

# MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística  
Universidade de São Paulo  
pam@ime.usp.br  
<http://www.ime.usp.br/~pam>

## Aula 21

4 de junho de 2025

# Sumário

- 1 Simulação dinâmica
- 2 Amostrador de Gibbs
- 3 Algoritmo de Metropolis-Hastings
- 4 Algoritmo EM

## Métodos MCMC

- Nesta seção vamos dar uma breve ideia sobre dois métodos usados para gerar amostras de uma dada função densidade de probabilidades.
- Suas origens remontam a Metropolis et al. (1953) e Hastings (1970), com interesse mais recente a partir dos artigos de Geman e Geman (1984) e Gelfand and Smith (1990).
- Estudaremos em especial o **amostrador de Gibbs (AG)** e o **algoritmo de Metropolis-Hastings (AMH)**.
- Atualmente, há uma vasta literatura sobre métodos MCMC (de **Markov chain Monte Carlo**), bem como aplicações nas mais diversas áreas.
- Referências: Gamerman e Lopes (2006), Gilks et al. (1996), Tanner (1996), Robert e Casella (2004) e Chen et al. (2000).
- Como o próprio nome indica, os algoritmos baseiam-se na teoria de **cadeias de Markov**. A idéia básica é que teremos que gerar uma tal cadeia e amostrar de sua distribuição estacionária, que supostamente coincidirá com a distribuição alvo da qual queremos amostrar. Para tanto teremos que construir adequadamente o núcleo dessa cadeia.

## O amostrador de Gibbs

- Suponha que temos uma densidade  $\pi(x)$ , possivelmente multivariada, da qual queremos retirar uma amostra.
- A idéia básica do Amostrador de Gibbs (AG) é construir uma cadeia de Markov cuja distribuição estacionária seja  $\pi(x)$ .
- Para tanto, é necessário construir o **núcleo (matriz) de transição** desta cadeia. Este será obtido a partir das distribuições condicionais completas.
- Suponha que  $\pi(\mathbf{x}) = \pi(x_1, \dots, x_p)$  e estamos interessados na marginal

$$\pi_1(x_1) = \int \cdots \int \pi(x_1, \dots, x_p) dx_2 \dots dx_p. \quad (1)$$

- Em particular queremos calcular a média ou variância de  $X_1$ . O cálculo de (1) pode ser complicado ou mesmo impossível. Obtida uma amostra de  $X_1$  podemos estimar a média  $E(X_1)$ , por exemplo, pela média amostral.

## O AG: duas variáveis

- Vamos considerar inicialmente o caso de duas variáveis, que chamaremos de  $X$  e  $Y$ , com densidade  $\pi(x, y)$ . Se conhecemos as condicionais  $\pi_{X|Y}(x|y)$  e  $\pi_{Y|X}(y|x)$ , então o AG gera uma amostra de  $\pi(x)$  (ou de  $\pi(y)$ ) do seguinte modo:

[1] Especificamos um valor inicial  $Y = y_0$ ; os demais valores são obtidos alternando-se entre as distribuições condicionais :

[2] Amostre de  $X$  e de  $Y$  conforme

$$\begin{aligned}x_j &\sim \pi_{X|Y}(x|Y_j = y_j), \\y_{j+1} &\sim \pi_{Y|X}(y|X_j = x_j).\end{aligned}\tag{2}$$

- Obtemos a **sequência de Gibbs**

$$y_0, x_0, y_1, x_1, \dots, y_k, x_k,\tag{3}$$

usando (2).

## O AG: duas variáveis

- O que se pode demonstrar é que, para  $k$  suficientemente grande,  $x_k$  é um valor de  $\pi(x)$ . Para obter uma amostra de tamanho  $m$ , digamos, podemos gerar  $m$  seqüências de Gibbs independentes e usar o valor final de cada seqüência.
- Obtida uma tal amostra, podemos estimar a densidade marginal de  $X$  por

$$\hat{\pi}(x) = \frac{1}{m} \sum_{i=1}^m \pi_{X|Y}(x|y_i), \quad (4)$$

onde cada  $y_i$  é o valor final de cada seqüência. Note que isto vem do fato que  $E[\pi_{X|Y}(x|y)] = \pi(x)$ .

- No caso discreto, a fórmula análoga é

$$\hat{P}(X = x) = \frac{1}{m} \sum_{i=1}^m P(X = x | Y_i = y_i). \quad (5)$$

- O que pode ser demonstrado é que o esquema iterativo (2) de fato produz uma amostra de  $\pi(x)$ .

## O AG: caso geral

- Vamos considerar a situação geral no contexto bayesiano, em que temos  $\pi(\theta)$  a densidade de interesse, com  $\theta = (\theta_1, \dots, \theta_p)'$ .
- Cada componente  $\theta_i$  pode ser um escalar, um vetor ou mesmo uma matriz. Suponha que podemos calcular as distribuições condicionais completas

$$\pi_i(\theta_i) = \pi(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p) = \pi(\theta_i | \theta_{-i}), \quad (6)$$

para  $i = 1, 2, \dots, p$ .

## O AG: caso geral

### Algoritmo AG:

- (1) Considere os valores iniciais  $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_p^{(0)})'$ ;
- (2) um novo valor  $\theta^{(j)} = (\theta_1^{(j)}, \dots, \theta_p^{(j)})'$  é obtido a partir de  $\theta^{(j-1)}$  por meio de gerações sucessivas de valores

$$\begin{aligned}\theta_1^{(j)} &\sim \pi(\theta_1 | \theta_2^{(j-1)}, \dots, \theta_p^{(j-1)}), \\ \theta_2^{(j)} &\sim \pi(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_p^{(j-1)}), \\ &\vdots \\ \theta_p^{(j)} &\sim \pi(\theta_p | \theta_1^{(j)}, \dots, \theta_{p-1}^{(j)}).\end{aligned}$$

- (3) Itere até convergência.

## O AG: caso geral

- Segue-se que os vetores  $\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(k)}, \dots$  são realizações de uma CM com núcleo de transição

$$p(\theta', \theta) = \pi(\theta_1 | \theta'_2, \dots, \theta'_p) \pi(\theta_2 | \theta_1, \theta'_3, \dots, \theta'_p) \dots \pi(\theta_p | \theta_1, \dots, \theta_{p-1}). \quad (7)$$

- É fácil ver que o esquema do AG define uma CM (o que acontece na iteração  $j$  depende somente da iteração  $j - 1$ ), e a cadeia é homogênea, pois podemos escrever

$$p(\theta^{(j-1)}, \theta^{(j)}) = \pi(\theta_1^{(j)} | \theta_2^{(j-1)}, \dots, \theta_p^{(j-1)}) \dots \pi(\theta_p^{(j)} | \theta_1^{(j)}, \dots, \theta_p^{(j-1)}), \quad (8)$$

que não varia com  $j$ .

## O algoritmo de Metropolis-Hastings

- Como antes, o objetivo é gerar uma amostra de uma distribuição  $\pi$ , por meio de uma cadeia de Markov. Construímos um núcleo de transição  $p(\theta, \phi)$  de forma que  $\pi$  seja a distribuição de equilíbrio da cadeia. Para tal, consideramos cadeias que satisfaçam a **condição de reversibilidade**

$$\pi(\theta)p(\theta, \phi) = \pi(\phi)p(\phi, \theta), \quad (9)$$

para todo  $(\theta, \phi)$ . Essa condição é suficiente para que  $\pi$  seja a distribuição de equilíbrio.

- Integrando ambos os membros, obtemos

$$\int \pi(\theta)p(\theta, \phi)dx = \int \pi(\phi)p(\phi, \theta)dx = \pi(\phi), \quad \text{para todo } \phi. \quad (10)$$

- O núcleo  $p(\theta, \phi)$  é constituído de dois elementos: um núcleo de transição arbitrário,  $q(\theta, \phi)$  e uma probabilidade  $\alpha(\theta, \phi)$ , tal que

$$p(\theta, \phi) = q(\theta, \phi)\alpha(\theta, \phi), \quad \theta \neq \phi. \quad (11)$$

## O algoritmo de Metropolis-Hastings

- O núcleo de transição define uma densidade  $p(\theta, \cdot)$ , para todos os valores distintos de  $\theta$ , logo há uma probabilidade positiva da cadeia ficar em  $\theta$ , dada por

$$p(\theta, \theta) = 1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi. \quad (12)$$

- De modo geral,

$$p(\theta, A) = \int_A q(\theta, \phi)\alpha(\theta, \phi)d\phi + I(\theta \in A)[1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi], \quad (13)$$

logo  $p(\cdot, \cdot)$  define uma distribuição mista para o novo estado  $\phi$  da cadeia de Markov. Para  $\theta \neq \phi$ , essa distribuição tem uma densidade e para  $\theta = \phi$ , essa distribuição atribui uma probabilidade positiva.

- A expressão mais comumente usada para a probabilidade de aceitação é

$$\alpha(\theta, \phi) = \min \left\{ 1, \frac{\pi(\phi)q(\phi, \theta)}{\pi(\theta)q(\theta, \phi)} \right\}. \quad (14)$$

## O algoritmo de Metropolis-Hastings

### Algoritmo de M-H

- (i) inicialize o número de iterações com  $j = 1$  e o valor inicial  $\theta^{(0)}$ ;
- (ii) mova a cadeia para um novo valor  $\phi$  gerado da densidade  $q(\theta^{(j-1)}, \cdot)$ ;
- (iii) calcule a probabilidade de aceitação do movimento,  $\alpha(\theta^{(j-1)}, \phi)$ , dada por (14); se o movimento for aceito,  $\theta^{(j)} = \phi$ ; se não,  $\theta^{(j)} = \theta^{(j-1)}$  e a cadeia não se move;
- (iv) mude o contador de  $j$  para  $j + 1$  e retorne a (ii) até convergência.

## O algoritmo de Metropolis-Hastings

- A etapa (iii) é realizada após a geração de um NA  $u$ , independente de todas as outras variáveis. Se  $u \leq \alpha$ , o movimento é aceito e se  $u > \alpha$  o movimento não é aceito. O núcleo de transição  $q$  define apenas uma proposta de movimento, que pode ou não ser confirmada por  $\alpha$ . O sucesso do método depende de taxas de aceitação não muito baixas (da ordem de 20%–50%) e de propostas  $q$  fáceis de simular.
- Algumas escolhas particulares:
  - (a) **Cadeias simétricas**:  $p(\theta, \phi) = p(\phi, \theta)$ , para todo par  $(\theta, \phi)$ . Essa é a versão original de Metropolis et al. (1953), que não depende de  $q$ , então (14) só depende de  $\pi(\phi)/\pi(\theta)$ , e nesse caso não é necessário conhecer a forma completa de  $\pi$ .
  - (b) **Passeio aleatório**: nesse caso a cadeia tem evolução dada por  $\theta^{(j)} = \theta^{(j-1)} + w_j$ , onde  $w_j$  é uma v.a. com distribuição independente da cadeia. Normalmente, toma-se  $w_j$  v.a.'s i.i.d. com densidade  $f_w$  e nesse caso  $q(\theta, \phi) = f_w(\phi - \theta)$ . Se  $f_w$  for simétrica (como no caso da normal e da  $t$ -Student), obtemos o caso (a).
  - (c) **Cadeias independentes**: a transição proposta é formulada independentemente da posição atual  $\theta$  da cadeia, ou seja,  $q(\theta, \phi) = f(\phi)$ .

## O algoritmo de Metropolis-Hastings

- **Metropolis-Hastings em Gibbs.** No amostrador de Gibbs as transições são baseadas nas distribuições condicionais completas das componentes de  $\theta$ . É possível que  $\pi$  tenha uma forma complicada, impossibilitando a geração de valores diretamente, mas alguma condicional completa  $\pi_i$  possa ser utilizada diretamente para a geração. Müller (1992) sugere que a geração dos componentes  $\theta_i$  para os quais não se pode gerar diretamente  $\pi_i$  seja feita por meio de uma subcadeia de M-H dentro do ciclo do amostrador de Gibbs.
- Para ver algoritmo de Metropolis, que posteriormente foi generalizado por Hastings (1970), veja Tanner (1996).
- Se  $q$  for **irreduzível e aperiódica** e  $\alpha(\theta, \phi) > 0$ , para todo  $(\theta, \phi)$ , então o algoritmo de M-H define uma cadeia de Markov aperiódica e irreduzível, com núcleo de transição  $p$  dada por (11) e distribuição de equilíbrio  $\pi$ . Veja Roberts e Smith (1994).

## Dados latentes e imputação

- Há dois algoritmos baseados no conceito de **dados latentes**, que podem ocorrer por diferentes mecanismos.
- Por exemplo, para facilitar o procedimento de amostragem da verossimilhança ou da densidade a posteriori de alguma variável  $Y$  para as quais temos observações, aumentamos os dados disponíveis introduzindo dados  $Z$ , chamados **latentes ou não observados**; esse é o caso do **algoritmo de dados aumentados** ou do **algoritmo EM**.
- Outros procedimentos que utilizam dados latentes são os vários métodos de imputação e os métodos de reamostragem e de reamostragem ponderada que foram tratados na Aula 20. O leitor interessado poderá consultar Tanner (1996) para detalhes.

## Dados latentes e imputação

- Originalmente, a ideia de imputação múltipla estava associada com valores omissos oriundos de respostas omissas em levantamentos amostrais. Esta formulação foi estendida para outras áreas.
- O método **MCMC** (*Markov Chain Monte Carlo*), descrito acima, é um exemplo de procedimento que visa mimetizar informação não observada.
- Esse tópico de observações omissas remonta a Wilks (1932) e Anderson (1957) e talvez a primeira formulação sistematizada seja aquela de Rubin (1977).
- Outras referências sobre esse tema incluem Rubin (1987), Little e Rubin (1987) e Rubin (1996).
- As técnicas que se valem de dados latentes também são úteis no caso de dados incompletos, que ilustramos por meio de um exemplo.

## Dados latentes e imputação - Exemplo

- Consideremos dados observados periodicamente (por exemplo, peixes presos em armadilhas).
- Sejam  $x_1, \dots, x_7$  as quantidades capturadas nos dias  $1, \dots, 7$  que consideraremos como **dados completos**.
- Agora imaginemos que as observações ocorrem apenas nos dias 1, 3, 4 e 7 de forma cumulativa, gerando os dados  
 $y_1 = x_1 + x_2, y_2 = x_3, y_3 = x_4 + x_5, y_4 = x_6 + x_7$ .
- Os valores  $y_1, y_2, y_3$  e  $y_4$  constituem os **dados incompletos**.
- Se considerarmos observações (não cumulativas) somente nos dias 1, 2, 5 e 7, então  $y_1 = x_1, y_2 = x_2, y_3 = x_5, y_4 = x_7$  constituem os dados incompletos, nesse caso.

## O algoritmo EM

- Consideremos dados  $\mathbf{x} = (x_1, \dots, x_n)^\top$  provenientes do modelo  $f(\mathbf{x}|\boldsymbol{\theta})$  obtidos com o objetivo de encontrar o estimador de máxima verossimilhança de  $\boldsymbol{\theta}$ , ou seja, de maximizar a verossimilhança  $L(\boldsymbol{\theta}|\mathbf{x})$ .
- Alternativamente, o objetivo poderia ser encontrar a moda da distribuição a posteriori  $p(\boldsymbol{\theta}|\mathbf{x})$ . Concentremo-nos inicialmente na maximização da verossimilhança.
- Suponhamos que os dados  $\mathbf{x}$  não sejam completamente observados mas que alguma função de  $\mathbf{x}$ , digamos  $\mathbf{y} = \mathbf{h}(\mathbf{x})$  o seja. Diremos que os elementos de  $\mathbf{x}$  constituem os **dados completos** enquanto aqueles de  $\mathbf{y}$  constituem os **dados incompletos**.

## O algoritmo EM

- A verossimilhança dos dados observados (incompletos)  $\mathbf{y}$  é

$$L(\boldsymbol{\theta}|\mathbf{y}) = \int_{\mathcal{X}(\mathbf{y})} L(\boldsymbol{\theta}|\mathbf{x})d\mathbf{x}, \quad (15)$$

em que  $\mathcal{X}(\mathbf{y})$  é a parte do espaço amostral  $\mathcal{X}$  de  $\mathbf{x}$  determinada pela restrição  $\mathbf{y} = \mathbf{h}(\mathbf{x})$ .

- O algoritmo EM é um procedimento iterativo segundo o qual encontramos o valor de  $\boldsymbol{\theta}$  que maximiza a verossimilhança dos dados observados,  $L(\boldsymbol{\theta}|\mathbf{y})$ , usando  $L(\boldsymbol{\theta}|\mathbf{x})$  de maneira conveniente. “Conveniente” aqui significa escolher  $L(\boldsymbol{\theta}|\mathbf{x})$  de tal forma que  $L(\boldsymbol{\theta}|\mathbf{y})$  seja obtida por meio de (15).

## O algoritmo EM - Exemplo

- Retomemos o problema da ligação genética discutido no Exemplo 2 da Aula 20, segundo o qual os dados  $\mathbf{y} = (y_1, y_2, y_3, y_4)^\top = (125, 18, 20, 34)^\top$  ocorrem com probabilidades  $\pi_1 = (2 + \theta)/4$ ,  $\pi_2 = \pi_3 = (1 - \theta)/4$ ,  $\pi_4 = \theta/4$ ,  $0 < \theta < 1$ .
- A verossimilhança dos dados incompletos é

$$L(\theta|\mathbf{y}) = \frac{(y_1 + \dots + y_4)!}{y_1! \dots y_4!} \left(\frac{2 + \theta}{4}\right)^{y_1} \left(\frac{1 - \theta}{4}\right)^{y_2 + y_3} \left(\frac{\theta}{4}\right)^{y_4},$$

cujo núcleo é

$$\left(\frac{2 + \theta}{4}\right)^{y_1} \left(\frac{1 - \theta}{4}\right)^{y_2 + y_3} \left(\frac{\theta}{4}\right)^{y_4}.$$

- Admitamos que  $\mathbf{x} = (x_1, \dots, x_5)^\top$  sejam os dados completos que ocorrem, respectivamente, com probabilidades  $[1/2, \theta/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4]$ , mas que só observamos  $y_1 = x_1 + x_2$ ,  $y_2 = x_3$ ,  $y_3 = x_4$  e  $y_4 = x_5$ , de modo que o núcleo da verossimilhança dos dados completos seja

$$L(\theta|\mathbf{x}) \propto \left(\frac{\theta}{4}\right)^{x_2 + x_5} \left(\frac{1 - \theta}{4}\right)^{x_3 + x_4}.$$

## O algoritmo EM - Exemplo

- Podemos simplificar as expressões acima obtendo

$$L(\theta|\mathbf{y}) \propto (2 + \theta)^{y_1} (1 - \theta)^{y_2 + y_3} \theta^{y_4}, \quad (16)$$

e

$$L(\theta|\mathbf{x}) \propto \theta^{x_2 + x_5} (1 - \theta)^{x_3 + x_4}. \quad (17)$$

- Neste caso, a expressão (15) se reduz a

$$L(\theta|\mathbf{y}) = \sum_{x_1, x_2 | x_1 + x_2 = 125} L(\theta | x_1, x_2, 18, 20, 34).$$

## O algoritmo EM - Exemplo

Nesse contexto, o algoritmo **EM** pode ser explicitado como:

- i) Escolher um valor inicial,  $\theta^{(0)}$ .
- ii) Obter a esperança condicional de  $\mathbf{x}$ , dado  $\mathbf{y}$  (passo **E**), ou seja, estimar os dados completos por meio de suas esperanças condicionais, dados  $\mathbf{y}$  e  $\theta^{(0)}$ , observando que  $E(x_3|\mathbf{y}, \theta^{(0)}) = 18$ ,  $E(x_4|\mathbf{y}, \theta^{(0)}) = 20$ ,  $E(x_5|\mathbf{y}, \theta^{(0)}) = 34$  e que

$$E(x_1|\mathbf{y}, \theta^{(0)}) = E(x_1|x_1 + x_2 = 125, \theta^{(0)}) = x_1^{(0)},$$

$$E(x_2|\mathbf{y}, \theta^{(0)}) = E(x_2|x_1 + x_2 = 125, \theta^{(0)}) = x_2^{(0)}.$$

Como

$$x_1|x_1 + x_2 = 125 \sim \text{Bin}(n, p_1), \quad n = 125, \quad p_1 = \frac{1/2}{1/2 + \theta/4} = \frac{2}{2 + \theta^{(0)}},$$

$$x_2|x_1 + x_2 = 125 \sim \text{Bin}(n, p_2), \quad n = 125, \quad p_2 = \frac{\theta^{(0)}}{2 + \theta^{(0)}},$$

obtemos  $x_1^{(0)} = 250/(2 + \theta^{(0)})$  e  $x_2^{(0)} = 125\theta^{(0)}/(2 + \theta^{(0)})$ . Neste passo, os dados completos estimados são  $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, 18, 20, 34)^\top$ .

- iii) O passo **M** consiste em maximizar a log-verossimilhança dos dados completos

$$\ell(\theta|\mathbf{x}^{(0)}) = x_1^{(0)} \log\left(\frac{1}{2}\right) + x_2^{(0)} \log\left(\frac{\theta}{4}\right) + x_3 \log\left(\frac{1-\theta}{4}\right) + x_4 \log\left(\frac{1-\theta}{4}\right) + x_5 \log\left(\frac{\theta}{4}\right),$$

que é proporcional a

$$(x_2^{(0)} + x_5) \log(\theta) + (x_3 + x_4) \log(1 - \theta).$$

## O algoritmo EM - Exemplo

- Derivando essa expressão em relação a  $\theta$ , obtemos

$$\frac{d\ell(\theta)}{d\theta} = \frac{x_2^{(0)} + x_5}{\theta} - \frac{x_3 + x_4}{1 - \theta};$$

- igualando-a a zero temos, finalmente,

$$\theta^{(1)} = \frac{x_2^{(0)} + x_5}{x_2^{(0)} + x_3 + x_4 + x_5} = \frac{x_2^{(0)} + 34}{x_2^{(0)} + 72}. \quad (18)$$

- De modo geral, dada a estimativa na iteração  $i$ ,  $\theta^{(i)}$ , estimamos os dados latentes por meio de

$$x_1^{(i)} = \frac{250}{2 + \theta^{(i)}}, \quad x_2^{(i)} = \frac{125\theta^{(i)}}{2 + \theta^{(i)}}$$

e atualizamos o estimador de  $\theta$  por intermédio de

$$\theta^{(i+1)} = \frac{x_2^{(i)} + 34}{x_2^{(i)} + 72}. \quad (19)$$

## O algoritmo EM - Exemplo

- Se tomarmos, por exemplo,  $\theta^{(0)} = 0,5$ , obteremos as iterações da Tabela 1, que produzem a estimativa  $\hat{\theta} = 0,62682$ .

Tabela 1: Iterações do algoritmo **EM** para o modelo genético

Iteração ( $i$ )	$\theta^{(i)}$
0	0,50000
1	0,60800
2	0,62400
3	0,62648
4	0,62677
5	0,62681
6	0,62682
7	0,62682

- As estimativas para as probabilidades  $\pi_i$  são,  $\hat{\pi}_1 = 0,6567$ ,  $\hat{\pi}_2 = \hat{\pi}_3 = 0,0933$ ,  $\hat{\pi}_4 = 0,1567$ .

## EM - caso geral

- O algoritmo **EM** pode ser usado para maximizar tanto a verossimilhança  $L(\theta|\mathbf{y})$  quanto a distribuição *a posteriori*  $p(\theta|\mathbf{y})$ .
- Neste caso, é preciso considerar a distribuição *a posteriori* aumentada  $p(\theta|\mathbf{y}, \mathbf{z}) = p(\theta|\mathbf{x})$  e a densidade  $p(\mathbf{z}|\mathbf{y}, \theta^{(i)})$ , que é a distribuição preditora dos dados latentes  $\mathbf{z}$ , condicional ao valor atual da moda e aos dados observados.
- No Exemplo, em que indicamos os passos necessários para a implementação do algoritmo **EM** em termos da verossimilhança, essa distribuição é a distribuição binomial com parâmetros  $n = 125$  e  $p = \theta^{(i)} / (2 + \theta^{(i)})$ .
- Para a densidade *a posteriori* as modificações são óbvias.

## EM - caso geral

Para a maximização da verossimilhança, os passos do algoritmo **EM** são:

i) **Passo E**: calcular

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}) = E[\ell(\boldsymbol{\theta}|\mathbf{x})|\mathbf{y}, \boldsymbol{\theta}^{(i)}], \quad (20)$$

ou seja, a esperança condicional da log-verossimilhança aumentada, considerando os dados  $\mathbf{y}$  e o valor atual  $\boldsymbol{\theta}^{(i)}$ .

ii) **Passo M**: obter o valor  $\boldsymbol{\theta}^{(i+1)}$  no espaço paramétrico que maximiza  $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)})$ .

iii) Iterar até a convergência, ou seja, até que  $\|\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)}\|$  ou  $|Q(\boldsymbol{\theta}^{(i+1)}, \boldsymbol{\theta}^{(i)}) - Q(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(i)})|$  sejam suficientemente pequenas.

No caso da distribuição *a posteriori*, considerar

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}) = \int_{\mathbf{z}} \log p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(i)}) d\mathbf{z}. \quad (21)$$

## EM - caso geral

Retornando ao Exemplo, temos

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E[(x_2 + x_5) \log \theta + (x_3 + x_4) \log(1 - \theta) | \mathbf{y}, \theta^{(i)}] \\ &= [E(x_2 | \mathbf{y}, \theta^{(i)}) + x_5] \log \theta + (x_3 + x_4) \log(1 - \theta). \end{aligned}$$

Lembremos que, aqui,

$$E(x_2 | \mathbf{y}, \theta^{(i)}) = \frac{125\theta^{(i)}}{2 + \theta^{(i)}}.$$

e maximizando  $Q(\theta, \theta^{(i)})$ , obtemos

$$\theta^{(i+1)} = \frac{E(x_2 | \mathbf{y}, \theta^{(i)}) + x_5}{E(x_2 | \mathbf{y}, \theta^{(i)}) + x_3 + x_4 + x_5}.$$

## Resultados teóricos

- Dempster et al. (1977) provam alguns resultados de convergência relativos ao algoritmo, mas com incorreções. As correções foram feitas por Boyles (1983) e Wu (1983).
- **Proposição 1.** Seja  $\ell(\theta)$  a log-verossimilhança dos dados observados. Então,  $\ell(\theta^{(i+1)}) \geq \ell(\theta^{(i)})$ , ou seja, toda iteração do algoritmo **EM** aumenta o valor da log-verossimilhança.
- O mesmo resultado vale para a densidade a posteriori.
- Dempster et al. (1977) definem um algoritmo **EM** generalizado, que seleciona  $\theta^{(i+1)}$  de modo que  $Q(\theta^{(i+1)}, \theta^{(i)}) > Q(\theta^{(i)}, \theta^{(i)})$ . Então, o enunciado da Proposição 1 também é aplicável a esse algoritmo generalizado.
- A convergência do algoritmo pode ser lenta, com uma taxa linear que depende da quantidade de informação sobre  $\theta$  disponível em  $L(\theta|\mathbf{y})$ .
- Existem métodos para acelerar o algoritmo. Por exemplo, Louis (1982) sugere um procedimento para alcançar uma taxa de convergência quadrática perto do máximo por meio do método de Newton-Raphson. Outra referência importante sobre esse tópico é Meng e Rubin (1993).

## Referências

- Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.
- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *The American Statistician*, bf 49, 327–335.
- Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo*. Boca Raton: Chapman & Hall.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6** , 721–741.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
- Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Second Edition. New York: Springer.
- Tanner, M.A. (1996). *Tools for Statistical Inference, 3rd Ed.*. New York: Springer.