INSTITUTO SUPERIOR DE AGRONOMIA

Modelos Matemáticos e Aplicações– 2015/16
Algumas resoluções de Exercícios de Estatística Multivariada

22. As diferentes unidades de medida das variáveis neste conjunto de dados `trigo` desaconselham uma ACP sobre a matriz de covariâncias.

(a) Eis o ajustamento da ACP sobre a matriz de correlações:

```
> trigo.acpR <- prcomp(trigo, scale=T)
> trigo.acpR
Standard deviations:
[1] 1.8435699 1.0078249 0.5428495 0.4008444 0.3608009

Rotation:
          PC1         PC2        PC3        PC4        PC5
x1  0.2927266 -0.80934768  0.2307521 -0.1736494 -0.4193647
x2 -0.4230070 -0.48228554 -0.6774251  0.3384100  0.1226362
x3  0.4996517 -0.03972188 -0.5052573 -0.5846198  0.3894934
x4 -0.4829956 -0.28611627  0.4324574 -0.3646951  0.6040123
x5 -0.5024338  0.17004891 -0.2134109 -0.6168807 -0.5408860
```

Cada coluna da matriz "`Rotation`" tem os coeficientes que definem cada CP (ou seja, que definem cada uma das combinações lineares das 5 variáveis originais $x1$ a $x5$).

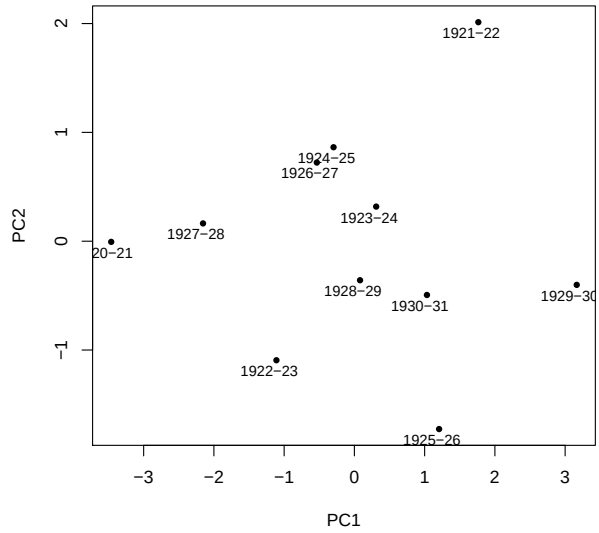
A proporção de variabilidade, e variabilidade cumulativa, explicada pelas cinco CPs é esta:

```
> summary(trigo.acpR)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     1.8436 1.0078 0.54285 0.40084 0.36080
Proportion of Variance 0.6797 0.2031 0.05894 0.03214 0.02604
Cumulative Proportion  0.6797 0.8829 0.94183 0.97396 1.00000
```

Assim, a redução para duas dimensões pode ser feita preservando mais de 88% da variabilidade total (inércia), percentagem que se eleva para quase 95% com 3 dimensões. Embora a dimensão inicial não fosse muito grande (a representação tradicional seria uma nuvem de $n=11$ pontos em $\mathbb{R}^5$), a redução de dimensionalidade efectuada com a ACP vai permitir a visualização em $\mathbb{R}^2$ (ou $\mathbb{R}^3$) do fundamental da inércia dessa nuvem de pontos.

(b) A nuvem de $n=11$ pontos, quando projectada no primeiro plano principal, é esta:

```
> plot(trigo.acpR$x, pch=16, cex=0.8)
> text(trigo.acpR$x-0.1, label=rownames(trigo), cex=0.8)
```



O pequeno número de observações permite que, neste caso, se interpretem as CPs com base nas observações mais extremas em cada CP. Inspeccionando esta nuvem projectada, vemos como em extremos opostos da primeira Componente Principal encontram-se os anos 1920-21 e 1929-30. Voltando aos dados originais, percebe-se que no primeiro caso temos uma campanha caracterizada por tempo seco (em Novembro-Dezembro, mas sobretudo em Julho, quase sem precipitação) e relativamente quente e com radiação elevada, quando comparada com a campanha 1929-30. O rendimento em 1920-21 e também (*ex-aequo* com 1927-28) o mais elevado de todos. A campanha de 1929-30, de tempo mais chuvoso (sobretudo em Julho) e frio, caracteriza-se também pelo menor rendimento das 11 campanhas registadas. A segunda CP distancia os anos de 1921-22 e 1925-26. O primeiro destes anos caracteriza-se por ser o ano com menor temperatura média em Julho, enquanto que 1925-26 caracteriza-se por ser o ano mais chuvoso em Novembro-Dezembro.

(c) A discussão do ponto anterior é completada com o pedido deste ponto. Eis os coeficientes de correlação entre variáveis originais e CPs da matriz de correlações:

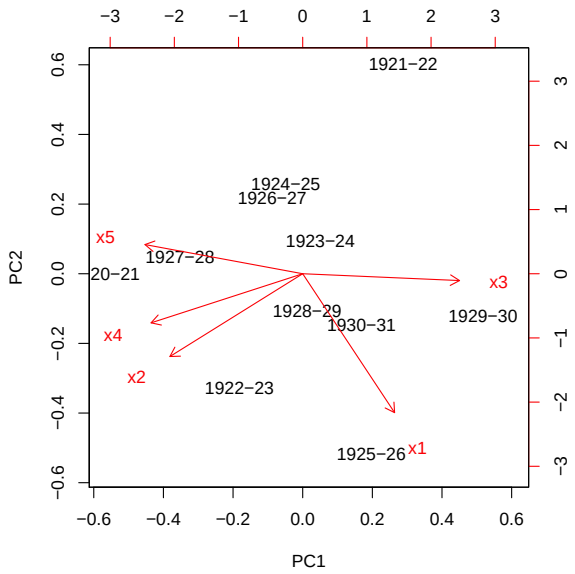
```
> round(cor(trigo, trigo.acpR$x), d=2)
     PC1   PC2   PC3   PC4   PC5
x1  0.54 -0.82  0.13 -0.07 -0.15
x2 -0.78 -0.49 -0.37  0.14  0.04
x3  0.92 -0.04 -0.27 -0.23  0.14
x4 -0.89 -0.29  0.23 -0.15  0.22
x5 -0.93  0.17 -0.12 -0.25 -0.20
```

Como seria de esperar, dadas as propriedades das CPs sobre os dados normalizados, a primeira CP tem correlações (em módulo) importantes com todas a variáveis, mas sobretudo com $x5$ (rendimento médio) e $x3$ (precipitação total em Julho). O facto destas correlações terem sinais opostos significa que a CP contrasta observações com altos rendimentos e baixa precipitação em Julho (como 1920-21) e, no outro extremo, observações com baixos rendimentos e Julhos chuvosos (como 1929-30).

Em relação à segunda CP, a correlação mais importante é com a variável $x1$ (precipitação em Novembro-Dezembro) e, em bastante menor medida, $x2$ (temperatura em Julho). Havendo

iguais sinais nestas duas correlações, a CP vai contrastar observações com valores elevados de ambas (como 1925-26) e observações com valores baixos de ambas (como 1921-22).
Eis o *biplot* resultante, que confirma os comentários acima feitos:

```
> biplot(trigo.acpR)
```



(d) As duas alterações referidas são mudanças lineares (afins) de escala, pelo que não afectam os resultados duma ACP sobre a matriz de correlações. Vejamos:

```
> trigo2 <- trigo
> trigo2[,4] <- trigo[,4]*0.75518263-0.02960342
> trigo2[,5] <- trigo[,5]/10
> trigo2
           x1   x2    x3        x4    x5
1920-21  87.9 19.6   1.0 1254.3287 2.837
1921-22  89.9 15.2  90.1  730.9872 2.377
1922-23 153.0 19.7  56.6 1021.7325 2.604
1923-24 132.1 17.0  91.0  976.4215 2.574
1924-25  88.8 18.3  93.7  870.6960 2.668
1925-26 220.9 17.8 106.9  971.1353 2.429
1926-27 117.7 17.8  65.5  833.6920 2.800
1927-28 109.0 18.3  41.8 1188.6279 2.837
1928-29 156.1 17.8  57.4  922.8036 2.496
1929-30 181.5 16.8 140.6  681.1451 2.166
1930-31 181.4 17.0  74.3  868.4304 2.437

> prcomp(trigo2, scale=T)
Standard deviations:
[1] 1.8435699 1.0078249 0.5428495 0.4008444 0.3608009

Rotation:
          PC1         PC2        PC3        PC4        PC5
x1  0.2927266 -0.80934768  0.2307521 -0.1736494 -0.4193647
```

```
x2 -0.4230070 -0.48228554 -0.6774251  0.3384100  0.1226362
x3  0.4996517 -0.03972188 -0.5052573 -0.5846198  0.3894934
x4 -0.4829956 -0.28611627  0.4324574 -0.3646951  0.6040123
x5 -0.5024338  0.17004891 -0.2134109 -0.6168807 -0.5408860
```

Assinale-se que, no caso de se ter optado por uma ACP sobre a matriz de covariâncias, estas mudanças lineares (afins) de escala iriam alterar os resultados.

23. Este exercício visa sobretudo chamar a atenção para alguns aspectos da ACP que podem suscitar confusão.

    (a) Embora uma ACP sobre a matriz de covariâncias não seja a opção mais adequada, dado haver variáveis com diferentes unidades de medida, faremos como é pedido no enunciado (e como foi feito por Kendall, no seu livro). Eis a cabeça da *data frame* e a síntese da variabilidade explicada pels CPs, na referida ACP:

```
> head(kendall)
  areia limo argila mat.org acidez
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

> summary(prcomp(kendall))
Importance of components:
                          PC1    PC2     PC3     PC4       PC5
Standard deviation     14.9613 2.8667 0.68736 0.50838 3.368e-15
Proportion of Variance  0.9616 0.0353 0.00203 0.00111 0.000e+00
Cumulative Proportion   0.9616 0.9969 0.99889 1.00000 1.000e+00
```

    Como se pode constatar, o último valor próprio da matriz de (co-)variâncias dos dados é nulo. Esse facto reflecte a existência duma dependência linear nas colunas da matriz de dados (multicolinearidade entre as variáveis): a presença das três categorias na estrutura, em percentagem, dos solos (teor arenoso, limoso e argiloso) significa que a soma das três primeiras colunas da matriz de dados dá sempre 100%. Comprovemos com o `R`:

```
> apply(kendall[,1:3],1,sum)
[1] 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
```

    Esta dependência linear exacta significa que qualquer vector que seja múltiplo escalar do vector $\vec{\mathbf{v}} = (1,1,1,0,0)^t$ anula a combinação linear $\mathbf{X}\vec{\mathbf{v}}$, e necessariamente também a combinação linear correspondente das colunas da matriz centrada: $\mathbf{X}^{\mathrm{c}}\vec{\mathbf{v}} = \mathbf{0}_n$. Logo, $\vec{\mathbf{v}}$ é vector próprio da matriz de covariâncias $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{\mathrm{c}t}\mathbf{X}^{\mathrm{c}}$, com valor próprio nulo, uma vez que:

$$\mathbf{S}\vec{\mathbf{v}} \;=\; \tfrac{1}{n-1}\mathbf{X}^{\mathrm{c}t}\underbrace{\mathbf{X}^{\mathrm{c}}\vec{\mathbf{v}}}_{=\mathbf{0}_n} \;=\; \mathbf{0}_p \qquad \Leftrightarrow \qquad \mathbf{S}\vec{\mathbf{v}} \;=\; 0\cdot\vec{\mathbf{v}}\ .$$

    O correspondente vector próprio tem de ser da forma $\alpha\vec{\mathbf{v}} = (\alpha,\alpha,\alpha,0,0)^t$. Para ser de norma 1, tem de ter-se $\alpha = \pm\frac{1}{\sqrt{3}} = \pm 0.5773503$. Confirmemos:

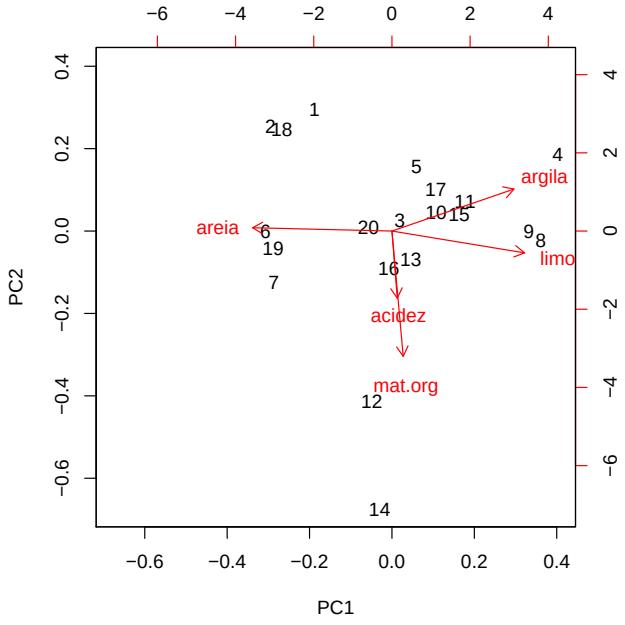
```
> prcomp(kendall)
[...]
```

```
Rotation:
                 PC1           PC2          PC3           PC4           PC5
areia   -0.7849449544 -0.223100113  0.02718830 -0.003901445 -5.773503e-01
limo     0.5870812022 -0.560792619  0.08607941 -0.010212850 -5.773503e-01
argila   0.1978637522  0.783892732 -0.11326771  0.014114294 -5.773503e-01
mat.org  0.0068110819 -0.145758773 -0.97995170  0.135656360  7.806256e-18
acidez   0.0007907581 -0.002131353 -0.13680723 -0.990595081 -5.681219e-17
```

A existência desta multicolinearidade não é um problema: é sempre possível excluir uma (ou mais, se necessário) variáveis do conjunto de dados, para eliminar a dependência linear nas colunas. No nosso exemplo, será preciso eliminar uma das três colunas correspondentes à composição dos solos. Repare-se que a eliminação de uma das outras duas colunas não só não resolve o problema da dependência linear, como conduz à perda de informação no conjunto de dados. Pelo contrário, a exclusão de uma das três primeiras colunas não perde informação, uma vez que é sempre possível recuperar o teor excluído a partir do conhecimento das outras duas variáveis.

(b) O comando `R` que produz o *biplot* pedido é:

```
> biplot(prcomp(kendall, scale=TRUE))
```



No gráfico obtido, os vectores correspondentes às variáveis `acidez` e `mat.org` surgem como praticamente colineares, o que sugere uma fortíssima correlação entre estas variávies (de sinal positivo, já que o sentido dos vectores é igual). No entanto, uma inspecção rápida à matriz de correlações dos dados mostra que essa correlação é, na realidade, quase nula:

```
> round(cor(kendall),d=3)
         areia   limo argila mat.org acidez
areia    1.000 -0.972 -0.828  -0.100 -0.023
limo    -0.972  1.000  0.674   0.213  0.024
argila  -0.828  0.674  1.000  -0.196  0.013
```

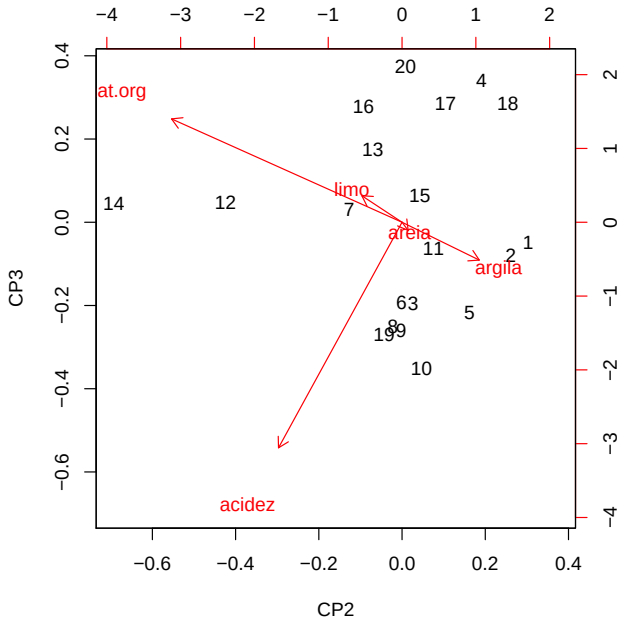
```
mat.org -0.100  0.213 -0.196   1.000  0.079
acidez  -0.023  0.024  0.013   0.079  1.000
```

Esta aparente contradição tem de significar que a aproximação bidimensional distorce bastante a relação entre os vectores representativos destas duas variáveis. Esta conclusão é também sustentada pelo comprimento bastante menor do vector que serve de marcador da variável `acidez`, quando comparado com os restantes vectores marcadores de variáveis. Recorde-se que, numa ACP sobre os dados normalizados, todas as variáveis têm variância 1, pelo que os respectivos vectores marcadores são de igual comprimento na representação em todas as dimensões. Pode confirmar-se esta afirmação através do cálculo dos coeficientes de correlação entre as variáveis originais e as CPs sobre os dados normalizados, que evidencia a forte correlação entre a variável `acidez` e a terceira CP:

```
> kendall.acpR <- prcomp(kendall, scale=TRUE)
> round(cor(kendall, kendall.acpR$x),d=3)
           PC1    PC2    PC3    PC4    PC5
areia   -0.996  0.024 -0.032 -0.082 -0.799
limo     0.948 -0.156  0.103  0.256  0.863
argila   0.873  0.300 -0.147 -0.356  0.456
mat.org  0.080 -0.896  0.402 -0.173  0.465
acidez   0.038 -0.480 -0.876  0.013  0.086
```

De forma mais elaborada, pode construir-se um *biplot* que utilize os marcadores de indivíduos e variáveis, não nas duas primeiras dimensões, mas na segunda e terceira (vemos pela matriz de correlações cruzadas que a variável `mat.org` está fortemente correlacionada com a CP2). Eis os comandos necessários (veja nos acetatos relativos ao *biplot* a forma de construir os marcadores de variáveis e indivíduos). O gráfico pode ser produzido pelo comando `biplot` do `R`, que aceita como primeiro argumento uma matriz de duas colunas com os marcadores de indivíduos (em baixo, as colunas 2 e 3 da matriz `kend.G`) e outra matriz de duas colunas com os marcadores de variáveis (colunas 2 e 3 da matriz `kend.H`).

```
> kendall.svd <- svd(scale(kendall))
> kend.G <- kendall.svd$u
> kend.H <- kendall.svd$v %*% diag(kendall.svd$d)
> colnames(kend.G) <- paste("CP",1:5,sep="")
> colnames(kend.H) <- paste("CP",1:5,sep="")
> rownames(kend.H) <- colnames(kendall)
> biplot(kend.G[,2:3], kend.H[,2:3])
```



Assinale-se que este *biplot* não deve ser usado para inspeccionar as principais características do conjunto de dados, uma vez que as CPs associadas (segunda e terceira) apenas explicam cerca de 40% da variabilidade total (inércia) dos dados. Mas é um instrumento que ilustra melhor a relação entre as variáveis `acidez` e `mat.org`.

(c) É pedido para repetir a ACP sobre a matriz de covariâncias, mas agora excluindo a primeira variável (`areia`), a fim de eliminar a multicolinearidade presente nos dados. Eis os resultados:

```
> kend2.acp <- prcomp(kendall[,-1])
> summary(kend2.acp)
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     9.3081 2.66338 0.68659 0.50837
Proportion of Variance 0.9172 0.07509 0.00499 0.00274
Cumulative Proportion  0.9172 0.99227 0.99726 1.00000
```

Tal como na análise inicial, com base nas 5 variáveis, as duas primeiras CPs são suficientes para explicar quase toda a variabilidade presente nos dados, o que não surpreeende: confirma-se que a nuvem de $n = 20$ pontos (agora em $\mathbb{R}^4$) é essencialmente bidimensional, ou seja, encontra-se aproximadamente num plano.

i. Eis os coeficientes de correlação entre cada variável original e cada CP, que evidenciam como a primeira CP está fortemente correlacionada com o teor limoso e a última CP com a acidez. De forma menos enfática, a terceira CP está bem correlacionada com a matéria orgânica. A segunda CP, de interpretação mais difícil, parece contrastar teor argiloso e matéria orgânica.

```
> round(cor(kendall[,-1], kend2.acp$x),d=3)
          PC1    PC2    PC3    PC4
limo    0.996 -0.086 -0.005  0.000
argila  0.735  0.677  0.026 -0.002
```

```
mat.org 0.174 -0.511  0.838 -0.086
acidez  0.024 -0.010  0.184  0.983
```

ii. A matriz cujas colunas são os coeficientes nas combinações lineares que definem as CPs (vectores de *loadings*, ou seja, vectores próprios da matriz de variâncias-covariâncias dos dados) é indicada de seguida:

```
> kend2.acp$rot
                PC1         PC2         PC3          PC4
limo    0.955785266 -0.28801387 -0.05901046  0.006348304
argila  0.293681185  0.94519099  0.14154620 -0.018166560
mat.org 0.014971757 -0.15381233  0.97857851 -0.136021005
acidez  0.001316516 -0.00194084  0.13735552  0.990519036
```

Caso fosse feita uma interpretação sumária das CPs, baseada apenas nestes coeficientes (como é prática corrente), haveria que associar cada CP a uma das variáveis (CP1 a `limo`, CP2 a `argila`, CP3 a `mat.org` e CP4 a `acidez`). Mas como se viu acima, estas interpretações não estão correctas (veja-se o caso da segunda CP). Aliás, esta associação de "a cada CP a sua variável" só poderia ser verdade se as variáveis originais fossem aproximadamente não correlacionadas entre si, uma vez que as CPs, por construção, são de correlação nula entre si. Mas a inspecção da matriz de correlações entre as 5 variáveis originais, dada no início, mostra que assim não é. Esta alínea ilustra os perigos de interpretações de CPs baseadas apenas nos vectores de *loadings*.

24. (a) A fim de visualizar o feixe de vectores que representa as 19 variáveis (centradas, mas não normalizadas) no espaço das variáveis, será necessário inspeccionar as variâncias de cada variável, bem como a matriz de correlações entre cada par de variáveis:

```
> diag(var(adelges))
    length      width    forwing    hinwing     spirac    antseg1    antseg2
14.1393590  4.0516923  1.6807628  0.8276667  0.1121795  0.1075321  0.1135321
  antseg3    antseg4    antseg5    antspin    tarsus3     tibia3     femur3
0.2235833  0.1485897  0.1483077  1.3326923  0.4122821  0.5789167  0.3435833
  rostrum     ovipos     ovspin       fold      hooks
0.7893333  0.3474615  3.8051282  0.2044872  0.2532051

> round(cor(adelges), d=2)
        length width forwing hinwing spirac antseg1 antseg2 antseg3 antseg4
length    1.00  0.93    0.93    0.91   0.52    0.80    0.85    0.79    0.84
width     0.93  1.00    0.94    0.94   0.49    0.82    0.86    0.83    0.86
forwing   0.93  0.94    1.00    0.93   0.54    0.86    0.89    0.85    0.86
hinwing   0.91  0.94    0.93    1.00   0.50    0.83    0.89    0.88    0.85
spirac    0.52  0.49    0.54    0.50   1.00    0.70    0.72    0.25    0.46
antseg1   0.80  0.82    0.86    0.83   0.70    1.00    0.92    0.70    0.75
antseg2   0.85  0.86    0.89    0.89   0.72    0.92    1.00    0.75    0.79
antseg3   0.79  0.83    0.85    0.88   0.25    0.70    0.75    1.00    0.75
antseg4   0.84  0.86    0.86    0.85   0.46    0.75    0.79    0.75    1.00
antseg5   0.85  0.88    0.86    0.88   0.57    0.84    0.91    0.79    0.80
antspin  -0.46 -0.50   -0.52   -0.49  -0.17   -0.32   -0.38   -0.50   -0.36
tarsus3   0.92  0.94    0.94    0.95   0.52    0.85    0.91    0.86    0.85
tibia3    0.94  0.96    0.96    0.95   0.49    0.85    0.91    0.88    0.88
femur3    0.95  0.95    0.95    0.95   0.45    0.82    0.89    0.88    0.88
rostrum   0.90  0.90    0.88    0.91   0.55    0.83    0.89    0.79    0.82
ovipos    0.69  0.65    0.69    0.62   0.81    0.81    0.86    0.41    0.62
ovspin    0.33  0.31    0.36    0.27   0.75    0.55    0.57    0.07    0.30
fold     -0.68 -0.71   -0.67   -0.74  -0.23   -0.50   -0.50   -0.76   -0.67
hooks     0.70  0.73    0.75    0.78   0.29    0.50    0.59    0.79    0.67
```

	antseg5	antspin	tarsus3	tibia3	femur3	rostrum	ovipos	ovspin	fold	hooks
length	0.85	-0.46	0.92	0.94	0.95	0.90	0.69	0.33	-0.68	0.70
width	0.88	-0.50	0.94	0.96	0.95	0.90	0.65	0.31	-0.71	0.73
forwing	0.86	-0.52	0.94	0.96	0.95	0.88	0.69	0.36	-0.67	0.75
hinwing	0.88	-0.49	0.95	0.95	0.95	0.91	0.62	0.27	-0.74	0.78
spirac	0.57	-0.17	0.52	0.49	0.45	0.55	0.81	0.75	-0.23	0.29
antseg1	0.84	-0.32	0.85	0.85	0.82	0.83	0.81	0.55	-0.50	0.50
antseg2	0.91	-0.38	0.91	0.91	0.89	0.89	0.86	0.57	-0.50	0.59
antseg3	0.79	-0.50	0.86	0.88	0.88	0.79	0.41	0.07	-0.76	0.79
antseg4	0.80	-0.36	0.85	0.88	0.88	0.82	0.62	0.30	-0.67	0.67
antseg5	1.00	-0.37	0.90	0.90	0.89	0.85	0.71	0.38	-0.63	0.67
antspin	-0.37	1.00	-0.47	-0.45	-0.44	-0.40	-0.20	-0.03	0.49	-0.42
tarsus3	0.90	-0.47	1.00	0.98	0.97	0.91	0.72	0.40	-0.66	0.70
tibia3	0.90	-0.45	0.98	1.00	0.99	0.92	0.71	0.36	-0.66	0.72
femur3	0.89	-0.44	0.97	0.99	1.00	0.92	0.68	0.30	-0.69	0.73
rostrum	0.85	-0.40	0.91	0.92	0.92	1.00	0.72	0.38	-0.63	0.69
ovipos	0.71	-0.20	0.72	0.71	0.68	0.72	1.00	0.78	-0.19	0.29
ovspin	0.38	-0.03	0.40	0.36	0.30	0.38	0.78	1.00	0.17	-0.03
fold	-0.63	0.49	-0.66	-0.66	-0.69	-0.63	-0.19	0.17	1.00	-0.77
hooks	0.67	-0.42	0.70	0.72	0.73	0.69	0.29	-0.03	-0.77	1.00

No feixe de vectores que representa as 19 variáveis, as variâncias são proporcionais ao quadrado do comprimento dos vectores. Assim, o vector associado à variável comprimento (`length`, ou `COM` no enunciado) é claramente mais comprido do que os restantes, seguido do comprimento dos vectores associados à largura (`width`, ou `LAR` no enunciado) e número de sedas do oviescapto (`ovspin`, ou `N` no enunciado). O feixe de vectores tem numerosos vectores que apontam em sentidos próximos, reflectindo as elevadas correlações existentes entre numerosos pares de variáveis. De facto, mais de metade (72) dos 171 diferentes pares de variáveis que podem ser formados a partir das 19 variáveis existentes têm correlações superiores a 0.8. Há duas variáveis, o número de sedas antenais (`antspin`, ou `S` no enunciado) e a existência ou não de prega anal (`fold`, ou `P` no enunciado), que têm numerosas correlações negativas com as restantes variáveis, pelo que as respectivas direcções no espaço das variáveis $\mathbb{R}^{40}$ são bastante diferentes das restantes variáveis.

Assim, pode adivinhar-se que uma ACP sobre a matriz de (co-)variâncias terá a primeira CP fortemente associada à variável comprimento, enquanto que a primeira CP numa ACP sobre a matriz de correlações irá privilegiar o feixe de variáveis fortemente correlacionadas.

(b) Eis a ACP sobre a matriz de correlações dos dados `adelges`:

```
> adel.acpR <- prcomp(adelges, scale=T)
> summary(adel.acpR)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     3.7230 1.5394 0.86515 0.70814 0.52743 0.51510 0.43969
Proportion of Variance 0.7295 0.1247 0.03939 0.02639 0.01464 0.01396 0.01018
Cumulative Proportion  0.7295 0.8542 0.89363 0.92002 0.93466 0.94863 0.95880
                          PC8    PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.3972 0.3750 0.35050 0.30409 0.27095 0.24462 0.20495
Proportion of Variance 0.0083 0.0074 0.00647 0.00487 0.00386 0.00315 0.00221
Cumulative Proportion  0.9671 0.9745 0.98097 0.98584 0.98970 0.99285 0.99506
```

A primeira CP corresponde a cerca de 73% da inércia total, as duas primeiras CPs a mais de 85% da inércia total, e as três primeiras CPs a quase 90% da inércia total.

i. Eis as correlações entre as três primeiras CPs sobre os dados normalizados e as 19 variáveis originais:

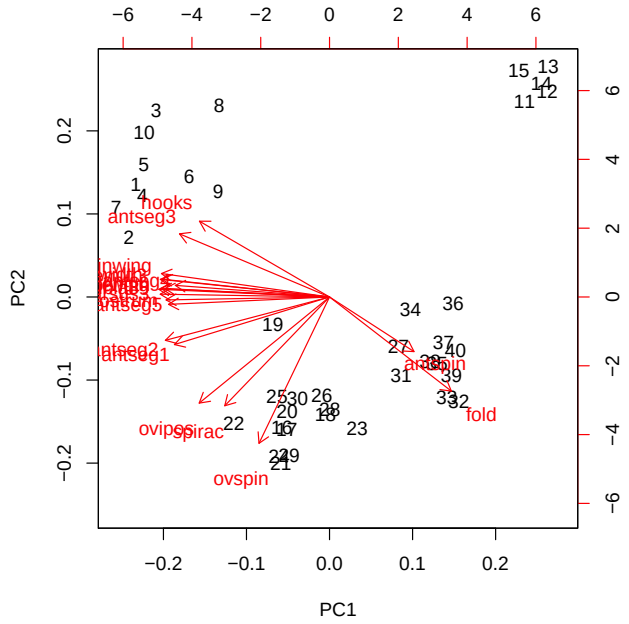
```
> round(cor(adel.acpR$x[,1:3], adelges), d=2)
    length width forwing hinwing spirac antseg1 antseg2 antseg3 antseg4 antseg5
PC1  -0.95 -0.96   -0.97   -0.97  -0.60   -0.89   -0.94   -0.86   -0.89   -0.93
PC2   0.05  0.10    0.05    0.13  -0.62   -0.27   -0.25    0.36    0.07   -0.04
PC3   0.02  0.01   -0.05    0.03  -0.16    0.03    0.00    0.05    0.14    0.09
    antspin tarsus3 tibia3 femur3 rostrum ovipos ovspin  fold hooks
PC1    0.49   -0.97  -0.98  -0.97   -0.94  -0.75  -0.41  0.70 -0.75
PC2   -0.31    0.02   0.04   0.10   -0.02  -0.61  -0.84 -0.54  0.44
PC3    0.80    0.03   0.07   0.10    0.07  -0.02  -0.13  0.04  0.05
```

A primeira CP tem fortes correlações (o sinal negativo é arbitrário) com quase todas as variáveis, sendo uma espécie de medida da dimensão global dos organismos de afídios. As mais baixas correlações dizem respeito a variáveis que estão mais associadas a CPs posteriores. Esta íntima associação entre CP1 e muitas variáveis seria de esperar, dado o número elevado de variáveis fortemente correlacionadas entre si. Recorde-se que a primeira CP sobre a matriz de correlações é a combinação linear que maximiza a soma de quadrados das correlações com cada variável original (vejam-se os acetatos, na discussão do critério alternativo optimizado pelas CPs sobre a matriz de correlações). Esta soma de quadrados de correlações ao quadrado é o valor próprio associado à primeira CP, que pode ser calculado, com base na informação já disponível, de duas formas alternativas (e equivalentes, a menos de erros de arredondamento): ou directamente; ou através do quadrado do desvio padrão associado à primeira CP:

```
> sum(cor(adel.acpR$x[,1], adelges)^2)
[1] 13.8606

> (adel.acpR$sdev^2)[1]
[1] 13.8606
```

Repare-se ainda que, sendo $p = 19$ também o traço da matriz de correlações (soma dos seus elementos diagonais), podemos dizer que a média destas correlações ao quadrado é a proporção de variabilidade total explicada pela primeira CP que consta do resumo apresentado pelo comando `summary`: $\frac{13.8606}{19} = 0.7295055$.
A CP2 tem uma correlação bastante forte com a variável `ovspin` (número de sedas do oviescapto, ou `N` no enunciado) e, em menor medida, também com as outras três das quatro variáveis finais, bem como com a variável `spirac` (número de espiráculos, ou `E` no enunciado). A CP3 parece estar bem correlacionada apenas com a variável `antspin` (número de sedas antenais, ou `S` no enunciado).

ii. A representação bidimensional corresponde a cerca de 85% da variabilidade total, o que é muito, sobretudo considerando tratar-se duma redução de dimensionalidade de 19 para 2 dimensões. Eis o *biplot* resultante:



As fortes correlações entre a maioria das variáveis originais e a primeira CP dos dados normalizados é visível no *biplot* (com a ressalva da aproximação, cuja qualidade está associada aos 85% de inércia associada às duas primeiras dimensões) no facto de os vectores correspondentes a um grupo nutrido de variáveis serem quase horizontais. O sentido destes marcadores de variáveis no *biplot* é arbitrário. As correlações entre este grupo de variáveis também devem ser elevadas (o que se pode confirmar na matriz de correlações). Da mesma forma, os ângulos obtusos entre os vectores deste grupo de variáveis e os vectores associados às variáveis `fold (P)` e `antspin (S)` sugere correlações negativas, facto que é igualmente confirmável na matriz de correlações.
Os marcadores de variáveis (vectores) mais verticais estão associados às variáveis mais fortemente correlacionadas com a segunda CP (`ovspin`, `spirac` e `ovipos`). A projecção ortogonal dos marcadores de indivíduos sobre as direcções definidas por estes vectores produz uma reconstrução aproximada dos valores dos indivíduos nessas variáveis, e permite evidenciar a separação entre os indivíduos na parte inferior do gráfico e o grupo de 5 indivíduos no canto superior direito (com os valores de 11 a 15). Em particular, o grupo de 5 indivíduos do canto superior direito corresponde a indivíduos com valores abaixo da média nestas variáveis (embora o sentido de cada eixo do biplot seja arbitrário, uma vez definido esse sentido nos gráficos os indivíduos que ficam do lado da seta são indivíduos de valor acima da média e os do lado contrário são indivíduos abaixo da média - uma vez que o centro de gravidade foi trasladado para a origem). Inspeccionado os dados, pode confirmar-se que os indíviduos 11 a 15 têm o menor valor observado nas variáveis `ovspin (N)` e `spirac (E)` (4 para todos, nas duas variáveis). Na variável `ovipos (OVI)` os valores correspondentes oscilam entre 2.3 e 2.7, sendo os cinco menores valores observados entre os 40 indivíduos. Analogamente, os indivíduos na parte inferior do gráfico (com números como 21, 22, 24 ou 29) são indivíduos com o valor mais elevado (10) na variável `ovspin (N)`. Neste *biplot*, os vectores representativos de todas as variáveis deveriam ter igual comprimento (uma vez que todas as variáveis

normalizadas têm desvio padrão igual). Na medida em que haja vectores mais curtos, têm de corresponder a variáveis que ficam menos bem representadas na projecção nestas duas dimensões. Em particular, a variável `antspin (S)` é representada por um vector bastante mais curto, facto que sugere que parte importante da informação dada por esta variável não está bem reflectida a duas dimensões. Esta conclusão é coerente com a alta correlação (0.80) entre a variável em causa e a terceira CP, como se viu acima.

iii. A projecção sobre o primeiro plano principal *não* é a indicada no *biplot* (onde as distâncias não correspondem às habituais distâncias euclideanas entre indivíduos, mas sim às respectivas distâncias de Mahalanobis). Mas as duas configurações não são, neste como em muitos outros casos, substancialmente diferentes. São visíveis quatro grandes grupos de indivíduos (com números de 1 a 10; de 11 a 15; de 16 a 29 - excepto o 19 e 27; e finalmente o 27 e de 31 a 40), aparecendo o indivíduo 19 como isolado. A separação dos segundo e terceiro destes grandes grupos já foi discutida em cima. O primeiro e quarto dos grandes grupos parecem definir-se essencialmente pela dupla de variáveis `hook (GAP)` e `antseg3 (AS3)` para um lado (individuos 1 a 10 têm valores elevados nestas variáveis e indivíduos do grupo 27+(31-40) têm valores baixos), contra a dupla de variáveis `antspin (S)` e `fold (P)` (valores elevados dos indivíduos do quarto grupo e baixo dos individuos do primeiro grupo). O individuo isolado (19), que surge mais próximo da origem, parece ter valores mais próximos da média no conjunto das variáveis.

iv. A proporção relativamente elevada da variabilidade explicada pelas duas primeiras CPs da matriz de correlações (cerca de 85%) reflecte o facto de neste conjunto de dados haver muitas variáveis com fortes correlações entre si. Veja-se a discussão no primeiro dos sub-pontos desta alínea.

v. Entre as 19 variáveis deste conjunto de dados há uma (`fold (P)`) que é na realidade uma variável dicotómica, uma vez que indica se os afídeos observados tinham, ou não, prega anal. Uma tal variável é de duvidosa presença num conjunto de dados a submeter a uma ACP. Não sendo claramente um erro a sua inclusão, como seria o caso duma variável categórica com categorias de posição totalmente arbitrária (em última análise, a variável `fold (P)` pode ser vista como uma variável de contagem do número de pregas anais), a verdade é que a ACP parte do pressuposto que as variáveis em causa são plenamente numéricas, privilegiando não apenas a ordem dos valores, mas também as suas escalas. Problema análogo existe com a variável `spirac (E)`, uma variável de contagem, mas para a qual os afídeos observados apenas tomam valores 4 e 5. Após a normalização, estas duas variáveis têm uma natureza semelhante (e os seus valores médios e variâncias são expressos por fórmulas que apenas dependem do número de observações com cada um dos dois valores possíveis). Até considerando o papel importante destas duas variáveis na definição das CPs, poderia ser interessante repetir a ACP sem a presença dessas variáveis. Mas tal hipótese não significa que as variáveis *devam* ser excluídas: a sua presença ou ausência muda a informação disponível e portanto é natural que altere os resultados.

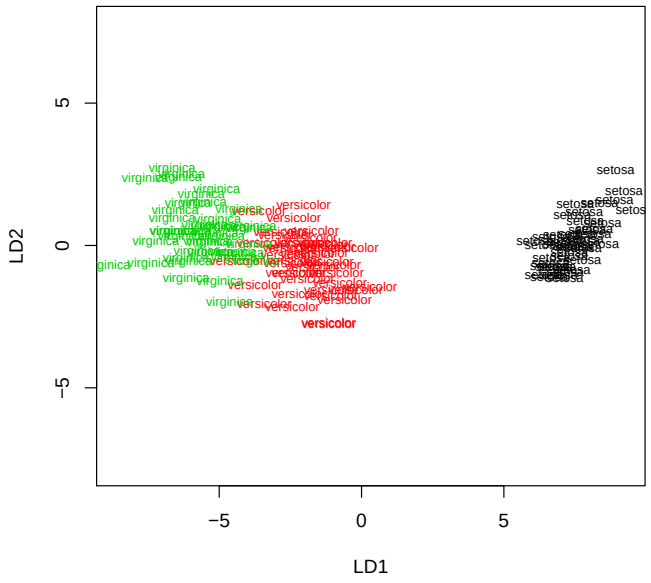
25. Feito nas aulas.

26. Considerando os dados da *data frame* `iris`:

    (a) Eis os comandos do `R` relevantes, e respectivos resultados:

```
> iris.lda <- lda(Species ~ .  , data=iris[c(1:40,51:90,101:140),])
```

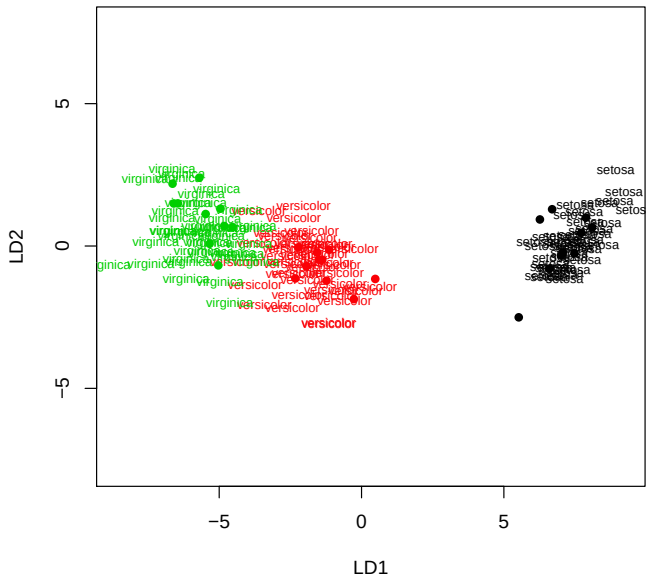
```
> iris.lda
Call: lda(Species ~ ., data = iris[c(1:40, 51:90, 101:140), ])
[...]
Coefficients of linear discriminants:
                    LD1        LD2
Sepal.Length  0.7863979 -0.4796486
Sepal.Width   1.5053009  2.6735737
Petal.Length -2.1434874 -0.1654659
Petal.Width  -2.7112210  1.7287670

Proportion of trace:
   LD1    LD2
0.9932 0.0068

> plot(iris.lda, col=as.numeric(iris[c(1:40, 51:90, 101:140),]$Species))
```



Havendo $k = 3$ grupos (as espécies dos lírios), não pode haver mais de $k - 1 = 2$ eixos com capacidade discriminante não nula. Embora a medida da qualidade dos eixos discriminantes usada pela função `lda` não corresponda directamente à que foi vista nas nossas aulas (em vez de indicar os valores próprios da matriz $\mathbf{W}^{-1}\mathbf{B}$, indica a sua proporção em relação à soma desses mesmos valores próprios), é evidente que apenas o primeiro eixo discriminante tem real capacidade discriminante, como fica patente na nuvem dos $3 \times 40 = 120$ pontos usados para definir os eixos discriminantes.

(b) Vejamos agora como ficariam, nos dois eixos discriminantes obtidos, os 30 lírios que foram deixados de fora do ajustamento. Na representação gráfica desses indivíduos do conjunto de validação, são usadas as cores das suas verdadeiras espécies (embora essa ingformação não seja usada para os posicionar nos eixos discriminantes), o que permite desde logo visualizar a qualidade da discriminação resultante.

```
> iris.ldaPred <- predict(iris.lda, new=data.frame(iris[c(41:50,91:100,141:150),]))
> points(iris.ldaPred$x, col=as.numeric(iris[c(41:50,91:100,141:150),]$Species), pch=16)
```



O bom resultado aparente no gráfico pode ser confirmado pedindo o objecto `class` da `list` produzida pelo comando `predict`, e será confirmado também na próxima alínea.

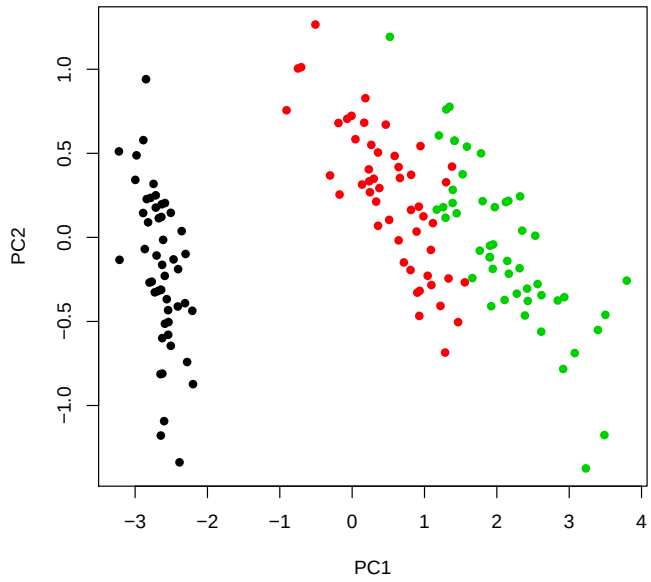
```
> iris.ldaPred$class
 [1] setosa     setosa     setosa     setosa     setosa     setosa
 [7] setosa     setosa     setosa     setosa     versicolor versicolor
[13] versicolor versicolor versicolor versicolor versicolor versicolor
[19] versicolor versicolor virginica  virginica  virginica  virginica
[25] virginica  virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

(c) Eis a tabela das classificações obtidas pelos eixos discriminantes que, como se pode verificar, classificam correctamente todas as 30 observações do conjunto de validação:

```
> table(iris[c(41:50,91:100,141:150),]$Species, iris.ldaPred$class)

           setosa versicolor virginica
setosa         10          0         0
versicolor      0         10         0
virginica       0          0        10
```

(d) Eis o primeiro plano principal (CPs 1 e 2), usando uma ACP sobre a matriz de covariâncias das quatro variáveis numéricas dos dados dos lírios, e evidenciando as verdadeiras espécies de cada observação:

```
> plot(prcomp(iris[,-5])$x[,1:2], pch=16, col=as.numeric(iris$Species))
```



Como se pode constatar, a separação entre espécies é já bastante evidente no primeiro plano principal, o que significa que as diferenças entre espécies estão entre as principais causas de variabilidade nas quatro variáveis morfométricas. No entanto, a optimização do critério de separação efectuada pela Análise Discriminante Linear significa que a separação das três espécies tem de ser melhor efectuada com os eixos discriminantes.

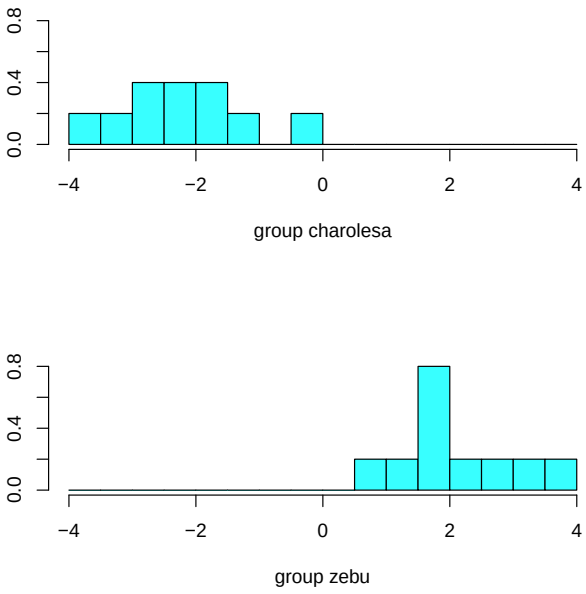
27. Neste exercício, existem apenas $k = 2$ classes (zebus e charolesas), o que significa que apenas existirá $k - 1 = 1$ eixo discriminante. Este facto introduz algumas especifidades aquando da apresentação dos resultados e da sua representação gráfica. Assinale-se também que para poder efectuar a ADL no R foi necessário converter a tabela numa *data frame* em que, quer as variáveis numéricas, quer o nível do factor, são indicadas na colunas:

```
> diday
    v1  v2   v3   especie
1  400 224 28.2      zebu
2  395 229 29.4      zebu
3  395 219 29.7      zebu
4  395 224 28.6      zebu
5  400 223 28.5      zebu
6  400 224 27.8      zebu
7  400 221 26.5      zebu
8  410 233 25.9      zebu
9  402 234 27.1      zebu
10 400 223 26.8      zebu
11 395 224 35.1 charolesa
12 410 232 31.9 charolesa
13 405 233 30.7 charolesa
14 405 240 30.4 charolesa
15 390 217 31.9 charolesa
16 415 243 32.1 charolesa
```

```
17 390 229 32.1 charolesa
18 405 240 31.1 charolesa
19 420 234 32.4 charolesa
20 390 223 33.8 charolesa
```

Eis os comandos do R relevantes, e respectivos resultados:

```
> diday.lda <- lda(especie ~ . , data=diday)

> diday.lda
Call: lda(especie ~ ., data = diday)
[...]
Coefficients of linear discriminants:
           LD1
v1 -0.01222210
v2 -0.09961473
v3 -0.84676160

> plot(diday.lda)
```



Como se pode constatar, o R considera o único eixo discriminante possível, de equação $y^c = -0.01222210\, v1^c - 0.09961473\, v2^c - 0.84676160\, v3^c$, e constrói os histogramas das observações de cada nível do factor neste único eixo (centrado). Esta representação gráfica salienta uma regra simples de separação de zebus e charolesas neste eixo: os zebus ficam com *scores* positivos e as charolesas com *scores* negativos. Esta regra simples daria uma classificação perfeita para as $n = 20$ observações usadas no ajustamento do eixo discriminante. A listagem de resultados produzida pelo R não indica a qualidade discriminante do único eixo, uma vez que a medida usada na função `lda` (proporção em relação à soma dos valores próprios não nulos de $\mathbf{W}^{-1}\mathbf{B}$) produziria sempre o valor 100% no caso de apenas existir um único eixo com capacidade discriminante.

Seria possível ajustar uma Regressão Logística, ou outro Modelo Linear Generalizado de resposta dicotómica, como forma alternativa de separar as duas classes (zebus e charolesas).

28. Este conjunto de dados fornece um exemplo duma discriminação linear pobre entre espécies. Nesta resolução, e dado o grande número de observações, foram usadas as primeiras 150 folhas de cada casta para determinar os resultados, que foram depois validadas com as 150 folhas (50 de cada casta) deixadas de fora na fase do ajustamento.

    (a) Eis os comandos e resultados obtidos.

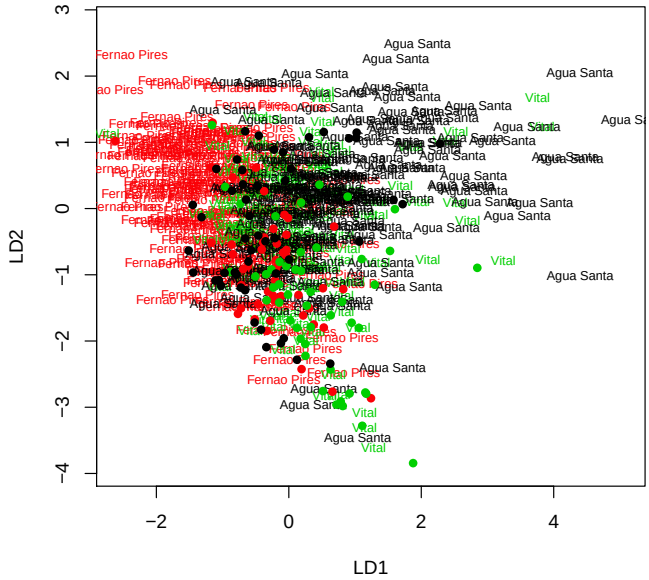
```
> library(MASS)
> vid.treino <- videiras[c(1:150, 201:350, 401:550),]
> vid.lda2 <- lda(Casta ~ . , data=vid.treino)

> vid.lda2
Call: lda(Casta ~ ., data = vid.treino)
[...]
Coefficients of linear discriminants:
              LD1         LD2
NLesq -0.61723332  0.10456287
NP    -0.11661377  0.39408107
NLdir -0.71622348  0.38512073
Area   0.04574428 -0.01027895

Proportion of trace:
   LD1    LD2
0.9588 0.0412
```

    Embora seja desde já possível constatar que a capacidade discriminante do primeiro eixo é muito superior (mais de 20 vezes superior) à do segundo eixo discriminante, a qualidade discriminatória destes eixos não é aparente a partir dos resultados listados acima. A construção da nuvem de pontos vai evidenciar a pobre capacidade discriminatória destes eixos. Mas primeiro vejamos a classificação das 50 folhas deixadas para conjunto de validação:

```
> vid.valid <- videiras[c(151:200, 351:400, 551:600),]
> vid.lda2Pred <- predict(vid.lda2, new=vid.valid)
> table(vid.valid[,"Casta"], vid.lda2Pred$class)

             Agua Santa Fernao Pires Vital
Agua Santa            9           23    18
Fernao Pires          1           12    37
Vital                 7            4    39
```

    Como se pode constatar, a maioria das folhas de Água Santa e Fernão Pires ficam mal classificadas (nas linhas estão as verdadeiras castas, e nas colunas as classes previstas pela função `lda`, uma vez que foi por essa ordem que os argumentos foram passados à função `table`). O gráfico seguinte ilustra essa situação:

```
> plot(vid.lda2, col=as.numeric(vid.treino[,1]))
> points(vid.lda2Pred$x, col=as.numeric(vid.valid[,1]), pch=16)
```



(b) Eis os resultados pedidos:

```
> vid.loadlda <- coef(vid.lda)
> t(vid.loadlda) %*% vid.loadlda
          LD1       LD2
LD1 0.7615636 0.4299989
LD2 0.4299989 0.4001284

> vid.scorelda <- predict(vid.lda)$x
> cor(vid.scorelda)
              LD1           LD2
LD1  1.000000e+00 -7.018464e-16
LD2 -7.018464e-16  1.000000e+00
```

Recorde-se que os eixos discriminantes são sempre não correlacionados entre si, mas os vectores de *loadings* (coeficientes das combinações lineares que definem esses eixos discriminantes) não são ortogonais entre si, mas sim $\mathbf{W}$-ortogonais. Esta situação distingue a ADL e a ACP (nesta última, além de correlação nula entre CPs tem-se também ortogonalidade usual entre vectores de *loadings*).

29. Eis uma possível resposta para a primeira parte do que é solicitado no enunciado:

```
> adl <- function(X,grupos){
        grupos <- as.factor(grupos)
        X <- as.matrix(X)
        k <- length(levels(grupos))
        n <- dim(X)[1]
        p <- dim(X)[2]
Ind <- model.matrix(aov(X[,1] ~ -1 + grupos))   % cria a matriz G indicada nos acetatos
PG <- Ind %*% solve(t(Ind)%*%Ind) %*% t(Ind)
Xc <- scale(X, scale=F)
```

```
B <- (t(Xc) %*% PG %*% Xc)/(n-1)
W <- (t(Xc) %*% (diag(n)-PG) %*% Xc)/(n-1)
valvec <- eigen(solve(W)%*%B)
val <- Re(valvec$val)[1:(k-1)]
loadings <- Re(valvec$vec)[,1:(k-1)]
if (k>2) {rownames(loadings) <- colnames(X)}
else if (k==2) {names(loadings) <- colnames(X)}
rownames(B) <- colnames(X)
colnames(B) <- colnames(X)
rownames(W) <- colnames(X)
colnames(W) <- colnames(X)
if (k>2) {colnames(loadings) <- paste("ED",1:(k-1),sep="")}
scores <- Xc %*% loadings
rownames(scores) <- rownames(X)
list(B=B,W=W,val=val,loadings=loadings,scores=scores)
}
```

Esta função não tem algumas limitações importantes (como a não validação do *input*, ou ainda a impossibilidade de especificar a ADL através duma fórmula, como no comando `lda`). No entanto, é uma primeira aproximação que produz resultados interessantes.

Repare-se na natureza dos argumentos de entrada: uma matriz ou *data frame* `X` com as variáveis numéricas, e um factor ou vector de texto com a designação dos subgrupos de observações que se pretende discriminar.

Eis um exemplo de aplicação aos dados do Exercício 28 (videiras), que permite identificar a pobre capacidade discriminante dos eixos, através do argumento de saída `val` que indica os valores próprios não nulos da matriz $\mathbf{W}^{-1}\mathbf{B}$:

```
> vid.treino <- videiras[c(1:150, 201:350, 401:550),]
> vid.adl <- adl(X=vid.treino[,-1], grupos=vid.treino[,1])
> vid.adl$val
[1] 0.62387847 0.02679995
```

Assim, o maior valor próprio da matriz $\mathbf{W}^{-1}\mathbf{B}$ é $\lambda_1 = 0.62387847$. Como foi visto nos acetatos, este é o valor do quociente $\frac{\mathbf{a}^t\mathbf{B}\mathbf{a}}{\mathbf{a}^t\mathbf{W}\mathbf{a}}$ que divide a variabilidade inter-classes no primeiro eixo discriminante, pela sua variabilidade intra-classes. O facto deste valor próprio ser inferior a 1 indica que neste eixo discriminante há mais variabilidade no seio das três classes (castas) do que há entre classes (castas) diferentes, o que não é bom para a capacidade discriminante do eixo.

Contraste-se esta situação com a que existe no caso dos dados dos lírios, onde o primeiro eixo discriminante (para a totalidade das $n = 150$ observações) tem uma variabilidade inter-classes muito maior do que a variabilidade intra-classes:

```
> adl(X=iris[,-5], grupos=iris[,5])$val
[1] 32.191929  0.285391
```

A função permite ainda quantificar a capacidade discriminante do único eixo discriminante do Exercício 27 (zebus e charolesas), no qual a variabilidade entre as duas classes é cerca de 5 vezes maior que a variabilidade no seio das classes.

```
> adl(X=diday[,-4], grupos=diday[,4])$val
[1] 5.095453
```

Com o auxílio desta função `adl` é também possível de confirmar a afirmação feita no final da resolução do Exercício 28 (videiras), de que os coeficientes (*loadings*) dos eixos discriminantes são **W**-ortogonais entre si (e não ortogonais no sentido usual). Eis a exemplificação com os dados do Exercício 28:

```
> vid.adl <- adl(X=vid.treino[,-1], grupos=vid.treino[,1])
> W <- vid.adl$W
> vid.load <- vid.adl$load
> t(vid.load) %*% W %*% vid.load
              ED1           ED2
ED1  1.094434e+00 -5.329071e-15
ED2 -8.881784e-16  3.163908e+00
```