Imen EH 🤓 //khaliwli l crédits amane

1-Crisp-DM

- **Définition:**

Crisp-DM : Cross Industry Standard Process for Data Mining

Méthodologie développée par IBM dans les année 60, son but est la réalisation de projets Datamining

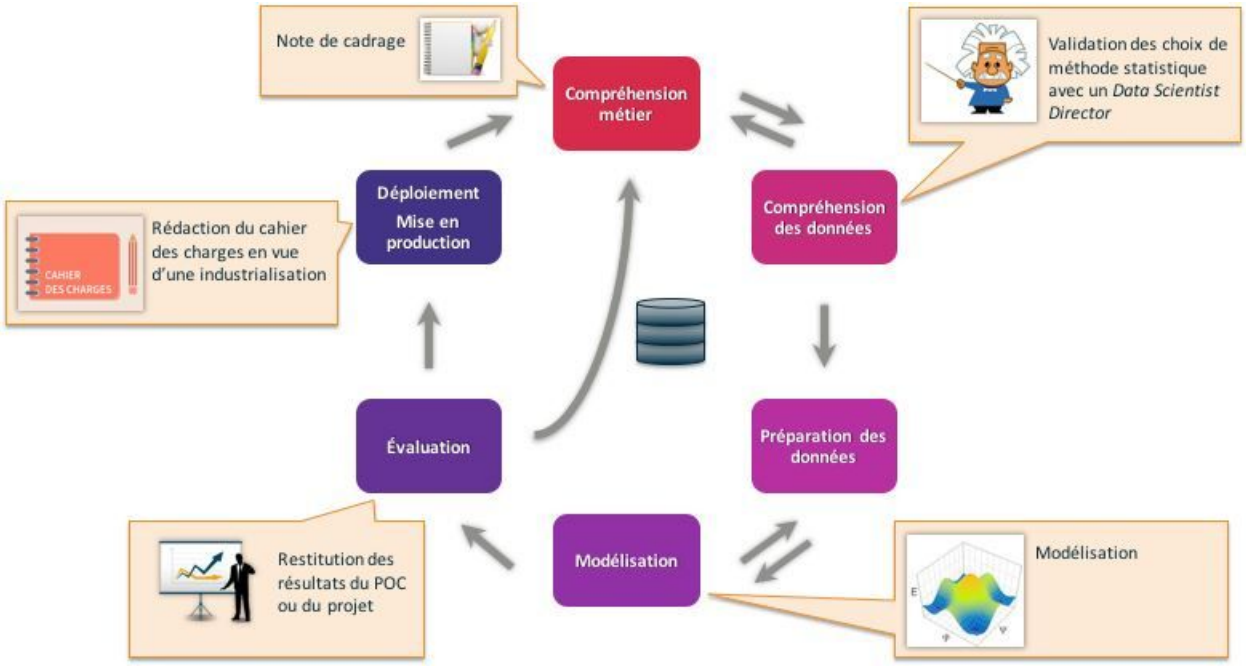
C'est la seule méthodologie efficace pour la réalisation de projets Data Science.

C'est un modèle de processus qui offre un aperçu du cycle de vie du Data Mining.

C'est aussi une méthodologie qui comprend des phases typiques d'un projet, et des tâches comprises dans chaque phase, et une explication des relations entre ces tâches.

- **Etapes:**

La méthodologie Crisp-dm se décompose en 6 étapes, où chaque étape à plusieurs sous-étapes:



1-Compréhension du problème métier (Business understanding) :

Dans cette étape, on vise à :

-déterminer les objectifs d'affaires

-identifier le problème qu'on voudrait résoudre

-évaluer la situation actuelle

-convertir en un problème de data mining :

Quels types de clients sont intéressés par notre produit (segmentation)

Quels sont les profils typiques de nos clients? (segmentation)

-élaborer un plan de projet

Exemple :

Si nous ne sommes pas familier avec le domaine médical et que nous voulons élaborer un projet de datamining pour la collecte de données médicales nous devons faire des recherches sur ce domaine et apprendre ses terminologies et détails pour qu'on puisse bien formuler le problème.

2-Compréhension des données (Data understanding) :

Dans cette étape, il faut bien définir le problème, donc, à bien préciser les données à analyser et à identifier leurs qualités et voir si elles répondent à nos besoins.

On peut aussi identifier de nouveaux problèmes d'où la nécessité de retourner à l'étape précédente.

Cette étapes englobe 5 tâches:

1) **Rassemblement de données (Gathering data)**
2) **Description de données (Describing data)**
3) **Exploration de données (Exploring data)**
4) **Vérification de la qualité de données (Verifying data quality)**
5) **Sélection des données :**
    a) Mettre en place une description du problème :
        i)Identifier les comportements de dépenses des femmes qui achètent

ii)des vêtements saisonniers

iii)Identifier les modèles de la faillite de détenteurs de cartes de crédit

b) Identifier les données pertinentes pour la description du problème

i)Données démographiques, données financières...

Les variables sélectionnées pour les données pertinentes doivent être indépendantes les unes des autres.

Les données peuvent provenir de nombreuses sources:

Internes (ERM / CRM / Data Warehouse..)

Externes (données commerciales/gouvernementales ..)

Créées (recherche)

3-Préparation des données :Construction du data hub (Data preparation) :

Convertir les données sous une forme qui sera compréhensible par l'outil de Data Mining.

1. Nettoyer les données sélectionnées pour une meilleure qualité
    a. Remplir les valeurs manquantes
    b. Identifier ou supprimer les valeurs aberrantes (salaire de la même personne différentes dans des fichiers différents)
    c. Résoudre la redondance causée par l'intégration des données
    d. Les données incohérentes correctes
2. Transformer les données
    a. Convertir des mesures différentes de données dans un échelle numérique unifié en utilisant des formulations mathématiques simples

4-Modélisation (Model building) :

La modélisation comprend le choix, le paramétrage et le test de différents algorithmes ainsi que leur enchaînement, qui constitue un modèle.

Ce processus est :

Descriptif : Pourquoi les choses se sont passées

Prédictif : Ce qui peut se passer

Prescriptif : Optimisation d'une situation future

Les techniques de DM : Association / Classification / Clustering / Prédictions / Les motifs séquentiels

5-Evaluation (Testing and evaluation) :

On vérifie si les résultats obtenus répondent aux objectifs énoncés dès le début du processus, en testant la robustesse et précision des modèles obtenus

Est-ce que le modèle répond aux objectifs métier?

Des objectifs métier importants non résolus?

Est-ce que le modèle est logique?

Est-ce que le modèle est actionnable?

Il devrait être possible de prendre des décisions après cette

étape.

Tous les objectifs importants doivent être atteints.

6-Déploiement (Deployment) :

Mise en production des modèles obtenus, son objectif est de mettre la connaissance obtenue par la modélisation dans une forme adaptée et de l'intégrer au processus de prise de décision.

En cours de suivi et d'entretien:

Évaluer la performance par rapport aux critères de réussite

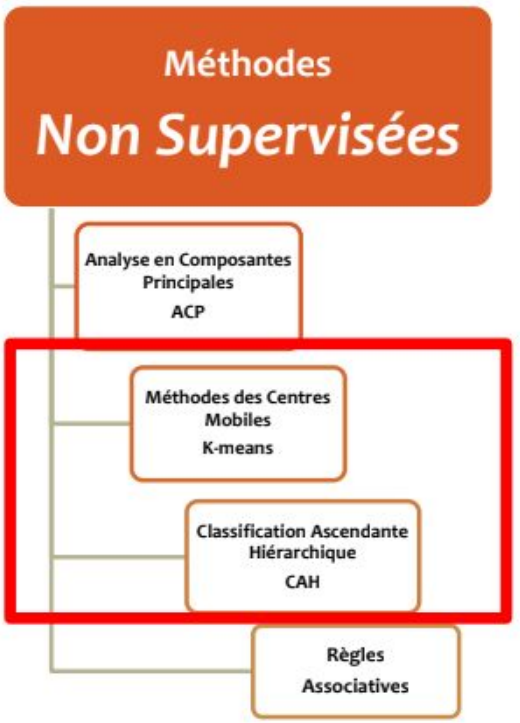
La réaction du marché et les changements des concurrents

! Remarque ! :

On passe le plus de temps dans les étapes 2 et 3 car elles sont très cruciales.

2-La ségmentation (clustering) > K-MEANS

La segmentation (clustering) est une technique non supervisée, elle fait partie des méthodes d'analyse descriptives



Utilité de la segmentation:

1-Banque/Assurance:

> Catégoriser la clientèle

> Regrouper les clients selon les critères et les caractéristiques communs : cibler les mailing

2-Médecine:

> Déterminer des segments de patients susceptibles d'être soumis à des protocoles thérapeutiques déterminés,

> Déterminer chaque segment regroupant tous les patients réagissant identiquement

> Retrouver les différents comportements similaires

3-Biologie – Zoologie – Ethologie – Sciences humaines:

> Expliquer les relations entre espèces, races, genres, familles,

> Retrouver de nouvelles répartitions

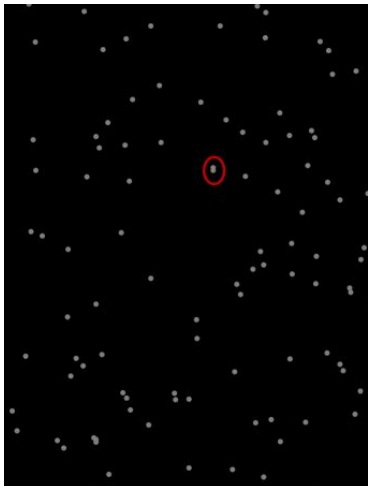
Profiling

- Analyse sémantique, sentimentale,
- Analyse et mesure de la tonalité d'un contenu textuel
- Catégorisation des concepts ou des entités nommés
- Construction d'agrégateur synthétique à partir des flux d'actualités

On retrouve des comportements similaires, dissimilaires, relation entre les groupes..

Objectifs de la segmentation: Le regroupement

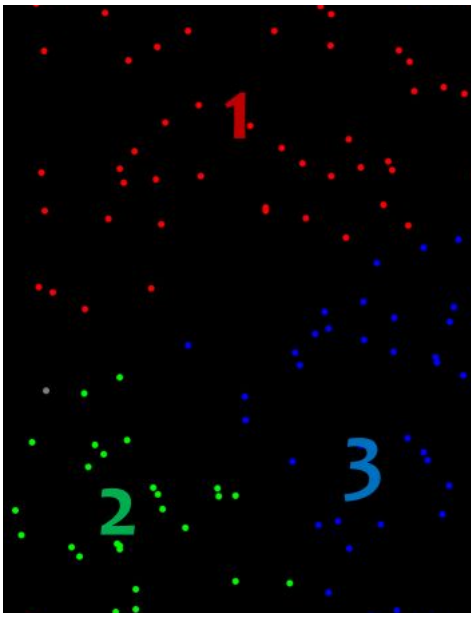
Problématique : Retrouver le nombre de clusters



Deux individus se ressemblent le plus SI

les points qui les représentent dans le nuage de points sont les plus proches

Nécessité d'une métrique de la distance /Euclidienne / de Mahalanobis / de Manhattan / de Ward ...



Critère d'évaluation d'une classification :
Inertie Totale

$$I_{tot} = I_{inter} + I_{intra}$$

Inertie intraclasse: somme des inerties totales de chaque classe

Inertie interclasse: moyenne (pondérée par la somme des poids de chaque classe) des carrés des distances des barycentres de chaque classe au barycentre global

Méthodologies:

Partitionnement	Hierarchique
K-MEANS	CAH

L'algorithme K-MEANS:

L'algorithme des K-moyennes est un algorithme qui permet de trouver des classes dans des données.

Les classes qu'il construit n'entretiennent jamais de relations hiérarchiques => une classe n'est jamais incluse dans une autre classe.

Il fonctionne en précisant le nombre de classes attendues.

Il calcule les distances Intra-Classe et Inter-Classe.

Il opère sur des variables continues car on calcule des distances.

POINTS FAIBLES DE K-MEANS

Le choix du nombre de groupes est subjectif dans le cas où le nombre de classes est inconnu au sein de l'échantillon.

L'algorithme du K-Means ne trouve pas nécessairement la configuration

optimale correspondant à la fonction objective minimale.

Les résultats de l'algorithme du K-Means sont sensibles à l'initialisation aléatoire des centres.

3-La ségmentation (clustering) > CAH

CAH > Classification Ascendante Hiérarchique

Principe algorithmique :

> Créer à chaque étape une partition obtenue en agrégeant 2 à 2 les éléments
les plus proches ! -- Eléments : individus ou groupe d'individus

> L'algorithme fournit une hiérarchie de partitions : arbre contenant
l'historique de la classification et permettant de retrouver n-1 partitions.

> Nécessité de se munir d'une métrique (distance euclidienne, chi2, Ward...)

> Nécessité de fixer une règle pour agréger un individu et un groupe d'individus (ou bien 2 groupes d'individus)

Le dendrogramme:

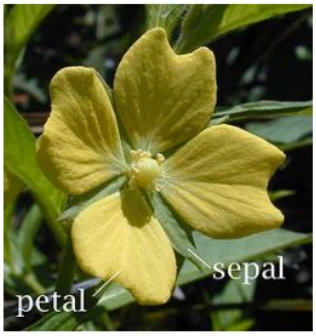
Durant les étapes d'un algorithme de classification hiérarchique, on est en train de construire
un dendrogramme.
Il indique les objets et classes qui ont été fusionnés à chaque itération et aussi la valeur du critère choisi pour chaque partition rencontrée.

1. **Il donne un résumé de la classification hiérarchique**
2. **Chaque palier correspond à une fusion de classes**
3. **Le niveau d'un palier donne une indication sur la qualité de la fusion correspondante**
4. **Toute coupure horizontale correspond à une partition**

4-Simulation KMEANS et CAH sur R

Base de donnée définie sur R : IRIS : Contient 3 types de fleurs + leurs caractéristiques

> iris



Data Mining-4GL © ESPRIT2018-2019

RGui - [R Console]

Fichier Edition Voir Misc Packages Fenêtres Aide

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1.0	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa

Variables :

Quantitatives (variables actives)

Qualitatives (variable décisionnelle)

> iris

> print(head(iris))

//afficher 6 premières lignes

> print(summary(iris))

//stat descriptives : min + max + valeur moyenne de chaque variable

```
147          6.3          2.5          5.0          1.9  virginica
148          6.5          3.0          5.2          2.0  virginica
149          6.2          3.4          5.4          2.3  virginica
150          5.9          3.0          5.1          1.8  virginica
> print(head(iris))
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1          3.5          1.4          0.2  setosa
2          4.9          3.0          1.4          0.2  setosa
3          4.7          3.2          1.3          0.2  setosa
4          4.6          3.1          1.5          0.2  setosa
5          5.0          3.6          1.4          0.2  setosa
6          5.4          3.9          1.7          0.4  setosa
> print(summary(iris))
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
```

Si on veut faire un segmentation on doit éliminer la variable décisionnelle

>iris_for_kmeans<-iris[,1:4] // affecter le dataset iris ayant toutes les lignes + colonnes de 1 à 4

```
> print(head(iris_for_kmeans))
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1          3.5          1.4          0.2
2          4.9          3.0          1.4          0.2
3          4.7          3.2          1.3          0.2
4          4.6          3.1          1.5          0.2
5          5.0          3.6          1.4          0.2
6          5.4          3.9          1.7          0.4
```

>iris_for_kmeans

>print(head(iris_for_kmeans))

On a préparé le dataset, maintenant on passe à l'application de l'algorithme kmeans

> **km <- kmeans(iris_for_kmeans, 3)** // k = 3

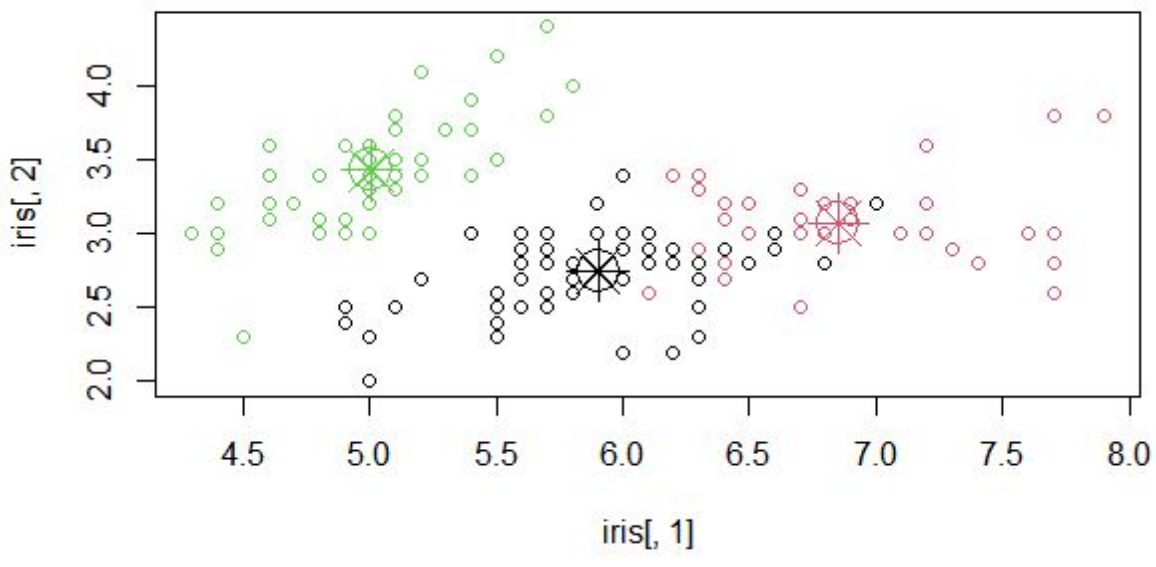
```
> km <- kmeans(iris_for_kmeans, 3)
> km
K-means clustering with 3 clusters of sizes 62, 38, 50

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.901613    2.748387     4.393548    1.433871
2     6.850000    3.073684     5.742105    2.071053
3     5.006000    3.428000     1.462000    0.246000

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [40] 3 3 3 3 3 3 3 3 3 3 3 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
 [79] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 1 1 2 2
[118] 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2 2 1

Within cluster sum of squares by cluster:
[1] 39.82097 23.87947 15.15100
 (between_SS / total_SS =  88.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
>
>
```

Représentation graphique : plot

>plot(iris[,1], iris[,3], col=km$cluster)

On peut rajouter les centres:

>points(km$centers[,c(1,)], col=1:3, pch=8, cex=2)



Comparaison du cluster kmeans avec le cluster réel:

```
> table(km$cluster, iris$Species)

    setosa versicolor virginica
1       0         48        14
2       0          2        36
3      50          0         0
```

	setosa	versicolor	virginica
Taux de classification	100%	96%	72%
% individus « mal classés »	0%	4%	28%
	10,67%		

Application de CAH sur IRIS:

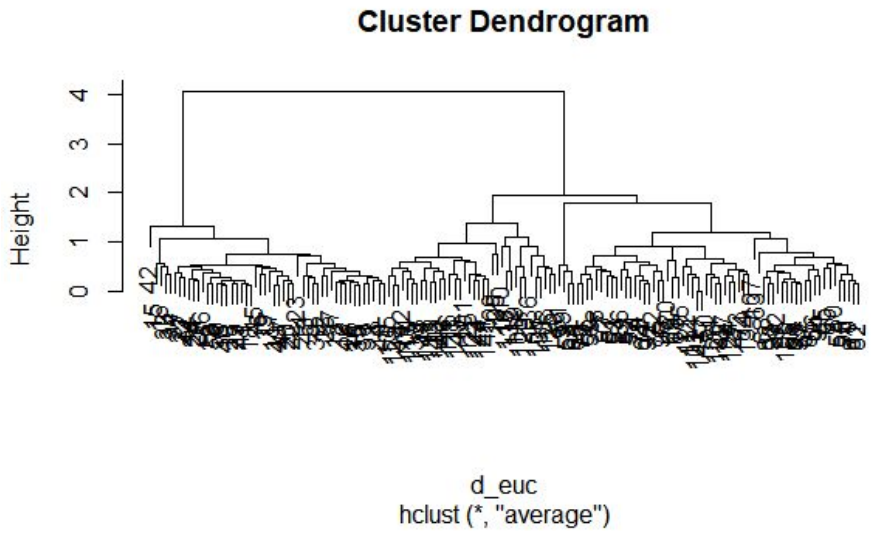
Application de la CAH sur la base IRIS en utilisant la distance euclidienne et les 4 variables de longueur et largeur des pétales et des sépales.

1. Calcul de la matrice des distances sur les colonnes de 1 à 4

> d_euc = dist(iris[,1:4], method = “euc”)

2. Application de la fonction hclust

> hc = hclust(d_euc,method ="ave")



> plot (hc)

//output : dendrogramme

3. Extraire – à partir du dendrogramme – la classification en 3 groupes :

> classe<-cutree(hc,3)

```
classe setosa versicolor virginica
     1     50          0         0
     2      0         50        14
     3      0          0        36
> |
```

	setosa	versicolor	virginica
Taux de classification	100%	100%	72%
% individus « mal classés »	0%	0%	28%
	9,33%		

AVANTAGES DE LA CAH

- Permet de classer : des individus, des variables, des moyennes de classes obtenues en sortie d'un algorithme des centres mobiles.
- si on classe des moyennes, on améliore les résultats si on connaît non seulement les moyennes des classes, mais aussi les inerties intraclasse et les effectifs des classes.
- S'adapte aux diverses formes de classes, par le choix de la distance.
- Permet de choisir le nombre de classes de façon optimale, grâce à des indicateurs de qualité de la classification en fonction du nombre de classes.

Inconvénients:

Complexité algorithmique non linéaire (en n^2 ou n^3, parfois n^2log(n))

Deux observations placées dans des classes différentes ne sont jamais comparées

OBJECTIFS DES TECHNIQUES DESCRIPTIVES

visent à mettre en évidence des **informations présentes** mais cachées par le volume des données

il n'y a **pas de variable « cible »** à prédire

projection du nuage de points sur un espace de dimension inférieure pour obtenir une visualisation de l'ensemble des liaisons entre : Individus, Variables… tout **en minimisant la perte d'information**

trouver dans l'espace de travail des **groupes homogènes** d'individus ou de variables

détection d'**associations** entre des objets

4-L'analyse ACP (Analyse en Composantes Principales) (séance6)

Analyse non supervisée > pas de variables à prédire / **Analyse descriptive**

Réduire les dimensions : le nb de variables > output 2dimensions

Chaque axe : écriture linéaire des variables

Données structurées sous forme matricielle.

Lignes : Observation / Individus Colonnes : Caractéristiques

Pourquoi on désire réduire les dimensions?

Dataset 10000lignes 10000colonnes, on va réduire le nombre de colonnes et ce en gardant les infos nécessaires.

Réduction:

-Application de la factorisation matricielle (transformation mathématique sur la matrice)

=> Application de l'ACP

Notre pb : pb à n dimension > output : espace 2D + minimisation de la perte d'informations + décomposition en valeurs propres > nouvel espace (combinaison linéaire de l'espace d'origine)

Taux d'infos : Fact1+Fact2

Déduction indiv/indiv : Carte des individus
var/var : Cercle de corrélation
indiv/var : superposition

Types de corrélations :

1) Corrélation positive : individus/variables ont un comportement similaire
   Angle < 90°
2) Corrélation négative : // // // // // dissimilaire
   Angle > 90°

3) Corrélation orthogonale : on ne peut pas faire une interprétation 90°

DES DONNÉES PAYS/PROTÉINES

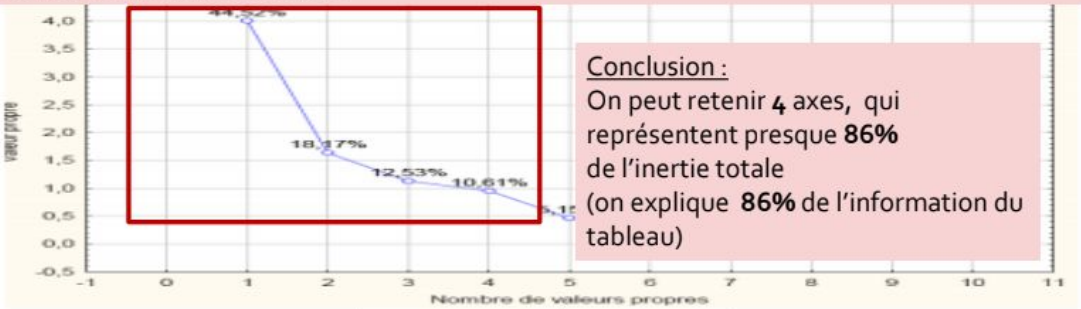
Pays	VR	VB	Oeufs	Lait	Poisson	Céréales	Starch	Noix	FL
Albanie	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Autriche	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
Belgique	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
Bulgarie	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Cheko.	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
Danemark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
Allemagne-E	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finlande	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Grèce	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hongrie	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
Irlande	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
Italie	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Pays-bas	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norvège	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
Pologne	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
Roumanie	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
Espagne	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
Suède	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
Suisse	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
Angleterre	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
Russie	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
Allemagne-O	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yougoslavie	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

Résultat de l'ACP :

CHOIX DU NOMBRE D'AXES A RETENIR

Valeur numéro	Val. Propres (matrice de corrél.) & stat. associées (proteines2) Variables actives seules Val Propre	% Total variance	Cumul Val Propre	Cumul %
1	4,006438	44,51597	4,006438	44,5160
2	1,634999	18,16666	5,641437	62,6826
3	1,127920	12,53244	6,769357	75,2151
4	0,954664	10,60738	7,724020	85,8224
5	0,463838	5,15376	8,187859	90,9762
6	0,325131	3,61257	8,512990	94,5888
7	0,271696	3,01785	8,784686	97,6066
8	0,116292	1,29213	8,900888	98,8988
9	0,099112	1,10124	9,000000	100,0000

Deux critères empiriques pour sélectionner le nombre d'axes :
- **Critère de Kaiser: on ne retient que les axes associés à des valeurs propre supérieures à 1**
- Critère du coude : sur l'évolution des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière. On sélectionne les axes avant le décrochement



Conclusion :
On peut retenir **4** axes, qui représentent presque **86%** de l'inertie totale
(on explique **86%** de l'information du tableau)

Les 9 lignes correspondent à une combinaison linéaire, où chaque ligne correspond à une composante principale.

Plans :

1+2 / 1+3 / 1+4 ..

On a choisi 4 axes selon le 2ème critère

(un critère à la fois)

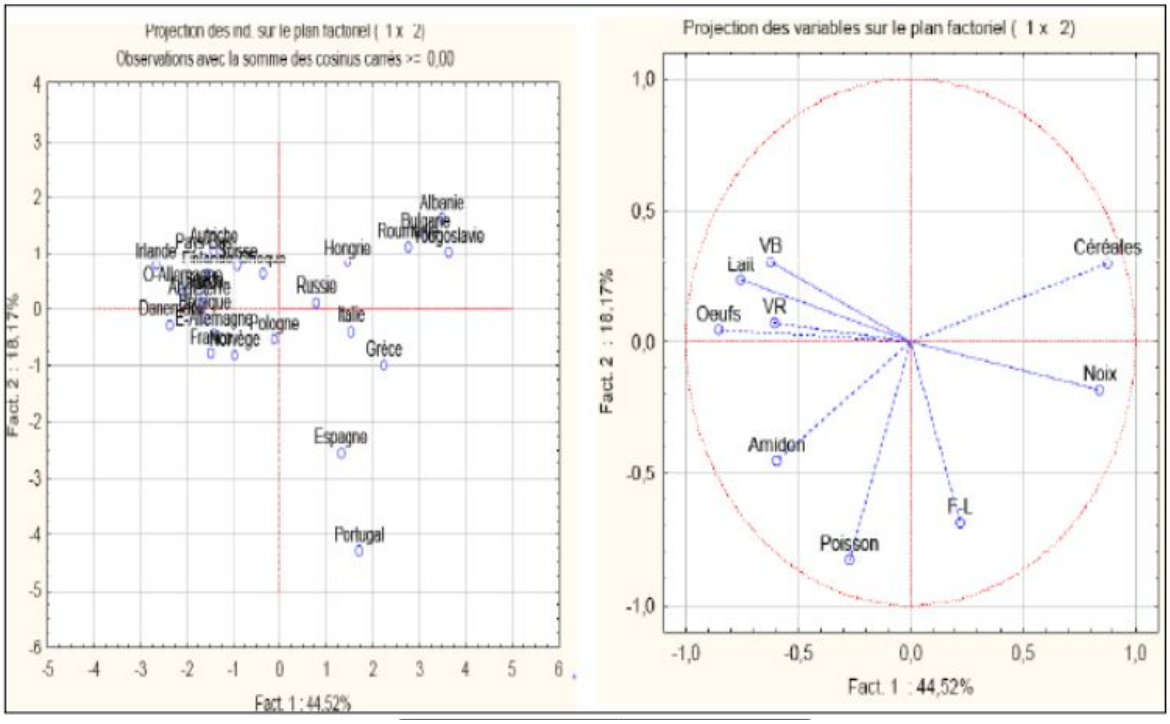
Projection du nuage de points:

Chaque nuage de points (variables et individus) est construit en projection sur les plans factoriels

Un plan factoriel est un repère du plan défini par deux des q axes factoriels retenus.

L'examen des plans factoriels permettra de visualiser les corrélations(relation) entre les variables et d'identifier les groupes d'individus ayant pris des valeurs proches sur certaines variables.

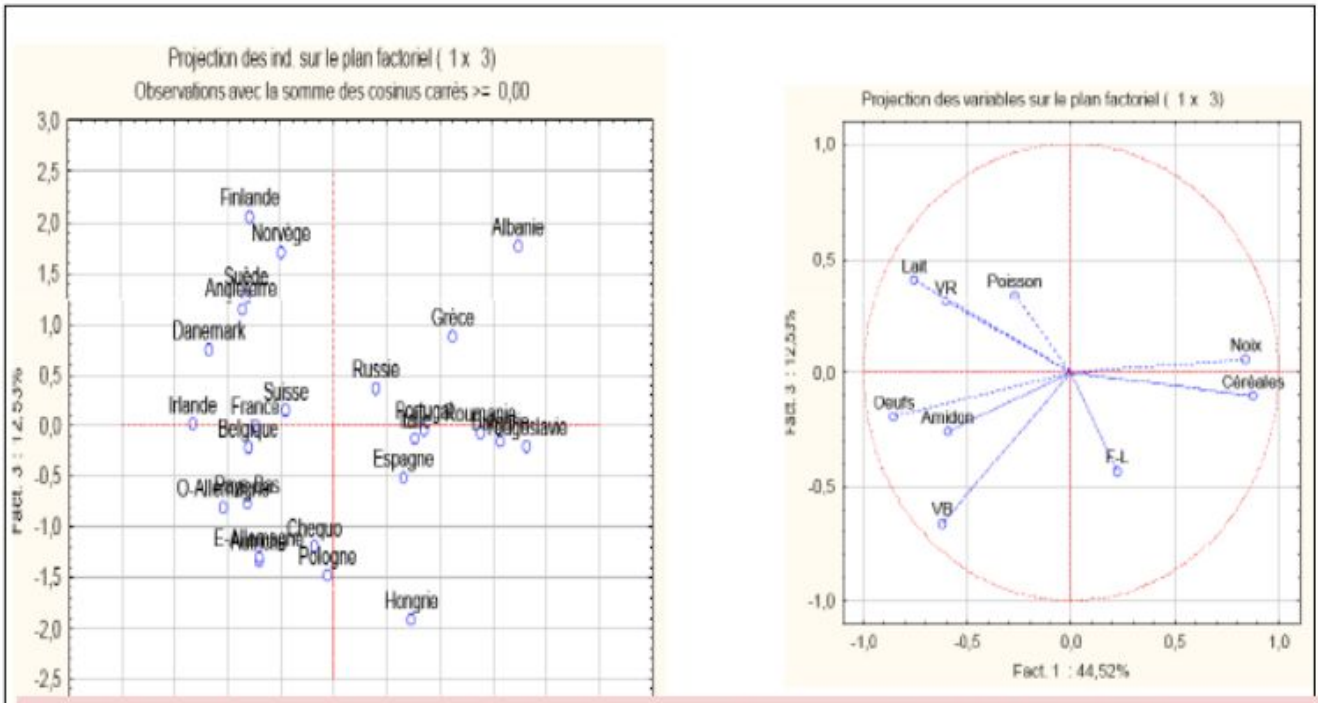
Plan factoriel 1x2: **Carte des individus** **Cercle de corrélation**



Italie-grèce:

Italie-Irlande:

Plan factoriel 1x3 : **Carte des individus** **Cercle de corrélation**



Plus les variables sont proches du bord du cercle et plus les variables sont bien représentées par le plan factoriel, c'est-à-dire que la variable est bien corrélée avec les deux facteurs constituant ce plan.

Etude de proximité entre les points :

1) Regarder les graphiques et analyser plus finement les proximités entre points.

2) Un point est dit bien représenté sur un axe ou un plan factoriel s'il est proche de sa projection sur l'axe ou le plan. S'il est éloigné, on dit qu'il est mal représenté.

Indicateur = angle formé entre le point et sa projection sur l'axe : au plus il est proche de 90 degrés, au moins le point est bien représenté

L'analyse se fera à l'aide des individus et variables contribuant le plus à l'axe : si une variable a une forte contribution positive à l'axe, les individus ayant une forte contribution positive à l'axe sont caractérisés par une valeur élevée de la variable.

COURS ACP ESPRIT

L’analyse en composantes principales

L’objectif de l’analyse en composantes principales (ou ACP) est purement descriptif : il s’agit « d’explorer » un ensemble d’observations rassemblées sous la forme d’un tableau de données indiquant pour chaque unité statistique les valeurs observées d’un certain nombre de variables quantitatives.

Interpréter les résultats :

1- Déterminer le nombre d’axes de l’analyse :

Pour répondre à cette question, il faut consulter le tableau des valeurs propres qui accompagne l’ACP. Les valeurs propres sont classées de façon décroissante. L’inertie de chaque axe et l’inertie cumulée figurent également dans ce tableau. Il y a deux manières pour déterminer le nombre d’axes à prendre en compte :

- Un critère absolu : ne retenir que les axes dont les valeurs propres sont supérieures à 1 (c'est le critère de Kaiser).
- Un critère relatif : retenir les valeurs propres qui “dominent” les autres, en se référant au graphique (critère du coude)

Il est important que les valeurs propres des axes retenus restituent une “bonne proportion” de l'analyse. Cela signifie que la somme de l'inertie expliquée par chacun des axes représente une partie importante de l'inertie totale. Cette somme est une mesure de la fiabilité de la lecture des mappings.

Les mapping de l’ACP sont les projections des variables et des individus sur un plan factoriel déterminé. On commencera par interpréter le premier plan factoriel (celui formé par les facteurs F1 et F2) car c’est celui qui concentre la plus grande partie de l’information du nuage. On ira voir ensuite et le cas échéant les autres plans factoriels. Sur un plan factoriel, on n’interprète que les variables et les individus qui sont bien représentés. Pour les individus, on utilisera les

contributions absolues et relatives alors que pour les variables, on n'interprètera que celles qui sont proches du cercle de corrélation.

La représentation des variables :

Ce mapping se distingue par la présence d'un cercle de corrélation. Sur un plan factoriel déterminé, on n'interprète que les variables qui sont bien représentées c'est à dire celles qui sont proches ou sur le cercle de corrélation.

On interprète deux types de positions :

Les positions des variables par rapport aux axes afin de déterminer quelles sont les variables qui « font les axes ». On va ainsi pouvoir nommer les axes en fonction des variables. Les positions des variables les unes par rapport aux autres.

Le coefficient de corrélation entre deux variables étant le cosinus de l'angle formé par les vecteurs on en déduit que :

- deux variables qui sont proches ou confondu (angle de 0°) sont corrélées positivement (coefficient de corrélation proche de 1),

- deux variables opposées (formant un angle de 180°) sont corrélées négativement (coefficient de corrélation proche de -1)

- deux variables positionnées à angle droit (angle de 90°) ne sont pas du tout corrélées (coefficient de corrélation égal à 0)

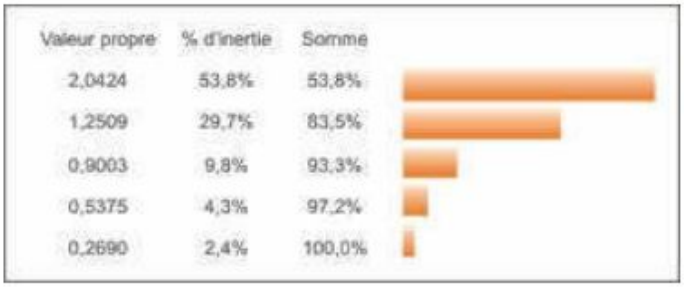
La représentation des individus :

Si l'ACP est réalisée sur un nombre d'individus faible, l'interprétation du nuage des individus est alors possible. Quel que soit le cas envisagé, on n'interprète, sur un plan factoriel déterminé, que les individus qui sont bien représentées. Pour cela, il faut aller voir leurs contributions absolues et relatives. Sous réserve d'une bonne représentation, la proximité de deux individus sur un plan factoriel est synonyme d'individus ayant un comportement similaire, c'est-à-dire ayant des réponses aux variables de l'analyse qui sont très proches. Si deux individus ont exactement les mêmes valeurs aux différentes variables, ils seront superposés sur les différents plans factoriels. De même, des individus ou des groupes d'individus s'opposant par rapport à un axe factoriel, s'opposeront par rapport aux variables qui « font » cet axe.

Exemple : l'entrepôt de données est constitué des notations moyennes des clients pour les différents magasins du groupe (Paris, Lyon, Marseille, Nice.)

Les variables sont : Choix proposés, Facilité pour trouver le produit, Disponibilité des vendeurs, Compétence des vendeurs, Courtoisie des vendeurs L'unité statistique est donc le magasin.

Les sorties de l'ACP :



Valeur propre	% d'inertie	Somme
2,0424	53,8%	53,8%
1,2509	29,7%	83,5%
0,9003	9,8%	93,3%
0,5375	4,3%	97,2%
0,2690	2,4%	100,0%

Sur le schéma précédent on remarque qu'en conservant les deux premiers axes on va expliquer 83,5% de l'inertie totale du nuage de point.

Quels sont les points qui nous intéressent ?

Les points les plus intéressants sont généralement ceux qui sont assez proches d'un des axes, et assez loin de l'origine. Ces points sont bien corrélés avec cet axe et sont les points explicatifs pour l'axe : Ce sont les points les plus "parlants" ; leur "vraie distance" de l'origine est bien représentée sur le plan factoriel1 . Dans le mapping ci-dessous, on voit clairement que Nice est extrêmement corrélé avec l'axe horizontal. De même, Paris et Reims notamment sont très bien corrélés à l'axe vertical.

La corrélation de chaque point sur un axe exprime la qualité de représentation du point sur l'axe. Elle prend des valeurs entre 0 (pas corrélé du tout) et 1 (fortement corrélé). Si cette valeur est proche de 1, alors le point est bien représenté sur l'axe.

Les points situés près du centre sont donc généralement mal représentés par le plan factoriel. Leur interprétation ne peut donc pas être effectuée avec confiance. Ainsi Nîmes et Strasbourg semble proche mais c'est peut-être le fruit d'une projection car ils sont peut-être opposés sur l'axe 3.

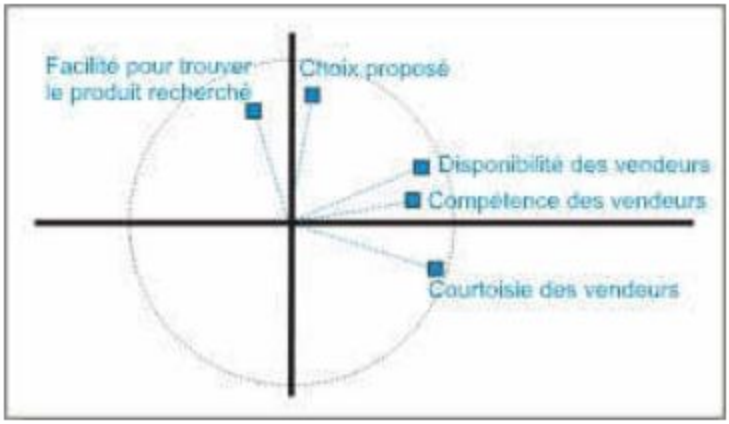
Comment interpréter les proximités ?

On s'intéresse donc essentiellement aux points bien représentés (i.e. situés loin du centre). Si deux points sont proches l'un de l'autre, il est probable que les réponses des individus qu'ils représentent soient très similaires. Il faut cependant se méfier : Il se peut que sur un axe ils soient très proches, alors que sur un autre ils seront très loin l'un de l'autre. Il faut donc les regarder par rapport à tous les axes qui ont été retenus pour l'analyse.

S'ils sont bien corrélés avec l'axe qui les montre proches, alors, on peut conclure qu'ils sont vraiment proches.

Est-ce qu'on peut donner un sens "réel" aux axes du mapping ?

Les axes factoriels sont des axes virtuels issus d'une synthèse entre les variables de l'analyse..



Dans notre exemple, nous pouvons constater que les points "disponibilité", "compétence" et "courtoisie" sont très proches du cercle de corrélation et donc très bien représentés sur le mapping. L'angle plutôt fermé (en partant de l'origine) que forment les points "compétence" et "disponibilité" indique que ces 2 variables sont assez bien corrélées entre elles.

En revanche, l'angle quasi droit formé par "compétence" et "choix" indique que ces deux variables sont indépendantes entre elles. Le fait que "compétence" soit proche de l'axe 1 indique qu'il est très bien représenté par cet axe. Comme il est très éloigné de l'axe 2, on peut conclure qu'il est peu représenté par cet axe. En ce qui concerne l'axe 2, le point "choix" est très bien corrélé avec l'axe. Le point "facilité" l'est également mais dans une moindre mesure.

De ces observations, nous pouvons conclure que l'axe 1 correspond plutôt à l'appréciation des vendeurs et notamment de leur compétence alors que l'axe 2 correspond plutôt à l'appréciation

du magasin et notamment du choix qu'il propose. Quelles autres conclusions tirer de notre analyse ?

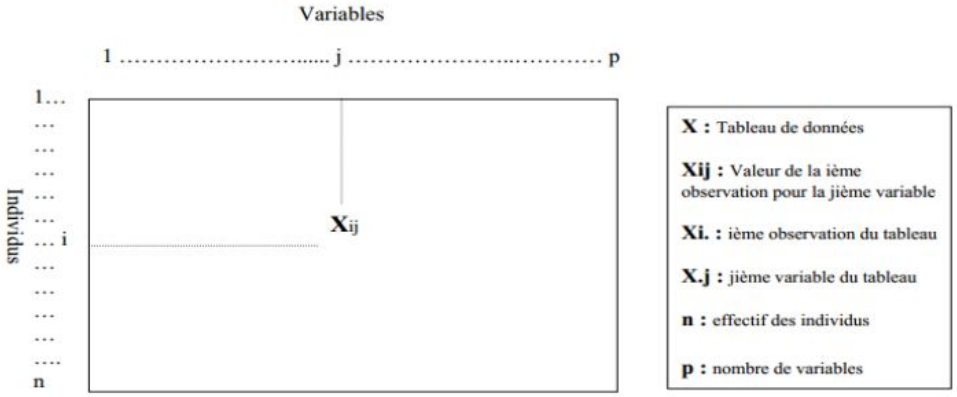
En synthétisant les informations issues des 5 variables analysées, notre mapping nous montre qu'il y a beaucoup d'efforts à faire en matière d'accueil et de renseignement des clients dans les magasins de Nice, Marseille, Amiens et Toulon. Ce dernier est également très peu apprécié en matière de choix. Les magasins de Paris, de Lyon et de Marseille sont appréciés de la clientèle pour le choix qu'ils proposent et la facilité pour trouver les produits recherchés. Lyon se distingue aussi par l'amabilité du personnel et peut être considéré comme le meilleur magasin parmi ceux qui ont fait l'objet de l'analyse. Ces conclusions sont confirmées par l'examen des tableaux de corrélations et de coordonnées des individus, fournis par le logiciel d'analyse.

FIN COURS ACP ESPRIT

MATRICE DE DONNÉES

On possède un tableau rectangulaire dont :

- ✓ **les colonnes sont des variables quantitatives**
  (mensurations, taux,...)
- ✓ **les lignes représentent des individus statistiques**
  (unités élémentaires telles que des êtres humains, des pays, des années...)



X : Tableau de données

Xij : Valeur de la ième observation pour la jième variable

Xi. : ième observation du tableau

X.j : jième variable du tableau

n : effectif des individus

p : nombre de variables

LES COMPOSANTES PRINCIPALES

Les composantes principales : permettent d'exprimer les variables initiales selon de nouveaux axes: les axes principaux, qui sont les vecteurs propres de la matrice :

- des covariances si on a des données hétérogènes, avec des ordres de grandeur différents

- des corrélations lorsque les unités de mesure ne sont pas les mêmes pour toutes les variables