Analyse des correspondances (TP Splus)

INRA ©1998 - Formation statistique - Module 7/8 (AFC TP Splus)-1

```
Splus  #appel de Splus

S-PLUS : Copyright (c) 1988, 1996 MathSoft, Inc.
S : Copyright AT&T.
Version 3.4 Release 1 for Sun SPARC, SunOS 5.3 : 1996
Working data will be in .Data

>library(multidim, first=T) # Appel de la librairie multidim.

>search()# Avec la fonction search, on vérifie que la librairie
         # multidim est en première position (fisrt=T) ; ainsi,
         # les fonctions appelées qui seront effectuées seront
         # bien celles de multidim et non celles de Splus qui
         # pourraient exister sous le même nom.

 [1] ".Data"
 [2] "/usr/local/prive/splus3.4/library/multidim/.Data"
 [3] "/usr/local/prive/splus3.4/library/inra/.Data"
 [4] "/usr/local/prive/splus3.4/splus/.Functions"
 [5] "/usr/local/prive/splus3.4/stat/.Functions"
 [6] "/usr/local/prive/splus3.4/s/.Functions"
 [7] "/usr/local/prive/splus3.4/s/.Datasets"
 [8] "/usr/local/prive/splus3.4/stat/.Datasets"
 [9] "/usr/local/prive/splus3.4/splus/.Datasets"
[10] "/usr/local/prive/splus3.4/library/trellis/.Data"
```

```
> ospef # Le fichier de données s'appelle ospef
        # dans la librairie multidim.
```

	dro	sci	let	med	pha	iut
age	4814	8941	8719	1583	959	558
oua	463	1118	1224	153	33	83
ind	3610	2608	3365	1634	728	110
art	2870	5091	5966	1583	562	261
gco	2777	2608	3212	1736	794	136
pco	4444	6209	7954	2655	1125	313
lib	9813	7823	9637	8985	3143	120
ing	4444	8195	7648	3880	1307	151
adm	14071	13535	18050	7147	2134	235
mes	833	993	1071	817	182	21
mso	555	621	765	357	66	16
tec	2129	3974	4436	1021	298	235
cam	7776	10058	13308	3931	1125	407
ins	1666	5464	6425	1481	579	141
emb	5184	8071	10861	2399	529	0
con	1481	2856	3059	613	132	177
ous	2314	4098	4436	766	182	516
min	278	869	918	153	17	37
mar	185	373	306	51	0	21
man	741	1863	1989	255	33	83
ser	741	993	1377	204	33	63
arm	2592	3104	3671	817	198	136

Les valeurs du tableau sont des effectifs d'étudiants dans la population française. En ligne, la catégorie socio-professionnelle des parents, en colonne la discipline suivie par les étudiants.

Catégories socio-professionnelles (CSP)

- *age* : Agriculteur exploitant
- *oua* : Ouvrier agricole
- *ind* : Industriel
- *art* : Artisan
- *gco* : Moyen et gros commerçant
- *lib* : Profession libérale
- *ing* : Ingénieur
- *adm* : Cadre de l'Administration
- *mes* : Profession médicale et salariée
- *mso* : Profession médicale et sociale
- *tec* : Technicien
- *cam* : Cadre administratif moyen
- *ins* : Instituteur
- *emb* : Employé de bureau
- *con* : Contremaître
- *ous* : Ouvrier spécialisé
- *min* : Mineur
- *mar* : Pêcheur
- *man* : Manœuvre
- *ser* : Personnel de service
- *arm* : Armée, Police

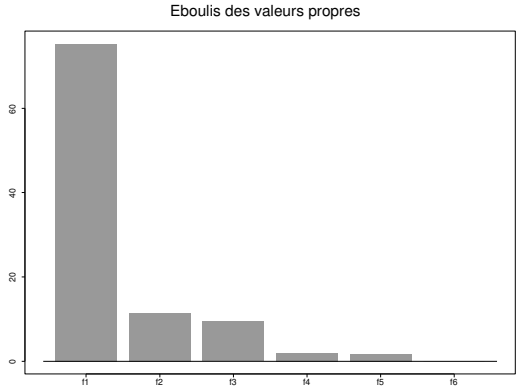
Disciplines

- *dro* : Droit
- *sci* : Sciences
- *let* : Lettres
- *med* : medecine
- *pha* : pharmacie

On cherche à répondre aux questions suivantes :

- la discipline suivie par l'étudiant renseigne-t-elle sur son origine sociale ?
- les étudiants d'une CSP donnée font-ils préférentiellement certains types d'études ?

```
> res <- afc(ospef) # Réalisation de l'afc par la fonction afc() et envoi
                    # des résultats dans le fichier ``res'' ; la fonction
                    # afc() peut s'appliquer à  n'importe quel data.frame
                    # créé sous Splus par :
                    # > ospef <- read.table("ospef",header=T).
Inertie totale:  0.06

> sum(res$values)   # Valeur exacte de l'inertie totale,
                    # soit la somme des valeurs propres.
[1] 0.06020051
> motif()           # Ouverture d'une fenêtre graphique.
> eboulis(res)      # Appel de la fonction eboulis() et visualisation
                    # de la diminution successive des valeurs propres.
```




```
> prexpl(res)

Pourcentage de inertie expliquee
             f1 f2 f3 f4  f5 tot
inertie exp 75 12  9  2   2 100
inertie cum 75 87 96 98 100 100

> round(res$values,4) # Édition des valeurs propres ($values) dans le
                      # fichier ``res'', arrondies (round) au 4ème chiffre.
     f1     f2     f3     f4     f5 f6
 0.0453 0.0069 0.0057 0.0012 0.0011  0

> round(sqrt(res$values),4) # Édition des valeurs singulières, racine
                            # carrée (sqrt) des valeurs propres.
     f1     f2     f3     f4     f5 f6
 0.2129 0.0832 0.0755 0.0345 0.0325  0
```


– La *règle du coude* indique que l'axe 1 est déjà un très bon résumé de la table de contingence avec 75,3 % de l'inertie totale. Ensuite, les 2 axes suivant sont à considérer si l'on veut aller plus loin dans l'analyse (soit 3 axes en tout).

– Les valeurs propres sont les carrés des valeurs singulières issues de la DVS.

```
> contri(res) # contri() est une fonction très complète qui édite
              # successivement :
              # - les cosinus carrés (qualité) des profils lignes
              #   avec les vecteurs principaux (f1, f2,..., axes >5) ;
              # - les valeurs cumulées de ces cosinus carrés dans
              #   les sous-espaces principaux correspondants
              #   (Axes1, Axes1a2,..., Axes1a5).
              # - la contribution des profils lignes à l'inertie
              #   de chacun des vecteurs principaux (f1, f2,..., f5) ;
              #   ORI est directement proportionnel à la distance du
              #   Khi2 du profil ligne au centre du nuage des lignes ;
              #   + 3 tableaux complétement analogues pour les
              #   profils colonnes.

 Cosinus carre lignes axes principaux

     f1  f2  f3  f4  f5 Axes >5  TOT
age 731  10 161  96   2       0 1000
oua 957   8  25   9   0       0 1000
ind 615 264  71  49   2       0 1000
art 870  97   7   3  23       0 1000
gco 821   0 119  25  35       0 1000
pco   3 171  85   7 734       0 1000
lib 976  19   3   2   0       0 1000
ing 134 704  17  22 123       0 1000
adm 437 484  73   0   6       0 1000
mes 740  82   4 100  73       0 1000
mso 355 113  82 258 191       0 1000
tec 945  18  27  10   1       0 1000
cam 441 314  92 102  52       0 1000
ins 498 361 108   0  32       0 1000
emb 416  36 547   0   0       0 1000
con 941   1  27  11  19       0 1000
ous 606   3 341  50   0       0 1000
min 883  76  19  10  12       0 1000
mar 745   5  43   1 207       0 1000
man 969   9   8   1  13       0 1000
ser 741 223   0  27   9       0 1000
arm 415 526   3   4  53       0 1000
TOT 753 115  95  20  18       0 1000
```


– Chaque profil ligne centré se décompose suivant la base orthonormée $\{\mathbf{v_s}\}$.

Dans la ligne i du premier tableau, on a les cosinus carré ($\times 1000$) du profil i avec les vecteurs principaux $\mathbf{v_s}$.

Dans le second tableau, les cosinus carré sont cumulés : on obtient la qualité de la représentation du profil i sur le s*ème* sous-espace principal dans la colonne s.

– Les cosinus carrés étant aussi les contributions relatives des facteurs aux individus, l'axe 1 contribue majoritairement à l'inertie des individus (par définition) mais certains sont mieux représentés pour certains axes :

ing pour l'axe 2
emb pour l'axe 3
pco pour l'axe 5

Cosinus carré lignes sous-espaces principaux

	Axe 1	Axes1a2	Axes1a3	Axes1a4	Axes1a5	TOT
age	731	741	902	998	1000	1000
oua	957	965	991	1000	1000	1000
ind	615	879	950	998	1000	1000
art	870	968	974	977	1000	1000
gco	821	821	940	965	1000	1000
pco	3	174	259	266	1000	1000
lib	976	995	998	1000	1000	1000
ing	134	838	855	877	1000	1000
adm	437	921	994	994	1000	1000
mes	740	822	827	927	1000	1000
mso	355	468	551	809	1000	1000
tec	945	962	989	999	1000	1000
cam	441	755	847	948	1000	1000
ins	498	859	968	968	1000	1000
emb	416	452	1000	1000	1000	1000
con	941	942	969	981	1000	1000
ous	606	609	950	1000	1000	1000
min	883	959	978	988	1000	1000
mar	745	749	793	793	1000	1000
man	969	978	987	987	1000	1000
ser	741	964	964	991	1000	1000
arm	415	941	943	947	1000	1000
TOT	753	868	963	982	1000	1000

Contribution des lignes aux facteurs

	f1	f2	f3	f4	f5	f>5	ORI
age	67	6	118	337	7	69	36
oua	25	1	5	9	0	20	86
ind	35	99	32	106	4	43	48
art	17	12	1	2	19	14	12
gco	33	0	37	38	60	30	35
pco	0	8	5	2	217	5	3
lib	479	61	12	40	1	370	124
ing	6	213	6	38	246	35	18
adm	30	217	40	1	18	52	12
mes	15	11	1	78	65	16	53
mso	1	2	2	26	22	2	11
tec	28	3	6	11	2	22	24
cam	6	28	10	53	31	10	4
ins	43	204	74	2	119	65	55
emb	32	18	329	0	1	57	28
con	31	0	7	14	28	25	40
ous	68	2	304	214	0	85	91
min	17	10	3	7	10	15	86
mar	6	0	3	0	71	6	86
man	39	2	3	1	22	31	82
ser	12	23	0	16	6	12	46
arm	9	78	1	3	51	17	21
TOT	1000	1000	1000	1000	1000	1000	1000

La base 1 000 est fixée facteur par facteur :

lib est responsable, à lui seul, de 47,9 % de l'inertie de l'axe 1 et de 37 % de l'inertie totale.

pco représente 0.0 % de l'inertie de l'axe 1 et 0.5 % de l'inertie totale.

```
 Cosinus carré colonnes axes principaux

     f1  f2  f3 f4 f5 f>5  TOT
dro 478 489  23  4  5   0 1000
sci 830 118   2 14 36   0 1000
let 805   8 125 11 52   0 1000
med 932  46   2 18  2   0 1000
pha 808  85  33 46 29   0 1000
iut 376   2 594 23  4   0 1000
TOT 753 115  95 20 18   0 1000

 Cosinus carré colonnes sous-espaces principaux

     f1 f1a2 f1a3 f1a4 f1a5  TOT
dro 478  967  991  995 1000 1000
sci 830  948  950  964 1000 1000
let 805  812  937  948 1000 1000
med 932  978  979  998 1000 1000
pha 808  892  925  971 1000 1000
iut 376  378  972  996 1000 1000
TOT 753  868  963  982 1000 1000

 Contribution des colonnes aux facteurs

      f1   f2   f3   f4   f5  f>5  TOT  ORI
dro   89  596   35   31   40  210  140   35
sci  161  150    3  102  302  283  146   27
let  119    7  147   60  331  337  111   17
med  403  129    6  301   40  120  325  141
pha  168  115   55  365  256   40  157  203
iut   60    2  755  141   30   11  120  577
     1000 1000 1000 1000 1000 1000 1000 1000
```

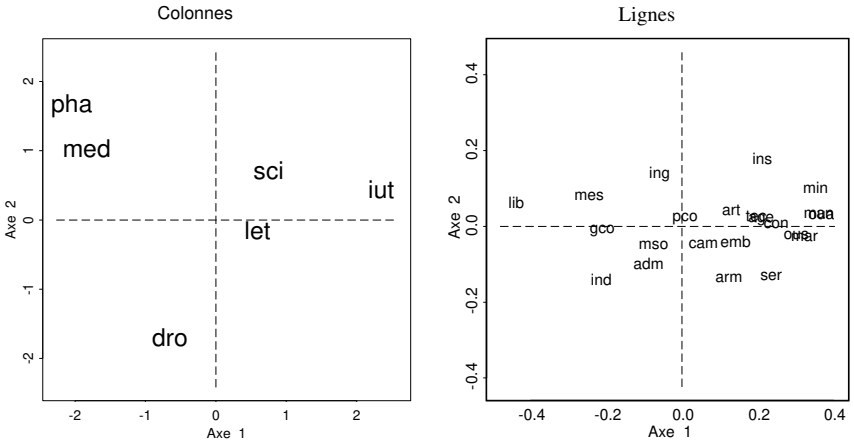

Pour les colonnes, on a un résultat semblable à ceux du transparent 6 :

On remarque que *iut* n'est bien représenté qu'à partir du sous-espace de dimension 3.

- Idem pour les colonnes :
  med contribue à 40.3 % de l'inertie de l'axe 1
  dro contribue à 59.6 % de l'inertie de l'axe 2
  iut contribue à 75.5 % de l'inertie de l'axe 3

  La dernière colonne exprime le carré de la distance d'un profil au profil moyen. C'est une mesure de l'originalité de ce profil.
- Dans tous les cas (lignes et colonnes), faire attention au fait que l'inertie relative peut être élevée mais faible en valeur absolue, en particulier sur les axes de rang élevé (axes 3, 4, 5, ...).

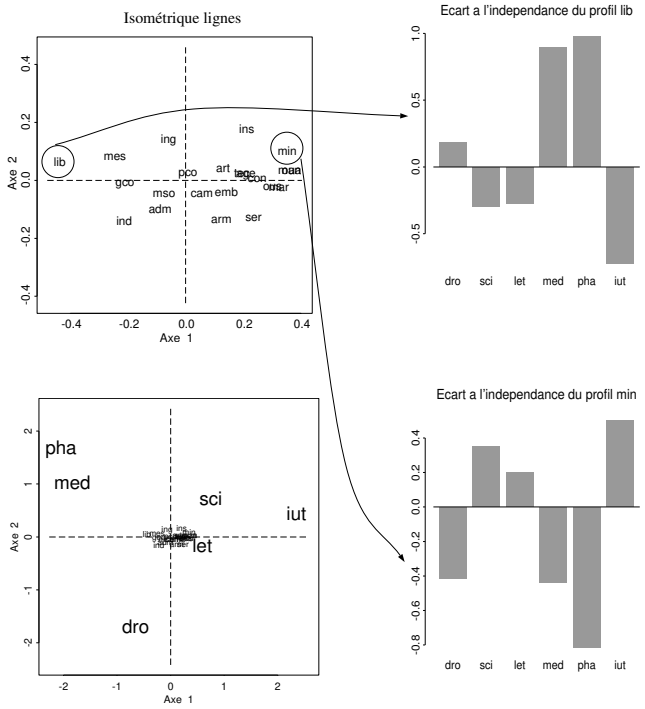
```
> biplot(res,1,2) # Représentation isométrique ligne des
  Selection: 1    # profils lignes et colonnes dans le
                  # sous-espace principal (1,2) en deux
                  # graphiques.
```



En AFC, la représentation barycentrique est aussi un biplot :

- Représentation *isométrique* lignes donc en $\lambda_s u_i^s$ pour les lignes et v_s^j pour les colonnes.
- Les points lignes sont au barycentre des points de coordonnées v_s^j (pour les poids : $f_{ij}/f_{i.},\ j = 1, \ldots, J$) formant le $ième$ profil colonne.

```
> plot(res$cmpr[,c(1,2)],type="n",xlab="Axe 1",ylab="Axe 2", main=
  "Isometrique Lignes",cex=0.9) # Représentation des profils lignes seuls.
  text(res$cmpr[,c(1,2)],labels=dimnames(ospef)[[1]])
  abline(h=0,v=0)

> tot <- rbind(res$cmpr,res$vectors)
  plot(tot[,c(1,2)],type="n",xlab="Axe 1",ylab="Axe 2",
       main="Isometrique Lignes",cex=0.9) # Représentation des profils lignes
                                          # et colonnes en un seul graphique.
  text(tot[,c(1,2)],labels=dimnames(tot)[[1]])
  abline(h=0,v=0)
```



– Phénomène d'échelle bien visible dans l'espace (1,2) où tous les points lignes sont rassemblés au centre tandis que les points colonnes sont plus dispersés.

– Sur l'axe 1, l'opposition de *lib* et de *min* correspond bien à des profils d'écart à l'indépendance complètement opposés pour toutes les catégories j (grande valeur des écarts à l'indépendance : $\pm 1 = \pm 100\%$ de la valeur attendue sous l'hypothèse d'indépendance : $f_{i.}f_{.j}$).

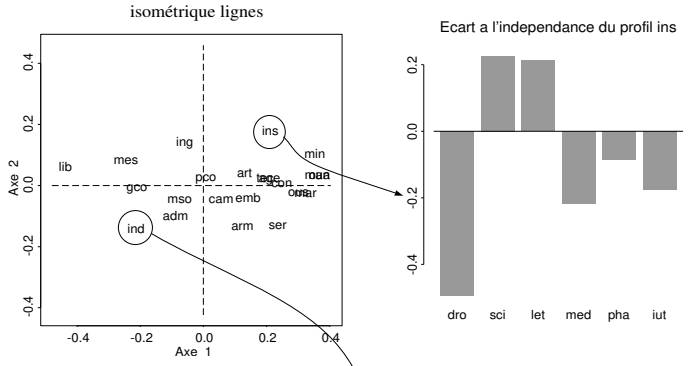
– Même phénomène sur l'axe 2 pour *ins* et *ind* mais avec de plus faibles variations des écarts à l'indépendance : ± 0.5.

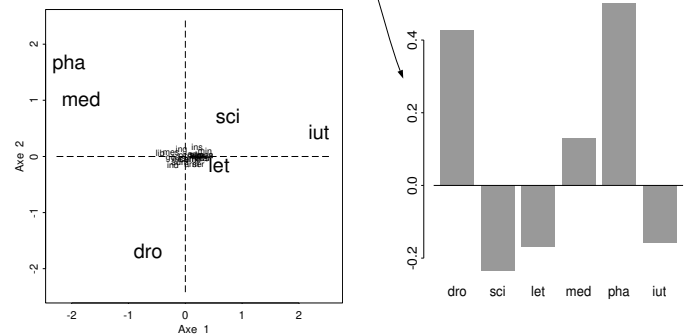
– *pco* est proche de l'origine car il a une faible originalité (ORI).

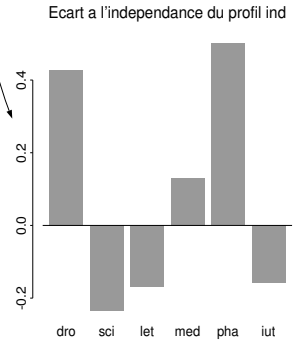
– Faire remarquer que les coordonnées des points lignes et colonnes permettent de calculer des approximations (de faible rang) des écarts à l'indépendance : $\sum_s \lambda_s u_i^s v_s^j = \dfrac{f_{ij}}{f_{i.}f_{.j}} - 1$

Exemple : (lib, med)
approximation $= -0.4 \times (-2) + 0.05 \times 1 = 0.85$
vraie valeur $= 0.89$ (*cf.* histogramme)

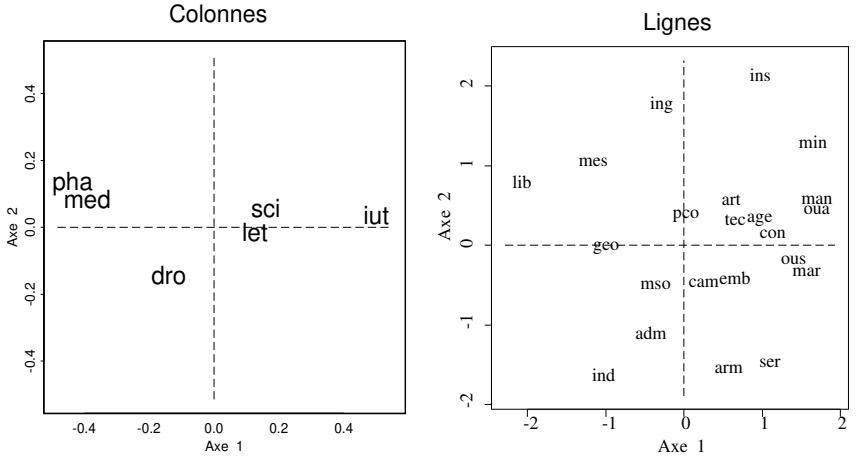
```
> fij <- ospef/sum(ospef)
  fi. <- as.vector(fij %*% rep(1, ncol(fij)))
  f.j <- as.vector(rep(1, nrow(fij)) %*% fij)

  ecart.indep <- diag(1/fi.) %*% fij %*% diag(1/f.j) - 1
  # Écarts à l'indépendance du profil ligne ``lib'' dans l'espace des lignes.
  dimnames(ecart.indep) <- dimnames(ospef)

  barplot(ecart.indep["lib",],names=dimnames(ecart.indep)[[2]],
          main="Ecart a l'independance du profil lib",cex=0.9)
  # Les profils ``min'', ``ins'' et ``in' s'obtiennent en faisant la
  # substitution appropriée de ``lib'' dans la fonction barplot().
```







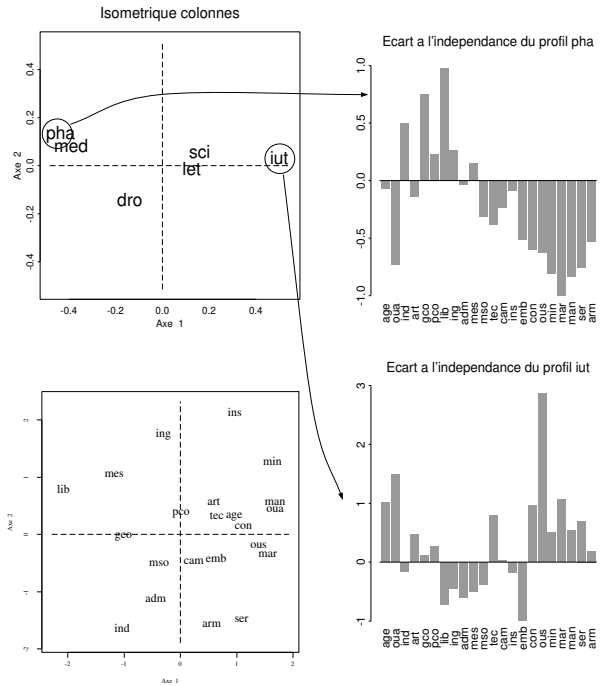
```
> biplot(res,1,2) # Représentation isométrique colonnes
  Selection: 2    # des profils lignes et colonnes dans le
                  # sous-espace principal (1,2) en deux
                  # graphiques.
```



– Représentation *isométrique* colonnes donc en $\lambda_s v_s^j$ pour les colonnes et u_i^s pour les lignes. C'est aussi un biplot. Remarquer les différences d'échelles des graphes.

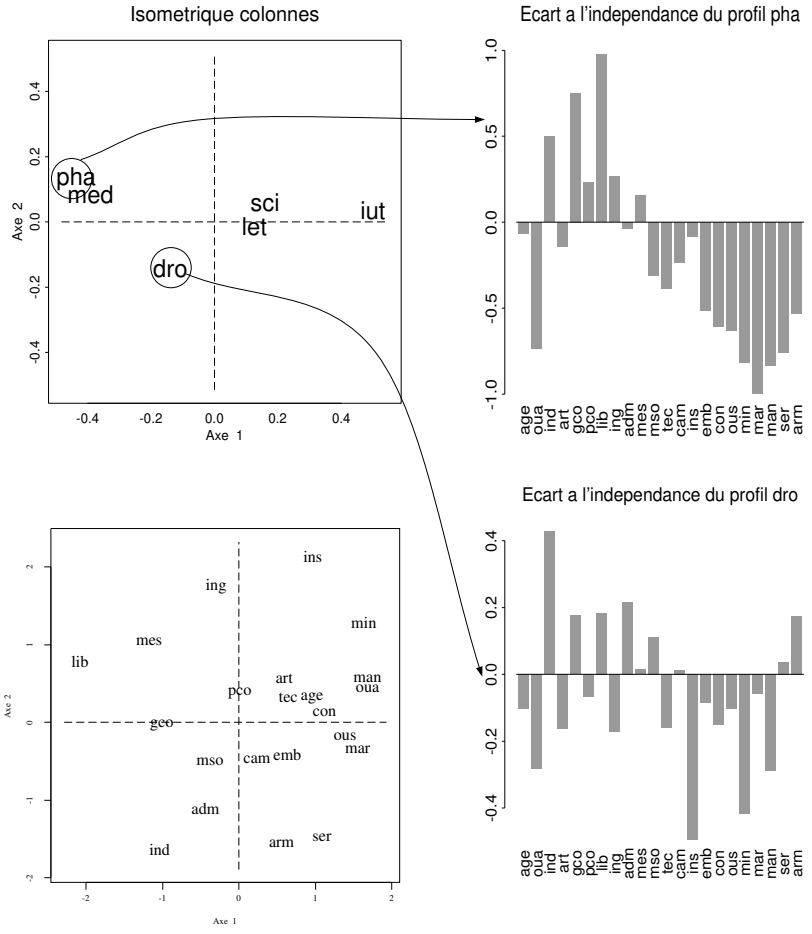
– Les points colonnes sont au barycentre des points de coordonnées u_i^s (pour les poids : $f_{ij}/f_{.j},\ i = 1, \ldots, I$) formant le *jème* profil ligne.

```
> plot(res$vectors %*% diag(sqrt(res$values))[,c(1,2)],
       type="n",xlab="Axe 1",ylab="Axe 2",main="Isometrique Colonnes",
       cex=0.9) # Représentation des profils colonnes seuls.
  text(res$vectors %*% diag(sqrt(res$values))[,c(1,2)],
       labels=dimnames(ospef)[[2]])
  abline(h=0,v=0)

> tot_rbind(res$vectors %*% diag(sqrt(res$values)),res$cmpr %*%
            diag(1/sqrt(res$values)))
  plot(tot[,c(1,2)],type="n",xlab="Axe 1",ylab="Axe 2",
       main="Isometrique Colonnes",cex=0.9)
  # Représentation des profils lignes et colonnes en un seul graphique.
  text(tot[,c(1,2)],labels=dimnames(tot)[[1]])
  abline(h=0,v=0)
```



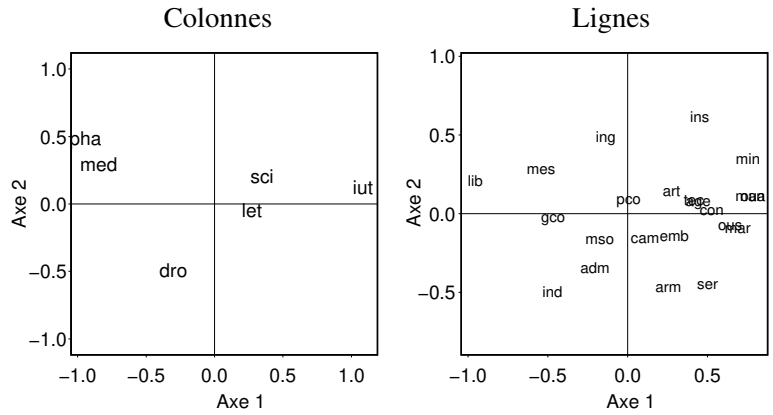
Sur l'axe 1, l'opposition de *pha* et de *iut* est cohérente avec l'antagonisme presque systématique des profils d'écart à l'indépendance pour toutes les catégories i.

```
> barplot(ecart.indep[,"pha"],names=dimnames(ecart.indep)[[1]],
          main="Ecart a l'independance du profil pha",cex=0.9)
  # Ecarts à l'indépendance du profil colonne ''pha'' dans l'espace
  # des colonnes ; les profils ''iut'' et ''dro'' s'obtiennent en
  # faisant la substitution appropriée de ''pha'' dans la fonction
  # barplot().
```

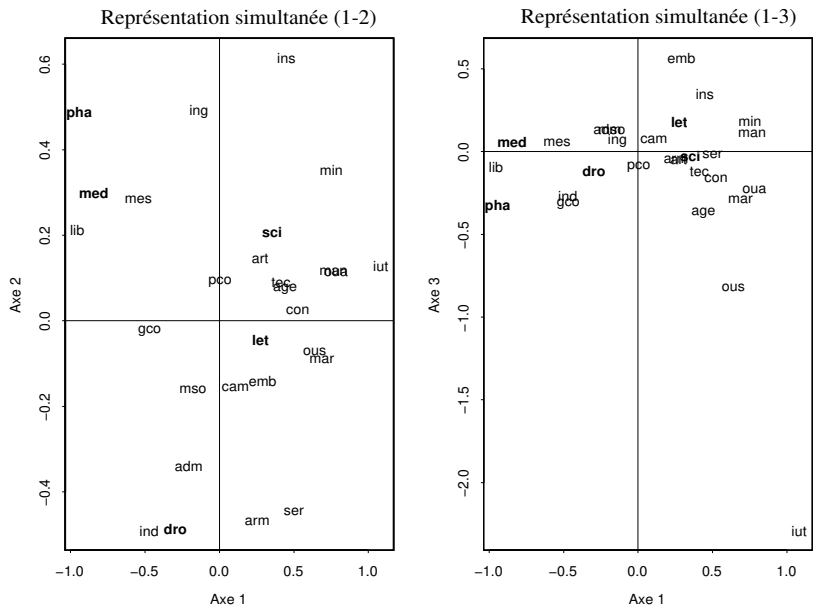


Même phénomène, en moins accentué, sur l'axe 2 pour *pha* vis-à-vis de *dro*.

```
> biplot(res,1,2) # Représentation isométrique des profils
                  # lignes et colonnes dans le sous-espace
   Selection: 3   # principal (1,2) en deux graphiques.
```



```
> par(mfrow=c(1,2))
  tot_rbind(res$cmpr %*% diag(res$values**-0.25),
            res$vectors %*% diag(res$values**0.25))
  plot(tot[,c(1,2)],type="n",xlab="Axe 1",ylab="Axe 2",
       main="Simultanee Lignes et Colonnes (1,2)",cex=0.9)
  # Représentation en un seul graphique des sous-espaces principaux
  # (1,2) et (1,3).
  text(tot[,c(1,2)],labels=dimnames(tot)[[1]],cex=0.9)
  abline(h=0,v=0)
  plot(tot[,c(1,3)],type="n",xlab="Axe 1",ylab="Axe 3",
       main="Simultanee Lignes et Colonnes (1,3)",cex=0.9)
  text(tot[,c(1,3)],labels=dimnames(tot)[[1]],cex=0.9)
  abline(h=0,v=0)
```



- Représentation simultanée ici en $\lambda_s u_i^s$ pour les lignes et $\lambda_s v_s^j$ pour les colonnes (axes 1-2 et 1-3). Les points ne sont plus des barycentres. C'est la représentation la plus utilisée dans les logiciels.
- L'avantage de cette représentation simplifiée est de permettre une interprétation plus facile des effets *lignes* et *colonnes* en terme de direction centrifuge.

  Par exemple,

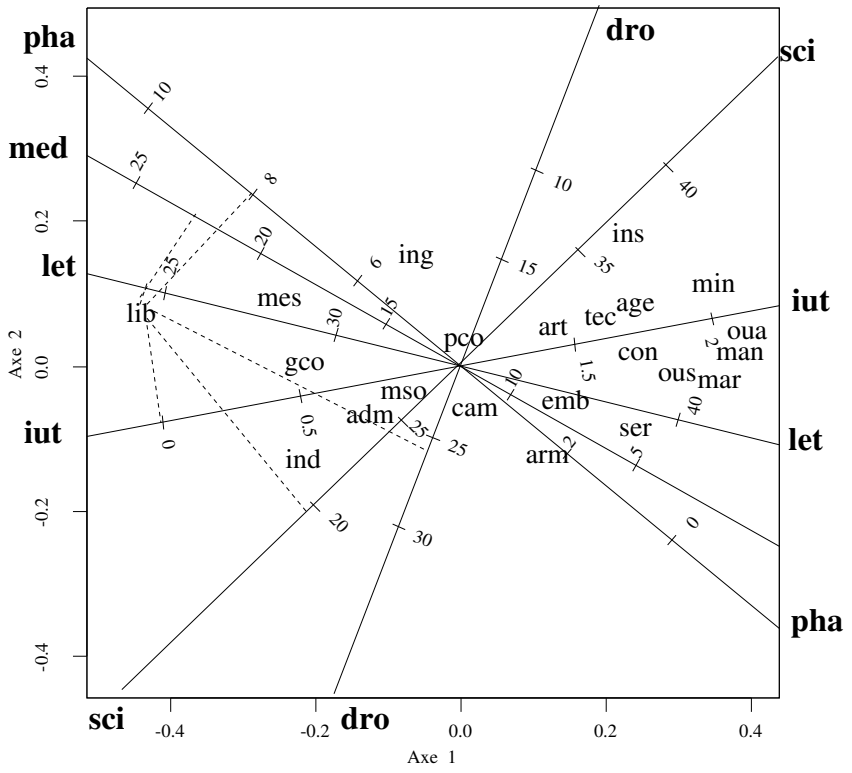
  lib est associé à *med* et *pha*

  ind, *adm* avec *dro*

  ins, *min* avec *sci*

  age, *ous* avec *iut*

```
> gabriel(res,1,2) # Représentation isométrique ligne où les
  Selection : 1    # axes sont les vecteurs propres gradués
                   # en profil ligne fij/fi. ; par définition,
                   # le centre de l'espace des lignes est
                   # occupé par le profil ligne moyen : f.j
                   # Comme pour les représentations dans les
                   # sous-espaces principaux, des biplots
                   # gradués isométrique colonne et simultané
                   # peuvent être obtenus par : Sélection = 2 ou 3.
```



– Le biplot gradué correspond ici à un biplot *isométrique* ligne où les axes sont les vecteurs propres (coordonnées des colonnes en v_s^j et gradués en $100 f_{ij}/f_{i.}$.

Pour une ligne i, les projections sur les axes donnent une approximation (de faible rang) de son profil : $f_{ij}/f_{i.} = \Pr(j|i)$ (probabilité de suivre les études j sachant que l'étudiant appartient à la catégorie socio-professionnelle i).

Exemple : $i = lib$

Projections sur :		Vraies valeurs $n_{ij}/n_{i.} = f_{ij}/f_{i.}$
med	22 %	0.2273
pha	8 %	0.0795
iut	0.1 %	0.0030
dro	26 %	0.2483
sci	20 %	0.1979
let	24 %	0.2438

Remarque :

– Les biplot précédents permettaient de calculer des approximations de $\dfrac{f_{ij}}{f_{i.}f_{.j}} - 1$. Par le jeu des graduations, on lit ici directement les approximations des pourcentages lignes.

– Un biplot gradué peut être réalisé à partir d'une représentation *isométrique* colonne où les axes sont gradués en 100 $f_{ij}/f_{.j}$. Les projections sont des valeurs approchées de 100 $\Pr(i|j)$.