Exercises - Modelos Matemáticos e Aplicações
Introduction to Multivariate Statistics - 2020-21

Note: The datasets for some of this Module's exercises can be found on the course webpage (Section *Materiais de Apoio*, *Módulo III*. The datasets are in a file called `dadosMulti.RData` (the extension indicates that this file may be loaded into an `R` session, with the command `load`). The file contains the following data frames: `santarem` (Exercise 7), `brix2` (Exercise 8), `trigo` (Exercise 10), `kendall` (Exercise 11), `adelges` (Exercise 12), `lobos` (Exercise 13) e `diday` (Exercise 15). The file also contains the data frame `lavagantes`, with the dataset discussed in the slides.

1 Matrices and Linear Algebra

1. Consider the linear space $\mathbb{R}^2$. Let M be a subspace of $\mathbb{R}^2$ spanned by vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Let N be the subspace of $\mathbb{R}^2$ spanned by vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.
   (a) Characterize the vectors that are in subspace M.
   (b) What is the orthogonal projection of the vector $\begin{bmatrix} c \\ d \end{bmatrix}$ onto the subspace M?
   (c) Characterize the vectors of subspace N.
   (d) What is the orthogonal projection of vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ onto the subspace N?
2. Consider the space $\mathbb{R}^n$ with its usual inner product $< \vec{\mathbf{x}}, \vec{\mathbf{y}} >= \vec{\mathbf{x}}^t \vec{\mathbf{y}}$.
   (a) Characterize the vectors in $\mathbb{R}^n$ that are orthogonal to the vector of n ones, $\mathbf{1}_n$.
   (b) Associate the points/vectors in $\mathbb{R}^n$ with sets of n observations on a given variable. From a statistical point of view, how can the elements of the subspace described in the previous question be interpreted?
3. Let $\vec{\mathbf{y}} \in \mathbb{R}^n$ be the vector representation of n observations of a given variable. Let $\vec{\mathbf{y}}^c \in \mathbb{R}^n$ be the corresponding centred vector.
   (a) Discuss the effect of a translation of the origin in the units of measurement of the variable (that is, $y_i \to a + y_i$) on the vectors $\vec{\mathbf{y}}$ and $\vec{\mathbf{y}}^c$.
   (b) Discuss the effect of a multiplicative change of scale ($y_i \to b\, y_i, \forall i$) on the vectors $\vec{\mathbf{y}}$ and $\vec{\mathbf{y}}^c$.
   (c) Discuss the effect of a linear transformation $y_i \to a + by_i, \forall i$, on the vectors $\vec{\mathbf{y}}$ and $\vec{\mathbf{y}}^c$.

   Now consider a second vector $\vec{\mathbf{x}} \in \mathbb{R}^n$ representing observations of a new variable on the same n individuals. Let $\vec{\mathbf{x}}^c$ be the corresponding centred vector.
   (d) Discuss the effect of different linear trasformations of the two variables ($x_i \to a + bx_i$ and $y_i \to c + dy_i$, $\forall i$) on the vectors that represent them in $\mathbb{R}^n$. Discuss the influence of those transformations on the statistical indicators covariance and correlation coefficient.

4. Consider the matrices $\mathbf{X}^t\mathbf{X}$ and $\mathbf{X}\mathbf{X}^t$, where $\mathbf{X}$ is an $n \times p$ matrix. Confirm that, if $\lambda_j \neq 0$ is an eigenvalue of $\mathbf{X}^t\mathbf{X}$, with corresponding eigenvector $\vec{\mathbf{c}}_j$, then $\mathbf{X}\vec{\mathbf{c}}_j$ is an eigenvector of matrix $\mathbf{X}\mathbf{X}^t$, with the same eigenvalue. Conversely, if $\lambda_j \neq 0$ is an eigenvalue $\mathbf{X}\mathbf{X}^t$ with corresponding eigenvector $\vec{\mathbf{b}}_j$, then $\mathbf{X}^t\vec{\mathbf{b}}_j$ is an eigenvector of $\mathbf{X}^t\mathbf{X}$, witht the same eigenvalue.

5. Use the Singular Value Decomposition of a matrix $\mathbf{Y}$, given by:

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

to show that if $\vec{\mathbf{w}}_i$ is a left singular vector associated with the singular value δ_i and $\vec{\mathbf{v}}_i$ is the corresponding right singular vector, then:

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

6. Consider a matrix $\mathbf{B}$ and the matrix of orthogonal projections onto the subspace spanned by the columns of $\mathbf{B}$, $\mathbf{P}_B = \mathbf{B}(\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t$. Using the Singular Value Decomposition of matrix $\mathbf{B}$, find an alternative expression for matrix $\mathbf{P}_B$. Comment.

2 Principal Component Analysis

7. In the 1973 Agricultural Statistics (Estatísticas Agrícolas) of Portugal's National Statistics Board (Instituto Nacional de Estatística, INE), productivities (in t/ha) of 9 agricultural products are given for each of the 20 municipalities of the Santarém distrcit. The data are shown below, and can be found in the `santarem` data frame, whch is available on the course website, in file `dadosACP.RData`.

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

Here is the variance-covariance matrix for this dataset:

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

(a) Consider a Principal Component Analysis on the covariance matrix of the data (that is, on the dataset in its original units).

i. Discuss the quality of the dimensionality reduction which can be achieved with PCA.

ii. Based on the results produced by the `prcomp` command, draw the 20-point scatterplot showing the municipalities on the plane defined by the first two principal components. Identify the 7 municipalities that appear on the right half of the plot. Also, identify the point that appears, by itself, in the top left corner.

iii. Calculate, using `R`, the coefficients of linear correlation between PC 1 and each of the nine original variables. Confirm the values of the three correlation coefficients between the first proncipal component and the variables `batata` (potato), `fava` (broadbeans) and `milho` (maize), using the formula shown in the slides. Repeat for the second principal component. Discuss.

iv. Try to interpret the nature of the first two principal components. Justify your reply.

v. Build the correspnding biplot and discuss it.

vi. Critically assess the Principal Components Analysis (PCA) that you carried out, discussing in particular the decision to use a covariance-matrix PCA.

(b) Now carry out a Principal Component Analysis of the normalized data, that is, based on the correlation matrix.

i. Discuss the quality of the reduction in dimensionality that can be obtained by a correlation-matrix PCA. Comment it, also taking into account the results of the PCA on the original data.

ii. Calculate the correlation coefficients between each of the original variables and each of the PCs that were now obtained. Is it necessary to standardize the variables in order to compute these correlation coefficients?

iii. Draw the relevant biplot and discuss it. In particular, try to interpret the nature of the first two principal components of the normalized data.

(c) Answer the following question by a user: "*which of the PCA variants should I use in this case*"?

8. In a study of greenhouse raspberries, 7 variables characterizing the properties of picked fruits were observed. Specifically, raspberries were collected from 14 different plants and their mean value for each plant were recorded, for the following variables: `Diametro` (diameter), `Altura` (height), `Peso` (weight), `Brix`, `pH`, a different measure of acidity, which will be called `Acidez`, and `Acucar` (sugar content). The resulting values are given in the data frame `brix2` (the dataset was already studied in Module II, but there is now the new variable `Acidez`):

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

(a) State, justifying your answer, whether a Principal Component Analysis on the covariance matrix is suitable for this dataset.

(b) State, justifying your answer, whether a Principal Component Analysis on the correlation matrix provides a suitable two-dimensional representation of the data, without substantial loss of information.

(c) Regardless of your answers above, build a biplot for the data. Discuss it.

(d) The 14 plants were not all observed on the same dates. The fruits from each plant were collected on five different dates:

Date	Plants
November 28	1,2,3,4
December 13	5,6,7,8
January 16	9,10,11,12
February 20	13
April 3	14

Are the different dates of collection reflected in the first principal plane of the standardized data? In your reply, identify which points in the scatterplot are associated with each date.

(e) *If your reply to the previous question was 'yes'* state, with justification, whether it would necessarily have to be the case that this sub-group structure is reflected in the firat principal plane. *If your answer was 'no'*, state why such structure does not have to be reflected in the first proncipal plane, given the optimizing properties of the first two principal components.

(f) Now assume that a new plant's raspeberries were observed, with the following mean values for each (in order) variable: 1.9, 2.0, 3.92, 8.1, 2.91, 1.48, 3.78. If you wish to represent this new observation on the first principal plane, what coordinates should it have? Justify your answer and draw the new point on the first principal plane. Confirm your answer, using `R`'s `predict` command, which also has a method for objects obtained resulting from PCAs obtained with the `prcomp` command. This command is used in a similar way to the `predict` command for linear, or generalized linear, models.

9. Consider the data for the production of corn in the US State of Iowa, already studied in Module II, and which can be found in the data frame `milho`.

(a) Which variant of PCA (covariance matrix or correlation matrix) do you consider suitable for this dataset? Justify your reply.

(b) How good is the dimensionality reduction provided by a PCA on the 10 standardized variables?

(c) Build a *biplot* for the correlation matrix PCA.

i. Comment the *biplot*, also taking into consideration the multiple linear regression submodel that was chosen by all the subset selection methods, and which resulted in modelling y based on the four predictors $x1$, $x2$, $x6$ and $x9$. Is it possible to make any comment regarding this choice, based on the *biplot*?

ii. Comment the following statement: "*The biplot suggests that variables* $x3$ *and* $x5$ *are strongly correlated, but this conclusion is not confirmed by the correlation matrix between the* 10 *variables*".

iii. Comment the following statement: "*Since this was a correlation matrix PCA, all the vectors representing the centred variables should be of equal length. The fact that variable* $x8$ *is represented in the biplot by a much shorter vector than all the rest suggests that this variable is poorly represented on the plane defined by the first two standardized PCs*".

10. An old study carried out in Belgium (Berce e Wilbaux, 1935 *Recherche Statistique des relations existant entre le rendement des plantes de grandes cultures et les facteurs météorologiques en Belgique*. Bull. Inst. Agron. Stn. Rech. Gembloux, **4**, 32–81), recorded $p = 5$ meteorological and agronomical variables throughout $n = 11$ agricultural seasons. The five variables were:

x_1	total rainfall in November and December (mm)
x_2	mean temperature in July (^{o}C)
x_3	total rainfall in July (mm)
x_4	radiation in July (mm of alcohol)
x_5	mean yield of durum wheat (quintals/ha)

The observed values were:

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

(a) Carry out a correlation matrix Principal Component Analysis for this dataset, identifying the five Principal Components.

(b) Build the best possible two-dimensional representation of the $n = 11$ point scatterplot in $\mathbb{R}^5$ for the data.

(c) Calculate the correlation coefficients between the first Principal Component and each of the five original variables. Interpret your results.

(d) Some units of measurement are now outdated. The most frequent units of measurement for yield are tons per hectare, which means that the values of variable $x5$ should be divided by 10. On the other hand, the metric system units for radiation are $MJ\,m^{-2}$, which means that to convert the values of variable $x4$ to these units, the following affine transformation is needed: $x_4^* = -0.02960342 + 0.75518263\,x_4$. How do these changes in units affect the replies to the above questions? Confirm your answer using `R`.

11. Consider the following data set, discussed by Kendall (*Multivariate Analysis*, Charles Griffin & Co., 1980, pg. 20), and with measurements for 20 soil samples:

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

(a) Carry out a covariance matrix Principal Component Analysis of the dataset. Explain the existence of a zero eigenvalue and the nature of the corresponding eigenvector.

(b) Build the biplot associated with the PCA on the standardized data. The relative positions of the vectors representing the variables `acidez` and `mat.org` (organic matter) suggests that these are two highly correlated variables. However, this fact is not confirmed by the correlation matrix between the original variables. How can this apparent contradiction be explained?

(c) Now drop the variable `areia` (sand content) from the data matrix. Repeat the covariance matrix PCA.

i. Calculate the correlation coefficient between each Principal Component and each variable.

ii. Compare the values obtained with the variable loadings in the linear combinations defining the PCs and note how the attept to interpret Principal Components only in terms of the coefficients (*loadings*) may be misleading.

12. In a study of winged aphids *Alate adelges* (D.F. Morrison, *Multivariate Statistical Methods*, p.477) measurements of 19 variables were taken on 40 individuals. The 19 observed variables, as well as the means and variances of the observed values were:

Name	Accronym	Description	$\overline{x}$	s^2
length	COM	body length	15.05	14.58
width	LAR	body width	7.14	4.05
forwing	CAA	fore-wing length	5.68	1.68
hinwing	CAP	hind-wing length	3.45	0.83
spirac	E	number of spiracles	4.88	0.11
antseg1	AS1	length of antennal segment I	1.86	0.11
antseg2	AS2	length of antennal segment II	1.69	0.11
antseg3	AS3	length of antennal segment III	2.25	0.22
antseg4	AS4	length of antennal segment IV	2.33	0.15
antseg5	AS5	length of antennal segment V	2.73	0.15
antspin	S	number of antennal spines	4.28	1.33
tarsus3	TAR	leg length, tarsus III	3.31	0.41
tibia3	TIB	leg length, tibia III	3.38	0.58
femur3	FEM	leg length femur III	2.57	0.34
rostrum	ROS	rostrum	5.58	0.79
ovipos	OVI	ovipositor	3.72	0.35
ovspin	N	number of ovipositor spines	7.80	3.81
fold	P	anal fold (no/yes - 0/1 variable)	0.73	0.20
hooks	GAP	number of hind-wing hooks	2.38	0.25

These were the observations:

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

(a) Briefly describe the main characteristics of the bundle of vectors representing the 19 centred, *but not standardized*, variables in the space of variables, $\mathbb{R}^{40}$.

(b) Carry out a correlation matrix Principal Component Analysis of the data.

   i. Attempt to interpret the first three principal components, based on the available information. Justify your comments.

   ii. Do you consider a two-dimensional graphical representation adequate? Justify your reply. Identify a variable whose representation on the first principal plane is not very good, justifying your answer.

iii. The projected scatterplot of points on the plane defined by the first two principal components seems to more or less clearly separate groups of individuals. Relate these groups to the original variables and comment.

iv. Datasets with 19 variables for which a correlation matrix PCA explains such a high proportion of total variance on the first 2 or 3 PCs are not frequent. How can this feature be justified in the case of this dataset?

v. Critically assess this PCA, taking into account the nature of some of these 19 variables. If you see some undesirable features, suggest alternatives.

3 Discriminant Analysis

13. The book by D.F. Morrison, *Multivariate Statistical Methods* (p.288), has data from a study involving nine morphometric variables on the skulls of wolves (*Canis lupus* L.): palatal length (X_1); postpalatal length (X_2); zygomatic width (X_3); palatal width outside the first upper molars (X_4); palatal width inside the second upper premolars (X_5); width between the postglenoid foramina (X_6); interorbital width (X_7); least width of the braincase (X_8); crown length of the first upper molar (X_9). All measurements are in mm. There are measurements for 25 individuals, who belong to 4 groups: (1) 6 Rocky Mountain males; (2) 3 Rocky Mountain females; (3) 10 Arctic males; and (4) 6 Arctic females. The data can be found in the data frame `lobos`, and are reproduced in the table below.

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

(a) Perform a Linear Discriminant Analysis with the `lda` command in `R`'s `MASS` package.

i. What is the first discriminant (canonical) variable? What is its discriminating capacity? Comment.

ii. Use R's `plot` command to visualize the scatterplots on the planes defined by the first three discriminant axes. Discuss your results.

iii. To which of the four groups would you associate a new set of observations, for a wolf of unknown sex and habitat, with the following values on the nine observed variables: 125, 104, 145, 81.1, 33.2, 68.2, 49.0, 43.3, 18.2? Use the command `predict`, which has a method for objects of class `lda`.

(b) Carry out a Principal Component Analysis on the set of observations of the 9 numerical variables, on the 25 individuals. In particular, assess the planes defined defined by each pair of PCs. Compare with the results of the LDA. Comment the discriminant capacity of the Principal Components.

14. Carry out a Discriminant Analysis of the 150 iris flowers of the data frame `iris`, obtaining linear functions to discriminate the three iris varieties. In particular,

(a) Use the first 40 individuals from each species to define the discriminant axes (i.e., as a *training set*).

(b) Classify the remaining 30 individuals (i.e., the *validation set*), using the discriminant axes defined above (use the classification provided by R's `predict` command).

(c) Build a table comparing the true species of the 30 observations in the validation set with these classifications produced by the Linear Discriminant Analysis. Discuss.

(d) Compare the projection of the 150 individuals on the first principal plane, defined by a Principal Component Analysis of the data. Discuss.

15. Three variables (v_1, v_2 e v_3) were observed on each of ten zebus and ten Charolais cattle. The resulting values (data in Diday *et. al.*, 1982) are shown below and are available in the data frame `diday`:

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

Perform a Linear Discriminant Analysis of the data and say whether you think the three variables provide a good discrimination of zebus and Charolais cattle.

16. Consider the `videiras` dataset, studied in Module II, with measurements of vineleaf surface area and lengths of main vein and left and right lateral veins, for $n=200$ leaves of each of three varieties.

(a) Perform a Linear Discriminant Analysis, seeking to discriminate the grape varieties based on the 4 observed numerical variables. Comment the result.

(b) Confirm that the vectors of coefficients (*loadings*) of the discriminant axes are not orthogonal to each other, but that the resulting new discriminant variables (vectors of *scores*) are uncorrelated to each other. **Note:** In `R`, the vectors of *loadings* can be obtained by applying the `coef` command to the results of the `lda` command; the vectors of *scores* result from applying the `predict` command to the results of `lda` and selecting object `x`.

17. Write an `R` function to carry out a Linear Discriminant Analysis. This function should accept as arguments:

- a matrix or *data frame* with the values of the variables;
- a vector or `factor` indicating to which of the k subgroups each observation belongs.

The function must compute and output:

- the matrix of between-class (inter-class) variability, $\mathbf{B}$;
- the matrix of within-class (intra-class) variability, $\mathbf{W}$;
- the eigenvalues and eigenvectors of matrix $\mathbf{W}^{-1}\mathbf{B}$;
- the discriminant axes (that is, the $k-1$ linear combination of the centred variables which are defined by the eigenvectors $\mathbf{W}^{-1}\mathbf{B}$ associated with non-zero eigenvalues).

If $k > 1$, the function should also output:

- the centres of gravity for each of the k scatterplots of points in each group, on each of the discriminant axes.
- the covariance matrices for each group, on all discriminant axes.

Note: Matrix $\mathbf{W}^{-1}\mathbf{B}$ is not symmetric, so that using the `R`'s `eigen` command may produce (artificially) complex eigenvalues and eigenvectors. The `Re` command may be used to extract the real part of these (false) complex numbers.