Ensimag 2nd year **Statistical Analysis and Document Mining**
MSIAM 1st year

Worksheet n°4

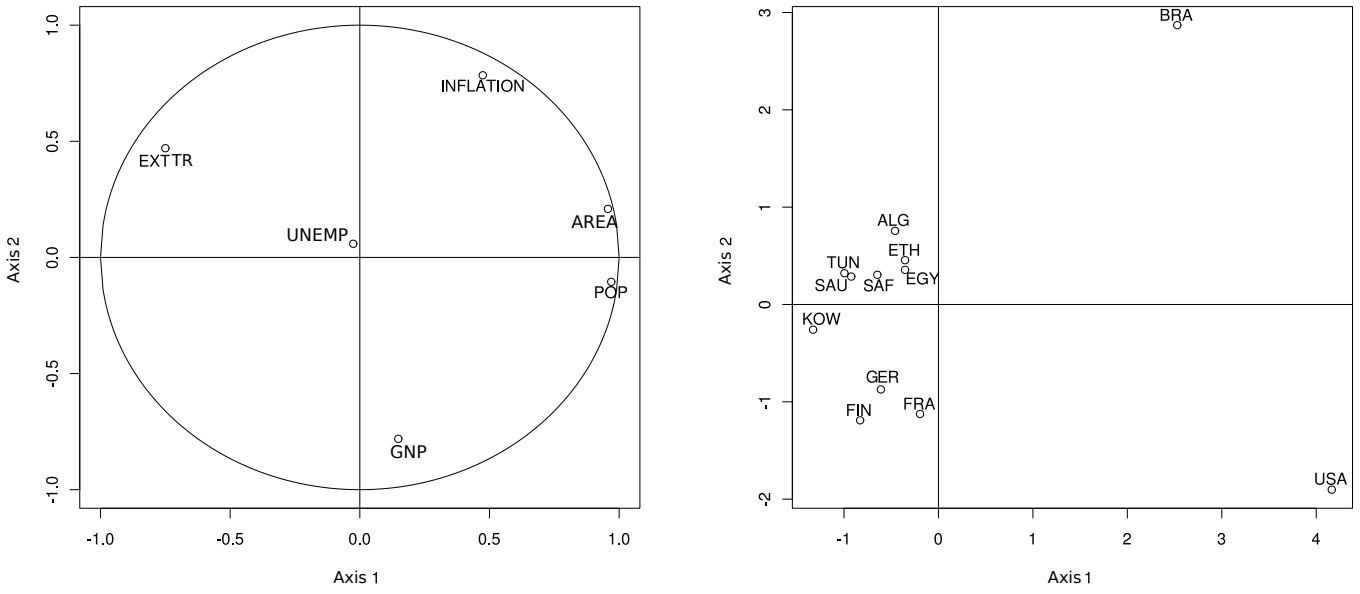
Exercise 1

Table 1 gives the measures of 6 economic indicators in 12 countries in 1991.
Interpret the results performed by a normalized PCA applied on this table. Results are summarized in figure 1. We also give the eigenvalues of the variance/covariance matrix.

i	1	2	3	4	5	6
λ_i	2.67	1.50	1.16	0.53	0.08	0.06

Country	Per capita GNP	Inflation (%)	Unemployment rate (%)	External trade (M $)	Population (M cap.)	Area (M km^2)
South Africa	2810	14,7	0,0	6,3	36,8	1,22
Algeria	1540	50,0	24,3	4,2	26,1	2,38
Germany	24130	3,5	5,1	23,5	81,0	0,35
Saudi Arabia	7328	4,4	0,0	22,1	14,8	2,15
Bresil	2400	440,8	4,8	10,5	153,0	8,51
Egypt	620	19,8	17,5	-5,9	56,1	1,00
USA	21890	4,2	6,7	-73,4	255,0	9,36
Ethiopia	110	35,7	0,0	-0,6	54,6	1,22
Finland	25800	4,1	7,7	2,2	5,0	0,33
France	21030	3,2	9,4	-10,1	57,2	0,55
Koweit	14000	3,3	0,0	20,0	1,3	0,02
Tunisia	1350	8,2	15,0	-0,9	8,6	0,16

Table 1: Economic data 1991



a) Variable space b) Samples space

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

Exercice 2

We are interested by some cars models from 2004. Each car is described by 11 variables in table 2.

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 2: Meaning of cars variabless

The aim of this exercise is to summarize and interpret the data using PCA.

Question 1

Using PCA performed by R we obtain:

```
>cars04.pca <- prcomp(cars04[,8:18], scale=TRUE)
>summary(cars04.pca)
                          PC1   PC2   PC3   PC4    PC5    PC6    PC7   PC8   PC9   PC10  PC11
Standard deviation       2.66  1.37  0.92  0.59   0.52   0.44   0.37  0.29  0.25  0.19   0.02
Proportion of Variance   0.64  0.17  0.07  0.03   0.02   0.01   0.01  0.00  0.00  0.00   0.00
Cumulative Proportion    0.64  0.81  0.89  0.92   0.95   0.96   0.98  0.99  0.99  0.99   1.00
```

a) What does the argument scale=TRUE do?

b) Does the representation in the first two principal components give a good idea of dataset variations?

Question 2

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

```
> round(cbind(cars04.pca$rotation[,1:2]%*%diag(cars04.pca$sdev[1:2]),
              cars04.pca$rotation[,1:2]),2)
             ???   ???   PC1   PC2
Retail     -0.79 -0.56 -0.28 -0.45
Dealer     -0.78 -0.56 -0.28 -0.46
Engine     -0.94  0.03 -0.34  0.02
Cylinders  -0.90 -0.18 -0.32 -0.15
Horsepower -0.89 -0.36 -0.32 -0.29
CityMPG     0.90 -0.05  0.32 -0.04
HighwayMPG  0.84 -0.02  0.30 -0.02
Weight     -0.90  0.19 -0.32  0.15
Wheelbase  -0.75  0.51 -0.27  0.42
Length     -0.74  0.50 -0.26  0.41
Width      -0.83  0.41 -0.30  0.34
```

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

Question 3

On Figure 2, the projection on the first 2 principal components of 50 cars models is plotted.

a) Interpret each quadrant of the Figure.

Can you describe which kind of car Audi RS 6, Ford Expedition 4.6 XLT and Nissan Sentra 1.8 are?

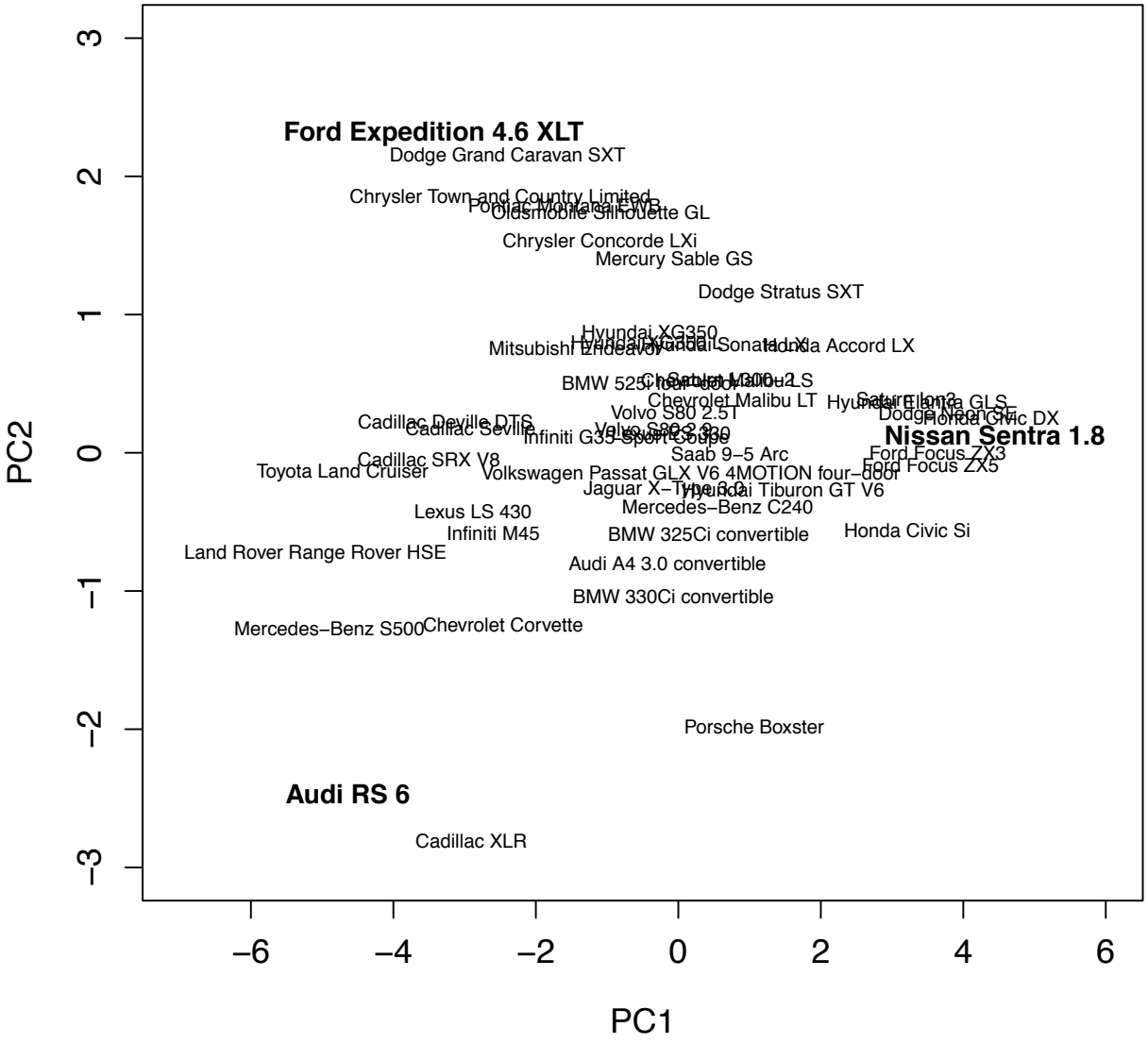


Figure 2: Projection of 50 cars models on the 2 first axes

Exercise 3

Prove Proposition 1. As a bonus, prove its Corollaries.

Proposition 1 *Let $x'_i = x_i - \bar{x}_i$ (for $i = 1, \dots, n$) be some centered sample in dimension p with covariance matrix Σ. Then the canonical inertia of these points $\frac{1}{n}\sum_{i=1}^{n} \|x'_i\|^2$ is $tr(\Sigma)$.*
Let Π some orthogonal projection (with the canonical dot product). Then the inertia of the projected points is $tr(\Sigma\Pi)$.

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

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 $a \in \mathbb{R}$, $a = \mathrm{tr}(a)$.

Exercise 4

Prove Proposition 2.

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 $\mathrm{Span}\left(\left\{X^{(1)}, \dots, X^{(p)}\right\}\right)$ is $X(X^TX)^{-1}X^T$ (or admit this result if it does not seem obvious).
- Write the maximization problem as

$$\max_{\substack{a \\ \|a\|=1}} \mathrm{tr}(\Sigma a a^T).$$

- Introduce a Lagrange multiplier ξ and cancel the gradient of

$$(a, \xi) \to a^T\Sigma a - \xi(a^Ta - 1).$$