ME705 - Prova 2 - 28/06/2022

Profa. Mariana Rodrigues Motta, DE - UNICAMP
Coloque nome e RA.

Considere as contagens anuais de grandes terremotos (ou seja, magnitude 7 e acima) para os anos 1900-2006 trabalhadas nas Listas 5, 6 e 7. A variável *counts* contém a contagem de $0, 1, 2, 3, \ldots, 41$ terremotos ao longo dos anos em estudo.

```
counts
 [1]  0  0  0  0  0  0  1  1  4  0  3  4  2  6  5 10  8  3  8  4  4  7  8
[24]  4  3  2  4  4  1  1  1  1  2  0  1  1  2  0  0  1  0  1
```

Considerando a função *kmeans* do pacote *mixtools* do R para **duas** classes, temos como saída

```
> kmeans(counts,2)
K-means clustering with 2 clusters of sizes 14, 28

Cluster means:
       [,1]
1 5.7142857
2 0.9642857

Clustering vector:
 [1] 2 2 2 2 2 2 2 2 1 2 2 1 2 1 1 1 1 2 1 1 1 1 1 1 2 2 1 1 2 2 2 2 2 2 2
[36] 2 2 2 2 2 2 2
```

onde *Cluster means* indica a média das observações dentro dos clusters 1 e 2, respectivamente. Lembre que a distribuição de Poisson com parâmetro λ é dada por

$$P(Y = k) = e^{-\lambda}\lambda^k/k!$$

1. (1,0 ponto) Escreva o modelo marginal de mistura de Poisson considerando duas classes usando p como probabilidade de mistura e os valores do parâmetro λ da distribuição de Poisson a partir dos resultados do *kmeans* com 2 classes.

2. (1,0 ponto) Seja $y_1, \ldots, y_n$ o número de terremotos (*counts*) observados, $n = 42$. Escreva a função de verossimilhança para modelo marginal de mistura de Poisson considerando duas classes e λ_1 e λ_2 a partir do resultado do *kmeans* com 2 classes.

3. (1,0 ponto) Considere o argumento de 'aumentar' os dados através das variáveis latentes $z_1, \ldots, z_n$, iid, tal que $z_i = 1$ se y_i pertence à classe 1 e zero caso contrário, onde $P(Z_i = 1) = p$ e $P(Z_i = 0) = 1-p$, p é a probabilidade de mistura. Escreva a função de verossimilhança com dados aumentados em $z_1, \ldots, z_n$ para a mistura de duas distribuições de Poisson considerando λ_1 e λ_2 a partir do resultado do *kmeans* com 2 classes.

4. Considerando a função *kmeans* do pacote *mixtools* do R para **três** classes, temos como saída

```
> kmeans(counts,3)
K-means clustering with 3 clusters of sizes 11, 6, 25

Cluster means:
      [,1]
1 3.818182
2 7.833333
3 0.720000

Clustering vector:
 [1] 3 3 3 3 3 3 3 3 1 3 1 1 3 2 1 2 2 1 2 1 1 2 2 1 1 3 1 1 3 3 3 3 3 3 3
[36] 3 3 3 3 3 3 3
```

onde *Cluster means* indica a média das observações dentro dos clusters 1, 2 e 3, respectivamente.

(a) (1,0 ponto) Escreva o modelo marginal de mistura de Poisson considerando três classes, com probabilidade de mistura p_1, p_2 e $p_3 = 1 - p_1 - p_2$, e valores do parâmetro λ da distribuição de Poisson de cada classe usando os resultados do *kmeans* para 3 classes.

(b) (1,0 ponto) Considere o argumento de 'aumentar' os dados através das variáveis latentes $z_1, \ldots, z_n$, iid, tal que $z_i = 1$ se y_i pertence à classe 1, $z_i = 2$ se y_i pertence à classe 2 e $z_i = 3$ se y_i pertence à classe 3, onde $P(Z_i = 1) = p_1$, $P(Z_i = 2) = p_2$ e $P(Z_i = 3) = 1 - p_1 - p_2$. Note que $z_i \sim Multinomial(3, p_1, p_2, p_3)$. Escreva a função de verossimilhança com dados aumentados em $z_1, \ldots, z_n$ para a mistura de três distribuições de Poisson com valores do parâmetro λ de cada classe usando os resultados do *kmeans* para 3 classes.

5. (1,0 ponto) Considere um modelo de mistura de Poisson com duas classes com dados aumentados em $z_1, \ldots, z_n$, iid. A seguir, escreva a distribuição condicional completa a posteriori de z_i.

6. (1,0 ponto) Escreva o algoritmo do amostrador de Gibbs para o modelo de mistura de Poisson com duas classes com dados aumentados em $z_1, \ldots, z_n$, iid. Para isso, use a notação $p(\lambda_1|else, \mathbf{y})$, $p(\lambda_2|else, \mathbf{y})$, $p(p|else, \mathbf{y})$ e $p(z_i|else, \mathbf{y})$ como as condicionais completas de λ_1, λ_2, p e z_i, respectivamente, e seja *else* o vetor de parâmetros sem o parâmetro de interesse da distribuição condicional completa. Por exemplo, para $p(\lambda_1|else, \mathbf{y})$, o interesse é λ_1 e *else* é igual a $\lambda_2, p, z_1, \ldots, z_n$. Use apenas a notação das condicionais completas, não precisa deduzí-las.

7. (1,0 ponto) Considere um modelo de mistura de Poisson com duas classes para $y_1, \ldots, y_n$ com dados aumentados em $z_1, \ldots, z_n$, iid, $n = 5$.

   (a) Para que serve o *burn-in*?

   (b) Por que tomamos amostras da distribuição a posteriori 'pulando' (*thinning*) elementos das cadeias de Markov obtidas pelo amostrador de Gibbs?

   (c) A seguir, considere a tabela abaixo, a qual associa a estatística R de Gelman a partir de duas cadeias para cada quantidade desconhecida do modelo estimada pelo amostrador de Gibbs. Note que para algumas quantidades R é bem maior do que 1. O que isso indica? Qual medida você tomaria para resolver o problema?

	R
λ_1	2.1
λ_2	1.1
p	3
z_1	2.5
z_2	1.2
z_3	1.1
z_4	1.6
z_5	1.3

8. Considere um modelo de mistura de Poisson com duas classes para $y_1, \ldots, y_n$ com dados aumentados em $z_1, \ldots, z_n$, iid, para os dados de incidência de terremotos descrito em (1), onde consideramos o objeto *counts* como a amostra em estudo. O código abaixo é utilizado para obter uma amostra da distribuição preditiva a posteriori de uma nova observação y_{n+1}, considerando que os objetos $lambda.1$, $lambda.2$, e p contém amostras a posteriori dos parâmetros λ_1, λ_2 e p após aplicar o *burn-in* e *thinning*. O objeto *deny[i,]* contém amostras da distribuição preditiva da contagem associada ao índice i. O código considera que observação futura y_{n+1} é independente da amostra $\mathbf{y}$.

```
## preditiva de y_{n+1}

yy = 0:max(counts)
ny = length(yy)
deny = matrix(0,ny,niter) ## densidade de y
for (i in 1:ny)
  deny[i,] = p*dpois(yy[i],lambda.1)+(1-p)*dpois(yy[i],lambda.2)
```

   (a) (1,0 ponto) Escreva um trecho de código para estimar $P(y_{n+1} = 0|\mathbf{y})$ usando a amostra da distribuição preditiva.

   (b) (1,0 ponto) Escreva um trecho de código para encontrar um intervalo de credibilidade de 95% para $P(y_{n+1} = 0|\mathbf{y})$ usando a amostra da distribuição preditiva.