Universidade Federal de São Carlos - UFSCar
Departamento de Estatística - DEs

Trabalho Final de Processos Estocásticos

Hélio Mota Ezequiel - 744855
Leticia A. Silva - 744860
Douglas Nestlehner- 752728
Eric Sato - 729739

Docente: Prof. Dr. Márcio Luis Lanfredi Viola

Conteúdo

1 Parte I: Construção do Modelo Markoviano de ordem 1

1.1 Situação problema de interesse utilizando cadeias de Markov.

A rotina de treino de uma pessoa é dividida em cinco partes: correr(1); pedalar(2); abdominal(3); flexão(4); e polichinelos(5), e a cada dia ela realiza apenas uma das atividades. Porém existem algumas limitações:

- Se a pessoa está correndo, então no dia seguinte ela irá correr, ou fazer abdominal, ou pedalar ou fazer flexão com a mesma probabilidade;
- Se a pessoa esta pedalando, portanto no dia seguinte ela irá fazer polichinelos ou correr ou flexão, todos com a mesma probabilidade;
- Se a pessoa estiver fazendo abdominal, por conta disso no dia seguinte ela irá pedalar ou fazer polichinelos;
- Se a pessoa estiver fazendo flexão, logo no dia seguinte ela irá, com 50% de chance, fazer abdominais ou decidirá entre flexão ou polichinelos com a mesma probabilidade;
- Se estiver fazendo polichinelos, então no dia seguinte ela irá, com a mesma probabilidade, correr ou fazer abdominal ou pedalar.

1.2 Cadeia de Markov associada ao problema proposto, matriz de transição da cadeia e o diagrama de transição de estados.

Primeiramente, vamos definir o processo estocástico de interesse, seja X_n a variável que especifica a atividade a ser realizado no n-ésimo dia. O espaço de estados S do processo pode ser descrito por $S = \{1, 2, 3, 4, 5\}$, onde 1 indica correr, 2 pedalar, 3 abdominal, 4 flexão e 5 polichinelos.

Através da modelagem do processo, sabemos que a probabilidade de fazer determinada atividade específica no dia n+1, X_{n+1}, depende apenas da atividade que foi feita no dia anterior, ou seja, depende apenas de X_n. Por conta disso, podemos modelar esse processo através de uma cadeia de Markov com espaço de estado $S = \{1, 2, 3, 4, 5\}$.

Agora, para caracterizar a cadeia, precisamos definir a matriz de transição P, ou seja, precisamos definir cada probabilidade de transição $P_{i,j}$ com base na situação problema de interesse que foi proposta, onde i equivale a linha e j a coluna. A probabilidade de transição é definida por:

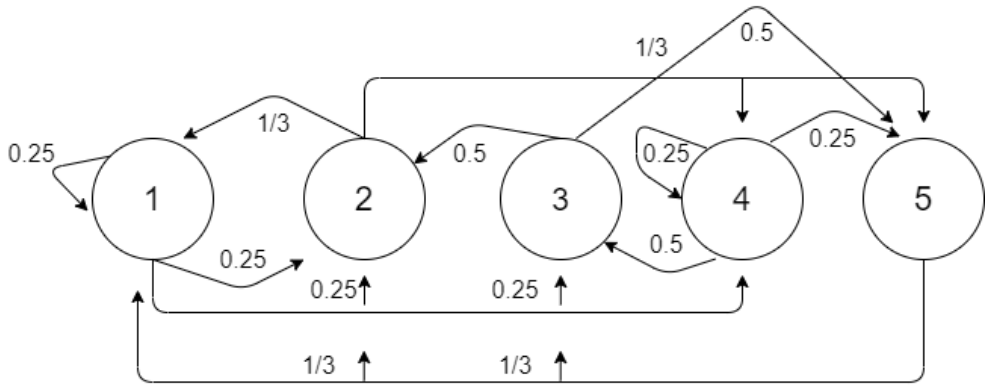
$$P_{i,j}(n) = \mathbb{P}\left(X_{n+1} = j | X_n = i\right)$$

onde $P_{ij}(n)$ é a probabilidade do processo ir para o estado j quando ele está no estado i no n-ésimo passo do processo.

Definindo cada probabilidade de transição através da nossa situação problema, chegamos a seguinte **Matriz de Transição:**

$$P = \begin{vmatrix} 0,25 & 0,25 & 0,25 & 0,25 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0,5 & 0 & 0 & 0,5 \\ 0 & 0 & 0,5 & 0,25 & 0,25 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \end{vmatrix}$$

Temos uma maneira visual para representar a matriz de transição P encontrada que é através do **Diagrama de transição de estados:**



Nesta representação, os círculos significam os estados da cadeia e as flechas que conectam os círculos representam as probabilidades de transição de um estado para o outro. O número em cima de cada flecha representa a probabilidade de transição entre os estados.

1.3 Classe de estados da cadeia de Markov e classificação de todos os estados da cadeia.

Analisando a cadeia para descobrir a(s) classe(S) de estados, vamos iniciar verificando quais estados são acessíveis um pelo outro. Assim,

- Os estados 1,2,3 e 4 são acessíveis pelo estado 1;
- Os estados 1,2,3 e 4 são acessíveis pelo estado 2;
- Os estados 1,2,3 e 4 são acessíveis pelo estado 3;
- Os estados 1,2,3 e 4 são acessíveis pelo estado 4;
- Os estados 1,2,3 e 4 são acessíveis pelo estado 5.

Podemos observar que todos os estados se comunicam entre si, portando temos apenas uma classe de estados definida por $S = \{1, 2, 3, 4, 5\}$. Desta forma, concluímos que esta é uma cadeia de Markov irredutível, visto que possui apenas uma única classe de estados.

Em seguida, classificaremos os estados da cadeia em recorrente, transitório ou absorvente.

- Correr (1): recorrente
- Pedalar (2): recorrente
- Abdominal (3): recorrente
- Flexão (4): recorrente
- Polichinelos (5): recorrente

Podemos classificar a cadeia em relação ao seu período, sendo:

- d(1): mdc(1,2,3,4,...)=1
- d(2): mdc(2,3,4,5,...)=1
- d(3): mdc(2,3,4,5,...)=1
- d(4): mdc(1,4,6,...)=1
- d(5): mdc(2,3,4,...)=1

Pode-se observar que esta é uma cadeia aperiódica, já que seus estados tem período igual a 1.

Apenas observando o resultado da classificação dos períodos, concluímos que esta cadeia é possível se obter a distribuição limite, já que, segundo o teorema visto em aula que diz que para a distribuição limite existir, ela precisa ser aperiódica e recorrente positiva. Desta maneira, de acordo com o teorema visto em aula, sabemos que se uma cadeia é irredutível, aperiódica e recorrente positiva, garantimos que

$$lim_{n\to\infty} P_{ij}^{n} = \pi_j$$

Portanto, conclui-se que existe a distribuição limite.

Assim como também, para a distribuição estacionária, verificamos que é possível encontrá-la.

1.4 Distribuição estacionária da cadeia.

Para obter a distribuição estacionária da cadeia, temos que resolver o seguinte cálculo através da equação de balanço geral $\pi P = \pi$. Assim, com a matriz de transição definida para este problema, temos que:

$$(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5) \begin{vmatrix} 0,25 & 0,25 & 0,25 & 0,25 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0,5 & 0 & 0 & 0,5 \\ 0 & 0 & 0,5 & 0,25 & 0,25 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \end{vmatrix} = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$$

Após resolvermos o sistema, encontramos os seguintes valores:

$$\pi = (\pi_1 = \tfrac{100}{513}, \pi_2 = \tfrac{2}{9}, \pi_3 = \tfrac{104}{513}, \pi_4 = \tfrac{28}{171}, \pi_5 = \tfrac{37}{171})$$

Escrevendo os resultados com 5 casas decimais, sem arredondamento:

$$\pi = (\pi_1 = 0,194931, \pi_2 = 0,222222, \pi_3 = 0,202729, \pi_4 = 0,163742, \pi_5 = 0,216374)$$

Com este resultado, obtemos que a probabilidade de o processo parar no estado 1 é igual a $0,194931$. Do mesmo modo, as probabilidades de o processo parar nos estados 2 é $0,222222$, no processo 3 é igual a $0,202729$. Assim, como nos processos 4 e 5, as probabilidades são, respectivamente, $0,163742$ e $0,216374$.

Com isso, podemos notar que o processo tem maior chance de parar no estado 2, ou seja, a pessoa tem maior chance de, na rotina de exercícios, estar pedalando no próximo dia e uma chance menor de estar fazendo flexão no dia seguinte.

1.5 Distribuição limite da cadeia.

Chamamos o limite $lim_{n\to\infty} P_{ij}^n$, quando ele existe e não depende de i, de **distribuição limite** do estado j. Como visto em aula, existe um resultado teórico garantindo que, para cadeias que são irredutíveis, aperiódicas e recorrente positivas, a distribuição limite tende para a distribuição estácionária da cadeia.

Logo, verificando o que acontece com a matriz de transição em n passos quando n cresce, ou seja, queremos estudar:

$$lim_{n\to\infty} \mathbb{P}(X_n = j | X_0 = i) = P_{ij}^n$$

Calculando P_{ij}^n para n = 5, 10 e 15:

$$P^5 = \begin{vmatrix} 0,194283 & 0,222400 & 0,203496 & 0,164374 & 0,215448 \\ 0,195129 & 0,220680 & 0,201808 & 0,164730 & 0,217653 \\ 0,194557 & 0,223749 & 0,201035 & 0,163202 & 0,217456 \\ 0,194111 & 0,221708 & 0,203955 & 0,162884 & 0,217342 \\ 0,196286 & 0,222604 & 0,203643 & 0,163316 & 0,214150 \end{vmatrix}$$

$$P^{10} = \begin{vmatrix} 0,194923 & 0,222226 & 0,202735 & 0,163744 & 0,216373 \\ 0,194931 & 0,222213 & 0,202723 & 0,163746 & 0,216386 \\ 0,194930 & 0,222229 & 0,202719 & 0,163744 & 0,216378 \\ 0,194937 & 0,222221 & 0,202731 & 0,163733 & 0,216377 \\ 0,194938 & 0,222223 & 0,202738 & 0,163744 & 0,216357 \end{vmatrix}$$

$$P^{15} = \begin{vmatrix} 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \\ 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \\ 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \\ 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \\ 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \end{vmatrix}$$

Portanto, após fazer os cálculos para o $n = 5$, 10 e 15, podemos verificar que quando o limite tende ao infinito, o resultado tende para a nossa distribuição limite, que será igual à distribuição estacionária. Podemos notar que com um $n = 15$, já conseguimos obter os resultados esperados, sendo eles os valores da nossa distribuição estacionária.

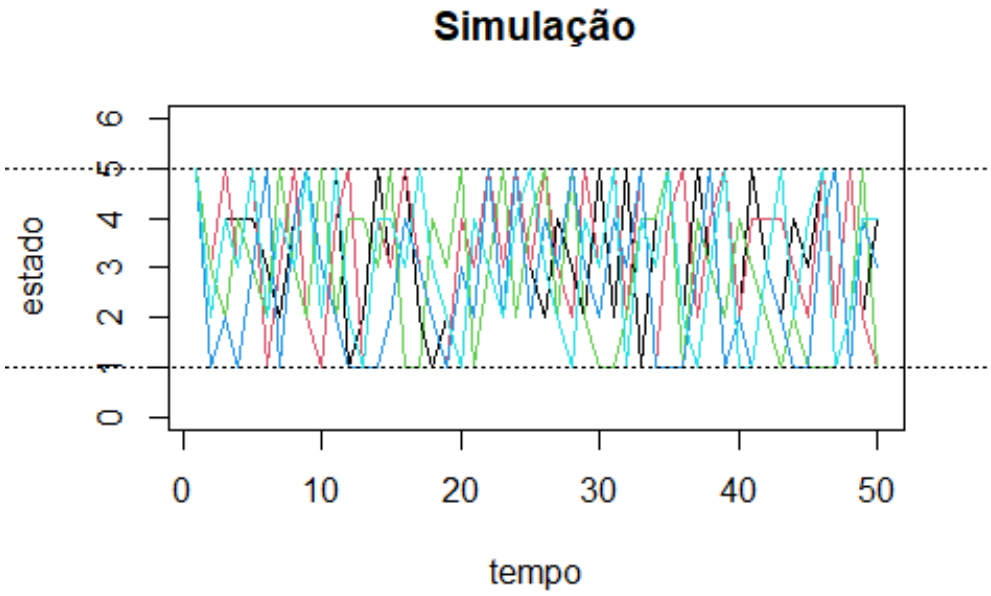
1.6 Algoritmo para simular uma amostra de tamanho n da cadeia de Markov estudada.

Código utilizado:

```
1
2  ##### Simulacao ######
3  #simular cadeias de Markov discretas de acordo com a matriz de transio P.
4  run.mc.sim <- function( P, num.iters = 50 ) {
5
6    # numero de possiveis estados
7    num.states <- nrow(P)
8
9    # armazena os astados X_t
10   states <- numeric(num.iters)
11
12   # estado inicial da cadeia
13   states[1] <- 5
14
15   for(t in 2:num.iters) {
16
17
18     p <- P[states[t-1], ]
19
20
21     states[t] <- which(rmultinom(1, 1, p) == 1)
22   }
23   return(states)
24 }
25
26
27
28 ##### Cadeia de Markov contruida #######
29
30
31 # A matriz de estudo
32 P <- t(matrix(c( 0.25, 0.25, 0.25, 0.25, 0,
33                  0.33, 0, 0, 0.33, 0.33,
34                  0, 0.5, 0, 0, 0.5,
35                  0, 0, 0.5, 0.25, 0.25,
36                  0.33, 0.33, 0.33, 0, 0), nrow=5, ncol=5))
37
38 # N de simulaes
39 num.chains <- 5 # Numero de simulaes
40 num.iterations <- 50 # Numero de iteraes
41 chain.states <- matrix(NA, ncol=num.chains, nrow=num.iterations)
42
43 # Simulao
44 for(c in seq_len(num.chains)){
45   chain.states[,c] <- run.mc.sim(P)
46 }
47
48 matplot(chain.states, main = "Simulao",sub = "Simulao de uma amostra de tamanho n, iniciando a
       ↪ cadeia no estado 1", type='l', lty=1, col=1:5, ylim=c(0,6), ylab='estado', xlab='tempo')
49 abline(h=1, lty=3)
50 abline(h=5, lty=3)
```

Listing 1: Código fonte em R

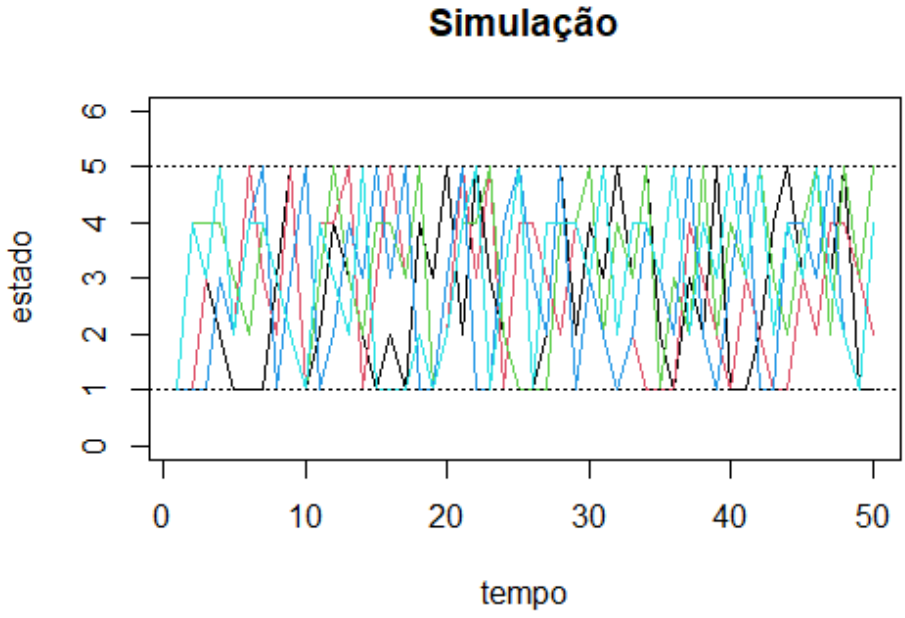
A seguir, será apresentado o resultado obtido através do algoritmo acima:



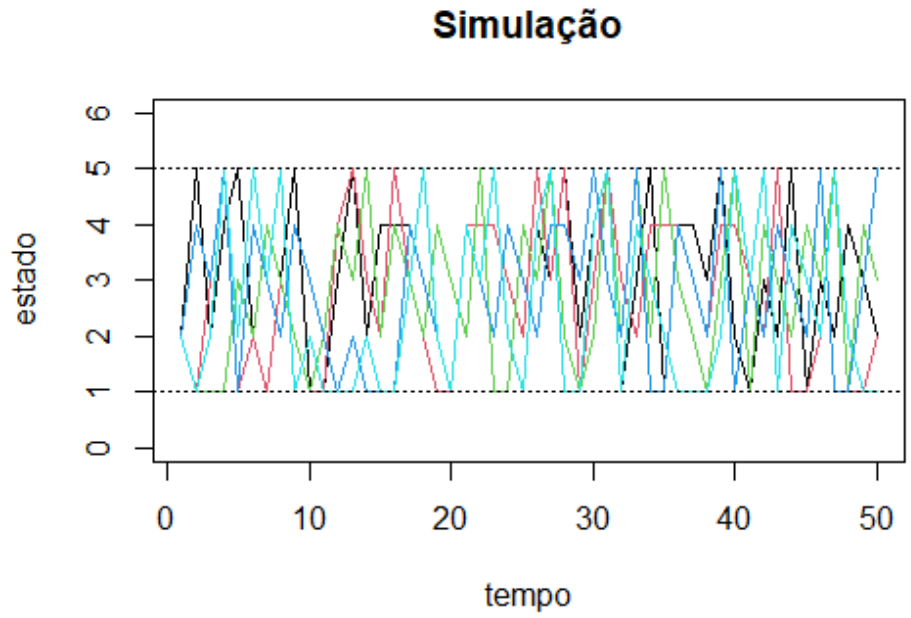
1.7 Simulação de uma amostra de tamanho n da cadeia de Markov em estudo utilizando diferentes valores inicias para a cadeia.

Utilizando o mesmo código utilizado na seção 1.6, vamos utilizar diferentes estados iniciais para gerar as seguintes simulações:

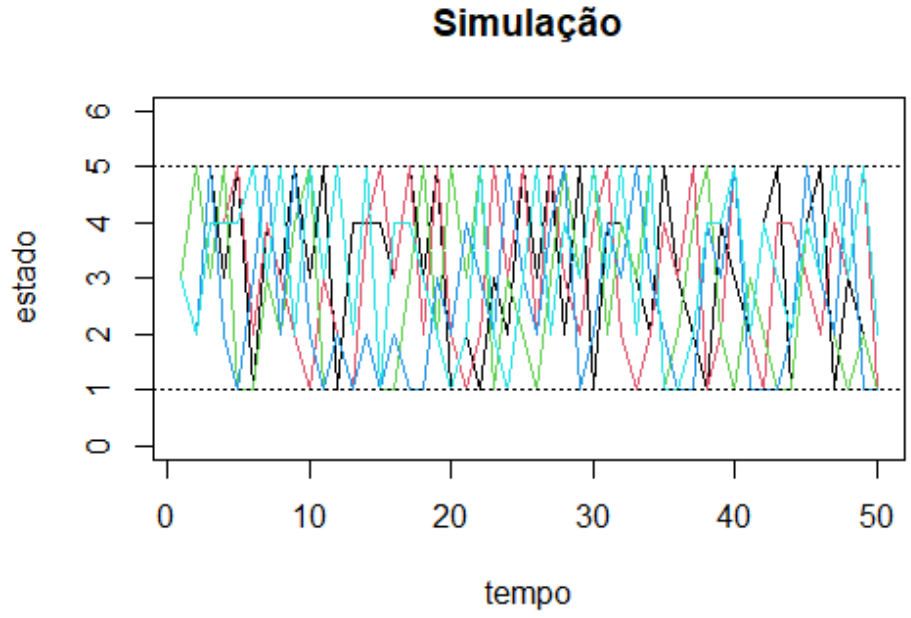
- Estado inicial = 1



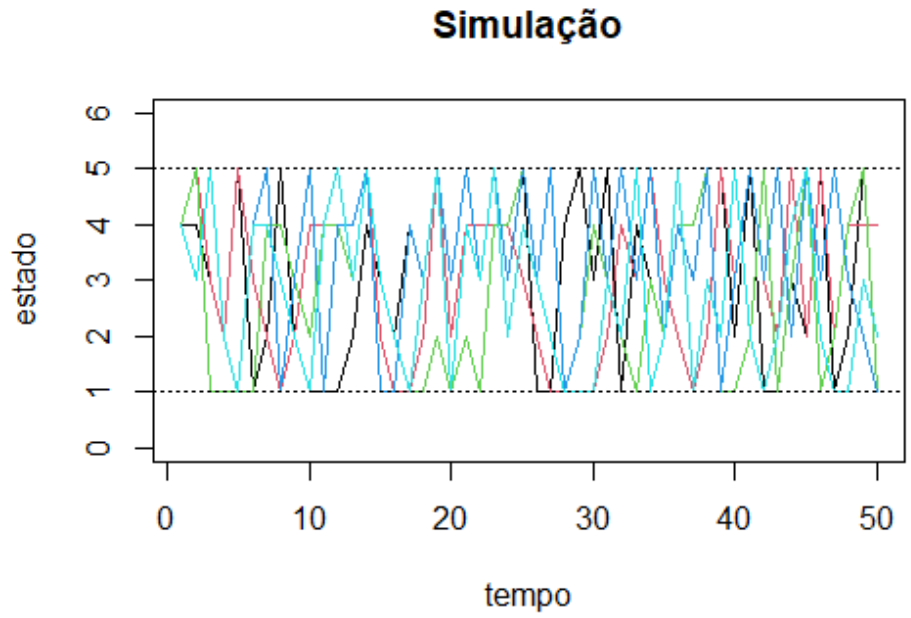
- Estado inicial = 2



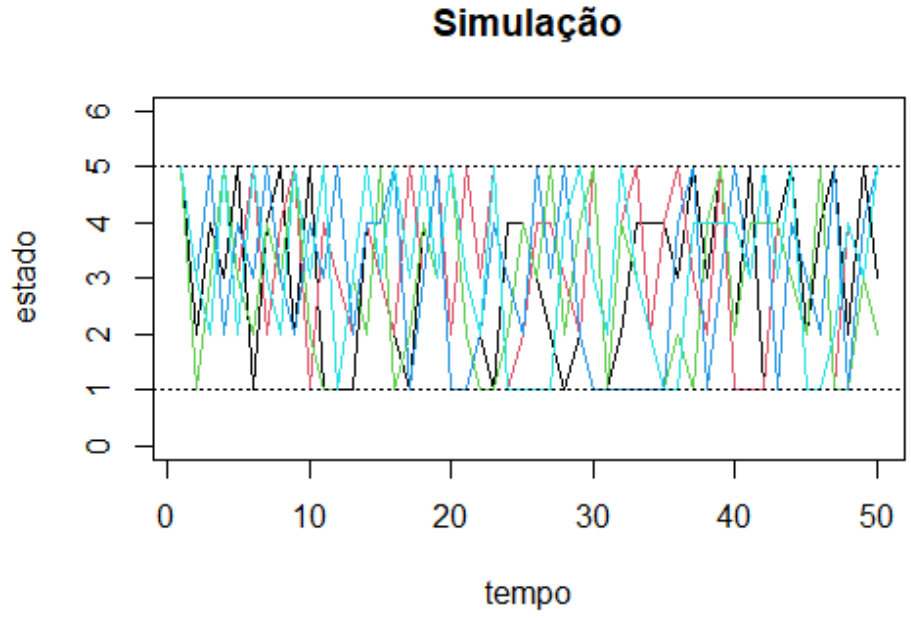
- Estado inicial = 3



- Estado inicial = 4



- Estado inicial = 5



1.8 Estimação da distribuição estacionária da cadeia.

A melhor forma para estimar a distribuição estacionária seria usar a matriz de transição (P) e verificar se a cadeia é irredutível, aperiódica e recorrente positiva.
Ao multiplicar a matriz transição (P) n vezes, temos que as colunas vão se convergir para o valor da distribuição estacionária, então para $n \geq 15$,

$$P^n = \begin{vmatrix} 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \\ 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \\ 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \\ 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \\ 0,194931 & 0,222222 & 0,202729 & 0,163742 & 0,216374 \end{vmatrix}$$

Podemos perceber com essa estimativa que temos os valores idênticos dos resultados teóricos, portanto, essa estimativa é quase certa que chegará em um valor muito próximo do resultado teórico que seria

$$\pi = (\pi_1 = 0,194931, \pi_2 = 0,222222, \pi_3 = 0,202729, \pi_4 = 0,163742, \pi_5 = 0,216374)$$

1.9 Tempos esperados de retorno a cada estado da cadeia.

A fórmula para o cálculo do tempo esperado de retorno é disponibilizada a seguir:

$$\mathbb{E}[T_i|X_0 = i] = \sum_{n=1}^{\infty} n \cdot \mathbb{P}(T_i = n|X_0 = i) = \sum_{n=1}^{\infty} n \cdot f_{ii}^{(n)}$$

Seja i um estado recorrente, dizemos que o estado i é **recorrente positivo** se $\mathbb{E}[T_i|X_0 = i] < \infty$ e que o estado i é **recorrente nulo** se $\mathbb{E}[T_i|X_0 = i] = \infty$.

Temos que todos os estados da cadeia é recorrente.

Fazendo o cálculo para cada estado da cadeia, temos que:

- Estado 1:

$$\begin{aligned} \mathbb{E}[T_1|X_0 = 1] = \sum_{n=1}^{\infty} n \cdot \mathbb{P}(T_1 = n|X_0 = 1) = \sum_{n=1}^{\infty} n \cdot f_{11}^{(n)} = \\ = 1 \cdot f_{11}^{(1)} + 2 \cdot f_{11}^{(2)} + 3 \cdot f_{11}^{(3)} + 4 \cdot f_{11}^{(4)} + 5 \cdot f_{11}^{(5)} = \\ = 1 \cdot 0,25 + 2 \cdot 0,25 \cdot \frac{1}{3} + 3 \cdot 0,25 \cdot 0,5 \cdot \frac{1}{3} + 4 \cdot 0,25 \cdot 0,5 \cdot 0,5 \cdot \frac{1}{3} \\ + 5 \cdot 0,25 \cdot 0,25 \cdot 0,5 \cdot 0,5 \cdot \frac{1}{3} = 0,6440 < \infty \end{aligned}$$

Logo, o estado 1 é recorrente positivo.

- Estado 2:

$$\begin{aligned}\mathbb{E}[T_2|X_0=2] &= \sum_{n=1}^{\infty} n\cdot\mathbb{P}(T_2=n|X_0=2) = \sum_{n=1}^{\infty} n\cdot f_{22}^{(n)} = \\ &= 1\cdot f_{22}^{(1)} + 2\cdot f_{22}^{(2)} + 3\cdot f_{22}^{(3)} + 4\cdot f_{22}^{(4)} + 5\cdot f_{22}^{(5)} = \\ &= 1\cdot 0 + 2\cdot\frac{1}{3}\cdot 0,25 + 3\cdot\frac{1}{3}\cdot 0,25\cdot 0,25 + 4\cdot\frac{1}{3}\cdot 0,25\cdot 0,5\cdot 0,5 \\ &+ 5\cdot\frac{1}{3}\cdot 0,25\cdot 0,25\cdot 0,5\cdot 0,5 = 0,3315 < \infty\end{aligned}$$

  Logo, o estado 2 é recorrente positivo.

- Estado 3:

$$\begin{aligned}\mathbb{E}[T_3|X_0=3] &= \sum_{n=1}^{\infty} n\cdot\mathbb{P}(T_3=n|X_0=3) = \sum_{n=1}^{\infty} n\cdot f_{33}^{(n)} = \\ &= 1\cdot f_{33}^{(1)} + 2\cdot f_{33}^{(2)} + 3\cdot f_{33}^{(3)} + 4\cdot f_{33}^{(4)} + 5\cdot f_{33}^{(5)} = \\ &= 1\cdot 0 + 2\cdot 0,5\cdot\frac{1}{3} + 3\cdot 0,5\cdot\frac{1}{3}\cdot 0,5 + 4\cdot 0,5\cdot\frac{1}{3}\cdot 0,25\cdot 0,25 \\ &+ 5\cdot 0,5\cdot\frac{1}{3}\cdot 0,25\cdot 0,25\cdot 0,5 = 0,6476 < \infty\end{aligned}$$

  Logo, o estado 3 é recorrente positivo.

- Estado 4:

$$\begin{aligned}\mathbb{E}[T_4|X_0=4] &= \sum_{n=1}^{\infty} n\cdot\mathbb{P}(T_4=n|X_0=4) = \sum_{n=1}^{\infty} n\cdot f_{44}^{(n)} = \\ &= 1\cdot f_{44}^{(1)} + 2\cdot f_{44}^{(2)} + 3\cdot f_{44}^{(3)} + 4\cdot f_{44}^{(4)} + 5\cdot f_{44}^{(5)} = \\ &= 1\cdot 0,25 + 2\cdot 0 + 3\cdot 0,5\cdot 0,5\cdot\frac{1}{3} + 4\cdot 0,25\cdot\frac{1}{3}\cdot 0,25\cdot 0,25 \\ &+ 5\cdot 0,5\cdot 0,5\cdot\frac{1}{3}\cdot 0,25\cdot\frac{1}{3} = 0,7355 < \infty\end{aligned}$$

  Logo, o estado 4 é recorrente positivo.

- Estado 5:

$$\begin{aligned}\mathbb{E}[T_5|X_0=5] &= \sum_{n=1}^{\infty} n\cdot\mathbb{P}(T_5=n|X_0=5) = \sum_{n=1}^{\infty} n\cdot f_{55}^{(n)} = \\ &= 1\cdot f_{55}^{(1)} + 2\cdot f_{55}^{(2)} + 3\cdot f_{55}^{(3)} + 4\cdot f_{55}^{(4)} + 5\cdot f_{55}^{(5)} = \\ &= 1\cdot 0 + 2\cdot\frac{1}{3}\cdot\frac{1}{3} + 3\cdot\frac{1}{3}\cdot 0,25\cdot\frac{1}{3} + 4\cdot\frac{1}{3}\cdot 0,25\cdot\frac{1}{3}\cdot 0,25 \\ &+ 5\cdot\frac{1}{3}\cdot 0,25\cdot 0,25\cdot\frac{1}{3}\cdot 0,25 = 0,3418 < \infty\end{aligned}$$

  Logo, o estado 5 é recorrente positivo.
  Portanto, como todos os estados são recorrentes positivo, podemos concluir que a cadeira de markov é recorrente positiva.

1.10 Estimativas para $\hat{P}_{ab}$ para diferentes tamanhos de amostras n.

Através do *Software* RStudio e códigos obtidos através do site *Github* que estão apresentados no final deste relatório na seção "Códigos". Simulamos os valores da matriz de transição $\hat{P}_{ab}$, utilizaremos um estimador de máxima verossimilhança para Q, que é dado por: $\hat{Q}(b \mid a) = \frac{N(ab)}{\sum_{b \in S} N(ab)}$, em que:

- $\hat{Q}(b \mid a)$, com a,b = (1,2,3,4,5), é a nossa matriz de transição que queremos estimar;
- $N(ab)$N(a b) corresponde ao número de vezes que observamos o estado b depois do estado a;
- $\sum_{b \in S} N(ab)$ corresponde à somatória do número de vezes que observamos qualquer outro estado.

Desta maneira, obtemos as seguintes matrizes para os tamanhos de n determinados:

- Para n=100

$$P = \begin{vmatrix} 0,35 & 0,30 & 0,20 & 0,15 & 0 \\ 0,25 & 0 & 0 & 0,4166667 & 0,3333333 \\ 0 & 0,60 & 0 & 0 & 0,40 \\ 0 & 0 & 0,5625 & 0,1875 & 0,25 \\ 0,35 & 0,25 & 0,4 & 0 & 0 \end{vmatrix}$$

- Para n=1000

$$P = \begin{vmatrix} 0,1894737 & 0,2473684 & 0,2315789 & 0,3315789 & 0 \\ 0,3463415 & 0 & 0 & 0,35122195 & 0,302439 \\ 0 & 0,475 & 0 & 0 & 0,525 \\ 0 & 0 & 0,4491979 & 0,2780749 & 0,2727273 \\ 0,3807339 & 0,3844037 & 0,3348624 & 0 & 0 \end{vmatrix}$$

- Para n=10000

$$P = \begin{vmatrix} 0,2570850 & 0,2591093 & 0,2368421 & 0,2469636 & 0 \\ 0,33336315 & 0 & 0 & 0,3313953 & 0,3349732 \\ 0 & 0,4955268 & 0 & 0 & 0,5044732 \\ 0 & 0 & 0,5081658 & 0,2280151 & 0,2638191 \\ 0,3305861 & 0,3324176 & 0,3369963 & 0 & 0 \end{vmatrix}$$

Através das estimativas realizadas, podemos perceber que conforme o tamanho amostral aumenta, mais próximo as estimativas para $\hat{P}_{ab}$ ficam da verdadeira probabilidade, ou seja, a matriz simulada fica cada vez mais próxima da verdadeira matriz conforme o tamanho da amostra aumenta.

2 Parte II: Análise de Dados através de Cadeias de Markov

Para a realização desta análise, foram esolhidos os dados disponíveis no site disponibilizado pelo professor, no qual foram selecionados os seguintes bancos de dados:

- MZ414421, onde os dados provém da localidade de Nova Iorque, nos Estados Unidos e foram coletados em 31 de março de 2021,
- MZ414283, em que os dados provém da localidade Virginia, também dos Estados Unidos, e foram coletados em maio de 2021.

Utilizando novamente uma parte do código já citado anteriormente, vamos estimar as probabilidades de transição, considerando cadeias de ordem 1,2 e 3. Nosso objetivo é comparar as matrizes obtidas para os dois bancos de dados diferentes.

Assim, iniciaremos nossa análise para as cadeias de ordem 1. Para este, nosso espaço de estados é definido por $S = (a, c, g, t)$

Assim, para o banco de dados MZ414421, temos a seguinte matriz, que denominaremos por $\hat{P}_1$:

$$\hat{P}_1 := \begin{array}{c} \\ a \\ c \\ g \\ t \end{array} \begin{array}{c} \begin{array}{cccc} a & c & g & t \end{array} \\ \begin{pmatrix} 0.2943905 & 0.2067138 & 0.18087456 & 0.2232774 \\ 0.3297358 & 0.1403052 & 0.07071083 & 0.3554150 \\ 0.2383471 & 0.1712397 & 0.16859504 & 0.3233058 \\ 0.2328376 & 0.1189652 & 0.24404156 & 0.3004685 \end{pmatrix} \end{array}$$

E para o banco de dados MZ414283, temos a seguinte matriz, chamada de $\hat{P}_2$:

$$\hat{P}_2 := \begin{array}{c} \\ a \\ c \\ g \\ t \end{array} \begin{array}{c} \begin{array}{cccc} a & c & g & t \end{array} \\ \begin{pmatrix} 0.2888055 & 0.2037977 & 0.18414661 & 0.2223449 \\ 0.3309701 & 0.1440299 & 0.06716418 & 0.3626866 \\ 0.2442244 & 0.1706271 & 0.16831683 & 0.3201320 \\ 0.2290605 & 0.1202364 & 0.24760546 & 0.2991645 \end{pmatrix} \end{array}$$

Podemos verificar, que as probabilidades de transição da cadeia considerando 1 ordem, a maior parte delas apenas difere depois da terceira casa decimal, assim como se considerarmos apenas as duas primeiras, com arredondamento, podemos verificar que elas passam a ser quase que iguais. Assim, podemos concluir que para uma cadeia de Markov de ordem 1 para este problema, as probabilidades de transição não diferem consideravelmente para os dois bancos de dados.

A seguir, apresentaremos as matrizes coniderando uma cadeia de ordem 2, com espaçoes de estados dado por: $S = (aa, ac, ag, at, ca, cg, cc, ct, ga, gc, gg, gt, ta, tc, tg, tt)$

Assim, para o banco de dados MZ414421, temos a seguinte matriz, denotada por $\hat{Q}_1$:

$$\hat{Q}_1 := \begin{array}{c} \\ aa \\ ac \\ ag \\ at \\ ca \\ cg \\ cc \\ ct \\ ga \\ gc \\ gt \\ gt \\ ta \\ tc \\ tg \\ tt \end{array} \begin{array}{c} \begin{array}{cccc} a & c & g & t \end{array} \\ \begin{pmatrix} 0.2798200 & 0.20030008 & 0.20330083 & 0.2243061 \\ 0.3450855 & 0.15064103 & 0.06623932 & 0.3173077 \\ 0.3076923 & 0.15628816 & 0.15628816 & 0.2722833 \\ 0.1869436 & 0.12561820 & 0.28783383 & 0.2769535 \\ 0.3047404 & 0.20090293 & 0.16930023 & 0.2234763 \\ 0.3421751 & 0.11140584 & 0.07161804 & 0.3713528 \\ 0.1789474 & 0.18421053 & 0.17894737 & 0.3526316 \\ 0.2502618 & 0.09947644 & 0.21256545 & 0.3256545 \\ 0.3079057 & 0.16782247 & 0.16227462 & 0.2385576 \\ 0.2895753 & 0.14864865 & 0.06370656 & 0.3918919 \\ 0.2078431 & 0.17647059 & 0.10392157 & 0.4137255 \\ 0.2116564 & 0.10633947 & 0.24539877 & 0.3108384 \\ 0.2703412 & 0.23622047 & 0.15923010 & 0.2230971 \\ 0.3304795 & 0.12500000 & 0.07534247 & 0.3458904 \\ 0.2078464 & 0.17696160 & 0.19949917 & 0.2988314 \\ 0.2555932 & 0.12542373 & 0.22644068 & 0.2874576 \end{pmatrix} \end{array}$$

E para o banco de dados MZ414283, temos a seguinte matriz, que denotaremos por $\hat{Q}_2$:

$$\hat{Q}_2 := \begin{array}{c} \\ aa \\ ac \\ ag \\ at \\ ca \\ cc \\ cg \\ ct \\ ga \\ gc \\ gg \\ gt \\ ta \\ tc \\ tg \\ tt \end{array} \begin{array}{c} \begin{array}{cccc} a & c & g & t \end{array} \\ \begin{pmatrix} 0.2744648 & 0.1926606 & 0.21024465 & 0.2117737 \\ 0.3391116 & 0.1657638 & 0.06500542 & 0.3131094 \\ 0.3105516 & 0.1546763 & 0.15707434 & 0.2721823 \\ 0.1838966 & 0.1292247 & 0.29025845 & 0.2842942 \\ 0.2998873 & 0.1927847 & 0.18489290 & 0.2040586 \\ 0.3445596 & 0.1036269 & 0.08031088 & 0.3808290 \\ 0.1944444 & 0.1888889 & 0.15555556 & 0.3611111 \\ 0.2376543 & 0.1152263 & 0.21193416 & 0.3220165 \\ 0.3108108 & 0.1797297 & 0.17027027 & 0.2445946 \\ 0.2765957 & 0.1566731 & 0.05802708 & 0.3926499 \\ 0.2078431 & 0.1745098 & 0.10196078 & 0.4000000 \\ 0.2092784 & 0.1041237 & 0.25154639 & 0.3257732 \\ 0.2571174 & 0.2411032 & 0.15569395 & 0.2241993 \\ 0.3525424 & 0.1237288 & 0.06610169 & 0.3694915 \\ 0.2222222 & 0.1687243 & 0.19588477 & 0.3078189 \\ 0.2527248 & 0.1287466 & 0.22615804 & 0.2690736 \end{pmatrix} \end{array}$$

Considerando uma cadeia de ordem 2, também notamos que muitas probabilidades são parecidas, novamente sendo diferentes a partir da terceira casa decimal, porém em menor número se considerado com as cadeias de ordem 1. Também podemos observar que para os dois bancos de dados, a maior probabilidade de transição se dá do estado **gc** para o estado **t**, ou seja, se estamos no estado **gc**, é mais provável que iremos a seguir para o estado **t**. Agora, considerando a menor probabilidade da cadeia, temos que esta se dá do estado **gc** para o estado **g**, novamente para as duas localidades.

E por fim, considerando uma cadeia de ordem 3, temos o seguinte espaço de estados:

$S = (aaa, aac, aag, aat, aca, acc, acg, act, aga, agc, agg, agt, ata, atc, atg, att, caa, cac, cag, cat, aca, ccc, ccg, cct, cga, cgc, cgg, cgt, cta, ctc, ctg, ctt, gaa, gac, gag, gat, gca, gcc, gcg, gct, gga, ggc, ggg, ggt, gta, gtc, gtg, gtt, taa, tac, tag, tat, tca, tcc, tcg, tct, tga, tgc, tgg, tgt, tta, ttc, ttg, ttt)$.

Assim, para o banco de dados MZ414421, temos a seguinte matriz, denotada por $\hat{T}_1$:

$$\hat{T}_1 := \begin{array}{c} \\ aaa \\ aac \\ aag \\ aat \\ aca \\ acc \\ acg \\ act \\ aga \\ agc \\ agg \\ agt \\ ata \\ atc \\ atg \\ att \\ caa \\ cac \\ cag \\ cat \\ cca \\ ccc \\ ccg \\ cct \\ cga \\ cgc \\ cgg \\ cgt \\ cta \\ ctc \\ ctg \\ ctt \\ gaa \\ gac \\ gag \\ gat \\ gca \\ gcc \\ gcg \\ gct \\ gga \\ ggc \\ ggg \\ ggt \end{array} \begin{array}{c} a \\ \left(\begin{array}{c} 0.2466488 \\ 0.3632959 \\ 0.2915129 \\ 0.1705686 \\ 0.3560372 \\ 0.3546099 \\ 0.1129032 \\ 0.2457912 \\ 0.3214286 \\ 0.2343750 \\ 0.1640625 \\ 0.1928251 \\ 0.2910053 \\ 0.3307087 \\ 0.2199313 \\ 0.2071429 \\ 0.2222222 \\ 0.2471910 \\ 0.3266667 \\ 0.2222222 \\ 0.2713178 \\ 0.3095238 \\ 0.2592593 \\ 0.2714286 \\ 0.3235294 \\ 0.4285714 \\ 0.1764706 \\ 0.2089552 \\ 0.2594142 \\ 0.3263158 \\ 0.1921182 \\ 0.2765273 \\ 0.3108108 \\ 0.5041322 \\ 0.2649573 \\ 0.1802326 \\ 0.2400000 \\ 0.3506494 \\ 0.3030303 \\ 0.2216749 \\ 0.3018868 \\ 0.2777778 \\ 0.1886792 \\ 0.2843602 \end{array} \right. \end{array} \begin{array}{c} c \\ \begin{array}{c} 0.19571046 \\ 0.17228464 \\ 0.14391144 \\ 0.11371237 \\ 0.19195046 \\ 0.11347518 \\ 0.17741935 \\ 0.10774411 \\ 0.15873016 \\ 0.14062500 \\ 0.18750000 \\ 0.08520179 \\ 0.21693122 \\ 0.12598425 \\ 0.18900344 \\ 0.14642857 \\ 0.28148148 \\ 0.15730337 \\ 0.16666667 \\ 0.14646465 \\ 0.21705426 \\ 0.09523810 \\ 0.18518519 \\ 0.12857143 \\ 0.11764706 \\ 0.17142857 \\ 0.14705882 \\ 0.19402985 \\ 0.20920502 \\ 0.05263158 \\ 0.20197044 \\ 0.11575563 \\ 0.10810811 \\ 0.07438017 \\ 0.17094017 \\ 0.09883721 \\ 0.22666667 \\ 0.09090909 \\ 0.09090909 \\ 0.08374384 \\ 0.17924528 \\ 0.11111111 \\ 0.22641509 \\ 0.06161137 \end{array} \end{array} \begin{array}{c} g \\ \begin{array}{c} 0.23324397 \\ 0.04119850 \\ 0.18450185 \\ 0.28093645 \\ 0.13312693 \\ 0.04964539 \\ 0.16129032 \\ 0.19191919 \\ 0.18650794 \\ 0.05468750 \\ 0.14062500 \\ 0.25560538 \\ 0.15343915 \\ 0.04724409 \\ 0.23024055 \\ 0.23214286 \\ 0.12592593 \\ 0.06741573 \\ 0.14000000 \\ 0.20202020 \\ 0.13178295 \\ 0.07142857 \\ 0.14814815 \\ 0.16428571 \\ 0.08823529 \\ 0.08571429 \\ 0.14705882 \\ 0.23880597 \\ 0.14644351 \\ 0.12631579 \\ 0.18719212 \\ 0.18971061 \\ 0.34684685 \\ 0.09090909 \\ 0.17948718 \\ 0.40116279 \\ 0.20000000 \\ 0.09090909 \\ 0.15151515 \\ 0.30541872 \\ 0.18867925 \\ 0.05555556 \\ 0.09433962 \\ 0.29383886 \end{array} \end{array} \begin{array}{c} t \\ \left. \begin{array}{c} 0.2305630 \\ 0.3108614 \\ 0.2324723 \\ 0.3076923 \\ 0.2260062 \\ 0.3758865 \\ 0.3548387 \\ 0.3198653 \\ 0.2023810 \\ 0.4296875 \\ 0.3750000 \\ 0.3318386 \\ 0.1798942 \\ 0.3937008 \\ 0.2096220 \\ 0.2750000 \\ 0.2370370 \\ 0.3426966 \\ 0.2666667 \\ 0.3131313 \\ 0.2248062 \\ 0.4285714 \\ 0.3333333 \\ 0.3000000 \\ 0.3235294 \\ 0.1714286 \\ 0.4411765 \\ 0.2089552 \\ 0.2552301 \\ 0.3052632 \\ 0.3054187 \\ 0.2958199 \\ 0.1351351 \\ 0.2148760 \\ 0.2564103 \\ 0.1686047 \\ 0.2200000 \\ 0.3116883 \\ 0.3636364 \\ 0.2709360 \\ 0.2358491 \\ 0.4000000 \\ 0.4339623 \\ 0.2274882 \end{array} \right) \end{array}$$

Continuação da matriz $\hat{T}_1$:

$$\hat{T}_1 := \begin{array}{c} \\ gta \\ gtc \\ gtg \\ gtt \\ taa \\ tac \\ tag \\ tat \\ tca \\ tcc \\ tcg \\ tct \\ tga \\ tgc \\ tgg \\ tgt \\ tta \\ ttc \\ ttg \\ ttt \end{array} \begin{array}{c} \begin{array}{cccc} a & c & g & t \end{array} \\ \left(\begin{array}{cccc} 0.2318841 & 0.25603865 & 0.20772947 & 0.1642512 \\ 0.2884615 & 0.14423077 & 0.07692308 & 0.3365385 \\ 0.1916667 & 0.19166667 & 0.19583333 & 0.3000000 \\ 0.2565789 & 0.08881579 & 0.23026316 & 0.2861842 \\ 0.3300971 & 0.18446602 & 0.12621359 & 0.2621359 \\ 0.3296296 & 0.13333333 & 0.07777778 & 0.3259259 \\ 0.3241758 & 0.15934066 & 0.10439560 & 0.3131868 \\ 0.1764706 & 0.10588235 & 0.28235294 & 0.2901961 \\ 0.2849741 & 0.17616580 & 0.20207254 & 0.2176166 \\ 0.3561644 & 0.16438356 & 0.05479452 & 0.3150685 \\ 0.1590909 & 0.27272727 & 0.18181818 & 0.3181818 \\ 0.2475248 & 0.09405941 & 0.17821782 & 0.3613861 \\ 0.2851406 & 0.17670683 & 0.11244980 & 0.2610442 \\ 0.3018868 & 0.15566038 & 0.07075472 & 0.3867925 \\ 0.2050209 & 0.16317992 & 0.07531381 & 0.4435146 \\ 0.1955307 & 0.12569832 & 0.20949721 & 0.3156425 \\ 0.2705570 & 0.23872679 & 0.14058355 & 0.2519894 \\ 0.3135135 & 0.12972973 & 0.05945946 & 0.3621622 \\ 0.2005988 & 0.14371257 & 0.17065868 & 0.3532934 \\ 0.2641509 & 0.12500000 & 0.24764151 & 0.2783019 \end{array} \right) \end{array}$$

E para o banco de dados MZ414283, temos a seguinte matriz denotada por $\hat{T}_2$:

$$\hat{T}_2 := \begin{array}{c} \\ aaa \\ aac \\ aag \\ aat \\ aca \\ acc \\ acg \\ act \\ aga \\ agc \\ agg \\ agt \\ ata \\ atc \\ atg \\ att \\ caa \\ cac \\ cag \\ cat \\ cca \\ ccc \\ ccg \\ cct \\ cga \\ cgc \\ cgg \\ cgt \\ cta \\ ctc \\ ctg \\ ctt \\ gaa \\ gac \\ gag \\ gat \\ gca \\ gcc \\ gcg \\ gct \\ gga \\ ggc \\ ggg \\ ggt \end{array} \begin{array}{c} a \\ \left(\begin{array}{c} 0.2116992 \\ 0.2976190 \\ 0.3345455 \\ 0.1702899 \\ 0.3194888 \\ 0.3464052 \\ 0.1500000 \\ 0.2422145 \\ 0.2818533 \\ 0.2093023 \\ 0.2519084 \\ 0.1982379 \\ 0.2864865 \\ 0.3692308 \\ 0.2157534 \\ 0.2517483 \\ 0.2593985 \\ 0.2748538 \\ 0.2926829 \\ 0.1823204 \\ 0.2481203 \\ 0.2250000 \\ 0.2580645 \\ 0.2517007 \\ 0.4285714 \\ 0.3529412 \\ 0.2142857 \\ 0.2307692 \\ 0.2467532 \\ 0.3839286 \\ 0.1796117 \\ 0.2651757 \\ 0.3043478 \\ 0.4285714 \\ 0.2539683 \\ 0.1878453 \\ 0.2377622 \\ 0.3580247 \\ 0.2000000 \\ 0.2315271 \\ 0.2830189 \\ 0.2696629 \\ 0.1923077 \\ 0.2352941 \end{array} \right. \end{array} \begin{array}{c} c \\ \begin{array}{c} 0.20334262 \\ 0.16269841 \\ 0.13818182 \\ 0.11594203 \\ 0.18849840 \\ 0.08496732 \\ 0.15000000 \\ 0.12456747 \\ 0.16216216 \\ 0.16279070 \\ 0.15267176 \\ 0.08810573 \\ 0.24324324 \\ 0.10000000 \\ 0.16780822 \\ 0.16083916 \\ 0.26691729 \\ 0.21052632 \\ 0.19512195 \\ 0.16574586 \\ 0.21052632 \\ 0.07500000 \\ 0.19354839 \\ 0.14965986 \\ 0.08571429 \\ 0.17647059 \\ 0.17857143 \\ 0.18461538 \\ 0.22510823 \\ 0.08035714 \\ 0.18446602 \\ 0.13099042 \\ 0.13478261 \\ 0.13533835 \\ 0.13492063 \\ 0.10497238 \\ 0.22377622 \\ 0.09876543 \\ 0.13333333 \\ 0.09359606 \\ 0.22641509 \\ 0.07865169 \\ 0.26923077 \\ 0.05392157 \end{array} \end{array} \begin{array}{c} g \\ \begin{array}{c} 0.22562674 \\ 0.05555556 \\ 0.17818182 \\ 0.30797101 \\ 0.15654952 \\ 0.06535948 \\ 0.23333333 \\ 0.19723183 \\ 0.22393822 \\ 0.05426357 \\ 0.16030534 \\ 0.25110132 \\ 0.13513514 \\ 0.05384615 \\ 0.19863014 \\ 0.21678322 \\ 0.15789474 \\ 0.08187135 \\ 0.14634146 \\ 0.20994475 \\ 0.19548872 \\ 0.17500000 \\ 0.12903226 \\ 0.16326531 \\ 0.11428571 \\ 0.08823529 \\ 0.14285714 \\ 0.29230769 \\ 0.13852814 \\ 0.08035714 \\ 0.17961165 \\ 0.19808307 \\ 0.30434783 \\ 0.09022556 \\ 0.23015873 \\ 0.33149171 \\ 0.22377622 \\ 0.06172840 \\ 0.06666667 \\ 0.23645320 \\ 0.21698113 \\ 0.06741573 \\ 0.03846154 \\ 0.29901961 \end{array} \end{array} \begin{array}{c} t \\ \left. \begin{array}{c} 0.2256267 \\ 0.3412698 \\ 0.2472727 \\ 0.2898551 \\ 0.1916933 \\ 0.4052288 \\ 0.3666667 \\ 0.3079585 \\ 0.1969112 \\ 0.4496124 \\ 0.2977099 \\ 0.3612335 \\ 0.2162162 \\ 0.4000000 \\ 0.2705479 \\ 0.2587413 \\ 0.2067669 \\ 0.3157895 \\ 0.2621951 \\ 0.2651934 \\ 0.2255639 \\ 0.4250000 \\ 0.3548387 \\ 0.2925170 \\ 0.3428571 \\ 0.2352941 \\ 0.3214286 \\ 0.1384615 \\ 0.2554113 \\ 0.3214286 \\ 0.3300971 \\ 0.2651757 \\ 0.1478261 \\ 0.2406015 \\ 0.2301587 \\ 0.2541436 \\ 0.1678322 \\ 0.3827160 \\ 0.4000000 \\ 0.2955665 \\ 0.1792453 \\ 0.4606742 \\ 0.4230769 \\ 0.2843137 \end{array} \right) \end{array}$$

Continuação da matriz $\hat{T}_2$:

$$\hat{T}_2 := \begin{array}{c} \\ gta \\ gtc \\ gtg \\ gtt \\ taa \\ tac \\ tag \\ tat \\ tca \\ tcc \\ tcg \\ tct \\ tga \\ tgc \\ tgg \\ tgt \\ tta \\ ttc \\ ttc \\ ttt \end{array} \begin{array}{c} \begin{array}{cccc} a & c & g & t \end{array} \\ \begin{pmatrix} 0.2118227 & 0.27093596 & 0.19704433 & 0.1773399 \\ 0.3861386 & 0.15841584 & 0.07920792 & 0.3267327 \\ 0.2090164 & 0.17213115 & 0.20901639 & 0.3196721 \\ 0.2468354 & 0.10443038 & 0.26582278 & 0.2405063 \\ 0.2941176 & 0.16608997 & 0.14878893 & 0.2422145 \\ 0.3357934 & 0.13653137 & 0.06273063 & 0.3247232 \\ 0.3085714 & 0.13714286 & 0.09714286 & 0.3200000 \\ 0.2023810 & 0.10714286 & 0.28174603 & 0.3015873 \\ 0.3269231 & 0.14903846 & 0.19711538 & 0.2163462 \\ 0.3287671 & 0.15068493 & 0.05479452 & 0.3561644 \\ 0.1025641 & 0.25641026 & 0.12820513 & 0.4102564 \\ 0.2064220 & 0.10550459 & 0.21100917 & 0.3715596 \\ 0.3074074 & 0.18148148 & 0.12222222 & 0.3000000 \\ 0.2682927 & 0.16585366 & 0.04878049 & 0.3804878 \\ 0.1848739 & 0.15966387 & 0.08403361 & 0.4327731 \\ 0.1951872 & 0.12032086 & 0.20053476 & 0.3582888 \\ 0.2803235 & 0.22371968 & 0.14016173 & 0.2075472 \\ 0.3333333 & 0.11640212 & 0.05820106 & 0.3756614 \\ 0.2259036 & 0.14457831 & 0.16566265 & 0.3524096 \\ 0.2455696 & 0.12151899 & 0.20506329 & 0.2759494 \end{pmatrix} \end{array}$$

Por fim, ao analisarmos uma cadeia de ordem 3, podemos perceber que as probabilidades de transição já diferem significamente, se comparado com as obtidas nas cadeias considerando ordens 1 e 2. Na matriz de transição , podemos perceber que o estado de transição que possui maior valor de probabilidade é o estado da trinca **gac** para o estado **a**, com valor $0,504132$, e o estado que apresenta a menor probabilidade de transição da cadeia é a transição da trinca **aac** para o estado **g**. Ou seja, se estivermos no estado **gac**, há uma grande probabilidade de o próximo elemento do RNA ser a letra **a**, em contrapartida, se estivermos na trinca **aac**, temos poucas chances de o próximo elemento ser a letra **g**.

Diferente do que acontecia nas outras matrizes considerando ordens 1 e 2, para uma cadeia de ordem 3, os valores de maior e menor probabilidade de transição da cadeia são diferentes. Assim, no caso da matriz , o maior valor de probabilidade de transição é da trinca **ggc** para o estado **t**, e a menor probabilidade se refere à trinca **ggg** para o estado **g**, esta última sendo um resultado interessante, pois se estivermos vindo de uma sequência que nos revela a trinca **ggg**, é muito pouco provável que no próximo estado eu tenha uma letra **g** novamente.

Podemos notar que, considerando uma cadeia de ordem 3, podemos vizualizar melhor as diferenças entre o sequenciamento do RNA do Sars-COV-2 nas duas diferentes localidades e em diferentes momentos. Seria preciso uma análise mais aprofundada sobre o assunto para podermos concluir se existe uma diferença significativa entre eles, porém olhando apenas a matriz de transição, podemos perceber que existe sim uma pequena diferença entre os valores.

3 Códigos

Os códigos e dados utilizados ao longo desse trabalho estão disponíveis em: https://github.com/Douglas-Nestlehner/Processos_Estocasticos

Referências

[1] E. Çinlar. **Introduction to stochastic processes.** Englewood Cliffs, N. J.: Prentice-Hall, 1975.

[2] Girardi, V.A.; FERREIRA, R.F.; PENA, R.F.O. **Inderência da conectividade neuronal via estimação de medidas da teoria da informação** Departamento de Estatística, Universidade Federal de São Carlos, Brasil, 2021. **Comandos da Teoria da Informação**. Disponível em: https://github.com/Victor-girardi/Comandos-da-Teoria-da-Informa-o. Acesso em 17 de junho de 2021.

[3] Notas de aula da matéria Processos Estocásticos.