Ensimag – Grenoble INP – UGA
Analyse Statistique Multidimensionnelle
Christophe Dutang

Année 2022-2023

christophe.dutang@grenoble-inp.fr

TD: Principal Component Analysis

Exercises 1 to 3 should be addressed in priority by students in tracks ISI, MMIS and IF.

Exercise 1

- If the orthogonal projection of a variable on one principal axis (here, two principal axes) is close to 1 (respectively -1), i.e., close to the edge of circle, this variable is highly positively (respectively negatively) correlated with this axis.
- Positively correlated variables are close; negatively correlated variables are at antipodal positions; orthogonal variables are uncorrelated. Caution: the correlation between variables must be investigated in a factorial plane (or subspace) only when they are well represented (close to the edge of the correlation circle) in this factorial plane (here the first factorial plane given by the first two principal axes 1 and 2). For example, the variable `UNEMP` (Unemployment rate) is not well represented in the first factorial plane, while the variable `POP` is well represented.
- Similarity/dissimilarity between individuals (in the first plane) is given by their prox- imity/distance in the factorial plane b). This similarity/dissimilarity is mainly supported by Axis 1, then Axis 2. Caution: Let's not look for any proximity between the individuals and the variables. Only the directions are important here.
- When a variable is not well represented (close to the origin of the first factorial plane like `UNEMP` before), that means the main difference between individuals is not explained by this variable (in the first plan).
- Similarly, when an individual is not well represented (close to the origin of the factorial plane like `EGY` (Egypt) in b)), that means the main difference between the values taken by the variables is not explained by this individual (in the first plan).
- Group of variables (highly) correlated with a given principal axis can be considered in order to give an "interpretation" to this axis. This interpretation must correspond to a notion that is common to all of these variables. Here, for example, variables `AREA` (Area) and `POP` (Population) are highly (positively) correlated with the Axis 1 (`EXTTR` (External Trade) is also (negatively) correlated with this axis but not as much as `AREA` and `POP`). Furthermore, both `AREA` and `POP` can be related to the notion of "size". But we must be careful with this type of interpretation which sometimes requires a good knowledge of the concerned variables.

Exercise 2

We are interested by the dataset `cars04` with some car models in 2004. Each car is described by 11 variables listed in table~1.

The aim of this exercise is to summarize and to interpret the data `cars04` using PCA by the following call

```
> cars04.pca <- prcomp(cars04, scale=TRUE)
> summary(cars04.pca)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.6655 1.3726 0.92181 0.59751 0.52482 0.44491 0.37486
Proportion of Variance 0.6459 0.1713 0.07725 0.03246 0.02504 0.01799 0.01277
Cumulative Proportion  0.6459 0.8171 0.89439 0.92685 0.95189 0.96988 0.98266
                           PC8     PC9    PC10    PC11
Standard deviation     0.29434 0.25766 0.19229 0.02811
```

Variable	Meaning
`Retail`	Builder recommended price(US$)
`Dealer`	Seller price (US$)
`Engine`	Motor capacity (liters)
`Cylinders`	Number of cylinders in the motor
`Horsepower`	Engine power
`CityMPG`	Consumption in city (Miles or gallon; proportional to km/liter)
`HighwayMPG`	Consumption on roadway (Miles or gallon)
`Weight`	Weight (pounds)
`Wheelbase`	Distance between front and rear wheels (inches)
`Length`	Length (inches)
`Width`	Width (inches)

Table 1: Variable list for `cars04`

```
Proportion of Variance 0.00788 0.00604 0.00336 0.00007
Cumulative Proportion  0.99053 0.99657 0.99993 1.00000
```

1. Using previous R traces, what does `scale=TRUE` mean?
2. Does the representation in the first two principal components give a good idea of dataset variations?

Principal components are linear combinations of the 11 variables. The coefficients of the first 2 principal components on these 11 variables are

```
> cars04.pca$rotation[,1:2]
```

```
                  PC1          PC2
Retail     -0.2637504 -0.468508698
Dealer     -0.2623186 -0.470146585
Engine     -0.3470805  0.015347186
Cylinders  -0.3341888 -0.078032011
Horsepower -0.3186023 -0.292213476
CityMPG     0.3104817  0.003365936
HighwayMPG  0.3065886  0.010964460
Weight     -0.3363294  0.167463572
Wheelbase  -0.2662100  0.418177107
Length     -0.2567902  0.408411381
Width      -0.2960546  0.312891350
```

3. Can you give an interpretation of each of these new variables?

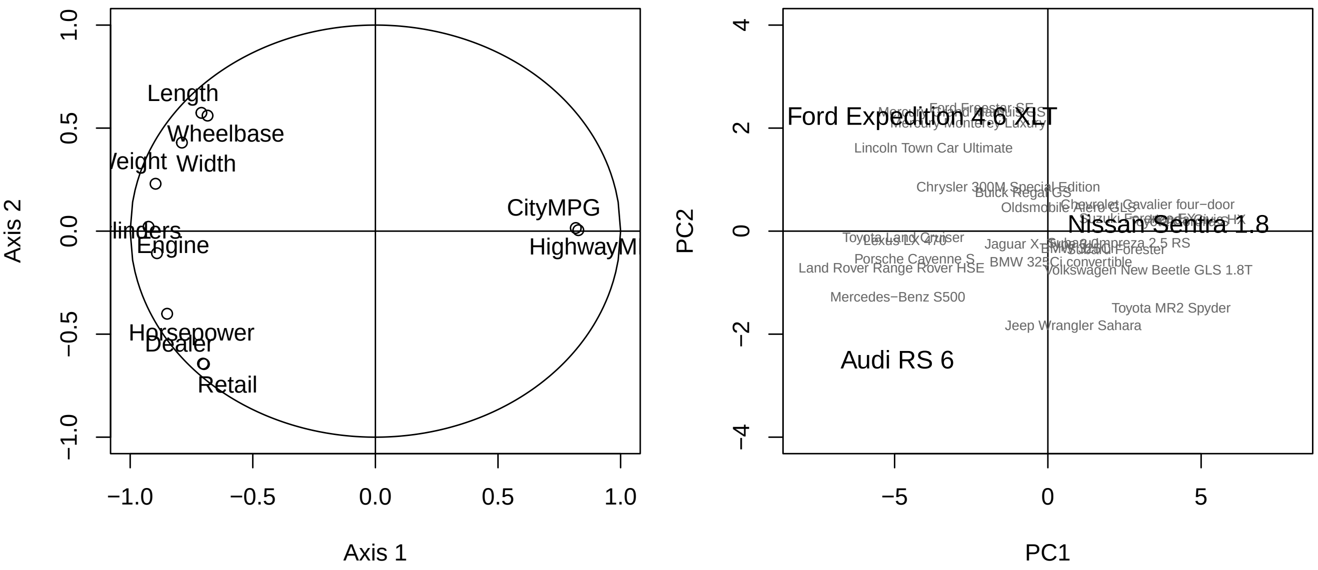
On Figure 1, the projection on the first two principal components of some cars models is plotted.

4. Interpret each quadrant of the Figure.
5. Can you describe which kind of car Audi RS 6, Ford Expedition 4.6 XLT and Nissan Sentra 1.8 are?

Exercise 3

Correspondence Analysis (CA) is an adaptation of PCA to study couples of qualitative variables. Let consider a couple of qualitative variables (X, Y) observed on n samples. The observations are denoted $((x_1, y_1), \dots, (x_n, y_n))$. Two PCAs will be performed.

- The first PCA considers the labels i of X as individuals. Each individual is described by conditional frequencies $f(Y = 1|X = i), \dots, f(Y = L|X = i)$ of values j of the variable Y given $X = i$.
- The second PCA considers the labels j of Y as individuals. Each individual is now characterized by conditional frequencies $f(X = 1|Y = j), \dots, f(X = K|Y = j)$.



(a) Variable space

(b) Individual space

Figure 1: Principal component representation in the first plane of the variable and of the sample spaces.

The interpretations of these two PCAs can be done as usual. The advantage of CA is its ability to represent both PCAs on the same graph. It allows to associate the values i of X with values j of Y using inner product between these two vectors.

- If the inner product is positive, it means that $(X = i, Y = j)$ is more frequent in the population than it would be under independence between X and Y.
- If the value is negative, it means that we would expect more couples $(X = i, Y = j)$ under independence property.

We propose to apply CA on results recorded after the first turn of presidential election in France in 2017. X represent the candidates and Y the overseas departments. Interpret the CA results.

N.B. *Candidate Lassalle was removed because he obtained quite small percentages of votes but with a very high relative variability.*

Exercise 4

1. Prove Proposition 1.

   Proposition 1. *Let $x'_i = x_i - \bar{x}_i$ (for $i = 1, \ldots, n$) be some centred sample in dimension p with covariance matrix Σ. Then the canonical inertia of these points $\dfrac{1}{n}\sum_{i=1}^{n} \|x'_i\|^2$ is tr(Σ).*

   Let Π some orthogonal projection (with the canonical dot product). Then the inertia of the projected points is tr$(\Sigma\Pi)$.

2. As a bonus, prove its Corollaries.

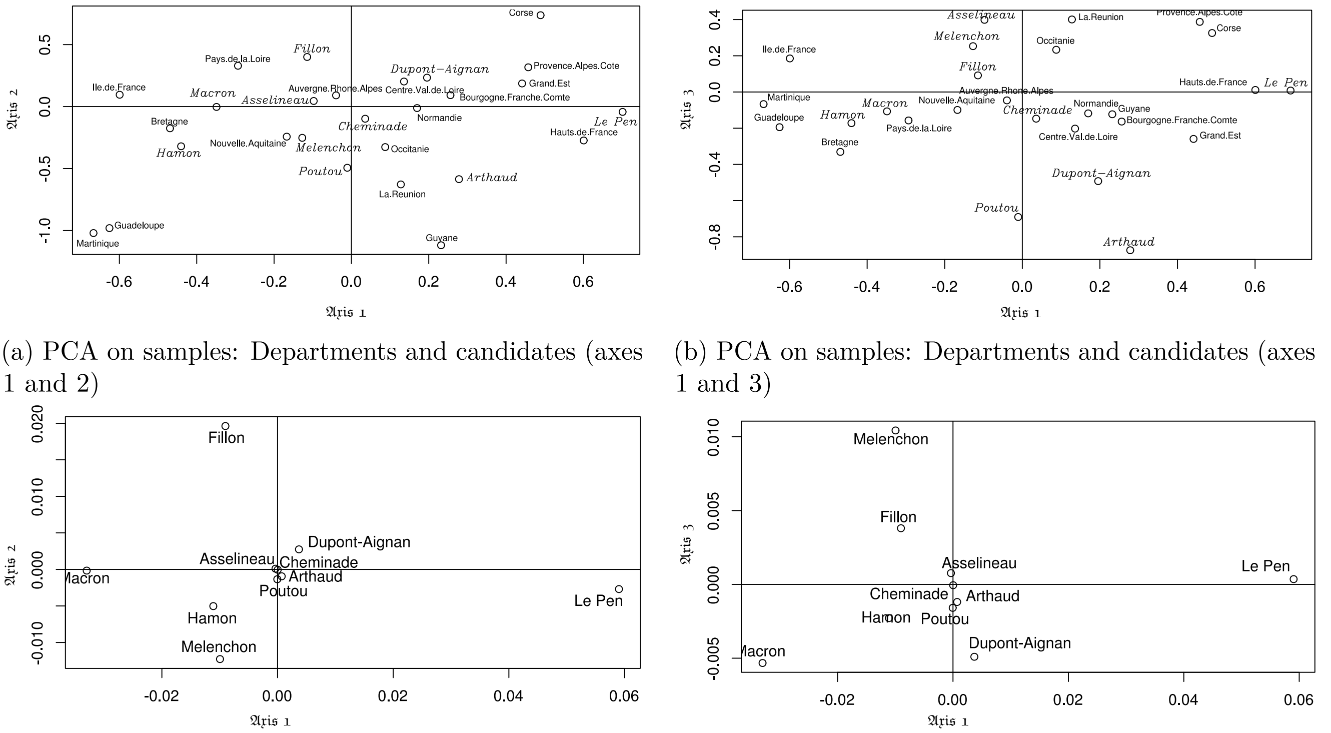
   Corollary 1. *As a consequence, for standardized samples, $\dfrac{1}{n}\sum_{i=1}^{n} \|x'_i\|^2 = p$.*

   Corollary 2. *The projected inertia on the sum of two orthogonal subspaces is the sum of the projected inertia on each subspace.*

   Hint: for any $a \in \mathbb{R}$, $a = \text{tr}(a)$.

Exercise 5

1. Prove Proposition 2.



(a) PCA on samples: Departments and candidates (axes 1 and 2)

(b) PCA on samples: Departments and candidates (axes 1 and 3)

(c) PCA on candidates considered as variables (axes 1 and 2)

(d) PCA on candidates considered as variables (axes 1 and 3)

Figure 2: Principal axes of the two PCA

Proposition 2. *We use the notations in Proposition 1.*
Let a_1 some vector with norm 1 such that $\Sigma a_1 = \lambda_1 a_1$, λ_1 being (one of) the highest eigenvalue of Σ. Then the projected inertia on the line $D_1 = Span(a_1)$ is maximal over projected inertia on all other possible lines.
Moreover, the projected inertia on D_1 is λ_1.

Hints:

- Use the results from multiple linear regression to prove that for any matrix X with linearly independent columns $X^{(1)}, \dots, X^{(p)}$, the matrix of the orthogonal projection on $\text{Span}\left(\left\{X^{(1)}, \dots, X^{(p)}\right\}\right)$ is $X(X^TX)^{-1}X^T$ (or admit this result if it does not seem obvious).
- Write the maximization problem as
$$\max_{a, \|a\|=1} \text{tr}(\Sigma a a^T).$$
- Introduce a Lagrange multiplier ξ and cancel the gradient of
$$(a, \xi) \to a^T \Sigma a - \xi(a^T a - 1).$$