Data Mining

Devoir surveillé, 6 novembre 2018

General advices:

- the responses must be **synthetic and brief**,
- the copy must be **clean and readable**.

Exercice 1. PCA.

The dataset `temperatures` describes 15 cities on 4 numerical variables:

- the latitude (North-South position),
- the longitude (East-West position),
- the mean temperature,
- the amplitude of temperatures.

The R code below gives the description of the cities in the dataset.

```
X[,]
```

```
##             Latitude Longitude MeanTemp Amplitude
## Bordeaux        44.5     -0.34    13.33      15.4
## Brest           48.2     -4.29    10.77      10.2
## Clermont        45.5      3.05    10.94      16.8
## Grenoble        45.1      5.43    10.98      18.6
## Lille           50.4      3.04     9.73      14.7
## Lyon            45.5      4.51    11.36      18.6
## Marseille       43.2      5.24    14.23      17.8
## Montpellier     43.4      3.53    13.89      17.1
## Nantes          47.1     -1.33    11.69      13.8
## Nice            43.4      7.15    14.84      15.2
## Paris           48.5      2.20    11.18      15.7
## Rennes          48.0     -1.41    11.13      13.1
## Strasbourg      48.4      7.45     9.72      18.6
## Toulouse        43.4      1.26    12.68      16.2
## Vichy           46.1      3.26    10.72      16.9
```

First part

Let $\mathbf{X}$ denote the $n \times p$ data matrix and $\mathbf{Z}$ denote the standardized data matrix.

1. Give the values of n and p.
2. The squared Euclidean distance between `Bordeaux` and `Brest` in $\mathbf{X}$ is equal to 63.52 and the squared Euclidean distance between `Bordeaux` and `Toulouse` is equal to 0.37. Interpret and compare these two distances using the variance of the columns of $\mathbf{X}$ below.

```
apply(X,2,var)
```

```
##  Latitude Longitude  MeanTemp Amplitude
##      5.27     11.01      2.57      5.41
```

3. Perform now the squared Euclidean distance between `Bordeaux` and `Brest` described in $\mathbf{Z}$. What is the advantage of using $\mathbf{Z}$ instead of $\mathbf{X}$ when calculating Euclidean distances ?

Second part

The **generalized** singular value decomposition (GSVD) of the matrix $\mathbf{Z}$ with the metrics $\frac{1}{n}\mathbb{I}_n$ on $\mathbb{R}^n$ and $\mathbb{I}_p$ on $\mathbb{R}^p$ gives:

$$\mathbf{Z} = \mathbf{U}\Lambda\mathbf{V}^T$$

with

- $\Lambda = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ the diagonal matrix of the singular values of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$ and $\frac{1}{n}\mathbf{Z}^T\mathbf{Z}$ and r the rank of $\mathbf{Z}$,
- $\mathbf{U}$ the $n \times r$ matrix of the r eigenvectors of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^T$ and $\frac{1}{n}\mathbf{U}^T\mathbf{U} = \mathbb{I}_r$
- $\mathbf{V}$ the $p \times r$ matrix of the r eigenvectors of $\frac{1}{n}\mathbf{Z}^T\mathbf{Z}$ and $\mathbf{V}^T\mathbf{V} = \mathbb{I}_r$.

Let us use this GSVD to perfom the principal component analysis (PCA on the **correlation matrix**) of the data matrix $\mathbf{X}$. Results of the GSVD are given below.

```
e <- gsvd(Z,rep(1/n,n), rep(1,p))
```

```
#singular values
e$d
```

```
## [1] 1.53 1.17 0.50 0.24
```

```
#two first left singular vectors
e$U[1:5,1:2] #5 first cities
```

```
##           dim1  dim2
## Bordeaux  0.16 -1.22
## Brest    -2.10 -1.24
## Clermont  0.11  0.43
## Grenoble  0.69  0.99
## Lille    -1.22  1.21
```

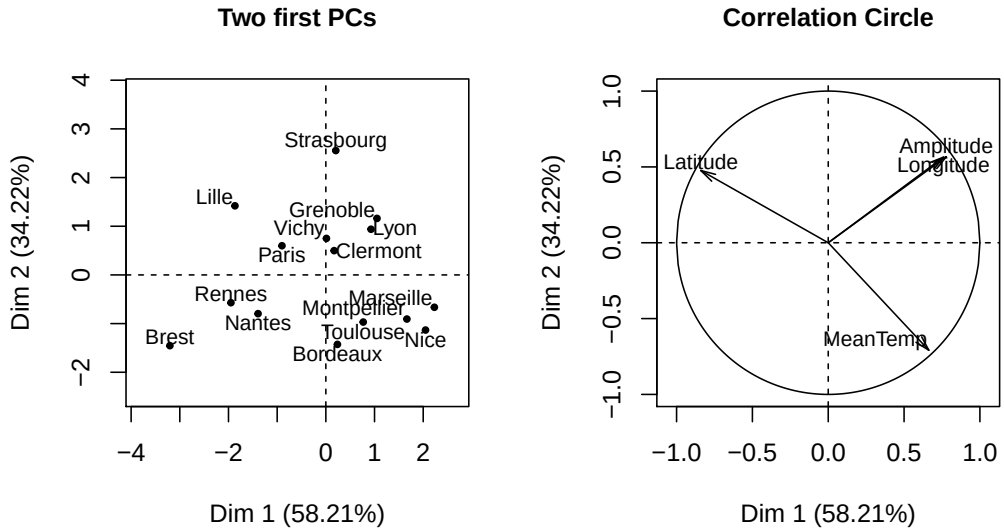
```
#two first right singular vectors
e$V[,1:2]
```

```
##            dim1  dim2
## Latitude  -0.55  0.41
## Longitude  0.50  0.48
## MeanTemp   0.43 -0.61
## Amplitude  0.51  0.48
```

1. Let $\mathbf{F}$ (resp. $\mathbf{A}$) denote the matrix of dimension $n \times r$ (resp. $p \times r$) of the factor coordinates of the n cities (resp. the p variables). Use the GSVD decomposition of $\mathbf{Z}$ to show that $\mathbf{F} = \mathbf{U}\Lambda$ and $\mathbf{A} = \mathbf{V}\Lambda$.
2. Perform (with **simple calculations**) the factor coordinates on the two first principal dimensions of:
   - the city `Bordeaux`,
   - the variable `Latitude`.
3. What is the correlation between `Latitude` and the first principal component ? What is the part of inertia explained by the two first principal components ?

Third part

The PCA gives the plots below.



1. Remind the general rule to interpret the angles between two variables in the correlation circle and give a **synthetic** interpretation the correlation circle above.
2. Remind the general property used interpret the position of the cities (left plot) according to the position of the variables in the correlation circle (right plot) and give a **synthetic** interpretation of the position of the group of cities `Marseille`, `Montpellier` and `Nice` on the plot above.

Exercice 2. MCA.

The dataset `customers` describes 28 customers of a repair garage on 5 categorical variables. Each variable corresponds to a question asked to the customers one week after they had picked up their car.

- Are you globally satisfied by the service? (Yes/No)
- Do you consider the problem is solved? (Yes/No/Don't know)
- How good was the welcome? (1 to 5)
- Is the quality/price ratio satisfactory? (Yes/No)
- Will you use our services again? (Yes/No/Don't know)

The R code below gives the desription of the 5 first customers in the dataset.

```
X[1:5,]
```

```
##    Satisfied Repaired Welcome Q.Price Come.back
## C1       Yes      Yes       5     Yes       Yes
## C2       Yes      Yes       4     Yes        dk
## C3       Yes      Yes       4     Yes        dk
## C4       Yes       dk       4     Yes        dk
## C5       Yes       dk       4     Yes       Yes
```

A short description of the 5 categorical variables is given below.

```
lapply(customers, table)
```

```
## $Satisfied
##
##  No Yes
##  13  15
##
## $Repaired
##
##  dk  No Yes
##   7   5  16
##
## $Welcome
##
## 1 2 3 4 5
## 6 4 7 7 4
##
## $Q.Price
##
##  No Yes
##  18  10
##
## $Come.back
##
##  dk  No Yes
##  13  11   4
```

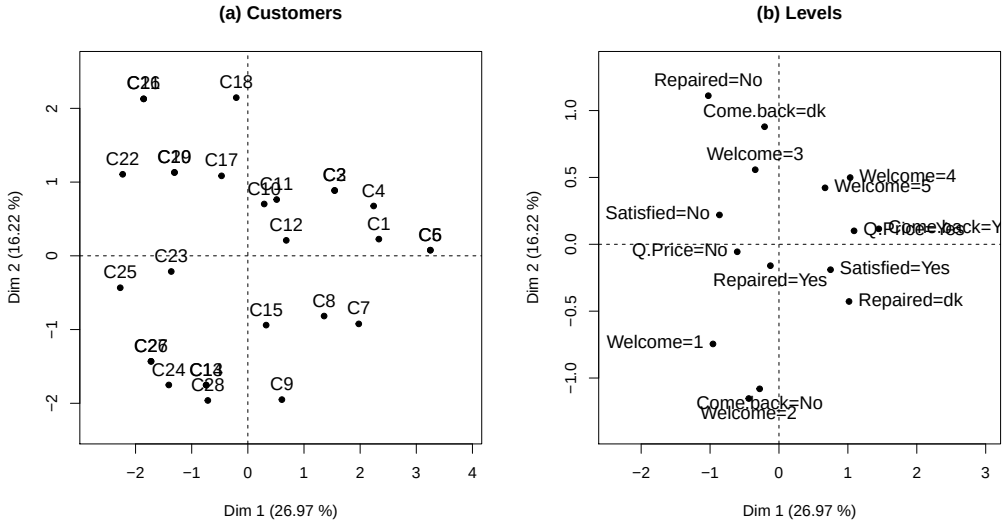
First part

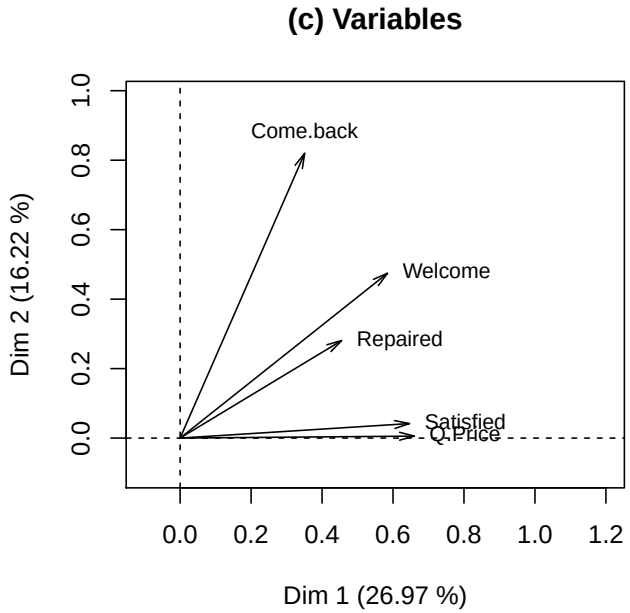
Let $\mathbf{X}$ denote the $n \times p$ categorical data matrix and $\mathbf{Z}$ denote the $n \times m$ centered disjonctive data matrix.

1. Give the values of n, p and m.
2. Give the value of the total inertia $I(\mathbf{Z})$. Which level and which variable contributes most to the total inertia ?

Second part

The multiple correspondance analysis (MCA) of the `customers` dataset gives the plots below.





1. We see that 26.96% (resp. 16.22%) of the inertia of $\mathbf{Z}$ is explained by the first principal component (res. the second principal component). What are then the two first eigenvalues of this MCA ?
2. Calculate the coordinate of the level `Come.back=Yes` on **the abscissa** of graph (b) using the previous question and the results above.

```
sel <- which(customers$Come.back=="Yes")
mca$ind$coord[sel,1, drop=FALSE]
```

```
##     dim 1
## C1   2.33
## C5   3.25
## C6   3.25
## C12  0.68
```

3. The coordinates of the levels `Come.back=dk`, `Come.back=No` and `Come.back=Yes` on the **ordinate** of graph (b) are given below. Calculate the coordinate of the variable `Come.back` on the ordinate axis of graph (c).

```
mca$levels$coord[13:15,2, drop=FALSE]
```

```
##                dim 2
## Come.back=dk    0.88
## Come.back=No   -1.08
## Come.back=Yes   0.12
```

Third part

1. Give the property used to interpret the proximity between two levels in (b) and the position of the customers in (a) according to the positions of the levels in (b). Interpret the proximity between `Repaired=No` and `Come.back=dk` in (b) and the position of the customers `C25` and `C23` in (a).
2. Give the property used to interpret the customers in (a) according to the positions of the variables in (c). Intepret the position of the customers in (a) according to the Y-axis coordinate of the variable `Come.back` in (c).