Pacote bgmm

...

Sobre o pacote

Usando o método “all”, temos nesse pacote duas formas de inicializar o algoritmo EM

Caso exista apenas uma variável, a inicialização é dada por

- mu = matrix(sapply(1:k, function(i) mean(dat[dat < quantile(dat, i/k) & dat > quantile(dat, (i - 1)/k)])), k, 1)
- cvar = array(sapply(1:k, function(i) var(dat[dat < quantile(dat, i/k) & dat > quantile(dat, (i - 1)/k)])), c(k, 1, 1))
- pi = rep(1/k, k)

Caso tenha mais de uma variável, será usado k-means 10 vezes e selecionar o melhor modelo

Em ambos os casos o critério de parada é $\frac{|\ell(\theta_k) - \ell(\theta_{k-1})|}{1 + |\ell(\theta_{k-1})|} \leq 10^{-5}$

Conjunto de dados utilizado

Temos 2 alelos, e 333 SNPs (polimorfismo único no nucleotídeo)

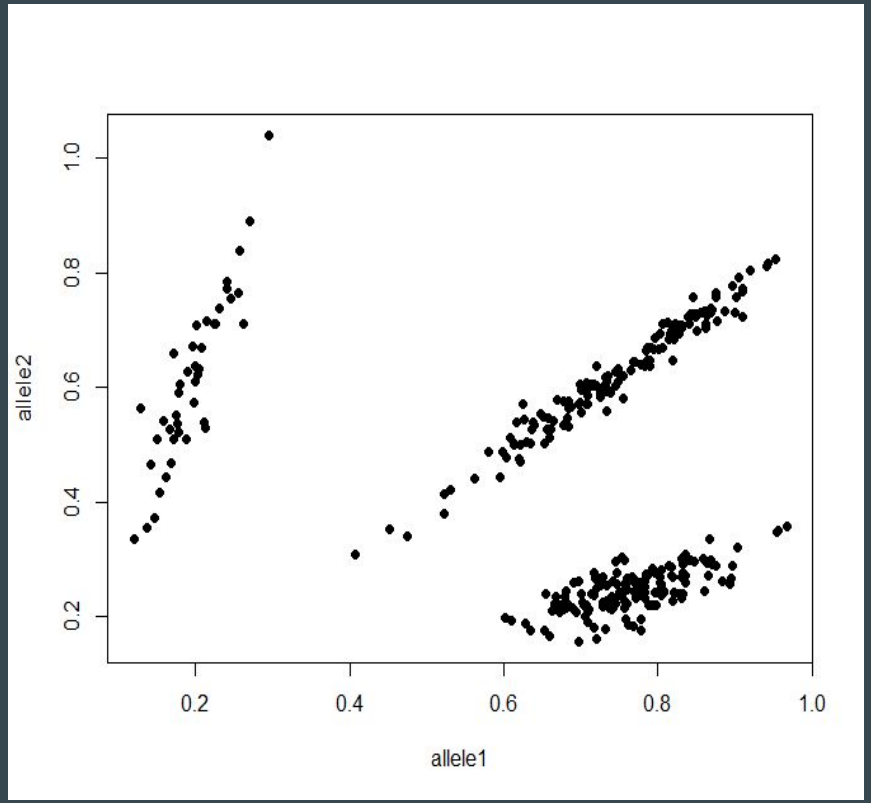
Valores referem-se à intensidade do sinal de luminescência para cada alelo nesse SNP

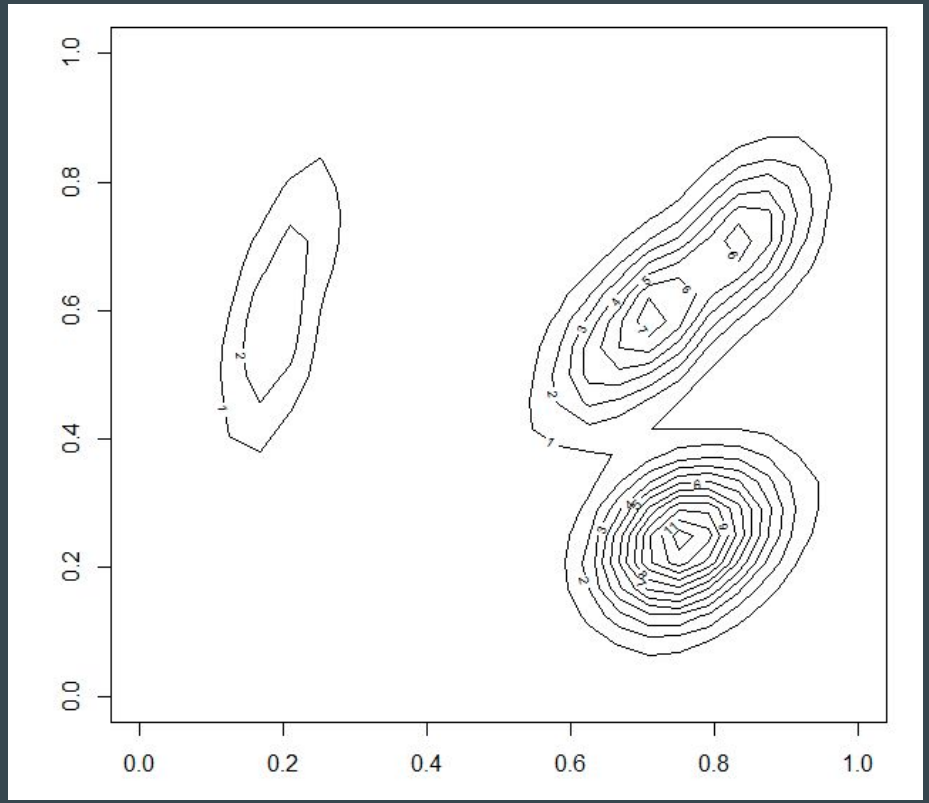
Cada SNP é caracterizado pela presença de um dos dois alelos possíveis (ou pela pela presença de ambos)

Conhecidamente são 3 grupos

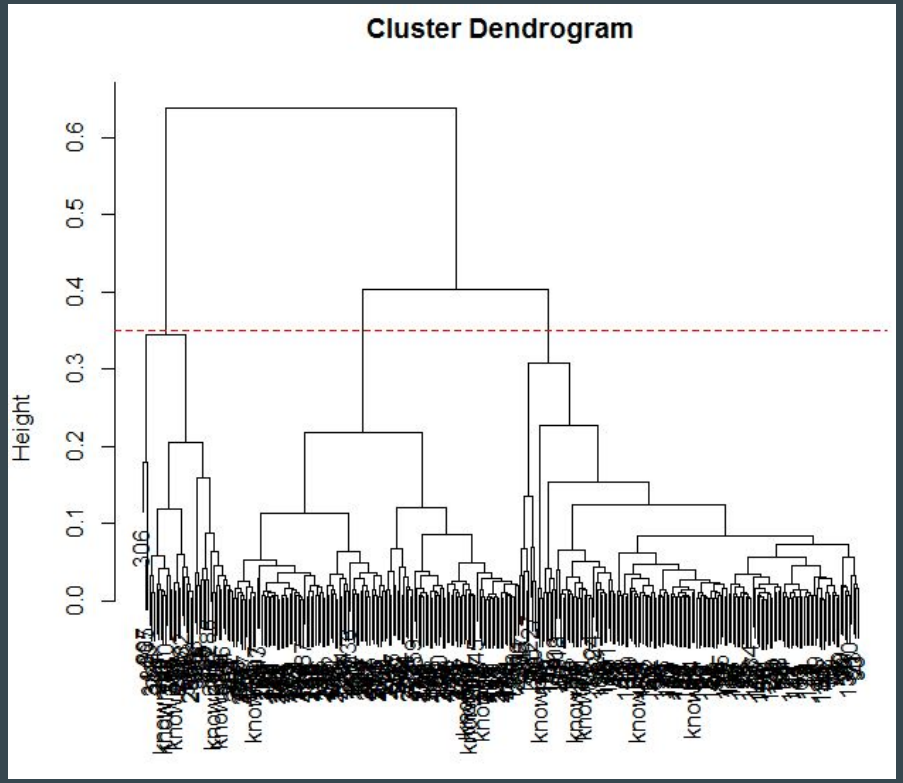
	allele1	allele2
1	0.740865	0.235506
2	0.681937	0.223097
3	0.796681	0.277226
4	0.771643	0.245674
5	0.737848	0.216432
6	0.723360	0.267854

Análise exploratória dos dados

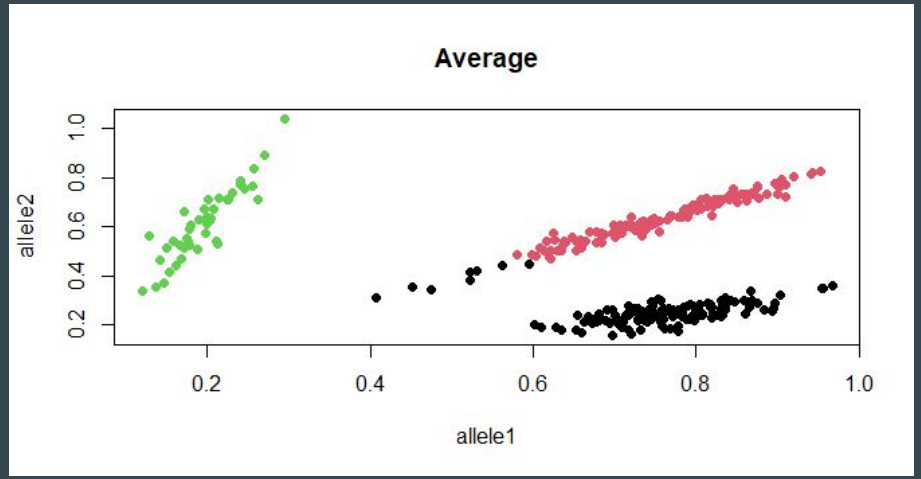




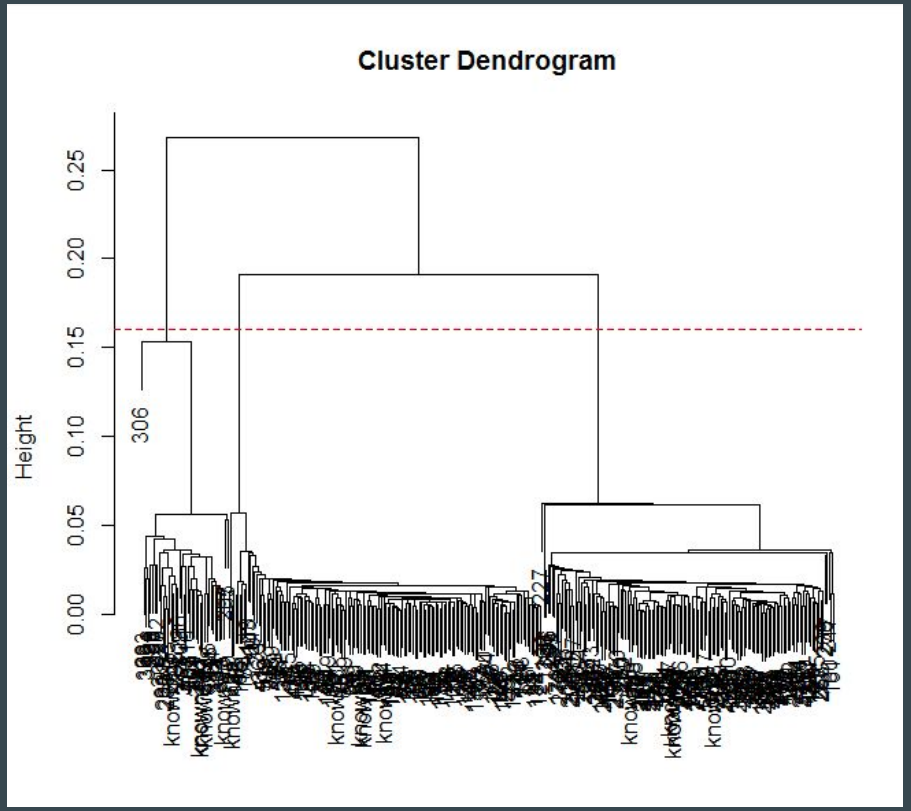
Método hierárquico

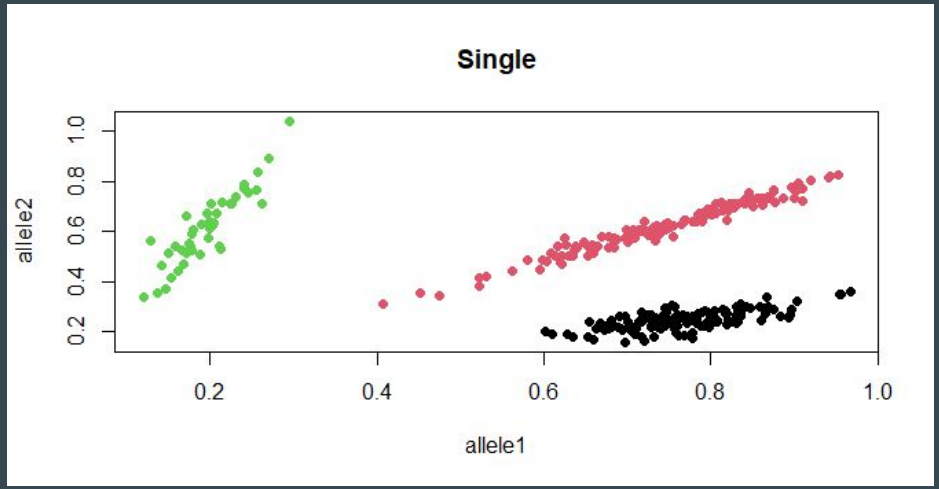



```
plot(hclust(dist(m), method = "average"))
 abline(h = .35, lty = "dashed", col = "red")
```




```
modelo = hclust(dist(m), method = "average")
cortado = cutree(modelo,k=3)
plot(m[,1], m[,2], col = cortado, xlab = "alelo 1", ylab = "alelo 2")
abline(h = .16, lty = "dashed", col = "red")
```

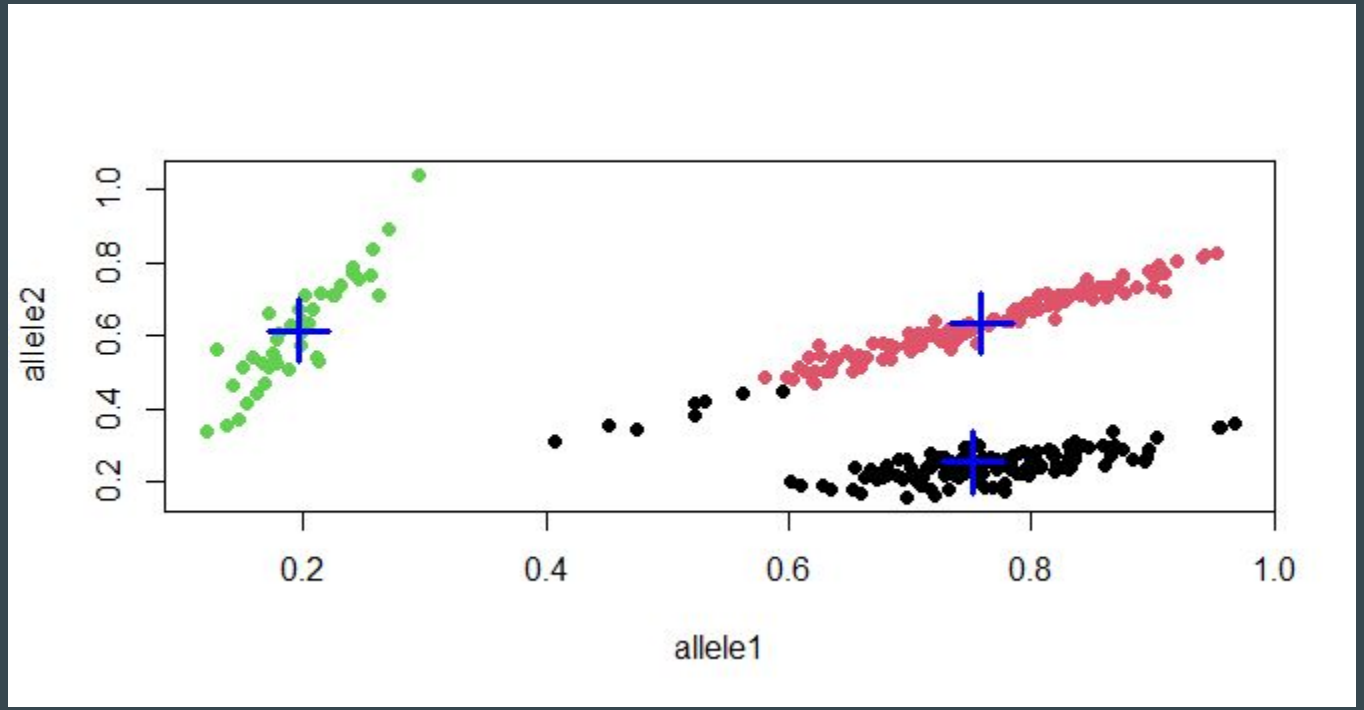





```
plot(hclust(dist(m), method = "single"))
 abline(h = .16, lty = "dashed", col = "red")
```

```
modelo = hclust(dist(m), method = "single")
cortado = cutree(modelo,k=3)
plot(m[,1], m[,2], col = cortado, xlab = "alelo 1", ylab = "alelo 2")
```

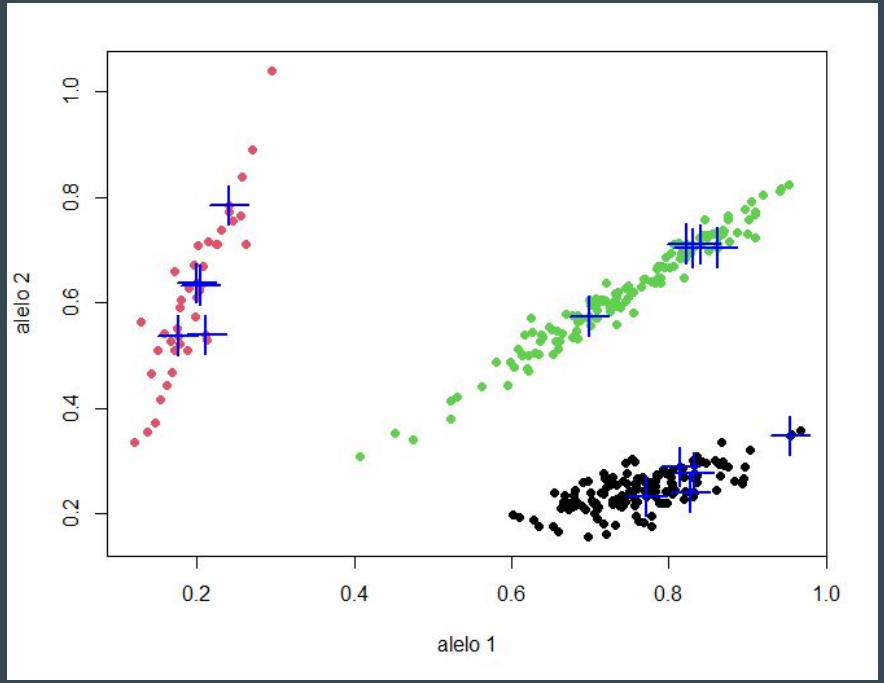
K-means



Mistura finita de normais

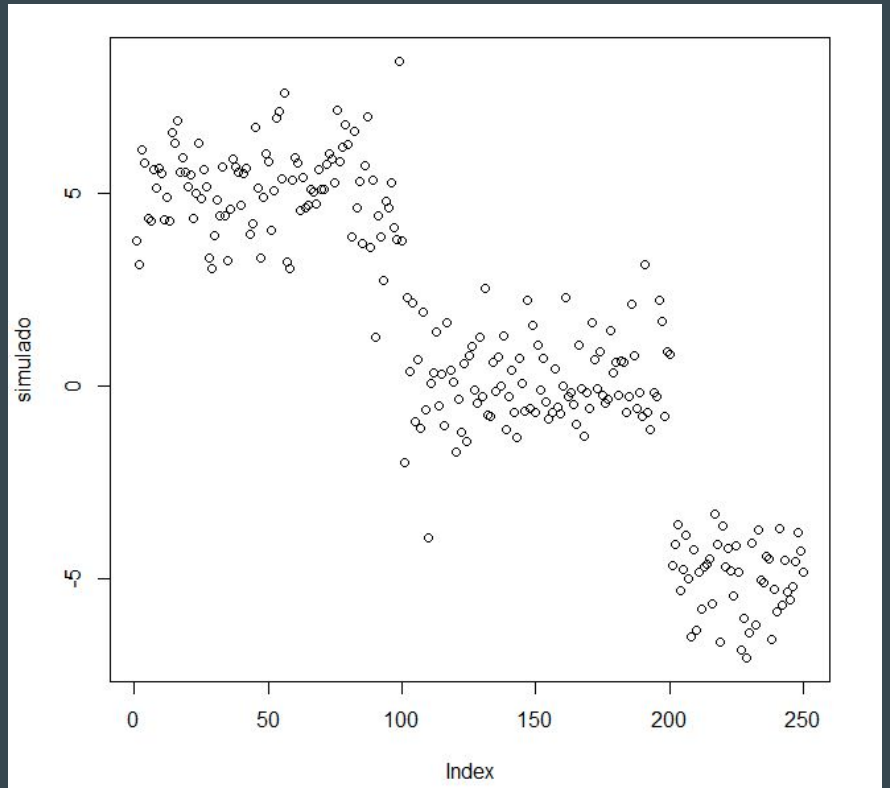
```
mistura <- unsupervised(m, k = 3)

plot(m[,1], m[,2],
col = apply(mistura$tij, 1, which.max),
xlab = "alelo 1", ylab = "alelo 2", pch = 19)

points(tail(m[,1], 15), tail(m[,2], 15), col = "blue", pch = 3)

mistura$mu
```

```
          [,1]      [,2]
[1,] 0.7449416 0.6197700
[2,] 0.1961935 0.6126590
[3,] 0.7648964 0.2455181
```



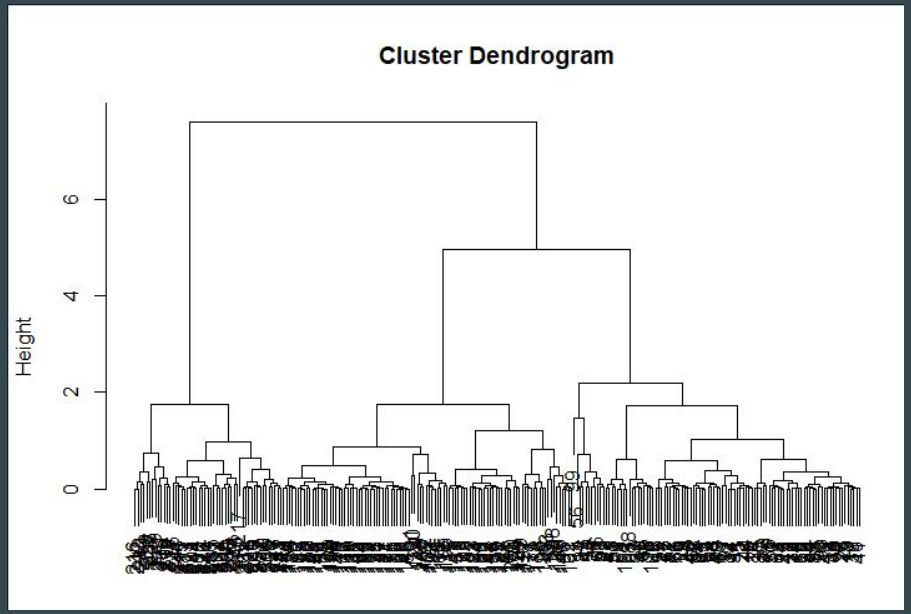
Simulação grupos bem separados

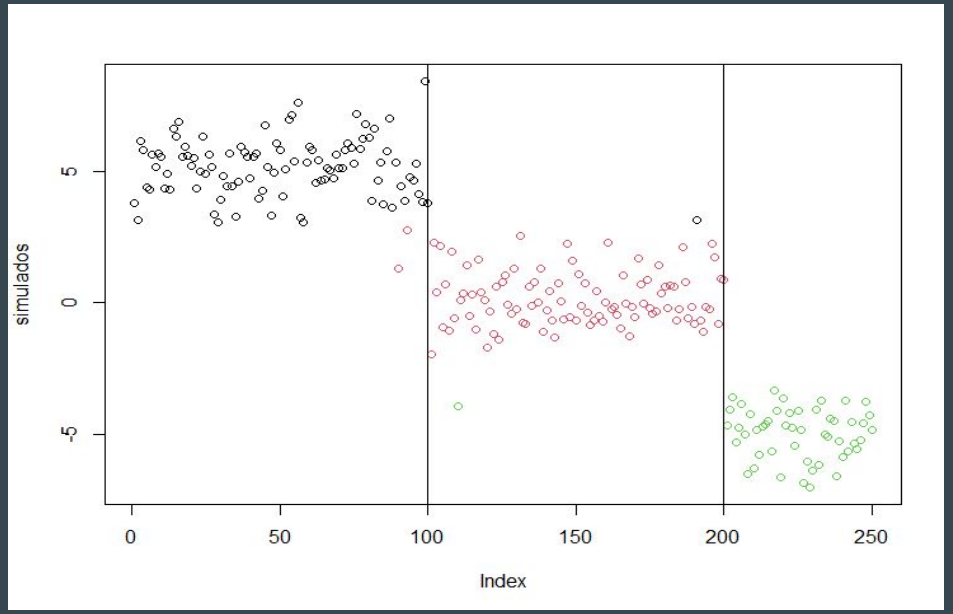
Foram geradas 100 amostras de uma normal com média 5, 100 de uma normal com média 0 e 50 com média -5, todas com variância igual a 1



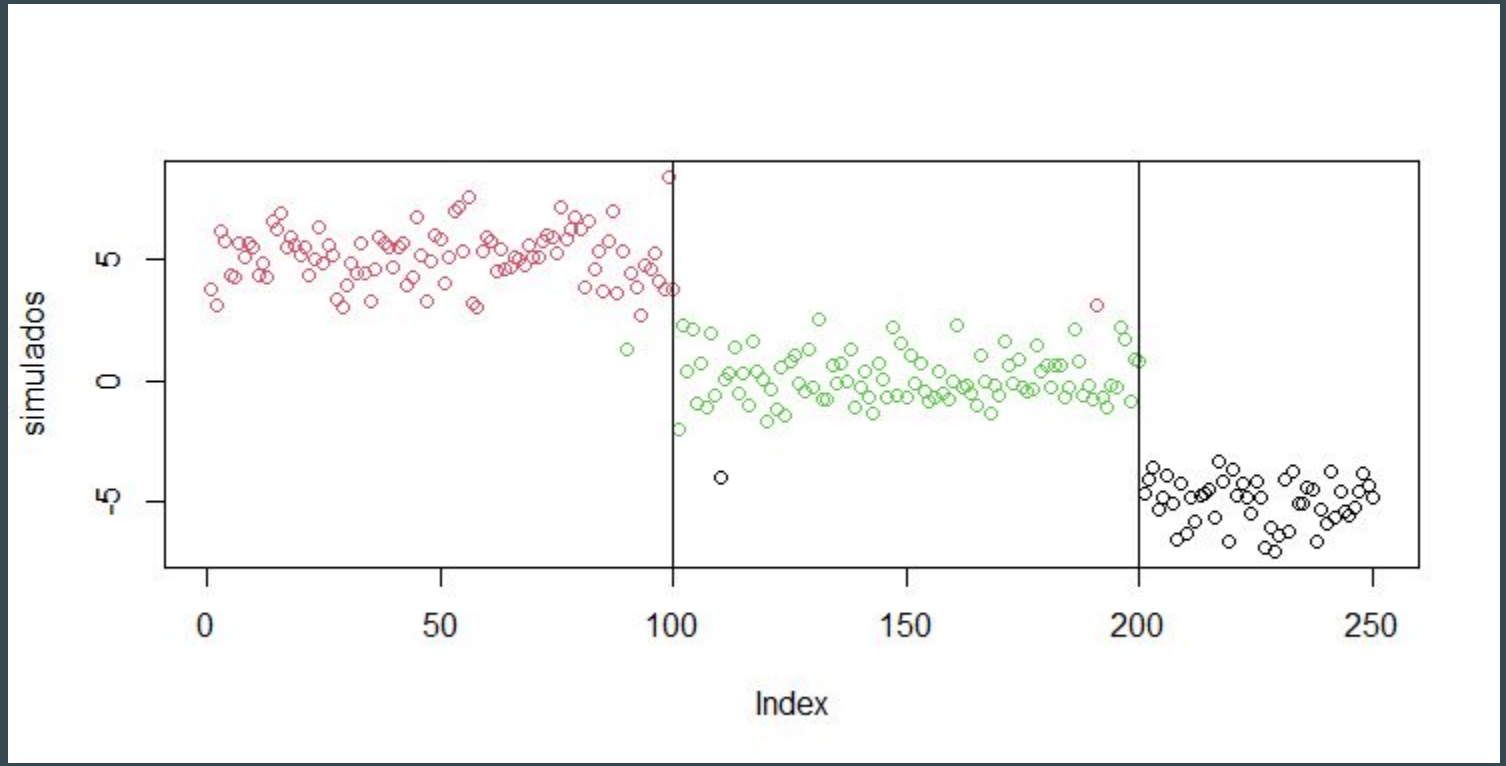
Hierárquico

method = “average”

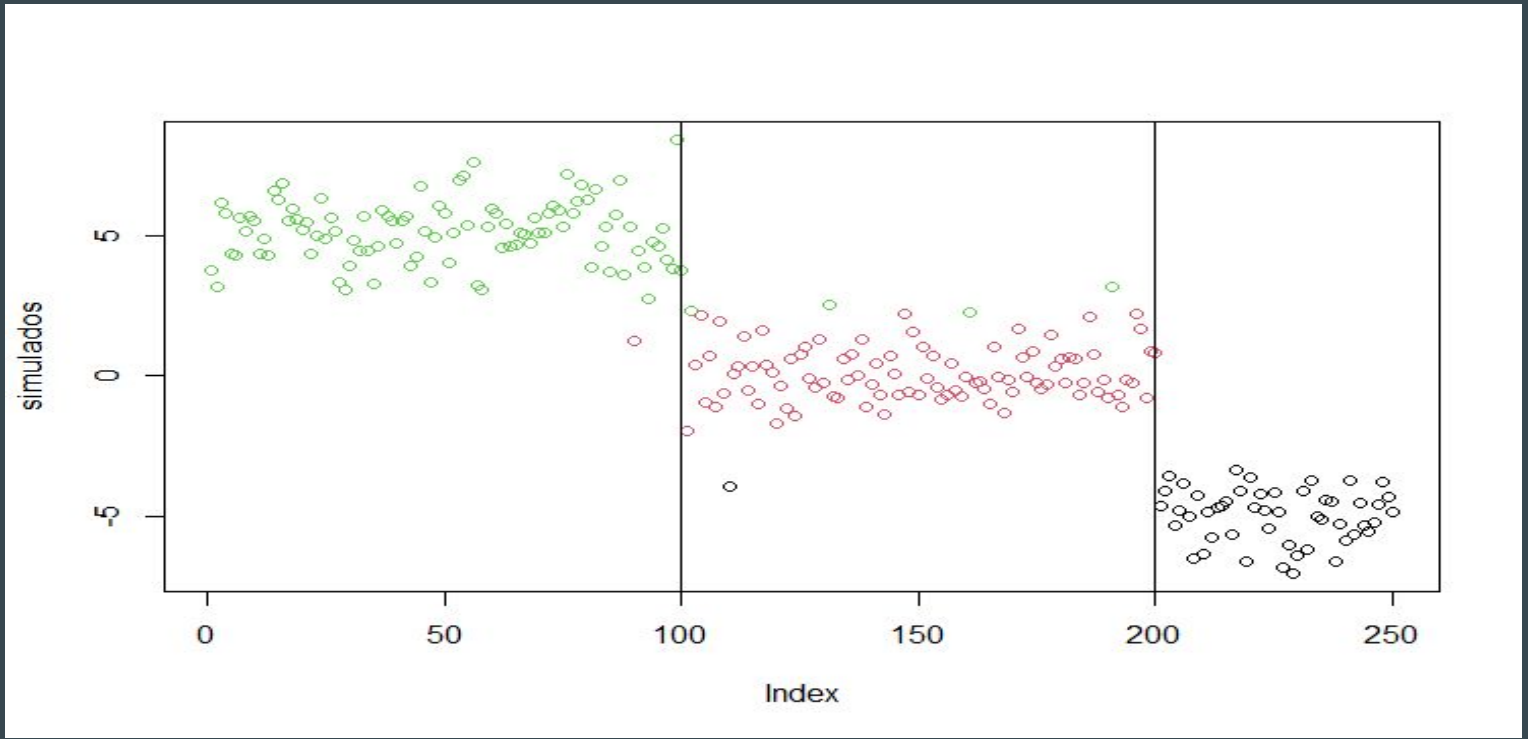




K-means

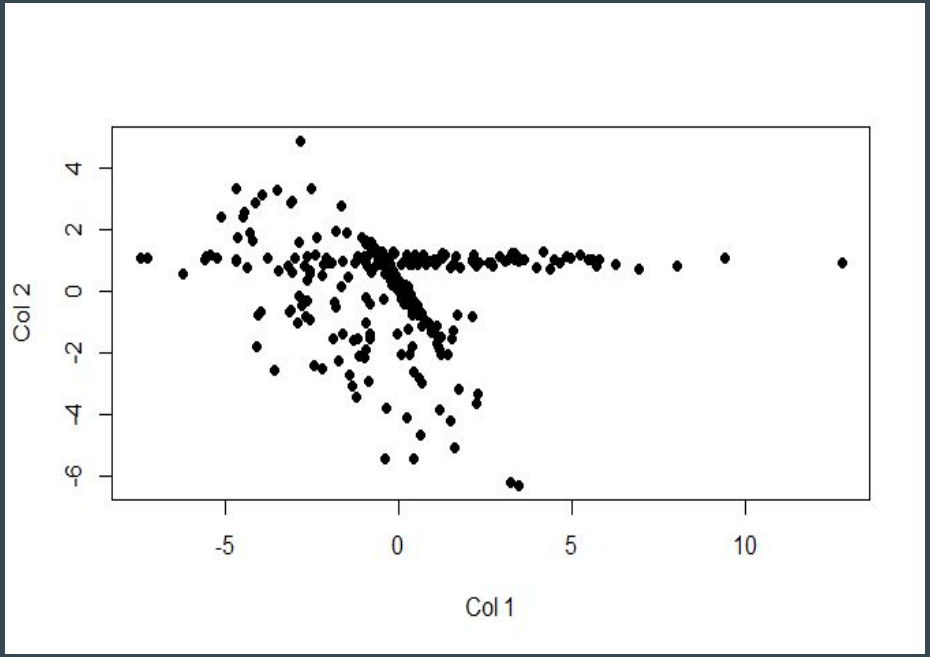


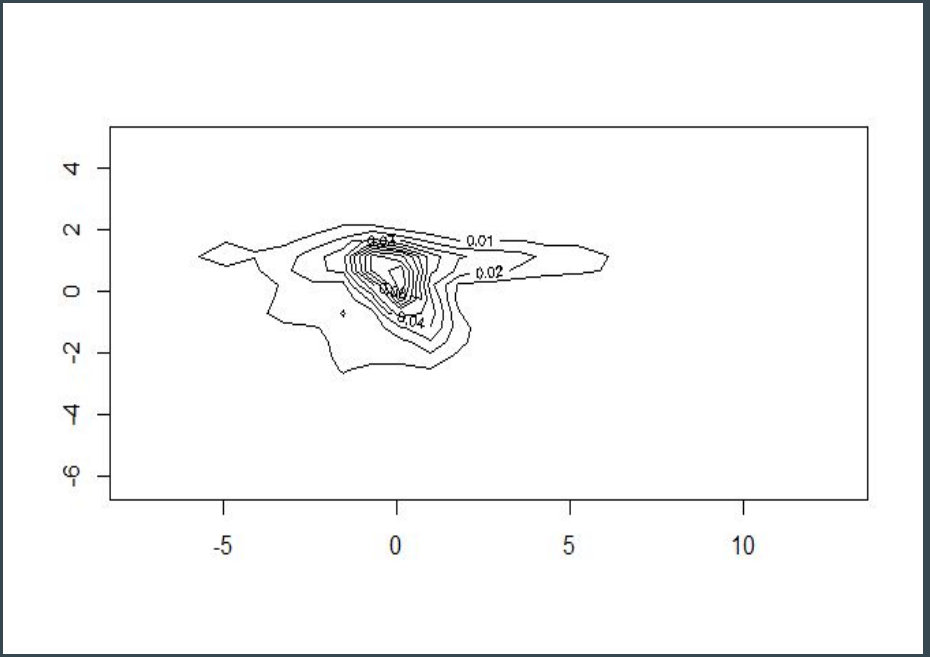
Mistura de normais



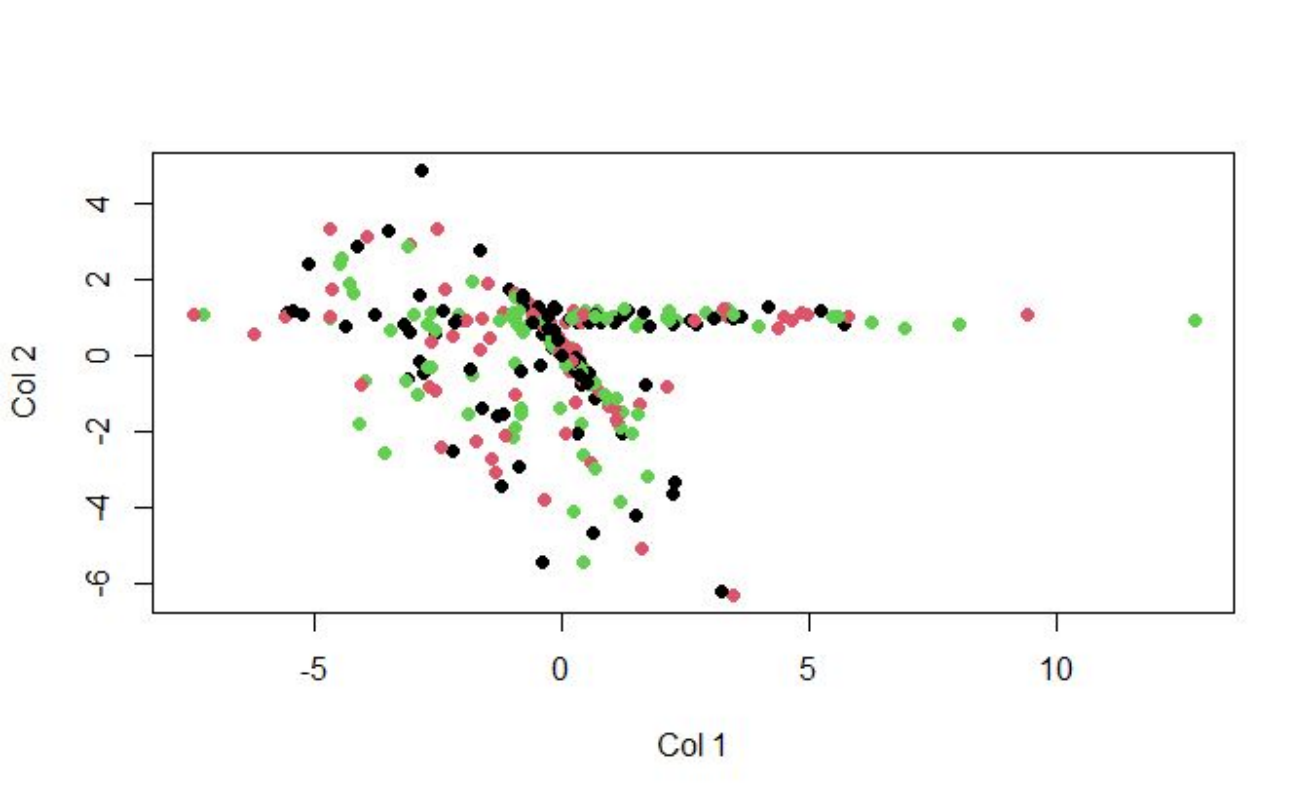
Simulação grupos mal separados

simulados <- bgmm::simulateData(d=2, k=3, n=300, mu=matrix(c(1,1,0,.25,-1,-1), ncol = 2,byrow = TRUE))

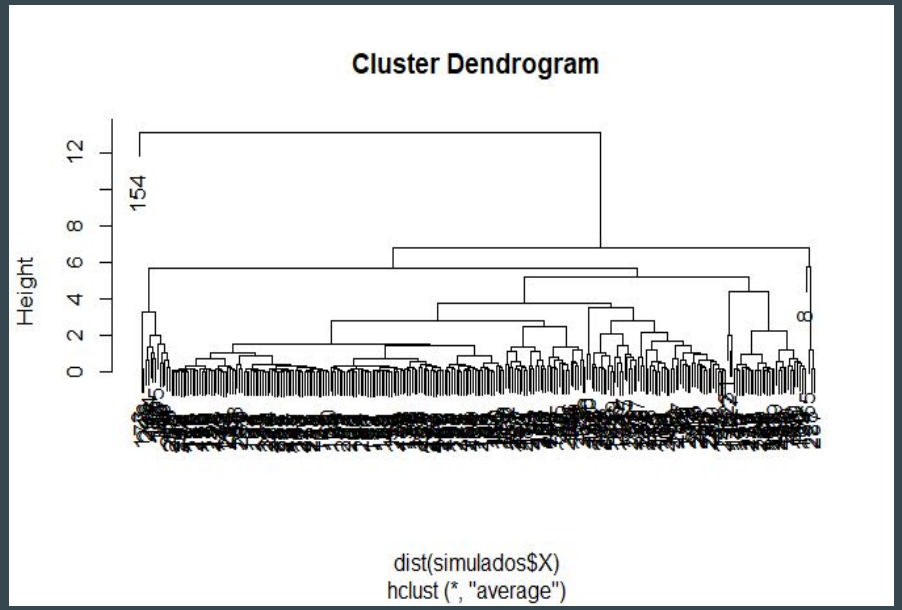


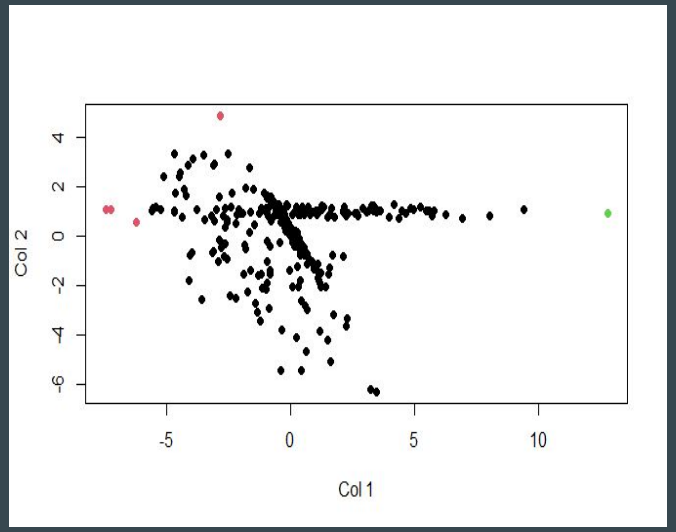


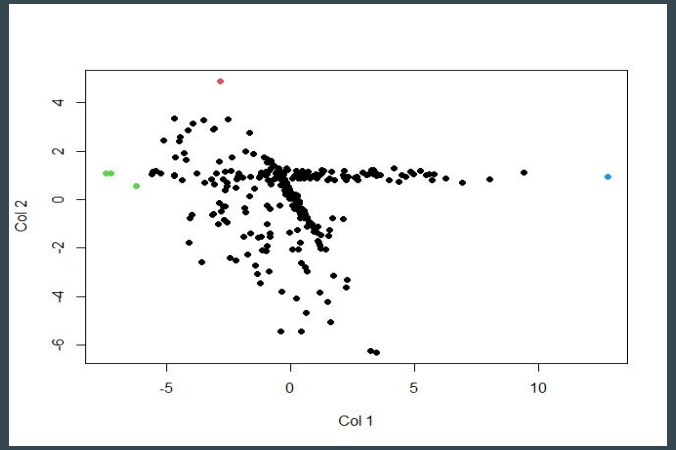
Separação real dos grupos



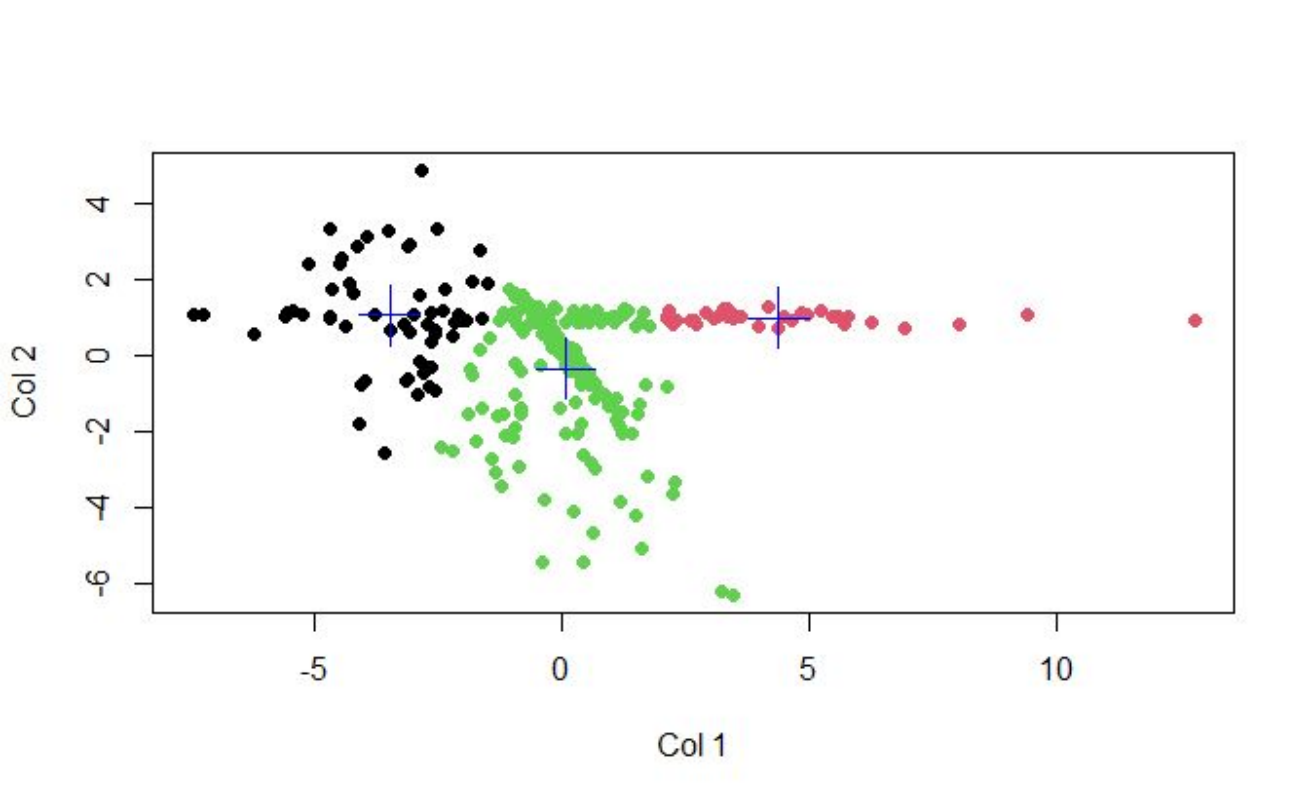
Hierárquico



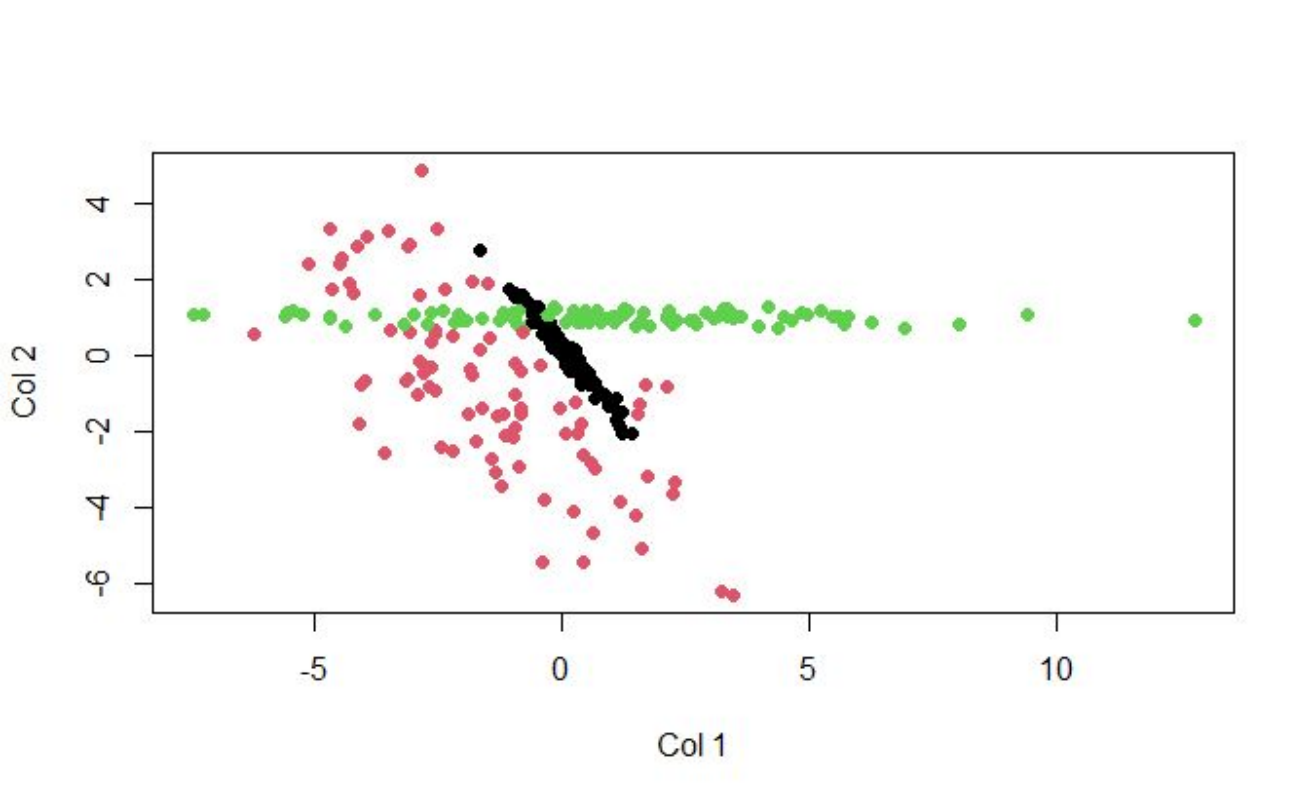




K-means



Misturas




```
> mod$mu
             [,1]       [,2]
[1,]  0.08762893  0.1006036
[2,] -1.45850285 -0.7622414
[3,]  0.99773258  1.0045836
```

Matrizes de confusão

hierárquico

	1	2	3
1	91	2	1
2	97	0	0
3	97	2	0

K-means

	1	2	3
1	21	35	38
2	1	0	96
3	38	0	61

mistura

	1	2	3
1	2	0	92
2	95	0	2
3	8	88	3