INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações (2018-19)

Teste Estatística Multivariada

19 de Junho, 2019 Duração: 2h30

A Floresta Experimental H.J. Andrews, no Estado norte-americano do Oregon, disponibiliza numerosos conjuntos de dados florestais. Um subconjunto de dados tem medições de concentração de 13 nutrientes em material vegetal de pequenas bacias hidrográficas. As 113 medições foram efectuadas em seis diferentes tipos de material vegetal, que definem um factor (`TYPE`), com níveis identificados pelas siglas `H` para herbáceas; `SL` para arbustos de pequeno porte; `SH` para arbustos de grande porte; `TB` para a casca de árvores, `TF` para a folhagem de árvores e `TW` para o cerne (madeira) de árvores. Os valores observados das concentrações de nutrientes são identificados pelos seus símbolos químicos. Estas concentrações são todas medidas em $mg\ kg^{-1}$, excepto o carbono (`C`) e o azoto (`N`), que são dados em percentagens. Em baixo indicam-se as médias, desvios padrão, valores mínimos e máximos de cada uma das variáveis numéricas, bem como os respectivos coeficientes de correlação.

	P	K	CA	MG	MN	CU	B	ZN	AL	FE	NA	C	N
$\overline{x}$	1143.49	5707.90	8058.27	948.93	118.91	4.34	11.43	15.96	143.60	40.99	32.55	48.25	0.53
s	1170.44	6071.35	7573.66	967.81	143.92	1.76	5.23	17.81	193.45	33.58	24.63	2.48	0.63
min	19.0	81.0	170.0	19.0	3.0	2.0	4.0	1.0	1.0	2.0	10.0	43.1	0.0
mx	6214	27850	39600	4669	770	12	24	124	735	323	154	55.90	2.31

	P	K	CA	MG	MN	CU	B	ZN	AL	FE	NA	C	N
P	1.000	0.796	0.478	0.879	0.626	0.465	0.617	0.468	0.319	0.155	0.283	-0.354	0.937
K	0.796	1.000	0.327	0.830	0.374	0.623	0.634	0.430	0.317	0.139	0.533	-0.474	0.844
CA	0.478	0.327	1.000	0.567	0.350	0.186	0.531	0.727	-0.067	-0.146	0.304	-0.394	0.408
MG	0.879	0.830	0.567	1.000	0.463	0.569	0.729	0.539	0.195	0.153	0.465	-0.514	0.904
MN	0.626	0.374	0.350	0.463	1.000	0.016	0.486	0.218	0.360	0.055	0.031	-0.033	0.536
CU	0.465	0.623	0.186	0.569	0.016	1.000	0.311	0.429	0.107	0.112	0.366	-0.371	0.485
B	0.617	0.634	0.531	0.729	0.486	0.311	1.000	0.429	0.343	-0.015	0.416	-0.305	0.683
ZN	0.468	0.430	0.727	0.539	0.218	0.429	0.429	1.000	-0.079	-0.038	0.520	-0.482	0.404
AL	0.319	0.317	-0.067	0.195	0.360	0.107	0.343	-0.079	1.000	-0.022	0.206	0.263	0.266
FE	0.155	0.139	-0.146	0.153	0.055	0.112	-0.015	-0.038	-0.022	1.000	0.128	-0.193	0.203
NA	0.283	0.533	0.304	0.465	0.031	0.366	0.416	0.520	0.206	0.128	1.000	-0.418	0.302
C	-0.354	-0.474	-0.394	-0.514	-0.033	-0.371	-0.305	-0.482	0.263	-0.193	-0.418	1.000	-0.369
N	0.937	0.844	0.408	0.904	0.536	0.485	0.683	0.404	0.266	0.203	0.302	-0.369	1.000

I [13 valores]

1. Foi efectuada uma Análise em Componentes Principais (ACP) sobre os dados normalizados das 13 concentrações de nutrientes, tendo-se obtido os seguintes resultados parciais:

```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.4572 1.3436 1.1712 1.02936 0.8941 0.73641 0.62170
Proportion of Variance 0.4645 0.1389 0.1055 0.08151 0.0615 0.04172 0.02973
Cumulative Proportion  0.4645 0.6033 0.7088 0.79034 0.8518 0.89356 0.92329
[...]
```

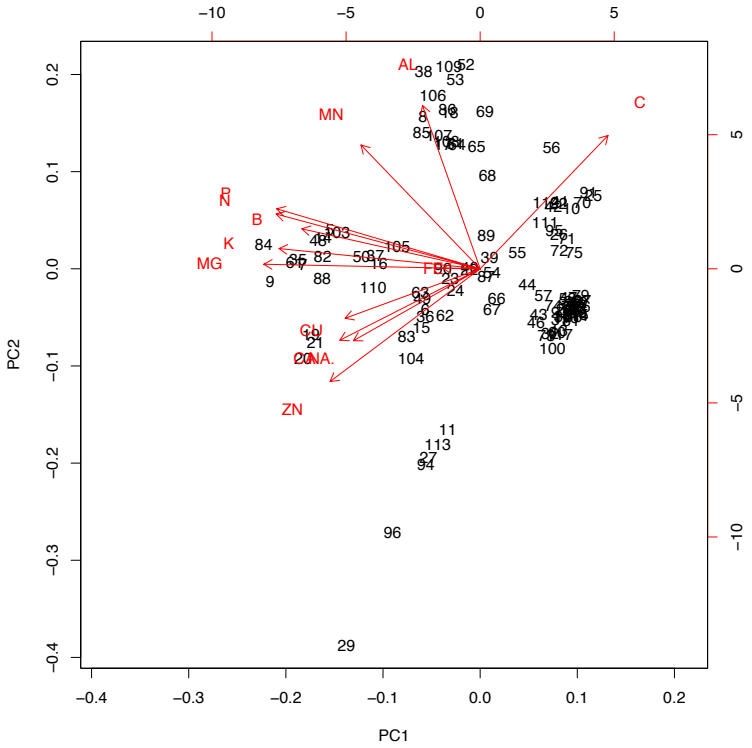
```
Rotation (n x k) = (13 x 13):
            PC1          PC2         PC3         PC4         PC5         PC6
P   -0.36327001  0.195969359 -0.01029241  0.18256538 -0.15277132  0.09438577
K   -0.35900357  0.066101621 -0.21998927 -0.13632457 -0.13668806 -0.16674548
CA  -0.25043219 -0.233184045  0.51061330  0.11060408  0.13072744  0.19283926
MG  -0.38622788  0.014900240 -0.03969922  0.07580000 -0.11064082 -0.10156130
MN  -0.21271191  0.404193399  0.27464306  0.31185535  0.19395577  0.15480598
CU  -0.24094680 -0.161523113 -0.33721604 -0.26009732 -0.48336879  0.43268441
B   -0.31827796  0.130513573  0.19387668 -0.10655345  0.11300119 -0.39171991
ZN  -0.26788976 -0.367949977  0.27614617 -0.07261102  0.10933086  0.47602226
AL  -0.10291257  0.533267134 -0.05608268 -0.48062640  0.24773944  0.14402488
FE  -0.06685579  0.001426644 -0.57840904  0.46376065  0.50368507  0.28786941
NA. -0.22624662 -0.234784462 -0.13948026 -0.49047678  0.52230572 -0.07270962
C    0.22813368  0.435078425  0.14214140 -0.17168553 -0.08071344  0.45259185
N   -0.36414610  0.178756596 -0.09405991  0.18015870 -0.19562236 -0.10404417
[...]
```

(a) Comente brevemente os resultados obtidos. Diga se considera adequada a opção por uma ACP sobre os dados normalizados.

(b) Com base na informação disponível até aqui, qual a componente principal mais correlacionada com a variável `C` (carbono), e qual o valor dessa correlação?

(c) Qual o valor médio dos quadrados das correlações de cada variável original com a primeira componente principal? Justifique e comente.

(d) Descreva e comente pormenorizadamente o seguinte *biplot*. Discuta, em particular, a observação número 29.



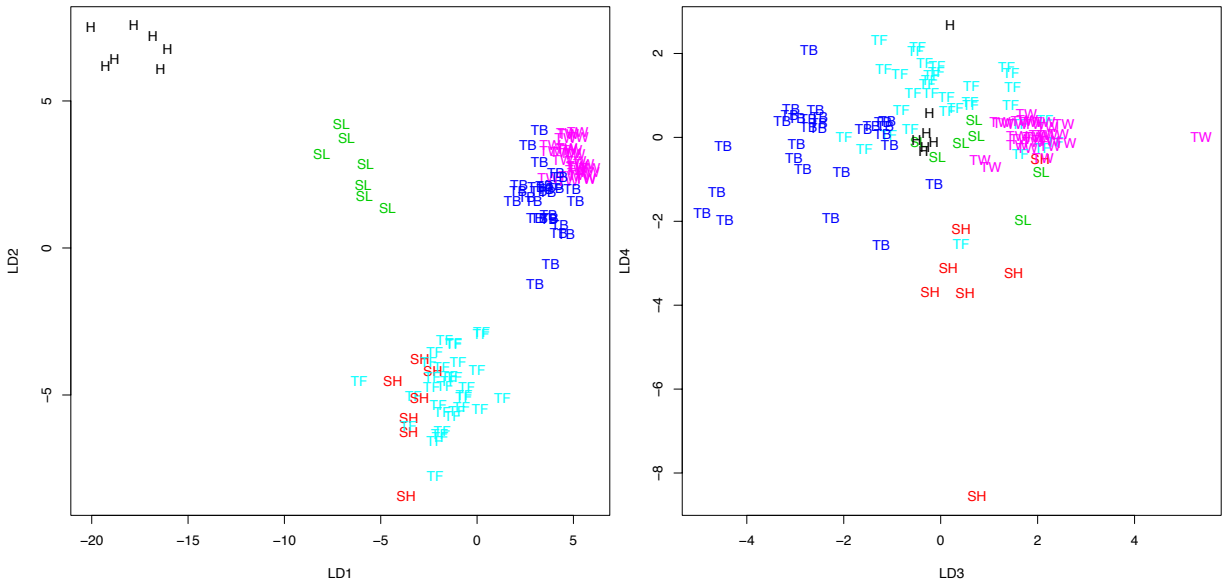
2. Seguidamente foi ajustada uma Análise Discriminante Linear, procurando discriminar os seis tipos de vegetação com base nas 13 concentrações numéricas nelas observadas. Eis alguns resultados parciais.

```
> TN025.lda
Call: lda(TN025sem1a4[, 11:23], group = TN025sem1a4$TYPE)
[...]
Coefficients of linear discriminants:
              LD1           LD2           LD3           LD4           LD5
P    1.868819e-03  7.149185e-04 -0.0008277943  1.422715e-03  5.439989e-04
K   -8.986185e-04  7.102793e-04 -0.0003159848  7.938650e-05  2.389478e-06
CA  -3.241653e-05 -1.846769e-04 -0.0001454056  3.296684e-05 -1.090798e-04
MG   9.076800e-04  1.514607e-03 -0.0008598320 -3.447690e-03 -8.326410e-04
MN  -2.652726e-03 -5.337732e-03  0.0036408206 -1.000389e-03 -1.435787e-03
```

```
CU   7.475112e-04  1.804773e-01  0.0145404241 -1.587132e-01 -6.092962e-02
B   -1.097782e-01 -1.436588e-01 -0.0603128202 -7.544791e-02  2.187124e-01
ZN   7.052104e-02 -1.154738e-02 -0.0227150881  9.669037e-03  2.632607e-02
AL  -3.490529e-03 -1.382937e-03 -0.0021217335  5.333779e-04 -5.785763e-03
FE   4.614523e-03  3.019507e-03  0.0124579441 -3.418548e-04 -9.331240e-03
NA. -5.041848e-02  1.067654e-02  0.0330457039  2.846635e-02  1.626117e-02
C   -1.710253e-02 -2.527090e-01 -0.4650363345  6.979374e-02  2.135671e-01
N   -3.167573e+00 -1.066442e+01  5.5245934459  1.722573e+00  1.053483e-01

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5
0.6332 0.2907 0.0515 0.0200 0.0046
```

(a) Em baixo, à esquerda, está o gráfico dos 113 pontos no plano definido pelos dois primeiros eixos discriminantes. Em baixo, direita, está a nuvem de pontos análoga, no plano definido pelos terceiro e quarto eixos discriminantes. Comente-os. Em particular, diga se considera que os terceiro e quarto eixos discriminantes podem ter algum papel útil na discriminação.



(b) A qualidade do primeiro eixo discriminante, através do critério dado nas aulas, 34.9538089. Interprete o significado deste valor e discuta-o, tendo em conta o gráfico da alínea anterior.

3. Seja $\mathbf{X}^c_{n\times p}$ uma matriz centrada de dados, e $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{ct}\mathbf{X}^c$ a respectiva matriz de (co-)variâncias. Seja $\mathbf{D}_{p\times p}$ a matriz diagonal cujos elementos diagonais so os recíprocos dos desvios-padrões das variáveis das colunas de $\mathbf{X}^c$, pelo que $\mathbf{Z} = \mathbf{X}^c\mathbf{D}$ é a matriz dos dados normalizados correspondentes a $\mathbf{X}^c$.

(a) Mostre que a matriz de correlações dos dados originais é dada por $\mathbf{R} = \mathbf{DSD}$.

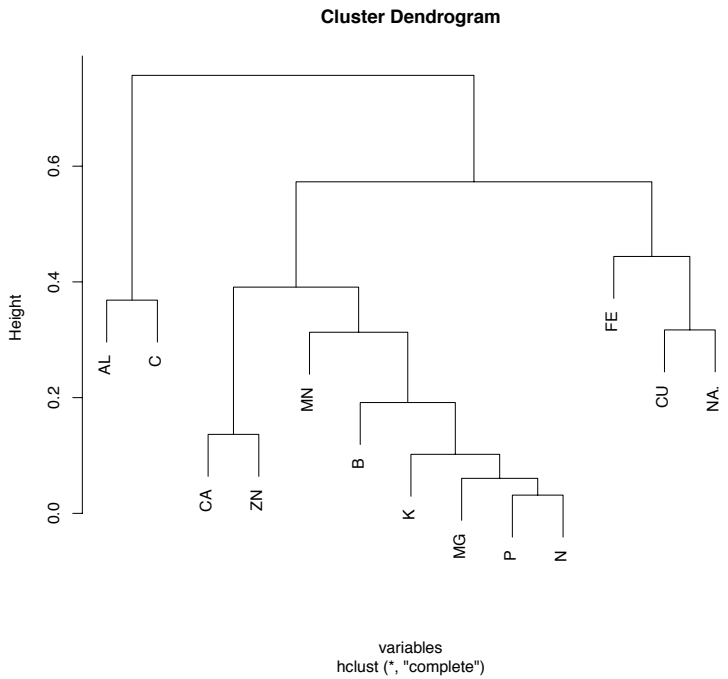
Seja G o factor que define k grupos de indivíduos (linhas) da matriz $\mathbf{X}^c$, e $\mathbf{P}_G$ a matriz de projecção ortogonal sobre o subespaço gerado pelas variáveis indicatrizes de cada um dos k nveis do factor. Sejam $\mathbf{B} = \frac{1}{n-1}\mathbf{X}^{ct}\mathbf{P}_G\mathbf{X}^c$ e $\mathbf{W} = \frac{1}{n-1}\mathbf{X}^{ct}(\mathbf{I}_n - \mathbf{P}_G)\mathbf{X}^c$ as habituais matrizes da variabilidade, respectivamente, inter-grupos e intra-grupos.

(b) Mostre que as matrizes de variabilidades inter- e intra-grupos dos dados normalizados so dadas, respectivamente, por $\mathbf{B}_R = \mathbf{DBD}$ e $\mathbf{W}_R = \mathbf{DWD}$.

(c) Mostre que se $\vec{\mathbf{b}}$ é um vector próprio de ${\mathbf{W}_R}^{-1}\mathbf{B}_R$, com valor próprio β, então $\mathbf{D}\vec{\mathbf{b}}$ é um vector próprio de $\mathbf{W}^{-1}\mathbf{B}$, associado ao mesmo valor próprio β. Discuta as consequências desse facto para os resultados de Análises Discriminantes Lineares sobre os dados originais, e sobre os dados normalizados das 13 concentrações de nutrientes. Comente.

II [7 valores]

1. Efectuou-se uma classificação hierárquica das variáveis usando o método do vizinho mais afastado (*complete*) com a dissemelhança entre variáveis dada por $d = (1 - r)/2$, onde r denota o coeficiente de correlação entre as variáveis, tendo-se obtido o dendrograma da figura abaixo.



(a) Comente o dendrograma obtido com base na informação disponibilizada e no método de classificação escolhido.

(b) Determine a distância cofenética entre as variáveis *Zinco* (ZN) e *Fósforo* (P).

(c) Considere as partições em dois e em três grupos que se obtêm a partir da análise classificatória hierárquica do conjunto das variáveis. Descreva essas partições e compare-as com o índice de RAND.

(d) Foi efectuada uma análise classificatória dos 113 dados normalizados com o método da inércia mínima tendo-se obtido três grupos homogéneos C_1, C_2 e C_3, com 19, 42 e 52 elementos, respectivamente. As distâncias entre esses grupos vêm dadas por $d(C_1, C_2) = 18.913$, $d(C_1, C_3) = 33.867$ e $d(C_2, C_3) = 23.138$. Determine o custo de fusão no último passo do algorítmo hierárquico.

2. Efectuaram-se duas análises classificatórias do conjunto de 3 pontos, $a = (0, 0)$, $b = (1, 0)$ e $c = (c_1, c_2)$, $c_1, c_2 > 0$, com o método hierárquico do vizinho mais próximo, tendo-se posteriormente efectuado um corte nos respectivos dendrogramas de modo a obter dois grupos para ambas as classificações. Na primeira análise classificatória foi usada a distância euclideana tendo-se obtido a partição $\{a, b\} \cup \{c\}$ e na segunda a distância do máximo tendo-se obtido a partição $\{a, c\} \cup \{b\}$. Identifique e represente no plano o conjunto de todos os possíveis pontos c.

3. Usando a fórmula de Lance-Williams mostre que o método hierárquico da inércia mínima não admite inversões.