Méthode des centres mobiles

Classification par partition - Les méthodes de réallocation

Ricco RAKOTOMALALA
Université Lumière Lyon 2

PLAN

1. Position du problème – La classification automatique
2. Algorithme K-Means – Méthode des centres mobiles
3. Cas des variables actives qualitatives
4. Fuzzy C-Means
5. Classification de variables
6. Conclusion
7. Bibliographie

La classification automatique

Constitution des groupes à partir des caractéristiques de proximité

Classification automatique

Typologie, apprentissage non-supervisé, clustering

Variables « actives », servent à la constitution des groupes.

Souvent (mais pas toujours) toutes quantitatives.

Modele	puissance	cylindree	vitesse	longueur	largeur	hauteur	poids	co2
PANDA	54	1108	150	354	159	154	860	135
TWINGO	60	1149	151	344	163	143	840	143
YARIS	65	998	155	364	166	150	880	134
CITRONC2	61	1124	158	367	166	147	932	141
CORSA	70	1248	165	384	165	144	1035	127
FIESTA	68	1399	164	392	168	144	1138	117
CLIO	100	1461	185	382	164	142	980	113
P1007	75	1360	165	374	169	161	1181	153
MODUS	113	1598	188	380	170	159	1170	163
MUSA	100	1910	179	399	170	169	1275	146
GOLF	75	1968	163	421	176	149	1217	143
MERC_A	140	1991	201	384	177	160	1340	141
AUDIA3	102	1595	185	421	177	143	1205	168
CITRONC4	138	1997	207	426	178	146	1381	142
AVENSIS	115	1995	195	463	176	148	1400	155
VECTRA	150	1910	217	460	180	146	1428	159
PASSAT	150	1781	221	471	175	147	1360	197
LAGUNA	165	1998	218	458	178	143	1320	196
MEGANECC	165	1998	225	436	178	141	1415	191
P407	136	1997	212	468	182	145	1415	194
P307CC	180	1997	225	435	176	143	1490	210
PTCRUISER	223	2429	200	429	171	154	1595	235
MONDEO	145	1999	215	474	194	143	1378	189
MAZDARX8	231	1308	235	443	177	134	1390	284
VELSATIS	150	2188	200	486	186	158	1735	188
CITRONC5	210	2496	230	475	178	148	1589	238
P607	204	2721	230	491	184	145	1723	223
MERC_E	204	3222	243	482	183	146	1735	183
ALFA 156	250	3179	250	443	175	141	1410	287
BMW530	231	2979	250	485	185	147	1495	231

Objectif de l'étude : Identifier les catégories (groupes) de voitures « similaires » (c.-à-d. qui se ressemblent au regard de leurs propriétés)

Objectif : identifier des groupes d'observations ayant des caractéristiques similaires (ex. comportement d'achats de clients, caractère « polluant » de véhicules, etc.)

On veut que :

(1) Les individus dans un même groupe se ressemblent le plus possible

(2) Les individus dans des groupes différents se démarquent le plus possible

Pourquoi ?

→ Identifier des structures sous-jacentes dans les données

→ Résumer des comportements

→ Affecter de nouveaux individus à des catégories

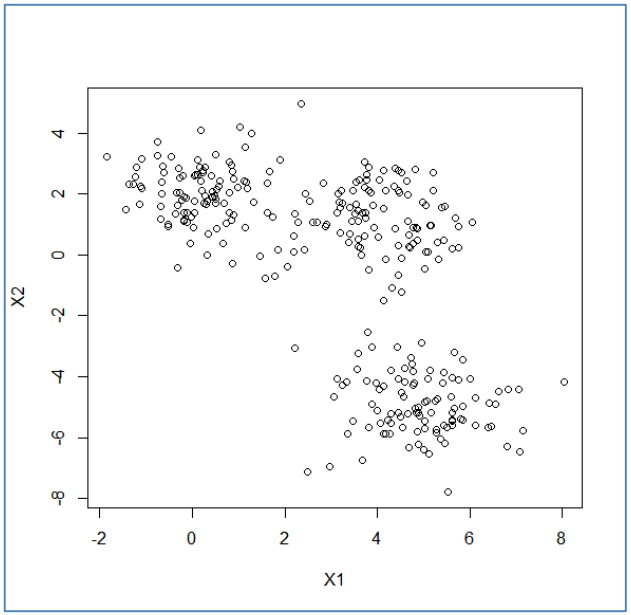
→ Identifier les cas totalement atypiques

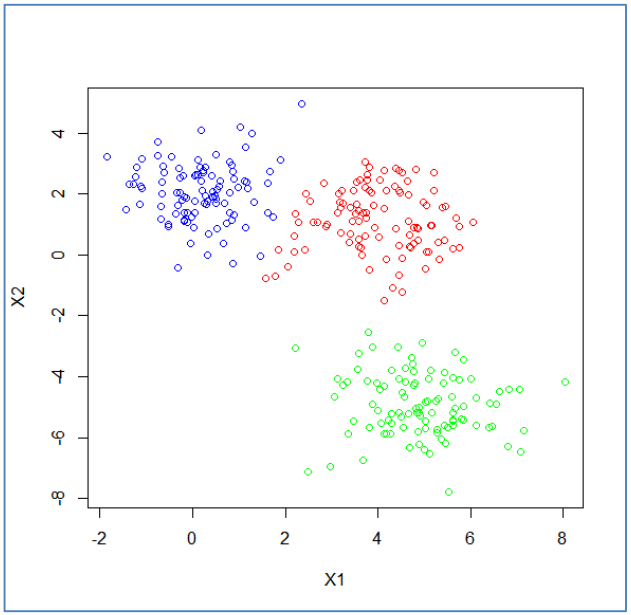
Classification automatique

Illustration graphique dans le plan

On « devine » les amas de points dans l'espace de représentation.

L'algorithme de classification automatique se charge de mettre en évidence les groupes « naturels » c.-à-d. qui se démarquent significativement les uns des autres.





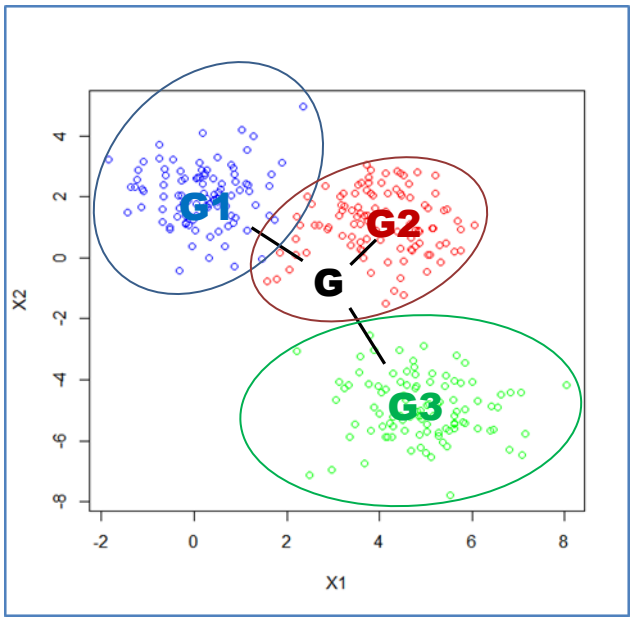
2 questions clés

1. Combien de groupes ?
2. Délimitation de ces groupes par le calcul

Caractérisation de la partition

Inertie intra-classes W

Donner un rôle crucial aux centres de classes



Remarque : les points étant rattachés à un groupe selon leur proximité avec le barycentre associé, les classes ont tendance à être convexes.

Relation fondamentale (Théorème d'Huygens)

$$\text{Inertie totale} = \text{Inertie inter - classes} + \text{Inertie intra - classe}$$

$$T = B + W$$

$$\sum_{i=1}^{n} d^2(i,G) = \sum_{k=1}^{K} n_k d^2(G_k,G) + \sum_{k=1}^{K}\sum_{i=1}^{n_k} d^2(i,G_k)$$

Dispersion des barycentres conditionnels autour du barycentre global. Indicateur de séparabilité des classes.

Dispersion à l'intérieur de chaque groupe. Indicateur de compacité des classes.

d() est une mesure de distance caractérisant les proximités entre les individus. Par ex. distance euclidienne ou euclidienne pondérée par l'inverse de la variance (*attention aux points aberrants*)

L'objectif de la classification automatique serait de minimiser l'inertie intra-classes W, à nombre de classes K fixé.

Classification par partition

Les méthodes de réallocation

Principales caractéristiques

- Fixer a priori le nombre de classes K
- Définir une partition de départ des données
- **Réallocation.** Déplacer les objets (observations) d'un groupe à l'autre pour obtenir une partition meilleure
- L'objectif (implicite ou explicite) est d'optimiser une mesure d'évaluation globale de la partition
- Fournit une partition unique des données

Mais peut être évolutive en fonction d'autres paramètres telle que le diamètre maximum des classes. Reste un problème ouvert souvent.

Souvent de manière aléatoire. Mais peut également démarrer à partir d'une autre méthode de partition ou s'appuyer sur des considérations de distances entre les individus (ex. les K individus les plus éloignés les uns des autres).

En faisant passer tous les individus, ou encore en tentant des échanges (plus ou moins) aléatoires entre les groupes.

La mesure W peut très bien faire office de fonction objectif.

On a une solution unique à K fixé. Et non pas une hiérarchie de partitions comme en CAH par ex.

Algorithme K-Means

Méthode des centres mobiles

Algorithme K-Means

Algorithme de Lloyd (1957), Forgy (1965), MacQueen (1967)

Algorithme particulièrement simple

Entrée : X (n obs., p variables), K #classes

Initialiser K centres de classes G_k

REPETER

Allocation. Affecter chaque individu à la classe dont le centre est le plus proche

Représentation. Recalculer les centres de classes à partir des individus rattachés

JUSQU'À Convergence

Sortie : Une partition des individus caractérisée par les K centres de classes Gk

Peut être K individus choisis au hasard. Ou encore, K moyennes calculées à partir d'une partition au hasard des individus en K groupes.

Variante MacQueen : remettre à jour les centres de classes à chaque individu traité. Accélère la convergence, mais le résultat dépend de l'ordre des individus.

Propriété fondamentale : l'inertie intra-classe diminue à chaque étape (nouvelles valeurs des barycentres conditionnels G_k)

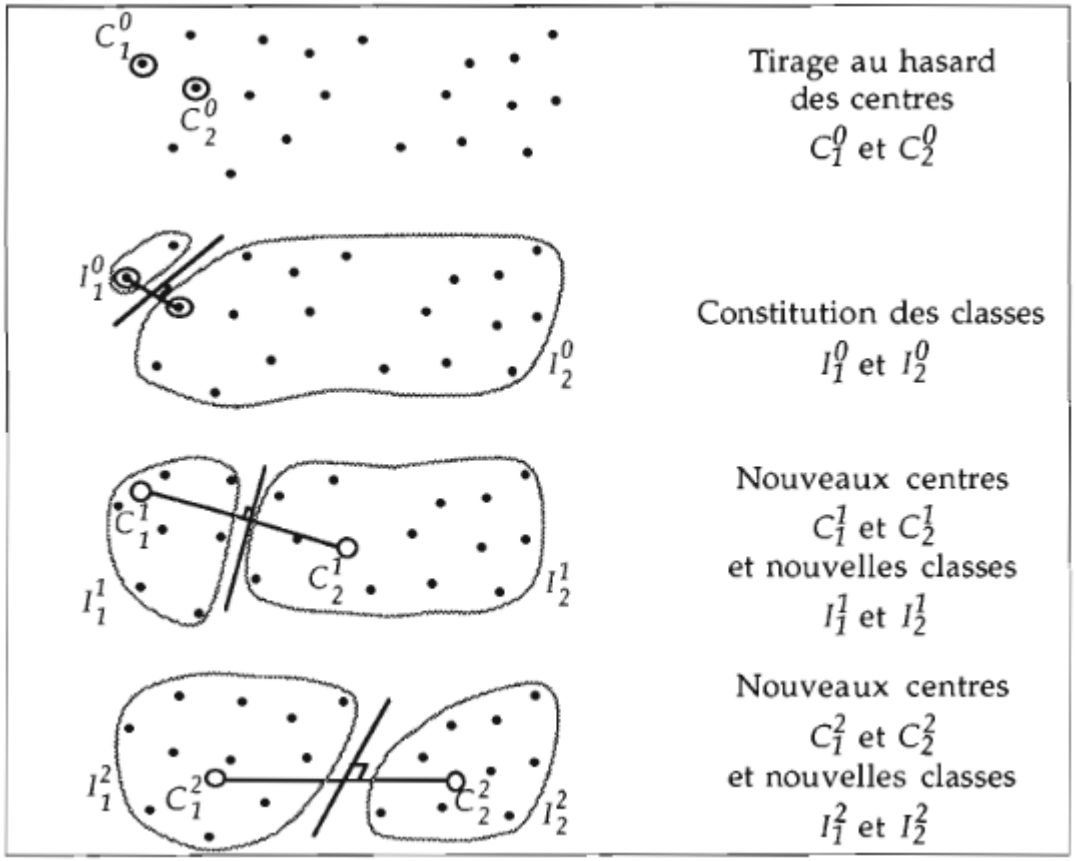
Nombre d'itérations fixé
Ou aucun individu ne change de classe
Ou encore lorsque W ne diminue plus
Ou lorsque les G_k sont stables

Le processus minimise implicitement l'inertie intra-classes W

(un réécriture sous forme d'optimisation explicite est possible. Cf. Gan et al., p. 163)

Algorithme K-Means

Un exemple



Lebart et al., 1995 ; page 149.

Algorithme K-Means

Avantages et inconvénients

Avantages

Scalabilité : Capacité à traiter les très grandes bases. Seuls les vecteurs des moyennes sont à conserver en mémoire centrale. Complexité linéaire par rapport au nombre d'observations (pas de calcul des distances deux à deux des individus, cf. CAH).

Inconvénients

Mais lenteur quand même parce que nécessité de faire passer plusieurs fois les observations.

L'optimisation aboutit à un minimum local de l'inertie intra-classes W.

La solution dépend du choix initial des centres de classes.

Essayer plusieurs configurations de départ et choisir celle qui aboutit à une solution minimisant W.

La solution peut dépendre de l'ordre des individus (MacQueen)

Mélanger aléatoirement les individus avant de les faire passer pour ne pas être dépendant d'une organisation non maîtrisée des observations.

Algorithme K-Means

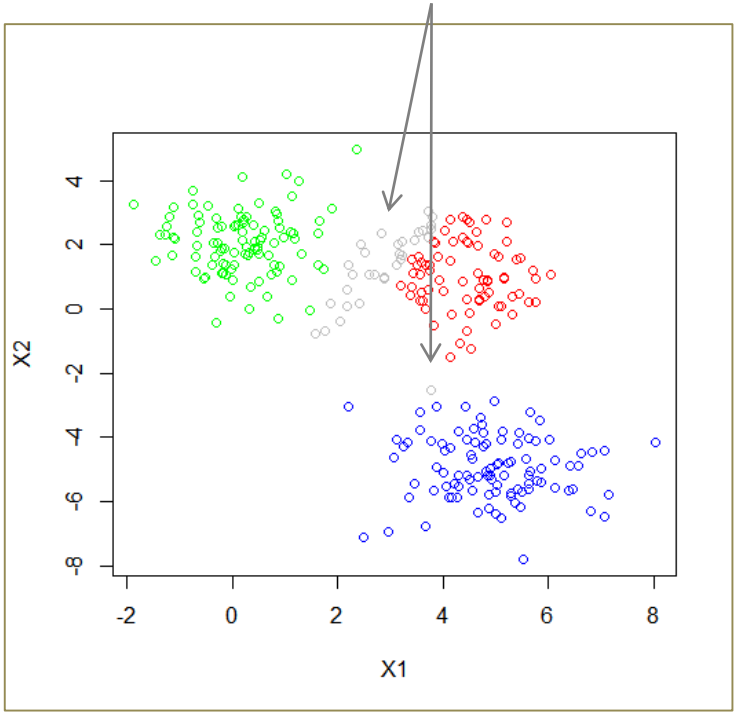
Notion de « formes fortes »

Deux (ou plusieurs) exécutions de l'algorithme sur les mêmes données peuvent aboutir à des solutions (légèrement) différentes. L'idée est de les croiser pour observer les regroupements stables, symptomatiques d'une véritable structuration des données c.-à-d. les formes fortes.

1ère exécution \ 2ème exécution	C1	C2	C3
C1	30	0	72
C2	0	99	1
C3	98	0	0

On observe les coïncidences entre les classes. C3 de la 1ère exécution correspond au C1 de la 2nde, etc.

Les zones d'indécisions (en gris) correspondent à des zones frontières entre les classes. « Formes faibles ».

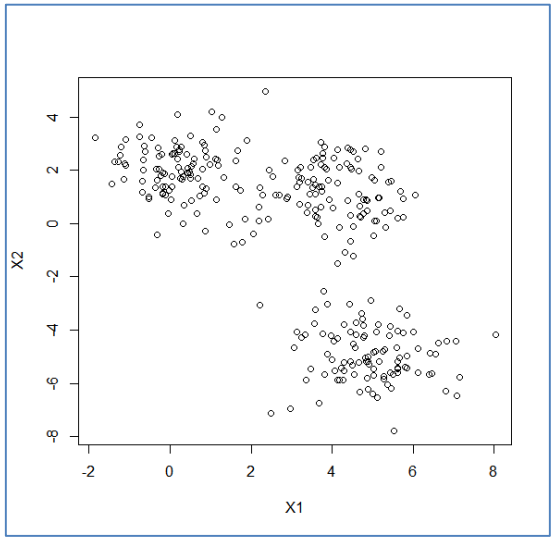


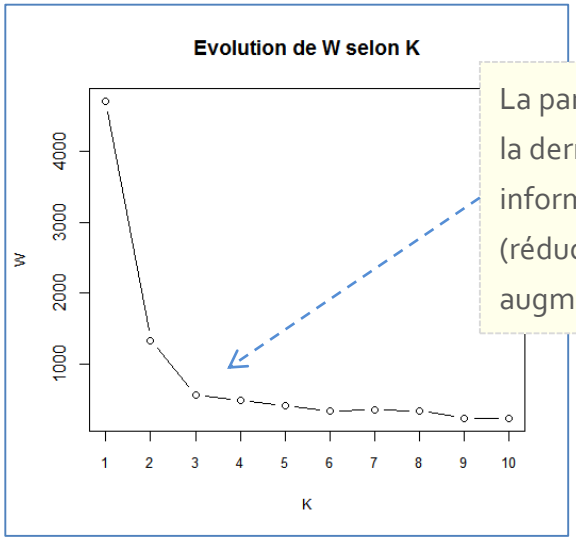
On peut multiplier les exécutions et les croisements, mais les calculs sont rapidement inextricables.

Algorithme K-Means

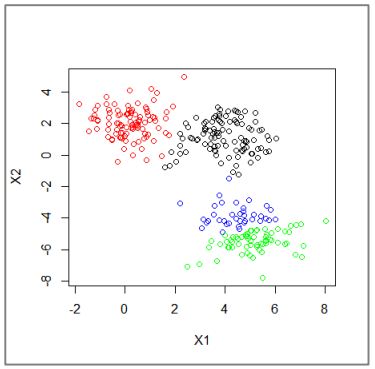
Détection du nombre de classes – Graphique de l'inertie intra-classes W

Principe : Une stratégie simple pour identifier le nombre de classes consiste à faire varier K et surveiller l'évolution de l'inertie intra-classes W. L'idée est de visualiser le « coude » où l'adjonction d'une classe ne correspond à rien dans la structuration des données.





La partition en K = 3 classes est la dernière à induire un gain informationnel significatif (réduction inertie intra ➔ augmentation de l'inertie inter)



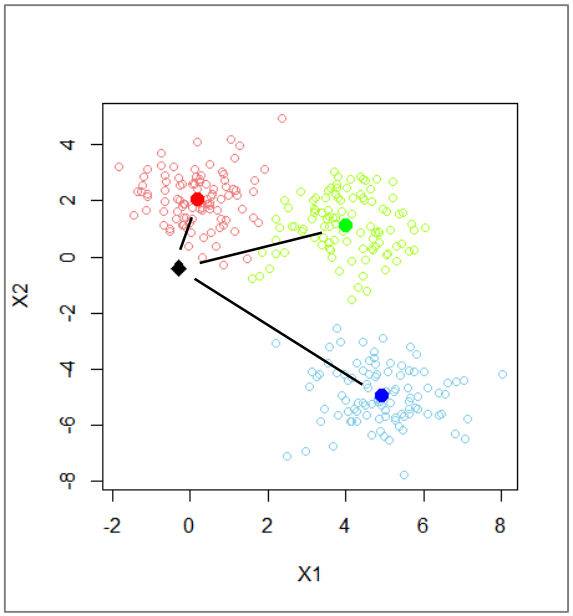
Solution en K = 4 classes, on se rend compte que la subdivision supplémentaire est artificielle.

Algorithme K-Means

Déploiement – Affectation des individus supplémentaires

Objectif : Rattacher un individu supplémentaire à une des classes.

Le principe doit être cohérent avec la démarche de modélisation.



Au sens de la distance aux barycentres conditionnels, l'individu supplémentaire « ♦ » est rattaché à la classe des « rouges ».

Solution 1 : Affecter l'individu à la classe dont le barycentre est le plus proche. L'approche est totalement cohérente avec l'algorithme des centres mobiles.

Solution 2 : Tenter de reproduire le processus d'affectation à l'aide d'un algorithme d'apprentissage supervisé, notamment l'analyse discriminante. QDA (quadratique) parce que les classes sont convexes, et peut être LDA (linéaire) si les classes sont de formes similaires. Utiliser le modèle en déploiement.

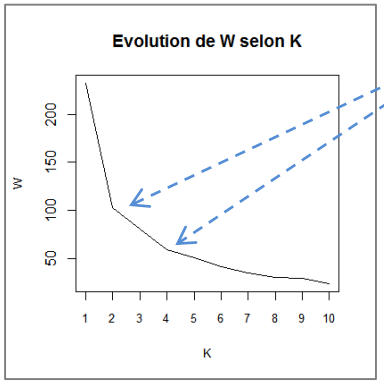
Ex. Pour les données dans le plan (à gauche), QDA arrive à reproduire parfaitement le mécanisme d'appartenance aux classes.

		Affectation QDA		
		C1	**C2**	**C3**
Classes K-Means	**C1**	102	0	0
	C2	0	100	0
	C3	0	0	98

Matrice de confusion en resubstitution.

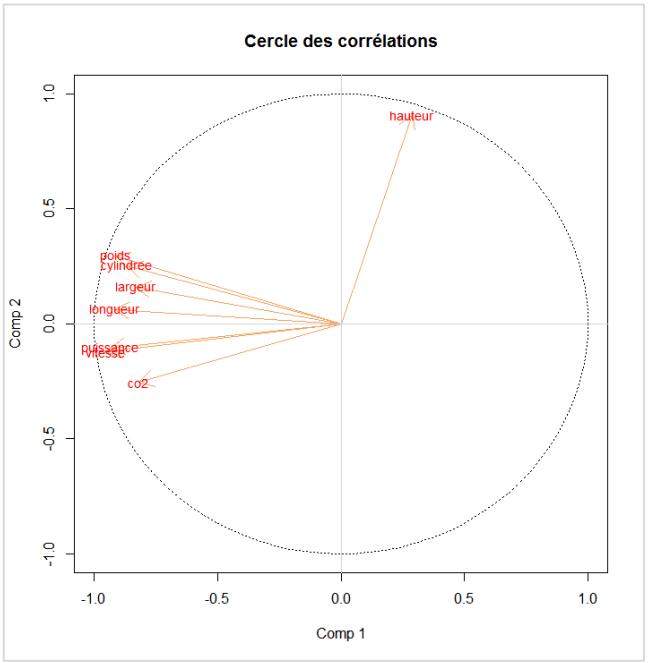
Algorithme K-Means

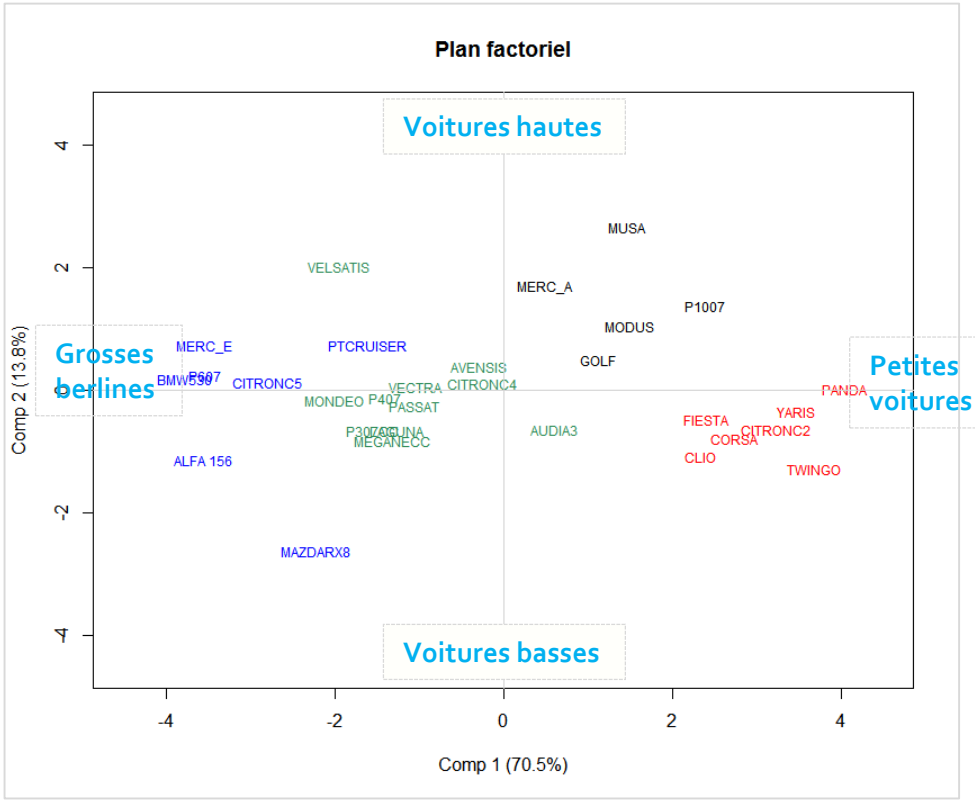
L'exemple des « Voitures »



Partition en 2 ou 4 classes possibles. On choisit K = 4 parce que cette solution sera confortée par les analyses complémentaires (ACP)

La classification semble tenir la route. Mais on constate qu'il y a des voitures à part (Vel Satis, Mazda Rx8), et certains rattachements posent question (Golf parmi les petits monospaces, PTCruiser parmi les berlines).





Le cercle des corrélations permet de comprendre le positionnement des observations dans le plan factoriel.

Cas des variables qualitatives

Stratégie pour le traitement des variables actives qualitatives

Distance du KHI-2 (1)

Passage au tableau des indicatrices

Un tableau de données (variables qualitatives) peut être transformé en tableau d'indicatrices, puis en tableaux de profils. Cf. le cours d'Analyse des Correspondances Multiples.

Tableau des canidés (Tenenhaus, 2006 ; page 254)

$j = 1,\ldots,p$

ID	Chien	Taille	Velocite	Affection
1	Beauceron	Taille++	Veloc++	Affec+
2	Basset	Taille-	Veloc-	Affec-
3	Berger All	Taille++	Veloc++	Affec+
4	Boxer	Taille+	Veloc+	Affec+
5	Bull-Dog	Taille-	Veloc-	Affec+
6	Bull-Mastif	Taille++	Veloc-	Affec-
7	Caniche	Taille-	Veloc+	Affec+
8	Labrador	Taille+	Veloc+	Affec+

$i = 1,\ldots,n$

$M = \sum_{j=1}^{p} m_j = 8$

$m_1 = 3$ $m_2 = 3$ $m_3 = 2$ $p = 3$

Chien	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+	Somme
Beauceron	0	0	1	0	0	1	0	1	3
Basset	1	0	0	1	0	0	1	0	3
Berger All	0	0	1	0	0	1	0	1	3
Boxer	0	1	0	x_{ik}	1	0	0	1	3
Bull-Dog	1	0	0	1	0	0	0	1	3
Bull-Mastif	0	0	1	1	0	0	1	0	3
Caniche	1	0	0	0	1	0	0	1	3
Labrador	0	1	0	0	1	0	0	1	3
Somme	3	2	3	3	3	2	2	6	24

$n = 8$

$n_1 = 3$

$\sum_{k=1}^{M} n_k = n \times p = 8 * 3 = 24$

La distance entre 2 individus peut être mesurée.
La barycentre a un sens, c'est le profil « moyen ».
La distance au barycentre peut être mesurée également.

$\frac{x_{ik}}{p}$

Chien	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+
Beauceron	0.000	0.000	0.333	0.000	0.000	0.333	0.000	0.333
Basset	0.333	0.000	0.000	0.333	0.000	0.000	0.333	0.000
Berger All	0.000	0.000	0.333	0.000	0.000	0.333	0.000	0.333
Boxer	0.000	0.333	0.000	0.000	0.333	0.000	0.000	0.333
Bull-Dog	0.333	0.000	0.000	0.333	0.000	0.000	0.000	0.333
Bull-Mastif	0.000	0.000	0.333	0.333	0.000	0.000	0.333	0.000
Caniche	0.333	0.000	0.000	0.000	0.333	0.000	0.000	0.333
Labrador	0.000	0.333	0.000	0.000	0.333	0.000	0.000	0.333
Profil moyen	0.125	0.083	0.125	0.125	0.125	0.083	0.083	0.250

Barycentre (O)

$\frac{n_k}{n \times p}$

Distance du KHI-2 (2)

Calcul des distances

$\frac{x_{ik}}{p}$

Chien	Taille-	Taille+	Taille++	Veloc-	Veloc+	Veloc++	Affec-	Affec+
Beauceron	0.000	0.000	0.333	0.000	0.000	0.333	0.000	0.333
Basset	0.333	0.000	0.000	0.333	0.000	0.000	0.333	0.000
Berger All	0.000	0.000	0.333	0.000	0.000	0.333	0.000	0.333
Boxer	0.000	0.333	0.000	0.000	0.333	0.000	0.000	0.333
Bull-Dog	0.333	0.000	0.000	0.333	0.000	0.000	0.000	0.333
Bull-Mastif	0.000	0.000	0.333	0.333	0.000	0.000	0.333	0.000
Caniche	0.333	0.000	0.000	0.000	0.333	0.000	0.000	0.333
Labrador	0.000	0.333	0.000	0.000	0.333	0.000	0.000	0.333
Profil moyen	0.125	0.083	0.125	0.125	0.125	0.083	0.083	0.250

Barycentre (O)

$\frac{n_k}{n \times p}$

Attention, les écarts entre modalités rares sont exacerbées

$$d^2(beauceron, basset) = \sum_{k=1}^{M} \frac{1}{\frac{n_k}{n \times p}} \left(\frac{x_{1k}}{p} - \frac{x_{2k}}{p} \right)^2 = \frac{1}{0.125}(0.000 - 0.333)^2 + \cdots + \frac{1}{0.250}(0.333 - 0.000)^2 = 5.778$$

$$d^2(basset, O) = \frac{1}{0.125}(0.333 - 0.125)^2 + \frac{1}{0.083}(0.333 - 0.083)^2 + \cdots + \frac{1}{0.250}(0.000 - 0.250)^2 = 2.111$$

Le « basset » est plus proche du « canidé moyen » que du « beauceron ».

Algorithme des K-Means

Distance du KHI-2

Algorithme toujours aussi simple

Entrée : X (n obs., p variables), K #classes

Initialiser K centres de classes G_k

REPETER

Allocation. Affecter chaque individu à la classe dont le centre est le plus proche

Représentation. Recalculer les centres de classes à partir des individus rattachés

JUSQU'À Convergence

Sortie : Une partition des individus caractérisée par les K centres de classes G_k

En utilisant la distance du KHI-2

Chaque classe peut être associée un profil moyen.

Algorithme des K-Modes

Autre approche pour traiter les variables qualitatives

Principe : (1) Définir une distance adaptée aux données qualitatives. (2) Une classe est représentée par un profil synthétique défini par les valeurs modales prises pour chaque variable active.

$$d(i,i') = \sum_{j=1}^{p} \delta(v_{ij}, v_{i'j}), \text{ où } \delta(i,i') = \begin{cases} 0 \text{ si } v_{ij} = v_{i'j} \\ 1 \text{ si } v_{ij} \neq v_{i'j} \end{cases}$$

Formule de calcul de la distance entre individus (v_{ij} *est la modalité prise par l'individu i pour la variable* V_j)

Entrée : X (n obs., p variables), K #classes

Initialiser K individus représentatifs des classes G_k (*choix de K individus pris au hasard*)

REPETER

Allocation. Affecter chaque individu à la classe dont le représentant est le plus proche

Représentation. Recalculer l'individu synthétique M_k représentatif de chaque classe constituée

JUSQU'À Convergence

Sortie : Une partition des individus caractérisée par les K modes de classes M_k

La description de l'individu M_k représentatif de la classe est constituée à partir du mode de chaque variable (pour les individus présents dans la classe).

Exemple

Chien	Taille	Velocite	Affection	Agressivite
Basset	Taille-	Veloc-	Affec-	Agress+
Bull-Dog	Taille-	Veloc-	Affec+	Agress-
Caniche	Taille-	Veloc+	Affec+	Agress-
Chihuahua	Taille-	Veloc-	Affec+	Agress-
Cocker	Taille+	Veloc-	Affec+	Agress+

Représentant	**Taille-**	**Veloc-**	**Affec+**	**Agress-**

Minimisation d'un critère similaire à W

$$Q = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d(i, M_k)$$

Remarque : Attention, ***les résultats peuvent être très instables****. Le mode – et donc la description de l'individu prototype – peut changer d'un coup avec une ou deux observations en plus ou en moins.*

Tandem Analysis

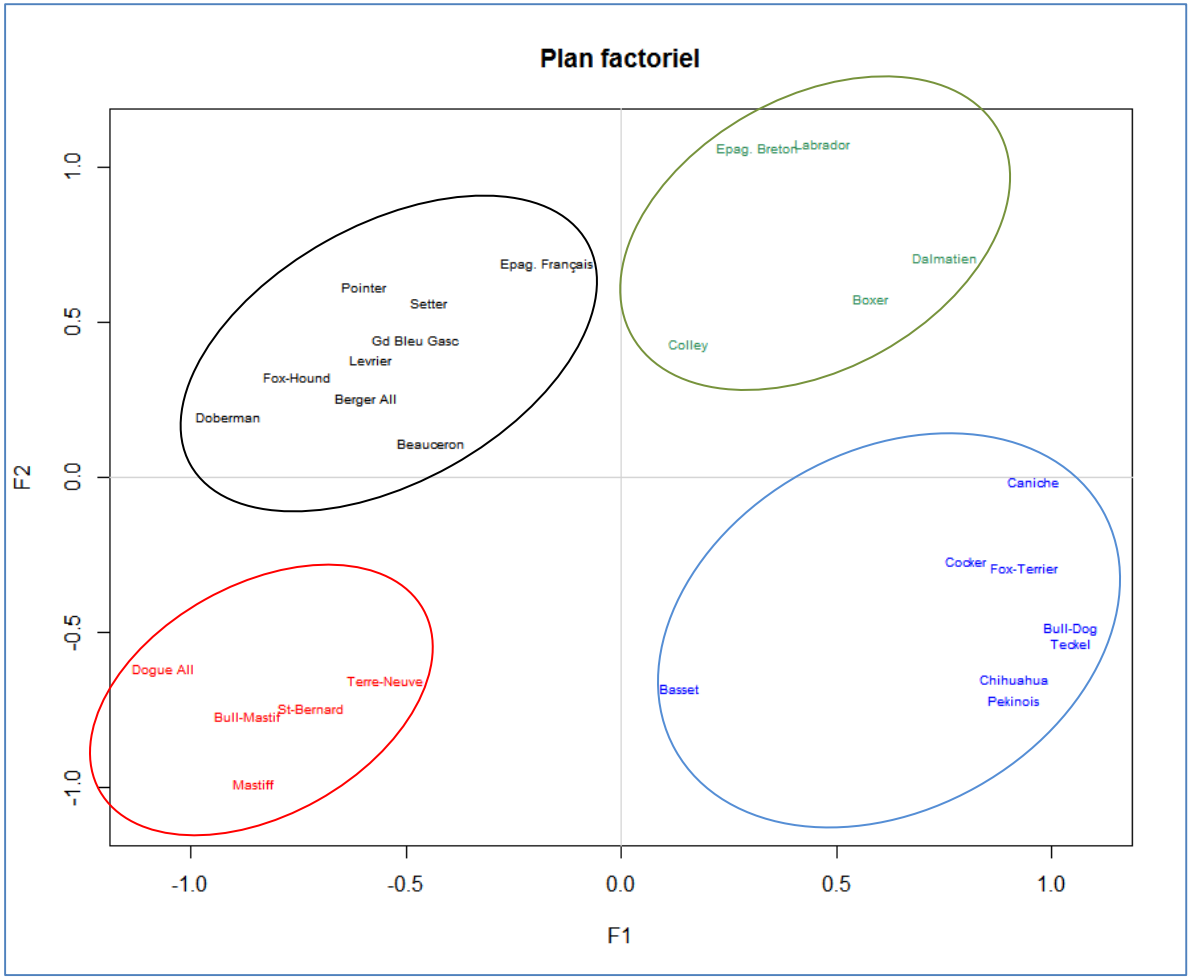
Se projeter dans un espace factoriel

Principe : Se projeter dans un espace factoriel (ACM : analyse des correspondances multiples) et réaliser les K-Means avec la distance euclidienne usuelle.

Avantage : L'approche peut être étendue à des problèmes avec variables actives mixtes (qualitatives et quantitatives) avec l'AFDM (Analyse factorielle des données mixtes).

Données « canidés »
(Tenenhaus, 2006 ; page 254)

Classification dans l'espace défini par le premier plan factoriel



Attention : N'utiliser qu'un nombre réduit de facteurs permet d'éliminer le « bruit » des données utilisées pour la classification. Mais ce choix devient un paramètre supplémentaire de l'algorithme.

Fuzzy C-Means

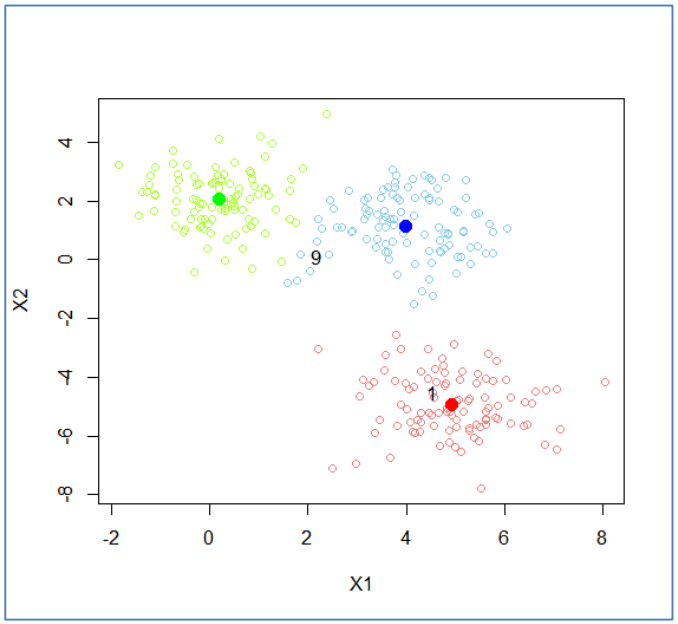
Plutôt qu'une affectation « sèche » (crisp) obligatoire à l'une des classes, introduire la notion de degré d'appartenance des individus aux classes

Fuzzy C-Means

Position du problème

Problème : Les K-Means affectent nécessairement un point à l'une des classes en leur accordant la même crédibilité. Or cela est éminemment discutable pour certains individus où le rattachement à une classe plutôt qu'une autre tient à très peu de choses dans le calcul des distances aux barycentres.

Solution : Introduire un indicateur de degré d'appartenance aux classes.



On se rend bien compte que selon l 'éloignement par rapport aux barycentres conditionnels, le degré d'appartenance des individus aux classes peut différer.

Ex.

N° point	Bleu	Rouge	Vert
1	0.011	0.983	0.006
9	0.472	0.105	0.423

Comment procéder pour obtenir ce type d'indicateurs ?

Ils doivent intervenir lors du classement, mais aussi durant la modélisation pour « lisser » le processus de constitution des classes (pondération du calcul des barycentres conditionnels)

Sachant de toute manière qu'il est toujours possible de procéder à une affectation « crisp » en prenant le max. des degrés d'appartenance.

Fuzzy C-Means

Algorithme (Dunn, 1973 ; Bezdek, 1981)

Principe : Introduction d'un tableau d'appartenance aux classes Ω de dimension (n x K) [n nombre d'obs., K nombre de classes].

➔ Les valeurs de Ω sont définies dans [0 ; 1], la somme de chaque ligne est égale à 1.

Entrée : X (n obs., p variables), K #classes

Assigner aléatoirement aux individus leurs degrés d'appartenances aux classes (Ω)

REPETER

Représentation. Calculer les barycentres G_k des classes en tenant compte des degrés d'appartenance

Allocation. Calculer pour chaque individu ses degrés d'appartenances aux classes

JUSQU'À Convergence

Sortie : Un tableau avec pour chaque individu son vecteur d'appartenance aux classes

$$G_{kj} = \frac{\sum_{i=1}^{n} \omega_{ik}^{m} x_j}{\sum_{i=1}^{n} \omega_{ik}^{m}}$$

La composante j (j = 1,…, p ; nombre de variables) du vecteur G_k

$$w_{ik} = \frac{1}{\sum_{l=1}^{K} \left(\frac{\|x_i - G_k\|}{\|x_i - G_l\|} \right)^{\frac{2}{m-1}}}$$

*Pour calculer ω_{ik} le degré d'appartenance à la classe **k** de l'individu n°**i**, on confronte bien sa distance à G_k avec sa distance aux autres barycentres conditionnels (G_l, l = 1…K)*

Lorsque la matrice d'appartenance aux classes des individus Ω n'est plus substantiellement modifiée.

Minimisation d'un critère similaire à W

$$Q = \sum_{i=1}^{n} \sum_{k=1}^{K} \omega_{ik}^{m} \times \|x_i - G_k\|^2$$

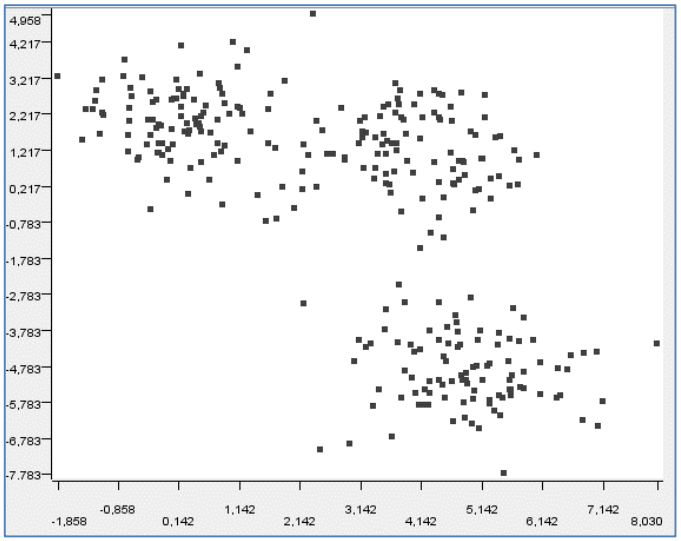
Le paramètre **m** « *fuzzifier* » (m ≥ 1) contrôle le degré de « fuzzification » du processus. Plus il est élevé, plus les appartenances aux classes sont lissées. A contrario, m= 1, on retombe sur un K-Means « crisp » (ω_{ik} = 0 ou 1). On fixe m = 2 de manière générale dans les logiciels.

!

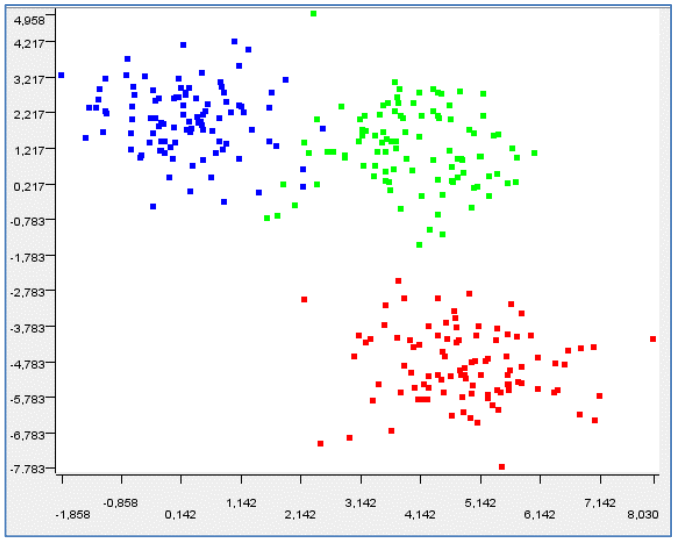
Fuzzy C-Means

Exemple dans le plan

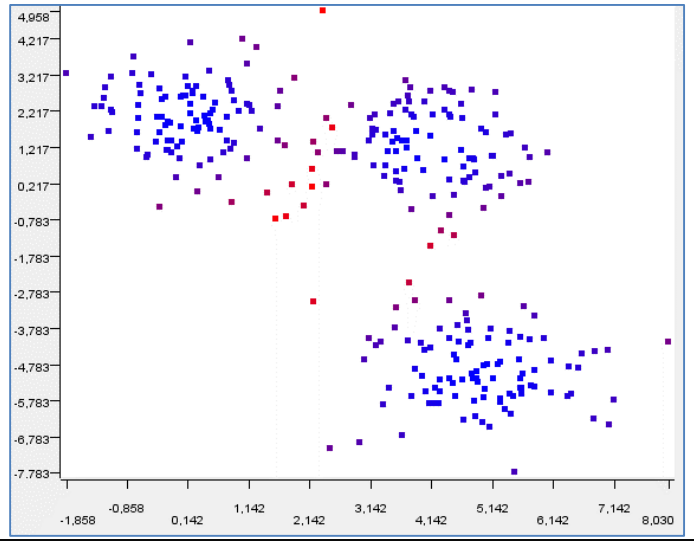
Positionnement des points dans le plan



Fuzzy C-Means sait construire une partition « crisp » en associant chaque individu à la classe maximisant le degré d'appartenance.

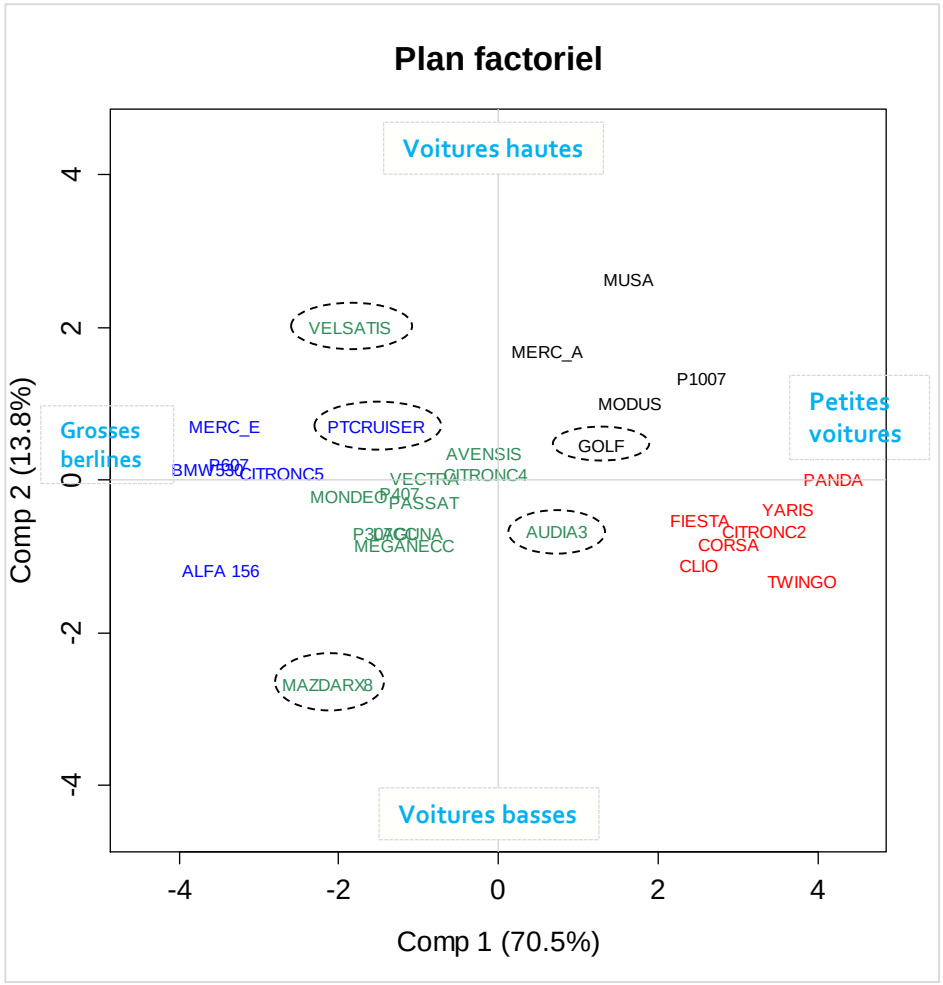


Mais il sait relativiser les résultats. Ici, points bleus, appartenance forte (~1) ; points rouges, appartenance faible (~1/3).



Fuzzy C-Means

Exemple des voitures



Résultats de Fuzzy C-Means représentés dans le premier plan factoriel. On notera que « Mazda RX8 » a changé de camp par rapport aux K-Means.

On se rend compte que le rattachement de certains véhicules à leur classe n'est pas vraiment tranchée (VELSATIS, MAZDA RX8, PT CRUISER, GOLF, AUDI A_3). On comprend pourquoi dans le plan factoriel !

Modele	cluster_0	cluster_1	cluster_2	cluster_3	Winner	MAX
P607	0.870	0.023	0.013	0.094	cluster_0	**0.870**
MERC_E	0.718	0.061	0.035	0.186	cluster_0	**0.718**
CITRONC5	0.878	0.020	0.011	0.091	cluster_0	**0.878**
PTCRUISER	0.409	0.187	0.081	0.323	cluster_0	**0.409**
BMW530	0.845	0.032	0.019	0.105	cluster_0	**0.845**
ALFA 156	0.613	0.094	0.065	0.228	cluster_0	**0.613**
GOLF	0.080	0.422	0.271	0.227	cluster_1	**0.422**
P1007	0.033	0.701	0.202	0.065	cluster_1	**0.701**
MUSA	0.052	0.755	0.106	0.087	cluster_1	**0.755**
MODUS	0.013	0.909	0.049	0.029	cluster_1	**0.909**
MERC_A	0.063	0.725	0.082	0.130	cluster_1	**0.725**
PANDA	0.033	0.172	0.736	0.059	cluster_2	**0.736**
TWINGO	0.020	0.074	0.866	0.039	cluster_2	**0.866**
CITRONC2	0.003	0.017	0.973	0.007	cluster_2	**0.973**
YARIS	0.012	0.067	0.897	0.025	cluster_2	**0.897**
FIESTA	0.028	0.129	0.775	0.068	cluster_2	**0.775**
CORSA	0.008	0.035	0.939	0.018	cluster_2	**0.939**
CLIO	0.038	0.136	0.740	0.086	cluster_2	**0.740**
AUDIA3	0.097	0.245	0.230	0.428	cluster_3	**0.428**
AVENSIS	0.113	0.177	0.083	0.628	cluster_3	**0.628**
P407	0.074	0.032	0.017	0.877	cluster_3	**0.877**
CITRONC4	0.106	0.165	0.081	0.649	cluster_3	**0.649**
MONDEO	0.288	0.113	0.072	0.526	cluster_3	**0.526**
VECTRA	0.068	0.045	0.023	0.865	cluster_3	**0.865**
PASSAT	0.099	0.060	0.032	0.808	cluster_3	**0.808**
VELSATIS	0.351	0.191	0.077	0.381	cluster_3	**0.381**
LAGUNA	0.058	0.023	0.014	0.905	cluster_3	**0.905**
MEGANECC	0.106	0.041	0.026	0.826	cluster_3	**0.826**
P307CC	0.209	0.060	0.035	0.695	cluster_3	**0.695**
MAZDARX8	0.355	0.135	0.116	0.395	cluster_3	**0.395**

Classification de variables

Regroupement en classes des variables

Classification de variables

Classification autour de variables latentes (Vigneau & Qannari)

Principe : Mettre en évidence une structuration des données via une identification des groupes de variables fortement liées

Entrée : X (n obs., p variables), K #classes

Initialiser les classes C_k avec K variables pris au hasard

REPETER

Allocation. Affecter chaque variable à la classe dont le représentant est le plus proche

Représentation. Recalculer la variable synthétique (U_k) représentative de chaque classe constituée

JUSQU'À Convergence

Sortie : Une partition des variables en K groupes caractérisés par les variables latentes U_k

Le carré du coefficient de corrélation r^2 peut faire office de mesure de proximité. La distance peut être comptabilisée avec $(1 - r^2)$.

On utilise la 1ère composante principale U_k de l'ACP pour résumer le groupe n°k de p_k variables. En effet U_k est telle qu'elle maximise

$$\lambda_k = \sum_{j=1}^{p_k} r^2(X_j, U_k)$$

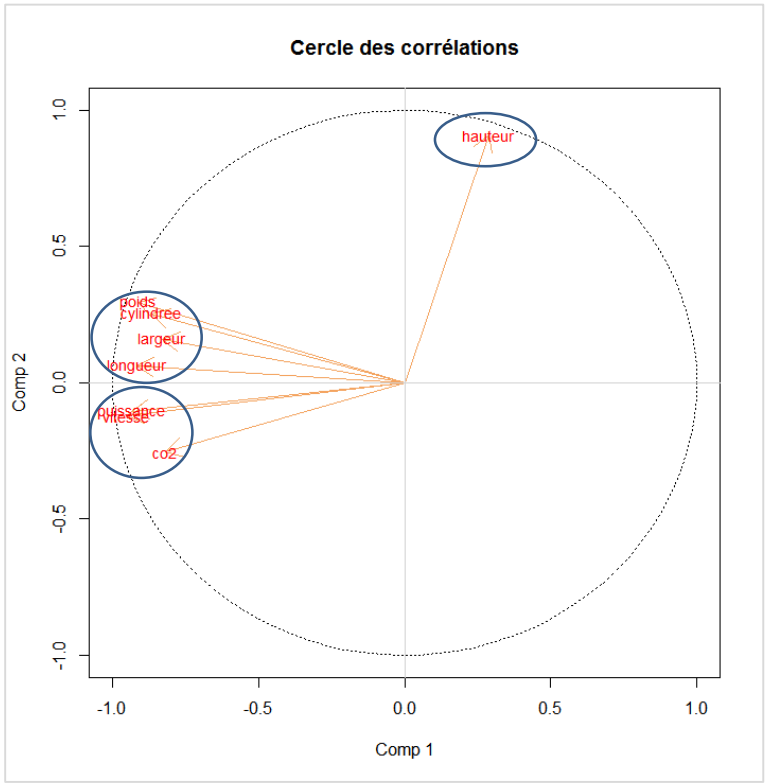
λ_k est la première valeur propre issue de la diagonalisation de la matrice des corrélations.

La première composante principale de l'ACP (variable latente) est le meilleur résumé que l'on peut avoir d'un groupe de variables (à l'instar du barycentre dans l'espace des individus).

Classification de variables

Exemple des voitures avec Tanagra – Partition en 3 groupes

Voici ce que nous dit le cercle des corrélations de l'ACP (1er plan factoriel)



Valeur propre associée à la 1ère composante du groupe

Cluster summary

Cluster	# Members	Variation Explained	Proportion Explained
1	3	2.7520	0.9173
2	4	3.4028	0.8507
3	1	1.0000	1.0000
Total		7.1548	0.8943

Fidélité de la représentation du groupe par la 1ère composante (λ_k/p_k). Indication de la compacité du groupe.

Corrélation² de la variable avec la variable latente de son groupe.

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R² ratio
1	puissance	0.9738	0.6520	0.0754
	vitesse	0.9037	0.7381	0.3676
	co2	0.8746	0.4181	0.2156
2	cylindree	0.7675	0.5932	0.5716
	longueur	0.9080	0.5903	0.2245
	largeur	0.8202	0.4181	0.3090
	poids	0.9070	0.6204	0.2449
3	hauteur	1.0000	0.1148	0.0000

Plus forte corrélation² de la variable avec la variable latente d'un des autres groupes.

$$1 - R^2 ratio = \frac{1 - R^2\ own}{1 - R^2\ next}$$

Degré d'appartenance à son propre groupe. Plus (1-R²) est proche de 0, mieux c'est ; > 1 pas bon du tout.

Cluster correlations -- Structure

Attribute	# membership	Cluster 1	Cluster 2	Cluster 3
puissance	1	0.9868	0.8074	-0.2870
cylindree	1	0.7702	0.8761	-0.0437
vitesse	2	0.9506	0.8591	-0.3567
longueur	1	0.7683	0.9529	-0.2718
largeur	1	0.6466	0.9056	-0.1803
hauteur	1	-0.3388	-0.1412	1.0000
poids	1	0.7877	0.9524	-0.0209
co2	1	0.9352	0.6466	-0.3316

Corrélation de chaque variable avec les variables latentes des groupes (on a le signe de la relation cette fois-ci).

Cluster 1 et Cluster 2 sont très proches au regard des corrélations.

Conclusion

- Les techniques de partitionnement par réallocation présentent l'avantage de la simplicité.

- Elles peuvent traiter de très grandes bases, mais sont lentes car requièrent plusieurs passages sur la base de données.
- Elles produisent des classes convexes, centrées sur les barycentres conditionnels (méthode des centres mobiles).
- La démarche peut être étendue aux cas des variables actives qualitatives ou mixtes.
- La démarche peut être étendue à la classification de variables.

- Le choix du nombre de classes K reste un problème ouvert.
- Résumer une classe par le barycentre n'est pas toujours pertinent, voir la notion de « medoid ». Cf. K-Medoids.

Bibliographie

Ouvrages

Chandon J.L., Pinson S., « Analyse typologique – Théorie et applications », Masson, 1981.

Diday E., Lemaire J., Pouget J., Testu F., « Eléments d'analyse de données », Dunod, 1982.

Gan G., Ma C., Wu J., « Data Clustering – Theory, Algorithms and Applications », SIAM, 2007.

L. Lebart, A. Morineau, M. Piron, « Statistique exploratoire multidimensionnelle », Dunod, 2000.

Tutoriels

« Classification automatique sous R », octobre 2015.

« Classification automatique sous Python », mars 2016.

« Classification de variables », mars 2008.

« Classification automatique – Déploiement de modèles », octobre 2008.