The Biotechnet event / Evénement Biotechnet

Kempinski Palace, Djibouti

25 Octobre 2022

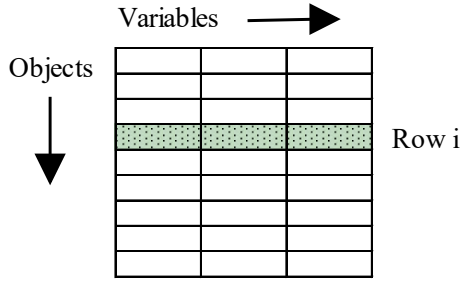
MODULE 3 - Analyse en composantes principales (ACP)

Pr. Tarik Ainane

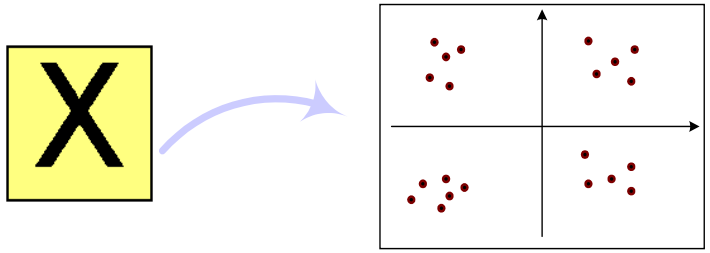
OBJECTIFS :

1 - Notions préliminaires.
2 - Interprétation des résultats.
3 - Exemples dans la recherche scientifique.

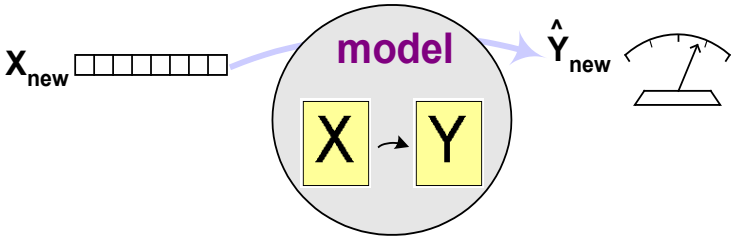
Objectifs des méthodes multivariées



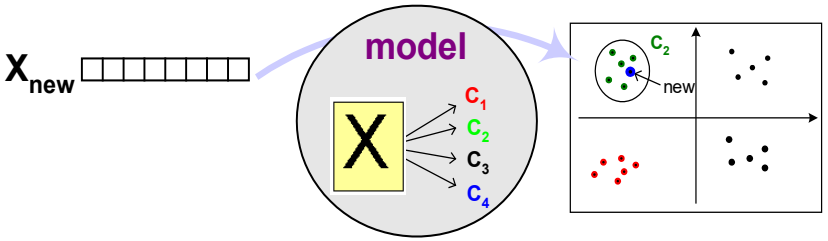
- Explorer et Décrire
  ACP

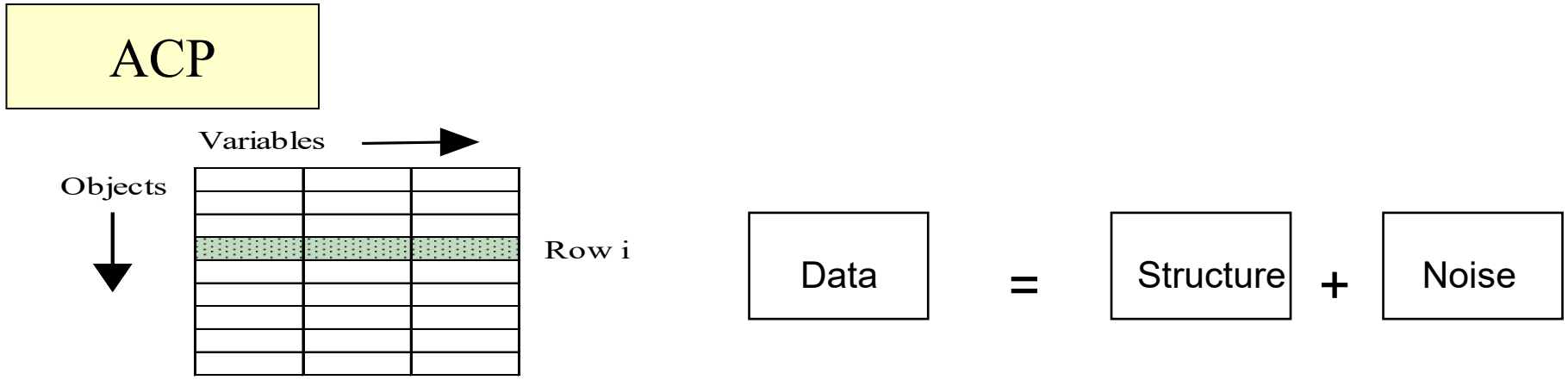


- Corréler et Prédire
  Régressions



- Caractériser et Classifier
  Classificassions et discrimination



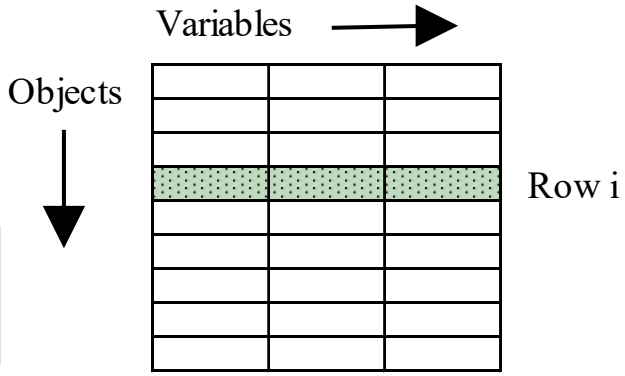


Objectifs de l'ACP (PCA)

Analyse en Composante Principale (Principal Component Analysis)

- Déterminer un moyen efficace de cartographier les échantillons (objets, individus)
  - Echantillons similaires proches les uns des autres;
  - Echantillons dissimilaires distants les uns des autres.
- Extraire le maximum d'information
- Réduire la dimension
- Créer de nouvelles variables

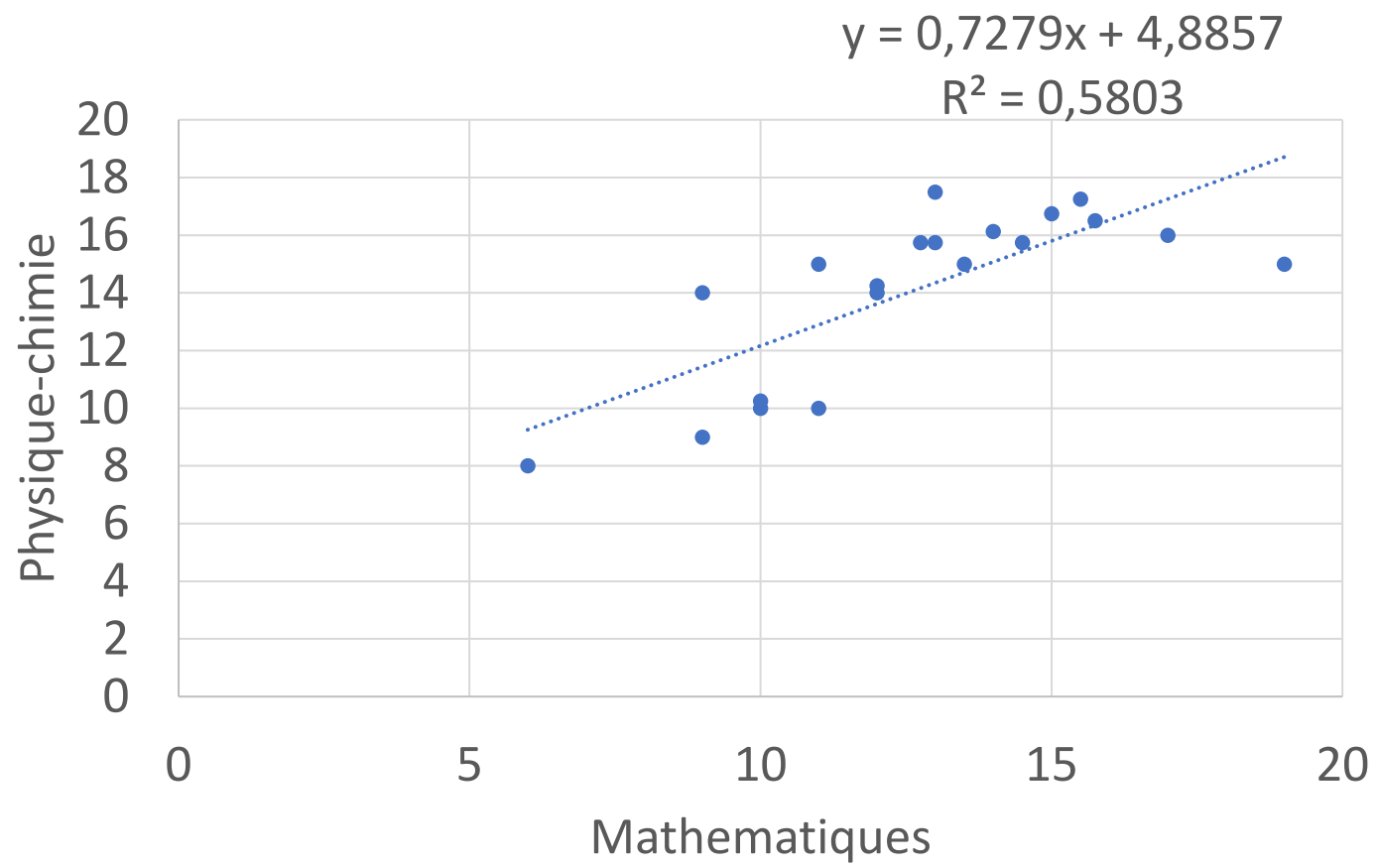
Tableau (notes des étudiants)
20 étudiants (individus)
6 variables (notes des matières



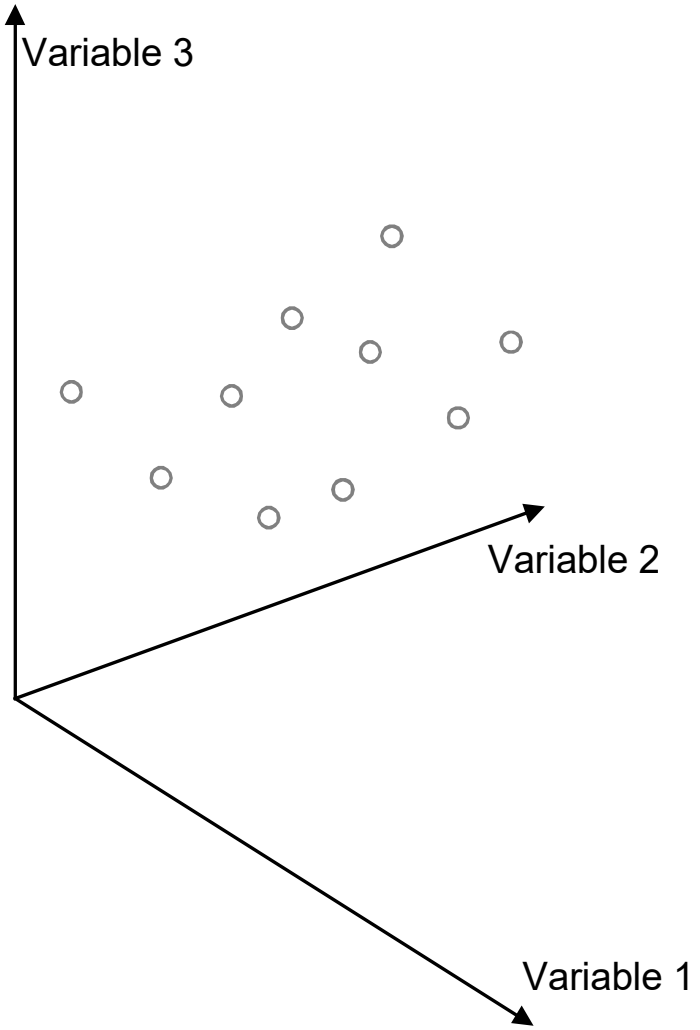
Chaque individu (lignes) a 6 coordonnées. Ce serait bien d'avoir une représentation plus simple

N°	Nom	Mathematique	Physique-chimie	Siences naturelles	Histoire-geographie	Littérature	Philosophie
1	Fatouma	17	16	17	16	16	16,5
2	Tarik	14,5	15,75	8	6	10	17
3	Hajar	19	15	12	16	12	15,5
4	Hasna	12,75	15,75	10	17	17	16
5	Hanine	13,5	15	12	6	8	14,5
6	Talal	10	10,25	10	10	10	12
7	Mohamed	15,5	17,25	13	12,5	13	17,75
8	Halima	13	17,5	12,5	12	12	18
9	Ali	10	10	14	10	7	12
10	Ayoub	12	14	12	6	8	7
11	Abdelaziz	9	9	12	17	16	12
12	Nabil	11	15	10	6	8	15
13	Khadija	11	10	10	16	17	18
14	Ahmed	14	16,125	8	6	6,98	14
15	Said	13	15,75	8	6,5	7,315	18
16	Houda	6	8	12	12	12	12
17	Saad	9	14	9	14	12	15
18	Sara	15,75	16,5	10	10	13,82	15
19	Abdellah	12	14,25	6	2	3,56	8
20	Majda	15	16,75	7	7	7,28	14

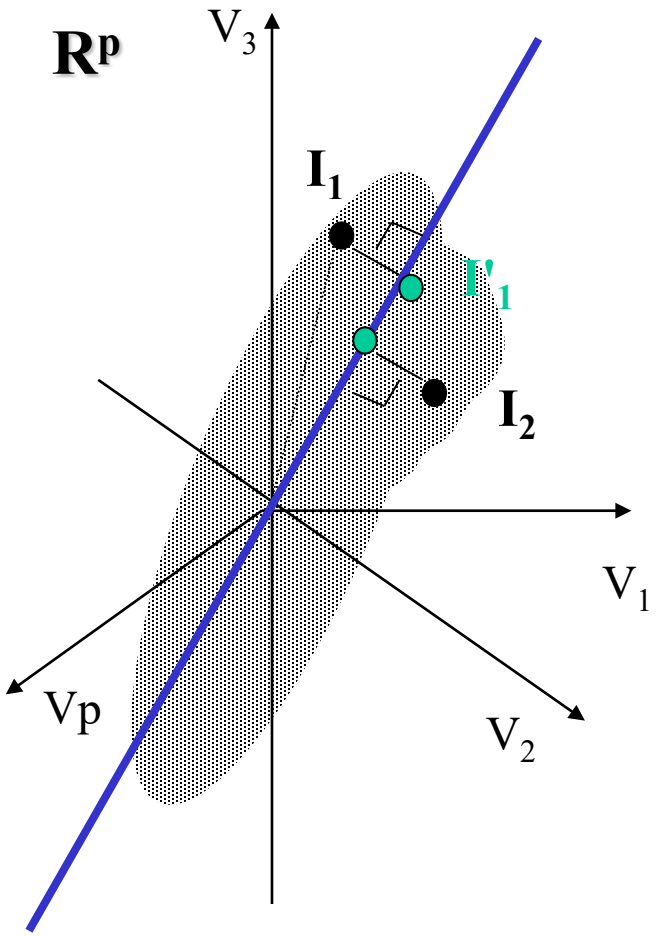
Corrélations entre 2 variables :
Physique-chimie
Mathématiques



Corrélations entre 3 variables :
Physique-chimie
Mathématiques
Sciences naturelles



Corrélations entre plusieurs variables :



Sous-espace à 1 dimension: Droite

Droite de projection qui va donner l'image la plus « réelle » du nuage de point,

Droite de projection qui va conserver au mieux les distances entre les individus (distance euclidienne),

Droite d'étirement maximum du nuage des points projetés,

Droite de variance (d'inertie) maximum du nuage des points projetés,

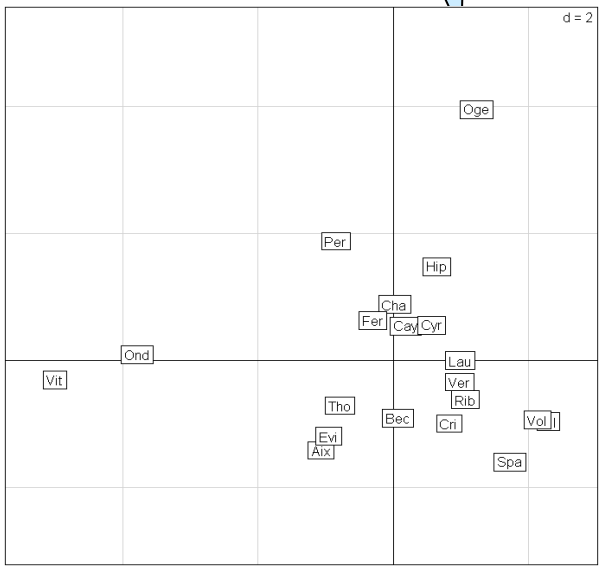
ACP

Représentation dans un plan comme si je n'avais que 2 variables

Tableau (notes des étudiants)
20 étudiants (individus)
6 variables (notes des matières

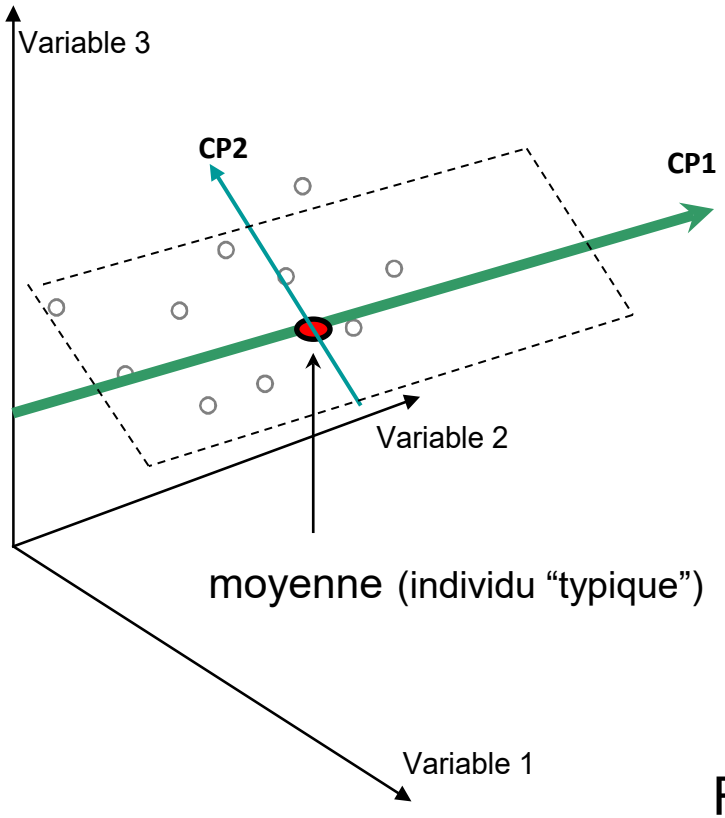
N°	Nom	Mathematique	Physique-chimie	Siences naturelles	Histoire-geographie	Littérature	Philosophie
1	Fatouma	17	16	17	16	16	16,5
2	Tarik	14,5	15,75	8	6	10	17
3	Hajar	19	15	12	16	12	15,5
4	Hasna	12,75	15,75	10	17	17	16
5	Hanine	13,5	15	12	6	8	14,5
6	Talal	10	10,25	10	10	10	12
7	Mohamed	15,5	17,25	13	12,5	13	17,75
8	Halima	13	17,5	12,5	12	12	18
9	Ali	10	10	14	10	7	12
10	Ayoub	12	14	12	6	8	7
11	Abdelaziz	9	9	12	17	16	12
12	Nabil	11	15	10	6	8	15
13	Khadija	11	10	10	16	17	18
14	Ahmed	14	16,125	8	6	6,98	14
15	Said	13	15,75	8	6,5	7,315	18
16	Houda	6	8	12	12	12	12
17	Saad	9	14	9	14	12	15
18	Sara	15,75	16,5	10	10	13,82	15
19	Abdellah	12	14,25	6	2	3,56	8
20	Majda	15	16,75	7	7	7,28	14

Merci l'ACP

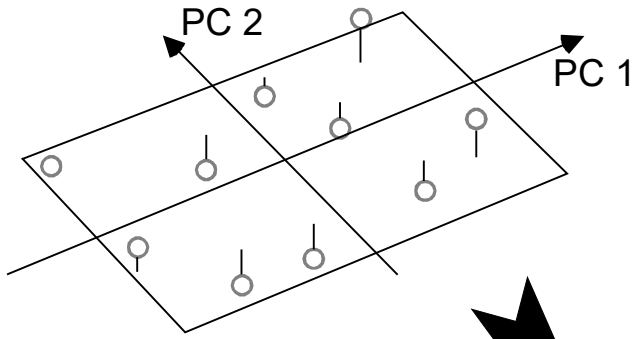


ACP

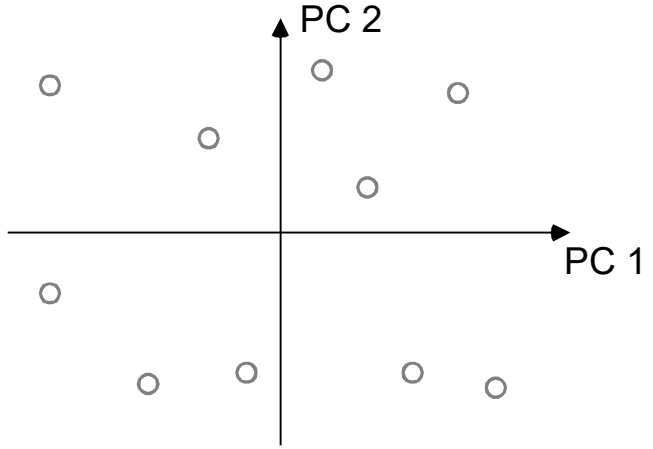
Plan principal



moyenne (individu “typique”)



Représentation classique du plan



Composantes principales
Facteurs, Axes

Principe de la méthode

Le principe de l'ACP est de réduire la dimension des données initiales (qui est p si l'on considère p variables quantitatives), en remplaçant les p variables initiales par q facteurs appropries ($q < p$). Les q facteurs cherchés sont des moyennes pondérées des variables initiales. Leur choix se fait en maximisant la dispersion des individus selon ces facteurs (variance maximum). Des techniques mathématiques appropriées permettent de réaliser tout cela de façon automatique et optimale,

Méthodologie

L' ACP est appliquée sur p variables quantitatives notées $X_1, \dots, X_j, \dots, X_p$ observées sur n individus notés $1, \dots, i, \dots, n$. L'observation de la variable X_j observées sur l'individu i est $x_{j.i}$.

Donc l'ensemble des informations se représente de la manière suivante :

	$\boldsymbol{X_1}$	...	$\boldsymbol{X_j}$	...	$\boldsymbol{X_p}$
1	$x_{1.1}$	...	$x_{j.1}$	...	$x_{p.1}$
...	...		...		...
$\boldsymbol{i}$	$x_{1.i}$	...	$x_{j.i}$	...	$x_{p.i}$
...	...		...		...
$\boldsymbol{n}$	$x_{1.n}$	...	$x_{j.n}$	...	$x_{j.p}$

Les q facteurs que l'on va définir, pour résumer l'information contenue dans le tableau initial, doivent maximiser la dispersion du nuage des observations. Généralement, lorsqu'on dispose d'un nuage d'observations en plusieurs dimensions, on parle d'inertie (somme des variances des variables considérées).

En passant de la dimension initiale p à la dimension réduite q, on perd, obligatoirement, de la dispersion, de l'inertie. L'idée est de choisir les facteurs convenables pour perdre le moins possible la dispersion.

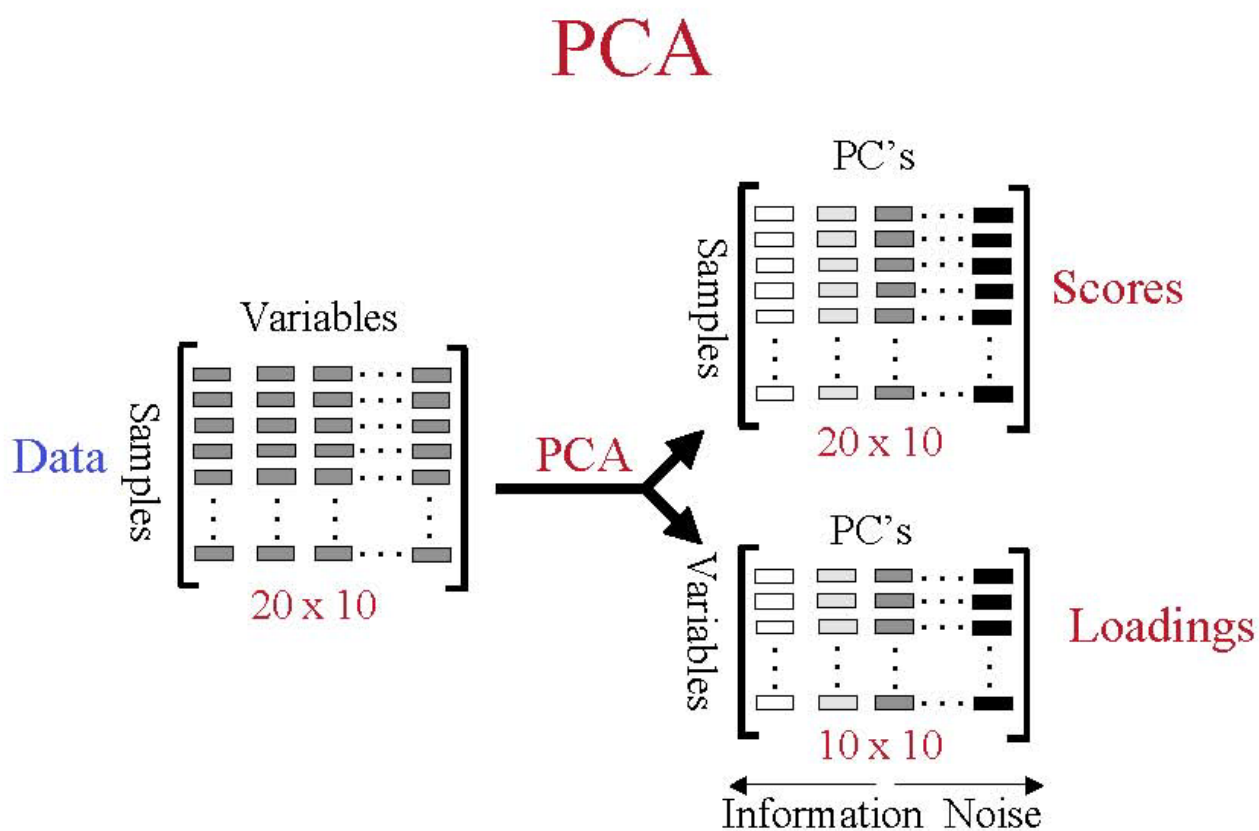
On cherche des combinaisons linéaires des variables initiales, appelées facteurs, ou encore composantes principales, s'écrivant sous la forme suivante :

$$C_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p$$

...

$$C_n = a_{1n}X_1 + a_{2n}X_2 + \cdots + a_{pn}X_p$$

ACP


PCA
PC's
Samples
Scores
20 x 10
Variables
Data
Samples
PCA
20 x 10
PC's
Variables
Loadings
10 x 10
Information
Noise

ACP

Interprétation des résultats

Regarder les **pourcentages de variation expliqués par chaque composante principale**.

Déterminer le nombre d'axes à examiner

→ Il faut examiner les axes jusqu'à obtenir une information « suffisante » (% de variation)

→ Il faut tenir compte de la forme dégressive des valeurs propres.

n	Valeur	Pourcent	Cumul
1	2.8807	72.02	72.02
2	0.6453	16.13	88.15
3	0.3897	9.74	97.89
4	0.0844	2.11	100.00

Variance totale = 4.0

ACP

Interprétation des résultats

Regarder la **structure des variables** à partir de leurs **corrélations** avec les axes principaux.

Qualité de la représentation

→ Cosinus carrés (COR)

→ Les variables sont d'autant mieux représentées sur le plan qu'elles sont proches du cercle

Structures des variables

→ Pour chaque axe, on regarde les variables qui lui sont les plus fortement corrélées.

→ On compare la position de ces variables les unes par rapport aux autres.

→ On peut ainsi interpréter cet axe.

→ On peut aussi étudier la position des variables par rapport aux deux axes et chercher une explication.

ACP

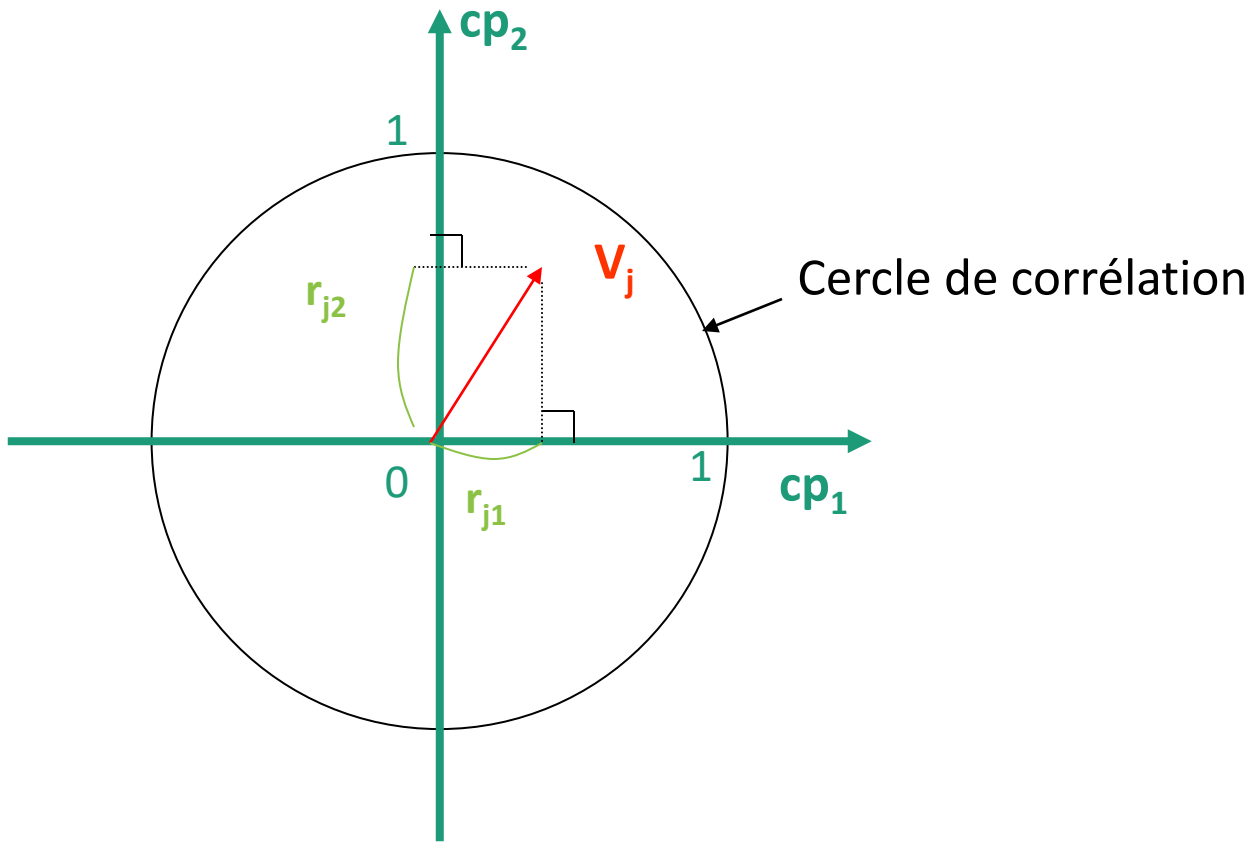
Interprétation des résultats

$$\cos^2\theta = (r_j^\alpha)^2$$

: Cosinus carré

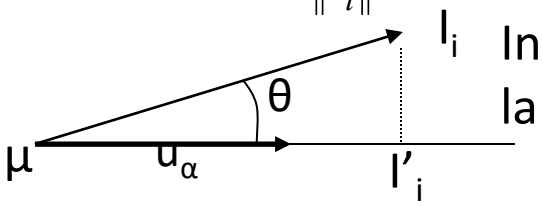
: Qualité de représentation de la variable j

coefficients de corrélation entre les variables et les composantes principales.



Regarder la répartition des individus à partir de leurs coordonnées sur les axes principaux et de la qualité de leur représentation.

- $\cos^2\theta = \frac{\|I_i'\|^2}{\|I_i\|^2}$: Cosinus carré
  : Qualité de représentation d'un individu i sur l'axe U_α



Indique dans quelle proportion l'axe α contribue à la représentation de l'individu i

→ Un individu sera bien représenté sur un axe s'il est proche de l'axe i.e. si le $\cos^2\theta$ est élevé et inversement.

→ Un individu sera bien représenté sur un plan si la somme des $\cos^2\theta$ est forte.

→ On ne peut pas interpréter les proximités d'individus mal représentés.

Les individus qui contribuent le plus à la formation des axes sont les individus qui ont les fortes coordonnées

- $\frac{(c_i^{\alpha})^2}{\lambda_{\alpha}}$: Contribution de l'individu i à l'axe principal α

  Indique dans quelle proportion l'individu i contribue à l'inertie λ_{α} du nuage projeté sur l'axe α

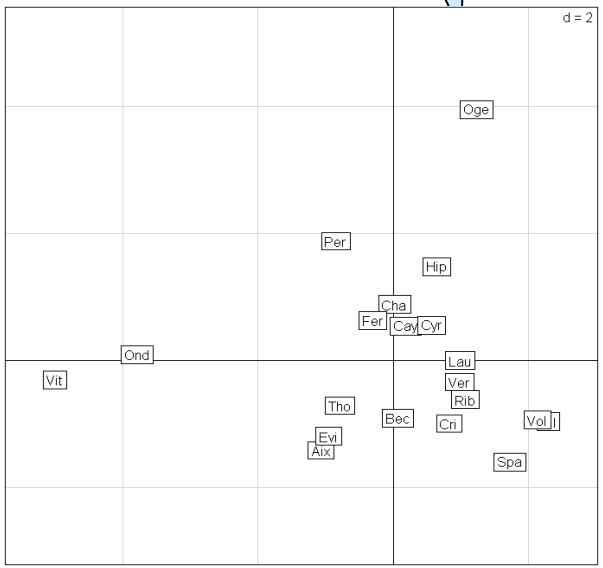
ACP

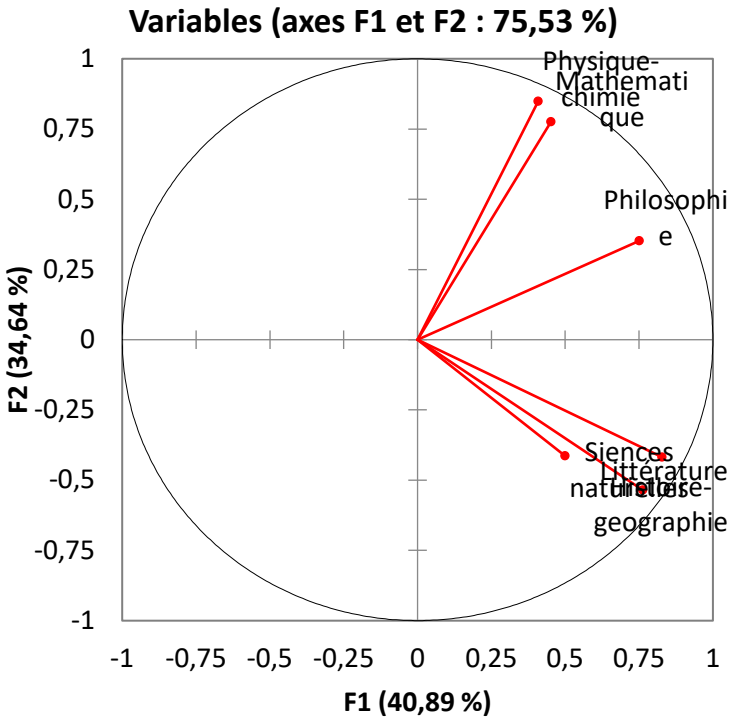
Représentation dans un plan comme si je n'avais que 2 variables

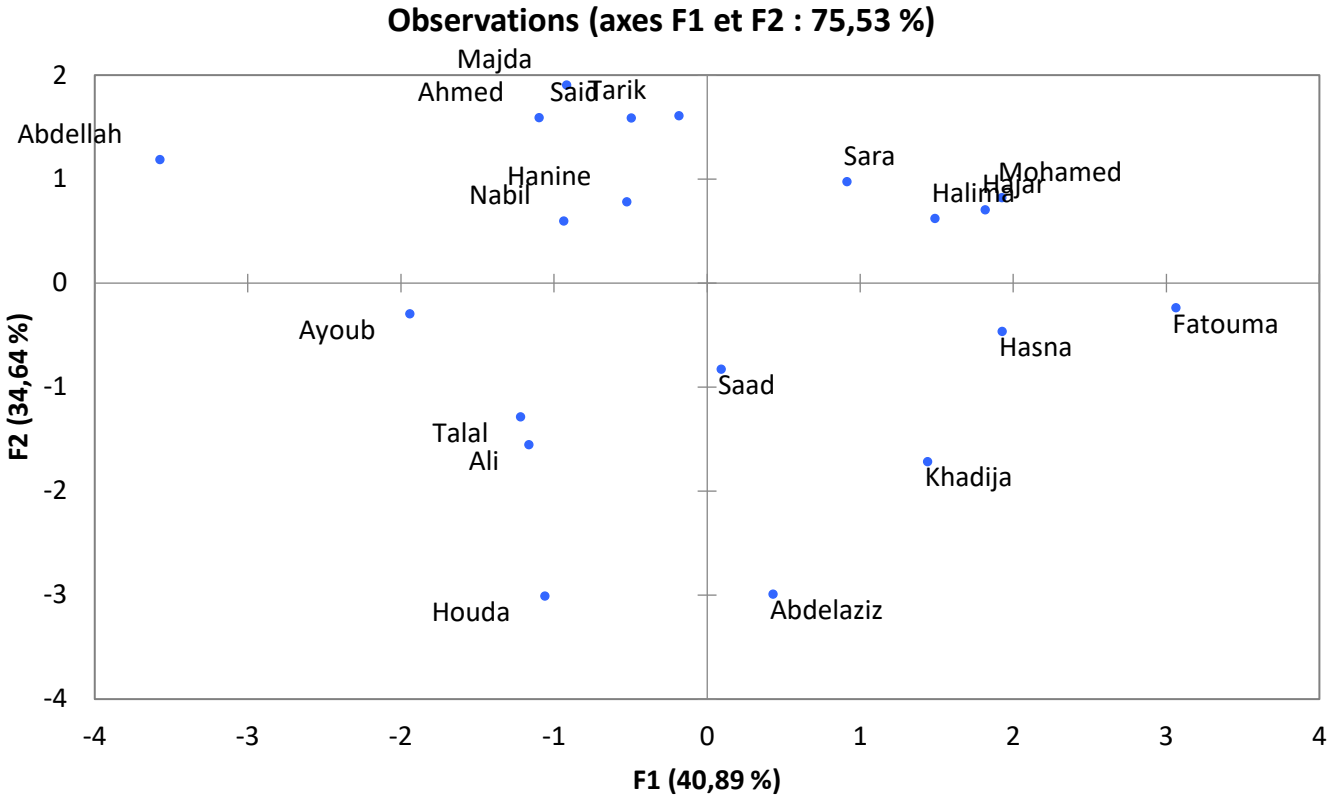
Tableau (notes des étudiants)
20 étudiants (individus)
6 variables (notes des matières

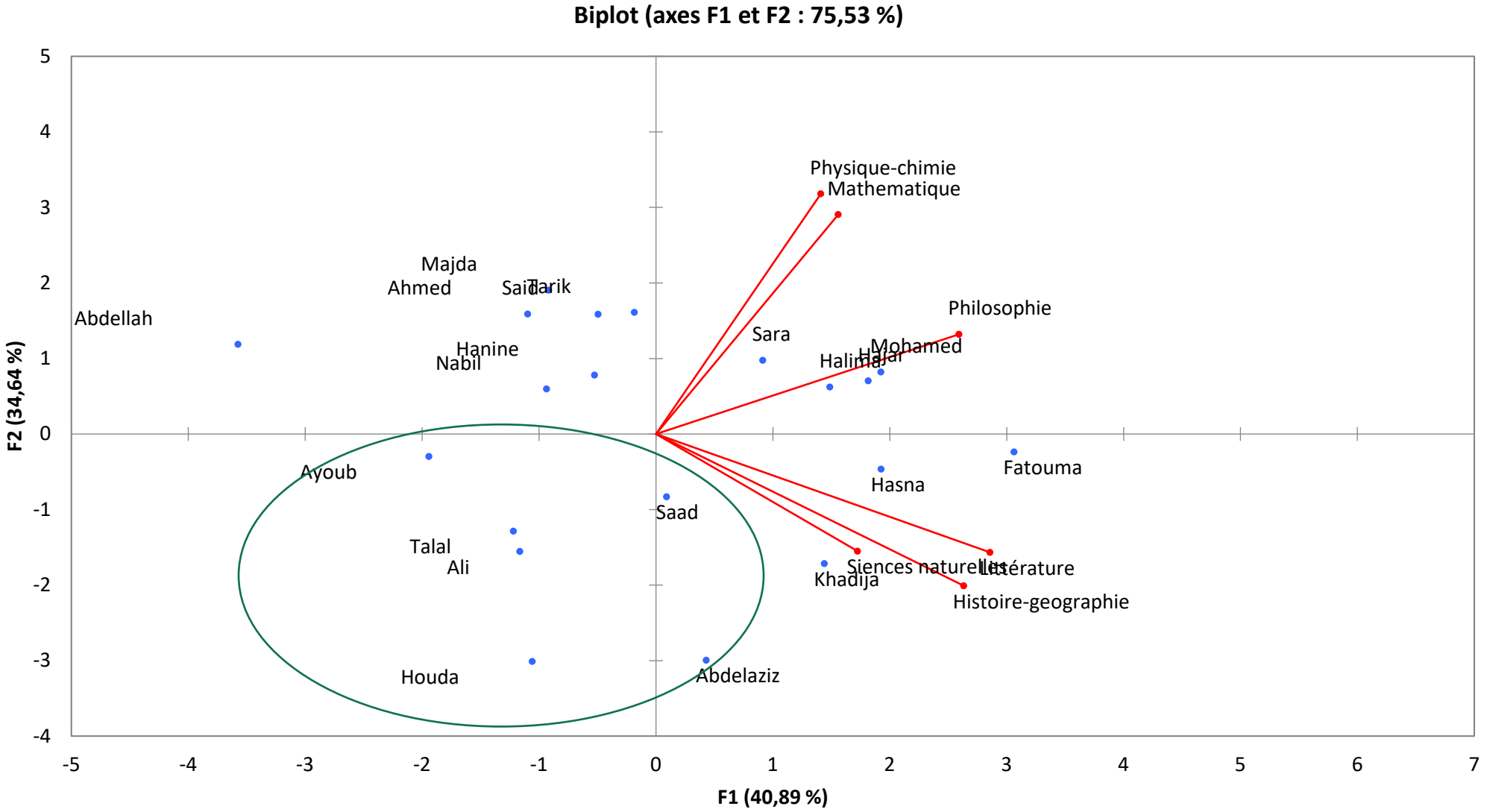
N°	Nom	Mathematique	Physique-chimie	Siences naturelles	Histoire-geographie	Littérature	Philosophie
1	Fatouma	17	16	17	16	16	16,5
2	Tarik	14,5	15,75	8	6	10	17
3	Hajar	19	15	12	16	12	15,5
4	Hasna	12,75	15,75	10	17	17	16
5	Hanine	13,5	15	12	6	8	14,5
6	Talal	10	10,25	10	10	10	12
7	Mohamed	15,5	17,25	13	12,5	13	17,75
8	Halima	13	17,5	12,5	12	12	18
9	Ali	10	10	14	10	7	12
10	Ayoub	12	14	12	6	8	7
11	Abdelaziz	9	9	12	17	16	12
12	Nabil	11	15	10	6	8	15
13	Khadija	11	10	10	16	17	18
14	Ahmed	14	16,125	8	6	6,98	14
15	Said	13	15,75	8	6,5	7,315	18
16	Houda	6	8	12	12	12	12
17	Saad	9	14	9	14	12	15
18	Sara	15,75	16,5	10	10	13,82	15
19	Abdellah	12	14,25	6	2	3,56	8
20	Majda	15	16,75	7	7	7,28	14

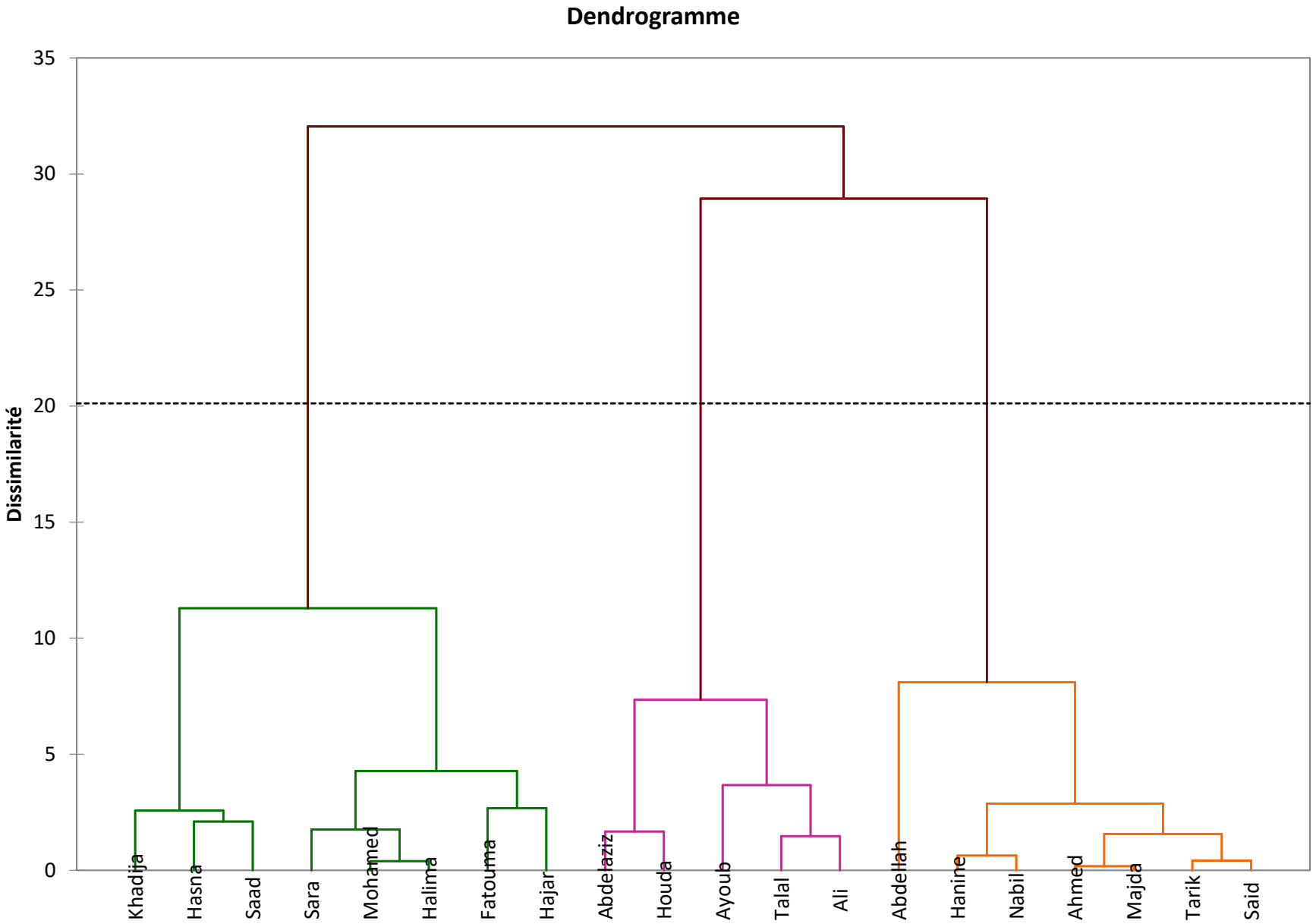
Merci l'ACP

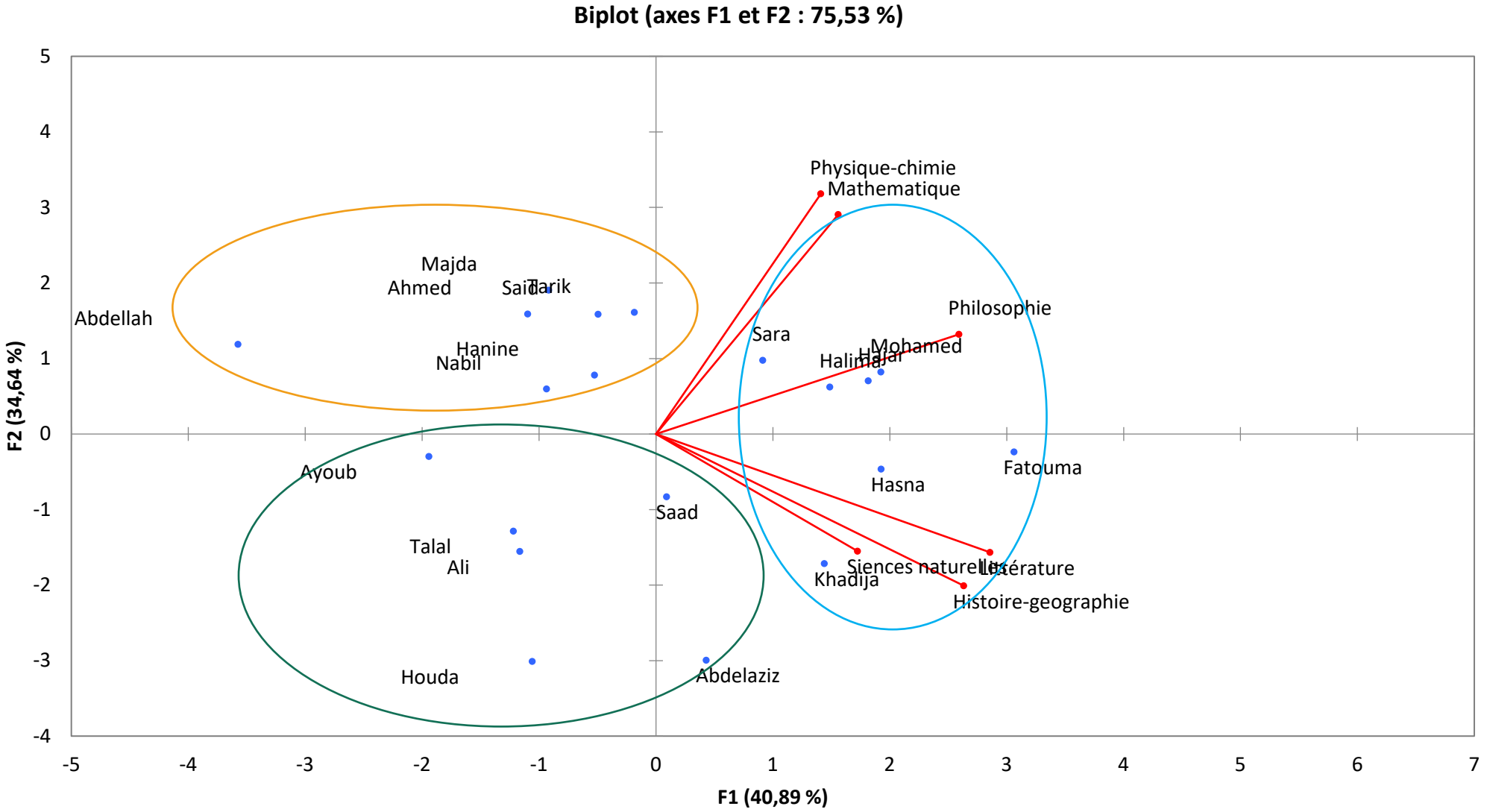



Variables (axes F1 et F2 : 75,53 %)
Physique-chimie
Mathematique
Philosophie
Sciences naturelles
Littérature
Histoire-geographie
F2 (34,64 %)
F1 (40,89 %)
1
0,75
0,5
0,25
0
-0,25
-0,5
-0,75
-1
-1
-0,75
-0,5
-0,25
0
0,25
0,5
0,75
1


Observations (axes F1 et F2 : 75,53 %)
Majda
Ahmed
Said
Tarik
Abdellah
Sara
Mohamed
Hanine
Nabil
Halima
Hajar
Ayoub
Fatouma
Hasna
Saad
Talal
Ali
Khadija
Houda
Abdelaziz
F2 (34,64 %)
F1 (40,89 %)
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4


Biplot (axes F1 et F2 : 75,53 %)
F2 (34,64 %)
F1 (40,89 %)
Physique-chimie
Mathematique
Philosophie
Mohamed
Sara
Halima
Hajar
Fatouma
Hasna
Saad
Littérature
Sciences naturelles
Khadija
Histoire-geographie
Majda
Ahmed
Said
Tarik
Abdellah
Hanine
Nabil
Ayoub
Talal
Ali
Houda
Abdelaziz


Dendrogramme
Dissimilarité
35
30
25
20
15
10
5
0
Khadija
Hasna
Saad
Sara
Mohamed
Halima
Fatouma
Hajar
Abdelaziz
Houda
Ayoub
Talal
Ali
Abdellah
Hanine
Nabil
Ahmed
Majda
Tarik
Said


Biplot (axes F1 et F2 : 75,53 %)
F2 (34,64 %)
F1 (40,89 %)
Physique-chimie
Mathematique
Philosophie
Sara
Mohamed
Halima
Hajar
Fatouma
Hasna
Saad
Siences naturelles
Khadija
Littérature
Histoire-geographie
Majda
Ahmed
Said
Tarik
Abdellah
Hanine
Nabil
Ayoub
Talal
Ali
Houda
Abdelaziz

FIN