



# L'analyse des données multivariées à l'aide du logiciel

L'analyse factorielle des correspondances multiples (A.F.C.m.)

**INRA - Formation Permanente  
Juillet 2005**

André Bouchier

Copyright 2005 André Bouchier

Permission is granted to make and distribute verbatim copies of this manual provided the copyright notice and this permission notice are preserved on all copies.

Permission is granted to copy and distribute modified versions of this manual under the conditions for verbatim copying, provided that the entire resulting derived work is distributed under the terms of a permission notice identical to this one.

# 1. Avant propos :

- Pré-recquis

Ce support de cours concerne la mise en oeuvre d'une **analyse factorielle des correspondances multiples** l'aide du logiciel de statistiques R. Ce n'est ni un cours de statistique ni une initiation à R. On suppose donc connues les bases de ce logiciel ainsi que quelques notions de statistiques concernant les analyses factorielles.

- Les fonctions utilisées :

Les fonctions utilisées pour cette analyse sont disponibles dans les bibliothèques standard de R et dans la bibliothèque **ade4**. Pour aider à la compréhension, l'écriture des «programmes» sera détaillée. Par la suite, vous pourrez condenser cette écriture. Mais n'oubliez pas de les commenter abondamment.

## 2. La méthode :

- Lorsque les variables sont quantitatives et les relations linéaires, la méthode appropriée est l'analyse en composantes principales.
- Lorsque le tableau de données croise 2 variables qualitatives, c'est un tableau de contingence . La méthode appropriée est l'analyse des correspondances simples.
- Mais lorsque les données sont un mélange de variables qualitatives et quantitatives, ou contenant des relations non linéaires, il n'existe pas de méthode spécifique.
- On transforme alors les données de façon à obtenir un tableau disjonctif. Ce tableau disjonctif peut-être considéré comme un tableau de fréquence. On peut alors le soumettre à l'analyse factorielle des correspondances.
- L'analyse factorielle des correspondances multiples permet d'obtenir un résumé pertinent d'un tableau de données disjonctif. Ce résumé peut-être projeté sur un plan. On visualise ainsi les relations les plus marquantes du tableau de données.

### 3. Les données d'exemple :

- Le tableau des données d'exemple **mtcars** contient 32 observations et 11 variables :

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models)

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (lb/1000)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

## 4. Lecture des données :

- Le tableau de données est fourni avec les bibliothèques standard de R

```
mtcars
```

```
      mpg  cyl  disp  hp  drat    wt  qsec  vs  am  gear  carb
Mazda RX4      21.0   6 160.0 110  3.90  2.620 16.46  0   1     4     4
Mazda RX4 Wag  21.0   6 160.0 110  3.90  2.875 17.02  0   1     4     4
Datsun 710     22.8   4 108.0  93  3.85  2.320 18.61  1   1     4     1
Hornet 4 Drive  21.4   6 258.0 110  3.08  3.215 19.44  1   0     3     1
Hornet Sportabout 18.7   8 360.0 175  3.15  3.440 17.02  0   0     3     2
Valiant        18.1   6 225.0 105  2.76  3.460 20.22  1   0     3     1
Duster 360     14.3   8 360.0 245  3.21  3.570 15.84  0   0     3     4
Merc 240D      24.4   4 146.7  62  3.69  3.190 20.00  1   0     4     2
Merc 230       22.8   4 140.8  95  3.92  3.150 22.90  1   0     4     2
Merc 280       19.2   6 167.6 123  3.92  3.440 18.30  1   0     4     4
Merc 280C     17.8   6 167.6 123  3.92  3.440 18.90  1   0     4     4
```

.../...

## 5. Codage des données :

- Ce tableau de données contient des valeurs quantitatives et qualitatives. Il nous faut, dans un premier temps, le transformer en données uniquement qualitatives.
- Chaque variable quantitative sera découpée en 3 classes d'effectifs égaux. Pour cela, nous utiliserons la fonction `codage()` (voir annexe)

```
source("c:/RAMV/codage.R")
mpg<-codage(mtcars$mpg)
disp<-codage(mtcars$disp)
hp<-codage(mtcars$hp)
drat<-codage(mtcars$drat)
wt<-codage(mtcars$wt)
qsec<-codage(mtcars$qsec)
```

- Les variables qualitatives (ou non modifiées) seront transformées en facteurs

```
cyl<-as.factor(mtcars$cyl)
am<-as.factor(mtcars$am)
gear<-as.factor(mtcars$gear)
vs<-as.factor(mtcars$vs)
carb<-as.factor(mtcars$carb)
```

## 6. Mise en forme des données codées :

- Le tableau des données codées :

```
don<-data.frame(mpg,cyl,disp,hp,drat,wt,qsec,vs,am,gear,carb)
row.names(don)<-row.names(mtcars)
don
```

	<i>mpg</i>	<i>cyl</i>	<i>disp</i>	<i>hp</i>	<i>drat</i>	<i>wt</i>	<i>qsec</i>	<i>vs</i>	<i>am</i>	<i>gear</i>	<i>carb</i>
<i>Mazda RX4</i>	2	6	2	2	2	1	1	0	1	4	4
<i>Mazda RX4 Wag</i>	2	6	2	2	2	2	1	0	1	4	4
<i>Datsun 710</i>	3	4	1	1	2	1	3	1	1	4	1
<i>Hornet 4 Drive</i>	2	6	2	2	1	2	3	1	0	3	1
<i>Hornet Sportabout</i>	2	8	3	2	1	2	1	0	0	3	2
<i>Valiant</i>	2	6	2	1	1	2	3	1	0	3	1
<i>Duster 360</i>	1	8	3	3	2	3	1	0	0	3	4
<i>Merc 240D</i>	3	4	2	1	2	2	3	1	0	4	2
<i>Merc 230</i>	3	4	1	1	3	2	3	1	0	4	2
					.../...						



## 7. Vérification du codage des données :

summary (don)

<i>mpg</i>	<i>cyl</i>	<i>disp</i>	<i>hp</i>	<i>drat</i>	<i>wt</i>	<i>qsec</i>	<i>vs</i>	<i>am</i>	<i>gear</i>	<i>carb</i>
1:11	4:11	1:11	1:11	1:11	1:11	1:11	0:18	0:19	3:15	1: 7
2:11	6: 7	2:10	2:11	2:11	2:10	2:10	1:14	1:13	4:12	2:10
3:10	8:14	3:11	3:10	3:10	3:11	3:11			5: 5	3: 3
										4:10
										6: 1
										8: 1

- Pour être pertinent, un découpage en classes doit respecter 3 principes :
  - 1) pas d'effectifs de classes trop déséquilibrés
  - 2) Des nombres de classes semblables pour toutes les variables
  - 3) Des découpages ayant une signification pour le chercheur
- La variable **carb** a des classes d'effectifs trop déséquilibrés. Cela rendra l'interprétation des résultats plus difficile.

## 8. Ajustement du codage :

- La variable `carb` a des classes d'effectifs trop déséquilibrés. Nous allons regrouper des modalités afin d'obtenir 3 classes :

- **classe 1:** 1 seul carburateur, 7 individus
- **classe 2:** 2 ou 3 carburateurs, 13 individus
- **classe 4:** plus de 3 carburateurs, 12 individus

```
don$carb<-as.numeric(don$carb)
don$carb[don$carb==3]<-2
don$carb[don$carb>3]<-4
don$carb<-as.factor(don$carb)
```

- On vérifie le codage :

```
summary(as.factor(don$carb))
```

```
1  2  4
7 13 12
```

## 9. Transformation des données en tableau disjonctif :

- Utilisation de la library `ade4` :

```
library(ade4)
```

- Création du tableau disjonctif :

```
disj<-acm.disjonctif (don)
```

- Un extrait du tableau disjonctif :

	<i>mpg.1</i>	<i>mpg.2</i>	<i>mpg.3</i>	<i>cyl.4</i>	<i>cyl.6</i>	<i>cyl.8</i>	<i>disp.1</i>	<i>disp.2</i>	<i>disp.3</i>
<i>Mazda RX4</i>	0	1	0	0	1	0	0	1	0
<i>Mazda RX4 Wag</i>	0	1	0	0	1	0	0	1	0
<i>Datsun 710</i>	0	0	1	1	0	0	1	0	0
<i>Hornet 4 Drive</i>	0	1	0	0	1	0	0	1	0
<i>Hornet Sportabout</i>	0	1	0	0	0	1	0	0	1
<i>Valiant</i>	0	1	0	0	1	0	0	1	0
<i>Duster 360</i>	1	0	0	0	0	1	0	0	1
<i>Merc 240D</i>	0	0	1	1	0	0	0	1	0
<i>Merc 230</i>	0	0	1	1	0	0	1	0	0
<i>Merc 280</i>	0	1	0	0	1	0	0	1	0
<i>Merc 280C</i>	0	1	0	0	1	0	0	1	0
<i>Merc 450SE</i>	1	0	0	0	0	1	0	1	0
				....	/	....			

## 10.L'A.F.C.m. - fonction `dudi.coa()` de la bibliothèque `ade4` :

- Les résultats de l'AFC sont stockés dans la variable `z`

```
z<-dudi.coa(df = disj, scannf = FALSE, nf = 3)
```

- L'éboulis des valeurs propres

```
inertie<-z$eig/sum(z$eig)*100  
barplot(inertie,ylab="% d'inertie",names.arg=round(inertie,2))  
title("Eboulis des valeurs propres en %")
```

- Les valeurs propres

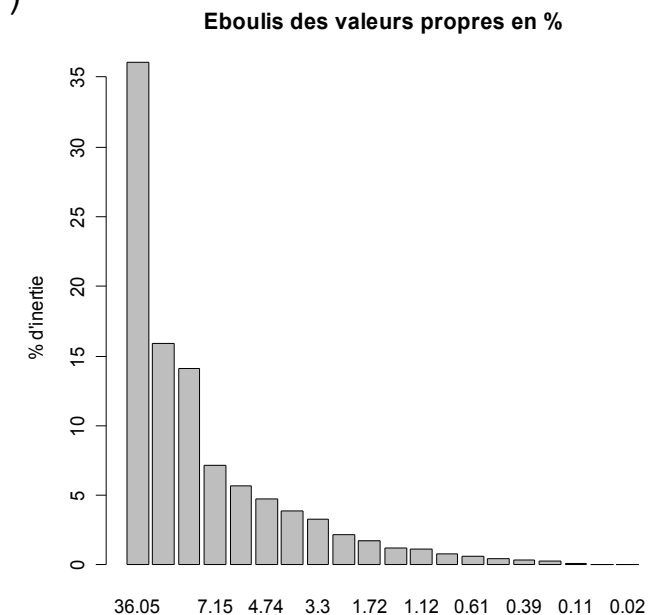
```
round(z$eig,3)
```

```
[1] 0.655 0.290 0.256 0.130 0.104
```

- Les valeurs propres en %

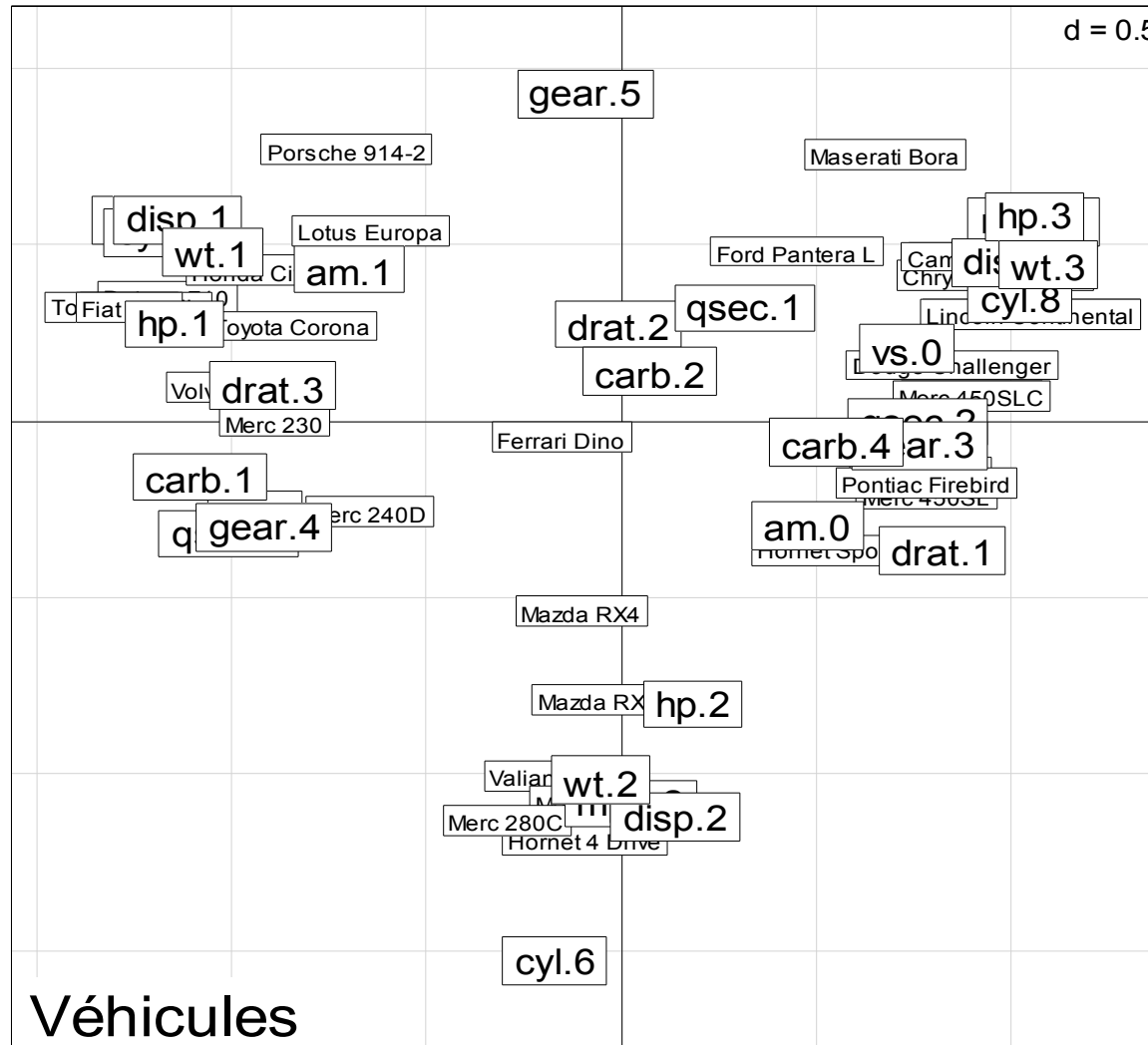
```
round(z$eig/sum(z$eig)*100,2)
```

```
[1] 36.05 15.95 14.10 7.15 5.72
```



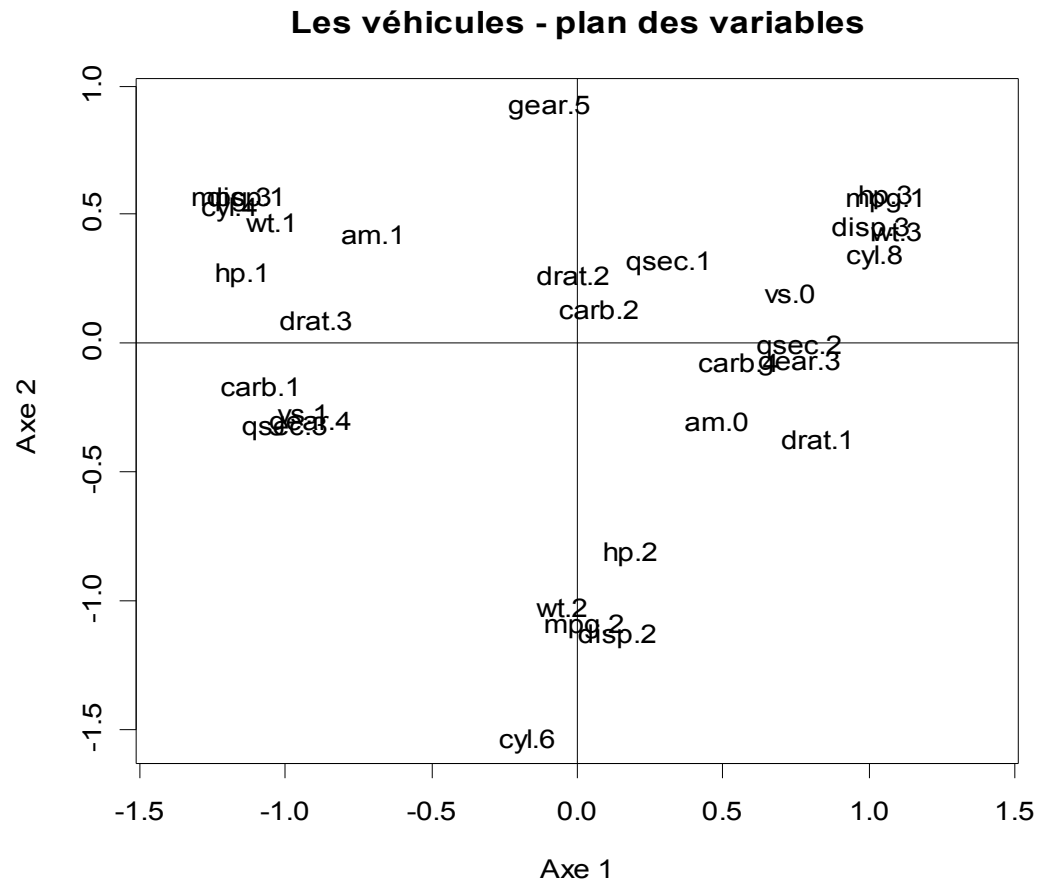
# 11. Une représentation graphique du plan factoriel :

```
scatter.coa(z, method = 1, sub = "Véhicules", posieig = "none")
```



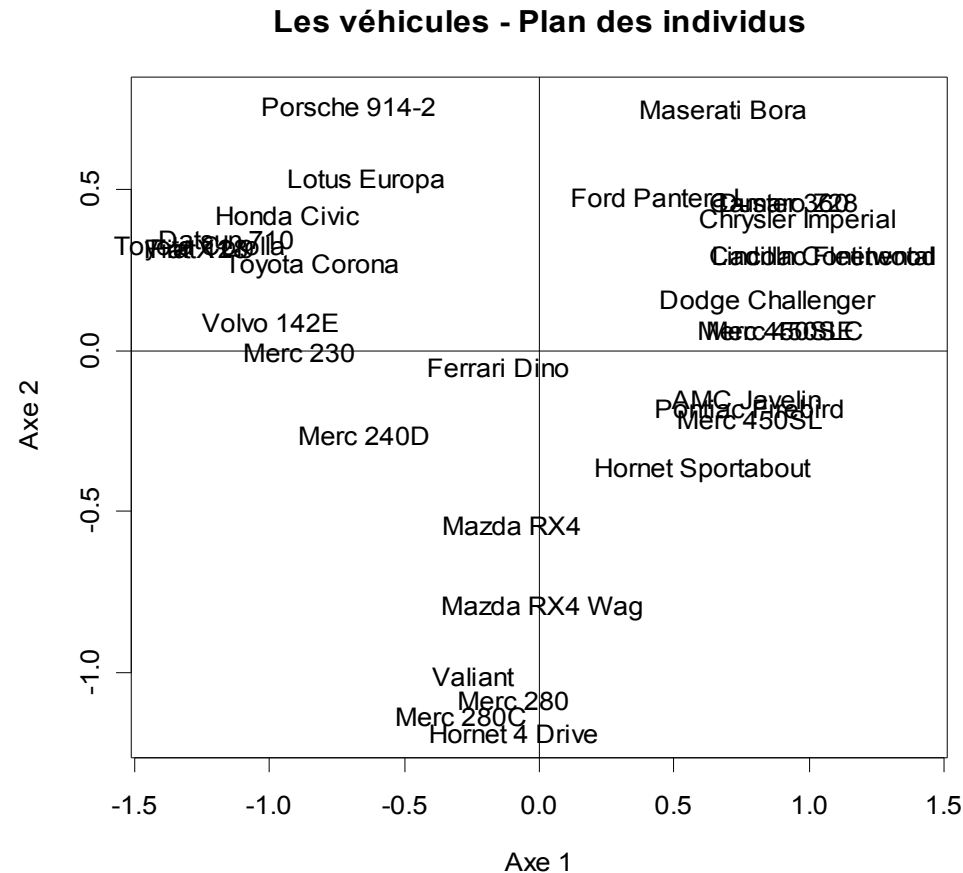
## 12. Une autre représentation du plan factoriel, les variables :

```
plot(z$co[,1],z$co[,2],type="n",xlab="Axe 1",ylab="Axe 2",xlim=c(-1.4,1.4))
text(z$co[,1], z$co[,2], label= colnames(disj))
title("Les véhicules - plan des variables")
abline(h=0,v=0)
```



### 13. Une autre représentation du plan factoriel, les individus :

```
plot(z$li[,1],z$li[,2],type="n",xlab="Axe 1",ylab="Axe 2",xlim=c(-1.4,1.4))  
text(z$li[,1], z$li[,2], label=row.names(disj))  
title("Les véhicules - Plan des individus")  
abline(h=0,v=0)
```



## 14. Interprétation des facteurs : contributions "relatives" des variables :

- Contributions des variables à la construction des axes :

```
inertia.dudi(z, col.inertia = T)$col.abs
```

	<i>Comp1</i>	<i>Comp2</i>	<i>Comp3</i>
<i>mpg.1</i>	533	345	12
<i>mpg.2</i>	0	1267	171
<i>mpg.3</i>	611	319	101
<i>cyl.4</i>	677	308	108
<i>cyl.6</i>	9	1608	582
<i>cyl.8</i>	633	163	62
<i>disp.1</i>	615	351	0
<i>disp.2</i>	8	1234	3
<i>disp.3</i>	487	218	3
	.../...		
<i>gear.3</i>	379	7	466
<i>gear.4</i>	436	106	1
<i>gear.5</i>	2	422	1273
<i>carb.1</i>	355	17	269
<i>carb.2</i>	3	25	256
<i>carb.4</i>	159	4	852
<b>Somme</b>	<b>10000</b>	<b>10000</b>	<b>10000</b>



## 15. Interprétation des facteurs : contributions "relatives" des individus :

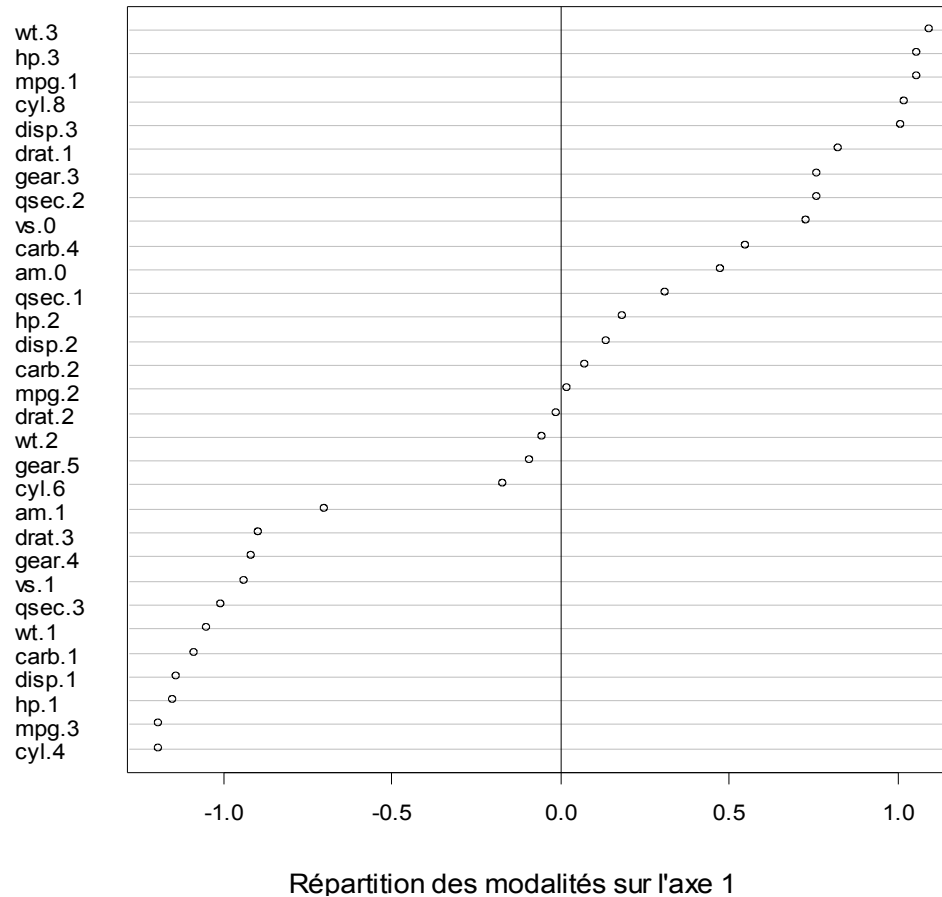
- Contributions des lignes à la construction des axes :

```
inertia.dudi(z, row.inertia = T)$row.abs
```

	<i>Axis1</i>	<i>Axis2</i>	<i>Axis3</i>
<i>Mazda RX4</i>	5	312	1245
<i>Mazda RX4 Wag</i>	0	674	1151
<i>Datsun 710</i>	645	135	25
<i>Hornet 4 Drive</i>	4	1516	65
<i>Hornet Sportabout</i>	173	145	0
<i>Valiant</i>	28	1087	219
<i>Duster 360</i>	392	234	58
<i>Merc 240D</i>	200	71	163
<i>Merc 230</i>	377	0	296
.../...			
<i>Fiat X1-9</i>	760	113	96
<i>Porsche 914-2</i>	237	638	202
<i>Lotus Europa</i>	197	311	492
<i>Ford Pantera L</i>	97	249	629
<i>Ferrari Dino</i>	11	2	2018
<i>Maserati Bora</i>	219	618	759
<i>Volvo 142E</i>	474	10	20
<b>Somme</b>	<b>10000</b>	<b>10000</b>	<b>10000</b>

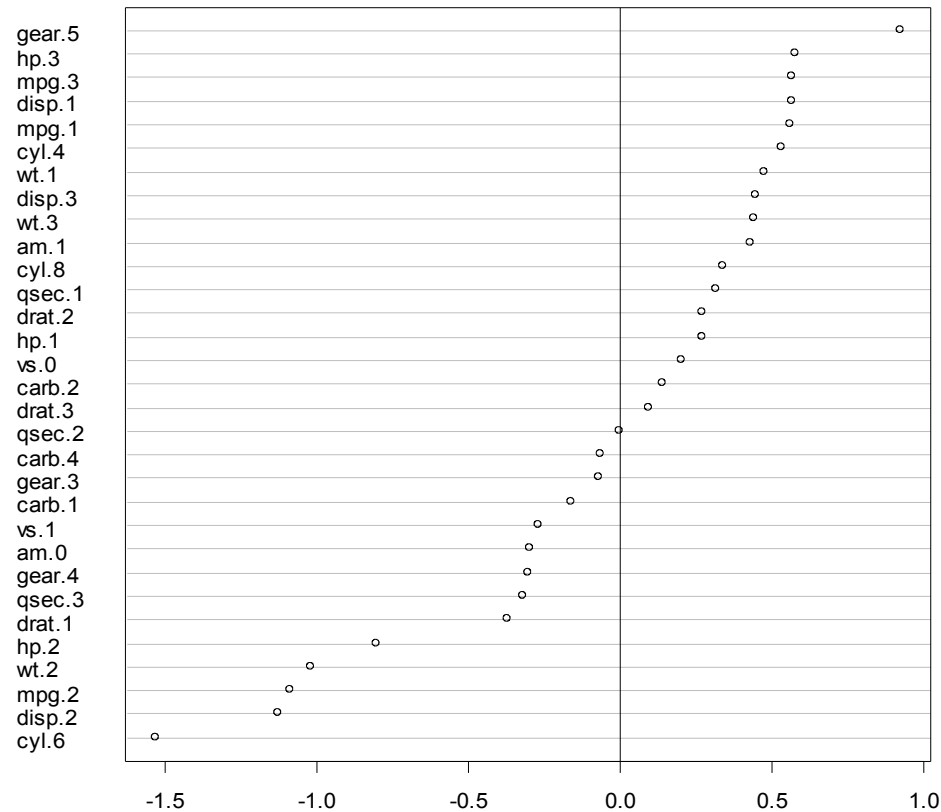
## 16.A.F.C. multiple, une aides à l'interprétation - axe 1 :

```
modal<-as.data.frame(z$co)
modal<-modal[sort.list(modal$Comp1),]
dotchart(modal[,1],labels = row.names(modal),cex=0.8)
title(sub="Répartition des modalités sur l'axe 1")
abline(v=0)
```



## 17.A.F.C. multiple, une aides à l'interprétation - axe 2 :

```
modal<-as.data.frame(z$co)
modal<-modal[sort.list(modal$Comp2),]
dotchart(modal[,2],labels = row.names(modal),cex=0.8)
title(sub="Répartition des modalités sur l'axe 2")
abline(v=0)
```



Répartition des modalités sur l'axe 2

## 18.L'A.F.C.s, le programme complet : préparation des données

```
#les données
library(datasets)
data(mtcars)
#codage des données
source("c:/RAMV/codage.R")
mpg<-codage(mtcars$mpg)
disp<-codage(mtcars$disp)
hp<-codage(mtcars$hp)
drat<-codage(mtcars$drat)
wt<-codage(mtcars$wt)
qsec<-codage(mtcars$qsec)
cyl<-as.factor(mtcars$cyl)
am<-as.factor(mtcars$am)
gear<-as.factor(mtcars$gear)
vs<-as.factor(mtcars$vs)
carb<-as.factor(mtcars$carb)
#création du tableau des données codées
don<-data.frame(mpg,cyl,disp,hp,drat,wt,qsec,vs,am,gear,carb)
row.names(don)<-row.names(mtcars)
#modification d'un codage
don$carb<-as.numeric(don$carb)
don$carb[don$carb==3]<-2
don$carb[don$carb>3]<-4
don$carb<-as.factor(don$carb)
#transformation en tableau disjonctif
library(ade4)
disj<-acm.disjonctif (don)
```

## 19.L'A.F.C.s, le programme complet : les plans factoriels

```
#l'AFC multiple
z<-dudi.coa(df = disj, scannf = FALSE, nf = 3)
#L'éboulis des valeurs propres
inertie<-z$eig/sum(z$eig)*100
barplot(inertie,ylab="% d'inertie",names.arg=round(inertie,2))
title("Eboulis des valeurs propres en %")
#Les valeurs propres en %
round(z$eig/sum(z$eig)*100,2)
#une représentation du plan factoriel
scatter.coa(z, method = 1, sub = "Véhicules", posieig = "none")
#une autre représentation du plan factoriel - les variables
plot(z$co[,1],z$co[,2],type="n",xlab="Axe 1",
      ylab="Axe 2",xlim=c(-1.4,1.4))
text(z$co[,1], z$co[,2], label= colnames(disj))
title("Les véhicules - plan des variables")
abline(h=0,v=0)
#une autre représentation du plan factoriel - les individus
plot(z$li[,1],z$li[,2],type="n",xlab="Axe 1",
      ylab="Axe 2",xlim=c(-1.4,1.4))
text(z$li[,1], z$li[,2], label=row.names(disj))
title("Les véhicules - Plan des individus")
abline(h=0,v=0)
```

## 20.L'A.F.C.s, le programme complet : aides à l'interprétation

```
#les contributions relatives des variables
inertia.dudi(z,col.inertia = T)$col.abs
#les contributions relatives des individus
inertia.dudi(z,row.inertia = T)$row.abs
#aide graphique à l'interprétation de l'axes 1
modal<-as.data.frame(z$co)
modal<-modal[sort.list(modal$Comp1),]
dotchart(modal[,1],labels = row.names(modal),cex=0.8)
title(sub="Répartition des modalités sur l'axe 1")
abline(v=0)
#aide graphique à l'interprétation de l'axes 2
modal<-as.data.frame(z$co)
modal<-modal[sort.list(modal$Comp2),]
dotchart(modal[,2],labels = row.names(modal),cex=0.8)
title(sub="Répartition des modalités sur l'axe 2")
abline(v=0)
```

## 21. Quelques références :

- **Logiciel R :**

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL

<http://www.R-project.org>.

- **Bibliothèque ade4 :**

Jean Thioulouse, Anne-Beatrice Dufour and Daniel Chessel (2004). ade4: Analysis of Environmental Data : Exploratory and Euclidean methods in Environmental sciences. R package version 1.3-3.

<http://pbil.univ-lyon1.fr/ADE-4>

Mailing list: <http://pbil.univ-lyon1.fr/ADE-4/adelist.html>

## 22. Annexes : la fonction codage()

```
codage<-function(nom)
#découpage en 3 classes d'effectifs égaux
{
  #calcul des bornes
  bornes<-quantile(nom, probs = c(0, 1/3,2/3,1), na.rm = TRUE,names = TRUE)
  #description des bornes et effectifs
  Amax<-aggregate(nom,list(Nom=cut(nom,bornes,include.lowest=T,label=F)),max)
  Amin<-aggregate(nom,list(Nom=cut(nom,bornes,include.lowest=T,label=F)),min)
  Afreq<-as.matrix(summary(as.factor(cut(na.omit(nom),bornes, include.lowest=T,
label=F))))
  limites<-as.data.frame(cbind(Amin[,1],Amin[,2],Amax[,2],Afreq))
  names(limites)<-c("Classe","Mini","Maxi","Effectif")
  #calcul du nombre de valeurs manquantes
  manques<-length(nom)-length(na.omit(nom))
  #impression des bornes
  cat(paste("Découpage de la variable ",deparse(substitute(nom))," - Nb de
valeurs manquantes : ", manques, "\n"))
  print(limites)
  #découpage de la variable
  varfac<-cut(nom,bornes,include.lowest=T,label=F)
  #transformation en facteur
  as.factor(varfac)
}
```



## 23. Table des matières

1. Avant propos :	3
2. La méthode :	4
3. Les données d'exemple :	5
4. Lecture des données :	6
5. Codage des données :	7
6. Mise en forme des données codées :	8
7. Vérification du codage des données :	9
8. Ajustement du codage :	10
9. Transformation des données en tableau disjonctif :	11
10. L'A.F.C.m. - fonction dudi.coa () de la bibliothèque ade4 :	12
11. Une représentation graphique du plan factoriel :	13
12. Une autre représentation du plan factoriel, les variables :	14
13. Une autre représentation du plan factoriel, les individus :	15
14. Interprétation des facteurs : contributions "relatives" des variables :	16
15. Interprétation des facteurs : contributions "relatives" des individus :	17
16. A.F.C. multiple, une aides à l'interprétation - axe 1 :	18
17. A.F.C. multiple, une aides à l'interprétation - axe 2 :	19
18. L'A.F.C.s, le programme complet : préparation des données.....	20
19. L'A.F.C.s, le programme complet : les plans factoriels.....	21
20. L'A.F.C.s, le programme complet : aides à l'interprétation.....	22
21. Quelques références :	23
22. Annexes : la fonction codage().	24