(Dis)similarités, distances et inerties

Cathy Maugis-Rabusseau

4modIA / INSA Toulouse & ENSEEIHT

2022-2023

Objectif

Données : On observe n individus décrits par p variables

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

• 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

⇒ besoin pour certains types de méthodes d'une notion de (dis)similarité entre individus et d'une mesure de séparabilité des classes.

Plan

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

2 Inerties

Plan

- 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 $s: \mathcal{X} \times \mathcal{X} \rightarrow [0,1]$ 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 $d(x_i,x_\ell)=\|x_i-x_\ell\|$

Distance ultramétrique

Une distance d est dite ${\bf ultram\'etrique}$ si elle satisfait l'inégalité ${\bf ultratriangulaire}$

$$\forall (x_i, x_\ell, x_m) \in \mathcal{X}^3, \ d(x_i, x_m) \leq \max \left[d(x_i, x_\ell), d(x_\ell, x_m) \right]$$

Plan

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

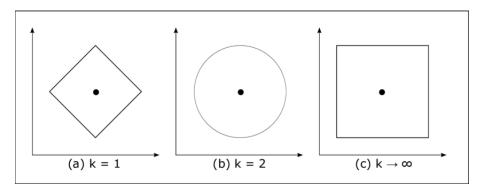
Norme L_q et norme infinie

- $x_i \in \mathcal{X} = \mathbb{R}^p$ pour tout $i = 1, \dots, n$
- 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) $d(x_i,x_\ell) = \|x_i x_\ell\|_1 = \sum_{i=1}^p |x_{ij} x_{\ell j}|$
- Norme infinie $d(x_i, x_\ell) = \max_{j=1,...,p} |x_{ij} x_{\ell j}|$
- ⇒ invariantes par translation mais sensibles à l'échelle des variables.

Norme L_q et norme infinie



Définitions

Moyennes par variable

$$\forall j \in \{1,\ldots,p\}, \ m_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

• Déviation absolue moyenne

$$\forall j \in \{1, \dots, p\}, \ s_j = \frac{1}{n} \sum_{i=1}^n |x_{ij} - m_j|$$

• z-score :

$$\forall j \in \{1,\ldots,p\}, \forall i \in \{1,\ldots,n\} \ z_{ij} = \frac{x_{ij} - m_j}{s_j}$$

Covariances

$$\forall (j,q) \in \{1,\ldots,p\}^2, \ \Sigma_{jq} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - m_j)(x_{iq} - m_q)$$

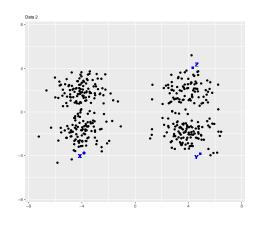
Formes quadratiques

Distances définies comme des formes quadratiques

$$d^2(x_i,x_\ell) = (x_i - x_\ell)' M(x_i - x_\ell)$$

- Norme euclidienne usuelle : $M = I_p$
- $M = \operatorname{diag}\left(\frac{1}{\sigma_1^2}, \dots, \frac{1}{\sigma_p^2}\right)$ où $\sigma_j^2 = \Sigma_{jj}$
- $M = \operatorname{diag}\left(\frac{1}{s_1^2}, \dots, \frac{1}{s_p^2}\right)$
- Distance de Mahalanobis : $M = \Sigma^{-1}$

Exemples



Distance euclidienne ||.||2

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 0.00 8.80 15.99 Y 8.80 0.00 8.46 Z 15.99 8.46 0.00

Oistance de Mahalanobis $\|.\|_{2}$

$$\Sigma = \left(\begin{array}{cc} 16.71 & -0.53 \\ -0.53 & 4.6 \end{array} \right)$$

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

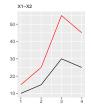
Basées sur le coefficient de corrélation

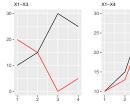
- Coefficient de corrélation de Pearson $\rho(x_i, x_\ell) \in [-1, 1]$
- Exemples de dissimilarités basées sur la corrélation

$$d(x_i, x_\ell) = 1 - \rho(x_i, x_\ell)$$

$$d(x_i, x_\ell) = 1 - |\rho(x_i, x_\ell)|$$

$$d(x_i, x_\ell) = 1 - \rho(x_i, x_\ell)^2$$





	X_2	X_3	X_4
$1 - \rho(X_1, .)$	0	2	0.02
$1 - \rho(X_1, .) $	0	0	0
$1 - \rho(X_1, .)^2$	0	0	0.04
$ X_1 _2$	33.9	37 41	3 74

Plan

- 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 $x_\ell \in \{0,1\}^p$:

$$\begin{array}{c|cccc} & 1 & 0 \\ \hline 1 & m_{11} & m_{10} \\ 0 & m_{01} & m_{00} \end{array}$$

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

Appariement simple
$$s(x_i, x_\ell) = \frac{m_{11} + m_{00}}{m_{11} + m_{00} + m_{01} + m_{00}}$$

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	М	Υ	N	Р	N	N	N
Marie	F	Υ	N	Р	N	Р	N
Pierre	M	Υ	Р	N	N	N	N
Anna	F	N	Р	N	Р	N	N

Jaccard :

• Appariement simple :

	Jules	Marie	Pierre	Anna		Jules	Marie	Pierre	Anna
Jules	1.0	0.5	0.50	0.00	Jules	1.00	0.71	0.71	0.29
Marie	0.5	1.0	0.20	0.00	Marie	0.71	1.00	0.43	0.29
Pierre	0.5	0.2	1.00	0.25	Pierre	0.71	0.43	1.00	0.57
Anna	0.0	0.0	0.25	1.00	Anna	0.29	0.29	0.57	1.00

Pour vdes ariables 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 $u = \text{nombre de variables où } x_i \text{ et } x_\ell \text{ ont la même modalité}$

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 $n \times \tilde{p}$:

- + distance pour variables binaires
 - Distance du χ^2 entre individus :

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

• Distance pour données quantitatives sur les coordonnées de l'ACM

Exemple des races de chien

	TAI	POI	VEL	INT	AFF	AGR
${\tt beauceron}$	3	2	3	2	2	2
ber_allem	3	2	3	3	2	2
caniche	1	1	2	3	2	1
teckel	1	1	1	2	2	1
epagn_bre	2	2	2	3	2	1
labrador	2	2	2	2	2	1

Appariement simple

	beauceron	ber_allem	caniche	teckel	epagn_bre	labrador
beauceron	1.00	0.83	0.17	0.33	0.33	0.50
ber_allem	0.83	1.00	0.33	0.17	0.50	0.33
caniche	0.17	0.33	1.00	0.67	0.67	0.50
teckel	0.33	0.17	0.67	1.00	0.33	0.50
epagn_bre	0.33	0.50	0.67	0.33	1.00	0.83
labrador	0.50	0.33	0.50	0.50	0.83	1.00

Exemple des races de chien

• Distance du χ^2 sur Z :

	beauceron	ber_allem	caniche	teckel	epagn_bre	labrador
beauceron	0.00	0.94	3.91	2.87	3.49	2.55
ber_allem	0.94	0.00	2.97	3.81	2.55	3.49
caniche	3.91	2.97	0.00	1.81	1.96	2.90
teckel	2.87	3.81	1.81	0.00	3.77	2.83
epagn_bre	3.49	2.55	1.96	3.77	0.00	0.94
labrador	2.55	3.49	2.90	2.83	0.94	0.00

Distance euclidienne sur les coefficients de l'ACM

	beauceron	ber_allem	caniche	teckel	epagn_bre	labrador
beauceron	0.00	0.98	1.70	1.70	1.38	1.25
ber_allem	0.98	0.00	1.46	2.04	1.25	1.77
caniche	1.70	1.46	0.00	1.21	1.23	1.70
teckel	1.70	2.04	1.21	0.00	1.90	1.72
epagn_bre	1.38	1.25	1.23	1.90	0.00	0.98
labrador	1.25	1.77	1.70	1.72	0.98	0.00

Plan

- 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

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

avec

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

et

$$d_{i\ell}^{(j)} = \left\{ \begin{array}{ll} \mathbbm{1}_{x_{ij} \neq x_{\ell j}} & \text{si } j \text{ variable binaire ou nominale} \\ \frac{|x_{ij} - x_{\ell j}|}{\max\limits_{1 \leq h \leq n} x_{hj} - \min\limits_{1 \leq h \leq n} x_{hj}} & \text{si } j \text{ est quantitative} \end{array} \right.$$

Exemple des données de maladie du coeur

```
library(VarSelLCM)
data(heart)
library(cluster)
DisGower<-daisy(heart, "gower")</pre>
as.matrix(DisGower)[1:5,1:5]
1 0.0000000 0.4727564 0.4393063 0.4083927 0.4915600
2 0.4727564 0.0000000 0.4715429 0.4840156 0.4741715
```

- 3 0.4393063 0.4715429 0.0000000 0.4202265 0.5048338 4 0.4083927 0.4840156 0.4202265 0.0000000 0.4168955
- 5 0.4915600 0.4741715 0.5048338 0.4168955 0.0000000

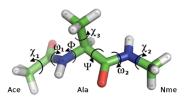
Et aussi ...

- Données compositionnelles (ex : en microbiome)
 - $x_i = (x_{i1}, \dots, x_{ip}) \in [0, 1]^p, \ \sum_{j=1}^p x_{ij} = 1$
 - Distance d'Aitchison

$$d(x_i, x_\ell) = \|\mathrm{CLR}(x_i) - \mathrm{CLR}(x_\ell)\|_2$$

avec $CLR(x_i) = (ln(\frac{x_{ij}}{g(x_i)}))_j$ et g(.) moyenne géométrique

- Données angulaires (ex : conformations de molécule)
 - $x_i = (x_{i1}, \dots, x_{ip}) \in]-\pi, \pi]^p$
 - $d(x_i, x_\ell) = \sum_{j=1}^{p} \left[\pi ||x_{ij} x_{\ell j}| \pi| \right]$
 - $d(x_i, x_\ell) = \sum_{j=1}^{p} 2(1 \cos(x_{ij} x_{\ell j}))$
 - ▶ ..



Quelques commandes

• Avec R

- daisy() [library(cluster)] : metric= "euclidian", "manhattan", "gower"
- Dans library(ade4) : dist.binary(), dist.quant(),
 dist.ktab()



- Avec Python
 - ▶ dist.euclidean [Numpy]
 - scipy.spatial.distance [SciPy]

Conclusion

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

Plan

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

2 Inerties

Inerties intra- / inter- classes

Définitions

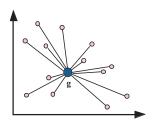
Soit d une distance euclidienne entre individus. Soit $\mathcal{P} = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ une partition des individus en K classes.

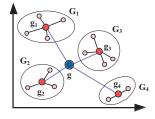
- Inertie totale : $I_T = \sum_{i=1}^n d(x_i, c)^2$ où $c = \frac{1}{n} \sum_{i=1}^n x_i$ est le centre de gravité du nuage de points
- Inertie interclasse : $I_{inter} = \sum_{k=1}^{K} |\mathcal{C}_k| \times d(m_k, c)^2$ où $m_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} x_i$ est le centre de gravité de la classe \mathcal{C}_k \Rightarrow variance des centres des classes
- Inertie intra-classe : $I_{intra} = \sum_{k=1}^{K} \sum_{i \in C_k} d(x_i, m_k)^2$ \Rightarrow variance des points d'une même classe

Propriété de Huygens

Propriété de Huygens

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





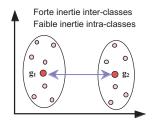
Bisson (2001)

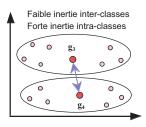
Objectif

on veut minimiser l'inertie intra-classe



maximiser l'inertie inter-classe





Bisson (2001)

References I

- [1] John Aitchison. "The statistical analysis of compositional data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2 (1982), pp. 139–160.
- [2] G. Bisson. "Catégorisation et textes". Atelier Applications, Apprentissage et Acquisition de Connaissances à partir de Textes électroniques (A3CTE), University Paris1. 2001.
- [3] John C Gower. "A general coefficient of similarity and some of its properties". In: *Biometrics* (1971), pp. 857–871.
- [4] Leonard Kaufman and Peter J Rousseeuw. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, 2009.
- [5] Jean Thioulouse and Stéphane Dray. "Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages". In: *Journal of Statistical Software* 22 (2007), pp. 1–14.