Comparing objects

K. Gibert

Ascendant hierarchical clustering

Hot points :

- Agregation criteria
  - Centroid criterion
  - Ward's criterion
- Distance between individuals
  - Only quantitative variables ⟹ Euclidean
  - Only qualitative variables ⟹ χ^2 (Benzécri 80)
  - Heterogenous variables ⟹ Gower (71)

    (compatibility measures)

    Gowda i Diday (91)

    Gibert's Mixed (91)

    Ichino i Yaguchi (94)

    Ralambondrainy (95)

Management of heterogeneous data matrices

Heterogeneous matrices faced. Several approaches [Anderberg 73]:

- **Variables partitioning**
- **Variables converting**
- **Compatibility measures**

Mixed metrics required

Gower [Gow 71]

$$s(i,i') = \frac{\sum_{k=1}^{K} w_k(i,i')\, s_k(i,i')}{\sum_{k=1}^{K} w_k(i,i')}$$

$$w_k(i,i') = \begin{cases} 0 & \text{if } (x_{ik} = missing) \text{ or } (x_{i'k} = \text{missing}) \\ 0 & \text{if } (X_k \text{ binary}) \text{ and } (x_{ik} = \text{false}) \text{ i } (x_{i'k} = \text{false}) \\ & \text{and negative absence of } X_k \text{ excluded} \\ 1 & \text{otherwise} \end{cases}$$

$$s_k(i,i') = \begin{cases} 1 - \frac{|\mathrm{x}_{ik} - \mathrm{x}_{i'k}|}{R_k} & \text{if } \mathrm{X}_k \text{ numerical} \\ 1 & \text{if } (\mathrm{X}_k \text{ qualitative}) \text{ and } (\mathrm{x}_{ik} = \mathrm{x}_{i'k}) \\ 0 & \text{if } (\mathrm{X}_k \text{ qualitative}) \text{ and } (\mathrm{x}_{ik} \neq \mathrm{x}_{i'k}) \end{cases}$$

Euclidean

$$d^2(i,i') = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Chi-2 distance

$$X^2 = \sum_{k=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{kj} - {}^{n_k n_j}\!/_{n}\right)^2}{{}^{n_k n_j}\!/_{n}}$$

Jacquard index (binary variables)

$$d(i,i') = \frac{n_{11} + n_{00}}{n}$$

Gowda-Diday [Gow 91]

$$D(i,i') = \sum_{k=1}^{K} D_k(i_k, i'_k) \quad \textbf{with} \quad D_k(i,i') = D_p(i,i') + D_s(i,i') + D_c(i,i')$$

Component Position

$$D_{kp}(i,i') = \begin{cases} \dfrac{|\,\mathrm{x_{ik}} - \mathrm{x_{i'k}}\,|}{R_k} & \text{if (} \mathrm{X}_k \text{ numerical) and } R_k \text{ is rang of } \mathrm{X}_k \\ 0 & \text{if (} \mathrm{X}_k \text{ qualitative)} \end{cases}$$

Component Span

$$D_{ks}(i,i') = \begin{cases} 0 & \text{if } \mathrm{X}_k \text{ numerical} \\ 0 & \text{if } \mathrm{X}_k \text{ qualitative and not nultivalued} \end{cases}$$

Component Content

$$D_{kc}(i,i') = \begin{cases} 0 & \text{if } \mathrm{X}_k \text{ numerical} \\ 0 & \text{if (} \mathrm{X}_k \text{ qualitative) and } (\mathrm{x_{ik}} = \mathrm{x_{i'k}}) \\ 1 & \text{if (} \mathrm{X}_k \text{ qualitative) and } (\mathrm{x_{ik}} \neq \mathrm{x_{i'k}}) \end{cases}$$

Mixed Metrics [Gib 91]

$$d^2_{(\alpha,\beta)}(i, i') = \alpha\, d_\varsigma{}^2(i, i') + \beta\, d_Q{}^2(i, i')$$

$$d_\varsigma^2(i,i') = \sum_{\forall k \in \varsigma} \frac{(x_{ik} - x_{i'k})^2}{s_k^2}$$

$$d_k^2(i,i') = \begin{cases} 0, & \text{if } x_{ik} = x_{i'k} \\ \dfrac{1}{I_{k^i}} + \dfrac{1}{I_{k^{i'}}}, & \text{otherwise, for compact } i \text{ and } i' \text{ with respect } X_k \\ \dfrac{(f_i^{k_s} - 1)^2}{I^{k_s}} + \displaystyle\sum_{j \neq s}^{n_k} \frac{(f_i^{k_j})^2}{I^{k_j}}, & \text{if } x_{ik} = c_s^k \text{, and extended } i' \text{ with respect } X_k \\ \displaystyle\sum_{j=1}^{n_k} \frac{(f_i^{k_j} - f_{i'}^{k_j})^2}{I^{k_j}}, & \text{for } i, i' \text{ extended with respect } X_k \end{cases}$$

- **Proposal [Gib 91]:**

$$\alpha = \frac{n_\zeta}{d_\zeta^2 max^*} \qquad \beta = \frac{n_Q}{d_Q^2 max^*}$$

Ralambondrainy [Ral95]

$$d^2(i,i') = \pi_1 d^2_{1/\sigma^2}(i,i') + \pi_2 d^2_{\chi^2}(i,i')$$

- Proposal [Ral 88] :
  - Standardisation by the inertia

$$\pi_1 = \frac{1}{Card(\zeta)} \qquad \pi_2 = \frac{1}{n_k - 1}$$

  - Standardisation by the norm

$$\pi_1 = \frac{1}{\sqrt{\sum\{\rho^2(X_k, X_{k'})/k, k' \in \zeta\}}} \qquad \pi_2 = \sqrt{n_k - 1}$$

Ichino-Yaguchi [Ichi 94]

Generalized Minkowski metrics, p-order (p >= 1)

$$d_p(i,i') = \sqrt[p]{\sum_{k=1}^{K} \left(\frac{\phi(x_{ik}, x_{i'k})}{|U_k|} \right)^p}$$

where $|U_k|$ **Normalizes (with *R_k* o *n_k*)**

$\phi(x_{ik}, x_{i'k})$ **Is a function of :**

- **Cartesian Joint**

$$|x_{ik} \oplus x_{i'k}| = \begin{cases} |x_{ik} - x_{i'k}|, & \text{if } X_k \text{ numerical} \\ 1, & \text{if } X_k \text{ categorical and } x_{ik} = x_{i'k} \\ 2, & \text{if } X_k \text{ categorical and } x_{ik} \neq x_{i'k} \end{cases}$$

- **Catesian Meet**

$$|x_{ik} \oplus x_{i'k}| = \begin{cases} |x_{ik} - x_{i'k}|, & \text{if } X_k \text{ numerical} \\ 1, & \text{if } X_k \text{ categorical and } x_{ik} = x_{i'k} \\ 2, & \text{if } X_k \text{ categorical and } x_{ik} \neq x_{i'k} \end{cases}$$

- **Cardinality of x, $|x_{ik}|$, (0 o 1)**
- **Factor** $\gamma \in [0,0.5]$

Example data Michalski example

	A	B	C	F	J	M	P	R	S	T
A	0	1.2488811	1.2438452	0.8309088	1.1683081	1.2438452	0.8309088	0.63954794	1.4049911	0.90644586
B	1.2488811	0	0.8309088	1.2438452	0.90644586	1.4352059	0.63954794	0.8309088	0.8309088	1.1683081
C	1.2438452	0.8309088	0	1.3999553	1.1330574	1.0474486	1.6920323	1.0474486	1.8229634	0.7201209
F	0.8309088	1.2438452	1.3999553	0	0.7201209	1.2388093	1.5006716	1.2388093	1.2488811	1.1330574
J	1.1683081	0.90644586	1.1330574	0.7201209	0	0.97191143	1.1632721	1.5762087	1.2942033	1.2488811
M	1.2438452	1.4352059	1.0474486	1.2388093	0.97191143	0	1.2488811	1.2085946	2.2661147	1.1632721
P	0.8309088	0.63954794	1.6920323	1.5006716	1.1632721	1.2488811	0	1.2488811	0.63451207	0.97191143
R	0.63954794	0.8309088	1.0474486	1.2388093	1.5762087	1.2085946	1.2488811	0	2.2661147	1.7675694
S	1.4049911	0.8309088	1.8229634	1.2488811	1.2942033	2.2661147	0.63451207	2.2661147	0	0.7201209
T	0.90644586	1.1683081	0.7201209	1.1330574	1.2488811	1.1632721	0.97191143	1.7675694	0.7201209	0

Ralambondrainy normalized by inertia pi1=0.5, pi2=0.2

	A	B	C	F	J	M	P	R	S	T
A	0	1.2560605	0.84210527	1.7314991	0.8254386	0.84210527	1.7314991	0.70877194	4.7996807	1.7481657
B	1.2560605	0	1.7314991	0.84210527	1.7481657	1.8648324	0.70877194	1.7314991	1.7314991	0.8254386
C	0.84210527	1.7314991	0	1.2893938	0.25	2.5017545	3.6244814	2.5017545	4.3242416	1.1393938
F	1.7314991	0.84210527	1.2893938	0	1.1393938	3.5244815	2.6017544	3.5244815	1.2560605	0.25
J	0.8254386	1.7481657	0.25	1.1393938	0	2.4850879	3.507815	2.618421	4.2075753	1.2560605
M	0.84210527	1.8648324	2.5017545	3.5244815	2.4850879	0	1.2560605	0.26666668	6.692663	3.507815
P	1.7314991	0.70877194	3.6244814	2.6017544	3.507815	1.2560605	0	1.2560605	3.391148	2.4850879
R	0.70877194	1.7314991	2.5017545	3.5244815	2.618421	0.26666668	1.2560605	0	6.692663	3.641148
S	4.7996807	1.7314991	4.3242416	1.2560605	4.2075753	6.692663	3.391148	6.692663	0	1.1393938
T	1.7481657	0.8254386	1.1393938	0.25	1.2560605	3.507815	2.4850879	3.641148	1.1393938	0

Ralambondrainy normalized by norm, pi1= 0,617, p2=2,236

	A	B	C	F	J	M	P	R	S	T
A	0	3.8718193	3.5263379	3.298699	3.3399992	3.5263379	3.298699	2.035626	7.0879183	3.4850378
B	3.8718193	0	3.298699	3.5263379	3.4850378	4.7894106	2.035626	3.298699	3.298699	3.3399992
C	3.5263379	3.298699	0	4.2444973	2.795085	4.415724	6.796831	4.415724	7.6610384	2.5674462
F	3.298699	3.5263379	4.2444973	0	2.5674462	5.678797	5.533758	5.678797	3.8718193	2.795085
J	3.3399992	3.4850378	2.795085	2.5674462	0	4.229385	5.492458	5.7200966	6.356665	3.8718193
M	3.5263379	4.7894106	4.415724	5.678797	4.229385	0	3.8718193	2.981424	10.58605	5.492458
P	3.298699	2.035626	6.796831	5.533758	5.492458	3.8718193	0	3.8718193	4.1880846	4.229385
R	2.035626	3.298699	4.415724	5.678797	5.7200966	2.981424	3.8718193	0	10.58605	6.9831696
S	7.0879183	3.298699	7.6610384	3.8718193	6.356665	10.58605	4.1880846	10.58605	0	2.5674462
T	3.4850378	3.3399992	2.5674462	2.795085	3.8718193	5.492458	4.229385	6.9831696	2.5674462	0

Gower

	A	B	C	F	J	M	P	R	S	T
A	0	0.625	0.625	0.5	0.625	0.625	0.5	0.375	0.625	0.5
B	0.625	0	0.5	0.625	0.5	0.75	0.375	0.5	0.5	0.625
C	0.625	0.5	0	0.625	0.5	0.5	0.875	0.5	0.75	0.375
F	0.5	0.625	0.625	0	0.375	0.625	0.75	0.625	0.625	0.5
J	0.625	0.5	0.5	0.375	0	0.5	0.625	0.75	0.5	0.625
M	0.625	0.75	0.5	0.625	0.5	0	0.625	0.5	1.0	0.625
P	0.5	0.375	0.875	0.75	0.625	0.625	0	0.625	0.375	0.5
R	0.375	0.5	0.5	0.625	0.75	0.5	0.625	0	1.0	0.875
S	0.625	0.5	0.75	0.625	0.5	1.0	0.375	1.0	0	0.375
T	0.5	0.625	0.375	0.5	0.625	0.625	0.5	0.875	0.375	0