Analyse Factorielle des Données Mixtes
Principes et pratique de l'AFDM

Prof. Mustapha RACHDI

Université Grenoble Alpes
UFR SHS, BP. 47
38040 Grenoble cedex 09
France
Bureau : C08 au Bât. Michel Dubois
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr

Position du problème

- Une méthodologie factorielle permettant d'inclure à la fois des variables quantitatives et qualitatives en tant qu'éléments actifs d'une même analyse (cf. Escofier, 1979) dans le cadre de l'AFCM.
- Cette approche se confond d'une part avec la méthodologie esquissée par Saporta (1990) dans le cadre de l'ACP, et d'autre part avec une AFM (Analyse Factorielle Multiple) dans laquelle chaque variable constitue un groupe à elle seule.
- L'ensemble de ces trois points de vue confère à la méthode proposée par Escofier (1979) le statut d'une méthode à part entière dotée de plusieurs bonnes propriétés et facile à mettre en oeuvre.
- Par ailleurs, en combinant cette méthode avec l'AFM, **il est possible** de réaliser une AFM sur des groupes de variables mixtes ce qui est une possibilité nouvelle.

AFDM

Dans toute la suite, de ce chapitre, nous nous consacrerons à l'AFDM :

1. Position du problème
2. AFDM - Calculs - Equivalences
3. Pratique de l'AFDM
4. Solution alternative : discrétisation + AFCM
5. R Packages : FactoMiner, ade4, PCAmixdata
6. Bibliographie

Position du problème

Construire un nouveau système de représentation (facteurs, axes factoriels : combinaisons linéaires des variables quantitatives et des indicatrices des variables qualitatives) qui permet synthétiser l'information

Position du problème

Variables « actives » quantitatives et /ou qualitatives

C = 5 quantitatives

D = 3 qualitatives

$i : 1, \ldots, n = 10$

Individus actifs

Modele	puissanc	longueu	hauteur	poids	CO2	origine	carburant	4X4
GOLF	75	421	149	1217	143	Europe	Diesel	non
CITRONC4	138	426	146	1381	142	France	Diesel	non
P607	204	491	145	1723	223	France	Diesel	non
VELSATIS	150	486	158	1735	188	France	Diesel	non
CITRONC2	61	367	147	932	141	France	Essence	non
CHRYS300	340	502	148	1835	291	Autres	Essence	non
AUDIA3	102	421	143	1205	168	Europe	Essence	non
OUTLAND	202	455	167	1595	237	Autres	Diesel	oui
PTCRUISER	223	429	154	1595	235	Autres	Essence	non
SANTA_FE	125	450	173	1757	197	Autres	Diesel	oui

- Le tableau de données comporte des caractères quantitatifs et des caractères qualitatifs.

Position du problème : questions !

- Quelles sont les automobiles qui se ressemblent ? (proximité entre les individus)
- Sur quelles caractéristiques sont fondées les ressemblances / dissemblances ?
- Quelles sont les relations entre les variables ?
- Et les relations entre les modalités, entre les modalités et les variables quantitatives ?

Objectif de l'analyse factorielle : rappel

1. Trouver un facteur C^1 qui soit le plus lié possible avec les variables originelles
2. Si la liaison n'est pas parfaite, trouver un second facteur qui explique l'information résiduelle (non prise en compte en 1.)
3. ... jusqu'au Qème facteur

Objectif de l'analyse factorielle

- **ACP** (Toutes les variables actives sont quantitatives) :

$$\lambda_1 = \sum_{j=1}^{C} R^2(C^1, X_j)$$

où R^2 est le carré du coefficient de corrélation.
C'est pour cela que dans le cercle des corrélations, on se concentre sur les variables proches du bord.

- **AFCM** (Toutes les variables actives sont qualitatives) :

$$\lambda_1 = \sum_{j=C+1}^{D} \eta^2(C^1, X_j)$$

où η^2 est le carré du rapport de corrélation.
On veut que les modalités d'une variable soient le plus écartés possible les unes des autres.

- Les variables actives sont **quantitatives et qualitatives** :

$$\sum_{j=1}^{C} R^2(C^1, X_j) + \sum_{j=C+1}^{D} \eta^2(C^1, X_j)$$

où $(0 \leq R^2 \leq 1)$ et $(0 \leq \eta^2 \leq 1)$: les deux types de variables jouent un rôle équilibré dans l'analyse !!! C'est un aspect primordial.

- Mais, Comment parvenir à ce résultat ?
(en exploitant les résultats de l'analyse factorielle).

AFDM via un programme d'ACP

- On peut obtenir les résultats de l'AFDM avec un programme réalisant une ACP
- Il faut simplement passer par une transformation judicieuse des données
- Equivalences avec l'ACP et l'AFCM

Transformation des données

Etape 1 – Codage disjonctif complet des variables qualitatives

$$k : 1,\ldots,C+M=P=12$$

$$\sum_{j=C+1}^{D} m_j = M = 7$$

$$j : 1,\ldots,C=5$$

$$m_1 = 3 \qquad m_2 = 2 \qquad m_3 = 2$$

Modele	puissance	longueur	hauteur	poids	CO2		orig_Autres	orig_Europ	orig_France	carb_Diese	carb_Essen	4X4_non	4X4_oui
GOLF	75	421	149	1217	143		0	1	0	1	0	1	0
CITRONC4	138	426	146	1381	142		0	0	1	1	0	1	0
P607	204	491	145	1723	223		0	0	1	1	0	1	0
VELSATIS	150	486	158	1735	188		0	0	1	1	0	1	0
CITRONC2	61	367	147	932	141	x_{ik}	0	0	1	0	1	1	0
CHRYS300	340	502	148	1835	291		1	0	0	0	1	1	0
AUDIA3	102	421	143	1205	168		0	1	0	0	1	1	0
OUTLAND	202	455	167	1595	237		1	0	0	1	0	0	1
PTCRUISER	223	429	154	1595	235		1	0	0	0	1	1	0
SANTA_FE	125	450	173	1757	197		1	0	0	1	0	0	1
Moyenne	162	444.8	153	1497.5	196.5	**N_k**	4	2	4	6	4	8	2
Ecart-Type	78.9	38.8	9.5	283.7	47.5	**p_k**	0.4	0.2	0.4	0.6	0.4	0.8	0.2

μ_k : moyenne de la variable

σ_k : écart-type

n_k : Effectif de la modalité

p_k : proportion = n_k/n

Transformation des données

Etape 2 – Standardisation différenciée des colonnes

Modele	puissance	longueur	hauteur	poids	CO2	orig_Autres	orig_Europe	orig_France	carb_Diesel	carb_Essenc	4X4_non	4X4_oui
GOLF	-1.103	-0.614	-0.419	-0.989	-1.127	0.000	2.236	0.000	1.291	0.000	1.118	0.000
CITRONC4	-0.304	-0.485	-0.733	-0.411	-1.148	0.000	0.000	1.581	1.291	0.000	1.118	0.000
P607	0.532	1.192	-0.838	0.795	0.558	0.000	0.000	1.581	1.291	0.000	1.118	0.000
VELSATIS	-0.152	1.063	0.524	0.837	-0.179	0.000	0.000	1.581	1.291	0.000	1.118	0.000
CITRONC2	-1.280	-2.007	-0.628	-1.993	-1.169	0.000	0.000	1.581	0.000	1.581	1.118	0.000
CHRYS300	2.256	1.476	-0.524	1.189	1.990	1.581	0.000	0.000	0.000	1.581	1.118	0.000
AUDIA3	-0.761	-0.614	-1.047	-1.031	-0.600	0.000	2.236	0.000	0.000	1.581	1.118	0.000
OUTLAND	0.507	0.263	1.466	0.344	0.853	1.581	0.000	0.000	1.291	0.000	0.000	2.236
PTCRUISER	0.773	-0.408	0.105	0.344	0.811	1.581	0.000	0.000	0.000	1.581	1.118	0.000
SANTA_FE	-0.469	0.134	2.094	0.915	0.011	1.581	0.000	0.000	1.291	0.000	0.000	2.236

$$z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k} \; ; k = 1, \ldots, C$$

$$z_{ik} = \frac{x_{ik}}{\sqrt{p_k}} \; ; k = C+1, \ldots, P$$

Et on peut lancer une ACP (non normée puisque les données sont déjà réduites) sur ces données transformées $\Longrightarrow$ on obtient les résultats de l'AFDM

Pourquoi ces transformations ?

(1) Variables quantitatives (inerties) — $I(\text{Variable}) = 1$ → Comme en ACP normée

⇒ Si toutes les variables sont quantitatives : AFDM = ACP

(2) Variables qualitatives (inerties) — $I(\text{Modalité}) = 1\text{-}p_k$, $I(\text{Variable}) = m_j\text{-}1$ → Comme en ACM (à un facteur près)

⇒ Si toutes les variables sont qualitatives : AFDM = ACM

L'AFDM est une "vraie" généralisation : dans le sens où l'ACP et l'AFCM en sont des cas particuliers.

Détails des calculs sous R

```
#chargement du fichier
autos <- read.table(file="AUTOS2005subset.txt",row.names=1,header=T,sep="\t")
print(summary(autos))

#fonction pour centrage-réduction
CR <- function(x){
  n <- length(x)
  m <- mean(x)
  v <- (n-1)/n*var(x)
  return((x-m)/sqrt(v))
}

#appliquer la fonction sur les variables continues
autos.cont <- data.frame(lapply(subset(autos,select=1:5),CR))
print(autos.cont)

#codage disjonctif complet
library(ade4)
autos.disc <- acm.disjonctif(subset(autos,select=6:8))

#fonction pour pondération des indicatrices
PF <- function(x){
  m <- mean(x)
  return(x/sqrt(m))
}

#appliquer la pondération sur les indicatrices
autos.disc.pond <- data.frame(lapply(autos.disc,PF))

#données transformées envoyées à l'ACP
autos.pour.acp <- cbind(autos.cont,autos.disc.pond)
rownames(autos.pour.acp) <- rownames(autos)
print(round(autos.pour.acp,3))
```

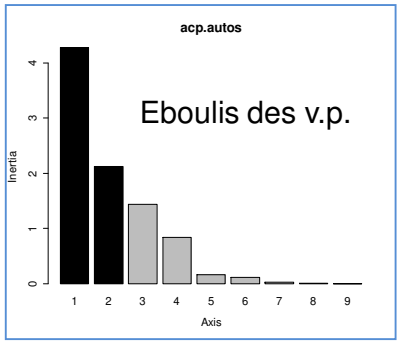
```
   puissance        longueur        hauteur          poids            CO2          origine    carburant  X4X4
Min.   : 61.0   Min.   :367.0   Min.   :143.0   Min.   : 932   Min.   :141.0   Autres:4   Diesel :6   non:8
1st Qu.:107.8   1st Qu.:422.2   1st Qu.:146.2   1st Qu.:1258   1st Qu.:149.2   Europe:2   Essence:4   oui:2
Median :144.0   Median :439.5   Median :148.5   Median :1595   Median :192.5   France:4
Mean   :162.0   Mean   :444.8   Mean   :153.0   Mean   :1498   Mean   :196.5
3rd Qu.:203.5   3rd Qu.:478.2   3rd Qu.:157.0   3rd Qu.:1732   3rd Qu.:232.0
Max.   :340.0   Max.   :502.0   Max.   :173.0   Max.   :1835   Max.   :291.0
```

```
    puissance   longueur    hauteur       poids         CO2
1  -1.1028751 -0.6140305 -0.4188539 -0.9885623 -1.12656547
2  -0.3042414 -0.4850325 -0.7329943 -0.4105793 -1.14762277
3   0.5324225  1.1919416 -0.8377078  0.7947266  0.55801841
4  -0.1521207  1.0629436  0.5235674  0.8370180 -0.17898704
5  -1.2803493 -2.0072090 -0.6282809 -1.9929839 -1.16868007
6   2.2564571  1.4757372 -0.5235674  1.1894466  1.98991471
7  -0.7606035 -0.6140305 -1.0471348 -1.0308537 -0.60013301
8   0.5070690  0.2631559  1.4659887  0.3436179  0.85282059
9   0.7732802 -0.4076337  0.1047135  0.3436179  0.81070599
10 -0.4690388  0.1341579  2.0942695  0.9145523  0.01052865
```

```
       origine.Autres origine.Europe origine.France carburant.Diesel carburant.Essence X4X4.non X4X4.oui
GOLF                0              1              0                1                0        1        0
CITRONC4            0              0              1                1                0        1        0
P607                0              0              1                1                0        1        0
VELSATIS            0              0              1                1                0        1        0
CITRONC2            0              0              1                0                1        1        0
CHRYS300            1              0              0                0                1        1        0
AUDIA3              0              1              0                0                1        1        0
OUTLAND             1              0              0                1                0        0        1
PTCRUISER           1              0              0                0                1        1        0
SANTA_FE            1              0              0                1                0        0        1
```

```
          puissance longueur hauteur  poids     CO2 origine.Autres origine.Europe origine.France carburant.Diesel carburant.Essence X4X4.non X4X4.oui
GOLF         -1.103   -0.614  -0.419 -0.989 -1.127          0.000          2.236          0.000           1.291           0.000    1.118    0.000
CITRONC4     -0.304   -0.485  -0.733 -0.411 -1.148          0.000          0.000          1.581           1.291           0.000    1.118    0.000
P607          0.532    1.192  -0.838  0.795  0.558          0.000          0.000          1.581           1.291           0.000    1.118    0.000
VELSATIS     -0.152    1.063   0.524  0.837 -0.179          0.000          0.000          1.581           1.291           0.000    1.118    0.000
CITRONC2     -1.280   -2.007  -0.628 -1.993 -1.169          0.000          0.000          1.581           0.000           1.581    1.118    0.000
CHRYS300      2.256    1.476  -0.524  1.189  1.990          1.581          0.000          0.000           0.000           1.581    1.118    0.000
AUDIA3       -0.761   -0.614  -1.047 -1.031 -0.600          0.000          2.236          0.000           0.000           1.581    1.118    0.000
OUTLAND       0.507    0.263   1.466  0.344  0.853          1.581          0.000          0.000           1.291           0.000    0.000    2.236
PTCRUISER     0.773   -0.408   0.105  0.344  0.811          1.581          0.000          0.000           0.000           1.581    1.118    0.000
SANTA_FE     -0.469    0.134   2.094  0.915  0.011          1.581          0.000          0.000           1.291           0.000    0.000    2.236
```

Détails des calculs sous R




```
[1] 4.27314 2.12189 1.43872 0.83637 0.16403 0.11447 0.03363 0.01581 0.00196
```

```
#acp avec le package ade4
library(ade4)
acp.autos <- dudi.pca(autos.pour.acp,center=T,scale=F,scannf=F)

#valeurs propres
print(round(acp.autos$eig,5))

#coordonnées ACP des variables : Gkh
#**** pour les quali -> calculs supplémentaires nécessaires ****
#récupérer coord. acp des modalités
moda <- acp.autos$co[6:12,1:2]

#fréquence des modalités
freq.moda <- colMeans(autos.disc)

#calcul des moyennes conditionnelles sur les 2 premiers facteurs
coord.moda <- moda[,1]*sqrt(acp.autos$eig[1]/freq.moda)
coord.moda <- cbind(coord.moda,moda[,2]*sqrt(acp.autos$eig[2]/freq.moda))
print(coord.moda)

#coordonnées des individus
print(round(acp.autos$li[,1:2],5))
```

$$r(F_h, X_k) = G_{kh}; k = 1, \ldots, C$$

$$\mu_{kh} = G_{kh} \times \sqrt{\frac{\lambda_h}{p_k}}; k = C+1, \ldots, C+M$$

	Comp1	Comp2
puissance	0.8193449	0.53918597
longueur	0.7961816	0.24488939
hauteur	0.5783768	-0.76358497
poids	0.9295162	0.06113128
CO2	0.8906616	0.37898421
origine.Autres	0.6145381	-0.09436935
origine.Europe	-0.5012586	0.01658939
origine.France	-0.2600947	0.08263888
carburant.Diesel	0.1063246	-0.40231276

	coord.moda	
origine.Autres	2.0085940	-0.21735125
origine.Europe	-2.3169688	0.05403519
origine.France	-0.8501096	0.19033366
carburant.Diesel	0.2837472	-0.75657040
carburant.Essence	-0.4256208	1.13485560
X4X4.non	-0.5181078	0.57354613
X4X4.oui	2.0724311	-2.29418453

	Axis1	Axis2
GOLF	-2.31780	-0.68727
CITRONC4	-1.44537	-0.12229
P607	0.77973	1.01856
VELSATIS	0.54106	-0.16006
CITRONC2	-3.27586	0.02512
CHRYS300	2.95770	2.62811
AUDIA3	-2.31613	0.79534
OUTLAND	2.25592	-1.84043
PTCRUISER	0.93181	1.09085
SANTA_FE	1.88894	-2.74794

Détails des calculs sous R

```
#carré des corrélations 1er facteur
r2 <- acp.autos$co[1:5,1]^2

#carré du rapport de corrélation, var. qualitatives
eta2 <- NULL
eta2[1] <- sum(acp.autos$co[6:8,1]^2)
eta2[2] <- sum(acp.autos$co[9:10,1]^2)
eta2[3] <- sum(acp.autos$co[11:12,1]^2)
```

$$r^2(F_h, X_j) = G_{jh}^2 ; j = 1, \ldots, C$$

$$\eta^2(F_h, X_j) = \sum_{k \in X_j} G_{kh}^2 ; j = C+1, \ldots, C+D$$

```
#valeurs à sommer
criteres <- c(r2,eta2)
names(criteres) <- colnames(autos)
print(criteres)
```

$r^2()$ $\eta^2()$

```
puissance  longueur   hauteur     poids       CO2    origine carburant      X4X4
0.67132603 0.63390520 0.33451975 0.86400031 0.79327801 0.69656654 0.02826231 0.25127747
```

```
#critère de l'AFDM – 1er facteur
lambda1 <- sum(criteres)
print(lambda1)
```

```
> print(lambda1)
[1] 4.273136
```

```
#confrontation avec résultat (v.p.) de l'ACP
#sur variables transformées – 1er facteur
print(acp.autos$eig[1])
```

```
> print(acp.autos$eig[1])
[1] 4.273136
```

L'inertie projetée sur le 1er facteur (ACP sur variables transformées) correspond bien au critère de l'AFDM : cqfd

Pratique de l'AFDM : Que lire et comment lire les résultats de l'AFDM ?

1. Détermination du nombre de facteurs à retenir
2. Caractérisation des facteurs par les variables : Analyse des relations entre les variables via les facteurs
3. Caractérisation des facteurs par les individus : Analyse des proximités entre les individus
4. Variables illustratives (supplémentaires) : Non utilisées pour la construction des facteurs, mais permettent de renforcer leur interprétation
5. Individus illustratifs (supplémentaires) : Par opposition aux individus "actifs' ' utilisés pour la construction des facteurs

Détermination du nombre de facteurs à retenir

Tableau des valeurs propres

Inertie(Totale) =
P – D = 12 – 3 = 9

Matrix trace = 9.00

λ_h

Axis	Val.Propre	% expliqué	% cumulé
1	4.27314	47.48%	47.48%
2	2.12189	23.58%	71.06%
3	1.43872	15.99%	87.04%
4	0.83637	9.29%	96.33%
5	0.16403	1.82%	98.16%
6	0.11447	1.27%	99.43%
7	0.03363	0.37%	99.80%
8	0.01581	0.18%	99.98%
9	0.00196	0.02%	100.00%
Tot.	9	-	-

H_{max} (nombre max de facteurs) =
P – D = 12 – 3 = 9

Part d'information restituée par le h-ème facteur : qualité (fidélité) de représentation sur le facteur

Part d'information restituée par les « h » premiers facteurs

Règle 1 : Kaiser - Guttman

Sélectionner les facteurs pour lesquels $\lambda_h > 1$ (« 1 » peut être vue également comme la moyenne des v.p.)

→ H = 3 facteurs ici

→ Mais ce critère est trop permissif, H est souvent trop grand

Règle 2 : Karlis – Saporta - Spinaki

Sélectionner les facteurs pour lesquels

$$\lambda_h > 1 + 2\sqrt{\frac{P-1}{n-1}} = 1 + 2\sqrt{\frac{12-1}{10-1}} = 3.211$$

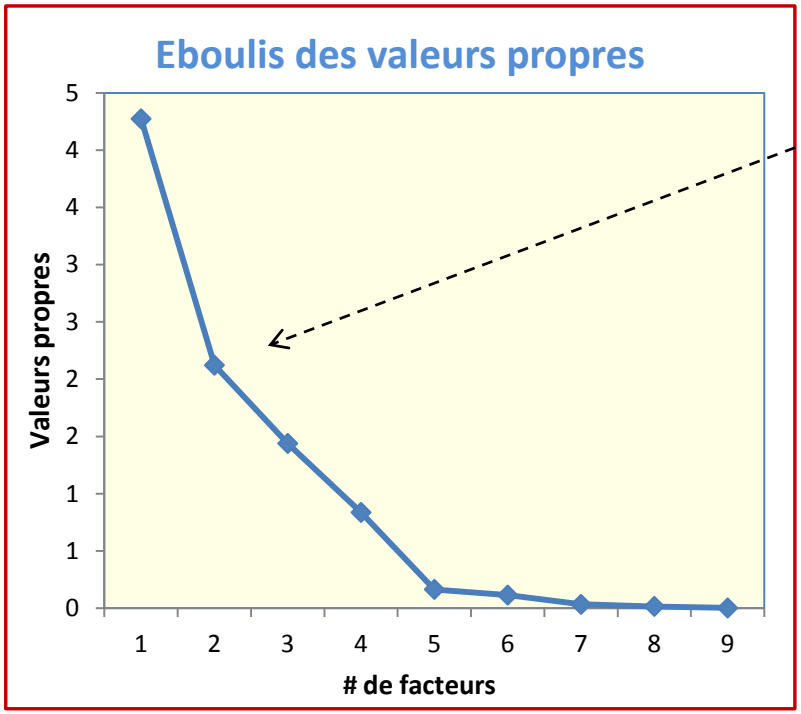
→ H = 1 facteur ici

→ Mais P est surévalué, certaines colonnes sont liées entres elles

→ Ce critère est trop restrictif dans l'AFDM

Détermination du nombre de facteurs à retenir

Eboulis des valeurs propres (scree plot) – Règle du coude



C'est ici que ça se passe !

H = 1 ou H = 2 : inclure le coude dans la sélection ou pas ?

Tout dépend de la v.p. associée au coude. Ici : λ_2 = 2.12189 # 23.58% de l'inertie. C'est beaucoup. On intègre le coude dans la sélection c.-à-d. H = 2

(*ça nous arrange aussi pour les graphiques…*)

Analyse des relations entre les variables via les facteurs

Analyse des relations entre les variables via les facteurs

Contribution de la variable au facteur

$$CTR_j(F_h) = \frac{r^2(F_h, X_j)}{\lambda_h} \quad \text{ou} \quad CTR_j(F_h) = \frac{\eta^2(F_h, X_j)}{\lambda_h}$$

Qualité de représentation d'une variable

$$COS^2_j(F_h) = r^2(F_h, X_j) \quad \text{ou} \quad COS^2_j(F_h) = \frac{\eta^2(F_h, X_j)}{m_j - 1}$$

En effet, $\sum_{h=1}^{H_{max}} r^2(F_h, X_j) = 1$ et $\sum_{h=1}^{H_{max}} \eta^2(F_h, X_j) = m_j - 1$

Attribute -	Facteur 1			Facteur 2		
	Coord.	CTR (%)	QLT % (Cumul %)	Coord.	CTR (%)	QLT % (Cumul %)
puissance (*)	0.67133	15.7%	67 % (67 %)	0.29072	13.7%	29 % (96 %)
longueur (*)	0.63391	14.8%	63 % (63 %)	0.05997	2.8%	6 % (69 %)
hauteur (*)	0.33452	7.8%	33 % (33 %)	0.58306	27.5%	58 % (92 %)
poids (*)	0.86400	20.2%	86 % (86 %)	0.00374	0.2%	0 % (87 %)
CO2 (*)	0.79328	18.6%	79 % (79 %)	0.14363	6.8%	14 % (94 %)
origine (**)	0.69657	16.3%	35 % (35 %)	0.01601	0.8%	1 % (36 %)
carburant (**)	0.02826	0.7%	3 % (3 %)	0.40464	19.1%	40 % (43 %)
4X4 (**)	0.25128	5.9%	25 % (25 %)	0.62012	29.2%	62 % (87 %)
Var. Expl.	4.27314		47 % (47 %)	2.12189		24 % (71 %)

$r^2(\,)$: puissance (*), longueur (*), hauteur (*), poids (*), CO2 (*)

$\eta^2(\,)$: origine (**), carburant (**), 4X4 (**)

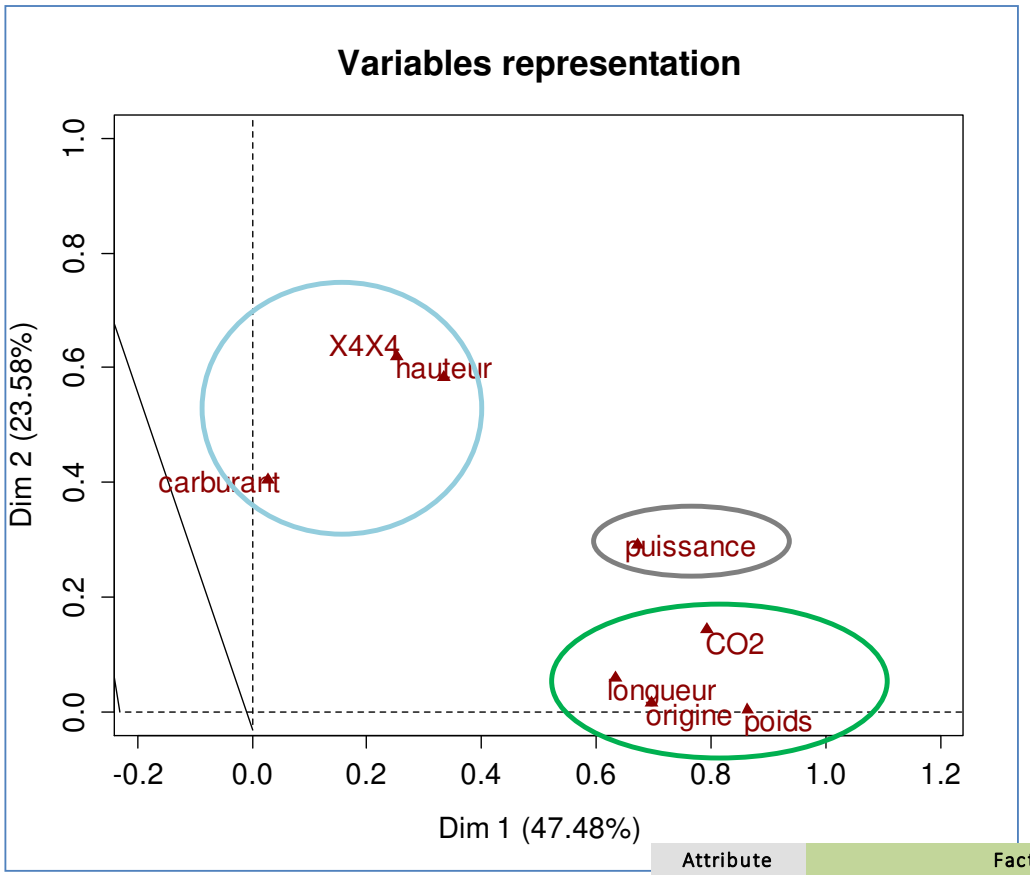
$$\lambda_1 = \sum_{j=1}^{C} r^2(F_1, X_j) + \sum_{j=C+1}^{C+D} \eta^2(F_1, X_j) = 4.27314$$

On sait que le premier facteur est déterminé par un lien entre (puissance, longueur, poids, CO2 et origine). Mais on ne sait pas dans quel sens s'établissent les liens.

Toutes les variables sont assez bien représentées sur les 2 premiers facteurs, sauf « origine » et « carburant » c.-à-d. une partie des informations véhiculées par ces variables ne sont pas restituées sur ces facteurs.

Graphique des “squared loadings”

Jauger en un coup d’oeil l’impact des variables sur les facteurs



Mis à part « puissance », le rattachement des variables aux facteurs est relativement tranché.

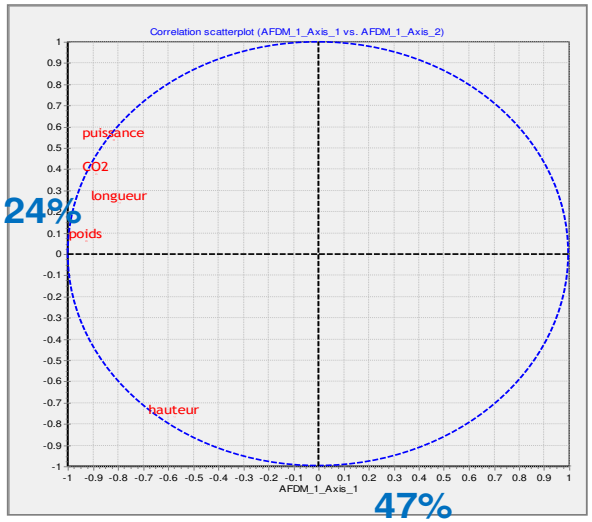
Attribute	Facteur 1			Facteur 2		
-	Coord.	CTR (%)	QLT % (Cumul %)	Coord.	CTR (%)	QLT % (Cumul %)
puissance (*)	**0.67133**	15.7%	*67 % (67 %)*	0.29072	13.7%	*29 % (96 %)*
longueur (*)	**0.63391**	14.8%	*63 % (63 %)*	0.05997	2.8%	*6 % (69 %)*
hauteur (*)	0.33452	7.8%	*33 % (33 %)*	**0.58306**	27.5%	*58 % (92 %)*
poids (*)	**0.86400**	20.2%	*86 % (86 %)*	0.00374	0.2%	*0 % (87 %)*
CO2 (*)	**0.79328**	18.6%	*79 % (79 %)*	0.14363	6.8%	*14 % (94 %)*
origine (**)	**0.69657**	16.3%	*35 % (35 %)*	0.01601	0.8%	*1 % (36 %)*
carburant (**)	0.02826	0.7%	*3 % (3 %)*	**0.40464**	19.1%	*40 % (43 %)*
4X4 (**)	0.25128	5.9%	*25 % (25 %)*	**0.62012**	29.2%	*62 % (87 %)*
Var. Expl.	4.27314		*47 % (47 %)*	2.12189		*24 % (71 %)*

Cercle des corrélations – Variables quantitatives

Sens du lien des variables avec les facteurs

Variables quantitatives - Corrélations

Attribute	Fact.1	Fact.2
puissance	-0.81935	0.53919
longueur	-0.79618	0.24489
hauteur	-0.57838	-0.76359
poids	-0.92952	0.06113
CO2	-0.89066	0.37898



1. (Puissance, longueur, CO2, poids) vont ensemble.
2. A «Facteur 1 » égal [c.-à-d. à « taille » égale],

Est-ce vrai ? Voyons les corrélations calculées sur les données originelles.

(1) OUI, corrélation brute

Y	X	r	r²	t	Pr(>|t|)
puissance	CO2	0.941	0.886	7.889	0.000
longueur	poids	0.911	0.829	6.227	0.000
poids	CO2	0.777	0.604	3.496	0.008
puissance	poids	0.754	0.569	3.248	0.012
puissance	longueur	0.742	0.551	3.130	0.014
longueur	CO2	0.727	0.529	2.994	0.017
hauteur	poids	0.465	0.216	1.486	0.176
hauteur	CO2	0.243	0.059	0.709	0.499
longueur	hauteur	0.193	0.037	0.555	0.594
puissance	hauteur	0.042	0.002	0.119	0.908

(2) OUI, corrélation partielle

Control variables

	Variable
1	AFDM_1_Axis_1

Partial correlation

N	Att.Y	Att.X	r	r²	t	p-value
1	puissance	hauteur	-0.9233	0.85248	-6.36023	0.00038

Sens du lien des modalités avec les facteurs

Sens du lien des modalités avec les facteurs

Opposition « origine = Europe » vs. « origine = Autres ». Surreprésentation des « 4x4 = oui » parmi les « origine = autres »

Opposition « carburant = essence » vs. « 4x4 = oui »

Discrete Attributes - Conditional means and contributions

Attribute	-	Fact.1 Mean	Fact.1 CTR (%)	Fact.2 Mean	Fact.2 CTR (%)
origine	Europe	2.3170	5.9	0.0540	0.0
	France	0.8501	1.6	0.1903	0.3
	Autres	-2.0086	8.8	-0.2174	0.4
	Tot.	-	16.3	-	0.8
carburant	Diesel	-0.2837	0.3	-0.7566	7.6
	Essence	0.4256	0.4	1.1349	11.4
	Tot.	-	0.7	-	19.1
4X4	non	0.5181	1.2	0.5735	5.8
	oui	-2.0724	4.7	-2.2942	23.4
	Tot.	-	5.9	-	29.2

Les contributions des modalités s'additionnent → contribution des variables (qui est cohérent avec le tableau des « squared loadings »).

Variables quantitatives - Corrélations

Attribute	Fact.1	Fact.2
puissance	-0.81935	0.53919
longueur	-0.79618	0.24489
hauteur	-0.57838	-0.76359
poids	-0.92952	0.06113
CO2	-0.89066	0.37898

Les valeurs des corrélations (variables quantitatives) ne peuvent pas être rapprochées directement avec les moyennes conditionnelles : il faut raisonner en termes de **directions**.

(Remarque : ADE4, après transformation des moyennes conditionnelles, place les modalités dans le même repère que les coordonnées des variables quantitatives – Voir plus loin).

Caractérisation des facteurs par les individus

Analyse des proximités entre les individus : Coordonnées des individus, Contribution et qualité de représentation

Contribution et qualité de représentation

Contribution : influence de l'individu « i » dans la construction du facteur « h »

Coordonnée de l'individu « i » sur le facteur « h »

F_{ih}

$$CTR_{ih} = \frac{F_{ih}^2}{n \times \lambda_h}$$

			Contribution		Qualité	
Modele	Coord.1	Coord.2	CTR.1	CTR.2	COS2.1	COS2.2
GOLF	2.32	-0.69	12.57	2.23	0.60	0.05
CITRONC4	1.45	-0.12	4.89	0.07	0.44	0.00
P607	-0.78	1.02	1.42	4.89	0.11	0.18
VELSATIS	-0.54	-0.16	0.69	0.12	0.06	0.01
CITRONC2	3.28	0.03	25.11	0.00	0.73	0.00
CHRYS300	-2.96	2.63	20.47	32.55	0.54	0.43
AUDIA3	2.32	0.80	12.55	2.98	0.58	0.07
OUTLAND	-2.26	-1.84	11.91	15.96	0.54	0.36
PTCRUISER	-0.93	1.09	2.03	5.61	0.18	0.25
SANTA_FE	-1.89	-2.75	8.35	35.59	0.31	0.65

COS² : qualité de représentation de l'individu « i » sur le facteur « h » (cumulable sur « h »)

$$\lambda_h = \frac{\sum_{i=1}^{n} F_{ih}^2}{n}$$

Lambda	4.27314	2.12189

Ex. $\lambda_1 = \frac{2.32^2 + 1.45^2 + \cdots + (-1.89)^2}{10} = 4.27314$

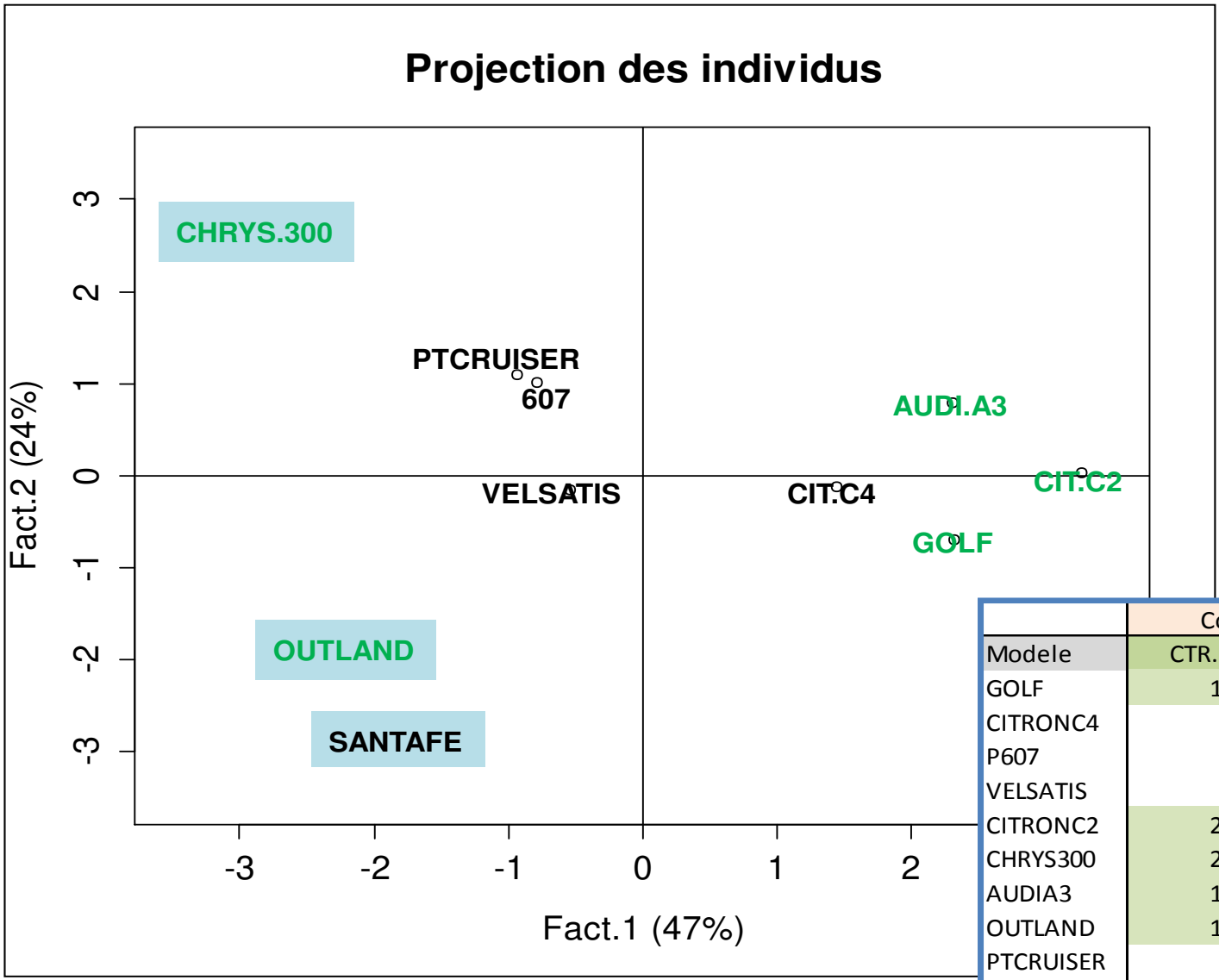
$$COS_{ih}^2 = \frac{F_{ih}^2}{\sum_{k=1}^{P} (z_{ik} - \bar{z}_k)^2} = \frac{F_{ih}^2}{\sum_{h=1}^{H_{max}} F_{ih}^2}$$

Carré de l'écart au barycentre du point « i »

Que l'on peut reproduire si on prend tous les facteurs (H_{max})

Caractérisation des facteurs par les individus

Coordonnées des individus



	Contribution	
Modele	CTR.1	CTR.2
GOLF	12.57	2.23
CITRONC4	4.89	0.07
P607	1.42	4.89
VELSATIS	0.69	0.12
CITRONC2	25.11	0.00
CHRYS300	20.47	32.55
AUDIA3	12.55	2.98
OUTLAND	11.91	15.96
PTCRUISER	2.03	5.61
SANTA_FE	8.35	35.59

Variables illustratives

Variables non utilisées lors de la construction des facteurs. Mais exploitées après coup pour mieux comprendre / commenter les résultats.

Ex. « Prix » est une caractéristique qui intègre des éléments subjectifs (marketing, etc.). Comment la situer par rapport aux caractéristiques objectives des véhicules.

Ex. « Surtaxe » est liée à la politique fiscale de l'administration. Comment la situer ?

Modele	prix	surtaxe
GOLF	19140	non
CITRONC4	23400	non
P607	40550	oui
VELSATIS	38250	oui
CITRONC2	10700	non
CHRYS300	54900	oui
AUDIA3	21630	non
OUTLAND	29990	oui
PTCRUISER	27400	oui
SANTA_FE	27990	oui

Variables illustratives quantitatives

Variables illustratives quantitatives

$$r_y(F_h) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(F_{ih} - \bar{F}_h)}{s_y \times s_{F_h}} = \frac{\frac{1}{n}\sum_{i=1}^{n} F_{ih}(y_i - \bar{y})}{s_y \times \sqrt{\lambda_h}}$$

Calculer les corrélations des variables supplémentaires avec les facteurs. c.-à-d. calculer le coefficient de corrélation entre les coordonnées des « n » individus sur les facteurs et les valeurs prises par la variable illustrative. Il est possible de les placer dans le cercle des corrélations.

CORR	Fact.1	Fact.2
Prix	-0.804	0.455

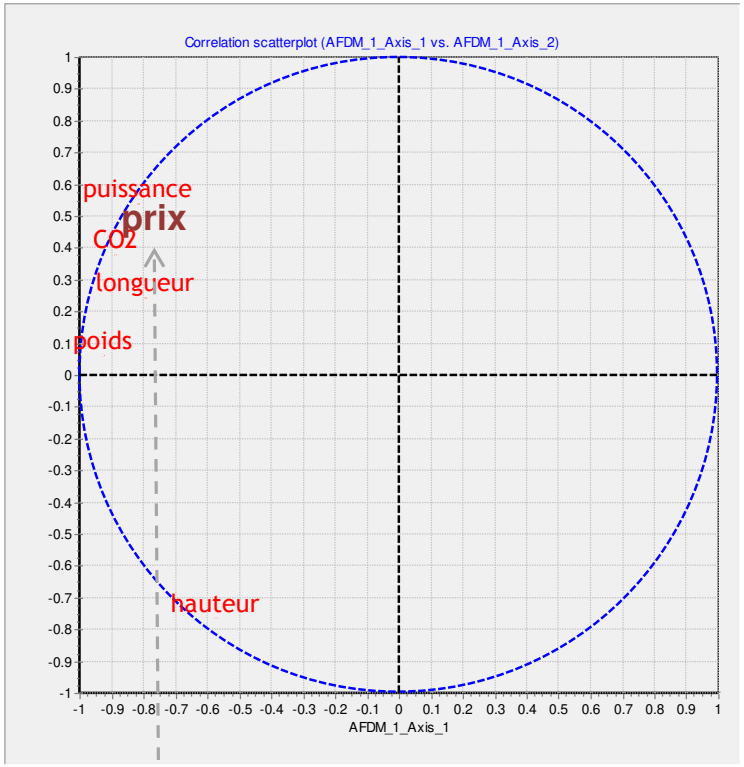
Tester la « significativité » du lien avec la statistique basée sur la transformation de Fisher

$$u_y = \sqrt{n-3} \times \left(\frac{1}{2} \ln \frac{1+r}{1-r} \right)$$

Lien significatif à (~) 5% si

$$|u_y| \geq 2$$

U.Fisher	Fact.1	Fact.2
Prix	-2.937	1.299



Le « prix » est surtout liée à la « taille » (longueur, puissance, etc. c.-à-d. ce qui caractérise le 1er facteur) de la voiture.

Variables illustratives qualititatives

$$\mu_{gh} = \frac{1}{n_g} \sum_{i:y_i=g} F_{ih}$$

		Fact.1		Fact.2	
surtaxe	**n_g**	Moyenne	Valeur.Test	Moyenne	Valeur.Test
non	4	2.339	**2.771**	0.003	0.005
oui	6	-1.559	**-2.771**	-0.002	-0.005

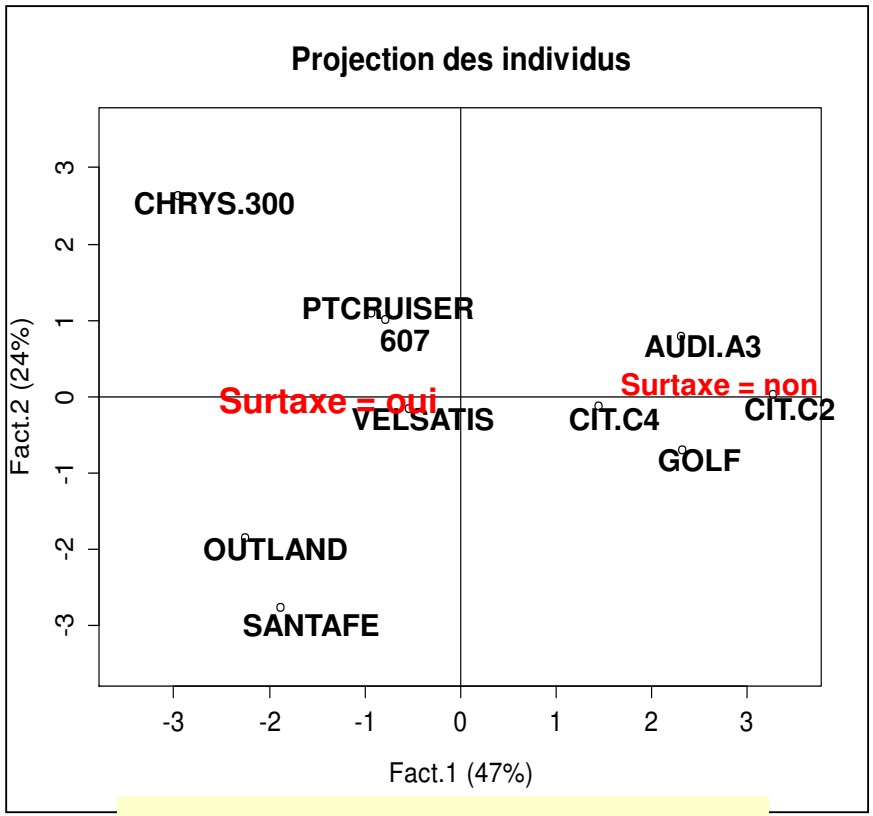
Comparer les moyennes des composantes conditionnellement aux groupes définis par les modalités de la variable illustrative qualitative.

Possibilité de tester la significativité de l'écart par rapport à l'origine (moyenne des composantes = 0) avec la « valeur test » (Morineau, 1984).

$$VT_{gh} = \frac{\mu_{gh} - \overline{F}_h}{\sqrt{\frac{n-n_g}{n-1} \times \frac{s^2_{F_h}}{n_g}}} = \frac{\mu_{gh} - 0}{\sqrt{\frac{n-n_g}{n-1} \times \frac{\lambda_h}{n_g}}}$$

Ecart significatif à (~) 5% si $|VT_{gh}| \geq 2$

Remarque : On pourrait également s'appuyer sur l'ANOVA pour comparer les moyennes, et/ou calculer le rapport de corrélation.



Conclusion : Les « grosses voitures » (au sens du 1er facteur) sont surtaxées.

Individus illustratifs (supplémentaires)

Par opposition aux individus “actifs” utilisés pour la construction des facteurs

Plusieurs raisons possibles :

- Des individus collectés après coup que l’on aimerait situer par rapport à ceux de l’échantillon d’apprentissage (les individus actifs).
- Des individus appartenant à une population différente (ou spécifique) que l’on souhaite positionner.
- Des observations s’avérant atypiques ou trop influentes dans l’AFDM que l’on a préféré écarter. On veut maintenant pouvoir juger de leur positionnement par rapport aux individus actifs.

Modele	puissance	longueur	hauteur	poids	CO2	origine	carburant	4X4
X-TRAIL	136	446	168	1520	190	Autres	Diesel	oui

Plutôt cas N°1 ici, on souhaite situer un véhicule supplémentaire.

Calculs pour un individu supplémentaire

(1)

Modele	puissance	longueur	hauteur	poids	CO2	orig_Autres	orig_Europe	orig_France	carb_Diesel	carb_Essence	4X4_non	4X4_oui
X-TRAIL	136	446	168	1520	190	1	0	0	1	0	0	1

(2)

Variable	puissance	longueur	hauteur	poids	CO2	orig_Autres	orig_Europe	orig_France	carb_Diesel	carb_Essence	4X4_non	4X4_oui
Fact.1	-0.396	-0.385	-0.280	-0.450	-0.431	-0.297	0.242	0.126	-0.051	0.063	0.108	-0.217
Fact.2	0.370	0.168	-0.524	0.042	0.260	-0.065	0.011	0.057	-0.276	0.338	0.242	-0.484

(3)

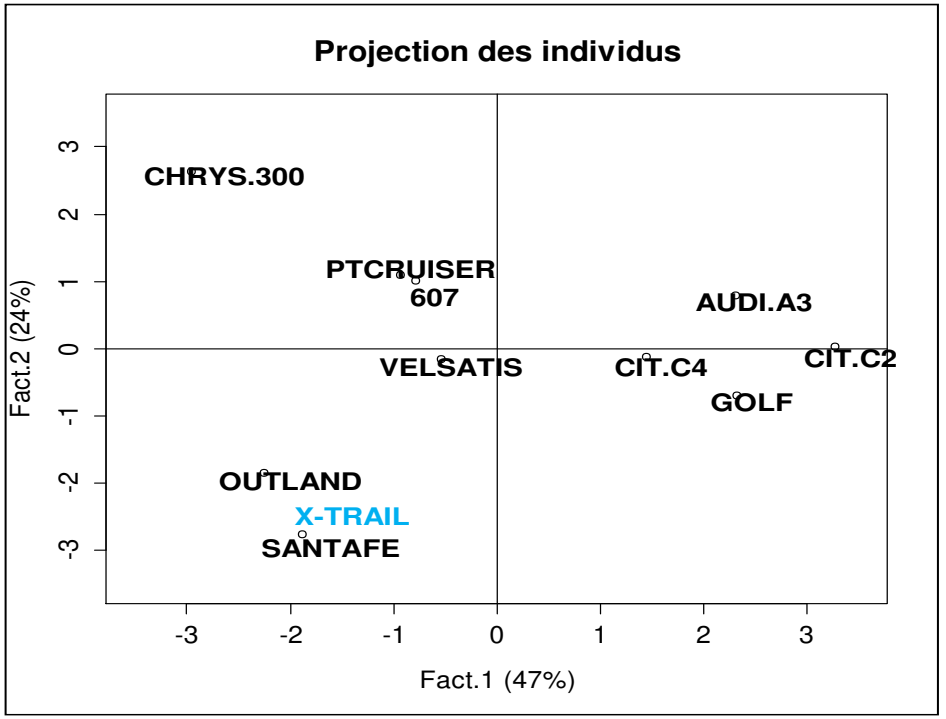
Variable	puissance	longueur	hauteur	poids	CO2	orig_Autres	orig_Europe	orig_France	carb_Diesel	carb_Essence	4X4_non	4X4_oui
Moyenne	162	444.8	153	1497.5	196.5	0.4	0.2	0.4	0.6	0.4	0.8	0.2
Ecart-type	78.88	38.76	9.55	283.75	47.49	0.63	0.45	0.63	0.77	0.63	0.89	0.45

Etapes :

1. Coder en 0/1 les variables qualitatives
2. Appliquer les coefficients fournis par l'analyse (vecteurs propres)
3. Non sans avoir centré (moyenne) et réduit (écart-type) les valeurs

$$F_{x-trail,1} = -0.396 \times \left(\frac{136-162}{78.88}\right) + \cdots - 0.217 \times \left(\frac{1-0.2}{0.45}\right) = -1.32$$

$$F_{x-trail,2} = 0.370 \times \left(\frac{136-162}{78.88}\right) + \cdots - 0.484 \times \left(\frac{1-0.2}{0.45}\right) = -2.51$$



Solution alternative : Discrétisation + AFCM

Une solution souvent citée dans la littérature

- ***Traitement en 2 étapes :***
  1. Découper en classes (discrétiser) les variables quantitatives
  2. Lancer une AFCM sur les variables ainsi homogénéisées
- ***Avantages :***
  1. Possibilité de prise en compte des relations non linéaires
  2. Forme de “nettoyage” des données en éliminant les valeurs extrêmes par ex.
  3. On a le choix du nombre de classes pour équilibrer les influences avec les autres variables qualitatives
  4. L’AFCM est bien maîtrisée et disponible dans de très nombreux logiciels
- ***Inconvénients*** :
  1. Découpage en classes $\Longrightarrow$ perte d’information : dommageable si les variables quantitatives sont nombreuses par rapport aux qualitatives ($C >> D$)
  2. La discrétisation en elle-même est un problème : combien de classes ? comment choisir les bornes de découpage ?

Solution alternative : Discrétisation + AFCM

Modele	puissance	longueur	hauteur	poids	CO2	origine	carburant	4X4
GOLF	A	A	B	A	A	Europe	Diesel	non
CITROEN.C4	B	A	A	A	A	France	Diesel	non
P607	B	B	A	B	B	France	Diesel	non
VELSATIS	B	B	B	B	B	France	Diesel	non
CITROEN.C2	A	A	A	A	A	France	Essence	non
CHRYS300	B	B	B	B	B	Autres	Essence	non
AUDIA3	A	A	A	A	A	Europe	Essence	non
OUTLAND	B	B	B	B	B	Autres	Diesel	oui
PTCRUISER	B	B	B	B	B	Autres	Essence	non
SANTA_FE	A	B	B	B	B	Autres	Diesel	oui

Seuil (médiane)	**144**	**439.5**	**148.5**	**1595**	**192.5**

Tableau de données
après transformation

Ex. de traitement : découpage en
2 classes de fréquence égales
Seuil de découpage = médiane

Résultats de l'ACM

Facteur 1 : opposition basée sur la « taille » (poids, longueur, CO2) → OUI
Facteur 2 : moins évident, opposition « Europe » vs. « France » basée sur la puissance ?

Factors characterization (active variables)

Values	Overall			Factor 1				Factor 2			
Attribute = Value	Mass	Sq.Dist	Inertia	coord	v.test	cos2	ctr (%)	coord	v.test	cos2	ctr (%)
CO2 = A	0.0500	1.5000	0.0750	1.18891	2.912	0.9423	12.05	0.11150	0.273	0.0083	0.29
CO2 = B	0.0750	0.6667	0.0500	-0.79261	-2.912	0.9423	8.03	-0.07434	-0.273	0.0083	0.19
-	-	-	-	-	-	Tot.ctr.	20.09	-	-	Tot.ctr.	0.48
poids = A	0.0500	1.5000	0.0750	1.18891	2.912	0.9423	12.05	0.11150	0.273	0.0083	0.29
poids = B	0.0750	0.6667	0.0500	-0.79261	-2.912	0.9423	8.03	-0.07434	-0.273	0.0083	0.19
-	-	-	-	-	-	Tot.ctr.	20.09	-	-	Tot.ctr.	0.48
longueur = A	0.0500	1.5000	0.0750	1.18891	2.912	0.9423	12.05	0.11150	0.273	0.0083	0.29
longueur = B	0.0750	0.6667	0.0500	-0.79261	-2.912	0.9423	8.03	-0.07434	-0.273	0.0083	0.19
-	-	-	-	-	-	Tot.ctr.	20.09	-	-	Tot.ctr.	0.48
origine = Europe	0.0250	4.0000	0.1000	1.28529	1.928	0.4130	7.04	1.05160	1.577	0.2765	12.73
origine = France	0.0500	1.5000	0.0750	0.29577	0.724	0.0583	0.75	-1.06064	-2.598	0.7500	25.91
origine = Autres	0.0500	1.5000	0.0750	-0.93841	-2.299	0.5871	7.51	0.53484	1.310	0.1907	6.59
-	-	-	-	-	-	Tot.ctr.	15.30	-	-	Tot.ctr.	45.23
puissance = A	0.0500	1.5000	0.0750	0.72629	1.779	0.3517	4.50	0.77907	1.908	0.4046	13.98
puissance = B	0.0750	0.6667	0.0500	-0.48419	-1.779	0.3517	3.00	-0.51938	-1.908	0.4046	9.32
-	-	-	-	-	-	Tot.ctr.	7.50	-	-	Tot.ctr.	23.30
4X4 = non	0.1000	0.2500	0.0250	0.26657	1.599	0.2842	1.21	-0.25865	-1.552	0.2676	3.08
4X4 = oui	0.0250	4.0000	0.1000	-1.06629	-1.599	0.2842	4.85	1.03460	1.552	0.2676	12.33
-	-	-	-	-	-	Tot.ctr.	6.06	-	-	Tot.ctr.	15.41
hauteur = B	0.0750	0.6667	0.0500	-0.55234	-2.029	0.4576	3.90	0.40320	1.481	0.2439	5.62
hauteur = A	0.0500	1.5000	0.0750	0.82850	2.029	0.4576	5.85	-0.60480	-1.481	0.2439	8.42
-	-	-	-	-	-	Tot.ctr.	9.75	-	-	Tot.ctr.	14.04
carburant = Diesel	0.0750	0.6667	0.0500	-0.18893	-0.694	0.0535	0.46	-0.08262	-0.304	0.0102	0.24
carburant = Essence	0.0500	1.5000	0.0750	0.28340	0.694	0.0535	0.68	0.12393	0.304	0.0102	0.35
-	-	-	-	-	-	Tot.ctr.	1.14	-	-	Tot.ctr.	0.59

Logiciels

- Les signes des vecteurs propres sont fixés arbitrairement, ils peuvent être différents d'un logiciel à l'autre. Ce n'est pas un problème.
- Le plus important est que les positions relatives entre les individus (proximités) et les liaisons entre variables/modalités soient préservées.
- Les packages dans R : FactoMiner, Ade4, PCAmixdata, ... , Tanagra

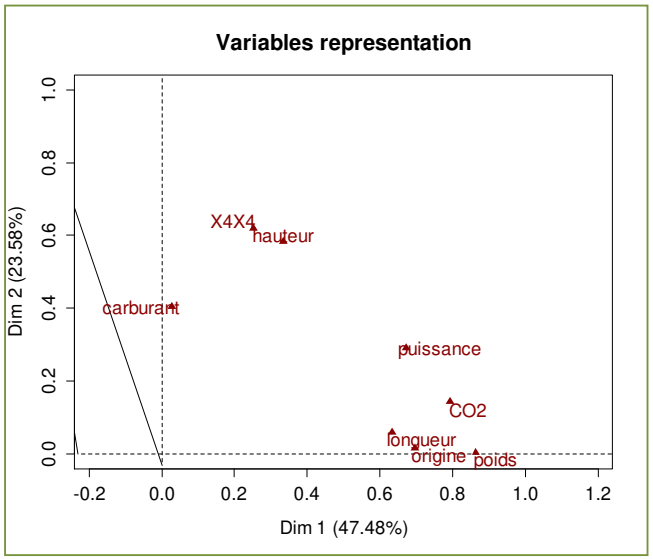
R – Package : FactoMiner

```
#chargement du package
library(FactoMineR)
#lancement de la procédure
afdm.autos <- FAMD(autos,ncp=2)
#affichage des résultats
print(summary(afdm.autos))
```

R – Package « FactoMineR »

Article de référence : Pagès (2004)

Avec quelques graphiques, dont « les squared loadings » (influence des variables sur les facteurs).




```
Call:
FAMD(autos, ncp = 2)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9
Variance               4.273   2.122   1.439   0.836   0.164   0.114   0.034   0.016   0.002
% of var.             47.479  23.577  15.986   9.293   1.823   1.272   0.374   0.176   0.022
Cumulative % of var.  47.479  71.056  87.042  96.335  98.157  99.429  99.803  99.978 100.000

Individuals
                 Dim.1    ctr   cos2    Dim.2    ctr   cos2
GOLF           | -2.318 12.572  0.601 | -0.687  2.226  0.053 |
CITROEN.C4     | -1.445  4.889  0.438 | -0.122  0.070  0.003 |
P607           |  0.780  1.423  0.105 |  1.019  4.889  0.180 |
VELSATIS       |  0.541  0.685  0.064 | -0.160  0.121  0.006 |
CITROEN.C2     | -3.276 25.113  0.732 |  0.025  0.003  0.000 |
CHRYS300       |  2.958 20.472  0.541 |  2.628 32.551  0.427 |
AUDIA3         | -2.316 12.554  0.582 |  0.795  2.981  0.069 |
OUTLAND        |  2.256 11.910  0.536 | -1.840 15.963  0.357 |
PTCRUISER      |  0.932  2.032  0.181 |  1.091  5.608  0.248 |
SANTA_FE       |  1.889  8.350  0.307 | -2.748 35.587  0.649 |

Continuous variables
                 Dim.1    ctr   cos2    Dim.2    ctr   cos2
puissance      |  0.819 15.710  0.671 |  0.539 13.701  0.291 |
longueur       |  0.796 14.835  0.634 |  0.245  2.826  0.060 |
hauteur        |  0.578  7.828  0.335 | -0.764 27.478  0.583 |
poids          |  0.930 20.219  0.864 |  0.061  0.176  0.004 |
CO2            |  0.891 18.564  0.793 |  0.379  6.769  0.144 |

Categories
                 Dim.1    ctr   cos2 v.test    Dim.2    ctr   cos2 v.test
Autres         |  2.009  8.838  0.846  2.380 | -0.217  0.420  0.010 -0.365 |
Europe         | -2.317  5.880  0.685 -1.681 |  0.054  0.013  0.000  0.056 |
France         | -0.850  1.583  0.303 -1.007 |  0.190  0.322  0.015  0.320 |
Diesel         |  0.284  0.265  0.071  0.504 | -0.757  7.628  0.508 -1.908 |
Essence        | -0.426  0.397  0.071 -0.504 |  1.135 11.442  0.508  1.908 |
non            | -0.518  1.176  0.431 -1.504 |  0.574  5.845  0.529  2.362 |
oui            |  2.072  4.704  0.431  1.504 | -2.294 23.380  0.529 -2.362 |
```

Bibliographie utilisée

Notre article de référence

(FactoMineR) Pagès, J., « Analyse Factorielle de Données Mixtes », Revue de Statistique Appliquée, Vol : 52, Issue : 4, pp. 93-111, **2004**.

D'autres articles à l'origine de packages pour R

(ADE4) Hill, M., Smith, A., « Principal Component Analysis of taxonomic data withmulti-state discrete characters », Taxon, 25, pp. 249-255, **1976**.

(PCAmixdata) Kiers, H.A.L., « Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables », Psychometrika, 56, pp. 197-212, **1991**.

Tutoriels accessible en ligne

Champelly, S., « Introduction à l'analyse multivariée (factorielle) sous R », Sept. 2005, pp. 37-42.

Tutoriel Tanagra, « Analyse Factorielle de données mixtes », Sept. 2012.

Husson, F., « Analyse factorielle de données mixtes avec FactoMineR (YouTube) », Avril 2013.