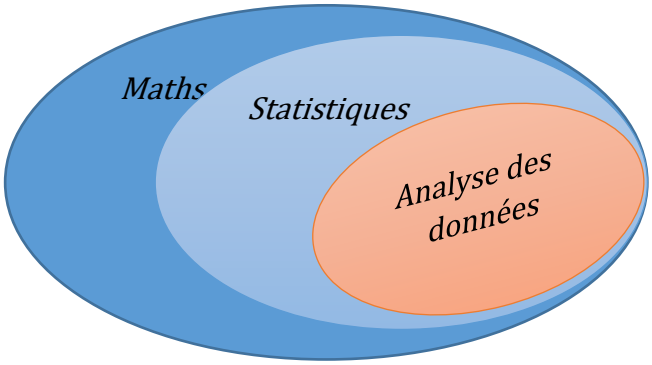
Analyse des données

Chapitre 1

1. Introduction Générale

L'analyse des données (**Data Analysis**) est une branche des mathématiques appliquées plus précisément de la statistique descriptive et multidimensionnelle.
Les différents types d'analyse de données se résument en :
Analyse descriptive, Analyse diagnostique, Analyse prédictive, Analyse prescriptive.



2. Types de données

Dans la plupart des applications, les données utilisées ne sont que les ***observations*** de n variables numériques (n assez grand) sur m individus (unités).
Les résultats obtenus c'est-à-dire la ***masse d'informations*** sont présentés en général dans une ***matrice (tableau)*** dont les ***coefficients représentent les données***.

3. Types de variables

Avant de présenter les différents types de variables, nous donnons une définition générale d'une variable.

3.1 Définition

Une ***variable*** represente une ***caractéristique***, une ***description***.
Deux types de variables :

- Variable **quantitative.**
- Variable **qualitative** ou bien catégorique.

4. Domaines d'application

Les domaines d'application des méthodes d'analyse de données sont nombreux et variés. Citons par exemple :

- L'industrie avec tous ses secteurs comme assurance, téléphonie, banque, contrôle de qualité, …..
- La recherche de documents

Analyse des données

- L'intelligence artificielle
- Traitement d'images et reconnaissance de formes
- Traitement du signal, biométrie,
- Les sciences humaines.

5. Objectifs

L'analyse de données consiste à traiter les données abondantes dans le but d'extraire les informations pertinentes. Ce traitement dans son sens large rassemble la représentation, l'analyse, la visualisation dans un meilleur espace et l'interprétation des données.
Pour ce faire, les chercheurs se sont fixés les objectifs principaux suivants :

- Comment ***regrouper*** ces données abondantes ?
- Comment ***visualiser*** les données dans le ***meilleur espace*** ?
- Comment ***interpréter graphiquement*** les résultats obtenus ?

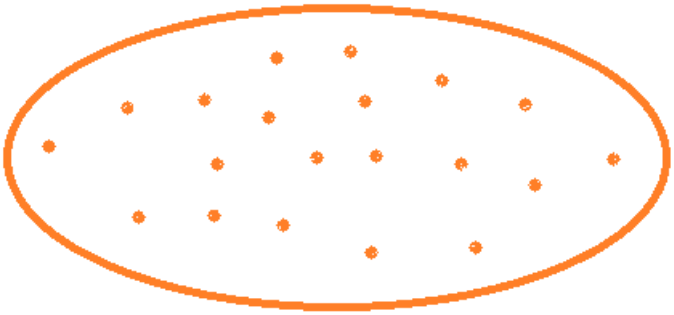
6. Types d'approches

Les différentes approches constituant l'analyse des données se classifient en deux grandes classes :

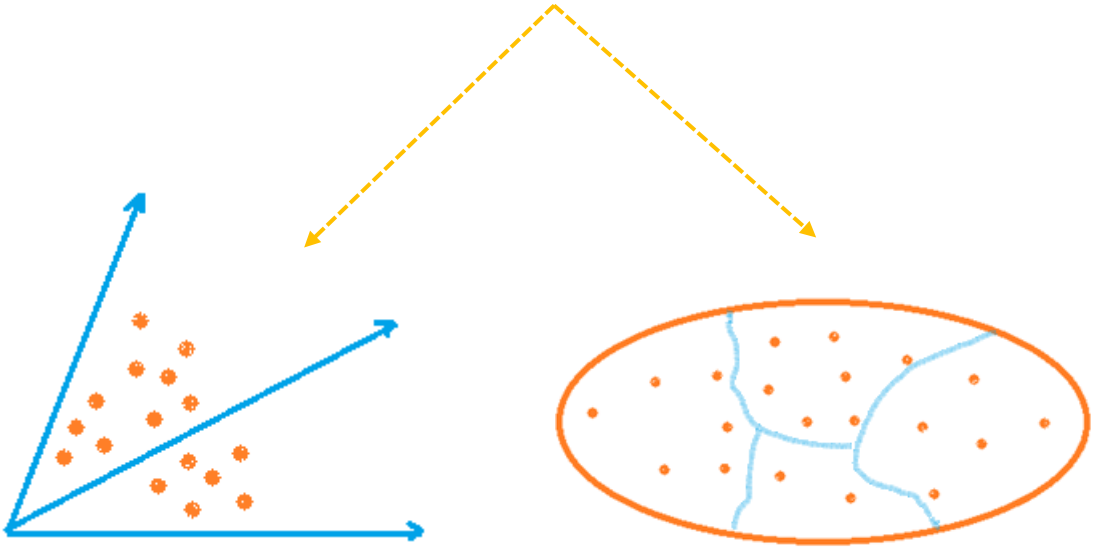
- Approches de classification.
- Approches factorielles.

Ensemble de feuilles de plantes.

Analyse des données



Espace des données



Représentation dans le meilleur espace ***Classification des données***

6.1 Approches de Classification

Les méthodes de cette classe visent à classifier automatiquement les données en classes (groupes ou clusters) homogènes selon leur degré de similarité ou bien dissimilarité. Ces méthodes à leurs tours se divisent en deux classes importantes :

- Classe des méthodes à apprentissage supervisé comme les arbres de décision, les réseaux de neurones, les chaines de Markov, les machines à vecteurs de supports (SVM), etc…..
- Classes des méthodes à apprentissage non supervisé, citons par exemple : les méthodes basées sur les graphes et hyperplan, la classification hiérarchique ascendante ou descendante, la méthode des centres mobiles, etc…

6.2 Approches Factorielles

Les méthodes de cette approche consistent à :

- ✓ ***Projeter*** les données sur un sous-espace de ***dimension inférieure*** à celle de l'espace donné initialement afin que nous puissions traiter les données et les
- ✓ ***Visualiser graphiquement*** les données pertinentes.

Dans cette classe, nous citons :

- Analyse en composantes principales (***ACP***). Méthode adaptée à des individus décrits par des variables quantitatives.
- Analyse factorielle des correspondance (***AFC***). Méthode adaptée à des individus décrits par des variables à modalités ou bien nominales.
- Analyse factorielle des correspondance multiples (***AFCM***) ou bien ***ACM***.
- Analyse factorielle discriminante (***AFD***). Méthode adaptée à des individus décrits par des variables quantitatives et appartenant à plusieurs classes.

7. Logiciels utilisés

Avec le développement de l'informatique, plusieurs logiciels de traitements des données ont été conçus et développés. Citons par exemple :

WinStat (CIRAD), SPSS (STATA), R, SAS, STATISTICA, MaxStat, SPAD,…

Remarque

Bien-sûr vu le nombre énorme des données, de nouvelles ***méthodes numériques*** (***algorithmes***) ont été développées pour traiter et visualiser cette masse de données. Toutes ces méthodes entrent dans le cadre ***sciences des données*** (***data sciences***).

Sciences des données = Statistiques + Sciences computationnelles

Analyse des données

Chapitre 2

Analyse en Composantes Principales

1. Introduction

L'analyse en composantes principales notée **ACP** est l'une des premières méthodes factorielles d'analyse multidimensionnelle.

2. Données

L'application de l'ACP se fait sur des données de type quantitatif. Ces données se présentent dans une matrice (tableau) X de taille (m, n) c'est-à-dire à m lignes et à n colonnes comme suit :

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^n \\ \vdots & \dots & \vdots & \dots & \vdots \\ x_i^1 & \dots & x_i^j & \dots & x_i^n \\ \vdots & \dots & \vdots & \dots & \vdots \\ x_m^1 & \dots & x_m^j & \dots & x_m^n \end{pmatrix}$$

m Individus

n Variables

n représente le nombre de variables observées sur les m individus.

Par conséquent, chaque x_i^j représente une **donnée** c'est-à-dire une **information.**

Ainsi :

✓ Les variables notées $X^1, X^2, \dots, X^n$ se présentent par des vecteurs de $\Re^m$.

C'est-à-dire, pour tout $j = 1, 2, \dots, n$.

$$\boldsymbol{X^j = (x_1^j,\ x_2^j,\ \dots\ ,\ x_m^j)^t \in \Re^m.}$$

Autrement dit, chaque variable est un point dans l'espace $\Re^m$.

✓ Les individus notés $X_1, X_2, \dots, X_m$ sont des vecteurs de $\Re^n$.

C'est-à-dire, pour tout $i = 1, 2, \dots, m$:

$$\boldsymbol{X_i = (x_i^1, x_i^2, \dots, x_i^n) \in \Re^n.}$$

De même, chaque individu est un point dans l'espace $\Re^n$.

Exemple

Considérons par exemples les **notes** obtenues par **5** étudiants dans **5** modules.

Ces informations peuvent être représentées dans la matrice des données suivante :

Analyse des données

Modules / *Etudiants*	**Maths**	**Algorithme**	**Base de Données**	**Génie Logiciel**	**Archi**
E_1	6	7	11	12	7.5
E_2	9	11	10	9	11
E_3	5	7	13	11	4
E_4	11	12	14	13	10
E_5	15.5	16	16	13	13

Dans cet exemple, la **donnée** (**l'information**) x_i^j est une **note**. C'est-à-dire :

Pour tout $i = 1, 2, \ldots, 5$ et $j = 1, 2, 3, 4, 5$,

x_i^j = la note obtenue par **l'étudiant** E_i dans **le module** j.

Analyse

A travers cet exemple, **nous souhaitons** :

- ✓ **Etudier la variabilité** observée sur les **individus** (étudiants) ou bien sur les **variables** (modules).
- ✓ **Réduire la taille** des données.
- ✓ **Représenter graphiquement** ou encore visualiser ces données.

2.1 Nuage des points

Il est clair que cette représentation matricielle des données dans une matrice (tableau) nous permet de lire la matrice de deux manières différentes selon ses lignes ou bien ses colonnes.

Ainsi, nous pouvons avoir :

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_m \end{pmatrix}.$$

Ou bien

$$X = (X^1 \quad \ldots \quad X^j \quad \ldots \quad X^n).$$

Ceci nous conduit à définir deux nuages de points :

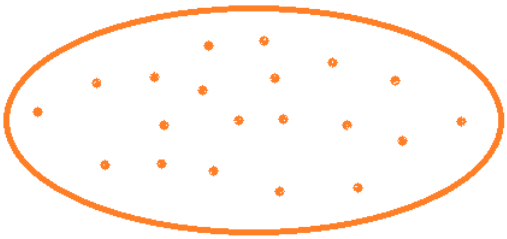
Analyse des données

- ***Nuage des individus***

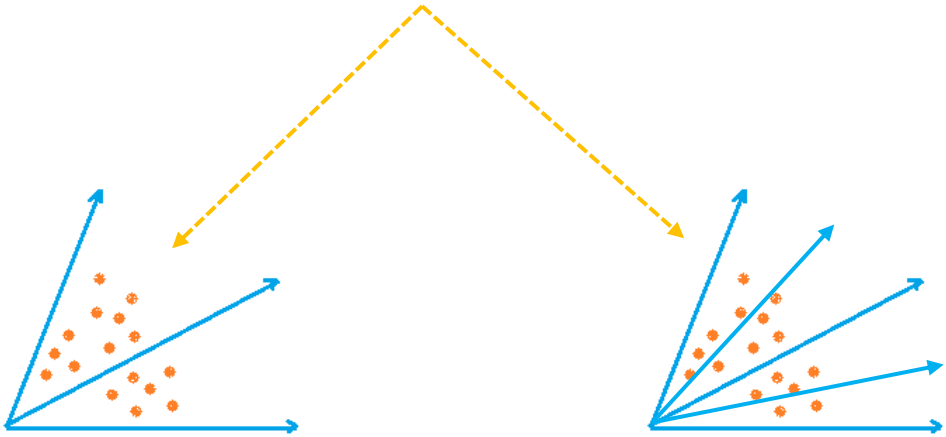
  C'est l'espace des points représentant les individus dont chaque axe represente une variable.

- ***Nuage des variables***

  C'est l'espace des points représentant les variables dont chaque axe represente un individu.



Espace des données



Du point de vue individus ***Du point de vue variables***

3. Objectifs

Les objectifs principaux de l'analyse en composantes principales sont :

- ***Décrire*** et ***Analyser*** la matrice des données selon ***ses lignes*** c'est-à-dire individus ou bien ***ses colonnes*** c'est-à-dire ses variables.

- ***Visualiser et Représenter*** graphiquement les données après avoir ***Réduit la dimension*** de l'espace initial c'est-à-dire le nombre de variables.

Pour ce faire,
Nous cherchons le ***meilleur sous-espace*** qui ***approche au mieux*** la matrice des données initiales.
C'est-à-dire :
Conserver au mieux **l'information pertinente** tout en ***minimisant la perte d'information.***