Bureau d’étude : Analyse en composantes principales

MICHELET Lucie / MIMOUNI Bilal / PICHOUX Manon / ALIX Simon

Classe : 3T4

Professeur : M. EL MAHBOUBY

Date de soumission : 21/01/2022

Sommaire

I- Contexte 3

II- Théorie de l'analyse en composantes principales 3

1) Montrons *que* $Var(Xv) = vTCov(X)v$ 3

2) Diagonalisation de la matrice Cov(X) 3

3) Définition du pourcentage de l'explication de la variance de la k-composante principale 4

4) Estimateur de la matrice de covariance de X à partir d'échantillons de X. 5

a. Montrons que $C := EcrTEcr$ 6

b. k-directions principales vk et décomposition en valeurs singulières de $Rcr = UVT$ 7

c. Projection de Rcr dans l'espace vectoriel engendré par les vecteurs de directions principales. 7

5) Critère de projection avec la règle de Kaiser 8

6) Comparaison des nouvelles variables Yk et les anciennes Xi 8

III- Applications concrète de ACP 9

A- Fleurs D'iris 9

B- Pizza 12

C- Boite crânienne 14

D- Election présidentiel 2017 15

E- CAC40 17

IV- Apprentissage renforcé avec l'algorithme des k-moyennes 19

I- Contexte

L'analyse en composantes principales ACP est un outil d'analyse qu'on utilise pour simplifier et tirer des conclusions de base de données. Mais mathématiquement une base de données n'est rien d'autre qu'une matrice que l'on appellera vecteur aléatoire X.

Or, on sait que toute l'information d'une matrice se retrouve dans ses valeurs propres et donc les valeurs singulières, c'est ce que l'on a pu expérimenter avec la compression d'image. La quasi-totalité de l'information *essentielle* se trouve portée par les plus grandes valeurs singulières.

Appliqué à notre problème de base de données, cela revient à trouver quel paramètre ou colonne est la plus *essentielle* à la compréhension de la totalité de la base de données. En pratique, c'est réduire l'espace vectoriel des données dans un nouvel espace vectoriel plus petit, de dimension 2 ou 3 pour pouvoir afficher les données dans un plan ou dans l'espace et donc les analyser plus simplement.

II- Théorie de l'analyse en composantes principales

1) Montrons que $Var(Xv) = v^T Cov(X) v$

L'objectif à présent est de montrer que $Var(Xv) = v^T Cov(X) v$ où $Cov(X)$ est la matrice de covariance du vecteur aléatoire X.

On sait que la covariance est

$$Cov(X) = X^T X$$

Donc on en déduit :

$$v^T Cov(X)\, v = v^T X^T X\, v$$
$$v^T Cov(X)\, v = (Xv)^T Xv$$
$$v^T Cov(X)\, v = Var(Xv)$$

2) Diagonalisation de la matrice Cov(X)

A présent, en justifiant que la matrice de covariance du vecteur aléatoire X, on souhaite démontrer que la k-composante principale de X est donnée par $Y_k = Xv_k$
En partant de $Cov(X) = VDV^T$, où V est la matrice orthogonale et D la matrice diagonale, on a :

La matrice $Cov(X)$ est symétrique si et seulement si

$$Cov(X) = Cov(X)^T$$

Donc :

- Les valeurs propres d'une matrice symétrique réelle A sont positives
- Si V_1 et V_2, deux vecteurs propres associés à λ_1 et λ_2, et $\lambda_1 \neq \lambda_2$ alors :

$$\begin{cases} A\,V_1 = \lambda_1 V_1 \\ A\,V_2 = \lambda_2 V_2 \end{cases}$$

$$\begin{cases} {}^T V_2 A V_1 = {}^T V_2 \lambda_1 V_1 = {}^T \lambda_1 V_1 V_2 = {}^T V_1 A V_2 = {}^T \lambda_2 V_1 V_2 \end{cases}$$

$$\Rightarrow \; {}^T V_1 V_2 = 0$$

V = (V_1 … V_P … V_n)
Alors $V^T A V = D$

Donc $\begin{pmatrix} {V_1}^T \\ \vdots \\ {V_p}^T \\ \vdots \\ {V_n}^T \end{pmatrix} A \left(V_1 \dots \dots V_p \dots . . V_n\right) = \begin{pmatrix} {V_1}^T \\ \vdots \\ {V_p}^T \\ \vdots \\ {V_n}^T \end{pmatrix} \left(AV_1 \dots \dots AV_p \dots . . AV_n\right)$

$$= \begin{pmatrix} {V_1}^T \\ \vdots \\ {V_p}^T \\ \vdots \\ {V_n}^T \end{pmatrix} \left(\lambda_1 V_1 \dots \dots \lambda_p V_p \dots . . \lambda_n V_n\right)$$

$$\begin{pmatrix} \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \lambda_p & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \lambda_n \end{pmatrix} = D$$

Donc V et D existent,
D'où $Cov(X) = VDV^T$

3) Définition du pourcentage de l'explication de la variance de la k-composante principale

Ainsi, pour tout k ∈ [1, n] et k entier.
On en déduit finalement que k-composante principale de X est donnée par $Y_k = Xv_k$:

On note à l'aide de l'énoncé : $Y_{k+1} = Xv_{k+1} = \sum_{i=1}^{n} v_{(k+1)i} X_i$

En procédant par récurrence :

Initialisation :

$$Y_1 = Xv_1$$

Hérédité :

Supposons que la propriété est vraie au rang n tel que $Y_n = Xv_n$
Montrons que $Y_{n+1} = Xv_{n+1}$
Donc avec l'énoncé, nous avons $Y_{n+1} = Xv_{n+1}$

Conclusion :

Ainsi, pour tout k ∈ [1, n] et k entier, nous avons $Y_k = Xv_k$

A présent, notons le vecteur v_k le k-ième vecteur colonne de la matrice orthogonal V.
On peut en déduire :

$$\begin{aligned} Var(Y_k) &= Var(X_{vk}) \\ &= v_k^T Cov(X) v_k \\ &= v_k^T V D V^T v_k \end{aligned}$$

$$V^T v_k = (v_1^T, v_2^T, \dots, v_n^T)^T v_k = \begin{pmatrix} (v_1)^T & vh \\ (v_2)^T & v_h \\ \vdots & \\ (v_n)^T & v_h \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

$$Var(Y_k) = (0, 0 \cdots 1 \cdots 0)\, D\, (0 \cdots 1 \cdots 0)^T = \lambda_k$$

On définit le pourcentage de l'explication de la variance de la k-composante principale tel que :

$$p_k = \frac{\lambda_k}{Trace(Cov(X))}$$

4) Estimateur de la matrice de covariance de X à partir d'échantillons de X.

Les résultats précédents supposent la connaissance du vecteur aléatoire X et par suite la connaissance des variables aléatoires X.
Cependant, il se peut que ces lois ne soient pas connues mais nous avons accès à des échantillons du vecteur aléatoire X.
L'objectif à présent est de donner un estimateur de la matrice de covariance de X à partir d'échantillons de X.

On notera le tableau suivant avec (Ind_1 , Ind_2 , … , Ind_m) avec m-échantillon du vecteur aléatoire X.

	X_1	X_2	⋯	X_n
Ind_1	X_{11}	X_{12}	⋯	X_{1n}
Ind_2	X_{21}	X_{22}	⋯	X_{2n}
⋮	⋮	⋮	⋱	⋮
Ind_m	X_{m1}	X_{m2}	⋯	X_{mn}

a. Montrons que $C := {E_{cr}}^T E_{cr}$

Tout d'abord, montrons que $C := {E_{cr}}^T E_{cr}$ est une approximation d'un estimateur de la matrice de covariance de X, en notant E_{cr} la matrice centrée-réduite associée à E dont les coefficients sont donnés par les variables aléatoires :

$$(E_{cr})_i = \frac{X_i - \overline{X_\iota}}{\sqrt{m-1}\overline{\sigma_\iota}}$$

On cherche à montrer que $Cov(E_{cr})$ est un estimateur avec un biais asymptomatique de $Cov(X)$ c'est-à-dire

$$\lim_{m\to\infty} E[Cov(E_{cr}) - Cov(X)] = 0$$

En reprenant

$$C = E_{cr}^T E_{cr}$$
$$C = Cov(E_{cr})$$

Soit,

$$E_{cr} = (\, E_{cr1}\; E_{cr2}\; E_{cr3} \dots\; E_{crm})$$

On a :

$$E(C) = E(E_{cr}^T E_{cr})$$
$$= E\left(\begin{pmatrix} E_{cr1}^T \\ E_{cr2}^T \\ \vdots \\ E_{crm}^T \end{pmatrix} \quad (\, E_{cr1} \quad E_{cr2} \quad E_{cr3} \dots \quad E_{crm})\right)$$

On reconnait que chaque terme de C est un produit scalaire entre $E_{cr\,i}^T$ et E_{crj} , que l'on explicite :

$$E(C_{ij}) = E(E_{cr\,i}^T E_{crj}) = E\left(\left(\frac{X_i - \overline{X_c}}{\sqrt{m-1}\overline{\sigma_\iota}}\right)^T . \left(\frac{X_i - \overline{X_c}}{\sqrt{m-1}\overline{\sigma_\iota}}\right)\right)$$

D'après l'énoncé

$$= \frac{E(\left[X_i - E\overline{(X_i)}\right]^T \left[X_j - E(X_j)\right]}{(m-1)\overline{\sigma_i}\overline{\sigma_j}}$$

Or, $proj_k(X_i, X_j) = \ E[(X_i - E(X_i))^T (X_j - E(X_j)]$
Donc,

$$E(C_{ij}) = \frac{Cov(X_i, X_j)}{(m-1)\overline{\sigma_i}\overline{\sigma_j}}$$

$$\lim_{m\to\infty} E(C_{ij}) = \ Cov(X_i, X_j)$$

Donc

$$E(C) = Cov(X)$$

D'où, C est un estimateur avec un biais asymptotique de $Cov(X)$.

b. k-directions principales v_k et décomposition en valeurs singulières de R_{cr}= $U \sum V^T$

Ensuite, on sait que $Var(Y_k) = \ \lambda_k$ avec λ_k les valeurs propres de E_{cr}.
Or, dans la décomposition SVD de E_{cr}, les valeurs singulières de E_{cr} sont les σ_k, donc par définition :

$$\sigma_k = \sqrt{\lambda_k}$$

D'où,

$$Var(Y_k) = \ {\sigma_k}^2$$

c. Projection de Rcr dans l'espace vectoriel engendré par les vecteurs de directions principales.

A présent, on souhaite montrer que les données centrées réduites R_{cr} peuvent être projetées dans l'espace vectoriel engendré par les vecteurs de directions principales : $Vec(v_1, \dots, v_k)$ et que la matrice des composantes projetées sur cet espace vectoriel s'écrit par blocs colonnes.

Les coefficients de E_{cr} sont réels donc il existe une décomposition SVD de $E_{cr} = U \sum V^T$
Où U et V sont des bases propres de E_{cr}.

On a :

$$E_{cr}\, u_i = \ \lambda_i\, u_i$$
$$E_{cr}\, u_i = \ {\sigma_i}^2\ u_i$$

Donc,

$$Cov(E_{cr}\, u_1, \dots, E_{cr}\, u_k)$$

Où $E_{cr}\, u_1$ est la projection de E_{cr} dans la base $Vect(v_1, v_2, \dots, v_k)$

On retrouve une ressemblance avec le théorème Ecart Young Minkowski, ou seulement les k première valeur propre porte l'essentiel de l'information d'une matrice.

5) Critère de projection avec la règle de Kaiser

Enfin, on veut montrer

$$\frac{1}{n}\sum_{i=1}^{n} Var(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i$$

Or, on observe que :

$$\sum_{i=1}^{n} \lambda_i = n$$

Donc,

$$\frac{1}{n}\sum_{i=1}^{n} Var(Y_i) = \frac{n}{n} = 1$$

Le critère de Kaiser consiste à dire que les k-premières composantes de Y représente une variance cumulée de δ % donc que l'ACP de paramètre k fournit δ % de l'information de la base de données.

$$\delta = \frac{1}{n}\sum_{i=1}^{k} {\sigma_i}^2$$

6) Comparaison des nouvelles variables Y_k et les anciennes X_i

Pour finir l'analyse, on veut justifier que $Cor_i(Cor(Y_1, x_i), \dots, Cor(Y_k, x_i)$ appartient à la boule unité B^k de $\mathbb{R}^k$.
Nous souhaitons donc montrer que la corrélation est bornée par 1.
On a,

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

A présent, posons une variable $Z = X + tY$
Donc,

$$Var(Z) = Var(X + tY)$$

$$Var(Z) = \sum_{1=1}^{n} ((X_i - \bar{X})^2 + t(Y_i - \bar{Y}))^2$$

$$\sum_{1=1}^{n} (X_i - \bar{X})^2 + 2t \sum_{1=1}^{n} (X_i - \bar{X})\,(Y_i - \bar{Y}) + t^2 \sum_{1=1}^{n} (Y_i - \bar{Y})^2$$

$$Var(X) + 2tCov(X,Y) + t^2 Var(Y)$$

Maintenant, nous procédons à la résolution d'un polynôme du second degré sachant que par définition, la Variance est positive ou nulle donc son discriminant est négatif :

$$\Delta = 4Cov(X,Y)^2 - 4Var(X)Var(Y) \leq 0$$

$$Cov(X,Y)^2 \leq Var(X)Var(Y)$$

$$|Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)} = \sigma_X \sigma_Y$$

$$\frac{|Cov(X,Y)|}{\sigma_X \sigma_Y} \leq 1$$

$$|Cor(X,Y)| \leq 1$$

Ainsi, on peut que conclure que pour tout k, Cor_i appartient à la boule unité de B^k de $\mathbb{R}^k$.

Pour conclure cette partie théorique, nous pouvons ajouter que le cercle de corrélation permet d'interpréter les relations entre les différentes caractéristiques des bases de données que nous allons étudier. Si les flèches relativement proches on peut dire qu'elles sont corrélées tandis que si les flèches sont opposées, elles sont inversement corrélées.

III- Applications concrète de ACP

Maintenant que nous avons vu les mathématiques derrière l'ACP et le lien fort avec la décomposition en valeur singulière SVD. Nous avons, sous python, codé différentes fonctions qui permettent d'analyser, de catégoriser et d'expliquer les phénomènes en question.

De la première partie découle deux matrices d'informations qui résument l'ensemble de la base de données X. La matrice de projection $proj$ et la matrice de corrélation cor, qui respectivement nous permettent de tracer les données dans un espace plus petit et d'étudier les corrélations entre les paramètres.

A- Fleurs D'iris

Premièrement, nous commençons par l'étude de la base de données de fleurs d'iris. Cette base de données est composée des différentes parties d'une fleur comme les sépales et les pétales.

Nous utilisons l'ACP sur une base de données où chaque colonne représente une des parties de la fleur où les lignes représentes les 150 mesures de fleurs. En effet, on retrouve pour chaque colonne les parties suivantes :

- Longueur du sépale
- Largeur du sépale
- La longueur du pétale
- La largeur du pétale

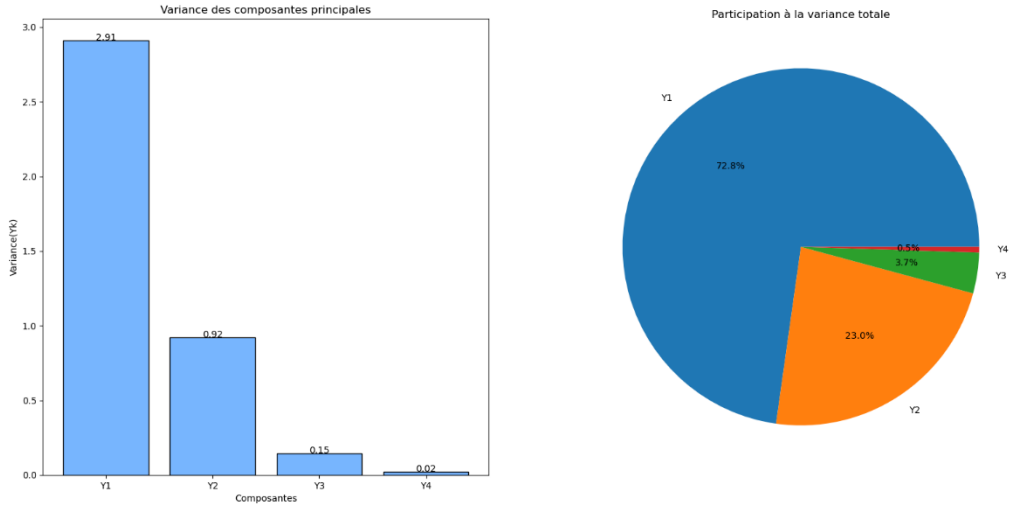


Figure 1 : Variance des composantes principales et leurs participations

Ici, l'objectif est retrouver les k-composantes principales pour pouvoir différencier les fleurs d'iris. Donc, sur la *Figure 1*, nous pouvons voir qu'il y a deux vecteurs principaux sur les quatre. Les deux derniers vecteurs restent négligeables, car en effet, les taux de participations à la variance totale sont infimes (3,7% et 0,5%) comparés aux deux vecteurs principaux (72,8% et 23%).

Ainsi, à partir de ces deux graphiques on peut facilement conclure sur le nombre de k-composantes principales, soit 2-composantes principales.

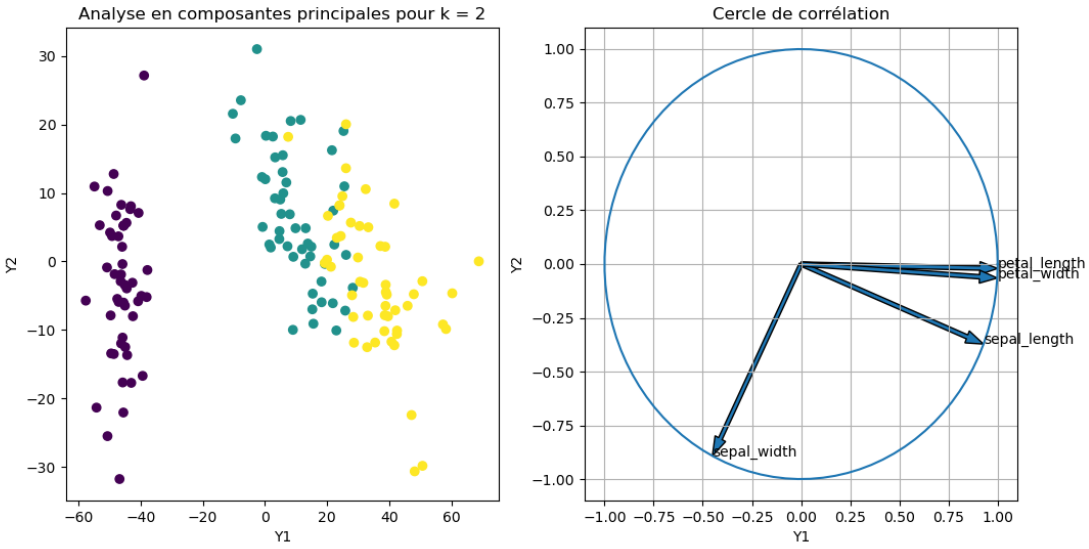


Figure 2 : Analyse en composantes principales et corrélation pour k = 2

Sur la figure ci-dessus, on peut voir que la largeur des pétales est fortement corrélée à la longueur de pétale. Pour vulgariser, cela signifie que plus le pétale de la fleur est long, plus le

pétale est large. Ces deux caractéristiques sont donc importantes pour la différenciation des espèces d'iris. Tandis que les deux autres caractéristiques des sépales ne sont pas corrélées.

De même, avec l'analyse en composantes principales on peut distinguer 3 types distincts de fleurs.

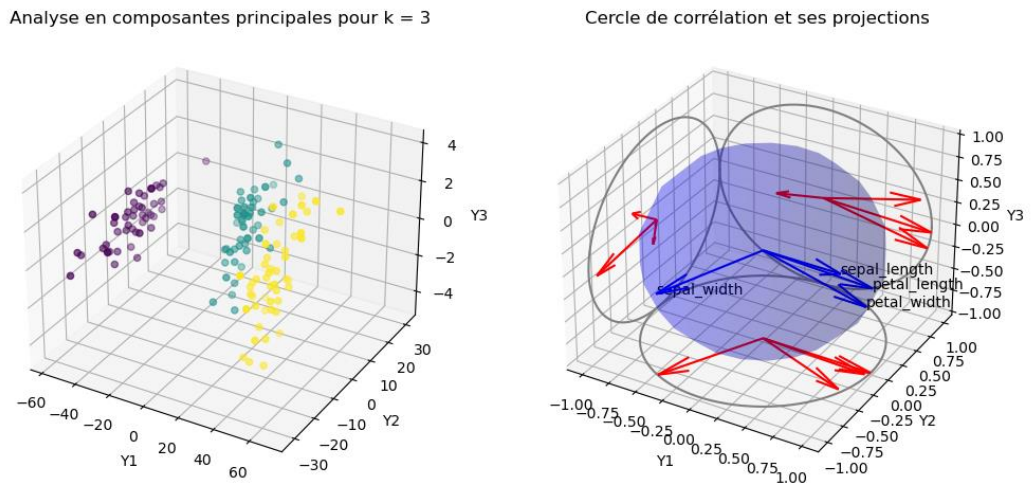


Figure 3 : Analyse en composantes principales et corrélation pour k = 3

Pour la *figure 3,* nous ne pouvons pas déduire davantage d'informations que sur le modèle 2D.

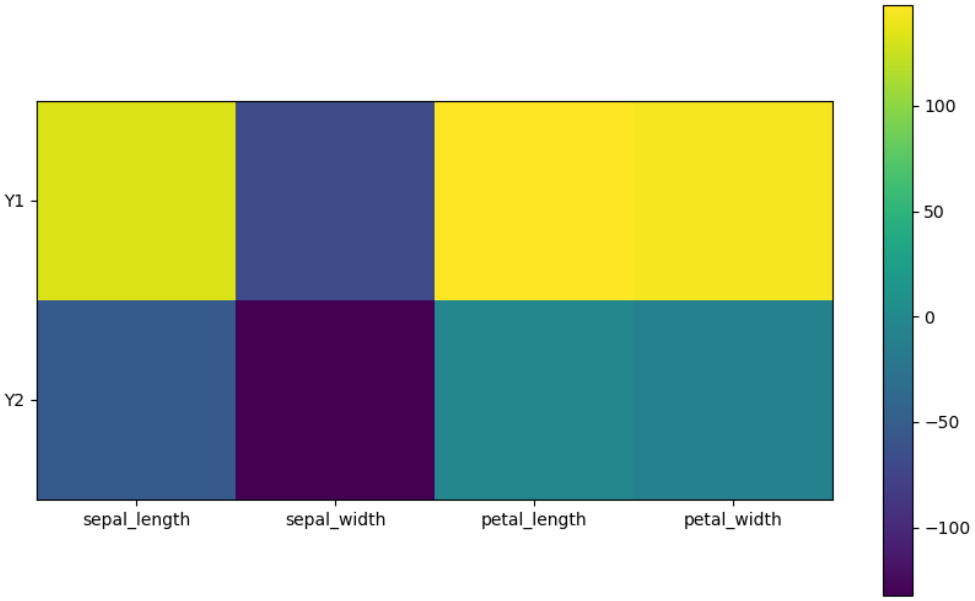


Figure 4 : Matrice de corrélation pour 2-composantes principales

Pour conclure l'étude de cette base de données, la matrice de corrélation ci-dessus présente les deux composantes principales Y_1 et Y_2 et la corrélation des caractéristiques de la fleur. On peut

voir qu'à 2-composantes principales, Y_1 et Y_2 sont différents. Donc, nous n'avons pas besoin de plus de 2-composantes principales.

B- Pizza

Ensuite, pour la seconde base de données, on étudie une base de données composées des différents caractéristiques et nutriments de la pizza.

Nous utilisons l'ACP sur cette base de données où chaque colonne désignent les caractéristiques nutritionnelles de la pizza et où les lignes désignent les 300 mesures de pizzas. On retrouve sur les colonnes les caractéristiques suivantes : la marque, l'eau, les protéines, le gras, « ash », le sodium, les glucides et les calories.

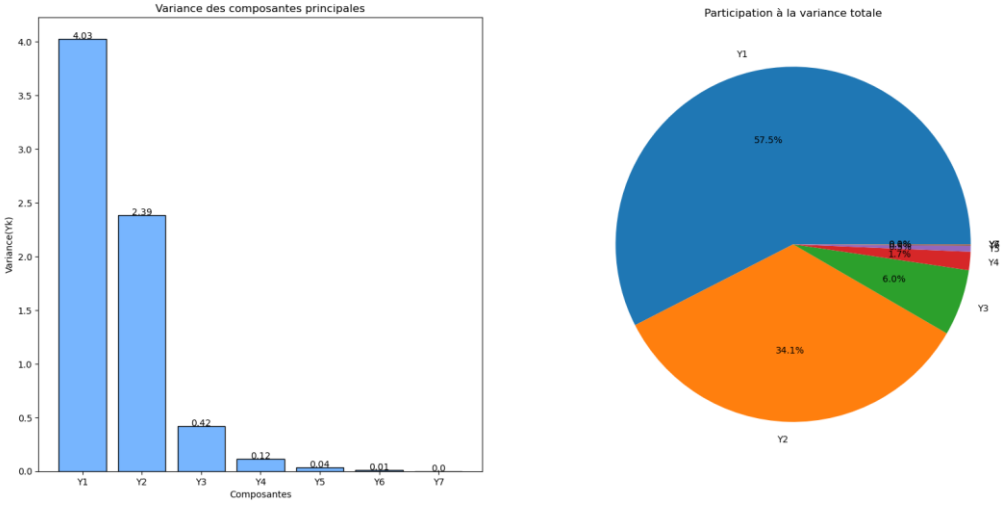


Figure 5 : Variance des composantes principales et leurs participations à la variance totale

De la même façon que pour la base de données Iris, on peut voir que deux caractéristiques se distinguent des autres. En effet, on peut voir deux composantes principales avec des taux respectifs de 57,5% et 34,1%.

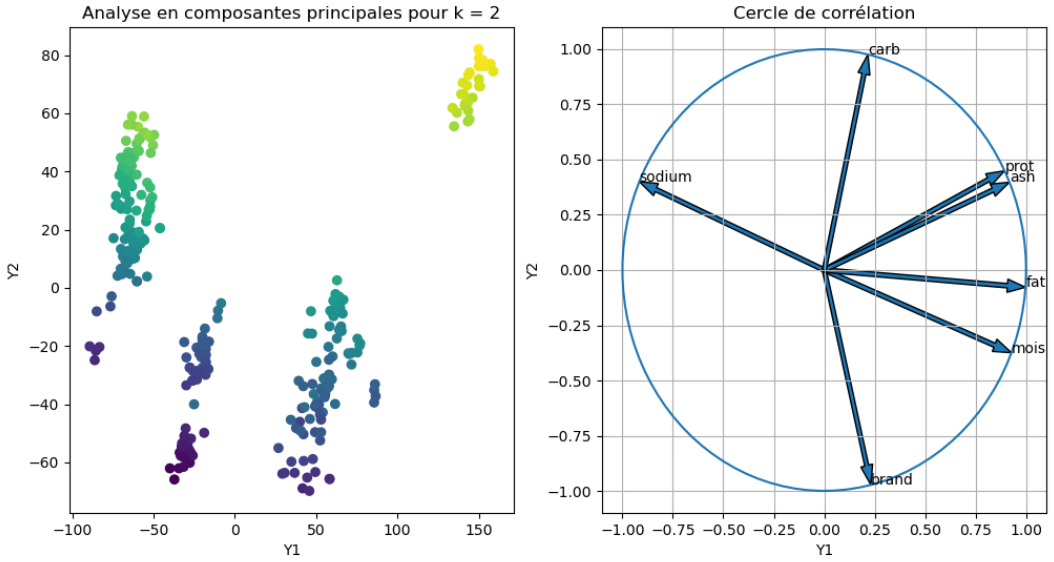


Figure 6 : Analyse en composantes principales et corrélation pour k = 2

On peut voir sur la *Figure 6* une importante corrélation entre le vecteur représentant des protéines et celui de « ash ». Cela signifie donc que les deux caractéristiques nutritionnelles sont intrinsèquement liées. Tandis que, pour les vecteurs représentant l'eau et le sel, eux, sont directement opposé ce qui signifie que plus il y a de sel, moins il a d'eau. Ce concept est cohérent avec la physique puisque le sel est réputé pour absorber l'eau liquide et la vapeur d'eau.

De même, sur l'analyse en composantes principales, on peut très clairement distinguer types distincts de pizzas.

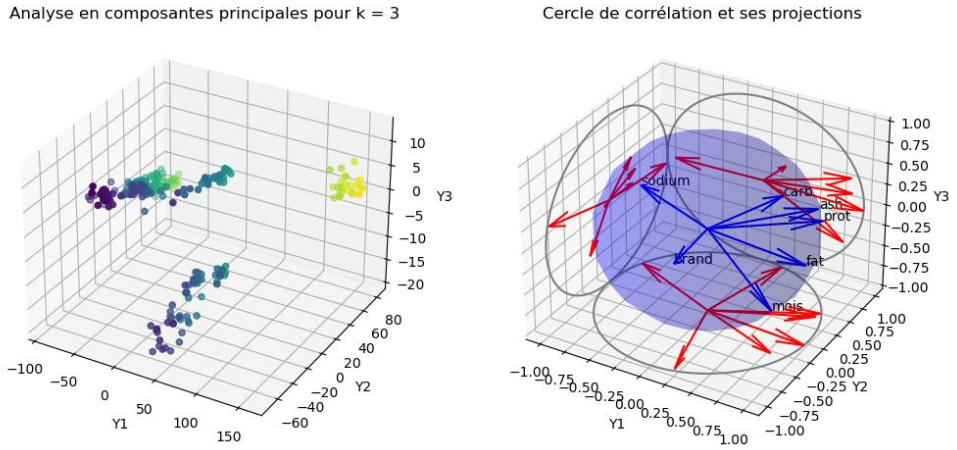


Figure 7 : Analyse en composantes principales et corrélation pour k = 3

Le modèle 3D quant à lui ne montre pas d'informations supplémentaires, mais confirme l'opposition deux vecteurs représentant l'eau et sel ainsi que la cohésion des vecteurs représentants les protéines et « ash ».

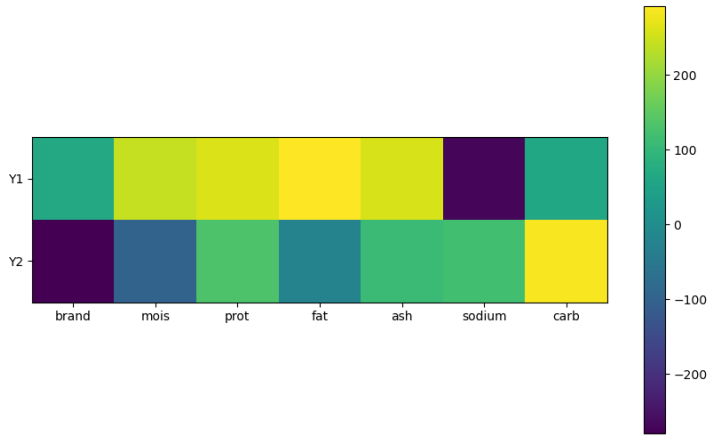


Figure 8 : Matrice de corrélation à deux composantes principales

La matrice de corrélation confirme les analyses puisque nous pouvons voir qu'il n'y aucune similitude entre les composantes principales Y_1 et Y_2. Donc nous n'avons pas besoin de plus de 2-composantes principales.

C- Boite crânienne

Le but de cette étude de donnée sur les différentes caractéristiques d'une boite crânienne est de déterminer la meilleure caractéristique pour différencier les types de boite crânienne.

Pour cela nous utiliserons les données récupérer sur la base de données « Howellmod.csv ». Nous utilisons l'ACP sur une base de données ou chaque colonne représente une caractéristique de la boite crânienne et chaque ligne représente les différentes mesures.

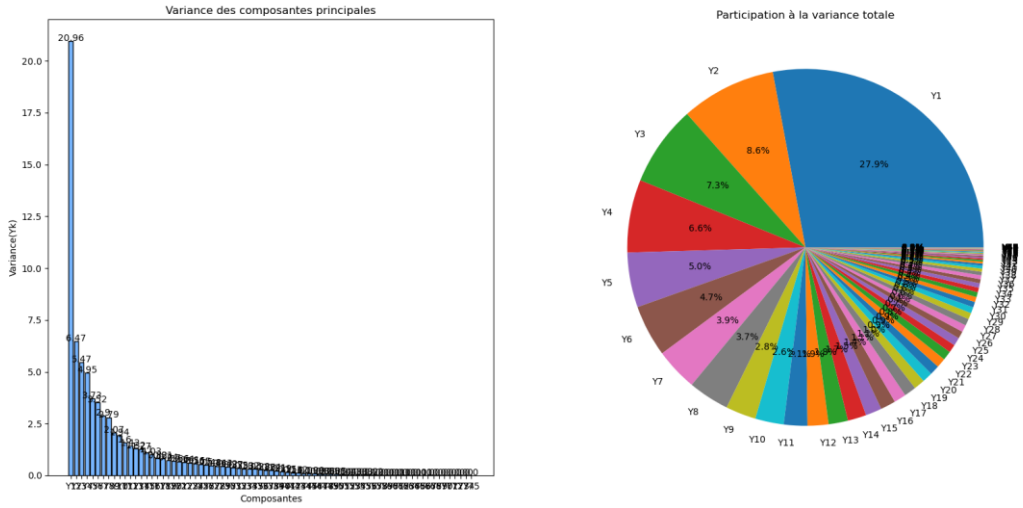
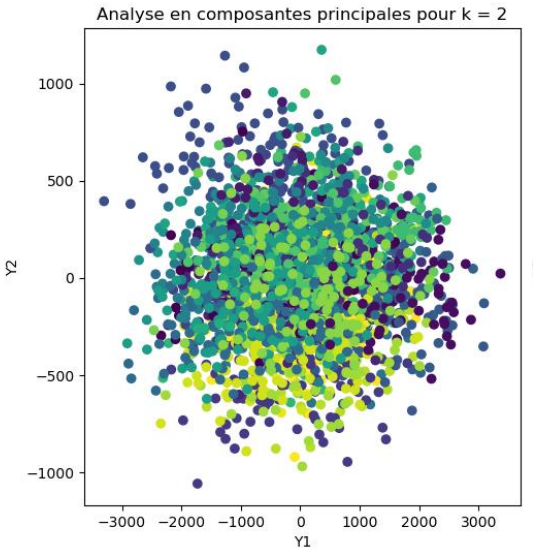
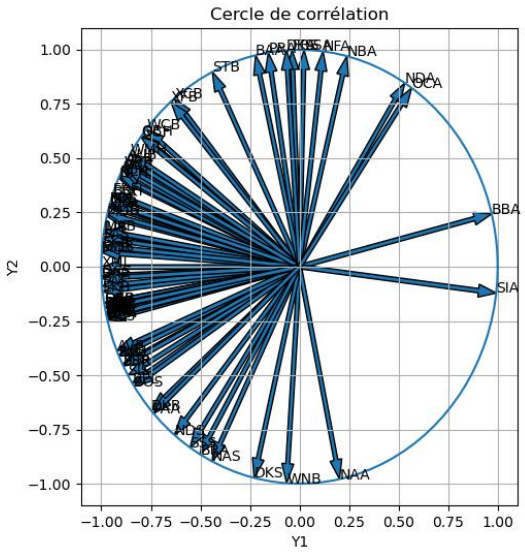
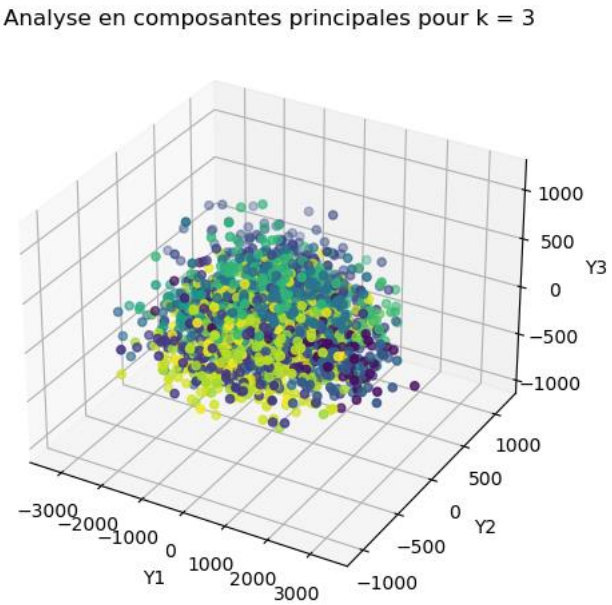


Figure 9 : Variance des composantes principales et leurs participations à la variance totale

Sur la variance des composantes principales, nous voyons très clairement une variance dominante et une participation de la variance totale de 27,9% sur plus de 2500 mesures de boite crânienne. Tout simplement, nous en déduisons une seule composante principale.







D- Election présidentiel 2017

La base de données élection présidentielle récupérée sur le site *data.gouv*[1] nous indique par commune la répartition des voix entre les différents candidats. Le but est de voir si, sans apriori sur les données, nous pouvons en tirer des conclusions sur des sujets d'actualité.

[1] https://www.data.gouv.fr/fr/pages/donnees-des-elections/

La base de données est conséquente car il y a 35000 communes en France donc nous n'afficherons pas tous les points avec l'ACP2D. Pour le cercle de corrélation on peut observer des résultats très intéressant.

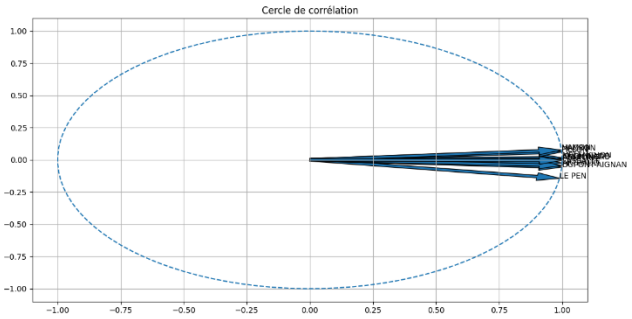


Figure 10 : cercle de corrélation entre orientation politique des communes de France

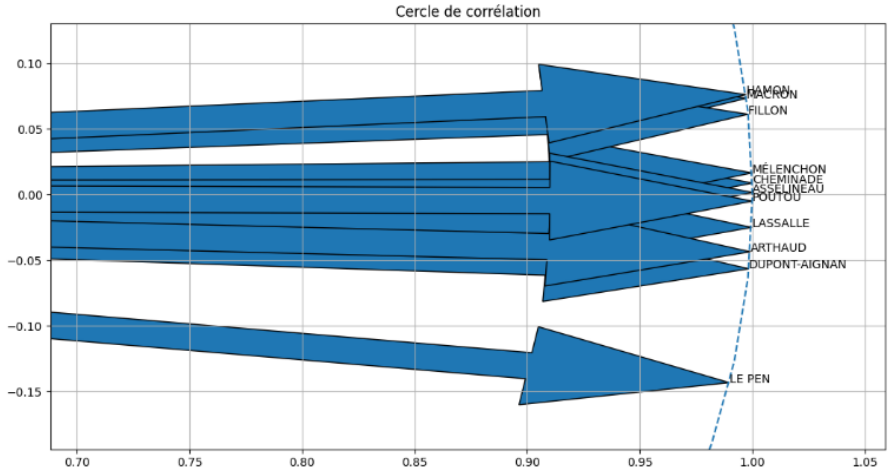


Figure 11 : zoom sur la figure 10

On voit distinctement 3 groupes se dessiner, qui correspondent au type de commune votée pour certain candidat. Par exemple ce sont les mêmes types de commune qui vote pour HAMON, FILLON et MACRON, leurs programmes étant proche ils touchent le même type d'électorat et donc on observe une forte corrélation.

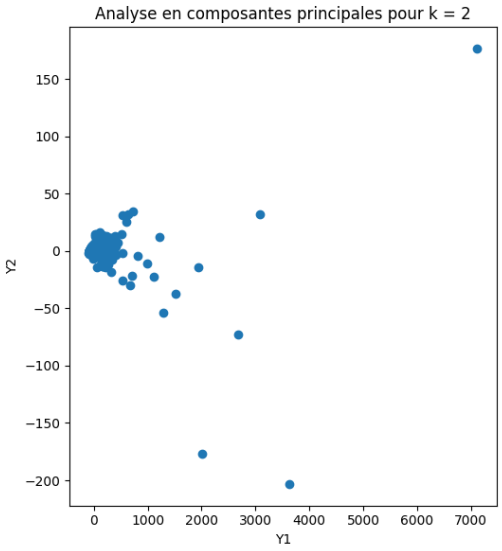


Figure 12: Répartition dans l'espace des communes de France en fonction de leur vote à l'élection présidentiel de 2017

Sur le nuage de point, on ne peut malheureusement pas en dire davantage car nos données ne sont pas labellisées et aucun groupe ne se forme spontanément.

On ne peut pas en dire d'avantage, une amélioration possible serait d'exprimer en plus de la répartition des votes dans les communes la taille de la commune. On pourrait peut-être voir avec des tranches de population bien définit des groupes se former et a partir de ca prédire en fonction de la taille d'une commune le candidat qui aura le plus de voix.

E- CAC40

Le but de cette étude de données sur la valeur boursière des 40 plus grosses entreprises françaises cotées en bourse est de déterminer s'il y a des liens entre les cours des différentes entreprises.

Pour cela, nous utiliserons les données récupérées sur le site *ABC bourse* [2]que nous devrons retravailler. Nous utilisons l'ACP sur une base de données où chaque colonne représente une entreprise du CAC40 et chaque ligne représente un jour entre le 18 décembre 2021 et le 18 janvier 2022.

Pour représenter la variation de la valeur de l'entreprise en bourse chaque jour nous prenons le prix du marché de la fermeture moins le prix de l'ouverture. Si le résultat est positif la côte de l'entreprise a augmenté.

[2] https://www.abcbourse.com/download/historiques?f=ex

La figure 13 nous montre que l'analyse en 3 composantes principales n'est pas nécessaire car les corrélations entre les axes principaux Y_2 et Y_3 sont identiques. On choisit donc de faire l'ACP avec $k = 2$.

La figure 14 est beaucoup plus intéressante et nous dit que certaines actions sont corrélées. L'exemple de 2 entreprises du même secteur d'activité qui sont impactées par un phénomène de pénurie ou autre. Mais pour d'autres, les flèches ont des directions opposées, dans notre cas si l'action PERNOD RICARD augmente (FR0000120693), l'action THALES diminue (FR0000121329).

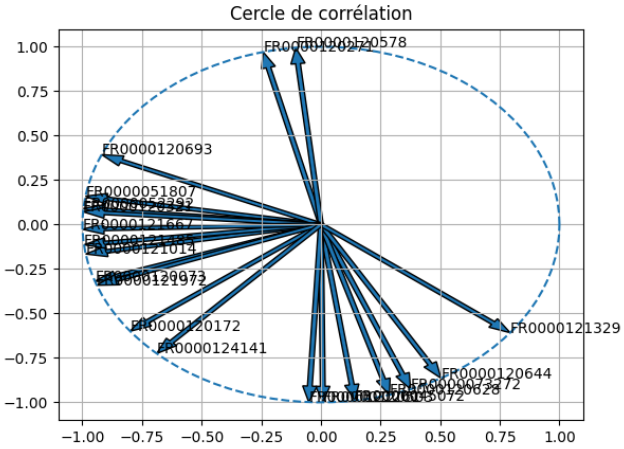


Figure 14 : corrélation entre les prix en bourse des entreprises de CAC40

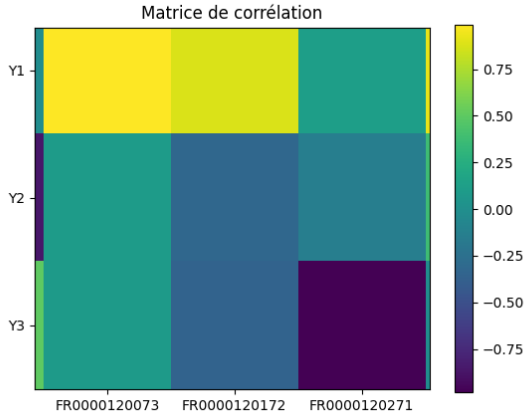


Figure 13 : extrait de la matrice de corrélation qui montre l'inutilité de prendre plus de composante principale

Pour les corrélations entre les valeurs boursières des entreprises, les corrélations sont simples à décrire. Avec la fonction *Approx* on peut afficher en 2 dimensions chaque ligne de la base de données, c'est-à-dire chaque jour du tableau.

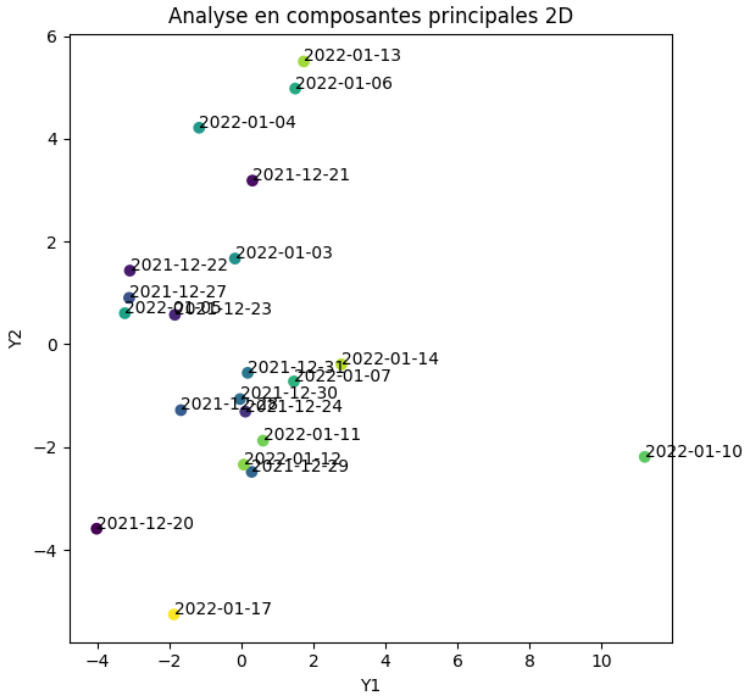


Figure 15 : classification des jours en fonction des variations globales du prix en bourse des entreprises du CAC40

On voit que les jours se répartissent en nuage, on peut dénombrer 3 catégories mais surtout on peut voir des points en périphérie des nuages. La base de données n'a pas de *label*, il est donc difficile de donner un sens à ce graphique contrairement à la figure 14 du cercle de corrélation.

Le cercle de corrélation de l'analyse en composante principale est très intéressant comparé à la projection de cette base sur ses composantes principales. Mais à partir de cette même base de données de *ABC bourse* on pourrait créer d'autres échantillons et en tirer d'autres conclusions.

IV- Apprentissage renforcé avec l'algorithme des k-moyennes

Après avoir vu l'efficacité de l'ACP pour séparer les différentes fleurs et autres données, nous attaquons au problème de classification de ces données. La question est simple sans *label*, c'est-à-dire sans étiquette pour savoir au préalable à quelle catégorie appartient la donnée en question, et s'il est possible tout de même en déduire des groupes.

La réponse est oui grâce à l'algorithme des k-moyennes, cela nous permet de mettre un pied dans l'apprentissage non-supervisé de machine Learning.

Nous testons l'algorithme sur le jeu de données des fleurs iris, nous savons qu'il existe au préalable 3 catégories de fleurs mais voyons ce qui se passe lorsque l'on cherche 2, 3 ou 4 catégories.

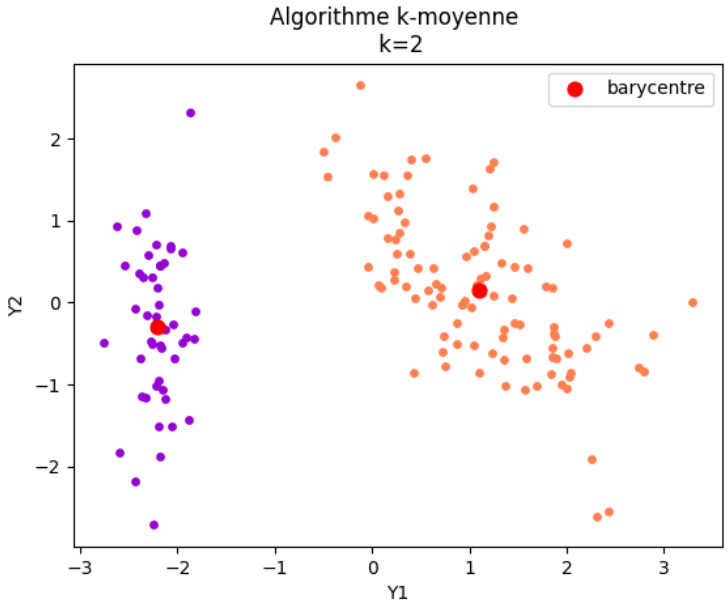


Figure 16: itération finale pour $\varepsilon = 0.01$, $val = 197$

Lorsque l'on cherche à séparer le nuage de point en 2 catégories, le résultat est plutôt satisfaisant. Le résultat semble converger vers cette solution à chaque fois, ce qui n'est pas le cas lorsque l'on cherche 3 catégories.

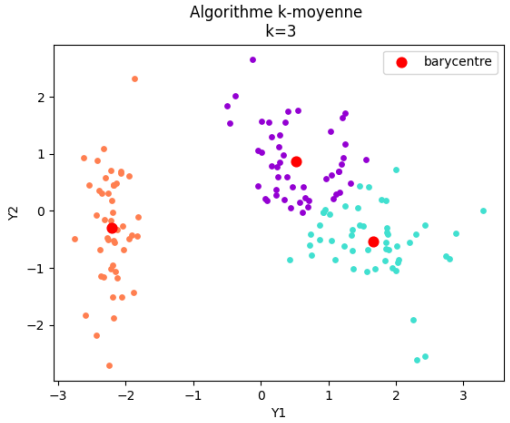


Figure 18: itération finale pour $\varepsilon = 0.01$, $val = 115$

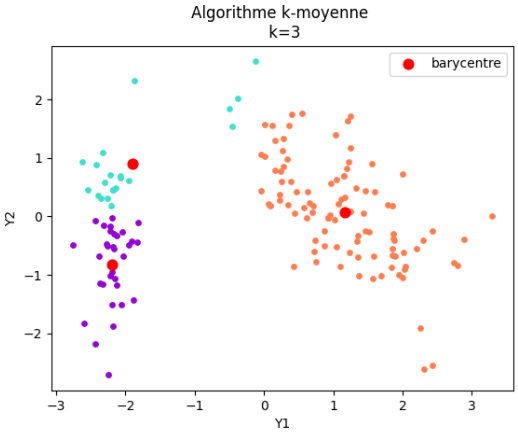


Figure 17: itération finale pour $\varepsilon = 0.01$, $val = 165$

On voit qu'il existe 2 positions d'équilibre pour séparer en 3 le nuage de point et c'est l'information val qui nous indique laquelle est la meilleure. Plus val est faible, meilleure est

sa classification car val représente la somme des distances cumulées de chaque point à son barycentre. Plus cette distance est faible, plus les nuages sont denses.

Pour $k = 3$, on va retenir la classification de gauche qui donne une valeur plus faible de val. On peut continuer de la même manière avec 4 catégories.

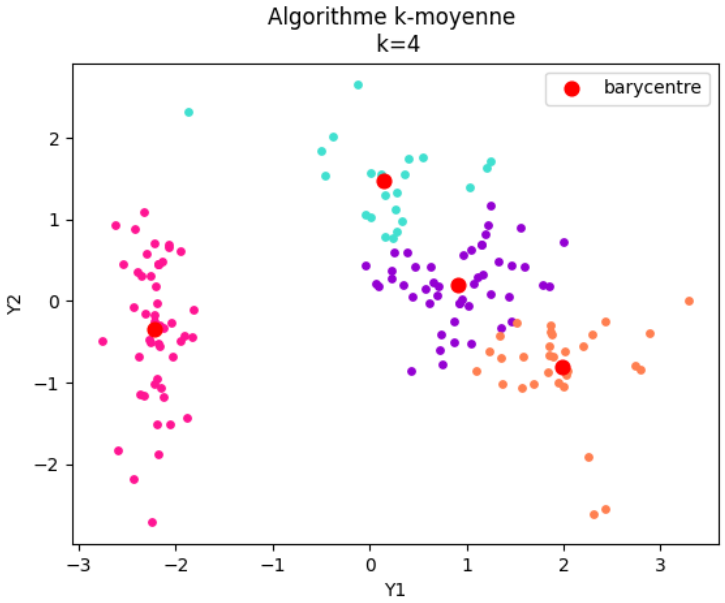


Figure 19: itération finale pour $\varepsilon = 0.01$, $val = 90$

Plus on augmente k, plus val va devenir faible, à l'extrême s'il y a autant de catégories que de points, val vaut 0. Ce critère est bon pour discriminer pour un même k quel équilibre est le meilleur mais pas pour connaitre le nombre idéal de catégories.

Pour cela, c'est le contexte et la nature des données qui aident, pour les fleurs d'iris les catégories représentent des variétés de fleurs. Pour d'autres études comme la boite crânienne, il n'y a pas de réponse évidente et aucune ne s'exclue.