Chapitre 5-2) (Dis)similarités, distances et inerties

Maxime El Masri

3 MIC / INSA Toulouse

2023-2024

Objectif

- Données : On observe n individus décrits par p variables

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \text{ avec } x_i = (x_{i1}, \ldots, x_{ip}) \in \mathcal{X}$$

- Objectif de la classification non supervisée :

  trouver une organisation en classes homogènes de n individus telle que

  - 2 individus d'une même classe **se ressemblent** plus que deux individus de classes différentes
  - les classes soient bien **séparées**

$\Longrightarrow$ besoin d'une notion de **(dis)similarité** entre individus et d'une mesure de séparabilité des classes.

Partie 2

1. (Dis)similarités et distances
   - Définitions
   - Pour des variables quantitatives
   - Pour des variables qualitatives
   - Pour des variables mixtes
2. Inerties

Partie 2

1. (Dis)similarités et distances
   - Définitions
   - Pour des variables quantitatives
   - Pour des variables qualitatives
   - Pour des variables mixtes

(Dis)similarité entre individus

Dissimilarité

Une **dissimilarité** est une fonction $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ telle que

- $\forall (x_i, x_\ell) \in \mathcal{X} \times \mathcal{X},\ d(x_i, x_\ell) = d(x_\ell, x_i)$ (symétrie)
- $d(x_i, x_\ell) = 0 \Leftrightarrow x_i = x_\ell$

Similarité

Une **similarité** (normée) est une fonction $s : \mathcal{X} \times \mathcal{X} \to [0, 1]$ telle que

- $\forall (x_i, x_\ell) \in \mathcal{X} \times \mathcal{X},\ s(x_i, x_\ell) = s(x_\ell, x_i)$ (symétrie)
- $s(x_i, x_\ell) = 1 \Leftrightarrow x_i = x_\ell$

Distances entre individus

Distance

Une **distance** est une dissimilarité d satisfaisant en plus l'inégalité triangulaire

$$\forall (x_i, x_\ell, x_m) \in \mathcal{X}^3,\ d(x_i, x_m) \leq d(x_i, x_\ell) + d(x_\ell, x_m)$$

La distance est dite euclidienne s'il existe une norme $\|.\|$ sur l'espace des variables telle que $d(x_i, x_\ell) = \|x_i - x_\ell\|$

Partie 2

1. (Dis)similarités et distances
   - Définitions
   - Pour des variables quantitatives
   - Pour des variables qualitatives
   - Pour des variables mixtes

Distances issues de normes

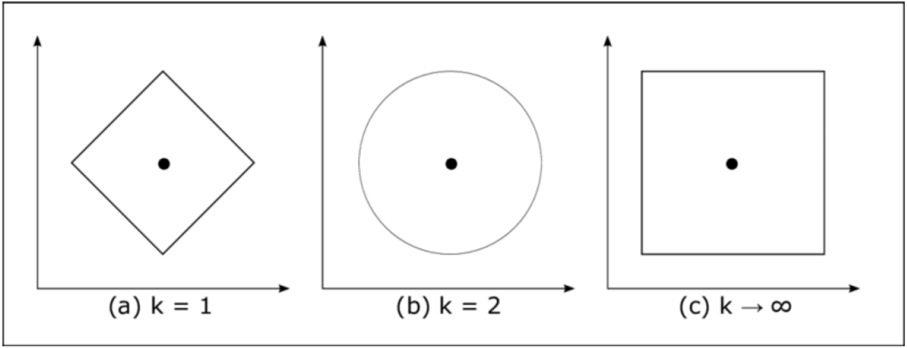
- $x_i \in \mathcal{X} = \mathbb{R}^p$ pour tout $i = 1, \ldots, n$
- Distance de Minkowski (norme L_q)

$$d(x_i, x_\ell) = \left(\sum_{j=1}^{p} |x_{ij} - x_{\ell j}|^q \right)^{\frac{1}{q}}$$

- Cas particuliers :
  - Distance euclidienne usuelle ($q = 2$) :
    $d(x_i, x_\ell) = \|x_i - x_\ell\|_2 = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{\ell j})^2}$
  - Distance de Manhattan ($q = 1$) : $d(x_i, x_\ell) = \|x_i - x_\ell\|_1 = \sum_{j=1}^{p} |x_{ij} - x_{\ell j}|$
- Norme infinie ($q \to +\infty$) : $d(x_i, x_\ell) = \max_{j=1,\ldots,p} |x_{ij} - x_{\ell j}|$

Norme L_1, L_2 et norme infinie

Boules unités



$\longrightarrow$ invariantes par translation mais sensibles à l'échelle des variables.

Définitions

- Moyennes de la variable $j \in \{1, \ldots, p\}$:

$$m_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

- Déviation absolue moyenne :

$$s_j = \frac{1}{n} \sum_{i=1}^{n} |x_{ij} - m_j|$$

- (Variances -)Covariances entre deux variables j et k :

$$\Sigma_{jk} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - m_j)(x_{ik} - m_k)$$

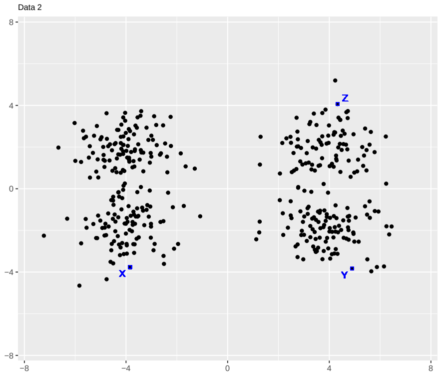
Distances issues de formes quadratiques

Distances définies comme des formes quadratiques

$$\forall (x_i, x_l) \in \mathcal{X},\ d^2(x_i, x_\ell) = (x_i - x_\ell)' M (x_i - x_\ell)$$

- Norme euclidienne usuelle : $M = I_p$
- $M = \text{diag}\left(\frac{1}{\sigma_1^2}, \dots, \frac{1}{\sigma_p^2}\right)$ où $\sigma_j^2 = \Sigma_{jj}$
- $M = \text{diag}\left(\frac{1}{s_1^2}, \dots, \frac{1}{s_p^2}\right)$
- Distance de Mahalanobis : $M = \Sigma^{-1}$, (Σ, matrice de variance-covariance) $\longrightarrow$ permet de réduire l'influence des données aberrantes

Exemples



- Distance euclidienne $\|.\|_2$

	X	Y	Z
X	0.00	8.74	11.31
Y	8.74	0.00	7.91
Z	11.31	7.91	0.00

- Distance de Manhattan $\|.\|_1$

	X	Y	Z
X	0.00	8.80	15.99
Y	8.80	0.00	8.46
Z	15.99	8.46	0.00

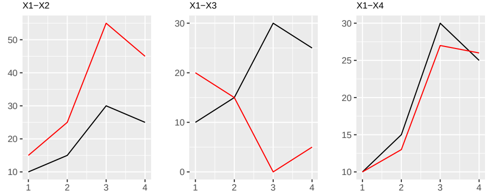
- Distance de Mahalanobis $\|.\|_{2,\Sigma^{-1}}$

$$\Sigma = \begin{pmatrix} 16.71 & -0.53 \\ -0.53 & 4.6 \end{pmatrix}$$

	X	Y	Z
X	0.00	2.14	4.28
Y	2.14	0.00	3.68
Z	4.28	3.68	0.00

Dissimilarités basées sur le coefficient de corrélation

- Coefficient de corrélation $\rho(x_i, x_\ell) \in [-1, 1]$
- Exemples de dissimilarités basées sur la corrélation
  - $d(x_i, x_\ell) = 1 - \rho(x_i, x_\ell)$
  - $d(x_i, x_\ell) = 1 - |\rho(x_i, x_\ell)|$
  - $d(x_i, x_\ell) = 1 - \rho(x_i, x_\ell)^2$



	X_2	X_3	X_4
$1 - \rho(X_1, .)$	0	2	0.02
$1 - \|\rho(X_1, .)\|$	0	0	0.02
$1 - \rho(X_1, .)^2$	0	0	0.04
$\|X_1 - .\|_2$	33,9	37,41	3,74

Partie 2

1. (Dis)similarités et distances
   - Définitions
   - Pour des variables quantitatives
   - Pour des variables qualitatives
   - Pour des variables mixtes

Pour des variables binaires

- Table de contingence entre 2 individus x_i et $x_\ell \in \{0,1\}^p$:

	1	0
1	m_{11}	m_{10}
0	m_{01}	m_{00}

- Variable binaire **symétrique**
  = pas d'influence sur le choix du codage $0-1$

Similarités :

Appariement simple	$s(x_i,x_\ell) = \frac{m_{11}+m_{00}}{m_{11}+m_{00}+m_{10}+m_{01}}$
Rogers et Tanimoto	$s(x_i,x_\ell) = \frac{m_{11}+m_{00}}{m_{11}+m_{00}+2(m_{10}+m_{01})}$
Sokal et Sneath	$s(x_i,x_\ell) = \frac{2(m_{11}+m_{00})}{2(m_{11}+m_{00})+m_{10}+m_{01}}$

- Variable binaire **asymétrique**
  = les valeurs 0-1 n'ont pas la même importance

Jaccard	$s(x_i,x_\ell) = \frac{m_{11}}{m_{11}+m_{10}+m_{01}}$
Dice	$s(x_i,x_\ell) = \frac{2m_{11}}{2m_{11}+m_{10}+m_{01}}$

Exemple

Nom	Sexe	Fièvre	Toux	Test1	Test2	Test3	Test4
Jules	M	P	N	P	N	N	N
Marie	F	P	N	P	N	P	N
Pierre	M	P	P	N	N	N	N
Anna	F	N	P	N	P	N	N

Tableaux des similarités :

- Jaccard :

	Jules	Marie	Pierre	Anna
Jules	1.0	0.5	0.50	0.00
Marie	0.5	1.0	0.20	0.00
Pierre	0.5	0.2	1.00	0.25
Anna	0.0	0.0	0.25	1.00

- Appariement simple :

	Jules	Marie	Pierre	Anna
Jules	1.00	0.71	0.71	0.29
Marie	0.71	1.00	0.43	0.29
Pierre	0.71	0.43	1.00	0.57
Anna	0.29	0.29	0.57	1.00

Pour des variables nominales

- Variables ayant plus de 2 modalités
  - Ex1 : couleur des yeux $\{$bleu, marron, vert$\}$
  - Ex2 : statut marital : $\{$marié, célibataire, pacsé,divorcé, veuf$\}$

- Coefficient d'appariement simple :

$$s(x_i, x_\ell) = \frac{u}{p}$$

où $u =$ nombre de variables où x_i et x_ℓ ont la même modalité

Pour des variables nominales

- Transformer la variable nominale en variables binaires (une par modalité)

Le tableau disjoint complet Z associé à $\underline{\mathbf{x}}$ de taille $n \times \tilde{p}$:

$$\underline{\mathbf{x}} = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 3 & 4 \\ 2 & 1 & 1 \\ \vdots & \vdots & \vdots \end{pmatrix} \Rightarrow Z = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$+$ utiliser une distance/dissimilarité pour variables binaires

- Distance du χ^2 entre individus :

$$d^2(x_i, x_\ell) = \frac{n}{p} \sum_{j=1}^{\tilde{p}} \frac{(Z_{ij} - Z_{\ell j})^2}{Z_{\cdot j}} \text{ avec } Z_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} Z_{ij}$$

- (Distance pour données quantitatives sur les coordonnées de l'ACM ($\simeq$ ACP pour variables quali/mixtes))

Partie 2

1. (Dis)similarités et distances
   - Définitions
   - Pour des variables quantitatives
   - Pour des variables qualitatives
   - Pour des variables mixtes

Cas des variables mixtes

- 1ère stratégie : tout transformer en variables de même nature
- 2ème stratégie : métrique de Gower

$$d(x_i, x_\ell) = \sum_{j=1}^{p} \delta_{i\ell}^{(j)} d_{i\ell}^{(j)} \Big/ \sum_{j=1}^{p} \delta_{i\ell}^{(j)}$$

avec

$$\delta_{i\ell}^{(j)} = \begin{cases} 0 & \text{si } \begin{cases} x_{ij} \text{ ou } x_{\ell j} \text{ est manquante} \\ x_{ij} = x_{\ell j} = 0 \text{ et } j \text{ variable binaire asymétrique} \end{cases} \\ 1 & \text{sinon.} \end{cases}$$

et

$$d_{i\ell}^{(j)} = \begin{cases} \mathbb{1}_{x_{ij} \neq x_{\ell j}} & \text{si } j \text{ variable binaire ou nominale} \\ \dfrac{|x_{ij} - x_{\ell j}|}{\max\limits_{1 \leq h \leq n} x_{hj} - \min\limits_{1 \leq h \leq n} x_{hj}} & \text{si } j \text{ est quantitative} \end{cases}$$

Conclusion

- Bien adapter le choix de la distance (dissimilarité) à
  - la nature des données étudiées
  - la définition de ressemblance entre individus dans le contexte
  - la méthode de clustering choisie
- Attention au comportement de la distance en grande dimension (beaucoup de variables)

Partie 2

1 (Dis)similarités et distances

- Définitions
- Pour des variables quantitatives
- Pour des variables qualitatives
- Pour des variables mixtes

2 Inerties

Inerties intra- / inter- classes

Définitions

Soit d une distance euclidienne entre individus.
Soit $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ une partition des individus en K classes.

- **Inertie totale** : $I_T = \sum_{i=1}^{n} d(x_i, c)^2$

  où $c = \frac{1}{n} \sum_{i=1}^{n} x_i$ est le centre de gravité du nuage de points

- **Inertie interclasse** : $I_{inter} = \sum_{k=1}^{K} |\mathcal{C}_k| \times d(m_k, c)^2$

  où $m_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} x_i$ est le centre de gravité de la classe $\mathcal{C}_k$

  $\Rrightarrow$ variance des centres des classes

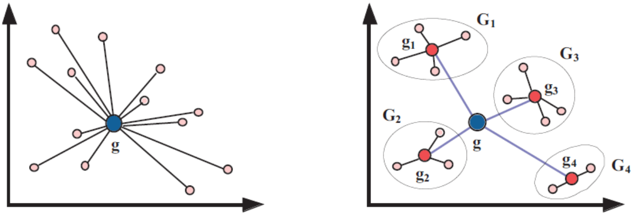
- **Inertie intra-classe** : $I_{intra} = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} d(x_i, m_k)^2$

  $\Rrightarrow$ variance des points d'une même classe

Propriété de Huygens

Propriété de Huygens

$$I_T = I_{inter} + I_{intra}$$



Bisson (2001)

Objectif : minimiser l'inertie intra-classe ($\Leftrightarrow$ maximiser l'inertie inter-classe)