Prétraitements
Réduction de dimensions

Analyse de données et Classification 2
ENSEEIHT - 3ème année Sciences du Numérique

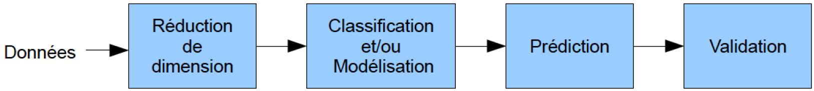
Contact :
Sandrine.Mouysset@irit.fr
sandrine.mouysset@toulouse-inp.fr

Analyse de Données 2 et Classification

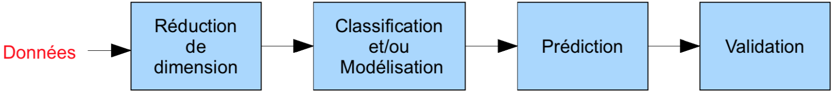
- 8 séances de CTD
- 6 séances de TP

Examens :
Examen écrit
Rendu de projet TP

Analyse de Données 2 et Classification



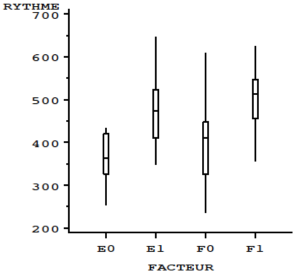
Chaîne d'analyse des données



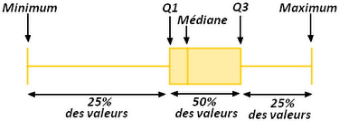
Chaîne d'analyse des données

Nature des données ?

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



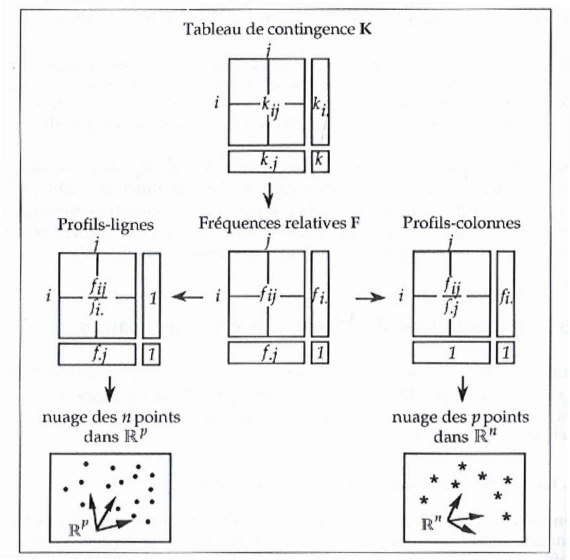
Boîte parallèle



Boîte à moustache (boxplot)

yeux/cheveux	brun	châtain	roux	blond	profil moyen
marron	11	20	4	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	21	100

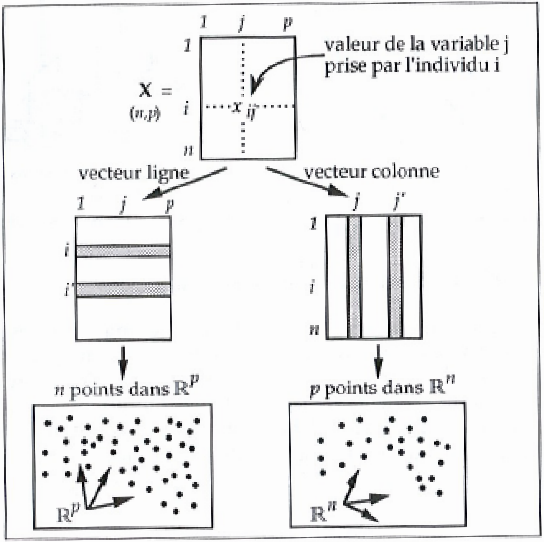
Table: Exemple de tableau de contingence



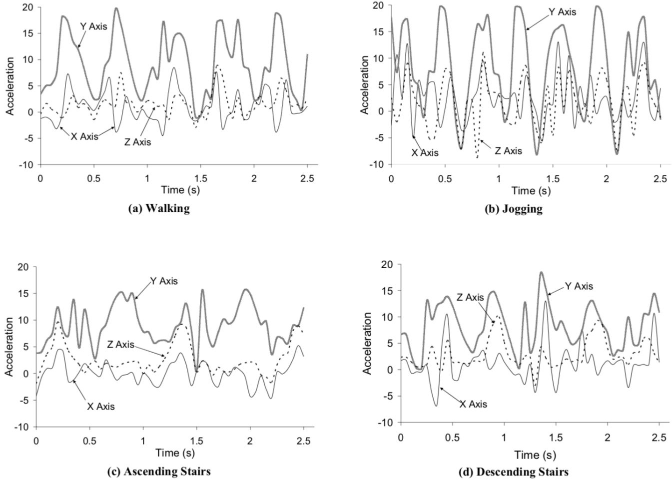
Transformations du tableau de contingence

	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

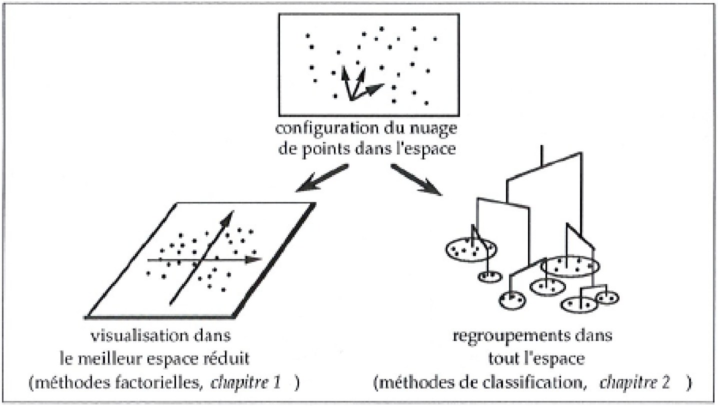
Tableau de Burt

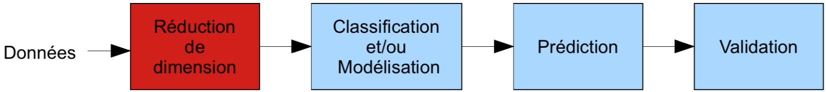


Principe de représentation graphique



Principe de représentation graphique


configuration du nuage
de points dans l'espace
visualisation dans
le meilleur espace réduit
(méthodes factorielles, *chapitre 1*)
regroupements dans
tout l'espace
(méthodes de classification, *chapitre 2*)



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)

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 $\frac{\Sigma_{jl}}{\sqrt{\Sigma_{jj}\Sigma_{ll}}}$

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.

- **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.

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

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

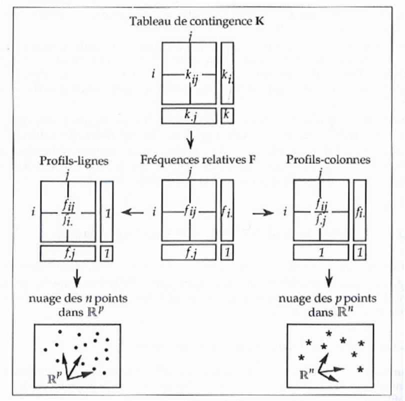
  Chaque valeur propre mesure la part de variance expliquée par l'axe factoriel correspondant.

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

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

Cas des variables qualitatives : Analyse Factorielle des Correspondances (A.F.C)

$\Rightarrow$ s'applique aux tableaux de contingences (tableau croisé de co-occurrence).



- Le **nombre** $f_{ij}/f_{i.}$ représente, la probabilité d'occuper la modalité j sachant que l'on détient la modalité i.
- Le **profil-ligne** $f_{i.}$ n'est rien d'autre que la loi de probabilité conditionnelle définie par i sur l'ensemble des colonnes.

Quelle métrique pour comparer les profils-lignes et les profils-colonnes?

Distance euclidienne entre deux profils-lignes i et i' ?

$$d^2(i,i') : \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

- Distance euclidienne usuelle entre deux profils-lignes traduit bien la ressemblance ou la différence entre 2 modalités
- sans tenir compte des effectifs totaux de ces modalités
- favorise les colonnes qui ont une masse importante

Distance χ^2

Distance du χ^2 entre les profils-lignes i et i' :

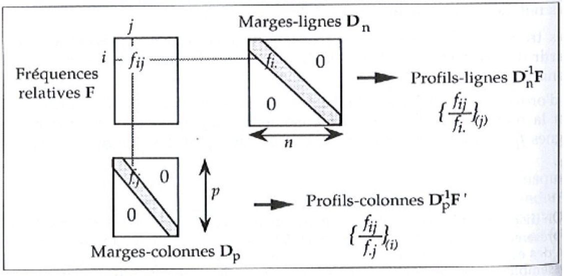
$$\chi^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

Distance χ^2 entre les profils-colonnes j et j' par :

$$\chi^2(j,j') = \sum_{j=1}^{p} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

- la pondération $1/f_{.j}$ équilibre l'influence des colonnes sur la distance entre les lignes
- rôle analogue à celui de la division par l'écart-type dans le cas de variables quantitatives.

But principal de l'AFC : le même que celui de l'ACP c-à-d lire l'information contenue dans un espace multidimensionnel par une réduction de la dimension de cet espace tout en conservant un maximum de l'information contenu dans l'espace de départ.



Fréquences, marges, profils

- F de dimensions $n \times p$ désigne le tableau des **fréquences relatives**;
- D_n de dimensions $n \times n$ est la matrice diagonale dont les éléments diagonaux sont les **marges en lignes** $f_{i.}$;
- D_p de dimensions $p \times p$ est la matrice diagonale des **marges en colonnes** $f_{.j}$.

Nuage de n points-lignes dans l'espace $\mathbb{R}^p$	Éléments de base	Nuage de p points-colonnes dans l'espace $\mathbb{R}^n$
$\mathbf{X} = \mathbf{D}_n{}^{-1}\mathbf{F}$ p coordonnées (point-ligne i) $\frac{f_{ij}}{f_{i.}}$, pour $j = 1, 2, \ldots, p$.	Analyse du tableau $\mathbf{X}$	$\mathbf{X} = \mathbf{D}_p{}^{-1}\mathbf{F}^T$ n coordonnées (point-colonne j) $\frac{f_{ij}}{f_{.j}}$, pour $i = 1, 2, \ldots, n$.
$\mathbf{M} = \mathbf{D}_p{}^{-1}$ $d^2(i, i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} (\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}})^2$	avec la métrique $\mathbf{M}$	$\mathbf{M} = \mathbf{D}_n{}^{-1}$ $d^2(j, j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} (\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}})^2$
$\mathbf{N} = \mathbf{D}_n$ masse du point i : $f_{i.}$	et le critère $\mathbf{N}$	$\mathbf{N} = \mathbf{D}_p$ masse du point j : $f_{.j}$

Table: Elements de base de l'analyse

Dans $\mathbb{R}^p$	Éléments de Construction	Dans $\mathbb{R}^n$
$\mathbf{S} = \mathbf{F}^T {\mathbf{D}_n}^{-1} \mathbf{F} \mathbf{D}_p^{-1}$	Matrice à diagonaliser	$\mathbf{T} = \mathbf{F} {\mathbf{D}_p}^{-1} \mathbf{F}^T \mathbf{D}_n^{-1}$
$\mathbf{S} u_\alpha = \lambda_\alpha u_\alpha$	Axe factoriel	$\mathbf{T} v_\alpha = \lambda_\alpha v_\alpha$
$\psi_\alpha = \mathbf{D}_n^{-1} \mathbf{F} {\mathbf{D}_p}^{-1} u_\alpha$ $\psi_{\alpha i} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.} f_{.j}} u_{\alpha j}$	Coordonnées factorielles	$\phi_\alpha = {\mathbf{D}_p}^{-1} \mathbf{F}^T {\mathbf{D}_n}^{-1} v_\alpha$ $\phi_{\alpha j} = \sum_{i=1}^{n} \frac{f_{ij}}{f_{i.} f_{.j}} v_{\alpha i}$

Table: Elements de construction de l'analyse

Les données temporelles sont souvent associées à des signaux continus dont il est intéressant d'étudier le contenu **fréquentiel** :

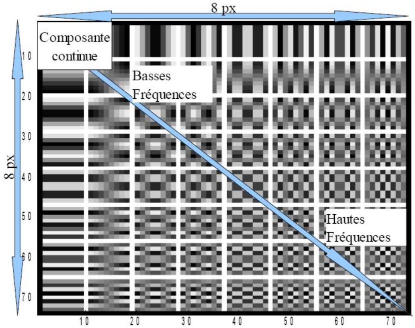
- contenu Fréquentiel pur : analyse de Fourier
  - Idée de l'analyse de Fourier : décomposer, dans le domaine fréquentiel, un signal en une infinité ou un nombre fini de fréquences.
- compromis Temps-Fréquence :
  - Transformée de Gabor,
  - Transformée en ondelettes.

Transformée en Cosinus Discret (DCT) : exprime une suite de nombreux points en termes de somme de fonctions cosinus oscillant à différentes fréquences.

DCT 1D

$$X_k = \sum_{n=0}^{N-1} x_n cos(\frac{\pi}{N}(n+\frac{1}{2})k), \;\; k = \{0, .., N-1\}. \tag{1}$$

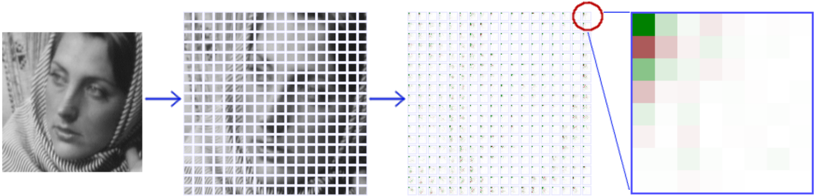
La DCT 1D est la plus utilisée. Cette transformation est l'exacte équivalent (à un facteur 2 près) d'une transformation de Fourier discrète de $4N$ données réelles.



DCT 2D d'une image

$$X_{k_1 k_2} = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} x_{n_1 n_2} cos(\frac{\pi}{N_1}(n_1 + \frac{1}{2})k_1) cos(\frac{\pi}{N_2}(n_2 + \frac{1}{2})k_2), \qquad (2)$$

où $x_{n_1 n_2}$ correspond au niveau de gris du pixel (n_1, n_2) et $k_1 = \{0, .., N_1 - 1\}$, $k_2 = \{0, .., N_2 - 1\}$.



Original image *Pixel blocks* *DCT coefficient blocks* *Single coefficient block*

$\Rightarrow$ Utilisée dans les compressions d'image jpeg, mpeg.

Transformée de Fourier discrète (TFD)

une suite de N termes $x(0), ..., x(N-1)$, la suite de N termes $X(0), ..., X(N-1)$, définis par :

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-2i\pi \frac{nk}{N}}, \ \forall k \in \{0, .., N-1\}.$$

En pratique :

- les N termes $x(n)$ peuvent être N échantillons d'un signal analogique échantillonné: $x_n = x(nT_o)$
- les N termes $X(k)$ correspondre à une approximation (à un facteur multiplicatif T_o près) de la transformée de Fourier de ce signal aux N points de fréquence : $f_k = kf_o/N, \forall k \in \{0, ..., N-1\}$.

Transformée de Fourier Rapide (notée FFT) est simplement une TFD calculée selon un algorithme permettant de réduire le nombre d'opérations et, en particulier, le nombre de multiplications à effectuer.

Idée : Transformation de Fourier par fenêtre (glissante) permettant de localiser simultanément en temps et en fréquence un signal en l'observant sur une fenêtre que l'on translate.

Transformée de Gabor :

$$TG(\tau, f) = \int_{-\infty}^{+\infty} e^{-\pi(u-\tau)^2} e^{-i2\pi f u} x(u) du \quad (3)$$

Par conséquent, cette transformée est une fonction de deux variables (τ, f), où τ désigne le temps et f la fréquence.

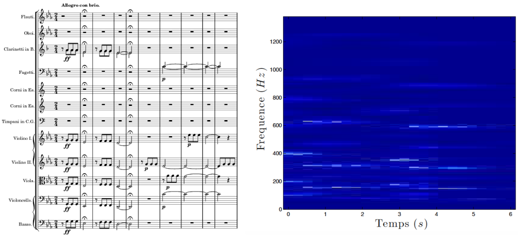


Figure: Représentation par spectrogrammes : sonagramme

Idée : utiliser un signal de base permettant une analyse :

- assez bien localisée temporellement,
- assez bien localisée fréquentiellement (dont le spectre d'amplitude est bien localisé)

$\Rightarrow$ **Ondelettes** : notion de "petite vague" oscillant à un endroit donné approximativement et à une fréquence donnée approximativement.

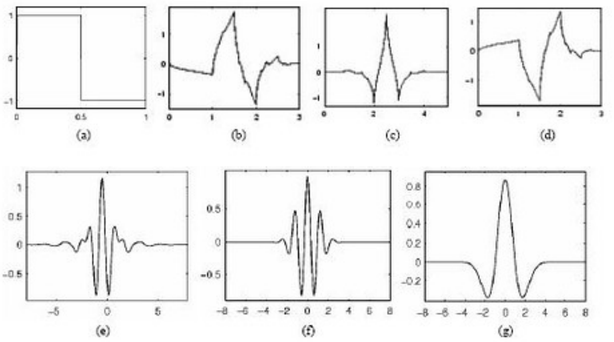


Figure 2.4 Wavelet families (a) Haar (b) Daubechies4 (c) Coiflet1 (d) Symlet2 (e) Meyer (f) Morlet (g) Mexican Hat.

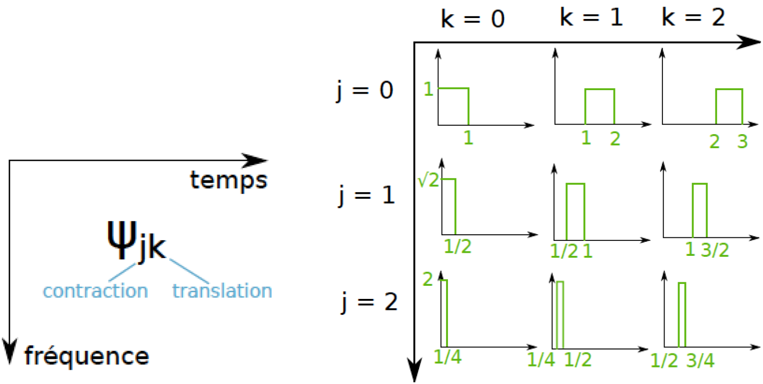
Tout signal f se décompose dans la base des ondelettes ψ_{jk} à deux indices (temps et fréquences) comme suit :

$$f = \sum_j \sum_k (f|\psi_{jk})\psi_{jk} \tag{4}$$

avec j indice des fréquences, k indice du temps et les fonctions de bases ψ_{jk} sont définies par :

$$\psi_{jk} = \underbrace{2^{\frac{j}{2}}}_{\text{normalisation}} \psi(\overbrace{2^j}^{\text{contraction}} x - \overbrace{k}^{\text{translation } 2^{-j}k})$$

où ψ est l'ondelette mère.


temps
Ψjk
contraction
translation
fréquence
k = 0
k = 1
k = 2
j = 0
j = 1
j = 2
1
1
1 2
2 3
√2
1/2
1/2 1
1 3/2
2
1/4
1/4 1/2
1/2 3/4