Chapitre 2

Principal Component Analysis (PCA)

Principal Component Analysis, abbreviated PCA, is a method for analyzing **a large amount of data**. The idea of this chapter is to identify **redundancies in the data**. By merging them, we then extract new data containing more information. The procedure we will study involves **diagonalization**, which you have seen in Linear Algebra. Thus, we will have access to new variables on which we will work.

2.1 Introduction

We will be led to study samples with a large number of individuals. To make things visual, suppose these individuals are birds and that we want to characterize the entire set of positions of a flock of birds at the exact moment of a photograph.

FIGURE 2.1 – Flock of birds. Its complex shape is difficult to describe simply.

Each bird has three data points corresponding to its position in space. The shape of a flock is difficult to describe. But if we had to do it, we could first say that it is elongated. Then

that there are two lateral lobes. This idea, of describing its main features, is exactly the very principle of Principal Component Analysis.

Considering now the following classical example :

FIGURE 2.2 – Front and side photographs of a dromedary.

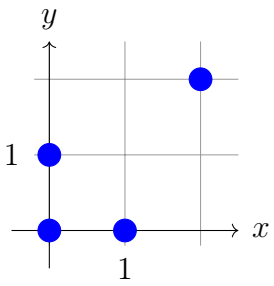
For the set of points forming a dromedary, and similarly to the flock of birds, we can realize that a side photograph, among all possible photographs, provides the most information. This is because it is the viewpoint where the set of points of the dromedary is captured with the greatest spread (the widest photograph, if you will).

A front photograph is not useless : although it contains less information, the information it provides does not appear in the side photograph.

With these two photographs, we therefore have a good idea of how to reconstruct a three-dimensional dromedary.

2.2 The Principle

Let us consider $n = 4$ individuals and $p = 2$ random variables, denoted X and Y. A realization of the sample is $(x_1, y_1) = (0, 0)$, $(x_2, y_2) = (1, 0)$, $(x_3, y_3) = (0, 1)$, $(x_4, y_4) = (2, 2)$. The point cloud is drawn alongside.



As in the first chapter, we calculate the averages, variances, covariance, and correlation coefficient : $\overline{x} = \overline{y} = \frac{3}{4}$ and $\sigma_x^2 = \sigma_y^2 = \frac{11}{16}$, $\sigma_{xy} = \frac{7}{16}$ and $\rho_{xy} = \frac{7}{11}$.

The variance-covariance and correlation matrices are written as :

$$R = \begin{pmatrix} 1 & \frac{7}{11} \\ \frac{7}{11} & 1 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \frac{11}{16} & \frac{7}{16} \\ \frac{7}{16} & \frac{11}{16} \end{pmatrix}$$

Before diagonalizing these matrices, let us recall some concepts from the Linear Algebra course.

Rappels 2.1: Linear Algebra : Diagonalization

A vector space is **a space of vectors equipped with two operations : "+" and "."**. The operation "+" is **internal**, allowing the addition of vectors. The operation "." is **external**, allowing multiplication by a scalar, i.e., by a real number, complex number, or a number from another field.

To form the algebra of square matrices, we introduce the multiplication of matrices "×". By introducing the notion of linear maps, this operation "×" can be interpreted as the **composition of the linear maps** associated with these matrices.

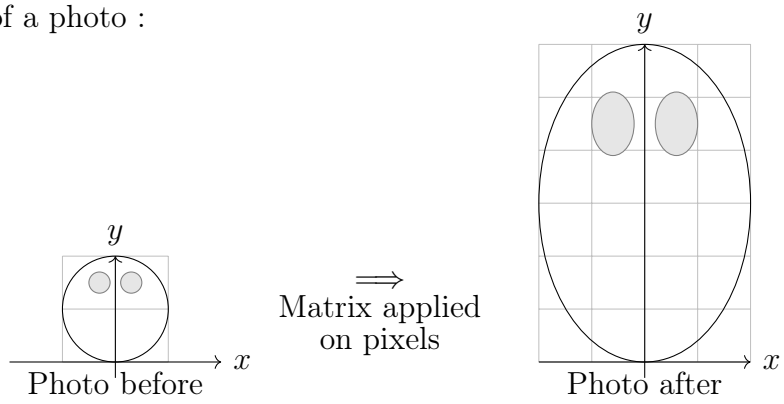
To characterize matrices, and thus the associated linear maps, we use **diagonalization**, which allows us to characterize endomorphisms of a vector space.

Simplifying the approach, one can view the diagonalization of a matrix as the expression of lines, and more generally vector subspaces, along which the associated linear map reduces to a homothety ; we then give the ratios λ. These sets are called **eigenspaces** and are such that all vectors u belonging to these sets are "stretched" by the linear map by a factor λ. For a matrix R, this corresponds to $Ru = \lambda u$ with λ fixed for each eigenspace. Obviously, this simplification is reductive because rotations are also endomorphisms of the plane.

Let us consider an endomorphism f of $\mathbb{R}^2$ and its matrix in the canonical basis $\mathcal{B} = \{e_1, e_2\}$:

$$\mathcal{M}(f) = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}.$$

The map f satisfies $f(e_1) = 2e_1$ and $f(e_2) = 3e_2$. Let us apply this linear transformation f to all pixels of a photo :



The two eigenspaces are Vect(e_1), associated with the eigenvalue 2, and Vect(e_2), associated with the eigenvalue 3. This example, of course very simple, helps to understand the notion of eigenspace.

We could also draw deformations of the plane for arbitrary axes, which may be oblique, and finally keep in mind that rotations by any angle are also linear maps.

Point méthode 2.2

To find the eigen values of a matrix R, we calculate the roots of the characteristic polynomial in λ : $det(R - \lambda I_n)$ with I_n being the identity matrix.

To find the eigen vectors associated to an eigen value λ of a matrix R, we determine a basis of $Ker(R - \lambda I_n)$, that is to solve the system in u such that $Ru = \lambda u$.

Let us return to the exercise and calculate the characteristic polynomial :

$$\left|R - \lambda I\right| = \begin{vmatrix} 1-\lambda & \frac{7}{11} \\ \frac{7}{11} & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - \left(\frac{7}{11}\right)^2 = \left(\frac{18}{11} - \lambda\right)\left(\frac{4}{11} - \lambda\right)$$

The polynomial is factored. Its roots are $\lambda_1 = \frac{18}{11}$ and $\lambda_2 = \frac{4}{11}$, each of multiplicity 1.

Now, let us calculate the eigenvector associated with λ_1. Let us set $u = \begin{pmatrix} a \\ b \end{pmatrix}$:

$$\begin{aligned} Ru = \frac{18}{11}u &\Leftrightarrow \begin{pmatrix} 1 & \frac{7}{11} \\ \frac{7}{11} & 1 \end{pmatrix} u = \frac{18}{11}u \\ &\Leftrightarrow \begin{pmatrix} 1 & \frac{7}{11} \\ \frac{7}{11} & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \frac{18}{11}\begin{pmatrix} a \\ b \end{pmatrix} \\ &\Leftrightarrow \begin{cases} a + \frac{7}{11}b = \frac{18}{11}a \\ \frac{7}{11}a + b = \frac{18}{11}b \end{cases} \\ &\Leftrightarrow \begin{cases} a - b = 0 \end{cases} \end{aligned}$$

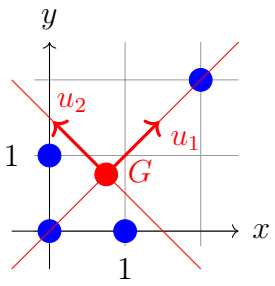
The eigenspace associated with the eigenvalue $\lambda_1 = \frac{18}{11}$ is :

$$\left\{ u \in \mathbb{R}^2 \quad | \quad u = \begin{pmatrix} a \\ a \end{pmatrix} = a\begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ with } a \in \mathbb{R} \right\} = Vect\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

In the continuation of this course, we normalize the eigenvector so that it is a unit vector, meaning its norm equals 1. It is easy to calculate the norm of $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ using the Pythagorean theorem. We take as eigenvector $u_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ for λ_1.

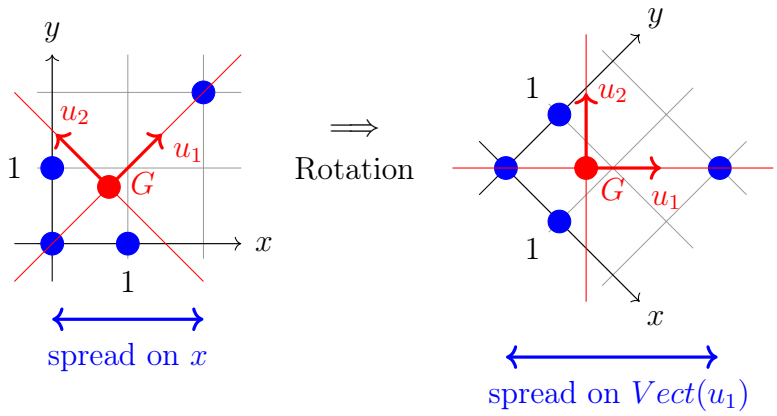
Similarly, the eigenspace associated with the eigenvalue $\lambda_2 = \frac{4}{11}$ is Vect $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$. We take the unit direction vector $u_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ for λ_2.

Let us plot them on the figure starting from the centroid G, which is located at the coordinates $(\overline{x}, \overline{y})$:



Signification :

Projected on the x-axis, our data are spread with a variance of $\sigma_x^2 = \frac{11}{16}$, and similarly for the y-axis. The diagonalization has found two axes, among which the one associated with the largest eigenvalue λ contains the **most information**. Indeed, if we project the data onto the axis generated by u_1, we see that the projections of the data are more spread out than along the x-axis. The axis generated by u_1 is thus, among all possible axes, **the axis that best separates the projections of the data** :



If we project the point cloud onto the axis generated by u_1, the projections of the farthest points are separated by a distance of $2\sqrt{2} \simeq 2.8$, whereas if we project onto the x-axis, the projections of the farthest points are separated by a distance of 2. Thus, we have found an axis that better separates the projections of the points.

We notice that individual 1 and individual 4 are strongly separated by this first new axis, and they are no longer separated on the second axis, as their projections on the second axis have the same value.

To quantify the information contained in our new axes, that is, our eigen axes, we need the following definition :

Définition 2.1 (Overall Quality of Explanation).
We call the **overall quality of explanation** for an eigen axis i the proportion of information carried by this axis. It is denoted by $oqe(i)$ and is equal to $\frac{\lambda_i}{p}$.

Proposition 2.3

When performing a PCA on the matrix R : $\sum_{i=1}^{p} \lambda_i = p$.

Démonstration. The trace of an endomorphism does not depend on the chosen basis. Thus, $Tr(R) = Tr(D)$ with $D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_p \end{pmatrix}_{p\times p}$ where $\sum_{i=1}^{p} oqe(i) = 1$. □

Returning back to our example and calculating the oqe of the two eigen axes : $oqe(1) = \frac{18}{22} \simeq 0,818 \simeq 81,8\%$ and $oqe(2) = \frac{4}{22} \simeq 0,182 \simeq 18,2\%$. Thus, the first eigen axis contains 81.8% of the information and the second eigen axis contains 18.2% of the information. We compare these values with those of the x-axis, which contains 50% of the information, as does the y-axis. We have indeed found an axis that carries more information.

Définition 2.2 (Change of basis matrix P).
Let u_i, for $i \in [\![1, p]\!]$, be the p unit eigenvectors of the matrix R (or Σ) being diagonalized. They form a basis. We concatenate them into :

$$P = mat_{\mathcal{A},\mathcal{N}} = \begin{pmatrix} \begin{pmatrix} \\ u_1 \\ \\ \end{pmatrix} \begin{pmatrix} \\ u_2 \\ \\ \end{pmatrix} \cdots \begin{pmatrix} \\ u_p \\ \\ \end{pmatrix} \end{pmatrix}_{p\times p}$$

P is the **change of basis matrix from the canonical basis $\mathcal{A}$, the old axes, to the eigenbasis $\mathcal{N}$, the new axes**.

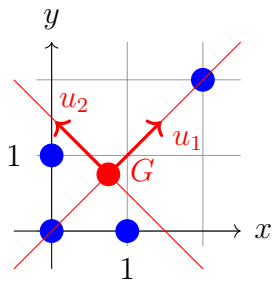
Définition 2.3 (The matrix F).
The matrix of coordinates of the n individuals in the new basis, which is the eigenbasis in PCA, is :

$$F = (f_{ij})_{i\in[\![1,n]\!],j\in[\![1,p]\!]} = M_s P$$

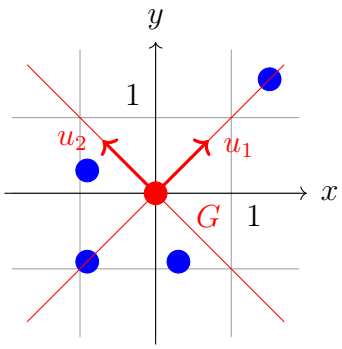
Let us return to our example and imagine that, for the 4 students, the random variable X represents the exam grade in "Multi-variable Functions" from semester 3, and the random variable Y represents the exam grade in "Probability" from semester 3. The idea of PCA is to find the "super grade" for each student that carries the most possible information. This grade is, for each student, the projection of the point representing them onto the direction of the eigenvector axis passing through the centroid G. It corresponds to the coordinate of the individual along the first principal axis.

To compute these new coordinates, we write the data matrices :

$$M = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 2 & 2 \end{pmatrix} \quad \text{et} \quad M_s = \begin{pmatrix} -\frac{3}{\sqrt{11}} & -\frac{3}{\sqrt{11}} \\ \frac{1}{\sqrt{11}} & -\frac{3}{\sqrt{11}} \\ -\frac{3}{\sqrt{11}} & \frac{1}{\sqrt{11}} \\ \frac{5}{\sqrt{11}} & \frac{5}{\sqrt{11}} \end{pmatrix}$$



Raw Data : M



Centered-Reduced Data M_s

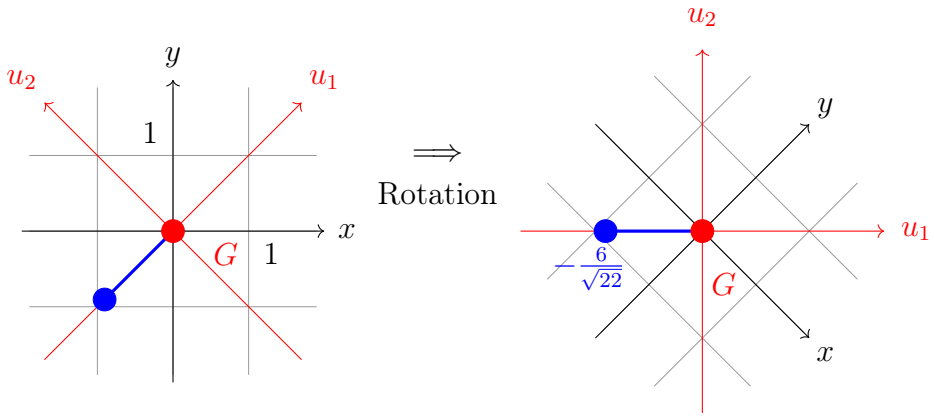
Then, we compute the coordinates of the individuals along the principal axes :

$$F = M_s P = \begin{pmatrix} -\frac{3}{\sqrt{11}} & -\frac{3}{\sqrt{11}} \\ \frac{1}{\sqrt{11}} & -\frac{3}{\sqrt{11}} \\ -\frac{3}{\sqrt{11}} & \frac{1}{\sqrt{11}} \\ \frac{5}{\sqrt{11}} & \frac{5}{\sqrt{11}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} -\frac{6}{\sqrt{22}} & 0 \\ -\frac{2}{\sqrt{22}} & -\frac{4}{\sqrt{22}} \\ -\frac{2}{\sqrt{22}} & \frac{4}{\sqrt{22}} \\ \frac{10}{\sqrt{22}} & 0 \end{pmatrix}$$

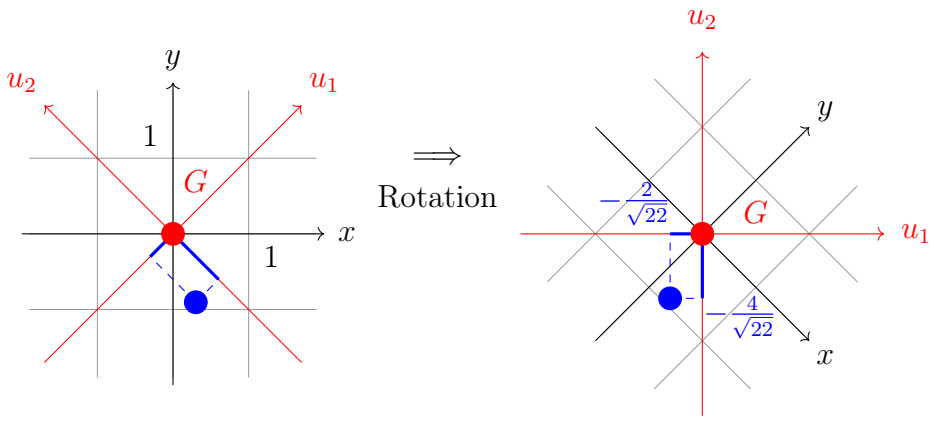
Let us now examine the matrix F of size $n \times p$: the i^{th} row corresponds to the i^{th} individual, and the j^{th} column corresponds to the j^{th} principal axis.

Let us look at the first row of F, which corresponds to individual 1 — the one who originally had the grades $(0, 0)$. This row contains their "new grades," that is, their projections onto the principal axes directed by u_1 (first column) and u_2 (second column). They obtained $-\frac{6}{\sqrt{22}} \simeq -1.28$ along axis u_1 and 0 along axis u_2, which constitute their new grades.

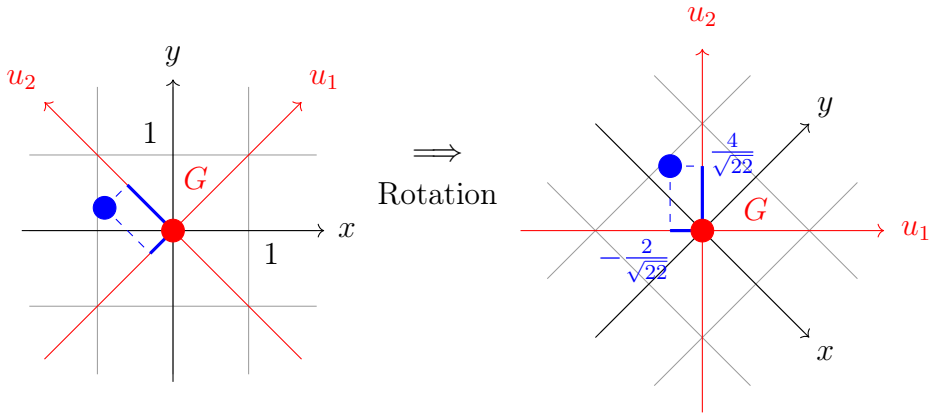
To visualize this, we plot the point corresponding to individual 1's new grades, which is $\left(-\frac{6}{\sqrt{22}}, 0\right) \simeq (-1.28, 0)$:



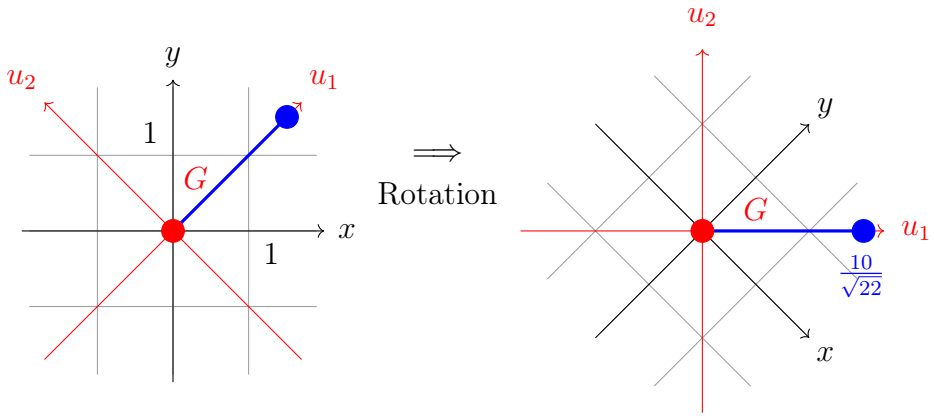
The new coordinates of individual 2 are $\left(-\frac{2}{\sqrt{22}}, -\frac{4}{\sqrt{22}}\right) \simeq (-0.43, -0,85)$:



The new coordinates of individual 3 are $\left(-\frac{2}{\sqrt{22}}, \frac{4}{\sqrt{22}}\right) \simeq (-0.43, 0, 85)$:



The new coordinates of individual 4 are $\left(\frac{10}{\sqrt{22}}, 0\right) \simeq (2.13, 0)$:



The grades of students 1 and 4 are positioned on the new axes. For example, individual 1 has their grades aligned with principal axis 1 ; we say that they are well represented by this axis. To quantify how well an individual is represented by a given axis, we define the following quantity :

Définition 2.4 (Quality of representation).
The **quality**, denoted by qlt, of an individual i with respect to a principal axis j is defined as follows :

$$qlt(\text{individual } i, \text{axis } j) = \frac{f_{i,j}^2}{\sum_{k=1}^{p} f_{i,k}^2}$$

Calculating the qlt for our four individuals :

- <u>individual 1</u> :
  $qlt(\text{individual 1, axis 1}) = \frac{\left(-\frac{6}{\sqrt{22}}\right)^2}{\left(-\frac{6}{\sqrt{22}}\right)^2 + 0^2} = 1 = 100\%$
  The grades of individual 1 lie on axis 1.
  $qlt(\text{individual 1, axis 2}) = \frac{(0)^2}{\left(-\frac{6}{\sqrt{22}}\right)^2 + 0^2} = 0 = 0\%$
  The grades of individual 1 are positioned perpendicular to axis 2.
- <u>Individual 2</u> :
  $qlt(\text{individual 2, axis 1}) = \frac{\left(-\frac{2}{\sqrt{22}}\right)^2}{\left(-\frac{2}{\sqrt{22}}\right)^2 + \left(-\frac{4}{\sqrt{22}}\right)^2} = 0.2 = 20\%$
  20% of individual 2's grades are explained by axis 1.
  $qlt(\text{individual 2, axis 2}) = \frac{\left(-\frac{4}{\sqrt{22}}\right)^2}{\left(-\frac{2}{\sqrt{22}}\right)^2 + \left(-\frac{4}{\sqrt{22}}\right)^2} = 0.8 = 80\%$
  80% of individual 2's grades are explained by axis 2.

- Individual 3 :
  $qlt(\text{individual 3, axis 1}) = \frac{\left(-\frac{2}{\sqrt{22}}\right)^2}{\left(-\frac{2}{\sqrt{22}}\right)^2+\left(\frac{4}{\sqrt{22}}\right)^2} = 0.2 = 20\%$
  20% of individual 3's grades are explained by axis 1.
  $qlt(\text{individual 3, axis 2}) = \frac{\left(\frac{4}{\sqrt{22}}\right)^2}{\left(-\frac{2}{\sqrt{22}}\right)^2+\left(\frac{4}{\sqrt{22}}\right)^2} = 0.8 = 80\%$
  80% of individual 3's grades are explained by axis 2.
- Individual 4 :
  $qlt(\text{individual 4, axis 1}) = \frac{\left(\frac{10}{\sqrt{22}}\right)^2}{\left(\frac{10}{\sqrt{22}}\right)^2+0^2} = 1 = 100\%$
  Individual 4's grades lie entirely along axis 1.
  $qlt(\text{individual 4, axis 2}) = \frac{0^2}{\left(\frac{10}{\sqrt{22}}\right)^2+0^2} = 0 = 0\%$
  Individual 4's grades are positioned perpendicular to axis 2.

REMARK : • The sum of the qlt values of an individual is indeed equal to 100%.

- The qlt represents the value of $\cos^2$ of the angle between the data point of the individual and the principal axis. Let us compute this angle for individual 2 and axis 1 :
  $\cos^2\left(\theta(\text{individual 2}, \text{axis 1})\right) = 0.2$
  So, $\theta(\text{individual 2}, \text{axis 1}) = \pm \arccos\left(\frac{1}{\sqrt{5}}\right)$

2.3 Tools of Interpretation

In order to interpret the principal axes, let us look at the correlations between the original variables (initial axes) and the new variables (principal axes) :

Définition 2.5 (The saturation matrix S).
The **saturation matrix** is defined as the matrix of correlation coefficients between the original variables (initial axes) and the new variables (principal axes).

:

$$S = (s_{ij})_{i\in[\![1,p]\!],j\in[\![1,p]\!]} = PD^{1/2} \qquad \text{avec} \quad D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_p \end{pmatrix}_{p\times p}$$

$$\overbrace{\phantom{\begin{pmatrix} s_{1,1} & \cdots & s_{1,j} & \cdots & s_{1,p} \end{pmatrix}}}^{p \text{ nnew variables= } p \text{ principal axes}}$$
$$p \text{ initial axes} \left\{ \begin{pmatrix} s_{1,1} & \cdots & s_{1,j} & \cdots & s_{1,p} \\ \vdots & & \vdots & & \vdots \\ s_{i,1} & \cdots & s_{i,j} & \cdots & s_{i,p} \\ \vdots & & \vdots & & \vdots \\ s_{p,1} & \cdots & s_{p,j} & \cdots & s_{p,p} \end{pmatrix} \right. = (s_{ij})_{ij} = S$$

The values $s_{i,j}$ being the coefficients of correlation, they range between -1 and 1.

Proposition 2.4

$$\sum_{i=1}^{p} s_{i,j}^2 = \lambda_j \text{ et } \sum_{j=1}^{p} s_{i,j}^2 = 1$$

Démonstration. Any real symmetric matrix admits an orthogonal diagonalization, so the vectors u_i, written in P, are all orthogonal. Moreover, we have normalized them. Therefore, $P^{-1} = P^t$, which means the transpose of P.

The diagonal entries of $SS^t = PD^{1/2}(PD^{1/2})^t = PD^{1/2}D^{1/2}P^t = PDP^t = R$ are equal to 1. Each of these diagonal entries of SS^t is equal to $\sum_{j=1}^{p} s_{i,j}^2$, hence the result.

The diagonal entries of $S^tS = (PD^{1/2})^tPD^{1/2} = D^{1/2}P^tPD^{1/2}$ are equal to λ_i. Each of these diagonal entries of S^tS is equal to $\sum_{i=1}^{p} s_{i,j}^2$, hence the result.

□

Let us return to the example we have been following from the beginning. The original variables are x and y, and the new variables are aligned with the principal axes u_1 and u_2. Let us compute the matrix S :

$$S = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{\frac{18}{11}} & 0 \\ 0 & \sqrt{\frac{4}{11}} \end{pmatrix} = \begin{pmatrix} \frac{3}{\sqrt{11}} & -\sqrt{\frac{2}{11}} \\ \frac{3}{\sqrt{11}} & \sqrt{\frac{2}{11}} \end{pmatrix} \simeq \begin{pmatrix} 90\% & -43\% \\ 90\% & 43\% \end{pmatrix}$$

According to this point cloud, the x-axis is correlated at 90% with the axis directed by u_1. The same goes for the y-axis. Moreover, the x-axis is correlated at -43% with the axis directed by u_1 and at 43% with the axis directed by u_2.

Thus, we indeed recover :

$\sum_{i=1}^{2} s_{i,1}^2 = \left(\frac{3}{\sqrt{11}}\right)^2 + \left(\frac{3}{\sqrt{11}}\right)^2 = \frac{18}{11} = \lambda_1$ et $\sum_{i=1}^{2} s_{i,2}^2 = \left(-\sqrt{\frac{2}{11}}\right)^2 + \left(\sqrt{\frac{2}{11}}\right)^2 = \frac{4}{11} = \lambda_2$

$\sum_{j=1}^{2} s_{1,j}^2 = \left(\frac{3}{\sqrt{11}}\right)^2 + \left(-\sqrt{\frac{2}{11}}\right)^2 = 1$ et $\sum_{j=1}^{2} s_{2,j}^2 = \left(\frac{3}{\sqrt{11}}\right)^2 + \left(\sqrt{\frac{2}{11}}\right)^2 = 1$

This example is simple and very useful, as it has allowed us to understand the general functioning of a Principal Component Analysis (PCA). However, an example with real-world data would lead us to a discussion about the **meaning of the principal axes**. That is what we will do next.

2.4 Complete Example of Analysis : Gender Equality in the European Union

In order to interpret the matrices obtained through a Principal Component Analysis, we will detail a complete example. We consider a study involving 20 countries of the European Union. For each country, we have data on different variables :

— **POP** : percentage of women in the active population.

— **ACT** : female employment rate (in %).

— **TPART** : percentage of women working part-time.

— **REMU** : gender pay gap in hourly wages (in %).

— **DOMES** : gender gap in household work (in hours per week).

— **MATER** : statutory length of maternity leave (in weeks).

— **DIPLOM** : number of women per 100 men with higher education degrees.

— **PARLEM** : percentage of women in national parliaments.

These data, from Eurostat, date from 2005 to 2010.

Here are the first rows of the table :

Data	POP	ACT	TPART	REMU	DOMES	MATER	DIPLOM	PARLEM
Germany	46	70,8	45,3	23,2	13,9	14	131,3	32
Austria	46	68,6	42,9	25,5	15,8	16	106,8	28
Belgium	45	60,8	41,5	9,1	13,6	15	142	38
Bulgaria	47	63,1	2,7	12,4	17,5	58	159,1	22
Danemark	47	77,1	37,9	17,7	10,1	18	137	37
...	...							

This partially displayed table corresponds to the **raw data**, denoted by M. The **rows** represent the countries of the European Union — these are our **individuals**. For each country, we have data corresponding to the **columns**.

These data have the following characteristics :

Data	POP	ACT	TPART	REMU	DOMES	MATER	DIPLOM	PARLEM
minimum	41	51,6	2,7	4,9	7,8	8	97,7	11
maximum	48	77,1	75,8	26,2	23,3	58	200,8	47
standard deviation	2,1	7,5	17,5	5,9	4,1	9,8	25,9	10,7
average	45,1	64,4	28,2	16,9	17,0	20,2	146,7	26,1

For example, the value in the first row corresponds to the smallest value in the POP column of the complete raw data table M.

We now center and scale the data :

	POP	ACT	TPART	REMU	DOMES	MATER	DIPLOM	PARLEM
Germany	0,43	0,85	0,98	1,07	-0,76	-0,63	-0,59	0,55
Austria	0,43	0,56	0,84	1,46	-0,29	-0,43	-1,54	0,18
Belgium	-0,05	-0,48	0,76	-1,32	-0,83	-0,53	-0,18	1,11
Bulgaria	0,90	-0,17	-1,46	-0,76	0,12	3,86	0,48	-0,38
Danemark	0,90	1,69	0,55	0,14	-1,68	-0,22	-0,37	1,02
...	...							

This large matrix M_s has zero mean in each column and unit variance in each column.

We diagonalize it : the eigenvalues sorted in decreasing order are as follows.

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
3,414	1,771	1,073	0,743	0,413	0,359	0,145	0,083

We can then compute the contributions of the axes :

axis 1 : $\frac{3,414}{8} \times 100 = 42\%$

axis 2 : $\frac{1,771}{8} \times 100 = 22\%$

axis 3 : $\frac{1,073}{8} \times 100 = 13\%$

We will therefore continue our study with these three axes, whose overall quality of explanation (oqe) is 77%.

Figure 2.4 shows the matrix R of variable correlations. The color code used is that of the Python Pandas library.

By looking at the largest values in absolute terms, we can make some initial groupings. For example, in the first column, POP, ACT, and DOMES are highly correlated (or anti-correlated).

We can associate with this group the variable TPART, which is strongly correlated with ACT, and the variable PARLEM, which is strongly correlated with DOMES.

Thus, it is likely that these five variables correspond to one principal component.

The last two variables, REMU and MATER, appear to be less or weakly correlated with the others. We can imagine that these variables will be described either by a common principal component or both together by the same principal component.

To refine our initial study, let us now examine the saturation matrix S :

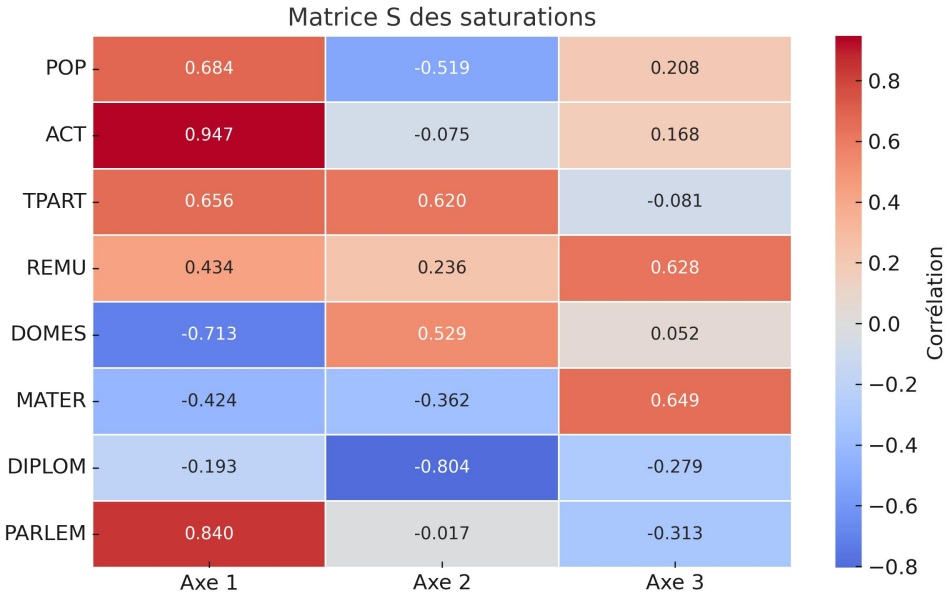


FIGURE 2.4 – Saturation matrix S, generated with Matplotlib and a heatmap using Seaborn.

This matrix allows us to identify the variables correlated with the principal components :

— **Axis 1** : it confirms that the variables POP, ACT, TPART, DOMES, and PARLEM explain this axis.

— **Axis 2** : we observe that DIPLOM mainly explains this axis.

— **Axis 3** : we observe that REMU and MATER mainly explain this axis.

This refines the initial impression we had by simply looking at the correlations between the variables.

Interpretation of the axes :

— Axis 1 appears as the axis of **parity and activity**. It ranks countries according to the role women play in public life : work and politics.

— Axis 2 appears as the axis of **university education**. It ranks countries according to the level of university education women can attain.

— Axis 3 appears as the axis of **discrimination** : It contrasts countries in terms of wage differences between men and women and differences related to maternity. The longer the legal maternity leave, the greater the potential risk of discrimination.

Définition 2.6 (Correlation Circles).
A **correlation circle** is the representation of points with coordinates $(s_{i,k_1}; s_{i,k_2})$ of the variables i, according to two principal components k_1 and k_2. These points lie inside a disk of radius 1.

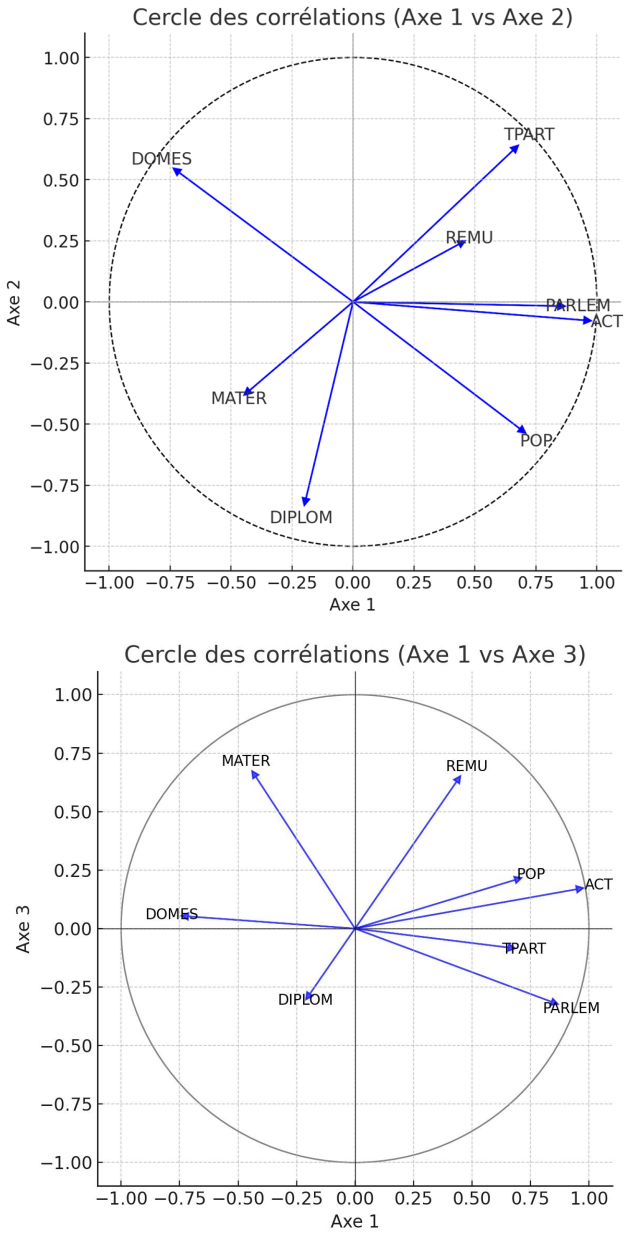


FIGURE 2.5 – Correlation circles : axis 1 vs axis 2 and axis 1 vs axis 3.
It is also sometimes interesting to plot axis 2 vs axis 3.

When studying these circles, only the variables close to the correlation circle are considered.

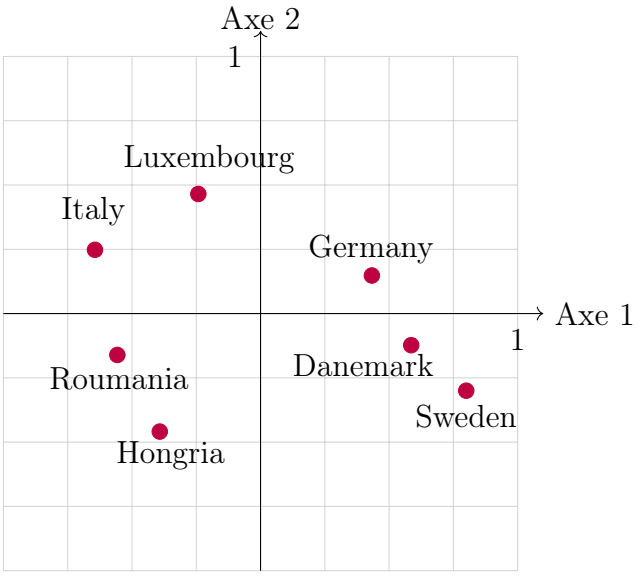
We find the previously mentioned information, namely that axis 1 is defined by the group of variables : PARLEM, ACT, TPART, DOMES, and POP. Axis 2 is clearly explained mainly by the variable DIPLOM. Finally, axis 3 is explained by the variables MATER and REMU.

To conclude this study, let us now examine the matrix F of the components of our individuals, i.e. of our countries according to the three principal axes :

	Axis 1		**Axis 2**		**Axis 3**	
Country	coord.	qlt	coord.	qlt	coord.	qlt
Germany	0,433	0,814	0,148	0,094	0,079	0,027
Austria	0,343	0,393	0,310	0,320	0,240	0,192
Belgium	0,147	0,090	0,018	0,001	-0,360	0,542
Bulgaria	-0,359	0,138	-0,545	0,317	0,493	0,260
Danemark	0,586	0,863	-0,123	0,038	0,014	0,001
Spain	-0,054	0,022	0,172	0,226	-0,142	0,154
Finland	0,521	0,528	-0,453	0,399	-0,065	0,008
France	0,118	0,076	0,077	0,032	0,163	0,146
Greece	-0,516	0,564	0,243	0,125	0,035	0,003
Hongria	-0,392	0,317	-0,459	0,435	0,029	0,002
...	...					

TABLE 2.1 – Matrix F and quality of the countries on the main axes

We select the individuals, i.e., the countries, that are the most relevant, such as those having a **high quality** and **extreme coordinates**. This allows us to identify the countries that are well represented by the principal components. Then, we plot the representative points of these countries according to the principal axes. For example, let us look at where they are located on axes 1 and 2 :



On this last plot, Germany, Denmark, and Sweden cluster together on axis 1 with positive values. These countries thus have high values in POP, ACT, TPART, PARLEM and take values well above average for DOMES (because DOMES is negatively correlated with POP, ACT, TPART, and PARLEM). Romania and Italy have very negative values on this parity and activity axis. Thus, in this category, we find Northern and Scandinavian European countries on one side, and Southern and Eastern European countries on the other.

If we continue this analysis on the other two axes : axis 2 vs axis 3, we would find that Belgium has the smallest wage gap, whereas the United Kingdom is among the highest.

Finally, one could draw a map of Europe with shades of red for countries having high values on axis 1, and shades of blue for countries having high values on axis 2.

2.5 Exercices

* = direct application of the course (should be solvable without any help) ;
** = more difficult (but should be solvable with some guidance if needed) ;
*** = challenging (bonus exercise for the most motivated students).

Exercice 8 (*).

Consider the matrix $R = \begin{pmatrix} 1 & \frac{1}{4} \\ \frac{1}{4} & 1 \end{pmatrix}$

1) Compute the matrix P as defined in the course.
2) Compute the *oqe* (proportion of explained variance) of the principal components.
3) Compute the matrix S.

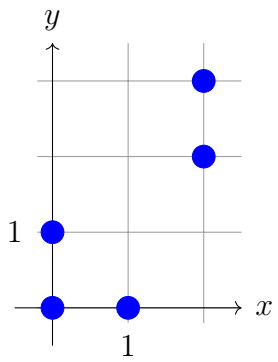
Exercice 9 (*).

Consider the matrix $F = \begin{pmatrix} -0.7 & 0,5 \\ 0.7 & -0.6 \\ -0.3 & 0.3 \\ -0.4 & 0 \end{pmatrix}$

Compute the *qlt* of the individuals identified in the rows of matrix F.

Exercice 10 (**).

Consider $n = 5$ individuals and $p = 2$ random variables X and Y. A realization of a sample is : $(x_1, y_1) = (0, 0)$, $(x_2, y_2) = (1, 0)$, $(x_3, y_3) = (0, 1)$, $(x_4, y_4) = (2, 2)$, $(x_5, y_5) = (2, 3)$:



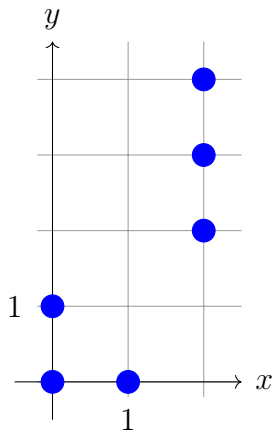
1) Compute the data matrix M, then the standardized matrix M_s.
2) Plot the standardized data according to x and y.
3) Compute the correlation matrix R, then compute the eigenvalues and eigenvectors of R. What is the matrix P ?
4) Compute the *oqe* of the principal axes.
5) Give the matrix F, then give the coordinates of individual 5 in the new axes.
6) Compute the *qlt* (quality of representation) of the individuals.

7) Give the saturation matrix S.

Exercice 11 (**).

Let $n = 6$ individuals and $p = 2$ random variables X and Y. They form the following points :

$$(x_1, y_1) = (0, 0)\,,\ (x_2, y_2) = (1, 0)\,,\ (x_3, y_3) = (0, 1)\,,\ (x_4, y_4) = (2, 2)\,,\ (x_5, y_5) = (2, 3)\,,\ (x_6, y_6) = (2, 4)$$



1) Compute the data matrix M, then the standardized matrix M_s.
2) Plot the standardized data according to x and y.
3) Compute the correlation matrix R, then compute the eigenvalues and eigenvectors of R. What is the matrix P ?
4) Compute the explained variances (qge) of the principal axes.
5) Give the matrix F, then give the coordinates of individual 5 in the new axes.
6) Compute the quality of representation, qlt, of the individuals.
7) Give the saturation matrix S.

Exercice 12 (**).

This exercise was part of the 2024 DE exam.

The goal of this problem is to perform an analysis of the following dataset. The chosen unit is the rate per one hundred thousand ; the variables represent different types of crimes and offenses in 20 U.S. states.

The initial data table is shown below. Note that a larceny is a minor theft without violence, and that an assault here means an armed assault, i.e., with violence.

STATE	Murder	Kidnapping	Theft	Assault	Rape	Larceny
ALABAMA	14.2	25.2	96.8	278.3	1135.5	1881.9
ALASKA	10.8	51.6	96.8	284.0	1331.7	3369.8
ARIZONA	9.5	34.2	138.2	312.3	2346.1	4467.4
ARKANSAS	8.8	27.6	83.2	203.4	972.6	1862.1
CALIFORNIA	11.5	49.4	287.0	358.0	2139.4	3499.8
COLORADO	6.3	42.0	170.7	292.9	1935.2	3903.2
CONNECTICUT	4.2	16.8	129.5	131.8	1346.0	2620.7
DELAWARE	6.0	24.9	157.0	194.2	1682.6	3678.4
FLORIDA	10.2	39.6	187.9	449.1	1859.9	3840.5
GEORGIA	11.7	31.1	140.5	256.5	1351.1	2170.2
HAWAII	7.2	25.5	128.0	64.1	1911.5	3920.4
IDAHO	5.5	19.4	39.6	172.5	1050.8	2599.6
ILLINOIS	9.9	21.8	211.3	209.0	1085.0	2828.5
INDIANA	7.4	26.5	123.2	153.5	1086.2	2498.7
IOWA	2.3	10.6	41.2	89.8	812.5	2685.1
KANSAS	6.6	22.0	100.7	180.5	1270.4	2739.3
KENTUCKY	10.1	19.1	81.1	123.3	872.2	1662.1
LOUISIANA	15.5	30.9	142.9	335.5	1165.5	2469.9
MAINE	2.4	13.5	38.7	170.0	1253.1	2350.7
MARYLAND	8.0	34.8	292.1	358.9	1400.0	3177.7

TABLE 2.2 – Crime rates for different types of crimes in 20 U.S. states

1) The correlation matrix R associated with this dataset is given below :

R	Murder	Kidnapping	Theft	Assault	Rape	Larceny
Murder	1.000		0.353	0.586	0.101	-0.106
Kidnapping	0.534	0.500	0.590	0.750	0.598	0.517
Theft	0.353	0.590	1.000	0.619	0.530	0.452
Assault	0.586	0.750	0.619	1.000	0.468	0.353
Rape	0.101	0.598	0.530	0.468	1.000	0.865
Larceny	-0.106	0.517	0.452	0.353	0.865	1.000

i) The above matrix is incomplete and contains several errors. Complete and correct it in red, justifying your approach.

ii) Analyze the correlation matrix.

2) The eigenvalues of the matrix R are given below :

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
3.486	1.422	0.473	0.294	0.219	0.106

We decide to retain only the first two axes for the rest of the analysis. What is the value of the *qge* ? What percentage of information is lost ?

3) A portion of the saturation matrix S is given below :

S	**Axis 1**	**Axis 2**
Murder	0.501	-0.774
Kidnapping	0.885	-0.161
Theft	0.788	-0.054
Assault	0.832	-0.343
Rape	0.802	0.491
Larceny	0.703	0.660

What is the value of $\sum_{j=1}^{6} s_{1,j}^2$? And of $\sum_{i=1}^{2} s_{i,3}^2$?

4) Draw the correlation circle in the plane defined by axes 1 and 2.

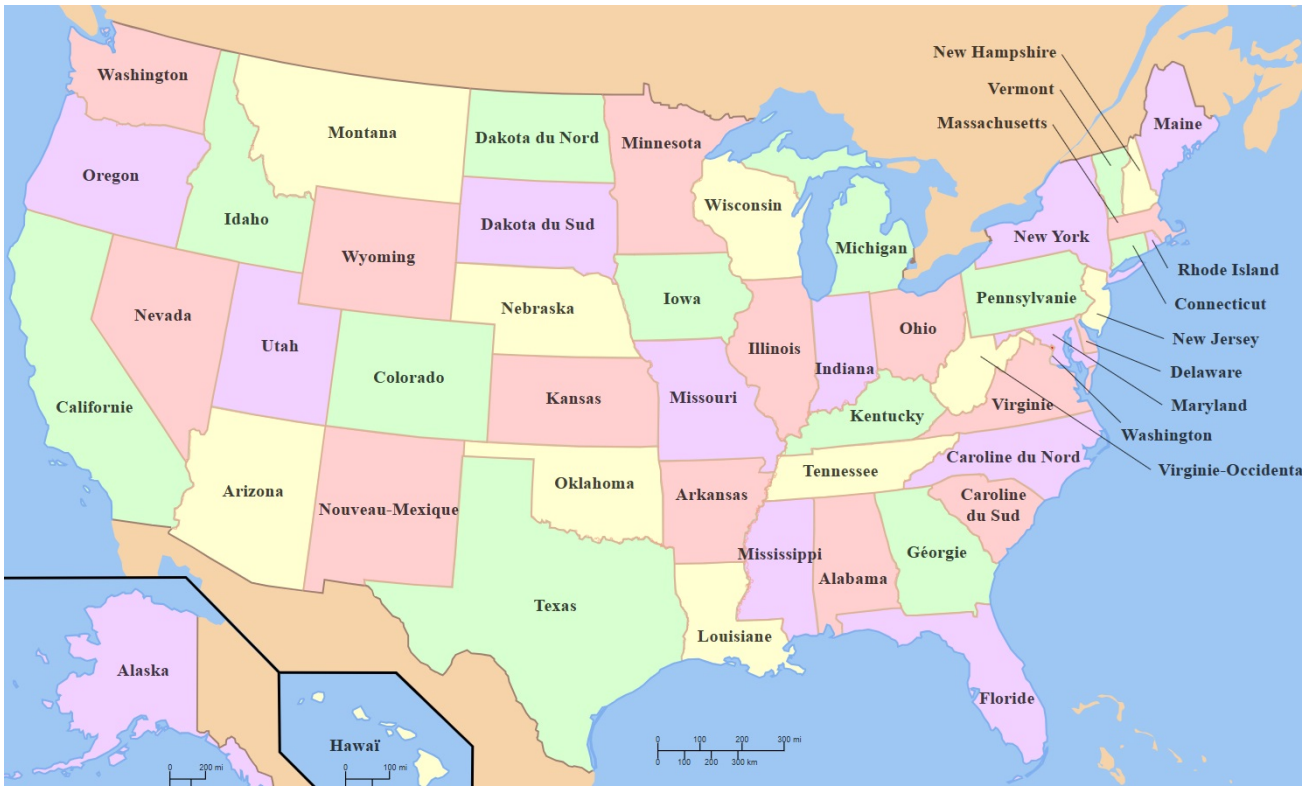
5) Interpret this graphic along with the saturation matrix, explaining your approach.

6) The table of principal components is given below :

State	**Coord. Axis 1**	**Qual. Axis 1**	**Coord. Axis 2**	**Qual. Axis 2**
ALABAMA	-0.090	0.040	-0.411	0.834
ALASKA	0.260	0.310	-0.115	0.061
ARIZONA	0.465	0.593	0.288	0.228
ARKANSAS	-0.262	0.544	-0.213	0.358
CALIFORNIA	0.702	0.931	-0.033	0.002
COLORADO	0.372	0.654	0.238	0.268
CONNECTICUT	-0.285	0.598	0.183	0.246
DELAWARE	0.056	0.039	0.268	0.881
FLORIDA	0.543	0.867	-0.006	0.000
GEORGIA	0.022	0.007	-0.250	0.865
HAWAII	0.001	0.000	0.374	0.627
IDAHO	-0.372	0.866	0.062	0.024
ILLINOIS	-0.030	0.010	-0.105	0.127
INDIANA	-0.208	0.755	-0.030	0.016
IOWA	-0.599	0.856	0.210	0.106
KANSAS	-0.202	0.904	0.064	0.091
KENTUCKY	-0.422	0.716	-0.241	0.233
LOUISIANA	0.139	0.089	-0.422	0.815
MAINE	-0.454	0.732	0.191	0.130
MARYLAND	0.364	0.501	-0.055	0.011

i) Select the most relevant states for the study of axes 1 and 2, explaining your approach.

ii) Summarize this study using the map of the United States.



Exercice 13 (***).

Download the file called "moyennesclimatiques2023.ods" from Moodle.

Perform a full Principal Component Analysis (PCA) using Python.

Interpret the principal components.