Analyse de données

Arnaud Poinas

Année 2024/2025

Table des matières

Informations sur l'UE

Modalités du cours :

- 13 séances de 2H de cours/TD
- 10 séances de 2H de TP en R
- 2 séances de 2H de TP d'introduction à SAS

Évaluation :

- 2 compte-rendus de TP
- 1 examen

Note finale $= \frac{1}{4}\text{CR}_1 + \frac{1}{4}\text{CR}_2 + \frac{1}{2}\text{Exam}$.

⚠ Il n'y a pas de seconde session.

Introduction

Définition 1

*On appelle **jeu de données** de n individus et p variables un tableau de taille $n \times p$:*

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}.$$

*Les lignes $e_i = (x_{i,1}, \cdots, x_{i,p})$ sont appelées les **individus** (ou les **observations**) et les colonnes $v_j = \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{pmatrix}$ sont appelées les **variables**.*

On distingue deux types de variable :

- Si $\forall i,\ x_{i,j} \in \mathbb{R}$, la variable v_j est dite **quantitative**.
- Si $\forall i,\ x_{i,j} \in \Omega$, où Ω est un ensemble fini de taille $N \in \mathbb{N}$, la variable v_j est dite **qualitative** (ou **facteur**) à N **modalités**.

Exemples :

- Le jeu de données "iris" de R contient 150 individus et 5 variables. Il y a 4 variables quantitatives et 1 variable qualitative à 3 modalités.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.10	3.50	1.40	0.20	setosa
2	4.90	3.00	1.40	0.20	setosa
3	4.70	3.20	1.30	0.20	setosa
⋮	⋮	⋮	⋮	⋮	⋮
150	5.90	3.00	5.10	1.80	virginica

- Un jeu de données qu'on utilisera plusieurs fois comme exemple est les données démographiques des 40 communes de Grand Poitiers. Il contient 22 variables quantitatives donnant la proportion d'individu de divers catégories d'âges et de sexe pour chaque commune, 2 variables quantitatives donnant le taux de natalité et mortalité de chaque commune, 1 variable quantitative donnant la proportion d'étudiant de chaque commune et 1 variable qualitative à trois modalités indiquant si la commune possède moins de 1000 habitants, entre 1000 et 5000 habitants ou plus de 5000 habitants.

On s'intéressera durant ce cours à faire de la **statistique exploratoire**, c'est à dire à chercher à résumé le jeu de données et voir s'il y a des comportements particuliers qui apparaissent. Le cours se découpera en deux parties :

- **Réduction de dimension :** Comment réduire le nombre de variables d'un jeu de données en perdant le moins d'information possible ? En particulier, comment transformer un jeu de données à $p \geqslant 3$ variables quelconques en un jeu de données à 2 variables quantitatives afin de pouvoir le visualiser par un nuage de point sur le plan ?
- **Classification :** Comment regrouper les individus en plusieurs groupes de sorte que deux individus d'un même groupe possèdent des comportements similaires et deux individus de groupes différents possèdent des comportements différents.

Chapitre 1

Réduction de dimension

I Centrage et réduction

⚠ On suppose dans cette section que l'on possède un jeu de données de n individus et p variables qui sont toutes quantitatives.

Définition 2

Pour une variable quantitative v_j, on note sa ***moyenne***

$$\bar{v}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$$

et on note le vecteur des moyennes

$$g = \begin{pmatrix} \bar{v}_1 \\ \vdots \\ \bar{v}_p \end{pmatrix} = \frac{1}{n}{}^t X 1_n \in \mathbb{R}^p$$

où $1_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ *est un vecteur colonne de taille n contenant que des 1.*

Remarque: g est le centre de gravité du nuage de point des e_i.

Définition 3

On appelle ***données centrées*** *la matrice*

$$Y = \begin{pmatrix} x_{1,1} - \bar{v}_1 & \cdots & x_{1,p} - \bar{v}_p \\ \vdots & \ddots & \vdots \\ x_{n,1} - \bar{v}_1 & \cdots & x_{n,p} - \bar{v}_p \end{pmatrix} = X - 1_n{}^t g.$$

Remarque: Le centre de gravité des données centrées est $0_p = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$, l'origine.

Définition 4

On définit la ***covariance*** *entre la variable v_i et v_j par*

$$\mathrm{cov}(v_i, v_j) = \frac{1}{n}\sum_{k=1}^{n}(x_{i,k} - \bar{v}_i)(x_{j,k} - \bar{v}_j) = \frac{1}{n}\langle v_i - \bar{v}_i 1_n, v_j - \bar{v}_j 1_n\rangle$$

et on note $\mathrm{var}(v_i) = \mathrm{cov}(v_i, v_i)$ *pour la* ***variance*** *de la variable v_i. On appelle* ***matrice de covariance*** *la matrice*

$$C = \begin{pmatrix} \mathrm{cov}(v_1, v_1) & \cdots & \mathrm{cov}(v_1, v_p) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(v_p, v_1) & \cdots & \mathrm{cov}(v_p, v_p) \end{pmatrix} = \frac{1}{n}{}^tYY.$$

⚠ Afin de distinguer la covariance entre des variables aléatoires et la covariance entre les variables d'un jeux de données on utilisera la notation cov avec un c minuscule pour désigner la covariance entre les variables d'un jeu de données. On fera la même chose pour la variance et la corrélation.

Définition 5

On appelle ***données centrées normalisées*** *la quantité*

$$Z = \begin{pmatrix} \frac{x_{1,1}-\bar{v}_1}{\sqrt{\mathrm{var}(v_1)}} & \cdots & \frac{x_{1,p}-\bar{v}_p}{\sqrt{\mathrm{var}(v_p)}} \\ \vdots & \ddots & \vdots \\ \frac{x_{n,1}-\bar{v}_1}{\sqrt{\mathrm{var}(v_1)}} & \cdots & \frac{x_{n,p}-\bar{v}_p}{\sqrt{\mathrm{var}(v_1)}} \end{pmatrix} = Y \begin{pmatrix} \frac{1}{\sqrt{\mathrm{var}(v_1)}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\mathrm{var}(v_p)}} \end{pmatrix}.$$

Définition 6

On définit la ***corrélation*** *entre la variable v_i et v_j par*

$$\mathrm{corr}(v_i, v_j) = \frac{\mathrm{cov}(v_i, v_j)}{\sqrt{\mathrm{var}(v_i)\mathrm{var}(v_j)}} \in [-1, 1].$$

On appelle ***matrice de corrélation*** *la matrice*

$$R := \begin{pmatrix} \mathrm{corr}(v_1, v_1) & \cdots & \mathrm{corr}(v_1, v_p) \\ \vdots & \ddots & \vdots \\ \mathrm{corr}(v_p, v_1) & \cdots & \mathrm{corr}(v_p, v_p) \end{pmatrix}.$$

Remarque: La matrice de corrélation est égale à la matrice de covariance des données centrées normalisées. C'est à dire que $R = \frac{1}{n}{}^tZZ$.

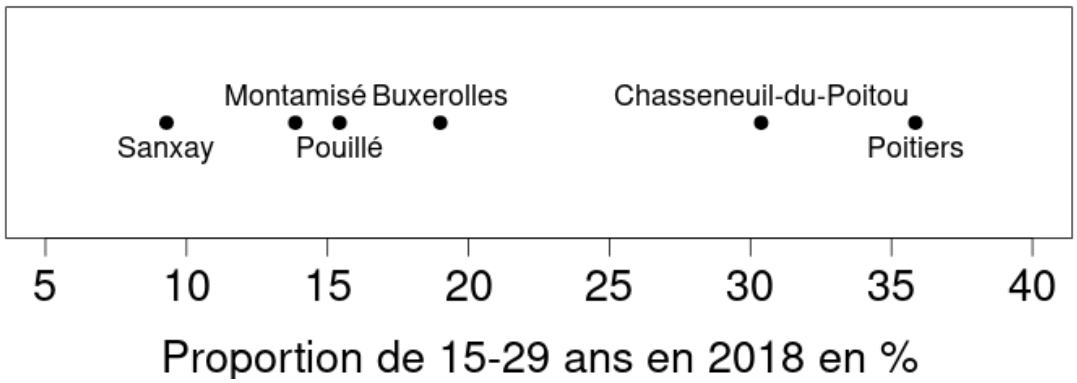
II Analyse en composantes principales

⚠ On suppose toujours que l'on possède un jeu de données de n individus et p variables qui sont toutes quantitatives.

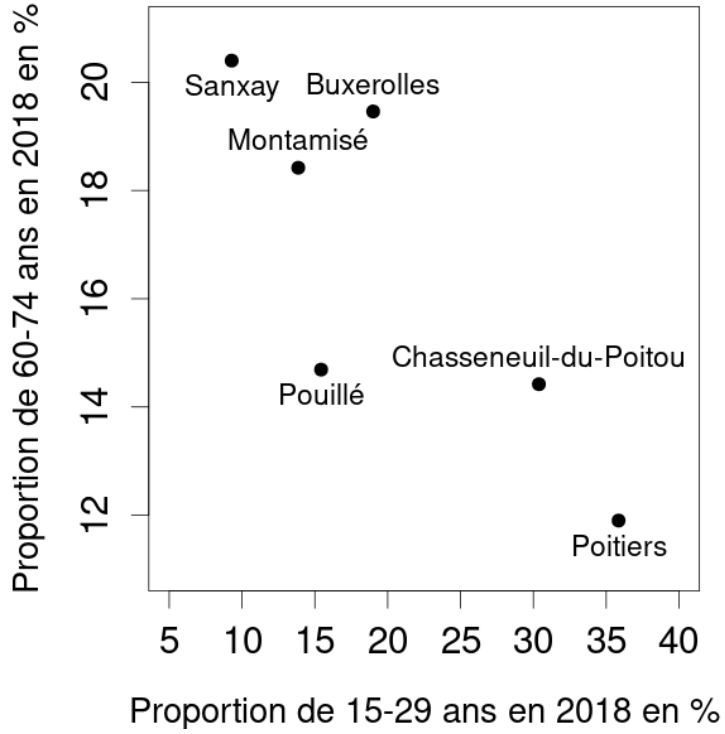
Représentation graphique d'un jeu de données

Chaque individu e_i correspond à un point $(x_{i,1}, \cdots, x_{i,p}) \in \mathbb{R}^p$. On peut donc représenter l'ensemble des individus par un nuage de n points en dimension p.

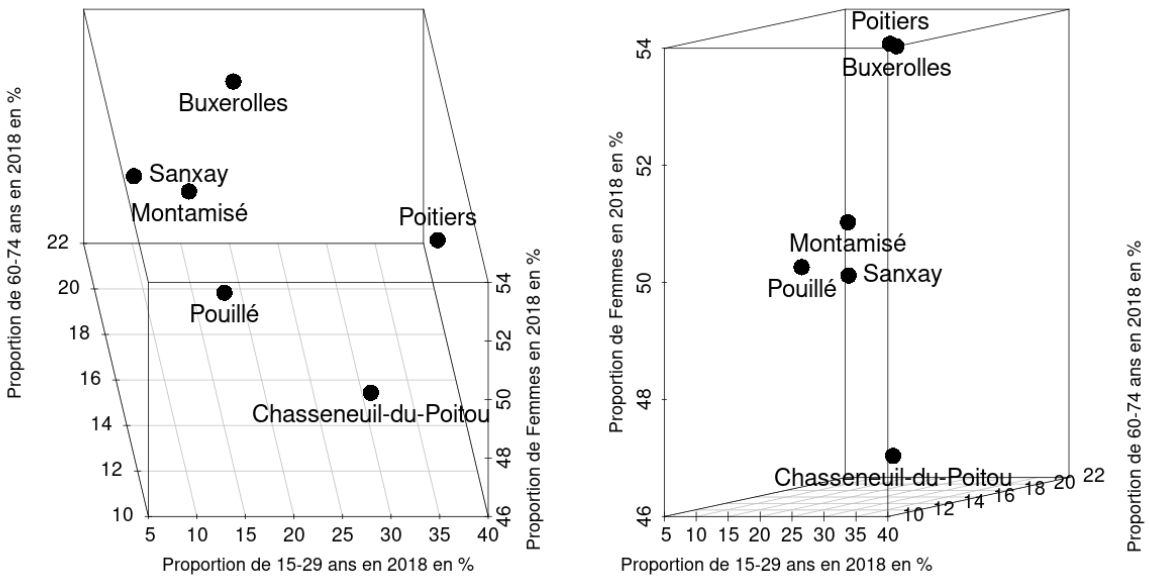
- Si $p = 1$ on peut représenter les données sur une droite.



- Si $p = 2$ on peut représenter les données dans le plan.



- Si $p = 3$ on peut représenter les données dans l'espace mais on est obligé de choisir un angle de vue.



Il vaut mieux choisir un angle de vue de sorte que les données soient les plus écartées possible afin de pouvoir observer au mieux les différences entre les individus. Par exemple, sur la figure de droite Poitiers et Buxerolles ont l'air d'avoir des valeurs similaires mais c'est juste à cause de l'angle de vue. On voit bien sur la figure de gauche que ce n'est pas le cas. On cherche donc un angle de vue (ce qui revient à projeter les données sur un espace de dimension 2) de sorte que les distances entre les individus sont les plus grandes possibles. Un tel principe peut se généraliser pour des données en dimension quelconque. On appelle ça l'**analyse en composante principale** ou **ACP**.

1 Choix de l'espace de projection

Les distances entre les points étant invariantes par translation, on commence toujours par centrer le jeu de données. En général on réduit aussi les données afin que toutes les variables aient le même poids dans le calcul des distances. On considère alors que la matrice des données X de taille $n \times p$ est centrée et réduite.

Remarque: Il est possible de ne pas normaliser les variables. On parle alors d'ACP non-normalisé.

Définition 7

On appelle ***inertie totale*** *du nuage de points* $\{e_1, \cdots, e_n\}$*, notée* I_{tot}*, la quantité*

$$I_{tot} = \frac{1}{n} \sum_{i=1}^{n} \|e_i - g\|_2^2.$$

avec $g = 0_n$ *pour des données centrées.*

Remarques:

- Si les données sont centrées alors $g = 0_n$ ce qui simplifie l'expression.
- Plus l'inertie est grande, plus les données sont écartées entre elles.
- Si on n'a qu'une seule variable ($p = 1$) alors l'inertie correspond à la variance de cette variable.

Proposition 8

$$I_{tot} = \sum_{j=1}^{p} \text{var}(v_j).$$

En particulier, pour des données centrées et réduites on a $I_{tot} = p$.

Démonstration :

$$I_{tot} = \frac{1}{n}\sum_{i=1}^{n} \|e_i - g\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{i,j} - g_j)^2 = \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - g_j)^2.$$

Or, comme g_j est la moyenne des $x_{i,j}$ alors $\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - g_j)^2$ est la variance de v_j d'où le résultat. ■

On considère maintenant un sous espace vectoriel H de $\mathbb{R}^p$ de dimension k et on note $P_H(e_i)$ la projection orthogonale de l'individu $e_i = (x_{i,1}, \cdots, x_{i,k})$ sur H.
Soit

$$I_H = \frac{1}{n}\sum_{i=1}^{n} \|P_H(e_i)\|_2^2,$$

l'inertie des projections des individus sur H. Le principe de l'ACP revient à chercher l'espace de dimension k qui maximise I_H.

Cas de la dimension 1 :
H est engendré par un vecteur unitaire u et la projection orthogonale sur H s'écrit $P_H(x) = \langle x, u\rangle u$. On cherche donc un vecteur unitaire u qui maximise

$$I_H = \frac{1}{n}\sum_{i=1}^{n} \|\langle e_i, u\rangle u\|^2 = \frac{1}{n}\sum_{i=1}^{n}\langle e_i, u\rangle^2.$$

Comme les e_i sont les lignes de X alors

$$Xu = \begin{pmatrix} \langle e_1, u\rangle \\ \vdots \\ \langle e_n, u\rangle \end{pmatrix}$$

d'où

$$I_H = \frac{1}{n}\|Xu\|_2^2 = \frac{1}{n}\langle Xu, Xu\rangle = \frac{1}{n}{}^t u\, {}^t X X u = {}^t u R u = \langle u, Ru\rangle,$$

où R est la matrice de corrélation des données. R est une matrice symétrique définie positive donc elle possède p valeurs propres réelles positives $0 \leqslant \lambda_p \leqslant \cdots \leqslant \lambda_1$ et on a $I_H = \langle u, Ru\rangle \leqslant \lambda_1\|u\|_2^2 = \lambda_1$ avec cas d'égalité lorsque u est un vecteur propre associé à λ_1. Ce résultat illustre le théorème plus général suivant que l'on admettra.

Théorème 9

Soient $\lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant 0$ les valeurs propres de la matrice de corrélation des données $R = \frac{1}{n}{}^t X X$ et $(u_1, \cdots, u_p)$ une base orthonormale de vecteurs propres associés. Alors, l'espace $H_k = Vect(u_1, \cdots, u_k)$ maximise I_H parmi tous les espaces de dimension k avec

$$I_{H_K} = \lambda_1 + \cdots + \lambda_k.$$

Remarques:

- On a $H_1 \subset H_2 \subset \cdots \subset H_p$. On dit que les solutions sont emboîtées.
- On a $H_{k+1} = H_k \oplus \text{Vect}(u_{k+1})$ avec $H_k \perp \text{Vect}(u_{k+1})$. On passe alors de l'espace H_k à l'espace H_{k+1} en rajoutant un axe orthogonal à H_k.

Définition 10

*Les vecteurs $u_1, \cdots, u_p$ sont appelés les **axes principaux** de l'ACP.*

Proposition 11

Le jeu de données projeté sur l'espace H_k s'écrit XU dans la base $(u_1, \cdots, u_k)$ où $U = (u_1|\cdots|u_k)$ est la matrice de taille $p \times k$ dont les colonnes sont les u_i.

Démonstration : On rappelle que la projection d'un point x sur la base orthonormée $(u_1, \cdots, u_k)$ s'écrit

$$P_H(x) = \langle x, u_1\rangle u_1 + \cdots + \langle x, u_k\rangle u_k$$

et donc les coordonnées de $P_H(x)$ dans cette base s'écrivent $(\langle x, u_1\rangle, \cdots, \langle x, u_k\rangle)$. Du coup, le jeu de données projeté va s'écrire

$$\begin{pmatrix} \langle e_1, u_1\rangle & \cdots & \langle e_1, u_k\rangle \\ \vdots & \ddots & \vdots \\ \langle e_n, u_1\rangle & \cdots & \langle e_n, u_k\rangle \end{pmatrix} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} (u_1|\cdots|u_k) = XU.$$

■

Définition 12

Soit $i \in \{1, \cdots, p\}$. La i-ème variable du jeu de données projeté sur H_k, quel que soit $k \geqslant i$, est tout le temps la même et est égale à

$$c_i = \begin{pmatrix} \langle e_1, u_i\rangle \\ \vdots \\ \langle e_n, u_i\rangle \end{pmatrix} = Xu_i.$$

*Les vecteurs $c_1, \cdots, c_p$ sont appelés les **composantes principales**.*

Proposition 13

$(c_1, \cdots, c_p)$ est une base orthogonale des vecteurs propres de la matrice $\frac{1}{n}X^tX$ et la valeur propre associée à c_i est λ_i.

Démonstration : Par définition de c_i, on a $\frac{1}{n}X^tXc_i = \frac{1}{n}X^tXXu_i$. Or, comme u_i est un vecteur propre de $\frac{1}{n}{}^tXX$ associé à la valeur propre λ_i on a $\frac{1}{n}{}^tXXu_i = \lambda_i$ et donc

$$\frac{1}{n}X^tXc_i = \lambda_i Xu_i = \lambda_i c_i.$$

Soient $i, j \in \{1, \cdots, p\}$, on a

$$\langle c_i, c_j\rangle = \langle Xu_i, Xu_j\rangle = \langle {}^tXXu_i, u_j\rangle = n\lambda_i\langle u_i, u_j\rangle = \begin{cases} 0 \text{ si } i \neq j, \\ n\lambda_i \text{ si } i = j, \end{cases}$$

en conséquence de l'orthonormalité des u_i. Cela montre que les c_i sont orthogonaux mais pas forcement orthonormés. ■

On peut alors résumer l'ACP par le tableau suivant :

Nuage de points	Lignes de X
Espace du nuage	$\mathbb{R}^p$
Matrice de poids	$\frac{1}{n} I_n$
Matrice de métrique	I_p
Axes principaux	Vecteurs propres u_i de $\frac{1}{n}{}^t XX$ vérifiant $\|u_i\|_2^2 = 1$ et $\langle u_i, u_j \rangle = 0$ si $i \neq j$
Comp. principales	Vecteurs propres c_i de $\frac{1}{n} X^t X$ vérifiant $\|c_i\|_2^2 = n\lambda_i$ et $\langle c_i, c_j \rangle = 0$ si $i \neq j$

La signification de ce qu'est la matrice de poids et la matrice de métrique sera clarifiée dans la section III.

Exemple: Si on applique l'ACP à deux dimensions sur les variables de proportion d'individus par catégories d'âges et de sexe alors on obtient le nuage de point projeté suivant.

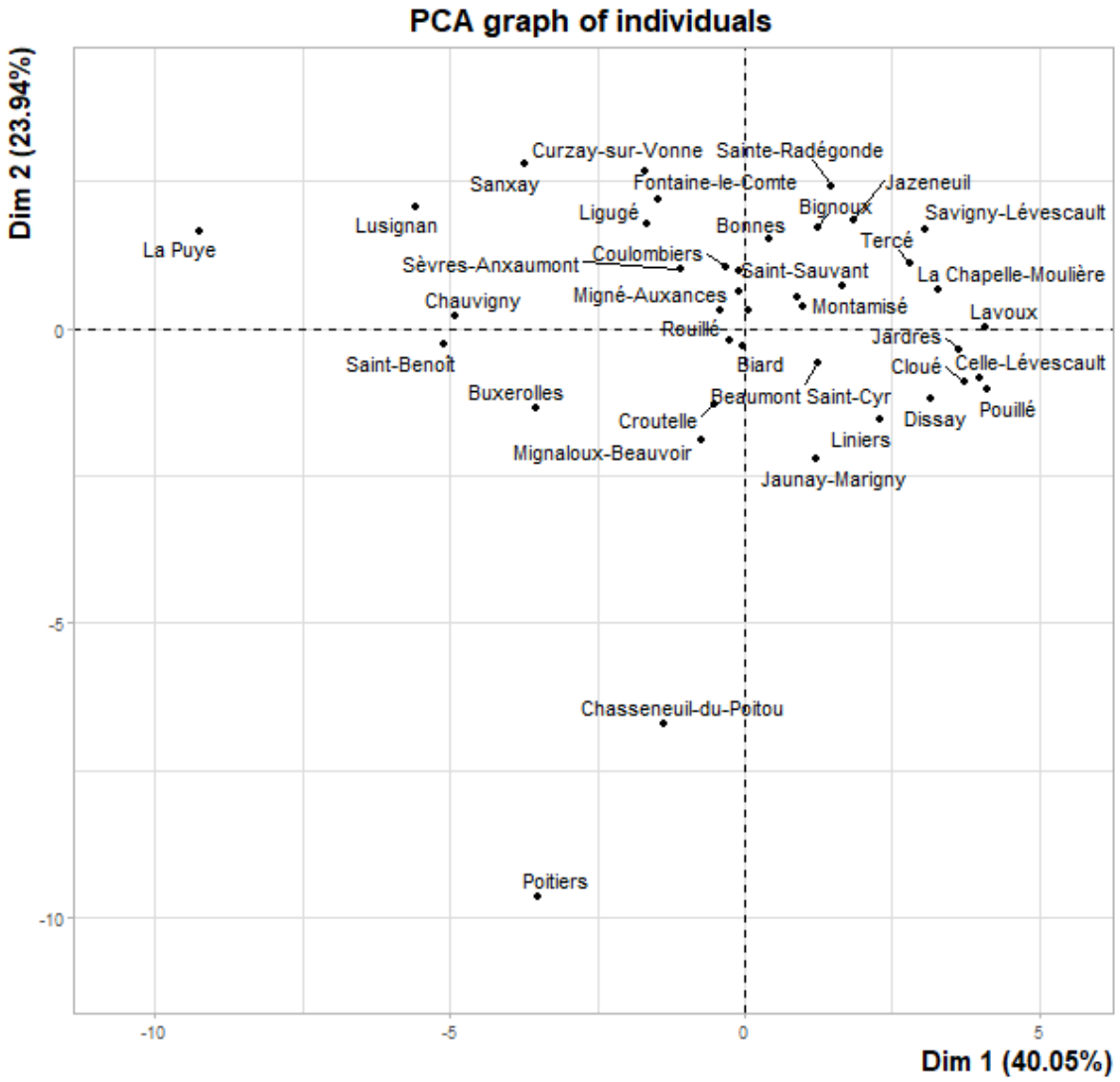


FIGURE 1.1 – Nuage de points des individus après projection sur l'espace H_2 des données de démographie des 40 communes de la communauté urbaine de Grand Poitiers.

2 Interprétation des résultats

a Interprétation pour les individus

On remarque que $I_{H_k} \leqslant I_{tot} = p$ pour tout k. C'est normal puisqu'une projection contracte les distances. En particulier, $I_{H_k} = I_{tot}$ si et seulement si toutes les données sont dans H_k et la projection ne les modifie pas. Le résultat suivant nous donne le nombre de dimension dont on a besoin pour avoir une représentation parfaite des données.

Proposition 14

Le plus petit k tel que $I_{H_k} = I_{tot}$ correspond à la dimension de l'espace vectoriel engendré par les variables $v_1, \cdots, v_p$ du jeu de données. En particulier, $k \leqslant \min(n, p)$.

Démonstration : Comme $I_{H_k} = \lambda_1 + \cdots + \lambda_k$ alors $I_{H_k} = I_{tot}$ si et seulement si $\lambda_{k+1} = \cdots = \lambda_p = 0$. Le plus petit k tel que $I_{H_k} = I_{tot}$ est donc le rang de la matrice $\frac{1}{n}{}^tXX$ qui est donc égal au rang de X et correspond alors à la dimension de $\mathrm{Vect}(v_1, \cdots, v_p)$ et est donc inférieur à $\min(n, p)$. ■

De façon générale, plus I_{H_k} est proche de I_{tot} moins la projection modifie les données et plus l'ACP est de bonne qualité. Cela motive la définition suivante

Définition 15

*On appelle **pourcentage d'inertie** pour l'espace H_k la quantité*

$$\frac{I_{H_k}}{I_{tot}} = \frac{\lambda_1 + \cdots + \lambda_k}{p} \in [0, 1].$$

*En particulier, la quantité $\frac{\lambda_k}{p}$ est appelée le **pourcentage d'inertie apporté par la k-ième dimension**.*

Exemple: On donne dans le tableau suivant les valeurs propres et pourcentages d'inerties pour chaque dimension de l'ACP sur les données démographiques des communes de grand Poitiers.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14 à Dim.22
Valeur propre	8.81	5.27	2.55	1.91	1.31	0.67	0.45	0.42	0.19	0.16	0.13	0.11	0.03	0.00
% d'inertie	40.05	23.94	11.60	8.66	5.94	3.06	2.04	1.90	0.86	0.71	0.60	0.49	0.15	0.00
% cumulé d'inertie	40.05	63.98	75.58	84.25	90.18	93.25	95.29	97.19	98.05	98.76	99.36	99.85	100.00	100.00

Une autre façon de visualiser ces valeurs propres est par un diagramme en barre appelé **scree plot**.

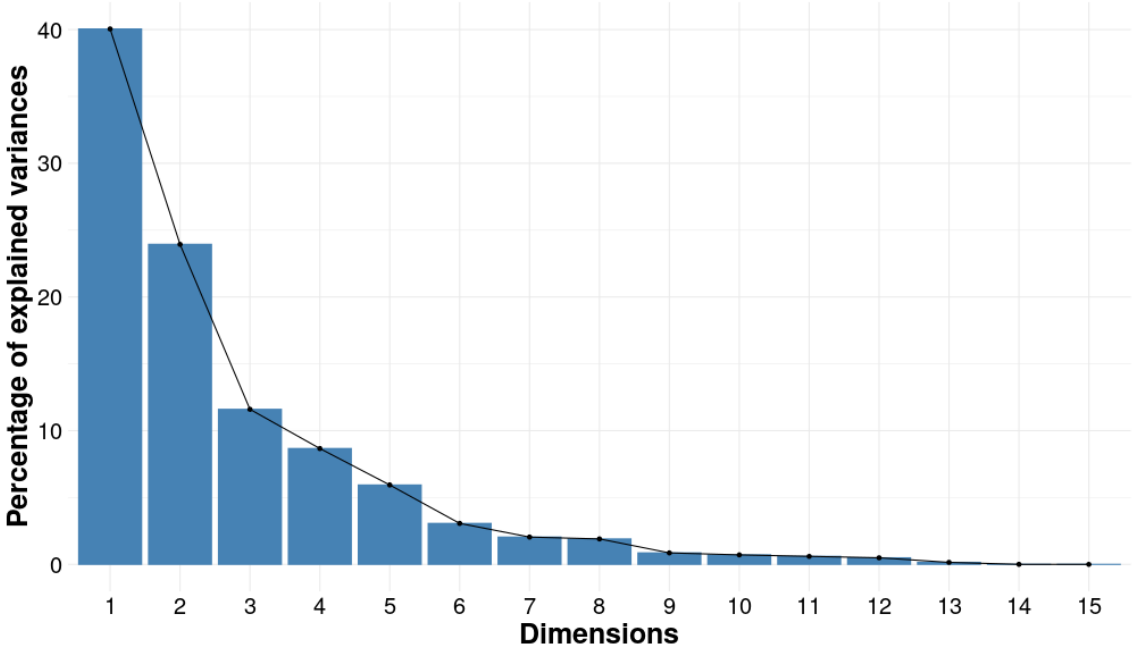


FIGURE 1.2 – Scree plot de l'ACP sur les données démographiques des communes de grand Poitiers.

Le pourcentage d'inertie donne une idée générale de la qualité d'une ACP mais ne dit rien sur la qualité de la représentation d'un individu donné. Pour un individu e_i, on souhaite que sa projection $P_{H_k}(e_i)$ soit la plus proche possible du point initial e_i. Cela motive la définition suivante.

Définition 16

On définit la ***qualité de représentation*** *de l'individu i par la quantité*

$$\frac{\text{Inertie de la projection de } e_i \text{ sur } H_k}{\text{Inertie de } e_i} = \frac{\|P_{H_k}(e_i)\|^2}{\|e_i\|^2} = \cos^2(\theta_{i,k}) \in [0,1],$$

où $\theta_{i,k}$ est l'angle entre $\overrightarrow{Oe_i}$ et H_K.

Plus $\cos^2(\theta_{i,k})$ est proche de 1, plus l'individu i est bien représenté par l'ACP. En particulier, on peut décomposer $\frac{\|P_{H_k}(e_i)\|^2}{\|e_i\|^2}$ en utilisant la base orthonormale de H_k formée par les axes principaux $u_1, \cdots, u_k$. Par le théorème de Pythagore, on a

$$\frac{\|P_{H_k}(e_i)\|^2}{\|e_i\|^2} = \sum_{j=1}^{k} \frac{\langle u_j, e_i\rangle^2}{\|e_i\|^2} = \sum_{j=1}^{k} \cos^2(\theta'_{i,j})$$

où $\theta'_{i,j}$ est l'angle entre u_j et e_i. La quantité $\cos^2(\theta'_{i,j})$ représente le **pourcentage d'inertie de l'individu e_i expliqué par la j-ème composante principale**.

Une dernière quantité intéressante à regarder est la contribution des individus à la construction de chaque axe :

Définition 17

On définit la ***contribution de l'individu i à l'axe*** *k par la quantité*

$$\text{contr}_{i,k} = \frac{\|P_{Vect(u_k)}(e_i)\|^2}{n\lambda_k} = \frac{\text{Inertie de } e_i \text{ projeté sur } u_k}{\text{Inertie totale des indivus projetés sur } u_k} \in [0,1].$$

Remarque: Pour tout k on a $\sum_{i=1}^{n} \text{contr}_{i,k} = 1$. La contribution est donc souvent représentée par un pourcentage.

En général, il n'est pas forcement souhaitable d'avoir un individu ayant une contribution excessive car cela peut créer des problèmes d'instabilité. Le fait de retirer un tel individu changerait énormément les résultats de l'analyse.

Exemple:

	$\cos^2$					Contribution				
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Beaumont Saint-Cyr	0.49	0.11	0.01	0.00	0.30	0.44%	0.16%	0.02%	0.02%	1.81%
Béruges	0.09	0.03	0.04	0.17	0.14	0.23%	0.13%	0.32%	2.05%	2.39%
Biard	0.00	0.01	0.00	0.46	0.01	0.00%	0.04%	0.00%	3.49%	0.14%
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Vouneuil-sous-Biard	0.00	0.09	0.00	0.33	0.19	0.00%	0.05%	0.00%	0.50%	0.43%

TABLE 1.1 – $\cos^2$ et contribution pour chaque dimension de l'ACP sur les données de démographie des 40 communes de Grand Poitiers. La contribution est représentée par un pourcentage.

b Interprétation pour les variables

De la même façon que pour les individus on peut aussi définir la qualité de représentation des variables.

Définition 18

Soient $v_1, \cdots, v_p$ les variables centrées réduites du jeu de données et $c_1, \cdots, c_p$ les composantes principales. On appelle ***qualité de représentation de la variable j par l'axe s*** *la quantité*

$$\frac{\text{Inertie de la projection de } v_j \text{ sur } c_s}{\text{Inertie de } v_j} = \frac{\|P_{Vect(c_s)}(v_j)\|^2}{\|v_j\|^2} = \cos^2(\theta_{j,s}) \in [0,1],$$

où $\theta_{j,s}$ est l'angle entre v_j et c_s. La ***qualité de la représentation de la variable j par l'ACP à k dimensions*** *est donc*

$$\frac{\|P_{Vect(c_1,\cdots,c_k)}(v_j)\|^2}{\|v_j\|^2} = \sum_{s=1}^{k} \cos^2(\theta_{j,s}).$$

Afin de pouvoir visualiser ces quantités ainsi que l'influence des variables du jeu de données sur les composantes principale des données projeté on va utiliser le résultat suivant liant les propriétés géométriques et statistiques des vecteurs formant les variables.

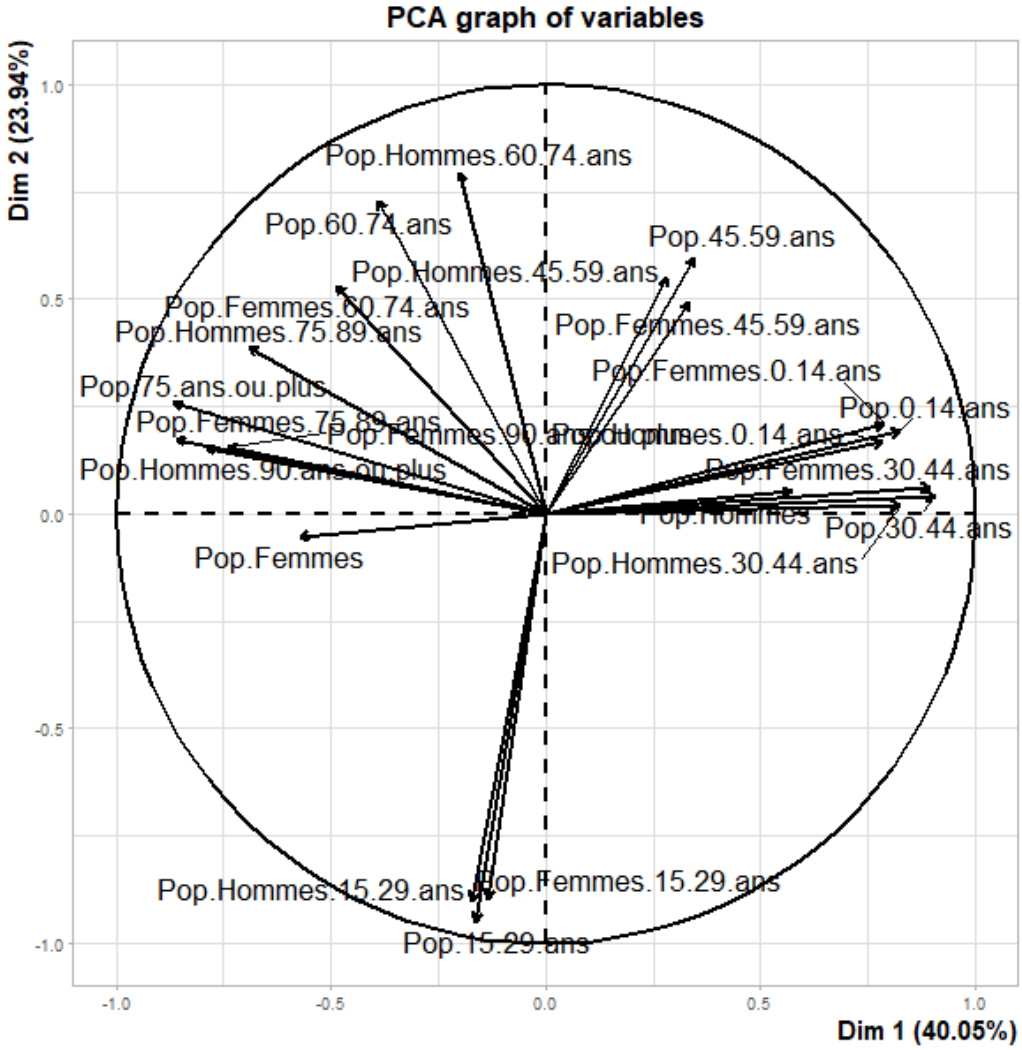
Proposition 19

Soient $v_i = \begin{pmatrix} x_{1,i} \\ \vdots \\ x_{n,i} \end{pmatrix}$ *et* $v_j = \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{pmatrix}$ *deux variables centrées. Alors,*

$$\frac{1}{n}\langle v_i, v_j \rangle = \frac{1}{n}\sum_{k=1}^{n} x_{k,i}x_{k,j} = \text{cov}(v_i, v_j).$$

En particulier, $\frac{1}{n}\|v_i\|^2 = \text{var}(v_i)$ et $\text{corr}(v_i, v_j) = \frac{\langle v_i, v_j \rangle}{\|v_i\|\|v_j\|} = \cos(\theta_{i,j})$ où $\theta_{i,j}$ est l'angle entre v_i et v_j.

Pour des variables centrées ont peut donc lire leurs variances en regardant leurs normes et on peut lire la corrélation entre deux variables en regardant le cosinus de leur angle. C'est ces astuces que l'on va utilisé pour visualiser l'influence des variables du jeu de données sur les composantes principale des données projeté. Pour une ACP à deux dimensions, on va représenter les variables par leur projection sur l'espace engendré par c_1 et c_2 divisée par $\|v_j\| = \sqrt{n}$. C'est ce qu'on appelle le **cercle de corrélation**.
Exemple: Le cercle de corrélation de l'ACP à deux dimensions sur les variables de proportion d'individus par catégories d'âges et de sexe est le suivant.



A AJOUTER COMM

On observe que les variables fortement corrélées positivement au premier axe sont les proportions d'individus de moins de 14 ans et des individus entre 30 et 44 ans alors que les variables fortement corrélées négativement au premier axe sont les proportions d'individus de 75 ans et plus. On peut donc en déduire que le premier axe sépare les communes avec une population jeune (qui seront plutôt sur la droite du nuage des individus) des communes avec une population âgée (qui seront plutôt sur la gauche du nuage des individus). On observe aussi que les variables fortement corrélées négativement au premier axe sont les proportions d'individus entre 15 et 29 ans. Du coup, le deuxième axe va surtout séparer les communes avec une forte population d'adolescent et jeunes adultes (qui seront plutôt en bas du nuage des individus) des communes avec une faible population d'adolescent et jeunes adultes (qui seront plutôt en haut du nuage des individus).

3 Représentation de variables et d'individus supplémentaires

Définition 20

Pour un jeu de données, on appelle ***variable supplémentaire*** *une variable que l'on n'inclue pas dans le calcul des composantes principales. Les variables utilisées dans le calcul des composantes principales sont appelées les* ***variables actives****. De la même façon, on appelle* ***individus actifs*** *les individus participants au calcul des composantes principales et* ***individus supplémentaires*** *les autres.*

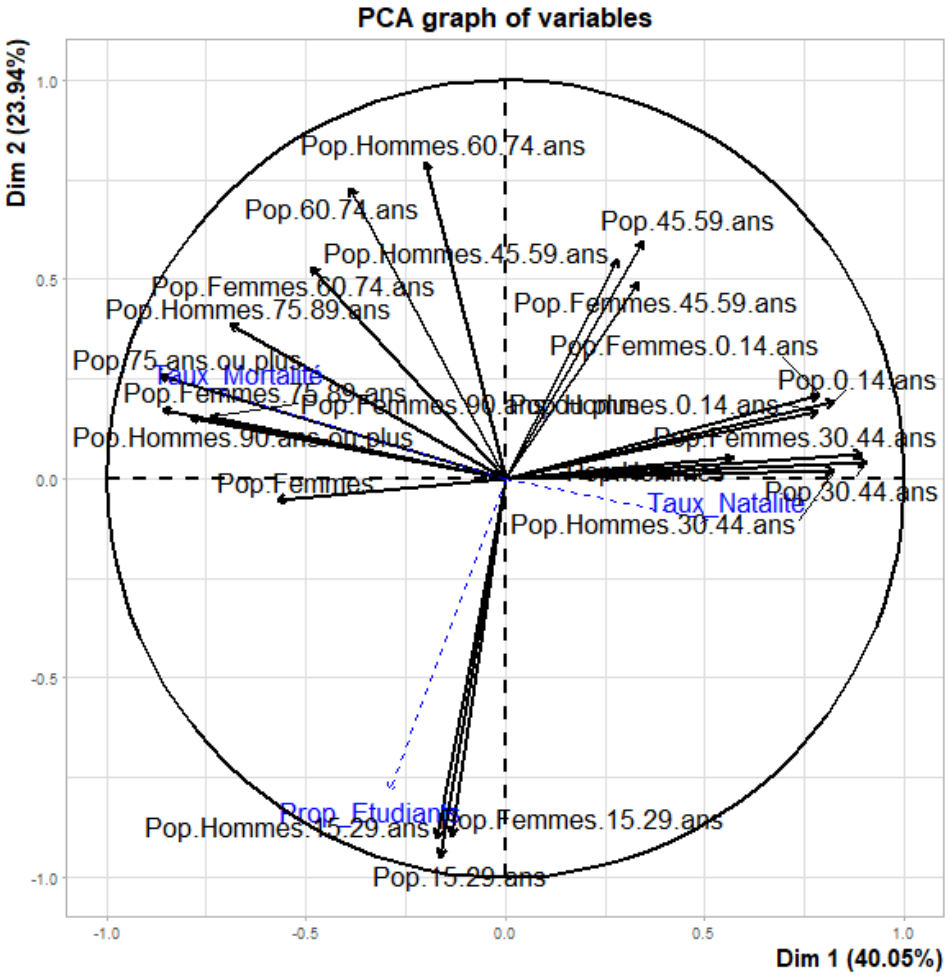
Remarques:

- Les variables supplémentaires sont souvent des variables dont on cherche à observer le comportement en fonction des variables actives.
- Les individus supplémentaires sont souvent des individus contenant des informations redondantes ou possédant des valeurs suspectes

Visualisation :

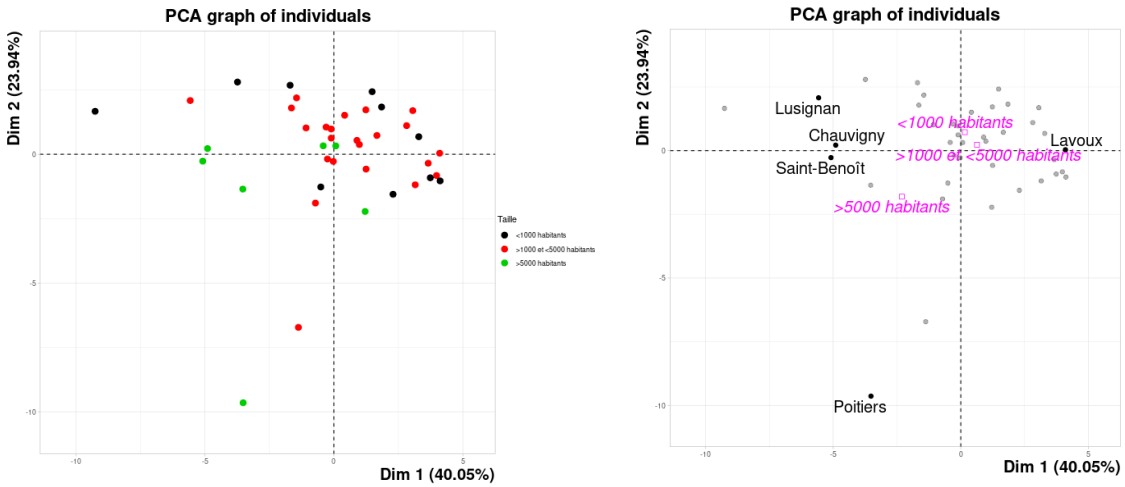
- On peut calculer les coordonnées d'un individu supplémentaire e_{n+1} **en centrant et réduisant ses variables en utilisant les moyennes et variances calculées sur les données actives** puis en projetant sur l'espace H_K.
- On représente une variable supplémentaire **quantitative** dans le cercle de corrélation en calculant sa projection sur les deux premières composantes principales et on interprète le résultat de la même façon que les variables actives.

  Exemple: Si on rajoute les variables de taux de natalité/mortalité et de proportion d'étudiants comme variable quantitative supplémentaire à l'ACP sur les données démographiques de Grand Poitiers on obtient le cercle de corrélation suivant.



On observe notamment que la proportion d'étudiant est fortement corrélée négativement au deuxième axe. Cela précise donc que cet axe sépare bien les communes avec beaucoup d'étudiant des communes avec peu d'étudiants.

- On représente une variable supplémentaire **qualitative** dans le nuage de point des individus soit en coloriant chaque point en fonction de sa modalité soit en indiquant le centre de gravité des points de chaque modalité.
  Exemple: On regarde la variable de taille des communes comme variable qualitative supplémentaire.



On observe que les communes de plus de 5000 habitants sont plutôt en bas à gauche du

nuage des individus. C'est donc plutôt des communes avec une population âgée et une forte proportion d'étudiant comparé aux autres.

III Analyse factorielle des correspondances

Dans cette section on considère un jeu de données de n individus contenant deux variables qualitatives : une variable à I modalités et une variable à J modalités. Comparé à l'ACP, on ne souhaite pas visualiser le comportement des individus mais on souhaite représenter visuellement les interactions entre les modalités des deux variables qualitatives.

Exemple: L'exemple fil rouge que nous utiliserons dans cette partie est des données de 717 observations de comportement des dauphins au large du port de Galveston en fonction des bateaux présents.
(Source : Piwetz S (2019) *Common bottlenose dolphin (Tursiops truncatus) behavior in an active narrow seaport.* PLOS ONE **14**(2))

	Type de bateau	Comportement du dauphin
1	Moyen	Recherche Nourriture
2	Moyen	Recherche Nourriture
3	Moyen	Recherche Nourriture
4	Moyen	Repos
5	Moyen	Repos
⋮	⋮	⋮
717	Tourisme	Social

Le jeu de données contient une variable qualitative **Type de bateau** possédant 5 modalités : *Aucun*, *Grand*, *Moyen*, *Pêcheur*, *Petit*, *Tourisme* et une variable qualitative **Comportement du dauphin** possédant 4 modalités : *Déplacement*, *Recherche de nourriture*, *Repos* et *Social*.

1 Table de contingence

Définition 21

On note $n_{i,j}$ le nombre d'individus prenant la i-ème modalité de la première variable et la j-ème modalité de la deuxième variable. La matrice de taille $I \times J$ définie par

$$N = \begin{pmatrix} n_{1,1} & \cdots & n_{1,J} \\ \vdots & \ddots & \vdots \\ n_{I,1} & \cdots & n_{I,J} \end{pmatrix}$$

est appelée la ***table de contingence*** *des données.*

Exemple:

		Comportement du dauphin			
		Recherche Nourriture	Repos	Social	Déplacement
Type de bateau	Grand	10	0	16	0
	Moyen	22	8	32	0
	Aucun	18	4	62	1
	Petit	28	17	115	8
	Tourisme	34	9	26	2
	Pêcheur	302	0	3	0

TABLE 1.2 – Table de contingence des données de comportement de dauphin.

Définition 22

On note $n_{i,\bullet}$ le nombre d'individus prenant la i-ème modalité de la première variable et $n_{\bullet,j}$ le nombre d'individus prenant la j-ème modalité de la deuxième variable. Les $n_{i,\bullet}$ et $n_{\bullet,j}$ sont appelés les ***effectifs marginaux****.*

Proposition 23

$$n_{i,\bullet} = \sum_{j=1}^{J} n_{i,j},\ n_{\bullet,j} = \sum_{i=1}^{I} n_{i,j} \text{ et } \sum_{i=1}^{I} n_{i,\bullet} = \sum_{j=1}^{J} n_{\bullet,j} = n$$

En particulier, on obtient les $n_{i,\bullet}$ en sommant les colonnes de la table de contingence et on obtient les $n_{\bullet,j}$ en sommant les lignes.

La table de contingence permet de transformer le jeu de données en un tableau de nombre. On peut alors représenter les modalités de la première variable par les lignes du jeu de données et les modalités de la seconde variable par les colonnes. Néanmoins, ce n'est pas une bonne idée comme illustré par l'exemple suivant.

Exemple: On considère le jeu de données fabriqué suivant sur le groupe sanguin de personnes atteint d'une maladie.

	Malade	Pas malade
A	5	10
B	10	20
AB	10	10
O	20	20

Si on représente les modalités de la première variable par les valeurs des lignes alors on a les coordonnées suivantes : A $= (5, 10)$, B $= (10, 20)$, AB $= (10, 10)$, O $= (20, 20)$ ce qui donne des points assez éloignés entre eux. Néanmoins, on voit que les groupes sanguins AB et O influencent les chances d'être malade de la même façon. On s'attendrait alors à ce que les modalités AB et O soient représentés par le même point. Même chose pour les modalités A et B.

Définition 24

On appelle ***matrice des profils lignes*** *et* ***matrice des profils colonnes*** *les matrices de taille $I \times J$ définies par*

$$N_l = \begin{pmatrix} \frac{n_{1,1}}{n_{1,\bullet}} & \cdots & \frac{n_{1,J}}{n_{1,\bullet}} \\ \vdots & \ddots & \vdots \\ \frac{n_{I,1}}{n_{I,\bullet}} & \cdots & \frac{n_{I,J}}{n_{I,\bullet}} \end{pmatrix} \quad et \quad N_c = \begin{pmatrix} \frac{n_{1,1}}{n_{\bullet,1}} & \cdots & \frac{n_{1,J}}{n_{\bullet,J}} \\ \vdots & \ddots & \vdots \\ \frac{n_{I,1}}{n_{\bullet,1}} & \cdots & \frac{n_{I,J}}{n_{\bullet,J}} \end{pmatrix}.$$

A partir des lignes de la matrice N_l on construit un nuage de I points dans $\mathbb{R}^J$ représentant les modalités de la première variable. De même, à partir des colonnes de la matrice N_c on construit un nuage de J points dans $\mathbb{R}^I$ représentant les modalités de la seconde variable. Comme pour l'ACP on note e_i les points du nuage de point étudié.
Exemple: On reprend la table de contingence

	Malade	Pas malade
A	5	10
B	10	20
AB	10	10
O	20	20

La matrice des profils lignes s'écrit alors

$$\begin{pmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \\ 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

Les modalités A et B sont donc représentées par le point $(1/3, 2/3)$ et les modalités AB et O par le point $(1/2, 1/2)$ ce qui est plus cohérent avec notre discussion précédente.

Définition 25

On note D_l et D_c les matrices diagonales des effectifs marginaux des lignes et des colonnes :

$$D_l = \begin{pmatrix} n_{1,\bullet} & & 0 \\ & \ddots & \\ 0 & & n_{I,\bullet} \end{pmatrix} \quad et \quad D_c = \begin{pmatrix} n_{\bullet,1} & & 0 \\ & \ddots & \\ 0 & & n_{\bullet,J} \end{pmatrix}.$$

Remarque: On a $N_l = D_l^{-1} N$ et $N_c = N D_c^{-1}$.

On va maintenant définir quel est le centre de notre nuage de point. Pour l'ACP c'était le centre de gravité qui correspond au "comportement moyen" des individus. Pour l'AFC, le choix du centre correspond au principe suivant.

Définition 26

*On dit qu'on a **indépendance statistique** entre la i-ème modalité de la première variable et la j-modalité de la seconde variable si*

$$\frac{n_{i,j}}{n} = \frac{n_{i,\bullet}}{n}\frac{n_{\bullet,j}}{n} \iff \frac{n_{i,j}}{n_{i,\bullet}} = \frac{n_{\bullet,j}}{n} \iff \frac{n_{i,j}}{n_{\bullet,j}} = \frac{n_{i,\bullet}}{n}.$$

Comme $\frac{n_{i,\bullet}}{n}$ est la proportion des individus possédant la modalité i de la première variable et $\frac{n_{i,j}}{n_{\bullet,j}}$ est la proportion des individus possédant la modalité i de la première variable parmi ceux possédant aussi la modalité j de la seconde variable alors on a l'interprétation suivante.

- Si $\frac{n_{i,j}}{n_{\bullet,j}} = \frac{n_{i,\bullet}}{n}$ alors le fait qu'un individu possède la modalité j de la seconde variable "n'influence pas les chances" qu'il possède aussi la modalité i de la première variable.
- Si $\frac{n_{i,j}}{n_{\bullet,j}} > \frac{n_{i,\bullet}}{n}$ alors le fait qu'un individu possède la modalité j de la seconde variable "augmente les chances" qu'il possède aussi la modalité i de la première variable.
- Si $\frac{n_{i,j}}{n_{\bullet,j}} < \frac{n_{i,\bullet}}{n}$ alors le fait qu'un individu possède la modalité j de la seconde variable "diminue les chances" qu'il possède aussi la modalité i de la première variable.

L'idée de l'AFC est donc de représenter le comportement de la i-ème modalité de la première variable par la façon dont le point

$$\left(\frac{n_{i,1}}{n_{i,\bullet}}, \cdots, \frac{n_{i,J}}{n_{i,\bullet}}\right)$$

s'écarte du point

$$\left(\frac{n_{\bullet,1}}{n}, \cdots, \frac{n_{\bullet,J}}{n}\right)$$

qui sera donc le centre de notre nuage de point. Afin de voir ce point comme un centre de gravité, on va affecté un poids à chacune des modalités.

Définition 27

*Pour N_l, la i-ème modalité (i.e. la i-ème ligne) est affectée du poids $\frac{n_{i,\bullet}}{n}$ et pour N_c, la j-ème modalité (la j-ème colonne) est affectée du poids $\frac{n_{\bullet,j}}{n}$. Les matrices $\frac{1}{n}D_l$ et $\frac{1}{n}D_c$ sont alors appelées les **matrices de poids** des profils-lignes et des profils colonnes.*

Remarque: Dans les deux cas, la somme des poids est égale à 1.

Proposition 28

- *Le centre de gravité des profils lignes pondérés par leur poids est*

$$g = \begin{pmatrix} \frac{n_{\bullet,1}}{n} \\ \vdots \\ \frac{n_{\bullet,J}}{n} \end{pmatrix}$$

- *Le centre de gravité des profils colonnes pondérés par leur poids est*

$$g = \begin{pmatrix} \frac{n_{1,\bullet}}{n} \\ \vdots \\ \frac{n_{J,\bullet}}{n} \end{pmatrix}$$

Démonstration : On fait la preuve juste pour les profils lignes. La j-ème coordonnée de g correspond à la moyenne des j-ème coordonnées des points du profil ligne pondéré par leur poids. Donc

$$g_j = \frac{\sum_{i=1}^{I} \frac{n_{i,\bullet}}{n} \frac{n_{i,j}}{n_{i,\bullet}}}{\sum_{i=1}^{I} \frac{n_{i,\bullet}}{n}} = \sum_{i=1}^{I} \frac{n_{i,j}}{n} = \frac{n_{\bullet,j}}{n}.$$

■

2 La métrique du χ^2

Comparé au cas de l'ACP classique qui cherche la projection qui conserve au mieux les distances usuelle entre les points, pour l'AFC on utilise une distance différente appelée **la distance du χ^2**. On illustre la nécessité d'utiliser une métrique différente avec l'exemple suivant :
Exemple: On reprend la table de contingence

	Malade	Pas malade
A	5	10
B	10	20
AB	10	10
O	20	20

La matrice des profils colonnes s'écrit alors

$$\begin{pmatrix} 1/9 & 1/6 \\ 2/9 & 2/6 \\ 2/9 & 1/6 \\ 4/9 & 2/6 \end{pmatrix}$$

Les modalités "malade" et "pas malade" sont donc représentées par les points $(\frac{1}{9}, \frac{2}{9}, \frac{2}{9}, \frac{4}{9})$ et $(\frac{1}{6}, \frac{2}{6}, \frac{1}{6}, \frac{2}{6})$ dont la distance euclidienne au carré est

$$\left(\frac{1}{9} - \frac{1}{6}\right)^2 + \left(\frac{2}{9} - \frac{2}{6}\right)^2 + \left(\frac{2}{9} - \frac{1}{6}\right)^2 + \left(\frac{4}{9} - \frac{2}{6}\right)^2 = \frac{5}{162} \approx 0.031.$$

Or, on a vu que les modalités A et B ainsi que les modalités AB et O influencent de la même façon les modalités "Malade" et "Pas malade". Donc, si on regroupe ces modalités ensemble on ne devrait pas changer l'écart entre les modalités "malade" et "pas malade". Après regroupement on obtient la table de contingence la table de contingence

	Malade	Pas malade
A ou B	15	30
AB ou O	30	30

ce qui donne $\begin{pmatrix} 1/3 & 1/2 \\ 2/3 & 1/2 \end{pmatrix}$ comme matrice des profils colonnes. Les modalités "malade" et "pas malade" sont donc maintenant représentées par les points $(\frac{1}{3}, \frac{2}{3})$ et $(\frac{1}{2}, \frac{1}{2})$ dont la distance euclidienne au carré est

$$\left(\frac{1}{3} - \frac{1}{2}\right)^2 + \left(\frac{2}{3} - \frac{1}{2}\right)^2 = \frac{1}{18} \approx 0.056$$

ce qui ne correspond pas à la distance que l'on avait avant.

Définition 29

On définit la ***distance du*** χ^2 *entre les profils lignes par*

$$d^2_{\chi^2}(e_i, e_{i'}) = \sum_{j=1}^{J} \frac{n}{n_{\bullet,j}} \left(\frac{n_{i,j}}{n_{i,\bullet}} - \frac{n_{i',j}}{n_{i',\bullet}}\right)^2 = n\langle e_i - e_{i'}, D_c^{-1}(e_i - e_{i'})\rangle$$

et la distance du χ^2 *entre les profils colonnes par*

$$d^2_{\chi^2}(e_j, e_{j'}) = \sum_{i=1}^{I} \frac{n}{n_{i,\bullet}} \left(\frac{n_{i,j}}{n_{\bullet,j}} - \frac{n_{i,j'}}{n_{\bullet,j'}}\right)^2 = n\langle e_j - e_{j'}, D_l^{-1}(e_j - e_{j'})\rangle.$$

Les matrices nD_c^{-1} *et* nD_l^{-1} *sont appelées les* ***matrices de métrique*** *des profils lignes et des profils colonnes.*

Remarques:

- Si $n_{\bullet,1} = \cdots = n_{\bullet,J} = \frac{n}{J}$ alors $nD_c^{-1} = JI_J$ et $d^2_{\chi^2}(e_i, e_{i'}) = J\|e_i - e_{i'}\|_2^2$.

- Comparé à la distance classique, on affecte un poids à chaque variable. Le fait de diviser par $\frac{n}{n_{\bullet,j}}$ permet d'équilibrer l'influence des colonnes sur la distance entre les lignes.

Exemple: On reprend la matrice des profils colonnes

$$\begin{pmatrix} 1/9 & 1/6 \\ 2/9 & 2/6 \\ 2/9 & 1/6 \\ 4/9 & 2/6 \end{pmatrix}$$

et on a que la distance du χ^2 au carré entre les modalités "malade" et "pas malade" s'écrit

$$\frac{105}{15}\left(\frac{1}{9}-\frac{1}{6}\right)^2+\frac{105}{30}\left(\frac{2}{9}-\frac{2}{6}\right)^2+\frac{105}{20}\left(\frac{2}{9}-\frac{1}{6}\right)^2+\frac{105}{40}\left(\frac{4}{9}-\frac{2}{6}\right)^2=\frac{49}{432}\approx 0.11.$$

Après regroupement de modalité on a obtenu la matrice des profils colonnes

$$\begin{pmatrix} 1/3 & 1/2 \\ 2/3 & 1/2 \end{pmatrix}$$

et la distance du χ^2 au carré entre les modalités "malade" et "pas malade" s'écrit

$$\frac{105}{45}\left(\frac{1}{3}-\frac{1}{2}\right)^2+\frac{105}{60}\left(\frac{2}{3}-\frac{1}{2}\right)^2=\frac{49}{432}\approx 0.11.$$

donc les distances du χ^2 sont bien préservées.

Proposition 30

Si deux colonnes j et j' de N_l sont proportionnelles, i.e. $\frac{n_{i,j}}{n_{\bullet,j}}=\frac{n_{i,j'}}{n_{\bullet,j'}}$ pour tout i, et qu'on combine les deux modalités associées en une alors la distance entre les profils-lignes reste inchangée.

Démonstration : Si on fusionne les deux dernières colonnes de N_l on obtient

$$N_l=\begin{pmatrix} \frac{n_{1,1}}{n_{1,\bullet}} & \cdots & \frac{n_{1,J-2}}{n_{1,\bullet}} & \frac{n_{1,J-1}+n_{1,J}}{n_{1,\bullet}} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{n_{I,1}}{n_{I,\bullet}} & \cdots & \frac{n_{I,J-2}}{n_{I,\bullet}} & \frac{n_{I,J-1}+n_{I,J}}{n_{I,\bullet}} \end{pmatrix}$$

et la distance entre deux profils lignes s'écrit

$$d^2_{\chi^2}(e_i,e_{i'})=\sum_{j=1}^{J-2}\frac{n}{n_{\bullet,j}}\left(\frac{n_{i,j}}{n_{i,\bullet}}-\frac{n_{i',j}}{n_{i',\bullet}}\right)^2+\frac{n}{n_{\bullet,J-1}+n_{\bullet,J}}\left(\frac{n_{i,J-1}+n_{i,J}}{n_{i,\bullet}}-\frac{n_{i',J-1}+n_{i',J}}{n_{i',\bullet}}\right)^2$$

De plus, on sait que

$$\frac{n_{i,J-1}}{n_{\bullet,J-1}}=\frac{n_{i,J}}{n_{\bullet,J}} \text{ et } \frac{n_{i',J-1}}{n_{\bullet,J-1}}=\frac{n_{i',J}}{n_{\bullet,J}}.$$

On obtient alors

$$
\begin{aligned}
&\frac{n}{n_{\bullet,J-1}+n_{\bullet,J}}\left(\frac{n_{i,J-1}+n_{i,J}}{n_{i,\bullet}}-\frac{n_{i',J-1}+n_{i',J}}{n_{i',\bullet}}\right)^2\\
&=\frac{n}{n_{\bullet,J-1}+n_{\bullet,J}}\left(\frac{n_{i,J}}{n_{i,\bullet}}\left(1+\frac{n_{\bullet,J-1}}{n_{\bullet,J}}\right)-\frac{n_{i',J}}{n_{i',\bullet}}\left(1+\frac{n_{\bullet,J-1}}{n_{\bullet,J}}\right)\right)^2\\
&=n(n_{\bullet,J-1}+n_{\bullet,J})\left(\frac{n_{i,J}}{n_{i,\bullet}\,n_{\bullet,J}}-\frac{n_{i',J}}{n_{i',\bullet}\,n_{\bullet,J}}\right)^2\\
&=\frac{n}{n_{\bullet,J-1}}\left(\frac{n_{i,J}\,n_{\bullet,J-1}}{n_{i,\bullet}\,n_{\bullet,J}}-\frac{n_{i',J}\,n_{\bullet,J-1}}{n_{i',\bullet}\,n_{\bullet,J}}\right)^2+\frac{n}{n_{\bullet,J}}\left(\frac{n_{i,J}}{n_{i,\bullet}}-\frac{n_{i',J}}{n_{i',\bullet}}\right)^2\\
&=\frac{n}{n_{\bullet,J-1}}\left(\frac{n_{i,J-1}}{n_{i,\bullet}}-\frac{n_{i',J-1}}{n_{i',\bullet}}\right)^2+\frac{n}{n_{\bullet,J}}\left(\frac{n_{i,J}}{n_{i,\bullet}}-\frac{n_{i',J}}{n_{i',\bullet}}\right)^2.
\end{aligned}
$$

■

Maintenant qu'on a définit les nuages de points, les poids associés et les métriques associées on peut alors définir l'inertie du jeu de donnée.

Définition 31

On définit l'inertie du nuage de point des profils lignes par

$$I_{tot}=\sum_{i=1}^{I}\frac{n_{i,\bullet}}{n}d^2_{\chi^2}(e_i,g)$$

et on définit l'inertie du nuage de point des profils colonnes par

$$I_{tot}=\sum_{j=1}^{J}\frac{n_{\bullet,j}}{n}d^2_{\chi^2}(e_i,g).$$

⚠ On ne centre pas les données ici !

Proposition 32

Dans les deux cas,

$$I_{tot}=\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{i,j}-\frac{n_{i,\bullet}n_{\bullet,j}}{n}\right)^2}{\frac{n_{i,\bullet}n_{\bullet,j}}{n}}.$$

Démonstration : On fait la preuve pour les profils lignes. On a

$$d^2_{\chi^2}(e_i,g)=\sum_{j=1}^{J}\frac{n}{n_{\bullet,j}}\left(\frac{n_{i,j}}{n_{i,\bullet}}-\frac{n_{\bullet,j}}{n}\right)^2$$

donc

$$
\begin{aligned}
I_{tot}&=\sum_{i=1}^{I}\frac{n_{i,\bullet}}{n}d^2_{\chi^2}(e_i,g)\\
&=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{i,\bullet}}{n}\frac{n}{n_{\bullet,j}}\left(\frac{n_{i,j}}{n_{i,\bullet}}-\frac{n_{\bullet,j}}{n}\right)^2\\
&=\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n}{n_{i,\bullet}n_{\bullet,j}}\left(n_{i,j}-\frac{n_{i,\bullet}n_{\bullet,j}}{n}\right)^2.
\end{aligned}
$$

■

Remarque: La quantité nI_{tot} est appelée la statistique du χ^2 des données et quantifie la dépendance entre les deux variables qualitatives.

3 ACP dans le cas général

On considère un jeu de données X de taille $n \times p$, pas forcement centré ou normalisé, avec une matrice de poids D et une matrice de métrique M. On suppose D et M être des matrices diagonales à termes positifs et on suppose que $\sum_i D_{i,i} = 1$. On note e_i les lignes de X. Les résultats suivants montrent comme se généralise l'ACP dans ce cadre.

Définition 33

On définit le centre de gravité g des individus pondérés par la matrice de poids D par

$$g = \sum_{i=1}^{n} D_{i,i} e_i = {}^tXD1_n.$$

On définit les données centrées par $Y = X - 1_n {}^tg$ et la matrice de covariance par $C = {}^tYDY$. Autrement dit,

$$C_{j,k} = \sum_{k=1}^{n} D_{i,i}(x_{i,j} - g_j)(x_{i,k} - g_k).$$

Définition 34

On définit le produit scalaire $\langle \cdot, \cdot \rangle_M$ et la norme $\| \cdot \|_M$ associés à la métrique M par

$$\langle u, v \rangle_M = \langle u, Mv \rangle \text{ et } \|u\|_M^2 = \langle u, u \rangle_M.$$

*On dit que $(u_1, \cdots, u_k)$ est une famille **M-orthonormée** si $\|u_i\|_M = 1$ pour tout i et $\langle u_i, u_j \rangle_M = 0$ pour tout $i \neq j$. On définit alors la **M-projection** d'un point x sur $H = Vect(u_1, \cdots, u_k)$ par*

$$P_H^M(x) = \langle x, u_1 \rangle_M u_1 + \cdots + \langle x, u_k \rangle_M u_k.$$

Proposition 35

La matrice CM possède une base de vecteurs propres M-orthonormée $(u_1, \cdots, u_p)$ de valeurs propres associées $\lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant 0$. De plus, l'espace H_k engendré par $(u_1, \cdots, u_k)$ maximise

$$I_H = \sum_{i=1}^{n} D_{i,i} \left\| P_H^M(e_i - g) \right\|_M^2$$

parmi les espaces vectoriels de dimension k avec $I_{H_k} = \sum_{i=1}^{k} \lambda_i$.

Démonstration : Admis ■

Dans le cas particulier de l'AFC, on a les propriétés supplémentaires suivantes.

Proposition 36

g est un vecteur propre de CM associé à la valeur propre 0. De plus, les matrices CM et ${}^tXDXM$ possèdent la même base M-orthonormée de vecteurs propres et les mêmes valeurs propres associées à la différence que g possède la plus grande valeur propre égale à 1.

Démonstration : On fait la preuve pour les profils lignes mais le même raisonnement s'applique aussi aux profils colonnes. Pour les profils lignes on a $X = N_l$, $n = I$, $p = J$, $D = \frac{1}{n} D_l$ et $M = nD_c^{-1}$. On commence par démontrer le lemme suivant.

Lemme 37

Le centre de gravité g vérifie les propriétés suivantes :

- $Mg = 1_J$;
- $\|g\|_M = 1$;
- $CMg = 0_J$.

Démonstration :

- Comme M est diagonale alors pour tout j on a $(Mg)_j = M_{j,j}g_j = n\frac{1}{n_{\bullet,j}}\frac{n_{\bullet,j}}{n} = 1$.
- $$\|g\|_M = \langle g, Mg\rangle = \langle g, 1_J\rangle = \sum_{j=1}^{J} \frac{n_{\bullet,j}}{n} = \frac{n}{n} = 1.$$
- On a $CMg = C1_J$. Or, comme C est une matrice de covariance, on a que pour tout i :
$$(C1_J)_i = \sum_{j=1}^{J} \mathrm{cov}(v_i, v_j) = \mathrm{cov}\left(v_i, \sum_{j=1}^{J} v_j\right).$$
Comme pour la matrice des profils lignes la somme des variables/colonnes est constante égale à 1_I alors $\mathrm{cov}\left(v_i, \sum_{j=1}^{J} v_j\right) = \mathrm{cov}\left(v_i, 1_I\right) = 0$ ce qui prouve le résultat. ■

En conséquence de ce lemme, on en déduit que g est vecteur propre unitaire de CM associé à la valeur propre 0. De plus, on peut réécrire la matrice de covariance C sous la forme

$$\begin{aligned} C &= {}^t(X - 1_n{}^t g)D(X - 1_n{}^t g) \\ &= {}^tXDX - {}^t(1_n{}^t g)DX - {}^tXD1_n{}^t g + {}^t1_n{}^t gD1_n{}^t g \\ &= {}^tXDX - g^t1_nDX - g^tg + {}^tg1_nD1_n{}^t g \\ &= {}^tXDX - 2g^tg + \langle 1_n, D1_n\rangle {}^t gg. \end{aligned}$$

Comme $\langle 1_n, D1_n\rangle$ correspond à la somme des termes de D qui est égale à 1 on obtient alors $C = {}^tXDX - g^tg$ et donc

$$0 = CMg = {}^tXDXMg - g^tgMg = {}^tXDXMg - g\|g\|_M \implies {}^tXDXMg = g.$$

g est bien un vecteur propre unitaire de ${}^tXDXM$ associé à la valeur propre 1. Maintenant, on complète g en une base M-orthonormée $(g, u_1, \cdots, u_{J-1})$ de vecteurs propres de CM. On a alors pour tout i

$${}^tXDXMu_i = CMu_i + g^tgMu_i = \lambda_i u_i + \langle g, u_i\rangle_M g.$$

Comme $\langle g, u_i\rangle_M = 0$ alors on en déduit que $(g, u_1, \cdots, u_{J-1})$ est aussi une base M-orthonormée de vecteurs propres de ${}^tXDXM$ associée aux même valeurs propres (sauf pour g). ■

Proposition 38

Les composantes principales s'écrivent $c_i = XMu_i$ et forment une base D-orthogonale de vecteurs propres de $XM{}^tXD$ associés aux mêmes valeurs propres que les u_i.

Démonstration : Soit u_i un vecteur propre de ${}^tXDXM$ pour la valeur propre λ_i et $c_i = XMu_i$. Alors,

$$XM{}^tXDc_i = XM{}^tXDXMu_i = \lambda_i XMu_i = \lambda_i c_i$$

et donc c_i est vecteur propre de $XM{}^tXD$ pour la même valeur propre λ_i. De plus, pour tout i, j on a

$$\langle c_i, c_j\rangle_D = \langle XMu_i, DXMu_j\rangle = \langle u_i, M{}^tXDXMu_j\rangle = \lambda_j\langle u_i, u_j\rangle_M$$

d'où $\langle c_i, c_j \rangle_D = 0$ si $i \neq j$ et $\|c_i\|_D = \lambda_i$ pour tout i. ■

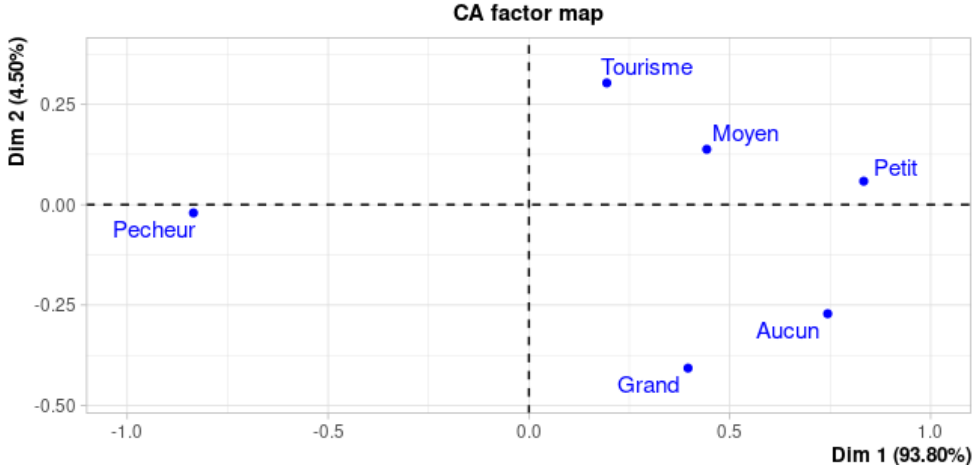
On peut alors résumer l'ACP de façon générale par le tableau suivant.

	ACP générale	AFC Profil ligne	AFC Profil colonne
Nuage de points	Lignes de $X \in \mathcal{M}_{n\times p}(\mathbb{R})$	$N_l = D_l^{-1}N$	${}^tN_c = D_c^{-1t}N$
Espace du nuage	$\mathbb{R}^p$	$\mathbb{R}^J$	$\mathbb{R}^I$
Matrice de poids	$D \in \mathcal{M}_n(\mathbb{R})$	$\frac{1}{n}D_l$	$\frac{1}{n}D_c$
Matrice de métrique	$M \in \mathcal{M}_p(\mathbb{R})$	nD_c^{-1}	nD_l^{-1}
Axes principaux	Vecteurs propres $u_i \neq g$ de ${}^tXDXM$ vérifiant $\|u_i\|_M^2 = 1$ et $\langle u_i, u_j \rangle_M = 0$ si $i \neq j$.	${}^tND_l^{-1}ND_c^{-1}$	$ND_c^{-1t}ND_l^{-1}$
Comp. principales	Vecteurs propres $c_i \neq XMg$ de $XM{}^tXD$ vérifiant $\|c_i\|_D^2 = \lambda_i$ et $\langle c_i, c_j \rangle_D = 0$ si $i \neq j$.	$D_l^{-1}ND_c^{-1t}N$	$D_c^{-1t}ND_l^{-1}N$

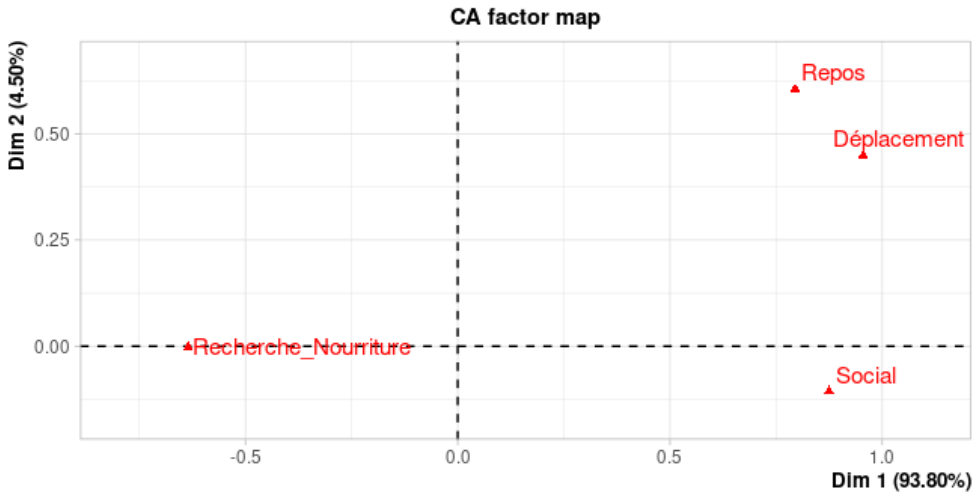
Remarques:

- Il y a au plus $\min(I, J) - 1$ valeurs propres non-nulles pour chaque matrice.
- Comme g est M-orthogonal aux H_k alors g est tout le temps projeté en l'origine.

Exemple: On obtient les résultats suivant pour l'AFC sur les données de comportement de dauphin.



(a) AFC sur les profils lignes



(b) AFC sur les profils colonnes

Du fait de la métrique considérée, on s'attend à ce que deux modalités d'une variable soient proche si elles sont influencées de façon similaire par les modalités de l'autre variable. On peut par exemple observer que les comportements *Repos* et *Déplacement* sont affectés de façon similaire par le type de bateau présent. On voit aussi que les bateaux de type *Tourisme*, *Petit* et *Moyen* entraînent des comportements similaires pour les dauphins.

4 Représentation duale

Proposition 39

Les matrices $D_l^{-1}ND_c^{-1}\ {}^tN$ et $D_c^{-1}\ {}^tND_l^{-1}N$ possèdent les mêmes valeurs propres non-nulles. Soient c_k et c'_k la k-ième composante principale de N_l et N_c et λ_k la valeur propre correspondante. Alors,

$$c_k(i) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{J} \frac{n_{i,j}}{n_{i,\bullet}} c'_k(j) \text{ et } c'_k(j) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{I} \frac{n_{i,j}}{n_{\bullet,j}} c_k(i).$$

Démonstration : On a $D_l^{-1}ND_c^{-1}\ {}^tNc_k = \lambda_k c_k$ par hypothèse. En multipliant les deux côtés par $D_c^{-1}\ {}^tN$ on obtient

$$D_c^{-1}\ {}^tND_l^{-1}ND_c^{-1}\ {}^tNc_k = \lambda_k D_c^{-1}\ {}^tNc_k$$

ce qui montre que $D_c^{-1}\ {}^tNc_k$ est un vecteur propre de $D_c^{-1}\ {}^tND_l^{-1}N$ associé à la même valeur propre que c_k. De plus, pour tout j, k on a

$$\begin{aligned}
&\langle D_c^{-1}\ {}^tNc_j, D_c^{-1}\ {}^tNc_k\rangle_{\frac{1}{n}D_c} \\
&= \frac{1}{n}\langle D_c^{-1}\ {}^tNc_j, D_cD_c^{-1}\ {}^tNc_k\rangle \\
&= \frac{1}{n}\langle c_j, ND_c^{-1}\ {}^tNc_k\rangle \\
&= \frac{1}{n}\langle c_j, D_lD_l^{-1}ND_c^{-1}\ {}^tNc_k\rangle \\
&= \frac{\lambda_k}{n}\langle c_j, D_lc_k\rangle \\
&= \lambda_k\langle c_j, c_k\rangle_{\frac{1}{n}D_l}. \\
&= \begin{cases} 0 \text{ si } j \neq k, \\ \lambda_k\|c_k\|^2_{\frac{1}{n}D_l} = \lambda_k^2 \text{ si } j = k. \end{cases}
\end{aligned}$$

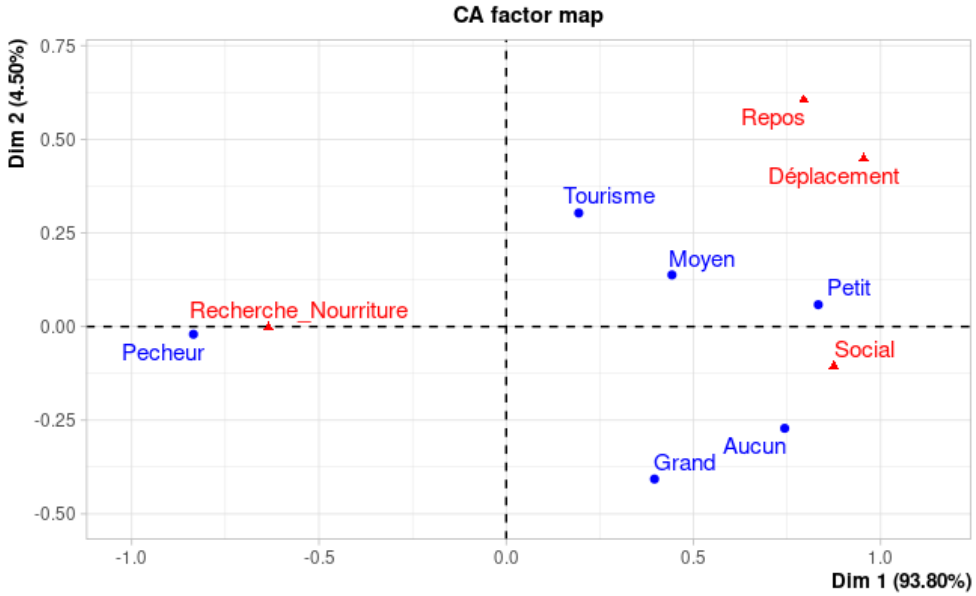
On peut alors conclure que les vecteurs $\frac{1}{\sqrt{\lambda_k}}D_c^{-1}\ {}^tNc_k$ forment une base $\frac{1}{n}D_c$-orthonormée avec $\|D_c^{-1}\ {}^tNc_k\|^2_{\frac{1}{n}D_c} = \lambda_k$ d'où $c'_k = D_c^{-1}\ {}^tNc_k$ et donc

$$c'_k(j) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{I} \frac{n_{i,j}}{n_{\bullet,j}} c_k(i).$$

On obtient ensuite c_k en fonction de c'_k avec le même raisonnement. ■

Ce résultat indique que, à un facteur $\frac{1}{\sqrt{\lambda_k}}$ près, la k-ième coordonnée d'une modalité i d'une variable est la moyenne des k-ième coordonnées des catégories de l'autre variable pondérées par les fréquences conditionnelles du profil de i. En conséquence, une modalité i va donc se situer du côté des modalités j pour lesquels les valeurs de $n_{i,j}$ sont les plus grandes.

Exemple: Si on superpose les deux AFC sur les données de comportement de dauphin on obtient le résultat suivant.



On peut alors observer que les dauphins on plutôt tendance à entre en train de chercher de la nourriture quand il y a un bateau de pêcheur dans les environs et plutôt tendance à avoir un comportement social quand il y a un petit bateau dans les environs.

Remarque: On peut définir le pourcentage d'inertie, les qualités de représentation et les contributions des modalités de la même façon que pour l'ACP mais en remplaçant les normes par des normes du χ^2.

IV Analyse des correspondances multiples

On considère maintenant un jeu de données avec n individus et p variables qualitatives. On note $J_1, \cdots, J_p$ le nombre de modalités de chaque variable et $J = J_1 + \cdots + J_p$ le nombre total de modalités. On souhaite généraliser le principe de l'AFC dans ce cadre là mais cette fois-çi on voudra aussi visualiser le comportement des individus.

1 Tableau disjonctif complet

Une façon alternative de travailler avec une variable qualitative à valeur dans un espace $\Omega = \{\omega_1, \cdots, \omega_N\}$ est de remplacer cette variable par les N indicatrices $\mathbb{1}_{\omega_1}, \cdots, \mathbb{1}_{\omega_N}$. Si l'on fait ça pour toutes la variables alors on obtient un jeu de données X de taille $n \times J$ tel que $x_{i,j} = 1$ si l'individu i possède la modalité j et 0 sinon.

Exemple :

Couleur des yeux	Sexe
Bleu	F
Marron	M
Marron	M
Vert	F
Bleu	M

$$\longrightarrow X = \left(\begin{array}{ccc:cc} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{array}\right)$$

La matrice de 0 et de 1 obtenue s'appelle le **tableau disjonctif complet** (ou TDC).

Proposition 40

- *La somme des lignes de X est le vecteur $p1_n$.*

- *La somme des colonnes de X est le vecteur $(n_1, \cdots, n_J)$ où n_j représente le nombre d'individu possédant la modalité j.*
- *La somme des éléments de X est pn.*

Proposition 41

$({}^tXX)_{i,j}$ est le nombre d'individus possédant la modalité i et la modalité j.

Démonstration : On remarque que $({}^tXX)_{i,j} = \sum_{k=1}^{n} x_{k,i}x_{k,j}$. Or, $x_{k,i}x_{k,j}$ est égal à 1 si l'individu k possède la modalité i et la modalité j et 0 sinon. D'où le résultat. ■

Corollaire 42

Si $p = 2$ alors

$$ {}^tXX = \begin{pmatrix} D_l & N \\ {}^tN & D_c \end{pmatrix}, $$

où N est la table de contingence des deux variables et D_l et D_c sont les matrices diagonales des effectifs lignes et colonnes.

Exemple: On reprend le TDC de l'exemple précédant.

Couleur des yeux	Sexe
Bleu	F
Marron	M
Marron	M
Vert	F
Bleu	M

$$ \longrightarrow N = \begin{pmatrix} 1 & 1 \\ 0 & 2 \\ 1 & 0 \end{pmatrix}, \; D_l = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ et } D_c = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}. $$

$$ {}^tXX = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 & 1 & 1 \\ 0 & 2 & 0 & 0 & 2 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 2 & 0 \\ 1 & 2 & 0 & 0 & 3 \end{pmatrix} $$

2 AFC du tableau disjonctif complet

Si on voit X comme une table de contingence à laquelle on applique l'AFC alors la matrice diagonale des effectifs lignes est pI_n et la matrice diagonale des effectifs colonnes est $D := \begin{pmatrix} n_1 & & 0 \\ & \ddots & \\ 0 & & n_J \end{pmatrix}$. Le tableau des profils lignes s'écrit alors $(pI_n)^{-1}X = \frac{1}{p}X$ et a comme matrice de poids $\frac{1}{pn}(pI_n) = \frac{1}{n}I_n$ et comme matrice de métrique pnD^{-1}. Le tableau des profils colonnes s'écrit quant à lui XD^{-1} et a comme matrice de poids $\frac{1}{pn}D^{-1}$ et comme matrice de métrique $pn(pI_n)^{-1} = nI_n$.

a Cas où p=2

Si on suppose que l'on a $p = 2$ variables qualitatives et que l'on fait l'AFC du tableau des profils colonnes de X on obtient

	Profil colonne
Nuage de points	Colonnes de XD^{-1} = lignes de $D^{-1t}X$
Espace du nuage	$\mathbb{R}^n$
Matrice de poids	$\frac{1}{2n}D$
Matrice de métrique	nI_n
Axes principaux	Vecteurs propres u_i de $\frac{1}{2}XD^{-1}\ {}^tX$ vérifiant $n\|u_i\|_2^2 = 1$ et $\langle u_i, u_j\rangle = 0$ si $i \neq j$
Comp. principales	Vecteurs propres c_i de $\frac{1}{2}D^{-1t}XX$ vérifiant $\|c_i\|^2_{\frac{1}{2n}D} = \lambda_i$ et $\langle c_i, c_j\rangle_D = 0$ si $i \neq j$

On peut alors montrer le résultat suivant reliant le résultat de l'AFC de la table de contingence et l'AFC du tableau disjonctif complet.

Proposition 43

Soit $\begin{pmatrix} u \\ v \end{pmatrix}$ *une composante principale de l'AFC des profils colonnes de* X*. Alors, à un facteur près,* u *est une composante principale de l'AFC des profils lignes de* N *et* v *est une composante principale de l'AFC du profil colonne de* N*.*

Démonstration : $\begin{pmatrix} u \\ v \end{pmatrix}$ est un vecteur propre de $\frac{1}{2}D^{-1t}XX$ associé à une valeur propre μ. Comme ${}^tXX = \begin{pmatrix} D_l & N \\ {}^tN & D_c \end{pmatrix}$ et $D = \begin{pmatrix} D_l & 0 \\ 0 & D_c \end{pmatrix}$ alors

$$\begin{aligned}
&\frac{1}{2}\begin{pmatrix} D_l^{-1} & 0 \\ 0 & D_c^{-1} \end{pmatrix}\begin{pmatrix} D_l & N \\ {}^tN & D_c \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \mu\begin{pmatrix} u \\ v \end{pmatrix} \\
\Leftrightarrow &\begin{pmatrix} I_{I_1} & D_l^{-1}N \\ D_c^{-1}\ {}^tN & I_{I_2} \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = 2\mu\begin{pmatrix} u \\ v \end{pmatrix} \\
\Leftrightarrow &\begin{pmatrix} u + D_l^{-1}Nv \\ D_c^{-1}\ {}^tNu + v \end{pmatrix} = 2\mu\begin{pmatrix} u \\ v \end{pmatrix} \\
\Leftrightarrow &\begin{cases} D_l^{-1}Nv = (2\mu - 1)u \\ D_c^{-1}\ {}^tNu = (2\mu - 1)v \end{cases}
\end{aligned}$$

On obtient alors par substitution

$$\begin{cases} D_l^{-1}ND_c^{-1t}Nu = (2\mu - 1)^2 u \\ D_c^{-1}\ {}^tND_l^{-1}Nv = (2\mu - 1)^2 v \end{cases}$$

ce qui prouve le résultat. ■

b Cas où p est quelconque

On généralise donc l'AFC pour une jeu de données à p variables en faisant l'AFC du tableau disjonctif complet. C'est ce qu'on appelle l'analyse en composante multiple ou ACM. L'AFC des profils lignes de X donne le nuage des individus et l'AFC des profils colonnes de X donne le nuage des modalités.

	Nuage des individus	Nuage des modalités
Nuage de points	Lignes de $X_l = \frac{1}{p}X$	Colonnes de $X_c = XD^{-1}$
Espace du nuage	$\mathbb{R}^J$	$\mathbb{R}^n$
Matrice de poids	$\frac{1}{n}I_n$	$\frac{1}{np}D$
Matrice de métrique	npD^{-1}	nI_n
Axes principaux	Vec. propres u_i de $\frac{1}{p}{}^tXXD^{-1}$ vérifiant $\|u_i\|^2_{npD^{-1}} = 1$ et $\langle u_i, u_j\rangle_{npD^{-1}} = 0$ si $i \neq j$	Vec. propres u'_i de $\frac{1}{p}XD^{-1}\,{}^tX$ vérifiant $n\|u'_i\|^2_2 = 1$ et $\langle u'_i, u'_j\rangle = 0$ si $i \neq j$
Comp. principales	Vec. propres c_i de $\frac{1}{p}XD^{-1}\,{}^tX$ vérifiant $\frac{1}{n}\|c_i\|^2_2 = \lambda_i$ et $\langle c_i, c_j\rangle = 0$ si $i \neq j$	Vec. propres c'_i de $\frac{1}{p}D^{-1t}XX$ vérifiant $\|c'_i\|^2_{\frac{1}{np}D} = \lambda_i$ et $\langle c'_i, c'_j\rangle_{\frac{1}{np}D} = 0$ si $i \neq j$

Remarque: De la même façon que pour l'AFC on peut superposer les deux nuages et interpréter les emplacements des individus relativement aux emplacements des variables. Les individus du nuage des individus vont avoir tendance à se trouver du côté des modalités du nuage des modalités qu'ils possèdent. Néanmoins, en général on ne superpose pas forcement les deux nuages car ils peuvent très vite devenir surchargés.

Les métriques utilisées pour l'ACM s'écrivent de la façon suivante :

Définition 44

Soient e_i et v_j les lignes et les colonnes du TDC.

- ***Distance entre les individus :***

$$d(e_i, e_{i'})^2 = \left\|\frac{1}{p}e_i - \frac{1}{p}e_{i'}\right\|^2_{npD^{-1}} = \sum_{j=1}^{n}\frac{pn}{n_j}\left(\frac{x_{i,j}}{p} - \frac{x_{i,j'}}{p}\right)^2 = \frac{n}{p}\sum_{j=1}^{I}\frac{(x_{i,j} - x_{i',j})^2}{n_j}.$$

- ***Distance entre les modalités :***

$$d(v_i, v_{j'})^2 = \left\|\frac{1}{n_j}v_j - \frac{1}{n_{j'}}v_{j'}\right\|^2_{nI_n} = n\sum_{i=1}^{n}\left(\frac{x_{i,j}}{n_j} - \frac{x_{i,j'}}{n_{j'}}\right)^2.$$

En particulier, la distance entre les individus correspond alors à une distance euclidienne où la j-ème coordonnée est renormalisée par $\frac{n}{pn_j}$. On peut d'ailleurs réécrire la distance entre les modalités de la façon suivante :

Proposition 45

Soit j et j' deux modalités. On note $n^{\neq}_{j,j'}$ le nombre d'individu prenant soit la modalité j soit la modalité j'. Alors, la distance entre j et j' vérifie

$$d(v_i, v_{j'})^2 = \frac{nn^{\neq}_{j,j'}}{n_j n_{j'}}$$

Démonstration : Comme $x_{i,j}$ est à valeur 0 ou 1 alors $x_{i,j}^2 = x_{i,j}$ d'où

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{x_{i,j}}{n_j}-\frac{x_{i,j'}}{n_{j'}}\right)^2 &= \sum_{i=1}^{n}\left(\frac{x_{i,j}}{n_j^2}+\frac{x_{i,j'}}{n_{j'}^2}-\frac{2x_{i,j}x_{i,j'}}{n_j n_{j'}}\right)\\ &= \frac{1}{n_j}+\frac{1}{n_{j'}}-\sum_{i=1}^{n}\frac{2x_{i,j}x_{i,j'}}{n_j n_{j'}}\\ &= \frac{n_j+n_{j'}-\sum_{i=1}^{n}2x_{i,j}x_{i,j'}}{n_j n_{j'}}\\ &= \frac{\sum_{i=1}^{n}x_{i,j}^2+\sum_{i=1}^{n}x_{i,j'}^2-\sum_{i=1}^{n}2x_{i,j}x_{i,j'}}{n_j n_{j'}}\\ &= \frac{\sum_{i=1}^{n}(x_{i,j}-x_{i,j'})^2}{n_j n_{j'}}\end{aligned}$$

Or, $(x_{i,j}-x_{i,j'})^2 = 1$ si l'individu i prend soit la modalité j soit la modalité j' et 0 sinon. Donc,

$$\sum_{i=1}^{n}\left(\frac{x_{i,j}}{n_j}-\frac{x_{i,j'}}{n_{j'}}\right)^2 = \frac{n_{j,j'}^{\neq}}{n_j n_{j'}}.$$

■

3 Le nuage des variables

On cherche à représenter graphiquement les liens entre les variables qualitatives du jeu de données et les composantes principales qui sont des variables quantitatives. Pour cela, on utilise une métrique appelée **rapport de corrélation** (ou η^2) pour quantifier la dépendance entre une variable qualitative et une variable quantitative.

Définition 46

*Soient $X = (x_1, \cdots, x_n)$ une variable qualitative à k modalités et $Y = (y_1, \cdots, y_n)$ une variable quantitative. On note n_i le nombre d'individus prenant la modalité i de X et $\bar{y}_i$ la moyennes de Y pour les individus prenant la modalité i. On note $\bar{y}$ la moyenne totale de y. On définit alors le **rapport de corrélation** entre X et Y, noté $\eta^2(X,Y)$, par*

$$\eta^2(X,Y) = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \in [0,1].$$

On représente alors une variable qualitative v_i dans le carré $[0,1]^2$ par le point $(\eta^2(v_i,c_1), \eta^2(v_i,c_2))$ afin de représenter visuellement la dépendance entre les composantes principales et les variables qualitatives du jeu de données. C'est ce qu'on appelle le **nuage des variables**.

Exemple: Pour illustrer le fonctionnement du η^2 on donne deux exemples ci-dessous.

On voit que le rapport de corrélation est proche de 0 quand les valeurs de la variable qualitative n'a quasiment aucune influence sur les valeurs de la variable quantitative et qu'il est proche de

1 quand les valeurs de la variable qualitative arrivent à bien séparer les valeurs de la variable quantitative.

4 Variables et individus supplémentaires

- Variable quantitative supplémentaire :
  On représente les variables quantitatives supplémentaires dans un cercle de corrélation par leurs projections sur l'espace engendré par les composantes principales de l'ACM.
- Variable qualitative supplémentaire :
  On calcule d'abord les rapports de corrélation entre le variables qualitatives supplémentaires et les composantes principales pour les représenter dans le nuage des variables. On représente ensuite leurs modalités dans le nuage des modalités en calculant les projections de leurs valeurs sur le TDC sur l'espace engendré par les axes principaux de l'AFC sur le profil colonne du TDC.
- Individus supplémentaires :
  On représente les individus supplémentaires dans le nuage des individus en calculant leurs valeurs sur le TDC puis en les projetant sur l'espace engendré par les axes principaux de l'AFC sur le profil ligne du TDC.

V Compléments

L'analyse des données mixtes

On a vu comment visualiser un jeu de données constitué uniquement de variables qualitatives ou quantitatives mais on n'a pas vu le cas des données mixtes. Les deux méthodes classiques pour gérer les données mixtes sont les suivantes :

- Transformer les variables quantitatives en variables qualitatives (en remplaçant leur valeurs par des intervalles) et faire une ACM.
- Transformer les variables qualitatives en leur tableau disjonctif complet et faire comme une AFC mais en utilisant une métrique bien particulière. C'est ce qu'on appelle l'AFDM. On pourra trouver le détail technique sur l'article Analyse factorielle des données mixtes de Jérôme Pagès et utiliser la fonction *FAMD* du package *FactomineR* sur R. L'interprétation des résultats se fait alors de façon similaire à ce qui a déjà été vu en ACP et ACM.

Détails sur l'interprétation des résultats d'une ACP/AFC/ACM

Pour plus d'informations sur comment bien effectuer et bien interpréter une ACP/AFC/ACM, je recommande de lire le chapitre 11 de [3]. Il rentre bien plus en détail sur l'interprétation que ce que j'ai pu faire en cours. A noter que ce qu'ils appellent "facteurs" d'une ACP c'est ce que j'appelle les "axes".

Bibliographie

[1] G. Saporta. *Probabilités, analyse des données et statistique.* Editions Technip, 2006.

[2] Husson F., Lê S., and Pagès J. *Analyse de données avec R.* Didact Statistique. Presses universitaires de Rennes, 2009.

[3] B. Escofier and J. Pagès. *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation*. Sciences sup : cours et études de cas : masters et écoles d'ingénieurs. Dunod, 2008.

Chapitre 2

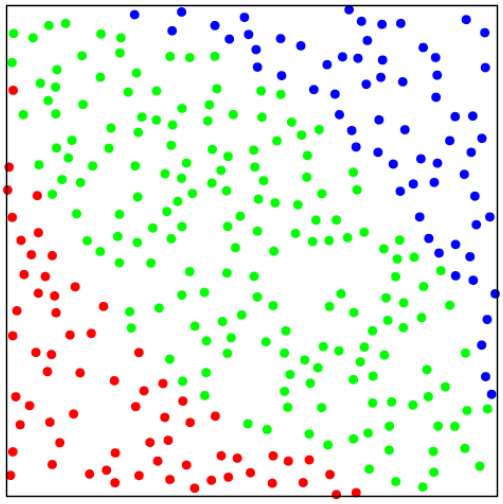
Classification des données

Introduction

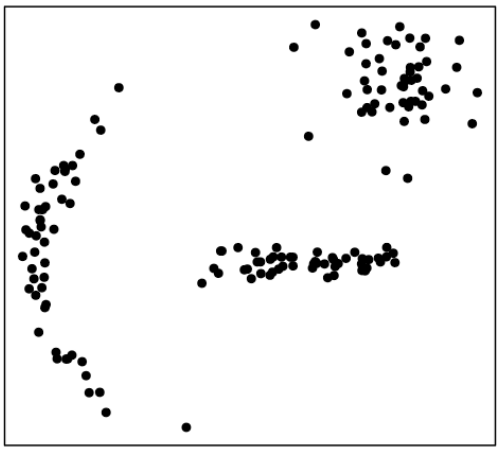
On considère qu'on a n individus $e_1, \cdots, e_n$ à valeur dans un espace Ω. En général on continuera de les voir comme les lignes d'un jeu de données mais ce ne sera pas toujours nécéssaire dans la suite du cours. Le but des méthodes de classification est de partitionner les individus en plusieurs sous-ensembles. On distingue deux types de problème de classification.

- **Classification supervisée :** A chaque individu e_i est associé un label y_i. La variable Y est donc une variable qualitative qui sépare les données en plusieurs sous-ensembles. On s'intéresse alors à prédire Y en fonction des e_i.
- **Classification non-supervisée :** Les individus ne sont pas déjà regroupé en catégories. On va donc chercher s'il y a une façon naturelle de regrouper les individus de sorte que deux individus dans une même classe soient similaires et deux individus dans une classe différente soient différents.

Exemple:



(a) Exemple de problème de classification supervisée.



(b) Exemple de problème de classification non supervisée.

FIGURE 2.1 – Sur la figure de droite on a des données déjà séparées en classe et on cherche une règle permettant de décider dans quelle classe doit se trouver une donnée. Sur la figure de gauche on a des données non séparées par une variable qualitative mais on peut observer une façon naturelle de les regrouper en classe.

Dans ce cours, on reste dans le domaine de la statistique exploratoire et **on se concentrera donc sur le cas de la classification non-supervisée**.

I Similarité entre individus

On considère qu'on a n individus $e_1, \cdots, e_n$ à valeur dans un espace Ω. On commence par définir une façon de quantifier à quel point ces individus sont similaires entre-eux.

Définition 47

*Soit $d : \Omega \times \Omega \to \mathbb{R}_+$. On dit que d est une **distance** sur Ω si elle vérifie pour tout $x, y, z \in \Omega$:*

- $d(x, y) = d(y, x)$ *; (Symétrie)*
- $d(x, y) = 0 \Leftrightarrow x = y$ *; (Séparation)*
- $d(x, z) \leqslant d(x, y) + d(y, z)$. *(Inégalité triangulaire)*

*S'il existe un produit scalaire tel que $d(x, y)^2 = \langle y - x, y - x\rangle$ alors on dit que d est une **distance euclidienne**.*
*Une fonction qui ne vérifie que les deux première propriétés est appelée une **dissimilarité**.*

Comparé au chapitre précédant où l'on devait forcement travailler avec des distances euclidiennes pour pouvoir faire des projections ce ne sera pas toujours nécessaire dans ce chapitre.
Exemples: Pour $e_i = (x_{i,1}, \cdots, x_{i,p}) \in \mathbb{R}^p$ on peut définir :

- La distance euclidienne classique :

$$d(e_i, e_j) = \|e_i - e_j\|_2 = \sqrt{\sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2}.$$

- La distance euclidienne sur les données centrées réduites :

$$d(e_i, e_j) = \sqrt{\sum_{k=1}^{p} \frac{(x_{i,k} - x_{j,k})^2}{\mathrm{var}(v_k)}}.$$

- La distance euclidienne associée à une matrice de métrique $M \in \mathcal{S}_p^+(\mathbb{R})$ (dont la métrique du χ^2) :

$$d(e_i, e_j) = \|e_i - e_j\|_M = \langle e_i - e_j, M(e_i - e_j)\rangle.$$

- La distance de Manhattan (non-euclidienne) :

$$d(e_i, e_j) = \|e_i - e_j\|_1 = \sum_{k=1}^{p} |x_{i,k} - x_{j,k}|.$$

- Les distances de Minkowski (non-euclidienne pour $l \neq 2$) :

$$d(e_i, e_j) = \|e_i - e_j\|_l = \left(\sum_{k=1}^{p} |x_{i,k} - x_{j,k}|^l\right)^{1/l} \quad \text{pour} \quad l \geqslant 1.$$

- La distance de Tchebychev (non-euclidienne) :

$$d(e_i, e_j) = \|e_i - e_j\|_\infty = \max_k |x_{i,k} - x_{j,k}|.$$

Pour donner une exemple d'une dissimilarité qui n'est pas une distance on considère $\Omega = \mathcal{P}(\{1, \cdots, n\})$ et la dissimilarité de Sørensen-Dice

$$d(A, B) = 1 - \frac{2|A \cap B|}{|A| + |B|},$$

où on note $|A|$ le cardinal de l'ensemble A. On a

$$d\Big(\{1\}, \{2\}\Big) = 1,\ d\Big(\{1\}, \{1, 2\}\Big) = \frac{1}{3} \text{ et } d\Big(\{2\}, \{1, 2\}\Big) = \frac{1}{3}$$
$$\Rightarrow\ d\Big(\{1\}, \{2\}\Big) > d\Big(\{1\}, \{1, 2\}\Big) + d\Big(\{2\}, \{1, 2\}\Big).$$

II Méthodes de classification générales

On commence par définir ce qu'est une partition.

Définition 48

*Une **partition** des individus en K classes est un choix de sous ensembles $P_1, \cdots, P_K$ (appelés les **classes** de la partition) de $\{e_1, \cdots, e_n\}$ tels que*

- $\forall i,\ P_i \neq \emptyset$,
- $\forall i, j$ *avec* $i \neq j$ *on a* $P_i \cap P_j = \emptyset$,
- $P_1 \cup \cdots \cup P_K = \{e_1, \cdots, e_n\}$

On note $P = \{P_1, \cdots, P_K\}$ une telle partition.

Remarque: Pour une partition $P = \{P_1, \cdots, P_K\}$ on a $|P_1| + \cdots + |P_K| = n$, où $|P_i|$ est le cardinal de P_i.

1 Partitionnement en k-means

Un critère usuel de classification pour quantifier la qualité d'une partition P est d'utiliser l'inertie comme mesure de la dispersion des éléments à l'intérieur et à l'extérieur d'une classe. On va donc utiliser les centres de gravité de chaque classes de la partition P que l'on note $g_1, \cdots, g_K$ défini par

$$g_i = \frac{1}{|P_i|} \sum_{e_j \in P_i} e_j.$$

Proposition 49

g est le centre de gravité des g_i pondérés par $|P_i|$:

$$g = \frac{1}{n} \sum_{i=1}^{K} |P_i| g_i.$$

Démonstration :

$$\frac{1}{n} \sum_{i=1}^{K} |P_i| g_i = \frac{1}{n} \sum_{i=1}^{K} \sum_{e_j \in P_i} e_j = \frac{1}{n} \sum_{j=1}^{n} e_j = g$$

■

Définition 50

Soit d une distance euclidienne.

- *On appelle* ***inertie intra-classe*** *la moyenne des inerties de chaque classe pondérées par les $|P_i|$:*

$$I_{intra} = \frac{1}{n}\sum_{i=1}^{K} |P_i| \times \frac{1}{|P_i|}\sum_{e_j \in P_i} d(e_j, g_i)^2 = \frac{1}{n}\sum_{i=1}^{K}\sum_{e_j \in P_i} d(e_j, g_i)^2$$

- *On appelle* ***inertie inter-classe*** *l'inertie des centres de gravité de chaque classe pondérés par leur taille :*

$$I_{inter} = \frac{1}{n}\sum_{i=1}^{K} |P_i|\ d(g_i, g)^2$$

Théorème 51 (*Théorème de Huygens*)

Soit d une distance euclidienne sur Ω et $\{e_1, \cdots, e_n\}$ un ensemble de points de Ω de centre de gravité g. Alors, pour tout $x \in \Omega$,

$$\frac{1}{n}\sum_{i=1}^{n} d(e_i, x)^2 = d(x, g)^2 + \frac{1}{n}\sum_{i=1}^{n} d(e_i, g)^2.$$

Démonstration :

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} d(e_i, g)^2 &= \frac{1}{n}\sum_{i=1}^{n}\langle e_i - g, e_i - g\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}(\langle e_i, e_i\rangle - 2\langle e_i, g\rangle + \langle g, g\rangle) \\
&= \frac{1}{n}\sum_{i=1}^{n}\langle e_i, e_i\rangle - 2\left\langle \frac{1}{n}\sum_{i=1}^{n} e_i, g\right\rangle + \langle g, g\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\langle e_i, e_i\rangle - 2\langle g, g\rangle + \langle g, g\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\langle e_i, e_i\rangle - \langle g, g\rangle
\end{aligned}$$

donc

$$\begin{aligned} d(x,g)^2 + \frac{1}{n}\sum_{i=1}^{n} d(e_i,g)^2 &= \langle g-x, g-x\rangle + \frac{1}{n}\sum_{i=1}^{n} d(e_i,g)^2 \\ &= \langle g,g\rangle + -2\langle g,x\rangle + \langle x,x\rangle + \frac{1}{n}\sum_{i=1}^{n}\langle e_i,e_i\rangle - \langle g,g\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n}\langle e_i,e_i\rangle - \frac{2}{n}\sum_{i=1}^{n}\langle e_i,x\rangle + \langle x,x\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\langle e_i,e_i\rangle - 2\langle e_i,x\rangle + \langle x,x\rangle\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\langle x-e_i, x-e_i\rangle^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} d(x,e_i)^2 \end{aligned}$$

■

Corollaire 52

$$I_{inter} + I_{intra} = \sum_{i=1}^{n} d(e_i,g)^2 = I_{total}$$

Démonstration :

$$\begin{aligned} \sum_{i=1}^{n} d(e_i,g)^2 &= \sum_{i=1}^{K}\sum_{e_j\in P_i} d(e_j,g)^2 \\ &= \sum_{i=1}^{K}\left(|P_i| d(g,g_i)^2 + \sum_{e_j\in P_i} d(e_j,g_i)^2\right) \\ &= I_{inter} + I_{intra}. \end{aligned}$$

■

Comme on cherche à avoir une partition avec des classes les moins dispersées possibles on cherche donc à **minimiser l'inertie intra-classe**. C'est ce qu'on appelle le **partitionnement en k-means**. Le corollaire précédant nous dit que minimiser l'inertie intra-classe est équivalent à maximiser l'inertie inter-classe.

⚠ Ce critère ne fonctionne que pour des distances euclidiennes et n'a de sens que si le nombre de classe est fixé. Sinon, on prendrait n classes contenant chacune un élément pour avoir une inertie intra-classe nulle.

On cherche alors à trouver une une partition à K éléments qui minimise l'inertie intra-classe. Si on note $p_{n,K}$ le nombre de partitions à K éléments d'un ensemble de taille n alors on a le résultat suivant.

Proposition 53

Soit $p_{n,K}$ le nombre de partitions à K éléments d'un ensemble de taille n. Alors

$$p_{n,K} \underset{n\to\infty}{\sim} \frac{K^n}{K!}.$$

Autrement dit, le nombre de partition à K éléments croit exponentiellement en le nombre d'individu. Ce n'est clairement pas faisable de tester toutes les partitions. De façon général, il n'existe pas d'algorithme rapide dû au résultat suivant.

Proposition 54

Soit $\Omega = \mathbb{R}^p$ avec $p \geqslant 2$, d une distance euclidienne sur Ω et $k = O(n^\varepsilon)$ pour $0 < \varepsilon < 1$. Dans ce cadre-là, le problème des k-means est NP-difficile.

On va donc établir un algorithme rapide permettant de trouver une solution pas forcement optimale mais qui aura quand même une inertie intra-classe faible. Cet algorithme est basé sur le résultat suivant.

Proposition 55

Soit P une partition de taille K et d'inertie intra-classe $I_{intra}(P)$. On note g_i le centre de gravité des individus dans P_i. Soit P' la partition telle que

$$P'_i = \{e_j \text{ t.q. } d(e_j, g_i) \leqslant d(e_j, g_k)\ \forall k \neq i\},$$

alors $I_{intra}(P') \leqslant I_{intra}(P)$

⚠ Si $d(e_j, g_i) = d(e_j, g_k)$ pour i et k on choisit au hasard si e_j se retrouve dans la partition P'_i ou P'_k.

Démonstration : Si on note g'_i le centre de gravité des individus dans P'_i alors

$$I_{intra}(P) = \frac{1}{n}\sum_{i=1}^{K}\sum_{e_j \in P_i} d(e_j, g_i)^2 \quad \text{et} \quad I_{intra}(P') = \frac{1}{n}\sum_{i=1}^{K}\sum_{e_j \in P'_i} d(e_j, g'_i)^2$$

Par le théorème de Huygens, on a pour tout i :

$$\sum_{e_j \in P'_i} d(e_j, g'_i)^2 \leqslant \sum_{e_j \in P'_i} d(e_j, g_i)^2.$$

De plus, par définition de P' on a

$$\sum_{i=1}^{K}\sum_{e_j \in P_i} d(e_j, g_i)^2 \geqslant \sum_{i=1}^{K}\sum_{e_j \in P'_i} d(e_j, g_i)^2$$

ce qui conclue le résultat. ■

Il en vient l'algorithme suivant :

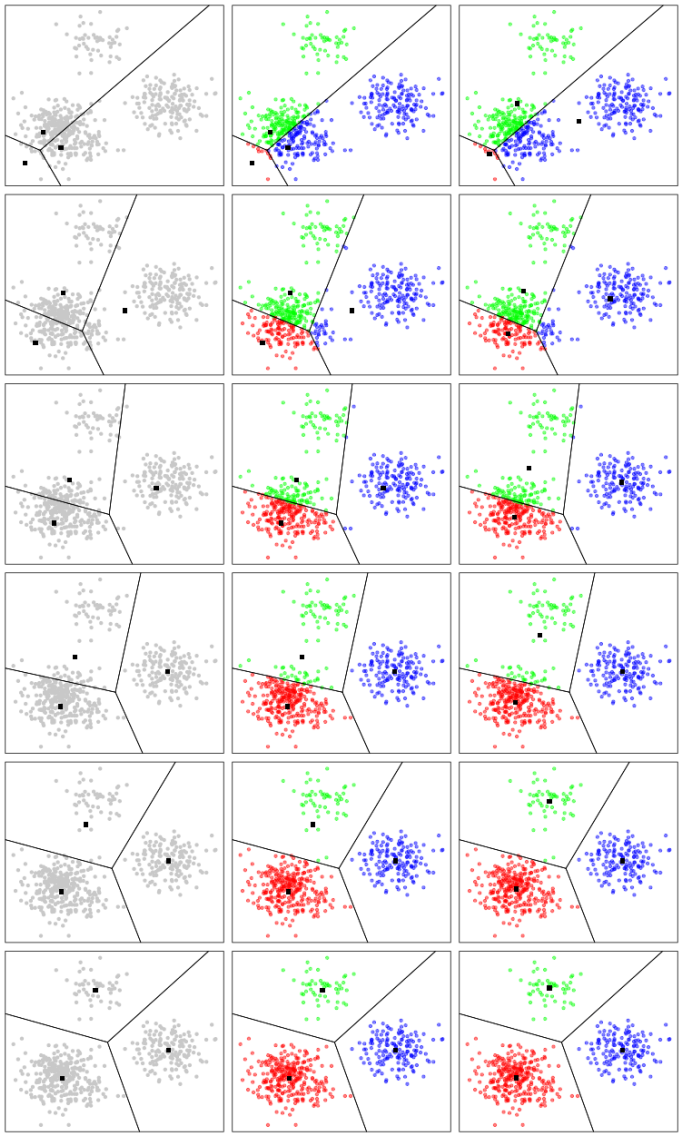
Algorithme 1 : Algorithme des moyennes mobiles

Entrées : $K \geqslant 0$; $e_1, \cdots, e_n \in \Omega$; d une distance euclidienne

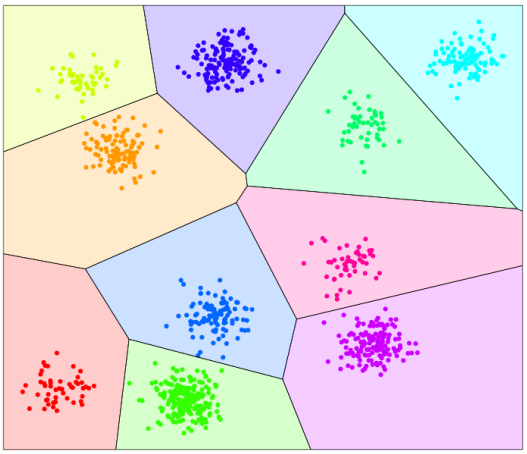
1 Choisir $\mu_1, \cdots, \mu_K$ distincts aléatoirement dans Ω.
2 Créer la partition P tel que $P_i = \{e_k \text{ t.q. } d(e_j, \mu_i) \leqslant d(e_j, \mu_k)\ \forall k \neq i\}$.
3 Affecter à chaque μ_i le centre de gravité des individus dans P_i.
4 Recommencer les étapes 2 et 3 jusqu'à ce que l'algorithme converge puis terminer et renvoyer P.

Cet algorithme termine forcement car l'inertie intra-classe de P diminue à chaque itération et il n'y a qu'un nombre fini de valeurs possibles pour P. Cet algorithme est très rapide mais il possède plusieurs gros défauts :

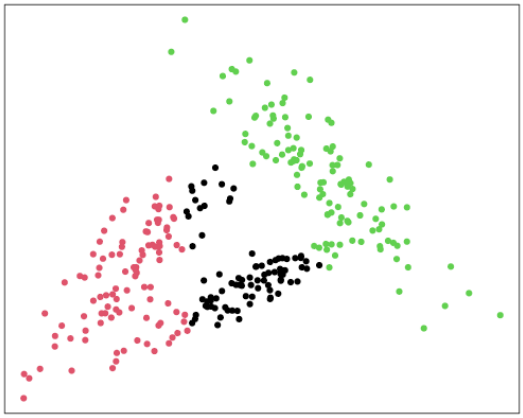
- Limité aux distances euclidiennes.
- Sensible aux valeurs absurdes.
- Fonctionne bien que pour des classes de forme circulaire et bien séparées.



(a) Exemple d'application de l'algorithme des moyennes mobiles avec 3 classes.



(b) Exemple d'un clustering à 10 classes avec l'algorithme des moyennes mobiles.



(c) Exemple où les k-means ne marchent pas.

FIGURE 2.2 – Trois exemples de partitionnement en k-means.

2 Partitionnement en k-medoïdes

Le partitionnement en k-médoïdes est l'analogue du partitionnement en k-means mais pour des dissimilarités.

Définition 56

Soit d une dissimilarité. On appelle ***médoïde*** *m des individus $\{e_1, \cdots, e_n\}$ l'individu e_i qui a la plus petite dissimilarité moyenne avec les autres individus :*

$$m = \underset{x \in \{e_1, \cdots, e_n\}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} d(x, e_i).$$

Remarque: Si $\Omega = \mathbb{R}$ et $d(x, y) = |y - x|$ alors la médoïde correspond à la médiane.

Le principe du partitionnement en k-médoïdes consiste à trouver la partition P de taille K qui

minimise la quantité

$$M = \sum_{i=1}^{K} \sum_{j \in P_i} d(e_j, m_i),$$

où m_i est la médoïde des individus dans P_i. Comme pour le partitionnement en k-means, il n'existe pas d'algorithme "rapide" permettant de résoudre ce problème. On utilise en général l'algorithme suivant renvoyant un résultat approximatif :

Algorithme 2 : PAM (Partitioning Around Medoids)

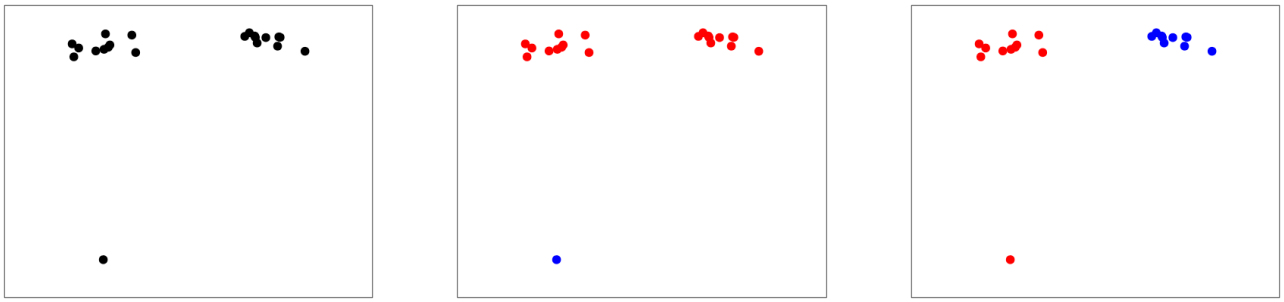
Entrées : $K \geqslant 0$; $e_1, \cdots, e_n \in \Omega$; d une dissimilarité

1 Choisir $\mu_1, \cdots, \mu_K$ distincts aléatoirement dans $\{e_1, \cdots, e_n\}$.
2 Créer la partition P tel que $P_i = \{e_k \text{ t.q. } d(e_k, \mu_i) \leqslant d(e_k, \mu_j)\ \forall j \neq i\}$.
3 Affecter à chaque μ_i la médoïde des individus dans P_i.
4 Recommencer les étapes 2 et 3 jusqu'à ce que l'algorithme converge puis terminer et renvoyer P.

Pour les mêmes raisons que la méthode des k-means, cet algorithme fait diminuer M à chaque itération jusqu'à ce qu'il converge vers un minimum local. Néanmoins, il y a plusieurs différences clés entre ces deux méthodes.

- La méthode des k-médoïdes avec la distance de Manhattan est moins affectée par les valeurs aberrantes que la méthode des k-means avec la distance euclidienne.
- L'algorithme PAM est en général beaucoup plus lent que l'algorithme des moyennes mobiles, surtout lorsque le nombre de données est très grand.
- L'algorithme PAM a juste besoin d'utiliser les valeurs des dissimilarité entre individus. On peut donc se limiter à juste utiliser la matrice de dissimilarité $(d(e_i, e_j))_{1 \leqslant i,j \leqslant n}$ comme entrée de l'algorithme.

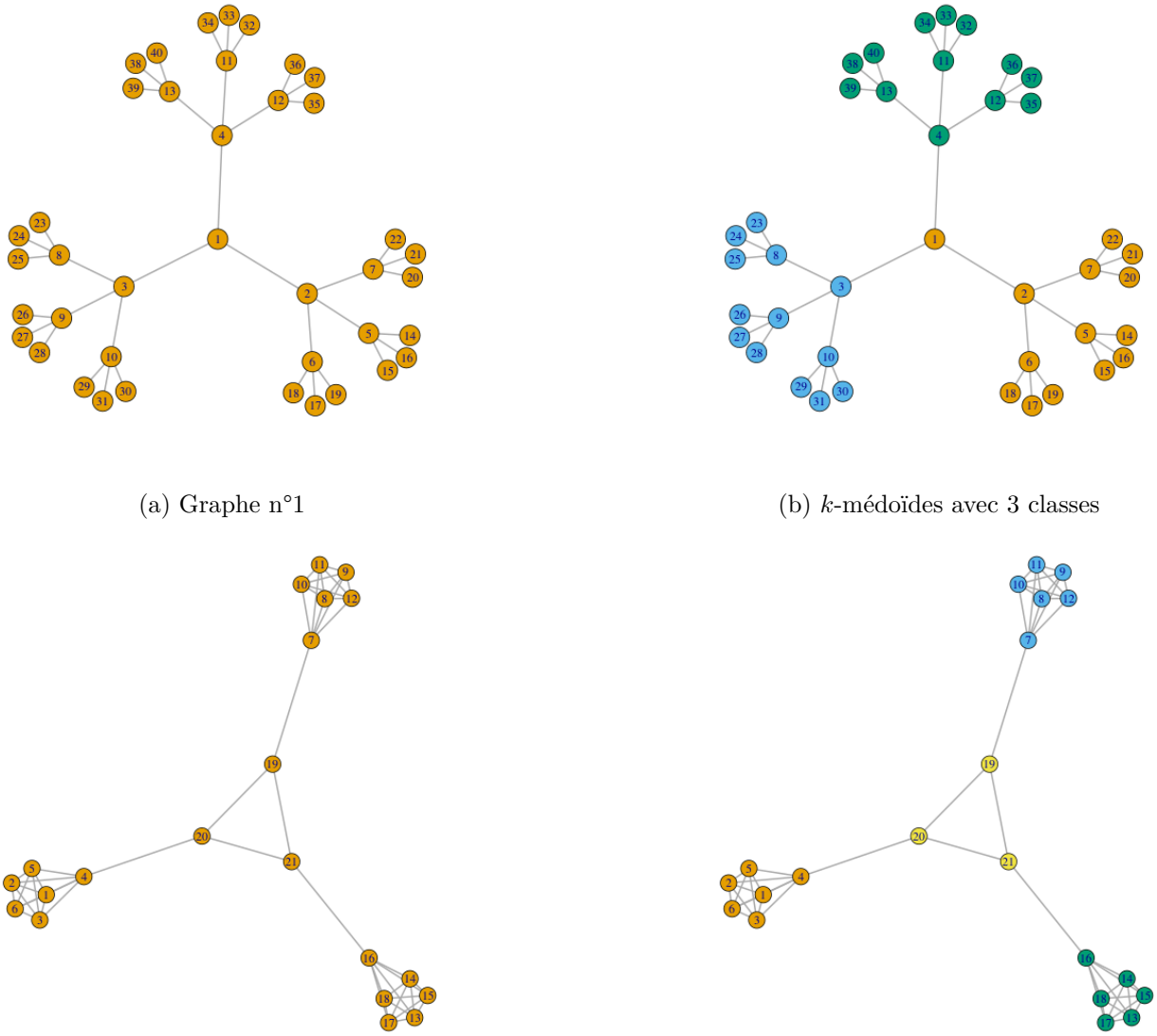
Exemple:



(a) Jeu de données (b) k-means à 2 classes (c) k-médoïdes à 2 classes

FIGURE 2.3 – Illustration de la différence entre l'utilisation des k-means et des k-médoïdes lorsqu'on a des valeurs absurdes.

Exemple: Comme la méthode des k-médoïdes ne se base que sur la matrice des dissimilarités entre les données alors on peut l'appliquer à n'importe quel type de données du temps qu'on peut définir une dissimilarité sur ces données. On peut par exemple l'appliquer pour classifier les sommets d'un graphe en prenant comme dissimilarité entre deux sommets le plus court chemin les reliant.



(a) Graphe n°1

(b) k-médoïdes avec 3 classes

(c) Graphe n°2

(d) k-médoïdes avec 4 classes

FIGURE 2.4 – Deux exemples de partitionnement des sommets d'un graphe avec les k-médoïdes.

3 Choix du nombre de classes pour les k-means

On s'intéresse à plusieurs méthodes pour choisir le nombre de classes K de la partition des données. On se concentre sur le cas des k-means mais les méthodes suivantes se généralisent bien dans le cas des k-médoïdes.

a Méthode du coude

Soit $I_{intra}(K)$ l'inertie intra-classe de la partition à K classe obtenue par le partitionnement en k-means. Plus on a de classe plus on va pouvoir diminuer l'inertie intra-classe donc $I_{intra}(K)$ va être décroissant par rapport à K. Néanmoins, à partir d'un certain rang K_0 on va avoir que pour tout $K \geqslant K_0$, $I_{intra}(K) \approx I_{intra}(K+1)$. Autrement dit, à partir de K_0 classes, le fait d'ajouter une classe n'a qu'un gain négligeable d'inertie intra-classe. On va donc choisir K_0 comme le nombre de classes de notre partition. Le nom de la méthode vient du fait que si on trace la courbe de $I_{intra}(K)$ en fonction de K alors ça va avoir une forme de bras et on choisit le nombre de classes qui forme le coude. (c.f. Figure 2.6)

b Méthode des silhouettes

On commence par définir les quantité suivantes qui vont quantifier à quel point un élément est bien partitionné.

Définition 57

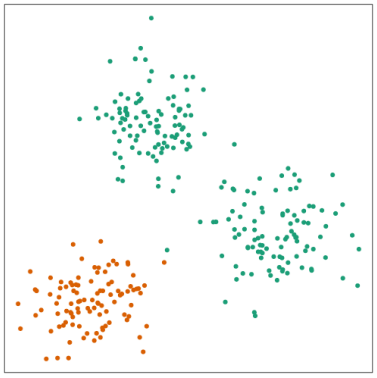
Soit P une partition en K classes et d une dissimilarité. On note $c(i) \in 1, \cdots, K$ la classe du i-ème individu. On définit les quantités suivantes :

- $a_i = \frac{1}{|P_{c(i)}|} \sum_{e_j \in P_{c(i)}} d(e_i, e_j)$, *la dissimilarité moyenne entre le i-ème individu et les autres individus de la même classe.*
- $\bar{d}(i,k) = \frac{1}{|P_k|} \sum_{e_j \in P_k} d(e_i, e_j)$, *la dissimilarité moyenne entre le i-ème individu et les individus de la classe P_k.*
- $b_i = \min_{k \neq c(i)} \bar{d}(i,k)$, *la dissimilarité moyenne entre le i-ème individu et les individus de la deuxième meilleur classe pour le i-ème individu.*
- $s_{i,K} = \frac{b_i - a_i}{\min(a_i, b_i)} \in [-1, 1]$, *appelé* ***la silhouette*** *de l'individu i.*
- $\bar{s}_K = \frac{1}{n} s_{i,K}$, *la silhouette moyenne des individus.*

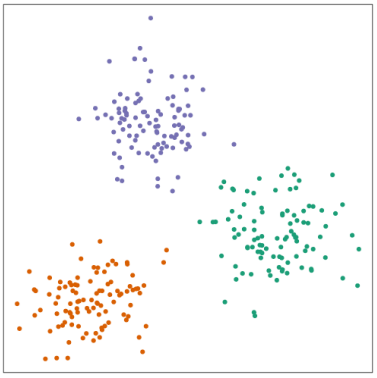
Plus $s_{i,K}$ est élevé plus l'individu i est bien partitionné. A l'inverse, une valeur de $s_{i,K}$ proche de 0 indique que l'individu i se trouve plutôt entre deux classes et une valeur très négative de $s_{i,K}$ indique un mauvais partitionnement. Le critère de cette méthode pour choisir le nombre de classes est de **choisir K qui maximise** $\bar{s}_K$. On utilise ensuite la quantité $SC = \max_k \bar{s}_K$, appelée **le coefficient de silhouette**, pour caractériser la structure obtenue par la classification :

- $0.75 < SC$: Structure forte
- $0.5 < SC \leqslant 0.75$: Structure modérée
- $0.25 < SC \leqslant 0.5$: Structure faible
- $SC \leqslant 0.25$: Pas de structure

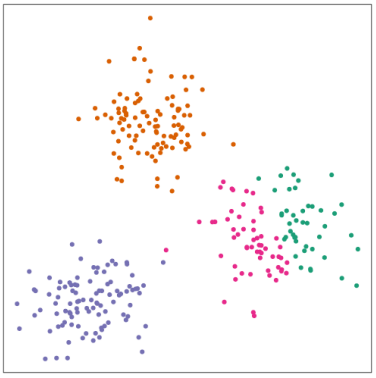
Exemples:



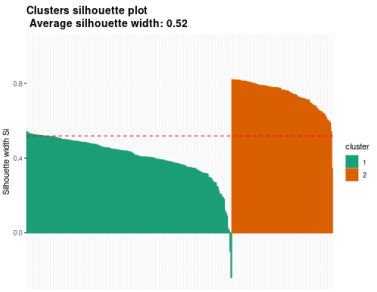
(a) k-means avec 2 classes



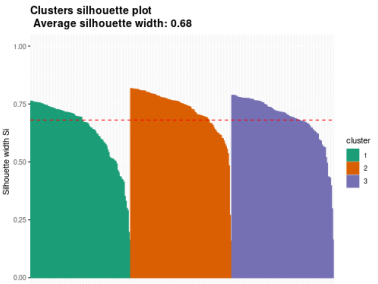
(b) k-means avec 3 classes



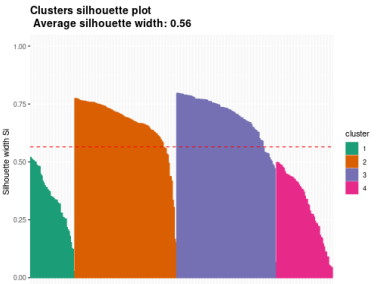
(c) k-means avec 4 classes



(d) Silhouettes avec 2 classes. On observe $SC = 0.52$.



(e) Silhouettes avec 3 classes. On observe $SC = 0.68$.



(f) Silhouettes avec 4 classes. On observe $SC = 0.56$.

FIGURE 2.5 – Exemples de silhouettes pour divers partitions d'un jeu de données.

c Statistique Gap

Soit $I_{intra}(K)$ l'inertie intra-classe de la partition à K classes obtenue par l'une des méthodes précédente. La statistique GAP est définie par

$$Gap(K) = \mathbb{E}[\log(I^*_{intra}(K))] - \log(I_{intra}(K)),$$

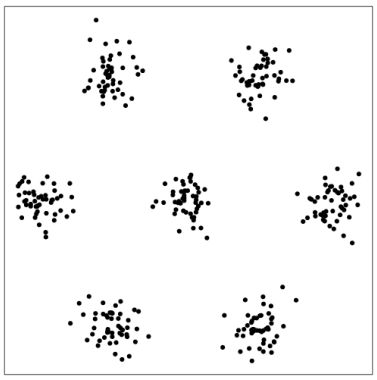
où $I^*_{intra}(K)$ est l'inertie intra-classe d'une partition à K classes de données simulées sous une bonne loi $\mathbb{P}^*$ générant des données sans cluster (typiquement, une loi uniforme). On s'attend à ce que plus le clustering est bon, plus la statistique $Gap(K)$ soit faible. Néanmoins, on ne connaît pas la valeur de $\mathbb{E}[\log(I^*_{intra}(K))]$. On va donc l'estimer en simulant N fois des données sous la loi $\mathbb{P}^*$ puis en les partitionnant avec la même méthode que pour le jeu de données initial. Si on note $I^*_{intra}(K, i)$ l'inertie intra-classe obtenue pour le i-ème jeu de données alors la statistique Gap s'écrit

$$Gap(K) = \frac{1}{N}\sum_{i=1}^{N} \log(I^*_{intra}(K, i)) - \log(I_{intra}(K)).$$

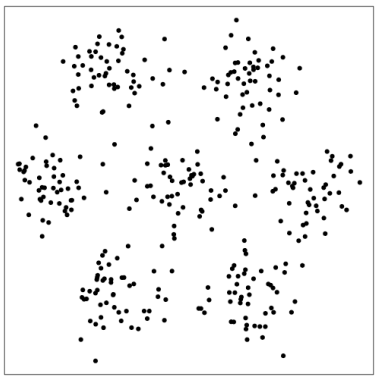
Afin de prendre en compte la variabilité dans les simulations, si on note s_K l'écart type des $\log(I^*_{intra}(K, i))$ alors on choisit comme nombre de classe le premier K tel que

$$Gap(K) \geqslant Gap(K+1) - \sqrt{1 + \frac{1}{N}} s_K.$$

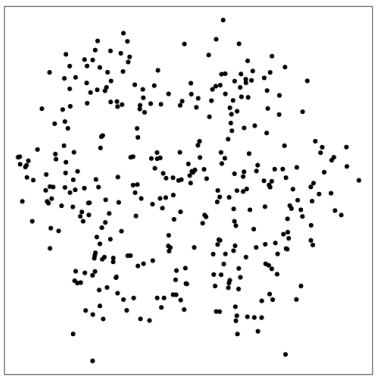
Exemples:



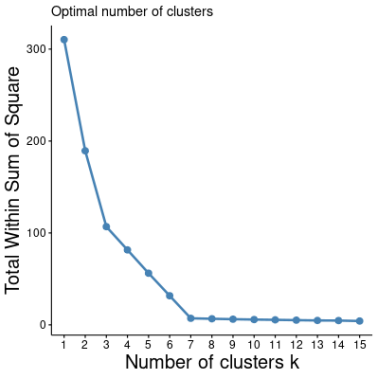
(a) Jeu de données n°1



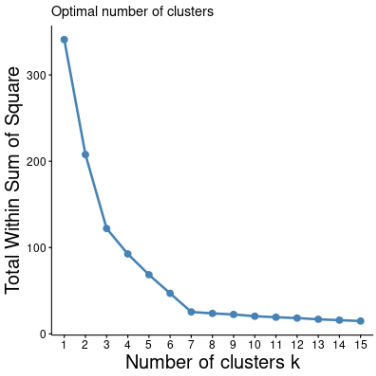
(b) Jeu de données n°2



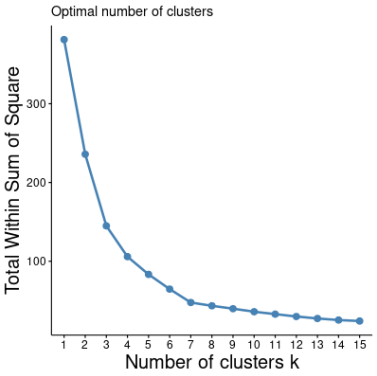
(c) Jeu de données n°3



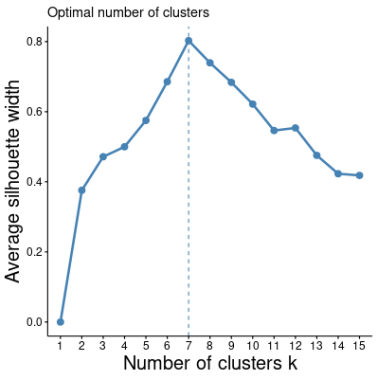
(d) Méthode du coude n°1



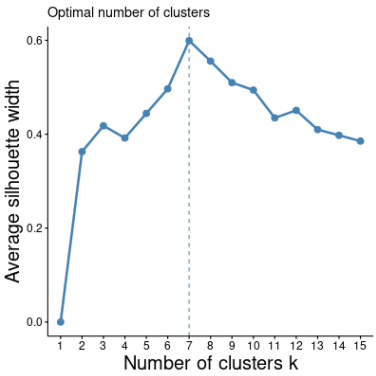
(e) Méthode du coude n°2



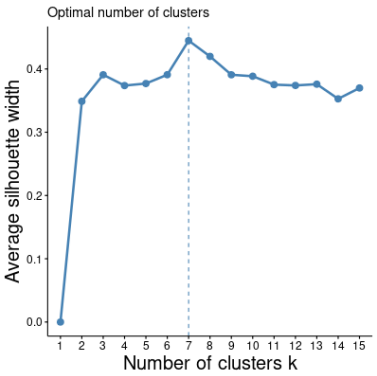
(f) Méthode du coude n°3



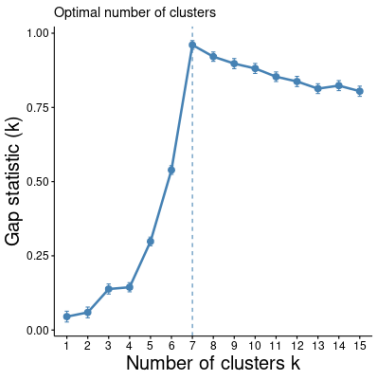
(g) Méthode des silhouettes n°1



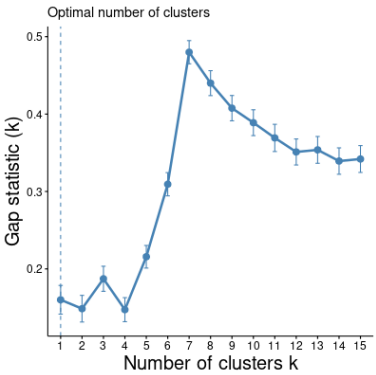
(h) Méthode des silhouettes n°2



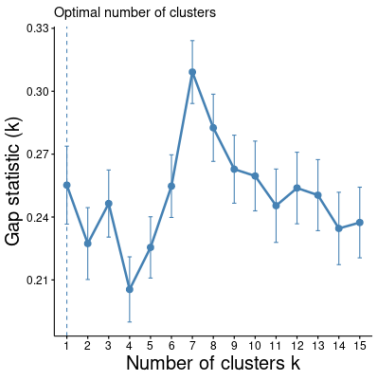
(i) Méthode des silhouettes n°3



(j) Statistique Gap n°1



(k) Statistique Gap n°2



(l) Statistique Gap n°3

FIGURE 2.6 – Exemple d'utilisation des trois méthodes de choix de classe présentées dans ce chapitre sur 3 jeux de données différents.

III Le clustering hiérarchique

Le principe du clustering hiérarchique consiste à commencer par considérer la partition en n classes où chaque individu est dans sa propre classe et itérativement fusionner les deux classes qui se ressemble le plus jusqu'à ne plus en avoir qu'une. Cette méthode va donc engendrer une partition en K classes pour tout $1 \leqslant K \leqslant n$. Pour cela on doit donc définir une dissimilarité entre les classes d'une partition.

1 Dissimilarité entre classes d'une partition

Définition 58

Soit d une dissimilarité et P_i, P_j deux classes d'une partition. On définit alors les mesures de dissimilarité D suivantes sur les classes :

- *Dissimilarité du plus proche voisin (single linkage) :*

$$D(P_i, P_j) = \min_{\substack{e_k \in P_i \\ e_l \in P_j}} d(e_k, e_l)$$

- *Dissimilarité du voisin le plus éloigné (complete linkage) :*

$$D(P_i, P_j) = \max_{\substack{e_k \in P_i \\ e_l \in P_j}} d(e_k, e_l)$$

- *Dissimilarité moyenne (average linkage) :*

$$D(P_i, P_j) = \frac{1}{|P_i||P_j|} \sum_{\substack{e_k \in P_i \\ e_l \in P_j}} d(e_k, e_l)$$

⚠ Attention à ne pas confondre la distance/dissimilarité utilisée entre les individus (que l'on notera avec un d minuscule) et la distance/dissimilarité utilisée entre les classes (que l'on notera avec un D majuscule).

Lorsque d est une distance euclidienne, on peut utiliser une distance particulière basée sur le résultat suivant :

Proposition 59

Soit d est une distance euclidienne. Soit P une partition avec au moins 2 classes et P' la partition obtenue en fusionnant P_1 et P_2. Alors,

$$n(I_{inter}(P) - I_{inter}(P')) = n(I_{intra}(P') - I_{intra}(P)) = \frac{|P_1||P_2|}{|P_1| + |P_2|} d^2(g_1, g_2).$$

Démonstration : On note $g_{1\cup 2}$ le centre de gravité des individus de $P_1 \cup P_2$. On a alors

$$g_{1\cup 2} = \frac{1}{[P_1| + |P_2|} \sum_{e_i \in P_1 \cup P_2} e_i = \frac{1}{[P_1| + |P_2|} \left(\sum_{e_i \in P_1} e_i + \sum_{e_i \in P_2} e_i \right) = \frac{|P_1| g_1 + |P_2| g_2}{[P_1| + |P_2|}$$

d'où

$$n(I_{inter}(P) - I_{inter}(P')) = |P_1| d^2(g_1, g) + |P_2| d^2(g_2, g) - (|P_1| + |P_2|) d^2(g_{1\cup 2}, g).$$

Or, on peut écrire

$$\begin{aligned}
&(|P_1|+|P_2|)d^2(g_{1\cup 2},g)\\
&=(|P_1|+|P_2|)\langle g_{1\cup 2}-g,g_{1\cup 2},g\rangle\\
&=(|P_1|+|P_2|)\left\langle \frac{|P_1|(g_1-g)+|P_2|(g_2-g)}{|P_1|+|P_2|},\frac{|P_1|(g_1-g)+|P_2|(g_2-g)}{|P_1|+|P_2|},\right\rangle\\
&=\frac{1}{|P_1|+|P_2|}(|P_1|^2d^2(g_1,g)+|P_2|^2d^2(g_2,g)+2|P_1||P_2|\langle g_1-g,g_2-g\rangle)\\
&=\frac{1}{|P_1|+|P_2|}(|P_1|^2d^2(g_1,g)+|P_2|^2d^2(g_2,g)+|P_1||P_2|\langle g_1-g,g_2-g_1\rangle+|P_1||P_2|d^2(g_1-g)\\
&\qquad\qquad +|P_1||P_2|\langle g_1-g_2,g_2-g\rangle+|P_1||P_2|d^2(g_2-g))\\
&=|P_1|d^2(g_1,g)+|P_2|d^2(g_2,g)+\frac{|P_1||P_2|}{|P_1|+|P_2|}(\langle g_1-g,g_2-g_1\rangle+\langle g_1-g_2,g_2-g\rangle)\\
&=|P_1|d^2(g_1,g)+|P_2|d^2(g_2,g)-\frac{|P_1||P_2|}{|P_1|+|P_2|}d^2(g_1,g_2).
\end{aligned}$$

On peut alors conclure que

$$n(I_{inter}(P)-I_{inter}(P'))=\frac{|P_1||P_2|}{|P_1|+|P_2|}d^2(g_1,g_2).$$

■

Définition 60

Soit d est une distance euclidienne. On définit la ***distance de Ward*** *D entre deux classes P_i et P_j d'une partition par*

$$D(P_i,P_j)=\frac{|P_i||P_j|}{|P_i|+|P_j|}d^2(g_i,g_j)$$

La distance de Ward correspond donc au gain d'inertie intra-classe (et à la perte d'inertie inter-classe) lorsqu'on fusionne les classes P_i et P_j. Comme on veut des partitions qui minimisent l'inertie intra-classe, il est donc logique de fusionner itérativement les paires de classes qui minimisent la distance de Ward.

2 Algorithme et dendrogramme

Algorithme 3 : Clustering ascendant

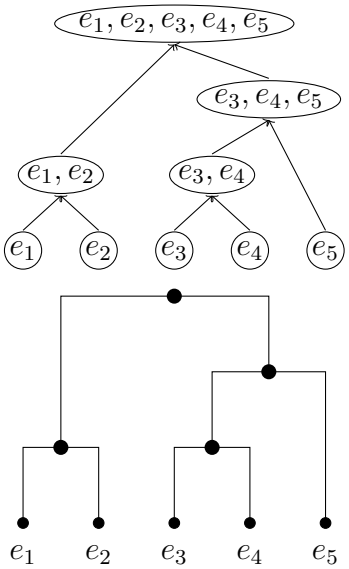
Entrées : $e_1,\cdots,e_n\in\Omega$, d : dissimilarité sur les individus, D : dissimilarité sur les classes d'une partition

1 Initialiser P en la partition triviale en n classes et créer une liste vide appelée *Fusion*.
2 Calculer les dissimilarités $D(P_i,P_j)$ entre toutes les paires de classes et choisir les classes P_I et P_J qui minimisent cette quantité.
3 Fusionner les classes P_I et P_J et ajouter le couple (P_I,P_J) à la liste *Fusion*.
4 Recommencer $n-1$ fois les étapes 2 et 3.
5 Terminer et renvoyer la liste *Fusion*.

On représente le résultat du clustering ascendant par un diagramme appelé **dendrogramme**. C'est un diagramme dont les nœuds correspondent aux classes obtenues par l'algorithme (en commençant par les classes avec un seul élément) et tel que lorsqu'on fusionne une classe P_i et une classe P_j on créé un nouveau nœud correspondant à la classe $P_i\cup P_j$ et on relie le nœud correspondant à la classe P_i et celui correspondant à la classe P_j à ce nouveau nœud.

Exemple:

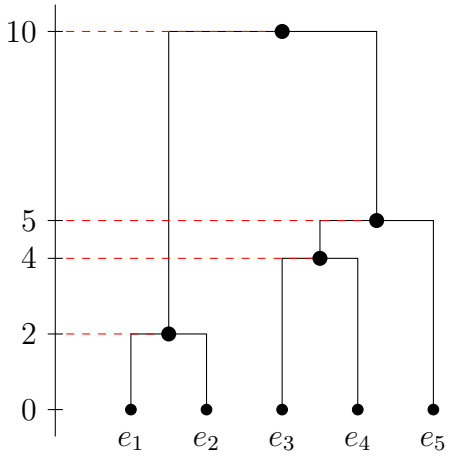
Si on considère 5 individus $e_1, \cdots, e_5$ et la liste *Fusion* avec les couples $(\{e_1\}, \{e_2\})$, $(\{e_3\}, \{e_4\})$, $(\{e_3, e_4\}, \{e_5\})$ et $(\{e_1, e_2\}, \{e_3, e_4, e_5\})$ alors on obtient le dendrogramme ci-contre.



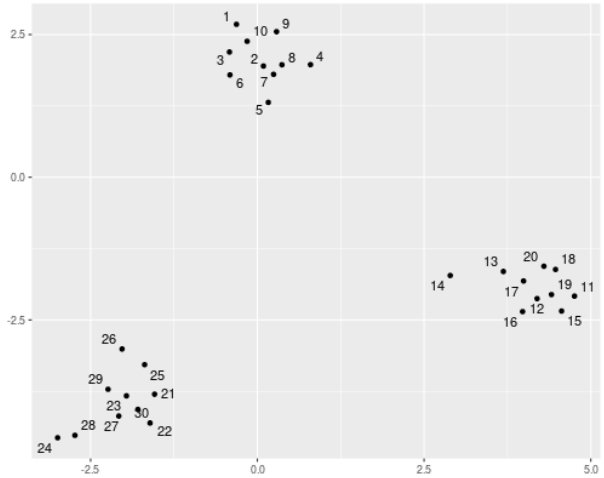
Une façon plus épurée (et plus facilement lisible quand on a beaucoup de données) de représenter un tel dendrogramme est représentée ci-contre.

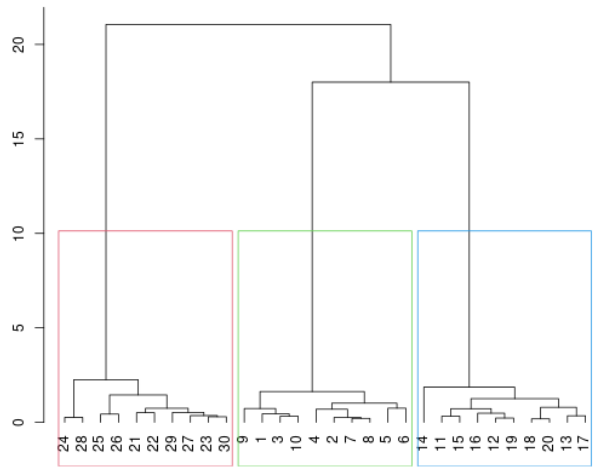
Il est aussi courant que lorsqu'on fusionne deux classes on place le nœud correspondant à une hauteur égale à la valeur de la dissimilarité entre les classes fusionnées.

Si on reprend la liste *Fusion* précédente avec les dissimilarités $D(\{e_1\}, \{e_2\}) = 2$, $D(\{e_3\}, \{e_4\}) = 4$, $D(\{e_3, e_4\}, \{e_5\}) = 5$ et $D(\{e_1, e_2\}, \{e_3, e_4, e_5\}) = 10$ alors on obtient le dendrogramme ci-contre.



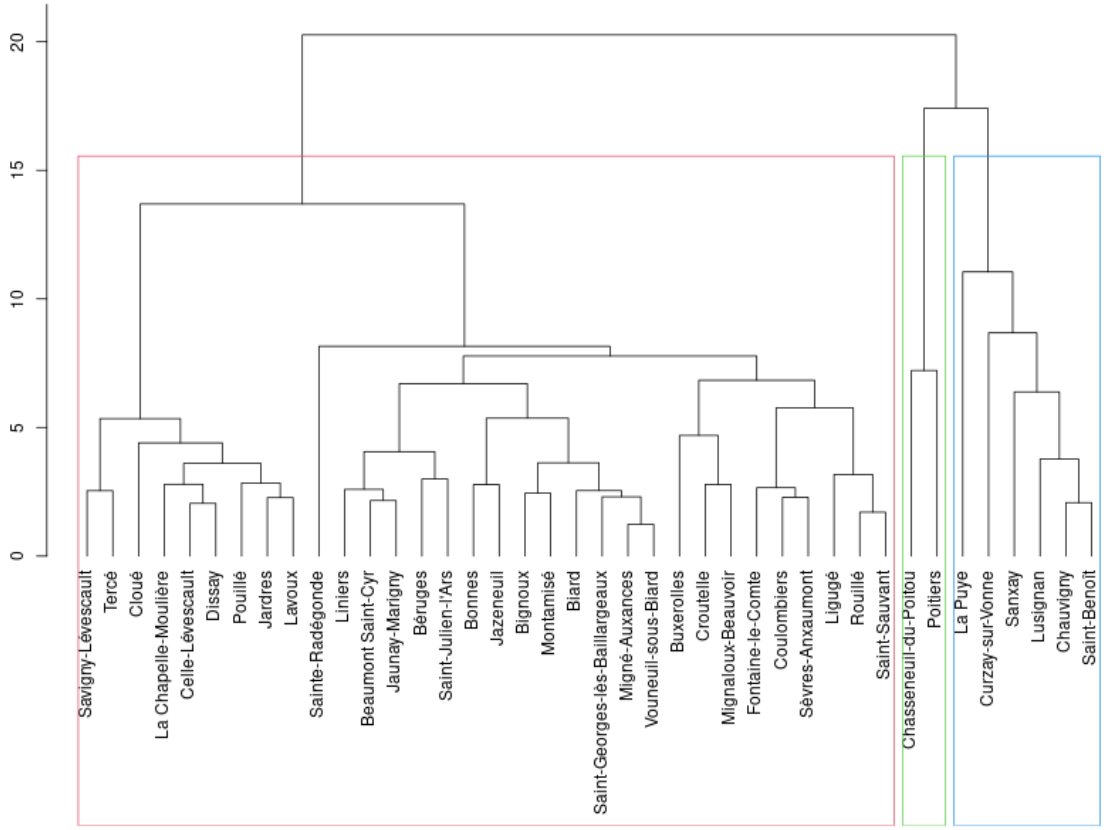
Remarque: On peut obtenir une classification en coupant l'arbre à une hauteur donnée. En général, il vaut mieux couper à l'endroit d'une grosse variation d'indice entre deux fusions de classe (ou sélectionner le nombre de classe avec l'une des méthodes vues précédemment).
Par exemple :

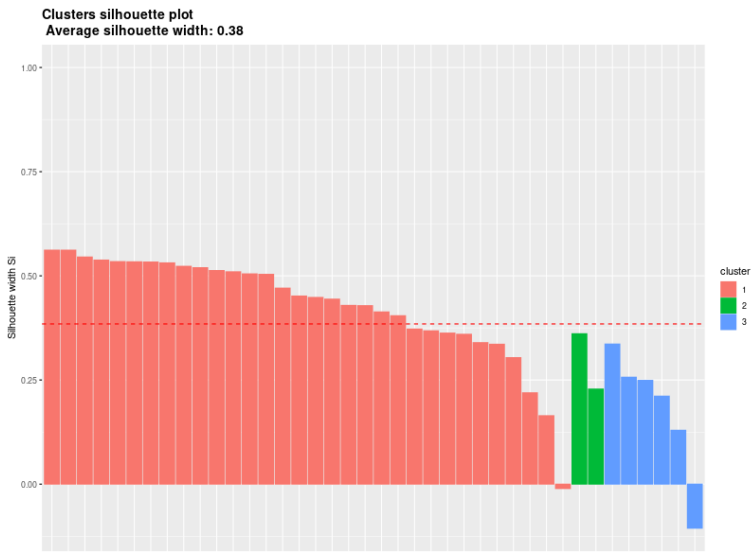




Exemple: Si on reprend les données centrées et réduites de démographie des communes de grand

Poitiers et qu'on leur applique le clustering ascendant avec la distance euclidienne classique entre les individus et la distance de Ward entre les classes, on obtient le dendrogramme et la classification en 3 classes ci-dessous, données avec le coefficient de silhouette correspondant.





On remarque que le coefficient de silhouette de 0.38 nous indique une structure assez faible pour cette classification. Elle ne vaut donc pas trop le coup d'être prise en compte. Cependant, même sans la classification le dendrogramme obtenue est une bonne représentation visuelle des distances entre les individus.

Application du clustering ascendant sur un example

On considère la matrice de dissimilarité suivante entre 5 individus et on veut appliquer l'algorithme de clustering ascendant avec la dissimilarité du plus proche voisin.

d	e_1	e_2	e_3	e_4	e_5
e_1	0	7	4	1	9
e_2	7	0	2	3	5
e_3	4	2	0	6	8
e_4	1	3	6	0	10
e_5	9	5	8	10	0

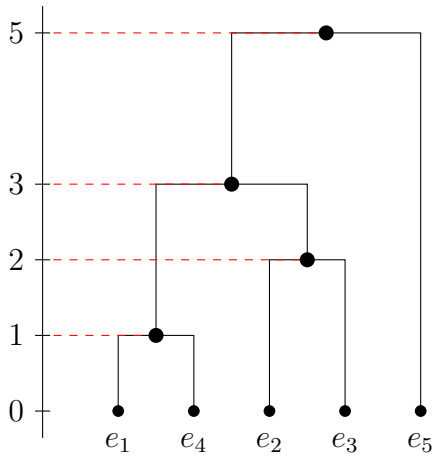
Tout d'abord, comme pour la dissimilarité entre deux ensembles composé d'un individu est égale à la distance entre les individus alors les classes avec la plus faible dissimilarité sont $\{e_1\}$ et $\{e_4\}$ avec $D(\{e_1\}, \{e_4\}) = 1$. On les fusionne puis on recalcule les dissimilarités entre les classes :

D	$\{e_1, e_4\}$	$\{e_2\}$	$\{e_3\}$	$\{e_5\}$
$\{e_1, e_4\}$	0	3	4	9
$\{e_2\}$	3	0	2	5
$\{e_3\}$	4	2	0	8
$\{e_5\}$	9	5	8	0

Les deux ensembles avec la dissimilarité la plus faible sont $\{e_2\}$ et $\{e_3\}$ avec $D(\{e_2\}, \{e_3\}) = 2$ donc on les fusionne et on recalcule les dissimilarités entre les classes :

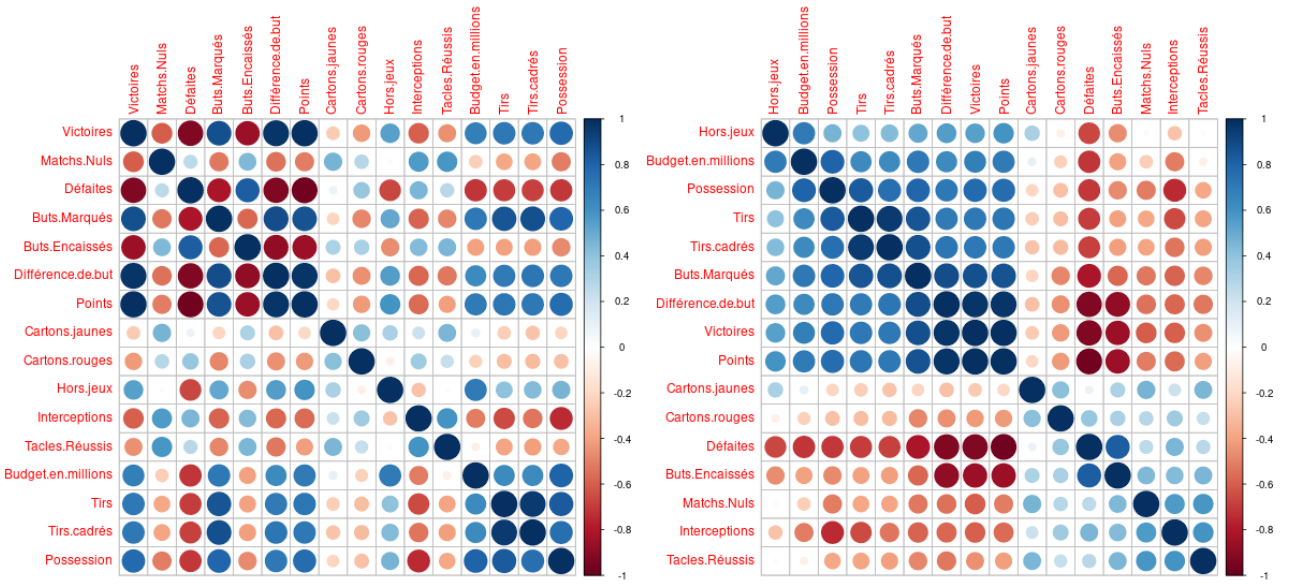
D	$\{e_1, e_4\}$	$\{e_2, e_3\}$	$\{e_5\}$
$\{e_1, e_4\}$	0	3	9
$\{e_2, e_3\}$	3	0	5
$\{e_5\}$	9	5	0

Les deux ensembles possédant la dissimilarité la plus faible sont $\{e_1, e_4\}$ et $\{e_2, e_3\}$ de dissimilarité $D(\{e_1, e_4\}, \{e_2, e_3\}) = 3$ donc on les fusionne. Il ne reste plus que les ensembles $\{e_1, e_2, e_3, e_4\}$ et $\{e_5\}$ avec $D(\{e_1, e_2, e_3, e_4\}, \{e_5\}) = 5$. On obtient alors le dendrogramme suivant :



3 Classification des variables

Comme on a vu que les méthodes de classification peuvent s'appliquer à n'importe quel type de données du temps que l'on peut définir une dissimilarité entre elles alors on peut aussi utiliser ces méthodes pour classifier les variables d'un jeu de données. On a déjà rencontrer de la classification de variable sans s'en rendre compte dans le TP n°3 quand on a utiliser la fonction *corrplot* sur R pour visualiser la matrice de corrélation des données de ligue 1. On a vu que lorsqu'on ajoute la paramètre *order="hclust"* à cette fonction alors les variables sont réordonnées de façon à améliorer la visualisation.



(a) Sans le paramètre *order="hclust"* (b) Avec le paramètre *order="hclust"*

FIGURE 2.7 – Utilisation du paramètre *hclust* dans la fonction *corrplot*.

La méthode utilisée pour réordonnée les variables est du clustering hiérarchique (avec la dissimilarité du voisin le plus éloigné entre les classes par défaut). La question se pose alors de la dissimilarité entre les variables à utiliser.

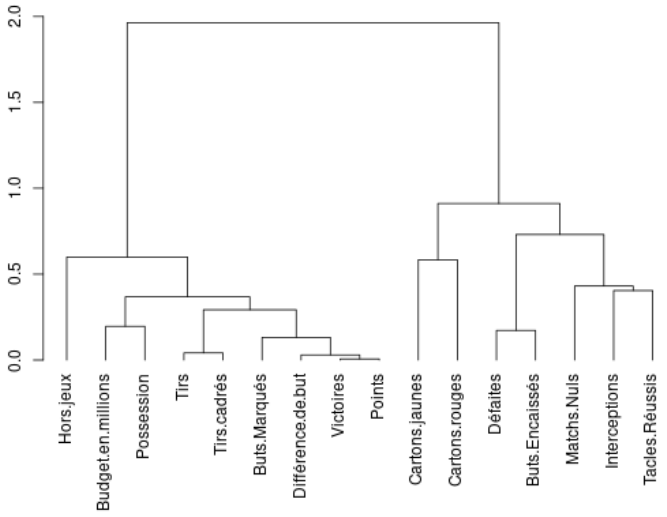
a Distance entre variables quantitatives

Il existe plein de façons de quantifier la similarité entre deux variables mais une des distances utilisées classiquement est $d(X,Y) = \sqrt{2 - 2\text{cor}(X,Y)} \in [0,2]$. C'est une distance euclidienne sur les variables réduites car si $\text{var}(X) = \text{var}(Y) = 1$ alors

$$\text{cov}(Y - X, Y - X) = \text{cov}(Y,Y) + \text{cov}(X,X) - 2\text{cov}(X,Y) = d^2(X,Y)$$

et $\langle u, v \rangle = \text{cov}(u,v)$ est un produit scalaire. On peut donc utiliser cette distance pour faire du clustering hiérarchique sur des variables avec la distance de Ward (et même du k-means !). Comme $d(X,Y)$ est maximal lorsque $\text{corr}(X,Y) = -1$ et égal à 0 lorsque $\text{corr}(X,Y) = 1$ alors lorsqu'on fait du clustering avec cette distance on va regrouper dans une même classe les variables fortement corrélées positivement et séparer les variables fortement corrélées négativement. Si à la place on souhaite regrouper les variables fortement corrélées positivement et négativement et séparer les variables décorrélées alors on peut prendre par exemple la dissimilarité $d(X,Y) = 1 - \text{corr}(X,Y)^2$.

Exemple: La dissimilarité utilisée par la fonction *corrplot* sur R afin de réordonner les variables pour la visualisation de la matrice de corrélation est $d(X,Y) = 1 - \text{corr}(X,Y)$. Si on trace le dendrogramme de la classification hiérarchique des variables des données de ligue 1 avec cette dissimilarité entre les variables et la dissimilarité du voisin le plus éloigné entre les classes on retrouve bien l'ordre des variables utilisé par la fonction *corrplot*. (c.f. Figure 2.7)



b Dissimilarité entre variables qualitatives

Il existe beaucoup de distances (et surtout de dissimilarités) pour de la classification de variables qualitatives. Je vais simplement en mentionner une basée sur la quantité suivante :

Définition 61

Soient X et Y deux variables qualitatives avec I et J modalités d'un jeu de n données. On note $\chi^2(X,Y)$ leur statistique du χ^2. On définit alors le V ***de Cramér*** *par*

$$V(X,Y) = \sqrt{\frac{\chi^2(X,Y)}{n \max(I-1, J-1)}} \in [0,1].$$

Remarque:

- On rappelle que $\frac{\chi^2(X,Y)}{n}$ correspond aussi à l'inertie totale de l'AFC sur les deux variables.
- $V = 0$ correspond à une indépendance parfaite entre les deux variables alors que $V = 1$ correspond au cas où la connaissance d'une variable détermine l'autre.
- La fonction $d(X,Y) = 1 - V(X,Y)$ est alors une dissimilarité. En faisant du clustering avec on va regrouper les variables très dépendantes entres-elles dans une même classe et séparer les variables indépendantes entres-elles.

Exemple:

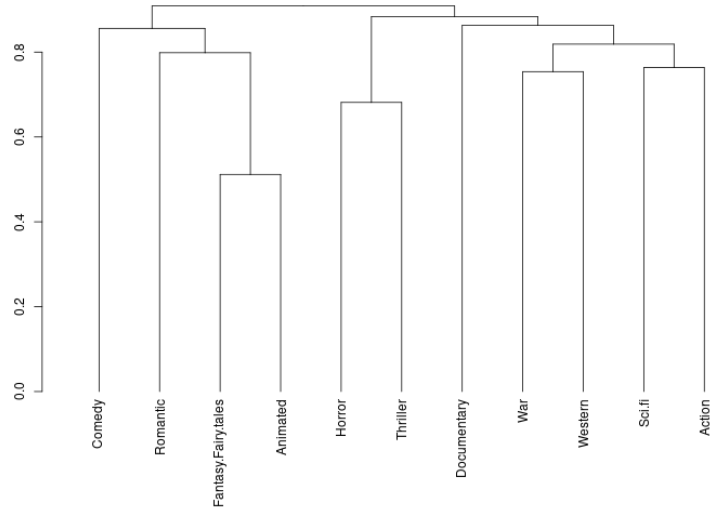


FIGURE 2.8 – Dendrogramme des variables de préférence de film pour l'exemple d'interprétation d'une ACM

IV Mélange de lois de probabilités

On se place maintenant dans un cadre de statistique inférentielle. On suppose que les individus d'une même classe sont issus d'une même loi et les individus de classes différentes sont issus de classes différentes. On va voir deux façons d'écrire ce problème mathématiquement.

1 Approche classification

Soit P une partition en K classes. On considère dans cette approche que les données d'une classe P_j ont été générées selon une loi de densité f_{θ_j}. Si on note $c(i) \in \{1, \cdots, K\}$ la classe du i-ème individu, alors la log-vraisemblance du modèle s'écrit

$$\mathcal{L}(\theta_1, \cdots, \theta_K, P) = \sum_{i=1}^{n} \log(f_{\theta_{c(i)}}(e_i)) = \sum_{j=1}^{K} \sum_{e_i \in P_j} \log(f_{\theta_j}(e_i)).$$

Le problème de clustering consiste donc à chercher la partition P de taille K et les paramètres $\theta_1, \cdots, \theta_K$ qui maximisent la vraisemblance.
Remarque:

- $\mathcal{L}(\theta_1, \cdots, \theta_K, P)$ a une forme similaire aux quantités que l'on chercher à minimiser pour les k-means et les k-médoïdes qui, pour rappel, sont :

$$\sum_{j=1}^{K} \sum_{e_i \in P_j} \|e_i - g_j\|^2 \text{ et } \sum_{j=1}^{K} \sum_{e_i \in P_j} d(e_i, m_j)$$

- Cette approche généralise les k-means dû au résultat suivant :

Proposition 62

On suppose que f_{θ_j} est la densité d'une loi normale $\mathcal{N}(\theta_j, \sigma^2 I_p)$ sur $\mathbb{R}^p$ avec $\sigma \geqslant 0$ connu. Alors, minimiser $\mathcal{L}(\theta_1, \cdots, \theta_K, P)$ est équivalent à résoudre le problème des k-means.

Démonstration :

$$\begin{aligned}\mathcal{L}(\theta_1, \cdots, \theta_K, P) &= \sum_{j=1}^{K} \sum_{e_i \in P_j} \log\left(\frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{1}{2\sigma^2}\|e_j - \theta_i\|_2^2\right)\right) \\ &= \sum_{j=1}^{K} \sum_{e_i \in P_j} -\frac{p}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{K}\sum_{e_i \in P_j} \|e_j - \theta_i\|_2^2.\end{aligned}$$

Comme σ est fixé, maximiser la vraisemblance revient donc à minimiser

$$\sum_{j=1}^{K}\sum_{e_i \in P_j} \|e_i - \theta_j\|_2^2.$$

Or, par le théorème de Huygens, on a que pour une partition P fixée cette quantité est minimale lorsque $\theta_j = g_j$, le centre de gravité des individus dans P_j. La log-vraisemblance est donc maximisée lorsque P est la partition qui minimise

$$\sum_{j=1}^{K}\sum_{e_i \in P_j} \|e_i - g_j\|_2^2,$$

correspondant à la partition qui résout le problème des k-means, et $\theta_j = g_j$ pour tout j. ■

Encore une fois, on utilise un algorithme renvoyant une solution approximative. Cette algorithme est similaire à celui utilisé pour les k-means et les k-médoïdes et fonctionne de la même façon.

Algorithme 4 : Classification EM

Entrées : $K \geqslant 0$, $e_1, \cdots, e_n \in \Omega$

1 Initialiser les θ_j par des valeurs choisies au hasard.
2 Affecter à P la partition $P_j = \{e_i \text{ t.q. } \log(f_{\theta_j}(e_i)) \geqslant \log(f_{\theta_k}(e_i)) \ \forall k \neq j\}$.
3 Affecter à chaque θ_i la valeur qui minimise la fonction $\theta \mapsto \sum_{e_i \in P_j} \log(f_\theta(e_i))$.
4 Recommencer les étapes 2 et 3 jusqu'à ce que l'algorithme converge puis terminer et renvoyer P et les θ_i.

Remarques:

- Dans le cas où f_{θ_j} est la densité d'une loi $\mathcal{N}(\theta_j, \sigma^2 I_p)$ alors l'algorithme de Classification EM est identique à l'algorithme des moyennes mobiles.
- En général on n'a pas forcement convergence de l'algorithme si on n'a pas de formule théorique pour le minimum des fonctions $\theta \mapsto \sum_{e_i \in P_j} \log(f_\theta(e_i))$. Du coup, on utilise des algorithmes d'optimisation qui peuvent potentiellement coûter cher en temps de calcul.

2 Approche estimation

On considère dans cette approche que la classe de chaque individu est choisie aléatoirement et que les données de la j-ème classe ont été générées selon une loi de densité f_{θ_j}. On note Z la variable aléatoire à valeur dans $\{1, \cdots, K\}$ correspondant au choix de la classe et $\pi_i = \mathbb{P}(Z = i)$, la probabilité d'appartenir à la classe i. On définit ensuite X une variable aléatoire telle que la loi de X sachant que $Z = i$ est de densité f_{θ_i} et on considère que les données observées sont issues de variables i.i.d. de même loi que X. La densité de X s'écrit alors

$$f_X(x) = \sum_{i=1}^{K} f_{X|Z=i}(x)\mathbb{P}(Z = i) = \sum_{i=1}^{K} \pi_i f_{\theta_i}(x).$$

C'est ce qu'on appelle un **mélange de lois**. La log-vraisemblance du modèle s'écrit alors

$$\mathcal{L}(\theta_1, \cdots, \theta_K, \pi_1, \cdots, \pi_K) = \sum_{i=1}^{n} \log(f_X(e_i)) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{K} \pi_j f_{\theta_j}(e_i)\right).$$

De plus, par la formule de Bayes on a

$$\mathbb{P}_{\theta,\pi}(Z = j | X = x) = \frac{\mathbb{P}(Z = j) f_{X|Z=j}(x)}{f_X(x)} = \frac{\pi_j f_{\theta_j}(x)}{\sum_{l=1}^{K} \pi_l f_{\theta_l}(x)}.$$

On note alors $p_{i,j} = \mathbb{P}_{\theta,\pi}(Z = j | X = e_i)$ la probabilité que l'individu i appartienne à la classe j. Une fois obtenu une estimation $\hat{\theta}_1, \cdots, \hat{\theta}_K, \hat{\pi}_1, \cdots, \hat{\pi}_k$ des paramètres, on estime alors $p_{i,j}$ par

$$\hat{p}_{i,j} = \frac{\hat{\pi}_j f_{\hat{\theta}_j}(e_i)}{\sum_{l=1}^{K} \hat{\pi}_l f_{\hat{\theta}_l}(e_i)}.$$

Pour obtenir une classification on affecte alors l'individu e_i à la classe P_j pour j qui maximise $\hat{p}_{i,j}$.

Afin d'estimer les paramètres du modèle, on remarque que l'on peut écrire pour n'importe quels estimateurs $\hat{\pi}_i$ et $\hat{\theta}_i$ des paramètres

$$\begin{aligned}
\mathcal{L}(\theta, \pi) &= \sum_{i=1}^{n} \log\left(\sum_{j=1}^{K} \pi_j f_{\theta_j}(e_i)\right) \\
&= \sum_{i=1}^{n} \sum_{l=1}^{K} \hat{p}_{i,l} \log\left(\sum_{j=1}^{K} \pi_j f_{\theta_j}(e_i)\right) \text{ car } \sum_{l=1}^{K} \hat{p}_{i,l} = 1 \text{ pour tout } i \\
&= \sum_{i=1}^{n} \sum_{l=1}^{K} \hat{p}_{i,l} \log\left(\frac{\pi_l f_{\theta_l}(e_i)}{p_{i,l}}\right) \\
&= \sum_{i=1}^{n} \sum_{l=1}^{K} \left(\hat{p}_{i,l} \log(\pi_l f_{\theta_l}(e_i)) - \hat{p}_{i,l} \log(p_{i,l})\right) \\
&= Q(\theta, \pi | \hat{\theta}, \hat{\pi}) - R(\theta, \pi | \hat{\theta}, \hat{\pi}).
\end{aligned}$$

Cette forme est utile car les fonctions Q et R vérifient les propriétés suivantes :

Proposition 63

- *Pour tout θ et π,*
$$R(\theta, \pi | \hat{\theta}, \hat{\pi}) \leqslant R(\hat{\theta}, \hat{\pi} | \hat{\theta}, \hat{\pi}).$$
- *On considère les $\widetilde{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{i,j}$ et les $\widetilde{\theta}_j$ qui maximisent la fonction*
$$\theta \mapsto \sum_{i=1}^{n} \hat{p}_{i,j} \log(f_\theta(e_i)).$$

Alors $(\widetilde{\theta}, \widetilde{\pi})$ maximise Q.

Démonstration : • On commence par écrire

$$R(\hat{\theta}, \hat{\pi} | \hat{\theta}, \hat{\pi}) - R(\theta, \pi | \hat{\theta}, \hat{\pi}) = \sum_{i=1}^{n} \sum_{l=1}^{K} \hat{p}_{i,l} \log(\hat{p}_{i,l}) - \hat{p}_{i,l} \log(p_{i,l}) = \sum_{i=1}^{n} \sum_{l=1}^{K} \hat{p}_{i,l} \log\left(\frac{\hat{p}_{i,l}}{p_{i,l}}\right).$$

En utilisant l'inégalité $\log(x) \geqslant 1 - 1/x$, vraie pour tout $x \geqslant 0$, on obtient

$$R(\hat{\theta}, \hat{\pi}|\hat{\theta}, \hat{\pi}) - R(\theta, \pi|\hat{\theta}, \hat{\pi}) \geqslant \sum_{i=1}^{n}\sum_{l=1}^{K}(\hat{p}_{i,l} - p_{i,l}) = n - n = 0$$

car

$$\forall i,\ \sum_{l=1}^{K}\hat{p}_{i,l} = \sum_{l=1}^{K} p_{i,l} = 1$$

- On a

$$Q(\theta, \pi|\hat{\theta}, \hat{\pi}) = \sum_{i=1}^{n}\sum_{l=1}^{K}\hat{p}_{i,l}\log(\pi_l) + \sum_{l=1}^{K}\sum_{i=1}^{n}\hat{p}_{i,l}\log(f_{\theta_l}(e_i))$$

donc Q s'écrit comme une somme de terme qui dépendent chacun d'un seul des paramètres. On en déduit alors que Q est maximisé pour les θ_j qui maximisent la fonction

$$\theta \mapsto \sum_{i=1}^{n}\hat{p}_{i,j}\log(f_\theta(e_i))$$

pour tout j et les $\pi_1, \cdots, \pi_n$ qui maximisent

$$(\pi_1, \cdots, \pi_n) \mapsto \sum_{i=1}^{n}\sum_{l=1}^{K}\hat{p}_{i,l}\log(\pi_l)$$

sous la contrainte que $\pi_j \geqslant 0$ pour tout j et $\sum_{j=1}^{K}\pi_j = 1$. Si on note $\widetilde{\pi}_1, \cdots, \widetilde{\pi}_n$ un tel maximum alors en utilisant la méthode des multiplicateurs de Lagrange on en déduit qu'il existe $\lambda \in \mathbb{R}$ tel que

$$\forall j,\ \sum_{i=1}^{n}\frac{\hat{p}_{i,j}}{\widetilde{\pi}_j} = \lambda.$$

Comme $\sum_{j=1}^{K}\hat{p}_{i,j} = 1$ pour tout i on en déduit

$$n = \sum_{j=1}^{K}\sum_{i=1}^{n}\hat{p}_{i,j} = \lambda\sum_{j=1}^{K}\widetilde{\pi}_j = \lambda$$

d'où $\widetilde{\pi}_j = \frac{1}{n}\sum_{i=1}^{n}\hat{p}_{i,j}$. ■

Corollaire 64

$$\mathcal{L}(\widetilde{\theta}, \widetilde{\pi}) \geqslant \mathcal{L}(\hat{\theta}, \hat{\pi}).$$

Autrement dit, $(\widetilde{\theta}, \widetilde{\pi})$ est un meilleur estimateur que $(\hat{\theta}, \hat{\pi})$

Il en vient l'algorithme suivant

Algorithme 5 : Algorithme Espérance-Maximisation (EM)

Entrées : $K \geqslant 0$, $e_1, \cdots, e_n \in \Omega$

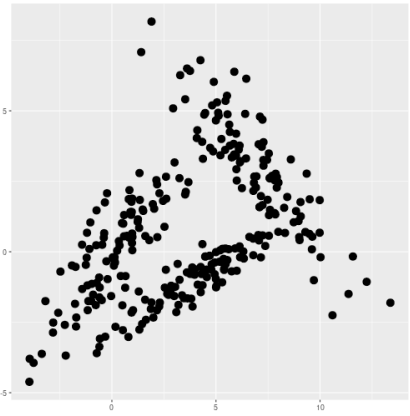
1 Initialiser les $\hat{\theta}_i$ et les $\hat{\pi}_i$.

2 Calculer les quantités $\hat{p}_{i,j} = \mathbb{P}_{\hat{\theta},\hat{\pi}}(Z = j|X = e_i) = \mathbb{E}_{\hat{\theta},\hat{\pi}}[\mathbb{1}_{Z=j}|X = e_i]$ (Espérance)

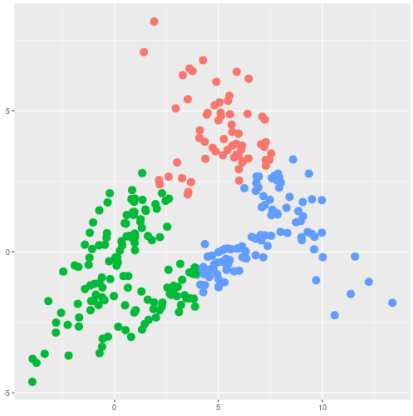
3 Affecter à chaque $\hat{\pi}_j$ la valeur $\frac{1}{n}\sum_{i=1}^{n}\hat{p}_{i,j}$ et à chaque $\hat{\theta}_j$ la valeur qui maximise la fonction $\theta \mapsto \sum_{i=1}^{n}\hat{p}_{i,j}\log(f_\theta(e_i))$ (Maximisation)

4 Recommencer les étapes 2 et 3 jusqu'à ce que l'algorithme converge et renvoyer $\hat{\theta}_1, \cdots, \hat{\theta}_K, \hat{\pi}_1, \cdots, \hat{\pi}_k$.

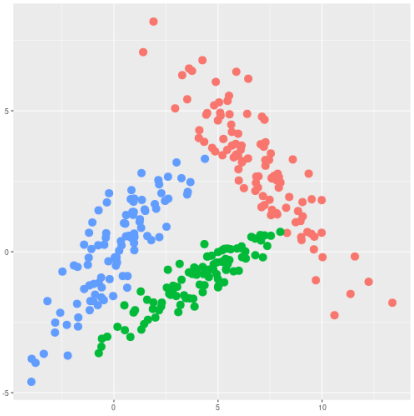
Exemple: L'une des grandes forces de l'algorithme EM est de pouvoir séparer des classes avec des formes ovales. On donne ci-dessous un exemple d'un tel jeu de données avec une classification par k-means et une classification avec l'algo EM en prenant des distributions normales.



(a) Jeu de données

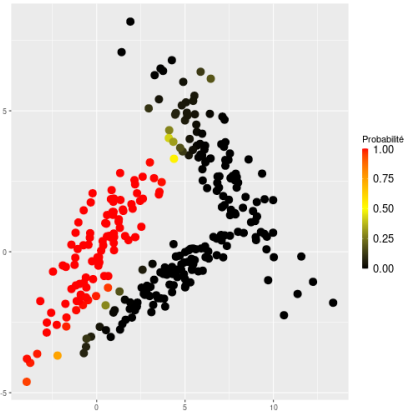


(b) Partitionnement en k-means

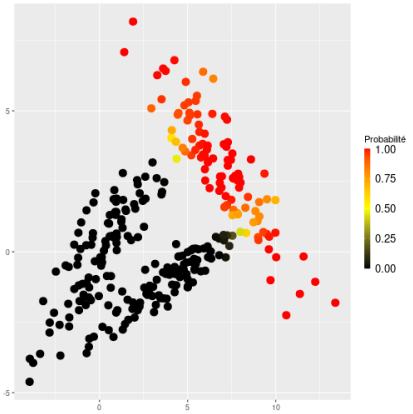


(c) Partitionnement avec l'algorithme EM

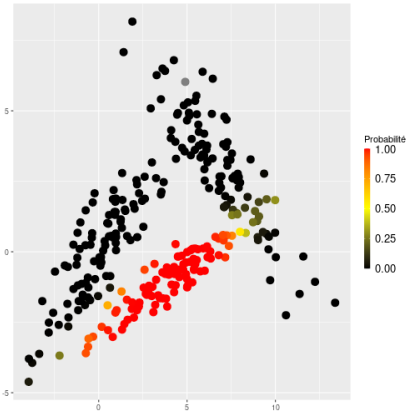
En particulier, l'algorithme EM renvoie une probabilité appartenance à chaque classe pour les individus ce qui permet notamment d'identifier les individus bien classés de ceux à cheval entre plusieurs classes.



(d) Probabilités d'appartenance à la première classe



(e) Probabilités d'appartenance à la deuxième classe



(f) Probabilités d'appartenance à la troisième classe

FIGURE 2.9 – Probabilités d'appartenance à chaque classe.

3 Choix du nombre de classe avec un critère d'information

Soit d_K la dimension de l'espace dans lequel est défini $(\theta_1, \cdots, \theta_K)$. On note alors $\nu_K = d_K + K - 1$ la dimension de l'espace dans lequel est défini $(\theta_1, \cdots, \theta_K, \pi_1, \cdots, \pi_K)$.

Définition 65

Soient $\hat{\theta}_1, \cdots, \hat{\theta}_K, \hat{\pi}_1, \cdots, \hat{\pi}_K$ une estimation des paramètres pour un modèle avec K classes.

- *On appelle* ***critère d'information d'Akaïke*** *la quantité*

$$AIC(K) = -2\mathcal{L}(\hat{\theta}, \hat{\pi}) + 2\nu_K$$

- *On appelle* ***critère d'information Bayésien*** *la quantité*

$$BIC(K) = -2\mathcal{L}(\hat{\theta}, \hat{\pi}) + \nu_K \log(n)$$

L'idée de ces critères d'information est de maximiser $\mathcal{L}(\hat{\theta}, \hat{\pi})$ mais en pénalisant l'utilisation d'un nombre de paramètre élevé (et donc un ν_K élevé). On les verra plus en détail dans le cours sur les modèles linéaires au second semestre. On choisit le nombre de classes en prenant K qui minimise l'un des deux critères d'information.

Exemples:

- Si f_{θ_i} est la densité d'un vecteur Gaussien dans $\mathbb{R}^p$ de loi $\mathcal{N}(\mu_i, \sigma^2 I_d)$ avec $\mu_i \in \mathbb{R}^p$ et $\sigma^2 \in \mathbb{R}_+^*$ non connu alors les paramètres sont $(\mu_1, \cdots, \mu_K, \sigma^2) \in (\mathbb{R}^p)^K \times \mathbb{R}_+^*$ d'où $d_K = pK + 1$.
- Si f_{θ_i} est la densité d'un vecteur Gaussien dans $\mathbb{R}^p$ de loi $\mathcal{N}_p(\mu_i, \Sigma_i)$ avec $\mu_i \in \mathbb{R}^p$ et $\Sigma_i \in \mathcal{S}_p^{++}(\mathbb{R})$ alors les paramètres sont $(\mu_1, \cdots, \mu_K, \Sigma_1, \cdots, \Sigma_K) \in (\mathbb{R}^p)^K \times \mathcal{S}_p^{++}(\mathbb{R})^K$. Comme une matrice symétrique de taille $p \times p$ est caractérisée par ses termes diagonaux et triangulaires inférieurs alors elle est définie par $\frac{p(p+1)}{2}$ paramètres d'où $d_K = dK + \frac{p(p+1)K}{2} = pK\frac{p+3}{2}$.

Bibliographie

[1] A.J. Izenman. *Modern Multivariate Statistical Techniques : Regression, Classification, and Manifold Learning*. Springer Publishing Company, Incorporated, 1 edition, 2008.

[2] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy. *Classification automatique des données. Environnement statistique et informatique.* Dunod Informatique, Paris, 1989.

[3] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

Chapitre 3

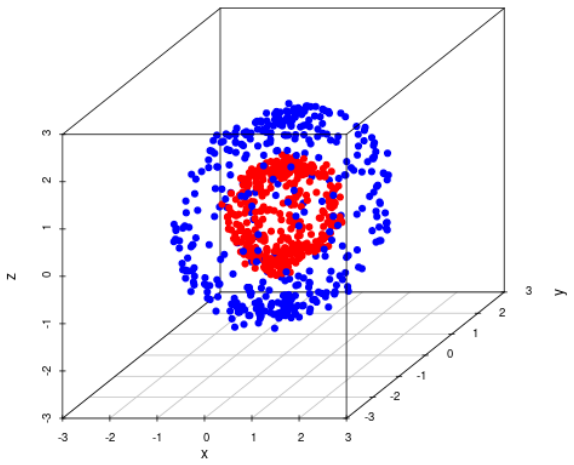
Compléments

I Méthodes non linéaires

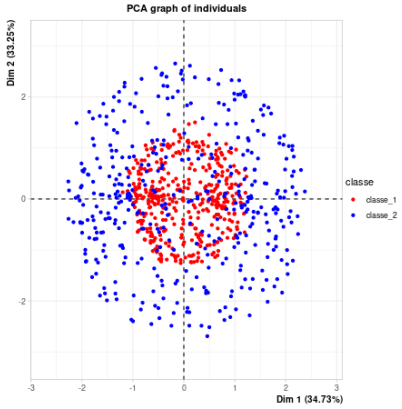
1 L'astuce du noyau

On considère un jeu de données X de n individus et p variables quantitatives.

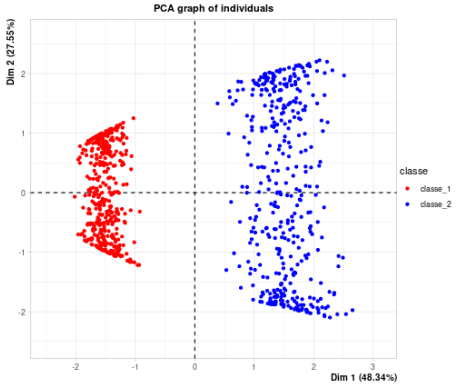
On rappelle que l'ACP cherche à projeter les données sur un sous-espace de sorte à conserver au maximum les distances entre les individus. Une projection étant une application linéaire cette méthode applique donc une transformation linéaire des données. Néanmoins, il y a certain types de données pas adaptés pour des méthodes linéaires.
Exemple: On considère le jeu de données suivant de 800 individus en dimension $p = 3$.



Ici les individus ont deux comportements très différents, néanmoins il n'est pas possible de projeter ces données en arrivant à bien séparer ces deux comportements. Par exemple, une ACP à deux dimensions donne le résultat suivant.



Néanmoins, si on rajoute à chaque individu $e_i = (x_{i,1}, x_{i,2}, x_{i,3})$ une quatrième variable $x_{i,4} = x_{i,1}^2 + x_{i,2}^2 + x_{i,3}^2$ et qu'on applique une ACP non normalisée sur les données alors on obtient le résultat suivant qui sépare bien les deux comportements qu'on observe dans les données.



L'idée principale des méthodes non-linéaires est donc de rajouter des variables en plus aux données comme fonction non-linéaire des variables déjà existantes afin d'arriver à mieux séparer les données.

On pose $\phi : \mathbb{R}^p \to \mathbb{R}^q$, avec $p < q$, une fonction qui envoie chaque individu dans un espace plus grand. On cherche donc à appliquer l'ACP sur le jeu de données

$$\phi(X) = \begin{pmatrix} \phi(e_1) \\ \vdots \\ \phi(e_n) \end{pmatrix}.$$

On pose la fonction

$$K(x, y) = \langle \phi(x), \phi(y) \rangle.$$

On remarque que

$$\|\phi(e_i)\|^2 = \langle \phi(e_i), \phi(e_i) \rangle = K(e_i, e_i)$$
et $\|\phi(e_i) - \phi(e_j)\|^2 = \|\phi(e_i)\|^2 + \|\phi(e_j)\|^2 - 2\langle \phi(e_i), \phi(e_j) \rangle = K(e_i, e_i) + K(e_i, e_i) - 2K(e_i, e_j)$.

Les distances entre toutes les individus peuvent donc s'exprimer en fonction de K et on n'a pas besoin de connaître les ϕ. De même, si on regarde le centre de gravité des données

$$g_\phi = \frac{1}{n} \sum_{i=1}^{n} \phi(e_i)$$

alors

$$\|g_\phi\|^2 = \left\langle \frac{1}{n}\sum_{i=1}^n \phi(e_i), \frac{1}{n}\sum_{j=1}^n \phi(e_j) \right\rangle = \frac{1}{n^2}\sum_{i,j=1}^n \langle \phi(e_i), \phi(e_j)\rangle = \frac{1}{n^2}\sum_{i,j=1}^n K(e_i, e_j)$$

et

$$\langle \phi(e_i), g_\phi\rangle = \left\langle \phi(e_i), \frac{1}{n}\sum_{j=1}^n \phi(e_j) \right\rangle = \frac{1}{n}\sum_{j=1}^n \langle \phi(e_i), \phi(e_j)\rangle = \frac{1}{n}\sum_{j=1}^n K(e_i, e_j)$$

donc

$$\begin{aligned}\langle \phi(e_i) - g_\phi, \phi(e_j) - g_\phi\rangle &= \langle \phi(e_i), \phi(e_j)\rangle + \|g\|^2 - \langle \phi(e_i), g\rangle - \langle g, \phi(e_j)\rangle \\ &= K(e_i, e_j) + \frac{1}{n^2}\sum_{k,l=1}^n K(e_k, e_l) - \frac{1}{n}\sum_{l=1}^n K(e_i, e_l) - \frac{1}{n}\sum_{k=1}^n K(e_k, e_j).\end{aligned}$$

En particulier, toutes les distances au centre de gravité peuvent donc aussi s'écrire en fonction de K et on n'a pas besoin de connaître ϕ pour faire une ACP.

Théorème 66 (*Théoreme de Mercer*)

Soit $K \in \mathbb{L}^2(\mathbb{R}^p \times \mathbb{R}^p, \mathbb{R})$ une fonction continue, symétrique ($K(x,y) = K(y,x)$) et de type positif, c'est à dire que

$$\sum_{i=1}^N \sum_{j=1}^N a_i K(x_i, x_j) a_j \geqslant 0$$

pour tout $N \in \mathbb{N}$, $a_1, \cdots, a_N \in \mathbb{R}$, $x_1, \cdots, x_n \in \mathbb{R}^p$. Alors il existe une fonction $\phi : \mathbb{R}^p \mapsto \mathbb{R}^{\mathbb{N}}$ telle que

$$K(x,y) = \langle \phi(x), \phi(y)\rangle.$$

*K est appelé un **noyau**.*

Exemple: Quelques exemples de noyau :

- Noyau linéaire : $K(x,y) = \langle x, y\rangle$
- Noyau polynomial : $K(x,y) = (1 + \langle x, y\rangle)^d$
- Noyau gaussien : $K(x,y) = e^{-\frac{\|y-x\|^2}{2\sigma^2}}$
- Noyau laplacien : $K(x,y) = e^{-\frac{\|y-x\|}{2\sigma^2}}$
- Noyau de Cauchy : $K(x,y) = \frac{1}{1+\frac{\|y-x\|^2}{\sigma^2}}$
- Noyau de Bessel : $K(x,y) = \frac{J_{\nu+1}(\sigma\|y-x\|)}{\|y-x\|^{-n(\nu+1)}}$

Pour le noyau linéaire la fonction ϕ associée est bien évidemment l'identité. Pour le noyau polynomial de degré 2 dans $\mathbb{R}^2$ on a

$$\begin{aligned}(1 + \langle x, y\rangle)^2 &= (1 + x_1y_1 + x_2y_2)^2 \\ &= 1 + 2x_1y_1 + 2x_2y_2 + x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 \\ &= \left\langle \begin{pmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix}, \begin{pmatrix} 1 \\ \sqrt{2}y_1 \\ \sqrt{2}y_2 \\ y_1^2 \\ y_2^2 \\ \sqrt{2}y_1y_2 \end{pmatrix} \right\rangle\end{aligned}$$

donc la fonction ϕ associée est

$$\phi((x_1, x_2)) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2).$$

2 L'ACP à noyau

Pour appliquer l'ACP non normalisé aux données $\phi(X)$ on rappelle que l'on commence par créer la matrice des données centrées $Y = \phi(X) - 1_n{}^t g_\phi$ puis on cherche à diagonaliser la matrice de covariance $C = \frac{1}{n}{}^t YY$. Lex axes principaux de l'ACP sont les vecteurs propres de C et les composantes principales sont les vecteurs propres de la matrice $\frac{1}{n}Y^tY$. Or, $(Y^tY)_{i,j} = \langle \phi(e_i) - g_\phi, \phi(e_j) - g_\phi \rangle$ qui ne dépend que de K. En particulier, si on pose la matrice de Graham

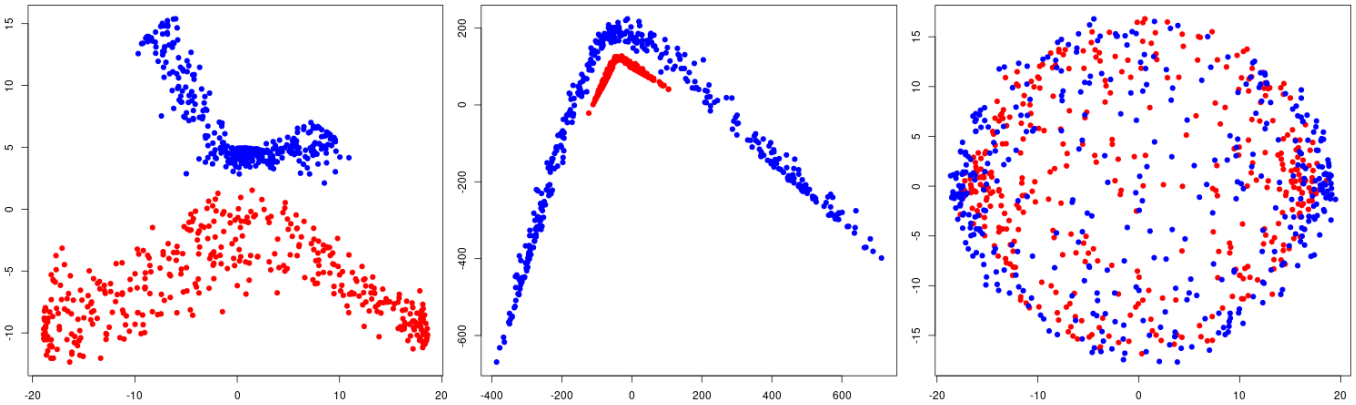
$$G = \begin{pmatrix} K(e_1,e_1) & \cdots & K(e_1,e_n) \\ \vdots & \ddots & \vdots \\ K(e_n,e_1) & \cdots & K(e_n,e_n) \end{pmatrix}$$

alors

$$Y^tY = G - \frac{1}{n}G1_{n\times n} - \frac{1}{n}1_{n\times n}G + \frac{1}{n^2}1_{n\times n}G1_{n\times n} \text{ où } 1_{n\times n} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

On peut donc calculer les composantes principales de l'ACP seulement en utilisant K et sans avoir besoin de préciser les ϕ.

Exemple: Si on reprend les données précédentes et qu'on applique l'ACP en utilisant trois noyaux différents on obtient les résultats suivants.



(a) Noyau Gaussien avec $\sigma^2 = 0.38$ (b) Noyau polynomial de degré 3 (c) Noyau de Bessel avec $n = \nu = 2$ et $\sigma = 1.3$

On peut voir que selon le noyau et les paramètres choisis on peut arriver à plus ou moins bien séparer les données avec des comportements différents. Le problème devint donc de choisir le noyau et les paramètres optimaux.

3 Le clustering spectral

A compléter

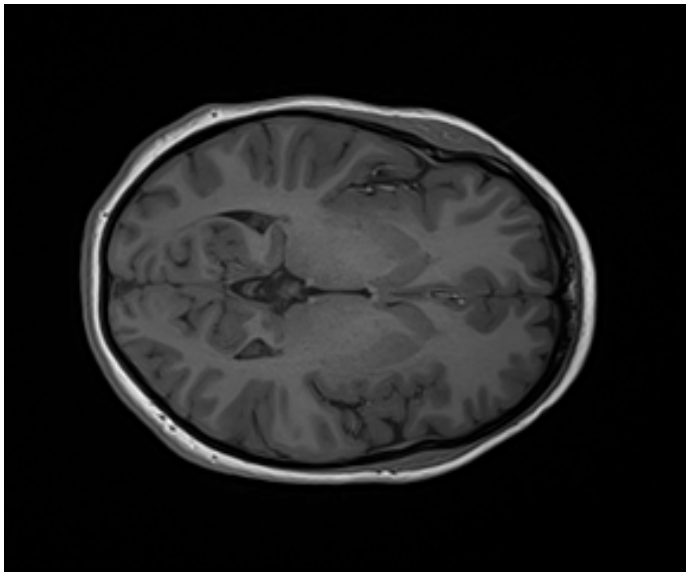
II Applications en machine learning

1 Application à des données d'image

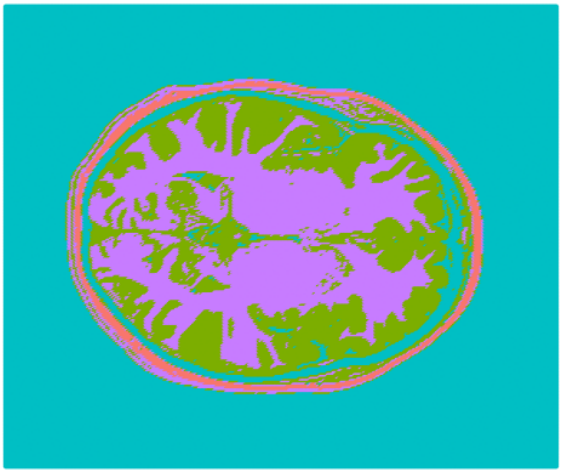
a Segmentation d'image

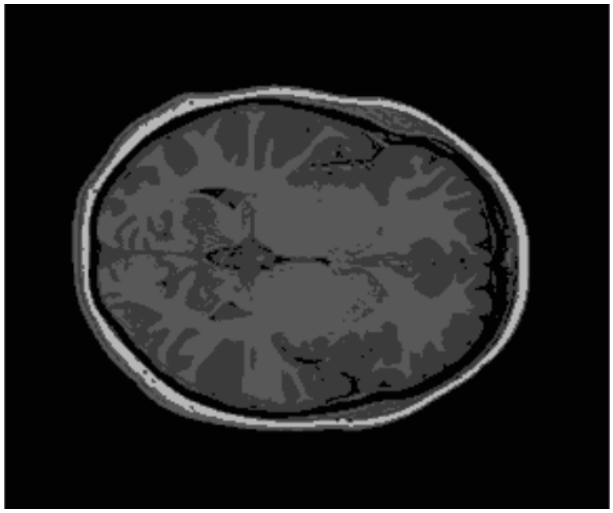
Un image en noir et blanc de $n \times m$ pixels peut être représentées par une matrice de $n \times m$ valeurs (ou un vecteur de nm valeurs) entre 0 et 1. Chaque valeur correspond au niveau de gris

d'un pixel avec 0 pour un pixel blanc et 1 pour un pixel noir. Comme exemple on considère une IRM d'un cerveau de taille 240×288 pixels :



Dans cette image, on souhaiterait isoler l'exterieur du crane (pixels noirs en dehors du crâne), le crâne (pixels blancs), le liquide céphalo-rachidien (pixels noirs à l'intérieur du crâne), la matière blanche (pixels gris clairs) et la matière grise (pixels gris foncés). Pour cela, on peut classifier les pixels avec les k-means selon leur valeurs. Dans cet exemple, cela revient à classifier $240 \times 288 = 69120$ individus de dimension 1. On donne les résultats ci-dessous.





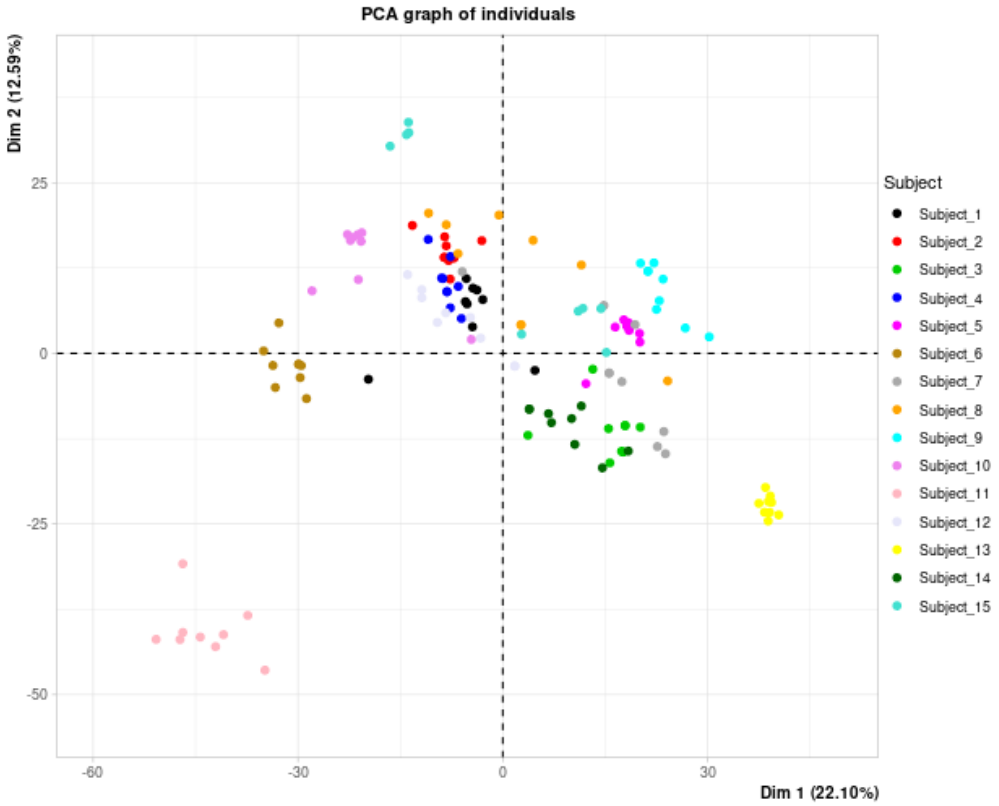
A gauche les pixels sont coloriés selon leur classe avec des couleurs arbitraires. A droite, les pixels sont coloriés selon leur classe avec la couleur correspondant au centre de gravité de leur classe.

b Classification d'image

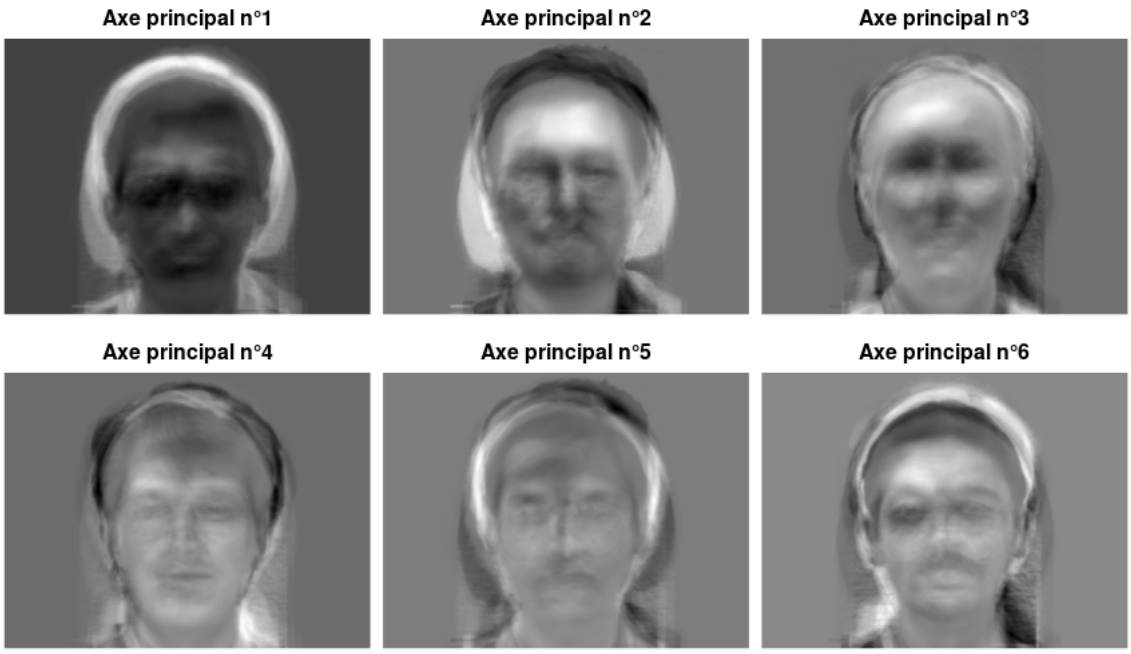
Un ensemble de k images en noir et blanc de même taille peut être représenté par un jeu de données X de taille $k \times nm$ où chaque ligne correspond à une image et chaque colonne correspond à un pixel.

Comme exemple on considère une jeu de données de 135 images de visages de 15 personnes différentes, chaque personne ayant 9 images.
(Source : https://www.kaggle.com/datasets/olgabelitskaya/yale-face-database)

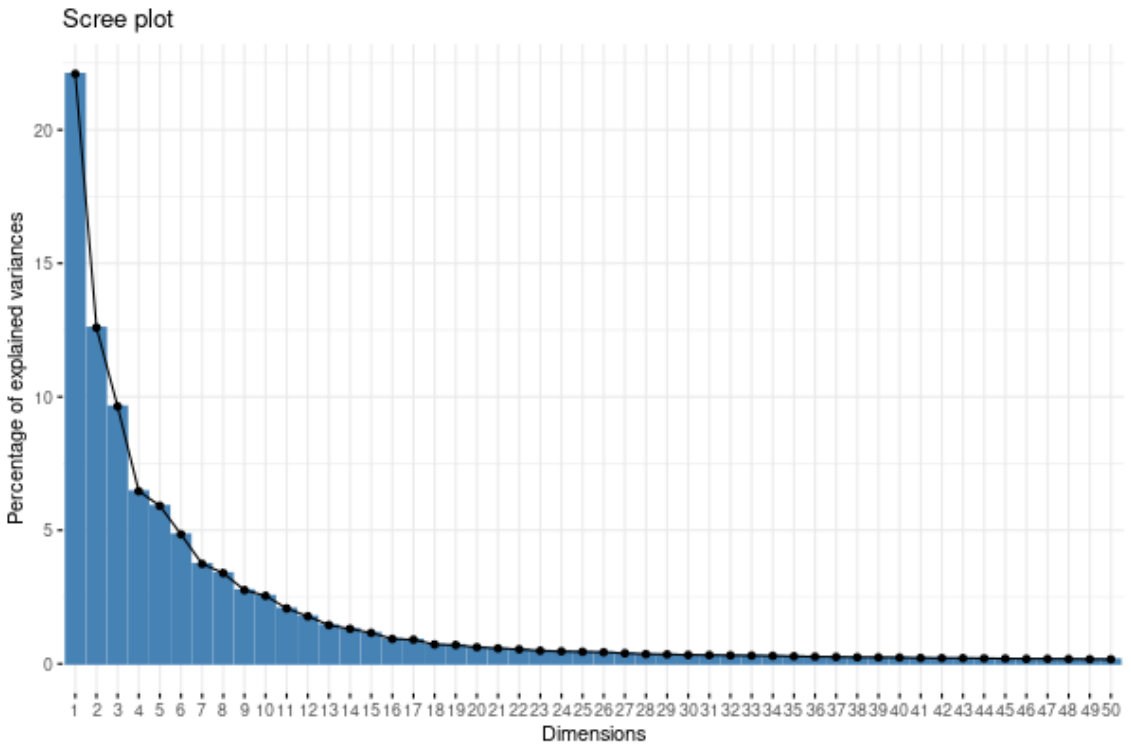
Les images sont de taille 241×320 ce qui donne un jeu de données de taille 135×77120. Appliquer l'ACP non normalisé sur ces données donne les résultats suivants.



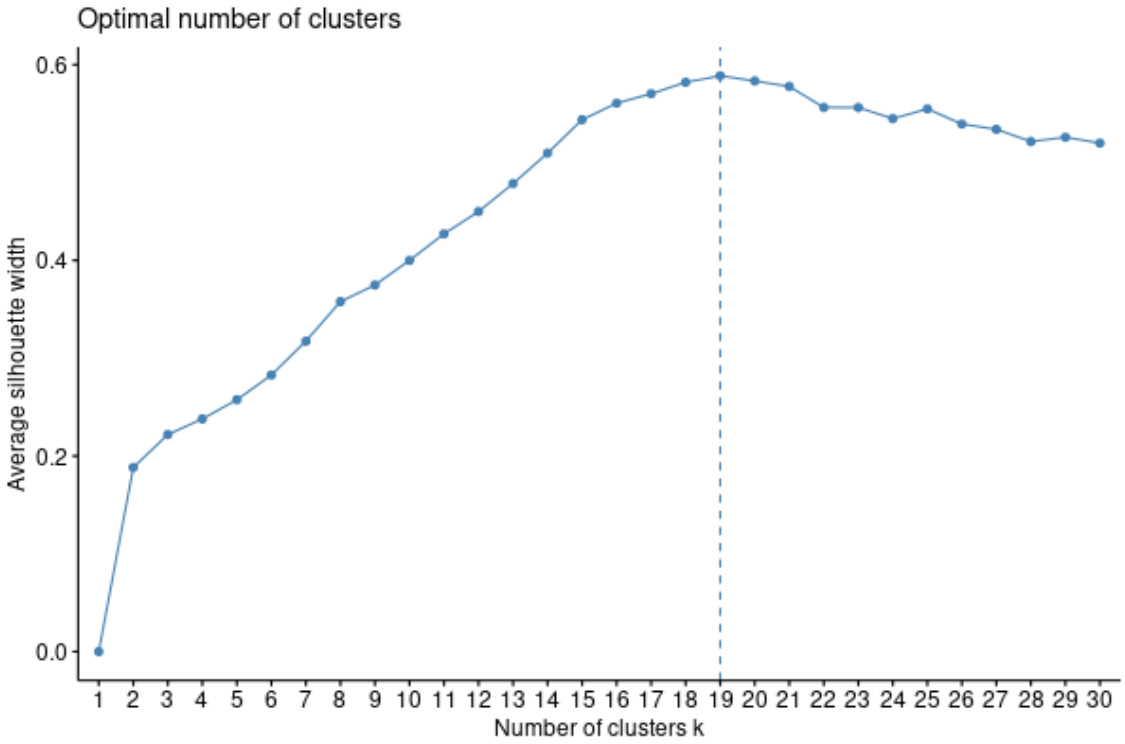
On observe que même en réduisant les 77120 variables à juste 2 on arrive déjà à garder quasiment 35% de l'information et on arrive même à distinguer les individus 6, 11 et 13 des autres. Cela suggère qu'en augmentant un peu le nombre de dimensions gardées on devrait pouvoir garder assez d'information pour bien séparer les individus tout en réduisant énormément la dimension du jeu de données. De plus, comme les axes principaux vivent dans le même espace que les individus, on peut alors essayer de les visualiser comme une image en regardant une carte de chaleur en niveau de gris de ces axes.



Afin de choisir le nombre de dimensions à garder on peut regarder la courbe de décroissance des valeurs propres et sélectionner un nombre de dimensions à partir duquel les valeurs propres ne décroissent quasiment plus.



Ici, on décide de garder 20 dimensions correspondant à un pourcentage d'inertie de 85.6% et c'est sur ces données projetées que l'on va appliquer l'ACP. Afin de sélectionner le nombre de classes on utilise la méthode des silhouettes qui va sélectionner 19 classes, un nombre assez proche de 15, le vrai nombre de personnes différentes dont les photos forme le jeu de données.



On donne ci-dessous le résultat de la classification en k-means dans 3 cas différents :

- Un cas où on fait la classification sur les 20 premières dimensions de l'ACP avec 19 classes, le nombre obtenu par la méthode des silhouettes.
- Un cas on fait la classification sur les 20 premières dimensions de l'ACP avec 15 classes, le nombre de classes correctes.
- Un cas on fait la classification sur les 50 premières dimensions de l'ACP avec 15 classes afin de voir si le fait d'augmenter la dimensions des données améliore le résultat de la classification.

FIGURE 3.1 – 19 classes et 20 dimensions

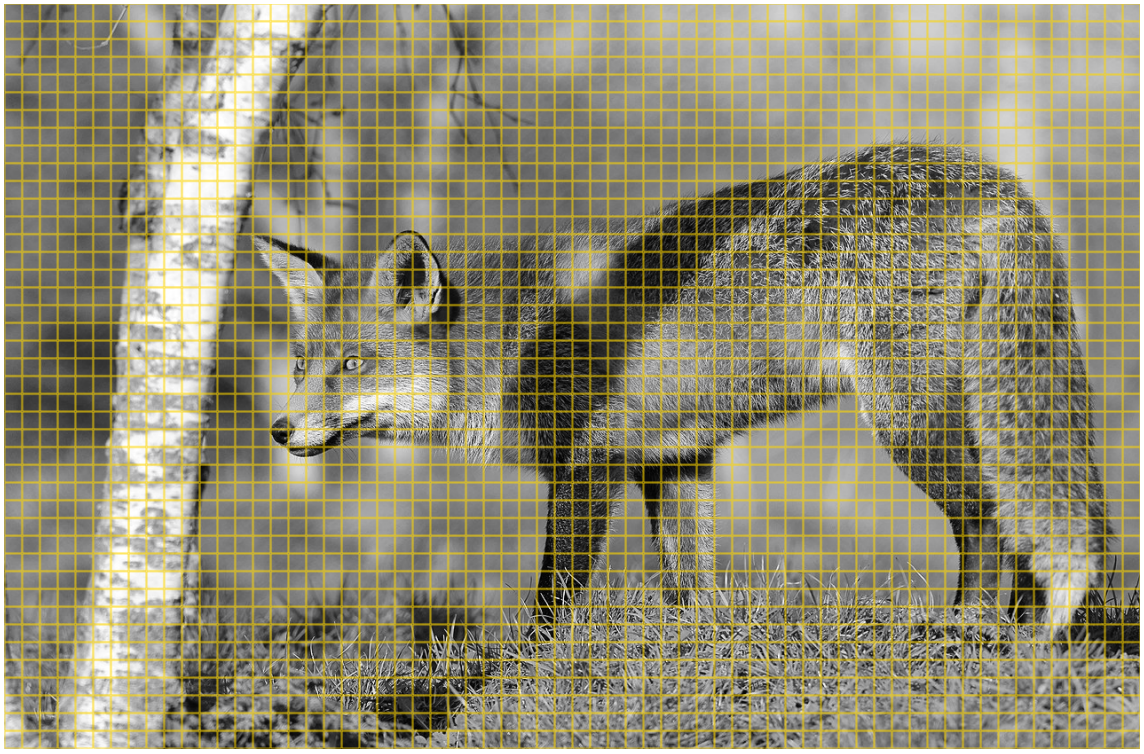
FIGURE 3.2 – 15 classes et 20 dimensions

FIGURE 3.3 – 15 classes et 50 dimensions

c Compression d'image

On s'intéresse maintenant à l'utilisation de l'ACP pour faire de la compression d'image. Afin de donner un exemple on considère l'image suivante de taille 1280×840 pixels.

On découpe l'image en $64 \times 42 = 2688$ mini-images (appelées des **patchs**) de taille 20×20.



Après conversion en vecteur on obtient un ensemble de 2688 vecteurs en dimension 400 que l'on peut voir comme une base de données de 2688 individus et 400 variables. Après avoir appliqué l'ACP à ce jeu de données on peut utiliser la formule de reconstruction suivante

Théorème 67 (*Eckart-Young*)

Soit $X \in \mathcal{M}_{n,p}(\mathbb{R})$. Pour une ACP à k dimensions, on note $u_1, \cdots u_k$ les axes principaux, $c_1, \cdots c_k$ les composantes principales et $\lambda_1, \cdots \lambda_k$ les valeurs propres associées. Alors, la

matrice

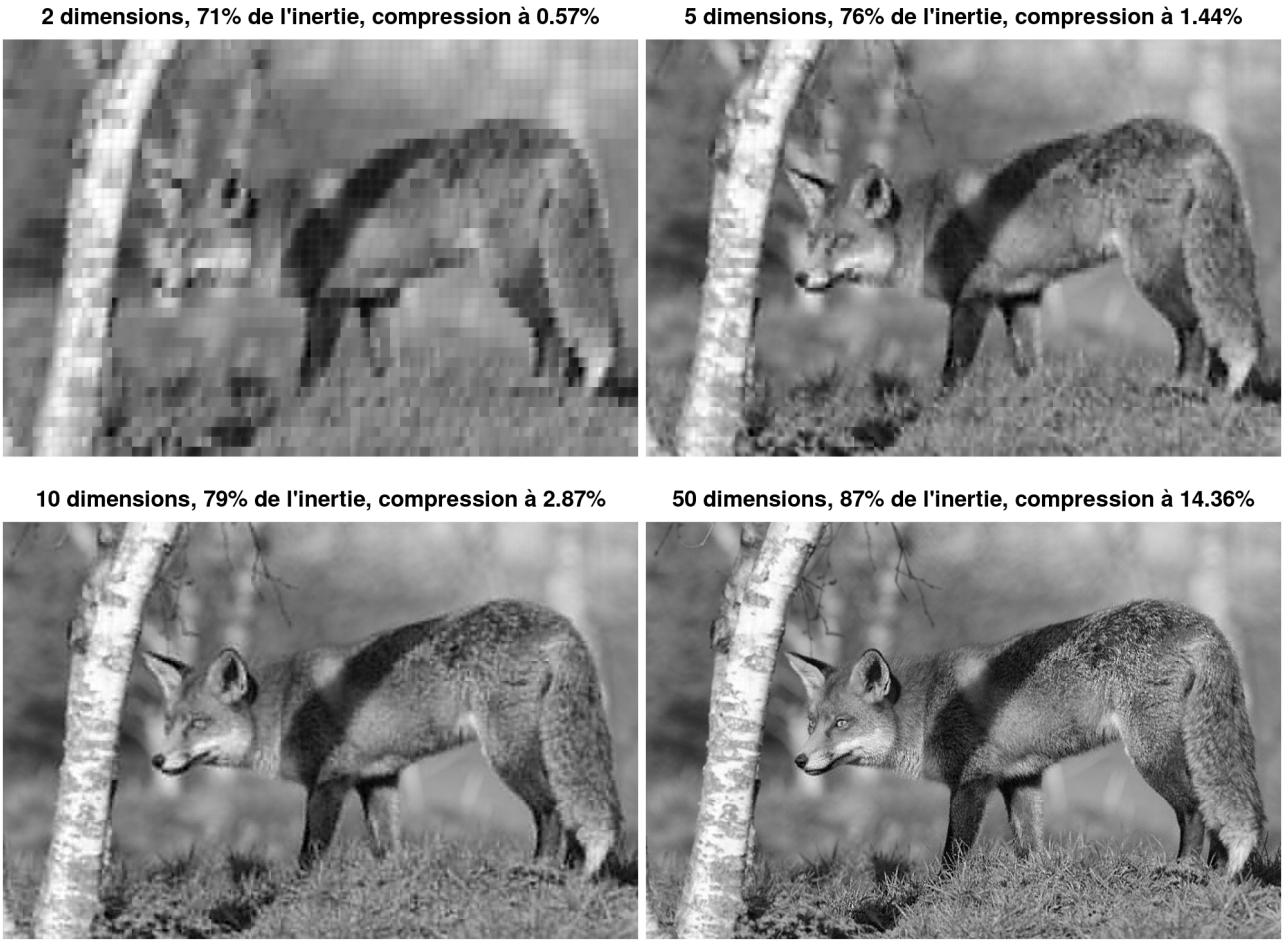
$$\widetilde{X} = \sum_{i=1}^{k} \sqrt{\lambda_i} c_i {}^t u_i$$

est la matrice qui minimise la quantité $\|X - Y\|_F$ *parmi toutes les matrices* Y *de rang* k *où* F *est la norme de Frobenius. Ce résultat reste vrai si on remplace la norme de Frobenius par la norme spectrale.*

On remarque que pour appliquer la formule de reconstruction on a besoin des c_i qui sont de dimension n, des u_i de dimension p et des λ_i de dimension 1. Vu que l'image complète possède $n \times p$ pixels on obtient alors un taux de compression pour une ACP à k dimensions de

$$\tau = \frac{k(n+p+1)}{np}.$$

Après avoir appliqué la formule de reconstruction on peut retransformer le jeu de données reconstruit en une image. Cela donne le résultat suivant.



2 Application à des données de texte

On s'intéresse au problème de la transformation d'un corpus de texte en un jeu de données afin de faire de la classification. Comme exemple on utilisera un corpus de 1490 morceaux d'articles de la BBC catégorisés comme étant des articles de business, de divertissement, de sport ou de tech.
(Source : https://www.kaggle.com/c/learn-ai-bbc)
Une méthode classique pour transformer un tel corpus de texte en un jeu de données est la suivante.

- Retirer toute la ponctuation des textes ainsi que les chiffres qui peuvent apparaître.
- Retirer les mots couramment utilisés (appelés les **stopwords**).
- Retirer les mots qui n'apparaissent que peu de fois dans l'ensemble des textes.

Une fois cela fait, on note n le nombre de documents dans le corpus de texte et p le nombre de mots différents apparaissant dans les textes après le pré-traitement décrit ci-dessus. On créé alors un jeu de données $X \in \mathcal{M}_{n,p}(\mathbb{R})$ tel que $x_{n,p}$ est une quantité donnant une information sur l'impact du p-ième mot sur le n-ième document. Une méthode courante est d'utiliser **le poids TfIdf** (Term Frequency - Inverse Document Frequency) :

Définition 68

*Soit $n_{i,j}$ le nombre de fois que le j-ème mot apparaît dans le i-ème document. On définit la **fréquence brut** du mot j dans le document i par*

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_j n_{i,j}}.$$

*On définit la **fréquence inverse de document** du mot j par*

$$\mathrm{idf}_j = \log\left(\frac{n}{\#\{i : n_{i,j} > 0\}}\right).$$

$\#\{i : n_{i,j} > 0\}$ correspond au nombre de documents contenant le mot j. Plus le nombre de document dans lesquels apparaît j est élevé, plus idf_j est proche de 0.
*Le **poids TfIdf** du mot j dans le texte i est alors définit par*

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j}\mathrm{idf}_j.$$

Dans l'exemple des articles de la BBC, après avoir appliqué cette méthode on se retrouve avec un jeu de données de taille 1490×4652. En appliquant une ACP à deux dimensions sur ce jeu de données on obtient le résultat suivant.

On voit que même avec juste 2 dimensions de gardées sur les 4652 et un pourcentage d'inertie d'à peine 1.1% on arrive assez bien à distinguer les divers classes. Cela suggère qu'il y a moyen en augmentant un peu la dimension de pouvoir avoir un jeu de données avec assez peu de variables mais qui permet de bien séparer les divers types de textes. En bonus, en inspectant la corrélation entre les variables (i.e. les mots) et les axes, on peut voir quels sont les mots qui arrivent au mieux à séparer les articles par catégorie.

government	said	people	market	growth	economy	economic	analysts
0.3120	0.2986	0.2784	0.2733	0.2710	0.2666	0.2616	0.2497

TABLE 3.1 – Mots les plus corrélés positivement avec l'axe 1

pop	dechy	mistake	reeves	doctor	places	roy	camp	grammy
-0.0632	-0.0632	-0.0630	-0.0630	-0.0628	-0.0628	-0.0627	-0.0625	-0.0625

TABLE 3.2 – Mots les plus corrélés négativement avec l'axe 1

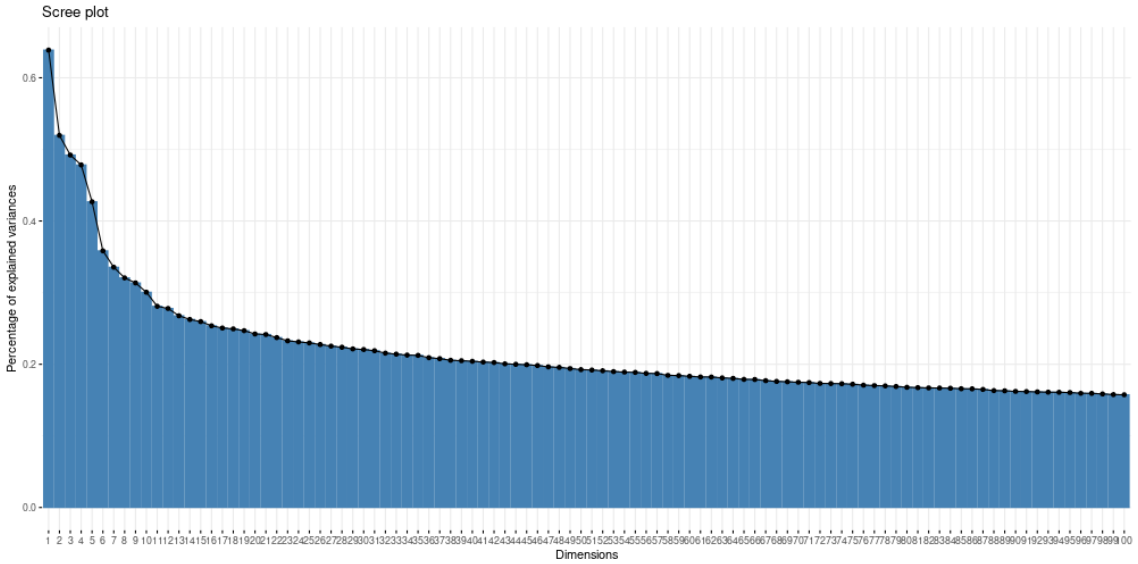
labour	tories	minister	government	election	blair	tory	leader
0.3302	0.3246	0.3199	0.3196	0.3180	0.3127	0.2990	0.2856

TABLE 3.3 – Mots les plus corrélés positivement avec l'axe 2

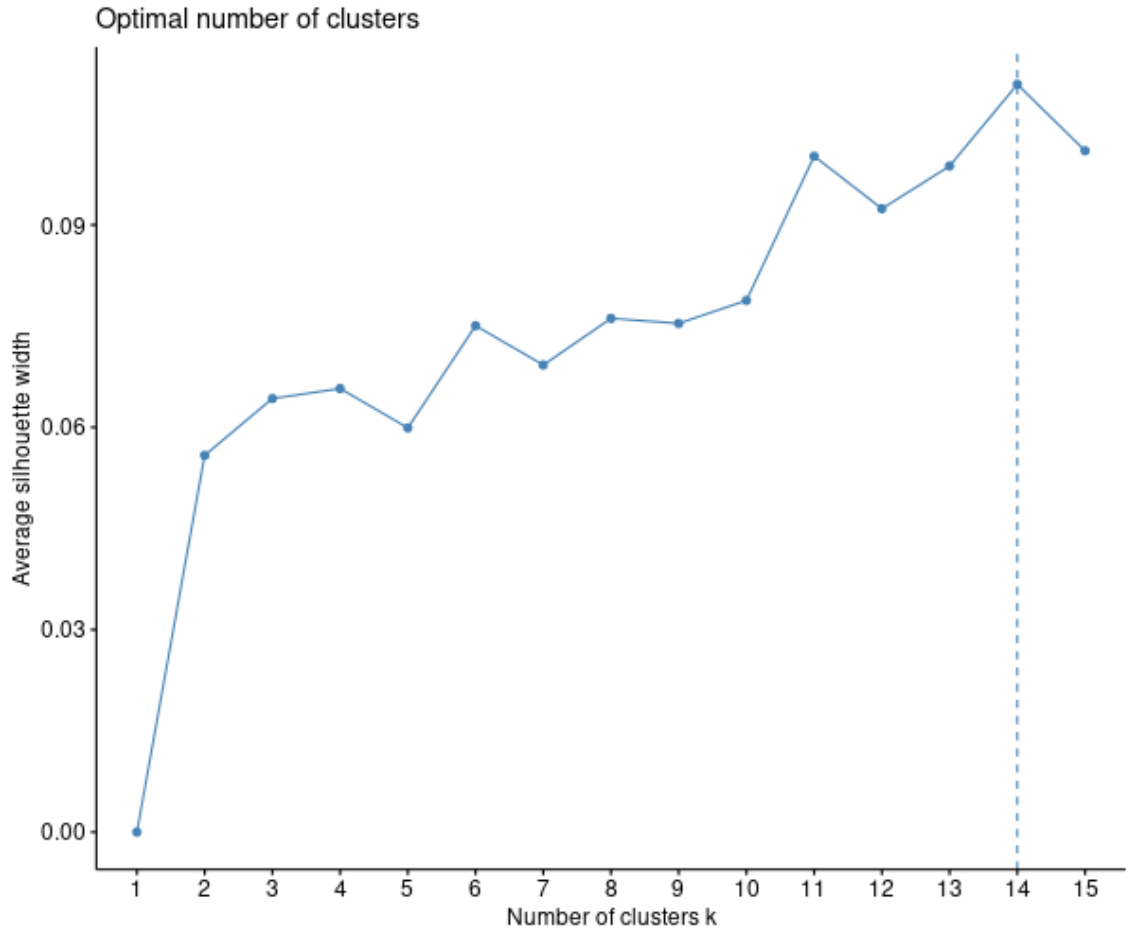
cinemas	connection	martin	firms	recording	interesting	dedicated	among	almost
-0.0929	-0.0929	-0.0929	-0.0929	-0.0929	-0.0928	-0.0926	-0.0925	-0.0922

TABLE 3.4 – Mots les plus corrélés négativement avec l'axe 2

On regarde maintenant le diagramme des valeurs propres afin de choisir le nombre de dimensions que l'on garde.



Dans ce cas-là, la décroissance des valeurs propres est très lente. J'ai décidé de garder les 50 premières dimensions. Même si le pourcentage d'inertie n'est que de 13%, au vu du nuage des individus sur les deux premières dimensions il y a bonne espoir que la classification fonctionne quand même assez bien. Malheureusement, le choix du nombre de classe par la méthode des silhouettes ne marche pas très bien.



C'est très probablement dû au fait que les individus sont très dispersés et il est difficile de bien identifier les classes de façon automatique. Si on triche un peu et qu'on applique les k-means avec 5 classes on obtient la classification suivante des articles.

	Classe n°1	Classe n°2	Classe n°3	Classe n°4	Classe n°5
business	325	0	4	7	0
entertainment	2	0	5	3	263
politics	10	1	3	258	2
sport	1	343	0	0	2
tech	4	4	239	7	7

On remarque que malgré les problèmes pointés précédemment, les k-means arrivent quand même à bien séparer la grande majorité des articles selon leur catégorie. Le pourcentage d'articles bien classés est de

$$\frac{325 + 343 + 239 + 258 + 263}{1490} \approx 95.83\%$$