

Analyse *en composantes principales* *avec les packages* *FactoMineR et dynGraph* *du logiciel 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. F. Husson, J. Josse, S. Le, J. Pages, Le package FactoMineR pour R ; <http://factominer.free.fr/>
4. S. Le, J. Durand, Le package dynGraph pour R ; <http://dyngraph.free.fr/>

Objectif de l'étude

Description d'une série de véhicules

Objectifs de l'étude

Ce document reproduit une étude décrite dans un précédent tutoriel, basée sur la procédure **princomp()** de R (<http://tutoriels-data-mining.blogspot.com/2009/05/analyse-en-composantes-principales-avec.html>). Nous utilisons la procédure **PCA** du package **FactoMineR** cette fois-ci. Nous complétons l'analyse avec une exploration graphique interactive à l'aide de **dynGraph**. **Les résultats sont absolument cohérents (heureusement !!!). On se rendra compte surtout que beaucoup de choses ont été mis en place pour nous faciliter la tâche (ex. calcul des coordonnées, des valeurs tests, etc. pour les variables supplémentaires).**

Nous reprenons une analyse décrite dans l'ouvrage de Saporta, pages 177 à 181. Le traitement des variables illustratives quantitatives a été rajoutée. Les justifications théoriques et les formules sont disponibles dans le même ouvrage, pages 155 à 177.

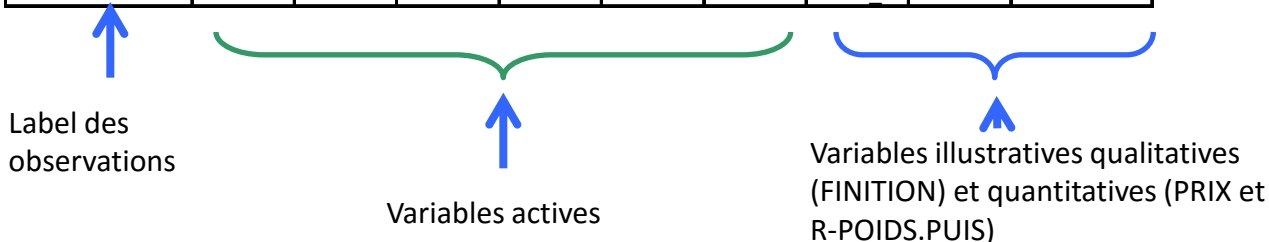
D'autres références ont été utilisées (Lebart et al., Dunod, 200 ; Tenenhaus, Dunod, 2006).

Traitements réalisés

- Réaliser une ACP sur un fichier de données.
- Afficher les valeurs propres. Construire les 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.
- Exploration graphique interactive à l'aide de dynGraph

Données disponibles

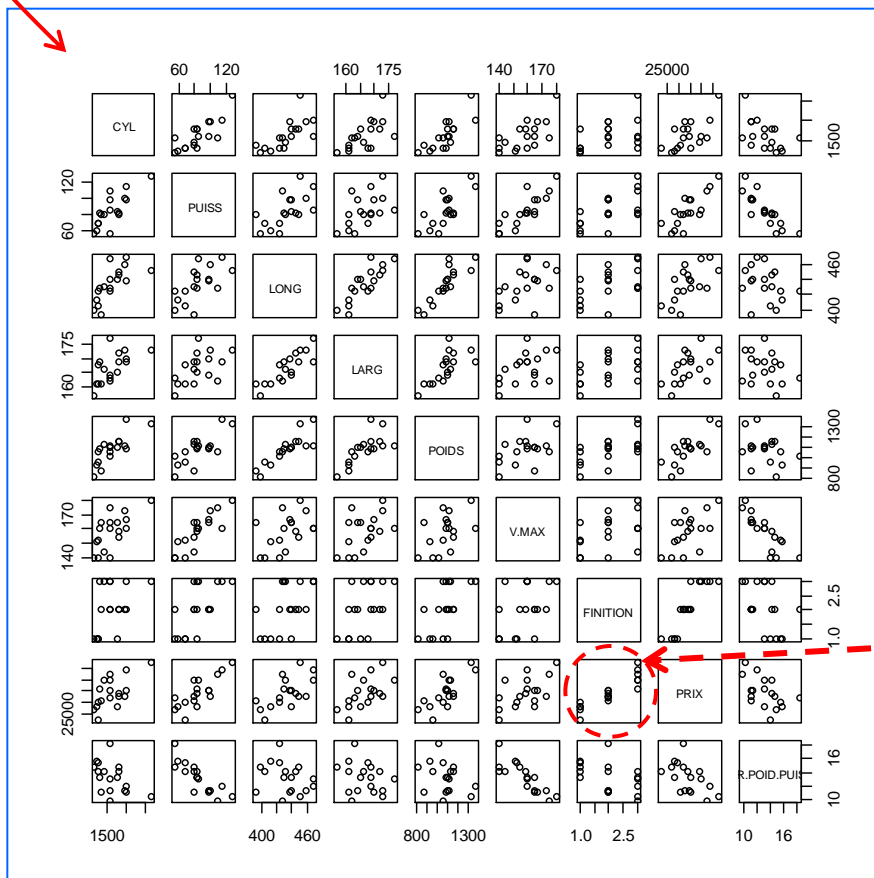
Modele	CYL	PUISS	LONG	LARG	POIDS	V-MAX	FINITION	PRIX	R-POID.PUIS
Alfasud TI	1350	79	393	161	870	165	2 B	30570	11.01
Audi 100	1588	85	468	177	1110	160	3 TB	39990	13.06
Simca 1300	1294	68	424	168	1050	152	1 M	29600	15.44
Citroen GS Club	1222	59	412	161	930	151	1 M	28250	15.76
Fiat 132	1585	98	439	164	1105	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(xlsx)
#changement de répertoire
setwd("c:/mon répertoire/...")
#chargement des données dans la première feuille de calcul
#les données sont dans la première feuille
autos.data <- read.xlsx(file="autos_acp_factominer_dyngraph.xls",header=T, sheetIndex=1)
#première colonne = label (étiquette) des observations
rownames(autos.data) <- autos.data[,1]
autos.data <- autos.data[,-1]
#qqq vérifications - affichage
print(autos.data)
#statistiques descriptives
summary(autos.data)
#nuages de points
pairs(autos.data)
```

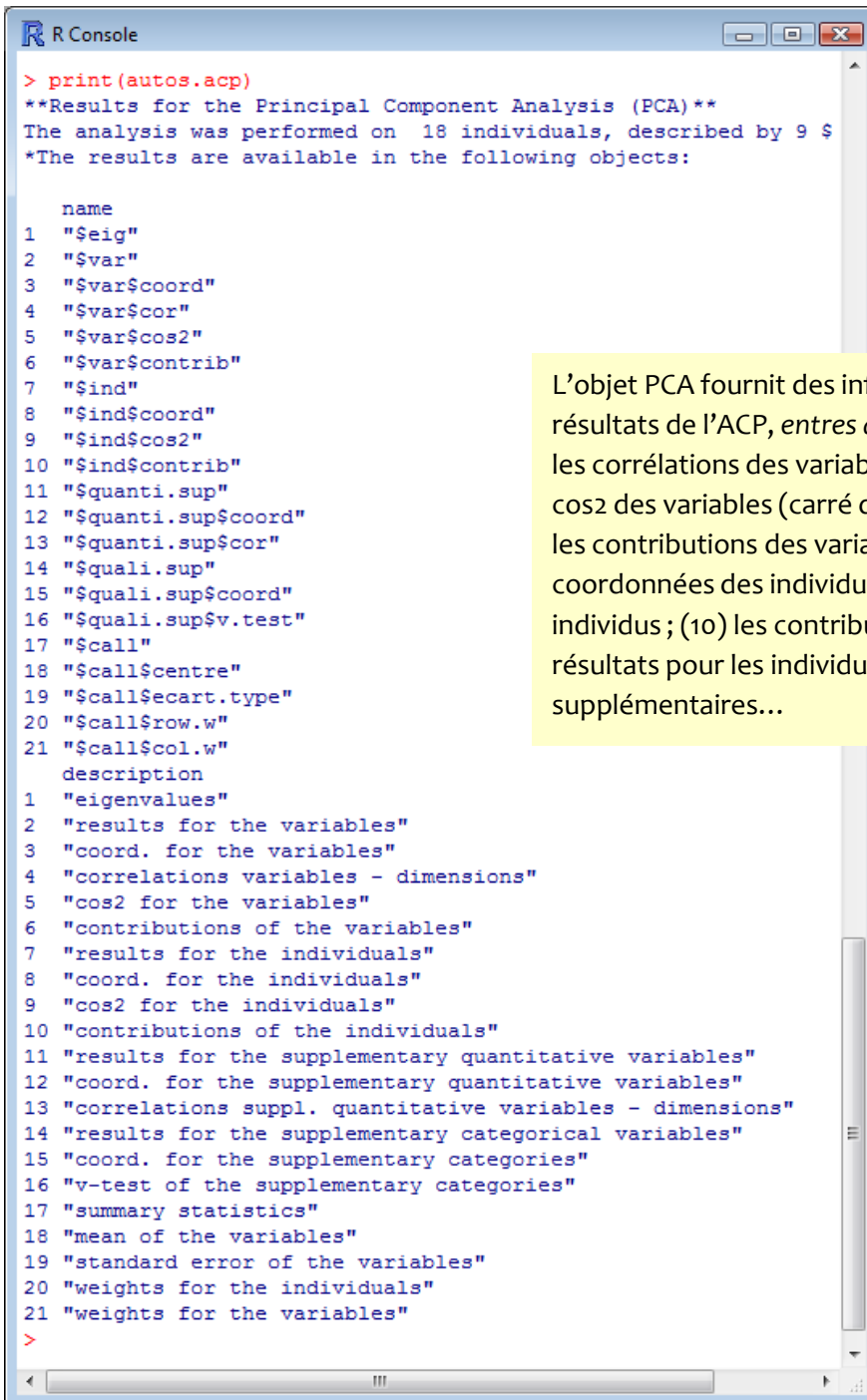


FINITION est une variable qualitative. En général, son introduction dans ce type de graphique n'est pas très indiquée. Néanmoins, on remarquera qu'on peut parfois en tirer des informations utiles : par exemple, ici, selon la finition, les prix sont différents.

Analyse en composantes principales

Utiliser la procédure « PCA » de FactoMineR - Résultats immédiats

```
#centrage et réduction des données --> scale.unit = T
#quanti.sup -> numéro des colonnes des var. quanti. supp.
#quali.sup -> numéro des colonnes des var. quali. supp.
#pas de graphiques pour l'instant -> graph = F
autos.acp <- PCA(autos.data,scale.unit = T,quanti.sup=8:9,quali.sup=7,graph=F)
#obtenir les propriétés de l'objet autos.acp
print(autos.acp)
```



```
R Console
> print(autos.acp)
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 18 individuals, described by 9 $
*The results are available in the following objects:

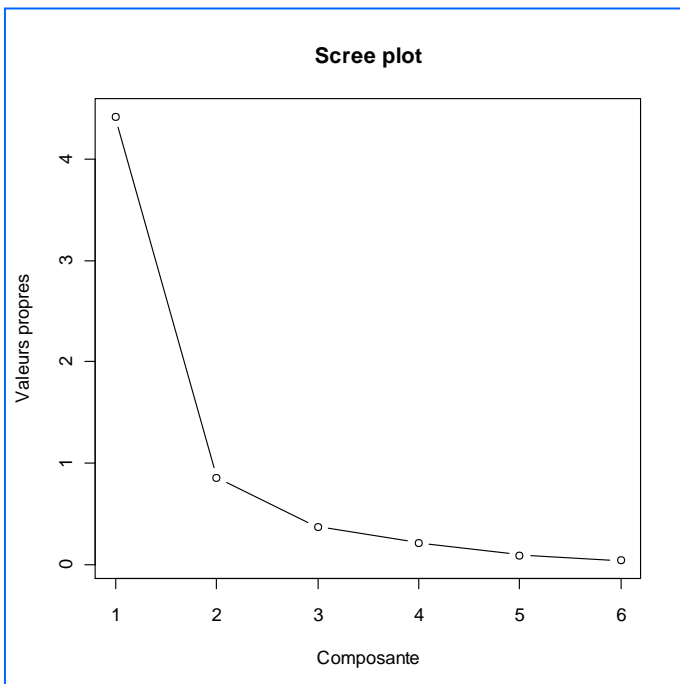
  name
1  "$eig"
2  "$var"
3  "$var$coord"
4  "$var$cor"
5  "$var$cos2"
6  "$var$contrib"
7  "$ind"
8  "$ind$coord"
9  "$ind$cos2"
10 "$ind$contrib"
11 "$quanti.sup"
12 "$quanti.sup$coord"
13 "$quanti.sup$cor"
14 "$quali.sup"
15 "$quali.sup$coord"
16 "$quali.sup$v.test"
17 "$call"
18 "$call$centre"
19 "$call$ecart.type"
20 "$call$row.w"
21 "$call$col.w"
description
1  "eigenvalues"
2  "results for the variables"
3  "coord. for the variables"
4  "correlations variables - dimensions"
5  "cos2 for the variables"
6  "contributions of the variables"
7  "results for the individuals"
8  "coord. for the individuals"
9  "cos2 for the individuals"
10 "contributions of the individuals"
11 "results for the supplementary quantitative variables"
12 "coord. for the supplementary quantitative variables"
13 "correlations suppl. quantitative variables - dimensions"
14 "results for the supplementary categorical variables"
15 "coord. for the supplementary categories"
16 "v-test of the supplementary categories"
17 "summary statistics"
18 "mean of the variables"
19 "standard error of the variables"
20 "weights for the individuals"
21 "weights for the variables"
>
```

L'objet PCA fournit des informations très complètes sur les résultats de l'ACP, *entres autres* : (1) les valeurs propres ; (4) les corrélations des variables avec les axes factoriels ; (5) les cos2 des variables (carré des corrélations) avec les axes ; (6) les contributions des variables aux axes ; (8) les coordonnées des individus ; (9) les cosinus carrés des individus ; (10) les contributions des individus ; (11 à 16) les résultats pour les individus et / ou variables supplémentaires...

Valeurs propres associées aux axes

Calcul, intervalles de confiance et Scree Plot

```
#obtenir les variances associées aux axes c.-à-d. les valeurs propres
val.propres <- autos.acp$eig[,1]
#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% (Saporta, page 172)
n <- nrow(autos.data)
val.basse <- val.propres * exp(-1.96 * sqrt(2.0/(n-1)))
val.haute <- val.propres * exp(+1.96 * sqrt(2.0/(n-1)))
#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.

```
R Console
> print(tableau,digits=3)
      B.Inf.  Val.  B.Sup
Comp.1 2.2571 4.4209 8.6591
Comp.2 0.4371 0.8561 1.6768
Comp.3 0.1905 0.3731 0.7307
Comp.4 0.1092 0.2139 0.4190
Comp.5 0.0474 0.0928 0.1818
Comp.6 0.0221 0.0433 0.0848
> |
```

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 #####
print(autos.acp$var$cor[,1:2],digits=2)

#carrés de la corrélation = cos2
print(autos.acp$var$cos2[,1:2],digits=2)

#cumul carrés de la corrélation
print(t(apply(autos.acp$var$cos2[,1:2],1,cumsum)),digits=2)

### cercle des corrélations - variables actives ###
plot(autos.acp,choix="var",title="Cercle des corrélations", invisible="quanti.sup")
```

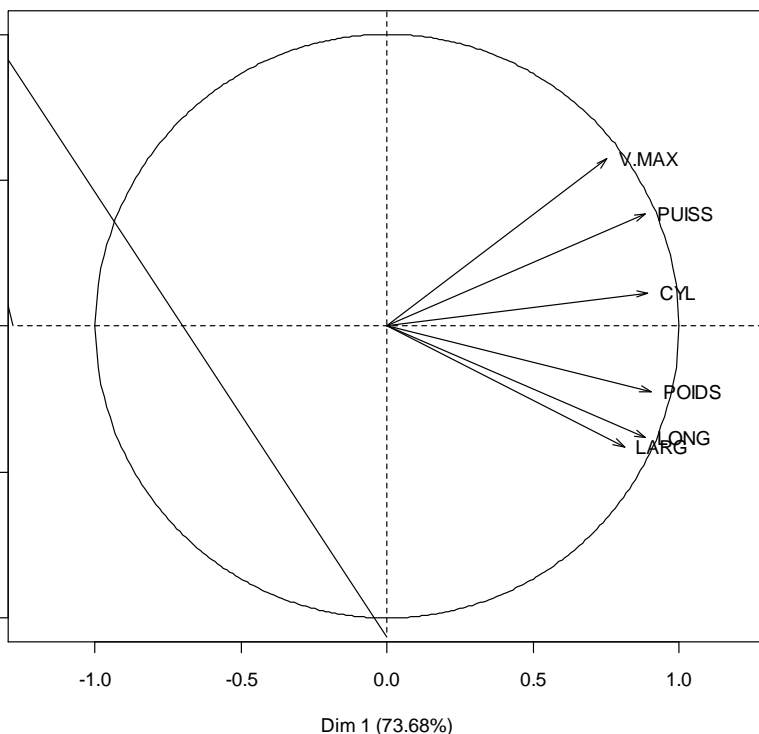
Corrélation variables - facteurs.

```
> print(autos.acp$var$cor[,1$
  Dim.1 Dim.2
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(autos.acp$var$cos2[, $
  Dim.1 Dim.2
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(autos.acp$va$
  Dim.1 Dim.2
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.

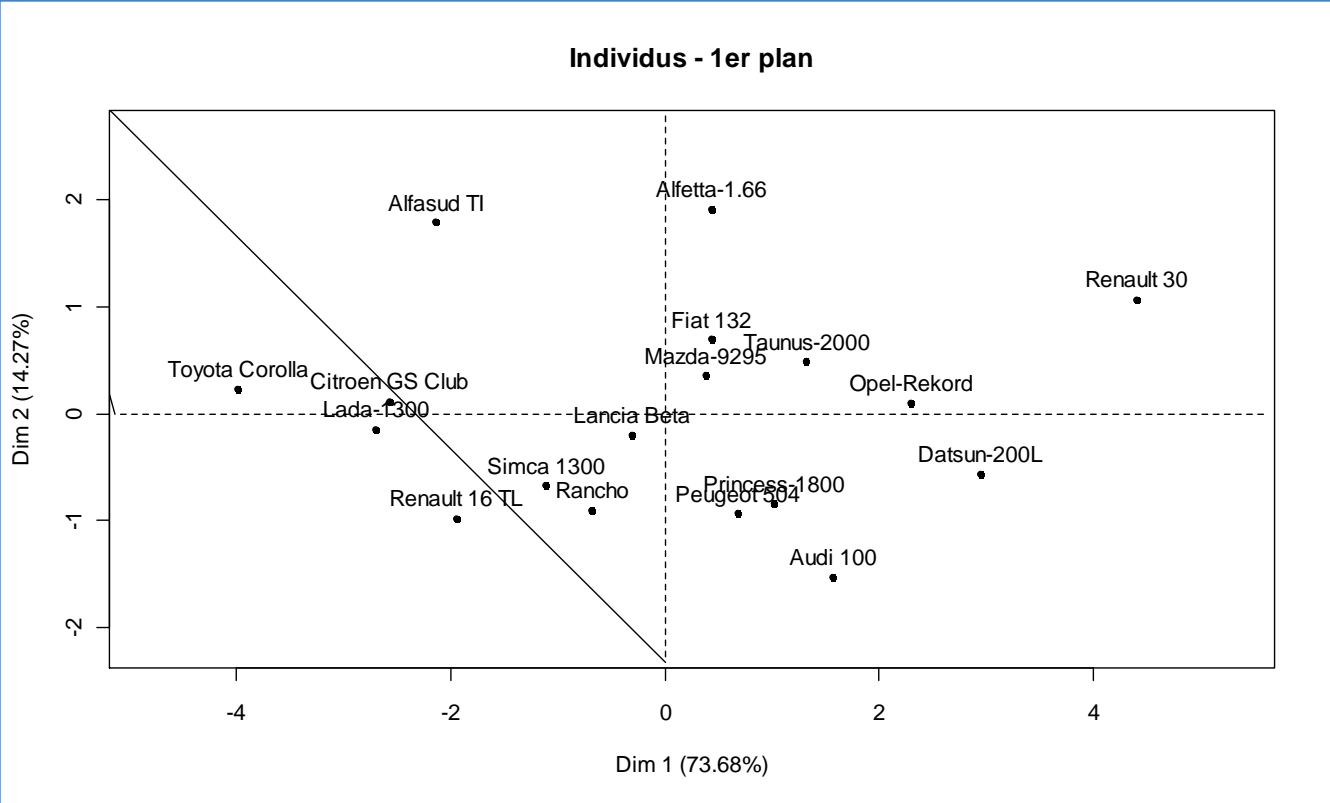
Cercle des corrélations



Carte des individus sur les 2 premiers axes

Individus actifs

```
# invisible = "quali" -> ne pas afficher les variables quali supplémentaires  
plot(autos.acp,choix="ind",title="Individus - 1er plan",invisible="quali")
```



Description automatique des axes

Identifier les variables qui influent le plus dans la définition des axes

```
#*****  
#*** Caractérisation automatique des axes - Husson, Le, Pages, pages 23 à 25  
#*****  
dimdesc(autos.acp, axes = 1:2, proba = 0.05)
```

```
> dimdesc(autos.acp, axes=1:2, proba=0.05)  
$Dim.1  
$Dim.1$quanti  
      correlation      p.value  
POIDS      0.9051875 2.429997e-07  
CYL        0.8934635 5.944907e-07  
PUISS      0.8868580 9.414656e-07  
LONG       0.8861548 9.870455e-07  
LARG       0.8135364 4.018386e-05  
PRIX       0.7724752 1.717743e-04  
V.MAX      0.7547104 2.948610e-04  
R.POID.PUIS -0.5890389 1.010705e-02  
  
$Dim.1$quali  
      R2      p.value  
FINITION 0.4024844 0.02101934  
  
$Dim.1$category  
      Estimate      p.value  
3_TB  1.516634 0.023024315  
1_M   -1.876151 0.008949921  
  
$Dim.2  
$Dim.2$quanti  
      correlation      p.value  
V.MAX      0.5735194 0.012830374  
R.POID.PUIS -0.6725451 0.002227728
```

Objectif : Identifier les variables qui pèsent le plus dans la définition des axes. Via un test de significativité de corrélation pour les variables quantitatives ; et une ANOVA à un facteur pour les variables qualitatives (couplée avec un test de Student pour chaque modalité).

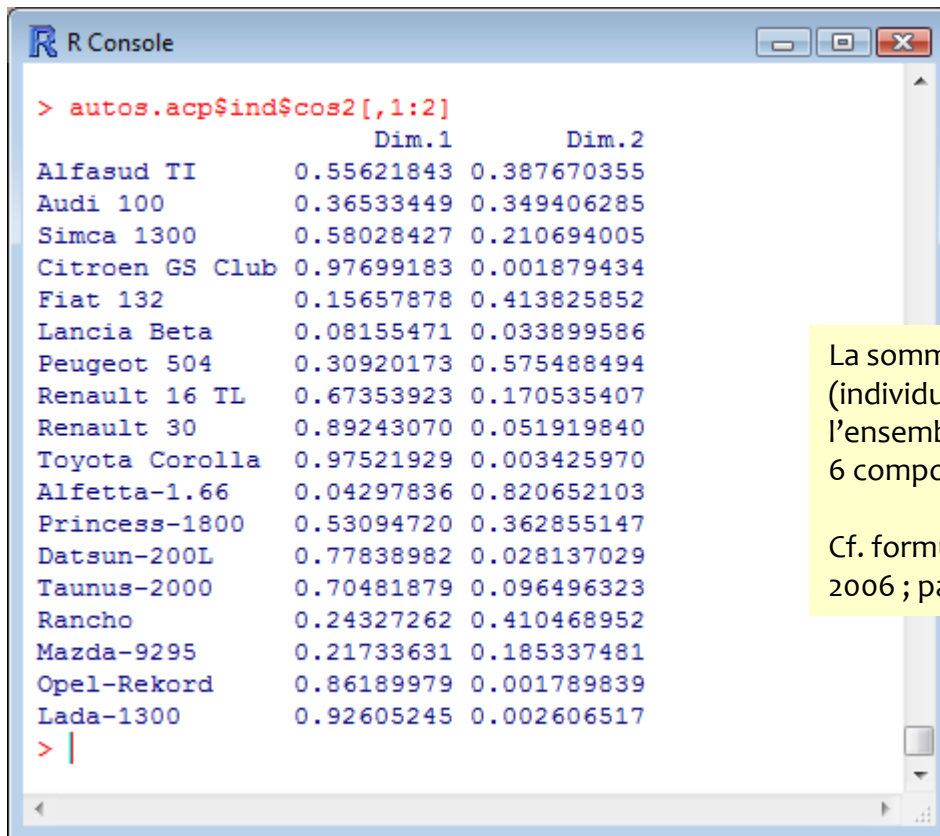
Attention, le « test » est biaisé pour les variables actives qui ont participé à la construction des axes. Les résultats sont donnés pour guider l'interprétation.

Voir F. Husson, S. Le, J. Pages, « Analyse de données avec R », PUR, 2009 ; pages 23 à 25.

COSINUS² des individus avec les composantes

Qualité de représentation des individus sur les composantes

```
#directement fournis par l'objet PCA  
#ici pour les 2 premiers axes  
autos.acp$ind$cos2[,1:2]
```



```
> autos.acp$ind$cos2[,1:2]  
          Dim.1   Dim.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  
> |
```

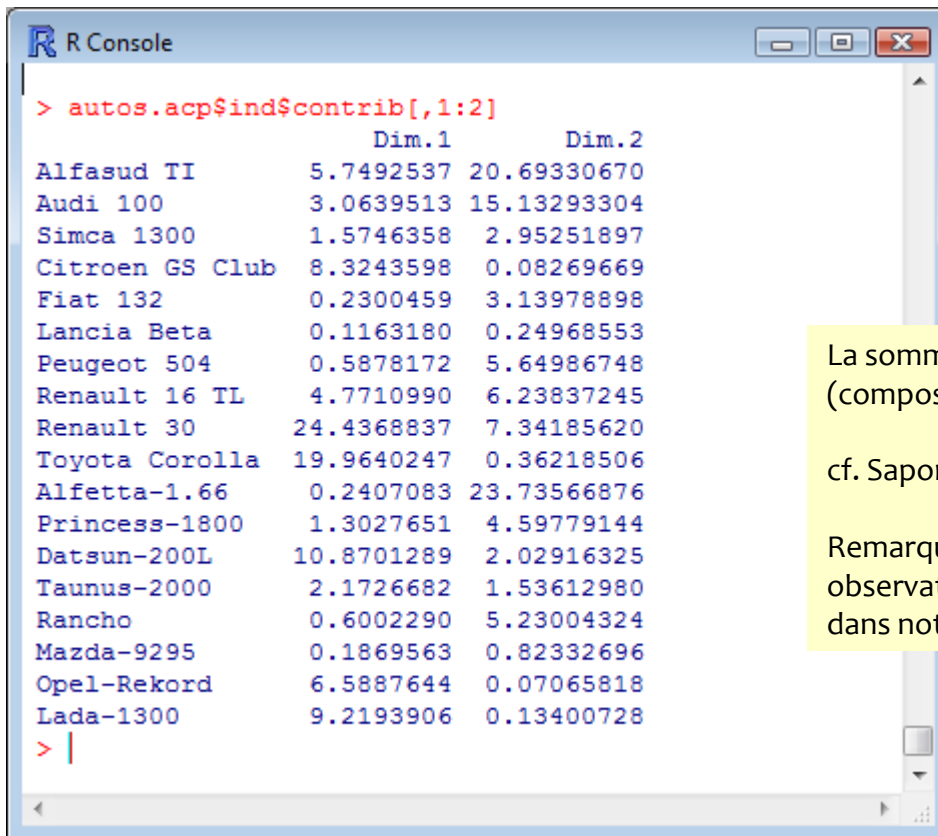
La somme pour chaque ligne (individu) vaut 1 si l'on prend l'ensemble des composantes (les 6 composantes).

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

```
#directement fournis par l'objet PCA  
#ici pour les 2 premiers axes  
autos.acp$ind$contrib[,1:2]
```



```
> autos.acp$ind$contrib[,1:2]  
          Dim.1      Dim.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.

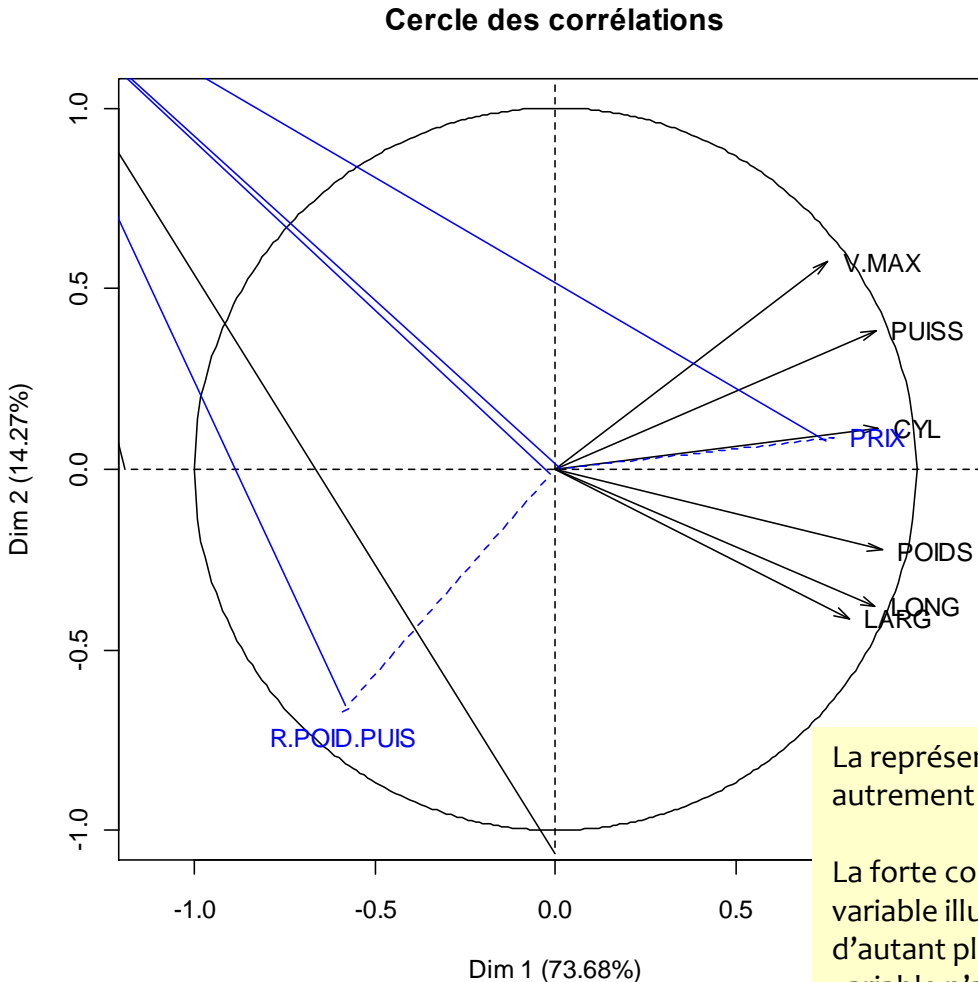
cf. Saporta, page 175.

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

Variables quantitatives illustratives

Positionnement dans le cercle des corrélations

```
*** cercle des corrélations - variables actives ET illustratives ***  
plot(autos.acp,choix="var",title="Cercle des corrélations")  
#corrélation et cosinus carré avec les deux premiers axes  
print(cbind(autos.acp$quanti.sup$cor[,1:2],autos.acp$quanti.sup$cos2[,1:2]))
```



La représentation simultanée est autrement plus instructive.

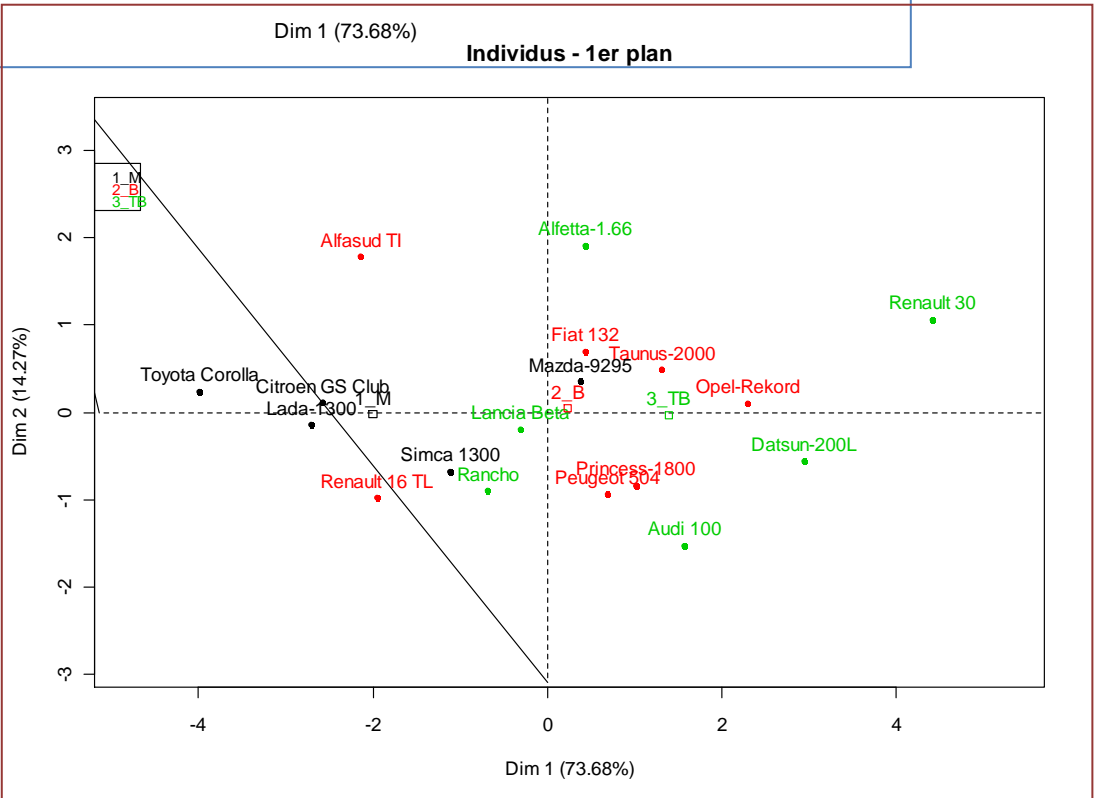
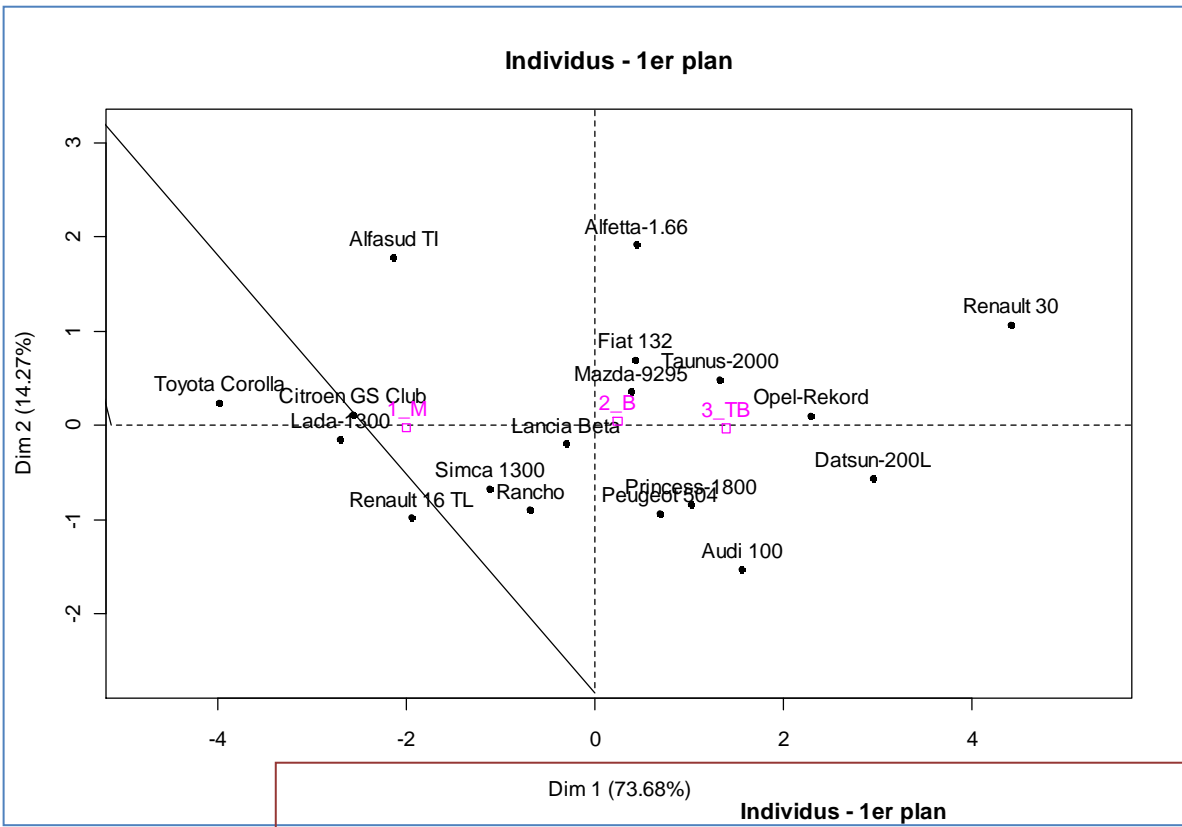
La forte corrélation entre une variable illustrative et un axe est d'autant plus intéressante que la variable n'a pas participé à la construction de l'axe.

```
R Console  
> print(cbind(autos.acp$quanti.sup$cor[,1:2],autos.acp$quanti.sup$cos2[,1:2]))  
          Dim.1      Dim.2      Dim.1      Dim.2  
PRIX      0.7724752  0.08670844  0.5967180  0.007518354  
R.POID.PUIS -0.5890389 -0.67254512  0.3469668  0.452316938  
> |
```

Variables qualitatives illustratives

Positionner les groupes associés aux modalités de la variable illustrative

```
#position dans le plan factoriel - barycentre pour chaque moda. de la var.quali
plot(autos.acp,choix="ind",title="Individus - 1er plan")
#coloriage différent des individus pour chaque moda. de la var.quali n°7
plot(autos.acp,choix="ind",title="Individus - 1er plan",habillage=7)
```



Variables qualitatives illustratives

Moyennes conditionnelles et valeurs test – Ellipses de confiance

```
#moyennes conditionnelles et valeur test
carac.quali <- cbind(autos.acp$quali.sup$coord[,1:2],autos.acp$quali.sup$vv.test[,1:2])
print(carac.quali)
#ellipse de confiance des modalités dans le plan factoriel
#créer un nouveau data.frame pour calculer les ellipse
new.autos <- cbind(as.data.frame(autos.data$FINITION),as.data.frame(autos.acp$ind$coord[,1:2]))
#calculer les ellipse de dispersion autour des barycentres dans le plan
finition.ellipse <- coord.ellipse(new.autos,bary=T)
#représentation graphique
plot(autos.acp,ellipse=finition.ellipse,habillage=7)
```

R Console

```
> carac.quali <- cbind(autos.acp$quali.sup$coord[,1:2],autos.acp$quali.sup$vv.test[,1:2])
> carac.quali
```

	Dim.1	Dim.2	Dim.1	Dim.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

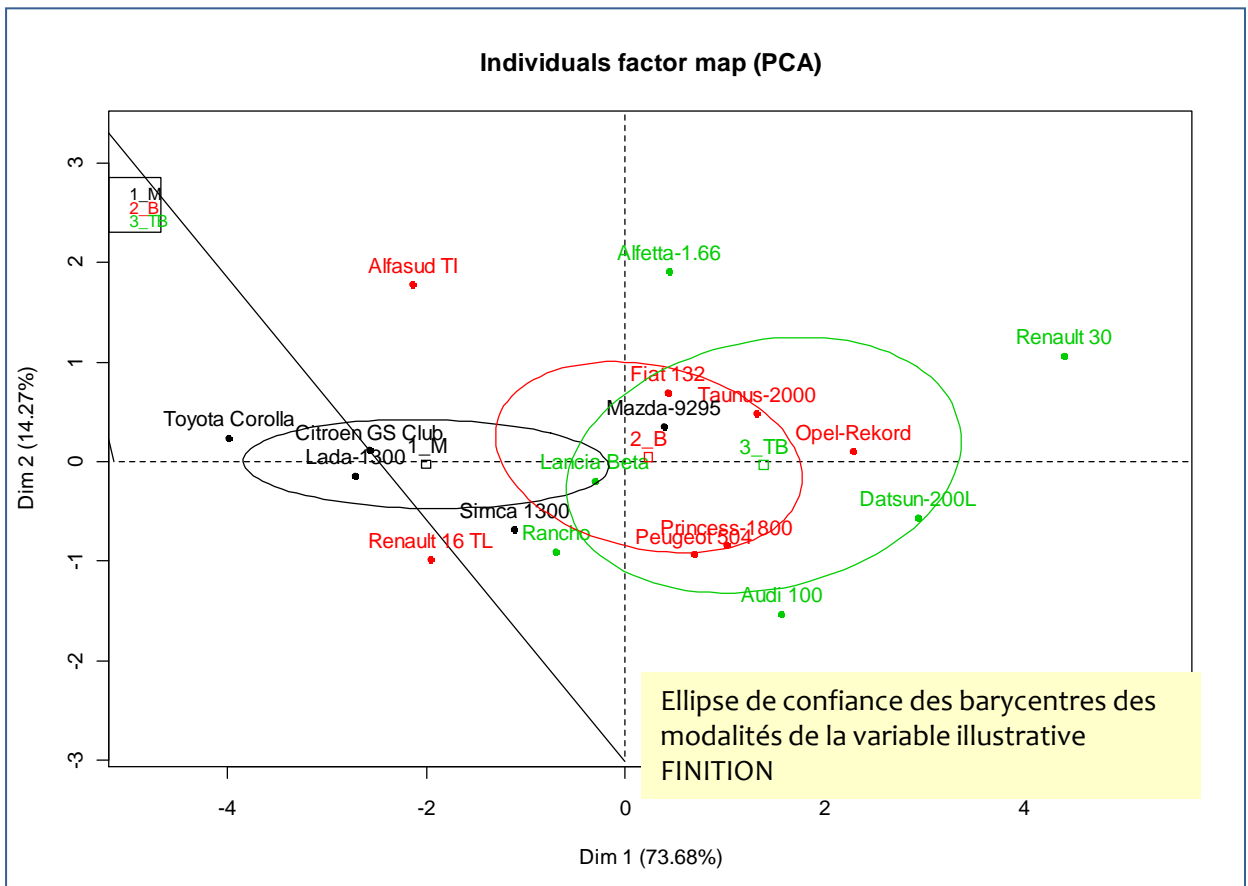
```
> |
```

	Moyennes conditionnelles	Valeurs test
1_M	-2.0003548	-2.4327167
2_B	0.2353131	0.3681035
3_TB	1.3924304	1.9307662

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.



Exploration graphique interactive

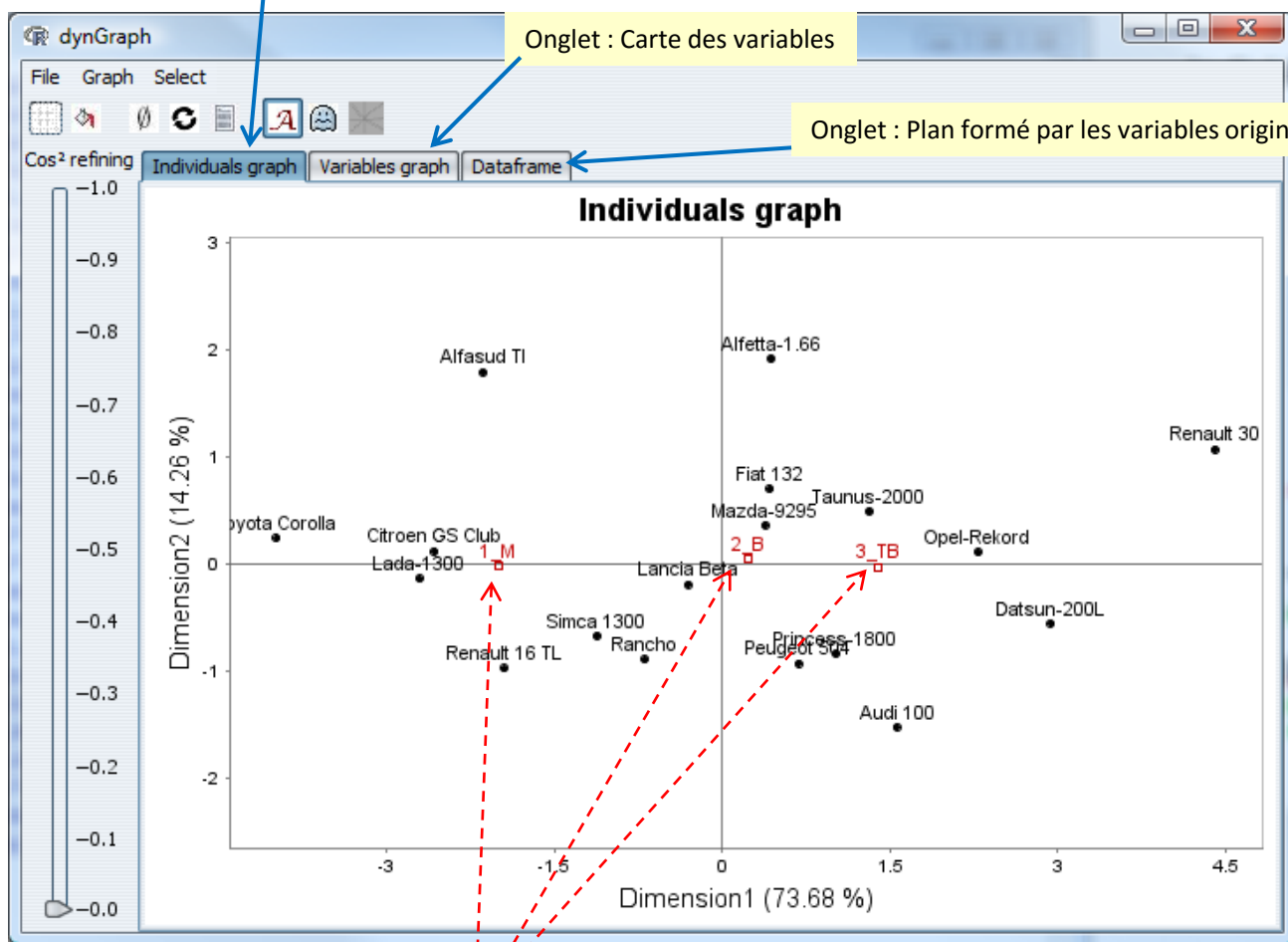
dynGraph – Démarrage et interface – Onglet « Carte des individus »

```
#*****  
# exploration graphique interactive avec dynGraph  
#*****  
  
#lancer l'outil d'exploration interactive  
dynGraph(autos.acp)
```

Onglet : Carte des individus

Onglet : Carte des variables

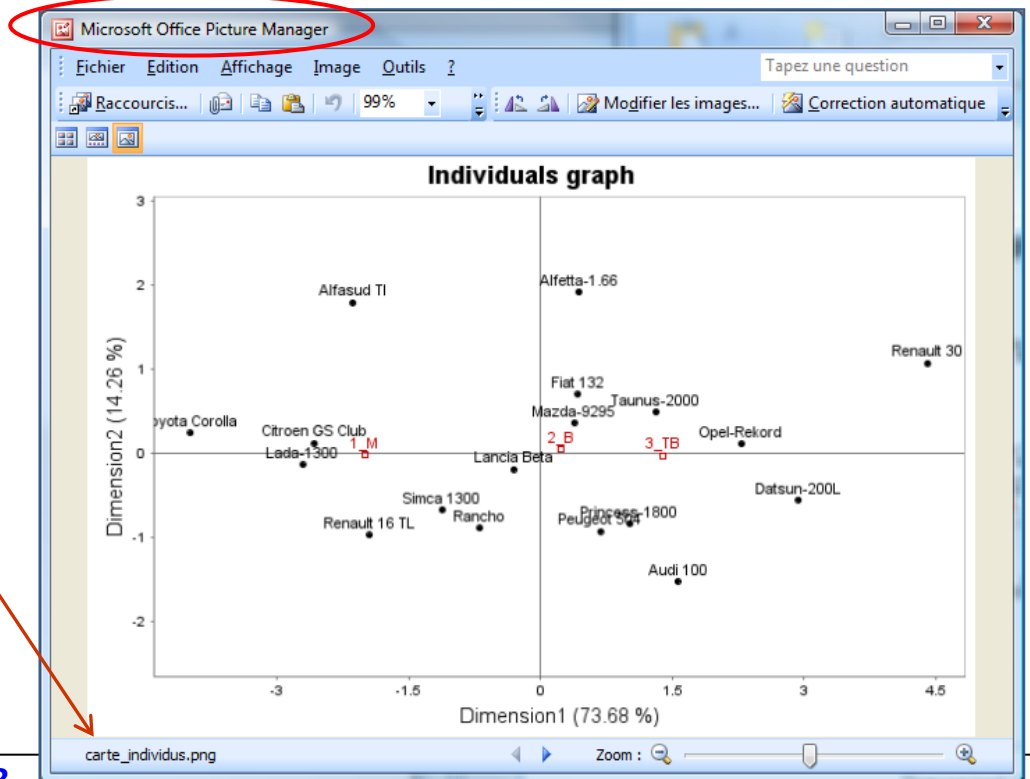
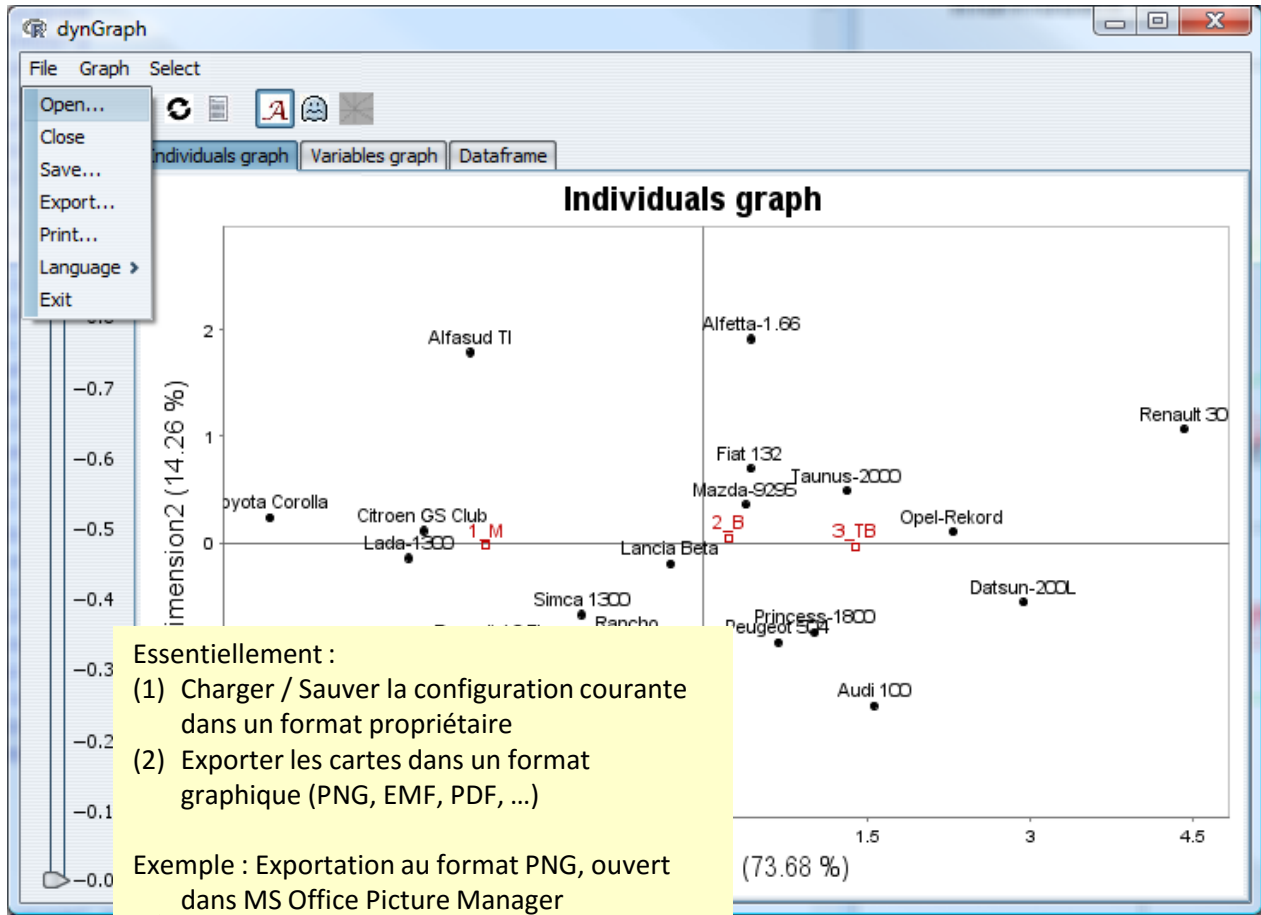
Onglet : Plan formé par les variables originelles



Les barycentres des modalités des variables illustratives sont positionnées automatiquement

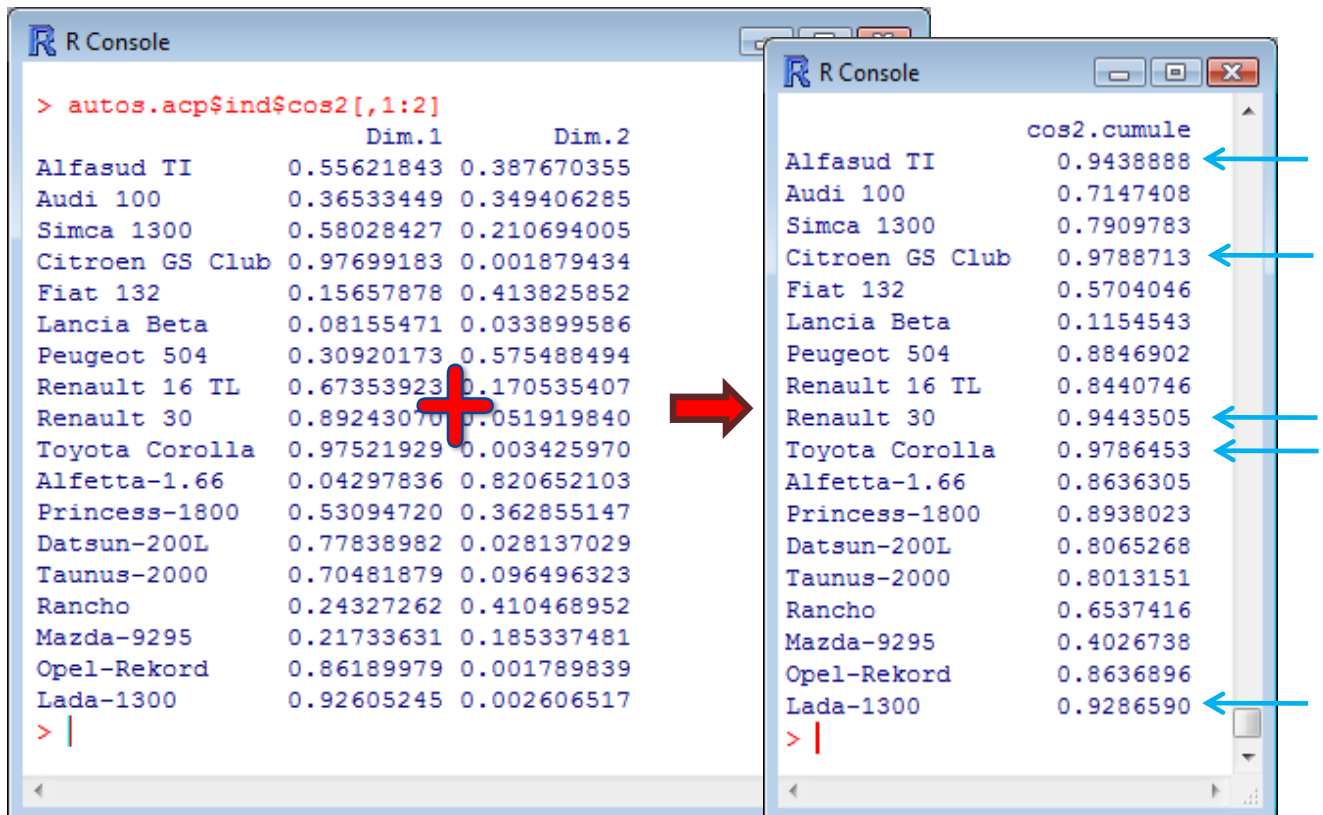
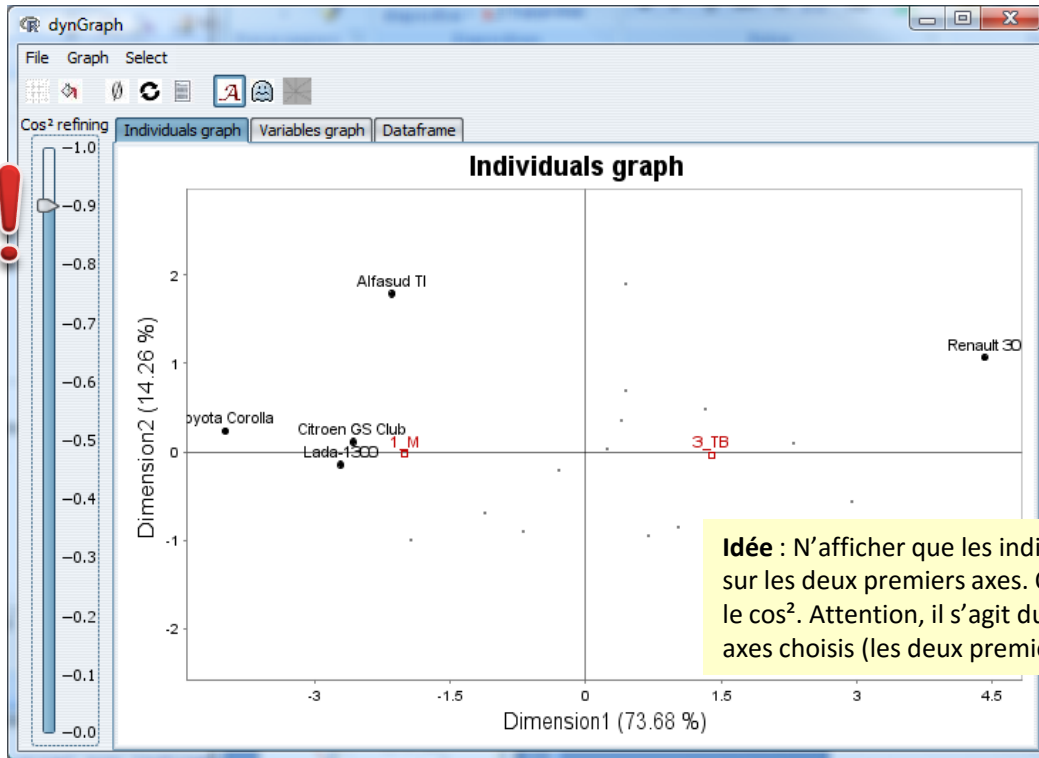
Exploration graphique interactive

dynGraph – Le menu « Fichier »

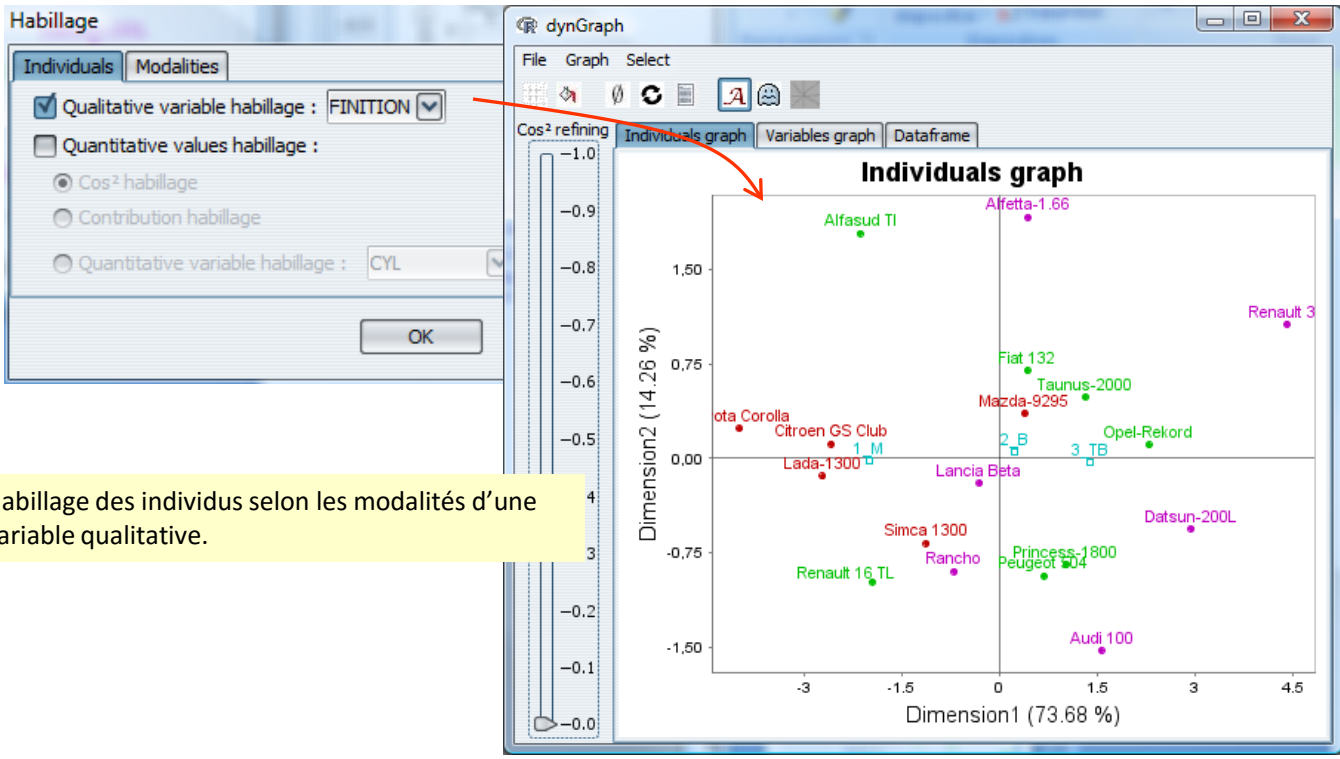


Exploration graphique interactive

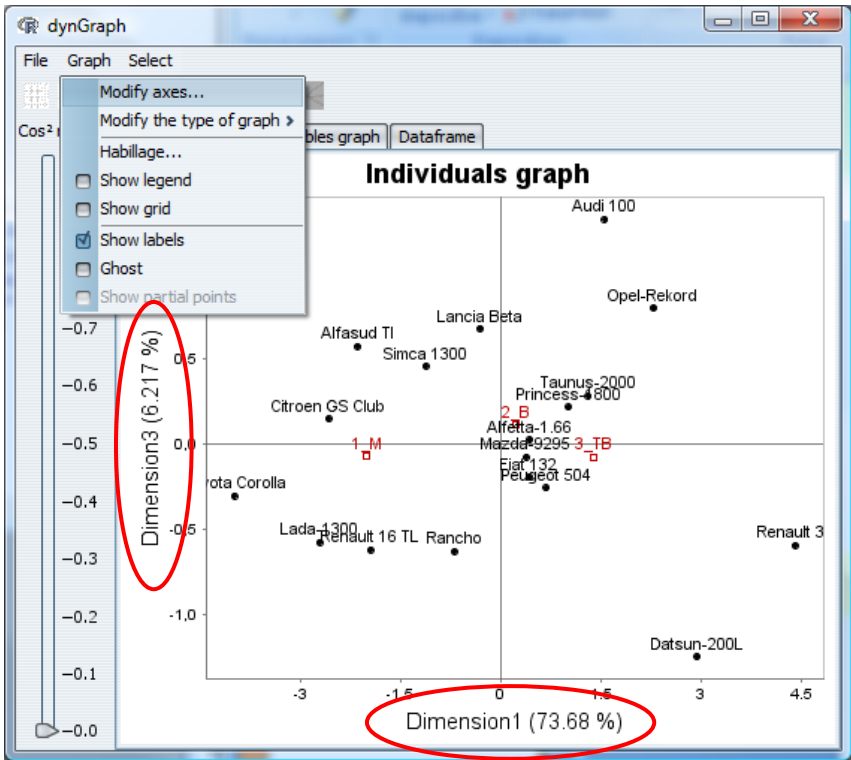
dynGraph – Filtrer les individus selon COS^2



Exploration graphique interactive dynGraph – Habillage des individus selon...



Habillage des individus selon les modalités d'une variable qualitative.

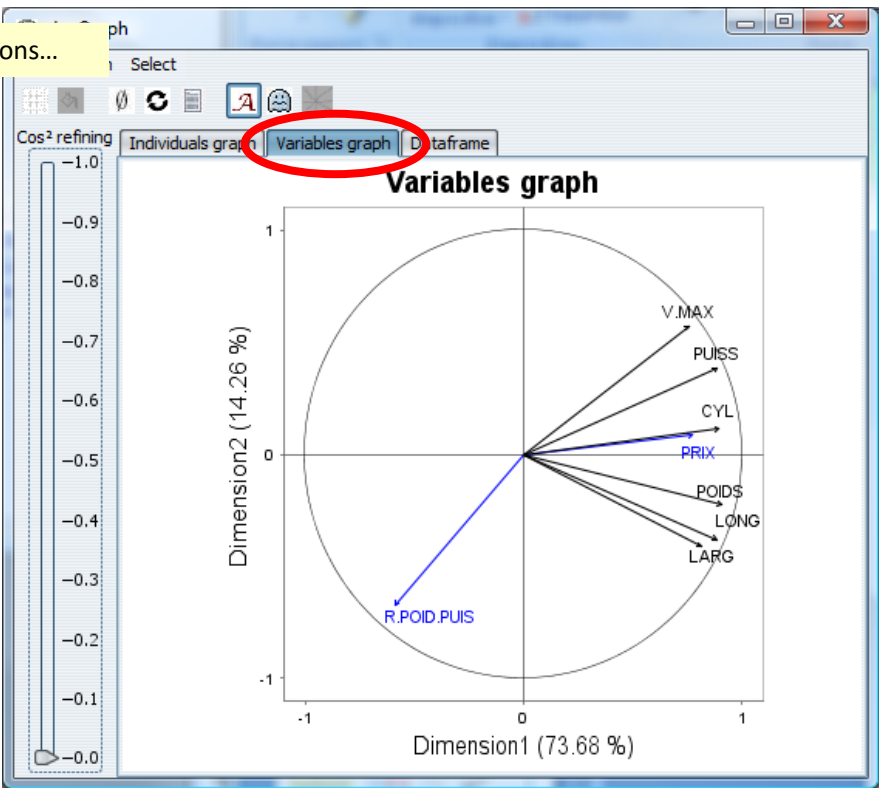


Bien entendu, nous avons la possibilité de choisir les axes factoriels.

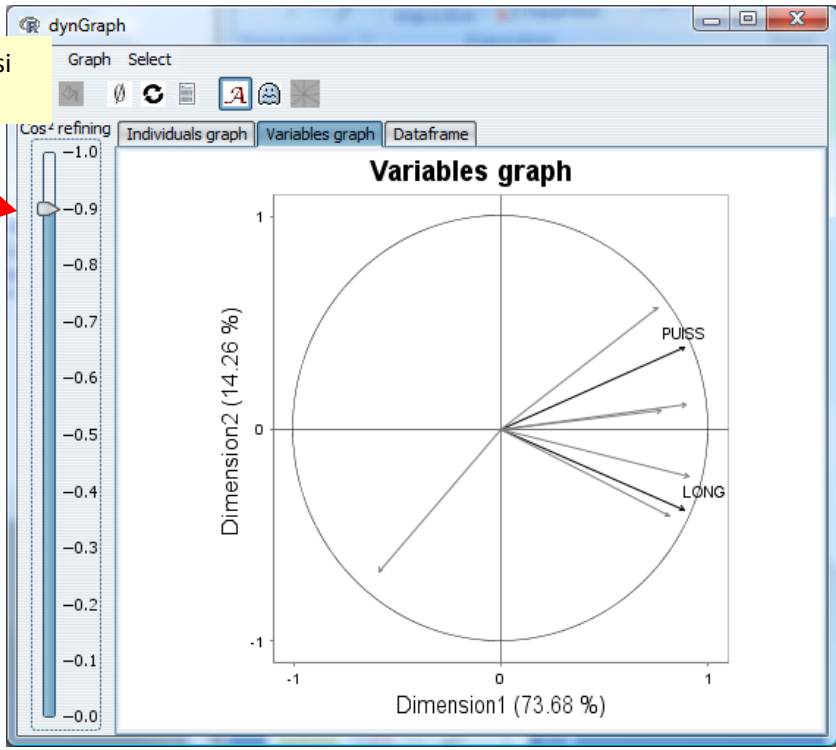
Exploration graphique interactive

dynGraph – Carte des variables

Cercle des corrélations...



Que l'on peut filtrer aussi avec le \cos^2



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