Chapitre 3

Analyse en Composantes Principales

Introduction

L'objet de ce cours est de donner quelques outils couramment employés en statistique pour traiter des données multidimensionnelles. Ces données correspondent souvent à l'observation de nombreuses variables aléatoires sur plusieurs individus, le mot individu étant à prendre en un sens très large. Ces données sont représentées sous forme d'un tableau ou chaque ligne représente les variables mesurées sur un individu. Le but est d'extraire le maximum d'informations de ce tableau de données. Suivant la nature de la question posée, et suivant la nature des données, plusieurs méthodes sont possibles.

Si les variables auxquelles on s'intéresse sont toutes des variables quantitatives, il s'agit d'un problème d'analyse en composante principale (ACP). S'il s'agit de deux variables qualitatives, on parle d'analyse factorielle des correspondances (AFC).

I. Définition et Objectif de l'ACP:

L'ACP est utilisée lorsqu'on observe sur n individus, p variables quantitatives X^1, X^2, . . ., X^p présentant des liaisons multiples que l'on veut analyser. Ces observations sont regroupées dans un tableau (matrice) rectangulaire X ayant n lignes (individus) et p colonnes (variables)

$$X = \begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^p \\ x_2^1 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & \vdots & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^p \end{pmatrix}$$

où x_i^j est la valeur observée de la j-ième variable X^j sur le i-ème individu de l'échantillon.

où x_i^j est la valeur observée de la j-ième variable X^j sur le i-ème individu de l'échantillon

L'ACP est une méthode de description et de résumer d'un tableau de données (n; p). Son objectif principal est de remplacer ce tableau de données par un tableau de dimension réduite (n; q) (q < p). Un des avantages de cette réduction de dimension est par exemple, de pouvoir obtenir des représentations graphiques des données. En effet, lorsque p = 2, chaque individu peut être représente par un point dans un plan, et le tableau de données initial peut être visualisé graphiquement par un nuage de points dans un plan. Dès que p supérieure ou égale à 4, une représentation graphique du nuage de points est difficile, et l'un des buts de l'ACP est de trouver "la

meilleure" représentation plane du nuage de points, ce qui revient à chercher un tableau (n; q = 2) qui approche "au mieux" le tableau de données initial. On cherche en particulier une représentation plane :

- qui minimise les déformations du nuage initial ;
- qui fait apparaitre les liaisons entre les variables initiales ;
- qui permet de résumer l'information contenue dans le tableau initial (n; p) dans un tableau de plus faible dimension (n; q), q < p, (en fait q = 2; 3).

Cette réduction va être obtenue en remplaçant les variables initiales X^j , j = 1,..., p, par un petit nombre de nouvelles variables C^j , j = 1,..., q, appelées composantes principales, qui sont non corrélées, et combinaisons linéaires des X^j . Ces nouvelles variables vont être obtenues en analysant la structure des covariances, ou des corrélations, entre les variables initiales.

Avant de décrire plus précisément la méthode, il faut en souligner quelques limites :

- l'ACP ne permet pas le traitement de variables qualitatives.
- l'ACP ne détecte que d'éventuelles liaisons linéaires entre variables.

L'ACP présente de nombreuses variantes selon les transformations apportées au tableau de données. Parmi ces variantes, l'ACP sur un tableau où les colonnes sont centrées et réduites, appelée ACP normée est la plus fréquemment utilisée.

Application

On étudie les consommations annuelles d'un groupe d'individus, exprimées dinars, de 8 denrées alimentaires (les variables), les individus étant 8 catégories socioprofessionnelles. Les données sont des moyennes par CSP :

	PAO	PAA	HUI	HUIA	POT	LEC	RAI	PLP
AGRI	167	1	163	23	41	8	6	6
SAAG	162	2	141	12	40	12	4	15
PRIN	119	6	69	56	39	5	13	41
CSUP	87	11	63	111	27	3	18	39
CMOY	103	5	68	77	32	4	11	30
EMPL	111	4	72	66	34	6	10	28

OUVR	**130**	**3**	**76**	**52**	**43**	**7**	**7**	**16**
INAC	**138**	**7**	**117**	**74**	**53**	**8**	**12**	**20**

Les individus sont : AGRI = Exploitants agricoles, SAAG= Salariés agricoles; PRIN = Professions indépendantes, CSUP = Cadres supérieurs, CMOY= Cadres moyens, EMPL= Employés, OUVR = Ouvriers, INAC = Inactifs.

Les 8 variables numériques sont : PAO = Pain ordinaire, PAA = Autre pain, HUI = huile d'olive, HUIA=Autres huiles, POT= Pommes de terre, LEC=Légumes secs, RAI=Raisin de tables, PLP= Plats préparés

1) **Comment peut-on faire une analyse séparément de chacune de ces 8 variables?**
2) **Comment analyser les liaisons entre 2 variables ?**
3) **Comment faire une étude simultanée des 8 variables?**

Corrigé :

1) **On faire une analyse séparément de chacune de ces 8 variables soit en faisant un graphique (diagramme en bâtons), soit en calculant des résumés numériques (moyenne arithmétiques, variance, écart type...).**

2) **On peut analyser les liaisons entre 2 variables (par exemple PAO et HUI), soit en faisant un graphique du type nuage de points, soit en calculant leur coefficient de corrélation linéaire, voire en réalisant la régression de l'une sur l'autre.**

3) **L'étude simultanée des 8 variables, ne peut pas se réaliser par les deux types d'analyses citées précédemment. La difficulté vient de ce que les individus ne sont plus représentés dans un plan de dimension 2, mais dans un espace de dimension 8 (chaque individu étant caractérisé par les 8 denrées alimentaires). Pour ce faire, l'Analyse en Composantes Principales permet de revenir à un espace de dimension réduite (par exemple, ici, 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent des données initiales.**

II. Données initiales

1) Nuage des points individus

On associe à chaque individu i, un vecteur xi contenant les valeurs de chaque variable pour l'individu considéré:

$$x_i' = (x_i^1, x_i^2, \cdots, x_i^p) \qquad (i\text{-ème ligne de la matrice } X)\,.$$

Chaque individu peut alors être représenté par un point dans R^p, appelée espace des individus.

a) Matrice des poids :

On affecte à chaque individu un poids p_i reflétant son importance par rapport aux autres individus avec:

$$p_i > 0 \text{ et } \sum_{i=1}^{n} p_i = 1$$

On appelle matrice des poids la matrice diagonale (n; n) dont les éléments diagonaux sont les poids pi. Elle sera notée

$$D = \mathrm{diag}(p_1, p_2, \cdots, p_n) = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & p_n \end{pmatrix}$$

Le cas le plus fréquent est de considérer que tous les individus ont la même importance : $p_i = 1/n$, pour tout i = 1,....,n. Si les individus sont par exemple des pays, on peut être amenée à prendre:

$$p_i = \frac{\text{Population du pays i}}{\text{Population totale}}\,.$$

b) Nuage des individus:

On appelle nuage des individus, l'ensemble des points xi munis de leurs poids : M = {f(x_i; p_i) ; i = 1,...,n }.

c) Barycentre du nuage de points**:**

Le point g de R^p dont les coordonnées sont les moyennes empiriques des variables $g' = (\bar{x}^1, \bar{x}^2, \ldots\ldots \bar{x}^p)$, est le centre de gravité (le barycentre) du nuage de points M. le point g représente l'individu moyen.

2) Données centrées réduites: Centrage et réduction des données.

Tout d'abord, il faut noter que le centrage des variables d'un tableau soumis à une A.C.P. (on retranche à chaque observation la moyenne de la variable correspondante) ne modifie en rien les résultats de l'A.C.P. En effet, on utilise comme critère la maximisation de la dispersion (de l'inertie) et la dispersion d'une variable n'est pas modifiée par son centrage. Comme il est plus commode de travailler avec des données centrées (les expressions manipulées sont plus simples à écrire), les A.C.P. sont systématiquement réalisées après centrage de chaque variable.

Dans la pratique, on peut ainsi faire soit une A.C.P. centrée (les variables x_j considérées sont seulement centrées), soit une A.C.P. réduite (les variables sont centrées et réduites : on divise chaque donnée centrée par l'écart-type de la variable correspondante).

On recommande l'A.C.P. seulement centrée lorsque les variables sont homogènes : même signification, même unité de mesure, même ordre de grandeur... C'est le cas de l'exemple traité au paragraphe précédent. Au contraire, on recommande l'A.C.P. réduite lorsque les variables sont hétérogènes, c'est à-dire dans les autres cas.

Les données centrées et réduites sont notées:

$$z_i^j = \frac{x_i^j - \bar{x}^j}{s_j}$$

(S_j l'écart type empirique de la variable)

Ce sont des données sans dimension. Elles sont regroupées dans un tableau

$$Z = \left[z^1, z^2, \cdots, z^p\right] = \begin{pmatrix} z_1^1 & \cdots & z_1^j & \cdots & z_1^p \\ z_2^1 & \cdots & z_2^j & \cdots & z_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_n^1 & \cdots & z_n^j & \cdots & z_n^p \end{pmatrix}$$

3) Matrice de corrélations empirique.

Notons r_{ij} les corrélations empiriques des variables, et R la matrice des corrélations empiriques

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

Avec R=Z'DZ

La matrice R est la matrice de variance-covariance des données centrées réduites et résume la structure des dépendances linéaires entre les p variables.

Suite de l'application :

On fait une ACP sur variables centrées-réduites des 8 variables numériques, le programme SAS permet d'obtenir les sorties ci-dessous:

Matrice des corrélations

	PAO	PAA	HUI	HUIA	POT	LEC	RAI	PLP
PAO	1.0000	-.7737	0.9262	-.9058	0.6564	0.8886	-.8334	-.8558
PAA	-.7737	1.0000	-.6040	0.9044	-.3329	-.6734	0.9588	0.7712
HUI	0.9262	-.6040	1.0000	-.7502	0.5171	0.7917	-.6690	-.8280
HUIA	-.9058	0.9044	-.7502	1.0000	-.4186	-.8386	0.9239	0.7198
POT	0.6564	-.3329	0.5171	-.4186	1.0000	0.6029	-.4099	-.5540
LEC	0.8886	-.6734	0.7917	-.8386	0.6029	1.0000	-.8245	-.7509
RAI	-.8334	0.9588	-.6690	0.9239	-.4099	-.8245	1.0000	0.8344
PLP	-.8558	0.7712	-.8280	0.7198	-.5540	-.7509	0.8344	1.0000

4) **Quels sont les objectifs de l'ACP?**
5) **Que signifie cette matrice et comment est-elle calculée ?**
6) **Quelle est la signification des deux valeurs mises en gras ?**
7) **Cette matrice est-elle diagonalisable ?**
8) **Donner sa trace. Que représente cette valeur?**

Corrigé:

4) L'intérêt de cette ACP est double:

- Réduire le nombre de variables décrivant les consommations des CSP en fournissant un petit nombre de nouvelles variables (les composantes principales) décrivant les CSP.

- Identifier des groupes de CSP ayant les mêmes types de consommation et décrire leurs consommations en denrées alimentaires

5) C'est la matrice des corrélations des variables (pour le calcul voir le cours)

6) -0.9058: la corrélation entre PAO et HUIA est forte et négative
0.9588: la corrélation entre PAA et RAI est forte et positive

7) Oui elle est diagonalisable car il s'git dune matrice symétrique

8) Trace R=P=8 (nombre de variables)

III. Composantes principales

1) Détermination des composantes principale

Les axes principaux sont déterminés en recherchant les vecteurs propres (u_1,.....u_p) de la matrice des corrélations R. en d'autre terme ceci revient à diagonaliser R.

Ainsi, on calcule les valeurs propres ($\lambda_1, \lambda_2 ....., \lambda_p$) de R qui seront classées par ordre décroissant $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_p$

On note U la matrice (carrée de dimension p) des vecteurs propres de R

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

- Z est la matrice des variables centrées réduites.
- U est la matrice des vecteurs propres de la matrice des corrélations

Les coordonnées factorielles des points individus sont données par les projections du point i sur l'axe principal avec:

$$C_i^{\alpha} = \sum_{k=1}^{\alpha} z_i^k u_i^k$$

• Les composantes principales dégagent la redondance entre les variables.
• La variance d'une composante principale est sa valeur propre associée.
• Chaque valeur propre mesure la part de la variance expliquée par l'axe factoriel correspondant.

Remarque : Mathématiquement, la détermination des axes factoriels se fait par diagonalisation de la matrice de variances-covariances, d'où le vocabulaire utilisé (valeurs propres, vecteurs propres).

Explication : Détermination des composantes principales

Centrer e réduire la matrice X des données initiales

$$X_{(n,p))} = \begin{pmatrix} X_1^1 & . & X_1^\alpha & . & X_1^p \\ . & . & . & . & . \\ . & . & . & . & . \\ X_n^1 & . & X_1^\alpha & . & X_n^p \end{pmatrix} \qquad Z = \frac{X - \bar{x}}{\sigma} \qquad Z = \begin{pmatrix} Z_1^1 & . & Z_1^\alpha & . & Z_1^p \\ . & . & . & . & . \\ . & . & . & . & . \\ Z_n^1 & . & Z_1^\alpha & . & Z_n^p \end{pmatrix}$$

Calculer R : La matrice des corrélations des variables R=Z'DZ

$$R(p,p) = \begin{pmatrix} 1 & r_{12} & . & . & r_{1p} \\ r_{21} & 1 & . & . & . \\ . & . & . & . & . \\ r_{p1} & . & & . & 1 \end{pmatrix} \text{ et } D(n,n) = \begin{pmatrix} \frac{1}{n} & . & 0 & . & 0 \\ . & . & . & . & . \\ . & . & \frac{1}{n} & . & . \\ . & . & . & . & . \\ 0 & . & 0 & . & \frac{1}{n} \end{pmatrix} \text{ matrice des poids des individus}$$

La dimension de R =p = nombre de variables

Calculer les valeurs propres ($\lambda_1, \lambda_2 ...., \lambda_p$) et les vecteurs propres (u_1,.....u_p)

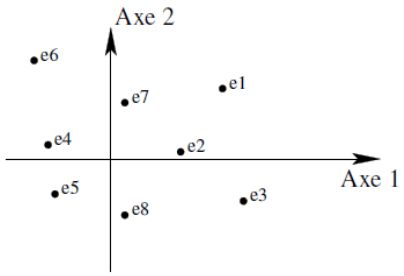
$$U = \begin{pmatrix} u_1^1 & . & u & . & u_1^p \\ . & . & . & . & . \\ . & . & . & . & . \\ u_n^1 & . & u_1^\alpha & . & u_n^p \end{pmatrix}$$ Matrice des vecteurs propres de R

Déterminer la matrice des composantes principales : $C = ZU$

$$C = ZU = \begin{pmatrix} Z_1^1 & . & Z_1^\alpha & . & Z_1^p \\ . & . & . & . & . \\ . & . & . & . & . \\ Z_n^1 & . & Z_1^\alpha & . & Z_n^p \end{pmatrix} \times \left(\begin{array}{ccccc} u_1^1 & . & u & . & u_1^p \\ . & . & . & . & . \\ . & . & . & . & . \\ u_n^1 & . & u_1^\alpha & . & u_n^p \end{array} \right) = \begin{pmatrix} C_1^1 & . & C_1^\alpha & . & C_1^p \\ . & . & . & . & . \\ . & . & . & . & . \\ C_n^1 & . & C_1^\alpha & . & C_n^p \end{pmatrix}$$

2) Représentation graphique des individus dans un plan principal

Pour deux composantes principales C^1 et C^2, on représente chaque individu "i" par un point d'abscisse C_i^1 et d'ordonnée C_i^2.



Représentation graphique des individus dans le premier plan principal.

Les nouveaux axes, appelés *axes factoriels*, sont choisis de la façon suivante :

- Le 1er axe factoriel, ou *axe principal d'inertie*, est la direction de "plus grand allongement" du nuage (en statistiques on dit : "de plus grande dispersion" ou "de plus grande inertie" du nuage).

Lorsque on projette les points du nuage sur cet axe, leurs projections sont plus dispersées qu'elles ne le seraient sur n'importe quel autre axe. L'axe factoriel F_1 est donc l'axe selon lequel est préservé, par projection, le maximum de la dispersion initiale des points du nuage. **Le fait que le nuage soit allongé précisément dans cette direction doit trouver une explication**. La nouvelle variable C^1 (la *composante principale n°1*) est le caractère selon lequel les individus se différencient le plus. Pourquoi ? Quelle signification peut bien avoir cette variable qui combine avec des poids plus ou moins importants (les coefficients a_i) les variables initiales mesurées sur les individus? Une étape fondamentale de l'ACP est l'interprétation de cette composante principale, qui se fera par l'examen de sa combinaison avec les variables de départ. On espère toujours pouvoir détecter dans cette nouvelle variable un *caractère complexe*, qui n'est pas directement mesurable par une seule quantité, mais

bien réel, comme par exemple la *santé* (pour des individus, pour des entreprises...), l'i*ndustrialisation* (d'une région...), la qualité du *jeu d'attaque* (pour un joueur de football, de tennis...), la *compétence dans les matières quantitatives* (pour un étudiant), etc.

-Le 2ème axe factoriel est la 2e direction d'allongement du nuage, c'est-à-dire celle qui explique, après le 1er axe, le maximum de l'inertie résiduelle. De plus le 2e axe est choisi orthogonal au 1er, ce qui traduit - comme nous le verrons- le fait que la 2e composante principale est non corrélée à la 1e (les vecteurs directeurs des 2 premiers axes ont un produit scalaire nul les 2 premières composantes principales ont une covariance nulle). Comme précédemment, on cherchera à donner un sens à cette 2e composante principale, en observant comment elle combine les variables de départ.

-Et ainsi de suite, jusqu'à avoir remplacé les m anciens axes par m nouveaux axes (les axes factoriels), portant des parts décroissantes de la dispersion initiale et dont les 2, 3 ou 4 premiers suffisent souvent à donner une image à peine déformée du nuage initial. C'est cette image **réduite donc beaucoup plus accessible à notre observation** que nous examinerons pour décrire et analyser les données du tableau initial.

3) Les facteurs à retenir :

Trois règles sont applicables :

- 1ere règle : la règle de Kaiser qui veut qu'on ne retienne que les facteurs aux valeurs propres supérieures à 1.
- 2eme règle : on choisit le nombre d'axe en fonction de la restitution minimale d'information que l'on souhaite. Par exemple, on veut que le modèle restitue au moins 80% de l'information.

Pour ces deux premières règles, on examine le tableau « Total Variance Explained ».

- 3eme méthode : le « Scree-test » ou test du coude. On observe le graphique des valeurs propres et on ne retient que les valeurs qui se trouvent à gauche du point d'inflexion. Graphiquement, on part des composants qui apportent le

moins d'information (qui se trouvent à droite), on relie par une droite les points presque alignés et on ne retient que les axes qui sont au dessus de cette ligne.

4) Contribution absolue :

La contribution absolue (CTR) du point "i" à l'inertie des projections sur l'axe α est définie par

$$\mathrm{CTA}(i,\alpha) = \frac{p_i (C_i^{\alpha})^2}{\lambda_{\alpha}}$$

Avec:

- λ_{α} : la valeur propre de l'axe α

-Pi: le point de l'individu i ($p_i = \frac{1}{n}$ où n est le nombre des individus).

- C_i^{α} : la coordonnée de l'individu i sur la composante principale α .

La contribution d'un individu à une composante est la part de la variance d'une composante principale qui provient d'un individu donné. Si cette contribution est supérieure de 2 à 4 fois à son poids ($\frac{p_i (C_i^{\alpha})^2}{\lambda_{\alpha}} \geq \beta p_i, 2 \leq \beta \leq 4$) l'individu définit la composante. Si elle est tr`es supérieure aux autres, on dit qu'il est surreprésenté et on peut avoir intérêt à mettre l'individu en donnée supplémentaire.

5) Qualité globale de la représentation:

L'inertie totale du nuage des individus (notée I(n)) est égale au nombre de variables de départ dans le cas d'une analyse sur données centrées-réduites.

En effet, On a:

$I(n) = \lambda_1 + \lambda_2 + ... + \lambda_p = \mathrm{Tr}(R) = p$ (TR=La trace d'une matrice est la somme de ses valeurs propres)

Le pourcentage d'inertie (ou "variance" du nuage ou "dispersion") expliquée par un axe factoriel permet d'évaluer en quelque sorte la quantité d'information recueillie par cet axe. Notons que l'inertie expliquée par un axe est égale à la *valeur propre* correspondante et que

la qualité de la représentation obtenue par k valeurs propres est mesurée par la proportion de l'inertie expliquée, elle vaut:

$$\frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

Le taux d'inertie absorbée par le premier plan est donnée par : $\frac{\lambda_1 + \lambda_2}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_1 + \lambda_2}{p}$

Si par exemple $\lambda_1 + \lambda_2$ est égal 90% de I(n), on en déduit que le nuage de points est aplati autour du premier plan principal.

Le taux d'inertie définit le pouvoir explicatif d'un facteur.

Suite de l'application:

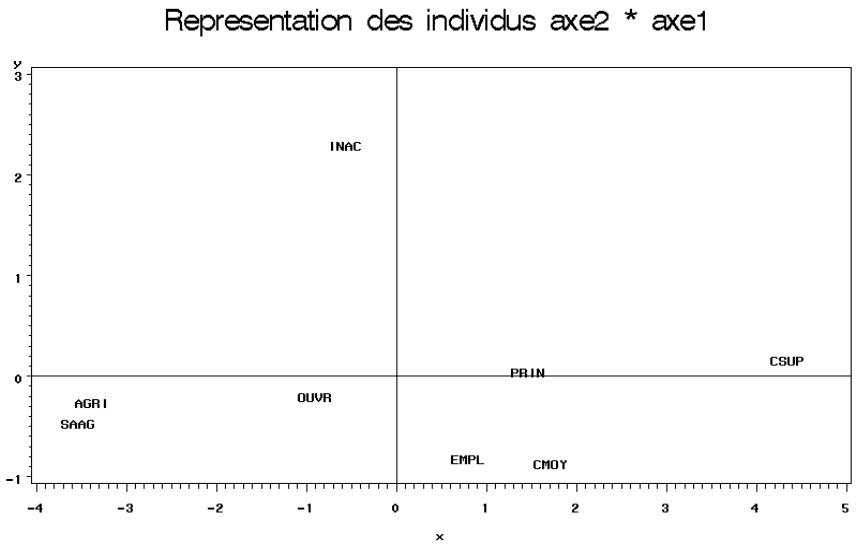
L'ACP sur variables numériques permet d'obtenir les valeurs propres suivantes :

	val.prop.	pct.var	pct.cum
1	6.20794684	0.7760	0.7760
2	0.87968139	0.1100	0.8860
3	0.41596112	0.0520	0.9379
4	0.30645467	0.0383	0.9763
5	0.16844150	0.0211	0.9973
6	0.01806771	0.0023	0.9996
7	0.00344677	0.0004	1.0000
8	0.00000000	0.0000	1.0000

9) **Combien d'axe faut-il retenir? Justifier votre réponse.**
10) **Quel est le pourcentage d'inertie absorbé par le premier plan factoriel?**

On donne ci-dessous les coordonnées des individus, leurs contributions aux axes, ainsi que la représentation graphique des individus sur les deux premiers axes.

CSP	C1	C2	Contr1	Contr2
AGRI	-3.37158	-0.24582	0.22917	0.00868
SAAG	-3.52171	-0.44740	0.25000	0.02875
PRIN	1.47203	0.05851	0.04356	0.00049
CSUP	4.35879	0.17611	0.38150	0.00445
CMOY	1.71808	-0.85665	0.05895	0.10525
EMPL	0.80653	-0.80853	0.10309	0.09392
OUVR	-0.89910	-0.18304	0.01629	0.00481
INAC	-0.56304	2.30681	0.00202	0.76402



11) Pour chacune des 2 premières composantes principales, donner la liste des individus qui contribuent à l'axe de manière significative

Corrigé:

9) La règle de Kaiser conduit à retenir les valeurs propres qui sont supérieures à 1, ce qui signifie ici la première. ($\lambda_1 = 6.20 > 1$)

10) On doit trouver 8 valeurs propres de la matrice des corrélations des variables de dimension 8 (il s'agit de 8 variables décrivant les consommations des CSP)

Le taux d'inertie cumulé des 2 premiers axes est donné par :

(6,20794684+0,87968139)/8=88.59%

Étant donné l'importance du taux d'inertie cumulé par rapport au nombre de variables initial qui est 8. On peut donc retenir les deux premiers axes.

11) On regarde les contributions aux axes, dont on veut qu'elles soient supérieures à 2 fois le poids ($p = \frac{1}{n} = \frac{1}{8}$). Comme il y a 8 individus(CSP), on veut des contributions supérieures à (2x (1/8)=0.25). On fait bien attention de séparer les coordonnées positives des coordonnées négatives en se reportant à la projection des individus.

$$contr \geq 2 \times p \quad avec \; p = \frac{1}{n} = \frac{1}{8}$$

Axe 1 :
***côté négatif : SAAG**
*** côté positif : CSUP;**

Axe 2 :
***côté négatif : aucun**
*** côté positif : INAC**

Interprétation:

- le premier axe oppose les cadres supérieurs (coté positif) à salariés agriculteurs (coté négatif) ce qui permet de dire que ces deux types de CSP n'ont pas le même comportement de consommation des denrées alimentaires.

-le deuxième axe est marqué par les inactifs du coté positif.

IV. Nuage de points variables

1. Corrélation entre composantes et variables initiales

Quand on travaille sur les variables centrées-réduites, la corrélation entre une composante principale c_k et une variable z^j est:

$$r(Z^j, c_k) = \frac{Cov(Z^j, c_k)}{\sqrt{V(c_k)}} = \frac{(Z^j)' Dc_k}{\sqrt{\lambda_k}}$$

et donc le vecteur des corrélations de c_k avec Z est:

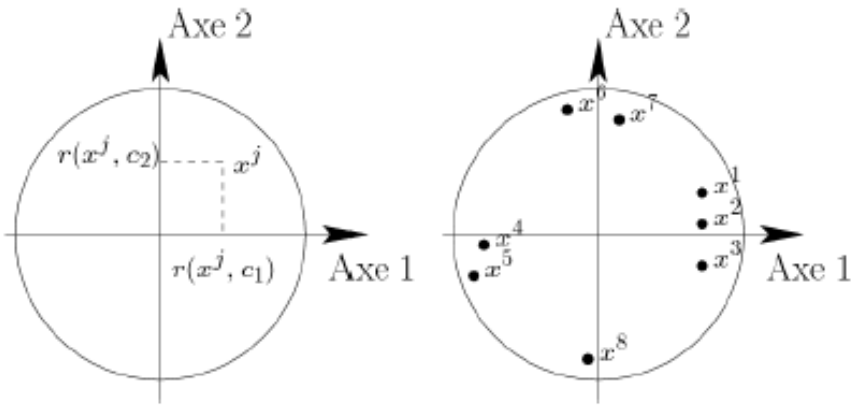
$$r(Z, c_k) = (r(z^1, c_k), \ldots, r(z^p, c_k))' = \frac{Z' Dc_k}{\sqrt{\lambda_k}}$$

$$Comme \quad Z' Dc_k = Z' DZu_k = Ru_k = \lambda_k u_k, on\ a\ \ finalement$$

$$r(Z, c_k) = \sqrt{\lambda}_k u_k$$

2. Cercle de corrélation:

C'est une représentation où, pour deux composantes principales, par exemple c_1 et c_2, on représente chaque variable z^j par un point d'abscisse $r(z^j, c_1)$ et d'ordonnée $r(z^j, c_2)$.



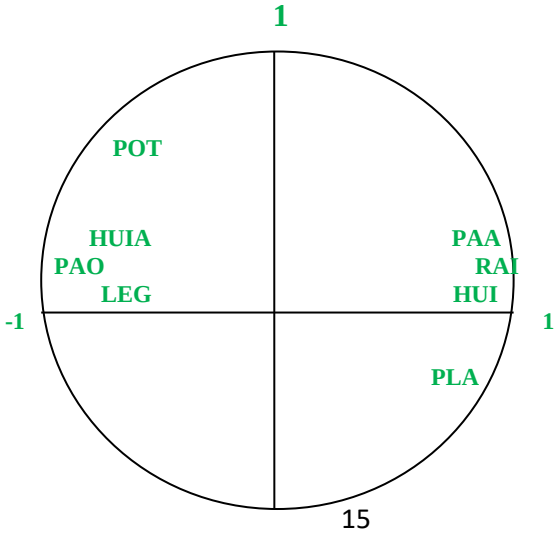
Effet « taille » cela arrive quand toutes les variables sont corrélées positivement avec la première composante principale. Cette composante est alors appelée facteur de « taille », le second facteur de « forme ».

Le cercle des corrélations permet de visualiser comment les variables sont corrélées (positivement ou négativement) avec les composantes principales. `A partir de l`a, on peut soit trouver une signification physique à chaque composante, soit montrer que les composantes séparent les variables en paquets.

Coordonnees des variables sur les axes

Pearson Correlation Coefficients, N = 8

	Prin1	Prin2
PAO	-0.97498	0.12927
PAA	0.86875	0.41323
HUI	-0.87004	0.18916
HUIA	0.93092	0.24415
POT	-0.61385	0.69764
LEC	-0.90898	0.12007
RAI	0.92949	0.30574
PLP	0.90114	-0.04711



Données cibles

- On applique l'ACP sur tous les tableaux rectangulaires numériques (Variables quantitatives), où chaque ligne est **un individu** et chaque colonne est **une variable.**
- Le but est de trouver :
- **La ressemblance** entre les individus.
- **La liaison** entre les variables

Question

- Quand peut-on dire que 2 individus se ressemblent en considérant l'ensemble des variables?
- Comment étudier un large ensemble d'individus?

Réponse:

- Regrouper les individus en partitions.

Objectifs

- L'ACP donc sert à :
- Faire une description des données.
- Explorer les données.
- Visualiser un ensemble important de données à l'aide d'un graphique simple.
- Synthétiser les données.

Nuage des individus

- Ayant K variables (descripteurs), considérons les cas suivants:
- K=1: représentation axiale.
- K=2: nuage de points.
- K=3: représentation en 3D (difficile mais faisable)
- K>3: impossibilité de représentation, néanmoins, le principe est simple.

- Maximiser la distance entre les individus:

$$d^2(i1,i2) = \sum_{k=1}^{K} (x_{i1k} - x_{i2k})^2$$

- Théorème de Pythagore.