Rapport Projet Calcul Scientifique et Analyse de Données
TP-Projet 1 : Se familiariser avec l´ Analyse en Composantes Principales (ACP)

Noms des auteurs :

EL BOUZEKRAOUI YOUNES
TOUBALI HADAOUI FAICAL

Département Sciences du Numérique - Première année
2019-2020

Table des matières

Table des figures

1 Introduction

L'objectif de cette première partie du projet est de se familiariser avec l'utilisation de de l'ACP Analyse en Composante Principales sur des données ayant des dimenssions quelconques.

2 Partie 1 : Visualisation des données

2.1 Question 1

Les données sur lequelles on a appliqué l'ACP lors du TP1 sont les trois couleurs rouge vert et bleu, RVB de chaque pixel composant l'image.
La matrice X correspond au tableau des données elle contient des données représentés par trois nombre : le niveau de rouge, de vert et de bleu de chaque pixel de l'image.
La dimension des données est : 206 x 345 x 3.

2.2 Question 2

Après avoir complété le script matlab visualisation.m, on obtient les figures suivantes :

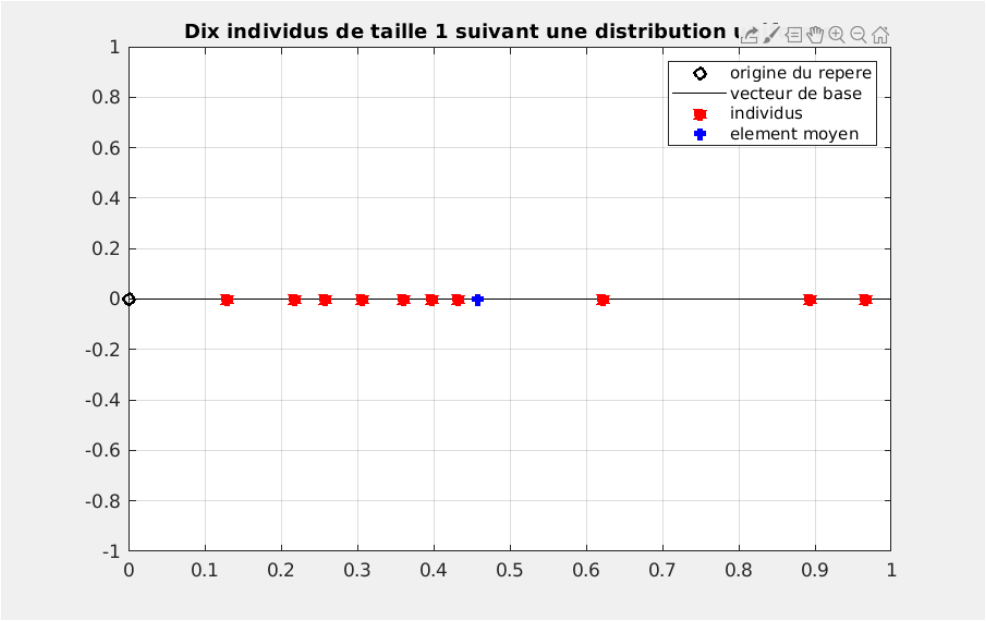

FIGURE 1 – Génération d'un tableau de 10 individus de taille 1 suivant une distribution uniforme

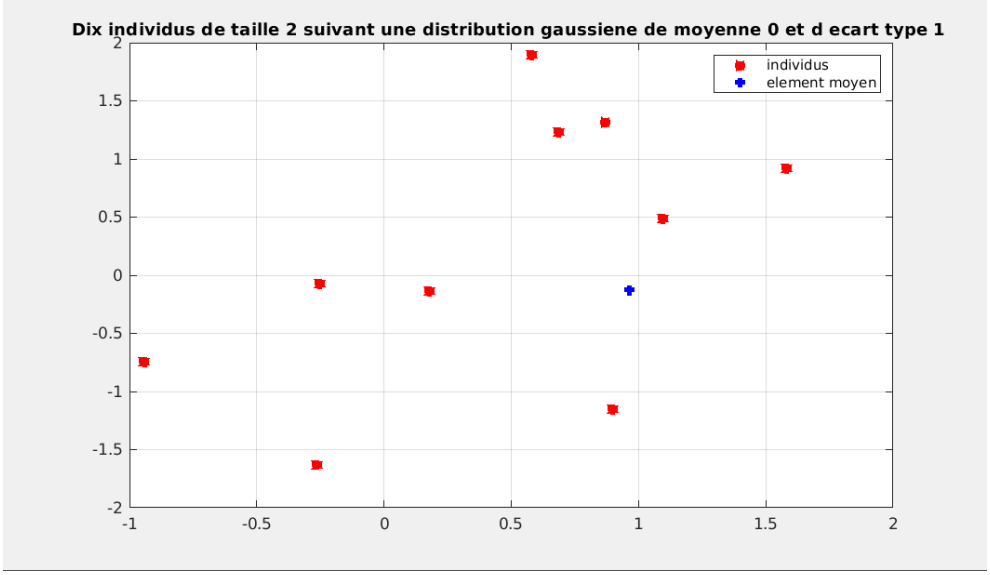

FIGURE 2 – Génération d'un tableau de 10 individus de taille 2 suivant une distribution gaussiene de moyenne 0 et d'ecart type 1

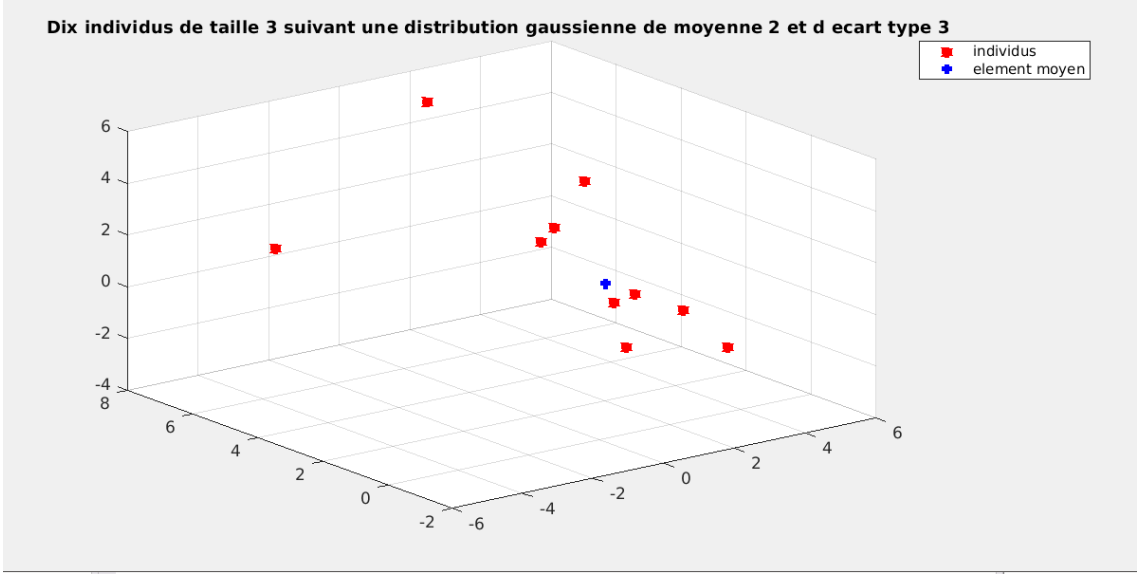

FIGURE 3 – Génération d'un tableau de 10 individus de taille 3 suivant une distribution gaussiene de moyenne 0 et d'ecart type 1

3 Partie 2 : L'Analyse en Composantes Principales

On dispose d'un tableau des données X qui appartient à R^{nxp} et qui est constitué de n individus représentés par p variables. On calcule la matrice de Variance/Covariance du tableau des données X ainsi que les vecteurs propres qui représentent les axes principaux, On réordonne ces axes par ordres décroissant du contrast qu'ils fournissent, puis on calcule la matrice C de l'echantillon X dans ce nouveau repère.

3.1 Question 3

En projettant les données sur les deux premiers axes de la base connonique puis sur les deux premiers axes principaux, on remarque que dans la base de l'ACP les données sont visiblement mieux distibués et plus facile à lire et à analyser. Ceci le montre les figures suivantes :

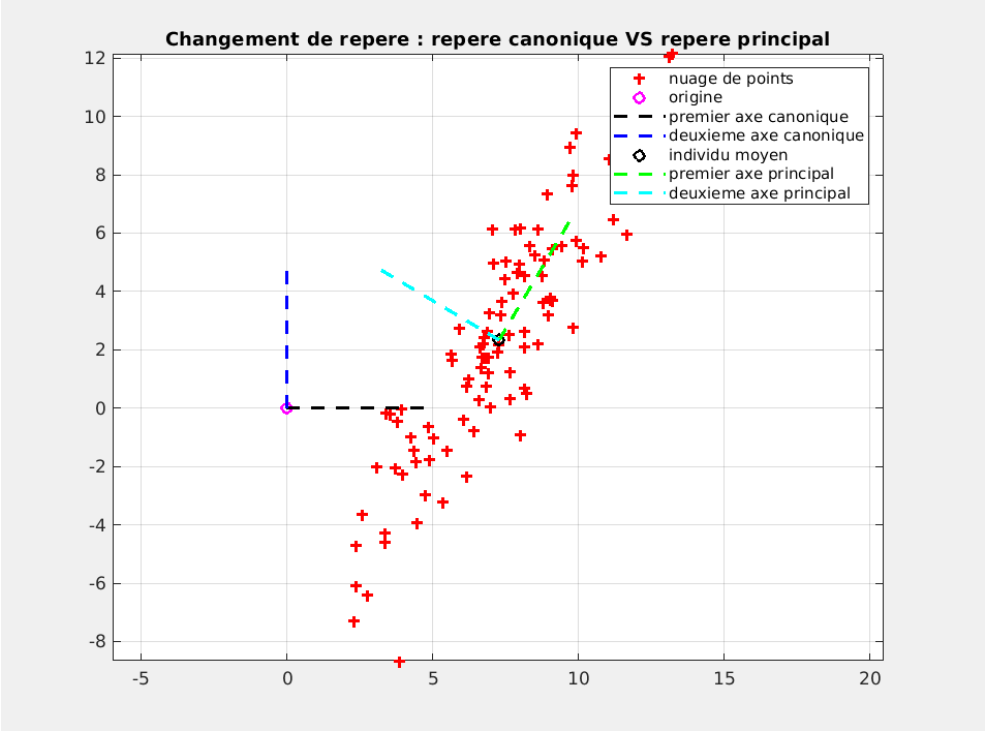

FIGURE 4 – La différence entre l'affichage sur le repère canonique et l'affichage sur le repère principal

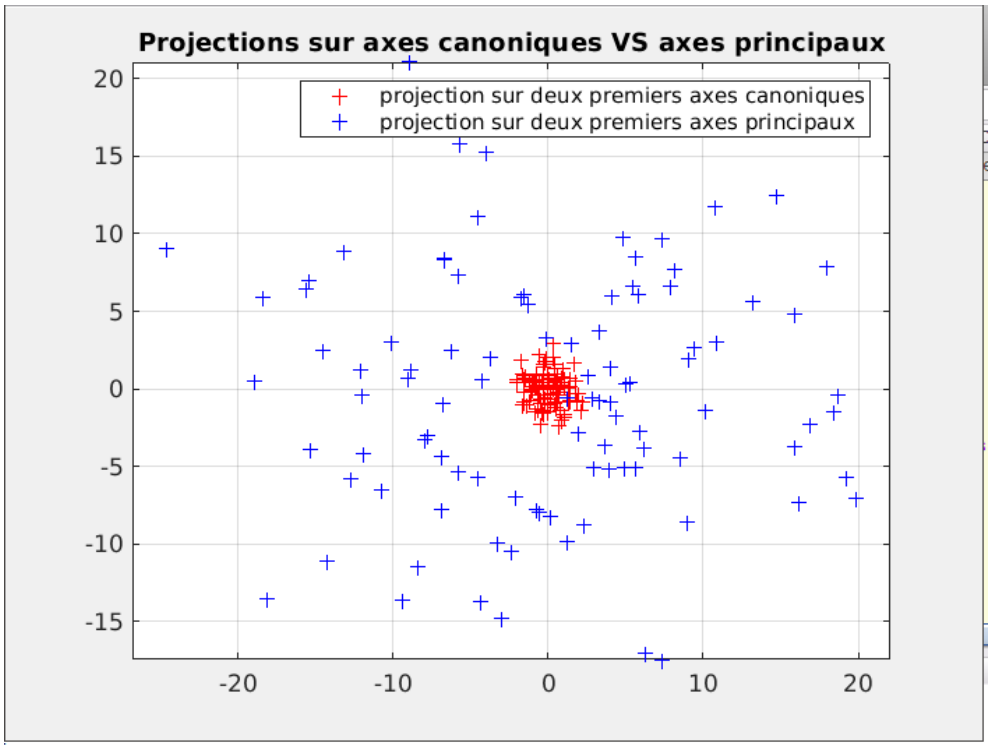

FIGURE 5 – nuage de points projetés sur les deux premiers axes canoniques et les deux premiers axes principaux

3.2 Question 4

Chaque valeur de la diagonnale de la matrice Sigma est proportionnel au pourcentage d'information que porte l'axe correspondant. Alors à partir de cette matrice Sigma on peut quantifier l'information contenue dans les q premières composantes principales. C'est l'information maximale qu'on peut extraire à partie de q axes.

4 Partie 3 : L'ACP et la classification de données

Dans cette partie on utilisera la methode de l'ACP pour classifier des données.

4.1 Question 5

Première classification (2 classes) :

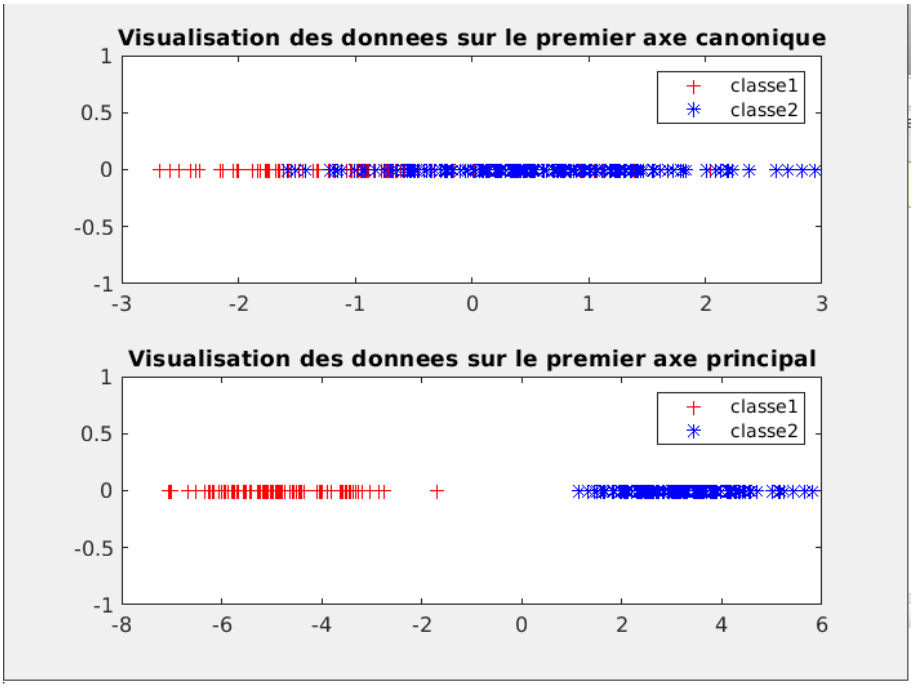

FIGURE 6 – Projection du nuage de points sur l'axe canonique et sur l'axe principal

Deuxième classification (4 classes) : On projette les données sur le premier axe principale puis sur le deuxième puis sur le troisième. A partir de la lecteur de la projection sur chaque axe tout seul on ne peut distinguer que 2 classes. Alors à partir seulement des 3 projections précédantes on peut faire une conjecture qu'ils existent uniquement 2 classes.

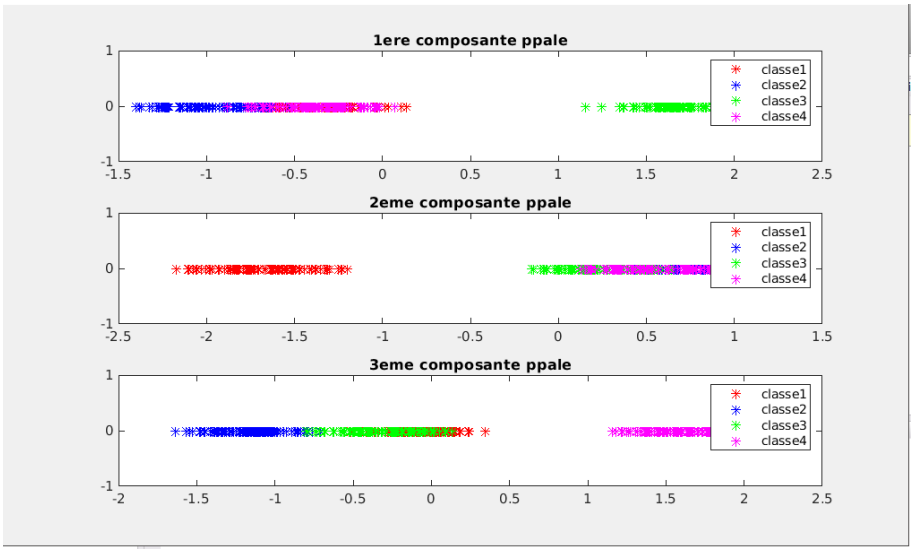

FIGURE 7 – Projection du nuage de points sur trois axes principaux

En projettant les données sur plan on peut identifier 3 classes, alors que dans l'espace on peut identifier 4 classes.

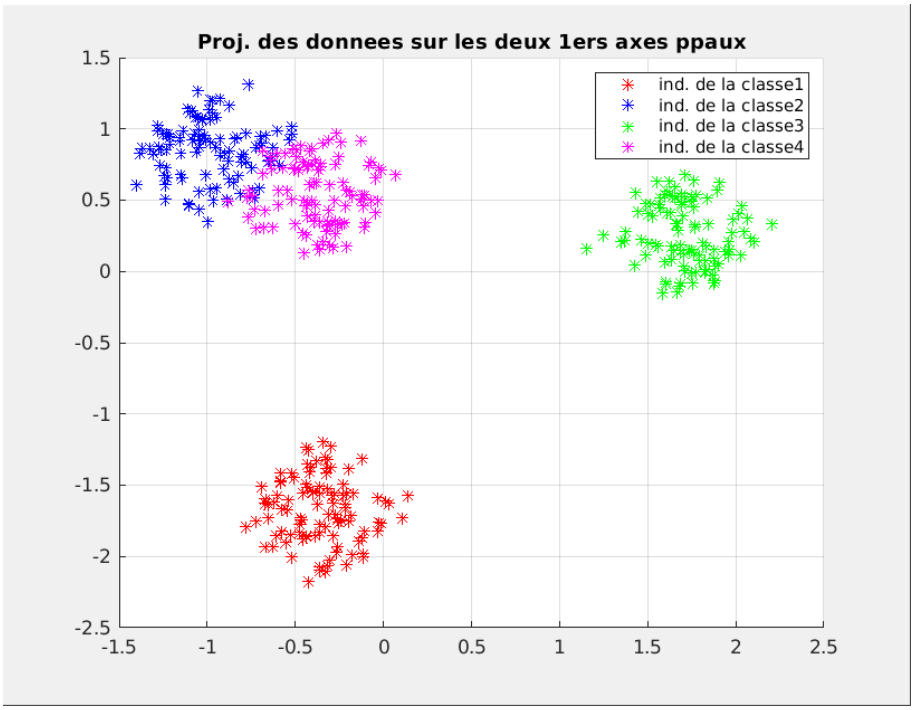

FIGURE 8 – Projection du nuage de points sur le plan constitué par les premiers deux axes principaux

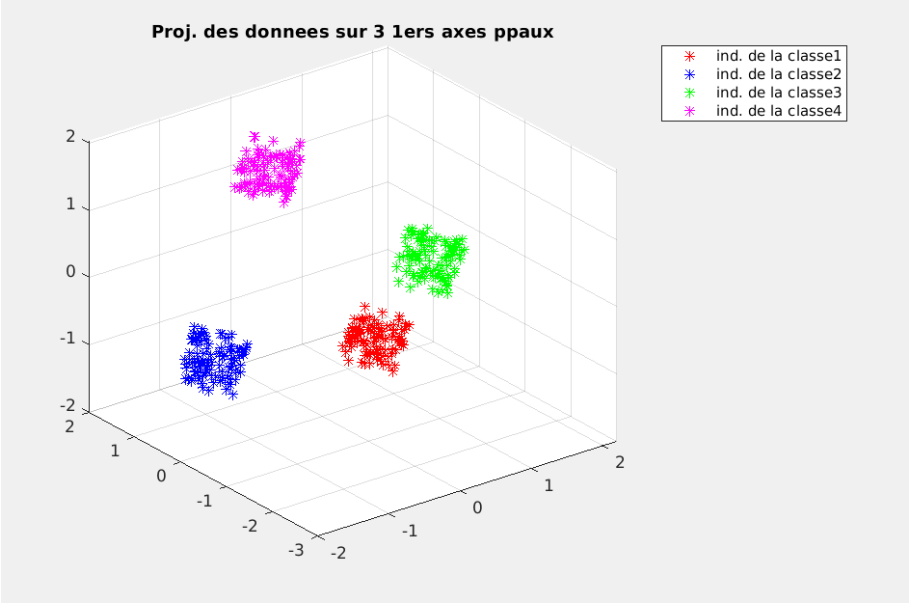

FIGURE 9 – Projection du nuage de points sur l'espace constitué par les trois premiers axes principaux

En comparant les figures du pourcentage d'informations pour la premiere classification en deux

groupe et cette dernière de 4 groupe on remarque qu'un seul axe suffisait pour la première car il comportait le maximum d'informations alors que pour la dernièreon remarque que 3 axes axes portent des pourcentages non negligeables d'information ce qui justifie le fait qu'on à besoin de 3 trois dimensions pour bien visualiser la classification des données.

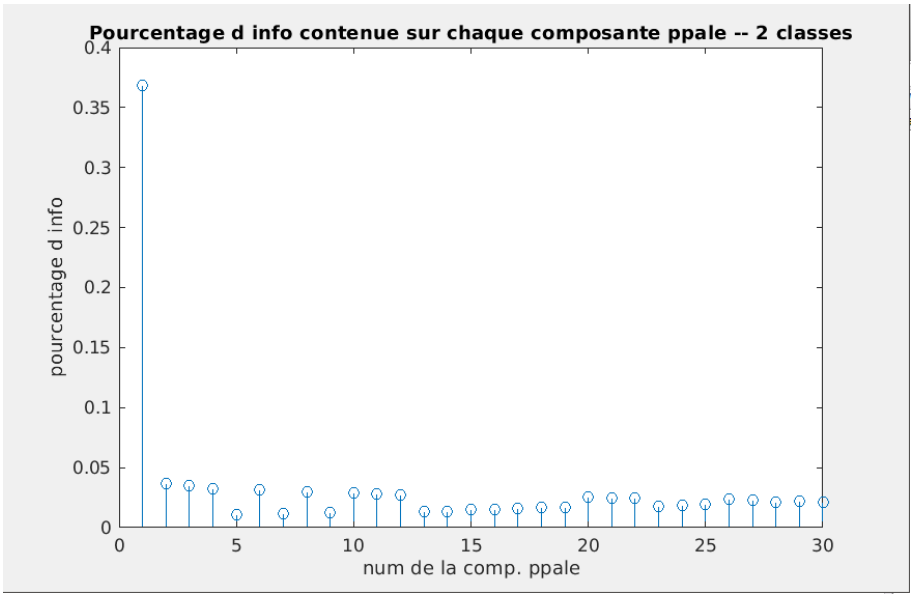

FIGURE 10 – Pourcentage d'information pour la première classification

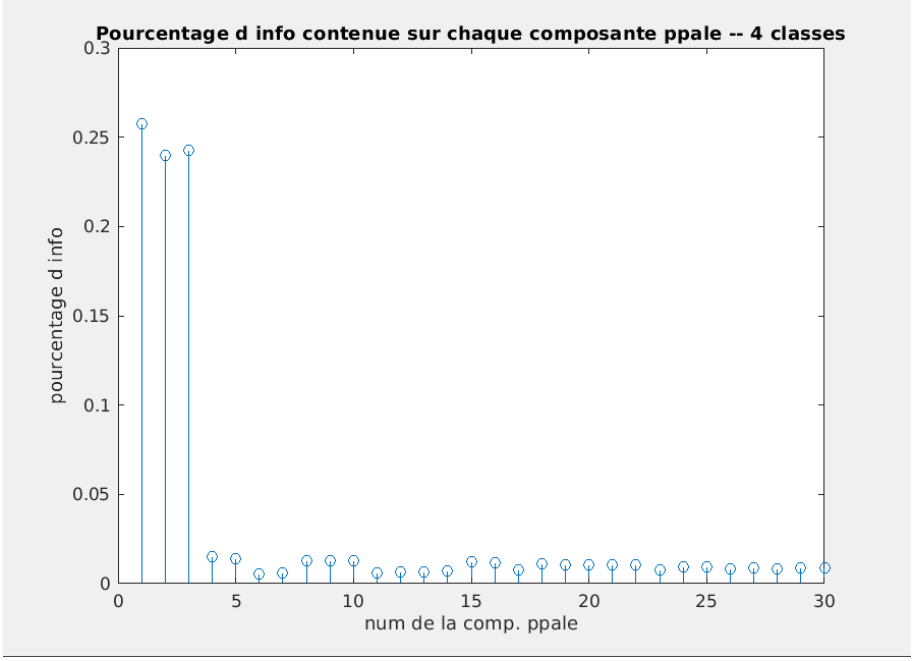

FIGURE 11 – Pourcentage d'information pour la deuxième classification

4.2 Question 6

Après avoir calculer les pourcentage d'information portés par chaque axe on a deduit qu'on aura besoin de au moins deux dimensions pour bien visualiser les données. En les representant, on constate l'existence de 6 classes que ce soit sur 2 ou 3 dimensions.

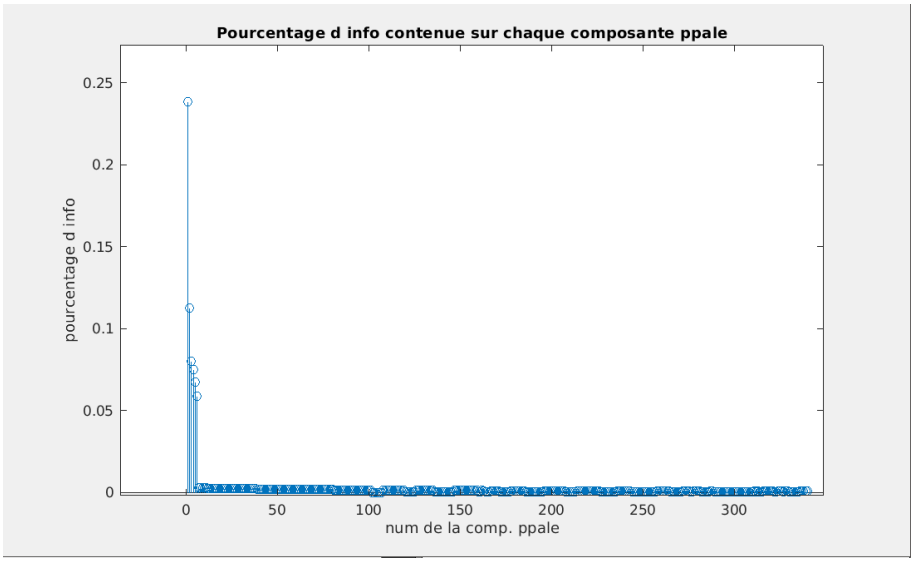

FIGURE 12 – Pourcentage d'information pour la classification

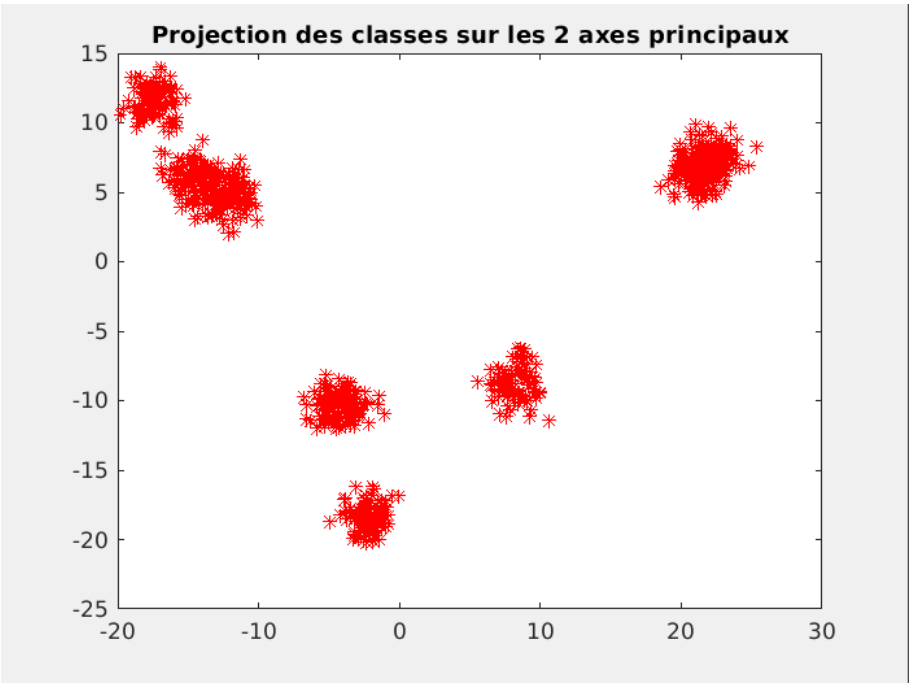

FIGURE 13 – Projection du nuage de points sur le plan constitué par les premiers axes principaux

4.3 Question 7

La même démarche de la question 6 applique à la transopsé du tableau des données permet de classer les données selon les variables. On trouve donc 6 classes d'individus.

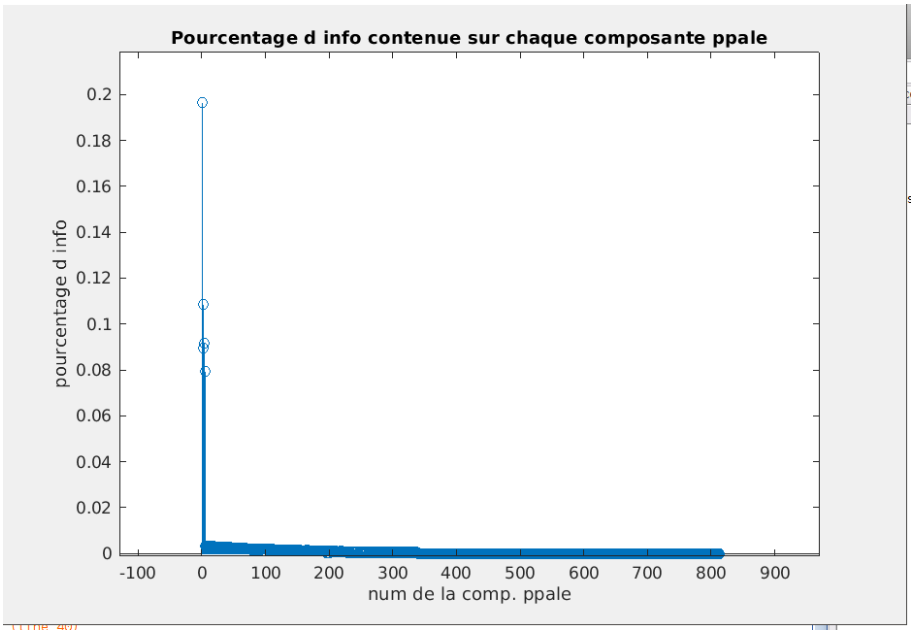

FIGURE 14 – Pourcentage d'information pour la classification

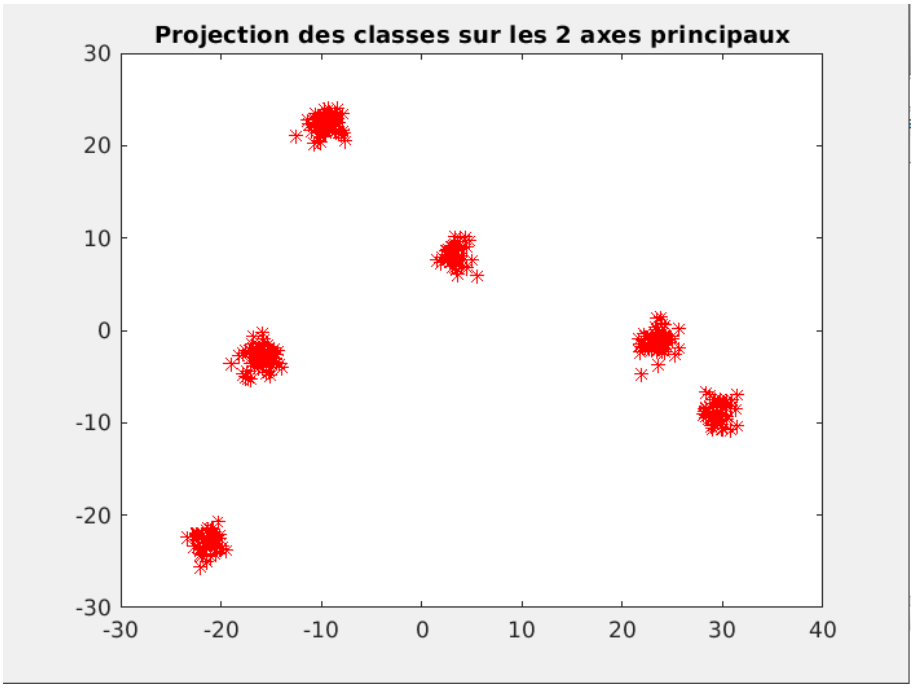

FIGURE 15 – Projection du nuage de points sur le plan constitué par les premiers axes principaux

5 Partie 4 : L'ACP et la méthode de la puissance itérée

5.1 Question 8

On considère Ψ une valeur propre de la matrice H^tH associé au vecteur propre x, on a alors :

$H^tH.x = \Psi.x$

En multipliant à droite par la matrice t^H des deux cotés, on obtient :

$HH^tH.x = \Psi.Hx$

On déduit alors que Ψ est une valeur propre de la matrice HH^t associé à la valeur propre $H.x$. On a alors l'équivalence mathématique suivante :

Ψ une valeur propre de la matrice H^tH associé au vecteur propre x si et seulement si Ψ une valeur propre de la matrice HH^t associé au vecteur propre $H.x$

Il suffit alors de connaitre les éléments propre de la matrice H^tH pour connaitre ceux de la matrice HH^t et vice versa.

5.2 Question 9

Voir Script puissance_itérée.m

5.3 Question 10

En théorie on a besoin plus de précision, et surtout un taux d'erreur qui tend vers 0, la methode de puissance itérées donne le resultat avec un taux erreur non nul qui dépend de épsilon et le nombre d'itérations maximal. Par contre la méthode eig donne le résultat directement avec un taux d'ereur plus faible,ce qui va nous pérmetre d'avoir une meuilleure estimation des valeurs et vecteurs propres et donc des axes principaux correctes et parsuite un taux d'erreur réduit en général dans la méthode de l'ACP.

5.4 Question 11

En éxecutant le script à chaque fois, on constate que l'erreur relative et le temps prit par l'algorithme pour trouver l'élèment propre dominant sont minimals pour la matrice H^tH.

```
Erreur relative pour la methode avec la grande matrice = 9.967e-09
Erreur relative pour la methode avec la petite matrice = 9.989e-09
Ecart relatif entre les deux valeurs propres trouvees = 1.52e-04
Temps pour une ite avec la grande matrice = 4.290e-03
Temps pour une ite avec la petite matrice = 1.132e-04
fx >>
```

FIGURE 16 – l'erreur relative et le temps prit par l'algorithme pour calculer l'élèment propre de chaque matrice