Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

18 julho 2023

Prova 3

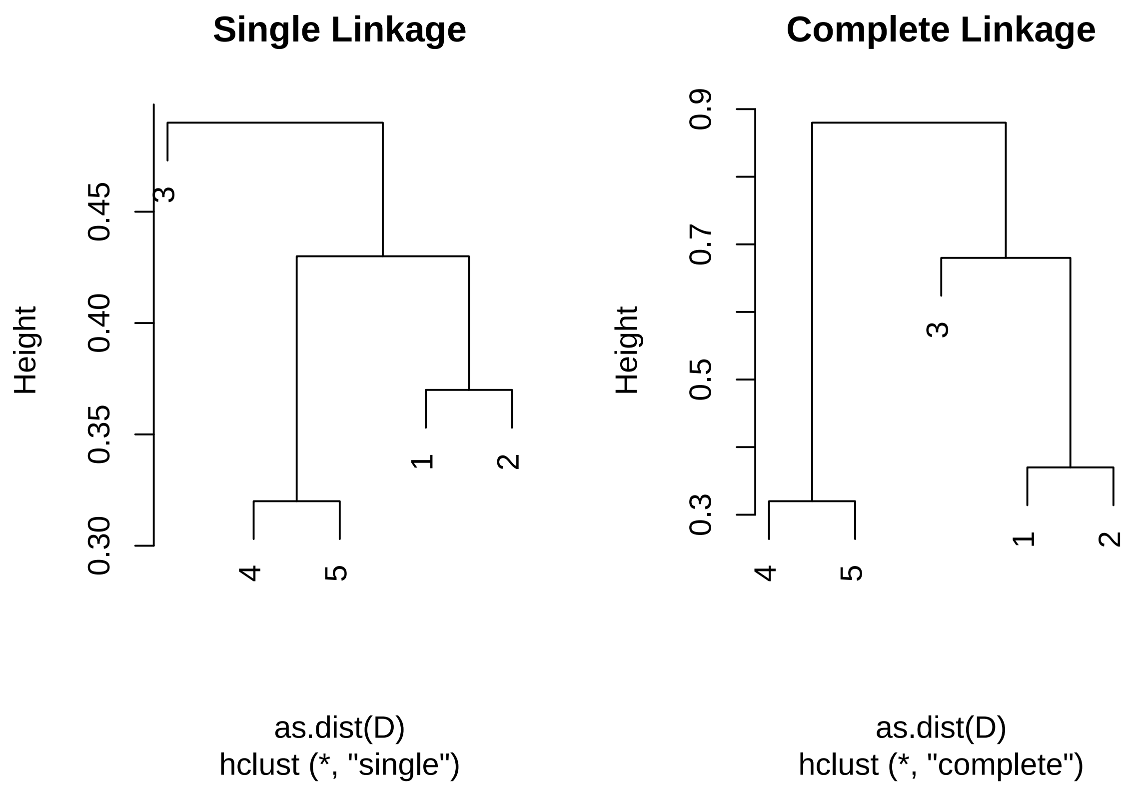
Prof. Dr. George von Borries

Análise Multivariada 1

Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

Página 01/39

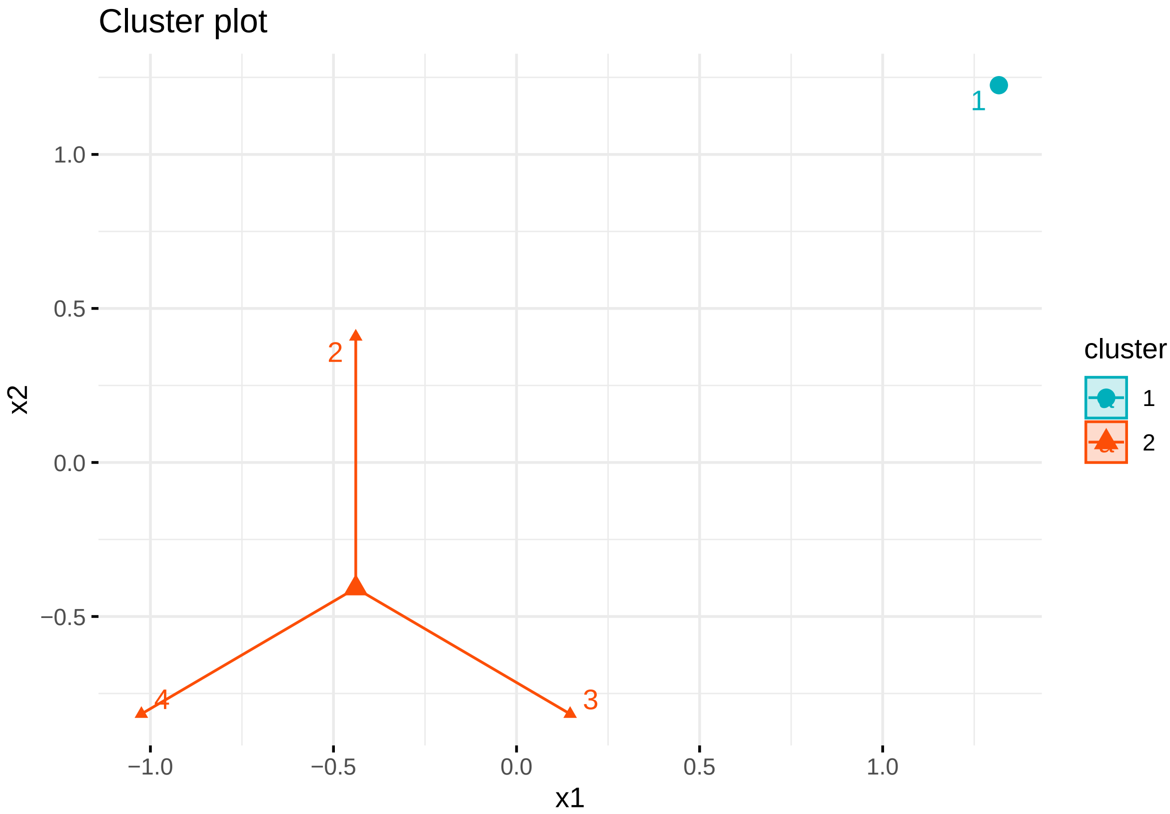
79. Johnson e Wichern - Exercício 12.7.



Analisando os dendogramas, percebemos que tanto as abordagens simples e completa agregaram os valores (1,2) e (4,5) no mesmo grupo, mas diferiram quanto a agregação do valor (3); no caso da agregação simples, o valor (3) foi caracterizado como um grupo robustamente separado dos dois demais grupos, enquanto na agregação completa, o *cluster* do valor (3) foi colocado como mais próximo do *cluster* dos valores (1,2), e esses mais distantes do *cluster* dos valores (4,5).

81. Johnson e Wichern - Exercício 12.12.

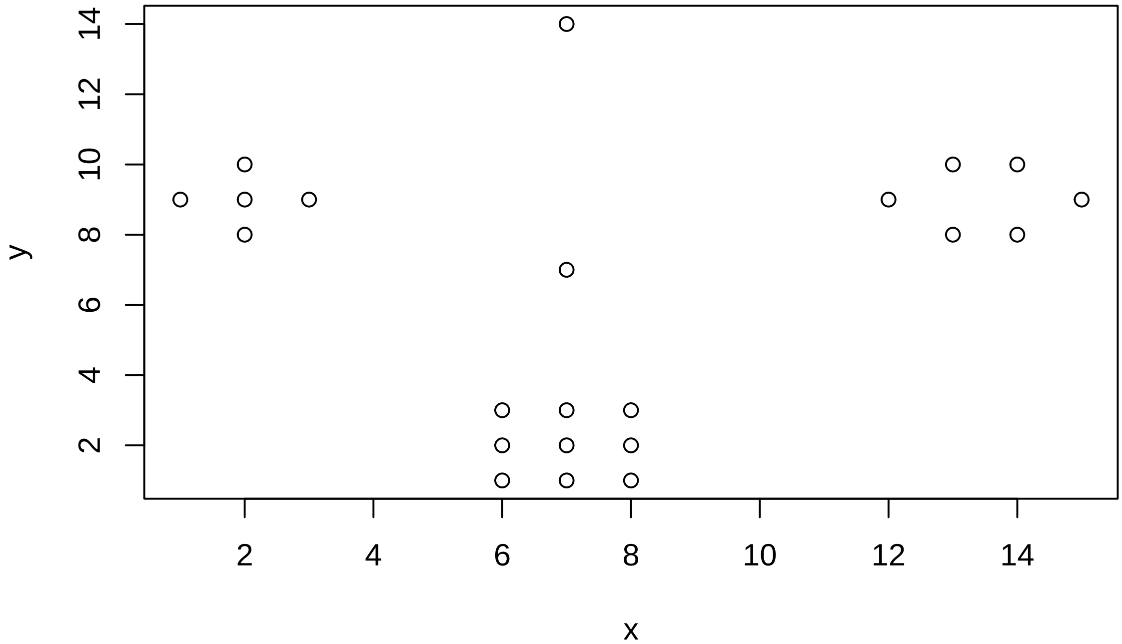
```
## Too few points to calculate an ellipse
## Too few points to calculate an ellipse
```



Conforme elucidado pelo prof. George, o algoritmo *k-means*, após decidir os centros dos grupos (neste caso, ele partiu do que eu defini manualmente inicialmente), itera os pontos afim de encontrar os centros e agrupar de tal modo que minimize a variabilidade dentro; e maximize a variabilidade entre os clusters. No caso, este ponto "ótimo" é o mesmo que o calculado no Exercício 12.11 do livro; portanto independente deu alterar os centros iniciais, o processo iterativo sempre vai retornar para este valor. Isto é verdade pelo número baixo de pontos e número alto de iterações permitidas. Conjuntos com muitos pontos e número de iterações reduzido por vezes irão produzir resultados aglomerativos diferentes.

83.

a)



Pela análise do gráfico, aparetam haver entre 3 a 5 grupos: sendo 3 grupos sólidos agrupados, e 2 *outliers* dispersos que provavelmente otimizariam formando um grupo para cada, ou ainda podem ser talvez agregados a algum dos 3 grupos mais robustos, porém aumentando assim sua dispersão.

b)

Irei apresentar os valores na forma corrida para caber melhor no documento, mas é bom observar que a forma "natural" destes valores são matrizes triangulares inferiores. Favor verificar o código para exibi-los estruturados.

Distâncias euclidianas:

1.4142136, 1, 1.4142136, 2, 7.8102497, 11, 12.0415946, 12.0415946, 13.0384048, 13.0384048, 14, 6.3245553, 7.8102497, 8.4852814, 9.2195445, 8.6023253, 9.2195445, 9.8994949, 9.4339811, 10, 10.6301458, 1, 2, 1.4142136, 6.4031242, 10.0498756, 11, 11.1803399, 12, 12.1655251, 13.0384048, 5.8309519, 8.0622577, 8.6023253, 9.2195445, 8.9442719, 9.4339811, 10, 9.8488578, 10.2956301, 10.8166538, 1, 1, 7.0710678, 10, 11.045361, 11.045361, 12.0415946, 12.0415946, 13, 5.3851648, 7.2111026, 7.8102497, 8.4852814, 8.0622577, 8.6023253, 9.2195445, 8.9442719, 9.4339811, 10, 1.4142136, 7.8102497, 10.0498756, 11.1803399, 11, 12.1655251, 12, 13.0384048, 5.0990195, 6.4031242, 7.0710678, 7.8102497, 7.2111026, 7.8102497, 8.4852814, 8.0622577, 8.6023253, 9.2195445, 6.4031242, 9, 10.0498756, 10.0498756, 11.045361, 11.045361, 12, 4.472136, 6.7082039, 7.2111026, 7.8102497, 7.6157731, 8.0622577, 8.6023253, 8.5440037, 8.9442719, 9.4339811, 7.0710678, 7.2111026, 8.4852814, 8.0622577, 9.2195445, 9.4339811, 7, 11.045361, 11, 11.045361, 12.0415946, 12, 12.0415946, 13.0384048, 13, 13.0384048, 1.4142136, 1.4142136, 2.236068, 2.236068, 3, 5.3851648, 8.4852814, 7.8102497, 7.2111026, 9.2195445, 8.6023253, 8.0622577, 10, 9.4339811, 8.9442719, 2, 1, 2.236068, 2.236068, 6.7082039, 9.8994949, 9.2195445, 8.6023253, 10.6301458, 10, 9.4339811, 11.4017543, 10.8166538, 10.2956301, 2.236068, 1, 2.236068, 6.0827625, 8.6023253, 7.8102497, 7.0710678, 9.2195445, 8.4852814, 7.8102497, 9.8994949, 9.2195445, 8.6023253, 2, 1.4142136, 7.6157731, 10.6301458, 9.8994949, 9.2195445, 11.3137085, 10.6301458, 10, 12.0415946, 11.4017543, 10.8166538, 1.4142136, 7.0710678, 9.4339811, 8.6023253, 7.8102497, 10, 9.2195445, 8.4852814, 10.6301458, 9.8994949, 9.2195445, 8.2462113, 10.8166538, 10, 9.2195445, 11.4017543, 10.6301458, 9.8994949, 12.0415946, 11.3137085, 10.6301458, 4.1231056, 4, 4.1231056, 5.0990195, 5, 5.0990195, 6.0827625, 6, 6.0827625, 1, 2, 1, 1.4142136, 2.236068, 2, 2.236068, 2.8284271, 1, 1.4142136, 1, 1.4142136, 2.236068, 2, 2.236068, 2.236068, 1.4142136, 1, 2.8284271, 2.236068, 2, 1, 2, 1, 1.4142136, 2.236068, 1, 1.4142136, 1, 1.4142136, 2.236068, 1.4142136, 1, 1, 2, 1

Distâncias 'Manhattan':

2, 1, 2, 2, 11, 11, 13, 13, 14, 14, 14, 8, 11, 12, 13, 12, 13, 14, 13, 14, 15, 1, 2, 2, 9, 11, 11, 13, 12, 14, 14, 8, 11, 12, 13, 12, 13, 14, 13, 14, 15, 1, 1, 10, 10, 12, 12, 13, 13, 13, 7, 10, 11, 12, 11, 12, 13, 12, 13, 14, 2, 11, 11, 13, 11, 14, 12, 14, 6, 9, 10, 11, 10, 11, 12, 11, 12, 13, 9, 9, 11, 11, 12, 12, 12, 6, 9, 10, 11, 10, 11, 12, 11, 12, 13, 10, 10, 12, 11, 13, 13, 7, 12, 11, 12, 13, 12, 13, 14, 13, 14, 2, 2, 3, 3, 3, 7, 12, 11, 10, 13, 12, 11, 14, 13, 12, 2, 1, 3, 3, 9, 14, 13, 12, 15, 14, 13, 16, 15, 14, 3, 1, 3, 7, 12, 11, 10, 13, 12, 11, 14, 13, 12, 2, 2, 10, 15, 14, 13, 16, 15, 14, 17, 16, 15, 2, 8, 13, 12, 11, 14, 13, 12, 15, 14, 13, 10, 15, 14, 13, 16, 15, 14, 17, 16, 15, 5, 4, 5, 6, 5, 6, 7, 6, 7, 1, 2, 1, 2, 3, 2, 3, 4, 1, 2, 1, 2, 3, 2, 3, 3, 2, 1, 4, 3, 2, 1, 2, 1, 2, 3, 1, 2, 1, 2, 3, 2, 1, 1, 2, 1

Distâncias de Mahalanobis:

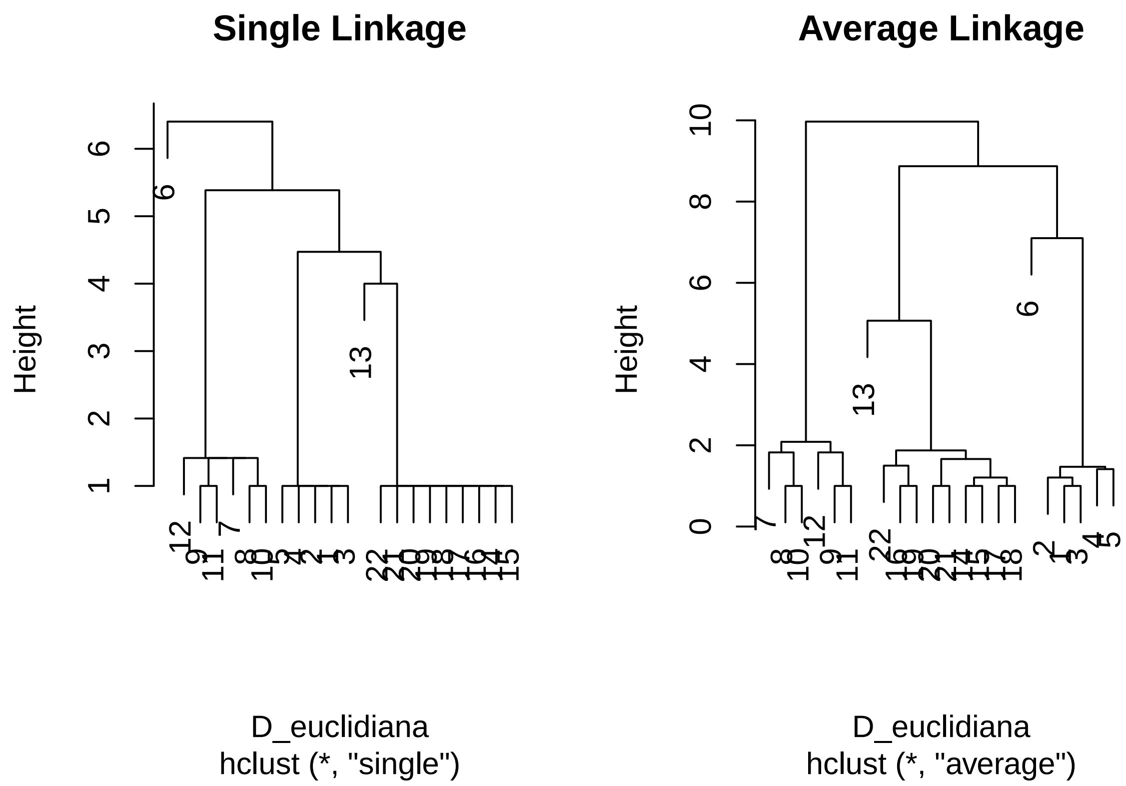
0.1089331, 0.0554805, 0.1355485, 0.2219218, 3.2670729, 6.7131345, 7.8962527, 8.2156376, 9.2699563, 9.6159566, 10.8741683, 2.4240299, 4.1896137, 4.8797448, 5.6808369, 5.1240362, 5.8274751, 6.6418749, 6.1919795, 6.9087261, 7.7364335, 0.0667603, 0.2670413, 0.1355485, 2.1890222, 5.7478824, 6.7131345, 7.2729452, 7.9891849, 8.575611, 9.6159566, 2.1874696, 4.5315583, 5.1240362, 5.8274751, 5.5861938, 6.1919795, 6.9087261, 6.7743499, 7.3934433, 8.1234976, 0.0667603, 0.0554805, 2.7233265, 5.548045, 6.6335101, 6.9262795, 7.8962527, 8.2156376, 9.3761961, 1.7871295, 3.6104434, 4.1896137, 4.8797448, 4.5315583, 5.1240362, 5.8274751, 5.5861938, 6.1919795, 6.9087261, 0.1089331, 3.3911514, 5.4817283, 6.6874063, 6.7131345, 7.9368412, 7.9891849, 9.2699563, 1.5203101, 2.8228491, 3.3887117, 4.0655352, 3.6104434, 4.1896137, 4.8797448, 4.5315583, 5.1240362, 5.8274751, 2.290541, 4.4939165, 5.4817283, 5.7478824, 6.6335101, 6.9262795, 7.9891849, 1.2611901, 3.142234, 3.6104434, 4.1896137, 4.0500412, 4.5315583, 5.1240362, 5.091369, 5.5861938, 6.1919795, 3.3887117, 3.3848461, 4.8797448, 4.1593228, 5.6808369, 5.7520648, 3.2712554, 7.9870936, 8.0779979, 8.2798631, 9.5092731, 9.6134851, 9.828658, 11.1649733, 11.282493, 11.5109736, 0.1089331, 0.1355485, 0.2620667, 0.3152975, 0.4993241, 1.5209755, 3.9215902, 3.3911514, 2.9716736, 4.709628, 4.192497, 3.7863269, 5.6311865, 5.1273632, 4.7345007, 0.2670413, 0.0554805, 0.3491371, 0.3152975, 2.3586004, 5.3377199, 4.709628, 4.192497, 6.2459707, 5.6311865, 5.1273632, 7.2877421, 6.6862656, 6.19575, 0.2959063, 0.0554805, 0.2620667, 1.9842103, 3.9217803, 3.2670729, 2.7233265, 4.5629898, 3.9215902, 3.3911514, 5.3377199, 4.709628, 4.192497, 0.2670413, 0.1355485, 3.0399231, 6.0767728, 5.3377199, 4.709628, 6.9717158, 6.2459707, 5.6311865, 8.0001795, 7.2877421, 6.6862656, 0.1089331, 2.6921485, 4.6874485, 3.9217803, 3.2670729, 5.3153503, 4.5629898, 3.9215902, 6.0767728, 5.3377199, 4.709628, 3.6048668, 6.1786718, 5.3153503, 4.5629898, 6.9267865, 6.0767728, 5.3377199, 7.8084219, 6.9717158, 6.2459707, 1.0704147, 1.068165, 1.1768763, 1.6579498, 1.6690078, 1.7910268, 2.3790055, 2.4033713, 2.538698, 0.0554805, 0.2219218, 0.0667603, 0.1355485, 0.3152975, 0.2670413, 0.3491371, 0.5421939, 0.0554805, 0.1089331, 0.0667603, 0.1355485, 0.2959063, 0.2670413, 0.3491371, 0.2620667, 0.1089331, 0.0667603, 0.4357322, 0.2959063, 0.2670413, 0.0554805, 0.2219218, 0.0667603, 0.1355485, 0.3152975, 0.0554805, 0.1089331, 0.0667603, 0.1355485, 0.2620667, 0.1089331, 0.0667603, 0.0554805, 0.2219218, 0.0554805

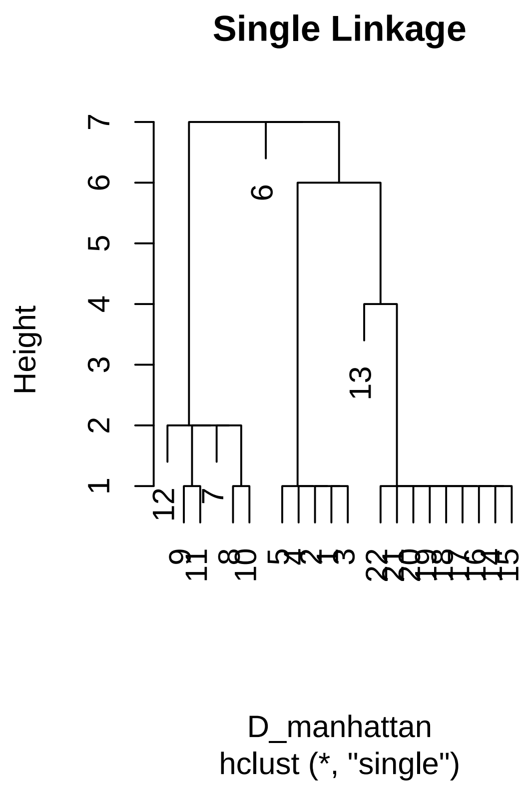
Apesar dos valores serem bem diferentes, isso se dá mais pelo método de cálculo de distância de cada uma das técnicas. A distância Euclidiana trabalha basicamente com a "distância bruta" entre um ponto e outro, literalmente medindo a distância linear. A distância Manhattan trabalha com distância absoluta entre as coordenadas dos pontos. A distância de Mahalanobis busca centralizar os dados, calculando as distâncias levando em consideração a correlação entre as dimensões.

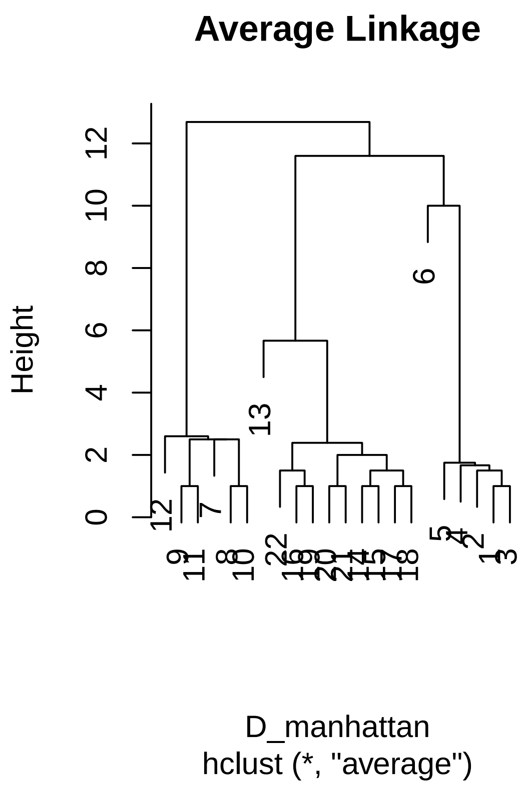
Portanto, apesar de improvável, é possível que mesmo com valores observados absolutamente distoantes, agrupar as variáveis segundo as três distâncias trabalhadas e em todos os casos, retornar os exatos mesmos clusters pros dados.

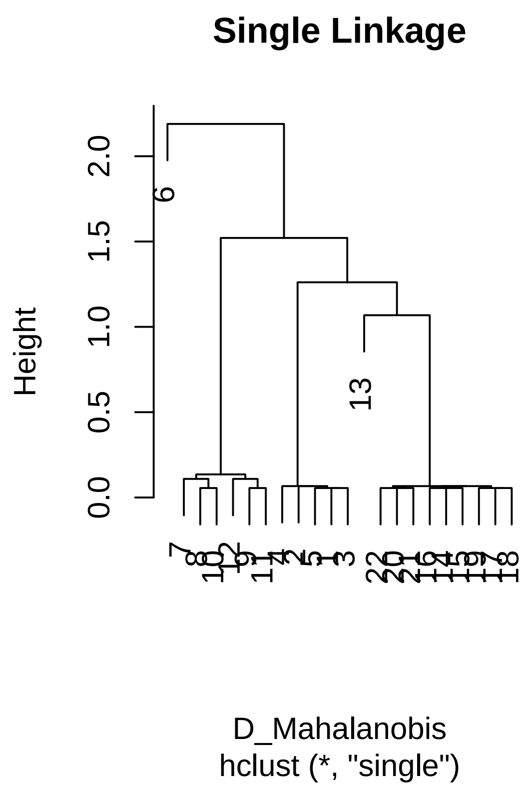
c)

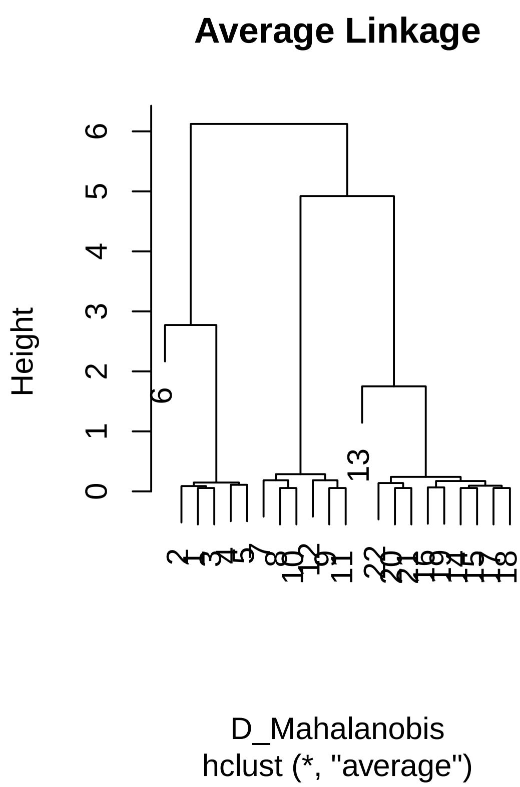
Irei apresentar corridamente três painéis, cada um composto por dois dendogramas (agregação simples e média), referentes respectivamente aos valores de distância Euclidiana, Manhattan e de Mahalanobis.







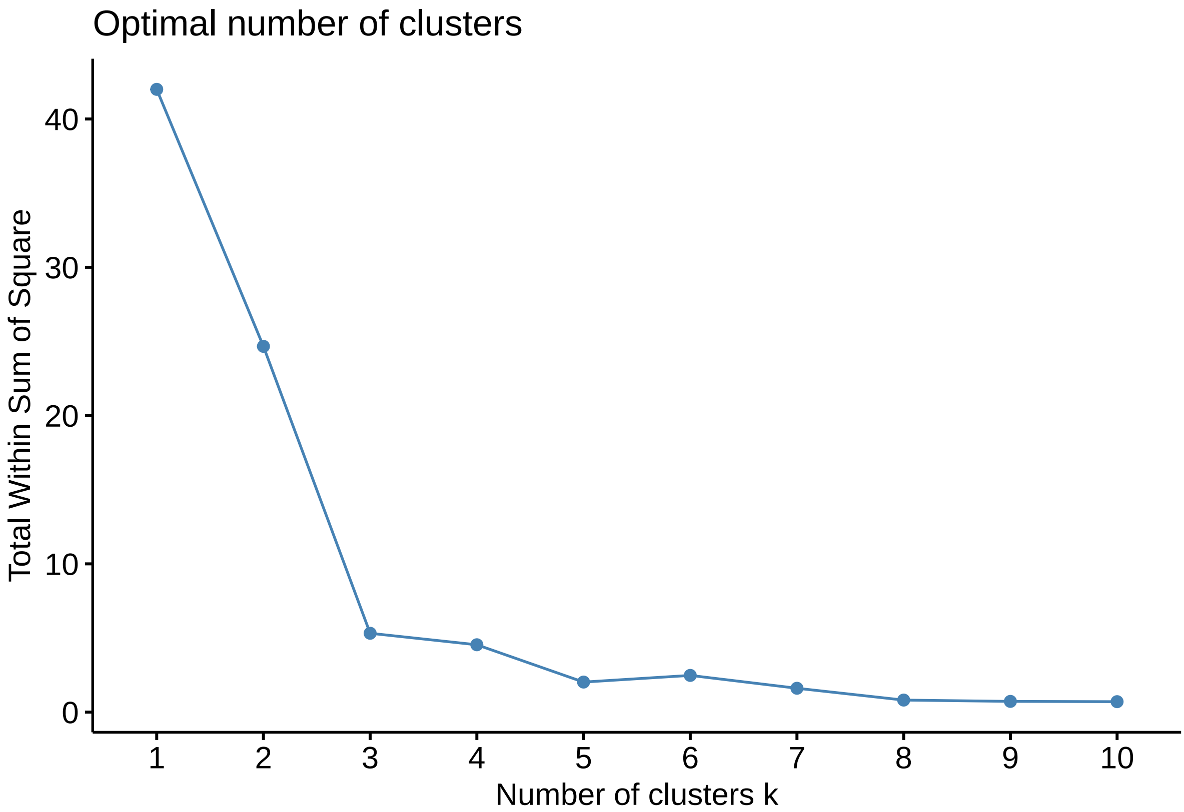




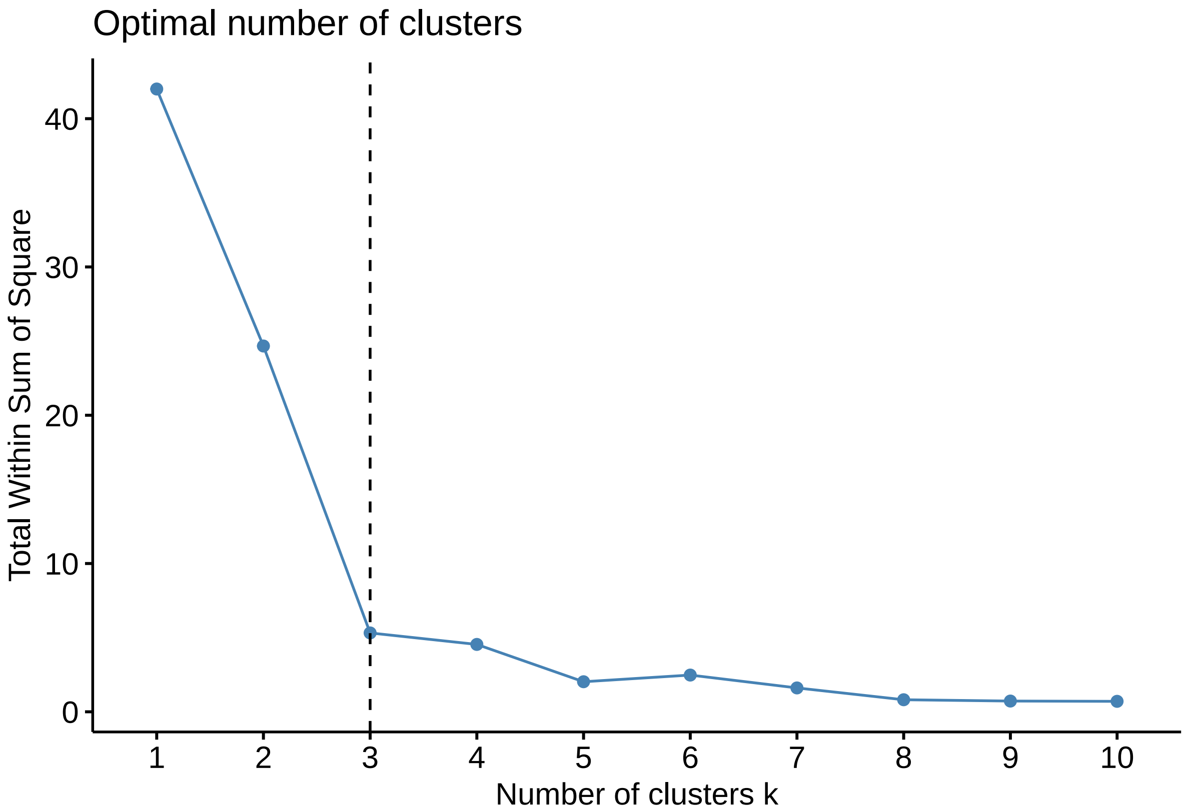
Em todos os dendogramas, foi confirmada a suspeita levantada no item (a); em que haviam 3 grupos aglomerativos bem definidos, e mais 2 grupos formados cada um por apenas um *outlier*. Cada dendograma teve seu formato específico, mas todos foram eficientes em agrupar os dados pelos seus similares.

d)

Primeiro, devemos identificar o número ideal de *clusters*, já que o *k-means* necessita que o usuário entre manualmente com o número de *clusters* que o algoritmo deve separar. Já foi visto anteriormente que o número é 3 ou 5, dependendo da abordagem que queira se fazer quanto aos *outliers*. Porém, irei também seguir a praxe deste algorítmo, que é *plotar* um gráfico que ajuda a determinar o número ideal de *clusters*.

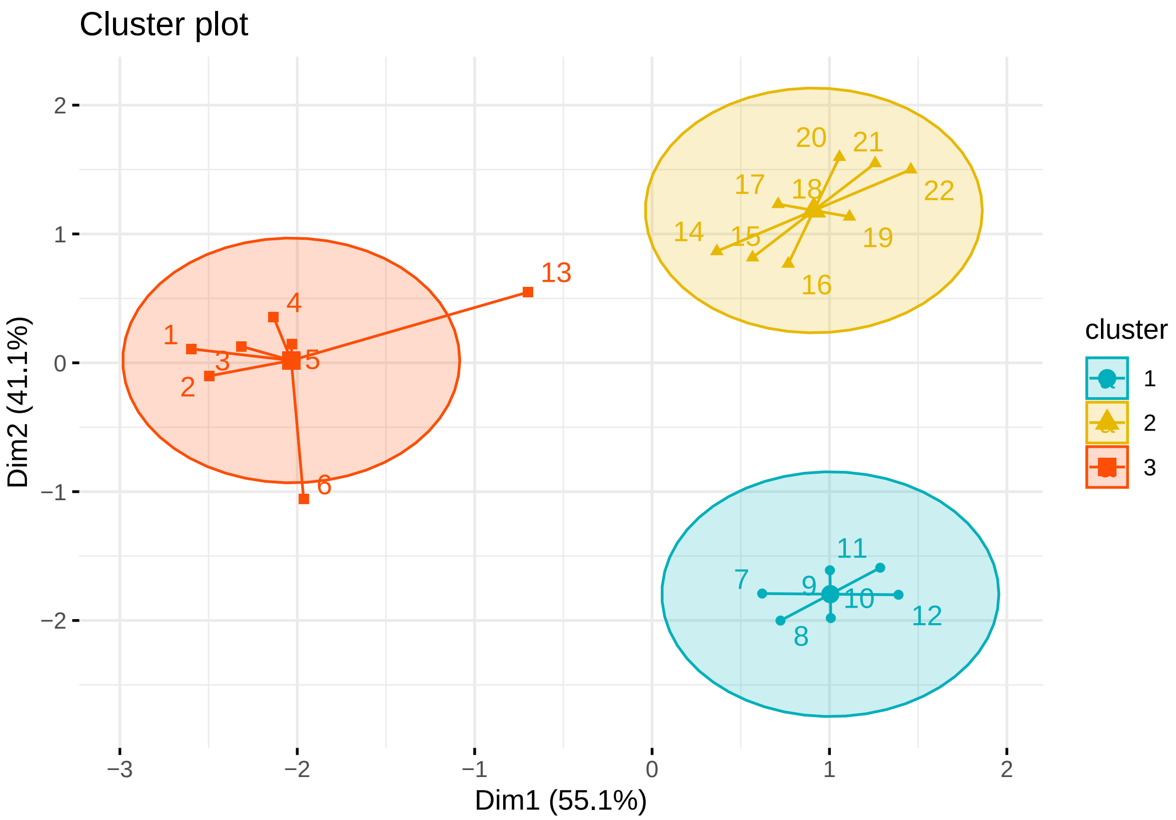


Pelo método de *elbow*, o número ideal são 3 *clusters*...


Optimal number of clusters
Total Within Sum of Square
0
10
20
30
40
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Portanto, executando o *k-means* para 3 *clusters*, iremos obter o seguinte resultado:

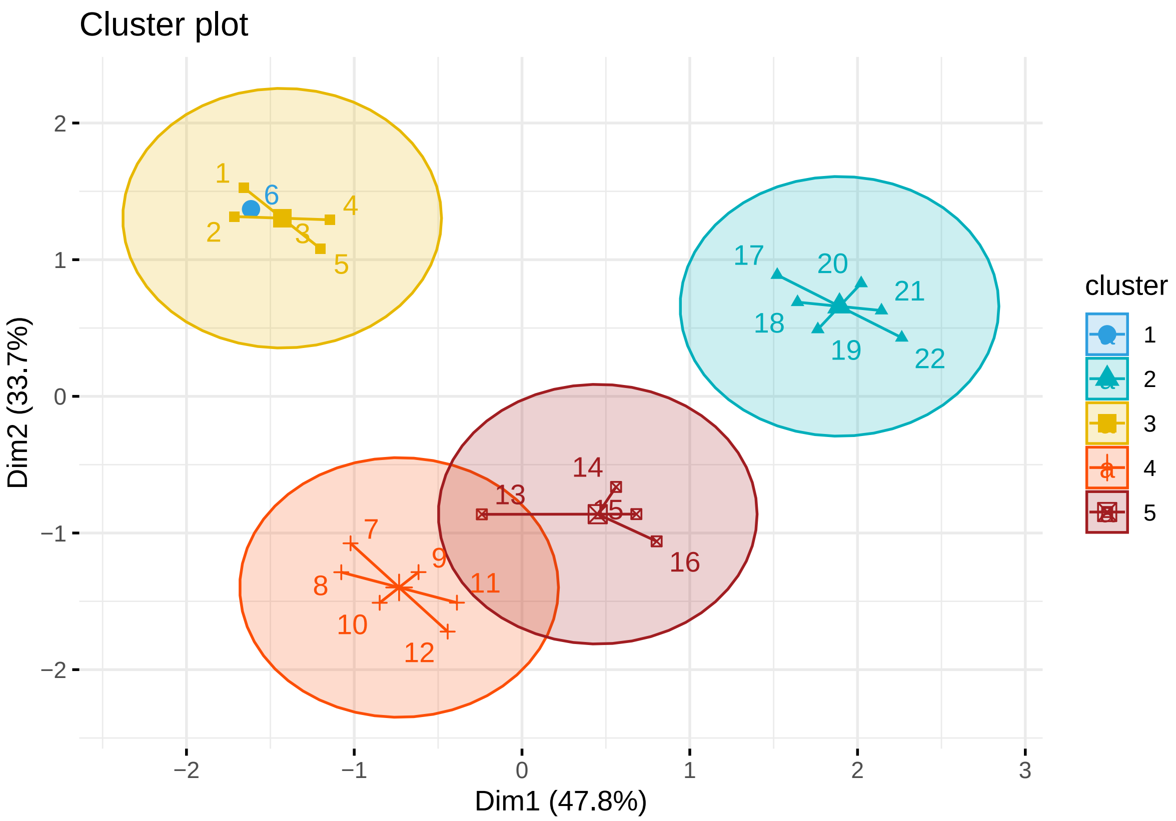
```
##   cluster          x          y
## 1       1  1.3728584  0.7004490
## 2       2 -0.1489924 -1.0973700
## 3       3 -0.9851742  0.8105195
```



Aqui, notamos que o *k-means* foi relativamente eficiente em classificar os dois *outliers* em um dos *clusters*, sem muita perda de generalização.

Porém, se quisermos forçar a mão e testar a aglomeração *k-means* com 5 grupos, este será o resultado:

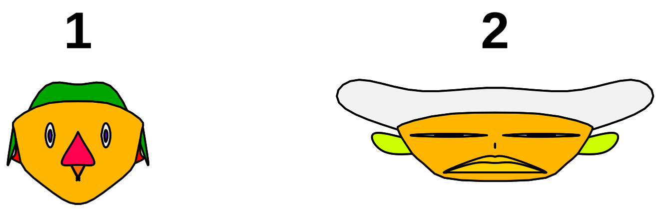
```
##   cluster ponto    x    y cluster
## 1       1   6.0  7.0 14.0    3.00
## 2       2  19.5  7.0  1.5    2.00
## 3       3   3.0  2.0  9.0    3.00
## 4       4   9.5 13.5  9.0    1.00
## 5       5  14.5  7.0  4.0    2.25
```



Em que notamos que o *k-means* não foi nada eficiente em identificar os *outliers* cada um como sendo um grupo robustamente separado dos outros três.

84.

a)

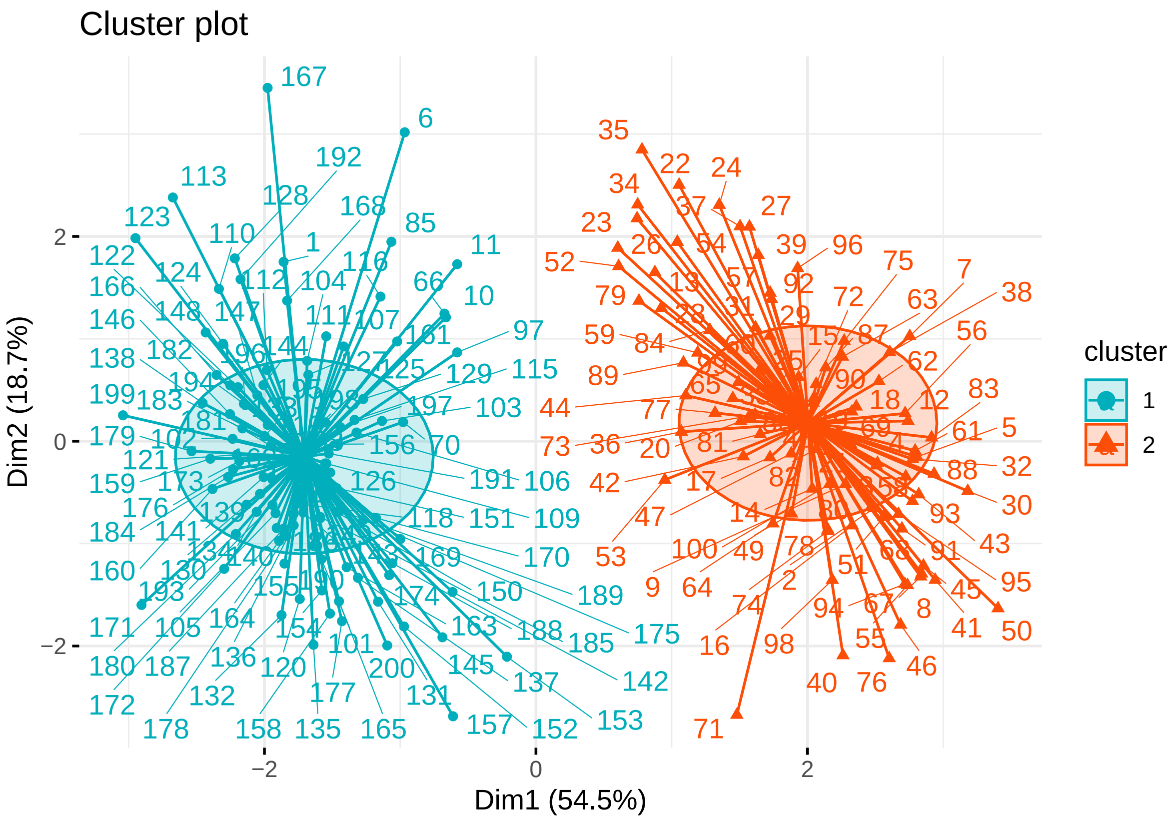


Apesar d'eu particularmente não gostar desse tipo de gráfico, por talvez trazer um ar de ridículo a um trabalho potencialmente sério, é inegável seu valor num exemplo como esse, em que conseguimos identificar de forma simples e didática a diferença entre os dois grupos de notas disponíveis, de forma muito mais visual que vetores numéricos ou gráficos potencialmente de interpretação complexa para o público leigo.

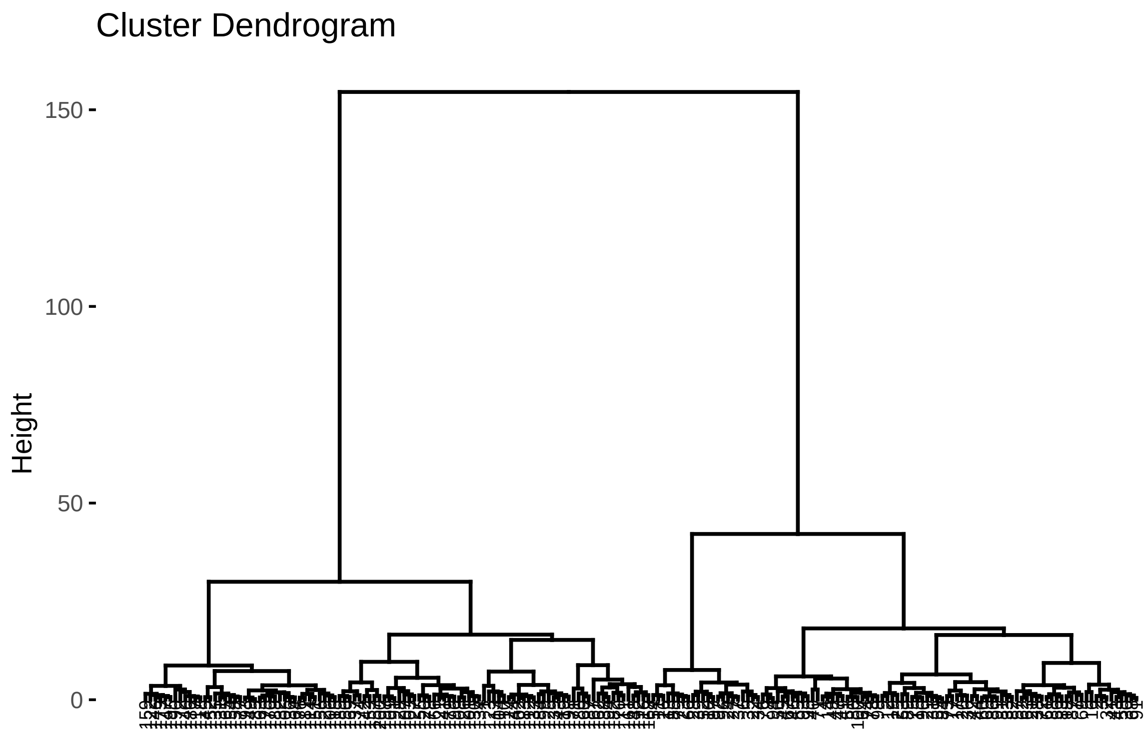
b)

Aqui, irei testar diferentes formas de agrupamento, para avaliar quais métodos performam melhor para este conjunto de dados.

***k-means*:**



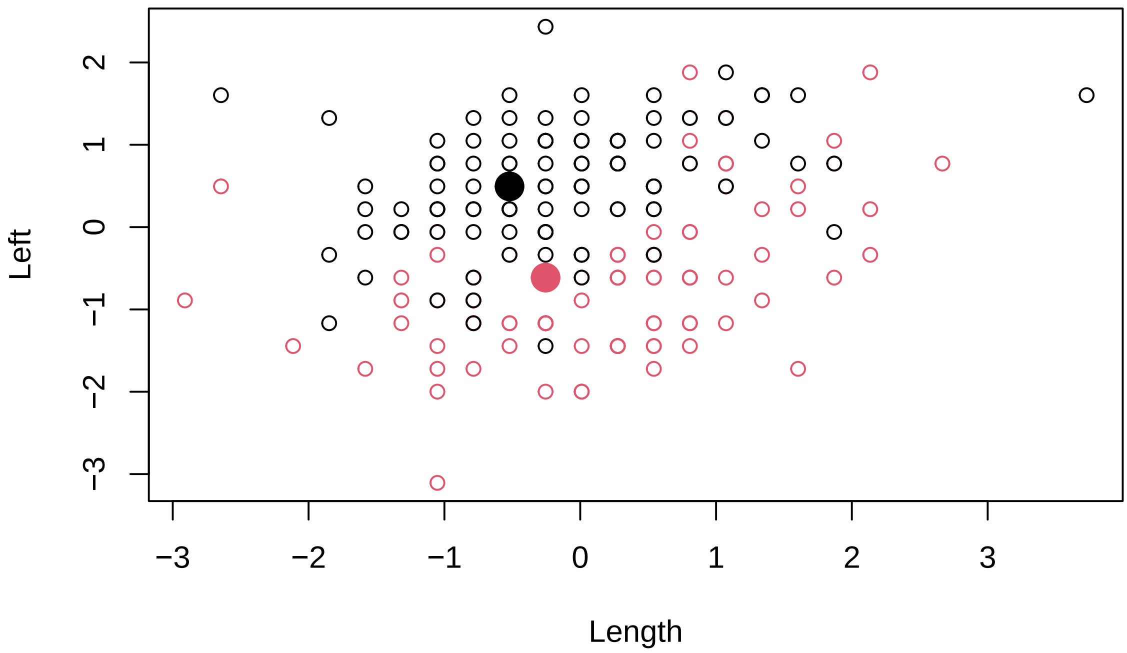
Aglomerativo:



Para um conjunto relativamente grande como esse, é praticamente impossível pela figura verificar onde está cada valor. Entretanto, ao verificar os dois grupos principais formados pelo dendograma e verificando os valores que foram agregados à eles, notamos que este foi extremamente eficiente em dividir as notas genuínas das falsificadas, com pouquíssimas observações sendo classificadas incorretamente.

Algorítmos não hierárquicos:

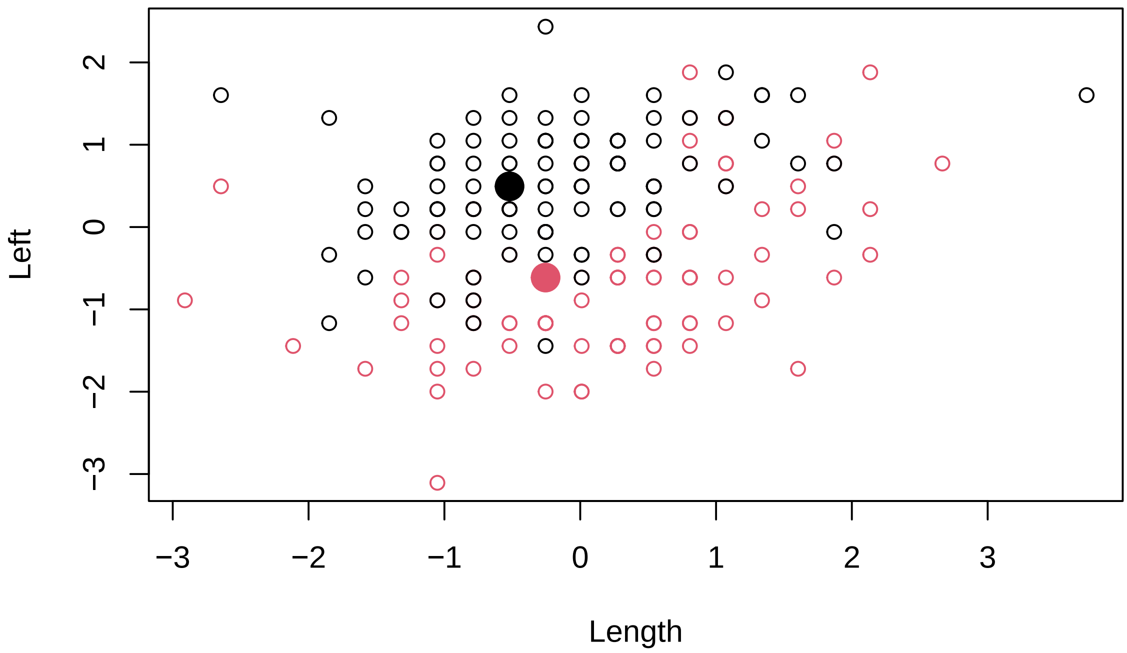
CLARA:



Var1	Freq
1	103
2	97

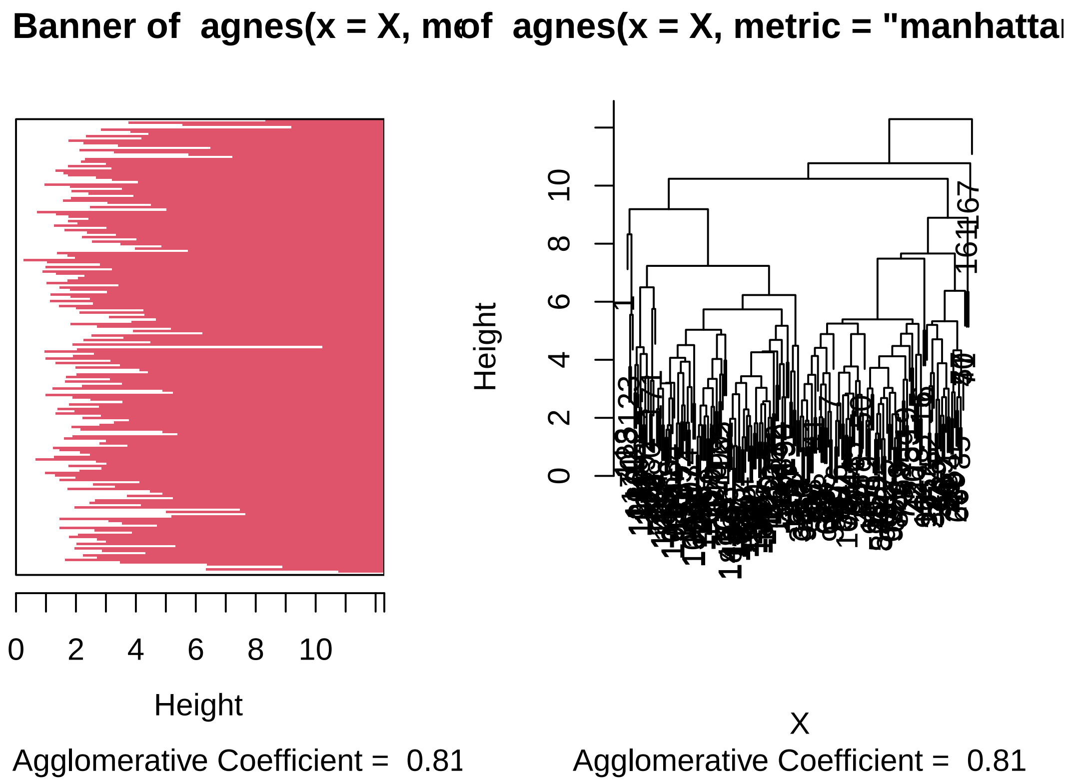
Notamos que CLARA agrupou apenas 3 valores errados, apontando 3 notas genuínas como falsificadas. Além disso, não apontou nenhuma falsificada como genuína.

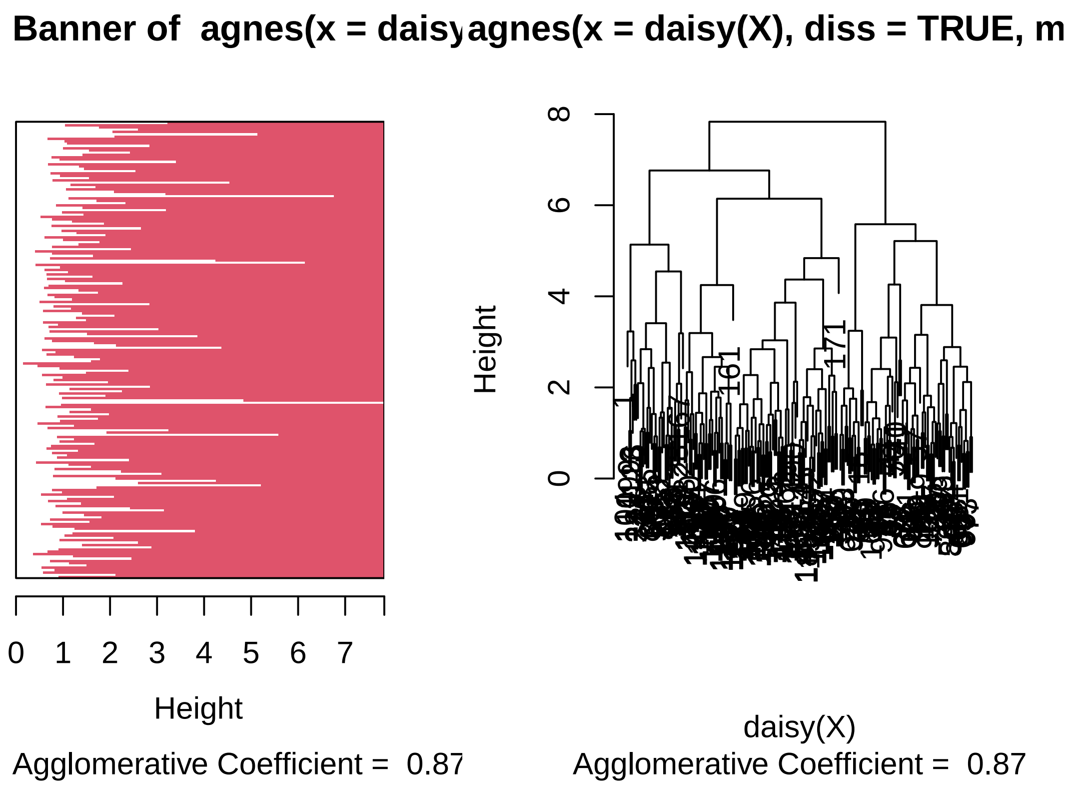
PAM:



Notamos que PAM também agrupou apenas 3 valores errados, também apontando 3 notas genuínas como falsificadas. Também não apontou nenhuma falsificada como genuína. O resultado foi idêntico ao retornado por PAM neste caso.

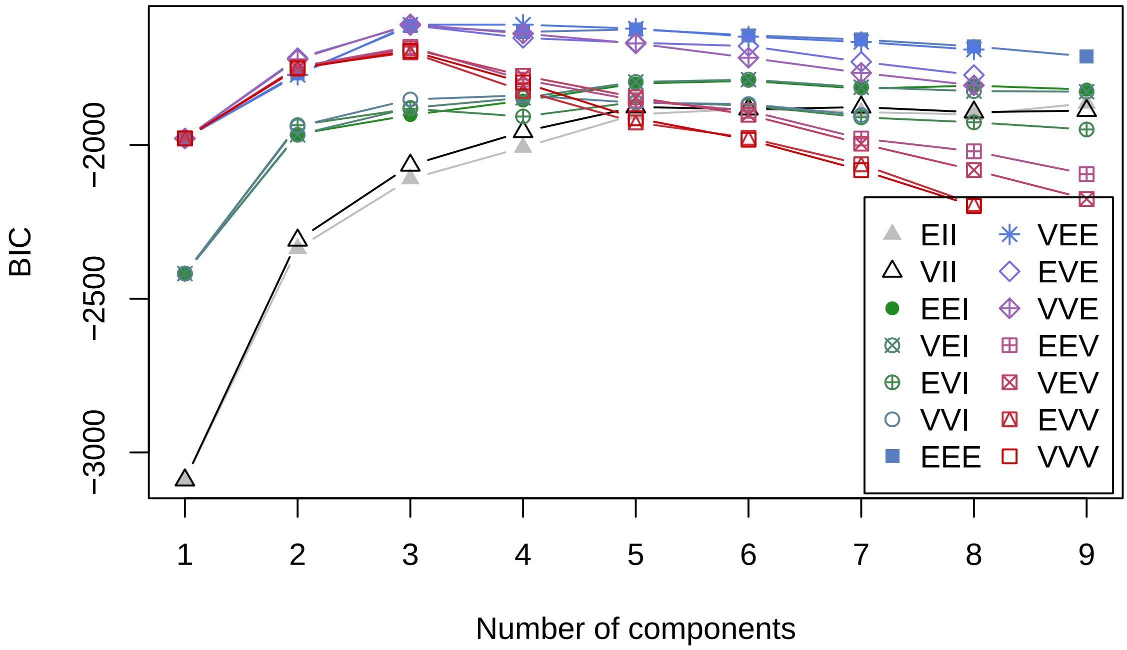
AGNES:


Banner of agnes(x = X, me of agnes(x = X, metric = "manhatta
Height
Height
X
Agglomerative Coefficient = 0.81
Agglomerative Coefficient = 0.81


Banner of agnes(x = daisy agnes(x = daisy(X), diss = TRUE, m
Height
Height
daisy(X)
Agglomerative Coefficient = 0.87
Agglomerative Coefficient = 0.87

Aqui, testamos tanto AGNES utilizando as distâncias de Manhattan com aglomeração simples no primeiro caso, e usando distâncias euclidianas com aglomeração completa no segundo caso. Em nenhum dos dois AGNES performou tão bem quanto CLARA e PAM para este conjunto de dados.

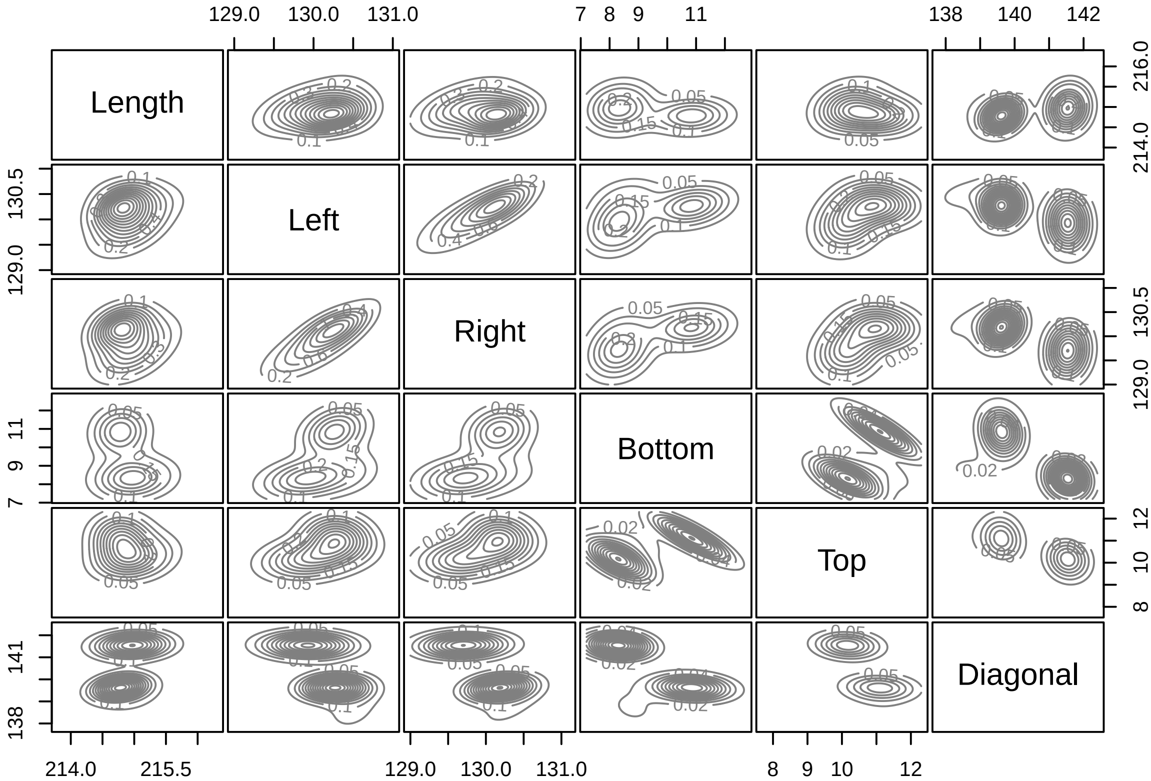
c)




```
## Best BIC values:
##                VVE,3        VEE,4        VEE,3
## BIC       -1607.574 -1608.767736 -1608.793746
## BIC diff      0.000    -1.194096    -1.220106
```

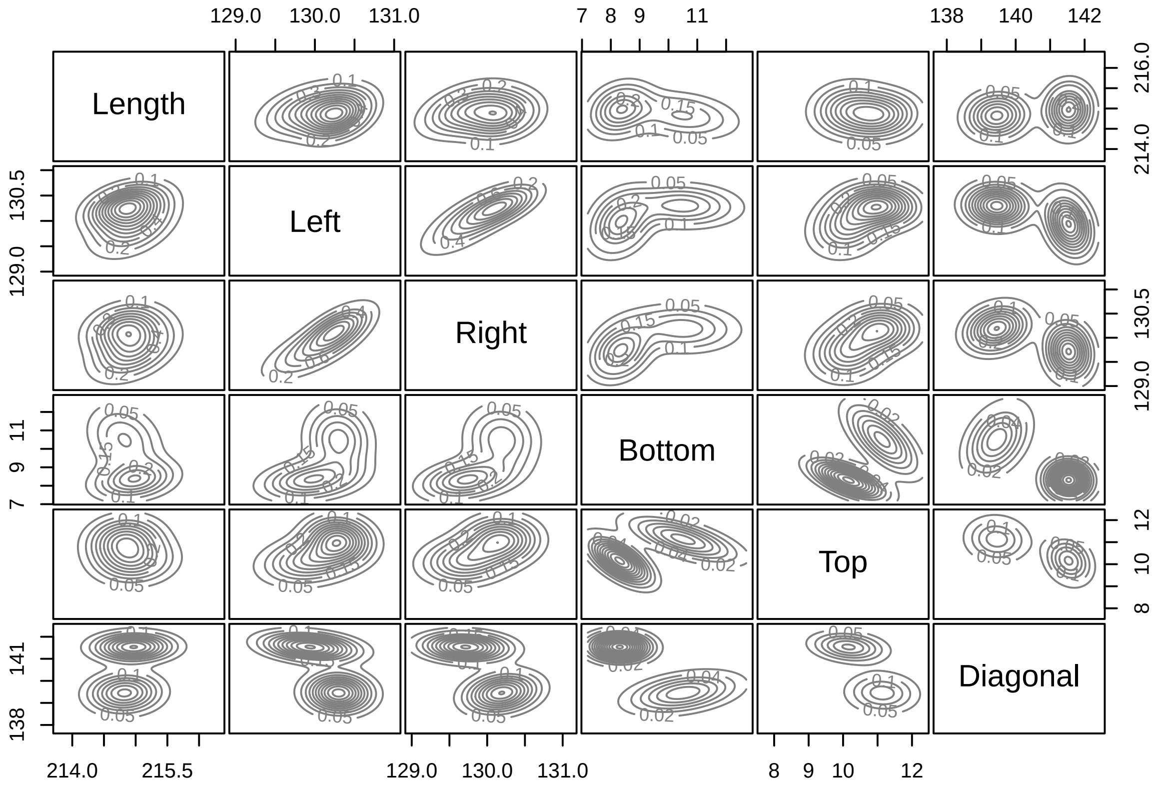
Para o método *mclust*, está indicando que o ideal seriam 3 agrupamentos, com o modelo *VVE*. Como sabemos que são apenas 2 grupos, temos que este método provavelmente não irá funcionar bem.

Seguindo a sugestão da BIC, iremos ajustar com o modelo *VVE* de 3 grupos



gsim	Freq
1	16
2	105
3	79

Percebemos que este foi o modelo que mais errou dos testados até agora. Apenas por fins didáticos, testarei o modelo mais 'complexo' VVV, forçando o número de *clusters* como igual à dois.



gsim	Freq
1	108
2	92

Percebemos que aqui, foi dissolvido o grupo 3 que possivelmente continham informações mais de "fronteira" entre os dois grupos mais sólidos, e estas foram diluídas entre os 2 grupos robustos existentes, com um dos grupos "ganhando" 3 itens, enquanto o outro ficando com o restante das 13 observações. Apesar do erro bruto não parecer tão grande, é um pouco decepcionante para um algoritmo tão robusto e pesado um resultado como este. Isso nos leva a suspeitar que as distribuições diferem bastante de uma normal multivariada, apesar deste não ser exatamente um pressuposto rígido deste modelo.

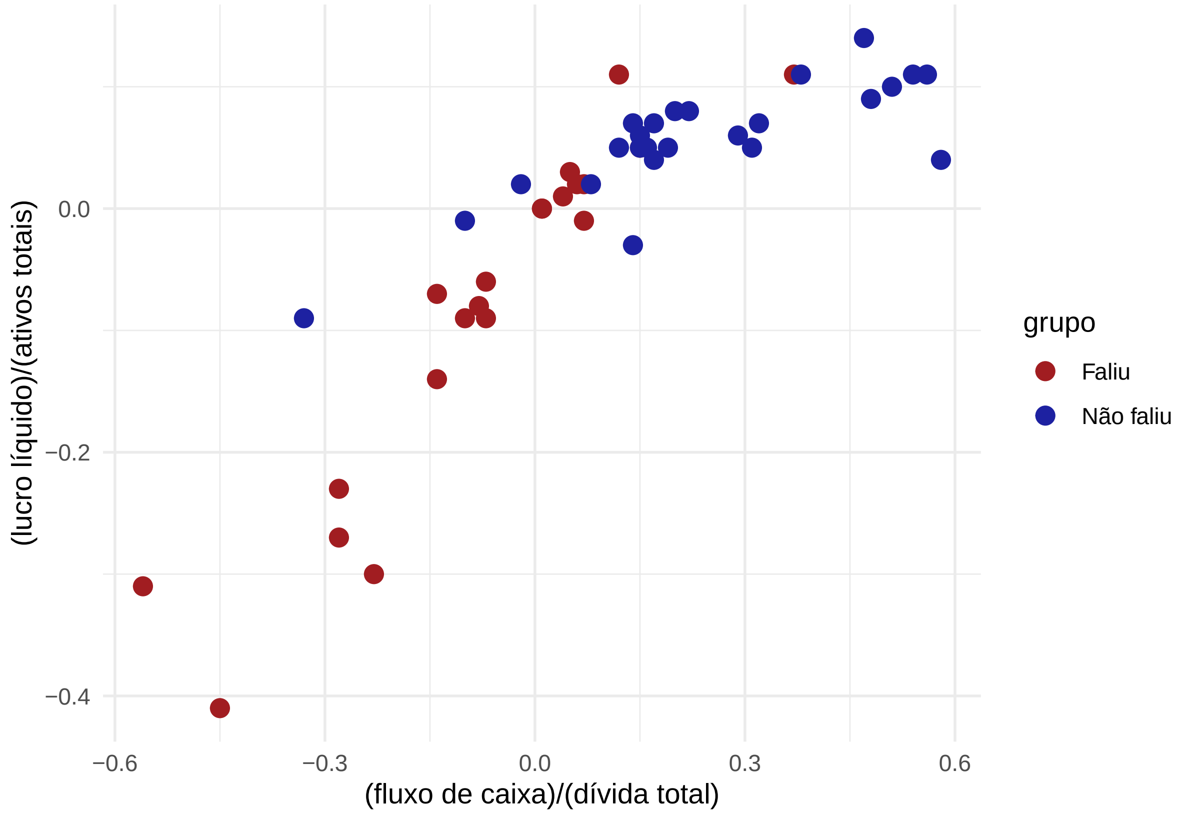
d)

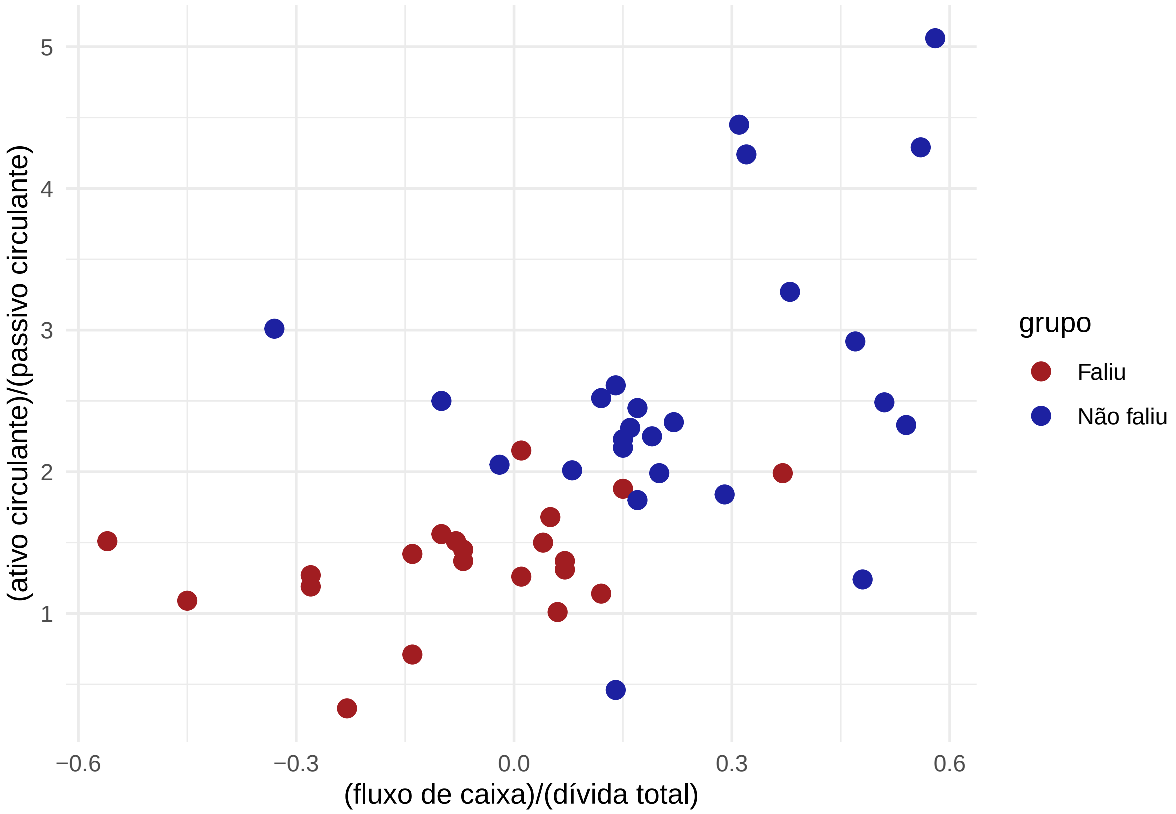
rand_kmeans	rand_clara	rand_pam	APER_mclustVVE3	APER_mclustVVV2	rand_mclustVVE3	rand_mclustVVV2
0.8456292	0.9406018	0.9406018	0.895	0.92	-0.0049758	-0.004616

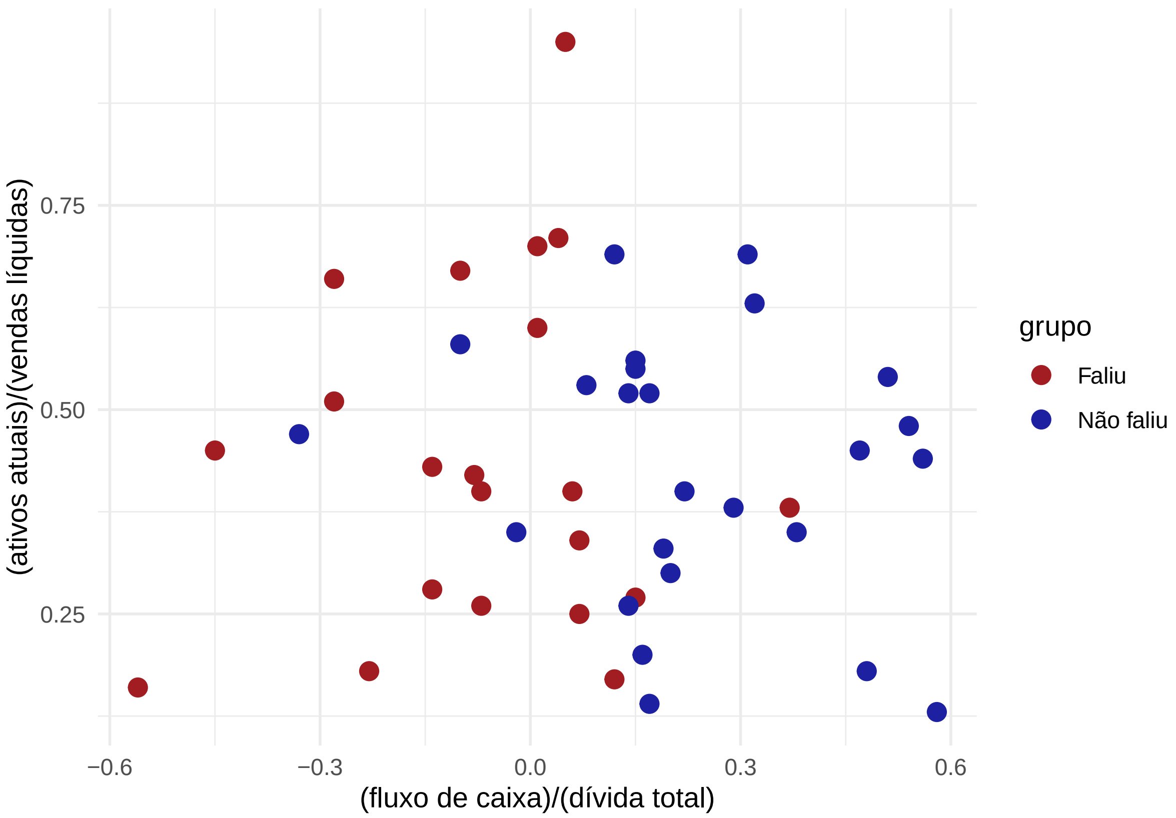
Aqui notamos que, de todos os algorítmos aqui testados, os que perfomaram melhor foram os (esquecidos e discriminados) CLARA e PAM.

89. Johnson e Wichern - Exercício 11.24.

a)







Em todos os gráficos, os pontos lembram a forma de elipsoides. Portanto, graficamente, não é possível rejeitar a normalidade bivariada dos dados.

b)

Considerando 1 como o grupo de empresas que faliram (falidos) e 2 como o grupo de empresas que não faliram ainda (ativos), temos os vetores de média μ_1', μ_2' dados respectivamente por: [-0.0690476, -0.0814286],[0.2352, 0.0556], e matrizes de covariância $S_1 =$

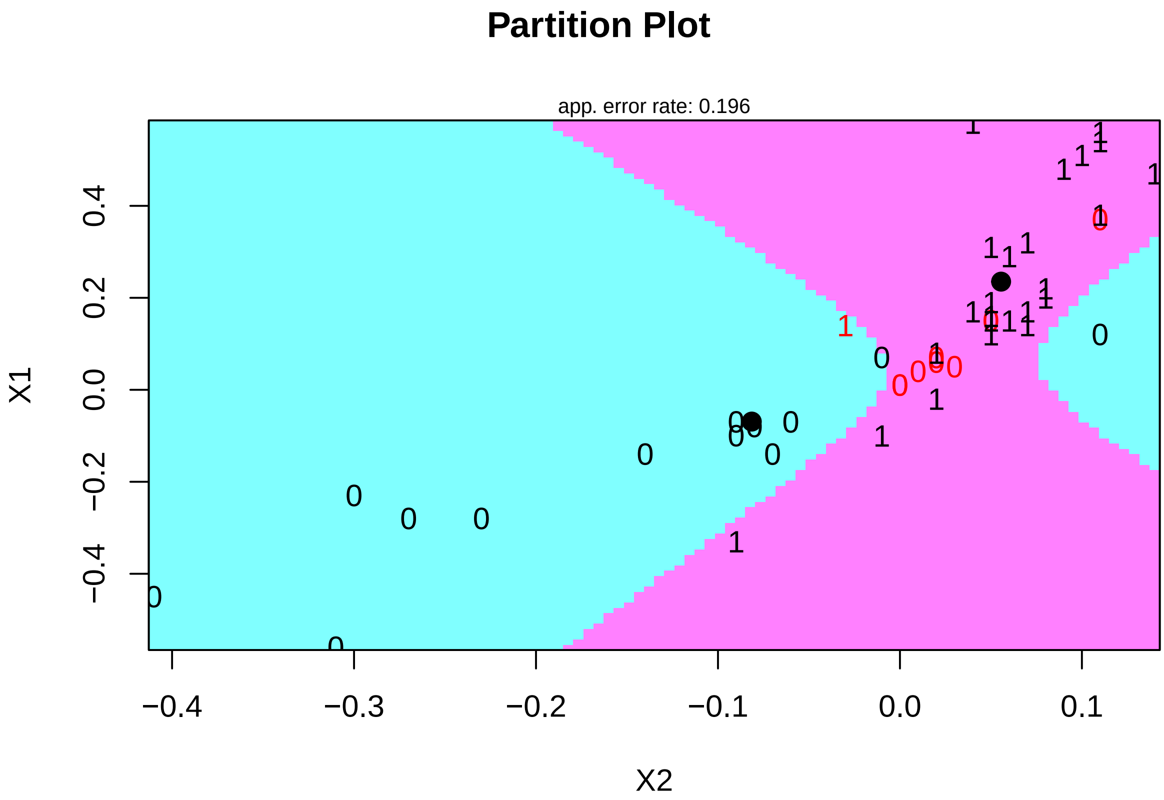
	X1	X2
X1	0.0441290	0.0284764
X2	0.0284764	0.0210029

e $S_2 =$

	X1	X2
X1	0.0470510	0.0085072
X2	0.0085072	0.0023757

c)

Como para este conjunto não rejeitamos a hipótese de normalidade multivariada (apesar de termos feito apenas análise gráfica), e, apesar de não termos testado a igualdade das variâncias, elas aparentam ser diferentes; portanto a abordagem mais adequada para este caso é a análise discriminante quadrática abaixo. No caso, foram definido custos e prioris iguais para ambos os grupos.



Matriz de confusão:

	0	1
0	13	8
1	1	24

Proporção de classificações corretas em cada grupo:

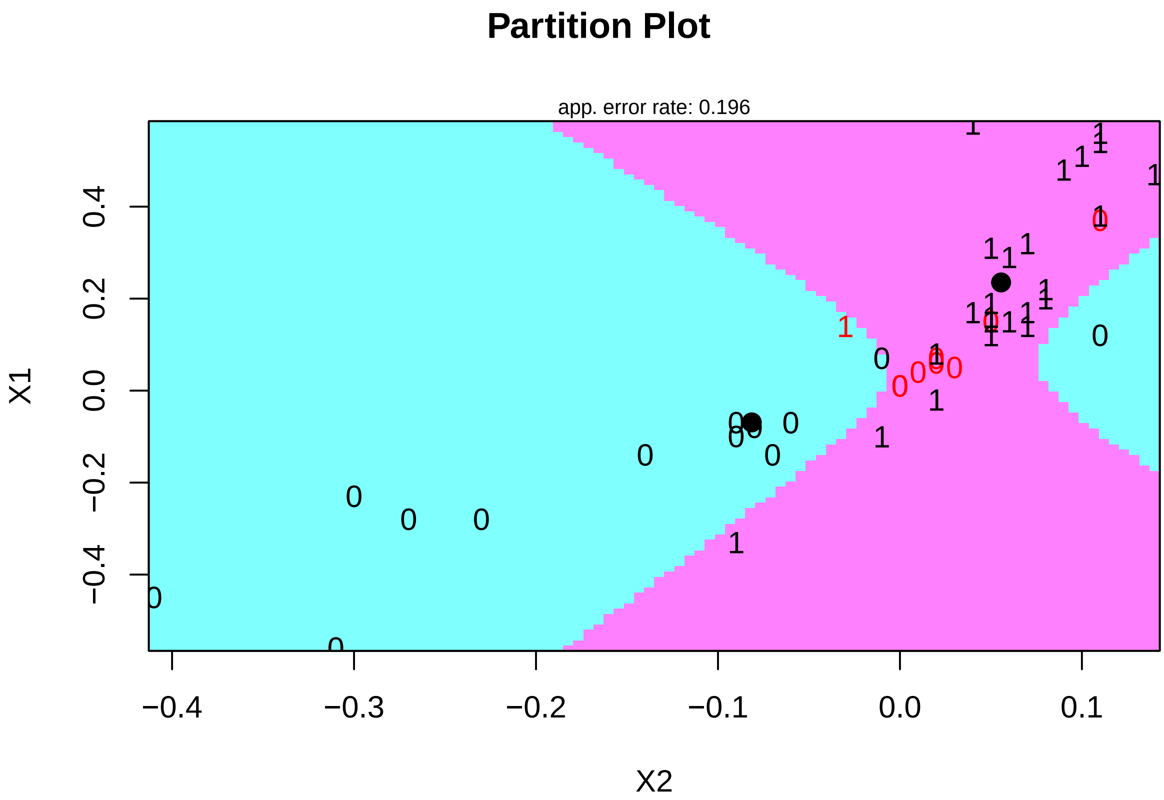
	x
0	0.6190476
1	0.9600000

Proporção total de classificação correta: 0.8043478

d)

O erro aparente (APER) deste conjunto foi calculado como sendo 0.1956522; enquanto que a estimação da taxa de erro aparente ($\hat{E}(AER)$) foi calculada como 0.2173913. Notamos que apesar de o erro estimado via validação cruzada Jackknife ter sido maior que o erro aparente, esta é uma estimativa mais robusta em comparação com o resultado sem validação cruzada.

e)



Matriz de confusão:

	0	1
0	9	12
1	0	25

Proporção de classificações corretas em cada grupo:

	x
0	0.4285714
1	1.0000000

Proporção total de classificação correta: 0.7391304

O erro aparente (APER):0.2608696

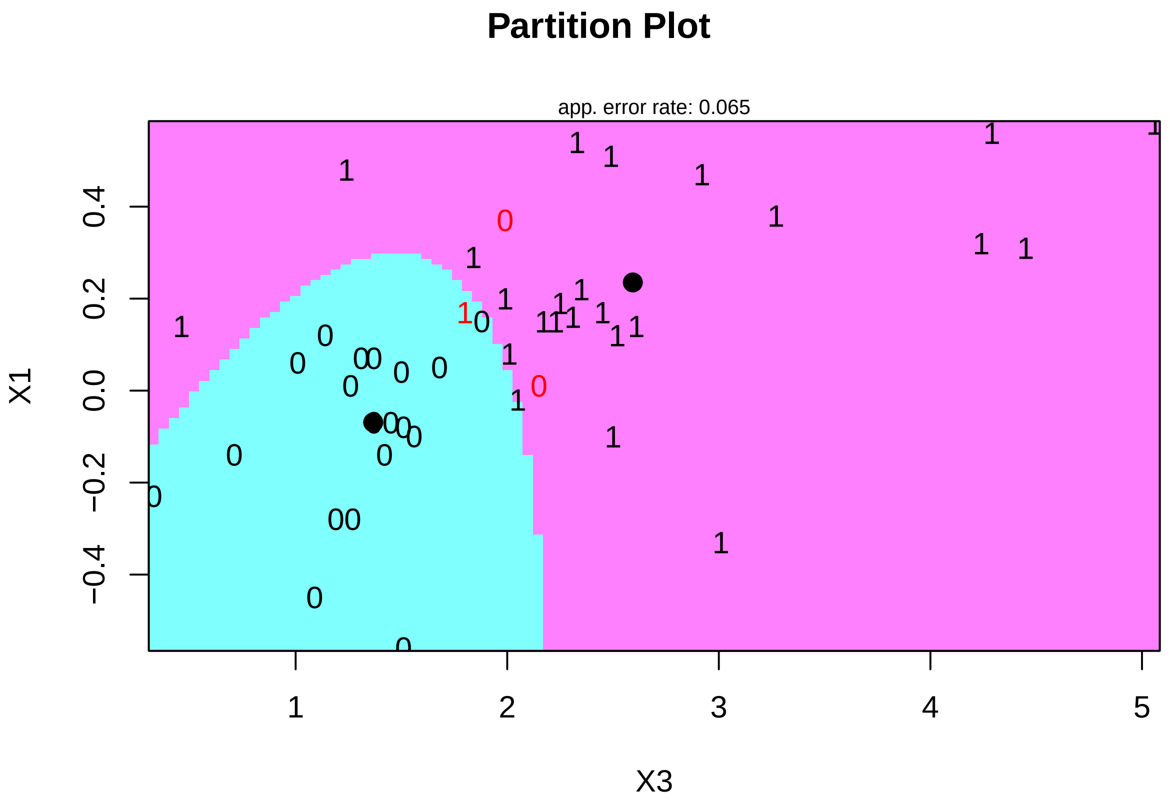
Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.2608696

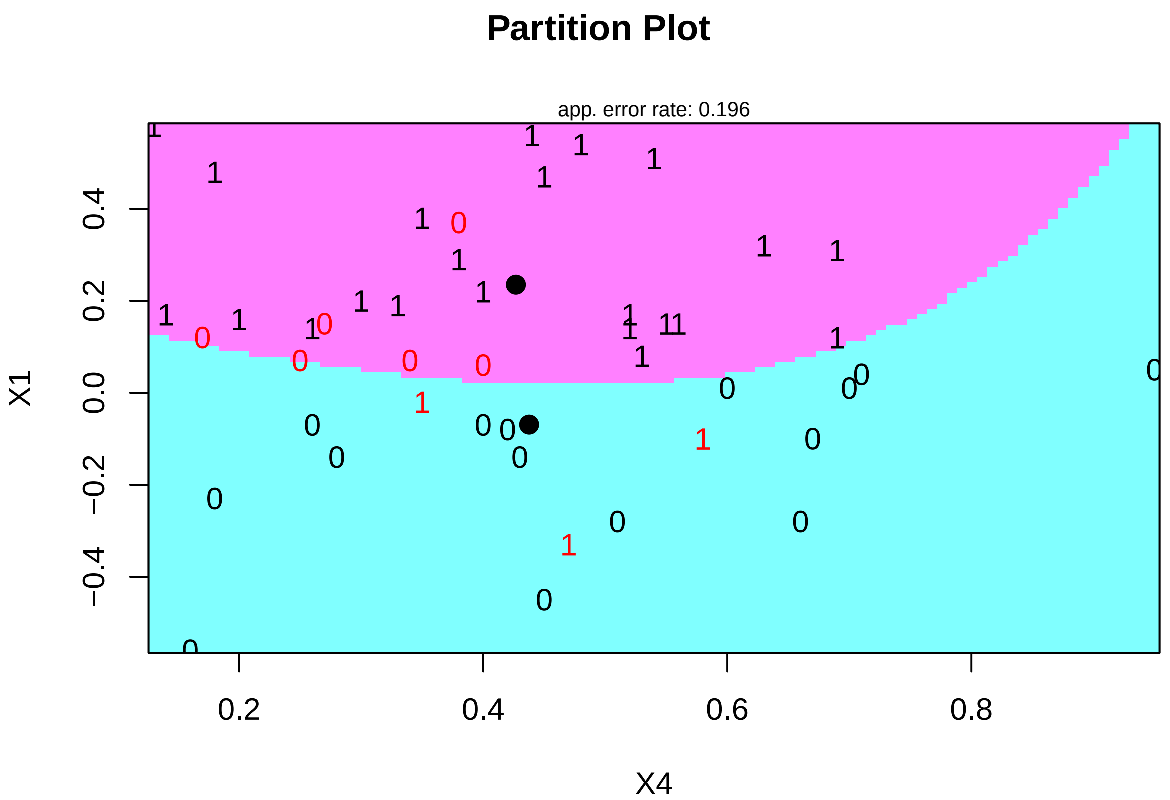
Analisando os APER e $\hat{E}(AER)$, concluímos que as prioris iguais ($p_1 = 0,5; p_2 = 0,5$) tem um erro de classificação inferior se comparado as prioris desiguais ($p_1 = 0,05; p_2 = 0,95$). Neste caso, notamos que tanto o APER quanto o $\hat{E}(AER)$ deram resultados idênticos.

f)

Como as matrizes S_1 e S_2 aparentam ser diferentes, esta técnica não aparenta ser a mais adequada. Entretanto, tomando como base apenas a performance do APER = 0.173913, até que a classificação por discriminantes lineares não ficou ruim, com resultados até melhores do que os obtido pelos discriminantes quadráticos.

g)


Partition Plot
app. error rate: 0.065
X1
X3


Partition Plot
app. error rate: 0.196
X1
X4

Vetores de média e matrizes de covariância para as variáveis (x1,x3):

Vetor média μ_1' = -0.0690476, 1.3666667

Vetor Média μ_3' = 0.2352, 2.5936

Matriz de covariância S_1 =

	X1	X3
X1	0.0441290	0.0344933
X3	0.0344933	0.1643033

Matriz de covariância S_3 =

	X1	X3
X1	0.0470510	0.0749305
X3	0.0749305	1.0467740

Análise discriminante quadrática, com prioris = (0,5;0,5), utilizando as variáveis (x1,x3):

Matriz de confusão:

	0	1
0	19	2
1	3	22

Proporção de classificações corretas em cada grupo:

	x
0	0.9047619
1	0.8800000

Proporção total de classificação correta: 0.8913043

Erro aparente (APER):0.1086957

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.1304348

Análise discriminante quadrática, com prioris = (0,05;0,95), utilizando as variáveis (x1,x3):

Matriz de confusão:

	0	1
0	4	17
1	0	25

Proporção de classificações corretas em cada grupo:

	x
0	0.1904762
1	1.0000000

Proporção total de classificação correta: 0.6304348

Erro aparente (APER):0.3695652

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.3913043

Vetores de média e matrizes de covariância para as variáveis (x1,x4):

Vetor média μ_1' = -0.0690476, 0.437619

Vetor Média μ_3' = 0.2352, 0.4268

Matriz de covariância S_1 =

	X1	X4
X1	0.0441290	0.0041474
X4	0.0041474	0.0445790

Matriz de covariância S_3 =

	X1	X4
X1	0.0470510	-0.0067035
X4	-0.0067035	0.0263810

Análise discriminante quadrática, com prioris = (0,5;0,5), utilizando as variáveis (x1,x4):

Matriz de confusão:

	0	1
0	17	4
1	4	21

Proporção de classificações corretas em cada grupo:

	x
0	0.8095238
1	0.8400000

Proporção total de classificação correta: 0.826087

Erro aparente (APER):0.173913

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.2173913

Análise discriminante quadrática, com prioris = (0,05;0,95), utilizando as variáveis (x1,x4):

Matriz de confusão:

	0	1
0	3	18
1	0	25

Proporção de classificações corretas em cada grupo:

	x
0	0.1428571
1	1.0000000

Proporção total de classificação correta: 0.6086957

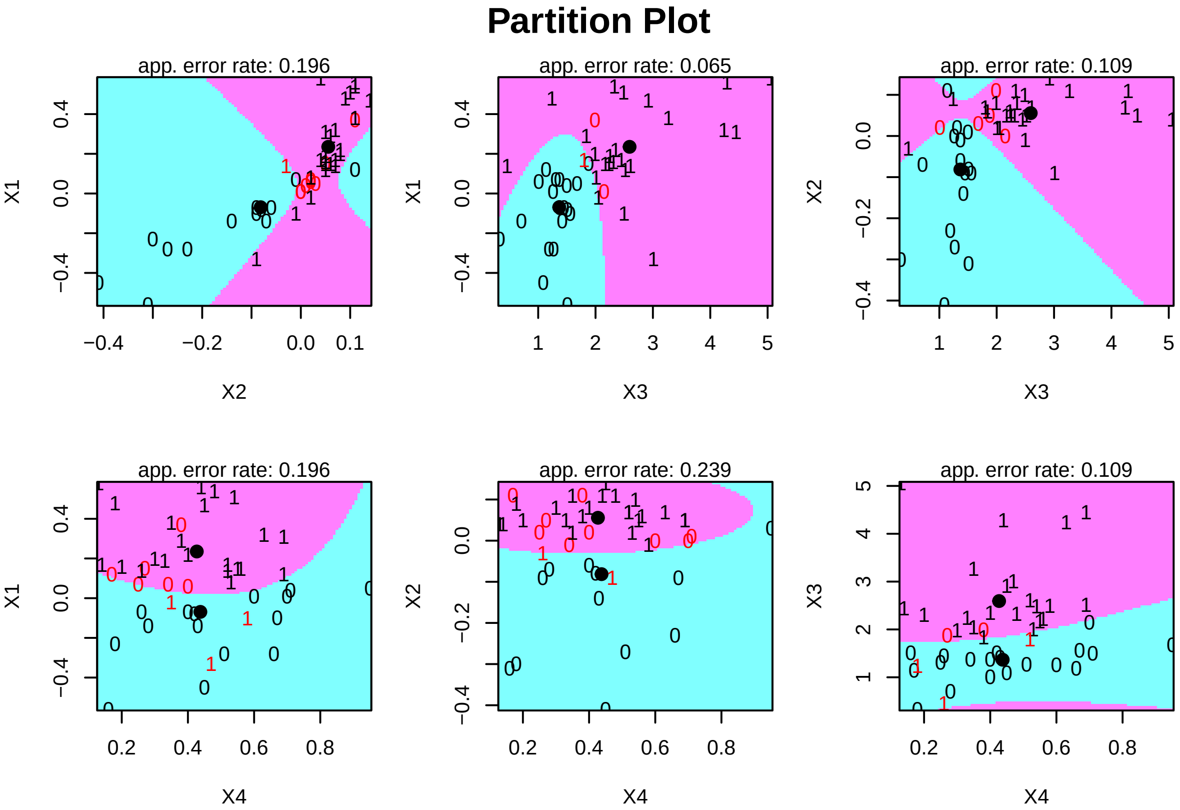
Erro aparente (APER):0.3913043

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.4565217

Conclusões:

De fato, os resultados encontrados foram bastante distintos para cada caso. Analisando somente os APER e $\hat{E}(AER)$, notamos que a análise em que foi observado o menor valor de ambos foi a análise executada utilizando as variáveis (x_1, x_3), com prioris iguais $(0,5; 0,5)$, enquanto que os maiores valores foram observados para o modelo em que utilizei as variáveis (x_1, x_4) com prioris desiguais $(0,05; 0,95)$. O modelo que menos variou estas duas estatísticas para ambas as prioris testadas $(0,5; 0,5)$ e $(0,05; 0,95)$ foi o modelo inicialmente testado com as variáveis (x_1, x_2). Com base nisso, podemos concluir que tanto a escolha das variáveis quanto a escolha das prioris, influenciam bastante na qualidade do modelo final.

h)



Vetores de média e matrizes de covariância para as variáveis (x1,x2,x3,x4):

Vetor média $\mu_1' =$ -0.0690476, -0.0814286, 1.3666667, 0.437619

Vetor Média $\mu_3' =$ 0.2352, 0.0556, 2.5936, 0.4268

Matriz de covariância $S_1 =$

	X1	X2	X3	X4
X1	0.0441290	0.0284764	0.0344933	0.0041474
X2	0.0284764	0.0210029	0.0260200	0.0034414
X3	0.0344933	0.0260200	0.1643033	0.0327817
X4	0.0041474	0.0034414	0.0327817	0.0445790

Matriz de covariância $S_3 =$

	X1	X2	X3	X4
X1	0.0470510	0.0085072	0.0749305	-0.0067035
X2	0.0085072	0.0023757	0.0085832	0.0001853
X3	0.0749305	0.0085832	1.0467740	0.0326328
X4	-0.0067035	0.0001853	0.0326328	0.0263810

Análise discriminante quadrática, com prioris = (0,5;0,5), utilizando as variáveis (x1,x2,x3,x4):

Matriz de confusão:

	0	1
0	19	2
1	1	24

Proporção de classificações corretas em cada grupo:

	x
0	0.9047619
1	0.9600000

Proporção total de classificação correta: 0.9347826

Erro aparente (APER):0.0652174

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.1086957

Análise discriminante quadrática, com prioris = (0,05;0,95), utilizando as variáveis (x1,x2,x3,x4):

Matriz de confusão:

	0	1
0	12	9
1	0	25

Proporção de classificações corretas em cada grupo:

	x
0	0.5714286
1	1.0000000

Proporção total de classificação correta: 0.8043478

Erro aparente (APER):0.1956522

Estimativa da taxa de erro aparente ($\hat{E}(AER)$): 0.2391304

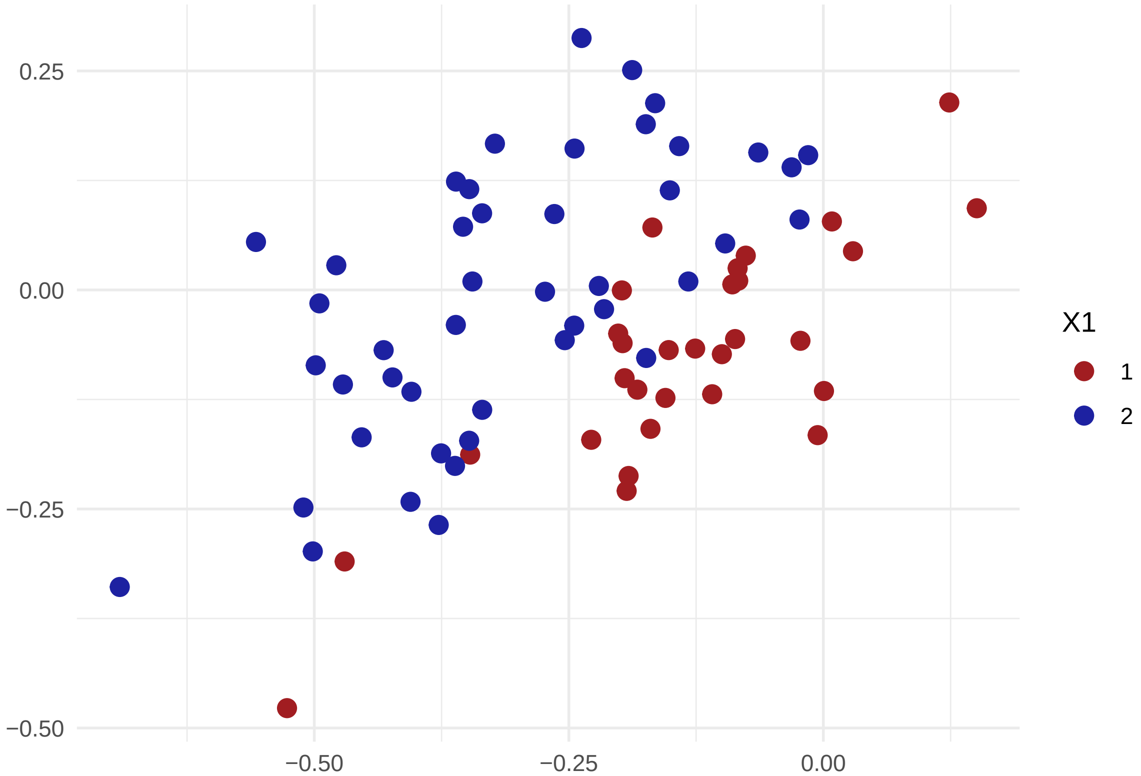
No caso da inclusão de todas as 4 variáveis, o classificador com prioris iguais produziu as melhores classificações (menores APER e $\hat{E}(AER)$). Também neste caso, o classificador com prioris $(0,05;0,95)$ produziu um APER significativamente maior que o mesmo modelo com prioris iguais, porém foram os menores valores se comparados com os valores observados nos demais modelos com prioris $(0,05;0,95)$.

Conclusões:

Isto nos leva a acreditar que a inclusão de mais variáveis foi bom para o modelo, produzindo os menores erros aparentes. Entretando, a diferença não foi tão substantiva assim, então, deve-se considerar questões como verba para coleta de tantas variáveis, complexidade da análise e viabilidade de novas coletas caso deseje-se seguir com o modelo mais preciso.

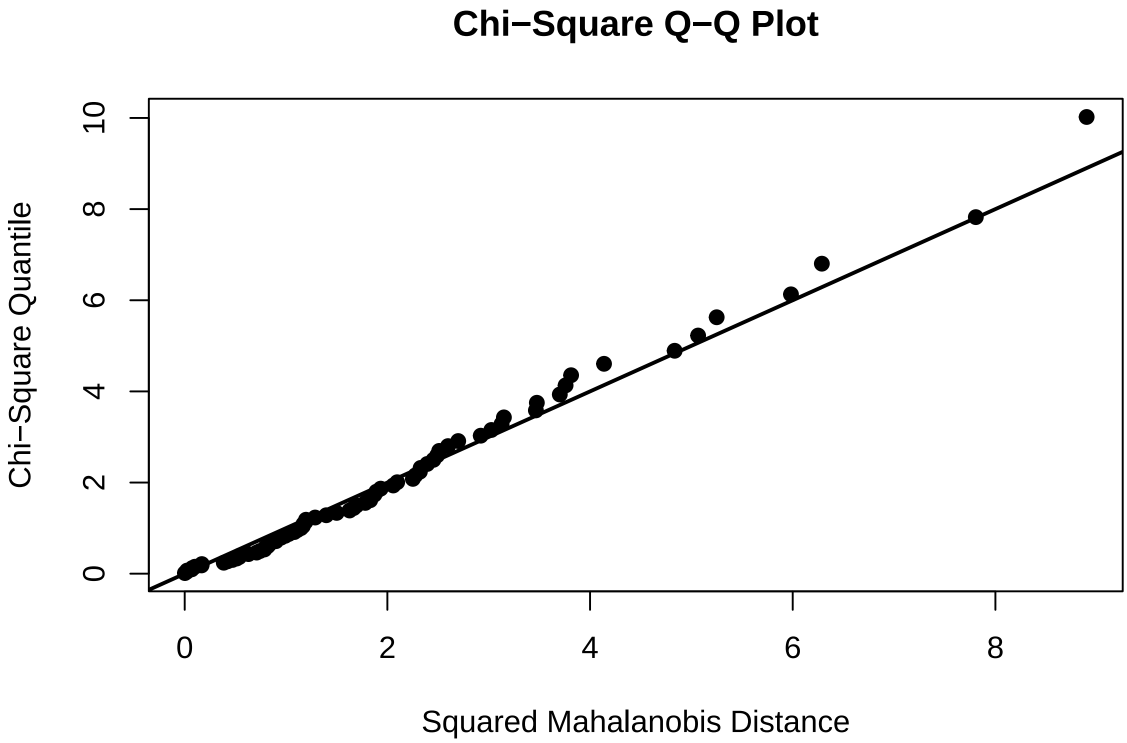
90. Johnson e Wichern - Exercício 11.32.

a)




```
##
##   Shapiro-Wilk normality test
##
## data:  dados$X2
## W = 0.98496, p-value = 0.5185
```

```
##
##   Shapiro-Wilk normality test
##
## data:  dados$X3
## W = 0.99255, p-value = 0.9428
```




```
##              Anderson-Darling test for Multivariate Normality
##
##    data : dados[, 2:3]
##
##    AD              : 0.7583493
##    p-value         : 0.2345765
##
##    Result  : Data are multivariate normal (sig.level = 0.05)
```

Através da análise visual, não é possível rejeitar a normalidade bivariada, visto que os pontos aparentam formar uma elipsoide. Foi realizado ainda testes de Shapiro-Wilk nas duas marginais, que também não rejeitaram a normalidade; univariada, neste caso. Foi ainda utilizado o teste de Anderson-Darling para normalidade multivariada do pacote `mvnTest`, que também não rejeitou a normalidade multivariada. Portanto, não temos evidências para rejeitar a hipótese de normalidade multivariada dos dados.

b)

Matriz de confusão:

FALSE	TRUE
20	1
12	19

Proporção de classificações corretas em cada grupo:

x
0.9523810
0.6129032

Proporção total de classificação correta: 0.75

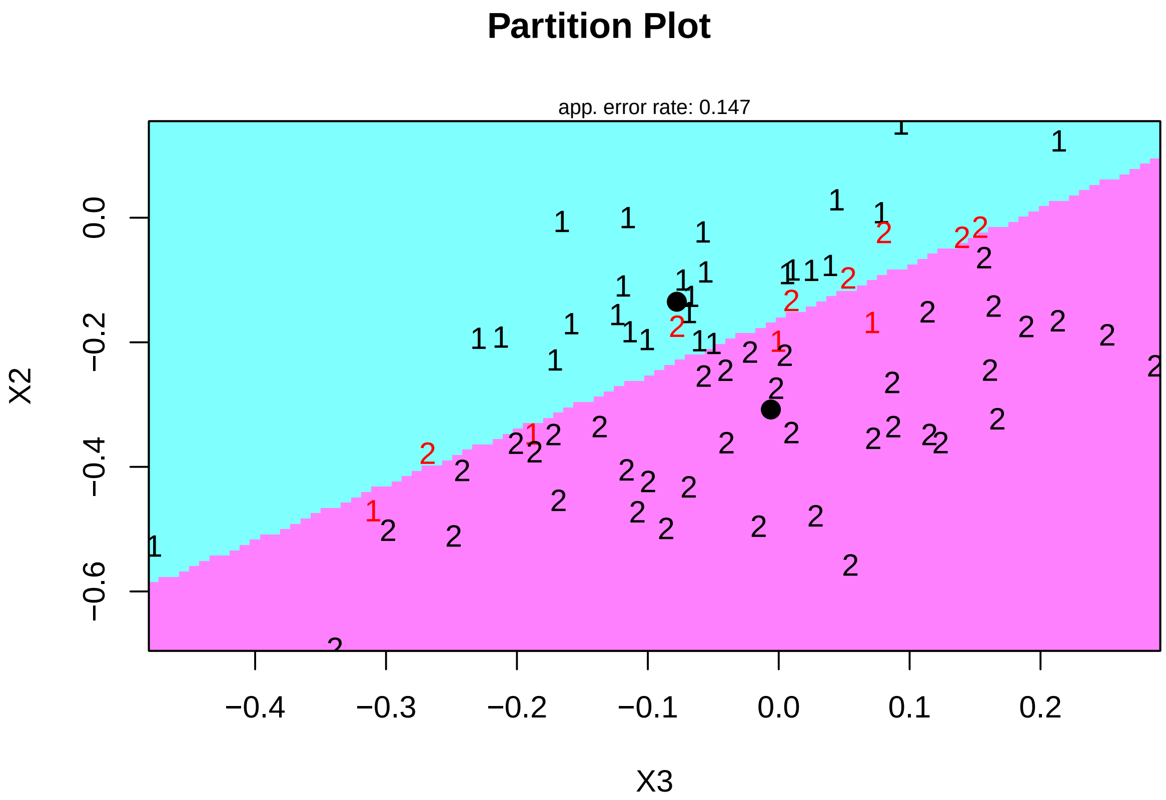
Com isso, temos que a taxa de erro do modelo pontual é de 25%. Esta é relativamente maior do que a encontrada pelos outros métodos de validação utilizados até agora.

c)

pop3	Freq
p1	10

Todas as 10 novas observações foram classificadas como percentence à população π_1

d)



Matriz de confusão:

FALSE	TRUE
19	2
5	26

Proporção de classificações corretas em cada grupo:

x
0.9047619
0.8387097

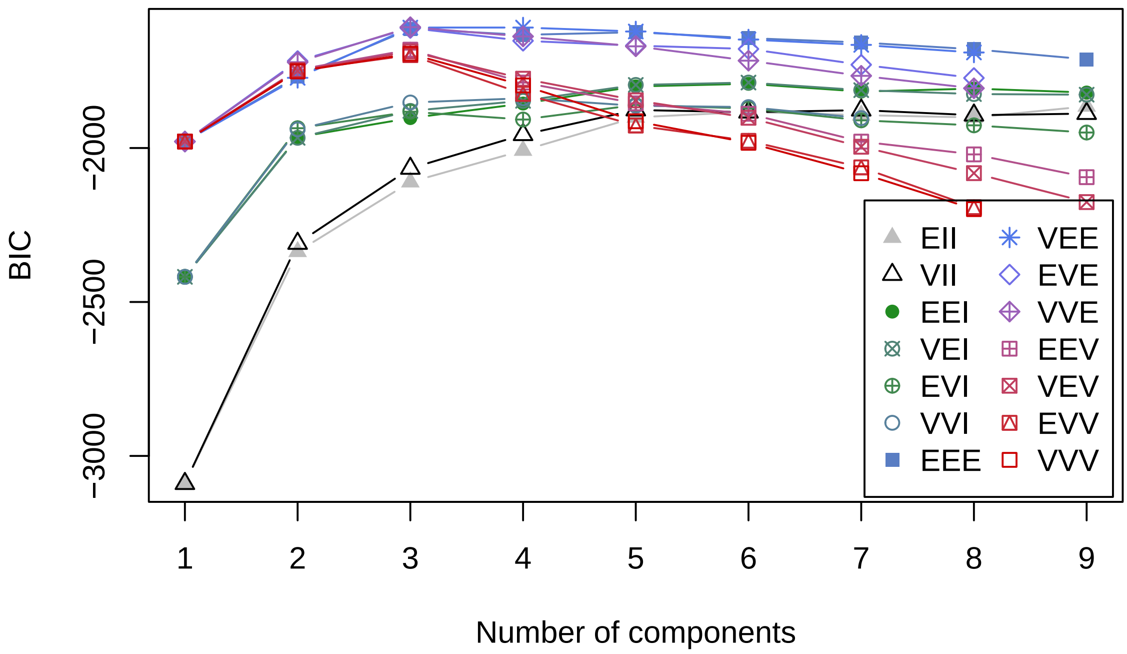
Proporção total de classificação correta: 0.8653846

Com isso, temos que a taxa de erro do modelo pontual é de 13.4615385%. Percebemos que a taxa de erro caiu consideravelmente ao escolher esta outra priori.

Além disso, todas as 10 novas observações foram novamente classificadas como pertencente à população π_1. Este é um resultado que não impressiona, visto que já haviam sido classificados assim com a priori igual, então era de se esperar que confirmasse este resultado com uma priori maior para a população 1.

91.

```
## Best BIC values:
##              VVE,3        VEE,4        VEE,3
## BIC      -1607.574 -1608.767736 -1608.793746
## BIC diff     0.000    -1.194096    -1.220106
```




```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVE (ellipsoidal, equal orientation) model with 3 components:
##
##  log-likelihood   n df       BIC      ICL
##       -663.3814 200 53 -1607.574 -1607.71
##
## Clustering table:
##  1  2  3
## 18 98 84

##
## Class          1  2  3
##   Genuína      2 98  0
##   Falsificada 16  0 84
```

A mistura de normais não operou tão bem quanto os discriminantes lineares e quadráticos. Enquanto nesses dois, 199 das 200 notas foram classificadas corretamente, o algorítmo de mistura de normais encontrou m=3 como o número ideal de clusters (sendo que neste caso sabemos que há apenas dois: genuínas e falsificadas). Com isso, classificou corretamente 182 das 200 notas. Interessante notar que não houve classificação de notas falsas como notas genuínas ou vice-versa; e sim algumas notas desses dois grupos foram classificadas em outro cluster, que seria talvez um cluster de "confusão", ou seja, notas em que não estava claro o suficiente se eram genuínas ou classificadas.

Erro aparente (APER) do modelo de discriminante linear: 0.005

Erro aparente (APER) do modelo de discriminante quadrático: 0.005

Erro aparente (APER) do modelo de mistura de normais: 0.09

Índice de Rand ajustado do modelo de mistura de normais: 0.8418856