TP report:
"Principal component analysis"

BRAHMIA ABDELBACET

February 5, 2021

1 introduction

1.1 Definition of Principal Component Analysis :

It is a basic method of analyzing data tables that include individuals and their quantitative variables.

1.2 The objective of Principal Component Analysis :

- Condensing the data contained in any table by an analysis of linear relations (correlations) between n variables
- Describe graphically a table of data of individuals as well as their large quantitative variables.

1.3 practical work

1.4 Presentation of the table

We have as table (or matrix) the one entitled "IRIS". The individuals in the latter correspond to the number of flowers (150 flowers) and the variables which represent the length and width of each of the petals and sepals (4 variables) and we have as a 5th column the families to which each flower belongs.

1.5 The steps followed

1.5.1 Centralization

which aims to set the origin of the axis system to center of gravity of the point cloud.
Code Matlab :

```
N = 150
moyenne= sum(IRIS)/N
for j=1:4
    Xc(:,j)= IRIS(:,j)-moyenne(j)
end
```

1.5.2 Reduction

to give the same importance to each of the variables so that the type of measurement units does not influence the analysis (with the use of reduced centered data):
Code Matlab :

```
for j=1:4
    jj(j)= [sqrt(sum(Xc(:,j).^2)/N)];
end
Y=[ (Xc(:,1)/jj(1)) (Xc(:,2)/jj(2))
    (Xc(:,3)/jj(3)) (Xc(:,4)/jj(4))]
```

1.5.3 Standardization

In this step we currency The matrix of the centered data reduced on Sqrt of N
Code Matlab :

```
Z=Y/sqrt(N)
```

1.5.4 Result of the PCA

Once the best projection has been determined, the results are generally represented by two types of graphs: The circle of correlations of the variables and the factorial plane of the individuals. The information of interest to individuals is mainly the distance between the points.

1.5.5 Factorial plan

Covariance variance matrix it is very useful for determining the factorial plane and the linear correlation
Code Matlab :

```
R=(Z'*Z)
```

The vectors and the eigenvalues (the axes and the main components) which will allow us to draw the cone diagram (Scree of the eigenvalues)
Code Matlab :

```
[vecteur,valeur]=eigs(R);
Vp=sort((eig(R).'),'descend')
```

In our example, we will result in this diagram:

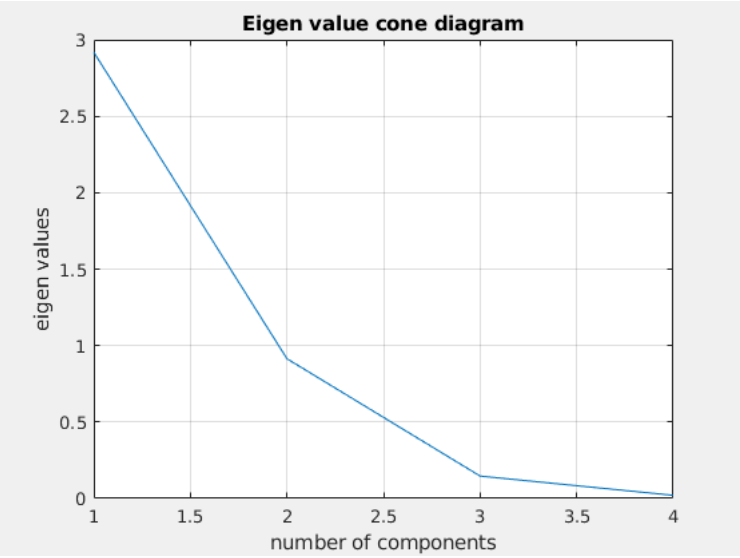

Figure 1: number of components

This representation makes it possible to determine the proportion of information contained in a plan.

Obtaining the plan This plane is the graphic representation of individuals in the new representation space, it is the product of the matrix normalized by the eigenvector.
Code Matlab :

```
Coord=Z*vecteur
```

The factorial plane of this lab is represented as follows:

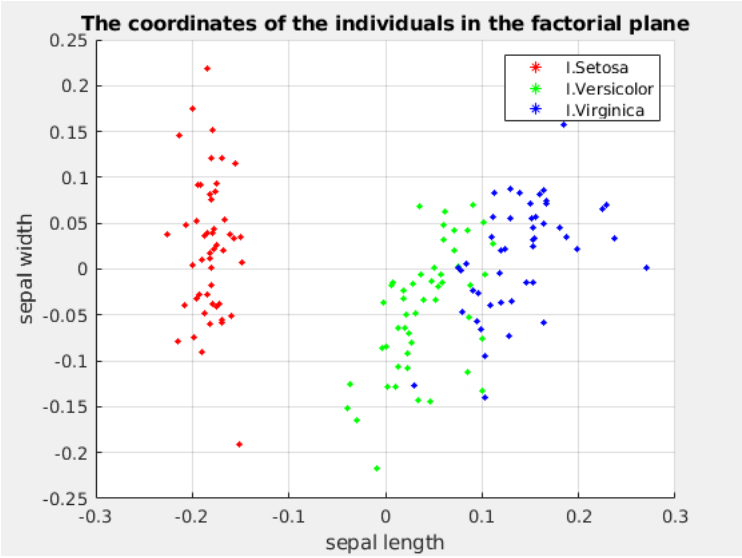

Figure 2: The coordinates of the individuals in the factorial plane

- Flowers that are near the center are poorly represented by the factorial plane
- Those that are close to one of the axes are well correlated with this axis and are the explanatory points for this axis.
- Those that are close to each other most resemble (are similar)

1.6 Correlation circle

The circle of correlations is the projection of the cloud of variables on the plane of the principal components.

Correlation coefficients Linear correlation coefficients are calculated between the old variables and new variables (factors).
Code Matlab :

```
for i=1:4
    for j=1:4
        L(j,i)=(sqrt(valeur(i,i)))*vecteur(j,i);
    end
end
```

We have as result, this circle of correlation:

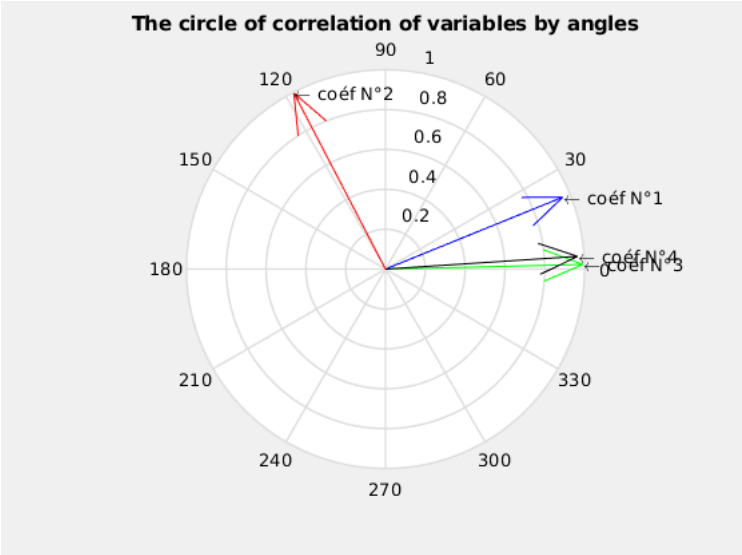

Figure 3: The circle of correlation of variables by angles

- The variables which are close to the circle (coef N ° 1 and coef N ° 2) are well represented.
- Variables which have a small angle between themselves are correlated.
- The points which are close together, the cosine of their angles tends towards 1 and therefore the 2 variables are positively correlated.
- The points which are opposite, the cosine of their angles tends towards -1 and therefore these variables are negatively correlated.