Rapport du TP1 du Projet de Calcul Scientifique et Analyse de données)

SE FAMILIARISER AVEC L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Auteurs :
M. LOTFI CHAIMAA
M. MESKINE HATIM
M. WISSAD MEHDI

20 mars 2020

Sommaire

Introduction

Ce projet a pour but de réaliser un mécanisme visant la reconnaissance de visages et d'images de la manière la plus efficace. Ce rapport traite la première partie qui consiste à se familiariser avec la méthode de l'Analyse en compoosantes principales (ACP).
Il sera réparti en quatre parties qui viseront à poser la problématique et d'expliquer le rôle de l'ACP .

Partie I : Visualiser les données

Lorsque l'on cherche à analyser un jeu de données, un réflexe naturel est d'essayer de visualiser ces données. En effet, leur distribution dans l'espace et la façon dont elles sont agencées les unes par rapport aux autres peuvent être des indices précieux sur leurs interactions.

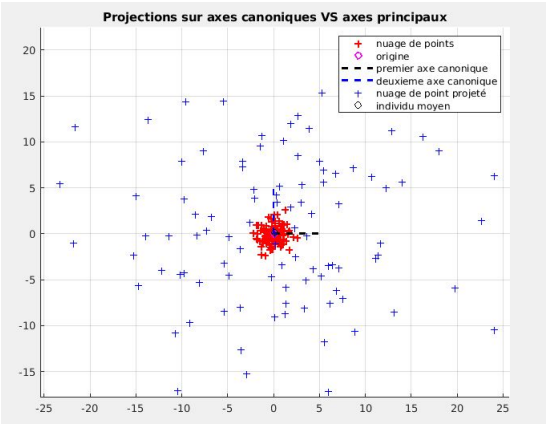
1) Dans le TP1 "Espace de représentation de Données " , les données sur lesquels on a appliqué l'ACP sont les 3 couleurs principales d'une image (Rouge , vert et bleu) . On cherchait ainsi à obtenir une image en niveau de gris.
Le tableau X contenait les données des trois canaux R, V et B . Il était ainsi composé de trois colonnes et de n lignes , avec n = nbr de pixels que contient l'image . Ici , on a n = 105600. Chaque composante de X représentait l'intensité lumineuse d'un pixel selon la couleur de la colonne. Les canaux R, V et B étaient quant à eux de dimension 264*400 (=105600).

2) Voir le code fourni sur le script Visualisation.m

Partie II : L'Analyse en composantes principales

L'analyse en composantes principales (ACP) est une méthode de la famille de l'analyse des données qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

3) La figure ci-dessous représente la projection des individus de X^c sur les deux premiers axes de la base canonique et sur les deux premiers axes principaux. On peut remarquer que les points sont très regroupés dans le repère canonique (en rouge) . Quant au repère principal , ils sont plus dispersés (croix bleus) et on arrive mieux ainsi à classifier les données . L'ACP permet ainsi de mieux faire apparaitre les différentes classes.



[0.5cm]

4) Le niveau d'information visualisée sur les q premiers axes principaux est quantifiable, grâce à la proportion de contraste qui s'écrit en fonction de la trace de Σ

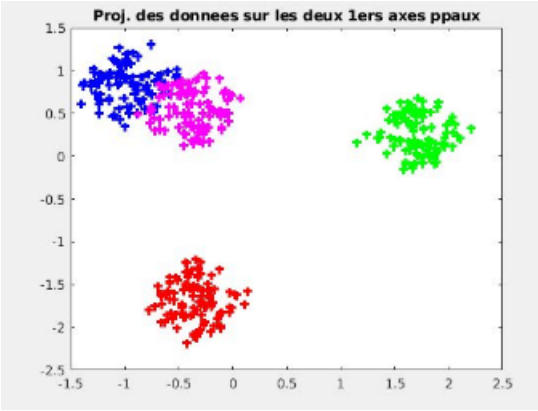
Partie III : L'ACP et la classification des données

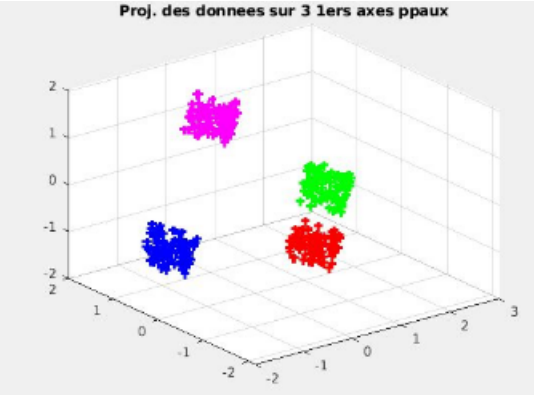
Cette partie s'interesse sur la classification de données qui est un domaine dans lequel on cherche à partitionner un ensemble de données en classes ou clusters, c'est-'a-dire, une partition de l'ensemble des données initial. En général, on définit une mesure de distance sur l'espace des individus. Des données proches doivent appartenir au même cluster quant aux données éloignées , ils doivent appartenir à des clusters différents. Quand les variables sont décorrélées, on peut utiliser la distance euclidienne pour mesurer l'écart entre les individus.

5)

— Chaque composante nous permet de detecter une classe . Les trois ensembles nous permettent finalement de détecter 3 classes .

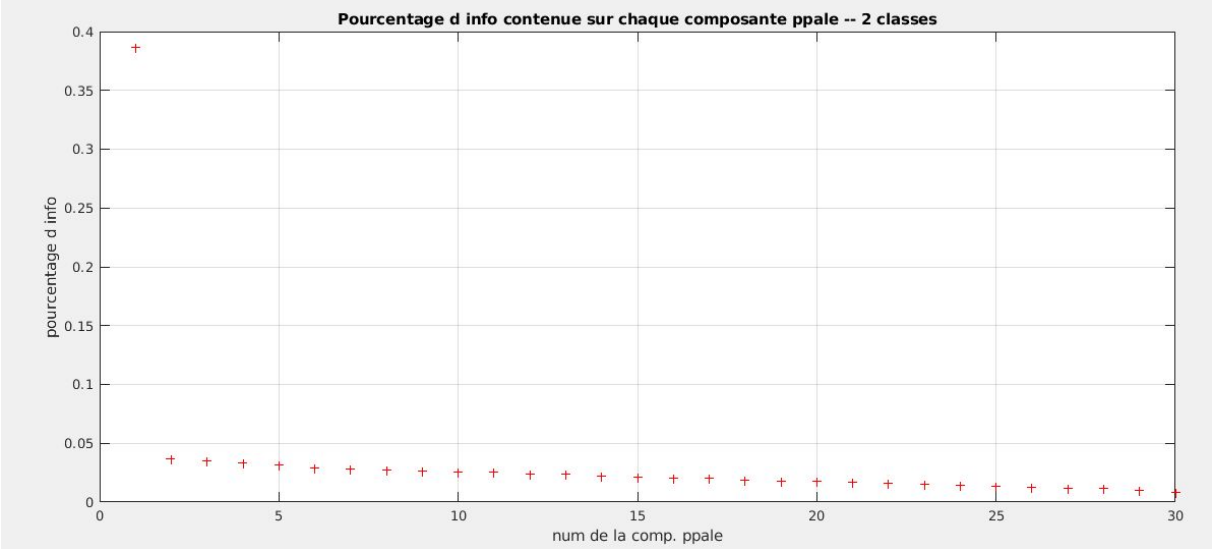
— Dans le plan , on détecte deux classes qui sont bien séparées , et deux autres qui se confondent un peu (figure de gauche). Pour ce qui est du cas dans l'espace , on détecte bien les 4 classes de notre jeu de données qui sont séparées entre elles (figure de droite) .

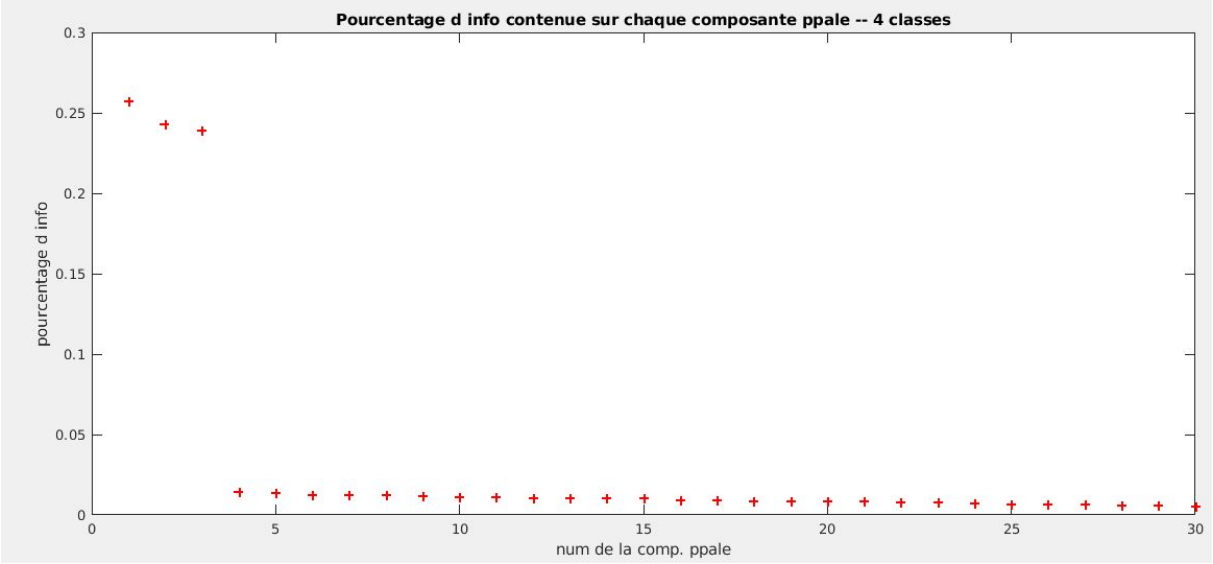




— En traçant la courbe qui montre le pourcentage d'information apporté par chaque composante principale, on remarque que le nombre des composantes principales avec un taux d'information élévé augmente avec le nombre de classes . C'est ce qu'on observe sur les deux figure ci-dessus : pour la classification en 2 groupes , on obtient que seule une com-

posante contient presque toute l'information (figure de dessus). Quant à la classification en 4 classes , les 3 premières composantes ont toutes un taux élevé (figure du dessous) .

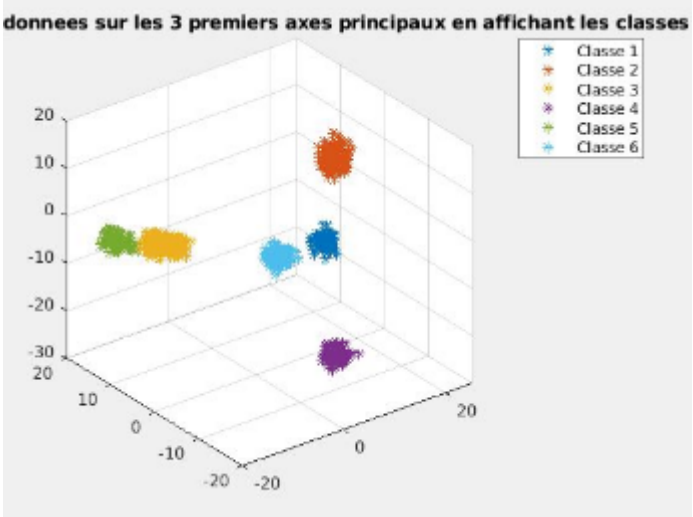




6)- Dans le nouveau jeu données , le script classes_individus.m nous permet de détecter six classes d'individus.

- On se limite à classifier les individus car pour classifier les variables cela reviendrait à considérer la transposée de X et effectuer le même travail . De plus , on obtient les même valeurs propres (car toute matrice possède les même valeurs propres que sa transposee) et donc le même nombre de classes.

7) On constate visuellement d'après les figures que ce jeu de données est partitionné en 6 classes



Remarque : Quand on applique la projection sur les 3 derniers axes principaux (4 ,5 et 6 èmes) , on obtient une visualisation plus claire par rapport celle obtenue par la projection sur les 3 premiers.

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

On pourrait utiliser la tès classique méthode de la puissance itérée avec déflation, qui renverrait les couples propres directement dans l'ordre voulu.

8) Soit une matrice rectangulaire H $\in R^{n*p}$. Soit (X,λ) un couple propre pour la matrice H^TH. On pose $Y = \frac{1}{\sqrt{\lambda}}HX$
Alors on a $HH^TY = \frac{1}{\sqrt{\lambda}}HH^THX$

$= \frac{1}{\sqrt{\lambda}}H(\lambda X)$

= $\lambda(\frac{1}{\sqrt{\lambda}}HX)$

= λY

Ainsi (Y,λ) est un couple propre de la matrice HH^T . De plus Y est en fonction de X , ainsi on peut conclure que le fait de connaitre les éléments propres de H^TH permet de connaitre les éléments propres de HH^T.

9) En exécutant le script Matlab puissance_iteree , on obtient les résultas suivants :

— Erreur relative pour la méthode avec la grande matrice = 9.899×10^{-9}
— Erreur relative pour la méthode avec la petite matrice = 9.935×10^{-9}
— Ecart relatif entre les deux valeurs propres trouvées = 2.54×10^{-9}
— Temps pour une iteration avec la grande matrice = 6.211×10^{-3}
— Temps pour une iteration avec la grande matrice = 2.452×10^{-4}

10) En théorie , il serait plus utile d'utiliser la méthode des puissane itérées pour calculer les éléments propres de Σ si le but est d'effectuer une ACP .En effet, elle permet de trier directement les vecteurs propres de Σ par ordre décroissant des valeurs propres associées, pour obtenir les axes principaux contrairement à la fonction eig qui calcule les valeurs prores sans les trier .

11) Pour minimiser le temps de calcul et la mémoire utilisée , il faudrait minimiser la taille de Σ . Or on sait que $\Sigma = X^t X$ avec X $\in R^{n*p}$. Donc la taille de $\Sigma = n$. Si n < p , alors on applique la méthode de la puissance itérée directement sur Σ. Par contre , si n > p , on appliquerait la méthode sur XX^t qui est de taille p . Et d'après la question 8, cela nous permettrait de connaitre également les valeurs propres de la matrice $X^t X = \Sigma$.