Méthodes d'analyse des données

exemples traités avec Splus

Didier Guerniou
Patrick Gasqui
Claire Chabanet

INRA Biométrie — Jouy-en-Josas
October 12, 1995

SOMMAIRE

Ce document présente la façon de mettre en œuvre avec Splus les méthodes les plus classiques d'analyse des données, et la façon d'interpréter les résultats, à partir d'exemples. Il ne remplace pas une formation, mais peut être utilisé en complément d'un cours. Les chapîtres d'analyse en composantes principales et d'analyse des correspondances sont indépendants l'un de l'autre, bien que certains principes généraux d'interprétation, valables aussi bien en analyse en composantes principales qu'en analyse des correspondances, ne soient pas détaillés à nouveau dans le cadre de l'analyse des correspondances.

- Analyse en composantes principales
- Analyse des correspondances
  simple
  multiple

- Classification

- ANNEXES
  I Valeurs propres et vecteurs propres : illustration graphique

  II Les sorties numériques de la fonction "acp" : coord, cta, ctr

  III Les tableaux en codage disjonctif complet

  IV Les fonctions spécifiques

- Bibliographie

Analyse en Composantes Principales

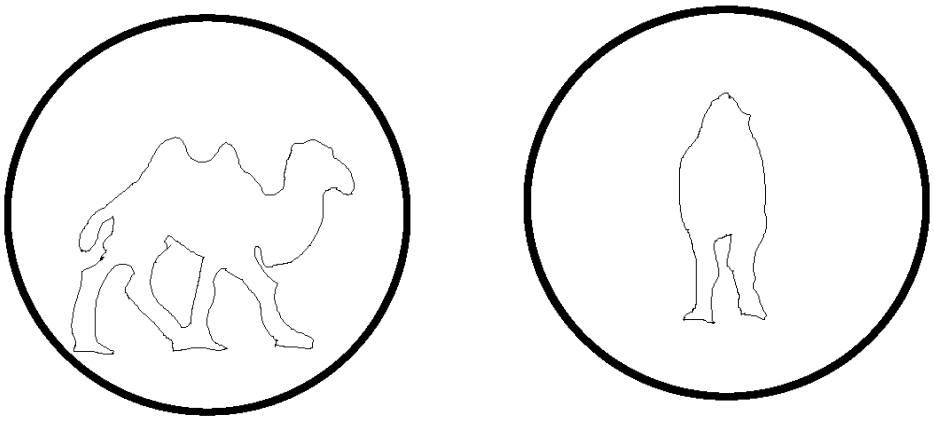
Le principe

L'Analyse en composantes principales est une méthode statistique descriptive : l'objectif est de représenter sous forme graphique le maximum de l'information contenue dans un tableau de données. Cette méthode s'applique à des **tableaux de données quantitatives**. Elle est intéressante dès que la taille du jeu de données n'en permet pas une interprétation directe. Elle est donc adaptée au dépouillement de grands jeux de données. Elle permet aussi une présentation synthétique de jeux de données, petits ou grands.

Le but est de **réduire l'espace de représentation** des données pour une meilleure compréhension : si p est le nombre de variables, chaque individu peut être représenté par un point dans un espace de dimension p, les coordonnées étant les valeurs observées. Pour réduire l'espace, on construit des variables synthétiques résumant l'information : elles sont des **combinaisons linéaires des variables initiales** et forment les axes du nouvel espace de représentation, la nouvelle origine étant le centre de gravité du nuage.

Les **axes factoriels** sont ordonnés, le premier axe étant celui qui véhicule la plus grande partie de l'information, c'est l'axe associé au pourcentage d'inertie le plus élevé, le dernier axe est l'axe qui présente la plus faible variabilité (il s'agit de la variabilité des projections des individus sur cet axe). Les premiers axes factoriels sont retenus, ce sont ceux qui expliquent la plus grande part de l'information, ceux qui déforment le moins possible le nuage initial. Les autres axes factoriels sont négligés, ils sont considérés comme "du bruit".

La projection choisie est celle qui restitue la plus grande proportion de l'information. Voici deux projections du même animal, la première permet de le reconnaitre sans ambiguité, c'est celle qui restitue le maximum d'inertie.



Le premier axe est l'axe sur lequel l'inertie de la projection est la plus grande. Le deuxième axe est perpendiculaire au premier. Autrement dit, le coefficient de corrélation linéaire entre les deux axes est nul, les deux axes apportent des informations complémentaires et non redondantes. Le deuxième axe est l'axe perpendiculaire au précédent, sur lequel l'inertie de la projection est maximale. Le troisième axe est l'axe perpendiculaire aux deux précédents, sur lequel l'inertie de la projection est maximale...

Un exemple

Résumé des principales commandes :

```
> brush(mamiferes2,hist=T)
> pairs(mamiferes2,panel=function(x,y)
         {points(x,y); lines(lowess(x,y))})
> library(afc.acp)
> acp.mamiferes_acp(mamiferes2,q.choix=2)
> caplot(acp.mamiferes,cols="n",points=2)
```

Les données

Le tableau de données contient des mesures effectuées sur 62 espèces de mamifères. Les sept premières colonnes sont des variables quantitatives : poids total, poids du cerveau, durée de sommeil léger, durée de sommeil paradoxal, durée de sommeil total, durée de vie maximum et temps de gestation. Ce sont des moyennes par espèce.

```
> mamifères
```

			___sommeil___				
	poids total	poids cerveau	leger	paradoxal	total	vie max	temps gestation
African elephant	6654.000	5712.00	NA	NA	3.3	38.6	645.0
African giant pouched rat	1.000	6.60	6.3	2.0	8.3	4.5	42.0
Arctic Fox	3.385	44.50	NA	NA	12.5	14.0	60.0
Arctic ground squirrel	0.920	5.70	NA	NA	16.5	NA	25.0
Asian elephant	2547.000	4603.00	2.1	1.8	3.9	69.0	624.0
Baboon	10.550	179.50	9.1	0.7	9.8	27.0	180.0
Big brown bat	0.023	0.30	15.8	3.9	19.7	19.0	35.0
Brazilian tapir	160.000	169.00	5.2	1.0	6.2	30.4	392.0
Cat	3.300	25.60	10.9	3.6	14.5	28.0	63.0
Chimpanzee	52.160	440.00	8.3	1.4	9.7	50.0	230.0
Chinchilla	0.425	6.40	11.0	1.5	12.5	7.0	112.0
Cow	465.000	423.00	3.2	0.7	3.9	30.0	281.0
Desert hedgehog	0.550	2.40	7.6	2.7	10.3	NA	NA
Donkey	187.100	419.00	NA	NA	3.1	40.0	365.0
Eastern American mole	0.075	1.20	6.3	2.1	8.4	3.5	42.0

Echidna	3.000	25.00	8.6	0.0	8.6	50.0	28.0
European hedgehog	0.785	3.50	6.6	4.1	10.7	6.0	42.0
Galago	0.200	5.00	9.5	1.2	10.7	10.4	120.0
Genet	1.410	17.50	4.8	1.3	6.1	34.0	NA
Giant armadillo	60.000	81.00	12.0	6.1	18.1	7.0	NA
Giraffe	529.000	680.00	NA	0.3	NA	28.0	400.0
Goat	27.660	115.00	3.3	0.5	3.8	20.0	148.0
Golden hamster	0.120	1.00	11.0	3.4	14.4	3.9	16.0
Gorilla	207.000	406.00	NA	NA	12.0	39.3	252.0
Gray seal	85.000	325.00	4.7	1.5	6.2	41.0	310.0
Gray wolf	36.330	119.50	NA	NA	13.0	16.2	63.0
Ground squirrel	0.101	4.00	10.4	3.4	13.8	9.0	28.0
Guinea pig	1.040	5.50	7.4	0.8	8.2	7.6	68.0
Horse	521.000	655.00	2.1	0.8	2.9	46.0	336.0
Jaguar	100.000	157.00	NA	NA	10.8	22.4	100.0
Kangaroo	35.000	56.00	NA	NA	NA	16.3	33.0
Lesser short-tailed shrew	0.005	0.14	7.7	1.4	9.1	2.6	21.5
Little brown bat	0.010	0.25	17.9	2.0	19.9	24.0	50.0
Man	62.000	1320.00	6.1	1.9	8.0	100.0	267.0
Mole rat	0.122	3.00	8.2	2.4	10.6	NA	30.0
Mountain beaver	1.350	8.10	8.4	2.8	11.2	NA	45.0
Mouse	0.023	0.40	11.9	1.3	13.2	3.2	19.0
Musk shrew	0.048	0.33	10.8	2.0	12.8	2.0	30.0
N. American opossum	1.700	6.30	13.8	5.6	19.4	5.0	12.0
Nine-banded armadillo	3.500	10.80	14.3	3.1	17.4	6.5	120.0
Okapi	250.000	490.00	NA	1.0	NA	23.6	440.0
Owl monkey	0.480	15.50	15.2	1.8	17.0	12.0	140.0
Patas monkey	10.000	115.00	10.0	0.9	10.9	20.2	170.0
Phanlanger	1.620	11.40	11.9	1.8	13.7	13.0	17.0
Pig	192.000	180.00	6.5	1.9	8.4	27.0	115.0
Rabbit	2.500	12.10	7.5	0.9	8.4	18.0	31.0
Raccoon	4.288	39.20	NA	NA	12.5	13.7	63.0
Rat	0.280	1.90	10.6	2.6	13.2	4.7	21.0
Red fox	4.235	50.40	7.4	2.4	9.8	9.8	52.0
Rhesus monkey	6.800	179.00	8.4	1.2	9.6	29.0	164.0
Rock hyrax (Hetero. b)	0.750	12.30	5.7	0.9	6.6	7.0	225.0
Rock hyrax (Procavia hab)	3.600	21.00	4.9	0.5	5.4	6.0	225.0
Roe deer	14.830	98.20	NA	NA	2.6	17.0	150.0
Sheep	55.500	175.00	3.2	0.6	3.8	20.0	151.0
Slow loris	1.400	12.50	NA	NA	11.0	12.7	90.0
Star nosed mole	0.060	1.00	8.1	2.2	10.3	3.5	NA
Tenrec	0.900	2.60	11.0	2.3	13.3	4.5	60.0
Tree hyrax	2.000	12.30	4.9	0.5	5.4	7.5	200.0
Tree shrew	0.104	2.50	13.2	2.6	15.8	2.3	46.0
Vervet	4.190	58.00	9.7	0.6	10.3	24.0	210.0
Water opossum	3.500	3.90	12.8	6.6	19.4	3.0	14.0
Yellow-bellied marmot	4.050	17.00	NA	NA	NA	13.0	38.0

Ces données soumises par Roger Johnson sont disponibles dans statlib@lib.stat.cmu.edu.[1]
Les noms d'espèces ont été traduits dans la suite du document.

[1] Statlib est une bibliothèque électronique de jeux de données, programmes, logiciels, fonctions S, etc mis à disposition des utilisateurs. On peut obtenir ces données via la messagerie électronique (ou via le serveur www dont l'URL est http://lib.stat.cmu.edu):
mail statlib@lib.stat.cmu.edu
subject : (return)
send sleep from general

Choix de l'analyse

L'objectif est d'explorer ces données, notamment d'étudier les liaisons entre les variables et d'obtenir une cartographie des espèces permettant de repérer d'éventuelles espèces différentes des autres, d'éventuels groupes d'espèces homogènes pour les variables étudiées. Toutes les variables sont quantitatives. Une ACP permet d'étudier ces données.

Premiers traitements

Les valeurs manquantes

Le tableau initial contient des valeurs manquantes. Les fonctions d'ACP (et d'AFC) n'autorisant aucune valeur manquante, il faut, avant toute chose, les supprimer.

La variable "sommeil total" est la somme des variables "sommeil léger" et "sommeil paradoxal". Les trois variables sont redondantes. On élimine la variable "sommeil paradoxal". On supprime ensuite les espèces ayant des données manquantes. Il reste alors 42 espèces de mamifères.

```
> mamiferes1_mamiferes[,c(1:3,5:7)]
> mamiferes1_na.omit(mamiferes1)
```

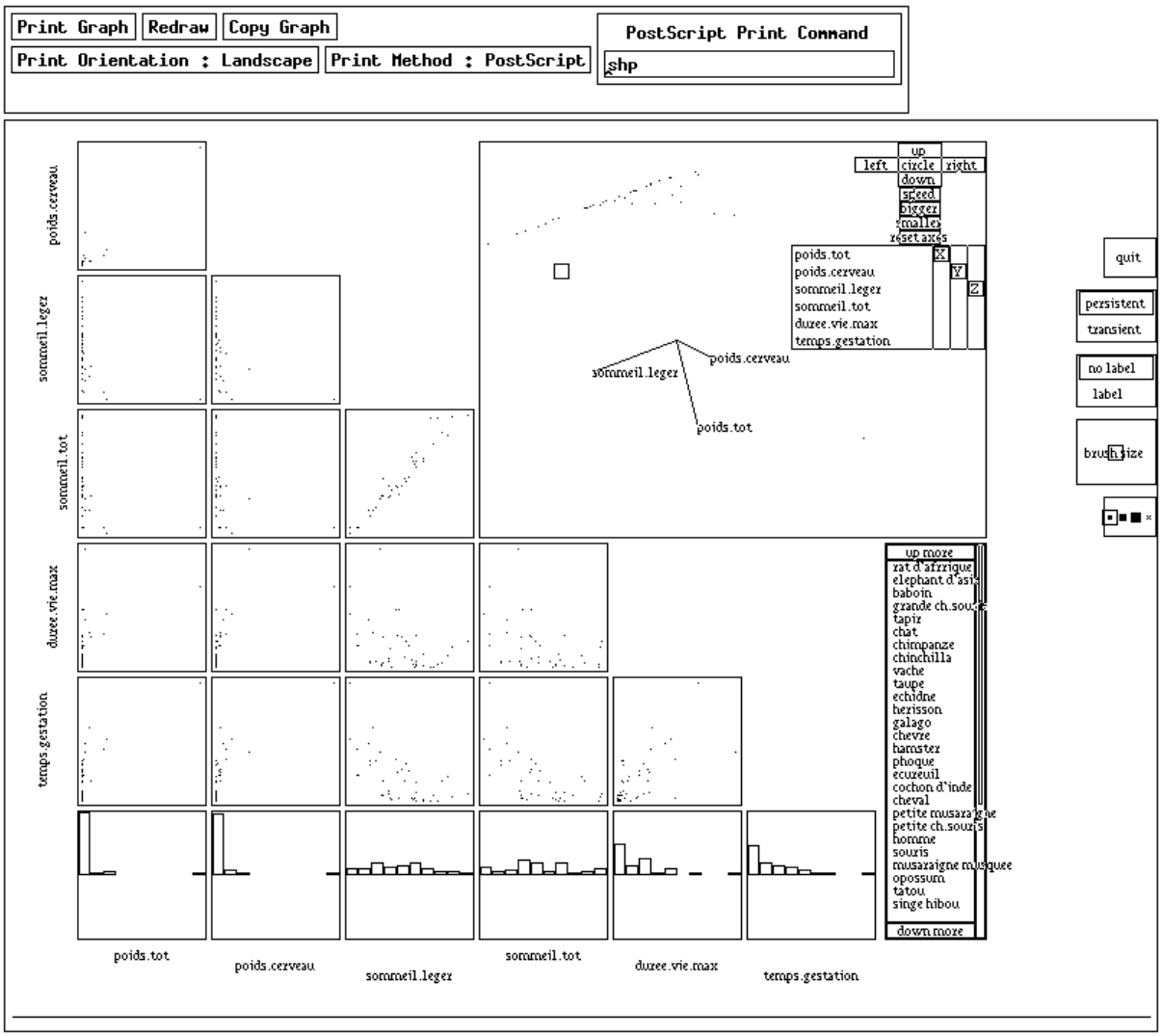
Visualisation graphique

Avant d'effectuer toute analyse, il est essentiel d'effectuer une première exploration graphique des données. Cela permet de détecter d'éventuelles erreurs de saisie, de prendre connaissance des données, des ordres de grandeur, de commencer une première interprétation. Ces graphiques complètent l'analyse et sont un support d'interprétation auquel on reviendra souvent.

Pour construire ces graphiques, il existe dans Splus de nombreuses possibilités de représentations. L'une d'entre elles est la fonction "brush". Elle permet de visualiser tous les nuages de points des variables prises deux à deux, et de représenter l'histogramme de chaque variable. Mais "brush" permet aussi d'interagir avec le graphique : on peut par exemple marquer des points ou des groupes de points sur l'un des graphiques, ils sont alors marqués sur l'ensemble des représentations ainsi que sur la liste des individus. De même, On peut "cliquer" sur certains individus de la liste et les voir apparaître marqués sur tous les graphiques.

Avant toute représentation graphique, il faut ouvrir une unité graphique (X11, motif, ...). Il faut aussi transformer le "data frame" mamiferes1 en matrice, avant d'utiliser "brush".

```
> motif()
> is.data.frame(mamiferes1)
T
> mamiferes2_as.matrix(mamiferes1)
> brush(mamiferes2,hist=T)
```



Certaines variables comme “sommeil total” et “sommeil léger” semblent linéairement corrélées. Les nuages croisant le poids total aux autres variables sont très difficiles (sinon impossibles) à lire. Ceci parce que des animaux comme l’éléphant d’Asie ou la vache sont très lourds par rapport à la plupart des autres animaux. On remarque d’ailleurs sur l’histogramme du poids total la très forte dissymétrie de cette variable.

L’analyse

Première analyse

Pour effectuer l’ACP, on utilise la fonction “acp” disponible dans la librairie “afc.acp” développée par le laboratoire de Biométrie de Montpellier.

```
> library(afc.acp)
> acp(mamiferes2,q.choix=2)
```

L'argument "q.choix=2" permet de réaliser une ACP centrée-réduite; on dit aussi une **ACP normée**, ou une ACP sur la matrice des corrélations. Par défaut, c'est l'argument q.choix=1 qui est choisi, c'est à dire qu'on réalise une ACP centrée, non normée, une ACP sur la matrice de variance-covariance. L'ACP non normée donne aux variables des poids d'autant plus élevés que leur variance est élevée. Elle donnerait un poids très grand à la variable "poids total", qui expliquerait à elle seule le premier axe. Une ACP non normée se justifie dans les cas où l'utilisateur souhaite donner un poids plus grand aux variables présentant de grandes variations. Cela peut être le cas, lorsque les variables sont de même nature, mesurées avec les mêmes unités, par exemple des poids ou des longueurs... L'ACP normée permet de donner le même poids à toutes les variables, quelque soit leur variance, l'unité de mesure utilisée... C'est une ACP normée qui s'impose ici.

La première étape d'une ACP normée consiste à centrer et réduire les données, c'est à dire à retrancher sa moyenne à chaque variable, et à diviser chaque variable par son écart-type. Les variables obtenues sont donc de moyenne nulle et de variance égale à 1. Elles auront le même poids dans l'analyse.

```
Moyennes
 poids.tot poids.cerveau sommeil.leger sommeil.tot
  100.8139      218.6838      8.742857    10.64286
 duree.vie.max temps.gestation
      19.37143        129.9405
```

La fonction affiche[2] les moyennes des différentes variables. Ceci donne un ordre de grandeur mais il ne faut pas interpréter les moyennes sans s'intéresser aux distributions : une distribution dissymétrique, ou même une seule valeur extrême peuvent avoir une très forte influence sur la moyenne. Le poids moyen, par exemple, 101 kg, n'est pas représentatif de la majorité des espèces car il est fortement influencé par les espèces les plus lourdes. L'examen de la représentation obtenue par "brush" est un complément indispensable à tout résumé numérique.

```
Matrice des correlations
                poids.tot poids.cerveau sommeil.leger
      poids.tot     1.000         0.956        -0.394
  poids.cerveau     0.956         1.000        -0.387
  sommeil.leger    -0.394        -0.387         1.000
    sommeil.tot    -0.343        -0.337         0.968
  duree.vie.max     0.470         0.629        -0.372
temps.gestation     0.714         0.734        -0.606
                sommeil.tot duree.vie.max temps.gestation
      poids.tot      -0.343         0.470           0.714
  poids.cerveau      -0.337         0.629           0.734
  sommeil.leger       0.968        -0.372          -0.606
    sommeil.tot       1.000        -0.382          -0.614
  duree.vie.max      -0.382         1.000           0.646
temps.gestation      -0.614         0.646           1.000
```

[2] Toutes les sorties de la fonction "acp" apparaitront dans un encadré de ce type, sur fond grisé.

L'argument q.choix=2 a été choisi, l'ACP consiste donc à étudier la matrice de corrélation (si q.choix=1, c'est la matrice de variance-covariance qui est étudiée).

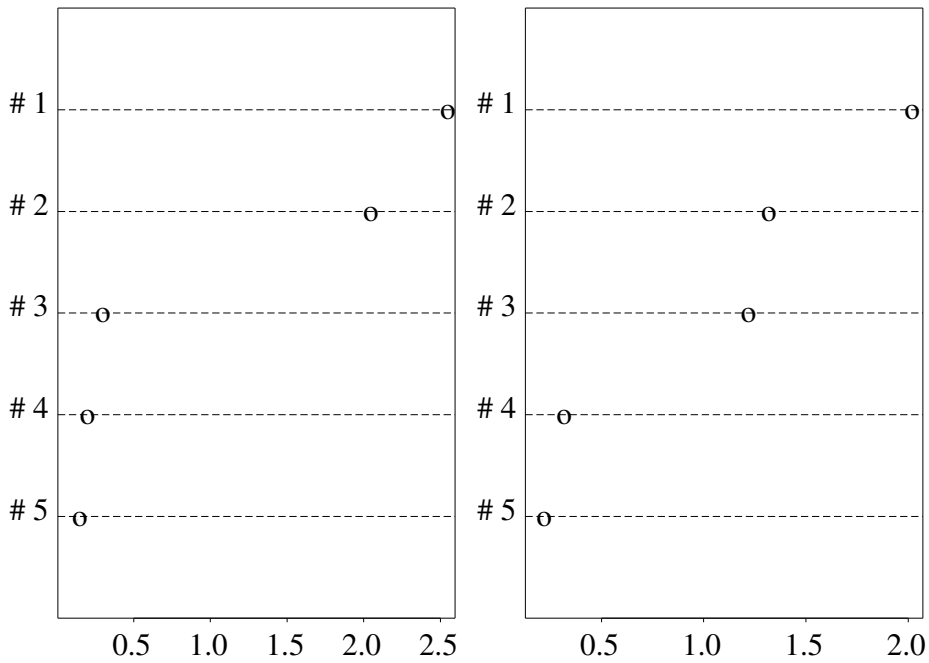
La matrice des corrélations contient les coefficients de corrélation de toute les variables prises deux à deux, elle résume les relations entre les différentes variables.

Choix du nombre d'axes

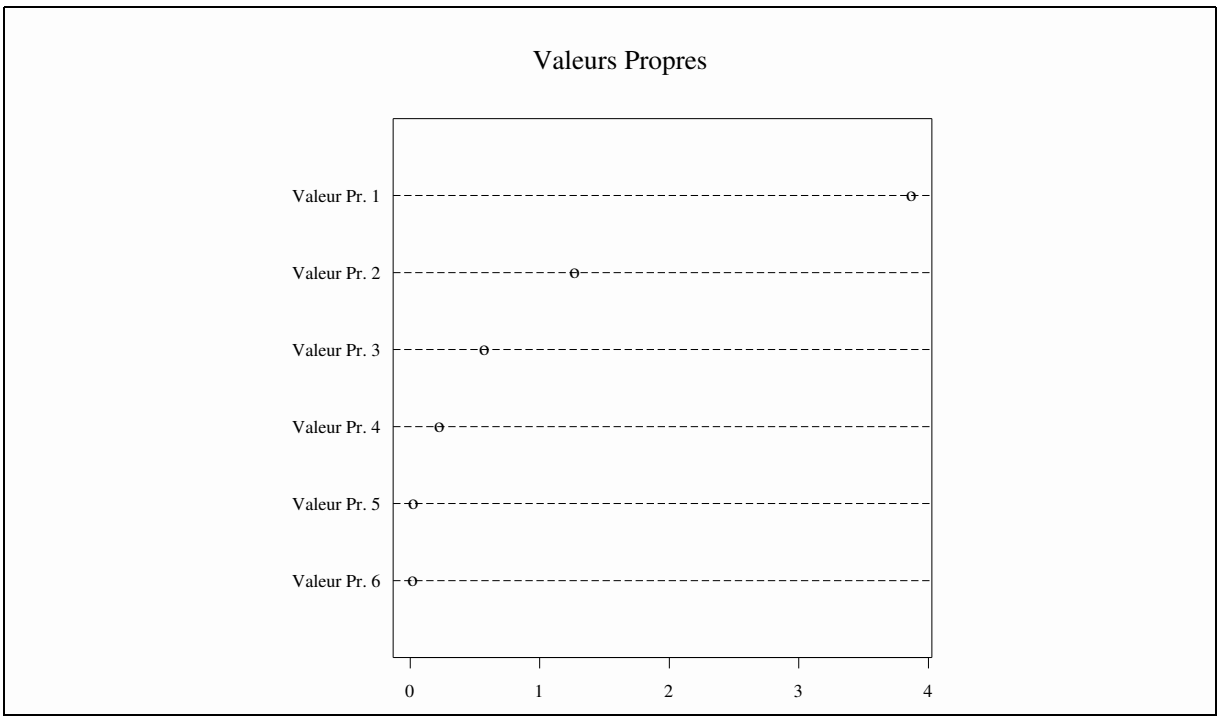
```
Combien de Composantes Voulez-Vous retenir ? (<= 6 ) 2
```

Splus attend une réponse : le nombre d'axes retenus dans l'analyse. Ce choix est important. En effet, si on retient un nombre d'axes insuffisant, on perd l'information contenue dans les axes restants. Par contre, si on retient un trop grand nombre d'axes, on donne une signification non fondée à des axes qui doivent être considérés comme "du bruit". Le choix du nombre d'axes peut être basé sur l'un ou l'autre des critères suivants :

- La dimension de l'espace des variables initiales est égale à p. Le nombre de composantes principales est égal à p (les p composantes principales s'obtiennent par rotation du système d'axes initial). La somme des valeurs propres est égale à p dans le cas d'une ACP normée. Sous l'hypothèse d'un nuage sphérique (nuage ne présentant aucune information, toutes les variables sont indépendantes), l'inertie moyenne absorbée par un axe est égale à 1, tous les axes ont la même inertie. **La valeur 1 peut donc servir de point de repère** : on peut interpréter comme du "bruit" tous les axes dont les valeurs propres sont inférieures à 1, et s'intéresser uniquement aux autres.
- D'autres critères de choix du nombre d'axes peuvent être retenus : on peut tenir compte de **la forme dégressive des valeurs propres**, en veillant à ne pas écarter une composante si la composante précédente a une valeur propre très proche. Par exemple dans le premier cas on ne retiendra que 2 axes, alors que dans le deuxième on en retiendra 3 :



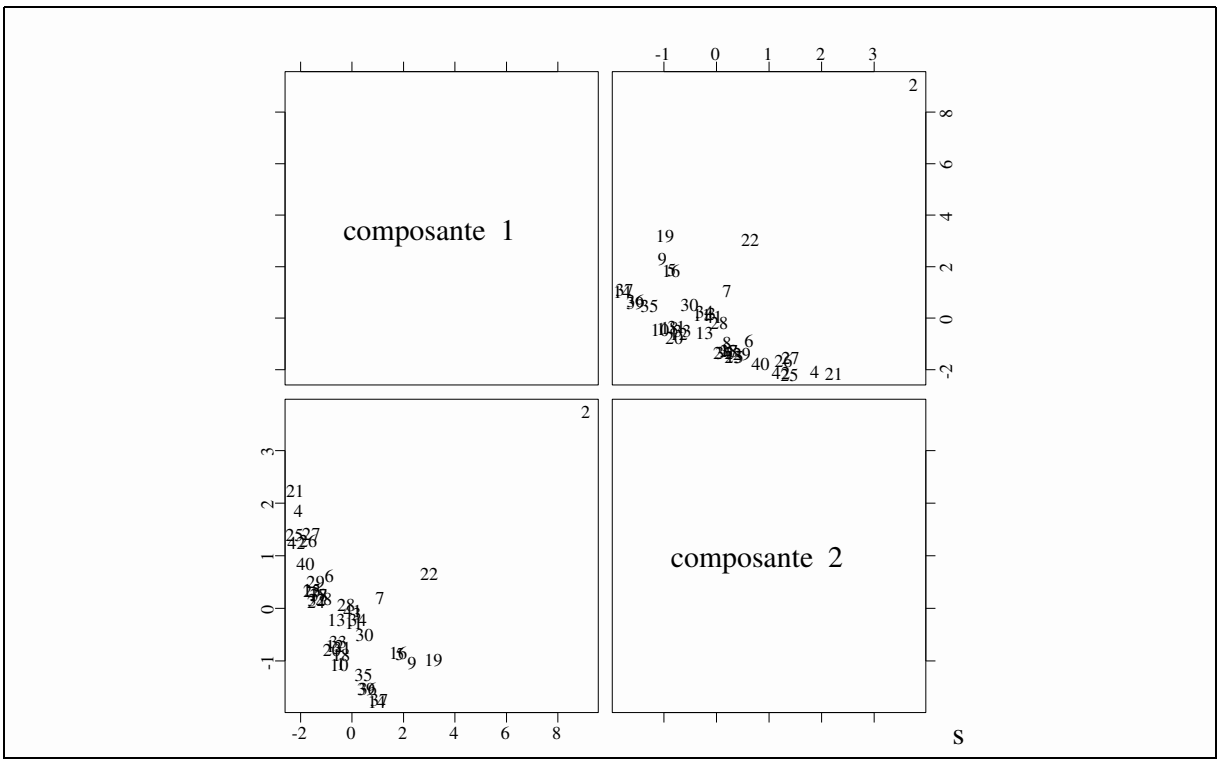
- On note $\lambda_1, ..., \lambda_p$ les p valeurs propres présentées par ordre décroissant. On a $\lambda_1 + \lambda_2 + ... + \lambda_p = p$ dans le cas d'une ACP normée. $\frac{\lambda_1}{\lambda_1+\lambda_2...\lambda_p}$ est le pourcentage d'inertie (de variabilité) expliquée par le premier axe, $\frac{\lambda_2}{\lambda_1+\lambda_2+...\lambda_p}$ le pourcentage d'inertie expliquée par le deuxième axe. Les axes étant orthogonaux par construction, on peut sommer ces pourcentages d'inertie. $\frac{\lambda_1}{\lambda_1+\lambda_2...\lambda_p} + \frac{\lambda_2}{\lambda_1+\lambda_2+...\lambda_p}$ représente par exemple le pourcentage d'inertie expliquée par les deux premiers axes. Le nombre d'axes peut être choisi de façon à ce que l'**inertie cumulée soit jugée suffisante** (85% par exemple). Ce peut être un critère de choix du nombre d'axes, qu'il faudra "aménager" en fonction de la forme dégressive des valeurs propres. C'est à dire que ce critère ne doit pas conduire à conserver un axe et exclure le suivant, si les valeurs propres (les inerties) sont sensiblement égales.



Ici nous ne retiendrons que les deux premiers axes : ils absorbent à eux deux 85% de l'inertie totale. La troisième composante représente moins de 10% de l'inertie totale et a une valeur propre proche de celle de la quatrième composante, elle est inférieure à 1. La première composante absorbe à elle seule 61% de l'inertie.

```
Valeurs propres :
 3.87216 1.27191 0.576325 0.228736 0.0291634 0.0217141
```

Projection sur les plans factoriels

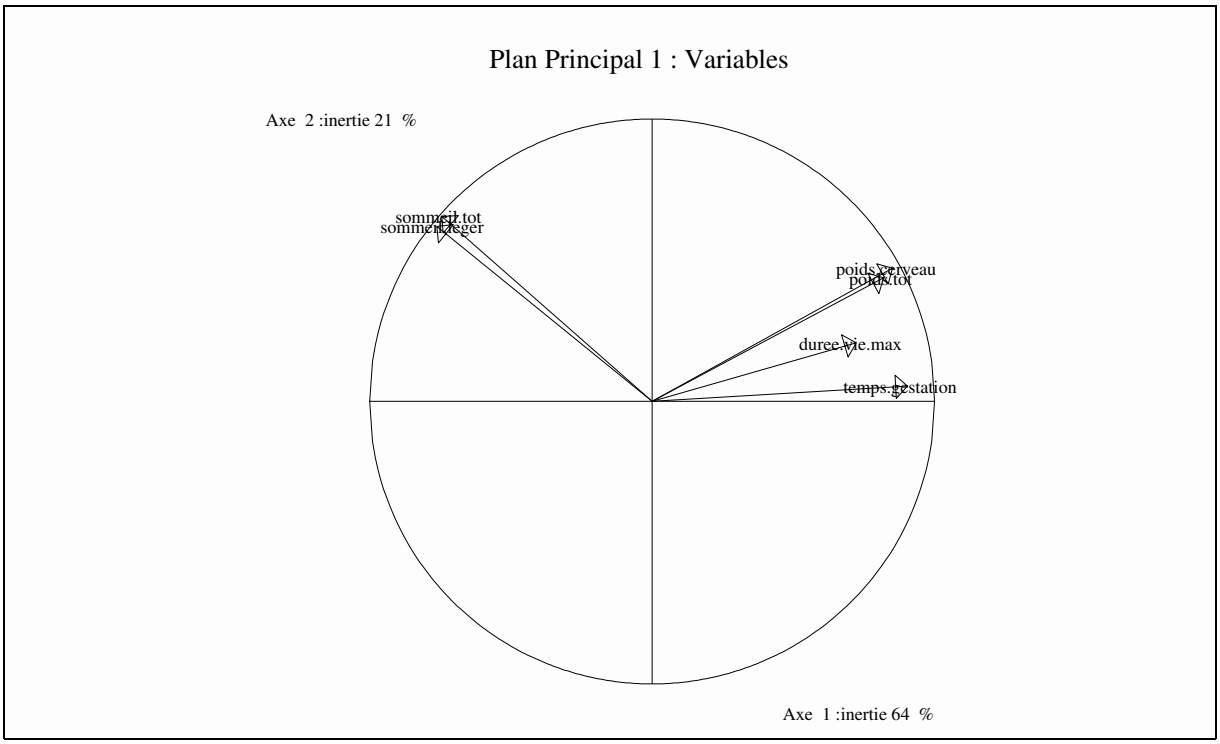


La fonction “acp” représente la projection des individus, ici des espèces, sur les plans factoriels (formés par les axes factoriels retenus, ici les axes 1 et 2).

Le point 2 est très à l’écart des autres espèces. Il s’agit de l’éléphant d’Asie, espèce ayant le poids total le plus élevé. Un examen des contributions absolues (voir plus loin) montre que l’éléphant d’Asie contribue très fortement à la détermination du premier axe. Cela groupe fortement les autres espèces à gauche, sur l’axe 1.

L’interprétation de cette carte des espèces ne pourra se faire qu’après avoir donné une signification aux axes.

Les cercles de corrélations



Voici le premier cercle des corrélations. Les variables sont projetées sur le plan formé par les deux premiers axes. On interprète sur ces graphiques les **distances des variables au cercle**, les **angles entre variables et axes**, et les **angles entre variables**.

Plus une variable est proche du cercle, meilleure est sa représentation sur le plan. On n'interprète pas dans ce plan une variable qui serait mal représentée, c'est à dire une variable éloignée du cercle.

Plus l'angle entre une variable et un axe est faible, plus la variable est corrélée avec cet axe, c'est à dire qu'elle est bien représentée sur cet axe. Deux variables ayant même direction et même sens sont corrélées positivement, tandis que deux variables ayant la même direction mais étant de sens opposés sont corrélées négativement. Deux variables perpendiculaires ont un coefficient de corrélation nul.

Ici on peut distinguer des groupes de variables : sommeil total et sommeil léger sont très proches, tout comme poids du cerveau et poids total. Ceci était visible sur la représentation graphique faite par "brush" ainsi que sur la matrice des corrélations, les deux coefficients de corrélation correspondants étant supérieurs à 0.95.

Coordonnées, contributions absolues et relatives

VARIABLES
(cta = cta * 1000 ctr = ctr * 1000)

	Axe 1			Axe 2		
	coord.	cta	ctr	coord.	cta	ctr
poids.tot	330.0	**174**	672	170.0	150	191
poids.cerveau	620.0	**187**	726	340.0	175	222
sommeil.leger	-2.9	**150**	581	2.3	302	384
sommeil.tot	-3.5	**143**	554	3.0	337	428
duree.vie.max	14.0	**135**	522	4.2	34	43
temps.gestation	110.0	**211**	**817**	6.9	2	**3**

INDIVIDUS
(cta = cta * 1000 ctr = ctr * 1000)

	Axe 1			Axe 2		
	coord.	cta	ctr	coord.	cta	ctr
rat d'afrique	-0.38	**1**	**80**	-1.000	20	**579**
elephant d'asie	9.10	**512**	822	3.800	265	140
baboin	0.24	**0**	140	-0.075	0	14
grande ch.souris	-2.00	**26**	525	1.900	67	447
tapir	1.90	**22**	569	-0.840	13	109
chat	-0.84	**4**	435	0.640	8	255
chimpanze	1.10	**8**	407	0.220	1	15
chinchilla	-0.91	**5**	770	0.220	1	44
vache	2.40	**34**	813	-1.000	19	150
taupe	-0.41	**1**	90	-1.000	20	562
echidne	0.15	**0**	7	-0.240	1	17
herisson	-0.58	**2**	244	-0.680	9	332
galago	-0.52	**2**	668	-0.190	1	92
chevre	1.10	**7**	257	-1.800	58	721
hamster	-1.50	**13**	840	0.390	3	58
phoque	1.90	**22**	665	-0.830	13	130
ecureuil	-1.20	**9**	865	0.280	1	45
cochon d'inde	-0.34	**1**	100	-0.850	14	638
cheval	3.20	**65**	897	-0.940	17	76
petite musaraigne	-0.71	**3**	280	-0.760	11	328
petite ch.souris	-2.10	**28**	436	2.300	96	495
homme	3.10	**58**	462	0.680	9	22
souris	-1.50	**13**	836	0.360	2	51
musaraigne musquee	-1.30	**10**	832	0.140	0	10
opossum	-2.20	**28**	675	1.400	38	298

```
                tatou  -1.60  16 542  1.300  32 357
         singe hibou  -1.50  14 446  1.400  39 416
         singe patas  -0.15   0  74  0.089   0  27
           phalanger  -1.30  11 811  0.520   5 125
              cochon   0.57   2 407 -0.470   4 281
               lapin  -0.30   1  83 -0.730  10 478
                 rat  -1.30  10 852  0.190   1  18
              renard  -0.46   1 235 -0.610   7 421
        singe rhesus   0.30   1 211 -0.190   1  88
  daman (hetero b.)   0.53   2 113 -1.200  29 616
   daman (procavia)   0.70   3 142 -1.500  42 659
              mouton   1.10   8 298 -1.700  55 675
              Tanrec  -1.20   9 862  0.270   1  44
   daman des arbres   0.63   2 128 -1.500  43 734
musaraigne des bois  -1.70  18 759  0.880  14 195
              grivet   0.11   0  19 -0.018   0   1
  opossum aquatique  -2.10  27 681  1.300  30 253
```

Le premier tableau concerne les variables, le deuxième concerne les individus (les différentes espèces). Dans ces tableaux, on trouve trois colonnes par axe. Ces trois colonnes sont coord., cta et ctr, ce qui correspond aux coordonnées, aux contributions absolues et aux contributions relatives. Plus de détail sur ces trois colonnes sont donnés en annexe.

La colonne coord est utilisée pour représenter graphiquement les individus dans les plans factoriels, et les variables sur les cercles des corrélations. Dans le cas des individus, il s'agit des nouvelles coordonnées, alors que pour les variables, la colonne coord affichée ne correspond pas aux coordonnées lues sur le cercle des corrélations, mais à une transformation de ces coordonnées. (voir détail en annexe)

Les colonnes cta et ctr sont ici multipliées par 1000, dans le seul but d'améliorer la lisibilité. La somme des contributions absolues pour un même axe est égale à 1000. **La contribution absolue représente la participation de chaque variable à la construction de l'axe**. Les six variables participent de façon équivalente à la construction du premier axe (le temps de gestation, le poids total et le poids du cerveau ont une contribution légèrement supérieure) : aucune variable n'a une contribution très faible (avec 6 variables, la contribution absolue moyenne est égale à 1000/6 soit 167.

En ce qui concerne la colonne ctr, on verrait, si on avait retenu tous les axes, que la somme de toutes les contributions relatives pour une même variable est égale à 1000. Pour chacune des variables, les **contributions relatives** sont les parts de variabilité expliquées sur chaque axe. Cette colonne ctr doit donc être interprétée comme la **qualité de représentation de la variable**

sur l'axe. Par exemple, la qualité de représentation du temps de gestation sur le premier axe est ici de 0.8 (elle ne peut dépasser 1), cette variable est mieux représentée sur le premier axe que sur le deuxième. Cela est visible sur le premier cercle des corrélations. La coordonnée de la variable "temps de gestation" est très élevée sur le premier axe, alors qu'elle est presque nulle sur le deuxième. (les coordonnées des variables sur les cercles des corrélations sont en fait les racines carrées des ctr, ce qui explique que l'on interprète les cercles des corrélations en termes de qualité de représentation).

En sommant la qualité de représentation sur le premier et le deuxième axe, on obtient la qualité de représentation sur le plan formé par les axes 1 et 2 que nous appellerons plan 1–2 (ou plan principal). La qualité de représentation du poids total sur le plan 1–2 est de 0.672+0.191 soit 0.863. Les 14% restant sont répartis sur les 4 axes que nous n'avons pas retenus dans l'analyse.

Les contributions absolues et les contributions relatives s'interprètent de la même façon pour les individus que pour les variables. **Les contributions absolues se lisent colonne par colonne, et s'interprètent comme la participation des individus ou des variables à la construction de chaque axe. Les contributions relatives se lisent ligne par ligne et s'interprètent comme la qualité de représentation de chaque individu ou variable, sur les axes**.

Bilan, critique et transformations

En observant rapidement le tableau, on s'aperçoit que l'éléphant d'Asie a une contribution absolue au premier axe supérieure à 500, c'est à dire qu'à lui tout seul, il influence plus la construction du premier axe que toutes les autres espèces réunies. Devant un tel phénomène, on utilise généralement une possibilité très intéressante qu'offre l'ACP : les **éléments supplémentaires** (individus ou variables).

Les variables ou individus supplémentaires n'interviennent pas dans la détermination des axes, mais on peut les projeter sur les plans factoriels, de façon à les situer par rapport aux autres variables ou aux autres individus.

Nous pourrions mettre l'éléphant d'Asie en individu supplémentaire.

```
> acp(mamiferes1,q.choix=2,isup=2)
```

Nous aurions alors constaté que le cheval contribuait très fortement à la détermination du premier axe. En considérant à son tour le cheval comme un individu supplémentaire, on aurait trouvé une contribution très forte de la vache à la détermination du premier axe ...

En fait, le problème vient du fait que la distribution de la variable poids est fortement dissymétrique, et que cette variable contribue fortement à la formation du premier axe, il en est de même pour le poids du cerveau.

```
> attach(mamiferes1)
```

mamiferes1 et mamiferes2 contiennent les mêmes données, mamiferes1 est un "data frame", mamiferes2 est une matrice. La fonction "attach" permet d'accéder directement aux variables du data fame.

```
> stem(poids.tot)
N = 42   Median = 2.25
Quartiles = 0.28, 10.55

Decimal point is at the colon

    0 : 0000011112345789
    1 : 0067
    2 : 05
    3 : 03556
    4 : 22
    5 :
    6 : 8
    7 :
    8 :
    9 :
   10 : 05
   11 :
   12 :
   13 :
   14 :
   15 :
   16 :
   17 :
   18 :
   19 :
   20 :
   21 :
   22 :
   23 :
   24 :
   25 :
   26 :
   27 : 7

High:   52.16   55.50   62.00   85.00  160.00  192.00  465.00
  521.00 2547.00
```

La transformation log, transformation fréquemment utilisée pour les variables pondérales, symétrise la distribution.

```
> stem(log(poids.tot))
N = 42   Median = 0.804719
Quartiles = -1.272966, 2.356126

Decimal point is at the colon

  -5 : 3
  -4 : 6
  -3 : 880
  -2 : 6331
  -1 : 63
  -0 : 97321
   0 : z05579
   1 : 12333449
   2 : 34
   3 : 3
   4 : 0014
   5 : 13
   6 : 13
   7 : 8
```

Le fait de symétriser la distribution a l'avantage de permettre une bien meilleure visualisation des nuages de points représentant le poids et une autre variable. Il est donc plus judicieux de travailler sur la variable log(poids.tot) plutôt que sur la variable poids.tot.

```
> mamiferes2[,"poids.tot"]_log(poids.tot)
```

La matrice de corrélation et le cercle des corrélations révèlent des corrélations très élevées entre poids total et poids du cerveau d'une part, entre sommeil léger et sommeil total d'autre part. On pouvait s'attendre à de telles valeurs : globalement, plus un animal est gros, plus son cerveau est gros. Les deux informations sont partiellement redondantes. Il serait plus pertinent d'étudier le poids total (ou son log) et le poids relatif du cerveau par rapport au poids total (poids cerveau / poids total). De même, temps de sommeil total et temps de sommeil léger sont partiellement redondants. Mieux vaut étudier le temps de sommeil total et la part relative du temps de sommeil léger (sommeil léger / sommeil total). Ce sont deux informations différentes.

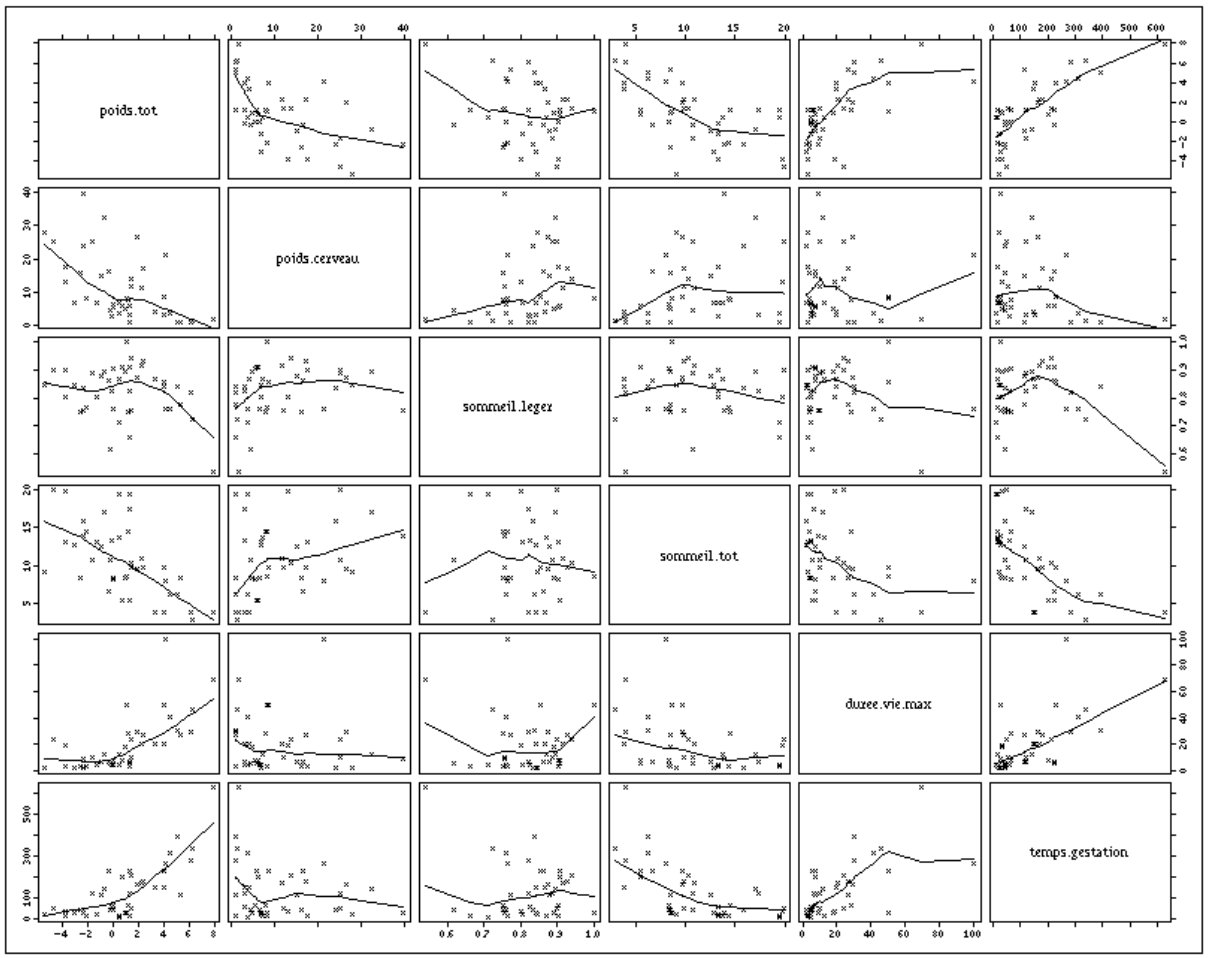
```
> mamiferes2[,"poids.cerveau"]_poids.cerveau/poids.tot
> mamiferes2[,"sommeil.leger"]_sommeil.leger/sommeil.tot
```

La première ACP effectuée sur les données brutes permet de "**débroussailler**" l'étude. Elle peut suggérer de considérer comme **supplémentaires** certains individus qui ont une influence déterminante sur sur les axes, d'**étudier séparément 2 populations d'individus** qui se distinguent visiblement l'une de l'autre. Elle peut suggérer, comme c'est le cas ici, des **transformations** des variables à étudier. Ces transformations sont choisies dans le but de symétriser les distributions, ou d'étudier une information plus pertinente que l'information véhiculée par les données brutes.

Cependant, dans cet exemple, les trois transformations auraient pu être basées sur une réflexion a priori. L'examen du graphique obtenu par la fonction "brush" montre clairement la dissymétrie

de la distribution des poids totaux, la corrélation élevée entre poids total et poids du cerveau, de même que la corrélation élevée entre temps de sommeil total et temps de sommeil léger.

```
>pairs(mamiferes2,panel=function(x,y)
        {points(x,y);lines(lowess(x,y))})
```



Les nuages de points représentant le poids total et les autres variables sont maintenant plus lisibles. On peut faciliter la visualisation des liaisons entre variables, en superposant un lissage à chacun des nuages.

Deuxième analyse

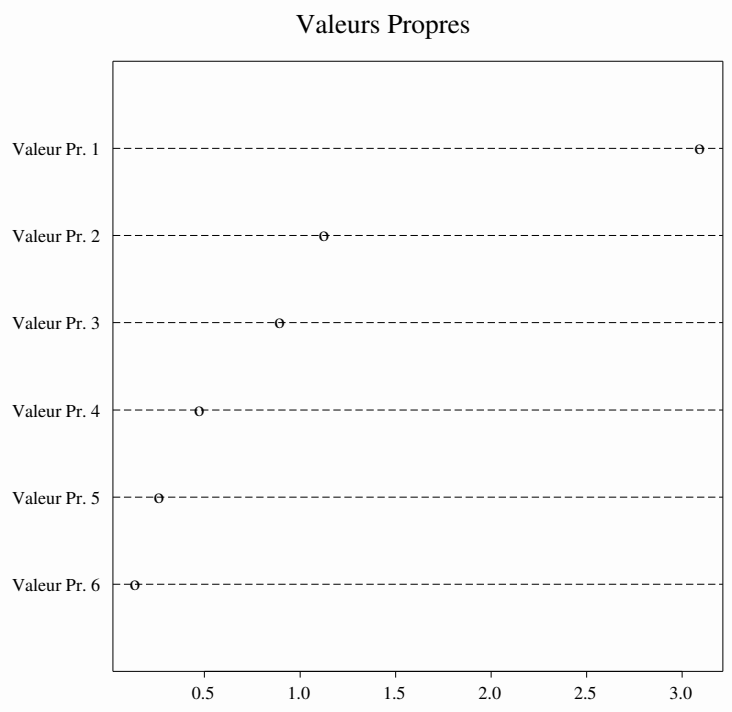
On recommence l'ACP à partir du jeu de données transformé.

```
>acp.mamiferes_acp(mamiferes2,q.choix=2)
```

```
Moyennes
poids.tot poids.cerveau sommeil.leger sommeil.tot duree.vie temps.gestation
0.8039947      11.15057     0.8226459    10.64286  19.37143 129.9405
```

```
Matrice des correlations
                poids.tot poids.cerveau sommeil.leger sommeil.tot
      poids.tot     1.000        -0.552        -0.256      -0.617
  poids.cerveau    -0.552         1.000         0.269       0.311
  sommeil.leger    -0.256         0.269         1.000      -0.030
    sommeil.tot    -0.617         0.311        -0.030       1.000
  duree.vie.max     0.647        -0.076        -0.152      -0.382
temps.gestation     0.759        -0.236        -0.211      -0.614
                duree.vie.max temps.gestation
      poids.tot         0.647           0.759
  poids.cerveau        -0.076          -0.236
  sommeil.leger        -0.152          -0.211
    sommeil.tot        -0.382          -0.614
  duree.vie.max         1.000           0.646
temps.gestation         0.646           1.000
Combien de Composantes Voulez-Vous retenir ? (<= 6 ) 3
```

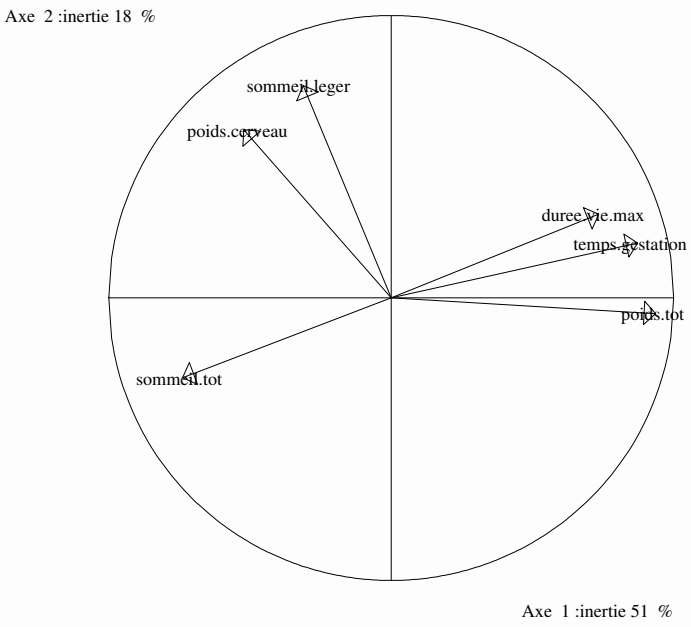
La matrice des corrélations révèle une liaison entre le poids total et la plupart des autres variables, notamment le temps de gestation.



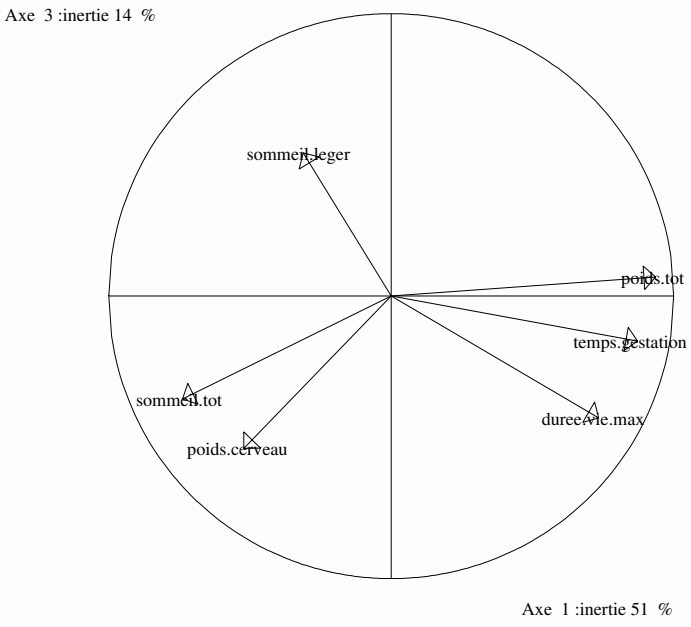
Ici nous retenons les trois premières composantes car les deuxième et troisième ont des valeurs propres très proches. Mieux vaut ne pas pas les dissocier. La valeur de la quatrième composante est suffisamment petite et suffisamment inférieure à la troisième pour qu'on ne la retienne pas.

```
 Valeurs propres :
3.09541 1.12841 0.896256 0.477127 0.264021 0.138773
```

Plan Principal 1 : Variables



Plan Principal 2 : Variables



La matrice des corrélations et le cercle des corrélations montrent la pertinence des transformations effectuées sur le poids du cerveau et le sommeil léger : les variables sont mieux réparties dans les cercles, le poids relatif du cerveau véhicule une information différente du poids total; la part relative du sommeil léger est une information différente de la durée totale du sommeil.

VARIABLES
(cta = cta * 1000 ctr = ctr * 1000)

	Axe 1			Axe 2			Axe 3		
	coord.	cta	ctr	coord.	cta	ctr	coord.	cta	ctr
poids.tot	2.840	**283**	**877**	-0.1680	**3**	3	0.2050	**5**	5
poids.cerveau	-4.970	**88**	272	5.6700	314	354	-5.1300	324	291
sommeil.leger	-0.028	**31**	97	0.0674	501	565	0.0454	287	257
sommeil.tot	-3.440	**177**	546	-1.3100	71	80	-1.6900	148	133
duree.vie.max	14.700	**175**	540	5.9500	78	88	-8.6400	208	186
temps.gestation	110.000	**246**	762	24.6000	34	38	-20.0000	28	25

INDIVIDUS
(cta = cta * 1000 ctr = ctr * 1000)

	Axe 1			Axe 2			Axe 3		
	coord.	cta	ctr	coord.	cta	ctr	coord.	cta	ctr
rat d'afrique	-0.3180	**1**	**48**	-0.9580	19	**438**	0.5230	7	**131**
elephant d'asie	5.6700	**248**	806	-1.1200	26	31	-2.2000	129	121
baboin	0.3120	**1**	41	1.3800	40	805	0.1480	1	9
grande ch.souris	-2.0200	**31**	607	-0.6330	8	60	-0.9590	24	137
tapir	2.6900	**56**	838	0.2470	1	7	0.5690	9	37
chat	-0.1170	**0**	7	-0.9630	20	482	-0.6380	11	212
chimpanze	1.6900	**22**	663	0.6740	10	106	-0.3190	3	24
chinchilla	-1.0200	**8**	727	0.4060	3	115	0.2220	1	34
vache	2.6800	**55**	893	0.0415	0	0	0.8410	19	88
taupe	-1.0800	**9**	330	-0.4510	4	58	-0.1400	1	6
echidne	0.2120	**0**	6	1.6300	56	368	0.8410	19	98
herisson	-0.1980	**0**	6	-2.3200	113	789	-0.4370	5	28
galago	-1.2200	**11**	424	1.2300	32	433	-0.2830	2	23
chevre	1.2700	**12**	440	0.3320	2	30	1.2800	43	445
hamster	-1.4200	**15**	574	-1.1800	29	394	-0.0582	0	1
phoque	2.5500	**50**	980	-0.1860	1	5	-0.2260	1	8
ecureuil	-2.1900	**37**	403	0.7090	11	42	-2.0800	115	361
cochon d'inde	-0.3780	**1**	68	0.1830	1	16	1.3600	49	884
cheval	3.5200	**96**	977	-0.3400	2	9	-0.1080	0	1
petite musaraigne	-2.2800	**40**	590	0.9780	20	109	-0.3580	3	15
petite ch.souris	-2.5900	**51**	640	0.9350	18	84	-1.2600	42	152
homme	2.8500	**62**	395	1.5400	50	116	-2.6900	192	352
souris	-2.1600	**36**	894	0.5350	6	55	0.2940	2	17
musaraigne musquee	-1.5300	**18**	680	-0.5300	6	81	0.6410	11	119
opossum	-1.1500	**10**	187	-2.1800	101	673	-0.4660	6	31
tatou	-0.5860	**3**	105	-1.0700	24	349	0.2360	1	17
singe hibou	-1.7600	**24**	395	1.3800	40	244	-1.2500	41	198
singe patas	0.2160	**0**	32	0.7980	13	434	0.4870	6	161
phalanger	-0.8720	**6**	424	-0.3010	2	50	0.5550	8	172
cochon	1.5000	**17**	563	-0.8500	15	181	0.4560	6	52
lapin	-0.1370	**0**	10	0.1470	0	11	1.1500	35	691
rat	-1.1500	**10**	577	-0.8850	17	339	0.3620	3	57
renard	-0.2070	**0**	34	-0.6990	10	387	-0.0419	0	1
singe rhesus	0.0484	**0**	1	1.5300	49	695	-0.7470	15	166
daman (hetero b.)	0.0441	**0**	1	0.8470	15	306	0.3950	4	66
daman (procavia)	0.6490	**3**	120	0.5980	8	102	1.4500	56	596
mouton	1.4900	**17**	545	0.0579	0	1	1.1900	38	350
Tanrec	-0.7360	**4**	267	-0.8970	17	396	0.7110	13	249

```
   daman des arbres  0.4690   2  71  0.6120   8 121  1.4100  53 644
musaraigne des bois -2.1200  34 850  0.2570   1  13 -0.6930  13  91
             grivet  0.2330   0  23  1.2900  35 707  0.3810   4  62
  opossum aquatique -0.8750   6  81 -2.7800 163 815 -0.5590   8  33
```

On cherche à donner une signification aux axes, en s'appuyant sur le tableau des contributions.

Les variables

cta : contribution absolue des variables à la construction des axes. Elle s'interprète colonne par colonne, c'est à dire axe par axe.

La première composante principale qui représente 51% de l'inertie du nuage est principalement déterminée par le poids total et le temps de gestation et, dans une moindre mesure, par le sommeil total et la durée de vie maximale. On retrouve sur la partie positive de l'axe 1 les individus ayant un grand poids total et/ou ayant une grande durée de gestation.

L'axe 2 est principalement expliqué par le sommeil léger et dans une moindre mesure par le poids du cerveau. Sur l'axe 2 on trouve dans la partie positive les espèces ayant une grande part de sommeil léger et dans la partie négative les individus ayant une grande part de sommeil paradoxal (par rapport aux autres espèces car la part de sommeil paradoxal est toujours inférieure à la part de sommeil léger).

L'interprétation de l'axe 3 est plus délicate car les contributions absolues sont plus partagées entre les différentes variables. De plus, les variables contribuant le plus fortement à cet axe contribuent également à la construction d'un autre axe.

ctr : contribution relative. Elle s'interprète ligne par ligne, c'est à dire variable par variable, comme la qualité de représentation d'une variable sur les différents axes.

Le poids total est très bien représenté sur l'axe 1, très mal représenté sur l'axe 2. Le sommeil léger est bien représenté sur l'axe 2 et sur l'axe 3. C'est donc au plan 2–3 qu'il faudra se reporter pour observer la répartition des différentes espèces selon la part relative du sommeil léger. On y interpréte également le poids relatif du cerveau.

Les individus

cta : contribution absolue des individus à la construction des axes. Elle s'interprète colonne par colonne, c'est à dire axe par axe.

L'éléphant et le cheval sont les deux espèces qui contribuent le plus fortement à la construction du premier axe. Au contraire, des espèces comme le chat ou le hérisson n'interviennent pas dans la construction de cet axe.

ctr : contribution relative. Elle s'interprète ligne par ligne, c'est à dire individu par individu, comme la qualité de représentation d'un individu sur les différents axes.

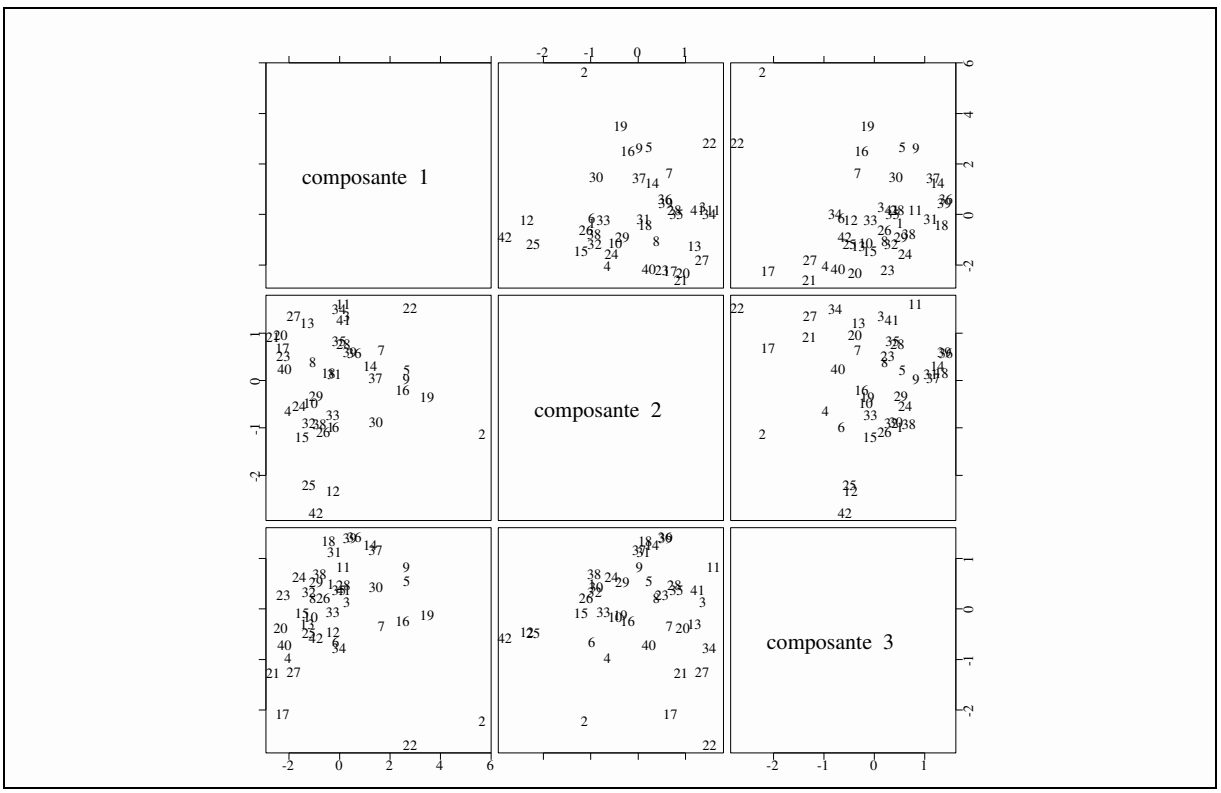
L'éléphant d'Asie est très bien représenté sur le premier axe (0.806), son poids total et son temps de gestation étant très élevés.

Pour confirmer les interprétations et éviter les erreurs, il faut toujours se référer aux données initiales. Par exemple, le cochon d'inde est très bien représenté sur l'axe 3 (0.884) et a une coordonnée positive. On peut vérifier sur les données que le cochon d'inde est un mamifère dont le poids relatif du cerveau est plus faible que la moyenne. On retrouve dans la partie positive de l'axe 3 les espèces ayant un petit cerveau.

coord. : La coordonnée d'un individu est toujours interprétable, même si sa qualité de représentation sur cet axe est mauvaise. Par exemple, pour le singe rhésus, sa coordonnée presque nulle indique des valeurs moyennes pour les variables déterminant l'axe 1 c'est à dire des valeurs moyennes pour l'ensemble des variables. Deux individus ayant des coordonnées similaires sur un axe (ou sur un plan) se ressemblent pour les variables qui déterminent l'axe (le plan).

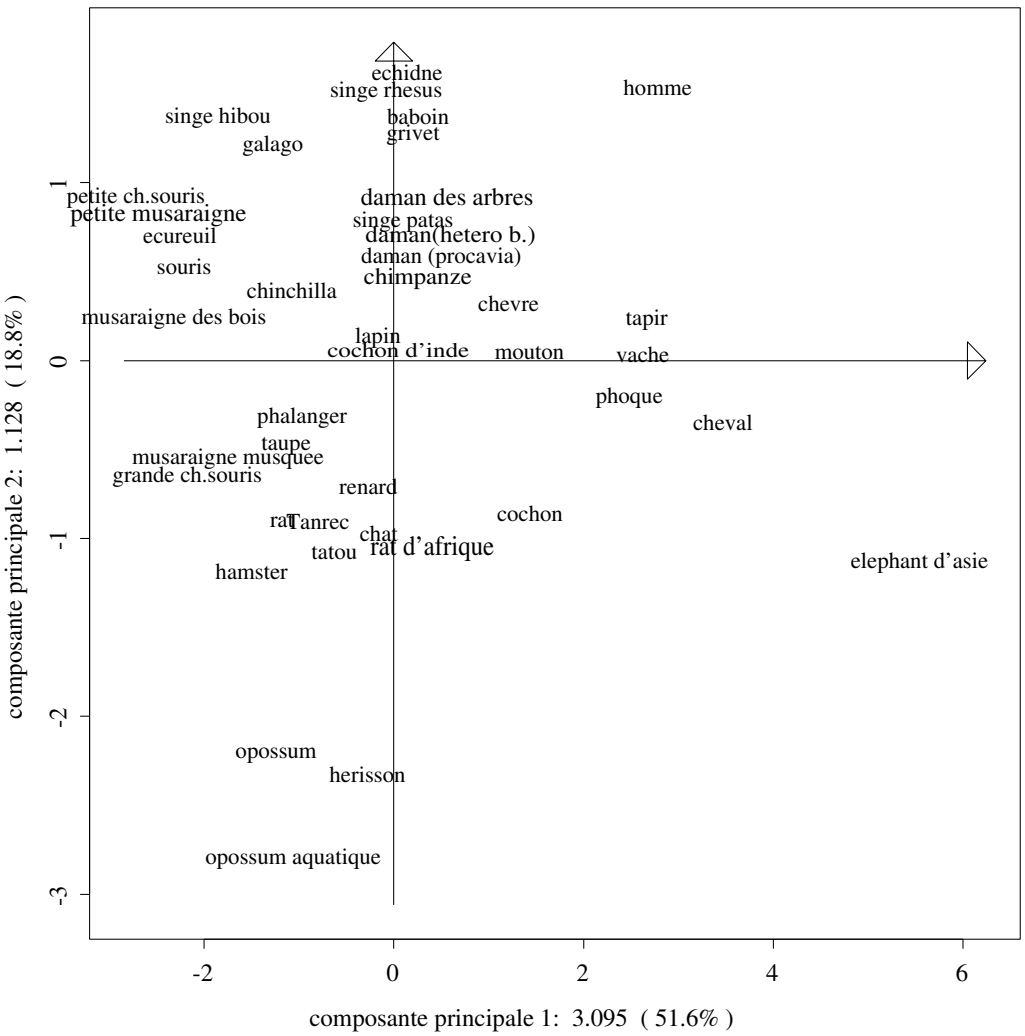
Lecture des plans factoriels

Après avoir donné une signification aux axes, on s'intéresse aux représentations des espèces sur les plans factoriels.



On peut améliorer et personnaliser ces représentations graphiques à l'aide de la fonction caplot.

```
> caplot(acp.mamiferes,cols="n",titre="ACP",points=2)
```



Certains points sont parfois superposés sur le graphique. La liste des points vus et des points cachés est affichée.

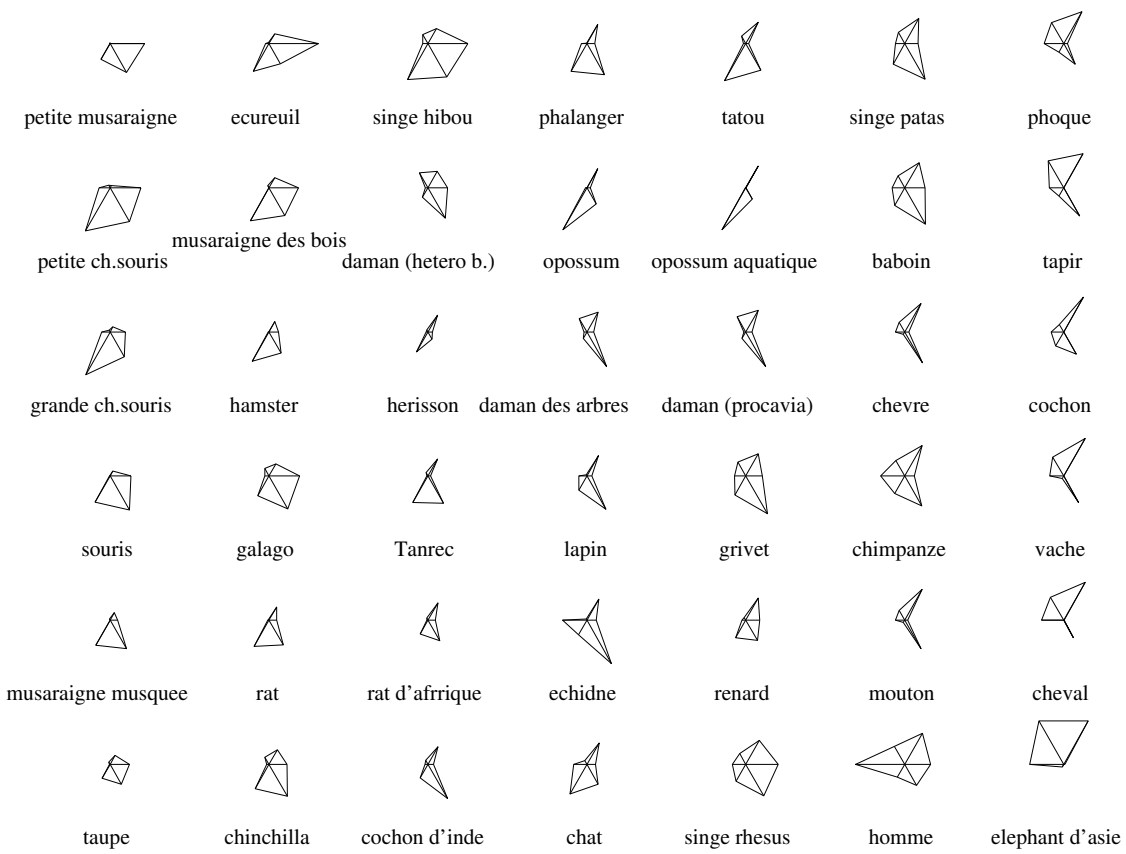
```
$ptvus:
 "petite ch.souris" "singe patas"  "daman (procavia)"
  "daman (procavia)" "lapin"            "chat"

$ptcaches:
 "petite musaraigne" "daman (hetero b.)" "chimpanze"
"daman des arbres"  "cochon d'inde"  "rat d'afrrique"
```

On a ajouté manuellement les points doubles (ceux qui sont masqués par un autre point).

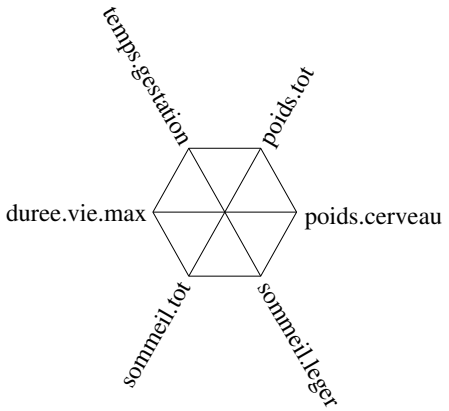
Les espèces sont ordonnées sur le premier axe, en fonction de leur poids total et de leur temps de gestation, sur le deuxième en fonction de la part relative du sommeil léger, et du poids relatif du cerveau.

Pour aller plus loin...

```
> ordre_order(poids.tot)
> stars(mamiferes2[ordre,c(2,1,6,5,4,3)])
```

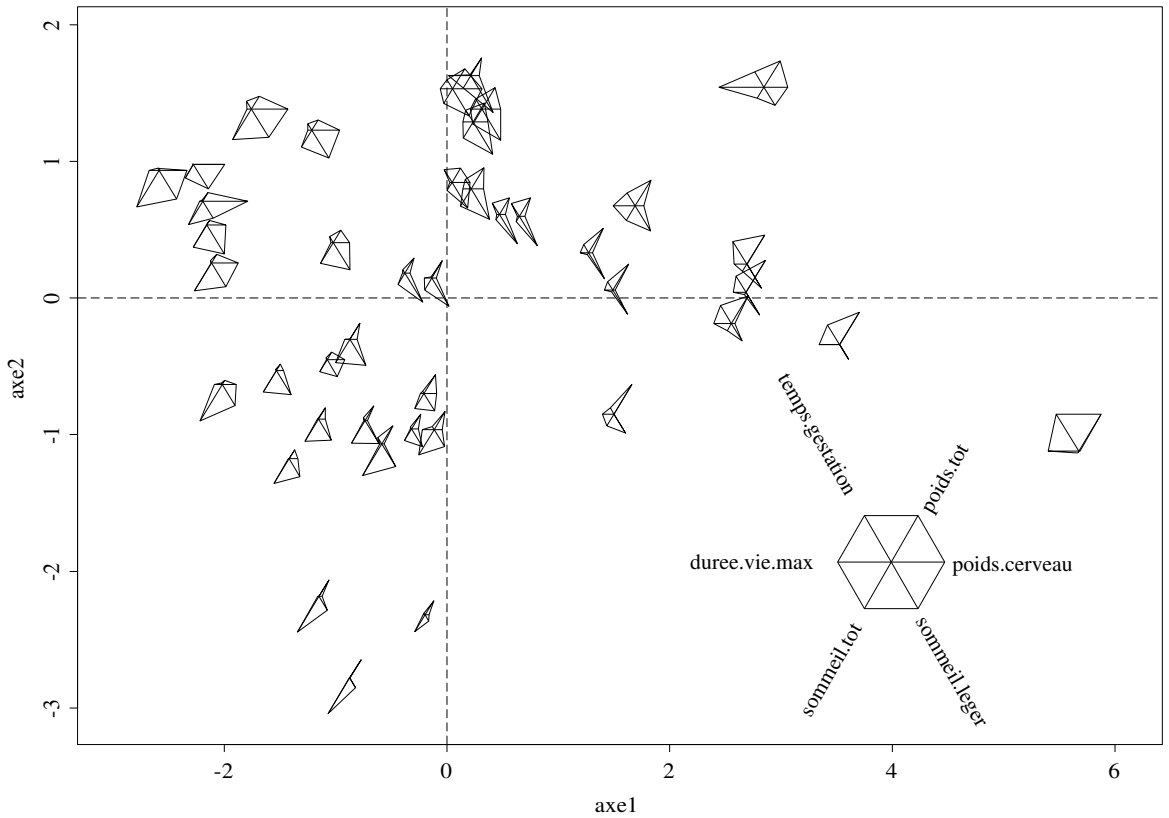


Chaque espèce est représentée par une étoile, la longueur des branches des étoiles étant proportionnelle aux valeurs prises par les variables. L'ordre des variables, c'est à dire l'ordre des branches a été modifié de façon à juxtaposer les variables liées entre elles. Les espèces sont ordonnées selon le poids total, et représentées colonne par colonne.



Ceci permet de représenter les données espèce par espèce, de repérer des espèces voisines, d'identifier les particularités de chacune des espèces. On peut aussi représenter les espèces sur le plan factoriel à l'aide de la fonction symbols.

```
> coords_acp.mamiferes$ind.cords
> mat_(mamiferes2[,c(2,1,6,5,4,3)])
> mat_scale(mat,center=apply(mat,2,min))
> mat_scale(mat,center=F,scale=apply(mat,2,max))
> symbols(coords[,1],coords[,2],stars=mat,inches=0.3,
    xlab="axe1",ylab="axe2")
```

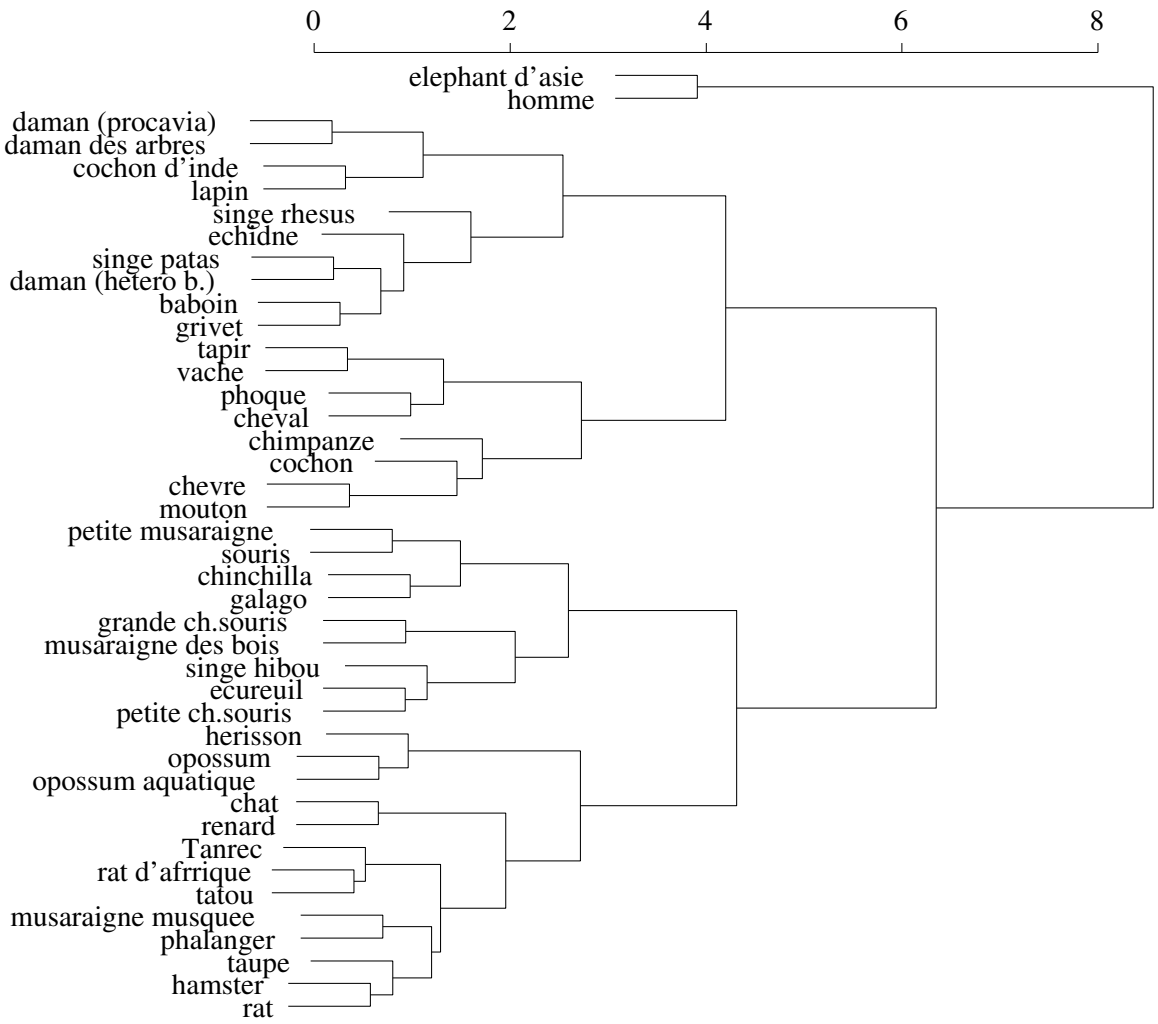


Le tapir, la vache, le phoque et le cheval qui sont proches sur le plan ont des caractéristiques semblables pour les différentes mesures : ils ont un grand poids total, une durée de sommeil total assez faible avec une grande proportion de sommeil léger, un temps de gestation élevé. Ils ont un poids relatif du cerveau très faible.

Une classification des lignes ou des colonnes après une analyse factorielle est souvent utile à l'interprétation.

A partir des coordonnées des individus sur les trois premiers axes, on calcule la matrice de distance entre individus (distance euclidienne). On trace le dendogramme obtenu par classification hiérarchique ascendante (méthode d'agrégation : lien complet).

```
> d_dist(coords)
> h_hclust(d)
> plclust(h,labels=dimnames(coords)[[1]])
```

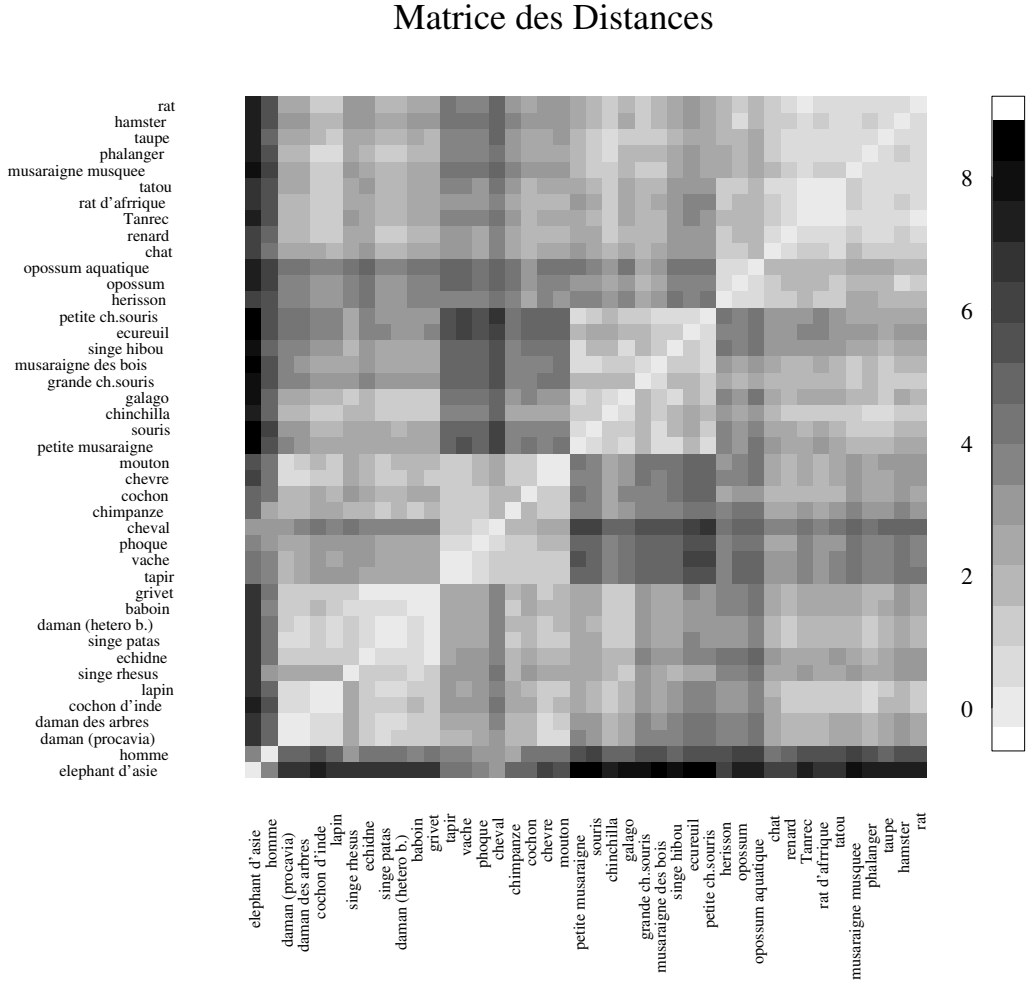


Ce graphique montre les ressemblances entre certaines espèces. Le daman, le daman des arbres, le cochon d'inde et le lapin, par exemple, sont des espèces voisines. Il permet de regrouper les espèces en classes homogènes. Plus on agrège les espèces entre elles, plus l'hétérogénéité est grande.

On peut obtenir une représentation graphique des distances :

```
> d_dist2full(d)
> d_d[h$order,h$order]
> nom_dimnames(mamiferes2)[[1]]
> dimnames(d)_list(nom,nom)
> dist.gris(d)
```

Les fonctions dist2full et dist.gris sont des fonctions écrites en langage Splus, elles sont présentées en annexe.



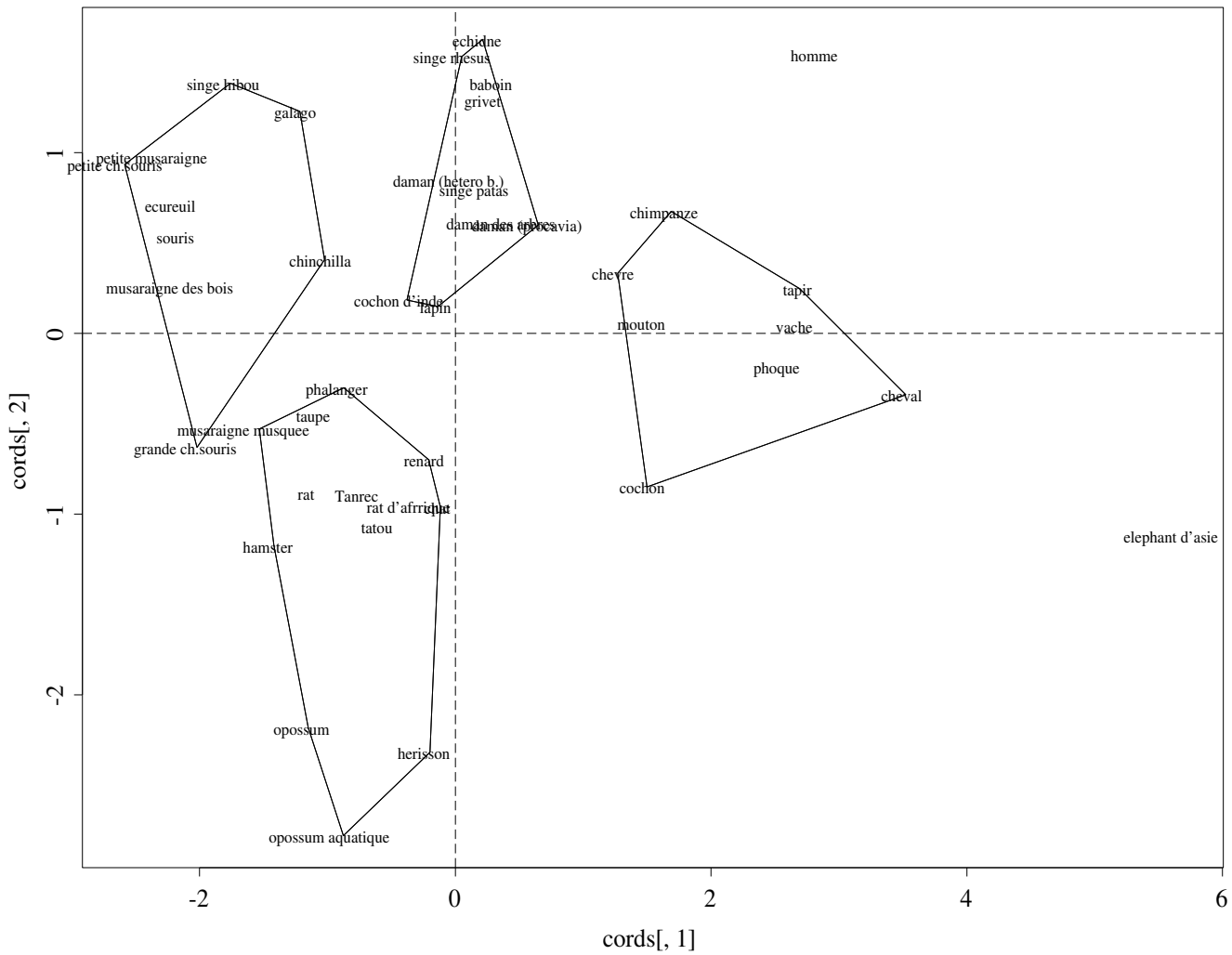
Les espèces sont ordonnées selon le résultat de la classification. On peut ainsi clairement distinguer des groupes d'espèces qui se ressemblent.

Cette représentation est beaucoup plus complète que le dendogramme. Elle permet de visualiser toutes les distances entre espèces, alors que le dendogramme ne permet pas d'évaluer la distance entre la petite chauve-souris et l'éléphant d'Asie par exemple. Si le cheval ressemble au mouton, à la chèvre, ..., au tapir, il se distingue des autres espèces de façon plus marquée que le mouton, la chèvre, ..., le tapir. Pour l'homme qui, comme l'éléphant d'Asie, est très différent des autres espèces, on peut remarquer que c'est du chimpanzé et du singe rhésus qu'il est le plus proche.

Le dendogramme étant le résultat d'une classification hiérarchique ascendante, le fait que l'homme et le chimpanzé ne s'agrègent qu'à la racine de l'arbre ne prouve pas qu'ils soient très différents. La représentation en grisés de la matrice de distances, permet, elle, de visualiser la distance entre l'homme et le chimpanzé.

On peut également représenter les différentes classes sur le plan factoriel. On a découpé l'ensemble des espèces en 6 classes et matérialisé chaque classe par l'enveloppe convexe regroupant toutes ses espèces. L'homme et l'éléphant d'Asie constituent deux classes à eux seuls.

```
> ct_cutree(h,k=6)
> plot(coords[,1],coords[,2],type="n")
> text(coords[,1],coords[,2],dimnames(coords)[[1]],cex=.7)
> for(i in 1:max(ct))
  {coords1_coords[ct==i,]
  hull1_chull(coords1[,1],coords1[,2])
  polygon(coords1[hull1,1],coords1[hull1,2],density=0)
  }
```



Analyse Factorielle des Correspondances

Le principe

L'Analyse Factorielle des Correspondances est très proche de l'Analyse en Composantes Principales et peut même en être considérée comme un cas particulier.

L'AFC diffère de l'ACP par le type de données qu'elle permet d'étudier, et par le type de distance implicitement utilisée : l'**ACP** est adaptée à l'étude des **variables quantitatives**, la méthode est par conséquent basée sur la **distance euclidienne**. L'**AFC** est, elle, adaptée à l'étude de **variables qualitatives**, c'est la **distance du chi2** qui est utilisée.

L'AFC permet l'étude de **tableaux de contingence**. Un tableau de contingence simple est le tableau des effectifs obtenus en croisant les modalités de deux variables qualitatives. L'**AFC simple** étudie ce type de tableaux, et permet de décrire les dépendances ou la correspondance entre deux ensembles de caractères. Le tableau obtenu par juxtaposition de plusieurs tableaux de contingence simples est un tableau de contingence multiple. Une AFC réalisée sur ce type de tableaux est une **AFC multiple**, elle permet de décrire la dépendance ou la correspondance entre plusieurs ensembles de caractères.

L'AFC permet aussi d'étudier tout autre type de tableau individus x variables, pour lequel la distance du chi2 a une signification, par exemple les **tableaux de notes**, ou les **tableaux de données binaires** (0–1, présence-absence). Les **tableaux en codage disjonctif complet**[3] sont des tableaux de données binaires, obtenus par codage et "éclatement" des données initiales. Ces tableaux permettent d'étudier simultanément des variables qualitatives et quantitatives. L'AFC sur tableau disjonctif complet est particulièrement bien adaptée aux traitements d'enquêtes.

[3] Voir en annexe.

Analyse Factorielle des Correspondances Simple

Un exemple

Résumé des principales commandes :

```
> apply(brevet,1,dotchart)
> total_apply(brevet,1,sum)
> apply(brevet[order(total),],2,dotchart)
> library(afc.acp)
> brevet.ca_ca(brevet)
> caplot(brevet.ca,rows="n",cols="p")
> caplot(brevet.ca,rows="p",cols="n")
```

Les données

Voici le tableau des nombres de brevets déposés dans 9 pays, de 1980 à 1986. C'est un tableau d'effectifs obtenu en croisant les modalités de 2 variables qualitatives, le pays et l'année. Il s'agit d'un tableau de contigence simple.

```
>brevet
          1980 1981 1982 1983 1984 1985  1986
     USA 3574 5820 6243 7486 8969 9726 11126
   Japon  853 2026 2980 3728 4192 5384  6259
     RFA 4087 5747 6002 6360 6604 7892  8710
  France 1634 2371 2252 2735 2797 2891  3304
    G.B. 1853 2146 2366 2652 2944 3076  3396
  Italie  206  490  616  748  863  969  1180
Pays-Bas  268  542  545  677  732  779   905
   Suede  302  478  551  659  764  737   843
  Suisse  699  947  912 1047 1059 1220  1193
```

Tableau extrait de "Exploration informatique et statistique des données" de M. Jambu.

L'examen visuel de ce tableau brut ne permet pas une interprétation complète. Il est bien difficile de savoir s'il y a homogénéité entre les profils d'évolution des nombres de brevets, si ce n'est repérer une augmentation d'une année sur l'autre. Le taux d'augmentation est-il constant, existe t-il des années déterminantes, ...? Autant de questions auquelles on est tenté de répondre en réalisant des graphiques adaptés.

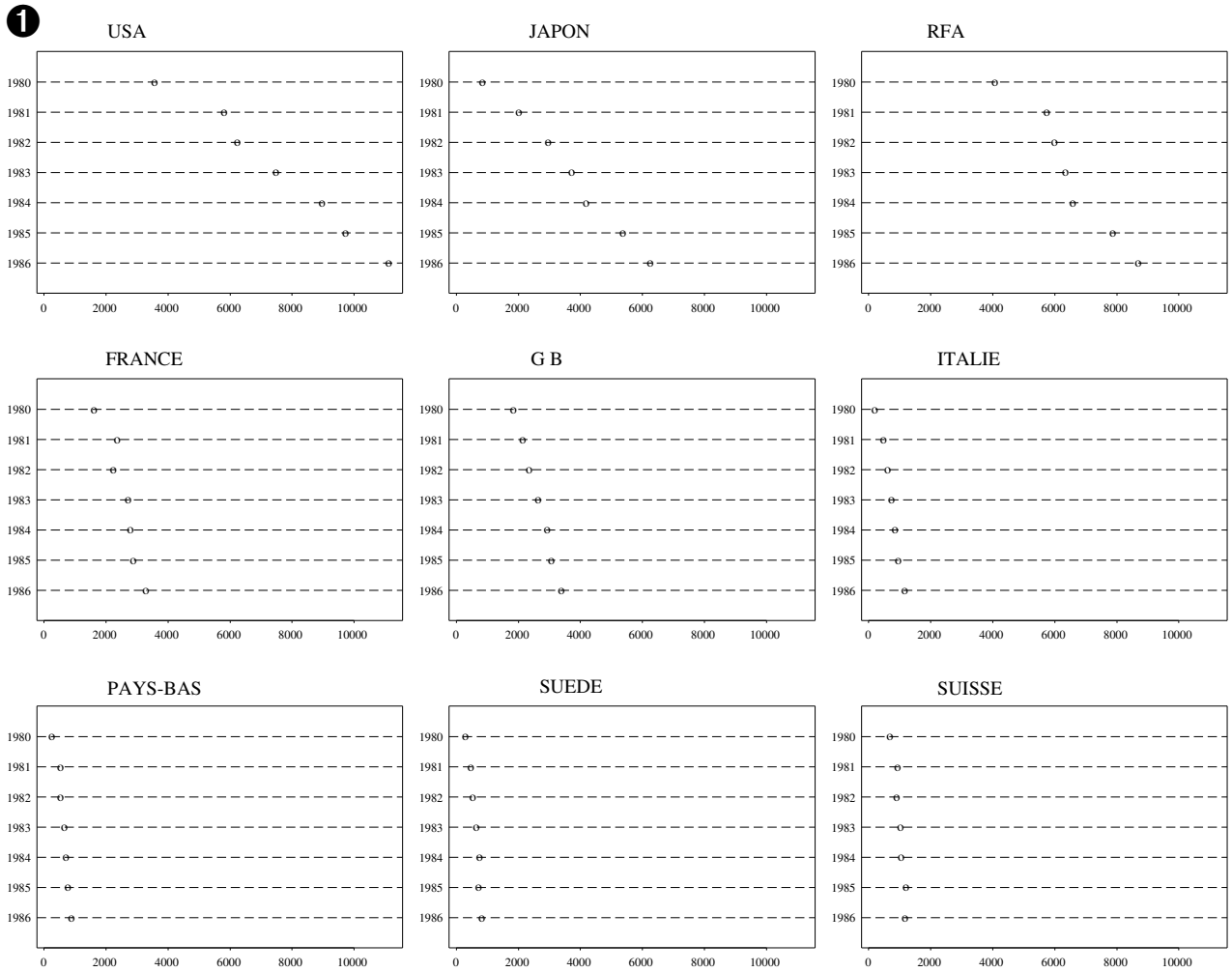
Une analyse factorielle des correspondances permet d'y répondre. L'AFC permet d'étudier des tableaux de contingence simples, c'est à dire des tableaux d'effectifs résultant du croisement de

2 variables qualitatives (un tableau de contingence multiple résulte du croisement de plus de 2 variables qualitatives). L'AFC permet ainsi une représentation synthétique des données.

L'objectif des représentations suivantes est d'observer les différences entre pays, l'évolution dans le temps et l'évolution par pays.

```
> par(mfrow=c(3,3))
> apply(brevet,1,dotchart,xlim=range(brevet))
```

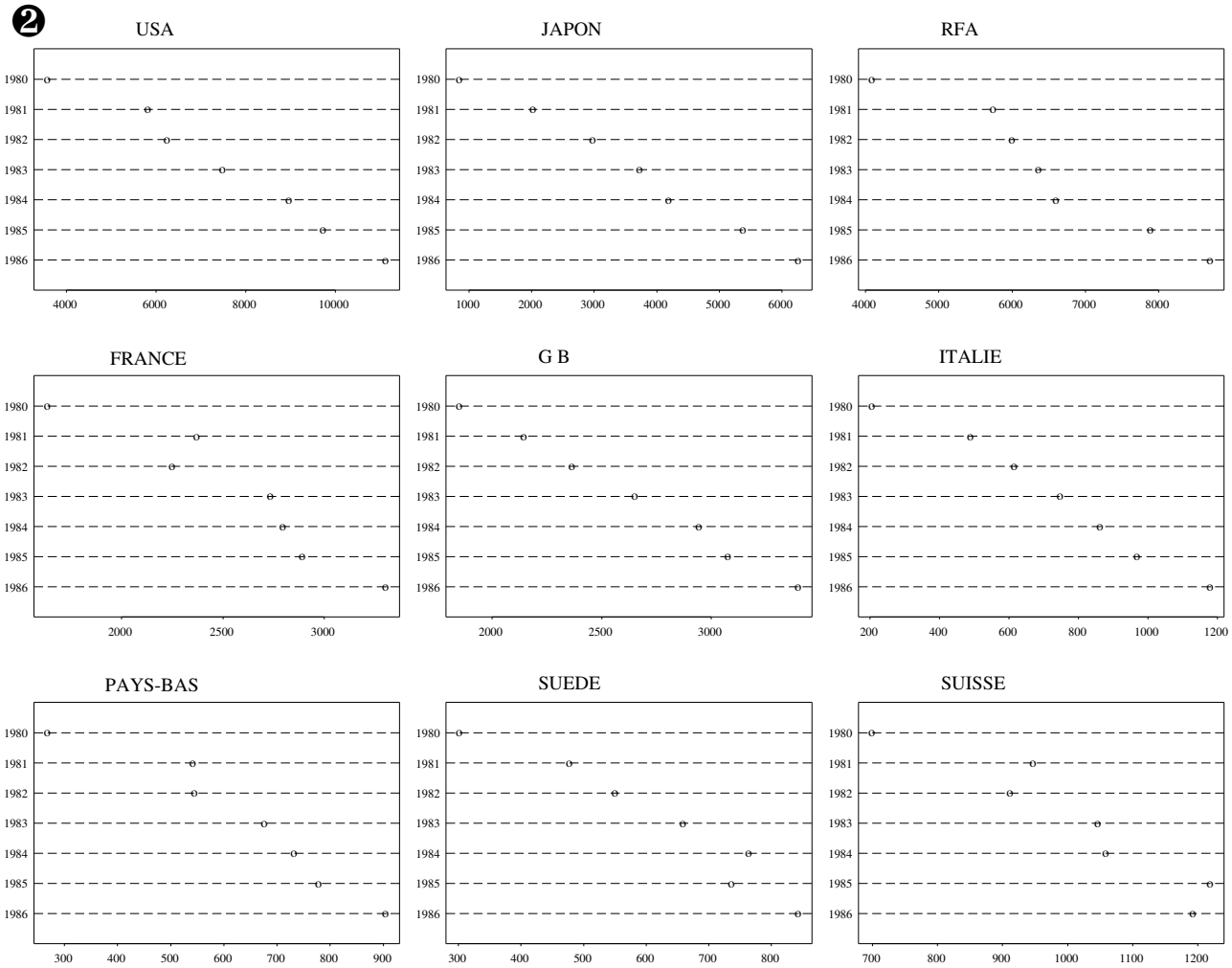
❶



Les profils des différnts pays sont représentés à la même échelle. Les noms des pays ont été ajoutés manuellement.

```
> apply(brevet,1,dotchart)
```

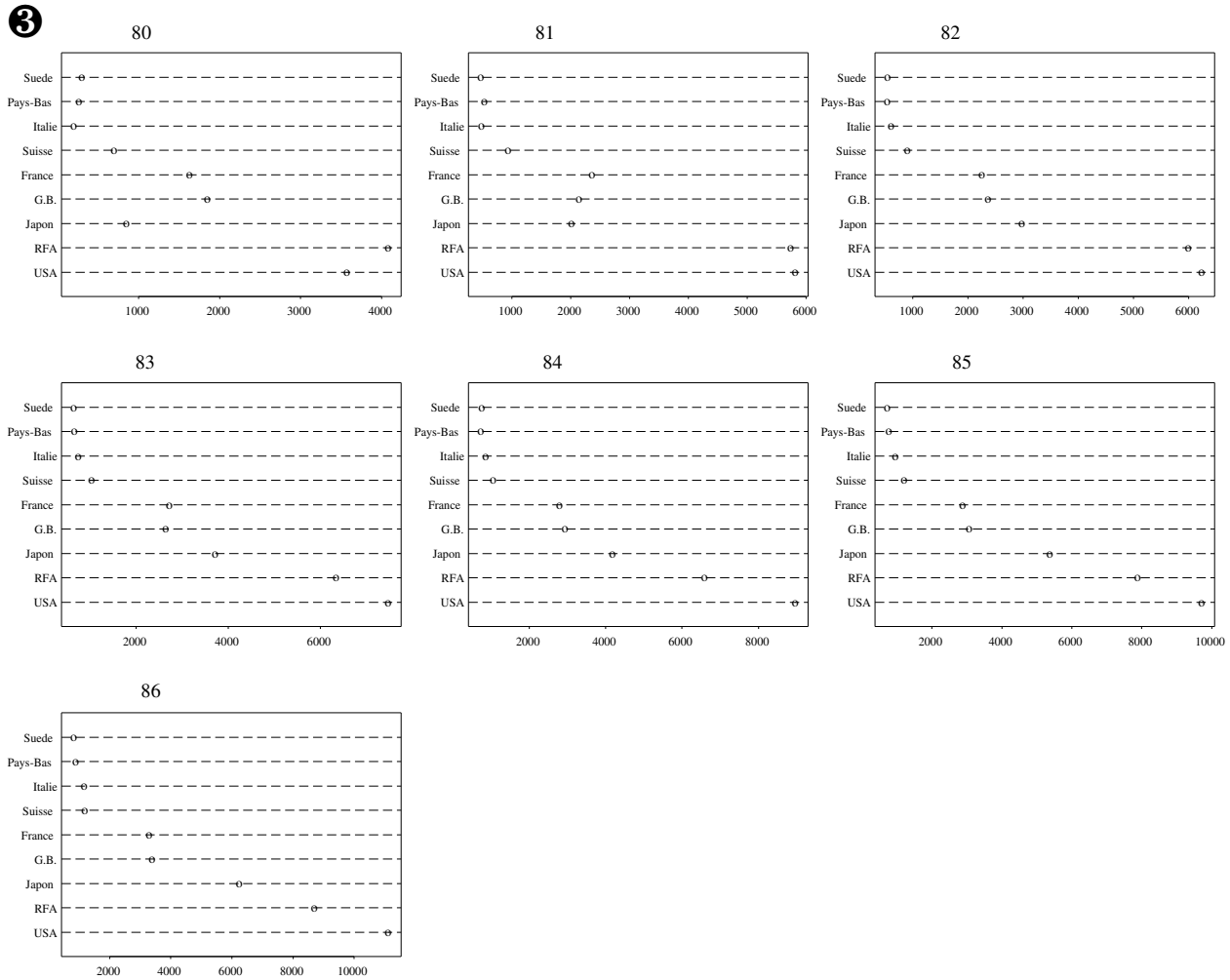
❷



L'échelle est déterminée pays par pays.

```
> total_apply(brevet,1,sum)
> apply(brevet[order(total),],2,dotchart)
```

❸

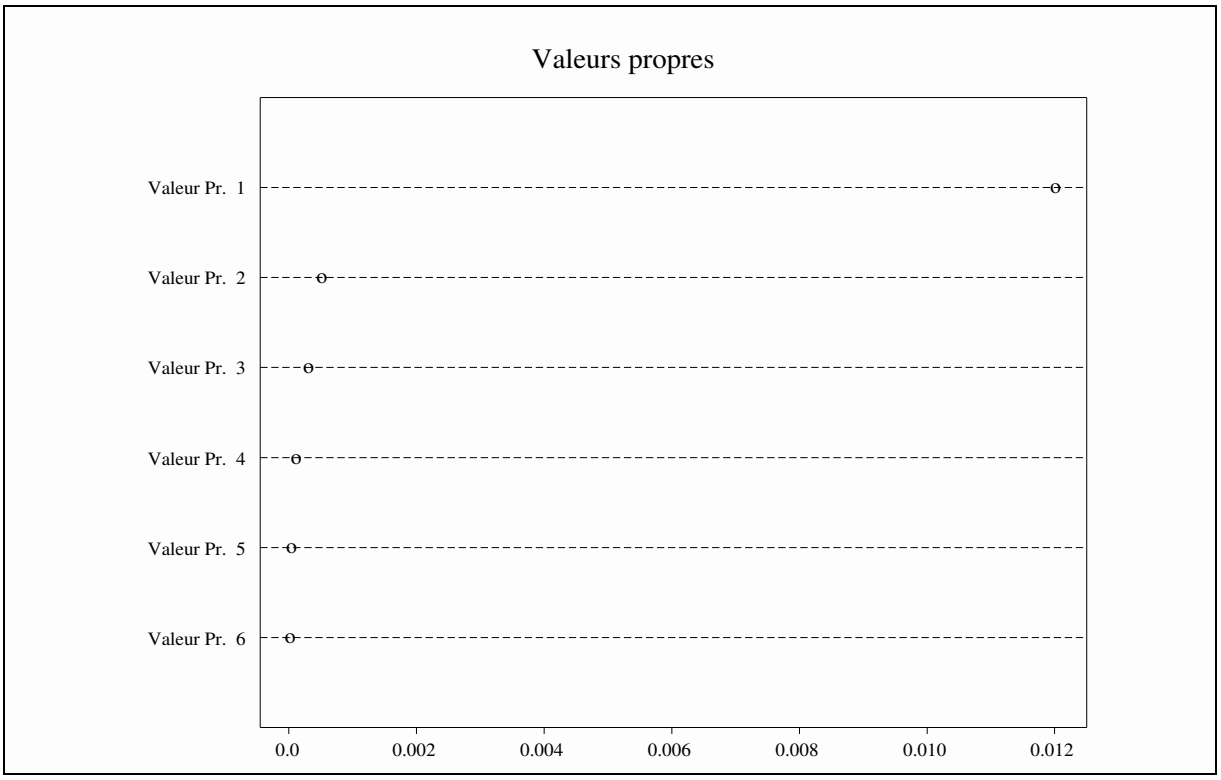


Les pays sont ordonnés en fonction du nombre total de dépots entre 80 et 86. On aurait aussi pu les ordonner en fonction du nombre de dépots en 80, ou en fonction du nombre de dépots en 86.

L'analyse

Pour réaliser une AFC, on utilise la fonction "ca"[4] ("correspondance analysis") de la librairie "afc.acp" développée par le laboratoire de Biométrie de Montpellier.

```
> brevet.ca_ca(brevet)
```



Le premier facteur absorbe presque toute l'inertie (92%). C'est donc sur cet axe qu'est résumée la plus grande part de l'information. Nous avons retenu les 2 premiers axes.

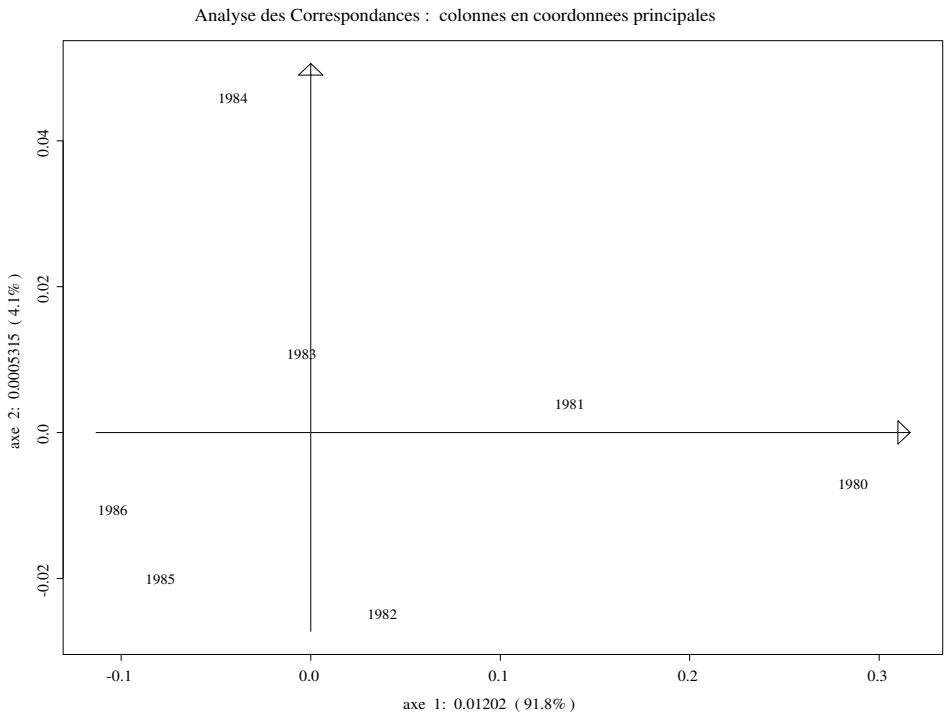
En AFC, les lignes et les colonnes jouent un rôle symétrique. On peut donc les représenter sur le même graphique.

Par défaut, la fonction caplot représente les lignes et les colonnes sur le même graphique, les lignes dans l'échelle des coordonnées standardisées (norme 1), les colonnes dans l'échelle des coordonnées principales (norme égale à la valeur propre). Les colonnes sont alors au centre de gravité des lignes. On peut aussi choisir de représenter les lignes au centre de gravité des colonnes. Le graphique obtenu par défaut dans notre exemple n'est pas lisible. On choisit donc de représenter les lignes et les colonnes sur des graphiques séparés.

[4] Toutes les sorties de la fonction "ca" apparaitront dans un encadré, sur fond grisé.

```
          COLONNES
    (cta = cta * 1000          ctr = ctr * 1000)

      Axe    1         Axe    2
      coord. cta ctr   coord. cta ctr
1980  0.2900 512 982 -0.0069   7   1
1981  0.1400 180 918  0.0040   3   1
1982  0.0390  16 627 -0.0250 144 253
1983 -0.0034   0  13  0.0110  32 138
1984 -0.0400  21 410  0.0460 636 550
1985 -0.0780  91 905 -0.0200 136  60
1986 -0.1000 181 977 -0.0110  42  10
```

> caplot(brevet.ca,rows="n",cols="p")



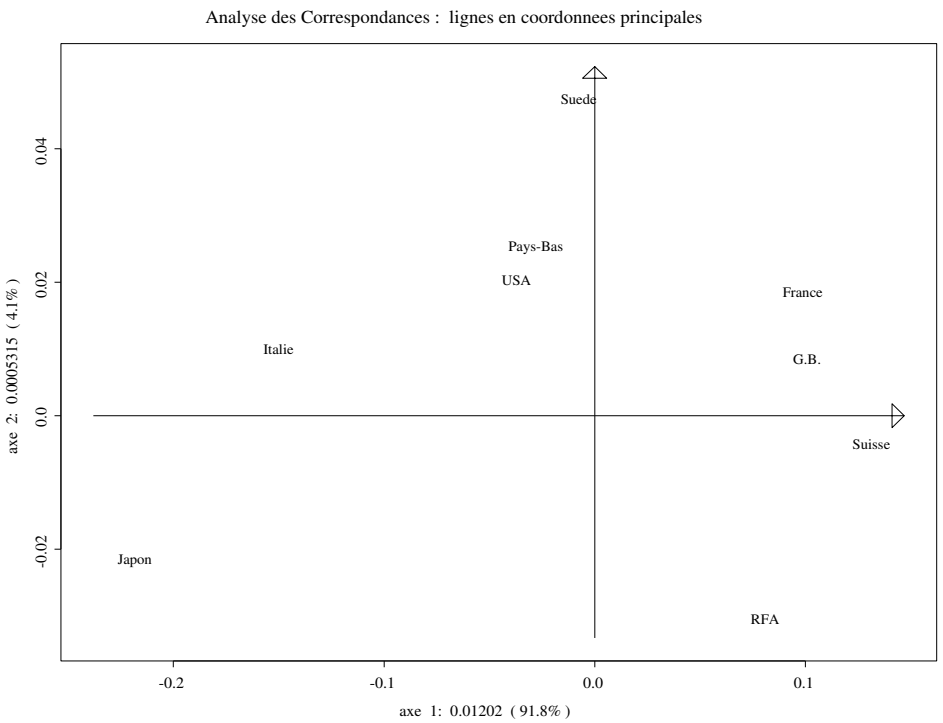
Les projections des années sur le premier axe sont ordonnées. Le premier axe traduit l'évolution moyenne du nombre de brevets de 80 à 86. Les années 80–81 et 85–86 sont bien représentées sur l'axe 1. On peut donc les interpréter en étudiant leur projection sur cet axe. L'écart entre 80 et 81, bien supérieur à l'écart entre 85 et 86 révèle une progression supérieure au début de la période étudiée.

```
             LIGNES
    (cta = cta * 1000           ctr = ctr * 1000)

      Axe    1         Axe    2
          coord. cta ctr  coord. cta ctr
     USA -0.0370  33 685  0.0200 229 213
   Japon -0.2200 546 987 -0.0210 120  10
     RFA  0.0810 138 870 -0.0300 433 120
  France  0.1000  84 900  0.0190  65  31
    G.B.  0.1000  88 842  0.0087  14   6
  Italie -0.1500  52 955  0.0100   5   4
Pays-Bas -0.0250   1 155  0.0260  30 161
   Suede -0.0055   0   8  0.0480 102 620
  Suisse  0.1300  58 939 -0.0042   1   1
```

En ce qui concerne les lignes, c'est le Japon qui contribue le plus à la construction du premier facteur. La RFA y participe également assez fortement. On trouve le Japon dans la partie négative et la RFA dans la partie positive. Le graphique 3 montre que le Japon est le pays qui a augmenté le plus rapidement le nombre de brevets déposés alors que la RFA, bien qu'ayant un nombre important de dépôts, a évolué relativement lentement.

Ce premier axe traduit donc la croissance du nombre de dépôts de brevets au cours de la période étudiée. A droite on trouve les années les plus faibles et les pays évoluant le plus lentement.

```
> caplot(brevet.ca,rows="p",cols="n")
```



Le deuxième axe a une inertie beaucoup plus faible que le premier (4%), mais certaines lignes ou colonnes y sont très bien représentées. Il est intéressant de l'étudier même si le pourcentage d'inertie expliquée est beaucoup plus faible.

1984 y est bien représentée. 1982 et 1983 également. Pour les pays, c'est surtout la Suède qui est bien représentée (ainsi que les USA et la RFA mais plus faiblement) (parmi les pays qui contribuent à la détermination de l'axe).

Les représentations graphiques faites en début d'analyse aident à interpréter le deuxième axe. L'année 1984 est une année charnière : certains pays comme la RFA et le Japon augmentent le rythme de croissance du nombre de dépôts de brevets, alors que des pays comme la Suède ou les USA voient leur vitesse de croissance diminuer (la Suède a même baissé en absolu de 84 à 85). Cette rupture en 1984 pour des pays comme la RFA, les USA, le Japon, la Suède peut être observée sur le graphique 2 (joindre les années 82, 84 et 85), elle est également "visible" sur la projection des années sur le plan 1–2 : 82, 83, 84 sont ordonnées sur le deuxième axe, 85 rompt cette progression, c'est donc entre 84 85 que le rythme de croissance est modifié.

Le plan 1–2 représente à lui seul 96% de l'inertie. Nous nous contentons donc d'étudier ces deux axes.

Analyse Factorielle des Correspondances Multiple

Un exemple

Résumé des principales commandes :

```
# codage des données en classes, puis :
> library(afc.acp)
> mamiferes.burt_burt(mamiferes3)
> mamiferes.burt_mamiferes.burt$burt.mat
> mamiferes.burt_mamiferes.burt[1:30,31:45]
> mamiferes.ca_ca(mamiferes.burt)
> caplot(mamiferes.ca,points=2)
```

Les données

Le tableau de données contient des mesures effectuées sur 62 espèces de mamifères. Les sept premières colonnes sont des variables quantitatives : poids total, poids du cerveau, durée de sommeil léger, durée de sommeil paradoxal, durée de sommeil total, durée de vie maximum et temps de gestation. Ce sont des moyennes par espèce. Les trois colonnes suivantes sont des indices de danger codés de 1 à 5 et qui représentent respectivement le danger de prédation, le danger pendant le sommeil et le danger global (le codage 1 correspondant au danger le plus faible, le codage 5 au danger le plus important).

> mamifères3

			___sommeil___				
	poids total	poids cerveau	leger	paradoxal	total	vie max	temps gestation
African elephant	6654.000	5712.00	NA	NA	3.3	38.6	645.0
African giant pouched rat	1.000	6.60	6.3	2.0	8.3	4.5	42.0
Arctic Fox	3.385	44.50	NA	NA	12.5	14.0	60.0
Arctic ground squirrel	0.920	5.70	NA	NA	16.5	NA	25.0
Asian elephant	2547.000	4603.00	2.1	1.8	3.9	69.0	624.0
Baboon	10.550	179.50	9.1	0.7	9.8	27.0	180.0
Big brown bat	0.023	0.30	15.8	3.9	19.7	19.0	35.0
Brazilian tapir	160.000	169.00	5.2	1.0	6.2	30.4	392.0
Cat	3.300	25.60	10.9	3.6	14.5	28.0	63.0
Chimpanzee	52.160	440.00	8.3	1.4	9.7	50.0	230.0
Chinchilla	0.425	6.40	11.0	1.5	12.5	7.0	112.0
Cow	465.000	423.00	3.2	0.7	3.9	30.0	281.0
Desert hedgehog	0.550	2.40	7.6	2.7	10.3	NA	NA
Donkey	187.100	419.00	NA	NA	3.1	40.0	365.0
Eastern American mole	0.075	1.20	6.3	2.1	8.4	3.5	42.0

Echidna	3.000	25.00	8.6	0.0	8.6	50.0	28.0
European hedgehog	0.785	3.50	6.6	4.1	10.7	6.0	42.0
Galago	0.200	5.00	9.5	1.2	10.7	10.4	120.0
Genet	1.410	17.50	4.8	1.3	6.1	34.0	NA
Giant armadillo	60.000	81.00	12.0	6.1	18.1	7.0	NA
Giraffe	529.000	680.00	NA	0.3	NA	28.0	400.0
Goat	27.660	115.00	3.3	0.5	3.8	20.0	148.0
Golden hamster	0.120	1.00	11.0	3.4	14.4	3.9	16.0
Gorilla	207.000	406.00	NA	NA	12.0	39.3	252.0
Gray seal	85.000	325.00	4.7	1.5	6.2	41.0	310.0
Gray wolf	36.330	119.50	NA	NA	13.0	16.2	63.0
Ground squirrel	0.101	4.00	10.4	3.4	13.8	9.0	28.0
Guinea pig	1.040	5.50	7.4	0.8	8.2	7.6	68.0
Horse	521.000	655.00	2.1	0.8	2.9	46.0	336.0
Jaguar	100.000	157.00	NA	NA	10.8	22.4	100.0
Kangaroo	35.000	56.00	NA	NA	NA	16.3	33.0
Lesser short-tailed shrew	0.005	0.14	7.7	1.4	9.1	2.6	21.5
Little brown bat	0.010	0.25	17.9	2.0	19.9	24.0	50.0
Man	62.000	1320.00	6.1	1.9	8.0	100.0	267.0
Mole rat	0.122	3.00	8.2	2.4	10.6	NA	30.0
Mountain beaver	1.350	8.10	8.4	2.8	11.2	NA	45.0
Mouse	0.023	0.40	11.9	1.3	13.2	3.2	19.0
Musk shrew	0.048	0.33	10.8	2.0	12.8	2.0	30.0
N. American opossum	1.700	6.30	13.8	5.6	19.4	5.0	12.0
Nine-banded armadillo	3.500	10.80	14.3	3.1	17.4	6.5	120.0
Okapi	250.000	490.00	NA	1.0	NA	23.6	440.0
Owl monkey	0.480	15.50	15.2	1.8	17.0	12.0	140.0
Patas monkey	10.000	115.00	10.0	0.9	10.9	20.2	170.0
Phanlanger	1.620	11.40	11.9	1.8	13.7	13.0	17.0
Pig	192.000	180.00	6.5	1.9	8.4	27.0	115.0
Rabbit	2.500	12.10	7.5	0.9	8.4	18.0	31.0
Raccoon	4.288	39.20	NA	NA	12.5	13.7	63.0
Rat	0.280	1.90	10.6	2.6	13.2	4.7	21.0
Red fox	4.235	50.40	7.4	2.4	9.8	9.8	52.0
Rhesus monkey	6.800	179.00	8.4	1.2	9.6	29.0	164.0
Rock hyrax (Hetero. b)	0.750	12.30	5.7	0.9	6.6	7.0	225.0
Rock hyrax (Procavia hab)	3.600	21.00	4.9	0.5	5.4	6.0	225.0
Roe deer	14.830	98.20	NA	NA	2.6	17.0	150.0
Sheep	55.500	175.00	3.2	0.6	3.8	20.0	151.0
Slow loris	1.400	12.50	NA	NA	11.0	12.7	90.0
Star nosed mole	0.060	1.00	8.1	2.2	10.3	3.5	NA
Tenrec	0.900	2.60	11.0	2.3	13.3	4.5	60.0
Tree hyrax	2.000	12.30	4.9	0.5	5.4	7.5	200.0
Tree shrew	0.104	2.50	13.2	2.6	15.8	2.3	46.0
Vervet	4.190	58.00	9.7	0.6	10.3	24.0	210.0
Water opossum	3.500	3.90	12.8	6.6	19.4	3.0	14.0
Yellow-bellied marmot	4.050	17.00	NA	NA	NA	13.0	38.0

	predation	danger.sommeil	danger.global
African elephant	3	5	3
African giant pouched rat	3	1	3
Arctic Fox	1	1	1
Arctic ground squirrel	5	2	3
Asian elephant	3	5	4

Baboon	4	4	4
Big brown bat	1	1	1
Brazilian tapir	4	5	4
Cat	1	2	1
Chimpanzee	1	1	1
Chinchilla	5	4	4
Cow	5	5	5
Desert hedgehog	2	1	2
Donkey	5	5	5
Eastern American mole	1	1	1
Echidna	2	2	2
European hedgehog	2	2	2
Galago	2	2	2
Genet	1	2	1
Giant armadillo	1	1	1
Giraffe	5	5	5
Goat	5	5	5
Golden hamster	3	1	2
Gorilla	1	4	1
Gray seal	1	3	1
Gray wolf	1	1	1
Ground squirrel	5	1	3
Guinea pig	5	3	4
Horse	5	5	5
Jaguar	1	1	1
Kangaroo	3	5	4
Lesser short-tailed shrew	5	2	4
Little brown bat	1	1	1
Man	1	1	1
Mole rat	2	1	1
Mountain beaver	3	1	3
Mouse	4	1	3
Musk shrew	4	1	3
N. American opossum	2	1	1
Nine-banded armadillo	2	1	1
Okapi	5	5	5
Owl monkey	2	2	2
Patas monkey	4	4	4
Phanlanger	2	1	2
Pig	4	4	4
Rabbit	5	5	5
Raccoon	2	2	2
Rat	3	1	3
Red fox	1	1	1
Rhesus monkey	2	3	2
Rock hyrax (Hetero. b)	2	2	2
Rock hyrax (Procavia hab)	3	2	3
Roe deer	5	5	5
Sheep	5	5	5
Slow loris	2	2	2
Star nosed mole	3	1	2
Tenrec	2	1	2
Tree hyrax	3	1	3
Tree shrew	3	2	2
Vervet	4	3	4
Water opossum	2	1	1

```
Yellow-bellied marmot         3              1              1
```

Le choix de l'analyse

Le but de l'étude est l'analyse des relations entre les indices de danger et les six variables quantitatives continues qui caractérisent les espèces.

L'étude simultanée des variables quantitatives continues et des indices de danger nécessite la codification des variables continues en classes. Pour étudier les liaisons entre deux ensembles de variables, les indices de danger, et les autres variables, on met en oeuvre une AFC multiple portant sur le tableau de contingence obtenu en croisant les deux ensembles de variables. Ce tableau est un sous-tableau du tableau de burt.

La préparation des données

```
> mamiferes3_mamiferes3[,c(1:3,5:10)]
> mamiferes3_na.omit(mamiferes3)
> attach(mamiferes3)
> mamiferes3[,"poids.cerveau"]_poids.cerveau/poids.tot
> mamiferes3[,"sommeil.leger"]_sommeil.leger/sommeil.tot
> detach()
```

La variable "sommeil.paradoxal" est supprimée, les durées de sommeil total et de sommeil léger déterminant complètement la durée de sommeil paradoxal. Les espèces pour lesquelles certaines données manquent sont supprimées. La variable "poids.cerveau" contient les poids relatifs du cerveau. La variable "sommeil.leger" contient la part relative du sommeil consacrée au sommeil léger.

On découpe en classes chacune des variables quantitatives. Les indices de danger étant codés de 1 à 5, on fait un découpage en 5 classes. Les nombres de modalités seront ainsi identiques d'une variable à l'autre. Il est important qu'ils ne soient pas trop différents d'une variable à l'autre, sous peine de donner un plus grand poids à certaines variables. Le nombre de modalités doit être suffisamment grand pour que l'interprétation soit intéressante, il ne doit pas être trop grand pour que l'interprétation ne soit pas trop complexe. Les effectifs des différentes modalités doivent être équilibrés, pour donner la même importance aux différentes modalités. Pour déterminer les bornes des classes, on fait une représentation en "branches et feuilles" de chacune des variables.

```
>stem(mamiferes3[,1])

N = 42   Median = 0.804719
Quartiles = -1.272966, 2.356126

Decimal point is at the colon

  -5 : 3
  -4 : 6
  -3 : 880
  -2 : 6331
  -1 : 63
```

```
-0 : 97321
 0 : z05579
 1 : 12333449
 2 : 34
 3 : 3
 4 : 0014
 5 : 13
 6 : 13
 7 : 8
```

On peut obtenir les 6 diagrammes en “branches et feuilles” en une seule commande.

```
> apply(mamiferes3[,1:6],2,stem)
```

Chacune des variables est découpée en 5 classes codées de 1 à 5.

```
> mamiferes3[,1]_cut(mamiferes3[,1],breaks=c(-6,-2,0,1.4,3.5,8))
> mamiferes3[,2]_cut(mamiferes3[,2],breaks=c(0,4,7,10,20,50))
> mamiferes3[,3]_cut(mamiferes3[,3],
    breaks=c(0.5,0.74,0.79,0.85,0.9,1.1))
> mamiferes3[,4]_cut(mamiferes3[,4],breaks=c(0,6,9,11,14,20))
> mamiferes3[,5]_cut(mamiferes3[,5],breaks=c(0,5,10,21,31,101))
> mamiferes3[,6]_cut(mamiferes3[,6],breaks=c(0,30,50,100,200,630))
```

On construit le tableau de Burt. C’est un tableau de contingence multiple ayant en lignes et en colonnes toutes les modalités des variables. Les différentes modalités sont toutes croisées deux à deux. Pour cela, on utilise la fonction “burt” de la librairie “afc.acp” développée par le laboratoire de Biométrie de Montpellier.

```
> library(afc.acp)
> mamiferes.burt_burt(mamiferes3)
```

Les noms des modalités sont générés automatiquement : “1a” siginfie “poids total faible”, “1e” signifie “poids total élevé” ...

Le tableau initial contenant 9 variables, chaque variable étant découpée en 5 modalités, le tableau de Burt est un tableau 45 x 45.

```
> mamiferes.burt_mamiferes.burt$burt.mat
> mamiferes.burt
```

```
    1a 1b 1c 1d 1e 2a ...8c 8d 8e 9a 9b 9c 9d 9e
1a   9  0  0  0  0  0      0  0  0  3  2  3  1  0
1b   0  8  0  0  0  1      0  1  0  0  5  2  1  0
1c   0  0 10  0  0  3      1  0  1  4  2  2  1  1
1d   0  0  0  6  0  0      2  2  1  1  1  0  3  1
1e   0  0  0  0  9  7      1  1  5  3  0  0  3  3
2a   0  1  3  0  7 11      1  1  5  4  1  0  3  3
.                                               .
.                                               .
.                                               .
8c   0  0  1  2  1  1      4  0  0  1  1  0  2  0
8d   0  1  0  2  1  0      0  4  0  0  0  0  4  0
8e   0  0  1  1  5  0      0  0  7  0  0  0  2  5
9a   3  0  4  1  3  8      1  0  0 11  0  0  0  0
9b   2  5  2  1  0  0      1  0  0  0 10  0  0  0
9c   3  2  2  0  0  0      0  0  0  0  0  7  0  0
9d   1  1  1  3  3  0      2  4  2  0  0  0  9  0
9e   0  0  1  1  3  0 ...  0  0  5  0  0  0  0  5
```

L'objectif de l'étude étant d'étudier les relations entre les indices de danger et les autres variables, on sélectionne le sous-tableau intéressant.

```
> mamiferes.burt_mamiferes.burt[1:30,31:45]
> mamiferes.burt
```

```
    7a 7b 7c 7d 7e 8a 8b 8c 8d 8e 9a 9b 9c 9d 9e
1a   3  0  2  2  2  7  2  0  0  0  3  2  3  1  0
1b   0  5  2  0  1  3  4  0  1  0  0  5  2  1  0
1c   1  5  2  0  2  5  3  1  0  1  4  2  2  1  1
1d   1  1  0  3  1  1  0  2  2  1  1  1  0  3  1
1e   3  0  1  2  3  2  0  1  1  5  3  0  0  3  3
2a   1  4  1  2  3  4  0  1  1  5  4  1  0  3  3
2b   0  1  4  1  3  4  2  1  0  2  0  1  5  1  2
2c   2  2  1  0  0  3  2  0  0  0  2  3  0  0  0
2d   3  1  0  4  1  4  1  1  3  0  3  1  1  4  0
2e   2  3  1  0  2  3  4  1  0  0  2  4  1  1  0
3a   0  3  1  0  1  2  1  0  0  2  2  1  0  1  1
3b   5  0  2  1  1  6  1  1  1  0  5  1  2  1  0
3c   1  2  2  2  3  5  2  0  0  3  2  2  2  2  2
3d   2  5  0  0  3  3  3  1  1  2  2  5  0  1  2
3e   0  1  2  4  1  2  2  2  2  0  0  1  3  4  0
4a   0  0  3  0  4  1  1  0  0  5  0  0  2  1  4
4b   3  2  1  2  2  3  2  2  1  2  3  2  1  3  1
4c   2  3  0  3  1  2  3  2  2  0  2  3  0  4  0
4d   0  2  1  2  2  6  0  0  1  0  0  2  4  1  0
```

```
4e  3  4  2  0  0  6  3  0  0  0  6  3  0  0  0
5a  1  3  4  2  1  9  2  0  0  0  3  3  4  1  0
5b  1  3  2  0  3  4  3  1  1  0  2  2  3  2  0
5c  1  3  0  1  3  2  2  0  1  3  1  3  0  1  3
5d  2  1  0  4  1  1  1  2  2  2  2  1  0  4  1
5e  3  1  1  0  1  2  1  1  0  2  3  1  0  1  1
6a  0  4  2  2  2  8  2  0  0  0  2  3  4  1  0
6b  3  1  2  0  1  4  2  0  0  1  3  2  1  0  1
6c  2  1  0  0  1  2  1  1  0  0  2  1  0  1  0
6d  0  4  1  3  3  2  2  1  4  2  1  3  1  4  2
6e  3  1  2  2  2  2  2  2  0  4  3  1  1  3  2
```

On donne des noms plus explicites aux différentes modalités (Dp pour danger de prédation, Ds pour danger pendant le sommeil, D pour danger global).

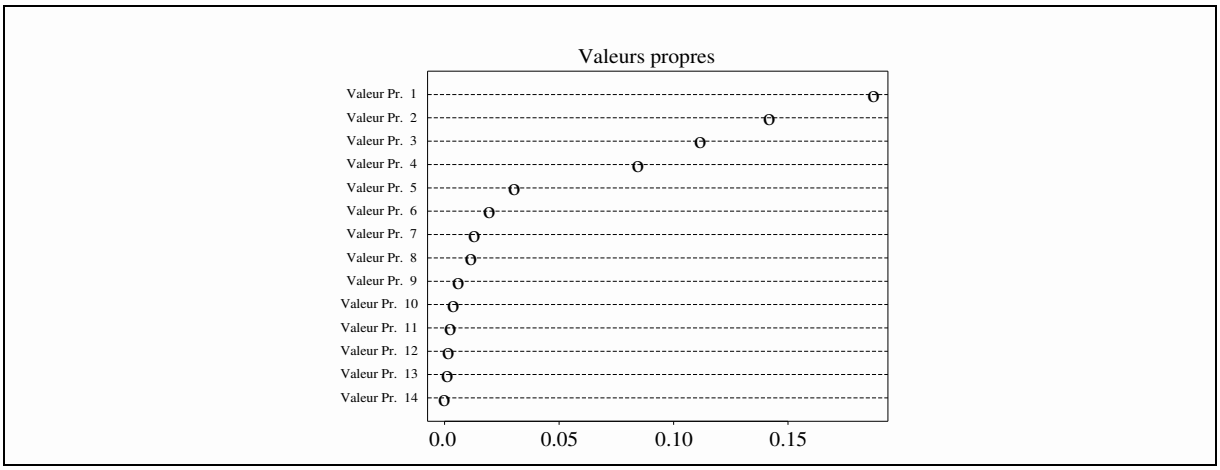
```
>nc_c(paste("Dp",1:5,sep=""),paste("Ds",1:5,sep=""),
     paste("D",1:5,sep=""))
>nl_c(paste("poids",1:5,sep=""),paste("cerv",1:5,sep=""),
     paste("som.l",1:5,sep=""),paste("som.t",1:5,sep=""),
     paste("d.vie",1:5,sep=""),paste("t.gest",1:5,sep=""))
>dimnames(mamiferes.burt)_list(nl,nc)
```

L'analyse

Visualisation des résultats

On utilise la fonction "ca" (correspondance analysis) de la librairie "afc.acp" , pour réaliser une AFC multiple.

```
> mamiferes.ca_ca(mamiferes.burt)
```



On retient les 4 premiers facteurs car il y a une chute d'inertie entre le quatrième et le cinquième axe.

COLONNES
(cta = cta * 1000 ctr = ctr * 1000)

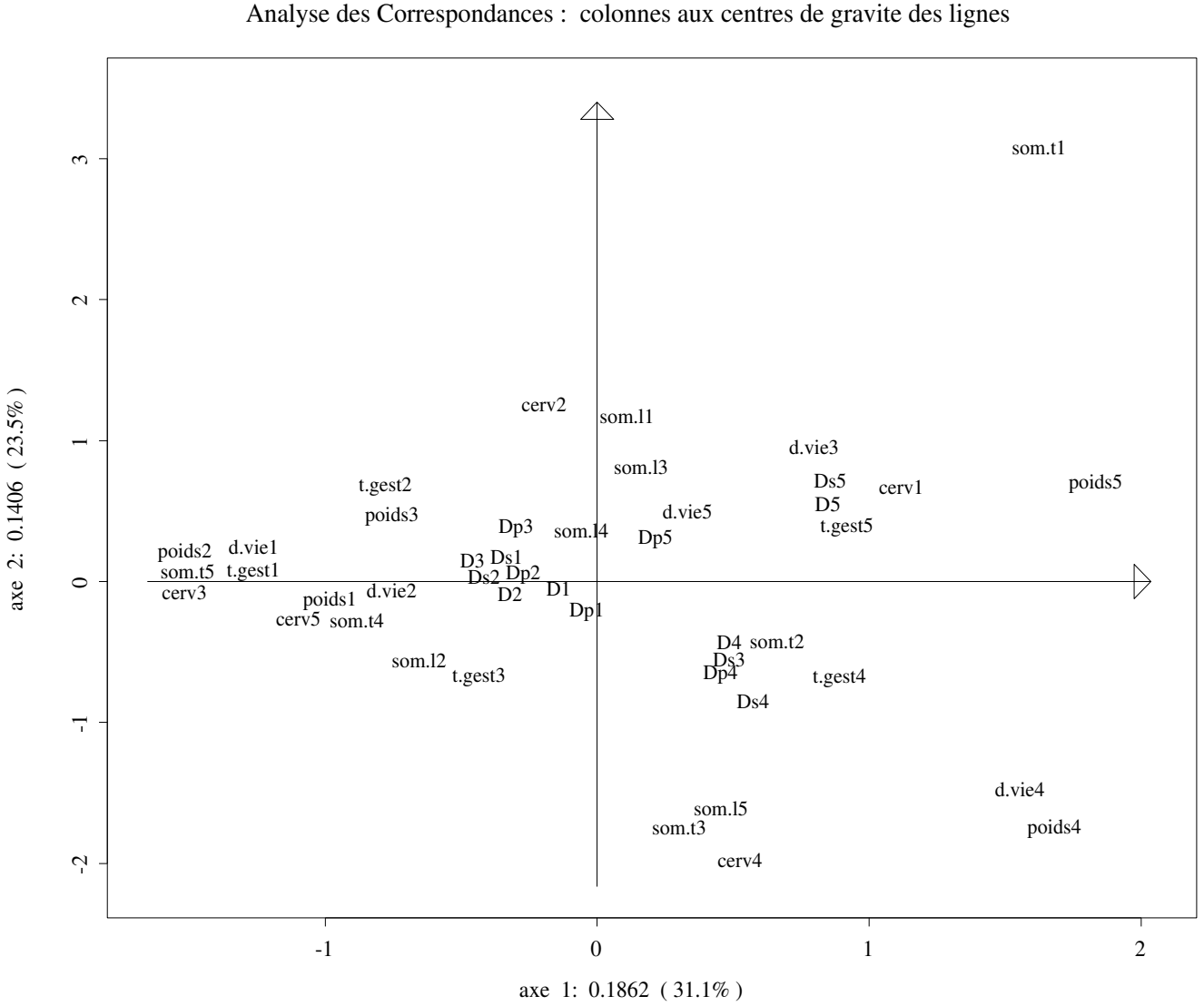
	Axe 1			Axe 2			Axe 3			Axe 4		
	coord.	cta	ctr	coord.	cta	ctr	coord.	cta	ctr	coord.	cta	ctr
Dp1	-0.03	**0**	**1**	-1.9e-01	17	**51**	0.660	248	**594**	-0.430	141	**253**
Dp2	-0.27	**34**	170	2.3e-02	0	1	0.082	5	15	0.540	304	671
Dp3	-0.29	**25**	159	4.0e-01	64	305	-0.290	42	160	-0.260	45	127
Dp4	0.46	**64**	240	-6.4e-01	160	455	-0.380	72	160	-0.190	23	38
Dp5	0.22	**19**	184	3.3e-01	54	396	-0.160	17	101	0.068	4	17
Ds1	-0.40	**126**	656	2.6e-02	1	3	-0.059	4	14	-0.200	67	158
Ds2	-0.34	**46**	350	-1.8e-05	0	0	0.100	6	29	0.290	73	253
Ds3	0.50	**42**	267	-5.4e-01	67	321	0.095	3	10	-0.058	1	4
Ds4	0.58	**58**	235	-8.4e-01	159	489	-0.270	21	50	0.300	34	62
Ds5	0.87	**224**	570	7.2e-01	207	398	0.120	8	11	-0.004	0	0
D1	-0.14	**9**	46	-4.3e-02	1	5	0.520	215	670	-0.260	69	162
D2	-0.38	**61**	364	-6.6e-02	2	11	0.110	8	29	0.420	168	451
D3	-0.45	**61**	199	1.6e-01	10	23	-0.820	340	656	-0.310	63	91
D4	0.49	**92**	515	-4.2e-01	91	382	-0.130	10	34	0.020	0	1
D5	0.81	**139**	500	7.7e-01	167	454	0.017	0	0	0.120	7	11

LIGNES
(cta = cta * 1000 ctr = ctr * 1000)

	Axe 1			Axe 2			Axe 3			Axe 4		
	coord.	cta	ctr	coord.	cta	ctr	coord.	cta	ctr	coord.	cta	ctr
poids1	-0.420	**33**	**371**	-0.0430	0	**4**	-0.080	2	**14**	-0.4700	95	**475**
poids2	-0.580	**58**	378	0.0044	0	0	-0.200	12	46	0.6700	170	498
poids3	-0.320	**22**	405	0.1800	9	131	0.093	3	34	0.1200	7	58
poids4	0.730	**69**	525	-0.6500	71	411	-0.096	2	9	0.0470	1	2
poids5	0.800	**122**	743	0.2700	18	85	0.220	16	57	-0.2900	36	96
cerv1	0.490	**56**	520	0.2500	20	141	0.140	8	42	0.0570	2	7
cerv2	-0.078	**1**	8	0.4700	57	292	-0.640	132	530	-0.2100	18	55
cerv3	-0.650	**45**	526	-0.0260	0	1	0.520	48	333	0.1600	6	32
cerv4	0.230	**10**	75	-0.7400	138	755	-0.013	0	0	-0.2100	20	63
cerv5	-0.440	**33**	426	-0.0390	0	3	0.220	14	109	0.2900	33	189
som.l1	0.054	**0**	5	0.4400	28	363	0.210	8	83	0.3100	24	182
som.l2	-0.280	**15**	115	-0.2100	11	65	0.300	29	135	-0.6300	169	596
som.l3	0.077	**1**	35	0.3100	27	567	-0.130	7	109	-0.0530	1	17
som.l4	-0.018	**0**	1	0.1400	5	46	0.290	31	204	0.5100	123	620
som.l5	0.200	**7**	44	-0.6000	81	381	-0.660	127	468	-0.0580	1	4
som.t1	0.710	**75**	245	1.2000	265	654	-0.370	34	67	-0.0900	3	4
som.t2	0.190	**8**	324	-0.1600	7	210	0.150	9	206	-0.0690	2	41
som.t3	0.140	**4**	35	-0.6500	108	768	0.110	4	21	0.2600	30	127
som.t4	-0.370	**21**	159	-0.1000	2	11	-0.720	131	593	-0.1100	4	14
som.t5	-0.620	**73**	482	0.0010	0	0	0.570	105	414	-0.0310	0	1
d.vie1	-0.570	**75**	638	0.0520	1	5	-0.280	30	153	-0.2200	25	94
d.vie2	-0.320	**20**	369	-0.0220	0	2	-0.190	12	136	0.0830	3	25
d.vie3	0.350	**21**	213	0.3600	29	228	0.120	4	26	0.4400	75	341
d.vie4	0.680	**78**	581	-0.5500	68	384	0.041	0	2	-0.0460	1	3
d.vie5	0.150	**3**	45	0.1900	6	72	0.580	73	679	-0.2500	19	130
t.gest1	-0.540	**61**	578	0.0350	0	3	-0.370	48	271	-0.0069	0	0
t.gest2	-0.330	**16**	261	0.2600	14	167	0.340	29	280	-0.2500	21	155
t.gest3	-0.180	**3**	51	-0.2500	7	96	0.580	49	544	-0.1400	4	31
t.gest4	0.390	**36**	301	-0.2500	19	120	-0.250	24	117	0.4100	87	323
t.gest5	0.400	**35**	473	0.1500	7	67	0.170	10	79	-0.2100	21	125

Les graphiques

```
> caplot(mamiferes.ca,points=2)
```



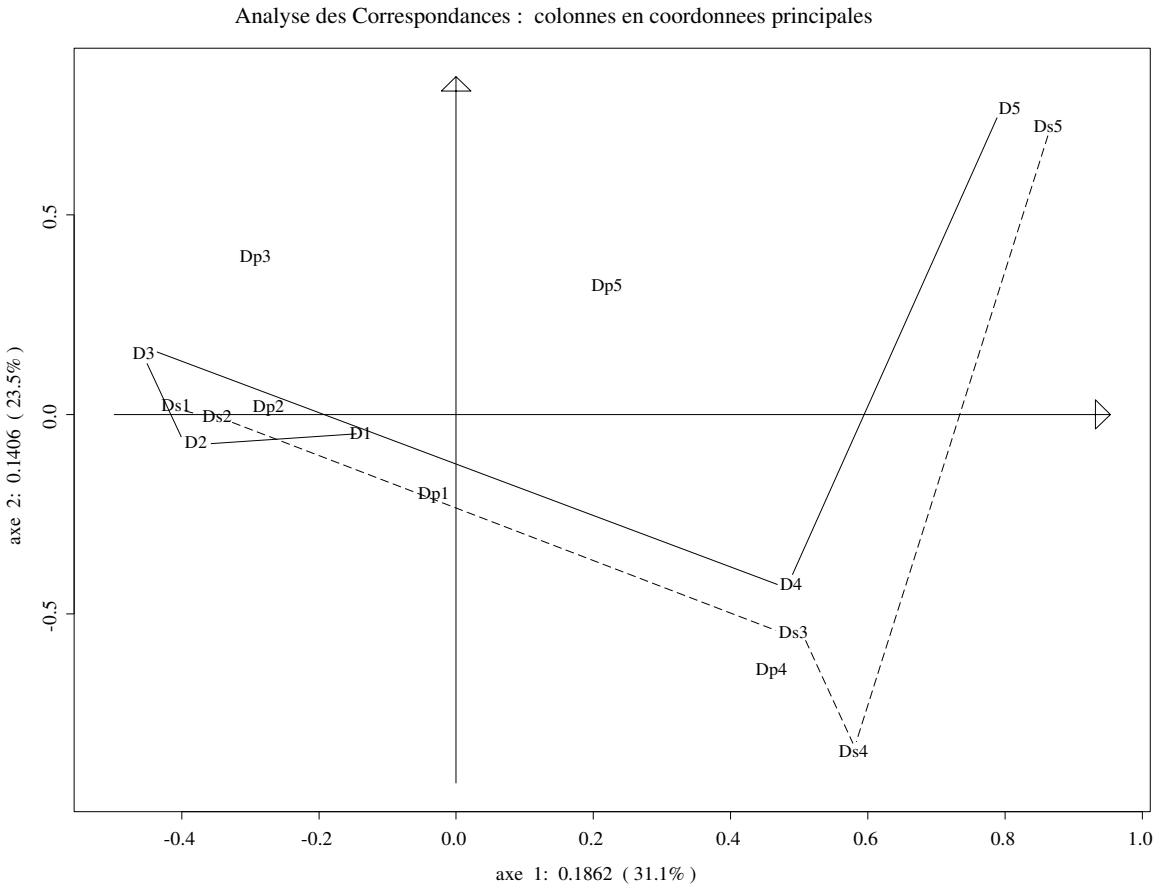

```
$ptvus:
"Ds5"     "t.gest1" "Dp2"     "som.t5"  "poids1"  "D4"

$ptcaches:
"D5"     "d.vie1" "Ds1"    "poids2" "cerv5"  "som.t2"
```

Les points cachés ont été reportés manuellement sur le graphique.

L'interprétation

Les 4 premiers axes factoriels expliquent 85% de l'inertie totale.

Les **modalités-colonne** contribuant le plus à la construction du premier axe sont celles dont la valeur "cta" est élevée, Ds5, D5 et Ds1. Les modalités de Dp contribuent très peu à la détermination de l'axe, et n'y sont pas bien représentées (ctr faibles). Cet axe oppose essentiellement les dangers de sommeil extrêmes, très faible et très fort, le danger global très fort étant associé au danger de sommeil très fort.
Le deuxième axe est essentiellement déterminé par les modalités DS5, DS4, D5, Dp4. C'est un axe qui oppose danger de sommeil fort et très fort, le danger de prédation fort étant associé au danger de sommeil fort, alors que le danger global très fort est associé au danger de sommeil très fort.

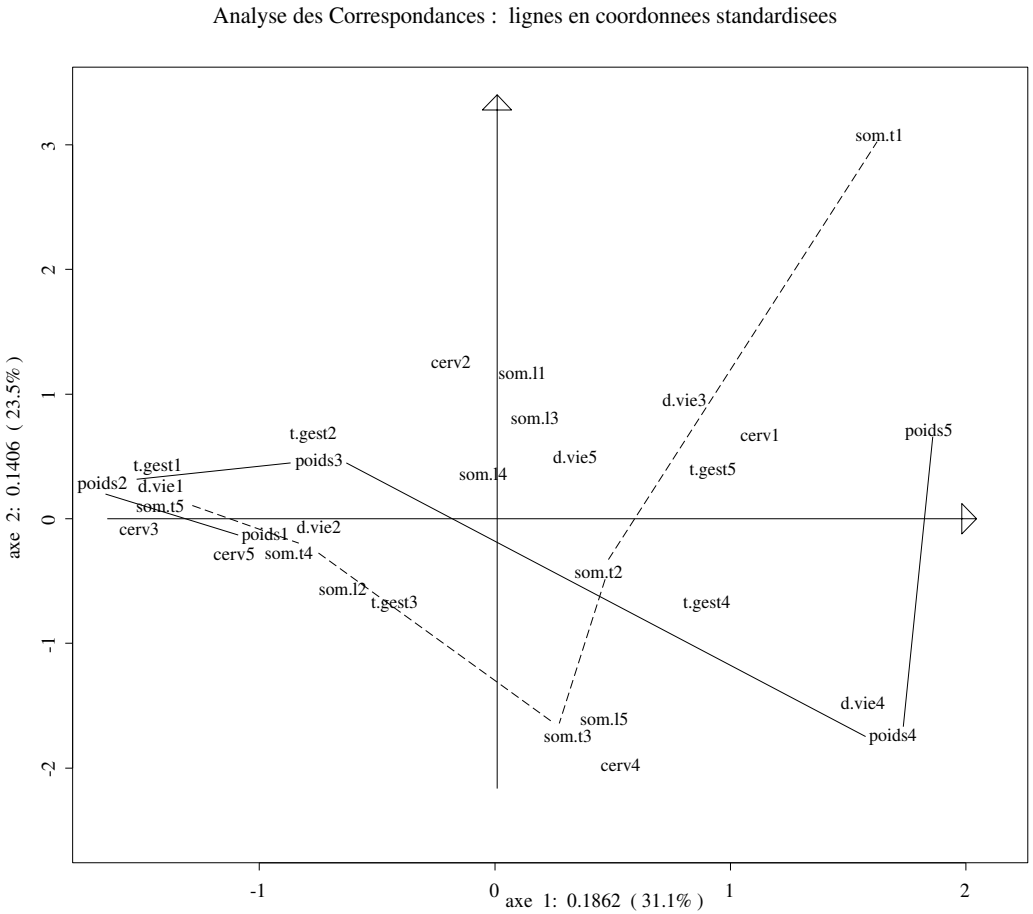


A eux seuls, ces 2 axes expliquent 54% de l'inertie totale. Le 1er plan permet de séparer les dangers très faibles, faibles et moyens (à gauche), les dangers forts (en bas à droite) et les dangers très forts (en haut à droite). Seul DS3, le danger sommeil moyen est associé à des dangers forts; Un danger moyen pendant le sommeil est en fait un danger fort. Dp5, le danger prédation très fort se situe au centre de gravité des 3 groupes, il se distingue de Ds5 et D5. Bien qu'associé à D5 et Ds5, il s'en distingue nettement. Sur le plan (1,2), les lignes brisées liant les modalités de D et de Ds sont parallèles, excepté le décalage de Ds3 vers Ds4.

Le pourcentage d'inertie expliqué par le 3ème axe est du même ordre de grandeur que celui du 2ème axe. Le 3ème axe traduit l'opposition entre le danger global très faible et moyen, et l'association du danger global très faible au danger prédation très faible.

Sur le plan (1,3) on observe le parallélisme du passage des modalités très faible, faible, moyen à fort pour D et Dp (voir le graphique plus loin). Seul Dp5 rompt cette cohérence. Le plan (1,3) permet d'étudier les dangers très faibles et très forts, ainsi que la progression très faible-faible-moyen. Le plan (3,4) est le plan idéal pour étudier cette progression.

On procède de la même façon pour étudier les **modalités-ligne**. Le 1er axe est essentiellement déterminé par poids5, et dans une moindre mesure, par l'opposition entre som.t1 et som.t5, l'opposition entre d.vie1 et d.vie4, par t.gest1, ... Sur le 1er axe, les modalités de poids, t.gest, d.vie sont positionnées par ordre croissant.



Ce 1er axe révèle les corrélations positives entre poids, durée de vie et temps de gestation, et la corrélation négative entre ces variables et le sommeil total.

L'interprétation simultanée des modalités-ligne et des modalités-colonne permet de mettre en évidence la liaison entre niveau de danger et poids, t.gest, d.vir, som.t : les espèces dont les niveaux de danger sont fort ou très fort, sont plutôt des espèces ayant un poids, un temps

de gestation et une durée de vie élevées (mais pas une durée de vie exceptionnellement élevée) et un sommeil total faible.

C'est le 2ème axe qui oppose les niveaux de danger fort et très fort. C'est donc sur cet axe qu'il faut rechercher les modalités qui pourraient être liées à cette différenciation. C'est essentiellement la modalité som.t1 qui contribue à l'explication de cet axe, suivie par cerv4, som.t3, som.l5, poids4, d.vie4, ...

som.t1, som.t2,som.t3 sont ordonnées sur l'axe 2. Le sommeil total très faible est caractéristique d'un danger très fort, alors qu'un sommeil total moyen est lié à un danger fort, le sommeil total faible est intermédiaire. (se reporter au tableau de contingence analysé). Ce 2ème axe est plus difficile à interpréter que le premier, bien qu'il explique 23% de l'inertie, ce qui n'est pas négligeable par rapport au 1er axe.

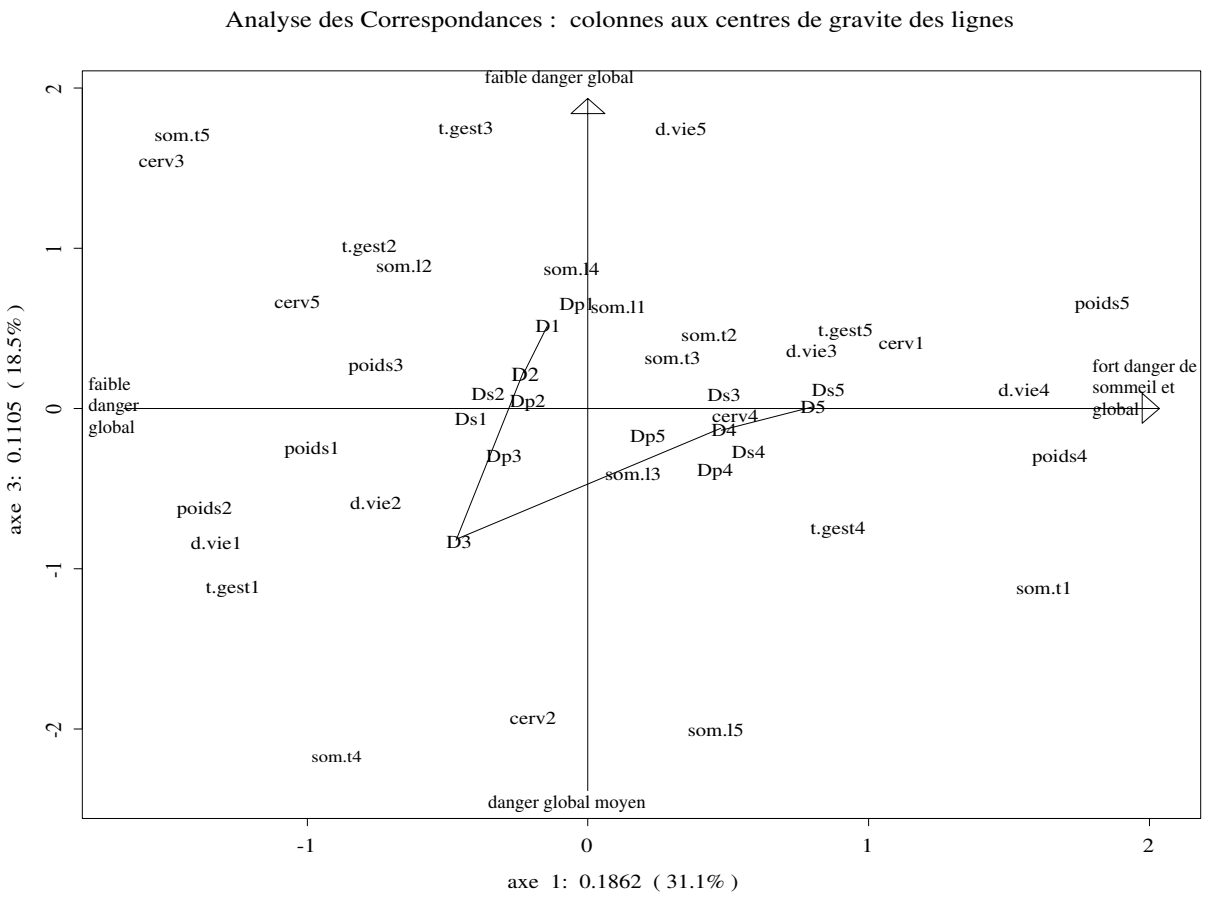
Le 3ème axe permet de différencier les niveaux très faible, faible et moyen du danger global et du danger de prédation. Cette différentiation est à lier à une opposition entre som.t4 et som.t5 : un niveau de danger très faible est lié à un temps de sommeil très élevé, un niveau de danger moyen est lié à un temps de sommeil élevé, une part relative du sommeil léger élevée, un poids relatif du cerveau faible.

L'axe 4 oppose les dangers très faibles et faibles.

Le premier axe est facile à interpréter, et pourtant, il est déterminé par un grand nombre de modalités. Au contraire, sur les axes 2, 3 et 4, quelques modalités se distinguent nettement des autres par leur "cta" particulièrement élevées. Le 1er axe traduit une évolution globale et concordante des variables, alors que les suivants traduisent une opposition entre quelques modalités.

Ici, chacun des trois indices de danger n'est pas entièrement représenté sur le même axe. Si on s'intéresse à un danger en particulier, il faut déterminer le meilleur plan de visualisation. Prenons l'exemple du danger global : pour avoir une représentation la plus complète possible des 5 niveaux de danger il sera intéressant de se reporter aux axes 1 et 3, c'est à dire au plan 1–3.

```
> caplot(mamiferes.ca,axes=c(1,3),points=2)
```



On a ajouté un nom aux axes à partir des interprétations faites précedemment. On a également relié certains points et ajouté le point caché D2.

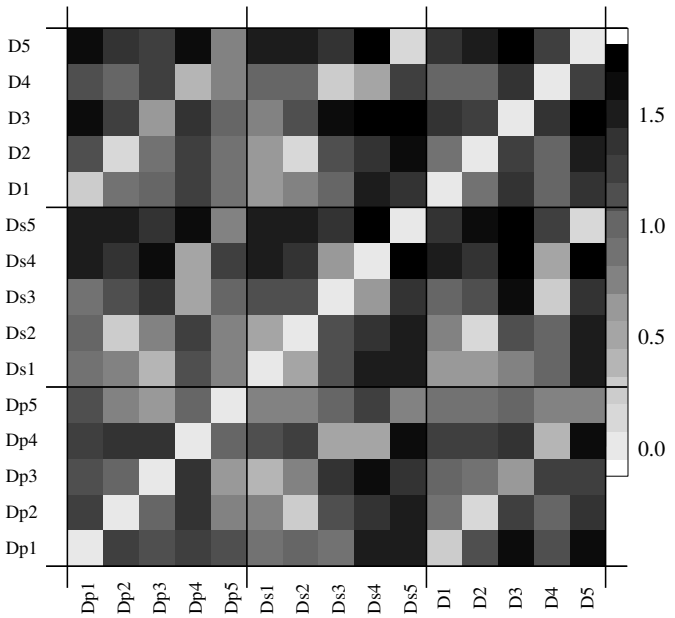
En reliant les différents points du danger global on visualise mieux leur évolution sur le graphique en fonction du niveau de danger.

Pour aller plus loin...

On conserve les coordonnées des modalités-colonne sur les 4 premiers axes. On calcule la matrice des distances euclidiennes entre ces modalités, que l'on représente graphiquement.

```
> coord.col_mamiferes.ca$var.cords[,1:4]
> noms.col_dimnames(coord.col)[[1]]
> dist.col_dist2full(dist(coord.col))
> dimnames(dist.col)_list(noms.col,noms.col)
> dist.gris(dist.col)
> abline(h=seq(1,by=5,length=4),lwd=2)
> abline(v=seq(1,by=5,length=4),lwd=2)
```

Matrice des Distances



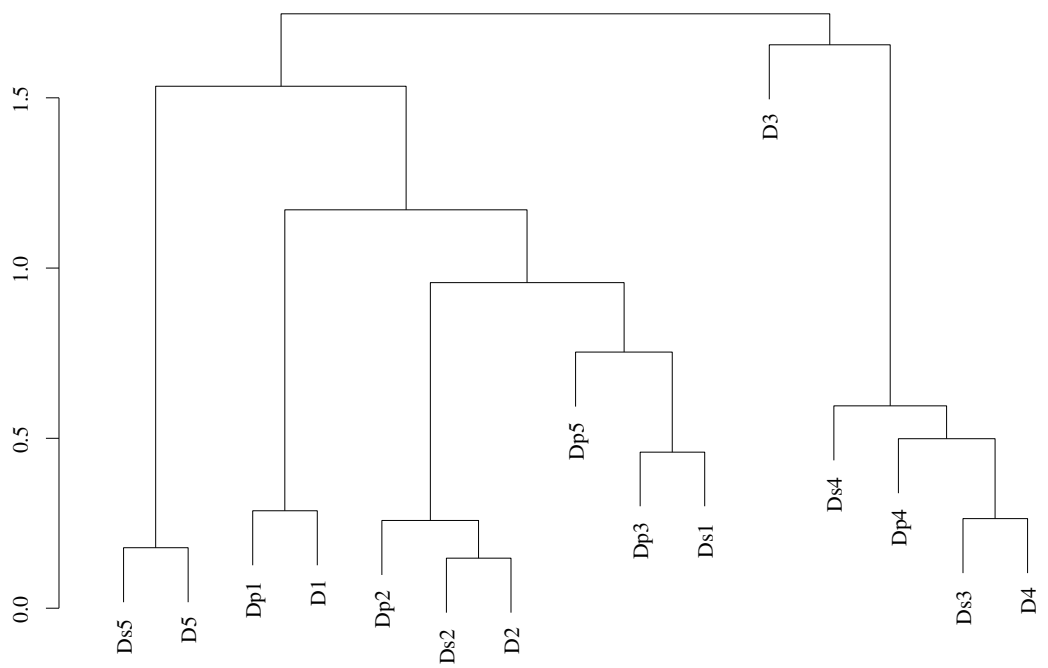
La représentation en grisé de cette matrice de distances, associée à l'examen de la répartition des modalités de danger sur le 1er plan montre que :

- Ds1 et Ds2 sont proches, Ds3 et Ds4 sont proches également,
- D1, D2, D3, D4, bien que différents, sont relativement proches les uns des autres, et éloignés de D5,
- les différentes modalités de Dp sont toutes différentes les unes des autres, il n'y a pas de regroupement possible; cependant, aucune d'entre elle n'est très éloignée d'une autre.

Ces remarques mettent en évidence un bon choix de l'échelle des indices de prédation, un choix de l'échelle des indices de sommeil tel que les indices 1 et 2 sont très (trop ?) proches l'un de l'autre, de même que 3 et 4, et un choix de l'échelle des indices de danger global et de danger de sommeil, tel qu'il existe une très grande différence entre les niveaux 1 à 4 d'une part, et 5 d'autre part. Globalement, il existe une bonne cohérence entre les 3 indices (parallélisme relatif des lignes brisées sur le 1er plan). Le 1er plan révèle une assez bonne cohérence entre danger global et danger sommeil, excepté pour Ds3 qui traduit plutôt un danger fort.

Voici sous forme de dendogramme, le résultat de la classification hiérarchique ascendante obtenue à partir de cette matrice de distance, en choisissant le lien moyen comme méthode d'agrégation.

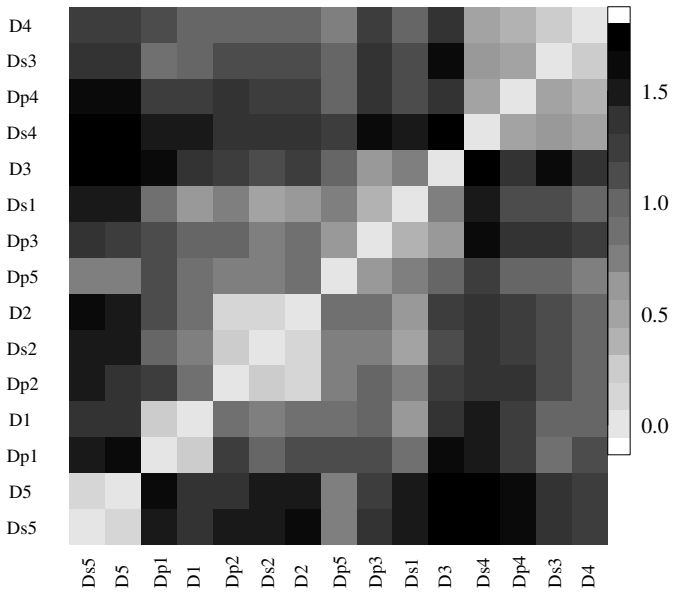
```
> hc.col_hclust(dist.col)
> plclust(hc.col,labels=noms.col)
```



On présente graphiquement la matrice des distances, dont les lignes et les colonnes sont réordonnées en fonction du résultat de la classification hiérarchique.

```
> dist.gris(dist.col[hc.col$order,hc.col$order])
```

Matrice des Distances

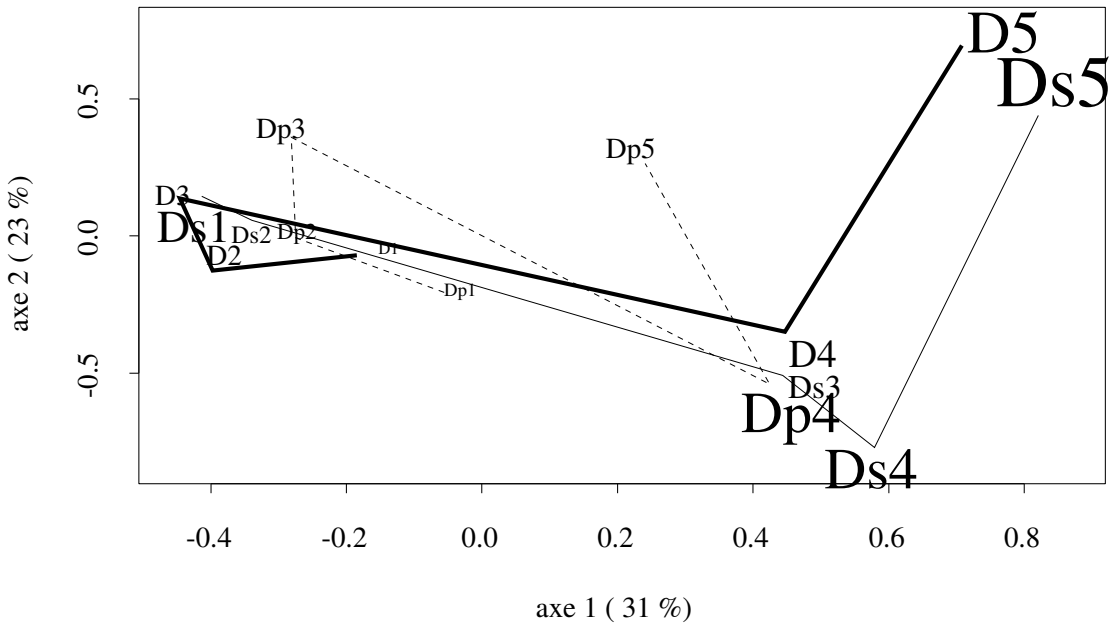


Le dendogramme et la nouvelle représentation graphique de la matrice des distances mettent en évidence :

- la forte concordance entre D2, Ds2 et Dp2.
- le fait que D1 exprime essentiellement un danger de prédation très faible,
- que D5 exprime essentiellement un danger de sommeil très fort,
- une assez bonne concordance entre D4, Ds4 et Dp4,,
- le fait que D3 n'a pas grand sens,
- et que Ds3 s'interprète plutôt comme un danger fort.

On peut personnaliser les représentations graphiques, de manière à faire ressortir certaines particularités. A l'aide de la fonction ploy.my[5], on a représenté les points avec une taille de caractères proportionnelle au maximum des contributions absolues sur les deux axes, ce qui permet de faire apparaitre les variables qui contribuent le plus à la détermination de l'un ou l'autre des axes. On peut envisager d'utiliser une taille des caractères proportionnelle aux contributions relatives, de manière à faire apparaître les variables les mieux représentées sur l'un ou l'autre des axes, ou encore sur le plan, si on somme les contributions relatives sur les deux axes.

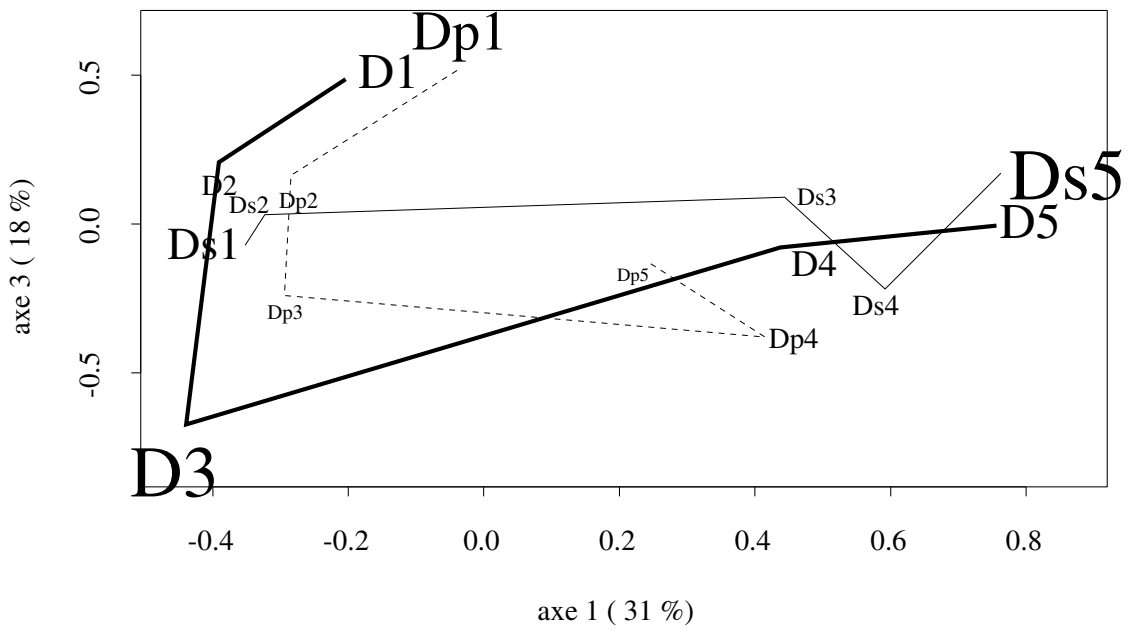
```
> plot.my(mamiferes.ca)
```



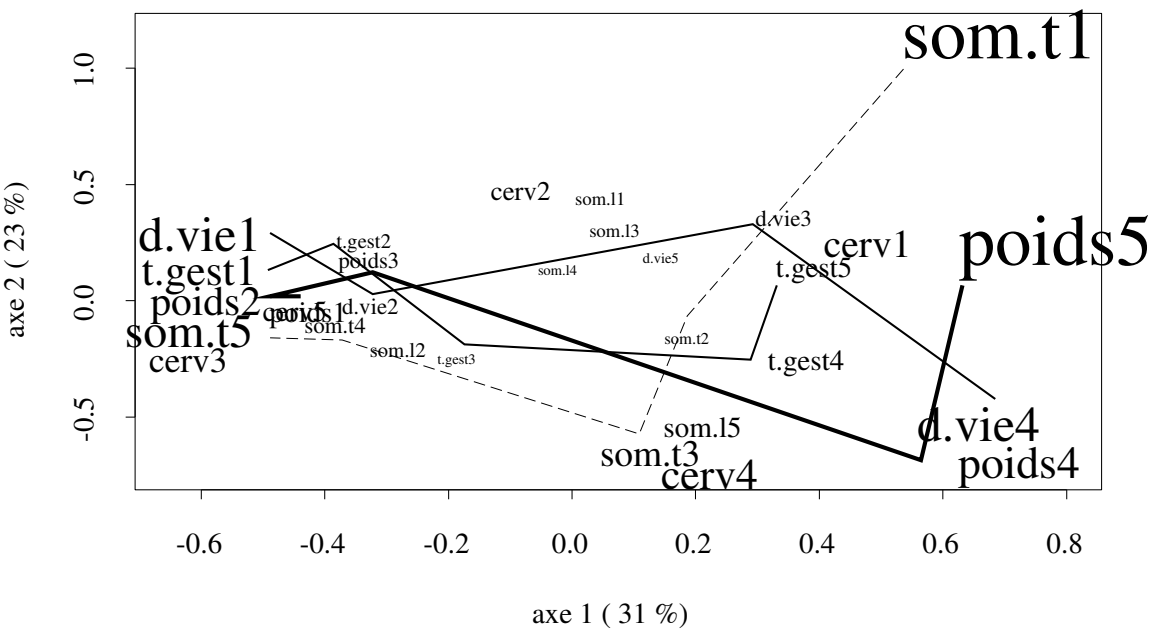
On a joint manuellement les modalités successives, pour chacun des types de danger, et pour certaines des modalités-ligne.

[5] présentée en annexe

```
> plot.my(mamiferes.ca,axes=c(1,3))
```




```
> plot.my(mamiferes.ca,rows="p",cols="n")
```



ANNEXES

I Valeurs propres et vecteurs propres : illustration graphique

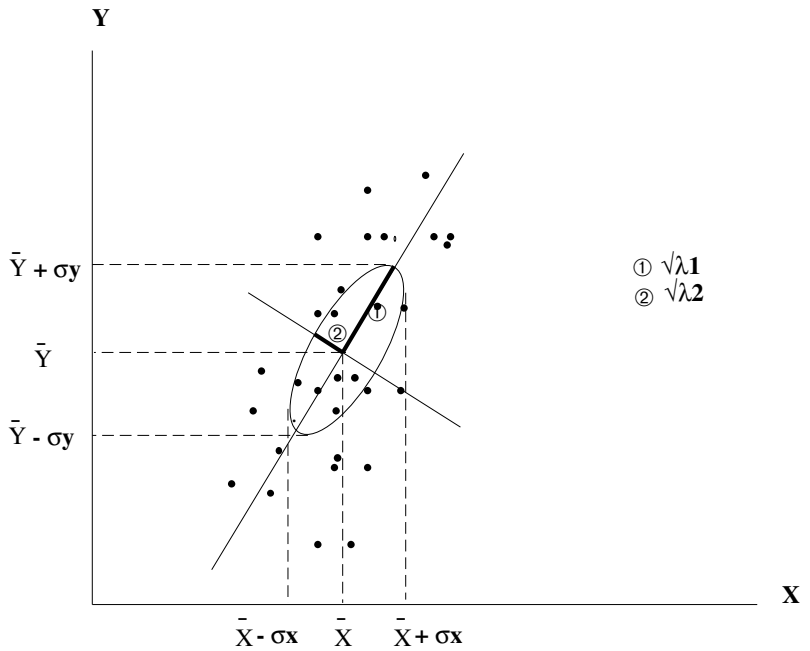
II Les sorties numériques de la fonction “acp” : coord, cta, ctr

III Les tableaux en codage disjonctif complet

IV Les fonctions spécifiques

ANNEXE 1 :
Valeurs propres et vecteurs propres : illustration graphique.

On considère le cas de 2 variables. Le nuage de points est donc représenté dans un plan.



La **matrice de variance-covariance** est représentée par l'**ellipse de dispersion**. Plus l'ellipse est applatie, plus la corrélation entre les deux variables est élevée. Un applatissement de l'ellipse dans la direction de la première bissectrice indique une corrélation positive, alors qu'un applatissement dans la direction de la deuxième bissectrice indique une corrélation négative. L'ellipse "serait un cercle", si les deux variables étaient indépendantes.

La projection de l'ellipse sur l'axe des abscisses est confondue avec l'intervalle $[\overline{x} - \sigma_x, \overline{x} + \sigma_x]$, sa projection sur l'axe des ordonnées est confondue avec l'intervalle $[\overline{y} - \sigma_y, \overline{y} + \sigma_y]$.

Le **grand axe** de l'ellipse est la direction de plus grande variabilité du nuage, c'est la **direction du vecteur propre** (de la matrice de variance-covariance) **associé à la plus grande valeur propre** λ_1. Le **petit axe** est la direction orthogonale à la précédente, qui est déterminée par la **direction du deuxième vecteur propre**, associé à la **deuxième valeur propre** λ_2 (ici la plus faible).

La **longueur du demi grand-axe** de l'ellipse est $\sqrt{\lambda_1}$, la **longueur du demi petit-axe** est $\sqrt{\lambda_2}$.

Le grand axe et le petit axe sont les deux axes factoriels d'une ACP réalisée sur ces données (deux variables).

Dans le cas de p variables, la matrice de variance-covariance a p valeurs propres, et p vecteurs propres associés. On classe les valeurs propres par ordre décroissant. Les vecteurs propres déterminent les directions des axes factoriels, de l'inertie la plus forte à l'inertie la plus faible.

ANNEXE 2 :
les sorties numériques de la fonction “acp” : coord, cta, ctr

D'un point de vu mathématique, une ACP est la diagonalisation d'une matrice de variance-covariance dans le cas d'une ACP non normée, d'une matrice de corrélation dans le cas d'une ACP normée.

Soient ϕ_i les p vecteurs propres de la matrice à diagonaliser, λ_i les p valeurs propres associées, et V les variances de chaque variable.

Dans la colonne coord, on a : $\sqrt{\lambda_i}.\sqrt{V}.\phi_i$
Dans la colonne cta : ϕ_i^2
Dans la colonne ctr : $\lambda_i.\phi_i^2$

On peut reconstituer la matrice de données, à l'aide des valeurs propres, des colonnees coord des individus et des variables.

Si on note
$\psi_i, \quad i = 1...p,$ les p vecteurs contenant les colonnes coord pour les variables ,
$\varphi_i, \quad i = 1...p,$ les p vecteurs contenant les colonnes coord pour les individus , et
$\lambda_i, \quad i = 1...p,$ les p valeurs propres, on reconstitue les données par :

$$X = \sum_{i=1}^{p} \sqrt{\lambda_i}\psi_i\varphi_i'$$

L'intérêt des valeurs présentées dans les colonnes coord, c'est de correspondre à la décomposition des données.

Mais : la colonne coord pour les variables ne contient pas les coordonnées des variables dans les cercles des corrélations (on lit d'ailleurs dans cette colonne des valeurs extèrieures à l'intervalle $[-1, 1]$). Les coordonnées des variables dans les cercles des corrélations s'obtiennent en calculant $\sqrt{\lambda_i}\phi_i$, il suffit donc de calculer la racine carrée de la colonne ctr, pour les obtenir.

ANNEXE 3 :
Les tableaux en codage disjonctif complet

Les méthodes d'analyse factorielle des correspondances restent valables pour tout tableau de nombres positifs, dès lors que la distance du chi2 entre les lignes du tableau a un sens (de même que la distance du chi2 entre les colonnes du tableau).

Si certaines des variables ne sont pas des variables ordinales mais des variables nominales (exemple couleur ...), l'analyse du tableau brut n'a pas de sens : même si on code les couleurs de 1 à 5, il n'est pas possible de donner un sens à l'ordre des modalités.

On peut, dans ce cas, coder chacune des variables et transformer le tableau en tableau disjonctif complet. Le tableau obtenu peut être étudié par une analyse des correspondances multiple.

Dans un tableau disjonctif complet, on trouve en ligne les individus, en colonne, les modalités des variables : à l'intersection d'une ligne et d'une colonne, on trouve 1 ou 0 selon que l'individu possède la modalité ou non. L'origine de la terminologie "tableau disjonctif complet" est la suivante : l'ensemble des valeurs d'un même individu pour les modalités d'une même variable, comporte la valeur 1 une fois (complet) et une fois seulement (disjonctif).

Voici par exemple le tableau mamiferes3 après codage en classes de chacune des variables.

```
> mamiferes3
```

	poids	cerv	s.leger	s.tot	vie.max	gest	pred	d.som	d.global
Tanrec	2	1	3	4	1	3	2	1	2
baboin	4	4	5	3	4	4	4	4	4
chat	3	3	2	5	4	3	1	2	1
cheval	5	1	1	1	5	5	5	5	5
chevre	4	2	4	1	3	4	5	5	5
chimpanze	5	3	4	3	5	5	1	1	1
chinchilla	2	4	4	4	2	4	5	4	4
cochon	5	1	2	2	4	4	4	4	4
cochon d'inde	3	2	5	2	2	3	5	3	4

. . .

Et le tableau disjonctif complet associé, obtenu à l'aide de la fonction "cod.disj".

```
> cod.disj(mamiferes3)
```

	poids1	poids2	poids3	poids4	poids5	cerv1	cerv2	...	danger global5
Tanrec	0	1	0	0	0	1	0		0
baboin	0	0	0	1	0	0	0		0
chat	0	0	1	0	0	0	0		0
cheval	0	0	0	0	1	1	0		1
chevre	0	0	0	1	0	0	1		1
chimpanze	0	0	0	0	1	0	0		0
chinchilla	0	1	0	0	0	0	0		0
cochon	0	0	0	0	1	1	0		0
coch d'inde	0	0	1	0	0	0	1		0

. . .

ANNEXE 4 :
Les fonctions spécifiques

Voici la définition des nouvelles fonctions utilisées dans le document.

dist2full

Cette fonction permet de transformer en matrice un tableau de distances produit par la fonction dist.[6]

```
dist2full <- function(dis)
      {
               n <- attr(dis, "Size")
               full <- matrix(0, n, n)
               full[lower.tri(full)] <- dis
               full + t(full)
      }
```

dist.gris

Cette fonction permet de représenter graphiquement des matrices de distances.[7]

```
dist.gris_function(matdist)
{
 p <- ncol(matdist)
 x <- seq(1, p + 1)
 y <- x
 z <- matdist
 if(length(dimnames(matdist)) == 0)
 dimnames(z) <- list(paste("VAR", seq(1, p)), paste("VAR", seq(1,p)))
 else dimnames(z) <- list(dimnames(matdist)[[2]], dimnames(matdist)[[2]])
 oldpar <- par(oma = c(6, 6, 0, 7), las = 1, pty = "s")
 image(x, y, z, xaxt = "n", yaxt = "n", las = 1, bty = "n")
 mtext(dimnames(z)[[1]], side = 1, line = 0.2, at = seq(1.5, p + 0.5, 1),
           srt=90,adj=1,cex=.7)
 mtext(dimnames(z)[[1]], side = 2, line = 2, at = seq(1.5, p + 0.5, 1),
           adj=1,cex=.7)
 title("Matrice des Distances ")
 image.legend(z, x = (p + 1)+p/8, y = p + 1, size = c(0.2, 4), nint = 15,
           hor = F)
 par(oldpar)
}
```

[6] Cette fonction est présentée dans l'aide en ligne de la fonction dist.

[7] Pour superposer une autre représentation à la matrice grisée, avec la même échelle, (ex. AFC multiple page 23), mettre la ligne `par(oldpar)` en commentaire.

plot.my

Cette fonction permet de représenter, dans les plans factoriels d'une analyse des correspondance, les projections des lignes et/ou des colonnes. Les tailles des noms des lignes et des colonnes sont proportionnelles à leurs contributions absolues ou relatives.

```
plot.my_function(ca,axes=c(1,2),ct="a",rows="n",cols="p")
{ax1_axes[1]; ax2_axes[2]
 scale.my_function(x) {x_x-min(x);2*x/max(x)}
 inertie_round(ca$valp/sum(ca$valp)*100)
 if (cols!="n")
 {x_ca$var.cords[,ax1] ; y_ca$var.cords[,ax2]
  if (ct=="r") {taille1_scale.my(ca$var.ctr[,ax1])
                taille2_scale.my(ca$var.ctr[,ax2])}
  if (ct=="a") {taille1_scale.my(ca$var.cta[,ax1])
                taille2_scale.my(ca$var.cta[,ax2])}
  plot(x,y,type="n",xlab=paste("axe",ax1,"(",inertie[ax1],"%)"),
                    ylab=paste("axe",ax2,"(",inertie[ax2],"%)"))
  text(x,y,names(x),cex=apply(cbind(taille1,taille2),1,max)+0.5) }
 if (rows!="n")
 {x_ca$ind.cords[,ax1] ; y_ca$ind.cords[,ax2]
  if (ct=="r") {taille1_scale.my(ca$ind.ctr[,ax1])
                taille2_scale.my(ca$ind.ctr[,ax2])}
  if (ct=="a") {taille1_scale.my(ca$ind.cta[,ax1])
                taille2_scale.my(ca$ind.cta[,ax2])}
  plot(x,y,type="n",xlab=paste("axe",ax1,"(",inertie[ax1],"%)"),
                    ylab=paste("axe",ax2,"(",inertie[ax2],"%)"))
  text(x,y,names(x),cex=apply(cbind(taille1,taille2),1,max)+0.5) }
invisible()
}
```

cod.disj

Cette fonction éclate un tableau de données codées, en tableau disjonctif complet.

```
cod.disj_function(don)
# l'argument doit etre une matrice labellee
{disjonc_NULL
 for (i in 1:ncol(don))
 {tab_table(dimnames(don)[[1]],don[,i])
 dimnames(tab)[[2]]_paste(dimnames(don)[[2]][i],1:ncol(tab),sep=".")
 disjonc_cbind(disjonc,tab)
 }
 disjonc
}
```

BIBLIOGRAPHIE

Ouvrages et documents généraux :

- Bouroche,J.M., Saporta,G. L'analyse des donnees, Que sais-je ?, PUF, 1989.
- Diday E., Lemaire J., Pouget J., Testu F., Eléments d'analyse de donnés, Dunod,
- Escofier B., Pagès J., Analyses factorielles simples et multiples, Dunod, 1988.
- Escoufier Y., De l'analyse en composantes principales à l'analyse des correspondances multiples, INRA, 1991.
- Jambu M., Exploration informatique et statistique des données, Dunod informatique, 1989.
- Mardia,K.V., Kent,J.T. and Bibby, J.M., Multivariate Analysis, Academic Press, London, 1979.
- Tomassone R., Dervin C., Masson J.P., Biométrie : modélisation de phénomènes biologiques, MASSON, 1993.

Ouvrages et documents spécifiques à l'ACP :

- Philippeau G., Comment interpréter les résultats d'une analyse en composantes principales?, SESI-ITCF, 1986.

Ouvrages et documents spécifiques à l'AFC :

- Benzécri J.-P. & F., Pratique de l'analyse des données, Analyse des correspondances, Dunod, 1980.
- Dervin C., Comment interpréter les résultats d'une analyse factorielle des correspondances?, SESI-ITCF, 1990.
- Greenacre M.J., Correspondence Analysis in Practice, Academic Press, London, 1993.
- Greenacre M.J., Theory and Application of Correspondance Analysis, Academic Press, London, 1984.

Ouvrages et documents spécifiques à la classification:

- Everitt, B. Cluster Analysis (second edition), Halsted, New York, 1980.
- Hartigan, J.A. Clustering Algorithms, Wiley New York, 1975.
- Hartigan, J.A. and Wong, M.A. A k-means clustering algorithm, Applied Statistics 28, 1979, 100-108.
- Celeux,G, Diday E., Govaert G., Lechevallier Y., Ralambondrainy H. Classification automatique des donnees, Dunod informatique, Paris, 1989.
- Gordon,A.D. Classification, Chapman et Hall, London, New- York, 1981.
- Jambu,M. Classification automatique pour l'analyse des données, Dunod, Paris, 1978.
- Roux,M., Algorithmes de classification, Masson, Paris, 1985.

Ouvrages et documents liés au logiciels Splus :

- Baumgartner M., Une introduction à S-Plus, Ecole Polytechnique Fédérale de Lausanne, 1994.
- Chambers J.M., Hastie T.J., Statistical models in S, 1991
- Everitt B.S., a handbook of Statistical Analyses using S-PLUS, Chapman & Hall, 1994

- Spector P., An introduction to S and S-Plus, Duxbury Press, 1994
- Venables W.N., Ripley B.D., Modern Applied Statistics with S-plus, Springer-Verlag, 1994