Universidade de Brasília
Departamento de Estatística

Cadeias Estocásticas de Ordem Variável
Um Estudo de Simulação e Aplicações

Pedro Caio Limeira de Miranda

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Pedro Caio Limeira de Miranda

Cadeias Estocásticas de Ordem Variável
Um Estudo de Simulação e Aplicações

Orientador: Lucas Moreira
Coorientador: Felipe Sousa Quintino

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Dedico este trabalho aos meus pais e ao meu irmão, por todo esforço em prol da minha educação e por sempre me darem a liberdade de estudar.

Agradecimentos

Agradeço, primeiramente, a Deus pela saúde, bênçãos e por todos os detalhes que me fizeram conseguir terminar este trabalho.

Aos meus pais, Maurício e Francineide, por todo o amor e dedicação na minha criação, além de todo esforço para minha educação.

Ao meu irmão, Victor, por ser uma grande inspiração e por sua dedicação ao sempre me ajudar em todas as áreas possíveis.

À minha namorada, Giovanna, por me acompanhar e apoiar durante toda a graduação, mesmo à distância.

Aos meus grandes amigos, Eduardo, Vinícius, Gilson, Lucão e Caio, por todos os momentos e aprendizados compartilhados; e a todos os meus amigos de infância.

Aos meus amigos da Estatística, Amanda, Laíza, Malu e André, pelo esforço conjunto e apoio mútuo.

Aos meus orientadores, Lucas e Felipe, por acreditarem no meu potencial, pela força e ajuda que me deram durante este trabalho.

Aos professores com quem tive o prazer de aprender, em especial, ao Bernardo que me ajudou na primeira experiência profissional.

Resumo

Este trabalho estuda as Cadeias Estocásticas de Ordem Variável, introduzidas por Rissanen (1983). Esta abordagem estende as Cadeias de Markov e foca na Teoria de Processos Estocásticos e Probabilidade Clássica. A principal ideia é que apenas uma parte do passado, denominada contexto, é relevante para prever o próximo símbolo, com seu comprimento determinado pelo próprio histórico. O problema de pesquisa central é analisar como diferentes métodos de estimação afetam a precisão das árvores de contextos estimadas, e como o modelo se comporta ao ser aplicado a dados climáticos e financeiros. O objetivo geral é analisar os métodos de estimação de árvores de contextos por meio de estudos simulados, entender o comportamento dos estimadores e aplicá-los na análise das mudanças no regime de chuvas e na precisão das previsões financeiras. A metodologia inclui a exploração de métodos de seleção de modelos e a estimação de árvores de contexto. As aplicações foram realizadas em dados climáticos do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET) e em dados financeiros importados do pacote "quantmod" do Software R. Os resultados mostram que as Cadeias Estocásticas de Ordem Variável são ferramentas valiosas para a previsão em áreas dinâmicas como climatologia e finanças, devido à sua capacidade de adaptar-se a contextos variáveis. Este estudo contribui para a compreensão e aplicação dessas técnicas em diferentes áreas, destacando a relevância de modelos analíticos avançados na formulação de políticas públicas e na gestão econômica.

Palavras-chaves: Cadeias Estocásticas de Ordem Variável; Árvores de Contexto; Meteorologia; Mudanças no Regime de Chuvas; Previsões; Mercado Financeiro.

Abstract

This work study the Stochastic Chains With Memory of Variable Length, introducted by Rissanen (1983). This approach extend the Markov Chains and focus on Stochastics Process Theory and Classical Probability. The main idea is that only one part of past, called context, is relevant to predict the next symbol, with your length determined by the own past. The central problem of research is to analyze how different estimation methods affects the precision of estimated context tree, and how the model behaves when applied to climatic and financial data. The overall objective is to analyze context tree estimation methods through simulated studies, understand the behavior of the estimators, and apply them to analyze changes in rainfall patterns and the accuracy of financial forecasts. The methodology includes the exploration of selection models methods and the context tree estimation. The applications were conducted using climatic data from the Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) of the Instituto Nacional de Meteorologia (INMET) and financial data imported from the "quantmod" package of the R software. The results show that Stochastic Chains With Memory of Variable Length are valuable tools for forecasting in dynamic areas such as climatology and finance, due to their ability to adapt to varying contexts. This study contributes to the understanding and application of those techniques in different areas, highlighting the relevance of advanced analytics models in the formulation of political policies and economic management.

Keywords: Stochastic Chain With Memory of Variable Length; Context Trees; Meteorology; Changes in Rainfall Patterns; Forecasts; Financial Market.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

Neste trabalho serão estudadas as Cadeias Estocásticas de Ordem Variável, introduzidas em Rissanen (1983), também chamadas de Cadeias com Memória de Alcance Variável, foram popularizadas na literatura matemática e estatística em Bühlmann e Wyner (1999). Essas cadeias são uma generalização das Cadeias de Markov, portanto, este trabalho possui enfoque na Teoria de Processos Estocásticos e Probabilidade Clássica.

A ideia das Cadeias Estocásticas de Ordem Variável é que apenas uma parte do passado é relevante para predizer o próximo símbolo. Essa sequência, porção do passado, é chamada de contexto e seu comprimento é função do próprio passado. Esses contextos podem ser representados por uma árvore, chamada de árvore de contextos, que será muito importante para os critérios de seleção de modelos. A vantagem de um modelo de ordem variável quando comparado às Cadeias de Markov de Ordem k, quando os passados considerados possuem comprimento $k \in \mathbb{N}$, é que eles são muito mais flexíveis e precisa-se estimar menos parâmetros. As áreas de aplicação de Cadeias Estocásticas de Ordem Variável são inúmeras, quais sejam linguística, bioinformática, música, neurobiologia e meteorologia.

Para a estimação das árvores de contextos, serão estudados e aplicados diversos critérios. Por exemplo, o Algoritmo Contexto, também introduzido em Rissanen (1983), que utiliza a divergência de Kullback-Leibler para podar a árvore de contextos até se obter a estimativa final da árvore. Além desse algoritmo, serão exploradas outras técnicas como a Máxima Verossimilhança Penalizada (PML), amplamente utilizada em diversas áreas da estatística. A PML inclui o Critério de Informação Bayesiano (BIC), que foi amplamente explorado em Csiszár e Talata (2006), a fim de estimar a árvore de contextos por meio do BIC.

O foco deste estudo é investigar como diferentes métodos de estimação afetam a precisão das árvores de contextos estimadas, e como as Cadeias Estocásticas de Ordem Variável se comportam ao serem aplicadas a dados climáticos e financeiros. Alguns trabalhos já utilizaram esses modelos em dados climáticos, como demonstrado em Quintino (2015) que aplicou para previsão de temperatura e em Ramos (2023) que usou a modelagem para dados de insolação. Em dados financeiros, as aplicações são mais frequentemente associadas às Cadeias de Markov, como em Delgado, Queiroz e Átila (2023). O objetivo geral é analisar esses métodos por meio de estudos simulados para compreender o comportamento dos estimadores e aplicá-los na análise das mudanças no regime de chuvas e na precisão das previsões financeiras.

Nesse sentido, será conduzido um estudo de simulação para avaliar o desempenho dos estimadores investigados. Posteriormente, será analisado o comportamento dessas técnicas em dados reais. O estudo de aplicação abordará duas áreas distintas: a primeira utilizando dados climáticos e a segunda utilizando dados financeiros. A primeira aplicação analisa o regime de chuvas no Rio Grande do Sul, motivado pelo desastre climático ocorrido em maio de 2024. Já a segunda aplicação, analisa o poder preditivo das Cadeias Estocásticas de Ordem Variável, reforçando a sua versatilidade. Todas as simulações e estimativas foram realizadas utilizando o software R, versão 4.3.1 (R Core Team, 2023).

Este trabalho está estruturado da seguinte forma: no Capítulo 1, são apresentados a introdução, o problema de pesquisa e os objetivos do trabalho. No capítulo 2, é abordada toda teoria necessária sobre Cadeias Estocásticas de Ordem Variável. No Capítulo 3, são detalhadas as técnicas e os métodos utilizados para o estudo. No Capítulo 4, são apresentados os resultados do estudo de simulação e aplicações. Por fim, no Capítulo 5, são discutidas as conclusões obtidas através dos resultados.

2 Fundamentação Teórica

2.1 Conceitos e Notações Básicas

Um processo estocástico é uma sequência ou, de forma mais geral, uma coleção, de variáveis aleatórias $X(t)$, $t \in T$, definidas em um espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$, em que, T é um subconjunto de $(-\infty, +\infty)$ e $X(t)$ assume valores em um mesmo espaço de estados $\mathcal{S}$. Um processo é dito de tempo discreto, se T é um subconjunto dos inteiros. Um processo é dito de tempo contínuo, se T é um intervalo de comprimento positivo (Hoel; Port; Stone, 1986).

Também é comum representar um processo estocástico $X(t)$, $t \in T$, com a seguinte notação $\{X_t\}_{t\in T}$. Além disso, nos estudos de Cadeias de Alcance Variável, é mais comum denotar o espaço de estados por um alfabeto $\mathcal{A}$.

Definição 2.1 *Seja $\{X_t\}_{t\in T}$ um processo estocástico em $T = \mathbb{Z}$ e com um alfabeto finito e discreto $\mathcal{A}$. Então, para todo $x_t, x_{t-1}, x_{t-2}, ... \in \mathcal{A}$ e para todo $t \in \mathbb{Z}$, se o processo segue a propriedade de Markov, tem-se*

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, ...) = P(X_t = x_t | X_{t-1} = x_{t-1}).$$

Os processos que satisfazem essa propriedade são chamados de Cadeias de Markov.

Definição 2.2 *Seja um processo $\{X_t\}_{t\in\mathbb{Z}}$ com valores em um alfabeto finito e discreto $\mathcal{A}$. Tal processo é uma Cadeia de Markov de Ordem k, se para um $k \in \{1, 2, ...\}$, tem-se*

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, ...) = P(X_t = x_t | X_{t-1} = x_{t-1}, ..., X_{t-k} = x_{t-k}).$$

Para todo $x_t, x_{t-1}, x_{t-2}, ... \in \mathcal{A}$ e para todo $t \in \mathbb{Z}$.

Pode-se definir Cadeias de Ordem Infinita. Para mais informações, recomenda-se a leitura de Matta e Garcia (2008). Os conceitos apresentados até aqui, serão a base das ideias e dos novos conceitos que serão apresentados em diante.

Dito isso, considere um alfabeto $\mathcal{A}$ discreto e finito, e $a_m^n = (a_m, ..., a_n)$ uma sequência de símbolos de $\mathcal{A}$, tal que $m, n \in \mathbb{Z}$ e $m \leq n$. O comprimento da sequência é dado por $\ell(a_m^n) = n - m + 1$. Caso a sequência seja vazia, então $a = \emptyset$ e $\ell(a) = 0$, isso se aplica ao caso em que $m > n$.

Sejam u e v sequências de símbolos em $\mathcal{A}$, então uv é a concatenação dessas sequências. Uma sequência v é sufixo de uma sequência w, se $w = uv$, para qualquer sequência u não vazia, isso é, $\ell(u) \geq 1$, tal que $\ell(v) \leq \ell(w)$. Nesse caso, é denotado que $v \prec w$. Já para uma sequência u que seja sufixo de w ou que u seja igual a w, denota-se $u \preceq w$. O maior sufixo de uma sequência é denotado por suf(.).

Exemplo 2.1 *Considere um alfabeto $\mathcal{A} = \{0, 1\}$, tal que $a = 111$, $b = 0011$ e $c = 11$ sejam sequências de elementos de $\mathcal{A}$. Então, $ab = 1110011$ é a concatenação de a e b. Além disso, perceba que c é sufixo tanto de a, quanto de b. O maior sufixo de ab, é $suf(ab) = 110011$.*

Denota-se por $\mathcal{A}^k$ o conjunto de todas as sequências de elementos, do alfabeto $\mathcal{A}$, de tamanho k. Dessa maneira, $\mathcal{A}^* = \bigcup_{k=0}^{\infty} \mathcal{A}^k$ é o conjunto de todas as sequências sob o alfabeto $\mathcal{A}$, sendo que $\mathcal{A}^0$ é a sequência vazia. Além disso, é denotado por $\mathcal{A}^\infty$ o conjunto de todas as sequências semi-infinitas em $\mathcal{A}$.

Visto isso, agora será possível estudar as árvores associadas aos processos. Um conjunto $\mathcal{T} \in \mathcal{A}^* \cup \mathcal{A}^\infty$ é uma árvore irredutível se satisfaz duas seguintes propriedades.

1. **Propriedade do Sufixo**: nenhuma sequência pertencente à árvore pode ser sufixo de outra sequência da mesma árvore. Isso é, para uma sequência $u \in \mathcal{T}$, não pode haver uma sequência $v \in \mathcal{T}$, tal que $v \prec u$.

2. **Irredutibilidade**: nenhuma sequência em $\mathcal{T}$ pode ser substituída por uma subsequência de si mesma (um sufixo próprio), sem que esse sufixo viole a Propriedade do Sufixo. Por exemplo, seja $w \in \mathcal{T}$ e s um sufixo próprio de w, isso é, $s \prec w$; se w for substituído por s, então haverá uma sequência v, tal que $s \prec v$. Dessa forma, violando a Propriedade do Sufixo.

Os elementos (sequências) da árvore são chamados de folhas; os nós internos são os sufixos das folhas; e os nós são o conjunto de todas as folhas e nós internos da árvore. Os descendentes de um nó interno s são todas as sequências as, $a \in \mathcal{A}$, que são nós. Uma árvore $\mathcal{T}$ é completa se cada nó interno tem exatamente $|\mathcal{A}|$ descendentes.

A cardinalidade, número de elementos do conjunto, de uma árvore é dada por $|\mathcal{T}|$ e a sua profundidade é dada por $h(\mathcal{T}) = max\{\ell(s)|s \in \mathcal{T}\}$. Observa-se que $|\mathcal{T}|$ informa o número de folhas da árvore. Uma árvore é limitada se $h(\mathcal{T}) < \infty$, caso contrário, a árvore é ilimitada.

A seguir, será apresentado um exemplo de uma árvore. Naturalmente, desenha-se a árvore de cima para baixo, mas para isso é preciso ler as sequências da direita para a esquerda. Ressalta-se que as sequências são dadas em ordem dos símbolos mais remotos aos mais recentes, por isso essa lógica.

Exemplo 2.2 *Considere um alfabeto* $\mathcal{A} = \{0, 1\}$ *e a árvore* $\mathcal{T} = \{11, 10, 100, 101, 000, 001\}$. *Essa árvore satisfaz a propriedade do sufixo, pois nenhuma sequência em* $\mathcal{T}$ *é sufixo de outra sequência. Também satisfaz a propriedade da irredutibilidade, pois todas as sequências podem ser substituídas por sufixos próprios sem que não violem a propriedade do sufixo. A árvore* $\mathcal{T}$ *é representada abaixo.*

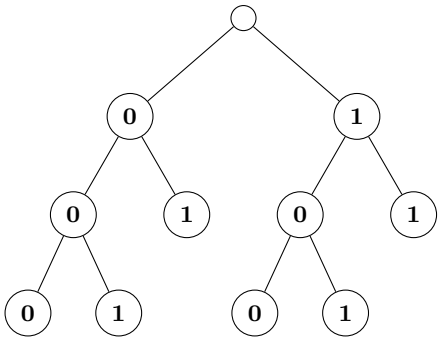


Figura 2.1: Exemplo de Árvore.

As folhas dessa árvore são os elementos de $\mathcal{T}$*, isso é, 11, 10, 100, 101, 000, 001; os nós internos são 0, 1, 00 e 01. Logo, os nós são 0, 1, 00, 01, 11, 10, 100, 101, 000, 001. Além disso, perceba que* $h(\mathcal{T}) = 3$*, portanto, limitada.*

Uma árvore pode ser truncada em k, de tal forma que

$$\mathcal{T}|_k = \{s \mid \ell(s) \leq k\} \ \cup \ \{s' \mid s' \prec s, \ \ell(s') = k \ \ e \ \ \ell(s) > k\}, \ s \in \mathcal{T}.$$

Para a árvore do Exemplo 2.2, a árvore truncada em $k = 2$, é dada por $\mathcal{T}|_2 = \{11, 10, 01, 00\}$.

Para as próximas definições, exceto caso seja especificado, considere um processo estacionário e ergódico $\mathbf{X} = \{X_t\}_{t\in\mathbb{Z}}$ em um alfabeto $\mathcal{A}$, tal que $|\mathcal{A}| < \infty$. Dada uma sequência $w \in \mathcal{A}^*$, seja $\mu_X(w)$ a probabilidade de observar w, tal que

$$\mu_X(w) = \begin{cases} P(X_1^{\ell(w)} = w), & se \ \ w \neq \emptyset; \\ 0, & se \ \ w = \emptyset. \end{cases} . \tag{2.1.1}$$

Seja $p_X(a|w)$ a probabilidade condicional, tal que

$$p_X(a|w) = \begin{cases} P(X_0 = a|X_{-\ell(w)}^{-1} = w), & se \;\; w \neq \emptyset; \\ \mu_X(a), & se \;\; w = \emptyset. \end{cases} \quad \forall a \in \mathcal{A}. \tag{2.1.2}$$

Definição 2.3 *Uma sequência $w \in \mathcal{A}^*$ é um contexto finito do processo $\boldsymbol{X}$, se são satisfeitas as seguintes características.*

1. $\mu_X(w) > 0$;
2. *Para toda sequência s que tem como sufixo w, isso é, $w \prec s$. Tem-se que:*

$$P(X_0 = a|X_{-\ell(s)}^{-1} = s) = p(a|w), \;\; \forall a \in \mathcal{A}.$$

3. *Por último, nenhum sufixo de w pode satisfazer 2.*

Um contexto infinito é uma sequência semi-infinita $s_{-\infty}^{-1}$ tal que seus sufixos finitos s_{-k}^{-1}, $k = 1, 2, ...$ têm probabilidade positiva, mas nenhum deles é um contexto do processo (Galves; Leonardi; Ost, 2022).

A árvore de contextos de $\mathbf{X}$ é o conjunto de todos os contextos do processo. De maneira que, $w \in \mathcal{T}$, se e somente se, w é um contexto para $\mathbf{X}$. Essa árvore é irredutível. Desse modo, agora é possível definir quando o processo $\mathbf{X}$ é compatível com a árvore de contextos $(\mathcal{T}, \bar{p})$.

Definição 2.4 *Uma árvore de contextos em $\mathcal{A}$ é probabilística, se o par ordenado $(\mathcal{T}, \bar{p})$ satisfaz*

1. *$\mathcal{T}$ é uma árvore de contextos irredutível;*
2. *$\bar{p} = \{p(.|s)|s \in \mathcal{T}\}$ é uma família de probabilidades de transição de elementos de $\mathcal{T}$ em $\mathcal{A}$.*

Definição 2.5 *Um processo $\boldsymbol{X}$ é compatível com a árvore probabilística de contextos $(\mathcal{T}, \bar{p})$, se*

1. *$\mathcal{T}$ é a árvore de contextos de $\boldsymbol{X}$;*
2. *$p_X(a|w) = p(a|w)$, para qualquer $w \in \mathcal{T}$ e $\forall a \in \mathcal{A}$.*

A árvore de contextos compatível com o processo $\mathbf{X}$ é denotada por $\mathcal{T}_X$. Tal processo é conhecido como Cadeia de Ordem Variável.

2.2 Inferência das Cadeias Estocásticas de Ordem Variável

Nesta seção, serão abordadas algumas propriedades, notações e conceitos para a estimação das probabilidades de transição da árvore de contextos do processo $\mathbf{X}$, através da máxima verossimilhança.

Considere uma amostra aleatória $X_1^n = \{X_1, X_2, ..., X_n\}$ do processo $\mathbf{X}$ e um inteiro d, tal que $d < n$. Dada uma sequência $w \in \mathcal{A}^*$, tal que $\ell(w) \leq d$, então, é denotado por $N_n(w, a)$ a quantidade de vezes em que wa aparece na amostra $X_{d-\ell(w)+1}^n$. Essa medida é dada por

$$N_n(w, a) = \sum_{t=d+1}^{n} \mathbb{1}\{X_{t-\ell(w)}^{t-1} = w, X_t = a\}. \tag{2.2.1}$$

Perceba que, a partir disso, é possível definir a quantidade de vezes em que w aparece na amostra $X_{d-\ell(w)+1}^{n-1}$, dada por

$$N_n(w) = \sum_{t=d+1}^{n} \mathbb{1}\{X_{t-\ell(w)}^{t-1} = w\} = \sum_{a \in \mathcal{A}} N_n(w, a). \tag{2.2.2}$$

Caso $d = \ell(w)$, então, a amostra considerada será X_1^n. Além disso, é importante destacar que nem sempre $N_n(w, a)$ será igual a $N_n(wa)$, já que as amostras consideradas são diferentes. Para $N_n(w, a)$, a amostra é $X_{d-\ell(w)+1}^n$, enquanto para $N_n(wa)$, a amostra é $X_{d-\ell(w)+1}^{n-1}$. Porém, assintoticamente, esse detalhe não impacta nos resultados.

Agora, é possível definir o estimador de máxima verossimilhança para as probabilidades de transição $p(a|w)$, $\forall w \in \mathcal{T}$ e $\forall a \in \mathcal{A}$.

2.2.1 Estimador de Máxima Verossimilhança

Considere uma cadeia de ordem variável $\mathbf{X}$ com elementos no alfabeto $\mathcal{A}$. Seja a árvore de contextos do processo $\mathcal{T}$ com profundidade $h(\mathcal{T}) \leq d$ e as probabilidades de transição $p(a|w)$, para $a \in \mathcal{A}$ e $w \in \mathcal{T}$.

Então, pelo princípio multiplicativo,

$$
\begin{aligned}
L(x_1^n, \mathcal{T}) &= P(X_1 = x_1,\ X_2 = x_2,\ ...,\ X_n = x_n) = P(X_1^n = x_1^n) \\
&= P(X_1^{n-1} = x_1^{n-1})\,.\,P(X_n = x_n | X_1^{n-1} = x_1^{n-1}) \\
&= P(X_1^{n-2} = x_1^{n-2})\,.\,P(X_{n-1} = x_{n-1} | X_1^{n-2} = x_1^{n-2})\,.\,P(X_n = x_n | X_1^{n-1} = x_1^{n-1}) \\
&\quad \vdots \\
&= P(X_1^d = x_1^d)\,.\,\prod_{i=1}^{k} P(X_{n-i+1} = x_{n-i+1} | X_1^{n-i} = x_1^{n-i}),\ \ n-k=d.
\end{aligned}
$$

Assumindo que $P(X_1^d = x_1^d) = 1$, e usando a função $N_n(w, a)$, tem-se

$$L(x_1^n, \mathcal{T}) = \prod_{w \in \mathcal{T}} \prod_{a \in \mathcal{A}} p(a|w)^{N_n(w,a)}. \tag{2.2.3}$$

Dessa forma, é possível provar que o estimador de máxima verossimilhança das probabilidades de transição é dado por

$$\hat{p}_n(a|w) = \frac{N_n(w,a)}{N_n(w)},\ \ N_n(w) > 0.$$

E qualquer distribuição em $\mathcal{A}$, se $N_n(w) = 0$ (Galves; Leonardi; Ost, 2022). Nesse caso, pode-se considerar que $\hat{p}_n(a|w) = \frac{1}{|\mathcal{A}|}$. Substituindo esse estimador em (2.2.3), tem-se que a máxima verossimilhança da amostra X_1^n, condicionada em X_1^d, para a árvore de contextos é

$$\hat{\mathrm{ML}}(x_1^n, \mathcal{T}) = \prod_{w \in \mathcal{T}} \prod_{a \in \mathcal{A}} \hat{p}_n(a|w)^{N_n(w,a)}. \tag{2.2.4}$$

Exemplo 2.3 *Considere uma cadeia de ordem variável* $\boldsymbol{X}$ *com elementos no alfabeto* $\mathcal{A} = \{0, 1\}$ *e com a árvore de contextos* $\mathcal{T}$ *com* $h(\mathcal{T}) \leq d$. *Para facilitar, considere* $d = \ell(w)$ *nas equações 2.2.1 e 2.2.2. Seja* 00010110100110 *uma amostra desse processo, calcule* $\hat{p}_n(1|01)$ *e* $\hat{p}_n(0|00)$.

1. $N_{14}(01, 1) = 2$ *e* $N_{14}(01) = 4$, *logo* $\hat{p}_{14}(1|01) = 2/4 = 1/2$;
2. $N_{14}(00, 0) = 1$ *e* $N_{14}(00) = 3$, *logo* $\hat{p}_{14}(0|00) = 1/3$.

Visto isso, agora é possível fazer inferência do modelo sob as árvores de contextos. Entretanto, na prática, não haverá o conhecimento dessa árvore que modela todo o problema. Portanto, o próximo objetivo é definir um conjunto de árvores que sejam factíveis. Para tanto, é preciso definir o que é uma árvore factível.

Definição 2.6 *Para uma árvore $\mathcal{T}$ com profundidade $h(\mathcal{T}) = d$ e amostra X_1^n, tal que $1 \leq d \leq n$. Essa árvore será factível, se*

1. $s \in \mathcal{T} \iff N_n(s) \geq 1$ *e* $\ell(s) \leq d$;
2. *Para qualquer sequência $w \in \mathcal{A}^*$, tal que $N_n(w) \geq 1$. Então, $w \preceq u$ ou $u \prec w$, $u \in \mathcal{T}$.*

O conjunto de todas as árvores factíveis é dado por $\mathcal{F}_n$.

2.3 Métodos de Estimação da Árvore de Contextos

Nesta seção serão abordados os métodos de seleção de árvores de contextos para o processo $\mathbf{X}$, como o Algoritmo Contexto e a Máxima Verossimilhança Penalizada. Esses critérios de seleção se baseiam, de forma simplória, em diferentes maneiras de se podar uma árvore, até que se tenha a árvore de contextos estimada mais próxima da real.

2.3.1 Algoritmo Contexto

O Algoritmo Contexto (AC) foi introduzido também em Rissanen (1983). Esse método considera a discrepância entre as probabilidades de transição associadas ao maior sufixo de um contexto e as probabilidades de transição associadas à concatenação de um elemento do alfabeto com esse maior sufixo, de maneira que, se essa discrepância for maior que um valor de corte predefinido, então os contextos serão mantidos. Caso contrário, eles serão podados.

O algoritmo parte da maior árvore factível e utiliza a divergência de Kullback-Leibler para construir a medida de discrepância entre o maior sufixo e seu descendentes. A divergência de Kullback-Leibler entre duas medidas de probabilidade p e q em $\mathcal{A}$, é definida por

$$D(p;q) = \sum_{a \in \mathcal{A}} p(a) \log \frac{p(a)}{q(a)};$$

em que, por convenção, $p(a) \log \frac{p(a)}{q(a)} = 0$, se $p(a) = 0$; e $p(a) \log \frac{p(a)}{q(a)} = +\infty$, se $q(a) = 0$ para $p(a) > 0$.

O conjunto de todas as sequências $w \in \mathcal{A}^*$ que aparecem ao menos uma vez na amostra é denotado por $\mathcal{V}_n$.

$$\mathcal{V}_n = \{w \in \mathcal{A}^* \mid N_n(w) \geq 1\}. \tag{2.3.1}$$

Desse modo, é possível definir a medida de discrepância $\Delta_n(w)$, para $w \in \mathcal{V}_n$.

$$\Delta_n(w) = \sum_{b\,:\,bw \in \mathcal{V}_n} N_n(bw) D(\hat{p}_n(.|bw); \hat{p}_n(.|w)), \quad b \in \mathcal{A}. \tag{2.3.2}$$

O valor de corte predefinido é dado por δ_n e para a consistência do algoritmo contexto, é necessária uma propriedade assintótica, em que, $(\delta_n)_{n\in\mathbb{N}}$ é uma sequência de números reais positivos, de tal modo que $\delta_n \to +\infty$ e $\delta_n/n \to 0$ quando $n \to +\infty$.

Agora, o algoritmo contexto pode ser descrito em 3 passos:

- Passo 1: obter a maior árvore factível $\mathcal{T}_{max} = \mathcal{T}_0$ com base na amostra.
- Passo 2: aplicar $\Delta_n(suf(w))$ para todo $w \in \mathcal{T}_0$, tal que $N_n(suf(w)) \geq 1$. Dessa maneira, se $\Delta_n(suf(w)) \leq \delta_n$, então todos os descendentes de $suf(w)$ são substituídos por $suf(w)$, isso é, a árvore foi podada. Caso contrário, mantêm-se os contextos. A árvore obtida nesse passo é denotada por $\mathcal{T}_1$.
- Passo 3: repetir o passo 2 para $\mathcal{T}_i$, $i \in \{1, 2, ...\}$, gerando a árvore $\mathcal{T}_{i+1}$. Até que não seja mais possível podar a árvore.

A árvore final estimada, é denotada por $\hat{\mathcal{T}}_c(x_1^n)$. No Passo 2, Rissanen considerava $N_n(suf(w)) > 1$. Essa árvore também pode ser obtida com auxílio da função recursiva $C_w(x_1^n)$, definida a seguir.

$$C_w(x_1^n) = \begin{cases} \max\{\mathbb{1}\{\Delta_n(w) > \delta_n\}, \max\limits_{b\in\mathcal{A}} C_{bw}(x_1^n)\}, & N_n(w) \geq 1 \text{ ou } l(w) < d; \\ 0, & N_n(w) < 1 \text{ e } l(w) \geq d. \end{cases} . \tag{2.3.3}$$

Definição 2.7 *A árvore $\hat{\mathcal{T}}_c(x_1^n)$ estimada pelo algoritmo contexto pode ser obtida pelo conjunto de sequências dadas por*

$$\hat{\mathcal{T}}_c(x_1^n) = \{w \in \mathcal{V}_n \mid C_w(x_1^n) = 0 \text{ e } C_u(x_1^n) = 1, \text{ para todo } u \prec w\}.$$

De maneira simples, uma sequência w pertence à árvore se não há ganho em olhar o seu passado, isso é, $C_w(x_1^n) = 0$. Além disso, todo u sufixo de w deve ser mantido, pois a medida de discrepância é significativamente alta, isso é, $C_u(x_1^n) = 1$.

O algoritmo contexto é consistente para qualquer sequência δ_n, tal que $\frac{\delta_n}{n} \to 0$, quando $n \to +\infty$. Além disso, $\frac{\delta_n}{\log\log n} > c_0$, para todo n, em que c_0 é uma constante dependendo do processo. Como essa constante é desconhecida, a escolha mais comum é usar $\delta_n = c \log n$, para $c > 0$. A consistência do Algoritmo Contexto e outros resultados podem ser vistos em Galves, Leonardi e Ost (2022).

O Algoritmo Contexto estudado até aqui foi o proposto em Rissanen (1983). Porém, outras versões também foram propostas, como em Galves e Leonardi (2008). Esse algoritmo utiliza uma medida de discrepância diferente, dada por

$$\Delta_n(w) = \max_{a\in\mathcal{A}} |\hat{p}_n(a|w) - \hat{p}_n(a|suf(w))|, \quad w \in \mathcal{A}^*. \tag{2.3.4}$$

Definição 2.8 *Dado $\delta > 0$ e $d < n$, a árvore estimada pelo Algoritmo Contexto em Galves e Leonardi (2008) é dada por*

$$\hat{\mathcal{T}}_n^{\delta,d} = \Bigg\{ w \in \bigcup_{k=1}^{d} \mathcal{A}^k \;\Big|\; \Delta_n(a\,suf(w)) > \delta \text{ para algum } a \in \mathcal{A} \text{ e } \Delta_n(uw) \le \delta \text{ para todo } u \in \bigcup_{k=1}^{d-\ell(w)} \mathcal{A}^k, N_n(uw) \ge 1 \Bigg\}.$$

Caso $\ell(w) = d$, então $\bigcup_{k=1}^{d-\ell(w)} \mathcal{A}^k = \emptyset$.

É importante salientar que as probabilidades de transição utilizadas em (2.3.4) são as probabilidades empíricas, dadas por $\hat{p}_n(a|w) = \frac{N(w,a)+1}{N(w)+|\mathcal{A}|}$.

Mais detalhes, como demonstrações de consistência dos algoritmos, podem ser vistos em Galves, Leonardi e Ost (2022).

2.3.2 Máxima Verossimilhança Penalizada

É comum utilizar critérios de Máxima Verossimilhança Penalizada (PML) para escolha de modelos estatísticos, como o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiana (BIC). Nesse trabalho, será estudado o BIC, além da possibilidade de generalizar a ideia a partir da função de penalidade $pen(n)$.

Considerando o conjunto de todas as árvores factíveis $\mathcal{F}_n$, a árvore escolhida pela máxima verossimilhança penalizada para a amostra x_1^n é dada por

$$\hat{\mathcal{T}}_{\mathrm{PML}}(x_1^n) = \arg\min_{\mathcal{T}\in\mathcal{F}_n} \Big\{ -\log \hat{\mathrm{ML}}(x_1^n, \mathcal{T}) \;+\; |\mathcal{T}|\, pen(n) \Big\}. \tag{2.3.5}$$

O BIC introduzido em Csiszár e Talata (2006), considera $pen(n) = \frac{(|\mathcal{A}|-1)}{2}\log n$.

A ideia do BIC é penalizar a máxima verossimilhança a partir do número de parâmetros livres do modelo vezes $\log n$. Para uma árvore completa, o número de parâmetros livres é $(|\mathcal{A}|-1)$. Entretanto, para árvores incompletas, o número de parâmetros livres deve ser menor, pois algumas probabilidades de transição devem ser iguais a zero. Portanto, Csiszár e Talata (2006) explica que por esses motivos o BIC não é exatamente o da literatura.

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

A árvore estimada através do BIC é a que obtiver o menor valor de $\mathrm{BIC}_{\mathcal{T}}(x_1^n)$, dentre todas as árvores factíveis. Porém, aplicar o BIC em todas as árvores factíveis pode ser impossível. Por isso, em Csiszár e Talata (2006) foi proposta uma maneira eficiente de obter o estimador via BIC, descrita a seguir.

Primeiro, para $w \in \mathcal{V}_n$, definido em (2.3.1), tem-se:

$$\hat{\mathrm{ML}}(x_1^n, w) = \prod_{a\in\mathcal{A}} \hat{p}_n(a|w)^{N_n(w,a)}. \tag{2.3.6}$$

Dessa forma, é possível escrever a função de máxima verossimilhança, a partir de (2.2.3) e (2.3.6). De modo que,

$$\hat{\mathrm{ML}}(x_1^n, \mathcal{T}) = \prod_{w\in\mathcal{T}} \hat{\mathrm{ML}}(x_1^n, w).$$

Essa manipulação algébrica permite decompor a função de máxima verossimilhança penalizada, definida em (2.3.5), da seguinte maneira

$$\begin{aligned} -\log \hat{\mathrm{ML}}(x_1^n, \mathcal{T}) \; + \; |\mathcal{T}|\, pen(n) &= -\log \prod_{w\in\mathcal{T}} \hat{\mathrm{ML}}(x_1^n, w) + |\mathcal{T}|\, pen(n) \\ &= \sum_{w\in\mathcal{T}} \big[-\log \hat{\mathrm{ML}}(x_1^n, w) + pen(n) \big]. \end{aligned}$$

A partir disso, é possível definir a função recursiva $V_w(x_1^n)$ e a função indicadora $\chi_w(x_1^n)$, para $w \in \mathcal{V}_n$, definido em (2.3.1).

$$V_w(x_1^n) = \begin{cases} \max\Big\{ e^{-pen(n)} \hat{\mathrm{ML}}(x_1^n, w), \prod_{b\in\mathcal{A}: bw\in\mathcal{V}_n} V_{bw}(x_1^n) \Big\}, & l(w) < d; \\ e^{-pen(n)} \hat{\mathrm{ML}}(x_1^n, w), & l(w) \geq d. \end{cases} ;$$

$$\chi_w(x_1^n) = \begin{cases} \mathbb{1}\Big\{ \prod\limits_{b\in\mathcal{A}:bw\in\mathcal{V}_n} V_{bw}(x_1^n) > e^{-pen(n)}\hat{\mathrm{ML}}(x_1^n, w)\Big\}, & l(w) < d; \\ 0, & l(w) \geq d. \end{cases} .$$

Por convenção, caso $\{b \in \mathcal{A} \,|\, bw \in \mathcal{V}_n\} = \emptyset$, então $V_w(x_1^n) = e^{-pen(n)}\hat{\mathrm{ML}}(x_1^n, w)$ e $\chi_w(x_1^n) = 0$. A partir dessas duas funções, é possível definir a árvore de contextos $\hat{\mathcal{T}}_{\mathrm{PML}}(x_1^n)$ estimada a partir da máxima verossimilhança penalizada.

$$\hat{\mathcal{T}}_{\mathrm{PML}}(x_1^n) = \{w \in \mathcal{V}_n \mid \chi_w(x_1^n) = 0 \ \ e \ \ \chi_u(x_1^n) = 1, \ \ para \ todo \ u \prec w\}\cdot \tag{2.3.7}$$

Esses resultados foram provados em Csiszár e Talata (2006) e podem ser revistos em Galves, Leonardi e Ost (2022). Perceba que, essa definição é similar ao Algoritmo Contexto, definido em (2.7). De fato, elas possuem uma relação interessante quando δ_n, valor de corte do Algoritmo Contexto, é menor que $pen(n)$. Em Garivier e Leonardi (2011), foi demonstrada seguinte proposição.

Proposição 2.9 *Para $n \geq 1$ e para todas as sequências x_1^n, se $\delta_n \leq pen(n)$, então*

$$\hat{\mathcal{T}}_{PML}(x_1^n) \preceq \hat{\mathcal{T}}_c(x_1^n).$$

Isso é, para toda sequência $v \in \hat{\mathcal{T}}_c(x_1^n)$, existe uma sequência $w \in \hat{\mathcal{T}}_{\mathrm{PML}}(x_1^n)$, tal que $w \preceq v$. Então, a árvore $\hat{\mathcal{T}}_c(x_1^n)$ será maior ou igual a $\hat{\mathcal{T}}_{\mathrm{PML}}(x_1^n)$, nessas condições.

3 Metodologia

Para estudar as Cadeias de Ordem Variável, serão abordados os métodos de seleção de modelos, isso é, de estimação das árvores de contextos, em diversos âmbitos. Desse modo, será apresentado um estudo de simulação para analisar o desempenho dos estimadores estudados. Posteriormente, será analisado o comportamento dessas técnicas em dados reais.

O primeiro critério de seleção a ser utilizado é o Algoritmo Contexto, introduzido em Rissanen (1983). Esse método considera a discrepância entre as probabilidades de transição associadas ao maior sufixo de um contexto e as probabilidades de transição associadas à concatenação de um elemento do alfabeto com esse maior sufixo. Para isso, é utilizada uma função recursiva. Além desse algoritmo, a Máxima Verossimilhança Penalizada, muito conhecida e utilizada entre diversas áreas da estatística, também pode ser utilizada como critério de seleção. Em Csiszár e Talata (2006) foi proposto um algoritmo para estimar o BIC em árvores de contexto. Essa ideia foi generalizada para outras penalizações em função do tamanho da amostra. Esses métodos de estimação foram explicados com mais detalhes na Seção 2.3.

As simulações e as estimativas foram feitas através do Software R, versão 4.3.1, (R Core Team, 2023). Foram feitas funções para o Algoritmo Contexto e para a Máxima Verossimilhança Penalizada. Desta maneira, os resultados podem ser replicados em outras máquinas, os códigos estão na Seção 5. Para as simulações, foi utilizada programação paralela, com o pacote "parallel" (R Core Team, 2023), em que a ideia é dividir a tarefa computacional quando essas são independentes, o que é muito comum em estudos de simulação. Dividir uma tarefa grande em menores tarefas, geralmente, ajuda no tempo de processamento.

O estudo de aplicação abordará duas áreas distintas: a primeira utilizando dados climáticos e a segunda utilizando dados financeiros. Para a primeira aplicação, os dados podem ser acessados em (INMET, 2024). O Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) possui dados meteorológicos diários em forma digital, conforme as normas técnicas internacionais da Organização Meteorológica Mundial, segundo o (INMET, 2024). Já a segunda aplicação, os dados financeiros foram importados do pacote "quantmod" (RYAN; ULRICH, 2024), que reúne dados de diversas fontes do mercado financeiro.

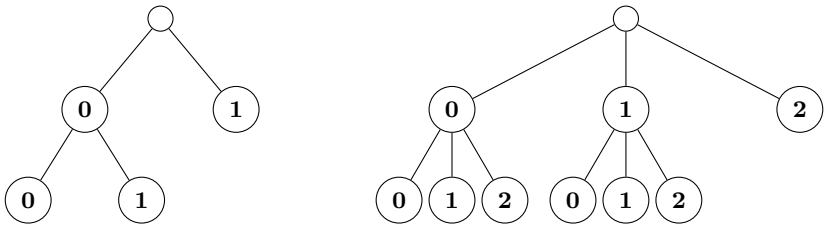
4 Resultados

Neste capítulo, serão analisados os resultados do estudo de simulação e do estudo de aplicações. Para isso, serão explorados os métodos de estimação com detalhes no estudo de simulação, a fim de verificar a correta implementação dos códigos e a assertividade dos métodos estudados, além de fornecer orientações para a aplicação prática. No estudo de aplicações, será observado o comportamento do modelo para contextos reais, de modo que os resultados possam ser úteis nas respectivas áreas estudadas. A primeira aplicação abordará o desastre climático no Rio Grande do Sul, ocorrido em maio de 2024; a segunda aplicação analisará a capacidade preditiva do modelo em dados financeiros.

4.1 Estudo de Simulação

Neste estudo de simulação será analisado o desempenho do Algoritmo Contexto e da Máxima Verossimilhança Penalizada em quatro Árvores de Contextos com características diferentes. Além disso, os dados foram simulados em quatro tamanhos de amostra diferentes: 500, mil, 5 mil e 10 mil. Dessa maneira, é possível compreender melhor o comportamento dos algoritmos em diferentes situações.

As duas primeiras árvores são mais simples, ou seja, são menores ou possuem um alfabeto menor. Porém, isso não significa que são mais fáceis de serem estimadas. Para a primeira árvore, considere o alfabeto $\mathcal{A} = \{0, 1\}$ e a Árvore $\mathcal{T}_1 = \{00, 10, 1\}$; para a segunda árvore, o alfabeto $\mathcal{A} = \{0, 1, 2\}$ e a árvore $\mathcal{T}_2 = \{00, 10, 20, 01, 11, 21, 2\}$. Logo, $h(\mathcal{T}_1) = h(\mathcal{T}_2) = 2$, porém, a Árvore 2 possui um alfabeto maior. A representação gráfica dessas duas árvores está na Figura 4.1.

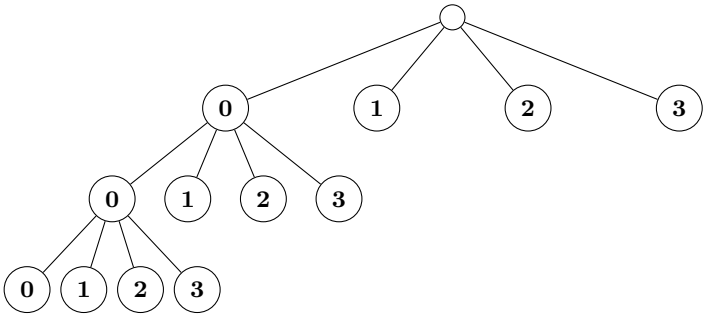


(a) Árvore 1 (b) Árvore 2

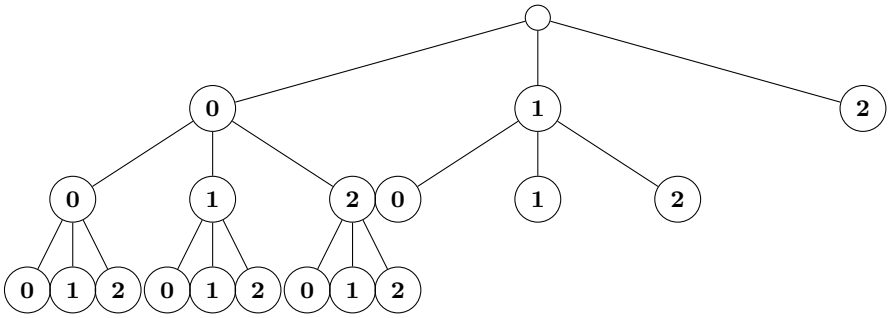
Figura 4.1: Árvores de Contextos 1 e 2.

Já a terceira e quarta Árvores de Contextos são maiores, como pode ser visto na Figura 4.2. A Árvore 3 possui um alfabeto maior $\mathcal{A} = \{0, 1, 2, 3\}$, além de maior profun-

didade: $\mathcal{T}_3 = \{000, 100, 200, 300, 10, 20, 0, 1, 2, 3\}$. Por último, para a quarta árvore, considere o alfabeto $\mathcal{A} = \{0, 1, 2\}$ e $\mathcal{T}_4 = \{000, 100, 200, 010, 110, 210, 020, 120, 220, 01, 11, 21, 2\}$, sendo a maior árvore das quatro, com treze contextos.



(c) Árvore 3



(d) Árvore 4

Figura 4.2: Árvores de Contextos 3 e 4.

4.2 Simulação de Monte Carlo

Para cada árvore e para cada tamanho de amostra foram geradas mil simulações de Monte Carlo. Desse modo, foi possível calcular a taxa de acertos de cada algoritmo. Considere N o número de simulações, isso é, $N = 1000$, e t_x a taxa de acertos. Assim, são geradas N estimativas da Árvore de Contextos $\mathcal{T}$: $\hat{\mathcal{T}}^{(1)}, \hat{\mathcal{T}}^{(2)}, ..., \hat{\mathcal{T}}^{(N)}$. Desse modo, é calculada a taxa

$$t_x = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\hat{\mathcal{T}}^{(i)} = \mathcal{T}\}.$$

A fim de analisar também a penalidade $pen(n)$ no estimador de Máxima Verossimilhança Penalizada, foram propostas 4 penalidades: $pen(n) = 2k$, que seria equivalente ao AIC; $pen(n) = k \log(n)$, o BIC; além de outras duas tentativas inspiradas no comporta-

mento das outras penalidades, $pen(n) = k\log(\log(n))$ e $pen(n) = k\sqrt{\log(n)}$. Além disso, o limiar δ_n no Algoritmo Contexto também foi estudado, seguindo os últimos trabalhos na área que indicam $\delta_n = c\log(n)$, para $c > 0$. Foram propostos dois limiares: $\delta_n = k\log(n)$; e um pouco maior $\delta_n = 1{,}5k\log(n)$. Considerando $k = \frac{(|\mathcal{A}|-1)}{2}$ para ambos métodos de seleção.

Tabela 4.1: Taxa de acertos por árvore e tamanho de amostra (n).

n	Árvore	$pen(n)$				δ_n	
		$2k$	$k\log(n)$	$k\log(\log(n))$	$k\sqrt{\log(n)}$	$k\log(n)$	$1{,}5\,k\log(n)$
500	Árvore 1	0,455	0,170	0,470	0,395	0,140	0,060
	Árvore 2	0,725	0,125	0,690	0,755	0,625	0,255
	Árvore 3	0,395	0,000	0,330	0,370	0,365	0,300
	Árvore 4	0,100	0,000	0,100	0,055	0,005	0,000
1000	Árvore 1	0,600	0,475	0,595	0,615	0,405	0,220
	Árvore 2	0,855	0,445	0,840	0,935	0,925	0,705
	Árvore 3	0,585	0,000	0,540	0,860	0,630	0,675
	Árvore 4	0,300	0,000	0,275	0,260	0,130	0,010
5000	Árvore 1	0,575	0,965	0,595	0,635	0,990	0,900
	Árvore 2	0,925	1,000	0,925	0,965	1,000	1,000
	Árvore 3	0,715	1,000	0,805	0,965	0,950	1,000
	Árvore 4	0,535	0,675	0,595	0,870	0,980	0,895
10000	Árvore 1	0,620	0,985	0,660	0,730	0,995	1,000
	Árvore 2	0,895	1,000	0,915	0,980	1,000	1,000
	Árvore 3	0,735	1,000	0,860	0,980	0,975	1,000
	Árvore 4	0,550	1,000	0,690	0,910	0,985	1,000

A partir da Tabela 4.1, é possível perceber que para a amostra de tamanho 500, a $pen(n) = k\log(n)$, ou seja, o critério de informação bayesiano (BIC), não apresentou resultados satisfatórios. Embora o Algoritmo Contexto tenha superado o BIC, não conseguiu superar os demais métodos. Para tamanho mil, o comportamento dos métodos foi parecido. Já para 5 mil e 10 mil, o comportamento parece mudar, o BIC e o Algoritmo Contexto passaram a ter ótimos resultados. Enquanto $pen(n) = 2k$ se manteve estável, o BIC e o Algoritmo Contexto passaram a ter quase 100% de acertos para todas as árvores.

Para observar melhor o comportamento dos métodos para cada árvore e para cada tamanho amostral, segundo a penalização e o limiar, será analisado o gráfico apresentado na Figura 4.3.

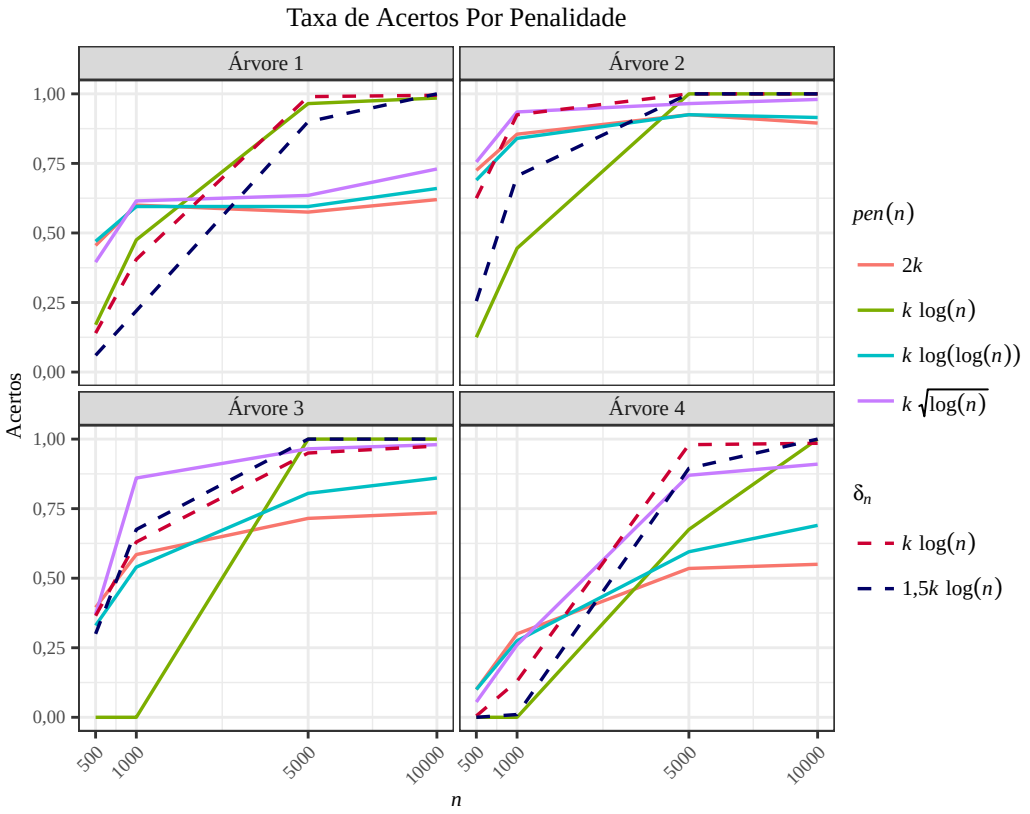


Figura 4.3: Taxa de acertos por penalidade e método de estimação.

Percebe-se o comportamento assintótico de assertividade dos algoritmos, independente da penalidade e do limiar. A primeira árvore, apesar de ser a menor, isso é, com menos contextos, apresentou resultados que negam sua simplicidade e reforçam a análise anterior.

Já a segunda árvore obteve bons resultados desde a menor amostra, exceto para o BIC e o limiar $\delta_n = 1{,}5k\log(n)$. No entanto, à medida que o tamanho da amostra aumentou, esses dois métodos superaram os demais.

A terceira árvore, a com maior alfabeto, teve maior variação a depender da penalidade e do limiar. A $pen(n) = k\sqrt{\log(n)}$ obteve bons resultados para todos os tamanhos amostrais, mas foi superado pelo BIC e pelo Algoritmo Contexto a partir de $n = 5000$.

Os resultados para a quarta árvore seguiram o mesmo padrão. No entanto, é evidente a dificuldade de se estimar esta que é a maior árvore, isso é, com mais contextos. O BIC que conseguia excelentes taxas de acertos a partir de $n = 5000$, conseguiu apenas em 10 mil. Um destaque positivo foi o Algoritmo Contexto com $\delta_n = k\log(n)$.

De forma geral, os algoritmos demonstraram um bom desempenho, consistente e alinhado com o comportamento assintótico esperado. Como complemento do estudo,

para analisar possíveis falhas, será realizada uma análise das probabilidades de transição estimadas nas simulações.

Desse modo, será analisado o Viés e a Raiz do Erro Quadrático Médio (RMSE) das probabilidades estimadas nas simulações. O Viés é a diferença entre o parâmetro a ser estimado e a média das estimativas feitas por meio das simulações. Ele indica uma tendência consistente das estimativas em se afastarem do valor real do parâmetro; por isso, espera-se valores próximos de zero.

Considere $\theta = p(a|w)$, a partir das N simulações de Monte Carlo, tem-se $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N$ e $\bar{\hat{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$, então

$$\text{Viés} = \theta - \bar{\hat{\theta}}.$$

O RMSE avalia a precisão das estimativas em relação ao parâmetro verdadeiro. Ele é calculado como a raiz quadrada da média dos quadrados das diferenças entre o parâmetro verdadeiro e suas estimativas; também é esperado valores próximos de zero.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta - \hat{\theta}_i)^2}.$$

Visto isso, as Tabelas 4.2, 4.3, 4.4 e 4.5 possuem as informações do Viés e RMSE para o tamanho de amostra 10 mil. Seguindo a lógica de uma matriz de transição, as linhas são as sequências "w" e as colunas o elemento "a".

Também será analisado o comportamento dessas duas métricas para diferentes tamanhos amostrais, a fim de investigar o comportamento assintótico, através dos gráficos apresentados nas Figuras 4.4, 4.5, 4.6 e 4.7.

Tabela 4.2: Viés e RMSE - Árvore 1.

	Viés		RMSE	
	0	**1**	**0**	**1**
00	-0,0008	0,0008	0,0105	0,0105
1	0,0001	-0,0001	0,0043	0,0043
10	0,0007	-0,0007	0,0069	0,0069

A partir da Tabela 4.2, é possível concluir que as probabilidades estimadas da árvore 1 foram próximas das verdadeiras probabilidades. O Viés foi próximo de zero, como deve ser em estimadores não-viesados e o RMSE indica um erro médio baixo em todas as simulações, também próximo de zero.

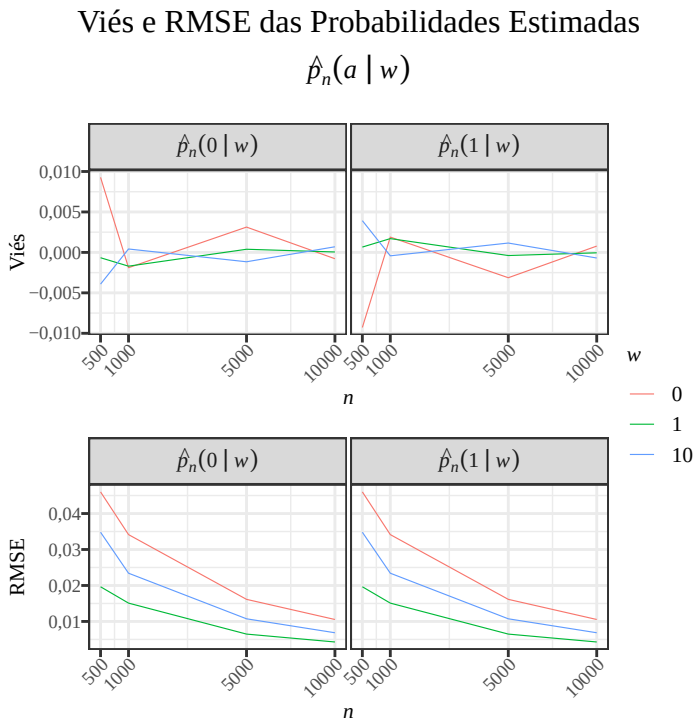


Figura 4.4: Viés e RMSE por tamanho de amostra - Árvore 1.

O gráfico da Figura 4.4 mostra uma clara tendência de redução do RMSE quando n cresce. Apesar do Viés e RMSE maiores para menores tamanhos de amostra, também foram baixos e próximos de zero.

Tabela 4.3: Viés e RMSE - Árvore 2.

	Viés			RMSE		
	0	**1**	**2**	**0**	**1**	**2**
00	0,0002	0,0003	-0,0005	0,0122	0,0089	0,0132
01	0,0000	-0,0007	0,0007	0,0208	0,0171	0,0201
10	0,0009	-0,0020	0,0011	0,0122	0,0129	0,0121
11	0,0006	0,0009	-0,0015	0,0164	0,0131	0,0156
2	-0,0003	-0,0001	0,0005	0,0066	0,0099	0,0081
20	0,0006	-0,0012	0,0005	0,0170	0,0117	0,0148
21	0,0003	-0,0006	0,0004	0,0118	0,0114	0,0103

De maneira análoga, a Tabela 3 apresenta resultados satisfatórios para as estimativas das probabilidades de transição da Árvore 2, com Viés próximo de zero e RMSE baixo.

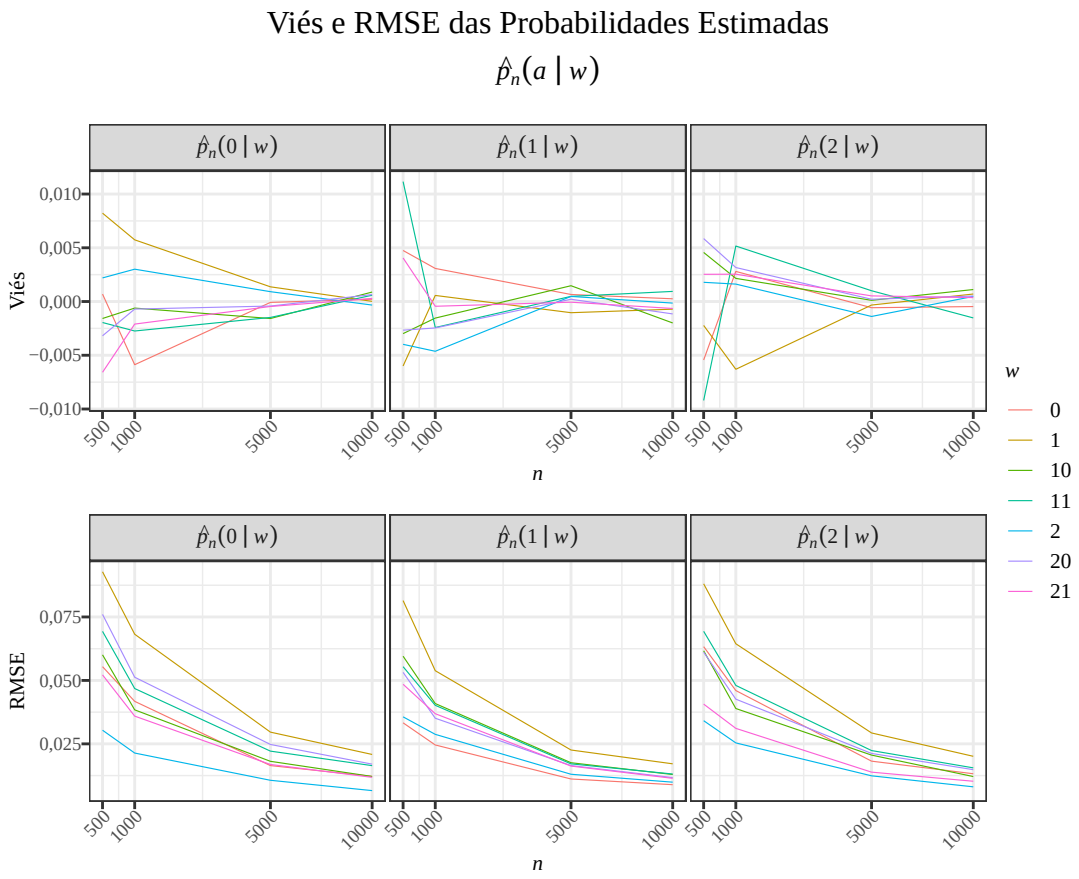


Figura 4.5: Viés e RMSE por tamanho de amostra - Árvore 2.

Para a Árvore 2, torna-se evidente o comportamento assintótico de assertividade do estimador. À medida que o tamanho da amostra aumenta, tanto o viés quanto o RMSE diminuem consideravelmente.

Tabela 4.4: Viés e RMSE - Árvore 3.

	Viés				RMSE			
	0	**1**	**2**	**3**	**0**	**1**	**2**	**3**
000	0,0021	-0,0009	-0,0003	-0,0009	0,0331	0,0204	0,0372	0,0208
1	-0,0000	0,0008	-0,0007	-0,0001	0,0090	0,0083	0,0095	0,0062
10	0,0014	-0,0006	-0,0005	-0,0003	0,0193	0,0156	0,0163	0,0121
100	0,0012	-0,0018	-0,0006	0,0011	0,0237	0,0274	0,0195	0,0227
2	-0,0012	-0,0002	0,0009	0,0005	0,0104	0,0064	0,0080	0,0072
20	-0,0006	-0,0006	0,0006	0,0006	0,0130	0,0143	0,0117	0,0113
200	0,0007	-0,0013	0,0025	-0,0019	0,0323	0,0210	0,0353	0,0189
3	0,0000	-0,0010	0,0001	0,0009	0,0089	0,0106	0,0098	0,0108
30	-0,0006	-0,0001	0,0002	0,0005	0,0218	0,0273	0,0150	0,0250
300	0,0002	-0,0031	0,0037	-0,0008	0,0492	0,0439	0,0359	0,0561

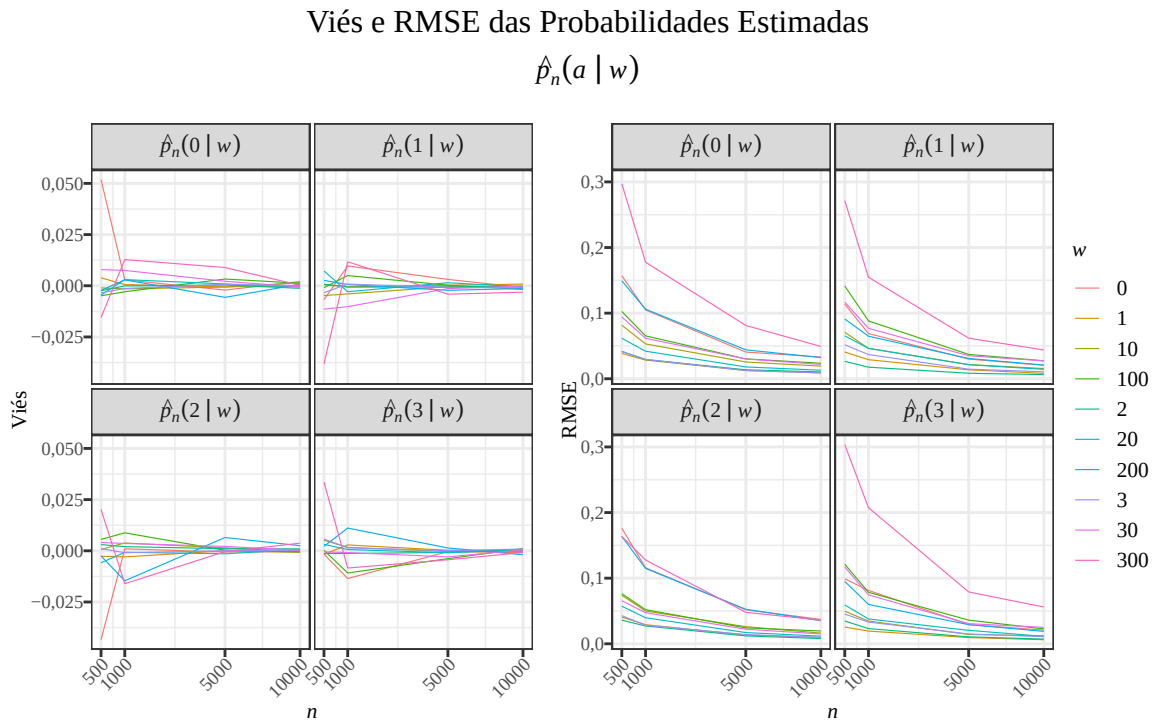


Figura 4.6: Viés e RMSE por tamanho de amostra - Árvore 3.

O comportamento para a Árvore 3 foi análogo; no entanto, o RMSE foi mais alto para amostras menores. Através da Tabela 4.4 e do respectivo gráfico, percebe-se uma dificuldade em estimar as probabilidades de transição $\hat{p}_n(a|w = 300)$.

Tabela 4.5: Viés e RMSE - Árvore 4.

	Viés			RMSE		
	0	**1**	**2**	**0**	**1**	**2**
000	0,0025	-0,0003	-0,0022	0,0319	0,0186	0,0308
01	0,0014	0,0012	-0,0026	0,0169	0,0126	0,0174
010	0,0000	0,0012	-0,0012	0,0221	0,0216	0,0217
020	0,0051	-0,0027	-0,0023	0,0327	0,0291	0,0322
100	0,0003	-0,0005	0,0002	0,0204	0,0275	0,0236
11	0,0000	-0,0006	0,0005	0,0194	0,0125	0,0172
110	0,0005	-0,0002	-0,0003	0,0249	0,0254	0,0199
120	-0,0019	0,0013	0,0005	0,0313	0,0270	0,0238
2	0,0002	-0,0006	0,0004	0,0065	0,0087	0,0079
200	-0,0000	0,0017	-0,0017	0,0259	0,0183	0,0278
21	-0,0005	0,0003	0,0003	0,0121	0,0103	0,0126
210	-0,0010	0,0004	0,0007	0,0184	0,0218	0,0212
220	-0,0035	0,0002	0,0033	0,0225	0,0259	0,0240

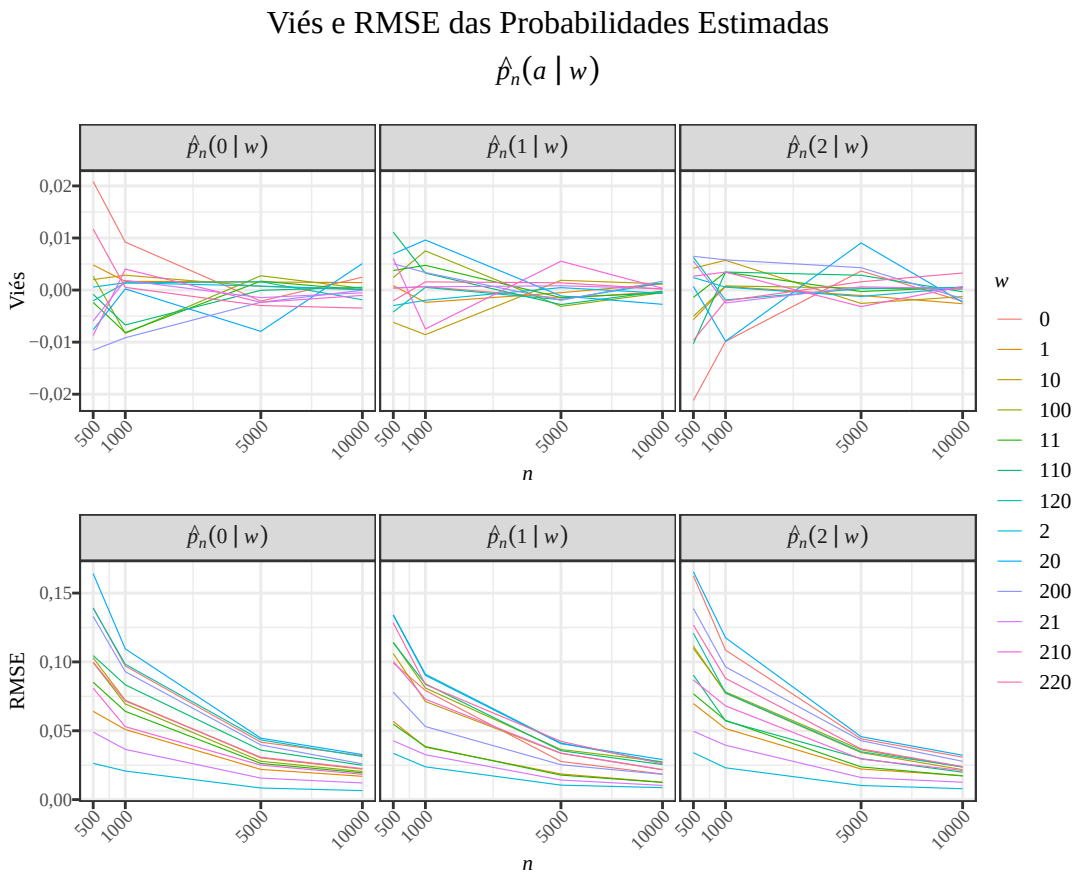


Figura 4.7: Viés e RMSE por tamanho de amostra - Árvore 4.

Por fim, a Árvore 4 obteve bons resultados, análogos às Árvores 1 e 2; o RMSE, que havia subido para a terceira árvore, voltou a ser mais baixo. Isso pode indicar que o tamanho do alfabeto seja uma dificuldade adicional para o estimador.

Em conclusão, a partir das simulações realizadas, torna-se evidente o poder tanto do Algoritmo Contexto quanto da Máxima Verossimilhança Penalizada. Além disso, ressalta-se a importância crítica da seleção cuidadosa do limiar e da penalidade. Observou-se também o comportamento assintótico dos métodos de estimação da árvore de contextos e do estimador das probabilidades de transição. Dessa maneira, é possível partir para um estudo de aplicação dos métodos estudados, visando sua implementação em contextos práticos e a análise de seu desempenho em cenários práticos com dados reais.

4.3 Estudo de Aplicações

Nesta seção, será analisado o comportamento das Cadeias Estocásticas de Ordem Variável em diferentes contextos reais. A primeira aplicação explorará dados climáticos

para investigar mudanças no padrão de precipitação no Rio Grande do Sul. Em seguida, será abordada a previsão do preço de ações no mercado financeiro. Esta abordagem visa destacar tanto a capacidade explicativa quanto a preditiva do modelo em cenários distintos.

4.3.1 Análise das Mudanças no Regime de Chuvas no Rio Grande do Sul

Em maio de 2024, o Rio Grande do Sul foi atingido pela maior tragédia socio-ambiental de sua história, destacando não apenas questões sociais, políticas e geológicas, mas também evidenciando os impactos das mudanças climáticas. O elevado volume de chuvas acarretou a inundação de diversas cidades do estado, inclusive a sua capital. Diante desses eventos devastadores, torna-se crucial analisar os dados climáticos para obter mais evidências das mudanças no clima e suas consequências para a região.

Dessa forma, serão aplicadas Cadeias Estocásticas de Ordem Variável para analisar mudanças nas árvores de contexto e nas probabilidades de transição ao longo dos últimos vinte anos. Este estudo utilizará dados de Precipitação Total Diária em Milímetros (mm), disponibilizados pelo INMET (2024), nas cidades gaúchas de Santa Maria e Caxias do Sul. A escolha de ambas cidades se deve à disponibilidade de dados mais abrangentes e completos, além disso, elas foram muito atingidas pelas chuvas. Santa Maria chegou a ser a cidade com maior volume de chuva durante o primeiro dia de maio:

> Santa Maria foi a cidade com o maior volume de chuva até as 14h desta quarta-feira (1º) em todo o mundo. Informações do Ogimet - site que reúne dados meteorológicos de vários centros mundiais -, apontavam três estações da cidade do centro gaúcho entre os 10 com maiores acumulados de água. (GaúchaZH, 2024b).

O volume de chuvas em Caxias do Sul no mês da tragédia foi o maior desde o início das medições:

> Choveu 845,3 milímetros em Caxias do Sul no mês de maio, de acordo com a estação do Instituto Nacional de Meteorologia (Inmet). É o mês com a maior precipitação dos últimos 93 anos - a medição ocorre desde 1931 no município. Convertendo a medida, o resultado é que 1,3 trilhão de litros de água caíram na cidade durante o mês. (GaúchaZH, 2024a).

Os gráficos presentes na Figura 4.8, mostram a Precipitação Diária para ambas cidades até o dia 22 de maio de 2024.

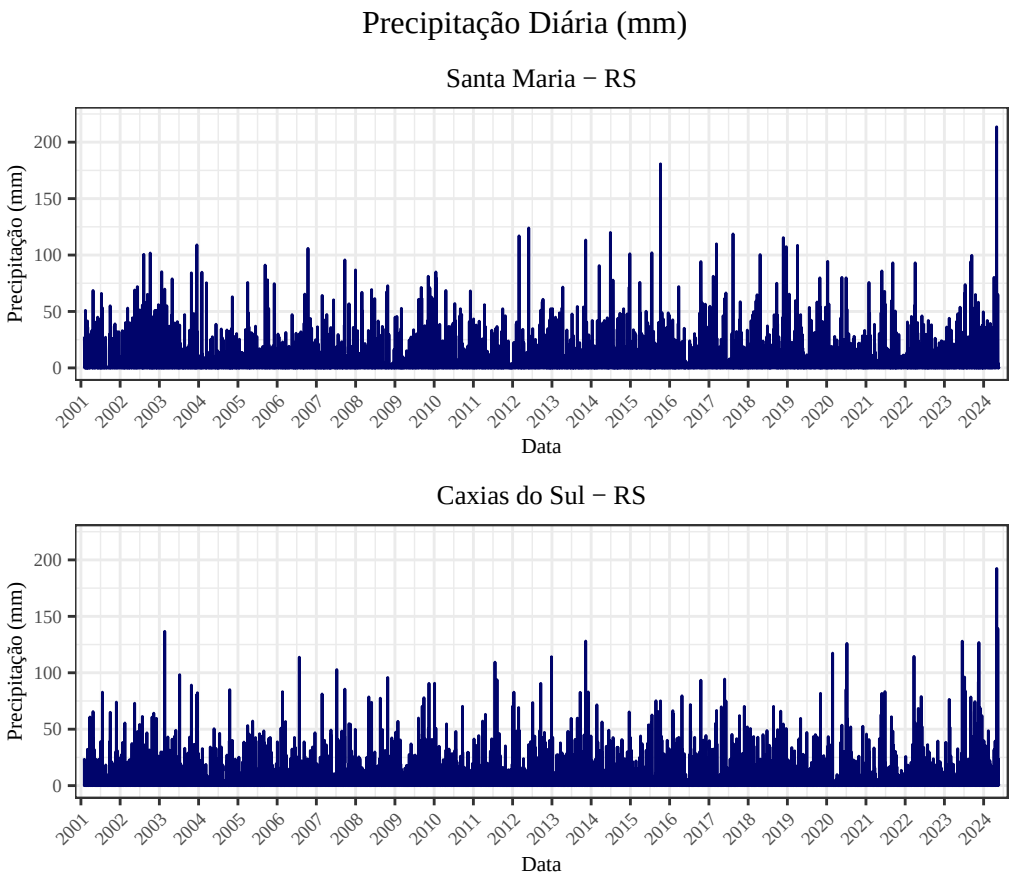


Figura 4.8: Precipitação Diária (2001-2024).

A análise dos gráficos revela que o maior volume diário de chuva registrado desde 2001 ocorreu em maio de 2024 para ambas as cidades. Além disso, percebe-se como a chuva está presente durante todo o ano, sem seguir um padrão sazonal claro.

Para a aplicação dos métodos, foi realizada a discretização da variável Precipitação, categorizando os níveis de chuva diária conforme apresentado na Tabela 4.6.

Tabela 4.6: Categorização da Precipitação Diária.

Classificação	**Intervalo**	**Alfabeto**
Chuva Fraca	[0; 2,5)	0
Chuva Moderada	[2,5; 10)	1
Chuva Forte	[10; 50)	2
Chuva Violenta	[50; ∞)	3

Para facilitar a interpretação e organização dos resultados, serão utilizados o intervalo e o alfabeto para representar a respectiva classificação.

A fim de analisar as mudanças no regime de chuvas na região, será comparado o

resultado da modelagem com dados de 2004 a 2014, com a modelagem com dados de 2014 a 2024. Isso é, serão comparados os últimos dez anos com os dez anos imediatamente anteriores. Os dados analisados foram até dia 22 de maio de 2024, após a tragédia. Essa análise pode ser interessante para detectar mudanças mais recentes e de curto prazo da precipitação no Rio Grande do Sul.

Com base na análise do estudo de simulação, foi escolhido $pen(n) = k\sqrt{\log(n)}$ e $\delta_n = k\log(n)$, pois esses parâmetros demonstraram um comportamento mais equilibrado para todos os tamanhos de amostra, além de proporcionarem boa taxa de acertos. Além disso, foi escolhido $d = 5$, pois os resultados tiveram no máximo essa profundidade.

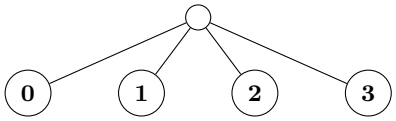


Figura 4.9: Árvore de Contextos (Santa Maria - RS, 2004-2014).

Para Santa Maria, no período de 2004 a 2024, tanto o Algoritmo Contexto quanto a Máxima Verossimilhança Penalizada, resultaram em uma Cadeia de Markov, representada pela Árvore na Figura 4.9. As probabilidades de transição também são de interesse e, por isso, foram representadas pela matriz de transição na Figura 4.10.

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0{,}82 & 0{,}07 & 0{,}10 & 0{,}01 \\ 0{,}72 & 0{,}11 & 0{,}15 & 0{,}02 \\ 0{,}62 & 0{,}15 & 0{,}19 & 0{,}04 \\ 0{,}49 & 0{,}21 & 0{,}25 & 0{,}05 \end{pmatrix} \end{array}$$

Figura 4.10: Matriz de transição (Santa Maria - RS, 2004-2014).

Percebe-se que a probabilidade de transição para Chuva Fraca é sempre a maior, independentemente da condição. Porém, quando há chuva violenta, essa probabilidade é menor.

Desse modo, serão comparados os resultados vistos com os da modelagem para o período de 2014 a 2024.

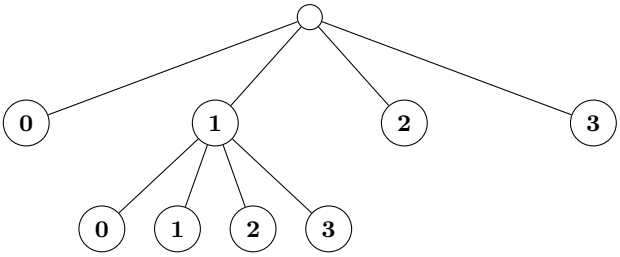


Figura 4.11: Árvore de Contextos PML (Santa Maria - RS, 2014-2024).

Como pode ser visto na Figura 4.11, a Árvore de Contextos estimada pela Máxima Verossimilhança Penalizada cresceu, indicando uma mudança no regime de chuvas na cidade. Basicamente, para se obter informações sobre a chuva amanhã, dado que hoje ocorreu chuva moderada, é necessário considerar também o nível de precipitação de ontem.

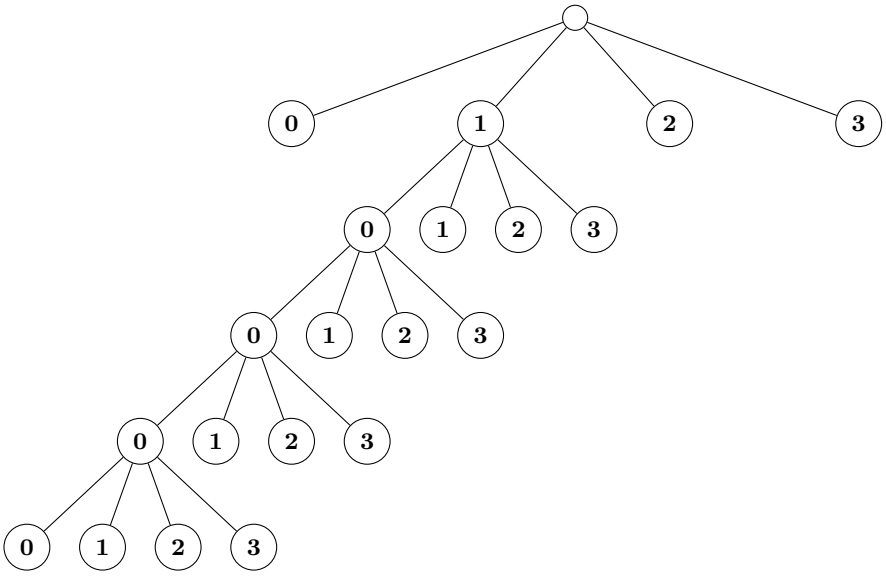


Figura 4.12: Árvore de Contextos AC (Santa Maria - RS, 2014-2024).

A Árvore estimada pelo Algoritmo Contexto, representada graficamente na Figura 4.12, segue a mesma ideia, sendo necessário considerar não apenas o ontem, mas os três últimos dias para se obter informações sobre a chuva amanhã. Essa característica é esperada no Algoritmo Contexto, que geralmente é mais sensível aos dados e pode gerar árvores mais complexas. Importante notar que ambos métodos obtiveram resultados semelhantes, apesar do Algoritmo Contexto considerar um passado mais distante.

As respectivas probabilidades de transição do período mostram um comportamento semelhante ao período passado, como pode ser visto na Matriz de Transição na

Figura 4.13.

	0	1	2	3
0	0,82	0,07	0,10	0,01
01	0,54	0,14	0,27	0,05
11	0,68	0,11	0,16	0,05
2	0,60	0,14	0,22	0,04
21	0,79	0,12	0,09	0,00
3	0,38	0,22	0,31	0,09
31	0,80	0,20	0,00	0,00

Figura 4.13: Matriz de transição PML (Santa Maria - RS, 2014-2024).

Porém, há uma probabilidade que chama atenção, a transição de um dia de chuva violenta para outro dia de chuva violenta, $p(3|3)$. Entre os anos de 2004 a 2014, conforme ilustrado na Figura 4.10, essa transição apresentava uma probabilidade de 0,05, ou seja, 5%. Nos últimos dez anos, essa probabilidade aumentou significativamente para 0,09, representando um crescimento considerável.

	0	1	2	3
0	0,82	0,07	0,10	0,01
00001	0,54	0,15	0,27	0,04
10001	0,38	0,50	0,00	0,12
1001	0,40	0,10	0,40	0,10
101	0,50	0,07	0,43	0,00
11	0,68	0,11	0,16	0,05
2	0,60	0,14	0,22	0,04
20001	0,78	0,00	0,00	0,22
2001	0,67	0,08	0,25	0,00
201	0,60	0,07	0,33	0,00
21	0,79	0,12	0,09	0,00
3	0,38	0,22	0,31	0,09
30001	0,00	0,00	0,67	0,33
3001	0,00	1,00	0,00	0,00
301	0,50	0,50	0,00	0,00
31	0,80	0,20	0,00	0,00

Figura 4.14: Matriz de transição AC (Santa Maria - RS, 2014-2024).

A matriz de transição do Algoritmo Contexto, representada na Figura 4.14, for-

nece algumas informações adicionais. As probabilidades de transição para a chuva fraca ficam mais baixas com os novos contextos. Por exemplo, as probabilidades $p(0|30001)$ e $p(0|3001)$ são iguais a zero. Visto isso, ambos algoritmos apontam para a mesma direção, uma possível mudança no regime de chuvas e o aumento da probabilidade de eventos extremos.

Em Caxias do Sul, os resultados foram semelhantes. De 2004 a 2014, a árvore estimada pela Máxima Verossimilhança Penalizada foi a mesma mostrada na Figura 4.9, ou seja, uma Cadeia de Markov. Contudo, o Algoritmo Contexto estimou uma árvore maior, representada na Figura 4.15. Conforme mencionado, esse método é mais sensível e tende, naturalmente, a estimar árvores com mais contextos.

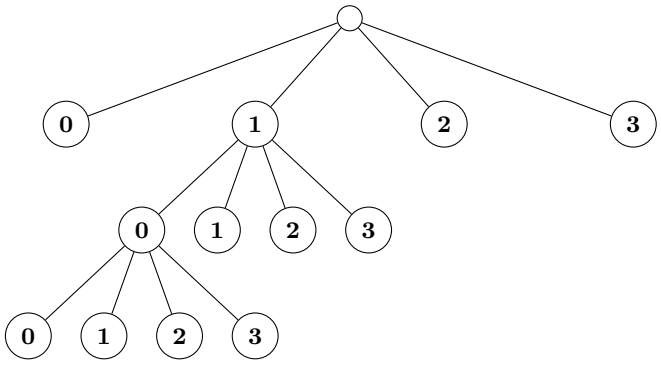


Figura 4.15: Árvore de Contextos AC (Caxias do Sul - RS, 2004-2014).

A matriz de transição referente ao período de 2004 a 2014, representada na Figura 4.16, assemelha-se a matriz de Santa Maria, especialmente na alta probabilidade de transição para chuva fraca. Entretanto, em dias de chuva violenta, o padrão é diferente: a probabilidade de transição para chuva forte é maior que para chuva fraca. Além disso, a transição de chuva forte para moderada ou forte é mais frequente em Caxias do Sul do que em Santa Maria.

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0{,}79 & 0{,}09 & 0{,}11 & 0{,}01 \\ 0{,}60 & 0{,}21 & 0{,}17 & 0{,}02 \\ 0{,}54 & 0{,}21 & 0{,}22 & 0{,}03 \\ 0{,}28 & 0{,}23 & 0{,}43 & 0{,}06 \end{pmatrix} \end{array}$$

Figura 4.16: Matriz de transição PML (Caxias do Sul - RS, 2004-2014).

A matriz de transição estimada com o Algoritmo Contexto, representada na Fi-

gura 4.17, apresenta mais contextos, conforme observado na análise das árvores. A probabilidade de chover moderadamente, dado que nos últimos 3 dias choveu moderadamente, fraco e moderadamente, respectivamente, é de 50%. De maneira semelhante, se o primeiro dia foi de chuva violenta, essa probabilidade seria de 100%, isso quer dizer que nessa amostra, só ocorreu esse tipo de transição.

	0	1	2	3
0	0,79	0,09	0,11	0,01
001	0,63	0,18	0,17	0,02
101	0,34	0,50	0,13	0,03
11	0,62	0,18	0,17	0,03
2	0,54	0,21	0,22	0,03
201	0,44	0,15	0,41	0,00
21	0,65	0,20	0,13	0,02
3	0,28	0,23	0,43	0,06
301	0,00	1,00	0,00	0,00
31	0,82	0,09	0,09	0,00

Figura 4.17: Matriz de transição AC (Caxias do Sul - RS, 2004-2014).

A árvore estimada para o período de 2014 a 2024 foi a mesma em ambos os métodos. Além disso, essa mesma árvore foi estimada pela Máxima Verossimilhança Penalizada em Santa Maria, durante o mesmo período, representada na Figura 4.11. Desse modo, é possível analisar as mudanças nas probabilidades de transição a partir da mesma matriz de transição, na Figura 4.18.

	0	1	2	3
0	0,77	0,10	0,12	0,01
01	0,54	0,22	0,22	0,03
11	0,68	0,13	0,16	0,03
2	0,52	0,21	0,24	0,03
21	0,80	0,07	0,10	0,04
3	0,34	0,16	0,37	0,12
31	0,91	0,09	0,00	0,00

Figura 4.18: Matriz de transição (Caxias do Sul - RS, 2014-2024).

Assim como em Santa Maria, a mudança mais significativa é na probabilidade de transição de um dia de chuva violenta para outro dia de chuva violenta, que aumentou

para 12%, o dobro do valor anterior.

A partir dessa análise, percebe-se que o regime de chuvas pode estar, realmente, mudando. Embora o Algoritmo Contexto tenha estimado uma árvore maior no período de 2004 a 2014 em Caxias do Sul, parece haver uma tendência temporal de aumento do número de contextos. Isso sugere uma maior complexidade no padrão das chuvas, indicando uma possível mudança no regime de chuvas. Logo, a análise da evolução do número de contextos ao longo do tempo pode fornecer mais evidências.

Para analisar a evolução do número de contextos ao longo do tempo e fornecer evidências adicionais, foram estimadas Árvores de Contextos de 2014 a 2024, utilizando uma janela de dados de dez anos. As especificações utilizadas na modelagem foram as mesmas utilizadas anteriormente. O processo de estimação pode ser dividido em duas etapas: primeiro, estimar a Árvore de Contextos com os dados de 22 de maio de 2004 a 22 de maio de 2014 e registrar o número de contextos; segundo, repetir o processo, começando no dia seguinte à data inicial e terminando no dia seguinte à data final, até alcançar a data final dos dados.

Esse método permite analisar se houve de fato um aumento no número de contextos ao longo do tempo e identificar quando esse aumento ocorreu. A partir da análise gráfica das Figuras 4.19 e 4.20.

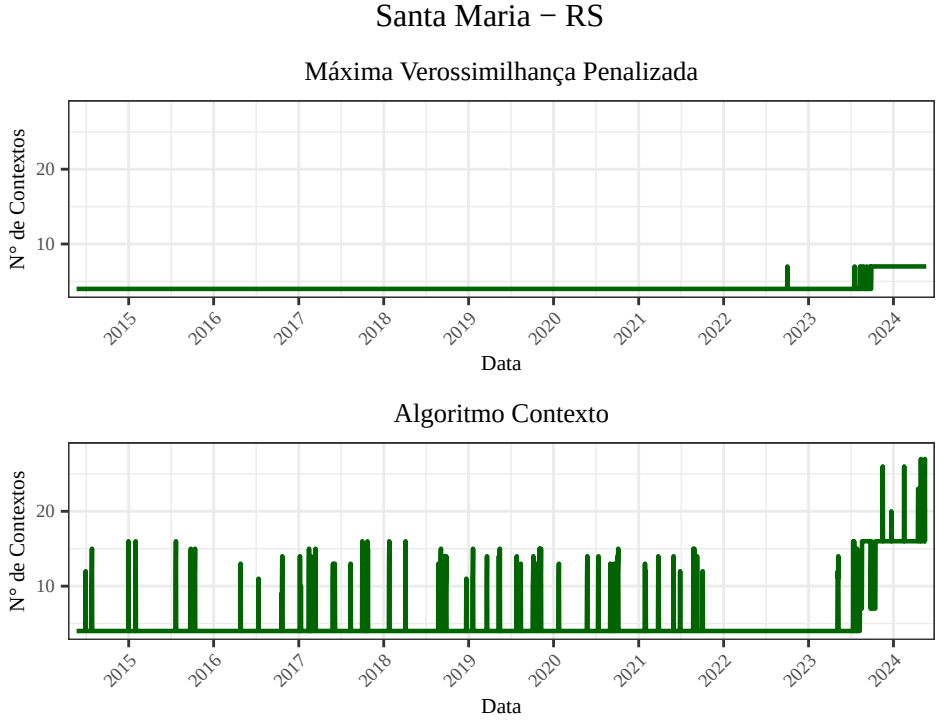


Figura 4.19: Número de Contextos (Santa Maria - RS).

Como evidenciado na Figura 4.19, em Santa Maria, o número de contextos aumentou após 2023 em ambos métodos de estimação. Isso sugere um novo padrão de chuvas a partir do segundo semestre de 2023. O número maior de contextos estimados pelo Algoritmo Contexto e a sua variação, reafirma a sensibilidade. Enquanto para a Máxima Verossimilhança Penalizada, o número de contextos quase constante e deixa clara a mudança após 2023. Entretanto, os dois métodos apontam para o mesmo padrão de mudança.

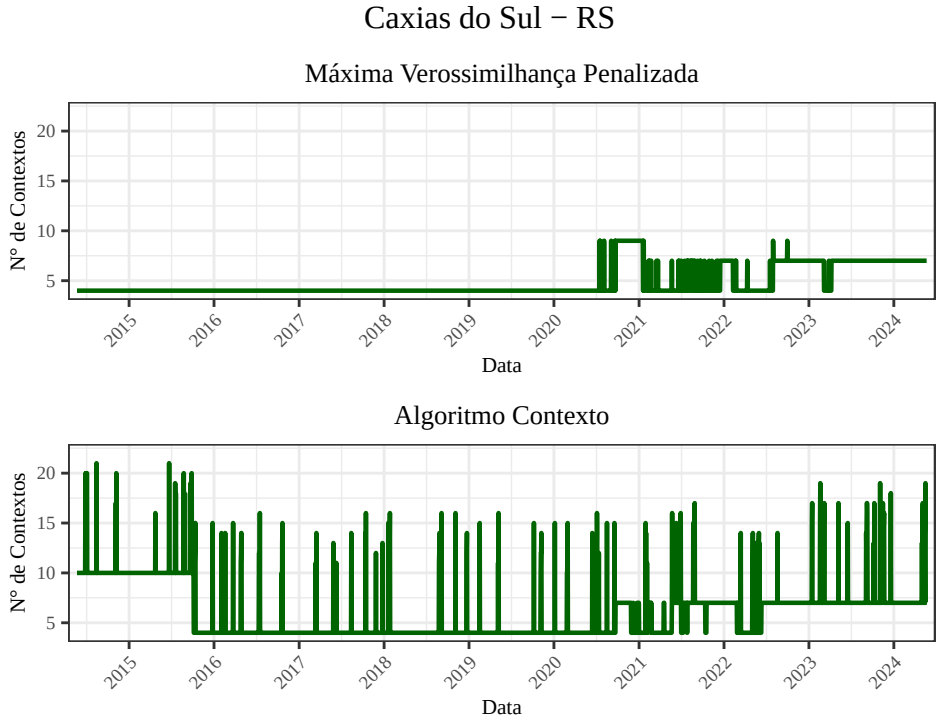


Figura 4.20: Número de Contextos (Caxias do Sul - RS).

Em Caxias do Sul, como visto na Figura 4.20, há uma maior instabilidade na estimação das Árvores de Contextos. Para o Algoritmo Contexto, assim como foi visto na Árvore da Figura 4.15, pouco antes de 2016 o número de contextos era maior, mas logo após esse período, volta a diminuir. Dessa maneira, o comportamento torna-se semelhante ao de Santa Maria, com um aumento no número de contextos após o segundo semestre de 2020. Além de uma clara mudança no nível do número de contextos após o segundo semestre de 2022, para ambos métodos de estimação.

Juntamente ao processo de estimação das Árvores, a partir das janelas de dados, foram feitas previsões para o dia seguinte. Dessa maneira, é possível analisar a acurácia das previsões considerando o "erro" fora da amostra. Os resultados estão na Tabela 4.7.

Tabela 4.7: Acurácia das Previsões de Chuva.

Método / Cidades	PML	AC
Santa Maria	0,63	0,61
Caxias do Sul	0,56	0,55

Considerando que a previsão utiliza apenas o histórico da própria precipitação, sem auxílio de covariáveis, os resultados estão dentro do esperado. Além disso, demonstram a instabilidade no regime de chuvas das cidades, principalmente em Caxias do Sul.

Portanto, nessa seção, foi possível analisar o regime de chuvas em Santa Maria e Caxias do Sul, cidades do Rio Grande do Sul, e evidenciar uma possível mudança no regime de chuvas do estado. Os resultados obtidos destacam uma clara alteração nos níveis de precipitação ao longo dos últimos anos. Diante disso, torna-se claro o atraso na aplicação de políticas públicas voltadas para a prevenção de tragédias climáticas no estado, ressaltando a necessidade urgente de novas estratégias.

4.3.2 Previsão Diária do Preço de Ações

Uma das aplicações comuns nos estudos de Cadeias de Markov é a previsão de séries temporais, como o preço de ações. Consequentemente, é possível estender esse conceito para Cadeias Estocásticas de Ordem Variável. Nesta aplicação, será analisada a capacidade preditiva do modelo utilizando as ações do Banco do Brasil (BBAS3).

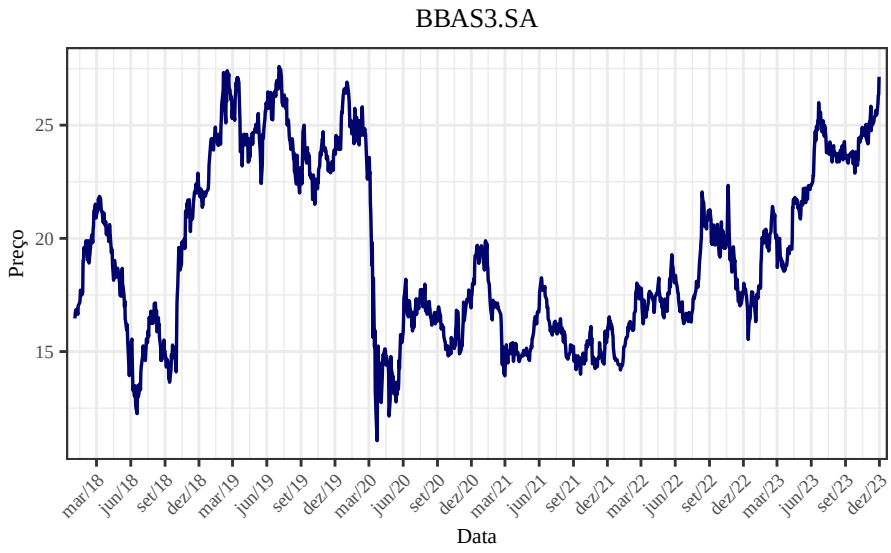


Figura 4.21: Preço de Fechamento da Ação BBAS3, 2018-2023.

Em termos simples, as ações, também conhecidas como papéis, representam uma parcela do capital social de uma empresa. Para este estudo, analisaremos o preço de fechamento das ações BBAS3, no período de janeiro de 2018 a dezembro de 2023, a partir do pacote "quantmod" (RYAN; ULRICH, 2024).

A partir da visualização gráfica da série temporal da BBAS3, na Figura 4.21, é possível notar a grande queda no período de anúncio da pandemia de COVID-19. Além disso, percebe-se um crescimento nos preços a partir de 2022.

Para o estudo de previsão, os dados serão divididos em conjunto de treinamento e teste. Para a validação do modelo, os dados de teste considerarão aproximadamente os últimos 12 meses. Portanto, no conjunto de treinamento será considerado o período de janeiro de 2018 a novembro de 2022, enquanto no conjunto de teste será considerado o período de novembro de 2022 a dezembro de 2023.

Além disso, é preciso discretizar o preço das ações para aplicar o modelo. Desse modo, serão considerados cinco intervalos de tamanho igual entre o valor máximo e o mínimo da série no período de treinamento. Em Delgado, Queiroz e Átila (2023), foram testadas variações dessa ideia para Cadeias de Markov. A Figura 4.22, que representa graficamente essa discretização, pode auxiliar na compreensão do processo.

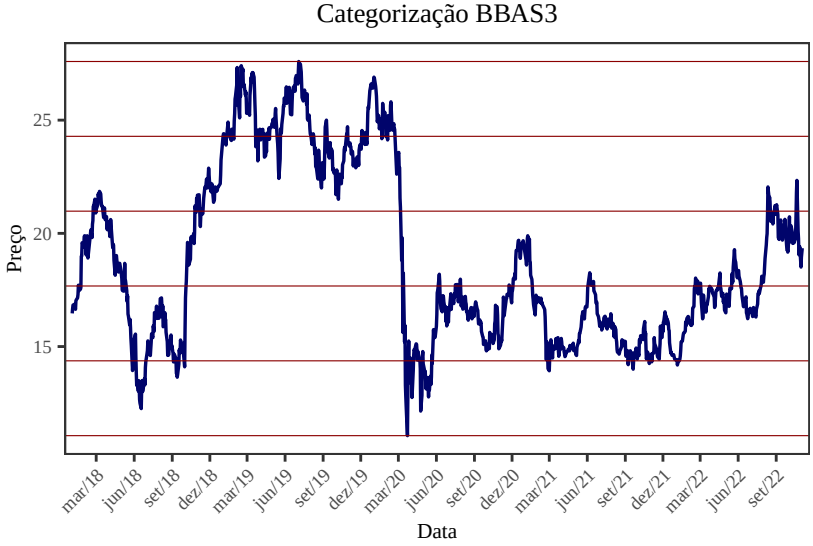


Figura 4.22: Categorização do Preço da Ação

A validação considerará a previsão para o dia seguinte, desse modo, a amostra de treino sempre crescerá um dia a cada previsão. Consequentemente, os intervalos podem mudar com o passar da inclusão de novos dados aos dados de treinamento. Na Figura 4.23, estão as previsões feitas pelo modelo em comparação com a série original.

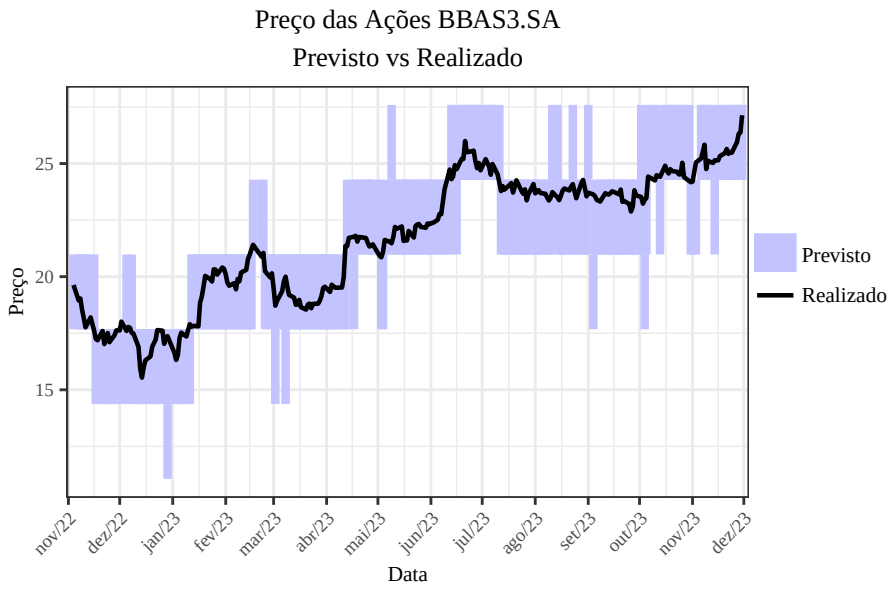


Figura 4.23: Previsões do Preço da Ação BBAS3.

Com base no gráfico, nota-se que as previsões acompanham bem a trajetória da série original, mesmo em variações rápidas de queda ou crescimento. É relevante destacar que, apesar da possibilidade de mudança nos intervalos com o passar do tempo, nesse caso não houve. Visto isso, é possível analisar de maneira mais clara as transições entre os intervalos.

Tabela 4.8: Categorização do Preço BBAS3.

Intervalo	**Alfabeto**
[11,065; 14,370)	A
[14,370; 17,675)	B
[17,675; 20,980)	C
[20,980; 24,285)	D
[24,285; 27,590)	E

Conforme a categorização presente na Tabela 4.8, foi elaborada a matriz de confusão das previsões, apresentada na Figura 4.24. A matriz de confusão é importante na avaliação de modelos de previsão, ao permitir visualizar em quais intervalos o modelo acertou ou errou. Além da Acurácia, também será analisado o Kappa proposto em Cohen (1960). O coeficiente Kappa de Cohen é uma medida de concordância que leva em consideração o acerto ao acaso. Quanto mais próximo de 1, maior é a concordância entre os valores previstos e os realizados, considerando o acerto ao acaso.

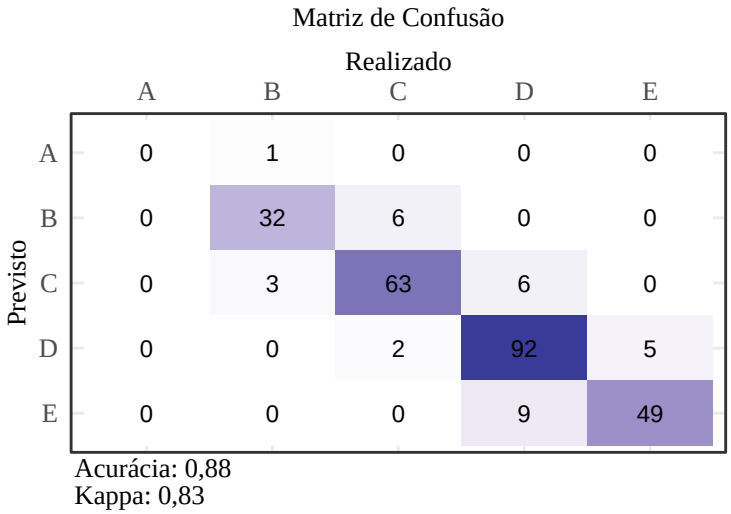


Figura 4.24: Matriz de Confusão das Previsões da BBAS3.

A Matriz de Confusão apresenta o comportamento esperado, evidenciado pela diagonal destacada, em que a maioria das previsões está corretas. Além disso, a Acurácia confirma um bom desempenho, com 88% de acertos. Assim como o coeficiente Kappa de 0,83 também demonstra um bom desempenho do modelo.

Ademais, a análise das probabilidades de transição é essencial para fornecer informações aos gestores financeiros. Por exemplo, a última matriz de probabilidades de transição estimada, apresentada na Figura 4.25, oferece informações essenciais para orientar as decisões futuras.

$$
\begin{array}{r} \\ A \\ B \\ BC \\ CC \\ CD \\ DC \\ DD \\ E \\ ED \end{array}
\begin{array}{c}
\begin{array}{ccccc} A & B & C & D & E \end{array} \\
\left(\begin{array}{ccccc}
0{,}75 & 0{,}25 & 0{,}00 & 0{,}00 & 0{,}00 \\
0{,}03 & 0{,}94 & 0{,}03 & 0{,}00 & 0{,}00 \\
0{,}00 & 0{,}48 & 0{,}52 & 0{,}00 & 0{,}00 \\
0{,}00 & 0{,}04 & 0{,}91 & 0{,}05 & 0{,}00 \\
0{,}00 & 0{,}00 & 0{,}27 & 0{,}73 & 0{,}00 \\
0{,}00 & 0{,}00 & 0{,}71 & 0{,}29 & 0{,}00 \\
0{,}00 & 0{,}00 & 0{,}04 & 0{,}90 & 0{,}06 \\
0{,}00 & 0{,}00 & 0{,}00 & 0{,}11 & 0{,}89 \\
0{,}00 & 0{,}00 & 0{,}00 & 0{,}60 & 0{,}40
\end{array}\right)
\end{array}
$$

Figura 4.25: Matriz de transição Ações BBAS3.

A partir da matriz de transição, observa-se que as probabilidades mais altas correspondem à transição dentro do mesmo intervalo. Por exemplo, a transição do contexto

CD para D possui uma probabilidade de 73%, a mais elevada entre as possíveis transições, mesmo considerando o passado C no contexto. As probabilidades associadas ao contexto BC também são interessantes, com transições para os intervalos B e C bastante próximas entre si. Essas informações são fundamentais para o estudo dos preços e para embasar as previsões.

Visto isso, a utilização de Cadeias Estocásticas de Ordem Variável para previsão do preço de ações pode fornecer informações valiosas na área financeira. O conhecimento das probabilidades de transição pode auxiliar na compreensão dos padrões dos preços das ações e, consequentemente, na tomada de decisões estratégicas.

5 Conclusão

Os resultados apresentados neste estudo têm implicações significativas para a gestão pública e para a tomada de decisão em políticas públicas, climáticas e econômicas. A análise das Cadeias Estocásticas de Ordem Variável revelou possíveis mudanças nos padrões de chuvas em Santa Maria e Caxias do Sul ao longo das últimas décadas. Essas mudanças não só impactam a infraestrutura urbana e rural, mas também têm potencial para influenciar diretamente a agricultura e a gestão de recursos hídricos nessas regiões. A capacidade de identificar e compreender essas alterações é crucial para o planejamento adequado das áreas urbanas e agrícolas, permitindo adaptações necessárias às novas condições climáticas emergentes.

Além disso, as acurácias moderadas das previsões de chuvas destacam a utilidade dos modelos estudados como ferramentas de suporte à decisão. No entanto, também evidenciam áreas onde melhorias podem ser implementadas, como a inclusão de covariáveis adicionais para aprimorar a precisão das previsões. Isso é particularmente relevante em cenários de mudanças climáticas, em que a variabilidade aumentada pode desafiar modelos baseados exclusivamente em dados históricos. A integração dessas análises pode proporcionar uma gestão mais adaptativa e resiliente, alinhada aos desafios contemporâneos de sustentabilidade e desenvolvimento socioeconômico.

A aplicação das Cadeias Estocásticas de Ordem Variável na previsão do preço de fechamento das ações do BBAS3, demonstrou uma metodologia robusta para gestores públicos interessados em compreender e antecipar tendências de mercado. Utilizando dados históricos disponíveis, esse método pode fornecer percepções valiosas para decisões de investimento e políticas econômicas, contribuindo para uma gestão financeira mais estratégica e informada.

Logo, este estudo não apenas identificou mudanças significativas nos regimes de chuvas em Santa Maria e Caxias do Sul, mas também destacou a importância de modelos analíticos avançados na formulação de políticas públicas climáticas e econômicas. A integração dessas análises pode proporcionar uma gestão mais adaptativa e resiliente, capaz de enfrentar os desafios contemporâneos de sustentabilidade e desenvolvimento socioeconômico eficazmente.

Futuras pesquisas podem ser realizadas para ampliar e aprimorar este estudo. Uma delas é a incorporação de mais variáveis climáticas nos modelos, buscando uma visão mais completa e integrada dos padrões climáticos. Isso poderia incluir não apenas variáveis

atmosféricas, mas também dados de oceanos e outros fatores ambientais relevantes. Ademais, expandir o estudo para incluir mais localidades em diferentes regiões climáticas do Brasil poderia oferecer uma compreensão mais abrangente dos padrões climáticos e suas mudanças ao longo do tempo.

No estudo de previsão dos preços de ações, seria interessante analisar outras formas de categorização, incluindo a análise do número de intervalos e suas distâncias. Além disso, considerar a inclusão de variáveis auxiliares poderia ser uma abordagem promissora para aprimorar a acurácia das previsões.

Referências

Barboza, F. L. *Cadeias estocásticas com memória de alcance variável*. Dissertação (Monografia) — Universidade Federal do Rio Grande do Norte, Natal, 2019. Disponível em: ⟨https://repositorio.ufrn.br/bitstream/123456789/34307/2/CadeiasEstocasticasMemoria_Barboza_2019.pdf⟩.

Bomfim, A. B. A. *Estudo de estimadores de árvores de contexto com aplicação em linguística*. Dissertação (Monografia) — Universidade de Brasília, Brasília, 2015. Disponível em: ⟨https://bdm.unb.br/bitstream/10483/11247/1/2015_AlexBarrosAzevedoBomfim.pdf⟩.

Bühlmann, P.; Wyner, A. J. Variable length markov chains. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 27, n. 2, p. 480–513, 1999.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.

Csiszár, I.; Talata, Z. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information theory*, IEEE, v. 52, n. 3, p. 1007–1016, 2006.

DELGADO, M. X. T.; QUEIROZ, J.; ÁTILA, O. Previsão de intervalos de preço no mercado de ações brasileiro usando cadeias de markov de tempo discreto. *Revista Brasileira de Computação Aplicada*, v. 15, n. 1, p. 34–47, 2023.

Galves, A.; LEONARDI, F. Exponential inequalities for empirical unbounded context trees. *In and out of equilibrium 2*, Springer, p. 257–269, 2008.

Galves, A.; Leonardi, F. G.; Ost, G. Statistical model selection for stochastic systems with applications to bioinformatics, linguistics and neurobiology. 2022.

Garivier, A.; Leonardi, F. Context tree selection: A unifying view. *Stochastic Processes and their Applications*, Elsevier, v. 121, n. 11, p. 2488–2506, 2011.

GaúchaZH. *Chuva de maio em Caxias do Sul é cem vezes maior que o consumo anual de água mineral do Brasil*. 2024. Disponível em: ⟨https://gauchazh.clicrbs.com.br/pioneiro/geral/noticia/2024/06/chuva-de-maio-em-caxias-do-sul-e-cem-vezes-maior-que-o-consumo-anual-de-agua-mineral-do-brasil-clwv7upch0041013ugjfbndh4.html⟩.

GaúchaZH. *Santa Maria foi a cidade com maior volume de chuva no mundo nesta quarta-feira*. 2024. Disponível em: ⟨https://gauchazh.clicrbs.com.br/ambiente/noticia/2024/05/santa-maria-foi-a-cidade-com-maior-volume-de-chuva-no-mundo-nesta-quarta-feira-clvob3qnw00io01fm479on5bv.html⟩.

Hoel, P. G.; Port, S. C.; Stone, C. J. *Introduction to stochastic processes*. [S.l.]: Waveland Press, 1986.

INMET. *Dados Históricos - Instituto Nacional de Meteorologia - INMET*. 2024. Disponível em: ⟨https://bdmep.inmet.gov.br/⟩, Acesso em: 23 de maio de 2024.

Matta, D. H. da; Garcia, N. L. *Algoritmos de estimação para Cadeias de Markov de alcance variavel: aplicações a detecção do ritmo em textos escritos*. Tese (Dissertação de Mestrado) — Universidade Estadual de Campinas, Campinas, 2008.

Quintino, F. S. *Aplicações de cadeias de ordem variável estocasticamente perturbadas*. Dissertação (Trabalho de Conclusão de Curso) — Universidade de Brasília, Brasília, 2015. Disponível em: ⟨https://www.bdm.unb.br/bitstream/10483/14390/1/2015_FelipeSousaQuintino.pdf⟩.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. Disponível em: ⟨https://www.R-project.org/⟩.

Ramos, A. d. A. *Modelagem de dados de insolação por meio de cadeias de ordem variável*. Dissertação (Trabalho de Conclusão de Curso) — Universidade de Brasília, Brasília, 2023.

Rissanen, J. A universal data compression system. *IEEE Transactions on Information Theory*, v. 29, n. 5, p. 656–664, 1983.

RYAN, J. A.; ULRICH, J. M. *quantmod: Quantitative Financial Modelling Framework*. [S.l.], 2024. R package version 0.4.26. Disponível em: ⟨https://CRAN.R-project.org/package=quantmod⟩.

Apêndice

A Código da Função do Algoritmo Contexto

Esta função retorna a Árvore de Contextos e respectiva Matriz de Transição.

```
arvore_rissanen <- function(dados, d, limiar){
  texto <- paste0(dados, collapse = "")
  alfabeto <- sort(unique(dados))
  alfabeto <- as.character(alfabeto)

  list2 <- combinacoes2 <- seqs <- NULL
  matrizes <- lapply(0:d, function(d){
    if(d==0){list <- replicate(d, list(alfabeto))
    combinacoes <- expand.grid(list)
    linhas <- do.call(paste0, rev(combinacoes))}
    else{list <- list2
    combinacoes <- combinacoes2
    linhas <- seqs}
    list2 <<- replicate(d+1, list(alfabeto))
    combinacoes2 <<- expand.grid(list2)
    seqs <<- do.call(paste0, rev(combinacoes2))
    matrix(seqs, ncol = length(alfabeto),
           dimnames = list(linhas, alfabeto), byrow = T)
  })

  remove(seqs, combinacoes2, list2)

  N_cont <- function(matriz, texto){
    if(nchar(matriz[[1]])==1){
      seqs <- strsplit(texto, split = "")[[1]]
    }
    else{
      seqs <- sapply(1:(nchar(texto) - nchar(matriz[[1]]) + 1),
                     function(X){
                       substr(texto, start = X,
                              stop = X + nchar(matriz[[1]]) - 1)})
    }
    N <- table(seqs)

    matriz[-which(matriz %in% names(N))] <- NA
```

```
  matriz[which(matriz %in% names(N))] <-
    N[matriz[which(matriz %in% names(N))]]
  class(matriz) <- "numeric"
  matriz[is.na(matriz)] <- 0
  return(matriz)
}

contagem <- lapply(matrizes, N_cont, texto)

# Probabilidades de Transião
# P = N(s,a)/N(s)
# N(s) = \sum_a N(s, a)
probs <- lapply(contagem, function(m) m/rowSums(m))
# do Csizar ele não considera P = 1/|A|
for(i in 1:(d+1)){
  probs[[i]][which(is.na(probs[[i]]))] <- 1/length(alfabeto)
}

SeqToValue <- function(matrizes, seq.){
  if(nchar(seq.)[1]>1){
    s <- substr(seq., 1, nchar(seq.)-1)
    a <- substr(seq., nchar(seq.), nchar(seq.))
    return(matrizes[[nchar(seq.)[[1]]]][s,a])
  }else{return(matrizes[[nchar(seq.)[[1]]]][1,seq.])}
}



delta_n <- function(w){
  if(nchar(w)+2<=d+1){
    p <- probs[[nchar(w)+2]][paste0(alfabeto, w),]

    q <- matrix(rep(if(nchar(w)==0){probs[[1]]
      }else{probs[[nchar(w)+1]][w,]}, length(alfabeto)),
      ncol = length(alfabeto), byrow = T)

    div <- p*log(p/q)
    div[which(p==0)] <- 0
    div[which(p > 0 & q==0)] <- Inf
    divergencia <- rowSums(div)
    rissanen <- if(nchar(w)==0){
      SeqToValue(contagem, paste0(alfabeto, w))*divergencia
```

```
    }else{
        SeqToValue(contagem, paste0(alfabeto, w))[,1]*divergencia
      }
    rissanen[is.nan(rissanen)] <- 0
    return(sum(rissanen))
  }else{NA}
}

V_n <- unlist(matrizes)[which(unlist(contagem)>0)]
V_n <- c("", V_n)
vetor_delta <- sapply(V_n, delta_n)

C_w <- function(sequencia, vetor, delta){
  if(nchar(sequencia)[1]>=d | if(sequencia==""){FALSE
    }else{SeqToValue(contagem, sequencia)==0}){
    return(0)
  }else{return(max(as.numeric(vetor[if(sequencia==""){1
    }else{sequencia}] > delta),
                max(sapply(paste0(alfabeto, sequencia),
                           function(x){
                             C_w(x, vetor, delta)}
                           ))))}
}

vetor_c <- sapply(V_n, function(x){
  C_w(x, vetor_delta, delta = limiar)})

sequencias <- names(vetor_c[vetor_c==TRUE])

d_max <- max(nchar(names(vetor_c[vetor_c==TRUE]))) + 1
# vai ser a profundidade máxima, então não precisamos analisar
# sequências maiores que isso

seqs_0 <- vetor_c[which(
  vetor_c[nchar(names(vetor_c))<=d_max]==FALSE)]
names(vetor_c)[1] <- "vazio"
arvore <- c()
i=1
for(seq in names(seqs_0)){
  if(all(vetor_c[if(nchar(seq)>1){
    (c(sapply(2:nchar(seq),
```

```
              function(s) substr(seq, s, nchar(seq))), "vazio"))
      }else{"vazio"}]==TRUE)){
      arvore[i] <- seq
      i = i + 1
    }
  }
  arvore <- sort(arvore)
  probs_arv <- t(sapply(arvore, function(x){
    probs[[nchar(x)+1]][x,]}))

  return(list("Árvore" = arvore,
              "Probs" = probs_arv)) # A árvore final
}
```

B Código da Função da Máxima Verossimilhança Penalizada

Esta função retorna a Árvore de Contextos e respectiva Matriz de Transição. Apesar do nome conter BIC, ela pode ser estendida para outros critérios através do argumento "pen".

```
arvore_bic <- function(dados, d, pen){
  texto <- paste0(dados, collapse = "")
  alfabeto <- sort(unique(dados))
  alfabeto <- as.character(alfabeto)

  list2 <- combinacoes2 <- seqs <- NULL
  matrizes <- lapply(0:d, function(d){
    if(d==0){list <- replicate(d, list(alfabeto))
    combinacoes <- expand.grid(list)
    linhas <- do.call(paste0, rev(combinacoes))}
    else{list <- list2
    combinacoes <- combinacoes2
    linhas <- seqs}
    list2 <<- replicate(d+1, list(alfabeto))
    combinacoes2 <<- expand.grid(list2)
    seqs <<- do.call(paste0, rev(combinacoes2))
    matrix(seqs, ncol = length(alfabeto),
           dimnames = list(linhas, alfabeto), byrow = T)
  })
```

```
remove(seqs, combinacoes2, list2)

N_cont <- function(matriz, texto){
  if(nrow(matriz)==1){
    seqs <- strsplit(texto, split = "")[[1]]
  }
  else{
    seqs <- sapply(1:(nchar(texto) - nchar(matriz[[1]]) + 1),
                   function(X){
                     substr(texto, start = X,
                            stop = X + nchar(matriz[[1]]) - 1)})
  }
  N <- table(seqs)

  matriz[-which(matriz %in% names(N))] <- NA
  matriz[which(matriz %in% names(N))] <-
    N[matriz[which(matriz %in% names(N))]]
  class(matriz) <- "numeric"
  matriz[is.na(matriz)] <- 0
  return(matriz)
}

contagem <- lapply(matrizes, N_cont, texto)

# Probabilidades de Transião
# P = N(s,a)/N(s)
# N(s) = \sum_a N(s, a)
probs <- lapply(contagem, function(m) m/rowSums(m))
# do Csizar ele não considera P = 1/|A|

SeqToValue <- function(matrizes, seq.){
  if(nchar(seq.)>1){
    s <- substr(seq., 1, nchar(seq.)-1)
    a <- substr(seq., nchar(seq.), nchar(seq.))
    return(matrizes[[nchar(seq.)[[1]]]][s,a])
  }else{return(matrizes[[nchar(seq.)[[1]]]][1,seq.])}
}

# Máxima verossimilhana
ML_w <-mapply(function(x,y) {
  apply(log(x)*y, 1, sum, na.rm=TRUE)}, x=probs, y=contagem)
```

```
V_w <- function(w, pen){
  if(nchar(w)[1] == d| ifelse(nchar(w)[1] == d, TRUE,
                              sum(sapply(paste0(alfabeto, w),
                                         function(x) {
                                           SeqToValue(contagem, x)
                                           })) == 0)){

    v <- -pen+if(nchar(w)>0){ML_w[[nchar(w)[1]+1]][w]
      }else{ML_w[[nchar(w)[1]+1]]}

    names(v) <- NULL
    return(v)

  }else{
    v <- max(-pen+if(nchar(w)>0){ML_w[[nchar(w)[1]+1]][w]
      }else{ML_w[[nchar(w)[1]+1]]},
             sum(sapply(paste0(alfabeto, w)[sapply(paste0(alfabeto, w),
                                                   function(x){
                                                     SeqToValue(contagem, x)
                                                     })>0], V_w, pen)))
    names(v) <- NULL
    return(v)}
}

w_in_vn <- names(unlist(ML_w))[-1][sapply(names(unlist(ML_w))[-1],
                                          function(x){
                                            SeqToValue(contagem, x)}
                                          )>0]
w_in_vn <- c("", w_in_vn)
vetor_v <- sapply(w_in_vn, V_w, pen)

xis <- sapply(names(vetor_v), function(v){
  ml <- -pen+if(nchar(v)>0){ML_w[[nchar(v)+1]][v]}else{ML_w[[nchar(v)+1]]}
  names(ml) <- NULL
  sum(vetor_v[names(which(sapply(paste0(alfabeto, v),
                                 function(x) SeqToValue(contagem, x),
                                 USE.NAMES = ifelse(nchar(v)==0, FALSE,
                                                    TRUE)) > 0))]) > ml}
)
xis[is.na(xis)] <- FALSE
```

```
  sequencias <- names(xis[xis==TRUE])

  d_max <- max(nchar(names(xis[xis==TRUE]))) + 1

  seqs_0 <- xis[which(xis[nchar(names(xis))<=d_max]==FALSE)]
  names(xis)[1] <- "vazio"
  arvore <- c()
  i=1
  for(seq in names(seqs_0)){
    if(all(xis[if(nchar(seq)>1){(c(sapply(2:nchar(seq),
                                          function(s){
                                            substr(seq, s, nchar(seq))
                                            }), "vazio"))
      }else{"vazio"}]==TRUE)){

      arvore[i] <- seq
      i = i + 1
    }
  }
  arvore <- sort(arvore)
  probs_arv <- t(sapply(arvore, function(x){
    probs[[nchar(x)+1]][x,]}))

  return(list("Árvore" = arvore,
              "Probs" = probs_arv)) # A árvore final
}
```