# TP-Projet 1 : Se familiariser avec l'Analyse en Composantes Principales (ACP)

Nom des auteurs
CAZES Noa
JAMES Christopher
MARTIN Cédric

Département Sciences du Numérique - Première année
2019-2020

# Table des matières



# Table des figures

# 1 Introduction

L'objectif de cette première partie de projet était d'utiliser l'Analyse en Composantes Principales afin de réduire les dimensions d'un problème dans le but de pouvoir visualiser des données (et notamment des corrélations entre données) dans le plan ou l'espace.

# 2 Visualisation des données

Ainsi on ne peut visualiser des données qui ne sont caractérisées que par 3, 2 ou 1 variable(s). Cependant la plupart des données s'exprime avec plus de 3 variables.

**Question 1** Retour aux données du TP1 d'Analyse de données

Les données sur lesquelles nous avons appliqué l'ACP dans le TP1 étaient des images.
Le tableau de données X correspondait alors aux trois canaux de couleur R, V et B de l'image.
Les dimensions correspondaient au nombre de pixels de l'image pour les lignes, avec trois colonnes, une pour chaque canal.

**Question 2**

Les graphiques issus de la visualisation des données qui s'expriment avec 1, 2 ou 3 variable(s) sont donnés ci-dessous :
La figure 1 présente un nombre sur la droite des réels.

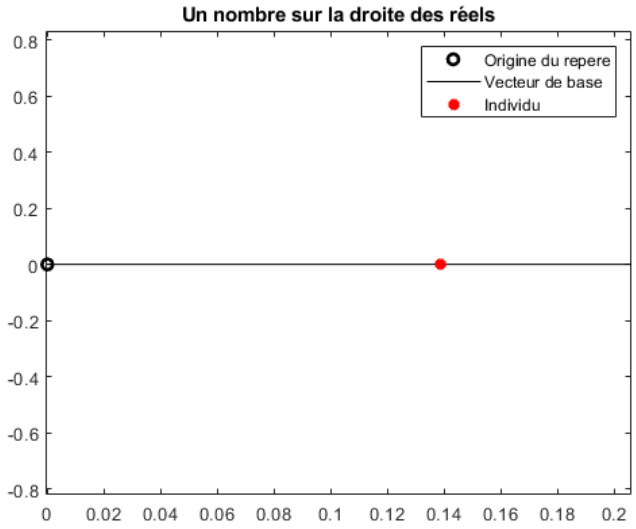


FIGURE 1 – Visualisation d'un nombre sur la droite des réels

La figure 2 présente un élément de $\mathbb{R}^2$ dans le plan.

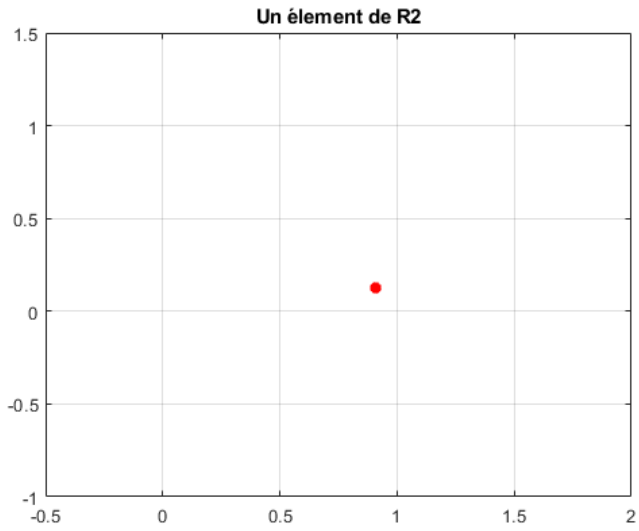


FIGURE 2 – Visualisation d'un élément de $\mathbb{R}^2$ dans le plan

La figure 3 présente un élément de $\mathbb{R}^3$ dans l'espace.

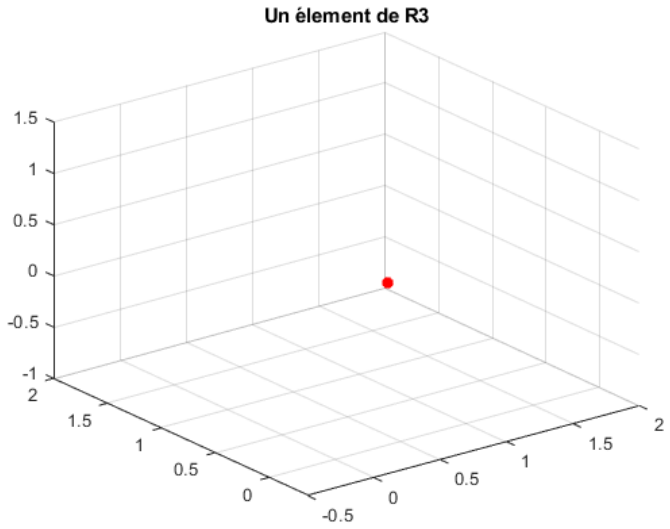


FIGURE 3 – Visualisation d'un élément de $\mathbb{R}^3$ dans l'espace

La figure 4 présente des nombres sur la droite des réels.

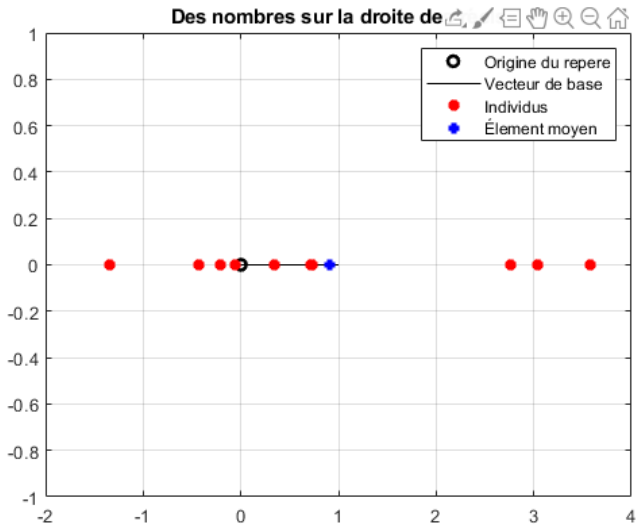


FIGURE 4 – Visualisation de nombres sur la droite des réels

La figure 5 présente des nombres dans le plan des réels.

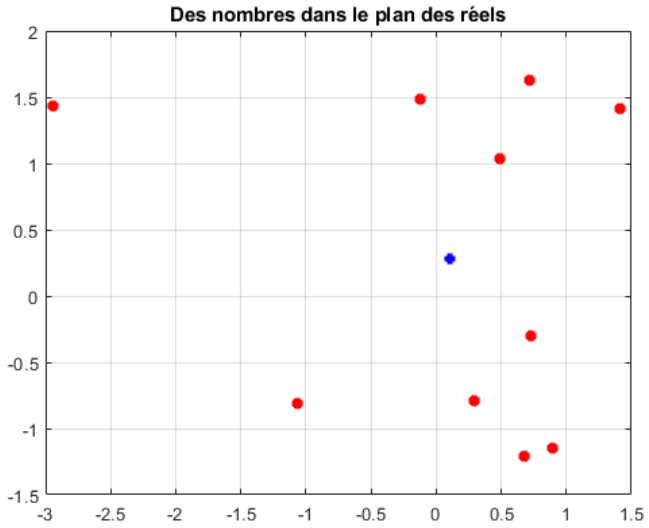


FIGURE 5 – Visualisation de nombres dans le plan des réels

La figure 6 présente des nombres dans l'espace des réels.

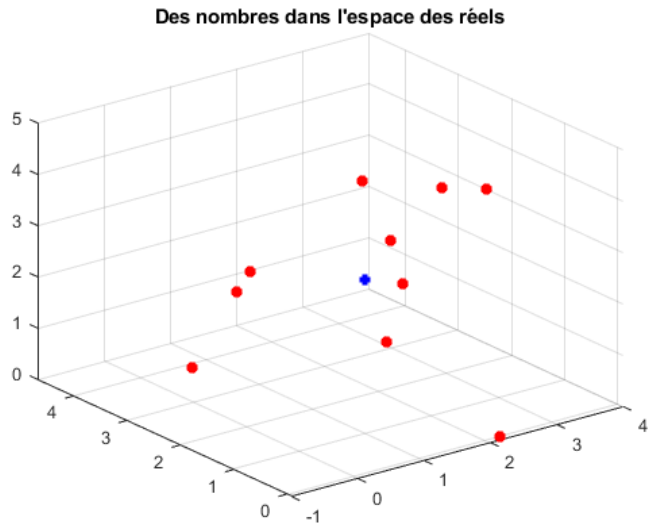


FIGURE 6 – Visualisation de nombres dans l'espace des réels

# 3 Analyse en Composantes Principales

Dans cette partie, on met en pratique l'ACP afin de diminuer la dimension des données pour visualiser au mieux ces dernières.

Elle se réalise en plusieurs étapes :

1. Représentation des données sous une forme matricielle $X$ appartenant à $\mathbb{R}^{n\times p}$
2. Calcul des données de l'individu moyen $\bar{x}$
3. Centrage des données autour de cet individu moyen
4. Détermination de la matrice de variance/covariance $\Sigma$
5. Diagonalisation de $\Sigma$ avec classement des éléments diagonaux par ordre croissant
6. Calcul de $C = X^c U$, la matrice $X^c$ dans la base orthonormée de vecteurs propres de $\Sigma$
   La colonne n°$j$ correspond à la projection des des individus sur le $j^{eme}$ axe principal.
   Les axes sont décorrélés (intérêt de la diagonalisation).
7. Sélection des axes principaux en fonction de la proportion de contraste associée

**Question 3**

On obtient les graphiques suivants après modifications du script $ACP.m$ :

La figure 7 confronte la représentation des données dans le repère canonique à celle dans le repère principal.

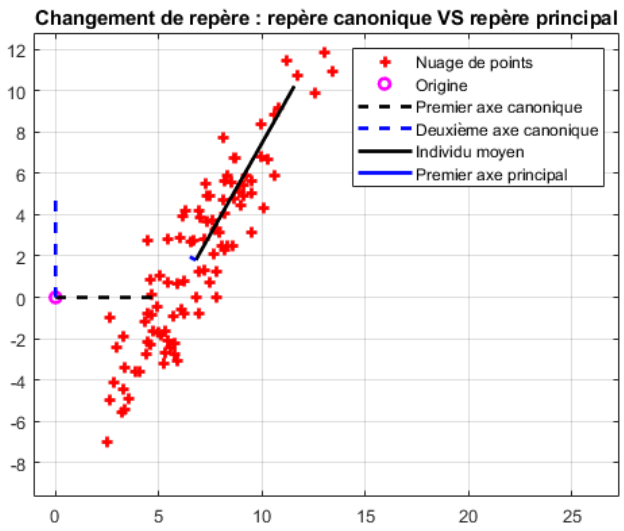


FIGURE 7 – Changement de repère : repère canonique VS repère principal

La figure 8 montre la projection des individus sur les axes principaux et aussi sur les axes canoniques.

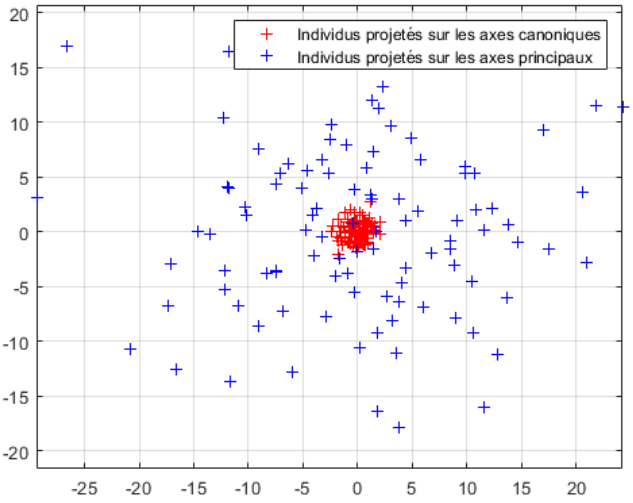


FIGURE 8 – Projection des individus sur les axes principaux et sur les axes canoniques

On remarque que le nuage de points projeté sur les axes principaux est plus dispersé - donc plus "riche en information" - que le nuage de points projeté sur les deux premiers axes canoniques.

**Question 4**

Une quantification de l'information contenue dans la 1ere composante principale est la plus grande valeur propre de $\Sigma$.

Une quantification de l'information contenue dans les 2 premières composantes principales est la somme des deux plus grandes valeurs propres de $\Sigma$.

...

Une quantification de l'information contenue dans les q premières composantes principales est la somme des q plus grandes valeurs propres de $\Sigma$.

# 4 L'ACP et la classification de données

On peut, par la suite, partitionner les données en clusters, en fonction d'une mesure de distance ou de proximité définie au préalable.

Après avoir complété le script matlab *classification.m* on obtient les graphiques suivants :

La figure 9 représente la visualisation des données, comportant deux classes, projetées sur le premier axe canonique et sur le premier axe principal.

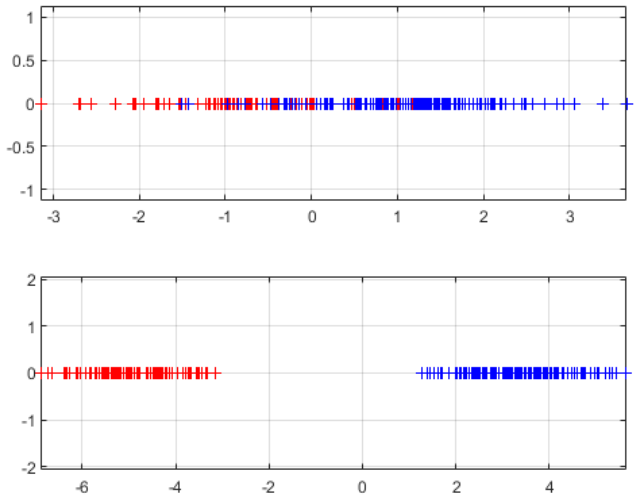


FIGURE 9 – Projection des données sur le premier axe canonique (figure du haut) et sur le le premier axe principal (figure du bas)

On remarque alors que le changement de repère est nécessaire pour pouvoir distinguer les deux classes.

La figure 10 montre le pourcentage d'information apporté par chaque composante principale.

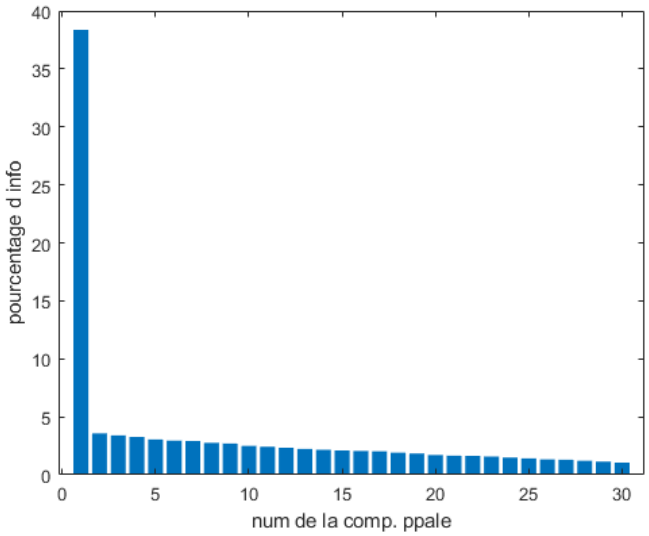


FIGURE 10 – Pourcentage d'information apporté par chaque composante principale.

C'est ainsi que cela confirme le fait qu'avec deux classes, la représentation des données suivant la première composante principale est suffisante, car, dans cette projection, est conservée une grande partie de l'information initiale.

La figure 11 représente la visualisation d'un nuage de points, comportant quatre classes, en projection sur le premier, puis le deuxième, puis le troisième axe principal.

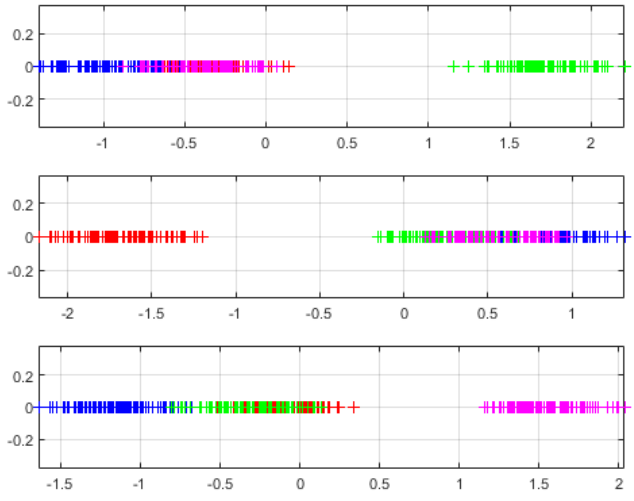

FIGURE 11 – Nuage de points, comportant quatre classes, en projection sur le premier (figure du haut), puis le deuxième (figure du milieu), puis le troisième (figure du bas) axe principal.

La figure 12 représente la visualisation d'un nuage de points, comportant quatre classes, dans le repère défini par les deux premiers axes principaux.

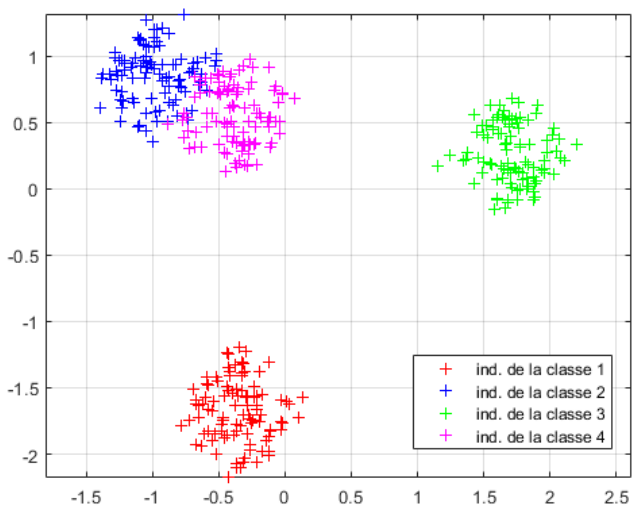


FIGURE 12 – Nuage de points, comportant quatre classes, dans le repère défini par les deux premiers axes principaux.

Ici, on pourrait se méprendre et observer seulement 3 classes, ainsi deux axes principaux ne sont pas suffisant pour conserver le maximum de l'information quond on dispose de 4 classes. On observe alors ce nuage sur les trois premiers axes principaux.

La figure 13 représente la visualisation d'un nuage de points, comportant quatre classes, dans le plan défini par les trois premiers axes principaux.

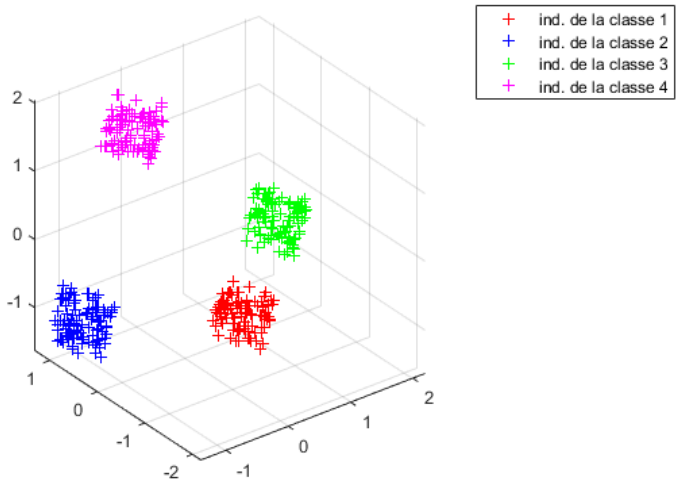


FIGURE 13 – Nuage de points, comportant quatre classes, dans le plan défini par les trois premiers axes principaux.

On distingue alors bien 4 classes en utilisant l'information contenue dans les trois premiers axes principaux.

La figure 14 montre le pourcentage d'information apporté par chaque composante principale (cas avec 4 classes).

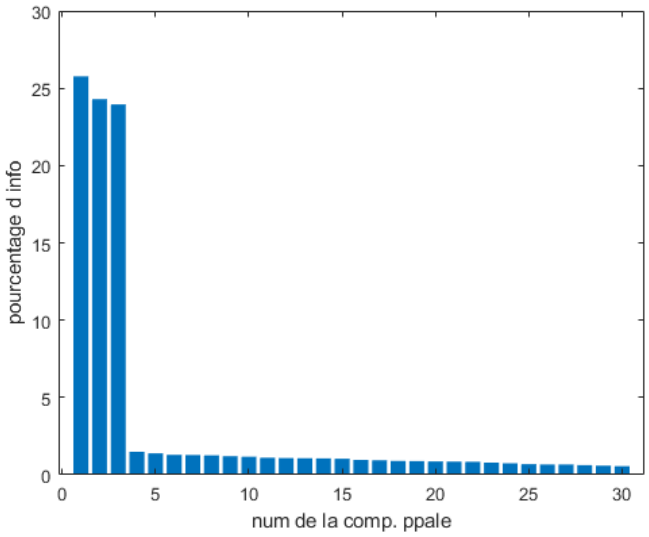


FIGURE 14 – Pourcentage d'information apporté par chaque composante principale (cas avec 4 classes).

Cela confirme le fait que, comme la majorité de l'information est contenue dans les trois premières composantes principales, celles-ci sont nécessaires à une visualisation exacte des données, contrairement au cas où on avait seulment deux classes.

Ainsi on pourrait conjecturer que pour un nombre $x$ de classes, on ait $x - 1$ composantes principales significatives.

**Question 6**

La figure 15 montre le pourcentage d'information apporté par chaque composante principale.

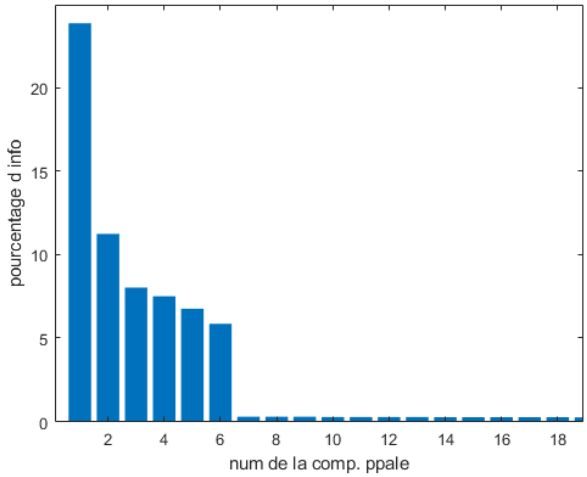


FIGURE 15 – Pourcentage d'information apporté par chaque composante principale.

On remarque que les 6 premières composantes principales contiennent une part importante de l'information. Comme on a 6 composantes principales essentielles, on peut conjucturer qu'on ait 7 classes (d'après la question 5). Vérifions cette conjecture.

La figure 16 représente la visualisation de données dans le repère défini par les deux premiers axes canoniques et le quatrième.

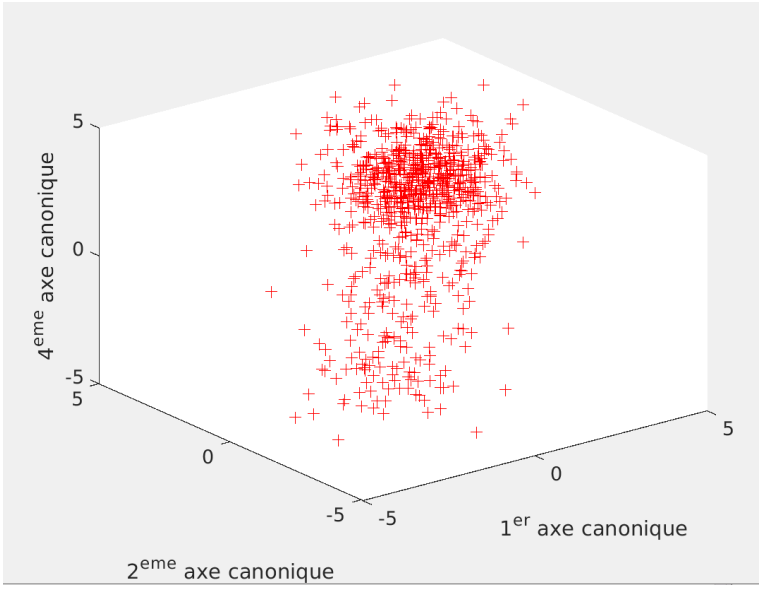


FIGURE 16 – Données dans le repère défini par les deux premiers axes canoniques et le quatrième.

La figure 17 représente la visualisation de données dans le repère défini par les deux premiers axes principaux et le quatrième.

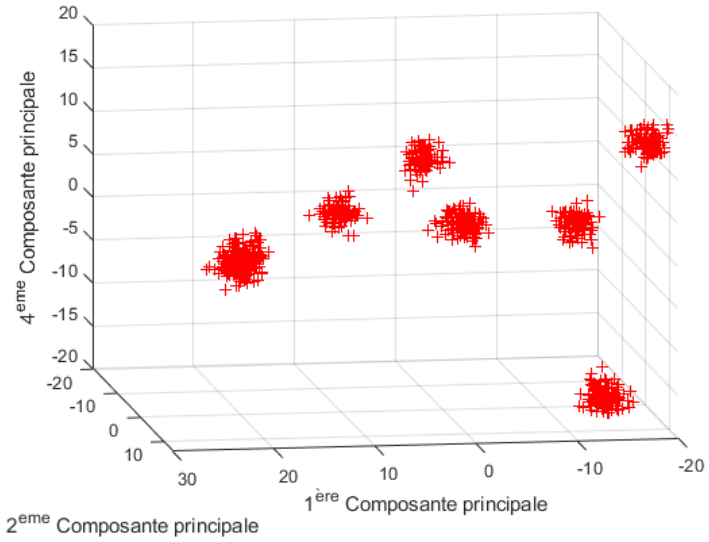


FIGURE 17 – Données dans le repère défini par les deux premiers axes principaux et le quatrième.

En visualisant, non pas sur les trois premiers axes principaux, mais sur les deux premiers et le quatrième, on peut distinguer 7 classes (remarque : on ne peut néanmoins pas distinguer de classes si on considère les axes canoniques (figure 16)).

**Question 7**

On veut classifier les variables.

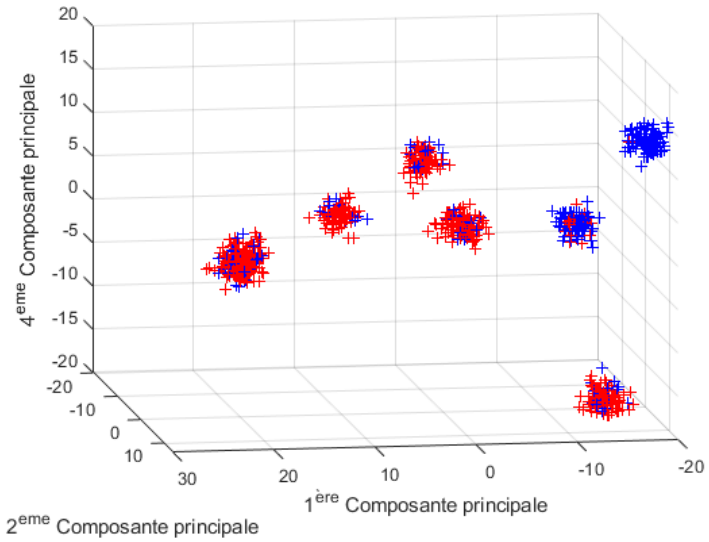


FIGURE 18 – Classification de variables.

# 5 L'ACP et la méthode de la puisance itérée

**Question 8**

Soit $x$ un vecteur propre de $H^T H$ et $\lambda$ sa valeur propre associée.

$$HH^T Hx = H\lambda x$$

$$H^T Hx = \lambda x$$

$$HH^T(Hx) = \lambda(Hx)$$

Ainsi $\lambda$ est aussi une valeur propre de $HH^T$, mais associée au vecteur propre $Hx$.

**Question 10**

*eig* est mieux car il renvoie tous les couples propres et non pas un seul comme avec la méthode de la puissance itérée sans déflation. Si on veut faire une ACP, pour réduire les dimensions de l'espace, on a besoin de tous les couples propres significatifs.

**Question 11**

On déduit des résultats obtenus sur Matlab qu'il faut l'appliquer sur la petite matrice, c'est-à-dire celle de plus petite dimension entre $n \times n$ et $p \times p$.

# 6 Conclusion

Les données ne peuvent se visualiser que dans, au maximum, un espace à trois dimensions. Or les données sont, en général, représentées par plus de 3 variables.

Donc il faut envisager une réduction de cet espace de variables, en ne gardant que celles qui contiennent un pourcentage important de l'information.

On peut réaliser des clusters qui permettent de rassembler des données qui présentent des points communs, que les donnnées des autres clusters ne présentent pas. On peut tout aussi réaliser des classes de variables.

L'obtention des valeurs et vecteurs propres de la matrice $\Sigma$ peut se faire à l'aide de la fonction Matlab *eig* ou de la méthode de la puissance itérée.