Exemples d'analyse en composantes principales

1.1.1 Mini-exemple

Ci-dessous, un tableau de notes attribuées à 9 sujets dans 5 matières.

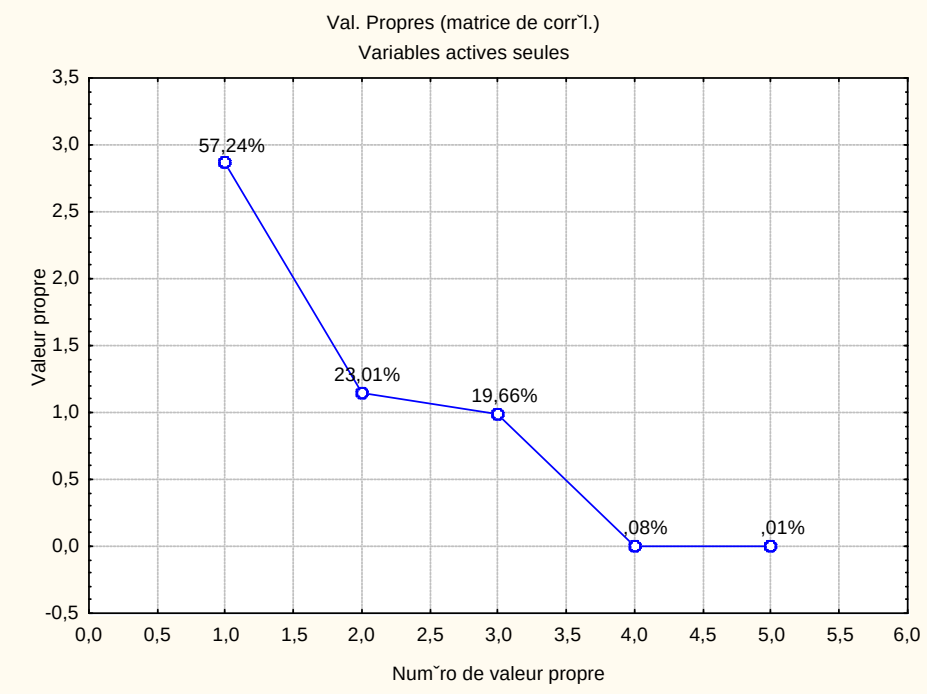
Sujet	Math	Sciences	Français	Latin	Musique
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

L'ACP étudie les lignes et les colonnes de la matrice centrée-réduite :

Sujet	Math	Sciences	Français	Latin	Musique
Jean	-1,0865	-1,2817	-1,5037	-1,6252	-1,0190
Aline	-0,4939	-0,6130	-0,6399	-0,7223	-0,6794
Annie	-1,0865	-0,9474	0,2239	-0,1806	0,0000
Monique	1,4322	1,5604	1,5197	1,8058	-1,0190
Didier	1,2840	1,3932	0,5119	0,7223	-0,3397
André	0,3951	0,0557	-1,3597	-1,0835	0,6794
Pierre	-1,2347	-0,9474	1,0878	0,5417	-0,3397
Brigitte	0,9877	0,8916	-0,4959	-0,1806	0,3397
Evelyne	-0,1975	-0,1115	0,6559	0,7223	2,3778

1.2 Valeurs propres et inerties

	Val. propr	Variance (%)	Variance cumul (%)
1	2,8618	57,24	57,24
2	1,1507	23,01	80,25
3	0,9831	19,66	99,91
4	0,0039	0,08	99,99
5	0,0004	0,01	100,00



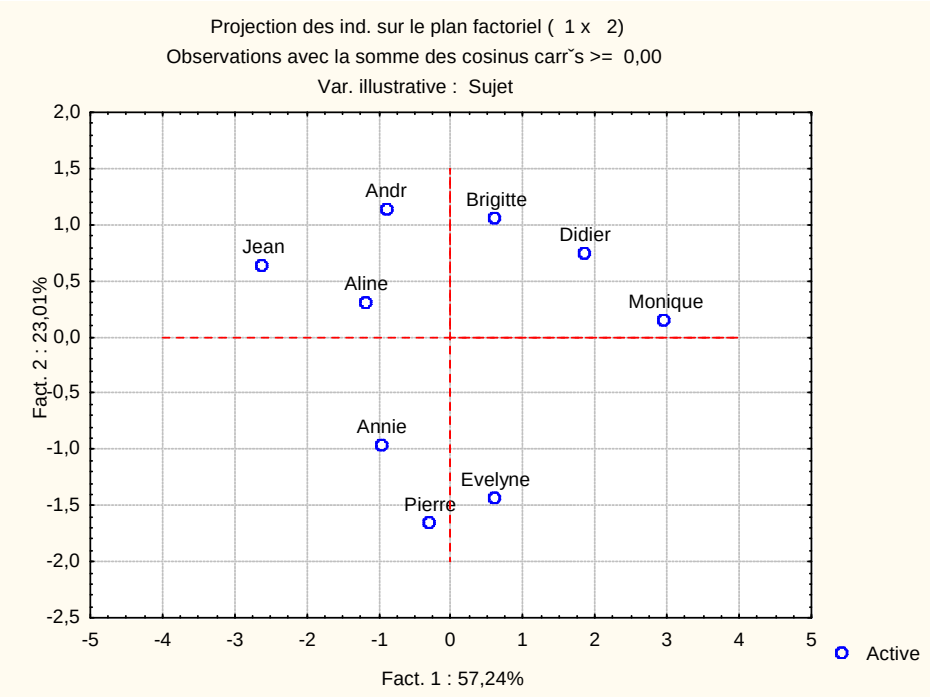
La variation totale (100%) est répartie selon 5 valeurs propres. D'où l'idée de ne garder que les valeurs propres (et directions propres) qui représentent au moins 20% de variation. Dans le cas d'une ACP normée, cela revient à conserver les valeurs propres supérieures à 1.
Variante : on observe une brusque décroissance des valeurs propres entre la 3è et la 4è valeur propre.
Au final, on décide de ne garder que trois valeurs propres.

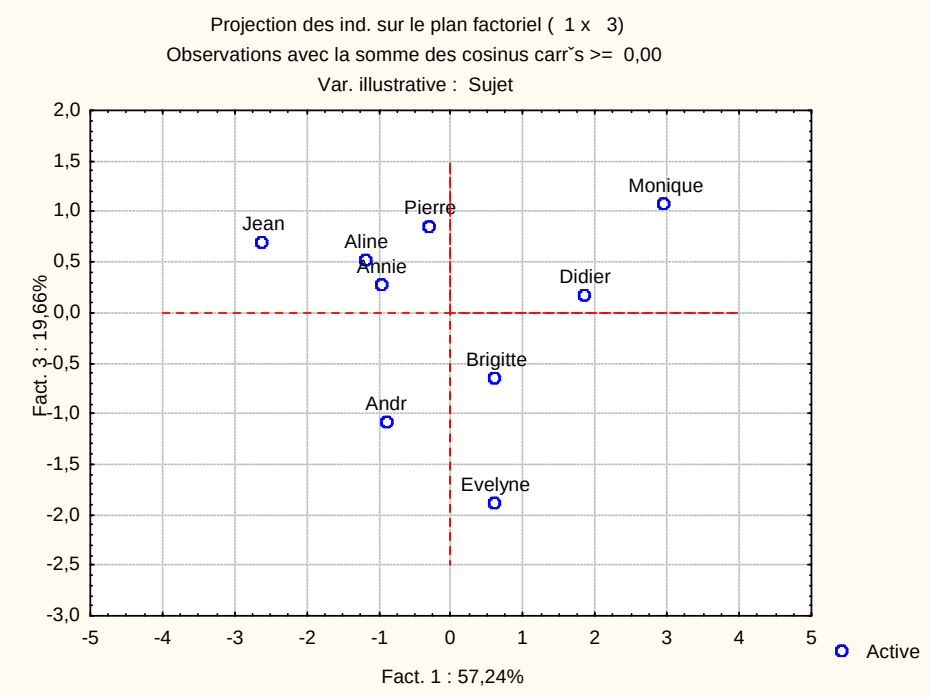
1.3 Résultats relatifs aux individus

1.3.1 Scores des individus

Les scores des individus sont les valeurs des composantes principales sur les individus :

	Fact. 1	Fact. 2	Fact. 3
Jean	-2,7857	0,6765	0,7368
Aline	-1,2625	0,3303	0,5549
Annie	-1,0167	-1,0198	0,2881
Monique	3,1222	0,1659	1,1442
Didier	1,9551	0,7879	0,1892
André	-0,9477	1,2014	-1,1401
Pierre	-0,3250	-1,7548	0,9095
Brigitte	0,6374	1,1298	-0,6919
Evelyne	0,6231	-1,5173	-1,9909





Contributions des individus

La contribution relative d'un individu i à la formation de la composante principale α est l'inertie relative de cet individu sur l 'axe factoriel k. Elle est définie par :

$$CTR_{\alpha}(i) = \frac{(Score\,de\,i\ sur\,l'axe\,\alpha)^2}{n\,\lambda_{\alpha}}$$

Par exemple : $CTR_1(Jean) = \frac{(-2,7857)^2}{9 \times 2,8618} = 0,3013$

Contributions des individus exprimées en pourcentages

Sujet	Fact. 1	Fact. 2	Fact. 3
Jean	30,13	4,42	6,14
Aline	6,19	1,05	3,48
Annie	4,01	10,04	0,94
Monique	37,85	0,27	14,80
Didier	14,84	5,99	0,40
André	3,49	13,94	14,69
Pierre	0,41	29,73	9,35
Brigitte	1,58	12,33	5,41
Evelyne	1,51	22,23	44,79

Qualités de la représentation des individus

La qualité de la représentation d'un individu i par la composante principale α est définie par :

$$QLT_{\alpha}(i) = \frac{(Score\,de\ i\ \ sur\ \ l'axe\ \alpha)^2}{\sum_{l}(Score\,de\ i\,sur\,l'axe\ \ l)^2}$$

Par exemple :

$$QLT_1(Jean) = \frac{(-2,7857)^2}{2,7857^2 + 0,6765^2 + ... + 0,0332^2} = \frac{(-2,7857)^2}{1,0865^2 + 1,2817^2 + 1,5037^2 + 1,6252^2 + 1,0190^2} = 0,8855$$

Cosinus carré :

Sujet	Fact. 1	Fact. 2	Fact. 3
Jean	0,8855	0,0522	0,0619
Aline	0,7920	0,0542	0,1530
Annie	0,4784	0,4813	0,0384
Monique	0,8786	0,0025	0,1180
Didier	0,8515	0,1383	0,0080
André	0,2465	0,3962	0,3568
Pierre	0,0263	0,7671	0,2061
Brigitte	0,1877	0,5898	0,2211
Evelyne	0,0583	0,3458	0,5954

Les qualités de représentation sont additives. Par exemple, la qualité de représentation d'un individu i par le plan 1-2 est donnée par :

$$QLT_{1,2}(i) = \frac{(Score\ de\ i\ sur\ l'axe\ 1)^2 + (Score\ de\ i\ sur\ l'axe\ 2)^2}{\sum_l (Score\ de\ i\ sur\ l'axe\ l)^2}$$

Pour le sujet 1 (Jean), la qualité de représentation par le plan factoriel 1-2 est : 0,8855+0,0522=0,9377.

1.4 Résultats relatifs aux variables

Saturation des variables

Les saturations des variables sont les coordonnées factorielles des variables. Elles sont égales au coefficients de corrélation entre les variables (centrées réduites) de départ et les scores des individus : $\phi_{j\alpha} = \rho(j, \psi_\alpha)$

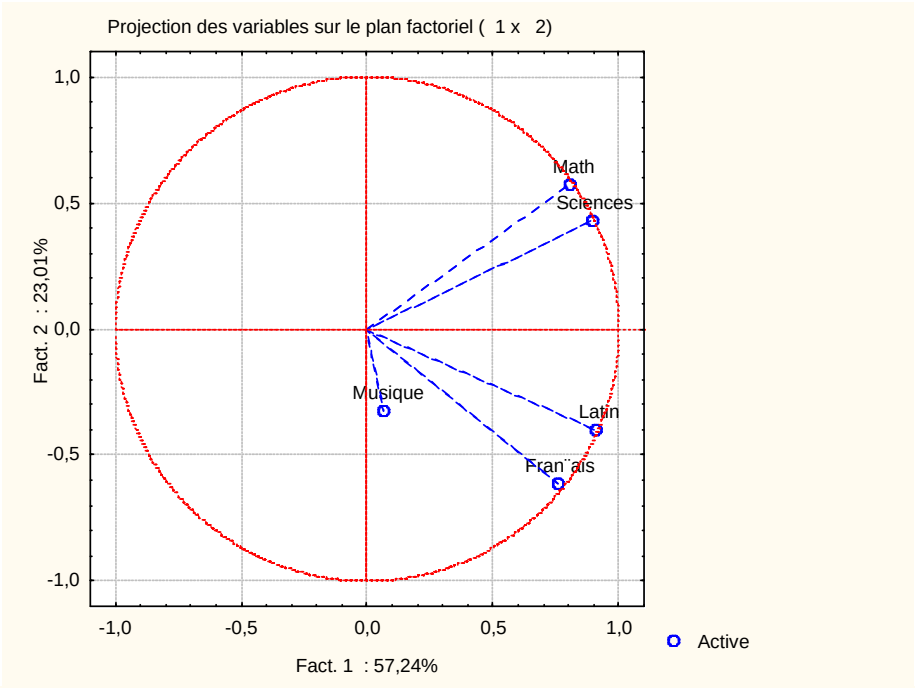
N.B. Les variables de départ sont centrées réduites, les scoress sont centrées, et de variances égales aux valeurs propres correspondantes. On peut donc retrouver les saturations à l'aide d'un calcul tel que :

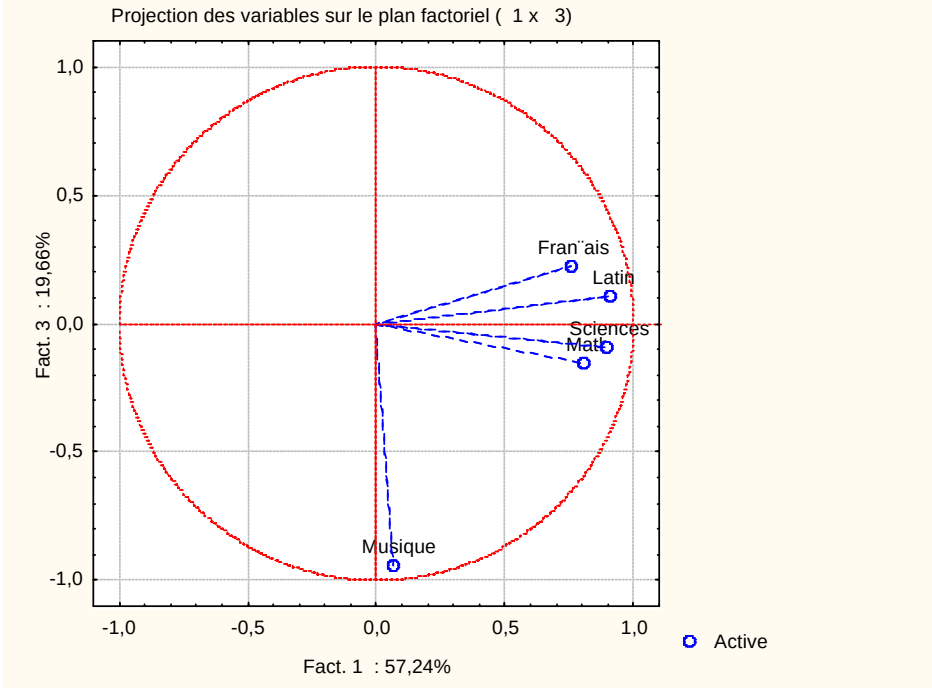
$$\phi_{jean,1} = \frac{(-1,0865)(-2,7857) + (-0,4939)(-1,2625) + (-1,0865)(-1,0168) + (1,4322)(3,1222)}{9\sqrt{2,8618}} +$$

$$\frac{(1,2840)(1,9551) + (0,3951)(-0,9478) + (-1,2347)(-0,3250) + (0,9877)(0,6373) + (-0,1975)(0,6231)}{9\sqrt{2,8618}}$$

Coord. factorielles des variables :

	Fact. 1	Fact. 2	Fact. 3
Math	0,8059	0,5714	-0,1534
Sciences	0,8970	0,4308	-0,0929
Français	0,7581	-0,6110	0,2257
Latin	0,9103	-0,3975	0,1084
Musique	0,0667	-0,3275	-0,9425





Contributions des variables

Les contributions des variables à la formation des composantes principales sont définies de la même façon que celles des individus. Par exemple : $CTR_1(Math) = \frac{0{,}8059^2}{2{,}8618} = 0{,}2269$

Contributions des variables

	Fact. 1	Fact. 2	Fact. 3
Math	0,2269	0,2837	0,0239
Sciences	0,2812	0,1613	0,0088
Français	0,2008	0,3245	0,0518
Latin	0,2895	0,1373	0,0120
Musique	0,0016	0,0932	0,9035

Qualités de la représentation des variables

La qualité de la représentation d'une variable par une composante principale est définie de la même façon que pour les individus :

$$QLT_{\alpha}(j) = \frac{(Saturation\ de\ j\ sur\ l'axe\ \alpha)^2}{\sum_{l}(Saturation\ de\ j\ sur\ \ l'axe\ l)^2} = (Saturation\ de\ j\ sur\ l'axe\ \alpha)^2$$

Mais, comme les variables *j* sont normées, la qualité est simplement le carré de la saturation de la variable par rapport à la composante principale.

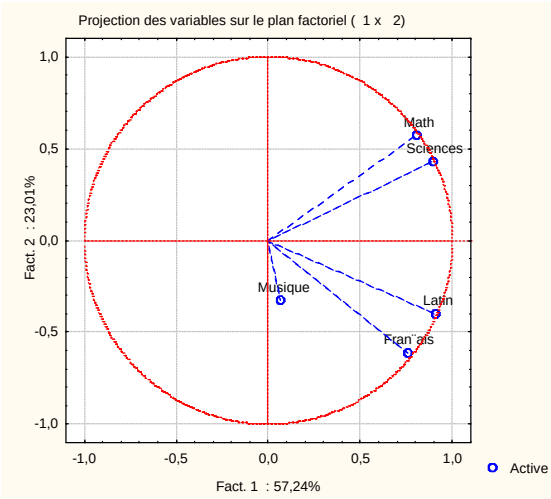
Comme dans le cas des individus, les qualités des représentations d'une variable selon les composantes principales s'additionnent. Le tableau ci-dessous donne les qualités de représentation selon la première composante principale, selon le plan des deux premières composantes et dans l'espace défini par les trois premières composantes.

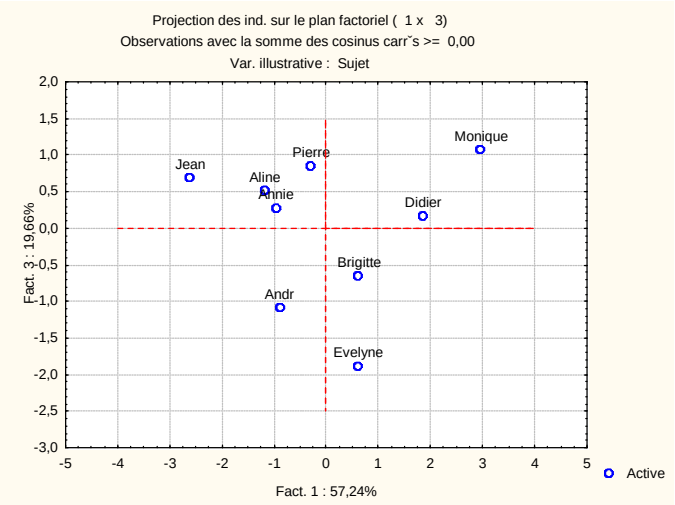
Communautés des variables

	Avec 1 facteur	Avec 2 facteurs	Avec 3 facteurs
Math	0,6495	0,9759	0,9995
Sciences	0,8046	0,9902	0,9988
Français	0,5747	0,9481	0,9990
Latin	0,8286	0,9866	0,9983
Musique	0,0044	0,1117	1,0000

1.5 Interprétation conjointe des plans factoriels des individus et des variables

On utilise la formule : $\phi_{j\alpha} = \rho(j, \psi_{\alpha})$





1.5.1 Enquête budget-temps

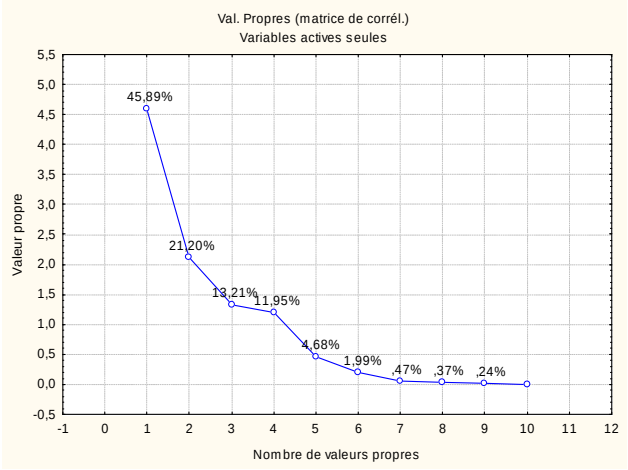
Il s'agit d'une enquête (ONU 1967) sur les budgets-temps (temps passé dans différentes activités au cours de la journée).

Le tableau suivant comprend 10 variables numériques et 4 variables catégorisées.
Les 10 variables numériques sont: le temps passé en: Profession, Transport, Ménage, Enfants, Courses, Toilette, Repas, Sommeil, Télé, Loisirs.
Les 4 variables catégorisées sont: Le sexe (1=Hommes 2=Femmes), l'activité (1=Actifs 2=Non Act. 9=Non précisé), l'état civil (1=Célibataires 2=Mariés 9=Non précisé), le Pays (1=USA 2=Pays de l'Ouest 3=Pays de l'Est 4=Yougoslavie).

Le code suivant est utilisé pour identifier les lignes:
H: Hommes, F: Femmes, A: Actifs, N: Non Actifs(ves), M: Mariés, C: Célibataires, U: USA, W: Pays de l'Ouest sauf USA, E : Est sauf Yougoslavie, Y: Yougoslavie

Les temps sont notés en centièmes d'heures. La première case en haut à gauche du tableau (HAU) indique que les Hommes Actifs des USA passent en moyenne 6 heures et 6 minutes (6 heures + 10/100 d'heure, soit 6 heures et 6mn) en activité PROFessionnelle. Le total d'une ligne (sur ces 10 variables numériques) est 2400 (24 heures).

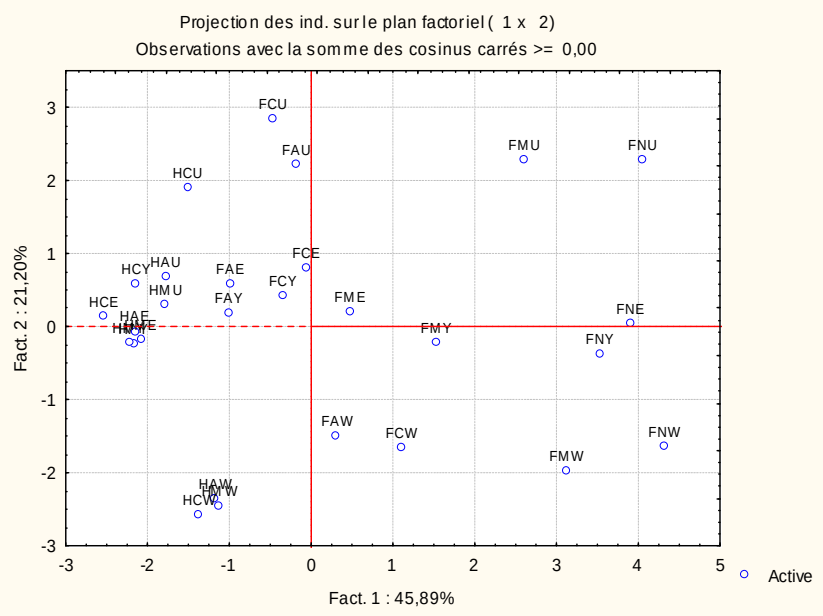
1.6 Valeurs propres et inerties



A priori, on peut choisir de retenir 4 composantes principales. On remarque également que la dernière valeur propre est 0. Cette propriété est due à une particularité de nos données : la somme des variables de départ est une constante, égale à 2400 sur chaque individu.

1.7 Résultats relatifs aux individus

Les résultats numériques nous donnerons successivement les scores des individus, leurs contributions à la formation des composantes principales et leurs qualités de représentation. Ces résultats permettent de construire et d'interpréter les plans factoriels des individus. Très souvent, les individus pris en compte pour une ACP sont en nombre très élevé et sont considérés comme anonymes. Les éléments qui suivent concernent évidemment les cas où ils ne le sont pas.



1.7.1 Qualité de la représentation

Même s'il s'agit du plan 1-2, les proximités entre individus doivent être interprétées avec prudence : deux points proches l'un de l'autre sur le graphique peuvent correspondrent à des individus éloignés l'un de l'autre. Pour interpréter ces proximités, il est nécessaire de tenir compte des qualités de représentation des individus.
Se méfier également **des individus proches de l'origine : mal représentés, ou proches de la moyenne**, ils ont, de toutes façons, peu contribué à la formation des axes étudiés.

1.7.2 Contributions des individus à la formation d'un axe

On relève, pour chaque axe, quels sont les individus qui ont la plus forte contribution à la formation de l'axe. Par exemple, on retient (pour l'analyse) les individus dont la contribution relative est supérieure à $\frac{100\%}{n}$. On note également si cette contribution intervient dans la partie positive ou dans la partie négative de l'axe.

Ainsi, pour l'exemple Budget-temps, on s'intéresse aux contributions relatives supérieures à $\frac{100\%}{28} = 3{,}57\%$. On pourra s'aider du tableau suivant pour interpréter la première variable factorielle :

-	+
HCE (4,98%)	FNW (14,5%)
HMY (3,84%)	FNU (12,8%)
HAY (3,64%)	FNE (11,95%)
HAE (3,59%)	FNY (9,73%)
	FMW (7,63%)
	FMU (5,31%)

On peut ainsi caractériser l'axe en termes d'opposition entre individus : ici, femmes autres que "femmes actives" v/s hommes actifs ou non précisé. Il peut également être intéressant d'étudier comment l'axe classe les individus.

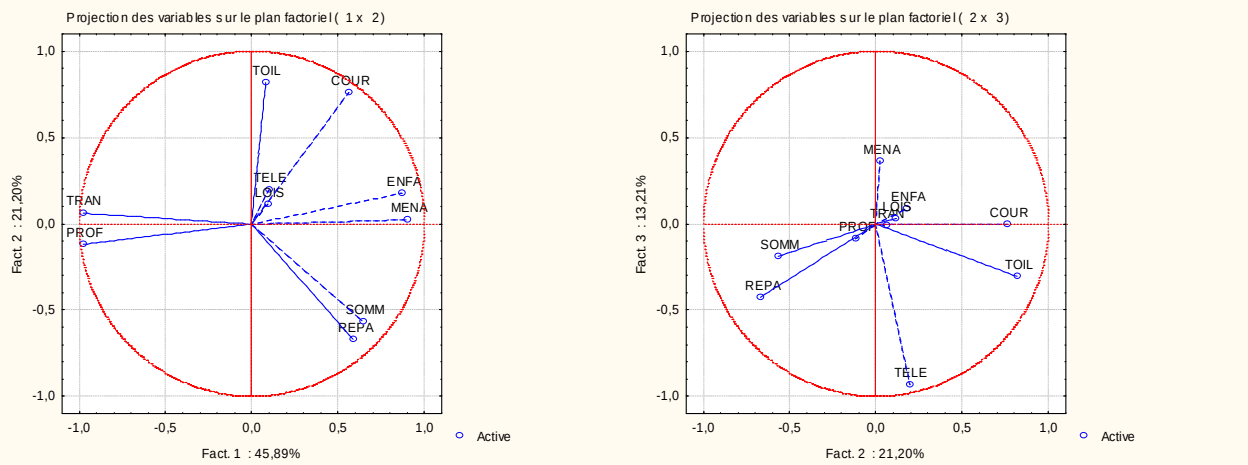
Si un individu a une contribution très forte à la formation d'un axe, on peut choisir de recommencer l'analyse en retirant cet individu, puis de l'introduire en tant qu'individu illustratif.

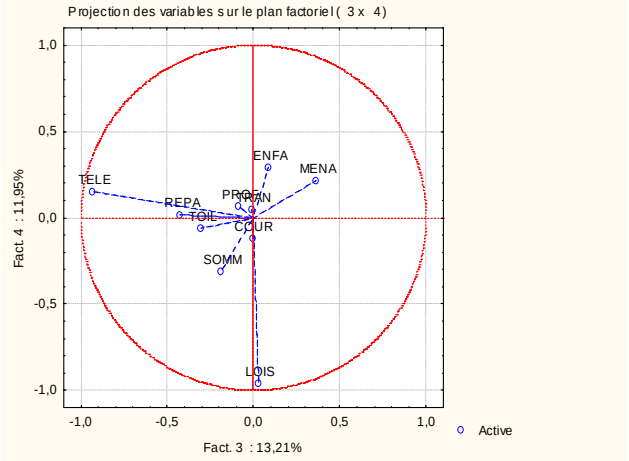
1.8 Résultats relatifs aux variables

Les résultats numériques donneront les saturations des variables (coordonnées des variables), leurs contributions à la formation des composantes principales et les qualités de leur représentation qui sont calculées, de façon

cumulative (qualité de la projection sur l'axe1, puis sur le plan 1-2, puis selon l'espace 1-2-3). Ces résultats permettent de construire et d'interpréter les plans factoriels des variables

1.9 Qualité de représentation





Les diagrammes représentant les projections des variables sur les axes factoriels nous fournissent plusieurs types d'informations :

- La longueur du vecteur représentant la variable est liée à la qualité de la représentation de la variable par sa projection dans ce plan factoriel : le carré de la longueur est la qualité de la représentation.

- Pour les variables bien représentées, l'angle entre deux variables est lié au coefficient de corrélation entre ces variables (si la représentation est exacte, le coefficient de corrélation est le cosinus de cet angle). Ceci permet de dégager des "groupes de variables" de significations voisines, des groupes de variables qui "s'opposent", des groupes de variables relativement indépendants entre eux.
-
- De même, pour les variables bien représentées, l'angle que fait la projection de la variable avec un axe factoriel est lié au coefficient de corrélation de cette variable et de l'axe factoriel. Cela permet d'interpréter les plans factoriels des individus.

- L'exemple des notes est un cas (fréquent en pratique) où toutes les variables sont corrélées positivement entre elles. Le premier axe factoriel correspond alors à une synthèse de l'effet commun à ces variables. Dans notre exemple, cela correspondrait au "niveau scolaire général" des sujets. Ce facteur a souvent une interprétation évidente et l'étude doit s'attacher à analyser les facteurs suivants. Ce phénomène est connu sous le nom d'"effet taille".

1.9.1 Contributions des variables

L'examen du tableau des contributions des variables peut permettre d'identifier des variables qui ont un rôle dominant dans la formation d'un axe factoriel. Pour l'exemple "Budget-Temps-ONU", on voit ainsi que les

variables PROF, TRAN, MENA, ENFA jouent un rôle prépondérant dans la formation du premier axe. En revanche, les axes factoriels N°3 et 4 représentent essentiellement les variables TELE et LOIS.

1.10 Variables et individus illustratifs

Plusieurs motifs peuvent nous pousser à déclarer certaines variables comme et/ou certains individus comme illustratifs.

Par exemple, lorsque des individus ou des variables ont une influence trop importante sur les résultats d'une ACP, on peut essayer de recommencer les calculs en les déclarant comme individus inactifs ou variables illustratives.

Les données correspondantes n'interviennent plus dans le calcul de détermination des composantes principales. En revanche, on leur applique les mêmes transformations qu'aux autres données afin de les ré-introduire dans les tableaux et graphiques de résultats.

Les variables illustratives peuvent également être des variables qualitatives. Dans l'exemple, nous disposons d'une variable catégorisée "sexe" et d'une variable "zone géographique". Il serait intéressant de faire apparaître sur les graphiques des points représentant les moyennes observées sur les deux sexes, ou les moyennes correspondant à chacune des 4 zones géographiques étudiées. Pour cela, pour la variable sexe par exemple, un individus « homme moyen » et un individu « femme moyenne » sont projetés sur les plans factoriels. On obtient alors des résultats tels que le suivant :