XXXIII Congresso de Iniciação Científica da Unicamp

Introdução Às Cadeias Estocásticas com Memória de Alcance Variável

Palavras chave: CEMAV, Redução de ordem, Esqueleto de cadeia de Markov

Caio Théodore Genovese Huss Oliveira - IMECC, UNICAMP
Christophe Frédéric Gallesco (orientador) - IMECC, UNICAMP

1 Introdução

O projeto expande o conhecimento sobre processos estocásticos adquirido na graduação, abordando o tema de cadeias estocásticas com memória de alcance variável (CEMAV, ou *VLMC*).

Inicialmente limitado ao estudo teórico e potencial aplicação de teorias já exploradas, o projeto teve seu escopo expandido quando foi identificada a possibilidade de desenvolver um algoritmo complementar à nova pesquisa do professor orientador[1].

Uma análise de dados foi prevista, dependente da disponibilidade de dados adequados. Infelizmente, não foi possível encontrar tal conjunto durante o projeto.

Cadeias estocásticas com memória de alcance variável foram introduzidas por Jorma Rissanen em 1983[2], como método de compressão de dados binários. Sua memória variável permite capturar alguns padrões complexos (de alta ordem) sem aumentar a ordem da cadeia inteira, tornando-se muito mais eficiente que uma cadeia de ordem fixa, em situações nas quais, para algumas sequências, o passado relevante é muito maior que na maioria das outras.

CEMAVs possuem aplicações em diversas áreas como, por exemplo, na linguística[3], microbiologia[4], e análise de comportamento na internet[5].

Uma redução de ordem para cadeias com transições proibidas foi desenvolvida em 2025[1], tendo o esqueleto[6] como um caso particular. O esqueleto limita-se a olhar para o passado de comprimento mínimo para que possa se determinar quais transições tem probabilidade nula, e no entanto compartilha propriedades com a cadeia original, como a irredutibilidade e periodicidade.

2 Métodos, Materiais e Cronograma

Todos os códigos foram desenvolvidos na linguagem R[7]. Os códigos para a estimação do esqueleto foram em parte inspirados pelo pacote *VLMC*[8]. Otimizações foram feitas ao longo do projeto com base nas práticas do livro *Advanced R*[9].

O projeto esteve em desenvolvimento de setembro de 2024 a julho de 2025, o fim adiantado é devido à conclusão de curso do aluno orientado.

Não houve impacto no desempenho acadêmico ou profissional do aluno orientado.

3 Resultados

Provas teóricas sobre continuidade, log-continuidade, não-nulidade fraca e forte, e renovação de CEMAVs foram replicadas a partir de exercícios de [10] e [11].

Uma propriedade de esqueletos foi encontrada, que acelera a análise da irredutibilidade da cadeia em alguns casos.

Duas Cadeias foram simuladas (esqueletos em vermelho):

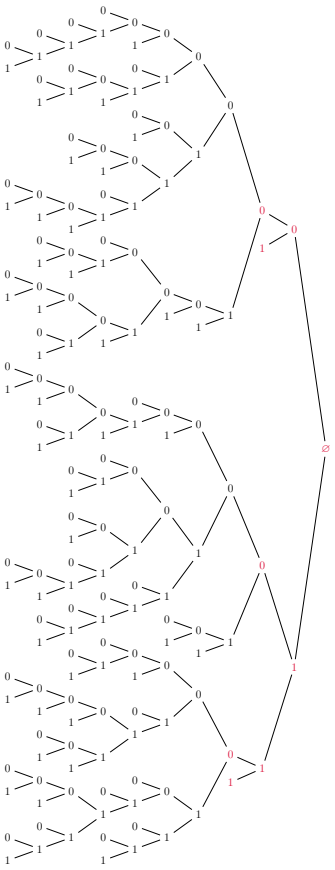

Figura 1: Árvore da cadeia 1

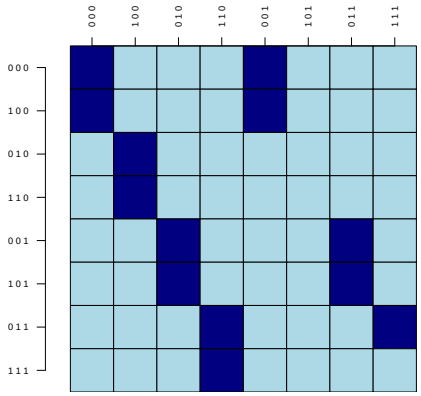

Figura 2: Matriz de transição do esqueleto da Cadeia 1 (transições permitidas em escuro)

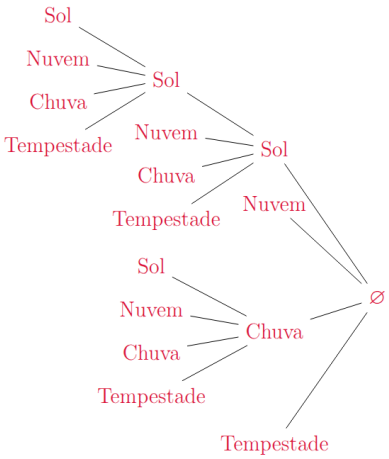

Figura 3: Esqueleto da cadeia 2

A árvore completa da cadeia 2 e as matrizes de transição não podem ser visualizadas apropriadamente por falta de espaço. A primeira cadeia é CEMAV, de ordem 10 com um esqueleto de ordem 3. A segunda é de ordem fixa 5, com um esqueleto de ordem 3. A segunda cadeia foi simulada para explorar alfabetos com mais de 2 símbolos. As reduções de tamanho das matrizes de transição *cadeia* $\rightarrow$ *esqueleto* foram de $1024 \times 1024 \rightarrow 8 \times 8$ (cadeia 1) e $1024 \times 1024 \rightarrow 64 \times 64$ (cadeia 2).

Dois algoritmos foram desenvolvidos para obter o esqueleto de uma cadeia estocástica: O primeiro é analítico, partindo de uma árvore já conhecida. O segundo, empírico, estimando o esqueleto a partir de dados. Uma descrição abreviada dos algoritmos é:

- Algoritmo Analitico:

    1. Transforme o vetor de probabilidades $p = (p_1, \cdots, p_a)$ associado a cada contexto em um vetor de transições permitidas $v = (v_1, \cdots, v_a), v_i = \mathbf{1}_{(0,1]}(p_i)$.
    2. Para cada nodo pai de folhas da árvore, corte todas suas folhas se seus vetores v forem identicos.

- Algoritmo Empírico:

    1. Escolha os valores dos critérios de sensibilidade, resultando em N_{min}.
    2. Para cada passado, começando pela raiz ($\varnothing$), estime o vetor de probabilidades.
    3. Se o passado tiver um número de ocorrências maior ou igual a N_{min} e não tiver nenhum $p_i = 1$, crie seus filhos adicionando cada um dos símbolos do alfabeto no começo e estime o vetor de probabilidades para eles. Repira até nenhum passado atender os pré-requisitos.
    4. Aplique o algortimo analítico na arvore obtida, levando apenas em conta os passados com ocorrências suficientes quando aplicando cortes.
    5. Passados sem o número mínimo de ocorrências recebem um vetor de transições permitidas $v = (1, 1, \cdots, 1)$.

Ambos algoritmos foram implementados na linguagem R, em um pacote a ser publicado posteriormente.

O algoritmo analítico será publicado em conjunto com os professores C. Gallesco (orientador) e D. Takahashi[6].

4 Discussão

As árvores simuladas foram "verificadas" com o pacote *VLMC*[8]. Para cada, 100.000 símbolos foram simulados, de forma reprodutível e disponível no relatório final. A simulação se tornou relativamente rápida com os códigos desenvolvidos, levando cerca de 2 segundos. Para economia de espaço e como as amostras obtidas já eram adequadas para o objetivo do projeto, não julgou-se necessário aumentar o tamanho.

Uma análise da irredutibilidade de uma matriz $M_{k\times k}$ envolve até aproximadamente k^3 produtos de vetor. O algoritmo analítico, no pior caso possível, para uma cadeia de ordem k, envolve $|\mathcal{A}|^k$ cortes, onde $\mathcal{A}$ é o alfabeto da cadeia, conjunto de símbolos que ela pode assumir.

A redução de custo com o esqueleto pode então ser estimada, em produtos e comparações de vetor, de $|A|^{3k}$ para $|A|^k|A|^{3d}$ operações, para um esqueleto de ordem d. Em casos $d << k$, se torna aproximadamente uma redução de ordem cúbica. Para a cadeia 1 simulada, esta redução é de 2^{30} para 2^{19}, um custo 2048 vezes menor.

O custo da estimação do algoritmo empírico ainda resta a ser estimado, mas como segue uma lógica similar a estimações de cadeia já existentes, a redução total deve ser próxima à já discutida.

O projeto foi uma ótima iniciação no mundo da pesquisa acadêmica, e permitiu avaliar os tipos de pesquisa pelos quais me interesso.

Referências

[1] Christophe Gallesco, Alexsandro Gallo e Daniel Yasumasa Takahashi. *Uniqueness of stationary compatible probability measure for chains of infinite order with forbidden transitions*. 2025. arXiv: 2507.16981 [math.PR]. URL: https://arxiv.org/abs/2507.16981.

[2] Jorma Rissanen. "A universal data compression system". Em: *Information Theory, IEEE Transactions on* 29 (out. de 1983), pp. 656–664. DOI: 10.1109/TIT.1983.1056741.

[3] Antonio Galves et al. "Context tree selection and linguistic rhythm retrieval from written texts". Em: *The Annals of Applied Statistics* (2012). DOI: 10.1214/11-AOAS511.

[4] Weinan Liao et al. "Alignment-free Transcriptomic and Metatranscriptomic Comparison Using Sequencing Signatures with Variable Length Markov Chains". Em: *Scientific Reports 6, Nature* (2016). DOI: https://doi.org/10.1038/srep37243.

[5] Jose Borges e Mark Levene. "Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions". Em: *IEEE Transactions on Knowledge and Data Engineering* 19.4 (2007), pp. 441–452. DOI: 10.1109/TKDE.2007.1012.

[6] Caio T. G. Huss Oliveira, Christophe F. Gallesco e Daniel Y. Takahashi. *Reduction algorithm for high order Markov chains (nome provisório)*. 2025 (em desenvolvimento).

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[8] Martin Maechler. *VLMC: Variable Length Markov Chains ('VLMC') Models*. R package version 1.4-4. 2024. URL: https://CRAN.R-project.org/package=VLMC.

[9] Roger M. Sauter. "Advanced R (2nd ed.)" Em: *Technometrics* 62.3 (2020), pp. 417–417. DOI: 10.1080/00401706.2020.1783959. eprint: https://doi.org/10.1080/00401706.2020.1783959. URL: https://doi.org/10.1080/00401706.2020.1783959.

[10] Antonio Galves e Eva Löcherbach. *Stochastic chains with memory of variable length*. 2008. arXiv: 0804.2050 [math.PR]. URL: https://arxiv.org/abs/0804.2050.

[11] Antonio Galves, Roberto Fernandez e Ferrari Galves. *Coupling, renewal and perfect simulation of chains of infinite order*. Jan. de 2001.