Cadeias de Markov de Alcance Variável: um Estudo de Convergência do Estimador

Carla V. Otero

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do grau
de
Mestrado em Estatística

Área de Concentração: **Probabilidade**
Orientadora: **Profa. Dra. Cláudia Peixoto**

-São Paulo, Março de 2005-

Ao meu pai José Otero (in memorian)

e a minha mãe Neide.

Conteúdo

Introdução **5**

1 Cadeias de Markov de Alcance Variável **7**

1.1 Definição 7

1.2 Representação Gráfica 8

2 Algoritmo do Contexto **13**

2.1 Descrição 13

2.1.1 Algoritmo do Contexto 14

2.2 Comentários 15

3 Simulação **17**

3.1 Introdução 17

3.2 Metodologia Utilizada 18

3.3 Resultados 20

3.3.1 Resultados para CMAV1 21

3.3.2 Resultados para CMAV2 26

4 Aplicação **31**

4.1 Introdução 31

4.2 Metodologia Utilizada 33

4.3 Resultados 34

5 Conclusão **37**

Introdução

As Cadeias de Markov de Alcance Variável (CMAV) têm sua origem na Teoria da Informação (Rissanen, [8]). Mais recentemente, sob uma abordagem estatística, esses modelos vêm sendo utilizados em substituição às cadeias de ordem fixa, em particular por Buhlman e Wyner, em 1999 [5], cujo trabalho fundamenta o presente estudo.

O modelo mais utilizado para processos estacionários, assumindo valores num espaço de estados categorizados $\mathcal{X}$, é a Cadeia de Markov Completa de ordem finita k, $k \geq 1$. Trata-se, portanto, de uma Cadeia de Markov estacionária definida em todo espaço $\mathcal{X}^k$.

Esse modelo não é apropriado do ponto de vista estatístico, pois o número de parâmetros a ser estimado cresce consideravelmente em relação à ordem da cadeia, isto é, teremos $(|\mathcal{X}| - 1)|\mathcal{X}|^k$ parâmetros, onde $|\mathcal{X}|$ é a cardinalidade do espaço $\mathcal{X}$. Por exemplo, mesmo no caso onde $|\mathcal{X}| = 2$, o número de parâmetros a ser estimado é 2^k, onde k é a ordem da cadeia. No caso de estar-se tratando de uma cadeia de ordem 10, 1.024 parâmetros devem ser estimados.

As CMAV, que também são Cadeias de Markov de ordem finita, minimizam o problema acima, pois não são cadeias plenas. Tais cadeias pressupõem que a memória relevante na previsão de um evento pode variar em tamanho, de acordo com o ocorrido, reduzindo, assim, sua ordem. Dessa forma, a quantidade de parâmetros a ser estimada é inferior a $(|\mathcal{X}| - 1)|\mathcal{X}|^k$, viabilizando a estimativa em questão.

O presente trabalho apresenta, no Capítulo 1, os principais conceitos e definições relativas às CMAV. No Capítulo 2 introduzimos o Algoritmo do Contexto (ver Buhlman & Wyner, [5]), método de estimativa das CMAV. O Capítulo 3 apresenta os resultados referentes às estimativas dos contextos para uma CMAV previamente gerada. Neste

capítulo as simulações foram realizadas para diversos parâmetros (*valor de corte*) do algoritmo, tendo como objetivo, o estudo do comportamento das estimativas. No Capítulo 4 apresentamos uma aplicação do modelo estudado, desenvolvento um modelo de *perfil de comportamento* para clientes com produto de crédito de longo prazo. No Capítulo 5 encontram-se as conclusões sobre o desempenho do algoritmo estudado.

Capítulo 1

Cadeias de Markov de Alcance Variável

Uma maneira de minimizar o número de parâmetros imposto pelas Cadeias de Markov de ordem fixa e finita k é utilizar um modelo que permite que a probabilidade condicional de um evento possa depender de um número variável e limitado de variáveis aleatórias. Esse é o intuito das definições que seguem.

1.1 Definição

Os processos a serem considerados, salvo menção explícita em contrário, serão Cadeias de Markov de ordem finita, estacionárias que assumem valores num espaço categorizado finito $\mathcal{X}$ e indexadas pelo conjunto $\mathbb{Z}$ dos números inteiros unido aos valores $\{-\infty, \infty\}$. Denotaremos o conjunto dos índices $\mathbb{Z} \cup \{-\infty, \infty\}$ por $\overline{\mathbb{Z}}$. Considere x_i^j a seqüência $x_j, \cdots, x_i$ de elementos de $\mathcal{X}$, com $i < j$, $i \in \overline{\mathbb{Z}}$, $j \in \overline{\mathbb{Z}}$. Um processo $(X_t)_{t \in \overline{\mathbb{Z}}}$ é uma Cadeia de Markov estacionária de ordem k se para todo $x_{-\infty}^t$ valer

$$\mathbb{P}[X_t = x_t | X_{-\infty}^{t-1} = x_{-\infty}^{t-1}] = \mathbb{P}[X_t = x_t | X_{t-k}^{t-1} = x_{t-k}^{t-1}], \ \forall t \in \overline{\mathbb{Z}}. \tag{1.1}$$

O que se deseja é permitir que o número de passos do passado seja variável e menor ou igual a k sem ter que ser sempre k, ou ainda que o número de passos dependa dos valores assumidos sem ter que ser necessariamente constante. As duas próximas definições foram elaboradas com esse objetivo.

Definição 1.1.1. Seja $(X_t)_{t \in \overline{\mathbb{Z}}}$ um processo estocástico estacionário assumindo valores num espaço finito $\mathcal{X}$. Dado $x_{-\infty}^0$ em $\mathcal{X}^\infty$, seja $\ell = \ell(x_{-\infty}^0)$ o menor valor de $k \in \overline{\mathbb{Z}}$, $k \geq 1$, tal que para todo $x_1 \in \mathcal{X}$ valha $\mathbb{P}[X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0] = \mathbb{P}[X_1 = x_1 | X_{-k+1}^0 = x_{-k+1}^0]$. No caso especial em que as variáveis são independentes tomaremos $\ell = 0$ ao invés de 1, que é o

menor k, $k \geq 1$, que satisfaz a equação acima. A projeção $x_{-\infty}^{0} \in \mathcal{X}^{\infty} \xrightarrow{c} x_{-\ell+1}^{0} \in \bigcup_{m=0}^{\infty} \mathcal{X}^{m}$ é chamada de uma *função contexto* e, para todo $t \in \mathbb{Z}$, $c(x_{-\infty}^{t-1})$ é chamado *o contexto para o valor* x_t. Chamaremos $\mathcal{C} = \{c(x_{-\infty}^{0}) \mid x_{-\infty}^{0} \in \mathcal{X}^{\infty}\}$ o *conjunto contexto* para a função contexto c.

Podemos, agora, definir uma Cadeia de Markov de Alcance Variável.

Definição 1.1.2. Seja $(X_t)_{t\in\mathbb{Z}}$ um processo estocástico estacionário assumindo valores num espaço finito $\mathcal{X}$ cuja função contexto é $c : \mathcal{X}^{\infty} \to \bigcup_{m=0}^{\infty} \mathcal{X}^{m}$. O máximo k, $0 \leq k \leq \infty$, do conjunto $\{\ell(x_{-\infty}^{0}) \mid x_{-\infty}^{0} \in \mathcal{X}^{\infty}\}$ chama-se *ordem* da função contexto c. Quando k é finito, dizemos que $(X_t)_{t\in\mathbb{Z}}$ é uma *Cadeia de Markov de Alcance Variável (CMAV) de ordem* k.

Naturalmente, para uma CMAV de ordem k, seu conjunto contexto é finito de cardinalidade máxima $|\mathcal{X}|^k$. Para cada um dos elementos do conjunto contexto teremos $|\mathcal{X}| - 1$ parâmetros independentes a serem estimados. Assim, a quantidade de parâmetros é $|\mathcal{X}|-1$ vezes a quantidade de elementos do conjunto contexto.

1.2 Representação Gráfica

Uma CMAV de ordem máxima k pode ser completamente descrita através de suas probabilidades de transição. A função contexto c será a projeção de $\mathcal{X}^{\infty}$ sobre $\mathcal{X}^{k}$, e o conjunto contexto dos estados passados relevantes é o produto cartesiano $\mathcal{X}^{k}$. No caso mais geral, considera-se a relação de equivalência em $\mathcal{X}^{k}$ definida por $\vec{x} = (x_1, \ldots, x_k) \equiv (y_1, \ldots, y_k) = \vec{y}$ se, e somente se, existir um* $z_{-\infty}^{-k-1}$ tal que $c(\vec{x}z_{-\infty}^{-k-1}) = c(\vec{y}z_{-\infty}^{-k-1})$, onde $\vec{x}z_{-\infty}^{-k-1}$ é a concatenação $x_1 \cdots x_k z_{-\infty}^{-k-1}$. Podemos identificar uma CMAV genérica de ordem k com o conjunto quociente de $\mathcal{X}^{k}$ pela relação de equivalência $\mathcal{X}^{k}/\equiv$.

Produtos cartesianos finitos são usualmente representados por árvores. Por exemplo, se desejarmos representar um produto cartesiano $A \times B \times C$, começamos por um nó raiz, a partir do qual emanam tantos nós quanto elementos do conjunto A. De cada nó originado no passo anterior, emanam tantos nós quanto os elementos de B e, de cada caminho assim

* Ou, alternativamente, se para todo $z_{-\infty}^{-k-1}$ vale $c(\vec{x}z_{-\infty}^{-k-1}) = c(\vec{y}z_{-\infty}^{-k-1})$.

obtido, raiz$\longrightarrow$ nó de $A \longrightarrow$ nó de B, emanam tantos nós quanto os elementos de C. O resultado é uma árvore formada por caminhos raiz$\longrightarrow$ nó de $A \longrightarrow$ nó de $B \longrightarrow$ nó de C.

Como identificamos uma CMAV de ordem k com um conjunto quociente $\mathcal{X}^k/\equiv$, podemos também identificá-la com uma subárvore da árvore que represente $\mathcal{X}^k$ da seguinte forma: partindo da raiz, buscamos todos os elementos de $\mathcal{X}$ que sejam primeiras componentes de $k-$uplas não equivalentes; depois, a partir de cada um desses nós, todos os elementos de $\mathcal{X}$ que sejam segundas componentes (tendo por primeira componente a do nó) de $k-$uplas não equivalentes e assim por diante. Quando, a partir de um nó, todos os elementos de $\mathcal{X}$ junto com o ramo até aí construído formarem $k-$uplas equivalentes, encerramos o ramo. O resultado obtido é a representação da CMAV de ordem k sob a forma de uma árvore.

O exemplo a seguir esclarece melhor os conceitos introduzidos até agora.

Exemplo 1.2.1. Vamos supor que $\mathcal{X} = \{\text{sol}, \text{nublado}, \text{chuva}, \text{temporal}\}$ e $k = 2$. Admitamos que a função contexto do processo seja:

$$c(x_{-\infty}^{0}) = \begin{cases} \text{sol}, & \text{se } x_0 = \text{sol e } x_{-\infty}^{-1} \text{ qualquer}, \\ \text{nublado}, & \text{se } x_0 = \text{nublado e } x_{-\infty}^{-1} \text{ qualquer}, \\ \text{chuva}, & \text{se } x_0 = \text{chuva e } x_{-\infty}^{-1} \text{ qualquer}, \\ \text{temporal}, & \text{se } x_0 = \text{temporal}, x_{-1} \in \{\text{sol, nublado, chuva}\} \text{ e } x_{-\infty}^{-2} \text{ qualquer}, \\ \text{temporal, temporal} & \text{se } x_0 = \text{temporal}, x_{-1} = \text{temporal e } x_{-\infty}^{-2} \text{ qualquer}. \end{cases}$$

A função contexto acima estabelece que se hoje há sol ($x_0 = \text{sol}$), então a probabilidade de X_1 estar em algum estado de $\mathcal{X}$ independe do que ocorreu ontem, anteontem,... $(x_{-1}, x_{-2}, \ldots)$. Isso também pode ser expresso escrevendo $\ell(\text{sol}\, x_{-\infty}^{-1}) = 1$, para qualquer $x_{-\infty}^{-1}$. O que acabou de ser dito vale também quando $x_0 = \text{nublado}$ ou $x_0 = \text{chuva}$. Quando $x_0 = \text{temporal}$ a situação é diferente, isto é, a probabilidade de X_1 asssumir um estado de $\mathcal{X}$ depende de quem precedeu x_0. Desta forma se x_{-1} for sol, nublado ou chuva, temos uma probabilidade de transição; caso x_{-1} seja temporal, teremos outra (presumivelmente,

nesse caso a probabilidade de X_1 ser chuva ou temporal seria maior que no caso anterior), ou seja $\ell(\text{temporal sol}\, x_{-\infty}^{-2}) = \ell(\text{temporal nublado}\, x_{-\infty}^{-2}) = \ell(\text{temporal chuva}\, x_{-\infty}^{-2}) = 1$ e $\ell(\text{temporal temporal}\, x_{-\infty}^{-2}) = 2$. Veja a figura a seguir:

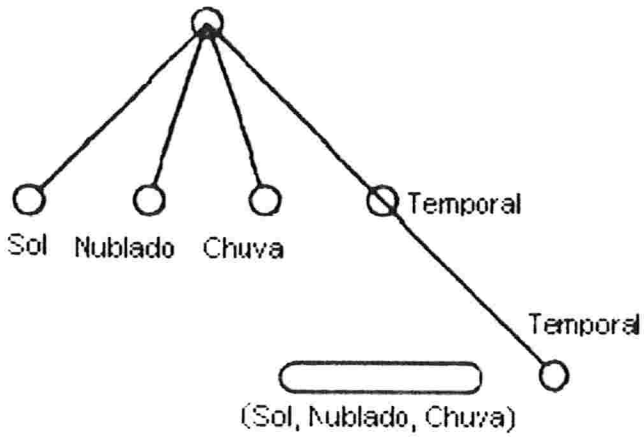

Figura 1.1 *Representação de uma CMAV sob a forma de árvore.*

Na figura acima, o retângulo arrendondado abaixo do primeiro nó formado por temporal indica que os pares (temporal, sol), (temporal,nublado), e (temporal,chuva) são equivalentes.

Um outro modo equivalente de se definir a árvore que representa uma CMAV é o seguinte:

Definição 1.2.1. Dada uma CMAV cuja função contexto é c, definimos sua *árvore contexto* como sendo

$$\tau = \tau_c = \{c(x_{-\infty}^0) \mid x_{-\infty}^0 \in \mathcal{X}^\infty\}.$$

A *árvore contexto de nós terminais* é definida como sendo

$$\tau^T = \tau_c^T = \{w \in \tau_c \mid \forall u \in \mathcal{X},\ wu \notin \tau_c\}.$$

Onde $wu = (w_{|w|}, \cdots, w_2, w_1, u_{|u|}, \cdots, u_2, u_1)$ é a concatenação dos vetores w e u.

Doravante, trabalharemos com a definição que acabamos de introduzir. Seu significado ficará mais claro através de um exemplo. Para tanto, voltemos ao exemplo (1.2.1).

De acordo com a definição, teremos $\tau = \{$sol, nublado, chuva, temporal, (temporal,temporal)$\}$. Representaremos τ graficamente por meio de uma árvore $\mathcal{A}_\tau$ definida como segue:

1. Independente de τ ser ou não vazio, $\mathcal{A}_\tau$ tem um único nó raiz.

2. Um elemento $(x_1, \ldots, x_n)$ de τ será representado em $\mathcal{A}_\tau$ como o ramo onde x_1 é o passo anterior da raiz, e x_i é o passo anterior de x_{i-1}, $2 \leqq i \leqq n$.

3. Todo ramo de $\mathcal{A}_\tau$ é obtido por uma das formas anteriormente descritas.

De acordo com o exposto, a representação gráfica de τ seria como a figura abaixo.

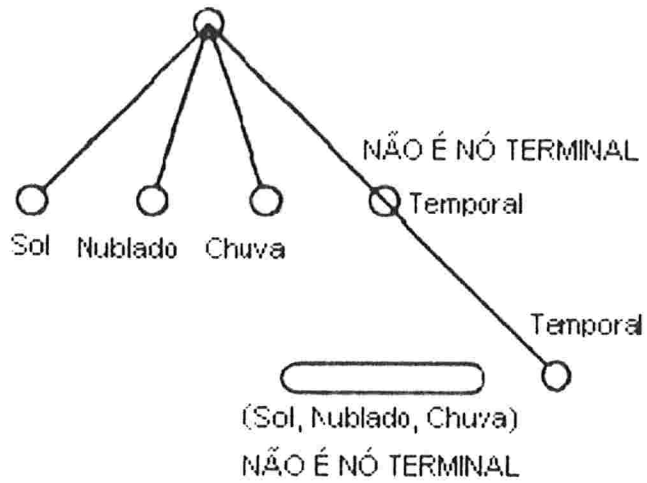

Figura 1.2 *Representação gráfica de τ.*

Para simplificar, usaremos τ ao invés de $\mathcal{A}_\tau$, isto é, o símbolo será usado tanto para designar o conjunto quanto sua representação gráfica.

A noção de árvore contexto de nós terminais será útil quando apresentarmos o Algoritmo do Contexto. Um nó é considerado terminal se, ao adicionarmos uma componente ($x_i \in \mathcal{X}$) a um ramo qualquer, esse novo ramo não pertencer à árvore contexto. Por exemplo, o ramo (temporal, temporal) é nó terminal, pois se adicionarmos uma componente, (temporal), por exemplo, o novo ramo (temporal, temporal, temporal) não pertence à arvore. Agora, o nó (temporal) não é terminal, pois ao adicionarmos a componente (temporal), o ramo (temporal, temporal) faz parte da árvore contexto. No exemplo em questão, $\tau^T = \{$sol, nublado, chuva, (temporal, temporal)$\}$ é o conjunto de nós terminais. Note que τ^T nada

mais é que a subárvore de τ da qual foram retirados os ramos não terminais. É importante notar que a árvore de nós terminais determina completamente a árvore contexto que a gerou, isto é, o conjunto τ_c pode ser reconstruído a partir do conjunto τ_c^T.

A árvore contexto para uma CMAV funciona como uma representação "palpável" para esse tipo de processo estocástico. Queremos com isso dizer que a árvore contexto está para a CMAV, assim como uma função de distribuição está para uma variável aleatória, isto é, a árvore contexto serve como um objeto que resume o processo da CMAV. Prosseguindo nesta analogia, as probabilidades condicionais associadas aos nós da árvore contexto podem ser pensadas como os análogos dos parâmetros de uma distribuição. É natural, então, que se busque estimadores para a árvore e as probabilidades condicionais, o que nos leva ao próximo capítulo.

Capítulo 2

Algoritmo do Contexto

Neste capítulo vamos apresentar o *Algoritmo do Contexto* que é um método para se estimar a árvore de contextos e as probabilidades condicionais associadas a uma CMAV. Embora ele possa ser formulado para seqüências de CMAV, o que permite sua utilização no estudo de processos estacionários em geral a valores num espaço categorizado finito (ver Ferrari e Wyner [3]), apresentaremos uma versão que se aplica a uma única CMAV.

2.1 Descrição

Seja P uma CMAV de ordem finita, para cada $n \in \mathbb{N}$, considere $x_1, \ldots, x_n$ uma realização finita de P, $c(\cdot)$ sua função contexto, e $p(\cdot \mid \cdot)$ sua probabilidade de transição. Antes de apresentarmos o algoritmo, que produzirá uma estimativa de $p(\cdot \mid \cdot)$ e c, necessitamos introduzir algumas convenções e notações.

Para $x = x_i^j$, vamos adotar que $|x| = j - i + 1$. Para $w \in \bigcup_{m \in \mathbb{N}} \mathcal{X}^m$, sejam

$$N(w) = \sum_{t=1}^{n-|w|+1} 1_{[X_t^{t+|w|-1} = w]}, \qquad N_{-k}(w) = \sum_{t=1}^{n-|w|+1-k} 1_{[X_t^{t+|w|-1} = w]}. \tag{2.1}$$

Para facilitar a exposição, vamos dizer que uma seqüência qualquer x_i^j é uma *palavra*(sequência de letras). Note que, dada uma palavra w, $N(w)$ é simplesmente a quantidade de vezes que a palavra ocorre na realização de $X_1, \ldots, X_n$ e, por sua vez, $N_{-k}(w)$ tem a mesma interpretação que $N(w)$, só que, agora, para $X_1, \ldots, X_{n-k}$.

Relembremos que, dados x, $w \in \bigcup_{m \in \mathbb{N}} \mathcal{X}^m$, xw é a concatenação de x com w. Vamos

definir, então,

$$\hat{P}(w) = \frac{N(w)}{n}, \qquad \hat{P}(x|w) = \frac{N(xw)}{N_{-|x|}(w)}. \tag{2.2}$$

Dado um elemento $u \in \mathcal{X}$,

$$\Delta_{wu} = \sum_{x \in \mathcal{X}} \hat{P}(x|wu) \log(\frac{\hat{P}(x|wu)}{\hat{P}(x|w)}) N(wu). \tag{2.3}$$

De posse dessas notações podemos, agora, apresentar o Algoritmo do Contexto.

2.1.1 Algoritmo do Contexto

A seguir, relacionaremos os passos necessários para a estimativa da árvore contexto.

Entradas: $x_1, \ldots, x_n$, presumivelmente dados de uma amostra, e K, um parâmetro fornecido pelo usuário (e que será discutido posteriormente).

Passos:

1. Para cada i, $1 \leqq i \leqq n$, e para cada $w \in \mathcal{X}^i$, calcule $N(w)$ (a quantidade de vezes que w aparece em $x_1, \ldots, x_n$). Se $2 \leqq N(w)$, coloque w dentro de uma árvore τ cuja inicialização era vazia.

   Após esta etapa, a árvore contém todas as realizações ($w \in \bigcup_{m \in \mathbb{N}} \mathcal{X}^m$) que ocorrem mais de uma vez na seqüencia $x_1, \ldots, x_n$.

2. Para cada elemento wu de τ, onde $u \in \mathcal{X}$, calcule Δ_{wu}. Se $\Delta_{wu} < K$, retire-o de τ.

   Ou seja, para cada elemento η obtido na primeira etapa, tomamos (se possível) como sendo $\eta = wu$, onde u é um elemento do espaço $\mathcal{X}$, e testamos se $\Delta_{wu} < K$. Se isso ocorrer, retiramos u da árvore e examinamos o próximo item. Os que sobrarem constituirão a nova árvore.

3. Repita o passo anterior até que não seja mais possível retirar nenhum elemento de τ.

   À árvore obtida no passo anterior, aplicamos novamente o mesmo procedimento (2), obtendo, a cada vez, uma nova árvore. Repetiremos esses procedimentos até que não consigamos retirar mais nenhum componente ($u \in \mathcal{X}$).

4. Para cada elemento $x \in \mathcal{X}$ e para cada elemento w de τ, calcule $\hat{P}(x|w)$ de acordo com a fórmula apresentada na equação (2.2).

Saídas: A árvore τ e a coleção de números $\hat{P}(x|w)$, onde x, $w \in \tau$, conforme definidos no último do passo.

2.2 Comentários

A árvore τ, obtida a partir do Algoritmo do Contexto, é a estimativa da estrutura da CMAV cuja amostra constituem os dados de entrada $x_1, \ldots, x_n$. Trata-se aqui de sua árvore contexto de nós terminais (ver definição (1.2.1)), que, como já foi comentado, é suficiente para se determinar a árvore da CMAV.

A função Δ_{wu} é um múltiplo da *entropia relativa*, ou *distância de Kullback-Leibler**, entre duas medidas finitas de probabilidade. Lembrando que a entropia relativa, ou distância de Kullback-Leibler, entre duas medidas finitas de probabilidade P e Q é dada por $D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log(P(x)/Q(x))$, vê-se que $\Delta_{wu} = N(wu)\, D(\,\hat{P}(\cdot|wu)\,||\,\hat{P}(\cdot|w)\,)$.

O valor de K, por sua vez, é heurísticamente determinado através de uma escala de χ^2 ou por critérios informacionais. Segundo Bühlmann e Mächler ([6]), K, como limiar para diferenças de desvios, é assintoticamente associado a uma distribuição $\frac{1}{2}\chi^2_{|\mathcal{X}|-1}$. Assim $K = K(\alpha) = \frac{1}{2}\chi^2_{|\mathcal{X}|-1,\alpha}$. Nos casos em que se tem aplicado CMAV à genética, temos $\mathcal{X} = \{A, C, G, T\}$ e, portanto, $|\mathcal{X}| - 1 = 3$, logo, para $\alpha = 5\%$ e $\alpha = 1\%$, temos, respectivamente, $K = 3.91$ e $K = 5.67$. Continuando com Bühlmann e Mächler ([6]), pode-se, de forma alternativa, determinar K a partir de critérios informacionais como, por exemplo, os de AIC e BIC. Os autores mostram que a escolha de K segundo um critério ou outro são bem semelhantes. Eles também citam, na página 19 de ([6]), que, por considerações assintóticas, é suficiente tomar $K = C\log(n)$, com $2|\mathcal{X}| + 4 < C$, porém ressaltam, no caso de amostras pequenas, devido ao fato da condição $K = C\log(n)$ não ter que ser necessária, que tal escolha pode produzir resultados inadequados, com árvores subestimadas. No próximo capítulo, iremos, justamente, testar a eficiência da função

* A entropia relativa não satisfaz a propriedade triangular, de sorte que não se trata realmente de uma distância no sentido usual do termo.

contexto para amostras de tamanhos variados, quando utilizamos $K = C\log(n)$, e K que minimiza o critério AIC.

Capítulo 3

Simulação

3.1 Introdução

Neste capítulo apresentamos os resultados referentes ao estudo do parâmetro K (*valor de corte* na estimativa da função contexto). Tal parâmetro tem impacto decisivo na estimativa de uma função contexto.

Bühlmann e Mächler (ver [6]) apresentaram duas sugestões de *valores de corte*, sobre as quais o estimador apresentaria um bom desempenho: $Clog(n)$, onde $C > 2|\mathcal{X}| + 4$, e K que minimiza o AIC. Ambas foram comparadas para amostras de tamanhos variados. Para complementar o estudo, outros *valores de corte* também foram avaliados.

Para os estudos mencionados acima foram realizadas 100 simulações para cada tamanho de amostra. As simulações foram desenvolvidas para duas CMAV diferentes, uma de ordem máxima 2, e outra de ordem máxima 3, ambas definidas no espaço categorizado $\mathcal{X} = \{0, 1\}$.

A fundamentação teórica para os procedimentos de simulação pode ser encontrada na Lei dos Grandes Números. Na sua forma forte, devida a Kolmogorov, ela afirma que se $X_1, \ldots, X_n, \ldots$ são variáveis aleatórias independentes, identicamente distribuídas e com primeiro momento finito, ou seja, para todo $1 \leq i$, $\mathbb{E}|X_i| = \mu < \infty$, então a seqüência de suas médias, isto é, das variáveis $\overline{X}_n = \sum_{i=1}^{n} X_i/n$, converge quase certamente para μ. Dela derivam vários resultados estatísticos, dos quais destacamos o de Glivenko e Cantelli sobre convergência das distribuições empíricas. Ele afirma que se $X_1, \ldots, X_n, \ldots$ são variáveis aleatórias independentes, identicamente distribuídas, e com função de distribuição F, então a seqüência de funções de distribuição empíricas definidas

por $\mathbb{F}_n(x) = \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}/n$ converge quase certamente e uniformemente para F, isto é, $\sup_{-\infty<x<\infty}|\mathbb{F}_n(x) - F(x)| \to 0$ quase certamente.

3.2 Metodologia Utilizada

Como mencionado, o estudo consistiu na estimativa de funções contexto para uma série de amostras simuladas previamente. As etapas envolvidas no processo são apresentadas abaixo.

- Foram escolhidas duas CMAV, uma de ordem 2, e outra de ordem 3. Ambas definidas no espaço $\mathcal{X} = \{0, 1\}$.

  As funções contexto adotadas, e suas probabilidades são as seguintes:

  1) CMAV1

$$c(x_{-\infty}^{n}) = \begin{cases} 1, & \text{se } x_n = 1 \text{ e } x_{-\infty}^{n-1} \text{ qualquer,} \\ 0,1 & \text{se } x_n = 0,\ x_{n-1} = 1 \text{ e } x_{-\infty}^{n-2} \text{ qualquer,} \\ 0,0 & \text{se } x_n = 0,\ x_{n-1} = 0 \text{ e } x_{-\infty}^{n-2} \text{ qualquer.} \end{cases}$$

$$\begin{aligned} \mathbb{P}[X_n = 1 \mid X_{n-1} = 1] &= 0,53 \\ \mathbb{P}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 1] &= 0,12 \\ \mathbb{P}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 0] &= 0,75 \end{aligned}$$

  2) CMAV2

$$c(x_{-\infty}^{n}) = \begin{cases} 1, & \text{se } x_n = 1 \text{ e } x_{-\infty}^{n-1} \text{ qualquer,} \\ 0,1 & \text{se } x_n = 0,\ x_{n-1} = 1 \text{ e } x_{-\infty}^{n-2} \text{ qualquer,} \\ 0,0,1 & \text{se } x_n = 0,\ x_{n-1} = 0,\ x_{n-2} = 1 \text{ e } x_{-\infty}^{n-3} \text{ qualquer,} \\ 0,0,0 & \text{se } x_n = 0,\ x_{n-1} = 0,\ x_{n-2} = 0 \text{ e } x_{-\infty}^{n-3} \text{ qualquer.} \end{cases}$$

$$\mathbb{P}[X_n = 1 \mid X_{n-1} = 1] = 0,10$$
$$\mathbb{P}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 1] = 0,81$$
$$\mathbb{P}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 0, X_{n-3} = 1] = 0,27$$
$$\mathbb{P}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 0, X_{n-3} = 0] = 0,65$$

- Em seguida, foram realizadas 100 simulações para cada tamanho de amostra. As simulações foram obtidas através de uma rotina em Java apresentada no Apêndice 2.

- De posse das amostras, passamos à estimativa das funções contexto através do programa estatístico R. As rotinas utilizadas também se encontram no Apêndice 2. Inicialmente, foram estimadas as funções contexto utilizando-se os *valores de corte* $Clog(n)$, onde $C = 2|\mathcal{X}| + 5$, e também aquele que minimiza o AIC. Nesse último caso, não foram estimados 100 Ks diferentes para cada amostra, mas sim, para 5 casos escolhidos aleatoriamente, e a sua média utilizada na estimativa das funções contexto de todas as outras amostras.

- Por último, foram realizadas as estimativas para outros Ks diferentes dos mencionados acima, com a finalidade de encontrar um *valor de corte* mais eficiente. Como verificamos, a partir dos primeiros resultados, que para $K = Clog(n)$, a funcão contexto era subestimada (ordem menor que a esperada), e para K que minimiza o AIC, a função era superestimada (ordem maior que a esperada), os *valores de corte* analisados nesta etapa apresentam um valor intermediário entre os dois mencionados; a saber: $log(n), (Clog(n))^{1/2}, ((Clog(n))^{1/2} + (Clog(n))^{1/3})/2$, e $(Clog(n))^{1/3}$, onde mantivemos $C = 2|\mathcal{X}|+5$. Esses *valores de corte* apresentam a seguinte distribuição:

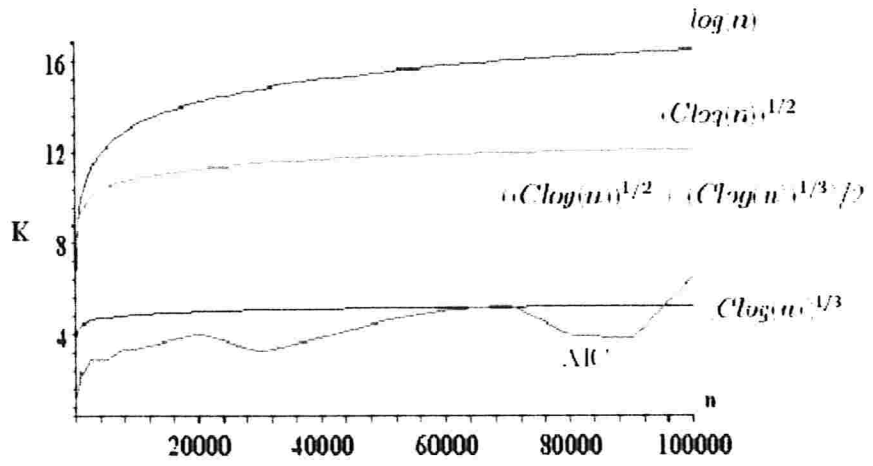

Figura 3.1 *Valores de corte utilizados*

3.3 Resultados

As tabelas a seguir apresentam os resultados consolidados dos estudos descritos em 3.2. Nelas, está a quantidade de casos que convergiram corretamente (C), a quantidade de árvores subestimadas (ordem 0 e 1 para a cadeia de ordem 2, e 0, 1, e 2, para a cadeia de ordem 3), e a quantidade de árvores que excederam a ordem da função contexto verdadeira, ou que, apesar de apresentarem a mesma ordem, não convergiram para a árvore correta (E), isso para amostras de tamanhos variados (n).

É importante ressaltar que a convergência de que estamos falando refere-se à função contexto, suas probabilidades podem apresentar uma flutuação em relação à distribuição simulada, principalmente para os casos de convergência com $n \leq 1.000$.

3.3.1 Resultados para CMAV1

$Clog(n)$					AIC				
n	**0**	**1**	**C**	**E**	**n**	**0**	**1**	**C**	**E**
50	100	0	0	0	50	3	3	25	69
100	100	0	0	0	100	0	0	8	92
250	100	0	0	0	250	0	0	0	100
500	100	0	0	0	500	0	0	4	96
750	100	0	0	0	750	0	0	0	100
1000	97	3	0	0	1000	0	0	2	98
1500	7	93	0	0	1500				
1750	0	0	100	0	1750				
2000	0	0	100	0	2000				
2500	0	0	100	0	2500	0	0	2	98
5000	0	0	100	0	5000	0	0	0	100
7500	0	0	100	0	7500	0	0	0	100
10000	0	0	100	0	10000	0	0	0	100
20000	0	0	100	0	20000	0	0	0	100
30000	0	0	100	0	30000				
40000	0	0	100	0	40000				
50000	0	0	100	0	50000				
60000	0	0	100	0	60000				
70000	0	0	100	0	70000				
80000	0	0	100	0	80000				
90000	0	0	100	0	90000				
100000	0	0	100	0	100000				

Tabela 3.1 *Resultados das simulações para $K = Clog(n)$, e K que minimiza o AIC.*

Nessa primeira tabela, apresentamos os resultados referentes à estimativa da função contexto quando $K = Clog(n)$, e K que minimiza o AIC (todos os resultados relativos à árvore de ordem 2). Nela, encontramos um melhor desempenho, quando n é pequeno, para K que minimiza o AIC, e quando n é grande, para $K = Clog(n)$.

Para esse *valor de corte* houve acerto para as 100 amostras a partir de 1.750 casos. Notar que isso ocorreu de forma abrupta, dado que até 1.500 casos não houve nenhum evento de acerto, apenas árvores subestimadas. Por exemplo, para $n = 1.500$, chegamos a 7 estimativas onde não foi encontrada nenhuma dependência com o passado (casos i.i.d., árvores de ordem 0), e o restante, 93 estimativas, onde a dependência resultou em árvores de ordem 1.

Para K que minimiza o AIC, encontramos alguns casos de acerto quando $n \leq 2.500$, porém em baixa frequência. Apenas para $n = 50$ o percentual de eventos de acerto é mais significativo - 25%. É importante observar que para esse n, e $K = Clog(n)$, não houve nenhum evento de acerto. Para as 100 simulações, o estimador apresentou como resultado uma variável i.i.d.

$log(n)$					$(Clog(n))^{1/2}$				
n	**0**	**1**	**C**	**E**	**n**	**0**	**1**	**C**	**E**
50	91	2	7	0	50	97	0	3	0
100	54	1	45	0	100	66	0	34	0
250	0	0	100	0	250	0	0	100	0
500	0	0	100	0	500	0	0	100	0
750	0	0	100	0	750	0	0	100	0
1000	0	0	100	0	1000	0	0	100	0
2500	0	0	100	0	2500	0	0	100	0
5000	0	0	100	0	5000	0	0	100	0
7500	0	0	100	0	7500	0	0	100	0
10000	0	0	100	0	10000	0	0	100	0
20000	0	0	100	0	20000	0	0	100	0
30000	0	0	100	0	30000	0	0	100	0
40000	0	0	100	0	40000	0	0	100	0
50000	0	0	100	0	50000	0	0	100	0
60000	0	0	100	0	60000	0	0	100	0
70000	0	0	100	0	70000	0	0	100	0
80000	0	0	100	0	80000	0	0	100	0
90000	0	0	100	0	90000	0	0	100	0
100000	0	0	100	0	100000	0	0	100	0

Tabela 3.2 *Resultados das simulações para* $K = log(n)$, *e* $K = (Clog(n))^{1/2}$.

Na tabela acima, apresentamos os resultados referentes à estimativa da função contexto quando $K = log(n)$, e $K = (Clog(n))^{1/2}$ (todos os resultados relativos à árvore de ordem 2). Para ambos os Ks, houve acerto para as 100 amostras a partir de 250 casos.

$((Clog(n))^{1/2} + (Clog(n))^{1/3})/2$				
n	**0**	**1**	**C**	**E**
50	91	2	7	0
100	37	1	62	0
250	0	0	100	0
500	0	0	100	0
750	0	0	100	0
1000	0	0	100	0
2500	0	0	100	0
5000	0	0	100	0
7500	0	0	100	0
10000	0	0	99	1
20000	0	0	100	0
30000	0	0	99	1
40000	0	0	99	1
50000	0	0	99	1
60000	0	0	98	2
70000	0	0	98	2
80000	0	0	97	3
90000	0	0	99	1
100000	0	0	98	2

$(Clog(n))^{1/3}$				
n	**0**	**1**	**C**	**E**
50	69	4	27	0
100	16	3	81	0
250	0	0	100	0
500	0	0	94	6
750	0	0	96	4
1000	0	0	91	9
2500	0	0	85	15
5000	0	0	83	17
7500	0	0	75	25
10000	0	0	70	30
20000	0	0	61	39
30000	0	0	51	49
40000	0	0	36	64
50000	0	0	29	71
60000	0	0	26	74
70000	0	0	26	74
80000	0	0	23	77
90000	0	0	18	82
100000	0	0	21	79

Tabela 3.3 *Resultados das simulações para* $K = ((Clog(n))^{1/2} + (Clog(n))^{1/3})/2$, *e* $K = (Clog(n))^{1/3}$.

Na tabela acima, apresentamos os resultados referentes à estimativa da função contexto quando $K = ((Clog(n))^{1/2} + (Clog(n))^{1/3})/2$, e $K = (Clog(n))^{1/3}$ (todos os resultados relativos à árvore de ordem 2).

Para $K = ((Clog(n))^{1/2} + (Clog(n))^{1/3})/2$ houve acerto para as 100 amostras a partir de 250 casos.

Para $K = (Clog(n))^{1/3}$, a maior incidência de acerto ocorre para n entre 250 e 1.000, a partir daí, essa incidência decai continuadamente, chegando a um percentual em torno de 20% para amostras com 90.000 e 100.000 casos.

3.3.2 Resultados para CMAV2

$Clog(n)$						AIC					
n	**0**	**1**	**2**	**C**	**E**	**n**	**0**	**1**	**2**	**C**	**E**
50	100	0	0	0	0	50	0	7	4	3	86
100	100	0	0	0	0	100	0	6	12	20	62
250	100	0	0	0	0	250	0	0	3	8	89
500	100	0	0	0	0	500	0	0	1	2	97
750	21	79	0	0	0	750	0	0	3	13	84
1000	0	100	0	0	0	1000	0	0	1	4	95
2500	0	100	0	0	0	2500	0	0	0	8	92
5000	0	87	13	0	0	5000	0	0	0	0	100
7500	0	0	100	0	0	7500	0	0	0	4	96
10000	0	0	100	0	0	10000	0	0	0	0	100
20000	0	0	55	45	0	20000	0	0	0	2	98
22500	0	0	12	88	0	22500					
23500	0	0	5	95	0	23500					
24500	0	0	2	98	0	24500					
25000	0	0	1	99	0	25000					
25500	0	0	0	100	0	25500					
26000	0	0	0	100	0	26000					
30000	0	0	0	100	0	30000					
40000	0	0	0	100	0	40000					
50000	0	0	0	100	0	50000					
60000	0	0	0	100	0	60000					

Tabela 3.4 *Resultados das simulações para $K = Clog(n)$, e K que minimiza o AIC.*

	$Clog(n)$				
n	**0**	**1**	**2**	**C**	**E**
70000	0	0	0	100	0
80000	0	0	0	100	0
90000	0	0	0	100	0
100000	0	0	0	100	0

Continuação da Tabela 3.4

Nessa tabela, apresentamos os resultados referentes à estimativa da função contexto quando $K = Clog(n)$, e K que minimiza o AIC (todos os resultados relativos à árvore de ordem 3). Nela, encontramos um melhor desempenho, quando n é pequeno, para K que minimiza o AIC, e quando n é grande, para $K = Clog(n)$.

Para esse *valor de corte* houve acerto para as 100 amostras a partir de 25.500 casos, sendo que a partir de 23.500 já encontramos, ao menos, 95% de eventos de acerto. A convergência, ao contrário do encontrado para a árvore de ordem 2, apresenta-se de forma menos abrupta. Notar que para $n \leq 500$, encontramos, apenas, casos i.i.d; para $n = 750$, 21 casos i.i.d., e 79 árvores de ordem 1; para n entre 1.000 e 2.500, apenas árvores de ordem 1; para $n = 5.000$, encontramos 87 árvores de ordem 1, e 13 de ordem 2; para n entre 7.500 e 10.000, apenas árvores de ordem 2; para n entre 20.000 e 25.000, árvores de ordem 2 e árvores de ordem 3 que convergiram corretamente; e, a partir daí, todas as árvores convergiram corretamente.

Para K que minimiza o AIC, encontramos alguns casos de acerto, mesmo para valores altos como 20.000, porém em baixa frequência. Apenas para $n = 100$ e $n = 750$, encontramos uma incidência de acerto superior a 10%.

$log(n)$						$(Clog(n))^{1/2}$					
n	**0**	**1**	**2**	**C**	**E**	**n**	**0**	**1**	**2**	**C**	**E**
50	32	68	0	0	0	50	53	43	0	0	0
100	5	95	0	0	0	100	9	91	0	0	0
250	0	92	8	0	0	250	0	95	5	0	0
500	0	37	62	1	0	500	0	37	62	1	0
750	0	3	86	11	0	750	0	3	86	11	0
1000	0	0	93	7	0	1000	0	0	88	12	0
2500	0	0	10	90	0	2500	0	0	5	95	0
5000	0	0	0	100	0	5000	0	0	0	100	0
7500	0	0	0	100	0	7500	0	0	0	100	0
10000	0	0	0	100	0	10000	0	0	0	100	0
20000	0	0	0	100	0	20000	0	0	0	100	0
30000	0	0	0	100	0	30000	0	0	0	100	0
40000	0	0	0	100	0	40000	0	0	0	100	0
50000	0	0	0	100	0	50000	0	0	0	100	0
60000	0	0	0	100	0	60000	0	0	0	100	0
70000	0	0	0	100	0	70000	0	0	0	100	0
80000	0	0	0	100	0	80000	0	0	0	100	0
90000	0	0	0	100	0	90000	0	0	0	100	0
100000	0	0	0	100	0	100000	0	0	0	100	0

Tabela 3.5 *Resultados das simulações para* $K = log(n)$, *e* $K = (Clog(n))^{1/2}$.

Na tabela acima, apresentamos os resultados referentes à estimativa da função contexto quando $K = log(n)$, e $K = (Clog(n))^{1/2}$ (todos os resultados relativos à árvore de ordem 3). Para ambos os Ks, houve acerto para as 100 amostras a partir de 5.000 casos.

Notar que quando $K = (Clog(n))^{1/2}$, a convergência apresenta-se levemente mais rápida (quando $K = 2.500$, há convergência em 95% dos casos).

$((Clog(n))^{1/2} + (Clog(n))^{1/3})/2$					
n	**0**	**1**	**2**	**C**	**E**
50	30	70	0	0	0
100	3	97	0	0	0
250	0	76	21	3	0
500	0	4	90	6	0
750	0	0	69	31	0
1000	0	0	67	33	0
2500	0	0	0	100	0
5000	0	0	0	100	0
7500	0	0	0	100	0
10000	0	0	0	99	1
20000	0	0	0	100	0
30000	0	0	0	99	1
40000	0	0	0	99	1
50000	0	0	0	100	0
60000	0	0	0	97	3
70000	0	0	0	95	5
80000	0	0	0	100	0
90000	0	0	0	98	2
100000	0	0	0	97	3

$(Clog(n))^{1/3}$					
n	**0**	**1**	**2**	**C**	**E**
50	10	86	1	2	1
100	0	87	9	4	0
250	0	31	54	13	2
500	0	0	60	36	4
750	0	0	35	68	7
1000	0	0	21	70	9
2500	0	0	0	85	15
5000	0	0	0	82	18
7500	0	0	0	76	24
10000	0	0	0	66	34
20000	0	0	0	48	52
30000	0	0	0	40	60
40000	0	0	0	29	71
50000	0	0	0	29	71
60000	0	0	0	21	79
70000	0	0	0	23	77
80000	0	0	0	9	91
90000	0	0	0	13	87
100000	0	0	0	6	94

Tabela 3.6 *Resultados das simulações para* $K = ((Clog(n))^{1/2} + (Clog(n))^{1/3})/2$, *e* $K = (Clog(n))^{1/3}$.

Nessa última tabela, apresentamos os resultados referentes à estimativa da função contexto quando $K = ((Clog(n))^{1/2} + (Clog(n))^{1/3})/2$, e $K = (Clog(n))^{1/3}$ (todos os resultados relativos à árvore de ordem 3).

Para $K = ((Clog(n))^{1/2} + (Clog(n))^{1/3})/2$ houve acerto para as 100 amostras a partir de 2.500 casos. A convergência, ao contrário do encontrado para a árvore de ordem 2, apresenta uma leve flutuação na quantidade de casos convergentes para $n \geq 10.000$.

Para $K = (Clog(n))^{1/3}$, a maior incidência de acerto ocorre para n entre próximo de 2.500, a partir daí essa incidência decai continuadamente, chegando a um percentual em torno de 10% para amostras com 80.000 e 100.000 casos.

Capítulo 4

Aplicação

4.1 Introdução

Neste capítulo apresentamos os resultados da aplicação de técnicas das CMAV em problema de estimativa de risco de inadimplência para produtos de crédito, especificamente na estimativa de um modelo de *perfil de comportamento*.

Um modelo de *perfil de comportamento* consiste numa metodologia de previsão do comportamento creditício de um cliente, tomador de um crédito qualquer. Esse modelo é usualmente indicado para produtos de crédito com prazo médio elevado e para produtos rotativos, pois o caráter prolongado desses produtos permite a coleta de informações suficientes para a análise em questão.

A informação proveniente do modelo de *perfil de comportamento* possibilita a tomada de decisões estratégicas relacionadas ao controle de qualidade da carteira de crédito de uma instituição financeira. Pode-se, por exemplo, cancelar o cheque especial de um cliente que apresenta alta probabilidade de permanência em excesso sobre o limite, ou ainda, reduzir o limite de crédito de um cliente que apresenta um risco moderado, ou, em contrapartida, aumentar o limite de um cliente que apresenta baixo risco de crédito.

No caso, o algoritmo foi aplicado a uma situação muito comum no mercado financeiro, a concessão de crédito parcelado. Este é muito utilizado, por exemplo, em vendas de bens a prazo.

Na operação de crédito parcelado, concede-se uma certa quantia de dinheiro a ser paga em prestações periódicas. Por exemplo, abre-se um crédito de R$ 100,00 a ser pago em 10 prestações de R$ 12,00. O problema do credor é a inadimplência e ele deve fazer o

possível para poder estimá-la com a maior precisão possível. A partir dessa estimativa ele calculará a taxa de juros do empréstimo, o número de prestações, etc.

Para tornar a discussão mais precisa, vamos supor que o crédito a ser concedido deve ser pago em 12 parcelas. Vamos representar um tal crédito por meio de uma palavra com 12 caracteres escolhidos entre 0 ou 1. Vamos arbitrar que a presença do 0 significa parcela quitada e a de 1, não paga. Por exemplo, 001000111000 representa um crédito cujas parcelas de número 3, 7, 8, e 9 foram pagas com atraso. Vamos também supor que uma parcela só pode ser paga quando todas as outras anteriores o forem.

O que se deseja obter é o perfil do comportamento dos tomadores de crédito. Por exemplo, dado que uma pessoa pagou a primeira parcela, qual a probabilidade dela inadimplir a segunda? Ou ainda, se uma pessoa pagou as seis primeiras, que probabilidade ela tem de pagar a sétima, etc. Em termos formais, representado um crédito por $X_1, \ldots, X_{12}$, quais são os valores de $\mathbf{P}[X_2 = 1 \mid X_1 = 0]$ e $\mathbf{P}[X_7 = 0 \mid X_1 = 0, \ldots, X_6 = 0]$?

Note que o problema refere-se a um espaço amostral finito, mas que a quantidade de cálculos envolvidos é muito grande, ficando inviável se desejarmos estudar situações onde o número de parcelas é grande, digamos 36.

O Algoritmo do Contexto pode nos auxiliar neste caso. A idéia foi a de colocar os vários casos de crédito observados sequencialmente. Mais precisamente, colocou-se o primeiro crédito ocupando as primeiras 12 posições, o segundo, ocupando as posições de índice 13 a 24 e assim por diante. Dessa forma, obtemos uma amostra sobre a qual pode-se aplicar o Algoritmo do Contexto.

Embora a convergência para o Algoritmo do Contexto tenha sido estabelecida sob hipóteses fortes, entre as quais está a de estacionariedade, não se sabe se essas condições são necessárias. O que se tem é um estimador, com boas propriedades assintóticas sob certas condições, mas isso não impede que ele não possa ser utilizado sob outras restrições. É algo semelhante a se utilizar um teste t sem que a distribuição populacional seja rigorosamente normal, procedimento muito comum na prática do dia-a-dia.

No presente trabalho, analisamos um conjunto de dados referentes ao histórico de pagamento de um crédito parcelado de longo prazo. O modelo foi desenvolvido sobre um espaço categorizado $\mathcal{X} = \{0, 1\}$, onde 0 representa os eventos de atraso até 30 dias (os

casos de pagamento em dia, ou antecipados constam nessa classe), e 1, os eventos de atraso superiores a 30 dias.

4.2 Metodologia Utilizada

O banco de dados escolhido apresentava, inicialmente, 132.037 casos, definidos no espaço categorizado $\mathcal{X} = \{0, 1\}$. Esse banco de dados foi montado de forma que o histórico de pagamento de um determinado cliente foi unido ao histórico de outro cliente. Formalizando temos, para dois clientes com histórico de pagamento $\vec{w}$ e $\vec{u}$, montamos um único vetor de eventos $wu = (w_{|w|}, \cdots, w_2, w_1, u_{|u|}, \cdots, u_2, u_1)$, que representa a concatenação do histórico dos dois clientes.

Ao todo, haviam 8.839 clientes com 12 a 36 parcelas cada.

Os procedimentos adotados no desenvolvimento do modelo e em sua análise de desempenho são apresentados abaixo.

- Primeiramente, foram selecionados (aleatoriamente) 98.836 casos, 75% do banco de dados inicial, para compor a amostra de desenvolvimento. Dentre esses casos, encontramos 6.619 clientes diferentes com uma quantidade de parcelas que varia de 12 a 36.

- O *valor de corte* utilizado foi o $(Clog(n))^{1/2}$, com $C = 2|\mathcal{X}| + 5$. Esse valor foi escolhido por apresentar uma eficiência levemente superior a $log(n)$ no caso da simulação para a árvore de ordem 3. Essa árvore foi escolhida como referência, pois, pela experiência que acumulamos em trabalhos semelhantes, modelos de *perfil de comportamento* apresentam ordem de dependência superior a 2.

- Após a estimativa do modelo, seu desempenho foi analisado. Para isso, para cada ramo da árvore contexto estimada, foi observada a frequência com que os eventos 0 e 1 apareciam. Esse acompanhamento foi realizado para cada cliente da amostra. O procedimento foi repetido ao longo do vetor de informações, e a média dos resultados comparada com a probabilidade da árvore estimada.

4.3 Resultados

A árvore resultante apresentou ordem 7. Sua árvore contexto, e suas probabilidades são as seguintes:

Árvore Contexto

$$\hat{c}(x_{-\infty}^{n}) = \begin{cases} 1,0 & \text{se } x_n = 1,\ x_{n-1} = 0 \text{ e } x_{-\infty}^{n-2} \text{ qualquer,} \\ 1,1,0 & \text{se } x_n = 1,\ x_{n-1} = 1,\ x_{n-2} = 0 \text{ e } x_{-\infty}^{n-3} \text{ qualquer,} \\ 1,1,1 & \text{se } x_n = 1,\ x_{n-1} = 1,\ x_{n-2} = 1 \text{ e } x_{-\infty}^{n-3} \text{ qualquer,} \\ 0,1 & \text{se } x_n = 0,\ x_{n-1} = 1 \text{ e } x_{-\infty}^{n-2} \text{ qualquer,} \\ 0,0,1 & \text{se } x_n = 0,\ x_{n-1} = 0,\ x_{n-2} = 1 \text{ e } x_{-\infty}^{n-3} \text{ qualquer,} \\ 0,0,0,1 & \text{se } x_n = 0,\ x_{n-1} = 0,\ x_{n-2} = 0,\ x_{n-3} = 1 \text{ e } x_{-\infty}^{n-4} \text{ qualquer,} \\ 0,0,0,0,1 & \text{se } x_n = 0,\ x_{n-1} = 0,\ x_{n-2} = 0,\ x_{n-3} = 0,\ x_{n-4} = 1 \\ & \text{e } x_{-\infty}^{n-5} \text{ qualquer,} \\ 0,0,0,0,0,1 & \text{se } x_n = 0,\ x_{n-1} = 0,\ x_{n-2} = 0,\ x_{n-3} = 0,\ x_{n-4} = 0, \\ & x_{n-5} = 1 \text{ e } x_{-\infty}^{n-6} \text{ qualquer,} \\ 0,0,0,0,0,0,1 & \text{se } x_n = 0,\ x_{n-1} = 0,\ x_{n-2} = 0,\ x_{n-3} = 0,\ x_{n-4} = 0, \\ & x_{n-5} = 0,\ x_{n-6} = 1 \text{ e } x_{-\infty}^{n-7} \text{ qualquer,} \\ 0,0,0,0,0,0,0 & \text{se } x_n = 0,\ x_{n-1} = 0,\ x_{n-2} = 0,\ x_{n-3} = 0,\ x_{n-4} = 0, \\ & x_{n-5} = 0,\ x_{n-6} = 0 \text{ e } x_{-\infty}^{n-7} \text{ qualquer.} \end{cases}$$

Probabilidades

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 1,\ X_{n-2} = 0] = 0,493$

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 1,\ X_{n-2} = 1,\ X_{n-3} = 0] = 0,621$

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 1,\ X_{n-2} = 1,\ X_{n-3} = 1] = 0,724$

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 1] = 0,221$

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 0,\ X_{n-3} = 1] = 0,129$

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 0,\ X_{n-3} = 0,\ X_{n-4} = 1] = 0,090$

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 0,\ X_{n-3} = 0,\ X_{n-4} = 0,\ X_{n-5} = 1] = 0,076$

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 0,\ X_{n-3} = 0,\ X_{n-4} = 0,\ X_{n-5} = 0,\ X_{n-6} = 1] = 0,055$

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 0,\ X_{n-3} = 0,\ X_{n-4} = 0,\ X_{n-5} = 0,\ X_{n-6} = 0,\ X_{n-7} = 1] = 0,060$

$\hat{\mathbb{P}}[X_n = 1 \mid X_{n-1} = 0,\ X_{n-2} = 0,\ X_{n-3} = 0,\ X_{n-4} = 0,\ X_{n-5} = 0,\ X_{n-6} = 0,\ X_{n-7} = 0] = 0,027$

Notemos que a árvore estimada apresenta uma interpretação condizente com o esperado. Ela mostra que, conforme o evento de atraso superior a 30 dias se distancia no tempo, a probabilidade de reincidência dessa ocorrência reduz significativamente. Por exemplo, um cliente que apresentou um atraso superior a 30 dias há sete meses atrás, e nenhum outro evento desses até o momento, apresenta uma probabilidade de atraso, relacionada à próxima parcela, de 5,5%, ao contrário do cliente que incorreu num evento de atraso há dois meses, que apresenta uma probabilidade de reincidência de 46,8%.

Como mencionado no item 4.2, a análise de desempenho do modelo foi desenvolvida através do estudo da frequência com que cada evento ocorria para cada cliente ($\hat{\mathbb{P}}_c$). A relação das frequências encontradas está relacionada abaixo.

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 1,\, X_{n-2} = 0] = 0,469$

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 1,\, X_{n-2} = 1,\, X_{n-3} = 0] = 0,611$

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 1,\, X_{n-2} = 1,\, X_{n-3} = 1] = 0,714$

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 0,\, X_{n-2} = 1] = 0,230$

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 0,\, X_{n-2} = 0,\, X_{n-3} = 1] = 0,160$

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 0,\, X_{n-2} = 0,\, X_{n-3} = 0,\, X_{n-4} = 1] = 0,136$

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 0,\, X_{n-2} = 0,\, X_{n-3} = 0,\, X_{n-4} = 0,\, X_{n-5} = 1] = 0,097$

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 0,\, X_{n-2} = 0,\, X_{n-3} = 0,\, X_{n-4} = 0,\, X_{n-5} = 0,\, X_{n-6} = 1] = 0,076$

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 0,\, X_{n-2} = 0,\, X_{n-3} = 0,\, X_{n-4} = 0,\, X_{n-5} = 0,\, X_{n-6} = 0,\, X_{n-7} = 1] = 0,055$

$\hat{\mathbb{P}}_c[X_n = 1 \,|\, X_{n-1} = 0,\, X_{n-2} = 0,\, X_{n-3} = 0,\, X_{n-4} = 0,\, X_{n-5} = 0,\, X_{n-6} = 0,\, X_{n-7} = 0] = 0,020$

Ao compararmos as duas estimativas, percebemos que estão muito próximas, indicando que as CMAV caracterizam um bom estimador para o cálculo da probabilidade de atraso de um determinado cliente.

Capítulo 5

Conclusão

Neste trabalho foram estudadas as Cadeias de Markov de Alcance Variável (CMAV), e sua aplicação através do Algoritmo do Contexto.

Inicialmente, foi analisado o desempenho do estimador para $K = Clog(n)$ (com $C = 2|\mathcal{X}| + 5$), e K derivado do AIC, para funções contexto de ordem 2 e 3. Nos dois casos, encontramos um melhor desempenho, quando n é pequeno, para K que minimiza o AIC, e quando n é grande, para $K = Clog(n)$.

Em relação aos demais *valores de corte* estudados ($log(n)$, $(Clog(n))^{1/2}$, $((C\log(n))^{1/3} + (C\log(n))^{1/2})/2$, e $(c\log(n)^{1/3})$, segunda etapa do trabalho, o melhor desempenho encontrado foi para os dois primeiros casos: $log(n)$, e $(Clog(n))^{1/2}$. Isso ocorreu tanto para a árvore de ordem 2 quanto para a de ordem 3. Os *valores de corte* com valores mais próximos do encontrado pelo AIC mostraram-se mais volúveis.

O estimador estudado apresentou um desempenho satisfatório para o problema de estimativa de risco de inadimplência para produto de crédito de longo prazo. A aplicação revelou uma estrutura probabilística que faz todo o sentido para um profissional da área de concessão de crédito. Este exemplo sugere que se deve procurar estender a validade da convergência do estimador resultante da aplicação do Algoritmo do Contexto para situações semelhantes às da aplicação acima.

Apêndice 1: Propriedades das Estimativas Geradas pelo Algoritmo do Contexto

Neste capítulo faremos apenas uma apresentação das demonstrações da propriedades relativas ao Algoritmo do Contexto. Não há pretensão de esmiuçamento das mesmas, apenas de sua apresentação em uma ordem mais linear do que aparecem nos artigos originais. Falaremos agora de seqüências de CMAVs.

Consideremos uma seqüência $(P_n)_{n\in\mathbb{N}}$ de elementos de $\mathcal{P}$, a reunião de todas a CMAV de ordem finita. Vamos supor que cada P_n tem função contexto c_n, de ordem k_n, que induz a árvore contexto $\tau_n = \tau_{c_n}$ e uma árvore de nós terminais $\tau_n^T = \tau_{c_n}^T$. Vamos empregar a notação $(X_{t,n})_{t\in\mathbb{Z}}$ para seu processo estocástico. Assumamos que as condições a seguir são verdadeiras:

A1: Escrevendo $P_n^{(r)}(A,w)$ para a probabilidade de transição em r passos iniciando em $w \in \mathcal{X}^{k_n}$ e terminando em $A \subseteqq \mathcal{X}^{k_n}$, isto é, $P_n^{(r)}(A,w) = \mathbb{P}_{P_n}[X_{r-k_n+1,n}^{r,n} \in A | X_{-k_n+1,n}^{0,n} = w]$, assuma que, para algum r e para algum $0 < \kappa$, valha

$$|P_n^{(r)}(A,w) - P_n^{(r)}(A,w')| < 1 - 2\kappa \tag{5.1}$$

para todos $A \subseteqq \mathcal{X}^{k_n}$ e todos w, $w' \in \mathcal{X}^{k_n}$.

A2: Escrevendo a norma L^1 de uma função $f : \mathcal{X} \longrightarrow \mathbb{R}$ como $||f||_1$, isto é, $||f||_1 = \sum_{x\in\mathcal{X}}|f(x)|$, seja ε_n o menor valor de $||P_n(\cdot\,|\,wu) - P_n(\cdot\,|\,w)||_1$ para todos os $w \in \bigcup_{j\in\mathbb{N}} \mathcal{X}^j$ e todos $u \in \mathcal{X}$ tais que $wu \in \tau_n^T$. Seja, também, b_n o menor valor de $P_n(w)$ para todos $w \in \tau_n^T$. Vamos supor que

$$\text{para algum } \beta,\ 0 < \beta < \infty, \text{ valha } b_n^{-1} = \mathbf{O}(\log(n)^{-(1/2+\beta)} n^{1/2}) \text{ e} \tag{5.2}$$

$$\text{para algum } \delta,\ 0 < \delta < \infty, \text{ valha } \varepsilon_n^{-1} = \mathbf{O}((\log(n)^{-(1+\delta)} n b_n)^{1/2}) \tag{5.3}$$

A3: A probabilidade de transição satisfaz

$$\frac{1}{\min_{x\in\mathcal{X},\, w\in\tau_n} p_n(x \mid w)} = \mathbf{O}(n) \tag{5.4}$$

Estas são as hipóteses que utilizaremos para provar os resultados deste capítulo que lidam com seqüências de CMAV. Para o caso especial e importante de uma única CMAV P, isto é, de uma seqüência constante $P_n = P$, basta supor que $p(x \mid w)$ nunca é nula para todo $x \in \mathcal{X}$ e todo $w \in \tau$, onde τ é a árvore contexto de P, para que as hipóteses **A1**, **A2** e **A3** sejam atendidas. De fato, todos os b_n são iguais a um único número, bem como todos os ε_n, o que faz com que **A2** seja satisfeita. O mesmo se dá com **A3**. Quanto a **A1**, o fato de não existir nenhuma probabilidade de transição nula faz com que nenhuma das parcelas da diferença $|P_n^{(r)}(A, w) - P_n^{(r)}(A, w')|$ possa ser igual a 1, o que acarreta um limite superior menor que 1.

Utilizaremos no teorema abaixo as seguintes observações:

OBS: 5.0.1. *A hipótese **A2** sobre a probabilidade estacionária mínima limita o tamanho da árvore terminal de contextos da forma* $|\tau_n^T| \leqq b_n^{-1} = \mathbf{O}(n^{1/2}/\log(n)^{1/2+\beta})$

OBS: 5.0.2. *A hipótese **A1** sobre os núcleos de transição está relacionada com o coeficiente de ergodicidade para processos de Markov estacionários (ver Iosifescu and Theodorescu [4] , Rajarashi [7] e Doukhan [2]). Ela acarreta que os processos de estado* $(Z_{t,n})_{t\in\mathbb{Z}}$, *onde* $Z_{t,n} = c_n(X_{-\infty,n}^{t,n})$ *e a CMAV* $(X_{t,n})_{t\in\mathbb{Z}}$ *são geometricamente* ϕ*-mixing com coeficientes de mixing limitados por*

$$\sup_{n\in\mathbb{N}} \phi_n(j) \leqq (1 - 2\kappa)^j \textit{ para todo } j \in \mathbb{N}$$

Vale a pena lembrar que a noção de ϕ-mixing, bem como as diversos outros conceitos de mixing tais como α-mixing, β-mixing etc... são introduzidos como formas de se medir a dependência de σ-álgebras. Formalmente, se temos um espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$ e $\mathcal{U}$ e $\mathcal{V}$ duas sub-σ-álgebras de $\mathcal{A}$, o *coeficiente de ϕ-mixing* entre $\mathcal{U}$ e $\mathcal{V}$ é dado por

$$\phi(\mathcal{U}, \mathcal{V}) = \sup\{|\mathbb{P}(V) - \mathbb{P}(U \mid V)| \,:\, U \in \mathcal{U},\ \mathbb{P}(U) \neq 0,\ V \in \mathcal{V}\}.$$

Naturalmente, o coeficiente de mixing entre duas variáveis aleatórias é o coeficiente de mixing entre as sub-σ-álgebras que elas geram. Maiores detalhes podem ser encontrados em [2], no primeiro capítulo.

A seguir enunciaremos o teorema de consistência do estimador obtido pelo Algoritmo do Contexto.

Teorema 5.0.1. Seja $X_{1,n}, \ldots, X_{n,n}$ com P_n satisfazendo as condições **A1**, **A2** e **A3**. Seja $\hat{P}(\cdot \mid \cdot)$ como na equação (2.2) e $\hat{c}(\cdot)$ a estimativa da função contexto dada pelo Algoritmo do Contexto. Então:

1. $\lim_{n\to\infty} \mathbb{P}[\hat{c}(\cdot) = c_n(\cdot)] = 1$.

2. $\sup_{x_{-\infty}^1 \in \mathcal{X}^\infty} |\hat{P}(\, x_1 | \hat{c}(x_{-\infty}^0)\,) - p_n(\, x_1 | c(x_{-\infty}^0)\,)| = \mathbf{o}_P(1)$ quando n tende ao infinito.

A demonstração desse teorema é extensa e técnica. Vamos realizá-la em etapas, através de um série de proposições.

Vamos iniciar enunciando um teorema que se encontra em Doukhan [2]. Ele será aplicado na proposição que o segue.

Teorema 5.0.2. (Teorema 4 de Doukhan [2], capítulo 1.4.2) Assuma que a seqüência seja β-mixing * e

1. $\mathbb{E}(X_t) = 0$.

2. $\exists\, \sigma^2 > 0,\ \forall m, \forall n,\ \frac{1}{m}\mathbb{E}(X_n + \cdots + X_{n+m})^2 \leqq \sigma^2$.

3. $|X_t| \leqq M$.

Então para todo $0 < \varepsilon$ e para todo número real não negativo $q \leqq \frac{n}{1+\theta}$, onde $\theta = \frac{\varepsilon^2}{4}$, temos:

$$\mathbb{P}[|\sum_{t=1}^{n} X_t| \geqq x] \leqq 4\exp(-\frac{(1-\varepsilon)x^2}{2(n\sigma^2 + qMx/3)}) + 2^n \frac{\beta_{[q\theta]-1}}{q}.$$

* A condição de limitação sobre ϕ-mixing implica a mesma condição para β-mixing porque $\beta \leqq \phi$ (ver a proposição 1 da página 4 de Doukhan [2]).

Com o concurso do teorema acima podemos provar a seguinte

Proposição 5.0.1. *Assumindo as hipóteses **A1** e **A2**, existe uma constante* $0 < G < \infty$ *tal que*

$$\max_{w \in \tau_n^T} \mathbb{P}[\,(N(w) - \mathbb{E}[N(w)]) > b_n n/3] = \mathbf{O}(n^{-\log(n)^{2\beta} G})$$

Dem: Já que $w \in \tau_n$, podemos escrever

$$N(w) = \sum_{t=1}^{n} 1_{[Z_{t,n}=w]}, \quad Z_{t,n} = c(X_{-\infty,n}^{t,n})$$

Pela hipótese **A1** (ver observação 5.0.2), $(Z_{t,n})_{t\in\mathbb{Z}}$ é ϕ-mixing com coeficientes de mixing limitados por $\sup_{n\in\mathbb{N}} \phi_n(j) \leqq (1-2\kappa)^j$. Aplicando agora o Teorema 5.0.2 com $q = \log(n)^{1+2\beta}$ e notando que $b_n n \geqq \text{const.} n^{1/2} \log(n)^{1/2+\beta}$ justificamos a afirmação. ■

Prosseguindo, vamos agora nos ocupar do resultado seguinte.

Proposição 5.0.2. *Para* $(Y_i)_{i\in\mathbb{N}}$ *com* $\mathbb{E}[Y_1] = p$, *onde colocamos, para* $i \in \mathbb{N}$, $I_i = \{$*tempo da i-ésima ocorrência de v em* $(X_t)_{t\in\mathbb{N}}\}$, $W_i = X_{I_i+1}$ *o símbolo que ocorre após a i-ésima ocorrência de v e* $Y_i = 1_{[W_i=x]}$. *Para* $a_n(k) = (\frac{\varepsilon_n}{2} - \gamma_n(k))^2$, *onde* $\gamma_n(k)^2 = C\log(n)/k$, *dentro das hipóteses **A1** e **A2**, se* $k \geqq \rho_n = b_n n/2$, *então, para alguma constante* $0 < F < \infty$, *temos*

$$\sup_{0<p<1} \mathbb{P}[\,|\sum_{i=1}^{k} \frac{Y_i}{k} - p|^2 > a_n(k)] = \mathbf{O}(n^{-\log(n)^{\delta} F})$$

Dem: A sequência W_i é estacionária ϕ-mixing com coeficientes de mixing limitados pelo mesmo limitante que serve para a sequência $(X_t)_{t\in\mathbb{N}}$. Pela hipótese **A1**, o processo $(X_t)_{t\in\mathbb{N}}$ tem coeficientes de mixing $\phi(j) \leqq (1-2\kappa)^j$ e os mesmos limitantes aplicam-se para os coeficientes de mixing do processo $(Y_i)_{i\in\mathbb{N}}$.

Utilizando o teorema 5.0.2 agora com $q = \log(n)^{1+\delta}$. Note que $k \geqq \rho_n = b_n n/2 \geqq \text{const.} n^{1/2} \log(n)^{1/2+\beta}$. Usando que, para $k \geqq \rho_n$, vale $k\, a_n(k) \geqq \text{const.} \log(n)^{1+\delta}$, justificamos a afirmação. ■

Consideremos agora os eventos

$$U_n = \{\text{existe } w \in \hat{\tau} \text{ tal que, para algum } u \in \bigcup_{n=1}^{\infty} \mathcal{X}^n \text{ tem-se } wu \in \tau_n \text{ e } wu \notin \hat{\tau}\}$$

$$O_n = \{\text{existe } w \in \hat{\tau} \text{ tal que, para algum } u \in \bigcup_{n=1}^{\infty} \mathcal{X}^n \text{ tem-se } wu \notin \tau_n \text{ e } wu \in \hat{\tau}\}$$

O evento U_n corresponde à subestimação e O_n à superestimação da árvore contexto. Observe que o *evento erro* $E_n = \{\tau_n \neq \hat{\tau}\}$ é a união de U_n com O_n.

Proposição 5.0.3. *Sob as hipóteses **A1** e **A2**, existem constantes* $0 < D_1, D_2 < \infty$ *tais que* $\mathbb{P}[U_n] = \mathbf{O}(\max\{n^{-\log(n)^{2\beta D_1}}, n^{-\log(n)^{\delta D_2}}\})$ *quando* n *tende ao infinito.*

Dem:

Particionemos U_n pelo evento $D_n = \{\rho_n \leqq N(w) \text{ para todo } w \in \tau_n^T\}$, onde ρ_n é uma constante a ser determinada. Como $\mathbb{P}[U_n] \leqq \mathbb{P}[U_n \cap D_n] + \mathbb{P}[D_n^c]$, limitaremos $\mathbb{P}[U_n]$ se limitarmos $\mathbb{P}[U_n \cap D_n]$ e $\mathbb{P}[D_n^c]$. Limitando-nos, sem perda de generalidade, a τ_n^T, podemos escrever

$$\begin{aligned}\mathbb{P}[U_n \cap D_n] &\leqq \sum_{wu \in \tau_n^T, u \in \mathcal{X}} \mathbb{P}[\Delta_{wu} < C\log(n),\ \rho_n \leqq N(wu)] \\ &= \sum_{wu \in \tau_n^T, u \in \mathcal{X}} \sum_{k=\rho_n}^{n} \sum_{j=k}^{n} \mathbb{P}[D(\hat{P}_{wu}||\hat{P}_w) < \frac{C\log(n)}{k},\ N(wu) = k,\ N(w) = j]\end{aligned} \tag{5.5}$$

Como é sabido (Cover e Thomas, [1]: ver Lemma 12.6.1, páginas 300 a 301), podemos limitar inferiormente a divergência pela distância L_1, isto é, $\frac{1}{2\ln 2}||\hat{P}_{wu} - \hat{P}_w||_1^2 \leqq D(\hat{P}_{wu}||\hat{P}_w)$. Também $||\hat{P}_{wu} - \hat{P}_w||_1^2 = 2(\hat{P}_{wu}(A) - \hat{P}_w(A))^2$, onde $A = \{x \in \mathcal{X} \mid \hat{P}_{wu} > \hat{P}_w\}$. Consequentemente

$$\begin{aligned}&\mathbb{P}[D(\hat{P}_{wu}||\hat{P}_w) < \frac{C\log(n)}{k},\ N(wu) = k,\ N(w) = j] \\ &\leqq \mathbb{P}[(\hat{P}_{wu}(A) - \hat{P}_w(A))^2 < \frac{C\log(n)}{k},\ N(wu) = k,\ N(w) = j]\end{aligned} \tag{5.6}$$

Em razão da hipótese **A2**, ou $\hat{P}_{wu}(A)$ está afastado de $P_{wu}(A)$, ou $\hat{P}_w(A)$ está afastado de $P_w(A)$. Colocando $\gamma_n(k)^2 = C\log(n)/k$, $\hat{P}_{wu}(x) = a$, $\hat{P}_w(x) = b$, $p_{wu}(x) = r$ e $p_w(x) = s$, onde x é um elemento de $\mathcal{X}$, nosso objetivo passa a ser provar que se $|a - b|$ é pequeno,

então ou $|r-a|$ é grande, ou $|s-b|$ é grande. Assumindo, sem perda de generalidade, que $s<r$ temos, por **A2**, que $r-s>\varepsilon_n$. Agora, se $b<s$, então $|a-b|<\gamma_n(k)$ acarreta $|a-r|>\varepsilon_n-\gamma_n(k)$. Além disso, se $b>r$, então temos que ter $|s-b|>\varepsilon_n$. Assim, se $s\leqq b\leqq r$, então ou $s\leqq b<s+(r-s)/2$, caso em que se tem $|r-a|>\varepsilon/2-\gamma_n(k)$; ou $r-(r-s)/2\leqq b\leqq r$, caso em que se tem $|s-b|>\varepsilon/2$. Juntando tudo, provamos que se $|\hat{P}_{wu}(x)-\hat{P}_w(x)|<\gamma_n(k)$, então ou $|\hat{P}_{wu}(x)-p_{wu}(x)|>\varepsilon/2-\gamma_n(k)$, ou $|\hat{P}_w(x)-p_w(x)|>\varepsilon/2-\gamma_n(k)$. Dessa forma, aplicando os resultados à desigualdade (5.6) e fazendo

$$a_n(k)=(\frac{\varepsilon_n}{2}-\gamma_n(k))^2 \tag{5.7}$$

temos que ter

$$\begin{aligned}
\mathbb{P}[D(\hat{P}_{wu}||\hat{P}_w)<\frac{C\log(n)}{k},\ N(wu)=k,\ N(w)=j]\\
\leqq \mathbb{P}[\sum_{x\in A}|\hat{P}_{wu}(x)-p_{wu}(x)|>a_n(k)^{1/2},\ N(wu)=k]\\
+\mathbb{P}[\sum_{x\in A}|\hat{P}_w(x)-p_w(x)|>a_n(k)^{1/2},\ N(w)=j]\\
\leqq |\mathcal{X}|\max_{x\in A}\mathbb{P}[\,|\hat{P}_{wu}(x)-p_{wu}(x)|>a_n(k)^{1/2},\ N(wu)=k]\\
+|\mathcal{X}|\max_{x\in A}\mathbb{P}[\,|\hat{P}_w(x)-p_w(x)|>a_n(k)^{1/2},\ N(w)=j]
\end{aligned} \tag{5.8}$$

Escolhendo agora $\rho_n=b_nn/2\geqq \text{const}.n^{1/2}\log(n)^{1/2+\beta}$, obtemos a seguinte desigualdade: $\max_{k\geqq\rho_n}\gamma_n^2(k)\leqq \text{const}.C\log(n)/(nb_n)$. Segue-se assim, pela hipótese assumida sobre ε_n, que

$$\min_{k\geqq\rho_n}a_n(k)=\min_{k\geqq\rho_n}(\frac{\varepsilon_n}{2}-\gamma_n(k))^2\geqq \text{const}.\frac{\log(n)^{1+\delta}}{nb_n}.$$

Portanto,

$$\min_{k\geqq\rho_n}k\,a_n(k)\geqq \text{const}.\log(n)^{1+\delta}$$

Trataremos os dois casos do lado direito da desigualdade (5.8) ao mesmo tempo. Seja, então $v=wu$ ou $v=w$ conforme o caso. Coloquemos $p=P_v(x)$ e $\hat{p}=\hat{P}(x)$. Desejamos achar um limitante superior para a probabilidade do evento $\{|p-\hat{p}|^2>a_n(k),\ N(v)=k\}$. Para tanto, considere a extensão de X_1^n para a sequência infinita $(X_t)_{t\in\mathbb{N}}$. Coloquemos, para $i\in\mathbb{N}$, $I_i=\{\text{tempo da } i\text{-ésima ocorrência de } v \text{ em } (X_t)_{t\in\mathbb{N}}\}$. Seja $W_i=X_{I_i+1}$ o

símbolo que ocorre após a i-ésima ocorrência de v. A distribuição marginal de probabilidade para W_1 em $\mathcal{X}$ é P_v. Seja $Y_i = 1_{[W_i=x]}$. Observando que

$$\{|\sum_{i=1}^{N(v)} \frac{Y_i}{N(v)} - p|^2 > a_n(k),\ N(v) = k\} \subseteqq \{|\sum_{i=1}^{k} \frac{Y_i}{k} - p|^2 > a_n(k)\}$$

vemos que estabelecemos o limite superior

$$\mathbb{P}[\,|p - \hat{p}|^2 > a_n(k),\ N(v) = k] \leqq \mathbb{P}[\,|\sum_{i=1}^{k} \frac{Y_i}{k} - p|^2 > a_n(k)] \tag{5.9}$$

Continuando com a prova da proposição 5.0.3, façamos $M_n = \exp(-F\log(n)^{1+\delta})$. Aplicando a proposição 5.0.2 à desigualdade (5.9), provamos que, para $k,\ j \geqq \rho_n$, temos

$$\max_{x\in A} \mathbb{P}[\,|\hat{P}_{wu}(x) - p_{wu}(x)| > a_n(k)^{1/2},\ N(wu) = k] = \mathbf{O}(M_n)$$
$$\max_{x\in A} \mathbb{P}[\,|\hat{P}_{w}(x) - p_{w}(x)| > a_n(k)^{1/2},\ N(w) = j] = \mathbf{O}(M_n)$$

Assim, junto com (5.5), (5.6) e a última afirmação, podemos escrever

$$\mathbb{P}[\,U_n \cap D_n] = |\mathcal{X}| \sum_{wu\in\tau_n^T,\, u\in\mathcal{X}} \sum_{k=\rho_n}^{n} \sum_{j=k}^{n} \mathbf{O}(M_n) \tag{5.10}$$

Levando em conta a observação (5.0.1), a desigualdade acima e a hipótese **A2**, podemos escrever

$$\begin{aligned} \mathbb{P}[\,U_n \cap D_n] &= \mathbf{O}(|\tau_n^T| n^2 M_n) = \mathbf{O}(b_n^{-1} n^2 \exp\left(-\text{const.}\log(n)^{1+\delta}\right)) \\ &= \mathbf{O}(n^{-\log(n)^{\delta}\text{const.}}) \end{aligned} \tag{5.11}$$

Para completar a prova da proposição 5.0.3, necessitamos limitar $\mathbb{P}[\,D_n^c]$. Pela hipótese **A2** podemos escrever

$$\begin{aligned} \mathbb{P}[\,D_n^c] &\leqq \sum_{w\in\tau_n^T} \mathbb{P}[\,N(w) < \rho_n] = \sum_{w\in\tau_n^T} \mathbb{P}[\,(N(w) - \mathbb{E}[N(w)]) < \rho_n - (n - |w| + 1)P_n(w)] \\ &\leqq \sum_{w\in\tau_n^T} \mathbb{P}[\,(N(w) - \mathbb{E}[N(w)]) < -b_n n/3] \\ &\leqq \sum_{w\in\tau_n^T} \mathbb{P}[\,(N(w) - \mathbb{E}[N(w)]) > b_n n/3] \end{aligned}$$

A proposição 5.0.1 permite escrever

$$\mathbb{P}[\,D_n^c] = \mathbf{O}(b_n^{-1} n^{-\log(n)^{2\beta}\text{const.}}) = \mathbf{O}(n^{-\log(n)^{2\beta}\text{const.}})$$

onde utilizamos a hipótese **A2** para a última estimativa. Utilizando agora a desigualdade (5.11), completa-se a prova da proposição 5.0.3. ■

Vamos agora considerar o evento de sobrestimação $O_n = \{\text{existe } w' \in \tau_n \text{ com } w = w'u \in \hat{\tau}, w = w'u \notin \tau_n \text{ para algum } u \in \bigcup_{m=1}^{\infty} \mathcal{X}^m\}$. Para algum w ser um elemento de $\hat{\tau}^T \subseteqq \hat{\tau}$ é necessário que $N(w) > 1$ e que $\Delta_w \geqq C\log(n)$. Levando em conta que Weinberger, Rissanen e Feder (ver [9]) estabelecem, para qualquer $w = w'u$ ($w' \in \tau_n$, $u \in \bigcup_{m=1}^{\infty} \mathcal{X}^m$) com $w = w'u \notin \tau_n$:

$$\mathbb{P}[O_n] \leqq \sum_{|w| \leqq \log(n)/\log(|\mathcal{X}|)} (n+1)^{-C+2|\mathcal{X}|} \leqq n^{-C+2|\mathcal{X}|+1}$$

(a razão para a última desigualdade é que, para qualquer m, não há mais do que $|\mathcal{X}|^m$ sequencias distintas w com $|w| = m$), é possível eliminar a restrição quanto ao alcance de w. Vejamos inicialmente a seguinte proposição:

Proposição 5.0.4. *Seja swv com $s \in \tau_n$, $w \in \bigcup_{m=1}^{\infty} \mathcal{X}^m$, $v \in \mathcal{X}$ e $swv \notin \tau_n$. Seja $O_n(swv) = \{\Delta_{swv} \geqq C\log(n), N(swv) > 1\}$. Seja $p_{\min}(n) = \min_{x\in\mathcal{X}, w\in\tau_n} p_w(x)$ e $\hat{\tau}_{\max}$ a árvore contexto maximal do primeiro passo do Algoritmo do Contexto. Então, sob as hipóteses **A1**, **A2** e **A3**, temos*

$$\mathbb{P}[O_n(swv)] \leqq \frac{1}{p_{\min}(n)} \mathbb{P}[sw \in \hat{\tau}_{\max}] n^{-C+2|\mathcal{X}|}$$

Dem: Seja $s \in \tau_n$ um contexto e $su = swv$ com $w \in \bigcup_{m=1}^{\infty} \mathcal{X}^m$, $v \in \mathcal{X}$ e $swv \notin \tau_n$. Nosso objetivo é limitar superiormente a probabilidade da sobrestimação em su. Relembremos algumas desigualdades que se acham em Weinberger, Rissanen e Feder (ver [9]). Fixemos, inicialmente, uma sequência x_1^n, uma realização de P_n. Determinemos uma lei de probabilidade dada por $Q_{su}(y_1^n \mid x_1^n)$ como abaixo

$$\log(Q_{su}(y_1^n \mid x_1^n)) = R_{sw}(y_1^n \mid S_s) + \sum_{x\in\mathcal{X}} \sum_{b\neq v} N_{y_1^n}(x \mid swb) \log(\hat{P}_{x_1^n}(x \mid sw))$$
$$+ \sum_{x\in\mathcal{X}} N_{y_1^n}(x \mid su) \log(\hat{P}_{x_1^n}(x \mid su))$$

Onde, $R_{sw}(y_1^n \mid S_s)$ é formalmente definido em Weinberger, Rissanen e Feder (ver [9]) como a soma da log-probabilidade de todos os símbolos que ocorrem em qualquer contexto

que não seja sw. Cumpre observar que, para toda sequência y_1^n com $N_{y_1^n}(sw) = 0$, a probabilidade Q_{su} de y_1^n é a mesma que a probabilidade P_n.

Para cada x_1^n, seja $\sigma_{x_1^n}$ o conjunto de todas as sequências y_1^n tais que $N_{y_1^n}(xsw) = N_{x_1^n}(xsw)$ e $N_{y_1^n}(xswv) = N_{x_1^n}(xswv)$, para todo $x \in \mathcal{X}$. Se $\Delta_{x_1^n}(swv) > C\log(n)$ segue-se de (A9) de Weinberger, Rissanen e Feder (ver [9]) que

$$P_n(\sigma_{x_1^n}) \leqq Q_{su}(\sigma_{x_1^n} \mid x_1^n)n^{-C}. \tag{5.12}$$

Necessitamos, neste ponto, introduzir uma nova medida de probabilidade dada por Q' no conjunto das sequências de alcance n que tenha relação com Q_{su}. Para tanto, para cada sequência y_1^t, seja x_0 o símbolo que ocorre após a primeira ocorrência de sw. Seja b_0 o símbolo imediatamente anterior à primeira ocorrência de sw. Assim, x_0 ocorre no contexto swb_0. Se $b_0 \neq v$, definamos

$$\log(Q'(y_1^n \mid x_1^n)) = \log(Q_{su}(y_1^t \mid x_1^n)) + \log(P_n(x_0 \mid sw)) - \log(\hat{P}_{x_1^n}(x_0 \mid sw)).$$

Quando $b_0 = v$, coloquemos

$$\log(Q'(y_1^n \mid x_1^n)) = \log(Q_{su}(y_1^t \mid x_1^n)) + \log(P_n(x_0 \mid sw)) - \log(\hat{P}_{x_1^n}(x_0 \mid swv)).$$

Assim, se $N_{y_1^n}(sw) < 2$ teremos, forçosamente, $P_n(y_1^n) = Q'(y_1^n \mid x_1^n)$. Segue-se também da definição de Q' que

$$Q_{su}(y_1^n \mid x_1^n) \leqq \frac{1}{p_{\min}(n)} Q'(y_1^n \mid x_1^n).$$

Assim, junto com a desigualdade (5.12), podemos escrever

$$P_n(\sigma_{x_1^n}) \leqq Q'(\sigma_{x_1^n}) \frac{1}{p_{\min}(n)} n^{-C}.$$

A construção de $\sigma_{x_1^n}$ e fato de que $N_{x_1^n}(sw) > 1$ acarretam

$$Q'(\sigma_{x_1^n} \mid x_1^n) \leqq Q'(y_1^n; N_{y_1^n}(sw) > 1 \mid x_1^n) = P_n(y_1^n; N_{y_1^n}(sw) > 1) = P_n(sw \in \hat{\tau}_{\max}).$$

Além disso, como existem no máximo $n^{2|\mathcal{X}|}$ classes distintas de $\sigma_{x_1^n}$, segue-se que

$$\mathbb{P}[\, O_n(swv)] = P_n(y_1^n; \Delta_{y_1^n}(swv) > C\log(n)) \frac{1}{p_{\min}(n)} \leqq \mathbb{P}[\, sw \in \hat{\tau}_{\max}] n^{-C+2|\mathcal{X}|}$$

■

As duas proposições a seguir finalizam os preparativos.

Proposição 5.0.5. *Sob as hipóteses **A1**, **A2** e **A3**, temos*

$$\sum_{n=1}^{\infty} \mathbb{P}[\,O_n]\log(n) < \infty$$

Dem: Aplicando a proposição 5.0.4, para swv, temos

$$\mathbb{P}[\,O_n] \leqq \sum_{swv} \mathbb{P}[\,O_n(swv)] = \mathbf{O}(n^{-C+2\mathcal{X}+1}) \sum_{swv} \mathbb{P}[\,sw \in \hat{\tau}_{\max}]$$

Seja L o número de sequências que ocorrem pelo menos duas vezes nos dados X_1^n. Temos

$$\sum_{swv} \mathbb{P}[\,sw \in \hat{\tau}_{\max}] \leqq |\mathcal{X}|\mathbb{E}[\,\sum_{sw} 1_{[sw \text{ ocorre pelo menos duas vezes em } X_1^n]}\,] \leqq |\mathcal{X}|\mathbb{E}[\,L] \leqq |\mathcal{X}|n^2.$$

Levando em conta agora que $C > 2|\mathcal{X}| + 4$, justificamos a proposição. ■

Vamos destacar o fato seguinte:

Proposição 5.0.6. *Sob as hipóteses **A1**, **A2** e **A3**, com o corte do Algoritmo do Contexto satisfazendo $K_n \sim C\log(n)$, onde $C > 4|\mathcal{X}| + 8$, temos*

$$\sum_{n=1}^{\infty} \mathbb{P}[\,\tilde{O}_n]\log(n) < \infty$$

Dem: Como já empregamos na proposição 5.0.3, temos $D(P\,||\,Q) \geqq \frac{1}{2}||P - Q||_1^2$. Logo $\tilde{\Delta}_w \leqq 2\Delta_w$ ■

Vamos agora reenunciar o primeiro resultado e prová-lo com os instrumentos disponíveis. **Teorema** Seja $X_{1,n}, \ldots, X_{n,n}$, com P_n satisfazendo as condições **A1**, **A2** e **A3**. Seja $\hat{P}(\cdot\,|\,\cdot)$ como na equação (2.2), e $\hat{c}(\cdot)$ a estimativa da função contexto dada pelo Algoritmo do Contexto. Então:

1. $\lim_{n\to\infty} \mathbb{P}[\hat{c}(\cdot) = c_n(\cdot)] = 1$.

2. $\sup_{x_{-\infty}^1 \in \mathcal{X}^\infty} |\hat{P}(\,x_1|\hat{c}(x_{-\infty}^0)\,) - p_n(\,x_1|c(x_{-\infty}^0)\,)| = \mathbf{o}_P(1)$ quando n tende ao infinito.

Dem: Podemos agora arrematar a prova do teorema 5.0.1. As proposições 5.0.3 e 5.0.5 acarretam o primeiro item do teorema. O segundo item segue-se do primeiro na mesma

linha da prova da proposição 5.0.3: trata-se de usar a partição com os conjuntos D_n e usar as proposições 5.0.2 e 5.0.1. ■

Para facilitar o enunciado do próximo teorema, vamos introduzir as notações a seguir.

Dadas duas medidas de probabilidade P e Q sobre $\mathcal{X}^\infty$, consideremos as medidas por elas induzidas sobre $\mathcal{X}^m$, $1 \leqq m < \infty$, isto é, as medidas que associam a um subconjunto mensurável B de $\mathcal{X}^m$ os valores $P\{x = (x_1, \ldots, x_n, \ldots) \in \mathcal{X}^\infty | (x_1, \ldots, x_m) \in B\}$ e $Q\{x = (x_1, \ldots, x_n, \ldots) \in \mathcal{X}^\infty | (x_1, \ldots, x_m) \in B\}$, respectivamente. Vamos indicar tais medidas por $P \circ \pi^{-1}_{1,\ldots,m}$ e $Q \circ \pi^{-1}_{1,\ldots,m}$, respectivamente. Então,

$$d_m(P \circ \pi^{-1}_{1,\ldots,m}, Q \circ \pi^{-1}_{1,\ldots,m}) = \sup_{x \in \mathcal{X}^m} |P \circ \pi^{-1}_{1,\ldots,m}\{x\} - Q \circ \pi^{-1}_{1,\ldots,m}\{x\}| \quad \text{e} \tag{5.13}$$

$$d(P, Q) = \sum_{m=1}^{\infty} 2^{-m} d_m(P \circ \pi^{-1}_{1,\ldots,m}, Q \circ \pi^{-1}_{1,\ldots,m}). \tag{5.14}$$

De posse dessas notações podemos enunciar nosso próximo teorema.

Teorema 5.0.3. Seja $X_{1,n}, \ldots, X_{n,n}$ com P_n satisfazendo as condições **A1**, **A2** e **A3**. Então,

1. Para $\hat{P}(\cdot \,|\, \cdot)$ como na equação (2.2), e $\hat{c}(\cdot)$ estimativa da função contexto dada pelo Algoritmo do Contexto, temos

   $$\lim_{n \to \infty} \mathbb{P}[\text{o conjunto } \{\hat{P}(\,\cdot | \hat{c}(x^0_{-\infty})\,) \,|\, x^0_{-\infty} \in \mathcal{X}^\infty\}$$
   $$\text{gera uma única medida de probabilidade estacionária } \hat{P}_n \in \mathcal{P}] = 1.$$

2. Para $\hat{P}_n$ como no item anterior, e $d(\cdot, \cdot)$ como na equação (5.14), temos $d(\hat{P}_n, P_n) = o_P(1)$ quando n tende ao infinito.

3. O processo $\hat{P}_n$ do primeiro item satisfaz

   $$\mathbb{P}[\,\hat{P}_n \text{ é } \phi - \text{mixing com coeficientes de mixing}$$
   $$\text{satisfazendo } \phi_{\hat{P}_n}(j) \leqq (1 - \kappa)^j, \text{ para todo } j \in \mathbb{N}] \longrightarrow 1 \text{ quando } n \to \infty.$$

   Em particular, o limite para os coeficientes $\phi_{\hat{P}_n}(j)$ é não-aleatório e o mesmo para todo natural n.

Dem: Os items 1 e 3 seguem da fórmula (5.15) abaixo, equanto que o item 2 decorre do teorema anterior.

$$\begin{aligned}
&\sup_{v,w,w'\in\mathcal{X}^{\hat{k}_n}} |\hat{P}_r^{(n)}(v,w) - \hat{P}_r^{(n)}(v,w')| \\
&= \sup_{v,w,w'\in\mathcal{X}^{\hat{k}_n}} |T(v\,|\,w;\, r,\, \hat{P}(\cdot\,|\,\cdot),\, \hat{c}(\cdot)) - T(v\,|\,w';\, r,\, \hat{P}(\cdot\,|\,\cdot),\, \hat{c}(\cdot))| \qquad (5.15)\\
&\leqq 1 - 2\kappa + \mathbf{o}(1) \text{ sobre } A_n
\end{aligned}$$

Vamos iniciar com uma forma análoga da hipótese **A1** para o processo estimado $\hat{P}_n$. O núcleo de transição para o passo r, $P_r^{(n)}(v,w) = \mathbb{P}_{P_n}[X_{r-k_n+i,n}^{r,n} = v\,|\,X_{-k_n+i,n}^{0,n} = w]$, da Cadeia de Markov $(X_{t-k_n+i,n}^{t,n})_{t\in\mathbb{Z}}$ que contém a CMAV P_n, pode ser caracterizado pelas probabilidades de transição $p_n(\cdot\,|\,\cdot)$ e pela função contexto $c_n(\cdot)$ da seguinte forma:

$$\begin{aligned}
T(v\,|\,w;\, r,\, p_n(\cdot\,|\,\cdot),\, c_n(\cdot)) &= P_r^{(n)}(v,w) \\
&= \sum_{x_1^r\in\mathcal{X}^r,\,(x_1^r w)_{a-k_n+1}^a = v} \prod_{i=0}^{r-1} p_n(x_{r-1}\,|\,c_n(x_1^{r-i-1}w)), \qquad (5.16)
\end{aligned}$$

onde $a = r + |w|$. Note que, para cada $n \in \mathbb{N}$, o processo $(Y_{t,n})_{t\in\mathbb{Z}} \sim \hat{P}_n$ é uma CMAV. Considere seu núcleo de transição de passo r, $\hat{P}_r^{(n)}(v,w) = \mathbb{P}_{\hat{P}_n}[Y_{r-\hat{k}_n+i,n}^{r,n} = v\,|\,Y_{-\hat{k}_n+i,n}^{0,n} = w]$ da Cadeia de Markov estendida $(Y_{t-\hat{k}_n+i,n}^{t,n})_{t\in\mathbb{Z}}$ da CMAV $\hat{P}_n$ de ordem $\hat{k}_n$. Essa transição é caracterizada por

$$T(v\,|\,w;\, r,\, \hat{P}_n(\cdot\,|\,\cdot),\, \hat{c}_n(\cdot)) = \hat{P}_r^{(n)}(v,w), \qquad r \geqq 1.$$

Vamos agora obter o análogo de **A1** para $\hat{P}_n$. Considere os conjuntos $A_n = \{\omega\,|\,\hat{c}(\cdot,\omega) = c_n(\cdot)\}$. Pelo item 1 do teorema 5.0.1, $\hat{k}_n = k_n$ em A_n e para w, $w' \in \mathcal{X}^{\hat{k}_n}$, temos

$$\begin{aligned}
&|T(v\,|\,w;\, r,\, \hat{P}(\cdot\,|\,\cdot),\, \hat{c}(\cdot)) - T(v\,|\,w';\, r,\, \hat{P}(\cdot\,|\,\cdot),\, \hat{c}(\cdot))| \\
&\leqq |T(v\,|\,w;\, r,\, p_n(\cdot\,|\,\cdot),\, c_n(\cdot)) - T(v\,|\,w';\, r,\, p_n(\cdot\,|\,\cdot),\, c_n(\cdot))| \\
&\quad + |T(v\,|\,w;\, r,\, \hat{P}(\cdot\,|\,\cdot),\, c_n(\cdot)) - T(v\,|\,w;\, r,\, p_n(\cdot\,|\,\cdot),\, c_n(\cdot))| \\
&\quad + |T(v\,|\,w';\, r,\, \hat{P}(\cdot\,|\,\cdot),\, c_n(\cdot)) - T(v\,|\,w';\, r,\, p_n(\cdot\,|\,\cdot),\, c_n(\cdot))| \quad \text{sobre } A_n.
\end{aligned}$$

Utilizando agora a hipótese **A1** para $T(\cdot\,|\,\cdot\,;\, r,\, p_n(\cdot\,|\,\cdot),\, c_n(\cdot))$ com relação ao real processo subjacente. Com relação aos outros termos, utilizaremos a finitude de r e $\mathcal{X}$, bem como

o item 2 do teorema 5.0.1 e a equação (5.16). Temos, então, como resultado o conteúdo de (5.15).

Assim, sobre A_n, $\hat{P}_n$, da forma que foi construída, segundo o exposto no item 2 desse teorema, está univocamente definida, é estacionária e ϕ-mixing, com coeficientes de mixing limitados superiormente como abaixo.

$$\phi_{\hat{P}_n}(j) \leqq (1-\kappa)^j, \text{ para todo } j \in \mathbb{N} \text{ sobre } A_n$$

(ver Rajarshi, [7], Lemma 2.1 ou Doukhan [2]).

Como, pelo item 1 do teorema 5.0.1, temos $\mathbb{P}[A_n] \to 1$ quando $n \to \infty$, os resultados desse teorema estão justificados.

■

Apêndice 2: Rotinas Empregadas

As rotinas† às quais se refere o capítulo anterior são as seguintes:

Gerador

```
class Gerador{

int anterior = 0, atual;
double prob;
double p100 = 0.12; // P(1/00) = 0.12
double p101 = 0.75; // P(1/01) = 0.75
double p11 = 0.53; // P(1/1) = 0.53

java.util.Random r = new java.util.Random();
void gera(int n){
prob = r.nextDouble();
if(prob > (p100 + p101 + p11)/3 )
atual = 0;
else
atual = 1;

System.out.print(atual+" ");
```

† Rotinas gentilmente cedidas por Mariana Pereira de Melo.

```
while(n > 0){
    prob = r.nextDouble();
    if(atual == 1){
        anterior = atual;
    if(prob > p11)
        atual = 0;
    else
        atual = 1;
    }
    else{
        if(anterior == 0){
        anterior = atual;
        if(prob > p100)
            atual = 0;
        else
            atual = 1;
    }
    else{
        anterior = atual;
        if(prob > p101)
            atual = 0;
        else
            atual = 1;
    }
    }
    System.out.print(atual+" ");
    n--;
    if( n%30 == 0)
        System.out.println();
```

```
    }
    }

    public static void main(String [] arg){
        Gerador g = new Gerador();
    g.gera(Integer.parseInt(arg[0]));
    }
} \acute{}
```

Rotina em "R" para o Algoritmo do Contexto

A rotina é a seguinte:

```
#######################################################################
                          "contextos"
Pede para que o usuário informe o diretório onde estão localizados
os arquivos. Pega a lista de arquivos de extensão ".bin"
localizados no diretório informado, e tenta ler cada arquivo.
Chama a função que contrói as tabelas para cada texto e
finalmente, chama a função que contrói as tabelas de comparação.
#######################################################################

contextos <- function(dir = paste(c(dirname(choose.files()),"/"),
collapse="")){

    mat <- c()
    textos <- list.files(dir,pattern=".txt")

## pega a lista de arquivos de formato '.txt'
```

```
    cat("########## TABELAS DE CONTEXTOS E
TRANSIÇÃO ##########\n\n",file=paste(c(dir,"saida.ctx"),collaps
e=""))

    for(i in 1:length(textos)){
        arquivo <-  paste(c(dir,textos[i]),
collapse="")
        cat("Carregando arquivo
",arquivo,"...\n")

    cat("\n",textos[i],"\n",file=paste(c(dir,"saida
.ctx"),collapse=""), append=TRUE)

        vet <- leitura(arquivo)

        Contexto(vet,length(texto),dir)
    }
    file.show(paste(c(dir,"saida.ctx"),collapse="")
) }

#######################################################################
                        "leitura"
Função que faz a leitura do arquivo devolvendo um vetor com o
texto
#######################################################################

leitura <- function(arquivo){

    texto <- readLines(arquivo)
```

```
    vet <- c()
    for(a in 1:length(texto)){
        vet <- c(vet,strsplit(texto[a],"[\s,
]+")[[1]])
    }

    vet
}

#####################################################################
                        "Contexto"
Função que estima, separadamente, a árvore contexto para cada
texto, e devolve uma matriz com a tabela de contextos e suas
respectivas probabilidades. Parâmetros:
* texto : vetor com o texto(codificado)
* nf    : número de frases do texto
* dir   : diretório onde se localiza o texto
#####################################################################

Contexto<- function(texto,nf,dir){
    prec=3            ## Precisão da probabilidade
(3 casas decimais. Pode ser alterado)
    vlmc <- c()       ## Variável que irá guardar o
VLMC do texto

    n <- length(alphabet(vlmc(texto)))
    c <- 2*n + 5
    k <-
(((c*log(length(texto),n))^(1/3))+((c*log(length(text
o),n))^(1/2)))/2
```

```
    if(n == 1)
        k <- c

    vlmcObject <- vlmc(texto, cutoff = k)

    prec <- min(prec, 7)          ## precisão
        vect <- vlmcObject$vlmc.vec  ## vetor que
armazena a árvore
        indice <- 2               ## índice da
posição corrente do vetor que armazena a arvore
        vetor <- c()              ## Vetor com
o contexto de determinado galho
    contextos <- c()              ## matriz
unidimencional com a tabela do de contextos

    aux <- contextos.rec(vlmcObject,c(), vetor,
indice, prec)
    contextos <- aux[-length(aux)]

    mat <-
matrix(contextos,length(contextos)/(vect[1]+3),vect[1 ]+3,byrow=T)
    colnames(mat) <-
c("Contextos",alphabet(vlmcObject)," n"," freq")
    vetAux <- mat[,vect[1]+3]
    mat[,vect[1]+3] <-
double.precisao(as.integer(vetAux)/sum(as.integer(vet Aux)),prec)
    cat("n: ",length(vlmcObject$y)," ","K:",k,"
AIC:",AIC(vlmcObject),"\n")
    cat("n: ",length(vlmcObject$y)," ","K:",k,"
AIC:",AIC(vlmcObject),"\n",file=paste(c(dir,"saida.ct
```

```
x"),collapse=""), append=TRUE)
    draw(vlmcObject)
    imprime(mat,diret=dir,colnames(mat)[-1])
}

#####################################################################
threshold.gen = 15
                        "contexto.rec"
Função recursiva que caminha pelos galhos da árvore contexto e
monta tabela de contexto sempre que chega em um terminal. Também
completa a árvore sempre que determinado galho não aparece na
árvore.
Parâmetros:
    * x     : VLMC do texto
    * cont  : símbolo atual
    * vet   : contexto até esse galho
    * i     : índice do vetor que armazena a árvore
    * prec  : precisão
#####################################################################

contextos.rec <- function(x,cont, vet, i, prec){
    vec <- x$vlmc.vec              ## Vetor que
armazena a árvore
    alphabet <- alphabet(x)        ## Vetor com o
alfabeto usado
        context <- c()             ## vetor que
armazenará o numero de vezes que determinado
                                   ## contexto
apareceu seguido por determinada letra do alfabeto usado
    terminal <- c()                ## irá guardar as
```

```
direções onde não foi encontrado um terminal
    contextos <- c()
        if(i<=length(vec) & vec[i]!=-1){ ## Se o
indice é menor que o tamanho do vetor e
                i <- i+1                ## se o nó
anterior não foi terminal
                vet <- c(cont, vet)
        atual <- 0
                while(atual < vec[1]){   ## Enquanto
atual é menor que o tamanho do alfabeto
                        context <- c(context, vec[i])
                        i <- i+1
            atual <- atual+1
                }
        atual <- 1
                while(atual <= vec[1]){ ## Enquanto
atual é menor que o tamanho do alfabeto
                        aux <- contextos.rec(x,atual,
vet, i, prec) ## recebe um vetor com o contexto

## em determinada direção até o terminal
            if(length(aux)==1)
## se só tem uma posição
                terminal <- c(terminal,atual)
## não chegou a um terminal nessa direção
            else
                contextos <-
c(contextos,aux[-length(aux)])
                        i <-
as.integer(aux[length(aux)])            ## atualiza o indice
```

```
            atual <- atual+1
## Segue na próxima direção
                }

        if(length(terminal) == vec[1])
## esse contexto é terminal
                    contextos <-
c(contextos,transicao(vet, context, prec,alphabet))
        else{
## encontra os outros galhos que estão faltando
            atual = 1
            while(atual <= length(terminal)){
                aux <-
busca(c(terminal[atual],vet),x$y, alphabet)
                contextos <- c(contextos,
transicao(c(terminal[atual],vet),aux,prec,alphabet))
                atual <- atual+1
            }
        }
        }
        else{
                i <- i+1
        }

    if(length(contextos)>0)
            c(contextos, i) ## Devolve um vetor com
os contextos terminais dos seus galhos,
                                ## e suas respectivas
probabilidaes, além da posição do vetor
                                ## que armazena a
```

```
árvore
    else
        i                ## se não tem galhos,
devolve 1 }

#####################################################################
                        "transicao"
Função que calcula as probabilidades de transição.
Parâmetros:
    * contexto  : vetor com o contexto
    * transicao : vetor com os valores de transição
    * prec      : precisão
    * alphabet  : vetor contendo o alfabeto usado
#####################################################################

transicao <- function(contexto, transicao,  prec, alphabet){
    contextos <- c()
        soma <- sum(as.integer(transicao))
    if(soma > 0){
        contextos <-
c(paste(alphabet[contexto],collapse=""))
        a <- 1
        while(a <= length(transicao)){
            prob <-
double.precisao(transicao[a]/soma,prec)
            contextos <- c(contextos,prob)
            a <- a+1
        }

        contextos <- c(contextos,soma,soma)
```

```
}


    contextos    ## devolve esse vetor que contém uma String na primeira posição,
                 ## o valor de cada probabilidade de transição nas demais posições
                 ## e o numero de vezes que esse contexto apareceu.

#####################################################################
                        "double.precisao"
Devolve o número real passado como parâmetro com um determinado
numero de casas decimais passado como parâmetro também.
#####################################################################

double.precisao <- function(num, precisao){
    if(!is.double(num)) stop("Esse nao é um numero
Real")
    num <- as.integer(num*(10^precisao))
    num <- num/(10^precisao)

    num
}

#####################################################################
                        "imprime"
Salva num arquivo de nome saida.ctx as tabelas de contextos e
transição.
Parâmetros:
    * matrix : matriz com as informações da tabela
    * diret  : diretório onde será salvo o arquivo
    * alph   : alfabeto ou nome dos textos
```

```
#####################################################################

imprime <- function(matrix,diret="",alph=c()){
    saida <- c()
    arquivo <-
paste(c(diret,"saida.ctx"),collapse="")
    max <- max(nchar(matrix[,1]),3)
    len <- length(matrix[1,])-1
    saida <- c(saida,"_")
    a <- 0
    while(a < max){
        saida <- c(saida,"__")
        a <- a+1
    }
    saida <- c(saida,"__")
    a <- 0
    while(a < len){
        saida <- c(saida,"_______")
        a <- a+1
    }
    saida <- c(saida,"\n|Cont.")
    a <- 2
    while(a < max){
        saida <- c(saida,"  ")
        a <- a+1
    }
    saida <- c(saida,"|")
    a <- as.integer((len*7-15)/2)
    while(a > 0){
        saida <- c(saida," ")
```

```
        a <- a-1
    }
    saida <- c(saida,"Probabilidades")
    a <- as.integer((len*7-15)/2)+14
    while(a < len*7-1){
        saida <- c(saida," ")
        a <- a+1
    }
    saida <- c(saida,"|\n|")
    if(length(alph)==len){
        a <- 0
        while(a < max){
            saida <- c(saida,"  ")
            a <- a+1
        }
        saida <- c(saida," | ")
        n <- 1
        while(n < len){
            prob <- paste(alph[n],collapse="")
            saida <- c(saida,prob)
            l <- nchar(prob)
            while(l < 5){
                saida <- c(saida," ")
                l <- l+1
            }
            saida <- c(saida,"  ")
            n <- n+1
        }
        prob <- strsplit(alph[n],NULL)[[1]]
        prob <- prob[-1]
```

```
        prob <- prob[-1]
        prob <- paste(prob,collapse="")
        saida <- c(saida,prob)
        l <- nchar(prob)
        while(l < 5){
            saida <- c(saida," ")
            l <- l+1
        }
        saida <- c(saida,"|\n|")
    }
    a <- 0
    while(a < max){
        saida <- c(saida,"__")
        a <- a+1
    }
    saida <- c(saida,"_|")
    a <- 0
    while(a < len-1){
        saida <- c(saida,"_______")
        a <- a+1
    }
    saida <- c(saida,"______|\n")

    a <- 1
    while(a <= length(matrix[,1])){
        context <-
strsplit(as.character(matrix[a,1]),NULL)[[1]]
        saida <- c(saida,"|")
        n <- 0
        while(n < (max-length(context))){
```

```
            saida <- c(saida,"  ")
        n <- n+1
        }
            saida <-
c(saida,paste(context,collapse=" "),"  | ")
        n <- 2
        while(n <= len+1){
            prob <- matrix[a,n]
            saida <- c(saida,prob)
            prob <-
strsplit(as.character(prob),NULL)[[1]]
            l <- length(prob)
            while(l < 5){
                saida <- c(saida," ")
                l <- l+1
            }
            saida <- c(saida,"| ")
            n <- n+1
        }
        saida <- c(saida,"\n")



        a <- a+1
    }

    saida <- c(saida,"|")
    a <- 0
    while(a < max){
        saida <- c(saida,"__")
        a <- a+1
```

```
    }
    saida <- c(saida,"_|")
    a <- 0
    while(a < len){
        saida <- c(saida,"______|")
        a <- a+1
    }
    cat(paste(saida, collapse=""),"\n",
file=arquivo, append=TRUE) }


busca <- function(contexto, vec, alphabet){
    contexto <- as.integer(contexto)
    aux <- c()
    for(i in alphabet)
        aux <- c(aux,0)
    i <- 1
    while(i <= length(vec)-length(contexto)){
        if(vec[i] == alphabet[contexto[1]]){
            OK = TRUE
            j = 2
            while(j<=length(contexto) && OK){
                if(alphabet[contexto[j]] !=
vec[i+j-1]){
                    OK = FALSE
                }
                j <- j+1
            }
            if(OK){
                c <- vec[i+j-1]
```

```
                n <- 1
                while(n < length(alphabet) &&
alphabet[n] != c)
                    n <- n+1
                aux[n] <- aux[n]+1
            }
        }
        i <- i+1
    }

    aux
}
```

Referências Bibliográficas

[1] Cover, T. M. and Thomas, J. A. 1991. *Elements of Information Theory.* Wiley, New York.

[2] Doukhan, P. 1994. *Mixing Properties and Examples.* Lecture Notes in Statist. 85. Springer, Berlin.

[3] Fiorenzo Ferrari and Abraham J. Wyner, 2003. *Estimation of General Stationary Processes by Variable Length Markov Chains*, Scandinavian Journal of Statistics, Vol 3, Issue 3, September 2003, p. 549.

[4] Iosifescu, M. and Theodorescu, R. 1969. *Random Processes and Learning.* Springer, Berlin.

[5] Peter Bühlmann and Abraham J. Wyner, 1999. *Variable length Markov chains*, The Annals of Statistics, Vol 27, No. 2—April 1999, p. 480.

[6] Peter Bühlmann and Martin Mächler, 2002. *Variable Length Markov Chains: Methodology, Computing and Software*, Seminar für Statistik Eidgenössische Technische Hochschule (ETH), Research Report No. 104 March 2002.

[7] Rajarshi, M. B. 1990. *Bootstrap in Markov-sequences based on estimates of transition density.* Ann. Inst. Statist. Math. 42 253–268.

[8] Rissanen, J. (1983). *A universal data compression system*, IEEE Trans. Inform. Theory IT - 29, 656-664.

[9] Weinberger, M. J., Rissanen, J. and Feder, M. 1995. *A universal finite memory source.* IEEE Trans. Inform. Theory IT-41 643–652.

[10] Withers, C. S. 1981. *Central limit theorems for dependent variables* I. Z. Wahrsch. Verw. Gebiete 57 509–534 Corrigendum 1983 63 555.