INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2020-21)

Teste de Estatística Multivariada

11 de Junho, 2021 Duração: 2h30

I [16 valores]

Foi efectuado um estudo da composição química de 64 colunas de gelo extraídas em diferentes localidades da região Ártica[1], que deu origem a 64 conjuntos de medições de 16 elementos ou compostos (`Al`, `Ti`, `Fe`, `Mn`, `Ca`, `Mg`, `Na`, `K`, `P` - todas em peso (%) - e `Ba`, `Sr`, `Cr`, `Ni`, `Sc`, `V` e `Zr` - em partes por milhão). Indicadores de síntese, bem como a matriz de correlações deste conjunto de dados são indicadas em baixo.

	Min.	1st Qu.	Mean	3rd Qu.	Max.	St.Dev
Al	2.260	6.250	7.223	8.805	10.450	2.349
Ti	0.098	0.372	0.429	0.522	0.712	0.161
Fe	2.390	4.760	5.419	6.168	9.050	1.302
Mn	0.012	0.036	0.161	0.154	1.097	0.243
Ca	0.190	0.308	0.449	0.460	4.040	0.469
Mg	0.400	0.785	0.906	1.040	1.710	0.243
Na	0.370	1.315	1.450	1.693	1.890	0.349
K	0.450	1.070	1.679	2.055	2.910	0.614
P	0.021	0.045	0.080	0.100	0.239	0.040
Ba	78.000	485.000	632.484	693.000	2506.000	456.340
Sr	46.000	89.000	129.844	161.250	231.000	46.641
Cr	18.000	70.750	83.203	101.000	195.000	31.232
Ni	23.800	40.625	51.230	51.400	191.900	25.130
Sc	4.700	12.350	14.833	16.925	38.300	4.739
V	56.000	161.250	195.984	234.250	291.000	52.937
Zr	43.000	131.250	154.953	184.250	264.000	51.885

```
> round(cor(arctic[,3:18]),d=2)
      Al    Ti    Fe    Mn    Ca   Mg    Na     K     P    Ba    Sr   Cr    Ni   Sc     V    Zr
Al  1.00  0.93 -0.03  0.25  0.05 0.71  0.41  0.94  0.38  0.57  0.83 0.75  0.32 0.73  0.70  0.91
Ti  0.93  1.00 -0.11  0.19  0.07 0.68  0.40  0.89  0.38  0.60  0.76 0.76  0.26 0.62  0.64  0.97
Fe -0.03 -0.11  1.00  0.16 -0.02 0.18 -0.16 -0.13  0.22 -0.16 -0.08 0.02  0.11 0.07 -0.08 -0.12
Mn  0.25  0.19  0.16  1.00  0.14 0.30  0.25  0.17  0.44 -0.01  0.36 0.10  0.11 0.12  0.14  0.18
Ca  0.05  0.07 -0.02  0.14  1.00 0.43  0.16  0.12  0.13 -0.06  0.24 0.17  0.14 0.02 -0.08  0.03
Mg  0.71  0.68  0.18  0.30  0.43 1.00  0.59  0.71  0.33  0.27  0.76 0.63  0.28 0.41  0.46  0.58
Na  0.41  0.40 -0.16  0.25  0.16 0.59  1.00  0.44  0.32  0.02  0.63 0.40  0.05 0.17  0.40  0.36
K   0.94  0.89 -0.13  0.17  0.12 0.71  0.44  1.00  0.37  0.49  0.83 0.69  0.30 0.64  0.68  0.87
P   0.38  0.38  0.22  0.44  0.13 0.33  0.32  0.37  1.00 -0.02  0.38 0.25 -0.04 0.16  0.12  0.39
Ba  0.57  0.60 -0.16 -0.01 -0.06 0.27  0.02  0.49 -0.02  1.00  0.47 0.40  0.47 0.69  0.36  0.65
Sr  0.83  0.76 -0.08  0.36  0.24 0.76  0.63  0.83  0.38  0.47  1.00 0.65  0.45 0.67  0.64  0.74
Cr  0.75  0.76  0.02  0.10  0.17 0.63  0.40  0.69  0.25  0.40  0.65 1.00  0.23 0.53  0.52  0.73
Ni  0.32  0.26  0.11  0.11  0.14 0.28  0.05  0.30 -0.04  0.47  0.45 0.23  1.00 0.77  0.38  0.34
Sc  0.73  0.62  0.07  0.12  0.02 0.41  0.17  0.64  0.16  0.69  0.67 0.53  0.77 1.00  0.65  0.71
V   0.70  0.64 -0.08  0.14 -0.08 0.46  0.40  0.68  0.12  0.36  0.64 0.52  0.38 0.65  1.00  0.65
Zr  0.91  0.97 -0.12  0.18  0.03 0.58  0.36  0.87  0.39  0.65  0.74 0.73  0.34 0.71  0.65  1.00
```

1. Uma análise preliminar dos dados foi efectuada com uma Análise em Componentes Principais sobre a *matriz de correlações*. Eis alguns dos resultados obtidos:

```
> summary(arctic.acp1)
Importance of components:
```

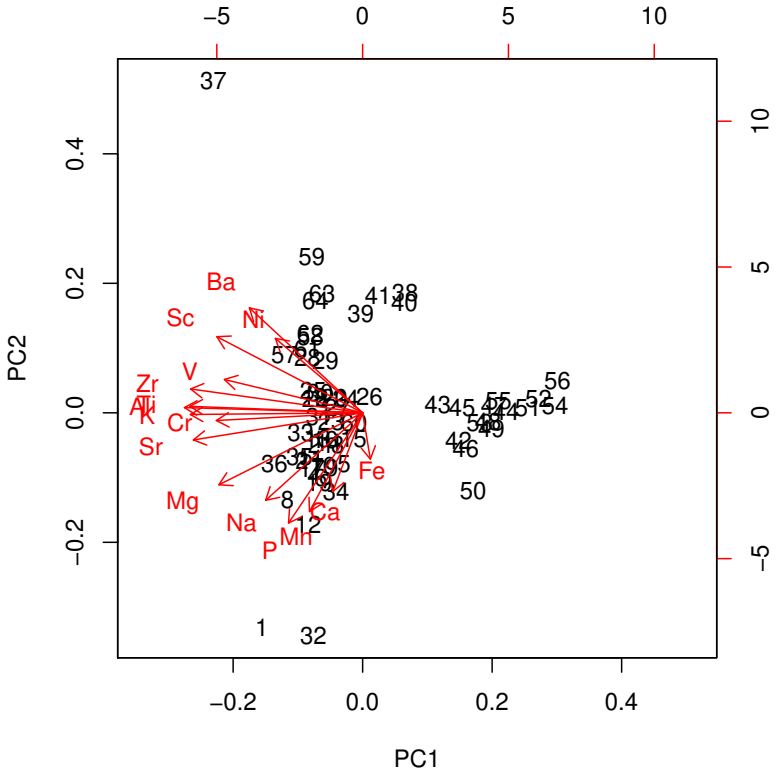
[1] N.C. Martinez, R.W. Murray, G.R. Dickens, and M. Molling (2009). *Discrimination of Sources of Terrigenous Sediment Deposited in the Central Arctic Ocean Through the Cenozoic*, Paleoceonography, Vol. 24, PA1210, doi:10.1029/2007PA001567, 2009

```
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9  [...]
Standard deviation     2.7899 1.3972 1.17937 1.07633 0.94507 0.89816 0.71258 0.67580 0.58056
Proportion of Variance 0.4865 0.1220 0.08693 0.07241 0.05582 0.05042 0.03174 0.02854 0.02107
Cumulative Proportion  0.4865 0.6085 0.69540 0.76781 0.82363 0.87405 0.90579 0.93433 0.95540

> round(cor(arctic.acp1$x[,1:3], arctic[,3:18]) , d=2)
       Al    Ti    Fe    Mn    Ca    Mg    Na     K     P    Ba    Sr    Cr    Ni    Sc     V    Zr
PC1 -0.95 -0.93  0.04 -0.28 -0.16 -0.77 -0.52 -0.92 -0.40 -0.61 -0.91 -0.78 -0.47 -0.78 -0.74 -0.92
PC2  0.03  0.03 -0.25 -0.53 -0.42 -0.39 -0.47 -0.01 -0.59  0.56 -0.15 -0.04  0.40  0.41  0.18  0.13
PC3 -0.07 -0.18  0.73  0.34  0.17  0.08 -0.31 -0.17  0.12  0.06  0.01 -0.12  0.60  0.36 -0.07 -0.13
```

(a) Comente a decisão de efectuar uma ACP sobre os dados estandardizados e discuta os resultados apresentados em cima.

(b) Eis o correspondente *biplot* bi-dimensional. Discuta-o.



(c) Qual o valor da soma de quadrados das correlações entra a quarta Componente Principal (CP) e cada uma das 16 variáveis originais?

(d) Indique um majorante (menor que 1) para a correlação entre qualquer das variáveis originais e a nona CP. Será possível que este valor seja atingido? Justifique as suas respostas.

2. Um investigador que não esteve envolvido na recolha dos dados apercebeu-se que a designação de cada uma das 64 observações começava sempre por um código de dois símbolos, como indicado em baixo:

```
> arctic$groups
 [1] 4C 4C 4C 3A 4C 2A 4C 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A
[33] 2A 2A 2A 2A 2A 2A 2A 4B 4B 2A 2A 2A 2A 2A 2A 2A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A
```

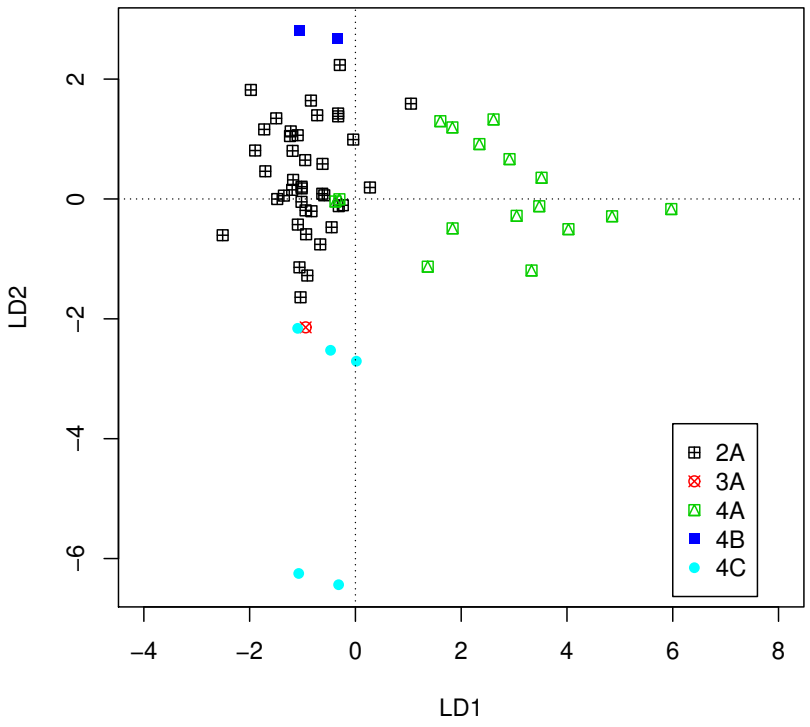
```
> summary(arctic$groups)
2A 3A 4A 4B 4C
40  1 16  2  5
```

Suspeitando que etiquetas idênticas correspondiam a observações recolhidas em condições análogas, decidiu discriminar os cinco grupos de observações através duma Análise Discriminante Linear das 16 variáveis numéricas.

(a) Em baixo indica-se a capacidade discriminante de cada um dos eixos discriminantes, obtidas usando a função `adl` do Exercício 17, que produz valores que correspondem às definições usadas nas aulas. Comente estes resultados.

```
> arctic.adl$val
[1] 2.50842423 1.79961417 0.79033773 0.06527797
```

(b) A figura em baixo dá a nuvem dos 64 pontos nos dois primeiros eixos discriminantes (com informação dada pelo comando `lda` do módulo `MASS`). Comente-a.



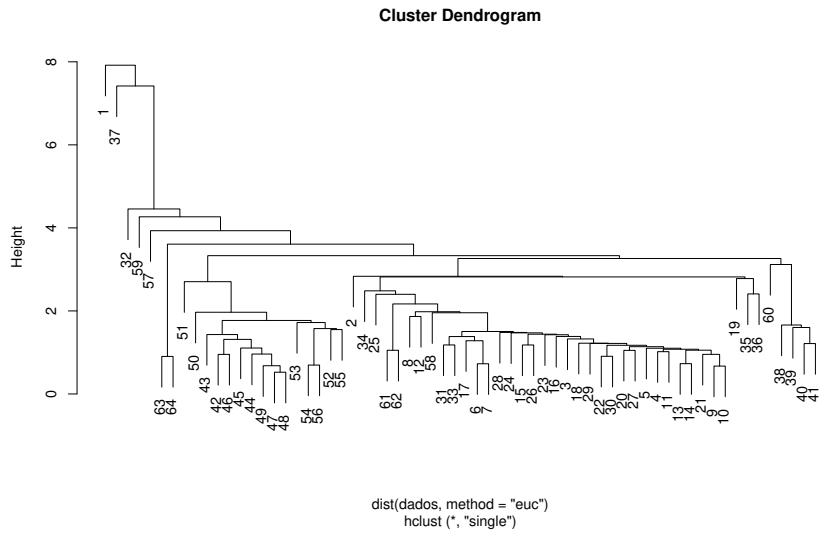
(c) Foram criadas duas novas observações, uma constituída pelos valores médios, e outra pelos terceiros quartis, das 64 observações em cada uma das 16 variáveis. Em baixo encontram-se os coeficientes nos eixos discriminantes obtidos pelo comando `lda`, para cada uma das duas novas observações:

```
> predict(arctic.lda, new=novos)$x
           LD1          LD2           LD3           LD4
1 -1.016129e-15 1.585299e-16 -3.785922e-16 -4.574619e-16
2 -1.489254e+00 5.195185e-01  4.977837e-02 -8.053507e-02
```

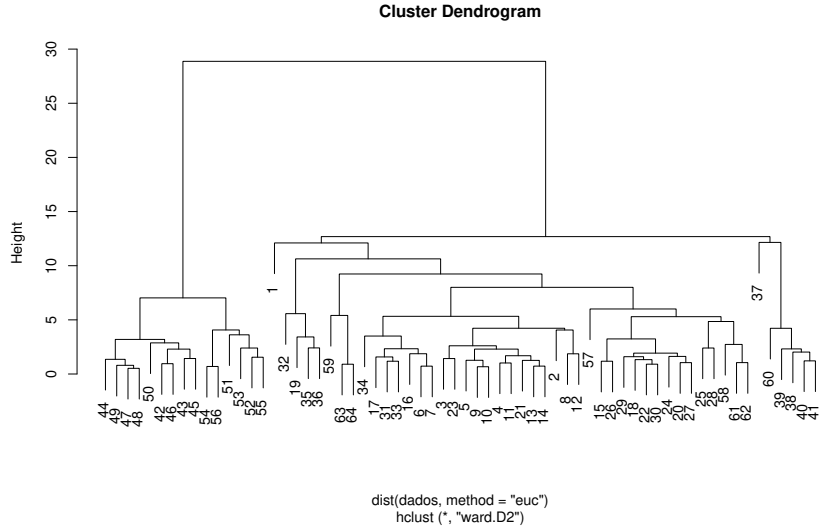
i. Identifique, justificando a sua resposta, qual a linha correspondente à observação resultante de tomar os valores médios.

ii. A qual dos grupos iria associar a outra observação? O que é que isso nos diz sobre os indivíduos que compõem esse grupo?

3. A partir dos dados estandartizados das 64 medições foi efectuada uma análise classificatória usando o método hierárquico do vizinho mais próximo com a distância euclideana, tendo-se obtido o dendrograma abaixo. Sabe-se ainda que a matriz de distâncias cofenéticas associada a este dendrograma contém os valores 7.42 e 7.92.



Seguidamente foi efectuada uma análise classificatória dos dados estandartizados das 64 medições usando o método da inércia mínima e a distância euclideana, tendo-se obtido o seguinte dendrograma e uma partição do conjunto das 64 medições em 5 grupos.



A informação sobre as distâncias entre os 5 grupos é apresentada na seguinte tabela (arredondada às duas casas decimais), onde a designação dos grupos respeita a ordem do dendrograma (da esquerda para a direita):

	C_1	C_2	C_3	C_4
C_2	15.58			
C_3	28.88	12.10		
C_4	16.81	12.13	12.28	
C_5	14.57	13.04	12.48	12.15

(a) De acordo com a informação disponibilizada, justifique quais das seguintes afirmações pode garantir que estão corretas:

i. A distância entre as medições 1 e 37 é igual a 7.92.

ii. A distância entre as medições 1 e 37 é maior ou igual a 7.92.

iii. A distância entre as medições 1 e 37 é menor que 7.92.

iv. A distância entre as medições 1 e 32 é maior ou igual a 15.34.

(b) Determine a distância cofenética entre as medições 1 e 37 definida pela análise classificatória com o método da inércia mínima. O que traduz o valor obtido do ponto de vista do método de agregação hierárquico utilizado.

(c) Sabe-se que o índice de RAND entre a classificação em 5 grupos dada pelas etiquetas **2A**, **3A**, **4A**, **4B** e **4C**, e a classificação em 5 grupos obtida pelo método da inércia mínima tem o valor de 0.5729167, e que existem 642 pares de medições que são classificadas em grupos distintos por estes dois processos de classificação. Determine o número de pares de medições que são classificados no mesmo grupo pelos dois processos de classificação.

(d) Constatou-se mais tarde que uma das medições com a etiqueta **4B** estava mal catalogada, tendo-lhe sido atribuída a etiqueta **4D**. Qual o valor do índice de RAND entre a nova classificação em 6 grupos dada pelas etiquetas e a classificação anterior em 5 grupos dada pelo método da inércia mínima?

(e) Posteriormente foi efetuada uma consolidação da partição do conjunto das 64 medições em 5 grupos, aplicando o método das k-médias móveis com sementes iniciais dadas pelos centros de gravidade desses grupos, tendo-se constatado que se obtinham as mesmas classes que no caso da inércia mínima. O que conclui? Este facto é suficiente para garantir que a partição em 5 grupos obtida pelo método da inércia mínima, minimiza o valor da inércia total intra-grupos de entre todas as partições em 5 grupos do conjunto das 64 medições ? Justifique

II [4 valores]

1. (a) Mostre que *não* é, em geral, verdade que o produto de duas matrizes simétricas $\mathbf{A}$ e $\mathbf{B}$ também seja simétrico. Indique uma condição necessária e suficiente para que o produto $\mathbf{AB}$ seja simétrico.

   (b) Considere uma Análise em Componentes Principais sobre a matriz de covariância. Admita que também existe uma estrutura em grupos dos indivíduos, definida pelos níveis de um factor, e com correspondente matriz de variabilidade intra-grupos $\mathbf{W}$. Encontre uma fórmula que indique a capacidade *discriminante* duma dada Componente Principal (centrada), $\mathbf{X}^c\vec{\mathbf{a}}$, que dependa apenas da variância dessa CP e da sua variabilidade intra-grupos, $\vec{\mathbf{a}}^t\mathbf{W}\vec{\mathbf{a}}$. Usando essa fórmula, indique um majorante para a variabilidade intra-grupos dessa CP.

2. (a) Prove que se na fórmula de Lance-Williams,

$$d_{ij,k} = \alpha_i d_{i,k} + \alpha_j d_{j,k} + \beta d_{i,j} + \gamma |d_{i,k} - d_{j,k}|,$$

   os parâmetros $\alpha_i, \alpha_j, \gamma$ são não negativos e verificam $\alpha_i + \alpha_j + \beta \geq 1$, então $d_{ij,k} \geq d_{i,j}$ para todo o grupo $C_k \neq C_i, C_j, C_{ij} (= C_i \cup C_j)$.

   (b) Deduza da alínea anterior que o método da inércia mínima não admite inversões.