Calcul Scientifique et Apprentissage Automatique

M1 IAFA-SECIL
Université Paul Sabatier

Contacts :
Sandrine.Mouysset@irit.fr
Thomas.Pellegrini@irit.fr

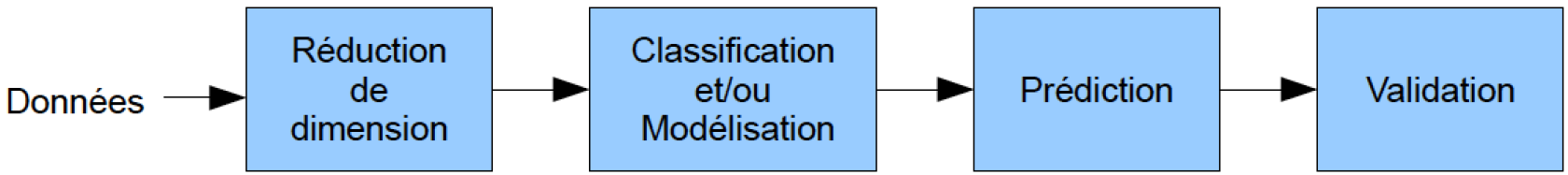
Plan de l'UE

- Prétraitement pour l'analyse de données / algèbre linéaire
- Apprentissage supervisé et non supervisé
- Introduction à l'optimisation : sans contrainte, avec contraintes d'égalité et moindres carrés
- Introduction aux réseaux de neurones
- Limites de l'Apprentissage Automatique
- Master Class

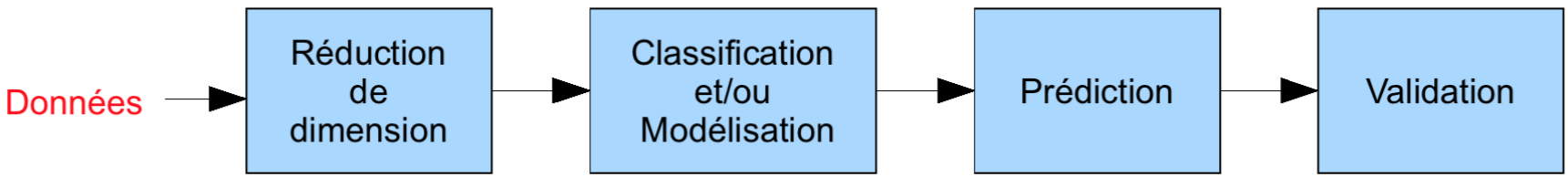
Nombre de séances :

- 9 cours + 3h (masterclass)
- 9 TD
- 9 TP

L'expression Analyse de données recouvre les techniques ayant pour objectif la description statistique des grands tableaux $n \times p$ de données



Chaîne d'analyse des données



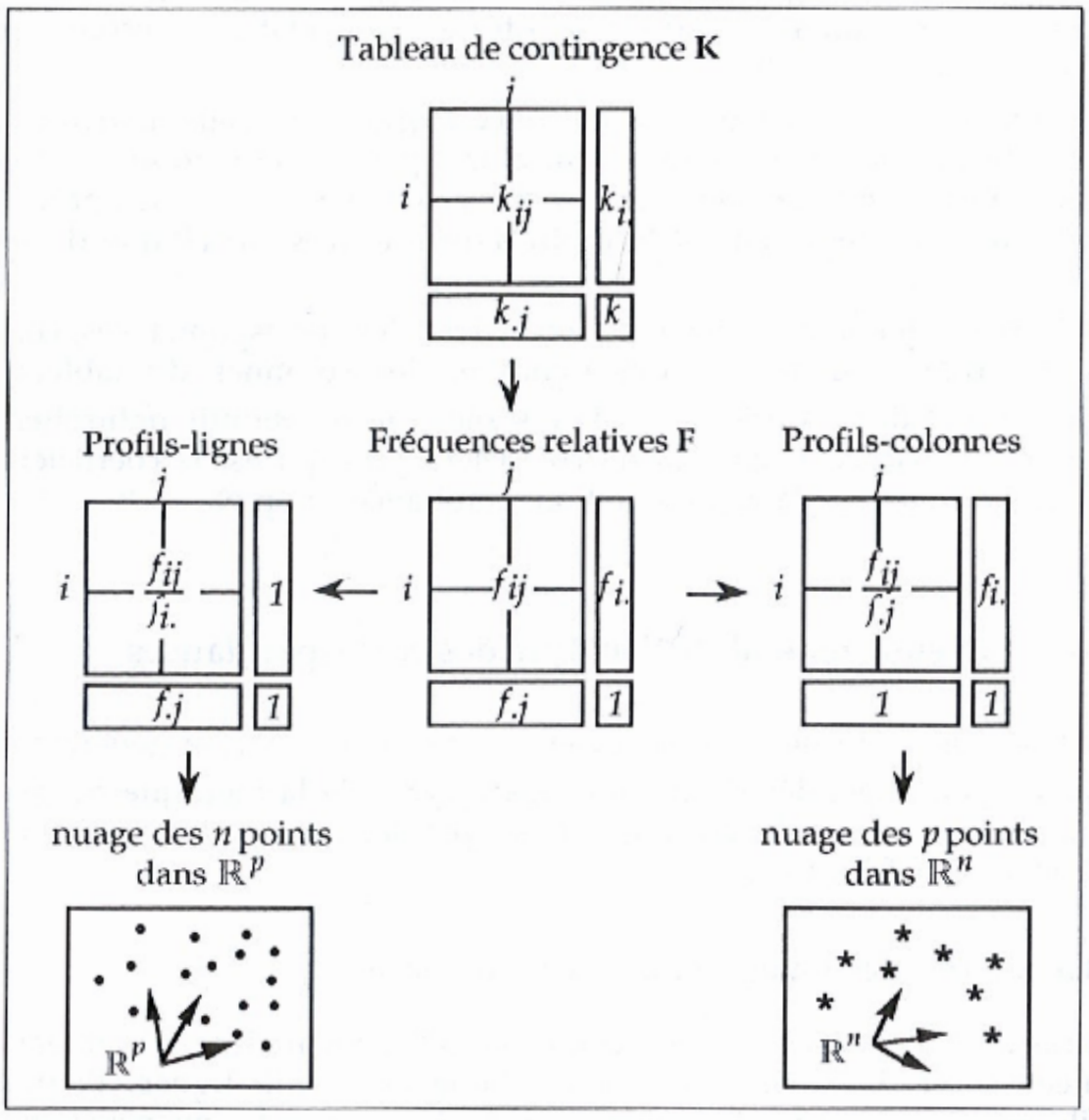
Chaîne d'analyse des données

Nature des données ?

- Qualitative : ordinale, nominale;
- Quantitative;
- Temporelle.

yeux/cheveux	brun	châtain	roux	blond	profil moyen
marron	11	20	5	1	37
noisette	3	9	2	2	16
vert	1	5	2	3	11
bleu	3	14	3	16	36
Profil moyen	18	48	12	22	100

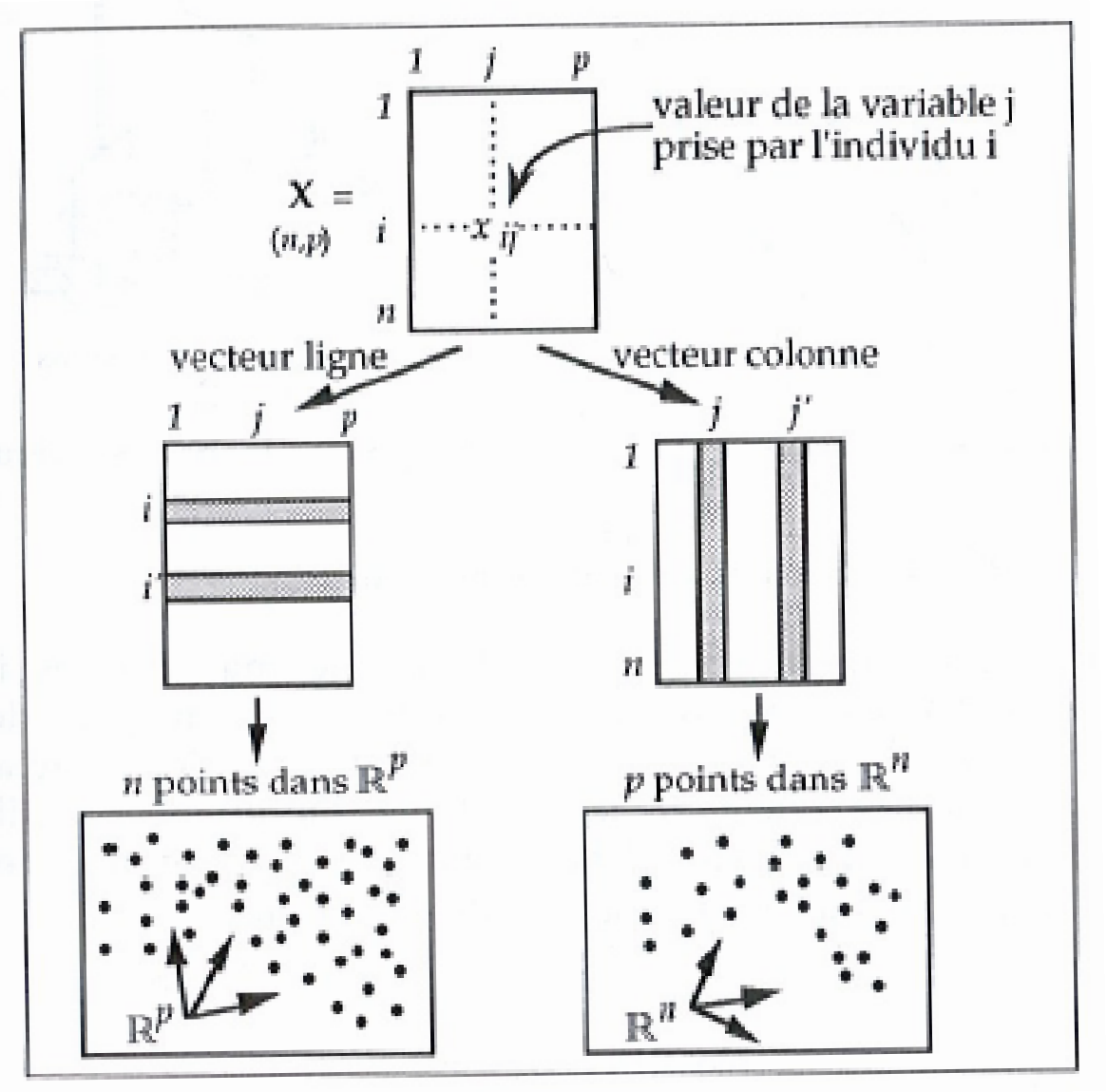
Table: Exemple de tableau de contingence



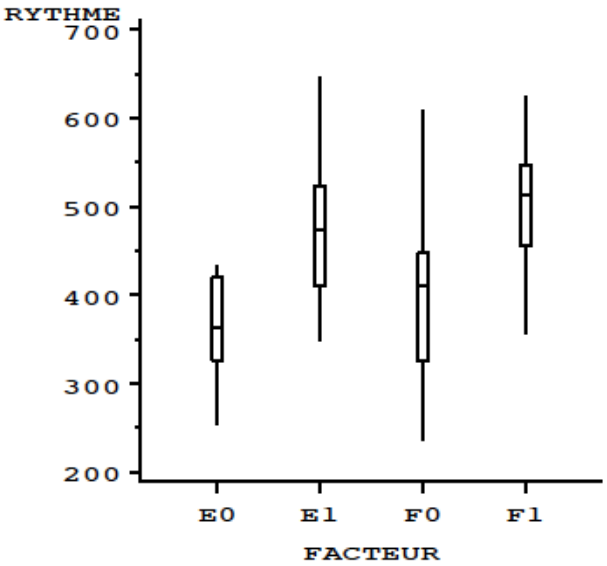
Transformations du tableau de contingence

	bacC	bacD	< 18	18ans	19ans	> 19	2ans	3ans	4ans
bacC	583	0	108	323	114	38	324	192	67
bacD	0	214	25	97	68	24	76	82	56
< 18	108	25	133	0	0	0	84	35	14
18ans	323	97	0	420	0	0	224	137	59
19ans	114	68	0	0	182	0	73	75	34
> 19	38	24	0	0	0	62	19	27	16
2ans	324	76	84	224	73	19	400	0	0
3ans	192	82	35	137	75	27	0	274	0
4ans	67	56	14	59	34	16	0	0	123

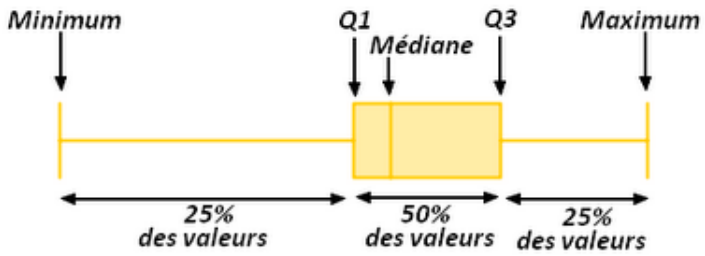
Tableau de Burt



Principe de représentation graphique

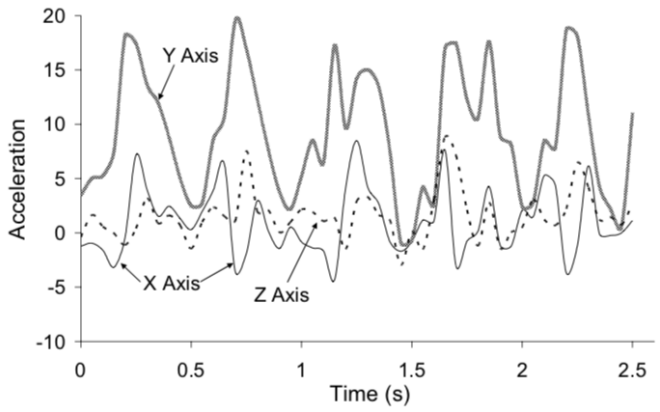


Boîte parallèle

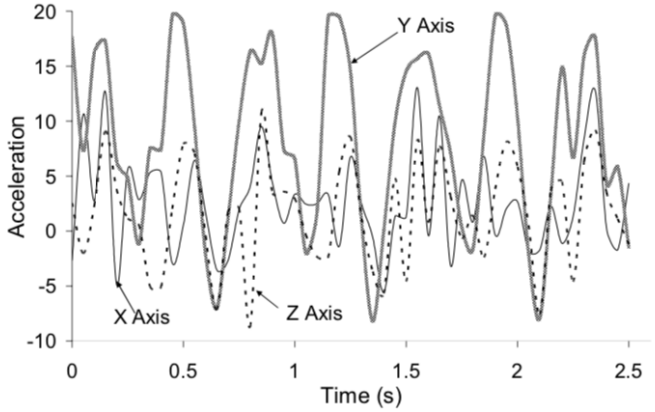


Boîte à moustache (boxplot)

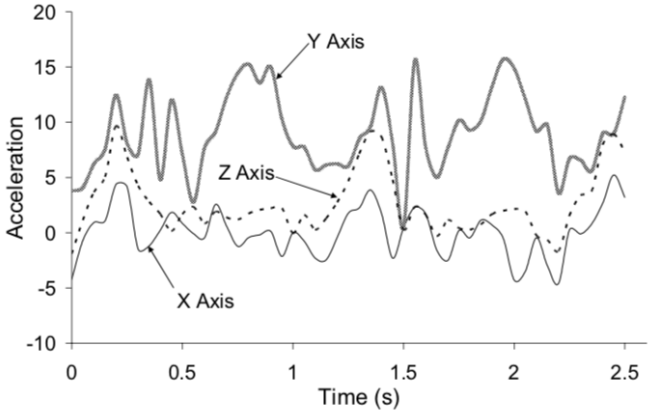
Cas de variables temporelles



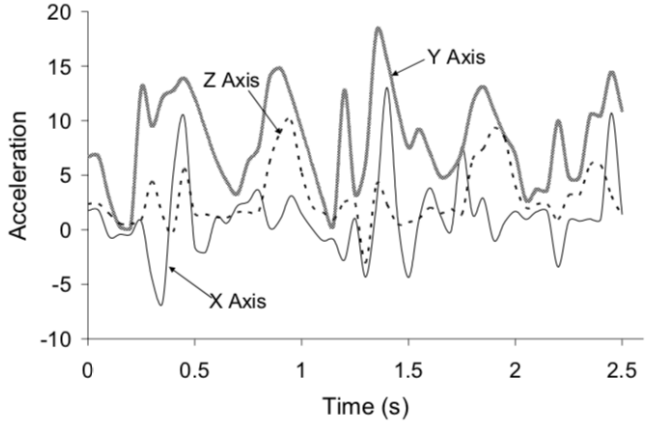
(a) Walking



(b) Jogging



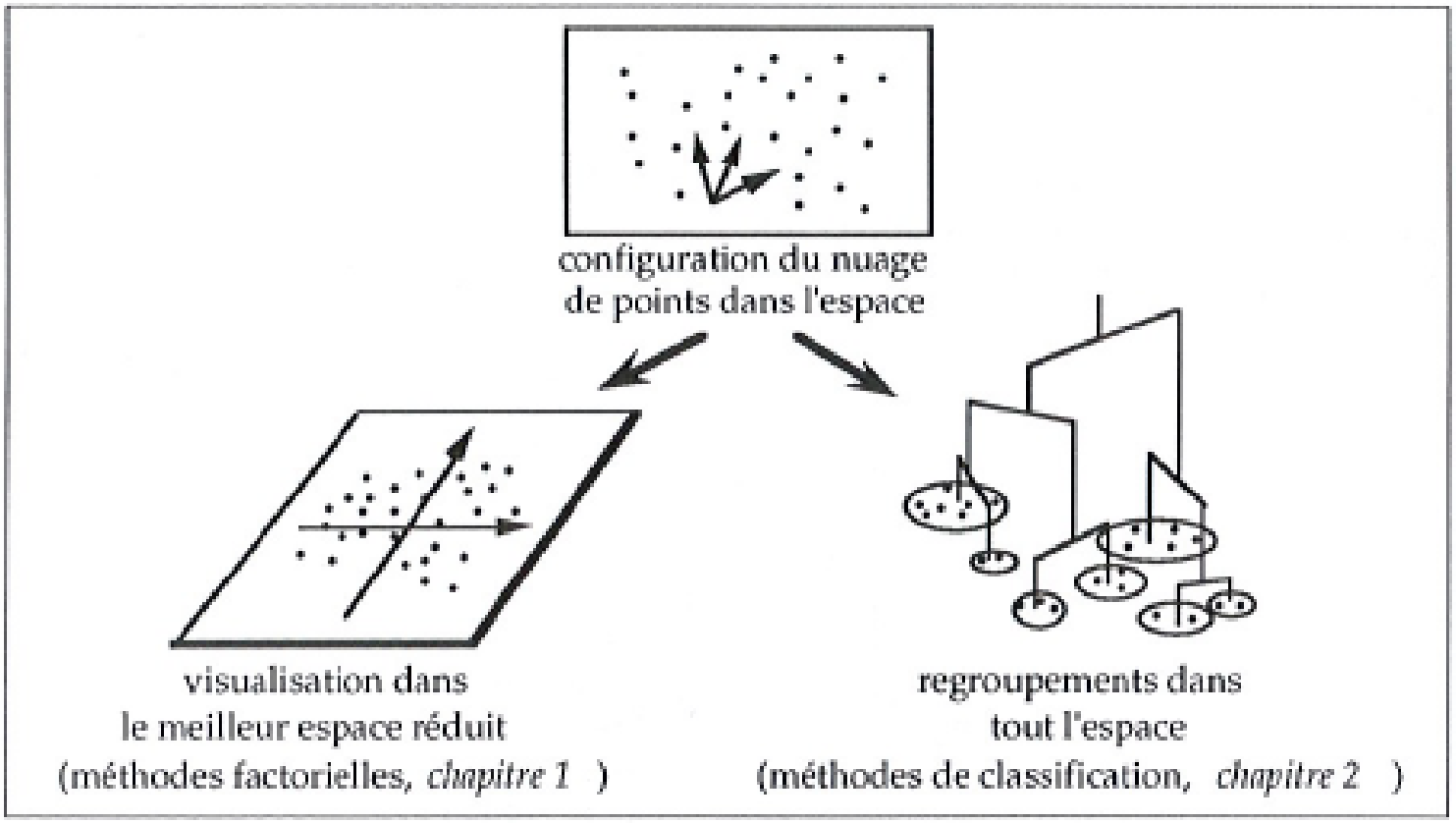
(c) Ascending Stairs

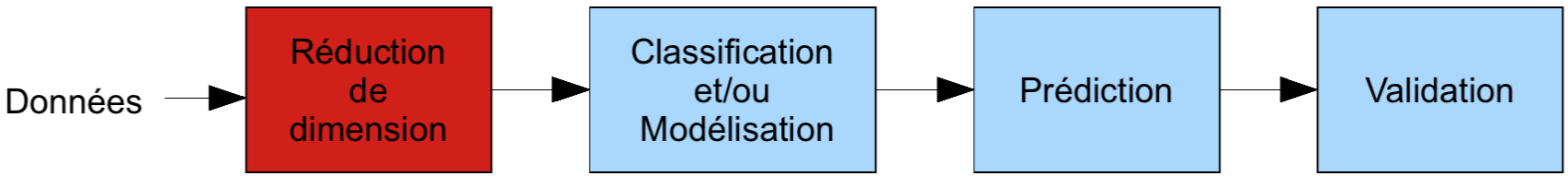


(d) Descending Stairs

Principe de représentation graphique

Vers les méthodes factorielles et les méthodes de classifications





Chaîne d'analyse des données

- Variables quantitatives : Analyse en Composantes Principales (A.C.P)
- Variables qualitatives : Analyse Factorielle des Correspondances (A.F.C)
- Variables temporelles : Analyse Fréquentielle (Analyse Hilbertienne)

Soit X le tableau des données de dimensions $n \times p$ d'élément x_{ij} construit à partir de n individus (ou observations/ unités statistiques/expériences) définis par p variables (ou facteurs/ mesures physiques).

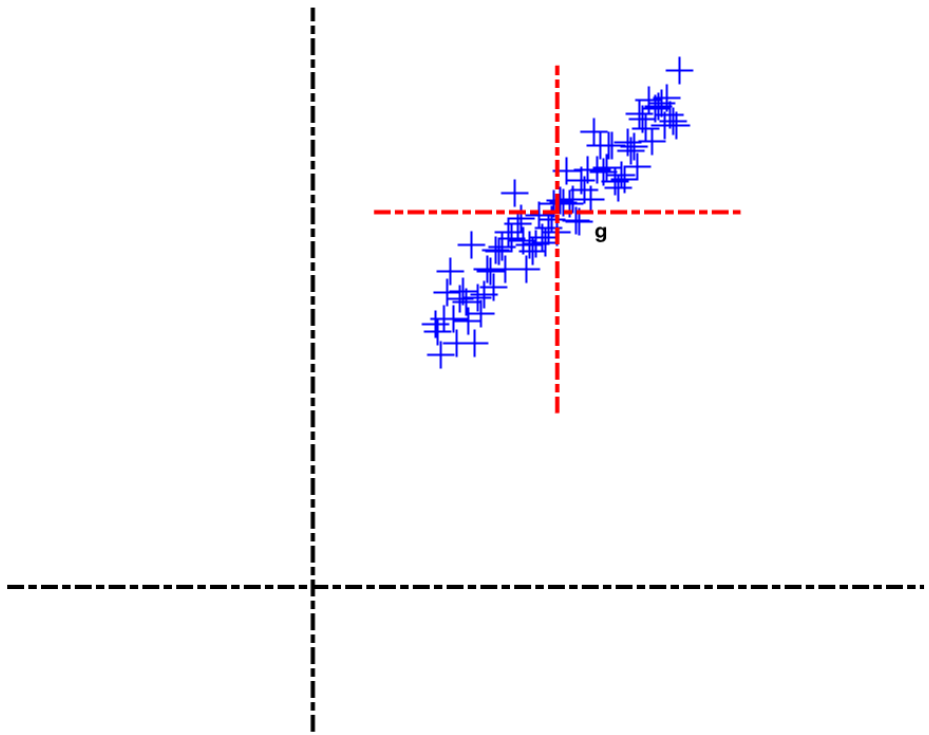
$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & x_{ij} & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

On définit l'*individu moyen* g par le vecteur de $\mathbb{R}^p$ par :

$$g = [\overline{x_1}, .., \overline{x_p}] \text{ avec } \overline{x_j} = \frac{1}{n} \sum_{k=1}^{n} x_{kj}, \ \forall j \in \{1, .., p\}.$$

Soit maintenant X_C le tableau centré en g de dimensions $n \times p$ défini par :

$$X_{C_{ij}} = x_{ij} - \overline{x_j}, \forall i \in \{1, .., n\}, \forall j \in \{1, .., p\}.$$



Centrer la matrice des données

Matrice de variance-covariance Σ

Soit la matrice des données $X \in \mathbb{R}^{n \times p}$. La matrice symétrique Σ de dimension $p \times p$ définie par :

$$\Sigma = \frac{1}{n} X_C^T X_C,$$

avec X_c matrice des données centrées.

- La covariance de la variable j et l, notée Σ_{jl}, sert à mesurer la liaison/dépendance des paramètres :

$$\Sigma_{jl} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x_j})(x_{il} - \overline{x_l})$$

- La variance de la variable j, notée Σ_{jj}, mesure l'écart au carré des données à la moyenne :

$$\Sigma_{jj} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2$$

Mesure de Corrélation

on définit aussi *la corrélation entre les variables X et Y*, indépendant des unités de mesure des variables :

$$-1 \leq Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \leq 1$$

- $Corr(X, Y) = 0$, les variables sont décorrélées, indépendantes c'est-à-dire étant donné X, on ne peut rien dire prédire sur la valeur de Y.
- $Corr(X, Y) = 1$, dépendance linéaire positive de X et Y.
- $Corr(X, Y) = -1$, dépendance linéaire négative de X et Y.

A partir de la matrice Σ, la corrélation entre les variables j et l correspond à $\frac{\Sigma_{jl}}{\sqrt{\Sigma_{jj}\Sigma_{ll}}}$

Problème : soit I une image de niveaux de gris dont les éléments $I_{ij} \in [0, 255]$. Soient α et β une paire de pixels voisins (horizontalement). Les intensités de 2 pixels voisins sont-elles corrélées ?

Résolution : Soit l'image I de taille 3×4 suivante :

$$I = \begin{pmatrix} 7 & 4 & 9 & 7 \\ 0 & 2 & 0 & 3 \\ 1 & 8 & 5 & 7 \end{pmatrix}$$

Décorrélation au + proche voisin :

$$I_d(i,j) \leftarrow I(i,j) - I(i+1,j)$$

Représentation des coefficients décorrélés $\alpha_d \in [-255, 255]$ et $\beta_d \in [-255, 255]$ par histogramme.

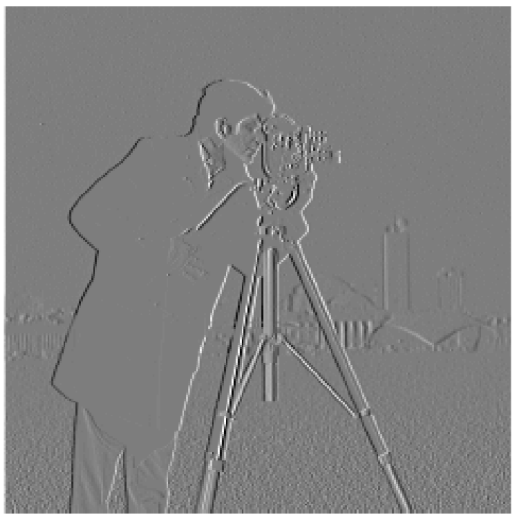
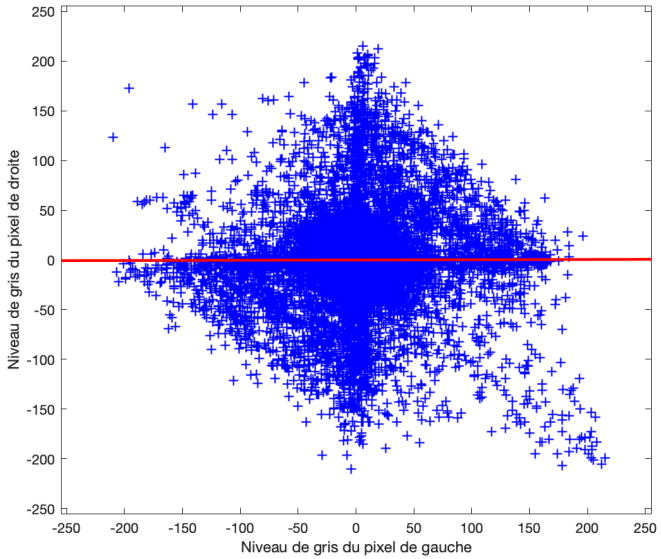
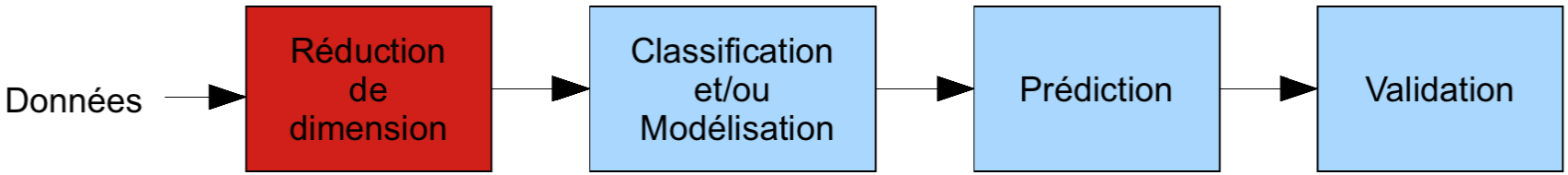


Image décorrélée



Mise en évidence de la décorrélation entre pixels voisins

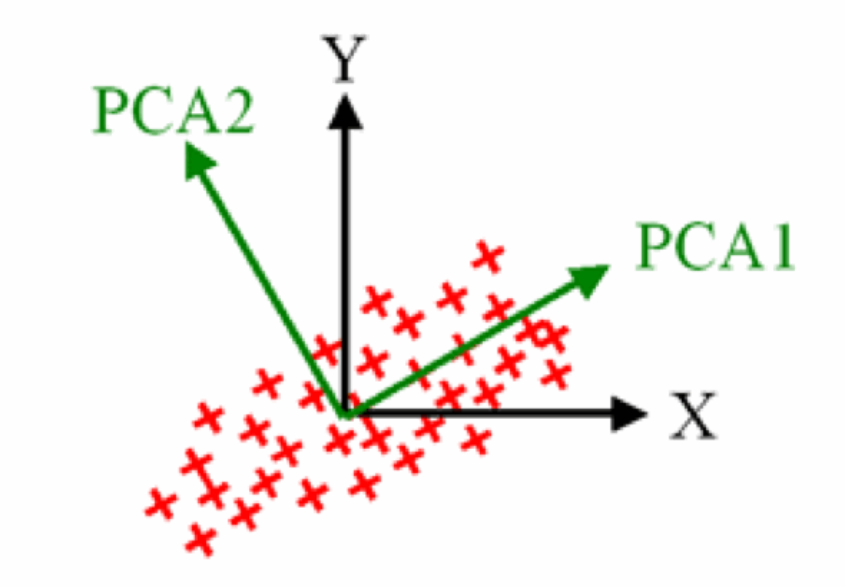


Chaîne d'analyse des données

- Variables quantitatives : **Analyse en Composantes Principales (A.C.P)**
- Variables qualitatives : Analyse Factorielle des Correspondances (A.F.C)
- Variables temporelles : Analyse Fréquentielle (Analyse Hilbertienne)

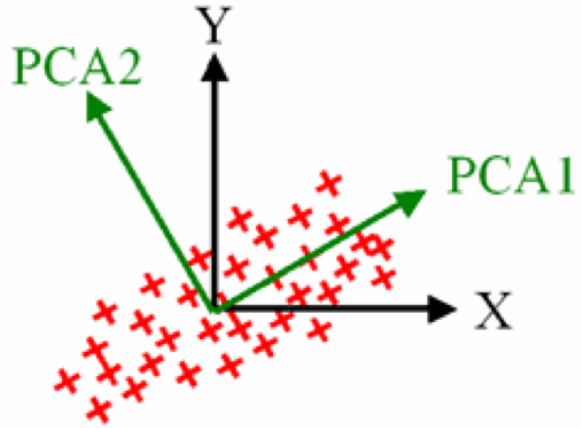
But

Trouver q composantes principales $C_1, .., C_q$ avec $q << p$ comme des nouvelles variables combinaison linéaire des variables d'origines $x_{.1}, ...x_{.p}$ telles que les C_k soient 2 à 2 non corrélées, de variance maximale, d'importance décroissante.



Exemple d'ACP

$\Rightarrow$ Toolbox Scikit-learn (*sklearn*)



Exemple d'ACP

- **Décomposition de la variance :**

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - g)^T(x_i - g)$$

où g est l'individu moyen et x_i est la $i$ème ligne de la matrice des données X.

- **Projection sur une droite :** L'opérateur de projection orthogonale, noté π, sur une droite de vecteur directeur unitaire v s'écrit :

$$\Pi = vv^T$$

avec $v^T v = 1$.

Recherche de la projection de variance maximale

Maximiser cette variance des observations projetées:

$$\max_{v} v^T \Sigma v \text{ avec } v^T v = 1$$

Solution : v est le vecteur propre de Σ associé à la plus grande valeur propre λ.

- Premier axe principal : correspond au vecteur propre associé à la plus grande valeur propre de la matrice de variance-covariance Σ
- Composantes principales : coefficients de projection des données sur les axes principaux

Recherche de la projection de variance maximale

Maximiser cette variance des observations projetées:

$$\max_{v} v^T \Sigma v \text{ avec } v^T v = 1$$

Solution : v est le vecteur propre de Σ associé à la plus grande valeur propre λ.

- Interprétation des vecteurs propres : La somme des valeurs propres correspond à la variance totale:

$$Tr(\Sigma) = \sigma^2 = \sum_{i=1}^{p} \lambda_i$$

  Chaque valeur propre mesure la part de variance expliquée par l'axe factoriel correspondant.

- Choix de la dimension q : La "qualité globale" des représentations est mesurée par la part d'inertie expliquée :

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{i=1}^{p} \lambda_i}.$$

Diagonalisation de matrice : En dimension finie, la diagonalisation revient à décrire une matrice à l'aide d'une **matrice diagonale**.
La diagonalisation concerne les matrices carrées ($n = p$). Dans la suite, on considère une matrice A carrée de dimension $n \times n$.

Calcul des coefficients de cette matrice diagonale ?

Valeurs propres, Vecteurs propres

- Les **valeurs propres** λ_i sont les racines du **polynôme caractéristique** :

$$\chi_A(\lambda) = det(A - \lambda I) \tag{1}$$

  Cette expression est un polynôme de degré n.
- Un vecteur X est un **vecteur propre de** A associé à la valeur propre λ si :

$$AX = \lambda X \Leftrightarrow (A - \lambda I_n) X = 0 \tag{2}$$

Diagonalisation

Si la matrice est diagonalisable alors il existe une matrice P inversible ($det(P) \neq 0$) ayant ses n vecteurs propres comme colonnes et la matrice $D = P^{-1}AP$ est alors diagonale. La diagonale de D est constituée des valeurs propres ordonnées dans le même ordre que les colonnes de P.
On dit alors que D et A sont **semblables**.

Déterminant d'une matrice:

Le **déterminant d'une matrice carrée** A est un scalaire noté $det(A)$ dont la définition n'est pas intuitive. Il peut être calculé, par exemple, en développant suivant la j^{eme} colonne :

$$det(A) = \left| \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \dots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \right| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} detA_{ij}$$

où A_{ij} est la matrice A à laquelle on a ôté la i^{eme} ligne et la j^{eme} colonne.

Pour étudier l'efficacité de capteurs de température sur des montres connectées, on considère les différences entre les mesures de deux capteurs de température par rapport à la température réelle. On obtient les données suivantes sur 4 individus.

Personne	Capteur 1	Capteur 2
Ind. 1	0	2
Ind. 2	-2	-1
Ind. 3	1	0
Ind. 4	1	-1

1. Existe-t-il une dépendance entre ces deux capteurs ? Expliquer votre réponse.
2. Calculer le premier axe principal de ces points.
3. Représenter les données et la premier axe principal.

Décomposition en valeurs singulières SVD

Soit $A \in \mathbb{R}^{q \times p}$ $(q \geq p)$ alors on peut décomposer A :

$$A = U \Sigma V^T$$

avec :

- $U \in \mathbb{R}^{q \times q}$ est formée de q vecteurs propres orthonormés associés aux q valeurs propres de AA^T.
- $V \in \mathbb{R}^{p \times p}$ est formée de p vecteurs propres orthonormés associés aux p valeurs propres de $A^T A$.
- $\Sigma \in \mathbb{R}^{q \times p}$ est une matrice rectangulaire dont les élements non nuls sur la diagonale sont les valeurs singulières $\sigma_i, i = \{1, .., q\}$ de A (sont les racines carrées des valeurs propres de $A^T A$ et de AA^T).

Propriétés :

- $rang(A) = p$ donc $\sigma_{p+1} = ... = \sigma_q = 0$.
- Si $A = U\Sigma V^T$ alors $A^T = V\Sigma U^T$ est une SVD de A^T.

Meilleure approximation de rang inférieur

Soit $A \in \mathbb{R}^{q \times p}$ de SVD $A = \sum_{i=1}^{p} \sigma_i u_i v_i^T$ avec $p = rang(A)$.

Si $k < p$ et $A^k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ alors A^k est la meilleure approximation de A de rang $k < r$ c'est-à-dire :

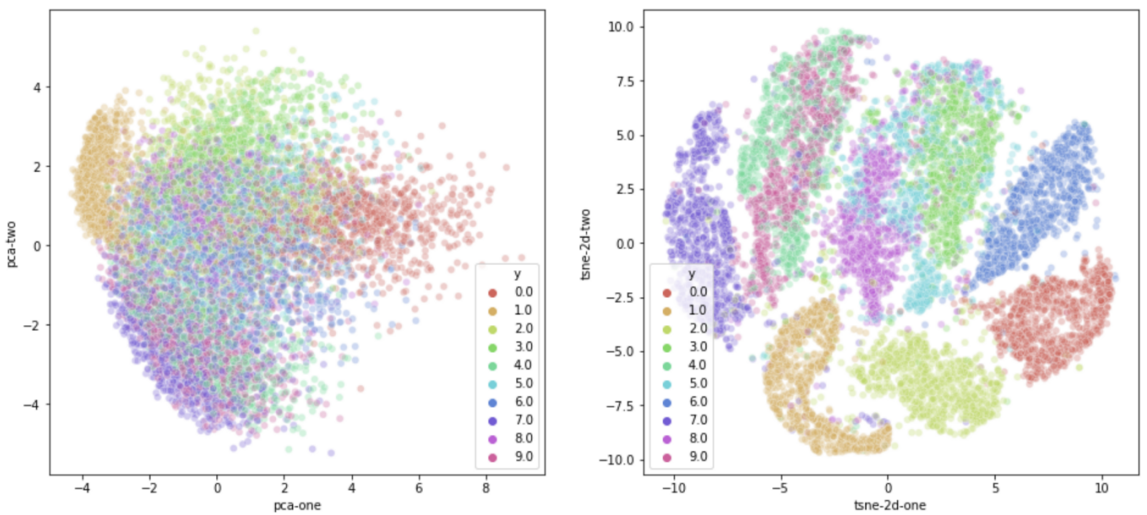
$$\min_{rg(D)=k} \|A - D\|_F = \|A - A^k\|_F$$

t-SNE (t-distributed stochastic neighbor embedding)

méthode non linéaire permettant de représenter un ensemble de points d'un espace à grande dimension dans un espace 2D ou 3D.

⇒ Conserver la proximité entre les points pendant la transformation : deux points qui sont proches (resp. éloignés) dans l'espace d'origine doivent être proches (resp. éloignés) dans l'espace de faible dimension.

⇒ interprétation probabiliste des proximités.



Exemples ACP et t-SNE