INSTITUTO SUPERIOR DE AGRONOMIA

**Modelos Matemáticos e Aplicações– 2020-21**

**Algumas resoluções de Exercícios de Estatística Multivariada**

# 1 Matrizes e Álgebra Linear

1. (a) Sabemos que os subespaços gerados por um conjunto de vectores são definidos como o conjunto de todas as combinações lineares desses vectores. No caso dum subespaço gerado por *um único vector*, essa definição significa *o conjunto de todos os múltiplos escalares desse vector*, e do ponto de vista geométrico, o subespaço resultante é uma recta que atravessa a origem. No nosso caso, trata-se do subespaço de todos os vectores da forma $\alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$, para qualquer $\alpha \in \mathbb{R}$. Assim, o subespaço M é o eixo dos $xx$.

   (b) Uma vez que o subespaço é gerado por um único vector, o vector $\vec{\mathbf{m}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, a matriz de projecção ortogonal sobre M é a matriz (de dimensao $2 \times 2$) da forma $\mathbf{P}_M = \vec{\mathbf{m}}(\vec{\mathbf{m}}^t\vec{\mathbf{m}})^{-1}\vec{\mathbf{m}}^t = \begin{bmatrix} 1 \\ 0 \end{bmatrix} (1)^{-1} \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. Logo, a projecção dum vector genérico $\begin{bmatrix} c \\ d \end{bmatrix}$ sobre o subespaço M (isto é, sobre o eixo dos $xx$), é da forma $\mathbf{P}_M \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} c \\ 0 \end{bmatrix}$. **NOTA:** Represente geometricamente a situação.

   (c) O subespaço N é o subespaço gerado pelo vector $\vec{\mathbf{n}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, ou seja, o subespaço de todos os múltiplos escalares de $\vec{\mathbf{n}}$, isto é, dos vectores da forma $\beta \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \beta \\ \beta \end{bmatrix}, \forall \beta \in \mathbb{R}$. Estes vectores caracterizam-se por ser da forma $\begin{bmatrix} x \\ y \end{bmatrix}$, com $y = x$, logo o subespaço N é a recta bissectriz dos quadrantes ímpares de $\mathbb{R}^2$.

   (d) A matriz de projecção ortogonal sobre N é a matriz $\mathbf{P}_N = \vec{\mathbf{n}}(\vec{\mathbf{n}}^t\vec{\mathbf{n}})^{-1}\vec{\mathbf{n}}^t = \begin{bmatrix} 1 \\ 1 \end{bmatrix} (2)^{-1} \begin{bmatrix} 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$. Logo, a projecção do vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ sobre o subespaço N (isto é, sobre a bissectriz dos quadrantes ímpares), é da forma $\mathbf{P}_N \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$. **NOTA:** Represente geometricamente a situação.

2. (a) Por definição, dois vectores dizem-se ortogonais quando o seu produto interno é nulo. Assim, os vectores de $\mathbb{R}^n$ ortogonais ao vector dos $n$ uns são os vectores $\vec{\mathbf{x}} \in \mathbb{R}^n$ cujo produto interno com o vector dos $n$ uns, o vector $\vec{\mathbf{1}}_n$, é nulo. Ora, o produto interno de qualquer vector $\vec{\mathbf{x}} = (x_1, x_2, ..., x_n)^t$ com o vector dos $n$ uns é (verifique!) a soma dos elementos de $\vec{\mathbf{x}}$, ou seja, $\vec{\mathbf{x}}^t\vec{\mathbf{1}}_n = \sum_{i=1}^n x_i$. Logo, os vectores de $\mathbb{R}^n$ ortogonais ao vector $\vec{\mathbf{1}}_n$ são os vectores cujos elementos somam zero. **NOTA:** o subespaço de vectores ortogonais a um dado vector, neste caso, ao vector $\vec{\mathbf{1}}_n$, designa-se o *complemento ortogonal* do subespaço gerado por esse vector.

   (b) Um vector $\vec{\mathbf{x}}$ com $n$ observações duma variável ter soma nula corresponde a uma variável em que a média amostral das observações é zero. A situação mais frequente em que surge

um vector deste tipo é aquela em que trabalhamos com um vector de dados centrados, do tipo $\vec{\mathbf{x}}^c = \begin{bmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{bmatrix}$.

3. Seja $\vec{\mathbf{y}}$ o vector de elemento genérico $y_i$, e $\vec{\mathbf{y}}^c$ o correspondente vector centrado, de elemento genérico $y_i - \overline{y}$.

   (a) O vector de elemento genérico $a + y_i$ é o vector $a\vec{\mathbf{1}}_n + \vec{\mathbf{y}}$. Centrar este vector corresponde a subtrair, a cada um dos seus valores, a média desses valores todos, ou seja, subtrair $\frac{1}{n}\sum_{i=1}^{n}(a+y_i) = \frac{1}{n}\sum_{i=1}^{n} a + \frac{1}{n}\sum_{i=1}^{n} y_i = a + \overline{y}$. Logo, o elemento genérico que se obtém centrando o vector $a\vec{\mathbf{1}}_n + \vec{\mathbf{y}}$ é apenas $(a + y_i) - (a + \overline{y}) = y_i - \overline{y}$. O vector que se obtém centrando $a\vec{\mathbf{1}}_n + \vec{\mathbf{y}}$ é o vector centrado original, $\vec{\mathbf{y}}^c$: a soma duma mesma constante aditiva não afecta os vectores centrados associados.

   (b) Se todos os valores do vector $\vec{\mathbf{y}}$ forem multiplicados por uma mesma constante $b$, obtem-se o vector $b\vec{\mathbf{y}}$, cujo elemento genérico é $b\,y_i$. Centrar este vector corresponde a substituir cada elemento $b\,y_i$ pela diferença desse elemento em relação à média global, que é $\frac{1}{n}\sum_{i=1}^{n} b\,y_i = b\,\overline{y}$. Assim, o elemento genérico do vector centrado é agora $b\,y_i - b\,\overline{y} = b(y_i - \overline{y})$. Logo, o vector centrado é $b\,\vec{\mathbf{y}}^c$.

   (c) Juntando os resultados das duas alíneas anteriores, o vector da transformação linear indicada no enunciado é agora $a\vec{\mathbf{1}}_n + b\vec{\mathbf{y}}$, e o seu correspondente vector centrado é $b\,\vec{\mathbf{y}}^c$.

   (d) Como se viu acima, se as $n$ observações originais $x_i$ constituem os elementos do vector $\vec{\mathbf{x}}$, o correspondente vector centrado $\vec{\mathbf{x}}^c$ tem elemento genérico $x_i - \overline{x}$. À transformação linear $x_i \to a + b\,x_i$ corresponde o vector $a\vec{\mathbf{1}}_n + b\vec{\mathbf{x}}$ e o correspondente vector centrado é $b\,\vec{\mathbf{x}}^c$. Analogamente, a um vector de $n$ observações numa outra variável, $\vec{\mathbf{y}}$, corresponde um vector centrado $\vec{\mathbf{y}}^c$ e, após uma transformação linear $y_i \to c + d\,y_i$ corresponde um vector $c\vec{\mathbf{1}}_n + d\,\vec{\mathbf{y}}$ e um vector centrado $d\,\vec{\mathbf{y}}^c$. Por outro lado, e como se viu nas aulas, a variância amostral das observações de $x$ é proporcional ao quadrado da norma do vector centrado $\vec{\mathbf{x}}^c$: $s_x^2 = (n-1)\|\vec{\mathbf{x}}^c\|^2$. Um resultado equivalente verifica-se para a variância amostral de $y$. Assim, após a transformação linear indicada no enunciado, tem-se $s_{a+bx}^2 = (n-1)\,\|b\,\vec{\mathbf{x}}^c\|^2 = b^2(n-1)\|\vec{\mathbf{x}}^c\|^2 = b^2\,s_x^2$ e, de forma equivalente, $s_{c+dy}^2 = d^2\,s_y^2$. Os desvios padrões são a raíz quadrada (positiva) da variância, pelo que $s_{a+bx} = |b|\,s_x$ e $s_{c+dy} = |d|\,s_y$.

   De forma análoga, a covariância entre as observações de $x$ e $y$ é proporcional ao produto interno entre os vectores $\vec{\mathbf{x}}^c$ e $\vec{\mathbf{y}}^c$: $cov_{x,y} = (n-1) < \vec{\mathbf{x}}^c, \vec{\mathbf{y}}^c >$. Assim, e usando as propriedades dos produtos internos, tem-se: $cov_{a+bx\,,\,c+dy} = (n-1) < b\vec{\mathbf{x}}^c, d\vec{\mathbf{y}}^c > = bd\,(n-1) < \vec{\mathbf{x}}^c, \vec{\mathbf{y}}^c > = bd\,cov_{x,y}$.

   Finalmente, tem-se que o coeficiente de correlação amostral entre $x$ e $y$ é dado por $r_{x,y} = \frac{cov_{x,y}}{s_x \cdot s_y} = \frac{<\vec{\mathbf{x}}^c, \vec{\mathbf{y}}^c>}{\|\vec{\mathbf{x}}^c\| \cdot \|\vec{\mathbf{y}}^c\|}$, ou seja, pelo cosseno do ângulo em $\mathbb{R}^n$ entre os vectores centrados $\vec{\mathbf{x}}^c$ e $\vec{\mathbf{y}}^c$. Para as variáveis após a transformação linear, tem-se: $r_{a+bx\,,\,c+dy} = \frac{cov_{a+bx\,,\,c+dy}}{s_{a+bx} \cdot s_{c+dy}} = \frac{<b\vec{\mathbf{x}}^c, d\vec{\mathbf{y}}^c>}{\|b\,\vec{\mathbf{x}}^c\| \cdot \|d\,\vec{\mathbf{y}}^c\|} = \frac{bd<\vec{\mathbf{x}}^c, \vec{\mathbf{y}}^c>}{|b| \cdot |d|\,\|\vec{\mathbf{x}}^c\| \cdot \|\vec{\mathbf{y}}^c\|} = sgn(bd) \cdot r_{x,y}$, onde $sgn(bd)$ indica o sinal do produto de $b$ e $d$. Quase sempre em aplicações práticas, estas duas constantes multiplicativas são positivas e, nesse caso, as transformações lineares deixam o coeficiente de correlação invariante, resultado que é conhecido das disciplinas introdutórias de Estatística. Geometricamente, o

resultado é evidente. Os vectores centrados $b\vec{\mathbf{x}}^c$ e $d\vec{\mathbf{y}}^c$ mantêm sempre a mesma direcção que os vectores $\vec{\mathbf{x}}^c$ e $\vec{\mathbf{y}}^c$; se $b > 0$ e $d > 0$, mantêm igualmente o mesmo sentido; logo, o ângulo entre esses vectores permanece igual, pelo que o seu coeficiente de correlação não muda.

4. Por definição, o vector $\vec{\mathbf{c}}_j \in \mathbb{R}^p$ ser vector próprio da matriz $\mathbf{X}^t\mathbf{X}$ $(p \times p)$, com valor próprio $\lambda_j$, significa que se verifica a igualdade $\mathbf{X}^t\mathbf{X}\vec{\mathbf{c}}_j = \lambda_j\vec{\mathbf{c}}_j$. Ora, multiplicando esta equação à esquerda por $\mathbf{X}$ tem-se:

$$\mathbf{X}^t\mathbf{X}\vec{\mathbf{c}}_j = \lambda_j\vec{\mathbf{c}}_j \qquad \Rightarrow \qquad (\mathbf{X}\mathbf{X}^t)\mathbf{X}\vec{\mathbf{c}}_j = \lambda_j\mathbf{X}\vec{\mathbf{c}}_j \ ,$$

o que significa que o vector $\mathbf{X}\vec{\mathbf{c}}_j \in \mathbb{R}^n$ é um vector próprio da matriz $\mathbf{X}\mathbf{X}^t$ $(n \times n)$, associado ao mesmo valor próprio $\lambda_j$. Embora não seja, em geral de norma 1, pode, como qualquer outro vector ser normalizado, dividindo-o pela sua norma, $\|\mathbf{X}\vec{\mathbf{c}}_j\|$.

Da mesma forma, e partindo da definição de que $\vec{\mathbf{b}}_j \in \mathbb{R}^n$ é um vector próprio de $\mathbf{X}\mathbf{X}^t$ $(n \times n)$, associado ao valor próprio $\lambda_j$, tem-se, multiplicando à esquerda por $\mathbf{X}^t$:

$$\mathbf{X}\mathbf{X}^t\vec{\mathbf{b}}_j = \lambda_j\vec{\mathbf{b}}_j \qquad \Rightarrow \qquad (\mathbf{X}^t\mathbf{X})\mathbf{X}^t\vec{\mathbf{b}}_j = \lambda_j\mathbf{X}^t\vec{\mathbf{b}}_j \ ,$$

pelo que $\mathbf{X}^t\vec{\mathbf{b}}_j \in \mathbb{R}^p$ é um vector próprio de $\mathbf{X}^t\mathbf{X}$ $(p \times p)$, associado ao mesmo valor próprio $\lambda_j$.

Uma implicação deste resultado é que as matriz $\mathbf{X}^t\mathbf{X}$ e $\mathbf{X}\mathbf{X}^t$ partilham os mesmos valores próprios não nulos: recorde-se que, a partir do Teorema da Decomposição Espectral, uma matriz simétrica de dimensão $k \times k$ tem $k$ valores próprios, associados a um conjunto de $k$ vectores próprios ortogonais. Como elas são de tamanhos diferentes ($\mathbf{X}^t\mathbf{X}$ é de dimensão $p \times p$ e $\mathbf{X}\mathbf{X}^t$ é de dimensão $n \times n$), os restantes valores próprios da maior destas matrizes terão de ser iguais a zero.

5. Tem-se:

$$\mathbf{Y}\vec{\mathbf{v}}_i = \left(\sum_{j=1}^{r} \delta_j\vec{\mathbf{w}}_j\vec{\mathbf{v}}_j^t\right)\vec{\mathbf{v}}_i = \sum_{j=1}^{r} \delta_j\vec{\mathbf{w}}_j\vec{\mathbf{v}}_j^t\vec{\mathbf{v}}_i \ .$$

Mas, numa Decomposição em Valores Singulares, os vector $\vec{\mathbf{v}}_j$ (que são as colunas da matriz $\mathbf{V}$) formam um conjunto ortonormado, ou seja, são de vectores de norma 1, ortogonais entre si. Isso significa que os produtos $\vec{\mathbf{v}}_j^t\vec{\mathbf{v}}_i$ são quase todos nulos, com a excepção de quando $i = j$, em que $\vec{\mathbf{v}}_i^t\vec{\mathbf{v}}_i = 1$. Logo, e como diz o enunciado, tem-se $\mathbf{Y}\vec{\mathbf{v}}_i = \sum_{j=1}^{r} \delta_j\vec{\mathbf{w}}_j\vec{\mathbf{v}}_j^t\vec{\mathbf{v}}_i = \delta_i\vec{\mathbf{w}}_i\vec{\mathbf{v}}_i^t\vec{\mathbf{v}}_i = \delta_i\vec{\mathbf{w}}_i$.

Como se viu nas aulas, a Decomposição em Valores Singulares de $\mathbf{Y}^t$ obtém-se trocando o papel dos vectores $\vec{\mathbf{w}}_j$ e $\vec{\mathbf{v}}_j$, ou seja, é dada por $\mathbf{Y}^t = \sum_{j=1}^{r} \delta_j\vec{\mathbf{v}}_j\vec{\mathbf{w}}_j^t$. Assim, por um raciocínio análogo, e tendo em conta que os vectores $\vec{\mathbf{w}}_j$ também formam um conjunto ortonormado, tem-se:

$$\mathbf{Y}^t\vec{\mathbf{w}}_i = \left(\sum_{j=1}^{r} \delta_j\vec{\mathbf{v}}_j\vec{\mathbf{w}}_j^t\right)\vec{\mathbf{w}}_i = \sum_{j=1}^{r} \delta_j\vec{\mathbf{v}}_j\vec{\mathbf{w}}_j^t\vec{\mathbf{w}}_i = \delta_i\vec{\mathbf{v}}_i\vec{\mathbf{w}}_i^t\vec{\mathbf{w}}_i = \delta_i\vec{\mathbf{v}}_i \ .$$

Em resumo, as imagens dos vectores singulares direitos de $\mathbf{Y}$, através de $\mathbf{Y}$, são um múltiplo escalar dos vectores singulares esquerdos (a constante sendo dada pelo correspondente valor singular). Analogamente, as imagens dos vectores singulares esquerdos de $\mathbf{Y}$, através de $\mathbf{Y}^t$, são um múltiplo escalar dos vectores singulares direitos (a constante sendo dada pelo correspondente valor singular).

6. Como para qualquer outra matriz $n \times p$, a matriz $\mathbf{B}$ admite uma Decomposição em Valores Singulares (DVS), da forma $\mathbf{B} = \mathbf{W}\Delta\mathbf{V}^t$, sendo $\mathbf{W}$ e $\mathbf{V}$ matrizes de colunas ortonormadas (de dimensão, respectivamente $n \times p$ e $p \times p$, admitindo que $\mathbf{B}$ é de característica $p \leq n$). Logo, a matriz de projecção ortogonal sobre o espaço das colunas de $\mathbf{B}$ (o subespaço $\mathcal{C}(\mathbf{B}) \subset \mathbb{R}^n$) é da forma $\mathbf{P}_B = \mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t$. Substituindo a DVS de $\mathbf{B}$, tendo em conta que a ortonormalidade das colunas de $\mathbf{W}$ e de $\mathbf{V}$ significa que $\mathbf{W}^t\mathbf{W} = \mathbf{I}_p$ e $\mathbf{V}^t\mathbf{V} = \mathbf{I}_p$, e usando a expressão para a inversa duma matriz simétrica, baseada na sua Decomposição Espectral (dada nas aulas), tem-se:

$$\begin{aligned}
\mathbf{P}_B &= \mathbf{W}\Delta\mathbf{V}^t \left[(\mathbf{W}\Delta\mathbf{V}^t)^t(\mathbf{W}\Delta\mathbf{V}^t)\right]^{-1} (\mathbf{W}\Delta\mathbf{V}^t)^t = \mathbf{W}\Delta\mathbf{V}^t \left[\mathbf{V}\Delta\mathbf{W}^t\mathbf{W}\Delta\mathbf{V}^t\right]^{-1} \mathbf{V}\Delta\mathbf{W}^t \\
&= \mathbf{W}\Delta\mathbf{V}^t \left[\mathbf{V}\Delta^2\mathbf{V}^t\right]^{-1} \mathbf{V}\Delta\mathbf{W}^t = \mathbf{W}\Delta\mathbf{V}^t \left[\mathbf{V}\Delta^{-2}\mathbf{V}^t\right] \mathbf{V}\Delta\mathbf{W}^t \\
&= \mathbf{W}\Delta\Delta^{-2}\Delta\mathbf{W}^t = \mathbf{W}\Delta^0\mathbf{W}^t = \mathbf{W}\mathbf{I}_p\mathbf{W}^t = \mathbf{W}\mathbf{W}^t \ .
\end{aligned}$$

Repare-se que a matriz de projecção ortogonal tem a forma simplificada, correspondente ao facto de os vectores singulares esquerdos e $\mathbf{B}$, ou seja, as colunas da matriz $\mathbf{W}$, constituírem uma base ortonormada do espaço gerado pelas colunas de $\mathbf{B}$.

## 2 Análise em Componentes Principais

7. (a) Começamos por ajustar a ACP pedida no enunciado, sobre os dados não normalizados:

```
> santarem.acp <- prcomp(santarem)
```

i. A proporção de variabilidade explicada apenas pela primeira CP é elevadíssima, $94,18\%$, como se pode verificar através do comando `summary`:

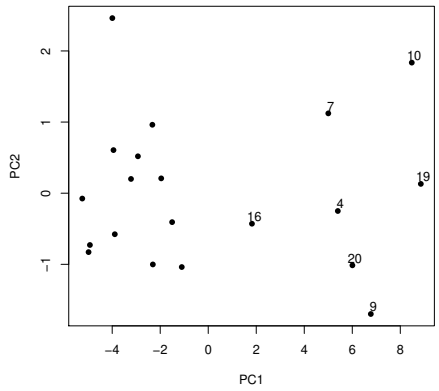
```
> summary(prcomp(santarem))
Importance of components:
                          PC1     PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     4.8666 1.03395 0.4134 0.3362 0.23305 0.16197 0.14394 0.07579 0.05609
Proportion of Variance 0.9418 0.04251 0.0068 0.0045 0.00216 0.00104 0.00082 0.00023 0.00013
Cumulative Proportion  0.9418 0.98433 0.9911 0.9956 0.99778 0.99882 0.99965 0.99987 1.00000
```

Às duas primeiras CPs corresponde quase $98,5\%$ da variabilidade total (inércia). Trata-se de valores muito elevados, mas como se verá de seguida, algo ilusórios.

ii. Os comandos seguintes permitem construir a nuvem de pontos no primeiro plano principal, e identificar (através da selecção dos pontos com o rato) os sete pontos na metade direita do gráfico, aos quais correspondem *scores* positivos na primeira Componente Principal:

```
> plot(santarem.acp$x[,1:2], pch=16)
> identify(santarem.acp$x[,1:2])
[1]  4  7  9 10 16 19 20
> rownames(santarem[c(4,7,9,10,16,19,20),])
[1] "Alpiarca"  "Chamusca"  "Coruche"  "Entroncamento"  "Santarem"  "Torres Novas"  "V.N.Barquinha"
```

O gráfico produzido é o seguinte:




De forma análoga, pode identificar-se o ponto no canto superior esquerdo (com o maior *score* na CP2), que corresponde à Golegã.

iii. Arredondando até 4 casas decimais, as correlações entre variáveis originais e CPs são:

```
> round(cor(santarem, santarem.acp$x), d=4)
             PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
trigo     0.1873 -0.0099  0.4037 -0.7764 -0.2424  0.0579  0.3650  0.0615  0.0038
milho     0.3423  0.9396 -0.0034 -0.0032  0.0018  0.0009  0.0025 -0.0042 -0.0002
centeio   0.2569 -0.0239  0.6387  0.4543 -0.2660 -0.4748  0.1268 -0.0753  0.0347
aveia     0.0910 -0.0574  0.5380 -0.1045  0.8059 -0.1249  0.1246 -0.0197 -0.0841
cevada    0.2854  0.0137  0.8738 -0.2872 -0.0804  0.0806 -0.2417 -0.0300  0.0034
fava     -0.0013  0.0354  0.5571  0.7525 -0.0315  0.3121  0.1498  0.0365 -0.0045
feijao    0.2836  0.4915  0.2847  0.2553 -0.2047 -0.3756 -0.2472  0.5164 -0.1450
grao      0.3308  0.1966  0.2950 -0.0515  0.7297 -0.0672 -0.0314  0.2243  0.4171
batata    0.9999 -0.0165 -0.0021  0.0006  0.0001  0.0003  0.0000  0.0000 -0.0001
```

A fórmula, dada nas aulas para a correlação entre a $i$-ésima variável, $x_i$, e a $j$-ésima Componente Principal, $z_j$, é:

$$r_{x_i,z_j} = \frac{\sqrt{\lambda_j}}{s_i}\, v_{ij} \ ,$$

onde $\lambda_j$ é a variância da CP $j$, $s_i$ o desvio padrão da $i$-ésima variável, e $v_{ij}$ o coeficiente (*loading*) da variável $x_i$ na CP $z_j$. Já foi dado em cima o desvio padrão da CP 1 (4.8666) e do enunciado consta a variância da cada variável (na diagonal da matriz de covariâncias). Assim, para a batata tem-se $s_9 = \sqrt{23.531} = 4.850876$; para a fava tem-se $s_6 = \sqrt{0.084} = 0.2898275$; e para o milho $s_2 = \sqrt{1.198} = 1.094532$. A fórmula utiliza ainda os coeficientes de cada variável na primeira CP, que são dados pela primeira coluna da matriz `Rotations`, criada pelo comando `prcomp`:

```
> santarem.acp$rot[,1]
      trigo         milho       centeio         aveia        cevada          fava
1.007776e-02  7.698885e-02  1.309135e-02  4.481334e-03  2.005914e-02 -7.561437e-05
     feijao          grao         batata
 7.315052e-03  7.868217e-03   9.966253e-01
```

Assim, o coeficiente de correlação entre a variável `batata` e a primeira CP é: $r_{x_9,z_1} = \frac{4.8666}{4.850876} \times 0.966253 = 0.9998558$, o que confirma o valor obtido acima. De forma análoga, a correlação entre a variável `fava` e a CP 1 é: $r_{x_6,z_1} = \frac{4.8666}{0.2898275} \times (-0.0007561437) = -0.00127$, que igualmente confirma o valor acima obtido. Finalmente, a correlação entre o `milho` e a primeira CP é: $r_{x_2,z_1} = \frac{4.8666}{1.094532} \times 0.07698885 = 0.3423143$. Destes valores resulta óbvio que a CP 1 é praticamente colinear com a variável `batata`, com a qual se

pode identificar. O facto da CP 1 ser quase ortogonal à variável `fava` (sendo a respectiva correlação quase nula) obriga assim a que a correlação entre `batata` e `fava` também tenha de ser quase nula, o que se pode confirmar calculando essa correlação a partir da matriz de covariâncias dada no enunciado: $r_{x_6,x_9} = \frac{cov_{x_6,x_9}}{s_6 \cdot s_9} = \frac{-0.003}{\sqrt{0.084 \times 23.351}} = -0.002$.
No que respeita à segunda CP, são necessários os *loadings* de cada variável, dados por:

```
> santarem.acp$rot[,2]
       trigo        milho      centeio        aveia       cevada         fava
-0.002495072  0.994802708 -0.005725971 -0.013293132  0.004544923  0.009935757
      feijao         grao       batata
 0.059656229  0.022008418 -0.077390188
```

Também neste caso se confirmam os valores obtidos com o auxílio do `R`, sendo evidente a forte correlação entre a CP 2 e a variável `milho` (0.9396).

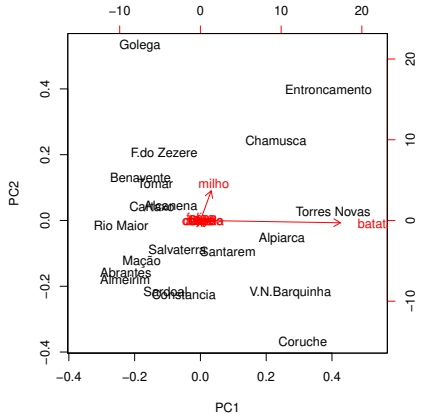
iv. Como se viu, a primeira CP é quase exclusivamente determinada pela variável produtividade da batata. Tal facto era previsível: está-se a efectuar uma ACP sobre a matriz de covariâncias, sendo a variável `batata` a que tem, de longe, a maior variância ($s_9 = 23.531$). Como se viu aquando da interpretação da ACP no espaço das variáveis ($\mathbb{R}^n$), a ACP procura a combinação linear de comprimento máximo dos vectores (centrados) representativos de cada variável. Ora, o comprimento de cada um desses vectores é proporcional ao desvio padrão da variável, pelo que havendo uma variável de variância muito maior que as restantes (cujo vector representativo é de longe o maior de todos), a combinação linear irá acompanhar esse vector, ou seja a CP resultante estará próxima dessa variável. Repare-se que esta conclusão não invalida que a CP 1 possa estar também fortemente correlacionada com outra variáveis originais (embora não seja o caso neste conjunto de dados), uma vez que poderia haver outras variáveis fortemente correlacionadas com a variável `batata`. Outra forma de pensar nesta conclusão é verificar que a variância de `batata`, só por si, corresponde a $93,57\%$ da soma das variâncias de todas as variáveis, ou seja, $93,57\%$ da inércia total da nuvem de pontos. Assim, a primeira CP nunca poderia explicar menos do que esta proporção da inércia total (já que uma variável individual é também uma combinação linear de todas as variáveis, com peso 1 associado a essa variável e 0 a todas as restantes) e, para exceder essa proporção, não se afastará muito da direcção definida pela variável `batata`. Pode ainda considerar-se a representação usual dos $n = 20$ pontos em $\mathbb{R}^9$, em que cada eixo corresponde a uma das variáveis originais (centradas). Neste caso, e dada a enorme variância da variável `batata`, é evidente que a forma da nuvem tem como principal direcção a definida pelo eixo associado à `batata`, pelo que a primeira CP acompanhará essa direcção.
No que respeita à CP 2, e como também já se viu, há uma correlação bastante elevada (0.9396) com a produtividade do `milho`. Mais uma vez, a variância da variável `milho` é um factor a ter em conta, já que é mais de 10 vezes superior à variância seguinte (da variável `cevada`). De qualquer forma, a correlação entre `milho` e CP 2 não pode ser ainda mais elevada, uma vez que as produtividades do milho e da batata ainda têm alguma correlação (0.3267) e, sendo CP 1 quase perfeitamente correlacionada com a `batata`, a segunda CP (que, por construção, tem de ser não correlacionada com a CP 1) não pode acompanhar de forma tão forte o `milho`.

v. O comando do `R` para construir o *biplot* pedido no enunciado é apenas:

```
> biplot(santarem.acp)
```

O resultado, dado em baixo, é uma aproximação bidimensional de grande qualidade à

nuvem de $n=20$ pontos em $\mathbb{R}^9$ correspondente aos dados originais (centrados), uma vez que a proporção da inércia total associada a estes dois eixos principais é muito elevada $(98,4\%)$.




O comprimento muito maior do marcador da variável `batata` reflecte a muito maior variância desta variável (já que o comprimento de cada marcador de variáveis é, nas 9 dimensões, proporcional ao desvio padrão dessa variável). De igual forma, o facto de os vectores marcadores das restantes variáveis (com exclusão do `milho`) serem quase invisíveis reflecte a muito mais baixa variância dessas variáveis. É visível a fortíssima correlação entre a primeira CP e a variável `batata`, uma vez que o marcador desta última variável está quase na horizontal (direcção a que corresponde a CP 1). Da mesma forma, o facto do marcador de `milho` estar quase na vertical reflecte a forte correlação entre essa variável e a CP2. Sabemos que, projectando ortogonalmente os marcadores de indivíduos sobre as direcções definidas pelos marcadores de cada variável reconstrói-se (aproximadamente, mas neste caso com grande qualidade) o valor de cada indivíduo em cada variável. Assim, é possível verificar que os sete concelhos do lado direito têm as mais elevadas produtividades de batata (além de terem *scores* elevados na CP1). Dois destes concelhos (Entroncamento e Chamusca) têm igualmente altas produtividades de milho (uma vez que a sua projecção sobre a direcção do marcador dessa variável coloca-os numa posição extrema). O mesmo se passa com a Golegã (embora este concelho não tenha uma elevada produtividade de batata). Já no que respeita a outros concelhos com elevada produtividade de batata, como Coruche e Vila Nova da Barquinha, verifica-se que a sua produtividade de milho é baixa. Estas afirmações podem ser confirmadas directamente na tabela.

vi. O facto de a batata dominar completamente a CP 1 e o milho dominar em grande medida a CP2, bem como o facto de as restantes culturas desempenharem um papel pouco importante na ACP, são consequências directas da opção por usar uma ACP sobre os dados originais, ou seja, sobre a matriz de covariâncias, uma vez que essas características reflectem as enormes diferenças entre as variabilidades associadas às produtividades dessas duas culturas, e delas com as restantes. Embora as unidades de medida das produtividades (t/ha) sejam iguais para todas as culturas (o que torna legítima a ACP sobre a matriz de covariâncias), estas diferenças assinaláveis de produtividades (e da respectiva variância) tornam previsível, e talvez pouco informativa, a variante de ACP usada. Merece atenção uma ACP sobre os dados normalizados (isto é, uma ACP sobre a matriz de correlações), como se fará no ponto seguinte.

(b) Consideremos agora a ACP sobre os dados normalizados:

```
> santarem.acpR <- prcomp(santarem, scale=TRUE)
```

i. Eis as proporções da inércia total explicadas pelas CPs sobre a matriz de correlações:

```
> summary(santarem.acpR)
Importance of components:
                          PC1    PC2    PC3   PC4     PC5     PC6     PC7    PC8     PC9
Standard deviation     1.6922 1.2951 1.1977 1.127 0.84353 0.68001 0.52376 0.4124 0.37011
Proportion of Variance 0.3182 0.1864 0.1594 0.141 0.07906 0.05138 0.03048 0.0189 0.01522
Cumulative Proportion  0.3182 0.5045 0.6639 0.805 0.88402 0.93540 0.96588 0.9848 1.00000
```

É evidente que as primeiras CPs explicam bastante menos da variabilidade total do que no caso da ACP sobre os dados originais. Tal facto, que costuma caracterizar as CPs sobre dados normalizados, reflecte o facto de que a normalização torna igual a variância de todas as variáveis (que passa a ser 1). Na habitual representação dos dados ($n$ pontos em $\mathbb{R}^p$), a normalização tende a tornar mais esférica a nuvem de pontos, pelo que as primeiras direcções principais explicam uma menor proporção da variabilidade total. Mesmo assim, as diferenças entre os dois conjuntos de valores são, neste exemplo, particularmente assinaláveis. Assim, mesmo 3 CPs sobre os dados normalizados apenas correspondem a cerca de 2/3 da variabilidade total.

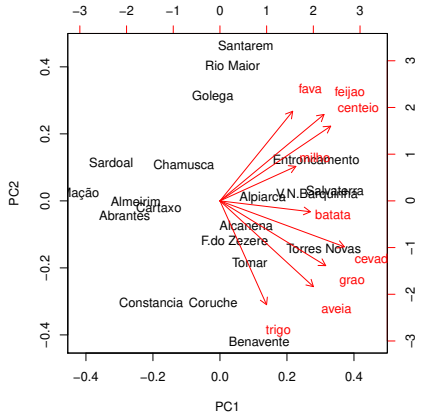
ii. As diferenças desta variante da ACP em relação à ACP sobre os dados originais são igualmente visíveis nas correlações entre cada variável individual e cada uma das CPs agora definidas, como se verifica em baixo. Registe-se que, sendo as correlações lineares invariantes a mudanças lineares de escala (a constante multiplicativa da normalização, $1/s_j$, é sempre positiva, pelo que nem mudanças de sinal haverá nas correlações), não é necessário normalizar as variáveis originais no cálculo destas correlações.

```
> round(cor(santarem, santarem.acpR$x), d=4)
           PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
trigo   0.2791 -0.6214  0.2133 -0.6227  0.1529 -0.0936  0.1711 -0.1956  0.0546
milho   0.4545  0.2064  0.6774  0.2572  0.2984 -0.3042  0.1588  0.1333 -0.0348
centeio 0.6629  0.4472 -0.3000 -0.3180 -0.1244  0.2225  0.2876  0.0551 -0.1369
aveia   0.5595 -0.5134 -0.4368  0.3779  0.0792  0.0905  0.1312  0.1345  0.2002
cevada  0.7449 -0.2748 -0.1417 -0.4451  0.1347 -0.0674 -0.3077  0.1638 -0.0852
fava    0.4354  0.5362 -0.5644  0.0124 -0.0098 -0.4231 -0.0357 -0.1345  0.0758
feijao  0.6230  0.5172  0.3473  0.0058  0.2404  0.3356 -0.1426 -0.1139  0.1412
grao    0.6331 -0.3877 -0.0386  0.6106 -0.0040  0.0471 -0.0469 -0.1866 -0.1880
batata  0.5415 -0.0651  0.4343 -0.0143 -0.7081 -0.0707 -0.0382  0.0147  0.0753
```

Assim, em vez de a primeira CP estar fortemente associada a uma variável individual, tem-se agora que a CP 1 tem correlações assinaláveis com a generalidade das variáveis. Este facto é natural, tendo em conta que (como se viu nas aulas) a primeira CP sobre a matriz de correlações é igualmente a combinação linear das variáveis originais que maximiza a soma dos quadrados das correlações com cada variável individual. Da mesma forma, a segunda CP não tem correlações muito elevadas com qualquer das culturas (embora trigo, aveia, fava e feijão tenham correlações mais importantes).

iii. O primeiro comentário a fazer ao *biplot* dos dados normalizados é que a sua qualidade deixa bastante a desejar, uma vez que a visualização apenas corresponde a cerca de metade da variabilidade total (0.5045). Assim, haverá que ter cautela nas afirmações que se produzem e confirmar quaisquer potenciais conclusões. Um segundo comentário diz respeito ao facto de que, na totalidade das 9 dimensões, os marcadores de variáveis deste *biplot* deveriam ter todos igual comprimento. Assim, as variáveis com marcadores

mais curtos (neste caso, sobretudo milho e batata) estão pior representadas nas duas CPs que são mostradas, tendo parte importante nas restantes direcções (não mostradas). Assim se explica também que, aparecendo o marcador da variável `batata` quase na horizontal, essa variável não tenha uma correlação elevada com a CP1 ($r = 0.5415$), havendo outras variáveis (como a cevada, centeio, grão, feijão) mais correlacionadas com a CP 1. Registe-se, aliás, a partir das correlações entre variáveis e CPs, que a variável `batata` tem uma correlação importante com a CP 7 (e em bastante menor medida, também com a CP 3).




Apesar destas limitações, pode ver-se no *biplot* que a totalidade das variáveis têm correlações interessantes com a primeira CP, como se constatou no ponto anterior. Além disso, o gráfico sugere que concelhos como o Entroncamento, Salvaterra de Magos, Vila Nova da Barquinha ou Torres Vedras têm produtividades elevadas na generalidade das culturas, enquanto que outros, como Mação ou o Sardoal, têm produtividades baixas na generalidade das culturas. Também parece ser o caso que um concelho como Benavente tem produtividades elevadas nos cereais (excepto o centeio) e grão, mas pouco distantes da média noutras culturas.

iv. A resposta a esta pergunta deve sempre partir de saber se todas as variáveis têm as mesmas unidades de medida (em cujo caso as duas variantes da ACP são legítimas) ou não (em cujo caso é aconselhável a normalização prévia dos dados, a fim de evitar resultados que sejam dependentes da escolha concreta de sistema de medidas usadas). Neste caso, é a partida legítimo optar, quer pela ACP sobre os dados originais, quer pela ACP sobre os dados normalizados. No entanto, as características específicas do conjunto de dados, acima discutidas, apontam para o maior interesse numa ACP sobre dados normalizados. Em sentido contrário, pode assinalar-se que a ACP sobre a matriz de correlações é menos eficaz na redução da dimensionalidade dos dados. Como sempre, nestas situações, os objectivos concretos do estudo acabam por ter um peso crucial na decisão. Pode sugerir-se que sejam efectuadas ambas as análises, uma vez que ambas contribuem para explorar a natureza dos dados e da informação neles contida.

8. The `brix2` dataset is small in size: it produces a $14 \times 7$ data frame.

   (a) A Covariance matrix PCA is *not* advisable, since the variables are in different units of measurement: some variables are lengths, others weights, another is in degrees Brix, one in the pH scale, another measures acidity. The results of a Covariance matrix PCA would be

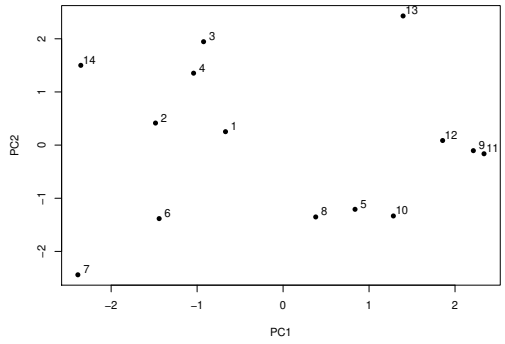
dependent on the precise choice of units of measurement, which is an undesirable effect.

(b) The results of a correlation matrix PCA are shown below:

```
> brix2.acp <- prcomp(brix2, scale=TRUE)
> brix2.acp
Standard deviations (1, .., p=7):
[1] 1.6674531 1.4405975 1.0692044 0.7565070 0.4768517 0.3424740 0.2900035

Rotation (n x k) = (7 x 7):
                PC1         PC2         PC3         PC4         PC5         PC6         PC7
Diametro -0.4596408  0.13164613 -0.49198976  0.20019323 -0.47136568  0.34642868 -0.38349703
Altura   -0.2584398 -0.40661389 -0.58588179  0.04735340  0.53980838  0.07792709  0.35342394
Peso     -0.3687679 -0.49905490  0.05821883 -0.14959125 -0.41905140 -0.64082704  0.05407468
Brix     -0.3375926  0.54981307  0.03673560  0.02181401 -0.18526037 -0.05222837  0.73814787
pH       -0.4405964  0.08780403  0.37280515  0.65294260  0.37799871 -0.19946897 -0.22400140
Acidez   -0.3303218 -0.39086752  0.52021613 -0.18561317 -0.06154355  0.63791789  0.14935249
Acucar   -0.4107834  0.32480343  0.01847072 -0.68850610  0.36031265 -0.11747485 -0.32825660

> summary(brix2.acp)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.6675 1.4406 1.0692 0.75651 0.47685 0.34247 0.29000
Proportion of Variance 0.3972 0.2965 0.1633 0.08176 0.03248 0.01676 0.01201
Cumulative Proportion  0.3972 0.6937 0.8570 0.93875 0.97123 0.98799 1.00000
```

The first two correlation matrix PCs account for just under 70% of the total variance of the *normalized* data, which is fairly good considering the reduction from 7 to 2 dimensions. Since in a normalized-data PCA all variables have variance 1, the total variance (inertia) of the full data set equals the number of variables, in our case, $p = 7$. Below is the resulting scatterplot, when projected onto the principle plane.

```
> plot(brix2.acp$x[,1:2], pch=16)
> text(brix2.acp$x[,1:2]+0.1, rownames(brix2))
```




Note that the plot has been centred, that is, the centre of gravity of the scatterplot is at the origin (0,0).

The proportion of total inertia accounted for by the first three PCs grows to almost 86%. With 3-D graphical software, most of the variability in the original 14-point scatterplot in $\mathbb{R}^7$ is retained.

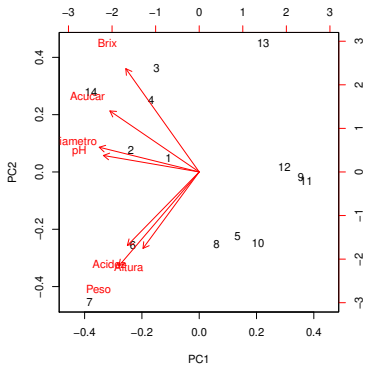
(c) Below is the two-dimensional biplot for this (normalized) dataset. The near-horizontal position of the vectors representing the variables `pH` and `Diametro` in the biplot suggests that are strongly correlated with the first principal component. The variable most strongly correlated with the second principal component would appear to be `Brix`, whose representative vector is the one that is closest to a vertical position. As with all other statements of this kind, we must be careful to check whether this is, in fact the case, since the first principal plane accounts of some 70% of the total variability and leaves the remaining 30% unaccounted for. Here are the correlations between individual variables (rows) and PCs (columns):

```
> round(cor(brix2, brix2.acp$x), d=3)
            PC1    PC2    PC3    PC4    PC5    PC6    PC7
Diametro -0.766  0.190 -0.526  0.151 -0.225  0.119 -0.111
Altura   -0.431 -0.586 -0.626  0.036  0.257  0.027  0.102
Peso     -0.615 -0.719  0.062 -0.113 -0.200 -0.219  0.016
Brix     -0.563  0.792  0.039  0.017 -0.088 -0.018  0.214
pH       -0.735  0.126  0.399  0.494  0.180 -0.068 -0.065
Acidez   -0.551 -0.563  0.556 -0.140 -0.029  0.218  0.043
Acucar   -0.685  0.468  0.020 -0.521  0.172 -0.040 -0.095
```

One noticeable fact resulting from these correlations is that some variables also have non-negligeable correlations with PCs 3 and 4, namely `Altura`, `Acidez`, `pH` and `Acucar`. These variables are therefore not very accurately represented in the first two dimensions. This fact is also related to the shorter length of their representative vectors in the biplot. In fact, we know that the length of the representative vectors is proportional to the standard deviation of each variable. In a correlation matrix PCA, the normalization of the data implies that all normalized variables have variance (and standard deviation) 1. Hence, if they were represented in all $p$ dimensions, their representative vectrs should all have equal length. In so far as some vectors are shorter than others in the 2-D biplot, this means that the shorter vectors have important components in other principal directions that are not shown.

In the biplot, individuals 14 and 1 to 4 seem to have large values in variables `pH` and `Diametro`, and also in variables `Acucar` (sugar) and `Brix`, which appear to be positively correlated with `pH` and `Diametro`. In contrast, individuals 5 and 8 to 12 appear to have low values in these variables. Also visible in the biplot is a group of three variables - `Acidez` (acidity), `Altura` (height) and `Peso` (weight) - which appear to be strongly correlated among themselves. Individuals 6 and 7 appear to have large values on these variables, in contrast to individual 13. Given the small size of the dataset, these statements can be directly confirmed in the question sheet.

```
> biplot(brix2.acp)
```



A 3-D biplot can be obtained (and rotated) using the package `pca3d`, with the command:

```
> pca3d(brix2.acp, biplot=TRUE)
```

(d) The five groups of points described in the question are not entirely well separated in the first principal plane, despite some intra-group proximity. For example, individuals 5 and 8 (which belong to the same group, associated with the December 13 date), appear to be closer to individual 10 (January 16) than to the other two individuals in the December date (6 and 7). This may, in part, be due to the variability that was lost with the two-dimensional representation, but see the answer to the next question.

(e) It is not necessary for a low-dimension representation to preserve homogeneity within groups of observed individuals. That is the goal of Discriminant Analyses, but PCA is unaware of the existence of subgroups on individuals, and these will only be preserved in low-dimensional representations insofar as the groups are reflected in the variability in the observed variables.

(f) The scores of the new observation on the principal plot are given by the linear combination of the new individual's *normalized* values (because a Correlation-matrix PCA was performed), using the loadings of the first and of the second PC. The loadings are given in the `Rotation` object of the output, namely in the first two columns of that output matrix, which are shown above (question 8b). The normalization of the new values for each variable must be done with the mean and standard deviation for each variable in the original dataset (used to perform the PCA). These are shown below:

```
> apply(brix2,2,mean)
Diametro   Altura     Peso     Brix       pH   Acidez   Acucar
1.957143 1.892857 3.735000 8.278571 2.878571 1.475714 3.628571
> apply(brix2,2,sd)
 Diametro    Altura      Peso      Brix        pH    Acidez    Acucar
0.1089410 0.1384768 0.3598451 0.2750624 0.1360349 0.1554504 1.2755684
```

These values are also stored in the list resulting from the `prcomp` command, in the objects `center` (US spelling) and `scale`, respectively:

```
> brix2.acp$center
Diametro   Altura     Peso     Brix       pH   Acidez   Acucar
1.957143 1.892857 3.735000 8.278571 2.878571 1.475714 3.628571
> brix2.acp$scale
 Diametro    Altura      Peso      Brix        pH    Acidez    Acucar
0.1089410 0.1384768 0.3598451 0.2750624 0.1360349 0.1554504 1.2755684
```

Thus, to obtain the coordinate on the first PC for the new observation, we would do the following calculations:

$$\begin{aligned}(-0.4596408) \times \frac{1.9 - 1.957143}{0.1089410} \quad &+ \quad (-0.2584398) \times \frac{2.0 - 1.892857}{0.1384768} + \ldots \\ \ldots \quad &+ \quad (-0.4107834) \times \frac{3.78 - 3.628571}{1.2755684} \;=\; -0.08895189\end{aligned}$$

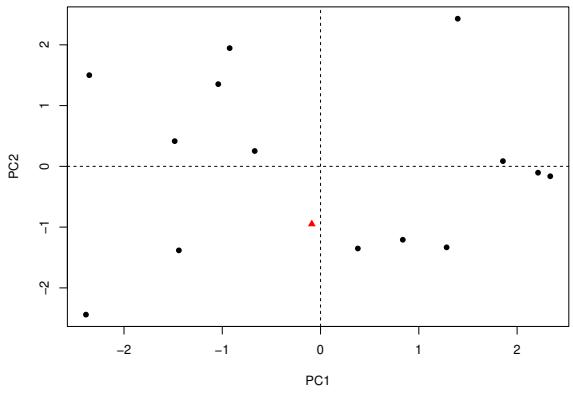
The second coordinate would be obtained in a similar fashion, but using the loadings for the second PC (the second column in the `Rotation` matriz). The resulting value is $-0.9491006$.

In R, the easiest way to obtain these coordinates is by using the `predict` command, indicating the new values in a data frame associated with the `new` argument (as in the Linear and Generalized Linear Models):

```
> predict(brix2.acp, new=data.frame(Diametro=1.9, Altura=2.0, Peso=3.92, Brix=8.1,
+                                   pH=2.91, Acidez=1.48,Acucar=3.78))
             PC1        PC2     PC3         PC4       PC5       PC6         PC7
[1,] -0.08895189 -0.9491006 -0.0865 -0.09543871 0.6981494 -0.45941 -0.0634021
```

The coordinate (score) on PC1 confirms our previous calculations. So the point representing the new individual on the principal plane will have coordinates (scores) $-0.08895189$ and $-0.9491006$. We can place that marker on the 2-D scatterplot, with the following commands:

```
> plot(brix2.acp$x[,1:2], pch=16)
> abline(v=0, lty=2)
> abline(h=0, lty=2)
> points(-0.08895189, -0.9491006, pch=17, col="red")
```




9. Neste Exercício consideram-se de novo os dados da *data frame* `milho`, já considerados nos Exercícios de Modelo Linear (Módulo II).

   (a) As variáveis têm diferentes unidades de medida ($^oF$, polegadas). Logo, devem-se normalizar as variáveis antes de efectuar a ACP, ou seja, deve optar-se por uma ACP sobre a matriz de correlações. Registe-se que, caso se optasse por uma ACP sobre as variáveis nas unidades originais (ou seja, uma ACP sobre a matriz de covariâncias), os seus resultados seriam diferentes usando as unidades além-Atlântico, ou usando as suas conversões para as unidades do sistema métrico internacional.

   (b) Eis os resultados produzidos pelo comando `summary`, relativos a uma ACP sobre as variáveis normalizadas:

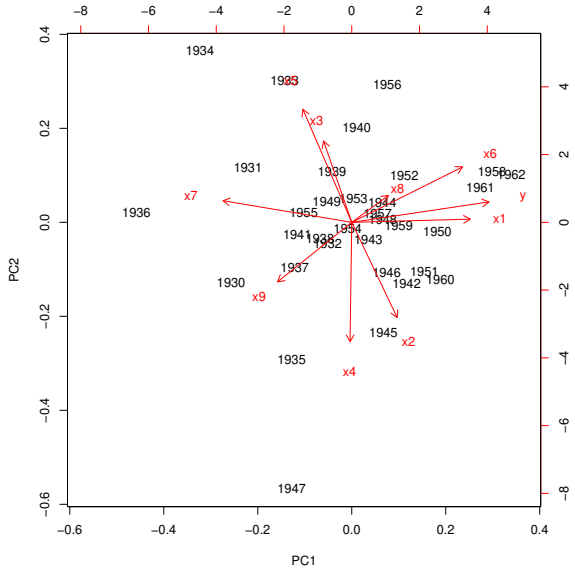
```
> milho.acp <- prcomp(milho, scale=TRUE)
> summary(milho.acp)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     1.7443 1.4483 1.1456 1.0152 0.86804 0.76880 0.64507
Proportion of Variance 0.3043 0.2098 0.1313 0.1031 0.07535 0.05911 0.04161
Cumulative Proportion  0.3043 0.5140 0.6452 0.7483 0.82367 0.88278 0.92439
```

```
                            PC8     PC9    PC10
Standard deviation     0.5622 0.55605 0.36174
Proportion of Variance 0.0316 0.03092 0.01309
Cumulative Proportion  0.9560 0.98691 1.00000
```

Mesmo retendo quatro Componentes Principais (o que já não permite a visualização), apenas se retém cerca de três quartos da variabilidade (inércia) total. Uma visualização dos dados no primeiro plano principal (definido pelas duas primeiras CPs) apenas reproduz pouco mais de metade da inércia original. Com uma visualização tri-dimensional, a proporção da inérica total que é retida sobe para cerca de 2/3.

Assim, não sendo de desprezar as possibilidades de visualização produzidas pela ACP, há que ter cautela na extracção de conclusões a partir das visualizações a 2 ou 3 dimensões, confirmando sempre as conclusões que esses gráficos a baixa dimensão parecem sugerir.

(c) Eis o *biplot* pedido no enunciado, em cuja interpretação deve ser sempre tida em conta a discussão da alínea anterior:




i. Sabemos que apenas cerca de metade da variabilidade (inércia) dos dados é preservada na representação bidimensional. Assim, as ilações que a seguir se enumeram devem ser sempre confirmadas numericamente com base na totalidade da informação disponível. Feita esta ressalva, enunciemos alguns aspectos da informação do biplot:

- A grande dispersão de sentidos dos vectores representativos das variáveis sugere uma grande diversidade de situações no que respeita às correlações entre pares de variáveis, desde variáveis fortemente correlacionadas (com sinal positivo ou negativo) a variáveis fracamente correlacionadas. Como exemplo de pares de variáveis aparentemente forte e positivamente correlacionadas citem-se $y$ e $x_1$. Um par de variáveis forte, mas negativamente correlacionadas, serão $x_5$ e $x_2$, cujos vectores representativos partilhaa uma direcção comum embora de sentido oposto. Como exemplo de par de variáveis fracamente correlacionadas cite-se o caso de $x_5$ e $x_6$, cujos vectores representativos são quase ortogonais. Vejamos se a matriz de correlações deste conjunto de dados confirma estas impressões:

```
> round(cor(milho), d=2)
      x1    x2    x3    x4    x5    x6    x7    x8    x9     y
```

```
x1  1.00  0.04 -0.04 -0.07 -0.26  0.40 -0.55  0.06 -0.08  0.75
x2  0.04  1.00 -0.12  0.33 -0.45 -0.03 -0.30  0.20  0.09  0.19
x3 -0.04 -0.12  1.00 -0.39  0.29  0.08  0.15  0.00  0.12 -0.10
x4 -0.07  0.33 -0.39  1.00 -0.35 -0.28 -0.16 -0.05  0.08 -0.15
x5 -0.26 -0.45  0.29 -0.35  1.00 -0.08  0.26  0.17 -0.23 -0.15
x6  0.40 -0.03  0.08 -0.28 -0.08  1.00 -0.51  0.02 -0.42  0.58
x7 -0.55 -0.30  0.15 -0.16  0.26 -0.51  1.00 -0.09  0.37 -0.58
x8  0.06  0.20  0.00 -0.05  0.17  0.02 -0.09  1.00 -0.32  0.21
x9 -0.08  0.09  0.12  0.08 -0.23 -0.42  0.37 -0.32  1.00 -0.34
y   0.75  0.19 -0.10 -0.15 -0.15  0.58 -0.58  0.21 -0.34  1.00
```

Confirma-se a grande diversidade de valores das correlações entre pares de variáveis neste conjunto de dados, que vão desde $r_{y,x_7} = -0.58$ até $r_{y,x_1} = 0.75$. Esta correlação positiva mais forte de todas tinha sido indiciada pelo *biplot* (que talvez deixasse prever uma correlação ainda mais forte). Confirma-se também a correlação quase nula entre $x_5$ e $x_6$ ($r_{x_5,x_6} = -0.08$). A correlação entre $x_2$ e $x_5$ ($r_{x_2,x_5} = -0.45$), sendo negativa e de relevo, é menos forte do que o *biplot* deixaria supôr. Estas discrepâncias estão associadas à proporção de variabilidade relativamente baixa associada às duas primeiras CPs, como já assinalado.

- As variáveis que parecem mais fortemente correlacionadas com a primeira CP são $y$, $x_1$ e $x_7$ (cujos vectores representativos estão quase na horizontal). A variável que parece mais fortemente correlacionada com a segunda CP é $x_4$, cujo vector representativo é quase vertical. A matriz de correlações entre variáveis originais e CPs confirma estas conclusões:

```
> round(cor(milho, milho.acp$x), d=2)
     PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10
x1  0.76  0.02 -0.36  0.05 -0.43  0.21 -0.05  0.10  0.00 -0.21
x2  0.29 -0.61  0.18  0.55  0.26 -0.14  0.31  0.04 -0.08 -0.11
x3 -0.18  0.52 -0.32  0.57  0.33  0.34 -0.16  0.11 -0.07  0.04
x4 -0.01 -0.77  0.27 -0.20  0.08  0.46 -0.15 -0.12 -0.23  0.01
x5 -0.31  0.73  0.31 -0.06 -0.08  0.29  0.37 -0.18 -0.06 -0.07
x6  0.71  0.36 -0.18 -0.13  0.36 -0.21 -0.14 -0.32 -0.13 -0.09
x7 -0.83  0.14 -0.03  0.07 -0.22 -0.27 -0.10  0.07 -0.39 -0.08
x8  0.23  0.17  0.71  0.48 -0.29 -0.06 -0.26 -0.14  0.08  0.01
x9 -0.48 -0.38 -0.57  0.32 -0.26  0.01  0.06 -0.34  0.07  0.05
y   0.88  0.13 -0.08  0.07 -0.23 -0.02  0.16  0.03 -0.25  0.23
```

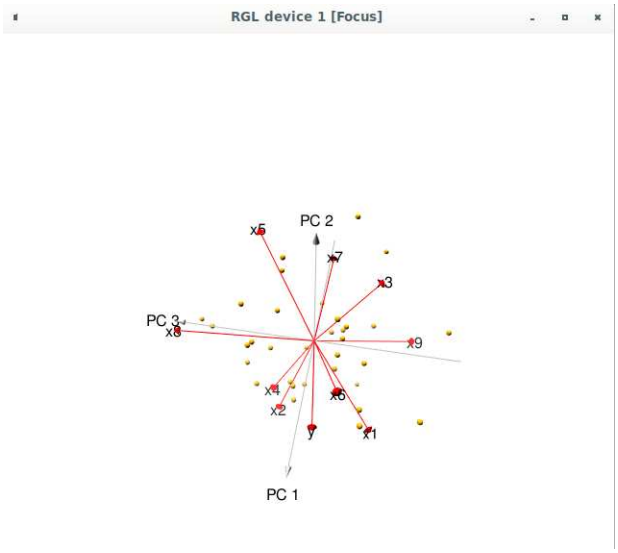
- O facto de na regressão linear múltipla de $y$ se ter seleccionado o submodelo com quatro preditores ($x_1$, $x_2$, $x_6$ e $x_9$) não é um facto que tenha relação directa com a ACP. Na regressão linear o objectivo último era prever valores de $y$. Na ACP é explicar variabilidade do conjunto de todas as variáveis. Os objectivos são diferentes. Não há nenhuma razão à partida que leve a pensar que variáveis com maior variabilidade tenham de ser bons preditores duma qualquer outra variável. Aliás, o primeiro plano principal é sempre este, qualquer que seja a variável resposta numa regressão linear múltipla envolvendo estas variáveis preditoras (o facto de $y$ estar bem correlacionado com a primeira CP é um acaso).

ii. A afirmação é verdadeira. De facto, os vectores representativos de $x_3$ e $x_5$ apontam (aproximadamente) no mesmo sentido. Mas a correlação entre essas variáveis é, na realidade, apenas 0.33, como vimos acima. Há um aspecto do *biplot* que indica o que deve ser a causa principal do problema. Tratando-se duma ACP sobre a matriz de correlações, os vectores representativos das variáveis têm, no espaço completo $\mathbb{R}^{10}$,

o mesmo comprimento (as variáveis normalizadas têm todas o mesmo desvio padrão: 1). No entanto, na representação bidimensional do *biplot* é visível que há vectores bastante mais curtos do que outros. Trata-se de vectores com componentes importantes nas 8 direcções que não saõ visíveis. Um desse vectores é $x_3$, que é assim um vector mal representado nas duas primeiras componentes principais. Este facto é visível se obtivermos a representação tri-dimensional do *biplot*, com o auxílio do módulo `pca3d`, no `R`. Usando o comando:

```
> pca3d(milho.acp, biplot=TRUE)
```

obtém-se uma janela gráfica que, rodada de forma conveniente, permite ver a seguinte representação do *biplot* a três dimensões, onde o plano definido pelas duas primeiras CPs está em profundidade à nossa frente, estando o eixo correspondente à CP3 visível na horizontal.




Nesta representação, pode ver-se que o ângulo entre os vectores que representam as variáveis $x_3$ e $x_5$ é, na realidade, bastante maior do que transparece no *biplot* a duas dimensões. De assinalar que há ainda sete dimensões não visualizadas (embora a variabilidade que lhes corresponde seja, no total, cerca de 25% da inércia total.

iii. Trata-se novamente duma afirmação verdadeira. Como se pode constatar, quer através da matriz de correlações entre variáveis originais e CPs, quer através da representação tri-dimensional do *biplot*, a variável $x_8$ está bastante fortemente correlacionada com a terceira CP. A sua projecção ortogonal sobre o primeiro plano principal (definido pelas CPs 1 e 2) perde boa parte do comprimento do vector (ou seja, da variabilidade associada à variável $x_8$), criando um vector projectado de pequena dimensão nesse plano.

10. 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
```

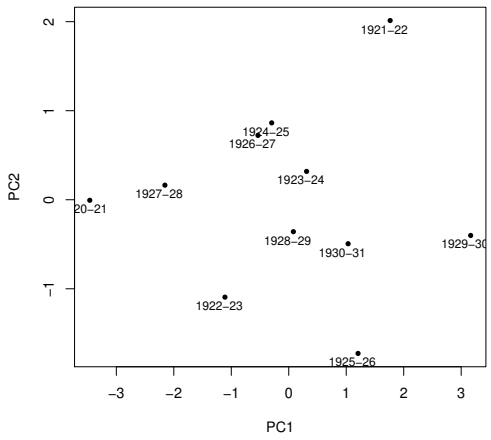
Cada coluna da matriz "`Rotation`" tem os coeficientes que definem cada CP (ou seja, que definem cada uma das combinações lineares das 5 variáveis originais $x1$ a $x5$).
A proporção de variabilidade, e variabilidade cumulativa, explicada pelas cinco CPs é esta:

```
> 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:

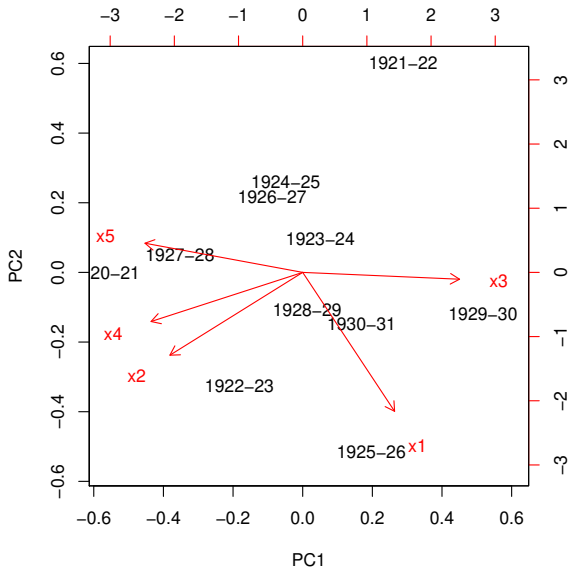
```
> round(cor(trigo, trigo.acpR$x), d=2)
     PC1   PC2   PC3   PC4   PC5
x1  0.54 -0.82  0.13 -0.07 -0.15
x2 -0.78 -0.49 -0.37  0.14  0.04
x3  0.92 -0.04 -0.27 -0.23  0.14
x4 -0.89 -0.29  0.23 -0.15  0.22
x5 -0.93  0.17 -0.12 -0.25 -0.20
```

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.

11. 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:

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

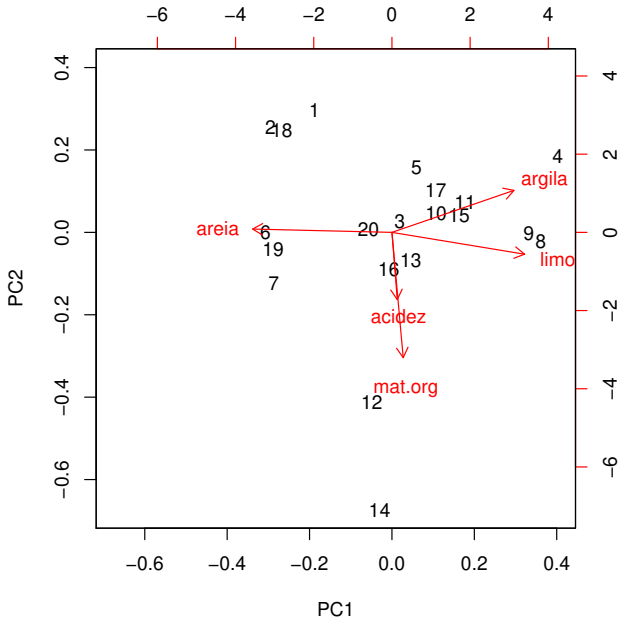
O correspondente vector próprio tem de ser da forma $\alpha\vec{\mathbf{v}}=(\alpha,\alpha,\alpha,0,0)^t$. Para ser de norma 1, tem de ter-se $\alpha=\pm\frac{1}{\sqrt{3}}=\pm 0.5773503$. Confirmemos:

```
> 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:

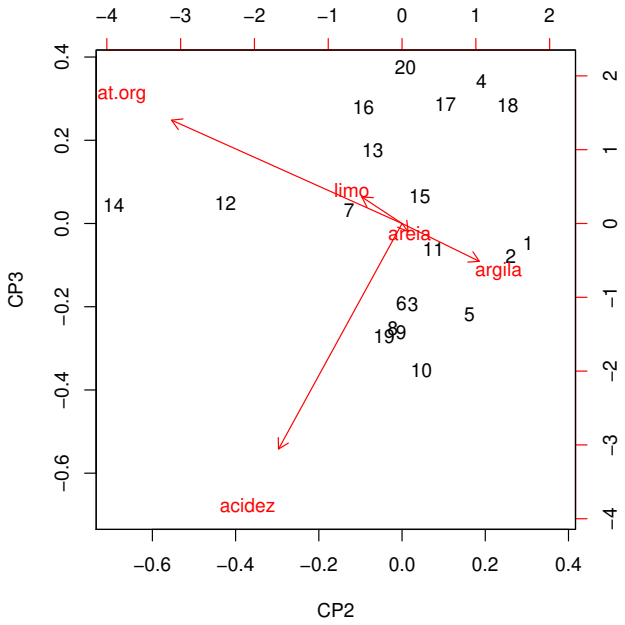
```
> round(cor(kendall),d=3)
         areia   limo argila mat.org acidez
areia    1.000 -0.972 -0.828  -0.100 -0.023
limo    -0.972  1.000  0.674   0.213  0.024
argila  -0.828  0.674  1.000  -0.196  0.013
mat.org -0.100  0.213 -0.196   1.000  0.079
acidez  -0.023  0.024  0.013   0.079  1.000
```

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*.

12. (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
```

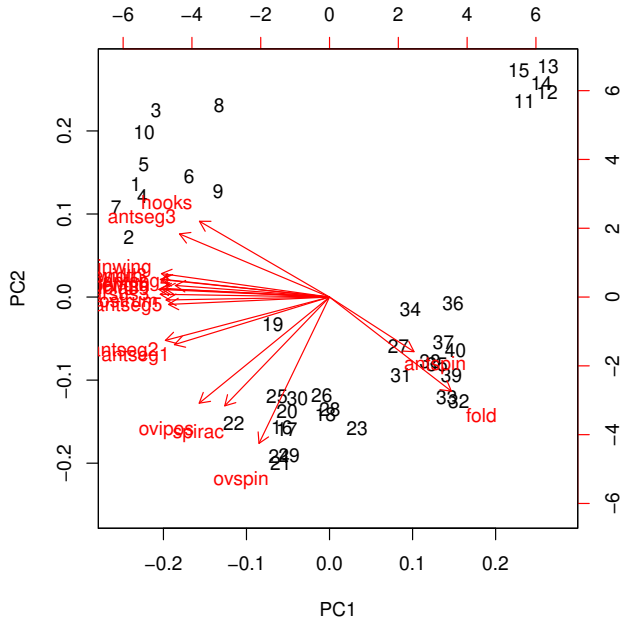
A primeira CP tem fortes correlações (o sinal negativo é arbitrário) com quase todas as variáveis, sendo uma espécie de medida da dimensão global dos organismos de afídios. As mais baixas correlações dizem respeito a variáveis que estão mais associadas a CPs posteriores. Esta íntima associação entre 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.

# 3 Análise Discriminante Linear

13. We need to load the `MASS` package, so that we can use the `lda` command, as follows:

```
> library(MASS)
```

(a) Here is the command to carry out the Linear Discriminant Analysis:

```
> lobos.lda <- lda(Grupos ~ . , data=lobos)
```

i. Among the output elements in the `list` produced by the `lda` command, `lobos.lda`, there is the object `scaling`, which is a matrix whose columns are the loadings for the discriminant axes:

```
> lobos.lda$scaling
           LD1        LD2         LD3
X1 -0.57520585  0.1763681 -0.01437166
X2 -0.25161855 -0.1319993 -0.05113646
X3  0.10067719  0.1809517 -0.11057288
X4  0.19937361  0.5451907  0.09317552
X5  0.21429307 -0.1745088  0.67165224
X6  0.18461849 -0.3544111 -0.70147042
X7 -0.03692864 -0.1041444  0.65096772
X8 -0.33219378 -0.2247208 -0.12674873
X9  1.76542145  1.8612996  0.21141396
```

There are only three discriminant axes since there are $k = 4$ groups of wolves. In the column with the heading `LD1` we find the coefficients of the linear combination (of the centred variables) which defines the first discriminant axis. In other words, the first (and best) discriminant axis is given by:

$$-0.57520585\,X_1 - 0.25161855\,X_2 + 0.10067719\,X_3 + ... + 1.76542145\,X_9$$

In order to establish its discriminant capacity, as defined in the course slides, we need to process further information. We know that the last otem in the displayed output, when we use the `lda` command, gives the proportion of the sum of discriminant capacities that is associated with each discriminant axis:

```
> lobos.lda
[...]
Proportion of trace:
   LD1    LD2    LD3
0.7436 0.2293 0.0272
```

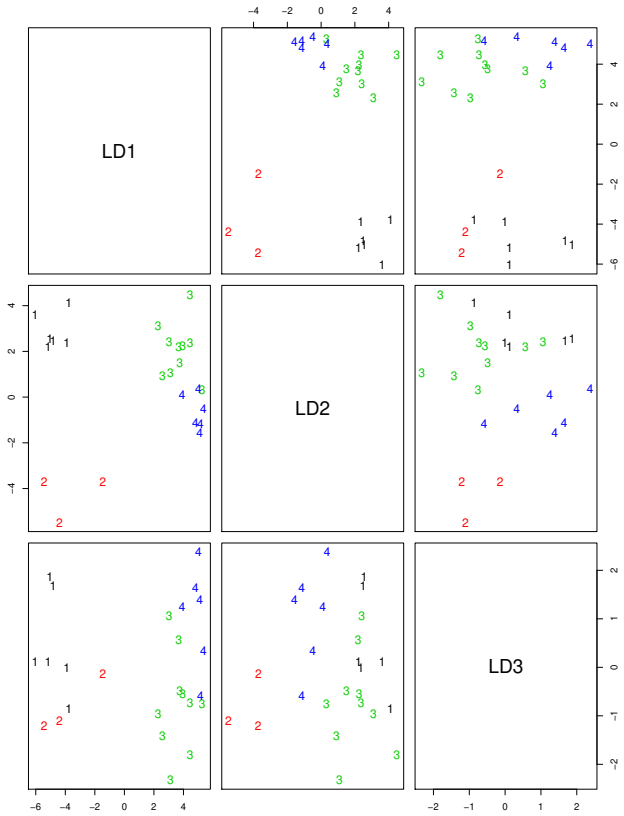
But this only tells us that the first discriminant axis is roughly three times as good as the second one in discriminant capacity. It does not measure the discriminant capacity of the first dsicriminant axis, which we defined as the ratio of between-class variability over within-class variability ($\frac{\mathbf{a}_1^t \mathbf{B} \mathbf{a}_1}{\mathbf{a}_1^t \mathbf{W} \mathbf{a}_1}$, where $\mathbf{a}_1$ is the vector of loadings shown above as `LD1` and $\mathbf{B}$ and $\mathbf{W}$ era, respectively, the between- and within-class variability matrices). This ratio corresponds to the largest eigenvalue of the matrix $\mathbf{W}^{-1}\mathbf{B}$ (using our class definitions). This eigenvalue can be obtained by multiplying the square of the values in the `svd` component of the output `lobos.lda` by the ratio $\frac{k-1}{n-k}$ which, in our case is defined using $k = 4$ and $n = 25$:

```
> lobos.lda$svd^2*3/21
[1] 20.5155273  6.3257322  0.7493596
```

Thus, the first discriminant axis has a between-group variability that is over 20.5 times larger than its within-group variability, indicating a strong discriminant capacity. The second discriminant axis still has a considerable discriminant capacity, with the between-class variability being almost six and a half times as big as the within-class variability. Only in the third and final discriminant axis is the within-class variability larger than the variability between groups.

ii. Since there are more than two discriminant axes ($k-1=3$), the `plot` method for objects of class `lda` automatically produces the 'matrix' of scatterplots requested in the question:

```
> plot(lobos.lda, col=as.numeric(lobos$Grupos))
```




As can be seen, discriminant axis 1 does a good job at separating the Rocky Mountain wolves to one side (groups 1 and 2, with negative scores) and the Arctic wolves on the other (groups 3 and 4, with positive scores). Discriminant axis 2 separates the males (groups 1 and 3, with positive scores) from the females (groups 2 and 4, mostly with negative scores), although ths separation is not quite as strong for Arctic wolves as it is for Rocky Mountain wolves. Thus, the scatterplot with `LD1` on the horizontal axis and `LD2` on the vertical axis leaves Rocky Mountain males as a separate group on the top left of the plot, Rocky Mountain females as an isolated group at the bottom left, Arctic males as a group on the top right and, with a slight overlap, Arctic females at the centre-right.

iii. To use the `predict` command, we must first create a data frame with the new wolf's observed values (using the same variable names as in the original data frame `lobos`):

```
> lobo.novo <- data.frame(X1=125, X2=104, X3=145, X4=81.1, X5=33.2, X6=68.2,
```

```
+                                X7=49.0, X8=43.3, X9=18.2)
> predict(lobos.lda, new=lobo.novo)
$class
[1] 1
[...]
$x
      LD1      LD2        LD3
1 -4.0297 1.532802 -0.4920112
```

The output object `class` is using `lda`'s classification rule to tell us that the new wolf should be considered as belonging to class 1, that is, a Rocky Mountain male. The object `x` gives the individual's coordinates on the three discriminant axes. By visually inspecting the scatteplots above, it is not hard to agree with this conclusion: the new wolf's representative point would be in the top left corner of the scatterplot discussed above.
To further help us explore the classification, it is also useful to see yet another object in the `lda` output list: the object `means` which gives us the centres of gravity of the individuals in each group:

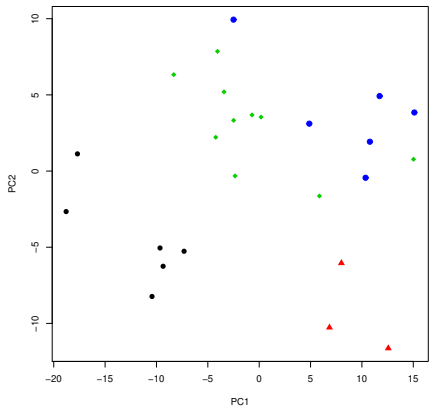
```
> lobos.lda$means
        X1        X2       X3       X4       X5       X6       X7       X8       X9
1 126.5000 108.16667 145.1667 82.11667 33.55000 67.28333 49.41667 43.20000 18.18333
2 117.3333 102.66667 128.6667 75.50000 31.10000 63.73333 43.70000 41.43333 16.90000
3 115.8000 100.80000 142.4000 81.68000 33.53000 67.07000 45.54000 39.99000 17.98000
4 110.8333  96.16667 137.0000 79.23333 32.33333 64.73333 46.18333 39.96667 17.38333
```

By comparing the new wolf's observed values with these group means for the nine variables, the classification appears reasonable.

(b) The question does not clarify whether a covariance-matrix PCA or a correlation-matrix PCA should be used. In this case, both options are available since all the variables have the same units of measurement: millimeters. Since the purpose is to see how well PCs can discriminate between the groups, both options were inspected and it was found that covariance-matrix provides PCs that best separate the subgroups. This is not entirely surprising, since covariance matrix PCA will use the variables with largest variances to define the first few PCs, and it is to be expected that this greater variability will help to distinguish the morphometric characteristics of the four groups.

Here is the covariance matrix PCA of the dataset, and the resulting scatterplot on the first principal plane (using both different colours and characters for each group):

```
> lobos.acp <- prcomp(lobos[,-10])
> summary(lobos.acp)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7    PC8    PC9
Standard deviation     9.7562 5.6691 2.92566 2.16911 1.82639 1.59417 1.34039 0.7726 0.4573
Proportion of Variance 0.6385 0.2156 0.05742 0.03156 0.02238 0.01705 0.01205 0.0040 0.0014
Cumulative Proportion  0.6385 0.8541 0.91155 0.94311 0.96549 0.98254 0.99459 0.9986 1.0000

> plot(lobos.acp$x[,1:2], col=as.numeric(lobos$Grupos), pch=15+as.numeric(lobos$Grupos))
```

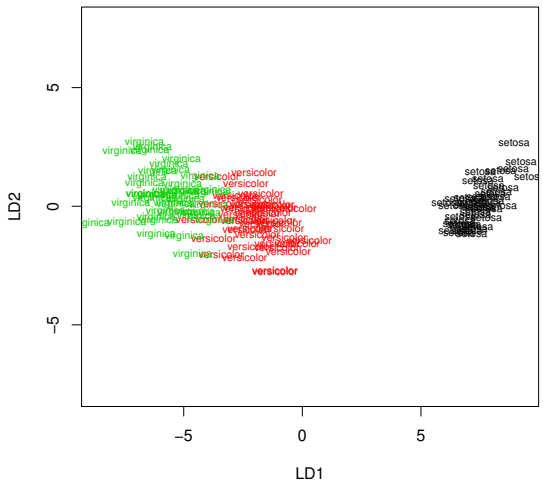



A quick assessment of the results shows that some separation of the four groups is visible on the first principal plane, but not nearly as good a job as the first two discriminant axes. This is not unexpected, since both sets of new variables are linear combinations of the original variables, but the discriminant axes were explicitely built to optimize the discriminant capacity, and so will always do at least as well as the PCs, but usually considerably better. The discriminant capacity of each PC can actually be measured, by talking the ratios $\frac{\vec{\mathbf{v}}_j^t \mathbf{B} \vec{\mathbf{v}}_j}{\vec{\mathbf{v}}_j^t \mathbf{W} \vec{\mathbf{v}}_j}$, where $\vec{\mathbf{v}}_j$ is the vector of loadings (eigenvector of matrix $\mathbf{S}$) of the $j$-th PC. It should be noted that there is no reason why subsequent PCs may not be helpful in the discrimination, nor is there any reason why the discriminating capacity of any PC (including the first) has to be larger than that of any subsequent PC. For this dataset, the discriminant capacity of PC1 is 2.1914 and that of PC2 is slightly larger: 2.6886.

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

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

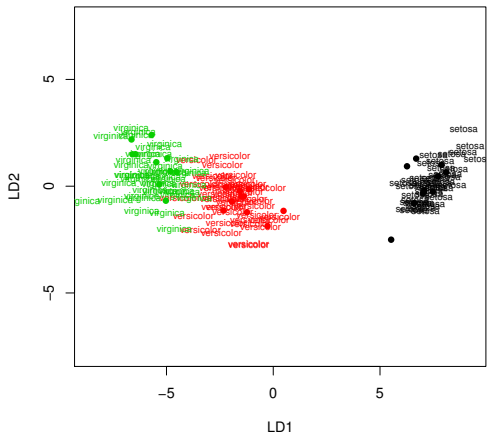
```
> iris.lda <- lda(Species ~ .  , data=iris[c(1:40,51:90,101:140),])

> 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.

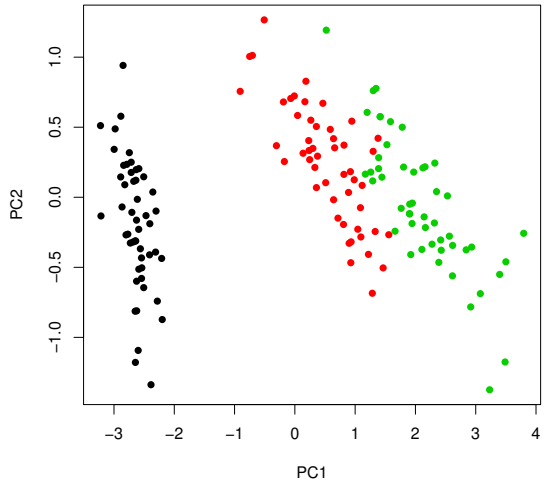
```
> iris.ldaPred$class
[1] setosa     setosa     setosa     setosa     setosa     setosa
[7] setosa     setosa     setosa     setosa     versicolor versicolor
[13] versicolor versicolor versicolor versicolor versicolor versicolor
[19] versicolor versicolor virginica  virginica  virginica  virginica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

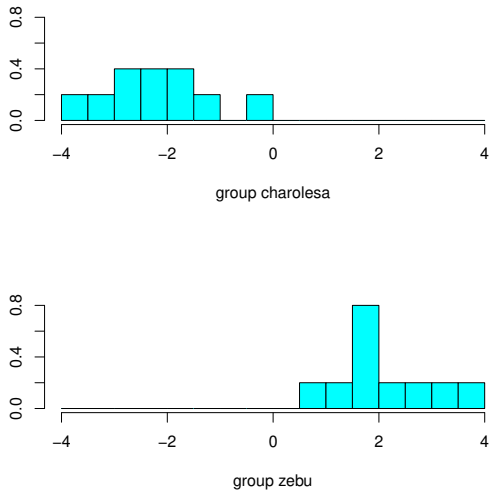
15. Neste exercício, existem apenas $k = 2$ classes (zebus e charolesas), o que significa que apenas existirá $k - 1 = 1$ eixo discriminante. Este facto introduz algumas especifidades aquando da apresentação dos resultados e da sua representação gráfica. Assinale-se também que para poder

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
[...]
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).

16. 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.

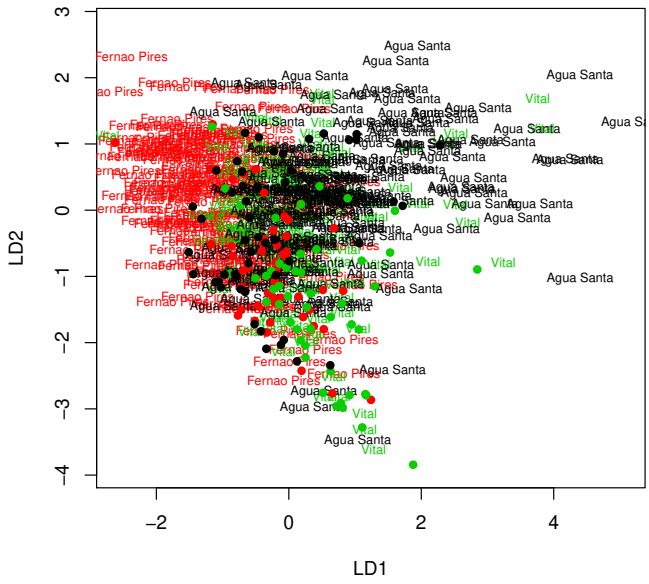
```
> library(MASS)
> vid.treino <- videiras[c(1:150, 201:350, 401:550),]
> vid.lda2 <- lda(Casta ~ . , data=vid.treino)

> 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*).

17. 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 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 16 (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 15 (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 16 (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 16:

```
> 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
```