisar a Máxima Verossimilhança pelo número de parâmetros a serem estimados estabelecendo um equilíbrio entre a verossimilhança e o número de parâmetros do modelo.

Um ponto importante do BIC é o tamanho amostral para estimativa do modelo. Primeiramente deve-se considerar que para testar um modelo de ordem k, por exemplo, as k primeiras observações são necessárias para definição do primeiro estado observado. Deve-se levar em conta que ao aumentar a ordem do modelo que está sendo testado, o número de parâmetros a serem estimados também aumenta. Tais parâmetros nos modelos nas cadeias são na verdade as probabilidades de transição. Ao se tomar uma amostra suficientemente pequena, pode-se obter valores observados pequenos para uma transição, ou até mesmo pode ser observada nenhuma contagem, o que comprometeria numa possível avaliação da verossimilhança.

Aqui, o processo $\mathbb{P}$ com VLMC $\mathcal{T}$ é descrito pelas probabilidades condicionais $\mathbb{P}(a|s)$,$a \in \mathcal{A}$,$s \in \mathcal{T}$, e $(|\mathcal{A}|-1)|\mathcal{T}|$ é o número de parâmetros livres quando a árvore $\mathcal{T}$ é completa. Para um processo com uma VLMC incompleta, as probabilidades de determinados ramos devem ser zero. É sabido (Csiszár e Shields, 2000) que para estimar a ordem de cadeias de Markov, o estimador BIC é consistente, sem qualquer restrição sobre as ordens. No teorema a seguir (Csiszar, I. & Talata, Zs.2006) é necessário um limite na profundidade das candidatas VLMC´s.

Teorema 3.1.1: *No caso de* $d(\mathcal{T}_0)<\infty$, *o estimador BIC é dado por;*

$$\hat{\mathcal{T}}_{BIC}\,(x_1^n) = arg \max_{\mathcal{T}\in\mathcal{F}_1(x_1^n,\,D(n))\cap\mathcal{I}} logML_{\mathcal{T}}(x_1^n) - Cdf(\mathcal{T}) log\ n(x_1^n).$$

com $D(n) = o(log\ n)$, *satisfazendo*

$$\hat{\mathcal{T}}_{BIC}\,(x_1^n) = \mathcal{T}_0$$

quase certamente quando $n \to \infty$.

em que $logML_{\mathcal{T}}(X_1^n)$ é a verossimihança da VLMC $\mathcal{T}$ dada a amostra (X_1^n) e $df(\tau)$ denota o número de graus de liberdade do modelo correspondente da VLMC $\mathcal{T}$. Obtendo-se então os valores do BIC, cria-se uma função que será maximizada na ordem que o BIC apresentar o maior valor.

Geralmente se utiliza $Cdf(\mathcal{T})=\frac{(|\mathcal{A}|-1)|\mathcal{T}|}{2}$, em que $|\mathcal{A}|$ é o tamanho do alfabeto (espaço de possíveis estados da cadeia), Csiszár e Talata utilizam a constante de penalização $\frac{(|\mathcal{A}|-1)|\mathcal{T}|}{2}$. Mas eles mostram que o estimador BIC é consistente para qualquer valor da constante de penalização $Cdf(\mathcal{T})$. Sendo assim, o critério BIC foi usado utilizando vários valores de C para escolher de maneira consistente a VLMC. Utilizamos o programa desenvolvido durante a dissertação chamado Probabilist Context Tree (PCT),que abordaremos nas próximas seções, para encontrar qual é o valor de C ótimo para estimar a melhor VLMC.

3.2 *Computação do Estimador BIC*

Corolário 3.2.1: *O vetor das probabilidades condicionais empríricas é dado por;*

$$\hat{\mathbb{P}}_{\hat{\mathcal{T}}}(a|s) = \frac{N_n(s,a)}{N_n(s)},\ a \in \mathcal{A},\ s \in \hat{\mathcal{T}},$$

que converge para as verdadeiras probabilidades condicionais $\mathbb{P}(a|s)$, $a \in \mathcal{A}$, $s \in \mathcal{T}_0$, *quase certamente quando* $n \to \infty$, *onde* $\hat{\mathcal{T}}$ *é o estimador BIC.*

A máxima verossimilhança $ML_{\mathcal{T}}(x_1^n)$ é dada por;

$$ML_{\mathcal{T}}(x_1^n) = \prod_{s \in \mathcal{T}} \tilde{\mathbb{P}}_{ML,s}(x_1^n), \text{ em que}$$

$$\tilde{\mathbb{P}}_{ML,s}(x_1^n) = \begin{cases} \prod\limits_{a \in \mathcal{A}} \left(\dfrac{N_n(s,a)}{N_n(s)} \right)^{N_n(s,a)}, & \text{se} \quad N_n(s) \geq 1 \\ 1, & \text{se} \quad N_n(s) = 0 \end{cases}$$

O estimador $\hat{\mathcal{T}}_{BIC}\,(x_1^n)$ no ***Teorema 3.1.1*** pode ser representado como:

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

onde $\tilde{\mathbb{P}}_s\,(x_1^n) = n^{-\frac{|\mathcal{A}|-1}{2}} \tilde{\mathbb{P}}_{ML,s}(x_1^n)$, no caso do BIC.

Seja x_1^n uma amostra de $\mathcal{T}$, para cada ramo $s \in S_D$ (o conjunto de todos os ramos com tamanho máximo D) tal que $N_n(s) \geq 1, D = D(n)$. Definindo de forma recurssiva começando pelas folhas da VLMC completa $\mathcal{A}^D$, o valor das variáveis $V_s^D(x_1^n)$ e $\mathcal{X}_{\text{s}}^{\text{D}}(x_1^n)$ como:

$$V_s^D(x_1^n) = \begin{cases} \max \left\{ \tilde{\mathbb{P}}_s(x_1^n), \prod\limits_{a \in \mathcal{A}: N_n(as) \geq 1} V_{as}^D(x_1^n) \right\}, & \text{se} \quad 0 \leq l(s) < D \\ \tilde{\mathbb{P}}_s(x_1^n), & \text{se} \qquad l(s) = D \end{cases}$$

E o indicador

$$\mathcal{X}_{\text{s}}^{\text{D}}(x_1^n) = \begin{cases} 1, & \text{se} \quad 0 \leq l(s) < D, e \prod\limits_{a \in \mathcal{A}: N_n(as) \geq 1} V_{as}^D(x_1^n) > \tilde{\mathbb{P}}_s(x_1^n) \\ 0, & \text{se} \quad 0 \leq l(s) < D, e \prod\limits_{a \in \mathcal{A}: N_n(as) \geq 1} V_{as}^D(x_1^n) \leq \tilde{\mathbb{P}}_s(x_1^n) \\ 0, & \text{se} \qquad\qquad\qquad\qquad\qquad\qquad l(s) = D. \end{cases} \cdot$$

Usando esses indicadores, para cada $s \in S_D, D = D(n)$ a estimação da árvore $\mathcal{T}_{\text{s}}^{\text{D}}(x_1^n)$ consiste dos ramos $u \succeq s$, Tais que;

$\mathcal{T}_{\text{s}}^{\text{D}}(x_1^n)$ igual a;

$$\{u \in S_D : \mathcal{X}_u^{\text{D}}(x_1^n) = 0, \mathcal{X}_v^{\text{D}}(x_1^n) = 1, \forall s \preceq v \preceq u, \text{ se } \mathcal{X}_{\text{s}}^{\text{D}}(x_1^n) = 1$$

$$\text{e igual a } \{s\} \text{ se } \mathcal{X}_{\text{s}}^{\text{D}}(x_1^n) = 0\,\}$$

A VLMC estimada $\tilde{\mathcal{T}}(x_1^n)$ *é igual a estimação da VLMC começando da raiz, ou seja,* $\tilde{\mathcal{T}}(x_1^n) = \mathcal{T}_{\phi}^{\text{D}}(x_1^n)$ ***(Ver demonstração em Csiszar, I. & Talata, Zs. 2006)***.

EXEMPLO: Temos que;

$$\tilde{P}_s(x_1^n) = n^{-\frac{|\mathcal{A}|-1}{2}} \tilde{P}_{ML,s}(x_1^n)$$

em que;

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

Considere $\mathcal{A} = \{1, 2, 3\}$ o espaço de estados, e seja uma cadeia máxima $\mathcal{T}_s^{\mathcal{D}}$ hipotética do tipo;

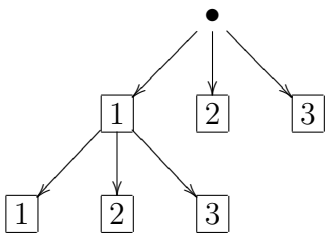


Figura 3: Exemplo de uma VLMC $\mathcal{T}_s^{\mathcal{D}}$ hipotética onde o espaço de estados é $A = \{1, 2, 3\}$

Para $s = \emptyset$ e $l(\emptyset) = 0$, temos que;

$$\tilde{P}_{s=\emptyset, l(\emptyset)=0}\left(x_1^n\right) = n^{-\frac{|\mathcal{A}|-1}{2}}$$

Para $s = \{1, 2, 3\}$ com $l(s) = 1$, temos que;

$$\tilde{P}_{s=1,l(s)=1}\left(x_1^n\right) = n^{-\frac{|\mathcal{A}|-1}{2}} . \left[\frac{N_n(1,1)}{N_n(1)}\right]^{N_n(1,1)} . \left[\frac{N_n(1,2)}{N_n(1)}\right]^{N_n(1,2)} . \left[\frac{N_n(1,3)}{N_n(1)}\right]^{N_n(1,3)}$$

$$\tilde{P}_{s=2,l(s)=1}\left(x_1^n\right) = n^{-\frac{|\mathcal{A}|-1}{2}} . \left[\frac{N_n(2,1)}{N_n(2)}\right]^{N_n(2,1)} . \left[\frac{N_n(2,2)}{N_n(2)}\right]^{N_n(2,2)} . \left[\frac{N_n(2,3)}{N_n(2)}\right]^{N_n(2,3)}$$

$$\tilde{P}_{s=3,l(s)=1}\left(x_1^n\right) = n^{-\frac{|\mathcal{A}|-1}{2}} . \left[\frac{N_n(3,1)}{N_n(3)}\right]^{N_n(3,1)} . \left[\frac{N_n(3,2)}{N_n(3)}\right]^{N_n(3,2)} . \left[\frac{N_n(3,3)}{N_n(3)}\right]^{N_n(3,3)}$$

Portanto, a decisão do indicador $\mathcal{X}_s^D(x_1^n)$ é dada por:

$$\mathcal{X}_{s=\emptyset}^D(x_1^n) =$$
$$\begin{cases} 1, & \text{se} \quad l(s) = 0, \text{e} \tilde{P}_{s=1,l(s)=1}\left(x_1^n\right) . \tilde{P}_{s=2,l(s)=1}\left(x_1^n\right) . \tilde{P}_{s=3,l(s)=1}\left(x_1^n\right) > n^{-\frac{|\mathcal{A}|-1}{2}} \\ 0, & \text{se} \quad l(s) = 0, \text{e} \tilde{P}_{s=1,l(s)=1}\left(x_1^n\right) . \tilde{P}_{s=2,l(s)=1}\left(x_1^n\right) . \tilde{P}_{s=3,l(s)=1}\left(x_1^n\right) < n^{-\frac{|\mathcal{A}|-1}{2}} \end{cases} .$$

Para $s = \{11, 21, 31\}$ com $l(s) = 2$, temos que;

$$\tilde{P}_{s=11,l(s)=2}\left(x_1^n\right) = n^{-\frac{|\mathcal{A}|-1}{2}} . \left[\frac{N_n(11,1)}{N_n(11)}\right]^{N_n(11,1)} . \left[\frac{N_n(11,2)}{N_n(11)}\right]^{N_n(11,2)} . \left[\frac{N_n(11,3)}{N_n(11)}\right]^{N_n(11,3)}$$

$$\tilde{P}_{s=21,l(s)=2}\left(x_1^n\right) = n^{-\frac{|\mathcal{A}|-1}{2}} . \left[\frac{N_n(21,1)}{N_n(21)}\right]^{N_n(21,1)} . \left[\frac{N_n(21,2)}{N_n(21)}\right]^{N_n(21,2)} . \left[\frac{N_n(21,3)}{N_n(21)}\right]^{N_n(21,3)}$$

$$\tilde{P}_{s=31,l(s)=2}\left(x_1^n\right) = n^{-\frac{|\mathcal{A}|-1}{2}} . \left[\frac{N_n(31,1)}{N_n(31)}\right]^{N_n(31,1)} . \left[\frac{N_n(31,2)}{N_n(31)}\right]^{N_n(31,2)} . \left[\frac{N_n(31,3)}{N_n(31)}\right]^{N_n(31,3)}$$

Portanto, a decisão do indicador $\mathcal{X}_s^D(x_1^n)$ é dada por:

$$\mathcal{X}_{s=\emptyset}^D(x_1^n) =$$
$$\begin{cases} 1, & \text{se} \quad l(s) > 0, \text{e} \tilde{P}_{s=11,l(s)=2}\left(x_1^n\right) . \tilde{P}_{s=21,l(s)=2}\left(x_1^n\right) . \tilde{P}_{s=31,l(s)=2}\left(x_1^n\right) > \tilde{P}_{s=1,l(s)=1}\left(x_1^n\right) \\ 0, & \text{se} \quad l(s) > 0, \text{e} \tilde{P}_{s=11,l(s)=2}\left(x_1^n\right) . \tilde{P}_{s=21,l(s)=2}\left(x_1^n\right) . \tilde{P}_{s=31,l(s)=2}\left(x_1^n\right) < \tilde{P}_{s=1,l(s)=1}\left(x_1^n\right) \end{cases} .$$

E temos que o indicador $\mathcal{X}_s^D(x_1^n) = 0$ se $l(s) = D$ em que $D = D(n)$.

4 *Assinatura de uma Amostra*

Em Duarte, Prates e Colosimo (2011) é proposta uma ferramenta para caracterizar amostras de VLMC e compará-las de maneira bastante eficaz chamada Assinatura da Amostra. Esta ferramenta é uma sequência de constantes de penalização do BIC que é função da amostra, de maneira que amostras da mesma VLMC, as sequências de constantes de penalização não se diferem muito uma das outras e que em amostras provindas de VLMC´s diferentes, as sequências de constantes de penalização sejam diferentes. A base para a construção da Assinatura de uma Amostra é dada pelo **Teorema 4.1**, apresentado em Galves et al (2011)que mostra que há uma mudança de regime da verossimilhança quando passamos de um modelo que subestima para outro que superestima a árvore verdadeira. Desta maneira esse resultado nos mostra que existe uma constante ótima C, associada à árvore verdadeira que pode ser encontrada através da mudança de comportamento da verossimilhança. Ele nos sugere, portanto, um algoritmo de escolha desta constante ótima. Para chegar até esta constante ótima esse algoritmo varre uma sequência de possíveis candidatas e é a sequência que chamamos de Assinatura da Amostra.

Considere a seguinte função:

$$c \in [0, +\infty) \longmapsto \hat{\tau}_{BIC}(X_1^n, c) \in T_n$$

Denote por C_n a classe das árvores campeãs, isto é;

$$\mathcal{C}_n = \{\tau_n^c = \hat{\tau}_{BIC}\,(X_1^n,c) : c \in[0,+\infty)\,\}$$

Defina a classe $\mathcal{C}$ de todas as árvores campeãs para uma amostra infinita como;

$$\mathcal{C}=\bigcup_{n\geq 1} \mathrm{C}_n$$

Teorema 4.1: *Assuma $X_1, ..., X_n$ como uma amostra de um processo estocástico ergódico $(\mathcal{T}^*, \mathrm{p}^*)$ com $\mathcal{T}^*$ finito. Então, seguem os seguintes resultados quase certamente quando, $n \to \infty$.*

1. *Para qualquer $\mathcal{T} \in C_n$ com $\mathcal{T} \prec \mathcal{T}^*$, existe uma constante $C(\mathcal{T}^*,\mathcal{T}) > 0$ tal que,*

$$\log L_{\mathcal{T}^*}(X_1^n) - \log L_{\mathcal{T}}(X_1^n) \geq C(\mathcal{T}^*,\mathcal{T})\mathrm{n}.$$

2. *Para algum $\mathcal{T} \prec \mathcal{T}' \in C_n$ com $\mathcal{T}^* \preceq \mathcal{T}$, existe uma constante $C(\mathcal{T}, \mathcal{T}') > 0$ tal que,*

$$\log L_{\mathcal{T}'}(X_1^n) - \log L_{\mathcal{T}}(X_1^n) \leq C(\mathcal{T}, \mathcal{T}')\log \mathrm{n}.$$

4.1 *Implementando o PCT*

A fim de implementar o PCT, primeiro é necessário um algoritmo para calcular o estimador BIC $\hat{\tau}_{BIC}(X_1^n, c)$ para qualquer constante $c > 0$. Isto pode ser feito de forma eficiente por meio do algoritmo Context Tree Weighted introduzido por Willems et al. (1995) e adaptado para o caso BIC por Csiszár e Talata (2006), o PCT foi implementado no software R durante o desenvolvimento dessa dissertação, sendo um dos objetivos alcançados desse projeto. No Apêndice 1 segue um exemplo do programa PCT.

Usando esse algorítmo, podemos computar o conjunto das árvores campeãs $\mathcal{C}$ realizando as seguintes etapas (Galves et al, 2011).

Procedimento de cálculo das Árvores campeãs:

1. Fixe $i = 0, l = 0$ e $u > 0$ grande o suficiente, tal que $\hat{\tau}_{BIC}(X_1^n, u)$ é a raiz da árvore.
2. Calcule $\tau_l = \hat{\tau}_{BIC}(X_1^n, u)$, defina $\tau_0 = \tau_l$ e $\tau_u =< raiz >$.
3. Enquanto $(\tau_l \neq < raiz >)$
   (a) Enquanto $(|u - l|) > \epsilon)$
      i. Enquanto $(\tau_l \neq \tau_u)\,\{a = u$ e $u = (l + u/2.\}$
      ii. $l = u$ e $u = a$.
   (b) $i = i + 1$.
   (c) $\tau_i = \tau_u$.

4.2 *Identificando a Assinatura de uma amostra*

O algoritmo PCT escolhe uma árvore campeã entre as canditadas, dada uma amostra usando o critério BIC. Mas também gera uma seqüência de árvores candidatas (campeãs) para cada valor da constante de penalização. Esta seqüência pode ser vista como uma evolução da amostra de acordo com o "preço"que se está disposto a pagar a fim de obter um modelo mais simples para os dados.

Dada uma amostra $X_1, X_2, ..., X_n$, a árvore campeã estimada, $\hat{\mathcal{T}}$ é uma função da amostra e da constante de penalização ótima C_{opt},

$$\hat{\mathcal{T}} = f(X_1^n, C_{opt})$$

O algoritmo gera uma sequência de constantes de penalização.

$$C_n > C_{n-1} > \ldots > C_{opt},$$

em que C_n é a constante de penalização da árvore com apenas a raiz (modelo independente). Logo, nos leva a uma sequência de árvores

$$\mathcal{T}_\phi = \mathcal{T}_n \prec \mathcal{T}_{n-1} \prec \mathcal{T}_{n-2} \prec \ldots \prec \mathcal{T}_{opt}$$

Chamamos a sequência $C_x = (C_n, C_{n-1}, C_{n-2}, \ldots, C_{opt})$ de **Assinatura da Amostra**.

Dada uma amostra $X_1, X_2, ..., X_n$, C_x é uma sequência de constantes de penalização que caracteriza esta amostra no seguinte sentido:

- Em amostras da mesma VLMC, as sequências de constantes de penalização não diferem muito umas das outras;
- Em amostras de duas VLMC´s diferentes as sequências de constantes de penalização são diferentes.

A seguir apresentaremos alguns resultados de simulação que mostram que a Assinatura de fato caracteriza uma amostra.

4.3 *Simulações de Assinaturas da Amostra*

Dadas duas amostras X_1^n e Y_1^n provenientes da mesma VLMC, podemos estimar o melhor modelo de acordo com os dados usando um algoritmo apropriado. O que nós queremos fazer é decidir se as duas amostras provêm da mesma cadeia ou não. Nas situações onde a VLMC estimada para cada amostra são muito diferentes, por exemplo, têm ordem diferente ou os galhos aparecem em lugares diferentes, podemos concluir que as amostras não foram retiradas da mesma cadeia, pois o algoritmo é consistente.

O problema aparece quando as VLMC´s estimadas são muito semelhantes e não somos capazes de decidir se eles são diferentes com base apenas na árvore final estimada. Neste caso, verifica-se que a assinatura da amostra pode ser muito útil. Nós apresentamos aqui algumas simulações que comprovam este fato.

O objetivo das simulações foi mostrar que as assinaturas podem ser usadas para identificar uma VLMC. Para atingir este objetivo, pensamos em algumas possibilidades, onde poderiamos cometer um erro. Se nas situações que iremos apresentar em seguida a assinatura da amostra é capaz de diferenciar as VLMC´s, nós acreditamos que isto é o suficiente para aceitar que a assinatura da amostra é capaz de diferenciar as amostras provenientes de diferentes VLMC´s.

Simulação 1:

Nesta simulação, tomaram-se duas VLMC´s (que nós chamamos A_1 e A_2) com a mesma estrutura de galhos e probabilidades de transição semelhantes e comparamos a assinatura da amostra delas. Como já observado, o que tínhamos em mente quando nós escolhemos essas duas VLMC´s semelhantes foi ver o que acontece em uma situação

extrema. Queríamos testar se a assinatura da amostra poderia erroneamente indicar que elas pertencem à mesma VLMC, quando, de fato, elas não são.

A VLMC é dada pela estrutura de contextos $\mathcal{T}$ e as probabilidades de transição da seguinte forma:

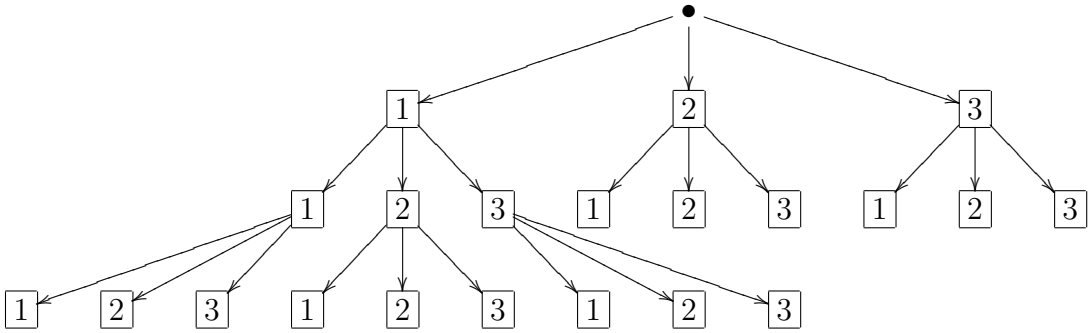


Figura 4 - Configuração das VLMC's A_1 e A_2, onde o espaço de estados é $A = \{1, 2, 3\}$

As probabilidades de transição da VLMC A_1, cuja configuração é mostrada na Figura 1, são dadas como se segue na Tabela 1. A primeira coluna representa o contexto da VLMC.

Tabela 1: Probabilidades de Transição da VLMC A_1

w	$p(1/w)$	$p(2/w)$	$p(3/w)$
111	0.20	0.24	0.56
211	0.19	0.80	0.01
311	0.10	0.40	0.50
121	0.19	0.39	0.42
221	0.49	0.47	0.04
321	0.48	0.48	0.04
131	0.38	0.38	0.24
231	0.10	0.15	0.75
331	0.50	0.24	0.26
12	0.09	0.09	0.82
22	0.32	0.32	0.36
32	0.40	0.40	0.20
13	0.12	0.12	0.76
23	0.27	0.27	0.46
33	0.49	0.49	0.02

Exemplo: No caso quando o contexto é de 111, a probabilidade de o próximo símbolo a ser 1 é de $0,2$, de ser 2 é $0,24$ a de ser 3 é $0,56$. E as probabilidades de transição da VLMC A_2 são:

Tabela 2: Probabilidades de Transição da VLMC A_2

w	$p(1/w)$	$p(2/w)$	$p(3/w)$
111	0.25	0.18	0.57
211	0.20	0.79	0.01
311	0.11	0.40	0.49
121	0.18	0.39	0.43
221	0.40	0.47	0.13
321	0.48	0.38	0.14
131	0.30	0.4	0.20
231	0.15	0.10	0.75
331	0.50	0.25	0.25
12	0.13	0.15	0.72
22	0.32	0.32	0.36
32	0.35	0.45	0.20
13	0.12	0.10	0.78
23	0.25	0.25	0.50
33	0.50	0.35	0.15

Foram reealizadas 500 simulações de Monte Carlo das VLMC´s A_1 e A_2 e encontrada a assinatura de cada amostra. Os resultados podem ser visualizados no Gráfico 1. Foi utilizado o algoritmo PCT para encontrar a seqüência de VLMC´s campeãs. Tem-se que os pontos nos Gráficos 1 e 2 representam um intervalo de 90% de confiança percentílico, depois de ordenados os valores das constantes de penalização de cada uma das 500 amostras monte carlo. Primeiro, observamos no Gráfico 1 que para ambas as VLMC´s, o algoritmo encontra a verdadeira VLMC com 15 galhos com base nos dados, mas não poderíamos dizer que eles foram gerados por diferentes fontes, pois além de ter a mesma estrutura de galhos as probabilidades de transição estimadas são praticamente as mesmas (p-valor $<$ 0,05) no Teste de razão de verossimilhança.

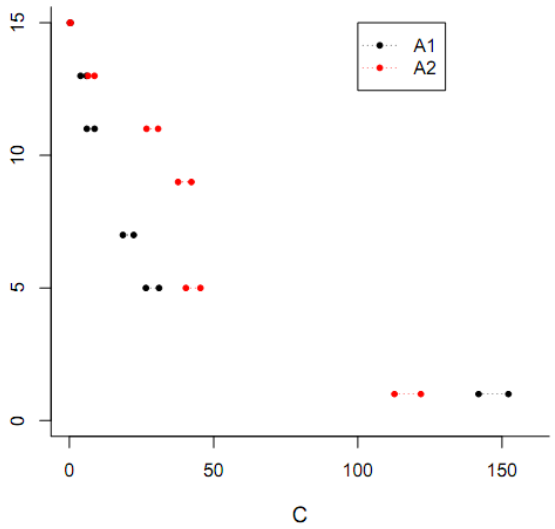
Mas é muito interessante observar que as assinaturas das amostras para cada VLMC são completamente diferentes. O valor da constante de penalização que é necessário escolher um modelo independente, a raiz, como a melhor é muito maior para a VLMC A_1 do que para a VLMC A_2. O oposto acontece para a VLMC com 5 galhos. A VLMC com 7 galhos não foi possível para a cadeia A_2 e a VLMC com 9 ramos, não foi possível para a VLMC A_1. A VLMC com 11 galhos é possível para ambos, mas é mais cara para A_1.

Assim, concluímos que, apesar da semelhança das VLMC´s A_1 e A_2 a assinatura da amostra é capaz de captar as diferenças entre eles. Observamos também que para a mesma VLMC, por exemplo para A_1 os valores das constantes de penalização não mudam significativamente de uma simulação para outra.

Simulação 2:

Foram geradas amostras de duas VLMC´s simétricas (chamadas de B_1 e B_2) com iguais probabilidades de transição, onde o alfabeto é $A = \{1, 2\}$. Os contextos da VLMC

Gráfico 1 - Simulação das Assinaturas das Amostras para as cadeias A1 e A2



B_1 são dados na Figura 5.

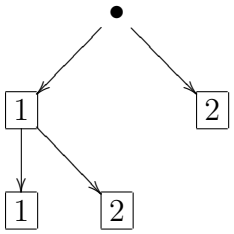


Figura 5: Configuração da VLMC B_1 onde o espaço de estados é $A = \{1, 2\}$

Tabela 3: Probabilidades de Transição da VLMC B_1

w	$p(1/w)$	$p(2/w)$
11	0.50	0.50
21	0.30	0.70
2	0.70	0.30

Os contextos da VLMC B_2 são dados na Figura 6:

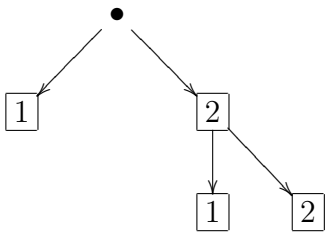


Figura 6: Configuração da VLMC B_2 onde o espaço de estados é $A = \{1, 2\}$

Tabela 4: Probabilidades de Transição da VLMC B_2

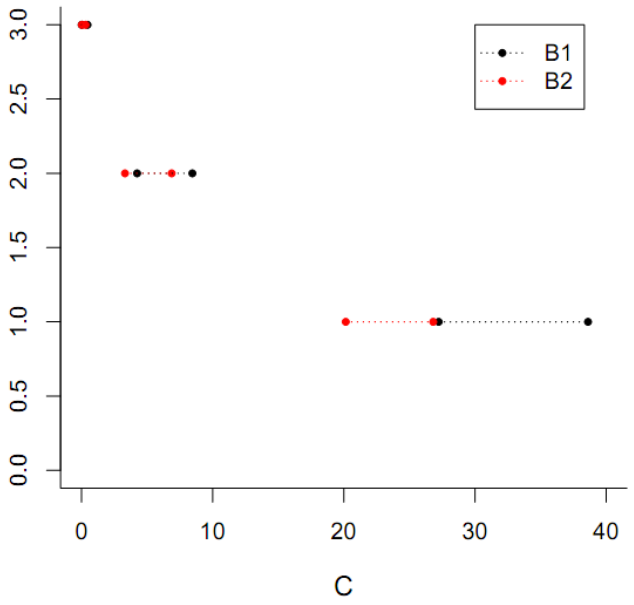
w	$p(1/w)$	$p(2/w)$
22	0.50	0.50
12	0.30	0.70
1	0.70	0.30

Também foram realizadas 500 simulações de Monte Carlo das cadeias B_1 e B_2 e encontradas as assinaturas de cada amostra usando o algoritmo PCT. Novamente, temos que as assinaturas das amostras de cada VLMC são completamente diferentes.

Observe no Gráfico 2 que a principal diferença ocorre quando o modelo independente (raiz) é considerado. Os valores da constante de penalização são maiores para a VLMC B_1 do que para a VLMC B_2. Isso acontece porque a verossimilhança da amostra gerada por B_1 com o modelo independente é muito maior do que a verossimilhança deste modelo usando a amostra gerada por VLMC B_2.

Portanto para a mesma cadeia as assinaturas da amostra são similares e para diferentes cadeias as assinaturas se diferem, de onde concluímos que a Assinatura de uma Amostra é capaz de identificá-las.

Gráfico 2: Simulação das Assinaturas das Amostras para as cadeias B1 e B2



5 *Modelagem para Medidas Repetidas*

As assinaturas de amostras podem ser tratadas como medidas repetidas no tempo. A constante de penalização é uma resposta contínua de uma medida repetida de diferentes números de galhos da cadeia e as respostas de diferentes sequências são independentes entre si, mas as respostas para a mesma sequência são correlacionadas. Especialmente para a resposta contínua, um grande número de abordagens para a análise de dados longitudinais têm sido propostos e desenvolvidos na literatura (Verbeke and Molenberghs, 2000; Diggle et al., 2002; Fitzmaurice et al., 2004).

Nos últimos 10 anos, houve um interesse crescente em estudos de medidas repetidas e na análise estatística de dados longitudinais. Esses estudos são definidos como estudos no qual a variável resposta é medida repetidamente, isto é, a variável resposta é medida

no mesmo indivíduo em várias ocasiões. Em estudos longitudinais, nos quais a variável resposta é contínua, o método de análise comumente usado envolve modelos lineares com erros correlacionados.

Técnicas estatísticas que assumem observações independentes como análise de regressão linear e análise de regressão logística, não podem ser usadas diretamente em estudos com medidas repetidas. Para análise de dados em estudos longitudinais são desenvolvidas técnicas estatísticas especiais, que levam em conta que as observações repetidas de cada indivíduo sejam correlacionadas.

5.1 *Equações de Estimação Generalizadas (GEE)*

Em dados longitudinais, um fator relevante é a estrutura de correlação existente no vetor de respostas repetidas e uma proposta que tem recebido considerável atenção nos últimos anos, já com diversas extensões, é das equações de estimação generalizadas (GEE) proposta por Liang e Zeger (1986). Esta metodologia é uma extensão multivariada da quase-verossimilhança, onde somente a relação entre a função do valor esperado da resposta e o preditor linear (função de ligação) e a relação entre a variância e a média, além do parâmetro de escala, necessitam ser especificadas.

A estratégia das GEE baseia-se em um modelo marginal para estimar os parâmetros de regressão considerando a correlação entre os indivíduos sem especificar a função de probabilidade conjunta. Esta técnica produz estimativas consistentes e assintoticamente normais sob a especificação correta da função de ligação. A questão da dependência proveniente das medidas repetidas da variável resposta é levada em consideração através da especificação de uma "estrutura de correlação de trabalho" $R_i(\alpha)$ denominado por Liang e Zeger (1986), nas equações quase-escore. Diferentes estruturas de correlação podem ser consideradas, sendo já implementadas nos principais pacotes estatísticos como R, SAS, STATA. Como exemplo, podemos mencionar a estrutura de correlação simétrica, onde as correlações entre as medidas subseqüentes são assumidas iguais, independente do comprimento do interperíodo.

O GEE pode ser pensado como uma extensão dos Modelos Lineares Generalizados para dados correlacionados. Em GEE, é necessário apenas especificar um modelo para a estrutura da média populacional e podem ser definidas para os modelos da família GLM. Uma forma de expressar a matriz variância-covariância de trabalho V_i de y_i, decomposta em termos de $R_i(\alpha)$ é apresentada abaixo:

$$V_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} \phi,$$

onde A_i é a matriz diagonal n_i x n_i e R_i é uma matriz simétrica positiva definida, que depende de um parâmetro ou vetor de parâmetros α para cada $y_i = (y_{i1}, y_{i2}, ..., y_{it})$ cujo objetivo é incorporar a correlação entre as medidas repetidas para a mesma unidade de análise.

Um procedimento fundamental das GEE é estimar β quando a matriz de variância-covariância de y_i, V_i é especificada corretamente. Liang e Zeger (1986) parametrizaram V_i em termos de um parâmetro desconhecido,α, e utilizou estimadores de momento para estimação do mesmo. A função de estimação GEE (função quase-escore) é dada por:

$$S(\beta) = \sum_{i=1}^{M} \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} (y_i - \mu_i) = \sum_{i=1}^{n} \boldsymbol{X}'_i \boldsymbol{R}_i(\boldsymbol{\alpha})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0,$$

onde $\mathbf{y}_i$ é o vetor resposta do i-ésimo indivíduo de dimensão n_i; μ_i, uma função de $\boldsymbol{\beta}$, é a correspondente a média, $\boldsymbol{\beta}$ é o vetor p-dimensional de parâmetros desconhecidos, $\frac{\partial \mu_i}{\partial \beta}$ é a derivada parcial de μ_i com respeito a β. Uma estimativa de β, $\hat{\beta}_{GEE}$, é obtida por iteração entre soluções para equação $S(\beta)$ e estimadores $\hat{\beta}$ consistente e assintoticamente normal como descrito por Liang e Zeger (1986).

A solução de $S(\beta)$ é dada por:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}'_i \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}'_i \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{y}_i\right),$$

uma generalização do estimador dos mínimos quadrados ponderada pela matriz $\boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})$, onde α é estimado pelos resíduos do modelo independente (Diggle et al., 2002).

A variância de $\widehat{\beta}$ pode ser obtida pelo estimador sanduiche:

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{M}_0^{-1} \boldsymbol{M}_1 \boldsymbol{M}_0^{-1},$$

onde

$$\begin{aligned} \boldsymbol{M}_0 &= \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{X}_i, \\ \boldsymbol{M}_1 &= \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)' \boldsymbol{R}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{X}_i. \end{aligned}$$

A $\widehat{Var}(\hat{\boldsymbol{\beta}})$ fornece estimadores consistentes para $Var(\hat{\boldsymbol{\beta}})$ mesmo quando $\boldsymbol{R}_i(\boldsymbol{\alpha})$ não é especificada corretamente.

O procedimento das GEE permite que a dependência do tempo possa ser especificada de diferentes maneiras. Como a matriz de correlação de trabalho de R é uma função de $\alpha, R(\alpha)$, diferentes estruturas podem ser consideradas para o mesmo, como por exemplo:

Tabela 5: Tipos de Matriz de Correlação de Trabalho

Estrutura de Correlação	Parâmetros
Independente	Nenhum Parâmetro a estimar
Permutável	α é um Escalar
Auto Regressiva	α é um Vetor
Estacionária	α é um Vetor
Não-Estacionária	α é uma Matriz
Não-Estacionária ou Não-Especificada	α é uma Matriz

Essas estruturas de correlação produzem estimadores mais eficientes quando a especificação é correta. Entretanto, pode acontecer que tal estrutura resulte em uma matriz não-positiva definida quando existe dados faltantes e/ou uma variação no número de observações por indivíduo.

Dentre várias técnicas para medidas repetidas, neste trabalho decidimos em favor do modelo marginal GEE devido à sua robustez e por conta da interpretação das quantidades médias da população, pois as mesmas são desconhecidas. A média populacional é a única estrutura necessária para o ajuste do modelo GEE.

6 *Aplicação*

Na história da língua Portuguesa tem sido relatado pelos gramáticos que houve uma mudança prosódica entre os séculos 17 e 18. Teyssier (1980) datou essa mudança na segunda metade do século 18 e Castro (2006), um século antes. Mais recentemente, Frota, Galves, Gonzalez-Lopez e Abaurre (2010) apresentaram evidências empíricas da mudança prosódica em Português, estudando o ritmo de textos escritos. Mas a leitura da fonologia de uma língua a partir de textos escritos é um desafio. Provas Fonéticas e fonologicas dessa mudança prosódica são ainda escassas e controversas. Portanto, mais estudos neste assunto são necessários.

Nós apresentamos aqui uma abordagem baseada no exemplo de assinatura de uma VLMC que nos dá mais evidências de que a mudança prosódica no Português Brasileiro ocorreu em torno de 1675 em textos do corpus Tycho Brahe.

Para atingir tais objetivos, articular-se análise qualitativa com análise quantitativa, reunindo de um lado a teoria da gramática gerativa e a fonologia prosódica e, de outro lado, a estatística descritiva e modelagem estocástica. Essa abordagem multidisciplinar une este trabalho às tendências mais recentes no campo dos estudos da mudança lingüística, aliando o recurso a grandes volumes de dados - possíveis graças à contribuição das tecnologias computacionais - a investigações teóricas solidamente fundadas, reunindo e fomentando a contribuição entre pesquisadores de formação variada - sintaticistas, fonólogos, estatísticos, probabilistas, teóricos da computação e lingüistas computacionais - contribuindo, por fim, na renovação das perspectivas para o campo.

O Corpus Histórico do Português Tycho Brahe é um corpus eletrônico anotado, com-

posto de textos em português escritos por autores nascidos entre 1435 e 1845 e está disponível para pesquisadores, em http://www.tycho.iel.unicamp.br. Definindo um alfabeto finito que codifique o ritmo das palavras dos textos podemos realizar a modelagem da VLMC associada ao padrão rítmico do texto. Assim se torna possível comparar os textos de diversos autores e em diversas épocas de acordo com seu padrão rítmico.

Dezesseis textos do Corpus Histórico do Português Tycho Brahe foram modernizados, ou seja, foram alterados para que fosse possível estabelecer um critério ortográfico e gramatical que permitisse codificar as palavras de acordo com a tonalidade de suas sílabas. O programa Sílaba2008, desenvolvido em Perl por Miguel Galves, etiqueta palavras segundo o número de sílabas e posição da sílaba tônica, retornando o texto codificado de acordo com o significado.

Antes de rodar o Sílaba2008 foi necessário executar nos textos outro programa desenvolvido em Perl, o Limpad, desenvolvido por Jesus Garcia, que retira os números e os colchetes dos textos modernizados, que aparecem antes de cada linha com o código do texto e a linha em questão. Por exemplo, antes das linhas do texto Nova Floresta de Manuel Bernardes (1675) aparece $[b\_003\_s\_1]$ simbolizando que é a linha $1(\_s\_1)$ do texto (b_003). A lista dos textos analisados e apresentados estão na tabela 6.

Tabela 6:Descrição dos textos do Corpus Histórico do Português Tycho Brahe

Code	Year	Author	Title
G008	1502	Pedro Magalhães de Gandavo	História da Província de Santa Cruz
P001	1510	Fernão Mendes Pinto	Perigrinação
S001	1556	Luis de Sousa	A vida de Frei Bartolameu dos Mártires
V002	1608	Antônio Vieira	Cartas
V004	1608	Antônio Vieira	História do Futuro
C003	1631	Antônio das Chagas	Cartas Espirituais
B003	1675	Manuel Bernardes	Nova Floresta
C001	1702	Cavaleiro de Oliveira Fco Xavier	Cartas
A001	1705	Matias Aires	Reflexões sobre a Vaidade dos Homens
C004	1714	Antônio da Costa	Cartas, Antônio da Costa
A004	1750	Marquesa de Alorna	Cartas
G003	1799	J. B. da Silva L. de Almeida Garret	Cartas, Almeida Garret
G004	1799	J. B. da Silva L. de Almeida Garret	Teatro, Almeida Garret
A003	1802	Marquês da Fronteira e d'Alorna	Memórias do Marquês da Fronteira e d'Alorna
B005	1826	Camilo Castelo Branco	Maria Moisés
O001	1836	Ramalho Ortigão	Cartas a Emília

Cada sílaba recebeu um código:

- 0. Sílaba fonética átona no meio ou final de palavras
- 1. Sílaba fonética tônica no meio ou final de palavras
- 2. Sílaba fonética átona no início de palavras
- 3. Sílaba fonética tônica no início de palavras
- 4. Final de frase

Exemplo:

A seleção brasileira venceu a argentina no último sábado.

A/se/le/ção bra/si/lei/ra ven/ceu a /ar/gen/ti/na no/úl/ti/mo sá/ba/do/.

2 0 0 1 2 0 1 0 2 1 2 0 0 1 0 2 1 0 0 3 0 0 4

Dado o objetivo de verificar possíveis diferenças rítmicas entre os textos disponíveis, foi verificado que todos os textos apresentaram uma VLMC da mesma forma, com 11 folhas. Procuramos então identificar diferenças nas probabilidades de transição.

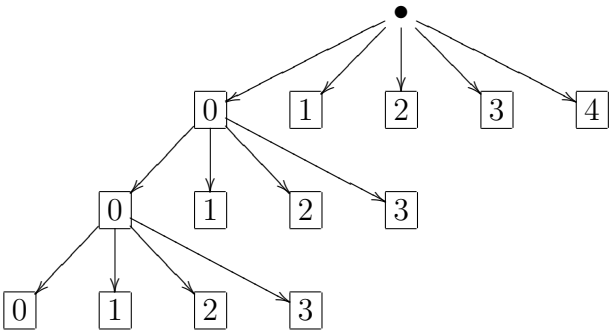


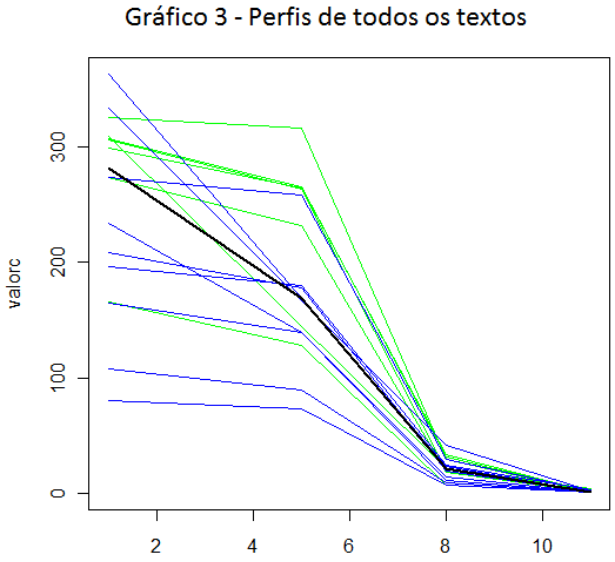
Figura 7: VLMC estimada dos textos do Corpus Tycho

Não foi encontrado nenhuma diferença significativa entre os textos, nem olhando para a estrutura da árvore estimada, uma vez que todas apresentaram a mesma VLMC como o melhor modelo, nem por suas probabilidades de transição, uma vez que o teste da razão de verossimilhança não deu provas de diferenças entre eles com p-valor inferior a 5%.

Com isso foi proposto encontrar um padrão de diferenciação de acordo com o rastro deixado pela constante de penalização, ou seja, pela assinatura da amostra, que como visto pelas simulações é uma técnica capaz de captar diferenças mais sutis.

As Assinaturas das Amostras dos 16 textos, como pode ser visto no Gráfico 3, sugerem que os textos são de fato diferentes. Para alguns deles, é muito mais caro aceitar um modelo independente do que para outros. Ou seja, o valor da constante de penalização para a árvore sem galhos é enorme, em alguns casos e é pequeno para outros textos.

A seqüência de árvores candidatas obtidas por quase todos os textos, uma VLMC com cinco folhas (fim da cadeia de Markov 1), a VLMC com oito folhas, onde os novos ramos saiu da folha "0"e, finalmente, a VLMC final como sendo a VLMC com 11 galhos com mais ramos que saem da folha "0 0". Alguns textos foram diretamente da árvore com 5 galhos para a árvore com 11 galhos. Mas o valor da constante de penalização é diferente



de um texto para outro, mostrando perfis distintos para cada texto, de acordo com a sua Assinatura da Amostra.

Desta forma, a comparação das Assinaturas das Amostras seria um meio adequado para mostrar as diferenças entre os textos, objetivo que não conseguimos alcançar se olharmos apenas para a melhor árvore estimada.

Aqui queremos apenas para verificar a diferença entre os grupos de textos ao longo dos anos. Isso poderia ser considerado como uma indicação de que houve uma mudança no ritmo de textos escritos e em que ano tal mudança tenha ocorrido.

Os textos foram primeiramente divididos em dois grupos distintos de acordo com o ano em que ele foi escrito. As comparações foram feitas entre quatro grupos, com o objetivo de tentar identificar se há diferenças entre os grupos e, se houver, em que ano aconteceu essa mudança. Também realizamos comparação através da divisão do conjunto de dados

em três grupos, mas não apresentaremos os resultados dessa comparação, pois não houve diferença significativa entre os grupos.

Primeira Comparação

Grupo 1: Textos entre os anos de 1502 á 1675

Grupo 2: Textos entre os anos de 1702 á 1836

Segunda Comparação

Grupo 1: Textos entre os anos de 1502 á 1702

Grupo 2: Textos entre os anos de 1705 á 1836

Terceira Comparação

Grupo 1: Textos entre os anos de 1502 á 1705

Grupo 2: Textos entre os anos de 1714 á 1836

Quarta Comparação

Grupo 1: Textos entre os anos de 1502 á 1714

Grupo 2: Textos entre os anos de 1750 á 1836

Entre essas quatro análises, somente no primeiro grupo de comparação se obteve diferença significativa entre os grupos. Em todos os outros grupos de comparação não se encontrou diferença significativa entre os grupos. Então vamos mostrar apenas os resultados para esse grupo de comparação.

Também foi observado que não há diferença significativa quando os grupos são divididos em três ocasiões diferentes no tempo, ou seja, não há duas mudanças na escrita ao longo do tempo.

Para verificar se existe diferença entre os textos antes e depois do ano de 1675, considere a seguinte notação.

Notação

Os textos foram divididos em dois grupos da seguinte maneira:

Grupo 1: Textos entre os anos de 1502 a 1675

Grupo 2: Textos entre os anos de 1702 a 1836

A Tabela 7 a seguir mostra a quantidade de textos por Folhas e Grupo. Nota-se que dentre os 16 textos, sendo 9 do grupo 2 e 7 do grupo 1, tivemos 3 textos sem informação sobre a Folha 5, onde dessas informações perdidas 1 é do grupo 1 e 2 do grupo 2.

Tabela 7 - Quantidade de Textos por Folhas e Grupos

Folhas	Grupos	
-	1	2
1	7	9
5	6	7
8	7	9
11	7	9

Verifica-se nos Gráficos 4 e 5 que o comportamento dos textos por Folhas em relação á variável resposta (valor C) mudou no que diz respeito a forma, e nota-se uma grande mudança na variabilidade dos valores da constante C da Folha 5 para a Folha 8.

Nota-se que a variabilidade da variável resposta se comporta de maneira similar entre as Folhas 1, 5 e 8 quando comparando os grupos, porém houve uma mudança brusca em relação a variabilidade da variável resposta do grupo 2 na Folha 11 quando comparado com a variabilidade da Folha 11 do grupo 1.

Como o interesse é somente verificar se existe diferença entre os grupos em relação a constante de penalização C, temos que;

Gráfico 4 - Boxplots do Valor da Constante C por Folhas

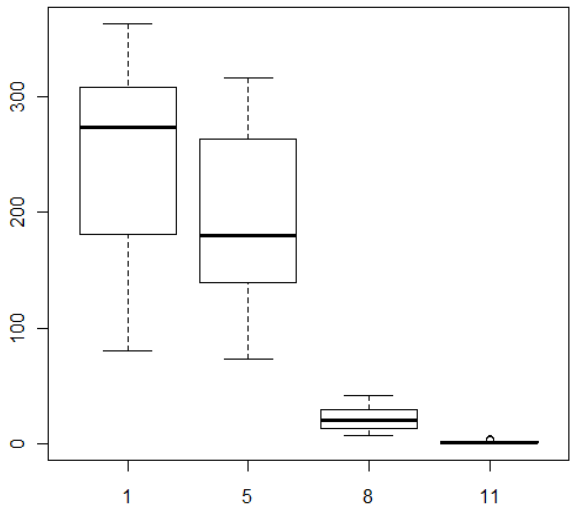
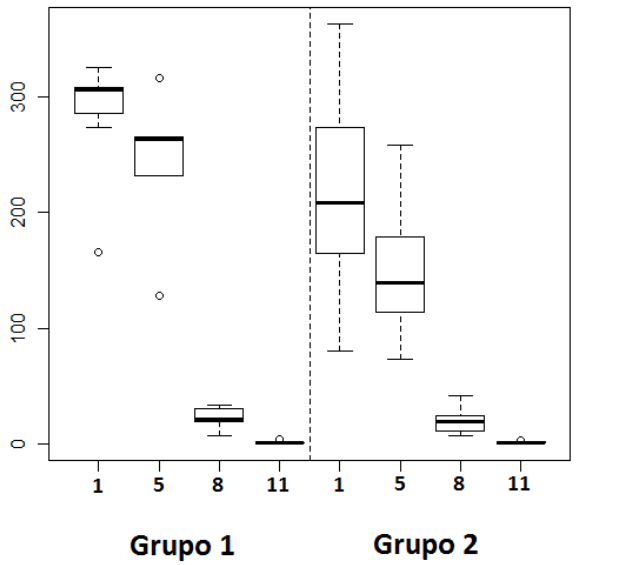


Gráfico 5 - Boxplots do Valor da Constante C por Folhas e Por Grupos



Seja Y_{ij} a resposta do i-ésimo texto do j-ésimo grupo. Um modelo satisfatório para analise de dados como este, chamado de Modelo Maximal, pode ser definido por:

$$E(Y_{ij}) = \beta_0 + \beta_1 I[Folha5]_{ij} + \beta_2 I[Folha8]_{ij} + \beta_3 I[Folha11]_{ij} + \beta_4 I[Grupo_2] +$$
$$\beta_5 (I[Folha5]_{ij} * I[Grupo_2]) + \beta_6 (I[Folha8]_{ij} * I[Grupo_2]) + \beta_7 (I[Folha11]_{ij} * I[Grupo_2]).$$

em que, $i = 1, 2, ..., n_j$ é o i-ésimo texto e $j = 1, 2, ...J$ é o j-ésimo grupo.

Os resultados do ajuste para o banco de dados desbalanceados:

Tabela 8 - Estimativas do Modelo Marginal GEE - AR(1)

Coeficientes	Estimativa	S.E.Naive	S.E.Robust	Z-Robust	P-Valor
Intercepto	284.53	19.17	18.88	15.06	0.0000
Folhas 5	-37.97	19.42	6.49	-5.84	0.0000
Folhas 8	-262.97	22.34	15.94	-16.49	0.0000
Folhas 11	-284.00	24.77	18.48	-15.36	0.0000
Grupo 2	-63.00	25.54	35.88	-1.75	0.0791
Folhas 5:Grupo 2	-17.86	26.20	18.88	-0.94	0.3440
Folhas 8:Grupo 2	52.38	29.54	34.14	1.53	0.1250
Folhas 11:Grupo 2	58.92	32.91	36.64	1.60	0.1078

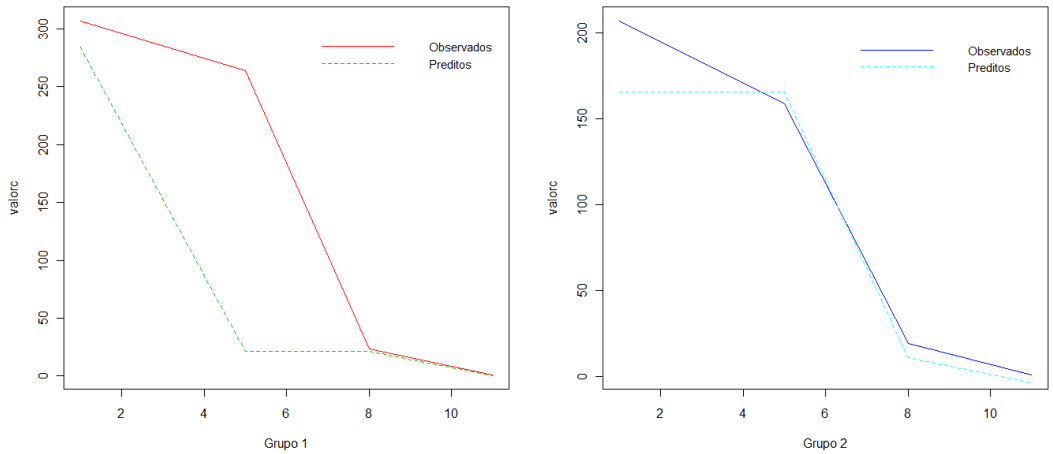
Tabela 9 - Matriz da Correlação de Trabalho

	1	2	3	4
1	1.0000000	0.5269896	0.2777181	0.1463546
2	0.5269896	1.0000000	0.5269896	0.2777181
3	0.2777181	0.5269896	1.0000000	0.5269896
4	0.1463546	0.2777181	0.5269896	1.0000000

Observa-se que todos os coeficiente das Folhas foram significativos, indicando que não existe diferença entre o comportamento da variável resposta nessas Folhas em relação á Folha 1. Porém não houve interação entre Folhas e Grupos. E verifica-se que existe diferença entre os grupos ao nível de significância de 10%. E a matriz de correlação mostra que realmente uma estrutua AR(1) é bem ajustado nesse caso, pois nota-se que a medida que as observações vão se distanciando uma da outra a correlação vai diminuindo entre elas.

E através do Gráfico 6, verifica-se que o modelo proposto não parece estar bem ajustado.

Gráfico 6 - Valores Observados vs Valores Preditos



Como não houve interação entre folhas e grupo, ajustamos um modelo sem interação, da forma:

$$E(Y_{ij}) = \beta_0 + \beta_1 I[Folha5]_{ij} + \beta_2 I[Folha8]_{ij} + \beta_3 I[Folha11]_{ij} + \beta_4 I[Grupo_2].$$

em que, $i = 1, 2, ..., n_j$ é o i-ésimo texto e $j = 1, 2, ...J$ é o j-ésimo grupo.

Os resultados do ajuste para o banco de dados desbalanceados:

Tabela 10 - Estimativas do Modelo GEE - AR(1)

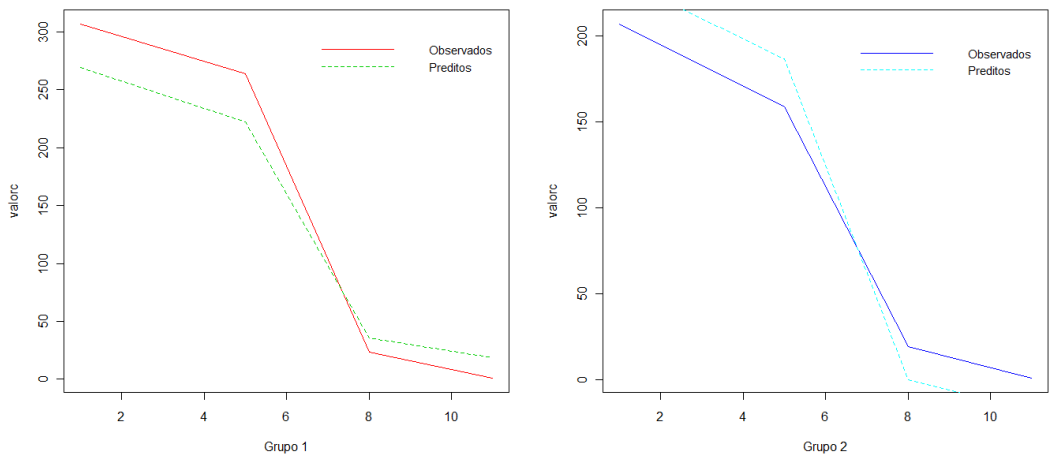
Coeficientes	Estimativa	S.E.Naive	S.E.Robust	Z-Robust	P-Valor
Intercepto	269.34	17.41	17.23	15.62	0.0000
Folhas 5	-46.65	13.26	9.39	-4.96	0.0000
Folhas 8	-233.36	14.99	19.19	-12.15	0.0000
Folhas 11	-250.92	16.85	20.65	-12.14	0.0000
Grupo 2	-36.03	20.08	16.94	-2.12	0.0334

Tabela 11 - Matriz da Correlação de Trabalho - Dados desbalanceados

	1	2	3	4
1	1.0000000	0.5548990	0.3079129	0.1708605
2	0.5548990	1.0000000	0.5548990	0.3079129
3	0.3079129	0.5548990	1.0000000	0.5548990
4	0.1708605	0.3079129	0.5548990	1.0000000

Os resultados do modelo ajustado sem a interação mostram que todos os coeficientes são significativos e o coeficiente dos grupos abaixou a significância para 3%. E percebe-se que em média o Grupo 2 tem um valor da constante C 36 menor do o valor de C do Grupo 1. E a matriz de correlação mostra que realmente uma estrutua AR(1) é bem ajustada nesse caso, pois nota-se que a medida que as observações vão se distanciando uma da outra a correlação vai diminuindo entre elas. O gráfico dos valores previstos vs valores observados mostra um modelo bem ajustado, indicando que há de fato um ponto de corte no ano de 1675, ou seja, houve uma mudança nos textos escritos, de acordo com suas características rítmicas captadas pela codificação proposta.

Gráfico 7 - Valores Observados vs Valores Preditos



7 *Conclusões*

Os resultados das simulações mostram que a Assinatura da Amostra identifica uma amostra de tamanho fixo n no sentido de que se as amostras são extraídas da mesma cadeia, as Assinaturas das Amostras dessas amostras são muito semelhantes, e se as amostras são retiradas de diferentes cadeias, as Assinaturas se diferem significativamente.

Podemos concluir que a assinatura da amostra é uma ferramenta muito interessante para comparar as seqüências discretas de informação modelado como Cadeias de Memória de Alcance Varável para uma amostra de tamanho fixo. E que utilizando as sequências de valores da constante penalizadora C (assinatura da amostra), fomos capazes de classificar, utilizando da metodologia das Equações de Estimações Generalizadas (GEE), os textos escritos do Português histórico de acordo com conjecturas linguística, o que não era possível fazer usando apenas o melhor modelo final para cada texto.

Nós alcançamos também o objetivo de propôr um programa para Estimação de VLMC para ser implementado no software R, que além de estimar consistentemente a VLMC de uma amostra, ainda constrói o seu gráfico, o que não é feito em outros programas que estimam VLMC.

8 *Apêndice 1*

A seguir será apresentado o algorítmo para a implementação da computação do estimador BIC (Csiszár & Talata, 2006) para a estimação de Cadeias de Memória de Alcance Variável.

Algoritmo BIC

Input: Vetor X: $\{X_1, X_2, ..., X_n\} \leftarrow$ Amostra

Output: Objeto Cadeia de Memória de Alcance Variável

Passo 1: Ler a amostra X e montar as matrizes de ordens$(1, 2, ..., log(n))$

$mount = function\ (log(n), X) \leftarrow$ Crie uma função que irá gerar as matrizes

$for\ (i\ in\ 1 : (length(X) - log(n) + 1))\ \{$

$X[i : (i + log(n) - 1)] \leftarrow$ cria as matrizes de ordens$(1, 2, ..., log(n))$

$\}\leftarrow$ Fim do comando for

Passo 2: Compare as Linhas iguais da VLMC e retire-as fazendo a contagem

Fazendo a leitura na Matriz de ordem $(1, 2, ..., log(n))$

if (Faça a comparação)$\{\leftarrow$ Verificando quais linhas da Matriz são iguais

Acresente o vetor das contagens na Matriz $\}\leftarrow$ Fim do comando if

return(nome da Matriz)

} Fim da função mount

Passo 3: Criando as Probabilidades futuras para um determinado nó das Matrizes de ordem $(1, 2, ..., log(n))$

Crie um list somente com a Matrize das Frequencias dos estados

Crie o vetor de uma parte do cálculo da verossimilhança $\left[\frac{N(s,a)}{N(s)}\right]^{N(s,a)}$

Crie o vetor de Probabilidades futuras $\frac{N(s,a)}{N(s)}$

$mount(1, X) \leftarrow$ Matriz inicial de Ordem 1

$for\ (i\ in\ 1 : (log(n) - 1))\ \{$

$mount((i + 1), X) \leftarrow$ Matriz de Ordem Maior

$for\ (j\ in\ 1 : nrow(mount(1, X)))\ \{$

$for\ (k\ in\ 1 : nrow(mount((i + 1), X)))\ \{$

if (compare somente as colunas dos estados de cada linha da Matriz de Ordem Maior (1: ncol(somente colunas dos estados da Matriz de Ordem Menor)) com as colunas dos estados de cada linha da Matriz de Ordem Menor)$\{\leftarrow$ Verificando quais linhas são iguais

$\}\leftarrow$ Fim do comando if

$\}\leftarrow$ Fim do comando for para o k

$\}\leftarrow$ Fim do comando for para o j

$mount(1, X) = mount((i + 1), X) \leftarrow$ Matriz de ordem maior já calculada é a nova matriz $mount(1, X)$

$mount((i+1),X)$=$cbind(mount((i+1),X)$,vetor das Probabilidades da verossimilhança , vetor de Probabilidades futuras)
$lista[[(i+1)]]$=$mount((i+1),X)$ ← Armazendo as Matrizes em list
}← Fim do comando for para o i

Passo 4: Fazendo a Multiplicação dos vetores $\left[\frac{N(s,a)}{N(s)}\right]^{N(s,a)}$
Crie um list para agrupar a Multiplicação dos vetores
$for\ (i\ in\ 2 : length(lista))\ \{$
$vetor$=$rep(0, nrow(lista[[(i-1)]]))$ ← Crie um vetor de 1´s de tamanho $nrow = lista[[i-1]]$ para não alterar a multiplicação dos vetores de cada Matriz
$for\ (k\ in\ 1 : nrow(lista[[(i-1)]]))\ \{$← Matriz de ordem menor
$for\ (l\ in\ 1 : nrow(lista[[i]]))\ \{$← Matriz de ordem maior
if (compara somente as colunas dos estados de cada linha da Matriz de Ordem Menor com as colunas dos estados de cada linha da Matriz de Ordem Maior (1: ncol(somente colunas dos estados da Matriz de Ordem Menor))){← Verificando quais linhas são iguais para verificar quais são os futuros
$vetor[k]$=$vetor[k] * lista[[i]][l,(i+2)]$ ← Fazendo a multiplicação dos futuros
}← Fim do comando if
}← Fim do comando for para o l
}← Fim do comando for para o k
$list[[i]]$=$vetor$ ← list = valores da multiplicação dos futuros
$lista[[i-1]]$=$cbind(lista[[i-1]], vetor)$
}← Fim do comando for para o i

Passo 5: Criando a Verossimilhança com os Vetores das Multiplicações criados no **Passo 4** em um determinado nó
Crie um vetor para agrupar os novos vetores das Multiplicações que serão multiplicado pela constante C
$for\ (j\ in\ 1 : nrow(lista[[1]]))\ \{$
$vetor[j]$=$(1/(sqrt(length(X)^{(|A|-1)}))) * lista[[1]][j,4]$ ← Fazendo a multiplicação
}← Fim do comando for
$lista[[1]]$=$cbind(lista[[1]], vetor)$
Foi separado, pois a Matriz de freqüência dos estados ($lista[[1]]$)tem o número de colunas diferente para a comparação com a Matriz maior.
$for\ (i\ in\ 2 : (length(lista)-1))\ \{$
$for\ (j\ in\ 1 : nrow(lista[[i]]))\ \{$
$vetor[j]$=$(1/(sqrt(length(X)^(Y-1))))) * lista[[i]][j,i+4]$ ← Fazendo a multiplicação
}← Fim do comando for para o j
$lista[[i]]$=$cbind(lista[[i]], vetor)$
}← Fim do comando for para o i

Passo 6: Montando os cálculos levando em consideração as folhas de um determinado nó
$Constante$=$(1/(sqrt(length(X)^{(|A|-1)})))$ ← String Nulo
Crie um vetor da Multiplicação dos strings ($folhas$) de ordem 1 que está na $lista[[1]]$
$soma$=$list()$ ← Crie um list para agrupar as Multiplicações das folhas de um determinado nó das Matrizes
$for\ (i\ in\ 2 : (length(lista)-1))\ \{$
Crie um vetor de 1´s para não alterar a multiplicação dos vetores de cada Matriz

$for\ (k\ in\ 1 : nrow(lista[[(i-1)]]))\{$

$for\ (l\ in\ 1 : nrow(lista[[i]]))\{$

if (compara somente as colunas dos estados de cada linha da Matriz de Ordem Menor com as colunas dos estados de cada linha da Matriz de Ordem Maior (1: ncol(somente colunas dos estados da Matriz de Ordem Menor))){← Verificando quais linhas são iguais para verificar quais são os futuros

$vetor[k]$=$vetor[k] * lista[[i]][l,(i+5)]$

}← Fim do comando if

}← Fim do comando for para o l

}← Fim do comando for para o k

$soma[[i]]$=$vetor$

$lista[[i-1]]$=$cbind(lista[[i-1]], vetor)$

}← Fim do comando for para o i

Passo 7: Montando os Indicadores

Crie um vetor para guardar os indicadores

$for\ (i\ in\ 1 : (length(lista)-2))\{$

$for\ (j\ in\ 1 : nrow(lista[[i]]))\{$

$if(lista[[i]][j, ncol(lista[[i]])] > lista[[i]][j, ncol(lista[[i]])-1])\{$← Fazendo as comparações

$vetor[j]$=1

}← Fim do comando if

$else$

$vetor[j]$=0

}← Fim do comando for para o j

$lista[[i]]$=$cbind(lista[[i]], vetor)$

}← Fim do comando for para o i

Passo 8: Fazendo os cortes da VLMC e Montando os galhos que farão parte da cadeia

Crie um list para armazenar a Árvore

$cont$=1

$for\ (i\ in\ 2 : length(lista))\{$

$for\ (j\ in\ 1 : nrow(lista[[i-1]]))\{$

$for\ (k\ in\ 1 : nrow(lista[[i]]))\{$

if (compara somente as colunas dos estados de cada linha da Matriz de Ordem Menor com as colunas dos estados de cada linha da Matriz de Ordem Maior (1: ncol(somente colunas dos estados da Matriz de Ordem Menor))){← Fazendo a comparação para o corte

$list[[cont]]$=$lista[[i]][k, ncol$=somente até o números de colunas dos estados]

$cont$=$cont+1$

}← Fim do comando if

}← Fim do comando for para o k

}← Fim do comando for para o j

}← Fim do comando for para o i

Passo 9: Buscando a Matriz de Transição

Crie um list para agrupar as Matrizes

$cont$=1

for (k in 1 : $length$(list criado no Passo 8)){

for (l in 1 : $nrow$($lista$[[$length$(list criado no Passo 8[[k]])+1]])){

if(sum(lista[[length(list criado no Passo 8[[k]]) + 1]][l, 1 : $length$(list cirado no Passo 8[[k]])]!=list criado no Passo 8[[k]]) == 0){← Fazendo a comparação

$list$[[$cont$]]=$lista$[[$length$(list criado no Passo 8[k]]) + 1]][l, c((1 : ($length$(list criado no Passo 8[[k]]) + 1)), ($length$(list criado no Passo 8[[k]]) + 4))], 2) ← Buscando as probabilidades de transição nas Matrizes

$cont$=$cont$ + 1

}← Fim do comando if

}← Fim do comando for para o l

}← Fim do comando for para o k

9 *Apêndice 2*

A seguir apresentaremos um exemplo de uma estimação de uma VLMC utilizando o programa PCT, a fim de se verificar que o nosso programa estima de forma eficiente e ainda nos auxilia na visualiazação gráfica da VLMC.

Seja uma VLMC verdadeira em que as probabilidades de transição $\mathbb{P}\left(X_0 = a \left| t\left(X_k^{-1} = x_k^{-1}\right)\right.\right.$, $\forall a \in \mathcal{A}, x \in \mathcal{A}^k$ e alfabeto finito $\mathcal{A} = \{1, 2, 3\}$ seja dada da seguinte maneira:

VLMC Verdadeira

Contextos	1	2	3
111	0.25	0.18	0.57
211	0.21	0.08	0.71
311	0.11	0.40	0.49
121	0.19	0.39	0.42
221	0.40	0.47	0.13
321	0.48	0.38	0.14
131	0.13	0.70	0.17
231	0.10	0.15	0.75
331	0.21	0.19	0.60
12	0.39	0.20	0.41
22	0.32	0.32	0.36
32	0.08	0.77	0.15
13	0.12	0.01	0.87
23	0.27	0.27	0.46
33	0.50	0.33	0.17

Dada essa VLMC, utilizamos o programa gera para nos fornecer uma amostra x_1^n que viesse dessa lei de formação dada acima, em que o tamanho n da amostra, nesse caso, é definido de forma que se tenha possibilidades lógicas de se encontrar uma matriz de transição com as devidas probabilidades associadas.

Dessa forma, teríamos uma amostra na qual sabíamos qual seria a forma da VLMC, ou seja, qual seria a matriz de transição e suas probabilidades que o programa PCT teria que recuperar. Logo diante da amostra x_1^n é que iniciamos o nosso programa PCT.

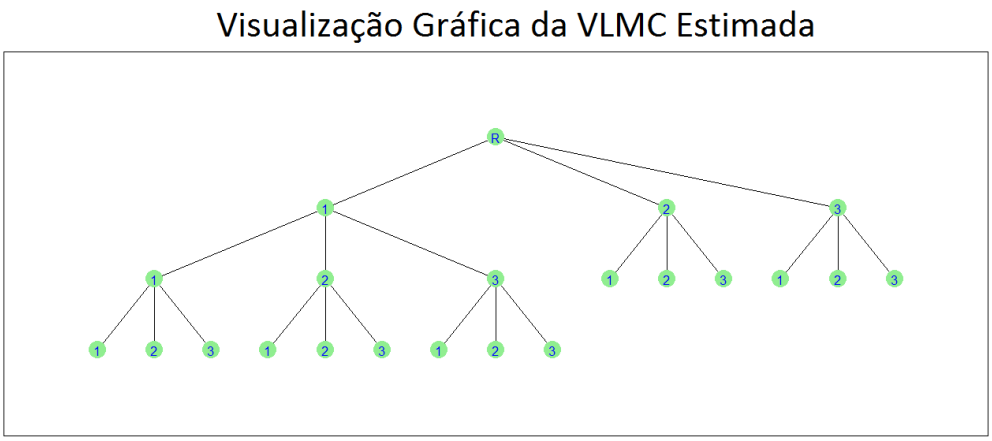
Aplicando o algoritmo do PCT como descrito na seção ***3.3*** e utlizando a amostra x_1^n fornecida pelo programa gera, de tamanho 10.000, o programa estimou uma VLMC com matriz de transição com probabilidades $\hat{\mathbb{P}}\left(X_0 = a \left| t\left(X_k^{-1} = x_k^{-1}\right)\right.\right., \forall a \in \mathcal{A}, x \in \mathcal{A}^k$ e alfabeto finito $\mathcal{A} = \{1, 2, 3\}$ da seguinte maneira:

VLMC Estimada			
Contextos	1	2	3
111	0.29	0.18	0.53
211	0.21	0.05	0.74
311	0.12	0.39	0.49
121	0.14	0.43	0.42
221	0.43	0.44	0.14
321	0.53	0.36	0.12
131	0.14	0.70	0.16
231	0.14	0.15	0.72
331	0.21	0.18	0.61
12	0.40	0.20	0.40
22	0.35	0.31	0.35
32	0.09	0.75	0.17
13	0.12	0.01	0.87
23	0.28	0.27	0.45
33	0.50	0.34	0.15

Percebe-se que com um tamanho de amostra de apenas 10.000, considerado razoavelmente baixo para problemas envolvendo VLMC, o programa PCT foi capaz de recuperar de forma eficaz a verdadeira VLMC.

Verifica-se que as probabilidades de transições estimadas são realmente muito próximas das verdadeiras. E com um tamanho de amostra maior, essas probabilidades se tornam praticamente iguais mesmo trabalhando com mais casas decimais.

E através do programa PCT podemos visualizar a VLMC através do gráfico da sua árvore de contexto, como segue na figura abaixo.



Portanto, verifica-se que o programa PCT implementado no software R, estima de forma eficiente uma VLMC e nos fornece uma boa visualização gráfica.

10 *Referências Bibliográficas*

Referências

[1] Bühlmann (2000) Model selection for variable length Markov chains and tuning the context algorithm. Ann. Inst. Statist. Math., 52(2).

[2] Bühlmann and A. J. Wyner (1999) Variable length Markov chains. Ann.Statist., 27, 1999.

[3] Castro, I. (2006). Introducao a historia do Portugues. Lisboa: Edies Colibri.

[4] Csiszar, I. & Talata, Zs. (2006) Context tree estimation for not necessarily nite memory processes, via BIC and MDL. IEEE Trans. Inform. Theory, 52(3)

[5] Csiszar, I. & Shields, Paul C. (2000) Consistency of the BIC Order Estimator. Eletronic Research Announcements of the Americam Mathematical Society: volume 5, Pages 123,127.

[6] Duarte, D., Galves, A. e Garcia, N. L. (2006) Markov approximation and consistent estimation of unbounded probabilistic suffix trees. Bulletin of the Brazilian Mathematical Society.

[7] Frota, S., Galves, C, Vigário, M.,Gonzalez-Lopez, V. , Abaurre, B (2010), The phonology of rhythm from Classical to Modern Portuguese. Submitted to Language Variation and Change.

[8] Galves, A., Galves, C., Garcia, N. and Leonardi, F. (2011) Context tree selection and linguistic rhythm retrieval from written texts. ArXiv: 0902.3619v2.

[9] K.Y., e Zegar, S.L. (1986) Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.

[10] Rissanen, J.(1983) A universal data compression system. IEEE Trans. Inform.Theory, 29(5).

[11] Teyssier, P. (1980). Histoire de la langue portugaise. Paris: PUF.