# Analyse de données

L. BELLANGER

Master 1 Ingénierie Statistique
Dpt de Mathématiques - Université de Nantes

## Plan





# 0. Introduction



## Un peu d'histoire …

**L'analyse des données multidimensionnelles** « à la française » s'est développée, à partir des années soixante-dix, sans faire référence au modèle probabiliste avec le développement de l'outil informatique.

Au XX ème siècle C. Spearman, H. Hotelling, R. Fisher, J.-P. Benzécri (1973), …

conception proprement géométrique.

On désigne sous cette appellation:
les différentes méthodes statistiques (descriptives)
d'un tableau de données
permettant de « voir » l'information
contenue dans de gros tableaux de données.

# Brève introduction

On présentera dans ce cours, 2 grands types de méthodes d'analyse de données, celles pour lesquelles

- **Le tableau $\mathbf{X}_{n\times p}$ n'a pas de structure**
  - C'est le domaine des **analyses factorielles simples** et des **méthodes de classification**.

  La traduction pratique de ces méthodes conduit à étudier les observations pour :
  - En **faire des cartes** qui définissent leurs positions relatives :

    ACP, AFC , AFCM.
  - **Les mettre dans des classes** :

    classification automatique (non supervisée) .
- **Le tableau $\mathbf{X}_{n\times p}$ a une structure**
  - C'est le domaine des **analyses canoniques** dans lesquelles :
    - Une ou plusieurs colonnes jouent des rôles différents : régression ou corrélation canonique, AFD.
    - Les lignes sont déjà regroupées en entités définies par l'appartenance a priori à des niveaux de facteurs : **discrimination**, classification automatique (supervisée).



# Brève introduction

- Les notions communes à toutes ces méthodes font intervenir des mesures de **dissemblance** et de **distance** ou de **ressemblance** et de **corrélation**.
- De surcroît, les lignes ou/et les colonnes peuvent être **actives** ou **passives** dans l'étude, participer à l'analyse ou simplement l'illustrer. Des objets nouveaux (lignes ou colonnes) peuvent être replacés dans l'analyse de base.



# Brève introduction … aux méthodes factorielles

- Méthodes destinées à fournir des représentations de l'information contenue ds des tableaux de données volumineux.
  - dites **multidimensionnelles** tq ACP et l'AFC, en opposition aux méthodes statistique descriptive ne traitant qu'1 ou 2 variables à la fois.
  - permettant la mise en relation de nbreuses variables par des représentations graphiques (**cartes**)
- Méthodes consistant à construire de **nouveaux caractères** (variables) **synthétiques** obtenus en combinant les variables initiales aux moyens de « facteurs ».
  - **Méthodes linéaire** car combinaison linéaire des variables initales.



# Brève introduction … aux méthodes factorielles

$$\mathbf{X} = \begin{bmatrix} x_1^1 & \cdots & x_1^j & \cdots & x_1^p \\ \vdots & \ddots & & \iddots & \vdots \\ x_i^1 & & x_i^j & & x_i^p \\ \vdots & \iddots & & \ddots & \vdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{bmatrix}$$

*individu $i$ : vecteur ligne dans $\mathbb{R}^p$*

*variable $j$ : vecteur colonne dans $\mathbb{R}^n$*

***2 nuages de points:***

- nuage des $n$ individus dans $\mathbb{R}^p$
- nuage des $p$ variables dans $\mathbb{R}^n$

# Brève introduction …
## aux méthodes factorielles

- **Principe**

1. **Calcul des distances entre lignes et entre colonnes**

   Ces proximités géométriques entre pts-lignes et pts-colonnes traduisent des associations statistiques soit entre les individus, soit entre les variables.

2. **Description des tableaux de distances associées à ces représentations géométriques** :

   méthodes factorielles ou classification automatique

**Objectif** : réduction de la dimension de l'espace dans lequel sont décrits les individus (resp. les variables ou les profils).



# Brève introduction …
## aux méthodes factorielles

- Elles reposent sur un **schéma de dualité:**

$$\begin{array}{ccc} \mathbb{R}^p & \xrightarrow{\mathbf{Q}} & \mathbb{R}^{p*} \\ \mathbf{X}^{\mathrm{T}} \uparrow & & \downarrow \mathbf{X} \\ \mathbb{R}^{n*} & \xleftarrow{\mathbf{D}} & \mathbb{R}^n \end{array}$$

  - Rigoureusement, $\mathbb{R}^{p*}$ est le dual de $\mathbb{R}^p$ (ensemble des applications linéaires de $\mathbb{R}^p$ dans $\mathbb{R}$), $\mathbb{R}^{n*}$ est le dual de $\mathbb{R}^n$ (ensemble des applications linéaires de $\mathbb{R}^n$ dans $\mathbb{R}$ ),
  - $\mathbf{Q}$ est vue comme la matrice d'une application linéaire définie par : $\left(\mathbf{Q}(\boldsymbol{e_i})\right)(\boldsymbol{e_{i'}}) = \langle \boldsymbol{e_i}, \boldsymbol{e_{i'}} \rangle_{\mathbf{Q}}$,

    **matrice produit scalaire** utilisée dans $\mathbb{R}^p$
  - $\mathbf{D}$ est vue comme la matrice d'une application linéaire définie par : $\left(\mathbf{D}(\boldsymbol{x^j})\right)(\boldsymbol{x^{j'}}) = \langle \boldsymbol{x^j}, \boldsymbol{x^{j'}} \rangle_{\mathbf{D}}$.

    **matrice produit scalaire** utilisée dans $\mathbb{R}^n$

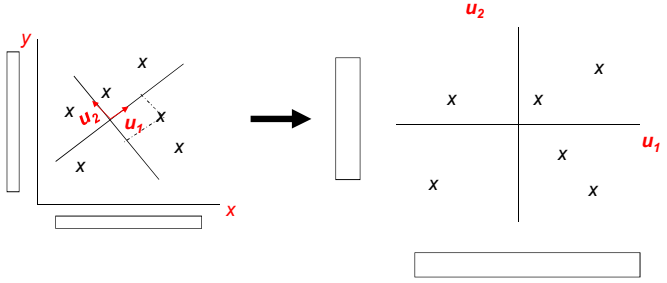


# Brève introduction …
## aux méthodes factorielles

- Techniques factorielles présentées dans la suite:
  - **Analyse en composantes principales (ACP):**

S'applique aux tableaux de type « variables-individus ».

**Objectif:** étudier la liaison entre $p$ variables et représenter les individus à l'aide d'un nb restreint de variables synthétiques.






# Brève introduction …
## aux méthodes factorielles

- Techniques factorielles présentées dans la suite:
  - **Analyse factorielle des correspondances (AFC) :**

S'applique aux tableaux de contingence ou logiques ou de mesures.

**Objectif:** étudier la liaison entre 2 variables qualitatives.

Fournir des représentations des associations entre lignes et colonnes.

Fondée sur la distance entre profils appelée *distance du Chi-deux*.

  - **Analyse des Correspondances Multiples (ACM)**

Extension du domaine d'application de l'AFC avec des procédures de calcul et des règles d'interprétations spécifiques.

lignes = individus,

colonnes = modalités de variables nominales (ex : catégories socio-professionellles)

# Brève introduction …
## aux méthodes factorielles

- Techniques factorielles présentées dans la suite:
  - **Analyse factorielle discriminante (AFD) :**

**Objectif:** construire des variables discriminantes permettant de différencier de façon optimale des groupes d'individus connus à priori.

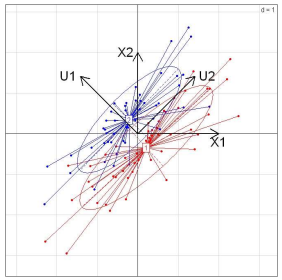



*Figure 1*

- *Si on projette les 100 observations constitués de 2 classes (rouge et bleu) sur les axes ($X_1$ et $X_2$) :*
  *=> il y a une superposition des pts des 2 classes*
  *=> on ne peut pas différencier les 2 classes.*
- *Par contre, si on projette sur $U_1$ et $U_2$, les obs de chaque classe sont beaucoup mieux séparées :*
  *$U_1$ $U_2$ constitue le meilleur repère pour séparer les 2 classes.*



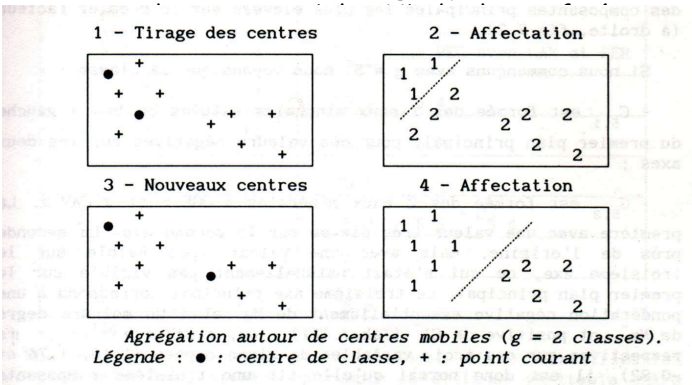
# Brève introduction …
## aux méthodes de classification

- **Objectif**:
  - regrouper les individus d'une population en un nombre limité de classes qui :
    - ne sont pas prédéfinies mais déterminées au cours de l'opération, contrairement aux classes du classement ;
    - regroupent les individus ayant des caractéristiques similaires et séparent les individus ayant des caractéristiques différentes.
- **2 grands types** :
  - Classification **par partition**
  - Classification **hiérarchique**



# Brève introduction …
## aux méthodes de classification

- Techniques classification présentées dans la suite:
  - **Classification par partition :** Regroupement d'observations

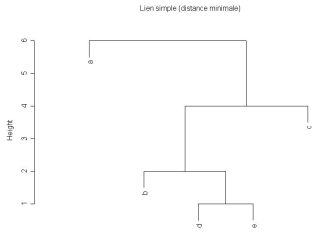



*Agrégation autour de centres mobiles (g = 2 classes).*
*Légende : ● : centre de classe, + : point courant*

# Brève introduction …
## aux méthodes de classification

- Techniques classification présentées dans la suite:
  - **Classification hiérarchique :**

conduisent à une chaîne de partitions emboîtées que l'on peut représenter par un arbre de classification appelé **dendrogramme**.



# Références


- L. Bellanger, R. Tomassone, *Exploration de données et méthodes statistiques : Data analysis & Data mining avec R. Collection Références Sciences, Editions* Ellipses, Paris, 2014.
- J.-M. Bouroche & G. Saporta, *L'analyse des données*. Presses Universitaires de France : Que sais-je ? 85, Paris, 1992.
- L. Lebart, A. Morineau, M. Piron, *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 2006.
- J.-P. Nakache, J. Confais, *Approche pragmatique de la Classification*. Editions Technip, Paris, 2005.
- G. Saporta, *Probabilités, Analyse des données*. Editions Technip, Paris, 2006.


17