Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

SELEÇÃO ESTATÍSTICA DE ÁRVORES DE CONTEXTO

Isadora Nascimento de Almeida

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

SELEÇÃO ESTATÍSTICA DE ÁRVORES DE CONTEXTO

Isadora Nascimento de Almeida
Orientador: Prof. Dr. Ricardo Felipe Ferreira

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Fevereiro de 2024

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

CONTEXT TREE SELECTION

Isadora Nascimento de Almeida
Advisor: Prof. Dr. Ricardo Felipe Ferreira

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
February 2024

Isadora Nascimento de Almeida

SELEÇÃO ESTATÍSTICA DE ÁRVORES DE CONTEXTO

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Isadora Nascimento de Almeida e aprovado pela banca examinadora.

Aprovado em 22 de Janeiro de 2024

Banca Examinadora:

- Prof. Dr. Ricardo Felipe Ferreira (Orientador)
- Prof. Dr. Luis Aparecido Milan
- Prof. Dr. Marcio Alves Diniz

Aos meus familiares e amigos por todo suporte

Agradecimentos

Agradeço a todos que estiveram ao meu lado durante esta jornada acadêmica e pessoal, contribuindo de diversas formas para a realização deste trabalho.

Primeiramente, agradeço à minha família: à minha mãe, Marinei Souza Nascimento de Almeida, e ao meu pai, José Leoncio Pereira de Almeida, por todo o apoio, amor e incentivo ao longo dos anos. À minha irmã Heloisa Nascimento de Almeida e prima Luana Almeida da Conceição, por todo apoio e compreensão.

Agradeço também a todos os familiares que, de alguma maneira, influenciaram e contribuíram para o meu crescimento e desenvolvimento ao longo da vida.

Ao meu grande amor, Gustavo Dasa, agradeço por sua compreensão, paciência e pelo suporte emocional durante os momentos desafiadores desta jornada.

Aos meus amigos de faculdade, Crystiane Souza, Matthews Martins e Juliano Cesar, compartilho minha gratidão pela parceria, momentos de estudos, aprendizados e descontrações que vivemos juntos.

Aos meus amigos de vida, Jessica Gomes, Tamara Rodrigues e Dennys Malta, agradeço pela amizade sincera, momentos de diversão, palavras de encorajamento e por estarem presentes, me apoiando e compreendendo as minhas ausências durante este período.

Não poderia deixar de expressar minha gratidão ao meu orientador, Ricardo Felipe Ferreira, por sua orientação e apoio que foram fundamentais para o desenvolvimento deste trabalho. Agradeço também à professora Daiane Aparecida Zuanetti e a todos os outros professores que cruzaram o meu caminho, por compartilharem seus conhecimentos, experiências e por serem fontes de inspiração ao longo desta jornada acadêmica.

A todos vocês, minha mais profunda gratidão. Este trabalho não seria possível sem o apoio e o amor de cada um de vocês. Obrigada por fazerem parte da minha história e por tornarem este momento tão especial e significativo.

"Na estatística, não existem fatos, apenas estimativas com diferentes graus de incerteza."

(George E. P. Box)

Resumo

Árvores de contextos são modelos que generalizam de maneira parcimoniosa os modelos Markovianos. Esses modelos foram introduzidos por Jorma Rissanen em 1983, como uma ferramenta eficiente na Teoria da Informação. Desde então, esses modelos têm sido amplamente utilizados em muitos campos da Probabilidade e Estatística tanto do ponto de vista teórico quanto aplicado. Observada uma amostra, um problema central em Estatística é o de estimar um modelo aos dados observados. Neste trabalho, estamos interessados em estudar alguns dos principais métodos discutidos pela literatura para a seleção estatística de árvores de contexto. Para isso, vamos realizar um estudo comparativo entre os métodos frequentistas de seleção de árvores de contexto (algoritmo contexto e suas variações) e o método Bayesiano, utilizando dados sintéticos obtidos via simulações. Além disso, ilustramos a performance dos métodos de seleção de modelos por meio de aplicações em dados reais relacionados à neurociência e genética.

Palavras-chave: *Árvore de contexto; Algoritmo contexto; Neurociência; Cadeia de Markov com memória de alcance variável.*.

Abstract

Context trees are models that parsimoniously generalize Markovian models. These models were introduced by Jorma Rissanen in 1983, as an efficient tool in Information Theory. Since then, these models have been widely used in many fields of Probability and Statistics from both theoretical and applied perspectives. Given a sample, a central problem in Statistics is to estimate a model to the observed data. In this work, we are interested in studying some of the main methods discussed in the literature for the statistical selection of context trees. To do this, we will conduct a comparative study between frequentist methods for context tree selection (context algorithm and its variations) and the Bayesian method, using synthetic data obtained via simulations. Additionally, we illustrate the performance of model selection methods through applications in real data related to neuroscience and genetics.

Keywords: *Context trees; Context algorithm; Neuroscience; Variable length Markov chains.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A análise estatística de séries temporais discretas é uma tarefa científica importante, com uma gama muito ampla de aplicações. A modelagem mais natural para a maioria das séries temporais discretas que apresentam uma estrutura temporal aparente se dá a partir de cadeias de Markov (Markov, 2006). No entanto, a descrição de uma cadeia de Markov de ordem d assumindo valores em um conjunto de tamanho m, requer a especificação de $m^d(m-1)$ parâmetros, o que faz do uso das cadeias de Markov não muito apropriado do ponto de vista da inferência da estatística. Como tem sido frequentemente notado na literatura (Raftery, 1985; Bühlmann e Wyner, 1999; Sarkar e Dunson, 2016), o principal obstáculo enfrentado pelas cadeias de Markov de ordem finita é que elas formam uma classe de modelos que não são estruturalmente ricos, uma vez que qualquer tipo de representação parcimoniosa do espaço de estados não é possível. Além disso, a dimensão do espaço de parâmetros cresce exponencialmente com o tamanho da memória. A falta de flexibilidade dessa classe de modelos dificulta, entre outras coisas, a obtenção de um preditor que apresente um bom equilíbrio entre a sua capacidade de estimar bem os dados e a sua complexidade.

Na teoria da informação, abordagens alternativas foram desenvolvidas para superar as dificuldades mencionadas. Cadeias de Markov com memória de alcance variável fornecem uma classe de modelos muito mais rica e muito mais flexível de cadeias. A principal característica dessa abordagem é que o tamanho da memória que determina a probabilidade de transição da cadeia depende de uma parcela relevante dos mais recentes símbolos que foram observados, fornecendo modelos mais parcimoniosos e de fácil interpretação. Essa classe de cadeias foi introduzida por Rissanen (1983a,b, 1986) e vem sendo amplamente utilizada em diversas áreas de pesquisa, incluindo compressão de dados (Weinber-

ger *et al.*, 1994; Willems *et al.*, 1995), aprendizado de máquina (Gabadinho e Ritschard, 2016), seleção de modelos (Mächler e Bühlmann, 2004; Bejerano e Yona, 2001) e predição (Merhav e Feder, 1998).

Rissanen (1983a) denominou a parte relevante de cada passado que determina a transição da cadeia de contexto. O conjunto de todos os contextos possuem uma propriedade, denominada propriedade do sufixo, que afirma que nenhum contexto é sufixo próprio de outro contexto. Essa propriedade permite representar o conjunto de todos os sufixos como uma árvore rotulada com raiz. Com essa representação, a cadeia estocástica subjacente é descrita pela árvore de todos os contextos e especificada por uma família de probabilidades de transição, que juntas definem o que denominamos árvore de contexto probabilística.

O termo cadeia de Markov de alcance variável foi cunhado na literatura estatística por (Bühlmann e Wyner, 1999) para designar os processos que permitem uma representação de árvore de contexto. Essa classe de modelos tem se mostrado útil na modelagem de muitos problemas reais como, por exemplo, em bioinformática (Bejerano e Yona, 2001; Leonardi, 2006) e linguística (Galves *et al.*, 2012; Abakuks, 2012).

Historicamente, a estimação da árvore de contexto de um processo tem sido considerada a partir de diferentes versões do algoritmo contexto, que foi introduzido por (Rissanen, 1983b). O algoritmo contexto é baseado na ideia de poda da árvore, i.e., a partir de uma amostra produzida pela cadeia com memória de alcance variável, inicializamos o algoritmo com a maior árvore de contextos candidatos para a amostra e, em seguida, iniciamos a poda da árvore a partir de uma regra de decisão pré-estabelecida (baseada em estimativas das probabilidades de transição) até obtermos a menor árvore de contextos que estime bem a amostra. Uma lista incompleta de artigos que abordaram o problema de estimação de árvores de contexto inclui (Ron *et al.*, 1996; Bühlmann e Wyner, 1999; Galves e Leonardi, 2008) e, para uma revisão de literatura, Galves e Löcherbach (2008). Outra abordagem para a seleção de árvores de contexto foi proposta por Csiszár e Talata (2006), que é baseada no critério de informação Bayesiano (BIC, do inglês, *Bayesian Information Criterion*). Nesse caso, para cada possível árvore de contexto, um critério é calculado com o balanço da qualidade do ajuste e a complexidade do modelo. Do ponto de vista da teoria da informação, esse procedimento pode ser interpretado como uma variação do princípio do comprimento mínimo da descrição (MDL, do inglês, *Minimum Description Length principle*). Outras abordagens e resultados podem ser encontrados

em Csiszár e Talata (2006) como, por exemplo, o algoritmo baseado na ponderação de árvores de contexto introduzido por Willems *et al.* (1995).

Em seu artigo seminal, Rissanen (1983a) mostrou a consistência fraca do estimador do algoritmo contexto no caso em que os contextos possuem tamanho limitado, isto é, quando as árvores de contexto são finitas. Em seguida, Bühlmann e Wyner (1999) provaram a consistência fraca do estimador do algoritmo contexto no caso em que o tamanho máximo dos contextos não é limitado por um valor fixo d, mas sim por um valor $d(n)$ que é função do tamanho n da amostra. Uma maneira diferente de provar a consistência para o caso de árvores de contextos finitas foi introduzido em Galves e Löcherbach (2008), usando limites da concentração das probabilidades de transição empíricas em torno do seu valor limitante. Esse resultado foi generalizado por Galves e Leonardi (2008) para o caso de árvores de contextos ilimitadas. O caso ilimitado também foi considerado por Csiszár e Talata (2006) e Leonardi (2007), que mostraram, respectivamente, a consistência forte dos estimadores das árvores de contexto estimadas a partir do critério de informação Bayesiano e a consistência fraca do estimador do algoritmo que penaliza a função de verossimilhança. Outros resultados que mencionan o BIC e do método de ponderação de árvores de contexto podem ser encontrados em Garivier e Leonardi (2011).

Recentemente, Papageorgiou *et al.* (2021) e Kontoyiannis *et al.* (2022) revisitaram os modelos de árvores de contextos e o método de ponderação de árvores de contexto a partir de um ponto de vista da inferência Bayesiana. Esses modelos foram denominados árvores de contextos Bayesianas (BCT, do inglês, *Bayesian Context Tree*) e desenvolvidos para séries temporais discretas. As árvores de contextos Bayesianas tem se mostrado eficazes em diversas tarefas estatísticas, incluindo seleção de modelos, estimação, predição e detecção de pontos de mudança (Papageorgiou *et al.*, 2021; Lungu *et al.*, 2022a,b). Além disso, um pacote implementado no software R está disponível (Kontoyiannis *et al.*, 2022). Kontoyiannis (2022) apresenta uma série de resultados teóricos que oferecem uma visão adicional sobre o desempenho de ferramentas estatísticas e metodológicas associadas com as árvores de contextos Bayesianas e que também fornecem uma justificativa teórica para sua aplicação prática.

Neste trabalho, pretendemos realizar um estudo comparativo entre os métodos frequentistas de seleção de árvores de contexto (algoritmo contexto e suas variações) e o método Bayesiano proposto por Kontoyiannis *et al.* (2022). Para isso, vamos utilizar dados sintéticos obtidos a partir de simulações. Além disso, aplicamos essas metodologias

em dados obtidos de problemas práticos relacionados à neurociência e genética. Portanto, esse estudo complementa os estudos sobre inferência estatística em modelos que podem ser descritos a partir de árvores de contexto.

Este trabalho está organizado da seguinte maneira. No próximo capítulo, são apresentadas as notações, definições e conceitos iniciais relacionados às árvores de contexto, bem como os modelos de árvore de contexto. No Capítulo 3, exploramos a inferência frequentista em árvores de contexto, incluindo a seleção de ordem de cadeia de Markov e alguns métodos de estimação. No Capítulo 4, introduzimos a abordagem Bayesiana para inferência em árvores de contexto, incluindo a ponderação dessas árvores e o algoritmo correspondente. No Capítulo 5, exploramos o estudo de simulação, desempenhos dos métodos em diferentes cenários e aplicações em dados eletrofisiológicos e genéticos. O Capítulo 6 encerra esta monografia com algumas considerações.

Capítulo 2

Árvores de contexto

Árvores de contexto são modelos que generalizam, de maneira parcimoniosa, os modelos Markovianos. Dessa forma, são utilizadas como uma ferramenta eficiente para análise e modelagem de processos estocásticos, visto que fornecem uma visão abrangente e detalhada do processo, facilitando a previsão, modelagem e simulação de seu comportamento. Neste capítulo, estamos interessados em definir algumas das notações, definições e conceitos que são utilizados nesta monografia.

2.1 Notações, definições e conceitos preliminares

Nesta seção, apresentamos as notações, definições e conceitos necessários para compreendermos o conceito de árvores de contexto. Começamos com a definição de processos estocásticos, seguida por algumas de suas propriedades. Além disso, também apresentamos algumas das notações que serão utilizadas ao longo do texto.

Informalmente, um processo estocástico é uma coleção de variáveis aleatórias que representam a evolução de um sistema ao longo do tempo. Diferente de um processo determinístico, um processo estocástico não possui apenas uma alternativa de evolução, isto é, à medida que o tempo avança, o processo estocástico gera uma sequência aleatória de observações. E para definir um processo estocástico, é preciso, antes definir alguns conceitos:

Definição 2.1 (Espaço de probabilidade) *Um espaço de probabilidade é uma tripla* $(\Omega, \mathcal{F}, P)$ *que permite modelar e quantificar a incerteza associada a um experimento aleatório, em que* Ω *é o conjunto de todos os possíves resultados do experimento aleatório*

e é denominado espaço amostral, consequentemente, cada $\omega \in \Omega$ é chamado de ponto amostral; $\mathcal{F}$ é denominado sigma-álgebra de eventos e trata-se de uma coleção de subconjuntos de Ω que satisfazem certas condições; e P é uma função que associa a cada evento de $\mathcal{F}$ um número real pertencente ao intervalo $[0, 1]$, essa função é denominada medida de probabilidade ou, simplesmente, probabilidade.

Definição 2.2 (Processo Estocástico) *Um processo estocástico $\{X_t : t \in \mathbb{T}\}$ é uma coleção de variáveis aleatórias definidas em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$ e são indexadas pelo conjunto $\mathbb{T}$.*

Um processo estocástico pode ser definido com um conjunto de indexação discreto ou contínuo, dependendo da natureza das variáveis aleatórias. Nesta monografia, assumimos que o processo estocástico associado ao problema em estudo é um processo com tempo discreto, ou seja, $\mathbb{T}$ é um conjunto enumerável. Nesse caso, $\mathbb{T} = \mathbb{Z}$, em que $\mathbb{Z}$ denota o conjunto dos números inteiros. Desse modo, para cada tempo $t \in \mathbb{Z}$, X_t é uma variável aleatória que representa o estado do sistema no instante de tempo t e assume valores do alfabeto (ou espaço de estados) discreto.

Dado o objetivo de modelar e prever o comportamento de uma sequência aleatória por meio de um processo estocástico, é importante entender o seu padrão de evolução e características estatísticas à medida que o tempo avança indefinidamente. Para tanto, apresentamos os conceitos de homogeneidade e estacionaridade de um processo estocástico.

Definição 2.3 (Homogeneidade) *Um processo estocástico $\{X_t : t \in \mathbb{Z}\}$ definido em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$ com valores em um alfabeto $\mathcal{A}$ finito é dito homogêneo quando as probabilidades de transição são invariantes ao longo do tempo, ou seja, os conjuntos de probabilidades condicionais que regem a dinâmica do sistema são independentes do valor de $t \in \mathbb{Z}$.*

Dessa forma, a propriedade de homogeneidade permite que as conclusões obtidas a partir de um determinado período de tempo sejam extrapoladas para outros períodos, pois não há mudança nas características estatísticas do processo.

Definição 2.4 (Estacionaridade) *Um processo estocástico $\{X_t : t \in \mathbb{Z}\}$ definido em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$ com valores em um alfabeto $\mathcal{A}$ finito é dito estacionário quando a distribuição conjunta das variáveis aleatórias é a mesma independentemente do tempo da escolha inicial.*

Portanto, um processo estocástico é estacionário quando suas propriedades estatísticas permanecem inalteradas ao longo do tempo. Dessa forma, observa-se que a estacionariedade é crucial para a estimação da dinâmica da cadeia e previsão de seu comportamento futuro, uma vez que podemos extrapolar informações passadas para o futuro.

Tendo em vista os objetivos da monografia, consideramos processos estocásticos homogêneos e estacionários a tempo discreto com alfabeto finito. Porém, antes de formalizar as ideias que serão desenvolvidas, é importante fixar algumas notações que serão utilizadas ao longo do texto.

Seja $\mathcal{A}$ um conjunto finito denominado alfabeto. Denotamos por $|\mathcal{A}|$ a cardinalidade do conjunto $\mathcal{A}$. Considerando m e n dois números inteiros tais que $m \leq n$, denotamos por w_m^n a sequência finita $(w_m, ..., w_n)$, por $w_{-\infty}^n$ a sequência infinita à esquerda $(\ldots, w_{n-1}, w_n)$ e por $w_m^{+\infty}$ a sequência infinita à direita $(w_m, w_{m+1}, \ldots)$, de forma que $w_i \in \mathcal{A}$ para todo $i \in \mathbb{Z}$.

Para qualquer $m \leq n$, a quantidade de símbolos pertencentes à sequência w_m^n é dada por $|w_m^n| = n - m + 1$. Além disso, para todo $n \in \mathbb{Z}$, vamos considerar que $w_{n+1}^n = \varnothing$, visto que $\left|w_{n+1}^n\right| = n - (n+1) + 1 = 0$. Considerando duas sequências w e $w^{'}$, denotamos por $ww^{'}$ a sequência de tamanho $|w| + |w^{'}|$ obtida a partir da concatenação de duas sequências. A junção de duas sequências é também estendida ao caso em que w representa uma sequência infinita à esquerda, isto é, $w = w_{-\infty}^{-1}$. Se n é um inteiro positivo e w uma sequência finita de elementos de em $\mathcal{A}$, denotamos por $w^n = ww \ldots w$ a junção de n vezes a sequência w.

Nesse sentido, denotamos o conjunto de todas as sequências finitas e o conjunto de todas as sequências infinitas à esquerda, respectivamente, por

$$\mathcal{A}^* = \bigcup_{j=0}^{+\infty} \mathcal{A}^j \quad \text{e} \quad \mathcal{A}^{-\mathbb{N}},$$

em que $\mathcal{A}^j$ é o conjunto de todas sequências de tamanho j com símbolos em $\mathcal{A}$. No caso em que $j = 0$, temos a sequência vazia denotada por $\varnothing$.

2.2 Cadeias de Markov

As cadeias de Markov são um caso particular de processos estocásticos, cuja transição entre estados depende de um passado de tamanho fixo e finito. Esses modelos foram

introduzidos pelo matemático russo Andrei Andreyevich Markov (Markov, 1906) e vem sendo amplamente utilizados em diversas áreas de pesquisa que apresentam uma dinâmica temporal, como finanças, marketing, compreensão de textos e entre outros.

Neste sentido, introduzimos os modelos Markovianos, começando pela definição de uma cadeia de Markov de ordem 1 e suas principais propriedades, seguida pela generalização de uma cadeia de Markov de ordem k.

Propriedade 2.5 (Propriedade de Markov) *Um processo estocástico* $\{X_t : t \in \mathbb{Z}\}$ *definido em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores em um alfabeto* $\mathcal{A}$ *finito é dito possuir a propriedade de Markov quando*

$$\mathbb{P}(X_t = x_t \mid X_0^{t-1} = x_0^{t-1}) = \mathbb{P}(X_t = x_t \mid X_{t-1} = x_{t-1}),$$

para quaisquer $t \in \mathbb{Z}$ *e* $x_t \in \mathcal{A}$.

A propriedade de Markov refere-se ao tamanho da memória de um processo estocástico. Assim, um processo estocástico que possui a propriedade de Markov é tal que a distribuição de probabilidade dos estados futuros do processo depende apenas do estado atual, ou uma quantidade finita de estados imediatamente anteriores.

Definição 2.6 (Cadeia de Markov de ordem 1) *Um processo estocástico* $\{X_t : t \in \mathbb{Z}\}$ *definido em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores em um alfabeto* $\mathcal{A}$ *finito é denominado uma cadeia de Markov quando possui a propriedade de Markov.*

Dessa forma, nas cadeias de Markov de ordem 1, a probabilidade da cadeia assumir um certo estado futuro do sistema, dado todos os estados passados assumidos por ela e o seu estado presente, depende apenas do seu estado presente. Consequentemente, assumindo uma cadeia homogênea com alfabeto finito, é possível descrever a dinâmica do sistema a partir de uma matriz quadrada de ordem $|\mathcal{A}| \times |\mathcal{A}|$, que denominaos de matriz de transição.

Definição 2.7 (Matriz de transição) *Uma Cadeia de Markov homogênea* $\{X_t : t \in \mathbb{Z}\}$ *de ordem* 1 *definida em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores em*

um alfabeto $\mathcal{A}$ finito é dita possuir uma matriz de transição Q quando

$$Q = \begin{array}{c} \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & \cdots & a_p \end{array} \\ \begin{bmatrix} Q(a_1 \mid a_1) & Q(a_2 \mid a_1) & \cdots & Q(a_p \mid a_1) \\ Q(a_1 \mid a_2) & Q(a_2 \mid a_2) & \cdots & Q(a_p \mid a_2) \\ \vdots & \vdots & \ddots & \vdots \\ Q(a_1 \mid a_p) & Q(a_2 \mid a_p) & \cdots & Q(a_p \mid a_p) \end{bmatrix} \end{array},$$

em que $Q(a^* \mid a) := \mathbb{P}(X_t = a^* \mid X_{t-1} = a)$, para quaisquer $t \in \mathbb{Z}$ e $a, a^* \in \mathcal{A}$.

A matriz de transição Q, conforme apresentada na Definição 2.7, descreve a transição de um sistema do estado presente (representado pelas linhas) para um estado futuro (representado pelas colunas). Cada elemento da matriz representa uma probabilidade de transição, garantindo que $0 \leq Q(. \mid a) \leq 1$, para todo a pertencente ao alfabeto. É importante observar que cada linha da matriz indica a probabilidade de transição de um estado $a \in \mathcal{A}$ para todos os estados do alfabeto. Portanto, as probabilidades de todas as transições em uma linha devem somar um.

Do ponto de vista estatístico, é importante reconhecer que alguns processos dependem não apenas do estado imediatamente anterior, mas sim de $k > 1$ passos no passado. Portanto, para lidar com esses casos, é necessário introduzir as cadeias de Markov de ordem k. Essas cadeias permitem capturar a dependência de longo prazo entre os estados do sistema, levando em consideração as informações dos k passos anteriores. Essa generalização é crucial para modelar com precisão uma variedade de fenômenos e sistemas complexos.

Definição 2.8 (Cadeia de Markov ordem k) *Um processo estocástico $\{X_t : t \in \mathbb{Z}\}$ definido em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$ com valores em um alfabeto $\mathcal{A}$ finito é caracterizado como uma cadeia de Markov de ordem k quando possui a propriedade de Markov com memória de alcance $k \in \mathbb{N}$, ou seja,*

$$\mathbb{P}(X_t = x_t \mid X_0^{t-1} = x_0^{t-1}) = \mathbb{P}(X_t = x_t \mid X_{t-k}^{t-1} = x_{t-k}^{t-1}),$$

para todo $t \in \mathbb{Z}$ e $x_t \in \mathcal{A}$.

Dessa forma, as cadeias de Markov de ordem k é um tipo especial de cadeia de Markov em que a probabilidade de transição entre estados depende dos k estados anteriores, mas

dispensa qualquer conhecimento anterior as estes. Semelhante às cadeias apresentadas anteriormente, as cadeias de Markov de ordem k também possuem as propriedades de homogeneidade e estacionaridade. Além disso, qualquer cadeia de Markov de ordem $k > 1$ pode ser reescrita como uma cadeia de Markov de ordem 1 com alfabeto apropriado.

Preposição 2.9 *Toda cadeia de Markov de ordem $k > 1$ definida em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$ com valores em um alfabeto $\mathcal{A}$ finito pode ser reescrita como uma cadeia de Markov de ordem 1 em $\mathcal{A}^k$, em que $\mathcal{A}^k$ é o conjunto de todas as sequências de símbolos de $\mathcal{A}$ com tamanho k.*

Exemplo 2.10 *Para ilustrar o conceito, consideremos um processo estocástico $\{X_t : t \in \mathbb{Z}\}$ definido em um espaço de probabilidade adequado $(\Omega, \mathcal{F}, P)$ com valores em um alfabeto $\mathcal{A} = \{0, 1\}$, que possui a propriedade de Markov com memória de alcance $k = 2$. Nesse caso,*

$$\mathbb{P}(X_t = x_t \mid X_0^{t-1} = x_0^{t-1}) = \mathbb{P}(X_t = x_t \mid X_{t-2} = x_{t-2}, X_{t-1} = x_{t-1}),$$

para todo $t \in \mathbb{Z}$ e $x_t \in \mathcal{A}$.

Observamos que a probabilidade do estado futuro depende apenas dos dois instantes imediatamente anteriores, refletindo uma dependência de curto prazo. Para descrever a dinâmica desse sistema de ordem $k = 2$, podemos definir um novo alfabeto $\mathcal{A}^2$, que consiste em todas as sequências de dois símbolos em $\mathcal{A}$, ou seja, $\mathcal{A}^2 = \{00, 01, 10, 11\}$. A matriz de probabilidade de transição Q é definida, neste caso, como

$$Q = \begin{array}{c} \\ 00 \\ 01 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} Q(0 \mid 00) & Q(1 \mid 00) \\ Q(0 \mid 01) & Q(1 \mid 01) \\ Q(0 \mid 10) & Q(1 \mid 10) \\ Q(0 \mid 11) & Q(1 \mid 11) \end{bmatrix} \end{array}.$$

Essa matriz representa a evolução do sistema a partir dos dois estados imediatamente anteriores (linhas) em direção ao estado futuro (colunas).

Portanto, a representação das cadeias de Markov de ordem k em termos de uma cadeia de Markov de ordem 1 com um alfabeto expandido $\mathcal{A}^k$ oferece uma abordagem conveniente para descrever a evolução de sistemas com dependências de curto prazo. Essa representação permite a aplicação de métodos e técnicas tradicionais das cadeias de

Markov de ordem 1 em problemas que originalmente possuíam uma memória mais longa. Além disso, a notação matricial fornece uma forma compacta e eficiente de representar as probabilidades de transição entre os estados.

2.3 Modelos de árvore de contexto

Uma árvore de contexto é um conjunto de sequências finitas ou infinitas à esquerda, que são sufixos de sequências passadas infinitas. Os sufixos são, portanto, subsequências do passado infinito que contêm informações suficientes para determinar a dinâmica da cadeia estocástica subjacente. Nesse sentido, compreender os conceitos de "sufixo" e "árvore" é fundamental para a apresentação de uma definição formal de árvore de contexto.

Definição 2.11 (Sufixo) *Dados m e n números inteiros positivos, uma sequência de estados s_{-m}^{-1} é **sufixo** de uma outra sequência w_{-n}^{-1} quando $m < n$ e $w_{-m}^{-1} = s_{-m}^{-1}$. Denotamos essa relação por $s_{-m}^{-1} \prec w_{-n}^{-1}$.*

Exemplo 2.12 *Para ilustrar o conceito abordado, consideramos a sequência infinita à esquerda de símbolos binários $w_{-\infty}^{-1} = \{...00101001010\}$. Nesse caso, cada sufixo é uma sequência finita formada a partir de um determinado ponto à esquerda até o instante de tempo $t = -1$. Alguns exemplos de sufixos para sequência $w_{-\infty}^{-1}$ são: "0", "10", "010", "001010".*

Definição 2.13 (Propriedade do Sufixo) *Um subconjunto τ de $\mathcal{A}^* \cup \mathcal{A}^{-\mathbb{N}}$ é uma **árvore** quando nenhuma sequência de estados $s \in \tau$ é um sufixo de outra sequência $w \in \tau$. Essa propriedade é denominada **propriedade do sufixo**.*

Exemplo 2.14 *Considere $\mathcal{A} = \{0, 1\}$ e dois subconjuntos τ_1 e τ_2 de $\mathcal{A}^* \cup \mathcal{A}^{-\mathbb{N}}$. Por exemplo, $\tau_1 = \{1, 110, 010, 00\}$ e $\tau_2 = \{1, 110, 01\}$. Observa-se que nenhuma sequência $s \in \tau_1$ é sufixo de outra sequência também pertencente a τ_1. Portanto, τ_1 possui a propriedade de sufixo apresentada na Definição 2.13, o que implica que τ_1 é uma árvore. Por outro lado, nota-se que a sequência $1 \prec 01$ e ambas pertencentes a τ_2. Assim, conclui-se que τ_2 não possui a propriedade de sufixo e, portanto, não pode ser considerada uma árvore.*

Na estrutura de uma árvore, as informações são organizadas de forma hierárquica e são compostas por dois elementos fundamentais: nós e arestas. Os nós são elementos individuais que compõem a árvore enquanto as arestas representam as conexões entre

esses nós. Cada nó em uma árvore pode ou não ter nós filhos. Um nó que não possui filhos é chamado de folha, enquanto os nós com filhos são chamados de nós internos. Além disso, o nó no topo da árvore é chamado de raiz, e a partir dela, é possível alcançar todos os outros nós da árvore. Na Figura 2.1, podemos ver a representação gráfica da estrutura de uma árvore.

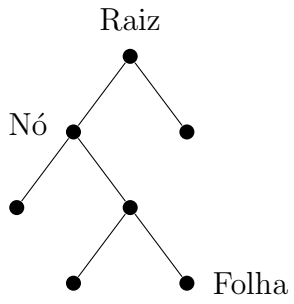


Figura 2.1: Representação gráfica da estrutura de uma árvore. Na imagem, temos a indicação da raiz, de um nó e uma folha.

Definição 2.15 (Contexto) *Seja* $\{X_t : t \in \mathbb{Z}\}$ *um processo estocástico definido em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores em um alfabeto* $\mathcal{A}$ *finito e* m *um número inteiro positivo. Uma sequência de estados* $s_{-m}^{-1} \in \mathcal{A}^*$ *é* ***contexto*** *de um processo* $\{X_t : t \in \mathbb{Z}\}$ *quando*

1. $P(s_{-m}^{-1}) > 0$;

2. *Para todo* $a \in \mathcal{A}$ *e todo* $w \in \mathcal{A}^{-\mathbb{N}}$ *tal que* $s \prec w$ *temos*

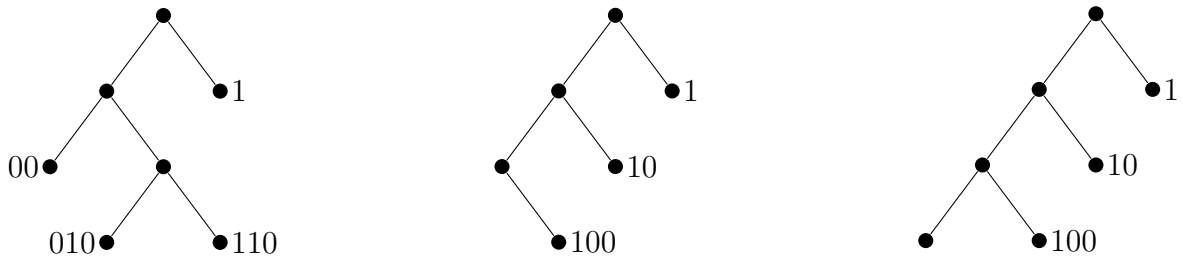
$$P(X_t = a \mid X_{-\infty}^{-1} = w) = P(X_t = a \mid X_{-m}^{-1} = s);$$

3. *Nenhum sufixo de* s *satisfaz 2.*

Definição 2.16 (Árvore completa e árvore de contexto) *Uma árvore* τ *é dita* ***completa*** *quando qualquer sequência infinita à esquerda* $w \in \mathcal{A}^{-\mathbb{N}}$ *tem um sufixo pertencente a* τ*. Pela propriedade do sufixo, temos que esse sufixo é único. Chamamos esse sufixo de contexto da sequência* w *e é denotado por* $c_\tau(w)$*. Uma árvore completa é denominada também uma* ***árvore de contexto****.*

Podemos observar que cada sequência finita $s \in \tau$ pode ser vista como um ramo que liga a raiz da árvore até uma folha. Na abordagem em questão, uma folha da árvore τ de tamanho $j \in \mathbb{N}$ é uma sequência de elementos w_{-j}^{-1} que pode ser representada de

forma gráfica por j arestas começando no topo da árvore e rotuladas, de cima para baixo, por $w_{-1}, w_{-2}, \ldots, w_{-j}$. Neste sentido, os nós mais perto da raiz representam os eventos mais recentes. No caso de sequências infinitas à esquerda, temos folhas infinitas. As sequências finitas ou infinitas w que são sufixos de s para qualquer $s \in \tau$, com $|w| < |s|$, são denominadas de nós, e quando eles são finitos, são denominados de nós internos. Os filhos do nó w são sequências aw, $a \in \mathcal{A}$. Além disso, a raiz é identificada como sequência vazia. Na Figura 2.2, podemos ver a representação de algumas árvores de maneira gráfica.



(a) τ_1 (b) τ_2 (c) τ_3

Figura 2.2: Representação gráfica de três árvores: $\tau_1 = \{1, 110, 010, 00\}$, $\tau_2 = \{100, 10, 1\}$ e $\tau_3 = \{1, 10, 100, \ldots\} \cup 0^\infty$. Todas satisfazem a propriedade do sufixo. As árvores τ_1 e τ_2 são finitas e a árvore τ_3 é infinita, por isso uma versão truncada de τ_3 é representada. A árvore τ_2 não é completa, pois a sequência infinita à esquerda $0^\infty := ...000$ não possui sufixo na árvore.

Definição 2.17 (Altura da árvore) *A **altura** de uma árvore τ é o tamanho da maior sequência $s \in \tau$, isto é,*

$$h(\tau) := \sup \{|s| : s \in \tau\}.$$

Em outras palavras, a altura da árvore τ é a maior distância, em número de nós, entre a raiz da árvore e uma de suas folhas. Caso $h(\tau) < +\infty$, dizemos que τ é limitada, isto é, todas as sequências da árvore τ são finitas e com tamanho menor ou igual a $h(\tau)$. Por outro lado, se $h(\tau) = +\infty$, dizemos que τ é ilimitada, ou seja, existe alguma sequência infinita em τ. A partir da Figura 2.2, vimos que as árvores τ_1 e τ_2 são limitadas, enquanto τ_3 é ilimitada. Além disso, tem-se $h(\tau_1) = h(\tau_2) = 3$.

Definição 2.18 (Árvore truncada) *Dados $\ell \in \mathbb{N}$ e uma árvore τ. Denotamos por $\tau|_\ell$ a árvore τ truncada no nível ℓ, ou seja*

$$\tau|_\ell = \{w \in \tau : |w| \leq \ell\} \cup \left\{w \in \mathcal{A}^\ell : w \prec u \text{ para algum } u \in \tau\right\}.$$

Na Figura 2.2 temos a árvore τ_3 infinita e uma versão truncada é representada. No

caso em questão, temos $\ell = 3$.

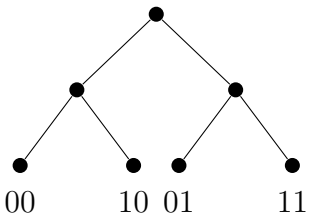
Podemos considerar sistemas nos quais a evolução ocorre a partir de um contexto específico de uma árvore τ em direção a algum símbolo pertencente ao alfabeto $\mathcal{A}$, com base em probabilidades de transição. Essa abordagem nos leva à definição de árvores de contexto probabilísticas.

Definição 2.19 (Árvore de contexto probabilística) *Uma* ***árvore de contexto probabilística*** *com símbolos em $\mathcal{A}$ é um par ordenado (τ, Q) tal que*

1. *τ é uma árvore de contexto;*
2. *$Q := \{Q(a \mid s) : a \in \mathcal{A} \text{ e } s \in \tau\}$ é uma família de probabilidades de transição sobre $\mathcal{A}$.*

A árvore de contexto probabilística é vista como uma representação de uma família de probabilidades de transição tais que, para cada passado, precisamos considerar apenas um contexto desse passado para definir a transição do sistema para o próximo estado. Podemos considerar processos estacionários e ergódicos que evoluem de acordo com árvores de contexto probabilísticas para modelar o sistema em questão.

Exemplo 2.20 *Suponha que um processo estocástico $\{X_t : t \in \mathbb{Z}\}$ definido em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$, possa ser descrito como uma cadeia de Markov de ordem $k = 2$, com valores no alfabeto $\mathcal{A} = \{0, 1\}$. Dessa forma, a árvore τ associada é uma árvore Markoviana completa, e a família de probabilidade de transição pode ser representada por uma matriz de ordem 2. A Figura 3.2 ilustra o par ordenado (τ, Q) que representa essa árvore de contexto probabilística.*



$$Q = \begin{array}{c} \\ 00 \\ 01 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} Q(0 \mid 00) & Q(1 \mid 00) \\ Q(0 \mid 01) & Q(1 \mid 01) \\ Q(0 \mid 10) & Q(1 \mid 10) \\ Q(0 \mid 11) & Q(1 \mid 11) \end{bmatrix} \end{array},$$

Figura 2.3: Representação gráfica da árvore de contexto $\tau = \{00, 01, 10, 11\}$. Note que τ é uma árvore completa, já que toda sequência infinita à esquerda possui sufixo em τ. Além disso, representamos de forma matricial a família de probabilidades de transição Q sobre $\mathcal{A}$.

No entanto, do ponto de vista estatístico, muitos processos requerem um alcance de memória maior do que $k = 2$. Para descrever uma cadeia de Markov de ordem k, que assume valores em um conjunto de tamanho m, é necessário especificar $m^k(k-1)$ parâmetros, o que torna o uso de cadeias de Markov inadequado do ponto de vista da inferência estatística. Dessa forma, abordagens alternativas foram desenvolvidas para superar os obstáculos enfrentados pelas cadeias de Markov de ordem superior. A principal característica das novas abordagens é que o tamanho da memória que determina a probabilidade de transição da cadeia depende de uma parcela variável dos mais recentes símbolos que foram observados. Formalmente, dizemos que processos dessa natureza são cadeias estocásticas com memória de alcance variável.

Definição 2.21 (Cadeia estocástica com memória de alcance variável) *Dizemos que uma cadeia estocástica $\{X_t : t \in \mathbb{Z}\}$ definida em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$ é **compatível** com uma árvore de contexto probabilística (τ, Q) quando para P-quase toda sequência infinita à esquerda $w \in \mathcal{A}^{-\mathbb{N}}$ e qualquer símbolo $a \in \mathcal{A}$ temos*

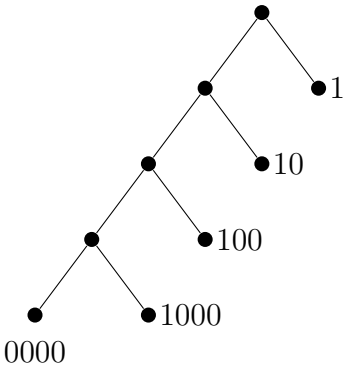
$$P\left(X_0 = a \mid X_{-\infty}^{-1} = w\right) = Q\left(a \mid c_\tau\left(w\right)\right),$$

*em que $c_\tau(w)$ é o contexto da sequência w. Em outras palavras, $c_\tau(w)$ representa a informação relevante do passado que determina a transição da cadeia. E essas cadeias são chamadas de **cadeias estocásticas com memória de alcance variável**.*

Exemplo 2.22 *Suponha que um processo estocástico $\{X_t : t \in \mathbb{Z}\}$ definido em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$, possa ser modelado como uma cadeia estocástica com memória de alcance variável com valores no alfabeto $\mathcal{A} = \{0, 1\}$. Além disso, considere que a informação relevante do passado, que influencia a transição da cadeia, é determinada pelo tempo até a ocorrência do primeiro símbolo 1. Por questões didáticas, vamos nos restringir a uma árvore τ de tamanho $h(\tau) = 4$. A Figura 2.4 ilustra o par ordenado (τ, Q) que representa essa árvore de contexto probabilística.*

Nesse caso, bastam 5 parâmetros para descrever a dinâmica da cadeia estocástica com alcance de memória variável de tamanho $h(\tau) = 4$, enquanto seriam necessários 16 parâmetros para descrever uma cadeia Markoviana com ordem $k = 4$.

A classe das cadeias estocásticas com memória de alcance variável foi introduzida por Rissanen (1983a). Essa classe ficou popular na comunidade estatística e probabilística



$$Q = \begin{array}{c} \\ 1 \\ 10 \\ 100 \\ 1000 \\ 0000 \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{bmatrix} Q(0 \mid 1) & Q(1 \mid 1) \\ Q(0 \mid 10) & Q(110) \\ Q(0 \mid 100) & Q(1 \mid 100) \\ Q(0 \mid 1000) & Q(1 \mid 1000) \\ Q(0 \mid 0000) & Q(1 \mid 0000) \end{bmatrix} \end{array},$$

Figura 2.4: Representação gráfica da árvore de contexto $\tau = \{1{,}10, 100, 1000, 0000\}$. Note que τ é uma árvore completa, já que toda sequência infinita à esquerda possui sufixo em τ. Além disso, representamos de forma matricial a família de probabilidades de transição Q definida sobre $\mathcal{A}$.

com o trabalho de Bühlmann e Wyner (1999) que cunhou o termo **cadeias de Markov com memória de tamanho variável** para se referir às cadeias compatíveis com árvores cujos contextos eram todos finitos. Mais tarde, surge o termo cadeias estocásticas com memória de alcance variável proposto por Galves e Löcherbach (2008) para se referir às cadeias que são compatíveis com árvores que podem possuir contextos infinitos.

Capítulo 3

Inferência frequentista em árvores de contexto

Neste capítulo, abordamos as principais etapas e técnicas empregadas na análise de dados utilizando modelos de árvore de contexto. Começando pela inferência dos parâmetros do modelo no contexto frequentista, exploramos como determinar as características do modelo que melhor estimam os dados observados. Além disso, consideramos crucial a tarefa de seleção das próprias árvores de contexto. Isso inclui, em particular, a seleção da ordem de uma cadeia de Markov, uma tarefa essencial na modelagem de dados reais a partir de processos Markovianos. Em seguida, detalhamos o uso do algoritmo Contexto para a estimação de árvores de contexto compatíveis com processos não necessariamente Markovianos, além de explorar o processo de estimação de árvores de contexto com base em critérios de penalização da verossimilhança. Esse capítulo oferece uma compreensão aprofundada das técnicas frequentistas usadas para extrair informações dos dados por meio das árvores de contexto.

3.1 Inferência dos parâmetros do modelo

Em inferência estatística, cadeias de Markov de ordem finita exigem muitos parâmetros para serem inferidos. Se uma cadeia estocástica $\boldsymbol{X} := \{X_t : t \in \mathbb{Z}\}$ é de memória com alcance variável compatível com uma árvore de contexto probabilística (τ, Q) de forma que $h(\tau) = d < \infty$, então basta uma coleção de $(|\mathcal{A}| - 1)\,|\tau|$ probabilidades de transição para descrever a cadeia. Assim, é vantajoso, inferencialmente, considerar cadeias compatíveis com árvores de contexto probabilísticas, visto que em um processo de estimação na

abordagem Markoviana a quantidade de parâmetros a serem estimados é $(|\mathcal{A}|-1)\,|\mathcal{A}|^d$, podendo ser bem superior ao da abordagem via cadeias estocásticas com memória de alcance variável a depender do tamanho do alfabeto.

Dados um número inteiro n e uma amostra x_1^n de símbolos de $\mathcal{A}$ geradas por um processo estocástico $\boldsymbol{X}$ com lei P compatível com uma árvore de contexto probabilística (τ, Q), o problema de inferência estatística está relacionado com a estimação da família de probabilidades de transição Q que determinam completamente a lei P do processo estocástico $\boldsymbol{X}$. Para qualquer sequência finita w de símbolos de $\mathcal{A}$ e para qualquer símbolo $a \in \mathcal{A}$, denotamos por $N_n(w,a)$ o número de ocorrências da sequência wa na amostra x_1^n, i.e.,

$$N_n(w,a) := \sum_{t=|w|+1}^{n} \mathbb{I}\left\{x_{t-|w|}^{t-1} = w, x_t = a\right\},$$

em que $\mathbb{I}$ representa a função indicadora e $N_n(w)$ denota a soma de $N_n(w,a)$ sobre todos os símbolos $a \in A$, ou seja,

$$N_n(w) = \sum_{a \in A} N_n(w,a).$$

Exemplo 3.1 *Seja* $\boldsymbol{X} := \{X_t : t \in \mathbb{Z}\}$ *um processo estocástico compatível com uma árvore de contexto probabilística definido em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores no alfabeto* $\mathcal{A} = \{0,1\}$ *e* $x_1^9 = (010010110)$ *uma amostra de símbolos de* $\mathcal{A}$ *geradas por* $\boldsymbol{X}$*. Dessa forma, temos* $N_n(w,a)$ *sendo o número de ocorrências da sequência* $01a$ *na amostra* x_1^9*, para qualquer* $a \in \mathcal{A}$*, isto é,*

$$N_n(01,0) = \sum_{t=3}^{9} \mathbb{I}\left\{x_{t-2}^{t-1} = 01, x_t = 0\right\} = 2;$$

$$N_n(01,1) = \sum_{t=3}^{9} \mathbb{I}\left\{x_{t-2}^{t-1} = 01, x_t = 1\right\} = 1.$$

Nesse sentido, $N_n(01)$ *é dado por*

$$N_n(01) = \sum_{a \in A} N_n(01,a) = 3.$$

Seja d um número inteiro positivo tal que $d < n$ e dada qualquer árvore de contexto probabilística (τ, Q) com $h(\tau) \leq d$, a função de verossimilhança [1] $\mathcal{L}_\tau\left(Q \mid x_1^n\right)$ condicionada

[1] A função de verossimilhança é utilizada para avaliar a plausibilidade de um conjunto de parâmetros em um modelo estatístico. Ela calcula a probabilidade ou densidade da amostra observada, assumindo que

em x_{-d+1}^{n} é dada por

$$\begin{aligned}
\mathcal{L}_\tau\left(Q \mid x_{-d+1}^{n}\right) &= P\left(X_{-d+1}^{n} = x_{-d+1}^{n}\right) \\
&= P\left(X_n = x_n \mid X_{-d+1}^{n-1} = x_{-d+1}^{n-1}\right) P\left(X_{-d+1}^{n-1} = x_{-d+1}^{n-1}\right) \\
&= Q\left(x_n \mid c_\tau\left(x_{-d+1}^{n-1}\right)\right) P\left(X_{-d+1}^{n-1} = x_{-d+1}^{n-1}\right) \\
&= Q\left(x_n \mid c_\tau\left(x_{-d+1}^{n-1}\right)\right) Q\left(x_{n-1} \mid c_\tau\left(x_{-d+1}^{n-2}\right)\right) P\left(X_{-d+1}^{n-2} = x_{-d+1}^{n-2}\right) \\
&\vdots \\
&= Q\left(x_n \mid c_\tau\left(x_{-d+1}^{n-1}\right)\right) \ldots Q\left(x_{d+1} \mid c_\tau\left(x_{-d+1}^{d}\right)\right) P\left(X_{-d+1}^{d} = x_{-d+1}^{d}\right).
\end{aligned}$$

Repare que a sequência x_{-d+1}^{d} se refere ao passado distante do processo subjacente à árvore τ. Dado que o processo é estacionário, temos a distribuição conjunta das variáveis aleatórias independentemente do tempo da escolha inicial. Portanto, podemos considerar que $P\left(X_{-d+1}^{d} = x_{-d+1}^{d}\right) = 1$.

Logo, a função de verossimilhança $\mathcal{L}_\tau\left(Q \mid x_1^n\right)$ condicionada sobre x_{-d+1}^{d} é dada por

$$\mathcal{L}_\tau\left(Q \mid x_1^n\right) = \prod_{w \in \tau} \prod_{a \in A} Q(a \mid w)^{N_n(w,a)}, \tag{3.2}$$

com a convenção que $0^0 = 1$.

Portanto, os estimadores de máxima verossimilhança das probabilidades de transição $\hat{Q}(a \mid w)$ são obtidos por meio da função de verossimilhança apresentada em 3.2. Seja $\mathcal{L}_\tau\left(Q \mid x_1^n\right)$ a função de verossimilhança condicionada sobre $x_{-\infty}^{d}$ e $\ell_\tau\left(Q \mid x_1^n\right)$ a função de log-verossimilhança associada definida como

$$\begin{aligned}
\ell_\tau\left(Q \mid x_1^n\right) &= \ln(\mathcal{L}_\tau\left(Q \mid x_1^n\right)) \\
&= \sum_{w \in \tau} \sum_{a \in \mathcal{A}} N_n(w,a) \ln(Q(a \mid w)).
\end{aligned}$$

Segue que o problema consiste em encontrar $Q \in \mathcal{Q}$, em que $\mathcal{Q}$ é a família de probabilidade de transição sobre $\tau \times \mathcal{A}$, que maximize a probabilidade de ocorrência dos dados

os parâmetros assumem um valor dado. O método de máxima verossimilhança consiste em encontrar os valores dos parâmetros do modelo estatístico que maximizam a probabilidade ou densidade de ocorrência da amostra observada.

observados, isto é,

$$\hat{Q}(a \mid w) := \arg\max_{Q \in \mathcal{Q}} \ell_\tau \left(Q \mid x_1^n\right),$$

com x_1^n fixado e sujeito a restrição

$$\sum_{a \in \mathcal{A}} Q(a \mid w) = 1, \quad \forall w \in \tau.$$

Uma vez que o problema de maximização envolve restrições, usamos o método de multiplicadores de Lagrange para determinar $\hat{Q}(a \mid w)$. Portanto, o problema se torna

$$\hat{Q}(a \mid w) := \arg\max_{Q \in Q_k} L(\lambda, Q),$$

sendo

$$L(\lambda, Q) = \ell_\tau \left(Q \mid x_1^n\right) + \lambda \left(1 - \sum_{a \in \mathcal{A}} Q(a \mid w)\right),$$

em que λ é o multiplicador de Lagrange associado à restrição.

O máximo da função Lagrangeana é encontrado analisando seus pontos críticos em relação às duas variáveis desconhecidas (λ, Q). Logo,

$$\frac{\partial L(\lambda, Q)}{\partial \lambda} = 1 - \sum_{a \in \mathcal{A}} Q(a \mid w), \quad \forall w \in \tau;$$

$$\frac{\partial L(\lambda, Q)}{\partial Q(a \mid w)} = \frac{N_n(w, a)}{Q(a \mid w)} - \lambda, \quad \forall a \in \mathcal{A} \text{ e } \forall w \in \tau.$$

Igualando a zero, obtemos

$$\sum_{a \in \mathcal{A}} \hat{Q}(a \mid w) = 1, \quad \forall w \in \tau; \tag{3.3}$$

$$\hat{Q}(a \mid w) = \frac{N_n(w, a)}{\lambda}, \quad \forall a \in \mathcal{A} \text{ e } \forall w \in \tau. \tag{3.4}$$

Das equações (3.3) e (3.4), segue que

$$\sum_{a \in \mathcal{A}} \frac{N_n(w, a)}{\lambda} = 1 \Rightarrow \lambda = \sum_{a \in \mathcal{A}} N_n(w, a), \quad \forall w \in \tau.$$

Consequentemente,

$$\hat{Q}(a \mid w) = \frac{N_n(w,a)}{N_n(w)} \quad \forall a \in \mathcal{A} \text{ e } \forall w \in \tau, \tag{3.5}$$

se $N_n(w) > 0$ e definimos $\hat{Q}(a \mid w) = 1/|\mathcal{A}|$ quando $N_n(w) = 0$. Além disso, quando o processo estocástico $\boldsymbol{X}$ é estacionário, é possível provar que o estimador de máxima verossimilhança $\hat{Q}(a \mid w)$ é fortemente consistente (ver Galves e Leonardi (2008)).

Substituindo os estimadores (3.5) em (3.2), obtemos que a probabilidade de processo estocástico, com determinada matriz de transição $\hat{Q}$, gerar uma amostra x_1^n é dada por

$$\hat{P}\left(X_1^n = x_1^n\right) = \prod_{w \in \tau} \prod_{a \in A} \hat{Q}(a \mid w)^{N_n(w,a)}.$$

Exemplo 3.6 *Sejam* $\boldsymbol{X_1} := \{X_{1t} : t \in \mathbb{Z}\}$ *e* $\boldsymbol{X_2} := \{X_{2t} : t \in \mathbb{Z}\}$ *processos estocásticos compatíveis com uma árvore de contexto probabilística e definidos em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores no alfabeto* $\mathcal{A} = \{0, 1\}$ *e seja a amostra* $x_1^{20} = (10010010010001111001)$. *Estamos interessados em estimar as probabilidades de transição associadas aos dois processos, em que* $\tau_1 = \{00, 01, 10, 11\}$ *e* $\tau_2 = \{00, 10, 1\}$ *são as árvores de contexto dos processos* $\boldsymbol{X_1}$ *e* $\boldsymbol{X_2}$, *respectivamente.*

(a) Árvore τ_1 (b) Árvore τ_2.

Figura 3.1: Representação gráfica das árvores de contexto $\tau_1 = \{00, 01, 10, 11\}$, associada à cadeia de Markov de ordem $k = 2$ e $\tau_2 = \{00, 10, 1\}$, associada à cadeia de Markov com memória de alcance variável, com $h(\tau_1) = h(\tau_2) = 2$.

Dada a amostra observada x_1^{20}, *devemos calcular as quantidades* $N_{20}(w, a)$ *e* $N_{20}(w)$, $\forall\ w \in \tau_1$ *e* τ_2 *e* $\forall a \in \mathcal{A}$, *representadas na Tabela 3.1.*

Tabela 3.1: Número de ocorrências da sequência wa na amostra x_1^{20}, para todo $w \in \tau_1$ e τ_2 e para todo $a \in \mathcal{A}$.

a	$N_{20}(00, a)$	$N_{20}(01, a)$	$N_{20}(10, a)$	$N_{20}(11, a)$	$N_{20}(1, a)$
0	1	3	5	1	5
1	5	1	0	2	3
$N_{20}(w)$	6	4	5	3	8

A partir dos valores apresentados na Tabela 3.1, conseguimos estimar as probabilidades de transição e, consequentemente, a matriz de transição associada a cada árvore.

$$\hat{Q}_1 = \begin{array}{c} \\ 00 \\ 01 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0,17 & 0,83 \\ 0,75 & 0,25 \\ 1 & 0 \\ 0,33 & 0,67 \end{bmatrix} \end{array}, \qquad \hat{Q}_2 = \begin{array}{c} \\ 00 \\ 10 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0,17 & 0,83 \\ 1 & 0 \\ 0,625 & 0,375 \end{bmatrix} \end{array}.$$

Figura 3.2: Representação matricial das matrizes de transição associadas as árvore de contexto $\tau_1 = \{00, 01, 10, 11\}$ e $\tau_2 = \{00, 10, 1\}$.

Uma vez estimada a matriz de transição, é possível estimar a probabilidade da amostra observada ter sido gerada pelo processo estocástico em questão, levando em consideração cada um dos casos. Consequentemente

$$P(X_1^n = x_1^n \mid \hat{Q}_1) = \prod_{w \in \tau_1} \prod_{a \in A} \hat{Q}_1(a \mid w)^{N_n(w,a)} = 0,0010;$$
$$P(X_1^n = x_1^n \mid \hat{Q}_2) = \prod_{w \in \tau_2} \prod_{a \in A} \hat{Q}_2(a \mid w)^{N_n(w,a)} = 0,00003.$$

Observe a importância de definir a árvore de contexto subjacente ao processo estocástico ao estimar as probabilidades de transição, já que árvores diferentes podem estar relacionadas a diferentes matrizes de transição e probabilidades. No entanto, na prática, não possuímos conhecimento prévio do modelo subjacente à amostra, ou seja, a árvore de contexto é desconhecida. Portanto, é necessário estudar métodos que permitam estimar a árvore de contexto e, consequentemente, as probabilidades de transição.

3.2 Seleção de árvores de contexto

O problema da seleção estatística de árvores de contexto está relacionado a escolher um modelo estatístico a partir de um conjunto de modelos candidatos, utilizando um critério pré-estabelecido. Em outras palavras, busca-se selecionar a árvore de contexto mais provável de ter gerado uma amostra x_1^n. Nesse sentido, o princípio MDL (do inglês *Minimum Description Length*) desempenha um papel fundamental.

O princípio MDL fundamenta-se na ideia de que o melhor modelo estatístico para uma determinada amostra de dados é aquele que proporciona a melhor compressão dos dados, ou seja, aquele que consegue representar a amostra de forma mais concisa segundo alguns critérios. Portanto, o processo de seleção estatística de árvores de contexto envolve a comparação entre diferentes modelos candidatos, visando escolher aquele que equilibra de forma ideal a capacidade de representação dos dados e a complexidade do próprio modelo. Dessa forma, a abordagem incorpora o princípio da parcimônia, permitindo que modelos mais complexos sejam preferidos somente quando eles realmente oferecem uma melhoria significativa na descrição dos dados.

Neste capítulo, estudamos três métodos de seleção baseados no princípio MDL e algumas de suas variações: a seleção da ordem de uma cadeia de Markov, o algoritmo contexto (Rissanen, 1983a) e suas variações (Galves *et al.*, 2012; Galves e Leonardi, 2008), o critério de penalização de verossimilhança (Leonardi, 2006) e suas variações (Csiszár e Talata, 2006).

Começamos considerando uma árvore compatível com uma cadeia de Markov e, portanto, o problema de seleção da árvore de contexto se restringe ao problema de estimação da ordem da cadeia.

3.2.1 Seleção da ordem de uma cadeia de Markov

Ao considerar cadeias de Markov, a seleção de modelos consiste em determinar a ordem k da cadeia. Nesse sentido, sabemos que a cadeia possui ordem finita, mas desconhecemos tal ordem.

Dados um número inteiro positivo n e uma amostra x_1^n de símbolos de $\mathcal{A}$ geradas por um processo estocástico $\boldsymbol{X}$ definido em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$ com valores no alfabeto $\mathcal{A}$ e compatível com uma cadeia de Markov de ordem k finita, vamos denotar a matriz de transição associada ao processo por

$$Q^k := \{Q(a \mid a_1^k) : a \in \mathcal{A} \quad \text{e} \quad a_1^k \in \mathcal{A}^k\}.$$

Assim, queremos definir o valor $\hat{k}$ tal que [2] $\hat{k} \ll n$ e o problema consiste em encontrar

[2] Em que $\hat{k} \ll n$ denota uma desigualdade estrita e significa que k é muito menor que n. Definimos que $\hat{k} \ll n \Rightarrow \hat{k} < \left\lfloor \frac{\ln_{|\mathcal{A}|} n}{2} \right\rfloor$, em que $\lfloor . \rfloor$ é função que retorna o maior inteiro .(Girardi, 2021).

o menor valor $\hat{k}$ para o qual Q^k seja suficientemente próxima de Q^{k+1}, ou seja,

$$\hat{k} = \min_k \{Q^k \approx Q^{k+1} : k \geq 0\},$$

em que Q^k denota a matriz de transição associada a um processo de ordem k e $Q^k \approx Q^{k+1}$ significa que

$$\max_{a \in \mathcal{A}} \max_{a^* \in \mathcal{A}} \left| Q^k \left(a \mid a_{-k}^{-1} \right) - Q^{k+1} \left(a \mid a_{-k}^{-1} a^* \right) \right| \leq \delta, \quad \forall a_{-k}^{-1} \in \mathcal{A}^k,$$

em que $\delta \in (0, 1)$ é o limiar fixado, ou seja, o valor crítico que não pode ser ultrapassado para considerarmos que Q^k é suficientemente próxima de Q^{k+1}. Perceba que ao considerar essa condição, estamos encontrando o menor valor de k para o qual a diferença entre Q^k e Q^{k+1} seja suficientemnte pequena para todas as combinações possíveis de a e a^*.

Segue que queremos testar, sequencialmente, a hipótese nula

$$H_0 : Q^k \left(a \mid a_{-k}^{-1} \right) = Q^{k+1} \left(a \mid a_{-k-1}^{-1} \right), \quad \forall a \in \mathcal{A} \quad \text{e} \quad \forall a_{-k-1}^{-1} \in \mathcal{A}^{k+1},$$

utilizando a estatística de teste

$$\Delta(a_{-k}^{-1}) = \max_{a \in \mathcal{A}} \max_{a^* \in \mathcal{A}} \left| Q^k \left(a \mid a_{-k}^{-1} \right) - Q^{k+1} \left(a \mid a_{-k}^{-1} a^* \right) \right|,$$

até ser encontrado um valor $\hat{k}$ para o qual a hipótese nula seja rejeitada. Neste caso, rejeitamos H_0 quando $\Delta(a_{-k}^{-1}) > \delta$ para algum $a_{-k}^{-1} \in \mathcal{A}^k$.

A fim de obter uma estimação factível para a ordem da cadeia, é fundamental levar em conta amostras que resultem em uma matriz de transição estocástica estimada.

Definição 3.7 (Amostra k-admissível) *Sejam n um número inteiro positivo e $x_1^n \in \mathcal{A}^n$ uma amostra de uma cadeia de Markov $\boldsymbol{X}$ definida em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$ com valores no alfabeto $\mathcal{A}$ e ordem k desconhecida e finita com $k < n$. A amostra x_1^n é dita k-admissível quando para toda sequência $a_1^k \in \mathcal{A}^k$ temos $N_n(a_1^k) \geq 1$.*

Assim, dada uma amostra x_1^n k-admissível gerada por uma cadeia de Markov $\boldsymbol{X}$, o problema de estimar a ordem k da cadeia pode ser resolvido por meio do seguinte algoritmo

1. Catalogar todas as sequências de tamanho $d(n) = \left\lfloor \frac{\ln_{|\mathcal{A}|} n}{2} \right\rfloor$ presentes na amostra x_1^n;

2. Definir $\delta \in (0,1)$ pequeno e $k = d(n) - 1$;

3. Para cada sequência a_{-k}^{-1} catalogada, calcule

$$\Delta(a_{-k}^{-1}) = \max_{a \in \mathcal{A}} \max_{a^* \in \mathcal{A}} \left| \hat{Q}^k \left(a \mid a_{-k}^{-1} \right) - \hat{Q}^{k+1} \left(a \mid a_{-k}^{-1} a^* \right) \right|;$$

4. Se $\max \Delta(a_{-k}^{-1}) \geq \delta$ com $a_{-k}^{-1} \in \mathcal{A}^k$, pare e defina $\hat{k} = k + 1$. Caso contrário, $\max \Delta(a_{-k}^{-1}) < \delta$ com $a_{-k}^{-1} \in \mathcal{A}^k$, descarte a ordem $k+1$ da cadeia, então definimos $k = k - 1$ e repetimos o processo a partir do passo (3).

Note que o processo começa considerando a maior ordem k para amostra x_1^n e diminui esse valor iterativamente até que sejam encontradas duas matrizes que não possam ser consideradas suficientemente próximas.

Exemplo 3.8 *Considere uma amostra* $x_1^{30} = 000000001101101010010101001010$ *obtida por meio de uma cadeia de Markov de ordem* $k = 1$ *definida em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores no alfabeto* $\mathcal{A} = \{0, 1\}$. *Suponha que o valor* k *seja desconhecido e que desejamos selecionar o modelo mais provável de ter gerado a amotra observada, ou seja, queremos estimar a ordem* k *da cadeia. Para tanto, consideramos o algoritmo definido.*

1. *Catalogar todas as sequências de tamanho* $d(n) = \left\lfloor \frac{\ln_2 30}{2} \right\rfloor = 2$.

Para a amostra em questão, observamos as seguintes sequências de tamanho 2*:* $00, 01, 10, 11$ *e, portanto, a amostra é* k*-admissível para* $k = 2$.

2. *Definimos* $\delta = 0,1$ *e* $k = 1$.

3. *Para cada sequência* a_{-k}^{-1} *catalogada, calcular*

$$\Delta(a_{-k}^{-1}) = \max_{a \in \mathcal{A}} \max_{a^* \in \mathcal{A}} \left| Q^k \left(a \mid a_{-k}^{-1} \right) - Q^{k+1} \left(a \mid a_{-k}^{-1} a^* \right) \right|$$

Dada a amostra observada, devemos, primeiramente, estimar, como indicado na Seção 3.1, as matrizes de transição que serão utilizadas no cálculo da estatística de teste. Temos

$$\hat{Q}^1 = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0,8182 & 0,1818 \\ 0,5714 & 0,4286 \end{bmatrix} \end{array}, \qquad \hat{Q}^2 = \begin{array}{c} \\ 00 \\ 10 \\ 01 \\ 11 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0,8235 & 0,1765 \\ 0,7500 & 0,2500 \\ 0,5000 & 0,5000 \\ 0,6667 & 0,3333 \end{bmatrix} \end{array}.$$

Neste sentido, primeiramente fixando $a = 0$ *e calculando a estatística de teste para todas as sequências catalogadas*

$$\begin{aligned} \Delta(a_{-k}^{-1}) &= \max\left\{|0,8182 - 0,8235|, |0,8182 - 0,75|, |0,5714 - 0,5|, |0,5714 - 0,6667|\right\} \\ &= |0,5714 - 0,6667| = 0,0953. \end{aligned}$$

Agora, consideramos $a = 1$*,*

$$\begin{aligned} \Delta(a_{-k}^{-1}) &= \max\left\{|0,1818 - 0,25|, |0,1818 - 0,1765|, |0,4286 - 0,5|, |0,4286 - 0,3333|\right\} \\ &= |0,4286 - 0,3333| = 0,0953. \end{aligned}$$

Logo, $\max\{0,0953; 0,0953\} = 0,0953 < 0,1$.

4. *Como* $\Delta(a_{-k}^{-1}) < 0,1$*, descartamos a ordem* $k+1 = 2$ *da cadeia e definimos um novo* $k = k - 1 = 0$ *e repetimos o processo a partir do passo (3).*

Repetimos o processo, considerando $k = 0$ *e comparando a matrizes* Q^0 *e* Q^1*, obtemos* $\Delta(a_{-k}^{-1}) = 0,1953 > 0,1$*. Portanto, paramos o processo e definimos* $\hat{k} = k + 1 = 1$.

3.2.2 Estimação via algoritmo Contexto

O algoritmo Contexto, introduzido por Rissanen (1983a), calcula, para cada nó de uma árvore, uma medida de discrepância entre a probabilidade de transição associada a esse nó e as respectivas probabilidades de transição associada aos filhos desse nó, ou seja, à concatenação de um único símbolo a esse nó. O algoritmo é iniciado catalogando todas as sequências de tamanho d e definindo a primeira árvore candidata como sendo aquela cujas folhas possuem tamanho d e tal que cada um de seus contextos apareça pelo menos uma vez na amostra. Em seguida, definimos $\mathcal{V}_n(x_1^n)$ como sendo o conjunto de todas as

sequências finitas $w \in \mathcal{A}^*$ que aparecem pelo menos uma vez na amostra x_1^n, isto é,

$$\mathcal{V}_n(x_1^n) := \{w \in \mathcal{A}^* : N_n(w) \geq 1\},$$

perceba que todos os nós da árvore que aparecem pelo menos uma vez na amostra pertencem ao conjunto $\mathcal{V}_n(x_1^n)$.

Similarmente ao algoritmo para seleção da ordem da cadeia de Markov, também precisamos definir um limiar δ como sendo um número positivo, em geral pequeno. Nesse caso, estabelecemos a seguinte regra de decisão: se uma medida de discrepância for maior do que δ, os contextos são mantidos na árvore, caso contrário, eles são podados, pois não existe ganho preditivo ao considerar um alcance maior. O procedimento é repetido até não haver mais podas na árvore.

A medida de discrepância utilizada no algoritmo Contexto proposto por Rissanen (1983a) é baseada na medida de divergência de Küllback-Leibler. Definidas duas medidas de probabilidades p e q sobre $\mathcal{A}$ temos a divergênca entre elas, denotada

$$D(p|q) := \sum_{a \in A} p(a) \ln\left(\frac{p(a)}{q(a)}\right),$$

em que, por convenção consideramos $p(a)\ln\left(\frac{p(a)}{q(a)}\right) = 0$ se $p(a) = 0$ e $p(a)\ln\left(\frac{p(a)}{q(a)}\right) = +\infty$ se $p(a) > q(a) = 0$.

Para uma dada sequência $w \in \mathcal{V}_n(x_1^n)$, Rissanen (1983a) definimos a seguinte medida de discrepância

$$\Delta_n(w) := \sum_{b:bw \in \mathcal{V}_n(x_1^n)} N_n(bw) D\left(\hat{p}(\cdot|bw)\,|\hat{p}(\cdot|w)\right).$$

Se $\Delta_n(w) < \delta$, então poda-se a folha, ou seja, descartamos todos os filhos associados ao nó w. Caso contrário, não realizamos a poda e tomamos $bw \in \hat{\tau}$, para todo $b \in \mathcal{A}$. Realizamos esse procedimento para toda sequência catalogada e repetimos o processo até que mais nenhuma poda seja executada. Ao final do processo, definimos como $\hat{\tau}_c(x_1^n)$ a árvore de contexto estimada.

Observe que a medida de discrepância $\Delta_n(w)$ para uma sequência $w \in \mathcal{V}_n(x_1^n)$ mede o quão importante é manter os filhos associados ao nó w para melhorar o poder preditivo da cadeia estocástica com alcance de memória variável. Definimos então o seguinte algoritmo:

1. Catalogar todas as sequências de tamanho d presentes na amostra x_1^n e definir a

árvore máxima de tamanho d garantindo que todos os nós pertençam a $\mathcal{V}_n(x_1^n)$;

2. Definir $\delta \in (0,1)$ pequeno;

3. Atribuir a cada nó w da árvore uma função indicadora $C_w(x_1^n)$. Inicializar as folhas com o valor zero e, para os demais nós, definir

$$C_w(x_1^n) = \max\left\{\mathbb{I}\{\Delta_n(w) > \delta\}, \max_{b\in\mathcal{A}} C_{bw}(x_1^n)\right\}.$$

Após a conclusão do algoritmo, cada nó é marcado com uma função indicadora. A árvore de contexto estimada é dada por

$$\hat{\tau}_c(x_1^n) = \{w \in \mathcal{V}_n(x_1^n) : C_w(x_1^n) = 0 \text{ e } C_u(x_1^n) = 1 \text{ para todo } u \prec w\} \tag{3.9}$$

Exemplo 3.10 *Considere a amostra* $x_1^{30} = 001000000110000111110111101000$ *obtida de um processo estocástico* $\boldsymbol{X}$ *compatível com uma árvore de contexto probabilística definida em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores no alfabeto* $\mathcal{A} = \{0,1\}$. *Estamos interessados em estimar a árvore de contexto subjacente ao processo* $\boldsymbol{X}$, *por meio do algoritmo Contexto. Considere* $d = 2$.

0. Definir o conjunto $\mathcal{V}_n(x_1^{30})$;

$$\mathcal{V}_n(x_1^{30}) = \{0, 1, 00, 01, 11, 10, 000, 001, 011, 110, 101, 010, 100, \cdots\}$$

1. Catalogar todas as sequências de tamanho $d = 2$ *presentes na amostra* x_1^{30} *e definir a árvore máxima de tamanho* d;

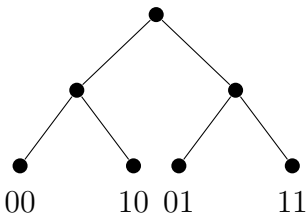


Figura 3.4: Árvore máxima de tamanho $d = 2$, tal que todos os nós pertençam a $\mathcal{V}_n(x_1^{30})$.

2. Definimos $\delta = 0,1$;

3. Inicialize todas as folhas como 0 *e atribua a cada nó da árvore a função indicadora*

$$C_w(x_1^{30}) = \max\left\{\mathbb{1}\{\Delta_n(w) > \delta\}, \max_{b\in\mathcal{A}} C_{bw}(x_1^{30})\right\}.$$

Considere, primeiramente, $w = 0$. *Temos*

$$\Delta_n(w) = \sum_{b:b01\in\mathcal{V}_n(x_1^{30})} N_n(b0) \sum_{a\in\mathcal{A}} \hat{p}(a \mid b0) ln \left(\frac{\hat{p}(a \mid b0)}{\hat{q}(a \mid 0)} \right) = 0,0889 \quad e$$

$$C_w(x_1^{30}) = \max\{0, \max\{0, 0\}\} = 0.$$

Repetimos o procedimento considerando $w = 1$ *e obtemos* $C_w(x_1^{30}) = 0$.

Considere a Figura 3.5 para visualizar os valores de $C_w(x_1^{30})$ *em cada nó da árvore máxima de tamanho* $d = 2$.

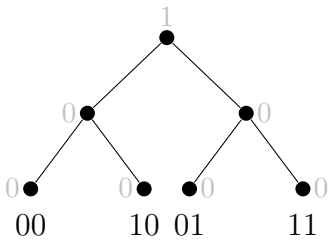


Figura 3.5: Árvore máxima de tamanho $d = 2$, tal que todos os nós pertençam a $\mathcal{V}_n(x_1^{30})$. Valores em preto indicam os contextos associados a essa árvore, enquanto valores em cinza indicam a função indicadora $C_w(x_1^{30})$ associada ao nó w.

A partir da Figura 3.5, concluímos que a árvore de contexto estimada por meio do algoritmo Contexto é dada por $\hat{\tau}_c(x_1^{30}) = \{0, 1\}$, *ou seja, observada a amostra* x_1^{30} *é plausível assumir que o processo estocástico subjacente é descrito como uma cadeia de Markov de ordem* $k = 1$.

De fato, a amostra foi obtida por meio de uma cadeia de Markov de ordem $k = 1$.

Embora a medida de discrepância Δ_n tenha sido originalmente proposta por Rissanen (1983a), outras possibilidades tem sido propostas na literatura. Neste trabalho, abordaremos as modificações do algoritmo Contexto proposta por Galves e Leonardi (2008). Dessa forma, definimos medida de discrepância como

$$\Delta_n(w) = \max_{a\in\mathcal{A}} |\hat{p}_n(a \mid w) - \hat{p}_n(a \mid suf(w))|,$$

em que $suf(w)$ representa o maior sufixo de w e $\hat{p}_n(a \mid w)$ denota a estimativa de máxima verossimilhança da probabilidade de transição associada à sequência w na amostra x_1^n.

A expressão $\Delta_n(w)$ calcula a distância máxima entre a probabilidade de transição empírica associada à sequência w e a mesma probabilidade empírica associada ao maior

sufixo de w. Essa medida permite avaliar se há ganhos preditivos ao considerar mais um passo no passado. Neste sentido, se $\Delta_n(w) < \delta$, significa que não há ganho preditivo ao considerar um passo a mais no passado. Nesse caso, definimos $suf(w)$ como o novo candidato a contexto da árvore. O procedimento é repetido até que observe um ganho preditivo significativo, ou seja, até que não seja possível fazer mais podas na árvore.

Em resumo, essa abordagem busca identificar a extensão do passado que é relevante para a predição, removendo contextos desnecessários e otimizando o modelo preditivo, reduzindo o número de parâmetros e custo computacional. Portanto, definimos o seguinte algoritmo Contexto:

1. Catalogar todas as sequências de tamanho d presentes na amostra x_1^n;

2. Definir $\delta \in (0, 1)$ pequeno;

3. Para cada sequência w catalogada, calcule

$$\Delta_n(w) = \max_{a \in \mathcal{A}} |\hat{p}_n(a \mid w) - \hat{p}_n(a \mid suf(w))| \,,$$

   se $\Delta_n(w) \leq \delta$, poda-se a folha e definimos $w = suf(w)$ como novo candidato a contexto. Repita o processo até não observar mais podas na árvore, i.e, $\Delta_n(w) > \delta$ e adicione w a árvore estimada $\hat{\tau}_{cm}(x_1^n)$

Após a conclusão do algoritmo, a árvore de contexto estimada é dada por

$$\hat{\tau}_{cm}(x_1^n) = \left\{ w \in \mathcal{A}_1^d : \Delta_n(w) > \delta \text{ e } \Delta_n(uw) \leq \delta, \quad \forall u \in \mathcal{A}_1^{d-|w|} \right\}, \tag{3.11}$$

em que $\mathcal{A}_1^d$ denota o conjunto com todas as sequências de tamanho 1 até d. Além disso, perceba que $\mathcal{A}_1^{d-|w|} = \emptyset$ quando $|w| = d$.

É possível provar que os estimadores obtidos por meio do algoritmo Contexto com as modificações propostas é fortemente consistente (ver Galves e Leonardi (2008)).

Exemplo 3.12 *Considere a amostra* $x_1^{30} = 00100000011000011111010111101000$ *obtida de um processo estocástico* $\boldsymbol{X}$ *compatível com uma árvore de contexto probabilística definida em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores no alfabeto* $\mathcal{A} = \{0, 1\}$. *Estamos interessados em estimar a árvore de contexto subjacente ao processo* $\boldsymbol{X}$, *por meio do algoritmo Contexto Modificado. Considere* $d = 2$.

1. *Catalogar todas as sequências de tamanho* $d = 2$ *presentes na amostra* x_1^{30}*;*

2. *Definimos* $\delta = 0,1$*;*

3. *Para cada sequência* w *catalogada, calcule*

$$\Delta_n(w) = \max_{a \in \mathcal{A}} |\hat{p}_n(a \mid w) - \hat{p}_n(a \mid suf(w))| .$$

Considere, primeiramente, $w = 00$*. Temos*

$$\Delta_n(w) = \max_{a \in \mathcal{A}} |\hat{p}_n(a \mid 00) - \hat{p}_n(a \mid 0)| = 0,0125,$$

como $\Delta_n(w) < \delta$*, definimos* $w = suf(w)$ *e calculamos*

$$\Delta_n(w) = \max_{a \in \mathcal{A}} |\hat{p}_n(a \mid 0) - \hat{p}_n(a)| = 0,1541,$$

como $\Delta_n(w) > \delta$*, adicionamos* w *na árvore estimada. Após a conclusão do algoritmo, concluímos que a árvore de contexto estimada por meio do algoritmo Contexto Modificado é dada por* $\hat{\tau}_{cm}(x_1^{30}) = \{0,1\}$*, ou seja, observada a amostra* x_1^{30} *é plausível assumir que o processo estocástico subjacente é descrito como uma cadeia de Markov de ordem* $k = 1$*.*

3.2.3 Estimação via critérios de penalização da verossimilhança

Um critério de penalização da verossimilhança é uma técnica estatística utilizada para ajustar modelos estatísticos complexos, levando em consideração tanto o ajuste aos dados observados (verossimilhança) quanto a complexidade do modelo. Essa penalização é necessária para evitar o *overfitting*, que ocorre quando um modelo se ajusta excessivamente aos dados de treinamento. A ideia básica é adicionar uma penalização à função de verossimilhança, e essa penalização aumenta à medida que a complexidade do modelo aumenta. Dessa forma, busca-se encontrar o equilíbrio entre um ajuste adequado aos dados e a complexidade do modelo.

Logo, a abordagem da máxima verossimilhança penalizada é baseada na função de verossimilhança definida em (3.2) com um termo de penalização apropriado. Dessa forma, a árvore estimada é aquela que maximiza a função de verossimilhança penalizada. Vale ressaltar que a função de verossimilhança (3.2) não precisa ser calculada sobre todas possíveis

árvores de contexto, mas apenas para aquelas árvores cujos contextos são observados na amostra. Essas árvores são denominadas admissíveis.

Definição 3.13 (Árvore de contexto admissível) *A árvore de contexto τ é dita ser* ***admissível*** *para a amostra x_1^n quando $|\tau| \leq d$, $N_n(w) \geq 1$ para todo $w \in \tau$ e para qualquer número inteiro j tal que $d \leq j \leq n-1$, existe uma sequência $w \in \tau$ de forma que w é sufixo de x_1^j.*

Seja $\mathcal{T}_n$ o conjunto de todas as árvores de contexto admissíveis para a amostra x_1^n. Os critérios que penalizam a função de verossimilhança são dadas por

$$\hat{\tau}_{pmv} := \arg\max_{\tau \in \mathcal{T}_n} \left\{ \log \hat{P}(X_1^n = x_1^n) - |\tau| pen(n) \right\},$$

em que $pen(n)$ é alguma função positiva tal que $pen(n) \to +\infty$ e $pen(n)/n \to 0$ quando $n \to \infty$.

Um dos métodos de penalização da verossimilhança comumente utilizado para estimar árvores de contexto a partir de uma amostra de uma série temporal é o Critério de Informação Bayesiano (BIC, do inglês *Bayesian Information Criterion*). Nesse caso, $pen(n) = c(|\mathcal{A}| - 1)\log(n)$ para alguma constante $c > 0$. A princípio, pode parecer praticamente impossível calcular $\hat{\tau}_{pmv}$, pois a maximização leva em consideração o conjunto de todas as árvores admissíveis, no entanto, Csiszár e Talata (2006) mostraram como obter um algoritmo simples e eficiente a partir da adaptação do *Context Tree Maximizing (CTM)*, proposto por Willems *et al.* (1995), que veremos em mais detalhes no próximo capítulo.

Galves *et al.* (2012) apresentam em seus estudos, o SMC (do inglês, *Smallest Maximizer Criterion*), um método de penalização que envolve otimizar a verossimilhança penalizada para diferentes valores de uma constante de penalização. Por meio desse processo, o SMC identifica um conjunto de "árvores campeãs", isto é, aquelas que maximizam a verossimilhança penalizada para cada valor possível da constante. O aspecto notável é que há uma transição no modo como a verossimilhança aumenta entre as árvores campeãs e as demais árvores, e a árvore que corresponde a essa transição é a árvore estimada pelo SMC. Isso torna o SMC um método eficaz para selecionar um modelo de árvore de contexto apropriado, baseado na maneira como a verossimilhança se comporta em relação às variações da constante de penalização.

Além disso, Garivier e Leonardi (2011) demonstraram que a árvore estimada por

meio da máxima verossimilhança penalizada é sempre menor ou igual à árvore obtida pelo algoritmo Contexto. Em outras palavras, para qualquer sequência $n \geq 1$ e todas as sequências x_1^n, se $\delta \leq pen(n)$, então $\hat{\tau}_{pmv}(x_1^n) \preceq \hat{\tau}c(x_1^n)$. Esse resultado indica que a abordagem da máxima verossimilhança penalizada produz árvores de contexto mais compactas em comparação com o algoritmo Contexto.

Exemplo 3.14 *Considere a amostra* $x_1^{30} = 001000000110000111110111101000$ *obtida de um processo estocástico* $\boldsymbol{X}$ *compatível com uma árvore de contexto probabilística definida em um espaço de probabilidade apropriado* $(\Omega, \mathcal{F}, P)$ *com valores no alfabeto* $\mathcal{A} = \{0, 1\}$. *Estamos interessados em estimar a árvore de contexto subjacente ao processo* $\boldsymbol{X}$, *por meio do critério de penalização de verossimilhança. Consideramos* $pen(n) = 0,5(|\mathcal{A}|-1)log(n)$ *como função de penalização. Considere* $d = 2$

Dada a amostra x_1^{30}, $\mathcal{T}_n = \{\{0, 1\}, \{00, 10, 1\}, \{0, 11, 01\}, \{00, 10, 01, 11\}\}$ *representa o conjunto de todas as árvores de contexto adimissíveis com* $d = 2$.

Nesse caso, o método consiste em calcular a função de verossimilhança penalizada para cada $\tau \in \mathcal{T}_n$, *de forma que*

$$\hat{\tau}_{pmv} := \arg\max_{\tau \in \mathcal{T}_n} \left\{ \log \hat{P}(X_1^{30} = x_1^{30}) - |\tau| 0,5 log(30) \right\},$$

Considere, primeiramete, $\tau = \{0, 1\}$. *Temos*

$$\hat{P}(X_1^{30} = x_1^{30}) = \prod_{w \in \tau} \prod_{a \in A} \hat{p}(a \mid w)^{N_n(w,a)} = 8,3667 \cdot 10^{-9}.$$

Logo,

$$\log \hat{P}(X_1^{30} = x_1^{30}) - 0,5 log(30) = -22,0002.$$

A Tabela 3.2 apresenta os valores da log verossimilhança e função de penalização para cada $\tau \in \mathcal{T}_n$.

Tabela 3.2: Valores necessários para o cálculo da árvore de contexto estimada via critério de penalização da verossimilhança, considerando $pen(n) = 0,5(|\mathcal{A}| - 1)log(n)$.

$\tau \in \mathcal{T}_n$	$\hat{\mathbf{P}}(\mathbf{X_1^{30}} = \mathbf{x_1^{30}})$	$\|\tau\|\mathbf{0,5log(30)}$
$\{0, 1\}$	$8,3667 \cdot 10^{-9}$	$3,4011$
$\{00, 10, 1\}$	$1,3302 \cdot 10^{-8}$	$5,1017$
$\{0, 01, 11\}$	$8,4007 \cdot 10^{-9}$	$5,1017$
$\{00, 10, 01, 11\}$	$1,3356 \cdot 10^{-8}$	$6,8023$

Dessa forma,

$$\hat{\tau}_{pmv} := \arg\max_{\tau \in \mathcal{T}_n} \{-22,0002; -23,2371; -23,6967; -24,9336\}\,.$$

Portanto, a árvore de contexto estimada por meio do critério de penalização de verossimilhança é dada por $\hat{\tau}_{pmv}(x_1^{30}) = \{0,1\}$*, ou seja, observada a amostra* x_1^{30} *é plausível assumir que o processo estocástico subjacente é descrito como uma cadeia de Markov de ordem* $k = 1$*.*

Capítulo 4

Inferência bayesiana em árvores de contexto

No contexto da análise de dados com modelos de árvore de contexto, este capítulo se dedica à abordagem de Inferência Bayesiana. Neste capítulo, exploramos como os princípios da Inferência Bayesiana podem ser aplicados para aprimorar a compreensão e interpretação dos dados. Em seguida, abordamos a ponderação de árvores de contexto, destacando como as considerações Bayesianas podem ser incorporadas para aprimorar a seleção do modelo. Por fim, apresentamos o algoritmo da árvore de contexto Bayesiana, que é uma ferramenta chave na aplicação dessa abordagem. Este capítulo oferece *insights* sobre como aproveitar os princípios da Inferência Bayesiana para obter resultados mais robustos.

4.1 Árvores de contexto Bayesianas

A classe de modelos denominada árvores de contextos Bayesianas (BCT, do inglês *Bayesian Context Tree*), introduzidas por Papageorgiou *et al.* (2021) e Kontoyiannis *et al.* (2022) têm se mostrado uma abordagem promissora para a modelagem de cadeias de Markov com memória de alcance variável, pois oferecem uma estrutura flexível e adaptativa que captura as dependências de longo prazo em séries temporais discretas.

Dada uma cadeia estocástica $\{X_t : t \in \mathbb{Z}\}$ definida em um espaço de probabilidade apropriado $(\Omega, \mathcal{F}, P)$ e associada a uma árvore de contexto probabilística (τ, Q), podemos definir a família de distribuição a *priori* em (τ, Q).

Definição 4.1 (Modelo a *priori*) *Seja $D \geq 0$ um número inteiro que representa a maior altura possível que uma árvore pertencente ao conjunto $\mathcal{T}(D)$ pode possuir, em que $\mathcal{T}(D)$ é o conjunto de todas as árvores de contexto sobre $\mathcal{A}$ com altura até D e $\mathcal{A}$ é um alfabeto finito tal que $|\mathcal{A}| \geq 2$. Dado um número real $\beta \in (0,1)$, definimos a distribuição a* priori *para uma árvore $\tau \in \mathcal{T}(D)$ por*

$$\pi(\tau) := \pi_D(\tau, \beta) = \alpha^{|\tau|-1}\beta^{|\tau|-L_D(\tau)}, \tag{4.2}$$

em que $\alpha := (1-\beta)^{1/|\mathcal{A}|-1}$ e $L_D(\tau)$ denota o número de folhas da árvore τ com tamanho D.

Note que a equação (4.2) define uma função de penalização que utiliza o termo $\alpha^{|\tau|-1}$ para penalizar modelos complexos de forma exponencial. Isso significa que, à medida que o tamanho do modelo $|\tau|$ aumenta, o termo $\alpha^{|\tau|-1}$ cresce exponencialmente, resultando em uma penalização mais intensa para modelos mais complexos. Além disso, o termo $\beta^{|\tau|-L_D(\tau)}$ adiciona uma penalidade às folhas com tamanho inferior a D.

É possível mostrar que (4.2) define uma distribuição de probabilidade. Em geral, é comum adotar $\beta \approx 1 - 2^{-|\mathcal{A}|+1}$ e, consequentemente, obter $\alpha \approx 1/2$, a menos que haja uma razão específica para optar por valores diferentes (ver Kontoyiannis *et al.* (2022)).

Dada uma árvore de contexto probabilística (τ, Q) que seja admissível para a amostra x_1^n, podemos definir uma distribuição *a priori* Dirichlet[1] com parâmetros $(1/2, 1/2, \ldots, 1/2)$, correspondente à *priori* de Jeffreys, para cada $Q(\cdot \mid s)$, $s \in \tau$. Assim,

$$\pi(Q \mid \tau) = \prod_{s \in \tau} \pi\left(Q(\cdot \mid s)\right),$$

em que

$$\pi\left(Q(\cdot \mid s)\right) := \frac{\Gamma(|\mathcal{A}|/2)}{\pi^{|\mathcal{A}|/2}} \prod_{j=1}^{|\mathcal{A}|} Q(j \mid s)^{-1/2} \propto \prod_{j=1}^{|\mathcal{A}|} Q\left(j \mid s\right)^{-1/2}.$$

Kontoyiannis *et al.* (2022) demonstra que os resultados podem ser generalizados de forma direta para qualquer seleção arbitrária de vetor de parâmetros da distribuição a *priori* de Dirichlet.

A verossimilhança marginal $P(x_1^n \mid \tau)$ dada uma árvore de contexto probabilística

[1] A distribuição Dirichlet é uma distribuição de probabilidade contínua multivariada definida em um espaço K-dimensional, onde K é um número inteiro positivo. Essa distribuição é frequentemente utilizada para modelar variáveis aleatórias que representam proporções ou frações em um conjunto de categorias.

(τ, Q) admissível para a amostra x_1^n é em que

$$\begin{aligned} P_e(N_n(s,\cdot)) &= P_e\left(N_n(s,1), N_n(s,2), \ldots, N_n(s,|A|)\right) \\ &= \frac{\prod_{j=1}^{|\mathcal{A}|}\left[(1/2)(3/2)\cdots(N(s,j)-1/2)\right]}{(|\mathcal{A}|/2)(|\mathcal{A}|/2+1)\cdots(|\mathcal{A}|/2+N_n(s)-1)}, \end{aligned}$$

com a convenção que qualquer produto vazio é tomado como sendo igual a 1.

Dessa forma, o foco principal é a distribuição a *posteriori* do modelo, representada por

$$\pi(\tau \mid x_1^n) = \frac{P(x_1^n \mid \tau)\pi(\tau)}{P(x_1^n)},$$

em que $P(x_1^n)$ representa a verossimilhança preditiva a *priori* de x_1^n, obtida somando-se as probabilidades conjuntas de x_1^n para todos os valores possíveis de τ, isto é,

$$P(x_1^n) = \sum_{\tau \in \mathcal{T}(D)} \pi(\tau) P(x_1^n \mid \tau) = \sum_{\tau \in \mathcal{T}(D)} \pi(\tau) \int P(x_1^n \mid Q, \tau)\pi(Q \mid \tau) dQ.$$

Observe que a dificuldade surge devido ao tamanho considerável da classe $\mathcal{T}(D)$ dos modelos com memória variável. O tamanho de $\mathcal{T}(D)$ cresce exponencialmente em relação a D de forma extremamente rápida, seguindo a seguinte relação

$$|\mathcal{T}(D)| \geq \sum_{d=0}^{D-1} 2^{|\mathcal{A}|^d} - D.$$

4.2 Ponderação da árvore de contexto

O algoritmo de ponderação da árvore de contexto (CTW, do inglês *Context Tree Weighting*) consegue computar $P(x_1^n)$ de forma exata e extremamente eficiente (Kontoyiannis *et al.*, 2022). Esse algoritmo é utilizado para modelar e prever sequências de dados, como séries temporais. Ele recebe várias entradas, incluindo o tamanho do alfabeto $\mathcal{A}$, a profundidade máxima do contexto D, a série temporal de observações x_1^n e x_{-D+1}^0 e o valor do parâmetro a *priori* $\beta \in (0,1)$ e é dividido em cinco passos:

1. Construção da árvore: O algoritmo começa construindo uma árvore máxima τ_{max}. As folhas dessa árvore representam todos os contextos possíveis de comprimento D, que aparecem em x_{-D+1}^n. Se algum nó da árvore estiver em uma profundidade menor que D e apenas alguns de seus filhos estiverem presentes na árvore, o algoritmo

adiciona os filhos restantes para garantir que τ_{max} seja uma árvore completa.

2. Cálculo do vetor de contagem: Em cada nó da árvore τ_{max}, incluvise nas folhas adicionais incluídas no passo anterior, o algoritmo calcula o vetor de contagem $N_n(s, \cdot)$, representando a frequência de ocorrência de cada símbolo do alfabeto até aquele ponto da árvore.

3. Cálculo da probabilidade estimada: Em cada nó da árvore τ_{max}, o algoritmo calcula a probabilidade estimada $P_e(N_n(s, \cdot))$, probabilidade condicional de observar o vetor de contagem $N_n(s, \cdot)$, dada a estrutura da árvore. Quando o vetor $N_n(s, \cdot)$ é identificamente nulo, definimos $P_e(N_n(s, \cdot)) = 1$.

4. Cálculo das probabilidades ponderadas: O algoritmo realiza uma etapa recursiva, percorrendo a árvore τ_{max} da raiz às folhas. Em cada nó, é calculada uma probabilidade ponderada $P_{w,s}$, uma combinação ponderada entre a probabilidade estimada $P_{e,s} := P_e(N_n(s, \cdot))$ e as probabilidades ponderadas dos nós filhos. Se o nó for uma folha, a probabilidade ponderada é simplesmente a probabilidade estimada $P_{e,s}$. Caso contrário, a probabilidade ponderada é calculada ponderando a probabilidade estimada $P_{e,s}$ com um fator β e somando as probabilidades ponderadas dos nós filhos,

$$P_{w,s} = \begin{cases} P_{e,s}, & \text{se } s \text{ é uma folha,} \\ \beta P_{e,s} + (1-\beta) \prod_{j=1}^{\mathcal{A}} P_{w,js}, & \text{caso contrário.} \end{cases}$$

5. Saída da probabilidade ponderada: A probabilidade ponderada final $P_{w,\lambda}$ é a probabilidade calculada na raiz λ da árvore τ_{max}, representa a probabilidade de previsão prévia das observações, considerando a estrutura da árvore de contexto e o parâmetro a *priori* β.

Note que o algoritmo CTW retorna a estimativa da verossimilhança a *priori* das observações, incorporando a estrutura da árvore de contexto e a distribuição a *priori* dos parâmetros. Logo, $P_{w,\lambda}$ representa a probabilidade conjunta das observações futuras, dada a estrutura da árvore, os parâmetros e as observações passadas.

Uma vantagem das árvores de contexto Bayesianas é que elas permitem encontrar, a partir do algoritmo descrito anteriormente, uma fórmula fechada para $P(x_1^n) := P\left(X_1^n = x_1^n\right)$. Formalizamos esse fato no teorema a seguir, cuja demonstração pode ser vista em Kontoyiannis *et al.* (2022).

Teorema 4.3 *A probabilidade ponderada $P_{w,\lambda}$ calculada na raíz calculada pelo CTW é exatamente a verossimilhança preditiva a* priori *das observações*

$$P_{w,\lambda} = P(X_1^n = x_1^n) = \sum_{\tau \in \mathcal{T}(D)} \pi(\tau) \int_Q P\left(x_1^n \mid x_{-D+1}^0, Q, \tau\right) \pi(Q \mid \tau) dQ.$$

4.3 Algoritmo da árvore de contexto Bayesiana

Uma vez que conseguimos obter, a partir do algoritmo descrito a seguir, $P(x_1^n) := P\left(X_1^n = x_1^n\right)$, podemos calcular a distribuição *a posteriori* $\pi\left(\tau \mid x_1^n\right)$, dada por

$$\pi\left(\tau \mid x_1^n\right) = \frac{P\left(x_1^n \mid \tau\right) \pi(\tau)}{P(x_1^n)}.$$

A partir da maior árvore admissível τ_0, começamos com as folhas e procedemos recursivamente até a raiz, em cada nó s da árvore τ_0 calculamos as seguintes probabilidades

$$P_{|\mathcal{A}|,s} := \begin{cases} P_{e,s} & \text{se } s \text{ é uma folha e } |s| = D, \\ \beta, & \text{se } s \text{ é uma folha e } |s| < D, \\ \max\left\{\beta P_{e,s},\ (1-\beta)\prod_{j=1}^{|A|} P_{|\mathcal{A}|,js}\right\}, & \text{se } s \text{ é um nó interno.} \end{cases} \tag{4.4}$$

Note que $P_{e,s}$ é obtido por meio do algoritmo CTW. Em seguida, começando na raiz e procedendo recursivamente com seus descendentes, para cada nós s: se o máximo em (4.4), pode ser obtido através do primeiro termo, então podamos todos os seus descendentes na árvore τ_0; caso contrário, repetimos o mesmo processo para cada um dos $|\mathcal{A}|$ filhos do nó s. Após, todos os nós terem sido verificados, obtemos a árvore estimada $\hat{\tau}_{BCT}$ com suas respectivas probabilidades estimadas. Kontoyiannis *et al.* (2022) mostra que a árvore estimada $\hat{\tau}_{BCT}$ é a árvore admissível com a maior probabilidade a posteriori. A seguir enunciamos esse resultado e sugerimos a leitura do artigo em questão para detalhes sobre sua demonstração.

Teorema 4.5 *Para todos os valores de $\beta > 1/2$, a árvore $\hat{\tau}_{BCT}$ produzida pelo algoritmo BCT é o modelo de árvore MAP, isto é, $\hat{\tau}_{BCT}$ é a árvore que maximiza a probabilidade a* posteriori, *(ou um dos modelos de árvore MAP, caso o máximo não seja alcançado de forma única).*

$$\pi(\hat{\tau}_{BCT} \mid x_1^n) = \max_{\tau \in \mathcal{T}(D)} \pi(\tau \mid x_1^n).$$

É importante notar que o algoritmo descrito não garante a unicidade de solução. Portanto, não podemos afirmar que a árvore gerada seja a única que fornece a maior probabilidade a *posteriori*. Em seu trabalho, Kontoyiannis *et al.* (2022) apresenta o k-BCT, uma generalização natural do BCT, utilizado para identificar as k árvores mais prováveis a *posteriori*, sendo $k > 1$. Dessa forma, podemos avaliar um conjunto de árvores que são prováveis de terem gerado os dados.

Dada a verossimilhança preditiva a *priori* $P(x_1^n)$ calculada pelo algoritmo CTW e os modelos mais prováveis a *posteriori* identificados pelos algoritmos BCT e k-BCT, a probabilidade a *posteriori* do modelo pode ser facilmente obtida por meio dos resultados desses algoritmos. De fato, se $D \geq 0$ é uma profundidade máxima fixada, $\beta \in [1/2, 1)$ e x_1^n uma série temporal com contexto inicial x_{-D+1}^0, então para qualquer árvore $\tau \in \mathcal{T}(D)$,

$$\pi(\tau \mid x_1^n) = \frac{P(x_1^n \mid \tau)\pi(\tau)}{P(x_1^n)} = \frac{\pi(\tau)\prod_{s\in\tau} P_{e,s}}{P_{w,\lambda}}. \tag{4.6}$$

Capítulo 5

Aplicações

Neste capítulo, testamos e avaliamos o desempenho dos estimadores de árvores de contexto por meio de simulações computacionais. Os estimadores considerados são o algoritmo contexto de Rissanen (1983a), a versão modificada do algoritmo contexto proposta por Galves e Leonardi (2008) e as árvores de contexto Bayesiana proposta por Kontoyiannis *et al.* (2022). No restante do capítulo, apresentamos uma breve descrição de como essas três técnicas serão utilizadas, e então apresentamos os resultados dos experimentos com dados simulados e com dados eletrofisiológicos e genéticos.

5.1 Simulação

Com o objetivo de compreender as propriedades, capacidades e limitações dos métodos de inferência frequentista e Bayesiana em diversos cenários, conduzimos simulações para comparar as árvores estimadas com as árvores geradoras das amostras. A seguir, detalhamos os cenários em estudo, assim como as hipóteses e considerações das simulações.

5.1.1 Hipóteses e Cenários

Pretendemos aplicar a metodologia estudada neste monografia em dados eletrofisológicos e genéticos. Portanto optamos por considerar, no estudo de simulação, árvores com alfabetos de tamanho 2 e 4, escolhidas para se adequarem às características particulares de cada cenário. No âmbito eletrofisilógico, os processos são caracterizados por um alfabeto $|\mathcal{A}| = 2$, que simboliza a presença (1) ou ausência (0) de disparos neuronais, enquanto na génetica, o alfabeto adotado tem $|\mathcal{A}| = 4$, representando as bases nitrogenadas: Adenina (0), Citosina (1), Guanina (2) e Timina (3).

Além disso, a escolha das matrizes de transição foram realizadas com base nas particularidades dos modelos neuronais e genéticos, buscando uma abordagem alinhada com as características de cada cenário. No contexto dos disparos neuronais, uma suposição é que um neurônio dispara quando seu potencial de membrana ultrapassa um certo limiar, em seguida, ele retorna ao seu potencial de repouso e se mantém nesse estado até que outro desequilíbrio eletrolítico ocorra. Dessa forma, após seu último disparo, o neurônio acumula potencial de membrana até que este seja suficientemente grande para permitir um disparo. Assim, é razoável considerar que a probabilidade de um neurônio disparar condicionada ao conhecimento da sua resposta temporal passada é tão maior quanto mais longe estiver o último disparo desse neurônio no passado em questão. Em relação aos processos genéticos, nossa intuição baseia-se na presença de quatro bases nitrogenadas que se conectam em triplas. No RNA mensageiro, cada tripla é chamada de códon, e cada códon especifica um aminoácido. Assim, considerando a relação entre triplas, é razoável supor que o processo subjacente é de ordem 2.

5.1.2 Estudo de simulação

Consideramos quatro árvores de contextos probabilísticas com estruturas diferentes:

- **Árvore 1:** árvore de contexto completa com alfabeto binário compatível com uma cadeia de Markov de ordem 2;
- **Árvore 2:** árvore de contexto completa com alfabeto binário compatível com uma cadeia de Markov de ordem 5;
- **Árvore 3:** árvore de contexto binária compatível com uma cadeia de Markov de alcance variável com ordem no máximo 6;
- **Árvore 4:** árvore de contexto completa com $|\mathcal{A}| = 4$ compatível com uma cadeia de Markov de ordem 2;
- **Árvore 5:** árvore de contexto com $|\mathcal{A}| = 4$ compatível com uma cadeia de Markov de alcance variável com ordem no máximo 3.

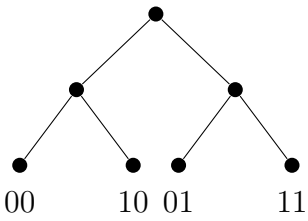
Para avaliar o desempenho dos métodos de estimação em cada um desses cenários, simulamos 100 amostras de diferentes tamanhos e calculamos a proporção de acertos dos

métodos. Os tamanhos de amostra considerados em cada cenário são: $n = 5000$, $n = 8000$, $n = 11000$ e $n = 14000$.

Adicionalmente, fixamos alguns parâmetros em cada método de estimação. No algoritmo de contexto de Rissanen, definimos $d = \left\lfloor \frac{\ln_{|\mathcal{A}|} n}{2} \right\rfloor$ e $\delta = \ln(n)$ conforme proposto por Galves *et al.* (2012). Para o algoritmo contexto modificado, estimamos as árvores considerando vários valores para δ, de modo que $\delta \in [0, 1]$ e escolhendo o valor que proporciona a maior proporção de acertos e também definimos $d = \left\lfloor \frac{\ln_{|\mathcal{A}|} n}{2} \right\rfloor$. No método Bayesiano, adotamos a sugestão de Kontoyiannis *et al.* (2022) e consideramos $\beta \approx 1 - 2^{|\mathcal{A}|+1}$.

Árvore 1

A Figura 5.1 apresenta a representação gráfica da árvore de contexto probabilística completa compatível com uma cadeia de Markov de ordem 2 cuja matriz de transição Q é considerada no processo de simulação.



$$Q = \begin{array}{c} \\ 00 \\ 01 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0,20 & 0,80 \\ 0,90 & 0,10 \\ 0,55 & 0,45 \\ 0,75 & 0,25 \end{bmatrix} \end{array}.$$

Figura 5.1: Representação gráfica da árvore de contexto $\tau = \{00, 01, 10, 11\}$ e matriz de transição Q.

Nesse cenário, fixamos os seguintes parâmetros: $d = 6$ nos três algoritmos; $\delta = 10$ no algoritmo contexto de Rissanen; $\delta = 0,1$ no algoritmo contexto de Galves e Leonardi (2008).

A Tabela 5.1 apresenta as proporções de acertos nos três algoritmos em relação a diferentes tamanhos de amostra, considerando a árvore de contexto probabilística (τ, Q) fixada.

Nota-se que o algoritmo contexto de Rissanen demonstra uma tendência de melhoria à medida que o tamanho da amostra aumenta, começando com uma proporção de acertos igual a $0,88$ para amostras de tamanho 5.000 e alcançando 1 para amostras maiores. Em contrapartida, o algoritmo contexto de Galves se mostrou mais sensível, e não acertou a árvore geradora das amostras em nenhuma das situações desse cenário. É relevante destacar que mesmo considerando um *grid* de valores para $\delta \in [0, 1]$, nenhum acerto foi

Tabela 5.1: Proporção de acertos dos algoritmos em estudo considerando uma árvore Markoviana completa de ordem 2 e parâmetros fixados.

n	Alg. Cont. Rissanen	Alg. Cont. Galves	Arv. Cont. Bayesiana
5.000	0,88	0	0,95
8.000	0,99	0	0,98
11.000	1	0	1
14.000	1	0	1

registrado nesse cenário. Por outro lado, a árvore de contexto Bayesiana mantém uma performance mais parecida em todos os tamanhos de amostra. Esses resultados sugerem que o desempenho do algoritmo contexto de Rissanen e da árvore Bayesiana são similares nesse cenário enquanto, como veremos a seguir, a modificação proposta por Galves e Leonardi (2008) apresenta dificuldade em recuperar a estrutura da árvore fixada para tamanho de amostras que não são suficientemente grandes.

Adicionalmente, é relevante mencionar que, para contextualização, apresentamos na Tabela 5.2 as árvores estimadas pelo algoritmo contexto modificado por Galves e Leonardi (2008) para os tamanhos de amostra 5000, 8000, 11000 e 14000 considerando a 69º amostra gerada, escolhido de forma aleatória. No entanto, vale ressaltar que essas estimativas não foram capazes de capturar a verdadeira estrutura da amostra.

Tabela 5.2: Árvores estimadas pelo algoritmo contexto modificado por Galves e Leonardi (2008), considerando a 69º amostra gerada para os diferentes tamanhos amostrais no cenário de árvore Markoviana completa de ordem 2.

n	$\hat{\tau}$	$h(\tau)$	$\lvert\hat{\tau}\rvert$
5.000	000000, 110000, 110100, 111010, 001110, 011110, 111110,110001, 101011, 111011, 110111, 001111, 110110, 001011, 101111	6	15
8.000	1111, 000000, 111000, 111100 111010, 000110, 000001, 111011	6	8
11.000	00000, 11110, 11011, 110000, 000011, 000111, 110111, 111111	6	8
14.000	000000, 110000, 000110, 111110, 000011, 101111, 011111, 111111	6	8

A partir da Tabela 5.2, é perceptível que, mesmo o algoritmo não tendo identificado a verdadeira relação entre a amostra e a árvore geradora nos cenários abordados, há uma

tendência de redução na estimativa da árvore de contexto à medida que o tamanho da amostra aumenta. Essa tendência sugere a possibilidade de convergência para a árvore correta à medida que o tamanho da amostra tende ao infinito.

Árvore 2

A Figura 5.2 apresenta a representação gráfica da árvore de contexto probabilística completa compatível com uma cadeia de Markov de ordem 5 cuja matriz de transição é Q. Nesse caso, fixamos as probabilidades de transição todas próximas de 1/2.

Nesse cenário, fixamos os seguintes parâmetros: $d = 6$ nos três algoritmos; $\delta = 10$ no algoritmo contexto de Rissanen; $\delta = 0,1$ no algoritmo contexto de Galves.

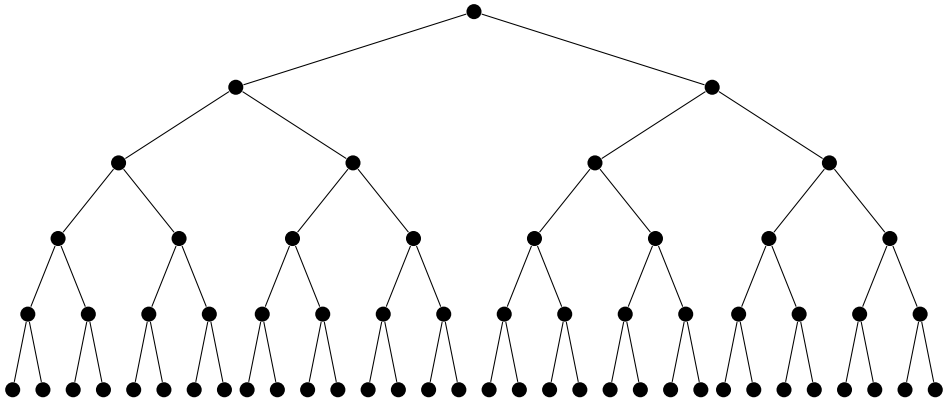
Note que ao considerar a matriz de transição Q apresentada na Figura 5.2, não estamos diferenciando os contextos, isto é, a probabilidade de transição para os estados "0" ou "1" independe dos estados visitados pelo processo nos últimos cinco instantes do passado. Dessa forma, é esperado que nenhum dos métodos em estudo seja capaz de capturar a verdadeira relação entre a amostra e a árvore geradora previamente fixada. De fato, a conclusão dos algoritmos é que o processo é independente, o que é esperado dada a estrutura da matriz de transição Q fixada.

A Tabela 5.3 apresenta as proporções de acertos nos três algoritmos em estudo em relação aos quatro tamanhos de amostra.

Tabela 5.3: Proporção de acertos dos algoritmos em estudo considerando uma arvore Markoviana completa de ordem 5 com probabilidades de transição próximas de 1/2.

n	Alg. Cont. Rissanen	Alg. Cont. Galves	Arv. Cont. Bayesiana
5.000	0	0	0
8.000	0	0	0
11.000	0	0	0
14.000	0	0	0

Observe que as probabilidades de transição todas próximas de 1/2 dificultam a correta interpretação dos padrões subjacentes, comprometendo a eficácia de todos os métodos utilizados. No entanto, os três métodos apresentaram estimações razoáveis, identificando para a maioria das amostras o processo subjacente como sendo uma sequência de variáveis aleatórias independentes.



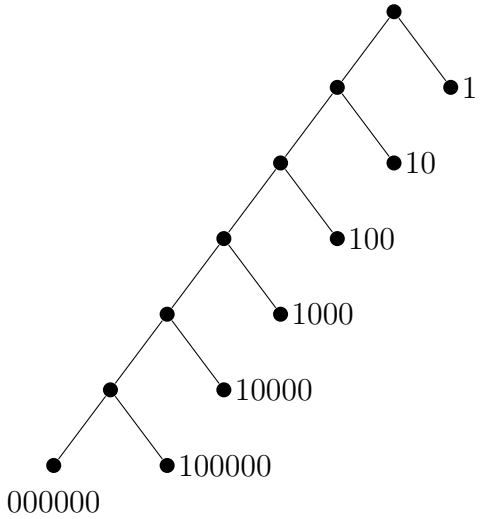
$$
Q = \begin{array}{c} \\ 00000 \\ 00001 \\ 00010 \\ 00011 \\ 00100 \\ 00101 \\ 00110 \\ 00111 \\ 01000 \\ 01001 \\ 01010 \\ 01011 \\ 01100 \\ 01101 \\ 01110 \\ 01111 \\ 10000 \\ 10001 \\ 10010 \\ 10011 \\ 10100 \\ 10101 \\ 10110 \\ 10111 \\ 11000 \\ 11001 \\ 11010 \\ 11011 \\ 11100 \\ 11101 \\ 11110 \\ 11111 \end{array}
\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \left[\begin{array}{cc}
0.4948819 & 0.5051181 \\
0.4556870 & 0.5443130 \\
0.4445373 & 0.5554627 \\
0.4284962 & 0.5715038 \\
0.4963631 & 0.5036369 \\
0.4228119 & 0.5771881 \\
0.4305150 & 0.5694850 \\
0.4108216 & 0.5891784 \\
0.4946520 & 0.5053480 \\
0.4754072 & 0.5245928 \\
0.4343737 & 0.5656263 \\
0.4429470 & 0.5570530 \\
0.4335178 & 0.5664822 \\
0.4389733 & 0.5610267 \\
0.4795045 & 0.5204955 \\
0.4269441 & 0.5730559 \\
0.4401309 & 0.5598691 \\
0.4708572 & 0.5291428 \\
0.4407185 & 0.5592815 \\
0.4693148 & 0.5306852 \\
0.4848891 & 0.5151109 \\
0.4252405 & 0.5747595 \\
0.4436026 & 0.5563974 \\
0.4347334 & 0.5652666 \\
0.4676612 & 0.5323388 \\
0.4779795 & 0.5220205 \\
0.4895633 & 0.5104367 \\
0.4526629 & 0.5473371 \\
0.4551908 & 0.5448092 \\
0.4271894 & 0.5728106 \\
0.4699663 & 0.5300337 \\
0.4618842 & 0.5381158
\end{array}\right] \end{array},
$$

Figura 5.2: Representação gráfica da árvore de contexto Markoviana de ordem 5 e matriz de transição Q.

Árvore 3

A Figura 5.3 apresenta a representação gráfica de uma árvore de contexto probabilística compatível com uma cadeia de Markov de alcance variável de ordem no máximo 6 cuja matriz de transição é Q.

Fixamos os parâmetros: $d = 6$ nos três algoritmos; $\delta = 10$ no algoritmo contexto de Rissanen; $\delta = 0,1$ no algoritmo contexto de Galves.



$$Q = \begin{array}{c} \\ 1 \\ 10 \\ 100 \\ 1000 \\ 10000 \\ 000000 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0,99 & 0,01 \\ 0,80 & 0,20 \\ 0,70 & 0,30 \\ 0,55 & 0,45 \\ 0,30 & 0,70 \\ 0,10 & 0,90 \end{bmatrix} \end{array},$$

Figura 5.3: Representação gráfica da árvore de contexto τ = {1,10, 100, 1000, 10000, 10000, 100000, 000000} e matriz de transição Q.

A Tabela 5.4 apresenta as proporções de acertos nos três algoritmos em estudo em relação aos quatro tamanhos de amostra.

Tabela 5.4: Proporção de acertos dos algoritmos em estudo considerando uma árvore de alcance variável e alfabeto binário.

n	Alg. Cont. Rissanen	Alg. Cont. Galves	Arv. Cont. Bayesiana
5.000	1	0	0,67
8.000	1	0	0,79
11.000	1	0	0,86
14.000	1	0	0,89

Observe que o algoritmo contexto de Rissanen demostrou um ótimo desempenho, independente do tamanho da amostra, sendo o melhor algoritmo em estudo para o cenário abordado. Enquanto o algoritmo contexto de Galves continuou não reconhecendo a verdadeira relação entre a amostra e a árvore geradora. Além disso, também notamos que a árvore de contexto Bayesiana apresentou uma tendência de melhoria à medida que o tamanho da amostra aumentava, começando com uma proporção de acertos igual a $0,67$

para amostras de tamanho 5.000 e alcançando $0,89$ para amostras maiores. Apresentamos na Tabela 5.5 as árvores estimadas pelo algoritmo contexto modificado por Galves e Leonardi (2008) para os diferentes tamanhos de amostra.

Tabela 5.5: Árvores estimadas pelo algoritmo contexto modificado por Galves e Leonardi (2008), considerando a 69^{o} amostra gerada para os diferentes tamanhos amostrais no cenário com árvore de alcance variável e alfabeto binário.

n	$\hat{\tau}$	$h(\tau)$	$\lvert\hat{\tau}\rvert$
5.000	111, 1110, 11100, 11101, 101011 11011, 000000, 100000, 111000, 110100, 110010, 111010, 110110, 111001, 110101, 001101, 110011	6	17
8.000	111, 1110, 11100, 11101, 11011, 000000, 100000, 111000, 101100, 111010, 100110, 110110, 110001, 111001, 110101, 101101, 110011	6	17
11.000	111, 1110, 11100, 11101, 11011, 000000, 100000, 111000, 101100, 111010, 110110, 111001, 110011	6	13
14.000	111, 1110, 11100, 11101, 11011, 000000, 100000, 111000, 110100, 110010, 111010, 110110, 111001, 110101, 110011	6	15

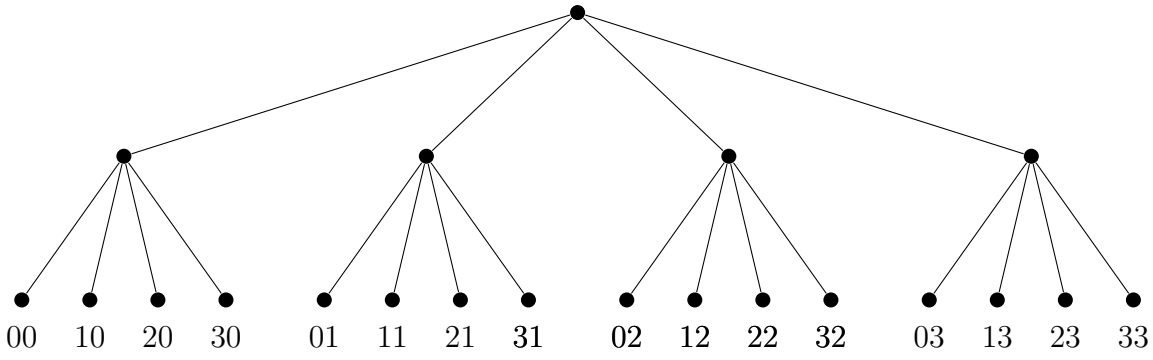
A partir da Tabela 5.5, podemos notar o mesmo comportamento observado no cenário com árvore Markoviana completa de ordem 2 apresentado na Tabela 5.2, isto é, há uma tendência de poda na estimativa da árvore à medida que o tamanho da amostra aumenta.

Árvore 4

Nesta seção, consideramos uma árvore de contexto probabilística completa compatível com uma cadeia de Markov de ordem 2, cujas probabilidades de transição seguem a proposta de Mendes (2018). A representação gráfica desta árvore e sua matriz de transição é fornecida na Figura 5.4.

Nesse cenário, fixamos os seguintes parâmetros: $d = 3$ nos três algoritmos; $\delta = 10$ no algoritmo contexto de Rissanen; $\delta = 0,1$ no algoritmo contexto de Galves e Leonardi (2008).

A Tabela 5.6 apresenta as proporções de acertos nos três algoritmos em estudo em relação aos quatro tamanhos de amostra.



$$Q = \begin{array}{c} \\ 00 \\ 10 \\ 20 \\ 30 \\ 01 \\ 11 \\ 21 \\ 31 \\ 02 \\ 12 \\ 22 \\ 32 \\ 03 \\ 13 \\ 23 \\ 33 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0,30 & 0,20 & 0,40 & 0,10 \\ 0,40 & 0,20 & 0,25 & 0,15 \\ 0,35 & 0,30 & 0,30 & 0,05 \\ 0,05 & 0,15 & 0,50 & 0,30 \\ 0,15 & 0,35 & 0,25 & 0,25 \\ 0,30 & 0,30 & 0,35 & 0,05 \\ 0,05 & 0,60 & 0,20 & 0,15 \\ 0,05 & 0,25 & 0,45 & 0,25 \\ 0,20 & 0,25 & 0,30 & 0,25 \\ 0 & 0,30 & 0,35 & 0,35 \\ 0,20 & 0,20 & 0,45 & 0,15 \\ 0 & 0,25 & 0,30 & 0,45 \\ 0 & 0,20 & 0,25 & 0,55 \\ 0,20 & 0,40 & 0,30 & 0,10 \\ 0,25 & 0,40 & 0,10 & 0.25 \\ 0,05 & 0,35 & 0,15 & 0,45 \end{bmatrix} \end{array}.$$

Figura 5.4: Representação gráfica da árvore de contexto Markoviana completa de ordem 2 com alfabeto $\mathcal{A} = \{0, 1, 2, 3\}$ e matriz de transição Q.

Tabela 5.6: Proporção de acertos dos algoritmos em estudo considerando uma árvore Markoviana completa de ordem 2 e alfabeto $\mathcal{A} = \{0, 1, 2, 3\}$.

n	Alg. Cont. Rissanen	Alg. Cont. Galves	Arv. Cont. Bayesiana
5.000	0,80	0	0,99
8.000	0,78	0	0,98
11.000	0,83	0	1
14.000	0,76	0	0,99

Observe que o tanto o algoritmo contexto de Rissanen quanto a árvore de contexto Bayesiana apresentaram um ótimo desempenho. Enquanto o algoritmo contexto de Rissanen se manteve estável, com taxa de acerto de aproximadamente 80% em todos os casos, a árvore de contexto Bayesiana começa em 0,99 para amostras de tamanho 5.000 e alcança total acerto em amostras de tamanho 11.000. Em contrapartida, o algoritmo contexto

modificado por Galves e Leonardi (2008) apresentou o pior desempenho, se comparada aos demais métodos em estudo, não conseguindo identificar a verdadeira relação entre a amostra e a árvore geradora. Apresentamos, na Tabela 5.7 as árvores estimadas pelo algoritmo contexto modificado por Galves e Leonardi (2008) para os tamanhos de amostra 5000, 8000, 11000 e 14000 considerando a 69º amostra gerada.

Tabela 5.7: Árvores estimadas pelo algoritmo contexto modificado por Galves e Leonardi (2008), considerando a 69º amostra gerada para os diferentes tamanhos amostrais no cenário com árvore Markoviana completa de ordem 2 e alfabeto $\mathcal{A} = \{0, 1, 2, 3\}$.

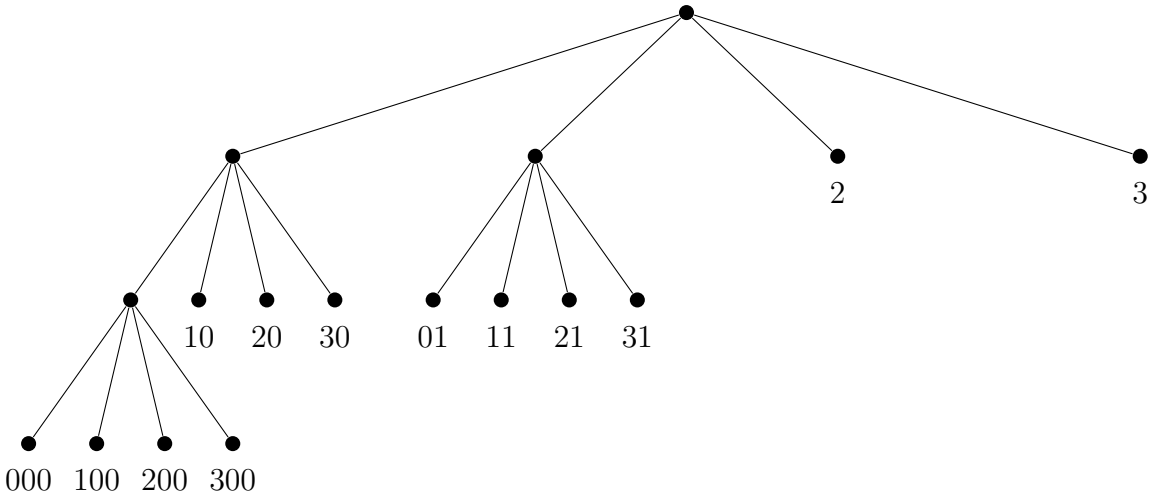
n	$\hat{\tau}$	$h(\tau)$	$\lvert\hat{\tau}\rvert$
5.000	11, 21, 12, 22, 13, 300, 101, 031, 010, 210, 120, 320, 030, 332, 203, 133	3	16
8.000	21, 31, 12, 22, 32, 33, 011, 003, 203, 310, 120, 320, 030, 330, 303, 013	3	16
11.000	11, 21, 31, 12, 22, 32, 03, 23, 33, 300, 310, 120, 320, 030, 301, 013	3	16
14.000	01, 11, 21, 31, 12, 32, 13, 33, 010, 310, 120, 320, 030, 322, 203	3	15

Analisando a Tabela 5.7, percebe-se que, mesmo nos cenários abordados nos quais o algoritmo não conseguiu identificar a verdadeira relação entre a amostra e a árvore geradora, há uma tendência de poda na estimativa da árvore à medida que o tamanho da amostra aumenta. Observe que, para amostras de tamanho 5.000, o algoritmo estimou corretamente 5 contextos, valor que aumenta conforme o tamannho amostral também aumenta, alcançando 8 para amostras de tamanho 14.000. Essa tendência sugere a possibilidade de convergência para a árvore correta à medida que o tamanho da amostra tende ao infinito.

Árvore 5

A Figura 5.5 apresenta a representação gráfica de uma árvore de contexto probabilística compatível com uma cadeia de Markov de alcance variável de ordem no máximo 3 cujas probabilidades da matriz de transição seguem a proposta de Mendes (2018).

Fixamos os parâmetros: $d = 3$ nos três algoritmos; $\delta = 10$ no algoritmo contexto de Rissanen; $\delta = 0,1$ no algoritmo contexto modificado.



$$Q = \begin{array}{c} \\ 2 \\ 3 \\ 31 \\ 21 \\ 11 \\ 01 \\ 30 \\ 20 \\ 10 \\ 300 \\ 200 \\ 100 \\ 000 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0,3 & 0,2 & 0,4 & 0,1 \\ 0,35 & 0,3 & 0,25 & 0,1 \\ 0,05 & 0,25 & 0,45 & 0,25 \\ 0,05 & 0,6 & 0,2 & 0,15 \\ 0,3 & 0,3 & 0,35 & 0,05 \\ 0,15 & 0,35 & 0,25 & 0,25 \\ 0,05 & 0,15 & 0,5 & 0,3 \\ 0,35 & 0,3 & 0,3 & 0,05 \\ 0,4 & 0,2 & 0,25 & 0,15 \\ 0 & 0,25 & 0,3 & 0,45 \\ 0,2 & 0,2 & 0,45 & 0,15 \\ 0 & 0,3 & 0,35 & 0,35 \\ 0,2 & 0,25 & 0,3 & 0,25 \end{bmatrix} \end{array}.$$

Figura 5.5: Representação gráfica da árvore de contexto $\tau = \{$ 2, 3, 21, 11, 01, 30, 20, 10, 300, 200, 100, 000$\}$ e matriz de transição Q.

A Tabela 5.8 apresenta as proporções de acertos nos três algoritmos em estudo em relação aos quatro tamanhos de amostra.

Tabela 5.8: Proporção de acertos dos algoritmos em estudo considerando uma árvore de alcance variável e alfabeto $\mathcal{A} = \{0, 1, 2, 3\}$.

n	Alg. Cont. Rissanen	Alg. Cont. Galves	Arv. Cont. Bayesiana
5.000	0,62	0	1
8.000	0,71	0	0,99
11.000	0,70	0	1
14.000	0,70	0	1

Com base nos dados da Tabela 5.8, observamos que, mais uma vez, os melhores resultados foram alcançados pelo algoritmo contexto de Rissanen e pela árvore de contexto

Bayesiana. Eles conseguiram estimar com precisão mais de 60% e 99% em todos os cenários, respectivamente. Em contrapartida, o algoritmo contexto modificado por Galves e Leonardi (2008) não teve êxito em identificar corretamente a verdadeira relação entre a amostra e a árvore geradora para os tamanhos amostrais considerados neste estudo.

A Tabela 5.9 apresenta as árvores estimadas pelo algoritmo contexto modificado por Galves e Leonardi (2008) para os tamanhos de amostra 5000, 8000, 11000 e 14000 considerando a 69 amostra gerada.

Tabela 5.9: Árvores estimadas pelo algoritmo contexto modificado por Galves e Leonardi (2008), considerando a 69º amostra gerada para os diferentes tamanhos amostrais no cenário com árvore de alcance variável e alfabeto $\mathcal{A} = \{0, 1, 2, 3\}$.

n	$\hat{\tau}$	$h(\tau)$	$\lvert\hat{\tau}\rvert$
5.000	20, 30, 01, 11, 21, 100, 300, 210, 031, 102, 012, 032, 113, 123, 033, 233	3	16
8.000	20, 30, 01, 11, 21, 31, 100, 300, 310, 332, 103, 203, 113, 323, 333	3	15
11.000	11, 21, 000, 100, 300, 210, 310, 320, 330, 231, 323, 133, 233, 333	3	14
14.000	1, 2, 3, 10, 20, 30, 100, 200, 300	3	9

Analisando a Tabela 5.9, percebe-se que, mesmo nos cenários abordados nos quais o algoritmo não conseguiu identificar a verdadeira relação entre a amostra e a árvore geradora, há uma tendência de poda na estimativa da árvore à medida que o tamanho da amostra aumenta. Observe que, ao lidar com amostras de tamanho 5.000, o algoritmo estimou uma árvore com 16 contextos, dos quais 7 foram estimados corretamente. Em contrapartida, ao ampliar para amostras de tamanho 14.000, o algoritmo conseguiu acertar 8 contextos em uma árvore estimada contendo 9 contextos. Essa tendência sugere a possibilidade de convergência para a árvore correta à medida que o tamanho da amostra tende ao infinito.

5.1.3 Discussão

Em todos os cenários analisados, tanto o algoritmo contexto de Rissanen quanto a árvore de contexto Bayesiana exibiram tendências de melhoria à medida que o tamanho

da amostra aumentava. Por outro lado, o algoritmo contexto modificado por Galves e Leonardi (2008) encontrou dificuldade em recuperar a verdadeira árvore de contexto. Todavia, observamos um aumento no número de contextos identificados corretamente à medida que o tamanho amostral aumenta. Essa observação sugere que, de maneira geral, a eficiência desses métodos está diretamente relacionada ao tamanho da amostra utilizada.

Além disso, a definição do melhor algoritmo varia conforme o cenário em estudo. Ao considerarmos os cenários com árvores de contexto binária compatível com cadeias de Markov de ordem 2, observamos comportamentos semelhantes entre os métodos de Rissanen e Bayesiano. Também não foram identificadas diferenças significativas entre os métodos quando analisamos o cenário com árvore de contexto probabilística compatível com uma cadeia de Markov de ordem 5, cujas probabilidades de transição são todas próximas de 1/2, no qual todos os métodos concordaram que o processo subjacente é uma sequência de variáveis aleatórias independentes. No entanto, o último cenário com alfabeto binário, que envolveu uma árvore de contexto probabilística compatível com uma cadeia de Markov de alcance variável com ordem no máximo 6, destacou as principais disparidades. Nesse caso, o algoritmo contexto de Rissanen demonstrou consistência ao acertar a verdadeira relação entre a amostra e a árvore geradora em todas as situações, enquanto a árvore de contexto Bayesiana enfrentou maior dificuldade na identificação. Por fim, no cenário que considera uma árvore de contexto probabilística com $|\mathcal{A}| = 4$ compatível com uma cadeia de Markov de ordem 2, o algoritmo contexto de Rissanen e a árvore de contexto Bayesiana se destacaram ao apresentar as melhores performances. Isso se deve ao fato de que o tamanho da amostra teve uma interferência mínima na proporção de acertos, que se manteve acima de 78% e 98%, respectivamente. Por outro lado, o algoritmo frequentista proposto por Galves e Leonardi (2008) não conseguiu identificar com precisão a verdadeira relação em nenhum dos casos considerando os tamanhos de amostra em estudo.

Essas conclusões revelam a importância de considerar a natureza do cenário em questão ao selecionar o método mais apropriado, evidenciando que não há uma abordagem única que se destaque em todos os cenários investigados. No entanto, preferimos adotar a árvore de contexto Bayesiana, dado que foi o método mais consistente, identificando mais de 60% da verdadeira relação em todos os cenários analisados. Além disso, a árvore de contexto Bayesiana também se destaca por não possuir restrições de altura máxima devido ao tamanho amostral e apresentou o melhor tempo de processamento computacional.

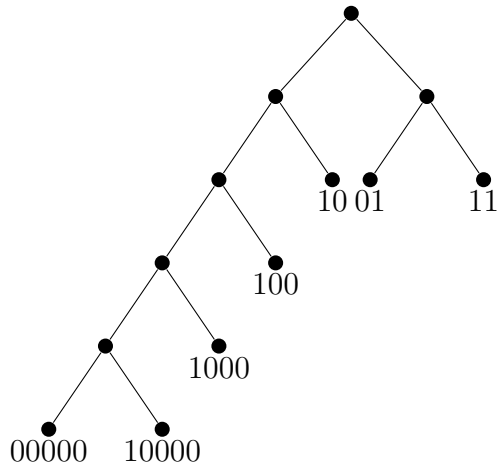
5.2 Aplicação em dados eletrofisiológicos

A implementação dos métodos, levando em consideração os dados eletrofisiológicos, foi conduzida por meio do registro utilizando eletrodos extracelulares em ratos e camundongos durante a realização de atividades livres. Os dados registram o tempo de disparo de 34 neurônios, conforme disponibilizados por Buzsaki Lab Petersen & Hernandez (2018) e o neurônio 9 a ser estudado foi escolhido por meio de um sorteio.

O neurônio em estudo apresenta 5592 disparos em um intervalo de observação de aproximadamente 3 horas (10792,39 segundos). Dado que o tempo médio para ocorrência de uma disparo foi de 1,92 segundos, a série foi discretizada observando, a cada 1,92 segundos, se houve ou não um disparo e caracterizando os eventos como: houve disparos no momento observado (1) e não houve disparo no momento observado.(0)

A seguir apresentamos os resultados obtidos pelo três métodos. Iniciando pela árvore estimada pelo algoritmo contexto de Rissanen, seguida pela modificada por Galves e por fim a árvore de contexto Bayesiana.

A Figura 5.6 apresenta a árvore estimada pelo algoritmo contexto de Rissamen.



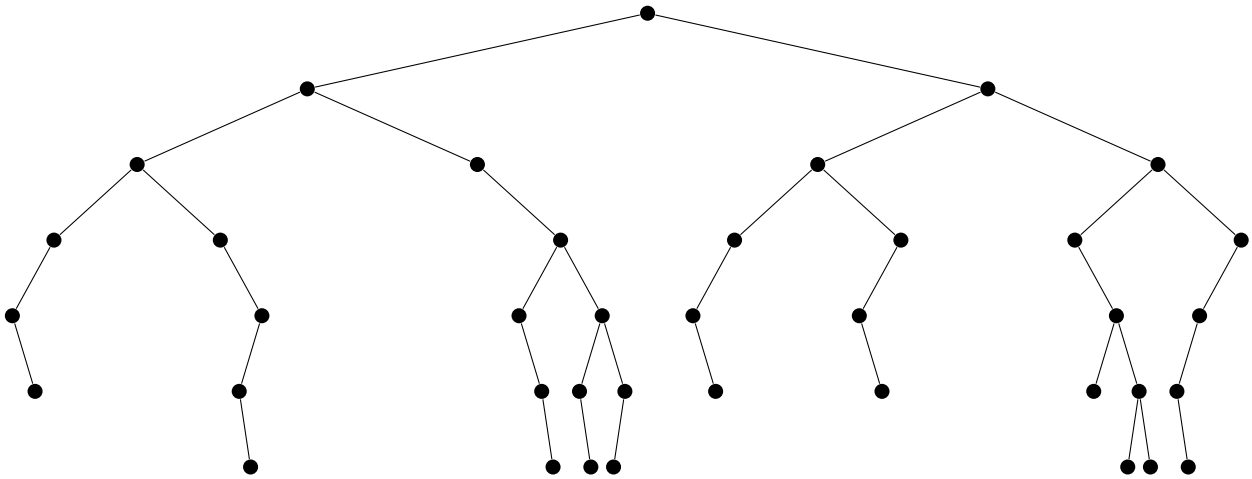
$$\hat{Q} = \begin{array}{c} \\ 01 \\ 11 \\ 10 \\ 100 \\ 1000 \\ 10000 \\ 00000 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0,57 & 0.43 \\ 0,43 & 0,57 \\ 0,64 & 0,36 \\ 0,70 & 0,30 \\ 0,75 & 0,25 \\ 0,85 & 0,15 \\ 0,70 & 0,30 \end{bmatrix} \end{array},$$

Figura 5.6: Representação gráfica da árvore de contexto estimada pelo algortimo de Rissamen $\hat{\tau} = \{01,11,10,100,1000,00000,10000\}$ e matriz de transição estimada $\hat{Q}$.

Observe que a árvore contexto estimada segue o comportamento esperado para um processo estocástico subjacente a dados eletrofisiológicos, i.e., atende a premissa de que um neurônio dispara quando seu potencial de membrana ultrapassa um certo limiar, em seguida, ele retorna ao seu potencial de repouso e se mantém nesse estado até que outro desequilíbrio eletrolítico ocorra. Nesse caso, a observação do primeiro disparo se revela, na maioria da vezes, suficiente para determinar a probabilidade do próximo, uma vez

que a probabilidade de um segundo disparo imediatamente após o primeiro é considerada baixa, alinhando-se à dinâmica natural dos processos neuronais em nossa modelagem.

A Figura 5.7 apresenta a árvore estimada pelo algoritmo contexto modificado por Galves e Leonardi (2008).



$$\hat{Q} = \begin{array}{c} \\ 10000 \\ 10001 \\ 10101 \\ 01011 \\ 101100 \\ 110110 \\ 101110 \\ 011110 \\ 011011 \\ 111011 \\ 100111 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \left[\begin{array}{cc} 0,72 & 0,28 \\ 0,51 & 0,49 \\ 0,39 & 0,61 \\ 0,46 & 0,54 \\ 0,80 & 0,20 \\ 0,79 & 0,21 \\ 0,47 & 0,53 \\ 0,60 & 0,40 \\ 0,48 & 0.52 \\ 0,16 & 0,84 \\ 0,38 & 0,62 \end{array}\right] \end{array},$$

Figura 5.7: Representação gráfica da árvore de contexto estimada pelo algortimo modificado por Galves e Leonardi (2008) $\hat{\tau} = \{10000, 10001, 10101, 01011, 101100, 110110, 101110, 011110, 011011, 111011, 100111\}$ e matriz de transição estimada $\hat{Q}$.

A árvore estimada pelo algoritmo contexto modificado por Galves e Leonardi (2008) apresentou um comportamento inesperado em relação ao processo estocástico subjacente aos dados eletrofisiológicos. Nesse caso, a observação do primeiro disparo não se revela, na maioria das vezes, suficiente para determinar a probabilidade do próximo, uma vez que a chance de um segundo disparo imediatamente após o primeiro não é considerada baixa.

É relevante destacar que esse comportamento não surpreende, dado que durante as

simulações realizadas, não foi observada consistência nos acertos da verdadeira relação entre a amostra e a árvore geradora. Portanto, a discrepância entre o comportamento previsto e o observado confirma as limitações do algoritmo em questão, indicando uma falta de coerência na identificação da dinâmica subjacente aos dados eletrofisiológicos analisados.

A Figura 5.8 apresenta a árvore contexto Bayesiana estimada.

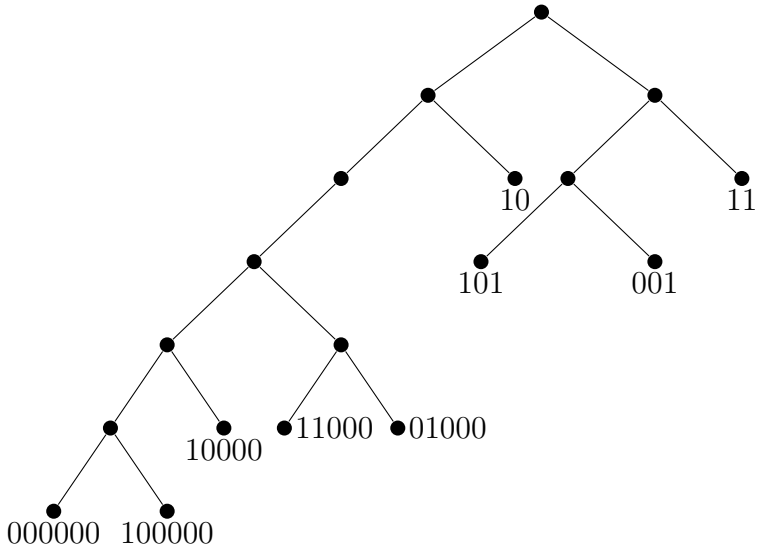


Figura 5.8: Representação gráfica da árvore de contexto Bayesiana estimada $\hat{\tau} = \{11, 10, 001, 101, 11000, 10000, 01000, 000000, 100000\}$.

A análise da árvore de contexto Bayesiana estimada revelou que, em grande parte, o comportamento esperado para um processo estocástico subjacente aos dados eletrofisiológicos de ratos e camundongos durante atividades livres foi seguido. Notadamente, observamos a esperada propensão para um neurônio disparar após o disparo inicial, refletindo a dinâmica natural prevista para esse tipo de processo.

Entretanto, é importante notar que, em alguns pontos, identificamos desvios em relação ao comportamento padrão aguardado para esse processo. Essas variações sugerem a presença de complexidades adicionais ou particularidades na dinâmica neural.

Com base nas simulações anteriormente conduzidas e na propensão do algoritmo contexto de Rissanen em lidar com amostras que possuem essas características em particular, isto é, tamanho de amostra aproximadamente 5.000 e possível comportamento de árvore com alcance variável, acreditamos que este tenha proporcionado a melhor estimação, considerando o cenário abordado. Os resultados dessas simulações podem ser visualizados na Tabela 5.4, que destaca as proporções de acertos dos três algoritmos analisados em diferentes tamanhos de amostra, considerando uma árvore de alcance variável e alfabeto

binário.

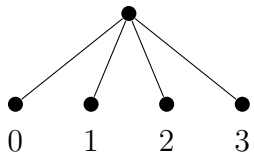
Considerando esses resultados, nossa confiança na eficácia do algoritmo contexto Rissanen, como a melhor opção para a estimação nesse cenário específico, é reforçada.

5.3 Aplicação em dados genéticos

O conjunto de dados derivado da genética refere-se ao *ítron* 7 do gene *α-fetoprotein* dos chimpanzés. Este conjunto foi analisado por Boys *et al.* (2000) e consiste em uma sequência de 1968 bases nitrogenadas: Adenina (A), Citosina (C), Guanina (G) e Timina (T), codificadas como 0, 1, 2 e 3, respectivamente. Este estudo é de particular importância, uma vez que o gene *α-fetoprotein* desempenha um papel crucial no desenvolvimento embrionário dos mamíferos e também parece ter influência no desenvolvimento de tumores.

A seguir apresentamos os resultados obtidos pelos três métodos em estudo. Iniciando pela árvore estimada pelo algoritmo contexto de Rissanen, seguida pela modificado por Galves e Leonardi (2008) e por fim a árvore de contexto Bayesiana.

A Figura 5.9 apresenta a árvore estimada pelo algoritmo contexto de Rissanen.



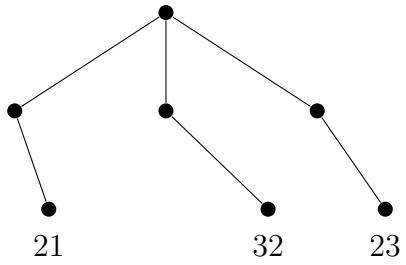
$$\hat{Q} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0,32 & 0,20 & 0,24 & 0,24 \\ 0,21 & 0,19 & 0,20 & 0,40 \\ 0,35 & 0,15 & 0,26 & 0,24 \\ 0,29 & 0,25 & 0,04 & 0,42 \end{bmatrix} \end{array}.$$

Figura 5.9: Representação gráfica da árvore de contexto estimada pelo algortimo contexto de Rissanen $\hat{\tau} = \{0, 1, 2, 3\}$ e matriz de transição estimada $\hat{Q}$.

Observe que a relação encontrada pelo algortimo contexto de Rissanen é uma cadeia de Markov completa de ordem 1, ou seja, é necessário apenas a informação referente a uma base nitrogenada imediatemente anterior para definir o próximo passo da cadeia.

A Figura 5.10 apresenta a árvore estimada pelo algoritmo contexto modificado por Galves e Leonardi (2008).

Note que a árvore de contexto estimada pelo algoritmo modificado por Galves e Leonardi (2008) não possui a propriedade de ser uma árvore completa. Isso significa que apresenta filhos adicionados sem que todos os seus irmãos estejam igualmente incluídos,



$$\hat{Q} = \begin{array}{c} \\ 21 \\ 32 \\ 23 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0,21 & 0,15 & 0,30 & 0,30 \\ 0,39 & 0,22 & 0,01 & 0,38 \\ 0,43 & 0,07 & 0,43 & 0,07 \end{bmatrix} \end{array}.$$

Figura 5.10: Representação gráfica da árvore de contexto estimada pelo algoritmo contexto modificado por Galves e Leonardi (2008) $\hat{\tau} = \{21, 32, 23\}$ e matriz de transição estimada $\hat{Q}$.

gerando nós internos que não possuem todos os seus filhos.

Ao reconhecer essa falta de completude na árvore de contexto gerada pelo algoritmo de Galves, percebe-se que essa árvore só seria completa caso os ramos ausentes fossem adicionados, o que resultaria em uma árvore Markoviana completa de ordem 2.

A Figura 5.11 apresenta a estimação da árvore de contexto Bayesiana.

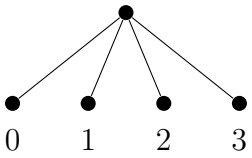


Figura 5.11: Representação gráfica da estimação da árvore de contexto Bayesiana.

Observa-se que a relação identificada pela estimação da árvore de contexto Bayesiana segue o padrão obtido pelo algoritmo contexto de Rissanen. Em outras palavras, essa relação é representada por uma cadeia de Markov de ordem 1, onde a informação referente a uma base nitrogenada imediatamente anterior é suficiente para determinar o próximo passo na cadeia. Essa concordância entre as duas abordagens reforça a consistência na modelagem e na compreensão do processo subjacente, destacando a confiabilidade dos resultados obtidos.

Além disso, com base nas simulações previamente conduzidas e considerando as características do cenário genético em análise, acreditamos que a árvore de contexto Bayesiana proporcionou a melhor estimação. Os resultados dessas simulações são apresentados na Tabela 5.6, que evidencia as proporções de acertos dos três algoritmos em estudo em diferentes tamanhos de amostra. Destaca-se que a análise foi realizada considerando uma árvore Markoviana completa de ordem 2 e um alfabeto $\mathcal{A} = \{0, 1, 2, 3\}$.

Capítulo 6

Considerações Finais

Neste trabalho, estudamos alguns métodos de inferência de árvores de contexto sob abordagens frequentistas e Bayesiana, explorando conceitos, técnicas e aplicações em diversos cenários. Os cenários considerados incluíram desde árvores compatíveis com cadeias de Markov de ordem completa quanto de memória com alcance variável sendo aplicados em estudos com dados eletrofisiológicos e genéticos. As simulações realizadas abordaram três dos métodos discutidos ao longo do trabalho: o algoritmo contexto proposto por Rissanen (1983a), a modificação sugerida por Galves e Leonardi (2008), e, por fim, a árvore de contexto Bayesiana desenvolvida por Kontoyiannis *et al.* (2022).

Na avaliação do desempenho dos algoritmos, notamos que o algoritmo contexto modificado por Galves e Leonardi (2008) apresentou melhorias à medida que o tamanho da amostra aumentava. Essa tendência sugere que esse método pode ser mais eficaz em cenários onde há uma grande quantidade de dados disponíveis. Em contrapartida, tanto o algoritmo contexto de Rissanen quanto a árvore de contexto Bayesiana não enfrentaram grandes desafios ao realizar as estimações com tamanhos amostrais a partir de 5000. Além disso, destacamos o método da árvore de contexto Bayesiana por não apresentar limitações quanto ao tamanho máximo da árvore a depender da quantidade de dados disponíveis.

No que diz respeito à aplicação computacional, o código está disponível no Google Colab Trabalho de Conclusão de Curso "Seleção Estatística de Árvores de Contexto".

Referências Bibliográficas

Abakuks, A. (2012). The synoptic problem: on Matthew's and Luke's use of Mark. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **175**(4), 959–975.

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

Boys, R., Henderson, D. A. e Wilkinson, D. (2000). Detecting homogeneous segments in dna sequences by using hidden markov models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **49**, 269–285.

Bühlmann, P. e Wyner, A. J. (1999). Variable length Markov chains. *The Annals of Statistics*, **27**(2), 480–513.

Csiszár, I. e Talata, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Transactions on Information Theory*, **52**(3), 1007–1016.

Gabadinho, A. e Ritschard, G. (2016). Analyzing state sequences with probabilistic suffix trees: The pst r package. *Journal of Statistical Software*, **72**, 1–39.

Galves, A. e Leonardi, F. (2008). Exponential inequalities for empirical unbounded context trees. *In and Out of Equilibrium*, **2**, 257–269.

Galves, A. e Löcherbach, E. (2008). Stochastic chains with memory of variable length. *arXiv preprint arXiv:0804.2050*.

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

Garivier, A. e Leonardi, F. (2011). Context tree selection: a unifying view. *Stochastic Processes and their Applications*, **121**(11), 2488–2506.

Girardi, V. A. (2021). Inferência da conectividade neuronal via estimação de medidas da teoria da informação. *Bacharelado em Estatística*, página 169.

Kontoyiannis, I. (2022). Context-tree weighting and Bayesian context trees: Asymptotic and non-asymptotic justifications. *arXiv preprint arXiv:2211.02676*.

Kontoyiannis, I., Mertzanis, L., Panotopoulou, A., Papageorgiou, I. e Skoularidou, M. (2022). Bayesian context trees: Modelling and exact inference for discrete time series. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **84**(4), 1287–1323.

Leonardi, F. (2007). Rate of convergence of penalized likelihood context tree estimators. *Manuscript*.

Leonardi, F. G. (2006). A generalization of the PST algorithm: modeling the sparse nature of protein sequences. *Bioinformatics*, **22**(11), 1302–1307.

Lungu, V., Papageorgiou, I. e Kontoyiannis, I. (2022a). Bayesian change-point detection via context-tree weighting. Em *2022 IEEE Information Theory Workshop (ITW)*, páginas 125–130. IEEE.

Lungu, V., Papageorgiou, I. e Kontoyiannis, I. (2022b). Change-point detection and segmentation of discrete data using Bayesian context trees. *arXiv preprint arXiv:2203.04341*.

Mächler, M. e Bühlmann, P. (2004). Variable length Markov chains: methodology, computing, and software. *Journal of Computational and Graphical Statistics*, **13**(2), 435–455.

Markov, A. A. (1906). Extension of the law of large numbers to dependent events. *Bulletin of the Society of the Physics Mathematics*, **2**, 155–156. (In Russian).

Markov, A. A. (2006). An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains. *Science in Context*, **19**(4), 591–600.

Mendes, G. H. C. (2018). Modelagem de sequências de dna por meio de cadeias de ordem variável. *Bacharelado em Estatística*, página 52.

Merhav, N. e Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory*, **44**(6), 2124–2147.

Papageorgiou, I., Kontoyiannis, I., Mertzanis, L., Panotopoulou, A. e Skoularidou, M. (2021). Revisiting context-tree weighting for Bayesian inference. Em *2021 IEEE International Symposium on Information Theory (ISIT)*, páginas 2906–2911. IEEE.

Raftery, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society: Series B (Methodological)*, **47**(3), 528–539.

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

Rissanen, J. (1983b). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, **11**(2), 416–431.

Rissanen, J. (1986). Complexity of strings in the class of Markov sources. *IEEE Transactions on Information Theory*, **32**(4), 526–532.

Ron, D., Singer, Y. e Tishby, N. (1996). The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning*, **25**, 117–149.

Sarkar, A. e Dunson, D. B. (2016). Bayesian nonparametric modeling of higher order Markov chains. *Journal of the American Statistical Association*, **111**(516), 1791–1803.

Weinberger, M. J., Merhav, N. e Feder, M. (1994). Optimal sequential probability assignment for individual sequences. *IEEE Transactions on Information Theory*, **40**(2), 384–396.

Willems, F. M., Shtarkov, Y. M. e Tjalkens, T. J. (1995). The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory*, **41**(3), 653–664.