Université de Technologie de Compiègne

Génie informatique

# Rapport du projet de SY09

*Authors :*

Antoine Collas et Julien Crauser

2 mai 2018

# 1 Cuisine

## Question 1

Le jeu de données présent dans le fichier recettes-pays.data contient 51 variables pour 26 individus. Nous pouvons donc noter que nous disposons de plus de variables que d'individus. Parmi ces 51 variables, une seule est qualitative (les origines des recettes) et 50 sont quantitaves. Ces dernières prennent leurs valeurs entre 0 et 1 (0 et 0.82 pour être plus précis). Dans ce jeux de données les recettes ont été agrégées par origine, il y a donc 26 origines (une par ligne de notre tableau individus-variables). De plus, le jeu de données ne présente aucunes valeurs manquantes. Cependant certaines origines sont à remarquer. En effet, il y a par exemple une distinction entre l'Afrique et le Maroc ou encore entre l'Asie et Chine, Japon, Vietnam, Thailande. De plus, les origines semblent être regroupées par région géographique mais des origines comme j̈uive,c̈adiens, sont présentes.

## Question 2

Nous réalisons une ACP sur le jeu de données. Nous obtenons 26 axes principaux : autant que d'individus. En effet, comme il y a moins d'individus que de variables, il est suffisant de prendre 26 axes pour représenter tous les individus.

Nous obtenons les pourcentages d'inertie expliquée visibles sur la figure 1. Les 3 premiers axes représentent 72% de la variance du nuage de points.

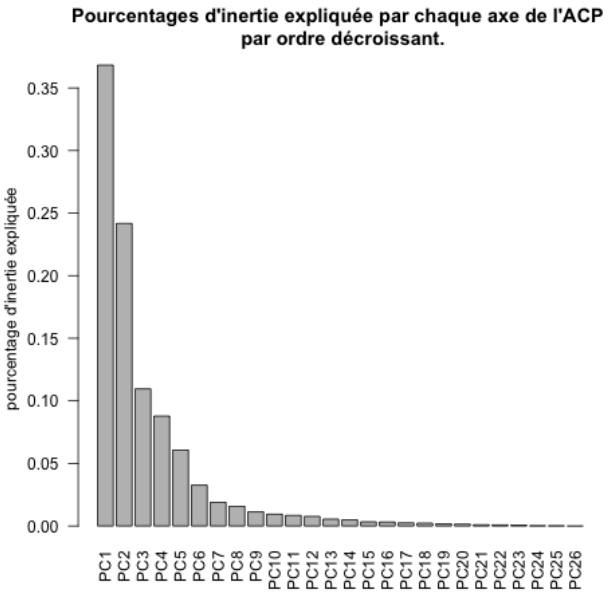
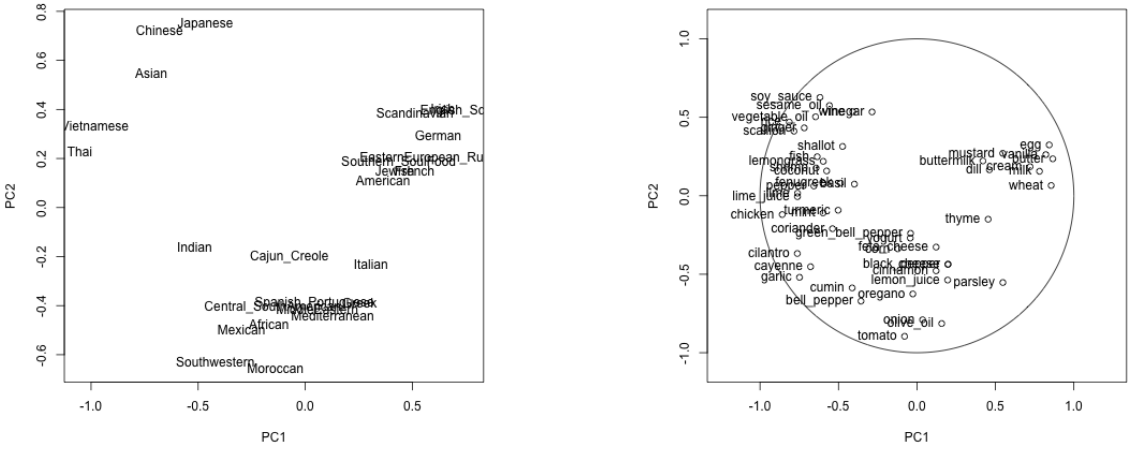



FIGURE 1 – Pourcentages d'inertie expliquée

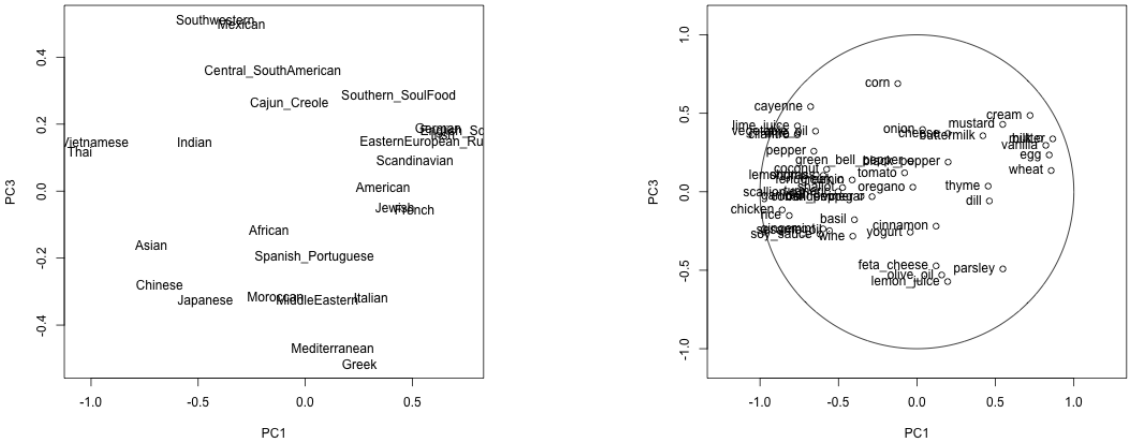
Comme les 3 premiers axes factoriels représentent 72% de la variance totale, nous analysons le nuage de points seulement avec les 3 premiers plans factoriels. Nous obtenons les plans factoriels des

figures 3, **??** et 4. Nous avons associé à chaque plan son cercle des corrélations.



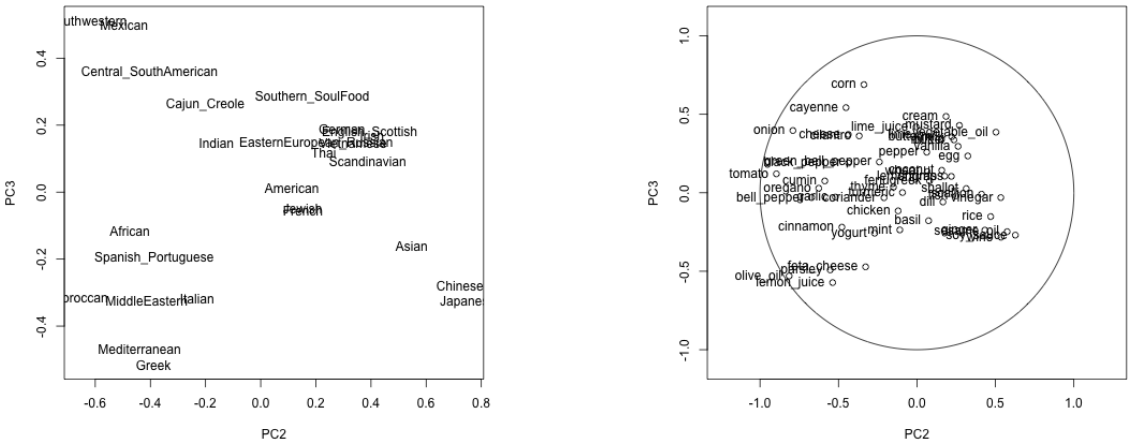

(a) Plan 1,2 (b) Cercle des corrélations

FIGURE 2 – Plan factoriel 1,2 avec son cercle associé




(a) Plan 1,3 (b) Cercle des corrélations

FIGURE 3 – Plan factoriel 1,3 avec son cercle associé




(a) Plan 2,3

(b) Cercle des corrélations

FIGURE 4 – Plan factoriel 2,3 avec son cercle associé

Dans le premier plan factoriel, nous observons que plusieurs groupes se forment. Les pays asiatiques se distinguent avec des ingrédients comme l'huile de sésame, l'huile de soja, le poisson, les échalottes... Un deuxième groupe se distingue : la cuisine occidentale avec des ingrédients comme le blé, les oeufs et le lait. Aussi, les pays consommant de la tomate et de l'huile d'olive sont exposés dans ce premier plan : Maroc, Mexique et Afrique.

Dans le deuxième plan factoriel, la Grèce et la méditerranée sont liées par leur consommation de féta, d'huile d'olive et de jus de citron. Ce plan permet aussi de voir la formation de deux sous groupes parmi les pays asiatiques. Le Vietnam et la Thaïlande consomment du poulet et du piment, ce qui est moins le cas de la Chine et du Japon.

Enfin le troisième plan factoriel met en évidence des pays consommant du maïs comme le Mexique.

## Question 3

Nous obtenons le dendrogramme de la figure 5.

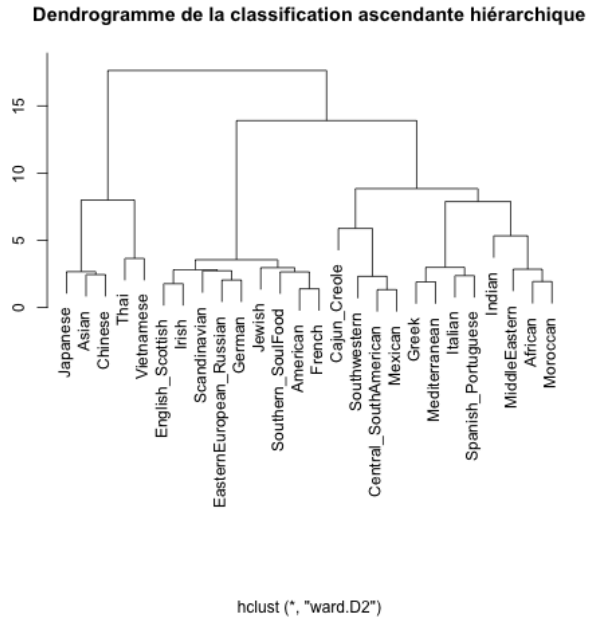



FIGURE 5 – Dendrogramme de la classification ascendante hiérarchique

Ce dendrogramme confirme plusieurs observations que nous avions déjà indiquées dans la deuxième question :

— les cuisines des pays asiatiques sont ensemble avec deux sous groupes : Japon-Chine, Thailande-Vietnam.

— les recettes de deux pays occidentaux, les Etats-Unis et la France, sont ensemble

— la cuisine méditerranée est regroupée avec les pays suivants : Grèce, Italie, Espagne, Portugal.

D'autres groupes que nous n'avions pas identifiés sont présents :

— les recettes des pays de l'europe de l'est avec l'Allemagne, la Russie et les pays de l'est.

— la cuisine des régions dans la partie inférieure au Texas : les Cadiens (présents au Texas et en Louisiane), Mexique, Amérique centrale, Amérique du sud.

— la cuisine d'Afrique et du Moyen-Orient (Afrique, Moyen-Orient et Maroc).

A un niveau plus macro nous avons 3 groupes :

— l'asie avec un indice de 7.5

— l'occident avec l'europe de l'est : un indice de 4 ce qui traduit une utilisation relativement homogène des ingrédients

— Amérique centrale et du sud, méditerranée, l'Inde, l'Afrique, avec un indice de 8 qui traduit une certaine hétérogénéité par rapport à l'occident par exemple

## Question 4

Nous commençons par chercher quel serait un b̈onïombre de classes à choisir :

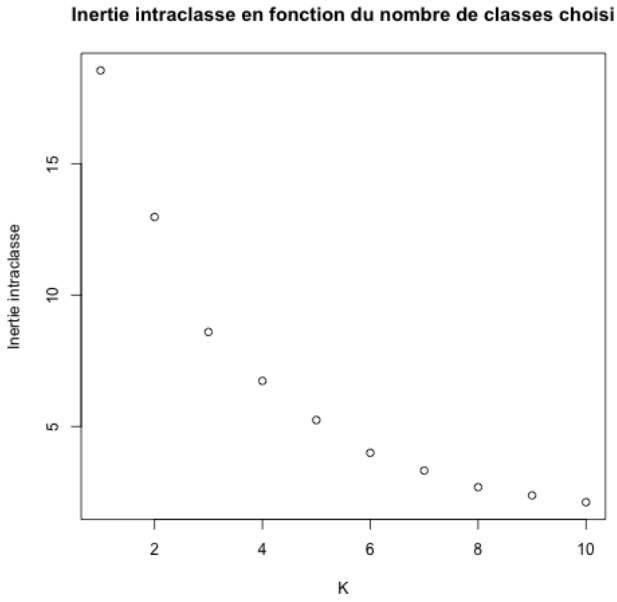



FIGURE 6 – Inertie intraclasse en fonction du nombre de classes choisi

En utilisant la règle du coude un nombre de classes est 3. Ce nombre est en adéquation avec la classification hiérarchique ascendante que nous avions obtenue dans la question précédente. En effet nous avions déjà identifié 3 groupes.

Nous obtenons la classification suivante :

| | |
|---|---|
| 1 | African, Cajun_Creole, Central_SouthAmerican, Greek, Indian, Italian, Mediterranean, Mexican, MiddleEastern, Moroccan, Southwestern, Spanish_Portuguese |
| 2 | Asian, Chinese, Japanese, Thai, Vietnamese |
| 3 | American, EasternEuropean_Russian, English_Scottish, French, German, Irish, Jewish, Scandinavian, Southern_SoulFood |

Cette classification est identique aux trois groupes que nous avions identifiés précedemment avec la CAH.

## Question 5

## Question 6

Le jeu de données qui est fourni contient un échantillon avec 2000 recettes (2000 lignes). Il y a 26 régions (les mêmes que dans le jeu de données précédent) et 51 variables. Les valeurs sont binaires, 0 pour l'absence et 1 pour la présence d'un ingrédient. Il n'y a aucune valeur manquante.

## Question 7

Nous transformons les données en tableau individu-variable : les individus sont les ingrédients et les variables sont les recettes.

Pour calculer la matrice de similarité/dissimilarité nous n'utilisons pas les distances euclidienne et manhattan. En effet, si nous prenons deux individus :
0 0 0 0 0 1 0
0 0 1 0 0 0 0
la distance euclidienne est $\sqrt{2}$ ce qui semble faible alors que les deux ingrédients n'ont jamais été utilisés ensemble. La distance euclidienne dépend de la fréquence d'utilisation de deux ingrédients. Deux ingrédients peu utilisés sont proches au sens euclidien car ils sont peu utilisés. La distance euclidienne ne mesure pas tellement le degré d'utilisation conjointe.

Une meilleure mesure de dissimilarité consiste à ne compter seulement les couples (un couple par variable) où il y a au moins un 1.

Nous notons :
— $M_{01}$ le nombre de couples (0,1)
— $M_{10}$ le nombre de couples (1,0)
— $M_{11}$ le nombre de couples (1,1)
Une mesure de dissimilarité est :

$$d = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}}$$

Cette mesure est comprise entre 0 et 1.
0 pour deux ingrédients qui sont systématiquement utilisés conjointement. 1 pour deux ingrédients qui ne sont jamais utilisés conjointement.

Cette mesure est implémentée dans R avec la fonction "dist" et l'argument method="binary" et est appelée distance de Jaccard.

## Question 8

En utilisant la distance de Jaccard et le critère de Ward nous obtenons la classification hiérarchique ascendante suivante :

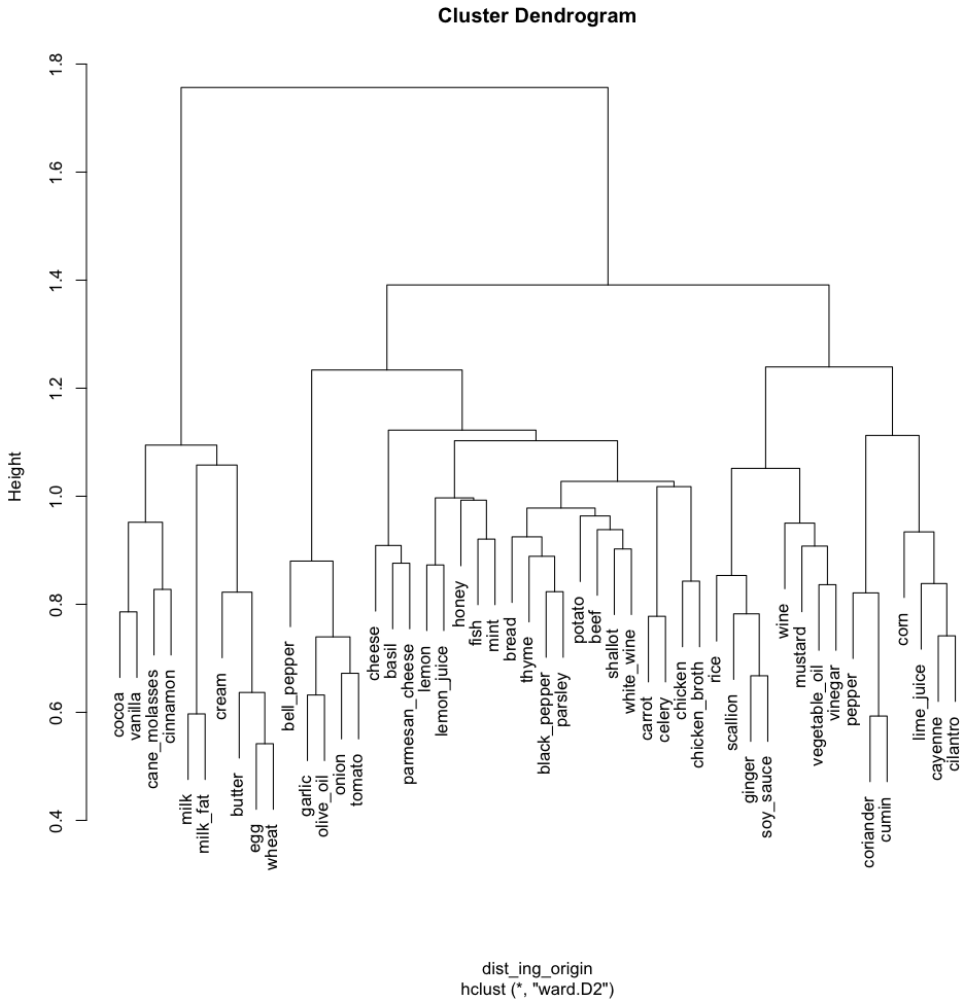



FIGURE 7 – Dendrogramme obtenue avec la distance de Jaccard et le critère de Ward

Trois classes d'ingrédients se distinguent :

— les ingrédients pour les desserts

— les ingédients pour le plat principal

— les ingrédients qui viennent s'ajouter au reste : les condiments, huiles et vin

## Question 9

En appliquant l'algorithme des K-médoïdes, avec K=3, au tableau de dissimilarités établi à la question 7 nous obtenons 3 médoides. Ces médoïdes nous donnent des ingrédients représentant chaque classe. Un ingrédient représentant la première classe est l'oeuf. L'onion représente la deuxième classe. Le gingembre représente la dernière classe.

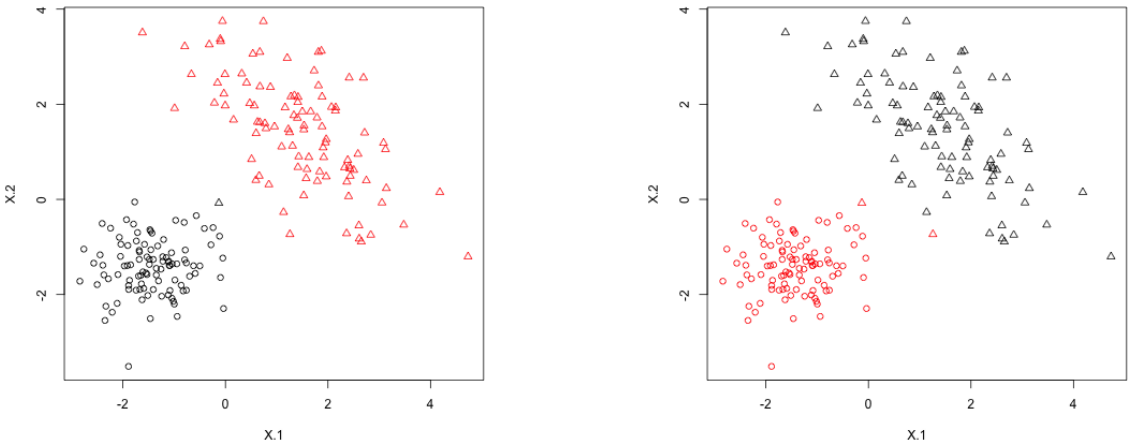
# 2 Classification par K-means avec distance adaptative

## 2.1 Programmation

## 2.2 Application

Nous testons l'algorithme des K-means à distance adpatative à l'aide de jeux de données synthétiques, puis, comparons les résultats avec l'algorithme des K-means classique (utilisant la distance euclidienne).

Jeu de données synthétiques 1 :



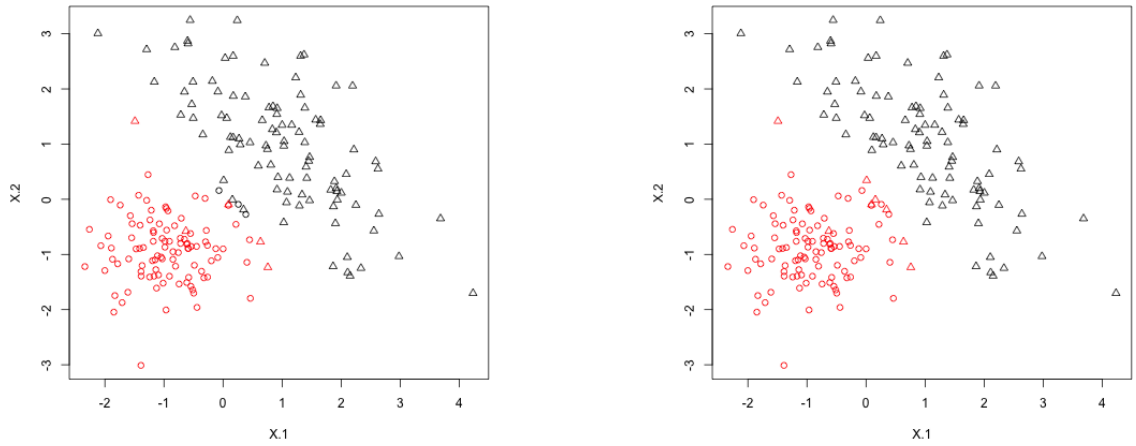

(a) Algorithme des K-means.

(b) Algorithme des K-means à distance adaptative.

FIGURE 8 – Clusterings effectués sur le jeu de données synthétiques 1. Les couleurs représentent les ensembles trouvés à l'aide des K-means, les symboles des points représentent les vraies partitions.

Les deux clusterings obtenus à l'aide des deux méthodes sont très proches. Le Rand index ajusté des K-means "classiques" est de 0.98 et Rand index ajusté des K-means à distance adaptative est de 0.96. Ces deux scores sont très bons puisque le maximum est de 1. Les partitions trouvées sont donc très proches des vraies partitions.

Jeu de données synthétiques 2 :



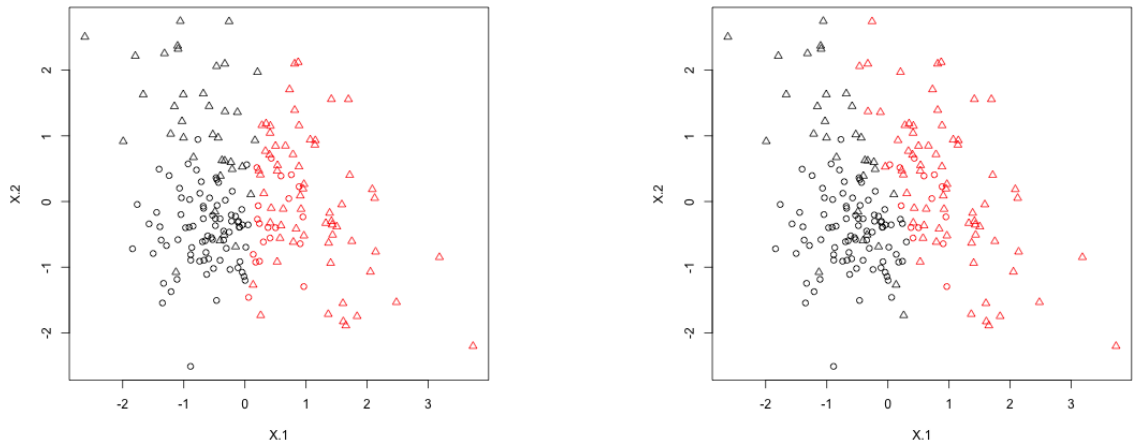

(a) Algorithme des K-means.

(b) Algorithme des K-means à distance adaptative.

FIGURE 9 – Clusterings effectués sur le jeu de données synthétiques 2. Les couleurs représentent les ensembles trouvés à l'aide des K-means, les symboles des points représentent les vraies partitions.

Les deux clusterings obtenus à l'aide des deux méthodes sont très proches. Les Rand index ajustés des K-means sont tous les deux de 0.85. Les partitions trouvées sont donc très proches des vraies partitions.

Jeu de données synthétiques 3 :




(a) Algorithme des K-means.

(b) Algorithme des K-means à distance adaptative.

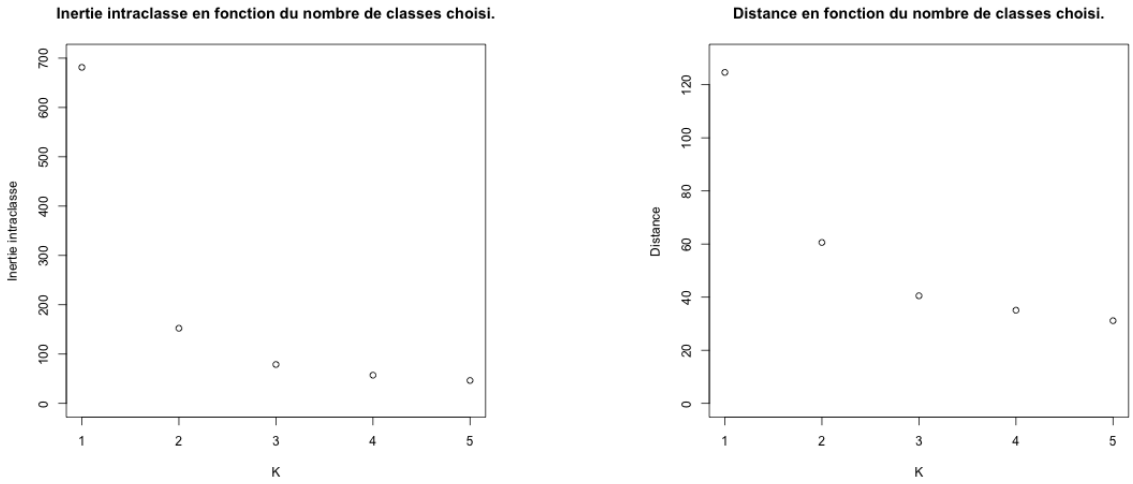
FIGURE 10 – Clusterings effectués sur le jeu de données synthétiques 3. Les couleurs représentent les ensembles trouvés à l'aide des K-means, les symboles des points représentent les vraies partitions.

Les deux clusterings obtenus à l'aide des deux méthodes sont cette fois assez différents. Les deux ensembles qui forment la vraie partition sont collés tout en ayant une dispersion très différente. En effet, l'un a une dispersion faible (inertie par rapport au centre de gravité de 0.17) alors que le second a une dispersion élevée (inertie par rapport au centre de gravité de 0.71). L'algorithme des K-means

"classiques" ne prend pas en compte la dispersion des classes et sépare donc mal les classes. L'algorithme des K-means à distance adaptative prend en compte cette dispersion et sépare donc mieux les données. Le Rand index ajusté des K-means "classiques" est de 0.04 et Rand index ajusté des K-means à distance adaptative est de 0.30. Les partitions trouvées sont donc assez éloignées des vraies partitions mais l'algorithme des K-means à distance adpatative fait bien mieux que les K-means "classiques".

Iris :

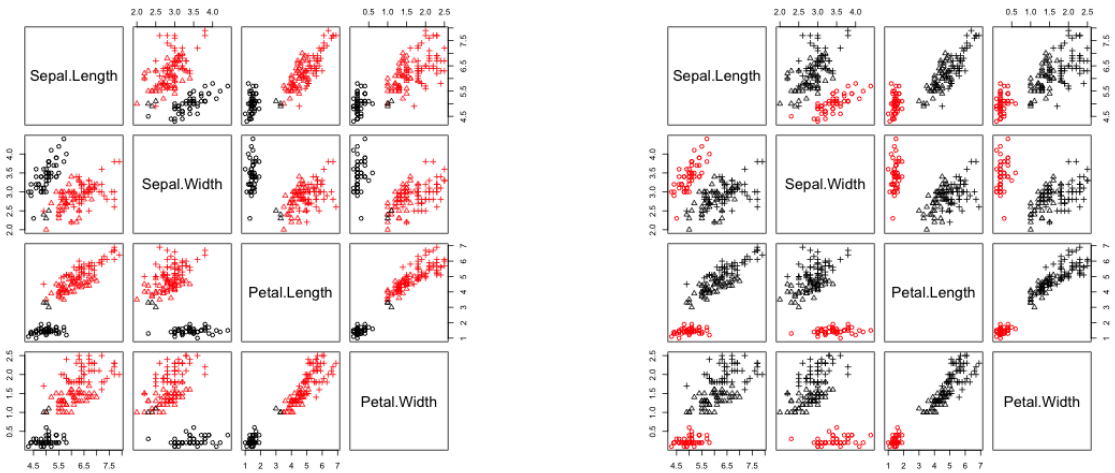
Nous déterminons les classifications avec les deux méthodes pour K=2,...,5. Nous obtenons les valeurs de criteres, en fonction de K, suivant :




(a) Algorithme des K-means. (b) Algorithme des K-means à distance adaptative.

FIGURE 11 – Valeurs des critères en fonction de K sur le jeu de données des Iris.

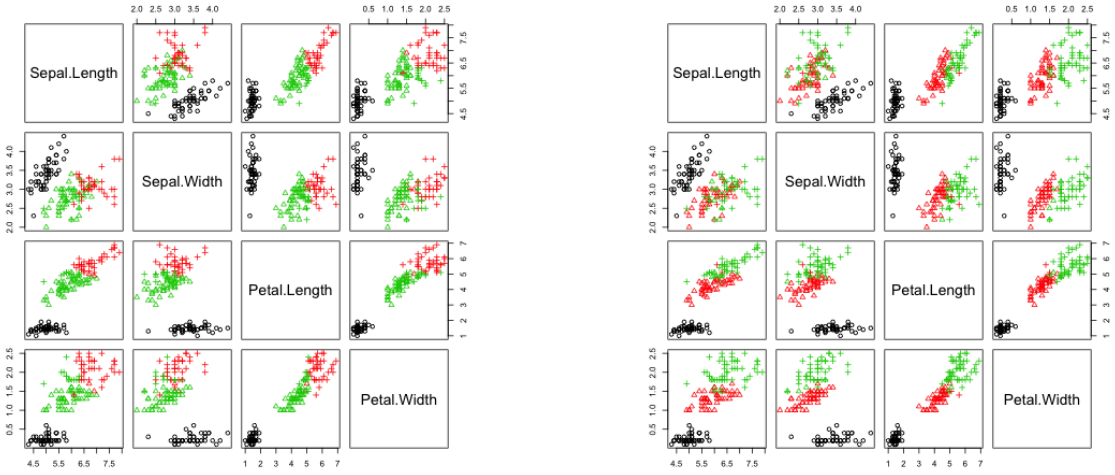
Pour les K-means un bon K est K=2 (en utilisant la méthode du coude). Pour les K-means à distance adaptative un bon K est plutôt 3 (toujours en utilisant la méthode du coude).




(a) Algorithme des K-means. (b) Algorithme des K-means à distance adaptative.

FIGURE 12 – Classification pour K=2 sur le jeu de données des Iris.

Pour K=2 nous obtenons deux classifications très similaires. En effet deux groupes d'individus se détachent très nettement : un premier avec l'espèce Setosa et un deuxième avec les deux autres espèces. La distance adaptative n'apporte donc rien à ce problème de classification pour K=2. Les deux index de Rand ajustés sont proches : 0.54 pour les K-means classiques et 0.57 pour les K-means à distance adaptative. Les index ne sont pas élevés puisque nous avons mis K=2 alors que le jeu comporte en réalité 3 classes.

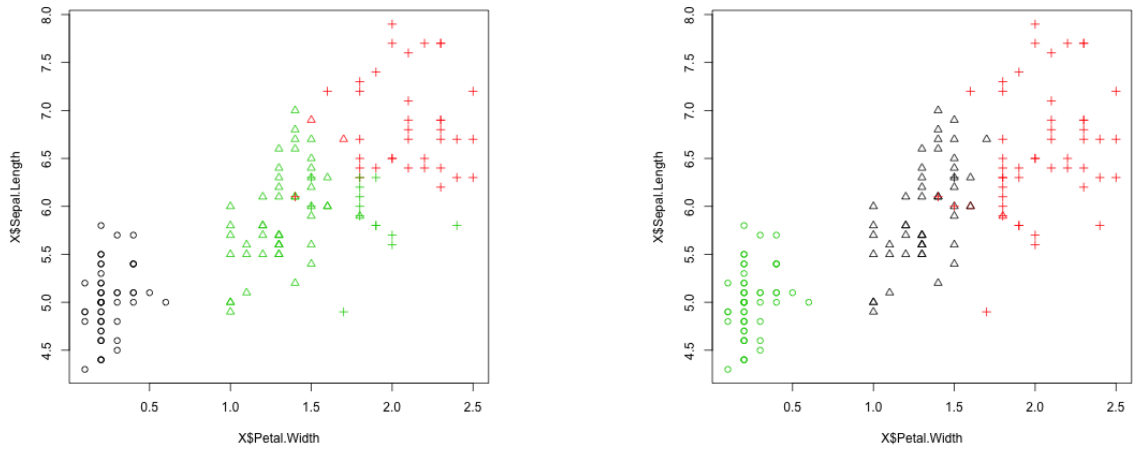



(a) Algorithme des K-means.

(b) Algorithme des K-means à distance adaptative.

FIGURE 13 – Classification pour K=3 sur le jeu de données des Iris.

Pour K=3, les deux classifications sont assez différentes. En effet, dans les deux cas, nous retrouvons un premier groupe avec l'espèce Setosa mais les deux autres groupes sont formés différemment. L'algorithme des K-means à distance adaptative sépare les données de manière très proche par rapport aux classes exactes. En effet nous avons un index de Rand ajusté de 0.87 alors que pour l'algorithme des k-means classique l'index de Rand est de 0.73. Cela s'explique par le fait les K-means à distance adaptative s'adaptent beaucoup mieux aux différentes dispersions des données. Les individus de l'espèce Virginica sont bien plus dispersés que ceux de l'espèce Versicolor. Cela est visible sur la figure 14.




(a) Algorithme des K-means.

(b) Algorithme des K-means à distance adaptative.

FIGURE 14 – Classification pour K=3 sur le jeu de données des Iris.(rond : Setosa, triangle : Versicolor, plus : Virginica).

Spam :

Nous commençons par traiter les données en utilisant une ACP. 99.8% de la variance totale est contenue dans les 2 premiers axes de l'ACP. Nous réduisons donc l'espace en un espace de dimension 2.

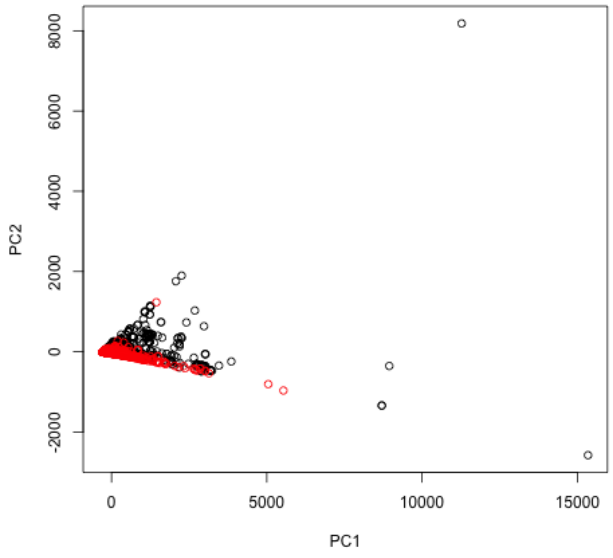



FIGURE 15 – Ensemble des données après ACP. Les couleurs représentent les 2 classes.

Nous pouvons observer sur la figure 15 que les spams et non-spams se superposent pour les valeurs négatives du deuxième axe. Cette superposition montre qu'il est difficile de trouver une séparation entre les spams et non-spams avec un algorithme non supervisé puisque les algorithmes non supervisés

fonctionnent par recherche de similarité.

En utilisant les K-means à distance adaptative implémenté précédemment nous trouvons la classification visible sur la figure 16. Malheureusement, le résultat n'est pas très bon puisque l'index de Rand ajusté est de 0.05.

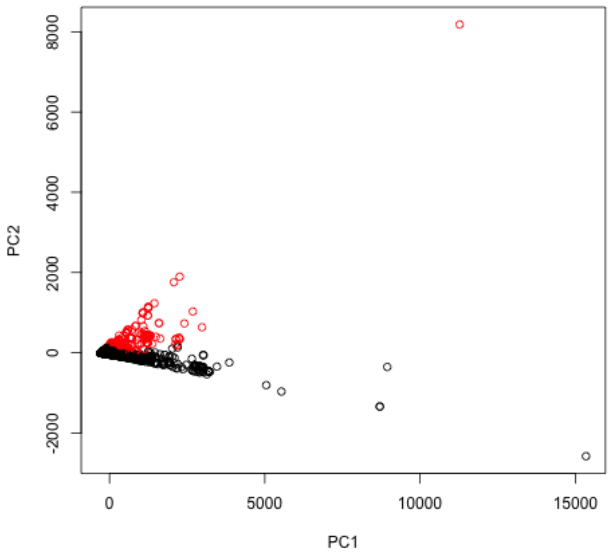



FIGURE 16 – Classification en utilisant les k-means à distance adaptative. Les couleurs représentent les 2 classes.

## 2.3 Justification

### Question 10

$$J(\{v_k, M_k\}_{k=1,...,K}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \left[ (x_i - v_k)^\top M_k (x_i - v_k) \right] \tag{1}$$

Soit $k \in \{1, ..., K\}$

$$\frac{\partial J(\{v_k, M_k\})}{\partial v_k} = -\sum_{i=1}^{n} z_{ik} (M_k + {M_k}^\top)(x_i - v_k) = -(M_k + {M_k}^\top) \sum_{i=1}^{n} z_{ik} (x_i - v_k) \tag{2}$$

Nous cherchons quand

$$\frac{\partial J(\{v_k, M_k\})}{\partial v_k} = 0 \tag{3}$$

Comme $M_k$ est quelconque, il faut que :

$$\sum_{i=1}^{n} z_{ik}(x_i - v_k) = 0 \tag{4}$$

En supposant $n_k \neq 0$ :

$$\implies n_k\overline{x_k} - \sum_{i \in P_k} v_k = 0 \; avec \; \overline{x_k} = \frac{1}{n_k}\sum_{i=1}^{n} z_{ik}x_i \tag{5}$$

$$\implies n_k\overline{x_k} - n_k v_k = 0 \tag{6}$$

$$\implies v_k = \overline{x_k} = \frac{1}{n_k}\sum_{i=1}^{n} z_{ik}x_i \tag{7}$$

**Question 11**

$$J(\{v_k, M_k\}_{k=1,...,K}) = \sum_{k=1}^{K}\sum_{i=1}^{n} z_{ik}\big[(x_i - v_k)^\top M_k(x_i - v_k)\big] \tag{8}$$

Soit $k \in \{1, ..., K\}$

$$\frac{\partial L(\{v_k, M_k, \lambda_k\})}{\partial M_k} = \frac{\partial J(\{v_k, M_k\})}{\partial M_k} - \frac{\partial\big[\sum_{q=1}^{K}\lambda_q(detM_q - \rho_q)\big]}{\partial M_k} \tag{9}$$

$$\frac{\partial J(\{v_k, M_k\})}{\partial M_k} = \sum_{i=1}^{n} z_{ik}\frac{\partial\big[(x_i - v_k)^\top M_k(x_i - v_k)\big]}{M_k} = \sum_{i=1}^{n} z_{ik}(x_i - v_k)(x_i - v_k)^\top \tag{10}$$

$$\frac{\partial\big[\sum_{q=1}^{K}\lambda_q(detM_q - \rho_q)\big]}{\partial M_k} = \lambda_k\frac{\partial detM_k}{\partial M_k} = \lambda_k detM_k({M_k}^{-1})^\top \tag{11}$$

donc,

$$\frac{\partial L(\{v_k, M_k, \lambda_k\})}{\partial M_k} = \sum_{i=1}^{n} z_{ik}(x_i - v_k)(x_i - v_k)^\top - \lambda_k detM_k({M_k}^{-1})^\top \tag{12}$$

Nous cherchons quand

$$\frac{\partial L(\{v_k, M_k, \lambda_k\})}{\partial M_k} = 0 \tag{13}$$

donc,

$$\lambda_k det M_k ({M_k}^{-1})^\top = \sum_{i=1}^{n} z_{ik}(x_i - v_k)(x_i - v_k)^\top \tag{14}$$

en supposant que $detM_k \neq 0$ et $\lambda_k \neq 0$

$$\iff ({M_k}^{-1})^\top = \frac{1}{\lambda_k det M_k} \sum_{i=1}^{n} z_{ik}(x_i - v_k)(x_i - v_k)^\top \tag{15}$$

en appliquant la transposée de chaque côté de l'équation et $(A+B)^\top = A^\top + B^\top$ et $(AB)^\top = B^\top A^\top$

$$\iff {M_k}^{-1} = \frac{1}{\lambda_k det M_k} \sum_{i=1}^{n} z_{ik}(x_i - v_k)(x_i - v_k)^\top \tag{16}$$

$$\iff {M_k}^{-1} = \frac{n_k}{\lambda_k det M_k} V_k \; avec \; V_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik}(x_i - v_k)(x_i - v_k)^\top \tag{17}$$

De plus, comme $J$ ne dépend pas de $\lambda_k$ :

$$\frac{\partial L(\{v_k, M_k, \lambda_k\})}{\partial \lambda_k} = detM_k - \rho_k \tag{18}$$

Nous cherchons quand,

$$\frac{\partial L(\{v_k, M_k, \lambda_k\})}{\partial \lambda_k} = 0 \tag{19}$$

$$\iff detM_k = \rho_k \tag{20}$$

Donc si $\rho_k \neq 0$ alors $detM_k \neq 0$, ce qui montre que l'hypothèse faite pour obtenir l'équation 15 est réaliste.

En combinant les équations 17 et 20, nous obtenons :

$${M_k}^{-1} = \frac{n_k}{\lambda_k \rho_k} V_k \; avec \; V_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik}(x_i - v_k)(x_i - v_k)^\top \tag{21}$$

Donc, (en supposant $n_k \neq 0$) :

$$\implies V_k = \frac{\lambda_k \rho_k}{n_k} {M_k}^{-1} \tag{22}$$

$$\implies det(V_k) = det(\frac{\lambda_k \rho_k}{n_k} {M_k}^{-1}) \tag{23}$$

En utilisant la multilinéarité du déterminant et le fait que $V_k$ soit une matrice carrée de taille $p$ :

$$\implies det(V_k) = \frac{{\lambda_k}^p {\rho_k}^p}{{n_k}^p} det({M_k}^{-1}) \tag{24}$$

De plus $det({M_k}^{-1}) = \frac{1}{det(M_k)} = \frac{1}{\rho_k}$

$$\implies \rho_k det(V_k) = \frac{{\lambda_k}^p {\rho_k}^p}{{n_k}^p} \tag{25}$$

$$\implies \frac{{n_k}^p}{{\lambda_k}^p} = \frac{{\rho_k}^p}{\rho_k det(V_k)} \tag{26}$$

$$\implies \frac{n_k}{\lambda_k \rho_k} = \frac{1}{{\rho_k}^{\frac{1}{p}} det(V_k)^{\frac{1}{p}}} \tag{27}$$

En combinant les équations 21 et 27 nous obtenons :

$${M_k}^{-1} = \frac{1}{{\rho_k}^{\frac{1}{p}} det(V_k)^{\frac{1}{p}}} V_k \ avec \ V_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} (x_i - v_k)(x_i - v_k)^\top \tag{28}$$

$$\implies {M_k}^{-1} = (\rho_k det(V_k))^{-\frac{1}{p}} V_k \ avec \ V_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} (x_i - v_k)(x_i - v_k)^\top \tag{29}$$