ANALYSE DES CORRESPONDANCES SIMPLE (AFC) ET MULTIPLE (ACM)

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

PLAN

L'analyse des correspondances simple (AFC)	**Transparent**
Présentation	3
Tableau de contingence, de fréquences relatives	4 à 6
Liaison entre variables qualitatives	7 à 9
Tableaux dérivés de la table N	10
L'AFC définie comme une ACP	11
Une définition équivalente de l'AFC	12
Caractéristiques de l'AFC	13 à 16
Pratique de l'AFC, éboulis des valeurs propres	17
Qualité de la représentation	18 à 19
Contributions	20
Biplots	21 à 26
Quelques formules	27 à 28
L'analyse des correspondances multiple (ACM)	
Présentation	29
Exemple de données	30 à 31
Tableau d'indicatrices	32
Tableau de Burt	33
ACM : AFC du tableau d'indicatrices	
Distance entre individus	34
Distance entre réponses	35
AFC du tableau Z	36
Résultats	37 à 40
Biplots	41
Variables illustratives	42
ACM : analyse du tableau de Burt	43
Résultats	44 à 45
Représentation isométrique ligne (ou colonne)	46

INRA ©1998 - Formation statistique - Module 7/8 (AFC - ACM)-2

L'analyse des correspondances simple : AFC

Méthode d'analyse de type ACP décrivant les liaisons entre 2 variables qualitatives. Elle utilise en entrée la table de contingence croisant ces deux variables.

Si plus de deux variables sont considérées, une extension de l'AFC doit être effectuée (cf. ACM).

INRA ©1998 - Formation statistique - Module 7/8 (AFC - ACM)-3

Tableau de contingence N

Le tableau N ventile un échantillon de taille n selon les modalités de 2 variables qualitatives : l'une en ligne avec I modalités et l'autre en colonne avec J modalités.

Notations

n_{11}	$\dots$	n_{1j}	$\dots$	n_{1J}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{i1}	$\dots$	n_{ij}	$\dots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{I1}	$\dots$	n_{Ij}	$\dots$	n_{IJ}	$n_{I.}$
$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.J}$	n

avec

$$\mathbf{I} = \{1, \dots, I\}$$
$$\mathbf{J} = \{1, \dots, J\}$$
$$n_{.j} = \sum_{i \in I} n_{ij}$$
$$n_{i.} = \sum_{j \in J} n_{ij}$$
$$n = \sum_{i \in I} \sum_{j \in J} n_{ij}$$

– Bien souligner que les données du tableau N correspondent à des effectifs : les n_{ij} sont les effectifs d'individus de l'échantillon présentant simultanément la modalité i de la variable I et la modalité j de la variable J.

– Une AFC sur tableau de contingence se justifie si :

. l'on est intéressé par l'analyse des relations entre les catégories de l'une et l'autre variable.

. l'on dispose de nombreuses modalités par variable (au moins trois) et suffisamment bien remplies (par regroupement au besoin).

– Notation : les modalités en lignes sont indicées par i et celles en colonnes en j alors que le rôle des deux variables dans l'analyse est complétement symétrique ; il n'y a pas *d'individus* ni de *variables* au sens de l'ACP.

Rappel : en ACP, les valeurs des lignes i et colonnes j, notées par convention x_i^j, correspondent à la valeur prise par l'individu i pour la variable j.

Exemple de tableau de contingence

	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?

Tableau des fréquences relatives F

On considère le tableau F des fréquences relatives obtenu en divisant l'effectif n_{ij} de chaque cellule du tableau N par l'effectif total n.

$$\begin{array}{ccccc|c} f_{11} & \dots & f_{1j} & \dots & f_{1J} & f_{1.} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{i1} & \dots & f_{ij} & \dots & f_{iJ} & f_{i.} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{I1} & \dots & f_{Ij} & \dots & f_{IJ} & f_{I.} \\ \hline f_{.1} & \dots & f_{.j} & \dots & f_{.J} & 1 \end{array}$$

avec

$$\begin{aligned} &\mathbf{I} = \{1, \dots, I\} \\ &\mathbf{J} = \{1, \dots, J\} \\ &f_{ij} = n_{ij}/n \\ &f_{.j} = \sum_{i \in I} f_{ij} \\ &f_{i.} = \sum_{j \in J} f_{ij} \\ &\sum_{i \in I} \sum_{j \in J} f_{ij} = 1 \\ &\sum_{i \in I} f_{i.} = \sum_{j \in J} f_{j.} = 1 \end{aligned}$$

Liaison entre variables qualitatives

On dit que deux variables sont liées (ou ne sont pas indépendantes) si la connaissance de la valeur de l'une modifie la distribution attendue de l'autre.

En terme de probabilité, l'indépendance s'écrit :

$\forall i \in I$ et $j \in J$:

$$\begin{array}{lcl} \Pr(i \text{ et } j) & = & \Pr(i)\Pr(j) \\ \dfrac{\Pr(i \text{ et } j)}{\Pr(i)\Pr(j)} & = & 1 \end{array}$$

On estime

$$\begin{array}{lll} \Pr(i \text{ et } j) & \text{par} & f_{ij} = n_{ij}/n, \\ \Pr(i) & \text{par} & f_{i.} = n_{i.}/n \\ \Pr(j) & \text{par} & f_{.j} = n_{.j}/n. \end{array}$$

Il y aura liaison entre les 2 variables si pour certains couples (i, j),

$$f_{ij} \neq f_{i.}f_{.j} \text{ significativement}$$

ou si (ce qui est équivalent)

$$\frac{f_{ij}}{f_{i.}f_{.j}} \neq 1 \text{ significativement}$$

– Bien montrer qu'on passe d'un tableau d'effectifs (n_{ij}) à des fréquences $(f_{ij}, f_{i.}, f_{.j})$ qui sont des estimations des probabilités $\Pr(i \text{ et } j)$, $\Pr(i)$ et $\Pr(j)$;
($\Longrightarrow$ attention aux conclusions si $n_{i.}$ ou $n_{.j}$ petits).

– Pratiquement, l'AFC va traiter des écarts à l'indépendance $(\dfrac{f_{ij}}{f_{i.}f_{.j}} - 1)$ ou encore $(\dfrac{f_{ij}}{f_{i.}} - f_{.j})$, $(\dfrac{f_{ij}}{f_{.j}} - f_{i.})$.

Liaison entre variables qualitatives

Le vecteur $f_{ij}/f_{i.}$ (avec $j \in J$) est appelé le i*ème profil ligne*.

Le vecteur $f_{ij}/f_{.j}$ (avec $i \in I$) est appelé le j*ème profil colonne*.

En cas d'indépendance tous les profils ligne (respectivement profils colonne) sont égaux au profil marginal $f_{.j}$ avec $j \in J$ (respectivement $f_{i.}$ avec $i \in I$).

$$\frac{f_{ij}}{f_{i.}} = f_{.j} \qquad \frac{f_{ij}}{f_{.j}} = f_{i.}$$

Les notions de liaisons s'expriment simplement pour les vrais valeurs des probabilités (les deux variables sont liées si pour certains couple (i,j) $\Pr(i \text{ et } j) \neq \Pr(i)\Pr(j)$.

En terme d'estimation, il ne suffit pas que les fréquences qui estiment ces probabilités soient telles que $f_{ij} \neq f_{i.}f_{.j}$ pour certains couples. Il faut en plus que cette différence soit significative. Cette significativité sera testée par la statistique du χ^2.

Le i*ème* profil correspond à la fréquence conditionnelle de j sachant i.

Une mesure globale de la liaison

Sous l'hypothèse d'indépendance, la quantité :

$$\chi^2 = \sum_{i,j} \left[\frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}} \right]$$

suit une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté.

Cette quantité s'écrit encore:

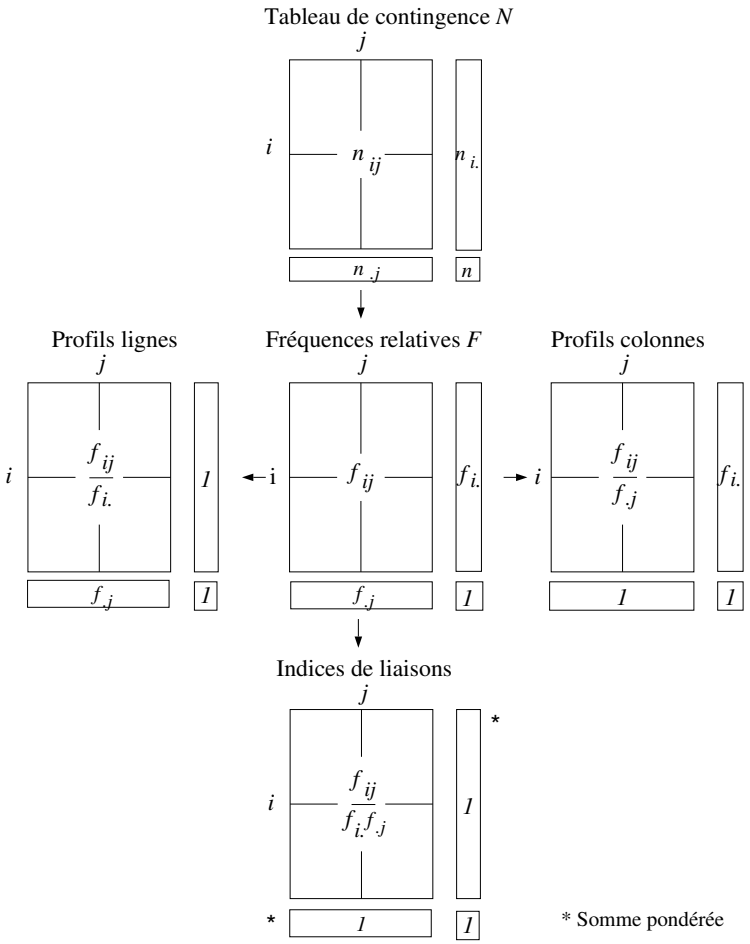
$$\chi^2 = n \sum_{i,j} f_{i.} f_{.j} \left(\frac{f_{ij}}{f_{i.} f_{.j}} - 1 \right)^2$$

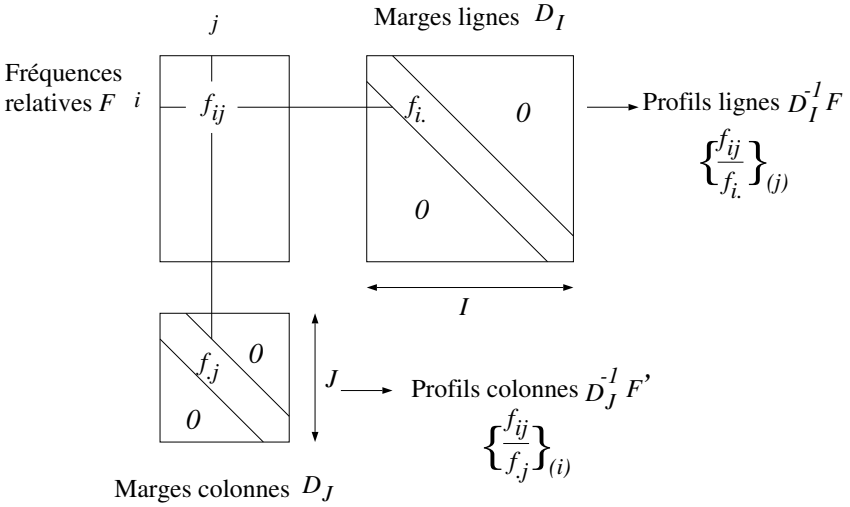
C'est une mesure de distance de la distribution observée à l'indépendance :

- Si le χ^2 est grand et supérieur à un seuil X_α, on rejette l'hypothèse d'indépendance : il y a liaison, due à des écarts significatifs entre $f_{ij}/(f_i f_j)$ et 1 pour au moins certains couples (i, j).
- Si le χ^2 est petit, on accepte l'hypothèse de non liaison : il n'y a rien à expliquer.

Vérifier les conditions d'utilisation du test du χ^2.

Tableaux dérivés de la table N





L'AFC définie comme une ACP

L'AFC est obtenue par l'ACP du tableau des indices de liaison $(f_{ij}/f_{i.}f_{.j})$ avec

- les poids $(f_{i.},\ i \in I)$: $D = D_I = diag(f_{i.})$,
- la métrique $M = D_J = diag(f_{.j})$.

Le tableau centré approximé par l'ACP a pour éléments $(f_{ij}/f_{i.}f_{.j} - 1)$ et a pour DVS :

$$f_{ij}/f_{i.}f_{.j} - 1 = \sum_{s=1}^{r} \lambda_s u_i^s v_j^s. \qquad (1)$$

Dans cette ACP,

- Les facteurs $\mathbf{u}_s$ sont orthonormés pour la métrique D_I,
- les facteurs $\mathbf{v}_s$ sont orthonormés pour la métrique D_J,
- l'inertie totale du tableau est égale à la mesure globale de liaison :

$$\frac{\chi^2}{n} = \sum_{ij} f_{i.}f_{.j} \left(\frac{f_{ij}}{f_{i.}f_{.j}} - 1 \right)^2$$

On appelle *représentation simultanée* de l'AFC la superposition des représentations isométriques des lignes et des colonnes sur un même graphique. Dans ces représentations, la coordonnée de la ligne i (resp. colonne j) sur l'axe s est $\lambda_s u_i^s$ (resp. $\lambda_s v_j^s$).

– Dans une table de contingence, les lignes et les colonnes jouent des rôles symétriques. Pour l'AFC, on remplace dans l'exposé de l'ACP le terme individu par *ligne* et le terme variable par *colonne*.

– Alors qu'en ACP on parle de facteurs $\mathbf{u}_s$ et de vecteurs principaux $\mathbf{v}_s$, ici on ne parle que de facteurs $\mathbf{u}_s$ dans $\mathbf{R}^I$ ou $\mathbf{v}_s$ dans $\mathbf{R}^J$.

– La justification de la représentation simultanée sera donnée ultérieurement.

Une définition équivalente de l'AFC

Autre façon de voir : l'AFC comme une double ACP :

- **L'ACP du tableau des profils lignes** $(f_{ij}/f_{i.},\ j \in J)$,
  - munis des poids $(f_{i.})$,
  - avec la métrique $D_J^{-1} = diag(1/f_{.j})$ (appellée *métrique du* χ^2).

  La DVS de cette ACP s'obtient en multipliant l'équation (1) de la DVS précédente par $f_{i.}$:

  $$f_{ij}/f_{i.} - f_{.j} = \sum_{s=1}^{r} \lambda_s u_i^s (f_{.j} v_j^s).$$

  Cette ACP fournit la même représentation des lignes que l'ACP des indices de liaison.

- **L'ACP du tableau des profils colonnes** $(f_{ij}/f_{.j},\ i \in I)$,
  - munis des poids $(f_{.j})$,
  - avec la métrique $D_I^{-1} = diag(1/f_{i.})$.

  La DVS de cette ACP s'obtient en multipliant l'équation (1) par $f_{.j}$:

  $$f_{ij}/f_{.j} - f_{i.} = \sum_{s=1}^{r} \lambda_s v_j^s (f_{i.} u_i^s).$$

  Cette ACP fournit la même représentation des colonnes que l'ACP des indices de liaison.

De plus ces trois ACP ont mêmes inerties totales, mêmes valeurs propres. Les contributions et indices de qualité sont aussi identiques.

- Une ligne i peut-être considérée comme la représentation du profil ligne i. Deux lignes i et i' sont proches si les profils lignes associés sont semblables. Un résultat analogue vaut pour les colonnes.
- Le profil ligne moyen $(f_{.j},\ j \in J)$ est le profil marginal ligne. Il est à l'origine des axes. Une ligne i proche de l'origine a un profil proche du profil moyen. Un résultat analogue vaut pour les colonnes.

Caractéristiques de l'AFC

– Le double centrage

– Les propriétés barycentriques

– l'équivalence distributionnelle

INRA ©1998 - Formation statistique - Module 7/8 (AFC - ACM)-13

Double centrage

– L'ACP opère un centrage des variables par leurs moyennes.

– L'AFC opère un centrage sur les colonnes qui est aussi un centrage sur les lignes.

En ACP, on est amené à étudier les proximités entre individus et les cosinus d'angles (les corrélations) entre variables centrées.

En AFC, on n'analyse pas de corrélations mais la configuration de deux nuages centrés des lignes et des colonnes.

$$\sum_i f_{i.}(\frac{f_{ij}}{f_{i.}f_{.j}} - 1) = 0 \qquad \text{et} \qquad \sum_j f_{.j}(\frac{f_{ij}}{f_{i.}f_{.j}} - 1) = 0$$

INRA ©1998 - Formation statistique - Module 7/8 (AFC - ACM)-14

Propriétés barycentriques

Dans la base des vecteurs principaux $s = 1, \ldots, r$

- *le point ligne* i, de coordonnées $\lambda_s u_i^s = \sum_j \frac{f_{ij}}{f_{i.}} v_s^j$ est au barycentre des points j de coordonnées v_s^j. Chaque point j est affecté du poids $\frac{f_{ij}}{f_{i.}}$. Les poids sont les éléments du profil de la *ième* ligne.

- *le point colonne* j, de coordonnées $\lambda_s v_s^j = \sum_i \frac{f_{ij}}{f_{.j}} u_i^s$ est au barycentre des points de coordonnées u_i^s. Chaque point i est affecté du poids $\frac{f_{ij}}{f_{.j}}$. Les poids sont les éléments du profil de la *jème* colonne.

Cette propriété est à la base de l'interprétation barycentrique de l'AFC.

Des exemples sont donnés par la suite.

Équivalence distributionnelle

Cette propriété est une propriété de stabilité de l'AFC : si deux lignes de la table de contingence sont proportionnelles, on ne modifie pas l'AFC de la table en remplaçant les deux lignes par leur somme. La propriété est vraie aussi pour les colonnes.

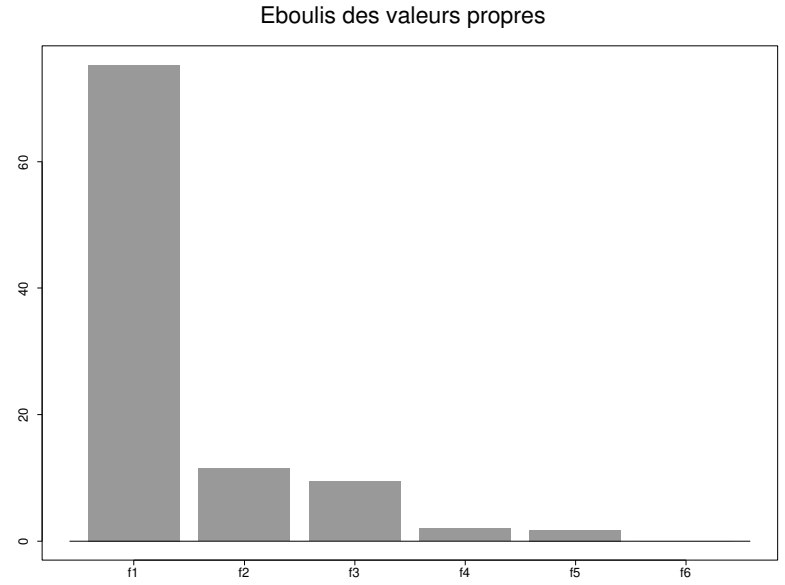
Cette propriété permet de regrouper des catégories d'une variable qui sont homogènes vis-à-vis des catégories de l'autre variable (mêmes profils).

Intéressant car la définition des catégories comporte une part d'arbitraire.

Redécouper en 2 une catégorie homogène ne change pas les résultats de l'AFC.

INRA ©1998 - Formation statistique - Module 7/8 (AFC - ACM)-16

Pratique de l'AFC



facteurs	f1	f2	f3	f4	f5	tot
valeurs singulieres	0.2129	0.0832	0.0755	0.0346	0.0325	
valeurs propres	0.0453	0.0069	0.0057	0.0012	0.0011	0.0602
inertie exp	75.3 %	11.5 %	9.5 %	2.0 %	1.8 %	100 %
inertie cum	75.3 %	86.8 %	96.3 %	98.2 %	100 %	100 %

– 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.

– $\sum \lambda_s^2 = \frac{\chi^2}{n}$ est une mesure globale de dépendance.

– Le nombre maximum de valeurs propres non nulles est le $min(I-1, J-1)$ soit ici 5. Dans un but descriptif, on n'interprète rarement plus de 5 facteurs.

Cosinus carre lignes
axes principaux

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

Cosinus carre lignes
sous-espaces principaux

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

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

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

Dans le second tableau, les cosinus carré sont cumulés : on obtient la qualité de la représentation du profil i sur le *sème* 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 carre colonnes axes principaux

     f1  f2  f3 f4 f5  TOT

dro 478 489  23  4  5 1000
sci 830 118   2 14 36 1000
let 805   8 125 11 52 1000
med 932  46   2 18  2 1000
pha 808  85  33 46 29 1000
iut 376   2 594 23  4 1000

TOT 753 115  95 20 18 1000

Cosinus carre colonnes sous-espaces principaux

     f1 f1a2 f1a3 f1a4  TOT

dro 478  967  991  995 1000
sci 830  948  950  964 1000
let 805  812  937  948 1000
med 932  978  979  998 1000
pha 808  892  925  971 1000
iut 376  378  972  996 1000

TOT 753  868  963  982 1000
```

Pour les colonnes, on a un résultat semblable à ceux du transparent 16 :

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

Contribution des lignes aux facteurs

	f1	f2	f3	f4	f5	TOT	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

Contribution des colonnes aux facteurs

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

– 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

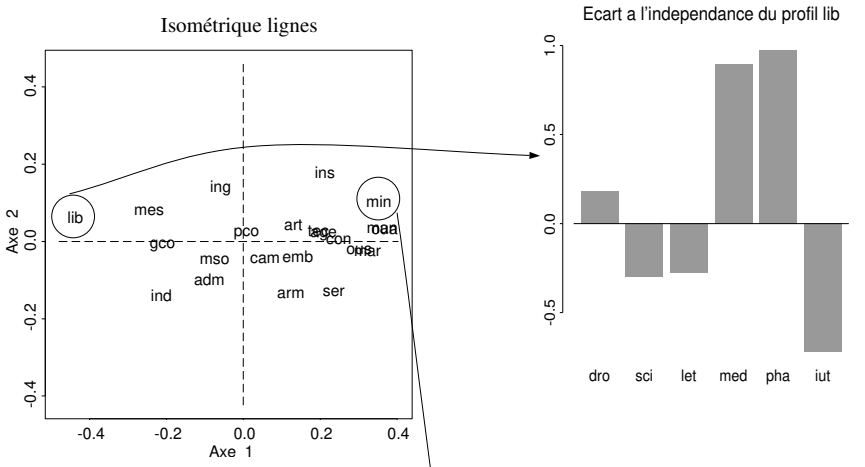
iut contribue à 75.5 % de l'inertie de l'axe 3

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, ...).
Ex : *pco* représente 21.7 % de l'inertie de l'axe 5.

– ORI = originalité d'un profil par rapport au profil marginal (centre du nuage).

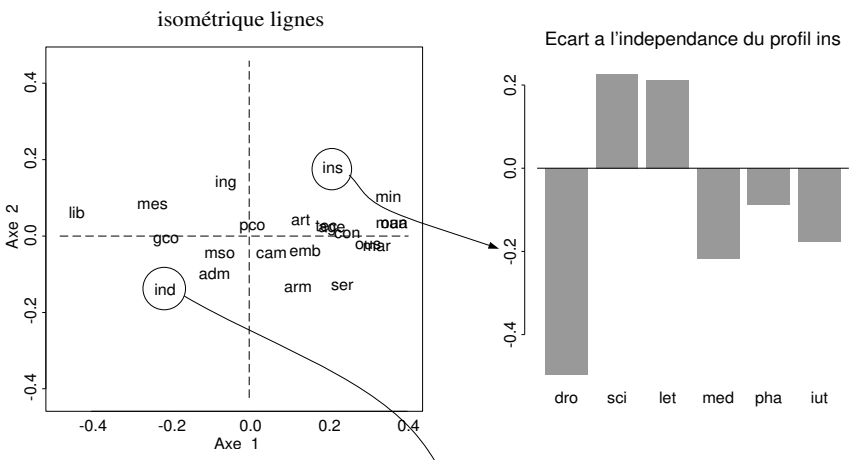
ORI = norme du vecteur - moyenne au sens de la distance du khi-deux.

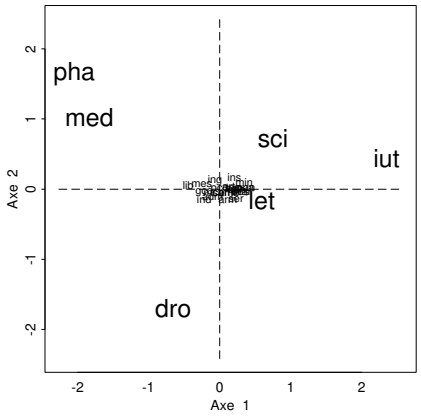


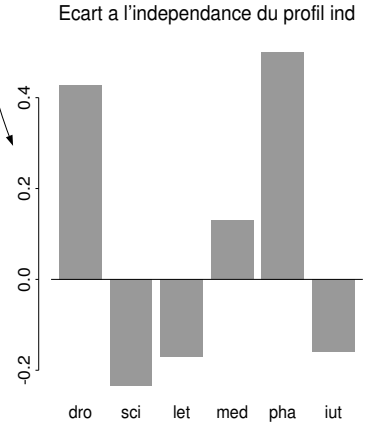
En AFC, la représentation barycentrique est aussi un biplot :

- 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 (cf. transparent 13).
- 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)



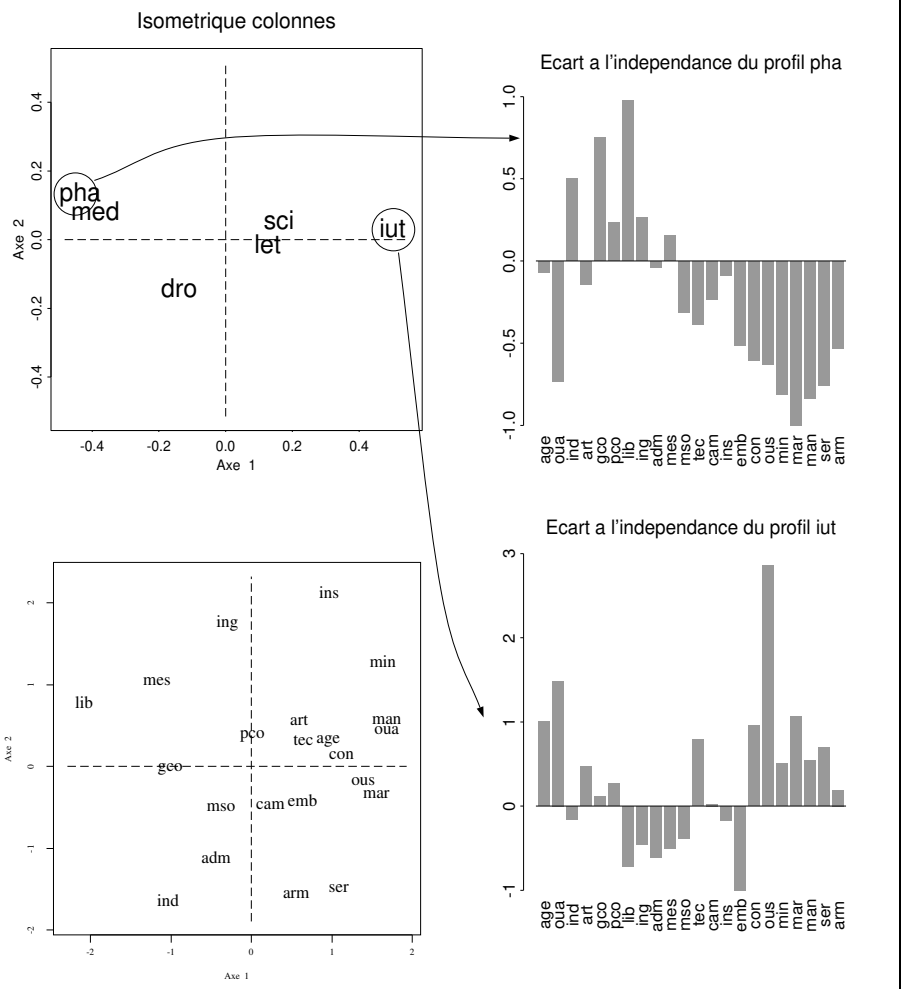




– *ins* est attiré par science et lettre.

– *ind* est attiré par droit et pharmacie.

Les représentations barycentriques sont souvent faites sur 2 graphiques séparés pour des raisons de lisibilité.

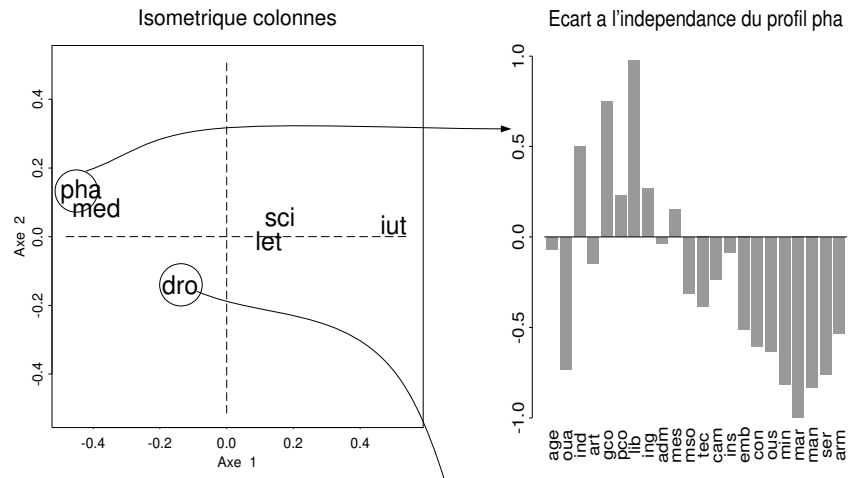


– Représentation *isométrique* colonnes donc en $\lambda_s v_s^j$ pour les colonnes et u_i^s pour les lignes. C'est aussi un biplot. Remarquer les différences d'échelles des graphes.

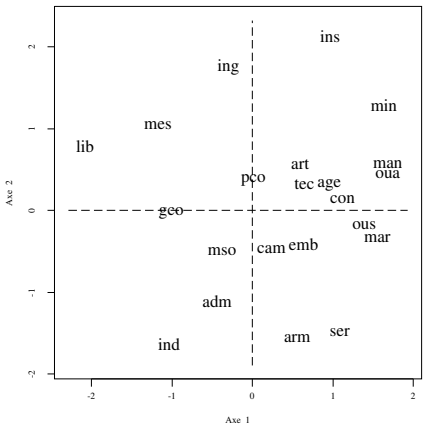
– Les points colonnes sont au barycentre des points de coordonnées u_i^s (pour les poids : $f_{ij}/f_{.j},\ i = 1, \ldots, I$) formant le *jème* profil ligne.

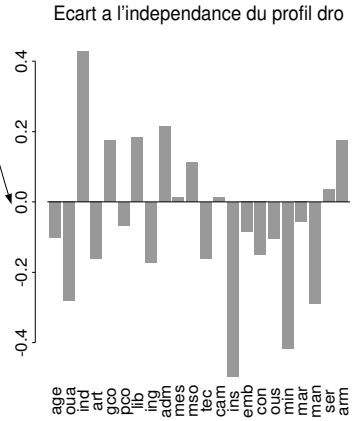
– Sur l'axe 1, l'opposition de *pha* et de *iut* est cohérente avec l'antagonisme presque systématique des profils d'écart à l'indépendance pour toutes les catégories i.

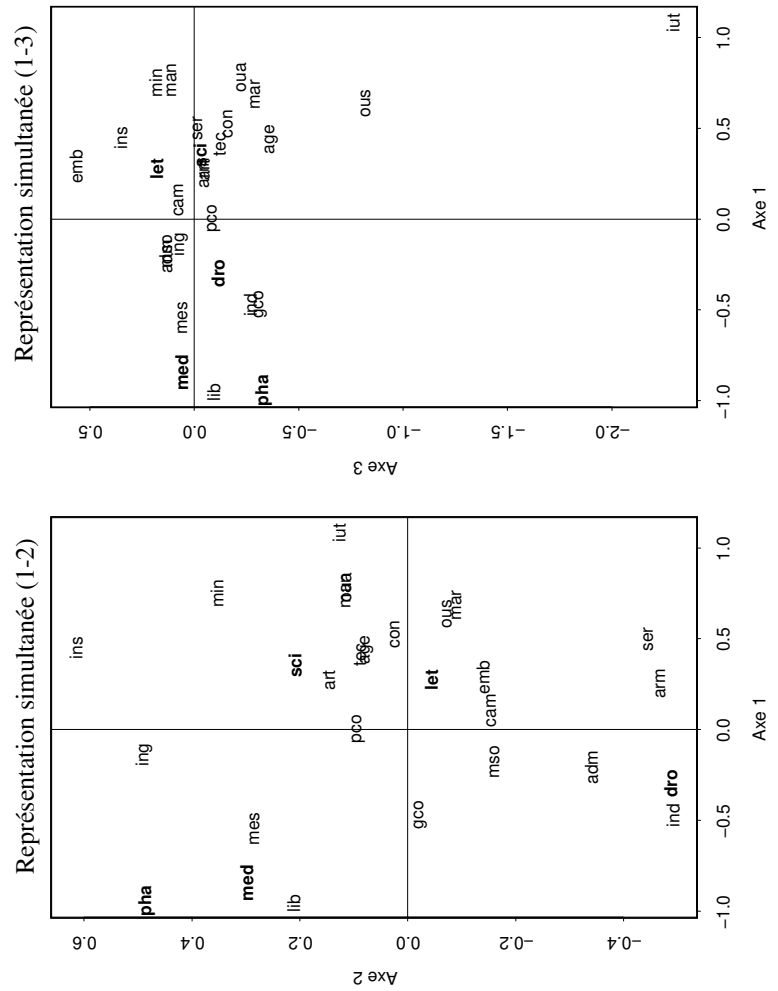
– Même phénomène, en moins accentué, sur l'axe 2 pour *pha* vis-à-vis de *dro*.


Isometrique colonnes
pha
med
sci
let
iut
dro
Axe 2
Axe 1
-0.4
-0.2
0.0
0.2
0.4

Ecart a l'independance du profil pha
1.0
0.5
0.0
-0.5
-1.0
age oua ind art gco pco lib ing adm mes mso tec cam ins emb con ous min mar man ser arm


ins
ing
min
mes
lib
art
man
oua
pco
tec age
con
geo
ous
mar
mso
cam emb
adm
arm
ser
ind
Axe 2
Axe 1


Ecart a l'independance du profil dro
0.4
0.2
0.0
-0.2
-0.4
age oua ind art gco pco lib ing adm mes mso tec cam ins emb con ous min mar man ser arm



– 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.

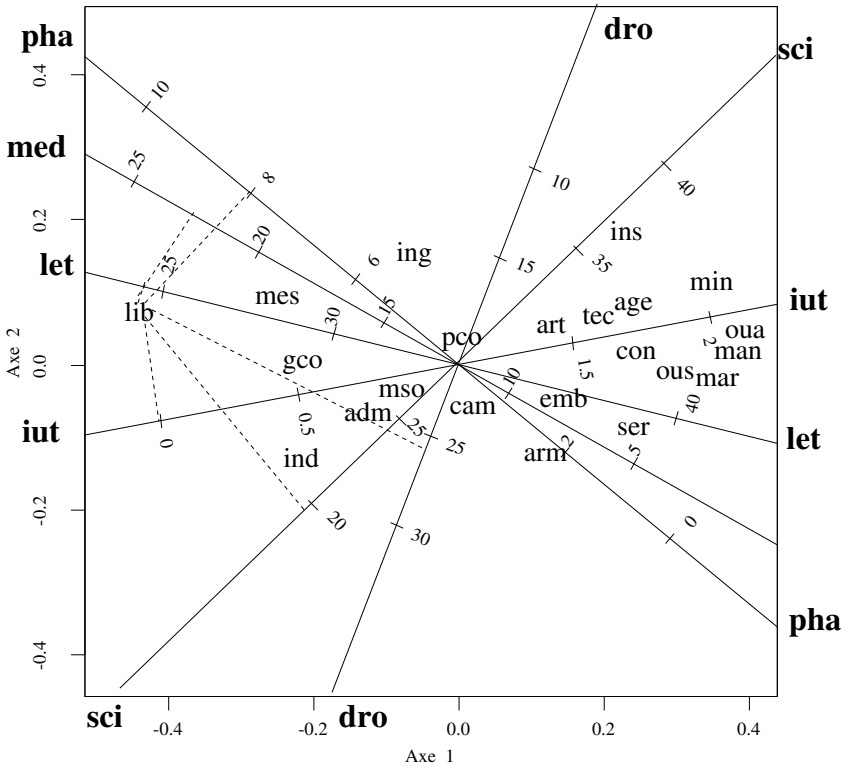
Par exemple,

lib est associé à *med* et *pha*

ind, *adm* avec *dro*

ins, *min* avec *sci*

age, *ous* avec *iut*



– 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 $100 f_{ij}/f_{i.}$.

$\Rightarrow$ 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 $\Pr(i|j)$.

AFC - Quelques formules

Soit une table de contingence $I \times J$ à i et j modalités où n_{ij} représente les effectifs d'individus partageant la modalité i et la modalité j.

L'analyse des lignes

Profil ligne moyen : $f_{.j} = n_{.j}/n,\ j = 1 \ldots, J$

Profil de la ligne i : $f_{ij}/f_{i.} = n_{ij}/n_{i.} = \Pr(j|i),\ j = 1 \ldots, J$

Écart à l'indépendance du profil de la ligne i : $f_{ij}/f_{i.} - f_{.j},\ j = 1 \ldots, J$

Distance du χ^2 de la ligne i au centre du nuage des lignes :

$$\sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 = \sum_j f_{.j} \left(\frac{f_{ij}}{f_{i.} f_{.j}} - 1 \right)^2 = \sum_s (\lambda_s u_i^s)^2$$

Contribution à l'inertie de la ligne i :

$$\sum_j f_{i.} f_{.j} \left(\frac{f_{ij}}{f_{i.} f_{.j}} - 1 \right)^2 = f_{i.} \sum_s (\lambda_s u_i^s)^2$$

Cosinus carré de la ligne i à l'axe s : $\dfrac{(\lambda_s u_i^s)^2}{\sum_s (\lambda_s u_i^s)^2}$

L'analyse des colonnes

Profil colonne moyen : $f_{i.} = n_{i.}/n,\ i = 1 \ldots, I$

Profil de la colonne j : $f_{ij}/f_{.j} = n_{ij}/n_{.j} = \Pr(i|j)$

Écart à l'indépendance du profil de la colonne j : $f_{ij}/f_{.j} - f_{i.},\ i = 1 \ldots, I$

Distance du χ^2 de la colonne j au centre du nuage des lignes :

$$\sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2 = \sum_i f_{i.} \left(\frac{f_{ij}}{f_{i.} f_{.j}} - 1 \right)^2 = \sum_s \left(\lambda_s v_s^j \right)^2$$

Contribution à l'inertie de la colonne j :

$$\sum_i f_{i.} f_{.j} \left(\frac{f_{ij}}{f_{i.} f_{.j}} - 1 \right)^2 = f_{.j} \sum_s \left(\lambda_s v_s^j \right)^2$$

Cosinus carré de la colonne j à l'axe s : $\dfrac{\left(\lambda_s v_s^j \right)^2}{\sum_s \left(\lambda_s v_s^j \right)^2}$

L'analyse des correspondances multiples : ACM

– Elle permet d'étudier les liaisons entre plus de deux variables qualitatives, c'est une généralisation de l'AFC.

– Pratiquement elle se ramène à l'AFC d'un tableau particulier (le tableau de Burt ou le tableau d'indicatrices) et elle ne prend en compte que les liaisons deux à deux entre les variables qualitatives.

– Objectifs :

. but descriptif :
bilan des liaisons entre les variables
bilan des ressemblances entre modalités
bilan des ressemblances entre individus (typologie)

. but de compression ou de recodification :
réduction du nombre de variables ou de modalités.

– Elle peut aussi s'appliquer aux variables quantitatives après leur codification en classes (mais cela implique une perte d'information). Elle permet donc l'étude simultanée de variables quantitatives discrétisées et qualitatives (non ordonnées).

– Elle permet de décrire des liaisons non linéaires entre variables quantitatives codifiées, ce que ne peut faire l'ACP. Pour discrétiser une variable quantitative, on cherche à créer des classes d'effectifs à peu près égaux, tout en tenant compte de l'intérêt de l'étude (certaines classes à effectif faible peuvent être importantes à conserver).

– Elle permet de générer une base orthogonale du sous-espace engendré par les indicatrices dans l'espace des variables. Lorsque les variables sont linéairement dépendantes on ne peut pas faire directement d'analyse discriminante.

– Cette méthode est très utilisée, mais elle ne s'interprète facilement que lorsque les variables partagent beaucoup de liaisons en commun (forte redondance entre les variables).

Un exemple de données

57 variétés de pommes sont caractérisées par 9 variables qualitatives codées :

- *arb* = forme de l'arbre, caractérisée par 4 modalités (1 = colonnaire ; 2 = spur ; 3 = étalé ; 4 = très étalé)
- *rec* = date de récolte, caractérisée par 3 modalités (1 = précoce ; 2 = intermédiaire ; 3 = tardif ou très tardif)
- *cal* = calibre du fruit, caractérisé par 3 modalités (1 = petit ou moyen ; 2 = gros ; 3 = très gros)
- *coul* = couleur dominante du fruit ou intensité de la couleur rouge additionnelle, caractérisée par 4 modalités (1 = jaune ou vert ; 2 = rouge-orangé ; 3 = rouge; 4 = rouge-violacé)
- *pour* = proportion de coloration rouge additionnelle, caractérisée par 4 modalités (1 = absente (0-25 %) ; 2 = faible (25-50 %) ; 3 = moyenne (50-75 %) ; 4 = forte (75-100 %))
- *type* = type de coloration du fruit, caractérisé par 3 modalités (1 = lavé ; 2 = lavé-strié; 3 = strié)
- *form* = forme du fruit, caractérisée par 3 modalités (1 = allongé ; 2 = intermédiaire ; 3 = aplati)
- *ferm* = fermeté du fruit, caractérisée par 3 modalités (1 = peu ferme ; 2 = moyen ; 3 = très ferme)
- *gout* = rapport sucre / acidité, caractérisé par 3 modalités (1 = doux ; 2 = équilibré ; 3 = acidulé)

Les variables concernent différents types de caractères : forme de l'arbre, date de récolte,description et qualité du fruit. Certaines ont des modalités ordonnées (ex : *pour*) et ne sont pas des variables qualitatives au sens strict.

La plupart des variétés sont génétiquement différentes, certaines sont issues de mutations et représentent la part la plus importante des variétés commercialisées aujourd'hui car elles apportent un petit *plus* par rapport à la variété standard dont elles dérivent, notamment par leur coloration.

Le but de l'analyse est d'étudier les liaisons entre ces différentes variables et la façon dont cet échantillon de variétés se structure en fonction de ces critères.

Extrait du tableau individus × variables

	arb	rec	cal	coul	pour	type	form	ferm	gout
Granny-spur	1	3	2	1	1	1	2	3	3
Granny-stand.	4	3	2	1	1	1	2	3	3
Boskoop-mut.	3	3	3	4	4	1	1	2	2
Boskoop-stand.	2	3	3	4	3	1	1	2	2
Gala-red	2	2	1	3	3	3	2	2	1
Gala	3	2	1	3	2	3	2	2	1
Elstar	3	1	1	2	2	3	2	2	3
Elista	3	1	1	3	3	1	2	2	3
Jonagold	3	2	3	3	2	2	2	2	3
Jonagored	3	2	3	4	3	3	2	2	3
New-Jonagold	3	2	3	2	3	1	2	2	3
Gloster	2	3	2	3	4	1	1	3	2
Gloster-Spur	1	3	2	3	4	1	1	3	2
Averdal	3	2	2	4	4	2	1	2	1
StarKrimson	1	2	2	4	4	3	1	2	1
Erovan	2	2	2	4	4	1	1	2	1
Goldenspur	1	2	1	1	1	1	1	2	2
Quemoni	3	2	1	1	2	1	1	2	2
Golden-972	3	2	1	1	1	1	1	2	2
Sinta	2	2	1	3	2	1	3	2	2
Fiesta	3	1	1	2	3	3	2	3	2
Arlet	3	1	1	3	3	2	2	2	2
Hilrome	4	3	3	3	4	1	2	3	2
Rome-Beauty	4	3	3	3	3	3	2	3	2
Melrose-mut.	3	3	2	2	4	1	3	3	2
Melrose	3	3	3	2	3	3	3	3	2

Remarquons que le codage est séquentiel à partir de 1 (c'est en principe la convention utilisée par la plupart des logiciels).

Les éléments du tableau sont les codes des modalités et non des effectifs.

Tableau d'indicatrices Z

	arb				rec			cal			
	arb1	arb2	arb3	arb4	rec1	rec2	rec3	cal1	cal2	cal3	...
Granny-spur	1	0	0	0	0	0	1	0	1	0	...
Granny-stand.	0	0	0	1	0	0	1	0	1	0	...
Boskoop-mut.	0	0	1	0	0	0	1	0	0	1	...
Boskoop-stand.	0	1	0	0	0	0	1	0	0	1	...
Gala-red	0	1	0	0	0	1	0	1	0	0	...
Gala	0	0	1	0	0	1	0	1	0	0	...
Elstar	0	0	1	0	1	0	0	1	0	0	...
Elista	0	0	1	0	1	0	0	1	0	0	...
Jonagold	0	0	1	0	0	1	0	0	0	1	...
Jonagored	0	0	1	0	0	1	0	0	0	1	...
New-Jonagold	0	0	1	0	0	1	0	0	0	1	...
...	...	...	...	...	...	...	...	...	...	...	...

Notations

Questions	1					k			K				Total
Réponses	1	.	j	.	m_1	.	j	.	1	.	j	.	m_K	$m = \sum_{k=1}^{K} m_k$
1	1	.	0	.	0				0	.	0	.	1	K
.	.	.	.	.	.				.	.	.	.	.	.
Individus i	.	.	.	.	.				.	.	.	.	.	K
.	.	.	.	.	.				.	.	.	.	.	.
n	0	.	1	.	0				1	.	0	.	0	K
Total	n_{11}	.	n_{j1}	.	$n_{m_1 1}$	.	n_{jk}	.	n_{1K}	.	n_{jK}	.	$n_{m_K K}$	$n \times K$

$$\sum_{j=1}^{m_k} n_{jk} = n$$

- Tableau d'indicatrices Z = tableau disjonctif complet.
- Les individus sont en lignes ; toutes les modalités de toutes les variables sont en colonnes.
- En ligne, un individu prend la valeur 1 pour la modalité à laquelle il appartient et la valeur 0 pour toutes les autres modalités de la variable considérée.
- Si l'on additionne les colonnes correspondant à une même variable on obtient une colonne de 1.
- La somme des valeurs d'une même ligne est une constante égale à K (nombre de variables).
- Ce tableau est à considérer principalement quand on s'intéresse aux individus.
- Terminologie utilisée :

  question = une variable qualitative
  réponse = une modalité de la variable

Tableau de Burt B

		arb				rec			cal			
		arb1	arb2	arb3	arb4	rec1	rec2	rec3	cal1	cal2	cal3	...
arb	arb1	4	0	0	0	0	2	2	1	3	0	...
	arb2	0	10	0	0	0	8	2	5	2	3	...
	arb3	0	0	34	0	12	14	8	22	7	5	...
	arb4	0	0	0	9	1	3	5	2	3	4	...
rec	rec1	0	0	12	1	13	0	0	9	4	0	...
	rec2	2	8	14	3	0	27	0	15	6	6	...
	rec3	2	2	8	5	0	0	17	6	5	6	...
cal	cal1	1	5	22	2	9	15	6	30	0	0	...
	cal2	3	2	7	3	4	6	5	0	15	0	...
	cal3	0	3	5	4	0	6	6	0	0	12	...
coul	coul1	2	1	8	2	1	8	4	8	4	1	...
	coul2	0	1	8	1	5	2	3	5	3	2	...
	coul3	1	6	14	3	6	11	7	16	4	4	...
	coul4	1	2	4	3	1	6	3	1	4	5	...
pour	pour1	2	0	3	2	0	3	4	4	3	0	...
	pour2	0	3	10	0	3	9	1	9	2	2	...
	pour3	0	4	14	2	7	8	5	12	2	6	...
	pour4	2	3	7	5	3	7	7	5	8	4	...
type	type1	3	7	16	6	5	14	13	16	8	8	...
	type2	0	1	8	1	4	5	1	6	3	1	...
	type3	1	2	10	2	4	8	3	8	4	3	...
form	form1	3	4	11	1	3	11	5	7	8	4	...
	form2	1	4	18	7	7	13	10	18	6	6	...
	form3	0	2	5	1	3	3	2	5	1	2	...
ferm	ferm1	0	1	2	0	1	2	0	1	1	1	...
	ferm2	2	8	25	3	11	24	3	23	8	7	...
	ferm3	2	1	7	6	1	1	14	6	6	4	...
gout	gout1	1	2	6	1	1	6	3	6	3	1	...
	gout2	2	8	19	3	9	15	8	17	8	7	...
	gout3	1	0	9	5	3	6	6	7	4	4	...

– Tableau des variables croisées deux à deux.

– Tableau symétrique en lignes et colonnes, constitué de blocs.

– Dans la diagonale on retrouve les effectifs des modalités.

– Inconvénient : on perd la notion d'individu.

– On a : $B = {}^tZZ$

– Les analyses de B et de Z donnent des résultats équivalents ; c'est-à-dire qu'on peut déduire d'une analyse les résultats de l'autre. Donc, en pratique, lorsque le tableau Z est très important (quelques dizaines ou centaines de milliers d'individus) par rapport aux capacités de calcul dont on dispose, il est intéressant d'effectuer l'AFC sur le tableau B (de dimension plus modeste).

– En analyse des correspondances, il est souhaitable que le remplissage de chaque table de contingence croisant deux variables vérifie les règles appliquées au calcul du χ^2 (un minimum d'effectif).

On peut montrer tous les χ^2 dans une matrice triangulaire (attention, ils n'ont pas tous le même nombre de degrés de liberté).

ACM : AFC du tableau des indicatrices Z Distance entre individus

Un individu est caractérisé par l'ensemble des valeurs des réponses qu'il a fournies aux K différentes questions.

Exemples :

– variétés proches : Baujade et Granny-standard

	arb	rec	cal	coul	pour	type	form	ferm	gout
Codage	1234	123	123	1234	1234	123	123	123	123
Baujade	0001	001	100	1000	1000	100	010	001	001
Granny-stand.	0001	001	010	1000	1000	100	010	001	001

– variétés éloignées : Granny-spur et Vista-bella

	arb	rec	cal	coul	pour	type	form	ferm	gout
Codage	1234	123	123	1234	1234	123	123	123	123
Granny-spur	1000	001	010	1000	1000	100	010	001	001
Vista-Bella	0010	100	100	0010	0001	010	001	100	010

Matrice $D_J = diag(\frac{n_{jk}}{n \times K})$

Carré de la distance entre 2 individus i et i' :

$$d^2(i,i') = \sum_{k \in \Delta_{ii'}} \frac{n}{K} \left(\frac{1}{n_{jk}} + \frac{1}{n_{j'k}} \right)$$

avec $\Delta_{ii'} \subset \{1, \ldots, K\}$ ensemble des questions auxquelles i et i' ont donné des réponses différentes ($jk \neq j'k$)

Analogie avec l'AFC : D_J est ici $diag(\frac{n_{jk}}{n \times K})$ et D_I est ici $\frac{1}{n} I_n$.

Deux individus sont d'autant plus éloignés qu'ils ont donné à de nombreuses questions des réponses *différentes et rares* (une modalité j intervient d'autant plus dans le calcul de la distance entre deux individus que son effectif n_{jk} est plus faible).

Exemples :

Les sommes marginales des colonnes étant :

```
arb1 arb2 arb3 arb4 rec1 rec2 rec3 cal1 cal2 cal3 coul1 coul2 coul3 coul4
   4   10   34    9   13   27   17   30   15   12    13    10    24    10

pour1 pour2 pour3 pour4 type1 type2 type3 form1 form2 form3 ferm1 ferm2
    7    13    20    17    32    10    15    19    30     8     3    38

ferm3 gout1 gout2 gout3
   16   10     32    15
```

Baujade et *Granny-stand.* diffèrent uniquement par leur réponse à la variable *cal*

$d^2 = \frac{57}{9} \left(\frac{1}{30} + \frac{1}{15} \right) = 0.633$ soit $d = 0.796$

Granny-spur et *Vista-Bella* ont donné des réponses différentes pour toutes les variables

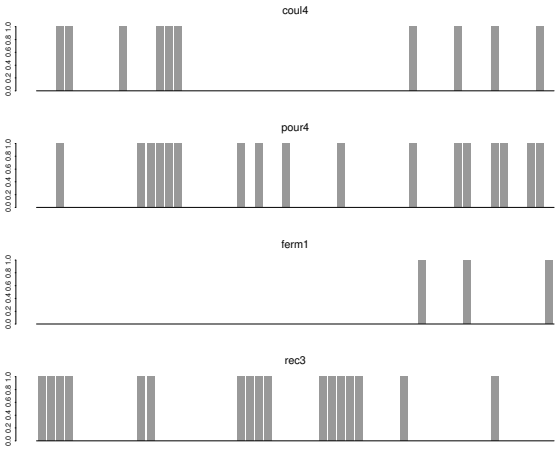
$$\begin{aligned} d^2 \;=\; \frac{57}{9} \;\; & \Big[\left(\tfrac{1}{4} + \tfrac{1}{34} \right) + \left(\tfrac{1}{13} + \tfrac{1}{17} \right) + \left(\tfrac{1}{30} + \tfrac{1}{15} \right) + \\ & \left(\tfrac{1}{13} + \tfrac{1}{24} \right) + \left(\tfrac{1}{7} + \tfrac{1}{17} \right) + \left(\tfrac{1}{32} + \tfrac{1}{10} \right) + \\ & \left(\tfrac{1}{30} + \tfrac{1}{8} \right) + \left(\tfrac{1}{3} + \tfrac{1}{16} \right) + \left(\tfrac{1}{32} + \tfrac{1}{15} \right) \Big] = 10.249 \text{ soit } d = 3.201 \end{aligned}$$

Distance entre réponses

La réponse à une question est caractérisée par l'ensemble des valeurs fournies par les individus.

Exemples :

profils proches coul4 et pour4 (couleur rouge-violacé et fort pourcentage de coloration additionnelle)
profils éloignés ferm1 et rec3 (fruit peu ferme et récolte tardive)



Matrice $D_I = \frac{1}{n} I_n$ car le poids des lignes est constant et égal à K.

Carré de la distance entre la réponse j à la question k et la réponse j' à la question k' :

$$d^2(j,j') = \frac{n}{n_{jk} n_{j'k'}} \delta_{kk'}$$

$\delta_{kk'}$ = nbre d'individus ayant répondu j à la question k mais pas j' à la question k', ou ayant répondu j' à la question k' mais pas j à la question k.

Deux réponses sont d'autant plus éloignées qu'elles sont *rares* (n_{jk} et $n_{j'k'}$ faibles) et qu'elles sont données par des individus différents ($\delta_{kk'}$ grand).

Exemples :

coul4 : 10 individus sur 57 ont donné cette réponse
8 ont aussi répondu pour4
2 ont répondu différemment à la question pourcentage

pour4 : 17 individus sur 57 ont donné cette réponse
8 ont aussi répondu coul4
9 ont répondu autrement

distance $d^2 = \frac{57}{10 \times 17} \times (2+9) = 3.688235$
$d = 1.920478$

ferm1 : 3 individus sur 57 ont donné cette réponse
aucun n'a répondu également rec3

rec3 : 17 individus sur 57 ont donné cette réponse
aucun n'a répondu ferm1

distance $d^2 = \frac{57}{3 \times 17} \times (3+17) = 22.35294$
$d = 4.72789$

AFC du tableau Z

L'AFC de Z s'obtient par la DVS de :

$$X = nF_z D_J^{-1} - E = \frac{1}{K} Z D_J^{-1} - E = \sum_s \lambda_s u_s {}^t v_s$$

pour les métriques D_I et D_J ; les vecteurs u_s sont D_I orthonormés, le vecteurs v_s sont D_J orthonormés.

n est le nombre d'individus

$F_z = \frac{1}{nK} Z$ est le tableau de fréquences associé à Z

D_J est la matrice diagonale des poids des colonnes de Z

K est le nombre de questions

$D_I = \frac{1}{n} I_n$ est la matrice diagonale des poids des lignes de Z

E est la matrice de dimensions $m \times n$ dont tous les éléments sont égaux à 1

La matrice d'éléments

$\frac{f_{ij}}{f_{i.} f_{.j} - 1}$ de l'AFC

est ici $D_I^{-1} F_z D_J^{-1} - E = nF_z D_J^{-1} - E$.

Résultats de l'analyse

– Inertie globale de l'AFC du tableau Z : $\dfrac{m-K}{K}$

– Contribution d'une modalité à l'inertie :

$\dfrac{1}{K}\left(1-\dfrac{n_{jk}}{n}\right)$ = contribution de la réponse j à la question k

K = nombre de questions (variables)
m = nombre total de réponses
n = nombre d'individus
n_{jk} = nombre d'individus ayant répondu j à la question k

– Contribution d'une question à l'inertie :

$\dfrac{1}{K}(m_k-1)$ = contribution de la question k

– La contribution d'une modalité est d'autant plus importante que la réponse est rare (n_{jk} petit).

Conséquence : La présence de modalités rares risque de gêner l'analyse. Si l'on doit coder une variable quantitative, on choisira un découpage en modalités d'effectifs comparables.
Les modalités à faible effectif n'interviennent légitimement que lorsque l'occurrence d'une telle modalité est un événement important.

– La contribution d'une question est proportionnelle au nombre de réponses possibles (m_k) diminué d'une unité.

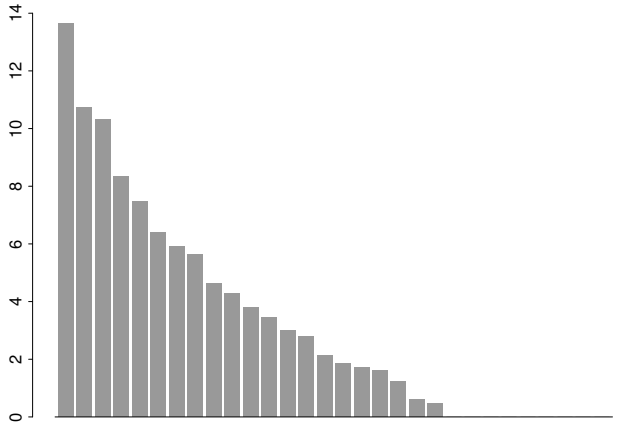
Conséquence : La présence de questions avec des nombres de modalités très supérieurs revient à donner à ces questions une grande importance. On choisit si possible des ensembles de réponses d'effectifs (m_k) assez semblables.

– Dans l'espace des facteurs les distances sont celles du χ^2.

Résultats de l'exemple

```
Inertie totale:  2.33

Valeurs propres

 f1   f2   f3   f4   f5   f6   f7   f8   f9   f10  f11  f12  f13  f14   f15
0.32 0.25 0.24 0.19 0.17 0.15 0.14 0.13 0.11 0.10 0.09 0.08 0.07 0.065 0.05

 f16  f17  f18  f19  f20  f21   f22     f23     f24     f25     f26     f27
0.04 0.04 0.04 0.03 0.01 0.01 1.5e-16 8.9e-17 7.7e-17 5.6e-17 4.4e-17 3.4e-17

  f28     f29      f30
2.9e-17 1.9e-18 -2.3e-17
```

Eboulis des valeurs propres




```
Pourcentage d'inertie expliquee par les facteurs:

                          f1 f2 f3 f4 f5 tot
          inertie expliquee 14 11 10  8  7  51
          inertie cumulee   14 24 35 43 51  51
```


– On choisit des questions censées être liées à un même phénomène d'intérêt et dont l'ensemble est assez homogène et exhaustif. Les premiers facteurs prenant en compte la plus grande part de la redondance entre les questions seront alors très liés au phénomène d'intérêt.

– **Remarques :**

. $2.33 = \dfrac{30-9}{9}$ avec $30 = m$ et $9 = K$.

. Le nombre maximum de valeurs propres non nulles est de $30 - 1 = 29$.

. Seuls figurent les résultats des 5 premiers facteurs.

– En ACM les pourcentages d'inertie expliquée par les premiers facteurs sont faibles, cela ne signifie pas que l'analyse soit sans intérêt.

Ici, l'éboulis présente un premier coude après 3 facteurs (soit 35 % de la variabilité) et un second après 5 facteurs (soit 51 %).

Contribution des modalités et des variables aux facteurs et à l'inertie totale:

	. f1	f2	f3	f4	f5	f>5.	TOT	ORI
arb1	. 51	23	15	12	106	49.	44	116
arb2	. 8	0	82	18	82	45.	39	41
arb3	. 30	1	21	5	17	23.	19	6
arb4	. 80	29	11	4	29	45.	40	47
arb	.169	53	129	39	234	162.	142	210
rec1	. 56	5	36	2	123	31.	37	30
rec2	. 9	25	65	21	41	19.	25	10
rec3	.108	18	24	22	3	25.	33	21
rec	.173	48	125	45	167	75.	95	61
cal1	. 35	15	19	3	4	27.	23	8
cal2	. 31	0	12	3	183	32.	35	25
cal3	. 10	40	8	23	147	37.	38	33
cal	. 76	55	39	29	334	96.	96	66
coul1	. 21	261	7	0	1	10.	37	30
coul2	. 9	39	91	2	18	47.	39	41
coul3	. 24	11	1	11	4	44.	28	12
coul4	. 29	49	121	14	0	33.	39	41
coul	. 83	360	220	27	23	134.	143	124

– Un facteur s'interprète en fonction des questions et des réponses qui y ont le plus contribué (la contribution d'une question est la somme des contributions de ses réponses). Il s'agit toujours, comme en AFC, de facteurs marquant des oppositions.

– ORI est un indice d'originalité proportionnel au carré de la distance d'un individu, d'une modalité ou d'une variable à l'origine.

– La contribution d'une variable à 3 modalités est de 95 millièmes (aux arrondis près).

– La contribution d'une variable à 4 modalités est de 142 millièmes (aux arrondis près).

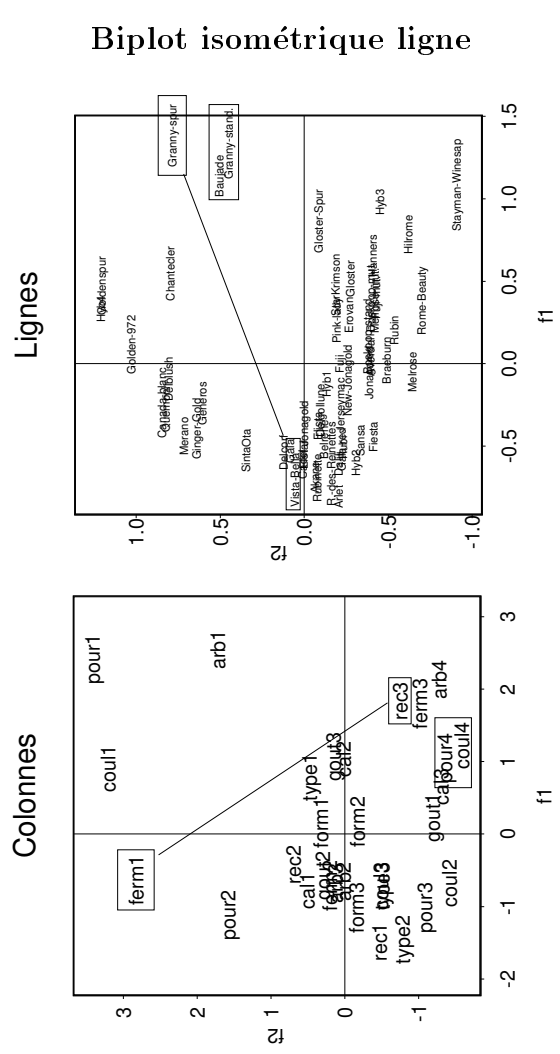
– Les questions qui influent le plus sur le premier facteur sont arb, rec, $pour$ et $ferm$. Le facteur $f1$ s'interprète en termes d'opposition entre certaines réponses à ces questions.

– La question form a très peu d'influence sur le facteur $f1$.

– On vérifiera la qualité de la représentation des réponses sur le tableau des cos^2 lignes - sous espaces principaux (non produit ici).

– Puisque les points ont des poids différents, le carré de la distance à l'origine (ORI) n'est pas proportionnel à sa contribution au plan. En jouant sur la grosseur des points on pourrait représenter ces contributions dans le graphe des projections sur un plan factoriel.

```
pour1  . 80  159   26    0    5   22.   42   63
pour2  . 32   64    1   15   46   42.   37   30
pour3  . 40   45   28    1    1   35.   31   16
pour4  . 37   60   71    5   30   24.   33   21
.........................................
pour   .189  328  126   21   82  123.  143  130
.........................................
type1  . 39   14    6   16    7   23.   21    7
type2  . 41   12    1   16   57   54.   39   41
type3  . 15    7    8   83    6   49.   35   25
.........................................
type   . 95   33   15  115   70  126.   95   73
.........................................
form1  .  1    4   96   13   43   34.   32   18
form2  .  2    2   68   28   32   21.   23    8
form3  . 16    0    1  247    1   36.   41   54
.........................................
form   . 19    6  165  288   76   91.   96   80
.........................................
ferm1  .  2   46    3  183    0   49.   45  158
ferm2  . 37    3   17   40    0   11.   16    4
ferm3  .101   32   51   15    0   21.   34   23
.........................................
ferm   .140   81   71  238    0   81.   95  185
.........................................
gout1  .  1   28   34   89    0   51.   39   41
gout2  . 20    6    6   81    5   20.   21    7
gout3  . 33    1   70   30    9   40.   35   25
.........................................
gout   . 54   35  110  200   14  111.   95   73
.........................................
       .1000 1000 1000 1000 1000 1000 1000 1000
```



– On observe que :

. Le facteur 1 sépare en gros les variétés précoces ($rec1$) des variétés tardives ($rec3$) et très fermes ($ferm3$), ces deux dernières modalités apportant une forte contribution à l'axe.

. Le facteur 2 sépare les variétés jaunes et vertes ($coul1$) sans teinte additionnelle ($pour1$) de toutes les autres.

. Les modalités $coul4$ et $pour4$, qui ont des profils semblables, ont des projections proches l'une de l'autre sur le plan principal. Les modalités $ferm1$ et $rec3$, qui ont des profils très différents, sont éloignées.

. Les variétés $Baujade$ et $Granny - stand.$ aux profils semblables, sont proches sur le plan, tandis que les variétés $Granny - spur$ et $Vista - Bella$, aux profils très différents sont éloignées.

. Dans l'espace des lignes et dans celui des colonnes la zone centra le est vide ; cela est dû ici au choix volontairement contrasté de l'échantillon des variétés et des caractéristiques retenues.

– Le barycentre des marqueurs des réponses d'une même question, pondérées par les effectifs des réponses, est le barycentre global (donc l'origine dans les représentations).

Conséquences:

. Lorsqu'une variable a deux modalités, les marqueurs sont aux extrémités d'un segment de droite contenant l'origine, les longueurs des segments étant inversement proportionnels aux effectifs des réponses.

. Dans la représentation ayant les réponses comme barycentre, une réponse est au barycentre des individus qui l'ont donnée.

. Dans la représentation ayant les individus comme barycentre, un individu est au barycentre des réponses qu'il a données.

Variables illustratives

On enrichit l'interprétation d'une ACM en rajoutant des variables supplémentaires ou illustratives.

Une réponse étant au barycentre des individus qui l'ont donnée, il est facile de représenter par ce moyen des réponses supplémentaires. Si ces réponses sont suffisamment éloignées du barycentre, alors cette question est liée aux variables actives, ou plus précisément, au résumé qui en est fait par les premiers facteurs.

En ACM, les variables illustratives sont qualitatives (ou quantitatives discrétisées).

Analyse du tableau de Burt B

- Profils : le tableau étant symétrique, les profils lignes sont identiques aux profils colonnes. Une seule ACP sera nécessaire pour l'analyse.
- Distances : les distances sont déformées régulièrement et faiblement si λ_1 est proche de λ_2.
- Métrique : D_B matrice diagonale du poids des lignes et des colonnes
- Les analyses (AFC) de B et de Z sont telles que les valeurs propres de l'AFC de B sont celles de l'AFC de Z au carré.
- Les pourcentages d'inertie expliquée sont plus forts que dans l'ACM de Z (nuage plus allongé sur le $1er$ axe.

On a :

$B = {}^tZZ$ et $Z = UD\,{}^tV$

$B = VD\,{}^tUUD\,{}^tV = VD^2\,{}^tV$

les DVS sont les mêmes, les valeurs singulières de B sont au carré.

Ou bien :

AFC de $Z \rightarrow$ DVS de $\frac{1}{K}ZD_J^{-1} = \sum_s \lambda_s \mathbf{u}_s\,{}^t\mathbf{v}_s$

AFC de $B \rightarrow$ DVS de $\frac{1}{nK^2}D_J^{-1}BD_J^{-1} = \left[\frac{1}{K}D_J^{-1}\,{}^tZ\right]D_I\left[ZD_J^{-1}\frac{1}{K}\right]$

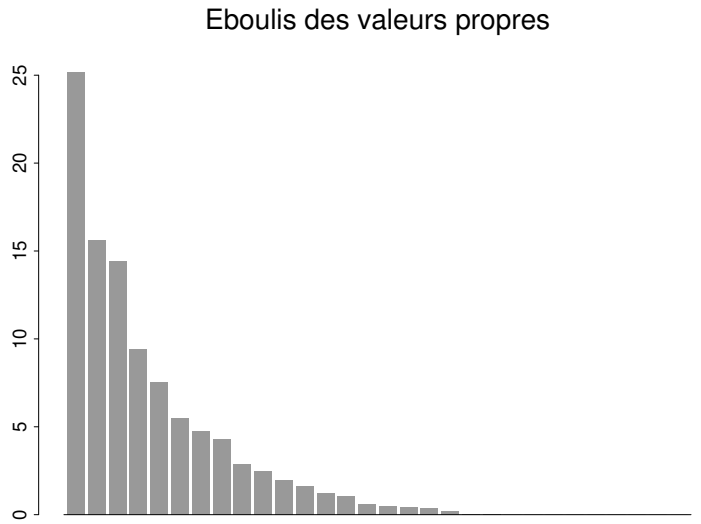
$= [\sum_s \lambda_s\,{}^t\mathbf{v}_s\mathbf{u}_s]D_I[\sum_l \lambda_l\,{}^t\mathbf{u}_l\mathbf{v}_l] = \sum_s \lambda_s^2\,{}^t\mathbf{v}_s\mathbf{v}_s.$

avec $D_I = \frac{1}{n}I_n$

la coordonnée passe de $\lambda_s u_i^s(Z)$ à $(\lambda_s)^2 u_i^s(B)$.

Résultats de l'ACM sur B

```
Inertie totale:  0.4

Valeurs propres :

       f1         f2         f3         f4         f5         f6         f7
0.1013181 0.06278225 0.05808789 0.0379203 0.03026723 0.02215653 0.01903697
       f8         f9        f10         f11        f12         f13
0.0173106 0.01164472 0.01000652 0.007828113 0.00646734 0.004877728
        f14         f15         f16         f17       f18          f19
0.004236716 0.002454416 0.001881952 0.001618166 0.0014098 0.0008408889
         f20          f21          f22          f23          f24          f25
0.0001971365 0.0001175645 6.301526e-18 4.152271e-18 3.240682e-18 1.861229e-18
         f26           f27           f28         f29           f30
4.874816e-19 -1.654147e-18 -3.902081e-18 -8.2177e-18 -9.546523e-18
```



L'inertie de l'AFC de Z est différente de celle de l'AFC de B (à comparer avec le transparent 38).

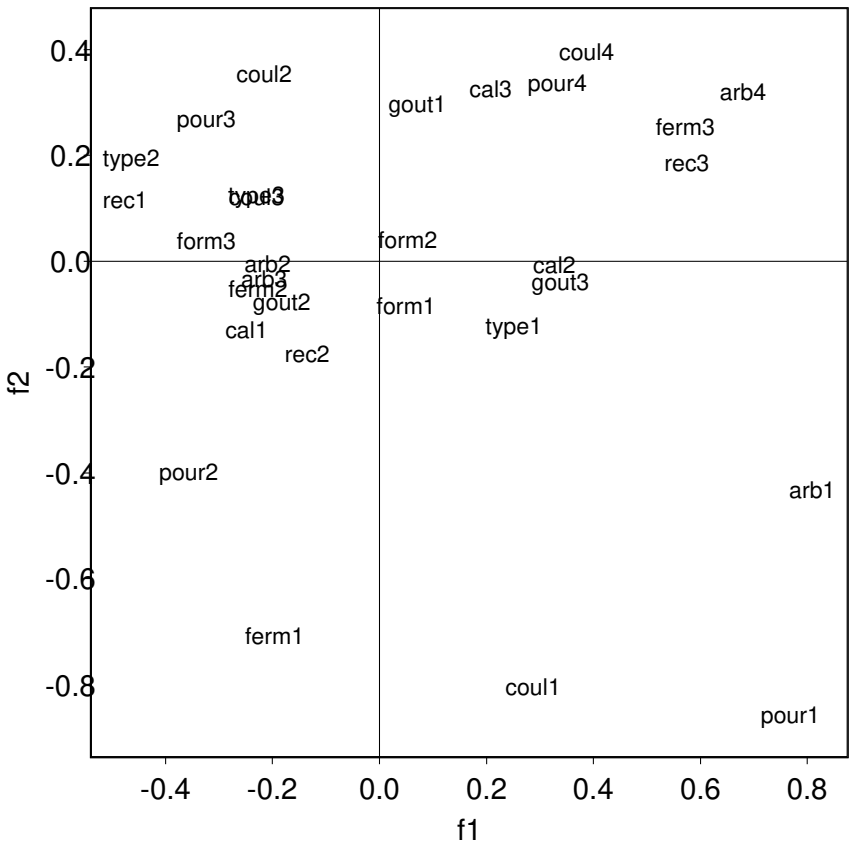
Pourcentage d'inertie expliquée par les facteurs :

	f1	f2	f3	f4	f5	tot
inertie expliquee	25	16	14	9	8	72
inertie cumulee	25	41	55	65	72	72

Contribution des modalités aux facteurs

	f1	f2	f3	f4	f5	f>5	TOT	ORI
arb1	51	23	15	12	106	45	40	108
arb2	8	0	82	18	82	43	34	36
arb3	30	1	21	5	17	21	18	6
arb4	80	29	11	4	29	43	41	49
rec1	56	5	36	2	123	27	37	31
rec2	9	25	65	21	41	16	25	10
rec3	108	18	24	22	3	25	43	27
cal1	35	15	19	3	4	29	23	8
cal2	31	0	12	3	183	28	32	23
cal3	10	40	8	23	147	40	34	31
coul1	21	261	7	0	1	6	48	40
coul2	9	39	91	2	18	48	36	39
coul3	24	11	1	11	4	54	24	11
coul4	29	49	121	14	0	33	43	46
pour1	80	159	26	0	5	16	53	82
pour2	32	64	1	15	46	43	35	29
pour3	40	45	28	1	1	39	32	17
pour4	37	60	71	5	30	22	38	24
type1	39	14	6	16	7	22	21	7
type2	41	12	1	16	57	58	34	37
type3	15	7	8	83	6	54	30	21
form1	1	4	96	13	43	36	29	17
form2	2	2	68	28	32	20	21	8
form3	16	0	1	247	1	31	36	49
ferm1	2	46	3	183	0	51	40	142
ferm2	37	3	17	40	0	11	19	5
ferm3	101	32	51	15	0	22	45	31
gout1	1	28	34	89	0	59	34	37
gout2	20	6	6	81	5	19	20	7
gout3	33	1	70	30	9	41	34	24
	1000	1000	1000	1000	1000	1000	1000	1000

Représentation Isométrique ligne (ou colonne)



– La méthode générale d'interprétation est semblable à celle d'une AFC. Les quantités approximées par le biplot entre une modalité k d'une question j et k' d'une question j' est le rapport de la fréquence de l'événement conjoint sur le produit des fréquences marginales :

$$\frac{\Pr[\text{répondre } k \text{ à } j \text{ et } k' \text{ à } j']}{\Pr[\text{répondre } k \text{ à } j] * \Pr[\text{répondre } k' \text{ à } j']} - 1$$

Remarque : dans le biplot symétrique les 2 nuages sont identiques et une seule figure permet l'interprétation du biplot.

– On peut aussi obtenir une représentation des individus dans l'ACM de B par le biais d'individus supplémentaires (cf. ACP etAFC).