Analyse en Composantes Principales

Introduction

- Les principales méthodes de l'analyse des données se séparent en deux groupes:
  - Les méthodes de classification,
  - Les méthodes factorielles.

Les méthodes de classification:

- Elles visent à réduire la taille de l'ensemble des individus en formant des groupes homogènes d'individus ou de variables.
- Ces groupes on les appelle aussi des classes, ou familles, ou segments, ou clusters.
- La classification est appelée aussi Segmentation ou Clustering.

Les méthodes factorielles:

- les méthodes factorielles consistent en la projection sur un espace de dimension inferieure pour obtenir une visualisation de l'ensemble des liaisons entre variables tout en minimisant la perte de l'information.

Les méthodes factorielles:

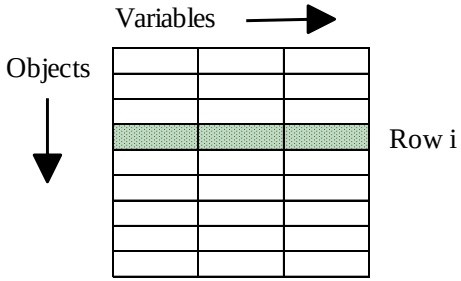
- Elles cherchent a réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques.
- **Si on travaille avec un tableau de variables numériques ou échelles, on utilisera l'analyse en composantes principales,**
- Si on travaille avec des variables qualitatives, on utilisera l'analyse des correspondances.

Les méthodes factorielles:

Les méthodes factorielles regroupent :

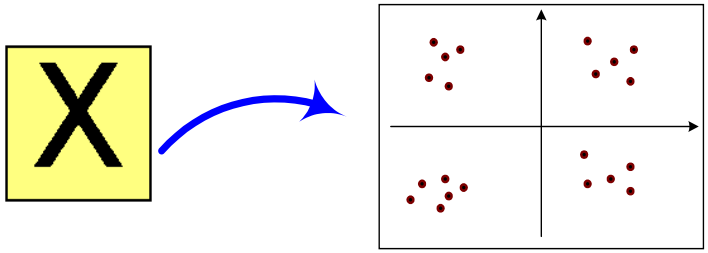
- **L'ACP : L'analyse en composantes principales**
- **L'AFC : L'analyse factorielle des correspondances**

Objectifs des méthodes multivariées



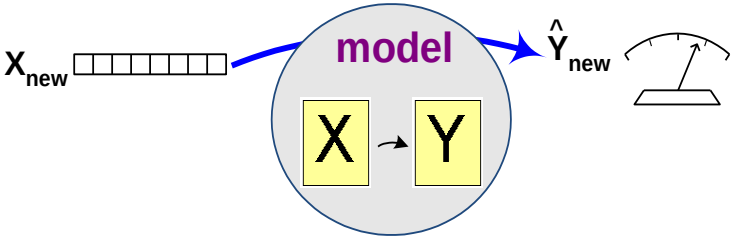
- Explorer et Décrire

  ACP



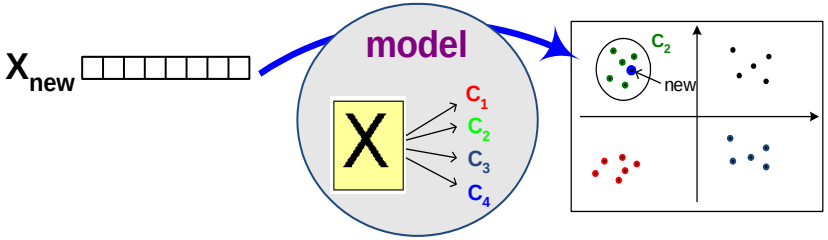
- Corréler et Prédire

  Régressions



- Caractériser et Classifier

  Classifications et discrimination



L’ACP

- L’ACP (Hotelling, 1933) a pour objectif de réduire le nombre de données, souvent très élevé, d’un tableau de données représenté, algébriquement, comme une matrice et, géométriquement comme un nuage de points.
- L’ACP se base sur l’étude des projections des points de ce nuage sur un axe (axe factoriel ou principal), un plan.
- On cherche donc, à réduire le nombre de descripteurs (variables) avec le minimum de perte d’information et préservant les relations existant déjà entre les différents descripteurs.

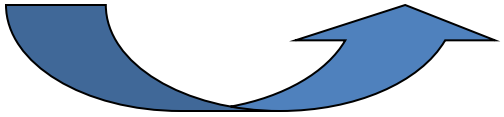
Déduction des coordonnées des individus (tableau 1) dans un espace de dimension deux (tableau 2)

Individus	Poids	Taille	Age	Note
1	45	150	13	14
2	50	160	13	15
3	50	165	13	16
4	60	175	15	9
5	60	170	14	10
6	60	170	14	7
7	70	160	14	8
8	65	160	13	13
9	60	155	15	17
10	65	170	14	11

Tableau1

Individus	**Axe 1?**	**Axe 2?**
1	-1,62	-0,20
2	-1,09	-0,52
3	-0,98	-0,72
4	1,27	0,09
5	0,67	-0,46
6	0,90	-0,90
7	0,81	0,35
8	-0,26	-0,16
9	-0,34	2,63
10	0,71	-0,10

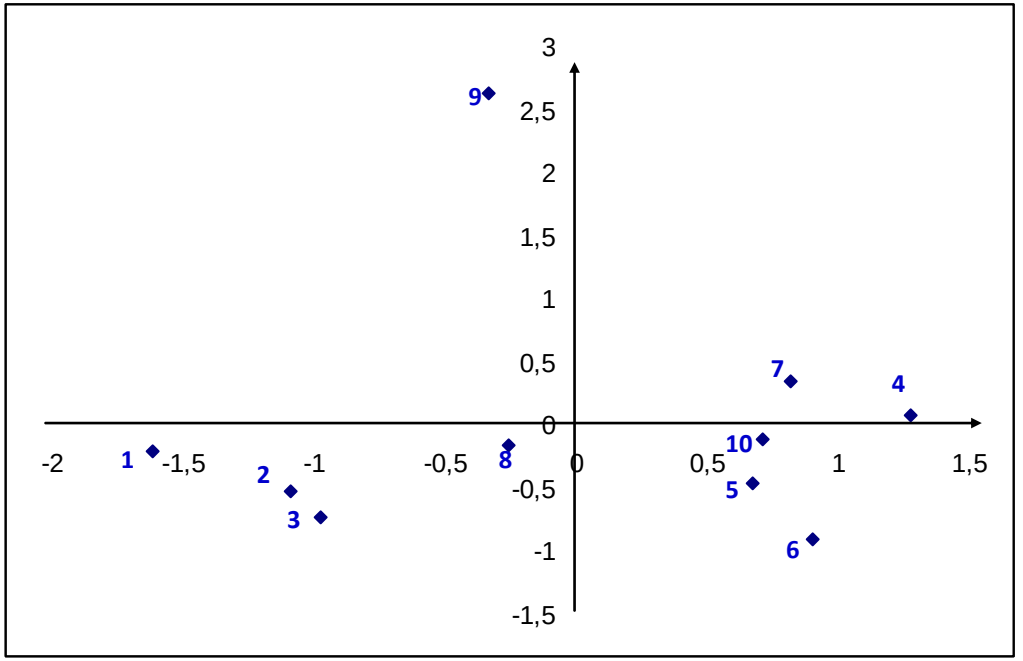
Tableau2



Quantité d'information?

Individus	Axe 1	Axe 2
1	-1,62	-0,20
2	-1,09	-0,52
3	-0,98	-0,72
4	1,27	0,09
5	0,67	-0,46
6	0,90	-0,90
7	0,81	0,35
8	-0,26	-0,16
9	-0,34	2,63
10	0,71	-0,10

Tableau2

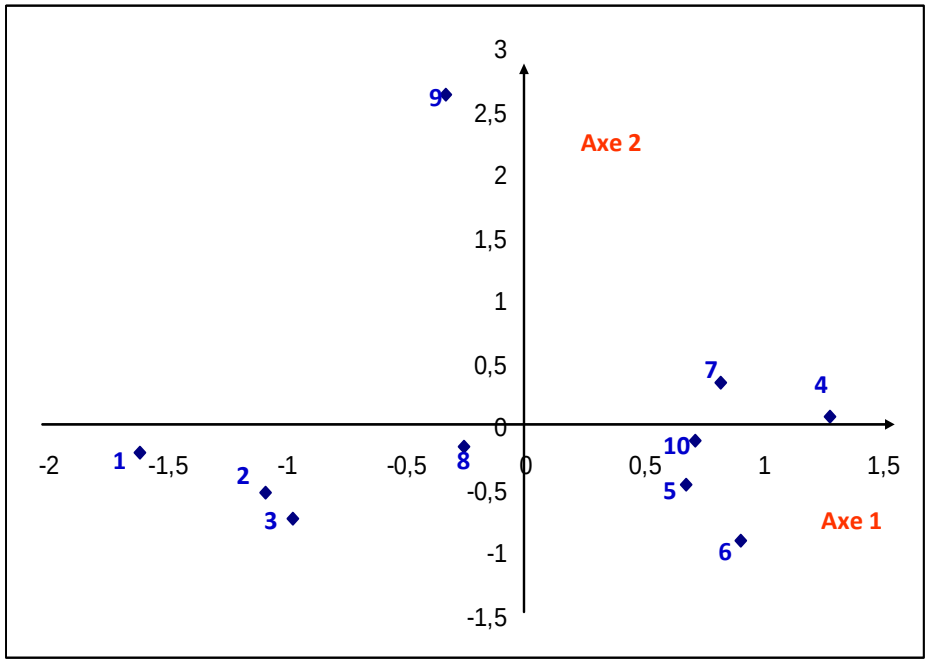


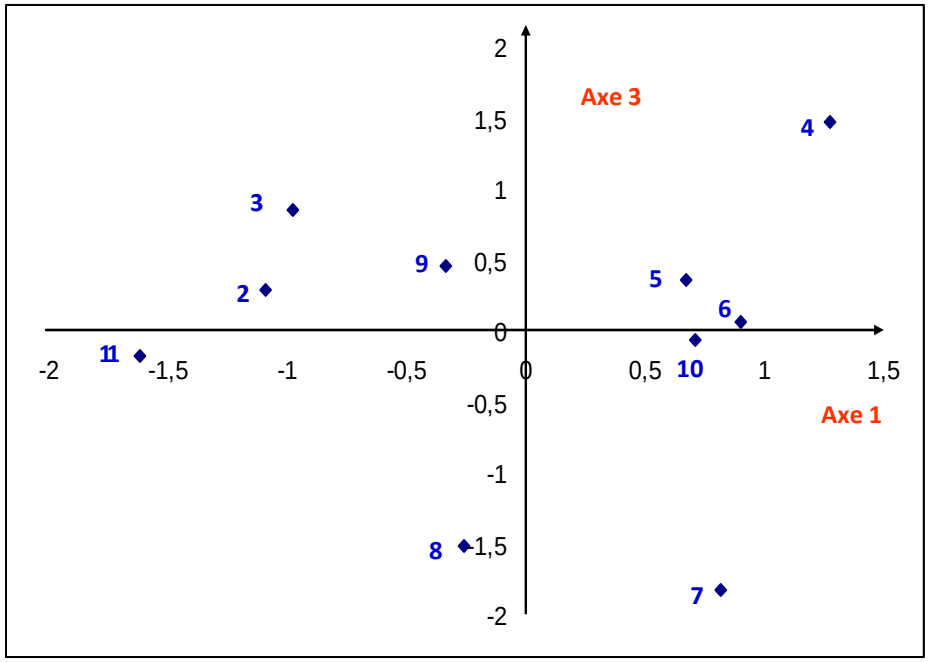
Graphe 1

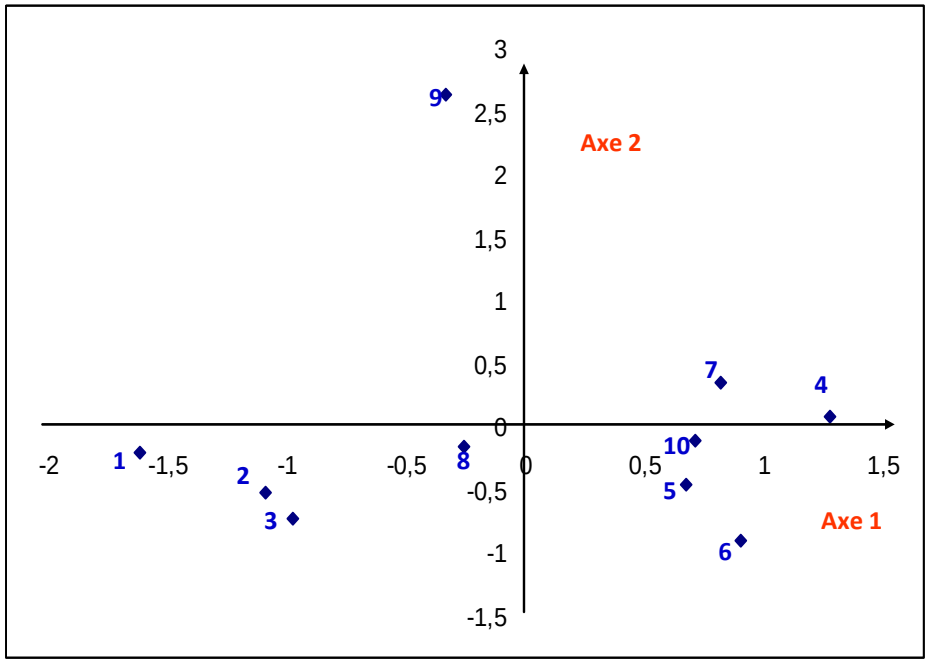
Quantité d'information restituée?

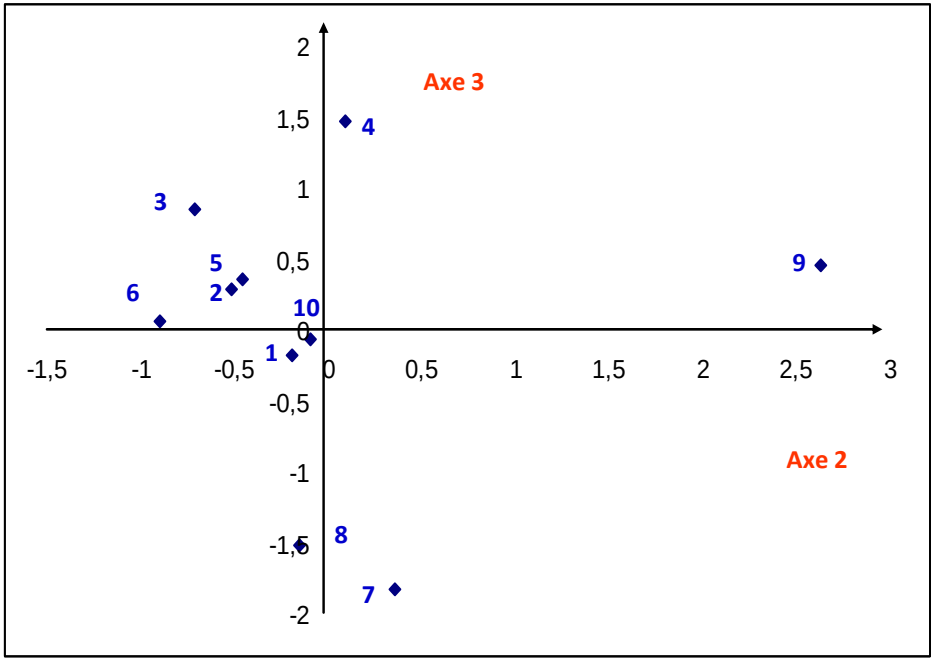
Individus	Axe 1	Axe 2	**Axe 3**
1	-1,62	-0,20	-0,17
2	-1,09	-0,52	0,30
3	-0,98	-0,72	0,86
4	1,27	0,09	1,48
5	0,67	-0,46	0,37
6	0,90	-0,90	0,07
7	0,81	0,35	-1,81
8	-0,26	-0,16	-1,51
9	-0,34	2,63	0,46
10	0,71	-0,10	-0,06

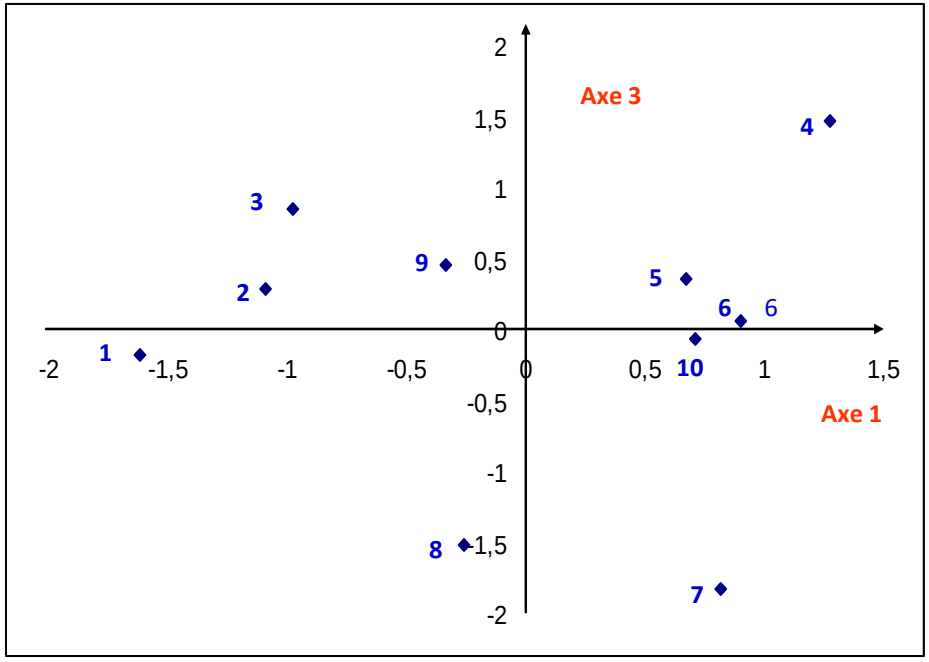
Peut-on améliorer l'image?

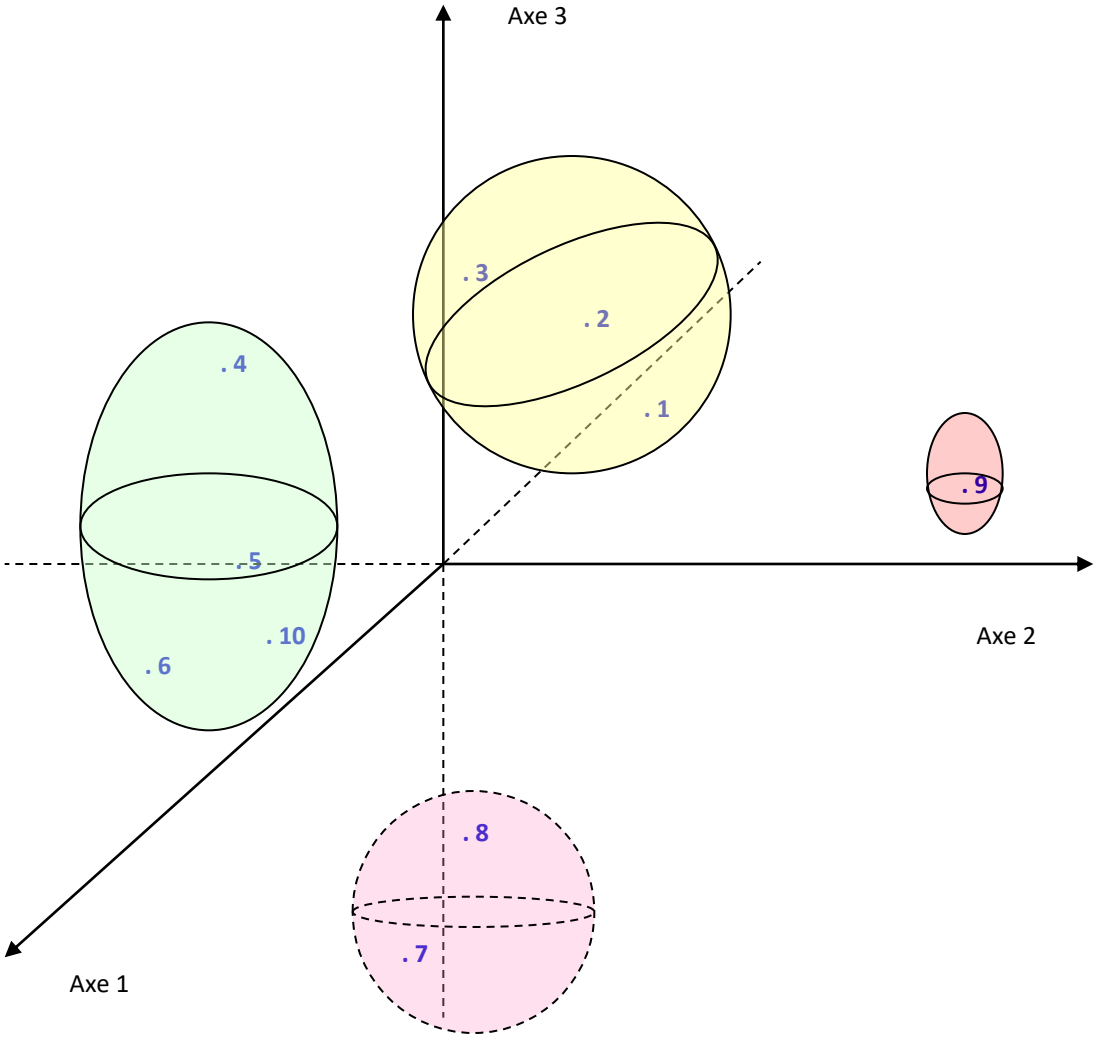





Axe 2
Axe 1
9
7
4
1
2
8
10
5
3
6


Axe 3
Axe 2
4
3
5
9
6
2
10
1
8
7


Axe 3
Axe 1
4
3
9
5
2
6
6
1
10
8
7


Axe 3
. 3
. 2
. 1
. 4
. 5
. 10
. 6
. 9
Axe 2
. 8
. 7
Axe 1

L'ACP : Algébriquement

- Il s'agit de chercher les valeurs propres maximales de la matrice des données et par conséquent leurs vecteurs propres associés qui représenteront leurs sous-espaces vectoriels (axes factoriels ou principaux).

Procédure de l'ACP:

- On cherche X' la transposée de la matrice X.
- On détermine les valeurs propres de la matrice symétrique X'X.
- Soient $\lambda_1, \lambda_2, \ldots, \lambda_p$ ses valeurs propres.
- On les classe $\lambda_1>\lambda_2>\lambda_3>\lambda_4>\ldots$.
- Alors $X'X=A\Lambda A^{-1}$ où

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_p \end{pmatrix}$$

Procédure de l'ACP:

- D'après les propriétés de la trace des matrices; on a: $tr(X'X) = tr(A\Lambda A^{-1}) = tr(AA^{-1}\Lambda) = tr(\Lambda)$
- Soit $tr(X'X) = tr(\Lambda)$
- En raison des valeurs numériques décroissantes de λ_1, λ_2, …, la somme des premières valeurs propres représente, souvent, une proportion importante de la trace de X'X.

Procédure de l'ACP:

- Ainsi, dans la pratique on peut se limiter à trouver les premières valeurs propres λ_1, λ_2, …, λ_q avec q assez inferieur à p.
- L'information perdue est alors relativement faible.
- Généralement, avec q=3, le taux de restitution est acceptable.

Procédure de l'ACP:

- Les valeurs propres trouvées, donnent les espaces propres associés à leurs vecteurs propres qui seront des droites vectorielles (on les appelles des axes factoriels ou des facteurs).
- L'ACP nous permet de traiter un très grand nombre de données (matrice) pour identifier un nombre relativement restreint de données (axes factoriels)

L'ACP géométriquement:

- **Géométriquement, on représente le tableau** comme un nuage de points.

- Lors de la projection, le nuage peut être déformé est donc serait différent du réel, alors les méthodes d'ajustement consistent en minimiser cette déformation et ce en maximisant les distances projetées.

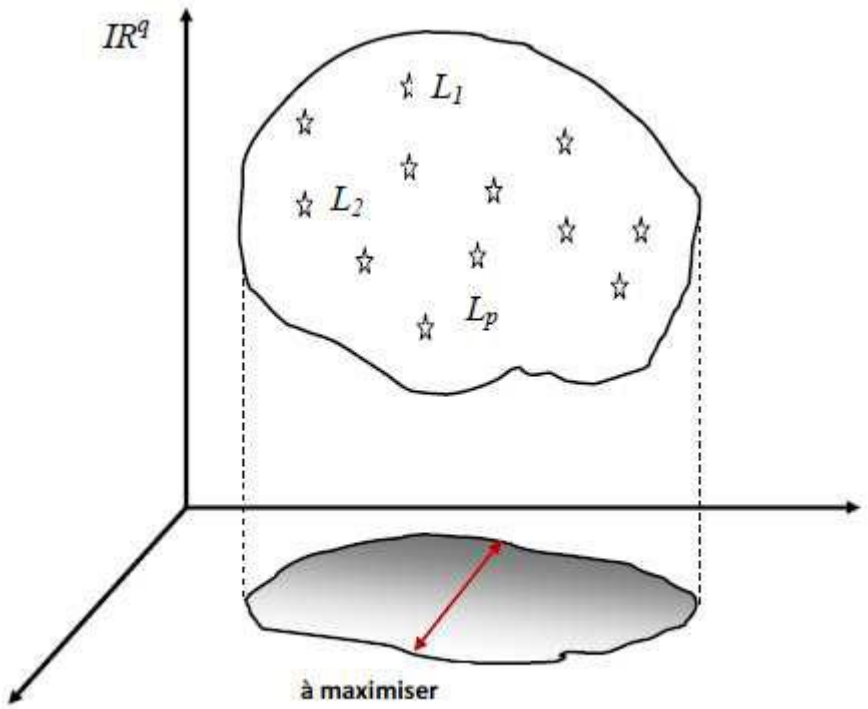


Schéma de travail

1. Du tableau de base

	X_1	$\cdots$	X_j	$\cdots$	X_p	M_i
1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1p}	M_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ip}	M_i
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
n	x_{n1}	$\cdots$	x_{nj}	$\cdots$	x_{np}	M_n
N_j	N_1	$\cdots$	N_j	$\cdots$	N_p	

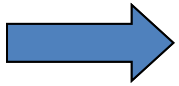
on déduit l'un des deux nuages possibles, individus ou variables.

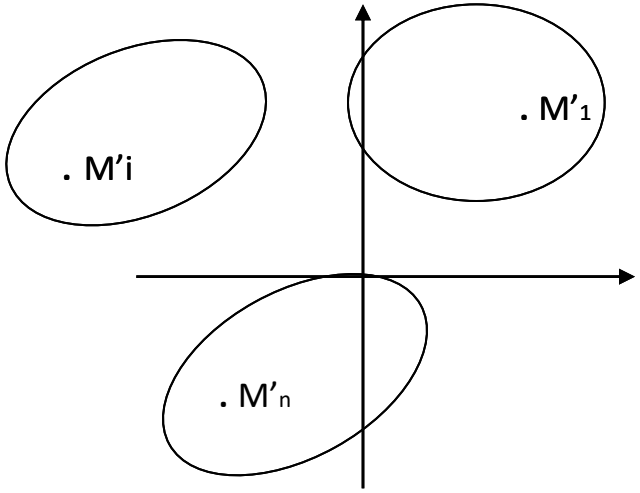
$$\{M_i, m_i \text{ où } i \text{ varie de } 1 \text{ à } n\} \qquad \{N_j, f_j \text{ où } j \text{ varie de } 1 \text{ à } p\}$$

2. On détermine ensuite **l'inertie I**, c'est-à-dire la dispersion du nuage par rapport à son centre de gravité.

3. Déterminer un espace de dimension faible dans lequel le nuage choisi sera projeté orthogonalement.

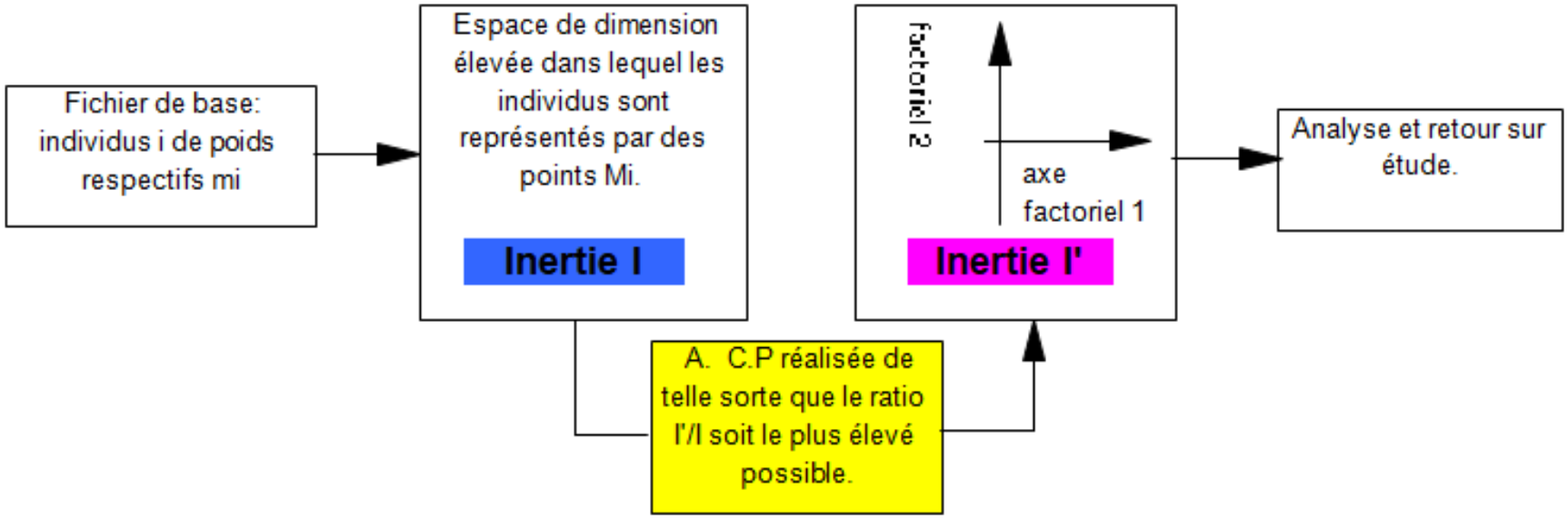
Individus	Axe 1	Axe 2
M'_1		
…		
M'_i		
…		
M'_n		





Ceci constitue un nouveau nuage de points $\left\{M'_i, m_i \text{ où } i \text{ varie de } 1 \text{ à } n\right\}$

pour lequel on détermine **l'inertie I'.** On compare I' avec I. Si le ratio est bon, on peut conserver l'image.


Fichier de base: individus i de poids respectifs mi
Espace de dimension élevée dans lequel les individus sont représentés par des points Mi.
Inertie I
factoriel 2
axe factoriel 1
Inertie I'
Analyse et retour sur étude.
A. C.P réalisée de telle sorte que le ratio I'/I soit le plus élevé possible.

L'inertie

La forme mathématique de l'inertie est la suivante:

$$I = \sum_{i=1}^{i=n} m_i \left\| GM_i \right\|^2$$

Lorsque les variables sont centrées, c'est-à-dire lorsqu'à chaque valeur on a enlevé la valeur moyenne, l'inertie est égale à la somme des variances des variables que l'on soumet à l'analyse. A ce titre **l'inertie est une généralisation de la notion de variance**.

$$I = \sum_{j=1}^{j=p} V\left(X_j\right)$$

Lorsque les points représentant les individus sont proches du centre de gravité, l'inertie est faible.

Matrice d'inertie

La réalisation d'une ACP est construite sur les qualités d'une matrice qui porte le nom de matrice d'inertie. Celle-ci est définie de la manière suivante:

$$M = \sum_{i=1}^{i=n} m_i GM_i \cdot GM_i^{'}$$

Chaque produit $GM_i \cdot GM_i^{'}$ s'exprime par la relation:

$$GM_i \cdot GM_i^{'} = \begin{pmatrix} x_{i1}^2 & x_{i1}x_{i2} & \cdots & x_{i1}x_{ip} \\ x_{i2}x_{i1} & x_{i2}^2 & \cdots & x_{i2}x_{ip} \\ \vdots & \vdots & & \vdots \\ x_{ip}x_{i1} & \cdots & \cdots & x_{ip}^2 \end{pmatrix}$$

et la matrice d'inertie par la relation:

$$\sum_{i=1}^{i=n} m_i GM_i \cdot GM_i^{'} =$$

$$\sum_{i=1}^{i=n} m_i \begin{pmatrix} x_{i1}^2 & x_{i1}x_{i2} & \cdots & x_{i1}x_{ip} \\ x_{i2}x_{i1} & x_{i2}^2 & \cdots & x_{i2}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{ip}x_{i1} & \cdots & \cdots & x_{ip}^2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{i=n} m_i x_{i1}^2 & \sum_{i=1}^{i=n} m_i x_{i1}x_{i2} & \cdots & \sum_{i=1}^{i=n} m_i x_{i1}x_{ip} \\ \sum_{i=1}^{i=n} m_i x_{i2}x_{i1} & \sum_{i=1}^{i=n} m_i x_{i2}^2 & \cdots & \sum_{i=1}^{i=n} m_i x_{i2}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{i=n} m_i x_{ip}x_{i1} & \cdots & \cdots & \sum_{i=1}^{i=n} m_i x_{ip}^2 \end{pmatrix} =$$

$$\begin{pmatrix} V(X_1) & \sum_{i=1}^{i=n} m_i x_{i1}x_{i2} & \cdots & \sum_{i=1}^{i=n} m_i x_{i1}x_{ip} \\ \sum_{i=1}^{i=n} m_i x_{i2}x_{i1} & V(X_2) & \cdots & \sum_{i=1}^{i=n} m_i x_{i2}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{i=n} m_i x_{ip}x_{i1} & \cdots & \cdots & V(X_p) \end{pmatrix}$$

1. Nous constatons que la **trace** de cette matrice, c'est-à-dire la somme de ses éléments diagonaux est égale à **l'inertie de système**.

Ainsi, avons-nous la possibilité de caractériser la dispersion du nuage par les valeurs propres d'une matrice. En effet la trace est un invariant égal à la somme des valeurs propres.

$$Tr(M) = \sum_{j=1}^{j=p} \sum_{i=1}^{i=n} m_i x_{ij}^2 = \sum_{j=1}^{p} V(X_j) = I = \sum_{j=1}^{p} \lambda_j$$

Parce que l'inertie est identifiée aux valeurs propres d'une matrice, il est normal de sélectionner les plus importantes pour conserver au mieux l'information. Rangeons celles-ci par ordre décroissant et sélectionnons les plus fortes.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

Le taux de restitution de l'information dans un plan est donné par:

$$\tau = \frac{\lambda_1 + \lambda_2}{\sum \lambda_j} \cdot 100$$

2. Le plan de projection est, donc engendré par deux vecteurs propres associés aux deux plus grandes valeurs propres. Soit à résoudre les équations:

$$\begin{cases} u_j \neq 0 \\ Mu_j = \lambda_j u_j \end{cases} \qquad \text{j variant de 1 à 2}$$

On se posera la question du type de repère:

- orthonormé;
- orthogonal;
- normé;
- quelconque

3. Les diverses projections

Lorsque le plan est défini, il reste à donner les diverses coordonnées. Pour cela, on utilise les relations:

- abscisse

$$\alpha_i = GM_i^{'} \cdot u_1$$

- ordonnée

$$\beta_i = GM_i^{'} \cdot u_2$$

Analyser un tableau de données : Variables quantitatives

Modele	CYL	PUISS	LONG	LARG	POIDS	V.MAX
Alfasud TI	1350	79	393	161	870	165
Audi 100	1588	85	468	177	1110	160
Simca 1300	1294	68	424	168	1050	152
Citroen GS Club	1222	59	412	161	930	151
Fiat 132	1585	98	439	164	1105	165
Lancia Beta	1297	82	429	169	1080	160
Peugeot 504	1796	79	449	169	1160	154
Renault 16 TL	1565	55	424	163	1010	140
Renault 30	2664	128	452	173	1320	180
Toyota Corolla	1166	55	399	157	815	140
Alfetta 1.66	1570	109	428	162	1060	175
Princess 1800	1798	82	445	172	1160	158
Datsun 200L	1998	115	469	169	1370	160
Taunus 2000	1993	98	438	170	1080	167
Rancho	1442	80	431	166	1129	144
Mazda 9295	1769	83	440	165	1095	165
Opel Rekord	1979	100	459	173	1120	173
Lada 1300	1294	68	404	161	955	140

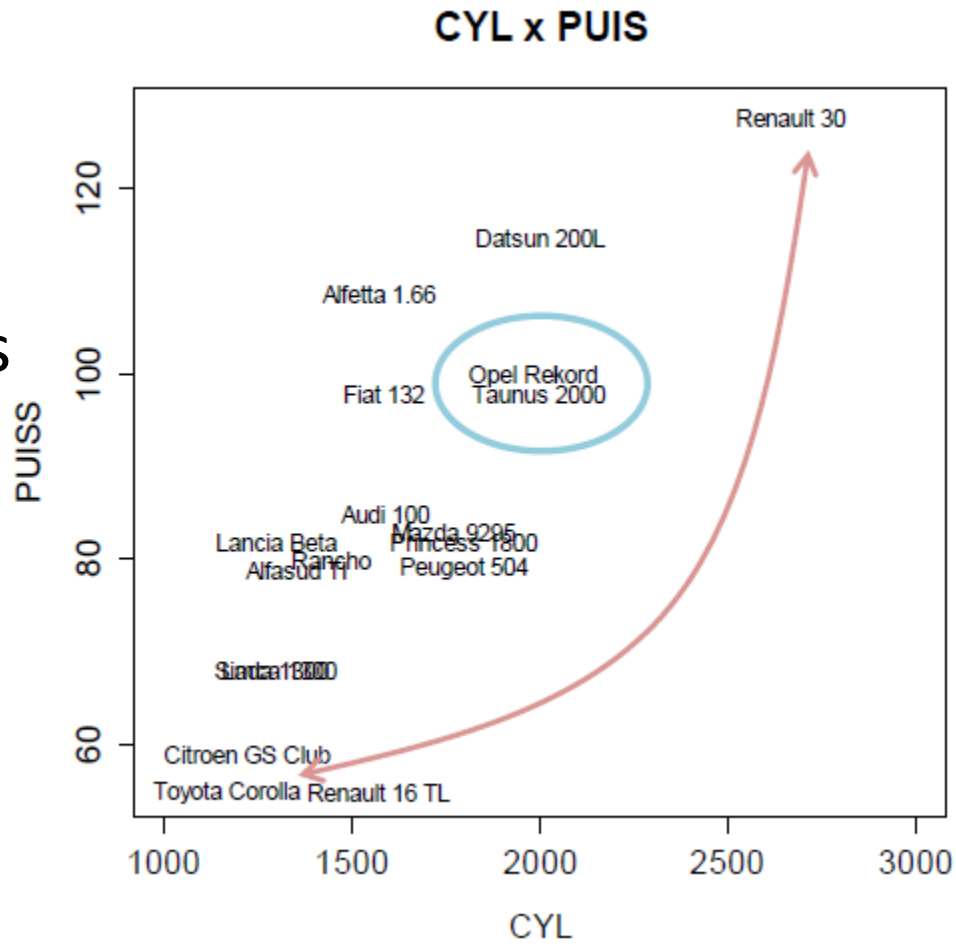
x_{ij}

Etude des individus

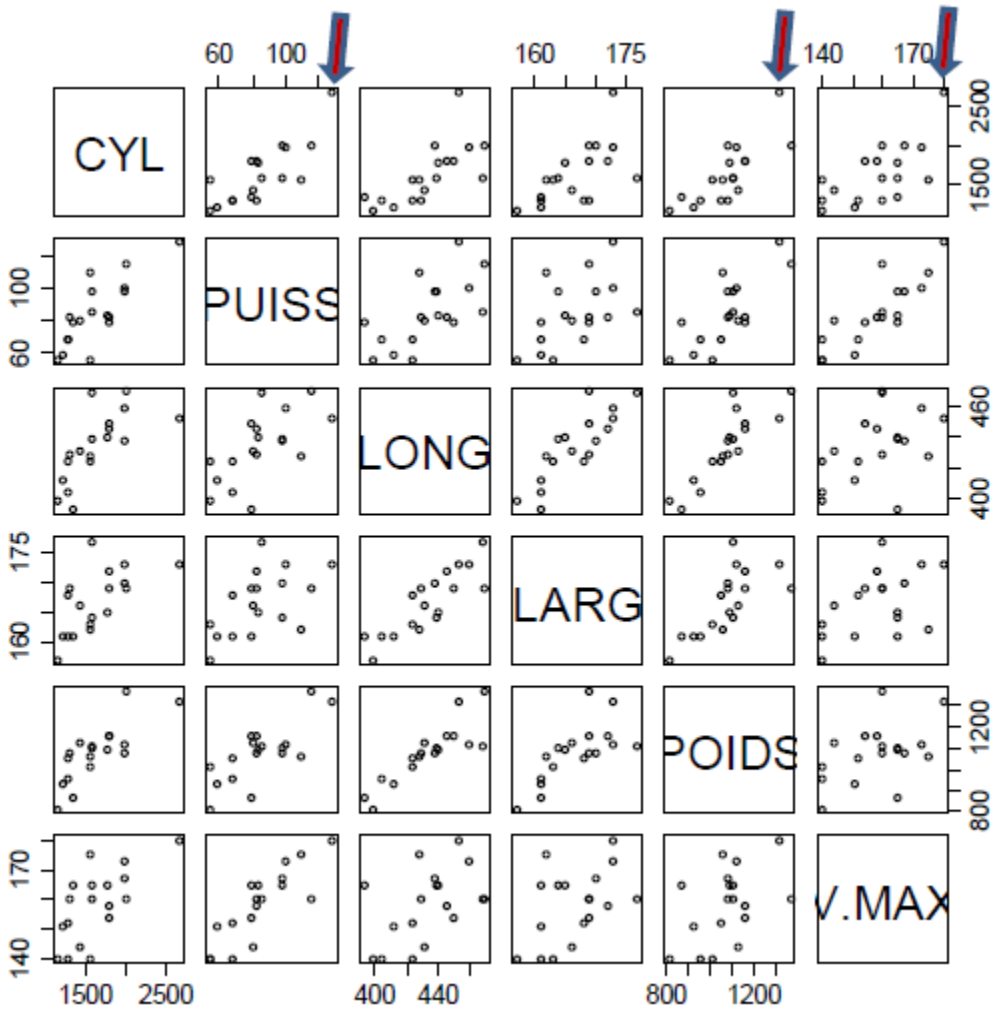
Etude des individus (2 variables)

Observations :

- Les variables CYL et PUISS sont liées.
- «Opel Reckord» et «Taunus 2000 (Ford)» ont le même profil (caractéristiques)
- «Renault 30» et «Toyota Corolla» ont des profils opposés…



Etude des individus (p >2)



Impossible de créer un nuage à «p» dimensions.

On pourrait croiser les variables 2 à 2, mais :

- Très difficile de surveiller plusieurs cadrans en même temps.
- Etiqueter les points rendrait le tout illisible.

Etude des individus : Notion d'inertie

- Principe : Construire un système de représentation de dimension réduite (q<< p) qui préserve les **distances entre les individus.**

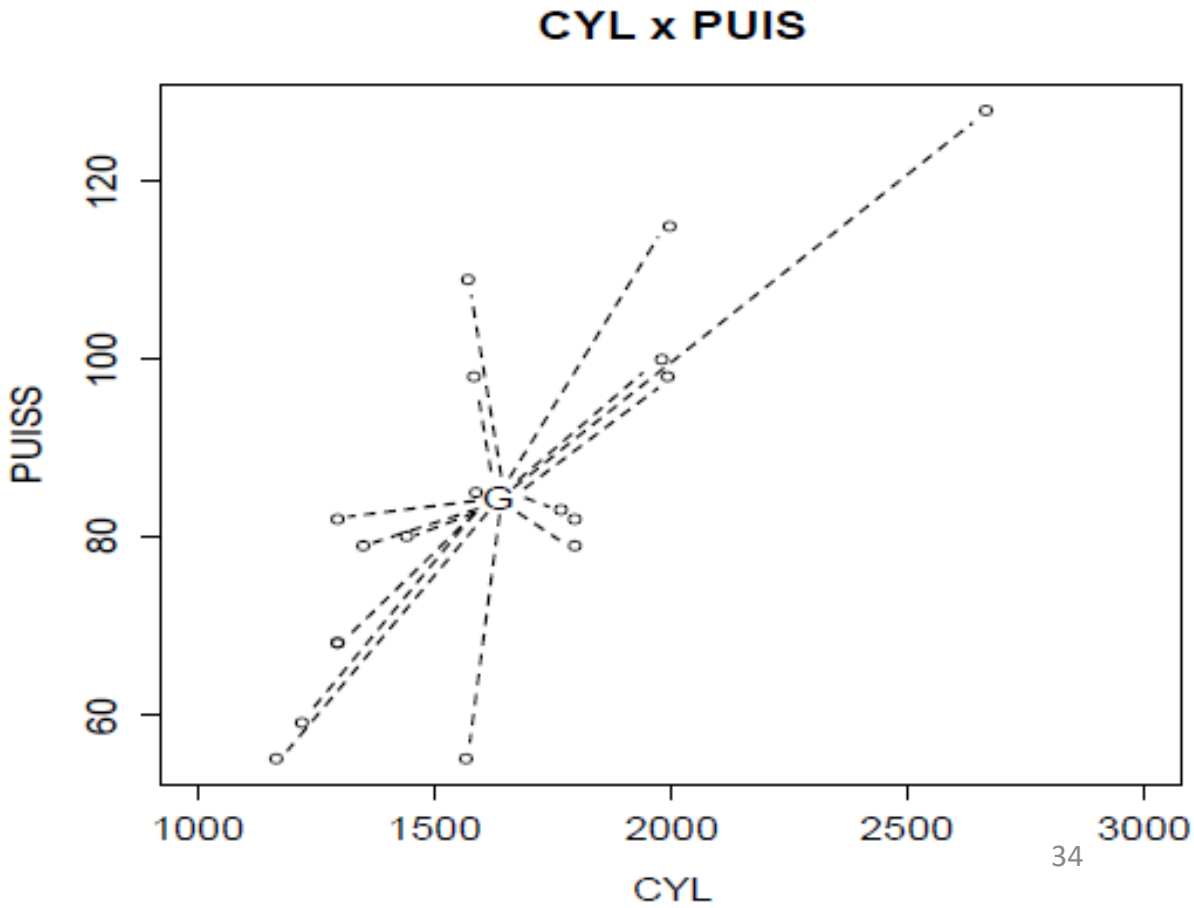
- Distance euclidienne entre 2 individus (i, i')

$$d^2(i,i') = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2$$

Etude des individus : Notion d'inertie

- L'inertie indique la dispersion autour du barycentre, c'est une variance calculée sur p dimensions.

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



Etude des individus : Régression orthogonale

Habituellement on (a) centre et (b) réduit les variables. On parle d'ACP normée.

a) Pour que G soit situé à l'origine [obligatoire]
b) Pour rendre comparables les variables exprimées sur des échelles (unités) différentes [non obligatoire]

Etude des individus : Régression orthogonale

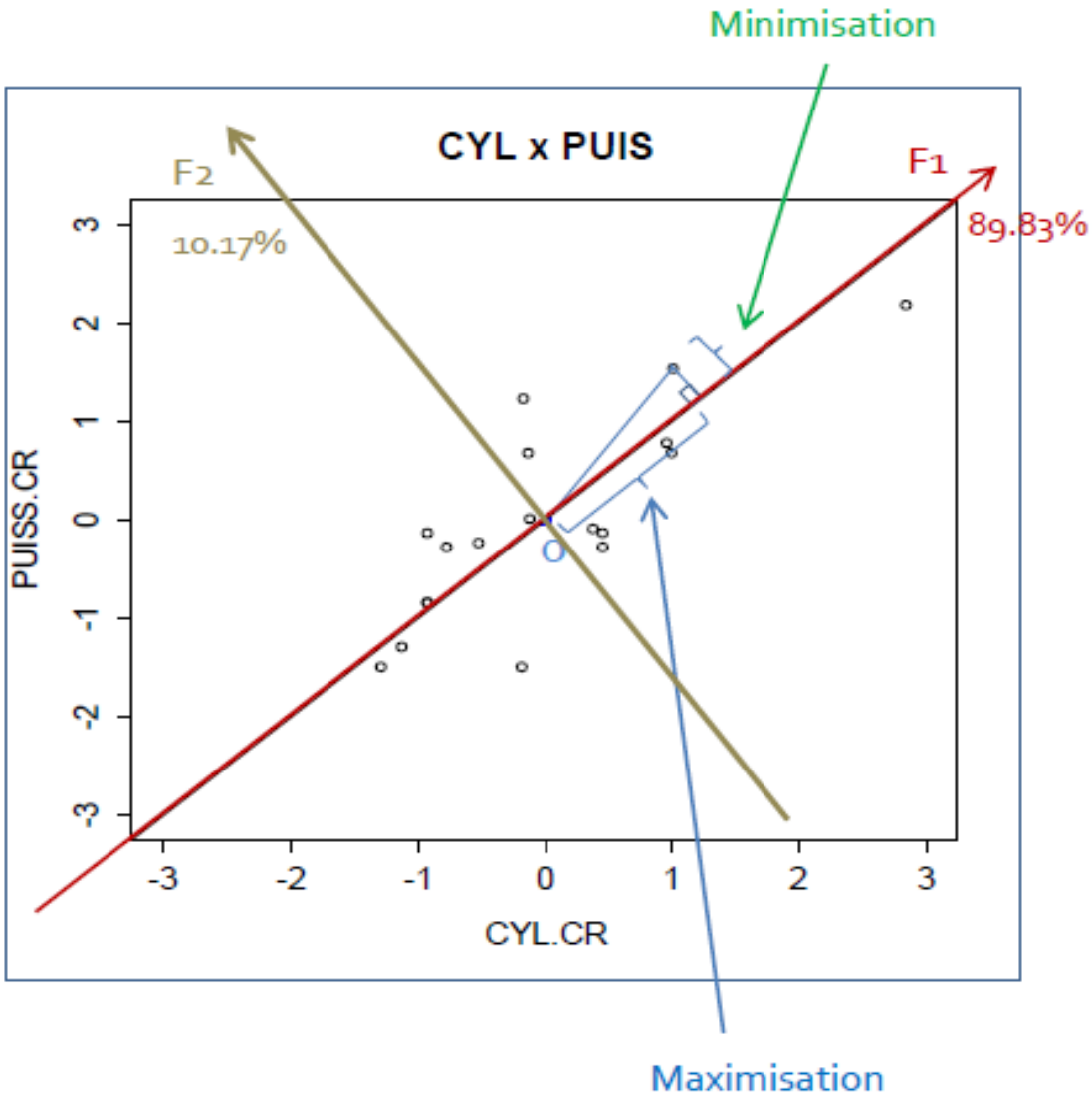
- Trouver la première composante F_1 qui maximise l'écartement global des points par rapport à l'origine :

$$\lambda_1 = \frac{1}{n}\sum_{i=1}^{n} F_{i1}^2 = 1.796628 \qquad \frac{\lambda_1}{I_p} = 89.83\%$$

- Trouver la 2nde composante F_2 qui traite l'inertie non-expliquée (résiduelle) par F_1 (par conséquent, F_2 est non corrélée avec F_1)

$$\lambda_2 = \frac{1}{n}\sum_{i=1}^{n} F_{i2}^2 = 0.203372 \qquad \left(\frac{\lambda_2}{I_p} = 10.17\%\right)$$

Etude des individus : Régression orthogonale



Etude des variables

Etude des variables : Matrice des corrélations

- Matrice des corrélations **R sur les données du tableau**

CORR	CYL	PUISS	LONG	LARG	POIDS	V.MAX
CYL	1	0.797	0.701	0.630	0.789	0.665
PUISS		1	0.641	0.521	0.765	0.844
LONG			1	0.849	0.868	0.476
LARG				1	0.717	0.473
POIDS					1	0.478
V.MAX						1

$$r_{jm} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{im} - \bar{x}_m\right)}{s_j \times s_m}$$

Etude des variables : Construction des composantes

- Construire la première composante F_1 qui permet de maximiser le carré de sa corrélation avec les variables

$$\lambda_1 = \sum_{j=1}^{p} r_j^2(F_1)$$

- Habituellement, Inertie totale = Somme des variances des variables
- Lorsque les données sont réduites (ACP normée), Inertie totale = Trace(R) = p

$$I_p = p \qquad \sum_{k=1}^{p} \lambda_k = p$$

Etude des variables : Exemple

- On a observé p (4) variables sur n (6) individus. Dans la pratique cela représente un tableau à np (24) entrées.

descripteur / Sujet	D_1	D_2	D_3	D_4
S_1	-11	-60	110	40
S_2	-12	-62	93	25
S_3	-15	-80	113	39
S_4	-14	-75	94	25
S_5	-14,5	-82	100	30
S_6	-13	-72	102	32

Etude des variable : Matrice de corrélation

La matrice de corrélation **R** montre que la variable 1 est fortement corrélée avec la variable 2 ; il en est de même pour les variables 3 et 4.

1	0,970	-0,064	0,094
--	1	-0,102	0,037
--	--	1	0,986
--	--	--	1

Etude des variables : réduction de dimension

- Les variables de départ sont remplacées par « des vecteurs propres » de la matrice des variances (covariance) ou de la matrice **R**, appelés **Composantes principales**.

- **Y-a-t-il un critère d'arrêt ?** généralement on s'arrête quand au moins 75% de la variance est expliquée par la variance cumulée par les CP.

Etude des variables : vecteur propre

- λ est une **valeur propre** de la matrice A si et seulement si $A\mathbf{v} = \lambda\mathbf{v}$
- Le vecteur **v** dans la relation ci-dessus est appelé **vecteur associé à** λ
- Les valeurs propres s'obtiennent en résolvant le système d'équations $\det(A- \lambda I) = 0$.
- Le nombre de valeurs propres, $\lambda_1 > \ldots > \lambda_p$, est égal au nombre de lignes = nombre de colonnes de la matrice A
- **Important : La somme** des valeurs propres de A est égale à la **variance** contenue dans l'ensemble des données.

Etude des variables : composantes principales

- D'un point de vue pratique les composantes principales s'écrivent

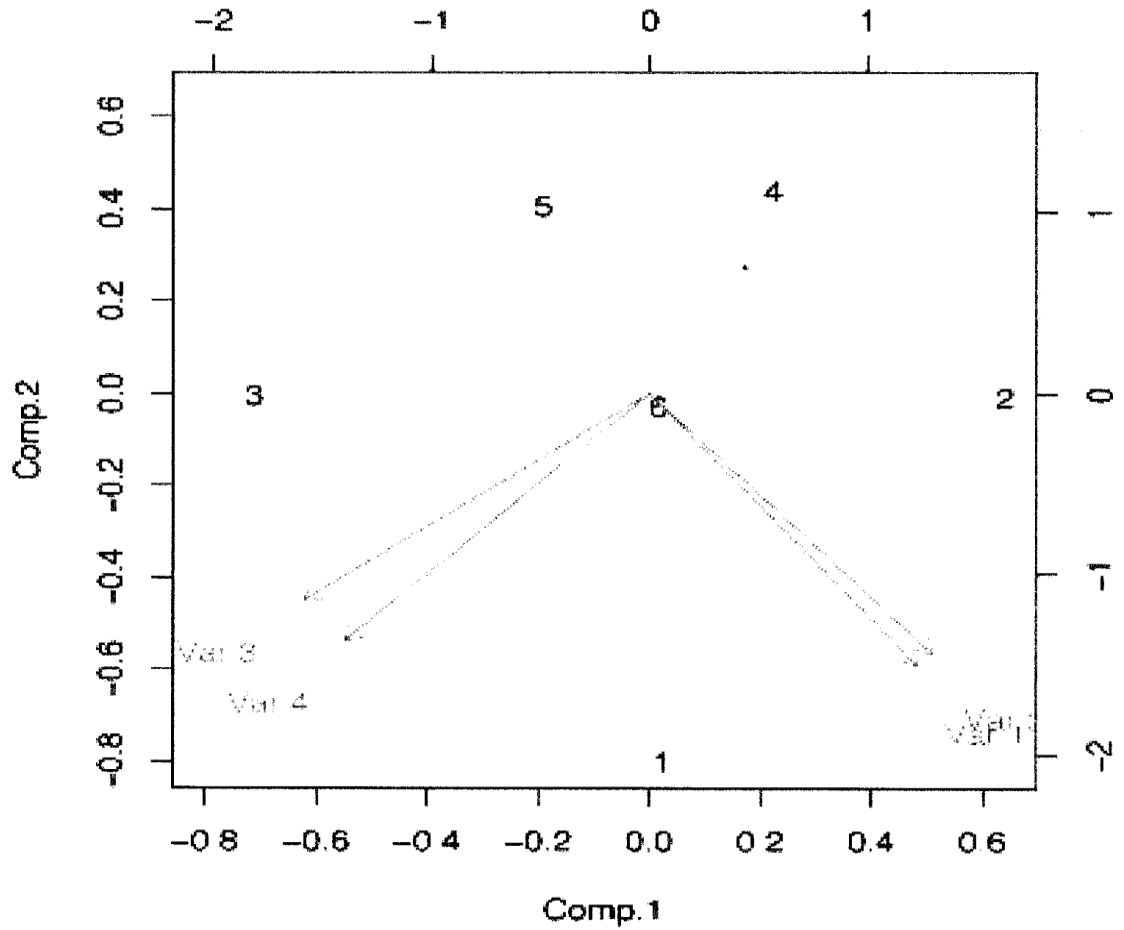
$$F_j = u_{j1}X_1 + \ldots + u_{jp}X_p$$

c'est-à-dire que F_j est une combinaison linéaire des variables initiales $X_1, \ldots, X_p$.

En plus de cet aspect calculatoire on doit pouvoir faire des affirmations sur la qualité de la réduction et la qualité de la représentation graphique.

Etude des variables : Représentation graphique

- Représentation des variables et individus dans le plan CP1, CP2 comme illustré ci-dessous



Etude des variables : Interprétation

- Chaque valeur propre représente la variance prise en compte par la composante principale correspondante.
- Pour l'exemple on obtient :

	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 premières composantes rendent compte de 0,5003+0,4917 = 0,9920 = 99,2 % de la variance totale.
- Ce qui veut dire que les 4 descripteurs peuvent être remplacés par les 2 premières composantes tout en préservant la quasi-totalité de l'information (réduction).

Résultats des calculs

- **Scores des individus** : il s'agit des valeurs prises par les composantes principales sur les individus.
- Ici

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

Résultats des calculs

- **Saturations des variables** : il s'agit des coefficients de corrélation entre les variables et les composantes principales.

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

- La première composante est surtout corrélée avec les deux derniers descripteurs

Résultats des calculs

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

- CTR(sujet 1, CP1)= $\frac{0,0771^2}{0,0771^2+\ldots+0,0556^2}=0,64$

- Qualité de la représentation :

pour sujet 1 et CP2

$$\text{QLT} = \frac{2,7515^2}{0,0771^2+\ldots+0,0166^2}=0,998$$

Résultats des calculs

- **Qualité de la représentation d'une variable** à la formation d'une CP : contribution de la première variable à la formation de la première composante principale

$$\textbf{CTR} = \frac{0{,}6288^2}{0{,}6288^2 + 0{,}6651^2 + \ldots + 0{,}7129^2} = 0{,}1976$$

Exemple 1

On interroge des chefs de service sur les qualités:

- X1: technicité;
- X2: polyvalence;
- X3: créativité

que possèdent ou non leurs collaborateurs. Les réponses sont données sur une échelle de valeurs comprises entre 0 et 4. Les résultats sont présentés dans le tableau suivant:

Individus	Technicité	Polyvalence	Créativité
1	3	4	4
2	1	0	0
3	2	0	0
4	3	2	4
5	2	0	4
6	1	2	0
7	2	2	0
8	1	2	4
9	2	4	4
10	1	0	4
11	2	2	4
12	2	4	0
13	3	4	0
14	3	2	0

Réaliser une A.C.P d'ordre 2 du nuage des individus.

Points représentant les individus	Vecteurs	Technicité	Polyvalence	Créativité
M1	GM1	1	2	2
M2	GM2	-1	-2	-2
M3	GM3	0	-2	-2
M4	GM4	1	0	2
M5	GM5	0	-2	2
M6	GM6	-1	0	-2
M7	GM7	0	0	-2
M8	GM8	-1	0	2
M9	GM9	0	2	2
M10	GM10	-1	-2	2
M11	GM11	0	0	2
M12	GM12	0	2	-2
M13	GM13	1	2	-2
M14	GM14	1	0	-2
G		0	0	0

Si à chaque individu on accorde le même poids égal à 1/14, l'inertie est égale à:

$$I = \sum_{i=1}^{i=14} m_i GM_i^{'} GM_i = \frac{1}{14} \sum_{i=1}^{i=14} GM_i^{'} GM_i = V(X_1) + V(X_2) + V(X_3) = \frac{96}{14}$$

Par exemple dans ce calcul,

- la variance de la première variable est égale à 8 / 14, alors que
- la variable numéro deux a une variance égale à 32 / 14 et que
- la variable numéro trois à une variance égale à 4.

C'est ce nombre qu'il faut essayer de restituer.

• Le meilleur plan de projection, est dirigé par 2 vecteurs appelés vecteurs propres de la matrice d'inertie associés aux deux plus grandes valeurs propres de la même matrice.

• Il faut noter que dans l'opération, on a réussi à identifier la quantité d'information aux valeurs propres d'une matrice.

•Lorsque les valeurs sont centrées, la matrice d'inertie est la matrice des variances/covariances;

•Lorsque les valeurs sont centrées et réduites, la matrice est la matrice des corrélations.

•La somme des éléments diagonaux de cette matrice est égale à l'inertie.

Dans notre exemple, cette matrice est égale à:

$$\mu = \frac{1}{14}\begin{pmatrix} 8 & 8 & 0 \\ 8 & 32 & 0 \\ 0 & 0 & 56 \end{pmatrix}$$

Les valeurs propres s'obtiennent par différentes méthodes. Ici, elles sont égales à:

$$\lambda_1 = 4 \quad \lambda_2 = 2{,}46 \quad \lambda_3 = 0{,}4$$

On note que

$$I = V(X_1) + V(X_2) + V(X_3) = \lambda_1 + \lambda_2 + \lambda_3 = 6{,}86$$

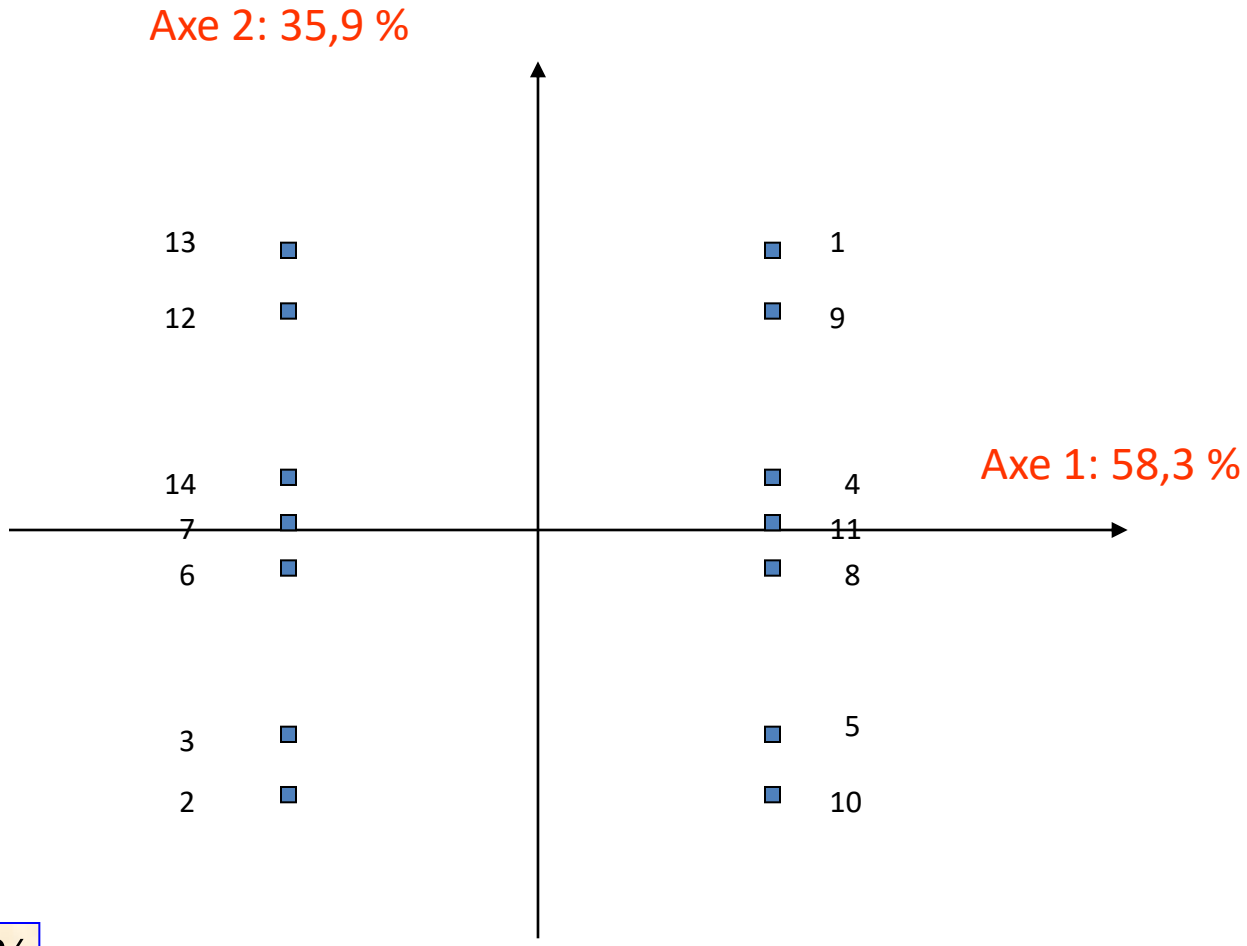
Le taux de restitution de l'information est égal à:

$$\tau = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} \cdot 100 = 94{,}17\,\%$$

Cela signifie que l'image de dimension deux que l'on va voir représente bien le nuage de points.

Si le taux de restitution est insuffisant, on ajoute un axe ou on réduit le nombre de variables que l'on soumet à l 'analyse. Nous comprenons mieux la nécessité d'étudier les diverses corrélations entre les variables.

Axe 1	Axe 2
2	2,11
-2	-2,11
-2	-1,92
2	0,19
2	-1,92
-2	-0,19
-2	0
2	-0,19
2	1,92
2	-2,11
2	0
-2	1,92
-2	2,11
-2	0,19



$$\tau = 94{,}17\,\%$$

Vecteurs	Technicité	Polyvalence	Créativité	Facteur 1	Facteur 2
GM1	1	2	2	2	2,11
GM2	-1	-2	-2	-2	-2,11
GM3	0	-2	-2	-2	-1,92
GM4	1	0	2	2	0,19
GM5	0	-2	2	2	-1,92
GM6	-1	0	-2	-2	-0,19
GM7	0	0	-2	-2	0
GM8	-1	0	2	2	-0,19
GM9	0	2	2	2	1,92
GM10	-1	-2	2	2	-2,11
GM11	0	0	2	2	0
GM12	0	2	-2	-2	1,92
GM13	1	2	-2	-2	2,11
GM14	1	0	-2	-2	0,19

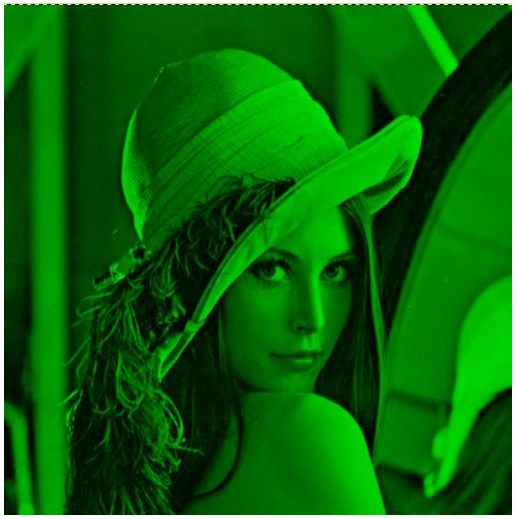
Pour donner un sens aux deux axes, on peut utiliser la corrélation:

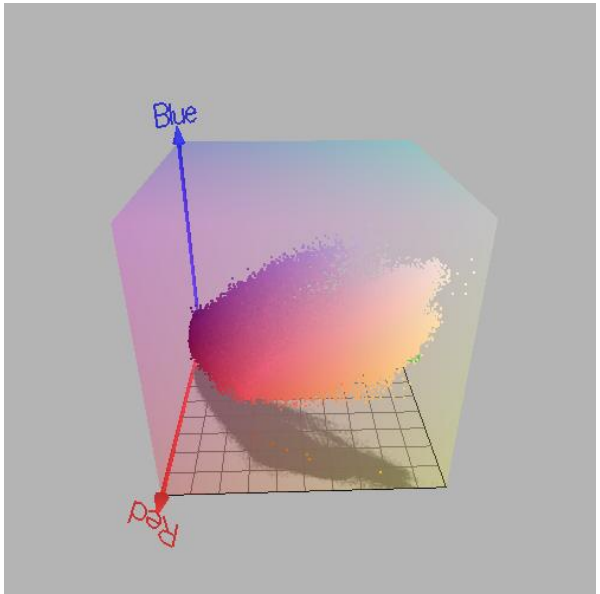
Corrélations	Technicité	Polyvalence	Créativité
Variables / axe 1	0	0	***1***
Variables / axe 2	***0,57***	***0,996***	0

Exemple 2: Imagerie

- Chaque image en couleur: rouge, vert, bleu (RGB), contient une information couleur sur l'intensité du rouge, l'intensité du vert et l'intensité du bleu.
- Il est donc possible de diviser par trois la taille d'une image, en ne conservant qu'un seul canal.
  - (RGB --> ACP1) , (x; y; z) -- > x')

Image RGB





ACP sur une image couleur

- Matrice de covariances:

$$P = \begin{pmatrix} var(R) & cov(R,G) & cov(R,B) \\ cov(G,R) & var(G) & cov(G,B) \\ cov(B,R) & cov(B,G) & var(B) \end{pmatrix}$$

- Matrice des vecteurs propres:

$$V = \begin{pmatrix} 0.614 & 0.588 & 0.526 \\ -0.581 & -0.114 & 0.806 \\ 0.5346 & -0.801 & 0.271) \end{pmatrix}$$

- Matrice des valeurs propres:

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

$$= \begin{pmatrix} 2718 & 0 & 0 \\ 0 & 110 & 0 \\ 0 & 0 & 11) \end{pmatrix}$$

- La conservation de l'axe principale permet d'expliquer plus 90% de l'information:

$$\frac{\lambda_1}{\sum_{i=1}^{3}(\lambda_i)} > \tau(0.90)$$

ACP sur une image couleur

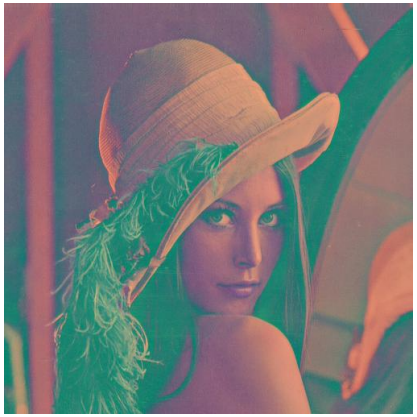
- Projection des données originales sur les axes factoriels:

$$P = \begin{pmatrix} RACP1 \\ RACP2 \\ RACP3 \end{pmatrix} = \begin{pmatrix} 0.614 & 0.588 & 0.526 \\ -0.581 & -0.114 & 0.806 \\ 0.5346 & -0.801 & 0.271) \end{pmatrix} . \begin{pmatrix} R \\ G \\ B \end{pmatrix}$$

- Exemple : pour l'axe principal (1)
- RACP1 = R*0.614 + G*0.588 + B*0.526

- Image originale.

- Image projetée sur les trois axes de l'ACP.



- Image projetée sur l'axe principale.

Exemple 3: notes sujet/matière

- 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

matrice centrée-réduite

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

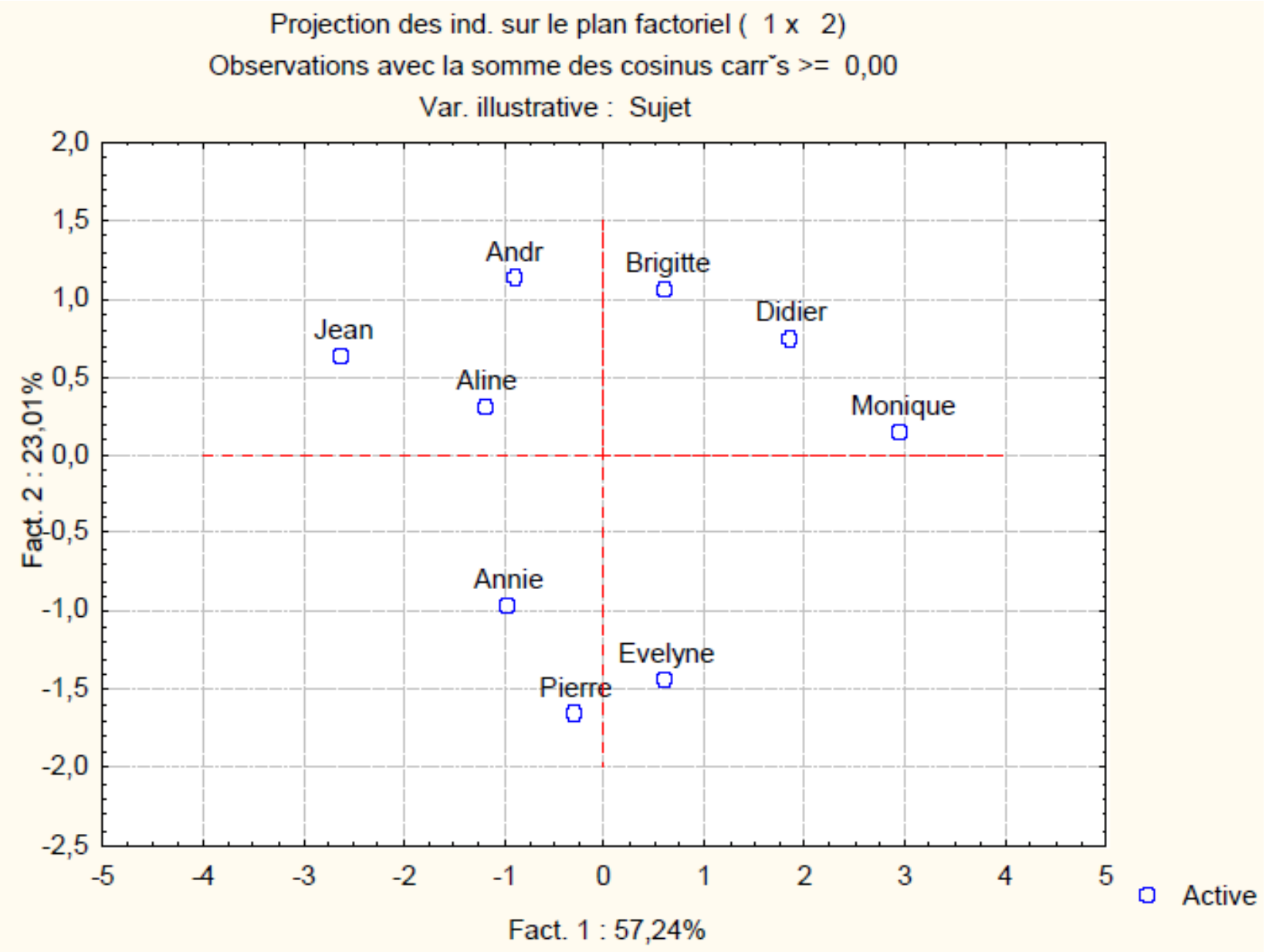
Valeurs propres et inerties

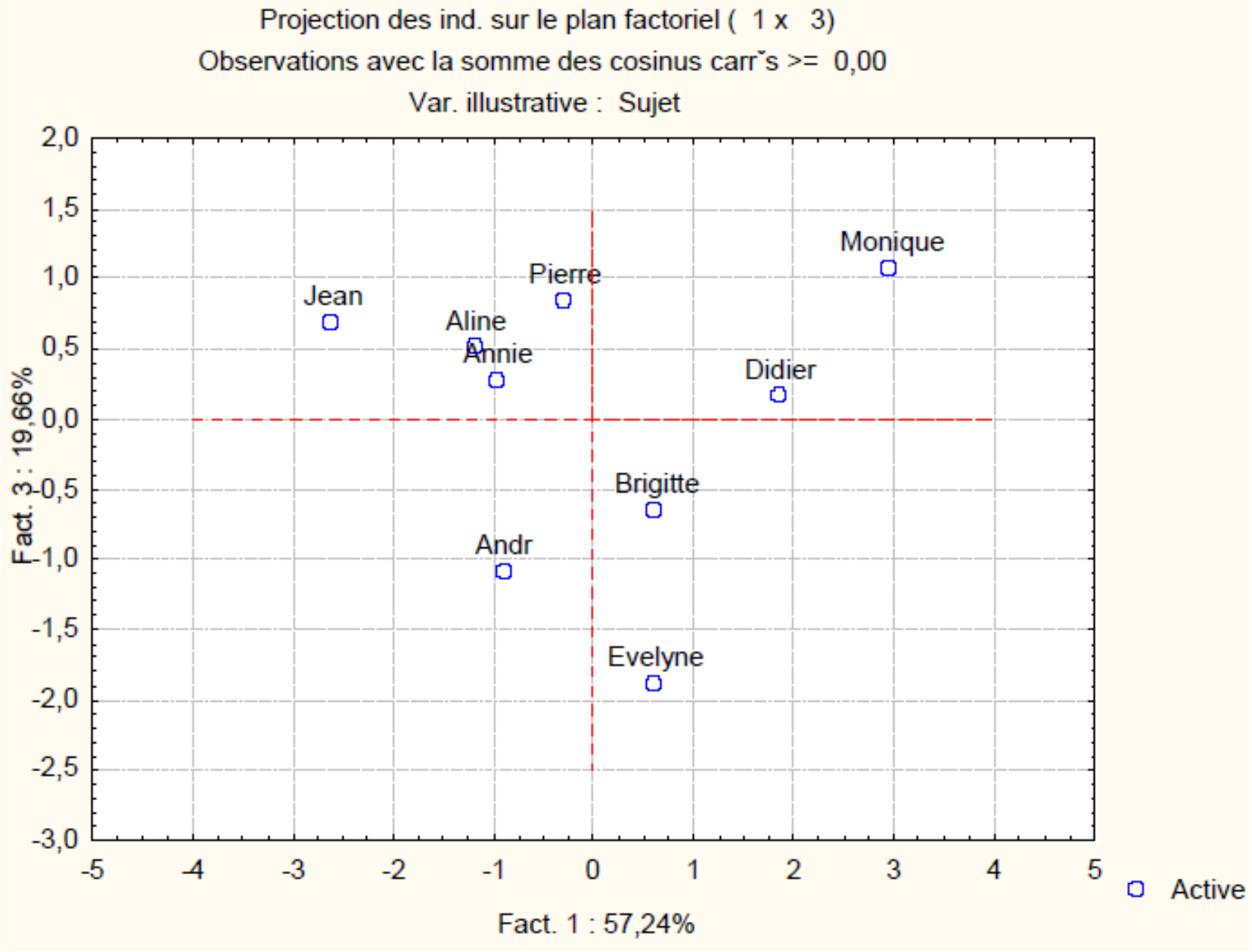
	Val. propr	Variance (%)	Variance cumul (%)
1	2,8618	57,24	57,24
2	1,1507	23,01	80,25
3	0,9831	19,66	99,91
4	0,0039	0,08	99,99
5	0,0004	0,01	100,00

Résultats relatifs aux individus

- **Scores des individus**
  - Les scores des individus sont les valeurs des composantes principales sur les individus :

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


Projection des ind. sur le plan factoriel (1 x 2)
Observations avec la somme des cosinus carr˜s >= 0,00
Var. illustrative : Sujet
Andr
Brigitte
Didier
Jean
Aline
Monique
Annie
Evelyne
Pierre
Fact. 2 : 23,01%
Fact. 1 : 57,24%
Active


Projection des ind. sur le plan factoriel (1 x 3)
Observations avec la somme des cosinus carr˜s >= 0,00
Var. illustrative : Sujet
Fact. 3 : 19,66%
Fact. 1 : 57,24%
Jean
Aline
Annie
Pierre
Monique
Didier
Brigitte
Andr
Evelyne
Active

Contributions des individus

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

$$CTR_{\alpha}(i) = \frac{(Score\ de\ i\ sur\ l'axe\ \alpha)^2}{n\,\lambda_{\alpha}}$$

- Par exemple : $$CTR_1(Jean) = \frac{(-2,7857)^2}{9 \times 2,8618} = 0,3013$$
- *Contributions des individus exprimées en pourcentages*

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

Qualités de la représentation des individus

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

$$QLT_{\alpha}(i) = \frac{(Score\,de\ i\ \ sur\ \ l'axe\ \alpha)^2}{\sum_{l}(Score\,de\ i\,sur\,l'axe\ \ l)^2}$$

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

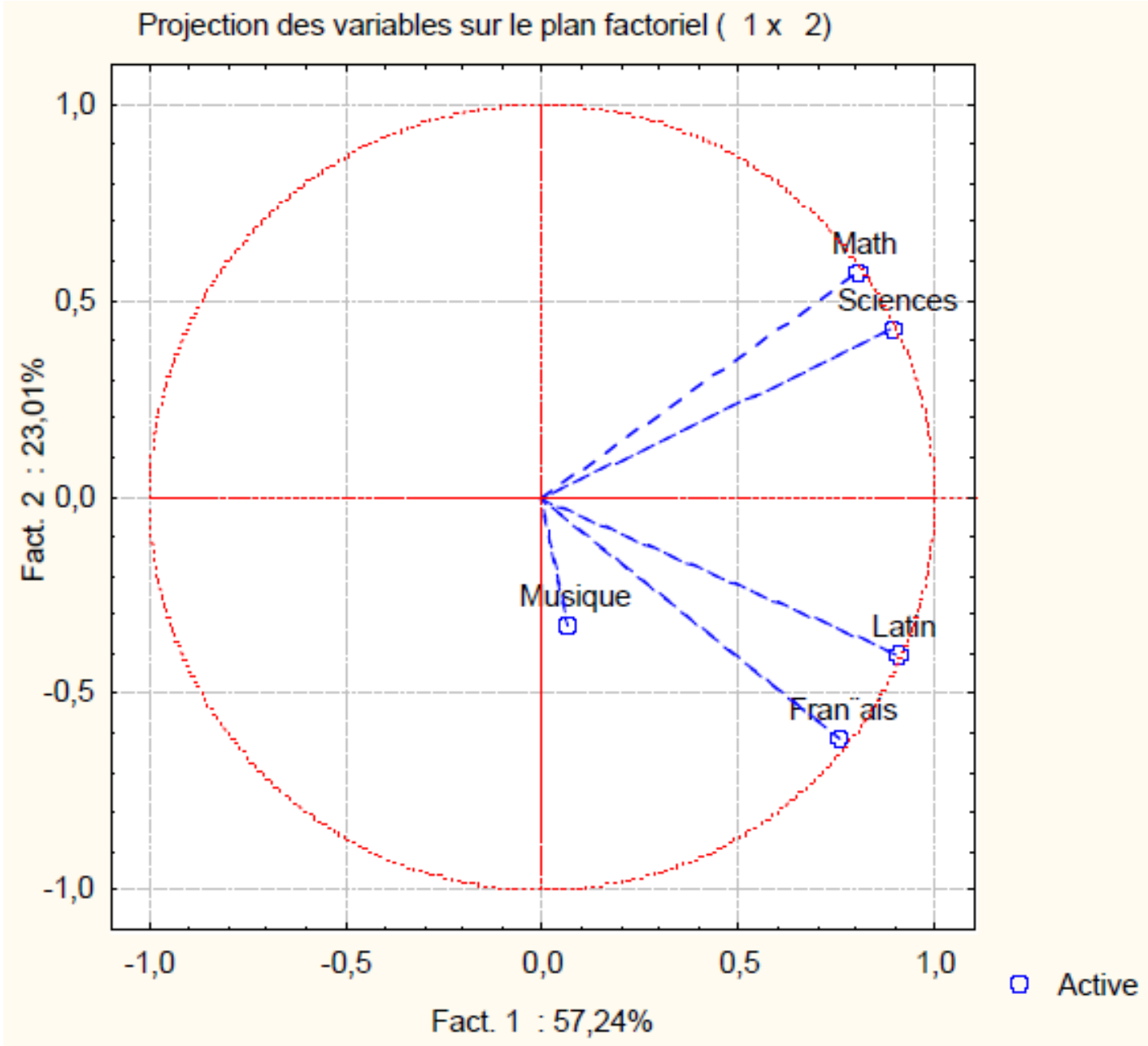
Résultats relatifs aux variables

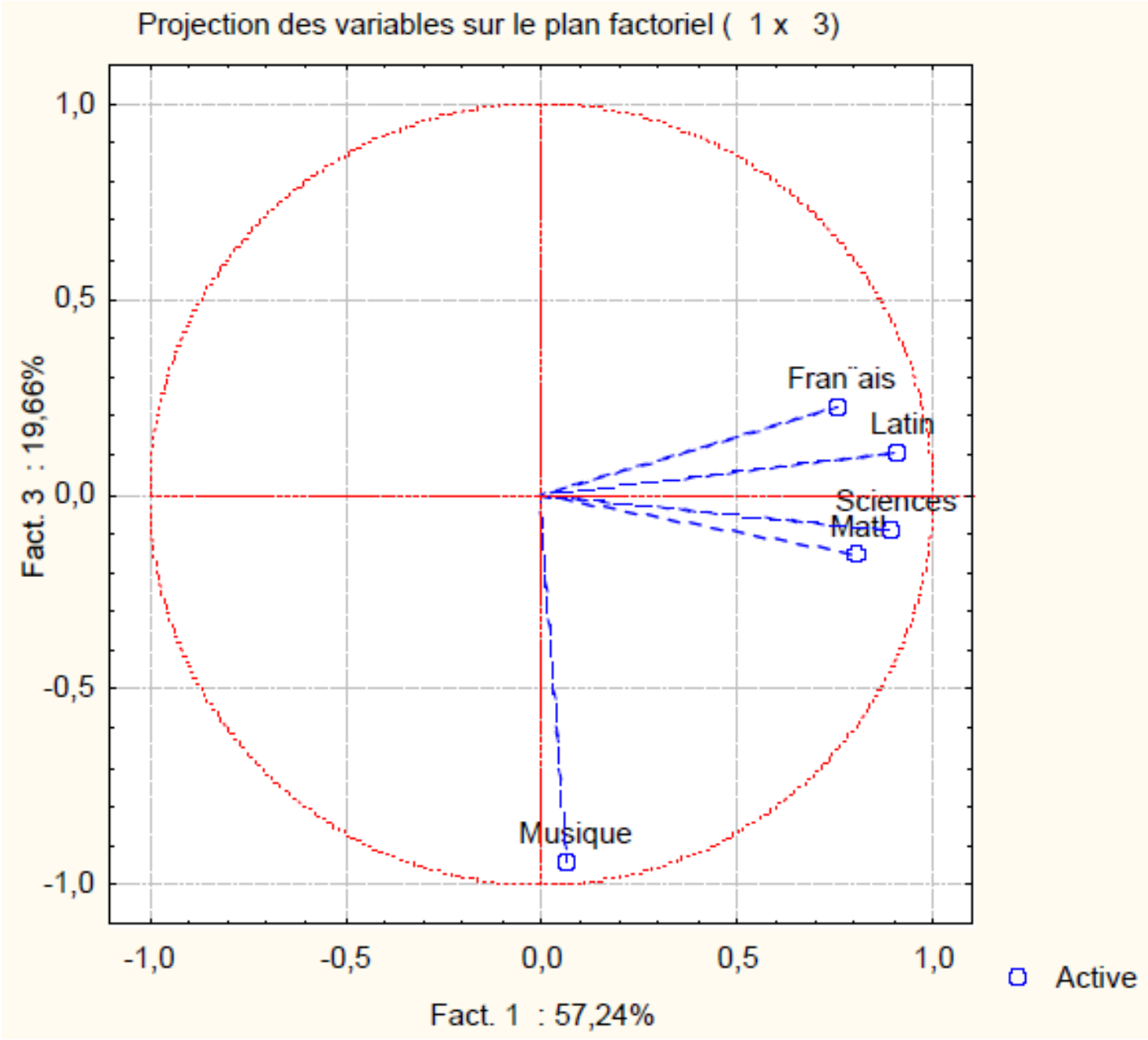
- ***Saturation des variables*** Les saturations des variables sont les coordonnées factorielles des variables. Elles sont égales au coefficients de corrélation entre les variables (centrées réduites) de départ et les scores des individus

$$\phi_{\text{Math},1} = \frac{(-1{,}0865)(-2{,}7857)+(-0{,}4939)(-1{,}2625)+(-1{,}0865)(-1{,}0168)+(1{,}4322)(3{,}1222)}{9\sqrt{2{,}8618}}+$$

$$\frac{(1{,}2840)(1{,}9551)+(0{,}3951)(-0{,}9478)+(-1{,}2347)(-0{,}3250)+(0{,}9877)(0{,}6373)+(-0{,}1975)(0{,}6231)}{9\sqrt{2{,}8618}}$$

	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


Projection des variables sur le plan factoriel (1 x 2)
1,0
0,5
0,0
-0,5
-1,0
Fact. 2 : 23,01%
Math
Sciences
Musique
Latin
Fran ais
-1,0
-0,5
0,0
0,5
1,0
Fact. 1 : 57,24%
Active


Projection des variables sur le plan factoriel (1 x 3)
1,0
0,5
0,0
-0,5
-1,0
Fact. 3 : 19,66%
Fran¨ais
Latin
Sciences
Math
Musique
-1,0
-0,5
0,0
0,5
1,0
Fact. 1 : 57,24%
Active

Contributions des variables

- Les contributions des variables à la formation des composantes principales sont définies de la même façon que
- celles des individus. Par exemple :

$$CTR_1(Math) = \frac{0{,}8059^2}{2{,}8618} = 0{,}2269$$

- *Contributions des variables*

	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 :

$$QLT_{\alpha}(j) = \frac{(Saturation\ de\ j\ sur\ l'axe\ \alpha)^2}{\sum_{l}(Saturation\ de\ j\ sur\ \ l'axe\ l)^2} = (Saturation\ de\ j\ sur\ l'axe\ \alpha)^2$$

- 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.

	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

Les objectifs de l'analyse des composantes principales

- Tableau individus/variables:
  - Visualiser le positionnement des individus les uns par rapport aux autres
  - Visualiser les corrélations entre les variables
  - Interpréter les axes factoriels

L'ACP en trois transparents (1)

- **Données**
  les données représentent les valeurs de p variables mesurées sur n individus ; les individus peuvent avoir un poids. En général on travaille sur des données centrées réduites Z (on retranche la moyenne et on divise par l'écart type).

- **Matrice de corrélation**
  c'est la matrice R de variance covariance des variables centrées réduites. Elle possède p valeurs propres:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$$

- **Facteurs principaux u_k**
  ce sont les vecteurs propres orthonormés de R (de dimension p) associes aux valeurs propres k. Leur j-ieme composante u_{kj} est le poids de la variable j dans la composante k.

- **Composantes principales c_k**
  ce sont les vecteurs Zu_k de dimension n. Leur i-ieme coordonnée c_{ki} est la valeur de la composante k pour l'individu i. Les c_k sont decorrélées.

L'ACP en trois transparents (2)

- **Nombre d'axes**
  on se contente souvent de garder les axes interprétables de valeur propre supérieure à 1. La qualité de la représentation retenue est mesuré par la part d'inertie expliquée par ces composantes.

- **Cercle des corrélations**
  il permet de visualiser comment les variables sont corrélées (positivement ou négativement) avec les composantes principales. A partir de la, on peut soit trouver une signification physique à chaque composante, soit montrer que les composantes séparent les variables en paquets. Seules les variables bien représentées (situées près du bord du cercle) doivent être interprétées.

L'ACP en trois transparents (3)

- **Représentation des individus pour un plan principal donné**,
  la représentation des projections des individus permet de conformer l'interprétation des variables. On peut aussi visualiser les individus aberrants (erreur de donnée ou individu atypique).

- **Contribution d'un individu à une composante**
  c'est la part de la variance d'une composante principale qui provient d'un individu donné. Si cette contribution est très supérieure aux autres, on peut avoir intérêt a mettre l'individu en donnée supplémentaire.

- **Qualité globale de la représentation**
  c'est la part de l'inertie totale I_g qui est expliquée par les axes principaux qui ont été retenus. Elle permet de mesurer la précision et la pertinence de l'ACP.

- **Qualité de la représentation d'un individu**
  elle permet de vérifier que tous les individus sont bien représentes par le sous-espace principal choisi; elle s'exprime comme le carré du cosinus de l'angle entre l'individu et sa projection orthogonale.