Analyse en Composantes Principales

Position du problème :
On a observé p variables sur n individus : protocole multivarié.

On cherche à remplacer ces p variables par q nouvelles variables résumant au mieux le protocole, avec $q <= p$, et si possible $q = 2$.

Mini-exemple : 6 sujets décrits par 4 variables :

Données :

Suj	X_1	X_2	X_3	X_4
s1	-11	-60	110	40
s2	-12	-62	93	25
s3	-15	-80	113	39
s4	-14	-75	94	25
s5	-14.5	-82	100	30
s6	-13	-72	102	32

Corrélations des variables prises deux à deux :

	X_1	X_2	X_3	X_4
X_1	1.0000	0.9701	-0.0635	0.0940
X_2	0.9701	1.0000	-0.1018	0.0373
X_3	-0.0635	-0.1018	1.0000	0.9856
X_4	0.0940	0.0373	0.9856	1.0000

Nuage de points obtenu en prenant 2 variables, 3 variables . . .

Données centrées réduites (sans correction ou avec correction) :

Suj	Z_1	Z_2	Z_3	Z_4
s1	1.5993	1.4197	1.0722	1.3648
s2	0.8885	1.1798	-1.2063	-1.1420
s3	-1.2439	-0.9798	1.4743	1.1977
s4	-0.5331	-0.3799	-1.0722	-1.1420
s5	-0.8885	-1.2198	-0.2681	-0.3064
s6	0.1777	-0.0200	0.0000	0.0279

Suj	Z_{1c}	Z_{2c}	Z_{3c}	Z_{4c}
s1	1.4600	1.2960	0.9788	1.2459
s2	0.8111	1.0770	-1.1012	-1.0425
s3	-1.1356	-0.8944	1.3459	1.0933
s4	-0.4867	-0.3468	-0.9788	-1.0425
s5	-0.8111	-1.1135	-0.2447	-0.2797
s6	0.1622	-0.0183	0.0000	0.0254

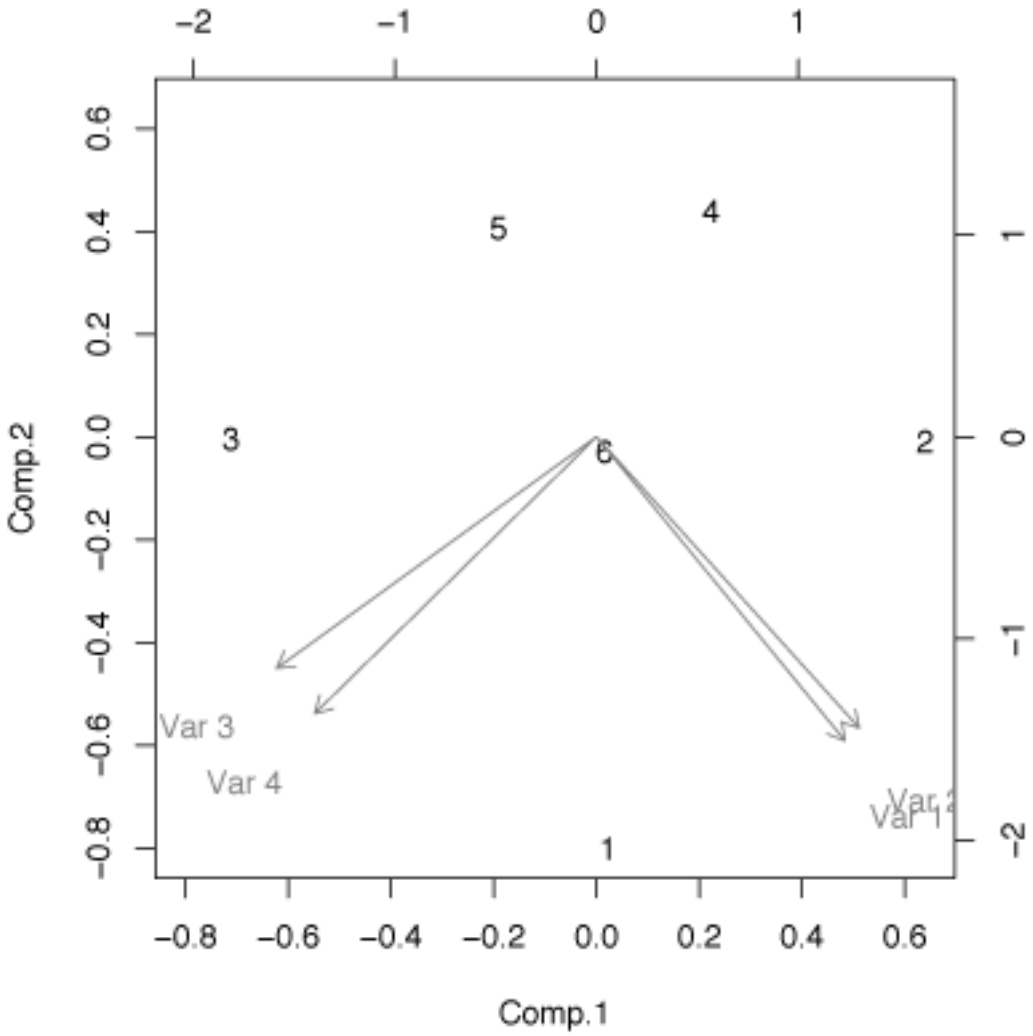
N.B. La matrice des corrélations reste inchangée.

Inertie totale du nuage avec les variables centrées réduites :

$I = \sum z_{ij}^2 = 6 \times 4 = 24$

Le principal résultat fourni par l'ACP :

Projection du nuage selon les deux premières composantes principales



Composantes principales : variables CP_1, CP_2, CP_3, CP_4 telles que :

• 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 éliminé
• idem pour CP_3 et CP_4
• Les variables CP_j sont des combinaisons des variables Z_j
• Les variables CP_j ne sont en général pas réduites
• Les variables CP_j sont deux à deux indépendantes : pour $j \neq k$, $\rho(CP_j, CP_k) = 0$.

Résultats numériques attendus :

– Valeurs propres

– Scores des individus
– Contributions des individus
– Inertie relative des individus
– Qualité de la représentation des individus

– Saturations des variables
– Contributions des variables
– Qualité de la représentation des variables

Valeurs propres : Chaque valeur propre représente la variance prise en compte par la composante principale correspondante

	CP_1	CP_2	CP_3	CP_4
Valeur propre	2.0011	1.8668	0.0317	0.0003
Prop. variance	0.5003	0.4917	0.0079	0.0001
Prop. cumulée	0.5003	0.9920	0.9999	1.0000

Ici, les deux premiers axes rendent comptent de 99.2% de la variance totale.

Scores des individus : valeurs prises par les variables CP_j sur les individus.

Suj	CP_1	CP_2	CP_3	CP_4
s1	0.0771	-2.7515	-0.0935	0.0166
s2	2.2153	-0.0327	0.1778	-0.0095
s3	-2.4608	-0.0173	0.2445	-0.0036
s4	0.7734	1.5097	0.0664	0.0219
s5	-0.6606	1.3926	-0.2592	0.0064
s6	0.0556	-0.1008	-0.1360	-0.0319

Valeurs propres : variances des variables CP_j.

Expressions des composantes principales comme combinaisons linéaires des variables de départ :

Var	CP_1	CP_2	CP_3	CP_4
Z_1	0.445	-0.548	-0.656	-0.267
Z_2	0.470	-0.525	0.690	0.166
Z_3	-0.572	-0.418	0.232	-0.667
Z_4	-0.504	-0.499	-0.199	0.676

Par exemple :

$CP_1 = 0.445\ Z_1 + 0.470\ Z_2$ - $0.572\ Z_3$ - $0.504\ Z_4$

et, par exemple, pour le premier sujet :

$0.0771 = 0.445 \times 1.5993 + 0.470 \times 1.4197 - 0.572 \times 1.0722 - 0.504 \times 1.3648$

Ce tableau peut aussi être lu dans l'autre sens :

$Z_1 = 0.445\ CP_1$ - $0.548\ CP_2$ - $0.656\ CP_3$ - $0.267\ CP4$

et, pour le premier sujet :

$1.5993 = 0.445 \times 0.0771 - 0.548 \times (-2.7515) - 0.656 \times (-0.0935) - 0.267 \times 0.0166$

Saturations des variables : coefficients de corrélation entre Z_j et CP_k

Var	CP_1	CP_2	CP_3	CP_4
Z_1	0.6288	-0.7687	-0.1169	-0.0048
Z_2	0.6651	-0.7366	0.1228	0.0030
Z_3	-0.8094	-0.5857	0.0413	-0.0119
Z_4	-0.7129	-0.7002	-0.0355	0.0121

Il existe un lien entre les deux tableaux précédents : Par exemple : $0.6288 = \sqrt{2.0011} \times 0.445$

Inertie relative d'un individu :

Carré de la distance de l'individu à l'origine divisé par l'inertie totale du nuage ;

Contribution (relative) d'un individu à la formation d'une composante principale :

Par exemple, pour s1 et CP_1 :

$$CTR = \frac{0.0771^2}{0.0771^2 + \cdots + 0.0556^2} = \frac{0.0771^2}{6 \times 2.0011} = 0.64\%$$

Qualité de la représentation d'un individu par une composante principale, par les composantes principales retenues :

Pour s1 et CP_2 :

$$QLT = \frac{2.7515^2}{0.0771^2 + 2.7515^2 + 0.0935^2 + 0.0166^2} = 0.9980$$

Pour s1 et CP_1, CP_2

$$QLT = \frac{0.0771^2 + 2.7515^2}{0.0771^2 + 2.7515^2 + 0.0935^2 + 0.0166^2} = 0.9988$$

Pour tous les individus :

Suj	QLT
s1	0.9988
s2	0.9936
s3	0.9902
s4	0.9983
s5	0.9725
s6	0.4044

Tous les individus, sauf le dernier, sont très bien représentés par les deux premières composantes principales.

Contribution d'une variable à la formation d'une composante principale :

Exemple : contribution de la première variable à la formation de la première composante principale

$$CTR = \frac{0.6288^2}{0.6288^2 + 0.6651^2 + 0.8094^2 + 0.7129^2}$$

$$CTR = \frac{0.6288^2}{2.011} = 0.1976$$

Qualité de la représentation d'une variable par une composante principale : carré du coefficient correspondant.

Qualité de la représentation d'une variable par les composantes principales retenues :

Pour Z_1 et CP_1, CP_2 :

$$QLT = \frac{0.6288^2 + 0.7687^2}{0.6288^2 + 0.7687^2 + 0.1169^2 + 0.0048^2}$$

$$QLT = 0.6288^2 + 0.7687^2 = 0.9863$$

Pour les 4 variables :

Var	QLT
Z_1	0.9863
Z_2	0.9849
Z_3	0.9982
Z_4	0.9985

Interprétation graphique : carré de la longueur du "vecteur" représentant la variable.

Interprétation des résultats

Scores des individus et saturations ne sont pas exprimées avec la même unité de mesure.

Interpréter chaque axe : part de la variance dont il rend compte, variables avec lesquelles il est corrélé.

Proximités entre individus : à interpréter avec prudence, ils peuvent prendre des valeurs très différentes sur des variables non représentées.

Individus proches de l'origine : ils ont, de toutes façons, peu contribué à l'inertie.

Interpréter plutôt les oppositions marquées entre individus.

Effet de masse ou de taille : lorsque toutes les variables ont entre elles des corrélations positives, le premier axe classe simplement les individus par valeurs des variables.

Variantes

– ACP non normée

– ACP pondérée : on affecte des poids aux individus

Analyse Factorielle des Correspondances

Position du problème :
On dispose d'un tableau de contingence comportant un grand nombre de lignes et de colonnes. On veut faire une étude de ces données, plus précise qu'un simple χ^2.

Pour chacune des variables, quelles sont les modalités qui se ressemblent, quelles sont celles qui s'opposent ?

Pour les couples de modalités des deux variables, quelles sont les modalités qui s'attirent, quelles sont celles qui se repoussent ?

Exemple "historique" : Données Caith

Couleur des yeux et couleur des cheveux pour 5387 enfants du comté de Caithness (Student, 1940)

	HFAI	HRED	HMEDIUM	HDARK	HBLACK
EBLUE	326	38	241	110	3
ELIGHT	688	116	584	188	4
EMEDIUM	343	84	909	412	26
EDARK	98	48	403	681	85

Valeur du Chi2 à 12 ddl : 1240.04
Niveau de significativité : inférieur à 10^{-5}

Etude descriptive : profils des lignes, des colonnes, taux de liaison

Fréquences conjointes f_{ij} et marginales $f_{i.}$, $f_{.j}$

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK	$f_{i.}$
EBLUE	0,0605	0,0071	0,0447	0,0204	0,0006	0,1333
ELIGHT	0,1277	0,0215	0,1084	0,0349	0,0007	0,2933
EMEDIUM	0,0637	0,0156	0,1687	0,0765	0,0048	0,3293
EDARK	0,0182	0,0089	0,0748	0,1264	0,0158	0,2441
$f_{.j}$	0,2701	0,0531	0,3967	0,2582	0,0219	1

Profil des lignes

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK	Total
EBLUE	0,454	0,053	0,336	0,153	0,004	1
ELIGHT	0,435	0,073	0,370	0,119	0,003	1
EMEDIUM	0,193	0,047	0,512	0,232	0,015	1
EDARK	0,075	0,037	0,306	0,518	0,065	1
Masse	0,270	0,053	0,397	0,258	0,022	1

Profil des colonnes

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK	Masse
EBLUE	0,224	0,133	0,113	0,079	0,025	0,133
ELIGHT	0,473	0,406	0,273	0,135	0,034	0,293
EMEDIUM	0,236	0,294	0,425	0,296	0,220	0,329
EDARK	0,067	0,168	0,189	0,490	0,720	0,244
Total	1	1	1	1	1	1

Taux de liaison t_{ij} :

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK
EBLUE	0,6810	-0,0031	-0,1539	-0,4067	-0,8093
ELIGHT	0,6122	0,3829	-0,0683	-0,5392	-0,8844
EMEDIUM	-0,2841	-0,1081	0,2917	-0,1006	-0,3309
EDARK	-0,7241	-0,3125	-0,2275	1,0056	1,9509

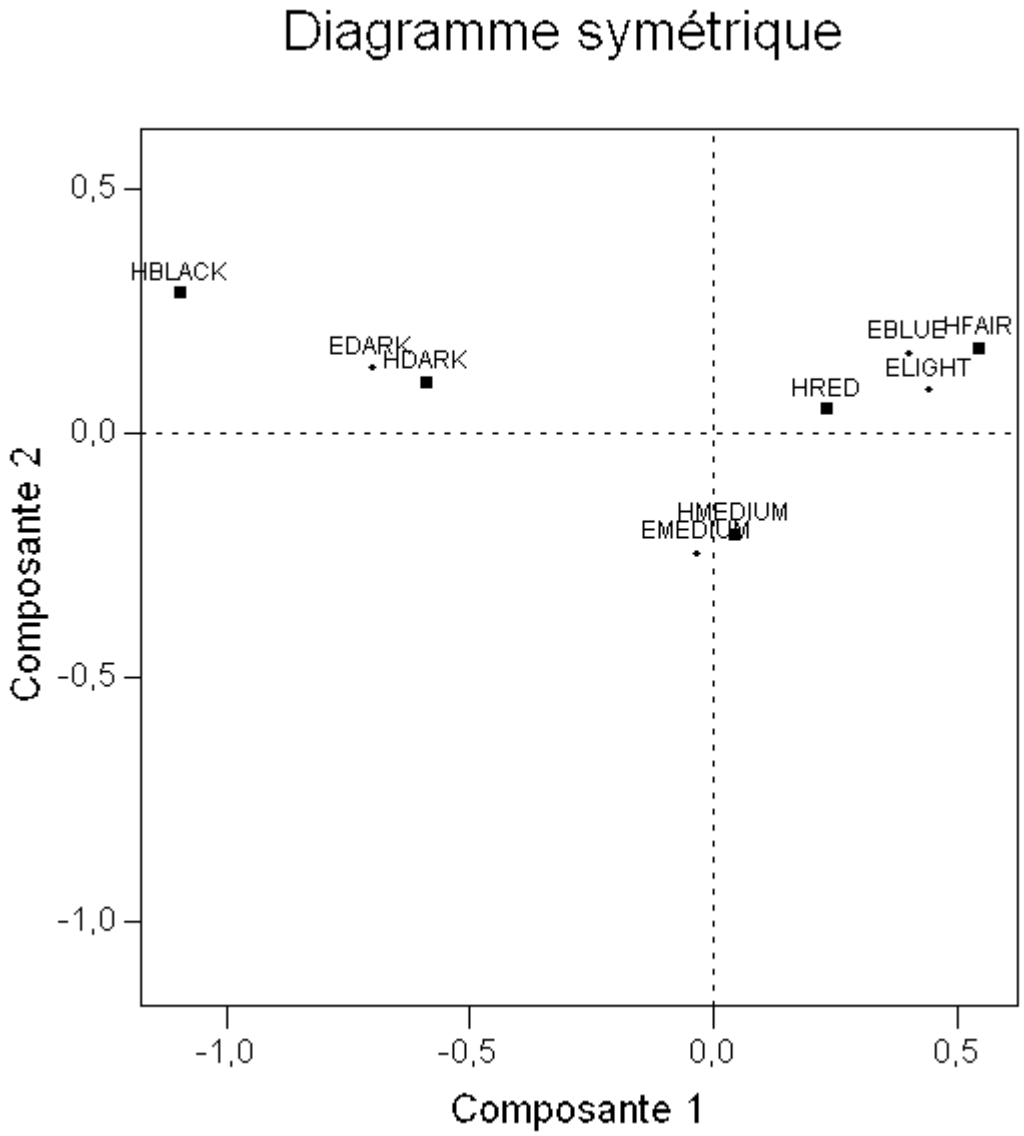
Définition : $t_{ij} = \frac{f_{ij} - f_{i.}f_{.j}}{f_{i.}f_{.j}}$

La moyenne des t_{ij}, pondérés par les coefficients $f_{i.}f_{.j}$, est nulle.

La moyenne des t_{ij}^2, pondérés par les coefficients $f_{i.}f_{.j}$, est le *carré moyen de contingence* Φ^2. On a :

$$\Phi^2 = \sum\sum \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = \frac{\chi^2}{N}$$

Analyse des correspondances proprement dite :



Le calcul des valeurs propres est fait à partir d'un tableau calculé à partir des profils ligne et des profils colonne. Nous ne le détaillerons pas ici.

Valeurs propres associées aux axes principaux :

La première valeur propre vaut 1, et n'est pas mentionnée dans les résultats. Les autres valeurs propres sont inférieures à 1.
Somme des valeurs propres : Φ^2.

Axe	Inertie	Proportion	Cumulé
1	0,1992	0,8656	0,8656
2	0,0301	0,1307	0,9963
3	0,0009	0,0037	1,0000
Total	0,2302		

Contributions relatives des modalités à la formation de l'inertie : *Inerties relatives*

	HFAIR	HRED	HMEDIUM	HDARK	HBLACK	Total
EBLUE	0,073	0,000	0,005	0,025	0,008	0,111
ELIGHT	0,129	0,010	0,002	0,096	0,022	0,259
EMEDIUM	0,031	0,001	0,048	0,004	0,003	0,088
EDARK	0,150	0,005	0,022	0,277	0,088	0,543
Total	0,383	0,016	0,078	0,401	0,122	1,000

Pour les deux premiers axes factoriels :

Contributions des lignes

Nom	Qual	Mass	Inert
EBLUE	0,979	0,133	0,111
ELIGHT	0,995	0,293	0,259
EMEDIUM	0,999	0,329	0,088
EDARK	1,000	0,244	0,543

$$\text{Qualité} = \frac{(\text{Coord. selon CP1})^2 + (\text{Coord. selon CP2})^2}{\sum(\text{Coord. selon les CP})^2}$$

$$\text{Exemple : } 0.979 = \frac{0.400^2 + 0.165^2}{0.400^2 + 0.165^2 + 0.064^2}$$

	—Composante 1—			—Composante 2—		
Nom	Coord	Corr	Contr	Coord	Corr	Contr
EBLUE	0,400	0,836	0,107	0,165	0,143	0,121
ELIGHT	0,441	0,956	0,286	0,088	0,039	0,076
EMEDIUM	-0,034	0,018	0,002	-0,245	0,981	0,657
EDARK	-0,703	0,965	0,605	0,134	0,035	0,145

Corr : qualité de la représentation de l'individu par sa projection sur l'axe.

$$\text{Qualité suivant CP i} = \frac{(\text{Coord. selon CP i})^2}{\sum(\text{Coord. selon les CP})^2}$$

$$\text{Exemple : } 0.836 = \frac{0.400^2}{0.400^2 + 0.165^2 + 0.064^2}$$

Contr : contribution relative d'un individu à la formation de l'inertie d'un axe.

$$\text{Contr(Individu i, Axe k)} = \frac{\text{Masse ligne i} \times (\text{Coord. indiv. i selon CP k})^2}{\text{Valeur propre relative à l'axe k}}$$

$$\text{Exemple : } 0.107 = \frac{0.400^2 \times 0.133}{0.1992}$$

Contribution des colonnes

Nom	Qual	Mass	Inert
HFAIR	1,000	0,270	0,383
HRED	0,803	0,053	0,016
HMEDIUM	1,000	0,397	0,078
HDARK	1,000	0,258	0,401
HBLACK	0,998	0,022	0,122

	—Composante 1—			—Composante 2—		
Nom	Coord	Corr	Contr	Coord	Corr	Contr
HFAIR	0,544	0,907	0,401	0,174	0,093	0,271
HRED	0,233	0,770	0,014	0,048	0,033	0,004
HMEDIUM	0,042	0,039	0,004	-0,208	0,961	0,572
HDARK	-0,589	0,969	0,449	0,104	0,030	0,093
HBLACK	-1,094	0,934	0,132	0,286	0,064	0,060

Interprétation des résultats

– Valeur propre proche de 1 : forte liaison entre lignes et colonnes
– Pour chaque axe : points-lignes et points-colonnes dont les contributions sont fortes. Par exemple : points dont la contribution est supérieure à la contribution moyenne. Voir aussi les points bien représentés
– Une modalité (ou un groupe de modalités proches) bien représentée : axe spécifique
– Points proches : à interpéter avec précaution. Intéressant si les points sont bien représentés.
– Etudier les associations points-lignes / points-colonnes proches