Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
IMD0601 - Bioestatística

Cadeias de Markov

Prof. Dr. Tetsu Sakamoto
Instituto Metrópole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

Baixe a aula (e os arquivos)

- Para aqueles que não clonaram o repositório:

```
> git clone https://github.com/tetsufmbio/IMD0601.git
```

- Para aqueles que já tem o repositório local:

```
> cd /path/to/IMD0601

> git pull
```

Essência do algoritmo bayesiano

$$Posteriori \propto Verossimilhança.\, Priori$$

Objetivo de uma inferência bayesiana ➔ Encontrar a **distribuição posteriori**.

Priors conjugados:

$$P(H|E) \propto \binom{n}{k} p^k (1-p)^{n-k} . \frac{1}{\beta(\alpha,\beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

$$P(H|E) \propto c p^{k+\alpha-1} (1-p)^{n-k+\beta-1}$$

Fácil de falar, difícil de fazer.

Determinar a distribuição posteriori pode ser uma tarefa árdua...

Pares conjugados verossimilhança multinomial e prior Dirichlet:

Verossimilhança

$$f(X_1, X_2, ..X_K) = \binom{n}{X_1 X_2 .. X_K} p_1^{X1} p_2^{X2} ... p_k^{Xk}$$

Posterior

$$\binom{n}{X_1 X_2 .. X_K} p_1^{X1} p_2^{X2} ... p_k^{Xk} \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

Prior

$$f(p_1, p_2, ..p_K) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

Fácil de falar, difícil de fazer.

Determinar a distribuição posteriori pode ser uma tarefa árdua...

Pares conjugados verossimilhança normal e prior normal (média) e gamma (variância):

Verossimilhança

$$\text{lik}(\mu,\phi) = \frac{\phi^{n/2}}{(2\pi)^{n/2}} e^{-\frac{\phi}{2}\sum_{i=1}^{n}(X_i-\mu)^2}$$

Posterior

$$f(\mu,\phi \mid X_1,\dots,X_n) \propto \phi^{n/2} e^{-\frac{\phi}{2}\sum_{i=1}^{n}(X_i-\mu)^2} \frac{1}{\tau} e^{-\frac{1}{2\tau^2}\mu^2} \phi^{\alpha-1} e^{-\left(\phi/\beta\right)}$$

Prior

$$f(\mu,\phi) = f_1(\mu)\, f_2(\phi) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}\mu^2} \cdot \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} \phi^{\alpha-1} e^{-\left(\phi/\beta\right)}$$

Distribuição posteriori

Envolvendo múltiplas variáveis

Difíceis de analisar;

Soluções analíticas não triviais;

Em bioinformática (e em outras áreas) ➜ busca por informações sobre a distribuição posteriori.

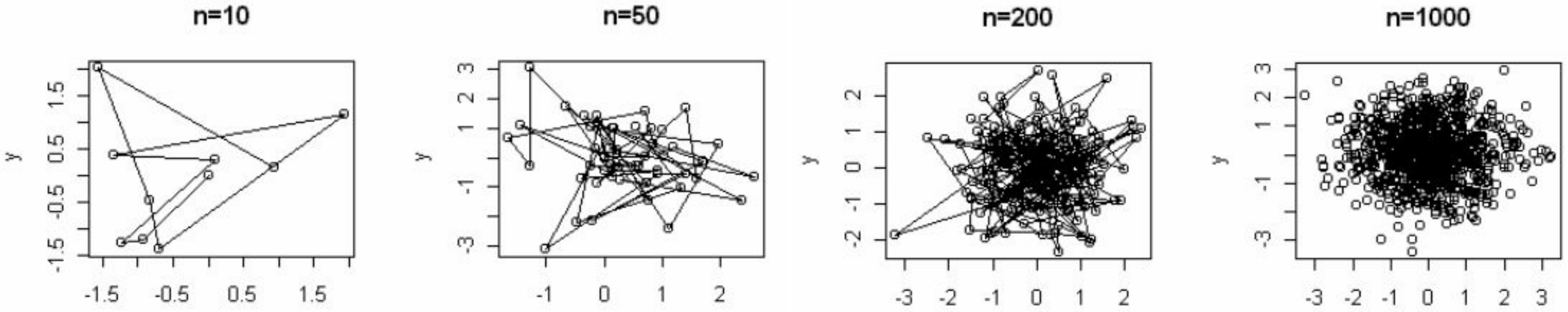
Solução ➜ MCMC

Markov Chain Monte Carlo (MCMC)

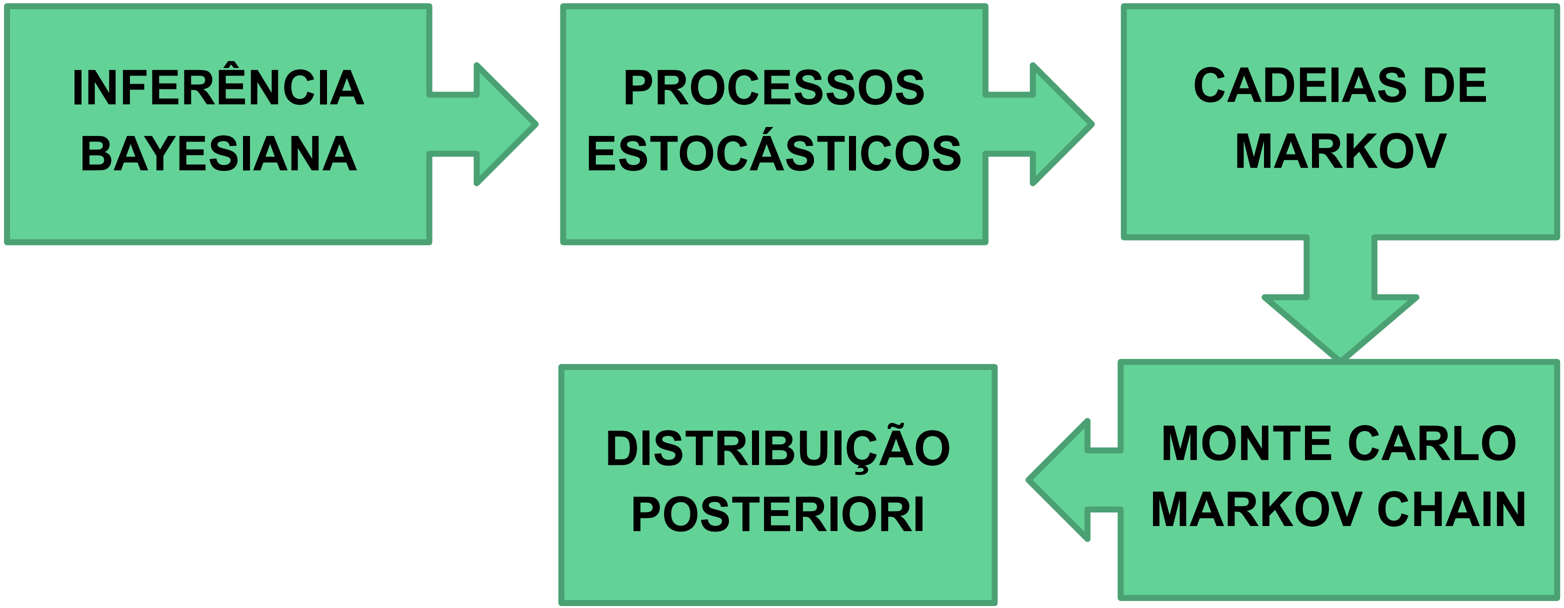
Um método de simulação que produz distribuição posteriori;

Abaixo uma simulação de uma distribuição normal bivariada;

- Aproximação se inicia de um ponto e passeia aleatoriamente no espaço por n ciclos.
- Cada amostra sucessiva são condicionalmente dependentes do outro.



Road map


INFERÊNCIA BAYESIANA
PROCESSOS ESTOCÁSTICOS
CADEIAS DE MARKOV
MONTE CARLO MARKOV CHAIN
DISTRIBUIÇÃO POSTERIORI

Modelos estocásticos

Incerteza e aleatoriedade

Modelo → forma de representar o comportamento e as propriedades de um sistema do mundo real;

Modelo determinístico:

- Input: não aleatório;
- Output: determinado pelo input, cálculo direto ou aproximação numérica;

Modelo estocástico:

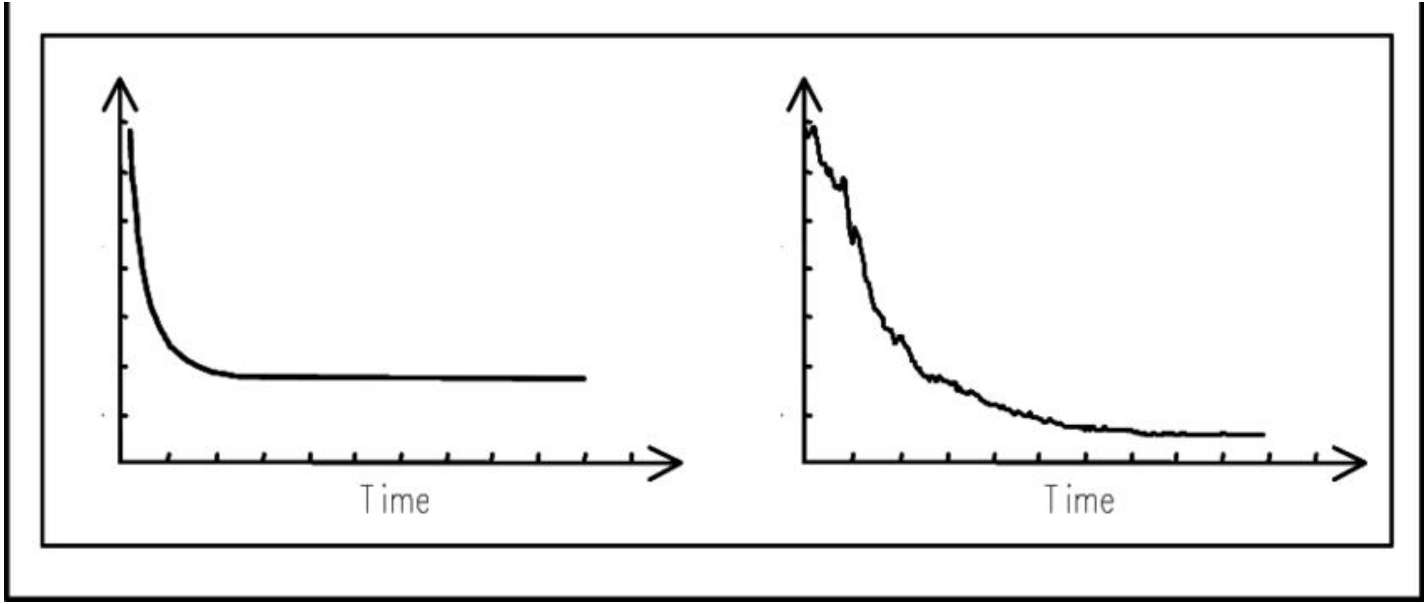
- Input: aleatório;
- Output: aleatório, simulação

Processos estocásticos

Modelo estocástico que evolui com o tempo (ou espaço) e gera uma série de valores.

Processo estocástico X Determinístico

Determinístico Estocástico



Processos estocásticos

Propriedades

Modelo estocástico que evolui com o tempo (ou espaço) e gera uma série de valores.

- Cada tempo (ou espaço) representa uma variável aleatória (X1, X2, X3, ..., Xn);
- Cada variável aleatória possui o mesmo espaço amostral;
- As variáveis aleatórias subsequentes são dependentes da “variável atual”.

Processos estocásticos

Categorias

- Arrival-time;
- **Processos de Markov**;

Processos de Markov

Definição

Processo estocástico;

Conjunto de variáveis aleatórios {X1, X2, ..., Xn} que modelam um processo ao longo do tempo e que representam medidas obtidas nesse espaço de tempo.

Tipo especial de dependência entre as variáveis: Xn+1 depende apenas do Xn.

Processo de Markov

$$P(X_{n+1}=x_{n+1}|X_n=x_n, X_{n-1}=x_{n-1}, \ldots X_0=x_0)=P(X_{n+1}=x_{n+1}|X_n=x_n)$$

Condição de Markov

Processo de Markov

Exemplo: Evolução do DNA

ATCGCCATCGAATACTCTAGCATG t=0

ATC**c**CCATCGAATACTCTAGCATG t=1

ATC**c**CCA**a**CGAATACTCTAGCATG t=2

ATC**c**CCA**a**CGAATAC**c**CTAGCATG t=3

A sequência no estado t=3 depende apenas da sequência no estado t=2 → **Processo de Markov**

Processo de Markov

Exemplo: Sequência do DNA

Padrão de nucleotídeos de 5’ ➔ 3’ nem sempre são independentes e pode ser modelado de forma que o nucleotídeo em uma posição seja dependente do nucleotídeo da posição anterior.

Modelo de Markov usado em uma aplicação não-MCMC ➔ HMM

Processos estocásticos

Classificações

Tempo (ou posição): índice da variável aleatória;

- Discreta;
- Contínua;

Espaço do estado: Valores que a variável aleatória pode assumir;

- Discreto;
- Contínuo;

Cadeias de Markov → Tempo discreto, Espaço discreto.

Random Walks

Cadeia de Markov simples

Estado x (x pode ser um inteiro qualquer) no tempo n;

No tempo n+1, o sistema moverá um ponto para cima ou para baixo, normalmente com a mesma probabilidade;

Cadeias de Markov

Modelos probabilísticos

Cadeia de Markov é um modelo probabilístico de um processo de Markov;

Requerimento:

Operações com matriz;

Básico de matriz

Uma tabela retangular de números;

No nosso contexto estamos interessados em **matrizes quadradas** e na **multiplicação de matrizes**;

Multiplicação de matriz ➔ Multiplicar os elementos da linha i da primeira matriz com os elementos da coluna j da segunda matriz para obter o elemento Cij.

$$A = \begin{bmatrix} a11 & a12 \\ a21 & a22 \end{bmatrix}$$

$$B = \begin{bmatrix} b11 & b12 \\ b21 & b22 \end{bmatrix}$$

$$C = AB = \begin{bmatrix} a11*b11+a12*b21 & a11*b12+a12*b22 \\ a21*b11+a22*b21 & a21*b12+a22*b22 \end{bmatrix} = \begin{bmatrix} c11 & c12 \\ c21 & c22 \end{bmatrix}$$

Básico de matriz em R

Em R, existe um tipo de objeto específico para lidar com matriz (matrix)

Declarando uma matriz em R:

```
matrix(data, nrow, ncol)
```

Por padrão, as colunas são preenchidas antes das linhas;

Outros parâmetros podem ser consultados no help do R.

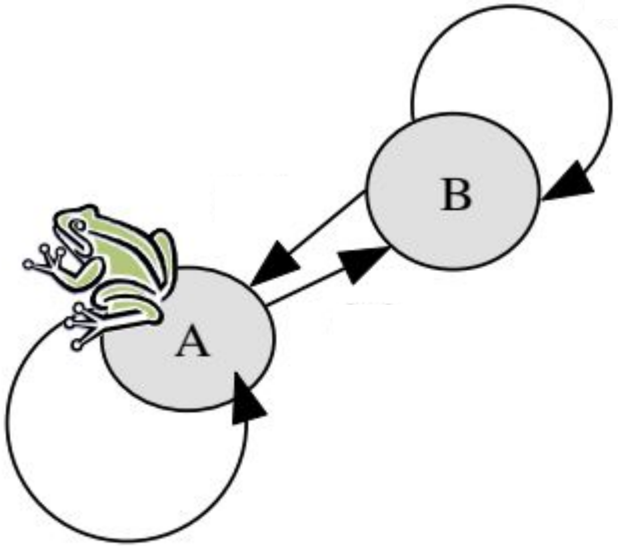
Um modelo simples de Cadeia de Markov

Suponha que um sapo vive solitariamente em um pequeno lago com duas folhas de vitória-régia;

O sapo fica apenas em uma das duas folhas (A e B);

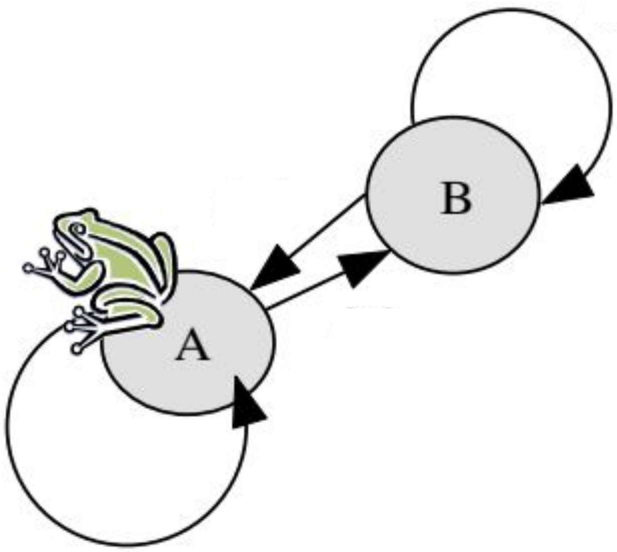
A cada 5 minutos, registramos em qual das folhas o sapo se encontrava.

Entre as observações, o sapo pode mover de uma folha para outra ou permanecer onde está.



Um modelo simples de Cadeia de Markov

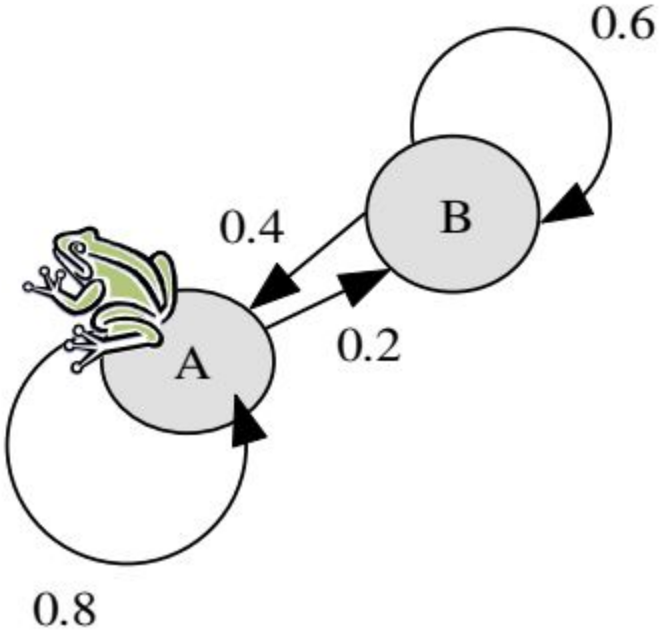
Se soubermos que o sapo se encontra na folha A, qual a probabilidade dele estar na folha A depois de 10 minutos?



Um modelo simples de Cadeia de Markov

O nosso modelo inicial sugere:

- Durante o primeiro intervalo (tempo 0 a tempo 1 ➔ 5 minutos):
  - Se o sapo está em A em t=0, a probabilidade de permanecer é de 0,8 e de mover é de 0,2;;
  - Se o sapo está em B em t=0, a probabilidade de permanecer é 0,6 e de mover é de 0,4.



Matriz de transição Para o primeiro movimento

$$\begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix}$$

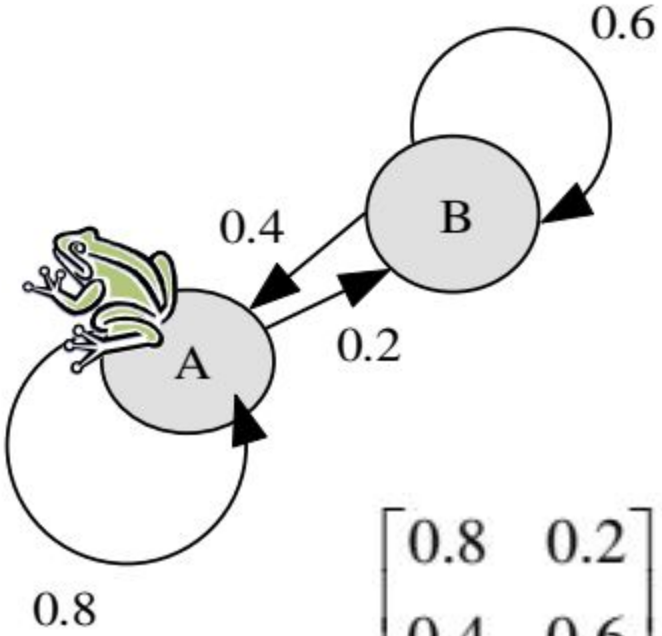
Um modelo simples de Cadeia de Markov

Probabilidade do sapo estar na folha A depois de dois períodos (10 minutos) dado que ele estava na folha A ➔ duas possibilidades:

- Início A ➔ A ➔ A;
- Início A ➔ B ➔ A;

P(sapo em A, t=2) = P(AA)P(AA) + P(AB)P(BA)

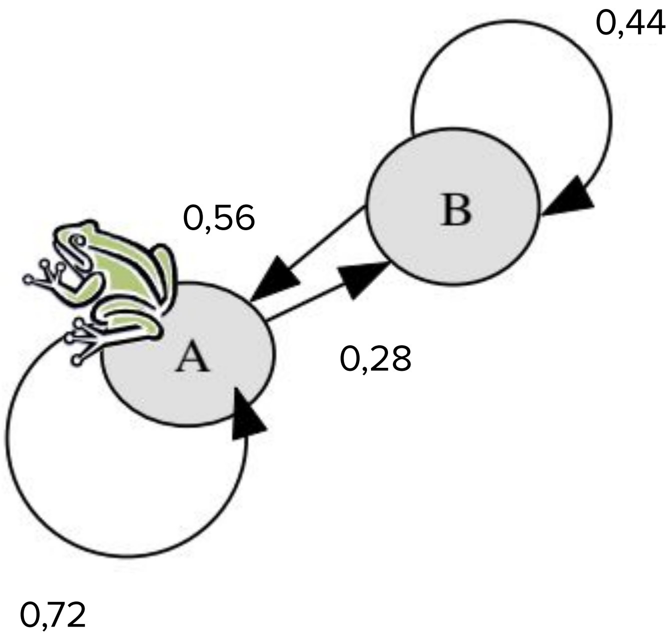
P(sapo em A, t=2) = 0,8*0,8 + 0,2*0,4 = 0,72



$$\begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix}$$

Um modelo simples de Cadeia de Markov

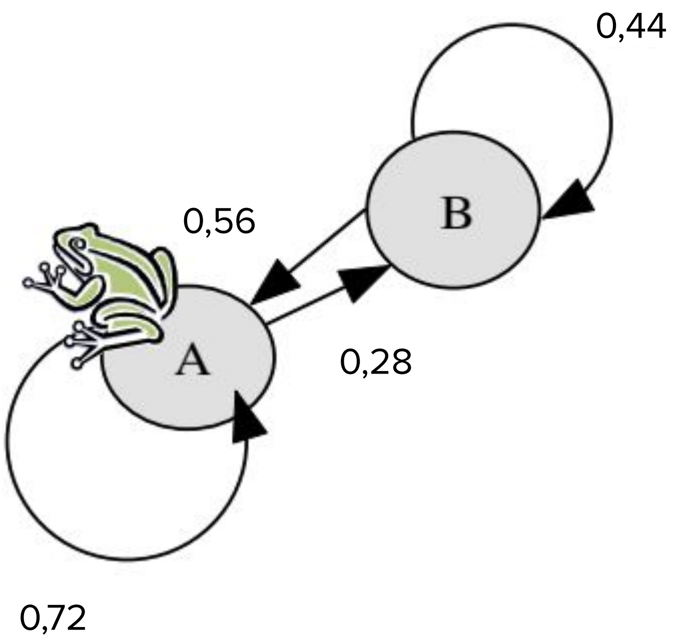
Para obter a próxima matriz de transição podemos utilizar o R e multiplicar a primeira matriz de transição com ela mesma:



Um modelo simples de Cadeia de Markov

Vamos observar o comportamento do modelo para tempos maiores.

A medida que nós aumentamos a potência da matriz de transição ocorre um comportamento peculiar...



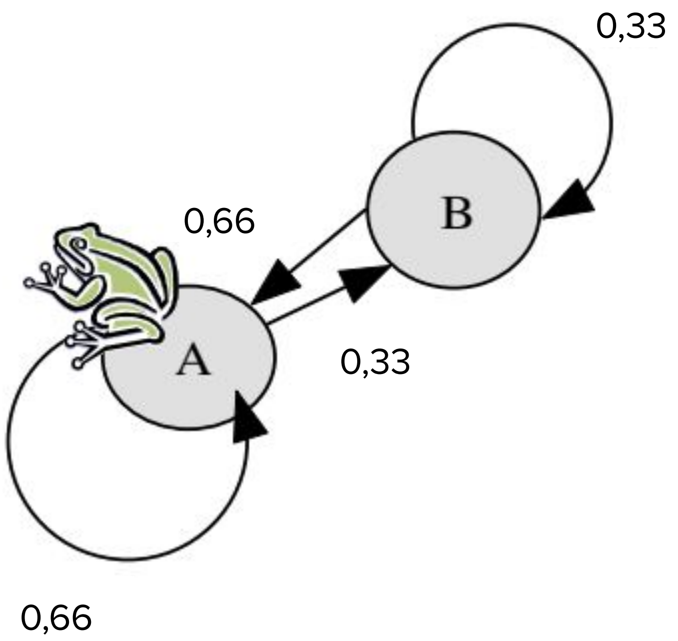
Um modelo simples de Cadeia de Markov

Vamos observar o comportamento do modelo para tempos maiores.

A medida que nós aumentamos a potência da matriz de transição ocorre um comportamento peculiar...

Depois de aproximadamente 20 intervalos, a matriz de transição não se altera.

Distribuição estacionária ➔ a cadeia convergiu para esta distribuição.



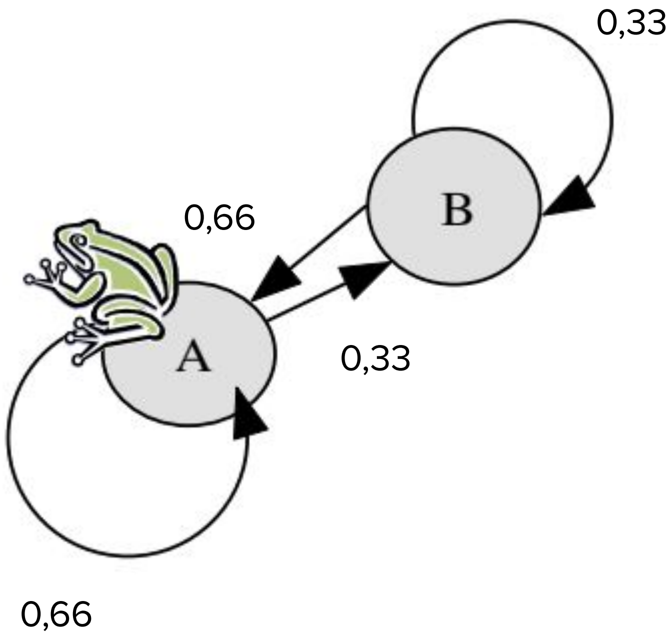
Um modelo simples de Cadeia de Markov

Distribuição a priori ➔ facilmente incorporada neste contexto:

Suponha que a probabilidade do sapo estar inicialmente em uma das folhas é a mesma:

p(A) = p(B) = 0,5

Podemos atualizar as probabilidades multiplicando-a pela matriz de transição;



Um modelo simples de Cadeia de Markov

Conceitos importantes:

- Cadeia de Markov como modelo probabilístico;
- Estados da cadeia de Markov (folha A e B em um dado tempo);
- Matriz de transição de um estado para o outro;
- Computar as probabilidades de transição em intervalos k>1;
- Distribuição estacionária ➔ Convergência.