Exercícios - Modelos Matemáticos e Aplicações
Introdução à Estatística Multivariada - 2016-17

1 Matrizes e Álgebra Linear

1. Considere o espaço $\mathbb{R}^2$. Seja M o subespaço de $\mathbb{R}^2$ gerado pelo vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Seja N o subespaço de $\mathbb{R}^2$ gerado pelo vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.
   (a) Caracterize os vectores do subespaço M.
   (b) Qual a projecção ortogonal do vector $\begin{bmatrix} c \\ d \end{bmatrix}$ sobre o subespaço M?
   (c) Caracterize os vectores do subespaço N.
   (d) Qual a projecção ortogonal do vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ sobre o subespaço N?

2. Considere o espaço $\mathbb{R}^n$ com o habitual produto interno.
   (a) Diga o que caracteriza os vectores de $\mathbb{R}^n$ que são ortogonais ao vector de n uns, $\mathbf{1}_n$.
   (b) Associe os pontos/vectores de $\mathbb{R}^n$ a possíveis conjuntos de n observações duma variável. Do ponto de vista estatístico, como pode caracterizar os elementos do subespaço caracterizado na alínea anterior?

3. Seja $\vec{\mathbf{y}} \in \mathbb{R}^n$ a representação vectorial de n observações duma variável. Seja $\vec{\mathbf{y}}^c \in \mathbb{R}^n$ o correspondente vector centrado.
   (a) Discuta o efeito duma translação da origem da escala de valores da variável (i.e., se cada $y_i \to a + y_i$) nos vectores $\vec{\mathbf{y}}$ e $\vec{\mathbf{y}}^c$.
   (b) Discuta o efeito duma mudança multiplicativa de escala ($y_i \to by_i, \forall i$) nos vectores $\vec{\mathbf{y}}$ e $\vec{\mathbf{y}}^c$.
   (c) Discuta o efeito duma transformação linear $y_i \to a + by_i, \forall i$, nos vectores $\vec{\mathbf{y}}$ e $\vec{\mathbf{y}}^c$.

   Considere agora um novo vector $\vec{\mathbf{x}} \in \mathbb{R}^n$ representando observações duma nova variável sobre os mesmos n indivíduos. Seja $\vec{\mathbf{x}}^c$ o respectivo vector centrado.
   (d) Discuta o efeito de diferentes transformações lineares nas duas variáveis ($x_i \to a + bx_i$ e $y_i \to c + dy_i, \forall i$) sobre os vectores que as representam em $\mathbb{R}^n$. Discuta a influência desse efeito nos indicadores estatísticos covariância e coeficiente de correlação.

4. Considere as matrizes $\mathbf{X}^t\mathbf{X}$ e $\mathbf{X}\mathbf{X}^t$, onde $\mathbf{X}$ é uma matriz $n \times p$. Verifique que se $\lambda_j \neq 0$ é um valor próprio de $\mathbf{X}^t\mathbf{X}$, com vector próprio associado $\vec{\mathbf{c}}_j$, então $\mathbf{X}\vec{\mathbf{c}}_j$ é um vector próprio da matriz $\mathbf{X}\mathbf{X}^t$, para o mesmo valor próprio. Conversamente, se $\lambda_j \neq 0$ é um valor próprio de $\mathbf{X}\mathbf{X}^t$ com vector próprio associado $\vec{\mathbf{b}}_j$, então $\mathbf{X}^t\vec{\mathbf{b}}_j$ é um vector próprio de $\mathbf{X}^t\mathbf{X}$, com o mesmo valor próprio associado.

5. Utilize a Decomposição em Valores Singulares duma matriz $\mathbf{Y}$, na forma:

$$\mathbf{Y} = \sum_{i=1}^{r} \delta_i \vec{\mathbf{w}}_i \vec{\mathbf{v}}_i^t$$

para mostrar que se $\vec{\mathbf{w}}_i$ é um vector singular esquerdo associado ao valor singular δ_i, e $\vec{\mathbf{v}}_i$ é um vector singular direito associado ao mesmo valor singular, então tem-se:

$$\mathbf{Y}\vec{\mathbf{v}}_i = \delta_i \vec{\mathbf{w}}_i \qquad \text{e} \qquad \mathbf{Y}^t \vec{\mathbf{w}}_i = \delta_i \vec{\mathbf{v}}_i$$

6. Considere uma matriz $\mathbf{B}$ e a matriz de projecção ortogonal sobre o subespaço gerado pelas colunas de $\mathbf{B}$, $\mathbf{P}_B = \mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t$. Utilizando a Decomposição em Valores Singulares da matriz $\mathbf{B}$, obtenha uma expressão alternativa para a matriz $\mathbf{P}_B$. Comente.

2 Análise em Componentes Principais

7. Nas Estatísticas Agrícolas do INE (1973) indicam-se as produtividades (em t/ha) de 9 produções agrícolas, nos 20 concelhos do distrito de Santarém. Os dados são os seguintes, e encontram-se na *data frame* `santarem`, disponível na página da disciplina, no ficheiro `dadosACP.RData`.

Concelho	Trigo	Milho	Centeio	Aveia	Cevada	Fava	Feijão	Grão	Batata
Abrantes	1.041	0.541	0.515	0.595	0.402	0.672	0.327	0.423	7.437
Alcanena	0.887	1.697	0.700	1.051	0.630	0.631	0.517	0.618	10.317
Almeirim	1.013	0.431	0.545	0.511	0.374	0.696	0.376	0.495	7.389
Alpiarça	1.293	1.803	0.891	0.413	1.094	0.591	0.518	0.500	17.678
Benavente	1.559	1.949	0.669	1.053	1.029	0.628	0.346	0.614	8.290
Cartaxo	0.925	1.600	0.544	0.696	0.460	0.657	0.352	0.469	9.071
Chamusca	1.103	3.144	0.379	0.321	0.423	0.542	0.543	0.442	17.199
Constância	1.516	0.524	0.321	0.562	0.571	0.474	0.381	0.485	11.271
Coruche	1.443	0.483	0.605	0.698	1.250	0.742	0.229	0.371	19.160
Entroncamento	1.023	4.120	0.716	0.621	0.707	1.057	0.533	0.700	20.600
F.do Zêzere	0.981	2.413	0.305	0.773	1.048	0.696	0.524	0.602	9.889
Golegã	1.223	3.777	0.646	0.330	0.763	0.763	0.672	0.311	8.113
Mação	0.839	0.772	0.306	0.362	0.260	0.600	0.293	0.420	8.468
Rio Maior	0.809	1.153	0.927	0.694	0.707	1.777	0.417	0.433	7.060
Salvaterra	1.509	1.100	1.034	0.697	1.582	1.138	0.636	0.516	10.791
Santarém	0.712	1.342	1.145	0.457	0.686	0.982	0.616	0.426	14.135
Sardoal	0.780	0.463	0.326	0.414	0.435	0.822	0.383	0.396	10.078
Tomar	1.000	1.928	0.430	0.863	1.080	0.913	0.404	0.687	9.320
Torres Novas	1.262	2.453	0.716	0.971	0.885	0.928	0.512	0.664	21.100
V.N.Barquinha	0.917	1.081	0.811	1.000	0.909	0.967	0.620	0.667	18.347

A matriz de variâncias-covariâncias destes dados é a seguinte:

```
> round(var(santarem), d=3)
         trigo  milho centeio  aveia cevada   fava feijao  grao batata
trigo    0.069  0.016   0.002  0.010  0.050 -0.021 -0.002 0.001  0.236
milho    0.016  1.198   0.017 -0.006  0.040  0.009  0.076 0.038  1.735
centeio  0.002  0.017   0.062  0.011  0.039  0.041  0.016 0.002  0.308
aveia    0.010 -0.006   0.011  0.057  0.034  0.012 -0.001 0.020  0.106
cevada   0.050  0.040   0.039  0.034  0.117  0.026  0.013 0.012  0.470
fava    -0.021  0.009   0.041  0.012  0.026  0.084  0.009 0.003 -0.003
feijao  -0.002  0.076   0.016 -0.001  0.013  0.009  0.016 0.003  0.167
grao     0.001  0.038   0.002  0.020  0.012  0.003  0.003 0.013  0.184
batata   0.236  1.735   0.308  0.106  0.470 -0.003  0.167 0.184 23.531
```

(a) Considere uma Análise de Componentes Principais sobre a matriz das covariâncias (isto é, sobre os dados originais).

i. Discuta a qualidade da redução de dimensionalidade possibilitada pela ACP.

ii. Construa a nuvem de 20 pontos (concelhos) sobre o plano definido pelas duas primeiras componentes principais. Identifique, no gráfico construído usando o comando `prcomp` do `R`, os 7 concelhos correspondentes aos pontos na metade direita do gráfico. Identifique ainda o ponto isolado no canto superior esquerdo.

iii. Calcule, usando o `R`, os coeficientes de correlação entre a CP 1 e as nove variáveis originais. Confirme os valores obtidos para os três coeficientes de correlação entre a primeira componente principal e as variáveis "batata","fava"e "milho", usando a fórmula dada nas aulas para estes coeficientes de correlação. Repita para a segunda componente principal. Comente.

iv. Procure interpretar a natureza das duas primeiras componentes principais, justificando as suas conclusões.

v. Construa o *biplot* correspondente e comente.

vi. Avalie criticamente a Análise de Componentes Principais (ACP) efectuada, discutindo em particular a opção por uma ACP sobre a matriz das covariâncias.

(b) Efectue agora uma Análise de Componentes Principais dos dados normalizados, ou seja, baseada na matriz de correlações.

i. Discuta a qualidade da redução de dimensionalidade possibilitada pela ACP sobre a matriz de correlações. Comente, tendo também em conta o resultado da ACP sobre os dados originais.

ii. Calcule os coeficientes de correlação entre cada uma das variáveis originais e cada uma das CPs agora obtidas. É necessário normalizar as variáveis para calcular os coeficientes de correlação?

iii. Construa o *biplot* correspondente e comente. Em particular, procure interpretar a natureza das duas primeiras componentes principais sobre os dados normalizados.

(c) Responda à pergunta de um utilizador: "*por qual das duas variantes de ACP devo optar*"?

8. Numa estudo sobre o cultivo de framboesas em estufa, observam-se 7 variáveis caracterizadoras de propriedades de frutos colhidos. Mais concretamente, recolhem-se framboesas em 14 plantas e determinam-se os valores médios desses 14 grupos de framboesas para as seguintes variáveis: `Diametro`, `Altura`, `Peso`, `Brix`, `pH`, uma outra medida de acidez, que será designada apenas por `Acidez`, e `Acucar`. Os valores obtidos são dados na *data frame* `brix2` (já estudada no Módulo II, mas com a nova variável `Acidez`) e foram:

Planta	Diametro	Altura	Peso	Brix	pH	Acidez	Acucar
1	2.0	2.1	3.71	8.4	2.78	1.39	5.12
2	2.1	2.0	3.79	8.4	2.84	1.49	5.40
3	2.0	1.7	3.65	8.7	2.89	1.51	5.38
4	2.0	1.8	3.83	8.6	2.91	1.44	5.23
5	1.8	1.8	3.95	8.0	2.84	1.62	3.44
6	2.0	1.9	4.18	8.2	3.00	1.74	3.42
7	2.1	2.2	4.37	8.1	3.00	1.68	3.48
8	1.8	1.9	3.97	8.0	2.96	1.57	3.34
9	1.8	1.8	3.43	8.2	2.75	1.46	2.02
10	1.9	1.9	3.78	8.0	2.75	1.54	2.14
11	1.9	1.9	3.42	8.0	2.73	1.26	2.06
12	2.0	1.9	3.60	8.1	2.71	1.18	2.02
13	1.9	1.7	2.87	8.4	2.94	1.32	3.86
14	2.1	1.9	3.74	8.8	3.20	1.46	3.89

(a) Diga, justificando, se uma Análise em Componentes Principais sobre a matriz de covariâncias é adequada para este conjunto de dados.

(b) Diga, justificando, se uma Análise em Componentes Principais sobre a matriz das correlações permite representar de forma adequada o conjunto dos dados em apenas duas dimensões, sem grande perda de informação.

(c) Independentemente da sua resposta na alínea anterior, construa um `biplot` para os dados em apreço. Comente.

(d) As 14 plantas não foram observadas todas nas mesmas datas, tendo os cortes sido efectuados em cinco datas diferentes:

Data de corte	Plantas
28 Novembro	1,2,3,4
13 Dezembro	5,6,7,8
16 Janeiro	9,10,11,12
20 Fevereiro	13
3 Abril	14

As diferentes datas de corte são reflectidas no primeiro plano principal resultante da análise anterior? Responda, identificando os pontos do gráfico da nuvem de pontos no primeiro plano principal.

(e) *Caso a sua resposta à alínea anterior seja afirmativa*, diga, justificando, se é obrigatório que o primeiro plano principal reflicta esse tipo de estrutura dos dados em sub-grupos. *Caso a sua resposta á alínea anterior seja negativa*, diga como se pode explicar que essa estrutura não esteja reflectida no primeiro plano principal, dadas as propriedades optimizadas pelas primeiras componentes principais.

(f) Admita agora que se procede a uma nova observação dos valores das 7 variáveis nas framboesas de uma nova planta e que se registaram os seguintes valores: 1.9, 2.0, 3.92, 8.1, 2.91, 1.48, 3.78. Se pretendesse representar esta nova observação no primeiro plano principal, quais as coordenadas que deveria associar-lhe? Justifique a sua resposta e represente o ponto no gráfico dado acima. Confirme a sua resposta, usando o comando `predict` do `R`, que tem um método para objectos resultantes de ACPs efectuadas com o comando `prcomp` (cuja utilização é semelhante à do comando `predict` para modelos lineares ou modelos lineares generalizados).

9. Considere os dados da produção de milho no estado do Iowa (EUA), nos anos 1930-1962, já estudados no Módulo II, e que se encontram na *data frame* `milho`.

(a) Qual a variante de ACP (sobre a matriz de covariâncias ou sobre a matriz de correlações) que considera adequada para o estudo destes dados? Justifique.

(b) Discuta a qualidade da redução de dimensionalidade que se obtém com uma ACP sobre as 10 variáveis normalizados.

(c) Construa um *biplot* correspondente à ACP sobre a matriz de correlações.

i. Comente o *biplot*, tendo também em conta o submodelo de regressão linear múltipla escolhido por todos os métodos de selecção de submodelos e que se reduziu a modelar y com base nos quatro preditores $x1$, $x2$, $x6$ e $x9$. É possível fazer algum comentário sobre esta escolha, com base no *biplot*?

ii. Comente a seguinte afirmação: "*O* biplot *sugere que as variáveis* $x3$ *e* $x5$ *são fortemente correlacionadas, mas essa conclusão não é confirmada pela matriz de correlações entre as* 10 *variáveis em causa*".

iii. Comente a seguinte afirmação: "*Tratando-se duma ACP sobre a matriz de correlações, todos os vectores que representam as variáveis centradas no espaço das variáveis deveriam ter igual comprimento. O facto de a variável* $x8$ *estar representada no* biplot *por um vector mais curto do que os restantes sugere que essa variável está mal representada nas duas primeiras CPs normalizadas*".

10. No âmbito dum estudo realizado na Bélgica em 1935 (Berce e Wilbaux, 1935 *Recherche Statistique des relations existant entre le rendement des plantes de grandes cultures et les facteurs météorologiques en Belgique*. Bull. Inst. Agron. Stn. Rech. Gembloux, **4**, 32–81), registaram-se os valores de $p = 5$ variáveis meteorológicas e agronómicas ao longo de $n = 11$ anos agrícolas. As cinco variáveis são:

x_1	precipitação total em Novembro e Dezembro (mm)
x_2	temperatura média em Julho (^{o}C)
x_3	precipitação total em Julho (mm)
x_4	radiação em Julho (mm de álcool)
x_5	rendimento médio da colheita de trigo candial (quintais/ha)

Os valores observados foram:

Campanha	$\vec{\mathbf{x}}_1$	$\vec{\mathbf{x}}_2$	$\vec{\mathbf{x}}_3$	$\vec{\mathbf{x}}_4$	$\vec{\mathbf{x}}_5$
1920-21	87.9	19.6	1.0	1661	28.37
1921-22	89.9	15.2	90.1	968	23.77
1922-23	153.0	19.7	56.6	1353	26.04
1923-24	132.1	17.0	91.0	1293	25.74
1924-25	88.8	18.3	93.7	1153	26.68
1925-26	220.9	17.8	106.9	1286	24.29
1926-27	117.7	17.8	65.5	1104	28.00
1927-28	109.0	18.3	41.8	1574	28.37
1928-29	156.1	17.8	57.4	1222	24.96
1929-30	181.5	16.8	140.6	902	21.66
1930-31	181.4	17.0	74.3	1150	24.37

(a) Efectue uma Análise em Componentes Principais sobre a matriz de correlações destes dados, determinando as cinco Componentes Principais dos dados.

(b) Construa a melhor representação possível, a duas dimensões, da nuvem de $n = 11$ pontos em $\mathbb{R}^5$ (anos agrícolas) que os dados definem.

(c) Calcule os coeficientes de correlação entre a primeira Componente Principal e as cinco variáveis originais. Interprete os resultados obtidos.

(d) Algumas das unidades de medida dos dados já não são usadas. As unidades de rendimento mais frequentes são toneladas por hectare, o que significa que os valores da variável $x5$ deverão ser divididos por 10. Por outro lado, as unidades de radiação do sistama métrico são MJm^{-2}, o que significa que para converter os valores na tabela da variável $x4$ para estas unidades, será necessário efectuar a seguinte transformação afim: $x_4^* = -0.02960342 + 0.75518263\, x_4$. Diga em que medida é que estas transformações afectam as suas respostas às alíneas anteriores. Confirme a sua resposta no `R`.

11. Considere o seguinte conjunto de dados, referido por Kendall (*Multivariate Analysis*, Charles Griffin & Co., 1980, pg. 20), e relativo a medições em 20 amostras de terras:

Amostra	Teor arenoso (%)	Teor em limo (%)	Teor em argila (%)	Matéria orgânica (%)	Acidez (pH)
1	77.3	13.0	9.7	1.5	6.4
2	82.5	10.0	7.5	1.5	6.5
3	66.9	20.6	12.5	2.3	7.0
4	47.2	33.8	19.0	2.8	5.8
5	65.3	20.5	14.2	1.9	6.9
6	83.3	10.0	6.7	2.2	7.0
7	81.6	12.7	5.7	2.9	6.7
8	47.8	36.5	15.7	2.3	7.2
9	48.6	37.1	14.3	2.1	7.2
10	61.6	25.5	12.9	1.9	7.3
11	58.6	26.5	14.9	2.4	6.7
12	69.3	22.3	8.4	4.0	7.0
13	61.8	30.8	7.4	2.7	6.4
14	67.7	25.3	7.0	4.8	7.3
15	57.2	31.2	11.6	2.4	6.5
16	67.2	22.7	10.1	3.3	6.2
17	59.2	31.2	9.6	2.4	6.0
18	80.2	13.2	6.6	2.0	5.8
19	82.2	11.1	6.7	2.2	7.2
20	69.7	20.7	9.6	3.1	5.9

(a) Efectue uma Análise de Componentes Principais sobre a matriz de Covariâncias destes dados. Explique a existência de um valor próprio nulo e a natureza do vector próprio correspondente.

(b) Construa um *biplot* associado à ACP sobre os dados normalizados. As posições relativas dos vectores representativos das variáveis `acidez` e `mat.org` (matéria orgânica) sugere que se trata de duas variáveis fortemente correlacionadas. No entanto, esse facto não é confirmado pela inspecção da matriz de correlações entre as variáveis originais. Como explicar esta aparente contradição?

(c) Elimine agora a variável `areia` (teor arenoso) da matriz de dados. Repita a ACP sobre a matriz de covariâncias.

i. Calcule o coeficiente de correlação entre cada Componente Principal e cada variável.

ii. Compare os valores obtidos com os coeficientes das variáveis nas combinações lineares que definem as CPs e veja como a tentativa de interpretar Componentes Principais apenas em função dos coeficientes (*loadings*) pode induzir em erro.

12. Num estudo dos afídios alados *Alate adelges* efectuaram-se medições de 19 variáveis sobre 40 indivíduos. As 19 variáveis observadas, bem como as médias e variâncias dos valores observados, foram as seguintes:

Nome	Acrónimo	Descrição	$\overline{x}$	s^2
length	COM	comprimento total do organismo	15.05	14.58
width	LAR	largura do corpo	7.14	4.05
forwing	CAA	comprimento da asa anterior	5.68	1.68
hinwing	CAP	comprimento da asa posterior	3.45	0.83
spirac	E	número de espiráculos	4.88	0.11
antseg1	AS1	comprimento do segmento de antena I	1.86	0.11
antseg2	AS2	comprimento do segmento de antena II	1.69	0.11
antseg3	AS3	comprimento do segmento de antena III	2.25	0.22
antseg4	AS4	comprimento do segmento de antena IV	2.33	0.15
antseg5	AS5	comprimento do segmento de antena V	2.73	0.15
antspin	S	número de sedas antenais	4.28	1.33
tarsus3	TAR	comprimento do tarso III	3.31	0.41
tibia3	TIB	comprimento da tíbia III	3.38	0.58
femur3	FEM	comprimento do fémur III	2.57	0.34
rostrum	ROS	rostrum	5.58	0.79
ovipos	OVI	oviescapto	3.72	0.35
ovspin	N	número de sedas do oviescapto	7.80	3.81
fold	P	prega anal (var. qualitativa 0/1)	0.73	0.20
hooks	GAP	número de ganchos da asa posterior	2.38	0.25

As observações obtidas foram as seguintes:

COM	LAR	CAA	CAP	E	AS1	AS2	AS3	AS4	AS5	S	TAR	TIB	FEM	ROS	OVI	N	P	GAP
21.2	11.0	7.5	4.8	5	2.0	2.0	2.8	2.8	3.3	3	4.4	4.5	3.6	7.0	4.0	8	0	3
20.2	10.0	7.5	5.0	5	2.3	2.1	3.0	3.0	3.2	5	4.2	4.5	3.5	7.6	4.2	8	0	3
20.2	10.0	7.0	4.6	5	1.9	2.1	3.0	2.5	3.3	1	4.2	4.4	3.3	7.0	4.0	6	0	3
22.5	8.8	7.4	4.7	5	2.4	2.1	3.0	2.7	3.5	5	4.2	4.4	3.6	6.8	4.1	6	0	3
20.6	11.0	8.0	4.8	5	2.4	2.0	2.9	2.7	3.0	4	4.2	4.7	3.5	6.7	4.0	6	0	3
19.1	9.2	7.0	4.5	5	1.8	1.9	2.8	3.0	3.2	5	4.1	4.3	3.3	5.7	3.8	8	0	3.5
20.8	11.4	7.7	4.9	5	2.5	2.1	3.1	3.1	3.2	4	4.2	4.7	3.6	6.6	4.0	8	0	3
15.5	8.2	6.3	4.9	5	2.0	2.0	2.9	2.4	3.0	3	3.7	3.8	2.9	6.7	3.5	6	0	3.5
16.7	8.8	6.4	4.5	5	2.1	1.9	2.8	2.7	3.1	3	3.7	3.8	2.8	6.1	3.7	8	0	3
19.7	9.9	8.2	4.7	5	2.2	2.0	3.0	3.0	3.1	0	4.1	4.3	3.3	6.0	3.8	8	0	3
10.6	5.2	3.9	2.3	4	1.2	1.0	2.0	2.0	2.2	6	2.5	2.5	2.0	4.5	2.7	4	1	2
9.2	4.5	3.7	2.2	4	1.3	1.2	2.0	1.6	2.1	5	2.4	2.3	1.8	4.1	2.4	4	1	2
9.6	4.5	3.6	2.3	4	1.3	1.0	1.9	1.7	2.2	4	2.4	2.3	1.7	4.0	2.3	4	1	2
8.5	4.0	3.8	2.2	4	1.3	1.1	1.9	2.0	2.1	5	2.4	2.4	1.9	4.4	2.3	4	1	2
11.0	4.7	4.2	2.3	4	1.2	1.0	1.9	2.0	2.2	4	2.5	2.5	2.0	4.5	2.6	4	1	2
18.1	8.2	5.9	3.5	5	1.9	1.9	1.9	2.7	2.8	4	3.5	3.8	2.9	6.0	4.5	9	1	2
17.6	8.3	6.0	3.8	5	2.0	1.9	2.0	2.2	2.9	3	3.5	3.6	2.8	5.7	4.3	10	1	2
19.2	6.6	6.2	3.4	5	2.0	1.8	2.2	2.3	2.8	4	3.5	3.4	2.5	5.3	3.8	10	1	2
15.4	7.6	7.1	3.4	5	2.0	1.9	2.5	2.5	2.9	4	3.3	3.6	2.7	6.0	4.2	8	1	3
15.1	7.3	6.2	3.8	5	2.0	1.8	2.1	2.4	2.5	4	3.7	3.7	2.8	6.4	4.3	10	1	2.5
16.1	7.9	5.8	3.7	5	2.1	1.9	2.3	2.6	2.9	5	3.6	3.6	2.7	6.0	4.5	10	1	2
19.1	8.8	6.4	3.9	5	2.2	2.0	2.3	2.4	2.9	4	3.8	4.0	3.0	6.5	4.5	10	1	2.5
15.3	6.4	5.3	3.3	5	1.7	1.6	2.0	2.2	2.5	5	3.4	3.4	2.6	5.4	4.0	10	1	2
14.8	8.1	6.2	3.7	5	2.2	2.0	2.2	2.4	3.2	5	3.5	3.7	2.7	6.0	4.1	10	1	2
16.2	7.7	6.9	3.7	5	2.0	1.8	2.3	2.4	2.8	4	3.8	3.7	2.7	5.7	4.2	10	1	2.5
13.4	6.9	5.7	3.4	5	2.0	1.8	2.8	2.0	2.6	4	3.6	3.6	2.6	5.5	3.9	10	1	2
12.9	5.8	4.8	2.6	5	1.6	1.5	1.9	2.1	2.6	5	2.8	3.0	2.2	5.1	3.6	9	1	3
12.0	6.5	5.3	3.2	5	1.9	1.9	2.3	2.5	3.0	5	3.3	3.5	2.6	5.4	4.3	8	1	2
14.1	7.0	5.5	3.6	5	2.2	2.0	2.3	2.5	3.1	5	3.6	3.7	2.8	5.8	4.1	10	1	2
16.7	7.2	5.7	3.5	5	1.9	1.9	2.5	2.3	2.8	5	3.4	3.6	2.7	6.0	4.0	10	1	2.5
14.1	5.4	5.0	3.0	5	1.7	1.6	1.8	2.5	2.4	5	2.7	2.9	2.2	5.3	3.6	8	1	2
10.0	6.0	4.2	2.5	5	1.6	1.4	1.4	2.0	2.7	6	2.8	2.5	1.8	4.8	3.4	8	1	2
11.4	4.5	4.4	2.7	5	1.8	1.5	1.9	1.7	2.5	5	2.7	2.5	1.9	4.7	3.7	8	1	2
12.5	5.5	4.7	2.3	5	1.8	1.4	1.8	2.2	2.4	4	2.8	2.6	2.0	5.1	3.7	8	0	2
13.0	5.3	4.7	2.3	5	1.6	1.4	1.8	1.8	2.5	4	2.7	2.7	2.1	5.0	3.6	8	1	2
12.4	5.2	4.4	2.6	5	1.6	1.4	1.8	2.2	2.2	5	2.7	2.5	2.0	5.0	3.2	6	1	2
12.0	5.4	4.9	3.0	5	1.7	1.5	1.7	1.9	2.4	5	2.7	2.7	2.0	4.2	3.7	6	1	2
10.7	5.6	4.5	2.8	5	1.8	1.4	1.8	2.2	2.4	4	2.7	2.6	2.0	5.0	3.5	8	1	2
11.7	5.5	4.3	2.6	5	1.7	1.5	1.8	1.9	2.4	5	2.6	2.5	1.9	4.6	3.4	8	1	2
12.8	5.7	4.8	2.8	5	1.6	1.4	1.7	1.9	2.3	5	2.3	2.5	1.9	5.0	3.1	8	1	2

(a) Descreva sucintamente as principais características do feixe de vectores que representa as 19 variáveis (centradas,*mas não normalizadas*) no espaço $\mathcal{R}^{40}$.

(b) Efectue uma Análise de Componentes Principais sobre a matriz de correlações dos dados.

i. Procure interpretar o significado das três primeiras componentes principais, à luz da informação disponível. Justifique.

ii. Diga se considera adequada uma representação gráfica bi-dimensional. Justifique. Identifique uma variável cuja representação no primeiro plano principal seja menos fidedigna. Justifique.

iii. Na projecção da nuvem de pontos no plano definido pelos dois primeiros eixos principais aparece com alguma nitidez uma arrumação dos 40 indivíduos em grupos. Relacione essa arrumação com as variáveis originais e comente.

iv. Não é muito frequente encontrar conjuntos de dados com 19 variáveis para os quais uma ACP sobre a matriz de correlações explique uma proporção tão elevada da variabilidade total nas 2 ou 3 primeiras CPs. O que pensa que pode justificar este facto, neste conjunto de dados?

v. Avalie criticamente a ACP efectuada neste conjunto de 19 variáveis, tendo em atenção a natureza de (algumas) delas. Sugira alternativas, no caso de considerar haver algum aspecto indesejável.

3 Exercícios de Análise Discriminante

13. Um estudo referido no livro de D.F. Morrison, *Multivariate Statistical Methods* (p.288), envolve nove variáveis morfométricas sobre crânios de lobos *Canis lupus* L.. São efectuadas medições sobre 25 individuos, repartidos por 4 grupos: 6 machos das Montanhas Rochosas, 3 fêmeas das Montanhas Rochosas, 10 machos do Ártico, 6 fêmeas do Ártico. Os dados obtidos (todas as variáveis em mm.) encontram-se na *data frame* `lobos`, e são reproduzidos na tabela seguinte.

X1	X2	X3	X4	X5	X6	X7	X8	X9	Grupo
126	104	141	81.0	31.8	65.7	50.9	44.0	18.2	1
128	111	151	80.4	33.8	69.8	52.7	43.2	18.5	1
126	108	152	85.7	34.7	69.1	49.3	45.6	17.9	1
125	109	141	83.1	34.0	68.0	48.2	43.8	18.4	1
126	107	143	81.9	34.0	66.1	49.0	42.4	17.9	1
128	110	143	80.6	33.0	65.0	46.4	40.2	18.2	1
116	102	131	76.7	31.5	65.0	45.4	39.0	16.8	2
120	103	130	75.1	30.2	63.8	44.4	41.1	16.9	2
116	103	125	74.7	31.6	62.4	41.3	44.2	17.0	2
117	99	134	83.4	34.8	68.0	40.7	37.1	17.2	3
115	100	149	81.0	33.1	66.7	47.2	40.5	17.7	3
117	106	142	82.0	32.6	66.0	44.9	38.2	18.2	3
117	101	144	82.4	32.8	67.5	45.3	41.5	19.0	3
117	103	149	82.8	35.1	70.3	48.3	43.7	17.8	3
119	101	143	81.5	34.1	69.1	50.1	41.1	18.7	3
115	102	146	81.4	33.7	66.4	47.7	42.0	18.2	3
117	100	144	81.3	37.2	66.8	41.4	37.6	17.7	3
114	102	141	84.1	31.8	67.8	47.8	37.8	17.2	3
110	94	132	76.9	30.1	62.1	42.0	40.4	18.1	3
112	94	134	79.5	32.1	63.3	44.9	42.7	17.7	4
109	91	133	77.9	30.6	61.9	45.2	41.2	17.1	4
112	99	139	77.2	32.7	67.4	46.9	40.9	18.3	4
112	99	133	78.5	32.5	65.5	44.2	34.1	17.5	4
113	97	146	84.2	35.4	68.7	51.0	43.6	17.2	4
107	97	137	78.1	30.7	61.6	44.9	37.3	16.5	4

(a) Efectue uma Análise Discriminante Linear, usando o comando `lda` do módulo `MASS` do `R`.

i. Qual é a primeira variável discriminante (canónica)? Qual a sua capacidade discriminante? Comente.

ii. Utilize o comando `plot` para visualizar as nuvens de pontos sobre os planos definidos pelos três eixos discriminantes. Comente os resultados.

iii. A qual das 4 classes associaria um novo conjunto de observações, respeitantes a um lobo de sexo e habitat desconhecidos, com os seguintes valores para as 9 variáveis: 125, 104, 145, 81.1, 33.2, 68.2, 49.0, 43.3, 18.2? Utilize o comando `predict`, que tem um método para objectos de classe `lda`.

(b) Efectue uma Análise em Componentes Principais sobre o conjunto das observações dos 25 individuos nas 9 variáveis numéricas. Em particular, analise os planos definidos por cada par de CPs. Compare com os resultados obtidos na ADL. Comente a capacidade discriminante das CPs.

14. Efectue uma Análise Discriminante sobre o conjunto dos 150 lírios da *data frame* `iris`, a fim de estudar as funções discriminantes de Fisher para as três variedades. Emparticular,

(a) Utilize os primeiros 40 indivíduos de cada espécie para definir os eixos discriminantes.

(b) Classifique os 30 restantes indivíduos (o conjunto de validação) utilizando os eixos discriminantes definidos na alínea anterior (pode utilizar a classificação feita pelo comando `predict` do `R`).

(c) Construa uma tabela comparando as espécies reais destas 30 observações do conjunto de validação com as classificações obtidas pela Análise Discriminante Linear. Comente.

(d) Compare com a projecção desses mesmos 150 indivíduos sobre o primeiro plano principal, definido por uma Análise em Componentes Principais dos dados. Comente.

15. Dez zebus e dez charolesas foram observados em três variáveis (v_1, v_2 e v_3). Os valores obtidos (dados de Diday *et. al.*, 1982, disponíveis na *data frame* `diday`) foram:

Zebus			Charolesas		
v_1	v_2	v_3	v_1	v_2	v_3
400	224	28.2	395	224	35.1
395	229	29.4	410	232	31.9
395	219	29.7	405	233	30.7
395	224	28.6	405	240	30.4
400	223	28.5	390	217	31.9
400	224	27.8	415	243	32.1
400	221	26.5	390	229	32.1
410	233	25.9	405	240	31.1
402	234	27.1	420	234	32.4
400	223	26.8	390	223	33.8

Efectue uma Análise Discriminante dos dados e diga se considera que as 3 variáveis observadas permitem uma boa discriminação entre zebus e charolesas.

16. Considere os dados `videiras`, estudados no Módulo II da disciplina, com medições de área foliar, e comprimentos das nervuras principal, lateral esquerda e lateral direita, para $n = 200$ folhas de cada uma de três castas.

(a) Efectue uma Análise Discriminante Linear, procurando discriminar as castas com base nas 4 variáveis numéricas observadas. Comente o resultado.

(b) Confirme que os vectores de coeficientes (*loadings*) dos eixos discriminantes não são ortogonais entre si, mas que as novas variáveis discriminantes (vectores de *socres*) são não correlacionados entre si. **Nota:** No `R`, os vectores de *loadings* podem ser obtidos aplicando a função `coef` ao resultado da função `lda`; os vectores de *scores* resultam de aplicar a função `predict` ao resultado da ADL, e seleccionando o objecto `x`.

17. Escreva uma função no `R` para fazer uma ADL. Esta função deve aceitar como argumentos:

    - uma matriz ou *data frame* com os valores das variáveis;
    - um vector (`factor`) indicando a qual de k grupos pertence cada observação.

    A função deve construir e devolver:

    - a matriz da variabilidade inter-classes $\mathbf{B}$;
    - a matriz da variabilidade intra-classes $\mathbf{W}$;
    - os valores e vectores próprios de $\mathbf{W}^{-1}\mathbf{B}$;
    - os eixos discriminantes (ou seja, as $k-1$ combinações lineares das variáveis centradas, definidos pelos vectores próprios de $\mathbf{W}^{-1}\mathbf{B}$ associados a valores próprios não nulos).

    Caso $k > 1$, a função deve ainda devolver:

    - os centros de gravidade das k nuvens de pontos de cada grupo, nos eixos discriminantes.
    - as matrizes de covariâncias para cada grupo, nos eixos discriminantes.

    Nota: A matriz $\mathbf{W}^{-1}\mathbf{B}$ não é simétrica, pelo que a utilização do comando `eigen` pode produzir valores e vectores próprios (artificialmente) complexos. Pode ser usada a função `Re` para extrair a parte real desses números numericamente complexos.