Analyse Des Données

Ouazza Ahmed

École Supérieure de Management, d'Informatique et de télécommunications
SUP MTI

2022-2023

Plan

- Introduction
- Analyse en Composante Principales ACP
- Analyse Factorielle des Correspondances AFC
- Méthodes de classification
- Analyse des Correspondances Multiples ACM

Introduction

• Dans toute étude statistique, la première démarche consiste à décrire et explorer l'ensemble des données qu'on dispose.

• Dans le cas où des données sont de grande dimension (c-à-d le nombre de variables ou de caractères est trop élevé), il est difficile de les visualiser. Alors il est nécessaire d'extraire l'information pertinente contenue dans l'ensemble des données, les techniques d'analyse des données répondent à ce besoin.

• Par conséquent, on peut définir l'analyse des données comme ensemble de méthodes descriptives ayant pour objectif de visualiser et résumer l'information contenue dans un grand tableau de données.

Introduction

Tableau de données: (Individu $\times$ Variable)

On suppose qu'on dispose p variables $X_1, X_2, ..., X_p$ observées sur n individus $I_1, I_2, ..., I_n$.

	X_1	...	X_j	...	X_p
I_1	x_{11}	...	x_{1j}	...	x_{1p}
.	.	...	.	...	.
.	.	...	.	...	.
.	.	...	.	...	.
I_i	x_{i1}	...	x_{ij}	...	x_{ip}
.	.	...	.	...	.
.	.	...	.	...	.
.	.	...	.	...	.
I_n	x_{n1}	...	x_{nj}	...	x_{np}

Introduction

La forme matricielle du tableau précédent est donnée comme suit:

$$X = (X_1, ..., X_p) = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

Introduction

Tableau de données: (Variable $\times$ Variable)

Dons ce cas, on cherche à croiser deux variables X et Y d'une population en dénombrant l'effectif correspondant à la conjonction «variable 1» et «variable 2», on parle de tableau de contingence:

X \ Y	y_1	...	y_j	...	y_K
x_1	n_{11}	...	n_{1j}	...	n_{1K}
.	.	...	.	...	.
.	.	...	.	...	.
.	.	...	.	...	.
x_i	n_{i1}	...	n_{ij}	...	n_{iK}
.	.	...	.	...	.
.	.	...	.	...	.
.	.	...	.	...	.
x_L	n_{L1}	...	n_{Lj}	...	n_{LK}

Analyse en Composante Principales ACP

ACP

• Analyse en composante principale ACP est une technique des statistiques descriptives destinée à l'analyse des données multidimensionnelles.

• Analyse en composante principale permet d'explorer **les liaisons entre variables** et **les ressemblances entre individus**.

Objectifs:

• L'ACP permet de réduire la dimension de l'espace des variables avec le minimum de perte d'information.(c-à-d on passe de $X_1, \cdots, X_p$ à $F_1, \cdots, F_k$ avec $k << p$, généralement $k = 2$ ou 3)

• Visualiser le positionnement des individus les uns par rapport aux autres (ressemblance) $\Rightarrow$ Notion de distance entre individus.

• Visualiser les corrélations entre les variables.

• Donner une interprétation aux facteurs (axes).

Les données:

Les données pour l'ACP sont généralement présentées sous la forme du tableau décrit dans les pages 4 et 5:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

- Les variables $X_1, ..., X_p$ sont supposées **quantitatives**.
- x_{ij} est la valeur de la variable X_j pour l'individu I_i.
- n désigne le nombre d'individus (observations) et p le nombre de variables.

ACP

Exemple du tableau des données pour ACP:

On considère les notes (de 0 à 20) obtenues par $n = 9$ étudiants dans $p = 5$ matières,

Etudiant	Math	Stat	Fran	Angl	Musique
I_1	6	6	5	5,5	8
I_2	4,5	5	7	7	14
I_3	6	7	11	9,5	11
I_4	14,5	14,5	15,5	15	5
I_5	14	14	12	12,5	6
I_6	16	15	5,5	5	7
I_7	5,5	7	14	11,5	10
I_8	13	12,5	8,5	9,5	12
I_9	6	8,5	13,5	13	18

Poids des individus:

On affecte à chaque individu I_i un poids p_i reflétant son importance par rapport aux autres individus, avec $0 < p_i < 1$ et $\sum_{i=1}^{n} p_i = 1$.

Le cas le plus fréquent est de considérer que tous les individus ont la même importance c-à-d : $p_i = \frac{1}{n}$

Poids des variables:

De la même manière, on peut définir des poids pour les variables, en affectant à chaque variable X_j un poids m_j reflétant son importance par rapport aux autres.

Le cas le plus fréquent est de considérer que tous les variables ont la même importance c-à-d : $m_j = \frac{1}{p}$

Le vecteur G des moyennes arithmétiques de chaque variable $G = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_p)'$ définit le point moyen, ou centre de gravité du nuage.

Notion de ressemblance: Critère de la distance Euclidienne

Définition 0.1

Deux individus se ressemblent s'ils possèdent des valeurs proches pour l'ensemble des variables.

Donc on parle d'une notion de proximité qui se traduit par une distance.

Ainsi, nous définissons la distance euclidienne entre deux individus I_i et I_j par :

$$d^2(I_i, I_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

Notion de ressemblance: Critère de la distance Euclidienne

Exemple:

Etudiant	Math	Stat	Fran	Angl	Musique
I_1	6	6	5	5,5	8
I_2	4,5	5	7	7	14
I_3	6	7	11	9,5	11
I_4	14,5	14,5	15,5	15	5
I_5	14	14	12	12,5	6
I_6	16	15	5,5	5	7
I_7	5,5	7	14	11,5	10
I_8	13	12,5	8,5	9,5	12
I_9	6	8,5	13,5	13	18

- $d^2(I_1, I_2) = (6-4.5)^2 + (6-5)^2 + (5-7)^2 + (5.5-7)^2 + (8-14)^2 = 45.5$
- $d^2(I_6, I_9) = (16-6)^2 + (15-8.5)^2 + (5.5-13.5)^2 + (5-13)^2 + (7-18)^2 = 391.2$

Liaison entre les variables:

Définition 0.2

Deux variables sont liées si elles ont un fort coefficient de corrélation linéaire.

Le coefficient de corrélation linéaire entre deux variables X_k et X_j est donné par :

$$Corr(X_j, X_k) = \frac{Cov(X_j, X_k)}{S_j S_k} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_{ij} - \overline{x}_j}{S_j} \right) \left(\frac{x_{ik} - \overline{x}_k}{S_k} \right)$$

Exemple:

$Corr(Math, Stat) = 0.982$

$Corr(Math, Fran) = 0.019$

Transformation des données (Problèmes des unités de mesure) :

Parfois, les variables contenues dans le tableau X peuvent être exprimées en différentes unités (cm, kg...)

$\Rightarrow$ Pour neutraliser ce problème des unités on remplace les données d'origine $X_1, ..., X_p$ par les données centrées-réduites. Ainsi chaque variables X_j , $j = 1, ..., p$ est remplacée par:

$$X_j^\star = \frac{X_j - \overline{x}_j}{S_j}$$

avec $\overline{x}_j$ et S_j sont respectivement la moyenne empirique et l'écart-type de la variable X_j.

ACP

Tableau centré-réduit:

	$X_1^\star$	...	$X_j^\star$	...	$X_p^\star$
I_1			.		
.			.		
.			.		
.			.		
I_i	.	...	$\frac{x_{ij}-\overline{x}_j}{S_j}$	...	.
.			.		
.			.		
.			.		
I_n			.		

Tableau centré-réduit:

$$X = (X_1^\star, ..., X_p^\star) = \begin{pmatrix} . & \cdots & . & \cdots & . \\ \vdots & & \vdots & & \vdots \\ . & \cdots & \frac{x_{ij} - \overline{x}_j}{S_j} & \cdots & . \\ \vdots & & \vdots & & \vdots \\ . & \cdots & . & \cdots & . \end{pmatrix}$$

Matrice de Covariance:

La matrice de covariance associée au tableau X est donnée par:

$$V = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_1, X_2) & Var(X_2) & \cdots & . \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_p) & \cdots & \cdots & Var(X_p) \end{pmatrix}$$

Matrice de Corrélation:

La matrice de corrélation associée au tableau X est donnée par:

$$R = \begin{pmatrix} 1 & Cor(X_1, X_2) & \cdots & Cor(X_1, X_p) \\ Cor(X_1, X_2) & 1 & \cdots & . \\ \vdots & \vdots & \ddots & \vdots \\ Cor(X_1, X_p) & \cdots & \cdots & 1 \end{pmatrix}$$

Remarque:

Si le tableau des données est centré-réduit, alors la matrice de covariance est égale à matrice de corrélation: $V = R$

Inertie totale:

• On appelle inertie la quantité d'information contenue dans un tableau de données.

• Une inertie nulle signifie que tous les individus sont presque identiques.

• L'inertie mesure la dispersion totale du nuage du points.

• L'inertie est la somme pondérée des carrés des distances des individus au centre de gravité G.

$$I_G = \frac{1}{n} \sum_{i=1}^{n} d^2(I_i, G)$$

Inertie totale:

L'inertie est aussi égale à la somme des variances des variables étudiées:

$$I_G = \sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} S_j^2 = Trace(V)$$

Dans le cas où X est centré-réduit on a:

$$I_G = Trace(V) = p$$

Nuages de points:

Dans l'ACP, on peut définir deux types de nuages de points:

1) Nuage des individus N_I

2) Nuage des variables N_v

Dans la suite, on suppose que le tableau des données X est centré-réduit.

C-à-d $\overline{x}_j = 0$ et $Var(X_j) = 1$ pour $j = 1, ..., p$

ACP

Avant de réaliser une ACP, on doit:

• Tester l'intérêt de l'ACP en vérifiant s'il est possible de compresser efficacement l'information disponible

• Mesurer le degré de redondance des données

On utilise deux critères:

⇒ Test de sphéricité de Bartlett

⇒ Indice KMO (Kaiser – Mayer – Olkin)

Test de sphéricité de Bartlett

Le test de sphéricité de Bartlett vérifie l'hypothèse nulle selon laquelle toutes les corrélations entre X_j et X_k (avec $j \neq k$) seraient égales à zéro. C-à-d on cherche a tester l'hypothèse suivante:

$$H_0 : R = \text{matrice identité}$$

$\Rightarrow$ c-à-d les variables sont deux à deux indépendantes

Où R est la matrice de corrélation

Remarque:

$\circ$ Si $|R| = 1$, les variables sont deux à deux orthogonales $\Rightarrow$ ACP inutile

$\circ$ Si $|R| = 0$, il y a une colinéarité parfaite $\Rightarrow$ Le premier facteur explique 100% de l'inertie totale

Test de sphéricité de Bartlett

- **Statistique de test:**

$$T = -\left(n - 1 - \frac{2p+5}{6}\right)\ln(|R|)$$

La statistique de test suit une loi du khi-deux χ^2 à $\frac{p(p-1)}{2}$ degrés de liberté.

Test de sphéricité de Bartlett

- **Règle de décision:**

Si $T \geq T_c$ alors on rejette l'hypothèse nulle H_0 au seuil α. Sinon on accepte H_0 au seuil α.

avec $T_c = \chi^2_{1-\alpha}(\frac{p(p-1)}{2})$ est le quantile d'ordre $(1-\alpha)$ d'une loi de χ^2 à $\frac{p(p-1)}{2}$ degrés de liberté.

Remarque:

Une deuxième manière de prendre la décision est de comparer $p-value = "sgnification"$ avec le seuil α:

Si $p-value < \alpha$ alors on rejette H_0 , sinon on accepte H_0.

Exemple: (voir SPSS)

Indice KMO (Kaiser – Mayer – Olkin)

KMO (ou MSA: Measure of sampling adequacy) est un indice d'adéquation de la solution factorielle, il indique jusqu'à quel point l'ensemble de variables retenu est un ensemble cohérent.

L'idée de KMO est de confronter la matrice des corrélations brutes avec la matrice des corrélations partielles.

On considère p variables quantitatives $X_1, \cdots, X_p$: la corrélation entre deux variables X_j et X_k est influencée par les $(p-2)$ autres. La corrélation partielle est utilisée pour mesurer la relation entre deux variables en retranchant l'influence de $(p-2)$ autres.

Indice KMO (Kaiser – Mayer – Olkin)

• Si la corrélation partielle est très faible en valeur absolue ($\simeq 0$),
c-à-d KMO $\simeq 1$, cela veut dire que la liaison est déterminée par les $(p-2)$ autres variables: dans ce cas l'ACP peut agir efficacement (réduction efficace de l'information).

• Si la corrélation partielle est plus élevée en valeur absolue,
c-à-d KMO $<< 1$, alors on conclut qu'il y a une relation directe entre les deux variables: dans ce cas, une compression efficace de l'information n'est pas possible.

Indice KMO (Kaiser – Mayer – Olkin)

- **Corrélation partielle**

On peut définir la corrélation partielle à partir de la matrice corrélation comme suit:

Soit $R^{-1} = (v_{ij})$ $(i, j \in \{1, \cdots, p\})$ la matrice inverse de la matrice de corrélation $R = (r_{ij})$, alors la matrice de corrélation partielle est $R_p = (a_{ij})$ $(i, j \in \{1, \cdots, p\})$ tel que:

$$a_{ij} = -\frac{v_{ij}}{\sqrt{v_{ii} \times v_{jj}}}$$

Indice KMO (Kaiser – Mayer – Olkin)

- **Indice KMO global**

L'indice KMO est définit par la formule suivante:

$$KMO = \frac{\sum_{i=1}^{p} \sum_{j \neq i}^{p} r_{ij}^2}{\sum_{i=1}^{p} \sum_{j \neq i}^{p} r_{ij}^2 + \sum_{i=1}^{p} \sum_{j \neq i}^{p} a_{ij}^2}$$

$KMO \in [0, 1]$, généralement une valeur de KMO de:

- moins de 0.5 est inacceptable
- 0.5 est misérable
- 0.6 est médiocre
- 0.7 est moyenne
- 0.8 est méritoire
- 0.9 est merveilleuse

Exemple: (voir SPSS)

Nuage des individus N_I

Nuage des individus N_I

Chaque individu I_i , $i \in \{1, 2, ..., n\}$ est un élément de $\mathbb{R}^p$

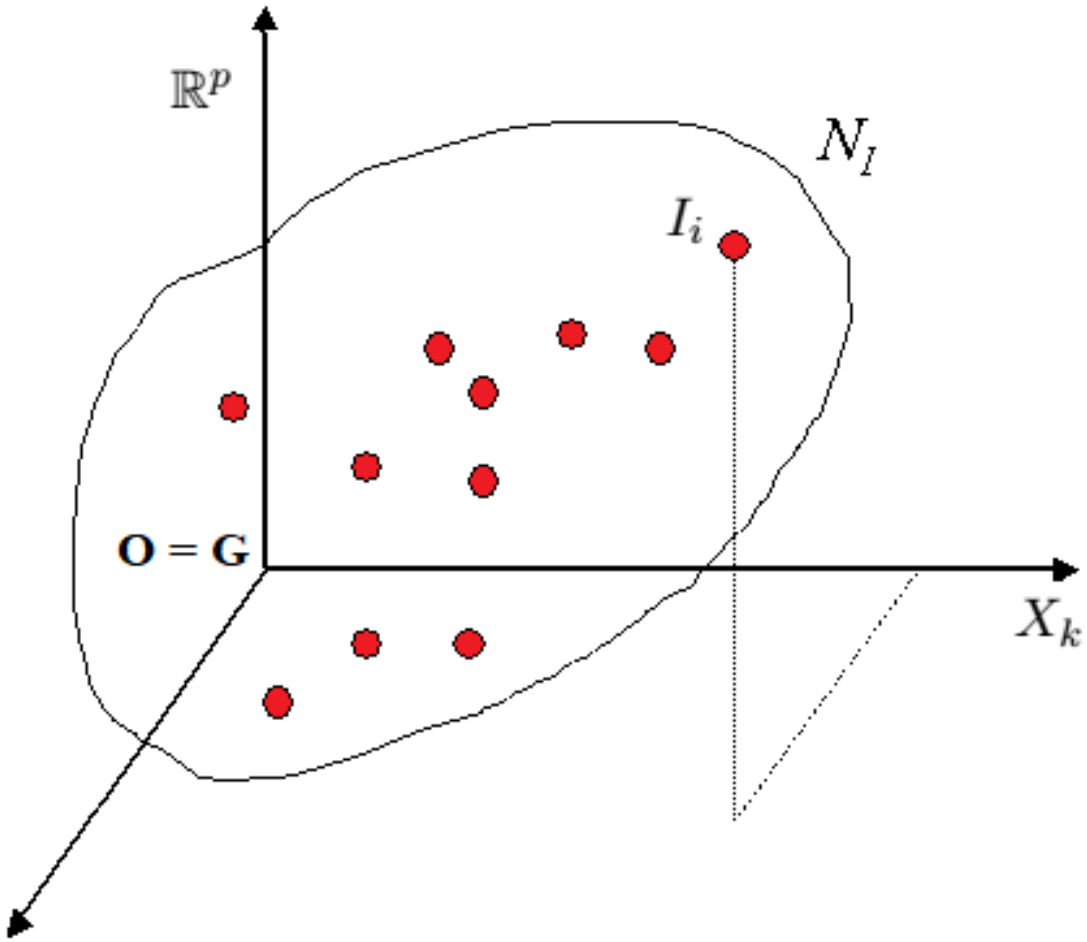


Figure 1: Nuage des individus N_I dans $\mathbb{R}^p$

Nuage des individus N_I

Ajustement du nuage des individus sur une droite U_1 :

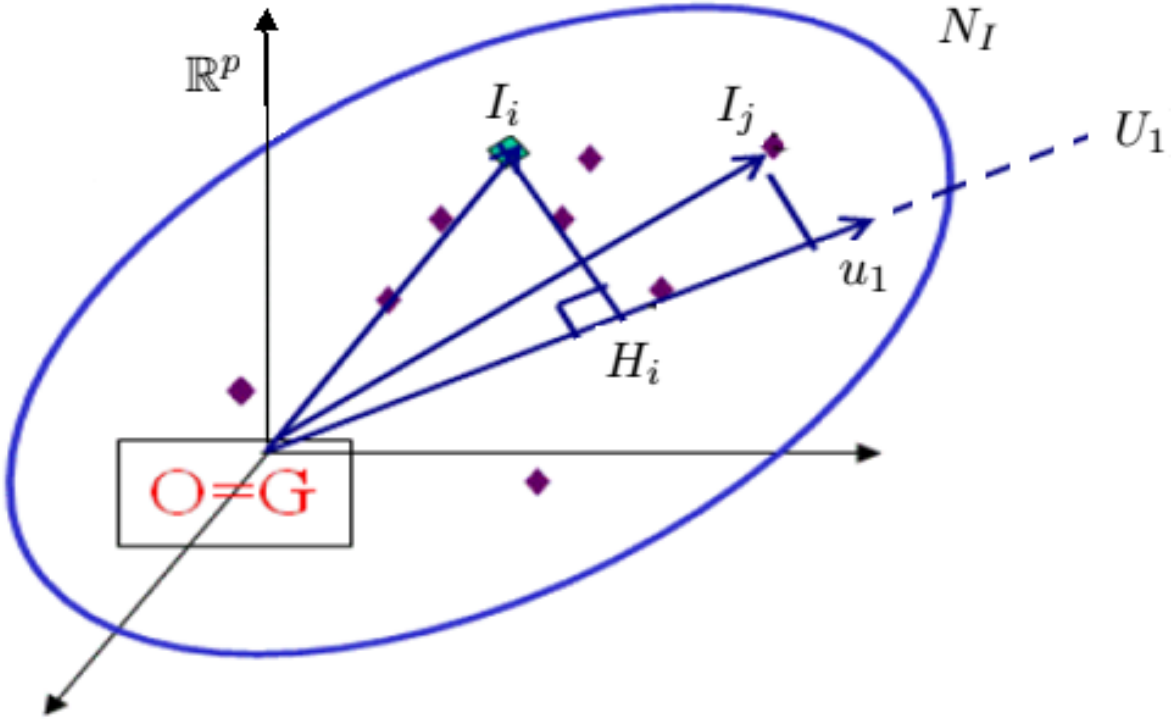


Figure 2: Ajustement du Nuage des individus N_I

Nuage des individus N_I

Ajustement du nuage des individus sur une droite U_1 :

Objectif :

On cherche l'axe u_1 passant le mieux possible au milieu du nuage N_I:
c-à-d on cherche à trouver u_1 tel que l'inertie du nuage N_I projeté sur u_1 soit maximum ,

$$\frac{1}{n}\sum_{i=1}^{n} OH_i^2$$

Ce qui revient à minimiser:

$$I(N_I, u_1) = \frac{1}{n}\sum_{i=1}^{n} d^2(I_i, H_i)$$

$u_1 \Rightarrow$ Axe d'inertie maximum.

On a :

$$d^2(O, I_i) = d^2(O, H_i) + d^2(H_i, I_i)$$

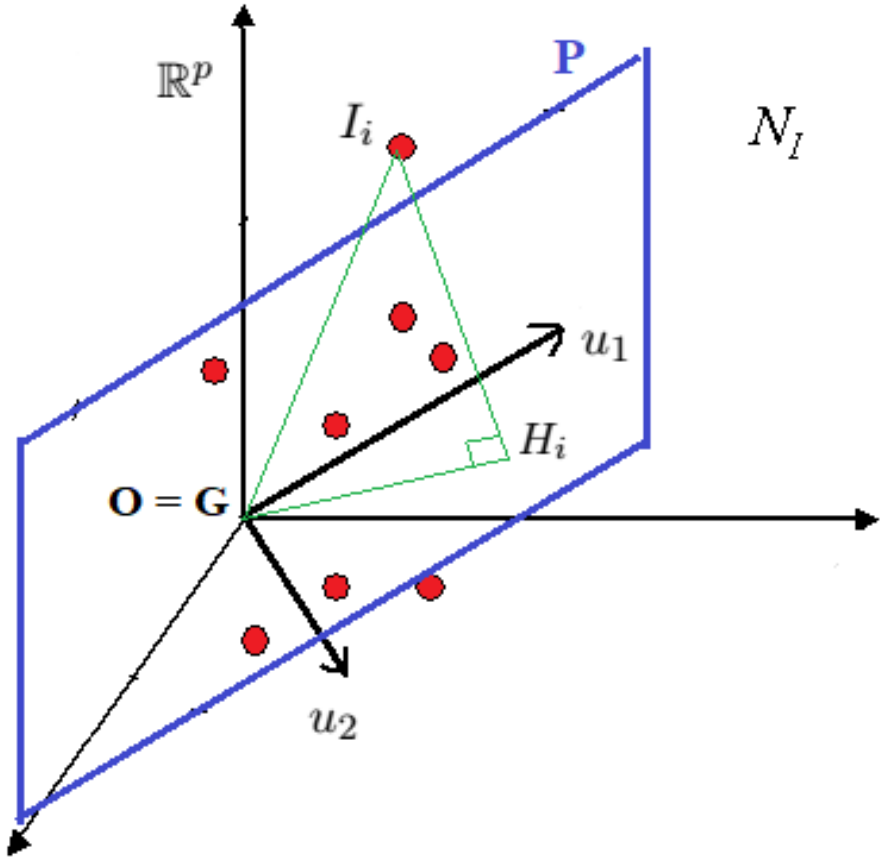
On déduit:

$$\frac{1}{n}\sum_{i=1}^{n} d^2(O, I_i) = \frac{1}{n}\sum_{i=1}^{n} d^2(O, H_i) + \frac{1}{n}\sum_{i=1}^{n} d^2(H_i, I_i)$$

Inertie totale $=$ Inertie expliquée par U_1 $+$ Inertie résiduelle

(Maximiser) (Minimiser)

Ajustement du nuage des individus sur un plan P:



Objectif:

Trouver un plan P tel que $\sum_{i=1}^{n} OH_i^2$ soit maximum.

Avec $u_1 \perp u_2$

$P \Rightarrow$ Plan d'inertie maximum.

Cas général: Ajustement du nuage des individus sur un éspace F_K de dimension K:

Objectif:

Trouver une suite d'axes $\{u_s, s = 1, ..., K\}$ orthogonaux d'inertie maximum.

Avec u_s est un vecteur unitaire de l'axe de rang s.

Soit H_i^s la projection de I_i sur u_s

Donc on cherche u_s tel que $\sum_{i=1}^{n}(OH_i^s)^2$ soit maximum.

Avec la contrainte $u_t \perp u_s$ pour $t < s$

Solution:

u_s est vecteur propre unitaire de la matrice de corrélation R associé à la valeur propre λ_s de rang s .

Définition 0.3

Les K axes factoriels $\{u_s\ ,\ s = 1, ..., K\}$ sont appelées les facteurs principaux.

Exemple:

Pour $K = 1 \Rightarrow$ Ajustement du nuage sur une droite $F_1 = U_1$

Dans ce cas, on a:

• L'axe U_1 passe par le centre de gravité $G = O$ du nuage de points N_I.

• L'axe U_1 est engendré par le vecteur normé u_1, vecteur propre de la matrice des corrélations R associé à la plus grande valeur propre λ_1.

• L'inertie expliquée par l'axe U_1 est égal à λ_1.

• La part d'inertie expliquée par le premier axe principal U_1 est égal à $\frac{\lambda_1}{p}$.

Remarquons que p est aussi égal à la somme des valeurs propes de la matrice de corrélation R.

Pour $K = 2 \Rightarrow$ Ajustement du nuage sur un plan $F_2 = P = (U_1 \times U_2)$

- U_1 et U_2 forment le même plan P
- U_2 perpendiculaire à U_1
- Le deuxième axe principal U_2 passe par le centre de gravité $G = O$ du nuage de points et engendré par le vecteur normé u_2, vecteur propre de la matrice des corrélations R associé à la deuxième plus grande valeur propre λ_2.
- U_2 est centrée, de variance λ_2, et non corrélée avec U_1.
- L'inertie expliquée par l'axe U_2 est égal à λ_2.
- L'inertie expliquée par le plan $P = (U_1 \times U_2)$ est égal à $\lambda_1 + \lambda_2$
- La part d'inertie expliquée par le plan $P = (U_1 \times U_2)$ est égal à $\frac{\lambda_1 + \lambda_2}{p}$.

Nuage des variables N_v

Nuage des variables N_v

On peut envisager le problème de la représentation des variables de façon complètement symétrique de celui des individus.

Ainsi, chaque variable X_j , $j \in \{1, 2, ..., p\}$ est un élément de $\mathbb{R}^n$.

La représentation du nuage N_v des variables se situe dans un espace à n dimensions, chaque dimension représentant un individu de la population totale.

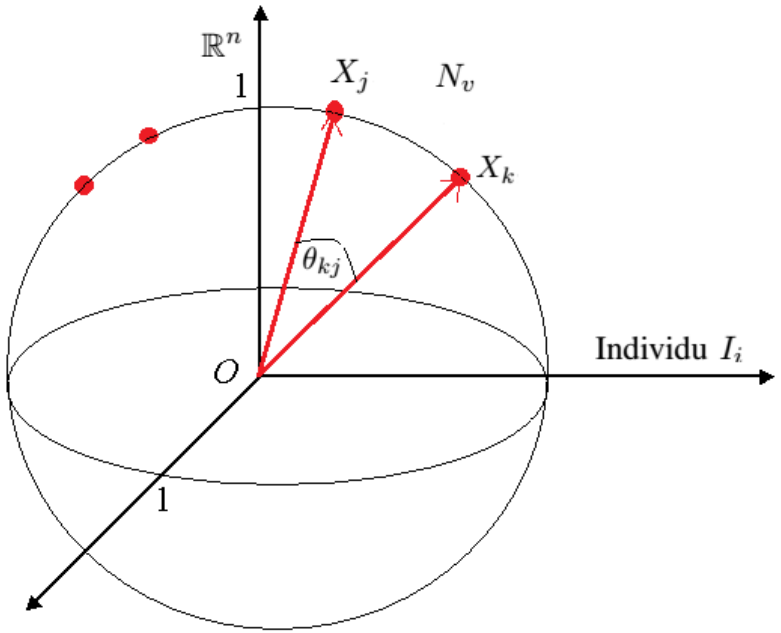


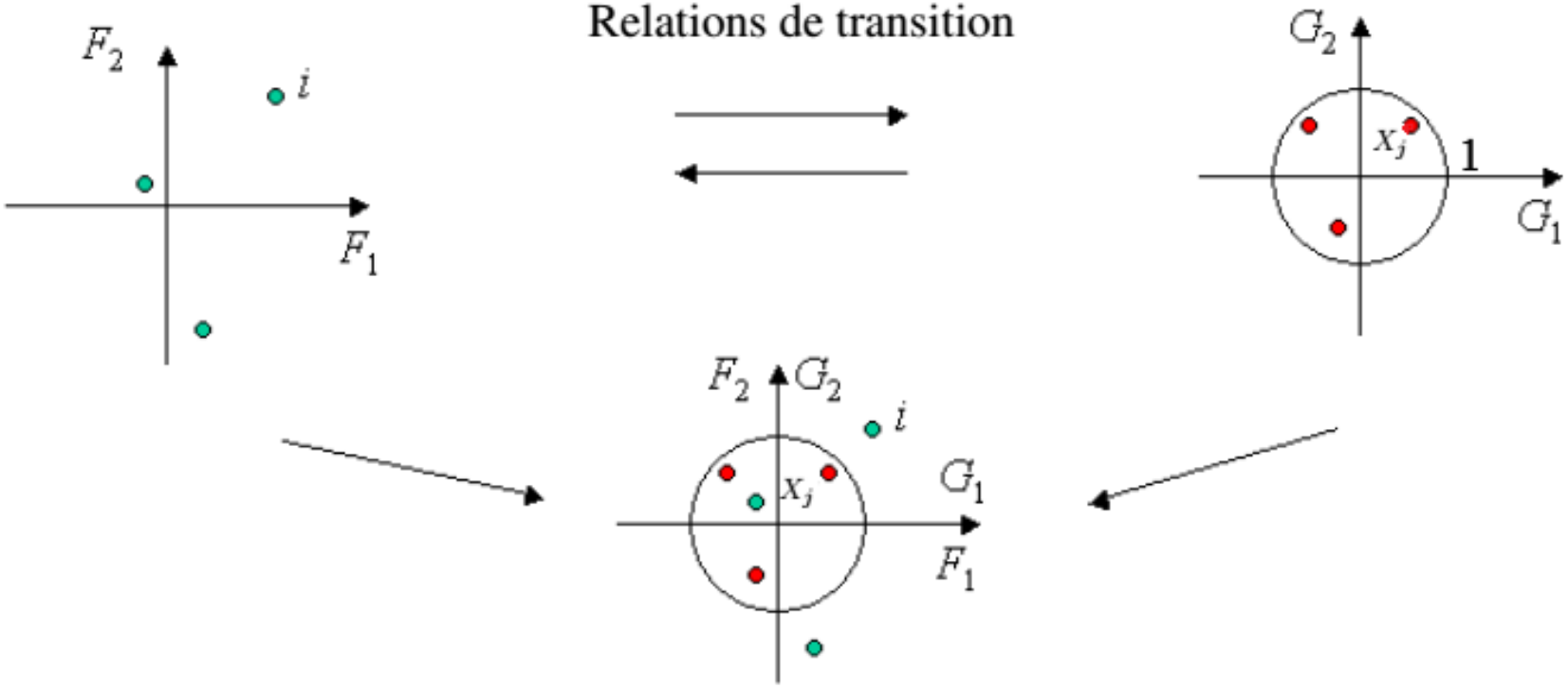
Figure 3: Nuage des variables N_v dans $\mathbb{R}^n$

Représentation simultanée

Nuage des individus N_I

Relations de transition

Nuage des variables N_v



Choix de nombre d'axe

Le but de l'ACP étant d'obtenir une représentation des individus dans un espace de dimension plus faible que p, la question se pose d'apprécier la perte d'information subie et de savoir combien de facteurs à retenir.

Pour choisir le nombre d'axe à retenir, on utilise les critères suivant:

- Critère de Kaisser:

On ne retient que les axes dont l'inertie est supérieur à 1 (inertie moyenne)

- Critère du coude:

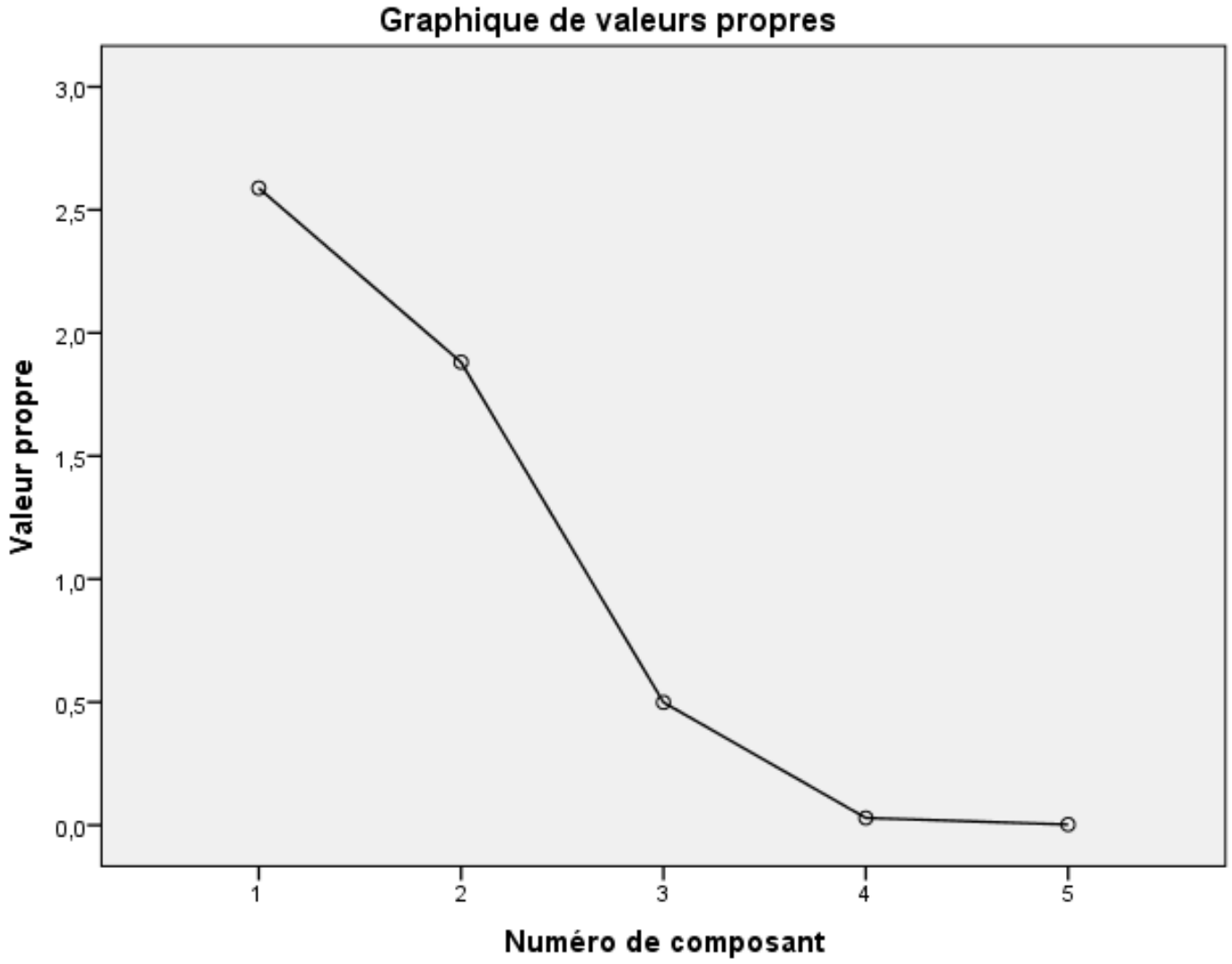
Sur l'éboulis des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière. On sélectionne les axes avant le décrochement.

- Pourcentage d'inertie:

Le pourcentage d'information expliqué par les axes doit être suffisamment grand (par exemple supérieur à 60-70%)

Choix de nombre d'axe

Éboulis des valeurs propres:



Qualité de représentation

Qualité de représentation d'un individu I_i par un axe s:

$$QLT_s(I_i) = \frac{(OH_i^s)^2}{(OI_i)^2} \tag{1}$$

Si $QLT_s(I_i)$ est proche de 1, alors l'individu I_i est bien représentée sur l'axe s.

Si $QLT_s(I_i)$ est proche de 0, alors l'individu I_i est mal représentée sur l'axe s.

Qualité de représentation

Qualité de représentation d'une variable X_j par un axe s:

$$QLT_s(X_j) = \frac{(OH_j^s)^2}{(OX_j)^2} = cor(X_j, v_s)^2$$

Si $QLT_s(X_j)$ est proche de 1, alors la variable X_j est bien représentée sur l'axe s.

Si $QLT_s(X_j)$ est proche de 0, alors la variable X_j est mal représentée sur l'axe s.

Interprétation

Dans la suite, on va définir quelques règles pour interpréter les résultats d'une ACP:

• Pour tout k et l tel que $k \neq l$, la projection du nuage N_I sur le plan principal engendré (u_k, u_l) est appelée **carte des individus**.

• Pour tout k et l tel que $k \neq l$, la projection du nuage N_v sur le plan principal engendré par (v_k, v_l) est appelée **carte des variables**.

• Un individu sera du côté des variables pour lesquelles il a de fortes valeurs, inversement il sera du côté opposé des variables pour lesquelles il a de faibles valeurs.

• Plus les valeurs d'un individu sont fortes pour une variable plus il sera éloigné de l'origine suivant l'axe factoriel décrivant le mieux cette variable.

• Deux individus à une même extrémité d'un axe (c-à-d éloignés de l'origine) sont proches (se ressemblent).

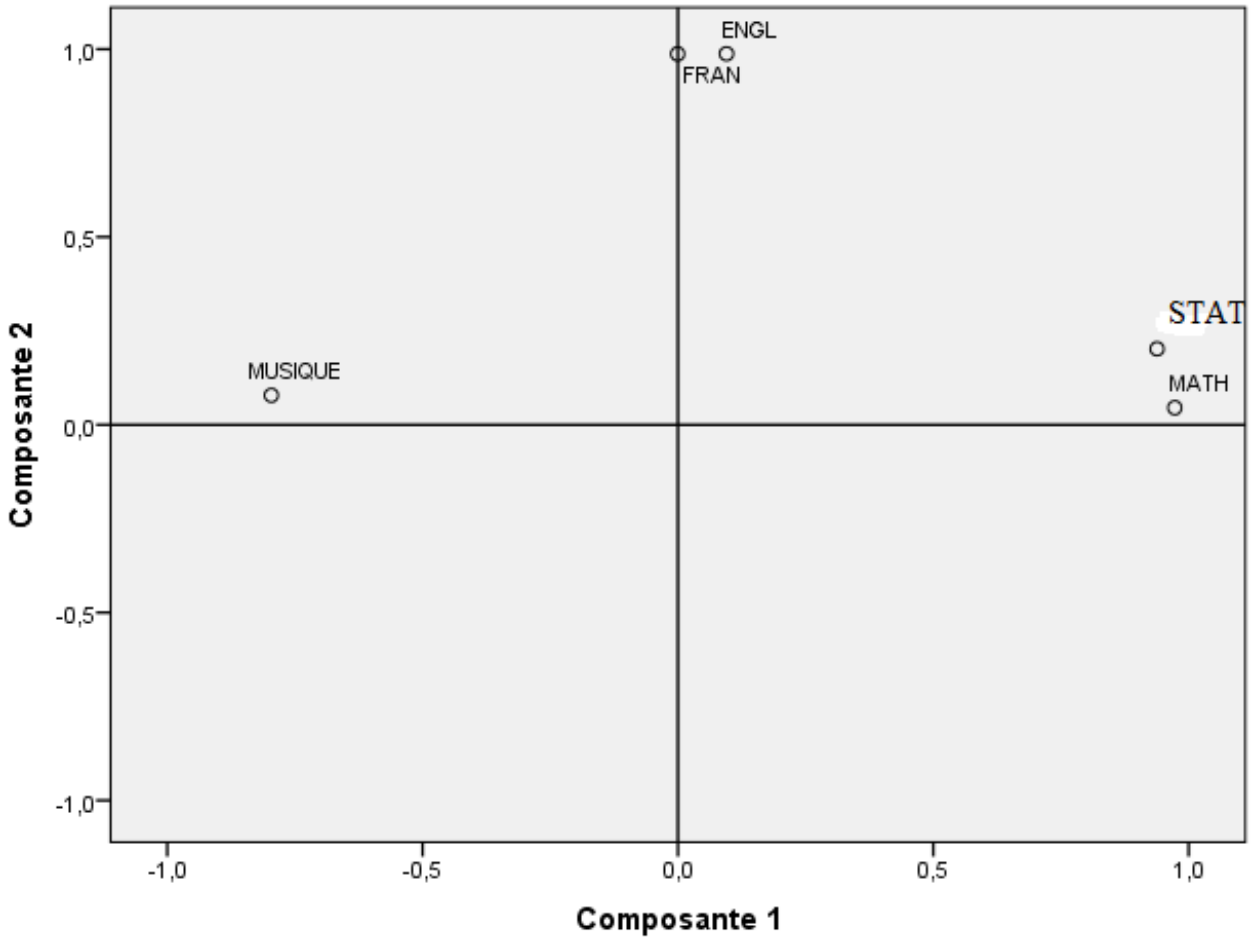
• Deux variables très corrélées positivement sont du même côté sur un axe.

Exemple

Etudiant	Math	Stat	Fran	Angl	Musique
I_1	6	6	5	5,5	8
I_2	4,5	5	7	7	14
I_3	6	7	11	9,5	11
I_4	14,5	14,5	15,5	15	5
I_5	14	14	12	12,5	6
I_6	16	15	5,5	5	7
I_7	5,5	7	14	11,5	10
I_8	13	12,5	8,5	9,5	12
I_9	6	8,5	13,5	13	18

Exemple

Carte des variables:



Exemple

Carte des individus: