

Analyse *en composantes principales* *avec R*

Ricco.Rakotomalala
<http://eric.univ-lyon2.fr/~ricco/cours>

Références :

1. G. Saporta, « Probabilités, Analyse de données et Statistique », Dunod, 2006 ; partie théorique, pages 155 à 177 ; partie pratique, pages 177 à 181.
2. Tutoriels Tanagra, « ACP – Description de véhicules », <http://tutoriels-data-mining.blogspot.com/2008/03/acp-description-de-vehicules.html> ; description des mêmes calculs sous le logiciel Tanagra.
3. A. Bouchier, « Statistique et logiciel R », <http://rstat.ouvaton.org> ; description théorique de l'ACP et mise en œuvre sous R avec le package ADE-4.

Objectif de l'étude

Description d'une série de véhicules

Objectifs de l'étude

Ce tutoriel reproduit sous le logiciel R, l'analyse menée dans l'ouvrage de Saporta, pages 177 à 181. De très légères modifications ont été introduites : traitement de variables illustratives quantitatives et traitement d'observations illustratives. Les justifications théoriques et les formules sont disponibles dans le même ouvrage, pages 155 à 177.

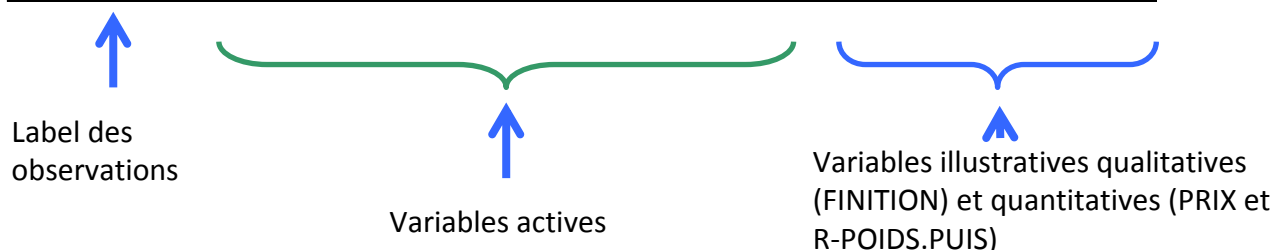
D'autres références ont été utilisées (Lebart et al., Dunod, 200 ; Tenenhaus, Dunod, 2006).

Traitements réalisés

- Réaliser une ACP sur un fichier de données.
- Afficher les valeurs propres. Construire le graphiques éboulis des valeurs propres.
- Construire le cercle de corrélations.
- Projeter les observations dans le premier plan factoriel.
- Positionner des variables illustratives quantitatives dans le cercle des corrélations.
- Positionner les modalités d'une variable illustrative catégorielle.
- Positionner des observations illustratives.

Individus actifs (Données disponibles)

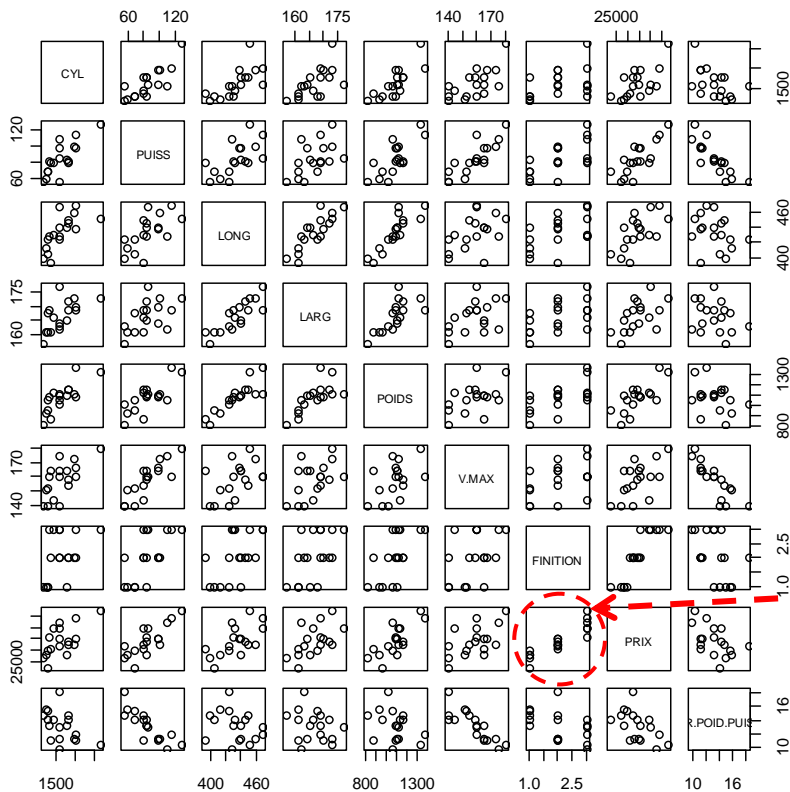
Modele	CYL	PUISS	LONG	LARG	POIDS	V-MAX	FINITION	PRIX	R-POID.PUIS
Alfasud TI	1350	79	393	161	870	165	2_B	30570	11.01
Audi 100	1588	85	468	177	1110	160	3_TB	39990	13.06
Simca 1300	1294	68	424	168	1050	152	1_M	29600	15.44
Citroen GS Club	1222	59	412	161	930	151	1_M	28250	15.76
Fiat 132	1585	98	439	164	1105	165	2_B	34900	11.28
Lancia Beta	1297	82	429	169	1080	160	3_TB	35480	13.17
Peugeot 504	1796	79	449	169	1160	154	2_B	32300	14.68
Renault 16 TL	1565	55	424	163	1010	140	2_B	32000	18.36
Renault 30	2664	128	452	173	1320	180	3_TB	47700	10.31
Toyota Corolla	1166	55	399	157	815	140	1_M	26540	14.82
Alfetta-1.66	1570	109	428	162	1060	175	3_TB	42395	9.72
Princess-1800	1798	82	445	172	1160	158	2_B	33990	14.15
Datsun-200L	1998	115	469	169	1370	160	3_TB	43980	11.91
Taunus-2000	1993	98	438	170	1080	167	2_B	35010	11.02
Rancho	1442	80	431	166	1129	144	3_TB	39450	14.11
Mazda-9295	1769	83	440	165	1095	165	1_M	27900	13.19
Opel-Rekord	1979	100	459	173	1120	173	2_B	32700	11.20
Lada-1300	1294	68	404	161	955	140	1_M	22100	14.04



Fichier de données

Importation, statistiques descriptives et graphiques

```
#bibliothèque lecture fichier excel
library(xlsReadWrite)
#changement de répertoire
setwd("D:/_Travaux/university/Cours_University/Supports_de_cours/Informatique/R/Tutoriels/acp")
#chargement des données dans la première feuille de calcul
#première colonne = label des observations
#les données sont dans la première feuille
autos <- read.xls(file="autos_acp_pour_r.xls",rowNames=T,sheet=1)
#qqv vérifications - affichage
print(autos)
#statistiques descriptives
summary(autos)
#nuages de points
pairs(autos)
#partition des données (var. actives et illustratives)
autos.actifs <- autos[,1:6]
autos.illus <- autos[,7:9]
#nombre d'observations
n <- nrow(autos.actifs)
print(n)
```

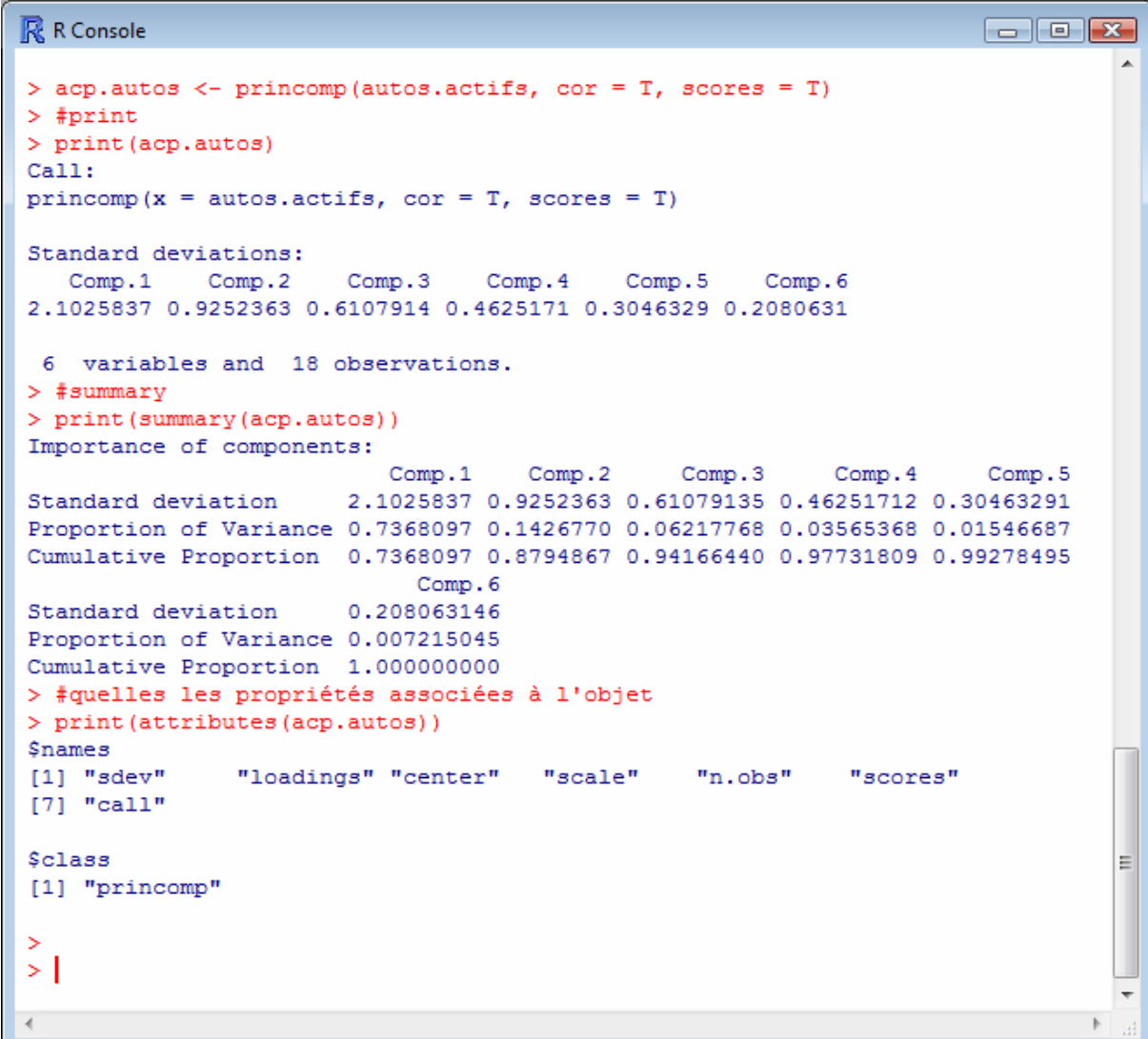


FINITION est une variable qualitative. En général, son introduction dans ce type de graphique n'est pas très indiquée. Néanmoins, on remarquera qu'on peut parfois en tirer des informations utiles : par exemple, ici, selon la finition, les prix sont différents.

Analyse en composantes principales

Utiliser la procédure « princomp » - Résultats immédiats

```
#centrage et réduction des données --> cor = T
#calcul des coordonnées factorielles --> scores = T
acp.autos <- princomp(autos.actifs, cor = T, scores = T)
#print
print(acp.autos)
#summary
print(summary(acp.autos))
#quelles les propriétés associées à l'objet ?
print(attributes(acp.autos))
```



```
R Console

> acp.autos <- princomp(autos.actifs, cor = T, scores = T)
> #print
> print(acp.autos)
Call:
princomp(x = autos.actifs, cor = T, scores = T)

Standard deviations:
  Comp.1   Comp.2   Comp.3   Comp.4   Comp.5   Comp.6
2.1025837 0.9252363 0.6107914 0.4625171 0.3046329 0.2080631

 6 variables and 18 observations.
> #summary
> print(summary(acp.autos))
Importance of components:
              Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
Standard deviation    2.1025837 0.9252363 0.61079135 0.46251712 0.30463291
Proportion of Variance 0.7368097 0.1426770 0.06217768 0.03565368 0.01546687
Cumulative Proportion 0.7368097 0.8794867 0.94166440 0.97731809 0.99278495
              Comp.6
Standard deviation    0.208063146
Proportion of Variance 0.007215045
Cumulative Proportion 1.000000000
> #quelles les propriétés associées à l'objet
> print(attributes(acp.autos))
$names
[1] "sdev"      "loadings" "center"   "scale"    "n.obs"    "scores"
[7] "call"

$class
[1] "princomp"

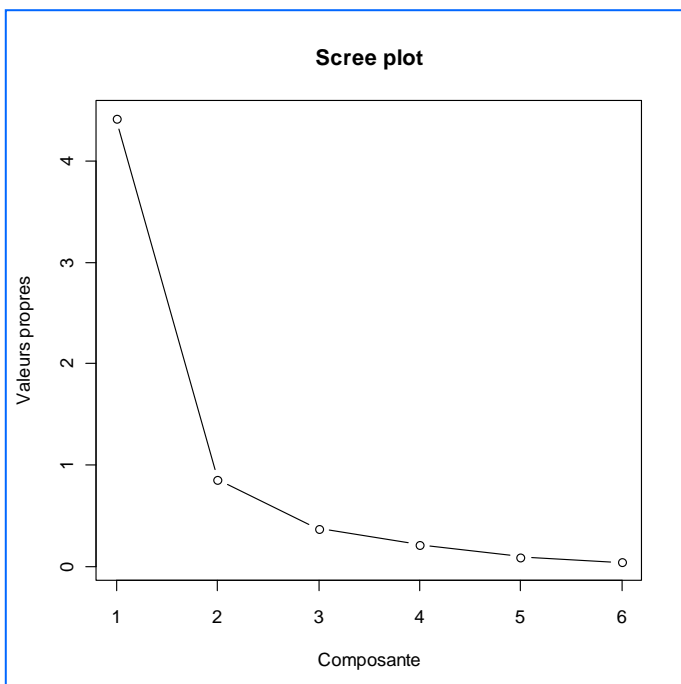
>
> |
```

PRINCOMP fournit les écarts-types associés aux axes. Le carré correspond aux variances = valeur propres. Nous avons également le pourcentage cumulé. Avec ATTRIBUTES, nous avons la liste des informations que nous pourrions exploiter par la suite.

Valeurs propres associés aux axes

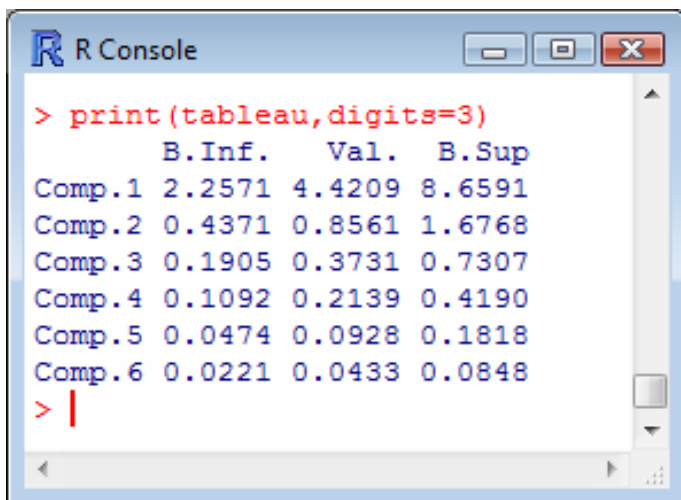
Calcul, intervalles de confiance et Scree Plot

```
#obtenir les variances associées aux axes c.-à-d. les valeurs propres
val.propres <- acp.autos$sdev^2
print(val.propres)
#scree plot (graphique des éboulis des valeurs propres)
plot(1:6,val.propres,type="b",ylab="Valeurs
propres",xlab="Composante",main="Scree plot")
#intervalle de confiance des val.propres à 95% (cf.Saporta, page 172)
val.basse <- val.propres * exp(-1.96 * sqrt(2.0/(n-1)))
val.haute <- val.propres * exp(+1.96 * sqrt(2.0/(n-1)))
#affichage sous forme de tableau
tableau <- cbind(val.basse,val.propres,val.haute)
colnames(tableau) <- c("B.Inf.", "Val.", "B.Sup")
print(tableau,digits=3)
```



Les deux premiers axes traduisent 88% de l'information disponible. On se rend compte ici qu'on pouvait s'en tenir uniquement au premier facteur.

Mais c'est moins pratique pour les graphiques ; on suspecte aussi un « effet taille » dans les données. On va donc conserver les deux premiers facteurs.



Les intervalles de confiance d'Anderson ne sont licites que si le nuage de points est gaussien. On ne l'affiche donc qu'à titre indicatif (cf. formules page 172 de Saporta).

Cercle des corrélations

Variables actives

```
##### corrélation variables-facteurs #####
c1 <- acp.autos$loadings[,1]*acp.autos$sdev[1]
c2 <- acp.autos$loadings[,2]*acp.autos$sdev[2]
#affichage
correlation <- cbind(c1,c2)
print(correlation,digits=2)
#carrés de la corrélation (cosinus²)
print(correlation^2,digits=2)
#cumul carrés de la corrélation
print(t(apply(correlation^2,1,cumsum)),digits=2)
### cercle des corrélations - variables actives ###
plot(c1,c2,xlim=c(-1,+1),ylim=c(-1,+1),type="n")
abline(h=0,v=0)
text(c1,c2,labels=colnames(autos.actifs),cex=0.5)
symbols(0,0,circles=1,inches=F,add=T)
```

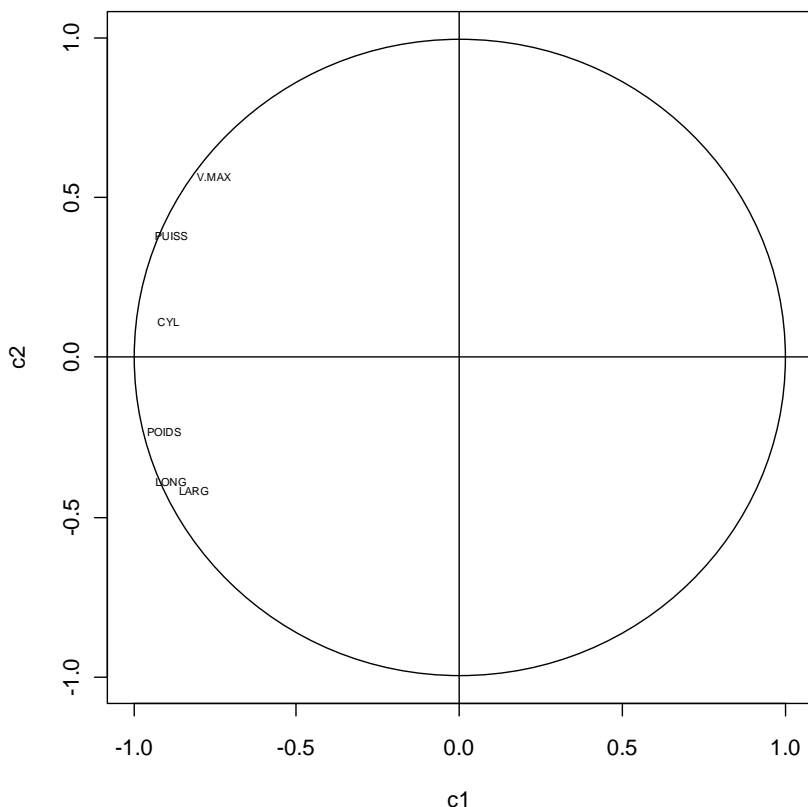
Corrélation variables - facteurs.

```
> print(correlation,digits$
      c1  c2
CYL  -0.89  0.11
PUISS -0.89  0.38
LONG  -0.89 -0.38
LARG  -0.81 -0.41
POIDS -0.91 -0.22
V.MAX -0.75  0.57
>
> #carrés de la corrélation$
> print(correlation^2,digi$
      c1  c2
CYL   0.80 0.013
PUISS 0.79 0.148
LONG  0.79 0.145
LARG  0.66 0.170
POIDS 0.82 0.050
V.MAX 0.57 0.329
>
> #cumul carrés de la corrélation$
> print(t(apply(correlation$
      c1  c2
CYL   0.80 0.81
PUISS 0.79 0.93
LONG  0.79 0.93
LARG  0.66 0.83
POIDS 0.82 0.87
V.MAX 0.57 0.90
```

Carré de la
corrélation.

Carré cumulé. Au
sixième axe, toutes les
valeurs sont égales à 1.

Cercle des corrélations.

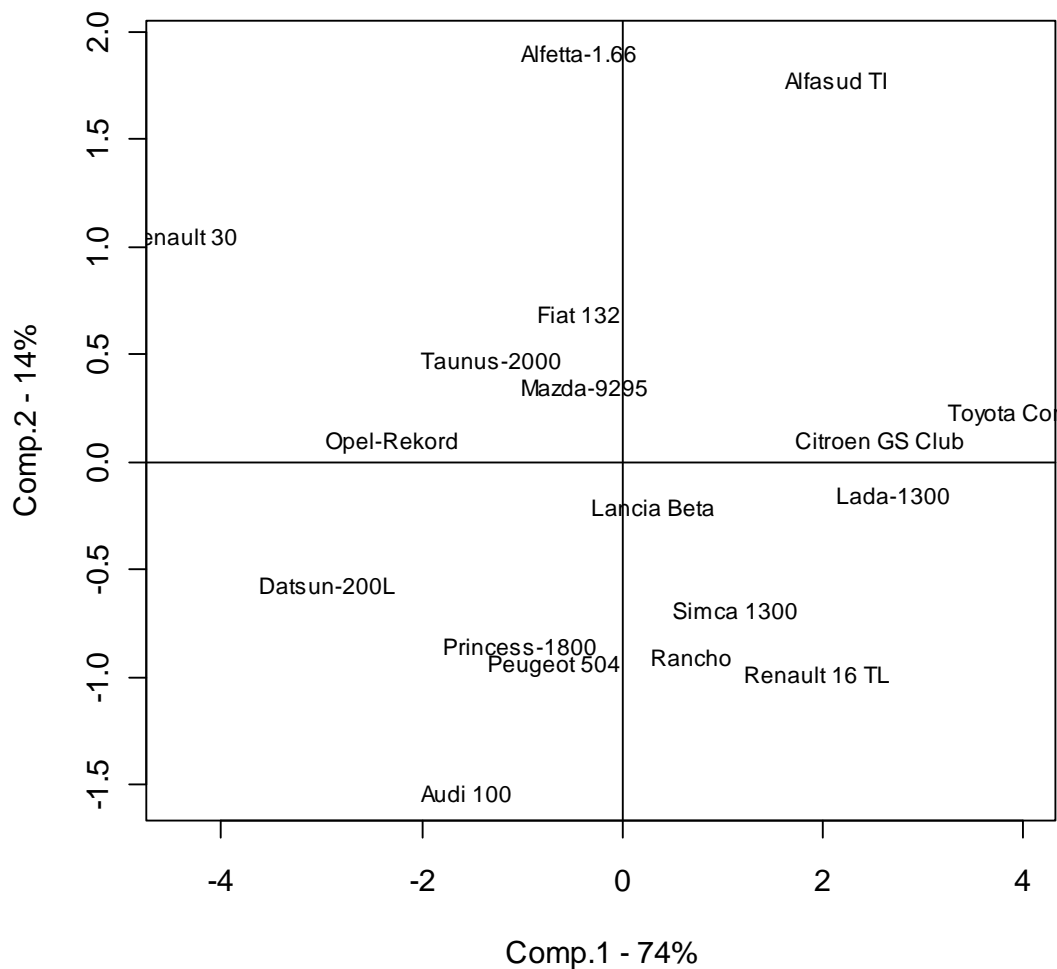


Carte des individus sur les 2 premiers axes

Individus actifs

```
#l'option "scores" demandé dans princomp est très important ici
plot(acp.autos$scores[,1],acp.autos$scores[,2],type="n",xlab="Comp.1 -
74%",ylab="Comp.2 - 14%")
abline(h=0,v=0)
text(acp.autos$scores[,1],acp.autos$scores[,2],labels=rownames(autos.actifs),cex=0.75)
```

Carte des individus dans le premier plan factoriel (88% de l'inertie)

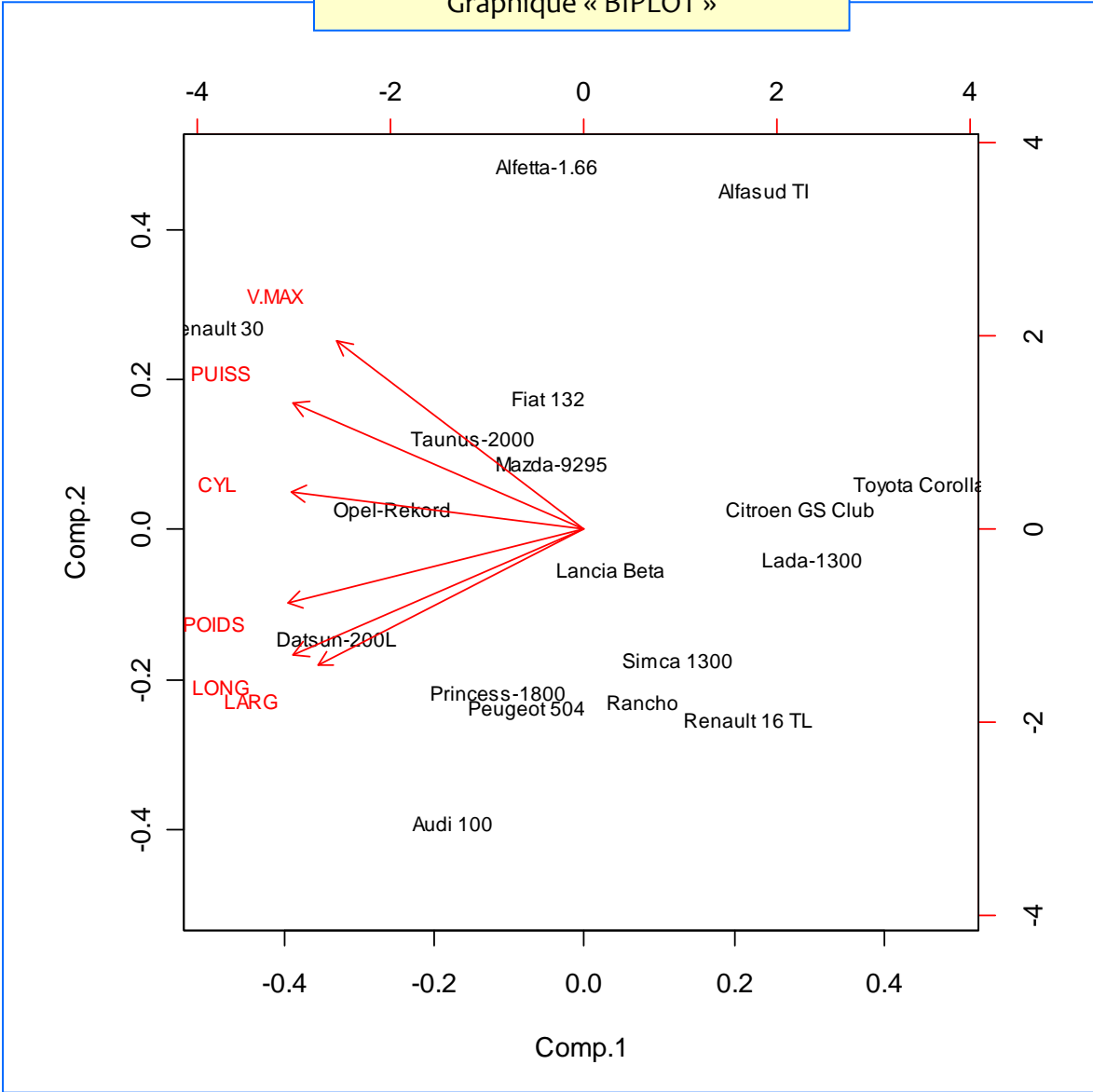


Carte des individus et des variables

Individus actifs et variables actives

```
*** représentation simultanée : individus x variables  
# cf. Lebart et al., pages 46-48  
#*****  
biplot(acp.autos,cex=0.75)
```

Graphique « BIPLLOT »

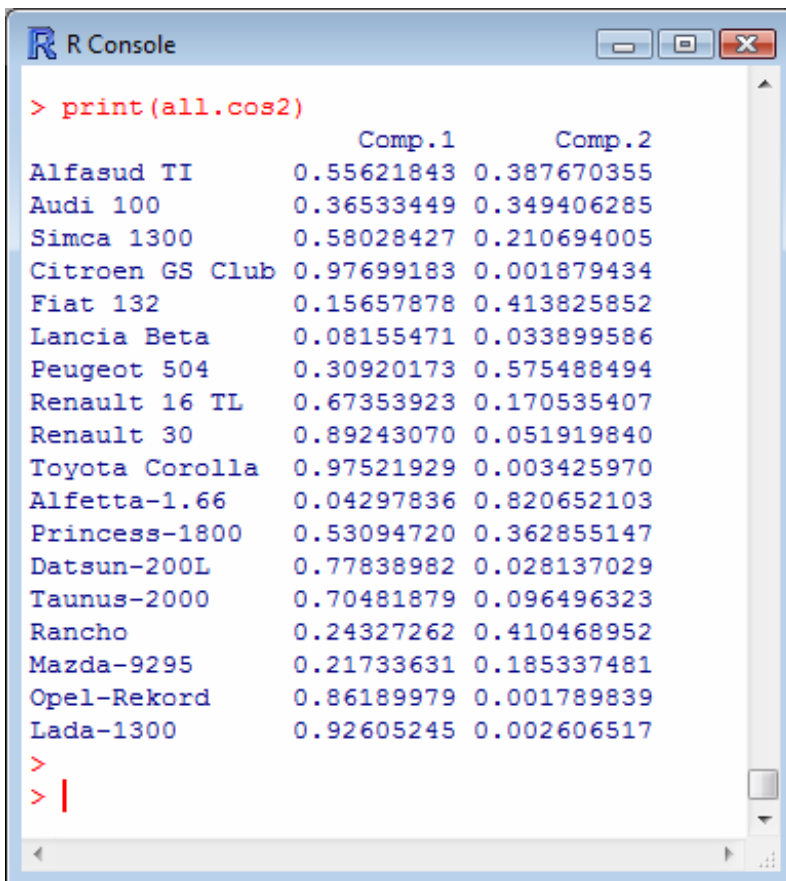


Attention : N'allons surtout pas chercher une quelconque proximité entre les individus et les variables dans ce graphique. Ce sont les directions qui sont importantes !!! (cf. principe et formules dans Lebart et al., pages 46 à 48)

COSINUS² des individus avec les composantes

Qualité de représentation des individus sur les composantes

```
#calcul du carré de la distance d'un individu au center de gravité
d2 <- function(x){return(sum(((x-acp.autos$center)/acp.autos$scale)^2))}
#appliquer à l'ensemble des observations
all.d2 <- apply(autos.actifs,1,d2)
#cosinus^2 pour une composante
cos2 <- function(x){return(x^2/all.d2)}
#cosinus^2 pour les composantes retenues (les 2 premières)
all.cos2 <- apply(acp.autos$scores[,1:2],2,cos2)
print(all.cos2)
```



```
> print(all.cos2)
```

	Comp.1	Comp.2
Alfasud TI	0.55621843	0.387670355
Audi 100	0.36533449	0.349406285
Simca 1300	0.58028427	0.210694005
Citroen GS Club	0.97699183	0.001879434
Fiat 132	0.15657878	0.413825852
Lancia Beta	0.08155471	0.033899586
Peugeot 504	0.30920173	0.575488494
Renault 16 TL	0.67353923	0.170535407
Renault 30	0.89243070	0.051919840
Toyota Corolla	0.97521929	0.003425970
Alfetta-1.66	0.04297836	0.820652103
Princess-1800	0.53094720	0.362855147
Datsun-200L	0.77838982	0.028137029
Taunus-2000	0.70481879	0.096496323
Rancho	0.24327262	0.410468952
Mazda-9295	0.21733631	0.185337481
Opel-Rekord	0.86189979	0.001789839
Lada-1300	0.92605245	0.002606517

```
>
> |
```

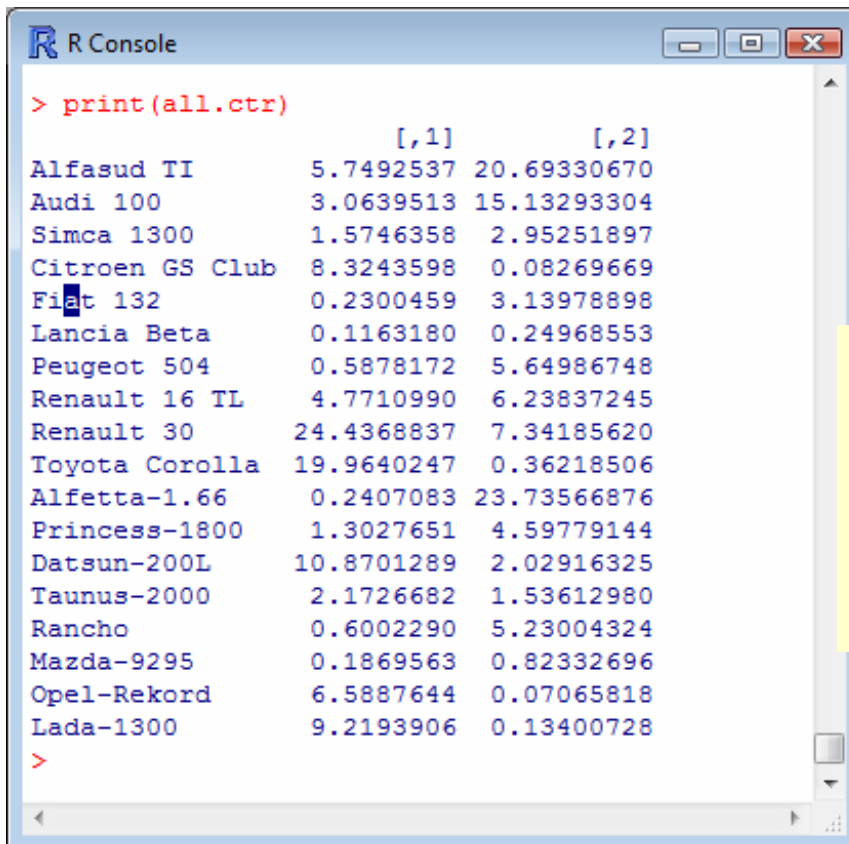
La somme pour chaque ligne (individu) vaut 1 si l'on prend l'ensemble des composantes (les 6 composantes).

Cf formules dans Tenenhaus, 2006 ; page 162.

CONTRIBUTION des individus aux composantes

Déterminer les individus qui pèsent le plus dans la définition d'une composante

```
#contributions à une composante - calcul pour les 2 premières
composantes
all.ctr <- NULL
for (k in 1:2){all.ctr <-
  cbind(all.ctr,100.0*(1.0/n)*(acp.autos$scores[,k]^2)/
    (acp.autos$sdev[k]^2))}
print(all.ctr)
```



```
> print(all.ctr)
           [,1]      [,2]
Alfasud TI    5.7492537 20.69330670
Audi 100      3.0639513 15.13293304
Simca 1300    1.5746358  2.95251897
Citroen GS Club 8.3243598 0.08269669
Fiat 132      0.2300459  3.13978898
Lancia Beta   0.1163180  0.24968553
Peugeot 504   0.5878172  5.64986748
Renault 16 TL 4.7710990  6.23837245
Renault 30    24.4368837  7.34185620
Toyota Corolla 19.9640247 0.36218506
Alfetta-1.66  0.2407083 23.73566876
Princess-1800 1.3027651  4.59779144
Datsun-200L   10.8701289  2.02916325
Taunus-2000   2.1726682  1.53612980
Rancho        0.6002290  5.23004324
Mazda-9295    0.1869563  0.82332696
Opel-Rekord   6.5887644  0.07065818
Lada-1300     9.2193906  0.13400728
>
```

La somme pour chaque colonne (composante) vaut 100.

cf. Saporta, page 175.

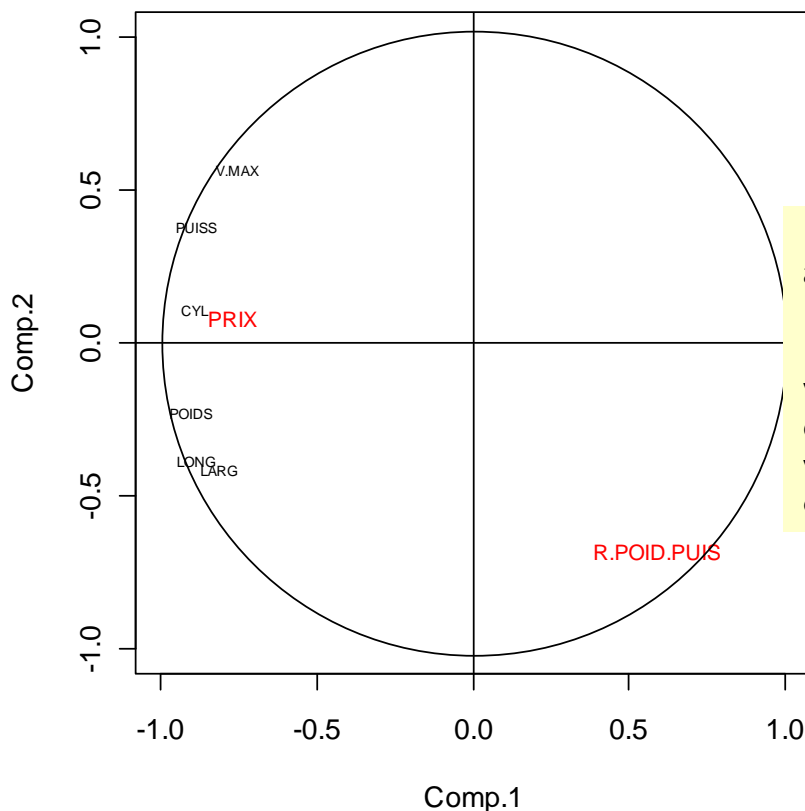
Remarque : toutes les observations ont le même poids dans notre exemple.

Variables quantitatives illustratives

Positionnement dans le cercle des corrélations

```
#####
#que faut-il penser de PRIX et R.POIS.PUIS ?
#####
#corrélation de chaque var. illustrative avec le premier axe
ma_cor_1 <- fonction(x){return(cor(x,acp.autos$scores[,1]))}
s1 <- sapply(autos.illus[,2:3],ma_cor_1)
#corrélation de chaque variable illustrative avec le second axe
ma_cor_2 <- fonction(x){return(cor(x,acp.autos$scores[,2]))}
s2 <- sapply(autos.illus[,2:3],ma_cor_2)
#position sur le cercle
plot(s1,s2,xlim=c(-1,+1),ylim=c(-1,+1),type="n",main="Variables
illustratives",xlab="Comp.1",ylab="Comp.2")
abline(h=0,v=0)
text(s1,s2,labels=colnames(autos.illus[2:3]),cex=1.0)
symbols(0,0,circles=1,inches=F,add=T)
#représentation simultanée (avec les variables actives)
plot(c(c1,s1),c(c2,s2),xlim=c(-1,+1),ylim=c(-1,+1),type="n",main="Variables
illustratives",xlab="Comp.1",ylab="Comp.2")
text(c1,c2,labels=colnames(autos.actifs),cex=0.5)
text(s1,s2,labels=colnames(autos.illus[2:3]),cex=0.75,col="red")
abline(h=0,v=0)
symbols(0,0,circles=1,inches=F,add=T)
```

Variables illustratives



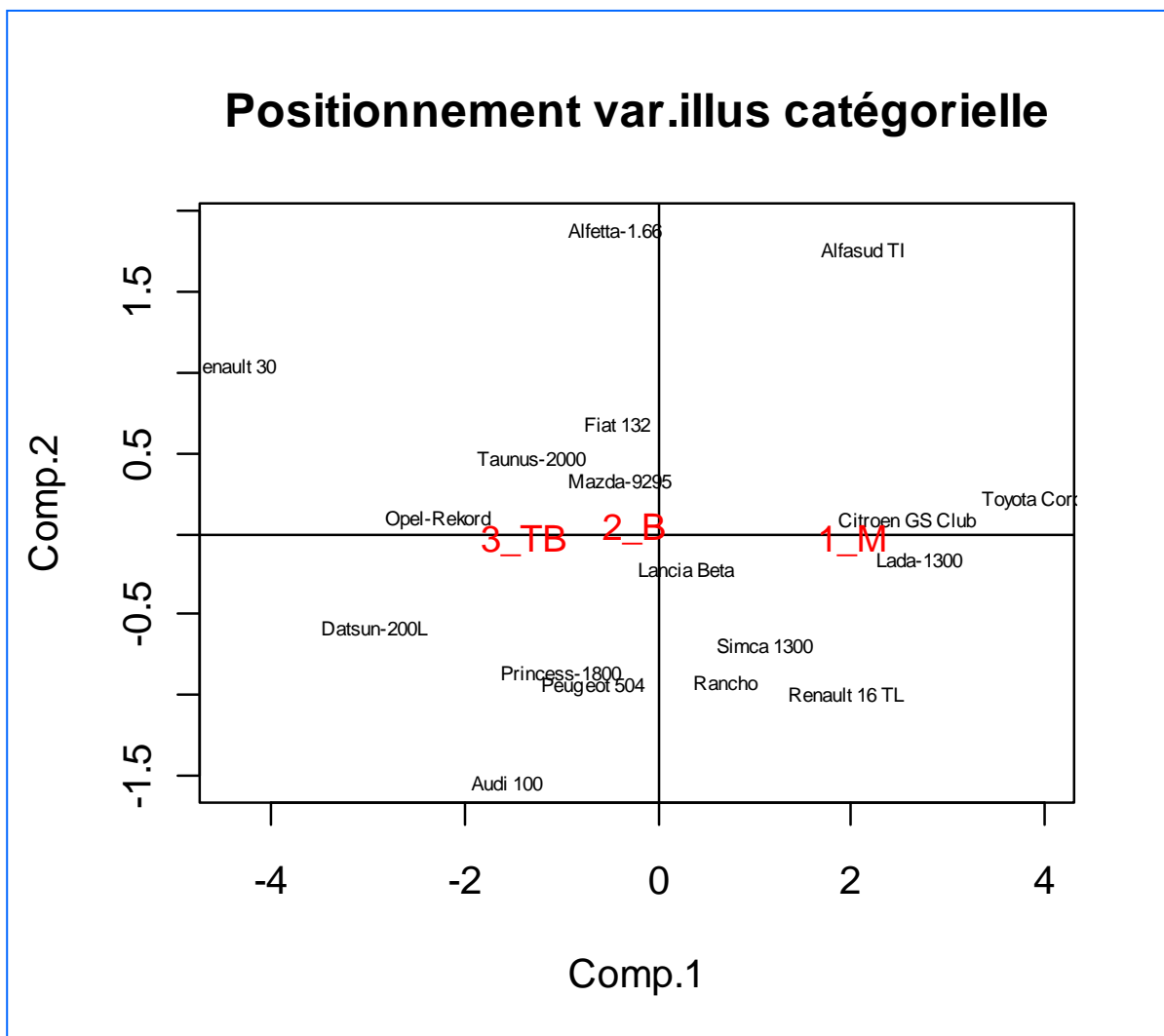
La représentation simultanée est autrement plus instructive.

La forte corrélation entre une variable illustrative et un axe est d'autant plus intéressante que la variable n'a pas participé à la construction de l'axe.

Variables qualitatives illustratives

Positionner les groupes associés aux modalités de la variable illustratives

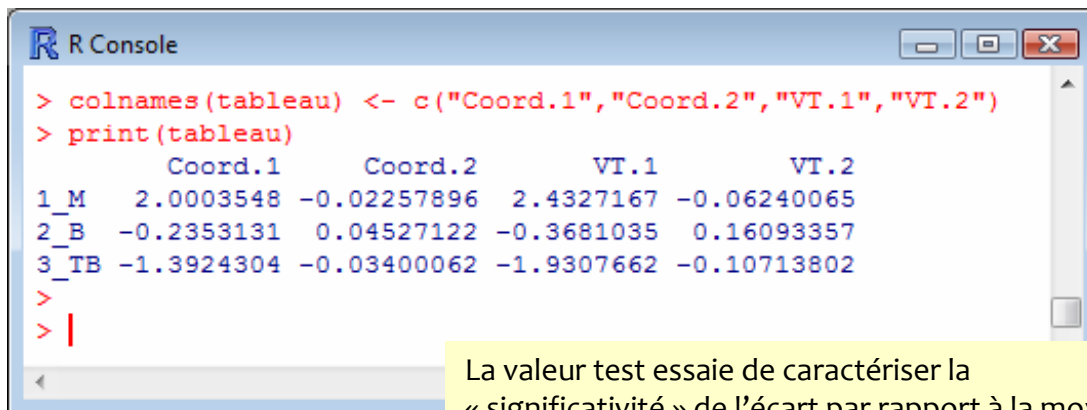
```
#####  
#positionner les modalités de la variable illustrative + calcul des valeurs test  
#####  
K <- nlevels(autos.illus[,"FINITION" ] )  
var.illus <- unclass(autos.illus[,"FINITION" ] )  
m1 <- c()  
m2 <- c()  
for (i in 1:K){m1[i] <- mean(acp.autos$scores[var.illus==i,1])}  
for (i in 1:K){m2[i] <- mean(acp.autos$scores[var.illus==i,2])}  
cond.moyenne <- cbind(m1,m2)  
rownames(cond.moyenne) <- levels(autos.illus[,"FINITION" ] )  
print(cond.moyenne)  
#graphique  
plot(c(acp.autos$scores[,1],m1),c(acp.autos$scores[,2],m2),xlab="Comp.1",ylab="Com  
p.2",main="Positionnement var.illus catégorielle",type="n")  
abline(h=0,v=0)  
text(acp.autos$scores[,1],acp.autos$scores[,2],rownames(autos.actifs),cex=0.5)  
text(m1,m2,rownames(cond.moyenne),cex=0.95,col="red")
```



Variables qualitatives illustratives

Moyennes conditionnelles et valeurs test

```
#####  
*** calcul des valeurs test ***  
#####  
#effectifs par modalité  
nk <- as.vector(table(var.illus))  
#valeur test par composante (les 2 premières)  
vt <- NULL  
for (j in 1:2){vt <- cbind(vt,cond.moyenne[,j]/sqrt((val.propres[j]/nk)*((n-  
nk)/(n-1))))}  
#affichage des valeurs  
tableau <- cbind(cond.moyenne,vt)  
colnames(tableau) <- c("Coord.1", "Coord.2", "VT.1", "VT.2")  
print(tableau)
```



```
R Console  
> colnames(tableau) <- c("Coord.1", "Coord.2", "VT.1", "VT.2")  
> print(tableau)  
      Coord.1  Coord.2  VT.1  VT.2  
1_M  2.0003548 -0.02257896  2.4327167 -0.06240065  
2_B  -0.2353131  0.04527122 -0.3681035  0.16093357  
3_TB -1.3924304 -0.03400062 -1.9307662 -0.10713802  
>  
> |
```

La valeur test essaie de caractériser la « significativité » de l'écart par rapport à la moyenne globale.

Cf. Saporta, page 177. On considère qu'il y a un écartement significatif lorsqu'elle est supérieure, en valeur absolue, à 2 voire 3.

Dans notre exemple, les véhicules se différencient véritablement par la FINITION sur le premier axe factoriel.

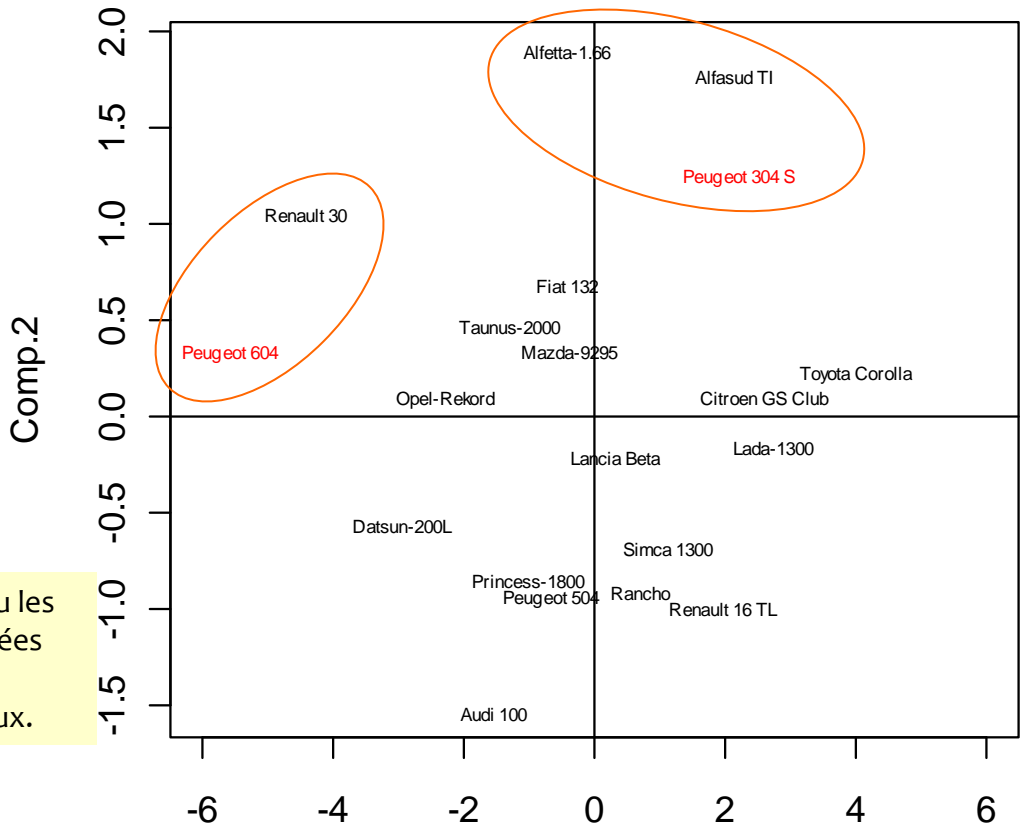
Individus illustratifs

Positionner des individus n'ayant pas participé à la construction des axes

Comment situer ces deux nouveaux 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

```
#chargement de la seconde feuille de calcul + vérification
ind.illus <- read.xls(file="autos_acp_pour_r.xls",rowNames=T,sheet=2)
summary(ind.illus)
#centrage et réduction en utilisant les moyennes et écarts-type de l'ACP
ind.scaled <- NULL
for (i in 1:nrow(ind.illus)){ind.scaled <- rbind(ind.scaled,(ind.illus[i,] -
acp.autos$center)/acp.autos$scale)}
print(ind.scaled)
#calcul des coordonnées factorielles (en utilisant les valeurs propres cf. loadings)
produit.scal <- function(x,k){return(sum(x*acp.autos$loadings[,k]))}
ind.coord <- NULL
for (k in 1:2){ind.coord <- cbind(ind.coord,apply(ind.scaled,1,produit.scal,k))}
print(ind.coord)
#*** projection des individus actifs ET illustratifs dans le premier plan factoriel
plot(c(acp.autos$scores[,1],ind.coord[,1]),c(acp.autos$scores[,2],ind.coord[,2]),xlim
=c(-6,6),type="n",xlab="Comp.1",ylab="Comp.2")
abline(h=0,v=0)
text(acp.autos$scores[,1],acp.autos$scores[,2],labels=rownames(autos.actifs),cex=0.5)
text(ind.coord[,1],ind.coord[,2],labels=rownames(ind.illus),cex=0.5,col="red")
```



Si on connaît un peu les voitures de ces années là, tout ça paraît cohérent. Tant mieux.

***Et on peut faire bien
d'autres choses encore...***