Análise Comparativa de Estimadores da Ordem de Cadeias de Markov

por

Paulo Angelo Alves Resende

Brasília – DF

2009

Análise Comparativa de Estimadores da Ordem de Cadeias de Markov

Dissertação apresentada ao Programa de Pós-Graduação em Matemática da Universidade de Brasília (UnB), como requisito parcial para obtenção do grau de MESTRE EM MATEMÁTICA.

por

Paulo Angelo Alves Resende

Orientador:
Cátia Regina Gonçalves

UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE MATEMÁTICA

Brasília – DF

2009

À minha mãe, Angela Maria.

Resumo

Neste trabalho estudamos o estimador da ordem de cadeias de Markov usando o critério *EDC* (*Efficient Determination Criterion*) com o termo de penalidade ótimo proposto por Dorea (2008). Realizamos uma análise comparativa das performances dos estimadores EDC_{opt}, *BIC* e *AIC*, baseada nos resultados de simulações computacionais realizadas.

Abstract

In what follows we study and analyze the Markov chain order estimator EDC (Efficient Determination Criterion) with the penalty function proposed by Dorea (2008). We also carry out extensive numerical simulations based on EDC, BIC and AIC, aiming to a detailed comparison of their features as well as their relative performance.

Sumário

Introdução

Os processos markovianos, em geral, vêm sendo utilizados como modelos aplicados em diversas áreas, tais como: economia (Silos 2006), geologia (Li 2007), ecologia (Balzter 2000), genética (Nuel 2007), meteorologia (Martell 1999), ciência da informação (Benoît 2005) e música (McAlpine, Miranda & Hoggar 1999). Uma boa parte dessas aplicações são tacitamente modeláveis usando Cadeias de Markov de ordem superior com espaços de estados finitos. Os casos onde os espaços de estados não são finitos são naturalmente aproximados para o caso discreto/finito em função das limitações computacionais e a necessidade de simplificação do modelo.

Em linhas gerais, uma Cadeia de Markov de ordem r caracteriza-se como um processo em que a informação num dado instante depende no máximo das informações nos r instantes anteriores.

Neste cenário, conhecer a ordem de dependência de um certo procedimento tem fundamental importância, não apenas para conhecer a dependência em si, mas principalmente para ser possível estimar outros parâmetros e encontrar a Cadeia de Markov superior que melhor se adapta, em certo sentido, ao problema em análise. Dessa forma, a questão da estimação da ordem de dependência surge como um problema natural e inevitável.

Soma-se ao problema de estimação da ordem, a limitação dos tamanhos das amostras em algumas aplicações, como por exemplo em sequências de tRNA[1] que possuem comprimento entre 74 a 95 aminoácidos (Lewin 2004) e em partituras musicais que são limitadas a poucas páginas.

Bartlett (1951) publicou um dos primeiros trabalhos sobre o problema de estimação da ordem de uma Cadeia de Markov, propondo um teste de hipóteses para testar a ordem máxima da cadeia. Seguindo a mesma linha, seu trabalho foi generalizado/aperfeiçoado por

[1] tRNA (Transfer RNA): Responsável por transportar aminoácidos para a síntese de proteínas.

Hoel (1954), Good (1955), Anderson & Goodman (1957) e Billingsley (1961).

Várias alternativas às técnicas de testes de hipóteses têm sido propostas. Tong (1975) propôs a aplicação do Critério de Informação de Akaike (AIC), apresentado por Akaike (1974) para seleção de modelos, para a determinação da ordem de uma Cadeia de Markov, com espaço de estados finito e assumindo a existência de um limitante superior conhecido para a ordem.

Basicamente, Akaike (1974) considerou o problema da seleção de um modelo, dentre K modelos possíveis, que melhor se aproxima do modelo verdadeiro e propôs um novo critério de informação que tem como base a informação média de Kullback-Leibler (Kullback 1959) e a razão de verossimilhança de Neyman-Pearson (vide Kendall, Stuart & Ord (1991) e Shao (2007)).

Apesar da indiscutível importância do trabalho de Akaike (1974) e da utilização do estimador AIC, como sugerido por Tong (1975), para estimação da ordem de cadeias em modelos de dados meteorológicos por Gates & Tong (1976) e Chin (1977), não se conhecia nenhuma demonstração rigorosa sobre as propriedades do procedimento AIC neste caso. Finalmente Katz (1981) derivou formalmente a distribuição assintótica do estimador AIC e demonstrou sua inconsistência para estimar a ordem de uma Cadeia de Markov. Nesse mesmo trabalho foi proposto, como alternativa fracamente consistente, um estimador baseado no Critério de Informação Bayesiano (BIC)[2], que foi um critério de informação criado por Schwarz (1978) para seleção de modelos, usando argumentos bayesianos. O critério proposto, basicamente, foi uma adaptação no termo de penalidade do AIC.

Vale ressaltar que, em um trabalho similar ao Katz, Shibata (1976) também demonstrou a inconsistência do estimador AIC para a ordem de processos auto-regressivos.

Csiszar & Shields (2000) demonstraram a consistência forte do estimador BIC sem a hipótese que a ordem desconhecida seja limitada.

Simultaneamente, Zhao, Dorea & Gonçalves (2001) generalizaram os estimadores AIC e BIC para a estimação da ordem r de uma Cadeia de Markov $X = \{X_n\}$ com espaço de estados finito E, apresentando o estimador EDC (Critério de Informação Eficiente) baseado na log-verossimilhança máxima e com certa liberdade para a escolha do termo de penalidade.

[2]Também conhecido por *Schwarz Information Criterion* (SIC).

Especificamente, de acordo com o critério proposto por Zhao et al, a ordem r é estimada por $\hat{r}_{EDC}$ definido por

$$\hat{r}_{EDC} = \text{argmin}\,\{EDC(k); k = 0, \ldots, K\} \tag{1}$$

e

$$EDC(k) = -2log\,\hat{L}(k) + \gamma(k)c_n, \tag{2}$$

onde $\hat{L}(k)$ é a função de máxima verossimilhança da amostra $(X_1, \ldots, X_n)$ da cadeia X, $\{c_n\}$ pode ser tomada como uma sequência de números positivos e $\gamma(k)$ pode ser qualquer função crescente em k.

No caso particular $c_n = 2$, $\gamma(k) = |E|^k(|E| - 1)$, o estimador EDC reduz-se ao estimador AIC, proposto por Akaike. No caso $c_n = log\,n$ e $\gamma(k) = |E|^k(|E| - 1)$ temos o BIC.

Sob a hipótese da existência de um limitante superior K, conhecido, para a ordem r e assumindo que a sequência $\{c_n\}$ satisfaz:

$$\frac{c_n}{log\,log\,n} \to \infty \;\; e \;\; \frac{c_n}{n} \to 0,$$

Zhao et al provaram a consistência forte do estimador EDC. Como casos particulares, obtiveram a consistência forte do estimador BIC e a inconsistência do AIC.

Posteriormente, Lopes (2005) estendeu o EDC para o caso de espaço de estados E enumerável. Dorea & Lopes (2006) derivaram taxas de convergência para o estimador EDC e Dorea & Zhao (2004) obtiveram limitantes exponenciais para a probabilidade de erro do estimador EDC.

Com a ampla possibilidade de escolha do termo de penalidade do EDC produzindo estimadores consistentes da ordem da cadeia, uma questão natural é a escolha do melhor termo de penalidade, ou seja, aquele que produziria um estimador consistente que, de certa forma, teria maior chance de acerto, ou ainda, a melhor performance.

Dorea (2008) demonstrou a consistência forte dos estimadores EDC sem assumir a existência de um limitante superior finito, K, da ordem e sob condições mais fracas sobre a

sequência $\{c_n\}$:

$$\liminf_{n\to\infty} \frac{c_n}{log\ log\ n} \geq \frac{2|E|}{|E|-1} \quad e \quad \lim_{n\to\infty} \frac{c_n}{n} = 0.$$

Além disso, propôs como estimador consistente ótimo, dentre a classe (1) considerada, aquele baseado no critério

$$EDC_{opt}(k) = -2log\ \hat{L}(k) + 2|E|^{k+1} log\ log\ n\ . \tag{3}$$

Ou seja, propôs a escolha de $\gamma(k) = |E|^k(|E|-1)$ e $c_n = \frac{2|E|}{|E|-1} log\ log\ n$ em (2).

Dorea mostrou teoricamente que a escolha do termo de penalidade em (3) produz um estimador consistente melhor do que o BIC.

Nosso interesse neste trabalho é fazer uma análise comparativa da performance desses estimadores através de simulações numéricas.

Inicialmente, no Capítulo 1, apresentamos um breve histórico, uma descrição mais minuciosa dos estimadores e estudamos em detalhes o trabalho de Dorea (2008), que nos auxiliará na análise da performance dos estimadores.

No Capítulo 2 descrevemos primeiramente os experimentos computacionais realizados com o objetivo de comparar a eficiência dos estimadores consistentes EDC_{opt} e BIC, e de analisar a performance do estimador não consistente AIC. Em seguida, apresentamos uma discussão, pautada na teoria estudada, sobre os resultados obtidos nas simulações, onde verificamos principalmente que o estimador ótimo EDC_{opt} apresenta uma performance substancialmente melhor que o BIC, e essa vantagem aumenta em função da complexidade do modelo considerado. Encerramos o capítulo, com a aplicação desses estimadores, num cenário real, na análise de uma peça musical.

Finalmente, apresentamos nossas conclusões sobre o trabalho realizado.

As informações sobre o programa computacional desenvolvido para as simulações, tais como ferramentas, linguagens e descrição de rotinas relevantes, estão no Apêndice A.

1 Fundamentação Teórica dos Estimadores

Neste capítulo nós consideramos a classe de estimadores EDC (Critério de Informação Eficiente) da ordem de Cadeias de Markov, com espaço de estados finito, baseados na log-verossimilhança máxima penalizada, que foi proposta por Zhao, Dorea & Gonçalves (2001) e que generaliza os estimadores clássicos AIC e BIC.

Na seção 1.1 apresentamos um breve histórico e uma descrição desses estimadores.

Na seção 1.2 estudamos em detalhes o trabalho de Dorea (2008), onde a consistência forte desses estimadores é demonstrada sob condições mais suaves do que as assumidas em Zhao, Dorea & Gonçalves (2001) e um termo de penalidade ótimo é proposto de tal forma a obter um estimador fortemente consistente de melhor performance.

1.1 Descrição e Breve Histórico

Considere uma Cadeia de Markov $X = \{X_n\}_{n\geq 1}$, de ordem desconhecida r, com espaço de estados $E = \{1,2,\ldots,N\}$ e probabilidades de transição

$$p(a_{r+1}|a_1^r) = P(X_{n+1} = a_{r+1}|X_{n-r+1}^n = a_1^r) = P(X_{n+1} = a_{r+1}|X_{n-r+1} = a_1,\ldots,X_n = a_r), \tag{1.1}$$

onde consideramos a notação

$$a_1^r = a_1^k a_k^r = (a_1,\ldots,a_r),\ se\ 1 \leq k \leq r.$$

Dada uma certa amostra $X_1^n = (X_1, \ldots, X_n)$ desta cadeia, o problema consiste em determinar a ordem r do processo, baseado nesta amostra.

Como hipótese inicial, assume-se a existência de um limitante superior conhecido para r, isto é

$$existe\, K\ (conhecido)\ \ tal\ que\ \ 0 \leq r \leq K. \tag{1.2}$$

Inicialmente, assumindo (1.2), foi proposto por Bartlett (1951) e Hoel (1954), utilizar testes de hipóteses para a determinação da ordem da cadeia.

O teste proposto por Bartlett testa a hipótese de que a cadeia tenha ordem máxima k, enquanto que Hoel testa a hipótese de que a Cadeia de Markov em questão tenha ordem máxima $k-1$ contra a hipótese de que a cadeia tenha ordem máxima k.

O teste de Hoel é baseado na estatística da razão de verossimilhança de Neyman-Pearson (vide, por exemplo, Shao (2007)) para testar hipóteses compostas:

$$\lambda = \frac{\hat{L}(k-1)}{\hat{L}(k)},$$

onde $\hat{L}(k)$ é a máxima verossimilhança estimada considerando verdadeira a hipótese $r = k$, dada por:

$$\hat{L}(k) = \prod_{a_1^{k+1}} \hat{p}(a_{k+1}|a_1^k)^{N(a_1^{k+1}|X_1^n)}, \tag{1.3}$$

assumindo $0^0 = \frac{0}{0}^0 = 1$, e

$$N(a_1^k|X_1^n) = \sum_{j=1}^{j=n-k+1} 1(X_j = a_1, \ldots, X_{j+k-1} = a_k). \tag{1.4}$$

Na sequência, $\hat{p}(a_{k+1}|a_1^k)$ é o estimador de máxima verossimilhança de (1.1). Usando a semelhança de (1.3) com o modelo multinomial (Anderson & Goodman 1957), ou uma simples verificação usando multiplicadores de Lagrange (Billingsley 1961), obtém-se

$$\hat{p}(a_{k+1}|a_1^k) = \frac{N(a_1^{k+1}|X_1^n)}{N(a_1^k|X_1^n)}. \tag{1.5}$$

Hoel (1954), supondo verdadeira a hipótese nula $H_0 : r = k-1$, verificou que

$$-2\log(\lambda) \sim \chi^2(|E|^{k-1}(|E|-1)^2). \tag{1.6}$$

Isto é, $-2log(\lambda)$ possui uma distribuição assintótica qui-quadrado com $|E|^{k-1}(|E|-1)^2$ graus de liberdade, onde $|E|$ é a cardinalidade do conjunto E. Para isso, utilizou a aproximação normal para distribuições multinomiais.

Tong (1975) propõe a aplicação do Critério de Informação de Akaike (AIC), apresentado por Akaike (1974) para seleção de modelos, para a determinação da ordem de uma Cadeia de Markov com espaço de estados finito.

Em linhas gerais, em seu trabalho, Akaike questiona a utilidade prática dos procedimentos de testes de hipóteses como métodos para a construção ou identificação de um modelo estatístico. Considerando o problema da seleção de um dos modelos $M_1, \ldots, M_K$ que melhor se aproxima do modelo verdadeiro M_r, Akaike propõe um novo critério de informação que tem como base a informação média de Kullback-Leibler (Kullback & Leibler (1951) e Kullback (1959)). Para a estimação desta divergência são utilizadas as propriedades assintóticas da razão de verossimilhança de Neyman-Pearson para testar hipóteses compostas e de estimadores de máxima verossimilhança (vide Billingsley (1961), Kendall, Stuart & Ord (1991) ou Rao (1973)).

Conforme sugerido por Tong (1975) e seguindo Lopes (2005), o problema de estimação da ordem de uma Cadeia de Markov de ordem desconhecida r, com espaço de estados finito e assumindo a hipótese (1.2), pode ser inserido no contexto de seleção de modelos da seguinte forma: denota-se por M_k a classe de processos estocásticos $X = \{X_n\}_{n\geq 1}$, com espaço de estados $E = \{1, 2, \ldots, N\}$, para o qual existe $k \geq 1$ tal que para todo $n \geq k$ $P(X_1 = a_1, \ldots, X_{n-1} = a_{n-1}, X_n = a_n) = P(X_1 = a_1, \ldots, X_k = a_k) \prod_{j=1}^{n-k} p_{a_j a_{j+1} \ldots a_{j+k-1}; a_{j+k}}$, para a matriz de transição apropriada $P = \left(p_{a_1 \ldots a_k; a_{k+1}}\right)$, onde $p_{a_1 \ldots a_k; a_{k+1}} = p_{a_1^k; a_{k+1}} = p(a_{k+1}|a_1^k)$, como denotado em (1.1). A classe de processos i.i.d. é denotada por M_0.

Desta maneira, a ordem de uma cadeia $X = \{X_n\}_{n\geq 1}$ em $M = \underset{k}{\cup} M_k$ é o menor inteiro r tal que, para algum $l \geq 1$, $X = \{X_n\}_{n\geq l}$ está em M_r.

Baseado numa amostra $X_1^n = (X_1, \ldots, X_n)$ de uma cadeia $X = \{X_n\}$ de ordem desconhecida r, pode-se estimar r selecionando-se a classe do modelo $M_{\hat{r}}$ em $M = \cup_k M_k$ que melhor se ajusta à M_r.

Assumindo (1.2), ou seja, $r \leq K$ (K conhecido), e admitindo que cada hipótese H_k : $\{a\ cadeia\ de\ Markov\ é\ de\ ordem\ k\}$ represente o modelo M_k, com matriz $P = \left(p_{a_1}^k; a_{k+1}\right)$ associada, deseja-se, então, selecionar sobre $M = \{M_0, M_1, \ldots, M_k\}$ o modelo $M_{\hat{r}}$ que melhor se ajusta a M_r.

Sob a hipótese H_k, com $k = 0, 1, \ldots, K$, a função de máxima verossimilhança é dada por (1.3), (1.4) e (1.5), ou seja,

$$\hat{L}(k) = \prod_{a_1^{k+1}} \left[\frac{N(a_1^{k+1}|X_1^n)}{N(a_1^k|X_1^n)} \right]^{N(a_1^{k+1}|X_1^n)},$$

onde $N(a_1^k|X_1^n)$, dado em (1.4), representa o número de ocorrências de a_1^k na amostra $(X_1, \ldots, X_n)$ e no caso $k = 0$ interpreta-se $N(a_1^k|X_1^n) = n$.

Assim, baseado nesta estatística, o Critério de Informação de Akaike, utilizado por Tong (1975) para selecionar a ordem que melhor se ajusta à ordem verdadeira r da cadeia é

$$AIC(k) = -2log\,\hat{L}(k) + 2\gamma(k), \tag{1.7}$$

onde $\gamma(k) = |E|^k(|E| - 1)$ é o número de parâmetros livres a serem estimados em H_k. A estimativa $\hat{r}$ de r é aquela que minimiza $AIC(k)$, dentre $k = 0, 1, \ldots, K$, ou seja,

$$\hat{r}_{AIC} = \text{argmin}\,\{AIC(k); k = 0, 1, \ldots, K\}. \tag{1.8}$$

Uma fundamentação mais detalhada dos trabalhos de Akaike e Tong pode ser encontrada em Lopes (2005).

Posteriormente, Katz (1981) obteve a distribuição assintótica do estimador AIC e mostrou

sua inconsistência para a estimação da ordem da cadeia, com a existência de uma probabilidade positiva de superestimar a ordem. Como uma alternativa ao procedimento AIC, Katz sugere o uso do Critério de Informação Bayesiano (BIC) proposto por Schwarz (1978) para a estimação da dimensão de um modelo.

O estimador BIC da ordem r de uma Cadeia de Markov X, sob a hipótese (1.2) e baseado numa amostra $X_1^n = (X_1, \ldots, X_n)$, pode ser descrito como

$$\hat{r}_{BIC} = \text{argmin}\,\{BIC(k); k = 0, 1, \ldots, K\},$$

onde

$$BIC(k) = -2log\,\hat{L}(k) + \gamma(k) log\, n\,,$$

com $\hat{L}(k)$ e $\gamma(k)$ definidos como no critério AIC.

Com a substituição da constante 2, no termo de penalidade do AIC em (1.7), pelo fator $log\, n$, que depende do tamanho da amostra e converge ao infinito a uma taxa suficientemente lenta, foi possível obter a consistência fraca do estimador $\hat{r}_{BIC}$, demonstrada por Katz (1981). No entanto, foi apontado por Katz, através de alguns experimentos computacionais modestos, a tendência do estimador BIC de subestimar a ordem da cadeia.

Mesmo depois dos trabalhos de Schwarz (1978) e Katz (1981), ficaram duas questões em aberto – a consistência forte do BIC e a possibilidade de se obter termos de penalidade "melhores". Csiszar & Shields (2000) responderam a primeira questão apresentando uma demonstração da consistência forte do estimador BIC, sem assumir a priori a existência de um limitante superior da ordem [hipótese (1.2)], mas deixaram explicitamente a segunda questão em aberto: *"it remains open whether smaller penalty terms suffice for consistency..."*.

Paralelamente, Zhao, Dorea & Gonçalves (2001), propuseram o estimador EDC (*Efficient Determination Criterion*) com uma certa liberdade para a escolha do termo de penalidade e incluíndo como casos particulares os estimadores AIC e BIC. Especificamente, r será estimado por $\hat{r}_{EDC}$, a estimativa mínima de EDC, ou seja,

$$\hat{r} = \text{argmin}\,\{EDC(k); k = 0, \ldots, K\}, \tag{1.9}$$

onde

$$EDC(k) = -2log\,\hat{L}(k) + \gamma(k)c_n, \tag{1.10}$$

com c_n podendo ser tomada como uma sequência de números positivos dependendo de n (ou, mais geral, como uma sequência de variáveis aleatórias positivas) e $\gamma(k)$ podendo ser tomada como qualquer função crescente em k.

Nos casos particulares: $c_n = 2$, $\gamma(k) = |E|^k(|E|-1)$ e $c_n = log\,n$, $\gamma(k) = |E|^k(|E|-1)$ temos os critérios AIC e BIC respectivamente.

Zhao, Dorea & Gonçalves (2001) provaram a consistência forte do estimador $\hat{r}_{EDC}$ para estimar a ordem r de uma Cadeia de Markov $X = \{X_n\}$, cujo processo derivado

$$\left\{Y^{(k)} = (X_n, \ldots, X_{n+k-1})\right\}_{k\geq 1}$$

é irredutível e recorrente positivo, assumindo a hipótese (1.2) e sob as seguintes condições para a sequência $\{c_n\}$ no termo de penalidade:

$$\frac{c_n}{log\,log\,n} \to \infty \;\; e \;\; \frac{c_n}{n} \to 0. \tag{1.11}$$

Em particular, obtiveram a consistência forte de $\hat{r}_{BIC}$.

Além disso, observaram que se ao invés de (1.11) assumirmos que $\{c_n\}$ é uniformemente limitada por uma constante, então $\hat{r}_{EDC}$ é inconsistente. Este é o caso do estimador $\hat{r}_{AIC}$.

Com isso, qualquer c_n satisfazendo as condições em (1.11), dá origem a um estimador fortemente consistente. Dessa forma, é natural pensar em qual "c_n" fornece o estimador com maior chance de acerto.

Recentemente, Dorea (2008), considerando $\gamma(k) = |E|^k(|E|-1)$, propôs o seguinte termo de penalidade como sendo ótimo dentro dessa classe de estimadores consistentes:

$$c_n = \frac{2|E|}{|E|-1} log\ log\ n\ .$$

Além disso, Dorea (2008) demonstrou, sob algumas condições de regularidade sobre X, a consistência forte do estimador EDC sem a hipótese (1.2) de limitação da ordem e assumindo as seguintes hipóteses (mais fracas) sobre c_n:

$$\liminf_{n\to\infty} \frac{c_n}{log\ log\ n} \geq \frac{2|E|}{|E|-1} \quad e \quad \limsup_{n\to\infty} \frac{c_n}{n} = 0.$$

Em particular, Dorea apresentou uma prova alternativa a de Csiszár-Shields (2000) para a consistência forte do estimador BIC sem a limitação (1.2).

1.2 EDC: Consistência e Termo de Penalidade Ótimo

Nesta seção, consideramos a classe de estimadores EDC, dados por (1.9) e (1.10), com $\gamma(k) = |E|^k(|E|-1)$ e $c_n > 0$ uma sequência de constantes, proposto por Zhao, Dorea & Gonçalves (2001) e que generaliza os estimadores AIC e BIC.

Como mencionamos no final da seção anterior, Dorea (2008) aborda a questão da escolha do termo de penalidade ótimo e ainda demonstra a consistência forte do estimador EDC, sob condições suaves de regularidade, sem a hipótese (1.2). A seguir apresentamos um estudo detalhado de seu trabalho.

1.2.1 Notações e Resultados Auxiliares

Suponha $X = \{X_n\}_{n\geq 1}$, uma Cadeia de Markov de ordem r, com probabilidades de transição

$$p(a_{r+1}|a_1^r) = P(X_{n+1} = a_{r+1}|X_{n-r+1}^n = a_1^r). \tag{1.12}$$

Para $k \geq r$, considere o processo $Y^{(k)} = \left\{Y_n^{(k)}\right\}_{n\geq 1}$, com $Y_n^{(k)} = (X_n, \ldots, X_{n+k-1}) \in E^k$.

Considerando $A_i = (a_{i,1}, \ldots, a_{i,k}) \in E^k$, temos que [1]

$$\begin{aligned}
&\mathrm{P}(\mathrm{Y}_{n+1}^{(k)} = A_{n+1} | Y_n^{(k)} = A_n, \ldots, Y_1^{(k)} = A_1) = \\
&= \mathrm{P}(\ (\mathrm{X}_{n+1}, \ldots, X_{n+k}) = (a_{n+1,1}, \ldots, a_{n+1,k}) | \\
&\qquad (\mathrm{X}_n, \ldots, X_{n+k-1}) = (a_{n,1}, \ldots, a_{n,k}), \ldots, (X_1, \ldots, X_{k-1}) = (a_{1,1}, \ldots, a_{1,k})).
\end{aligned}$$

Considerando apenas os casos possíveis, isto é, $a_{i,j} = a_{i-1,j+1}$, e denotando $a_{i+j-1} = a_{i,j}$, então

$$\begin{aligned}
&\mathrm{P}(\mathrm{Y}_{n+1}^{(k)} = A_{n+1} | Y_n^{(k)} = A_n, \ldots, Y_1^{(k)} = A_1) = \\
&= \mathrm{P}(\mathrm{X}_{n+1} = a_{n+1} | X_1^n = a_1^n) \\
&= \mathrm{P}(\mathrm{X}_{n+1} = a_{n+1} | X_{n-r+1}^n = a_{n-r+1}^n) \\
&= \mathrm{P}(\mathrm{X}_{n-r+2}^{n+1} = a_{n-r+2}^{n+1} | X_{n-r+1}^n = a_{n-r+1}^n) \\
&= \mathrm{P}(\mathrm{Y}_{n+1}^{(k)} = A_{n+1} | Y_n^{(k)} = A_n).
\end{aligned}$$

Assim, concluímos que $Y^{(k)}$ é uma Cadeia de Markov homogênea de primeira ordem, com probabilidades de transição

$$P(Y_{n+1}^{(k)} = a_2^{k+1} | Y_n^{(k)} = a_1^k) = p(a_{k+1} | a_1^k) = p(a_{k+1} | a_{k-r+1}^k). \tag{1.13}$$

Assim, se $X = \{X_n\}_{n \geq 1}$ é uma Cadeia de Markov de ordem r e espaço de estados E, então, para $k \geq r$ o processo $Y^{(k)} = \left\{ Y_n^{(k)} \right\}_{n \geq 1}$, onde $Y_n^{(k)} = (X_n, \ldots, X_{n+k-1}) \in E^k$, é chamado Cadeia de Markov k-derivada de X.

Podemos induzir a recorrência e aperiodicidade nas cadeias derivadas da seguinte forma:

Proposição 1.1. *Se as probabilidades de transição da Cadeia de Markov X, de ordem r, são estritamente positivas e $k \geq r$, então a cadeia k-derivada $Y^{(k)}$ é irredutível e aperiódica. Consequentemente, ergódica.*

Demonstração. Como o espaço de estados, E, de X é finito, temos que o espaço de estados da cadeia k-derivada, E^k, também é finito. Por outro lado, para quaisquer dois estados $a^{(k)} =$

[1] Supostamente, Doob foi o primeiro a sugerir essa adaptação em Doob (1966) páginas 89 e 185.

a_1^k e $b^{(k)} = b_1^k$, temos por (1.13)

$$\begin{aligned} &\mathrm{P}(\mathrm{Y}_2^{(k)} = a_2^k b_1, \ldots, Y_k^{(k)} = a_k b_1^{k-1}, Y_{k+1}^{(k)} = b_1^k | Y_1^{(k)} = a_1^k) \\ &\quad = \mathrm{p}(\ \mathrm{b}_1 | a_1^k) \cdot \cdots \cdot p(b_k | a_k b_1^{k-1}) \\ &= \begin{cases} p(b_1|a_1^r) \ldots p(b_r|a_r b_1^{r-1}) & > 0, \quad \text{para } k = r \\ p(b_1|a_{k-r+1}^k) \ldots p(b_r|b_{k-r}^{k-1}) & > 0, \quad \text{para } k > r. \end{cases} \end{aligned}$$

Assim, todos os estados se comunicam e portanto a Cadeia de Markov é irredutível. Como o espaço de estados é finito, segue que ela é recorrente positiva e portanto ergódica (vide, por exemplo, Kannan (1979)).

Além disso, para todo $a \in E$, $A = (a, a, \ldots, a)$, temos que $P(Y_{n+1}^{(k)} = A | Y_n^{(k)} = A) > 0$, usando a irredutibilidade, concluímos que $Y^{(k)}$ é aperiódica. □

Vale observar que a recíproca da proposição anterior não é verdadeira.

Uma questão natural é a relação entre as cadeias derivadas de X. Quanto a ergodicidade, podemos ter:

Proposição 1.2. *Se a cadeia k-derivada de X é ergódica, com distribuição de equilíbrio (estacionária) $\pi_k(a_1^k)$, então a cadeia $(k+1)$-derivada possui distribuição estacionária dada por*

$$\pi_{k+1}(a_1^{k+1}) = \pi_k(a_1^k) p(a_{k+1} | a_{k-r+1}^k). \tag{1.14}$$

Demonstração. Como por hipótese $Y^{(k)}$ possui a distribuição de equilíbrio e estacionária π_k, então temos

$$\pi_k(a_1^k) = \sum_{b_1^k} \pi_k(b_1^k) p(a_1^k | b_1^k)$$

(vide, por exemplo, Kannan (1979)).

Como $p(a_1^k | b_1^k) = 0$ para $a_1^{k-1} \neq b_2^k$,

$$\begin{aligned}\pi_k(a_1^k) &= \sum_{a_0} \pi_k(a_0 a_1^{k-1}) p(a_1^k | a_0 a_1^{k-1}) \\ &= \sum_{a_0} \pi_k(a_0 a_1^{k-1}) p(a_k | a_0 a_1^{k-1}). \end{aligned} \tag{1.15}$$

Para a Cadeia de Markov $(k+1)$-derivada definimos

$$\pi_{k+1}(a_1^{k+1}) = \pi_k(a_1^k) p(a_{k+1} | a_1^k). \tag{1.16}$$

Então, substituindo (1.15) em (1.16), temos

$$\pi_{k+1}(a_1^{k+1}) = \sum_{a_0} \pi_k(a_0 a_1^{k-1}) p(a_k | a_0 a_1^{k-1}) p(a_{k+1} | a_1^k). \tag{1.17}$$

Daí novamente, aplicando (1.16) em (1.17), com um ajuste nos sub-índices, obtemos

$$\begin{aligned}\pi_{k+1}(a_1^{k+1}) &= \sum_{a_0} \pi_k(a_0 a_1^{k-1}) p(a_k | a_0 a_1^{k-1}) p(a_{k+1} | a_1^k) \\ &= \sum_{a_0} \pi_{k+1}(a_0 a_1^k) p(a_{k+1} | a_1^k) \\ &= \sum_{b_0^k} \pi_{k+1}(b_0^k) p(a_{k+1} | b_0^k). \end{aligned}$$

Logo π_{k+1} dada por (1.14) é uma distribuição estacionária para $Y^{(k+1)}$. □

Para simplificar a notação, vamos utilizar

$$\pi(a_1^k) = \pi_k(a_1^k) \text{ e } \pi(a_1^{k+l}) = \pi_{k+l}(a_1^{k+l}).$$

Note que esta notação está bem definida pois, através do domínio, é possível fazer a distinção entre as distribuições.

Por indução, temos:

Corolário 1.3. *Se a cadeia r-derivada de X é ergódica então, para $l > r$, a k-derivada de X*

possui distribuição estacionária dada por

$$\pi(a_1^k) = \pi(a_1^r)p(a_{r+1}|a_1^r)\dots p(a_k|a_{k-r}^{k-1}). \tag{1.18}$$

Das proposições 1.1 e 1.2 segue:

Corolário 1.4. *Se X é uma Cadeia de Markov, de ordem r, cujas probabilidades de transição são estritamente positivas, ou seja, $p(a_{r+1}|a_1^r) > 0$, $\forall\, a_1^{r+1} \in E^{r+1}$, então para todo $k \geq r$ a cadeia $Y^{(k)}$ possui distribuição de equilíbrio (estacionária) dada por (1.18), onde $\pi(a_1^r)$ indica a distribuição de equilíbrio de $Y^{(r)}$.*

Intuitivamente vemos que uma cadeia de ordem r pode ser modelada por uma cadeia de ordem $k > r$ sem qualquer perda. Esse resultado (Corolário 1.4) mostra que a ergodicidade é preservada neste caso.

O Lema abaixo é necessário para detalhar os resultados de Dorea (2008). Embora seja um resultado simples, é de grande importância, pois com ele é possível relacionar as diversas formas de contagem de uma determinada sequência de eventos.

Lema 1.5. *Considerando a notação $N(a_1^k) = N(a_1^k|X_1^n)$, definida em (1.4), temos que*

$$N(a_1^k) = \sum_{a_0} N(a_0 a_1^k) + 1(X_1 = a_1, \dots, X_k = a_k)$$

e

$$N(a_1^k) = \sum_{a_{k+1}} N(a_1^k a_{k+1}) + 1(X_{n-k+1} = a_1, \dots, X_n = a_k). \tag{1.19}$$

Mais ainda, se $l > 0$, por indução segue que:

$$N(a_1^k) = \sum_{a_{1-l}^0} N(a_{1-l}^0 a_1^k) + \sum_{i=0}^{l-1} \sum_{a_{-i}^0} 1(X_1^{i+k+1} = a_{-i}^0 a_1^k)$$

e

$$N(a_1^k) = \sum_{a_{k+1}^{k+l}} N(a_1^k a_{k+1}^{k+l}) + \sum_{i=0}^{l-1} \sum_{a_{k+1}^{k+1+i}} 1(X_{n-k-i}^n = a_1^k a_{k+1}^{k+1+i}). \tag{1.20}$$

Demonstração. Usando a definição de $N(a_1^k)$

$$\begin{aligned}
\mathrm{N}(\mathrm{a}_1^k) &= \sum_{j=1}^{j=n-k+1} 1(X_j = a_1, \ldots, X_{j+k-1} = a_k) \\
&= \left(\sum_{j=2}^{j=n-k+1} \sum_{a_0} 1(X_{j-1} = a_0, X_j = a_1, \ldots, X_{j+k-1} = a_k) \right) + 1(X_1 = a_1, \ldots, X_k = a_k) \\
&= \left(\sum_{a_0} \sum_{i=1}^{i=n-k} 1(X_i = a_0, X_{i+1} = a_1, \ldots, X_{i+k} = a_k) \right) + 1(X_1 = a_1, \ldots, X_k = a_k) \\
&= \left(\sum_{a_0} N(a_0 a_1^k) \right) + 1(X_1 = a_1, \ldots, X_k = a_k).
\end{aligned}$$

Analogamente,

$$\begin{aligned}
\mathrm{N}(\mathrm{a}_1^k) &= \sum_{j=1}^{j=n-k+1} 1(X_j = a_1, \ldots, X_{j+k-1} = a_k) \\
&= \left(\sum_{j=1}^{j=n-k} \sum_{a_{k+1}} 1(X_j = a_1, \ldots, X_{j+k-1} = a_k, X_{j+k} = a_{k+1}) \right) + 1(X_{n-k+1} = a_1, \ldots, X_n = a_k) \\
&= \left(\sum_{a_{k+1}} \sum_{j=1}^{j=n-k} 1(X_j = a_1, \ldots, X_{j+k-1} = a_k, X_{j+k} = a_{k+1}) \right) + 1(X_{n-k+1} = a_1, \ldots, X_n = a_k) \\
&= \left(\sum_{a_{k+1}} N(a_1^k a_{k+1}) \right) + 1(X_{n-k+1} = a_1, \ldots, X_n = a_k).
\end{aligned}$$

□

Em resultados subsequentes, também vamos utilizar a seguinte adaptação da desigualdade das médias.

Lema 1.6 (Desigualdade das Médias)**.** *Supondo $a_i > 0$ e $e_i > 0$, $i = 1..l$, então*

$$\frac{\sum\limits_i e_i}{\sum\limits_{i=1}^{l} e_i \frac{1}{a_i}} \leq \sqrt[\sum_i e_i]{\prod_{i=1}^{l} a_i^{e_i}} \leq \frac{\sum\limits_{i=1}^{l} e_i a_i}{\sum\limits_i e_i}.$$

Nas demonstrações dos resultados da próxima sub-seção, vamos precisar de algumas relações da função verossimilhança. Essas relações estão intimamente ligadas aos resultados apresentados, pois tratam do comportamento local de L e $\hat{L}$.

Lema 1.7. *Seja X uma Cadeia de Markov de ordem r com probabilidades de transição dadas em (1.12). Então*

(a) para $k \geq r$, $\log L(k+1) = \log L(k) + o(\delta_n)$ e $\log L(k+1) = \log L(r) + o(\delta_n)$;

(b) $\log \hat{L}(k) = \sum\limits_{a_1^{l+1}} N(a_1^{l+1}) log \frac{N(a_{1+l-k}^{l+1}|X_1^n)}{N(a_{1+l-k}^{l}|X_1^n)} + o(\delta_n)$ para todo $l \geq 0$ e $0 \leq k < l$;

(c) $\hat{L}(k+1) \geq \hat{L}(k) + o(\delta_n)$, $k \geq 0$,

onde $L(k)$ é a função verossimilhança de X supondo a ordem k e $\hat{L}$ é a máxima verossimilhança, como definida em (1.3). Além disso, $o(\delta_n)$ significa que $\frac{o(\delta_n)}{\delta_n} \to 0$ sempre que $\delta_n \to \infty$. Neste caso, $o(\delta_n)$ é limitado em n.

Demonstração. (a) Por definição

$$L(k+1) = \prod_{b_1^{k+2}} p(b_{k+2}|b_1^{k+1})^{N(b_1^{k+2})}. \tag{1.21}$$

Como $k \geq r$, temos que $p(b_{k+2}|b_1^{k+1}) = p(b_{k+2}|b_2^{k+1})$ e substituindo em (1.21) obtemos

$$L(k+1) = \prod_{b_1^{k+2}} p(b_{k+2}|b_2^{k+1})^{N(b_1^{k+2})}.$$

Agrupando adequadamente e usando o Lema 1.5 obtemos

$$\begin{aligned} L(k+1) &= \prod_{b_2^{k+2}} p(b_{k+2}|b_2^{k+1})^{\sum_{b_1} N(b_1^{k+2})} \\ &= \prod_{b_2^{k+2}} p(b_{k+2}|b_2^{k+1})^{N(b_2^{k+2})-1(b_2^{k+2}=X_1^{k+1})} \\ &= \left(\prod_{b_2^{k+2}} p(b_{k+2}|b_2^{k+1})^{N(b_2^{k+2})}\right) \cdot \left(\prod_{b_2^{k+2}} p(b_{k+2}|b_2^{k+1})^{-1(b_2^{k+2}=X_1^{k+1})}\right). \end{aligned} \quad (1.22)$$

Chamando β o segundo fator do último membro de (1.22) e usando a definição de $L(k)$ obtemos

$$L(k+1) = L(k) \cdot \beta.$$

Tomando logaritmo e considerando que β não depende de n temos

$$log\, L(k+1) = log\, L(k) + o(\delta_n).$$

(b) Por (1.3) e (1.5) temos

$$\hat{L}(k) = \prod_{b_1^{k+1}} \left(\frac{N(b_1^{k+1})}{N(b_1^k)}\right)^{N(b_1^{k+1})},$$

e pelo Lema 1.5, segue

$$\hat{L}(k) = \prod_{b_1^{k+1}} \left(\frac{N(b_1^{k+1})}{N(b_1^k)}\right)^{\sum_{b^0_{-(l-k)+1}} N(b^0_{-(l-k)+1} b_1^k) + \sum_{i=0}^{l-1} \sum_{b^0_{-i}} 1(b^0_{-i} b_1^k = X_1^{i+k+1})}.$$

Considerando $a_1^{l+1} = b^{k+1}_{-(l-k)+1}$ e $\sum\sum 1(b^0_{-i} b_1^k = X_1^{i+k+1}) = c(a_1^{l+1})$, temos

$$\hat{L}(k) = \prod_{a_1^{l+1}} \left(\frac{N(a_{1+l-k}^{l+1})}{N(a_{1+l-k}^{l})}\right)^{N(a_1^{l+1})} \prod_{a_1^{l+1}} \left(\frac{N(a_{1+l-k}^{l+1})}{N(a_{1+l-k}^{l})}\right)^{c(a_1^{l+1})}. \quad (1.23)$$

Chamando o segundo fator em (1.23) de β e tomando logaritmo, obtemos

$$log\,\hat{L}(k) = \sum_{a_1^{l+1}} N(a_1^{l+1}) log\,\frac{N(a_{1+l-k}^{l+1})}{N(a_{1+l-k}^{l})} + o(\delta_n).$$

(c) Para provar que $\hat{L}(k+1) \geq \hat{L}(k) + o(\delta_n)$, seguiremos as ideias da demonstração do Teorema 1 de Dorea & Zhao (2004), páginas 3689-3697.

Para $k \geq 0$, temos

$$log\,\hat{L}(k+1) = \sum_{a_1^{k+2}} log\,N(a_1^{k+2})\frac{N(a_1^{k+2})}{N(a_1^{k+1})}\,.$$

Como por (1.19) $N(a_1^{k+1}) = \sum_{a_{k+2}} N(a_1^{k+2}) + 1(X_{n-k} = a_1, \ldots, X_n = a_{k+1})$,

$$\begin{aligned} log\,\hat{L}(k) &= \sum_{a_1^{k+1}} log\,N(a_1^{k+1})\frac{N(a_1^{k+1})}{N(a_1^{k})} \\ &= \sum_{a_1^{k+2}} log\,N(a_1^{k+2})\frac{N(a_2^{k+2})}{N(a_2^{k+1})} + o(\delta_n). \end{aligned}$$

Daí segue

$$\begin{aligned} log\,\hat{L}(k) - log\,\hat{L}(k+1) &= \sum_{a_1^{k+2}} N(a_1^{k+2}) log\left[\frac{N(a_2^{k+2})}{N(a_2^{k+1})}\frac{N(a_1^{k+1})}{N(a_1^{k+2})}\right] + o(\delta_n) \\ &= \sum_{a_1^{k+2}} N(a_1^{k+1})\frac{N(a_1^{k+2})}{N(a_1^{k+1})} log\left[\frac{N(a_2^{k+2})}{N(a_2^{k+1})}\frac{N(a_1^{k+1})}{N(a_1^{k+2})}\right] + o(\delta_n) \\ &= \sum_{a_1^{k+1}} N(a_1^{k+1}) \sum_{a_{k+2}} \frac{N(a_1^{k+2})}{N(a_1^{k+1})} log\left[\frac{N(a_2^{k+2})}{N(a_2^{k+1})}\frac{N(a_1^{k+1})}{N(a_1^{k+2})}\right] + o(\delta_n). \end{aligned}$$

Note que $\sum_{a_{k+2}} \frac{N(a_1^{k+2})}{N(a_1^{k+1})} < \infty$. Assim, usando a desigualdade de Jensen temos:

$$\begin{aligned}
&\sum_{a_{k+2}} \frac{N(a_1^{k+2})}{N(a_1^{k+1})} log \left[\frac{N(a_2^{k+2})}{N(a_2^{k+1})} \frac{N(a_1^{k+1})}{N(a_1^{k+2})} \right] + o(\delta_n) \\
&\qquad \leq \quad log \left[\sum_{a_{k+2}} \frac{N(a_1^{k+2})}{N(a_1^{k+1})} \frac{N(a_2^{k+2})}{N(a_2^{k+1})} \frac{N(a_1^{k+1})}{N(a_1^{k+2})} \right] + o(\delta_n) \\
&\qquad \leq \quad log \left[\sum_{a_{k+2}} \frac{N(a_2^{k+2})}{N(a_2^{k+1})} \right] + o(\delta_n) \\
&\qquad \leq \quad log\, 1 + o(\delta_n) = o(\delta_n).
\end{aligned}$$

Assim, concluímos que

$$log\, \hat{L}(k) - log\, \hat{L}(k+1) \leq o(\delta_n).$$

□

Observação 1.8. *Como $\hat{L}(k)$ depende apenas da amostra, que é conhecida, podemos determinar melhor seu comportamento. Nesse sentido, o item (c) do Lema anterior pode ser mais específico:*

$$\hat{L}(k+1) \geq \hat{L}(k)$$

Demonstração. Pela definição e pela demonstração do item (b) do lema, temos respectivamente

$$\hat{L}(k+1) = \prod_{a_1^{k+2}} \left(\frac{N(a_1^{k+2})}{N(a_1^{k+1})} \right)^{N(a_1^{k+2})} \tag{1.24}$$

e

$$\begin{aligned}
\hat{L}(k) &= \prod_{b_1^{k+1}} \left(\frac{N(b_1^{k+1})}{N(b_1^k)} \right)^{N(b_1^{k+1})} \\
&= \prod_{b_1^{k+1}} \left(\frac{N(b_1^{k+1})}{N(b_1^k)} \right)^{\sum_{b_0} \left[N(b_0 b_1^{k+1}) + 1(X_1^{k+1} = b_1^{k+1}) \right]} \\
&= \prod_{b_0 b_1^{k+1}} \left(\frac{N(b_1^{k+1})}{N(b_1^k)} \right)^{N(b_0 b_1^{k+1})} \prod_{b_1^{k+1}} \left(\frac{N(b_1^{k+1})}{N(b_1^k)} \right)^{1(X_1^{k+1} = b_1^{k+1})} \\
&= \prod_{a_1^{k+2}} \left(\frac{N(a_2^{k+2})}{N(a_2^{k+1})} \right)^{N(a_1^{k+2})} \left(\frac{N(X_1^{k+1})}{N(X_1^k)} \right). \qquad (1.25)
\end{aligned}$$

Para mostrar o desejado, basta obter $\frac{\hat{L}(k+1)}{\hat{L}(k)} \geq 1$. Usando (1.24) e (1.25), temos

$$\frac{\hat{L}(k+1)}{\hat{L}(k)} = \prod_{a_1^{k+2}} \left(\frac{N(a_1^{k+2})}{N(a_1^{k+1})} \frac{N(a_2^{k+1})}{N(a_2^{k+2})} \right)^{N(a_1^{k+2})} \left(\frac{N(X_1^k)}{N(X_1^{k+1})} \right).$$

Usando o Lema 1.6, segue

$$\frac{\hat{L}(k+1)}{\hat{L}(k)} \geq \left\{ \frac{\left(\sum_{a_1^{k+2}} N(a_1^{k+2}) \right) + 1}{\sum_{a_1^{k+2}} \left[N(a_1^{k+2}) \frac{N(a_1^{k+1}) N(a_2^{k+2})}{N(a_1^{k+2}) N(a_2^{k+1})} \right] + \frac{N(X_1^{k+1})}{N(X_1^k)}} \right\}^{\left(\sum_{a_1^{k+2}} N(a_1^{k+2}) \right) + 1}. \qquad (1.26)$$

Para o numerador (e expoente) do segundo membro de (1.26), usando o Lema 1.5, temos

$$\begin{aligned}\left(\sum_{a_1^{k+2}} N(a_1^{k+2})\right)+1 &= \sum_{a_1^{k+1}}\sum_{a_{k+2}} N(a_1^{k+1}a_{k+2})+1\\ &= \sum_{a_1^{k+1}}\Big(N(a_1^{k+1})-1(X_{n-k}^n=a_1^{k+1})\Big)+1\\ &= \sum_{a_1^{k+1}} N(a_1^{k+1})-1+1\\ &= \sum_{a_1^{k+1}} N(a_1^{k+1}). \qquad (1.27)\end{aligned}$$

Da mesma forma para o denominador, obtemos

$$\begin{aligned}&\left(\sum_{a_1^{k+2}} N(a_1^{k+2})\frac{N(a_1^{k+1})N(a_2^{k+2})}{N(a_1^{k+2})N(a_2^{k+1})}\right)+\frac{N(X_1^{k+1})}{N(X_1^k)}\\ &\quad=\left(\sum_{a_1^{k+2}}\frac{N(a_1^{k+1})N(a_2^{k+2})}{N(a_2^{k+1})}\right)+\frac{N(X_1^{k+1})}{N(X_1^k)}\\ &\quad=\sum_{a_2^{k+2}}\frac{N(a_2^{k+2})}{N(a_2^{k+1})}\sum_{a_1}N(a_1a_2^{k+1})+\frac{N(X_1^{k+1})}{N(X_1^k)}\\ &\quad=\sum_{a_2^{k+2}}\frac{N(a_2^{k+2})}{N(a_2^{k+1})}\Big[N(a_2^{k+1})-1(X_1^k=a_2^{k+1})\Big]+\frac{N(X_1^{k+1})}{N(X_1^k)}\\ &\quad=\sum_{a_2^{k+2}}N(a_2^{k+2})-\sum_{a_2^{k+2}}1(X_1^k=a_2^{k+1})\frac{N(a_2^{k+2})}{N(a_2^{k+1})}+\frac{N(X_1^{k+1})}{N(X_1^k)}\\ &\quad=\sum_{a_2^{k+2}}N(a_2^{k+2})-\sum_{a_{k+2}\neq X_{k+1}}\frac{N(X_1^ka_{k+2})}{N(X_1^k)}-\frac{N(X_1^{k+1})}{N(X_1^k)}+\frac{N(X_1^{k+1})}{N(X_1^k)}\\ &\quad=\sum_{a_2^{k+2}}N(a_2^{k+2})-\sum_{a_{k+2}\neq X_{k+1}}\frac{N(X_1^ka_{k+2})}{N(X_1^k)}. \qquad (1.28)\end{aligned}$$

Aplicando (1.28) e (1.27) em (1.26), segue

$$\frac{\hat{L}(k+1)}{\hat{L}(k)} \geq \left(\frac{\sum\limits_{a_1^{k+1}} N(a_1^{k+1})}{\sum\limits_{a_2^{k+2}} N(a_2^{k+2}) - \sum\limits_{a_{k+2} \neq X_{k+1}} \frac{N(X_1^k a_{k+2})}{N(X_1^k)}} \right)^{\sum\limits_{a_1^{k+1}} N(a_1^{k+1})} \geq 1.$$

□

Também serão utilizados os próximos três teoremas, o primeiro pode ser encontrado em Dacunha-Castelle, Duflo & McHale (1986) e os outros em Meyn & Tweedie (1993).

Teorema 1.9 (Lei Forte dos Grandes Números)**.** *Suponha Z uma Cadeia de Markov irredutível, recorrente positiva, com espaço de estados finito E e distribuição estacionária π. Considere ainda $f : E \to \mathbb{R}$ e $g : E \to (0, \infty)$ contínuas, então*

$$\frac{\sum f(Z_j)}{\sum g(Z_j)} \underset{q.c.}{\to} \frac{E_\pi(f(Z_j))}{E_\pi(g(Z_j))}.$$

Teorema 1.10 (Teorema do Limite Central)**.** *Se Z é uma Cadeia de Markov ergódica com espaço de estados finito E e distribuição estacionária π, $g : E \to \mathbb{R}$, $S_n(g) = \sum\limits_{j=1}^{n} g(Z_j)$ e $\theta_g^2 = E_\pi(g^2(Z_1)) + 2 \sum\limits_{j=2}^{n} E_\pi(g(Z_1)g(Z_j)) > 0$, então*

$$\frac{S_n(g) - E_\pi(S_n(g))}{\sqrt{n\theta_g^2}} \underset{d}{\to} \mathcal{N}(0,1).$$

Teorema 1.11 (Lei do Logaritmos Iterado)**.** *Se Z é uma Cadeia de Markov ergódica com espaço de estados finito E e distribuição estacionária π, $g : E \to \mathbb{R}$, $S_n(g) = \sum\limits_{j=1}^{n} g(Z_j)$ e $\theta_g^2 = E_\pi(g^2(Z_1)) + 2 \sum\limits_{j=2}^{n} E_\pi(g(Z_1)g(Z_j))$, então*

(a) Se $\theta_g^2 = 0$, quase certamente

$$\lim_{n \to \infty} \frac{1}{\sqrt{n}} [S_n(g) - E_\pi(S_n(g))] = 0.$$

(b) Se $\theta_g^2 > 0$, quase certamente

$$\limsup_{n\to\infty} \frac{S_n(g) - E_\pi(S_n(g))}{\sqrt{2\theta_g^2 n log\, log\, n}} = 1$$

e

$$\liminf_{n\to\infty} \frac{S_n(g) - E_\pi(S_n(g))}{\sqrt{2\theta_g^2 n log\, log\, n}} = -1.$$

1.2.2 Resultados Principais

Daqui para frente, vamos assumir que X é uma Cadeia de Markov de ordem r, com espaços de estados E finito, $|E| \geq 2$ e probabilidades de transição estritamente positivas, ou seja,

$$p(a_{r+1}|a_1^r) > 0,\ \forall\ a_1^{r+1} = (a_1, \ldots, a_{r+1}) \in E^{r+1}. \tag{1.29}$$

Lembremos que pelas proposições 1.1, 1.2 e corolários 1.3 e 1.4, segue que as cadeias k-derivadas $Y^{(k)}$, $k \geq r$, são irredutíveis e ergódicas e se $\pi(a_1^r)$ é a distribuição de equilíbrio estacionária para a cadeia r-derivada $Y^{(r)}$, então para $k > r$ a k-derivada $Y^{(k)}$ tem distribuição estacionária $\pi(a_1^k) = \pi_k(a_1^k)$ dada por (1.18), ou seja,

$$\pi(a_1^k) = \pi(a_1^r)p(a_{r+1}|a_1^r)\ldots p(a_k|a_{k-r}^{k-1}).$$

Lema 1.12. *Se X é uma Cadeia de Markov de ordem r satisfazendo* (1.29) *então, $\forall\ k \geq r$ e $\forall\ a_1^{k+1} \in E^{k+1}$, temos*

$$\limsup_{n\to\infty} \frac{\left[N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)\right]^2}{n log\, log\, n} = 2\pi(a_1^{k+1})(1 - p(a_{k+1}|a_1^k)) \tag{1.30}$$

quase certamente. Onde $\pi(a_1^{k+1})$ é a distribuição estacionária da cadeia k-derivada $Y^{(k)}$.

Demonstração. Considere

$$g(Y_j^{(k+1)}) = 1(Y_j^{(k+1)} = a_1^{k+1}) - 1(Y_j^{(k)} = a_1^k)p(a_k+1|a_1^k) \tag{1.31}$$

e

$$S_{n-k}(g) = \sum_{j=1}^{n-k} g(Y_j^{(k+1)}).$$

Usando a definição de $N(a_1^k)$ e g, obtemos

$$\begin{aligned} S_{n-k}(g) &= N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}^k|a_1^k) + 1(Y_{n-k+1}^{(k)} = a_{n-k+1}^n)p(a_{k+1}^k|a_1^k) \\ &= N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}^k|a_1^k) + o(\delta_n). \end{aligned} \tag{1.32}$$

Indicando por E_π a esperança relativo à distribuição estacionária π, então de (1.31) e (1.18)

$$\begin{aligned} \mathrm{E}_\pi(g(Y_j^{(k+1)})) &= \pi(a_1^{k+1}) - \pi(a_1^k)p(a_{k+1}|a_1^k) \\ &= 0\,, \end{aligned}$$

e daí

$$E_\pi(S_{n-k}) = 0. \tag{1.33}$$

Da mesma forma, temos de (1.31) e (1.18)

$$\begin{aligned} E_\pi(g^2(Y_j^{(k+1)})) &= \pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k))^2 + \pi(a_1^k)(1-p(a_{k+1}|a_1^k))p(a_{k+1}|a_1^k)^2 \\ &= \pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k))\left[(1-p(a_{k+1}|a_1^k)) + p(a_{k+1}|a_1^k)\right] \\ &= \pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k)). \end{aligned} \tag{1.34}$$

Para calcular $E(g(Y_1^{(k+1)}) \cdot g(Y_j^{(k+1)}))$ com $j > 1$, consideremos $\mathscr{F}_{j+k-1} = \sigma(X_1, \ldots, X_{j+k-1})$,

então, como $Y_j^{(k+1)}$ é $\mathscr{F}_{j+k-1}$-mensurável, temos

$$\begin{aligned} E\left[g(Y_1^{(k+1)})\cdot g(Y_j^{(k+1)})\right] &= E\left[E(g(Y_1^{(k+1)})\cdot g(Y_j^{(k+1)})|\mathscr{F}_{j+k-1})\right] \\ &= E\left\{g(Y_1^{(k+1)})E\left[1(Y_j^{(k+1)}=a_1^{k+1})-1(Y_j^{(k)}=a_1^k)p(a_{k+1}|a_1^k)|\mathscr{F}_{j+k-1}\right]\right\}. \quad (1.35) \end{aligned}$$

Mas,

$$\begin{aligned} \mathrm{E}\left[1(Y_j^{(k+1)}=a_1^{k+1})|\mathscr{F}_{j+k-1}\right] &= \mathrm{E}\left[1(Y_j^{(k)}=a_1^k)1(X_{j+1}=a_{k+1})|\mathscr{F}_{j+k-1}\right] \\ &= 1(\mathrm{Y}_j^{(k)}=a_1^k)p(a_{k+1}|a_1^k). \end{aligned}$$

Logo substituindo em (1.35) segue

$$E\left[g(Y_1^{(k+1)})\cdot g(Y_j^{(k+1)})\right]=0. \quad (1.36)$$

Agora, usando (1.34) e (1.36), obtemos

$$\begin{aligned} \theta_g^2 &= E_\pi(g^2(Y_1^{(k+1)}))+2\sum_{j=2}^{n}E\left[g(Y_1^{(k+1)})\cdot g(Y_j^{(k+1)})\right] \\ &= \pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k)). \quad (1.37) \end{aligned}$$

Aplicando (1.37), (1.32) e (1.33) no Teorema 1.11, considerando $t=n-k$, e como $\theta_g^2>0$ (por (1.29)) temos

$$\begin{aligned} 1 &= \limsup_{n\to\infty}\frac{S_t(g)-E(S_t(g))}{\sqrt{2\theta_g^2 t\log\log t}} \\ &= \limsup_{n\to\infty}\frac{N(a_1^{k+1})-N(a_1^k)p(a_{k+1}|a_1^k)+1(Y_{n-k+1}^{(k)}=a_{n-k+1}^n)}{\sqrt{2\pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k))t\log\log t}} \\ &= \limsup_{n\to\infty}\frac{N(a_1^{k+1})-N(a_1^k)p(a_{k+1}|a_1^k)}{\sqrt{2\pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k))t\log\log t}}. \end{aligned}$$

Pela continuidade da função $h(x)=x^2$ e usando que $t=\left(\frac{n-k}{n}\right)n$, temos que

$$
\begin{aligned}
1 &= \left\{ \limsup_{n\to\infty} \frac{N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)}{\sqrt{2\pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k))t\log\log t}} \right\}^2 \\
&= \limsup_{n\to\infty} \left\{ \frac{\left[N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)\right]^2}{2\pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k))t\log\log t} \right\} \\
&= \limsup_{n\to\infty} \left\{ \frac{\left[N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)\right]^2}{2\pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k))n\log\log\left(\frac{n-k}{n}\right)n\ \frac{n-k}{n}} \right\}.
\end{aligned}
$$

Usando a continuidade de $\log\log x$, e das propriedades de lim sup obtemos

$$
\begin{aligned}
1 &= \limsup_{n\to\infty} \left\{ \frac{\left[N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)\right]^2}{2\pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k))n\log\log\left(\frac{n-k}{n}\right)n\ \frac{n-k}{n}} \right\} \\
&= \left[\frac{1}{2\pi(a_1^{k+1})(1-p(a_{k+1}|a_1^k))} \right] \limsup_{n\to\infty} \left\{ \frac{\left[N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)\right]^2}{n\log\log n} \right\}
\end{aligned}
$$

e portanto temos (1.30).

□

Observação 1.13. *Sob as mesmas hipóteses do Lema 1.12, podemos ainda ter*

$$
\liminf_{n\to\infty} \frac{\left[N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)\right]^2}{n\log\log n} = 0. \tag{1.38}
$$

Demonstração. Usando a definição de $N(a_1^k)$, para $n > k$, podemos verificar que

$$
\begin{aligned}
\phi_{n+1} &:= \frac{N(a_1^{k+1}|X_1^{n+1}) - N(a_1^k|X_1^{n+1})p(a_{k+1}|a_1^k)}{\sqrt{(n+1)\log\log(n+1)}} \\
&= \frac{N(a_1^{k+1}|X_1^n) + 1(X_{n-k+1}^{n+1} = a_1^{k+1}) - N(a_1^k|X_1^n)p(a_{k+1}|a_1^k) - 1(X_{n-k+1}^{n+1} = a_1^{k+1})p(a_{k+1}|a_1^k)}{\sqrt{(n+1)\log\log(n+1)}} \\
&= \frac{N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)}{\sqrt{(n+1)\log\log(n+1)}} + \frac{1(X_{n-k+1}^{n+1} = a_1^{k+1}) - 1(X_{n-k+1}^{n+1} = a_1^{k+1})p(a_{k+1}|a_1^k)}{\sqrt{(n+1)\log\log(n+1)}}.
\end{aligned}
$$

Daí dado $\varepsilon > 0$, para n suficientemente grande,

$$
\begin{aligned}
|\phi_n - \phi_{n+1}| &= \left| \frac{N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)}{\sqrt{n log\, log\, n}} - \frac{N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)}{\sqrt{(n+1) log\, log\, (n+1)}} \right. \\
&\qquad \left. - \frac{1(X_{n-k+1}^{n+1} = a_1^{k+1}) - 1(X_{n-k+1}^{n+1} = a_1^{k+1})p(a_{k+1}|a_1^k)}{\sqrt{(n+1) log\, log\, (n+1)}} \right| \\
&\leq \left| \frac{N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)}{\sqrt{n log\, log\, n}} - \frac{N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)}{\sqrt{(n+1) log\, log\, (n+1)}} \right| \\
&\qquad + \left| \frac{1(X_{n-k+1}^{n+1} = a_1^{k+1}) - 1(X_{n-k+1}^{n+1} = a_1^{k+1})p(a_{k+1}|a_1^k)}{\sqrt{(n+1) log\, log\, (n+1)}} \right| \\
&\leq \varepsilon + \left| \frac{1(X_{n-k+1}^{n+1} = a_1^{k+1}) - 1(X_{n-k+1}^{n+1} = a_1^{k+1})p(a_{k+1}|a_1^k)}{\sqrt{(n+1) log\, log\, (n+1)}} \right| \\
&\leq 2\varepsilon. \qquad (1.39)
\end{aligned}
$$

Usando raciocínio semelhante ao usado na prova do Lema 1.12, mas aplicando o Teorema 1.11 para o lim inf, podemos verificar que

$$\liminf_{n\to\infty} \phi_n = -\sqrt{2\pi(a_1^{k+1})(1 - p(a_{k+1}|a_1^k))}. \qquad (1.40)$$

Além disso, aplicando o Teorema 1.11 para o *limsup* obtemos

$$\limsup_{n\to\infty} \phi_n = \sqrt{2\pi(a_1^{k+1})(1 - p(a_{k+1}|a_1^k))}. \qquad (1.41)$$

Assim, dado $\varepsilon > 0$, pode-se tomar n_0 tal que $n > n_0$ implica que $|\phi_n - \phi_{n+1}| < \varepsilon$. Além disso, de (1.40) e (1.41), temos que existe $n_1 > n_0$ tal que $\phi_{n_1} > 0$ e $\phi_{n_1+1} \leq 0$. Usando (1.39), obtemos que $|\phi_{n_1} - \phi_{n_1+1}| < \varepsilon$. Logo

$$
\begin{aligned}
\varepsilon &> |\phi_{n_1} - \phi_{n_1+1}| = \phi_{n_1} - \phi_{n_1+1} \\
&= |\phi_{n_1} - 0| + |0 - \phi_{n_1+1}| \\
&> |\phi_{n_1} - 0| = |\phi_{n_1}|.
\end{aligned}
$$

Assim concluímos que

$$\liminf_{n\to\infty} \frac{\left[N(a_1^{k+1}) - N(a_1^k)p(a_{k+1})\right]^2}{n log\, log\, n} = \liminf_{n\to\infty} |\phi_{n_1}|^2 = 0.$$

□

Teorema 1.14. *Se X é uma Cadeia de Markov de ordem r satisfazendo* (1.29)*, então para* $k \geq r$ *temos quase certamente*

(a)

$$\limsup_{n\to\infty} \frac{log\, \hat{L}(k) - log\, L(k)}{log\, log\, n} = \gamma(k), \tag{1.42}$$

(b)

$$log\, \hat{L}(k) - log\, L(k) \geq o(\delta_n), \tag{1.43}$$

onde $\gamma(k) = |E|^k(|E|-1)$ *é o número de parâmetros livres, considerando o modelo de ordem k.*

Demonstração. (a) Da definição de $L(k)$ e $\hat{L}(k)$ temos

$$\begin{aligned} log\, \hat{L}(k) - log\, L(k) &= \sum_{a_1^{k+1}} N(a_1^{k+1}) log\, \frac{N(a_1^{k+1})}{N(a_1^k)} - \sum_{a_1^{k+1}} N(a_1^{k+1}) log\, p(a_{k+1}|a_1^k) \\ &= -\sum_{a_1^{k+1}} N(a_1^{k+1}) log\, \frac{N(a_1^k)p(a_{k+1}|a_1^k)}{N(a_1^{k+1})} \\ &= -\sum_{a_1^{k+1}} N(a_1^{k+1}) log\, \left(1 + z_n(a_1^{k+1})\right), \end{aligned} \tag{1.44}$$

onde $z_n(a_1^{k+1}) = \frac{N(a_1^k)p(a_{k+1}|a_1^k) - N(a_1^{k+1})}{N(a_1^{k+1})}$. Notemos que, como $\frac{N(a_1^{k+1})}{N(a_1^k)} \underset{q.c.}{\to} p(a_{k+1}|a_1^k)$ então $z_n(a_1^{k+1}) \underset{q.c.}{\to} 0$.

Considerando o desenvolvimento em série de Taylor em torno de 1 para $log\, x$ em (1.44), temos que

$$\begin{aligned} log\,\hat{L}(k) - log\,L(k) &= -\sum_{a_1^{k+1}} N(a_1^{k+1}) z_n(a_1^{k+1}) + \\ &\quad + \frac{1}{2}\sum_{a_1^{k+1}} N(a_1^{k+1}) \left[z_n(a_1^{k+1})\right]^2 - \\ &\quad - \sum_{a_1^{k+1}} R(a_1^{k+1}), \end{aligned} \tag{1.45}$$

com

$$\lim_{z_n(a_1^{k+1})\to 0} \frac{R(a_1^{k+1})}{\left(z_n(a_1^{k+1})\right)^2} = 0. \tag{1.46}$$

Usando o Lema 1.5 na primeira parcela de (1.45), temos

$$\begin{aligned} -\sum_{a_1^{k+1}} N(a_1^{k+1}) z_n(a_1^{k+1}) &= -\sum_{a_1^k}\sum_{a_{k+1}} \left[N(a_1^k) p(a_{k+1}|a_1^k) - N(a_1^{k+1})\right] \\ &= -\sum_{a_1^k} \left[N(a_1^k) - \left(N(a_1^k) - 1(X_{n-k+1}^n = a_1^k)\right)\right] \\ &= -\sum_{a_1^k} 1(X_{n-k+1}^n = a_1^k) \\ &= -1. \end{aligned} \tag{1.47}$$

Agora, como $\frac{N(a_1^{k+1})}{n} \underset{q.c.}{\to} \pi(a_1^{k+1})$, usando o Lema 1.12 obtemos

$$\begin{aligned} \limsup_{n\to\infty} \frac{N(a_1^{k+1})\left(z_n(a_1^{k+1})\right)^2}{log\,log\,n} &= \limsup_{n\to\infty} \frac{N(a_1^{k+1}) - N(a_1^k) p(a_{k+1}|a_1^k)}{n log\,log\,n} \frac{n}{N(a_1^k)} \\ &= 2\pi(a_1^{k+1})(1 - p(a_{k+1}|a_1^k)) \frac{1}{\pi(a_1^{k+1})} \\ &= 2(1 - p(a_{k+1}|a_1^k)) \end{aligned} \tag{1.48}$$

e

$$\begin{aligned}\limsup_{n\to\infty}\frac{\left(z_n(a_1^{k+1})\right)^2}{log\ log\ n} &= \limsup_{n\to\infty}\left[\frac{N(a_1^{k+1})\left(z_n(a_1^{k+1})\right)^2}{log\ log\ n}\frac{n}{N(a_1^{k+1})}\frac{1}{n}\right] \\ &= 0. \end{aligned} \tag{1.49}$$

Assim, de (1.48) segue

$$\begin{aligned}\limsup\frac{1}{2}\sum_{a_1^{k+1}}\frac{1}{\pi(a_1^{k+1})}\frac{\left(z_n(a_1^{k+1})\right)^2}{nlog\ log\ n} &= \frac{1}{2}\sum_{a_1^{k+1}}\frac{1}{\pi(a_1^{k+1})}2\pi(a_1^{k+1})\left(1-p(a_{k+1}|a_1^k)\right) \\ &= \sum_{a_1^{k+1}}\left(1-p(a_{k+1}|a_1^k)\right) \\ &= |E|^k(1-|E|) \\ &= \gamma(k). \end{aligned} \tag{1.50}$$

Por outro lado, de (1.46) e (1.49) segue

$$\begin{aligned}\limsup\sum_{a_1^{k+1}}\frac{R(a_1^{k+1})}{log\ log\ n} &= \limsup\sum_{a_1^{k+1}}\frac{R(a_1^{k+1})}{\left(z_n(a_1^{k+1})\right)^2}\frac{log\ log\ n}{R(a_1^{k+1})} \\ &= 0. \end{aligned} \tag{1.51}$$

Logo, de (1.45), (1.47), (1.50) e (1.51) obtemos (1.42).

(b) Temos de (1.45) e (1.47) que

$$\begin{aligned} log\,\hat{L}(k) - log\,L(k) &= -1+\frac{1}{2}\sum_{a_1^{k+1}}N(a_1^{k+1})\left(z_n(a_1^{k+1})\right)^2-\sum_{a_1^{k+1}}R(a_1^{k+1}) \\ &\geq -1-\sum_{a_1^{k+1}}R(a_1^{k+1}). \end{aligned}$$

Assim, para provar (1.43) basta mostrarmos que

$$\sum_{a_1^{k+1}} R(a_1^{k+1}) = o(\delta_n). \tag{1.52}$$

Para isto, basta observarmos que, como $\frac{N(a_1^{k+1})}{n} \underset{q.c.}{\to} \pi(a_1^{k+1})$, usando (1.38) da Observação 1.13 temos que

$$\begin{aligned} \liminf_{n\to\infty} \frac{\left(z_n(a_1^{k+1})\right)^2}{log\ log\ n} &= \liminf_{n\to\infty} \left\{ \frac{\left[N(a_1^{k+1}) - N(a_1^k)p(a_{k+1}|a_1^k)\right]^2}{nlog\ log\ n} \frac{n^2}{\left[N(a_1^{k+1})\right]^2} \frac{1}{n} \right\} \\ &= 0. \end{aligned}$$

Daí usando (1.46) segue que

$$\liminf_{n\to\infty} \sum_{a_1^{k+1}} \frac{R(a_1^{k+1})}{log\ log\ n} = 0. \tag{1.53}$$

Logo, de (1.46), (1.51) e (1.53) obtemos (1.52) e consequentemente

$$log\ \hat{L}(k) - log\ L(k) \geq o(\delta_n).$$

□

Observação 1.15. *Sob as mesmas hipóteses do Teorema 1.14 e repetindo o mesmo raciocínio da prova da parte* (a) *deste teorema, substituindo* lim sup *por* lim inf *e usando a Observação 1.13 no lugar do Lema 1.12 podemos mostrar*

$$\liminf_{n\to\infty} \frac{log\ \hat{L}(k) - log\ L(k)}{log\ log\ n} = 0.$$

Para simplificar a notação utilizada nos próximos resultados, considere:

$$\delta(k) = \sum_{a_1^{r+1}} \pi(a_1^r) p(a_{r+1}|a_1^r) log \frac{p(a_{r+1}|a_1^r)}{q(a_{r+1}|a_{r-k+1}^r)}, \tag{1.54}$$

onde

$$q(a_{r+1}|a_1^r) = p(a_{r+1}|a_1^r)$$

e

$$q(a_{r+1}|a_{r-k+1}^r) = \frac{\sum_{a_1^{r-k}} \pi(a_1^r) p(a_{r+1}|a_1^r)}{\sum_{a_1^{r-k}} \pi(a_1^r)}, \ para\ 0 \leq k < r. \tag{1.55}$$

Uma motivação para a definição de $q(a_{r+1}|a_{r-k+1}^r)$ é escrever uma probabilidade com dependência menor que r em termos das probabilidades conhecidas. Podemos ainda ver que, quase certamente,

$$\begin{aligned}
\lim \frac{N(b_1^{k+1})}{N(b_1^k)} &= \lim_{n\to\infty} \frac{\sum_{b^0_{-(r-k)+1}} N(b^{k+1}_{-(r-k)+1}) + o(\delta_n)}{\sum_{b^0_{-(r-k)+1}} N(b^k_{-(r-k)+1}) + o(\delta_n)} \\
&= \lim_{n\to\infty} \frac{\sum_{b^0_{-(r-k)+1}} \frac{N(b^k_{-(r-k)+1})}{n} \frac{N(b^{k+1}_{-(r-k)+1})}{N(b^k_{-(r-k)+1})}}{\sum_{b^0_{-(r-k)+1}} \frac{N(b^k_{-(r-k)+1})}{n}} \\
&= q(a_{r+1}|a_{r-k+1}^r).
\end{aligned} \tag{1.56}$$

Teorema 1.16. *Se X é uma Cadeia de Markov de ordem r satisfazendo* (1.29) *então, se* $0 \leq k < r$, *temos*

$$\lim_{n\to\infty} \frac{log\, \hat{L}(r) - log\, \hat{L}(k)}{n} = \delta(k),\ q.c. \tag{1.57}$$

e

$$\delta(k) > 0\ e\ \delta(k) \geq \delta(k+1). \tag{1.58}$$

Demonstração. Seja $0 \leq k < r$. Para provar (1.57), notemos que de (1.20) do Lema 1.5 podemos obter

$$\begin{aligned}
log\,\hat{L}(r) - log\,\hat{L}(k) &= \sum_{a_1^{r+1}} N(a_1^{r+1}) log\,\frac{N(a_1^{r+1})}{N(a_1^r)} - \sum_{a_1^{k+1}} N(a_1^{k+1}) log\,\frac{N(a_1^{k+1})}{N(a_1^k)} \\
&= \sum_{a_1^{r+1}} N(a_1^{r+1}) log\,\frac{N(a_1^{r+1})}{N(a_1^r)} - \sum_{a_1^{k+1}} \left[\sum_{a_{k+2}^{r+1}} N(a_1^{k+1} a_{k+2}^{r+1}) + o(\delta_n) \right] log\,\frac{N(a_1^{k+1})}{N(a_1^k)} \\
&= \sum_{a_1^{r+1}} N(a_1^{r+1}) log\,\frac{N(a_1^{r+1})}{N(a_1^r)} - \sum_{a_1^{k+1}} \left[\sum_{a_{k+2}^{r+1}} N(a_1^{k+1} a_{k+2}^{r+1}) \right] log\,\frac{N(a_1^{k+1})}{N(a_1^k)} + o(\delta_n) \\
&= \sum_{a_1^{r+1}} N(a_1^{r+1}) log\,\frac{N(a_1^{r+1})}{N(a_1^r)} \frac{N(a_1^k)}{N(a_1^{k+1})} + o(\delta_n). \qquad (1.59)
\end{aligned}$$

Agora, $lim \frac{N(a_1^k)}{N(a_1^{k+1})} \frac{N(a_1^{r+1})}{N(a_1^r)} \neq 1$ para algum a_1^{r+1}, caso contrário a ordem da Cadeia de Markov seria menor ou igual a $k < r$. Como $\frac{N(a_1^{r+1})}{n} \underset{q.c.}{\to} p(a_{r+1}|a_1^r)$, $\frac{N(a_1^{r+1})}{N(a_1^r)} \underset{q.c.}{\to} \pi(a_1^r)$ (usando o Teorema 1.9) então, juntamente com (1.56), segue que

$$\begin{aligned}
\lim_{n\to\infty} \frac{log\,\hat{L}(r) - log\,\hat{L}(k)}{n} &= \lim_{n\to\infty} \sum_{a_1^{r+1}} \frac{N(a_1^{r+1})}{n} log\,\frac{N(a_1^{r+1})}{N(a_1^r)} \frac{N(a_1^k)}{N(a_1^{k+1})} \\
&= \sum_{a_1^{r+1}} p(a_{r+1}|a_1^r)\pi(a_1^r) log\,\frac{p(a_{r+1}|a_1^r)}{q(a_{r+1}|a_{r-k+1}^r)} \\
&= \delta(k).
\end{aligned}$$

Para provar (1.58), primeiramente segue da desigualdade de Jensen que

$$\begin{aligned}
\delta(k) &= \sum_{a_1^{k+1}} p(a_{k+1}|a_1^k)\pi(a_1^k) log\, \frac{p(a_{k+1}|a_1^k)}{q(a_{k+1}|a_1^k)} \\
&= -\sum_{a_1^k} \pi(a_1^k) \sum_{a_{k+1}} p(a_{k+1}|a_1^k) log\, \frac{q(a_{k+1}|a_1^k)}{p(a_{k+1}|a_1^k)} \\
&> -\sum_{a_1^k} \pi(a_1^k) log \left[\sum_{a_{k+1}} p(a_{k+1}|a_1^k) \frac{q(a_{k+1}|a_1^k)}{p(a_{k+1}|a_1^k)} \right] \\
&= -\sum_{a_1^k} \pi(a_1^k) log \sum_{a_{k+1}} q(a_{k+1}|a_1^k) \\
&= 0\,,
\end{aligned}$$

pois $\sum_{a_{k+1}} q(a_{k+1}|a_1^k) = 1$.

A última desigualdade de (1.58) segue de (1.57) e da parte (c) do Lema 1.7, ou seja,

$$\begin{aligned}
\delta(k) &= \lim_{n\to\infty} \frac{log\hat{L}(r) - log\hat{L}(k)}{n} \\
&= \lim_{n\to\infty} \frac{log\hat{L}(r) - log\hat{L}(k) + log\hat{L}(k+1) - log\hat{L}(k+1)}{n} \\
&= \lim_{n\to\infty} \frac{log\hat{L}(r) - log\hat{L}(k+1)}{n} + \lim_{n\to\infty} \frac{log\hat{L}(k+1) - log\hat{L}(k)}{n} \\
&\geq \lim_{n\to\infty} \frac{log\hat{L}(r) - log\hat{L}(k+1)}{n} \\
&= \delta(k+1).
\end{aligned}$$

□

Teorema 1.17. *Seja X uma Cadeia de Markov com espaço de estados E, tal que $|E| \geq 2$ e satisfazendo* (1.29)*. Considere o critério EDC em* (1.9), (1.10) *com $\gamma(k) = |E|^k(|E|-1)$ e $\{c_n\}$ uma sequência de constantes reais, $c_n > 0$.*

(a) Se $k \geq r$, então quase certamente

$$\liminf_{n\to\infty} \frac{EDC(k+1) - EDC(k)}{log\, log\, n} = -2\gamma(k+1) + \left(\liminf_{n\to\infty} \frac{c_n}{log\, log\, n} \right) (|E|-1)\gamma(k) \tag{1.60}$$

e

$$\limsup_{n\to\infty} \frac{EDC(k+1) - EDC(k)}{log\, log\, n} = 2\gamma(k) + \left(\limsup_{n\to\infty} \frac{c_n}{log\, log\, n} \right) (|E|-1)\gamma(k). \tag{1.61}$$

(b) Se $0 \leq k < r$*, então quase certamente*

$$\lim_{n\to\infty} \frac{EDC(k) - EDC(r)}{n} = 2\delta(k) + [\gamma(k) - \gamma(r)] \lim_{n\to\infty} \frac{c_n}{n}. \tag{1.62}$$

Demonstração. (a) Se $k \geq r$, do Lema 1.7 temos $log\, L(k+1) = log\, L(k) + o(\delta_n)$. Então da definição (1.10) do EDC segue

$$\begin{aligned} \text{EDC(k+1) - EDC(k)} &= \text{-2 log } \hat{L}(k+1) + \gamma(k+1)c_n + 2log\, \hat{L}(k) - \gamma(k)c_n \\ &= \text{-2}\left[log\, \hat{L}(k+1) - log\, L(k+1)\right] + 2\left[log\, \hat{L}(k) - log\, L(k)\right] \\ &\quad + \text{c}_n\left[\gamma(k+1) - \gamma(k)\right] + o(\delta_n). \end{aligned}$$

Como $\gamma(k) = |E|^k(|E| - 1)$ segue

$$\begin{aligned} \text{EDC(k+1) - EDC(k)} &= \text{-2}\left[log\, \hat{L}(k+1) - log\, L(k+1)\right] + 2\left[log\, \hat{L}(k) - log\, L(k)\right] \\ &\quad + \text{c}_n\gamma(k)(|E| - 1) + o(\delta_n). \end{aligned}$$

Daí, usando (1.42) do Teorema 1.14, temos

$$\begin{aligned} \liminf_{n\to\infty} \tfrac{EDC(k+1)-EDC(k)}{log\, log\, n} &= \text{-2} \limsup_{n\to\infty} \tfrac{log\, \hat{L}(k+1) - log\, L(k+1)}{log\, log\, n} \\ &\quad +2 \liminf_{n\to\infty} \tfrac{log\, \hat{L}(k) - log\, L(k)}{log\, log\, n} \\ &\quad + \gamma(k)(|E| - 1) \liminf_{n\to\infty} \tfrac{c_n}{log\, log\, n} \\ &= \text{-2}\gamma(k+1) + \gamma(k)(|E| - 1) \liminf_{n\to\infty} \tfrac{c_n}{log\, log\, n} \end{aligned}$$

e assim (1.60) está provado.

Analogamente, usando (1.42) obtemos

$$\begin{aligned} \limsup_{n\to\infty} \tfrac{EDC(k+1)-EDC(k)}{log\, log\, n} &= \text{-2} \liminf_{n\to\infty} \tfrac{log\, \hat{L}(k+1) - log\, L(k+1)}{log\, log\, n} \\ &\quad +2 \limsup_{n\to\infty} \tfrac{log\, \hat{L}(k) - log\, L(k)}{log\, log\, n} \\ &\quad + \gamma(k)(|E| - 1) \limsup_{n\to\infty} \tfrac{c_n}{log\, log\, n} \\ &= 2\gamma(k) + \gamma(k)(|E| - 1) \liminf_{n\to\infty} \tfrac{c_n}{log\, log\, n} \end{aligned}$$

e segue (1.61).

(b) Para $0 \leq k < r$, temos usando (1.57) do Teorema 1.16

$$\begin{aligned}\lim_{n\to\infty} \frac{EDC(k)-EDC(r)}{n} &= 2 \lim_{n\to\infty} \frac{log\hat{L}(r) - log\hat{L}(k)}{n} + [\gamma(k)-\gamma(r)] \lim_{n\to\infty} \frac{c_n}{n} \\ &= 2\delta(k) + [\gamma(k)-\gamma(r)] \lim_{n\to\infty} \frac{c_n}{n},\end{aligned}$$

e portanto (1.62) está provado.

□

Corolário 1.18. *Seja X uma Cadeia de Markov, satisfazendo as mesmas hipóteses do Teorema 1.17 e $c_n > 0$ satisfazendo*

$$\liminf_{n\to\infty} \frac{c_n}{log\, log\, n} \geq \frac{2|E|}{|E|-1} \quad e \quad \limsup_{n\to\infty} \frac{c_n}{n} = 0. \tag{1.63}$$

(a) Se $k > r$ então, quase certamente

$$\liminf_{n\to\infty} \frac{EDC(k)-EDC(r)}{log\, log\, n} \geq 0 \tag{1.64}$$

e

$$\limsup_{n\to\infty} \frac{EDC(k)-EDC(r)}{log\, log\, n} > 2\gamma(r)(k-r)(|E|+1) > 2\gamma(r), \tag{1.65}$$

com o limite em (1.65) monótono crescente em k.

(b) Se $0 \leq k < r$, então quase certamente

$$\lim_{n\to\infty} \frac{EDC(k)-EDC(r)}{n} = 2\delta(k), \tag{1.66}$$

com o limite (1.66) monótono decrescente em k.

Demonstração. (a) Seja $k > r$. Aplicando as hipóteses (1.64) em (1.60) no Teorema 1.17, obtemos

$$\begin{aligned}\liminf_{n\to\infty} \frac{EDC(k+1)-EDC(k)}{log\, log\, n} &\geq -2\gamma(k+1) + \frac{2|E|}{|E|-1}(|E|-1)\gamma(k) \\ &= -2\gamma(k+1) + 2|E|\gamma(k) \\ &= -2|E|^{k+1}(|E|-1) + 2|E||E|^k(|E|-1) \\ &= 0. \end{aligned} \tag{1.67}$$

Agora, como $EDC(k)-EDC(r)=\sum_{j=r}^{k-1}[EDC(j+1)-EDC(j)]$. Aplicando (1.67) repetidas vezes, obtemos (1.64). De forma semelhante, aplicando (1.63) em (1.61) obtemos

$$\begin{aligned}\limsup_{n\to\infty}\frac{EDC(k+1)-EDC(k)}{log\,log\,n} &\geq 2\gamma(k)+\frac{2|E|}{|E|-1}(|E|-1)\gamma(k)\\ &= 2\gamma(k)(1+|E|). \qquad (1.68)\end{aligned}$$

Novamente, aplicando repetidas vezes (1.68), como para $k>r$ $\gamma(k)>\gamma(r)$ podemos obter

$$\begin{aligned}\limsup_{n\to\infty}\frac{EDC(k)-EDC(r)}{log\,log\,n} &= \sum_{i=r}^{k-1}\limsup_{n\to\infty}\frac{EDC(i+1)-EDC(i)}{log\,log\,n}\\ &\geq 2(1+|E|)\sum_{i=r}^{k-1}\gamma(i)\\ &> 2(1+|E|)\gamma(r)(k-r)\end{aligned}$$

e (1.65) está provado.

(b) Segue de (1.63) e do Teorema 1.16. □

Corolário 1.19. *Sob as mesmas hipóteses do Teorema 1.17, o estimador EDC, com termo de penalidade positivo que satisfaça (1.63), é fortemente consistente. Reciprocamente, um estimador baseado na verossimilhança penalizada que não satisfaça (1.63) não é fortemente consistente.*

Demonstração. Nessas hipóteses, usando o Corolário 1.18, temos de (a) que, quase certamente, $\hat{r}_{EDC}\leq r$ e por (b) temos que $\lim\hat{r}_{EDC}\geq r$, onde concluímos a igualdade.

Por outro lado, se $\liminf\frac{c_n}{log\,log\,n}<\frac{2|E|}{|E|-1}$, então, usando (1.60) do Teorema 1.17, temos quase certamente

$$\begin{aligned}\liminf\frac{EDC(r+1)-EDC(r)}{log\,log\,n} &= -2\gamma(r+1)+\left(\liminf\frac{c_n}{log\,log\,n}\right)(|E|-1)\gamma(r)\\ &< 0,\end{aligned}$$

o que indica que poderá ocorrer superestimação da ordem. Além disso, se $\limsup\frac{c_n}{n}=c>0$,

temos por (1.62) do teorema 1.17 que

$$\limsup \frac{EDC(r-1)-EDC(r)}{n} = 2\delta(r-1)+[\gamma(r-1)-\gamma(r)]\,c,$$

que não garante a consistência, pois poderá ter casos em que $|\,[\gamma(r-1)-\gamma(r)]\,c| > 2\delta(r-1)$. □

Segue como consequência imediata o seguinte corolário.

Corolário 1.20. *O estimador AIC não é fortemente consistente.*

Corolário 1.21. *O estimador BIC é fortemente consistente.*

Demonstração. Para o estimador BIC temos $c_n = log\, n$ e temos

$$\liminf_{n\to\infty} \frac{log\, n}{log\, log\, n} = \infty > \frac{2|E|}{|E|-1}$$

e

$$\limsup_{n\to\infty} \frac{log\, n}{n} = 0,$$

logo as hipóteses do Corolário 1.19 estão satisfeitas. □

Corolário 1.22. *Sob as hipóteses do Corolário 1.18, o termo de penalidade ótimo é*

$$c_n \cdot \gamma(k) = \frac{2|E|}{|E|-1} log\, log\,(n) \cdot (|E|-1)|E|^k. \tag{1.69}$$

Demonstração. O termo de penalidade deve ser o menor possível para evitar subestimação da ordem e grande o suficiente para ter a consistência forte. Neste caso, pelas condições do Corolário 1.18, o menor termo assintótico é $\frac{2|E|}{|E|-1} log\, log\,(n) \cdot (|E|-1)|E|^k$. □

Em consequência desse resultado, segue:

Corolário 1.23. *O estimador BIC penaliza mais que necessário.*

Como pode ser verificado na demonstração do Corolário 1.22, o fato do BIC penalizar mais que o necessário gera uma maior tendência desse estimador a subestimar a ordem.

Vale ressaltar também que o Corolário 1.21 é o teorema da consistência forte do BIC, apresentada por Csiszar & Shields (2000).

1.3 Considerações

Se $0 \leq k < r$, pela equação (1.59) na prova do Teorema 1.16 e pela definição de $N(a_1^{r+1})$ em (1.4), temos que

$$log\,\hat{L}(r) - log\,\hat{L}(k) = \sum_{a_1^{r+1}} \left(\sum_{j=1}^{j-k} 1(X_j^{j+k} = a_1^{k+1}) \right) log\,\frac{N(a_1^{r+1})}{N(a_1^r)} \frac{N(a_1^k)}{N(a_1^{k+1})} + o(\delta_n).$$

Considerando a cadeia $(r+1)$-derivada, $Y^{(r+1)}$ e tomando $g : E \to \mathbb{R}$,

$$g(Y_j^{(r+1)}) = \sum_{a_1^{r+1}} 1(Y_j = a_1^{k+1}) log\,\frac{N(a_1^{r+1})}{N(a_1^r)} \frac{N(a_1^k)}{N(a_1^{k+1})} \;\; e\; S_n(g) = \sum_{j=1}^{n} g(Y_j),$$

temos

$$log\,\hat{L}(r) - log\,\hat{L}(k) = \sum_{j=1}^{j=n} g(Y_j^{(r+1)}) + o(\delta_n) = S_n(g) + o(\delta_n).$$

Assim, segue do Teorema 1.10

$$\frac{log\,\hat{L}(r) - log\,\hat{L}(k) - n\delta(k)}{\sqrt{n\theta_g^2}} \underset{d}{\to} \mathcal{N}(0,1),$$

onde

$$\begin{aligned} \theta_g^2 \;\; &= \sum_{a_1^{r+1}} \pi(a_1^{r+1}) \left(log\,\tfrac{p(a_{r+1}|a_1^r)}{q(a_{r+1}|a_{r-k+1}^r)} \right)^2 \\ &\quad + 2 \sum_{j=2}^{n} \sum_{a_1^{r+1}, b_1^{r+1}} P(Y_1^{(k+1)} = a_1^{r+1}) P(Y_j^{(k+1)} = b_1^{r+1}) log\,\tfrac{p(a_{r+1}|a_1^r)}{q(a_{r+1}|a_{r-k+1}^r)}\; log\,\tfrac{p(b_{r+1}|b_1^r)}{q(b_{r+1}|b_{r-k+1}^r)}. \end{aligned}$$

Analisando o comportamento para n suficientemente grande e fixo, podemos então concluir da argumentação acima e do Teorema 1.16: Se X é uma Cadeia de Markov de ordem r satisfazendo (1.29) então, para $k < r$

(i) $-2log\,\hat{L}(k)+2log\,\hat{L}(r)=2n\delta(k)+o(n)$ e

(ii) $-2log\,\hat{L}(k)+2log\,\hat{L}(r)\sim\mathcal{N}(2n\delta(k),n\theta_g^2)$.

Estas conclusões despertam o interesse em se conhecer melhor $\delta(k)$. O Teorema 1.16 mostra que $\delta(k)$ deve ser positivo para $k<r$ e, pela sua definição, é possível notar que pode ser arbitrariamente próximo de 0. Entretanto, podemos ter um limitante superior para $\delta(k)$, limitando-o no intervalo, relativamente pequeno, $(0, log\,|E|)$. Isto é: conforme definido,

$$\delta(k)<log\,|E|\,.$$

De fato, usando as definições (1.54) e (1.55) de $\delta(k)$ e $q(a_{r+1}|a_{r-k+1}^r)$, respectivamente, e o Lema 1.6 podemos obter

$$
\begin{aligned}
\delta(k) &= log\left[\prod_{a_1^{r+1}}\left(\frac{p(a_{r+1}|a_1^r)}{q(a_{r+1}|a_{r-k+1}^r)}\right)^{p(a_{r+1}|a_1^r)\pi(a_1^r)}\right] \\
&\leq log\left[\sum_{a_1^{r+1}}\left(\frac{p(a_{r+1}|a_1^r)p(a_{r+1}|a_1^r)\pi(a_1^r)}{\frac{\sum_{b_1^{r-k}}\pi(b_1^{r-k}a_{r-k+1}^r)p(a_{r+1}|b_1^{r-k}a_{r-k+1}^r)}{\sum_{b_1^{r-k}}\pi(b_1^{r-k}a_{r-k+1}^r)}}\right)\right] \\
&= log\left[\sum_{a_{r-k+1}^{r+1}}\left(\frac{\sum_{a_1^{r-k}}p(a_{r+1}|a_1^r)p(a_{r+1}|a_1^r)\pi(a_1^r)}{\frac{\sum_{b_1^{r-k}}\pi(b_1^{r-k}a_{r-k+1}^r)p(a_{r+1}|b_1^{r-k}a_{r-k+1}^r)}{\sum_{b_1^{r-k}}\pi(b_1^{r-k}a_{r-k+1}^r)}}\right)\right] \\
&< log\left[\max_{a_1^{r+1}}(p(a_{r+1}|a_1^r))\sum_{a_{r-k+1}^{r+1}}\left(\frac{\sum_{a_1^{r-k}}p(a_{r+1}|a_1^r)\pi(a_1^r)}{\frac{\sum_{b_1^{r-k}}\pi(b_1^{r-k}a_{r-k+1}^r)p(a_{r+1}|b_1^{r-k}a_{r-k+1}^r)}{\sum_{b_1^{r-k}}\pi(b_1^{r-k}a_{r-k+1}^r)}}\right)\right] \\
&\leq log\left[\sum_{a_{r-k+1}^{r+1}}\sum_{b_1^{r-k}}\pi(b_1^{r-k}a_{r-k+1}^r)\right] \\
&= log\,|E|\,. \qquad (1.70)
\end{aligned}
$$

Assim, podemos justificar a escolha do termo de penalidade ótimo em (1.69) no Corolário 1.22 alternativamente da seguinte forma:

Considere X_r, uma Cadeia de Markov de ordem r satisfazendo (1.2) e $|E| = N \geq 2$. Para $k < r$ e usando a argumentação acima temos que para o estimador indique corretamente a ordem é necessário a desigualdade

$$\begin{aligned}(\gamma(r)-\gamma(k))c_n &\leq -2log\,\hat{L}(k)+2log\,\hat{L}(r)\\ &\sim \mathcal{N}(2n\delta(k),n\theta_g^2)\end{aligned} \tag{1.71}$$

Nesse sentido, como $\delta(k)\in(0,log\,|E|\,)$ é arbitrário, deve-se tomar o menor c_n assintoticamente que garanta a consistência forte para que a desigualdade ocorra para n pequeno. Neste caso, $c_n=\frac{|E|}{|E|-1}log\,log\,n$.

Por outro lado, um termo pequeno pode causar uma tendência a superestimação da ordem para uma cadeia X_d, de ordem $d<r$. Entretanto, quando $(\gamma(r)-\gamma(k))c_n>2n\delta(k)$, a partir de (1.71), o erro de subestimação de X_r é aproximadamente

$$P\left(\mathcal{N}(2n\delta(k),n\theta_g^2)<(\gamma(r)-\gamma(k))c_n\right)>P\left(\mathcal{N}(2n\delta(k),n\theta_g^2)<2n\delta(k)\right)=0,5$$

enquanto, usando (1.6), X_d tem menos de 50% de chance de superestimação (erro).

Portanto, tomar o termo menor assintoticamente (que garanta a consistência forte) é uma boa escolha para antecipar o fim da tendência a subestimar e, por outro lado, não induz uma tendência a superestimar.

2 Análise Comparativa dos Estimadores

Neste capítulo são apresentados os resultados obtidos em simulações realizadas com o objetivo de comparar os estimadores fortemente consistentes BIC e EDC_{opt} (EDC com termo de penalidade ótimo) definidos por (1.9) e (1.69) no capítulo anterior, além de analisar o comportamento do estimador, inconsistente, AIC dado por (1.8).

Vale ressaltar que, esse tipo de comparação seria praticamente impossível de ser feita toda teoricamente. Isso porque levaria a contas exageradamente grandes que dependeriam das probabilidades de transição desconhecidas[1]. Além disso, não faria sentido simplificar as expressões tomando comportamentos assintóticos.

As simulações computacionais foram realizadas considerando os casos em que a ordem varia de 0 a 6 ($r = 0..6$) e o tamanho do espaço de estados varia de 2 a 10 ($N = 2..10$), perfazendo 63 casos. Para cada caso foram consideradas mil cadeias de Markov, geradas aleatoriamente. E para cada Cadeia de Markov foi gerada 1 amostra de tamanho 100 milhões. Foram consideradas "sub-amostras" desta, tomando-se os fragmentos da posição inicial até tamanhos pré-definidos. Isso não só dá uma sensação de aproximação, do ponto de vista teórico, mas traz um grande benefício computacional, pois desta forma as contagens de um fragmento de amostra são feitas a partir das contagens do último fragmento computado.

Os casos foram escolhidos em função das capacidades computacionais. Os tamanhos das amostras foram determinados empiricamente, na busca de valores mais adequados para a comparação dos estimadores.

[1] Esse fato foi observado por Katz (1981), que diz: *"Analytical expressions for exact distributions of $\hat{k}_{aic}$ and $\hat{k}_{bic}$ (as a function of the sample size n) are not available and, in any event, would probably be too complicated to be very useful."*

Esses números, embora não aparentem muita expressividade, são consideráveis. No caso de maior complexidade ($r = 6$ e $N = 10$), temos uma Cadeia de Markov com $(10-1)10^6 = 9.000.000$ parâmetros e, para este caso, nos testes realizados, os estimadores necessitam de amostras superiores a 100 milhões para acusarem a ordem corretamente!

As simulações geraram ao todo $22.050.000$[2] resultados para análise. Dessa forma, foram considerados relatórios sumarizados, com o foco na comparação direta entre os métodos, e distribuições dos valores calculados para cada método.

A seguir, na seção 2.1, descrevemos os objetivos e a metodologia dos experimentos realizados e na seção 2.2 apresentamos uma análise dos resultados obtidos nas simulações. Para finalizar, apresentamos na seção 2.3 um exemplo simples de aplicação desses estimadores na análise de peças musicais sugerido por McAlpine, Miranda & Hoggar (1999).

2.1 Definição dos Experimentos Computacionais

Objetivos

- Conhecer o comportamento dos estimadores em amostras “pequenas” e “grandes”;
- Identificar, para cada caso, os tamanhos em que os estimadores acertam 50%;
- Comparar a eficiência dos estimadores em relação a ordem (r), tamanho do espaço de estados (N) e tamanho da amostra (n);

Metodologia

Para cada $(r,N) \in \{0,\ldots,6\} \times \{2,\ldots,10\}$, gerar 1000 cadeias de Markov, de forma aleatória. Para cada cadeia, gerar uma amostra de tamanho 100 milhões. Para cada “sub-amostra” desta, calcular e salvar os valores da log-verossimilhança e ordens indicadas pelos estimadores $\hat{r}_{EDC}$, $\hat{r}_{BIC}$ e $\hat{r}_{AIC}$.

Usando o banco de dados criado, gerar relatórios e gráficos apropriados, a fim de auferir conclusões.

[2] Resultado de $349 * 1000 * 63$

Utilizar o procedimento proposto por Raftery (1985) para a geração de modelos de Cadeias de Markov por permitir uma maior representatividade. Na geração das amostras utilizar a biblioteca/algoritmo de aleatoriedade proposto por Park & Miller (1988).

Indicadores Comparativos

- Porcentagens de acertos para cada caso (r,N,n) considerado;
- Porcentagens de acertos sumarizados, para casos onde são fixados r ou N ou n;
- Porcentagens de acertos para todos os casos, considerados conjuntamente;
- Para os níveis de sumarização descritos nos itens anteriores, considerar as porcentagens de acerto (erro) de um certo estimador, quando os outros acertam (erram) – esse indicador dá uma noção de quais são plenamente "substituíveis" por outros;
- Gráficos para cada caso (r,N), considerando a porcentagem de acerto em função do tamanho da amostra n;
- Gráficos das distribuições dos estimadores em casos específicos (r,N,n).

2.2 Análise dos Resultados Obtidos nas Simulações

A seguir apresentamos algumas conclusões obtidas após a análise dos resultados dos experimentos realizados.

2.2.1 O estimador EDC_{opt} é mais eficiente que o BIC

Em todos os casos simulados, o EDC_{opt} apresentou maior proporção de acertos que o BIC para qualquer tamanho de amostra. A exceção foi para os casos onde $|E| = 2$ e em certo intervalo do tamanho amostral. Nestes casos, os termos de penalidade do BIC podem ser menores que os do EDC_{opt}, o que justifica o resultado obtido.

Análise dos Indicadores

A Tabela 2.1 apresenta resultados obtidos no caso $|E| = 4$ e $r = 1$, onde a coluna n representa o tamanho da amostra, "$<$", "$=$" e "$>$", respectivamente, representam as proporções de subestimação, acerto e superestimação para cada n.

Tabela 2.1: Distribuições de Acertos dos Estimadores EDC_{opt} e BIC para o caso $|E| = 4$ e $r = 1$

n	EDC_{opt}			BIC		
	$<$	$=$	$>$	$<$	$=$	$>$
10	98,70%	1,30%	0%	99,10%	0,90%	0%
25	90,20%	9,80%	0%	91,40%	8,60%	0%
68	50,60%	49,40%	0%	60,30%	39,70%	0%
775	0%	100,00%	0%	0,10%	99,90%	0%
900	0%	100,00%	0%	0%	100,00%	0%

Da mesma forma, a Tabela 2.2 apresenta os resultados para o caso $|E| = 10$ e $r = 1$. Como pode ser verificado, em ambos casos, o EDC_{opt} apresenta melhor performance que o BIC. No de menor complexidade ($|E| = 4$) as proporções de acertos são semelhantes, para o caso de maior complexidade ($|E| = 10$) o EDC_{opt} necessitou de pouco mais da metade do tamanho da amostra para acertar mais de 50% dos casos [considerando a mediana da distribuição de acertos como indicador de performance]. Esse distanciamento se verifica a medida que a complexidade (número de parâmetros livres) aumenta. Na Tabela 2.3 está representado, para cada caso, o tamanho de amostra mínimo em que cada estimador acertou pelo menos 50%, a última coluna tem a proporção $prop := \frac{n \text{ em que BIC acerta } 50\%}{n \text{ em que } EDC_{opt} \text{ acerta } 50\%}$, que indica o "quanto o EDC_{opt} é melhor que o BIC".

Tabela 2.2: Distribuições de Acertos dos Estimadores EDC_{opt} e BIC para o caso $|E| = 10$ e $r = 1$

n	EDC_{opt}			BIC		
	$<$	$=$	$>$	$<$	$=$	$>$
218	99,80%	0,20%	0%	100,00%	0,00%	0%
425	40,90%	59,10%	0%	100,00%	0,00%	0%
450	28,90%	71,10%	0%	99,90%	0,10%	0%
600	3,10%	96,90%	0%	91,10%	8,90%	0%
775	0,10%	99,90%	0%	48,20%	51,80%	0%
950	0%	100,00%	0%	15,40%	84,60%	0%
1812	0%	100,00%	0%	0%	100,00%	0%

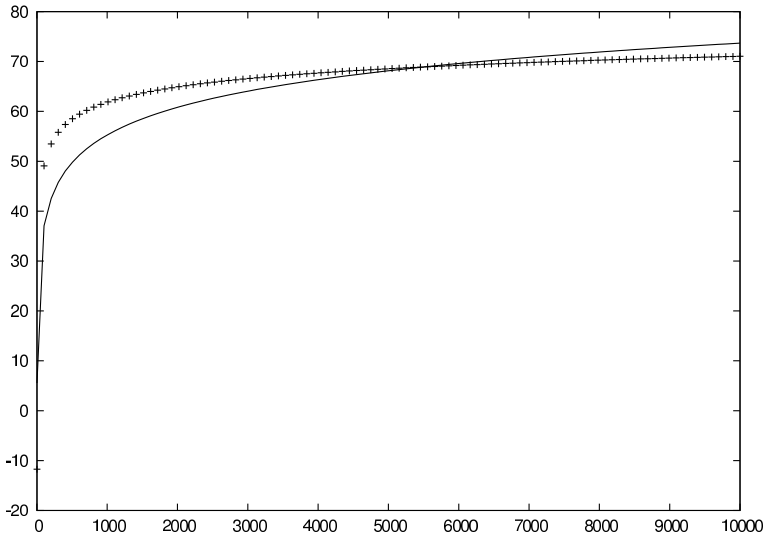
Tabela 2.3: Tamanhos de Amostras Mínimos em que os Estimadores EDC_{opt} e BIC acertam mais que 50%

r	$\|E\|$	n EDC_{opt}	n BIC	*prop*
1	2	76	50	0,65
	3	53	50	0,94
	4	72	85	1,18
	5	100	150	1,50
	6	143	225	1,57
	7	200	337	1,68
	8	250	475	1,90
	9	337	625	1,85
	10	425	775	1,82
2	2	1125	925	0,82
	3	1125	1500	1,33
	4	2000	3250	1,62
	5	3125	5750	1,84
	6	5750	11250	1,95
	7	8000	16875	2,10
	8	10625	23750	2,23
	9	16875	40000	2,37
	10	23750	62500	2,63
3	2	10625	11250	1,05
	3	10000	15625	1,56
	4	23750	45000	1,89
	5	47500	100000	2,10
	6	93750	212500	2,26
	7	162500	400000	2,46
	8	293750	775000	2,63
	9	525000	1437500	2,73
	10	637500	1812500	2,84
4	2	32500	37500	1,15
	3	81250	137500	1,69
	4	225000	475000	2,11
	5	600000	1437500	2,39
	6	1562500	4250000	2,72
	7	2875000	8125000	2,82
	8	4750000	13750000	2,89
5	2	143750	175000	1,21
	3	337500	650000	1,92
	4	1687500	4000000	2,37
	5	5625000	15000000	2,66
6	2	400000	525000	1,31
	3	2000000	4000000	2,00
	4	11875000	28750000	2,42

Nota-se que os casos onde $|E| = 2$ e $r \in \{1,2\}$ ou $|E| = 3$ e $r = 1$ são atípicos – o estimador BIC apresenta melhor performance que o EDC_{opt} – isso é justificável pois, para n pequeno (se $|E| = 2$, n no intervalo $[5,4500]$ é suficiente, como exemplo, veja o Gráfico 2.1, com os respectivos termos de penalidade, considerando $k = 3$ e $|E| = 2$), o termo de penalidade do BIC é menor que o do EDC_{opt} e, para as ordens considerados ($r = 0,\ldots,6$), o tamanho nessas proporções garante uma probabilidade pequena de superestimação para ambos os casos. Assim, o estimador com o menor termo de penalidade tem uma maior chance de acerto. Esses três casos específicos não contradizem o Corolário 1.22, pois para n

maior, o termo de penalidade do BIC é maior que o do EDC_{opt}, e o corolário toma o menor termo assintótico.

Figura 2.1: Termos de penalidade do BIC (contínuo) e EDC_{opt} (pontilhado) para $|E| = 2$ e $k = 3$ em função de n



À medida que aumenta a complexidade (número de parâmetros livres) dos modelos, aumenta também a diferença entre as proporções de acerto do EDC_{opt} e BIC. Isso ocorre pois, modelos mais complexos exigem tamanhos de amostras maiores, e nestes casos os termos de penalidade de ambos se distanciam substancialmente, refletindo essa diferença nas proporções de acerto. Isso nos leva a concluir que, para casos ainda mais complexos que os simulados, o EDC_{opt} deverá apresentar uma proporção de acerto ainda maior que o *BIC*.

Para os outros casos considerados de $|E|$ e r, podemos observar que o comportamento dos estimadores é semelhante ao descrito aqui no caso $r = 1$, conforme podemos ver nas Tabelas 2.4 e 2.5 para os casos $(r, |E|) = (3, 4)$ e $(r, |E|) = (4, 5)$.

Tabela 2.4: Distribuições de Acertos dos Estimadores EDC_{opt} e BIC para o caso $|E| = 4$ e $r = 3$

n	EDC_{opt}			BIC		
	$<$	$=$	$>$	$<$	$=$	$>$
1562	99,90%	0,10%	0,00%	100,00%	0,00%	0,00%
2375	98,80%	1,20%	0,00%	99,90%	0,10%	0,00%
23125	50,20%	49,80%	0,00%	65,10%	34,90%	0,00%
9375000	0,00%	100,00%	0,00%	0,60%	99,40%	0,00%
23750000	0,00%	100,00%	0,00%	0,00%	100,00%	0,00%

Tabela 2.5: Distribuições de Acertos dos Estimadores EDC_{opt} e BIC para o caso $|E| = 4$ e $r = 5$

n	EDC_{opt}			BIC		
	$<$	$=$	$>$	$<$	$=$	$>$
6500	100,00%	0,00%	0,00%	100,00%	0,00%	0,00%
32500	99,80%	0,20%	0,00%	100,00%	0,00%	0,00%
68750	93,60%	6,40%	0,00%	99,70%	0,30%	0,00%
600000	49,80%	50,20%	0,00%	66,40%	33,60%	0,00%
1437500	33,50%	66,50%	0,00%	49,90%	50,10%	0,00%
16875000	7,00%	93,00%	0,00%	13,81%	86,19%	0,00%
100000000	0,00%	100,00%	0,00%	3,40%	96,60%	0,00%

2.2.2 Para *n* suficientemente pequeno, todos os estimadores têm tendência a subestimar

Katz (1981) por três vezes sugere que os estimadores BIC e AIC subestimam em amostras pequenas – "$\ldots \hat{k}_{BIC}$ *seldom overestimates the true order.*", "*...A simple modification of the BIC procedure could reduce this tendency to underfit.*", "*...except for* $n = 50$*, the AIC procedure virtually never underfits...*" – Esse comportamento foi verificado nas simulações para os estimadores considerados.

A explicação para esse fato segue da seguinte argumentação: $2[log\,\hat{L}(r) - log\,\hat{L}(l)] = 2n\delta(l) + o(n)$ se $l < r$ (Teorema 1.16), e portanto $EDC(l) - EDC(r) = 2n\delta(l) + o(n) - c_n(\gamma(r) - \gamma(l))$. Como $\delta(l)$ é relativamente pequeno (conforme (1.70)), $c_n(\gamma(r) - \gamma(l)) > 2n\delta(l) + o(n)$ para *n* suficientemente pequeno, o que leva à subestimação. Obviamente, isso não pode ocorrer para *n* qualquer. O que justifica a não ocorrência de subestimações da ordem para *n* grande.

A Tabela 2.6 apresenta a distribuição de acertos para alguns casos em tamanhos variados.

Observa-se que o comportamento do AIC inverte para certo *n* que depende da complexidade e, como será visto na próxima seção, o AIC pode ter probabilidade consideravelmente pequena de superestimação para *n* grande.

Tabela 2.6: Distribuições de Acertos dos Estimadores EDC_{opt}, BIC e AIC

r	$\|E\|$	n	EDC_{opt}			BIC			AIC		
			$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
1	3	10	80,30%	19,70%	0,00%	73,90%	26,10%	0,00%	63,10%	36,90%	0,00%
		22	72,20%	27,80%	0,00%	67,40%	32,60%	0,00%	39,80%	59,90%	0,30%
		168	15,00%	85,00%	0,00%	15,80%	84,20%	0,00%	3,30%	93,50%	3,20%
		1375	0,60%	99,40%	0,00%	0,80%	99,20%	0,00%	0,10%	96,20%	3,70%
		3125	0,00%	100,00%	0,00%	0,10%	99,90%	0,00%	0,00%	96,20%	3,80%
		5000	0,00%	100,00%	0,00%	0,00%	100,00%	0,00%	0,00%	97,10%	2,90%
1	4	10	98,70%	1,30%	0,00%	99,10%	0,90 %	0,00%	96,10%	3,90%	0,00%
		131	18,40%	81,60%	0,00%	27,50%	72,50 %	0,00%	1,60%	98,40%	0,00%
		212	7,30%	92,70%	0,00%	12,20%	87,80 %	0,00%	0,20%	99,70%	0,10%
		975	0,00%	100,00%	0,00%	0,00%	100,00%	0,00%	0,00%	99,90%	0,10%
2	3	17	100,00%	0,00 %	0,00%	100,00%	0,00 %	0,00%	99,90%	0,10%	0,00%
		137	96,50 %	3,50 %	0,00%	97,30 %	2,70 %	0,00%	58,59%	41,30%	0,10%
		175000	1,70 %	98,30 %	0,00%	3,50 %	96,50%	0,00%	0,10%	99,80%	0,10%
		4750000	0,00 %	100,00%	0,00%	0,00 %	100,00%	0,00%	0,00%	99,90%	0,10%
2	5	137	100,00%	0,00%	0,00%	100,00%	0,00%	0,00%	99,70%	0,30%	0,00%
		400	99,90%	0,10%	0,00%	100,00%	0,00%	0,00%	67,00%	33,00%	0,00%
		650	99,00%	1,00%	0,00%	100,00%	0,00%	0,00%	50,10%	49,90%	0,00%
		750	97,00%	3,00%	0,00%	9990%	0,10%	0,00%	47,00%	53,00%	0,00%
		3125	49,90%	50,10%	0,00%	68,10%	31,90%	0,00%	20,90%	79,10%	0,00%
		6000	35,80%	64,20%	0,00%	49,00%	51,00%	0,00%	13,70%	86,30%	0,00%
		106250	5,40%	94,60%	0,00%	10,60%	89,40%	0,00%	0,50%	99,50%	0,00%
		187500	4,10%	95,90%	0,00%	6,40%	93,60%	0,00%	0,00%	100,00%	0,00%
		837500	0,00%	100,00%	0,00%	1,50%	98,50%	0,00%	0,00%	100,00%	0,00%
		2000000	0,00%	100,00%	0,00%	0,00%	100,00%	0,00%	0,00%	100,00%	0,00%
3	10	17500	100,00%	0,00%	0,00%	100,00%	0,00%	0,00%	99,39%	0,61%	0,00%
		81250	99,39%	0,61%	0,00%	100,00%	0,00%	0,00%	61,26%	38,74%	0,00%
		131250	91,49%	8,51%	0,00%	100,00%	0,00%	0,00%	49,24%	50,76%	0,00%
		637500	49,94%	50,06%	0,00%	77,07%	22,93%	0,00%	20,62%	79,38%	0,00%
		1812500	30,63%	69,37%	0,00%	49,74%	50,26%	0,00%	12,31%	87,69%	0,00%
		11250000	10,31%	89,69%	0,00%	19,51%	80,49%	0,00%	0,60%	99,40%	0,00%
		13750000	7,90%	92,10%	0,00%	18,21%	81,79%	0,00%	0,00%	100,00%	0,00%
		100000000	0,00%	100,00%	0,00%	6,41%	93,59%	0,00%	0,00%	99,20%	0,80%
4	4	2000	100,00%	0,00%	0,00%	100,00%	0,00%	0,00%	99,80%	0,20%	0,00%
		9000	99,90%	0,10%	0,00%	100,00%	0,00%	0,00%	81,90%	18,10%	0,00%
		15625	98,40%	1,60%	0,00%	99,90%	0,10%	0,00%	68,70%	31,30%	0,00%
		37500	88,60%	11,40%	0,00%	96,90%	3,10%	0,00%	49,70%	50,30%	0,00%
		225000	49,30%	50,70%	0,00%	64,90%	35,10%	0,00%	20,90%	79,10%	0,00%
		475000	36,60%	63,40%	0,00%	49,40%	50,60%	0,00%	13,30%	86,70%	0,00%
		16250000	2,80%	97,20%	0,00%	7,60%	92,40%	0,00%	0,00%	100,00%	0,00%
		100000000	0,10%	99,90%	0,00%	0,40%	99,60%	0,00%	0,00%	100,00%	0,00%

2.2.3 Comportamento do estimador AIC

Mesmo com a inconsistência do estimador AIC e a existência de alternativas fortemente consistentes, este estimador vêm sendo utilizado em muitas aplicações nas mais diferentes áreas (Yamaoka, Nakagawa & Uno (2005), Hoon, Imoto & Miyano (2002), Rose, Dick, Viken & Kaprio (2001), dentre os mais recentes).

Assim, a questão, já levantada por Kuha (2004), por exemplo, sobre a eficácia da utilização do estimador AIC merece ainda ser estudada. Neste sentido, realizamos as simulações numéricas com o objetivo de analisar especialmente a performance do AIC em amostras de

tamanho finito (não necessariamente grande) e a probabilidade de superestimação do AIC, apontada por Katz (1981).

Em linhas gerais, nos resultados obtidos para todos os casos considerados, verificou-se um comportamento padrão entre as proporções de acertos dos estimadores, sendo que o AIC subestimava para amostras bem pequenas, apresentava maior quantidade de acertos para amostras pequenas, mantendo proporção de acerto próximo de sua distribuição limite. Os tamanhos das amostras em que esses comportamentos apareciam dependiam substancialmente da complexidade dos casos considerados.

Para n grande e casos de menor complexidade, o AIC manteve uma taxa de acerto maior que 67%. Para casos de maior complexidade, essa taxa se manteve em valores **próximos de** 100% – que não contradiz a sua inconsistência, pois a probabilidade de superestimação, embora positiva, pode ser pequena.

Análise dos Indicadores

A Tabela 2.7 apresenta o caso $r = 1$ e $|E| = 2$. Verifica-se que o AIC, EDC_{opt} e BIC acertam, respectivamente, $42,92\%$, $28,65\%$ e $39,53\%$, para $n = 10$. Entretanto, com $n = 5000000$, o AIC acerta apenas $67,17\%$, enquanto ambos outros acertam $100,00\%$ dos casos. Mas, para um caso de maior complexidade (Tabela 2.8), $r = 1$ e $|E| = 6$, o AIC, EDC_{opt} e BIC acertam, respectivamente, $43,92\%$, $0,50\%$ e $0,00\%$, para $n = 45$ e para $n = 5000000$, o AIC acerta $100,00\%$.

Tabela 2.7: Distribuições de Acertos dos Estimadores Para o Caso $r = 1$ e $|E| = 2$

n	EDC_{opt}			BIC			AIC		
	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
10	70,45	28,65	0,89	53,68	39,53	6,78	45,20	42,92	11,87
57	52,99	46,91	0,10	43,40	54,80	1,79	24,95	50,90	24,15
375000	0,59	99,41	0,00	0,59	99,41	0,00	0,30	70,06	29,64
475000	0,59	99,41	0,00	0,59	99,41	0,00	0,00	68,77	31,23
5000000	0,00	100,00	0,00	0,00	100,00	0,00	0,00	67,17	32,83

Cabe observar que o AIC acerta primeiro pois, $2(log\,\hat{L}(r) - log\,\hat{L}(l)) = 2n\delta(l) + o(n)$ se $l < r$ (Teorema 1.16), e portanto $EDC(l) - EDC(r) = 2n\delta(l) + o(n) - c_n(\gamma(r) - \gamma(l))$.

Tabela 2.8: Distribuições de Acertos dos Estimadores Para o Caso $r = 1$ e $|E| = 6$

n	EDC_{opt}			BIC			AIC		
	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
22	100,00	0,00	0,00	100,00	0,00	0,00	99,80	0,20	0,00
45	99,50	0,50	0,00	100,00	0,00	0,00	56,08	43,92	0,00
200	11,87	88,13	0,00	46,70	53,30	0,00	0,00	100,00	0,00
5000000	0,00	100,00	0,00	0,00	100,00	0,00	0,00	100,00	0,00

Assim, é necessário n suficientemente grande para $2n\delta(l) + o(n) > c_n(\gamma(r) - \gamma(l))$ e o estimador não subestimar a ordem. Neste caso, quanto menor o fator c_n no termo de penalidade, menor o n necessário para que isso ocorra. Como o AIC tem um termo de penalidade menor que o EDC_{opt} e BIC, ele acerta primeiro.

Por outro lado, o AIC erra mesmo para n substancialmente grande. Isso vem do fato dele ser inconsistente (Katz 1981). Basicamente, $2[log\,\hat{L}(l) - log\,\hat{L}(r)] \sim \chi^2(\gamma(l) - \gamma(r))$ se $l > r$ (Billingsley 1961). Enquanto isso, o termo de penalidade do AIC é constante, resultando em $AIC(l) - AIC(r) \sim -\chi^2(\gamma(l) - \gamma(r)) + 2(\gamma(l) - \gamma(r))$. Logo, como $P(\chi^2(\gamma(l) - \gamma(r)) > 2(\gamma(l) - \gamma(r))) > 0$, temos uma probabilidade positiva de $AIC(l) < AIC(r)$, levando à superestimação da ordem.

Entretanto, $P(\chi^2(t) > 2t)$ pode ser muito pequena se $t = \gamma(l) - \gamma(r)$ for grande, o que ocorre em modelos mais complexos. Para exemplificar, calculamos alguns valores de $P(\chi^2(\gamma(l) - \gamma(r)) > 2(\gamma(l) - \gamma(r)))$ na Tabela 2.9 [3].

Nas simulações, consideramos o limitante superior para a ordem igual a 7 (i.e. $K = 7$). Utilizando as mesmas contas realizadas por Katz (1981), temos que, assintoticamente,

$$\begin{aligned} P(\hat{r}_{aic} > r) = \sum_{i=r}^{K} P(\hat{r}_{aic} = i) \quad &\leq \sum_{i=r}^{K} P\left(2[log\,\hat{L}(i+1) - log\,\hat{L}(i)] > 2(\gamma(i+1) - \gamma(i))\right) \\ &\cong \sum_{i=r}^{K} P\left(\chi^2[\gamma(i+1) - \gamma(i)] > 2(\gamma(i+1) - \gamma(i))\right) \end{aligned}$$

Assim, para o caso considerado por Katz, $|E| = 2$ e $r = 1$, a probabilidade assintótica do AIC superestimar é

[3] Foi utilizado o programa R para o cálculo numérico.

Tabela 2.9: Probabilidades Calculadas para a Distribuição χ^2

$\|E\|$	l	$\gamma(l)-\gamma(l-1)$	$P(\chi^2(\gamma(l)-\gamma(l-1)) > 2(\gamma(l)-\gamma(l-1)))$	Probabilidade
$\|E\|=2$	2	2	$P(\chi^2(\gamma(2)-\gamma(1)) > 2(\gamma(2)-\gamma(1)))$	$0,135335$
	3	4	$P(\chi^2(\gamma(3)-\gamma(2)) > 2(\gamma(3)-\gamma(2)))$	$0,0915782$
	4	8	$P(\chi^2(\gamma(4)-\gamma(3)) > 2(\gamma(4)-\gamma(3)))$	$0,0423801$
	5	16	$P(\chi^2(\gamma(5)-\gamma(4)) > 2(\gamma(5)-\gamma(4)))$	$0,00999978$
	6	32	$P(\chi^2(\gamma(6)-\gamma(5)) > 2(\gamma(6)-\gamma(5)))$	$0,000659928$
	7	64	$P(\chi^2(\gamma(7)-\gamma(6)) > 2(\gamma(7)-\gamma(6)))$	$0,00000361702$
$\|E\|=3$	2	12	$P(\chi^2(\gamma(2)-\gamma(1)) > 2(\gamma(2)-\gamma(1)))$	$0,0203410$
	3	36	$P(\chi^2(\gamma(3)-\gamma(2)) > 2(\gamma(3)-\gamma(2)))$	$0,000340357$
	4	108	$P(\chi^2(\gamma(4)-\gamma(3)) > 2(\gamma(4)-\gamma(3)))$	$0,00000000333$
	5	324	$P(\chi^2(\gamma(5)-\gamma(4)) > 2(\gamma(5)-\gamma(4)))$	$< 10^{-10}$
	6	972	$P(\chi^2(\gamma(6)-\gamma(5)) > 2(\gamma(6)-\gamma(5)))$	$< 10^{-10}$
	7	2916	$P(\chi^2(\gamma(7)-\gamma(6)) > 2(\gamma(7)-\gamma(6)))$	$< 10^{-10}$
$\|E\|>6$	2	150	$P(\chi^2(\gamma(2)-\gamma(1)) > 2(\gamma(2)-\gamma(1)))$	$< 10^{-10}$
	3	900	$P(\chi^2(\gamma(3)-\gamma(2)) > 2(\gamma(3)-\gamma(2)))$	$< 10^{-10}$
	4	5400	$P(\chi^2(\gamma(4)-\gamma(3)) > 2(\gamma(4)-\gamma(3)))$	$< 10^{-10}$
	5	32400	$P(\chi^2(\gamma(5)-\gamma(4)) > 2(\gamma(5)-\gamma(4)))$	$< 10^{-10}$
	6	194400	$P(\chi^2(\gamma(6)-\gamma(5)) > 2(\gamma(6)-\gamma(5)))$	$< 10^{-10}$
	7	1166400	$P(\chi^2(\gamma(7)-\gamma(6)) > 2(\gamma(7)-\gamma(6)))$	$< 10^{-10}$

$$\begin{aligned}\sum_{i=2}^{7} P(\hat{r}_{aic}=i) &> P\left(2[log\,\hat{L}(2)-log\,\hat{L}(1)] > 2[\gamma(2)-\gamma(1)]\right)\\ &= P\left(\chi^2(\gamma(2)-\gamma(1)) > 2[\gamma(2)-\gamma(1)]\right)\\ &\cong 0,13\end{aligned}$$

e

$$\begin{aligned}\sum_{i=2}^{7} P(\hat{r}_{aic}=i) &< \sum_{i=2}^{7} P\left(2[log\,\hat{L}(i)-log\,\hat{L}(i-1)] > 2[\gamma(i)-\gamma(i-1)]\right)\\ &= \sum_{i=2}^{7} P\left(\chi^2(\gamma(i)-\gamma(i-1)) > 2[\gamma(i)-\gamma(i-1)]\right)\\ &\cong 0,13+0,09+0,05=0,27\end{aligned}$$

Por outro lado, se considerarmos o caso $|E|=6$ e $r=1$, essa probabilidade é

$$\begin{aligned}\sum_{i=2}^{7} P(\hat{r}_{aic}=i) &< \sum_{i=2}^{7} P\left(2[log\,\hat{L}(i)-log\,\hat{L}(i-1)] > 2[\gamma(i)-\gamma(i-1)]\right)\\ &= \sum_{i=2}^{7} P\left(\chi^2(\gamma(i)-\gamma(i-1)) > 2[\gamma(i)-\gamma(i-1)]\right)\\ &\cong 6\cdot 10^{-10}\end{aligned}$$

Isso justifica os resultados encontrados nas simulações, onde o AIC aparenta convergir para a ordem verdadeira para os modelos mais complexos.

Observa-se que Katz desenvolveu as contas apenas para um caso simples, em que o AIC apresenta uma probabilidade substancial de superestimação. Os outros casos não foram mencionados. Isso e outras indicações induzem ao pensamento errôneo: "O AIC erra muito sempre."

2.3 Um Exemplo de Aplicação

Nas simulações realizadas foram consideradas amostras geradas por algoritmos, que representavam modelos markovianos "perfeitos". Entretanto, como observou Akaike (1974), a hipótese da existência de uma Cadeia de Markov, estacionária, que gerou a amostra pode mudar completamente o comportamento dos estimadores. Por isso, é interessante observar como os métodos se comportam em "dados reais".

Uma aplicação simples e interessante de Cadeias de Markov de ordem superior é a proposta por McAlpine, Miranda & Hoggar (1999), que sugere a utilização desse procedimento na modelagem de músicas. Isso pode ser utilizado não apenas para gerar músicas aleatoriamente, mas também para analisar/classificar composições existentes e gerar novas músicas a partir dessas.

Nesse sentido, foi escolhido, a "Serenata N^o 13" de Mozart, em função da sua grande quantidade de notas musicais para a voz considerada (total de 21233). Os resultados estão na Tabela 2.10.

Como pode ser observado, os resultados da Tabela 2.10, são satisfatórios para assumir a ordem como maior ou igual a 3, mas não para assumi-la como 3. Mesmo assim, é possível notar que os comportamentos dos estimadores foram semelhantes aos verificados nas simulações: para n pequeno todos subestimaram a ordem; o AIC teve melhor performance no início; o EDC_{opt} foi mais eficiente que o BIC.

No caso considerado $|E| = 7$, a probabilidade de superestimação do AIC é pequena e o tamanho da amostra não foi grande o suficiente para a ocorrência de superestimação.

Tabela 2.10: Ordens Indicadas pelos Estimadores para a “Serenata N^o 13” de Mozart

n	EDC_{opt}	BIC	AIC
91	0	0	0
101	0	0	1
161	1	0	1
171	1	0	1
181	1	1	1
581	1	1	2
1261	2	1	2
2851	2	2	2
4561	2	2	3
12871	3	2	3
21231	3	2	3

Este exemplo de aplicação a dados reais é bastante simples e a análise do comportamento dos estimadores em dados reais mais relevantes, como por exemplo dados meteorológicos, será objeto de estudos futuros.

Conclusão

As simulações realizadas indicaram que o estimador EDC_{opt} tem melhor performance que o BIC e que essa diferença aumenta em função da complexidade das Cadeias de Markov em análise.

Como também verificado por Katz (1981) para os estimadores AIC e BIC, observou-se uma tendência desses estimadores e do EDC_{opt} a subestimar a ordem quando o tamanho da amostra não é suficientemente grande.

Vale ressaltar que Katz (1981), argumentando a inconsistência do AIC, realizou simulações apenas para o caso $|E| = 2$ e $r = 1$, em que o AIC apresenta probabilidade substancial de superestimação. Entretanto, verificamos no nosso trabalho que para casos de maior complexidade essa probabilidade pode ser consideravelmente pequena, até mesmo insignificante.

Referências Bibliográficas

Akaike, H. 1974. "A new look at the statistical model identification." *Automatic Control, IEEE Transactions on* 19(6):716–723.

Anderson, T. W. & Leo A. Goodman. 1957. "Statistical Inference about Markov Chains." *The Annals of Mathematical Statistics* 28(1):89–110.

Balzter, Heiko. 2000. "Markov chain models for vegetation dynamics." *Ecological Modelling* 126(2-3):139–154.

Bartlett, M. S. 1951. "The frequency goodness of fit test for probability chains." *Proceedings of the Cambridge Philosophical Society* .

Benoît, Gerald. 2005. "Application of Markov chains in an interactive information retrieval system." *Inf. Process. Manage.* 41(4):843–857.

Billingsley, Patrick. 1961. "Statistical Methods in Markov Chains." *The Annals of Mathematical Statistics* 32(1):12–40.

Chin, E. H. 1977. "Modelling daily precipitation occurrence process with Markov chain." *Water Resources Res.* 13:949–956.

Csiszar, Imre & Paul C. Shields. 2000. "The Consistency of the *BIC* Markov Order Estimator." *The Annals of Statistics* 28(6):1601–1619.

Dacunha-Castelle, Didier, Marie Duflo & David McHale. 1986. *Probability and Statistics*. Vol. II Springer.

Doob, J. L. 1966. *Stochastic Processes (Wiley Publications in Statistics)*. John Wiley & Sons Inc.

Dorea, C. C. Y. 2008. "Optimal penalty term for EDC Markov chain order estimator." *Annales de l'Institut de Statistique de l'Universite de Paris (l'ISUP)* 52:15–26.

Dorea, C. C. Y. & J. S. Lopes. 2006. "Convergence Rates for Markov Chain Order Estimates Using EDC Criterion." *Bulletin of the Brazilian Mathematical Society* 37:561–570.

Dorea, C. C. Y. & L. Zhao. 2004. "Exponential Bounds for the Rate of Convergence of the EDC Criterion." *In: IX Congreso LatinoAmericano de Probabilidad y Estadistica Matematica* .

Feller, William. 1968. *An Introduction to Probability Theory and Its Applications, Volume 1*. Wiley.

Gates, P. & H. Tong. 1976. "On Markov chain modeling to some weather data." *J. Appl. Meteor.* 15:1145–1151.

Good, I. J. 1955. "The Likelihood Ratio Test for Markoff Chains." *Biometrika* 42(3/4):531–533.

Hoel, Paul G. 1954. "A Test for Markoff Chains." *Biometrika* 41(3/4):430–433.

Hoon, Michiel J. L., Seiya Imoto & Satoru Miyano. 2002. Inferring Gene Regulatory Networks from Time-Ordered Gene Expression Data Using Differential Equations. In *DS '02: Proceedings of the 5th International Conference on Discovery Science*. London, UK: Springer-Verlag pp. 267–274.

Kannan, D. 1979. *Introduction to Stochastic Processes*. Elsevier Science.

Katz, Richard W. 1981. "On Some Criteria for Estimating the Order of a Markov Chain." *Technometrics* 23(3):243–249.

Kendall, Maurice, Alan Stuart & Keith J. Ord. 1991. *Advanced Theory of Statistics: Classical Inference and Relationship*. Vol. 2 6th ed. Oxford, UK: Oxford University Press.

Kuha, Jouni. 2004. "AIC and BIC: Comparisons of Assumptions and Performance." *Sociological Methods Research* 33(2):188+.

Kullback, S. 1959. *Information theory and statistics*. New York: John Wiley and Sons.

Kullback, S. & R. A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22(1):79–86.

Lewin, Benjamin. 2004. *Genes VIII*. Upper Saddle River, NJ: Pearson Prentice Hall.

Li, Weidong. 2007. "A Fixed-Path Markov Chain Algorithm for Conditional Simulation of Discrete Spatial Variables." *Mathematical Geology* .

Lopes, Jaques Silveira. 2005. Determinação da Ordem de uma Cadeia de Markov Usando o Critério EDC PhD thesis Universidade de Brasília, UNB, Brasil.

Martell, David L. 1999. "A Markov chain model of day to day changes in the Canadian forest fire weather index." *International Jornal of Wildland Fire* 9:265–273.

McAlpine, Kenneth, Eduardo Miranda & Stuart Hoggar. 1999. "Making Music with Algorithms: A Case-Study System." *Comput. Music J.* 23(2):19–30.

Meyn, S. P. & R. L. Tweedie. 1993. *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

Nuel, Gregory. 2007. "Numerical Solutions for Patterns Statistics on Markov Chains." *Statistical Applications in Genetics and Molecular Biology* 5(1):26.

Park, S. K. & K. W. Miller. 1988. "Random number generators: good ones are hard to find." *Commun. ACM* 31(10):1192–1201.

Raftery, Adrian E. 1985. "A Model for High-order Markov Chains." *J. R. Statist. Soc. B.* .

Rao, C. R. 1973. *Linear Statistical Inference and its Applications*. 2nd ed. New York: J. Wiley and Sons.

Rose, R J, D M Dick, R J Viken & J Kaprio. 2001. "Gene-environment interaction in patterns of adolescent drinking: regional residency moderates longitudinal influences on alcohol use." *Clinical and Experimental Research* pp. 637–43.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6(2):461–464.

Shao, J. 2007. *Mathematical Statistics*. New York: Springer Verlag.

Shibata, R. 1976. "Selection of the Order of an Autoregressive model by Akaike's Information Criterion." *Biometrika* 63:117–126.

Silos, Pedro. 2006. "Assessing Markov chain approximations: A minimal econometric approach." *Journal of Economic Dynamics and Control* 30(6):1063–1079.

Tong, H. 1975. "Determination of the Order of a Markov Chain by Akaike's Information Criterion." *Journal of Applied Probability* 12(3):488–497.

Yamaoka, Kiyoshi, Terumichi Nakagawa & Toyozo Uno. 2005. "Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations." *Journal of Pharmacokinetics and Pharmacodynamics* .

Zhao, L., C. Dorea & C. Gonçalves. 2001. "On Determination of the Order of a Markov Chain." *Statistical Inference for Stochastic Processes* 4(3):273–282.

APÊNDICE A – Recursos Computacionais Utilizados

Sem dúvidas, a grande dificuldade em se gerar simulações em escala tecnicamente significativa se reside na criação do ambiente computacional adequado e eficiente.

Dentro desse problema, podemos citar:

- **A escolha da linguagem** – Geralmente linguagens mais fáceis de utilizar não são as mais eficientes computacionalmente. Por outro lado, em alguns casos, linguagens eficientes são de difícil manutenção e programação. Para contornar esse problema, pode-se utilizar diferentes linguagens em rotinas distintas;
- **Adequação do volume de dados** – A geração de muitos dados necessita de maior espaço para armazenagem e maior capacidade computacional para gerenciá-lo. Deve-se considerar a quantidade estritamente necessária para os resultados desejados.

A seguir, apresentamos uma pequena parcela do trabalho realizado na criação dos programas e *scripts* para a geração das simulações, relatórios e gráficos. Para isso foi considerado as seguintes premissas:

- **Eficiência** – Rápido computacionalmente;
- **Escalabilidade** – Possibilidade de aumentar a velocidade agregando mais poder de processamento.

A.1 Programa

Dentre as linguagens avaliadas (“R”, “C”, “C++”, “Perl”, “Python” e “PHP”), notou-se, indubitavelmente, que a linguagem “C” é a que melhor atendia as premissas postas. Além disso, como os procedimentos são relativamente simples, não haveria grande impacto na facilidade de programação. Para a geração de relatórios foi utilizado a ferramenta “AWK”; os gráficos foram gerados utilizando-se do aplicativo “GnuPlot”; os dados armazenados no banco de dados “PostgreSQL”.

Para solucionar o problema da escalabilidade, o programa foi dividido em pequenos módulos[1], que podem trabalhar em vários computadores em paralelo[2].

No banco de dados, foram criadas tabelas para salvar as seguintes informações[3]:

- **Cadeias de Markov** – Com suas respectivas matrizes de transição;
- **Amostras** – Amostras geradas pelas cadeias;
- **Log-verossimilhanças Estimadas** – Valores de $\hat{L}(k)$, $k = 0..7$, de cada tamanho de certa amostra;
- **Ordens Estimadas** – De cada Log-verossimilhança estimada;
- **Tarefas** – Utilizadas para orientar os trabalhos dos módulos.

Nas simulações principais, foram salvas apenas as ordens estimadas para cada estimador juntamente com alguns valores da log-verossimilhança.

A.1.1 Descrição das Principais Rotinas

As rotinas principais criadas foram:

- **Simular** – Desempenha o trabalho principal, gerando os modelos aleatórios e salvando os resultados numéricos;

[1] Arquitetura *dashboard*
[2] Trabalho em *cluster*
[3] Entidades.

- **Gerar Relatório** – Cria os indicadores apresentados nesse trabalho;
- **Gerar Gráfico** – Cria figuras para facilitar a identificação de "padrões" de comportamento.

Além dessas, foram geradas outras rotinas que auxiliaram nos testes e verificações. Estas não são enfatizadas nessa dissertação, mas também estão disponíveis.

A geração de relatórios e gráficos é realizada computando diretamente no banco de dados [usando a linguagem SQL]. A geração do relatório em Latex é feita utilizando o aplicativo AWK. Como a rotina "Simular" é a principal nesse trabalho, apresentamos a descrição do seu funcionamento abaixo.

1. Recupera no banco de dados uma tarefa a ser executada, obtendo os parâmetros da ordem e tamanho do espaço de estados;
2. Cria na memória a matriz de transição. Para cada probabilidade condicionada (linha da matriz) é gerada a distribuição particionando aleatoriamente, de forma uniforme o intervalo $[0,1]$ e considerando as partições de forma ordenada (quando utilizado o modelo proposto por Raftery (1985) a lógica é a mesma);
3. Gera uma amostra na memória com comprimento de 100 milhões. Para iniciar a amostra é sempre considerado o condicionamento por "000...0";
4. Para cada "sub-amostra":
   (a) Atualiza a matriz de contagem [aqui há o ganho ao considerar as "sub-amostras"];
   (b) Calcula os valores da log-verossimilhança (de 0 a 7) e salva no banco;
   (c) Calcula os estimadores $\hat{r}_{edc}$, $\hat{r}_{bic}$ e $\hat{r}_{aic}$, salvando-os no banco;

A.2 Estimativas

O tempo de execução de cada rotina independente varia com o tipo da tarefa e/ou com o volume de dados envolvido. Para fins comparativos, apresentamos os tempos aproximados

medidos[4] na Tabela A.1.

Tabela A.1: Tempos de execução das rotinas

Rotina	Tempo Aproximado de Execução
Simular	5 a 180 segundos
Gerar Relatório	1 segundo a 2 dias
Gerar Gráfico	1 a 30 segundos

Pode-se rodar diversas rotinas em um mesmo computador, considerando uma por processador. Além disso, as rotinas podem trabalhar em computadores distintos ao mesmo tempo. Isso aumenta consideravelmente a velocidade dos trabalhos.

A.3 Ambiente Utilizado

Para esse trabalho foi utilizado computadores com processadores AMD Turion 64. O sistema operacional utilizado foi o Ubuntu Linux. Dentre as principais ferramentas destacamos o compilador GCC versão 4.3.1 e a libc6 versão 2.7. Os números aleatórios foram obtidos a partir das bibliotecas propostas por Park & Miller (1988).

[4]Em um computador mono-processado AMD Turion 64 X2 800 Mhz.