UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

INFERÊNCIA ESTOCÁSTICA E MODELOS DE MISTURA DE DISTRIBUIÇÕES

Dissertação de Mestrado

Regis Nunes Vargas

Porto Alegre, 1° de dezembro de 2011.

Dissertação submetida por Regis Nunes Vargas como requisito parcial para a obtenção do grau de Mestre em Matemática pelo Programa de Pós-Graduação em Matemática do Instituto de Matemática da Universidade Federal do Rio Grande do Sul.

Professor Orientador:
Dr. Cleber Bisognin

Banca Examinadora:
Dr. Cleber Bisognin - PPGMAT(UFRGS)
Dr. Flávia Tereza Giordani - UFSC
Dr. Marcio Valk - UFRGS
Drª. Sílvia Regina Costa Lopes - PPGMAT(UFRGS)

Data da Defesa: 1° de dezembro de 2011.

Resumo

Neste trabalho apresentamos os resultados de consistência e normalidade assintótica para o estimador de máxima verossimilhança de uma Cadeia de Markov ergódica. Além disso apresentaremos os Modelos de Mistura de Distribuição Independente e um dos casos de Modelos de Mistura Dependente: os Modelos Ocultos de Markov. Estimaremos os parâmetros destes modelos a partir do método da máxima verossimilhança e abordaremos o critério de seleção através do cálculo do AIC e BIC.

Abstract

This paper presents the results of consistency and asymptotic normality for the maximum likelihood estimator of the ergodic Markov chain. In addition we present the Independent Mixture Models and one case of Dependent Mixture Models: the Hidden Markov Models. We estimate the parameters of these models from the maximum likelihood method and discuss the selection criteria by calculating the AIC and BIC.

Conteúdo

Capítulo 1

Introdução

Quando uma sequência de dados observáveis nos é apresentada podemos pressupor quais modelos seriam mais adequados a eles. Neste trabalho abordaremos três tipos de modelagem de dados: Cadeias de Markov, Modelo de Mistura de Distribuições Independente e Modelo Oculto de Markov (*Hidden Markov Model*).

Para que a teoria proposta por este trabalho seja apresentada de forma clara,x precisaremos conhecer uma série de conceitos relacionados ao estudo de processos estocásticos, em especial, as Cadeias de Markov à tempo discreto. Neste trabalho, nossa preocupação com estes pré-requisitos é traduzida na série de definições, propriedades e teoremas que serão apresentados logo de início ao leitor. Desta forma pretendemos possibilitar ao leitor clareza com relação aos assuntos que serão tratados após esta abordagem inicial e, se possível, fazer com que o mesmo não sinta a necessidade de buscar em outras bibliografias esclarecimentos para eventuais dúvidas que possam surgir durante a leitura deste trabalho. A seguir, tentaremos explicar de forma breve uma importante relação entre as probabilidades de transição de uma Cadeia de Markov e o Método de Máxima Verossimilhança.

Relacionaremos Inferência Estatística e Processos Estocásticos a partir da possibilidade de estimarmos as probabilidades de transição de uma Cadeia de Markov finita à tempo discreto através do Método de Máxima Verossimilhança. Devemos, no entanto, atentarmos para o fato de que este método seja adequado a uma realização (caminho) que obedece a propriedade markoviana. Isto é, precisamos lembrar que na realização de uma Cadeia de Markov "o estado atual depende única e exclusivamente do estado imediatamente anterior", o leitor interessado em saber como é feita esta contextualização pode consultar a Equação 2.4 na página 12. Uma abordagem inicial sobre este assunto é feita em Atuncar (2009, pág. 44-55), no qual é apresentado o estimador de máxima verossimilhança para as probabilidades de transição.

Atuncar (2009) também estima a distribuição invariante através da Lei Forte dos Grandes Números. A consistência de tal estimador é provada em Guttorp (1995), onde também é provada a sua normalidade assintótica. Uma abordagem rigorosa sobre Inferência Estocástica, aplicação das ideias da Inferência Estatística em Processos Estocásticos, é feita em Billingsley (1961). Além de provar as propriedades, consistência e normalidade assintótica dos estimadores Billingsley (1961), também apresenta algumas aplicações para processos discretos e contínuos.

Além da modelagem de dados através da Cadeia de Markov, também apresentaremos os Modelos de Mistura de Distribuições, uma das motivações para a utilização de tais modelos reside no fato de que quando tentamos adequar um modelo estatístico à um conjunto de dados

amostrais partindo do pressuposto de que a sequência em questão é proveniente de uma única distribuição, ocorre em muitos casos superdispersão, isto é, ocorre uma grande disparidade entre a variância amostral e a variância do modelo, chamada variância teórica.

O Modelo de Mistura de Distribuições surge então como uma alternativa valiosa na tentativa de solucionar tal problema, uma vez que a partir da hipótese de que existam subpopulações dentro de uma população geral, conseguimos diminuir consideravelmente a superdispersão. Cada uma dessas subpopulações são representadas por uma quantidade finita (neste trabalho abordaremos o caso finito) de distribuições componentes. Se as variáveis aleatórias com relação às quais as distribuições componentes estão associadas forem consideradas independentes entre si, então estaremos modelando o conjunto de dados pelo chamado Modelo de Mistura de Distribuições Independente. Se para estas mesmas variáveis aleatórias for considerada uma relação de dependência (relação esta que poderá ser indicada pela função de autocorrelação amostral), então trabalharemos com um dos casos de Modelo de Mistura de Distribuições Dependente: o Modelo Oculto de Markov. É importante destacar que, neste trabalho, utilizaremos, especificamente as distribuições componentes do tipo Poisson.

Um Modelo de Mistura de Distribuições Independente, caso finito, consiste na combinação linear de funções densidade (ou massa) de probabilidade, neste trabalho apresentaremos uma combinação linear de funções que obedecem à uma distribuição do tipo Poisson. Cada uma dessas funções possui um parâmetro, a ser estimado através do Método de Máxima Verossimilhança. Tais parâmetros serão estimados por rotinas implementadas em ambiente R. A utilização do Modelo de Mistura de Distribuições Independente pressupõe que os dados amostrais possuem independência entre si. Porém, algumas vezes, a função de autocorrelação amostral nos indica dependência. O que nos sugere o uso de um modelo de mistura dependente, um destes modelos é o Modelo Oculto de Markov.

O Modelo Oculto de Markov consiste em estimar, a partir dos dados amostrais, as probabilidades de transição e a distribuição inicial da Cadeia de Markov, que supostamente é o processo gerador dos dados amostrais. Assim um Modelo Oculto de Markov é composto por uma parte observável (amostra) e uma parte não observável (processo gerador), a qual neste caso, supõe-se ser uma Cadeia de Markov. Estimaremos os parâmetros das distribuições componentes e as probabilidades de transições pelo Método de Máxima Verossimilhança, também através de rotinas implementadas em ambiente R.

Quanto as propriedades assintóticas de um estimador de máxima verossimilhança para um Modelo Oculto de Markov, podemos citar Baum e Petrie (1966) que prova a consistência e a normalidade assintótica em um conjunto de estados finito. Leurox (1992), estabelece a consistência de forma geral, Bickel e Ritov (1996), provam a normalidade assintótica local e Rydén (1994) que propõe uma nova classe de estimadores e prova a normalidade assintótica sob certas condições de regularidade.

A primeira seção do Capítulo 2, tem por objetivo, estabelecer os conceitos iniciais sobre Cadeias de Markov. A maior parte desta seção possui como base teórica duas referências: Norris (2004) e Karlin e Taylor (1975). Excetuando-se as Definições: 2.5, 2.6 e 2.8; as quais foram obtidas de Zucchini e MacDonald (2009). Além disso, foram obtidos de Guttorp (1995): Teorema 2.3, Definição 2.23, Teorema 2.5 e Teorema 2.6. Guttorp (1995) foi a principal referência para a produção da segunda seção do Capítulo 2.

No Capítulo 3, apresentaremos o Modelo de Mistura de Distribuições Independente do tipo Poisson, caso finito, algumas de suas propriedades e a estimativa de seus parâmetros usando o

ambiente R.

No Capítulo 4, apresentaremos o Modelo Oculto de Markov do tipo Poisson, caso finito, referenciaremos algumas das propriedades de seu estimador de máxima verossimilhança e estimaremos seus parâmetros utilizando o ambiente R.

Em ambos os Capítulos, 3 e 4, usa-se o ambiente R para obter as estimativas dos parâmetros dadas a partir da maximização numérica da Função de Verossimilhança. Para o exemplo 3.2 utiliza-se o programa MMIX e para o exemplo 4.3 utiliza-se uma série de programas obtidos de Zucchini e MacDonald (2009). Também aborda-se brevemente, um método de seleção baseado no cálculo do AIC e BIC.

No Capítulo 5, os dados *Old Faithful Geyser* (ver Zucchini e MacDonald, 1997) serão modelados pela Cadeia de Markov, pelo Modelo de Mistura de Distribuições Independente e pelo Modelo Oculto de Markov. A seleção do "melhor" modelo, será feita a partir do cálculo do AIC e BIC, ver Zucchini e MacDonald (2009).

Capítulo 2

Cadeia de Markov: conceitos iniciais e inferência estocástica

O objetivo deste capítulo é apresentar definições e propriedades sobre Cadeia de Markov, necessárias no decorrer do trabalho. Além disso, apresentaremos alguns resultados relacionados a inferência estocástica: os estimadores de máxima verossimilhança para as probabilidades de transição e algumas de suas propriedades, tais como sua consistência e sua distribuição assintótica.

2.1 Definições e Propriedades

Nesta seção apresentaremos algumas definições e teoremas relacionados à Cadeia de Markov como a *Propriedade de Markov* e o *Teorema Ergódico*. O leitor interessado em maiores detalhes pode consultar Norris (2004) e Karlin e Taylor (1975). A seguir apresentamos a definição de um processo estocástico.

Definição 2.1. Um *Processo Estocástico* é uma família de variáveis aleatórias $(X_t)_{t\in T}$, no qual todas as variáveis aleatórias estão definidas no mesmo espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$, sendo $T \neq \varnothing$ um conjunto de índices, Ω o espaço amostral, $\mathcal{A}$ a classe de eventos aleatórios (σ-álgebra) e $\mathbb{P} : \mathcal{A} \longrightarrow [0, 1]$ a medida de probabilidade.

Se o conjunto dos índices T for um subconjunto de $\mathbb{R}$, ou mesmo o próprio $\mathbb{R}$, diz-se que o Processo Estocástico é à tempo contínuo. Se T for finito ou enumerável, por exemplo, $T = 1, \cdots, n$, $T = \mathbb{Z}$ ou $T = \mathbb{N}$, obtemos um Processo Estocástico à tempo discreto, o qual será o alvo de nosso estudo.

Notação: neste trabalho consideraremos $T = \mathbb{N} = \{0, 1, 2, \cdots\}$.

Para que a Cadeia de Markov fique bem definida é indispensável estabelecer claramente o que significa a propriedade markoviana. O embasamento teórico daqui até a definição da Cadeia de Markov propriamente dita foi obtido de Norris (2004, páginas 1 e 2).

Dizemos que um Processo Estocástico é markoviano se satisfaz a seguinte propriedade. A Definição 2.2 a seguir foi apresentada por Norris (2004).

Definição 2.2. Seja $(X_t)_{t\in\mathbb{N}}$, um *Processo Estocástico,* à tempo discreto. Dizemos que este processo satisfaz a propriedade de Markov se

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

em que $x_{t+1}, \cdots, x_0$ são denominados os estados do processo.

Denominamos $\mathbb{P}(X_{t+1} = x_{t+1}|X_t = x_t)$ de probabilidade de transição do estado x_t para o estado x_{t+1}. A probabilidade de transição é comumente denotada por:

$$p_{x_t x_{t+1}} = \mathbb{P}(X_{t+1} = x_{t+1}|X_t = x_t).$$

A seguir, definimos a matriz estocástica de uma Cadeia de Markov, cujas entradas são suas probabilidades de transição.

Definição 2.3. Seja $(X_t)_{t\in\mathbb{N}}$ um Processo Estocástico com espaço de estados S. Seja $\boldsymbol{\mathcal{P}}$ uma matriz em que cada elemento é da forma $p_{xy} = \mathbb{P}(X_{t+1} = y|X_t = x)$, $x, y \in S$. Isto é, a $x-$ésima linha da matriz é a distribuição de probabilidade condicional dos valores de $[X_{t+1} = y]$ sob a condição $[X_t = x]$. As quantidades p_{xy} satisfazem $0 \leq p_{xy} \leq 1$, para todo x, $y \in S$ e $\sum_{y\in S} p_{xy} = 1$. Assim $\boldsymbol{\mathcal{P}}$ é chamada *matriz linha-estocástica*. Especificamente neste trabalho, chamamos $\boldsymbol{\mathcal{P}}$ apenas *matriz estocástica*.

Para que um processos estocástico $(X_t)_{t\in\mathbb{N}}$ seja uma Cadeia de Markov a seguinte Definição deve ser satisfeita.

Definição 2.4. Seja $(X_t)_{t\in\mathbb{N}}$ um Processo Estocástico com espaço de estado S. Seja $\boldsymbol{\lambda} = (\lambda_x : x \in S)$ uma medida de de probabilidade em S. O Processo Estocástico $(X_t)_{t\in\mathbb{N}}$ é uma *Cadeia de Markov* com distribuição inicial $\boldsymbol{\lambda}$ se

i) X_0 tem distribuição $\boldsymbol{\lambda}$. Isto é $\lambda_x = \mathbb{P}(X_0 = x)$ para todo $x \in S$;

ii) $(X_t)_{t\in\mathbb{N}}$ satisfaz a Definição 2.2.

Denotamos uma Cadeia de Markov com matriz de transição $\boldsymbol{\mathcal{P}}$ e distribuição inicial $\boldsymbol{\lambda}$ por Markov($\boldsymbol{\lambda}$,$\boldsymbol{\mathcal{P}}$).

A Definição 2.4, por si só, não é uma ferramenta útil em grande parte das demonstrações. O teorema a seguir é considerado uma segunda Definição para a Cadeia de Markov e foi apresentado em Norris (2004, página 2).

Teorema 2.1. *Um processo estocástico a tempo discreto* $(X_t)_{t\in\mathbb{N}}$ *é Markov(*$\boldsymbol{\lambda}$,$\boldsymbol{\mathcal{P}}$*) se, e somente se, para todo* $(x_1, \cdots, x_N) \in S$

$$\mathbb{P}(X_0 = x_0, \cdots, X_N = x_N) = \lambda_{x_0} p_{x_0 x_1} \cdots p_{x_{N-1} x_N}.$$

A seguir apresentamos o Teorema 2.2 (ver Norris, 2004). Este reforça a ideia de que uma Cadeia de Markov possui a propriedade de perda de memória, isto é, o estado atual depende exclusivamente do estado imediatamente anterior.

Notação: $\boldsymbol{\delta}_x = (\delta_{xy} : y \in S)$, para todo $x \in S$, em que δ_{xy} é dado por

$$\delta_{xy} = \begin{cases} 1, & \text{se } x = y; \\ 0, & \text{se } x \neq y. \end{cases}$$

Teorema 2.2. *Seja $(X_t)_{t\in\mathbb{N}}$ Markov($\boldsymbol{\lambda}$, $\boldsymbol{\mathcal{P}}$). Então, condicionada em $X_m = x$, $(X_{m+n})_{n\in\mathbb{N}}$ é Markov($\boldsymbol{\delta}_x, \boldsymbol{\mathcal{P}}$) e independente das variáveis aleatórias $X_0, \cdots, X_m$.*

A seguir definimos um vetor de probabilidade $\boldsymbol{u}$ de mesma dimensão do espaço de estados. Aqui consideraremos a dimensão do espaço de estados finita (ver Zucchini e MacDonald, 2009).

Definição 2.5. Seja $S = \{1, \cdots, m\}$, o espaço de estados de uma Cadeia de Markov. Definimos o vetor de probabilidade $\boldsymbol{u}$ por

$$\boldsymbol{u}(t) = (\mathbb{P}(X_t = 1), \cdots, \mathbb{P}(X_t = m)).$$

A partir da Definição 2.5 temos que o vetor de probabilidade $\boldsymbol{u}$ satisfaz o seguinte resultado

$$\begin{aligned}
\mathbf{u}(t)\boldsymbol{\mathcal{P}} &= (\sum_{x=1}^{m} \mathbb{P}(X_t = x)p_{x1}, \cdots, \sum_{x=1}^{m} \mathbb{P}(X_t = x)p_{xm}) \\
&= (\sum_{x=1}^{m} \mathbb{P}(X_t = x, X_{t+1} = 1), \cdots, \sum_{x=1}^{m} \mathbb{P}(X_t = x, X_{t+1} = m)) \\
&= (\mathbb{P}(X_{t+1} = 1), \cdots, \mathbb{P}(X_{t+1} = m)) = \mathbf{u}(t+1). \qquad (2.1)
\end{aligned}$$

A equação (2.1) apresenta uma relação de recorrência. Isto é se conhecermos a distribuição inicial $\boldsymbol{u}(0)$ da Cadeia de Markov e sua matriz de transição $\boldsymbol{\mathcal{P}}$, então podemos encontrar $\boldsymbol{u}(t+1)$ a partir da seguinte expressão.

$$\boldsymbol{u}(t+1) = \boldsymbol{u}(0)\boldsymbol{\mathcal{P}}^{t+1},$$

onde $\boldsymbol{\mathcal{P}}^{t+1}$, representa a matriz $\boldsymbol{\mathcal{P}}$ multiplicada por ela mesma $t+1$ vezes.

A seguir definimos Cadeias de Markov homogêneas e Cadeias de Markov estacionárias. Ver Zucchini e MacDonald (2009, p. 18).

Definição 2.6. Seja $(X_t)_{t\in\mathbb{N}}$ Markov($\boldsymbol{\lambda}$, $\boldsymbol{\mathcal{P}}$) com espaço de estados S. Diremos que esta Cadeia de Markov é *homogênea* se para $x,\ y \in S$ e $t,\ s \in \mathbb{N}$, $\mathbb{P}(X_{t+s} = y|X_s = x)$ independe de s. Ou seja

$$\mathbb{P}(X_{t+s} = y|X_s = x) = \mathbb{P}(X_t = y|X_s = x) = \mathbb{P}(X_t = y|X_0 = x) = p_{xy}^{(t)}$$

Uma notação comum para indicar o retorno a um estado x após t passos, em uma cadeia homogênea, é dada por:

$$\mathbb{P}(X_{t+s} = x|X_s = x) = \mathbb{P}(X_t = x|X_0 = x) = p_{xx}^{(t)}$$

Antes de definirmos a Cadeia de Markov estacionária definiremos a distribuição estacionária.

Definição 2.7. Seja $\boldsymbol{\lambda} = (\lambda_x : x \in S)$ um vetor satisfazendo $\lambda_x \geq 0$, para todo $x \in S$ e $\sum_{x\in S} \lambda_x = 1$. Dizemos que $\boldsymbol{\lambda}$ é uma *distribuição estacionária* se

$$\boldsymbol{\lambda}\boldsymbol{\mathcal{P}} = \boldsymbol{\lambda}.$$

A seguir definimos a Cadeia de Markov estacionária. A qual obedece uma condição adicional com relação a cadeia homogênea.

Definição 2.8. Seja $(X_t)_{t\in\mathbb{N}}$ Markov($\boldsymbol{\lambda}$, $\boldsymbol{\mathcal{P}}$) homogênea (ver Definição 2.6). Se além disso sua distribuição inicial $\mathbf{u}(0) = \boldsymbol{\lambda}$ for estacionária, isto é $\boldsymbol{\lambda}\boldsymbol{\mathcal{P}} = \boldsymbol{\lambda}$, então diremos que esta Cadeia de Markov é uma *Cadeia de Markov estacionária.*

É possível definir relação e classe de equivalência para o espaço de estados de uma Cadeia de Markov. Apresentaremos algumas definições relacionadas a esta questão.

Definição 2.9. Dizemos que o estado x é levado a y e escrevemos $x \to y$ se:

$$\mathbb{P}_x(X_t = y \text{ para algum } t \geq 0) > 0.$$

A seguinte relação de equivalência é de grande relevância para uma maior compreensão em demonstrações futuras.

Definição 2.10. Dizemos que x *se comunica com* y e escrevemos $x \leftrightarrow y$, se $x \to y$ e $y \to x$.

A prova de que a relação acima é, de fato, uma relação de equivalência pode ser obtida em Norris (2004, página 10). A seguir definimos classe comunicante.

Definição 2.11. Um conjunto que possui apenas estados que se comunicam é dito uma *classe comunicante.*

A seguir definimos classe fechada.

Definição 2.12. Uma classe comunicante F é chamada de *classe fechada* se satisfaz a seguinte condição:

$$x \in F \text{ e } x \to y, \text{ então } y \in F.$$

Sequencialmente definimos Cadeia de Markov irredutível.

Definição 2.13. Seja $(X_t)_{t\in\mathbb{N}}$ Markov($\boldsymbol{\lambda}, \boldsymbol{\mathcal{P}}$) com espaço de estados S. Dizemos que $(X_t)_{t\in\mathbb{N}}$ é uma *Cadeia de Markov irredutível* se todos os seus estados se comunicam. Isto é se S é uma classe fechada.

Em seguida trataremos da propriedade de recorrência e transiência de um estado pertencente ao espaço de estados de uma Cadeia de Markov.

Definição 2.14. Dizemos que um estado x é *recorrente* se

$$\mathbb{P}(X_n = x \text{ para um número infinitos de n's}) = 1.$$

Definição 2.15. Dizemos que um estado x é *transiente* se

$$\mathbb{P}(X_n = x \text{ para um número infinitos de n's}) = 0.$$

O Teorema 2.3 nos auxiliará na demonstração da consistência dos estimadores das probabilidades de transição. Maiores detalhes ver Guttorp (1995, pág. 59).

Teorema 2.3. *Seja $(X_t)_{t\in\mathbb{N}}$ uma Cadeia de Markov com espaço de estado S. Seja S_T o conjunto de seus estados transientes e S_p o conjunto de seus estados recorrentes, nós temos que se S é o espaço de estados, então $S = S_T \oplus S_p$ e $S_p = \Sigma C_i$ em que C_i são classes disjuntas e fechadas. Os sinais $\oplus$ e Σ são utilizados para indicar união disjunta.*

Demonstração. Seja $x \in S_p$, defina $C = \{y \in S_p : x \to y\}$. Pela recorrência $p_{xx}^{(t)} > 0$ para uma infinidade de $t's$, então $x \in C$.

Nós primeiro mostraremos que C é classe, isto é se $y \in C$ e $z \in C$, então $y \leftrightarrow z$. Pela recorrência junto à definição de C temos que $x \leftrightarrow y$ e $x \leftrightarrow z$, então $y \leftrightarrow z$.

Agora veremos que C é fechada, isto é se $y \in C$ e $y \to z$, então $z \in C$. Pela recorrência de y e pelo fato de $y \in C$, temos, respectivamente, que $y \to z \to y \to z$ e $x \to y \to z$. O que respectivamente mostra que z é recorrente e $x \to z$, ou seja $z \in C$.

Por último concluiremos que ΣC_i é união disjunta. Para isto, considere C e D classes fechadas contidas em S_p tais que $\exists\ x \in C \cap D$. Tome $y \in C$. Uma vez que C é classe $x \leftrightarrow y$, isto mais o fato de D ser fechado nos dá que $y \in D$. Assim concluímos que $C \subseteq D$. De modo análogo pode-se mostrar que $D \subseteq C$. Ou seja, $C = D$. □

A seguir, definimos o tempo de primeira passagem e o tempo de parada os quais são importantes na demonstração das propriedades dos estimadores de verossimilhança das probabilidades de transição.

Definição 2.16. Definimos o *Tempo de Primeira Passagem* ao estado $x \in S$ como $T_x^{(1)} = min\{t > 0 : X_t = x\}$ e o tempo de *r-ésima passagem ao estado* $x \in S$ como $T_x^{(r)} = min\{t > T_x^{(r-1)} : X_t = x\}$

A seguir a definição do tempo de parada para uma Cadeia de Markov.

Definição 2.17. Seja $(X_t)_{t\in\mathbb{N}}$ Markov($\boldsymbol{\lambda}$, $\boldsymbol{\mathcal{P}}$). Uma variável aleatória $T : \Omega \longrightarrow \{0, 1, 2\cdots\}$ é um *Tempo de Parada* se o evento $[T = t]$ depende somente de $X_0, X_1, X_2, \cdots, X_t$ para $t = 0, 1, 2, \cdots$.

Para compreendermos melhor este conceito vejamos os seguintes exemplos.

Exemplo 2.1. O *tempo de absorção* $H^A(\omega) = inf\{n \geq 0 : X_n(\omega) \in A\}$, em que $\omega \in \Omega$ e A é um subconjunto do espaço de estados, é um tempo de parada, pois dado $t \in \mathbb{N}$ o evento $[H^A = t]$ depende apenas de $X_0, X_1, X_2, \cdots, X_t$.

Exemplo 2.2. O *tempo de última saída* $S^A(\omega) = max\{n \geq 0 : X_n(\omega) \in A\}$, em que $\omega \in \Omega$ e A é um subconjunto do espaço de estados, não é um tempo de parada, pois dado $t \in \mathbb{N}$ o evento $[H^A = t]$ depende de $X_t, X_{t+1}, X_{t+2}, \cdots$

A seguir definimos o tempo médio de recorrência.

Definição 2.18. *O Tempo Médio* de recorrência é definido como $\mu_x = \sum_{t=0}^{\infty} t p_{xx}^{(t)}$.

Note que $\mu_x \geq 0$. Se além disso, o tempo médio de recorrência for finito temos a seguinte definição.

Definição 2.19. Dizemos que um estado é *Recorrente Positivo* se ele é recorrente e $\mu_x < \infty$.

A seguir definimos estado aperiódico e Cadeia de Markov aperiódica, ver Norris (2004, página 40).

Definição 2.20. Seja $(X_t)_{t\in\mathbb{N}}$ Markov($\boldsymbol{\lambda}$, $\boldsymbol{\mathcal{P}}$) com espaço de estados S. Dizemos que $x \in S$ é um estado aperiódico se $p_{xx}^t > 0$ para todo t suficientemente grande. Se todo estado $x \in S$ for aperiódico, então $(X_t)_{t\in\mathbb{N}}$ será chamada *Cadeia de Markov aperiódica.*

A seguir definimos Cadeia de Markov Ergódica.

Definição 2.21. Uma Cadeia de Markov é dita *Ergódica* se for irredutível, aperiódica e todos os seus estados forem recorrentes positivos.

Para definirmos o limite da probabilidade de ocupação necessitamos inicialmente definir o número de visitas a algum estado $x \in S$.

Definição 2.22. Definimos o *número de visitas ao estado* $x \in S$ em $t \in \mathbb{N}$ passos como

$$N_x(t) = \sum_{k=1}^{t} I(X_k = x),$$

em que

$$I(X_k = x) = \begin{cases} 1, & se\ X_k = x; \\ 0, & se\ X_k \neq x. \end{cases}$$

Seja $(X_t)_{t\in\mathbb{N}}$ uma Cadeia de Markov com espaço de estados S e $x \in S$ um estado qualquer. Segundo Guttorp (1995) o *limite da probabilidade de ocupação* é a proporção do "tempo gasto" no estado x ao longo dos infinitos passos. A seguir definimos o limite da probabilidade de ocupação.

Definição 2.23. O *limite da probabilidade de ocupação* é definido como $\lim_{t\to\infty} \frac{N_x(t)}{t}$.

O Teorema 2.4 a seguir apresenta a Propriedade Forte de Markov, maiores detalhes ver Norris (2004, pág. 20).

Teorema 2.4. *(Propriedade Forte de Markov): Suponha que* $(X_t)_{t\in\mathbb{N}}$ *seja Markov(*$\boldsymbol{\lambda}$*,*$\boldsymbol{\mathcal{P}}$*) e seja* T *um tempo de parada de* $(X_t)_{t\in\mathbb{N}}$*.* $(X_{T+t})_{t\in\mathbb{N}}$*, condicionada em* $[T < \infty]$ *e* $[X_T = x]$*, é Markov(*$\boldsymbol{\delta}_x, \mathbb{P}$*) e independente de* $X_0, X_1, ..., X_t$*.*

Lema 2.1. Para $r = 2, 3, ...$, $T_x^{(r)} - T_x^{(r-1)}$, condicionada em $T_x^{(r-1)} < \infty$ e independente de $\{X_m : m \leq T_x^{(r-1)}\}$ e

$$\mathbb{P}(T_x^{(r)} - T_x^{(r-1)} = t | T_x^{(r-1)} < \infty) = \mathbb{P}_x(T_x = t).$$

Assim,

$$E_x(T_x^{(r)} - T_x^{(r-1)}) = \sum_{t=0}^{\infty} t\mathbb{P}_x(T_x = t) = \sum_{t=0}^{\infty} t p_{xx}^{(t)} = \mu_x.$$

Maiores detalhes ver Norris(2004, página 25).

O teorema a seguir trata da distribuição estacionária de uma Cadeia de Markov irredutível e com estados recorrentes positivos. A sua demonstração pode ser encontrada em Guttorp (1995, pág. 37)

Teorema 2.5. *Uma Cadeia de Markov irredutível possui distribuição estacionária se, e somente se, for recorrente positiva. A distribuição estacionária é única e dada por* $\pi_x = \mu_x^{-1}$.

A seguir apresentamos, o Teorema Ergódico para Cadeia de Markov. Maiores detalhes ver Guttorp (1995).

Teorema 2.6. *(Teorema Ergódico). O limite da probabilidade de ocupação (ver Definição 2.23) de uma cadeia ergódica é* $\frac{1}{\mu_x} = \pi_x$ *(com probabilidade 1).*

Demonstração. Suponha que a cadeia comece em x. Sejam $T_x^{(1)}, T_x^{(2)}, ...$ os sucessivos tempos de passagem da cadeia pelo estado x. Então, pelo Lema 2.1, $T_x^{(1)}, T_x^{(2)} - T_x^{(1)}, T_x^{(3)} - T_x^{(2)}, ...$ são independentes e identicamente distribuídas, com distribuição $p_{xx}^{(t)}$. Além disso, $\mu_x = E(T_x^{(r)} - T_x^{(r-1)}) = \sum_{t=0}^{\infty} t p_{xx}^{(t)} < \infty$, pois todos os estados são recorrentes positivos. Pela Lei Forte dos Grandes Números temos, com probabilidade 1, que

$$\lim_{r\to\infty} \frac{T_x^{(1)} + (T_x^{(2)} - T_x^{(1)}) + (T_x^{(3)} - T_x^{(2)}) + ... + (T_x^{(r)} - T_x^{(r-1)})}{r} = \lim_{r\to\infty} \frac{T_x^{(r)}}{r} = \mu_x,$$

por outro lado também vale

$$T_x^{(N_x(t))} \leq t \leq T_x^{(N_x(t)+1)}.$$

Além disso, $N_x(t) \longrightarrow \infty$ quando $t \longrightarrow \infty$ uma vez que o estado x é revisitado infinitas vezes, então

$$\frac{N_x(t)}{t} \leq \frac{N_x(t)}{T_x^{(N_x(t))}} \longrightarrow \frac{1}{\mu_x} \tag{2.2}$$

com probabilidade 1, e

$$\frac{N_x(t)+1}{t} \geq \frac{N_x(t)+1}{T_x^{(N_x(t)+1)}} \longrightarrow \frac{1}{\mu_x} \tag{2.3}$$

com probabilidade 1.

Assim, de (2.2) e (2.3), temos

$$\frac{N_x(t)}{t} \longrightarrow \frac{1}{\mu_x}.$$

Pelo Teorema 2.5 temos que $\frac{1}{\mu_x} = \pi_x$. □

2.2 Verossimilhança para Cadeias de Markov

Nesta seção iremos estabelecer alguns conceitos sobre Inferência Estocástica isto é, a utilização de conceitos da Inferência Estatística em Processos Estocásticos. Inicialmente estabeleceremos a definição de *função de verossimilhança* para variáveis aleatórias independentes e identicamente distribuídas (ver Rohatgi, 1976).

Posteriormente sua definição será modificada para o caso de dependência encontrado em uma Cadeia de Markov. Isto é, na Cadeia de Markov sabe-se que "o futuro depende do presente" tanto

esta relação de dependência quanto a *Propriedade markoviana* serão levadas em consideração na definição da função de verossimilhança *para a Cadeia de Markov.*

A partir da função de verossimilhança *para a Cadeia de Markov* obteremos, com base em Atuncar (2009), capítulo 3 o Estimador de Máxima Verossimilhança para uma Cadeia de Markov ergódica de primeira ordem. Cabe a observação de que a "Cadeia de Markov Fortemente Ergódica" (ver Atuncar 2009, página 3) neste trabalho é denominada "Cadeia de Markov Ergódica".

Além disso apresentaremos a consistência do Estimador de Máxima Verossimilhança para a Cadeia de Markov e o fim desta seção será discutido o seu comportamento assintótico.

A seguir definimos a função de verossimilhança para variáveis aleatórias independentes.

Definição 2.24. Seja $(X_t)_{t=1}^n$ uma amostra aleatória, isto é, variáveis aleatória independentes e identicamente distribuídas com função massa de probabilidade ou função densidade de probabilidade $f(x_t, \boldsymbol{\theta})$ $t = 1, \cdots, n$ em que $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$. Definimos a *Função de Verossimilhança* $L(\boldsymbol{\theta}, \boldsymbol{x})$ como

$$L(\boldsymbol{\theta}, \boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{t=1}^{n} f_{X_t}(x_t, \boldsymbol{\theta}).$$

Note que a *função de verossimilhança* é uma função de $\boldsymbol{\theta}$ supondo $\boldsymbol{x} = (x_1, \cdots, x_n)$ conhecido. O *Estimador de Máxima Verossimilhança* será o valor $\hat{\boldsymbol{\theta}}(\mathbf{x})$ que maximiza a função $L(\boldsymbol{\theta}, \boldsymbol{x})$.

Deste ponto em diante consideramos $(X_t)_{t\in\mathbb{N}}$ uma Cadeia de Markov com espaço de estados $S = \{0, 1, \cdots, M\}$, $M \in \mathbb{N}$ finito. Observe que

$$\mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0) = \mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}).$$

Assim não poderemos utilizar diretamente a Definição 2.24, uma vez que nesta definição a função é obtida a partir de uma amostra aleatória.

A seguir apresentaremos a definição de *função de verossimilhança* para Cadeia de Markov segundo Atuncar (2009, capítulo 3).

Considere $\boldsymbol{x} = (x_0, x_1, \cdots, x_t) \in S$ uma realização de tamanho $t+1$, $t \in \mathbb{N}$ finito, de uma Cadeia de Markov. Definimos $\boldsymbol{\theta} = (p_{00}, \cdots, p_{MM})$, a função de verossimilhança, denotada $L(\boldsymbol{\theta}, \boldsymbol{x})$, é dada por

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{x}) &= L((p_{00}, ..., p_{MM}), (x_1, ..., x_t)) = \mathbb{P}(X_0 = x_0, X_1 = x_1, X_2 = x_2, ..., X_t = x_t) \\ &= \mathbb{P}(X_0 = x_0).\mathbb{P}(X_1 = x_1 | X_0 = x_0)...\mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}, ..., X_0 = x_0) \\ &= \mathbb{P}(X_0 = x_0).\mathbb{P}(X_1 = x_1 | X_0 = x_0)...\mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}) \\ &= \lambda_{x_0}.p_{x_0x_1}.p_{x_1x_2}.p_{x_2x_3}...p_{x_{t-1}x_t} = \lambda_{x_0} \prod_{x,y=0}^{M} p_{xy}^{t_{xy}}, \end{aligned} \tag{2.4}$$

em que $\lambda_{x_0} = \mathbb{P}(X_0 = x_0)$, p_{xy} representa a probabilidade de transição do estado x para o estado y, em um passo, e t_{xy} representa o número de transições do estado x para o estado y, em um passo.

A jgualdade (2.4) é válida. Para garantirmos isto precisamos observar primeiramente que, fixados x, y consecutivos, há um agrupamento dos fatores p_{xy} que aparecem do lado esquerdo,

isto é mantemos a base (comum) e somamos os expoentes. Esta soma é exatamente o valor t_{xy}. Agora, fixando a nossa atenção para o lado direito da mesma igualdade, é possível observar que se a escolha, no produtório, for de valores x, y sequenciais que não aparecem na sequência amostral. Então $t_{xy} = 0$, e, consequentemente, $p_{xy}^{t_{xy}} = p_{xy}^0 = 1$. É importante observar também que p_{xy} não pode ser nulo. De fato isto será garantido ao definirmos, mais adiante, o espaço de estados observados.

O *Estimador de Máxima Verossimilhança* para $\boldsymbol{\theta}$ é definido como $\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{\boldsymbol{\theta}}{max} L(\boldsymbol{\theta}, \boldsymbol{x})$. Note que maximizar $L(\boldsymbol{\theta}, \boldsymbol{x})$ é o mesmo que maximizar $\pounds(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{x}) = \log L(\boldsymbol{\theta}, \boldsymbol{x})$. Aplicando o logaritmo na expressão que aparece do lado direito da igualdade (2.4) temos

$$\pounds(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{x}) = \log L(\boldsymbol{\theta}, \boldsymbol{x}) = \log(\lambda_{x_0}) + \sum_{x,y=0}^{M} t_{xy} \log(p_{xy}),$$

para tal utilizaremos o método de Lagrange (ver observação 2.1).

Observação 2.1. O método de Lagrange baseia-se na igualdade entre o gradiente $\nabla f(\cdot)$ de uma função f, com o gradiente $\nabla g(\cdot)$ de sua função restrição g, multiplicada por um fator real α. Ou seja: um ponto P é um ponto crítico de f se $\exists\ \alpha$ real tal que

$$\nabla f(P) = \alpha \nabla g(P).$$

Queremos estimar os $(M+1)^2$ parâmetros

$$\boldsymbol{\theta} = \begin{pmatrix} p_{00} & & p_{0M} \\ & \ddots & \\ p_{M0} & & p_{MM} \end{pmatrix}$$

da função $\pounds(\boldsymbol{\theta}, \boldsymbol{x})$.

Isto é, temos uma função $\pounds(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{x})$, de $(M+1)^2$ variáveis, a qual pode ser maximizada para cada $\boldsymbol{x}$, pelo Método de Lagrange, usando como restrição $\sum_{y=0}^{M} p_{xy} - 1 = 0$. Assim queremos encontrar μ tal que para todo $y \in \{0, 1, \cdots, M\}$,

$$\frac{\partial \pounds(\boldsymbol{\theta}, \boldsymbol{x})}{\partial p_{xy}} = \frac{\partial(\sum_{y=0}^{M} p_{xy} - 1)}{\partial p_{xy}}, \text{ ou seja } \frac{\partial \pounds(\boldsymbol{\theta}, \boldsymbol{x})}{\partial p_{xy}} = \alpha 1.$$

Assim $\alpha = t_{xy}\dfrac{1}{p_{xy}}$, logo $\alpha p_{xy} = t_{xy}$, então $\displaystyle\sum_{y=0}^{M} \alpha p_{xy} = \sum_{y=0}^{M} t_{xy}$, ou seja $\alpha \displaystyle\sum_{y=0}^{M} p_{xy} = t_x$, desta forma $\alpha = t_x$,

em que t_x representa o número de visitas ao estado x.

Sendo S o espaço de estados, $\hat{S} = \{x \in S : t_x \geq 1\}$ é chamado espaço de estados observado. Se $x, y \in \hat{S}$, então o Estimador de Máxima Verossimilhança, denotado por $\hat{p}_{xy}$, é dado por

$$t_x \hat{p}_{xy} = t_{xy}, \text{ ou seja } \hat{p}_{xy} = \frac{t_{xy}}{t_x}. \tag{2.5}$$

Note que se o espaço de estados é finito, então $\hat{S}$ também o é.

Exemplo 2.3. (Zuchini e Macdonald 2009, página 20)

Considere uma Cadeia de Markov com espaço de estados $S = \{1, 2, 3\}$ e com uma realização (caminho) 2332111112 3132332122 3232332222 3132332212 3232132232 3132332223 3232331232 3232331222 3232132123 3132332121.

Para esta realização temos:

$$\begin{aligned} &t_1 = 18,\ t_2 = 41,\ t_3 = 40,\\ &t_{11} = 4,\ t_{12} = 7,\ t_{13} = 6,\\ &t_{21} = 8,\ t_{22} = 10,\ t_{23} = 24,\\ &t_{31} = 6,\ t_{32} = 24,\ t_{33} = 10. \end{aligned}$$

Os estimadores das probabilidades de transição são dados por:

$$\begin{aligned} &\hat{p}_{11} = \frac{t_{11}}{t_1 - 1},\ \hat{p}_{12} = \frac{t_{12}}{t_1 - 1},\ \hat{p}_{13} = \frac{t_{13}}{t_1 - 1},\\ &\hat{p}_{21} = \frac{t_{21}}{t_2},\ \hat{p}_{22} = \frac{t_{22}}{t_2},\ \hat{p}_{23} = \frac{t_{23}}{t_3},\\ &\hat{p}_{31} = \frac{t_{31}}{t_3},\ \hat{p}_{32} = \frac{t_{32}}{t_3},\ \hat{p}_{33} = \frac{t_{33}}{t_3}. \end{aligned}$$

Note que para os estimadores das probabilidades de transição $\hat{p}_{1j}$, j=1, 2, 3 diminui-se em uma unidade o denominador, fizemos isso para garantir que $p_{11} + p_{12} + p_{13} = 1$, esta técnica é aplicada para o último estado da realização. Para uma realização de tamanho suficientemente grande é indiferente usar t_x ou $t_x - 1$ no denominador do último estado x.

Logo os valores estimados para as probabilidades de transição são dados por:

$$\begin{aligned} &\hat{p}_{11} = \frac{4}{17},\ \hat{p}_{12} = \frac{7}{17},\ \hat{p}_{13} = \frac{6}{17},\\ &\hat{p}_{21} = \frac{8}{42}, \hat{p}_{22} = \frac{10}{42}, \hat{p}_{23} = \frac{24}{42},\\ &\hat{p}_{31} = \frac{6}{40},\ \hat{p}_{32} = \frac{24}{40},\ \hat{p}_{33} = \frac{10}{40}. \end{aligned}$$

Estabelecidos os estimadores para as probabilidades de transição, o próximo passo é verificar as propriedades de consistência e normalidade assintótica dos mesmos.

Proposição 2.1. Seja $(X_t)_{t\in\mathbb{N}}$ uma Cadeia de Markov com parte observada $\hat{S}$ e matriz de transição estimada por $\hat{\boldsymbol{\mathcal{P}}}$ na qual os elementos $\hat{p}_{xy}$ são dados pela expressão 2.5 da página 13. $(X_t)_{t\in\mathbb{N}}$ tem uma classe de estados transientes, e precisamente uma classe fechada $\hat{S}_p$ de estados recorrentes. Aqui $\hat{\boldsymbol{\mathcal{P}}}$ representa a "matriz de transição" formada pelos estimadores acima.

Demonstração. Seja $m \in \mathbb{N}$. Se $x_m \in \hat{S}$ e $x_{m'} \in \hat{S}$, então $t_{x_m} \geq 1$ e $t_{x_{m'}} \geq 1$, sendo assim se $m < m'$ saindo de x_m, em algum momento, chegaremos em $x_{m'}$. Ou seja: $x_m \to x_{m'}$ sempre que $m < m'$. Seja x_{m_0} o primeiro estado recorrente. Construa $\hat{S}_p$ da seguinte forma: todo x_m tal que $m < m_0$, não pertence ao conjunto $\hat{S}_p$. Como x_{m_0} é recorrente, $p^t_{x_{m_0}x_{m_0}} > 0$ para

uma infinidade de $t's$. Por esta razão e pelo que foi discutido no início deste parágrafo temos que $x_{m_0} \leftrightarrow x_{m_0+k}$, para todo inteiro não negativo k. Então, definindo $\hat{S}_p = \{x_{m_0}, x_{m_{0+1}}, ..., x_t\}$ temos que $\hat{S}_p$ é fechado e também irredutível, assim pelo teorema 2.3 da página 9 temos que $S_T = S - \hat{S}_p$. □

A seguir definimos a *Cadeia de Passos*. O Lema 2.1 estabelece sua ergodicidade. Estes resultados serão importantes na demonstração do Teorema 2.7 da página 17.

Definição 2.25. Seja $(X_t)_{t\in\mathbb{N}}$ uma Cadeia de Markov. Defina $(Y_t)_{t\in\mathbb{N}}$, a *Cadeia de Passos,* por $Y_t = (X_t, X_{t+1})$.

Lema 2.1. A Cadeia de Passos (ver Definição 2.25) $(Y_t)_{t\in\mathbb{N}}$ é uma Cadeia de Markov com espaço de estados

$$\tilde{S} = \{(x, y) \in S^2 : p_{xy} > 0\}$$

distribuição inicial $\boldsymbol{\tilde{p}_0}$ dada por

$$\tilde{p}_0(x, y) = p_0(x)p_{xy}. \tag{2.6}$$

As entradas da matriz de transição $\tilde{\boldsymbol{\mathcal{P}}} = (\tilde{p}_{wz,xy})$ são dadas por $\tilde{p}_{wz,xy} = I(x = z)p_{xy}$. Além disso se $(X_t)_{t\in\mathbb{N}}$ é ergódica e $\mathbf{p_0} = \boldsymbol{\pi}$, então $(Y_t)_{t\in\mathbb{N}}$ é também ergódica com distribuição estacionária inicial $\tilde{\boldsymbol{\pi}}$ dada por $\tilde{\pi}(x, y) = \pi_x p_{xy}$.

Demonstração. Temos que verificar que

$$\mathbb{P}(Y_t = (x, y)|Y_{t-1} = (w_1, z_1), ..., Y_0 = (w_t, z_t)) = I(x = z_1)p_{xy} = \mathbb{P}(Y_t = (x, y)|Y_{t-1} = (w_1, z_1)),$$

ou seja, devemos mostrar que a Cadeia de Passos satisfaz a propriedade markoviana. Primeiramente note que

$$\begin{aligned}
&\mathbb{P}(Y_t = (x, y)|Y_{t-1} = (w_1, z_1), ..., Y_0 = (w_t, z_t)) \\
&= \mathbb{P}(X_{t+1} = y, X_t = x|X_t = z_1, X_{t-1} = w_1, X_{t-1} = z_2, \cdots, X_0 = w_t) \\
&= \mathbb{P}(X_{t+1} = y|X_t = x, X_t = z_1, X_{t-1} = w_1, \cdots, X_0 = w_t)\mathbb{P}(X_t = x|X_t = z_1, X_{t-1} = w_1, \cdots, X_0 = w_t) \\
&= \mathbb{P}(X_{t+1} = y|X_t = x, X_t = z_1)\mathbb{P}(X_t = x|X_t = z_1, X_{t-1} = w_1, X_{t-1} = z_2) \\
&= I(x = z_1)\mathbb{P}(X_{t+1} = y|X_t = x) = I(x = z_1)p_{xy}.
\end{aligned}$$

Por outro lado temos que

$$\begin{aligned}
&\mathbb{P}(Y_t = (x, y)|Y_{t-1} = (w_1, z_1)) \\
&= \mathbb{P}(X_{t+1} = y, X_t = x|X_t = z_1, X_{t-1} = w_1) \\
&= \mathbb{P}(X_{t+1} = y|X_t = x, X_t = z_1, X_{t-1} = w_1)\mathbb{P}(X_t = x|X_t = z_1, X_{t-1} = w_1) \\
&= \mathbb{P}(X_{t+1} = y|X_t = x, X_t = z_1)\mathbb{P}(X_t = x|X_t = z_1, X_{t-1} = w_1) \\
&= I(x = z_1)\mathbb{P}(X_{t+1} = y|X_t = x) = I(x = z_1)p_{xy}.
\end{aligned}$$

Note que a combinação dos dois resultados acima nos garante a propriedade markoviana. Para obtermos a distribuição inicial da *Cadeia de Passos* percebemos que

$$\begin{aligned}\tilde{p}_0(x,y) = \mathbb{P}(Y_0 = (x,y)) &= \mathbb{P}(X_0 = x, X_1 = y) \\ &= \mathbb{P}(X_0 = x)\mathbb{P}(X_1 = y | X_0 = x) = p_0(x)p_{xy}.\end{aligned}$$

Portanto a *Cadeia de Passos* satisfaz a propriedade markoviana e

$$\tilde{p}_0(x,y) = p_0(x)p_{xy}, \text{ para todo } (x,y) \in \tilde{S}.$$

Supondo $(X_t)_{t\in\mathbb{N}}$ uma Cadeia de Markov ergódica com $\mathbf{p_0} = \boldsymbol{\pi}$, temos que $p_0(x) = \mathbb{P}(X_0 = x) = \pi_x$. Assim, sejam $(x,y) \in \tilde{S}$ quaisquer e denote $\tilde{\pi}(x,y) = \mathbb{P}(Y_0 = (x,y))$ a distribuição inicial da Cadeia de Passos $(Y_t)_{t\in\mathbb{N}}$, então

$$\tilde{\pi}(x,y) = \mathbb{P}(Y_0 = (x,y)) = \mathbb{P}(X_0 = x, X_1 = y) = \mathbb{P}(X_0 = x)\mathbb{P}(X_1 = y | X_0 = x) = \pi_x p_{xy}.$$

Verificamos que se $\tilde{\pi}_{(x,y)} = \pi_x p_{xy}$ é a distribuição estacionária da Cadeia de Passos $(Y_t)_{t\in\mathbb{N}}$. Para isso, temos que verificar se

$$\sum_{w\in S}\sum_{z\in S} \tilde{\pi}(w,z)\tilde{p}_{wz,xy} = \tilde{\pi}(x,y) \ \forall \ (x,y) \in \tilde{S}.$$

De fato,

$$\begin{aligned}\sum_{w\in S}\sum_{z\in S} \tilde{\pi}(w,z)\tilde{p}_{wz,xy} &= \sum_{w\in S}\sum_{z\in S} \pi(w)p_{wz}I(x=z)p_{xy} \\ &= \sum_{w\in S} \pi(w)p_{wz} \sum_{z\in S} I(x=z)p_{xy} \\ &= \pi_z \sum_{z\in S} I(x=z)p_{xy} \\ &= \sum_{z\in S} I(x=z)\pi_z p_{xy}\pi_x p_{xy} = \tilde{\pi}(x,y).\end{aligned}$$

Além disso,

$$\sum_{y\in S}\sum_{x\in S} \tilde{\pi}(x,y) = \sum_{y\in S}\sum_{x\in S} \pi_x p_{xy} = \sum_{y\in S} \pi_y = 1.$$

□

A seguir utilizando a definição de *Cadeia de Passos* e o Teorema Ergódico, será demonstrado a consistência forte para o estimador $\hat{\boldsymbol{\mathcal{P}}}$. A notação $\hat{p}_{xy}(t)$ representará o estimador de máxima verossimilhança para a probabilidade de transição p_{xy} baseada em uma realização de tamanho t. Além disso utilizaremos a seguinte notação alternativa

$$\hat{p}_{xy} = \frac{t_{xy}}{t_x} = \frac{N_{xy}(t)}{N_x(t)}$$

para a equação 2.5 da página 13. Onde

$$N_x(t) = \sum_{k=1}^{t} I(X_k = x), \text{ e } N_{xy}(t) = \sum_{k=1}^{t} I(X_k = x)I(X_{k+1} = y)$$

Teorema 2.7. *Seja $(X_t)_{t\in\mathbb{N}}$ uma Cadeia de Markov ergódica e S seu espaço de estados. Então, $\hat{p}_{xy} \longrightarrow p_{xy}$ com probabilidade 1, quando $t \longrightarrow \infty$, para todo $x, y \in S$, independentemente de sua distribuição inicial.*

Demonstração. Sejam $x, y \in S$ tais que $p_{xy} = \mathbb{P}(X_{t-1} = x|X_t = y) = 0$. Então t_{xy}, que representa o número de transições do estado x para o estado y, em um passo, é zero. Logo $\mathbb{P}(\hat{p}_{xy}(t) = 0) = 1$. Portanto, necessitamos considerar $(x,\, y) \in \tilde{S}$. Seja

$$N_{xy}(t) = \sum_{k=1}^{t} I(Y_k = (x, y)),$$

o número de transições do estado x para o estado y numa realização de tamanho t.

Pelo Lema 2.1 e pelo Teorema Ergódico (Teorema 2.6) temos que $(Y_t)_{t\in\mathbb{N}}$ é ergódica e

$$\frac{1}{t}N_{xy}(t) \longrightarrow \tilde{\pi}(x, y) = \pi_x p_{xy},$$

com probabilidade 1. Ou seja,

$$\lim_{t\to\infty} \frac{1}{t}N_{xy}(t) = \pi_x p_{xy}, \text{ sendo assim } \sum_{y\in S}(\lim_{t\to\infty} \frac{N_{xy}(t)}{t}) = \sum_{y\in S} \pi_x p_{xy}$$

$$, \text{ logo } \lim_{t\to\infty} \sum_{y\in S} \frac{N_{xy}(t)}{t} = \pi_x, \text{ ou seja } \lim_{t\to\infty} \frac{N_x(t)}{t} = \pi_x.$$

E consequentemente

$$\lim_{t\to\infty} \frac{N_{xy}(t)}{t}\frac{t}{N_x(t)} = \pi_x p_{xy}\frac{1}{\pi_x}, \text{ isto é } \lim_{t\to\infty} \frac{N_{xy}(t)}{N_x(t)} = p_{xy}, \text{ então } \lim_{t\to\infty} \hat{p}_{xy} = p_{xy}$$

com probabilidade 1. □

A seguir, apresentamos a Definição de uma variável aleatória $W_x^{(m)}$ baseada no m-ésimo tempo de retorno a um estado $x \in S$. A partir de $W_x^{(m)}$ será construída $Q_{xy}(t)$. A partir de $Q_{xy}(t)$ será construído o vetor $Q_x(t)$. O Lema 2.2 trata da distribuição de tais vetores, bem como sua relação de independência. Maiores detalhes ver Guttorp(1995, pág. 62 e 63).

Definição 2.26. Seja $W_x^{(m)} = X_{1+T_x^{(m)}}$, $m \in \mathbb{N}$, a entrada do primeiro estado após o m-ésimo retorno ao estado x, e escrevemos

$$Q_{xy}(t) = \sum_{m=1}^{[t\pi_x]} I(W_x^{(m)} = y)$$

e finalmente defina o vetor

$$Q_x(t) = (Q_{xy}(t), y \in S), \text{ para cada } x \in S.$$

Para que esta definição fique mais clara. Consideremos o seguinte exemplo.

Exemplo 2.4. Considere a seguinte realização de uma Cadeia de Markov a três estados, ver a tabela a seguir, com distribuição estacionária dada por $\pi = \left(\frac{1}{3}, \frac{1}{4}, \frac{5}{12}\right)$ · Encontre $Q_{12}(21)$.

1	2	2	3	1	1	2	3	3	3	1	1	1	2	2	2	3	3	3	3	1	1
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

Tabela 2.1: realização de uma Cadeia de Markov a três estados, a primeira linha representa a realização da Cadeia e a segunda linha o tempo no qual cada estado ocorre

Utilizando $W_x^{(m)} = X_{1+T_x^{(m)}}$, $m \in \mathbb{N}$ e $Q_{xy}(t) = \sum_{m=1}^{[t\pi_x]} I(W_x^{(m)} = y)$ temos que

$$\begin{aligned}
Q_{12}(21) &= \sum_{m=1}^{[21\frac{1}{3}]} I(W_1^{(m)} = 2) = \sum_{m=1}^{7} I(W_1^{(m)} = 2) \\
&= I(W_1^{(1)} = 2) + I(W_1^{(2)} = 2) + I(W_1^{(3)} = 2) + I(W_1^{(4)} = 2) \\
&\quad + I(W_1^{(5)} = 2) + I(W_1^{(6)} = 2) + I(W_1^{(7)} = 2) \\
&= I(X_{1+T_1^{(1)}} = 2) + I(X_{1+T_1^{(2)}} = 2) + I(X_{1+T_1^{(3)}} = 2) + I(X_{1+T_1^{(4)}} = 2) \\
&\quad + I(X_{1+T_1^{(5)}} = 2) + I(X_{1+T_1^{(6)}} = 2) + I(X_{1+T_1^{(7)}} = 2) \\
&= I(X_5 = 2) + I(X_6 = 2) + I(X_{11} = 2) + I(X_{12} = 2) \\
&\quad + I(X_{13} = 2) + I(X_{21} = 2) + I(X_{22} = 2) \\
&= 0 + 1 + 0 + 0 + 1 + 0 + 0.
\end{aligned}$$

Assim $Q_{12}(21) = 2$.

O lema 2.2 a seguir trata da independência e da distribuição dos vetores $Q_x(t)$, para $x \in S$.

Lema 2.2. Os vetores $Q_x(t)$, em que $x \in S$ são independentes tendo distribuição multinomial com amostra de tamanho $[t\pi_x]$.

Demonstração. Primeiramente verificamos que $W_x^{(m)}$ são independentes com $\mathbb{P}(W_x^{(m)} = y) = p_{xy}$. Pela Propriedade Forte de Markov, temos que

$$\mathbb{P}(X_{1+T_x^{(m)}} = y | T_x^{(m)} < \infty) = \mathbb{P}(X_{1+T_x^{(m)}} = y | X_{T_x^{(m)}} = x, T_x^{(m)} < \infty) = p_{xy}.$$

Também, pela Propriedade Forte de Markov, $(X_{t+T_x^{(m)}})_{t\in\mathbb{N}}$ é independente de $X_0, \cdots, X_{T_x^{(m)}}$(ver Teorema 2.4). Assim, temos que $W_x^{(m)}$, com $x \in S$, são independentes. Note que, definido Q_x temos que, para algum $y \in S$, $I(W_x^{(m)} = y) = 1$.

Vale também: $\mathbb{P}(I(W_x^{(m)} = y) = 1) = \mathbb{P}(W_x^{(m)} = y) = p_{xy}$. Consequentemente,

$$\mathbb{P}(\sum_{m=1}^{[t\pi_x]} I_{y_1}(W_x^{(m)}) = k_1) = \frac{[t\pi_x]!}{([t\pi_x] - k_1)!k_1!} p_{xy_1}^{k_1},$$

onde k_1 é um natural não nulo menor que $[t\pi_x]$.

Analogamente,

$$\mathbb{P}(\sum_{m=1}^{[t\pi_x]} I_{y_1}(W_x^{(m)}) = k_1, \sum_{m=1}^{[t\pi_x]} I_{y_2}(W_x^{(m)}) = k_2)$$

$$= \frac{[t\pi_x]!}{([t\pi_x]-k_1)!k_1!} p_{xy_1}^{k_1} \frac{([t\pi_x]-k_1)!}{([t\pi_x]-k_1-k_2)!k_2!} p_{xy_2}^{k_2} = \frac{[t\pi_x]! p_{xy_1}^{k_1} p_{xy_2}^{k_2}}{([t\pi_x]-k_1-k_2)!k_1!k_2!}.$$

Como estamos trabalhando com realizações de tamanho t finito, $[t\pi_x]$ é finito. Portanto, repetindo o procedimento anterior uma quantidade finita de vezes obtemos

$$\mathbb{P}(Q_x(t) = (k_1, \cdots, k_n)) = \frac{[t\pi_x]! p_{xy_1}^{k_1} ... p_{xy_n}^{k_n}}{k_1! ... k_n!}, \text{ onde } \sum_{l=1}^{n} k_l = [t\pi_x] \text{ e } \sum_{y\in S} p_{xy} = 1.$$

O que mostra que $Q_x(t)$ tem distribuição multinomial. O Teorema 2.4 garante a independência entre as variáveis $X_{1+T_x^{(m)}}$, $m \in \mathbb{N}$. Como $Q_x(t)$, $x \in S$ são funções de $X_{1+T_x^{(m)}}$, $m \in \mathbb{N}$ e $X_{1+T_x^{(m)}}$, $m \in \mathbb{N}$ são independentes entre si, então $Q_x(t)$, $x \in S$ são independentes entre si. □

O Teorema 2.8 a seguir, apresenta a normalidade assintótica dos estimadores de máxima verossimilhança das probabilidades de transição de uma Cadeia de Markov.

Teorema 2.8. *Seja $(X_t)_{t\in\mathbb{N}}$ uma Cadeia de Markov Ergódica. Então, independente da distribuição inicial, para todo $x, y \in S$,*

$$[N_x(t)]^{\frac{1}{2}}(\hat{p}_{xy}(t) - p_{xy}) \xrightarrow{d} \mathcal{N}(0, p_{xy}(1-p_{xy}))$$

Demonstração. Uma vez que $\frac{t\pi_x}{N_x(t)} \longrightarrow 1$, temos que

$$[[N_x(t)]^{\frac{1}{2}}(\frac{N_{xy}(t)}{N_x(t)} - p_{xy})] = [[N_x(t)]^{\frac{-1}{2}}(N_{xy}(t) - p_{xy}N_x(t))] =$$

$$[\frac{t^{\frac{1}{2}}\pi_x^{\frac{1}{2}}}{N_x(t)^{\frac{1}{2}}}(\frac{N_{xy}(t) - p_{xy}.N_x(t)}{(t\pi_x)^{\frac{1}{2}}})] \longrightarrow [\frac{N_{xy}(t) - p_{xy}.N_x(t)}{(t\pi_x)^{\frac{1}{2}}}].$$

Assim basta mostrar que

$$\frac{N_{xy}(t) - p_{xy}N_x(t)}{(t\pi_x)^{\frac{1}{2}}} \xrightarrow{d} \mathcal{N}(0, p_{xy}(1-p_{xy})). \tag{2.7}$$

Assumiremos as seguintes aproximações e posteriormente mostraremos que, de fato, elas são adequadas.

$$N_{xy}(t) = \sum_{k=0}^{t} I(X_k = x)I(X_{k+1} = y) \simeq \sum_{m=1}^{[t\pi_x]} I(W_x^m = y) = Q_{xy}(t)$$

$$N_x(t) = \sum_{k=0}^{t} I(X_k = x) = \sum_{y\in S} N_{xy}(t) \simeq \sum_{y\in S} Q_{xy}(t) = [t\pi_x].$$

Assim trocando $N_{xy}(t)$ por $Q_{xy}(t)$ e $N_x(t)$ por $[t\pi_x]$ em (2.7) temos que mostrar

$$\frac{Q_{xy}(t) - p_{xy}[t\pi_x]}{(t\pi_x)^{\frac{1}{2}}} \xrightarrow{d} \mathcal{N}(0, p_{xy}(1-p_{xy})).$$

Para mostrar esta convergência usamos o *Teorema Central do Limite* (ver James 2008, página 241). Uma vez que $Q_{xy}(t)$ é a soma de binomiais independentes, $E[Q_{xy}(t)] = [t\pi_x]p_{xy}$ e $Var[Q_{xy}(t)] = [t\pi_x]p_{xy}(1-p_{xy})$, o *Teorema Central do Limite* nos garante que

$$\frac{Q_{xy}(t) - E[Q_{xy}(t)]}{\sqrt{Var[Q_{xy}(t)]}} \xrightarrow{d} \mathcal{N}(0,1).$$

Então,

$$\frac{Q_{xy}(t) - [t\pi_x]p_{xy}}{\sqrt{[t\pi_x]p_{xy}(1-p_{xy})}} \xrightarrow{d} \mathcal{N}(0,1).$$

Equivalentemente

$$\sqrt{p_{xy}(1-p_{xy})}\frac{Q_{xy}(t) - [t\pi_x]p_{xy}}{\sqrt{[t\pi_x]p_{xy}(1-p_{xy})}} \xrightarrow{d} \mathcal{N}(0, p_{xy}(1-p_{xy})).$$

Dessa forma

$$\frac{Q_{xy}(t) - [t\pi_x]p_{xy}}{\sqrt{[t\pi_x]}} \xrightarrow{d} \mathcal{N}(0, p_{xy}(1-p_{xy})).$$

Precisamos mostrar ainda que estas aproximações são adequadas, no sentido de que

$$D_t = (t\pi_x)^{\frac{-1}{2}}(N_{xy}(t) - N_x(t)p_{xy} - Q_{xy}(t) + [t\pi_x]p_{xy}) \xrightarrow{\mathbb{P}} 0.$$

Para $x, y \in S$ fixados, seja $Z_m = I(W_x^{(m)} = y) - p_{xy}$, $m \in \mathbb{N}$, e seja $S_t = \sum_{m=1}^{t} Z_m$. As Z_m's são iid (ver Guttorp 1995, página 63) e a média de cada uma delas é dada por

$$\begin{aligned} E(Z_m) &= E(I(W_x^{(m)} = y) - p_{xy}) = E(I(W_x^{(m)} = y)) - E(p_{xy}) \\ &= \mathbb{P}[W_x^{(m)} = y] - p_{xy} = \mathbb{P}[X_{T_x(m)+1} = y] - p_{xy} \\ &= \mathbb{P}[X_{T_x^{(m)}+1} = y | X_{T_x^{(m)}} = x] - p_{xy} = p_{xy} - p_{xy} = 0. \end{aligned}$$

Assim, concluímos que as Z_m's são iid com média zero. Chame $\sigma^2 = Var(Z_x)$ e $K = E(Z_x)^4$. Também temos

$$S_{N_x(t)} = \sum_{m=1}^{N_x(t)} Z_m, \text{ isto é } S_{N_x(t)} = \sum_{m=1}^{N_x(t)} (I(W_x^{(m)} = j) - p_{xy})$$

$$S_{[t\pi_x]} = \sum_{m=1}^{[t\pi x]} Z_m, \text{ ou seja } S_{[t\pi_x]} = \sum_{m=1}^{[t\pi x]} (I(W_x^{(m)} = y) - p_{xy}).$$

Então,

$$D_t = (t\pi_x)^{\frac{-1}{2}}(N_{xy}(t) - N_x(t)p_{xy} - Q_{xy}(t) + [t\pi_x]p_{xy}) =$$

$$(t\pi_x)^{\frac{-1}{2}}(\sum_{m=1}^{N_x(t)} I(W_x^{(m)} = y) - \sum_{m=1}^{N_x(t)} p_{xy} - (\sum_{m=1}^{[t\pi_x]} I(W_x^{(m)} = y) - \sum_{m=1}^{[t\pi_x]} p_{xy})) =$$

$$(t\pi_x)^{\frac{-1}{2}}(\sum_{m=1}^{N_x(t)} (I(W_x^{(m)} = y) - p_{xy}) - \sum_{m=1}^{[t\pi_x]} (I(W_x^{(m)} = y) - p_{xy})) =$$

$$(t\pi_x)^{\frac{-1}{2}}(S_{N_x(t)} - S_{[t\pi_x]}).$$

Dado $r \in \mathbb{R}$ temos, pela *Lei de* , (ver James 2008, página 17):

$$\mathbb{P}(|D_t| > \epsilon) = \mathbb{P}(|D_t| > \epsilon, |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}) + \mathbb{P}(|D_t| > \epsilon, |N_x(t) - [t\pi_x]| > rt^{\frac{1}{2}}) \leq$$

$$\mathbb{P}(|D_t| > \epsilon, |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}) + \mathbb{P}(|N_x(t) - [t\pi_x]| > rt^{\frac{1}{2}}),$$

no qual, a última desigualdade é válida, pois para quaisquer eventos A e B temos

$$\mathbb{P}(A \cap B) \leq \mathbb{P}(A).$$

Note que

$$\mathbb{P}(|D_t| > \epsilon, |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}) = \mathbb{P}(|D_t| > \epsilon \mid |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}})\mathbb{P}(|N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}})$$

Analisemos $\mathbb{P}(|D_t| > \epsilon \mid |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}})$.

$$\mathbb{P}(|D_t| > \epsilon \mid |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}) = \mathbb{P}(|D_t|^4 > \epsilon^4 \mid |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}) \leq$$

$$\frac{1}{\epsilon^4} E(|D_t|^4 \mid |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}).$$

A relação acima é válida pela *Desigualdade de Tchebychev*, James (2008, página 125)
Além disso, pela definição de D_t

$$\frac{1}{\epsilon^4} E(|D_t|^4 \mid |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}) = \frac{1}{\epsilon^4} E((S_{N_x(t)} - S_{[t\pi_x]})^4 \mid |N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}).$$

Uma vez que $S_m - S_{[t\pi_x]}$ é a soma de $|m - [t\pi_x] + 1|$ parcelas Z_k, nós temos, usando $E(\sum_{k=1}^t Z_k)^4 = tE(Z_k)^4 + 3t(t-1)[Var(Z_k)]^2 = tK + 3t(t-1)\sigma^4$, que

$$E(S_m - S_{[t\pi_x]})^4 = |m - [t\pi_x] + 1|K + 3|m - [t\pi_x] + 1||m - [t\pi_x]|\sigma^4 \leq$$

$$(|m - [t\pi_x] + 1|)K + 3(|m - [t\pi_x] + 1|)|m - [t\pi_x]|\sigma^4. \tag{2.8}$$

Fazendo $m = N_x(t)$, usando $|N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}$, juntamente com a desigualdade triangular temos que $|N_x(t) - [t\pi_x] + 1| \leq |N_x(t) - [t\pi_x]| + 1 \leq rt^{\frac{1}{2}} + 1$. Continuando a desenvolver a desigualdade (2.8)

$$(rt^{\frac{1}{2}} + 1)K + 3(rt^{\frac{1}{2}} + 1)rt^{\frac{1}{2}}\sigma^4 \leq$$

$$(rt^{\frac{1}{2}} + 1)K + 3(rt^{\frac{1}{2}} + 1)(rt^{\frac{1}{2}} + 1)\sigma^4 \leq$$

$$(rt^{\frac{1}{2}} + 1)K + 3(rt^{\frac{1}{2}} + 1)^2\sigma^4.$$

Assim,

$$\mathbb{P}(|D_t| > \epsilon) \leq \frac{(rt^{\frac{1}{2}} + 1)K + 3(rt^{\frac{1}{2}} + 1)^2\sigma^4}{\epsilon^4 t^2 \pi_i^2}\mathbb{P}(|N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}) + \mathbb{P}(|N_x(t) - [t\pi_x]| > rt^{\frac{1}{2}})$$

Pelo Teorema 2.6, temos que, dado $\epsilon > 0$, $\exists$, com probabilidade 1, $t_0 \in \mathbb{N}$ tal que $|N_x(t) - [t\pi_x]| < \epsilon t$, $\forall t \geq t_0$.

Então, fazendo $r = \epsilon t^{\frac{1}{2}}$, temos que, com probabilidade zero, $|N_x(t) - [t\pi_x]| > rt^{\frac{1}{2}}$, $\forall t \geq t_0$. E consequentemente, com probabilidade 1, $|N_x(t) - [t\pi_x]| \leq rt^{\frac{1}{2}}$, $\forall t \geq t_0$.

Assim $\mathbb{P}(|D_t| > \epsilon) \longrightarrow 0$ quando $t \longrightarrow \infty$. □

A seguir, apresentamos um exemplo envolvendo a disposição entre vogais e consoantes em um texto no idioma russo.

Exemplo 2.5. (Guttorp 1995, página 64). Consideremos um texto russo com 20000 caracteres. Neste texto é analizado cada caracter (vogal ou consoante) e o caracter subsequente (vogal ou consoante). Os dados são apresentados a seguir:

	Próxima é vogal	Próxima é consoante	Total
Vogal	1106	7532	8638
Consoante	7533	3829	11362
Total	8639	11361	20000

Defina

$$X_t = \begin{cases} 0, & \text{se a } t\text{−ésima letra é vogal;} \\ 1, & \text{se a } t\text{−ésima letra é consoante.} \end{cases}$$

Logo a matriz de transição é dada por

$$\mathcal{P} = \begin{pmatrix} \hat{p}_{00} & \hat{p}_{01} \\ \hat{p}_{10} & \hat{p}_{11} \end{pmatrix},$$

no qual

$$\hat{p}_{10} = \frac{t_{10}}{t_1} = \frac{7533}{11361} = 0,663.$$
$$\hat{p}_{01} = \frac{t_{01}}{t_0} = \frac{7532}{8639} = 0,872.$$
$$\hat{p}_{11} = 1 - \hat{p}_{10} = 0,337.$$
$$\hat{p}_{00} = 1 - \hat{p}_{01} = 0,128.$$

Capítulo 3

Modelo de Mistura de Distribuições Independente

Supor que um conjunto de dados amostrais é proveniente de uma única população, gera em muitos casos, superdispersão, ou seja, ocorre uma grande disparidade entre a variância teórica (calculada a partir do modelo) e a variância amostral (calculada a partir da sequência amostral). Pressupor a existência de subpopulações dentro de uma população geral, pode minimizar tal disparidade e é aí que reside a principal motivação no uso dos Modelos de Mistura de Distribuições. A seguir, apresentaremos um breve histórico, obtido a partir de um levantamento bibliográfico feito no decorrer deste trabalho.

Segundo McLachlan e Peel (2000, página 35), a história dos modelos de mistura finita remonta a mais de um século com Pearson (1894), cujos modelos são baseados em uma mistura de duas compontentes normais univariadas. A possibilidade de estabelecer um modelo a partir de distribuições componentes, foi apresentada de forma implícita nos trabalhos de Quetelet (1846, 1852) e mencionado de forma explícita por Galton (1869); veja também Stigler (1986, capítulo 10).

Uma abordagem interessante com relação a primeira referência a mistura de populações pode ser obtida, com maiores detalhes, em Billard (1997), segundo a qual uma das teorias de Falkner (1892) baseava-se na variação dos preços de determinados produtos ou serviços que ocorria durante o ano. Para isto, considerava-se discriminadamente os gastos médios de uma população em determinados produtos ou serviços, somando-se estas médias, estabelecia-se o total de gastos, ao qual era atribuído 100% e consequente, a média de gastos para cada um dos produtos ou serviços era escrito como um percentual deste valor. Isto é, estabelecia-se qual o nível de participação de um determinado gasto no todo. Acompanhava-se a variação destes gastos percentuais durante um ano e calculava-se o aumento ou diminuição do valor total. Cálculo este que forneceu a base para o que chamamos hoje de *Índice de Preços ao Consumidor.*

Falkner baseava-se no cálculo separado, a partir da variação de preços de produtos ou serviços, de diversas médias (média entre o maior e menor valor, média de preços diários, média dos preços na abertura de cada trimestre,...) porém, não considerava a diferença entre o poder aquisitivo de cada indivíduo. Holmes (1892), considerou que a simples comparação entre médias de preço de produtos ou serviços era inadequada, uma vez que deveria ser considerada a disparidade de riqueza entre as pessoas, introduzindo assim o *Conceito de Mistura de Popu-*

lações. Ou seja, a população era subdividida de acordo com o padrão de vida dos indivíduos e calculava-se esse índice anual separadamente.

Este capítulo, irá formalizar conceitos com relação a um dos casos de modelos de mistura finita, o Modelo de Mistura de Distribuições Independente, e a estimação de seus parâmetros, através da maximização da *função de verossimilhança.*

Definições e Notações.

O Modelo de Mistura de Distribuições Independente, segundo Zuchini e Macdonald (2009, página 6), consiste de um número finito m de distribuições componentes associando-se para cada uma dessas distribuições um valor $0 < \delta_i < 1$, $1 \leq i \leq m$, denominadas as probabilidades de mistura, restritas à condição $\sum_1^m \delta_i = 1$. A seguir, apresentamos a definição formal dos modelos de mistura de distribuição independente.

Definição 3.1. Sejam $Y_1, \cdots, Y_m$ variáveis aleatórias independentes, todas contínuas (ou discretas), definidas no mesmo espaço de probabilidade $(\Omega, \mathcal{A}, \mathbb{P})$, com respectivas funções densidade (ou massa) de probabilidade, $f_1(x), \cdots, f_m(x)$ associadas a respectivas probabilidades de mistura: $0 \leq \delta_i \leq 1$, $i = 1, \cdots, m$. Sob a condição $\sum_1^m \delta_i = 1$. A variável aleatória X satisfaz o Modelo de Mistura de Distribuições Independente se sua função densidade (ou massa) de probabilidade satisfaz a equação:

$$f_X(x) = \sum_{i=1}^{m} \delta_i f_i(x), \tag{3.1}$$

em que m é o número de distribuições envolvidas.

Na Proposição 3.1 a seguir, verificamos que o modelo de distribuição independente satisfaz as propriedades de uma função de densidade (ou massa) de probabilidade.

Proposição 3.1. O Modelo de Mistura de Distribuições Independente, ver Definição 3.1, satisfaz as condições para uma função densidade (ou massa) de probabilidade.

Demonstração. Sejam $Y_1, \cdots, Y_m$ variáveis aleatórias com função densidade (ou massa) de probabilidade $f_i(\cdot)$, com suporte A_i, para $i = 1, \cdots, m$, e X uma variável aleatória com função densidade (ou massa) de probabilidade satisfazendo a equação 3.1. Então o suporte de $f_X(\cdot)$ é dado por $A = \cup_{i=1}^m A_i$. Considerando $Y_1, \cdots, Y_m$ variáveis aleatórias discretas temos que:

$$\begin{aligned}\sum_{x\in A} f_X(x) &= \sum_{x\in A}\sum_{i=1}^{m}\delta_i f_i(x)\\ &= \sum_{x\in A_1}\delta_1 f_1(x) + \cdots + \sum_{x\in A_m}\delta_m f_m(x)\\ &= \sum_{i=1}^{m}\delta_i = 1.\end{aligned}$$

Considerando $Y_1, \cdots, Y_m$ variáveis aleatórias contínuas temos que:

$$\begin{aligned}\int_A f_X(x)dx &= \int_A \sum_{i=1}^{m}\delta_i f_i(x)dx\\ &= \sum_{i=1}^{m}\int_A \delta_i f_i(x)dx\\ &= \int_A \delta_1 f_1(x)dx + \cdots + \int_A \delta_m f_m(x)dx\\ &= \delta_1\int_{A_1} f_1(x)dx + \cdots + \delta_m\int_{A_m} f_m(x)dx\\ &\quad + \underbrace{\delta_1\int_{A-A_1} f_1(x)dx + \cdots + \delta_m\int_{A-A_m} f_m(x)dx}_{0}\\ &= \delta_1 + \cdots + \delta_m = 1.\end{aligned}$$

Além disso $f_X(\cdot)$, dada pela equação 3.1, é definida como a combinação linear de função densidade (ou massa) de probabilidade $f_i(\cdot)$, com coeficientes lineares $0 \leq \delta_i \leq 1$, para $i = 1, \cdots, m$. Então $f_X(\cdot) \geq 0$. Portanto $f_X(\cdot)$ satisfaz as propriedades de função densidade de probabilidade ou função massa de probabilidade. □

Observação 3.1. Caso $Y_1, \cdots, Y_m$ sejam variáveis aleatórias discretas, temos que:

$$f_X(x) = \mathbb{P}(X = x) = \sum_{i=1}^{m}\mathbb{P}(X = x|C = i)\mathbb{P}(C = i),$$

em que

$$C = \begin{cases} 1, \text{ com probabilidade } \delta_1 = 1 - \sum_{i=2}^{m}\delta_i;\\ 2, \text{ com probabilidade } \delta_2 = 1 - \delta_1 - \sum_{i=3}^{m}\delta_i;\\ \vdots\\ m, \text{ com probabilidade } \delta_m = 1 - \sum_{i=1}^{m-1}\delta_i. \end{cases}$$

A Proposição 3.2 a seguir apresenta o valor esperado de uma variável aleatória X que segue o Modelo de Mistura de Distribuições Independente, ver Definição 3.1.

Proposição 3.2. Seja X uma variável aleatória que segue o Modelo de Mistura de Distribuições Independente, ver Definição 3.1, então o valor esperado de X é dado por:

$$E(X) = \sum_{i=1}^{m} \delta_i E(Y_i), \tag{3.2}$$

no qual $Y_1, \cdots, Y_m$ são variáveis aleatórias independentes.

Demonstração. Seja $Y_1, \cdots, Y_m$ satisfazendo a Definição 3.1 e suponha que $E|Y_i| < \infty$, para todo $i = 1, \cdots, m$. Seja $f_i(\cdot)$ a função massa de probabilidade ou função densidade de probabilidade da variável aleatória Y_i, com suporte A_i, para cada $i = 1, \cdots, m$. Então $f_X(x)$ tem suporte $A = \bigcup_{i=1}^{m} A_i$.

Se $Y_1, \cdots, Y_m$ são variáveis aleatórias discretas, então

$$\begin{aligned} E(X) = \sum_{x \in A} x f_X(x) &= \sum_{x \in A} x\delta_1 f_1(x) + \cdots + \sum_{x \in A} x\delta_m f_m(x) \\ &= \delta_1 \sum_{x \in A_1} x f_1(x) + \cdots + \delta_m \sum_{x \in A_m} x f_m(x) \\ &= \delta_1 E(Y_1) + \cdots + \delta_m E(Y_m) \\ &= \sum_{i=1}^{m} \delta_i E(Y_i). \end{aligned}$$

Se $Y_1, \cdots, Y_m$ são variáveis aleatórias contínuas, então,

$$\begin{aligned} E(X) = \int_A x f_X(x) dx &= \int_A x \sum_{i=1}^{m} \delta_i f_i(x) dx \\ &= \sum_{i=1}^{m} \delta_i \int_A x f_i(x) dx \\ &= \delta_1 \int_A x f_1(x) dx + \cdots + \delta_m \int_A x f_m(x) dx \\ &= \delta_1 \int_{A_1} x f_1(x) dx + \cdots + \delta_m \int_{A_m} x f_m(x) dx \\ &\quad + \underbrace{\delta_1 \int_{A - A_1} x f_1(x) dx + \cdots + \delta_m \int_{A - A_m} x f_m(x) dx}_{0} \\ &= \delta_1 E(Y_1) + \cdots + \delta_m E(Y_m) \\ &= \sum_{i=1}^{m} \delta_i E(Y_i). \end{aligned}$$

□

Observação 3.2. Denotaremos $\mathbb{N}^* = \{1, 2, 3, 4, 5, \cdots\}$.

A Proposição 3.3 a seguir, apresenta o k-ésimo momento de uma variável aleatória que segue o Modelo de Mistura de Distribuições Independente, ver Definição 3.1.

Proposição 3.3. Seja X uma variável aleatória que segue o Modelo de Mistura de Distribuições Independente, ver Definição 3.1.Suponha $E|Y_i^k| < \infty$, onde $k \in \mathbb{N}^*$ fixo. Então

$$E(X^k) = \sum_{i=1}^{m} \delta_i E(Y_i^k).$$

Demonstração. Segue a demonstração da Proposição 3.2 substituindo X por X^k. □

Proposição 3.4. Seja X uma Variável Aleatória que segue o Modelo de Mistura de Distribuições Independente, ver Definição 3.1, então a variância de X é dada por:

$$Var(X) = \sum_{i=1}^{m} \delta_i Var(Y_i) + \sum_{j=1}^{m-1} \sum_{i=j+1}^{m} \delta_j \delta_i (EY_j - EY_i)^2, \tag{3.3}$$

onde $m \geq 2$

Demonstração. Prova por indução em m. Primeiramente vamos verificar que a expressão (3.3) vale para $m = 2$. Pelas Proposições 3.1 e 3.3, temos que:

$$\begin{aligned}
Var(X) &= \delta_1 E(Y_1^2) + \delta_2 E(Y_2^2) - [\delta_1 E(Y_1) + \delta_2 E(Y_2)]^2 \\
&= \delta_1 E(Y_1^2) + \delta_2 E(Y_2^2) - \delta_1^2 E^2(Y_1) - 2\delta_1\delta_2 E(Y_1)E(Y_2) - \delta_2^2 E^2(Y_2) \\
&= \delta_1 E(Y_1^2) + \delta_2 E(Y_2^2) - \delta_1\delta_1 E^2(Y_1) - 2\delta_1\delta_2 E(Y_1)E(Y_2) - \delta_2\delta_2 E^2(Y_2) \\
&= \delta_1 E(Y_1^2) + \delta_2 E(Y_2^2) - \delta_1(1-\delta_2) E^2(Y_1) - 2\delta_1\delta_2 E(Y_1)E(Y_2) - \delta_2(1-\delta_1) E^2(Y_2) \\
&= \delta_1 E(Y_1^2) + \delta_2 E(Y_2^2) - \delta_1 E^2(Y_1) + \delta_1\delta_2 E^2(Y_1) - 2\delta_1\delta_2 E(Y_1)E(Y_2) \\
&\quad -\delta_2 E^2(Y_2) + \delta_2\delta_1 E^2(Y_2) \\
&= \delta_1(E(Y_1^2) - E^2(Y_1)) + \delta_2(E(Y_2^2) - E^2(Y_2)) + \delta_1\delta_2(E^2(Y_1) \\
&\quad -2E(Y_1)E(Y_2) + E^2(Y_2)) \\
&= \delta_1 Var(Y_1) + \delta_2 Var(Y_2) + \delta_1\delta_2 (E(Y_1) - E(Y_2))^2.
\end{aligned}$$

Suponhamos que a igualdade é valida para algum $m = k,\ k \in \mathbb{N}$. Isto é, suponha que existe um k, natural, tal que para toda variável aleatória X que segue um modelo de mistura de distribuições a k componentes vale

$$Var(X) = \sum_{i=1}^{k} \delta_i Var(Y_i) + \sum_{j=1}^{k-1} \sum_{i=j+1}^{k} \delta_j \delta_i (EY_j - EY_i)^2.$$

Suponha agora que X seja uma variável aleatória que segue um Modelo de Mistura de Distribuições Independente a $k+1$ componentes. Devemos mostrar que

$$Var(X) = \sum_{i=1}^{k+1} \delta_i Var(Y_i) + \sum_{j=1}^{k} \sum_{i=j+1}^{k+1} \delta_j \delta_i (EY_j - EY_i)^2.$$

Pelas Proposições 3.1 e 3.3 temos que

$$Var(X) = \sum_{i=1}^{k+1} \delta_i EY_i^2 - \left(\sum_{i=1}^{k+1} \delta_i EY_i \right)^2.$$

Então

$$Var(x) = \delta_{k+1} EY_{k+1}^2 + \sum_{i=1}^{k} \delta_i EY_i^2 - \left(\delta_{k+1} EY_{k+1} + \sum_{i=1}^{k} \delta_i EY_i \right)^2$$

Desenvolvendo o quadrado da soma e usando o fato $\sum_{i=1}^{k+1} \delta_i = 1$ temos

$$\begin{aligned} Var(X) &= \delta_{k+1} EY_{k+1}^2 + \sum_{i=1}^{k} \delta_i EY_i^2 - \delta_{k+1}^2 E^2 Y_{k+1} - 2\delta_{k+1} EY_{k+1} \sum_{i=1}^{k} \delta_i EY_i - \left(\sum_{i=1}^{k} \delta_i EY_i \right)^2 \\ &= \delta_{k+1} EY_{k+1}^2 - \delta_{k+1}(1 - \delta_1 - \cdots - \delta_k) E^2 Y_{k+1} + \sum_{i=1}^{k} \delta_i EY_i^2 - \left(\sum_{i=1}^{k} \delta_i EY_i \right)^2 \\ &\quad -2\delta_1 \delta_{k+1} EY_1 EY_{k+1} - 2\delta_2 \delta_{k+1} EY_2 EY_{k+1} - \cdots - 2\delta_k \delta_{k+1} EY_k EY_{k+1}. \end{aligned}$$

Consequentemente

$$\begin{aligned} Var(X) &= \delta_{k+1} EY_{k+1}^2 - \delta_{k+1} E^2 Y_{k+1} + \sum_{i=1}^{k} \delta_i EY_i^2 - \left(\sum_{i=1}^{k} \delta_i EY_i \right)^2 \\ &\quad +\delta_1 \delta_{k+1} E^2 Y_{k+1} + \delta_2 \delta_{k+1} E^2 Y_{k+1} + \delta_3 \delta_{k+1} E^2 Y_{k+1} + \cdots + \delta_k \delta_{k+1} E^2 Y_{k+1} \\ &\quad -2\delta_1 \delta_{k+1} EY_1 EY_{k+1} - 2\delta_2 \delta_{k+1} EY_2 EY_{k+1} - \cdots - 2\delta_k \delta_{k+1} EY_k EY_{k+1}. \end{aligned}$$

Colocando δ_{k+1} em evidência nas duas primeiras parcelas

$$\begin{aligned} Var(X) &= \delta_{k+1} VarY_{k+1} + \sum_{i=1}^{k} \delta_i EY_i^2 - \left(\sum_{i=1}^{k} \delta_i EY_i \right)^2 \\ &\quad +\delta_1 \delta_{k+1} E^2 Y_{k+1} + \delta_2 \delta_{k+1} E^2 Y_{k+1} + \delta_3 \delta_{k+1} E^2 Y_{k+1} + \cdots + \delta_k \delta_{k+1} E^2 Y_{k+1} \\ &\quad -2\delta_1 \delta_{k+1} EY_1 EY_{k+1} - 2\delta_2 \delta_{k+1} EY_2 EY_{k+1} - \cdots - 2\delta_k \delta_{k+1} EY_k EY_{k+1}. \end{aligned}$$

Somando e subtraindo o valor δ_{k+1}

$$\begin{aligned} Var(X) &= \delta_{k+1}VarY_{k+1} + \sum_{i=1}^{k-1}\delta_i EY_i^2 + (\delta_k+\delta_{k+1})EY_k^2 - \delta_{k+1}EY_k^2 \\ &\quad - \left(\sum_{i=1}^{k-1}\delta_i EY_i + (\delta_k+\delta_{k+1})EY_k - \delta_{k+1}EY_k\right)^2 \\ &\quad +\delta_1\delta_{k+1}E^2Y_{k+1} + \delta_2\delta_{k+1}E^2Y_{k+1} + \delta_3\delta_{k+1}E^2Y_{k+1} + \cdots + \delta_k\delta_{k+1}E^2Y_{k+1} \\ &\quad -2\delta_1\delta_{k+1}EY_1EY_{k+1} - 2\delta_2\delta_{k+1}EY_2EY_{k+1} - \cdots - 2\delta_k\delta_{k+1}EY_kEY_{k+1}. \end{aligned}$$

Desenvolvendo o quadrado da soma

$$\begin{aligned} Var(X) &= \delta_{k+1}VarY_{k+1} + \sum_{i=1}^{k-1}\delta_i EY_i^2 + (\delta_k+\delta_{k+1})EY_k^2 - \delta_{k+1}EY_k^2 \\ &\quad - \left(\sum_{i=1}^{k-1}\delta_i EY_i + (\delta_k+\delta_{k+1})EY_k\right)^2 + 2\left(\sum_{i=1}^{k-1}\delta_i EY_i + (\delta_k+\delta_{k+1})EY_k\right)\delta_{k+1}EY_k \\ &\quad -\delta_{k+1}^2E^2Y_k + \delta_1\delta_{k+1}E^2Y_{k+1} + \delta_2\delta_{k+1}E^2Y_{k+1} + \delta_3\delta_{k+1}E^2Y_{k+1} + \cdots + \delta_k\delta_{k+1}E^2Y_{k+1} \\ &\quad -2\delta_1\delta_{k+1}EY_1EY_{k+1} - 2\delta_2\delta_{k+1}EY_2EY_{k+1} - \cdots - 2\delta_k\delta_{k+1}EY_kEY_{k+1}. \end{aligned}$$

Usando a hipótese de indução em

$$\sum_{i=1}^{k-1}\delta_i EY_i^2 + (\delta_k+\delta_{k+1})EY_k^2 - \left(\sum_{i=1}^{k-1}\delta_i EY_i + (\delta_k+\delta_{k+1})EY_k\right)^2,$$

temos

$$\begin{aligned} Var(X) &= \delta_{k+1}VarY_{k+1} + \sum_{i=1}^{k-1}\delta_i VarY_i + (\delta_k+\delta_{k+1})VarY_k - \delta_{k+1}EY_k^2 \\ &\quad + \sum_{j=1}^{k-2}\sum_{i=j+1}^{k-1}\delta_j\delta_i(EY_j - EY_i)^2 + \sum_{i=1}^{k-1}\delta_i(\delta_k+\delta_{k+1})(EY_i - EY_k)^2 \\ &\quad +\delta_1\delta_{k+1}E^2Y_{k+1} + \delta_2\delta_{k+1}E^2Y_{k+1} + \delta_3\delta_{k+1}E^2Y_{k+1} + \cdots + \delta_k\delta_{k+1}E^2Y_{k+1} \\ &\quad -2\delta_1\delta_{k+1}EY_1EY_{k+1} - 2\delta_2\delta_{k+1}EY_2EY_{k+1} - \cdots - 2\delta_k\delta_{k+1}EY_kEY_{k+1} \\ &\quad +2\left(\sum_{i=1}^{k-1}\delta_i EY_i + \delta_k EY_k + \delta_{k+1}EY_k\right)\delta_{k+1}EY_k - \delta_{k+1}^2E^2Y_k. \end{aligned}$$

A seguir, completaremos quadrados e faremos algumas operações convenientes

$$\begin{aligned}
Var(X) &= \sum_{i=1}^{k+1}\delta_i VarY_i + \delta_{k+1}VarY_k - \delta_{k+1}EY_k^2 \\
&+\sum_{j=1}^{k-2}\sum_{i=j+1}^{k-1}\delta_j\delta_i(EY_j - EY_i)^2 + \sum_{i=1}^{k-1}\delta_i\delta_k(EY_i - EY_k)^2 \\
&+\sum_{i=1}^{k-1}\delta_i\delta_{k+1}(EY_i - EY_k)^2 + (\delta_1\delta_{k+1}E^2Y_{k+1} + \delta_2\delta_{k+1}E^2Y_{k+1} + \cdots + \delta_k\delta_{k+1}E^2Y_{k+1} \\
&-2\delta_1\delta_{k+1}EY_1EY_{k+1} - 2\delta_2\delta_{k+1}EY_2EY_{k+1} - \cdots - 2\delta_k\delta_{k+1}EY_kEY_{k+1} \\
&+\delta_1\delta_{k+1}E^2Y_1 + \delta_2\delta_{k+1}E^2Y_2 + \cdots + \delta_k\delta_{k+1}E^2Y_k) \\
&-\delta_1\delta_{k+1}E^2Y_1 - \delta_2\delta_{k+1}E^2Y_2 - \cdots - \delta_k\delta_{k+1}E^2Y_k \\
&+2\left(\sum_{i=1}^{k-1}\delta_iEY_i + \delta_kEY_k + \delta_{k+1}EY_k\right)\delta_{k+1}EY_k - \delta_{k+1}^2E^2Y_k.
\end{aligned}$$

Expandindo a última parcela. Terminando o completamento de quadrados. Usando

$$\sum_{j=1}^{k-2}\sum_{i=j+1}^{k-1}\delta_j\delta_i(EY_j - EY_i)^2 + \sum_{i=1}^{k-1}\delta_i\delta_k(EY_i - EY_k)^2 = \sum_{j=1}^{k-1}\sum_{i=j+1}^{k}\delta_j\delta_i(EY_j - EY_i)^2$$

e $VarY_k = EY_k^2 - E^2Y_k$ segue que

$$\begin{aligned}
Var(X) &= \sum_{i=1}^{k+1}\delta_i VarY_i + \delta_{k+1}EY_k^2 - \delta_{k+1}E^2Y_k - \delta_{k+1}EY_k^2 \\
&+\sum_{j=1}^{k-1}\sum_{i=j+1}^{k}\delta_j\delta_i(EY_j - EY_i)^2 + \sum_{i=1}^{k}\delta_i\delta_{k+1}(EY_i - EY_{k+1})^2 + \sum_{i=1}^{k-1}\delta_i\delta_{k+1}(EY_i - EY_k)^2 \\
&-\delta_1\delta_{k+1}E^2Y_1 - \cdots - \delta_{k-1}\delta_{k+1}E^2Y_{k-1} - \delta_k\delta_{k+1}E^2Y_k \\
&+2\delta_1\delta_{k+1}EY_1EY_k + \cdots + 2\delta_{k-1}\delta_{k+1}EY_{k-1}EY_k + 2\delta_k\delta_{k+1}E^2Y_k \\
&+2\delta_{k+1}^2E^2Y_k - \delta_{k+1}^2E^2Y_k.
\end{aligned}$$

Novamente completando quadrados. Usando

$$\sum_{j=1}^{k-1}\sum_{i=j+1}^{k}\delta_j\delta_i(EY_j - EY_i)^2 + \sum_{i=1}^{k}\delta_i\delta_{k+1}(EY_i - EY_{k+1})^2 = \sum_{j=1}^{k-1}\sum_{i=j+1}^{k+1}\delta_j\delta_i(EY_j - EY_i)^2$$

e reduzindo os termos semelhantes temos que

$$
\begin{aligned}
Var(x) &= \sum_{i=1}^{k+1} \delta_i Var Y_i + \sum_{j=1}^{k-1} \sum_{i=j+1}^{k+1} \delta_j \delta_i (EY_j - EY_i)^2 \\
&+ \sum_{i=1}^{k-1} \delta_i \delta_{k+1} (EY_i - EY_k)^2 + \delta_k \delta_{k+1} E^2 Y_k - \delta_{k+1} E^2 Y_k + \delta_{k+1}^2 E^2 Y_k \\
&-(\delta_1 \delta_{k+1} E^2 Y_1 + \cdots + \delta_{k-1} \delta_{k+1} E^2 Y_{k-1} \\
&-2\delta_1 \delta_{k+1} EY_1 EY_k - \cdots - 2\delta_{k-1} \delta_{k+1} EY_{k-1} EY_k \\
&+\delta_1 \delta_{k+1} E^2 Y_k + \cdots + \delta_{k-1} \delta_{k+1} E^2 Y_k) \\
&+\delta_1 \delta_{k+1} E^2 Y_k + \cdots + \delta_{k-1} \delta_{k+1} E^2 Y_k.
\end{aligned}
$$

Sendo assim

$$
\begin{aligned}
Var(X) &= \sum_{i=1}^{k+1} \delta_i Var Y_i + \sum_{j=1}^{k-1} \sum_{i=j+1}^{k+1} \delta_j \delta_i (EY_j - EY_i)^2 - \delta_{k+1} E^2 Y_k \\
&+ \sum_{i=1}^{k-1} \delta_i \delta_{k+1} (EY_i - EY_k)^2 + \delta_k \delta_{k+1} E^2 Y_k + \delta_{k+1}^2 E^2 Y_k \\
&- \sum_{i=1}^{k-1} \delta_i \delta_{k+1} (EY_i - EY_k)^2 + \sum_{i=1}^{k-1} \delta_i \delta_{k+1} E^2 Y_k.
\end{aligned}
$$

Colocando δ_{k+1} em evidência nas parcelas que não são acompanhadas por somatório e usando $\sum_{i=1}^{k-1} \delta_1 = 1 - \delta_{k+1} - \delta_k$

$$
\begin{aligned}
Var(X) &= \sum_{i=1}^{k+1} \delta_i Var Y_i + \sum_{j=1}^{k-1} \sum_{i=j+1}^{k+1} \delta_j \delta_i (EY_j - EY_i)^2 \\
&-\delta_{k+1}(1 - \delta_k - \delta_{k+1}) E^2 Y_k + \delta_{k+1}(1 - \delta_k - \delta_{k+1}) E^2 Y_k.
\end{aligned}
$$

E então

$$
Var(X) = \sum_{i=1}^{k+1} \delta_i Var Y_i + \sum_{j=1}^{k-1} \sum_{i=j+1}^{k+1} \delta_j \delta_i (EY_j - EY_i)^2.
$$

□

Como próximo passo, estamos interessados em estimar os parâmetros das distribuições das variáveis aleatórias $Y_1, \cdots, Y_m$ e as probabilidades de mistura $\delta_1, \cdots, \delta_m$. Para isso, definimos a *função de verossimilhança* dos Modelos de Mistura de Distribuições Independente.

Definição 3.2. Sejam $\boldsymbol{x} = (x_1, \cdots, x_n)$ uma sequência de observações de tamanho n. Seja X uma variável aleatória com função densidade (ou massa) de probabilidade $f_X(x)$ seguindo o

Modelo de Mistura de Distribuições Independente, ver Definição 3.1, com $\boldsymbol{\theta} \in \Theta$, onde Θ é o espaço paramétrico. A função de verossimilhança correspondente à sequência de observações é dada por:

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{x}) = \prod_{j=1}^{n} \sum_{i=1}^{m} \delta_i f_i(x_j, \theta_{\mathbf{i}}), \tag{3.4}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta_m})$ é o vetor de parâmetros das distribuições componentes e $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_m)$ é o vetor formado pelas probabilidades de mistura.

A seguir apresentamos um exemplo da função de verossimilhança do Modelo de Misturas de Distribuições Independente, ver Definição 3.1, no qual as variáveis aleatórias possuem distribuições Poisson.

Exemplo 3.1. Seja $Y_1, \cdots, Y_m$ variáveis aleatórias independentes, em que Y_i possui distribuição de Poisson com parâmetro λ_i, para cada $i = 1, \cdots, m$. A função de verossimilhança de uma variável aleatória X que segue o Modelo de Mistura de Distribuições Independente, ver Definição 3.1, é dado por:

$$L(\boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{x}) = \prod_{j=1}^{n} \sum_{i=1}^{m} \delta_i \frac{\lambda_i^{x_j} e^{-\lambda_i}}{x_j!}.$$

Considerando $m = 1$, temos que $\delta_1 = 1$ e que $p_1(x) = \frac{\lambda_1^x e^{-\lambda_1}}{x!}$. Sendo assim:

$$L(\lambda_1, \mathbf{x}) = \prod_{i=1}^{n} \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!}.$$

Considerando $m = 2$, temos que $\delta_2 = 1 - \delta_1$, e $p_1(x) = \frac{\lambda_1^x e^{-\lambda_1}}{x!}$, e que $p_2(x) = \frac{\lambda_2^x e^{-\lambda_2}}{x!}$. Sendo assim:

$$L(\lambda_1, \lambda_2, \delta_1, \mathbf{x}) = \prod_{i=1}^{n} (\delta_1 \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!} + (1 - \delta_1) \frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!}).$$

Definição 3.3. O estimador de máxima verossimilhança do vetor de parâmetros $(\boldsymbol{\theta}, \boldsymbol{\delta})$ é o vetor $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\delta}})$ que maximiza a função de verossimilhança $L(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{x})$ ou o logaritmo da função de verossimilhança, denotado por $\pounds(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{x}) = \log(L(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{x}))$.

É possível, para casos simples, escrever um código em R para calcular e maximizar a *função de verossimilhança* de uma distribuição de mistura. No exemplo a seguir (ver Zucchini e MacDonald 2009, p. 12) utilizamos o código MMIX da seção A.1. Observe que o MMIX, otimiza a função $\log L$, porém sabemos que otimizar L é o mesmo que otimizar $\log L$.

Exemplo 3.2. Considere um modelo de mistura de m distribuições independentes do tipo Poisson, com respectivas médias $\lambda_1, \cdots, \lambda_m$. Nossa sequência de observações, de tamanho $n = 107$ será baseada no registro dos tremores de terra de magnitude maior ou igual a sete (7.0), no mundo, que ocorreram entre 1900-2006. Os dados são apresentados no Apêndice A.2 na página 56. Vamos considerar um Modelo de Mistura de Distribuições Independente com

m distribuições em que as variáveis aleatórias $Y_1, \cdots, Y_m$ possuem distribuição Poisson com parâmetro λ_j, para $j = 1, \cdots, m$ respectivamente.

A seguir a Tabela 3.1 contendo as estimativas obtidas a partir da maximização da *função de verossimilhança* com o programa MMIX utilizando o software R.

Modelo	p	i	δ_i	λ_i	$-\log L$	Média	Variância	AIC	BIC
$m = 1$	1	1	1	19,364	391,918	19,364	19,364	785,837	788,510
$m = 2$	3	1	0,676	15,777	360,369	19,364	46,182	726,738	**734,756**
		2	0,324	26,84					
$m = 3$	5	1	0,278	12,736	356,848	19,364	51,169	723,697	737,061
		2	0,593	19,785					
		3	0,13	31,629					
$m = 4$	7	1	0,093	10,584	356,733	19,364	51,638	727,467	746,177
		2	0,354	15,528					
		3	0,437	20,969					
		4	0,116	32,079					
Amostra						19,364	51,573		

Tabela 3.1: Resultados obtidos para os *modelos de mistura de distribuições independentes* para a amostra de tremores de terra. O número de distribuições componentes consideradas é $m \in \{1, \cdots, 4\}$, p representa o número de parâmetros a serem estimados, δ_i, para $i = 1, \cdots, m$, as probabilidades de mistura e λ_i o parâmetro da distribuição Poisson, para cada $i = 1, \cdots, m$. Nas duas últimas colunas apresentamos os valores dos critérios de seleção de modelos AIC e BIC (ver Observação 3.3 a seguir).

Observação 3.3. Os Critérios de Seleção de Modelos satisfazem as seguintes equações

$$AIC = -2 \log L + 2p$$

e

$$BIC = -2 \log L + p \log T,$$

no qual p é o número de parâmetros a serem estimados, $L(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{x})$ é a *função de verossimilhança* do modelo e T é o número de observações. O modelo a ser selecionado, dito "melhor modelo", é o que possui menor AIC e BIC. Neste exemplo não é possível determiná-lo, utilizando os dois critérios em conjunto, pois o menor AIC ocorre para $m = 3$ e o menor BIC ocorre para $m = 2$. O que podemos fazer é escolher o critério de seleção mais adequado. Quando uma sequência de dados amostrais apresentar mais de 100 observações o critério mais adequado é o BIC. Neste caso estamos trabalhando com 107 observações, ver tabela A.1 na página 56. Assim iremos selecionar o modelo de distribuição indepente do tipo Poisson que utiliza 2 componentes.

A Figura 3.1 a seguir apresenta os gráficos da função densidade de probabilidade de cada modelo comparado ao histograma dos dados amostrais, $m \in \{1, 2, 3, 4\}$.

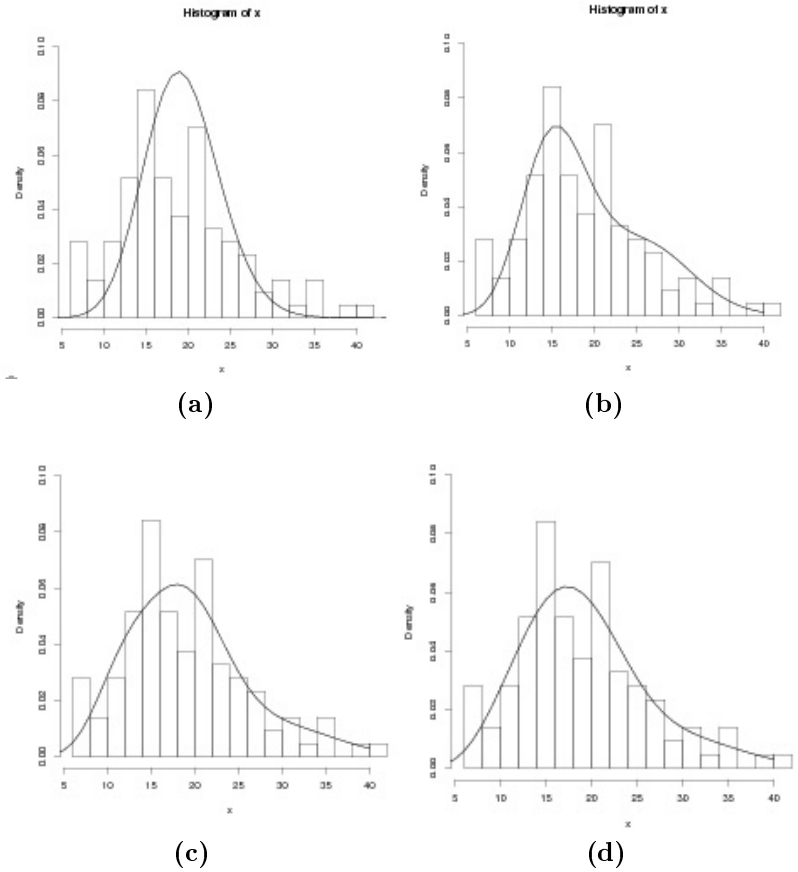


Figura 3.1: Histograma da amostra de tremores de terra e suas respectivas funções densidades de probabilidades dos modelos de mistura de distribuições independente (ver Definição 3.1). (a) $m = 1$; (b) $m = 2$; (c) $m = 3$; (d) $m = 4$.

Zucchini e MacDonald (2009) observa que não é considerado aqui, dependência entre os dados amostrais. Assunto que será detalhado no capítulo a seguir.

Capítulo 4

Modelos Ocultos de Markov

No capítulo anterior utilizamos o Modelo de Mistura de Distribuições Independente para modelagem dos dados "número de tremores de terra com magnitude maiores ou iguais a 7 ocorridos entre os anos de 1900 e 2006", ver Apêndice A.2. Porém, como mostra o gráfico a seguir, estes dados são correlacionados. O que nos sugere o uso de um modelo de mistura dependente. Neste capítulo abordaremos um destes modelos: o Modelo Oculto de Markov.

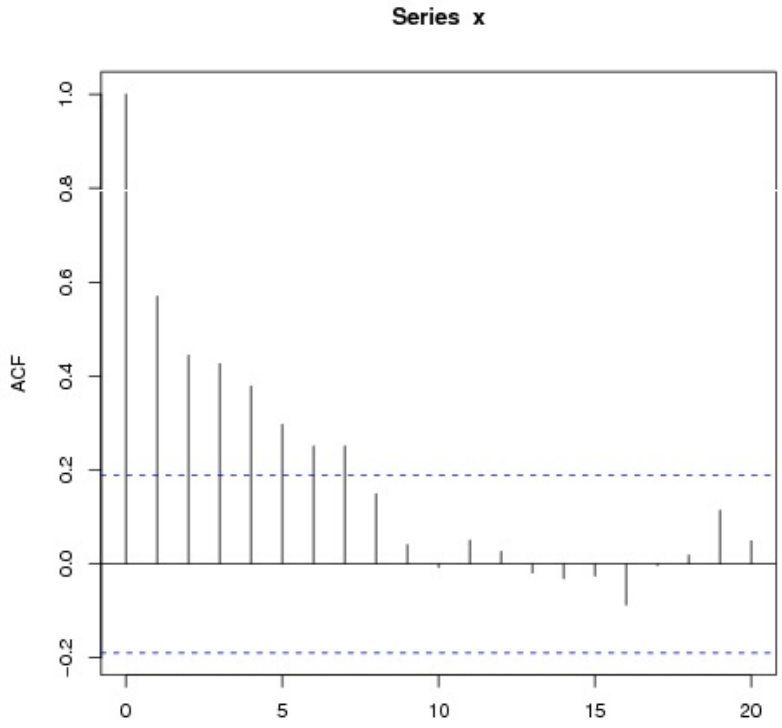
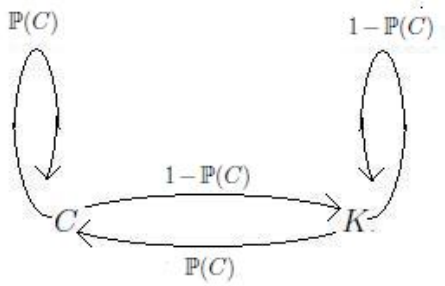


Figura 4.1: Série de tremores de terra: função de autocorrelação amostral.

Este tipo de modelagem de dados pressupõe um processo gerador, markoviano, o qual está oculto do observador. Isto é, supõe-se que a sequência de observações fornecida é gerada por um processo markoviano não observável (oculto). Para fixar esta ideia veja o exemplo a seguir.

Exemplo 4.1. *(Rabiner, 1989).* Suponha que você está em uma sala onde uma cortina o impede de enxergar o que ocorre. Do outro lado da cortina ocorre o lançamento de uma ou mais moedas. A pessoa responsável por tal experimento não contará a você o que exatamente ocorre. Ou seja, é feito um experimento oculto de lançamento de moedas. Suponha ainda que lhe é fornecida uma sequência de caras e coroas: $KKKCCKCCC\cdots KKCC$, em que C representa cara e K coroa.

Uma das formas de modelarmos esta situação é considerar que ocorre o lançamento de uma única moeda viciada. Neste caso se a probabilidade de ocorrer cara é dada por $\mathbb{P}(C)$, temos o seguinte diagrama para o processo markoviano gerador da sequência de observações.



Note que o processo acima não possui memória. O que o caracteriza como um processo markoviano degenerado, ou seja, o processo markoviano no qual o estado atual independe dos outros estados.

A seguir, a formalização do Modelo Oculto de Markov.

Considerando $(Z_t)_{t\in\mathbb{N}^*}$ a sequência de dados amostrais, $(X_t)_{t\in\mathbb{N}^*}$ uma Cadeia de Markov a tempo discreto, estabelecemos a seguinte definição para o Modelo Oculto de Markov (ver Zucchini e MacDonald, 2009).

Definição 4.1. Um *Modelo Oculto de Markov* $(Z_t)_{t\in\mathbb{N}^*}$ satisfaz as seguintes equações

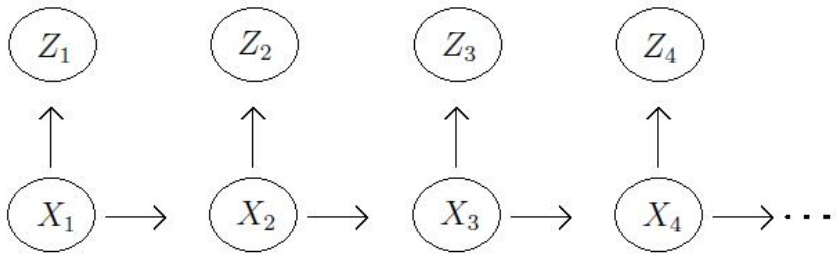
$$\mathbb{P}(X_t|X^{(t-1)}) = \mathbb{P}(X_t|X_{t-1}) \tag{4.1}$$

e

$$\mathbb{P}(Z_t|Z^{(t-1)}, X^{(t)}) = \mathbb{P}(Z_t|X_t), \tag{4.2}$$

em que $t \in \mathbb{N}^*$, com $Z^{(t)} = \{Z_1, \cdots, Z_t\}$ e $X^{(t)} = \{X_1, \cdots, X_t\}$ as histórias do tempo 1 até o tempo t.

Note que, quando $(X_t)_{t\in\mathbb{N}^*}$ é conhecido, a distribuição de $(Z_t)_{t\in\mathbb{N}^*}$ depende somente do estado atual X_t e não depende dos estados anteriores. Esta relação pode ser representada através do seguinte diagrama.



Se a Cadeia de Markov $(X_t)_{t\in\mathbb{N}^*}$ possui um espaço de estados S finito com m estados, $(Z_t)_{t\in\mathbb{N}^*}$ é um Modelo Oculto de Markov com m estados. Neste trabalho consideramos espaço de estados finitos.

Notação

Considerando um conjunto de observações discretas nós denotaremos, para $x \in \{1, 2, \cdots, m\}$,

$$f_x(z) = \mathbb{P}(Z_t = z | X_t = x)$$

Isto é, $f_x(z)$ é a função densidade (ou massa) de probabilidade de X_t se a Cadeia de Markov está no estado x no tempo t.

Na Seção 4.1 a seguir apresentamos as distribuições marginais dos Modelos Ocultos de Markov.

4.1 Distribuições Marginais

Nesta seção estamos interessados em encontrar as distribuições marginais do Modelo de Markov Oculto $(Z_t)_{t\in\mathbb{N}^*}$. Vamos primeiramente, verificar para o caso univariado.

Observação 4.1. Os valores $(Z_t)_{t\in\mathbb{N}^*}$ são obtidos através da sequência de observações. Isto é se o $t-$ésimo dado observado é z então $Z_t = z$.

Caso Univariado

Seja $(X_t)_{t\in\mathbb{N}^*}$ uma Cadeia de Markov com espaço de estados $S = \{1, \cdots, m\}$, em que $z \in \{z_1, \cdots, z_t\}$ representa a sequência de observações fornecida. Defina $u_x(t) = \mathbb{P}(X_t = x)$ para $t = 1, \cdots, T$, a $x-$ésima coordenada do vetor $\boldsymbol{u}(t) = (\mathbb{P}(X_t = 1), \mathbb{P}(X_t = 2), \cdots, \mathbb{P}(X_t = m))$, considere $(X_t)_{t\in\mathbb{N}^*}$ uma Cadeia de Markov com distribuição inicial $\boldsymbol{u}(1) = (u_1(1), \cdots, u_m(1))$ e $(Z_t)_{t\in\mathbb{N}^*}$ um Modelo Oculto de Markov. Assim

$$\mathbb{P}(Z_t = z) = \sum_{x=1}^{m} \mathbb{P}(X_t = x)\mathbb{P}(Z_t = z | X_t = x) = \sum_{x=1}^{m} u_x(t) f_x(z), \tag{4.3}$$

em que $x \in \mathbb{N}^*, 1 \leq x \leq m$ e $z \in \mathbb{R}$.

A expressão (4.3) pode ser convenientemente reescrita na notação matricial

$$\mathbb{P}(Z_t = z) = (u_1(t), \cdots, u_m(t)) \begin{pmatrix} f_1(z) & & 0 \\ & \ddots & \\ 0 & & f_m(z) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \boldsymbol{u}(t)\boldsymbol{\mathcal{D}}(z)\mathbf{1'}.$$

Utilizando a equação $\boldsymbol{u}(t) = \boldsymbol{u}(1)\boldsymbol{\mathcal{P}}^{t-1}$ da página 7 temos que

$$\mathbb{P}(Z_t = z) = \mathbf{u}(1)\boldsymbol{\mathcal{P}}^{t-1}\boldsymbol{\mathcal{D}}(z)\mathbf{1'},$$

em que

$$\boldsymbol{\mathcal{D}}(z) = \begin{pmatrix} f_1(z) & & 0 \\ & \ddots & \\ 0 & & f_m(z) \end{pmatrix}. \tag{4.4}$$

A equação se mantém para uma cadeia homogênea, e se além disso, a cadeia for estacionária temos

$$\mathbb{P}(Z_t = z) = \boldsymbol{\delta}\boldsymbol{\mathcal{D}}(z)\mathbf{1'},$$

em que $\boldsymbol{\delta}$ é a distribuição inicial, estacionária, da Cadeia de Markov $(X_t)_{t\in\mathbb{N}^*}$.

Caso Bivariado

Agora estabeleceremos uma expressão para a distribuição marginal bivariada. Aqui também consideraremos apenas o caso discreto.

$$\begin{aligned}
&\mathbb{P}(X_t{=}x_t, X_{t+k}{=}x_{t+k}, Z_t{=}z_t, Z_{t+k}{=}z_{t+k}) \\
&= \mathbb{P}(X_t)\mathbb{P}(X_{t+k} = x_{t+k}|X_t = x_t)\mathbb{P}(Z_t = z_t|X_{t+k} = x_{t+k}, X_t = x_t) \\
&\quad \times\mathbb{P}(Z_{t+k} = z_{t+k}|Z_t = z_t, X_{t+k} = x_{t+k}, X_t = x_t) \\
&= \mathbb{P}(X_t = x_t)\mathbb{P}(X_{t+k} = x_{t+k}|X_t = x_t)\mathbb{P}(Z_t = z_t|X_t = x_t)\mathbb{P}(Z_{t+k} = z_{t+k}|X_{t+k} = x_{t+k}), \qquad (4.5)
\end{aligned}$$

no qual a equação (4.5) vale pela Definição 4.1. Assim

$$\begin{aligned}
\mathbb{P}(Z_t{=}z, Z_{t+k}{=}v) &= \sum_{x=1}^{m}\sum_{y=1}^{m}\mathbb{P}(X_t = x, Z_t = z, X_{t+k} = y, Z_{t+k} = v) \\
&= \sum_{x=1}^{m}\sum_{y=1}^{m}\mathbb{P}(X_t = x)\mathbb{P}(Z_t = z|X_t = x) \\
&\quad \times \mathbb{P}(X_{t+k} = y|X_t = x)\mathbb{P}(Z_{t+k} = v|X_{t+k} = y) \\
&= \sum_{x=1}^{m}\sum_{y=1}^{m}u_x(t)f_x(z)p_{xy}^k f_y(v),
\end{aligned}$$

em que $x, y \in \mathbb{N}^*, 1 \leq x, y \leq m$ e $z, v \in \mathbb{R}$. Escrevendo a soma anterior como um produto de matrizes temos

$$\mathbb{P}(Z_t = z, Z_{t+k} = v) = \boldsymbol{u}(t)\boldsymbol{\mathcal{D}}(z)\boldsymbol{\mathcal{P}}^k\boldsymbol{\mathcal{D}}(v)\mathbf{1}', \tag{4.6}$$

no qual $\boldsymbol{\mathcal{D}(z)}$ é dado na equação (4.4) e $\boldsymbol{\mathcal{P}}$ é a matriz de transição da Cadeia de Markov $(X_t)_{t\in\mathbb{N}^*}$.

Se a Cadeia de Markov é estacionária, a equação (4.6) acima reduz-se a

$$\mathbb{P}(Z_t = z, Z_{t+k} = v) = \boldsymbol{\delta}\boldsymbol{\mathcal{D}}(z)\boldsymbol{\mathcal{P}}^k\boldsymbol{\mathcal{D}}(v)\mathbf{1}',$$

em que $\boldsymbol{\delta}$ é a distribuição estacionária e inicial da Cadeia de Markov $(X_t)_{t\in\mathbb{N}^*}$.

A seguir a expressão geral.

Expressão Geral para a Distribuição Marginal

Dados $\{z_0, z_1, z_2, \cdots, z_k\}$ pertencentes ao Modelo oculto de Markov e $\{x_0, x_1, x_2, \cdots, x_k\}$ pertencentes ao espaço de estados da Cadeia de Markov $(X_t)_{t\in\mathbb{N}^*}$, podemos encontrar a expressão geral para distribuição marginal, de dimensão $n+1$ com $n \in \mathbb{N}^*$ qualquer, da seguite forma

$$
\begin{aligned}
&\mathbb{P}(Z_t = z_0, Z_{t+t_1} = z_1,\\
&\cdots, Z_{t+t_1+\cdots+t_k} = z_k) \;=\; \sum_{x_0,\cdots,x_k\in S} \mathbb{P}(X_t = x_0, X_{t+t_1} = x_1, \cdots,\\
&\qquad X_{t+t_1+\cdots+t_k} = x_k, Z_t = z_0, \cdots, Z_{t+t_1+\cdots+t_k} = z_k)\\
&\qquad = \sum_{x_0,\cdots,x_k\in S} \mathbb{P}(X_t = x_0)\mathbb{P}(X_{t+t_1} = x_1|X_t = x_0)\cdots\\
&\qquad \mathbb{P}(X_{t+t_1+\cdots+t_k} = x_k|X_{t+t_1+\cdots+t_{k-1}} = x_{k-1})\cdots\\
&\qquad \mathbb{P}(Z_t = z_0|X_t = x_0)\cdots\mathbb{P}(Z_{t+t_1+\cdots+t_k} = z_k|X_{t+t_1+\cdots+t_k} = x_k)\\
&\qquad = \sum_{x_0,\cdots,x_k\in S} u_{x_0}p^{t_1}_{x_0x_1}p^{t_2}_{x_1x_2}\cdots p^{t_k}_{x_{k-1}x_k}f_{x_0}(z_0)\cdots f_{x_k}(z_k)\\
&\qquad = \sum_{x_0,\cdots,x_k\in S} u_{x_0}f_{x_0}(z_0)p^{t_1}_{x_0x_1}f_{x_1}(z_1)p^{t_2}_{x_1x_2}\cdots p^{t_k}_{x_{k-1}x_k}f_{x_k}(z_k)\\
&\qquad = \boldsymbol{u}\boldsymbol{\mathcal{D}}(z_0)\boldsymbol{\mathcal{P}}^{t_1}\boldsymbol{\mathcal{D}}(z_1)\cdots\boldsymbol{\mathcal{P}}^{t_k}\boldsymbol{\mathcal{D}}(z_k)
\end{aligned}
$$

Se a Cadeia de Markov é estacionária, a expressão geral para a distribuição marginal se reduz a

$$\mathbb{P}(Z_t = z_0, Z_{t+t_1} = z_1, Z_{t+t_1+\cdots+t_k} = z_k) = \boldsymbol{\delta}\boldsymbol{\mathcal{D}}(z_0)\boldsymbol{\mathcal{P}}^{t_1}\boldsymbol{\mathcal{D}}(z_1)\cdots\boldsymbol{\mathcal{P}}^{t_k}\boldsymbol{\mathcal{D}}(z_k),$$

em que $\boldsymbol{\delta}$ é a distribuição estacionária e inicial da Cadeia de Markov $(X_t)_{t\in\mathbb{N}^*}$.

Na Seção 4.2 a seguir, apresentamos o método de estimação de máxima verossimilhança para os Modelos Ocultos de Markov.

4.2 Estimadores de Máxima Verossimilhança

Nesta seção apresentamos os estimadores de máxima verossimilhança para os modelos ocultos de Markov baseado em uma amostra de observações consecutivas $(z_1, \cdots, z_T)$ de tamanho T. Inicialmente apresentamos um exemplo (ver Zucchini e MacDonald (2009, p. 35)).

Exemplo 4.2. Modelo Oculto de Markov Bernoulli a dois estados.

Considere um modelo Modelo Oculto de Markov Bernoulli a dois estados, com matriz de transição

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix},$$

e espaço de estados $S = \{0, 1\}$ em que a distribuição estacionária é dada por $\boldsymbol{\delta} = \frac{1}{3}(1, 2)$ e distribuições de estado dependente dada por

$$\begin{aligned} \mathbb{P}(Z_t = z | X_t = 1) &= \frac{1}{2},\ z \in S \\ \mathbb{P}(Z_t = 1 | X_t = 2) &= 1. \end{aligned}$$

Utilizando a Lei de Probabilidade Total e a Definição 4.1, podemos calcular $\mathbb{P}(Z_1 = 1, Z_2 = 1, Z_3 = 1)$ como segue

$$\begin{aligned} \mathbb{P}(Z_1{=}1, Z_2{=}1, Z_3{=}1) &= \sum_{x=1}^{2}\sum_{y=1}^{2}\sum_{w=1}^{2} \mathbb{P}(X_1{=}x, X_2{=}y, X_3{=}w, Z_1{=}1, Z_2{=}1, Z_3{=}1) \\ &= \sum_{x=1}^{2}\sum_{y=1}^{2}\sum_{w=1}^{2} \mathbb{P}(X_1{=}x)\mathbb{P}(X_2{=}y|X_1{=}x)\mathbb{P}(X_3{=}w|X_2{=}y) \\ &\quad \times \mathbb{P}(Z_1{=}1|X_1{=}x)\mathbb{P}(Z_2{=}1|X_2{=}y)\mathbb{P}(Z_3{=}1|X_3{=}k) \\ &= \sum_{x=1}^{2}\sum_{y=1}^{2}\sum_{w=1}^{2} \delta_x p_{xy} p_{yw} f_x(1) f_y(1) f_w(1). \end{aligned}$$

A Tabela 4.1 a seguir apresenta os cálculos parciais para a obtenção da probabiidade $\mathbb{P}(Z_1 = 1, Z_2 = 1, Z_3 = 1)$.

x	y	w	δ_x	p_{xy}	p_{yw}	$f_x(1)$	$f_y(1)$	$f_w(1)$	*Produto*
1	*1*	*1*	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{96}$
1	*1*	*2*	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{96}$
1	*2*	*1*	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{96}$
1	*2*	*2*	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{2}$	*1*	*1*	$\frac{6}{96}$
2	*1*	*1*	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{2}$	*1*	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{96}$
2	*1*	*2*	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{2}$	*1*	$\frac{1}{2}$	*1*	$\frac{2}{48}$
2	*2*	*1*	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{1}{4}$	*1*	*1*	$\frac{1}{2}$	$\frac{6}{96}$
2	*2*	*2*	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{3}{4}$	*1*	*1*	*1*	$\frac{18}{48}$
								Soma	$\frac{29}{48}$

Tabela 4.1: Cálculos dos valores $\mathbb{P}(X_1 = x, X_2 = y, X_3 = w, Z_1 = 1, Z_2 = 1, Z_3 = 1)$, em que $x, y, w \in \{1, 2\}$.

Pela Tabela acima é visto que $\mathbb{P}(Z_1 = 1, Z_2 = 1, Z_3 = 1) = \frac{29}{48}$. É possível notar ainda que o maior elemento da última coluna é $\frac{18}{48}$.

Um Modelo Oculto de Markov não é um Processo de Markov

Um Modelo Oculto Markov não satisfaz, em geral, a propriedade de Markov. Apresentamos um caso em que falha esta propriedade utilizando o Exemplo 4.2. Neste caso, sabemos que $\mathbb{P}(Z_1 = 1, Z_2 = 1, Z_3 = 1) = \frac{29}{48}$. Usando a Lei de Probabilidade Total, com o auxílio da tabela fornecida neste exemplo, podemos calcular o seguinte.

$$\begin{aligned}\mathbb{P}(Z_2 = 1) &= \sum_{x=1}^{2} \mathbb{P}(Z_2 = 1|X_t = x)\mathbb{P}(X_t = x) \\ &= \sum_{x=1}^{2} f_x(1)\delta_x = \frac{1}{2}\frac{1}{3} + 1\frac{2}{3} = \frac{5}{6}.\end{aligned}$$

Usando novamente a Lei de Probabilidade Total,

$$\begin{aligned}\mathbb{P}(Z_1 = 1, Z_2 = 1) &= \sum_{x=1}^{2}\sum_{y=1}^{2} \mathbb{P}(X_1 = x, X_2 = y, Z_1 = 1, Z_2 = 1) \\ &= \sum_{x=1}^{2}\sum_{y=1}^{2} \mathbb{P}(X_1 = x)\mathbb{P}(X_2 = y|X_1 = x)\mathbb{P}(Z_1 = 1|Z_1 = i)\mathbb{P}(Z_2 = 1|X_2 = y) \\ &= \sum_{x=1}^{2}\sum_{y=1}^{2} \delta_x p_{xy} f_x(1) f_y(1).\end{aligned}$$

x	y	δ_x	p_{xy}	$f_x(1)$	$f_y(1)$	Produto
1	1	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{24}$
1	2	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{2}{24}$
2	1	$\frac{2}{3}$	$\frac{1}{4}$	1	$\frac{1}{2}$	$\frac{2}{24}$
2	2	$\frac{2}{3}$	$\frac{3}{4}$	1	1	$\frac{12}{24}$
					Soma	$\frac{17}{24}$

Assim,

$$\mathbb{P}(Z_1 = 1, Z_1 = 1) = \frac{17}{24}.$$

Analogamente,

$$\mathbb{P}(Z_2 = 1, Z_3 = 1) = \frac{17}{24}$$

Então

$$\mathbb{P}(Z_3 = 1|Z_2 = 1, Z_1 = 1) = \frac{\mathbb{P}(Z_3 = 1, Z_2 = 1, Z_1 = 1)}{\mathbb{P}(Z_2 = 1, Z_1 = 1)} = \frac{\frac{29}{48}}{\frac{17}{24}} = \frac{29}{34}.$$

$$\mathbb{P}(Z_3 = 1|Z_2 = 1) = \frac{\mathbb{P}(Z_2 = 1, Z_3 = 1)}{\mathbb{P}(Z_2 = 1)} = \frac{\frac{17}{24}}{\frac{5}{6}} = \frac{17}{20}.$$

Uma vez que $\frac{29}{34} = \mathbb{P}(Z_3 = 1|Z_2 = 1, Z_1 = 1) \neq \mathbb{P}(Z_3 = 1|Z_2 = 1) = \frac{17}{20}$, a propriedade markoviana não é satisfeita. A seguir apresentaremos a função de verossimilhança para os Modelos Ocultos de Markov.

Considere um Modelo Oculto de Markov com distribuição inicial $\boldsymbol{\delta}$ e uma sequência de observações $(z_1, \cdots, z_T) \in \mathbb{R}^T$ supostamente gerada por este modelo. Pode-se obter sua verossimilhança a partir da proposição a seguir. Note que L_T é função das probabilidades de transição p_{xy} pertencentes a matriz de transição $\boldsymbol{\mathcal{P}}$ da Cadeia de Markov $(X_t)_{t\in\mathbb{N}^*}$, com espaço de estados $S = \{1, \cdots, m\}$ associada ao modelo e dos parâmetros relativos às funções densidade (ou massa) de probabilidade componentes do Modelo Oculto de Markov $(Z_t)_{t\in\mathbb{N}^*}$

Proposição 4.1. Seja $(X_t)_{t\in\mathbb{N}^*}$ a Cadeia de Markov, com espaço de estados $S = \{1, \cdots, m\}$, associada ao Modelo Oculto de Markov $(Z_t)_{t\in\mathbb{N}^*}$. O qual é gerador da seguinte sequência $(z_1, \cdots, z_T) \in \mathbb{R}^T$. Seja $\boldsymbol{\delta}$ é a distribuição inicial de $(X_t)_{t\in\mathbb{N}^*}$. A função de verossimilhança $L_T(\boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{z})$ é dada por

$$L_T(\boldsymbol{\delta}, \boldsymbol{\theta}, \mathbf{z}) = \boldsymbol{\delta}\boldsymbol{\mathcal{D}}(z_1)\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_2)\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_3)\cdots\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_T)\mathbf{1}',$$

em que $\mathbf{1}'$ é a matriz coluna com m coordenadas "iguais a um".

Se $\boldsymbol{\delta}$ for a distribuição estacionária, então

$$L_T = \boldsymbol{\delta}\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_1)\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_2)\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_3)\cdots\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_T)\mathbf{1}'.$$

Demonstração. Seja $\{Z_t : t \in \mathbb{N}^*\}$ um Modelo Oculto de Markov e $z_1, z_2, \cdots, z_T$ uma sequência de observações geradas por este modelo. Aplicando a Lei de Probabilidade Total junto com a definição de Modelo Oculto de Markov, temos:

$$\begin{aligned}
\mathbb{P}(Z_1 = z_1, \cdots, Z_T = z_T) &= \sum_{x_1,\cdots,x_T=1}^{m} \mathbb{P}(X_1 = x_1, \cdots, X_T = x_T, Z_1 = z_1, \cdots, Z_T = z_T) \\
&= \sum_{x_1,\cdots,x_T=1}^{m} \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2|X_1 = x_1)\cdots\mathbb{P}(X_T = x_T|X_{T-1} = x_{T-1}) \\
&\quad \times \mathbb{P}(Z_1 = z_1|X_1 = x_1)\cdots\mathbb{P}(Z_T = z_T|X_T = x_T) \\
&= \sum_{x_1,\cdots,x_T=1}^{m} = \delta_{x_1}p_{x_1x_2}\cdots p_{x_{T-1}x_T}f_{x_1}(z_1)\cdots f_{x_T}(z_T) \\
&= \sum_{x_1,\cdots,x_T=1}^{m} = \delta_{x_1}f_{x_1}(z_1)p_{x_1x_2}f_{x_2}(z_2)\cdots p_{x_{T-1}x_T}f_{x_T}(z_T)
\end{aligned}$$

$$
\begin{aligned}
&= \begin{pmatrix} \delta_1 & \cdots & \delta_m \end{pmatrix}
\begin{pmatrix} f_1(z_1) & & \\ & \ddots & \\ & & f_m(z_1) \end{pmatrix}
\begin{pmatrix} p_{11} & & p_{1m} \\ & \ddots & \\ p_{m1} & & p_{mm} \end{pmatrix} \\
&\quad \begin{pmatrix} f_1(z_2) & & \\ & \ddots & \\ & & f_m(z_2) \end{pmatrix} \cdots
\begin{pmatrix} p_{11} & & p_{1m} \\ & \ddots & \\ p_{m1} & & p_{mm} \end{pmatrix} \\
&\quad \begin{pmatrix} f_1(z_T) & & \\ & \ddots & \\ & & f_m(z_T) \end{pmatrix}
\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\
&= \boldsymbol{\delta}\boldsymbol{\mathcal{D}}(z_1)\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_2)\cdots\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_T)\mathbf{1}'.
\end{aligned}
$$

Se $\boldsymbol{\delta}$ for a distribuição estacionária, então, decorrendo diretamente do fato de que $\boldsymbol{\delta}\boldsymbol{\mathcal{P}} = \boldsymbol{\delta}$, temos que

$$\mathbb{P}(Z_1 = z_1, \cdots, Z_T = z_T) = \boldsymbol{\delta}\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_1)\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_2)\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_3)\cdots\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_T)\mathbf{1}'$$

□

As propriedades assintóticas para o estimador de máxima verossimilhança de um Modelo Oculto de Markov são consideradas em Baum e Petrie (1966), Leurox (1992), Bickel e Ritov (1996) e Rydén (1994). Baum e Petrie (1996) provam resultados de consistência e normalidade assintótica destes estimadores para o caso no qual o modelo assume valores em um conjunto de estados finito. Leurox (1997) estabelece a consistência de forma geral. Bickel e Ritov (1996) estabelecem a normalidade assintótica local. Rydén (1994) propõe uma nova classe de estimadores e prova a normalidade assintótica sob certas condições de regularidade.

4.3 Estimação de Máxima Verossimilhança: maximização direta

Nós vimos que a função de verossimilhança de um Modelo Oculto de Markov com m componentes é dada por:

$$L_T = \mathbb{P}(Z_1 = z_1, \cdots, Z_T = z_T) = \boldsymbol{\delta}\boldsymbol{\mathcal{D}}(z_1)\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_2)\cdots\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_T)\mathbf{1}',$$

em que $\boldsymbol{\delta}$ é a distribuição estacionária (neste caso, $\mathbb{P}(X_1 = x) = \delta_x$), e $\boldsymbol{\mathcal{D}}(z)$ uma matriz diagonal $m \times m$ onde o $x-$ésimo elemento da diagonal é dado por $\mathbb{P}(Z_t = z|X_t = x) = f_x(z)$. Estaremos interessados em estimar os parâmetros das funções densidade (ou massa) de probabilidade $f_x(z), x \in \{1, \cdots, m\}$ e a distribuição $\boldsymbol{\delta}$.

Nós podemos calcular $L_T = \boldsymbol{\alpha_T}1'$ recursivamente por:

$$\boldsymbol{\alpha_1} = \boldsymbol{\delta}\boldsymbol{\mathcal{D}}(z_1)$$

e

$$\boldsymbol{\alpha_t} = \alpha_{t-1}\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_t), \text{ para } t = 2, 3, \cdots, T.$$

Se assumirmos a estacionariedade, isto é $\boldsymbol{\delta} = \boldsymbol{\delta}\boldsymbol{\mathcal{P}}$, então a relação recursiva será dada por:

$$\boldsymbol{\alpha}_0 = \boldsymbol{\delta}$$

e

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{D}}(z_t), \text{ para } t = 2, 3, \cdots, T.$$

Para estimar os parâmetros via maximização direta da função de verossimilhança será utilizada a estratégia de Zucchini and MacDonald (2009). Esta usa o escalonamento do vetor de probabilidades $\boldsymbol{\alpha_t}$. Defina para cada $t = 0, 1 \cdots, T$, o vetor $\boldsymbol{\phi_t} = \frac{\boldsymbol{\alpha_t}}{\omega_t}$, em que $\omega_t = \sum_{i=1}^{m} \alpha_t(i) = \alpha_t \mathbf{1'}$.

Primeiramente, veremos algumas consequências das definições de $\boldsymbol{\phi_t}$ e ω_t, assumindo $\boldsymbol{\delta}$ distribuição estacionária. Observe que

$$\omega_0 = \alpha_0 \mathbf{1'} = \boldsymbol{\delta}\mathbf{1'} = 1$$

$$\boldsymbol{\phi_0} = \boldsymbol{\delta}$$

$$\omega_t \phi_t = \omega_{t-1}\phi_{t-1}\boldsymbol{B}_t, \text{ em que } \boldsymbol{B}_t = \boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{P}}(z_t). \tag{4.7}$$

Assim, $L_T = \omega_T = \prod_{t=1}^{T} \frac{\omega_t}{\omega_{t-1}}$

Da equação 4.7 segue que

$$\omega_t = \omega_{t-1}\phi_{t-1}\boldsymbol{B}_t\mathbf{1'}.$$

E então nós concluímos que

$$L_T = \omega_T = \prod_{t=1}^{T} \frac{\omega_t}{\omega_{t-1}}$$

e

$$\log(L_T) = \sum_{t=1}^{T} \log\left(\frac{\omega_t}{\omega_{t-1}}\right) = \sum_{t=1}^{T} \log(\boldsymbol{\phi_{t-1}}\boldsymbol{B}_t\mathbf{1'}),$$

logo, queremos estimar os parâmetros das funções densidade (ou massa) de probabilidade $f_x(z), x \in \{1, \cdots, m\}$ e a distribuição $\boldsymbol{\delta}$.

O cálculo de $\log L_T$, logaritmo da *Verossimilhança* pode ser resumido na forma de um algoritmo. Note que $\boldsymbol{\mathcal{P}}$ e $\boldsymbol{\mathcal{P}}(z_t)$ são matrizes $m \times m$, $\boldsymbol{v}$ e $\boldsymbol{\phi_t}$ são vetores com m coordenadas, u é um escalar, e l é o escalar que acumula $\log L_T$. Assumindo $\boldsymbol{\delta}$, distribuição estacionária.

Defina

$$\boldsymbol{\delta} \longrightarrow \phi_0 \text{ e } 0 \longrightarrow l, \text{ para } t = 1, 2, \cdots, T$$

$$\boldsymbol{\phi_{t-1}}\boldsymbol{\mathcal{P}}\boldsymbol{\mathcal{P}}(z_t) \longrightarrow \boldsymbol{v}$$

$$\boldsymbol{v}1' \longrightarrow u$$

$$l + \log u \longrightarrow l$$

$$\frac{\boldsymbol{v}}{u} \longrightarrow \boldsymbol{\phi_t}$$

Retorne l

O valor $\log(L_T)$ requisitado será então dado pelo valor final de l. Veja Zucchini e MacDonald (2009, página 239-252) para a implementação deste algoritmo em código R.

Maximização sujeita à restrições

Considerando o caso de um modelo oculto de Markov do tipo Poisson, as restrições relevantes são:
i) As médias λ_x não podem ser negativas;
ii) As linhas da matriz $\boldsymbol{\mathcal{P}}$ somam 1 e, além disso cada uma deve ser formada por elementos p_{xy} tais que $0 < p_{xy} < 1$. Veja Definição 2.3 na página 6.

Para estimar os parâmetros λ_x e p_{xy} utilizaremos a função "nlm" a qual baseia-se em parâmetros não limitados. Sendo assim, é necessário estabelecer as seguintes transformações:

$$\eta_x = \log(\lambda_x), \text{ para as médias e}$$

$$\tau_{xy} = \log\left(\frac{p_{xy}}{p_{xx}}\right), \text{ para as probabilidades de transição.}$$

Para ilustrar esta situação, consideremos o caso em que $x, y \in \{1, 2, 3\}$. Neste caso define-se, respectivamente o vetor e a matriz seguintes:

$$\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3) = (\log\lambda_1, \log\lambda_2, \log\lambda_3)$$

$$\boldsymbol{T} = \begin{pmatrix} - & \tau_{12} & \tau_{13} \\ \tau_{21} & - & \tau_{23} \\ \tau_{31} & \tau_{32} & - \end{pmatrix} = \begin{pmatrix} - & \log\frac{p_{12}}{p_{11}} & \log\frac{p_{13}}{p_{11}} \\ \log\frac{p_{21}}{p_{22}} & - & \log\frac{p_{23}}{p_{22}} \\ \log\frac{p_{31}}{p_{33}} & \log\frac{p_{32}}{p_{33}} & - \end{pmatrix}$$

Note que os elementos da diagonal principal não são colocados, pois a restrição $\sum_{x=1}^{3} p_{xy} = 1$, torna desnecessário a estimação de três parâmetros. Uma vez que estimando, por exemplo p_{12} e p_{13}, podemos obter a estimativa de p_{11} por $p_{11} = 1 - p_{12} - p_{13}$. λ_x ep_{xy} são chamados *parâmetros do processo*, enquanto η_x e τ_{xy} são chamados *parâmetros de trabalho.*

Para retornar aos parâmetros do processo utilizaremos as seguintes relações:

$$\lambda_x = e^{\eta_x},$$

$$p_{xx} = \frac{1}{1 + \sum_{y \neq x} e^{\tau_{xy}}} \text{ e}$$

$$p_{xy} = \frac{e^{\tau_{xy}}}{1 + \sum_{y \neq x} e^{\tau_{xy}}}.$$

O leitor interessado na obtenção do resultado acima pode consultar Aitchison (1982).

Assim, a estimação dos parâmetros pode ser feita via maximização numérica da função de verossimilhança. Porém existem alguns problemas que podem ocorrer neste cálculo, por exemplo "*overflow*", isto é a *Verossimilhança* aproxima-se de zero rapidamente. A seguir será discutida algumas técnicas que contornam este problema.

Iniciando Valores para as Iterações

É possível estabelecer funções em R que estimam os parâmetros de um Modelo Oculto de Markov do tipo Poisson, estacionário a m componentes, veja Zucchini and Macdonald (2009, página 239-252). Nesta mesma bibliografia, na página 50, há uma seção que nos mostra como deve ser a inicialização de valores para um Modelo Oculto de Markov de 2 e 3 componentes sobre como iniciar os valores, para dois e três componentes, nestes algoritmos.

Primeiramente o vetor de médias λ: para o caso em que $m = 2$, faça a média amostral dos valores observados e estabeleça valores próximos e equidistantes para esta média. Para o caso em que $m = 3$, utilize como valores iniciais o quartil inferior, a mediana e o quartil superior dos valores observados.

Já no caso das estimativas das probabilidades de transição p_{xy} utilize valores comuns (por exemplo 0.01 ou 0.05), para os valores que não estão na diagonal principal da matriz formada pelas probabilidades de transição. A seguir, mostramos como estimar os parâmetros de um Modelo Oculto de Markov, estacionário a 2 e a 3 componentes.

Exemplo 4.3. Utilizaremos os dados amostrais do exemplo 3.2 na página 32. Na seção A.3 mostramos como utilizar as funções que estimam os parâmetros de um modelo Modelo Oculto de Markov do tipo Poisson com duas e três componentes.

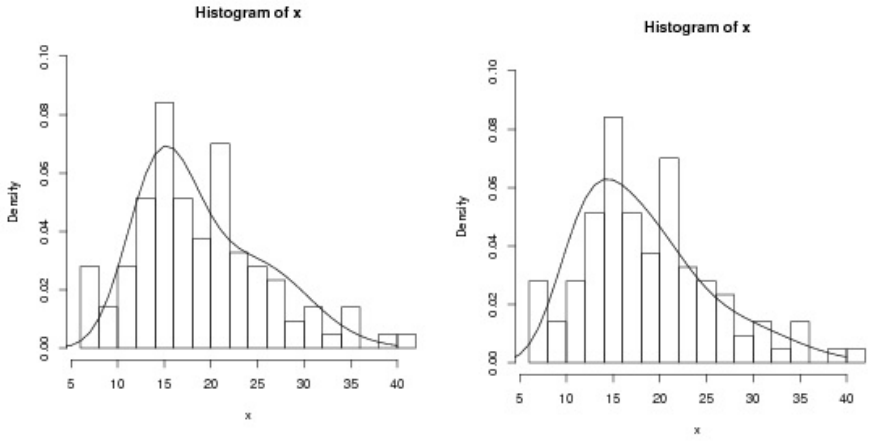
A Tabela 4.2 a seguir apresenta os valores calculados para $-logL_T$ e dos critérios de seleção de modelos AIC e BIC para os Modelos Ocultos de Markov e modelos de mistura de distribuições independente para $m \in \{2, 3, 4\}$. Os valores são comparados como na Tabela 3.1.

Modelo	k	$-\log L_T$	AIC	BIC
Mistura Independente ($m = 2$)	3	360,369	726,7	734,8
Mistura Independente ($m = 3$)	5	356,8489	723,7	737,1
Mistura Independente ($m = 4$)	7	356,7337	727,5	746,2
Modelo Oculto de Markov ($m = 2$)	4	342,3183	692,6	703,3
Modelo Oculto de Markov ($m = 3$)	9	329,4603	**676,9**	**701**

Tabela 4.2: valores dos critérios de seleção de modelos AIC e BIC e de $-\log L_T$, para $m \in 2, 3, 4$, em que k é o número de parâmetros a serem estimados.

Pelos critérios de seleção de modelos AIC e BIC o melhor modelo ajustado para os dados é o Modelo Oculto de Markov com $m = 3$.

A seguir os gráficos das densidades dos Modelos Ocultos de Markov junto ao histograma das observações.



(a) **(b)**

Figura 4.2: Dados sobre tremores de terra: histograma comparado às densidades dos Modelos Ocultos de Markov, estacionário, do tipo Poisson. (a) $m = 2$; (b) $m = 3$.

Capítulo 5

Aplicação

Neste capítulo apresentaremos uma aplicação dos modelos de Cadeia de Markov, Capítulo 2. Modelo de Mistura de Distribuições Independente, Capítulo 3 e o modelo oculto de Markov, Capítulo 4. Os dados apresentados nesta seção são *Old Faithful Geyser* (ver Zucchini e MacDonald 1997, p. 207), de tamanho $n = 299$, e são apresentados na Tabela A.2 na página 57. É importante observar que nesta bibliografia não são utilizadas modelos de mistura do tipo Poisson, mas modelos do tipo Binomial. Neste capítulo o nosso objetivo não é encontrar o "melhor modelo", mas sim aplicar as metodologias apresentadas no decorrer deste trabalho.

A amostra baseia-se no registro de 299 erupções sucessivas de uma antigo Geyser em atividade. O período de análise está compreendido entre os dias 1° e 15 de agosto de 1985. As erupções de curta duração são denotadas por zero, e as erupções de longa duração serão denotadas por 1. Iremos apresentar os critérios de seleção de modelos AIC e BIC de cada modelo e ao final iremos compará-los e escolheremos o modelo de menor AIC e BIC. Lembre que:

$$AIC = -2 \log L + 2p,$$

em que p é o número de parâmetros a ser estimado. E que

$$BIC = -2 \log L + p \log T,$$

em que T é o número de observações.

A seguir a modelagem dos dados por Cadeia de Markov.

Cadeia de Markov

Suponha que o conjunto de dados em questão seja uma realização (caminho) de uma Cadeia de Markov a dois estados, estacionária. A partir dos estimadores de máxima verossimilhança (ver equação (2.5), p. 13) e da função de Verossimilhança $L(\boldsymbol{\theta}, \boldsymbol{x}) = L((p_{00}, ..., p_{MM}), (x_1, ..., x_t)) = \mu_{x_0} \prod_{x,y=0}^{M} p_{xy}^{t_{xy}}$ da página 12, pode-se calcular o conjunto de dados apresentados na tabela a seguir.

x	y	t_{xy}	t_x	$\hat{p}_{xy}$	p	$-\log L$	AIC	BIC
0	0	0	105	0	2	-135,2378175	274,475635	281,8765222
0	1	104	105	0,99047619				
1	0	105	194	0,541237113				
1	1	89	194	0,458762887				

Tabela 5.1: estimativa dos parâmetros de um modelo baseado em uma Cadeia de Markov estacionária a dois estados

A seguir a modelagem dos dados pelo Modelo de Mistura de Distribuições Independente.

Modelo de Mistura de Distribuições Independente

Supondo que o conjunto de dados em questão são não-correlacionados, utilizando o programa MMIX da seção A.1 obtemos as seguintes estimativas para uma mistura de distribuições do tipo Poisson para $m = 2, 3$ componentes.

m	i	δ_i	λ_i	p	$-\log L$	AIC	BIC
2	1	0,443	0,648	3	277,922	557,843	561,543
	2	0,556	0,648				
3	1	0	1,118	5	288,041	586,082	604,584
	2	0,133	0,648				
	3	0,866	0,648				

Tabela 5.2: estimativa dos parâmetros para Modelos de Mistura de Distribuições Independente do tipo Poisson a duas e três componentes.

A seguir a modelagem de dados pelo Modelo Oculto de Markov.

Modelo Oculto de Markov

Suponha que a sequência de dados observáveis é gerada por um processo oculto de Markov. Pode-se, a partir das funções de Zucchini e MacDonald (2009, página 239-252), estimar os parâmetros do modelo oculto de Markov para $m = 2, 3$ componentes. Veja a seguinte tabela.

m	i	p	δi	λi	$-\log L$	AIC	BIC
2	1	4	0,545	0,648	277,921	563,843	578,644
	2		0,454	0,648			
3	1	9	0,244	0,648	277,921	573,843	607,147
	2		0,355	0,648			
	3		0,400	0,648			

Tabela 5.3: estimativa dos parâmetros para Modelos Ocultos de Markov do tipo Poisson a duas e a três componentes.

A Tabela a seguir faz um comparativo entre os três modelos a partir dos valores do AIC e BIC.

Modelo	AIC	BIC
Cadeia de Markov	**274,475**	**281,876**
Mistura Independente a 2 componentes	557,843	561,543
Mistura Independente a 3 componentes	586,082	604,584
Modelo Oculto de Markov a 2 componentes	563,843	578,644
Modelo Oculto de Markov a 3 componentes	573,843	607,147

Tabela 5.4: comparação entre os modelos a partir dos critérios de seleção de modelos AIC e BIC.

Note que o modelo a ser selecionado é o que pressupõe que os dados são uma realização (caminho) de uma Cadeia de Markov a dois estados, estacionária. Uma vez que este, é o modelo que possui menor AIC e BIC.

Capítulo 6

Conclusões

No Capítulo 2 apresentamos os conceitos iniciais sobre a teoria das Cadeias de Markov a tempo discreto e espaço de estados finito. Além disso, apresentamos resultados relacionados à inferência estocástica, formalizamos uma série de conceitos e apresentamos a prova detalhada de diversas proposições e Lemas. Apresentamos o estimador de verossimilhança para uma Cadeia de Markov e, para o caso em que esta é ergódica, discutimos a prova detalhada da consistência e da normalidade assintótica de tal estimador.

No Capítulo 3 apresentamos o Modelo de Mistura de Distribuições Independente. Provamos que a expressão que define o modelo de mistura satisfaz as propriedades de uma função densidade (ou massa) de probabilidade. Apresentamos uma expressão geral para a variância de uma variável aleatória com distribuição de mistura independente. Estimamos os parâmetros de tal modelo quando aplicado à modelagem dos registros de tremores de terra, apresentados no Apêndice A.1.

No Capítulo 4 apresentamos os Modelos Ocultos de Markov, suas distribuições marginais e a estimação de parâmetros pelo método da máxima verossimilhança. Ao final deste capítulo foi feita a seleção do "melhor modelo"através dos critérios de seleção de modelos AIC e BIC. Comparados inclusive às modelagens feitas no Capítulo 3.

No Capítulo 5 apresentamos uma aplicação que utiliza os três tipos de modelagem abordadas neste trabalho: Cadeia de Markov, Modelos de Mistura de Distribuições Independente e Modelos Ocultos de Markov. Além disso escolhemos o modelo mais adequado utilizando os critérios de seleção de modelos AIC e BIC.

Ao longo deste trabalho houve uma preocupação em apresentar de forma clara e rigorosa alguns resultados da teoria de Inferência Estocástica e Distribuições de Mistura. Além disso, sempre que possível, buscamos exemplos de aplicação da teoria apresentada, muitas vezes, valendo-se de programas produzidos a partir do ambiente R.

Bibliografia

[1] Aitchison, J. *The statistical analysis of compositional data.* 1982. J. Roy. Statist. Soc. B 44, 139-177.

[2] Atuncar, G. S.-*Conceitos Básicos de Processos Estocásticos.* Minas Gerais; 2009. Departamento de Estatística, Universidade Federal de Minas Gerais.

[3] Baum, L. E., Petrie, T.-*Statistical Inference for Probabilistic Functions of Finite State Markov Chains.* 1966. Ann. Math. Statist. 37, 1554-1563.

[4] Bickel, P. J., Ritov, Y.-*Inference in Hidden Markov Models I: Local asymptotic normality in the stationary case.* 1996. Bernoulli 2, 199-228.

[5] Billard, L.-*A Voyage of discovery.* 1997. Journal of the American Satistical Association.

[6] Billingsley, P.-*Inference for Markov Processes.* Chicago; 1961. University of Chicago Press.

[7] Douc, R., Molines, E., Olsson, J., Handel, R.-*Consistency of the Maximum Likelihood Estimator for General Hidden Markov Models.* 2011. The Annals of Statistics, volume 39, N° 1, 474-513.

[8] Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G.-*Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge; 1998. Cambridge University Press.

[9] Falkner, R. P.-*The theory and Practice of Price Statistics.* 1892. Journal of the American Statistical Association.

[10] Galton, F.-*Heredity Genius: An Inquiry into Its Laws and Consequences.* Londres; 1869. Macmillan.

[11] Guttorp, P.-*Stochastic Modeling of Scientific Data.* Seattle; 1995. Chapman and Hall.

[12] Holmes, G. K.-*Measures of Distribution.* 1892. Journal of the American Statistical Association 3, 141-157.

[13] James, B. R.-*Probabilidade: um curso em nível intermediário.* Rio de Janeiro; 2008. Coleção Projeto Euclides, IMPA.

[14] Karlin, S. and Taylor, H. M.-*A First Course in Stochastic Processes.* New York; 1975. Academic Press.

[15] Leroux, B.G. -*Maximum-penalized-likelihood Estimation for Hidden Markov Models.*. 1992. Stoch. Processes Appl. 40, 127-143.

[16] Leroux, B.G. and Puterman, M.L.-*Maximum-penalized-likelihood Estimation for Independent and Markov-dependent Mixture Models.* 1992. Biometrics 48, 545-558.

[17] McLachlan, G. J. and Peel, D.-*Finite Mixture Models.* New York; 2000. Wiley Series in Probability and Statistics.

[18] Pearson, K.-*Contributions to the Theory of Mathematical Evolution.* 1894. Philosophical Transactions of the Royal Society of London.

[19] Quetelet, A.-*Lettres à S.A.R. le Duc Régnant de Saxe-Cobourg an Gotha, la théorie des probabilité, appliquée aux sciences morales et-politiques.* 1846. Brussels: Hayes

[20] Quetelet, A.- *Sur quelques propritiétés curieuses que présentent les résultats d'une serie d'observations, faites dans la vue de déterminer une constante, lorsque les chances de rencontrer des écarts en plus et en moins sont égales et independantés les unes des autres.* 1852. Bulletins de l'Académie royale des sciences, des lettres et des beaux-arts de Belgique 19, 303-317,.

[21] Norris, J. R.-*Markov Chains.* New York; 2004. Cambridge Series on Statistical and Probabilistic Mathematics, Cambridge University.

[22] Rabiner, L. R.-. *A tutorial on hidden markov models and selected applications in speech recognition.* February 1989. Proceedings of the IEEE, Vol. 77 number 2.

[23] Rydén, T.-*Consistent and asymptotically normal parameter estimates for hidden Markov models.* 1994. Ann. Statist. 22, 1884-1895.

[24] Rohatgi, V. K.-*An Introduction to Probability Theory and Mathematical Statistics.* New York; 1976. Wiley Series in Probability and Statistics.

[25] Stigler, S. M.-*The History of Statistics.* Cambridge, Massachusetts; 1986. Belknap Press of Harward University Press.

[26] Zucchini, W. and MacDonald, I. L.- *Hidden Markov and Other Models for Discrete-valued Time Series.* Londres, 1997. Chapman and Hall.

[27] Zucchini, W. and MacDonald, I. L.-*Hidden Markov Models for Time Series: an introduction using R.* Londres, 2009. Chapman and Hall.

Apêndice A

Neste apêndice iremos apresentar o programa MMIX, desenvolvidos no ambiente R. Os dados amostrais que foram utilizados no decorrer do trabalho. E por último veremos como utilizar os códigos para estimar os parâmetros de um Modelo Oculto de Markov do tipo Poisson, estacionário, a duas e a três componentes.

A.1 Programa MMIX

Nesta seção será apresentado o programa MMIX que foi utilizado para estimar os parâmetros da *função de verossimilhança* obtida para as distribuições de mistura do exemplo 3.2 na página 32.

```
"MMIX"<-function(serie,m)
{ #Independent Mixture Models
#Poisson Distributions
tab<-list(serie=serie,m=m)
#START POINT
if(m==1){
tmp<-10
inf<-0
sup<-30
}
if(m==2){
tmp<-c(10,20,0.5)#diferentes nas 2 primeiras posições
inf<-c(0,0,0)
sup<-c(30,30,1)
}
if(m==3){
tmp<-c(10,15,25,0.5,0.3)#diferentes nas 3 primeiras posições
inf<-c(0,0,0,0,0)
sup<-c(40,40,40,1,1)
}
if(m==4){
tmp<-c(10,15,25,30,0.1,0.3,0.5)#diferentes nas 4 primeiras posições
inf<-c(0,0,0,0,0,0,0)
sup<-c(40,40,40,40,1,1,1)
}
#Maximized Likelihood
```

```
xp<-nlminb(start=tmp, obj=LIKEMMIXM, tab=tab,lower=inf,upper=sup)$parameters
Result(xp)

}
"LIKEMMIXM"<-function(x,tab)
{
#Likelihood Function
y<-tab$serie
m<-tab$m
if(m==1){plike<--sum(log((((x[1])^y)*exp(-x[1]))/factorial(y)))}
if(m==2){
plike<--sum(log(x[3]*(((x[1])^y)*exp(-x[1]))/factorial(y)+
(1-x[3])*(((x[2])^y)*exp(-x[2]))/factorial(y)))
}
if(m==3){
plike<--sum(log(x[4]*(((x[1])^y)*exp(-x[1]))/factorial(y)+
x[5]*(((x[2])^y)*exp(-x[2]))/factorial(y)+
(1-x[4]-x[5])*(((x[3])^y)*exp(-x[3]))/factorial(y)))
}
if(m==4){
plike<--sum(log(x[5]*(((x[1])^y)*exp(-x[1]))/factorial(y)+
x[6]*(((x[2])^y)*exp(-x[2]))/factorial(y)+x[7]*(((x[3])^y)
*exp(-x[3]))/factorial(y)+
(1-x[5]-x[6]-x[7])*(((x[4])^y)*exp(-x[4]))/factorial(y)))
}
return(plike)
}

"LIKEMIXPLOT"<-
function(serie)
{
#Likelihood Function Plot
tab<-list(serie=serie)
lambda1<-seq(0,30,by=0.5)
lambda2<-seq(0,30,by=0.5)
t<-length(lambda1)
Like<-matrix(nrow=t^2,ncol=3)
k<-1
for(i in 1:t){
for(j in 1:t){
x<-c(lambda1[i],lambda2[j],0.6)
Like[k,]<-c(lambda1[i],lambda2[j],LIKEMMIX(x,tab))
k<-k+1
}
}
```

```
return(Like)
}
"Result"<-
function(x)
{
#Result Return
m<-(length(x)+1)/2
if(m==1){
lambda<-x
delta<-1
media<-x
v<-x
}
if (m>=2){
lambda<-x[1:m]
delta<-c(x[(m+1):length(x)],1-sum(x[(m+1):length(x)]))
media<-sum(lambda*delta)
v<-sum(delta*(lambda+lambda^2))-(media)^2

}
return(list(lambda=lambda,delta=delta,mean=media,var=v))
}
```

A.2 Dados Amostrais

Esta seção tem por objetivo apresentar alguns dos conjuntos de dados que foram utilizados no decorrer do trabalho.

O conjunto de dados da Tabela A.1 foi obtido seguindo o seguinte procedimento: a cada ano, entre 1900-2006, registrou-se quantos foram os tremores de terra, no planeta, de magnitudes maiores ou iguais a sete (ver Zucchini e MacDonald 2009, página 12).

13	14	8	10	16	26	32	27	18	32	36	24	22	23	22	18	25	21	21	14
8	11	14	23	18	17	19	20	22	19	13	26	13	14	22	24	21	22	26	21
23	24	27	41	31	27	35	26	28	36	39	21	17	22	17	19	15	34	10	15
22	18	15	20	15	22	19	16	30	27	29	23	20	16	21	21	25	16	18	15
18	14	10	15	8	15	6	11	8	7	18	16	13	12	13	20	15	16	12	18
15	16	13	15	16	11	11													

Tabela A.1: Registro dos tremores de terra de magnitude maiores ou iguais a sete que ocorreram entre 1900-2006.

A Tabela A.2 nos fornece o registro de 299 erupções sucessivas de um antigo geyser em atividade (ver Zucchini e MacDonald 1997, página 207). O período de análise está compreendido

entre os dias 1° e 15 de agosto de 1985. As erupções de curta duração são denotadas por zero, e as longas por 1.

1	0	1	1	1	0	1	1	0	1	0	1	0	1	1
0	1	0	1	1	0	1	0	1	0	1	0	1	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0
1	0	1	0	1	0	1	0	1	0	1	1	1	1	1
0	1	0	1	0	1	0	1	1	0	1	0	1	1	1
0	1	1	1	1	1	0	1	1	1	0	1	0	1	0
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
0	1	0	1	0	1	1	0	1	0	1	0	1	0	1
0	1	1	1	0	1	1	1	1	1	1	1	0	1	1
1	1	1	0	1	1	1	1	1	1	1	0	1	0	1
0	1	0	1	0	1	0	1	1	1	1	1	1	0	1
0	1	0	1	0	1	1	1	0	1	0	1	0	1	1
0	1	0	1	1	1	1	0	1	0	1	0	1	0	1
1	1	0	1	0	1	0	1	1	0	1	1	0	1	1
1	0	1	0	1	0	1	0	1	1	0	1	1	1	1
1	1	1	0	1	0	1	0	1	1	1	1	0	1	1
0	1	1	1	0	1	1	0	1	0	1	1	1	0	1
0	1	1	1	1	1	0	1	1	1	0	1	0	1	0
1	1	0	1	0	1	1	1	1	1	1	1	1	0	1
0	1	0	1	0	1	0	1	0	1	0	1	1	0	

Tabela A.2: Antigo Geyser em Atividade.

A.3 Aplicação de algoritmos para o Modelo Oculto de Markov

Vejamos como utilizar os códigos, desenvolvidos em ambiente R, para estimar os parâmetros de um Modelo Oculto de Markov do tipo Poisson, estacionário, a duas e a três componentes. As funções utilizadas estão em Zucchini e MacDonald (2009, página 239).

Modelo Oculto de Markov do tipo Poisson com duas componentes

Inicialmente, estabeleça os seguintes valores.

```
m<-2
lambda<-c(18,22)
gamma<- matrix(c(0.99,0.01, 0.01,0.99), nrow = 2, ncol=2)
```

```
Carregue a função pois.HMM.pn2pw e rode-a: pois.HMM.pn2pw(m,lambda,gamma).
```

```
Será retornado: [1]  2.890372  3.091042 -4.595120 -4.595120
```

```
Escreva: parvect<-c(2.890372,3.091042,-4.595120,-4.595120).
```

Defina o vetor x com os dados do exemplo 3.2 na página 32.

```
Carregue a função pois.HMM.pw2pn, carregue a função pois.HMM.mllk e rode-a:

pois.HMM.mllk(parvect,x,m). Será retornado: [1] 369.5683.

Defina o vetor mllk<-369.5683, lambda0<-lambda, gamma0<-gamma.

Carregue a função pois.HMM.mle, e rode-a: pois.HMM.mle(x,m,lambda0,gamma0).
Será retornado:

$lambda
[1] 15.47223 26.12534

$gamma
          [,1]       [,2]
[1,] 0.9340403 0.06595975
[2,] 0.1285076 0.87149239

$delta
[1] 0.6608184 0.3391816

$code
NULL

$mllk
[1] 342.3183

$AIC
[1] 692.6365

$BIC
[1] 703.3278
```

Isto é, encontramos as seguintes estimativas

$$\hat{\boldsymbol{\mathcal{P}}} = \begin{pmatrix} 0.9340403 & 0.06595975 \\ 0.1285076 & 0.87149239 \end{pmatrix}$$

$$\hat{\boldsymbol{\delta}} = (0.6608184, 0.3391816)$$

$$\hat{\boldsymbol{\lambda}} = (15.47223, 26.12534)$$

Modelo Oculto de Markov do tipo Poisson com três componentes

Seguindo os passos anteriores iniciando com os seguintes valores

```
m<-3
lambda<-c(14.75,23.5,32.25)
gamma<- matrix(c(0.9,0.05, 0.05,0.05,0.9,0.05,0.05,0.05,0.95), nrow = 3, ncol=3)
```

e fazendo algumas alterações convenientes chegamos as seguintes estimativas.

$$\hat{\boldsymbol{\mathcal{P}}} = \begin{pmatrix} 0.9546245 & 0.02444223 & 0.02093325 \\ 0.04976616 & 0.89936798 & 0.05086586 \\ 0 & 0.19664207 & 0.80335793 \end{pmatrix}$$

$$\hat{\boldsymbol{\delta}} = (0.4436, 0.4045, 0.1519)$$

$$\hat{\boldsymbol{\lambda}} = (13.14573, 19.72101, 29.71437)$$