USTHB - Faculté d'Informatique

Département *d'IA et SD*
M1 : MIV
Enseignant : N. Laiche

Analyse des données

Chapitre 0

1- Introduction Générale

L'analyse de données est un ensemble d'approches de type descriptif et multidimensionnel. Elle consiste à représenter, analyser, visualiser et à interpréter les données. Ces dernières sont souvent de taille énorme.

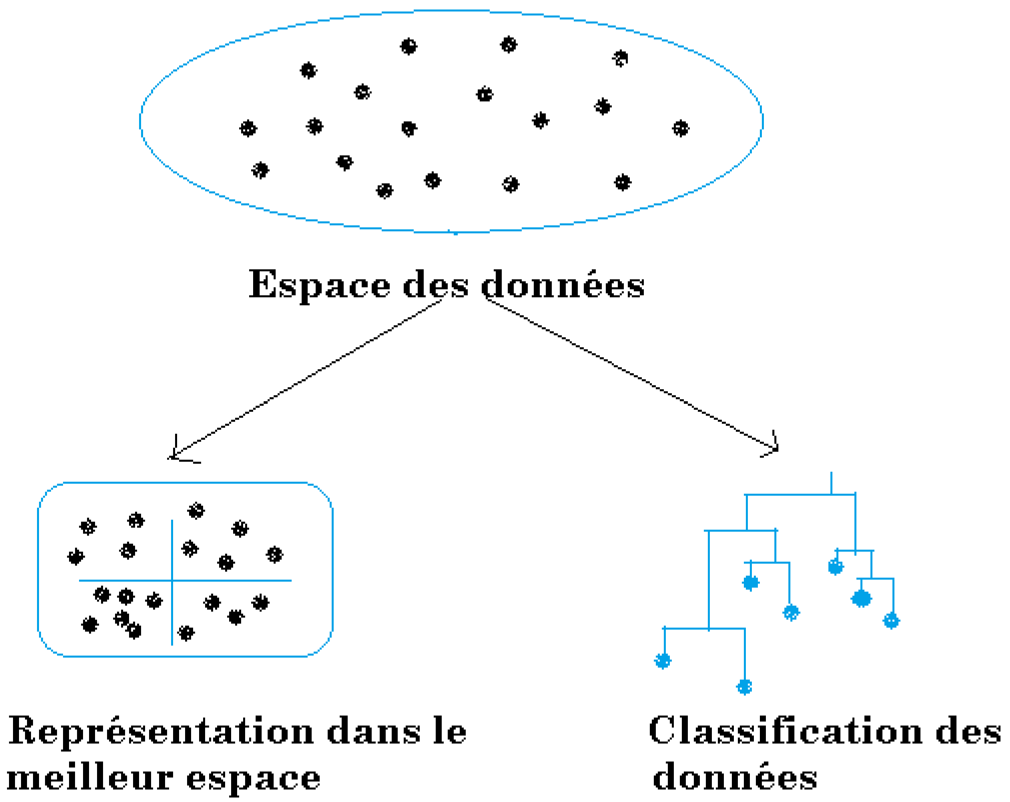
2- Types de Données

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 présentées en général dans une matrice dont les coefficients représentent les données.

3- Types d'Approches

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

1- Approches de classification.
2- Approches factorielles.



3.1 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 :

1- 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), ..ect.

2- Classe 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.

3.2 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. Dans cette classe, nous citons :

1- Analyse en composantes principales (ACP).
2- Analyse factorielle des correspondances (AFC, AFCM).

Chapitre 1

Analyse en Composantes Principales

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

1. **Données à analyser**

   L'ACP s'applique sur des données de type quantitatif. Ces données se présentent dans une matrice X de taille (m,n). n représente le nombre de variables observées sur les m individus.

$$X = \begin{pmatrix} x_1^1 & \cdots & x_1^j & \cdots & x_1^n \\ \vdots & & \vdots & & \vdots \\ x_i^1 & \cdots & x_i^j & \cdots & x_i^n \\ \vdots & & \vdots & & \vdots \\ x_m^1 & \cdots & x_m^j & \cdots & x_m^n \end{pmatrix} \updownarrow m \text{ individus}$$

$$\underleftrightarrow{\quad n \text{ variables} \quad}$$

Ainsi,

les variables notées $X^1, X^2, \ldots, X^n$ se présentent par des vecteurs de $\Re^m$ c'est à dire

$$X^j = {}^t(x_1^j, x_2^j, \ldots, x_m^j) \in \Re^m \quad \text{pour} \quad j = 1, 2, \ldots, n \text{, et}$$

les individus notés $X_1, X_2, \ldots, X_m$ sont des vecteurs de $\Re^n$ c'est-à-dire

$$X_i = (x_i^1, x_i^2, \ldots, x_i^n) \in \Re^n \quad \text{pour} \quad i = 1, 2, \ldots, m.$$

Exemple Considérons la matrice des notes attribuées à 5 étudiants dans 4 modules.

$X =$

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

Chaque x_i^j représente une donnée c'est-à-dire une information. Dans notre cas, c'est une note.

x_i^j = la note obtenue par l'étudiant E_i dans le module j pour $i = 1, 2, \ldots, 5$ et $j = 1, 2, 3, 4$.

But

- Il s'agit d'étudier la variabilité observée sur les individus ou bien sur les variables.
- De réduire la taille des données.

- La représentation matricielle des données dans une matrice nous permet de définir deux nuages de points selon les lignes de la matrice ou bien ses colonnes :

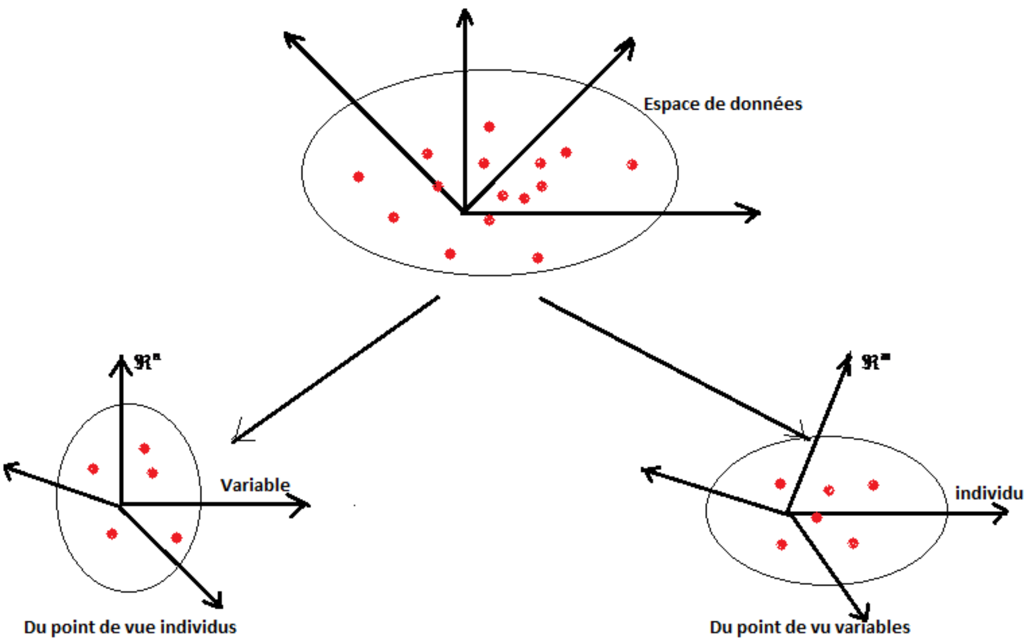


Figure 1. Nuages des données.

i) Nuage des individus C'est l'espace des points représentant les individus et chaque axe représente une variable.

ii) Nuage des variables C'est l'espace des points représentant les variables. Chaque axe représente un individu.

2. Objectifs

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

1- ***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. Ceci dit bien-sur après avoir réduit le nuage des points.

2- ***Réduire la dimension*** de l'espace initial dans le but de ***visualiser*** et de ***représenter*** graphiquement les données. Pour ce faire, nous cherchons le ***meilleur sous-espace*** qui ***approche au mieux*** la matrice des données initiales tout en ***minimisant la perte d'informations.***

D(X, Y) = a*d1(X,Y)+ b*d2(X,Y) : distance pondérée.

Remarque

A chaque individu X_i un poids positif w_i lui est affecté tel que : $\sum_{i=1}^{m} w_i = 1$.

Ce poids reflète l'importance de l'individu X_i par rapport aux autres individus.

Dans la pratique, le même poids est associé à tous les individus, il est égal à $\mathbf{1/m}$.

Ce qui nous conduit a définir la matrice des poids définie comme suit :

$D = (d_{ii})_i$ avec $d_{ii} = w_i$ pour tout $i = 1, 2, \ldots, m$.

$$D = \begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_m \end{pmatrix} = \begin{pmatrix} 1/m & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/m \end{pmatrix}.$$

3. Caractéristiques Numériques

Dans ce qui suit, nous donnons citons quelques caractéristiques de base liées à l'étude de l'analyse en composantes principales :

1- *Moyenne arithmétique*

La moyenne arithmétique $\overline{X^k}$ calculée sur une variable X^k est donnée par :

$$\frac{1}{m} \cdot \sum_{j=1}^{m} x_j^k \quad \text{ou bien} \quad \sum_{j=1}^{m} w_j \cdot x_j^k \quad \text{d'une manière générale.}$$

2- *Centre de gravité*

Le centre de gravité du nuage de points est défini par le vecteur :

$$g = \left(\overline{X^1}, \overline{X^2}, \ldots, \overline{X^j}, \ldots, \overline{X^n}\right) \in \Re^n.$$

Où $\overline{X^j}$ est la moyenne arithmétique de la variable X^j pour $j = 1, 2, \ldots, n$.

3- *Variance*

La variance d'une variable X^k est définie par :

$$Var(X^k) = \frac{1}{m} \cdot \sum_{j=1}^{m} \left(x_j^k - \overline{X^k}\right)^2.$$

Ou bien d'une manière plus générale :

$$Var(X^k) = \sum_{j=1}^{m} w_j \cdot \left(x_j^k - \overline{X^k}\right)^2.$$

4- *Ecart-type*

L'écart-type d'une variable X^k n'est que la racine carrée de sa variance :

$$\sigma(X^k) = S_k = \sqrt{Var(X^k)}.$$

5- *Inertie*

L'inertie d'un nuage de points est la quantité d'informations contenue dans la matrice de données. Elle représente la dispersion. Elle est donnée par la somme pondérée des variances des n variables :

$$I = \sum_{j=1}^{n} w_j \cdot Var(X^j).$$

4. Mesures de liaison

Avant de représenter les mesures de liaison qui nous permettent d'étudier les relations entre les variables, nous donnons quelques rappels.

4.1 Rappels

i) Produit scalaire

Soient u et v deux vecteurs quelconques de $\Re^n$, alors leur produit scalaire noté $\langle u,\, v\rangle$ ou bien $u \cdot v$ est défini par :

$$\langle u,\, v\rangle = \|u\| \cdot \|v\| \cdot \cos(u,\, v) = \langle v,\, u\rangle .$$

Le produit scalaire de u et v peut être défini directement en fonction des coordonnées des vecteurs en question par la relation suivante :

$$\langle u,\, v\rangle = {}^t u \cdot v = u \cdot {}^t v = \sum_{i=1}^{n} u_i \cdot v_i .$$

Remarques

- Si : les deux vecteurs sont orthogonaux, alors leur produit scalaire est nul.
- Si : $\|u\| = \|v\| = 1$, alors : $\langle u,\, v\rangle = \cos(u,\, v)$.
- Généralement, le produit scalaire se défini relativement à une matrice comme suit :

$$\langle u,\, v\rangle_M = {}^t u \cdot M \cdot v .$$

ii) Dérivation matricielle

- ***Dérivée d'une matrice par rapport à une variable***

Soit A une matrice de taille $(n,\, p)$ c'est-à-dire $A = (a_{i\,j})_{i,\,j}$ pour $i = 1, 2, \ldots, n$ et $j = 1, 2, \ldots, p$, alors : la dérivé de A par rapport à une variable t est définie par :

$$\frac{d\,A}{d\,t} = \frac{d\,(a_{i\,j})}{d\,t} \text{ pour tout } i,\, j .$$

En posant : $\frac{d(a_{ij})}{dt} = b_{ij}$ pour tout i, j , alors : $\frac{d\,A}{d\,t} = B$.

Exemple

Soit A une matrice carrée de dimension 3 :

$$A = \begin{pmatrix} t & 1 & -1/t \\ 0 & -2t & t \\ 1 & t^3 & -t \end{pmatrix}.$$

Alors la dérivée de A par rapport à t est donnée par :

$$\frac{d\,A}{d\,t} = \begin{pmatrix} dt/dt & d1/dt & d(-1/t)/dt \\ d0/dt & d(-2t)/dt & dt/dt \\ d1/dt & dt^3/dt & d(-t)/dt \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1/t^2 \\ 0 & -2 & 1 \\ 0 & 3t^2 & -1 \end{pmatrix}.$$

- ***Dérivée d'une fonction par rapport à une matrice***

Soit f une fonction des éléments de la matrice $A = (a_{i\,j})_{i,\,j}$ pour $i = 1, 2, \ldots, n$ et $j = 1, 2, \ldots, p$, alors :

$$\frac{\partial\, f}{\partial\, A} = \frac{\partial\, f}{\partial\,(a_{i\,j})} \quad \text{pour tout } i = 1, 2, \ldots, n \text{ et } j = 1, 2, \ldots, p.$$

Exemples

1) Soient X et Y deux vecteurs de $\Re^n$, et f une fonction réelle définie par :

$$f\,(X, Y) = \langle X,\, Y \rangle = {}^t X \cdot Y \text{, alors :}$$

$$f\,(X, Y) = \sum_{i=1}^{n} x_i \cdot y_i\,.$$

Par suite,

$$\frac{\partial\, f}{\partial\, X} = \frac{\partial\, f}{\partial\, x_i} = \frac{\partial\left(\sum_{i=1}^{n} x_i \cdot y_i\right)}{\partial\, x_i} \quad \text{pour : } i = 1, 2, \ldots, n\,.$$

Ainsi,

$$\frac{\partial\, f}{\partial\, X} = \begin{pmatrix} \dfrac{\partial\left(\sum_{i=1}^{n} x_i \cdot y_i\right)}{\partial\, x_1} \\ \dfrac{\partial\left(\sum_{i=1}^{n} x_i \cdot y_i\right)}{\partial\, x_2} \\ \vdots \\ \dfrac{\partial\left(\sum_{i=1}^{n} x_i \cdot y_i\right)}{\partial\, x_n} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial\,(x_1 \cdot y_1 + x_2 \cdot y_2 + \ldots + x_n \cdot y_n)}{\partial\, x_1} \\ \dfrac{\partial\,(x_1 \cdot y_1 + x_2 \cdot y_2 + \ldots + x_n \cdot y_n)}{\partial\, x_2} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \dfrac{\partial\,(x_1 y_1 + x_2 \cdot y_2 + \ldots + x_n \cdot y_n)}{\partial\, x_n} \end{pmatrix}.$$

Ce qui donne :

$$\frac{\partial f}{\partial X} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = Y \text{.}$$

De même, nous pouvons définir la dérivée de f par rapport au vecteur Y :

$$\frac{\partial f}{\partial Y} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = X \text{.}$$

2) Soit f une fonction factorielle définie par : $f(X) = Y$ avec X et Y deux vecteurs de $\Re^n$, alors :

$$\frac{\partial f}{\partial X} = \frac{\partial Y}{\partial X} = \frac{\partial Y}{\partial x_i} \quad \text{pour tout } i = 1, 2, \ldots, n \text{ .}$$

Or pour tout i fixé,

$$\frac{\partial Y}{\partial x_i} = \frac{\partial y_j}{\partial x_i} \quad \text{pour tout } j = 1, 2, \ldots, n \text{ .}$$

Alors :
la dérivée de f par rapport au vecteur X est donnée par :

$$\frac{\partial f}{\partial X} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}.$$

3) $\dfrac{\partial\left({}^tX \cdot M \cdot Y\right)}{\partial M} = X \cdot {}^tY$ **.**

4) $\dfrac{\partial\left({}^tX \cdot M \cdot X\right)}{\partial X} = M \cdot X + {}^tM \cdot X = (M + {}^tM) \cdot X$ **.**

Si : M est une matrice symétrique, alors :

$$\frac{\partial\left({}^{t}X \cdot M \cdot X\right)}{\partial X} = 2M \cdot X \,.$$

5) Soit A une matrice carrée d'ordre n : $A = (a_{i\,j})_{i,\,j}$ pour $i,\,j = 1, 2, \ldots, n$, alors :

$$trace(A) = \sum_{i=1}^{n} a_{i\,i} \,.$$

En posant : $f\,(A) = trace(A)$, alors :

$$\frac{\partial f}{\partial A} = \frac{\partial f}{\partial (a_{i\,j})} = \frac{\partial\left(\sum_{i=1}^{n} a_{i\,i}\right)}{\partial (a_{i\,j})} \quad \text{pour } i,\,j = 1, 2, \ldots, n \,.$$

Or,

$$\frac{\partial\left(\sum_{i=1}^{n} a_{i\,i}\right)}{\partial (a_{i\,j})} = \begin{cases} 1 & si : i = j \\ 0 & si : i \neq j \end{cases} \quad \textbf{pour tout } i,\,j = 1, 2, \ldots, n \,.$$

D'où,

$$\frac{\partial f}{\partial A} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} = Id \,.$$

4.2 Mesures de liaison

i) Pour mesurer la proximité entre les individus, nous utilisons en général la distance Euclidienne comme suit,

$$d\left(X_i,\, X_j\right) = \left(\sum_{k=1}^{n} (x_i^k - x_j^k)^2\right)^{1/2} .$$

ii) Pour juger de la liaison entre les variables c'est-à-dire pour mesurer la dispersion du nuage de points, nous appliquons les deux mesures suivantes :

- ***Covariance***
- Elle permet de préciser le sens de la liaison entre les variables, elle est définie par :

$$\mathrm{cov}(X,Y) = \frac{1}{m} \cdot \sum_{j=1}^{m} \left(x_j - \overline{X}\right) \cdot \left(y_j - \overline{Y}\right),$$

pour X et Y deux variables de $\Re^m$.

Remarque

La covariance n'est que $\left\langle \tilde{X},\, \tilde{Y} \right\rangle_M$ pour $\tilde{X}$ et $\tilde{Y}$ les deux variables centrées de X et Y respectivement et M est la matrice des poids.

- ***Coefficient de corrélation***

  Ce n'est qu'une normalisation de la covariance. Il permet de mesurer l'intensité de la liaison linéaire entre les variables, il est défini par :

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)} \cdot \sqrt{\operatorname{var}(Y)}} .$$

Remarque

Le coefficient de corrélation prend ses valeurs entre -1 et 1. Plus précisément, le coefficient de corrélation mesure le cosinus de l'angle inscrit entre les variables. En effet,

$$\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)} \cdot \sqrt{\operatorname{var}(Y)}} = \frac{\left\langle \tilde{X}, \tilde{Y} \right\rangle_M}{\left\| \tilde{X} \right\|_M \cdot \left\| \tilde{Y} \right\|_M} = COS\,(\theta_{\tilde{X},\tilde{Y}}) .$$

D'où,

$$\rho_{X,Y} = Cos\,(X,Y) .$$

Définition

Deux variables sont dites décorrélées si leur coefficient est nul, et elles sont linéairement liées si leur coefficient de corrélation est de module égal à 1.

Exemple