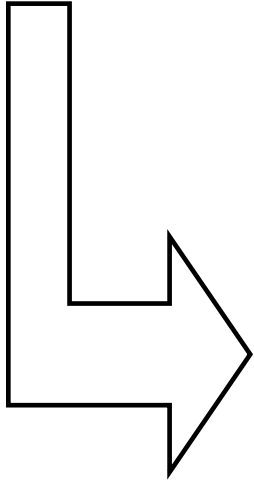


Méthodes d'Analyse des Données : Analyse Multidimensionnelles

Méthodes Factorielles & Méthodes de Classification



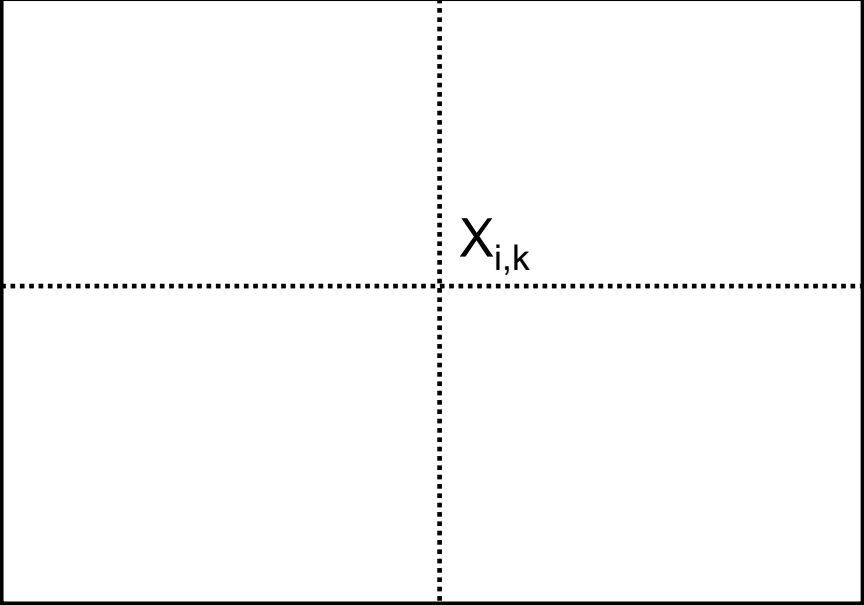
- Analyse en Composantes Principales
- Analyse factorielle des Correspondances
- Analyse factorielle des Correspondances multiples

L'Analyse en Composantes Principales (ACP)

Tableau : Individus , variables quantitatives

$X(n,p)$

	1	k	p
Individu i		$X_{i,k}$	
n			

A diagram of a data matrix $X(n,p)$. It is represented as a rectangle with a solid border. A horizontal dotted line and a vertical dotted line intersect inside the rectangle. The label 'Individu i' is placed to the left of the horizontal line. The label 'X_{i,k}' is placed at the intersection of the two dotted lines. The number '1' is at the top-left corner, 'k' is at the top-right corner, and 'p' is at the top-right corner of the rectangle. The number 'n' is at the bottom-left corner.

n : nombres d'individus, p : nombre de variables.

La distance entre deux individus i et j : $\sum_{k=1}^p (x_{ik} - x_{jk})^2$

La liaison entre deux variables h et k est mesurée par le coefficient de corrélation linéaire :

$$r(h, k) = \frac{\text{cov}(h, k)}{\sqrt{\text{var}(h) \text{var}(k)}} = \frac{1}{n} \sum_{i=1}^n \frac{(x_{ih} - \bar{x}_h)(x_{ik} - \bar{x}_k)}{s_h s_k}$$

$$\bar{x}_h = \frac{\sum_{i=1}^n x_{ih}}{n}$$

$$\bar{x}_k = \frac{\sum_{i=1}^n x_{ik}}{n}$$

$$s_h = \sqrt{\frac{\sum_{i=1}^n (x_{ih} - \bar{x}_h)^2}{n}}$$

$$s_k = \sqrt{\frac{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}{n}}$$

Objectifs de l'ACP :

Identifier des typologies de variables

Variables corrélées positivement, variables corrélées négativement

Identifier des typologies d'individus

Les individus qui se ressemblent, les individus différents

Existe-t-il des groupes homogènes d'individus ?

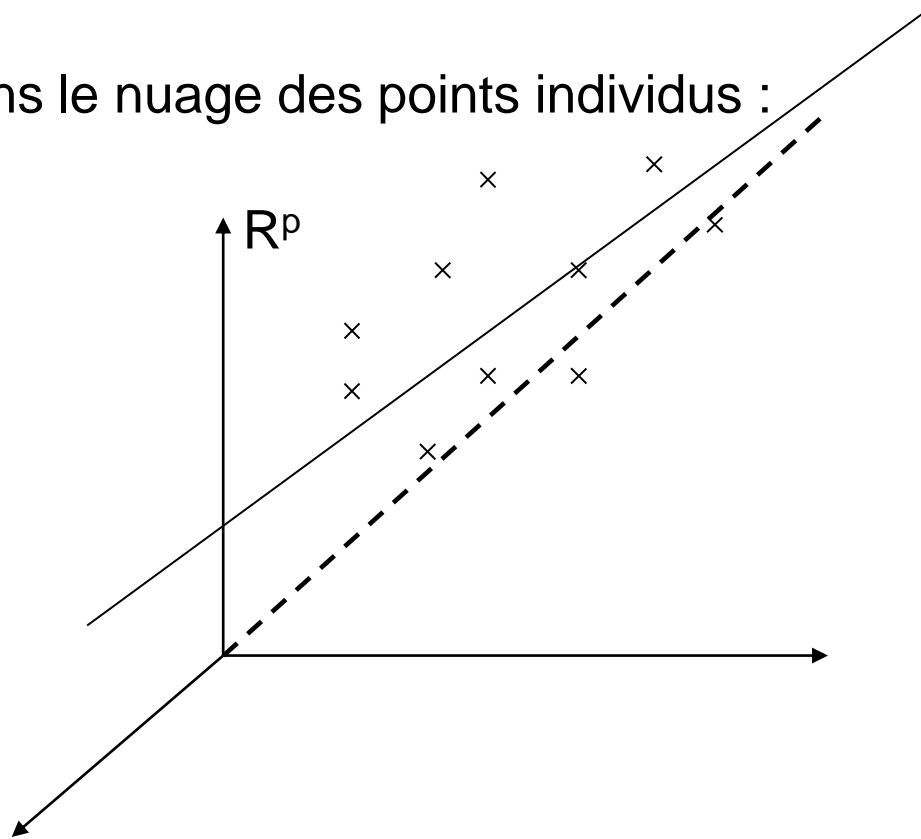
Objectifs de l'ACP :

Réduction de la taille du tableau : $X_{n,p} \rightarrow E$: sous esp de dim q

- Classification des variables selon la qualité de représentation dans E
- Classification des variables selon leurs contributions aux axes de E
- Classification des individus selon la qualité de représentation dans E
- Classification des individus selon leurs contributions aux axes de E

Transformation des données

Considérons le nuage des points individus :



La droite telle que le nuage projeté est proche du nuage initial ne passe pas nécessairement par le centre du repère

D'où le centrage des variables : rendre le centre du repère au centre de gravité du nuage.

Si les variables sont données dans le même système d'unité :

$$X(n,p) \rightarrow Y(n,p)$$

$$y_{i,k} = \frac{x_{i,k} - \bar{x}_k}{\sqrt{n}}$$

L'application de l'analyse factorielle consiste à diagonaliser la matrice $Y'Y$ qui est égale à la matrice des variance empiriques :

$$(Y'Y)_{h,k} = \sum_{i=1}^n Y'_{hi} Y_{ik} = \sum_{i=1}^n Y_{ih} Y_{ik}$$

$$y_{i,k} = \frac{x_{i,k} - \bar{x}_k}{\sqrt{n}} \quad y'_{h,i} = y_{ih} = \frac{x_{i,h} - \bar{x}_h}{\sqrt{n}}$$

$$(Y'Y)_{h,k} = \sum_{i=1}^n \frac{(x_{ih} - \bar{x}_h)(x_{ik} - \bar{x}_k)}{n} = \text{cov}(h, k)$$

Si les variables ont des : $\left\{ \begin{array}{l} \text{des unités différentes} \\ \text{des niveaux très différents} \end{array} \right.$

$$X(n, p) \rightarrow Z(n, p) : z_{i,k} = \frac{x_{i,k} - \bar{x}_k}{\sqrt{ns_k}}$$

L'ACP normée présente des propriétés très intéressantes :

La matrice $Z'Z$ est la matrice des coefficients de corrélation linéaire

$$(Z'Z)_{h,k} = \sum_{i=1}^n Z'_{hi} Z_{ik} = r(h, k)$$

$$z_{i,k} = \frac{x_{i,k} - \bar{x}_k}{\sqrt{ns_k}} \quad z'_{h,i} = \frac{x_{i,h} - \bar{x}_h}{\sqrt{ns_h}}$$

$$(Z'Z)_{h,k} = \sum_{i=1}^n \frac{(x_{ih} - \bar{x}_h)(x_{ik} - \bar{x}_k)}{ns_h s_k} = r(h, k)$$

L'analyse en composantes principales normée, consiste à faire l'analyse factorielle sur le tableau $Z(n,p)$

Recherche des valeurs propres et vecteurs propres de la matrice de corrélation linéaire : $Z'Z$

L'ACP normée

Le nuage des individus

La distance entre deux individus i et j : $\sum_{k=1}^p (z_{ik} - z_{jk})^2$

La projection d'un individu i sur l'axe u : $Z_i u$

Z_i étant la i ème ligne du tableau $z(n,p)$ et u vecteur propre unitaire de $Z'Z$

Le nuage des variables

La norme de chaque variable z_k est égale à 1, en effet :

$$\|z_k\|^2 = \sum_{i=1}^n \left(\frac{x_{ik} - \bar{x}_k}{\sqrt{ns_k}} \right)^2 = \frac{s_k^2}{s_k^2} = 1$$

Les variables centrées réduites appartiennent à la sphère de rayon 1

La distance au carré entre deux variables z_k et z_l est :

$$d(z_k, z_l)^2 = \sum_{i=1}^n \left(\frac{x_{ik} - \bar{x}_k}{\sqrt{ns_k}} - \frac{x_{il} - \bar{x}_l}{\sqrt{ns_l}} \right)^2 = 2(1 - r(k, l))$$

Si $r(k,l) = 1$, alors $d(z_k, z_l)^2 = 0$: les variables z_k et z_l sont confondues.

Si $r(k,l) = -1$ alors $d(z_k, z_l)^2 = 4$: les variables z_k et z_l sont diamétralement opposées.

Si $r(k,l) = 0$, alors $d(z_k, z_l)^2 = 2$: z_k et z_l appartiennent à des diamètres perpendiculaires.

Propriétés

Pour chaque vecteur propre unitaire v_α de la matrice ZZ' :

Moyenne(v_α) = 0 et Variance(v_α) = $1/n$

$$\text{En effet : } v_\alpha = \frac{1}{\sqrt{\mu_\alpha}} Z u_\alpha \Rightarrow v_{\alpha i} = \frac{1}{\sqrt{\mu_\alpha}} \sum_{h=1}^p z_{ih} u_{\alpha h}$$
$$\sum_{i=1}^n v_{\alpha i} = \frac{1}{\mu_\alpha} \sum_{i=1}^n \sum_{h=1}^p z_{ih} u_{\alpha h} = \frac{1}{\mu_\alpha} \sum_{h=1}^p \left(\sum_{i=1}^n z_{ih} \right) u_{\alpha h} = 0$$

$$V(v_\alpha) = \frac{\sum_{i=1}^n v_{\alpha i}^2}{n} = \frac{v_\alpha' v_\alpha}{n} = \frac{1}{n} \quad \text{Le vecteur } v_\alpha \text{ étant unitaire}$$

$v'z_k = r(v, z_k)$: la projection de chaque variable z_k sur un vecteur propre unitaire v est égale au coefficient de corrélation entre v et z_k

$$\text{cor}(v_\alpha, z_k) = \frac{\text{cov}(v_\alpha, z_k)}{\sqrt{V(v_\alpha)} \sqrt{V(z_k)}} = \frac{1}{n} \frac{\sum_{i=1}^n v_{\alpha i} z_{ki}}{\frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}}} = \sum_{i=1}^n v_{\alpha i} z_{ki} = v_\alpha' z_k$$

Il est possible de montrer également que $\text{cor}(z_k, z_l) = z'_k z_l$

Le nuage des points variables et celui des points individus ont la même inertie :

$$\sum_{k=1}^p \lambda_k = \sum_{i=1}^n \mu_i = p$$

F_λ vecteur des projections des individus sur l'axe u associé à la valeur propre λ : $F_\lambda = Zu$

G_λ vecteur des projections des variables sur l'axe v associé à la même valeur propre λ : $G_\lambda = Z'v$

$$\text{Or } u = \frac{1}{\sqrt{\lambda}} Z'v \Rightarrow Zu = \frac{1}{\sqrt{\lambda}} ZG_\lambda \Rightarrow F_\lambda = \frac{1}{\sqrt{\lambda}} ZG_\lambda$$

$$F_\lambda(i) = \frac{1}{\sqrt{\lambda}} \sum_{k=1}^p Z_{i,k} G_\lambda(k) = \frac{1}{\sqrt{\lambda}} \sum_{k=1}^p \left(\frac{X_{i,k} - \bar{X}_k}{s_k \sqrt{n}} \right) G_\lambda(k)$$

De même $G_{\lambda} = \frac{1}{\sqrt{\lambda}} Z' F_{\lambda} \Rightarrow G_{\lambda}(k) = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^n Z'_{k,i} F_{\lambda}(i)$

$$G_{\lambda}(k) = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^n \left(\frac{X_{i,k} - \bar{X}_k}{s_k \sqrt{n}} \right) F_{\lambda}(i)$$

Aide à l'interprétation pour l'ACP normée

Inertie liée aux facteurs

Chaque facteur : vecteur propre \rightarrow valeur propre (inertie portée par le facteur)

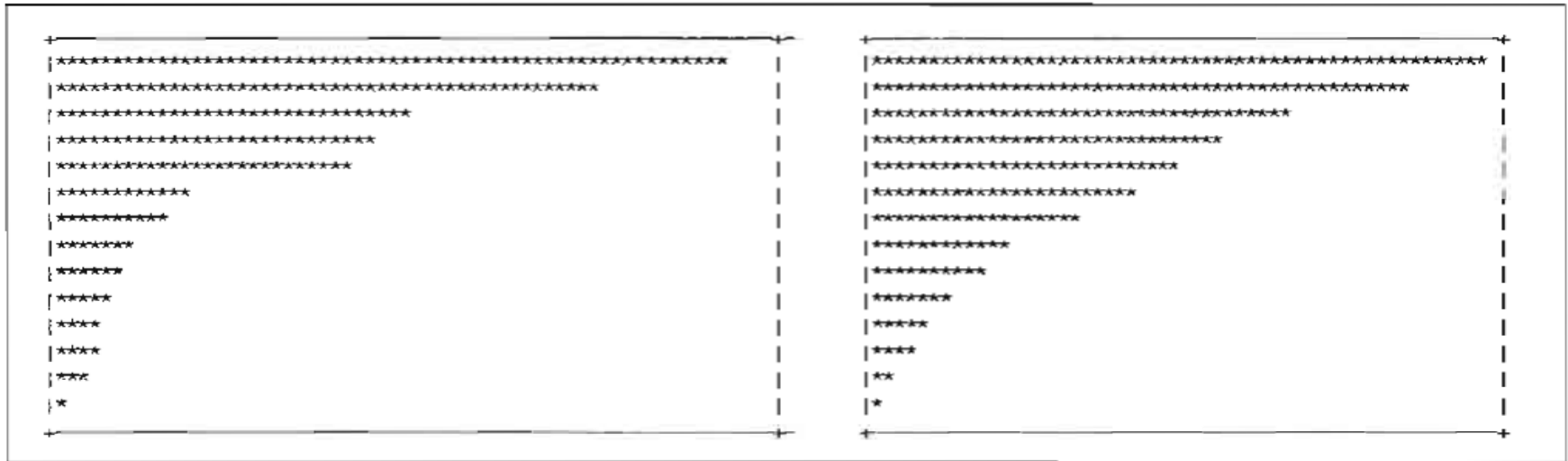
Pourcentage d'inertie d'un axe donné :
$$\frac{\lambda_k}{\sum_{h=1}^p \lambda_h} = \frac{\mu_k}{\sum_{j=1}^n \mu_j}$$

Le pourcentage d'inertie d'un sous espace est la somme des pourcentages d'inertie de chaque axe qui le constitue

Comment choisir la dimension du sous espace ?

- Pourcentage d'inertie du sous espace
- Qualité de représentation ($\cos(\theta)^2$) des points individus et des points variables

Exemples de décroissance des variables propres



(1) Paliers dans la décroissance des valeurs propres

(2) Décroissance régulière des valeurs propres

Dimension du sous espace :

(1) : les axes dont les inerties décroissent d'une manière irrégulière

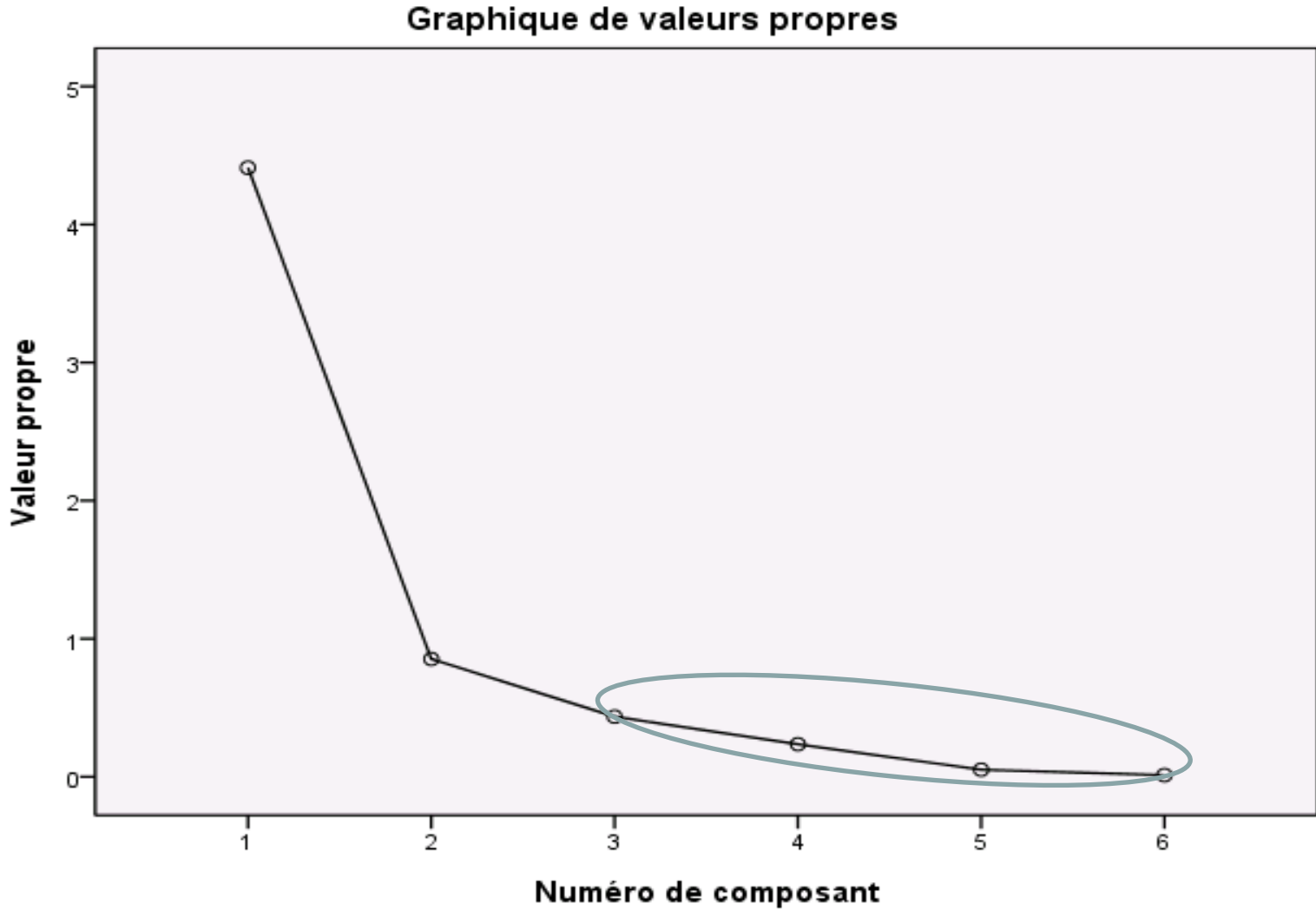
Dim E : nbre de val propres qui décroissent d'une manière irrégulière

(2) : les variables sont modérément corrélées et la forme du nuage des points est régulière, l'apport de l'analyse factorielle dans ce cas n'est pas très significatif. Il est possible de considérer les axes dont les inerties sont supérieures à 1

Dim E : nbre de val propres supérieures à 1

1 étant la moyenne des valeurs propres dans le cas de l'ACP normée

Aspect quantitatif dans la détermination de la dimension de E



Idée : éliminer les val propres faibles
presque égales

$$\text{R\`egle : } \frac{\text{Inertie Intra}}{\text{Inertie Totale}} < 0,05$$

Nuage des variables

La qualité de projection de la variable K sur l'axe v_α :

$$\cos^2(K) = \text{cor}(K, v_\alpha)^2$$

En ACP normée, toutes les variables sont de norme 1

Les variables fortement corrélées à un axe vont contribuer à la définition de cet axe, en effet, la contribution de la variable k à l'axe v_α :

$$\text{ctr}(k, v_\alpha) = \frac{(v_\alpha' Z_k)^2}{\lambda_\alpha} = \frac{\text{cor}(k, v_\alpha)^2}{\lambda_\alpha}$$

Les variables qui contribuent le plus à la définition d'un axe, sont celles dont les projections sur cet axe sont les plus élevées.

Pour un axe donné, on distingue :

- Les variables fortement corrélées positivement
- Les variables fortement corrélées négativement

Dans la projection d'un nuage de variables dans un plan factoriel, seules les variables proches du cercles unité peuvent être interprétées.

Les variables assez éloignées du cercle unité, sont mal représentées sur ce plan factoriel et donc ne peuvent être interprétées.

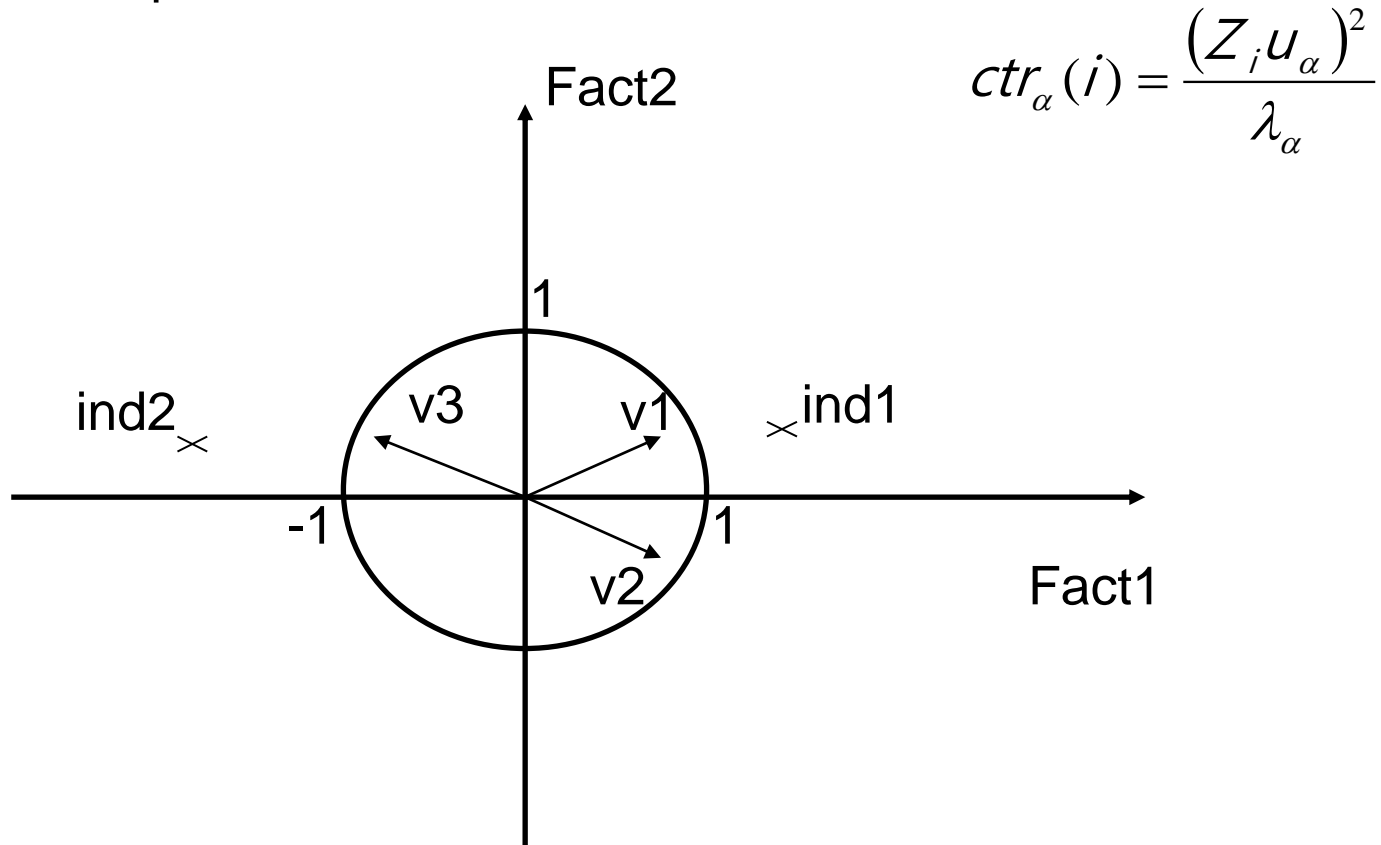
Ces variables sont caractérisées par des projections faibles sur les deux axes du plan factoriel.

Nuage des individus

Qualité de projection d'un individu i sur l'axe u_α :

$$\cos(i)^2 = \frac{(Z_i u_\alpha)^2}{\|Z_i\|^2}, \text{ } Z_i \text{ étant la } i\text{ème ligne du tableau } Z$$

Les individus ayant les plus grandes projections sur un axe donné participent le plus à sa définition :



Interprétations

Ind1 a une projection assez importante sur Fact1

Ind2 a une projection assez importante (négative) sur Fact1

On dit que Ind1 et Ind2 sont opposés selon l'axe Fact1

Fact1 est corrélé positivement avec V1 et V2 et corrélé négativement avec v3

Ind1 présente des valeurs élevées pour V1 et V2, et des valeurs faibles pour V3

Ind2 présente des valeurs faibles pour V1 et V2, et des valeurs élevées pour V3

Individus et Variables supplémentaires

X	X^+
X_+	

Il est possible d'ajouter au tableau des données X :

- des individus supplémentaires X_+
- des variables supplémentaires X^+

Transformation des données ajoutées :

Individu supplémentaire :
$$Z_{+i,j} = \frac{x_{+i,j} - \bar{x}_j}{\sqrt{ns_j}}$$

Variables supplémentaire :
$$Z_{i,j}^+ = \frac{x_{i,j}^+ - \bar{x}_j^+}{\sqrt{ns_j^+}}$$

\bar{x}_j^+ et s_j^+ étant respectivement la moyenne et l'écart types de la variable j ajoutée.

La projection d'un individu ajouté sur l'axe u : $Z_+(i)u$

La projection d'une variable ajoutée sur l'axe v : $Z^+(j)v$

Variable qualitative supplémentaire

Il est possible d'ajouter une variable qualitative à m modalités :

chaque modalité est un individu supplémentaire où il est possible de calculer la moyenne de chaque variable

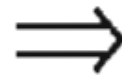
variables
continues
actives

variable nominale
supplémentaire
à 2 modalités

modalité 1
(homme)

modalité 2
(femme)

	taille	poids	sexe
1	150	45	2
	168	68	1
	175	72	1
	178	70	2
i	185	70	1
	160	53	2
	165	49	2
	180	90	1
	175	65	2
10	174	72	2



taille	poids
168	68
175	72
185	70
180	90

taille	poids
150	45
178	70
160	53
165	49
175	65
174	72



lignes
supplém.

177	75
167	59

177	75
-----	----

167	59
-----	----

Exemple : la variable sexe à deux modalités

Application de l'ACP normée

$$X_{n,p} \longrightarrow Z_{n,p} \text{ (Centrage et réduction)}$$

On calcule les valeurs propres et vecteurs propres (u) de $Z'Z$

On détermine la dimension du sous espace : E

On calcule les vecteurs propres (v) de ZZ'

Vecteur des projections des variables sur v : $Z'v$

Vecteur des projection des individus sur u , Zu

Appréciation de la qualité de projection des variables sur le sous espace

Appréciation de la qualité de projection des individus sur le sous espace

Identifier les variables qui contribuent le plus à chaque axe

Identifier les individus qui contribuent le plus à chaque axe

Statistiques descriptives

	Moyenne	Ecart-type	n analyse
Cylindrée	2722,54	1516,445	24
Puissance	206,67	155,721	24
Vitesse	214,71	56,572	24
Poids	1486,58	387,507	24
Largeur	1838,42	220,842	24
longueur	4277,83	581,497	24

Qualité de représentation

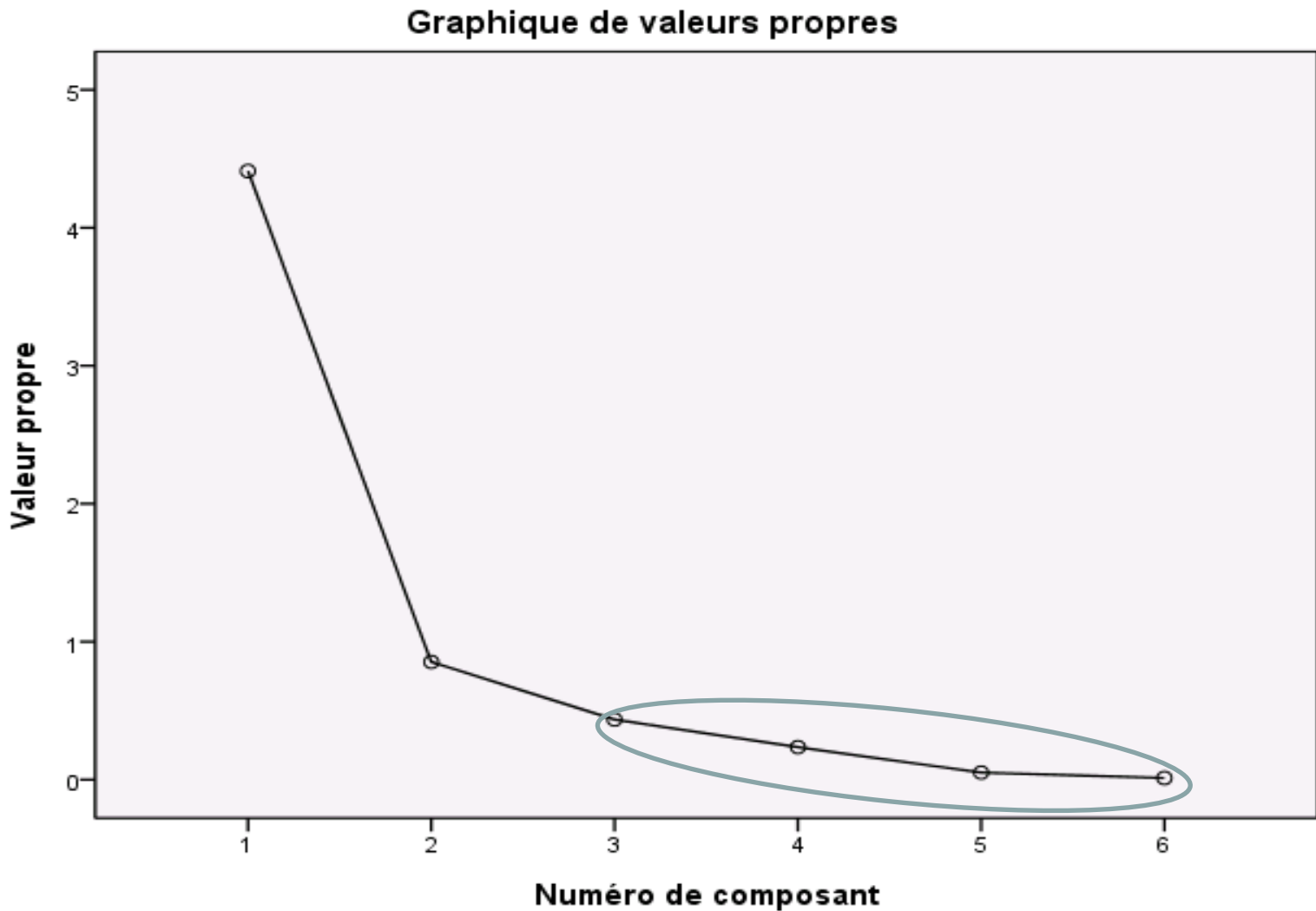
	Initial	Extraction
Cylindrée	1,000	,942
Puissance	1,000	,977
Vitesse	1,000	,900
Poids	1,000	,904
Largeur	1,000	,654
longueur	1,000	,887

Méthode d'extraction : Analyse en composantes principales.

Variance totale expliquée

Composante	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	4,411	73,521	73,521	4,411	73,521	73,521
2	,853	14,223	87,745	,853	14,223	87,745
3	,436	7,261	95,006			
4	,236	3,931	98,937			
5	,051	,857	99,794			
6	,012	,206	100,000			

Méthode d'extraction : Analyse en composantes principales.



Choix de la dimension du sous espace : Règle : $\frac{Inertie\ Intra}{Inertie\ Totale} < 0,05$

Test de Kaiser Meyer Olkin KMO

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2}$$

r_{ij} coefficient de corrélation linéaire entre la variable i et la variable j

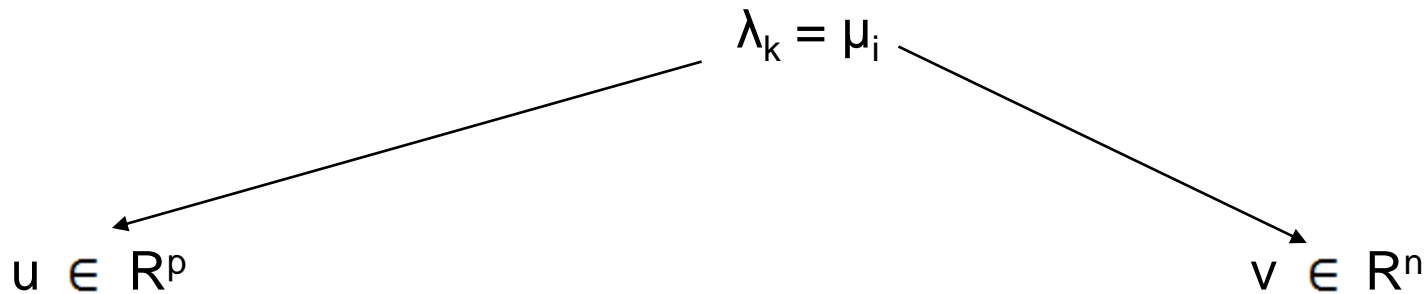
a_{ij} coefficient de corrélation linéaire partiel entre la variable i et la variable j

KMO	Évaluation	KMO	Evaluation
0,9	Merveilleux	0,6	Médiocre
0,8	Méritoire	0,5	Misérable
0,7	Moyen	Moins de 0,5	Inacceptable

$$msa_i = \frac{\sum_j r_{ij}^2}{\sum_j r_{ij}^2 + \sum_j a_{ij}^2}$$

Les matrices $Z'Z$ et ZZ' ont les même valeurs propres non nulles.

Cette propriété permet la représentation simultanée des points individus et des points variables dans un même plan factoriel



Chaque plan factoriel de R^p ou de R^n est associé aux mêmes valeurs propres

$$u = \frac{1}{\sqrt{\lambda}} Z' v \text{ et } v = \frac{1}{\sqrt{\lambda}} Z u$$