Analyse Factorielle des Correspondances
(TP SAS)

Bilan des méthodes

Donnée	Quantitative	Qualitative
Tableau	Individus $\times$ Variables	Variable1 $\times$ Variable2
		Individus $\times$ Variables
Exemples	forsy	CSP, Pommes
Analyse	ACP	AFC, ACM
	Classification	
Programmes		CSP.sas, Pommes.sas

Préciser le type des variables sur lesquelles on travaille.

variables quantitatives variables continues comme le poids, la taille, ...

variables qualitatives variables nominales: couleur, sexe, ..., et variables ordinales: niveau d'étude, ...

Selon les variables les traitements ne seront pas les mêmes.

Insister sur la nécessité de faire des pré-traitements avant de se lancer dans l'analyse. Utiliser

- des statistiques élémentaires adaptées au type des variables,
- rechercher les particularités (individus aberrants, influents, ...)
- des méthodes graphiques uni, bi ou multivariables,

Les tableaux avec des zéros structurels ne sont pas bien adaptés à l'utilisation de l'AFC.

Ce qui compte, c'est l'idée qu'on se fait de la distance entre individus (ex.: équivalence distributionnelle, est-ce que la distance entre 1 et 2 est équivalente à la distance entre 10 et 20).

On peut faire des ACP sur des variables qualitatives ordonnées (il est conseillé de créer une variable ayant des modalités d'effectifs à peu près égaux).

Fichiers de données : CSP.dat

	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

Le fichier donne la répartition en effectif des catégories socio-professionnelles (CSP) des parents en fonction des études supérieures suivies par leurs enfants ainsi que la signification des identificateurs lignes et colonnes.

Les différentes catégories d'études et les différentes catégories socio-professionnelles sont:

Études		Catégories socio-professionnelles	
dro	DROIT	*age*	Agriculteur exploitant
sci	SCIENCES	*oua*	Ouvrier agricole
let	LETTRE	*ind*	Industriel
med	MEDECINE	*art*	Artisan
pha	PHARMACIE	*gco*	Moyen et gros commerçant
iut	IUT	*pco*	Petit commerçant
		lib	Profession libérale
		ing	Ingénieur
		adm	Cadre de l'administration
		mes	Profession médicale et salarié
		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	Manoeuvre
		ser	Personnel de service
		arm	Armée, Police

Lecture du fichier

```
Data test;
 infile '../Pgm-s/csp.dat' firstobs=2 expandtabs;
 input origine$ dro sci let med pha iut;
run;


Data CSP;
 set test;
 length etude $4.;
 array tmp{*} dro--iut;
    do i=1 to dim(tmp);
       poids=tmp{i};
       call vname(tmp{i},etude);
       output;
    end;
 keep origine etude poids;
run;
```


Création d'un tableau SAS qu'on nomme *test* à partir du fichier externe *csp.dat*.
Les tableaux créés par SAS sont du type lignes × colonnes avec en lignes des *individus* et en colonnes des *variables*.
L'option *firstobs=2* impose la lecture du fichier à partir de la 2e ligne et l'option *expandtabs* indique que les données ne sont pas séparées par des blancs mais par des tabulations.
Le tableau *test* contient 7 colonnes ou variables *origine$ dro sci let med pha iut* correspondant à chacune des colonnes du fichier *csp.dat*.
La première variable (*origine*) est de type alphanumérique. On la fait suivre par ($). Les autres variables sont de type numérique.
Dans la deuxième partie du programme on va transformer le tableau *test.sas* en tableau de contingence (*CSP*) ne comportant plus que 3 variables,

- deux variables qualitatives : la variable *origine* qui contient les modalités relatives à l'origine socio-professionnelle des parents et la variable *etude* qui contient les modalités relatives au type d'étude suivie par les étudiants.
- une variable numérique *poids* qui contient le nombre étudiants n_{ij} dont les parents appartiennent à la CSP j et qui suivent les études i.

Pour cela on va utiliser une boucle aprés avoir créé un vecteur.

array tmp{*} dro - - iut;	on range les variables *dro*, *sci*, ..., *iut* dans un vecteur *tmp*, avec *tmp(1)=dro*, *tmp(2)=sci*, ..., *tmp(6)=iut*.
do i=1 to dim(tmp);	on commence une boucle de $i = 1$ à $i = 6$, dimension du vecteur *tmp* (= nombre de variables numériques).
poids=tmp(i);	on affecte à la nouvelle variable *poids* le contenu numérique de *tmp(i)*, c'est-à-dire selon la valeur de i, le contenu numérique de la variable *dro*, ...
call vname(tmp(i),etude);	on affecte à la nouvelle variable *etude* le label de *tmp(i)*, c'est-à-dire la valeur *dro* pour *tmp(1)*, la valeur *sci* pour *tmp(2)*, ...
output;	on remplit le tableau *CSP*.
end;	fin de la boucle.
keep origine etude poids;	on ne garde dans le tableau *CSP* que ces 3 variables.

Analyse préliminaire

```
Proc Freq data=CSP order=data;
 table origine*etude;
 weight poids;
run;

TABLE OF ORIGINE BY ETUDE

ORIGINE     ETUDE

Frequency|
Percent  |
Row Pct  |
Col Pct  |DRO     |SCI     |LET     |MED     |PHA     |IUT     | Total
---------+--------+--------+--------+--------+--------+--------+
age      |   4814 |   8941 |   8719 |   1583 |    959 |    558 | 25574
         |   1.37 |   2.54 |   2.48 |   0.45 |   0.27 |   0.16 |  7.27
         |  18.82 |  34.96 |  34.09 |   6.19 |   3.75 |   2.18 |
         |   6.52 |   8.99 |   7.36 |   3.75 |   6.77 |  14.61 |
---------+--------+--------+--------+--------+--------+--------+
oua      |    463 |   1118 |   1224 |    153 |     33 |     83 |  3074
         |   0.13 |   0.32 |   0.35 |   0.04 |   0.01 |   0.02 |  0.87
         |  15.06 |  36.37 |  39.82 |   4.98 |   1.07 |   2.70 |
         |   0.63 |   1.12 |   1.03 |   0.36 |   0.23 |   2.17 |
---------+--------+--------+--------+--------+--------+--------+
ind      |   3610 |   2608 |   3365 |   1634 |    728 |    110 | 12055
         |   1.03 |   0.74 |   0.96 |   0.46 |   0.21 |   0.03 |  3.43
         |  29.95 |  21.63 |  27.91 |  13.55 |   6.04 |   0.91 |
         |   4.89 |   2.62 |   2.84 |   3.87 |   5.14 |   2.88 |
...
---------+--------+--------+--------+--------+--------+--------+
ing      |   4444 |   8195 |   7648 |   3880 |   1307 |    151 | 25625
         |   1.26 |   2.33 |   2.17 |   1.10 |   0.37 |   0.04 |  7.28
         |  17.34 |  31.98 |  29.85 |  15.14 |   5.10 |   0.59 |
         |   6.02 |   8.24 |   6.46 |   9.19 |   9.23 |   3.95 |
---------+--------+--------+--------+--------+--------+--------+
Total       73781    99465   118397    42221    14159     3820  351843
            20.97    28.27    33.65    12.00     4.02     1.09  100.00
```


Calcul des profils ligne et colonne

Le calcul des profils ligne et des profils colonne se fait avec la procédure *FREQ*. L'option *order=data* permet de conserver l'ordre de catégories CSP telles qu'elles sont dans le tableau *CSP* (par défaut, l'ordre est alphabétique).

L'instruction *table* permet de spécifier les variables dont les modalités vont apparaitre en lignes et en colonnes dans le tableau de sortie.

L'instruction *weight=poids* indique le poids de chacune des cellules du tableau.

La sortie fournit un tableau croisant les modalités de la variable *CSP* (en ligne) et de la variable *ETUDE* (en colonne). Ici on ne présente qu'une partie du tableau.

Dans chaque cellule, on trouve l'effectif (n_{ij}), la fréquence (n_{ij}/n), l'élément du profil ligne ($n_{ij}/n_{i.}$) et l'élément du profil colonne ($n_{ij}/n_{.j}$) de la cellule exprimé en pourcentages.

On trouve, au bout de chaque ligne, sous l'intitulé *Total*, l'effectif ($n_{i.}$) et le profil marginal ($n_{i.}/n$).

On trouve, au bout de chaque colonne, sous l'intitulé *Total*: l'effectif ($n_{.j}$) et le profil ($n_{.j}/n$) marginal.

```
Proc Freq data=CSP order=data;
 table origine*etude /noprint chisq;
 weight poids;
run;


STATISTICS FOR TABLE OF ORIGINE BY ETUDE

Statistic                     DF     Value        Prob
------------------------------------------------------
Chi-Square                   105 21181.127       0.001
Likelihood Ratio Chi-Square  105 20904.564       0.001
Mantel-Haenszel Chi-Square     1   284.429       0.001
Phi Coefficient                      0.245
Contingency Coefficient              0.238
Cramer's V                           0.110

Sample Size = 351843
```


On utilise la même procédure (*FREQ*) que précédemment pour calculer différentes statistiques.

L'option *order=data* permet de conserver l'ordre des CSP telles qu'elles sont dans le tableau *CSP* (par défaut, l'ordre est alphabétique).

L'instruction *table* permet de spécifier les variables dont les modalités vont apparaître en lignes et en colonnes dans le tableau de sortie.

Le calcul des différents coefficients d'association se fait avec l'option *chisq* (l'option *noprint* évite la sortie du tableau).

L'instruction *weight=poids* indique le poids de chacune des cellules du tableau.

La sortie donne la valeur de plusieurs statistiques et de plusieurs coeff d'association :

Chi-square statistique $Q_p = \sum_i \sum_j (n_{ij} - e_{ij})^2 / e_{ij}$, avec $e_{ij} = n_{i.} n_{.j} / n$. Cette statistique suit un χ^2 à $(R-1)(C-1)$ degré de liberté, avec R le nombre de lignes et C le nombre de colonnes.

Likelihood Ratio Chi-Square statistique $G^2 = 2 \sum_i \sum_j n_{ij} \ln(n_{ij}/e_{ij})$. Suit la même distribution que Q_p avec les mêmes dl.

Mantel-Haenszel Chi-Square statistique $Q_{\mathrm{MH}} = (n-1)r^2$ avec r^2 le coefficient de corrélation entre (par défaut) les rangs des modalités des lignes et ceux des colonnes. Ce coefficient n'a de sens que si les modalités des variables possèdent un ordre. Cette statistique suit un χ^2 à 1 dl. Ce coefficient n'a pas de sens ici parce que les modalités des variables ne sont pas ordonnées.

Phi Coefficient $\phi = (n_{11} n_{22} - n_{12} n_{21}) / \sqrt{n_{1.} n_{2.} n_{.1} n_{.2}}$ pour une table 2×2, et $\phi = \sqrt{Q_p / n}$ dans les autres cas. Cette statistique varie entre -1 et 1 pour les tables 2×2 et entre 0 et 1 dans les autres cas.

Contingency Coefficient coefficient d'association $P = \sqrt{Q_p / (Q_p + n)}$. Ce coefficient est dérivé de Q_p et varie entre 0 et 1.

Cramer's V coefficient d'association dérivé de Q_p qui varie entre -1 et 1. $V = \phi$ pour les tables 2×2 et $V = \sqrt{(Q_p/n) / \min(R-1, C-1)}$ dans les autres cas.

On rejette l'hypothèse d'indépendance, allons plus loin avec l'AFC.

AFC

```
Proc Corresp data=CSP outf=tabfreq noprint;
  tables origine,etude;
  weight poids;
run;

goptions htext=2.5 hby=3 nodisplay;

Proc Datasets library=work;
  delete graphe gout/ memtype=catalog;
run;
quit;

Proc sort data=tabfreq;
 by origine;
run;

Proc Gplot data=tabfreq(where=(_type_ ='RP')) gout=graphe;
 symbol1 v=circle i=join l=3 h=5;
 axis1 label=(a=90);
 plot etude*percent /frame autovref lvref=33 vaxis=axis1;
 by origine;
run;
quit;

Goptions display;

%inc 'panels.sas';

%panels(rows=6,cols=4,plots=22,order=DOWN,
        equate=Y,gin=graphe,gout=graphe);
```


Programme de calcul et de sortie des profils ligne.

Utilisation de la procédure *Corresp* pour construire les profils lignes et colonnes et les stocker dans un tableau SAS *tabFreq*. Cette procédure sera vue plus en détail ultérieurement.

Options graphiques permettant de sortir les graphiques:

$xhtext = 2.5$	taille par défaut du texte des graphiques
$hby = 3$	taille par défaut de la variable de boucle (*by*)
nodisplay	n'affiche pas sur l'écran les graphiques produits

Efface du répertoire de travail, le catalogue (*graphe*) devant contenir les graphiques réalisés par la suite du programme. Cette opération n'est pas utile la première fois que l'on lance le programme. Par contre elle est nécessaire si l'on lance plusieurs fois le programme.

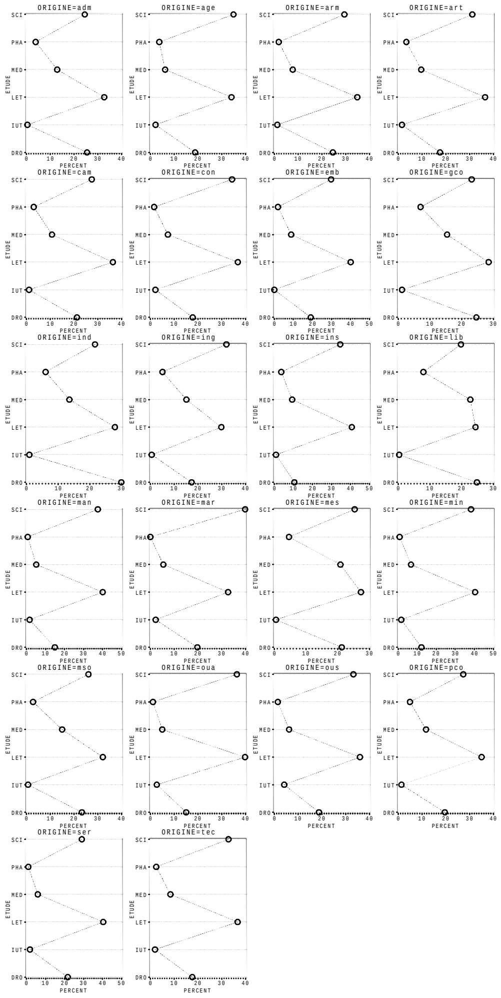
Tri du tableau *tabfreq* par ordre croissant selon les modalités de la variable *origine* (lignes du tableau), avec la procédure *Sort*.

Tracer des graphiques avec la procédure *Gplot*. Utilisation d'une partie du tableau SAS *tabfreq*: on ne s'intéresse qu'aux profils ligne (utilisation de l'option *where=(_type_ ='RP')*). Les graphiques sont stockés dans le catalogue *graphe*. Type de symbole utilisé: cercle de taille 5 relié par une ligne de taille 3. rotation de 90 degrés du label de l'axe des ordonnées. Tracer des profils ligne avec en ordonnée les modalités de la variable *etude* et en abscisse les pourcentages. Chaque graphique est entouré d'un cadre, et quadrillé. Un graphique pour chacune des modalités de la variable *origine*. Utilisation de l'instruction *by*.

Rétablit l'affichage.

Chargement d'une macro SAS permettant de mettre plusieurs graphiques sur une même page.

Utilisation de la macro : trace 22 graphiques dans une grille de 6 lignes et 4 colonnes.



Commenter les graphiques :

– rapprocher des profils analogues

– comparer au profil marginal

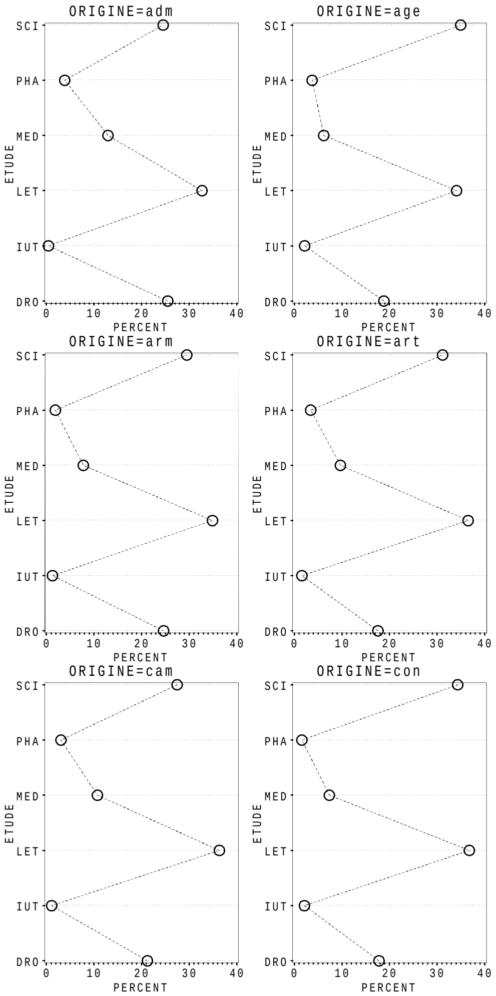
```
Proc sort data=tabFreq;
 by etude;
run;

Goptions nodisplay;

Proc Gplot data=TabFreq(where=(_type_ ='CP')) gout=graphe;
 symbol1 v=circle i=join l=3 h=5;
 axis1 label=(a=90);
 plot origine*percent /frame autovref lvref=33 vaxis=axis1;
 by etude;
run;
quit;

Goptions display;

%panels(rows=3,cols=2,plots=6,order=DOWN,
        equate=Y,gin=graphe,gout=graphe);

quit;
```


Programme de calcul et de sortie des profils colonne.

Même instructions que pour les profils ligne.

Le tracé des graphiques se fait sur une partie du fichier *tabfreq*. Ne contient que les profils colonne (*where=(_type_ ='CP')*).

Sortie sur une grille 3 lignes et 2 colonnes.



Commentaire à faire sur l'application.

```
Proc Corresp data=CSP outc=graphe dim=3;
 table origine,etude;
 weight poids;
run;
```

```
Inertia and Chi-Square Decomposition

Singular  Principal Chi-
Values    Inertias  Squares Percents   15   30   45   60   75
                                     ----+----+----+----+----+---
0.21290   0.04533     15948  75.29% *************************
0.08322   0.00693   2436.52  11.50% ****
0.07551   0.00570    2006.1   9.47% ***
0.03454   0.00119    419.73   1.98% *
0.03246   0.00105   370.738   1.75% *
          -------   -------
          0.06020   21181.1 (Degrees of Freedom = 105)
```


L'analyse factorielle des correspondances se fait avec la procédure *corresp*. les options utilisés sont:

data = nom du tableau utilisé.
outc = nom du tableau dans lequel sont stockés un certain nombre d'informations permettant entre autre de tracer les graphes.
dim = le nombre d'axes retenus (ici 3).

Indique la variable ligne et la variable colonne qui sont utilisées pour construire le tableau de contingence.

Indique la variable contenant la valeur des différentes cellules du tableau de contingence.

La procédure produit une sortie très *verbeuse*.

En colonne on a :

Singular Values les valeurs singulières

Principal Inertias les valeurs propres (égales au carré des valeurs singulières) associées à chaque axe (inertie) et l'inertie totale. L'inertie est égale χ^2/n. Ici, l'inertie totale est égale $21181.1/351843 = 0.06020$

Chi-Squares une décomposition du χ^2, sa valeur totale et le nombre de degrés de liberté

Percents le pourcentage d'inertie associé à chaque axe

SAS fournit aussi un éboulis des pourcentages d'inertie.

Row Coordinates

	Dim1	Dim2	Dim3
adm	0.093103	-.097952	-.037983
age	-.204953	0.024170	0.096188
arm	-.119086	-.134009	0.009826
art	-.127387	0.042609	0.011147

Summary Statistics for the Row Points

	Quality	Mass	Inertia
adm	0.993601	0.156809	0.051659
age	0.901919	0.072686	0.069400
arm	0.943464	0.029894	0.016967
art	0.974484	0.046421	0.014376

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2	Dim3
adm	0.029987	0.217257	0.039678
age	0.067359	0.006132	0.117947
arm	0.009353	0.077523	0.000506
art	0.016619	0.012170	0.001012

Squared Cosines for the Row Points

	Dim1	Dim2	Dim3
adm	0.437071	0.483784	0.072746
age	0.730793	0.010163	0.160963
arm	0.415048	0.525590	0.002826
art	0.870434	0.097384	0.006665

Row Coordinates:

Coordonnées des lignes sur les différents axes retenus (*Dim1* … *Dim3*).

Summary Statistics for the Row Points:

Quality =Qualité de la représentation des lignes dans le sous espace de dimension 3. La valeur correspond à la somme des contributions relatives (*Squared cosines*) des 3 premiers axes sur la modalité. Pour la modalité $age = 0.7307930 + 0.010163 + 0.160963 = 0.901919$

Mass = masse affectée au point-ligne et qui correspond à la fréquence marginale $n_{i.}/n = f_{i.}$. Pour la modalité $age = 25574/351843 = 0.072686$

Inertia = contribution du point-ligne à l'inertie totale.

Partial Contributions to Inertia for the Row Points:

contribution du point-ligne à l'inertie de l'axe. La somme en colonne est égale à 1 (100 %), la somme en ligne est égale à *Quality*.

Squared Cosines for the Row Points:

Qualité de la représentation des points-lignes par les différents axes.

Column Coordinates

	Dim1	Dim2	Dim3
DRO	0.138733	-.140311	0.030680
IUT	-.500889	0.037051	0.629807
LET	-.126509	-.012292	-.049852
MED	0.390065	0.086375	-.017204

Summary Statistics for the Column Points

	Quality	Mass	Inertia
DRO	0.990697	0.209699	0.140204
IUT	0.972333	0.010857	0.120362
LET	0.937254	0.336505	0.111173
MED	0.979480	0.120000	0.325426

Partial Contributions to Inertia for the Column Points

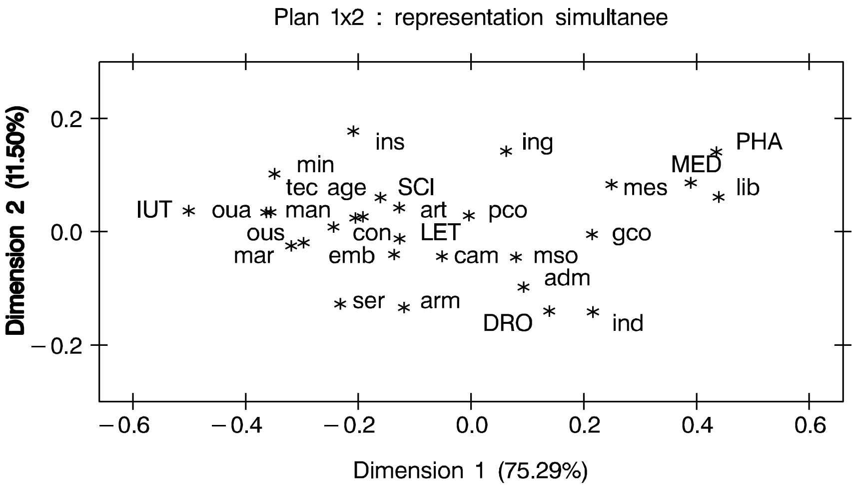
	Dim1	Dim2	Dim3
DRO	0.089043	0.596155	0.034619
IUT	0.060095	0.002152	0.755312
LET	0.118817	0.007341	0.146671
MED	0.402806	0.129280	0.006229

Squared Cosines for the Column Points

	Dim1	Dim2	Dim3
DRO	0.478186	0.489126	0.023386
IUT	0.375929	0.002057	0.594346
LET	0.804704	0.007596	0.124953
MED	0.931969	0.045698	0.001813

Mêmes explications que pour les lignes.

```
options nodate center;
title 'Plan 1x2 : representation simultanee';

%plotit(data=graphe,plotvars=dim2 dim1,labelvar=_name_,
        datatype=corresp,
        method=print,post=plan12.eps,
        gopprint=dev=psepsf gsfmode=replace gaccess=gsasfile,
        monochro=black);
```

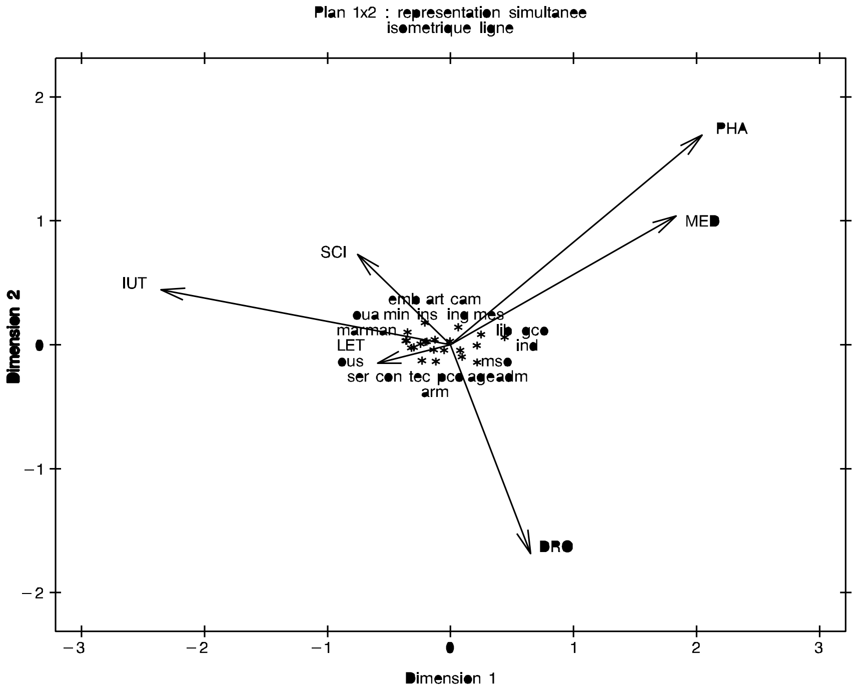


Représentation simultanée des lignes et des colonnes.

C'est la représentation par défaut fournie par SAS.

```
* -------- Représentation simultanée isométrique ligne;

title 'Plan 1x2 : representation simultanee';
title2 ' isometrique ligne';


Proc corresp data=CSP outc=graphe dim=5 profile=row noprint;
 table origine,etude;
 weight poids;
run;

%plotit(data=graphe,plotvars=dim2 dim1,labelvar=_name_,
        datatype=row,method=print,post=plan12l.eps,
        gopprint=dev=psepsf gsfmode=replace gaccess=gsasfile,
        monochro=black);
```

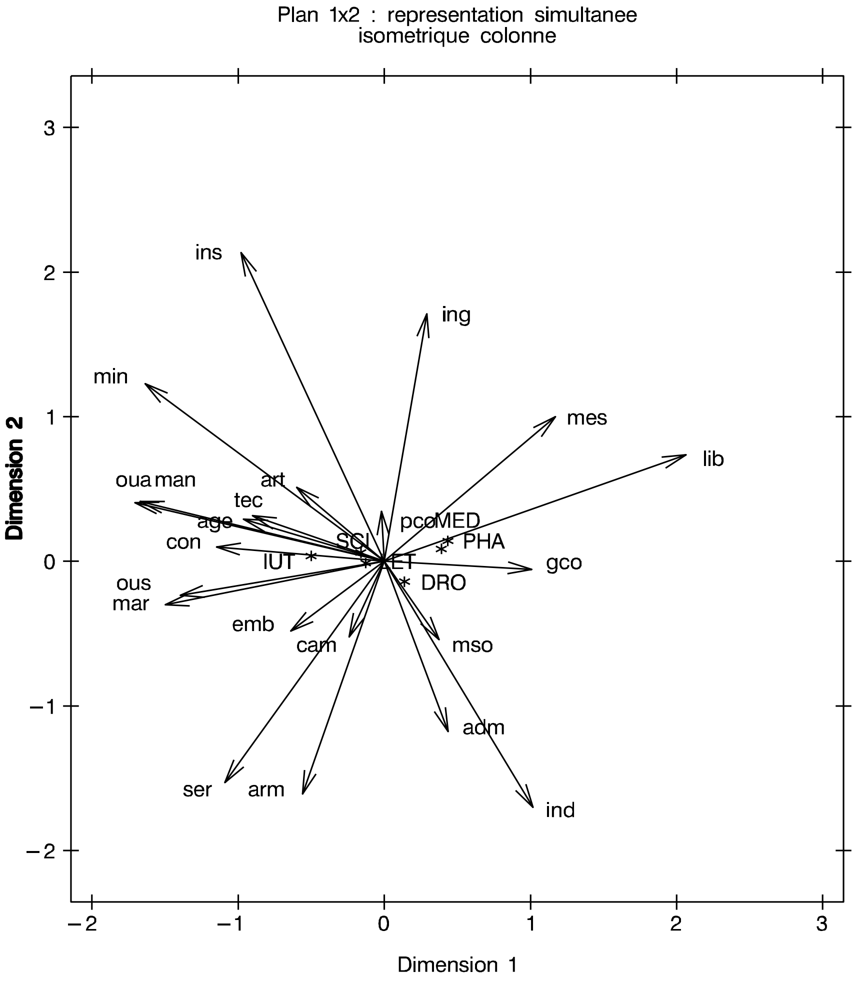

On a la reprsentation isométrique grâce à l'option *profile=row*. C'est le paramètre *datatype=row* (au lieu de *=corresp*) qui donne les flèches dans le représentation.



Plan 1x2 : representation simultanee
isometrique ligne
PHA
MED
SCI
IUT
emb art cam
oua min ins ing mes
marman
lib gce
LET
ind
ous
mse
ser con tec pco ageadm
arm
DRO
Dimension 1
Dimension 2

16

```
* -------- Représentation simultanée isométrique colonne;

title 'Plan 1x2 : representation simultanee';
title2 ' isometrique colonne';

Proc corresp data=CSP outc=graphe dim=5 profile=column noprint;
 table origine,etude;
 weight poids;
run;


%plotit(data=graphe,plotvars=dim2 dim1,labelvar=_name_,
        datatype=column,method=print,post=plan12c.eps,
        gopprint=dev=psepsf gsfmode=replace gaccess=gsasfile,
        monochro=black);
```


Plan 1x2 : representation simultanee
isometrique colonne
Dimension 1
Dimension 2
−2
−1
0
1
2
3
ins
ing
min
mes
lib
oua
man
art
tec
age
pco
MED
PHA
con
SCI
IUT
LET
gco
DRO
ous
mar
emb
cam
mso
adm
ser
arm
ind