2. Clustering validity indices

This section explains the internal and external indices that were implemented and used in this work. The indices are presented according to the type of analysis these indices perform. Fig. 1 shows the organization of the indices. Before introducing the each group of indices, preliminary definitions, which some indices have in common, are listed. The indices have in their name the rule to choice a specific partition, i.e.:

- [↑] indicates that partitions with the greatest value of the index are desired.
- [↓] indicates that partitions with the lowest value of the index are desired.
- [diff ↑] indicates that is a difference-like index that grows as long as the number of clusters increase.
- [diff ↓] indicates that is a difference-like index that decreases as long as the number of clusters increase.

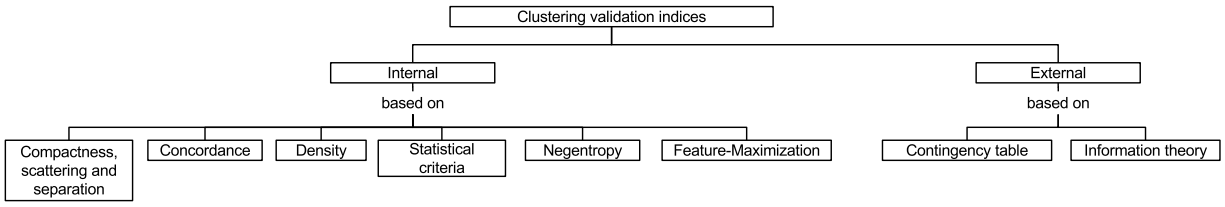

Figure 1: Clasification of internal and external indices according to the analysis these indices perform.

2.1. Internal indices

2.1.1. Indices based on compactness, scattering and separation

- ***Preliminary definitions:***
  - **Data matrix $X_{N\times p}$:** Dataset matrix that contains N observations x_i of p features.
  - **Obtained partition P_o:** Clustering of K clusters resulted from the clustering algorithm.
  - Each cluster c_k can be represented by a subset $X^{\{k\}}_{n_k\times p}$, which contains the n_k observations belonging to this cluster.
  - **Barycenter μ^k of c_k**:

$$\mu^k = \frac{1}{n_k} \sum_{x_i \in X^{\{k\}}} x_i \tag{1}$$

  - **Data barycenter μ**:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2}$$

– **Total dispersion**: The scatter matrix $\boldsymbol{T}$ is defined by:

$$\boldsymbol{T} = (\boldsymbol{X} - \boldsymbol{\mu}_n)'(\boldsymbol{X} - \boldsymbol{\mu}_n) \tag{3}$$

Where $\boldsymbol{\mu}_n$ a $n \times p$ matrix generated repeating the row vector $\boldsymbol{\mu}$ n times along the rows.

The total dispersion or total sum of squares (TSS) is the trace of the matrix $\boldsymbol{T}$.

– **Within-group dispersion**: The within-group scatter matrix $\boldsymbol{WG^{\{k\}}}$ for the group k is defined by:

$$\boldsymbol{WG^{\{k\}}} = (\boldsymbol{X^{\{k\}}} - \boldsymbol{\mu}^{\boldsymbol{k}}_{\boldsymbol{n_k}})'(\boldsymbol{X^{\{k\}}} - \boldsymbol{\mu}^{\boldsymbol{k}}_{\boldsymbol{n_k}}) \tag{4}$$

Where $\boldsymbol{\mu}^{\boldsymbol{k}}_{\boldsymbol{n_k}}$ is a $n_k \times p$ matrix generated repeating the row vector $\boldsymbol{\mu^k}$, n_k times along the rows.

The within-cluster dispersion or within-group sum of squares (WGSS) is the sum of the traces of matrices $\boldsymbol{WG^{\{k\}}}$.

– **Between-group dispersion**: This metric measures the dispersion of the clusters between each other. The between-group scatter matrix is defined by:

$$\boldsymbol{BG = B'B} \tag{5}$$

Where $\boldsymbol{B}$ is the matrix formed by $\left(\boldsymbol{\mu^k} - \boldsymbol{\mu}\right)$ for each cluster in each row, generating a $K \times p$ matrix.

– WGSS, BGSS and TSS can be expresed as a sum of square distances:

$$\text{WGSS} = \sum_{k=1}^{K} \text{WGSS}^{\{k\}} = \sum_{k=1}^{K} \sum_{\boldsymbol{x_i} \in \boldsymbol{X^{\{k\}}}} \|\boldsymbol{x_i} - \boldsymbol{\mu^k}\|^2 \qquad \text{TSS} = \sum_{i=1}^{n} \|\boldsymbol{x_i} - \boldsymbol{\mu}\|^2 \qquad \text{BGSS} = \sum_{k=1}^{K} n_k \|\boldsymbol{\mu^k} - \boldsymbol{\mu}\|^2$$

Where $\|\cdot\|$ represents the Euclidean distance.

– **Columns vectors $\boldsymbol{v_j}$:** These vector are obtained extracting from $\boldsymbol{X_{n \times p}}$ the p columns vectors.

$$\boldsymbol{X_{n \times p}} = \left[\boldsymbol{v_1}\ \boldsymbol{v_2}\ \ldots\ \boldsymbol{v_p}\right]_{n \times p} \tag{6}$$

– **Columns vectors $\boldsymbol{v_j}$ per cluster k:** These vector are obtained extracting from $\boldsymbol{X^{\{k\}}_{n_k \times p}}$ the p columns vectors.

$$\boldsymbol{X^{\{k\}}_{n_k \times p}} = \left[\boldsymbol{v^{\{k\}}_1}\ \boldsymbol{v^{\{k\}}_2}\ \ldots \boldsymbol{v^{\{k\}}_p}\right]_{n_k \times p} \tag{7}$$

– **Number of pairs of points n_T, n_W and n_B:** In total there are n_T distinct pairs of observations in X.

$$n_T = \binom{n}{2} = \frac{n(n-1)}{2} \tag{8}$$

Similarly, if only the observations of the cluster k are considered, there are n_W^k distinct pairs of observations in $\boldsymbol{X}^{\{k\}}$.

$$n_W^{\{k\}} = \binom{n_k}{2} = \frac{n_k(n_k-1)}{2} \tag{9}$$

n_W is obtained adding all the distinct pairs within the clusters:

$$n_W = \sum_{k=1}^{K} n_W^{\{k\}} = \sum_{k=1}^{K} \frac{n_k(n_k-1)}{2} \tag{10}$$

Finally, n_B, the number of distinct pairs between clusters can be calculated by:

$$n_T = n_W + n_B \qquad n_B = n_T - n_W = \sum_{k<k'} n_k n_k' \tag{11}$$

– **Sum of distances s_W and s_B:** The sum of the n_W within-cluster distances s_W can be calculated by:

$$s_W = \sum_{k=1}^{K} \sum_{\substack{\boldsymbol{x_i},\boldsymbol{x_j} \in \boldsymbol{X}^{\{k\}} \\ i<j}} \|\boldsymbol{x_i} - \boldsymbol{x_j}\| \tag{12}$$

Similarly, the sum of the n_B between-cluster distances s_B can be calculated by:

$$s_B = \sum_{k<k'} \sum_{\substack{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}} \\ \boldsymbol{x_j} \in \boldsymbol{X}^{\{k'\}} \\ i<j}} \|\boldsymbol{x_i} - \boldsymbol{x_j}\| \tag{13}$$

– **Point Symmetry-Distance of a point with respect a cluster k:** Based on the distance between the symmetric point of $\boldsymbol{x_i}$ with respect to the centroid of a cluster k (i.e., $2\boldsymbol{\mu^k} - \boldsymbol{x_i}$) and the m-nearest points in the cluster k (Bandyopadhyay and Saha, 2008).

$$d_{ps}(\boldsymbol{x_i}, c_k) = \frac{1}{m} \sum \min_{\boldsymbol{x_j} \in \boldsymbol{X}^{\{k\}}} (m) \left\{ \left\| (2\boldsymbol{\mu^k} - \boldsymbol{x_i}) - \boldsymbol{x_j} \right\| \right\} \tag{14}$$

Where $\sum \min_{\boldsymbol{x_j} \in \boldsymbol{X}^{\{k\}}} (m) \{\cdot\}$ computes the sum of the m lowest values of $\{\cdot\}$.

• ***Indices based on compactness, scattering and separation***

– **Ball-Hall [diff $\uparrow$]:** This index is defined as the average of within-cluster dispersions weighted by the inverse of n_k (Ball and Hall, 1965a):

$$\text{Ball-Hall} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} WGSS^{\{k\}} \tag{15}$$

– **Banfeld-Raftery [$\downarrow$]:** Banfeld-Raftery is defined as the weighted sum of the logarithms of the within-

cluster dispersion (Banfield and Raftery, 1993):

$$\text{Banfeld-Raftery} = \sum_{k=1}^{K} n_k \log\left(\frac{WGSS^{\{k\}}}{n_k}\right) \tag{16}$$

– **C index** [↓]: This measure is based on the within-group distances, as well as on their maximum and minimum possible value (Hubert and Schultz, 1976):

$$\text{C} = \frac{s_W - s_{\min}}{s_{\max} - s_{\min}} \tag{17}$$

Where $s_{\min}$ and $s_{\max}$ are the sum of the n_W smallest and largest distances between all the pairs of points of $\boldsymbol{X_{N\times p}}$.

– **Calinski-Harabasz** [↑]: The idea behind this index is to evaluate partitions based on the average between- and within-cluster sum of squares (Caliński and Harabasz, 1974):

$$\text{Calinski-Harabasz} = \frac{n-K}{K-1}\frac{BGSS}{WGSS} \tag{18}$$

– **Davies-Bouldin** [↓]: Two terms are defined in order to estimate this index: δ_k, the mean distance of the cluster k points to their barycenter $\boldsymbol{\mu^k}$ and $\Delta_{kk'}$, the distance between the barycenters $\boldsymbol{\mu^k}$ and $\boldsymbol{\mu^{k'}}$ (Davies and Bouldin, 1979):

$$\delta_k = \frac{1}{n_k}\sum_{k=1}^{K}\sum_{\boldsymbol{x_i}\in \boldsymbol{X^{\{k\}}}} \|\boldsymbol{x_i} - \boldsymbol{\mu^k}\| \qquad \Delta_{kk'} = \|\boldsymbol{\mu^{k'}} - \boldsymbol{\mu^k}\| \tag{19}$$

$$\text{Davies-Bouldin} = \frac{1}{K}\sum_{k=1}^{K}\max_{k'\neq k}\left(\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}}\right) \tag{20}$$

– **Det-Ratio** [**diff** ↓]: This is based the determinant of the total dispersion matrix and the sum of the within-group dispersion matrices (Scott and Symons, 1971).

$$\text{Det-Ratio} = \frac{\det(\boldsymbol{T})}{\det(\boldsymbol{WG})} \tag{21}$$

– **Dunn's indices** [↑]: These indices are based on two estimators, one focus on separation and the another in cohesion (compactness).

$$\text{Dunn index} = \min_{1\leq k\leq c}\left\{\min_{\substack{1\leq k'\leq c\\ j\neq i}}\left\{\frac{\delta_s\left(\boldsymbol{X^{\{k\}}}, \boldsymbol{X^{\{k'\}}}\right)}{\max_{1\leq k\leq c}\Delta_d(\boldsymbol{X^{\{k\}}})}\right\}\right\} \tag{22}$$

Where δ_s is the separation estimator and Δ_d is the cohesion estimator. In (Bezdek and Pal, 1998) 17

variants of the original Dunn index (Dunn, 1974) were proposed. These variants and the original Dunn index are produced by combining one of the 6 ways to compute δ_s and one of the 3 ways to compute Δ_d:

$$\delta_1\left(\boldsymbol{X}^{\{k\}}, \boldsymbol{X}^{\{k'\}}\right) = \min_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} \left\{ \min_{\boldsymbol{x}_j \in \boldsymbol{X}^{\{k'\}}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \right\} \tag{23}$$

$$\delta_2\left(\boldsymbol{X}^{\{k\}}, \boldsymbol{X}^{\{k'\}}\right) = \max_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}, \boldsymbol{x}_j \in \boldsymbol{X}^{\{k'\}}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \tag{24}$$

$$\delta_3\left(\boldsymbol{X}^{\{k\}}, \boldsymbol{X}^{\{k'\}}\right) = \frac{1}{n_k n_{k'}} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} \sum_{\boldsymbol{x}_j \in \boldsymbol{X}^{\{k'\}}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \tag{25}$$

$$\delta_4\left(\boldsymbol{X}^{\{k\}}, \boldsymbol{X}^{\{k'\}}\right) = \|\boldsymbol{\mu}^k - \boldsymbol{\mu}^{k'}\| \tag{26}$$

$$\delta_5\left(\boldsymbol{X}^{\{k\}}, \boldsymbol{X}^{\{k'\}}\right) = \frac{1}{n_k + n_{k'}} \left(\sum_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} \|\boldsymbol{x}_i - \boldsymbol{\mu}^{k'}\| + \sum_{\boldsymbol{x}_j \in \boldsymbol{X}^{\{k'\}}} \|\boldsymbol{x}_j - \boldsymbol{\mu}^{k}\| \right) \tag{27}$$

$$\delta_6\left(\boldsymbol{X}^{\{k\}}, \boldsymbol{X}^{\{k'\}}\right) = \max\left\{ \max_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} \left\{ \min_{\boldsymbol{x}_j \in \boldsymbol{X}^{\{k'\}}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \right\}, \max_{\boldsymbol{x}_j \in \boldsymbol{X}^{\{k'\}}} \left\{ \min_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \right\} \right\} \tag{28}$$

$$\Delta_1\left(\boldsymbol{X}^{\{k\}}\right) = \max_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} \left\{ \max_{\boldsymbol{x}_j \in \boldsymbol{X}^{\{k\}}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \right\} \tag{29}$$

$$\Delta_2\left(\boldsymbol{X}^{\{k\}}\right) = \frac{1}{n_k(n_k - 1)} \sum_{\boldsymbol{x}_i \neq \boldsymbol{x}_j \in \boldsymbol{X}^{\{k\}}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \tag{30}$$

$$\Delta_3\left(\boldsymbol{X}^{\{k\}}\right) = \frac{2}{n_k} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} \|\boldsymbol{x}_i - \boldsymbol{\mu}^{k}\| \tag{31}$$

In this work, we refers a specific combination Δ_i and δ_j by Dunnij. The original Dunn index is denoted by Dunn11.

– **Ksq DetW** [**diff** $\uparrow$]: Index that is also based in within dispersion matrices (Marriott, 1971).

$$\text{Ksq DetW} = K^2 \det(\boldsymbol{WG}) \tag{32}$$

– **Log Det Ratio** [**diff** $\downarrow$]: This index is the logarithm of the Det-Ratio index (Scott and Symons, 1971).

$$\text{Log Det Ratio} = n \log(\text{Det-Ratio}) = n \log\left(\frac{\det(\boldsymbol{T})}{\det(\boldsymbol{WG})}\right) \tag{33}$$

– **Log SS Ratio** [**diff** $\downarrow$]:

$$\text{Log SS Ratio} = \log\left(\frac{\text{BGSS}}{\text{WGSS}}\right) \tag{34}$$

– **McClain-Rao** [$\downarrow$]: This measure is defined as the quotient between the mean within-cluster and between-

cluster distances (McClain and Rao, 1975):

$$\text{McClain-Rao} = \frac{\frac{s_W}{n_W}}{\frac{s_B}{n_B}} = \frac{n_B}{n_W}\frac{s_W}{s_B} \tag{35}$$

– **PBM index** [↑]: Also called I index in literature (Pakhira et al., 2004).

$$\text{PBM} = \left(\frac{1}{K} \times \frac{E_T}{E_W} \times D_B\right) \tag{36}$$

Where D_B is maximum distance between two cluster barycenters. E_W is the sum of the distances of the cluster objects to its barycenter. E_T is the sum of the distances of the objects to the entire data barycenter. Their formulas are:

$$E_T = \sum_{i=1}^{n} \|\boldsymbol{x_i} - \boldsymbol{\mu}\| \qquad E_W = \sum_{k=1}^{K} \sum_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}} \|\boldsymbol{x_i} - \boldsymbol{\mu^k}\| \qquad D_B = \max_{k<k'} \|\boldsymbol{\mu^k} - \boldsymbol{\mu^{k'}}\|$$

– **Point-Biserial** [↑]: Index based on the statistics point-biserial, where the continuous variable A is a set of n_T possible distinct distances of the dataset. The binary variable B equals 1 if the corresponding distance of A is between two points of the same cluster. B equal 0 otherwise. (Milligan, 1981)

$$\text{Point-Biserial} = \left(\frac{s_W}{n_W} - \frac{s_B}{n_B}\right)\frac{\sqrt{n_W n_B}}{n_T} \tag{37}$$

– **Ratkowsky-Lance** [↑]: This index is based on the between-groups and total dispersion matrix (Ratkowsky and Lance, 1978). A term in denominator is included to avoid the monotone increasing as the number of clusters increases.

$$\text{Ratkowsky-Lance} = \sqrt{\frac{\overline{R}}{K}} = \frac{\overline{c}}{\sqrt{K}} \tag{38}$$

$$\overline{c}^2 = \overline{R} = \frac{1}{p}\sum_{j=1}^{p} \frac{BGSS_j}{TSS_j} \tag{39}$$

Where $\text{BGSS}_j = b_{jj}$ and $\text{TSS}_j = t_{jj}$ are the j-element of the diagonal of BGSS and of TSS, respectively.

– **Ray-Turi** [↓]: Ray-Turi is based in the same denominator used the Davies-Bouldin index ($\Delta_{kk'}$, the distance between the barycenters $\boldsymbol{\mu^k}$ and $\boldsymbol{\mu^{k'}}$) (Ray and Turi, 1999).

$$\text{Ray-Turi} = \frac{1}{n}\frac{WGSS}{\min_{k\neq k'} \Delta_{kk'}} \tag{40}$$

– **Scott-Symons** [↓]: This index is the weighted sum of the logarithms of the determinants of the variance-

covariance matrix of each cluster (Scott and Symons, 1971):

$$\text{Scott-Symons} = \sum_{k=1}^{K} n_k \log\det\left(\frac{\boldsymbol{WG^{\{k\}}}}{n_k}\right) \tag{41}$$

– **SD index** [↓]: This index is based on the average scattering and the total separation of clusters (Halkidi et al., 2000).

$$\text{SD} = \alpha S + D \tag{42}$$

$$S = \frac{\frac{1}{K}\sum_{k=1}^{K}\left\|\boldsymbol{v^{\{k\}}}\right\|}{\|\boldsymbol{v}\|} \qquad D = \frac{\max_{k\neq k'}\Delta_{kk'}}{\min_{k\neq k'}\Delta_{kk'}} \times \sum_{k=1}^{K}\frac{1}{\sum_{\substack{k'=1\\k'\neq k}}^{K}\Delta_{kk'}} \tag{43}$$

Where α equals to the value of D obtained for the partition with the greatest number of cluster, and $\boldsymbol{v}$ and $\boldsymbol{v^{\{k\}}}$ are the variances vector of the columns vector $\boldsymbol{v_j}$ and $\boldsymbol{v_j}^{\{k\}}$, respectively:

$$\boldsymbol{v^{\{k\}}} = \left[\text{Var}\left(\boldsymbol{v_1^{\{k\}}}\right), ..., \text{Var}\left(\boldsymbol{v_p^{\{k\}}}\right)\right] \qquad \boldsymbol{v} = \left[\text{Var}\left(\boldsymbol{v_1}\right), ..., \text{Var}(\boldsymbol{v_p})\right] \tag{44}$$

– **Silhouette index** [↑]: Silhouette is based on two terms computed for each sample $\boldsymbol{x_i}$: $a_{\text{SIL}}(i)$ and $b_{\text{SIL}}(i)$. $a_{\text{SIL}}(i)$ represents the mean distance of object i to the other objects in its cluster, while $b_{\text{SIL}}(i)$ is the smallest mean distance of an object i to the other clusters (Ratkowsky and Lance, 1978).

$$\text{Silhouette} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{\boldsymbol{x_i}\in\boldsymbol{X^{\{k\}}}}\frac{b_{\text{SIL}}(i)-a_{\text{SIL}}(i)}{\max\left(a_{\text{SIL}}(i), b_{\text{SIL}}(i)\right)} \tag{45}$$

$$a_{\text{SIL}}(i) = \frac{1}{n_k - 1}\sum_{\substack{\boldsymbol{x_i},\boldsymbol{x_j}\in\boldsymbol{X^{\{k\}}}\\ \boldsymbol{x_i}\neq\boldsymbol{x_j}}}\left\|\boldsymbol{x_i}-\boldsymbol{x_j}\right\| \qquad b_{\text{SIL}}(i) = \min_{k'\neq k}\left\{\frac{1}{n_{k'}}\sum_{\boldsymbol{x_j}\in\boldsymbol{X^{\{k'\}}}}\left\|\boldsymbol{x_i}-\boldsymbol{x_j}\right\|\right\} \tag{46}$$

– **Sym** [↑]**:** Adaptation of the PBM index using the point symmetric distance (Bandyopadhyay and Saha, 2008).

$$\text{Sym} = \frac{\max_{k<k'}\|\boldsymbol{\mu^k}-\boldsymbol{\mu^{k'}}\|}{K\sum_{k=1}^{K}\sum_{\boldsymbol{x_i}\in\boldsymbol{X^{\{k\}}}}d_{ps}(\boldsymbol{x_i}, c_k)} \tag{47}$$

– **Modified Davies–Bouldin** [↓]**:** This index is a modification of the Davies–Bouldin index (Saha and Bandyopadhyay, 2009), which uses the point symmetric distance.

$$\delta_k = \frac{1}{n_k}\sum_{\boldsymbol{x_i}\in\boldsymbol{X^{\{k\}}}}d_{ps}(\boldsymbol{x_i}, c_k) \tag{48}$$

– **Modified Dunn11** [↑]**:** This index is a modification of the Dunn11 (Saha and Bandyopadhyay, 2009),

which uses the point symmetric distance.

$$\Delta\left(\boldsymbol{X}^{\{k\}}\right) = \max_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} d_{ps}(\boldsymbol{x_i}, c_k) \tag{49}$$

– **Modified Dunn33** [↑]**:** This index is a modification of the Dunn33 (Saha and Bandyopadhyay, 2009), which uses the point symmetric distance.

$$\Delta_3\left(\boldsymbol{X}^{\{k\}}\right) = \frac{2}{n_k} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} d_{ps}(\boldsymbol{x_i}, c_k) \tag{50}$$

– **Trace W** [**diff** ↑]: Index previously studied in (Edwards and Cavalli-Sforza, 1965).

$$\text{Trace W} = \text{Tr}(\boldsymbol{WG}) = WGSS \tag{51}$$

– **Trace WiB** [**diff** ↑]: This index was studied in (Friedman and Rubin, 1967).

$$\text{Trace W} = \text{Tr}(\boldsymbol{WG}^{-1}\boldsymbol{BG}) \tag{52}$$

– **Wemmert-Gançarski** [↑]: The Wemmert-Gancarski index is built using proportions of distances between the points and the clusters barycenters (equation 54).

$$\text{Wemmert-Gançarski} = \frac{1}{n}\sum_{k=1}^{K} n_k J_k \tag{53}$$

$$J_k = \max\left\{0,\ 1 - \frac{1}{n_k}\sum_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} R(\boldsymbol{x_i})\right\} \qquad R(\boldsymbol{x_i}) = \frac{\left\|\boldsymbol{x_i} - \boldsymbol{\mu^k}\right\|}{\min_{k' \neq k}\left\|\boldsymbol{x_i} - \boldsymbol{\mu^{k'}}\right\|} \tag{54}$$

– **Xie-Beni index**[↓]: It is defined as the quotient between the mean quadratic error and the minimum of the minimal squared distances between the points in the clusters. (Xie and Beni, 1991)

$$\text{Xie-Beni} = \frac{1}{n}\frac{WGSS}{\min_{k<k'} \delta_1\left(\boldsymbol{X}^{\{k\}}, \boldsymbol{X}^{\{k'\}}\right)} \tag{55}$$

– **CVNN** [↓]**:** This is a more recent index, which has a intercluster separation based on L nearest neighbors (in the original paper was used the letter k instead of L to denote the nearest neighbors) and a compactness measure that does not take into account the barycenter in order to deal with arbitrary-shaped data (Liu

et al., 2013).

$$\text{Sep}(K, v) = \max_k \left\{ \frac{1}{n_k} \sum_{j=1}^{n_k} \frac{q_j}{L} \right\} \qquad \text{Comp}(K, v) = \Delta_2 \left(\boldsymbol{X}^{\{k\}} \right) = \frac{1}{n_k(n_k - 1)} \sum_{\boldsymbol{x_i} \neq \boldsymbol{x_j} \in \boldsymbol{X}^{\{k\}}} ||\boldsymbol{x_i} - \boldsymbol{x_j}|| \tag{56}$$

Where q_j is the number of nearest neighbors which are not in the same cluster k of object j. CVNN is then defined as the sum normalized compatness and separation:

$$\text{CVNN} = \text{Sep}_{\text{norm}}(K, v) + \text{Comp}_{\text{norm}}(K, v) \tag{57}$$

$$\text{Sep}_{\text{norm}}(K, v) = \frac{\text{Sep}(K, v)}{\max_{K_{min} < K' < K_{min}} \text{Sep}(K', v)} \qquad \text{Comp}_{\text{norm}}(K, v) = \frac{\text{Comp}(K, v)}{\max_{K_{min} < K' < K_{min}} \text{Comp}(K', v)} \tag{58}$$

– **WB** [↓]**:** Index proposed by (Zhao et al., 2009a).

$$\text{WB} = K \frac{WGSS}{BGSS} \tag{59}$$

– **Xu** [↓]**:** Index proposed by (Xu, 1997b).

$$\text{Xu} = \log \frac{K \times WGSS^{\frac{p}{2}}}{BGSS^{\frac{p}{2}}} \tag{60}$$

– **CS** [↓]**:** Index whose numerator estimates the cohesion by the cluster diameters. The separation is estimated in the denominator by nearest neighbor distances. (Chou et al., 2004)

$$\text{CS} = \frac{\sum_{k=1}^{K} \frac{1}{n_k} \sum_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}} \max_{\boldsymbol{x_j} \in \boldsymbol{X}^{\{k\}}} ||\boldsymbol{x_i} - \boldsymbol{x_j}||}{\sum_{k=1}^{K} \min_{\boldsymbol{\mu^{k'}}} ||\boldsymbol{\mu^k} - \boldsymbol{\mu^{k'}}||} \tag{61}$$

– **Score function** [↑]**:** Index maily based on a compactness term call wcd and on a separation term called bcd (Saitta et al., 2007). As the term of separation becomes larger than the term of compactness, the upper exponent becomes infinite, like the inferior exponent, resulting in an index equal to 1.

$$\text{SF} = 1 - \frac{1}{\exp^{\exp^{bcd - wcd}}} \tag{62}$$

$$\text{bcd} = \frac{\sum_{k=1}^{K} n_k ||\boldsymbol{\mu^k} - \boldsymbol{\mu}||}{n \times K} \qquad \text{wcd} = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}} ||\boldsymbol{x_i} - \boldsymbol{\mu^k}|| \tag{63}$$

– **COP** [↓]**:** Index whose compactness term is based on the distance from the objects to their cluster, and

whose separation term is the furthest distance between points in different clusters (Gurrutxaga et al., 2010).

$$\text{COP} = \frac{1}{n}\sum_{k=1}^{K} n_k \frac{\text{intra}_{COP}(k)}{\text{inter}_{COP}(k)} \tag{64}$$

$$\text{intra}_{COP}(k) = \frac{1}{n_k}\sum_{\boldsymbol{x_i}\in \boldsymbol{X}^{\{k\}}} \|\boldsymbol{x_i} - \boldsymbol{\mu^k}\| \qquad \text{inter}_{COP}(k) = \min_{\boldsymbol{x_i}\notin \boldsymbol{X}^{\{k\}}} \max_{\boldsymbol{x_j}\in \boldsymbol{X}^{\{k\}}} \|\boldsymbol{x_i} - \boldsymbol{x_j}\| \tag{65}$$

– **SV** [↑]**:** This index estimates the separation criterion in the numerator by finding the closest distance between clusters centroids. Its compactness term is based on the 10 percent of the lowest distances from the cluster objects to their centroid (Žalik and Žalik, 2011). Operators $\sum \min_{\boldsymbol{x_j}\in \boldsymbol{X}^{\{k\}}}(n)$ and $\sum \max_{\boldsymbol{x_j}\in \boldsymbol{X}^{\{k\}}}(n)$ are used to compute the compactness term.

$$\text{SV} = \frac{\sum_{k=1}^{K} \min_{\boldsymbol{\mu}^{k'}} \left\{\|\boldsymbol{\mu^k} - \boldsymbol{\mu}^{k'}\|\right\}}{\sum_{k=1}^{K} \frac{1}{n_k} \sum \max_{\boldsymbol{x_j}\in \boldsymbol{X}^{\{k\}}}(0.1n_k) \left\{\|\boldsymbol{x_i} - \boldsymbol{\mu^k}\|\right\}} \tag{66}$$

Where $\sum \max_{\boldsymbol{x_j}\in \boldsymbol{X}^{\{k\}}}(m)\{\cdot\}$ computes the sum of the m greatest values of $\{\cdot\}$.

– **OS** [↑]**:** This index uses the same compactness term that the SV index, but uses a more complex separation term that involves a function called ov (Žalik and Žalik, 2011). This function requires two terms: $a_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k})$ and $b_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k})$. $a_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k})$ is the sum of distances between pairs in the cluster k and $b_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k})$ is the sum of the n_k minimum distances between a point in a cluster k and points in other clusters.

$$\text{OS} = \frac{\sum_{k=1}^{K} \sum_{\boldsymbol{x_i}\in \boldsymbol{X}^{\{k\}}} \text{ov}(\boldsymbol{x_i}, \boldsymbol{\mu^k})}{\sum_{k=1}^{K} \frac{1}{n_k} \sum \max_{\boldsymbol{x_j}\in \boldsymbol{X}^{\{k\}}}(0.1n_k) \left\{\|\boldsymbol{x_i} - \boldsymbol{\mu^k}\|\right\}} \tag{67}$$

Where:

$$\text{ov}(\boldsymbol{x_i}, \boldsymbol{\mu^k}) = \begin{cases} \frac{a_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k})}{b_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k})} & if \quad \frac{b_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k}) - a_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k})}{b_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k}) + a_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k})} < 0.4 \\ 0 & otherwise \end{cases} \tag{68}$$

$$b_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k}) = \frac{1}{n_k} \sum \min_{\boldsymbol{x_j}\notin \boldsymbol{X}^{\{k\}}}(n_k) \left\{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|\right\} \qquad a_{OS}(\boldsymbol{x_i}, \boldsymbol{\mu^k}) = \frac{1}{n_k} \sum_{\boldsymbol{x_j}\in \boldsymbol{X}^{\{k\}}} \|\boldsymbol{x_i} - \boldsymbol{x_j}\| \tag{69}$$

– **AR_{SD}** [↑]**:** This index is based on a compactness measure (CM_{SD}), which is obtained using the standard deviation and the maximum diameter of each clusters, and separation measure ($\text{DM}_{\text{clusters}}$) based on a penalty function that is calculated using recursive algorithm (Wani and Riyaz, 2016b).

$$\text{AR}_{\text{SD}} = \text{CM}_{\text{SD}} + \text{DM}_{\text{clusters}} = \text{CM}_{\text{SD}} - Fn \tag{70}$$

Where:

$$\mathrm{CM_{SD}} = \frac{1}{n_k} \sum_{k=1}^{K} \frac{md_k - \mathrm{SD}_k}{md_k} \tag{71}$$

$$md_k = \max_{\boldsymbol{x_i}, \boldsymbol{x_j} \in \boldsymbol{X^{\{k\}}}} \left\{ ||\boldsymbol{x_i} - \boldsymbol{x_j}|| \right\} \tag{72}$$

Where SD_k is the standard deviation of the cluster k and the steps to compute Fn are:

1. Set the penalty function Fn = 0 and a variable i=K.
2. Select any cluster from the list of K clusters.
3. Compute the average of minimum distance with the cluster:

$$d_{\min(\text{intra})} = \frac{1}{n_k} \sum_{\boldsymbol{x_i} \in \boldsymbol{X^{\{k\}}}} \min_{\boldsymbol{x_j} \in \boldsymbol{X^{\{k\}}}} \left\{ ||\boldsymbol{x_i} - \boldsymbol{x_j}|| \right\} \tag{73}$$

4. Compute between clusters distances:

$$d_{\min(\text{inter})} = \min_{\boldsymbol{x_i} \in \boldsymbol{X^{\{k\}}} \cap \boldsymbol{x_j} \notin \boldsymbol{X^{\{k\}}}} \left\{ ||\boldsymbol{x_i} - \boldsymbol{x_j}|| \right\} \tag{74}$$

5. Compute the penalty function for the selected cluster:

$$\mathrm{fn} = \begin{cases} 0 & if \quad d_{\min(\text{inter})} > 2d_{\min(\text{intra})} \\ a & if \quad d_{\min(\text{inter})} \leq 2d_{\min(\text{intra})} \\ a \times \frac{d_{\min(\text{inter})} - d_{\min(\text{intra})}}{d_{\min(\text{intra})}} & otherwise \end{cases} \tag{75}$$

   Where a was set experimentally to 0.1.
6. Accumulate the penalty function in Fn. $Fn = Fn + fn$
7. Remove the selected cluster from the list of clusters and decrement i.
8. Repeat from step 2 until $i < 2$.

– **STR [↑]:** This index is based on two components: cluster compactness E(K) and and cluster separation D(K) (Starczewski, 2017b). E(K) is defined as the ratio of the total scatter to the within scatter of the clusters. Its formula is:

$$\mathrm{STR}(K) = [E(K+1) - E(K)] \times [D(K+2) - D(K+1)] \tag{76}$$

$$E(K) = \frac{E_0}{E_K} = \frac{\sum_{j=1}^{n} ||\boldsymbol{x_j} - \boldsymbol{\mu}||}{\sum_{k=1}^{K} \sum_{\boldsymbol{x_i} \in \boldsymbol{X^{\{k\}}}} ||\boldsymbol{x_i} - \boldsymbol{\mu^k}||} \qquad D(K) = \frac{D_{Kmax}}{D_{Kmin}} = \frac{\max_{\boldsymbol{\mu^k}, \boldsymbol{\mu^{k'}} \in P_o} \left\{ ||\boldsymbol{\mu^k} - \boldsymbol{\mu^{k'}}|| \right\}}{\min_{\boldsymbol{\mu^k}, \boldsymbol{\mu^{k'}} \in P_o} \left\{ ||\boldsymbol{\mu^k} - \boldsymbol{\mu^{k'}}|| \right\}} \tag{77}$$

– **Bhargavi-Gowda Index [diff ↑]:** This index is based on two ratios, the ratio of WGSS and BGSS called "sum-of-squares ratio", and the ratio of intra-cluster distances and inter-cluster distances called "intra-inter

ratio" (Bhargavi and Gowda, 2015).

$$\text{Bhargavi-Gowda} = \left| \frac{WGSS}{BGSS} \times TSS - \frac{C_{\text{Intra}}}{C_{\text{Inter}}} - (n-k) \right| \tag{78}$$

$$C_{\text{Intra}} = \sum_{k=1}^{K} \sqrt{\sum_{i=1}^{n_k} \sum_{j=1}^{p} \left(x_{ij} - \boldsymbol{\mu}_j^{\boldsymbol{k}} \right)} \qquad C_{\text{Inter}} = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{K} \sum_{k'=1\,,\,k' \neq k}^{K} \sqrt{\sum_{j=1}^{p} \left(\boldsymbol{\mu}_j^{\boldsymbol{k}} - \boldsymbol{\mu}_j^{\boldsymbol{k'}} \right)^2} \right] \tag{79}$$

– **CVM [↑]:** This index favours solutions where the size of the cluster cores is maximized together with their separation (De Morsier et al., 2015b). It is defined as:

$$\text{CVM} = \left(\sum_{k=1}^{K} \frac{d_{\text{in}}(k)}{R_{in}(k)} \right) \left(\sum_{k=1}^{K} d_{\text{inter}}(k) \right) \tag{80}$$

Where $d_{\text{in}}(k)$, $d_{\text{inter}}(k)$ and $R_{in}(k)$ are, respectively, the maximal distance among the samples in the cluster k, the core separability defined as the minimum distance between a cluster sample to the closest sample from another cluster and the core homogeneity that is the ratio between the maximum and the average smallest inner distances inside a cluster.

$$d_{\text{in}}(k) = \max_{\boldsymbol{x_i}\,,\,\boldsymbol{x_j} \in \boldsymbol{X}^{\{k\}}} \left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\|^2 \qquad d_{\text{inter}}(k) = \max_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}\,,\,\boldsymbol{x_j} \in \boldsymbol{X}^{\{k'\}}} \left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\|^2 \tag{81}$$

$$R_{in}(k) = \frac{\max_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}} \left(\min_{\boldsymbol{x_j} \in \boldsymbol{X}^{\{k\}}} \left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\|^2 \right)}{\frac{1}{n_k} \sum_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}} \left(\min_{\boldsymbol{x_j} \in \boldsymbol{X}^{\{k\}}\,,\,\boldsymbol{x_i} \neq \boldsymbol{x_j}} \left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\|^2 \right)} \tag{82}$$

$R_{in}(k)$ takes values from 1 to ∞.

2.1.2. Indices based on concordance:

- ***Preliminary definitions:***

– **Number of concordant pairs of objects s_+:** The number of times that the distance between a pair of objects from the same cluster is lower than the distance between a pair of objects from different cluster:

$$s_+ = \frac{1}{2} \sum_{k=1}^{K} \sum_{\substack{\boldsymbol{x_i}\,,\,\boldsymbol{x_j} \in \boldsymbol{X}^{\{k\}} \\ \boldsymbol{x_i} \neq \boldsymbol{x_j}}} \frac{1}{2} \sum_{k'=1}^{K} \sum_{\substack{\boldsymbol{x_a} \in \boldsymbol{X}^{\{k'\}} \\ \boldsymbol{x_b} \notin \boldsymbol{X}^{\{k'\}}}} \delta \left(\|\boldsymbol{x_i} - \boldsymbol{x_j}\| < \|\boldsymbol{x_a} - \boldsymbol{x_b}\| \right) \tag{83}$$

Where $\delta(\cdot)$ equals 1 if the inner inequality is satisfied and equals 0 if not.

– **Number of discordant pairs of objects s_-:** The number of times that the distance between a pair of

objects from the same cluster is lower than the distance between a pair of objects from different cluster:

$$s_- = \frac{1}{2}\sum_{k=1}^{K}\sum_{\substack{\boldsymbol{x_i}\,\boldsymbol{x_j}\in \boldsymbol{X^{\{k\}}}\\ \boldsymbol{x_i}\neq\boldsymbol{x_j}}}\frac{1}{2}\sum_{k'=1}^{K}\sum_{\substack{\boldsymbol{x_a}\in \boldsymbol{X^{\{k'\}}}\\ \boldsymbol{x_b}\notin \boldsymbol{X^{\{k'\}}}}}\delta\left(\|\boldsymbol{x_i}-\boldsymbol{x_j}\|>\|\boldsymbol{x_a}-\boldsymbol{x_b}\|\right) \tag{84}$$

- ***Indices based on concordance***

  - **Baker-Hubert Gamma** [↑]: (Rohlf, 1974).

$$\text{Baker-Hubert Gamma} = \frac{s_+ - s_-}{s_+ + s_-} \tag{85}$$

  - **G plus** [↓]: This index is defined as the proportion of discordant pairs among all the pairs of distinct points (Rohlf, 1974).

$$\text{G (+)} = \frac{2s_-}{n_T(n_T-1)} \tag{86}$$

  - **Tau** [↑]: s_+ and s_- do no count if a between-cluster distance and a within-cluster distance are equal. This index includes in the denominator a term that corrects this issue (Rohlf, 1974).

$$\tau = \frac{s_+ - s_-}{\sqrt{n_B n_W\left(\frac{n_T(n_T-1)}{2}\right)}} \tag{87}$$

2.1.3. Density-based indices:

- ***Indices based on density***

  - S_{Dbw} **validity index** [↓]:

    The density $\gamma_{kk'}$ of a given point, relative to two clusters c_k and c'_k, is equal to the number of points in these two clusters whose distance to this point is less than σ (Halkidi et al., 2001).

$$\gamma_{kk'}(\boldsymbol{x}) = \sum_{\boldsymbol{x_i}\in \boldsymbol{X^{\{k\}}}\cup \boldsymbol{x_i}\in \boldsymbol{X^{\{k'\}}}}\delta\left(\|\boldsymbol{x_i}-\boldsymbol{x}\|>\sigma\right) \tag{88}$$

$$\sigma = \frac{1}{K}\sqrt{\sum_{k=1}^{K}\left\|\boldsymbol{v^{\{k\}}}\right\|} \tag{89}$$

    Where $\delta(\cdot)$ equals 1 if the inner inequality is satisfied and equals 0 if not. S_{Dbw} is defined by:

$$S_{Dbw} = S + \text{Ş} \tag{90}$$

$$\text{Ş} = \frac{2}{K(K-1)}\sum_{k<k'}R_{k,k'} \qquad R_{k,k'} = \frac{\gamma_{kk'}\left(\boldsymbol{H_{kk'}}\right)}{\max\left(\gamma_{kk'}\left(\boldsymbol{\mu^k}\right),\gamma_{kk'}\left(\boldsymbol{\mu^{k'}}\right)\right)} \tag{91}$$

Where $\boldsymbol{H_{kk'}}$ is the midpoint if the barycenters $\boldsymbol{\mu^k}$ and $\boldsymbol{\mu^{k'}}$.

2.1.4. Statistical-based indices:

- ***Preliminary definitions:***
  - **Assumption:** The data is modeled by n gaussians, each with identical variance σ and different means $\boldsymbol{\mu^k}$ (This assumption is also called "identical spherical assumption") (Pelleg et al., 2000).
  - **Probability of an object i is an element of cluster c_k under the maximum likelihood:**

$$P(\boldsymbol{x_i} \in c_k) = \frac{n_k}{n} \qquad P(\boldsymbol{x_i}) = \frac{n_k}{n} \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\sigma^2} \left\|\boldsymbol{x_i} - \boldsymbol{\mu^i}\right\|\right) \tag{92}$$

  $\boldsymbol{\mu^i}$ is the mean of the cluster to which $\boldsymbol{x_i}$ belongs.

  - **Log-likelihood:**

$$l(D) = \log \prod_i P(\boldsymbol{x_i}) \tag{93}$$

  - **Log-likelihood detailed formula:**

$$l(D) = \sum_{k=1}^{K} \left[n_k \left(\log \frac{n_k}{n} - \frac{p}{2} \log\left(2\pi\sigma^2\right) \right) - \frac{1}{2\sigma} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}^{\{k\}}} \left\|\boldsymbol{x_i} - \boldsymbol{\mu^k}\right\|^2 \right] \tag{94}$$

- ***Statistical-based Indices***
  - **BIC** [↑]: The version from (Pelleg et al., 2000) was implemeted.

$$\text{BIC} = l(D) - \frac{(p+1)K}{2} \log(n) \tag{95}$$

  - **AIC** [↑]: The version from (Pelleg et al., 2000) was implemeted.

$$\text{AIC} = l(D) - (p+1)K \tag{96}$$

2.1.5. Negentropy-based indices:

- ***Preliminary definitions:***
  - **Assumption:** the concept of negentropy measures the average distance to normality of the clusters in the partition (Lago-Fernández and Corbacho, 2010a).
- ***Indices based on negentropy***

– **Negentropy increment** ($\Delta J_B(\Omega, X)$) [↓]:

$$\text{Negentropy increment} = \frac{1}{2}\sum_{k=1}^{K}\left(p_k \log|\Sigma_k|\right) - \sum_{k=1}^{K} p_k \log p_k \tag{97}$$

Where $p_k = n_k/n$ and Σ_k is the covariance matrix of the cluster k.

– **Negentropy increment corrected** ($\Delta J_U(\Omega, X)$) [↓]: When the previous index is estimated with few samples produces an biased value. With a correction term this issue is supplied:

$$\text{Negentropy increment corrected} = \text{Negentropy increment} + \frac{1}{2}\sum_{k=1}^{K} p_k C\left(n_k, p\right) \tag{98}$$

$$C\left(n_k, p\right) = -p \log\frac{2}{n_k - 1} - \sum_{j=1}^{p}\psi\left(\frac{n_k - j}{2}\right) \tag{99}$$

Where $\psi\left(x\right) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\ln\Gamma(x)\right)$ is the digamma function.

– **Negentropy increment + variance** [↓]: The variance of $\Delta J_U(\Omega, X)$ can be estimated as:

$$\sigma_s^2(\Delta J_U) \approx \frac{1}{4}\sum_{k=1}^{K} p_k^2 \sum_{j=1}^{p}\psi'\left(\frac{n_k - j}{2}\right) \tag{100}$$

Where $\psi'\left(x\right)$ is the first derivative of the digamma function.
$\Delta J_U(\Omega, X)$ does not take into account the variance and because of it is estimated on a finite sample, the estimation could satisfy $\Delta J(\Omega_1, X) < \Delta J(\Omega_2, X)$ for the true distribution of X, but using the finite sample of X the estimation could satisfy $\Delta J_U(\Omega_1, X) > \Delta J_U(\Omega_2, X)$. To solve this problem we include in the analysis the variance of $\Delta J_U(\Omega_1, X)$. Two partitions are considered equivalent if:

$$\Delta J_U(\Omega_2, X) + \sigma_s^2(\Delta J_U(\Omega_2, X)) > \Delta J_U(\Omega_1, X) - \sigma_s^2(\Delta J_U(\Omega_1, X)) \tag{101}$$

In such cases, the simplest (lower number of clusters) partition is selected.

2.1.6. Feature-Maximization-based indices:

- ***Preliminary definitions:***

– **Assumption:** Feature maximization is an unbiased measure which can be used to estimate the quality of a classification whether it be supervised or unsupervised. In unsupervised classification (i.e. clustering), this measure exploits the properties (i.e. the features) of cluster associated data (Lamirel et al., 2016b).

– **Feature F-measure $FF_k(f)$ associated to cluster k:** is defined as the harmonic mean of the feature recall $FR_k(f)$ and of the feature predominance $FP_k(f)$ (Where subindex k refers to the feature f evaluated for

the cluster k).

$$FF_k(f) = \frac{2 \times FP_k(f) \times F2\_k(f)}{FP_k(f) + FR_k(f)} \qquad FP_k(f) = \frac{\sum_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}} \boldsymbol{x_i}^f}{\sum_{f' \in F_k} \sum_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}} \boldsymbol{x_i}^{f'}} \qquad FR_k(f) = \frac{\sum_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}} \boldsymbol{x_i}^f}{\sum_{k=1}^{K} \sum_{\boldsymbol{x_i} \in \boldsymbol{X}^{\{k\}}} \boldsymbol{x_i}^f}$$

Where $\boldsymbol{x_i}^f$ is the value of the feature f for the object i and F_k contains all the features associated with the cluster k. $FP_k(f)$ measures the ability of f to describe cluster k. $FR_k(f)$ measures the ability of f to discriminate k form other clusters.

– **Set S_k of relevant characteristic features of the cluster k:**

$$S_k = \left\{ f \in F_k \,|\, FF_k(f) > \overline{FF}(f) \text{ and } FF_k(f) > \overline{FF}_D \right\} \tag{102}$$

Where $\overline{FF}(f)$ is the arithmetic mean of the F-measure for the feature f and $\overline{FF}_D$ is the arithmetic mean of the $\overline{FF}(f)$ of all features in data.

$$\overline{FF}(f) = \sum_{k=1}^{K} FF_k(f) \qquad \overline{FF}_D = \sum_{f=1}^{F} \overline{FF}(f) \tag{103}$$

Features with $FF_k(f) < \overline{FF}_D$ are discarded.

– **Set S of relevant characteristic features for the partition:**

$$S = \bigcup_{k=1}^{K} S_k \tag{104}$$

– **Constrast $G_k(f)$ of the feature f for a given cluster k:**

$$G_k(f) = \frac{FF_k(f)}{\overline{FF}(f)} \tag{105}$$

• ***Indices based on feature-maximization***

– **PC** [↑]: This index is based on the maximization of arithmetic mean of the contrast of selected features:

$$PC_K = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{f \in S_k} G_k(f) \tag{106}$$

– **EC** [↑]: EC index is based on the maximization of the average weighted compromise between the contrast of active features and the inverted contrast of passive features:

$$EC_K = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\frac{|s_k|}{n_k} \sum_{f \in S_k} G_k(f) + \frac{|\overline{s_k}|}{n_k} \sum_{f \in S_k} \frac{1}{G_k(f)}}{|s_k| + |\overline{s_k}|} \right) \tag{107}$$

Where $|s_k|$ and $|\overline{s_k}|$ are the number of selected and not selected features associated with the cluster k, respectively.

2.2. External indices

2.2.1. External indices based on a contingency table between the obtained partition and the ground partition

Preliminary definitions:

- **Obtained partition P_o:** Clustering resulted from the clustering algorithm.
- **Ground partition P_{GT}:** Predefined classes labels of the dataset.
- **Contingency table:** Matrix $\boldsymbol{M}$ used to compare the partitions P_o and P_{GT}. P_o and P_{GT} are ways of partitioning S into disjointed subsets $U = \{U_1, U_2, ..., U_R\}$ and $V = \{V_1, V_2, ..., V_C\}$, respectively. In this way, $\boldsymbol{M}$ is a $R \times C$ matrix and each element $n_{i,j}$ represents the number of elements that U_i and V_j have in common (Hubert and Arabie, 1985a).

Table 1: Contingency table

U \V	V_1	V_2	...	V_C	Sums
U_1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,C}$	a_1
U_2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,C}$	a_2
...	...	...	...	...	...
U_R	$n_{R,1}$	$n_{R,2}$	...	$n_{R,C}$	a_R
Sums	b_1	b_2	...	b_C	N

- **Number of pairs yy:** the number of pair of points that were grouped in the same cluster in P_o and belongs to the same cluster in P_{GT}.

$$yy = \frac{1}{2}\sum_{i=1}^{R}\sum_{j=1}^{C} n_{i,j}(n_{i,j} - 1) \tag{108}$$

- **Number of pairs yn:** the number of pair of points that were grouped in the same cluster in P_o and in different clusters in P_{GT}.

$$yn = \frac{1}{2}\left(\sum_{j=1}^{R} a_j^2 - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{i,j}^2\right) \tag{109}$$

- **Number of pairs ny:** the number of pair of points that were grouped in different clusters in P_o and in the same cluster in P_{GT}.

$$ny = \frac{1}{2}\left(\sum_{j=1}^{C} b_j^2 - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{i,j}^2\right) \tag{110}$$

- **Number of pairs nn:** the number of pair of points that were grouped in different clusters in P_o and in different

clusters in P_{GT}.

$$nn = \frac{1}{2}\left(n^2 + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{i,j}^2 - \left(\sum_{i=1}^{R} a_i^2 + \sum_{j=1}^{C} b_j^2\right)\right) \tag{111}$$

- **Total number of pairs of points n_T:**

$$n_T = \frac{n(n+1)}{2} = yy + yn + ny + nn \tag{112}$$

- **Indicator variable X_a associated to the partition P_a:** Defined as a binary random variable that indicates if a pair of points i and j are in the same cluster in P_a. This variable takes the value of 1 if a pair of points are grouped in the same cluster and the value of 0 otherwise.

- **Mean μ_{X_a} of the indicator variable X_a:**

$$\mu_{X_a} = \frac{1}{n_T}\sum_{i<j} X_a(i,j) \tag{113}$$

- **Standard deviation σ_{X_a} of the indicator variable X_a:**

$$\sigma_{X_a} = \frac{1}{n_T}\sum_{i<j} [X_a(i,j)^2 - \mu_{X_a}] \tag{114}$$

- **Indicator variables for P_o and P_{GT}:**

$$\mu_{X_o} = \frac{1}{n_T}\sum_{i<j} X_o(i,j) = \frac{yy+yn}{n_T} \qquad \sigma^2_{X_o} = \frac{1}{n_T}\sum_{i<j}\left[X_a(i,j)^2 - \mu^2_{X_O}\right] = \frac{yy+yn}{n_T} - \left(\frac{yy+yn}{n_T}\right)^2 \tag{115}$$

$$\mu_{X_{GT}} = \frac{1}{n_T}\sum_{i<j} X_{GT}(i,j) = \frac{yy+ny}{n_T} \qquad \sigma^2_{X_{GT}} = \frac{1}{n_T}\sum_{i<j}\left[X_{GT}(i,j)^2 - \mu^2_{X_{GT}}\right] = \frac{yy+ny}{n_T} - \left(\frac{yy+ny}{n_T}\right)^2 \tag{116}$$

- **Precision:** is defined as the ratio **of** yy (the number of pairs of points that were grouped in the same cluster in P_o and belongs to the same cluster in P_{GT}) **to** $yy + ny$ (the number of pairs of points that in P_{GT} belongs to the same cluster). In other words, precision is the portion of point pairs rightly grouped according to P_{GT}.

$$\text{Precision} = \frac{yy}{yy+ny} \tag{117}$$

- **Recall:** is defined as the ratio **of** yy **to** $yy + yn$ (the number of pairs of points that where grouped in the same cluster in P_o). In other words, from all pairs that where grouped in the same cluster in P_o ($yy + yn$) what is the

proportion that belongs to the same cluster in P_{GT}.

$$\text{Recall} = \frac{yy}{yy + yn} \tag{118}$$

Indices based on the contingency table

- **F-measure**: This index is based on the well-known metric of the supervised learning, F-score. The F-measure is the harmonic mean of the precision and recall coefficients (Larsen and Aone, 1999):

$$\text{F-Measure} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{119}$$

- **Folkes-Mallows**: This index is defined as (Fowlkes and Mallows, 1983):

$$\text{Folkes-Mallows} = \frac{yy}{\sqrt{(yy + yn) \times (yy + ny)}} = \sqrt{\text{Precision} \times \text{Recall}} \tag{120}$$

- **Hubert** Γ: This index is based on the correlation between the indicator variables of P_o and P_{GT}. This index can be defined in terms of the confusion matrix (Hubert, 1979).

$$\text{Hubert } \Gamma = \frac{n_T \times yy - (yy + yn)(yy + ny)}{\sqrt{(yy + yn)(yy + ny)(nn + yn)(nn + ny)}} \tag{121}$$

- **Jaccard**: This index is defined as (Rousseeuw and Kaufman, 1990):

$$\text{Jaccard} = \frac{yy}{yy + yn + ny} \tag{122}$$

- **Kulczynski**: This index is defined as (Murguía and Villaseñor, 2003):

$$\text{Kulczynski} = \frac{1}{2}\left(\frac{yy}{yy + ny} + \frac{yy}{yy + yn}\right) = \frac{1}{2}\left(\text{Precision} + \text{Recall}\right) \tag{123}$$

- **McNemar**: It is an adaptation of the non-parametric test of McNemar for the comparison of frequencies between two paired samples (Eliasziw and Donner, 1991).

$$\text{McNemar} = \frac{nn - ny}{\sqrt{nn + yy}} \tag{124}$$

- **Phi**: Phi index is based on the correlation between two dichotomic variables (Cramér, 2016).

$$\text{Phi} = \frac{yy \times nn - yn \times ny}{(yy + yn)(yy + ny)(yn + nn)(ny + nn)} \tag{125}$$

- **Rand**: (Rand, 1971)

$$\text{Rand} = \frac{yy + nn}{n_T} \tag{126}$$

- **Adjusted Rand index (ARI)**: Because Rand have a score different to 0.0 when the label assignments were random (uniform), an adjusted version of the rand index (ARI) was proposed. ARI is the difference of the Rand index and its expected value under the null hypothesis, which assumes a generalized hypergeometric distribution. To see more details refer to (Hubert and Arabie, 1985b).

$$\text{ARI} = \frac{Rand - E[Rand]}{max(Rand) - E[Rand]} \tag{127}$$

$$E[Rand] = \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{N}{2}} \qquad max(Rand) = \frac{1}{2} \frac{\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}}{\binom{N}{2}} \tag{128}$$

- **Rogers-Tanimoto**:

$$\text{Rogers-Tanimoto} = \frac{yy + nn}{yy + nn + 2(yn + ny)} \tag{129}$$

- **Russel-Rao**: is defined as the proportion of concordances between the two partitions (Rao, 1948):

$$\text{Russel-Rao} = \frac{yy}{\sqrt{n_T}} \tag{130}$$

- **Sokal-Sneath**: Sokal-Sneath is defined as the proportion of concordances between the two partitions:

$$\text{Sokal-Sneath} = \frac{yy}{yy + 2(yn + ny)} \tag{131}$$

2.2.2. External indices based on information theory and intuitive concepts:

Preliminary definitions:

- **Entropy $H(A)$ of partition P_a:** P_a is a way of partitioning S into disjointed subsets $A = \{A_1, A_2, ..., A_M\}$ and the entropy of P_a is the amount of uncertainty.

$$H(A) = -\sum_{i=1}^{M} P(i) log(P(i)) \tag{132}$$

Where $P(i) = |A_i|/N$ is the probability that an element of S falls into cluster A_i.

- **Conditional entropy $H(V/U)$:** is the conditional entropy of the C classes given the R cluster assignments.

$$H(V/U) = -\sum_{k=1}^{R} \frac{n_k}{n} \sum_{c=1}^{C} \frac{n_{c,k}}{n_k} log\left(\frac{n_{c,k}}{n_k}\right) \tag{133}$$

Where $n_{c,k}$ is the number of elements of the class c in the cluster k and n_k is the number of elements of the cluster k.

- **Probability $P(i, j)$:** is the probability that an element falls into a cluster A_i of a partition P_a and a cluster B_i of a partition P_b.

$$P(i, j) = \frac{|A_i \cap B_j|}{n} \tag{134}$$

Indices based on information theory

- **Conditional Entropy:** (Shannon, 2001)

$$\text{Conditional Entropy} = H(V/U) \tag{135}$$

- **Homogeneity:** Based on that a cluster must contain only elements of a single class (Rosenberg and Hirschberg, 2007).

$$\text{Homogeneity} = 1 - \frac{H(V/U)}{H(V)} \tag{136}$$

- **Completeness:** This index is symmetrical to homogeneity. This metric assumes that a class must be represented only by a single cluster (Rosenberg and Hirschberg, 2007).

$$\text{Completeness} = 1 - \frac{H(U/V)}{H(U)} \tag{137}$$

- **V-Measure:** defined as the harmonic mean of homogeneity and completeness (Rosenberg and Hirschberg, 2007).

$$\text{V-Measure} = 2 \times \frac{\text{Homogeneity} \times \text{Completeness}}{\text{Homogeneity} + \text{Completeness}} \tag{138}$$

- **Mutual information:** (Strehl and Ghosh, 2002)

$$\text{MI(U,V)} = \sum_{i=1}^{R} \sum_{i=1}^{C} P(i, j) \log \left(\frac{P(i, j)}{P(i)P(j)} \right) \tag{139}$$

- **Normalized mutual information:** (Strehl and Ghosh, 2002)

$$\text{NMI(U,V)} = \frac{\text{MI(U,V)}}{\sqrt{H(U)H(V)}} \tag{140}$$

- **Adjusted mutual information:** Proposed by (Vinh et al., 2009)

$$\text{AMI(U,V)} = \frac{\text{MI} - E[\text{MI}]}{max\left[H(U), H(V)\right] - E[\text{MI}]} \tag{141}$$

Where $E[\mathrm{MI}]$ is the expected mutual information and is defined by:

$$E[\mathrm{MI}] = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{n_{i,j}=max(a_i+b_j-n,0)}^{min(a_i,b_j)} \frac{n_{i,j}}{n} log\left(\frac{n \times n_{i,j}}{a_i \times b_j}\right) \frac{a_i!b_j!(n-a_i)!\left(n-b_j\right)!}{n!n_{i,j}!\left(a_i-n_{i,j}\right)!\left(b_j-n_{i,j}\right)!\left(n-a_i-b_j-n_{i,j}\right)!} \quad (142)$$

3. Experimental setup

This section explains the comparative study performed to analyze the clustering validation indices previously introduced. First, a recompilation of 26 synthetic and 23 real datasets have been chosen as benchmark datasets, aiming to consider the more general problems in cluster analysis with quantitative data. Then, six clustering algorithms have been used to create the candidate partitions: K-means (Jain, 2010), DBSCAN (Ester et al., 1996), Expectation Maximization (Dempster et al., 1977), Birch (Zhang et al., 1996), Sting (Wang et al., 1997), and Lamda (Bedoya et al., 2014). A brief explanation of the algorithms is presented in section 3.2. Each algorithm has at least one input parameter that the user must specify. A combination of parameters was performed per algorithm, generating the multiple partitions. Additional to this, three standardizations of features were used to analyze the effect of the scale in clustering validation. In total, 178737 partitions were generated. Both external and internal clustering validity indices were computed for each candidate partition. However, indices based on concordance and discordance were only estimated for datasets that contains less than 500 samples due to their high computational requirements. Then, a comparative methodology was carried out, analyzing the matching among internal and external indices. Following the ideas of (Gurrutxaga et al., 2011; Arbelaitz et al., 2013), external indices work as reference metrics, and internal indices should select as best partitions those that the external indices have selected.

3.1. Datasets

Datasets from several studies were recollected and those that were used in major number of times were selected (Arbelaitz et al., 2013; Wani and Riyaz, 2016b; Zhao and Fränti, 2014; Liu et al., 2013; Gurrutxaga et al., 2011). We tried to include the more general cases that can be found in clustering literature of quantitative data. The characteristics of these synthetic and real datasets are shown in table 2. Most of the synthetic datasets are two-dimensional datasets, therefore is possible to observe graphically the results. Real datasets were mainly chosen from the work of (Arbelaitz et al., 2013) due to one of the main goal of the present work is to extend that analysis.

3.2. Clustering algorithms

The first five algorithms were selected based on a clustering algorithm categorization done by (Fahad et al., 2014), who analyzed the ability of clustering algorithms to handle big data and divided the algorithms into five categories (partitional, hierarchical, density-based, grid-based, and model-based). A representative algorithm was chosen per each category. In this work, the same algorithms of the partitional, hierarchical and model-based categories were kept, but the density-based and grid-based algorithms were replaced by DBSCAN and STING. This decision was taken due