Analyse en Composantes Principales (ACP)

Principes et pratique de l'ACP

Ricco RAKOTOMALALA

Université Lumière Lyon 2

PLAN

Position du problème

Construire un nouveau système de représentation

(composantes principales, axes factoriels, facteurs : combinaisons linéaires des variables originelles)

qui permet synthétiser l'information

Variables « actives » quantitatives c.-à-d. seront utilisées pour la construction des facteurs

$j : 1, \ldots, p$

Les données « autos »

(Saporta, 2006 ; page 428)

$i : 1, \ldots, n$

Individus actifs

Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
Alfasud TI	1350	79	393	161	870	165
Audi 100	1588	85	468	177	1110	160
Simca 1300	1294	68	424	168	1050	152
Citroen GS Club	1222	59	412	161	930	151
Fiat 132	1585	98	439	164	1105	165
Lancia Beta	1297	82	429	169	1080	160
Peugeot 504	1796	79	449	169	1160	154
Renault 16 TL	1565	55	424	163	1010	140
Renault 30	2664	128	452	173	1320	180
Toyota Corolla	1166	55	399	157	815	140
Alfetta 1.66	1570	109	428	162	1060	175
Princess 1800	1798	82	445	172	1160	158
Datsun 200L	1998	115	469	169	1370	160
Taunus 2000	1993	98	438	170	1080	167
Rancho	1442	80	431	166	1129	144
Mazda 9295	1769	83	440	165	1095	165
Opel Rekord	1979	100	459	173	1120	173
Lada 1300	1294	68	404	161	955	140

x_{ij}

Questions :

(1) Quelles sont les véhicules qui se ressemblent ? (proximité entre les individus)

(2) Sur quelles variables sont fondées les ressemblances / dissemblances

(3) Quelles sont les relations entre les variables

Position du problème (1)

Analyse des proximités entre les individus

Que voit-on dans ce graphique ?

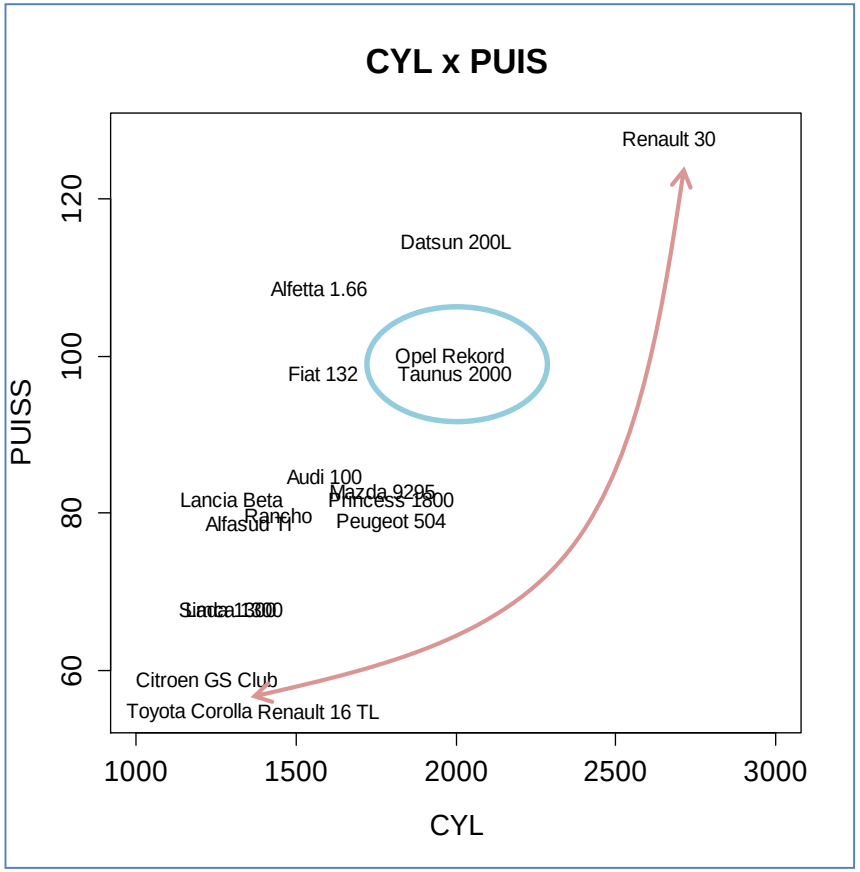
1. Les variables CYL et PUISS sont liées.
2. « Opel Reckord » et « Taunus 2000 (Ford) » ont le même profil (caractéristiques)
3. « Renault 30 » et « Toyota Corolla » ont des profils opposés…

Un graphique ne fait que révéler des informations présentes dans le tableau de données !

Modele	CYL	PUISS
Toyota Corolla	**1166**	**55**
Citroen GS Club	1222	59
Simca 1300	1294	68
Lada 1300	1294	68
Lancia Beta	1297	82
Alfasud TI	1350	79
Rancho	1442	80
Renault 16 TL	1565	55
Alfetta 1.66	1570	109
Fiat 132	1585	98
Audi 100	1588	85
Mazda 9295	1769	83
Peugeot 504	1796	79
Princess 1800	1798	82
Opel Rekord	**1979**	**100**
Taunus 2000	**1993**	**98**
Datsun 200L	1998	115
Renault 30	**2664**	**128**

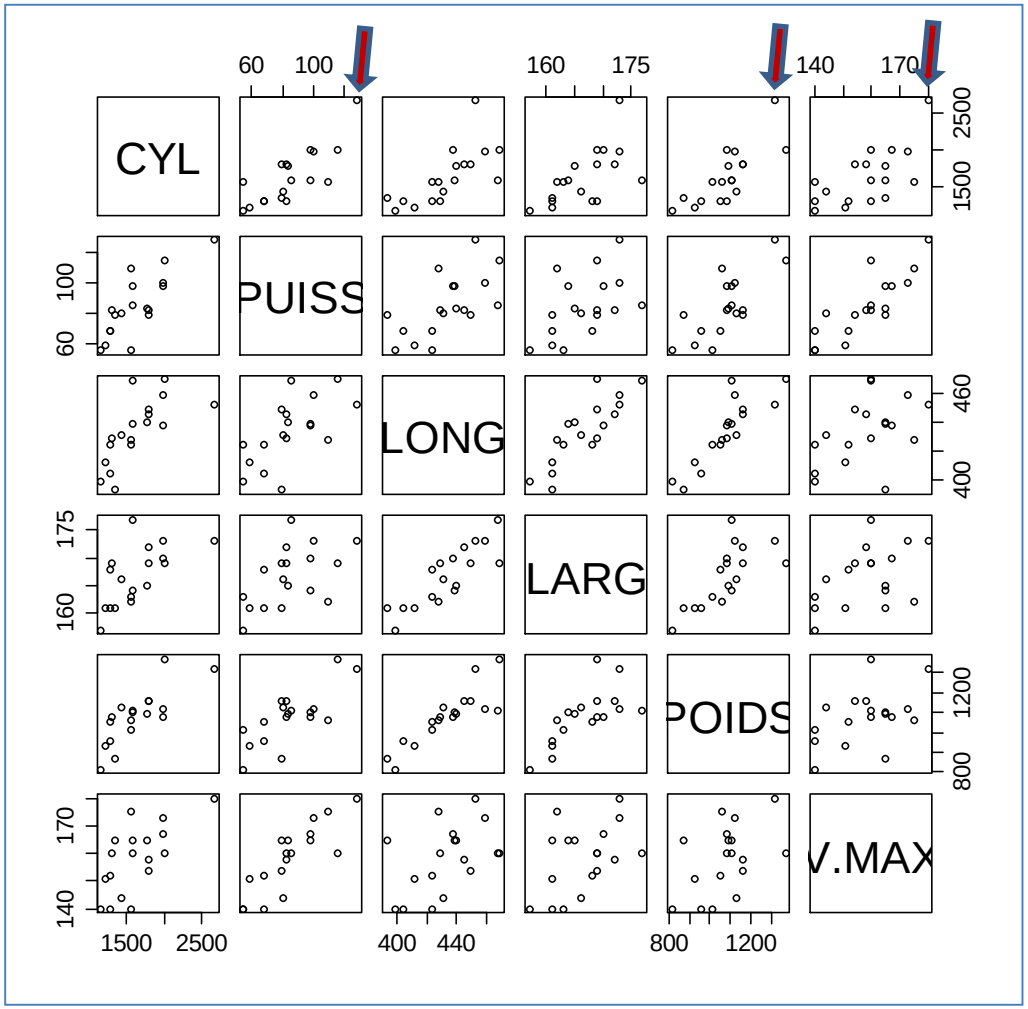
Tableau trié selon CYL

Positionnement des individus (2 variables)



Que faire si on veut prendre en compte (p > 2) variables **simultanément** ?

Positionnement des individus (p >2)



Impossible de créer un nuage à « p » dimensions.

On pourrait croiser les variables 2 à 2, mais :

1. Très difficile de surveiller plusieurs cadrans en même temps.
2. Etiqueter les points rendrait le tout illisible.

Ce type de représentation n'est utile que pour effectuer un diagnostic rapide et repérer les points atypiques.

Ex. Renault 30 : le plus gros moteur, la plus puissante, une des plus lourdes, la plus rapide.

Positionnement des individus – Principe de l'ACP (1) – Notion d'inertie

Principe : Construire un système de représentation de dimension réduite (q << p) qui préserve les **distances** entre les individus. On peut la voir comme une compression avec perte (contrôlée) de l'information.

Distance euclidienne entre 2 individus (i, i')

$$d^2(i,i') = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2$$

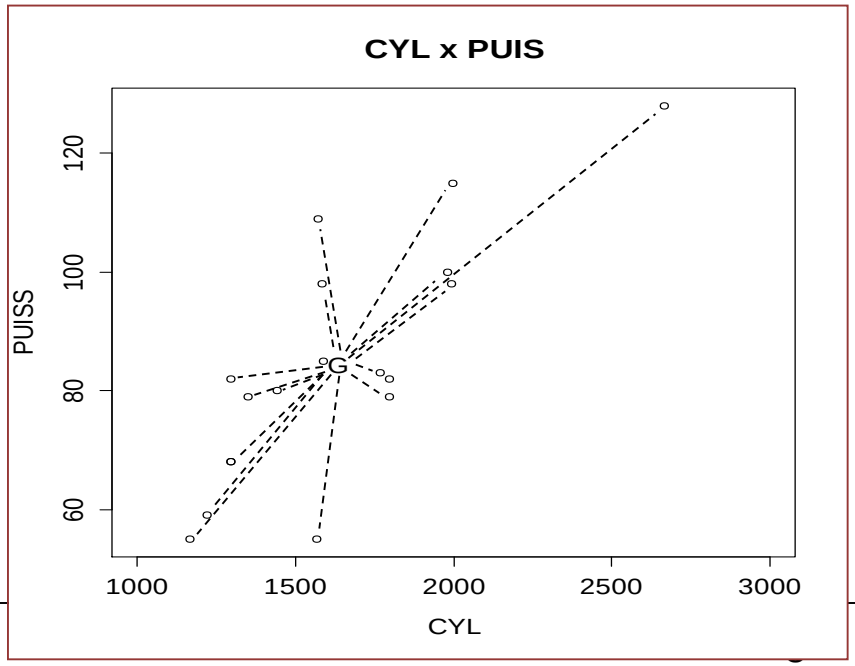
Un critère global : distance entre l'ensemble des individus pris 2 à 2, **inertie** du nuage de points dans l'espace originel. Elle traduit la quantité d'information disponible.

$$I_p = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d^2(i,i')$$

Autre écriture de l'inertie : écart par rapport au barycentre G (vecteur constitué des moyennes des p variables)

$$I_p = \frac{1}{n} \sum_{i=1}^{n} d^2(i,G)$$

L'inertie indique la dispersion autour du barycentre, c'est une variance multidimensionnelle (calculée sur p dimensions)



Positionnement des individus – Principe de l'ACP (1) – Régression orthogonale

Habituellement on (a) centre et (b) réduit les variables. On parle d'ACP normée.

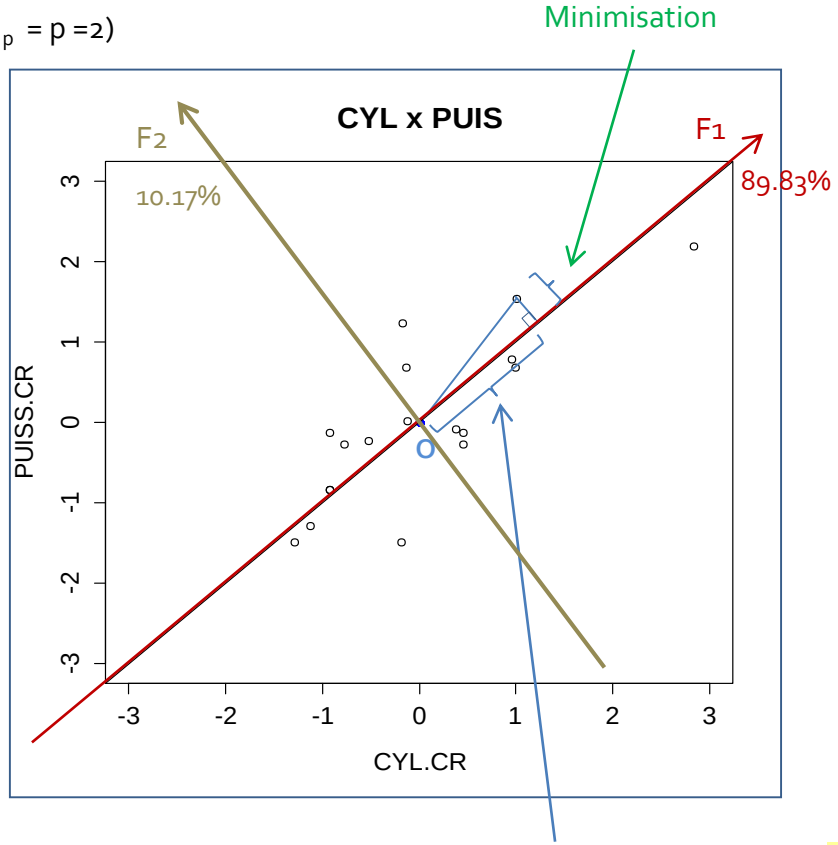
(a) Pour que G soit situé à l'origine [obligatoire]

(b) Pour rendre comparables les variables exprimées sur des échelles (unités) différentes [non obligatoire]

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad \left\{ \begin{array}{l} \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ s_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2 \end{array} \right.$$

Cas particulier de 2 variables c.r.

(I_p = p =2)



(1) Trouver la première composante F_1 qui maximise l'écartement global des points par rapport à l'origine :

$$\lambda_1 = \frac{1}{n}\sum_{i=1}^{n} F_{i1}^2 = 1.796628$$

$\frac{\lambda_1}{I_p} = 89.83\%$ est la part d'inertie expliquée par le 1er axe factoriel (ou 1ère composante)

(2) Trouver la 2nde composante F_2 qui traite l'inertie non-expliquée (résiduelle) par F_1 (par conséquent, F_2 est non corrélée avec F_1)

$$\lambda_2 = \frac{1}{n}\sum_{i=1}^{n} F_{i2}^2 = 0.203372 \qquad \left(\frac{\lambda_2}{I_p} = 10.17\%\right)$$

(3) Et bien évidemment : $\sum_{k=1}^{p} \lambda_k = 1.797 + 0.203 = 2 = I_p$

Les inerties expliquées s'additionnent. Prendre tous les « p » facteurs possibles permet de récupérer toute l'information disponible !

Préservation des proximités dans le repère réduit

(1) Les proximités entre individus sont préservées si on prend un nombre q de composantes suffisamment représentatives (en terme de % d'inertie exprimée)

(2) Si on prend les « p » facteurs, on retrouve les distances dans le repère originel

Distances dans le repère originel (variables centrées et réduites)

$$d^2(1,2) = (-1.2814 - (-1.1273))^2 + (-1.4953 - (-1.2933))^2 = 0.06455$$
$$d^2(2,6) = 1.14415$$
$$d^2(1,6) = 1.72529$$

	Modele	CYL	PUISS
1	Toyota Corolla	-1.2814	-1.4953
2	Citroen GS Club	-1.1273	-1.2933
3	Simca 1300	-0.9292	-0.8389
4	Lada 1300	-0.9292	-0.8389
5	Lancia Beta	-0.9209	-0.1319
6	Alfasud TI	-0.7751	-0.2834
7	Rancho	-0.5219	-0.2329
8	Renault 16 TL	-0.1835	-1.4953
9	Alfetta 1.66	-0.1697	1.2316
10	Fiat 132	-0.1284	0.6761
11	Audi 100	-0.1202	0.0196
12	Mazda 9295	0.3779	-0.0814
13	Peugeot 504	0.4522	-0.2834
14	Princess 1800	0.4577	-0.1319
15	Opel Rekord	0.9558	0.7771
16	Taunus 2000	0.9943	0.6761
17	Datsun 200L	1.0081	1.5346
18	Renault 30	2.8408	2.1911

Données centrées et réduites

	Modele	F1 (89.83%)	F2 (10.17%)
1	Toyota Corolla	1.9635	0.1513
2	Citroen GS Club	1.7117	0.1174
3	Simca 1300	1.2502	-0.0639
4	Lada 1300	1.2502	-0.0639
5	Lancia Beta	0.7444	-0.5580
6	Alfasud TI	0.7484	-0.3477
7	Rancho	0.5337	-0.2044
8	Renault 16 TL	1.1871	0.9276
9	Alfetta 1.66	-0.7509	-0.9909
10	Fiat 132	-0.3873	-0.5689
11	Audi 100	0.0711	-0.0989
12	Mazda 9295	-0.2097	0.3248
13	Peugeot 504	-0.1194	0.5201
14	Princess 1800	-0.2304	0.4169
15	Opel Rekord	-1.2254	0.1263
16	Taunus 2000	-1.1812	0.2250
17	Datsun 200L	-1.7980	-0.3723
18	Renault 30	-3.5581	0.4594

Coordonnées dans le repère factoriel

Si on ne tient compte que de la 1ère composante ($\lambda_1 = 89.83\%$), les distances sont approximées. On constate néanmoins que les proximités sont assez bien respectées (globalement).

$$d^2_{\{F_1\}}(1,2) = (1.9335 - 1.7117)^2 = 0.06340$$
$$d^2_{\{F_1\}}(2,6) = 0.92783$$
$$d^2_{\{F_1\}}(1,6) = 1.147632$$

Si on tient compte des 2 composantes, on retrouve les distances exactes entre les individus.

$$d^2_{\{F_1,F_2\}}(1,2) = (1.9635 - 1.7117)^2 + (0.1513 - 0.1174)^2 = 0.06455$$
$$d^2_{\{F_1,F_2\}}(2,6) = 1.14415$$
$$d^2_{\{F_1,F_2\}}(1,6) = 1.72529$$

Une des questions clés de l'ACP est de définir le nombre de composantes « q » à retenir pour obtenir une approximation suffisamment satisfaisante !!!

Position du problème (2)

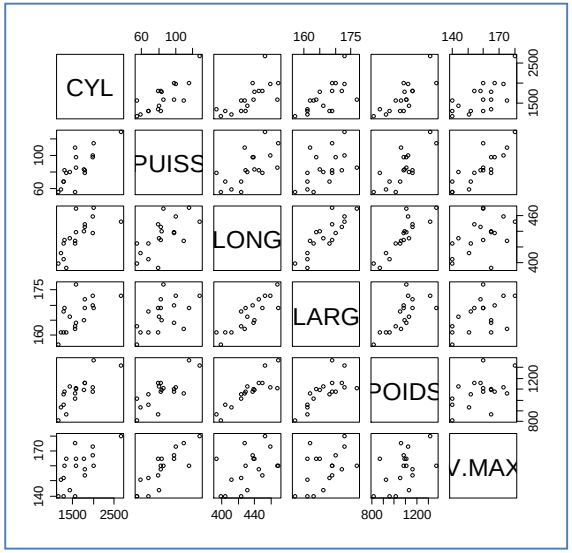
Analyse des relations entre les variables

Le coefficient de corrélation mesure la liaison (linéaire) entre deux variables X_j et X_m

$$r_{jm} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{im} - \bar{x}_m\right)}{s_j \times s_m}$$

Matrice des corrélations **R** sur les données « autos »

CORR	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.797	0.701	0.630	0.789	0.665
PUISS		1	0.641	0.521	0.765	0.844
LONG			1	0.849	0.868	0.476
LARG				1	0.717	0.473
POIDS					1	0.478
V.MAX						1



Elle traduit numériquement ce que l'on peut observer dans les graphiques croisés des variables

On peut essayer de la réorganiser manuellement pour mieux faire apparaître les « blocs » de variables mais….

	POIDS	CYL	PUISS	LONG	LARG	V.MAX
POIDS	1.000	0.789	0.765	0.868	0.717	0.478
CYL		1.000	0.797	0.701	0.630	0.665
PUISS			1.000	0.641	0.521	0.844
LONG				1.000	0.849	0.476
LARG					1.000	0.473
V.MAX						1.000

(1) Ce ne sera jamais parfait

(2) La manipulation est inextricable dès que le nombre de variables est élevé

Construire la première composante F_1 qui permet de maximiser le carré de sa corrélation avec les variables de la base de données

$$\lambda_1 = \sum_{j=1}^{p} r_j^2(F_1)$$

Habituellement, Inertie totale = Somme des variances des variables
Lorsque les données sont réduites (ACP normée), Inertie totale = Trace(R) = p

$$I_p = p$$

⇒ Part d'inertie expliquée par F_1 = $\frac{\lambda_1}{p}$

De nouveau, on observe la décomposition de l'information en composantes non corrélées (orthogonales)

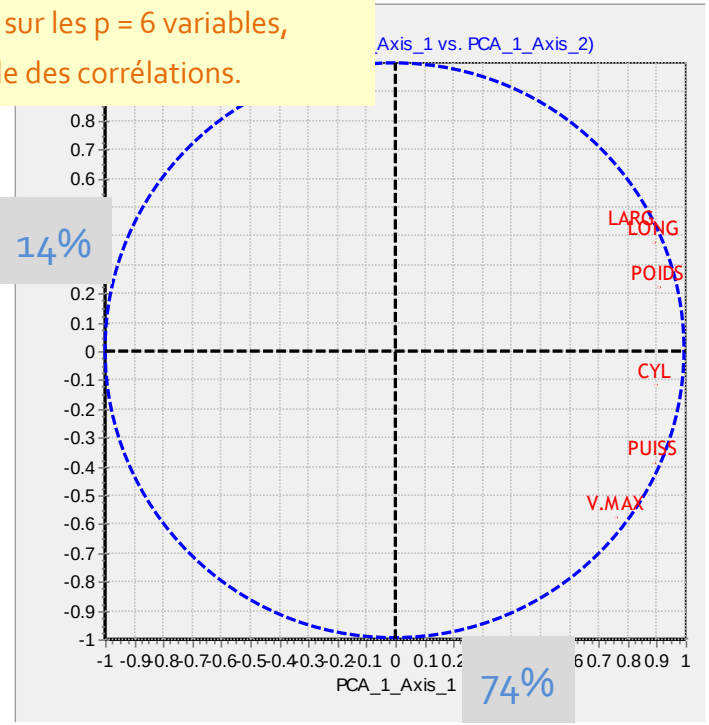
$$\sum_{k=1}^{p} \lambda_k = p$$

Exemple de traitement pour les p = 6 variables de la base de données

Axis	Eigen value	Proportion (%)	Cumulative (%)
1	4.421	73.68%	73.68%
2	0.856	14.27%	87.95%
3	0.373	6.22%	94.17%
4	0.214	3.57%	97.73%
5	0.093	1.55%	99.28%
6	0.043	0.72%	100.00%
Tot.	6	-	-

Relations entres variables – Principe de l'ACP (2) – Approximation des corrélations

ACP sur les p = 6 variables, cercle des corrélations.



Liaison de la variable « poids » avec le 1er axe

$$r_{poids}(F_1) = 0.905 \quad \text{et} \quad r^2_{poids}(F_1) = 0.819$$

La représentation de la variable n'est pas complète, on a besoin d'un second facteur F_2

$$r_{poids}(F_2) = 0.225 \quad \text{et} \quad r^2_{poids}(F_2) = 0.050$$

Si on exploite tous les « p » facteurs

$$\sum_{k=1}^{p} r^2_{poids}(F_k) = 0.819 + 0.050 + \cdots = 1$$

L'ACP produit aussi une approximation dans l'espace des variables (**approximation des corrélations**)

[Ex. si on ne prend en compte que « q = 1 » facteur]

$$\begin{cases} r_{poids,cyl} = 0.789 \\ r_{poids,cyl}(F_1) = \sum_{k=1}^{q=1} r_{poids}(F_k) \times r_{cyl}(F_k) = 0.90519 \times 0.89346 = 0.809 \end{cases}$$

Approximation assez bonne parce que POIDS et CYL sont bien représentées sur le 1er facteur

$$\begin{cases} r_{poids,v.max} = 0.478 \\ r_{poids,v.max}(F_1) = 0.90519 \times 0.75471 = 0.683 \end{cases}$$

Approximation mauvaise parce que V.MAX est mal représentée sur le premier facteur [(0.75471^2)=57% de l'information seulement]

Calculs

Les mains dans le cambouis : comment sont obtenus les résultats de l'ACP ?

Construire un ensemble de composantes (F_1, F_2, …, F_k,…), combinaisons linéaires des variables originelles (centrées et réduites), dont on peut apprécier la qualité de restitution de l'information à travers l'inertie reproduite (λ_k)

$$\begin{cases} F_1 = a_{11}z_1 + a_{21}z_2 + \cdots + a_{p1}z_p \ (\lambda_1) \\ \vdots \\ F_k = a_{1k}z_1 + a_{2k}z_2 + \cdots + a_{pk}z_p \ (\lambda_k) \\ \vdots \\ F_p = a_{1p}z_1 + a_{2p}z_2 + \cdots + a_{pp}z_p \ (\lambda_p) \end{cases}$$

Comment obtenir les coefficients « a_{jk} » à partir des données ?

Qui permettent de calculer les coordonnées des individus dans le repère factoriel, et de juger de leur proximité dans les différents plans factoriels

$$F_{ik} = a_{1k}z_{i1} + a_{2k}z_{i2} + \cdots + a_{pk}z_{ip}$$

Valeur de la variable Z_2 (X_2 après centrage et réduction) pour l'individu n°i

Que l'on interprétera en calculant leur corrélations (et autres indicateurs dérivés : CTR et COS²) avec les variables originelles (X_1, X_2, …, X_p)

$$r_{x_j}(F_k)$$

Plus la corrélation est élevée en valeur absolue, plus forte est l'influence de la variable sur le facteur

Calcul via la diagonalisation de la matrice des corrélations

Calcul uniquement dans l'espace des variables,
mais résultats disponibles pour les deux points de vue (individus et variables)

```r
#chargement du fichier de données
autos <- read.table(file="autos.txt",sep="\t",row.names=1,header=

#calcul de la matrice des corrélations
autos.cor <- cor(autos)
print(autos.cor)

#trace de la matrice = inertie totale
print(sum(diag(autos.cor)))

#diagonalisation avec la fonction eigen
autos.eigen <- eigen(autos.cor)
print(autos.eigen)

#calcul des corrélations des variables avec les composantes
cor.factors <- NULL
for (j in 1:ncol(autos)){
  rf <- sqrt(autos.eigen$values[j])*autos.eigen$vectors[,j]
  cor.factors <- cbind(cor.factors,rf)
}
rownames(cor.factors) <- colnames(autos)
colnames(cor.factors) <- paste("F",1:ncol(autos),sep="")

#affichage des 2 premières composantes seulement
print(cor.factors[,1:2])
```

	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1.0000000	0.7966277	0.7014619	0.6297572	0.7889520	0.6649340
PUISS	0.7966277	1.0000000	0.6413624	0.5208320	0.7652930	0.8443795
LONG	0.7014619	0.6413624	1.0000000	0.8492664	0.8680903	0.4759285
LARG	0.6297572	0.5208320	0.8492664	1.0000000	0.7168739	0.4729453
POIDS	0.7889520	0.7652930	0.8680903	0.7168739	1.0000000	0.4775956
V.MAX	0.6649340	0.8443795	0.4759285	0.4729453	0.4775956	1.0000000

```
> print(sum(diag(autos.cor)))
[1] 6
```

Valeurs propres = λ_k

```
> print(autos.eigen)
$values
[1] 4.42085806 0.85606229 0.37306608 0.21392209 0.09280121 0.04329027

$vectors
           [,1]       [,2]        [,3]       [,4]       [,5]        [,6]
[1,] -0.4249360  0.1241911 -0.35361252  0.8077865 -0.1515800 -0.05889517
[2,] -0.4217944  0.4157739 -0.18492049 -0.3577920  0.2937346 -0.63303302
[3,] -0.4214599 -0.4118177  0.06763394 -0.2797523 -0.7305690 -0.19029153
[4,] -0.3869222 -0.4460870  0.60486812  0.2115694  0.4781901 -0.10956624
[5,] -0.4305120 -0.2426758 -0.48439601 -0.3017114  0.3045584  0.58081220
[6,] -0.3589443  0.6198626  0.48547226 -0.0735743 -0.1886551  0.45852167
```

Vecteurs propres = a_{jk}

	F1	F2
CYL	-0.8934635	0.1149061
PUISS	-0.8868580	0.3846891
LONG	-0.8861548	-0.3810287
LARG	-0.8135364	-0.4127359
POIDS	-0.9051875	-0.2245325
V.MAX	-0.7547104	0.5735194

Corrélations variables x facteurs

$$r_{x_j}(F_k) = \sqrt{\lambda_k} \times a_{jk}$$

Calcul via la décomposition en valeurs singulières de la matrice des données centrées et réduites

Montre bien le caractère dual de l'analyse

z_{ij}

```
> print(head(autos.cr,3))
                CYL       PUISS        LONG       LARG       POIDS      V.MAX
Alfasud TI -0.7750989 -0.28335818 -1.8850808 -1.0973453 -1.5690068  0.5697604
Audi 100   -0.1201633  0.01963869  1.6058095  2.0010414  0.2341614  0.1459717
Simca 1300 -0.9292014 -0.83885242 -0.4421794  0.2581989 -0.2166306 -0.5320903
```

```
#affichage des 3 premières obs. de Z
print(head(autos.cr,3))

#décomposition en valeurs singulières
svd.autos <- svd(autos.cr)
print(svd.autos,digits=3)

#calcul des inerties associées aux composantes
print(svd.autos$d^2/nrow(autos))
```

On obtient les coordonnées factorielles des individus avec $F_{ik} = \delta_k \times u_{ik}$

```
> print(svd.autos,digits=3)
$d
[1] 8.921 3.925 2.591 1.962 1.292 0.883

$u
        [,1]    [,2]     [,3]     [,4]    [,5]    [,6]
 [1,] -0.2398 -0.4549 -0.22068 -0.10290  0.2332 -0.0611
 [2,]  0.1750  0.3890 -0.50756  0.10771 -0.1149  0.3707
 [3,] -0.1255  0.1718 -0.17620  0.08542  0.2904 -0.3079
 [4,] -0.2885 -0.0288 -0.05733  0.00884 -0.1755 -0.2985
 [5,]  0.0480 -0.1772  0.07459  0.31991 -0.2039  0.0421
 [6,] -0.0341  0.0500 -0.26079  0.28331  0.3444 -0.2267
 [7,]  0.0767  0.2377  0.09911 -0.10352 -0.1614 -0.1743
 [8,] -0.2184  0.2498  0.23909 -0.32122 -0.2268 -0.1231
 [9,]  0.4943 -0.2710  0.22904 -0.43176  0.2901 -0.0498
[10,] -0.4468 -0.0602  0.11698 -0.13511 -0.2154  0.3726
[11,]  0.0491 -0.4872 -0.00963  0.38675 -0.1301  0.0614
[12,]  0.1141  0.2144 -0.08359 -0.15463  0.1430 -0.2095
[13,]  0.3297  0.1424  0.48005  0.39350 -0.0421  0.0649
[14,]  0.1474 -0.1239 -0.10906 -0.29671  0.0516  0.2867
[15,] -0.0775  0.2287  0.24250  0.18231  0.2918  0.1377
[16,]  0.0432 -0.0907  0.02917 -0.05244 -0.4078 -0.3838
[17,]  0.2567 -0.0266 -0.30732 -0.12044 -0.2619  0.1775
[18,] -0.3036  0.0366  0.22163 -0.04902  0.2954  0.3211

$v
     [,1]   [,2]   [,3]    [,4]   [,5]   [,6]
[1,] 0.425 -0.124  0.3536 -0.8078 -0.152  0.0589
[2,] 0.422 -0.416  0.1849  0.3578  0.294  0.6330
[3,] 0.421  0.412 -0.0676  0.2798 -0.731  0.1903
[4,] 0.387  0.446 -0.6049 -0.2116  0.478  0.1096
[5,] 0.431  0.243  0.4844  0.3017  0.305 -0.5808
[6,] 0.359 -0.620 -0.4855  0.0736 -0.189 -0.4585
```

Principe de la SVD

$$Z = U \Delta V^T \quad \text{avec} \quad \begin{cases} Z\,\vec{v}_k = \delta_k\,\vec{u}_k \\ Z^T\vec{u}_k = \delta_k\,\vec{v}_k \end{cases}$$

V correspond aux vecteurs propres c.-à-d. les coef. a_{jk}

Calcul des inerties : $\lambda_k = \dfrac{\delta_k^2}{n}$

```
> print(svd.autos$d^2/nrow(autos))
[1] 4.42085806 0.85606229 0.37306608 0.21392209 0.09280121 0.04329027
```

Pratique de l'ACP

Que lire et comment lire les résultats de l'ACP ?

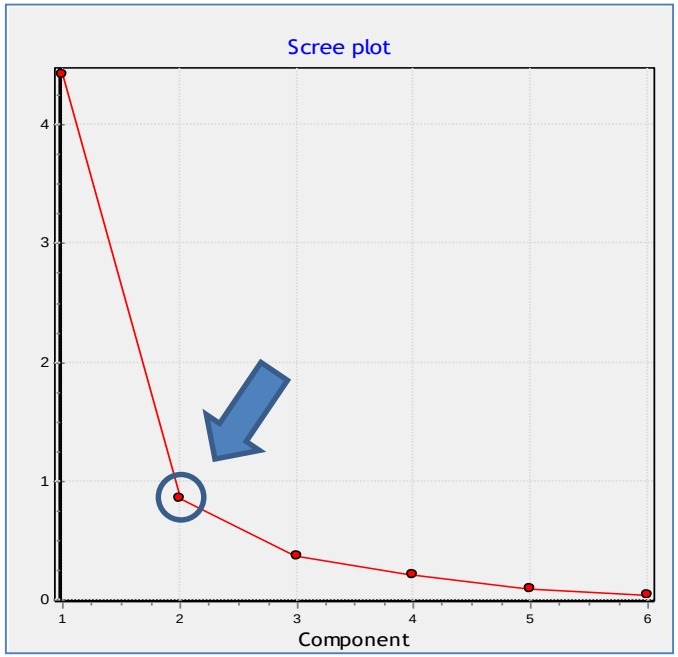
Détermination du nombre de composantes à retenir

(1) Tableau des valeurs propres

Axis	Eigen value	Difference	Proportion (%)	Cumulative (%)
1	4.420858	3.564796	73.68%	73.68%
2	0.856062	0.482996	14.27%	87.95%
3	0.373066	0.159144	6.22%	94.17%
4	0.213922	0.121121	3.57%	97.73%
5	0.092801	0.049511	1.55%	99.28%
6	0.04329	-	0.72%	100.00%
Tot.	6	-	-	-

Indications : (1) sur l'importance des composantes, (2) sur l'évolution de l'importance cumulée, (3) sur la qualité de l'information restituée par les « q » premiers facteurs.

(2) Eboulis des valeurs propres : « scree plot »



« Règle du coude » de Cattell, négliger les composantes qui emmènent peu d'informations additionnelles. Très performante lorsqu'il y a des « blocs » de variables. Fournit surtout des scénarios de solutions.

Problème : Intégrer le coude dans la sélection ? Ici, q = 2 ou q = 1 ? Tout dépend de la valeur propre associée au coude, si elle est faible, il faut exclure la composante associée.

Mais, en pratique, (a) il faut au moins « q = 2 » afin de pouvoir réaliser les représentations graphiques; (b) il faut aussi pouvoir interpréter les composantes.

Règle de Kaiser-Guttman : si les variables sont indépendantes deux à deux, les valeurs propres λ_k seraient toutes égales à 1.

Remarque 1 : cette règle ne tient pas compte du tout des caractéristiques des données.

Remarque 2 : On peut aussi voir le seuil « 1 » comme la moyenne des valeurs propres.

Axis	Eigen value	Difference	Proportion (%)	Cumulative (%)
1	**4.420858**	3.564796	73.68%	73.68%
2	0.856062	0.482996	14.27%	87.95%
3	0.373066	0.159144	6.22%	94.17%
4	0.213922	0.121121	3.57%	97.73%
5	0.092801	0.049511	1.55%	99.28%
6	0.04329	-	0.72%	100.00%
Tot.	6	-	-	-

Règle de Karlis-Saporta-Spinaki : rendre la règle plus stricte en tenant compte des caractéristiques (n et p) des données.

$$\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}$$

A droite, nous avons 2 x écart-type des v.p. sous Ho ≈ un test unilatéral à 5%

$$seuil = 1 + 2\sqrt{\frac{p-1}{n-1}} = 1 + 2\sqrt{\frac{6-1}{18-1}} = 2.08465$$

Axis	Eigen value	Difference	Proportion (%)	Cumulative (%)
1	**4.420858**	3.564796	73.68%	73.68%
2	0.856062	0.482996	14.27%	87.95%
3	0.373066	0.159144	6.22%	94.17%
4	0.213922	0.121121	3.57%	97.73%
5	0.092801	0.049511	1.55%	99.28%
6	0.04329	-	0.72%	100.00%
Tot.	6	-	-	-

Test des « bâtons brisés » de Frontier (1976) et Legendre-Legendre (1983) : si l'inertie était répartie aléatoirement sur les axes, la distribution des v.p. suivrait la loi des « bâtons brisés ».

Problème : les tables sont rarement accessibles. Heureusement les valeurs critiques à 5% peuvent être obtenues très facilement.

$$b_k = \sum_{m=k}^{p} \frac{1}{m}$$

La composante est validée si : $\lambda_k > b_k$

$$b_1 = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} = 2.45$$

Eigenvalues - Significance		
Axis	Eigenvalue	Broken-stick critical values
1	4.420858	2.45
2	0.856062	1.45
3	0.373066	0.95
4	0.213922	0.616667
5	0.092801	0.366667
6	0.04329	0.166667

$$b_3 = \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} = 0.95$$

Toutes les approches sont cohérentes : q = 1 seul facteur semble suffire dans cette étude. Par commodité (hum, pas si sûr, cf. interprétation et rotation des axes), on en choisira q = 2.

Caractérisation des composantes par les variables

Analyse des relations entre les variables via les composantes

Corrélation : degré de liaison de la variable avec la composante

$$\sum_{j=1}^{p} r_{x_j}^2(F_k) = \lambda_k$$

Contributions : influence de la variable dans la définition de la composante (rarement fournie car redondante avec CORR et COS²)

$$CTR_{jk} = \frac{r_{x_j}^2(F_k)}{\lambda_k}; \sum_{j=1}^{p} CTR_{jk} = 1$$

Cosinus carré : qualité de représentation de la variable sur la composante. On peut la cumuler sur les q premières composantes.

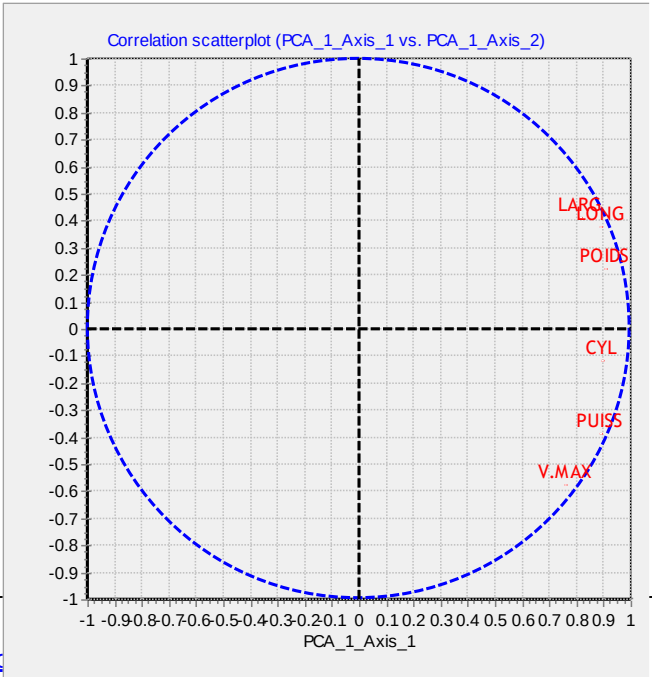
$$COS_{jk}^2 = r_{x_j}^2(F_k)$$

$$COS_{jq}^2 = \sum_{k=1}^{q} COS_{jk}^2$$

$$\sum_{k=1}^{p} COS_{jk}^2 = 1$$

	Axis_1			Axis_2		
	Corr.	CTR (%)	COS² (%)	Corr.	CTR (%)	COS² (%)
POIDS	0.905	19%	*82 % (82 %)*	0.225	6%	*5 % (87 %)*
CYL	0.893	18%	*80 % (80 %)*	-0.115	2%	*1 % (81 %)*
PUISS	0.887	18%	*79 % (79 %)*	-0.385	17%	*15 % (93 %)*
LONG	0.886	18%	*79 % (79 %)*	0.381	17%	*15 % (93 %)*
LARG	0.814	15%	*66 % (66 %)*	0.413	20%	*17 % (83 %)*
V.MAX	0.755	13%	*57 % (57 %)*	-0.574	38%	*33 % (90 %)*
Var. Expl.	4.42086		*74 % (74 %)*	0.85606		*14 % (88 %)*

On utilise souvent le « cercle des corrélations » pour obtenir une vision synthétique immédiate.



On observe (axe 1 : 74%) un « effet taille » marqué, que l'on peut lier à « l'encombrement / gamme » des véhicules ; mais aussi (axe 2 : 14%), une caractérisation par les performances (sportivité).

A nombre de composantes fixé, on peut comparer les corrélations brutes calculées (en bleu) sur les données originelles, et celles estimées à partir du repère factoriel (en vert). Nous avons choisi q = 2 pour les données « AUTOS ».

$$\hat{r}_q(x_j, x_{j'}) = \sum_{k=1}^{q} r_{x_j}(F_k) \times r_{x_{j'}}(F_k)$$

Entre parenthèses la différence entre les corrélations ➔

Original, reproduced and residual correlations

	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	-	0.7966 0.8366 (-0.0400)	0.7015 0.7480 (-0.0465)	0.6298 0.6794 (-0.0497)	0.7890 0.7830 (0.0060)	0.6649 0.7402 (-0.0753)
PUISS	0.7966 0.8366 (-0.0400)	-	0.6414 0.6393 (0.0020)	0.5208 0.5627 (-0.0419)	0.7653 0.7164 (0.0489)	0.8444 0.8899 (-0.0456)
LONG	0.7015 0.7480 (-0.0465)	0.6414 0.6393 (0.0020)	-	0.8493 0.8782 (-0.0289)	0.8681 0.8877 (-0.0196)	0.4759 0.4503 (0.0257)
LARG	0.6298 0.6794 (-0.0497)	0.5208 0.5627 (-0.0419)	0.8493 0.8782 (-0.0289)	-	0.7169 0.8291 (-0.1122)	0.4729 0.3773 (0.0957)
POIDS	0.7890 0.7830 (0.0060)	0.7653 0.7164 (0.0489)	0.8681 0.8877 (-0.0196)	0.7169 0.8291 (-0.1122)	-	0.4776 0.5544 (-0.0768)
V.MAX	0.6649 0.7402 (-0.0753)	0.8444 0.8899 (-0.0456)	0.4759 0.4503 (0.0257)	0.4729 0.3773 (0.0957)	0.4776 0.5544 (-0.0768)	-

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
CYL	0.89346	80 % (80 %)	-0.11491	1 % (81 %)
PUISS	0.88686	79 % (79 %)	-0.38469	15 % (93 %)
LONG	0.88615	79 % (79 %)	0.38103	15 % (93 %)
LARG	0.81354	66 % (66 %)	0.41274	17 % (83 %)
POIDS	0.90519	82 % (82 %)	0.22453	5 % (87 %)
V.MAX	0.75471	57 % (57 %)	-0.57352	33 % (90 %)
Var. Expl.	4.42086	74 % (74 %)	0.85606	14 % (88 %)

L'approximation sera d'autant meilleure que les variables sont bien représentées dans le repère sélectionné.

COS² des variables cumulé pour les 2 premières composantes

Caractérisation des composantes par les individus

Analyse des proximités entre individus via leurs coordonnées factorielles

N.B. $I_p = \frac{1}{n}\sum_{i=1}^{n} d^2(i,G) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} z_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} d_i^2$

d_i^2 indique la part de l'individu dans l'inertie totale (dans l'espace originel – variables c.r.). C'est le carré de la distance à l'origine.

Caractérisation des facteurs à l'aide des individus – Coordonnées, contributions et cos²

(1) Coordonnée factorielle de l'individu F_{ik} (permet de situer le positionnement relatif des observations).

(4) Les cos² s' additionnent. Qualité des représentations sur les q = 2 premiers facteurs.

Lecture : Les véhicules se caractérisent par l'encombrement (axe 1, illustrés par les véhicules {9, 10, 13}) et la performance (axe 2, avec surtout {1, 2, 11}).

Remarque : {6, 16 et 5} sont mal représentés sur les q = 2 premières composantes parce qu'ils ne se distinguent ni par l' encombrement (proche de la moyenne) ni par la performance (se situent dans la moyenne).

		Axe 1			Axe 2			
N°	Modele	Coord.	CTR	Cos²	Coord.	CTR	Cos²	SUM(COS²)
1	Alfasud TI	-2.139	6%	56%	-1.786	21%	39%	94%
2	Audi 100	1.561	3%	37%	1.527	15%	35%	71%
3	Simca 1300	-1.119	2%	58%	0.675	3%	21%	79%
4	Citroen GS Club	-2.574	8%	98%	-0.113	0%	0%	98%
5	Fiat 132	0.428	0%	16%	-0.696	3%	41%	57%
6	Lancia Beta	-0.304	0%	8%	0.196	0%	3%	12%
7	Peugeot 504	0.684	1%	31%	0.933	6%	58%	88%
8	Renault 16 TL	-1.948	5%	67%	0.980	6%	17%	84%
9	Renault 30	4.410	24%	89%	-1.064	7%	5%	94%
10	Toyota Corolla	-3.986	20%	98%	-0.236	0%	0%	98%
11	Alfetta 1.66	0.438	0%	4%	-1.912	24%	82%	86%
12	Princess 1800	1.018	1%	53%	0.842	5%	36%	89%
13	Datsun 200L	2.941	11%	78%	0.559	2%	3%	81%
14	Taunus 2000	1.315	2%	70%	-0.487	2%	10%	80%
15	Rancho	-0.691	1%	24%	0.898	5%	41%	65%
16	Mazda 9295	0.386	0%	22%	-0.356	1%	19%	40%
17	Opel Rekord	2.290	7%	86%	-0.104	0%	0%	86%
18	Lada 1300	-2.709	9%	93%	0.144	0%	0%	93%

(2) Contribution : indique l'influence de l'individu dans la définition du facteur

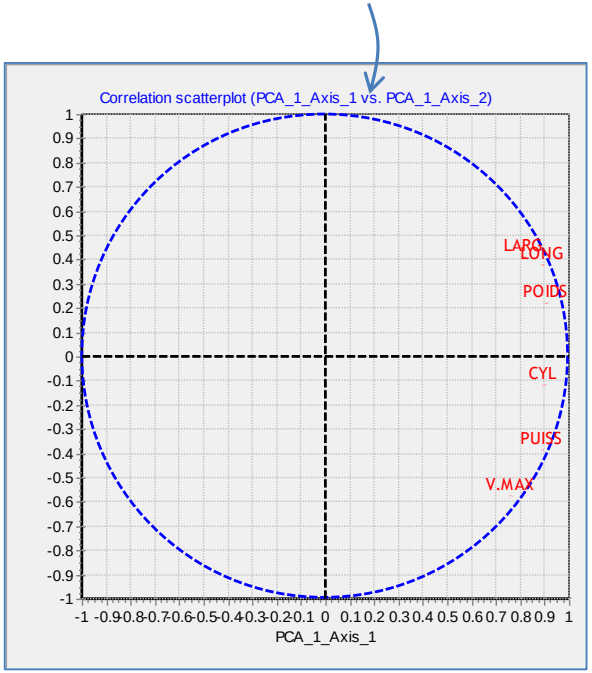
$$CTR_{ik} = \frac{F_{ik}^2}{n \times \lambda_k}; \sum_{i=1}^{n} CTR_{ik} = 1$$

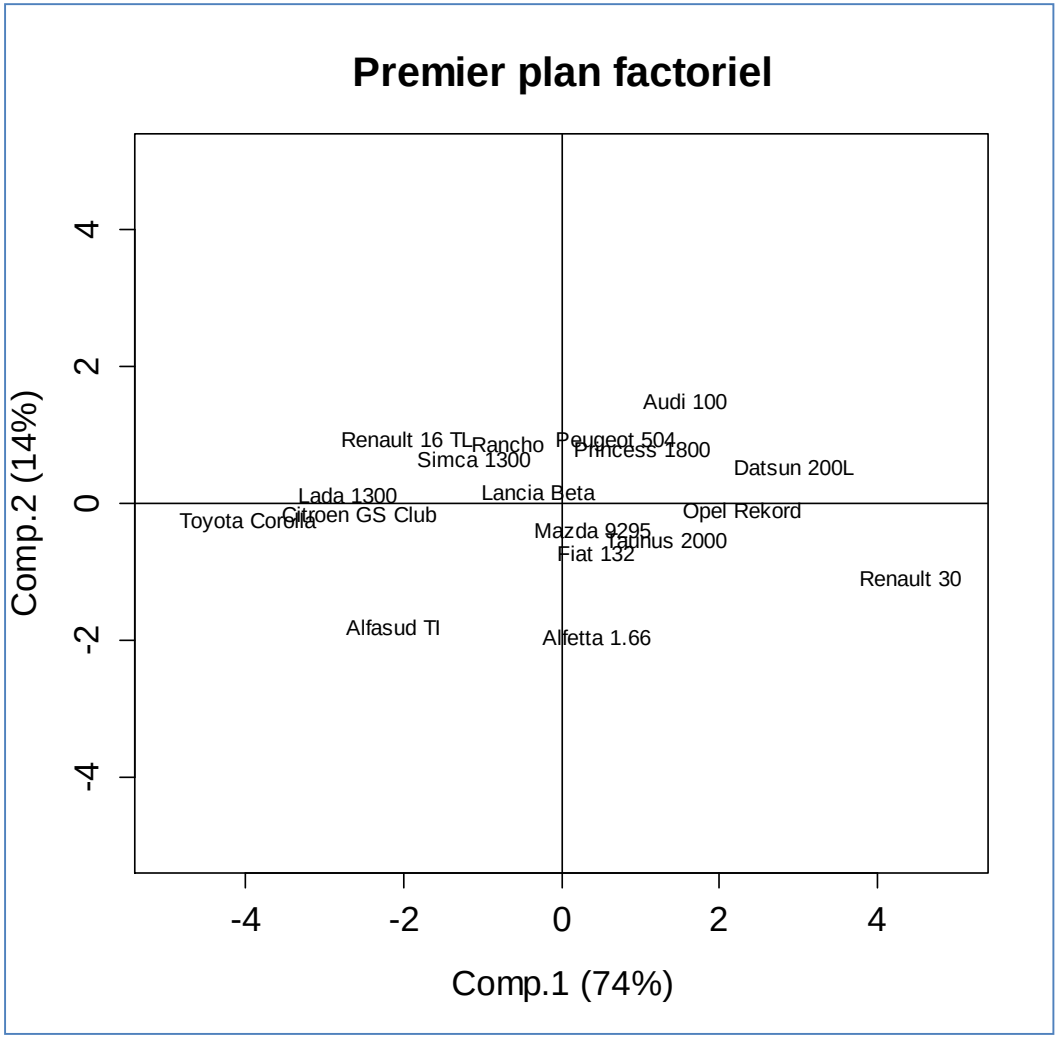
(3) Cos² : indique la qualité de la représentation de l'individu sur le facteur (fraction de son inertie restituée par le facteur)

$$COS_{ik}^2 = \frac{F_{ik}^2}{d_i^2}; \sum_{j=1}^{p} COS_{ik}^2 = 1$$

Ce graphique fait en très grande partie la popularité de l'ACP. On peut y juger visuellement des proximités (dissemblances) entre les individus.

Et on peut comprendre le pourquoi des proximités en considérant dans le même temps le cercle des corrélations.





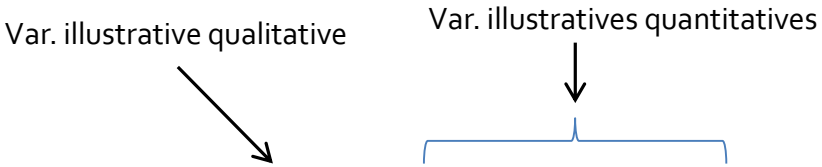
Remarque : certains proposent de mêler les deux représentations dans un graphique dit « biplot ». Attention, les proximités individus-variables n'ont pas vraiment de sens. Ce sont les directions qui importent dans ce cas.

Variables illustratives

Renforcer l'interprétation des composantes

Variables non utilisées pour la construction des composantes. Mais utilisées après coup pour mieux comprendre / commenter les résultats.

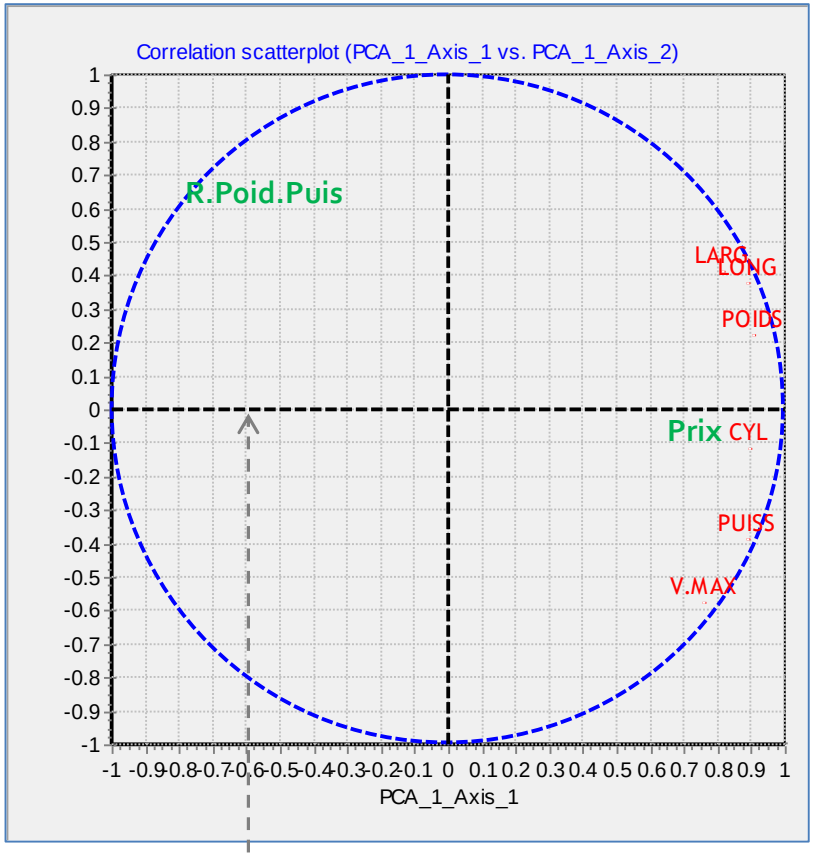
Ex. Les caractéristiques intrinsèques des véhicules sont les variables actives (largeur, poids, puissance, etc.). En illustratives, on utilise des variables introduisant des considérations subjectives (prix, gamme) ou calculées après coup pour une meilleure interprétation (rapport poids/puissance).



Modele	FINITION	PRIX	R.POID.PUIS
Alfasud TI	2_B	30570	11.01
Audi 100	3_TB	39990	13.06
Simca 1300	1_M	29600	15.44
Citroen GS Club	1_M	28250	15.76
Fiat 132	2_B	34900	11.28
Lancia Beta	3_TB	35480	13.17
Peugeot 504	2_B	32300	14.68
Renault 16 TL	2_B	32000	18.36
Renault 30	3_TB	47700	10.31
Toyota Corolla	1_M	26540	14.82
Alfetta-1.66	3_TB	42395	9.72
Princess-1800	2_B	33990	14.15
Datsun-200L	3_TB	43980	11.91
Taunus-2000	2_B	35010	11.02
Rancho	3_TB	39450	14.11
Mazda-9295	1_M	27900	13.19
Opel-Rekord	2_B	32700	11.20
Lada-1300	1_M	22100	14.04

$$r_y(F_k) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(F_{ik} - \bar{F}_k)}{s_y \times s_{F_k}} = \frac{\frac{1}{n}\sum_{i=1}^{n} F_{ik}(y_i - \bar{y})}{s_y \times \sqrt{\lambda_k}}$$

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	Comp.1	Comp.2
PRIX	0.772	-0.087
R.POID.PUIS	-0.589	0.673

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

$$|u_y| \geq 2$$

SIGNIF.	Comp.1	Comp.2
PRIX	**3.975**	-0.337
R.POID.PUIS	**-2.619**	**3.158**

Le rapport poids/puissance n'est pas lié positivement avec le poids parce que les voitures lourdes sont comparativement plus puissantes.

$$\mu_{gk} = \frac{1}{n_g} \sum_{i:y_i=g} F_{ik}$$

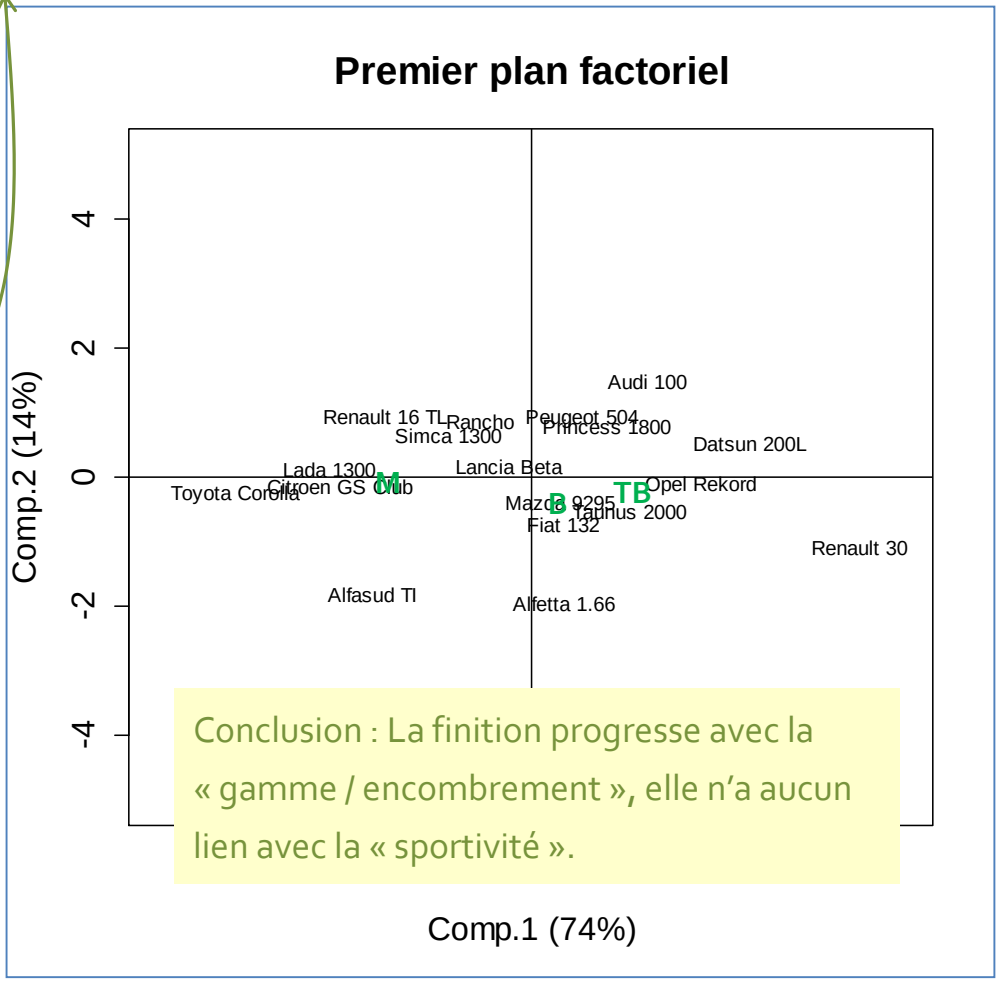
		Comp.1		Comp.2	
FINITION	n_g	Moyenne	Valeur.Test	Moyenne	Valeur.Test
1_M	5	-2.0004	-2.43	0.0226	0.06
2_B	7	0.2353	0.37	-0.0453	-0.16
3_TB	6	1.3924	1.93	0.0340	0.11

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).

$$VT_{gk} = \frac{\mu_{gk} - \bar{F}_k}{\sqrt{\frac{n-n_g}{n-1} \times \frac{s^2_{F_k}}{n_g}}} = \frac{\mu_{gk} - 0}{\sqrt{\frac{n-n_g}{n-1} \times \frac{\lambda_k}{n_g}}}$$

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



Conclusion : La finition progresse avec la « gamme / encombrement », elle n'a aucun lien avec la « sportivité ».

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

Individus illustratifs (supplémentaires)

Positionner de nouveaux individus

Plusieurs raisons possibles :

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

Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
Peugeot 604	2664	136	472	177	1410	180
Peugeot 304 S	1288	74	414	157	915	160

Plutôt cas n°2 ici, on souhaite situer 2 Peugeot supplémentaires (même s'il y a déjà la Peugeot 504 parmi les individus actifs).

Calculs pour les individus illustratifs

Description des véhicules

Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
Peugeot 604	2664	136	472	177	1410	180
Peugeot 304 S	1288	74	414	157	915	160

Moyenne	1631.667	84.611	433.500	166.667	1078.833	158.278
Ecart-type	363.394	19.802	21.484	5.164	133.099	11.798

Moyennes et écarts-type calculés sur l'échantillon d'apprentissage (individus actifs, n = 18).

Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
Peugeot 604	2.8408	2.5951	1.7920	2.0010	2.4881	1.8411
Peugeot 304 S	-0.9457	-0.5359	-0.9076	-1.8719	-1.2309	0.1460

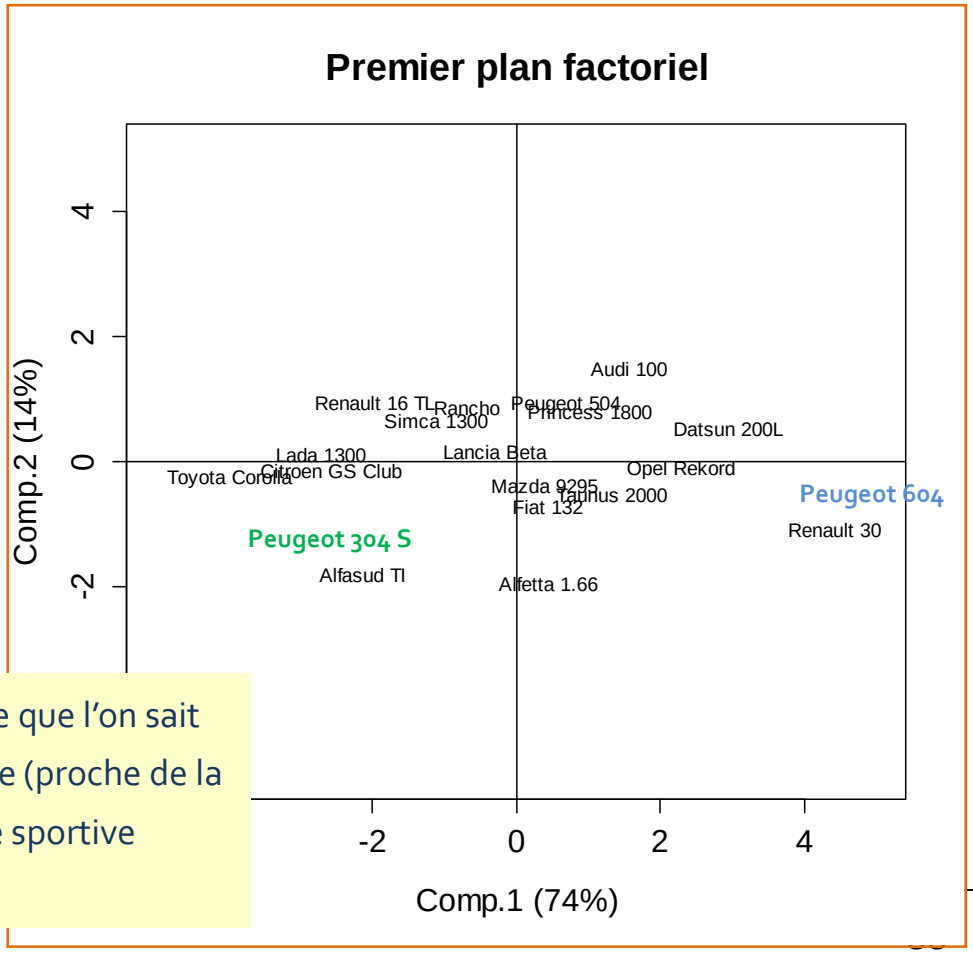
Description après centrage-réduction

Attribute	Comp.1	Comp.2
CYL	0.424936	-0.1241911
PUISS	0.4217944	-0.4157739
LONG	0.4214599	0.4118177
LARG	0.3869222	0.446087
POIDS	0.430512	0.2426758
V.MAX	0.3589443	-0.6198626

Coefficients des fonctions de projection = vecteurs propres issus de l'ACP

Modele	Comp.1	Comp.2
Peugeot 604	5.5633	-0.3386
Peugeot 304 S	-2.2122	-1.2578

Coordonnées factorielles des individus illustratifs : produit scalaire entre description (c.r.) et vecteurs propres.



Les positionnements confirment ce que l'on sait de ces véhicules : « 604 », statutaire (proche de la Renault 30); « 304 S », plutôt petite sportive (proche de l'Alfasud)

Rotation des axes factoriels

Objectif : obtenir des composantes plus facilement interprétables

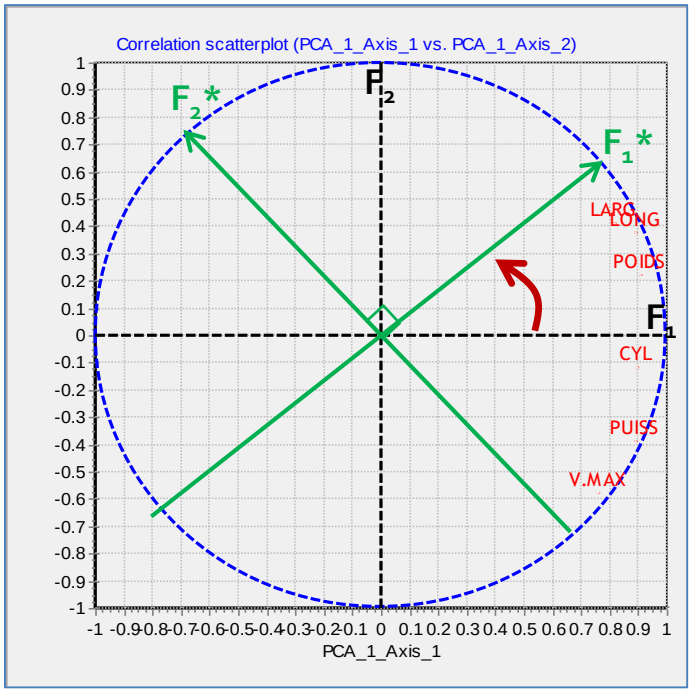
La composante est issue d'une maximisation globale $\lambda_1 = \sum_{j=1}^{p} r_j^2(F_1)$

Mais, il se peut très bien que chaque variable présente individuellement une corrélation « moyenne » avec la composante. L'interprétation est difficile.

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
LONG	0.88615	79 % (79 %)	0.38103	15 % (93 %)
LARG	0.81354	66 % (66 %)	0.41274	17 % (83 %)
POIDS	0.90519	82 % (82 %)	0.22453	5 % (87 %)
CYL	0.89346	80 % (80 %)	-0.11491	1 % (81 %)
V.MAX	0.75471	57 % (57 %)	-0.57352	33 % (90 %)
PUISS	0.88686	79 % (79 %)	-0.38469	15 % (93 %)
Var. Expl.	4.42086	74 % (74 %)	0.85606	14 % (88 %)

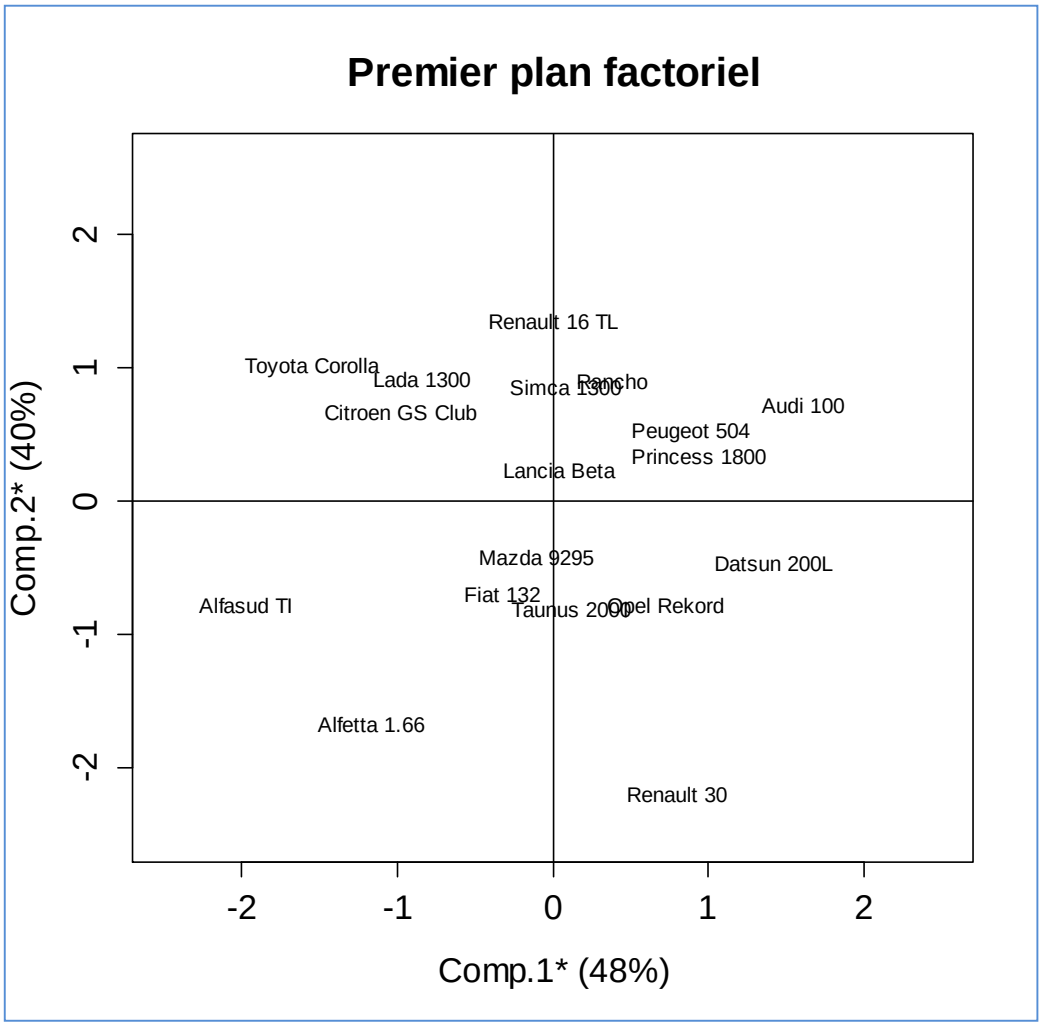
L'idée est de faire pivoter les facteurs (à nombre de facteurs fixés) de manière à rendre plus tranchées (clarifier) les liaisons ou non-liaisons des variables avec l'un des facteurs (on peut préserver ou non l'orthogonalité). L'interprétation des facteurs est facilité.

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
LONG	0.91748	84 % (84 %)	-0.2978	9 % (93 %)
LARG	0.88379	78 % (78 %)	-0.22608	5 % (83 %)
POIDS	0.8286	69 % (69 %)	-0.42801	18 % (87 %)
CYL	0.59598	36 % (36 %)	-0.67549	46 % (81 %)
V.MAX	0.18928	4 % (4 %)	-0.92881	86 % (90 %)
PUISS	0.41314	17 % (17 %)	-0.87397	76 % (93 %)
Var. Expl.	2.87114	48 % (48 %)	2.40578	40 % (88 %)



Méthode : Rotation VARIMAX (orthogonale)
Principe : maximiser la variance des carrés des corrélations intra-facteurs (c.-à-d. les rendre les plus différents possibles les uns des autres).
Cf. https://onlinecourses.science.psu.edu/stat505/node/86

L'explication globale est préservée (88% de l'inertie), mais la répartition entre les composantes a été modifiée (**48% - 40%** vs. 74% - 14%)



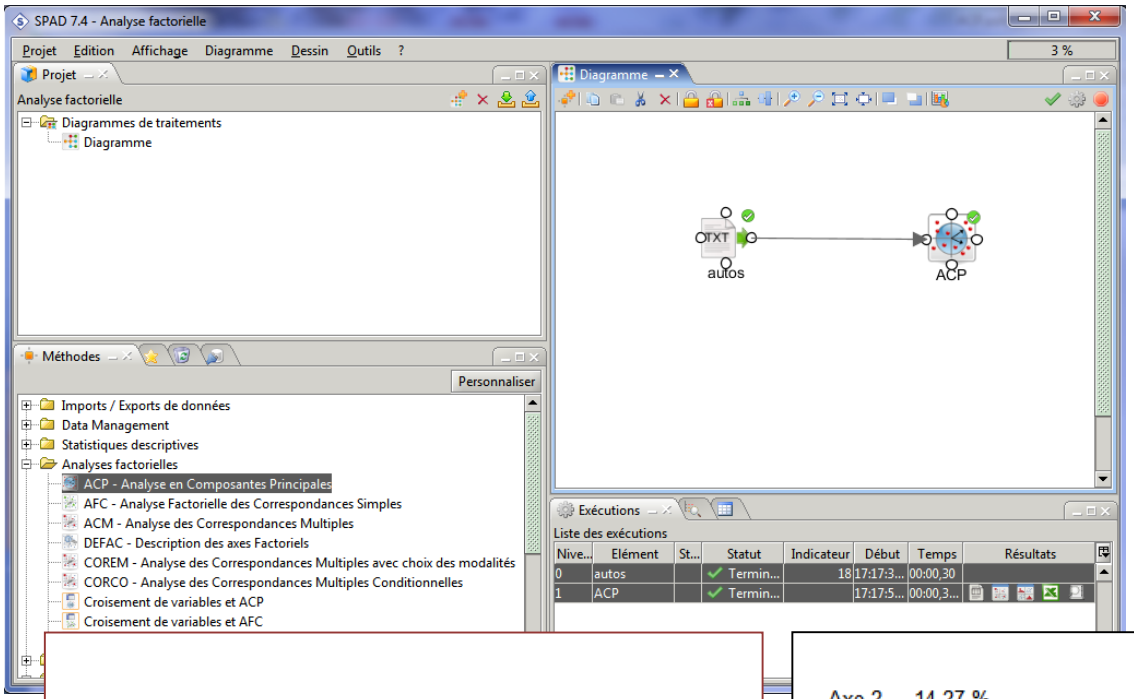
On retrouve les 2 dimensions : « encombrement / gamme » vs. « performances ».

Finalement, plus que par sa taille, la Renault 30 se distingue avant tout par ses performances (cylindrée, puissance et surtout v.max).

Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
Alfasud TI	1350	79	393	161	870	165
Toyota Corolla	1166	55	399	157	815	140
Lada 1300	1294	68	404	161	955	140
Citroen GS Club	1222	59	412	161	930	151
Simca 1300	1294	68	424	168	1050	152
Renault 16 TL	1565	55	424	163	1010	140
Alfetta 1.66	1570	109	428	162	1060	175
Lancia Beta	1297	82	429	169	1080	160
Rancho	1442	80	431	166	1129	144
Taunus 2000	1993	98	438	170	1080	167
Fiat 132	1585	98	439	164	1105	165
Mazda 9295	1769	83	440	165	1095	165
Princess 1800	1798	82	445	172	1160	158
Peugeot 504	1796	79	449	169	1160	154
Renault 30	**2664**	**128**	**452**	**173**	**1320**	**180**
Opel Rekord	1979	100	459	173	1120	173
Audi 100	1588	85	468	177	1110	160
Datsun 200L	1998	115	469	169	1370	160

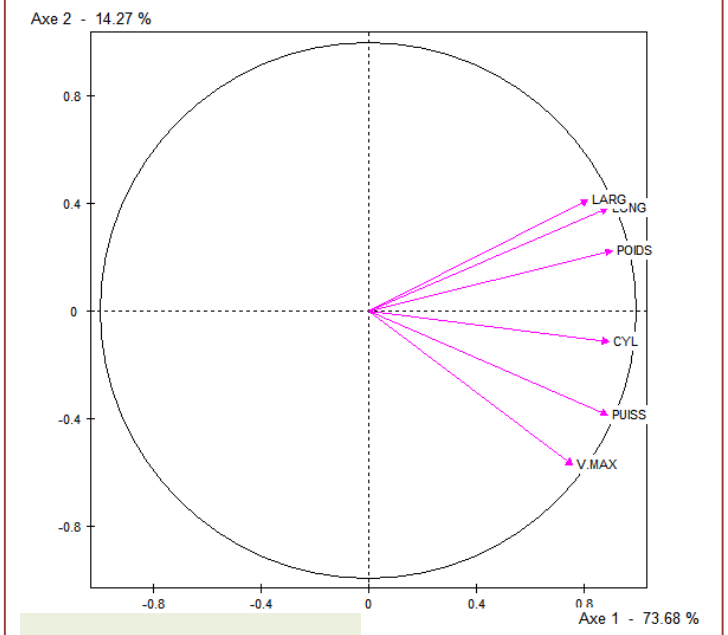
Les logiciels

Préambule : 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 variables (corrélations) soient préservées.

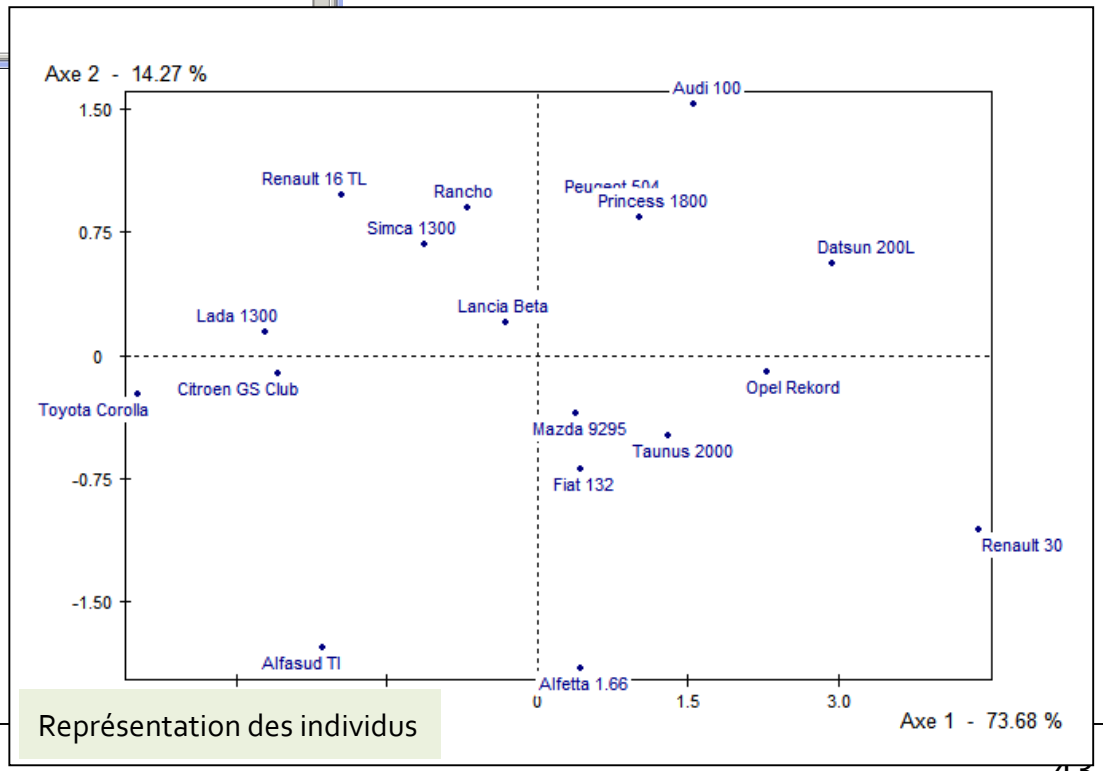


SPAD

La référence de l'analyse de données « à la française ».



Cercle des corrélations



Représentation des individus

La seconde est préférable car fournit des sorties plus détaillées et réalise les rotations

```
proc factor data = mesdata.autos
corr
method = principal
n=2
rotate=varimax
plots=all;
var cyl puiss long larg poids v_max;
run;
```

(1) Résultats avant rotation

Représentation du facteur		
	Factor1	Factor2
CYL	0.89346	0.11491
PUISS	0.88686	0.38469
LONG	0.88615	-0.38103
LARG	0.81354	-0.41274
POIDS	0.90519	-0.22453
V_MAX	0.75471	0.57352

Variance expliquée par chaque facteur	
Factor1	Factor2
4.4208581	0.8560623

Cos² cumulé sur les 2 premières composantes.

Valeurs estimées finales des facteurs communs : Total = 5.276920					
CYL	PUISS	LONG	LARG	POIDS	V_MAX
0.81148051	0.93450289	0.93045316	0.83219237	0.86977918	0.89851225

(2) Résultats après rotation

Matrice de rotation des facteurs (changement de base), θ = -41.25° »

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Procédure FACTOR
Méthode de rotation : Varimax

Matrice de transformation orthogonale		
	1	2
1	0.75185	0.65934
2	-0.65934	0.75185

Caractéristique du facteur de rotation		
	Factor1	Factor2
CYL	0.59598	0.67549
PUISS	0.41314	0.87397
LONG	0.91748	0.29780
LARG	0.88379	0.22608
POIDS	0.82860	0.42801
V_MAX	0.18928	0.92881

Variance expliquée par chaque facteur	
Factor1	Factor2
2.8711381	2.4057822

Valeurs estimées finales des facteurs communs : Total = 5.276920					
CYL	PUISS	LONG	LARG	POIDS	V_MAX
0.81148051	0.93450289	0.93045316	0.83219237	0.86977918	0.89851225

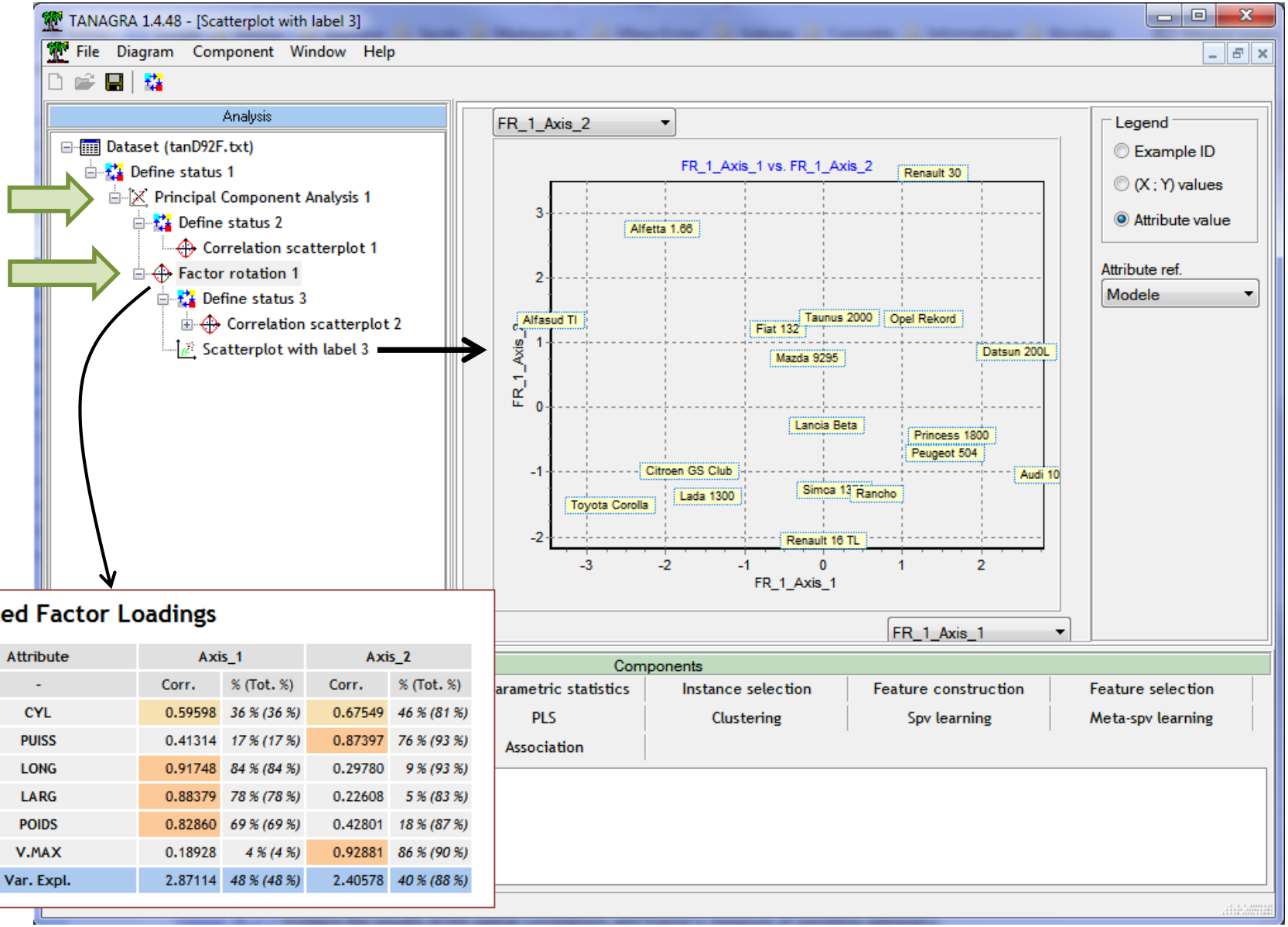
R

Avec pléthore de packages : ade4, ca, FactoMineR, psych, etc.

```
#exemple avec le package psych qui propose la rotation VARIMAX
library(psych)
library(GPArotation)
autos.varimax <- principal(autos,nfactors=2,rotate="varimax")
print(autos.varimax,digits=4)
```

```
> print(autos.varimax,digits=4)
Principal Components Analysis
Call: principal(r = autos, nfactors = 2, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
         RC1    RC2     h2      u2
CYL    0.5977 0.6739 0.8115 0.18852
PUISS 0.4154 0.8729 0.9345 0.06550
LONG   0.9182 0.2954 0.9305 0.06955
LARG   0.8844 0.2238 0.8322 0.16781
POIDS 0.8297 0.4259 0.8698 0.13022
V.MAX 0.1917 0.9283 0.8985 0.10149

                       RC1    RC2
SS loadings           2.8802 2.3967
Proportion Var        0.4800 0.3994
Cumulative Var        0.4800 0.8795
Proportion Explained  0.5458 0.4542
Cumulative Proportion 0.5458 1.0000

Test of the hypothesis that 2 components are sufficient.

The degrees of freedom for the null model are  15  and the objective function was  6.7143
The degrees of freedom for the model are 4  and the objective function was  1.2005
The total number of observations was  18  with MLE Chi Square =  15.4066  with prob <  0.003928

Fit based upon off diagonal values = 0.9934
```

Les tests basés sur le rapport de vraisemblance sont plus adaptés aux techniques de « factor analysis » (cf. la doc de psych)



Rotated Factor Loadings

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
CYL	0.59598	36 % (36 %)	0.67549	46 % (81 %)
PUISS	0.41314	17 % (17 %)	0.87397	76 % (93 %)
LONG	0.91748	84 % (84 %)	0.29780	9 % (93 %)
LARG	0.88379	78 % (78 %)	0.22608	5 % (83 %)
POIDS	0.82860	69 % (69 %)	0.42801	18 % (87 %)
V.MAX	0.18928	4 % (4 %)	0.92881	86 % (90 %)
Var. Expl.	2.87114	48 % (48 %)	2.40578	40 % (88 %)

Plus loin avec l'ACP (1) :

Techniques de ré-échantillonnage pour la détection du nombre de facteurs

Attention, ce sont des procédures purement mécaniques. Les résultats doivent être validés par l'interprétation des facteurs.

Ex. Toutes pousseront (à tort, on le sait maintenant) à négliger le 2nd facteur pour les données AUTOS

Analyse parallèle

Déterminer la distribution des λ_k sous Ho (absence de lien entre les variables)

Démarche :

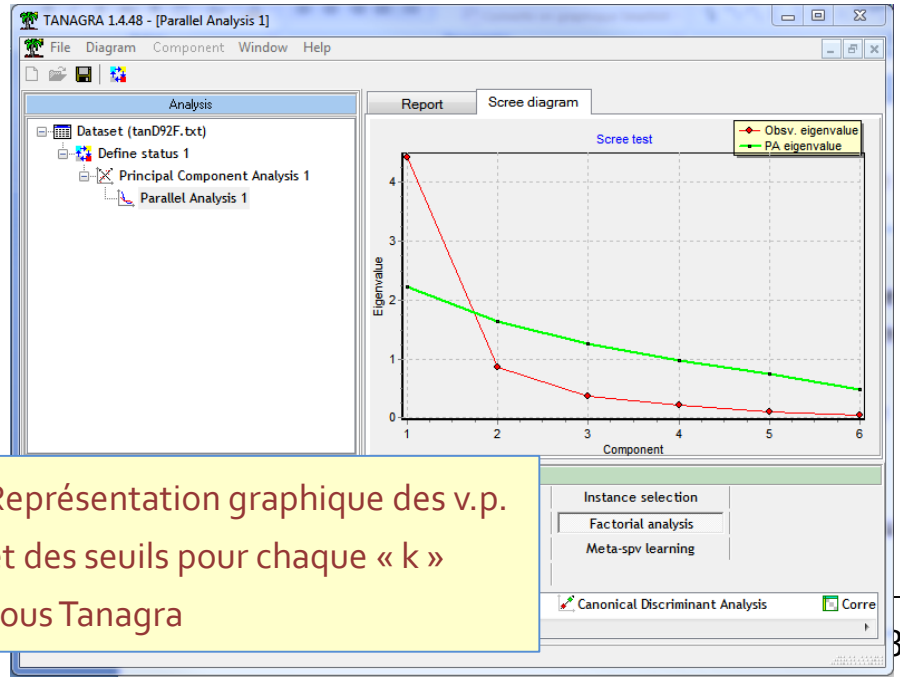
1. Mélanger aléatoirement les valeurs à l'intérieur des colonnes, en traitant les colonnes de manière indépendante → le lien entre les variables est complètement cassé (on est sous Ho)
2. Réaliser l'ACP sur cette nouvelle version des données, collecter les v.p.
3. Répéter T fois les opérations (1) et (2)
4. On obtient pour chaque k une collection de v.p., on en déduit la moyenne μ_k qui sert de seuil critique
5. On décide que la composante k est pertinente si $\lambda_k > \mu_k$

Variante : On peut aussi générer aléatoirement des colonnes de valeurs suivant une gaussienne avec même moyenne et écart-type.

Variante : Plutôt que la moyenne, on peut aussi prendre le quantile d'ordre 0.95 pour un test unilatéral à 5%

Component	Eigenvalue	(0.95) Critical value
1	4.4209	2.2255
2	0.8561	1.6438
3	0.3731	1.2513
4	0.2139	0.9783
5	0.0928	0.7357
6	0.0433	0.4874

Données « AUTOS », seuil critique : quantile d'ordre 0.95 des v.p. sous Ho



Représentation graphique des v.p. et des seuils pour chaque « k » sous Tanagra

Analyse bootstrap

Evaluer la significativité des v.p. successifs c.-à-d. « $\lambda_k > 1$ » significativement ?

Démarche :

1. Effectuer un tirage aléatoire **avec remise** de n observations parmi n (certains individus se répètent ainsi)
2. Réaliser l'ACP sur cette nouvelle version des données, collecter les v.p.
3. Répéter T fois les opérations (1) et (2)
4. On obtient pour chaque k une collection de v.p., on en déduit la distribution empirique. Pour chaque « k », on calcule le quantile d'ordre 0.05 ($\lambda_k^{0.05}$)
5. On déduit que la composante k est pertinente si $\lambda_k^{0.05} > 1$

L'idée est de procéder à un test de significativité à 5% (la v.p. est-elle significativement plus grande que 1 ?)

Component	(0.05) Lower bound
1	3.692685
2	0.513354
3	0.229661
4	0.096113
5	0.037946
6	0.006305

Données « AUTOS » : seul le premier facteur est significatif, le quantile d'ordre 0.05 des v.p. issus du bootstrap est > 1

Remarque : ceci étant, il faut être prudent. Si on s'intéresse à l'intervalle de confiance bootstrap à 90%, on se rend compte que celui du 2nd facteur couvre la valeur « 1 ».

Component	(0.05) Lower bound	(0.95) Upper bound
1	3.692685	4.975922
2	0.513354	1.42439
3	0.229661	0.577211
4	0.096113	0.295001
5	0.037946	0.122611
6	0.006305	0.052366

Analyse bootstrap

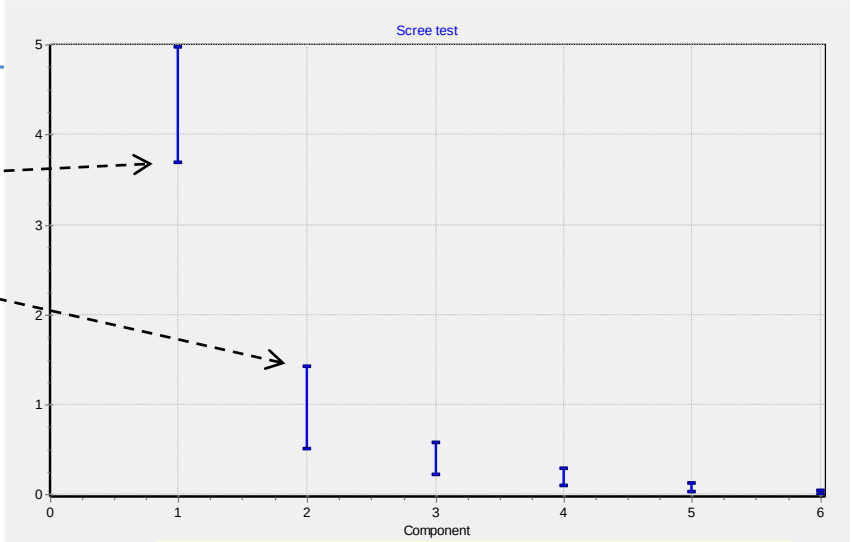
Evaluer le recouvrement entre les λ_k successifs. La composante « k » est pertinente si $\lambda_k > \lambda_{k+1}$ significativement.

Démarche :

1. Effectuer un tirage aléatoire **avec remise** de n observations parmi n (certains individus se répètent ainsi)
2. Réaliser l'ACP sur cette nouvelle version des données, collecter les v.p.
3. Répéter T fois les opérations (1) et (2)
4. On obtient pour chaque k une collection de v.p., on en déduit la distribution empirique. Pour chaque « k », on calcule l'intervalle de confiance à 90% avec les bornes = les quantiles d'ordre 0.05 et 0.95 $[\lambda_k^{0.05} ; \lambda_k^{0.95}]$
5. On déduit que la composante k est pertinente si $\lambda_k^{0.05} > \lambda_{k+1}^{0.95}$ c.-à-d. la borne basse de « k » est-elle plus grande que la borne haute de « k+1 » (y a-t-il un décalage significatif ?).

On retrouve l'idée du coude : est-ce que la composante « k » amène de l'information additionnelle significative par rapport aux suivantes ?

Component	(0.05) Lower bound	(0.95) Upper bound
1	3.692685	4.975922
2	0.513354	1.42439
3	0.229661	0.577211
4	0.096113	0.295001
5	0.037946	0.122611
6	0.006305	0.052366



Données « AUTOS » : dans le tableau des intervalles de confiance, on effectue les comparaisons en décalé avec les quantiles successifs.

Graphiquement, on voit mieux (Tanagra)

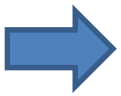
Plus loin avec l'ACP (2) :
Test de sphéricité de Bartlett
Indice MSA (KMO)

Tester l'intérêt de l'ACP en vérifiant s'il est possible de compresser efficacement l'information disponible

Mesurer le degré de redondance des données

Test de sphéricité de Bartlett

Basée sur l'analyse de la matrice des corrélations R

Ho : les variables sont deux à deux indépendantes  R = matrice unité

2 situations extrêmes (rappel : |R| = produit des valeurs propres de R)

|R| = 1, les variables sont deux à deux orthogonales, ACP inutile, impossible de résumer l'information

|R| = 0, il y a une colinéarité parfaite (le 1er facteur explique 100% de l'inertie totale)

Statistique de test : $$B = -\left(n - 1 - \frac{2p+5}{6}\right)\ln|R| \equiv \chi^2\left(\frac{p(p-1)}{2}\right)$$

Attention ! Quand « n » est grand, rejet quasi systématique de Ho car les ddl ne tiennent pas compte de « n »

R	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.797	0.701	0.630	0.789	0.665
PUISS	0.797	1	0.641	0.521	0.765	0.844
LONG	0.701	0.641	1	0.849	0.868	0.476
LARG	0.630	0.521	0.849	1	0.717	0.473
POIDS	0.789	0.765	0.868	0.717	1	0.478
V.MAX	0.665	0.844	0.476	0.473	0.478	1

Données « AUTOS »

Déterminant	0.001213

<<< les variables sont fortement redondantes

B	95.11988
ddl	15
p-value	<0.00001

Conclusion : rejet de Ho, les variables ne sont pas indép. 2 à 2. Il est possible de compresser l'information avec l'ACP. Efficacement même si l'on en juge la valeur de |R|.

Effectivement, on a vu que F_1 représentait 74% de l'info dispo.

Remarque : une variante de ce test peut être utilisée pour détecter le nombre de composantes « significatives », mais elle s'avère trop permissive en pratique.

MSA : Measure of sampling adequacy [ou indice KMO (Kaiser – Mayer – Olkin)]

Idée du MSA : confronter la matrice des corrélations brutes avec la matrice des corrélations partielles. Si MSA ≈ 1, l'ACP peut agir efficacement parce que corrélations partielles sont quasi-nulles (en valeur absolue) ; si MSA << 1, problème car pas de redondance entre les variables.

Corrélations brutes entre les variables prises 2 à 2 (r_{jm})

R	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.797	0.701	0.630	0.789	0.665
PUISS	0.797	1	0.641	0.521	0.765	0.844
LONG	0.701	0.641	1	0.849	0.868	0.476
LARG	0.630	0.521	0.849	1	0.717	0.473
POIDS	0.789	0.765	0.868	0.717	1	0.478
V.MAX	0.665	0.844	0.476	0.473	0.478	1

Corrélations partielles c.-à-d. après avoir retranché l'influence des (p-2) autres (r^*_{jm})

PARTIAL R	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.107	-0.060	0.109	0.320	0.189
PUISS	0.107	1	-0.083	-0.334	0.652	0.841
LONG	-0.060	-0.083	1	0.582	0.530	0.090
LARG	0.109	-0.334	0.582	1	0.127	0.331
POIDS	0.320	0.652	0.530	0.127	1	-0.611
V.MAX	0.189	0.841	0.090	0.331	-0.611	1

Les corrélations brutes et partielles sont majoritairement différentes, les relations entre 2 variables quelconques sont fortement déterminées par les autres.

$$MSA = \frac{\sum_j \sum_{m \neq j} (r_{jm})^2}{\sum_j \sum_{m \neq j} (r_{jm})^2 + \sum_j \sum_{m \neq j} (r^*_{jm})^2}$$

Données « AUTOS »

MSA = 0.740

MSA est plus un indice de compressibilité de l'information qu'un indicateur de l'intérêt d'une ACP pour un fichier de données !

Interpretation of the KMO as characterised by Kaiser. Meyer & Olkin (1974)

KMO Value	Degree of Common Variance
0.90 to 1.00	Marvelous
0.80 to 0.89	Meritorious
0.70 to 0.79	Middling
0.60 to 0.69	Mediocre
0.50 to 0.59	Miserable
0.00 to 0.49	Don't Factor

http://peoplelearn.homestead.com/Topic20-FACTORanalysis3a.html

MSA par variable : détecter les variables qui se détachent des autres, ceux dont la corrélation partielle reste proche de la corrélation brute c.-à-d. qui participent peu à la redondance globale → plus l'indice est faible, plus la variable est faiblement liée globalement aux autres.

$$MSA_j = \frac{\sum_{m \neq j} (r_{jm})^2}{\sum_{m \neq j} (r_{jm})^2 + \sum_{m \neq j} (r^*_{jm})^2}$$

r

R	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.797	0.701	0.630	0.789	0.665

r*

PARTIAL R	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.107	-0.060	0.109	0.320	0.189

$$MSA_{CYL} = \frac{0.797^2 + 0.701^2 + \cdots + 0.665^2}{(0.797^2 + 0.701^2 + \cdots + 0.665^2) + (0.107^2 + (-0.06)^2 + \cdots + 0.189^2)} = 0.940$$

	MSA
CYL	0.940
PUISS	0.674
LONG	0.803
LARG	0.784
POIDS	0.693
V.MAX	0.598

« CYL » est la variable la plus liée à l'ensemble des autres

« V.MAX » est celle qui participe le moins à la tendance collective

Plus loin avec l'ACP (3) :

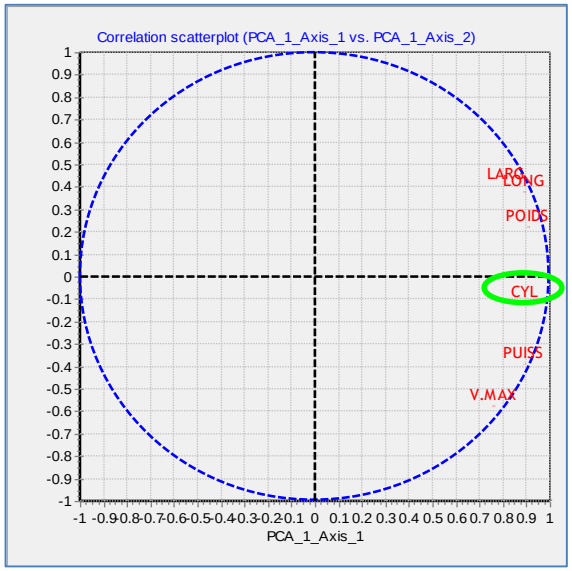
ACP à partir d'une matrice des corrélations partielles

Retrancher l'influence d'une ou plusieurs variables qui pèsent sur toutes les autres dans l'analyse

Une manière de gérer « l'effet taille » qui écrase souvent la 1ère composante

Problème de « l'effet taille » en ACP

« Effet taille » : parfois, une ou plusieurs variables pèsent sur l'ensemble des autres et déterminent fortement les résultats, notamment en pesant exagérément sur la 1ère composante qui semblent concentrer toute l'information disponible.



Ex. « CYL » pèse sur l'ensemble des variables, déterminant le 1er facteur, et entraînant avec lui l'ensemble des autres variables.

Solution : analyser les relations entre les variables après avoir retranché (en contrôlant) l'influence de ou des variables incriminées c.-à-d. au lieu de diagonaliser la matrice des corrélations brutes, baser l'ACP sur les corrélations partielles (la nature de l'information traitée est différente)

Corrélation brutes

CORR	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.797	0.701	0.630	0.789	0.665
PUISS		1	0.641	0.521	0.765	0.844
LONG			1	0.849	0.868	0.476
LARG				1	0.717	0.473
POIDS					1	0.478
V.MAX						1

Corrélations partielles / CYL

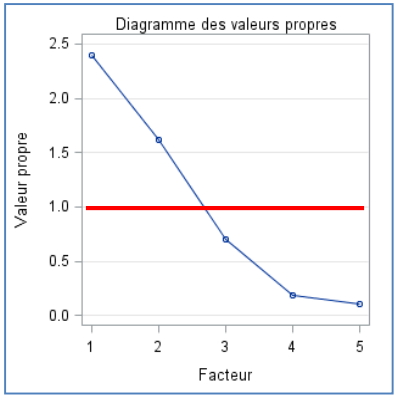
CORR/CYL	PUISS	LONG	LARG	POIDS	V.MAX
PUISS	1	0.192	0.041	0.368	0.697
LONG		1	0.736	0.719	0.018
LARG			1	0.461	0.093
POIDS				1	-0.102
V.MAX					1

Matrice à diagonaliser

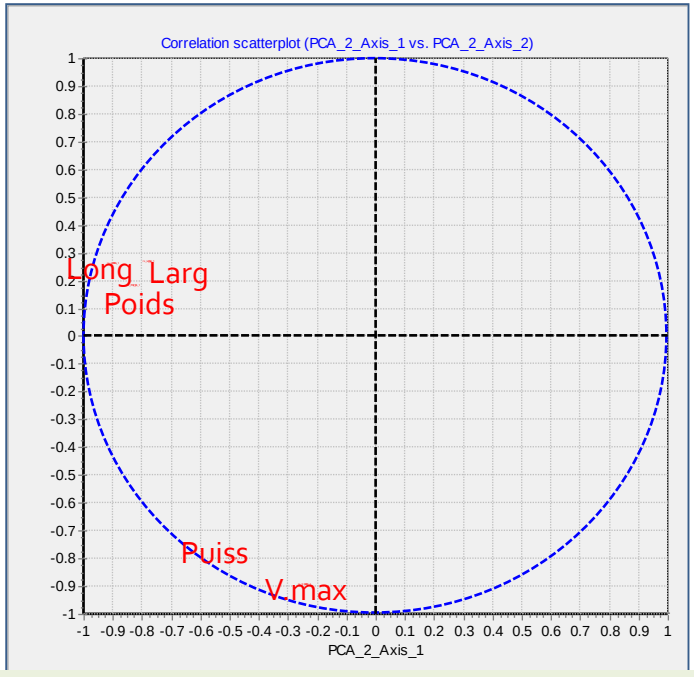
Ex. Voir la différence entre les corrélations « PUISS x POIDS » selon que l'on contrôle ou non l'influence de CYL ; voir aussi la relation « POIDS x V.MAX »

ACP à partir de la matrice des corrélations partielles

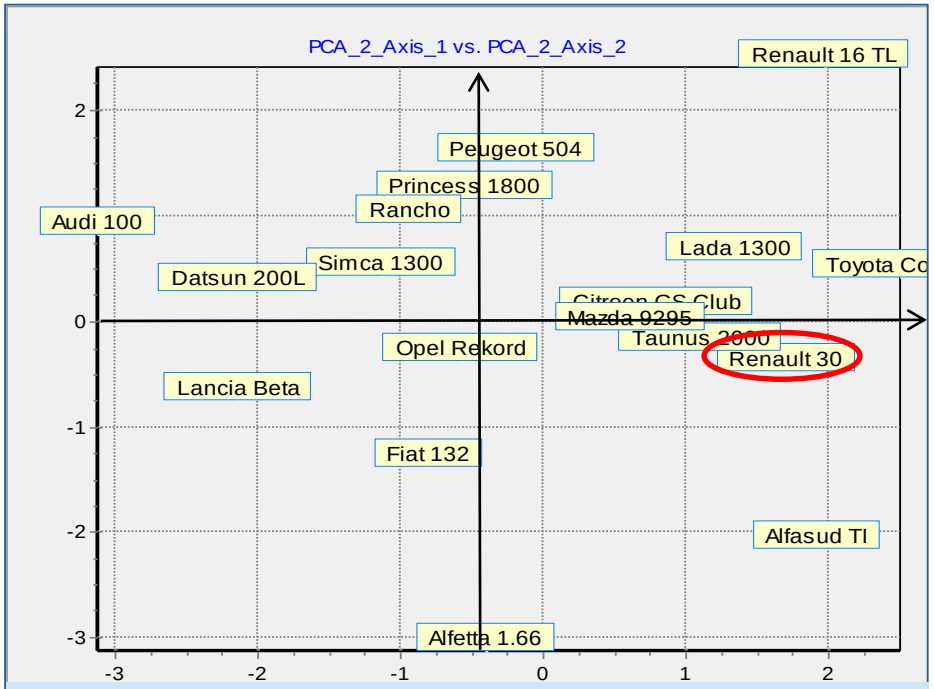
```
proc factor data = mesdata.autos
corr
method = principal
n=2
plots=all;
var puiss long larg poids v_max;
partial cyl;
run;
```



Il y a 2 composantes à analyser, de manière évidente maintenant : avec $\lambda_1 = 2.41$ et $\lambda_2 = 1.61$.



Les 2 dimensions apparaissent nettement. (1) Encombrement / gamme : à cylindrée égale, on distingue les voitures longues/larges/lourdes des autres. (2) Performances : la puissance et la rapidité caractérisent les véhicules. Sans qu'il soit nécessaire de post-traiter cette fois-ci (rotation des axes).



De nouveaux types de résultats apparaissent.

Ex. Eu égard à sa cylindrée, la RENAULT 30 :

1. N'est pas si encombrante (moins que les autres même)
2. N'est pas si performante (dans la moyenne simplement)

Ex. L'Alfasud Ti est une petite teigneuse

Ex. Le moteur de la Renault 16 TL est vraiment sous exploité

Plus loin avec l'ACP (4) :
Analyse en Facteurs Principaux

Plutôt que de s'intéresser à la variabilité totale des variables, analyser la variabilité partagée

Approche a priori préférable lorsque l'on cherche à structurer l'information

On s'intéresse aux techniques non-itératives seulement

Gestion des variables non pertinentes en ACP

L'ACP cherche à reproduire toute la variabilité des données, c'est pour cette raison que la somme des COS^2 des variables est égale à 1 lorsqu'on prend en compte tous les facteurs.

Cette idée n'est pas toujours pertinente quand des variables totalement étrangères à l'étude s'immiscent dans le fichier de données. Elles pèsent de manière indue sur les résultats.

Ex. AUTOS + 6 variables N(0, 1) générées aléatoirement.

Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX	RND1	RND2	RND3	RND4	RND5	RND6
Alfasud TI	1350	79	393	161	870	165	1.029	0.109	0.267	-0.949	0.053	-1.144
Audi 100	1588	85	468	177	1110	160	0.379	0.367	-1.054	0.106	1.316	1.632
Simca 1300	1294	68	424	168	1050	152	-0.346	0.737	-0.056	-0.430	0.274	1.044
Citroen GS Club	1222	59	412	161	930	151	0.849	0.494	-0.320	0.729	0.637	0.793
Fiat 132	1585	98	439	164	1105	165	-1.425	-0.210	0.535	-0.200	-1.983	0.832
Lancia Beta	1297	82	429	169	1080	160	0.890	0.071	-1.308	-0.971	-0.131	-1.134
Peugeot 504	1796	79	449	169	1160	154	-0.860	-0.001	-0.315	-1.329	-0.605	1.299
Renault 16 TL	1565	55	424	163	1010	140	2.379	-1.367	-0.633	-1.448	0.609	-0.175
Renault 30	2664	128	452	173	1320	180	-0.578	0.705	-1.304	2.124	-1.132	-2.091
Toyota Corolla	1166	55	399	157	815	140	0.381	0.330	0.121	-1.285	0.570	-1.628
Alfetta 1.66	1570	109	428	162	1060	175	0.204	0.287	-2.117	-1.675	0.111	2.773
Princess 1800	1798	82	445	172	1160	158	0.216	0.549	-0.619	-0.096	-1.632	-0.066
Datsun 200L	1998	115	469	169	1370	160	0.603	0.914	1.403	-0.371	-1.892	0.681
Taunus 2000	1993	98	438	170	1080	167	-0.326	0.857	-0.565	1.455	0.370	-0.656
Rancho	1442	80	431	166	1129	144	-0.787	0.948	-1.389	-0.377	-0.139	-0.721
Mazda 9295	1769	83	440	165	1095	165	-0.931	1.222	-0.133	-1.090	-1.201	0.187
Opel Rekord	1979	100	459	173	1120	173	0.984	0.831	-0.138	0.545	-1.730	1.459
Lada 1300	1294	68	404	161	955	140	0.986	0.791	-0.100	0.037	0.221	0.764

Ces variables additionnelles vont masquer les relations existantes entre les autres

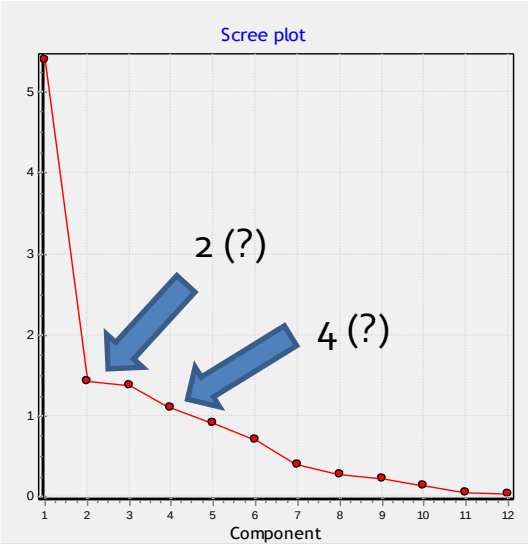
ACP sur AUTOS avec 6 variables « bruitées »

Attribute	Axis_1		Axis_2		Axis_3		Axis_4	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
CYL	-0.8847	78 % (78 %)	-0.0295	0 % (78 %)	-0.1463	2 % (80 %)	-0.2681	7 % (88 %)
PUISS	-0.8912	79 % (79 %)	-0.0157	0 % (79 %)	-0.0217	0 % (79 %)	0.0356	0 % (80 %)
LONG	-0.8452	71 % (71 %)	-0.2290	5 % (77 %)	0.3167	10 % (87 %)	-0.0911	1 % (88 %)
LARG	-0.7702	59 % (59 %)	-0.3696	14 % (73 %)	-0.0191	0 % (73 %)	-0.1024	1 % (74 %)
POIDS	-0.8905	79 % (79 %)	-0.0488	0 % (80 %)	0.1851	3 % (83 %)	-0.1536	2 % (85 %)
V.MAX	-0.7541	57 % (57 %)	-0.1145	1 % (58 %)	-0.0820	1 % (59 %)	0.1318	2 % (61 %)
RND1	0.4695	22 % (22 %)	-0.3550	13 % (35 %)	0.0293	0 % (35 %)	-0.5753	33 % (68 %)
RND2	-0.4413	19 % (19 %)	0.4068	17 % (36 %)	-0.2630	7 % (43 %)	0.5380	29 % (72 %)
RND3	0.0474	0 % (0 %)	0.7548	57 % (57 %)	0.3744	14 % (71 %)	-0.3263	11 % (82 %)
RND4	-0.5592	31 % (31 %)	0.0730	1 % (32 %)	-0.6058	37 % (69 %)	-0.0873	1 % (69 %)
RND5	0.6468	42 % (42 %)	-0.4995	25 % (67 %)	-0.3422	12 % (78 %)	0.1718	3 % (81 %)
RND6	-0.0811	1 % (1 %)	-0.3246	11 % (11 %)	0.7132	51 % (62 %)	0.4404	19 % (81 %)
Var. Expl.	5.40111	45 % (45 %)	1.42714	12 % (57 %)	1.36642	11 % (68 %)	1.09082	9 % (77 %)

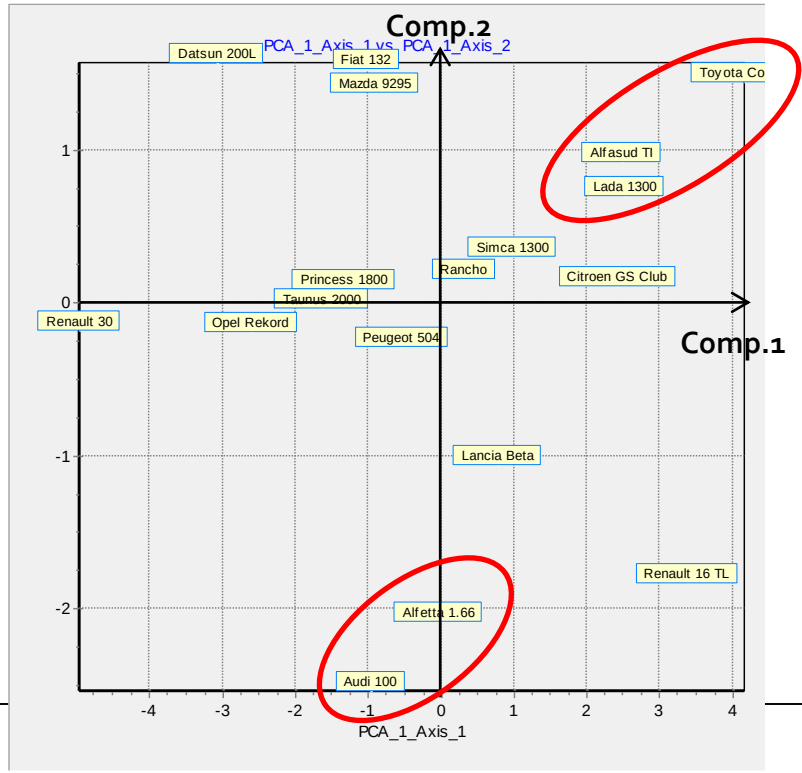
L'effet taille subsiste : « encombrement / gamme »

La « sportivité » est complètement noyée au milieu (masquée par) des variables « rnd »

Nombre de composantes



Par ex., l'Alfasud et l'Afetta n'ont rien à faire avec ces voisins !!!



Remarque : En réalité, il faut trouver la dimension « sportivité » sur le 5ème facteur, que nous n'avons pas retenu avec les critères usuels.

En théorie, l'analyse en facteur principaux correspond à une démarche de modélisation. On cherche à construire des facteurs (F_1, F_2, ..., F_q) [on parle aussi de « variables latentes »] qui permettent de reproduire au mieux les variables originelles.

$$\begin{cases} x_1 = a_{11}F_1 + a_{12}F_2 + \cdots a_{1q}F_q + e_1 \\ \cdots \\ x_p = a_{p1}F_1 + a_{p2}F_2 + \cdots a_{pq}F_q + e_p \end{cases}$$

e_j sont des termes d'erreur puisqu'une modélisation n'est jamais parfaite.

En pratique, il s'agit simplement de diagonaliser une variante de la matrice des corrélations mettant en exergue la variance partagée entres les variables. L'approche fournit des résultats très similaires à ceux de l'ACP. C'est pour cette raison qu'elles sont souvent confondues d'ailleurs.

Quelques définitions et formules

- **R^2_j** est la « communalité » (communality). Il s'agit du coefficient de détermination de la régression de X_j sur les (p-1) autres variables. Ainsi, R^2_j correspond à la part de variance de X_j expliquée par les autres. Cette quantité doit être modélisée.
- u_j est « l'uniqueness », $u_j = 1 - R^2_j$. C'est la proportion de la variance de X_j non expliquée par les autres variables. Elle ne doit pas être modélisée.

Analyse en Facteurs Principaux (AFP)

Analyse en facteurs principaux [AFP] (Outil « Principal Factor Analysis » dans Tanagra)

Matrice des corrélations (usuelle)

RND4 et RND5 sont fortuitement (très malencontreusement) corrélées avec les variables initiales. C'est une difficulté supplémentaire.

	CYL	PUISS	LONG	LARG	POIDS	V_MA	RND1	RND2	RND3	RND4	RND5	RND6
CYL	1	0.80	0.70	0.63	0.79	0.66	-0.28	0.24	-0.06	0.57	-0.55	-0.13
PUISS	0.80	1	0.64	0.52	0.77	0.84	-0.38	0.35	-0.11	0.43	-0.58	0.07
LONG	0.70	0.64	1	0.85	0.87	0.48	-0.29	0.22	-0.01	0.31	-0.48	0.31
LARG	0.63	0.52	0.85	1	0.72	0.47	-0.19	0.25	-0.19	0.48	-0.26	0.11
POIDS	0.79	0.77	0.87	0.72	1	0.48	-0.36	0.27	-0.01	0.35	-0.62	0.09
V_MA	0.66	0.84	0.48	0.47	0.48	1	-0.31	0.27	-0.19	0.38	-0.45	0.11
RND1	-0.28	-0.38	-0.29	-0.19	-0.36	-0.31	1	-0.44	0.02	-0.18	0.37	-0.02
RND2	0.24	0.35	0.22	0.25	0.27	0.27	-0.44	1	0.08	0.43	-0.25	0.01
RND3	-0.06	-0.11	-0.01	-0.19	-0.01	-0.19	0.02	0.08	1	-0.03	-0.38	-0.02
RND4	0.57	0.43	0.31	0.48	0.35	0.38	-0.18	0.43	-0.03	1	-0.14	-0.24
RND5	-0.55	-0.58	-0.48	-0.26	-0.62	-0.45	0.37	-0.25	-0.38	-0.14	1	-0.02
RND6	-0.13	0.07	0.31	0.11	0.09	0.11	-0.02	0.01	-0.02	-0.24	-0.02	1

Matrice H = (h_{jm}) à diagonaliser pour la PFA. On a remplacé « 1 » par les communalités dans la diagonale.

	CYL	PUISS	LONG	LARG	POIDS	V_MA	RND1	RND2	RND3	RND4	RND5	RND6
CYL	**0.84**	0.80	0.70	0.63	0.79	0.66	-0.28	0.24	-0.06	0.57	-0.55	-0.13
PUISS	0.80	**0.92**	0.64	0.52	0.77	0.84	-0.38	0.35	-0.11	0.43	-0.58	0.07
LONG	0.70	0.64	**0.93**	0.85	0.87	0.48	-0.29	0.22	-0.01	0.31	-0.48	0.31
LARG	0.63	0.52	0.85	**0.88**	0.72	0.47	-0.19	0.25	-0.19	0.48	-0.26	0.11
POIDS	0.79	0.77	0.87	0.72	**0.92**	0.48	-0.36	0.27	-0.01	0.35	-0.62	0.09
V_MA	0.66	0.84	0.48	0.47	0.48	**0.88**	-0.31	0.27	-0.19	0.38	-0.45	0.11
RND1	-0.28	-0.38	-0.29	-0.19	-0.36	-0.31	**0.34**	-0.44	0.02	-0.18	0.37	-0.02
RND2	0.24	0.35	0.22	0.25	0.27	0.27	-0.44	**0.39**	0.08	0.43	-0.25	0.01
RND3	-0.06	-0.11	-0.01	-0.19	-0.01	-0.19	0.02	0.08	**0.45**	-0.03	-0.38	-0.02
RND4	0.57	0.43	0.31	0.48	0.35	0.38	-0.18	0.43	-0.03	**0.61**	-0.14	-0.24
RND5	-0.55	-0.58	-0.48	-0.26	-0.62	-0.45	0.37	-0.25	-0.38	-0.14	**0.69**	-0.02
RND6	-0.13	0.07	0.31	0.11	0.09	0.11	-0.02	0.01	-0.02	-0.24	-0.02	**0.51**

R^2_{cyl} = 0.84 → coefficient de détermination de la régression de CYL sur les autres variables (PUISS, LONG, …, RND6). Seule cette fraction doit être intégrée dans la modélisation.

AFP sur les données AUTOS + RND – Tableau des valeurs propres

Moyenne des valeurs propres, peut être utilisée comme seuil pour la sélection des facteurs

Somme des valeurs de la diagonale principale de la matrice à diagonaliser

Eigen values

Matrix trace	8.364373
Average	0.697031

Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	5.229521	4.218050	62.52 %		62.52 %
2	1.011471	0.035886	12.09 %		74.61 %
3	0.975585	0.252347	11.66 %		86.28 %
4	0.723238	0.184633	8.65 %		94.92 %
5	0.538606	0.275195	6.44 %		101.36 %
6	0.263410	0.167558	3.15 %		104.51 %
7	0.095852	0.089953	1.15 %		105.66 %
8	0.005899	0.049727	0.07 %		105.73 %
9	-0.043828	0.047582	-0.52 %		105.21 %
10	-0.091409	0.060599	-1.09 %		104.11 %
11	-0.152009	0.039954	-1.82 %		102.30 %
12	-0.191963	-	-2.30 %		100.00 %
Tot.	8.364373	-	-	-	-

On en sélectionne 4 si on se fie à ce critère

H n'est pas semi-définie positive, il est normal que l'on puisse obtenir des v.p. négatives

On modélise plus que la variance commune, une correction dans l'autre sens est nécessaire

Au final, on modélise bien la variabilité commune

AFP sur les données AUTOS + RND - 4 facteurs sélectionnés

R^2_j à modéliser

R^2_j restituée sur les 4 premiers facteurs = somme du carré des « loadings »

Les variables initiales sont plutôt bien modélisées

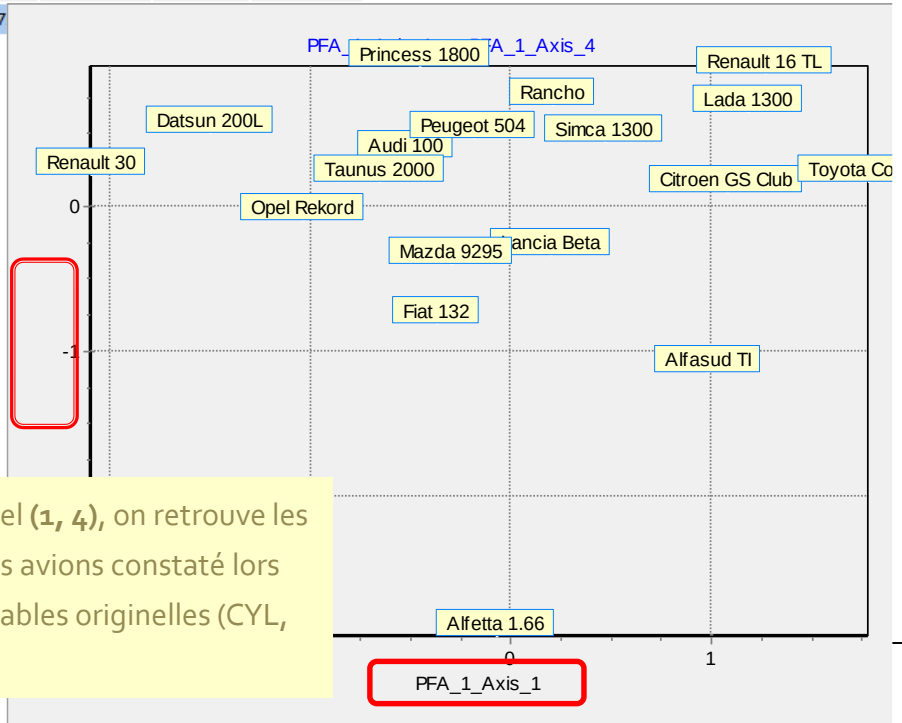
La « sportivité / performance » n'apparaît que sur le 4ème facteur

Factor Loadings [Communality Estimates]

Attribute	Communality Estimates		Axis_1		Axis_2		Axis_3		Axis_4	
-	Prior	Final	Corr.	Sq. (Cumul.)	Corr.	Sq. (Cumul.)	Corr.	Sq. (Cumul.)	Corr.	Sq. (Cumul.)
CYL	0.84305	0.80309	-0.87446	*0.76 (0.76)*	0.08499	*0.01 (0.77)*	-0.15299	*0.02 (0.80)*	0.08817	*0.01 (0.80)*
PUISS	0.92294	0.91244	-0.89312	*0.80 (0.80)*	0.11705	*0.01 (0.81)*	-0.06350	*0.00 (0.82)*	-0.31151	*0.10 (0.91)*
LONG	0.92590	0.95807	-0.85078	*0.72 (0.72)*	-0.36730	*0.13 (0.86)*	0.26576	*0.07 (0.93)*	0.16945	*0.03 (0.96)*
LARG	0.87988	0.86076	-0.76842	*0.59 (0.59)*	-0.42381	*0.18 (0.77)*	-0.12532	*0.02 (0.79)*	0.27381	*0.07 (0.86)*
POIDS	0.92397	0.87964	-0.89437	*0.80 (0.80)*	-0.09491	*0.01 (0.81)*	0.20490	*0.04 (0.85)*	0.16956	*0.03 (0.88)*
V.MAX	0.88229	0.85665	-0.74997	*0.56 (0.56)*	0.05841	*0.00 (0.57)*	-0.18954	*0.04 (0.60)*	-0.50483	*0.25 (0.86)*
RND1	0.33669	0.23869	0.41423	*0.17 (0.17)*	-0.23107	*0.05 (0.22)*	-0.05592	*0.00 (0.23)*	0.10284	*0.01 (0.24)*
RND2	0.39437	0.24426	-0.39190	*0.15 (0.15)*	0.28431	*0.08 (0.23)*	-0.09067	*0.01 (0.24)*	0.04031	*0.00 (0.24)*
RND3	0.45084	0.43613	0.04606	*0.00 (0.00)*	0.41717	*0.17 (0.18)*	0.44354	*0.20 (0.37)*	0.25149	*0.06 (0.44)*
RND4	0.61007	0.60126	-0.52443	*0.28 (0.28)*	0.18178	*0.03 (0.31)*	-0.46991	*0.22 (0.53)*	0.26905	*0.07 (0.60)*
RND5	0.68805	0.73010	0.61822	*0.38 (0.38)*	-0.40149	*0.16 (0.54)*	-0.43195	*0.19 (0.73)*	0.01109	*0.00 (0.73)*
RND6	0.50632	0.41872	-0.07745	*0.01 (0.01)*	-0.40138	*0.16 (0.17)*	0.41041	*0.17 (0.34)*	-0.28841	*0.08 (0.42)*
Var. Expl.	8.36437	7.93981	5.22952	*63 % (63 %)*	1.01147	*12 % (75 %)*	0.97			

Toujours « encombrement / gamme », mais RND4 et RND5 pèsent toujours beaucoup trop.

Attention : « loadings ≠ corrélation». Elles correspondent aux coefficients standardisés de la régression de chaque variable avec les facteurs (dans les faits, la lecture est très similaire à celle de l'ACP).



Dans le plan factoriel **(1, 4)**, on retrouve les proximités que nous avions constaté lors de l'ACP sur les variables originelles (CYL, ..., V.MAX)

AFP et Rotation VARIMAX - 4 facteurs sélectionnés

Factor rotation 1

Factors rotation	
Method	VARIMAX
# factors	4
Reproduced correlations	0
Sort variables according to loadings	1

Le principe de la rotation des axes reste valable : on effectue une rotation pour 4 facteurs.

Results

Rotated Factor Loadings

Attribute	Communality Estimates		Axis_1		Axis_2		Axis_3		Axis_4	
-	Prior	Final	Corr.	Sq. (Cumul.)	Corr.	Sq. (Cumul.)	Corr.	Sq. (Cumul.)	Corr.	Sq. (Cumul.)
LONG	0.92590	0.95807	0.90233	*0.81 (0.81)*	0.32470	*0.11 (0.92)*	-0.07687	*0.01 (0.93)*	-0.18038	*0.03 (0.96)*
LARG	0.87988	0.86076	0.87692	*0.77 (0.77)*	0.20204	*0.04 (0.81)*	0.19957	*0.04 (0.85)*	0.10545	*0.01 (0.86)*
POIDS	0.92397	0.87964	0.78972	*0.62 (0.62)*	0.46708	*0.22 (0.84)*	-0.19446	*0.04 (0.88)*	0.00133	*0.00 (0.88)*
CYL	0.84305	0.80309	0.60569	*0.37 (0.37)*	0.59224	*0.35 (0.72)*	-0.01180	*0.00 (0.72)*	0.29214	*0.09 (0.80)*
V.MAX	0.88229	0.85665	0.24441	*0.06 (0.06)*	0.85341	*0.73 (0.79)*	0.26134	*0.07 (0.86)*	-0.01744	*0.00 (0.86)*
PUISS	0.92294	0.91244	0.42328	*0.18 (0.18)*	0.85341	*0.73 (0.91)*	0.05704	*0.00 (0.91)*	0.04134	*0.00 (0.91)*
RND5	0.68805	0.73010	-0.28343	*0.08 (0.08)*	-0.57287	*0.33 (0.41)*	0.56606	*0.32 (0.73)*	0.03407	*0.00 (0.73)*
RND3	0.45084	0.43613	-0.07240	*0.01 (0.01)*	-0.03740	*0.00 (0.01)*	-0.65482	*0.43 (0.44)*	0.02657	*0.00 (0.44)*
RND6	0.50632	0.41872	0.16670	*0.03 (0.03)*	0.04499	*0.00 (0.03)*	0.04771	*0.00 (0.03)*	-0.62179	*0.39 (0.42)*
RND4	0.61007	0.60126	0.35708	*0.13 (0.13)*	0.29677	*0.09 (0.22)*	0.10445	*0.01 (0.23)*	0.61219	*0.37 (0.60)*
RND2	0.39437	0.24426	0.14562	*0.02 (0.02)*	0.36487	*0.13 (0.15)*	-0.13367	*0.02 (0.17)*	0.26844	*0.07 (0.24)*
RND1	0.33669	0.23869	-0.13703	*0.02 (0.02)*	-0.43752	*0.19 (0.21)*	0.15377	*0.02 (0.23)*	-0.06955	*0.00 (0.24)*
Var. Expl.	8.36437	7.93981	3.09346	*37 % (37 %)*	2.91601	*35 % (72 %)*	0.95916	*11 % (83 %)*	0.97118	*12 % (95 %)*

L'inertie expliquée par les 4 facteurs reste la même après rotation.

Encombrement / gamme

Sportivité / performances, avec RND5 qui s'immisce malheureusement

On peut oublier...

Analyse de Harris

Exacerber les corrélations en les divisant par les « uniqueness »

Harris – Matrice à diagonaliser et Tableau des valeurs propres

$$h^*_{jm} = \frac{h_{jm}}{\sqrt{u_j \times u_m}}$$

Souligner d'autant plus les corrélations qu'elles concernent des variables fortement liées aux autres ($R^2_j \approx 1 \rightarrow u_j \approx 0$)

	CYL	PUISS	LONG	LARG	POIDS	V_MA	RND1	RND2	RND3	RND4	RND5	RND6
CYL	5.35	7.24	6.50	4.59	7.22	4.89	-0.88	0.76	-0.20	2.29	-2.48	-0.45
PUISS	7.24	11.94	8.49	5.41	10.00	8.87	-1.67	1.61	-0.53	2.49	-3.72	0.37
LONG	6.50	8.49	12.55	9.00	11.57	5.10	-1.30	1.03	-0.07	1.82	-3.17	1.62
LARG	4.59	5.41	9.00	7.33	7.50	3.98	-0.67	0.91	-0.75	2.23	-1.35	0.46
POIDS	7.22	10.00	11.57	7.50	12.10	5.05	-1.58	1.28	-0.06	2.05	-4.02	0.48
V_MAX	4.89	8.87	5.10	3.98	5.05	7.48	-1.12	1.00	-0.76	1.75	-2.37	0.46
RND1	-0.88	-1.67	-1.30	-0.67	-1.58	-1.12	0.51	-0.70	0.03	-0.35	0.82	-0.04
RND2	0.76	1.61	1.03	0.91	1.28	1.00	-0.70	0.64	0.14	0.88	-0.57	0.02
RND3	-0.20	-0.53	-0.07	-0.75	-0.06	-0.76	0.03	0.14	0.82	-0.06	-0.93	-0.04
RND4	2.29	2.49	1.82	2.23	2.05	1.75	-0.35	0.88	-0.06	1.56	-0.40	-0.55
RND5	-2.48	-3.72	-3.17	-1.35	-4.02	-2.37	0.82	-0.57	-0.93	-0.40	2.21	-0.04
RND6	-0.45	0.37	1.62	0.46	0.48	0.46	-0.04	0.02	-0.04	-0.55	-0.04	1.03

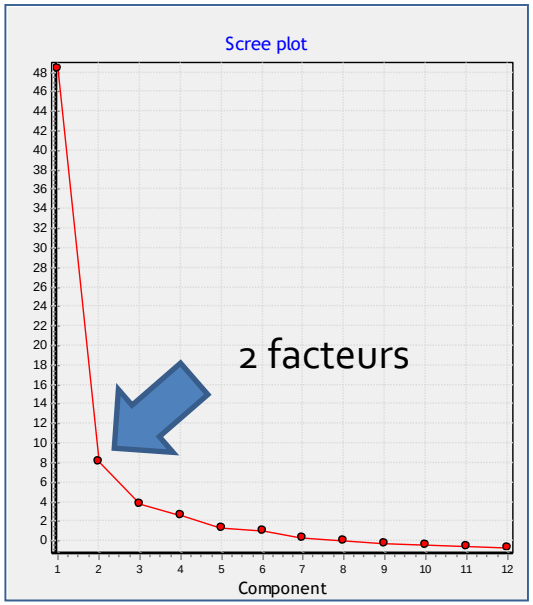
Eigen values

Matrix trace	63.592147
Average	5.299346

Seuil (possible) = Moyenne des valeurs sur la diagonale de la matrice à traiter = TRACE / p

Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	48.363500	40.296952	76.05 %		76.05 %
2	8.066548	4.333778	12.68 %		88.74 %
3	3.732770	1.044507	5.87 %		94.61 %
4	2.688263	1.378686	4.23 %		98.83 %
5	1.309577	0.335532	2.06 %		100.89 %
6	0.974045	0.611522	1.53 %		102.43 %
7	0.362522	0.327260	0.57 %		103.00 %
8	0.035262	0.315022	0.06 %		103.05 %
9	-0.279760	0.094832	-0.44 %		102.61 %
10	-0.374592	0.209262	-0.59 %		102.02 %
11	-0.583854	0.118280	-0.92 %		101.10 %
12	-0.702134	-	-1.10 %		100.00 %
Tot.	63.592147	-	-	-	-

Eboulis des valeurs propres



Les deux points de vue convergent pour une solution en q = 2 facteurs

Harris – Tableau des « loadings » et représentation des individus

R^2_j restituée sur les q = 2 premiers facteurs : somme du carré des « loadings »

Estimations plutôt bonnes individuellement et globalement, car : $6.03354 = \sum_{j=1}^{p} \hat{R}_j^2$

➔ (6.03354/8.36437) = 72% de la variabilité initiale a été reproduite.

Factor Loadings [Communality Estimates]

Attribute	Communality Estimates		Axis_1		Axis_2	
-	Prior	Final	Corr.	Sq. (Cumul.)	Corr.	Sq. (Cumul.)
CYL	0.84305	0.76902	0.86673	*0.75 (0.75)*	-0.13341	*0.02 (0.77)*
PUISS	0.92294	0.92234	0.87969	*0.77 (0.77)*	-0.38533	*0.15 (0.92)*
LONG	0.92590	0.92109	0.88809	*0.79 (0.79)*	0.36384	*0.13 (0.92)*
LARG	0.87988	0.75899	0.79279	*0.63 (0.63)*	0.36122	*0.13 (0.76)*
POIDS	0.92397	0.87330	0.91588	*0.84 (0.84)*	0.18563	*0.03 (0.87)*
V.MAX	0.88229	0.81191	0.72424	*0.52 (0.52)*	-0.53609	*0.29 (0.81)*
RND1	0.33669	0.15944	-0.37559	*0.14 (0.14)*	0.13554	*0.02 (0.16)*
RND2	0.39437	0.12657	0.32900	*0.11 (0.11)*	-0.13539	*0.02 (0.13)*
RND3	0.45084	0.01274	-0.07587	*0.01 (0.01)*	0.08354	*0.01 (0.01)*
RND4	0.61007	0.23796	0.47121	*0.22 (0.22)*	-0.12616	*0.02 (0.24)*
RND5	0.68805	0.38490	-0.60269	*0.36 (0.36)*	0.14717	*0.02 (0.38)*
RND6	0.50632	0.05529	0.12942	*0.02 (0.02)*	0.19632	*0.04 (0.06)*
Unweighted Var. Expl.	8.36437	6.03354	5.16276	*62 % (62 %)*	0.87078	*10 % (72 %)*

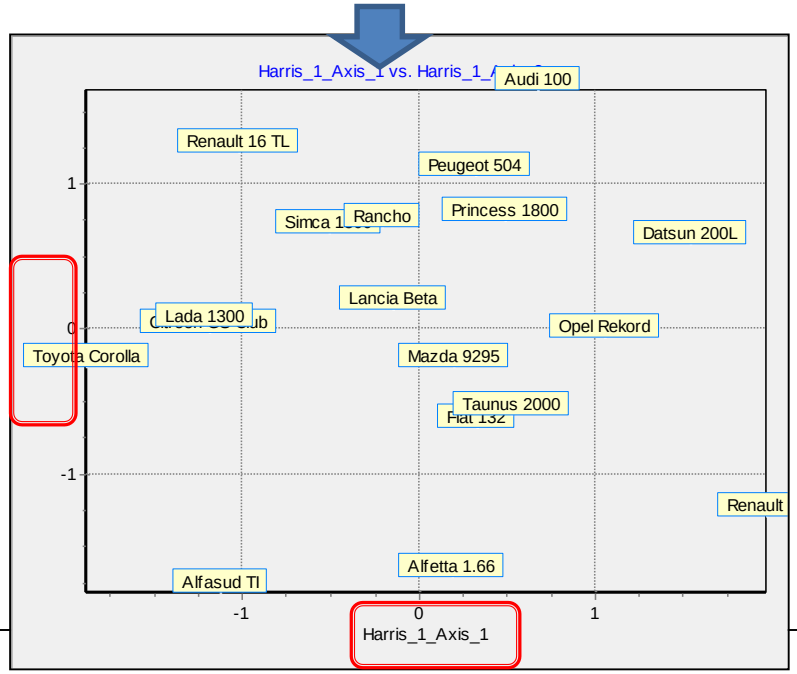
Le rôle de V.MAX et sa relation avec PUISS apparaît dès le 2nd facteur maintenant.

On retrouve les résultats de l'ACP sur variables originelles sans qu'il soit nécessaire de procéder à une rotation des axes. Résultat confirmé par la représentation des individus.

Variance non pondéré expliquée : somme des carrés des loadings des variables avec le facteur :

$(0.86673^2 + \ldots + 0.12942^2) = 5.16276$

Et, 5.16276 + 0.87078 = 6.03354



« Factor scores » - Les coefficients des fonctions de projection

L'analyse en facteurs principaux et l'analyse de Harris fournissent les coefficients permettant de projeter les individus dans le repère factoriel… avec une information supplémentaire : la crédibilité de la fonction de projection.

Factor Score Coefficients

Squared Multiple Corr. of the Variables with Each Factor			0.9797421	0.8897044
Attribute	Mean	Std-dev	Axis_1	Axis_2
CYL	1631.6666667	363.3944903	0.1118729	-0.0937558
PUISS	84.6111111	19.8021853	0.2312499	-0.5515015
LONG	433.5000000	21.4844905	0.2427979	0.5415751
LARG	166.6666667	5.1639778	0.1336967	0.3316641
POIDS	1078.8333333	133.0990650	0.2440358	0.2692996
V.MAX	158.2777778	11.7983311	0.1246392	-0.5023101
RND1	0.2026111	0.9125674	-0.0114707	0.0225371
RND2	0.4241111	0.5716308	0.0110049	-0.0246564
RND3	-0.4291667	0.7927809	-0.0027989	0.0167796
RND4	-0.2902778	0.9949314	0.0244806	-0.0356844
RND5	-0.3491111	0.9778135	-0.0391393	0.0520346
RND6	0.2138333	1.2347276	0.0053107	0.0438613

Indicateur de fiabilité du facteur. Correspond au carré de la corrélation entre la variable latente théorique (à estimer) et son estimation par le facteur (cf. doc SAS).

Plus il est proche de 1, plus le facteur est crédible ; plus il s'éloigne de 1, moins intéressant est le facteur. Selon certaines références, **≥ 0.7** indique une bonne stabilité.

Paramètres pour le centrage et réduction des variables

Coefficients permettant de calculer les coordonnées factorielles des individus (éventuellement supplémentaires) à partir de leur description.

Bibliographie

Les ouvrages incontournables sur l'analyse de données

Escofier B., Pagès J., « Analyses factorielles simples et multiples », Dunod, 2008.

Lebart L., Morineau A., Piron M., « Statistique exploratoire multidimensionnelle », Dunod, 3ème édition, 2000.

Saporta G., « Probabilités, Analyse des Données et Statistique », Technip, 2006.

Tenenhaus M., « Statistique : Méthodes pour décrire, expliquer et prévoir », Dunod, 2006.

Tutoriels et supports de cours (innombrables sur le web) avec, entres autres,

Tutoriel Tanagra, http://tutoriels-data-mining.blogspot.fr/ ; voir la section « Analyse Factorielle ».

Les plus complets (Tanagra, code source R, SAS, etc.), certains traitant le fichier AUTOS :

- « ACP – Description de véhicules » (Mars 2008)
- « Analyse en composantes principales avec R » (Mai 2009)
- « ACP avec R – Détection du nombre d'axes » (Juin 2012)
- « ACP sous R – Indice KMO et test de Bartlett » (Mai 2012)
- « ACP sur corrélations partielles (suite) » (Juin 2012)
- « ACP avec Tanagra – Nouveaux outils » (Juin 2012)
- « Analyse en facteurs principaux » (Sept. 2012)