Misturas e Algoritmo EM

Renato Martins Assunção

DCC - UFMG

Misturas

- Os modelos básicos de distribuições que dispomos são flexíveis mas não dão conta de tudo que ocorre.
- Será raro que um conjunto de instancias seja muito bem modelado por uma das poucas distribuições que aprendemos.
- Temos duas alternativas:
  - Aumentar o nosso "dicionário de distribuições" criando uma looooooooonga lista de distribuições para ajustar aos dados reais.
  - Misturar os tipos básicos já definidos para ampliar a classe de distribuições disponíveis para analise.
- Uma forma de misturar é construir um modelo de regressão: Cada indivíduo tem uma distribuição que é modulada pelas suas variáveis independentes ou features.
- E quando não tivermos covariáveis mas tivermos claramente dados vindos de 2 ou mais distribuições?

Um exemplo

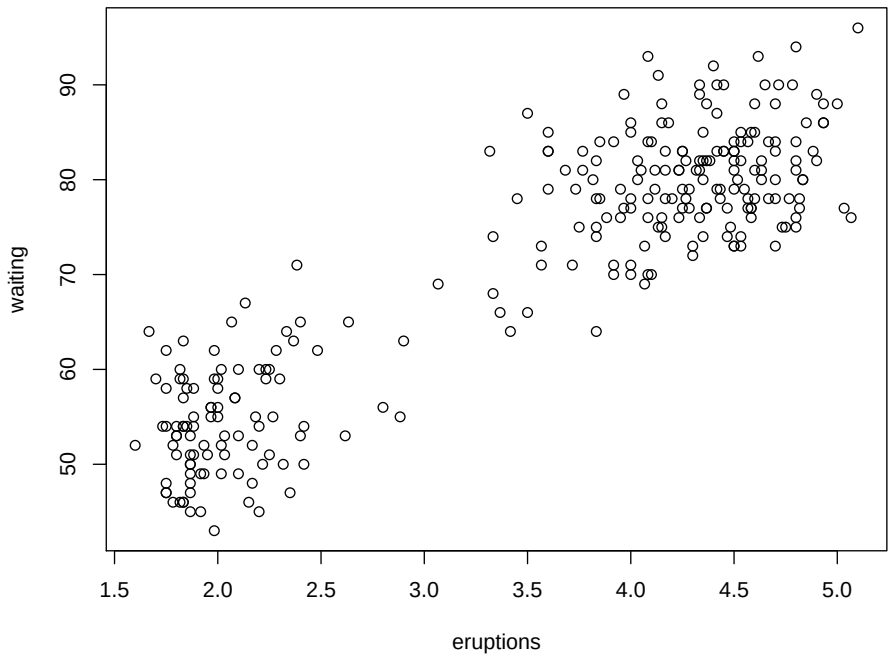


Figura: Dados de $n = 272$ erupções da geyser Faithful do Parque Yellowstone nos EUA. No eixo horizontal, a duração de cada erupção. No eixo vertical, temos o intervalo entre a erupção em questão e a erupção seguinte. Parece que existem duas normais bivariadas misturadas.

Mistura de 3 normais

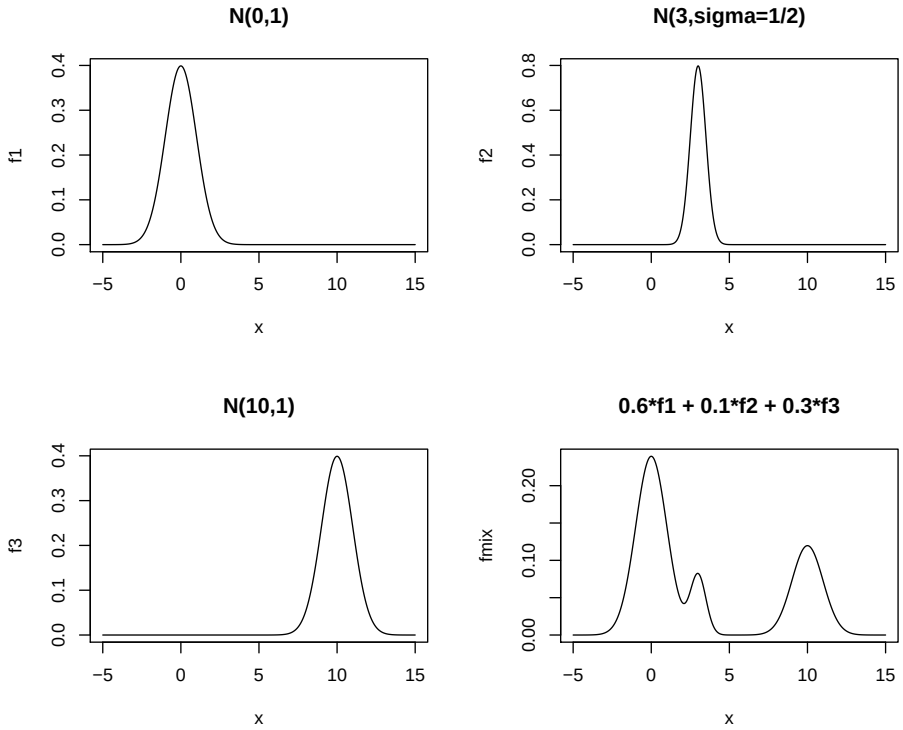


Figura: Mistura de 3 normais.

Amostra desta mistura

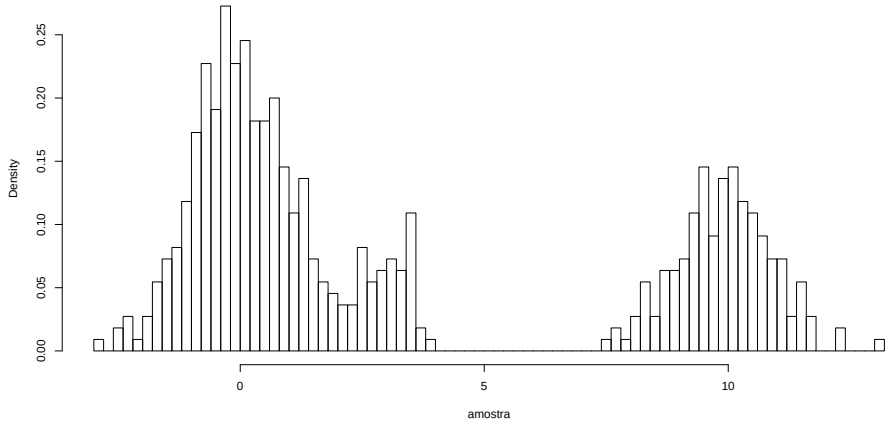


Figura: Amostra de $n = 550$ de dados vindos da densidade mistura de 3 normais.

Sobrepondo a densidade

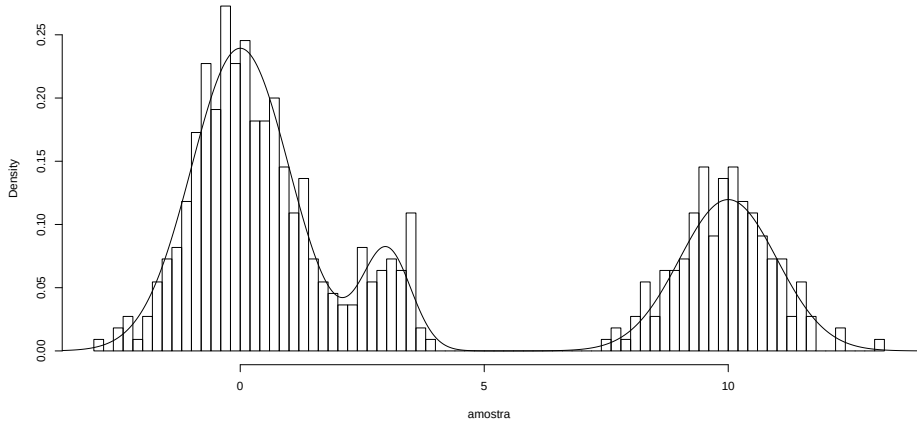


Figura: Amostra anterior com a densidade da mistura sobreposta.

Mistura no caso de v.a. discreta

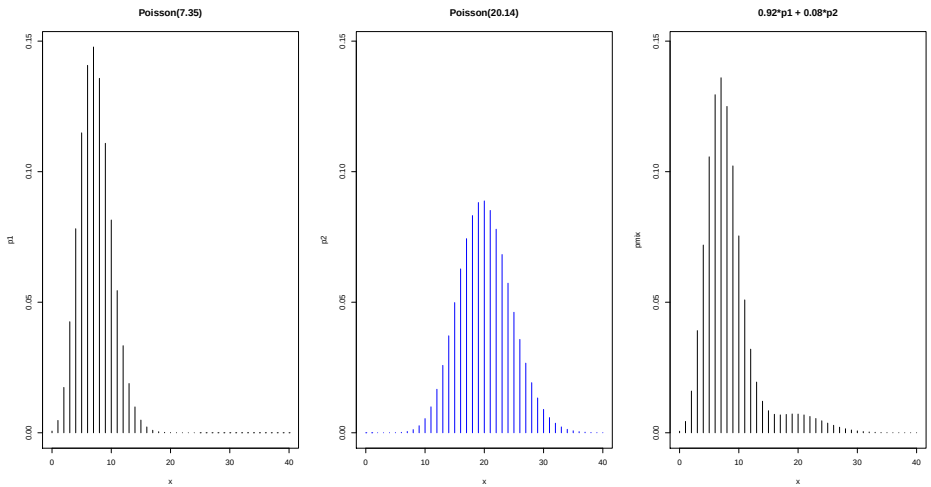


Figura: Mistura: 92% vêm de uma Poisson($\lambda = 7.35$) e os outros 8% vêm de Poisson($\lambda = 20.1$)

Amostra da mistura de e Poissons

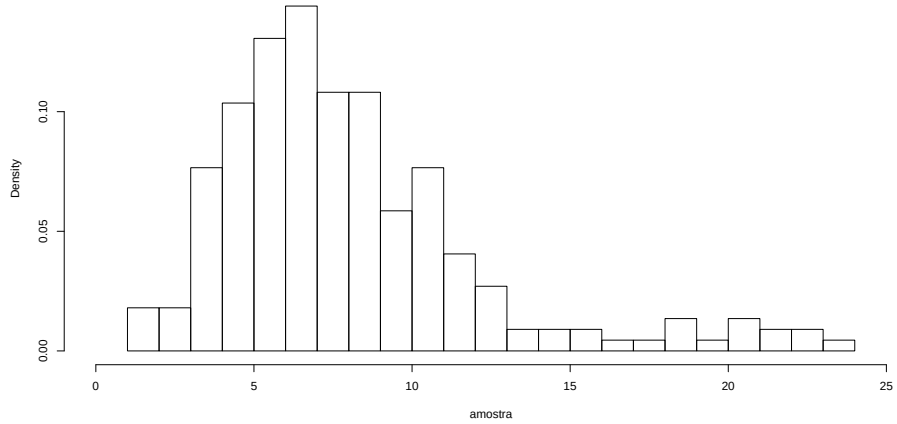


Figura: Amostra de $n = 222$ casos de parto cesáreo com complicações graves. Dados adaptados de Xiao et. al. (1999). Mistura: 92% vêm de uma Poisson($\lambda = 7.35$) e os outros 8% vêm de Poisson($\lambda = 20.1$).

Misturas: caso contínuo

- Estamos olhando o atributo Y
- Suponha que temos três sub-populações: 1, 2 e 3
- Represente as medições nas diferentes sub-populações como v.a.'s Y_1, Y_2, e Y_3.
- As sub-populações sao diferentes $\rightarrow$ as v.a.'s tem densidades diferentes
- As densidades são: $f_1(y)$, $f_2(y)$ e $f_3(y)$ e as respectivas distribuições acumuladas são $F_1(y)$, $F_2(y)$ e $F_3(y)$.
- Assim, $F_1'(y) = f_1(y)$, $F_2'(y) = f_2(y)$ e $F_3'(y) = f_3(y)$.
- Exemplo: $1 \rightarrow N(0, 1)$, densidade $2 \rightarrow N(3, \sigma^2 = 1/2^2)$ e $3 \rightarrow N(10, 1)$

Mistura de 3 normais

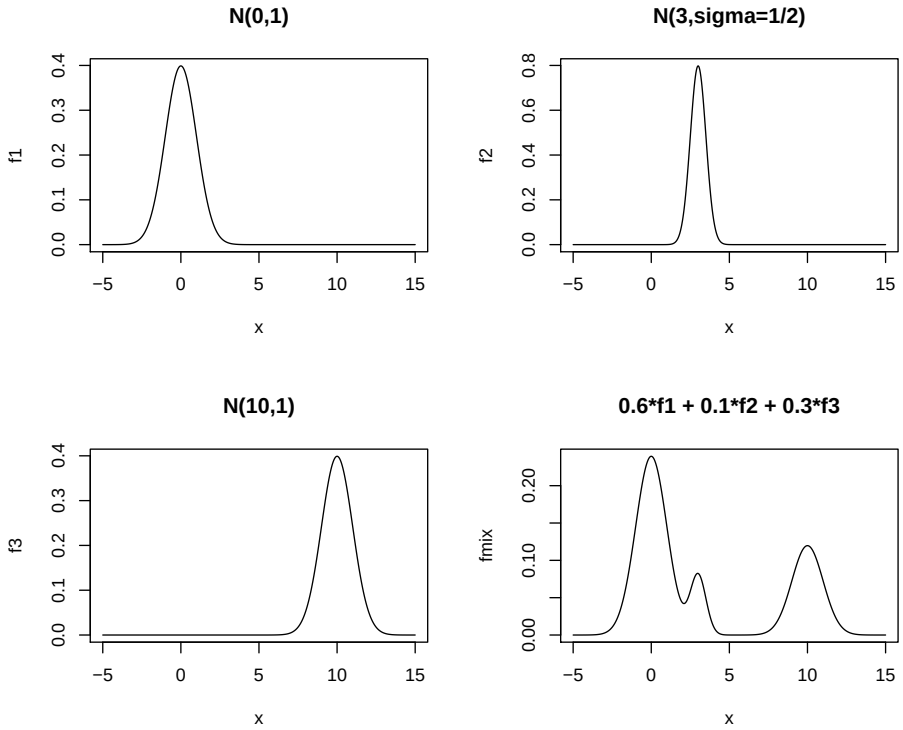


Figura: Mistura de 3 normais.

Misturas: caso contínuo

- A variável medida é representada por Y.
- Qual a distribuição de probabilidade da v.a. Y?
- Se o individuo vier da população 1, Y terá a mesma distribuição que a v.a. Y_1
- Se vier da população 2, $Y \sim Y_2$
- Se vier da população 3, $Y \sim Y_3$
- O individuo da população mistura vem de UMA das três populações aleatoriamente.
- Ele vem das 3 populações com as seguintes probabilidades:
  - Vem da população 1 com probabilidade θ_1
  - Vem da população 2 com probabilidade θ_2
  - Vem da população 3 com probabilidade θ_3
- Com $\theta_1 + \theta_2 + \theta_3 = 1$

Distribuição de mistura

- Assim, a medição Y tem a seguinte estrutura aleatória:
- y tem a mesma distribuição que Y_1 com probab θ_1 ou, de forma mais compacta:
  - $Y \sim Y_1$ com probabilidade θ_1
  - $Y \sim Y_2$ com probabilidade θ_2
  - $Y \sim Y_3$ com probabilidade θ_3
- Qual a densidade de Y?
- Usamos a formula da probabilidade total para calcular $\mathbb{F}(y) = \mathbb{P}(Y \leq y)$.
- Vamos condicionar no resultado de qual população ele foi amostrado e a seguir somamos (de forma ponderada) sobre as três possíveis populações.

Probab total para $\mathbb{F}(y)$

- Temos

$$\mathbb{F}(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(Y \leq y \text{ e vem de alguma pop})$$

- Temos a igualdade de eventos

$$\begin{aligned}[Y \leq y] &= [Y \leq y \cap \text{ vem de pop 1}] \cup [Y \leq y \cap \text{ vem de pop 2}] \\ &\cup [Y \leq y \cap \text{ vem de pop 3}]\end{aligned}$$

- Como os eventos são disjuntos, a probab da união é a soma das probabs:

$$\begin{aligned}\mathbb{F}(y) &= \mathbb{P}(Y \leq y \cap \text{ vem de pop 1}) + \mathbb{P}(Y \leq y \cap \text{ vem de pop 2}) + \mathbb{P}(Y \leq y \\ &= \mathbb{P}(Y \leq y|\text{pop 1})\mathbb{P}(\text{pop 1}) + \mathbb{P}(Y \leq y|\text{pop 2})\mathbb{P}(\text{pop 2}) + \mathbb{P}(Y \leq y|\text{po} \\ &= \mathbb{P}_1(Y \leq y)\theta_1 + \mathbb{P}_2(Y \leq y)\theta_2 + \mathbb{P}_3(Y \leq y)\theta_3 \\ &= \mathbb{F}_1(y)\theta_1 + \mathbb{F}_2(y)\theta_2 + \mathbb{F}_3(y)\theta_3\end{aligned}$$

- $\mathbb{F}(y)$ é uma média ponderada das dist acumuladas $\mathbb{F}_i(y)$ das componentes da mistura

..

- Revendo
$$\mathbb{F}(y) = \mathbb{F}_1(y)\theta_1 + \mathbb{F}_2(y)\theta_2 + \mathbb{F}_3(y)\theta_3$$
- $\mathbb{F}(y)$ é uma média ponderada das dist acumuladas $\mathbb{F}_i(y)$ das componentes da mistura
- Outra maneira de dizer isto é: a distribuição acumulada da mistura é a mistura das distribuições acumuladas.
- A dist acumulada não é intuitiva. A densidade é mais interpretável.

..

- Se temos a distribuição acumulada, podemos obter a densidade de Y derivando $\mathbb{F}(y)$:

$$\begin{aligned} f(y) &= \mathbb{F}'(y) = \mathbb{F}'_1(y)\theta_1 + \mathbb{F}'_2(y)\theta_2 + \mathbb{F}'_3(y)\theta_3 \\ &= f_1(y)\theta_1 + f_2(y)\theta_2 + f_3(y)\theta_3 \end{aligned}$$

- A densidade da mistura Y é a mistura das densidades das componentes Y_1, Y_2 e Y_3.

Densidade da mistura é a mistura das densidades

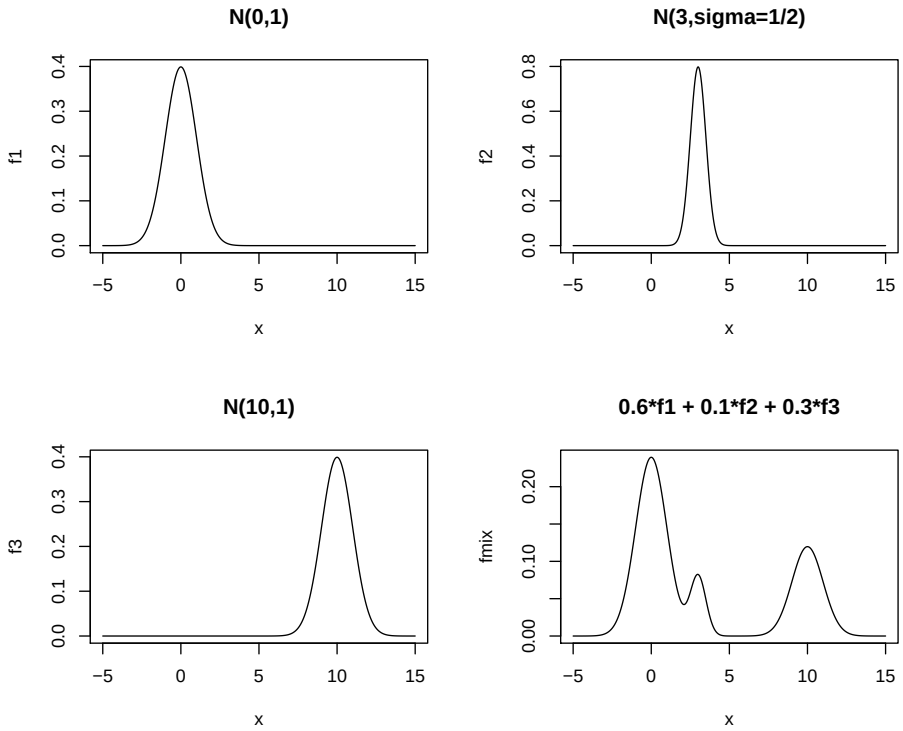


Figura: Mistura de 3 normais.

Gerando amostra de uma mistura

- Queremos gerar uma amostra de tamanho $n = 550$ da mistura de três normais.
- Algoritmo:

```
for(i in 1:550){
    Selecione a pop k = 1, 2 ou 3 com probabs p1, p2, p3
    Y = um valor da normal da pop k
}
```

- Script R:

```
## gerando amostra da mistura (n=550)
## 3 subpops normais, probabs = c(0.6, 0.1, 0.3)
## numero de cada subpop
num <- rmultinom(n=1, size=550, prob=c(0.6, 0.1, 0.3))
num # gerou (321, 56, 173)
amostra <- c(rnorm(num[1]), rnorm(num[2], 3, 1/2), rnorm(num[3]
```

Misturas de v.a.'s discretas

- Os resultados são os mesmos do caso contínuo.
- Suponha que Y seja uma mistura de tres v.a.'s discretas: Y_1, Y_2, Y_3 (por exemplo, 3 Poissons)
- As 3 v.a.'s tem distribuição acumuladas $F_i(y)$ e função de probabilidade $p_i(y) = \mathbb{P}(Y_i = y)$ para $i = 1, 2, 3$
- Então, a distribuição acumulada da mistura Y é dada por

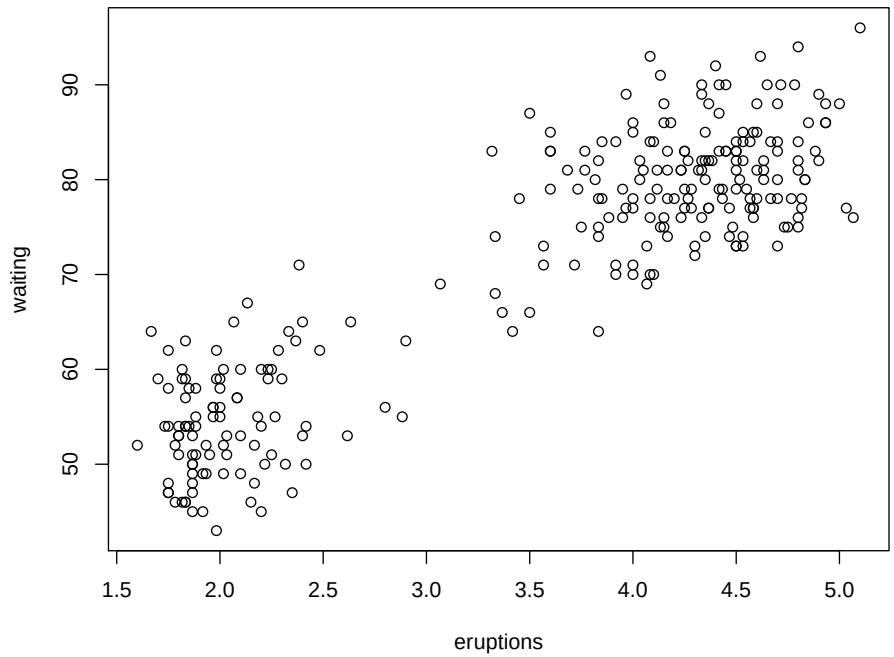
$$\mathbb{F}(y) = \mathbb{P}(Y \leq y) = \mathbb{F}_1(y)\theta_1 + \mathbb{F}_2(y)\theta_2 + \mathbb{F}_3(y)\theta_3$$

- Idêntico ao caso contínuo
- A função de massa de probabilidade é dada por

$$p(y) = \mathbb{P}(Y = y) = p_1(y)\theta_1 + p_2(y)\theta_2 + p_3(y)\theta_3$$

Erupção novamente

- Voltemos aos dados de erupção do geyser Faithful.



- Aparentemente temos duas normais bivariadas misturadas nestes dados.
- Olhando os dados, podemos chutar grosseiramente os valores dos parâmetros de cada componente

Misturas de normais multivariadas

- Componente 1, no canto inferior esquerdo do gráfico:
  - Vetor de valores esperados: $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}) = (2.1, 52)$
  - Matriz de covariância:

$$\Sigma_1 = \begin{bmatrix} \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{12} \\ \rho_1\sigma_{11}\sigma_{12} & \sigma_{12}^2 \end{bmatrix} = \begin{bmatrix} (0.25)^2 & 0.3\sigma_{11}\sigma_{12} \\ 0.3\sigma_{11}\sigma_{12} & 4^2 \end{bmatrix}$$

- Componente 2, no canto superior direito do gráfico:
  - Vetor de valores esperados: $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22}) = (4.5, 80)$
  - Matriz de covariância:

$$\Sigma_2 = \begin{bmatrix} \sigma_{21}^2 & \rho_2\sigma_{21}\sigma_{22} \\ \rho_2\sigma_{21}\sigma_{22} & \sigma_{22}^2 \end{bmatrix} = \begin{bmatrix} (0.35)^2 & 0.7\sigma_{21}\sigma_{22} \\ 0.7\sigma_{21}\sigma_{22} & 5^2 \end{bmatrix}$$

- Proporção do componente 1: 35% ou $\alpha = 0.40$

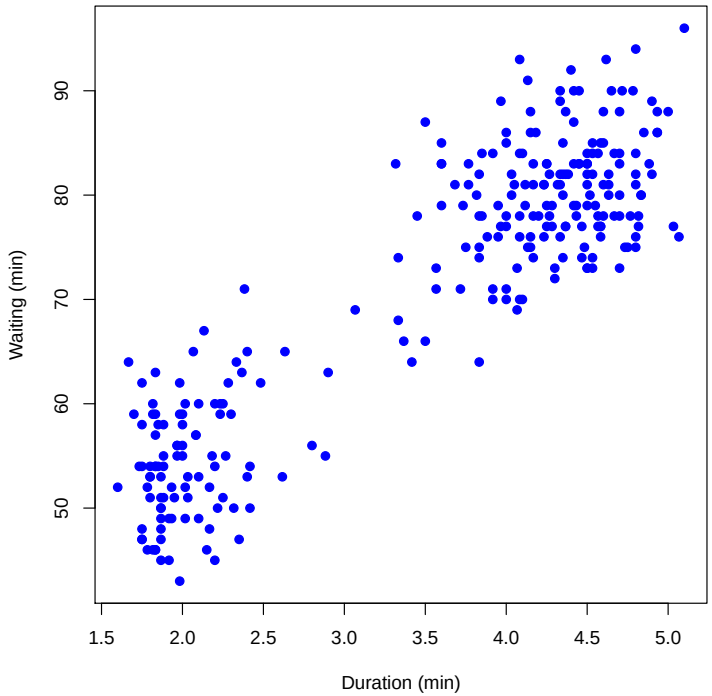
Misturas de normais multivariadas

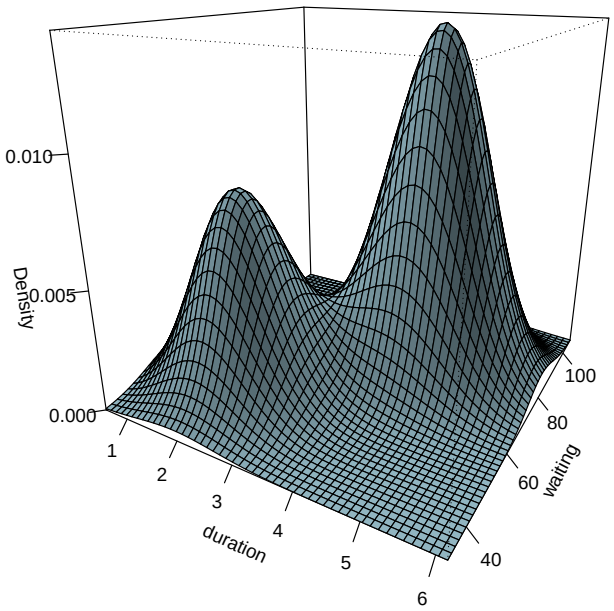
- Densidade conjunta do vetor bivariado $\mathbf{Y} = (Y_1, Y_2)$ é uma mistura de duas densidades gaussianas bivariadas.

$$f(\mathbf{y}) = f(y_1, y_2) = \theta_1 f_1(y_1, y_2) + \theta_2 f_2(y_1, y_2)$$

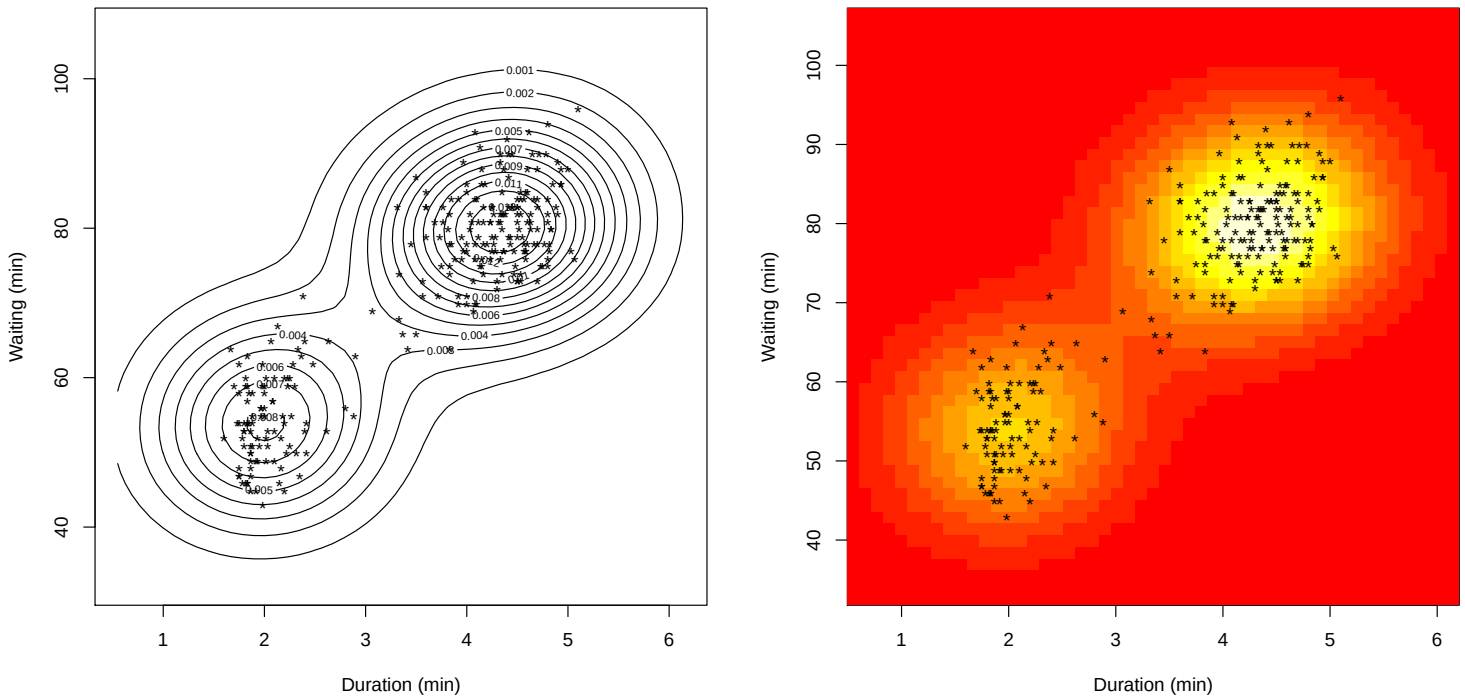
- onde $\theta_1 + \theta_2 = 1$ com $\theta_1 \geq 0$ e $\theta_2 \geq 0$
- e com $f_1(y_1, y_2)$ sendo a densidade do componente 1 (uma normal bivariada) e $f_2(y_1, y_2)$ sendo a densidade do componente 2 (também uma normal bivariada).

Densidade da mistura

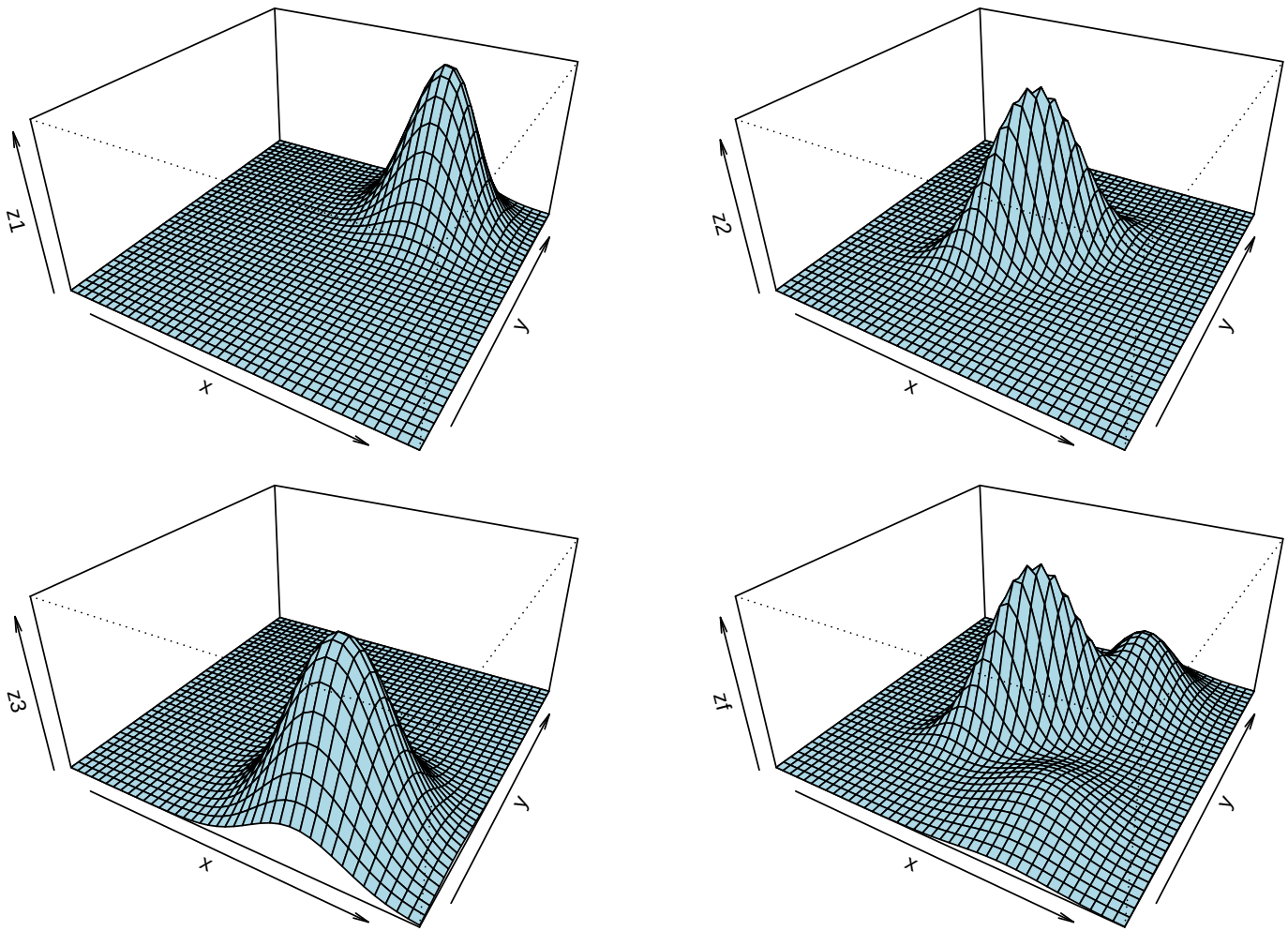




Densidade da mistura



Densidade da mistura de 3 normais bivariadas



Gerando dados de uma mistura

- Input: Número de grupos k
- Input: Densidade de cada grupo: $f_1(\mathbf{y}), f_2(\mathbf{y}), \ldots, f_k(\mathbf{y})$
- Input: proporções de cada grupo: $\theta_1, \theta_2, \ldots, \theta_k$
- Gerar amostra de mistura $\theta_1 f_1(\mathbf{y}) + \ldots \theta_k f_k(\mathbf{y})$ de k componentes: fácil.
- Passo 1: Escolha uma das k componentes ao acaso com probabilidades $\theta_1, \ldots \theta_k$.
- Passo 2: Selecione Y da distribuição $f_i(\mathbf{y})$ da componente i selecionada no passo anterior.
- Isto é, dado o mecanismo (o modelo) aleatório, podemos gerar dados sintéticos.

Ajuste de mistura

- Mas o problema REALMENTE relevante é o contrário.
- Como ajustar um modelo de mistura a dados observados?
- Isto é, recebemos os dados e queremos inferir qual o modelo que foi usado para gerá-los.
- Não é tão simples...
- Vamos supor que o número de componentes K é conhecido.

Ajuste de mistura

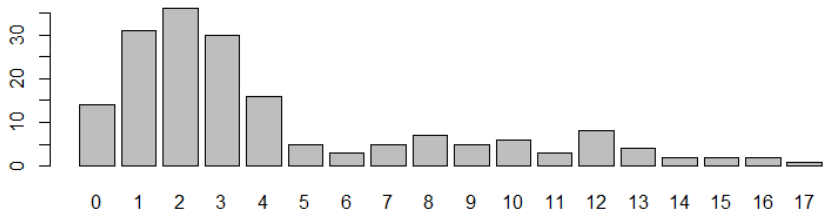
- SE:
  - Sabemos número K de componentes (digamos, 3)
  - Sabemos a CLASSE da distribuição de probab de cada componente (digamos, normal)
- então podemos usar o algoritmo EM para ajustar o modelo.
- A seleção de k é feita via técnicas de escolha de modelos:
- ajustamos vários modelos com diferentes k e escolhemos o "melhor".
- Veremos seleção de modelos mais tarde neste curso...

EMV com fatores latentes

- Nem sempre é fácil obter o EMV: problemas de otimização.
- Um problema difícil é quando temos variáveis latentes ou ocultas (hidden or latent states).
- Exemplos: mixture problems em diversas áreas como imagens, textos, etc...Factor analysis.
- Vamos estudar o algoritmo EM em problemas simples de misturas.

Exemplo

- Estágios iniciais de uma praga agrícola numa floresta industrial.
- Região de cultivo dividida em 180 blocos de 100 árvores cada.
- Contamos as árvores infestadas em cada bloco.
- A cauda estende-se por uma faixa muito longa para vir de uma única Poisson. Talvez uma mistura de duas: uma Poisson($\lambda_a \approx 2$) e uma Poisson($\lambda_b \approx 10$)



Um problema de mistura

- Uma proporção α dos dados vem de uma Poisson com parâmetro λ_a.
- A proporção $1 - \alpha$ restante vem de uma Poisson com parâmetro λ_b.
- Queremos inferir sobre $\boldsymbol{\theta} = (\lambda_a, \lambda_b, \alpha)$.
- Como fazer isto?
- Seria muito fácil SE SOUBÉSSEMOS A QUAL GRUPO CADA OBSERVAÇÃO PERTENCE: bastaria ajustar uma Poisson separadamente a cada um dos dois grupos de dados.
- Infelizmente não sabemos isto: observamos apenas os dados numéricos e não sua classe.
- MAS, como seria no caso em que conhecêssemos os rótulos dos grupos?

Se soubéssemos

- O vetor de dados com a informação completa, da contagem e do rótulo do grupo, pode ser representado por

$$(\mathbf{y}, \mathbf{z}) = (y_1, \ldots, y_{180}, z_1, \ldots, z_{180})$$

- onde y_i é a contagem da árvore i
- z_i é o tipo ddo bloco.
- $z_i = 0$ se o i-ésimo bloco for composto por árvores do tipo resistente e portanto a contagem vem de uma Poisson com parâmetro λ_a.
- $z_i = 1$ se o i-ésimo bloco NÃO for do tipo resistente $\mapsto y_i \sim \text{Poisson}(\lambda_b)$.
- Os dados REALMENTE observados são apenas $\mathbf{y} = (y_1, \ldots, y_{180})$.
- As variáveis em $\mathbf{z} = (z_1, \ldots, z_{180})$ são chamadas de variáveis latentes ou ocultas (hidden, latent)
- O vetor de parâmetros $\boldsymbol{\theta}$ é $\boldsymbol{\theta} = (\lambda_a, \lambda_b, \alpha)$.

O modelo de probabilidade

- (y_i, z_i) é um vetor composto por duas v.a.'s discretas com distribuição conjunta dada por

$$\begin{aligned}
\mathbb{P}(y_i = y, z_i = 0) &= \mathbb{P}(y_i = y | z_i = 0)\mathbb{P}(z_i = 0) = \frac{\lambda_a^y}{y!} e^{-\lambda_a} \cdot (1 - \alpha) \\
\mathbb{P}(y_i = y, z_i = 1) &= \mathbb{P}(y_i = y | z_i = 1)\mathbb{P}(z_i = 1) = \frac{\lambda_b^y}{y!} e^{-\lambda_b} \cdot \alpha
\end{aligned}$$

- Isto é, para $z = 0$ ou $z = 1$, temos

$$\mathbb{P}(y_i = y, z_i = z) = \left[\frac{\lambda_a^y e^{-\lambda_a}}{y!}(1 - \alpha)\right]^{1-z} \left[\frac{\lambda_b^y e^{-\lambda_b}}{y!}\alpha\right]^z \qquad (1)$$

A verossimilhança completa

- Estamos supondo que os blocos são independentes.
- A verossimilhança de $\boldsymbol{\theta} = (\lambda_a, \lambda_b, \alpha)$ baseada nos DADOS COMPLETOS é

$$L^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \prod_{i=1}^{180} \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!}(1-\alpha) \right)^{1-z_i} \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!}\alpha \right)^{z_i}$$

- Tomando log temos a log-verossimilhança

$$\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \sum_{i=1}^{180} (1-z_i) \log \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!}(1-\alpha) \right) + z_i \log \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!}\alpha \right)$$

O EMV com dados completos

- O EMV de $\boldsymbol{\theta} = (\lambda_a, \lambda_b, \alpha)$ no caso da informação completa $(\mathbf{y}, \mathbf{z})$ estar disponível é muito simples (exercício):

$$\begin{aligned}
\hat{\alpha} &= \frac{1}{180}\sum_{i=1}^{180} z_i \\
\hat{\lambda}_a &= \frac{\sum_{i=1}^{180} y_i(1-z_i)}{\sum_{i=1}^{180}(1-z_i)} = \text{média dos blocos com } z_i = 0 \\
\hat{\lambda}_b &= \frac{\sum_{i=1}^{180} y_i z_i}{\sum_{i=1}^{180} z_i} = \text{média dos blocos com } z_i = 1
\end{aligned}$$

- Se pelo menos tivéssemos o vetor completo $(\mathbf{y}, \mathbf{z})\ldots$
- Mas o que temos é apenas o vetor $\mathbf{y}$ das contagens.
- Precisamos da versossmilhança de $\alpha, \lambda_a, \lambda_b$ usando APENAS $\mathbf{y}$.

Verossimilhança marginal de **y**

- Como os blocos são independentes, basta encontrar a distribuição da contagem (y_i) do i-ésimo bloco.

$$\begin{aligned}\mathbb{P}(Y_i = y) &= \mathbb{P}(Y_i = y, Z_i = 0) + \mathbb{P}(Y_i = y, Z_i = 1) \\ &= \alpha \frac{\lambda_a^y e^{-\lambda_a}}{y!} + (1-\alpha)\frac{\lambda_b^y e^{-\lambda_b}}{y!}\end{aligned}$$

- Com isto, obtemos a verossimilhança baseada apenas nos dados realmente observados

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{180} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{180} \left(\frac{\alpha \lambda_a^{y_i} e^{-\lambda_a}}{y_i!} + \frac{(1-\alpha)\lambda_b^{y_i} e^{-\lambda_b}}{y_i!} \right)$$

- Esta função já não é tão simples de ser maximizada (na verdade, neste toy example, ela é muito simples).
- O algoritmo EM vem em nosso socorro (especialmente em problemas mais complicados).

A distribuição de $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$

- A primeira coisa a se fazer é obter a distribuição $\mathbb{P}(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ dos dados faltantes $\mathbf{Z}$ condicionados nos valores $\mathbf{y}$ observados.
- Temos

$$\mathbb{P}(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^{180} \mathbb{P}(z_i|y_i, \boldsymbol{\theta}) = \prod_{i=1}^{180} \frac{\mathbb{P}(y_i, z_i|\boldsymbol{\theta})}{\mathbb{P}(y_i|\boldsymbol{\theta})}$$

- Como não sabemos quem é $\mathbf{z}$, vamos deixá-lo aleatório e tomar o seu valor esperado!!
- Passo 1 do algoritmo EM: calcular o valor ESPERADO da log-verossimilhança baseada nos dados completos DEIXANDO OS DADOS FALTANTES COMO ALEATÓRIOS.

A distribuição de $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$

- Mais precisamente, calculamos a log-verossimilhança ($\log L^c$) de $\boldsymbol{\theta}$ baseada nos dados completos:

$$\begin{aligned}
\ell^c(\boldsymbol{\theta}|\mathbf{y},\mathbf{z}) &= \log L^c(\boldsymbol{\theta}|\mathbf{y},\mathbf{z}) \\
&= \log\left[\prod_{i=1}^{180}\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right)^{1-z_i}\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)^{z_i}\right] \\
&= \sum_{i=1}^{180}\left[(1-z_i)\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right) + z_i\log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)\right]
\end{aligned}$$

- A seguir, substituímos os valores z_i pelas variáveis aleatórias Z_i fazendo com que $l^c(\boldsymbol{\theta}|\mathbf{y},\mathbf{Z})$ seja a variável aleatória:

$$\ell^c(\boldsymbol{\theta}|\mathbf{y},\mathbf{Z}) = \sum_{i=1}^{180}\left[(1-Z_i)\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right) + Z_i\log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)\right]$$

..

- Precisamos agora calcular o valor esperado de $\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})$.
- Observe que, em $\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})$ estamos deixando $\mathbf{y}$ fixado em seus valores observados na amostra.
- A ÚNICA coisa aleatória em $\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})$ é o vetor $\mathbf{Z}$.
- Então, ao calcular a esperança de $\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})$ precisamos lembrar que calculamos uma esperança condicionada a $\mathbf{Y} = \mathbf{y}$.
- Assim,

$$\mathbb{E}\left[\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})|\mathbf{Y} = \mathbf{y}\right] =$$

$$\sum_{i=1}^{180}\left[\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right)\mathbb{E}(1-Z_i|\mathbf{Y}=\mathbf{y}) + \log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y})\right] \tag{2}$$

Outra sutileza...

- O valor de $\boldsymbol{\theta}$ na verossimilhança $\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})$ é um valor $\boldsymbol{\theta}$ arbitrário pertencente ao espaço paramétrico $\boldsymbol{\theta} heta_0$.
- MAs, ao calcular $\mathbb{E}(\mathbf{Z}_i|\mathbf{Y} = \mathbf{y})$ em (2), precisamos usar ALGUM VALOR para o parâmetro $\boldsymbol{\theta}$.
- Vamos usar um valor inicial $\boldsymbol{\theta}^{(0)}$ para o parâmetro.
- Para deixar tudo bastante explícito, vamos usar uma notação um pouco mais carregada reescrevendo (2) como:

$$\mathbb{E}\left[\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)}\right] =$$

$$\sum_{i=1}^{180} \mathbb{E}\left(1 - Z_i|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)}\right) \log\left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!}(1-\alpha)\right)$$

$$+ \mathbb{E}\left(Z_i|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)}\right) \log\left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!}\alpha\right) \quad (3)$$

Assim...

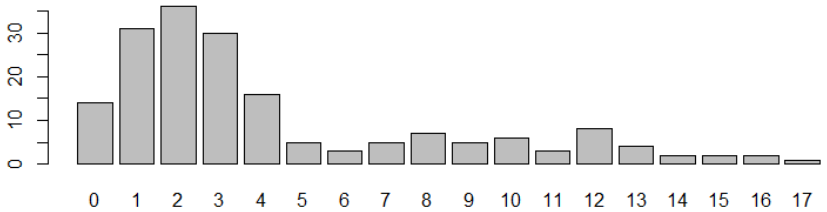
- Queremos calcular

$$\mathbb{E}\left[\ell^c(\boldsymbol{\theta}|\mathbf{y},\mathbf{Z})|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}^{(0)}\right] = \sum_{i=1}^{180}\mathbb{E}\left(1-Z_i|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}^{(0)}\right)\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right) + \mathbb{E}\left(Z_i|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}^{(0)}\right)\log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right) \quad (4)$$

- $\mathbf{Y}$ está fixado no seu valor observado $\mathbf{y}$.
- A esperança de Z_i usa um VALOR INICIAL E FIXO $\boldsymbol{\theta}^{(0)}$ para o parâmetro desconhecido.
- $\boldsymbol{\theta}$ é um valor genérico do parâmetro.
- $\mathbf{Z}$ é o vetor aleatório que torna a função $l^c(\boldsymbol{\theta}|\mathbf{y},\mathbf{Z})$ uma variável aleatória.
- Vamos denotar $\boldsymbol{\theta} = (\lambda_a, \lambda_b, \alpha)$ e $\boldsymbol{\theta}^{(0)} = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha^{(0)})$

Escolhendo $\boldsymbol{\theta}^{(0)}$

- O valor inicial $\boldsymbol{\theta}^{(0)} = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha^{(0)})$ pode ser obtido fazendo uma inspeção grosseira dos dados.
- Por exemplo, considerando o gráfico de barras para as 180 contagens, podemos chutar $\boldsymbol{\theta}^{(0)} = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha^{(0)}) = (2, 10, 0.30)$



$\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}^{(0)})$

- Para calcular $\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}^{(0)})$ lembramos que Z_i depende apenas de Y_i e que Z_i é uma variável aleatória binária. Portanto,

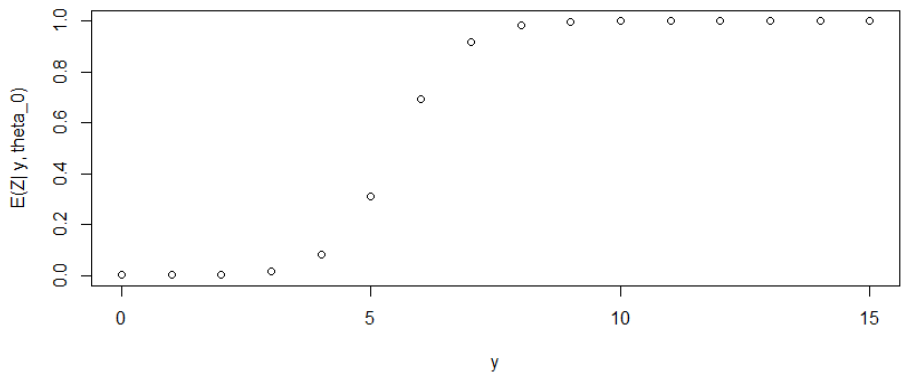
$$\begin{aligned}
\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}^{(0)}) &= \mathbb{P}(Z_i=1|Y_i=y_i,\boldsymbol{\theta}^{(0)}) \\
&= \frac{\mathbb{P}(Z_i=1,Y_i=y_i|\boldsymbol{\theta}^{(0)})}{\mathbb{P}(Z_i=1,Y_i=y_i|\boldsymbol{\theta}^{(0)})+\mathbb{P}(Z_i=0,Y_i=y_i|\boldsymbol{\theta}^{(0)})} \\
&= \frac{\left(\frac{\lambda_b^{(0)y_i}}{y_i!}e^{-\lambda_b^{(0)}}\right)\cdot\alpha_0}{\frac{\lambda_b^{(0)y_i}}{y_i!}e^{-\lambda_b^{(0)}}\cdot\alpha_0+\frac{\lambda_a^{(0)y_i}}{y_i!}e^{-\lambda_a^{(0)}}\cdot(1-\alpha_0)} \\
&= \frac{\lambda_b^{(0)y_i}e^{-\lambda_b^{(0)}}\alpha_0}{\lambda_b^{(0)y_i}e^{-\lambda_b^{(0)}}\alpha_0+\lambda_a^{(0)y_i}e^{-\lambda_a^{(0)}}(1-\alpha_0)}
\end{aligned}$$

- onde $\boldsymbol{\theta}^{(0)}=(\lambda_a^{(0)},\lambda_b^{(0)},\alpha_0)$.

$\mathbb{E}(Z_i|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)})$

- Por exemplo, considerando $\boldsymbol{\theta}^{(0)} = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha^{(0)}) = (2, 10, 0.30)$,
- temos

$$\begin{aligned}
\mathbb{E}(Z_i|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)}) &= \mathbb{P}(Z_i = 1|Y_i = y_i, \boldsymbol{\theta}^{(0)}) \\
&= \frac{\lambda_b^{(0)^{y_i}} e^{-\lambda_b^{(0)}} \alpha_0}{\lambda_b^{(0)^{y_i}} e^{-\lambda_b^{(0)}} \alpha_0 + \lambda_a^{(0)^{y_i}} e^{-\lambda_a^{(0)}} (1 - \alpha_0)} \\
&= \frac{10^{y_i} e^{-10} * 0.30}{10^{y_i} e^{-10} 0.30 + 2^{y_i} e^{-2} (1 - 0.30)}
\end{aligned}$$



$\mathbb{E}(Z_i|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)})$

- Tendo o valor de $\mathbb{E}(Z_i|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)})$ podemos então calcular $\mathbb{E}\left[l^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)}\right]$
- Este último valor pode ser calculado em pontos arbitrários $\boldsymbol{\theta}$ se $\boldsymbol{\theta}^{(0)}$ é fixado.
- Vamos denotar:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y}) = \mathbb{E}\left[l^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z})|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)}\right] \quad (5)$$

- Esta expressão é crucial no algoritmo EM.
- Lembre-se: $\boldsymbol{\theta}^{(0)}$ é um chute inicial e fixo para o parâmetro, $\boldsymbol{\theta}$ é um valor arbitrário para o parâmetro e os Z's são os rótulos dos grupos das observações.
- $\boldsymbol{\theta}^{(0)}$ será atualizado ao longo das iterações, como explicaremos em breve.

$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$

- Os dados $\mathbf{y}$ estarão fixos ao longo dasiterações do algoritmo EM.
- É importante perceber que $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$ é função de DOIS valores para o parâmetro $\boldsymbol{\theta}$ e $\boldsymbol{\theta}^{(0)}$.
- Por exemplo, suponha que $\boldsymbol{\theta} = (\lambda_a, \lambda_b, \alpha) = (1.2, 8.5, 0.30)$
- e que $\boldsymbol{\theta}_0 = (\lambda_a^{(0)}, \lambda_b^{(0)}, \alpha^{(0)}) = (1.8, 10, 0.45)$.
- O vetor $\mathbf{y}$ com 180 posições contém as contagens
- Os seguintes comandos no R calculam o valor de $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$:

```
theta <- c(1.2, 8.5, 0.30)
theta0 <- c(1.8, 10, 0.45)
Ez <- 1/(1+(theta0[1]/theta0[2]))*exp(-(theta0[1]-theta0[2])*((1-theta0[3])/theta0[3]))
Q <- sum(log(dpois(y,theta[1])*(1-theta[3])) * (1-Ez) + log(dpois(y,theta[2]) *theta[3]) * Ez)
```

M-step

- O primeiro passo é chamado E-step: trata-se de obter a expressão $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$ onde $\boldsymbol{\theta}^{(0)}$ é um valor inicial usado para calcular $E(Z|Y = y)$ e $\boldsymbol{\theta}$ é um valor arbitrário $\boldsymbol{\theta}$.
- O segundo passo do algoritmo EM é chamado M-step.
- Lembre-se: $\boldsymbol{\theta}^{(0)}$ é um valor inicial fixado pelo usuário.
- No passo M, encontramos o valor de $\boldsymbol{\theta}$ que maximiza $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$
- Isto é, encontramos o valor $\boldsymbol{\theta}_1$ do argumento $\boldsymbol{\theta}$ que maximiza $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$ para $\boldsymbol{\theta}^{(0)}$ fixo:

$$\boldsymbol{\theta}_1 = \arg_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$$

Solução exata

- No caso de mistura de Poissons, esta maximização é muito simples.
- Vamos escrever $\mathbb{E}(Z_i|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}^{(0)})$ simplesmente como $\mathbb{E}(Z_i)$.
- Então

$$\begin{aligned}
\hat{\alpha} &= \frac{1}{180}\sum_{i=1}^{180}\mathbb{E}(Z_i) \\
\hat{\lambda}_a &= \frac{\sum_{i=1}^{180} y_i(1-\mathbb{E}(Z_i))}{\sum_{i=1}^{180}(1-\mathbb{E}(Z_i))} \\
\hat{\lambda}_b &= \frac{\sum_{i=1}^{180} y_i\mathbb{E}(Z_i)}{\sum_{i=1}^{180}\mathbb{E}(Z_i)}
\end{aligned}$$

- Esta expressão é quase idêntica ao caso de dados completos (compare as expressões nos dois casos)

Resumo

- Começamos com um valor de $\boldsymbol{\theta}^{(0)}$ inicial para o parâmetro $\boldsymbol{\theta}$.
- Calculamos $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$ como uma função de $\boldsymbol{\theta}$ (com $\boldsymbol{\theta}^{(0)}$ fixo).
- A seguir, maximizamos $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$ com respeito a $\boldsymbol{\theta}$ obtendo $\boldsymbol{\theta}_1$.
- O processo é iterado:
  - calculamos $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_1, \mathbf{y})$ (passo E)
  - A seguir, maximizamos em $\boldsymbol{\theta}$ para obter $\boldsymbol{\theta}_2$ (passo M)
- Grande vantagem: Terminamos também com estimativa de $\mathbb{P}(Z_i = 1)$, a probabilidade de cada observação pertencer ao grupo 1.
- Este processo iterativo converge para o EMV de $\boldsymbol{\theta}$. Convergência pode ser lenta.
- O que muda de problema para problema é a expressão de $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y})$.

Exemplo

- Terminar EM para o caso Poisson no R
- Caso normal multivariado: ver wikipedia.
- Mas por quê o algoritmo EM funciona? Existe uma prova de que o EM converge para um máximo local (ou global) da log-verossimilhança, como veremos a seguir.

Função convexa

- Função $g(x)$ é uma função convexa se a curva está sempre abaixo da secante.
- Ou então: se a reta tangente em cada ponto está abaixo da curva.
- Ou então se a derivada $g'(x)$ é crescente.
- Ou então se a derivada segunda é positiva (ou melhor, não-negativa).
- Exemplo clássico: $g(x) = x^2$.
- Para quê tantas caracterizações? Generalizar para funções de várias variáveis.
- $g(\mathbf{x})$ é convexa se a MATRIZ de derivadas segundas $D^2 g$ é definida positiva: $\mathbf{x}^t \; D^2 g \; \mathbf{x} > 0$ para todo ponto $\mathbf{x}$.

Desigualdade de Jensen

- Desigualdade fundamental em probabilidade: Jensen
- Seja X uma v.a. qualquer com $E(X) = \mu$
- Seja $g(x)$ uma função convexa.
- Crie uma nova v.a. $Y = g(X)$.
- Então $E(Y) = E(g(X)) \geq g(\mu) = g(E(X))$
- Exemplo: $E(g(X)) = E(X^2) \geq g(E(X)) = [E(X)]^2 = \mu^2$
- Função g é côncava se $-g$ é convexa. No caso côncavo, desigualdade é invertida.
- Função LOG é côncava: $E(\log(X)) \leq \log[E(X)]$

Notação

- Seja $(\mathbf{y}, \mathbf{z})$ o vetor de dados completos com densidade $f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$.
- Vamos também denotar $\ell^c(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) = \log f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$.
- Seja $f(\mathbf{y}|\boldsymbol{\theta}) = \int_{\mathcal{Z}} f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z}$ a densidade marginal de $\mathbf{Y}$.
- Esta é também a log-verossimilhança de $\boldsymbol{\theta}$ baseada apenas nos dados observados $\mathbf{y}$.
- Isto é, $\ell(\boldsymbol{\theta}|\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta})$.
- Seja

$$k(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}) = \frac{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta})}$$

  a densidade condicional de $\mathbf{Z}$ dados as observações $\mathbf{y}$.
- Vamos usar a letra k para denotar esta densidade condicional e assim evitar mais usos da letra f para densidades.

A verossimilhança

- Suponha que temos um valor inicial $\boldsymbol{\theta}^{(0)}$ para o parâmetro $\boldsymbol{\theta}$.
- Como $k(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(0)})$ é uma densidade de probabilidade, sua integral sobre $\mathcal{Z}$, os valores possíveis de $\mathbf{z}$, é igual a 1:

$$1 = \int_{\mathcal{Z}} k(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(0)}) d\mathbf{z}$$

- Assim, vamos multiplicar a verossimilhança com os dados observados por 1:

$$\begin{aligned} \ell(\boldsymbol{\theta}|\mathbf{y}) &= \log f(\mathbf{y}|\boldsymbol{\theta}) \\ &= \log f(\mathbf{y}|\boldsymbol{\theta}) \int_{\mathcal{Z}} k(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(0)}) d\mathbf{z} \end{aligned}$$

- Como $\log f(\mathbf{y}|\boldsymbol{\theta})$ náo depende de $\mathbf{z}$, podemos passá-la para dentro da integral.

..

- Temos então

$$\begin{aligned}
\ell(\boldsymbol{\theta}|\mathbf{y}) &= \log f(\mathbf{y}|\boldsymbol{\theta}) \\
&= \int_{\mathcal{Z}} \log f(\mathbf{y}|\boldsymbol{\theta})\ k(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(0)}) d\mathbf{z} \\
&= \int_{\mathcal{Z}} \log \left[\frac{f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})}{k(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(0)})} \right] k(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(0)}) d\mathbf{z}
\end{aligned}$$

- Note como no primeiro termo temos o vetor genérico $\boldsymbol{\theta}$ na função k no denominador mas temos $\boldsymbol{\theta}^{(0)}$ na segunda aparição da função k na integral acima.

Continuando de onde paramos:

- $$\begin{aligned}
\ell(\boldsymbol{\theta}|\mathbf{y}) &= \int_{\mathcal{Z}} \log\left[\frac{f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})}{k(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(0)})}\right] k(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(0)})d\mathbf{z} \\
&= \int_{\mathcal{Z}} \left[\log f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta}) - \log k(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(0)})\right] k(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(0)})d\mathbf{z} \\
&= \int_{\mathcal{Z}} \log f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})\; k(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(0)})d\mathbf{z} - \int_{\mathcal{Z}} \log k(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})\; k(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(0)})d\mathbf{z} \\
&= \mathbb{E}_{\boldsymbol{\theta}^{(0)}}\left[\log f(\mathbf{y},\mathbf{Z}|\boldsymbol{\theta}) \;\middle|\; \mathbf{y},\boldsymbol{\theta}^{(0)}\right] - \mathbb{E}_{\boldsymbol{\theta}^{(0)}}\left[\log k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}) \;\middle|\; \mathbf{y},\boldsymbol{\theta}^{(0)}\right]
\end{aligned}$$
- Na última linha, usamos um sub-índice para indicar o valor do parâmetro usado no cálculo da esperança.
- Esperança de quê? O que é aleatório aqui? Colocamos $\mathbf{Z}$ em maiúscula para indicar que este vetor $\mathbf{Z}$ é aleatório, enquanto $\mathbf{y}$ permanece em minúscula já que as observações estão fixas nos seus valores observados na amostra.
- As duas esperanças são tomadas com respeito à densidade condicional $k(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(0)})$.

Repetindo...

- Encontramos que

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \mathbb{E}_{\boldsymbol{\theta}^{(0)}}\left[\log f(\mathbf{y},\mathbf{Z}|\boldsymbol{\theta}) \;\middle|\; \mathbf{y},\boldsymbol{\theta}^{(0)}\right] - \mathbb{E}_{\boldsymbol{\theta}^{(0)}}\left[\log k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}) \;\middle|\; \mathbf{y},\boldsymbol{\theta}^{(0)}\right]$$

- Vamos definir a notação

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)},\mathbf{y}) = \mathbb{E}_{\boldsymbol{\theta}^{(0)}}\left[\log f(\mathbf{y},\mathbf{Z}|\boldsymbol{\theta}) \;\middle|\; \mathbf{y},\boldsymbol{\theta}^{(0)}\right] = \int_{\mathcal{Z}} \log f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})\; k(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(0)})d\mathbf{z}$$

- Curiosamente, para encontramos o máximo em $\boldsymbol{\theta}$ da verossimilhança $\ell(\boldsymbol{\theta}|\mathbf{y})$, nós vamos maximizar apenas o primeiro termo $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)},\mathbf{y})$. Esta maximização será o passo M do algoritmo.
- Já veremos que este procedimento leva realmente a um ponto de máximo de $\ell(\boldsymbol{\theta}|\mathbf{y})$.

Algoritmo EM

- Seja $\boldsymbol{\theta}^{(m)}$ uma estimativa do vetor de parâmetros no passo m com $\boldsymbol{\theta}^{(0)}$ sendo uma estimativa inicial.
- **Passo E:** no passo *Expectation*, calcule

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}, \mathbf{y}) = \mathbb{E}_{\boldsymbol{\theta}^{(m)}} \left[\log f(\mathbf{y}, \mathbf{Z}|\boldsymbol{\theta}) \;\; | \;\; \mathbf{y}, \boldsymbol{\theta}^{(0)} \right]$$

  onde a esperança é calculada com respeito à densidade condicional $k(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(0)})$.
- **Passo M:** no passo *Maximization*, obtenha

$$\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}, \mathbf{y})$$

Propriedades

- Sob suposições, pode ser mostrado que $\boldsymbol{\theta}^{(m)}$ converge em probabilidade para o MLE quando $m \rightarrow \infty$.
- Nós não vamos mostrar este resultado mas outro, mais facilmente demonstrável.
- Vamos mostrar que a log-verossimilhança $\ell(\boldsymbol{\theta}^{(m)}|\mathbf{y})$ aumenta a medida que m cresce.
- Isto é, vamos provar que $\ell(\boldsymbol{\theta}^{(m+1)}|\mathbf{y}) \geq \ell(\boldsymbol{\theta}^{(m)}|\mathbf{y})$.

Teorema

- A sequência $\boldsymbol{\theta}^{(m)}$ definida pelo algoritmo EM satisfaz $\ell(\boldsymbol{\theta}^{(m+1)}|\mathbf{y}) \geq \ell(\boldsymbol{\theta}^{(m)}|\mathbf{y})$.
- **Prova:** Como $\boldsymbol{\theta}^{(m+1)}$ maximiza $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}, \mathbf{y})$, nós temos

$$Q(\boldsymbol{\theta}^{(m+1)}|\boldsymbol{\theta}^{(m)}, \mathbf{y}) \geq Q(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)}, \mathbf{y}) \,.$$

- Isto é,

$$\mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log f(\mathbf{y}, \mathbf{Z}|\boldsymbol{\theta}^{(m+1)}) \;\middle|\; \mathbf{y}, \boldsymbol{\theta}^{(m)}\right] \geq \mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log f(\mathbf{y}, \mathbf{Z}|\boldsymbol{\theta}^{(m)}) \;\middle|\; \mathbf{y}, \boldsymbol{\theta}^{(m)}\right]$$

- Voltando à decomposição de $\ell(\boldsymbol{\theta}|\mathbf{y})$, se provarmos que o segundo termo satisfaz

$$\mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log k(\mathbf{Z}|\mathbf{y}, \boldsymbol{\theta}^{(m+1)}) \;\middle|\; \mathbf{y}, \boldsymbol{\theta}^{(m)}\right] \leq \mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log k(\mathbf{Z}|\mathbf{y}, \boldsymbol{\theta}^{(m)}) \;\middle|\; \mathbf{y}, \boldsymbol{\theta}^{(m)}\right]$$

teremos provado o teorema.

Continuação da prova do teorema...

- Mas provar que

$$\mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}^{(m+1)}) \mid \mathbf{y},\boldsymbol{\theta}^{(m)}\right] \leq \mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}^{(m)}) \mid \mathbf{y},\boldsymbol{\theta}^{(m)}\right]$$

- é equivalente a provar que

$$\mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}^{(m+1)}) \mid \mathbf{y},\boldsymbol{\theta}^{(m)}\right] - \mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}^{(m)}) \mid \mathbf{y},\boldsymbol{\theta}^{(m)}\right] \leq 0$$

- ou ainda, que

$$\mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}^{(m+1)}) - \log k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}^{(m)}) \mid \mathbf{y},\boldsymbol{\theta}^{(m)}\right] \leq 0$$

- Usando propriedade básica dos logaritmos, temos de provar que:

$$\mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log \frac{k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}^{(m+1)})}{k(\mathbf{Z}|\mathbf{y},\boldsymbol{\theta}^{(m)})} \mid \mathbf{y},\boldsymbol{\theta}^{(m)}\right] \leq 0$$

Continuação da prova do teorema...

- Vamos aplicar a desigualdade de Jensen com a função log: temos $\mathbb{E}[\log(X)] \leq \log(\mathbb{E}[X])$ para qq v.a. X.Assim,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left[\log\left(\frac{k(\mathbf{Z}|\boldsymbol{\theta}^{(m+1)},\mathbf{y})}{k(\mathbf{Z}|\boldsymbol{\theta}^{(m)},\mathbf{y})}\right)\right] &\leq \log \mathbb{E}_{\boldsymbol{\theta}^{(m)}}\left(\frac{k(\mathbf{Z}|\boldsymbol{\theta}^{(m+1)},\mathbf{y})}{k(\mathbf{Z}|\boldsymbol{\theta}^{(m)},\mathbf{y})}\right) \\
&= \log \int_{\mathcal{Z}} \frac{k(\mathbf{z}|\boldsymbol{\theta}^{(m+1)},\mathbf{y})}{k(\mathbf{z}|\boldsymbol{\theta}^{(m)},\mathbf{y})}\, k(\mathbf{z}|\boldsymbol{\theta}^{(m)},\mathbf{y})\, d\mathbf{z} \\
&= \log \int_{\mathcal{Z}} k(\mathbf{z}|\boldsymbol{\theta}^{(m+1)},\mathbf{y})\, d\mathbf{z} \\
\text{pois } k \text{ é densidade} &= \log(1) = 0
\end{aligned}$$

- Isto conclui a demonstração do teorema.

De volta à mistura de distribuiçãoes

- Vamo derivar agora um caso geral de misturas sem especificar a distribuição e deixando o número de classes ser maior que 2.
- Suponha que temos dados i.i.d $\mathbf{y} = (y_1, \ldots, y_n)$ que vêm de uma distribuição que é uma mistura de k classes ou populações básicas.
- As classes possuem distribuições $f_1(|\phi_1), \ldots, f_k(|\phi_k)$.
- Usualmente, todas são membros de uma mesma classes tais como todas eleas serem gaussianas com diferentes parâmetros.
- Entretanto, isto não é uma imposição do modelo de mistura e o algoritmo EM funcionaria se tivéssemos, digamos, gaussianas misturadas com gamas.

De volta à mistura de distribuiçãoes

- Cada observação y_i vem independentemente da classe-distribuição j com probabilidade α_j.
- Claramente, temos $\alpha_j \geq 0$ e com $\alpha_1 + \ldots + \alpha_k = 1$.
- Seja $\boldsymbol{\theta} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_k, \alpha_1, \ldots, \alpha_k)$ o parâmetro sobre o qual queremos fazer inferência.
- Vamos agora definir as variáveis latentes (ocultas, não-observadas). Elas vão determinar de qual população veio cada uma das observações.
- Seja Z_i uma variável discreta com valores possíveis $1, 2, \ldots, k$ e probabilidades associadas $(\alpha_1, \ldots, \alpha_k)$.
- Isto é, Z_i é um ensaio multinomial $\mathcal{M}(k; \boldsymbol{\alpha})$ com $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$.
- Assumimos que as variáveis Z_i são independentes.

De volta à mistura de distribuiçãoes

- Assim, $Z_i = j$ se a observação y_i veio da população j com densidade $f_j(y; \boldsymbol{\theta}_j)$.
- Vamos definir a variável indicadora desse evento: $I[Z_i = j]$
- Como cada observação vem de uma única população, temos $1 = I[Z_i = 1] + I[Z_i = 2] + \ldots + I[Z_i = k]$ para todo $i = 1, \ldots, n$.
- Vamos então usar o algorimto EM.
- Precisamos da verossmilhança dos dados completos $(\mathbf{y}, \mathbf{z})$.

Verossmilhança dos dados completos

- A densidade conjunta dos dados, caso os rótulos z_i fossem conhecidos, é igual a

$$
\begin{aligned}
f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) &= f(\mathbf{z}|\boldsymbol{\theta})\ f(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) \\
&= \prod_i \left(\prod_j \alpha_j^{I[z_i=j]} \right) \prod_i f(y_i|z_i, \boldsymbol{\theta}) \\
&= \prod_i \left[\prod_j \alpha_j^{I[z_i=j]} f(y_i|\boldsymbol{\phi}_j)^{I[z_i=j]} \right] \\
&= \prod_i \left[\prod_j \left(\alpha_j f(y_i|\boldsymbol{\phi}_j) \right)^{I[z_i=j]} \right] \\
&= \prod_{i;z_i=1} [\alpha_1 f(y_i|\boldsymbol{\phi}_1)] \ldots \prod_{i;z_i=k} [\alpha_k f(y_i|\boldsymbol{\phi}_k)]
\end{aligned}
$$

Log-Verossmilhança dos dados completos

- Para evitar uma notação muito carregada vamos escrever $f(y_i|\phi_1) = f_j(y_i)$.
- Assim, a log-verossimilhança é

$$\begin{aligned}
\ell^c(\boldsymbol{\theta}|\mathbf{y},\mathbf{z}) &= \log f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta}) \\
&= \log \prod_i \left[\prod_j (\alpha_j f_j(y_i))^{I[z_i=j]} \right] \\
&= \sum_i \sum_j [I[z_i = j] \log(\alpha_j f_j(y_i))]
\end{aligned}$$

- Para o algoritmo EM, precisamos primeiro substituir os z_i na expressão acima pelas v.a.'s Z_i.
- Em seguida, tomamos a esperança da expressão resultante com respeito á densidade condicional de de $\mathbf{Z}$ dadas as observações $\mathbf{y}$ e um valor inicial $\boldsymbol{\theta}^{(0)}$ para o parâmetro.

Esperança da Log-Verossmilhança dos dados completos

- Dada a expressão linear da log-verossimilhança, temos

$$\mathbb{E}\left[\ell^c(\boldsymbol{\theta}|\mathbf{y},\mathbf{Z})\right] = \sum_i \sum_j \left[\log(\alpha_j f_j(y_i))\ \mathbb{E}(I[Z_i = j])\right]$$

- Note que trocamos $\mathbf{z}$ por $\mathbf{Z}$ para indicar agora que eles representam variáveis aleatórias.
- Note também que $\mathbb{E}(I[Z_i = j]) = \mathbb{E}_{\boldsymbol{\theta}^{(0)}}(I[Z_i = j] \mid \mathbf{y})$ será calculada assumindo que o parâmetro é igual a $\boldsymbol{\theta}^{(0)}$ e também assumindo que os dados $\mathbf{y}$ são conhecidos.
- Como $I[z_i = j]$ é uma variável aleatória binária, $\mathbb{E}(I[Z_i = j]) = \mathbb{P}(Z_i = j \mid y_i, \boldsymbol{\theta}^{(0)})$

$\gamma_{ij} = \mathbb{P}(Z_i = j \mid y_i)$

- Queremos
$$\gamma_{ij} = \mathbb{P}(Z_i = j \mid y_i, \boldsymbol{\theta}^{(0)}) .$$
- O que é esta probabilidade γ_{ij} ?
- Supondo que:
  - conhecemos as distribuições de cada população (isto é, conhecemos $\phi_j^{(0)}$
  - conhecemos as frequências $\alpha_j^{(0)}$ com que cada população aparece
  - e conhecendo o valor y_i que apareceu
- queremos então obter as probabilidades de que esta observação y_i:
  - tenha vindo da população 1: $\gamma_{i1} = \mathbb{P}(Z_i = 1 \mid y_i, \boldsymbol{\theta}^{(0)})$
  - ou tenha vindo da população 2: $\gamma_{i2} = \mathbb{P}(Z_i = 2 \mid y_i, \boldsymbol{\theta}^{(0)})$
  - ...
  - ou que tenha vindo da população k: $\gamma_{ik} = \mathbb{P}(Z_i = k \mid y_i, \boldsymbol{\theta}^{(0)})$.

$\gamma_{ij} = \mathbb{P}(Z_i = j \mid y_i)$

- Para obter estas probabilidades, usamos a definição de probabilidade condicional:

$$\begin{aligned}
\gamma_{ij} &= \mathbb{P}(Z_i = j \mid y_i, \boldsymbol{\theta}^{(0)}) \\
&= \frac{f(Y_i = y_i, I[Z_i = j] | \boldsymbol{\theta}^{(0)})}{f(Y_i = y_i | \boldsymbol{\theta}^{(0)})} \\
&= \frac{f(Y_i = y_i, I[Z_i = j] | \boldsymbol{\theta}^{(0)})}{\sum_g f(Y_i = y_i, I[Z_i = g] | \boldsymbol{\theta}^{(0)})} \\
&= \frac{\alpha_j f_j(y_i | \boldsymbol{\theta}^{(0)})}{\sum_g \alpha_g f_g(y_i | \boldsymbol{\theta}^{(0)})} \\
&= \frac{\alpha_j f_j(y_i | \boldsymbol{\phi}_j^{(0)})}{\sum_g \alpha_g f_g(y_i | \boldsymbol{\phi}_g^{(0)})}
\end{aligned}$$

- Na expressão acima estamos manipulando a distribuição conjunta de uma v.a. discreta (Z_i) e uma v.a. que pode ser contínua (Y_i).
- Não estudamos este tipo de manipulação no curso mas ele é válido.

$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}, \mathbf{y})$

- Seja $\boldsymbol{\theta}^{(m)}$ uma estimativa do vetor de parâmetros
- De posse da expressão $\gamma_{ij} = \mathbb{P}(Z_i = j \mid y_i, \boldsymbol{\theta}^{(m)})$ podemos seguir com o algoritmo EM.
- Estes valores γ_{ij} são simples números reais se tivermos y_i e $\boldsymbol{\theta}^{(m)}$.
- Podemos tratá-los como constantes no próximo passo.
- No algoritmo EM, vamos precisar de

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}, \mathbf{y}) &= \sum_i \sum_j [\mathbb{P}(Z_i = j|y_i) \ \log(\alpha_j f_j(y_i))] \\ &= \sum_i \sum_j [\gamma_{ij} \ \log(\alpha_j f_j(y_i))] \end{aligned}$$

Algoritmo EM

- Seja $\boldsymbol{\theta}^{(m)}$ uma estimativa do vetor de parâmetros. Obtenha

$$\gamma_{ij} = \frac{\alpha_j f_j(y_i|\boldsymbol{\phi}_j^{(m)})}{\sum_g \alpha_g f_g(y_i|\boldsymbol{\phi}_g^{(m)})}$$

- **Passo E:** calcule

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}, \mathbf{y}) = \sum_i \sum_j \left[\gamma_{ij} \log(\alpha_j f_j(y_i|\boldsymbol{\phi}_j))\right]$$

- **Passo M:** maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}, \mathbf{y})$ em $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}, \mathbf{y})$$

- Este último passo vai variar de problema para problema dependendo das densidades $f_j(y_i|\boldsymbol{\phi}_j))$ envolvidas.
- Em vários casos, como em que as f's são gaussianas, a maximização pode ser exata.