TP-Projet 1

Se familiariser avec l'Analyse en Composantes Principales (ACP)

Première Année, Département SN

Hamza MOUDDENE
Issam HABIBI
Younes SAOUDI

April 4, 2020

Contents

Introduction

Ce document représente le rapport de la première partie du projet de calcul scientifique et analyse de donnée. Cette première partie s'intéresse à l'étude de l'ACP (L'analyse en composante principale).

Cet outil classique et puissant, permet l'étude des échantillons d'énormes tailles caractérisés par un nombre très élévés de variables , en les visualisant sur des espaces de faibles dimensions. La méthode essaye aussi de récupérer un pourcentage extrêmement fiable de l'information apportée par l'échantillon étudié en se basant sur des outils de l'algèbre linéaire.

L'ACP permet aussi une classification claire des classes contenues dans l'échantillon et facilite le partitionnement des données en clusters et cette étude peut être menée grâce à des outils algorithmiques qui diffèrent en terme performance et de complexité.

Partie 1 : Visualiser les données

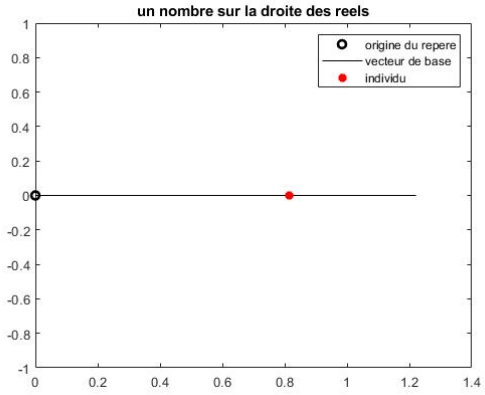
Cette première partie s'intéresse à la visualisation des données qui s'expriment grâce à un unique nombre ou des données représentées par deux ou trois nombres. Ces données peuvent être visualisées naturellement puisqu'elles se modélisent par des vecteurs de taille inférieure à 3 , ce n'est généralement pas le cas pour la majorité des données étudiées et ceci nécessite l'intervention de l'ACP qui serait présentée dans les parties suivantes .

Question 1 : Quelles étaient les données sur lesquelles on a appliqué l'ACP pendant le TP 1 d'analyse de données "Espace de représentation des couleurs" ? Expliquer formellement à quoi correspondait le tableau de données X dans ce TP. Quelles étaient les dimensions des données ?

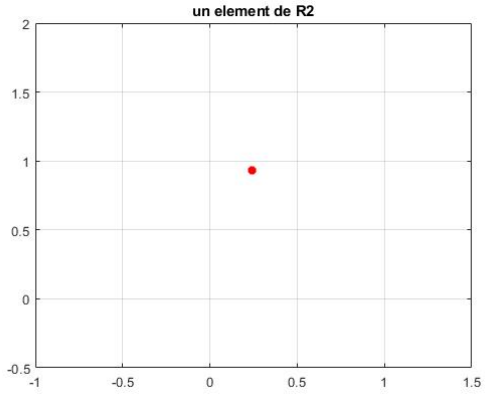
Réponse : Lors du premier TP d'analyse de données "Espace de représentation des couleurs", nous avions traité une image sous forme de matrices tridimensionelles représentant les différents niveau de gris, afin d'appliquer l'ACP, nous avons vectorisé la matrice tridimensionelle dans une seule matrice X de taille Mn,3 sachant que n est un entier très grand qui varie selon le jeu de donnés dont nous disposons. Dans le cas de l'image autumn.tif,par exemple, la dimension était n = 71070.

Question 2 : Completer le script de visualisation.

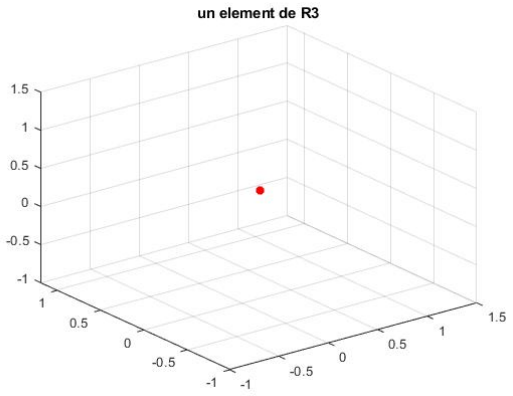
Visualisation : On Visualise dans un premier temps respectivement sur les figures 1,2 et 3 un individu de taille 1,2 et 3. Ensuite, On augmente le nombre d'individus à visualiser en gardant le même nombre de variables , on parle donc d'un "nuage d'individus" qu'on affichera sur les figures 4,5 et 6 en précisant l'individu moyen.



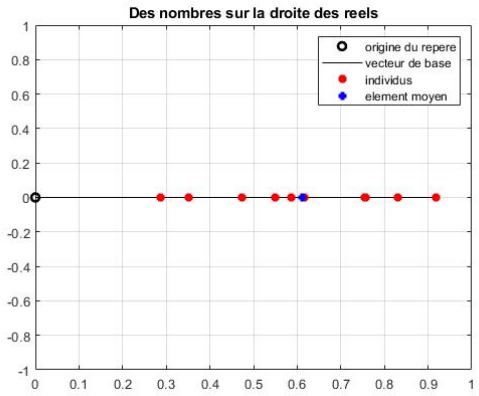
(a) Figure 1



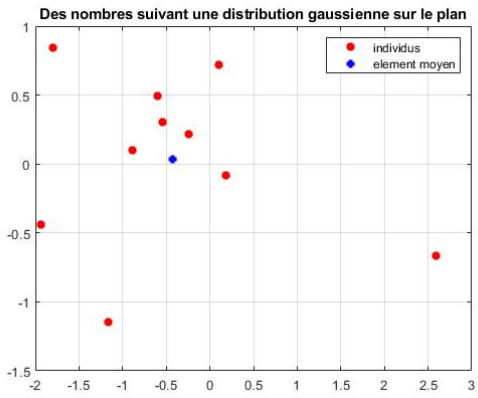
(b) Figure 2



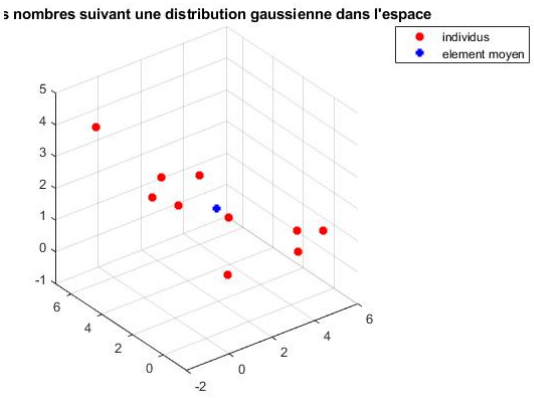
(a) Figure 3



(b) Figure 4



(a) Figure 5



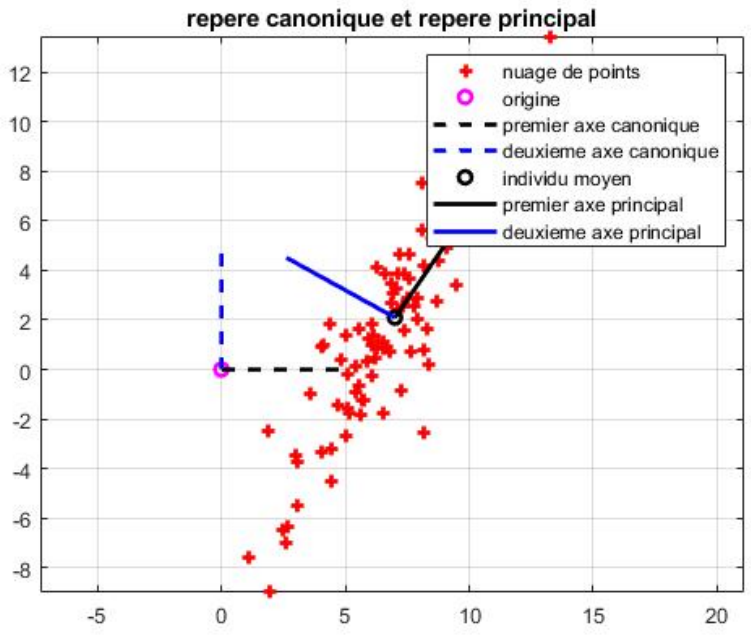
(b) Figure 6

Partie 2 : L'analyse en composantes principales

La visualisation des données de dimension très grande peut être classiquement faite par l'ACP. cet outil consiste à changer le repère de l'espace des données en un repère dont l'origine est la donnée moyenne (l'individu moyen dans notre cas particulier) et le nombre d'axes de la nouvelle base est inférieur ou égal à 3. En projetant l'ensemble des individus sur 1,2 ou 3 de ces axes ,triés par l'ordre décroissant de l'information qu'ils permettent d'obtenir, On arrive à approximer notre tableau de données dans un espace de faible dimension. Ceci revient à diagonaliser la matrice de variance/covariance du problème pour que les axes de la nouvelle base soient décorrélés entre eux, puis centrer et projeter chacun des individus sur les nouveaux axes.

Question 3 : Comparer la projection sur les axes canoniques avec la projection sur les axes principax.

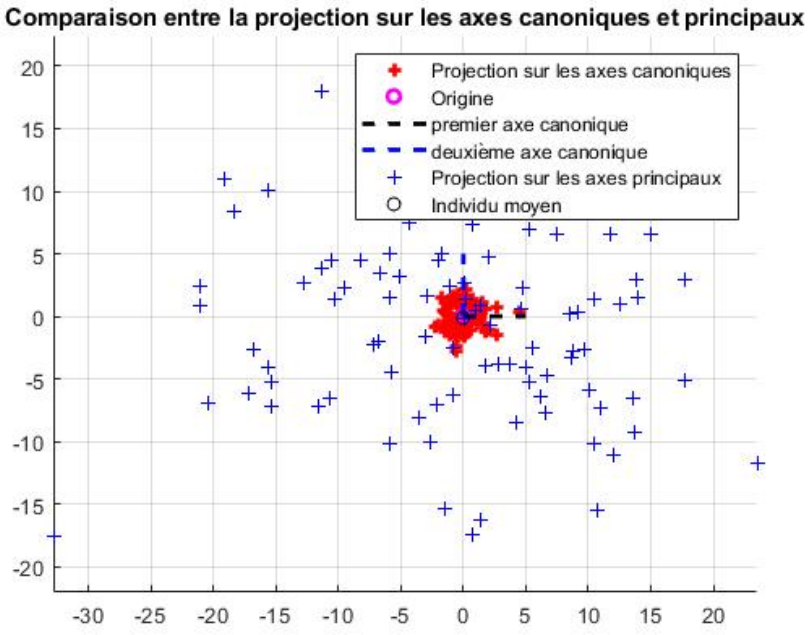
Réponse : Après avoir fait ce travail, on arrive à faire une comparaison entre le repère canonique et le nouveau repère définit par les axes principaux et l'élément moyen .



On remarque clairement dans la figure si dessus que les deux axes du nouveau repère sont en effet ortogonaux et que le nuage des points est bien répartis sur ce nouveau repère.

Pour bien visualiser l'interêt de la projection sur les axes principaux , on dispose d'un jeu de données des

100 individus dont chaque individu est caractérisé par 10 variables. On traçant la projection des individus sur les axes canoniques et sur les axes principaux dans une même figure, on obtient le résultat suivant :



Il paraît donc que les points sont beacoup mieux réparti en projetant sur les axes principaux , ceci assure donc une meuilleur classification des données et on arrivera plus facilement à détecter les clusters de notre jeu de données , ce qui serait le but de la partie suivante .

Question 4 : Comment peut-on quantifier l'information contenu dans les q premières composantes principales à partir de la matrice Σ .

Réponse : La quantification de l'information contenue dans les q première compostantes principale ce fait en calculant la somme des valeurs propres de ces compostantes sur la trace de la matrice Σ .

Partie 3 : L'ACP et la classification des données

Cette partie s'intéresse au partitionnement des données en certains classes . chaque classe doit contenir des données qui partage une homogénéité et qui diffèrent d'un manière ou une autre des données qui figures sur les autres classes. Cette classification peut être faite grâce à la distance euclidienne pour mesurer l'écart entre chaque individu . Ces classe doivent apparaître dans la visualisation en utilisant l'ACP.

Dans un premier temps , on a étudié un jeu de données qui contient deux classes . En projetant ces données sur le premier axe canonique puis sur le premier axe principal comme le montre la Figure 4.1, nous somme arrivés à visualier ces deux classes . La visualisation est plus claire et précise dans le deuxième cas , on conclut que le partitionnement des données est mieux avec l'ACP.

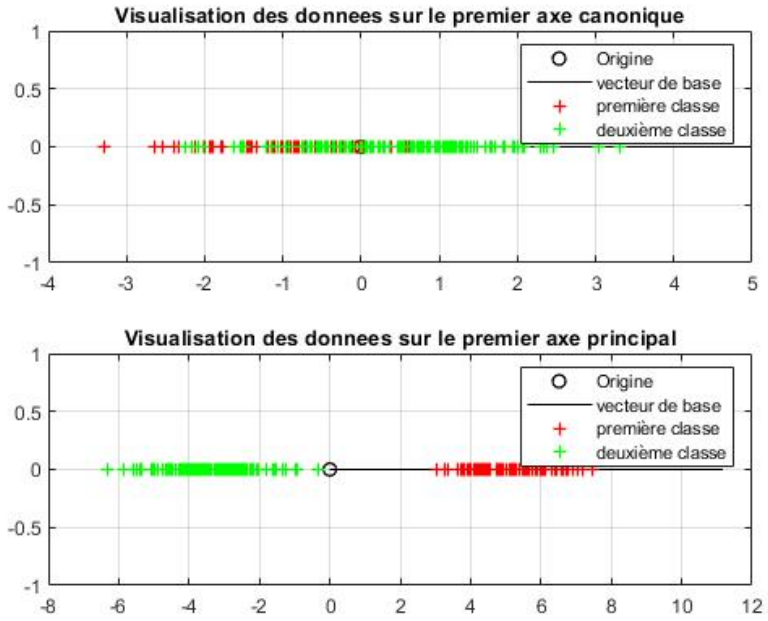


Figure 4.1

Ensuite, En traçant la courbe qui montre le pourcentage d'information apporté par chaque composante principale, on remarque bien dans la figure 4.2 que la première composante principale contient le plus d'information alors que le reste des composantes contiennent quasiment la même quantité d'information.

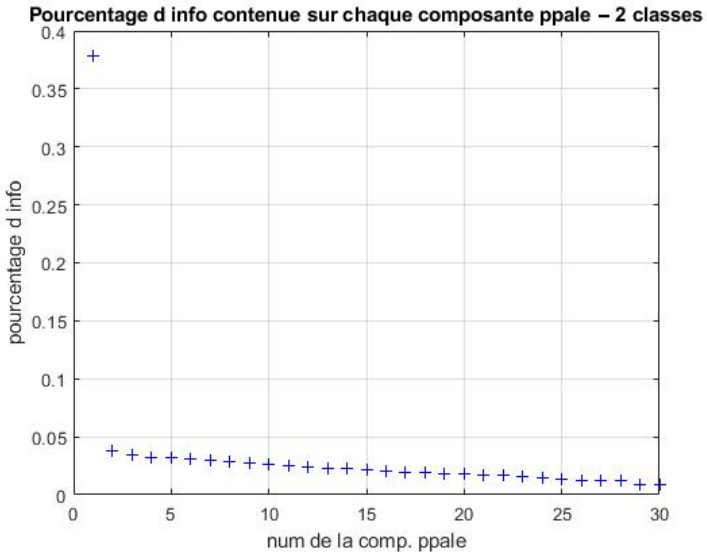
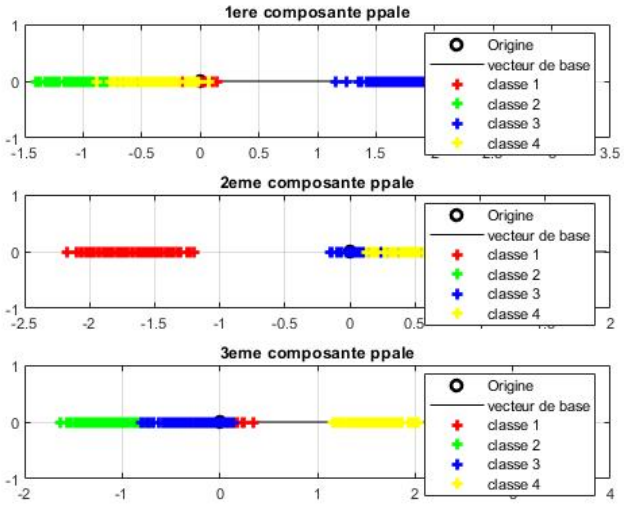


Figure 4.2

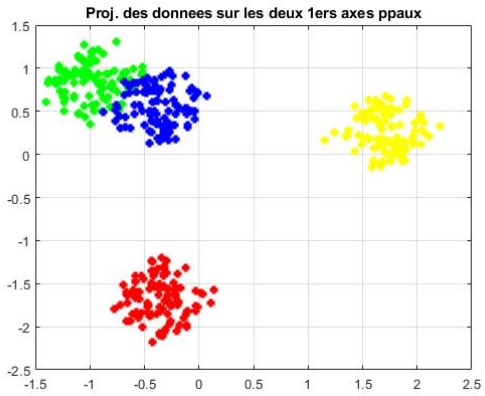
On introduit maintenant un jeu de données contenant 4 classes représentés par 4 tableaux de données différents , concaténés au sein du même tableau de données et on souhaite détecter ces classes avec l'ACP.

Question 5 : Combien de classes est-on capable de détecter avec chaque composante principale? Combien de classes peut on détécter dans le plan? dans l'espace?

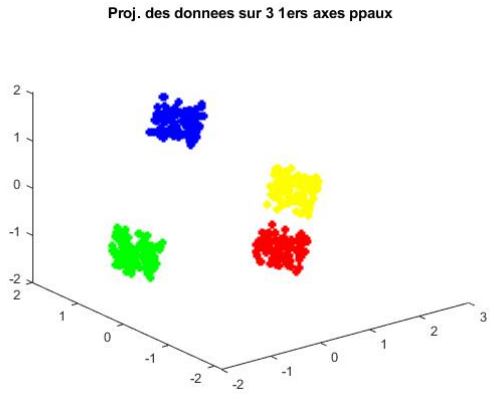
Réponse : La figure ci-dessous représente la projection des données sur les 3 premiers axex principaux , la première et la troisième composante principale nous permet de visualiser les 4 classes (vert, jaune, rouge et blue) alors que la deuxième ne détecte que 3 classes. On regroupant les résultats obtenus grâce aux trois composantes , on retrouve les 4 classes de notre jeu de données.



Si on s'intéresse à la projection des données sur deux ou trois axes principaux , autrement dit sur un plan ou un espace , On retrouve comme le montre les figures (a) et (b) les 4 classes de notre jeu de données .



(a)



(b)

En traçant la courbe montrant le pourcentage d'information contenu dans chaque composante principale dans ce cas (Figure 4.4) , on remarque que les 3 premières composantes principales contiennet une quantité semblable d'information beaucoup plus élévée que le reste des composantes. On conclut donc que si l'échantillon étudié contient un nombre élévé de classes, on aura besoin de plusieurs composantes principale pour récupérer l'information qu'il contient .

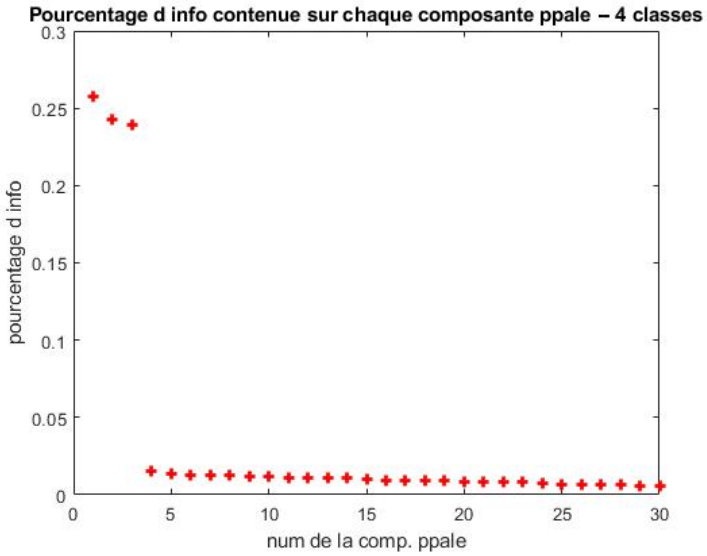


Figure 4.4

Question 6 : Identifiers les classes d'un tableau de données X proposé par l'énoncé

Réponse : Quand on est devant un jeu de données dont les différentes classes ne sont pas précisées, on devrait commencer par trouver le nombre de composantes principales nécessaires pour récupérer une quan-

tité d'information suffisante des données. Ceci peut être fait en tenant compte de la proportion de contraste fournies par les éléments propre de la matrice de variance/covariance . En faisant ça et en traçant la courbe représentant le pourcentage d'information contenu dans chaque composante comme le montre la figure ci-dessus .

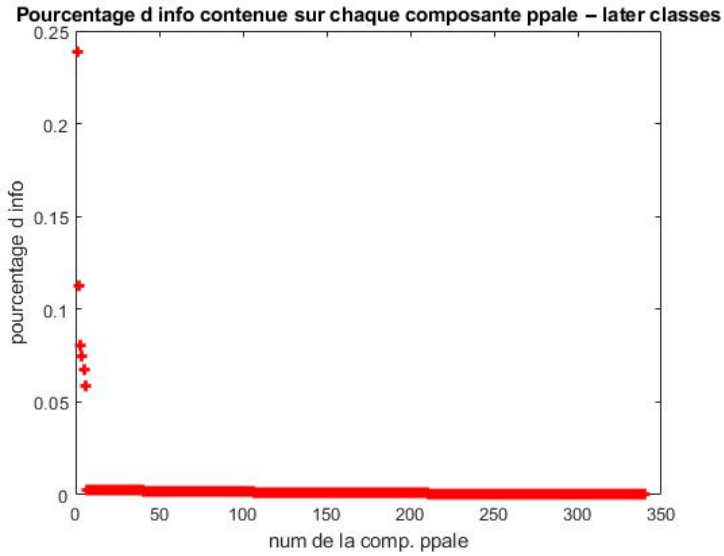
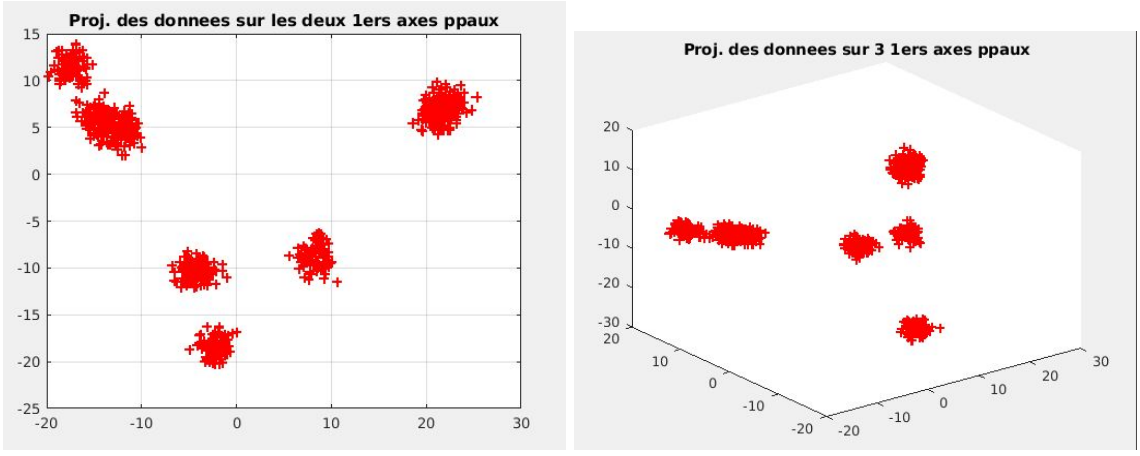


Figure 4.5

Il parait donc que les 6 premières composantes conitennet suffisemment d'information sur le modèle étudié.Si on se contente dans un premier temps seulement des 3 premières compostantes et on reprend le travail fait précedemment en projetant la matrice centrée sur les 3 axes principaux et puis en affichant ces composantes principales sur le plan et l'espace.



(a) (b)

On remarque visuellement selon les figures (a) et (b) que ce jeu donnée est pour le moment partitionné en 6 classes et on peut donc utiliser la fonction "kmeans" de matlab comme l'énoncé propose, cette fonction

correspond à l'algorithme de k-moyennes et elle permet de visualiser les classes comme le montre la figure ci-dessous.

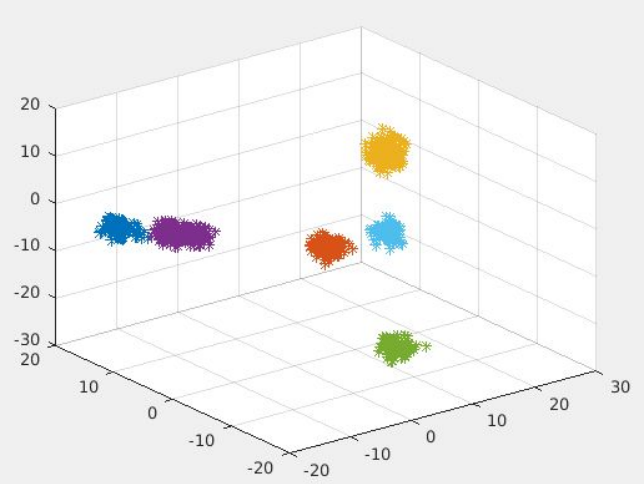
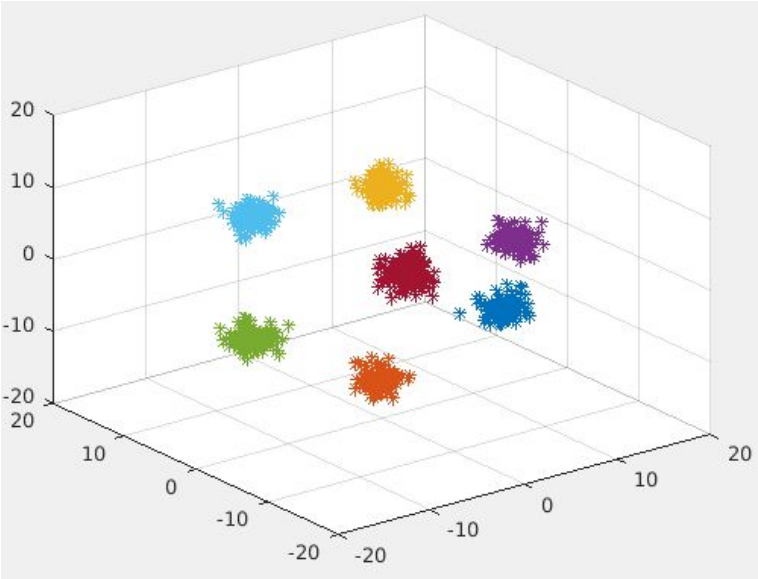


Figure 4.7

Cependant, si on se focalise un peu sur les autres composantes 4,5 et 6 , on remarque que le jeu de données contient 7 classes et pas seulement 6 . Ceci montre que même si les 3 premières composantes comportent un pourcentage plus élevé d'information que les 3 dernières, il est toujours possible qu'elles conitennet une disperssion de classe plus claire.Les 7 classes sont regroupées sur la figure suivante :

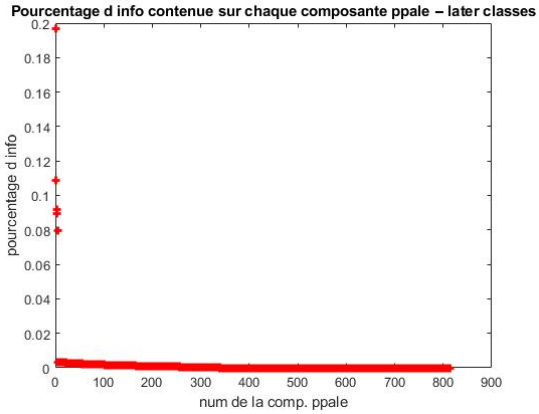


Question 7 : Identifier les classes de variables dans le même jeu de données que la question 6

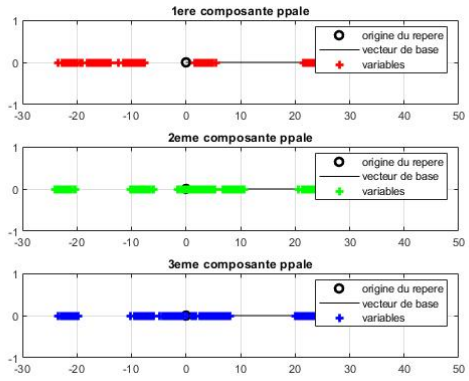
Réponse : Pour visualiser les clusters des variables, il suffit de transposer la matrice X qui représente le tableau de données et de reprendre le même travail qu'avant, les variables jouerons donc le même rôle que

les individus dans les questions précedentes .

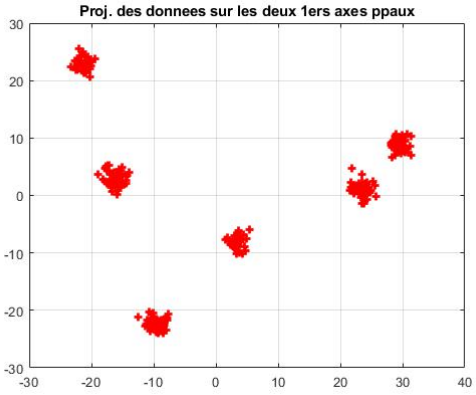
Les figures (a),(b),(c) et (d) montrent que les variables se composent de 6 classes en se basant sur les 3 premières composantes principales.



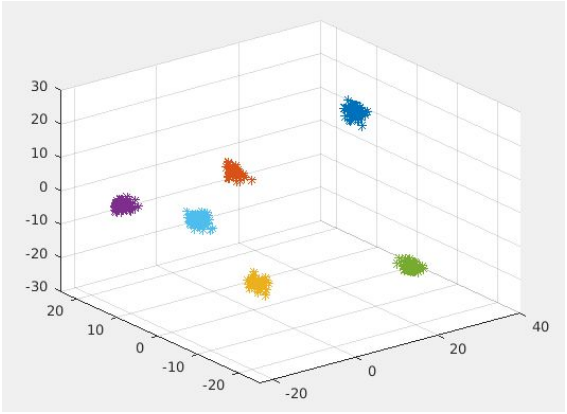
(a)



(b)



(a)



(b) En utilisant kmeans

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

Il n'est pas obligatoire d'utiliser les fonctions pré-définies de matlab (comme eig et sort) pour pouvoir récupérer et trier les éléments propore de la matrie Σ . En effet, on peut aussi avoir recours à la méthode de la puissance itérée qui nous renvoie les éléments propres dans ce même ordre .

Question 8 : Pourquoi connaître les éléments propre de H^tH permer de connaître les éléments propres de ${}^tHH$.

Réponse : Soit λ une valeur propre de la matrice H^tH et x le vecteur propre associé. .
On a : $H^tHx = \lambda$x
Donc : ${}^tHH^tH$x $= \lambda^tH$x
Donc : ${}^tHHy = \lambda$y avec $y = {}^tHx$
d'où λ est aussi une valeur propre de ${}^tHH$ et y est le vecteur propre associé qu'on peut préciser vu qu'on connaît x . Le même raisonnement est valable dans le sens inverse et il permer de conclure.

Question 9 : Comparer la méthode des puissances itérée pour les matrices A^tA et ${}^tAA$

Réponse : Cette comparaison s'effectue en calculent l'erreur relative pour les deux méthodes, l'écart relatif entre les deux valeurs propres trouvés et les temps d'itération .

```
Erreur relative pour la methode avec la grande matrice = 9.926e-09
Erreur relative pour la methode avec la petite matrice = 9.911e-09
Ecart relatif entre les deux valeurs propres trouvees = 1.66e-09
Temps pour une ite avec la grande matrice = 5.290e-03
Temps pour une ite avec la petite matrice = 3.960e-04
fx >>
```

On conlcut donc que l'érreur relative est quasiment la même dans les deux cas, que l'écart relatif est très faible entre les deux valeurs propres obtenues et que la méthode de la puissance itérée est plus rapide avec la matrice ${}^tAA$.

Question 10 : Est-il plus utile en théorie d'utiliser la fonction eig de matlab ou la méthode de la puissance itérée pour calculer les éléments propres de Σ? .

Réponse : Il nous paraît qu'il est plus utile d'utiliser la méthode de la puissance itérée car elle trie les éléments propres en ordre décroissant au fur et a mesure alors que la méthode eig de matlab les renvoit avec leurs ordre initiale dans la diagonale de la matrie Σ et c'est à nous de les trier.

Question 11 : En choisissant d'appliquer la méthode de la puissance itérée pour calculer les éléments propre de Σ? . Sur quelle matrice doit on appliquer la méthode pour optimiser le temps de calcule et de mémoire ?

Réponse : La matrice Σ peut s'écrire sous la forme suivante :

$$\Sigma = (1/n)^t XcXc$$

et selon la question 8 , les deux matrices ${}^tXcXc$ et $Xc{}^tXc$ ont les même éléments propres donc il suffit de choisir entre eux la matrice de la plus faible taille .