

# Introduction aux analyses multidimensionnelles

## Présentation de l'enseignement

### ***Bibliographie***

Bry X., Analyses factorielles simples, 1995, Ed. Economica  
Bry X., Analyses factorielles multiples, 1996, Ed. Economica  
Busca D., Toutain S., Analyse factorielle simple en sociologie, De Boeck, 2009, Bruxelles  
Cibois P. : L'analyse factorielle, 2000, PUF, Coll. Que sais-je N° 2095  
Crucianu M., Asselin de Beauville J-P., Boné R. : Méthodes factorielles pour l'analyse de données  
Doise W., Clémence A., Lorenzi-Cioldi F. : Représentations sociales et analyses de données, 1992, PUG, Grenoble  
Ecoffier B., Pagès J.C, Analyses factorielles simples et multiples, 1998, Dunod  
Lebart L., Morineau A., Piron M. : Statistique exploratoire multidimensionnelle, 2000, Dunod  
Rouanet H., Le Roux B. : Analyse des données multidimensionnelles, 1993, Dunod

### ***Autres sources de documentation***

Site internet de ce cours :  
<http://geai.univ-brest.fr/~carpentier/>

Autres sites à visiter :  
Le site de l'enseignement de Statistiques de l'Université de Paris 5  
<http://piaget.psycho.univ-paris5.fr/Statistiques/>  
Documents rédigés par R. Palm (aux formats Postscript et Pdf) :  
<http://www.fsagx.ac.be/si/NotesdeStatetInfo.htm>

### ***Programme***

Analyse d'un protocole multinumérique. Nuage euclidien. Inertie et variance d'un nuage. Directions principales d'un nuage. Analyse en composantes principales.  
Description d'un tableau de contingence : effectifs, fréquences, taux de liaison. Coefficient de contingence. Analyse factorielle des correspondances.  
Analyse des correspondances multiples. Tableau disjonctif de Burt. Nuage des modalités. Nuage des individus, des patrons.  
Classification ascendante hiérarchique. Distances et indices de similarité. Indices d'agrégation.

Travaux dirigés en salle d'ordinateurs. Réalisation d'analyses multidimensionnelles à l'aide d'un logiciel de traitements statistiques : analyse en composantes principales, analyse factorielle des correspondances,

analyse des correspondances multiples, classifications ascendantes hiérarchiques. Interprétation des résultats numériques et graphiques fournis par le logiciel.

Contrôle des connaissances : (contrôle continu)

Examen écrit (2 heures) (70%) + Evaluation de TD (30%)

# 1 Analyse en composantes principales ou ACP

## 1.1 Introduction

On a observé  $p$  variables sur  $n$  individus. On dit qu'il s'agit d'un protocole multivarié.

A la différence de la régression linéaire, aucune variable ne joue ici un rôle particulier. On s'intéresse à l'étude de la *variabilité* observée sur l'ensemble des individus ou l'ensemble des variables, avec l'idée suivante :

*trouver des variables abstraites, en petit nombre, reproduisant de la façon la moins déformée possible la variabilité observée.*

Du point de vue des variables : on cherche à remplacer les  $p$  variables par  $q$  nouvelles variables résumant au mieux le protocole, avec  $q \leq p$  et si possible  $q=2$ . Nous verrons que l'ACP permet de résumer un ensemble de variables corrélées en un nombre réduit de variables (appelées facteurs) non corrélées.

Du point de vue des individus : chaque individu est représenté par un point dans un espace de dimension  $p$ . On peut calculer les distances (euclidiennes) entre deux individus, entre un individu et le point moyen du nuage, etc. On cherche alors à trouver une projection des individus dans un espace de dimension  $q \leq p$ , respectant au mieux les distances entre les individus (une "carte", la moins déformée possible).

## 1.2 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

### *Données centrées réduites :*

En général, les variables retenues pour décrire les individus sont exprimées avec des unités différentes, et ne sont pas directement comparables entre elles. Dans la plupart des cas, on procède donc à un centrage-réduction des variables de départ.

Autrement dit, on remplace chaque variable  $X_i$  par la variable centrée réduite associée :

$$Z_i = \frac{X_i - \text{moyenne}(X_i)}{\text{écart type}(X_i)}$$

Cette nouvelle variable  $Z_i$  est sans unité. Elle a pour moyenne 0 et pour écart type 1.

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

On définit ainsi  $p$  variables  $Z_1, Z_2, \dots, Z_p$ .

La somme des valeurs de chaque colonne est nulle (données centrées, moyenne nulle pour chaque variable). La somme des carrés des valeurs de chaque colonne est 9 (données réduites, donc d'écart type égal à 1).

### ***Nuage des individus - Inertie du nuage***

On peut représenter un protocole bivarié à l'aide d'un nuage de points dans un plan. De manière analogue, on peut représenter chacun des individus du tableau précédent par un point d'un *espace géométrique à 5 dimensions* (autant que de variables). Il s'agit alors d'un *espace multidimensionnel*.

Le nuage des individus est l'ensemble des 9 points correspondant aux 9 sujets, pris dans un espace de dimension 5 (le nombre de variables).

La variabilité observée entre les 9 sujets est mesurée par *l'inertie* du nuage de points (vocabulaire issu de la mécanique) par rapport au point  $O$ , origine des coordonnées, et également point moyen du nuage.

L'inertie totale du nuage est :  $\sum OM_i^2 = \sum \sum z_{ij}^2 = n \times p = 9 \times 5 = 45$ .

Inertie (absolue) de l'individu  $i$  :  $OM_i^2$ .

Inertie relative de l'individu  $i$  :  $Inr_i = \frac{OM_i^2}{\sum_j OM_j^2}$

L'inertie relative d'un individu est d'autant plus grande que les valeurs des variables observées sur cet individu sont "loin de la moyenne".

### ***Inertie le long d'un axe, inertie selon un sous-espace***

L'inertie (absolue) de l'individu  $i$  le long d'un axe  $D$  est  $OH_i^2$ , où  $H_i$  est la projection orthogonale du point  $M_i$  sur l'axe  $D$ . L'inertie relative correspondante est  $\frac{OH_i^2}{\sum_j OH_j^2}$ .

On peut définir de même l'inertie selon un plan, un espace de dimension 3, etc.

**Nuage des variables**

De façon duale, on peut considérer les 5 points correspondant aux 5 variables, dans un espace de dimension 9 (le nombre des individus).

L'inertie absolue de chaque variable est  $n$ , son inertie relative est  $\frac{1}{p}$ .

**Corrélations des variables prises deux à deux :**

	Math	Sciences	Français	Latin	Musique
Math	1,0000	0,9825	0,2267	0,4905	0,0112
Sciences	0,9825	1,0000	0,3967	0,6340	0,0063
Français	0,2267	0,3967	1,0000	0,9561	0,0380
Latin	0,4905	0,6340	0,9561	1,0000	0,0886
Musique	0,0112	0,0063	0,0380	0,0886	1,0000

Comme les variables sont centrées réduites, la corrélation entre la variable  $Z_k$  et la variable  $Z_l$  est

simplement  $\frac{1}{n} \sum_i z_{ik} z_{il}$ .

Dans notre exemple, toutes les variables sont corrélées positivement. La corrélation est forte entre les 2 premières, et entre la 3<sup>e</sup> et la 4<sup>e</sup>. La cinquième est faiblement corrélée aux autres variables.

**1.2.1 Analyse en composantes principales (normée)****1.2.1.1 Aperçu sur les bases mathématiques de l'ACP**

*Ce paragraphe pourra être ignoré en première lecture.*

Abandonnons provisoirement l'exemple précédent et explorons quelques situations comportant un très petit nombre de variables et d'observations;

Ces exemples pourront aussi être explorés à l'aide du classeur Excel [Valeurs-propres.xls](#).

Considérons un exemple avec 2 variables et 3 individus. Par exemple :

	V1	V2
i1	1	6
i2	2	3
i3	3	4

Les variables centrées réduites associées à V1 et V2 sont données par :

	X <sub>1</sub>	X <sub>2</sub>
i1	-1,225	1,336
i2	0	-1,069
i3	1,225	-0,267

On cherche une variable U :

- combinaison linéaire de X<sub>1</sub> et X<sub>2</sub>
- représentant au mieux la variabilité observée sur les 3 individus, c'est-à-dire de variance maximale.

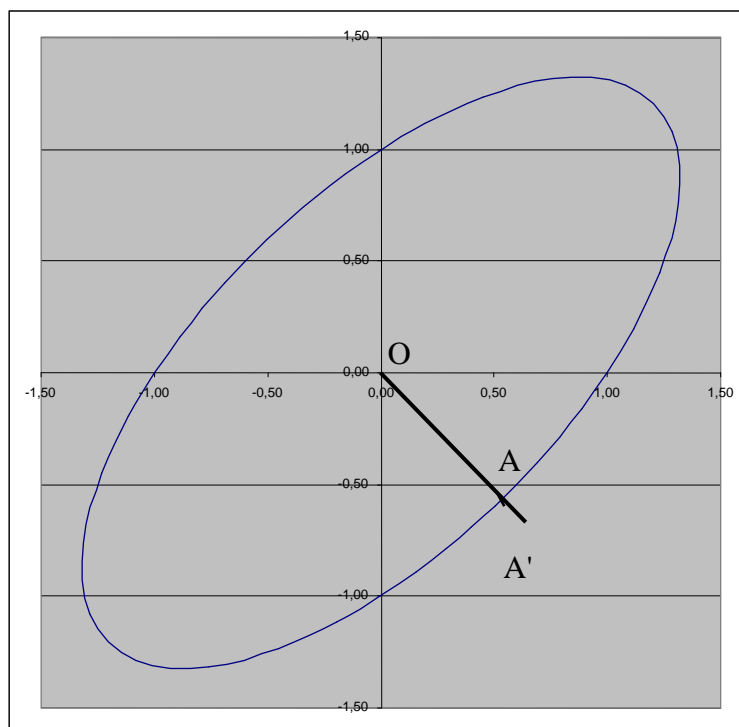
Pour que le problème posé ait un sens, il faut en outre poser une condition sur les coefficients de la combinaison linéaire U. Ainsi, nous recherchons  $U = t_1 X_1 + t_2 X_2$ , avec la condition  $t_1^2 + t_2^2 = 1$ , telle que  $\text{Var}(U)$  soit maximale.

Rappel :  $\text{Var}(U) = t_1^2 \text{Var}(X_1) + t_2^2 \text{Var}(X_2) + 2 t_1 t_2 \text{Cov}(X_1, X_2)$

Comme  $X_1$  et  $X_2$  sont centrées réduites, on a ici :  $\text{Var}(U) = t_1^2 + t_2^2 + 2 r t_1 t_2$ , où  $r$  désigne le coefficient de corrélation des variables  $X_1$  et  $X_2$ . Sur l'exemple proposé,  $r = -0,65$ .

Le dessin ci-dessous représente les points M de coordonnées  $(t_1, t_2)$  tels que  $\text{Var}(U)=1$  (mais ne vérifiant pas nécessairement  $t_1^2 + t_2^2 = 1$ ). La courbe ainsi obtenue est l'ellipse d'inertie du nuage de points. L'un des deux points de cette ellipse les plus proches de l'origine est le point A, et c'est dans la direction (OA) que l'on trouve la solution au problème posé : il suffit de prendre le point A' tel que  $\overrightarrow{OA'} = \frac{1}{OA} \overrightarrow{OA}$ . La

variance ainsi obtenue est égale à  $\frac{1}{OA^2}$ . Elle est aussi appelée *valeur propre* associée au *vecteur propre* ou axe factoriel  $\overrightarrow{OA'}$



Considérons maintenant un exemple comportant 3 variables et 4 individus.

Les variables centrées réduites sont par exemple données par :

	X1	X2	X3
i1	-1,183	1,606	-0,507
i2	-0,507	-1,147	0,169
i3	0,169	-0,229	-1,183
i4	1,521	-0,229	1,521

Le tableau ou *matrice* des corrélations ( $r_{ij}$ ) entre ces variables est :

$$S = \begin{bmatrix} 1 & -0,43 & 0,66 \\ -0,43 & 1 & -0,27 \\ 0,66 & -0,27 & 1 \end{bmatrix}$$

Comme précédemment, on recherche une combinaison linéaire :

$$U = t_1 X_1 + t_2 X_2 + t_3 X_3 \quad \text{avec } t_1^2 + t_2^2 + t_3^2 = 1$$

telle que  $\text{Var}(U)$  soit maximale.

Or, la variance de  $U$  s'exprime à l'aide de  $t_1, t_2, t_3$  et des coefficients de corrélation  $r_{ij}$ . La multiplication des matrices est une opération qui a été définie par les mathématiciens de telle façon qu'en introduisant la

matrice colonne  $T = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}$ , on peut écrire :  $\text{Var}(U) = 'T S T$ .

On peut alors déterminer la première valeur propre et le premier axe factoriel à l'aide de la méthode itérative suivante :

- On part d'une matrice colonne  $T_0$  arbitraire ;
- On calcule le produit de matrices  $S T_0$  ; c'est une matrice colonne, que l'on normalise (somme des carrés des coefficients égale à 1). On obtient ainsi  $T_1$
- On répète l'opération avec la matrice  $T_1$ .

Au bout de quelques itérations, la matrice obtenue ne diffère pratiquement plus de la matrice obtenue à l'étape précédente. On obtient ainsi une matrice  $T$  et un nombre  $\lambda$  tels que :

$$S T = \lambda T \quad \text{et} \quad \text{Var}(U) = \lambda$$

Exécutées à l'aide d'Excel sur les données fournies en exemple, ces opérations donnent les résultats suivants :

T0	1,000	0,000	0,000
T1	0,787	-0,336	0,517
T2	0,675	-0,432	0,598
T3	0,652	-0,459	0,603
T4	0,647	-0,469	0,601
T5	0,646	-0,472	0,600
T6	0,646	-0,473	0,599
T7	0,645	-0,474	0,599

Variance	1,000
	1,855
	1,920
	1,923
	1,923
	1,923
	1,923
	1,923
	1,923

S*T0	1,000	-0,427	0,657
S*T1	1,270	-0,812	1,126
S*T2	1,252	-0,882	1,159
S*T3	1,244	-0,901	1,156
S*T4	1,242	-0,908	1,154
S*T5	1,242	-0,910	1,153
S*T6	1,241	-0,911	1,152
S*T7	1,241	-0,912	1,152

La première valeur propre est ainsi 1,923, et la première composante principale est :

$$CP_1 = 0,645 X_1 - 0,474 X_2 + 0,599 X_3.$$

On peut ensuite poursuivre l'étude en calculant les résidus de la régression des variables  $X_1, X_2$  et  $X_3$  par rapport à  $CP_1$  et en poursuivant la méthode sur ces nouvelles variables.

### 1.2.1.2 Valeurs propres et vecteurs propres. Composantes principales

On reprend l'exemple introduit au paragraphe 1.1.

On cherche à projeter le nuage de points dans des espaces géométriques de dimension 1 (un axe), 2 (un plan), 3 (un espace "ordinaire"), ... de manière à déformer le moins possible les distances entre les individus, et donc la variabilité observée.

En fait, la méthode mathématique utilisée travaille axe par axe, chaque axe portant le nom de *axe factoriel* :

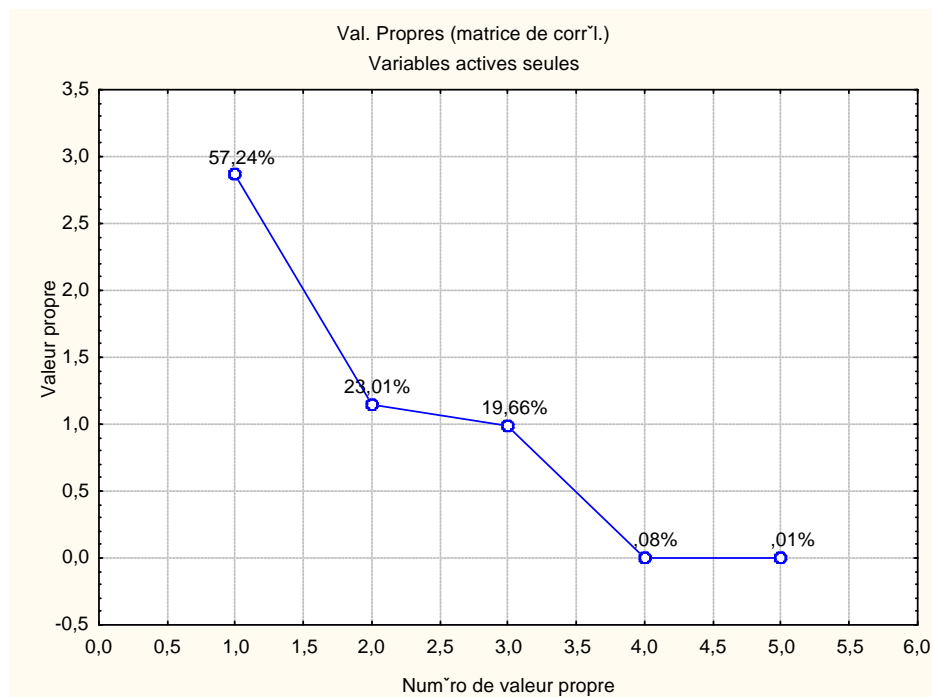
- Le meilleur axe est le premier axe factoriel ;
- Le meilleur plan est défini par l'axe factoriel précédent et un deuxième axe orthogonal au précédent.

La projection des individus sur ces axes produit de nouvelles variables appelées *composantes principales*. Les composantes principales  $CP_1, CP_2, \dots, CP_p$  sont des variables obtenues comme combinaisons linéaires des variables de départ, et qui vérifient les propriétés suivantes :

- $CP_1$  représente la direction de plus grande dispersion du nuage de points.
- $CP_2$  représente la direction de plus grande dispersion des résidus, une fois l'effet de  $CP_1$  pris en compte
- même chose pour  $CP_3, CP_4$ , etc
- Les variables  $CP_k$  sont indépendantes : si  $k \neq l$ , alors  $Cov(CP_k, CP_l) = 0$
- Les variables  $CP_k$  ne sont en général pas réduites : la variance de la composante principale  $CP_k$  est égale à la k-ième valeur propre.

Le terme de "valeur propre" (en anglais : eigenvalue) appartient au domaine de l'algèbre linéaire. Il s'agit en fait des valeurs propres de la matrice des corrélations. Mathématiquement, on dit que la matrice des corrélations et la matrice diagonale des valeurs propres sont semblables : elles représentent la même information (l'inertie du nuage de points) dans deux systèmes d'axes orthonormés différents.

	Val. propr	% Total variance	Cumul Val. propr	Cumul %
1	2,8618	57,24	2,86	57,24
2	1,1507	23,01	4,01	80,25
3	0,9831	19,66	5,00	99,91
4	0,0039	0,08	5,00	99,99
5	0,0004	0,01	5,00	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 3<sup>e</sup> et la 4<sup>e</sup> valeur propre. Au final, on décide de ne garder que trois valeurs propres.

### 1.2.1.3 Résultats relatifs aux individus

#### *Coordonnées ou scores des individus*

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

*Coordonnées factorielles des ind., basées sur les corrélations (crucianu-1-1.sta)*

Var. illustrative : Sujet

	Fact. 1	Fact. 2	Fact. 3	Sujet
1	-2,7857	0,6765	0,7368	Jean
2	-1,2625	0,3303	0,5549	Aline
3	-1,0167	-1,0198	0,2881	Annie
4	3,1222	0,1659	1,1442	Monique
5	1,9551	0,7879	0,1892	Didier
6	-0,9477	1,2014	-1,1401	André
7	-0,3250	-1,7548	0,9095	Pierre
8	0,6374	1,1298	-0,6919	Brigitte
9	0,6231	-1,5173	-1,9909	Evelyne

On peut remarquer que Statistica nous propose un second tableau de résultats, nommé *Scores Factoriels*. Ces derniers sont obtenus à partir des coordonnées factorielles, en divisant chaque coordonnée par la racine carrée de la valeur propre correspondante.

#### *Contributions des individus*

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

$$CTR(S_i, CP_k) = \frac{(\text{Score de } S_i \text{ selon } CP_k)^2}{\sum_j (\text{Score de } S_j \text{ selon } CP_k)^2} = \frac{(\text{Score de } S_i \text{ selon } CP_k)^2}{n \lambda_k}$$

$$\text{Par exemple : } CTR(S_1, CP_1) = \frac{(-2,7857)^2}{2,7857^2 + 1,2625^2 + \dots + 0,6231^2} = \frac{(-2,7857)^2}{9 \times 2,8618} = 0,3013$$

*Contributions des ind., basées sur les corrélations (crucianu-1-1.sta)*

Var. illustrative : Sujet

	Fact. 1	Fact. 2	Fact. 3	Sujet
1	30,13	4,42	6,14	Jean
2	6,19	1,05	3,48	Aline
3	4,01	10,04	0,94	Annie
4	37,85	0,27	14,80	Monique
5	14,84	5,99	0,40	Didier
6	3,49	13,94	14,69	André
7	0,41	29,73	9,35	Pierre
8	1,58	12,33	5,41	Brigitte
9	1,51	22,23	44,79	Evelyne

#### *Qualités de la représentation des individus: cosinus carrés*

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

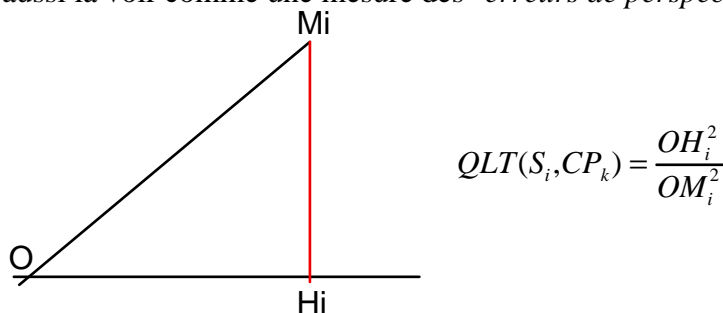


$$QLT(S_i, CP_k) = \frac{(\text{Score de } S_i \text{ selon } CP_k)^2}{\sum_l (\text{Score de } S_i \text{ selon } CP_l)^2} = \frac{(\text{Score de } S_i \text{ selon } CP_k)^2}{\text{Inertie}(S_i)}$$

Par exemple :

$$QLT(S_1, CP_1) = \frac{(-2,7857)^2}{2,7857^2 + 0,6765^2 + \dots + 0,0332^2} = \frac{(-2,7857)^2}{1,0865^2 + 1,2817^2 + 1,5037^2 + 1,6252^2 + 1,0190^2} = 0,8855$$

Géométriquement, la qualité de la représentation d'un individu i par la composante principale k est égale à  $\cos^2 \theta$ , où  $\theta$  est l'angle  $(\overrightarrow{OM_i}, \overrightarrow{CP_k})$ . Elle mesure la "déformation" due à la projection sur la composante principale  $CP_k$ . On peut aussi la voir comme une mesure des "erreurs de perspective".



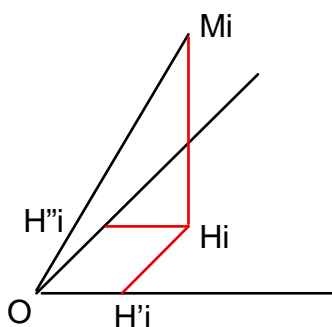
Cosinus carrés, basés sur les corrélations (crucianu-1-1.sta)

Var. illustrative : Sujet

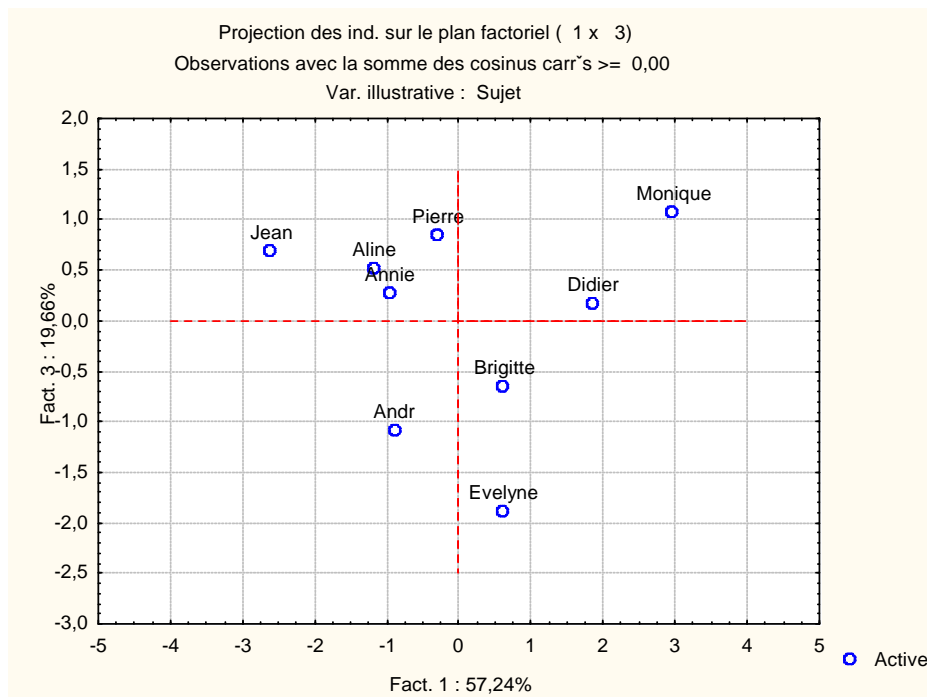
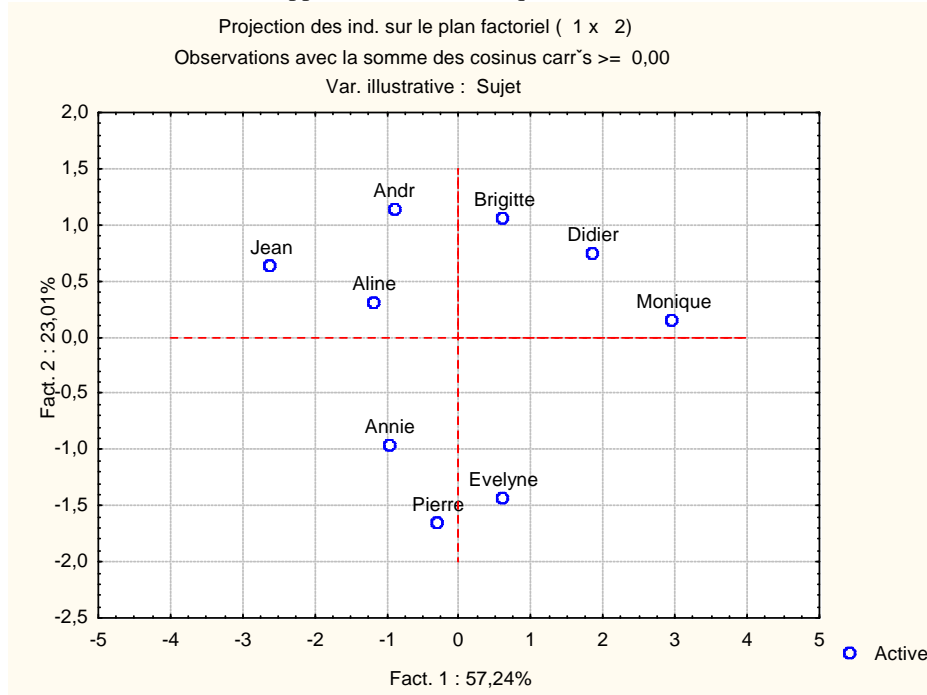
	Fact. 1	Fact. 2	Fact. 3	Sujet
1	0,8855	0,0522	0,0619	Jean
2	0,7920	0,0542	0,1530	Aline
3	0,4784	0,4813	0,0384	Annie
4	0,8786	0,0025	0,1180	Monique
5	0,8515	0,1383	0,0080	Didier
6	0,2465	0,3962	0,3568	André
7	0,0263	0,7671	0,2061	Pierre
8	0,1877	0,5898	0,2211	Brigitte
9	0,0583	0,3458	0,5954	Evelyne

Les qualités de représentation sont additives. Par exemple, la qualité de représentation d'un individu i par le plan (CP1, CP2) est donnée par :

$$QLT(S_i, CP_1; CP_2) = \frac{(\text{Score de } S_i \text{ selon } CP_1)^2 + (\text{Score de } S_i \text{ selon } CP_2)^2}{\sum_l (\text{Score de } S_i \text{ selon } CP_l)^2}$$



Pour le sujet 1 (Jean), la qualité de représentation par le plan factoriel 1x2 est :  $0,8855 + 0,0522 = 0,9377$ . Cette valeur représente le carré du cosinus de l'angle que fait  $\overrightarrow{OM_i}$  avec le plan (CP1, CP2).



### 1.2.1.4 Résultats relatifs aux variables

#### *Coordonnées ou saturations des variables*

Les saturations des variables sont les coefficients de corrélation entre les variables (centrées réduites) de départ et les variables factorielles.

$$SAT(Z_j, CP_k) = \rho(Z_j, CP_k)$$

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

$$SAT(Z_1, CP_1) = \frac{(-1,0865)(-2,7857) + (-0,4939)(-1,2625) + (-1,0865)(-1,0168) + (1,4322)(3,1222) + (1,2840)(1,9551) + (0,3951)(-0,9478) + (-1,2347)(-0,3250) + (0,9877)(0,6373) + (-0,1975)(0,6231)}{9\sqrt{2,8618}}$$

*Coord. factorielles des var., basées sur les corrélations (crucianu-1-1.sta)*

	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 :

$$CTR(Z_i, CP_k) = \frac{(Saturation\ de\ Z_i\ selon\ CP_k)^2}{\sum_j (Saturation\ de\ Z_j\ selon\ CP_k)^2} = \frac{(Saturation\ de\ Z_i\ selon\ CP_k)^2}{\lambda_k}$$

Par exemple :  $CTR(Z_1, CP_1) = \frac{0,8059^2}{2,8618} = 0,2269$

*Contributions des var., basées sur les corrélations (crucianu-1-1.sta)*

	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 :

$$QLT(Z_i, CP_k) = \frac{(Saturation\ de\ Z_i\ selon\ CP_k)^2}{\sum_l (Saturation\ de\ Z_i\ selon\ CP_l)^2} = (Saturation\ de\ Z_i\ selon\ CP_k)^2$$

Mais, comme les variables  $Z_i$  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, basées sur les corrélations (crucianu-1-1.sta)*

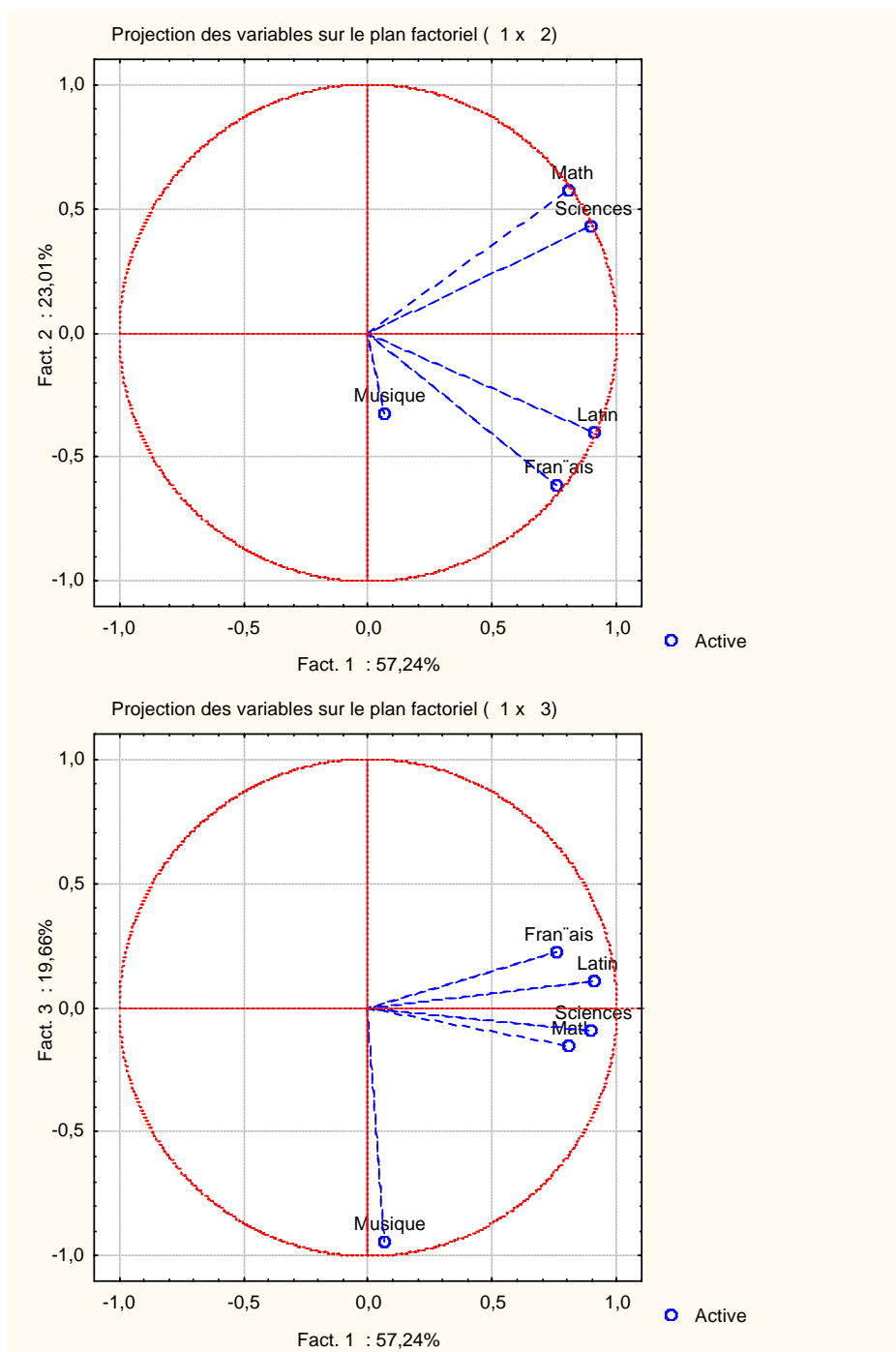
	Avec 1 facteur	Avec 2 facteurs	Avec 3 facteurs
Math	0,6495	0,9759	0,9995

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

Graphiquement, la qualité de la représentation d'une variable dans le plan (CP<sub>1</sub>, CP<sub>2</sub>) est le carré de la norme (longueur) du vecteur représentant cette variable (projection de cette variable dans le plan).

Sur notre exemple, on constate que toutes les variables ont une coordonnée positive selon le premier axe. Un résultat de ce type (saturations de même signe pour toutes les variables sur le premier axe) est obtenu chaque fois que les variables sont toutes corrélées positivement entre elles. Le premier axe représente alors l'effet conjoint de l'ensemble des variables et la variable factorielle correspondante est appelée "effet taille".

**Représentation des variables :**



### 1.2.1.5 Résultats relatifs à l'analyse elle-même :

#### Coefficients des variables :

Le tableau des coefficients des variables ("loadings" en anglais) peut être lu de deux façons :

- il permet de calculer les valeurs des composantes principales à partir des variables centrées réduites de départ
- il permet de retrouver les valeurs des variables centrées réduites de départ à partir des valeurs des composantes principales.

*Vecteurs propres de la matrice de corrélation (crucianu-1-1.sta)*

Variables actives seules

	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5
Math	0,4764	0,5326	-0,1548	-0,3030	0,6112
Sciences	0,5302	0,4016	-0,0936	0,5168	-0,5308
Français	0,4481	-0,5696	0,2276	0,4775	0,4414
Latin	0,5381	-0,3706	0,1093	-0,6416	-0,3868
Musique	0,0394	-0,3053	-0,9505	0,0390	0,0140

Les valeurs de ce tableau sont obtenues en divisant les saturations des variables par la racine carrée de la valeur propre de la composante principale correspondante. Mathématiquement, ce tableau est la matrice de "changement de base orthonormée" permettant de passer des variables  $Z_i$  aux composantes principales  $CP_k$  ou vice-versa. On observera que :

- chaque ligne représente un vecteur de norme 1
- chaque colonne représente un vecteur de norme 1
- deux "vecteurs ligne" quelconques sont orthogonaux
- deux "vecteurs colonne" quelconques sont orthogonaux

Pour l'individu 1, les variables de départ ont pour valeurs :

Math	Sciences	Français	Latin	Musique
-1,0865	-1,2817	-1,5037	-1,6252	-1,0190

On retrouve ainsi le score de cet individu sur la première composante principale :

$$CP_{11} = (-1,0865)(0,4764) + (-1,2817)(0,5302) + (-1,5037)(0,4481) + (-1,6252)(0,5381) + (-1,0190)(0,0394) = -2,7857$$

Pour l'individu 1, les scores sur les 5 composantes principales sont :

Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5
-2,7857	0,6764	0,7368	-0,0482	-0,0332

On retrouve ainsi la valeur de la première composante principale sur cet individu :

$$CP_{11} = (-2,7857)(0,4764) + (0,6764)(0,5326) + (0,7368)(-0,1548) + (-0,0482)(-0,3030) + (-0,0332)(0,6112) = -1,0865$$

Les valeurs propres pourraient également être calculées à partir du tableau, comme variances des composantes principales. Autrement dit, on pourrait à l'aide du tableau des coefficients, retrouver tous les résultats indiqués ci-dessus.

Ce tableau permet également de retrouver les saturations des variables, en multipliant les coefficients correspondant à chaque facteur par la racine carrée de la valeur propre correspondante.

Par exemple, pour la première variable et la première composante principale :

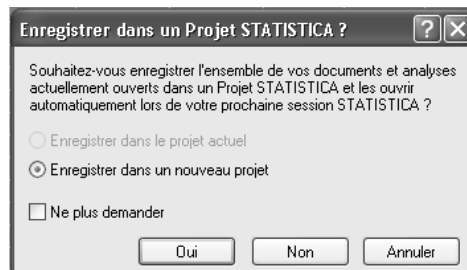
$$SAT(Z_1, CP_1) = 0,4764 \times \sqrt{2,8618} = 0,8059$$

## 1.3 Analyse en composantes principales avec Statistica

### 1.3.1 Quelques indications sur Statistica version 9

#### 1.3.1.1 Qu'est-ce qu'un fichier de projet ?

Outre les feuilles de données, graphiques, rapports, macros, classeurs, etc., Statistica 9 propose d'enregistrer notre travail sous forme de fichier de projet (extension .spf). Ce format de fichier est destiné à enregistrer un instantané à un moment donné de notre session de travail. On peut enregistrer un projet avec l'item correspondant du menu Fichier, ou en réponse à une fenêtre de dialogue affichée lorsque l'on quitte Statistica :



Mais attention ! Par défaut, le fichier de projet ne contient pas les documents qui sont enregistrés sur le disque, mais seulement des liens vers les fichiers qui contiennent ces documents. Ces liens cessent d'ailleurs de fonctionner dès que le fichier correspondant est renommé. Pour changer ce comportement, veuillez à activer le bouton radio "Intégrer le document dans le projet" du dialogue d'enregistrement.



### 1.3.1.2 Le menu Fichier - Ouvrir une URL ...


Le menu Fichier - Ouvrir une URL... affiche un dialogue permettant de saisir l'adresse d'une page Web. Par exemple :



On peut alors naviguer sur les différentes pages du site et, le cas échéant, ouvrir directement un document Statistica disponible en ligne en cliquant sur un lien vers ce document.

Mais, il faut penser à enregistrer dans notre répertoire de travail (Mes Documents, Bureau, etc.), le document ouvert de cette façon, sinon, Statistica se plante lorsqu'on quitte le logiciel. Et, c'est alors la commande "Enregistrer sous..." qu'il convient d'utiliser, pour éviter que l'enregistrement conduise à un message d'erreur ou se fasse dans le cache du navigateur.

### 1.3.1.3 Reprendre / ré-exécuter une analyse après l'avoir refermée

La plupart des icônes de dossiers des classeurs élaborés sous Statistica 9 comportent une flèche rouge : . On peut alors ré-exécuter les traitements correspondants en faisant un clic droit sur l'icône du dossier et en sélectionnant le menu convenable. Mais cette fonctionnalité disparaît si l'on modifie la hiérarchie des dossiers dans le classeur, par exemple.

### 1.3.1.4 Les rapports

Comme dans les versions précédentes, on peut demander à ce que les résultats des traitements soient écrits dans un rapport, en plus des feuilles de résultats. Statistica 9 offre deux sortes de rapports :

- les rapports au format Statistica, qui peuvent être inclus dans un classeur, enregistrés au format Statistica (fichiers .str) ou au format .rtf lisible par Word ;
- les rapports au format Word, dans lesquels les résultats peuvent être insérés soit sous forme d'objets Statistica, soit sous forme de tableaux de Word.

Ce dernier format peut être commode, par exemple pour reprendre des résultats de Statistica dans un mémoire. On notera que le dialogue Outils - Options comporte un item permettant de spécifier le niveau de détail désiré dans les rapports.

Les rapports (de l'un ou l'autre format) peuvent être insérés dans un classeur Statistica. Attention cependant au volume des classeurs ainsi produits. Le logiciel prévoit d'ailleurs un menu "Classeur - Compresser les graphiques et rapports", s'affichant lorsque l'objet actif est un classeur, pour limiter ce volume.

### 1.3.1.5 Le menu Données - Filtre automatique

Le menu Données - Filtre automatique qui s'affiche lorsque l'objet actif est une feuille de données peut être pratique lorsqu'on travaille sur des feuilles comportant de nombreuses observations. En revanche, les items de menu Données - Filtre automatique - Définir comme filtre de sélection et Données - Filtre automatique - Masquer les Obs. exclues des analyses peuvent conduire à des erreurs et même à des pertes de données.

### 1.3.2 Présentation des données étudiées

#### Références

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:

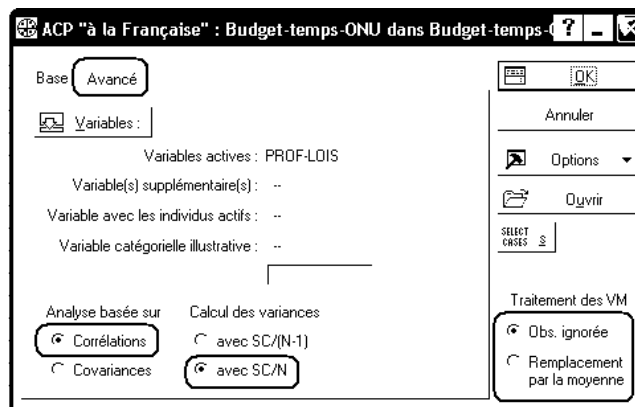
H: Hommes, F: Femmes, A: Actifs, N: Non Actifs(ves), M: Mariés, C: Célibataires, U: USA, W: Pays de l'Ouest sauf USA, E : Est sauf Yougoslavie, Y: Yougoslavie

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).

### 1.3.3 Traitement des données avec Statistica

Ouvrez le classeur Budget-temps-ONU.stw et observez les données saisies.

Pour effectuer l'ACP, nous utilisons le menu Statistiques - Techniques exploratoires multivariées - ACP "à la française".



La fenêtre de dialogue permet de spécifier les variables qui participeront à l'analyse. Elle permet également d'indiquer les différentes options choisies pour le traitement.

Utilisez l'onglet "Avancé" de cette fenêtre.

- Comment seront traitées les valeurs manquantes ? Ici, les données ne comportent pas de valeur manquante.

- L'analyse sera-t-elle basée sur les covariances ou sur les corrélations ? Sur l'exemple traité ici, la question mérite d'être posée, car toutes les données sont exprimées avec la même unité. Cependant, l'étude menée à partir des covariances ferait surtout apparaître les variables qui combinent valeurs élevées et fortes variations, telles que PROF par exemple. Le paragraphe précédent concernait l'ACP normée, c'est-à-dire l'ACP basée sur les corrélations. Nous dirons ultérieurement quelques mots sur l'ACP non normée.

- Utilise-t-on les variances et covariances non corrigées (SC/N) ou les variances et covariances corrigées (SC/(N-1)). Dans le cas d'une ACP normée, les deux méthodes fournissent des résultats presque identiques : seuls les scores des individus sont légèrement modifiés. En fait, l'ACP est une méthode descriptive et non une méthode inférentielle. Elle est effectuée dans un but exploratoire : on étudie les



données pour elles-mêmes, et non en vue d'une généralisation à une population. C'est pourquoi l'utilisation des variances non corrigées est généralement justifiée.

Cliquez ensuite sur le bouton OK.

N.B. Ne fermez pas l'analyse en cours pendant la suite des manipulations. Ainsi, vous n'aurez pas à indiquer de nouveau les options ci-dessus, vos résultats seront cohérents entre eux et se rassembleront dans un même classeur.

### 1.3.3.1 Statistiques descriptives - Matrice des corrélations

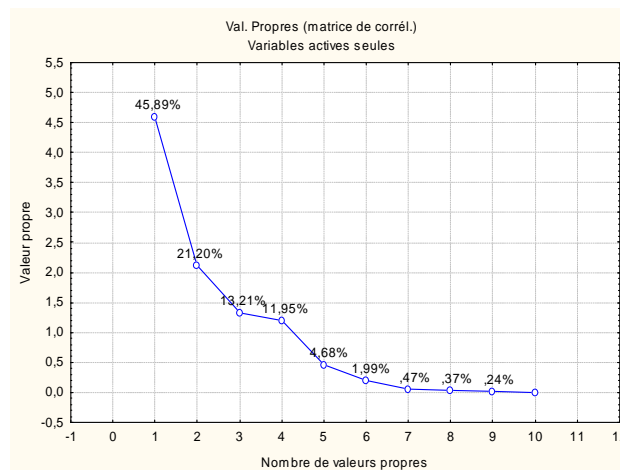
Ces résultats peuvent être obtenus à l'aide de l'onglet "Descriptives".

Par exemple, la matrice des corrélations est ici :

Variable	Corrélations (Budget-temps-ONU dans Budget-temps-ONU.stw)									
	PROF	TRAN	MENA	ENFA	COUR	TOIL	REPA	SOMM	TELE	LOIS
PROF	1,00	0,93	-0,91	-0,87	-0,66	-0,11	-0,45	-0,54	-0,06	-0,19
TRAN	0,93	1,00	-0,87	-0,81	-0,50	-0,08	-0,61	-0,70	-0,04	-0,11
MENA	-0,91	-0,87	1,00	0,86	0,50	-0,04	0,36	0,43	-0,21	-0,11
ENFA	-0,87	-0,81	0,86	1,00	0,54	0,12	0,37	0,28	0,12	-0,11
COUR	-0,66	-0,50	0,50	0,54	1,00	0,59	-0,18	-0,03	0,22	0,24
TOIL	-0,11	-0,08	-0,04	0,12	0,59	1,00	-0,36	-0,22	0,32	0,07
REPA	-0,45	-0,61	0,36	0,37	-0,18	-0,36	1,00	0,82	0,32	-0,04
SOMM	-0,54	-0,70	0,43	0,28	-0,03	-0,22	0,82	1,00	0,02	0,21
TELE	-0,06	-0,04	-0,21	0,12	0,22	0,32	0,32	0,02	1,00	-0,10
LOIS	-0,19	-0,11	-0,11	-0,11	0,24	0,07	-0,04	0,21	-0,10	1,00

### 1.3.3.2 Choix des valeurs propres

Affichez d'abord le tableau des valeurs propres et le diagramme correspondant.



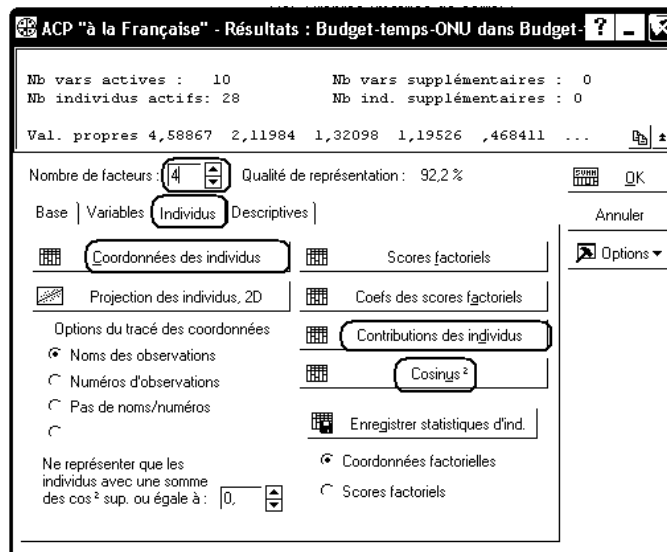
Pour cela, cliquez sur les boutons "Valeurs propres" et "Tracé des valeurs propres" de l'onglet "Base". Dans notre cas, on peut choisir de retenir 4 composantes principales. Dans les manipulations qui suivent, on indiquera donc 4 dans la zone d'édition "nombre de facteurs".

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.

Pour les résultats relatifs aux individus et aux variables, on utilisera les onglets "Individus" et "Variables" de préférence à l'onglet "Base".

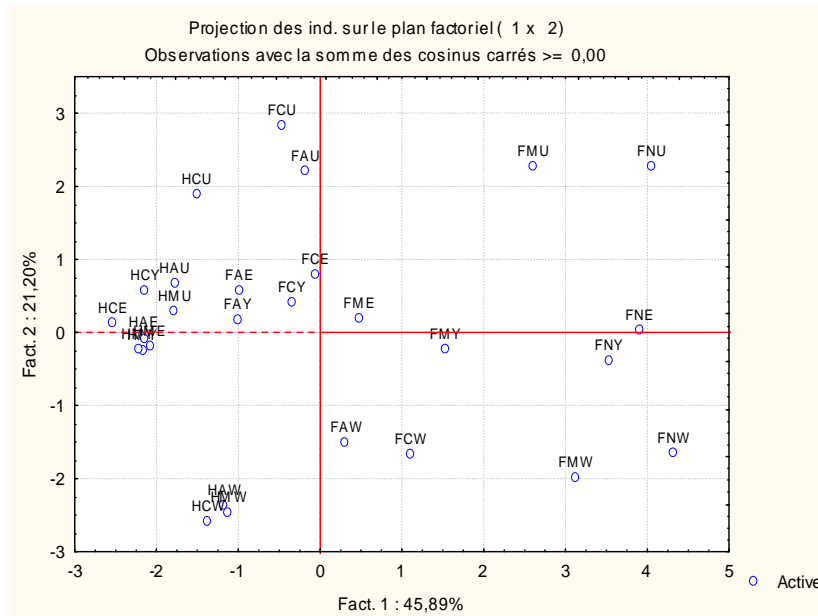
### 1.3.3.3 Résultats relatifs aux individus

On pourra obtenir successivement les scores des individus, leurs contributions à la formation des composantes principales et leurs qualités de représentation en utilisant les boutons "Coordonnées des individus", "Contributions des individus", "Cosinus<sup>2</sup>".



On peut ensuite obtenir les projections du nuage des individus selon les premiers axes factoriels à l'aide du bouton "Projection de individus, 2D". Lorsque les individus ne sont pas anonymes (c'est le cas ici), il est utile d'étiqueter chaque point. Plusieurs méthodes sont possibles :

- Utiliser les identifiants d'individus figurant dans la première colonne du tableau de données
- Utiliser les numéros des observations
- Utiliser les étiquettes indiquées dans la variable "illustrative" : ces étiquettes peuvent être des identifiants des individus, mais peuvent également représenter un groupe d'appartenance, etc.



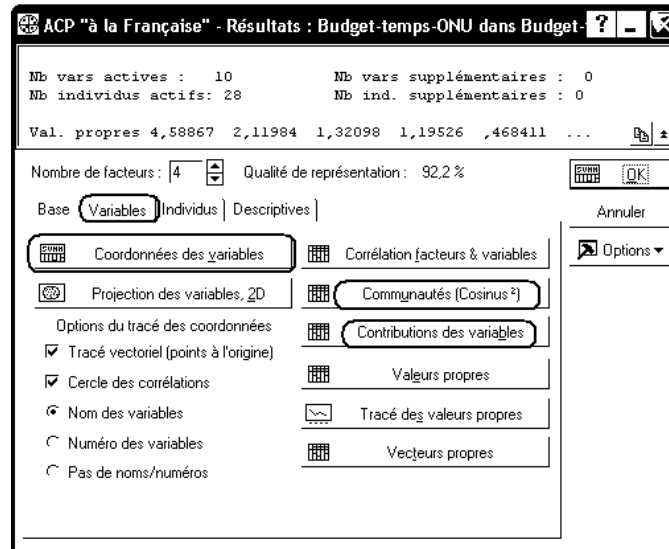
### 1.3.3.4 Résultats relatifs aux variables

Activons ensuite l'onglet "Variables".

On obtient les saturations des variables en cliquant sur le bouton "Coordonnées des variables" ou le bouton "Corrélation facteurs et variables" : dans le cas d'une ACP normée, ces deux traitements fournissent le même résultat.

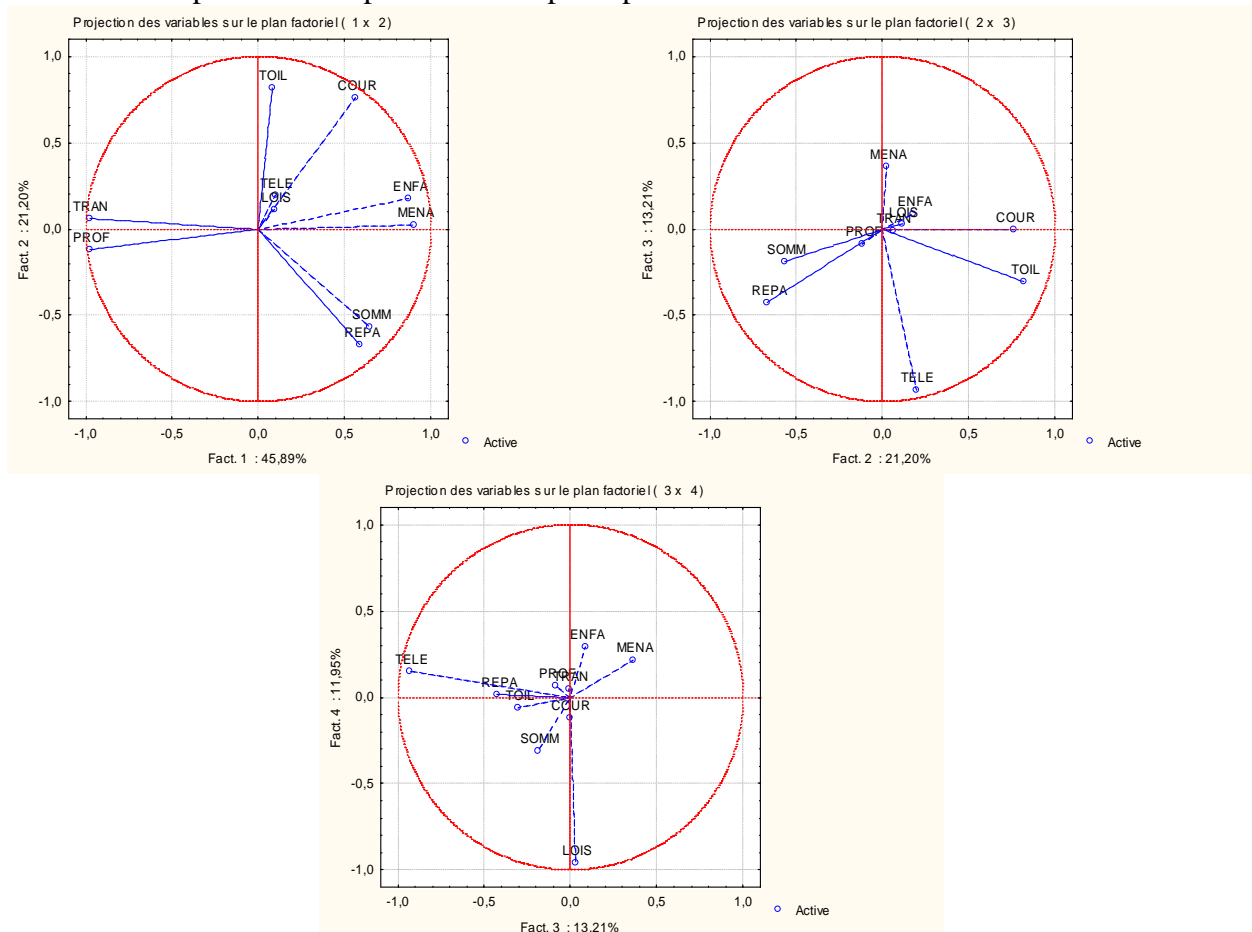
On obtient leurs contributions à la formation des composantes principales en utilisant le bouton "Contributions des variables".

Les qualités de représentation sont calculées, de façon cumulative (qualité de la projection selon CP1, puis selon le plan (CP1,CP2), puis selon l'espace (CP1,CP2,CP3) en utilisant le bouton "Communautés (Cosinus<sup>2</sup>)".



### Représentation des variables

Le bouton "Projection des variables, 2D" permet d'obtenir les diagrammes représentant les projections des variables selon les plans définis par deux axes principaux.



### 1.3.3.5 Coefficients des variables

Les coefficients des variables (c'est-à-dire la matrice permettant de passer des variables centrées réduites aux variables principales et vice-versa) seront obtenus à l'aide du bouton "Vecteurs propres" de l'onglet "Variables".

### 1.3.4 Variables supplémentaires et individus inactifs avec Statistica

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

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 supplémentaires.

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 réintroduire dans les tableaux et graphiques de résultats.

Avec Statistica, il est simple de déclarer une variable comme variable supplémentaire : le premier dialogue de l'ACP prévoit une zone d'édition pour cela. Pour déclarer des individus comme "inactifs", il est nécessaire de construire une variable supplémentaire, qui ne contiendra que deux modalités, et d'utiliser les zones d'édition "Variable avec individus actifs" et "Code des individus actifs".

Dans une étude de psychologie sociale, il arrive fréquemment que l'intérêt du chercheur se porte sur les variations et les oppositions entre groupes de sujets plutôt que sur les variations individuelles. Pour obtenir des résultats concernant ces groupes, on peut ajouter au tableau les individus inactifs, avec comme valeurs des variables, les moyennes observés sur les groupes. Ainsi, ces moyennes peuvent participer à l'interprétation des plans factoriels, mais n'ont pas participé à leur construction.

Dans l'exemple que nous traitons, 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.

#### 1.3.4.1 Calcul des moyennes par sexe, par zone géographique

##### *Faire une copie de la feuille de données Budget-Temps-ONU*

Veillez à ce que la feuille de données Budget-Temps-ONU soit un élément terminal de la hiérarchie des objets du classeur.

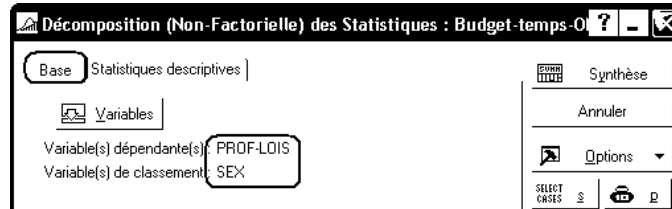
A l'aide du bouton droit de la souris, pointez l'icône de la feuille dans le volet gauche du classeur et utilisez le menu Extraire dans une fenêtre indépendante - Copie.

Insérez ensuite cette fenêtre comme objet du classeur, et renommez-la Budget-avec-Moyennes. Insérez à cette feuille six lignes supplémentaires, qui serviront à accueillir les moyennes par sexe et par zone géographique.

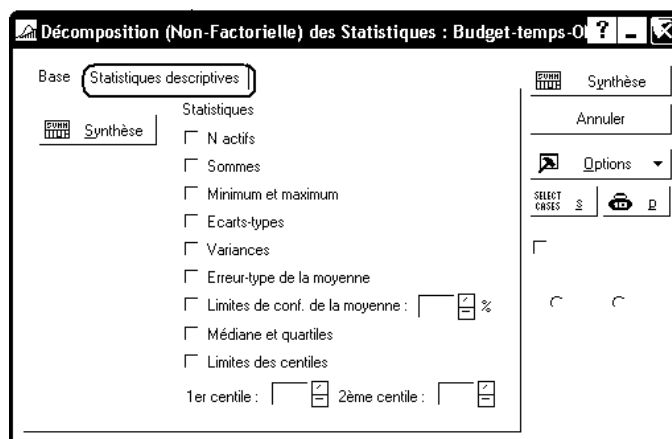
## *Calcul des moyennes d'une variable, selon les groupes définis par la variable catégorisée SEX*

Utilisez le menu Statistiques Elémentaires - Statistiques décomposées.

Sous l'onglet "Base", indiquez les 10 premières variables comme variables dépendantes, et SEX variable de classement :



Sous l'onglet "Statistiques descriptives", dé-selectionnez l'ensemble des boîtes à cocher :



Copiez ensuite les deux lignes de moyennes obtenues et collez-les dans la feuille Budget-avec-moyennes, comme observations 29 et 30.

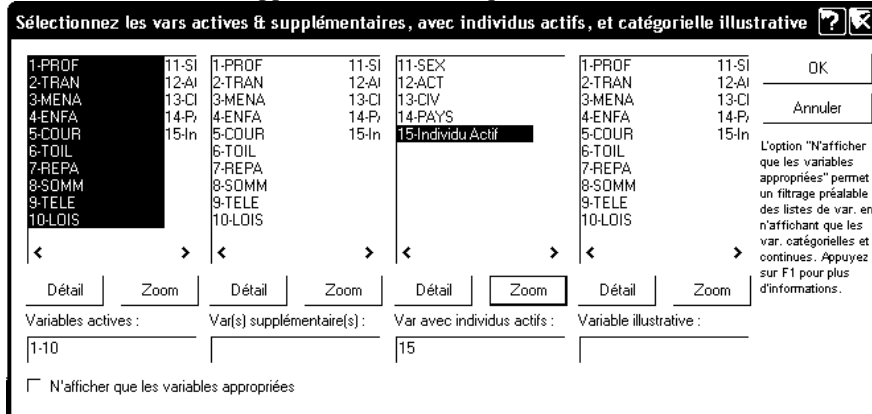
Attribuez à ces deux lignes les noms d'observations : Hommes et Femmes.

Procédez de même pour les moyennes par zone géographique. On obtient, dans l'ordre, les USA, l'Ouest, la Yougoslavie et l'Est.

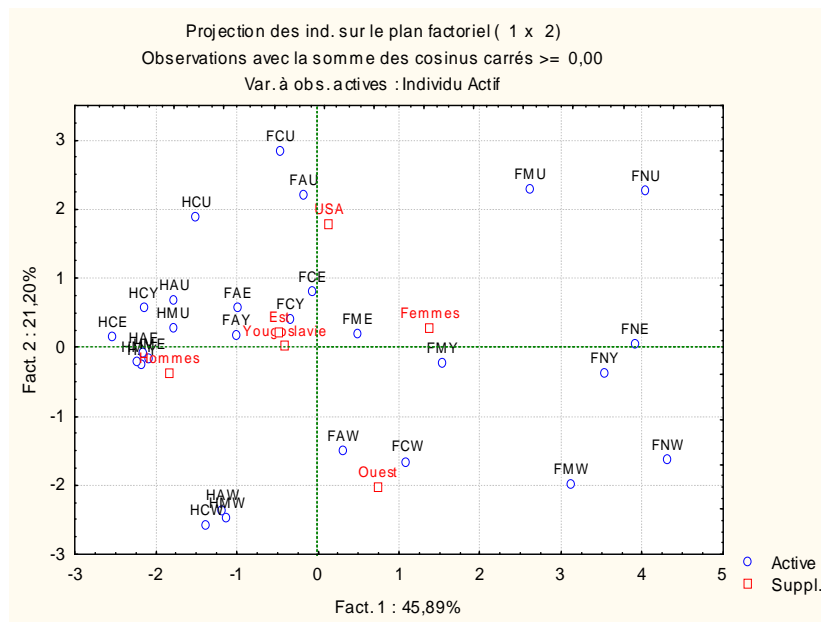
### **1.3.4.2 ACP avec les moyennes par sexe et par zone géographique comme individus supplémentaires**

Introduisez dans la feuille de données "Budget-avec-moyennes" une variable supplémentaire : "Individus actifs", valant 1 sur les 28 premières observations, et 0 sur les 6 moyennes qui suivent.

Rendez active cette feuille de données et refaites une ACP en déclarant en remplissant le premier dialogue comme suit :



Vous pouvez ainsi obtenir des résultats tels que le suivant :



### 1.3.5 Calculer les données centrées réduites

On sait que l'ACP normée travaille sur les données centrées réduites dérivées des données de base. Les dialogues du module "Techniques Exploratoires Multivariées" ne fournissent pas ces données. On peut cependant les obtenir de la façon suivante :

Faites une nouvelle copie de la feuille de données "Budget-temps-ONU" et réinsérez-la dans le classeur. Renommez-la Budget-centre-reduit

Affichez cette feuille et utilisez le menu Données - Centrer-réduire... pour remplacer les 10 premières variables par les variables centrées réduites associées.

## 1.4 Interpréter les résultats d'une ACP

### 1.4.1 Examen des valeurs propres. Choix du nombre d'axes

On examine les résultats relatifs aux valeurs propres. Plusieurs critères peuvent nous guider :

- "méthode du coude" on examine la courbe de décroissance des valeurs propres pour déterminer les points où la pente diminue de façon brutale ; seuls les axes qui précèdent ce changement de pente seront retenus.
- si l'analyse porte sur  $p$  variables et  $n > p$  individus, la variation totale est répartie sur  $p$  axes. On peut alors choisir de conserver les axes dont la contribution relative est supérieure à  $\frac{100\%}{p}$ , ce qui revient, pour une ACP normée, à conserver les valeurs propres supérieures à 1.

## 1.4.2 Interpréter les résultats relatifs aux 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.

### 1.4.2.1 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 supplémentaire.

### 1.4.2.2 Projections des individus dans un plan factoriel

Même s'il s'agit du plan (CP1, CP2), les proximités entre individus doivent être interprétées avec prudence : deux points proches l'un de l'autre sur le graphique peuvent correspondre 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.

## 1.4.3 Interpréter les résultats relatifs aux variables

### 1.4.3.1 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.

### 1.4.3.2 Analyse des projections des variables sur les plans factoriels

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.
- 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. La première variable factorielle est alors appelée "*effet taille*".

### 1.4.4 Quelques règles d'interprétation plus générales

Les commentaires qui suivent proviennent, pour l'essentiel, de l'ouvrage de W. Doise et al. cité en bibliographie.

La technique en composantes principales reproduit avec parcimonie la variation totale d'un grand nombre de variables (pour fixer les idées, dans les cas les plus courants: de 10 à 40) en un nombre sensiblement plus restreint de dimensions (généralement: de 2 à 6). L'échantillon des individus doit être au moins aussi important que le nombre de variables, mais si possible de quatre à cinq fois plus important.

L'analyse implique nécessairement une certaine perte d'informations par rapport aux réponses des individus. Elle fournit en contrepartie une vision bien structurée et immédiatement accessible de la manière dont les variables covarient, s'opposent, ou sont entre elles indépendantes.

La saturation de chaque variable sur chaque dimension indique la contribution de la variable à la dimension en question. Les saturations sont d'autant plus élevées que les variables correspondantes contribuent à donner un sens à la dimension. Le carré d'une saturation fournit la proportion de variance commune de la variable correspondante qui est expliquée par la dimension (ainsi, une saturation de 0.80 indique que 64 % de la variation de la variable est expliquée par la dimension). On ne considère généralement, aux fins de l'interprétation des dimensions, que les saturations atteignant la valeur de  $\pm 1-0.30$  (ce qui correspond approximativement à 10 % de variance expliquée).

Le signe de la saturation est un élément important, tout comme il l'est dans l'examen des corrélations entre deux variables. Deux variables ayant des saturations de même signe (positif ou négatif) sur une dimension, covarient sur cette dimension. Si les saturations ont des signes opposés, elles contribuent de manière opposée à la signification de la dimension.



On distingue habituellement trois types de dimensions (ou facteurs, ceci s'appliquant aussi bien à la technique en facteurs communs). La première dimension décrit la direction principale du faisceau de corrélations. Cette dimension est le plus souvent un facteur général, sur lequel toutes les variables ont des saturations positives et relativement élevées. Elle décrit donc une source de variation traversant l'ensemble de la population analysée: la dimension est présente chez tous les individus mais, fait important, à des degrés différents.

Les dimensions successives seront soit des dimensions de groupes, soit spécifiques. Les dimensions de groupes sont constituées par deux ou plus de deux variables qui covarient sur une dimension. Lorsque des signes positifs et négatifs sont présents sur la même dimension, on parle de facteurs de groupe bipolaires (par opposition à unipolaires).

Enfin, les facteurs spécifiques sont ceux qui ne comportent que des saturations élevées pour une variable à la fois. Habituellement, l'utilisateur arrête l'analyse avant l'apparition de telles dimensions.

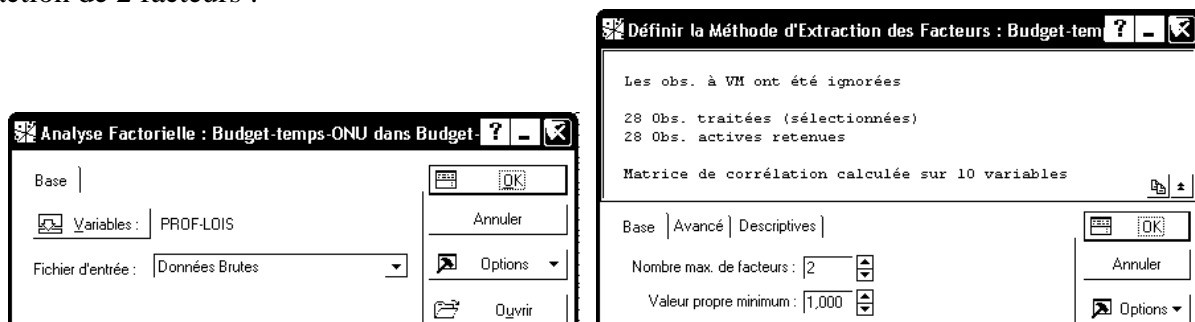
## 1.5 Autres méthodes produisant des résultats analogues à ceux de l'ACP

### 1.5.1 L'analyse factorielle classique

Dans la méthode de base, le principe est le même que pour l'ACP. Mais, après avoir fixé le nombre d'axes principaux à retenir, on peut faire une rotation des axes de manière à augmenter les corrélations entre les nouveaux "facteurs" et certaines variables de départ. Dans Statistica, cette méthode est disponible sous le menu Statistiques - Techniques exploratoires Multivariées - Analyse Factorielle.

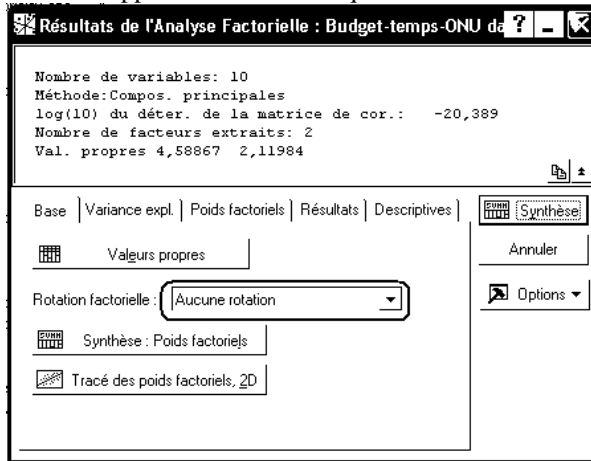
Rendez active la feuille de données Budget-Temps-ONU et utilisez le menu Statistiques - Techniques exploratoires Multivariées - Analyse Factorielle.

Sélectionnez comme précédemment les 10 premières variables comme variables actives, et demandez l'extraction de 2 facteurs :



Laissez les choix par défaut de Statistica pour les autres paramètres, tels que la méthode d'extraction (onglet "Avancé").

On constate que, lorsque liste à choix "Rotation factorielle" est positionnée sur "Aucune rotation", les résultats de la méthode (valeurs propres, poids factoriels, tracé des poids factoriels) sont identiques à ceux de l'ACP sur les deux premiers axes.

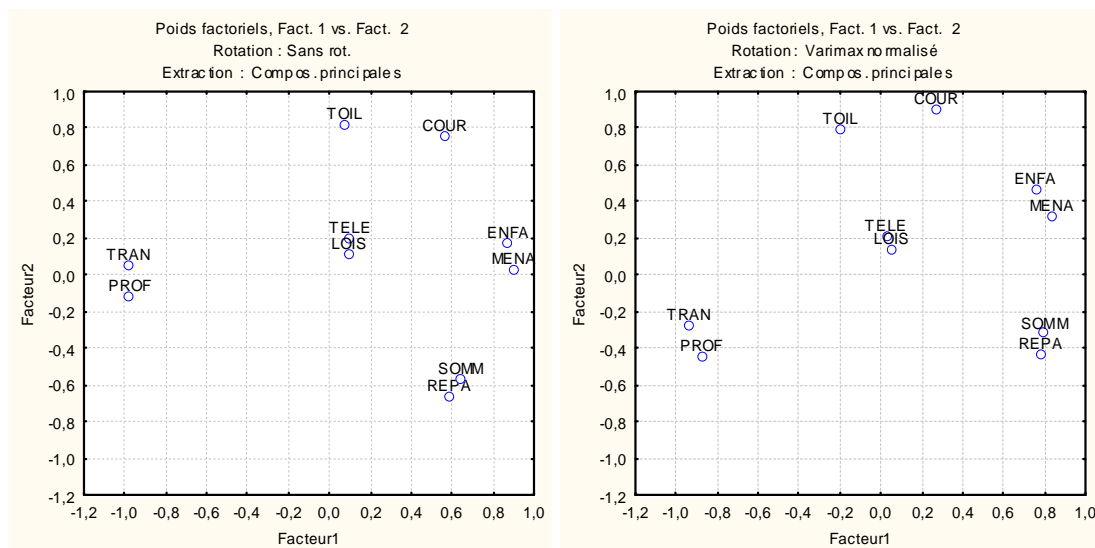


La rotation varimax produit très peu d'effet. En revanche, si l'on choisit la méthode "varimax normalisé", les facteurs sont légèrement modifiés :

- La variance expliquée par le premier plan factoriel est la même que précédemment, c'est-à-dire 6,71 ;
- Celle qui est expliquée par le premier facteur est 4,30 au lieu de 4,59 ;
- Celle qui est expliquée par le second facteur est 2,40 au lieu de 2,11.

On obtient ainsi un premier facteur mieux corrélé avec REPA et SOMM, pendant que le second facteur est mieux corrélé avec COUR.

Sur le schéma ci-dessous, on peut montrer que le graphique de droite est obtenu à partir du graphique de gauche en appliquant une rotation centrée à l'origine, et d'angle 20°, dans le sens inverse des aiguilles d'une montre.



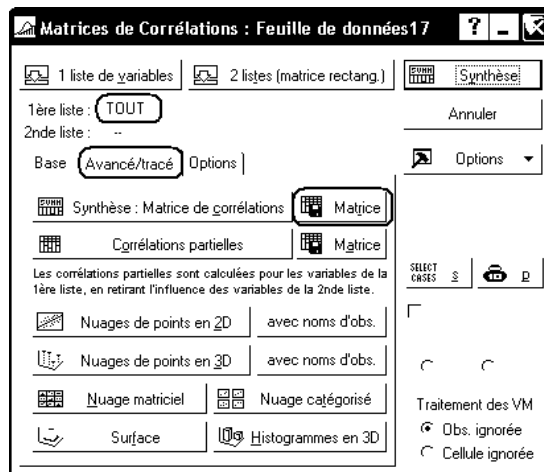
### 1.5.2 L'analyse de proximité

L'analyse des proximités (ou positionnement multidimensionnel, en anglais : multidimensional scaling ou MDS) vise à représenter les proximités entre n objets à l'aide d'un nuage de points d'un espace de dimension assez faible (en général 2 ou 3). Lorsque les objets sont des variables statistiques, dont les proximités mutuelles sont données par la matrice des corrélations, le MDS métrique produit des résultats analogues à ceux de l'ACP.

Le menu Statistiques - Méthodes Exploratoires Multivariées - Analyse de Proximité de Statistica exécute une analyse non métrique, généralisation de la méthode au cas où les proximités entre objets sont calculées à partir de variables ordinales, pas véritablement numériques.

L'analyse de proximité utilise le tableau des corrélations comme données d'entrée. Mais celui-ci doit être un objet de type "matrice" au sens de Statistica.

Utilisez le menu Statistiques - Statistiques Élémentaires - Matrice des Corrélations, sélectionnez les 10 variables PROF, ..., LOIS et, sous l'onglet "Avancé", cliquez sur le bouton "Matrice".

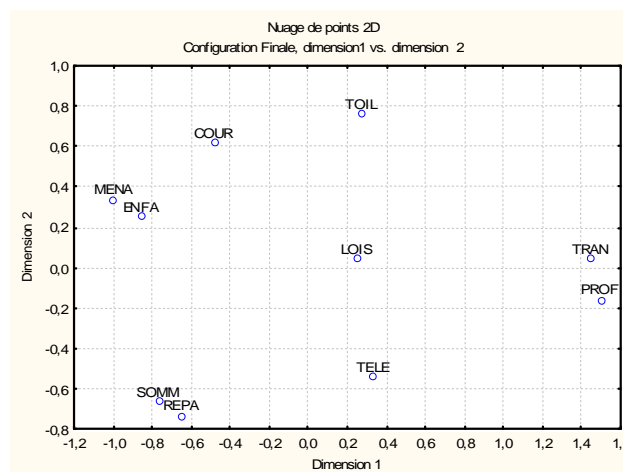


Statistica écrit les résultats dans une nouvelle feuille de données, de type "matrice". Notez la présence des 4 dernières lignes du fichier, avec leur structure particulière. Si on enregistre cette feuille, Statistica lui donne l'extension .smx au lieu de l'extension classique .sta des feuilles de calcul. Notez que l'on peut aussi ajouter cette feuille au classeur en cours.

Veillez à ce que la feuille de données active soit celle contenant la matrice des corrélations et utilisez le menu Statistiques - Méthodes Exploratoires Multivariées - Analyse de proximité.

On garde le nombre de dimensions par défaut, c'est-à-dire 2.

Cliquez ensuite deux fois sur le bouton OK, puis affichez le "graphique de la configuration finale". On obtient le schéma suivant, qui est assez proche des représentations obtenues précédemment :



## 1.6 Exemples et exercices

### 1.6.1 Le cas "Basket"

On s'intéresse au profil de 18 basketteurs de 14 ans. Ils ont passé un certain nombre de tests relatifs aux qualités physiques requises pour la pratique de cette discipline.

TAI : taille en cm

VIT : vitesse sur 30 m (en secondes)

DET : détente verticale en cm : sauter le plus haut possible, le bras tendu

PAS : passe en mètres : lancer un ballon de basket le plus loin possible

LEG : endurance, en litres/mn/kg : test Le Luc Léger

STA : adresse statique, en nombre de paniers.

La variable VIT est codée systématiquement avec un signe "-" afin que, comme pour les autres variables, une valeur élevée traduise une bonne performance.

Source : Institut National du Sport et de l'Education Physique (I.N.S.E.P.) - Extrait d'un fichier traité par Marion Wolf pour la Fédération Française de Basket-Ball

SUJET	TAI	VIT	DET	PAS	LEG	STA
I1	170	-4	77	15	63,7	17
I2	181	-5	49	15	45,1	11
I3	192	-5,1	50	16,1	46,2	15
I4	173	-4,1	70	15,5	63,5	17
I5	170	-4	70	12,5	64,3	19
I6	175	-4,3	72	12,4	61,6	18
I7	170	-4,4	70	12	65,6	10
I8	168	-4	76	11	64	7
I9	166	-4	76	10	64	8
I10	181	-5,3	48	15,2	50,2	10
I11	186	-4,7	55	15,5	51	14
I12	180	-4,6	50	12	51,7	16
I13	185	-4,8	50	12,8	49,7	19
I14	192	-5	48	11,5	45,6	17
I15	191	-4,9	45	11,3	45,9	16
I16	192	-4,9	43	10,5	48,9	18
I17	192	-5,1	50	10,5	45	16
I18	195	-5,3	50	15,1	47,1	19

On réalise une ACP normée sur ces données. Les résultats fournis par Statistica (ou Excel) sont les suivants :

Données centrées réduites et inerties relatives des individus (Excel)

SUJET	TAI	VIT	DET	PAS	LEG	STA	Inertie
I1	-1,1447	1,3863	1,5461	0,9983	1,2003	0,5695	7,76%
I2	-0,0058	-0,7836	-0,7661	0,9983	-1,1159	-1,0076	4,13%
I3	1,1332	-1,0006	-0,6836	1,5458	-0,9789	0,0438	5,65%
I4	-0,8341	1,1694	0,9680	1,2472	1,1754	0,5695	5,80%
I5	-1,1447	1,3863	0,9680	-0,2461	1,2750	1,0953	6,53%
I6	-0,6270	0,7354	1,1332	-0,2959	0,9388	0,8324	3,59%
I7	-1,1447	0,5184	0,9680	-0,4950	1,4369	-1,2705	5,96%
I8	-1,3518	1,3863	1,4635	-0,9928	1,2377	-2,0591	11,71%
I9	-1,5589	1,3863	1,4635	-1,4905	1,2377	-1,7962	12,48%
I10	-0,0058	-1,4346	-0,8487	1,0978	-0,4808	-1,2705	5,40%
I11	0,5120	-0,1326	-0,2707	1,2472	-0,3812	-0,2191	1,95%
I12	-0,1093	0,0844	-0,6836	-0,4950	-0,2940	0,3067	0,84%
I13	0,4084	-0,3496	-0,6836	-0,0968	-0,5431	1,0953	2,09%
I14	1,1332	-0,7836	-0,8487	-0,7439	-1,0536	0,5695	4,27%
I15	1,0297	-0,5666	-1,0965	-0,8434	-1,0163	0,3067	4,09%
I16	1,1332	-0,5666	-1,2616	-1,2416	-0,6427	0,8324	5,41%
I17	1,1332	-1,0006	-0,6836	-1,2416	-1,1284	0,3067	5,24%
I18	1,4438	-1,4346	-0,6836	1,0481	-0,8669	1,0953	7,09%

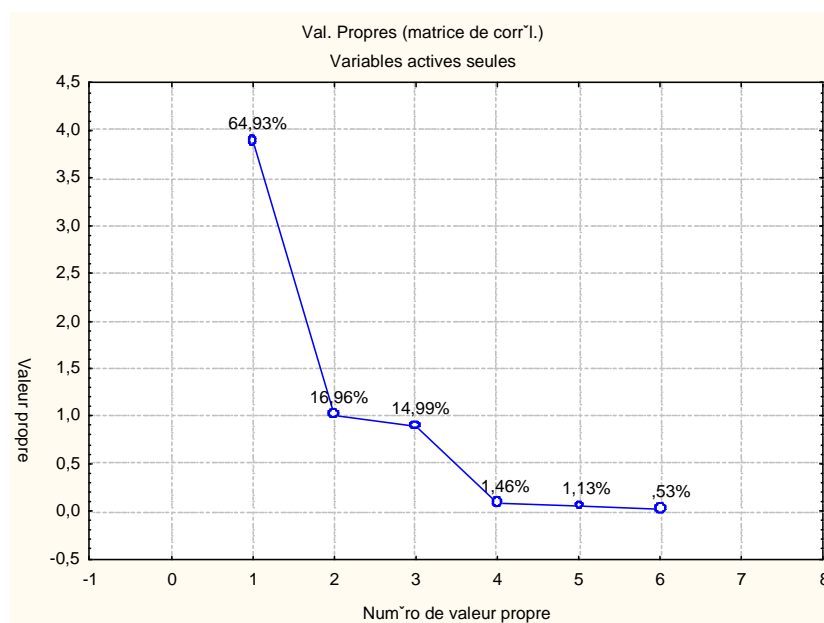
Corrélations (Basket.sta)

	TAI	VIT	DET	PAS	LEG	STA
TAI	1,0000	-0,8833	-0,8974	0,1054	-0,9241	0,4630
VIT	-0,8833	1,0000	0,9108	-0,2217	0,9206	-0,1748
DET	-0,8974	0,9108	1,0000	-0,0760	0,9498	-0,2969
PAS	0,1054	-0,2217	-0,0760	1,0000	-0,1230	0,1278
LEG	-0,9241	0,9206	0,9498	-0,1230	1,0000	-0,2621
STA	0,4630	-0,1748	-0,2969	0,1278	-0,2621	1,0000

Val. Propres (matrice de corrél.) & stat. associées (Basket.sta)

Variables actives seules

	Val. propr	% Total	Cumul	Cumul
		variance	Val. propr	%
1	3,8960	64,9331	3,8960	64,9331
2	1,0174	16,9573	4,9134	81,8904
3	0,8992	14,9862	5,8126	96,8766
4	0,0877	1,4613	5,9003	98,3378
5	0,0678	1,1304	5,9681	99,4682
6	0,0319	0,5318	6,0000	100,0000



Coordonnées factorielles des ind., basées sur les corrélations (Basket.sta)

	Fact. 1	Fact. 2	Fact. 3
I1	-2,3534	-1,6298	0,3578
I2	1,1832	-0,0810	-1,6450
I3	2,0077	-1,0247	-0,9373
I4	-1,7791	-1,7315	0,1988
I5	-2,1211	-0,7309	1,3785
I6	-1,5300	-0,4875	1,0646
I7	-2,2965	0,6306	-0,6465
I8	-3,1679	1,3129	-0,8662
I9	-3,2637	1,6115	-0,4352
I10	1,1787	-0,0647	-2,0121
I11	0,7097	-0,8580	-0,7993
I12	0,3985	0,4097	0,4446
I13	1,1835	-0,2256	0,8347
I14	1,9067	0,7151	0,6446
I15	1,7906	0,9374	0,5114

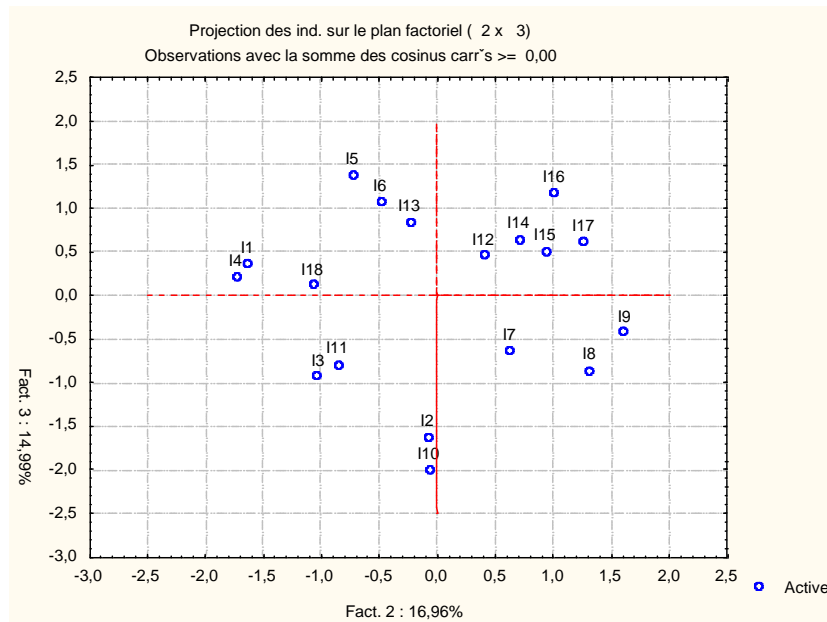
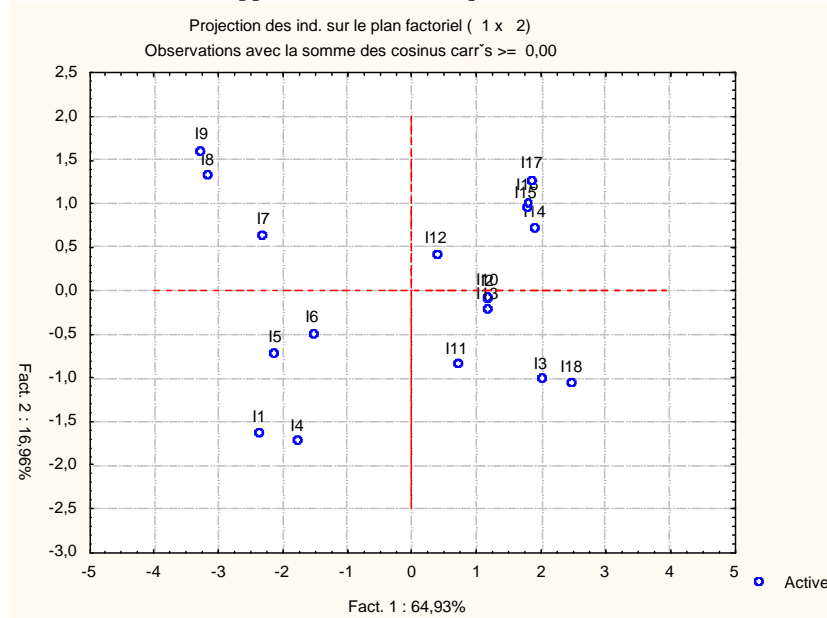
I16	1,8079	1,0146	1,1800
I17	1,8643	1,2658	0,6147
I18	2,4808	-1,0639	0,1118

Contributions des ind., basées sur les corrélations (Basket.sta)

	Fact. 1	Fact. 2	Fact. 3
I1	7,90	14,50	0,79
I2	2,00	0,04	16,72
I3	5,75	5,73	5,43
I4	4,51	16,37	0,24
I5	6,42	2,92	11,74
I6	3,34	1,30	7,00
I7	7,52	2,17	2,58
I8	14,31	9,41	4,64
I9	15,19	14,18	1,17
I10	1,98	0,02	25,01
I11	0,72	4,02	3,95
I12	0,23	0,92	1,22
I13	2,00	0,28	4,30
I14	5,18	2,79	2,57
I15	4,57	4,80	1,62
I16	4,66	5,62	8,60
I17	4,96	8,75	2,33
I18	8,78	6,18	0,08

Cosinus carrés, basées sur les corrélations (Basket.sta)

	Fact. 1	Fact. 2	Fact. 3
I1	0,6606	0,3168	0,0153
I2	0,3140	0,0015	0,6070
I3	0,6605	0,1721	0,1440
I4	0,5055	0,4788	0,0063
I5	0,6377	0,0757	0,2693
I6	0,6033	0,0613	0,2922
I7	0,8189	0,0618	0,0649
I8	0,7934	0,1363	0,0593
I9	0,7905	0,1927	0,0141
I10	0,2384	0,0007	0,6945
I11	0,2396	0,3503	0,3040
I12	0,1742	0,1841	0,2167
I13	0,6197	0,0225	0,3083
I14	0,7892	0,1110	0,0902
I15	0,7251	0,1987	0,0591
I16	0,5592	0,1761	0,2383
I17	0,6139	0,2830	0,0667
I18	0,8035	0,1478	0,0016



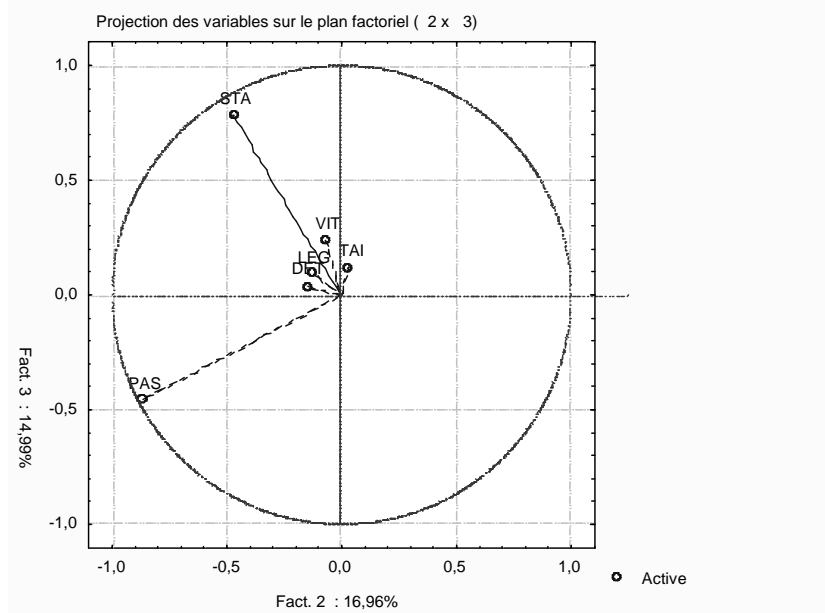
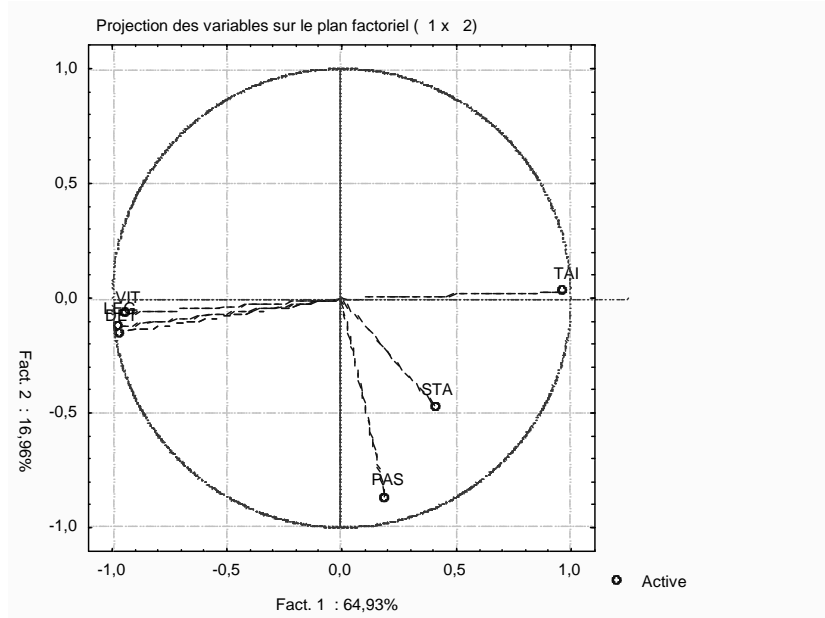
Corrél. facteur-var. (poids fact.), basées sur corrélations (Basket.sta)

	Fact. 1	Fact. 2	Fact. 3
TAI	0,9676	0,0292	0,1209
VIT	-0,9450	-0,0701	0,2349
DET	-0,9617	-0,1484	0,0359
PAS	0,1919	-0,8667	-0,4593
LEG	-0,9695	-0,1248	0,0922
STA	0,4065	-0,4721	0,7801

Contributions des var., basées sur les corrélations (Basket.sta)

	Fact. 1	Fact. 2	Fact. 3
TAI	0,2403	0,0008	0,0162
VIT	0,2292	0,0048	0,0614
DET	0,2374	0,0216	0,0014
PAS	0,0094	0,7383	0,2347
LEG	0,2412	0,0153	0,0095
STA	0,0424	0,2191	0,6768

	Avec 1 facteur	Avec 2 facteurs	Avec 3 facteurs
TAI	0,9362	0,9370	0,9516
VIT	0,8930	0,8979	0,9531
DET	0,9249	0,9469	0,9482
PAS	0,0368	0,7880	0,9990
LEG	0,9399	0,9555	0,9640
STA	0,1652	0,3882	0,9968



Vecteurs propres de la matrice de corrélation (Basket.sta)  
 Variables actives seules

	Fact. 1	Fact. 2	Fact. 3	Fact. 4	Fact. 5	Fact. 6
TAI	0,4902	0,0290	0,1275	0,5993	-0,3593	0,5044
VIT	-0,4788	-0,0695	0,2477	-0,3305	-0,7150	0,2900
DET	-0,4872	-0,1471	0,0379	0,6993	-0,1617	-0,4737
PAS	0,0972	-0,8592	-0,4844	-0,0556	-0,0922	0,0776
LEG	-0,4912	-0,1237	0,0972	0,1821	0,5473	0,6335
STA	0,2059	-0,4681	0,8227	-0,0798	0,1594	-0,1728



- 1) Examiner la matrice des corrélations entre les variables. Faites un commentaire.
- 2) Examen du nuage de points : quels sont les sujets dont l'inertie est la plus forte ? Quels sont ceux dont l'inertie est la plus faible ?
- 3) On choisit de ne conserver que 3 composantes principales. Justifier ce choix.
- 4) a) Quels sont les sujets qui contribuent le plus fortement à la formation du premier axe principal ? Indiquez également si leur contribution intervient dans la partie positive ou dans la partie négative de l'axe.
- b) Citez deux sujets qui sont bien représentés par leur première composante principale. Quels sont les deux sujets les plus mal représentés par cette composante ?
- 5) Analysez, de la même façon, le deuxième, puis le troisième axe principal.
- 6) a) Quelles sont les variables les plus fortement corrélées avec la première composante principale. Interprétez cette composante à l'aide de ces variables.
- b) De même, donnez une interprétation des deuxième et troisième composantes principales.

### 1.6.2 Le cas Psychométrie

Pour 20 élèves (sujets s1 à s20), on a relevé les notes obtenues à cinq épreuves individuelles : Combinatoire (Comb), Probabilités (Prob), Logique (Logi), notées de 0 à 10, QI verbal (QI, notes de 85 à 125) et Mathématiques (Math), notée de 0 à 20.

Pour chaque sujet, on dispose de deux informations : Pédagogie avec deux modalités p1 (moderne) et p2 (traditionnelle), Milieu avec deux modalités m1 (favorisé) et m2 (défavorisé).

	Comb	Prob	Logi	QI	Math	Peda	Milieu
s1	3,9	4,1	6	99	8	p1	m1
s2	5	5	5,2	122	10	p1	m1
s3	5,3	8,5	8,6	108	14	p1	m1
s4	8,3	6,2	7,2	125	18	p1	m1
s5	5,5	6	6,9	108	5	p1	m2
s6	6,6	7,7	5,8	113	7	p1	m2
s7	5,5	3	5,8	94	10	p1	m2
s8	2,2	4,5	3,3	85	9	p1	m2
s9	5,3	4,5	8,3	112	10	p1	m2
s10	5,3	6,4	6,5	125	12	p1	m2
s11	4,6	4,6	5,2	108	14	p1	m2
s12	3,7	4,1	7,2	91	15	p1	m2
s13	4,1	6,7	7,1	91	6	p2	m1
s14	2,7	4,5	3	109	9	p2	m1
s15	6,8	4,5	7,1	125	12	p2	m1
s16	2,7	3,7	6,9	94	13	p2	m1
s17	5,4	8,9	7,3	120	15	p2	m1
s18	6,2	4,7	4,4	112	7	p2	m2
s19	2,5	4,7	7,2	106	11	p2	m2
s20	2,4	4,4	5,2	91	12	p2	m2

1) Saisir les données dans Statistica sous une forme convenant à la réalisation d'une analyse en composantes principales..

2) Réalisez une analyse en composantes principales normée, sur les 4 variables Comb, Prob, Logi et Math.

Déterminez notamment la matrice des corrélations, les valeurs propres, les scores, contributions et qualités des individus sur les deux premières composantes, les coefficients des variables et les saturations, contributions et qualité des variables (2 premières composantes). Réalisez le graphique des individus et celui des variables par rapport aux deux premiers axes principaux.

3) Examiner et commenter le tableau des corrélations.

4) Les variables Comb et Proba apparaissent proches sur le graphique. Quel est pourtant leur coefficient de corrélation ? Comment peut-on l'expliquer ?

5) Les points s8 et s14 apparaissent très proches sur le graphique. Est-ce le cas dans la réalité ? Même question pour s9 et s15.

6) Comment les variables contribuent-elles à la formation de l'axe CP1 ? Comment cet axe classe-t-il les individus ?

7) Comment les variables contribuent-elles à la formation de l'axe CP2 ? Décrire cet axe en termes d'oppositions entre variables, en termes d'oppositions entre individus.

8) a) Réalisez le graphique des individus en étiquetant les points à l'aide des modalités de la variable Pédagogie, puis en étiquetant les points à l'aide des modalités de la variable Milieu. Interprétez les graphiques obtenus.

b) Calculez les moyennes des variables observées dans les 4 groupes définis par les combinaisons de modalités des variables Pédagogie et Milieu. Ajoutez ces moyennes comme observations supplémentaires dans la feuille de données Statistica, puis reprenez l'ACP en déclarant ces valeurs comme individus supplémentaires. Réalisez un graphique des individus affichant ces individus supplémentaires.

9) L'étude limitée aux deux premières composantes vous paraît-elle suffisante ? Comment souhaiteriez-vous poursuivre cette étude ?

### 1.6.3 Le cas "Budget-temps Multimédia"

Le CESP (Centre d'Étude des Supports de Publicité) a relevé, dans son Enquête Budget-temps Multimédia de 1991/1992 auprès de 17 665 personnes, des descripteurs de fréquentation de divers médias (radio, télévision, presse) et des temps d'activités quotidiennes (cf. Boeswillwald, 1992). Ont été également relevées de nombreuses caractéristiques socioéconomiques, parmi lesquelles l'âge, le sexe, l'activité, le niveau d'éducation, et le lieu de résidence de ces personnes, ce qui a conduit à créer 96 catégories en croisant ces divers critères.

Nous nous intéressons seulement ici à la sous-population des hommes actifs, soit 27 groupes qui seront, pour cet exemple, les "individus". On cherche à connaître les associations entre les temps consacrés à différentes activités par les "individus" observés et à étudier les liens entre ces familles d'activités et les caractéristiques de base des individus.

L'étude originale se proposait d'étudier le lien entre les activités quotidiennes et la fréquentation de divers médias (presse, radio, télévision, cinéma). Pour ce faire, elle faisait intervenir les caractéristiques socio-

économiques (variables nominales) et les habitudes de fréquentation des médias (variables numériques continues) en tant que variables supplémentaires. Mais ces données ne sont pas présentes ici.

L'ensemble des données se trouve dans la feuille de données Statistica Budget-temps-multimedia.sta du serveur de TD. Ci-dessous figurent quelques indications pour la lecture de ce tableau :

Les 27 "individus" (qui sont en réalité dans le cadre de cet exemple des groupes d'individus) sont repérés par un identificateur en 4 caractères:

- le 1er caractère est l'âge du groupe (1=jeune, 2=moyen, 3=âgé)
- le 2ème caractère est ici toujours égal à 1 (car il s'agit ici d'une sélection d'hommes actifs)
- le 3ème est le niveau d'éducation (1=primaire, 2=secondaire, 3=supérieur)
- le 4ème est le type d'agglomération ( 1=communes rurales; 2=villes moyennes; 3=villes importantes; 4=agglomération parisienne; 5,6,7 = groupes mixtes).

La signification des 16 variables actives est la suivante :

Somm.....	Sommeil
Repo .....	Repos
Reps .....	Repas chez soi
Repr .....	Repas restaurant
Trar .....	Travail rémunéré
Ména .....	Ménage
Visi .....	Visite à amis
Jard .....	Jardinage, Bricolage
Lois .....	Loisirs extérieur
Disq .....	Disque cassette
Lect .....	Lecture livre
Cour .....	Courses démarches
Prom .....	Promenade
A pi .....	Déplacement à pied
Voit .....	Déplacement en Voiture
Fréq .....	Fréquentation Média

On lit par exemple sur la première ligne du tableau que le groupe '1111' (jeunes, actifs, peu instruits, ruraux) consacre en moyenne par jour 463,8 minutes au "sommeil", 23,8 minutes à des activités regroupées sous la rubrique "repos", 107,3 minutes pour les "repas chez soi", etc.

Analysez ces données à l'aide d'une ACP, en suivant la méthode d'interprétation qui a été indiquée en cours.

N.B. Bien que la décroissance des valeurs propres soit relativement progressive, on étudiera essentiellement les deux premières composantes principales.

Créez des variables nominales supplémentaires Age, Niveau d'éducation, Catégorie d'agglomération et, pour chacune d'elle, réalisez un graphe de projection des individus en utilisant comme étiquettes les modalités de la variable. Essayez d'interpréter les graphes ainsi obtenus.

### 1.6.4 Le cas Sleep

Références . [Crucianu] p. 19, qui fait lui-même référence à un article publié dans Science en 1976 par T. Allison et D. Cicchetti et à des données accessibles à l'adresse <http://www.stat.ucl.ac.be/ISdidactique/Rhelp/library/psy/html/sleep.html>.

L'exemple qui suit est extrait d'une étude sur les relations qu'entretient le sommeil des mammifères avec différents facteurs morphologiques et écologiques.

L'ensemble étudié est constitué des représentants typiques de 62 espèces de mammifères variés, de la taupe à l'éléphant, décrits par 10 variables numériques. Chaque individu est d'abord caractérisé par des mesures concernant le poids du corps en kilogrammes, le poids du cerveau en grammes, le nombre d'heures de sommeil sans rêve par jour, le nombre d'heures de sommeil avec rêves, la somme des deux types de sommeil, la durée de vie maximale en années, et la durée de la période de gestation en jours.

Trois indices ont été calculés :

- Un indice de prédation : 1= faible risque d'être chassé par un prédateur à 5 = fort risque.
- Un indice d'exposition pendant le sommeil : 1= animal dormant dans une tanière très protégée, 5 = animal très exposé aux prédateurs pendant son sommeil
- Un indice de dangerosité, obtenu à partir des indices précédents et d'autres informations, décrivant dans quelle mesure le mammifère peut être mis en danger par d'autres animaux.

Ouvrez la feuille de données sleep.sta et observez les données saisies.

Traitez ces données à l'aide d'une ACP normée et interprétez les résultats, en utilisant essentiellement les résultats relatifs aux variables, et les deux premières dimensions factorielles.

Vous devriez parvenir aux résultats suivants :

On observe que toutes les variables sont relativement bien représentées par les 2 premiers axes factoriels. On observe également qu'aucune variable n'a un rôle dominant dans l'orientation des axes factoriels. Trois groupes de variables apparaissent : un premier groupe concernant directement le sommeil, un deuxième groupe de variables liées à l'évaluation du danger et un troisième groupe relatif aux caractéristiques physiques.

Le premier axe factoriel oppose le groupe "sommeil" aux deux autres groupes : les temps de sommeil les plus longs sont observés chez les mammifères qui sont le moins en danger.

Le deuxième axe factoriel montre une autre opposition, moins forte, entre le groupe "danger" et le groupe "caractéristiques physiques" : il existe, globalement, une corrélation négative entre la taille du mammifère et le danger encouru.

L'élément le plus évident dans le diagramme de projection des individus est la position excentrée des individus 1 et 5 (éléphants d'Afrique et d'Asie). Pour l'essentiel, l'examen du diagramme des individus confirme l'analyse proposée à partir de l'examen des variables.

Reprenez alors l'étude en plaçant déclarant ces deux individus comme individus inactifs.

### 1.6.5 Le cas disciplines-différentiateurs

On a demandé à 11 étudiants ce qu'ils pensaient de 15 disciplines scientifiques au moyen de 6 paires d'adjectifs antonymes. Les 11 étudiants appartiennent au DEA de didactique des disciplines scientifiques et sont ou seront des enseignants scientifiques. Les 15 disciplines sont :

1-algèbre	6-éthologie	11-physique nucléaire
2-astrologie	7-informatique	12-psychologie
3-biologie moléculaire	8-linguistique	13-science
4-didactique	9-médecine	14-sociologie
5-écologie	10-neurologie	15-technologie

Les 6 paires d'adjectifs utilisés sont appelées des différenciateurs sémantiques d'Osgood. Ce sont :

I	précis (1)-imprécis (5)
II	dur (1)-mou (5)

- III    subjectif (1)-objectif (5)
- IV     faux (1)-vrai (5)
- V      faible (1)-fort (5)
- VI     fantaisiste (1)-sérieux (5)

Pour un étudiant donné, une discipline donnée et une paire d'adjectifs donnée l'opinion exprimée sur la discipline par l'étudiant au moyen du différentiateur est une note qui peut prendre 5 valeurs :

- 1— association forte avec le premier terme du différentiateur
- 2— préférence pour le premier terme du différentiateur
- 3— absence d'opinion
- 4— préférence pour le second terme du différentiateur
- 5— association forte avec le second terme du différentiateur

Le tableau de données du fichier Disciplines-Differentiateurs.stw indique le score moyen obtenu par chaque discipline sur chaque paire d'adjectifs.

N.B. L'étiquette retenue pour désigner chaque couple est le second terme du différentiateur.

1) Traitez ces données par une analyse en composantes principales normée, *en plaçant l'astrologie comme individu supplémentaire*.

Calculez notamment à l'aide de Statistica le tableau des corrélations, celui des valeurs propres, les scores, contributions et qualités de représentation des individus et les saturations, contributions et qualités de représentation des variables.

Réalisez la représentation des individus et celle des variables dans le premier plan factoriel.

2) Etude du tableau des valeurs propres

- a) A quoi correspond la somme des valeurs propres ?
- b) On choisit de n'étudier que les deux premières composantes principales. Justifier ce choix en analysant le tableau des valeurs propres.

3) Etude du tableau des corrélations. Quelles sont les variables le plus fortement corrélées entre elles ? Y a-t-il des variables pratiquement non corrélées ?

4) Etude des qualités de représentation dans le premier plan principal. Quel est l'individu le moins bien représenté par le premier plan principal ? Quel est l'individu le mieux représenté ?

5) Etude du nuage des individus.

- a) Quels sont les individus dont la contribution à la formation de la première composante principale est supérieure à la moyenne ? Pour chacun d'eux, préciser le signe de la coordonnée correspondante. Caractériser cet axe en termes d'opposition entre individus.
- b) Même question pour la deuxième composante principale.

6) Etude du nuage des variables

- a) La représentation graphique des variables montre qu'elles sont toutes très bien représentées dans le plan (CP1, CP2). Justifier cette affirmation.
- b) Quelles sont les deux variables qui sont le plus fortement corrélées à la première variable principale ?
- c) Même question pour la deuxième variable principale.
- d) Deux variables sont pratiquement indépendantes de la 2<sup>e</sup> variable principale. Lesquelles ?
- e) A propos de cet exemple, peut-on parler "d'effet de taille" ?

7) L'individu "Astrologie" a été placé en individu supplémentaire dans l'analyse.

- a) Quel rôle joue un tel individu dans le déroulement des calculs nécessaires à l'exécution de l'ACP ?

- b) Pour quelles raisons a-t-on choisi de placer en individu supplémentaire ?  
 c) Commenter les valeurs numériques obtenues et la position de cet individu sur le graphique.

N.B. Les résultats fournis par cette ACP ne constituent évidemment en aucune façon un jugement de valeur sur les disciplines citées. Les conclusions éventuelles peuvent tout au plus porter sur les opinions des 11 sujets interrogés...

### 1.6.6 Le cas Salaire-genre

Pour un échantillon de 40 sujets, on dispose des informations suivantes :

- Le genre (0 = masculin, 1 = féminin)
- Le nombre d'années d'étude (de 8 à 17), codé ED
- L'âge (de 30 à 63 ans), codé AGE
- Le salaire en Euros (de 837 à 6250 Euros), codé SAL.

Une cinquième variable (Actifs) contient le code 1 pour les 40 observations et le code 0 pour les deux dernières lignes, qui contiennent les moyennes par sexe observées sur les données.

Les données sont rassemblées dans le classeur Statistica Salaire-genre.stw.

1) Traitez ces données en réalisant une ACP normée sur les variables ED, AGE et SAL ; le sexe est utilisé comme variable illustrative, et la variable Actifs est utilisée pour identifier les individus statistiques actifs.

2) Etude du tableau des corrélations

Commenter le tableau des corrélations entre les 3 variables SAL, ED et AGE.

3) Etude du tableau des valeurs propres

a) A quoi correspond la somme des valeurs propres ?

b) On choisit de n'étudier que les deux premières composantes principales. Justifier ce choix en analysant le tableau des valeurs propres.

4) Etude du nuage des variables

a) La représentation graphique des variables montre qu'elles sont toutes très bien représentées dans le plan (CP1, CP2). Justifier cette affirmation.

b) Quelle est la variable le plus fortement corrélée à la première variable principale ?

c) Question analogue pour la deuxième variable principale.

d) L'une des variables est pratiquement indépendante de la 2<sup>e</sup> variable principale. Laquelle ?

5) On a placé les moyennes par genre comme individus supplémentaires dans l'analyse.

a) Quel rôle jouent de tels individus dans le déroulement des calculs nécessaires à l'exécution de l'ACP ?

b) Commenter les valeurs numériques obtenues et la position de ces individus sur le graphique.

N.B. On n'a aucune information sur la population d'où est issu l'échantillon observé ici. On se gardera donc de toute généralisation des résultats observés sur cet exemple.

### 1.6.7 Le cas "Odors"

Une équipe de chercheurs s'est intéressée à l'effet de la culture sur la catégorisation des odeurs<sup>1</sup>. Pour l'expérience décrite ci-dessous, trois groupes de sujets ont été recrutés : 30 étudiants de l'Université de

<sup>1</sup> C. Chrea, D. Valentin, C. Sulmont-Rossé, D. Hoang-Nguyen, H. Abdi, Semantic, Typicality and Odor Representation : A Cross-cultural Study, Chem. Senses V. 30 no. 1, pp 37-49, 2005

C. Chrea, D. Valentin, C. Sulmont-Rossé, D. Hoang-Nguyen, H. Abdi, Odeurs et catégorisation : notre représentation mentale des odeurs est-elle universelle ou dépendante de notre culture ? in Actes de la 3<sup>ème</sup> journée du Sensoriel, Paris, France, 2005.

Bourgogne (Dijon - France), 30 étudiants de l'Université de Dallas (Texas - USA) et 30 étudiants de l'Institut Polytechnique de Danang (Vietnam).

Le matériel était constitué de 40 odorants fournis par une société spécialisée. Chacun de ces odorants était présenté aux sujets, qui devaient évaluer la typicalité de son odeur par rapport à 11 catégories (animal, pâtisserie, sucrerie, nettoyant, cosmétique, fleur, fruit, pharmacie, moisissure, nature, épice).

Les sujets donnaient leurs réponses sur des échelles de Likert en 7 points (1 = non typique, 7 = tout à fait typique).

On construit le protocole des moyennes observées pour le score de typicalité de chaque odorant, pour l'ensemble des participants. Les valeurs observées sont les suivantes :

	animal	pâtisserie	sucrerie	nettoyant	cosmétique	fleur	fruit	pharmacie	moisissure	nature	épice
ambre	1,36	1,84	2,19	2,36	4,04	2,65	1,96	2,66	2,32	2,27	2,54
anis	1,23	2,34	5,20	1,76	1,77	2,22	2,23	2,65	1,51	2,02	3,26
abricot	1,32	2,55	4,70	2,10	3,00	3,41	4,01	2,28	1,30	2,33	1,66
cassis	1,28	2,40	5,49	1,86	2,04	2,73	4,42	2,08	1,52	2,15	1,57
beurre	2,42	2,65	2,01	1,92	1,97	1,77	1,71	1,85	2,39	2,16	1,86
urine de chat	3,59	1,26	1,37	2,16	1,76	1,53	1,42	2,12	3,32	2,45	1,73
cannelle	1,34	3,27	3,59	1,69	2,29	2,23	1,89	2,07	1,49	2,63	4,83
civette	4,52	1,21	1,24	2,09	1,29	1,50	1,29	1,73	3,39	2,80	1,34
clou de girofle	1,62	2,05	1,93	1,91	1,77	2,00	1,62	3,15	2,04	2,23	4,28
biscuits	1,12	4,70	4,81	1,62	3,07	2,43	2,61	1,99	1,34	1,71	2,35
détergent	1,70	1,39	1,42	4,38	2,93	2,03	1,53	2,34	1,92	2,13	1,56
eucalyptus	1,26	1,26	2,69	3,19	1,89	1,71	1,50	5,10	1,68	2,24	1,65
gingembre	1,51	2,16	2,61	2,89	2,11	2,19	2,42	2,37	1,87	2,91	3,41
noisette	2,23	3,27	2,94	1,45	1,67	1,47	1,82	1,68	2,76	2,01	2,44
miel	2,34	2,22	2,59	2,25	1,53	2,35	1,77	2,24	2,22	2,83	2,12
jasmin	1,97	1,57	1,87	2,99	2,93	4,28	1,94	2,29	2,42	2,94	1,71
lavande	1,31	1,33	1,74	4,02	2,39	3,59	1,84	2,39	2,11	3,15	2,07
cuir	2,32	1,49	1,40	3,09	2,06	1,44	1,41	2,71	2,75	2,25	2,17
mangue	1,49	1,73	2,53	2,91	3,09	3,33	3,28	2,31	2,16	2,62	1,91
melon	1,19	2,57	5,30	1,70	2,51	2,87	5,81	1,78	1,28	2,61	1,48
lait	1,48	3,51	3,77	1,74	2,41	2,44	2,72	2,04	1,67	2,32	2,55
anti-mites	1,78	1,57	1,53	4,45	2,25	1,76	1,52	2,64	2,43	1,85	1,48
moisi	2,10	1,17	1,21	2,36	1,26	1,75	1,32	1,96	5,10	4,22	1,74
champignon	2,34	1,17	1,45	2,31	1,44	1,53	1,35	2,21	4,46	3,36	1,67
musc	2,31	1,41	1,73	2,85	3,54	2,56	1,75	2,20	2,00	2,01	1,71
noix de muscade	1,80	1,95	1,79	2,31	1,63	2,10	1,63	2,58	2,94	2,26	3,82
fleur d'oranger	1,40	1,85	1,92	3,46	2,95	3,89	2,14	2,44	1,94	2,68	1,68
arachide	1,73	3,15	2,65	1,90	1,72	1,78	1,95	2,02	2,58	2,39	2,53
pin	1,44	1,47	1,68	3,09	2,89	2,28	1,58	2,81	2,70	2,85	1,98
ananas	1,20	2,58	4,83	2,03	2,26	2,67	5,40	1,85	1,53	2,54	1,75
rose	1,30	1,86	2,66	3,21	3,87	3,82	2,13	1,86	1,49	2,46	1,34
savon	1,21	1,35	1,50	4,70	4,04	2,27	1,54	2,07	1,36	1,90	1,40
fraise	1,09	3,36	5,75	1,66	2,58	2,70	4,84	2,20	1,17	2,26	1,65
thé	2,07	1,73	1,46	1,89	1,86	2,28	1,43	2,14	2,58	2,85	2,40
truffe	3,28	1,74	1,56	1,81	1,61	1,59	1,47	2,00	3,29	2,57	1,73
vanille	1,12	3,46	3,81	2,09	3,41	2,96	2,40	1,85	1,42	2,28	2,90
violette	1,53	1,66	3,06	2,83	3,14	3,26	2,26	2,10	2,28	2,37	1,53
noix	1,58	3,21	2,31	1,45	1,50	1,60	1,92	2,43	2,42	2,54	3,72
pyrole	1,17	1,47	2,79	2,43	2,14	1,93	1,49	5,23	1,41	1,67	1,88
boisé	2,11	1,66	1,74	2,35	1,78	2,27	1,52	2,24	3,06	3,39	2,15

Les données du tableau sont rassemblées dans la feuille Odors-moyennes du classeur Statistica Odors.stw.

1) Réaliser une ACP normée sur ce tableau de données.

2) Examiner le tableau des valeurs propres et la représentation graphique correspondante. Selon le choix du critère utilisé, de 3 à 5 composantes principales pourraient être retenues. Justifier.

- 3) a) Quels sont les 8 individus (odorants) qui contribuent le plus à la formation du premier axe ? Quel est le signe de la coordonnée correspondante ?
- b) De même, quelles sont les 5 catégories qui contribuent le plus à la formation du premier axe ? Quel est le signe des saturations correspondantes ?
- c) Faire une synthèse des résultats trouvés concernant le premier axe factoriel.
- 4) Etudier de même le second axe.
- 5) Quelle est la variable (catégorie) la plus mal représentée dans le premier plan factoriel ? Justifier à partir des tableaux des résultats et à partir de la représentation graphique.

## **1.7 Variantes et extensions de la méthode**

### **1.7.1 ACP pondérée, ACP non normée**

Dans certains cas, il peut être pertinent de pondérer les individus. Par exemple, il peut s'agir de regrouper les observations identiques. Ou encore, dans une ACP relative à des données socio-économiques sur des entités géographiques telles que des régions ou des départements, il peut être pertinent de pondérer chaque observation par une donnée démographique (nombre d'habitants).

Il est également possible de réaliser l'ACP sur les covariances des variables de départ, au lieu d'utiliser les corrélations. Le poids d'une variable dépend alors de son écart type, alors que dans l'ACP normée, toutes les variables ont le même poids.

### **1.7.2 ACP avec rotation**

Par construction, les composantes principales sont des abstractions mathématiques et ne possèdent pas nécessairement de signification intuitive. Après avoir réalisé l'ACP, il peut parfois être intéressant de définir d'autres variables en effectuant une combinaison linéaire des composantes principales retenues, à l'aide d'une "rotation". L'objectif est généralement d'augmenter les saturations, c'est-à-dire les corrélations entre ces nouveaux "facteurs" et certaines variables de départ. Les nouveaux "facteurs" ainsi obtenus perdent les propriétés des facteurs principaux. Par exemple, le premier d'entre eux ne correspond plus à la direction de plus grande dispersion du nuage des individus. En revanche, la part de variance expliquée par les facteurs retenus reste identique. Il existe différents critères (varimax, quartimax, equamax, etc) permettant d'obtenir une rotation conduisant à des saturations proches de 1 ou -1, ou au contraire proches de 0.

Cette possibilité n'est pas disponible dans la méthode "ACP à la française" de Statistica. En revanche, on peut l'utiliser en utilisant le module "Analyse factorielle" convenablement paramétré.

## **1.8 Travail à rendre par mail**

Dans le cadre d'une recherche nommée "Identité humaine – Identité animale"<sup>2</sup>, un groupe de travail en sciences sociales de l'Université de Lausanne a cherché à mieux cerner les caractéristiques perçues généralement comme typiques des êtres humains ou typiques des animaux, ainsi que les caractéristiques perçues comme communes aux deux espèces.

<sup>2</sup> Ricciardi Joos Paola, DES THEORIES IMPLICITES DE L'HOMME. Comparer l'être humain et l'animal dans des contextes intra- et inter-espèces. Thèse présentée à la Faculté des sciences sociales et politiques de l'Université de Lausanne (Suisse), en cotutelle avec l'Université de Paris X-Nanterre (France), Lausanne, 2004



En particulier, les chercheurs ont mené une enquête dans laquelle les sujets évaluaient soit pour l'homme, soit pour l'animal (un mammifère), les conséquences fâcheuses ou bénéfiques du clonage reproductif. Le questionnaire comportait deux séries de 9 questions, l'une portant sur l'individu à naître, l'autre sur l'espèce. Pour chaque question, la personne interrogée devait évaluer sur une échelle en 6 points (-3, -2, -1, +1, +2, +3) l'effet du clonage sur telle ou telle capacité de l'individu ou de l'espèce. Un exemple de questionnaire de ce type est donné en annexe (fichier Enquete-Clonage-Annexe.pdf).

Les 9 questions sont les suivantes :

- 1 - Conséquences fâcheuses sur le développement de ses capacités motrices
- 2 - Conséquences fâcheuses sur le développement de son système respiratoire
- 3 - Conséquences fâcheuses sur le développement de ses capacités d'apprentissage
- 4 - Conséquences fâcheuses sur le développement de sa morphologie générale
- 5 - Conséquences fâcheuses sur le développement de son système immunitaire
- 6 - Conséquences fâcheuses sur le développement de ses capacités d'imitation
- 7 - Conséquences fâcheuses sur le développement de ses capacités musculaires
- 8 - Conséquences fâcheuses sur le développement de son système nerveux
- 9 - Conséquences fâcheuses sur le développement de ses capacités de reproduction

La population interrogée est constituée d'étudiants de première année de psychologie de l'Université de Lausanne (âge moyen : 20 ans). Environ la moitié des sujets ont répondu au questionnaire relatif à l'homme et à l'espèce humaine, pendant que l'autre moitié répondait au questionnaire relatif à l'animal. 187 réponses ont ainsi été recueillies.

L'ensemble des 187 réponses a été saisi dans la feuille de données "Données" du classeur Enquete-Clonage.stw. Quelques remarques relatives à ce tableau de données :

- Ce tableau ne distingue pas les réponses relatives à l'homme de celles relatives à l'animal. Les traitements envisagés ci-dessous porteront sur l'ensemble des réponses.
- Les variables e1 à e9 désignent les réponses aux 9 questions relatives aux conséquences pour l'espèce, les variables i1 à i9 désignent les réponses aux 9 questions relatives aux conséquences pour l'individu à naître. Pour ces 18 variables, les valeurs ont été *centrées et réduites*.

1) Etudier le tableau des corrélations entre les 18 variables. Quelle remarque peut-on faire ?

2) On réalise une ACP sur les 18 variables.

Au vu du tableau des valeurs propres, combien d'axes factoriels faudrait-il étudier ? Justifier.  
N.B. L'étude ci-dessous portera sur les trois premiers axes.

3) Etude des qualités de représentation dans l'espace formé par les trois premiers axes.

Y a-t-il des questions particulièrement mal représentées dans l'espace défini par les trois premières composantes principales ? Y a-t-il des questions nettement mieux représentées que les autres ?

4) Etude de la première variable factorielle.

a) Quel est le signe des coordonnées des 18 variables selon le premier axe factoriel ? Pouvaient-on s'attendre à un résultat de ce type ? Comment appelle-t-on cet effet ?

b) Comment peut-on qualifier les contributions des variables à la formation de cet axe ?

5) Etude de la deuxième variable factorielle.

Comment se répartissent les signes des coordonnées des variables selon le deuxième axe factoriel ? Comment peut-on décrire cet axe en termes d'opposition entre variables ?

## 6) Etude de la troisième variable factorielle.

Quelles sont les variables dont la contribution à la formation du troisième axe factoriel est supérieure à la moyenne ? Indiquer le signe de la coordonnée correspondante. Comment peut-on décrire cet axe ?

7) On poursuit l'étude en effectuant deux analyses en composantes principales, la première portant sur les 9 questions relatives à l'espèce, la seconde sur les 9 questions relatives à l'individu.

Commenter les tableaux des valeurs propres obtenus pour ces deux ACP.

8) A partir de l'étude menée dans les questions 1 à 7, justifier l'affirmation des auteurs :

*"La pertinence du regroupement des 18 questions en deux scores moyens - l'un basé sur les 9 questions posées sur l'individu et l'autre basé sur les 9 questions posées sur l'espèce - est soutenue par les résultats d'une analyse factorielle en composante principale effectuée sur l'ensemble des 18 questions (les 9 posées pour l'individu et les 9 posées pour l'espèce). (...) En outre, une ACP effectuée sur les 9 questions posées sur l'individu puis une autre ACP effectuée sur les 9 questions posées sur l'espèce concluent toutes deux à une échelle unidimensionnelle."*

Travail à rendre par mail à votre enseignant (Francois.Carpentier@univ-brest.fr) :

- Un classeur Statistica contenant les résultats numériques de l'ACP et les graphiques.
- Un fichier Word ou un rapport Statistica contenant votre interprétation des résultats.