Universidade de Brasília

Critério de Determinação Eficiente para Estimação de Cadeias de Markov de Partição Mínima

Diego Felipe Sôlha Pereira

Orientadora: Profa. Dra. Cátia Regina Gonçalves

Departamento de Matemática
Universidade de Brasília

Dissertação apresentada como requisito parcial para obtenção do grau de *Mestre em Matemática*

Brasília, 13 de Dezembro de 2021

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Matemática

Critério de Determinação Eficiente para Estimação de Cadeias de Markov de Partição Mínima

por

Diego Felipe Sôlha Pereira

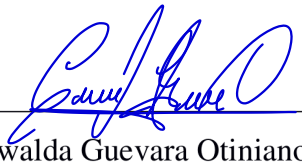
Dissertação apresentada ao Departamento de Matemática da Universidade de Brasília como parte dos requisitos para obtenção do grau de

MESTRE EM MATEMÁTICA

Brasília, 13 de dezembro de 2021.

Comissão Examinadora:

Profa. Dra. Cátia Regina Gonçalves - MAT/UnB (Orientadora)



Profa. Dra. Cira Etheowalda Guevara Otiniano – EST/UnB (Membro)

Profa. Dra. Verónica Andrea González López – UNICAMP (Membro)

Dedico este trabalho ao meu irmão Breno Ricardo que simboliza as qualidades da resistência e da perseverança e à Lalinha que atualmente está se divertindo no céu dos cachorros.

Agradecimentos

Gostaria de agradecer a Deus por tudo que fez e faz por mim e pela minha família.

Ao meu pai, à minha mãe e ao meu irmão, meus maiores heróis e inspirações nessa vida e a toda minha família.

Aos meus amigos Fernando e Saulo pelo companheirismo e força de sempre.

Aos meus "hupamigos", Amadeus, Ângelo, Guilherme, Gabriel, José e Leonardo, pessoas maravilhosas que me fizeram ter certeza de que fazia sentido estudar e ensinar matemática.

À minha orientadora, professora Cátia, que, com todo seu carisma e irreverência ao ensinar, me mostrou uma nova perspectiva de estudo da matemática por meio da Probabilidade e por todo apoio, dedicação e trabalho empenhado para que fosse possível finalizarmos esta dissertação.

À minha namorada Bruna por todo amor e companheirismo e por ser uma força propulsora que me incentiva a sair da zona de conforto e melhorar a cada dia.

Ás professoras Verónica Andrea González-López e Cira Etheowalda Guevara Otiniano por aceitarem participar da banca da defesa e pelas sugestões de melhoria do trabalho.

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil.

Resumo

Neste trabalho, estudamos o Modelo de Markov com Partição, estabelecido na literatura como um modelo de Cadeia de Markov de ordem finita, no qual o espaço de estados é particionado em classes de equivalência formadas por estados com as mesmas probabilidades de transição. Apresentamos em detalhes as propriedades de consistência do método de estimação da Partição Mínima, baseado no Critério de Informação Bayesiano (*BIC*), proposto por García e González-López (2017). Além disso, fundamentados nesse estudo, propomos a utilização do Critério de Determinação Eficiente (*EDC*), que estende o *BIC*, para a seleção do Modelo de Markov de Partição Mínima e estabelecemos a consistência da partição estimadora sob condições suaves.

Palavras-chave: Cadeias de Markov de ordem superior, Critério de Informação Bayesiano, Critério de Determinação Eficiente.

Abstract

In this work, we study the Partition Markov Model, established in the literature as a finite order Markov Chain model, in which the state space is partitioned into equivalence classes formed by states with the same transition probabilities. We present in details the consistency properties of the minimum partition estimation method, based on the Bayesian Information Criterion (*BIC*), proposed by García and González-López (2017). Furthermore, based on this study, we propose the use of the Efficient Determination Criterion (*EDC*), which extends the *BIC*, for the selection of the Minimum Partition Markov Model and the consistency of the estimating partition is established under mild conditions.

Keywords: High-Order Markov Chains, Bayesian Information Criterion, Efficient Determination Criterion.

Sumário

Lista de Tabelas **xiv**

Introdução **1**

1 Cadeias de Markov com Partição **4**
1.1 Preliminares 4
1.2 Cadeias de Markov de ordem superior 7
1.3 Cadeias de Markov com Partição 14
1.3.1 Definição e Notações 15
1.3.2 Máxima Verossimilhança Modificada 22

2 Estimação pelo Critério de Informação Bayesiano **26**
2.1 O *BIC* em Cadeias de Markov com Partição 27
2.2 Consistência 32

3 Estimação pelo Critério de Determinação Eficiente **42**
3.1 O *EDC* em Cadeias de Markov com Partição 43
3.2 Consistência 48
3.3 Seleção da Partição Mínima 54
3.4 A busca por um termo de penalidade ótimo 60

Conclusão **62**

Bibliografia **65**

Lista de Tabelas

1.1 Probabilidades de Transição 19

Introdução

Os processos com estrutura Markoviana têm sido amplamente utilizados na modelagem de fenômenos nas mais diferentes áreas.

Nas situações em que a informação em um dado instante depende das informações em um número fixo de instantes anteriores, os modelos de Cadeias de Markov de ordem fixa têm se mostrado eficientes em inúmeras aplicações, como, por exemplo, em meteorologia ([20]), em genética ([3]), entre muitas outras.

Em outras situações, o modelo mais apropriado pode ser aquele em que a ordem da Cadeia de Markov não é fixa, denominados modelos de Cadeias de Markov de Comprimento Variável (VLMC) ([2], [6]). Neste caso, a memória da cadeia depende dos denominados *contextos* e as sequências de estados anteriores com o mesmo contexto possuem as mesmas probabilidades de transição, sendo, então, agrupados.

Do ponto de vista da estimação, a vantagem de considerar um modelo de Cadeia de Markov de Comprimento Variável sobre um modelo de Cadeia de Markov completa de ordem fixa é a redução e flexibilização do número de parâmetros a serem estimados, que no modelo de Cadeia de Markov completa de ordem fixa M e espaço de estados A finito, cresce exponencialmente com a ordem M.

Inspirados pelas ideias do VLMC e considerando que "em base de dados com estrutura Markoviana é frequentemente observado um grau considerável de redundância, o que significa que diferentes sequências de símbolos têm o mesmo efeito sobre a lei de probabilidade do processo " ([13], tradução nossa), García e González-López, em [11], visando minimizar o número de parâmetros a serem estimados para modelar o processo através dessa redundância, introduziram o então chamado Modelo de Markov Mínimo (*Minimal Markov Model*), que posteriormente em [12] e [13] foi denominado Modelo de Markov com Partição (*Partition Markov Model*).

Nesse modelo a redundância é minimizada considerando-se uma partição $\mathcal{L}^*$ do espaço $S = A^M$, gerada por uma relação de equivalência na qual elementos de S com as mesmas probabilidades de transição são equivalentes.

Esse modelo engloba tanto o modelo de Cadeia de Markov (completa) de ordem finita, quanto o modelo VLMC e pode apresentar um número menor de parâmetros a serem estimados. Exemplos de diferentes aplicações e estudos relacionados aos Modelos de Markov com Partição podem ser encontrados em [9], [10], [12] e [14].

Em [13], para processos estacionários com espaço de estados finito, García e González-López propuseram a utilização do Critério de Informação Bayesiano (*BIC*) para construir uma estratégia consistente que consiste em encontrar, a partir de n observações da cadeia, o modelo com partição contendo um número mínimo de partes, para representar a lei de probabilidade do processo. Um algoritmo para a seleção da partição estimadora da Partição Mínima $\mathcal{L}^*$, baseado no *BIC*, é apresentado pelos autores e é demonstrado que a sequência de partições produzidas pelo algoritmo converge quase certamente para a partição $\mathcal{L}^*$.

O *BIC* é um critério de seleção de modelos amplamente utilizado em diferentes situações. No contexto de cadeias de Markov de ordem finita desconhecida foi utilizado por Katz em [19] para a obtenção de um estimador da ordem da cadeia e em [5] uma prova formal da consistência forte (quase certa) desse estimador é apresentada. Nesse sentido, em [23], Zhao, Dorea e Gonçalves propuseram a utilização do Critério de Determinação Eficiente (*EDC*), que generaliza o *BIC*, para a estimação da ordem de uma Cadeia de Markov e, assumindo que um limitante superior da ordem da cadeia é conhecido, demonstraram a consistência forte do estimador, sob condições suaves.

Posteriormente, em [7], Dorea demonstrou a consistência forte do *EDC* para a estimação da ordem da cadeia, sem a necessidade da hipótese do conhecimento prévio de um limitante, sob condições mais brandas sobre o termo de penalidade e propôs um termo de penalidade ótimo para o *EDC*, propondo teoricamente que o *EDC* com termo de penalidade ótimo pode ser um critério de informação fortemente consistente que permita uma estimação mais eficiente que o *BIC*, isto é, diminuindo a subestimação da ordem verdadeira da cadeia com um tamanho amostral menor.

Diante disso, duas questões surgem naturalmente. É possível utilizar, de forma similar a [23] o *EDC* para a estimação da Partição Mínima do Modelo de Markov com Partição, preservando a consistência forte, ou seja, estimando a Partição Mínima de forma quase certa com uma amostra suficientemente grande? Se sim, é possível propor algum termo de penalidade ótimo, como em [7], de tal forma a obter um estimador consistente para a Partição Mínima que seja, em algum sentido, mais eficiente que o *BIC*?

Neste trabalho, baseados em [13] e [23], nos propomos a responder a primeira questão, utilizando o *EDC* para a obtenção de um estimador consistente mais geral para a Partição Mínima de uma Cadeia de Markov com Partição.

A análise da segunda questão sobre a busca por um possível termo de penalidade ótimo para o *EDC* na estimação da Partição Mínima, que, em algum sentido de estimação, possa se mostrar mais eficiente que o *BIC* e a suavização das condições para a consistência do estimador proposto, baseado no *EDC*, serão objetos de estudos futuros. De qualquer modo, uma breve discussão sobre o assunto é apresentada na Seção 3.4, sob a ótica dos trabalhos [7], [8] e [13].

Assim, no Capítulo 1 apresentamos inicialmente conceitos preliminares gerais e os principais elementos relativos ao estudo de Cadeia de Markov de ordem superior. Mais ainda, baseados no artigo de García e González-López [13], apresentamos o conceito de Cadeia de Markov com Partição, as principais notações utilizadas e a dedução da função de máxima verossimilhança modificada, que será empregada na definição dos critérios *BIC* e *EDC* para a estimação da Partição Mínima da cadeia.

No Capítulo 2 apresentamos em detalhes a estratégia baseada no *BIC* proposta por García e González-López e os principais resultados obtidos em [13].

No Capítulo 3, motivados e seguindo as mesmas ideias de García e González-López em [13], estabelecemos o conceito do critério *EDC* no contexto do Modelo de Markov com Partição e obtemos, sob as mesmas condições sobre o termo de penalidade do *EDC*, dadas em [23], a consistência forte do *EDC* para a estimação da Partição Mínima do Modelo de Markov com Partição. Além disso, apresentamos uma versão estendida, utilizando o *EDC*, do algoritmo para a seleção da Partição Mínima apresentado em [13] e estabelecemos, ainda, a consistência forte da sequência de partições geradas por esse algoritmo.

Por fim, são apresentadas as Conclusões do estudo realizado nesta dissertação.

Capítulo 1

Cadeias de Markov com Partição

O objetivo central deste capítulo é apresentar o conceito e os principais elementos do Modelo de Markov com Partição, que é o objeto de estudo desta dissertação.

Além disso, apresentamos inicialmente alguns conceitos e resultados preliminares que serão utilizados nos capítulos posteriores.

Assim, na Seção 1.1, apresentamos conceitos e resultados gerais básicos e, na Seção 1.2, expomos os principais elementos relativos às Cadeias de Markov de ordem superior.

Finalmente, na Seção 1.3, baseados no artigo de García e González-López [13], apresentamos o conceito de Cadeia de Markov com Partição, as notações a serem utilizadas e concluímos com a dedução da função de máxima verossimilhança modificada, a qual será empregada para a definição dos critérios de informação *BIC* e *EDC* para a estimação da Partição Mínima da cadeia que serão apresentados e estudados no Capítulo 2 e no Capítulo 3, respectivamente.

As referências bibliográficas utilizadas neste capítulo são: [1], [2], [4], [7], [13], [15], [16], [17], [18] e [22].

1.1 Preliminares

Nesta seção, apresentamos alguns conceitos e resultados básicos, que serão úteis no decorrer desta dissertação.

As referências bibliográficas desta seção são: [1], [13], [17] e [18].

Definição 1.1. Seja $(\Omega, \mathcal{F}, \mathbb{P})$, um espaço de probabilidade. Dizemos que uma sequência de eventos $A_n \in \mathcal{F}$, $n = 1, 2 \ldots$, ocorre *quase certamente quando* $n \to \infty$, se

$$\mathbb{P}(\text{ocorrência de } A_n \text{ para todo } n \text{ suficientemente grande}) = 1, \tag{1.1}$$

ou equivalentemente, se

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty}A_k\right)=\mathbb{P}\left(\liminf_{n\to\infty}A_n\right)=1.$$

Definição 1.2 (Entropia Relativa)**.** Sejam $p=\Big(p(x_1),\ldots,p(x_n)\Big)$ e $q=\Big(q(x_1),\ldots,q(x_n)\Big)$ duas medidas de probabilidade discretas em $\mathcal{X}=\{x_1,\ldots,x_n\}\subset\mathbb{R}$. Definimos a *entropia relativa de p com relação a q* como sendo

$$D\Big(p||q\Big)=\sum_{x\in\mathcal{X}}p(x)\ln\left(\frac{p(x)}{q(x)}\right). \tag{1.2}$$

A entropia relativa também é chamada de divergência (ou distância) de Kullback-Leibler e pode ser interpretada como uma medida de proximidade entre duas distribuições. No entanto, ela não define uma métrica, pois não é simétrica.

A seguir, apresentamos duas propriedades básicas de entropia relativa que serão utilizadas nos próximos capítulos. Um estudo mais detalhado sobre entropia relativa pode ser encontrado, por exemplo, em [17].

Proposição 1.1 (Positividade da Entropia Relativa)**.** Sejam $p=\Big(p(x),x\in\mathcal{X}\Big)$ e $q=\Big(q(x),x\in\mathcal{X}\Big)$ duas distribuições de probabilidade discretas em $\mathcal{X}=\{x_1,x_2,\ldots,x_n\}\subset\mathbb{R}$, então

$$D\Big(p||q\Big)\geq 0 \tag{1.3}$$

e a igualdade ocorre se, e somente se, $p(x)=q(x)\ ,\ \forall\, x\in\mathcal{X}$

Demonstração. Vamos utilizar a conhecida desigualdade

$$\ln(x)\geq 1-\frac{1}{x},\forall\, x>0, \tag{1.4}$$

onde a igualdade ocorre se, e somente se, $x=1$.

Assim, usando (1.4) em (1.2), obtemos

$$\begin{aligned}D\Big(p||q\Big)&=\sum_{i=1}^{n}p(x_i)\ln\left(\frac{p(x_i)}{q(x_i)}\right)\\&\geq\sum_{i=1}^{n}p(x_i)\left(1-\frac{q(x_i)}{p(x_i)}\right)=0\end{aligned}$$

Além disso, se $p(x) = q(x)$, $\forall\, x \in \mathcal{X}$, é imediato que $D\Big(p||q\Big) = 0$. Por outro lado, se $D\Big(p||q\Big) = 0$, então temos

$$\sum_{i=1}^{n} p(x_i) \left[\ln\left(\frac{p(x_i)}{q(x_i)}\right) - \left(1 - \frac{q(x_i)}{p(x_i)}\right)\right] = 0.$$

Agora, por (1.4), temos que todas as parcelas da soma anterior são não negativas. Logo, devemos ter $\ln\left(\dfrac{p(x_i)}{q(x_i)}\right) = 1 - \dfrac{q(x_i)}{p(x_i)}$, $\forall\, i = 1,2,\ldots,n$. Mas a igualdade em (1.4) só é válida para $x = 1$. Portanto, segue que $p(x_i) = q(x_i)$, $\forall\, i = 1,\ldots,n$

□

Observação 1.1. Como consequência da Proposição 1.1, podemos obter o seguinte resultado, conhecido como *Desigualdade de log-soma*:

Sejam $a_1,\ldots,a_n$, $b_1,\ldots,b_n$ números reais positivos. Se $a = \sum_{i=1}^{n} a_i$ e $b = \sum_{i=1}^{n} b_i$, então

$$\sum_{i=1}^{n} a_i \ln\left(\frac{a_i}{b_i}\right) \geq a \cdot \ln\left(\frac{a}{b}\right) \tag{1.5}$$

e a igualdade ocorre se, e somente se, $\dfrac{a_i}{b_i} = \dfrac{a}{b}$, constante, $\forall\, i = 1,\ldots,n$.

De fato, basta considerar $p_i = \dfrac{a_i}{a}$ e $q_i = \dfrac{b_i}{b}$, $i = 1,2,\ldots,n$. Então $p = (p_1,\ldots,p_n)$ e $q = (q_1,\ldots,q_n)$ são distribuição de probabilidades discretas com $p_i > 0$ e $q_i > 0$.

Logo, de (1.3) segue que

$$\sum_{i=1}^{n} \frac{a_i}{a} \ln\left(\frac{a_i}{a}\frac{b}{b_i}\right) \geq 0.$$

Ou seja,

$$\frac{1}{a}\sum_{i=1}^{n} a_i \ln\left(\frac{a_i}{b_i}\right) - \ln\left(\frac{a}{b}\right) \geq 0$$

e temos (1.5).

Além disso, a igualdade ocorre se, e somente se, $p_i = \dfrac{a_i}{a} = \dfrac{b_i}{b} = q_i$, $\forall\, i = 1,\ldots,n$, ou seja, $\dfrac{a_i}{b_i} = \dfrac{a}{b}$, constante, $\forall\, i = 1,\ldots,n$. ◊

Proposição 1.2. Sob as mesmas hipóteses da Proposição 1.1, temos que

$$D\Big(p||q\Big) \leq \sum_{x\in\mathcal{X}} \frac{\Big(p(x) - q(x)\Big)^2}{q(x)}. \tag{1.6}$$

Demonstração. Podemos reescrever (1.4) com $\frac{1}{x}$ no lugar de x e obter

$$\ln(x) \leq x-1, \forall\, x>0. \tag{1.7}$$

Usando (1.7) em (1.2) temos

$$\begin{aligned} D\Big(p||q\Big) &\leq \sum_{x\in\mathcal{X}} p(x)\left(\frac{p(x)}{q(x)}-1\right) \\ &= \sum_{x\in\mathcal{X}} \frac{[p(x)]^2 - p(x)\cdot q(x)}{q(x)}. \end{aligned}$$

Completando quadrados na expressão do lado direito e como $\sum_{x\in\mathcal{X}} p(x) = \sum_{x\in\mathcal{X}} q(x) = 1$, obtemos

$$\begin{aligned} D\Big(p||q\Big) &\leq \sum_{x\in\mathcal{X}} \frac{\Big(p(x)-q(x)\Big)^2}{q(x)} + \sum_{x\in\mathcal{X}} \Big(p(x)-q(x)\Big) \\ &= \sum_{x\in\mathcal{X}} \frac{\Big(p(x)-q(x)\Big)^2}{q(x)}. \end{aligned}$$

□

1.2 Cadeias de Markov de ordem superior

Nesta seção, apresentaremos conceitos e resultados sobre Cadeias de Markov de ordem superior que serão necessários para o desenvolvimento dos próximos capítulos.

As referências bibliográficas utilizadas nesta seção são: [4], [7], [13], [16], [18] e [22].

Para o que segue, consideramos $\{X_t\} = \{X_t,\, t = 1,2,\dots\}$, um processo estocástico a tempo discreto definido sobre um mesmo espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$ e com espaço de estados discreto A, ou seja, $A \subset \mathbb{R}$ é enumerável e $\mathbb{P}(X_t \in A,\, t = 1,2,\dots) = 1$.

No contexto da Teoria da Informação, o espaço de estados A é comumente chamado o *alfabeto* sobre o qual as variáveis aleatórias X_t, $t \geq 1$, tomam seus valores.

Por facilidade, para $m < p$ e $l < j$, inteiros positivos, denotamos

$$\begin{aligned} &X_m^p = (X_m, X_{m+1}, \ldots, X_p); \\ &a_m^p = (a_m, a_{m+1}, \ldots, a_p), \text{ onde } a_i \in A, \forall\, i \geq 1; \\ &a_m^p b_l^j = (a_m, a_{m+1}, \ldots, a_p, b_l, b_{l+1} \ldots, b_j),\ a_i,\ b_i \in A,\ \forall\, i \geq 1. \end{aligned} \tag{1.8}$$

Definição 1.3. Um processo estocástico $\{X_t\}$ a tempo discreto e espaço de estados discreto A é chamado uma *Cadeia de Markov de ordem $M \in \mathbb{N}$, homogênea ao longo do tempo*, se, para $a_i \in A$, $i = 1, 2, \ldots, n+1$, satisfaz

(i) $\mathbb{P}(X_{n+1} = a_{n+1} \mid X_1^n = a_1^n) = \mathbb{P}\left(X_{n+1} = a_{n+1} \mid X_{n+1-M}^n = a_{n+1-M}^n\right)$, $\forall\, n \geq M$ (Propriedade de Markov);

(ii) $\mathbb{P}\left(X_{n+1} = a_{n+1} \mid X_{n+1-M}^n = a_{n+1-M}^n\right) = \mathbb{P}\left(X_{m+1} = a_{n+1} \mid X_{m+1-M}^m = a_{n+1-M}^n\right)$, $\forall\, n, m \geq M$ (Homogênea ao longo do tempo).

Neste trabalho, consideramos sempre cadeias homogêneas no tempo. Desse modo, frequentemente, por simplicidade, omitimos o termo "homogêneas no tempo".

Se $\{X_t\}$ é uma Cadeia de Markov de ordem M e espaço de estados A, chamamos de *probabilidades de transição* da cadeia as probabilidades

$$P(a \mid a_1^M) := \mathbb{P}(X_{M+1} = a \mid X_1^M = a_1^M) = \mathbb{P}(X_{n+1} = a \mid X_{n+1-M}^n = a_1^M),\ \forall\, n \geq M \text{ e } a \in A. \tag{1.9}$$

Se $M = 1$, chamamos $\{X_t\}$ simplesmente de Cadeia de Markov (C.M.). Assumimos conhecidos os resultados básicos da teoria clássica de Cadeia de Markov que serão utilizados em diversas situações neste texto. Como referências para essa teoria, citamos [16] e [18]. Uma forma de estudar as propriedades de uma Cadeia de Markov de ordem M, através das propriedades de uma C.M. é através dos processos k-derivados de $\{X_t\}$ definidos a seguir.

Definição 1.4. Dada $\{X_t\}$, uma Cadeia de Markov de ordem M e espaço de estados A, definimos $\left\{Y_t^{(k)}\right\}$, com $Y_t^{(k)} = X_t^{t+k-1} = (X_t, \ldots, X_{t+k-1}) \in A^k$, *o processo k-derivado de* $\{X_t\}$, com $k \in \mathbb{N}$.

Mostramos a seguir que o processo k-derivado de $\{X_t\}$ é uma C.M. para $k \geq M$.

Proposição 1.3. Se $\{X_t\}$ é uma Cadeia de Markov de ordem M e espaço de estados A, então, para $k \geq M$, seu processo k-derivado, $\left\{Y_t^{(k)}\right\}$, é uma Cadeia de Markov de ordem 1, chamada de *Cadeia de Markov k-derivada de* $\{X_t\}$, com espaço de estados A^k.

Demonstração. Primeiramente, como o espaço de estados A de $\{X_t\}$ é discreto, então A^k, o espaço de estados de $\left\{Y_t^{(k)}\right\}$, também é discreto.

Para provar que o processo $\left\{Y_t^{(k)}\right\}$, com $k \geq M$, satisfaz a Propriedade de Markov, isto é, $\forall\, n \geq 1$,

$$\mathbb{P}\Big(Y_{n+1}^{(k)} = a_{n+1}^{(k)} \;\Big|\; Y_n^{(k)} = a_n^{(k)}, \ldots, Y_1^{(k)} = a_1^{(k)}\Big) = \mathbb{P}\Big(Y_{n+1}^{(k)} = a_{n+1}^{(k)} \;\Big|\; Y_n^{(k)} = a_n^{(k)}\Big),$$

com $a_i^{(k)} = (a_{i,1}, \ldots, a_{i,k}) \in A^k$, observe primeiramente que a probabilidade

$$\mathbb{P}\Big(Y_{n+1}^{(k)} = a_{n+1}^{(k)} \;\Big|\; Y_n^{(k)} = a_n^{(k)}, \ldots, Y_1^{(k)} = a_1^{(k)}\Big) = \mathbb{P}\Big(X_{n+k} = a_{n+1,k}, \ldots, X_{n+1} = a_{n+1,1} \;\Big|\; (X_n, \ldots, X_{n+k-1}) = (a_{n,1}, \ldots, a_{n,k}), \ldots, (X_1, \ldots, X_k) = (a_{1,1}, \ldots, a_{1,k})\Big)$$

é positiva somente para os $a_i^{(k)}$, tais que $a_{i,j} = a_{i-1,j+1}$, quando $i \in \{2,3,\ldots,n+1\}$ e $j \in \{1,2,\ldots,k-1\}$. Por facilidade de notação, denote $a_{i+j-1} = a_{i,j}$, $\forall\, i \in \{1,\ldots,n+1\}$ e $j \in \{1,\ldots,k\}$.

Assim, podemos escrever

$$\begin{aligned}
&\mathbb{P}\Big(Y_{n+1}^{(k)} = a_{n+1}^{(k)} \;\Big|\; Y_n^{(k)} = a_n^{(k)}, \ldots, Y_1^{(k)} = a_1^{(k)}\Big) = \\
&\mathbb{P}(X_{n+k} = a_{n+k}, \ldots, X_{n+1} = a_{n+1} \mid X_{n+k-1} = a_{n+k-1}, \ldots, X_1 = a_1) = \\
&\mathbb{P}\left(X_{n+k} = a_{n+k} \mid X_1^{n+k-1} = a_1^{n+k-1}\right).
\end{aligned}$$

Agora, como $\{X_t\}$ satisfaz a Propriedade de Markov (i) da Definição 1.3, segue que

$$\begin{aligned}
\mathbb{P}\Big(Y_{n+1}^{(k)} = a_{n+1}^{(k)} \;\Big|\; Y_n^{(k)} = a_n^{(k)}, \ldots, Y_1^{(k)} = a_1^{(k)}\Big) &= \mathbb{P}\Big(X_{n+k} = a_{n+k} \mid X_{n+k-M}^{n+k-1} = a_{n+k-M}^{n+k-1}\Big) \\
&= \mathbb{P}\Big(X_{n+k} = a_{n+k} \mid X_n^{n+k-1} = a_n^{n+k-1}\Big) \\
&= \mathbb{P}\Big(X_{n+1}^{n+k} = a_{n+1}^{n+k} \mid X_n^{n+k-1} = a_n^{n+k-1}\Big) \\
&= \mathbb{P}\Big(Y_{n+1}^{(k)} = a_{n+1}^{(k)} \;\Big|\; Y_n^{(k)} = a_n^{(k)}\Big).
\end{aligned}$$

Além disso, usando os mesmos argumentos e notações anteriores e como $\{X_t\}$ é homogênea no tempo, ou seja, satisfaz o item (ii) da Definição 1.3, podemos obter

$$\begin{aligned}\mathbb{P}\left(Y_{n+1}^{(k)} = a_{n+1}^{(k)} \;\middle|\; Y_n^{(k)} = a_n^{(k)}\right) &= \mathbb{P}\left(X_{n+k} = a_{n+k} \mid X_{n+k-M}^{n+k-1} = a_{n+k-M}^{n+k-1}\right) \\ &= P\left(a_{n+k} \mid a_{n+k-M}^{n+k-1}\right) \\ &= \mathbb{P}\left(X_{k+1} = a_{n+k} \mid X_1^k = a_n^{n+k-1}\right) \\ &= \mathbb{P}\left(X_2^{k+1} = a_{n+1}^{n+k} \mid X_1^k = a_n^{n+k-1}\right) \\ &= \mathbb{P}\left(Y_2^{(k)} = a_{n+1}^{(k)} \;\middle|\; Y_1^{(k)} = a_n^{(k)}\right).\end{aligned}$$

Ou seja, para $k \geq M$, $\left\{Y_t^{(k)}\right\}$ é uma Cadeia de Markov homogênea no tempo.

□

Observe, por (1.9), que as probabilidades de transição dos casos possíveis de $\left\{Y_t^{(k)}\right\}$ são dadas, para todo $t \geq 1$, por

$$\mathbb{P}\left(Y_{t+1}^{(k)} = a_2^{k+1} \;\middle|\; Y_t^{(k)} = a_1^k\right) = P(a_{k+1} \mid a_{k-M+1}^k). \tag{1.10}$$

Em particular, para o caso $k = M$, obtemos

$$\mathbb{P}\left(Y_{t+1}^{(M)} = a_2^{M+1} \;\middle|\; Y_t^{(M)} = a_1^M\right) = P(a_{M+1} \mid a_1^M). \tag{1.11}$$

Assim, associamos as probabilidades de transição de uma Cadeia de Markov $\{X_t\}$ de ordem M, com as probabilidades de transição de sua Cadeia de Markov M-derivada $\left\{Y_t^{(M)}\right\}$, que é uma Cadeia de Markov de ordem 1.

Para a estimação em Cadeias de Markov, é natural estudar o comportamento da cadeia após um número n grande de realizações, ou seja, seu comportamento limite. Na proposição a seguir, veremos uma condição suficiente sobre as probabilidades de transição de uma cadeia $\{X_t\}$ de ordem M, com espaço de estados finito, para que as suas cadeias k-derivadas, $k \geq M$, sejam ergódicas. Para isso, utilizamos os conceitos e resultados clássicos da teoria de Cadeias de Markov.

Proposição 1.4. Seja $\{X_t\}$ uma Cadeia de Markov de ordem M e espaço de estados finito A. Se as suas probabilidades de transição são estritamente positivas, isto é,

$$P(a \mid a_1^M) > 0, \forall\, a, a_i \in A, \tag{1.12}$$

então, $\forall\, k \geq M$, a sua cadeia k-derivada $\left\{Y_t^{(k)}\right\}$ é irredutível, aperiódica e recorrente positiva, consequentemente, ergódica.

Demonstração. Sejam $a^{(k)} = a_1^k \in A^k$ e $b^{(k)} = b_1^k \in A^k$, dois elementos quaisquer do espaço de estados de $\left\{Y_t^{(k)}\right\}$. Então, podemos obter

$$\mathbb{P}\left(Y_{k+1}^{(k)} = b^{(k)} \;\middle|\; Y_1^{(k)} = a^{(k)}\right) \geq \mathbb{P}\left(Y_{k+1}^{(k)} = b_1^k, Y_k^{(k)} = a_k b_1^{k-1}, \ldots, Y_2^{(k)} = a_2^k b_1 \;\middle|\; Y_1^{(k)} = a_1^k\right).$$

Por outro lado pela Proposição 1.3, $\left\{Y_t^{(k)}\right\}$ é uma Cadeia de Markov de ordem 1, segue que

$$\mathbb{P}\left(Y_{k+1}^{(k)} = b^{(k)} \;\middle|\; Y_1^{(k)} = a^{(k)}\right) \geq \prod_{j=1}^{k} \mathbb{P}\left(Y_{j+1}^{(k)} = a_{j+1}^k b_1^j \;\middle|\; Y_j^{(k)} = a_j^k b_1^{j-1}\right),$$

em que convencionamos $a_{k+1}^k b_1^k = b_1^k$, $a_k^1 b_1^0 = a_1^k$, $a_k^k = a_k$ e $b_1^1 = b_1$.

Agora, por (1.10), para $k > M$, temos

$$\mathbb{P}\left(Y_{k+1}^{(k)} = b^{(k)} \;\middle|\; Y_1^{(k)} = a^{(k)}\right) \geq \prod_{j=1}^{M} P\left(b_j \mid a_{k-M+j}^k b_1^{j-1}\right) \cdot \prod_{j=M+1}^{k} P\left(b_j \mid b_{j-M}^{j-1}\right)$$

e, usando (1.11), para $k = M$, temos

$$\mathbb{P}\left(Y_{M+1}^{(M)} = b^{(M)} \;\middle|\; Y_1^{(M)} = a^{(M)}\right) \geq \prod_{j=1}^{M} P\left(b_j \mid a_j^M b_1^{j-1}\right).$$

Assim, pela hipótese (1.12), segue que, $\forall\, k \geq M$,

$$\mathbb{P}\left(Y_{k+1}^{(k)} = b^{(k)} \;\middle|\; Y_1^{(k)} = a^{(k)}\right) > 0, \forall\, a^{(k)}, b^{(k)} \in A^k.$$

Ou seja, todos os estados da cadeia $\left\{Y_t^{(k)}\right\}$ comunicam-se entre si e a cadeia é irredutível.

Logo, como seu espaço de estados A^k é finito, pois A é finito, e a cadeia é irredutível, temos que $\left\{Y_t^{(k)}\right\}$ é recorrente positiva.

Por fim, para $(a,a,,\ldots,a) \in A^k$, temos que

$$\mathbb{P}\left(Y_2^{(k)} = (a,a,\ldots,a) \;\middle|\; Y_1^{(k)} = (a,a,\ldots,a)\right) = \mathbb{P}\left(X_{M+1} = a \;\middle|\; X_M = a, \ldots, X_1 = a\right) > 0.$$

Logo, $(a,a,,\ldots,a)$ é aperiódico e, consequentemente, a cadeia $\left\{Y_t^{(k)}\right\}$ é aperiódica.

□

A seguir, apresentamos a definição de Cadeia de Markov de ordem superior irredutível sob a concepção de [4].

Definição 1.5. Uma Cadeia de Markov $\{X_t\}$ de ordem M e espaço de estados finito A é dita *irredutível*, se $\left\{Y_t^{(M)}\right\}$ sua cadeia M-derivada é uma $C.M.$ irredutível, ou seja, para quaisquer $a^{(M)} = a_1^M \in A^M$ e $b^{(M)} = b_1^M \in A^M$ tais que

$$\mathbb{P}\left(Y_t^{(M)} = a^{(M)}\right) > 0, \text{ para algum } t \geq 1 \text{ e } \mathbb{P}\left(Y_t^{(M)} = b^{(M)}\right) > 0, \text{ para algum } t \geq 1,$$

temos que $a^{(M)}$ se comunica com $b^{(M)}$, isto é, existe $t > 1$ tal que

$$\mathbb{P}\left(Y_t^{(M)} = b^{(M)} \mid Y_1^M = a^{(M)}\right) > 0.$$

Pela Proposição 1.4, é direto observar que, se as probabilidades de transição da cadeia são estritamente positivas, então a cadeia é irredutível.

A seguir, apresentamos dois resultados adaptados de outros trabalhos que tratam sobre a relação entre as as probabilidades de transição $P(a \mid a_1^M)$ da cadeia e seus respectivos estimadores. O primeiro é uma adaptação do Corolário 2 de [4] e o segundo é uma adaptação do Lema 2 de [7].

Denotamos, $\forall\, F \in \mathcal{F}$, a indicadora do evento F, $I(F)$, definida por

$$I(F)(\omega) = \begin{cases} 1, \text{ se } \omega \in F \\ 0, \text{ se } \omega \notin F \end{cases}$$

Proposição 1.5. Sejam $\{X_t\}$ uma Cadeia de Markov de ordem M, irredutível, com espaço de estados finito A, e x_1^n, n valores amostrados da cadeia, com $n > k$. Então, dado $\varepsilon > 0$, existe $\alpha > 0$ dependendo das probabilidades de transição da cadeia, $P(\,\cdot \mid \cdot)$, tal que, quase certamente quando $n \to \infty$,

$$\left| \frac{\sum_{t=1}^{n-k+1} I(X_t^{t+k-1} = a_1^k)}{\sum_{t=1}^{n-k+1} I(X_t^{t+k-2} = a_1^{k-1})} - P(a_k \mid a_{k-M}^{k-1}) \right| \leq \sqrt{\frac{\varepsilon \cdot \ln\left(\sum_{t=1}^{n-k+1} I(X_t^{t+k-2} = a_1^{k-1})\right)}{\sum_{t=1}^{n-k+1} I(X_t^{t+k-2} = a_1^{k-1})}}, \quad (1.13)$$

para todo k tal que $M < k \leq \alpha \ln(n)$ e para aqueles $a_1^k \in A^k$ ocorrendo em uma amostra suficientemente grande.

Demonstração. Ver o Corolário 2 de [4].

□

Lema 1.1. Sejam $\{X_t\}$ uma Cadeia de Markov de ordem M, com espaço de estados finito A e x_1^n, n valores amostrados, com $n > k \geq M$ e $a_1^{k+1} \in A^{k+1}$. Se $\left\{Y_t^{(M)}\right\}$, a cadeia M-derivada de $\{X_t\}$, é érgódica e π é sua distribuição estacionária, então quase certamente

$$\limsup_{n\to\infty} \frac{\left(\sum_{t=1}^{n-k} I(X_t^{k+t-1} = a_1^k, X_{k+t} = a_{k+1}) - P(a_{k+1} \mid a_{k-M+1}^k)\left(\sum_{t=1}^{n-k+1} I(X_t^{k+t-1} = a_1^k)\right)\right)^2}{n \cdot \ln(\ln(n))} = 2\pi\left(a_1^{k+1}\right)\left(1 - P(a_{k+1} \mid a_{k-M+1}^k)\right),$$

onde denotamos $\pi\left(a_1^{k+1}\right) = \pi(a_1^M) \cdot P(a_{M+1} \mid a_1^M) \cdots P(a^{k+1} \mid a_{k-M+1}^k)$.

Demonstração. Ver o Lema 2 de [7]. □

Neste trabalho, concentramos nosso estudo ao caso em que a Cadeia de Markov de ordem superior é estacionária, conforme a definição a seguir.

Definição 1.6. Uma processo estocástico $\{X_t\}$ é chamado *estacionário* se, para cada $i \geq 0$ e $m \geq 1$ inteiros e quaisquer índices $1 \leq t_1 < t_2 < \cdots < t_m$, os vetores $(X_{i+t_1}, X_{i+t_2}, \ldots, X_{i+t_m})$ e $(X_{t_1}, X_{t_2}, \ldots, X_{t_m})$ possuem a mesma distribuição conjunta.

Observação 1.2. Seja $\{X_t\}$ Cadeia de Markov de ordem $M \in \mathbb{N}$, estacionária e $\left\{Y_t^{(k)}\right\}$ sua Cadeia de Markov k-derivada com $k \geq M$.

(a) Segue diretamente da Definição 1.6:

(a.1) $\forall\, n, p \geq 1$, X_n e X_p são igualmente distribuídos;

(a.2) $\forall\, n \geq 1$, temos que $Y_n^{(k)}$ tem a mesma distribuição que $Y_1^{(k)}$.
Em particular, para $k = M$, $Y_n^{(M)} = (X_n, \ldots, X_{n+M-1})$ tem a mesma distribuição de $Y_1^{(M)} = (X_1, \ldots, X_M)$.

(b) Neste caso, por facilidade, vamos adotar as seguintes notações simplificadas:

(b.1) $S = A^M$, o espaço de estados da C.M. $\left\{Y_t^{(M)}\right\}$ M-derivada de $\{X_t\}$;

(b.2) $\forall\, a \in A$, $s \in S$ e $t \geq M+1$,

$$P(a \mid s) = \mathbb{P}\left(X_t = a \mid X_{t-M}^{t-1} = s\right) = \mathbb{P}\left(X_{M+1} = a \mid X_1^M = s\right), \tag{1.14}$$

$$P(s,a) = \mathbb{P}\left(X_{t-M}^{t-1} = s, X_t = a\right) = \mathbb{P}\left(X_1^M = s, X_{M+1} = a\right), \tag{1.15}$$

$$P(s) = \mathbb{P}\left(X_{t-M}^{t-1} = s\right) = \mathbb{P}\left(X_1^M = s\right) = \mathbb{P}\left(Y_t^{(M)} = s\right); \tag{1.16}$$

(b.3) Se $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ é uma partição arbitrária de S, então denotamos, $\forall\, L \in \mathcal{L}$, $\forall\, a \in A$,

$$P(L,a) = \sum_{s \in L} P(s,a), \tag{1.17}$$

$$P(L) = \sum_{s \in L} P(s), \tag{1.18}$$

$$P(a \mid L) = \frac{P(L,a)}{P(L)}, \text{ se } P(L) > 0. \tag{1.19}$$

$\Diamond$

Finalizamos esta seção apresentando um resultado básico sobre convergência de estimadores em uma C.M.

Lema 1.2. Seja $\{Y_t\}$, uma C.M. irredutível, aperiódica e recorrente positiva, com espaço de estados finito S e $s \in S$, então

$$\lim_{n\to\infty} \frac{\sum_{t=1}^{n} I(Y_t = s)}{n} = \pi(s), \text{ quase certamente,}$$

onde $\pi(s)$ é a distribuição estacionária da cadeia.

Em particular, se $\{Y_t\}$ é estacionária,

$$\lim_{n\to\infty} \frac{\sum_{t=1}^{n} I(Y_t = s)}{n} = \mathbb{P}(Y_1 = s), \text{ quase certamente.}$$

Demonstração. Para a primeira parte da prova, ver o Corolário 6 do Capítulo 2 de [16].

No caso em que $\{Y_t\}$ é estacionária, pela Definição 1.6, em particular, a distribuição de Y_t independe de t. Quando tal condição é satisfeita, a distribuição inicial da cadeia é a distribuição estacionária (ver [16], por exemplo), logo $\pi(s) = \mathbb{P}(Y_1 = s)$, concluindo a demonstração.

$\square$

1.3 Cadeias de Markov com Partição

Nesta seção, baseados no artigo de García e González-López [13], introduzimos o conceito de Cadeias de Markov com Partição e apresentamos alguns resultados preliminares que utilizamos nos próximos capítulos.

As referências bibliográficas desta seção são: [2], [13] e [15].

1.3.1 Definição e Notações

Daqui para frente, consideramos somente as Cadeias de Markov de ordem M que são estacionárias e utilizamos as notações introduzidas na seção anterior.

Assim, considere $\{X_t\}$ uma Cadeia de Markov de ordem M ($M \in \mathbb{N}$), estacionária e com espaço de estados finito A.

Neste caso, pela Proposição 1.3 e a Observação 1.2, a cadeia M-derivada $\left\{Y_t^{(M)}\right\}$ é uma C.M. homogênea no tempo, com espaço de estados finito $S = A^M$, tal que as variáveis aleatórias $Y_t^{(M)}$, $t \geq 1$, são identicamente distribuídas com distribuição $P(s)$, $s \in S$, dada em (1.16).

Se, além disso, as probabilidades de transição forem estritamente positivas, ou seja, $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S$, temos que a C.M. M-derivada $\left\{Y_t^{(M)}\right\}$ é irredutível, aperiódica e recorrente positiva, pela Proposição 1.4.

Na próxima proposição, seguindo [13], introduzimos uma relação de equivalência sobre o espaço de estados $S = A^M$ da cadeia M-derivada $\left\{Y_t^{(M)}\right\}$ sobre a qual está baseado o conceito de Cadeia de Markov com Partição, que definimos logo em seguida.

Proposição 1.6. Seja $\{X_t\}$, uma Cadeia de Markov de ordem M, estacionária, com espaço de estados finito A, tal que $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S$. A relação $\sim_p$ definida por

$$s \sim_p r \iff P(a \mid s) = P(a \mid r),\ \forall\, a \in A \tag{1.20}$$

é uma relação de equivalência sobre o espaço de estados $S = A^M$ da cadeia M-derivada $\left\{Y_t^{(M)}\right\}$.

Demonstração. É imediato que a relação (1.20) é reflexiva e simétrica.

Agora, sejam s, r, $q \in S$, tais que $s \sim_p r$ e $r \sim_p q$. Então, por (1.20) segue que

$$P(a \mid s) = P(a \mid r) \text{ e } P(a \mid r) = P(a \mid q),\ \forall\, a \in A.$$

Logo, $P(a \mid s) = P(a \mid q)$, $\forall\, a \in A$, ou seja, $s \sim_p q$.

Assim, a relação $\sim_p$ é também transitiva.

□

Observe que a relação $\sim_p$ particiona o espaço de estados $S = A^M$ em classes de equivalência disjuntas. Desta forma, podemos definir:

Definição 1.7. Dada $\{X_t\}$, uma Cadeia de Markov de ordem M, estacionária, com espaço de estados finito A e probabilidades de transição $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S$, dizemos que $\{X_t\}$ é uma *Cadeia de Markov com Partição* $\mathcal{L}^* = \{L_1^*, L_2^*, \ldots, L_{|\mathcal{L}^*|}^*\}$, se $\mathcal{L}^*$ é a partição de S determinada pela relação de equivalência $\sim_p$, definida em (1.20). Usualmente $\mathcal{L}^*$ é denominada a *Partição Mínima* da cadeia.

Note que neste caso, temos

$$s, r \in L_i^* \implies P(a \mid s) = P(a \mid r), \forall\, a \in A$$

e

$$s \in L_i^* \text{ e } r \in L_j^*, i \neq j \implies \exists\, a \in A \text{ tal que } P(a \mid s) \neq P(a \mid r).$$

Como motivação para o uso do Modelo de Markov com Partição, García e González-Lopéz, destacam a redução do número de parâmetros necessários para especificar uma Cadeia de Markov de ordem M e, consequentemente, a diminuição do número de parâmetros a serem estimados.

Especificamente, para uma Cadeia de Markov $\{X_t\}$ de ordem M, os parâmetros a serem estimados, considerando seu modelo completo, são suas probabilidades de transição. Tais probabilidades podem ser agregadas na matriz estocástica reduzida de sua cadeia M-derivada:

$$\begin{array}{c} \\ s_1 \\ s_2 \\ \vdots \\ s_{|S|} \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & \cdots & a_{|A|} \end{array} \\ \begin{pmatrix} P(a_1 \mid s_1) & P(a_2 \mid s_1) & \cdots & P(a_{|A|} \mid s_1) \\ P(a_1 \mid s_2) & P(a_2 \mid s_2) & \cdots & P(a_{|A|} \mid s_2) \\ \vdots & \vdots & \cdots & \vdots \\ P(a_1 \mid s_{|S|}) & P(a_2 \mid s_{|S|}) & \cdots & P(a_{|A|} \mid s_{|S|}) \end{pmatrix}. \end{array} \tag{1.21}$$

Por ser uma matriz estocástica, temos que $\sum_{j=1}^{|A|} P(a_j \mid s_i) = 1$, $\forall\ 1 \leq i \leq |S| = |A|^M$. Desse modo, basta estimar as $(|A|-1)$ probabilidades de cada linha, $\hat{P}(a_j \mid s_i)$, visto que o j_0-ésimo elemento restante da linha será naturalmente estimado por $1 - \sum_{j \neq j_0}^{|A|} \hat{P}(a_j \mid s_i)$.

Assim, ao todo, considerando o modelo completo, temos um total de $|A|^M \cdot (|A|-1)$ parâmetros livres. Uma quantidade que cresce exponencialmente com a ordem M da cadeia, o que dificulta a estimação em cadeias de ordem muito grande.

Entretanto, é possível que tenhamos elementos em S, equivalentes com respeito à relação de equivalência $\sim_p$ definida em (1.20). Sendo esse o caso, teríamos que as linhas na matriz estocástica referentes a tais elementos seriam iguais e bastaria estimar as $(|A|-1)$ probabilidades de transição de uma dessas linhas.

Logo, no caso do Modelo de Markov com Partição $\mathcal{L}^* = \{L_1^*, L_2^*, \ldots, L_{|\mathcal{L}^*|}^*\}$, a matriz estocástica da cadeia poderia ser sintetizada, reduzindo linhas redundantes, da seguinte forma:

$$\begin{array}{c} \\ L_1^* \\ L_2^* \\ \vdots \\ L_{|\mathcal{L}^*|}^* \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & \cdots & a_{|A|} \end{array} \\ \begin{pmatrix} P(a_1 \mid L_1^*) & P(a_2 \mid L_1^*) & \cdots & P(a_{|A|} \mid L_1^*) \\ P(a_1 \mid L_2^*) & P(a_2 \mid L_2^*) & \cdots & P(a_{|A|} \mid L_2^*) \\ \vdots & \vdots & \cdots & \vdots \\ P(a_1 \mid L_{|\mathcal{L}^*|}^*) & P(a_2 \mid L_{|\mathcal{L}^*|}^*) & \cdots & P(a_{|A|} \mid L_{|\mathcal{L}^*|}^*) \end{pmatrix} \end{array}. \tag{1.22}$$

Seguindo o raciocínio anterior, o número de parâmetros livres a serem estimados totaliza $|\mathcal{L}^*|\cdot(|A|-1)$. Como $\mathcal{L}^*$ é uma partição de S, então $|\mathcal{L}^*| \leq |S| = |A|^M$. Logo temos, em geral, uma redução do número de parâmetros a serem estimados.

Note também que, neste caso, os elementos de cada linha da matriz (1.22) são tais que para cada parte $L^* \in \mathcal{L}^*$, se $s \in L^*$, então

$$P(a \mid s) = P(a \mid L^*), \forall\, a \in A. \tag{1.23}$$

De fato, se $s \in L^*$, então para todo $a \in A$,

$$P(a \mid s) = P(a \mid s_i), \forall\, i = 1, 2, \ldots, |L^*|,$$

e por (1.14) a (1.19), podemos escrever

$$\frac{P(s,a)}{P(s)} = \frac{P(s_i,a)}{P(s_i)}, \forall\, i = 1, 2, \ldots, |L^*|,$$

o que implica que

$$P(a \mid L^*) = \frac{\sum_{i=1}^{|L^*|} \left[\frac{P(s,a)P(s_i)}{P(s)}\right]}{\sum_{i=1}^{|L^*|} P(s_i)} = P(a \mid s), \forall\, a \in A \text{ e } \forall\, s \in L^*.$$

Dessa forma, considerando-se o número de parâmetros livres a serem estimados para o modelo completo, teremos sempre um número total fixo de $|A|^M \cdot (|A|-1)$ parâmetros, enquanto, no Modelo com Partição, tal quantidade depende da quantidade de elementos da partição associada, apresentando, assim, uma estrutura mais flexível.

O Modelo de Markov com Partição não é o único que apresenta essas vantagens em comparaçao ao modelo completo. Outra classe de modelos é a das Cadeias de Markov de Alcance Variável (VLMC, sigla de seu nome em inglês "Variable Length Markov Chains").

Tal modelo, detalhado em [2], por exemplo, também considera Cadeias de Markov de ordem finita. Entretanto, a ordem da cadeia varia conforme o que se chama de "contexto", isto é, a depender dos últimos estados da cadeia, que podem apresentar tamanhos diferentes. Essa noção é matematicamente sintetizada pelo conjunto $\mathcal{T}$, denominado *árvore de contexto*.

Especificamente, $\mathcal{T}$ é um conjunto composto por sequências de elementos do espaço de estados A da cadeia, com comprimentos variáveis.

No caso da ordem da Cadeia de Markov ser finita, temos $|\mathcal{T}|< \infty$ e podemos definir $M = max\{\ell(\tau) \mid \tau \in \mathcal{T}\}$, onde ℓ é a função comprimento. Nesse caso, $\mathcal{T}$ caracteriza-se pelo fato de que qualquer sequência $a_1^n = s$ de elementos de A com comprimento maior ou igual a M possui algum elemento de $\mathcal{T}$ como seu sufixo de maior tamanho possível. Tal elemento, denotado por $c(s)$, é denominado o *contexto* da sequência.

Nesse modelo, as probabilidades de transição de Cadeias de Markov de ordem finita dependem, não de sequências de elementos de A com um tamanho fixo, mas dos contextos presentes em $\mathcal{T}$, que têm tamanhos variáveis, isto é, seja $\ell_s = \ell\Big(c(s)\Big)$, as probabilidades de transição são

$$\mathbb{P}\Big(X_{n+1} = a \mid X_1^n = s\Big) = \mathbb{P}\Big(X_{n+1} = a \mid X_{n-\ell_s+1}^n = c(s)\Big) = \mathbb{P}\Big(X_{\ell_s+1} = a \mid X_1^{\ell_s} = c(s)\Big) \coloneqq P^{\mathcal{T}}\Big(a \mid c(s)\Big).$$

Assim, seguindo a lógica apresentada anteriormente, o número de parâmetros a serem estimados seria dado por $|\mathcal{T}|\cdot(|A|-1)$.

O exemplo a seguir indica, de forma simplificada, que modelos VLMC de ordem finita podem ser apropriadamente identificados com o modelo de Cadeia de Markov com Partição introduzido na Definição 1.7.

Exemplo 1.1. Seja $\{X_t\}$, uma Cadeia de Markov de ordem finita, com espaço de estados $A = \{0,1,2\}$, estacionária, com probabilidades de transição estritamente positivas e árvore de contexto $\mathcal{T} = \{0,1,01,02,12,22\}$. Note que $M = max\{\ell(t) \mid t \in \mathcal{T}\} = 2$.

Neste caso, para identificar tal cadeia com o Modelo de Markov com Partição observe que a cadeia é de ordem $M = 2$.

De fato, seja $n \geq 2$ e a_1^n, uma sequência qualquer de elementos de A, então $c\left(a_1^n\right) = c\left(a_{n-1}^n\right)$, já que ambas as sequências, a_1^n e a_{n-1}^n, possuem o mesmo sufixo de maior tamanho possível em $\mathcal{T}$. Desse modo, temos

$$\mathbb{P}\Big(X_{n+1} = a \mid X_1^n = a_1^n\Big) = P^{\mathcal{T}}\Big(a \mid c(a_1^n)\Big) = P^{\mathcal{T}}\Big(a \mid c(a_{n-1}^n)\Big) = \mathbb{P}\Big(X_3 = a \mid X_1^2 = a_{n-1}^n\Big),$$

ou seja, a ordem da cadeia é $M = 2$.

Tomemos, inicialmente, a partição de $A^M = A^2$ que contenha elementos com mesmo contexto, isto é, $\mathcal{L} = \{L_0, L_1, L_{01}, L_{02}, L_{12}, L_{22}\}$, onde definimos $L_\tau = \{s \in A^M \mid c(s) = \tau\}$.

Assim, temos: $L_0 = \{00, 10, 20\}$, $L_1 = \{11, 21\}$, $L_{01} = \{01\}$, $L_{02} = \{02\}$, $L_{12} = \{12\}$ e $L_{22} = \{22\}$.

Agora, observe que $s \sim_p r$, $\forall\ s, r \in L_\tau$, e $\forall\ \tau \in \mathcal{T}$, pois as probabilidades de transição de sequências com o mesmo contexto são iguais. Além disso, para $s_\tau \in L_\tau$ e $s_{\tau'} \in L_{\tau'}$, temos duas possibilidades:

(i) $s_\tau \nsim_p s_{\tau'}$, $\forall\ \tau \neq \tau' \in \mathcal{T}$ e neste caso, $\mathcal{L} = \mathcal{L}'$ é a partição definida pela relação de equivalência $\sim_p$ dada em (1.20);

(ii) existem partes distintas de $\mathcal{L}$ com elementos equivalentes entre si, isto é, elementos que possuem as mesmas probabilidades de transição. Neste caso, a partição $\mathcal{L}^*$ é dada pela união de todas as partes de $\mathcal{L}$ que possuem elementos com as mesmas probabilidades de transição entre si, satisfazendo a Definição 1.7.

✓

A seguir, nos moldes do Exemplo 1.1, podemos notar a redução do número de parâmetros livres que a concepção do Modelo de Markov com Partição pode gerar para a estimação, em comparação com a concepção do modelo completo, ou de VLCM.

Exemplo 1.2. Seja $\{X_t\}$, uma Cadeia de Markov de ordem 2, com espaço de estados $A = \{0, 1, 2\}$, estacionária, com as seguintes probabilidades de transição:

Tabela 1.1 Probabilidades de Transição

s	00	10	20	01	11	21	02	12	22
$P(0 \mid s)$	0.4	0.4	0.4	0.3	0.2	0.2	0.4	0.3	0.2
$P(1 \mid s)$	0.3	0.3	0.3	0.3	0.2	0.2	0.3	0.3	0.2
$P(2 \mid s)$	0.3	0.3	0.3	0.4	0.6	0.6	0.3	0.4	0.6

Considerando o modelo de Cadeia de Markov completa temos um total de parâmetros a serem estimados igual a $|A|^M \cdot (|A|-1) = 3^2 \cdot (3-1) = 18$.

Entretanto, podemos identificar a cadeia pelo modelo VLMC com a árvore de contexto $\mathcal{T} = \{0, 1, 01, 02, 12, 22\}$, apresentada no Exemplo 1.1. Considerando esse modelo, o número total de parâmetros livres se reduz a $|\mathcal{T}| \cdot (|A|-1) = 6 \cdot (3-1) = 12$.

Por fim, analisando as probabilidades de transição $P(\ \cdot \mid s)$, observamos as seguintes equivalências entre os estados em A^2:

$$\begin{aligned}
&00 \sim_p 10 \sim_p 20 \sim_p 02;\\
&01 \sim_p 12;\\
&11 \sim_p 21 \sim_p 22;\\
&00 \nsim_p 01 \nsim_p 11 \nsim_p 00.
\end{aligned}$$

Desse modo, pela Definição 1.7, $\{X_t\}$ é uma Cadeia de Markov com Partição $\mathcal{L}^* = \{\overline{00}, \overline{01}, \overline{11}\} = \{L_1^*, L_2^*, L_3^*\}$, onde $L_1^* = \overline{01} = \{00, 10, 20, 02\}$, $L_2^* = \overline{01} = \{01, 12\}$ e $L_3^* = \overline{11} = \{11, 21, 22\}$. Nesse modelo, a quantidade de parâmetros a serem estimados diminui para $|\mathcal{L}^*| \cdot (|A| - 1) = 3 \cdot (3 - 1) = 6$. ✓

Com as motivações e vantagens do uso do modelo de Cadeias de Markov com Partição apresentadas, concentramos nosso interesse no problema de estimar a Partição Mínima $\mathcal{L}^*$ do modelo, o qual será abordado nos próximos capítulos.

Para isso, vamos estabelecer mais uma lista de notações a serem utilizadas, por simplicidade.

Observação 1.3. Considere $\{X_t\}$, uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária, tal que $P(a \mid s) > 0$, $\forall\ a \in A$ e $\forall\ s \in S = A^M$ e seja $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$, uma partição arbitrária de S.

(a) Dada uma amostra da cadeia, de tamanho $n > M$, $X_1^n = (X_1, \ldots, X_n)$, denotamos

$$N_n(s, a) = \sum_{t=1}^{n-M} I(X_t^{M+t-1} = s, X_{M+t} = a),$$

o número de ocorrências, na amostra, da sequência s sucedida pelo elemento a;

$$N_n(s) = \sum_{t=1}^{n-M} I(X_t^{M+t-1} = s),$$

o número de ocorrências, na amostra, da sequência s sucedida por qualquer elemento de A;

$$N_n(L, a) = \sum_{s \in L} N_n(s, a),$$

o número de ocorrências, na amostra, de sequências da parte $L \in \mathcal{L}$ sucedidas pelo elemento a;

$$N_n(L) = \sum_{s \in L} N_n(s),$$

o número de ocorrências, na amostra, de sequências da parte $L \in \mathcal{L}$ sucedidas por qualquer elemento de A.

(b) Para todos $1 \leq i < j \leq |\mathcal{L}|$, denotamos

$$\mathcal{L}^{ij} = \{L_1, \ldots, L_{i-1}, L_{ij}, L_{i+1}, \ldots, L_{j-1}, L_{j+1}, \ldots, L_{|\mathcal{L}|}\},$$

onde $L_{ij} = L_i \cup L_j$.

Note que como $L_i \cap L_j = \emptyset$, então para todo $a \in A$, temos

$$P(L_{ij}, a) = \sum_{s \in L_{ij}} P(s,a) = \sum_{s \in L_i} P(s,a) + \sum_{s \in L_j} P(s,a) = P(L_i, a) + P(L_j, a)$$

e, de maneira análoga, podemos obter

$$\begin{aligned} P(L_{ij}) &= P(L_i) + P(L_j) \\ N_n(L_{ij}, a) &= N_n(L_i, a) + N_n(L_j, a) \\ N_n(L_{ij}) &= N_n(L_i) + N_n(L_j). \end{aligned}$$

(c) Considerando, de maneira mais geral, se $T \subset \{1, 2, \ldots, |\mathcal{L}|\}$ é um conjunto de índices, denotamos $\mathcal{L}^T$, a partição que une as partes de $\mathcal{L}$ com índices em T, isto é,

$$L_T = \bigcup_{k \in T} L_k.$$

Neste caso, seguindo o mesmo raciocínio de (b), podemos obter

$$\begin{aligned} P(L_T, a) &= \sum_{k \in T} P(L_k, a) \\ P(L_T) &= \sum_{k \in T} P(L_k) \\ N_n(L_T, a) &= \sum_{k \in T} N_n(L_k, a) \\ N_n(L_T) &= \sum_{k \in T} N_n(L_k). \end{aligned}$$

$\Diamond$

Finalizamos esta seção, provando a consistência forte, isto é, a convergência quase certa, dos estimadores para $P(L)$ e $P(L,a)$, descritos em [13], apresentando os estimadores de máxima verossimilhança de $P(a \mid L)$ para cada L uma parte de uma partição arbitrária $\mathcal{L}$ e $a \in A$ e deduzimos a expressão da função de Máxima Verossimilhança modificada

1.3.2 Máxima Verossimilhança Modificada

Com as notações introduzidas na Observação 1.3 (a) e na Observação 1.2, provamos na sequência a consistência forte dos estimadores para $P(L)$ e $P(a,L)$, apresentados em [13], para cadeias estacionárias.

Proposição 1.7. Seja $\{X_t\}$ uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária, tal que $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S = A^M$ e sejam x_1^n, n valores amostrados, com $n > M$. Se $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ é uma partição de S, então para $L \in \mathcal{L}$ e $a \in A$, temos

$$\lim_{n\to\infty} \frac{N_n(L)}{n} = P(L) \text{ e } \lim_{n\to\infty} \frac{N_n(L,a)}{n} = P(L,a), \text{ quase certamente.}$$

Demonstração. Observe que $N_n(s) = \sum_{t=1}^{n-M} I(X_t^{M+t-1} = s) = \sum_{t=1}^{n-M} I\left(Y_t^{(M)} = s\right)$, onde $\left\{Y_t^{(M)}\right\}$ é a Cadeia de Markov M-derivada de $\{X_t\}$.

Desse modo, temos, $\forall\, s \in S = A^M$,

$$\lim_{n\to\infty} \frac{N_n(s)}{n} = \lim_{n\to\infty} \frac{\sum_{t=1}^{n-M} I\left(Y_t^{(M)} = s\right)}{n} = \lim_{n\to\infty} \frac{\sum_{t=1}^{n-M} I\left(Y_t^{(M)} = s\right)}{n-M} \cdot \frac{n-M}{n}.$$

Mas, pela Proposição 1.4, $\left\{Y_t^{(M)}\right\}$ é uma C.M. irredutível, aperiódica e recorrente positiva e da Observação 1.2 temos que, $\forall\, t \geq 1$, $\mathbb{P}(Y_t^{(M)} = s) = P(s)$, $\forall\, s \in S$. Assim, pelo Lema 1.2, segue que

$$\lim_{n\to\infty} \frac{\sum_{t=1}^{n-M} I\left(Y_t^{(M)} = s\right)}{n-M} = \lim_{m\to\infty} \frac{\sum_{t=1}^{m} I\left(Y_t^{(M)} = s\right)}{m} = P(s), \text{ quase certamente.}$$

Portanto, para $L \in \mathcal{L}$, como $|L| < \infty$, podemos obter

$$\lim_{n\to\infty} \frac{N_n(L)}{n} = \lim_{n\to\infty} \sum_{s\in L} \frac{N_n(s)}{n} = \sum_{s\in L} \lim_{n\to\infty} \frac{N_n(s)}{n} = \sum_{s\in L} P(s) = P(L),$$

ou seja, $\lim_{n\to\infty} \frac{N_n(L)}{n} = P(L)$, quase certamente.

De forma análoga, podemos provar que $\lim_{n\to\infty} \frac{N_n(L,a)}{n} = P(L,a), \forall\, a \in A$ e $L \in \mathcal{L}$.

□

Considere $\{X_t\}$ uma Cadeia de Markov com Partição $\mathcal{L}^*$, de ordem M, conforme a Definição 1.7.

Seja X_1^n, uma amostra de tamanho $n > M$ da cadeia $\{X_t\}$ e considere as notações introduzidas na Observação 1.3 (a) e na Observação 1.2.

Proposição 1.8. Para cada $L \in \mathcal{L}^*$, com $N_n(L) > 0$, o estimador de máxima verossimilhança de $P(a \mid L)$, para cada $a \in A$, é dado por

$$\hat{P}_n(a \mid L) = \frac{N_n(L,a)}{N_n(L)} \tag{1.24}$$

e a função de máxima verossimilhança modificada baseada em n valores observados x_1^n, da Cadeia de Markov $\{X_t\}$, assumindo uma partição arbitrária $\mathcal{L}$ de S, é dada por

$$MV(\mathcal{L}, x_1^n) = \prod_{a\in A,\ L\in\mathcal{L}} \left(\frac{N_n(L,a)}{N_n(L)}\right)^{N_n(L,a)}, \tag{1.25}$$

onde o produtório é tomado para $L \in \mathcal{L}$ tal que $N_n(L) > 0$.

Demonstração. Primeiramente, como X_1^n é uma amostra da Cadeia de Markov $\{X_t\}$ de ordem M, homogênea no tempo, a função de verossimilhança associada é obtida, usando a regra do produto e as propriedades (i) e (ii) da Definição 1.3, da seguinte forma

$$\begin{aligned}\mathbb{P}(X_1^n = x_1^n) &= \mathbb{P}(X_n = x_n \mid X_{n-M}^{n-1} = x_{n-M}^{n-1}) \cdot P(x_{n-1} \mid x_{n-M-1}^{n-2}) \cdots P(x_{M+1} \mid x_1^M) \cdot P(x_1^M) \\ &= P(x_n \mid x_{n-M}^{n-1}) \cdot P(x_{n-1} \mid x_{n-M-1}^{n-2}) \cdots P(x_{M+1} \mid x_1^M) \cdot P(x_1^M).\end{aligned}$$

Ou seja, para n valores observados x_1^n, temos

$$\mathbb{P}(X_1^n = x_1^n) = P(x_1^M) \cdot \prod_{a\,\in\,A,\ s\,\in\,S} P(a \mid s)^{N_n(s,a)}. \tag{1.26}$$

Agora, considerando que $\{X_t\}$ é uma Cadeia de Markov com Partição $\mathcal{L}^*$, então, por (1.23), temos para cada $L \in \mathcal{L}^*$,

$$P(a \mid s) = P(a \mid s') = P(a \mid L), \forall\, s, s' \in L.$$

Assim, agrupando em (1.26) as probabilidades de transição presentes em uma mesma parte, temos que a função de verossimilhança dos valores observados x_1^n da amostra da Cadeia de

Markov com Partição $\mathcal{L}^*$ é dada por

$$f\Big(P(a \mid L) \mid x_1^n\Big) = P(x_1^M) \cdot \prod_{a \in A,\ L \in \mathcal{L}^*} P(a \mid L)^{\sum_{s \in L} N_n(s,a)} = P(x_1^M) \cdot \prod_{a \in A,\ L \in \mathcal{L}^*} P(a \mid L)^{N_n(L,a)}. \tag{1.27}$$

Observe que se $N_n(L) = 0$, então $N_n(L,a) = 0$, $\forall\ a \in A$. Assim, (1.27) reduz-se a

$$f\Big(P(a \mid L) \mid x_1^n\Big) = P(x_1^M) \cdot \prod_{a \in A,\ L \in \mathcal{L}^*} P(a \mid L)^{N_n(L,a)}, \tag{1.28}$$

em que o produtório é considerado sobre $L \in \mathcal{L}^*$ tais que $N_n(L) > 0$.

Como $\ln(\cdot)$ é uma função crescente, os estimadores $\hat{P}_n(a \mid L)$ que maximizam a função $f\Big(P(a \mid L) \mid x_1^n\Big)$ são os mesmos que maximizam a função log-verossimilhança que é dada por

$$\ln\left(f\Big(P(a \mid L) \mid x_1^n\Big)\right) = \ln\Big(P(x_1^M)\Big) + \sum_{a,\ L} \ln\Big(P(a \mid L)\Big) \cdot N_n(L,a), \tag{1.29}$$

para $L \in \mathcal{L}^*$ tal que $N_n(L) > 0$.

Agora, suponha que $A = \{a_1, ...a_{|A|}\}$. Se $N_n(L_0) > 0$, para algum $L_0 \in \mathcal{L}^*$, então existe algum $a_{L_0} \in A$ tal que $N_n(L_0, a_{L_0}) > 0$. Lembrando que $\sum_{a \in A} P(a \mid L) = 1$, $\forall\ L \in \mathcal{L}^*$, então $P(a_{L_0} \mid L) = 1 - \sum_{a \neq a_{L_0}} P(a \mid L)$.

Assim, podemos reescrever (1.29) como

$$\begin{aligned}
\ln\left(f\Big(P(a \mid L) \mid x_1^n\Big)\right) &= \ln\Big(P(x_1^M)\Big) + \sum_{a \neq a_{L_0},\ L} \ln\Big(P(a \mid L)\Big) N_n(L,a) + \sum_{L} \ln\Big(P(a_{L_0} \mid L\Big) N_n(L, a_{L_0}) \\
&= \ln\Big(P(x_1^M)\Big) + \sum_{a \neq a_{L_0},\ L} \ln\Big(P(a \mid L)\Big) N_n(L,a) + \sum_{L} \ln\Big(1 - \sum_{a \neq a_{L_0}} P(a \mid L)\Big) N_n(L, a_{L_0}).
\end{aligned}$$

Derivando parcialmente a expressão acima em relação a $P(a \mid L_0)$ e igualando a 0, para cada $a \neq a_{L_0}$, obtemos a seguinte relação:

$$\frac{N_n(L_0,a)}{\hat{P}_n(a \mid L_0)} - \frac{N_n(L_0, a_{L_0})}{1 - \sum\limits_{a \neq a_{L_0}} \hat{P}_n(a \mid L_0)} = 0.$$

Como, $N_n(L_0, a_{L_0}) > 0$, obtemos que

$$\hat{P}_n(a \mid L_0) = \frac{\left(1 - \sum\limits_{a \neq a_{L_0}} \hat{P}_n(a \mid L_0)\right) \cdot N_n(L_0,a)}{N_n(L_0, a_{L_0})} = \frac{\hat{P}_n(a_{L_0} \mid L_0) \cdot N_n(L_0,a)}{N_n(L_0, a_{L_0})}, \tag{1.30}$$

mas, para que $\sum_{a\in A} \hat{P}_n(a \mid L_0) = 1$, devemos ter por (1.30) que

$$\hat{P}_n(a_{L_0} \mid L_0) = \frac{N_n(L_0, a_{L_0})}{N_n(L_0)}. \tag{1.31}$$

Assim, por (1.30) e (1.31), segue que

$$\hat{P}_n(a \mid L_0) = \frac{N_n(L_0, a)}{N_n(L_0)}, \forall\, a \in A.$$

Repetindo o mesmo processo, $\forall\, L \in \mathcal{L}^*$ tal que $N_n(L) > 0$, obtemos que o estimador de máxima verossimilhança de $P(a \mid L)$ é

$$\hat{P}_n(a \mid L) = \frac{N_n(L, a)}{N_n(L)}, \forall\, a \in A, \forall\, L \in \mathcal{L}^* \text{ com } N_n(L) > 0. \tag{1.32}$$

Assim, a função de máxima verossimilhança modificada da Cadeia de Markov $\{X_t\}$ com Partição $\mathcal{L}^*$, baseada nos n valores observados x_1^n, quando uma partição $\mathcal{L}$ é assumida, é dada por

$$MV(\mathcal{L}, x_1^n) = \prod_{a\in A,\ L\in\mathcal{L}} \left(\frac{N_n(L, a)}{N_n(L)} \right)^{N_n(L,a)}, \tag{1.33}$$

em que o produtório é considerado sobre toda $L \in \mathcal{L}$ tal que $N_n(L) > 0$.

□

Observe que, pela Proposição 1.7, segue que o estimador $\hat{P}_n(a \mid L)$, definido em (1.24), é fortemente consistente.

Concluímos este capítulo, observando que na Seção 1.1, apresentamos conceitos e resultados gerais básicos que serão necessários para o formalismo técnico das demonstrações dos principais teoremas do trabalho.

Na Seção 1.2, apresentamos o conceito de processo k-derivado para permitir a utilização de resultados básicos de Cadeias de Markov de ordem 1 no contexto de Cadeias de Markov de ordem superior M e apresentamos alguns conceitos e resultados já conhecidos de cadeias de ordem superior, que serão importantes para auxiliar no estabelecimento dos principais resultados de estimação do Modelo de Markov com Partição.

Por fim, na Seção 1.3, apresentamos o modelo de Cadeia de Markov com Partição e suas vantagens em relação a outros modelos para Cadeias de Markov de ordem superior. Obtemos os estimadores de máxima verossimilhança dos parâmetros do modelo e a máxima verossimilhança modificada do modelo, que será necessária para definição dos critérios de informação que serão utilizados para estimação da Partição Mínima do modelo nos próximos capítulos.

Capítulo 2

Estimação pelo Critério de Informação Bayesiano

O amplamente conhecido Critério de Informação Bayesiano para a seleção de um modelo, dentre K modelos, $M_1, \ldots, M_K$, que melhor aproxima o modelo verdadeiro, baseia-se na medida de similaridade entre um modelo M_k e o modelo verdadeiro, dada por

$$BIC(k) = \ln\left(MV(k)\right) - \frac{d_k \ln(n)}{2},$$

onde $MV(k)$ é a função de máxima verossimilhança modificada do modelo associada a uma amostra de tamanho n da população e d_k a dimensão do modelo M_k, isto é, o número de parâmetros a serem estimados no modelo.

Dessa forma, seleciona-se o modelo M_{k^*} tal que

$$k^* = \underset{k \in \{1,2,\ldots,K\}}{\operatorname{argmax}} \; BIC(k).$$

No contexto de Cadeias de Markov com Partição, em [13], García e González-López propuseram a utilização do Critério de Informação Bayesiano para construir uma estratégia consistente que se propõe, a partir de n observações da cadeia, a encontrar o modelo, dentre a classe de Modelos de Markov com Partição, que contém um número mínimo de partes para representar a lei de probabilidade do processo.

Neste capítulo, nós apresentamos em detalhes conceitos e resultados estabelecidos por García e González-López [13] que garantem a consistência teórica da estimação da Partição Mínima do Modelo de Markov com Partição.

Assim, na Seção 2.1 apresentamos o critério *BIC* no contexto de Cadeias de Markov com Partição e apresentamos conceitos e resultados auxiliares para a estimação consistente da Partição Mínima da cadeia.

Na Seção 2.2, apresentamos os principais resultados em [13] que demonstram a consistência forte do *BIC* na estimação da Partição Mínima do Modelo de Markov com Partição, isto é, quase certamente quando $n \to \infty$, o critério auxilia de forma efetiva na estimação da Partição Mínima.

2.1 O *BIC* em Cadeias de Markov com Partição

Considere $\{X_t\}$ uma Cadeia de Markov com Partição $\mathcal{L}^*$, conforme a Definição 1.7, de ordem M e espaço de estados A finito. Lembrando que, neste caso, $\{X_t\}$ é estacionária e assumimos que

$$P(a \mid s) > 0, \forall\, a \in A \text{ e } \forall\, s \in S = A^M.$$

Dados n valores observados x_1^n da cadeia e assumindo uma partição $\mathcal{L} = \{L_1, \ldots, L_{|\mathcal{L}|}\}$ arbitrária, dentre a classe $\mathcal{P}$ de todas as partições de S, vimos na Proposição 1.8, que a função de máxima verossimilhança modificada é dada por

$$MV(\mathcal{L}, x_1^n) = \prod_{a \in A,\ L \in \mathcal{L}} \left(\frac{N_n(L,a)}{N_n(L)} \right)^{N_n(L,a)}, \tag{2.1}$$

onde o produtório é tomado para $L \in \mathcal{L}$ tal que $N_n(L) > 0$.

Dessa forma, García e González-López ([13]) definiram o *BIC do Modelo* $\{X_t\}$ *com Partição* $\mathcal{L}$ por

$$BIC(\mathcal{L}, x_1^n) = \ln\left(MV(\mathcal{L}, x_1^n)\right) - \frac{(|A|-1)\cdot|\mathcal{L}|}{2} \cdot \ln(n), \tag{2.2}$$

como medida de similaridade entre o modelo assumindo a partição $\mathcal{L}$ e o modelo verdadeiro com partição $\mathcal{L}^*$.

Note que, neste caso, a constante $(|A|-1)|\mathcal{L}|$ no termo de penalidade de (2.2), refere-se ao número de parâmetros livres a serem estimados.

Utilizando (2.2), García e González-López propuseram, em [13], obter como partição estimadora $\mathcal{L}_n^*$ para $\mathcal{L}^*$ aquela que maximiza (2.2) dentre todas as partições em $\mathcal{P}$. Ou seja,

$$\mathcal{L}_n^* = \underset{\mathcal{L} \in \mathcal{P}}{\operatorname{argmax}} \{BIC(\mathcal{L}, x_1^n)\}. \tag{2.3}$$

No entanto, na prática, encontrar o máximo global (2.3) dentre todas as partições possíveis em $\mathcal{P}$ seria provavelmente impossível, mesmo para um espaço de estados $S = A^M$, da cadeia M-derivada, de tamanho moderado.

Dessa forma, os citados autores propuseram a determinação de uma partição estimadora $\hat{\mathcal{L}}_n^*$ que, quase certamente quando $n \to \infty$, se iguale a $\mathcal{L}^*$ e satisfaça (2.3). A estratégia para obter $\hat{\mathcal{L}}_n^*$, a qual será detalhada na Seção 3.3, é iniciar a busca por uma dada partição inicial, como por exemplo, o próprio espaço S, e então reduzir o tamanho da partição inicial passo a passo, de tal forma a unir determinadas partes de maneira apropriadamente escolhida.

A estratégia proposta baseia-se nos conceitos de *parte boa* e *partição boa*, conforme a definição a seguir.

Definição 2.1. Seja$\{X_t\}$ uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária tal que $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S = A^M$. Se $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$, é uma partição de S, então dizemos que:

(i) $L \in \mathcal{L}$ é uma *parte boa*, se

$$\forall\, s, s' \in L,\ P(a \mid s) = P(\, a \mid s'),\ \forall\, a \in A, \text{ isto é, } s \sim_p s'.$$

(ii) $\mathcal{L}$ é dita *partição boa*, se, $\forall\, L \in \mathcal{L}$, L é uma parte boa.

Exemplo 2.1.

(a) $\mathcal{L} = S$ é uma partição boa de S, pois, $\forall\, s \in \{s\}$, $s \sim_p s$.

Aqui cometemos um pequeno abuso de notação com $\mathcal{L} = S$. No rigor matemático, temos $\mathcal{L} = \{\{s\} \mid s \in S\}$.

(b) A partição $\mathcal{L}^*$ da Definição 1.7 é uma partição boa, pois é a partição gerada por $\sim_p$. Logo, $\forall\, s, s' \in L$, com $L \in \mathcal{L}^*$, $s \sim_p s'$.

(c) A partição $\mathcal{L} = \{L_0, L_1, L_{01}, L_{02}, L_{12}, L_{22}\}$ do Exemplo 1.1 é uma partição boa.

De fato, como os elementos de cada parte possuem o mesmo contexto, apresentam as mesmas probabilidades de transição, ou seja, $\forall\, s, s' \in L_\tau$, com $\tau \in \mathcal{T}$, $s \sim_p s'$.

Logo cada parte $L_\tau \in \mathcal{L}$ é uma parte boa e, consequentemente, $\mathcal{L}$ é uma partição boa.

✓

Observação 2.1. Seja $\mathcal{L} = \{L_1, \ldots, L_{|\mathcal{L}|}\}$, uma partição boa para a Cadeia de Markov $\{X_t\}$ de ordem M, satisfazendo as condições da Definição 2.1.

(a) Repetindo os mesmos argumentos utilizados para obter (1.23), podemos mostrar que, se $L \in \mathcal{L}$, então

$$P(a \mid s) = P(a \mid L), \forall\, s \in L.$$

(b) Pela Observação 1.3 segue que: se $L_i, L_j \in \mathcal{L}$ são tais que $P(\,\cdot \mid L_i) = P(\,\cdot \mid L_j)$, então a partição

$$\mathcal{L}^{ij} = \{L_1, \ldots, L_{i-1}, L_{ij}, L_{i+1}, \ldots, L_{j-1}, L_{j+1}, \ldots, L_{|\mathcal{L}|}\}, \text{ com } L_{ij} = L_i \cup L_j,$$

é também uma partição boa.

Mais ainda, se $P(\,\cdot \mid L_k) = P(\,\cdot \mid L_\ell)$, $\forall\, k, \ell \in T \subset \{1, 2, \ldots, |\mathcal{L}|\}$, com $k \neq \ell$, considere $L_T = \bigcup_{k \in T} L_k$ e $\mathcal{R}_T = \{L_k \mid k \in T\}$. Então a partição

$$\mathcal{L}^T = (\mathcal{L} \setminus \mathcal{R}_T) \cup \{L_T\}$$

que une as partes de $\mathcal{L}$ com índice em T em uma única parte L_T, também é uma partição boa.

$\Diamond$

Dessa forma, a estratégia proposta por García e González-López para obter a partição estimadora $\hat{\mathcal{L}}_n^*$ que, quase certamente quando $n \to \infty$, satisfaz (2.3), se baseia no método de construção de partição a partir de partes boas, com o objetivo de reduzir o tamanho da partição passo a passo.

Especificamente, duas partes boas L_i e L_j de uma mesma partição arbitrária $\mathcal{L}$ (não necessariamente uma partição boa) serão unidas para formar uma nova partição $\mathcal{L}^{ij}$, considerada de acordo com as notações da Observação 2.1, se

$$BIC(\mathcal{L}^{ij}, x_1^n) > BIC(\mathcal{L}, x_1^n), \tag{2.4}$$

para x_1^n valores observados do processo

Para estabelecer a consistência de (2.4) e, consequentemente, do estimador $\mathcal{L}_n^*$ dado em (2.3), o seguinte resultado auxiliar, apresentado em [13], será necessário.

Proposição 2.1. Sejam $\{X_t\}$ uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária e $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S = A^M$, $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$, uma partição de S e x_1^n, n valores observados do processo, com $n > M$.

Se $L \in \mathcal{L}$ é uma parte boa, então, dado $\delta > 0$, temos, quase certamente quando $n \to \infty$,

$$\left| \frac{N_n(L,a)}{N_n(L)} - P(a \mid L) \right| \leq \sqrt{\frac{\delta \cdot \ln(n)}{N_n(L)}}. \tag{2.5}$$

Demonstração. Pela Proposição 1.4, $\{X_t\}$ é irredutível, pois as probabilidades de transição da cadeia são estritamente positivas. Assim, usando a Proposição 1.5, dado $\varepsilon > 0$, existe $\alpha > 0$ dependendo das probabilidades de transição $P(\cdot \mid \cdot)$ da cadeia tal que, quase certamente quando $n \to \infty$,

$$\left| \frac{\sum_{t=1}^{n-k+1} I(X_t^{t+k-1} = a_1^k)}{\sum_{t=1}^{n-k+1} I(X_t^{t+k-2} = a_1^{k-1})} - P(a_k \mid a_{k-M}^{k-1}) \right| \leq \sqrt{\frac{\varepsilon \cdot \ln\left(\sum_{t=1}^{n-k+1} I(X_t^{t+k-2} = a_1^{k-1}) \right)}{\sum_{t=1}^{n-k+1} I(X_t^{t+k-2} = a_1^{k-1})}}, \tag{2.6}$$

para todo k tal que $M < k \leq \alpha \ln(n)$ e para aqueles $a_1^k \in A^k$ ocorrendo em uma amostra suficientemente grande.

Considerando n grande, tal que $\alpha \ln(n) \geq M+1$, podemos tomar $k = M+1$. Assim, para $s = a_1^M \in S$ e $a = a_{M+1}$, um elemento de A, podemos reescrever (2.6) da seguinte forma

$$\left| \frac{\sum_{t=1}^{n-M} I(X_t^{t+M-1} = s, X_{t+M} = a)}{\sum_{t=1}^{n-M} I(X_t^{t+M-1} = s)} - P(a \mid s) \right| \leq \sqrt{\frac{\varepsilon \cdot \ln\left(\sum_{t=1}^{n-M} I(X_t^{t+M-1} = s) \right)}{\sum_{t=1}^{n-M} I(X_t^{t+M-1} = s)}} \tag{2.7}$$

e, consequentemente, temos

$$\left| \frac{N_n(s,a)}{N_n(s)} - P(a \mid s) \right| \leq \sqrt{\frac{\varepsilon \cdot \ln(N_n(s))}{N_n(s)}}. \tag{2.8}$$

Mas, como $P(a \mid s) > 0$, $\forall\ s \in S$ e $a \in A$, então segue que $N_n(s) > 0$ quase certamente, conforme $n \to \infty$ e podemos concluir que (2.8) é válida, $\forall\ s \in S$ e $a \in A$, quase certamente quando $n \to \infty$.

Logo, dado $\delta > 0$, escolhendo $\varepsilon = \dfrac{\delta}{|A|^{2M}}$, por (2.8), temos

$$-\sqrt{\frac{\delta \cdot \ln(N_n(s))}{|A|^{2M} \cdot N_n(s)}} \leq \frac{N_n(s,a)}{N_n(s)} - P(a \mid s) \leq \sqrt{\frac{\delta \cdot \ln(N_n(s))}{|A|^{2M} \cdot N_n(s)}}$$

e, como $N_n(s) > 0$, segue que

$$-\frac{\sqrt{\delta \cdot \ln(N_n(s)) \cdot N_n(s)}}{|A|^M} \leq N_n(s,a) - P(a \mid s) \cdot N_n(s) \leq \frac{\sqrt{\delta \cdot \ln(N_n(s)) \cdot N_n(s)}}{|A|^M}.$$

Pela Observação 2.1 (a), temos que $P(a \mid s) = P(a \mid L)$, $\forall\, s \in L$, pois L é uma parte boa. Então, como $|L| < \infty$, segue que

$$-\sum_{s\in L} \frac{\sqrt{\delta \cdot \ln(N_n(s)) \cdot N_n(s)}}{|A|^M} \leq \sum_{s\in L} N_n(s,a) - \sum_{s\in L} P(a \mid L) \cdot N_n(s) \leq \sum_{s\in L} \frac{\sqrt{\delta \cdot \ln(N_n(s)) \cdot N_n(s)}}{|A|^M},$$

o que implica

$$-\frac{\sqrt{\delta}}{|A|^M} \sum_{s\in L} \sqrt{N_n(s) \cdot \ln(N_n(s))} \leq N_n(L,a) - P(a \mid L) \cdot N_n(L) \leq \frac{\sqrt{\delta}}{|A|^M} \sum_{s\in L} \sqrt{N_n(s) \cdot \ln(N_n(s))}.$$

Mas, como, $N_n(s) \leq n$ e $N_n(s) \leq m = \max\{N_n(s) \mid s \in L\}$, podemos obter

$$-\frac{\sqrt{\delta \cdot \ln(n)} \cdot \sqrt{m} \cdot |L|}{|A|^M} \leq N_n(L,a) - P(a \mid L) \cdot N_n(L) \leq \frac{\sqrt{\delta \cdot \ln(n)} \cdot \sqrt{m} \cdot |L|}{|A|^M}.$$

Agora, temos que $|L| \leq |A|^M$, pois $L \subset S$, e $m = \max\{N_n(s) \mid s \in L\} \leq \sum_{s\in L} N_n(s) = N_n(L)$. Logo, segue que

$$-\sqrt{\delta \cdot \ln(n)} \cdot \sqrt{N_n(L)} \leq N_n(L,a) - P(a \mid L) \cdot N_n(L) \leq \sqrt{\delta \cdot \ln(n)} \cdot \sqrt{N_n(L)}.$$

Portanto, como $N_n(L) > 0$ podemos concluir que, quase certamente quando $n \to \infty$,

$$-\sqrt{\delta \cdot \ln(n)} \cdot \frac{\sqrt{N_n(L)}}{N_n(L)} \leq \frac{N_n(L,a)}{N_n(L)} - P(a \mid L) \leq \sqrt{\delta \cdot \ln(n)} \cdot \frac{\sqrt{N_n(L)}}{N_n(L)},$$

ou seja,

$$\left| \frac{N_n(L,a)}{N_n(L)} - P(a \mid L) \right| \leq \sqrt{\frac{\delta \cdot \ln(n)}{N_n(L)}}.$$

□

2.2 Consistência

Nesta seção, apresentamos em detalhes os resultados obtidos em [13] que estabelecem a consistência forte da utilização do *BIC* para a estimação da Partição Mínima de uma Cadeia de Markov com Partição.

O primeiro resultado mostra que o *BIC* é um critério fortemente consistente para decidir se duas partes boas devem ser unidas, conforme descrito na seção anterior em (2.4).

Teorema 2.1. Seja $\{X_t\}$ uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária, tal que $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S = A^M$ e sejam x_1^n, n valores observados de $\{X_t\}$. Se $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ é uma partição de S, para a qual existem $i < j \in \{1, 2, \ldots, |\mathcal{L}|\}$ tais que L_i e L_j são partes boas, então

$$P(a \mid L_i) = P(a \mid L_j),\ \forall\, a \in A \tag{2.9}$$

se, e somente se, quase certamente quando $n \to \infty$,

$$BIC(\mathcal{L}^{ij}, x_1^n) > BIC(\mathcal{L}, x_1^n). \tag{2.10}$$

Demonstração. Primeiramente, como vamos comparar $BIC(\mathcal{L}^{ij}, x_1^n)$ e $BIC(\mathcal{L}, x_1^n)$ quando $n \to \infty$, podemos considerar n suficientemente grande tal que $N_n(L) > 0$, para toda L parte de S. Desse modo, substituindo (2.1) em (2.2), podemos obter

$$BIC(\mathcal{L}, x_1^n) = \sum_{a \in A,\ L \in \mathcal{L}} N_n(L,a) \ln\left(\frac{N_n(L,a)}{N_n(L)}\right) - \frac{(|A|-1)\cdot|\mathcal{L}|}{2} \cdot \ln(n)$$

e

$$BIC(\mathcal{L}^{ij}, x_1^n) = \sum_{a \in A,\ L \in \mathcal{L}^{ij}} N_n(L,a) \ln\left(\frac{N_n(L,a)}{N_n(L)}\right) - \frac{(|A|-1)\cdot|\mathcal{L}^{ij}|}{2} \cdot \ln(n).$$

Denotando $\mathcal{L}^0 = \mathcal{L} \setminus \{L_i, L_j\}$ e observando que $|\mathcal{L}^{ij}| = |\mathcal{L}| - 1$, podemos reescrever as expressões acima do seguinte modo

$$\begin{aligned} BIC(\mathcal{L}, x_1^n) &= \sum_{a \in A,\ L \in \mathcal{L}^0} N_n(L,a) \ln\left(\frac{N_n(L,a)}{N_n(L)}\right) - \frac{(|A|-1)\cdot|\mathcal{L}|}{2} \cdot \ln(n) \\ &+ \sum_{a \in A} N_n(L_i,a) \ln\left(\frac{N_n(L_i,a)}{N_n(L_i)}\right) + \sum_{a \in A} N_n(L_j,a) \ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right) \end{aligned} \tag{2.11}$$

e

$$BIC(\mathcal{L}^{ij},x_1^n) = \sum_{a\in A,\ L\in\mathcal{L}^0} N_n(L,a)\ln\left(\frac{N_n(L,a)}{N_n(L)}\right) - \frac{(|A|-1)\cdot|\mathcal{L}|}{2}\cdot\ln(n) \\ + \frac{(|A|-1)}{2}\cdot\ln(n) + \sum_{a\in A} N_n(L_{ij},a)\ln\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right). \tag{2.12}$$

Agora, subtraindo (2.12) de (2.11), obtemos

$$BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^{ij},x_1^n) = \sum_{a\in A}\left\{N_n(L_i,a)\ln\left(\frac{N_n(L_i,a)}{N_n(L_i)}\right) + N_n(L_j,a)\ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right)\right. \\ \left. - N_n(L_{ij},a)\ln\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right)\right\} - \frac{(|A|-1)}{2}\cdot\ln(n). \tag{2.13}$$

Por um lado, suponha que (2.9) está satisfeita. Consequentemente, temos que L_{ij} é uma parte boa de S e

$$P(\,\cdot \mid L_{ij}) = P(\,\cdot \mid L_i) = P(\,\cdot \mid L_j). \tag{2.14}$$

Agora, pela Proposição 1.8, $\dfrac{N_n(L_{ij},a)}{N_n(L_{ij})}$ é o estimador de máxima verossimilhança de $P(a \mid L_{ij})$. Logo, temos

$$\prod_{a\in A}\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right)^{N_n(L_{ij},a)} \geq \prod_{a\in A}\Big(P(a \mid L_{ij})\Big)^{N_n(L_{ij},a)}$$

e de (2.13), segue que

$$BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^{ij},x_1^n) \leq \sum_{a\in A}\left\{N_n(L_i,a)\ln\left(\frac{N_n(L_i,a)}{N_n(L_i)}\right) + N_n(L_j,a)\ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right)\right. \\ \left. - N_n(L_{ij},a)\ln\Big(P(a \mid L_{ij})\Big)\right\} - \frac{(|A|-1)}{2}\cdot\ln(n).$$

Como $N_n(L_{ij},a) = N_n(L_i,a) + N_n(L_j,a)$ e $P(a \mid L_{ij}) = P(a \mid L_i) = P(a \mid L_j)$, $\forall\ a \in A$, podemos reescrever a inequação anterior como

$$BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^{ij},x_1^n) \leq \sum_{a\in A}\left\{ N_n(L_i,a)\left(\ln\left(\frac{N_n(L_i,a)}{N_n(L_i)}\right) - \ln\Big(P(a \mid L_i)\Big)\right)\right.$$
$$\left. + N_n(L_j,a)\left(\ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right) - \ln\Big(P(a \mid L_j)\Big)\right)\right\} - \frac{(|A|-1)}{2}\cdot\ln(n)$$

$$= \sum_{a\in A}\left\{ N_n(L_i,a)\ln\left(\frac{\frac{N_n(L_i,a)}{N_n(L_i)}}{P(a \mid L_i)}\right) + N_n(L_j,a)\ln\left(\frac{\frac{N_n(L_j,a)}{N_n(L_j)}}{P(a \mid L_j)}\right)\right\} - \frac{(|A|-1)}{2}\cdot\ln(n)$$

$$= N_n(L_i)\sum_{a\in A}\left\{\frac{N_n(L_i,a)}{N_n(L_i)}\ln\left(\frac{\frac{N_n(L_i,a)}{N_n(L_i)}}{P(a \mid L_i)}\right)\right\} + N_n(L_j)\sum_{a\in A}\left\{\frac{N_n(L_j,a)}{N_n(L_j)}\ln\left(\frac{\frac{N_n(L_j,a)}{N_n(L_j)}}{P(a \mid L_j)}\right)\right\}$$
$$- \frac{(|A|-1)}{2}\cdot\ln(n).$$

Agora, usando o conceito de entropia relativa na Definição 1.2, temos que

$$BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^{ij},x_1^n) \leq N_n(L_i)D\left(\frac{N_n(L_i,\cdot)}{N_n(L_i)}\middle|\middle|P(\cdot \mid L_i)\right) + N_n(L_j)D\left(\frac{N_n(L_j,\cdot)}{N_n(L_j)}\middle|\middle|P(\cdot \mid L_j)\right)$$
$$- \frac{(|A|-1)}{2}\cdot\ln(n). \quad (2.15)$$

Mas, por (1.6) da Proposição 1.2 e pela Proposição 2.1, dado $\delta > 0$, temos, quase certamente quando $n \to \infty$, que

$$D\left(\frac{N_n(L_i,\cdot)}{N_n(L_i)}\middle|\middle|P(\cdot \mid L_i)\right) \leq \sum_{a\in A}\frac{\left(\frac{N_n(L_i,a)}{N_n(L_i)} - P(a \mid L_i)\right)^2}{P(a \mid L_i)} \leq \sum_{a\in A}\frac{\frac{\delta\ln(n)}{N_n(L_i)}}{P(a \mid L_i)} \quad (2.16)$$

e

$$D\left(\frac{N_n(L_j,\cdot)}{N_n(L_j)}\middle|\middle|P(\cdot \mid L_j)\right) \leq \sum_{a\in A}\frac{\left(\frac{N_n(L_j,a)}{N_n(L_j)} - P(a \mid L_j)\right)^2}{P(a \mid L_j)} \leq \sum_{a\in A}\frac{\frac{\delta\ln(n)}{N_n(L_j)}}{P(a \mid L_j)}. \quad (2.17)$$

Então, usando (2.16) e (2.17) em (2.15), segue que, quase certamente quando $n \to \infty$, temos

$$BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n) \leq N_n(L_i)\frac{\delta \ln(n)}{N_n(L_i)} \cdot \sum_{a \in A} \frac{1}{P(a \mid L_i)} + N_n(L_j)\frac{\delta \ln(n)}{N_n(L_j)} \cdot \sum_{a \in A} \frac{1}{P(a \mid L_j)} - \frac{(|A|-1)}{2}\ln(n). \tag{2.18}$$

Agora, como A é finito e $P(a \mid L) > 0$, $\forall\, a \in A$ e $\forall\, L \in \mathcal{L}$, usando (2.9) concluímos que $p = \min_{a \in A}\{P(a \mid L_i)\} = \min_{a \in A}\{P(a \mid L_j)\} > 0$.

Logo, por (2.18), temos

$$BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n) \leq \frac{2|A|\delta \ln(n)}{p} - \frac{(|A|-1)}{2}\ln(n).$$

Assim, escolhendo $\delta < \dfrac{p(|A|-1)}{4|A|}$, obtemos

$$BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n) < \left(\frac{(|A|-1)}{2} - \frac{(|A|-1)}{2}\right)\ln(n) = 0.$$

Portanto, segue (2.10), ou seja, quase certamente quando $n \to \infty$, temos

$$BIC(\mathcal{L}^{ij}, x_1^n) > BIC(\mathcal{L}, x_1^n).$$

Por outro lado, suponha que (2.10) é satisfeita. Por facilidade denote

$$r_n(L, a) = \frac{N_n(L, a)}{n} \qquad \text{e} \qquad r_n(L) = \frac{N_n(L)}{n}. \tag{2.19}$$

Então, para n suficientemente grande, dividindo (2.13) por n, obtemos

$$\frac{BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^{ij}, x_1^n)}{n} = \sum_{a \in A}\left\{ r_n(L_i, a)\ln\left(\frac{r_n(L_i, a)}{r_n(L_i)}\right) + r_n(L_j, a)\ln\left(\frac{r_n(L_j, a)}{r_n(L_j)}\right) - r_n(L_{ij}, a)\ln\left(\frac{r_n(L_{ij}, a)}{r_n(L_{ij})}\right)\right\} - \frac{(|A|-1)}{2} \cdot \frac{\ln(n)}{n}. \tag{2.20}$$

Como, $r_n(L_i)$, $r_n(L_j)$, $r_n(L_i, a)$ e $r_n(L_j, a)$, $\forall\, a \in A$, são não negativos, $r_n(L_i, a) + r_n(L_j, a) = r_n(L_{ij}, a)$ e $r_n(L_i) + r_n(L_j) = r_n(L_{ij})$, usando (1.5), temos que cada parcela do somatório em

(2.20) é não negativa. Logo, para n suficientemente grande,

$$\frac{BIC(\mathcal{L},x_1^n)-BIC(\mathcal{L}^{ij},x_1^n)}{n} \geq -\frac{(|A|-1)}{2}\cdot\frac{\ln(n)}{n}.$$

Ou seja,

$$\lim_{n\to\infty}\frac{BIC(\mathcal{L},x_1^n)-BIC(\mathcal{L}^{ij},x_1^n)}{n} \geq 0, \text{ quase certamente.} \tag{2.21}$$

Mas, pela hipótese (2.10), temos que, quase certamente quando $n\to\infty$, $BIC(\mathcal{L},x_1^n) < BIC(\mathcal{L}^{ij},x_1^n)$. Assim, quase certamente,

$$\lim_{n\to\infty}\frac{BIC(\mathcal{L},x_1^n)-BIC(\mathcal{L}^{ij},x_1^n)}{n} \leq 0. \tag{2.22}$$

Logo, por (2.21) e (2.22), segue que $\lim_{n\to\infty}\frac{BIC(\mathcal{L},x_1^n)-BIC(\mathcal{L}^{ij},x_1^n)}{n}=0, q.c.$ Assim, aplicando o limite quando $n\to\infty$ em (2.20), concluímos que, quase certamente,

$$\lim_{n\to\infty}\sum_{a\in A}\left\{r_n(L_i,a)\ln\left(\frac{r_n(L_i,a)}{r_n(L_i)}\right)+r_n(L_j,a)\ln\left(\frac{r_n(L_j,a)}{r_n(L_j)}\right)-r_n(L_{ij},a)\ln\left(\frac{r_n(L_{ij},a)}{r_n(L_{ij})}\right)\right\}$$
$$-\lim_{n\to\infty}\frac{(|A|-1)}{2}\cdot\frac{\ln(n)}{n}=0.$$

Ou seja,

$$\lim_{n\to\infty}\sum_{a\in A}\left\{r_n(L_i,a)\ln\left(\frac{r_n(L_i,a)}{r_n(L_i)}\right)+r_n(L_j,a)\ln\left(\frac{r_n(L_j,a)}{r_n(L_j)}\right)-r_n(L_{ij},a)\ln\left(\frac{r_n(L_{ij},a)}{r_n(L_{ij})}\right)\right\}=0, q.c.$$

Agora, como o somatório é sobre A, finito, e $\ln(x)$ é contínua sobre $x>0$, podemos utilizar as propriedades de limite e a Proposição 1.7 para concluir que

$$\sum_{a\in A}\left\{P(L_i,a)\ln\left(\frac{P(L_i,a)}{P(L_i)}\right)+P(L_j,a)\ln\left(\frac{P(L_j,a)}{P(L_j)}\right)-P(L_{ij},a)\ln\left(\frac{P(L_{ij},a)}{P(L_{ij})}\right)\right\}=0. \tag{2.23}$$

Finalmente, como $P(L_i)$, $P(L_j)$, $P(L_i,a)$ e $P(L_j,a)$, $\forall\, a\in A$, são não negativos, $P(L_i,a)+P(L_j,a)=P(L_{ij},a)$ e $P(L_i)+P(L_j)=P(L_{ij})$, segue de (1.5) que cada parcela do somatório acima é não negativa, mas, por (2.23), isso implica que,

$$\forall\, a\in A,\ \frac{P(L_i,a)}{P(L_i)}=\frac{P(L_j,a)}{P(L_j)}.$$

Portanto, obtemos (2.9), ou seja,

$$\forall\, a \in A,\, P(a \mid L_i) = P(a \mid L_j).$$

□

Seguindo o raciocínio utilizado na demonstração do teorema anterior, é possível utilizar o mesmo critério, para decidir se duas ou mais partes devem ser unidas. É o que estabelece o corolário a seguir.

Corolário 2.1. Sejam $\{X_t\}$ uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária, tal que $P(a \mid s) > 0$, $\forall\ a \in A$ e $\forall\ s \in S = A^M$ e x_1^n, $n > M$ valores observados do processo. Se $\mathcal{L} = \{L_1, \ldots, L_{|\mathcal{L}|}\}$ é uma partição de S com K partes boas, $\{L_{i_k}\}_{k=1}^{K}$ e $T \subset \{1, \ldots, K\}$ é um subconjunto de índices das partes boas, então

$$P(a \mid L_{i_k}) = P(a \mid L_{i_l}),\, \forall\, a \in A,\, k, l \in T \tag{2.24}$$

se, e somente se, quase certamente quando $n \to \infty$,

$$BIC(\mathcal{L}^T, x_1^n) > BIC(\mathcal{L}, x_1^n), \tag{2.25}$$

onde $\mathcal{L}^T$ denota a partição que une as $|T|$ partes boas em $L_T = \bigcup_{k \in T} L_{i_k}$.

Demonstração. Primeiramente, observe que, seguindo a mesma ideia utilizada para encontrar (2.13), para toda parte L de S, podemos obter, para n suficientemente grande tal que $N_n(L) > 0$, que

$$\begin{aligned} BIC(\mathcal{L}, x_1^n) - BIC(\mathcal{L}^T, x_1^n) = \sum_{a \in A} \Bigg\{ \sum_{k \in T} \left[N_n(L_{i_k}, a) \ln\left(\frac{N_n(L_{i_k}, a)}{N_n(L_{i_k})} \right) \right] \\ - N_n(L_T,\, a) \ln\left(\frac{N_n(L_T, a)}{N_n(L_T)} \right) \Bigg\} - \frac{(|A|-1)(|T|-1)}{2} \cdot \ln(n). \end{aligned} \tag{2.26}$$

Agora, por um lado, assumindo que (2.24) é válida, segue que L_T é uma parte boa e $P(a \mid L_T) = P(a \mid L_{i_k})$, $\forall\, k \in T$.

Pela Proposição 1.8, temos que $\dfrac{N_n(L_T, a)}{N_n(L_T)}$ é o estimador de máxima verossimilhança de $P(a \mid L_T)$. Desse modo, realizando os passos análogos aos passos da demonstração do Teorema

2.1 que deduziram (2.15) e lembrando que $\sum_{k\in T} N_n(L_{i_k},a) = N_n(L_T)$, concluímos que

$$BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^T,x_1^n) \leq \sum_{a\in A}\left\{\sum_{k\in T}\left[N_n(L_{i_k},a)\ln\left(\frac{N_n(L_{i_k},a)}{N_n(L_{i_k})}\right)\right] - N_n(L_T,a)\ln\Big(P(a\mid L_T)\Big)\right\} - \frac{(|A|-1)(|T|-1)}{2}\cdot\ln(n).$$

Ou seja,

$$BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^T,x_1^n) \leq \sum_{k\in T} N_n(L_{i_k})D\left(\frac{N_n(L_{i_k,\cdot})}{N_n(L_{i_k})}\middle\|P(\,\cdot\mid L_{i_k})\right) - \frac{(|A|-1)(|T|-1)}{2}\cdot\ln(n).$$

Assim, seguindo o mesmo raciocínio da demonstração do Teorema 2.1, dado $\delta > 0$ e denotando $p = \min\{P(a\mid L_{i_k}) \mid a\in A,\, k\in T\}$, podemos obter que, quase certamente quando $n\to\infty$,

$$\begin{aligned} BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^T,x_1^n) &\leq \frac{|T||A|\delta}{p}\ln(n) - \frac{(|A|-1)(|T|-1)}{2}\ln(n) \\ &= \ln(n)\left(\frac{|T||A|\delta}{p} - \frac{(|A|-1)(|T|-1)}{2}\right). \end{aligned}$$

Portanto, escolhendo $\delta < \dfrac{p}{|T||A|}\dfrac{(|A|-1)(|T|-1)}{2}$, concluímos que

$$BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^T,x_1^n) < \ln(n)\left(\frac{|T||A|}{p}\frac{p}{|T||A|}\frac{(|A|-1)(|T|-1)}{2} - \frac{(|A|-1)(|T|-1)}{2}\right) = 0,$$

ou seja, quase certamente quando $n\to\infty$, temos

$$BIC(\mathcal{L}^T,x_1^n) > BIC(\mathcal{L},x_1^n)$$

e obtemos (2.25).

Por outro lado, assuma que (2.25) é válida.

Por (2.26), temos que

$$\frac{BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^T,x_1^n)}{n} = \sum_{a\in A}\left\{\sum_{k\in T}\left[r_n(L_{i_k},a)\ln\left(\frac{r_n(L_{i_k},a)}{r_n(L_{i_k})}\right)\right] - r_n(L_T,\,a)\ln\left(\frac{r_n(L_T,\,a)}{r_n(L_T)}\right)\right\} - \frac{(|A|-1)(|T|-1)}{2}\cdot\frac{\ln(n)}{n}. \quad (2.27)$$

Agora, como $r_n(L_{i_k},a)$ e $r_n(L_{i_k})$, $\forall\ a \in A$ e $\forall\ k \in T$, são não negativos, $\sum_{k\in T} r_n(L_{i_k},a) = r_n(L_T,a)$ e $\sum_{k\in T} r_n(L_{i_k}) = r_n(L_T)$, então por (1.5) na Observação 1.1, temos que, cada parcela do somatório sobre A em (2.27) é não negativa.

Logo, concluímos que

$$\lim_{n\to\infty} \frac{BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^T,x_1^n)}{n} \geq 0,\ q.c. \tag{2.28}$$

Mas, pela hipótese, (2.25), segue de (2.28) que

$$\lim_{n\to\infty} \frac{BIC(\mathcal{L},x_1^n) - BIC(\mathcal{L}^T,x_1^n)}{n} = 0,\ q.c. \tag{2.29}$$

Assim, de (2.27) e (2.29) e pela Proposição 1.7, segue que

$$\sum_{a\in A}\left\{\sum_{k\in T} P(L_{i_k},a)\ln\left(\frac{P(L_{i_k},a)}{P(L_{i_k})}\right) - P(L_T,a)\ln\left(\frac{P(L_T,a)}{P(L_T)}\right)\right\} = 0. \tag{2.30}$$

Agora, como $P(L_{i_k},a)$ e $P(L_{i_k})$ são não negativos, $\forall\ a \in A$ e $\forall\ k \in T$, $P(L_T,a) = \sum_{k\in T} P(L_{i_k},a)$ e $P(L_T) = \sum_{k\in T} P(L_{i_k})$, novamente (1.5) implica que cada parcela do somatório em A é não negativa, o que implica por (2.30) que cada parcela não negativa do somatório em A é também nula.

Portanto, segue que

$$P(a \mid L_{i_k}) = P(a \mid L_{i_l}),\ \forall\ a \in A,\ k,\ l \in T$$

e obtemos (2.24). □

O corolário a seguir estabelece a consistência do *BIC* para decidir quando duas ou mais partes boas não devem ser unidas, aspecto importante para a estratégia de estimação que será detalhada na Seção 3.3.

Corolário 2.2. Sob as mesmas hipóteses do Corolário 2.1, temos que: se

$$P(a \mid L_{i_k}) \neq P(a \mid L_{i_l}), \text{ para algum } a \in A \text{ e } k,\ l \in T,$$

então, quase certamente quando $n \to \infty$,

$$BIC(\mathcal{L},x_1^n) > BIC(\mathcal{L}^T,x_1^n).$$

Demonstração. Por (2.27) e por (1.5), temos que, quase certamente

$$\lim_{n\to\infty}\frac{BIC(\mathcal{L},x_1^n)-BIC(\mathcal{L}^T,x_1^n)}{n}=\sum_{a\in A}\left\{\sum_{k\in T}P(L_{i_k},a)\ln\left(\frac{P(L_{i_k},a)}{P(L_{i_k})}\right)-P(L_T,a)\ln\left(\frac{P(L_T,a)}{P(L_T)}\right)\right\}>0.$$

Assim, quase certamente, para todo n suficientemente grande,

$$\frac{BIC(\mathcal{L},x_1^n)-BIC(\mathcal{L}^T,x_1^n)}{n}>0,$$

ou seja, $BIC(\mathcal{L},x_1^n)>BIC(\mathcal{L}^T,x_1^n)$, quase certamente quando $n\to\infty$.

□

Finalmente, podemos mostrar que o critério *BIC* é fortemente consistente para estimar a Partição Mínima de uma Cadeia de Markov com Partição.

Teorema 2.2. Seja$\{X_t\}$ uma Cadeia de Markov com Partição $\mathcal{L}^*$ de ordem M, com espaço de estados finito A, estacionária, tal que $P(a\mid s)>0$, $\forall\, a\in A$ e $\forall\, s\in S=A^M$.

Dados n valores amostrados x_1^n da cadeia, com $n>M$, então, quase certamente quando $n\to\infty$,

$$\mathcal{L}_n^*:=\underset{\mathcal{L}\in\mathcal{P}}{\operatorname{argmax}}\{BIC(\mathcal{L},x_1^n)\}=\mathcal{L}^*,$$

onde $\mathcal{P}$ denota a coleção de todas as possíveis partições de S e $\mathcal{L}^*$ é a partição boa mínima, conforme a Definição 1.7.

Demonstração. Vamos denotar por $\mathcal{P}^*$, o conjunto de todas as partições boas de S. Como A é finito, então a coleção $\mathcal{P}$ também é finita.

Primeiramente, observe que se $\mathcal{L}_r=\{R_1,R_2,\ldots,R_{|\mathcal{L}_r|}\}\in\mathcal{P}\setminus\mathcal{P}^*$ é uma partição qualquer que não seja partição boa, então $\mathcal{L}_r$ possui ao menos uma parte que não satisfaz o item (i) da Definição 2.1.

Suponha sem perda de generalidade que $R_1=\{s_1,s_2,\ldots,s_m\}$ seja tal parte. Então, existem $1\le i<j\le m$ tais que, para algum $a\in A$, temos $P(a\mid s_i)\neq P(a\mid s_j)$.

Considere $\mathcal{L}_r^{'}=\{\{s_1\},\{s_2\},\ldots,\{s_m\},R_2,R_3,\ldots,R_{|\mathcal{L}_r|}\}\neq\mathcal{L}_r$.

Como, $\bigcup_{i=1}^{m}\{s_i\}=R_1$, cada $\{s_i\}$ é uma parte boa, $\forall\, i=1,2,\ldots,m$, e $P(a\mid s)=P(a\mid\{s\})$, $\forall\, a\in A$ e $\forall\, s\in S$, então pelo Corolário 2.2, segue que, quase certamente quando $n\to\infty$,

$$BIC(\mathcal{L}_r^{'},x_1^n)>BIC(\mathcal{L}_r,x_1^n),$$

ou seja, $\mathcal{L}_n^*\neq\mathcal{L}_r$, quase certamente, para todo n suficientemente grande.

Como $\mathcal{L}_r$ é uma partição qualquer do conjunto finito $\mathcal{P} \setminus \mathcal{P}^*$, então

$$\mathcal{L}_n^* \in \mathcal{P}^*, \text{ quase certamente quando } n \to \infty. \tag{2.31}$$

Agora, considere $\mathcal{L}_b = \{B_1, B_2, \ldots, B_{|\mathcal{L}_b|}\} \in \mathcal{P}^* \setminus \{\mathcal{L}^*\}$, isto é, uma partição boa, diferente da Partição Mínima da Definição 1.7.

Como $\mathcal{L}_b$ é partição boa, então contém apenas partes boas. Porém, não é a partição $\mathcal{L}^*$, gerada pela relação de equivalência $\sim_p$, logo, existe uma parte $B_i \in \mathcal{L}_b$, com $i = 1, 2, \ldots, |\mathcal{L}_b|$, tal que, $B_i \neq [s^i]$, onde $[s^i]$ é a classe de equivalência de $s^i \in B_i$. Ou seja, existe, $s^j \in B_j$, com $j \neq i$, tal que $s^j \sim_p s^i$. Desse modo, como B_i e B_j são partes boas, temos que, existe $i \neq j$, tal que

$$P(a \mid B_i) = P(a \mid s^i) = P(a \mid s^j) = P(a \mid B_j), \forall\, a \in A,$$

onde a primeira e a última igualdades seguem da Observação 2.1 e a segunda igualdade segue de $s^j \sim_p s^i$.

Assim, pelo Teorema 2.1, sem perda de generalidade, supondo $i < j$, temos que, quase certamente quando $n \to \infty$,

$$BIC({\mathcal{L}_b}^{ij}, x_1^n) > BIC(\mathcal{L}_b, x_1^n),$$

ou seja, $\mathcal{L}_n^* \neq \mathcal{L}_b$, quase certamente, para todo n suficientemente grande.

Por (2.31) e por $\mathcal{L}_b$ ser uma partição qualquer do conjunto finito $\mathcal{P}^* \setminus \{\mathcal{L}^*\}$, segue que $\mathcal{L}_n^* = \mathcal{L}^*$ quase certamente quando $n \to \infty$.

□

Finalizamos este capítulo, observando que foram apresentados os principais conceitos e resultados estabelecidos por García e González-López em [11] e [13] que garantem a consistência teórica da estimação da Partição Mínima $\mathcal{L}^*$ do Modelo de Markov com Partição através do *BIC*, por meio da partição estimadora $\mathcal{L}_n^*$ proposta em ambos os trabalhos.

Além disso, os resultados apresentados serão fundamentais para a obtenção da consistência prática da estimação por meio da partição estimadora $\hat{\mathcal{L}}_n^*$. A estratégia proposta por García e González-López para o encontro de tal partição é detalhada no Capítulo 3.

Capítulo 3

Estimação pelo Critério de Determinação Eficiente

A utilização do *BIC* para obtenção de um estimador consistente para a ordem de uma Cadeia de Markov de ordem superior foi proposta por Katz (1981) em [19]. Uma prova formal da consistência forte desse estimador é apresentada em [5].

Nesse contexto, em [23], Zhao, Dorea e Gonçalves propuseram o Critério de Determinação Eficiente (*EDC*), que generaliza o *BIC*, para a estimação da ordem de uma Cadeia de Markov e demonstraram a consistência forte do estimador, sob condições suaves.

Especificamente, assumindo que a ordem verdadeira M da cadeia é tal que $0 \leq M \leq M^*$, onde $M^* > 0$ é previamente conhecido, o estimador de M, baseado em uma amostra de tamanho n da cadeia, é dado por

$$\hat{M}_n = \text{argmax}\{EDC(k),\ k = 0, 1, \ldots, M^*\} \tag{3.1}$$

e

$$EDC(k) = \ln(MV(k)) - \gamma(k) \cdot c_n, \tag{3.2}$$

onde $\gamma(k)$ é uma função estritamente crescente com a ordem da cadeia k e o termo $\{c_n\}$ é uma sequência de números reais positivos (ou de maneira mais geral, uma sequência de variáveis aleatórias positivas) apropriadamente escolhida. A vantagem desse critério é a flexibilidade na escolha do termo de penalidade.

A consistência forte do estimador foi estabelecida em [23], sob as condições

$$\frac{c_n}{n} \xrightarrow[n\to\infty]{} 0 \quad q.c \text{ e} \quad \frac{c_n}{\ln(\ln(n))} \xrightarrow[n\to\infty]{} +\infty \quad q.c. \tag{3.3}$$

Observe que, considerando $\gamma(k) = \frac{|A|^k(|A|-1)}{2}$ e $c_n = \ln(n)$, o *EDC* reduz-se ao *BIC*.

Neste capítulo, motivados pelo trabalho de Garcia e González-López ([13]), cujos resultados foram apresentados no Capítulo 2, propomos a utilização do *EDC* para a obtenção de um estimador mais geral para a Partição Mínima de uma Cadeia de Markov com Partição e, sob as mesmas condições sobre o termo de penalidade dadas em (3.3), estabelecemos a consistência forte do estimador proposto.

Assim, na Seção 3.1 estabelecemos o conceito do critério *EDC* no contexto do Modelo de Markov com Partição e apresentamos alguns resultados auxiliares para a estimação da Partição Mínima pelo critério *EDC*.

Na Seção 3.2, demonstramos, para o *EDC*, resultados análogos aos da Seção 2.2 que demonstram a consistência forte do critério na estimação da Partição Mínima do Modelo de Markov com Partição.

Em seguida, na Seção 3.3, abordamos o problema da estimação da Partição Mínima $\mathcal{L}^*$ por meio da partição estimadora $\mathcal{L}_n^*$ que maximiza o critério *EDC* dentre todas as partições possíveis. Para tanto, demonstramos a convergência para $\mathcal{L}^*$ da sequência de partições $\{\hat{\mathcal{L}}_n^*\}$ gerada pelo algoritmo de seleção proposto por Garcia e González-López em [11], adaptado ao critério *EDC*.

Por fim, na Seção 3.4, apresentamos uma breve discussão sobre a busca por um termo de penalidade ótimo no *EDC* para estimação da Partição Mínima $\mathcal{L}^*$, nos moldes dos trabalhos de Dorea em [7] e de Dorea, Resende e Gonçalves em [8].

As referências bibliográficas utilizadas neste capítulo são: [7], [8], [11], [13] ,[21] e [23].

3.1 O *EDC* em Cadeias de Markov com Partição

Seja $\{X_t\}$ uma Cadeia de Markov com Partição $\mathcal{L}^*$, conforme a Definição 1.7, de ordem M e espaço de estados finito A. Neste caso, em especial, $\{X_t\}$ é estacionária e assumimos $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S = A^M$.

Dados n valores observados x_1^n, $n > M$, da cadeia e assumindo uma partição arbitrária $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ dentre a classe $\mathcal{P}$ de todas as partições de S, o *EDC do Modelo* $\{X_t\}$, *com Partição* $\mathcal{L}$, é definido por

$$EDC(\mathcal{L}, x_1^n) = \ln\left(ML(\mathcal{L}, x_1^n)\right) - \gamma(|\mathcal{L}|) \cdot c_n, \tag{3.4}$$

onde $MV(\mathcal{L}, x_1^n)$ é a função de máxima verossimilhança modificada para o modelo da Cadeia de Markov com Partição $\mathcal{L}$, dada em (2.1), $\gamma(k)$ é uma função real qualquer estritamente

crescente em k e $\{c_n\}$ é uma sequência de números reais positivos (ou, de forma mais geral, uma sequência de variáveis aleatórias positivas).

Substituindo (2.1) em (3.4), do mesmo modo que foi feito com o *BIC*, podemos obter a seguinte expressão para o *EDC*:

$$EDC(\mathcal{L}, x_1^n) = \sum_{\substack{a \in A,\, L \in \mathcal{L}, \\ N_n(L) > 0}} N_n(L,a) \cdot \ln\left(\frac{N_n(L,a)}{N_n(L)}\right) - \gamma(|\mathcal{L}|) \cdot c_n. \tag{3.5}$$

Neste caso, a partição estimadora $\mathcal{L}_n^*$ para $\mathcal{L}^*$ é tal que

$$\mathcal{L}_n^* = \underset{\mathcal{L} \in \mathcal{P}}{\operatorname{argmax}} \{EDC(\mathcal{L}, x_1^n)\}. \tag{3.6}$$

Note que, de fato, o *EDC* estende o *BIC*, pois $\gamma(k) = \dfrac{(|A|-1)k}{2}$ é uma função estritamente crescente e $c_n = \ln(n)$ é uma sequência de valores positivos a partir de $n = 2$. Além disso, $c_n = \ln(n)$ satisfaz as condições (3.3).

Como já mencionado na Seção 2.1 para o critério *BIC*, encontrar uma partição que maximize o critério *EDC* dentre todas as partições possíveis de S pode ser impossível no caso em que S tenha tamanho moderado.

Assim, seguimos a mesma estratégia adotada em [13] proposta por Garcia e González-López, que se baseia no método de redução de partição pela união de partes boas até encontrar uma partição estimadora $\hat{\mathcal{L}}_n^*$ que, quase certamente quando $n \to \infty$, se iguale a $\mathcal{L}^*$ e satisfaça (3.6).

Nesse sentido, estendemos os resultados de Garcia e González-López apresentados no Capítulo 2 para o *EDC*. Para isso, necessitamos de alguns resultados auxiliares que apresentamos a seguir.

Primeiramente, adaptamos o Lema 2 apresentado por Dorea em [7], para o Modelo de Markov com Partição.

Lema 3.1. Sejam $\{X_t\}$, uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária, tal que $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S = A^M$, $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$, uma partição de S e x_1^n, n valores observados da cadeia, com $n > M$. Se $L \in \mathcal{L}$ é uma parte boa, então, para $a \in A$ e $s \in S$, temos

$$\limsup_{n \to \infty} \frac{\Big(N_n(s,a) - P(a \mid s) \cdot N_n(s)\Big)^2}{N_n(L)\ln(\ln(n))} = \frac{2P(a \mid s)(1 - P(a \mid s))P(s)}{P(L)},\ q.c. \tag{3.7}$$

Demonstração. Como, por hipótese, $P(a \mid s) > 0$, pela Proposição 1.4, temos que $\left\{Y_t^{(M)}\right\}$, a Cadeia de Markov M-derivada, é ergódica e como $\{X_t\}$ é estacionária, temos que, $\forall\, t \geq 1$, $\mathbb{P}(X_t^{t+M-1} = s) = P(s) = \pi(s)$, $\forall\, s \in S = A^M$.

Assim, podemos aplicar o Lema 1.1, considerando, $k = M$, $a_{M+1} = a \in A$ e $a_1^M = s \in S = A^M$, e obter que quase certamente, temos

$$\limsup_{n\to\infty} \frac{\left(\sum_{t=1}^{n-M} I(X_t^{M+t-1} = s, X_{M+t} = a) - P(a \mid s) \left(\sum_{t=1}^{n-M+1} I(X_t^{M+t-1} = s) \right) \right)^2}{n \cdot \ln(\ln(n))} = 2P(s)P(a \mid s)\,(1 - P(a \mid s))\,.$$

Agora, como $N_n(s,a) = \sum_{t=1}^{n-M} I(X_t^{M+t-1} = s, X_{M+t} = a)$ e $N_n(s) = \sum_{t=1}^{n-M} I(X_t^{M+t-1} = s)$, podemos reescrever o limite acima e obter que quase certamente

$$\limsup_{n\to\infty} \left\{ \frac{\Big(N_n(s,a) - P(a \mid s)N_n(s)\Big)^2}{n \cdot \ln(\ln(n))} - 2\frac{P(a \mid s)I(X_{n-M+1}^n = s)}{\ln(\ln(n))} \frac{N_n(s,a) - P(a \mid s)N_n(s)}{n} + \frac{\Big(P(a \mid s)I(X_{n-M+1}^n = s)\Big)^2}{n \cdot \ln(\ln(n))} \right\} = 2P(s)P(a \mid s)\,(1 - P(a \mid s))\,.$$

Denotando

$$A_n = 2\frac{P(a \mid s)I(X_{n-M+1}^n = s)}{\ln(\ln(n))} \frac{N_n(s,a) - P(a \mid s)N_n(s)}{n} \text{ e } B_n = \frac{\Big(P(a \mid s)I(X_{n-M+1}^n = s)\Big)^2}{n \cdot \ln(\ln(n))},$$

podemos reescrever a expressão anterior como

$$\limsup_{n\to\infty} \left[\frac{\Big(N_n(s,a) - P(a \mid s)N_n(s)\Big)^2}{n \cdot \ln(\ln(n))} - A_n + B_n \right] = 2P(s)P(a \mid s)\,(1 - P(a \mid s))\,. \tag{3.8}$$

Como $\lim_{n\to\infty} \dfrac{P(a \mid s)I(X_{n-M+1}^n = s)}{\ln(\ln(n))} = 0$, quase certamente, pois o numerador é limitado e como $\left|\dfrac{N_n(s,a) - P(a \mid s)N_n(s)}{n}\right| \leq 1$, pois $-n \leq -N_n(s) \leq N_n(s,a) - P(a \mid s)N_n(s) \leq N_n(s,a) \leq n$,

então segue que

$$\lim_{n\to\infty} A_n = \lim_{n\to\infty} 2\frac{P(a \mid s)I(X_{n-M+1}^n = s)}{\ln(\ln(n))}\frac{N_n(s,a) - P(a \mid s)N_n(s)}{n} = 0,\ q.c \tag{3.9}$$

e

$$\lim_{n\to\infty} B_n = \lim_{n\to\infty} \frac{\Big(P(a \mid s)I(X_{n-M+1}^n = s)\Big)^2}{n \cdot \ln(\ln(n))} = 0,\ q.c. \tag{3.10}$$

Logo, por (3.8) a (3.10), temos que quase certamente,

$$\limsup_{n\to\infty} \frac{\Big(N_n(s,a) - P(a \mid s)N_n(s)\Big)^2}{n \cdot \ln(\ln(n))} = 2P(s)P(a \mid s)\left(1 - P(a \mid s)\right).$$

Agora, pela Proposição 1.7, $\lim\limits_{n\to\infty} \dfrac{N_n(L)}{n} = P(L) > 0$, logo $\lim\limits_{n\to\infty} \dfrac{n}{N_n(L)} = \dfrac{1}{P(L)} > 0,\ q.c.$ Então, segue que

$$\begin{aligned}\limsup_{n\to\infty} \frac{\Big(N_n(s,a) - P(a \mid s)N_n(s)\Big)^2}{N_n(L) \cdot \ln(\ln(n))} &= \limsup_{n\to\infty} \left[\frac{\Big(N_n(s,a) - P(a \mid s)N_n(s)\Big)^2}{n \cdot \ln(\ln(n))} \cdot \frac{n}{N_n(L)}\right] \\ &= \frac{2P(s)P(a \mid s)\left(1 - P(a \mid s)\right)}{P(L)},\ q.c.\end{aligned}$$

□

Usando o lema anterior, obtemos o resultado a seguir que será utilizado para a demonstração da consistência forte do *EDC* a qual será apresentada na próxima seção.

Proposição 3.1. Sejam $\{X_t\}$, uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária tal que $P(a \mid s) > 0,\ \forall\ a \in A$ e $\forall\ s \in S = A^M$, $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$, uma partição de S e x_1^n, n valores observados da cadeia, com $n > M$. Se $L \in \mathcal{L}$ é uma parte boa tal que $N_n(L) > 0$, então para $a \in A$, existe uma constante R, dependente apenas das probabilidades da cadeia, tal que, quase certamente quando $n \to \infty$, temos

$$\left(\frac{N_n(L,a)}{N_n(L)} - P(a \mid L)\right)^2 \leq \frac{R\ln(\ln(n))}{N_n(L)}. \tag{3.11}$$

Demonstração. Pelo Lema 3.1, dados $a \in A$ e $s \in S$, temos

$$\limsup_{n\to\infty} \frac{\Big(N_n(s,a) - P(a \mid s)N_n(s)\Big)^2}{N_n(L) \cdot \ln(\ln(n))} = C(a,s,L),\ q.c, \tag{3.12}$$

onde

$$C(a,s,L) = \frac{2P(s)P(a \mid s)(1-p(a \mid s))}{P(L)}.$$

Considere,

$$C = \max\{C(a,s,L) \mid a \in A, s \in S, L \in \mathcal{L}\},$$

que está bem definido, pois $A \times S \times \mathcal{L}$ é finito.

Assim, para $L \in \mathcal{L}$ uma parte boa, de (3.11), dados $a \in A$ e $s \in L$, existe $K(s,a) \geq 1$ suficientemente grande, tal que, $\forall\, n \geq K(s,a)$,

$$\frac{\Big(N_n(s,a) - P(a \mid s)N_n(s)\Big)^2}{N_n(L) \cdot \ln(\ln(n))} \leq C+1, \, q.c,$$

o que implica que, para $n \geq K(a) = \max\{K(s,a),\ s \in L\}$, temos que, $\forall\, s \in L$,

$$-\sqrt{(C+1) \cdot N_n(L)\ln(\ln(n))} \leq N_n(s,a) - P(a \mid s)N_n(s) \leq \sqrt{(C+1) \cdot N_n(L)\ln(\ln(n))}, \, q.c. \tag{3.13}$$

Agora, como L é uma parte boa temos que $P(a \mid s) = P(a \mid L)$, se $s \in L$. Então, como $N_n(L,a) = \sum_{s \in L} N_n(s,a)$ e $N_n(L) = \sum_{s \in L} N_n(s)$, considerando a soma $\forall\, s \in L$ em (3.13), obtemos, para todo n suficientemente grande,

$$-|L|\sqrt{(C+1) \cdot N_n(L)\ln(\ln(n))} \leq N_n(L,a) - P(a \mid L)N_n(L) \leq |L|\sqrt{(C+1) \cdot N_n(L)\ln(\ln(n))}, \, q.c.$$

Assim, como $N_n(L) > 0$, para todo n suficientemente grande, segue que

$$\left(\frac{N_n(L,a)}{N_n(L)} - P(a \mid L)\right)^2 \leq \frac{|L|^2 \cdot (C+1) \cdot \ln(\ln(n))}{N_n(L)}, \, q.c. \tag{3.14}$$

Agora, basta escolher $R = |S|^2 \cdot (C+1)$, que depende somente das probabilidades da cadeia e é tal que $|L|^2 \cdot (C+1) \leq R$. Portanto, de (3.14), segue que, quase certamente quando $n \to \infty$,

$$\left(\frac{N_n(L,a)}{N_n(L)} - P(a \mid L)\right)^2 \leq \frac{R\ln(\ln(n))}{N_n(L)}.$$

□

3.2 Consistência

Nesta seção, seguindo os mesmos argumentos utilizados por Garcia e González-López, estabelecemos resultados análogos aos do Teorema 2.1 e Teorema 2.2, para o *EDC*.

Iniciamos provando que, sob condições suaves, o *EDC* é um critério fortemente consistente para decidir se duas partes boas devem ser unidas.

Teorema 3.1. Sejam $\{X_t\}$, uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária, com $P(a \mid s) > 0, \forall\, a \in A,\ s \in S = A^M$ e x_1^n, n valores observados da cadeia, com $n > M$. Seja $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$, uma partição de S para a qual existem $i < j \in \{1, 2, \ldots, |\mathcal{L}|\}$ tais que L_i e L_j são partes boas. Então, para o *EDC* definido em (3.4), com $\{c_n\}$ satisfazendo

$$\lim_{n\to\infty} \frac{c_n}{n} = 0,\ q.c \ \text{ e } \lim_{n\to\infty} \frac{c_n}{\ln(\ln(n))} = +\infty,\ q.c, \tag{3.15}$$

temos que

$$P(a \mid L_i) = P(a \mid L_j),\ \forall\, a \in A \tag{3.16}$$

se, e somente se, quase certamente quando $n \to \infty$,

$$EDC(\mathcal{L}^{ij}, x_1^n) > EDC(\mathcal{L}, x_1^n). \tag{3.17}$$

Demonstração. Seja n suficientemente grande tal que quase certamente $N_n(L) > 0$, para toda $L \in \mathcal{P}$

Por um lado, suponha que (3.16) é satisfeita. Por (3.5), temos

$$\begin{aligned} EDC(\mathcal{L}, x_1^n) - EDC(\mathcal{L}^{ij}, x_1^n) = &\ln\left(\prod_{a\in A}\left(\frac{N_n(L_i,a)}{N_n(L_i)}\right)^{N_n(L_i,a)}\right) + \ln\left(\prod_{a\in A}\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right)^{N_n(L_j,a)}\right) \\ &- \ln\left(\prod_{a\in A}\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right)^{N_n(L_{ij},a)}\right) - \Big(\gamma\big(|\mathcal{L}|\big) - \gamma\big(|\mathcal{L}^{ij}|\big)\Big)c_n. \end{aligned}$$

Agora, pela Proposição 1.8, segue que

$$\prod_{a\in A}\left(\frac{N_n(L_ij,a)}{N_n(L_{ij})}\right)^{N_n(L_{ij},a)} \geq \prod_{a\in A} P(a \mid L_{ij})^{N_n(L_{ij},a)}.$$

Logo, podemos obter que

$$\begin{aligned} EDC(\mathcal{L}, x_1^n) - EDC(\mathcal{L}^{ij}, x_1^n) \leq &\sum_{a\in a} N_n(L_i,a)\ln\left(\frac{N_n(L_i,a)}{N_n(L_i)}\right) + \sum_{a\in a} N_n(L_j,a)\ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right) \\ &- \sum_{a\in A} N_n(L_{ij},a)\ln\Big(P(a \mid L_{ij})\Big) - \Big(\gamma\big(|\mathcal{L}|\big) - \gamma\big(|\mathcal{L}^{ij}|\big)\Big)c_n. \end{aligned}$$

Como $N_n(L_{ij},a) = N_n(L_i,a) + N_n(L_j,a)$ e por hipótese $P(a \mid L_i) = P(a \mid L_j) = P(a \mid L_{ij})$, $\forall\, a \in A$, usando o conceito de entropia relativa dado na Definição 1.2, podemos reescrever a desigualdade anterior da seguinte forma:

$$EDC(\mathcal{L},x_1^n) - EDC(\mathcal{L}^{ij},x_1^n) \leq N_n(L_i)D\left(\frac{N_n(L_i,\cdot)}{N_n(L_i)}\middle\|P(\cdot \mid L_i)\right) + N_n(L_j)D\left(\frac{N_n(L_j,\cdot)}{N_n(L_j)}\middle\|P(\cdot \mid L_j)\right) - \Big(\gamma\big(|\mathcal{L}|\big) - \gamma\big(|\mathcal{L}^{ij}|\big)\Big)c_n. \quad (3.18)$$

Agora, pela Proposição 1.2 e Proposição 3.1, temos que existe uma constante R, dependente apenas das probabilidades do processo, tal que, quase certamente quando $n \to \infty$, temos

$$D\left(\frac{N_n(L_i,\cdot)}{N_n(L_i)}\middle\|P(\cdot \mid L_i)\right) \leq \sum_{a\in A} \frac{\left(\frac{N_n(L_i,a)}{N_n(L_i)} - P(a \mid L_i)\right)^2}{P(a \mid L_i)} \leq \sum_{a\in A} \frac{R\ln(\ln(n))}{N_n(L_i)P(a \mid L_i)}$$

e

$$D\left(\frac{N_n(L_j,\cdot)}{N_n(L_j)}\middle\|P(\cdot \mid L_j)\right) \leq \sum_{a\in A} \frac{\left(\frac{N_n(L_j,a)}{N_n(L_j)} - P(a \mid L_j)\right)^2}{P(a \mid L_j)} \leq \sum_{a\in A} \frac{R\ln(\ln(n))}{N_n(L_j)P(a \mid L_j)}.$$

Desse modo, aplicando as desigualdades acima em (3.18), segue que, quase certamente, para todo n suficientemente grande,

$$EDC(\mathcal{L},x_1^n) - EDC(\mathcal{L}^{ij},x_1^n) \leq \sum_{a\in A} \frac{R\ln(\ln(n))}{P(a \mid L_i)} + \sum_{a\in A} \frac{R\ln(\ln(n))}{P(a \mid L_j)} - \Big(\gamma\big(|\mathcal{L}|\big) - \gamma\big(|\mathcal{L}^{ij}|\big)\Big)c_n.$$

Considere $p = \min\{P(a \mid L_i) = P(a \mid L_j),\, \forall\, a \in A\}$. Como A é finito e por hipótese, $\forall\, a \in A$, $P(a \mid L_i) = P(a \mid L_j) > 0$, temos que $p > 0$ e da desigualdade anterior obtemos

$$\frac{EDC(\mathcal{L},x_1^n) - EDC(\mathcal{L}^{ij},x_1^n)}{\ln(\ln(n))} \leq \frac{2|A|R}{p} - \Big(\gamma\big(|\mathcal{L}|\big) - \gamma\big(|\mathcal{L}^{ij}|\big)\Big) \cdot \frac{c_n}{\ln(\ln(n))}.$$

Ou seja, podemos escrever

$$EDC(\mathcal{L}^{ij},x_1^n) - EDC(\mathcal{L},x_1^n) \geq O\Big(\ln(\ln(n))\Big) + \Big(\gamma\big(|\mathcal{L}|\big) - \gamma\big(|\mathcal{L}^{ij}|\big)\Big) \cdot c_n. \quad (3.19)$$

Agora, como $\gamma(\cdot)$ é estritamente crescente e $|\mathcal{L}^{ij}| < |\mathcal{L}|$, temos que $\Big(\gamma\big(|\mathcal{L}|\big) - \gamma\big(|\mathcal{L}^{ij}|\big)\Big) > 0$.

Portanto, como por (3.15), $\lim\limits_{n\to\infty} \dfrac{c_n}{\ln(\ln(n))} = +\infty$, $q.c$, de (3.19), segue que, quase certamente quando $n \to \infty$,

$$EDC(\mathcal{L}^{ij},x_1^n) > EDC(\mathcal{L},x_1^n)$$

e temos (3.17).

Reciprocamente, suponha que (3.17) seja satisfeita quase certamente quando $n \to \infty$.

Para n suficientemente grande, por (3.15) e usando as notações em (2.19), obtemos

$$\frac{EDC(\mathcal{L},x_1^n)-EDC(\mathcal{L}^{ij},x_1^n)}{n}=\sum_{a\in A}\left\{r_n(L_i,a)\ln\left(\frac{r_n(L_i,a)}{r_n(L_i)}\right)+r_n(L_j,a)\ln\left(\frac{r_n(L_j,a)}{r_n(L_j)}\right)\right.$$
$$\left.-r_n(L_{ij},a)\ln\left(\frac{r_n(L_{ij},a)}{r_n(L_{ij})}\right)\right\}-\left(\gamma\left(\left|\mathcal{L}\right|\right)-\gamma\left(\left|\mathcal{L}^{ij}\right|\right)\right)\cdot\frac{c_n}{n}. \quad (3.20)$$

Seguindo os mesmos argumentos utilizados na demonstração do Teorema 2.1 para obter (2.21), e como pela hipótese (3.15) temos $\lim_{n\to\infty}\frac{c_n}{n}=0$, $q.c$, podemos concluir de (3.20) que, quase certamente,

$$\lim_{n\to\infty}\frac{EDC(\mathcal{L},x_1^n)-EDC(\mathcal{L}^{ij},x_1^n)}{n}\geq\lim_{n\to\infty}\left(-\left(\gamma\left(\left|\mathcal{L}\right|\right)-\gamma\left(\left|\mathcal{L}^{ij}\right|\right)\right)\cdot\frac{c_n}{n}\right)=0.$$

Assim, pela hipótese (3.17), segue que

$$\lim_{n\to\infty}\frac{EDC(\mathcal{L},x_1^n)-EDC(\mathcal{L}^{ij},x_1^n)}{n}=0,\ q.c.$$

Agora, seguindo os passos finais da demonstração do Teorema 2.1, podemos obter (3.16).

□

Observação 3.1. Note que, no caso que $c_n=\ln(n)$, o EDC reduz-se ao BIC e como, neste caso $\{c_n\}$ satisfaz (3.15), o Teorema 2.1 segue do Teorema 3.1. Além disso, observe a importância das condições sobre $\{c_n\}$, para a demonstração do teorema. A condição $\lim_{n\to\infty}\frac{c_n}{\ln(\ln(n))}=+\infty$ é fundamental para demonstração da ida e a condição $\lim_{n\to\infty}\frac{c_n}{n}=0$ é fundamental para demonstração da volta do teorema. ◊

De forma análoga, podemos obter como consequência do Teorema 3.1, as respectivas versões do Corolário 2.1 e do Corolário 2.2 para o EDC.

Corolário 3.1. Sejam $\{X_t\}$, uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária, com $P(a\mid s)>0$, $\forall\ a\in A$ e $\forall s\in S=A^M$ e x_1^n, $n>M$ valores observados do processo. Se $\mathcal{L}=\{L_1,L_2,\ldots,L_{|\mathcal{L}|}\}$ é uma partição de S com K partes boas, $\{L_{i_k}\}_{k=1}^K$ e $T\subset\{1,\ldots,K\}$ é um subconjunto do conjunto de índices das partes boas com $|T|>2$ e se a condição (3.15) é satisfeita, então

$$P(a\mid L_{i_k})=P(a\mid L_{i_l}),\ \forall\ a\in A,\ k,l\in T$$

se, e somente se, quase certamente quando $n \to \infty$,

$$EDC(\mathcal{L}^T, x_1^n) > EDC(\mathcal{L}, x_1^n).$$

Demonstração. Segue os mesmos passos da prova do Corolário 2.1, observando que a Proposição 3.1 é utilizada aqui, nos mesmos moldes em que a Proposição 2.1 foi utilizada no caso do *BIC*.

□

Corolário 3.2. Sob as mesmas hipóteses do Corolário 3.1, temos que se

$$P(a \mid L_{i_k}) \neq P(a \mid L_{i_l}) \text{ para algum } a \in A \text{ e } k, l \in T,$$

então quase certamente quando $n \to \infty$,

$$EDC(\mathcal{L}, x_1^n) > EDC(\mathcal{L}^T, x_1^n).$$

Demonstração. Por hipótese (3.15), temos $\lim_{n\to\infty} \frac{c_n}{n} = 0$, *q.c.* Assim, podemos seguir os mesmos argumentos utilizados na demonstração do Corolário 2.2 e o resultado segue. □

Observamos que o Corolário 3.1 estabelece a consistência do *EDC* para decidir se duas ou mais partes boas devem ser unidas e o Corolário 3.2 estabelece a consistência do *EDC* para decidir quando duas ou mais partes boas não devem ser unidas. Essa característica vai ser fundamental para o a obtenção da consistência prática do algoritmo de seleção proposto por García e González-López em [11], que será adaptado ao *EDC* na Seção 3.3.

É possível, ainda, mostrar um resultado semelhante ao Corolário 2.2 aplicado ao caso em que $\mathcal{L}$ seja uma partição boa.

Lema 3.2. Seja $\{X_t\}$, uma Cadeia de Markov de ordem M, com espaço de estados finito A, estacionária, com $P(a \mid s) > 0, \forall\, a \in A$ e $\forall\, s \in S = A^M$ e seja $\mathcal{L} = \{L_1, L_2, \ldots, L_{\mathcal{L}}\}$, uma partição boa de S. Considere x_1^n , $n > M$ valores observados da cadeia. Se $\mathcal{G} = \{G_1, G_2, \ldots, G_{|\mathcal{G}|}\}$ é uma partição de S que não é boa e é tal que $\forall\, L \subset \mathcal{L}$, $L \subset G$, para algum $G \in \mathcal{G}$, e a condição (3.3) é satisfeita, então

$$EDC(\mathcal{L}, x_1^n) > EDC(\mathcal{G}, x_1^n), \text{ quase certamente quando } n \to \infty.$$

Demonstração. Seja $T_i \subset \{1, 2, \ldots, |\mathcal{L}|\}$, com $i = 1, 2, \ldots, |\mathcal{G}|$, o conjunto de índices tal que $L_k \subset G_i$, $\forall\, k \in T_i$, então $L_{T_i} = \bigcup_{k \in T_i} L_k \subset G_i$.

Agora, como $\mathcal{G}$ é partição, $G_i \cap G_j = \emptyset$, $\forall\, i \neq j$, segue que

$$G_i \setminus \bigcup_{k \in T_i} L_k = \emptyset, \forall\, i = 1, \ldots, |\mathcal{G}|. \tag{3.21}$$

De fato, caso contrário se existisse $s' \in G_i \setminus \bigcup_{k \in T_i} L_k$, com $s' \in S$, teríamos $s' \in L_p$ para algum $p \in \{1, 2, \ldots, |\mathcal{L}|\} \setminus T_i$. Mas, por hipótese $L_p \subset G_j$ para algum $j = 1, 2, \ldots, |\mathcal{G}|$ com $j \neq i$, pois $p \notin T_i$. Assim, teríamos $s' \in G_i \cap G_j$, o que é um absurdo.

Assim, temos de (3.21) que $G_i = \bigcup_{k \in T_i} L_k = L_{T_i}$ e, por (3.5), podemos escrever

$$\begin{aligned} EDC(\mathcal{L}, x_1^n) &= \sum_{a \in A} \sum_{L \in \mathcal{L}} N_n(L,a) \ln\left(\frac{N_n(L,a)}{N_n(L)}\right) - \gamma(|\mathcal{L}|) \cdot c_n \\ &= \sum_{a \in A} \sum_{i=1}^{|\mathcal{G}|} \sum_{k \in T_i} N_n(L_k,a) \ln\left(\frac{N_n(L_k,a)}{N_n(L_k)}\right) - \gamma(|\mathcal{L}|) \cdot c_n \end{aligned}$$

e

$$EDC(\mathcal{G}, x_1^n) = \sum_{a \in A} \sum_{i=1}^{|\mathcal{G}|} N_n(L_{T_i},a) \ln\left(\frac{N_n(L_{T_i},a)}{N_n(L_{T_i})}\right) - \gamma(|\mathcal{G}|) \cdot c_n.$$

Seguindo a notação dada em (2.19), obtemos

$$\frac{EDC(\mathcal{L}, x_1^n) - EDC(\mathcal{G}, x_1^n)}{n} =$$
$$\sum_{a \in A} \left\{ \sum_{i=1}^{|\mathcal{G}|} \left[\sum_{k \in T_i} \left(r_n(L_k,a) \ln\left(\frac{r_n(L_k,a)}{r_n(L_k)}\right)\right) - r_n(L_{T_i},a) \ln\left(\frac{r_n(L_{T_i},a)}{r_n(L_{T_i})}\right)\right]\right\} - \left(\gamma(|\mathcal{L}|) - \gamma(|\mathcal{G}|)\right)\frac{c_n}{n}.$$

Logo, como $\{c_n\}$ satisfaz a condição (3.3), usando a Proposição 1.7, podemos obter que, quase certamente,

$$\lim_{n \to \infty} \frac{EDC(\mathcal{L}, x_1^n) - EDC(\mathcal{G}, x_1^n)}{n} =$$
$$\sum_{a \in A} \left\{ \sum_{i=1}^{|\mathcal{G}|} \left[\sum_{k \in T_i} \left(P(L_k,a) \ln\left(\frac{P(L_k,a)}{P(L_k)}\right)\right) - P(L_{T_i},a) \ln\left(\frac{P(L_{T_i},a)}{P(L_{T_i})}\right)\right]\right\}.$$

Denotando, $C(i,a) = \sum_{k \in T_i} \left(P(L_k,a) \ln\left(\frac{P(L_k,a)}{P(L_k)}\right)\right) - P(L_{T_i},a) \ln\left(\frac{P(L_{T_i},a)}{P(L_{T_i})}\right)$, segue de (1.5) que temos $C(i,a) \geq 0$, $\forall\, a \in A$ e $i = 1, 2, \ldots, |\mathcal{G}|$. Mas, como $\mathcal{G}$ não é uma partição boa, então

existem $a_0 \in A$, e $k,l \in T_{i_0}$ para algum $i_0 = 1,2,\ldots,|\mathcal{G}|$, tais que $\frac{P(L_k,a_0)}{P(L_k)} \neq \frac{P(L_l,a_0)}{P(L_l)}$. Portanto, $C(i_0,a_0) > 0$.

Como $C(i,a) \geq 0$, $\forall\, a \in A$ e $i = 1,2,\ldots,|\mathcal{G}|$ e $C(i_0,a_0) > 0$, então $\sum_{a\in A}\sum_{i=1}^{|\mathcal{G}|} C(i,a) > 0$. Desse modo, quase certamente,

$$\lim_{n\to\infty} \frac{EDC(\mathcal{L},x_1^n) - EDC(\mathcal{G},x_1^n)}{n} > 0,$$

ou seja, quase certamente quando $n \to \infty$, temos

$$EDC(\mathcal{L},x_1^n) > EDC(\mathcal{G},x_1^n).$$

□

Finalmente, utilizando os resultados obtidos anteriormente, podemos mostrar que a partição estimadora $\mathcal{L}_n^*$ definida em (3.6) e em (3.4), sob as hipóteses (3.15), é consistente para a estimação da Partição Mínima $\mathcal{L}^*$ do Modelo de Markov com Partição.

Teorema 3.2. Sejam $\{X_t\}$, uma Cadeia de Markov com Partição $\mathcal{L}^*$ de ordem M, com espaço de estados finito A, estacionária, tal que $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S = A^M$ e x_1^n, $n > M$ valores observados da cadeia. Se a condição em (3.15) é satisfeita, então, quase certamente quando $n \to \infty$,

$$\mathcal{L}_n^* := \underset{\mathcal{L}\in\mathcal{P}}{\operatorname{argmax}}\{EDC(\mathcal{L},x_1^n)\} = \mathcal{L}^*,$$

onde $\mathcal{P}$ denota a coleção de todas as possíveis partições de S e $\mathcal{L}^*$ é a Partição Mínima da cadeia, conforme a Definição 1.7.

Demonstração. Denotando por $\mathcal{P}^*$, o conjunto de todas as partições boas de S. A demonstração será feita em dois passos:

(i) $\mathcal{L}_n^* \neq \mathcal{L}_r$, quase certamente quando $n \to \infty$, $\forall\, \mathcal{L}_r \in \mathcal{P} \setminus \mathcal{P}^*$;

(ii) $\mathcal{L}_n^* \neq \mathcal{L}_b$, quase certamente quando $n \to \infty$, $\forall\, \mathcal{L}_b \in \mathcal{P}^* \setminus \{\mathcal{L}^*\}$.

Pela Definição 1.1, queremos mostrar que $\mathbb{P}(O^*) = 1$, onde $O^* = \liminf_{n\to\infty}\{\omega \in \Omega \mid \mathcal{L}_n^* = \mathcal{L}^*\}$. Portanto, se (i) e (ii) valem, então o teorema está demonstrado.

De fato, denote o evento $O_{\mathcal{L}} = \liminf_{n\to\infty}\{\omega \in \Omega \mid \mathcal{L}_n^* \neq \mathcal{L}\}$ para $\mathcal{L} \in \mathcal{P}$. Por (i) e (ii), $\mathbb{P}(O_{\mathcal{L}}) = 1$, $\forall\, \mathcal{L} \neq \mathcal{L}^*$, logo $\mathbb{P}(O) = 1$, onde $O = \bigcap_{\mathcal{L}\neq\mathcal{L}^*} O_{\mathcal{L}}$. Fixando $\omega \in O$, temos que $\omega \in O_{\mathcal{L}}$, $\forall\, \mathcal{L} \neq \mathcal{L}^*$,

ou seja, existe $N_{\mathcal{L}}(\omega) \in \mathbb{N}$, suficientemente grande tal que $\mathcal{L}_n^*(\omega) \neq \mathcal{L}$ para todo $n > N_{\mathcal{L}}(\omega)$, com $\mathcal{L} \neq \mathcal{L}^*$. Tomando $N(\omega) = \max\{N_{\mathcal{L}}(\omega) \mid \mathcal{L} \neq \mathcal{L}^*\}$, então, para todo $n > N(\omega)$,

$$\mathcal{L}_n^*(\omega) \neq \mathcal{L}, \forall\, \mathcal{L} \neq \mathcal{L}^*,$$

ou seja, $\mathcal{L}_n^*(\omega) = \mathcal{L}^*$, para todo $n > N(\omega)$, logo $\omega \in O^*$. Como ω foi tomado arbitrariamente no conjunto O, mostramos que $O \subset O^*$, o que implica que $\mathbb{P}(O^*) = 1$.

Para mostrarmos (i), considere $\mathcal{L}_r = \{R_1, R_2, \ldots R_{|\mathcal{L}_r|}\} \in \mathcal{P} \setminus \mathcal{P}^*$, uma partição qualquer que não seja boa e $\mathcal{L} = \left\{\{s\} \in S\right\}$ a partição boa dada pelo próprio conjunto S. Como, $\forall\, \{s\} \in \mathcal{L}$, $\{s\} \subset R$, para algum $R \in \mathcal{L}_r$, então, pelo Lema 3.2, quase certamente quando $n \to \infty$,

$$EDC(\mathcal{L}, x_1^n) > EDC(\mathcal{L}_r, x_1^n),$$

ou seja, $\mathcal{L}_n^* \neq \mathcal{L}_r$ e temos (i).

Agora, para provar (ii), considere $\mathcal{L}_b = \{B_1, B_2, \ldots, B_{|\mathcal{L}_b|}\} \in \mathcal{P}^* \setminus \{\mathcal{L}^*\}$, isto é, uma partição boa diferente da Partição Mínima $\mathcal{L}^*$ da Definição 1.7. Por não ser a Partição Mínima, existem partes boas B_i e $B_j \in \mathcal{L}_b$, com $i < j$, tal que

$$P(a \mid B_i) = P(a \mid B_j), \forall\, a \in A.$$

Então, pelo Teorema 3.1 segue que, quase certamente quando $n \to \infty$,

$$EDC(\mathcal{L}_b^{ij}, x_1^n) > EDC(\mathcal{L}_b, x_1^n),$$

ou seja, $\mathcal{L}_n^* \neq \mathcal{L}_b$ e temos (ii). □

3.3 Seleção da Partição Mínima

Assim como foi observado no Capítulo 2, para o caso do *BIC*, o estimador $\mathcal{L}_n^*$, proposto em (3.6), é a partição que maximiza o $EDC(\mathcal{L}, x_1^n)$, dentre todas as partições possíveis do conjunto $S = A^M$.

Em [21], por exemplo, vemos que se $|S| = k$, o número de partições diferentes possíveis do conjunto S é dado por B_k, o k-ésimo número de Bell. Tais números crescem de maneira muito rápida. Para se ter uma noção, os números de Bell seguem a seguinte fórmula recursiva:

$$B_{m+1} = \sum_{k=0}^{m} \binom{m}{k} B_k,\ m \geq 1 \text{ e } B_1 = 1.$$

Desse modo, para um conjunto com um número ímpar $2m+1$ de elementos, por exemplo, podemos obter

$$B_{2m+1} = \sum_{k=0}^{2m} \binom{2m}{k} B_k \geq B_{2m} + (2m)B_{2m-1} \geq (2m)B_{2m-1} + (2m)B_{2m-1} = (4m)B_{2(m-1)+1} \geq$$
$$(4m)(4(m-1))B_{2(m-2)+1}$$

e assim, sucessivamente, podemos concluir que $B_{2m+1} \geq 4^m m!$.

Portanto, com essa simples cota inferior para os números de Bell, já é possível verificar um crescimento no mínimo fatorial, quanto maior for o número de elementos de um conjunto.

Exemplo 3.1. Em uma Cadeia de Markov $\{X_t\}$ com espaço de estados finito A, onde $|A|=2$ e ordem $M=4$. Sendo $S=A^M$, temos $|S|=|A|^M=2^4=16$. Nesse simples caso, se $\mathcal{P}$ é o conjunto das possíveis partições de S, então com auxílio de uma tabela com os números de Bell disponível em [21], por exemplo, temos que $|\mathcal{P}|=B_{16}=10480142147$. ✓

No exemplo acima, se fôssemos tentar estimar a Partição Mínima da cadeia, a partir de uma dada amostra, seria necessário de algum modo listar e calcular o *EDC* para 10480142147 Modelos com Partição e finalmente identificar, dentre todas essas possíveis partições aquela que maximiza o *EDC*. Fica claro que tal busca se torna inviável na prática, mesmo em um exemplo simples como o apresentado. Nesse sentido, Garcia e González-López propuseram em [11] e reapresentaram em [13] uma estratégia para encontrar um estimador da Partição Mínima da cadeia que não demande uma análise de tantos modelos.

Nesta seção, apresentamos o mesmo algoritmo proposto em [11], adaptado para o *EDC* e provamos que a sequência de partições $\hat{\mathcal{L}}_n^*$, gerada pelo algoritmo, quase certamente quando $n \to \infty$, converge à Partição Mínima $\mathcal{L}^*$, dada pela Definição 1.7.

Vale observar que quase certamente, para n suficientemente grande, temos que $\hat{\mathcal{L}}_n^*$ é a partição $\mathcal{L}_n^*$, definida em (3.6), que maximiza o *EDC* dentre todas as partições possíveis de S.

Para isso, seja $\{X_t\}$ uma Cadeia de Markov com Partição $\mathcal{L}^*$, conforme a Definição 1.7, de ordem M, com espaço de estados finito A, com $P(a \mid s) > 0$, $\forall\, a \in A$ e $\forall\, s \in S = A^M$.

Assim, como em [13], dados $n > M$ valores observados da cadeia, denotamos para cada partição $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ e para todos $i, j \in \{1, 2, \ldots, |\mathcal{L}|\}$, com $i < j$,

$$d_{\mathcal{L}}(i,j) := \frac{1}{c_n} \sum_{a \in A} \left\{ N_n(L_i,a) \ln\left(\frac{N_n(L_i,a)}{N_n(L_i)}\right) + N_n(L_j,a) \ln\left(\frac{N_n(L_j,a)}{N_n(L_j)}\right) - N_n(L_{ij},a) \ln\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right) \right\}. \tag{3.22}$$

Em linhas gerais, a ideia do algoritmo é iniciar a busca, a partir de uma partição boa inicial previamente escolhida. A partir dessa partição, procuram-se duas partes que compartilhem

as mesmas probabilidades de transição. Caso isso aconteça, unem-se as duas partes e a nova partição permanece boa. Repete-se esse processo até obter-se a partição boa mínima de acordo com o critério escolhido.

Especificamente, baseado no resultado do Teorema 3.1, dada uma partição boa $\mathcal{L} = \{L_1, \ldots, L_{|\mathcal{L}|}\}$, procuram-se duas partes L_i e L_j de tal modo que $EDC(\mathcal{L}, x_1^n) - EDC(\mathcal{L}^{ij}, x_1^n) < 0$.

Ou seja, em termos de (3.22) e observando que $|\mathcal{L}^{ij}| = |\mathcal{L}| - 1$, faz-se uma busca para encontrar $i, j \in \{1, \ldots, |\mathcal{L}|\}$, com $i < j$, tais que

$$d_{\mathcal{L}}(i, j) < \Big(\gamma(|\mathcal{L}|) - \gamma(|\mathcal{L}| - 1)\Big). \tag{3.23}$$

Sendo esse o caso, repete-se o mesmo processo agora para $\mathcal{L}^{ij}$, a partição que une essas duas partes. O processo segue até que se obtenha uma partição final $\hat{\mathcal{L}}_n^*$ na qual não existam duas partes que satisfaçam tal critério.

Dessa forma, o algoritmo adaptado segue o seguinte fluxograma geral:

Algoritmo (A): Seleção da Partição Mínima $\mathcal{L}^*$

Entrada: $\mathcal{L}$
Saída: $\hat{\mathcal{L}}_n^*$
início
  $i \leftarrow 0$, $j \leftarrow 1$, $m \leftarrow |\mathcal{L}|$
  enquanto $i < m-1$ **faça**
    $i \leftarrow i+1$
    $j \leftarrow i$
    enquanto $j < m$ **faça**
      $j \leftarrow j+1$
      $d \leftarrow d_{\mathcal{L}}(i,j)$
      enquanto $d < \Big(\gamma(m) - \gamma(m-1)\Big)$ **faça**
        $L_i \leftarrow L_i \cup L_j$
        se $j \leq m-1$ **então**
          para $l = j$ **até** $m-1$ **faça**
            $L_l \leftarrow L_{l+1}$
          fim
        fim
        $m \leftarrow m-1$, $\mathcal{L} \leftarrow \{L_1, \ldots, L_m\}$, $i \leftarrow 1$, $j \leftarrow 2$, $d \leftarrow d_{\mathcal{L}}(i,j)$
        se $m < 2$ **então**
          pare e **retorna** $\hat{\mathcal{L}}_n^* = \{L_1, \ldots, L_m\}$
        fim
      fim
    fim
  fim
  retorna $\hat{\mathcal{L}}_n^* = \{L_1, \ldots, L_m\}$
fim

Finalmente, no teorema a seguir, podemos provar a convergência da sequência de partições $\{\hat{\mathcal{L}}_n^*\}$, produzida pelo Algoritmo (A), na estimação de $\mathcal{L}^*$, a Partição Mínima do Modelo de Markov com Partição.

Teorema 3.3. Seja $\{X_t\}$, uma Cadeia de Markov com Partição $\mathcal{L}^*$ de ordem M, de acordo com a Definição 1.7. Dados $n > M$ valores observados x_1^n da cadeia, então a sequência $\hat{\mathcal{L}}_n^*$ de partições geradas pelo Algoritmo (A), dada uma partição boa inicial $\mathcal{L}$ e assumindo a condição (3.3), converge, quase certamente quando $n \to \infty$, para $\mathcal{L}^*$.

Demonstração. Seja $\mathcal{P}$ o conjunto de todas as partições de S e $\mathcal{P}^*$ o conjunto das partições boas de S. Assim como na demonstração do Teorema 3.2, vamos demonstrá-lo em dois passos:

(i) $\hat{\mathcal{L}}_n^* \neq \mathcal{L}_r$, quase certamente quando $n \to \infty$, $\forall\, \mathcal{L}_r \in \mathcal{P} \setminus \mathcal{P}^*$;

(ii) $\hat{\mathcal{L}}_n^* \neq \mathcal{L}_b$, quase certamente quando $n \to \infty$, $\forall\, \mathcal{L}_b \in \mathcal{P}^* \setminus \{\mathcal{L}^*\}$.

Dada uma partição boa inicial $\mathcal{L}$, observe que pelo Algoritmo (A), cada parte de $\mathcal{L}$, ou permaneceu inalterada ao longo do processo e é também uma parte de $\hat{\mathcal{L}}_n^*$, ou se uniu a alguma outra parte durante o processo, gerando uma parte formada por união de elementos de $\mathcal{L}$. Assim, todo elemento de $\mathcal{L}$ é subconjunto (próprio, ou não) de algum elemento de $\hat{\mathcal{L}}_n^*$.

Além disso, o algoritmo "engrossa" a partição boa inicial $\mathcal{L}$ em um número p finito de passos: $\mathcal{L} = \mathcal{L}_0 \to \mathcal{L}_1 \to \ldots \to \mathcal{L}_p = \hat{\mathcal{L}}_n^*$, onde temos $\mathcal{L}_{k+1} = \mathcal{L}_k^{ij}$, $\forall\, k = 0, \ldots, p-1$, seguindo a notação da Observação 1.3.

Como $d_{\mathcal{L}_k}(i,j) < \Big(\gamma(|\mathcal{L}_k|) - \gamma(|\mathcal{L}_k|-1)\Big)$ em cada um desses passos, então, por (3.23), temos

$$EDC(\hat{\mathcal{L}}_n^*, x_1^n) > EDC(\mathcal{L}_{p-1}, x_1^n) > \cdots > EDC(\mathcal{L}, x_1^n). \tag{3.24}$$

Para provar (i), considere $\mathcal{L}_r \in \mathcal{P} \setminus \mathcal{P}^*$, uma partição que não seja boa. Por um lado, se $\mathcal{L}_r$ não é uma partição que foi obtida unindo as partes de $\mathcal{L}$, então pela própria construção do algoritmo $\hat{\mathcal{L}}_n^* \neq \mathcal{L}_r$, quase certamente quando $n \to \infty$.

Se, por outro lado, $\mathcal{L}_r$ é tal que "engrosse" $\mathcal{L}$ da forma descrita acima, então, por (3.24) e pelo Lema 3.2, quase certamente quando $n \to \infty$, temos

$$EDC(\hat{\mathcal{L}}_n^*, x_1^n) > EDC(\mathcal{L}, x_1^n) > EDC(\mathcal{L}_r, x_1^n),$$

ou seja, $\hat{\mathcal{L}}_n^* \neq \mathcal{L}_r$, $q.c$ quando $n \to \infty$.

Agora, para provar (ii), observe primeiramente que o algoritmo é finalizado quando $d_{\hat{\mathcal{L}}_n^*}(i,j) \geq \Big(\gamma(|\hat{\mathcal{L}}_n^*|) - \gamma(|\hat{\mathcal{L}}_n^*|-1)\Big)$, $\forall\, i < j \in \{1,2,\ldots,|\hat{\mathcal{L}}_n^*|\}$, ou seja, quando

$$EDC(\hat{\mathcal{L}}_n^*, x_1^n) \geq EDC(\hat{\mathcal{L}}_n^{*ij}, x_1^n),\ \forall\, i < j \in \{1,2,\ldots,|\hat{\mathcal{L}}_n^*|\}. \tag{3.25}$$

Seja $\mathcal{L}_b = \{B_1, B_2, \ldots, B_{|\mathcal{L}_b|}\} \in \mathcal{P}^* \setminus \{\mathcal{L}^*\}$, uma partição boa distinta da Partição Mínima $\mathcal{L}^*$ da cadeia, caracterizada na Definição 1.7.

Desse modo, existem $i < j \in \{1,2,\ldots,|\mathcal{L}_b|\}$ tais que $P(a \mid B_i) = P(a \mid B_j)$, $\forall\, a \in A$. Então, pelo Teorema 3.1, quase certamente quando $n \to \infty$, temos

$$EDC(\mathcal{L}_b, x_1^n) < EDC(\mathcal{L}_b^{ij}, x_1^n),$$

ou seja, $\mathcal{L}_b$ não satisfaz (3.25) quase certamente, para todo n suficientemente grande e, consequentemente, temos (ii).

Portanto, de (i) e (ii) segue que $\hat{\mathcal{L}}_n^* = \mathcal{L}^*$, quase certamente quando $n \to \infty$.

□

Corolário 3.3. Dados x_1^n, n valores observados de uma Cadeia de Markov com Partição $\mathcal{L}^*$, de ordem M, de acordo com a Definição 1.7, com $n > M$, então $\hat{\mathcal{L}}_n$ dado pelo Algoritmo 3.1 [11], ou seja, considerando o *BIC*, converge, quase certamente quando $n \to \infty$, para $\mathcal{L}^*$.

Demonstração. Basta considerar, no Algoritmo (A) de Seleção da Partição Mínima $\mathcal{L}^*$, $\gamma(k) = \frac{(|A|-1)k}{2}$ e $c_n = \ln(n)$ que o critério *EDC* reduz-se ao *BIC*. Assim, a convergência segue do Teorema 3.3.

□

A seguir, considerando o caso descrito no Exemplo 3.1, vamos comparar as diferenças na quantidade de modelos a serem analisados utilizando-se o Algoritmo (A) com a situação em que analisamos todos os modelos, como seria necessário utilizando o Teorema 3.2 literalmente.

Exemplo 3.2. Suponha uma Cadeia de Markov como no Exemplo 3.1. Como vimos na demonstração do Teorema 3.3, o Algoritmo (A) "engrossa" a partição boa inicial $\mathcal{L}$ em um número $p \leq |\mathcal{L}|-1 \leq |S|-1$ de passos: $\mathcal{L} = \mathcal{L}_0 \to \mathcal{L}_1 \to \ldots \to \mathcal{L}_p = \hat{\mathcal{L}}_n^*$.

No caso em que teríamos o maior número de passos possíveis, a partição boa inicial é $\mathcal{L}_0 = S$ e realizar-se-iam $|S|-1$ passos até a última partição, que seria uma partição contendo apenas um elemento, o próprio S. Neste caso, para cada partição $\mathcal{L}_k$, calcula-se $d_{\mathcal{L}_k}(i,j)$ no máximo para toda as combinações (i,j) tais que $i < j \in \{1,2,\ldots,|\mathcal{L}_k|\}$, exceto, obviamente, para a última partição que conteria apenas uma parte.

Para cada função $d_{\mathcal{L}_k}(i,j)$, calcula-se também a diferença $\Big(\gamma(|\mathcal{L}_k|) - \gamma(|\mathcal{L}_k|-1)\Big)$. Em suma, realizam-se ao todo no processo um número máximo $D_{|S|}$ de comparações entre modelos, por meio de funções $d_{\mathcal{L}_k}(i,j)$ e constantes $\Big(\gamma(|\mathcal{L}_k|) - \gamma(|\mathcal{L}_k|-1)\Big)$, onde

$$\begin{aligned} D_{|S|} &= \binom{|\mathcal{L}_0|}{2} + \binom{|\mathcal{L}_1|}{2} + \cdots + \binom{|\mathcal{L}_{|S|-2}|}{2} \\ &= \binom{|S|}{2} + \binom{|S|-1}{2} + \cdots + \binom{2}{2} = \frac{(|S|+1)(|S|)(|S|-1)}{6}. \end{aligned}$$

Comparando com o Exemplo 3.1, enquanto o número de modelos a serem analisados é de $|\mathcal{P}| = B_{16} = 10480142147$, o número máximo de comparações entre modelos realizadas pelo Algoritmo (A) seria de $D_{16} = \frac{17 \cdot 16 \cdot 15}{6} = 680$. ✓

A partir desse simples exemplo, fica claro o quanto a utilização do Algoritmo (A) permite diminuir consideravelmente a quantidade de modelos a serem analisados e torna o processo de estimação mais viável.

3.4 A busca por um termo de penalidade ótimo

Acabamos de mostrar que, não apenas com o uso do *BIC*, mas utilizando o critério *EDC*, mais geral, que permite diversas combinações de funções γ e sequências $\{c_n\}$ para compor o termo de penalidade do critério, é possível realizar uma estimação fortemente consistente da Partição Mínima do Modelo de Markov com Partição. Desse modo, é natural questionar se há vantagens ou desvantagens na escolha dentre esses diferentes termos de penalidades.

No contexto de estimação da ordem de uma Cadeia de Markov de ordem r finita, desconhecida, Dorea (2008), em [7], obteve a consistência forte do estimador obtido utilizando-se do *EDC*, sob condições mais fracas do que (3.3), dadas por

$$\limsup_{n\to\infty} \frac{c_n}{n} = 0 \quad \text{e} \quad \liminf_{n\to\infty} \frac{c_n}{\ln(\ln(n))} \geq \frac{2|A|}{|A|-1} \tag{3.26}$$

e propôs o termo de penalidade ótimo para o $EDC(k)$, descrito em (3.2), considerando

$$\gamma(k) = \frac{|A|^k(|A|-1)}{2} \quad \text{e} \quad c_n = \frac{2|A|}{|A|-1}\ln(\ln(n))\,. \tag{3.27}$$

Em [7], argumenta-se teoricamente que, em pequenas amostras, o *BIC* pode apresentar uma tendência de subestimar a ordem da cadeia, indicando que seu termo de penalidade poderia estar exagerando na penalização de cadeias com ordens não pequenas.

Nesse sentido, buscou-se um termo de penalidade mais brando, porém, que ainda satisfizesse condições para que a estimação apresentasse uma consistência forte, de modo que, para $k < r$, tivéssemos $EDC(k) - EDC(r) < BIC(k) - BIC(r)$. Assim, se $EDC(k) > EDC(r)$, o que levaria, em geral, a uma subestimação da ordem da cadeia pelo *EDC*, então $BIC(k) > BIC(r)$ e o *BIC* também subestimaria. Porém, o contrário não valeria necessariamente, ou seja, poderia acontecer que $BIC(k) > BIC(r)$, o que acarretaria uma subestimação da ordem da cadeia pelo *BIC*, mas pelo *EDC* isso não ocorrer, já que seria possível $EDC(k) \leq EDC(r)$.

Nesse sentido, dentre as possíveis escolhas das funções γ, estritamente crescentes, e de c_n satisfazendo (3.26), Dorea (2008) propõe então as escolhas em (3.27) para obter um termo de penalidade ótimo como sendo o mais brando possível.

Tal comportamento é observado, não só teoricamente em [7], mas também por meio de simulações numéricas em [8]. Neste último trabalho observou-se, em geral, nos casos estudados,

que tanto o *EDC* com o termo de penalidade ótimo quanto o *BIC* tendem a subestimar a ordem da cadeia em pequenas amostras, entretanto, o *EDC* comete tal equívoco de estimação numa proporção menor de simulações mantendo-se fixo um mesmo tamanho de amostra. Além disso, o *EDC* com tal termo, alcançava uma proporção alta de acerto na estimação de maneira mais rápida que o *BIC*, isto é, com um menor tamanho amostral.

Considerando os resultados apresentados em [7] e [22], é natural pensarmos em obter conclusões semelhantes para o *EDC* adaptado ao contexto de Cadeias de Markov com Partição, que apresentamos na Seção 3.1.

Nesse contexto, faz-se necessária a análise de alguns aspectos específicos para a adaptação da escolha do termo de penalidade ótimo. Observamos, por exemplo, que ao estimar a ordem da cadeia, os possíveis erros são subestimação ou superestimação da ordem verdadeira. Já ao estimar uma partição de um conjunto, até poderíamos estabelecer um paralelo de subestimação e superestimação em relação a cardinalidade da partição, ou seja, estimar partições com mais ou menos elementos que a Partição Mínima da cadeia. Porém, existem várias partições distintas, de um mesmo conjunto, com o mesmo número de partes.

Além disso, na aplicação ao modelo de padrões de navegação na internet apresentado na Seção 4 de [13], além do Algoritmo (A) com o *BIC* no lugar do *EDC*, são testadas outras variações de estratégias para a busca da partição que maximiza o *BIC*. Neste sentido, é natural questionar: como essas variações se comportam com mudanças no termo de penalidade do critério *EDC*?

A análise dessas e de outras possíveis questões sobre a melhor escolha do termo de penalidade para o *EDC* e a possibilidade de obtermos o mesmo resultado do Teorema 3.3, sob as condições mais fracas (3.26), serão objeto de estudos futuros.

Finalizamos este capítulo concluindo que é possível estender o *BIC* para o *EDC* para realizar uma estimação consistente da Partição Mínima de uma Cadeia de Markov com Partição, seja teórica, tomando a partição $\mathcal{L}_n^*$ que maximiza o critério, ou prática, tomando a partição $\hat{\mathcal{L}}_n^*$ retornada pelo algoritmo de seleção de modelos apresentado.

Para tanto, estendemos a estratégia proposta por García e González-López para estimação, agora utilizando o *EDC* no lugar do *BIC*, e observamos que, em termos de demonstração, a diferença crucial reside na utilização da Proposição 3.1 no lugar da Proposição 2.1 como resultado auxiliar para demonstração do Teorema 3.1 que estende o Teorema 2.1.

Conclusão

Com este trabalho, foi possível estudar e apresentar conceitos, resultados e exemplos, já estabelecidos por García e González-López a respeito do modelo de Cadeia de Markov com Partição, buscando estabelecer os subsídios necessários para a continuidade da pesquisa no assunto.

Foi possível também, no Capítulo 3, apresentar a extensão do *BIC* para o *EDC* na estimação do modelo, mantendo sua consistência forte, sob as condições (3.3) para a sequência $\{c_n\}$ do termo de penalidade do critério.

Tal extensão foi feita, adaptando a estratégia e os resultados propostos por García e González-López em [11] e [13], para o uso do *EDC*, de forma quase direta.

Partindo de uma partição boa inicial, são unidas partes boas que compartilhem as mesmas probabilidades de transição, de forma consistente através do *EDC* (Teorema 3.1) e partes boas que não devem ser unidas, ou seja, possuem distribuições de transição diferentes, também não são unidas de forma consistente, pelo *EDC* (Corolário 3.2).

A partir disso foi possível obter a consistência teórica da estimação da Partição Mínima do modelo pelo *EDC* (Teorema 3.2) e a consistência prática (Teorema 3.3), por meio de uma partição estimadora retornada pelo Algoritmo (A), que estende o Algoritmo 3.1 de seleção de modelos apresentado em [11] para o uso do *EDC*.

A diferença crucial, em termos técnicos, para demonstração dos resultados estendidos, reside na utilização da Proposição 3.1 no lugar da Proposição 2.1 como resultado auxiliar para demonstração do principal teorema do trabalho, o Teorema 3.1.

Buscando um paralelo com o termo de penalidade ótimo do *EDC* para estimação da ordem de uma cadeia de Markov apresentado em [7], fica como sugestão para futuros trabalhos a busca por um termo de penalidade ótimo, sob algum sentido, para o *EDC* na estimação da Partição Mínima de uma Cadeia de Markov com Partição. Outra possibilidade de trabalho futuro é buscar implementar computacionalmente a estimação da Partição Mínima do modelo em bases de dados de diferentes tamanhos, com diferentes termos de penalidade do *EDC*, incluindo o *BIC*, para tornar passível uma análise comparativa de eficiência entre os critérios.

Por fim, espera-se que este trabalho possa servir como uma boa referência em português para um estudo inicial sobre o modelo de Cadeias de Markov com Partição que permita, posteriormente, um entendimento de trabalhos mais específicos sobre o tema.

Bibliografia

[1] Brémaud, P. (1988). *An Introduction to Probabilistic Modeling*. New York: Springer-Verlag.

[2] Bühlmann, P. & Wyner, A. (1999). Variable Length Markov Chains. *The Annals of Statistics*, 27, 480–513.

[3] Corander, J., Ekdahl, M., & Koski, T. (2009). Bayesian Unsupervised Learning DNA Regulatory Binding Regions. *Advances in Artificial Intelligence*, 2009.

[4] Csizár, I. (2002). Large-Scale Typicality of Markov Sample Paths and Consistency of MDL Order Estimators. *IEEE Trans. Inf. Theory*, 48, 1616–1628.

[5] Csizár, I. & Shields, P. C. (2000). The consistency of the BIC Markov order estimator. *Annals of Statistics*, 28, 1601–1619.

[6] Csizár, I. & Talata, Z. (2006). Context Tree Estimation for Not Necesarily Finite Memory Processes, via BIC and MDL. *IEEE Trans. Inf. Theory*, 52, 1007–1016.

[7] Dorea, C. (2008). Optimal Penalty Term for EDC Markov Chain Order Estimator. *Ann de l'Inst Stat l'Univ de Paris*, 52, 15–26.

[8] Dorea, C., Resende, P., & Gonçalves, C. (2015). Comparing the Markov Order Estimators AIC, BIC and EDC. In H. Kim, M. Amouzegar, & S. Ao (Eds.), *Transactions on Engineering Technologies* (pp. 41–54). Dordrecht, Netherlands: Springer.

[9] Fernandez, M., García, J. E., & González-López, V. A. (2015). Multivariate Markov chain predictions adjusted with copula models. *New Trends in Stochastic Modeling and Data Analysis*, 1, 389–394.

[10] García, J. E. & Fernandez, M. (2013). Copula based Model Correction for Bivariate Bernoulli Financial Series. *AIP Conference Proceedings*, 1558, 1487–1490.

[11] García, J. E. & González-López, V. A. (2010). Minimal Markov Models. *arXiv:1002.0729v1*.

[12] García, J. E. & González-López, V. A. (2013). Detecting Regime Changes in Markov Models. In *Proceedings of the Sixth Workshop on Information Theoretic Methods in Science and Engineering* Tokyo, Japan.

[13] García, J. E. & González-López, V. A. (2017). Consistent Estimation of Partition Markov Models. *Entropy*, 19, 1–15.

[14] García, J. E., González-López, V. A., & Hirsh, I. D. (2015). Copula-Based Prediction of Economic Movements. In *Proceedings of the 13th International Conference of Numerical Analysis and Applied Mathematics 2015 (ICNAAM 2015)*, volume 1738 (pp. 140005(1–4)). Rhodes, Greece.

[15] Grimmett, G. & Stirzaker, D. (1992). *Probability and Random Processes*. Oxford: Oxford University Press, 2nd edition.

[16] Hoel, P. G., Port, S. C., & Stone, C. J. (1972). *Introduction to Stochastic Processes*. Boston: Houghton Mifflin.

[17] Johnson, O. (2004). *Information Theory and The Central Limit Theorem*. London: Imperial College Press.

[18] Kannan, D. (1979). *An Introduction to Stochastic Processes*. New York: North-Holland.

[19] Katz, R. W. (1981). On Some Criteria for Estimating the Order of a Markov Chain. *Technometrics*, 23, 243–249.

[20] Martell, D. L. (1999). A Markov Chain Model of Day to Day Changes in the Canadian Forest Fire Weather Index. *International Journal of Wildland Fire*, 9, 265–273.

[21] Mező, I. (2020). *Combinatorics and Number Theory of Counting Sequences*. Boca-Raton: CRC Press.

[22] Resende, P. (2009). Análise Comparativa de Estimadores da Ordem de Cadeias de Markov. Master's thesis, Universidade de Brasília.

[23] Zhao, L. C., Dorea, C., & Gonçalves, C. (2001). On Determination of the Order of a Markov Chain. *Statistical Inference for Stochastic Processes*, 4, 273–282.