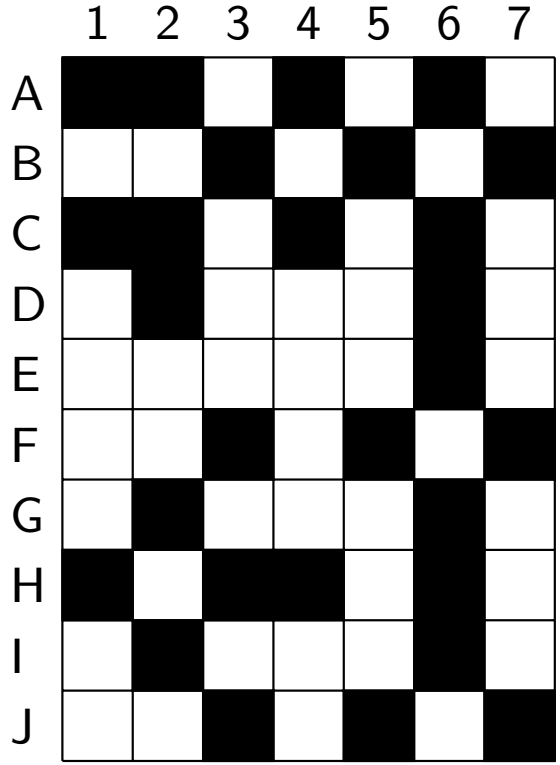
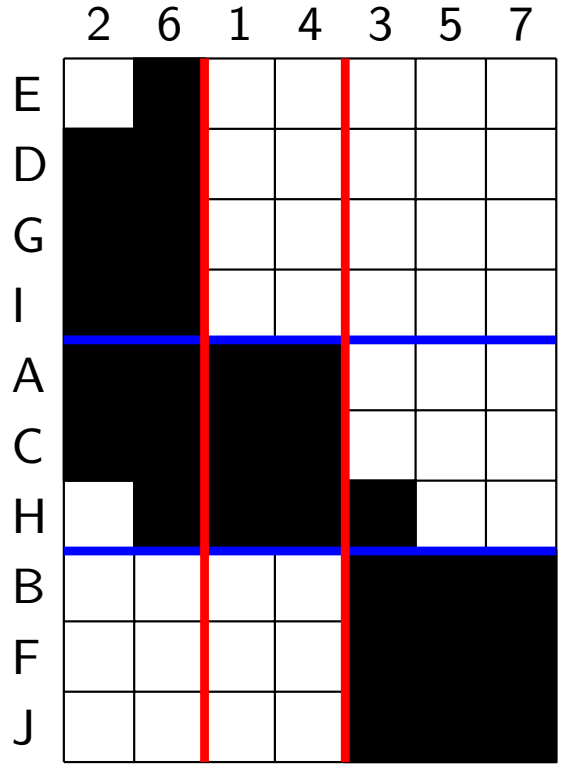
Co-clustering indices

Valérie Robert & Yann Vasseur & Vincent Brault

Working Group on Model-Based Clustering
26 Octobre 2021

Co-clustering Context





Desired properties of a co-clustering index:

1. proportional to the number of miss-classified cells.
2. symmetrical regarding two co-partitions.
3. varying between 0 and 1, where 1 means a perfect match and 0 leads to all worst scenarios, including independence.
4. be independent from label switching: if two co-partitions are equal when the labels are swaped, the score should remain unchanged.
5. computable within a reasonable time.

Proposed co-Clustering indices

- **Co-clustering Adjusted Rand Index (CARI)**

$$\mathrm{CARI}((\boldsymbol{z},\boldsymbol{w}),(\boldsymbol{z'},\boldsymbol{w'})) = \frac{\sum_{p,q}\binom{n_{p,q}^{zwz'w'}}{2} - \sum_{p}\binom{n_{p,\bullet}^{zwz'w'}}{2}\sum_{q}\binom{n_{\bullet,q}^{zwz'w'}}{2}\Big/\binom{I\times J}{2}}{\frac{1}{2}\left[\sum_{p}\binom{n_{p,\bullet}^{zwz'w'}}{2} + \sum_{q}\binom{n_{\bullet,q}^{zwz'w'}}{2}\right] - \left[\sum_{p}\binom{n_{p,\bullet}^{zwz'w'}}{2}\sum_{q}\binom{n_{\bullet,q}^{zwz'w'}}{2}\right]\Big/\binom{I\times J}{2}}.$$

- **Normalized Classification Error (NCE) inspired of CE (Lomet et al.)**

$$\mathrm{dist}_{(I,H)\times(J,L)}((\boldsymbol{z},\boldsymbol{w}),(\boldsymbol{z'},\boldsymbol{w'})) = \min_{\sigma\in\mathfrak{S}(\{1,\ldots,H\})}\min_{\tau\in\mathfrak{S}(\{1,\ldots,L\})}\left(1 - \frac{1}{I\times J}\sum_{i,j,h,\ell} z_{ih}z'_{i\sigma(h)}w_{j\ell}w'_{j\tau(\ell)}\right),$$

$$\mathrm{NCE}((\boldsymbol{z},\boldsymbol{w}),(\boldsymbol{z'},\boldsymbol{w'})) = 1 - \frac{\mathrm{dist}_{(I,H)\times(J,L)}((\boldsymbol{z},\boldsymbol{w}),(\boldsymbol{z'},\boldsymbol{w'}))}{\frac{1}{H}+\frac{1}{L}-\frac{1}{HL}}$$

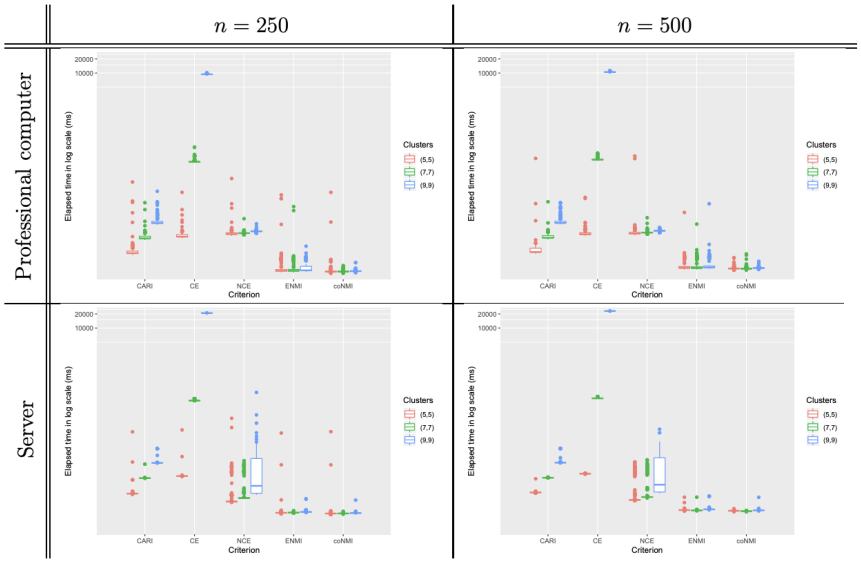
- **Co-clustering Normalized Mutual Information (CoNMI) inspired of ENMI (Wyse et al.)**

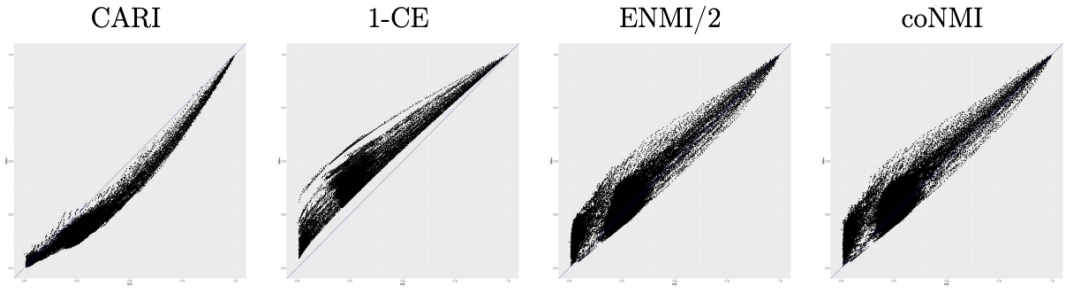
$$\mathrm{coNMI}((\boldsymbol{z},\boldsymbol{w});(\boldsymbol{z'},\boldsymbol{w'})) = \frac{\mathrm{coMI}((\boldsymbol{z},\boldsymbol{w});(\boldsymbol{z'},\boldsymbol{w'}))}{\max(\mathcal{H}(\boldsymbol{z},\boldsymbol{w}),\mathcal{H}(\boldsymbol{z'},\boldsymbol{w'}))}$$

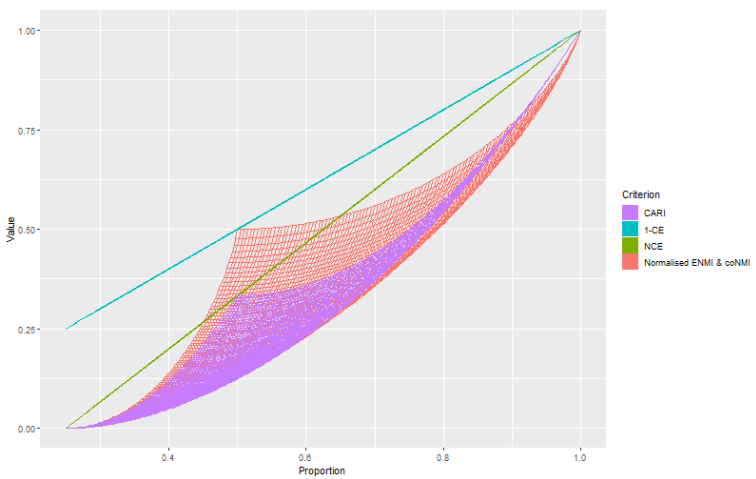
with

$$\mathrm{coMI}((\boldsymbol{z},\boldsymbol{w});(\boldsymbol{z'},\boldsymbol{w'})) = \sum_{p,q}\frac{n_{p,q}^{zwz'w'}}{IJ}\log\left(\frac{n_{p,q}^{zwz'w'}IJ}{n_{p,\bullet}^{zwz'w'}n_{\bullet,q}^{zwz'w'}}\right)$$

Simulations to compare co-clustering indices







	CARI	CE (Lomet, 2012)	NCE	ENMI (Wyse et al., 2017)	coNMI
1. Proportion of cells misclassified	Moderately	✓	✓	✗	✗
2. Symmetry	✓	✓	✓	✓	✓
3.1 Maximum limit equal to 1	✓	If we use $1 - CE$	✓	✓	✓
3.2 Minimum limit equal to 0	Asympto-tically	✗	✓	✓	✓
4. Label switching	✓	✓	✓	✓	✓
5. Execution time	Reasonable	Impossible as soon as $\max(H, L) > 9$	Reasonable	✓	✓

Robert, V., Vasseur, Y., & Brault, V. (2021). Comparing high-dimensional partitions with the Co-clustering Adjusted Rand Index. Journal of Classification, 38(1), 158-186.