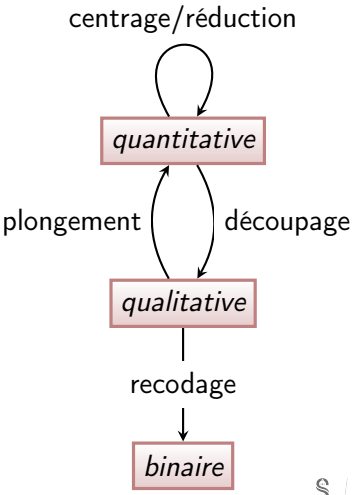
SY09 – Analyse de données et *Data Mining*

Cours n° 2 – Description élémentaire

Sylvain Rousseau

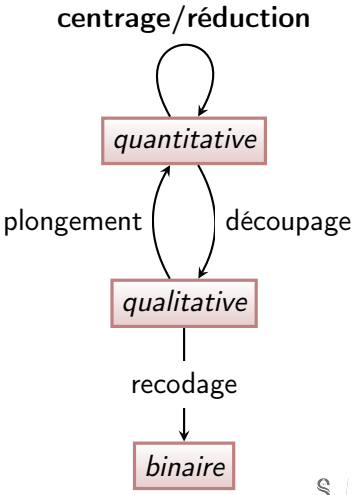
Printemps 2019

Transformation de variables



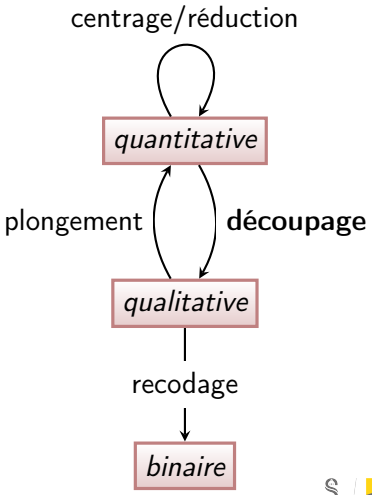
Centrage/Réduction

- Centrage
  - On définit $y_i = x_i - \overline{x}$
  - On a donc $\overline{y} = 0$
- Standardisation
  - On définit $y_i = \frac{x_i - \overline{x}}{s_x^*}$
  - On a donc $\overline{y} = 0$ et $s_y^* = 1$
- Renormalisation MinMax
  - On définit : $y_i = \frac{x_i - \min_i x_i}{\max_i x_i - \min_i x_i}$
  - On a $-1 \leq y_i \leq 1$
- Renormalisation Max
  - On définit $y_i = \frac{x_i}{\max_i |x_i|}$
  - On a $-1 \leq y_i \leq 1$, $x_i = 0 \Rightarrow y_i = 0$



Découpage

- Soit X une variable quantitative $V_X =]a; b]$
- On choisit $a = c_0 < \cdots < c_n = b$
- On définit Y telle que $V_Y = \{M_1, \ldots, M_n\}$
- $Y = M_i$ ssi $X \in]c_{i-1}; c_i]$



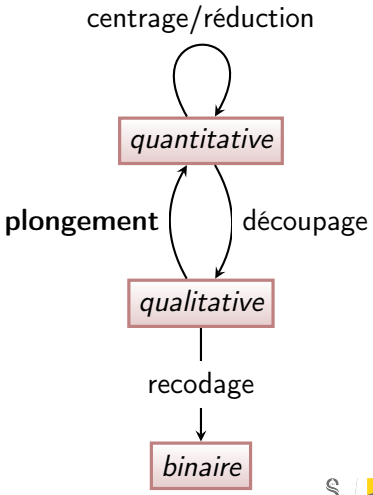
Plongement

Soit X une variable qualitative telle que

$$V_X = \{\text{chat}, \text{chien}, \text{chose}\}$$

On associe une vecteur à chaque modalité. Par exemple :

$$Y = \begin{cases} (1.2, -1.6, 0.4) & \text{si } X = \text{chat} \\ (0.1, -1.1, 1.4) & \text{si } X = \text{chien} \\ (0.4, 1.5, -0.6) & \text{si } X = \text{chose} \end{cases}$$



Recodage

Soit X une variable qualitative $V_X = \{R, G, B\}$

- Nominal : codage disjonctif complet (ou *one hot encoding*)

	v
1	B
2	R
3	B
4	G
5	R

	R ?	G ?	B ?
1	0	0	1
2	1	0	0
3	0	0	1
4	0	1	0
5	1	0	0

- Ordinale : codage additif $R \leq G \leq B$

	v
1	B
2	R
3	B
4	G
5	R

	$\geq R$	$\geq G$	$\geq B$
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	0
5	1	0	0

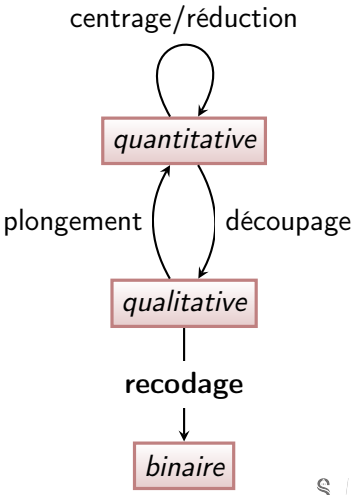


Tableau de proximités

- Une **proximité** P sur une population Ω est une fonction qui à chaque couple d'individus associe une nombre positif :

$$P : \Omega \times \Omega \longrightarrow \mathbb{R}^+$$

- Si la population Ω est finie, P est une matrice de nombre positifs
- Deux grandes familles de proximités :
  - Dissimilarité (distance) : Plus c'est grand, plus c'est loin
  - Similarité : Plus c'est grand, plus c'est près

Dissimilarité

Exemple

- Distances mesurées sur une carte (`eurodist`)

	Athens	Barcelona	Brussels	Calais	Cherbourg	Cologne	Copenhagen
Athens	0	3313	2963	3175	3339	2762	3276
Barcelona	3313	0	1318	1326	1294	1498	2218
Brussels	2963	1318	0	204	583	206	966
Calais	3175	1326	204	0	460	409	1136
Cherbourg	3339	1294	583	460	0	785	1545
Cologne	2762	1498	206	409	785	0	760
Copenhagen	3276	2218	966	1136	1545	760	0

- Remarques :
  1. La diagonale est nulle
  2. La matrice est symétrique

Dissimilarité

Définition

- La fonction d :

$$d : \Omega \times \Omega \longrightarrow \mathbb{R}^+$$

est une **dissimilarité** sur Ω si elle vérifie :

  - $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$ (séparation)
  - $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ (symétrie)

- Si Ω est finie, équivalent à
  1. Diagonale nulle
  2. Matrice symétrique

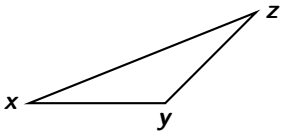
La distance

- La fonction d :

$$d : \Omega \times \Omega \longrightarrow \mathbb{R}^+$$

est une **distance** sur Ω si elle vérifie :

- $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$ (séparation)
- $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ (symétrie)
- $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ (inégalité triangulaire)



La distance ultramétrique

- La fonction d :

$$d : \Omega \times \Omega \longrightarrow \mathbb{R}^+$$

est une **distance ultramétrique** sur Ω si elle vérifie :

- $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$ (séparation)
- $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ (symétrie)
- $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \Omega \quad d(\boldsymbol{x}, \boldsymbol{z}) \leq \max(d(\boldsymbol{x}, \boldsymbol{y}), d(\boldsymbol{y}, \boldsymbol{z}))$ (inégalité ultramétrique)

- Plus contraignante car

$$\max(d(\boldsymbol{x}, \boldsymbol{y}), d(\boldsymbol{y}, \boldsymbol{z})) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$$

La distance euclidienne

Il existe un espace euclidien ($\mathbb{R}^p$) qui réalise les distances observées

La fonction d :

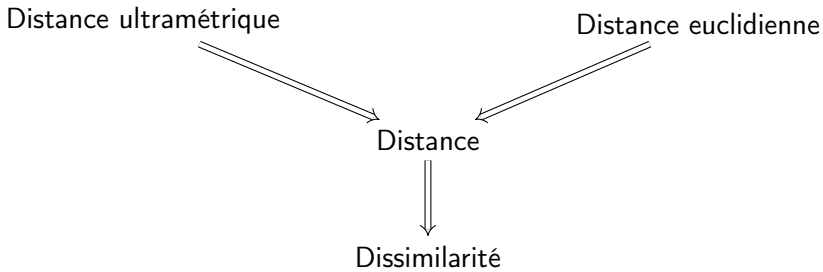
$$d : \Omega \times \Omega \longrightarrow \mathbb{R}^+$$

est une **distance euclidienne** sur Ω telle qu'il existe un entier k et un plongement p de Ω dans $\mathbb{R}^k$ tels que

$$\forall \omega, \omega' \in \Omega, \quad d(\omega, \omega') = \left\| p(\omega) - p(\omega') \right\|_2 .$$

Récapitulatif

Liens entre les différentes distances

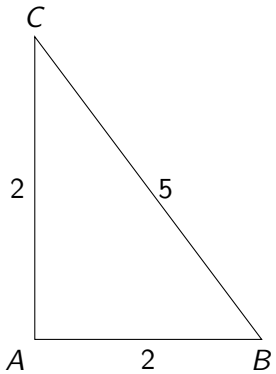


Contre-exemple

Une dissimilarité n'est pas forcément une distance

- La dissimilarité en A, B et C est bien définie
- L'inégalité triangulaire n'est pas vérifiée

$$BC \nleq AC + AB$$



Contre-exemple

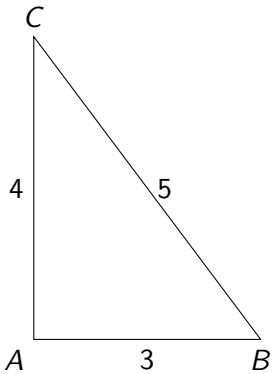
Une distance n'est pas forcément ultramétrique

- Inégalité triangulaire

$$BC \leq AC + AB$$

- Pas une ultramétrique

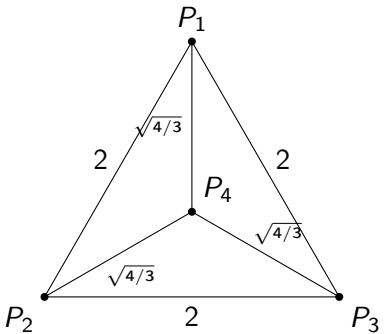
$$BC \not\leq \max(AC, AB)$$



Contre-exemple

Une distance n'est pas forcément euclidienne

- Géométriquement on doit avoir :
  $P_1P_4 = \sqrt{4/3} \approx 1.15$
- Et si on impose $P_4P_i = 1.1$?



Similarité

Exemple

- Similarité : Plus c'est grand, plus c'est près
- Exemple de notes allant de 0 (pas de ressemblance) à 10 (forte ressemblance)

parfums / parfums	1	2	3	4	5
1	–				
2	3	–			
3	5	8	–		
4	2	7	1	–	
5	9	3	5	7	–

- Symétrie, diagonale constante maximale

La fonction s :

$$s : \Omega \times \Omega \longrightarrow \mathbb{R}^+$$

est une **similarité** sur Ω si elle vérifie

- $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \quad s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x})$ (symétrie)
- $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega, \boldsymbol{x} \neq \boldsymbol{y} \quad s(\boldsymbol{x}, \boldsymbol{x}) = s_{\max} \quad \text{avec} \quad s_{\max} \geq s(\boldsymbol{x}, \boldsymbol{y})$

Transformations

- Équivalence similarité/dissimilarité

$$d(\boldsymbol{x}, \boldsymbol{y}) = s_{\max} - s(\boldsymbol{x}, \boldsymbol{y})$$

- Symétrisation

$$d(\boldsymbol{x}, \boldsymbol{y}) = \text{Moyenne}(\delta(\boldsymbol{x}, \boldsymbol{y}), \delta(\boldsymbol{y}, \boldsymbol{x}))$$

- Dissimilarité en distance (il manque l'inégalité triangulaire)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{x}, \boldsymbol{y}) + c \cdot \mathbb{1}_{\boldsymbol{x} \neq \boldsymbol{y}}$$

Construction de tableaux de proximité

À partir d'un tableaux individus-variables :

	variable 1	$\cdots$	variable j	$\cdots$	variable p
individu 1	x_{11}		x_{1j}		x_{1p}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
individu i	x_{i1}		x_{ij}		x_{ip}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
individu n	x_{n1}		x_{nj}		x_{np}

	variable 1	$\cdots$	variable j	$\cdots$	variable p
individu 1	x_{11}		x_{1j}		x_{1p}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
individu i	x_{i1}		x_{ij}		x_{ip}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
individu n	x_{n1}		x_{nj}		x_{np}

Distances entre variables quantitatives

- Euclidienne
  Distance classique « ligne droite »
  $\sqrt{\sum_j (x^j - y^j)^2} = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}$
- Euclidienne pondérée
  Pondéré par une matrice diagonale D
  $\sqrt{(\mathbf{x}-\mathbf{y})^T D(\mathbf{x}-\mathbf{y})}$
- Mahalanobis
  Généralisation avec S définie positive
  $\sqrt{(\mathbf{x}-\mathbf{y})^T S^{-1}(\mathbf{x}-\mathbf{y})}$
- Manhattan ou L_1
  $\sum_j |x_j - y_j|$
- Chebychev ou L_∞
  $\max_j |x^j - y^j|$
- Minkowski ou L_p
  $\left(\sum_j |x^j - y^j|^p\right)^{1/p}$

Distances pour variables qualitatives

Si X et Y sont deux variables qualitative avec I et J modalités chacune

- Distance du χ^2

$$D^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(N_{ij} - \frac{N_{i.} N_{.j}}{n}\right)^2}{\frac{N_{i.} N_{.j}}{n}}$$

- Conversion en binaire, tableaux disjonctif complet

Similarités pour données binaires

- Tableau de contingence

$$x = (1, 0, 1, 0, \ldots, 0)$$
$$y = (0, 0, 1, 1, \ldots, 0)$$

	1	0
1	a	b
0	c	d

- $x = y \iff b = c = 0$

Indice	$s(x, y)$
Csekanowski, Sorensen, Dice	$\frac{2a}{2a+b+c}$
Hamman	$\frac{(a+d)-(b+c)}{a+b+c+d}$
Jaccard	$\frac{a}{a+b+c}$
Kulezynsk	$\frac{a}{a+b}$
Ochiai	$\frac{a}{[(a+b)(a+c)]^{1/2}}$

Statistiques élémentaires

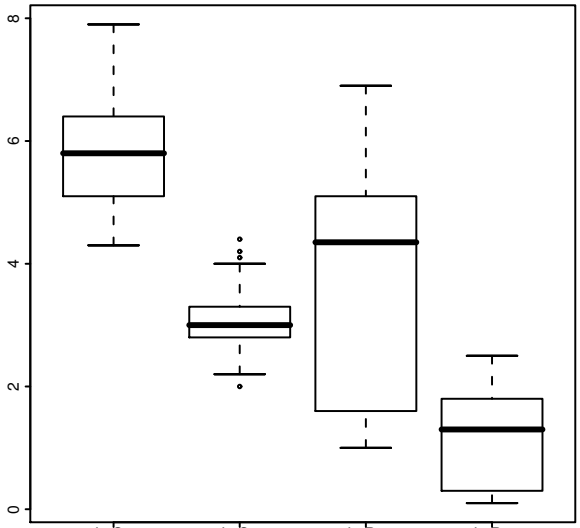
- Position
  - Moyenne empirique
  - Maximum
  - Minimum
  - Médiane
  - Quantile
- Dispersion
  - variance empirique
  - étendue interquartile

Statistiques élémentaires des données Iris

Command R : `summary(iris)`

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

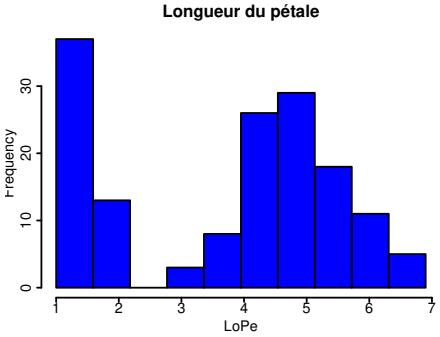
Diagrammes en boites

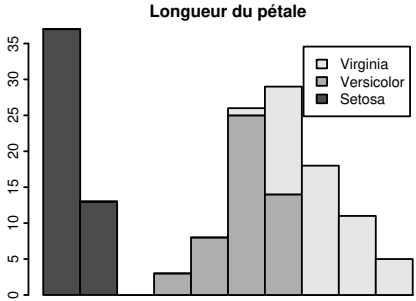


Histogramme

1. Partition de l'étendue en K classes
   - Règle empirique de Sturges : $K \approx 1 + \log_2 n$
2. Table de fréquence
3. Regroupement éventuel

Histogramme : longueur du pétale des données Iris





Estimation de densité

Si X est de densité f :

$$\begin{aligned} f(x) = F'(x) &\approx \frac{F(x+h/2) - F(x-h/2)}{h} \\ &= \frac{1}{h} \Pr(x - h/2 \leq X \leq x + h/2) \\ &= \frac{1}{h} \cdot \frac{\#\{i \mid x - h/2 \leq x_i \leq x + h/2\}}{n} \\ &= \frac{1}{h} \cdot \frac{\#\{i \mid (x - x_i)/h \in [-1/2, 1/2]\}}{n} \\ &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \end{aligned}$$

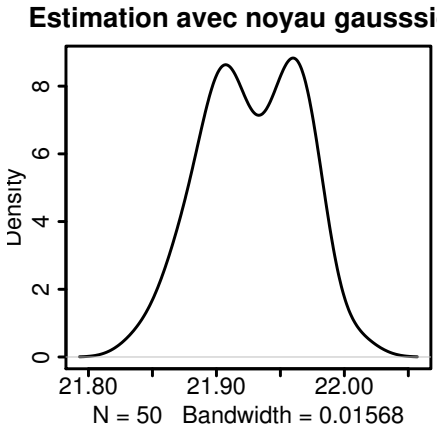
avec $K(x) = \mathbb{1}_{[-1/2,1/2]}(x)$

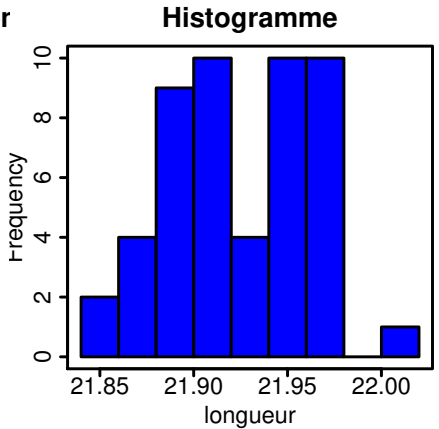
Exemple de noyaux

- Rectangulaire : $K(x) = \mathbb{1}_{[-0.5,+0.5]}(x)$
- Triangulaire : $K(x) = (1 - |x|) \cdot \mathbb{1}_{[-1,+1]}(x)$
- Gaussien : $K(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}x^2)$
- Epanechnikov : $K(x) = \frac{3}{4\sqrt{5}}(1 - x^2/5) \cdot \mathbb{1}_{[-\sqrt{5},+\sqrt{5}]}(x)$
- Lejeune : $K(x) = \frac{105}{64}(1 - x^2)^2(1 - 3x^2) \cdot \mathbb{1}_{[-1,+1]}(x)$

Exemple d'estimation de la densité

21.86	21.92	21.91	21.97	22.01	21.84	21.90	21.91	21.98	21.96
21.88	21.91	21.92	21.95	21.95	21.90	21.89	21.91	21.89	21.95
21.92	21.91	21.93	21.98	21.97	21.87	21.87	21.96	21.96	21.96
21.90	21.89	21.91	21.98	21.95	21.87	21.90	21.97	21.95	21.94
21.90	21.89	21.97	21.97	21.97	21.93	21.92	21.97	21.94	21.95



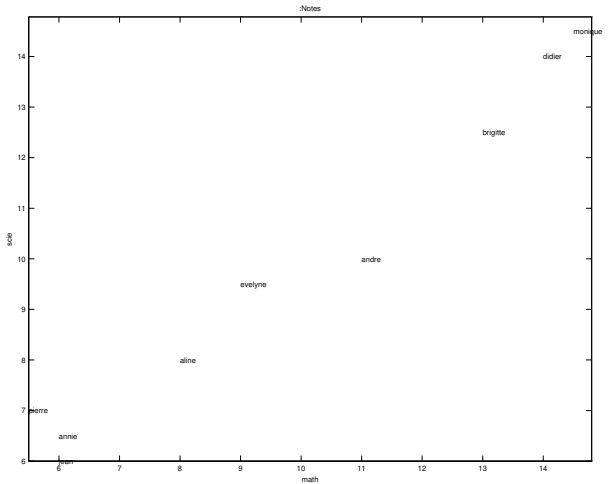


Graphique de dispersion

jean	6.0	6.0
alin	8.0	8.0
anni	6.0	7.0
moni	14.5	14.5
didi	14.0	14.0
andr	11.0	10.0
pier	5.50	7.0
brig	13.0	12.5
evel	9.0	9.5

Graphique de dispersion

jean	6.0	6.0
alin	8.0	8.0
anni	6.0	7.0
moni	14.5	14.5
didi	14.0	14.0
andr	11.0	10.0
pier	5.50	7.0
brig	13.0	12.5
evel	9.0	9.5

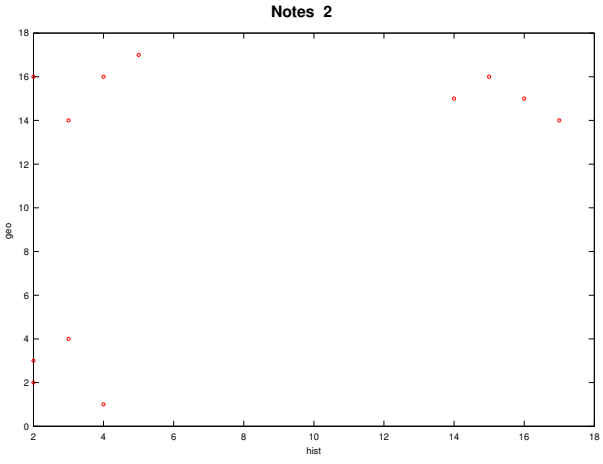


Graphique de dispersion (suite)

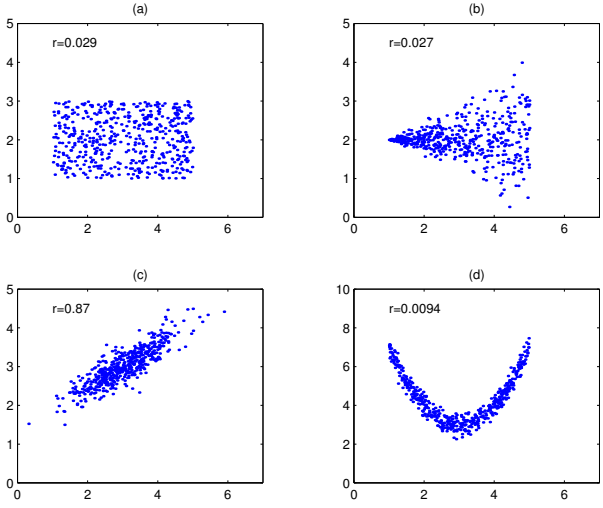
1	15	16
2	2	3
3	4	16
4	16	15
5	3	4
6	3	14
7	4	1
8	17	14
9	5	17
10	2	2
11	14	15
12	2	16

Graphique de dispersion (suite)

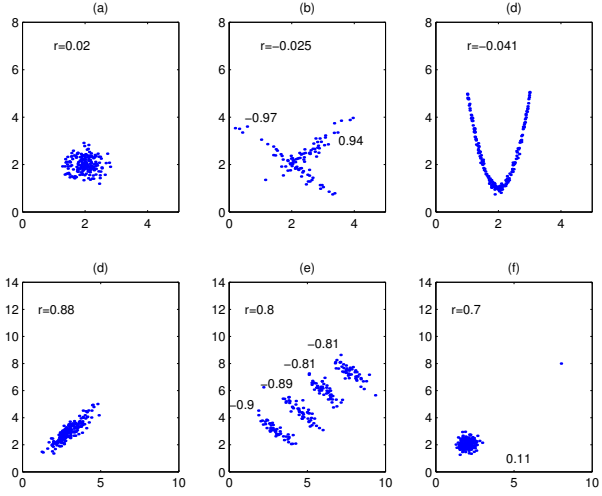
1	15	16
2	2	3
3	4	16
4	16	15
5	3	4
6	3	14
7	4	1
8	17	14
9	5	17
10	2	2
11	14	15
12	2	16



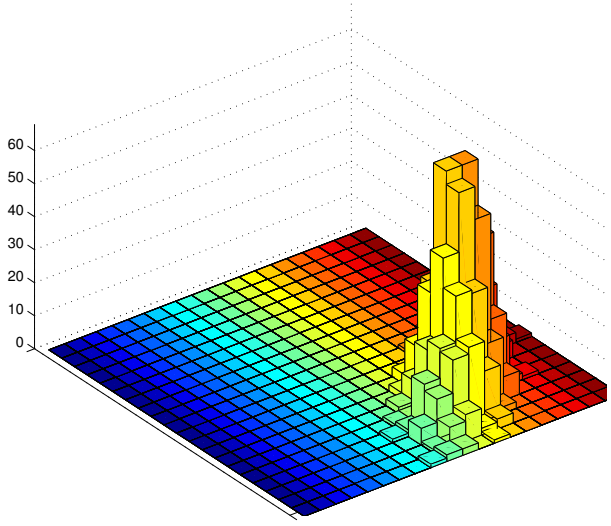
Exemple de corrélation



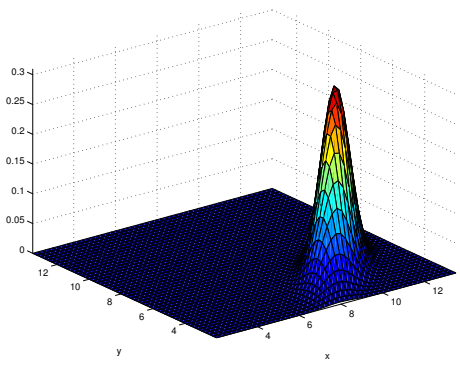
Exemple de corrélation (suite)

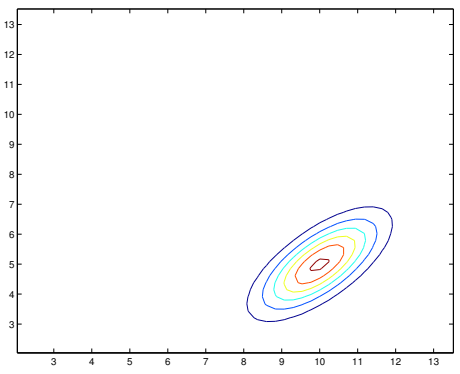


Histogramme bidimensionnel



Estimation de densité bidimensionnelle





Covariance et corrélation des données Iris

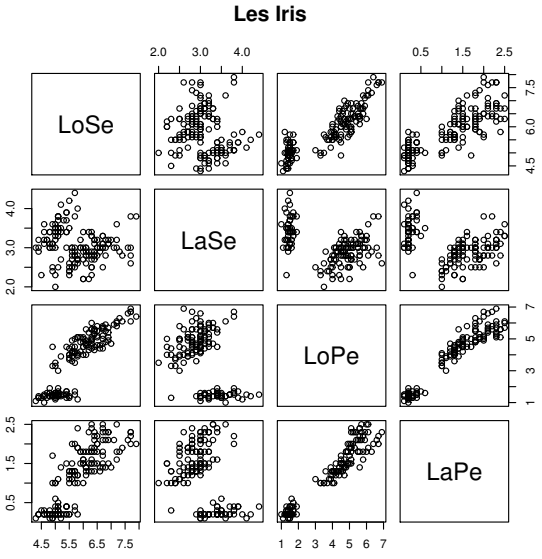
Matrice de covariance

	LoSe	laSe	LoPe	laPe
LoSe	0.69	−0.04	1.3	0.52
laSe	−0.04	0.19	−0.3	−0.12
LoPe	1.27	−0.33	3.1	1.30
laPe	0.52	−0.12	1.3	0.58

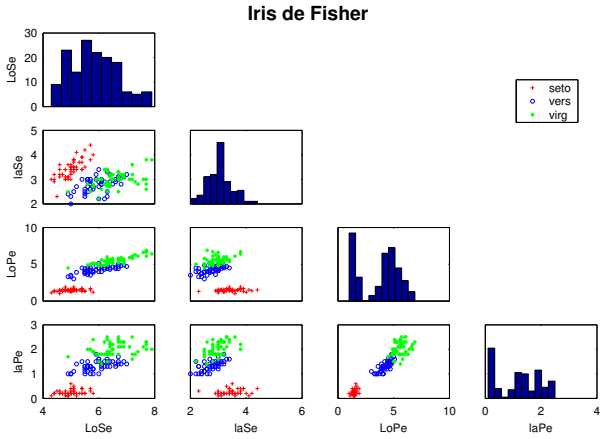
Matrice de corrélation

	LoSe	laSe	LoPe	laPe
LoSe	1.00	−0.12	0.9	0.82
laSe	−0.12	1.00	−0.4	−0.37
LoPe	0.87	−0.43	1.0	0.96
laPe	0.82	−0.37	1.0	1.00

Graphique matriciel



Graphique matriciel avec variable qualitative



Fléau de la dimension

- Le nombre de directions d'angle $\geq \frac{\pi}{3}$

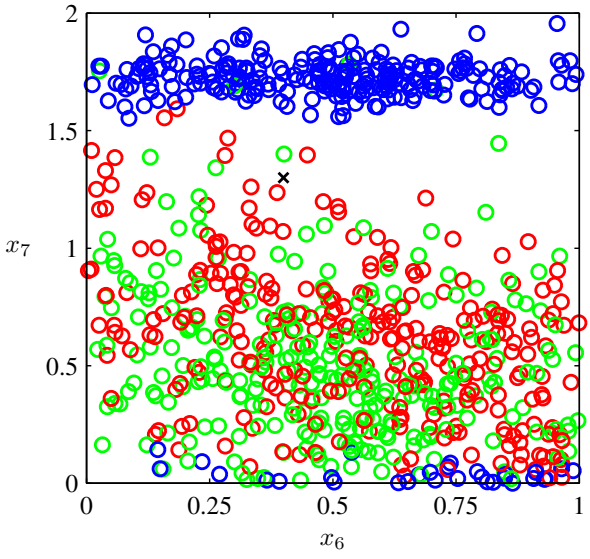
$$\text{nombre de directions} \gtrsim \left(\frac{2}{\sqrt{3}}\right)^n$$

n	2	100	1000
	6	1765781	$> 10^{62}$

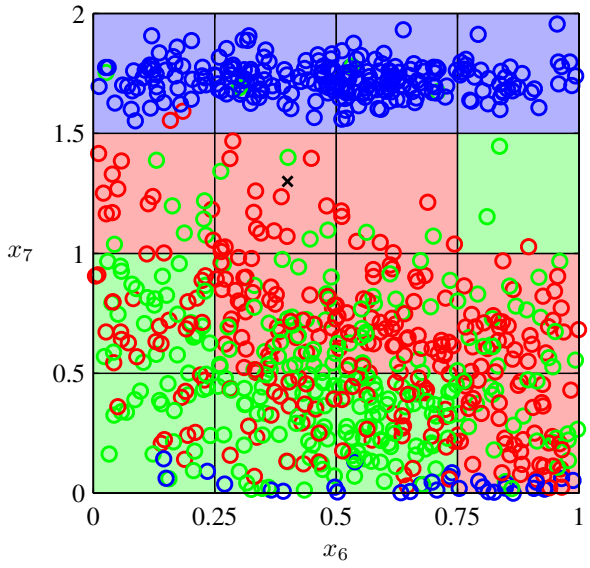
- Pourcentage de points de $[-1,1]^n$ situées dans $[-r,+r]^n$

		n				
		1	2	5	10	100
	0.50	0.50	0.25	0.031	0.00098	7.910^{-31}
r	0.75	0.75	0.56	0.24	0.056	3.210^{-13}
	0.95	0.95	0.90	0.77	0.60	0.0059

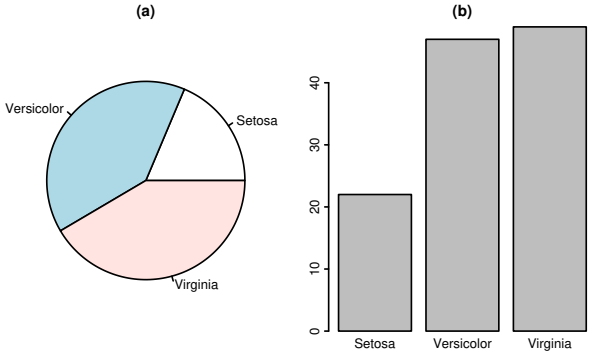
Fléau de la dimension : Oil flow data (1)



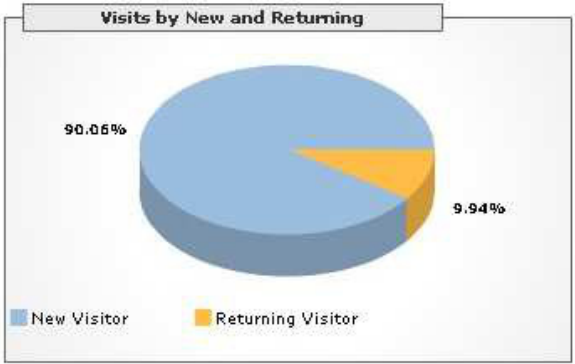
Fléau de la dimension : Oil flow data (2)



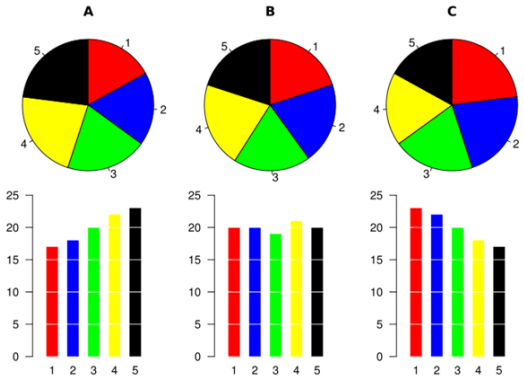
Description de la variable Espèce pour les Iris



Pie-chart : les défauts de ce type de représentation



Pie-chart : la lecture d'un angle est difficile



Pie-chart : Trop de modalités