Fundação Getúlio Vargas
Escola de Matemática Aplicada

Rener Oliveira

Inferência Estatística Trabalho 2: Algoritmo EM

Setembro de 2020

Sumário

Sumário **1**

1 O Algoritmo **2**
1.1 Glossário (Notações) 2

2 Exemplo das Moedas **3**
2.1 Passo E 5
2.2 Passo M 6
2.3 Conclusões 8
2.4 Simulação computacional 9

3 Demonstração da Monotonicidade **11**

4 Comentários Finais **12**

Referências **14**

1 O Algoritmo

1.1 Glossário (Notações)

- $\vec{X}$: Dados observados;
- $\vec{Z}$: Dados faltantes;
- Ω: Espaço de parâmentros;
- $\theta^{(j)}$: Estimador de θ na iteração j do algoritmo EM;
- $\mathcal{L}(\theta; \vec{X}, \vec{Z})$: Verossimilhança dos dados completos;
- $\mathcal{L}(\theta; \vec{X})$: Verossimilhança dos dados observados (incompletos);
- $f(x, z|\theta)$: Distribuição conjunta dos dados completos ($f(x, z|\theta) = \mathcal{L}(\theta; \vec{X}, \vec{Z})$);
- $g(x|\theta)$: Distribuição dos dados observados ($g(x|\theta) = \mathcal{L}(\theta; \vec{X})$).

O Estimador de Máxima Verossimilhança (EMV) em muitas situações práticas pode ser difícil de ser computado. Um exemplo recorrente é quando temos em nossa amostra um subconjunto de dados faltantes *("missing data")*;
Uma solução pra esse problema é o famoso Algoritmo EM[4][7] (*"Expectation-Maximization"*) que é uma método iterativo para aproxima o EMV nessas situações de dados faltantes.
De forma geral, queremos um estimador para o vetor de parâmetros $\theta \in \Omega$ $\mathcal{L}(\theta; \vec{X})$.
$\hat{\theta}_{EMV} = \arg\max_{\theta \in \Omega} \mathcal{L}(\theta; \vec{X})$
Mas $\mathcal{L}(\theta; \vec{X}) = \displaystyle\int f(x, z|\theta) d\vec{Z}$ e não iremos trabalhar com a maximização direta dessa integral, mas usaremos o seguinte processo iterativo:

- Passo "E": Dado um $\theta^{(j)}$, o passo *Expectation*, consiste em "eliminar" de certa forma a lacuna dos dados faltantes, usando o valor esperado da log-verossimilhança dos dados completos com respeito a $\theta^{(j)}$[1] e $\vec{X}$. Ou seja, definimos uma função $Q(\theta|\theta^{(j)})$, onde:

$$Q(\theta|\theta^{(j)}) = E[\ln \mathcal{L}(\theta; \vec{X}, \vec{Z})|\theta^{(j)}, \vec{X}]$$

[1] Como condição inicial $\theta^{(0)}$ podemos pegar qualquer vetor de Ω

- Passo "M": inicial de *Maximization* este passo consiste em encontrar o valor que maximiza a função acima. Este valor será plugado como $\theta^{(j+1)}$ e a iteração continuará. Formalmente:[7]

$$\theta^{(j+1)} = \arg\max_{\theta \in \Omega} Q(\theta|\theta^{(j)})$$

De fato, não é difícil mostrar que

$$\mathcal{L}(\theta|\vec{X}) \geq Q(\theta|\theta^{(j)}) = E[\ln \mathcal{L}(\theta; \vec{X}, \vec{Z})|\theta^{(j)}, \vec{X}],$$

as notas de [5] mostram isso para o caso discreto e o artigo [6] usa este fato para o caso contínuo. Tendo essa desigualdade, é fácil ver que o algoritmo ao maximizar $Q(\theta|\theta^{(j)})$ está maximizando $\mathcal{L}(\theta|\vec{X})$ que é nosso objetivo. Daremos mais detalhes na seção de demonstração da monotonicidade do método.

2 Exemplo das Moedas

Problema (Transcrição):
Suponha que temos duas moedas, Moeda 1 e Moeda 2 de modo que $Pr(Cara|Moeda = 1) = p_1$ e $Pr(Cara|Moeda = 2) = p_2$; Suponha agora que fazemos o seguinte experimento:

(i) Selecionamos uma moeda aleatoriamente com probabilidade 1/2;

(ii) Lançamos a moeda selecionada m vezes;

(iii) Repetimos (i) e (ii) n vezes.

Podemos representar os dados advindos deste experimento como:

$$\begin{array}{ccccc} X_{11} & ... & X_{1m} & & M_1 \\ X_{21} & ... & X_{2m} & & M_2 \\ \vdots & ... & \vdots & & \vdots \\ X_{n1} & ... & X_{nm} & & M_n \end{array}$$

onde os X_{ij} são variáveis de Bernoulli que guardam o resultados do lançamento da moeda e $M_i \in \{1, 2\}$ é a variável aleatória que guarda qual moeda foi selecionada na i-ésima rodada do experimento.

Desenvolveremos aqui um esquema EM para aproximar o EMV de $\theta = (p_1, p_2)$ quando desconhecemos os valores de M_i.
Este é um problema clássico, conhecido como Binomial Mixture [5], na qual se tem um conjunto de tipos de moedas com probabilidades de dar cara diferentes, seleciona-se uma dessas moedas e realizam-se experimentos binomiais (bernoulli repetidamente). No final, ficamos com o conjunto de observações dos resultados, mas não sabemos qual tipo da moeda que gerou cada resultado, e o objetivo da aplicação do método EM é estimar o vetor de probabilidades dos tipos da moeda.
Nosso problema é um caso particular da referência [5], pois temos apenas dois tipos de moedas, e a probabilidade de escolher uma ou outra é igual a $1/2$. O caso geral é bem interessante, pois além de explorar uma quantidade variável de experimentos a cada rodada (aqui temos fixo m) ele explora também o desconhecimento das probabilidades de escolha entre os tipos da moeda, que passam a incorporar o vetor de parâmetros a ser estimado.
Seguem algumas definições que usaremos:
Notações e Definições

- θ_i: p_1 se $M_i = 1$, ou p_2 se $M_i = 2$;
- $S_i = \sum_{j=1}^{m} X_{ij}$, neste caso, S_i tem distribuição binomial de parâmetros m e θ_i;
- $\vec{X}$: Vetor de dados incompletos $(S_1, \ldots, S_n)$
- $\vec{M} = (M_1, \ldots, M_n)$ (dados faltantes)

O que estamos fazendo é sumarizando a informação matricial dos experimentos em um vetor que contem a quantidade de caras de cada experimento. Este vetor será composto por distribuições binomiais, e tiraremos proveito disso para derivar a fórmula iterativa de $\theta^{(r)}$.
O processo que seguiremos para aproximar o EMV de θ é o seguinte:

- Escolher $\theta^{(0)}$ qualquer em $\in (0,1) \times (0,1)$;
- (Passo E)Computamos a função $Q(\theta, \theta^{(j)}) = E[\ln \mathcal{L}(\theta; \vec{X}, \vec{M}) | \theta^{(j)}, \vec{X}]$
- (Passo M) Escolhemos $\theta^{(j+1)} = \underset{\theta \in \Omega}{\arg\max}\, Q(\theta | \theta^{(j)})$

- Repetimos os dois itens anteriores até a condição de parada, que pode ser, atingimento de tolerância, $|\theta^{(j+1)} - \theta^{(j)}| < \varepsilon$, para algum ε inicialmente definido, ou quando um número máximo de iterações pré-definido é atingido.

Para realizar os passos E e M, vamos derivar a função $Q(\theta, \theta^{(j)})$ explicitamente.

2.1 Passo E

Primeiramente vamos escrever a verossimilhança dos dados completos.
Para cada i, temos[2]
$P(X_i, M_i|\theta)$

[3]$= P(X_i|M_i, \theta) \cdot P(M_i|\theta)$
$= \frac{1}{2} \cdot P(X_i|M_i, \theta)$

Como $X_i = S_i$ é binomial de parâmetros m e θ_i, temos que:

$P(X_i, M_i|\theta) = \frac{1}{2} \cdot Bin(X_i, \theta_i)$,

onde $Bin(X_i, \theta_i) = \binom{m}{S_i}\theta_i^{S_i}(1-\theta_i)^{m-S_i}$. Estamos usando uma notação similar[4] às notas de [5], porém omitimos o m de $Bin(X_i|m, \theta_i)$ pois no nosso caso é uma valor fixo para todo i.
Dessa forma, a verossimilhança será:
$\mathcal{L}(\theta; \vec{X}, \vec{M}) = \prod_{i=1}^{n} P(X_i, M_i|\theta) =$
$\prod_{i=1}^{n} \frac{1}{2} Bin(X_i, \theta_i)$
Tomando o logaritmo natural (log do produto é a soma dos logs), teremos:
$\ln \mathcal{L}(\theta; \vec{X}, \vec{M}) = n \ln \frac{1}{2} + \sum_{i=1}^{n} \ln Bin(X_i, \theta_i)$

[2]Como $\vec{X} = (S_1, ..., S_n)$, definiremos $X_i = S_i$ para todo i de 1 a n
[3]Probabilidade Condicional $P(A, B|C) = P(A|B, C) \cdot P(B|C)$
[4]Na verdade estamos cometendo um abuso de notação usando $Bin(X_i, \theta_i)$ como função que é igual a probabilidade de uma binomial (m, θ_i) ser igual a X_i

Temos então, por definição:
$Q(\theta,\theta^{(j)}) = E[\mathcal{L}(\theta;\vec{X},\vec{M})|\vec{X},\theta^{(j)}]$
$= E\left[n\ln\frac{1}{2} + \sum_{i=1}^{n} \ln Bin(X_i,\theta_i) \middle| \vec{X},\theta^{(j)}\right]$
$n\ln\frac{1}{2} + \sum_{i=1}^{n} E[\ln Bin(X_i,\theta_i)|\vec{X},\theta^{(j)}]$
Finalizamos então a etapa *Expectation*, computando a função Q:

$$Q(\theta,\theta^{(j)}) = n\ln\frac{1}{2} + \sum_{i=1}^{n} E[\ln Bin(X_i,\theta_i)|\vec{X},\theta^{(j)}] \tag{1}$$

2.2 Passo M

Vamos agora, maximizar a função acima. Usaremos derivação, ao fazer isso estamos supondo algumas condições de regularidade na função Q, mas isso será melhor detalhado na seção seguinte. Por enquanto, vamos aceitar que podemos fazer isso sem problemas.
Nosso objetivo nessa etapa é, encontrar:

$$\theta^{(j+1)} = \arg\max_{\theta\in(0,1)^2}\left(n\ln\frac{1}{2} + \sum_{i=1}^{n} E[\ln Bin(X_i,\theta_i)|\vec{X},\theta^{(j)}]\right)$$

Mas como $n\ln\frac{1}{2}$ não depende de θ, a expressão acima é igual a:

$$\arg\max_{\theta\in(0,1)^2}\left(\sum_{i=1}^{n} E[\ln Bin(X_i,\theta_i)|\vec{X},\theta^{(j)}]\right)$$

O processo de maximização consiste em computar θ_1 e θ_2 tal que $\partial Q/\partial\theta_1 = 0$ e $\partial Q/\partial\theta_2 = 0$; teremos então o vetor $\theta^{(j+1)} = (\theta_1,\theta_2)$
Vamos calcular, por simplicidade, apenas $\partial Q/\partial\theta_1$, e veremos que o processo para θ_2 é completamente análogo.
Como vimos acima, o argmax de Q foi reduzimo para um expressão mais simples. Vamos trabalhar então com a derivada dessa expressão:

$$\frac{\partial}{\partial\theta_1}\left(\sum_{i=1}^{n} E[\ln Bin(X_i,\theta_i)|\vec{X},\theta^{(j)}]\right) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta_1}E[\ln Bin(X_i,\theta_i)|\vec{X},\theta^{(j)}]$$

Note que, fixado i, o termo $E[\ln Bin(X_i, \theta_i)|X_i, \theta^{(j)}]$[5]pode ser escrito como:

$$\ln[Bin(X_i, \theta_1)] \cdot P(M_i = 1|X_i, \theta^{(j)}) + \ln[Bin(X_i, \theta_2)] \cdot P(M_i = 2|X_i, \theta^{(j)}), \quad (2)$$

na qual, pela definição de probabilidade conjunta condicional, temos:

$$P(M_i = 1|X_i, \theta^{(j)}) = \frac{P(M_i = 1, X_i|\theta^{(j)})}{P(X_i|\theta^{(j)})} =$$

$$\frac{P(M_i = 1)P(X_i|\theta^{(j)})}{P(X_i|\theta_1^{(j)})P(M_i = 1) + P(X_i|\theta_2^{(j)})P(M_i = 2)}$$

Mas por hipótese, $P(M_i = 1) = P(M_i = 2) = \frac{1}{2}$, assim, cancelamos todos esses termos e obtemos:

$$P(M_i = 1|X_i, \theta^{(j)}) = \frac{P(X_i|\theta^{(j)})}{P(X_i|\theta_1^{(j)}) + P(X_i|\theta_2^{(j)})} =$$

$$\frac{Bin(X_i, \theta_1^{(j)})}{Bin(X_i, \theta_1^{(j)}) + Bin(X_i, \theta_2^{(j)})} \quad (3)$$

Para fins computacionais[5], computaremos a quantidade acima usando seu valor explícito:

$$P(M_i = 1|X_i, \theta^{(j)}) = \left[1 + \left(\frac{\theta_2^{(j)}}{\theta_1^{(j)}}\right)^{X_i} \left(\frac{1 - \theta_2^{(j)}}{1 - \theta_1^{(j)}}\right)^{m - X_i}\right]^{-1} \quad (4)$$

Para $P(M_i = 2|X_i, \theta^{(j)})$ as expressões são análogas.

Note o termo $P(M_i = 1|X_i, \theta^{(j)})$ da expressão (2) é constante em relação a θ_1. Note também que a segunda parcela, não dependem de θ_1, logo, ao derivarmos a expressão com respeito a θ_1, teremos:

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_1} E[\ln Bin(X_i, \theta_i)|\vec{X}, \theta^{(j)}]$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_1} \ln[Bin(X_i, \theta_1)] \cdot P(M_i = 1|X_i, \theta^{(j)})$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_1} \ln[Bin(X_i, \theta_1)] \cdot \frac{Bin(X_i, \theta_1^{(j)})}{Bin(X_i, \theta_1^{(j)}) + Bin(X_i, \theta_2^{(j)})}$$

[5]Note que trocamos $\vec{X}$ por X_i pois pela independência dos experimentos, o único elemento do vetor $\vec{X}$ com informações de interesse é $X_i = S_i$.

Fazendo $B_i = \dfrac{Bin(X_i, \theta_1^{(j)})}{Bin(X_i, \theta_1^{(j)}) + Bin(X_i, \theta_2^{(j)})}$, temos:

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_1} E[\ln Bin(X_i, \theta_i) | \vec{X}, \theta^{(j)}]$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_1} \ln[Bin(X_i, \theta_1)] \cdot B_i.$$

Com alguns cálculos, que omitirei, é possível chegar em:

$$\frac{\partial}{\partial \theta_1} \ln[Bin(X_i, \theta_1)] = \frac{X_i - \theta_1 m}{\theta_1(1-\theta_1)};$$

Continuando, teremos a derivada igual a:

$$\sum_{i=1}^{n} \frac{X_i - \theta_1 m}{\theta_1(1-\theta_1)} \cdot B_i$$

Queremos θ_1 que zere a expressão acima. Note que $\theta_{\in}(0,1) \Rightarrow \theta_1(1-\theta_1) \neq 0$, logo, podemos encontrar tal valor, fazendo:

$$\sum_{i=1}^{n} (X_i - \theta_1 m) B_i = 0 \Rightarrow$$

$$\sum_{i=1}^{n} X_i B_i - \theta_1 m B_i = 0 \Rightarrow$$

$$\sum_{i=1}^{n} X_i B_i - \theta_1 m \sum_{i=1}^{n} B_i = 0 \Rightarrow$$

$$\theta_1 = \frac{\sum_{i=1}^{n} X_i B_i}{m \sum_{i=1}^{n} B_i}$$

Onde $B_i = P(M_i = 1 | X_i, \theta^{(j)})$ que pode ser computado pelas expressões (3) ou (4). A expressão acima é o nossa primeira coordenada de $\theta^{(j+1)}$. Para encontrar o segundo valor, os passos e resultados são análogos.

2.3 Conclusões

A sequência do método EM será dada por $\theta^{(j+1)} = (\theta_1, \theta_2)$, onde:

$\theta_1 = \dfrac{\sum_{i=1}^{n} X_i B_i}{m \sum_{i=1}^{n} B_i}$ na qual B_i depende de $\theta^{(j)}$ e é dado por (3) ou (4), e

$\theta_2 = \dfrac{\sum_{i=1}^{n} X_i B_i'}{m \sum_{i=1}^{n} B_i'}$ na qual B_i' é dado por:

$$\frac{Bin(X_i, \theta_2^{(j)})}{Bin(X_i, \theta_1^{(j)}) + Bin(X_i, \theta_2^{(j)})}$$

ou

$$\left[1 + \left(\frac{\theta_1^{(j)}}{\theta_2^{(j)}}\right)^{X_i} \left(\frac{1-\theta_1^{(j)}}{1-\theta_2^{(j)}}\right)^{m-X_i}\right]^{-1}$$

2.4 Simulação computacional

Fizemos uma simulação usando Python 3.7.6. Geramos uma tabela de experimentos onde as duas moedas tem probabilidades $p_1 = 0.3$ e $p_2 = 0.6$ de dar cara. O valor inicial foi escolhido aleatoriamente e chutou um valor menor que 0.2 para as duas quantidades. Note que apesar de estar próximo de p_1, está bem longe de p_2.

Usamos $n = 300$ e $m = 30$, ou seja, 300 experimentos de 30 lançamentos. Por simplicidade de implementação, a condição de parada é apenas um número máximo pré-fixado de iterações, mas o ideal seria implementar uma tolerância para o aumento da log-verossimilhança.

Os resultados são apresentados através de dois gráficos: o primeiro, exibido na Figura (1) compara o método EM com o estimador de máxima verossimilhança de p_1 e p_2, que foram computados como a proporção de caras em que cada moeda registrou, omitiremos aqui a derivação desse estimador. Podemos ver que com poucas iterações, o método convergiu para o MLE (*"Maximum Likelihod Estimator"*). Além disso, $p1\_MLE \approx 0.3063$ e $p2\_MLE \approx 0.6165$, valores bem próximos dos reais.

É de se esperar que o método EM convirja também para os valores reais dos parâmetros, que é uma afirmação que se sustenta dada a consistência do MLE da bernoulli/binomial, que não demonstraremos aqui, assim, no limite, o MLE converge em probabilidade para o valor real.

O segundo gráfico, exibido da Figura (2) é quase idêntico ao citado acima, mas agora o contraste é feito com o valor real dos parâmetros. Note que o método EM superestima o valor de p_2, mas isso ocorre pois o próprio MLE faz essa ligeira superestimação; Já que o método EM converge para o MLE, estamos sujeitos à esse tipo de incerteza do próprio MLE.

Note também nas figuras que o valor inicial (aleatório) foi bem distante do valor do MLE do parâmetros, mas o processo iterativo foi muito eficiente,

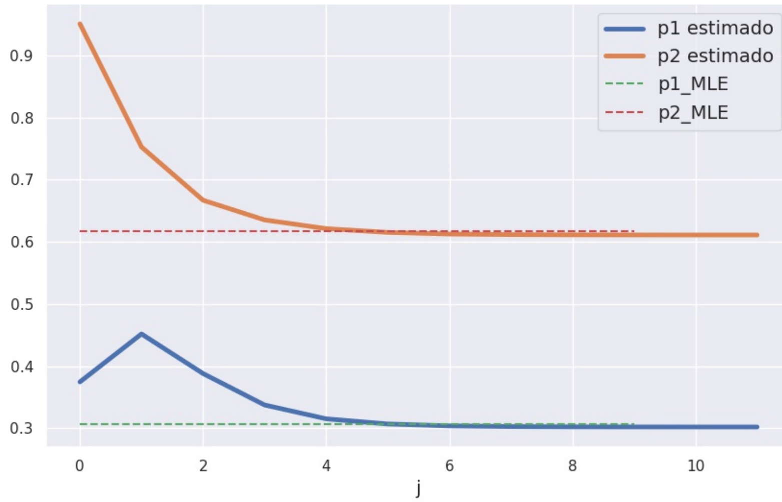

Figura 1: Simulação Método EM vs. MLE(n=300, m=30)

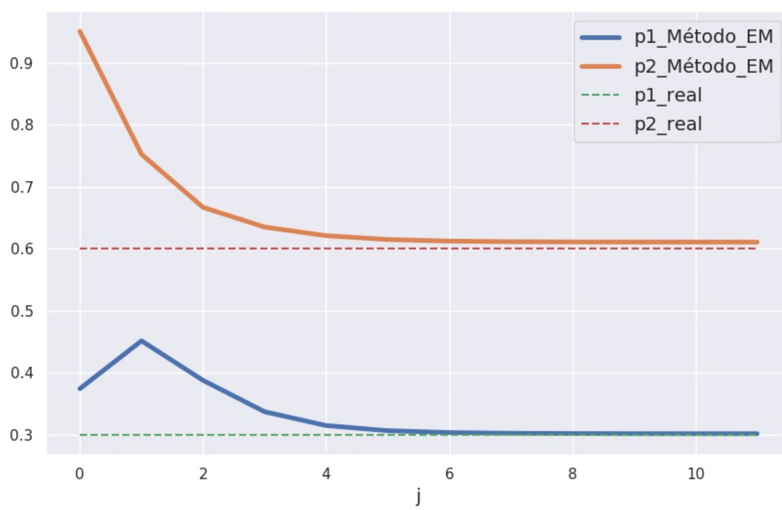

Figura 2: Simulação Método EM vs. Valor Real(n=300, m=30)

e com poucos passos convergiu, podemos perceber visualmente que após a quinta iteração (j=5) as curvas se aproximam de uma reta.

3 Demonstração da Monotonicidade

Queremos provar o Teorema 7.2.20 de [1]:
Teorema 7.2.20 (Adaptado): A sequência $\{\theta^{(j)}\}$ definida como $\theta^{(j+1)} = \arg\max_{\theta\in\Omega} Q(\theta|\theta^{(j)})$, onde $Q(\theta|\theta^{(j)}) = E[\ln \mathcal{L}(\theta; \vec{X}, \vec{Z})|\theta^{(j)}, \vec{X}]$ satisfaz:

$$\mathcal{L}(\theta^{(j+1)}; \vec{X}) \geq \mathcal{L}(\theta^{(j)}; \vec{X})$$

Demostração:
OBS: Usaremos as notações e definições do glossário da seção 1.
Seja $k(z|\theta, x) = \dfrac{f(x, z|\theta)}{g(x|\theta)}$ a distribuição de $\vec{Z}$ dados θ e $\vec{X}$. Sabendo que $\mathcal{L}(\theta; \vec{X}, \vec{Z}) = f(x, z|\theta)$ e $\mathcal{L}(\theta; \vec{X}) = g(x|\theta)$, temos que:
$\ln k(z|\theta, x) = \ln \mathcal{L}(\theta; \vec{X}, \vec{Z}) - \ln \mathcal{L}(\theta; \vec{X})$,
Logo:
$\ln \mathcal{L}(\theta; \vec{X}) = \ln \mathcal{L}(\theta; \vec{X}, \vec{Z}) - \ln k(z|\theta, x)$
Tomando o valor esperado[1] com respeito à distribuição de $k(z|\theta^{(j)}, x)$, o primeiro membro permanecerá como está, pois não há termos de $\vec{Z}$ livres; Assim teremos:
$\ln \mathcal{L}(\theta; \vec{X}) = E[\ln \mathcal{L}(\theta; \vec{X}, \vec{Z})|\theta^{(j)}, \vec{X}] - E[\ln k(z|\theta, x)|\theta^{(j)}, \vec{X}]$
Vamos definir $H(\theta, \theta^{(j)}) := -E[\ln k(z|\theta, x)|\theta^{(j)}, \vec{X}]$. É chamada de função de entropia[7] em outros contextos, mas vamos manter aqui como uma simples definição para simplificação de escrita.
Temos da última equação a seguinte identidade:

$$\ln \mathcal{L}(\theta; \vec{X}) = Q(\theta, \theta^{(j)}) + H(\theta, \theta^{(j)}) \tag{5}$$

Que vale para todo θ no espaço de parâmetros, e em particular para $\theta^{(j)}$. Assim, podemos escrever:
$\ln \mathcal{L}(\theta^{(j)}; \vec{X}) = Q(\theta^{(j)}, \theta^{(j)}) + H(\theta^{(j)}, \theta^{(j)})$
Como $\theta^{(j+1)}$ é argmax de $Q(\theta, \theta^{(j)})$, temos por definição que $\forall\theta \in \Omega, Q(\theta^{(j+1)}, \theta^{(j)}) \geq Q(\theta, \theta^{(j)})$. Portanto $Q(\theta^{(j+1)}, \theta^{(j)}) \geq Q(\theta^{(j)}, \theta^{(j)})$, o que prova o item (a) do exercício 7.32 de Casella[2].
Nosso objetivo é provar que $\ln \mathcal{L}(\theta^{(j+1)}; \vec{X}) \geq \ln \mathcal{L}(\theta^{(j)}; \vec{X})$, e pela identidade (5), Se provarmos que $Q(\theta^{(j+1)}, \theta^{(j)}) \geq Q(\theta^{(j)}, \theta^{(j)})$ (já feito) e $H(\theta^{(j+1)}, \theta^{(j)}) \geq H(\theta^{(j)}, \theta^{(j)})$, o Teorema fica demonstrado.
Resta-nos então provar $H(\theta^{(j+1)}, \theta^{(j)}) \geq H(\theta^{(j)}, \theta^{(j)})$ que é o item (b) do exercícios já citado.

Usando a Desigualdade de Jensen[3] para funções côncavas (e a concavidade de ln), podemos mostrar facilmente a dica do item (b) do exercício que afirma que, se f e g são funções de densidade de probabilidade, temos:

$$\int \ln[f(x)]g(x)dx \leq \int \ln[g(x)]g(x)dx \tag{6}$$

Tomemos então $E[\ln k(z|\theta, x)|\theta^{(j)}, \vec{X}]$. Por definição, temos:
$E[\ln k(z|\theta, x)|\theta^{(j)}, \vec{X}] = \int \ln k(z|\theta, x) \ln k(z|\theta^{(j)}, \vec{X})d\vec{Z}$
e de (6), temos:
$\int \ln k(z|\theta, x) \ln k(z|\theta^{(j)}, \vec{X})d\vec{Z}$
$\leq \int \ln k(z|\theta^{(j)}, \vec{X}) \ln k(z|\theta^{(j)}, \vec{X})d\vec{Z}$
$= E[\ln k(z|\theta^{(j)}, \vec{X})|\theta^{(j)}, \vec{X}]$,
onde a última igualdade vem da definição de esperança condicional.
Com isso, concluímos que
$\forall\theta \in \Omega;\ E[\ln k(z|\theta, x)|\theta^{(j)}, \vec{X}] \leq E[\ln k(z|\theta^{(j)}, \vec{X})|\theta^{(j)}, \vec{X}]$
Em particular:
$E[\ln k(z|\theta^{(j+1)}, x)|\theta^{(j)}, \vec{X}] \leq E[\ln k(z|\theta^{(j)}, \vec{X})|\theta^{(j)}, \vec{X}]$
$\Rightarrow -E[\ln k(z|\theta^{(j+1)}, x)|\theta^{(j)}, \vec{X}] \geq -E[\ln k(z|\theta^{(j)}, \vec{X})|\theta^{(j)}, \vec{X}]$
$\Rightarrow H(\theta^{(j+1)}, \theta^{(j)}) \geq H(\theta^{(j)}, \theta^{(j)})$
O que conclui a demonstração do **Teorema**.
■

4 Comentários Finais

O método EM é bastante utilizado em aplicações de machine learning por exemplo, em casos de missing data já mencionados. Entretanto nem tudo são flores, e ele não pode ser aplicado em todos os casos de dados faltantes. Nos casos em que o percentual de dado faltante represente muito do dado total, a estimação pode não ficar boa.
Além disso, vimos na questão anterior que a iteração é monótona e não-decrescente, o que garante que a sequência convirja para um mínimo local, mas não dá nenhuma garantia de convergência para mínimo global, o que pode ser um problema nos casos em que temos vários pontos críticos, A

convergência global nesses casos passa a depender do valor inicial $\theta^{(0)}$, o que não é bom.

Referências

[1] George Casella and Roger Berger. *Statistical Inference*, pages 326–329. Duxbury Resource Center, June 2001.

[2] George Casella and Roger Berger. *Statistical Inference*, page 361. Duxbury Resource Center, June 2001.

[3] George Casella and Roger Berger. *Statistical Inference*, page 190. Duxbury Resource Center, June 2001.

[4] M.H. DeGroot and M.J. Schervish. *Probability and Statistics, 4th ed.*, pages 434–439. Addison-Wesley, 2012.

[5] Jing Luan. Binomial mixture model with expectation maximum (em) algorithm. `https://medium.com/@jingluan.xw/binomial-mixture-model-with-expectation-maximum` 2016. [Online; Acesso 13 de Setembro 2020].

[6] Ajit Singh. The em algorithm. *Recuperado de: http://www. cs. cmu. edu/˜ awm/15781/assignments/EM. pdf*, 2005.

[7] Wikipedia. Expectation–maximization algorithm. `http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization%20algorithm`, 2020. [Online; Acesso 13 de Setembro 2020].