Clustering Indices

Bernard Desgraupes
University Paris Ouest
Lab Modal'X

November 2017

Contents

1 Internal clustering criteria

1.1 Algebraic background and notations

Let us denote by A the data matrix: each row is an observation O_i corresponding to an individual and each column represents a variable observed for all the individuals.

There are N observations and p variables. The size of matrix A is $N \times p$.

The data are assumed to be partitioned in K groups (or *clusters*). Let us denote by P the vector representing a partition of the data: it is an integer vector with values between 1 and K. The size of P is equal to the number N of observations. For each index i ($1 \leq i \leq N$), the coordinate P_i is equal to the number k ($1 \leq k \leq K$) of the cluster the observation O_i belongs to.

The cluster C_k can be represented by a submatrix $A^{\{k\}}$ of matrix A made of the rows of A whose index i is such that $P_i = k$. If n_k denotes the cardinal of C_k, the matrix $A^{\{k\}}$ has size $n_k \times p$ and one has the relation $\sum_k n_k = N$. Let us denote by I_k the set of the indices of the observations belonging to the cluster C_k:

$$I_k = \{i \,|\, O_i \in C_k\} = \{i \,|\, P_i = k\}.$$

The matrix $A^{\{k\}}$ can also be denoted formally as $A_{\{I_k\}}$.

Let us denote by $\mu^{\{k\}}$ the barycenter of the observations in the cluster C_k and by μ the barycenter of all the observations. $\mu^{\{k\}}$ and μ are row-vectors with length p: they are the means of the rows of the matrices $A^{\{k\}}$ and A respectively:

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

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

where x_i designates the row of index i in A.

1.1.1 Total dispersion

Each column vector V_j ($1 \leq j \leq p$) of the matrix A can be interpreted as a sample of size N of the j-th observed variable. Let us center each of these vectors with respect to its mean by setting $v_j = V_j - \mu_j$. If X is the matrix formed by the centered vectors v_j, the scatter matrix T is the matrix defined by

$$T = {}^tX\, X.$$

The general term of T is:

$$t_{ij} = \sum_{l=1}^{N} (a_{li} - \mu_i)(a_{lj} - \mu_j) \tag{3}$$

The matrix T is equal to N times the variance-covariance matrix of the family of column vectors $(V_1, \ldots, V_p)$. The general term of T can thus also be written as

$$t_{ij} = N \times \mathrm{Cov}(V_i, V_j). \tag{4}$$

In particular, the diagonal terms are N times the variances of the vectors V_i:

$$t_{ii} = N \times \mathrm{Var}(V_i). \tag{5}$$

One can also write:

$$t_{ij} = {}^t(V_i - \mu_i)(V_j - \mu_j) \tag{6}$$

where, by a slight abuse of notation, μ_i and μ_j are here identified with the vectors $\mu_i \mathbf{1}$ and $\mu_j \mathbf{1}$ respectively.

The scatter matrix is a square symmetric matrix of size $p \times p$. As it is of the form ${}^tX\,X$, the quadratic form it represents is positive semi-definite. Indeed, if one takes any vector v in $\mathbb{R}^p$:

$${}^tvTv = {}^tv{}^tX\,Xv = {}^t(Xv)\,(Xv) = ||Xv||^2 \geq 0 \tag{7}$$

In particular, the eigenvalues and the determinant of the scatter matrix are also greater than or equal to 0. If $N > p$ and if the matrix X has maximal rank p, the form is in fact positive definite.

The total scattering TSS (*total sum of squares*) is the trace of the matrix T:

$$TSS = \mathrm{Tr}(T) = N \sum_{j=1}^{p} \mathrm{Var}(V_j) \tag{8}$$

Geometric interpretation: let us denote by $M_1, \ldots, M_N$ the points of the space $\mathbb{R}^p$ representing all the observations: the coordinates of M_i are the coefficients of the i-th row of the data matrix A. Similarly, let us denote by G the barycenter of these points: its coordinates are the coefficients of the vector μ. One can easily prove the following relations:

$$\begin{aligned} \mathrm{Tr}(T) &= \sum_{i=1}^{N} ||M_i - G||^2 &(9)\\ &= \frac{1}{N} \sum_{i<j} ||M_i - M_j||^2 &(10) \end{aligned}$$

It means that the trace of T, in other words the total scattering TSS, is equal to the scattering (sum of the squared distances) of the points around the barycenter. The second equality shows that this quantity is also the sum of the distances between all the pairs of points, divided by N.

1.1.2 Within-group scatter

There are similar definitions for the different clusters C_k: each column vector $V_j^{\{k\}}$ of the matrix $A^{\{k\}}$ represents a sample of size n_k of the j-th observed variable.

For each cluster C_k, one defines the within-group scatter matrix (abbreviated as *WG*). If $\mu^{\{k\}}$ designates the barycenter of the observations in cluster k and $X^{\{k\}}$ is the matrix formed by the centered vectors $v_j^{\{k\}} = V_j^{\{k\}} - \mu_j^{\{k\}}$, the within-group scatter matrix is defined by the following relation:

$$WG^{\{k\}} = {}^tX^{\{k\}}\,X^{\{k\}} \tag{11}$$

and its general term is defined as:

$$w_{ij}^{\{k\}} = {}^t\big(V_i^{\{k\}} - \mu_i^{\{k\}}\big)\big(V_j^{\{k\}} - \mu_j^{\{k\}}\big) \tag{12}$$

In terms of variance and covariance, by analogy with the relations (4) and (5), the coefficients of the matrix $WG^{\{k\}}$ can also be written as:

$$\begin{cases} w_{ij}^{\{k\}} & = n_k \times \mathrm{Cov}\big(V_i^{\{k\}}, V_j^{\{k\}}\big) \\ w_{ii}^{\{k\}} & = n_k \times \mathrm{Var}\big(V_i^{\{k\}}\big) \end{cases} \tag{13}$$

The matrices $WG^{\{k\}}$ are square symmetric matrices of size $p \times p$. Let us denote by WG their sum for all the clusters:

$$WG = \sum_{k=0}^{K} WG^{\{k\}} \tag{14}$$

As was the case with the matrix T seen in section 1.1.1, the matrices $WG^{\{k\}}$ represent a positive semi-definite quadratic form Q_k and, in particular, their eigenvalues and their determinant are greater than or equal to 0.

The *within-cluster dispersion*, noted $WGSS^{\{k\}}$ or $WGSS_k$, is the trace of the scatter matrix $WG^{\{k\}}$:

$$WGSS^{\{k\}} = \mathrm{Tr}(WG^{\{k\}}) = \sum_{i \in I_k} ||M_i^{\{k\}} - G^{\{k\}}||^2 \tag{15}$$

The within-cluster dispersion is the sum of the squared distances between the observations $M_i^{\{k\}}$ and the barycenter $G^{\{k\}}$ of the cluster.

Finally the *pooled within-cluster sum of squares* WGSS is the sum of the within-cluster dispersions for all the clusters:

$$WGSS = \sum_{k=0}^{K} WGSS^{\{k\}} \tag{16}$$

The abovementioned geometric interpretation remains true at the level of each group: in each cluster C_k, the sum of the squared distances from the points of the cluster to their barycenter is also the sum of the squared distances between all the pairs of points in the cluster, divided par n_k. In other words:

$$\begin{aligned} WGSS^{\{k\}} &= \sum_{i \in I_k} ||M_i^{\{k\}} - G^{\{k\}}||^2 & (17) \\ &= \frac{1}{n_k} \sum_{i<j \in I_k} ||M_i^{\{k\}} - M_j^{\{k\}}||^2 & (18) \end{aligned}$$

Inverting the formula, one gets:

$$\begin{aligned} \sum_{i \neq j} ||M_i^{\{k\}} - M_j^{\{k\}}||^2 &= 2 \sum_{i<j} ||M_i^{\{k\}} - M_j^{\{k\}}||^2 \\ &= 2n_k \sum_{i \in I_k} ||M_i^{\{k\}} - G^{\{k\}}||^2 \\ &= 2n_k \, WGSS^{\{k\}} & (19) \end{aligned}$$

1.1.3 Between-group scatter

The between-group dispersion measures the dispersion of the clusters between each other. Precisely it is defined as the dispersion of the barycenters $G^{\{k\}}$ of each cluster with respect to the barycenter G of the whole set of data.

Let us denote by B the matrix formed in rows by the vectors $\mu^{\{k\}} - \mu$, each one being reproduced n_k times ($1 \leq k \leq K$). The between-group scatter matrix is the matrix

$$BG = {}^tB\,B. \tag{20}$$

The general term of this matrix is:

$$b_{ij} = \sum_{k=1}^{K} n_k(\mu_i^{\{k\}} - \mu_i)(\mu_j^{\{k\}} - \mu_j) \tag{21}$$

The between-group dispersion BGSS is the trace of this matrix:

$$\begin{aligned} BGSS = \mathrm{Tr}(BG) &= \sum_{k=1}^{K} n_k\, {}^t(\mu^{\{k\}} - \mu)(\mu^{\{k\}} - \mu) \\ &= \sum_{k=1}^{K} n_k\, ||\mu^{\{k\}} - \mu||^2 \\ &= \sum_{k=1}^{K} n_k \sum_{j=0}^{p} (\mu_j^{\{k\}} - \mu_j)^2 \end{aligned} \tag{22}$$

Geometrically, this sum is the weighted sum of the squared distances between the $G^{\{k\}}$ and G, the weight being the number n_k of elements in the cluster C_k:

$$BGSS = \sum_{k=1}^{K} n_k ||G^{\{k\}} - G||^2. \tag{23}$$

1.1.4 Pairs of points

The observations (rows of the matrix A) can be represented by points in the space $\mathbb{R}^p$. Several quality indices defined in section 1.2 consider the distances between these points. One is led to distinguish between pairs made of points belonging to the same cluster and pairs made of points belonging to different clusters.

In the cluster C_k, there are $n_k(n_k - 1)/2$ pairs of distinct points (the order of the points does not matter). Let us denote by N_W the total number of such pairs:

$$N_W = \sum_{k=1}^{K} \frac{n_k(n_k - 1)}{2} \tag{24}$$

$$= \frac{1}{2}\left(\sum_{k=1}^{K} n_k^2 - \sum_{k=1}^{K} n_k\right) \tag{25}$$

$$= \frac{1}{2}\left(\sum_{k=1}^{K} n_k^2 - N\right) \tag{26}$$

The total number of pairs of distinct points in the data set is

$$N_T = \frac{N(N-1)}{2} \tag{27}$$

Since $N = \sum_{k=1}^{K} n_k$, one can write :

$$\begin{aligned} N_T = \frac{N(N-1)}{2} &= \frac{1}{2}\left(\sum_{k=1}^{K} n_k\right)^2 - \frac{1}{2}\sum_{k=1}^{K} n_k \\ &= \frac{1}{2}\left(\sum_{k=1}^{K} n_k^2 + 2\sum_{k<k'} n_k n_{k'}\right) - \frac{1}{2}\sum_{k=1}^{K} n_k \\ &= N_W + \sum_{k<k'} n_k n_{k'} \end{aligned} \tag{28}$$

Let us denote by N_B the number of pairs constituted of points which do not belong to the same cluster, one has $N_T = N_W + N_B$ and consequently:

$$N_B = \sum_{k<k'} n_k n_{k'}. \tag{29}$$

In the remainder, I_B will denote the set of the N_B pairs of between-cluster indices and I_W the set of the N_W pairs of within-cluster indices.

1.2 Internal indices

The following sections provide the precise definitions of the various internal quality indices which have been proposed by various authors in order to determine an optimal clustering. They are sorted in alphabetical order. These indices, also called quality indices, are all denoted by the same letter $\mathcal{C}$. Let us also denote by d the distance function between two points (usually the ordinary euclidean distance).

Table 1 summarizes the existing indices, their name in the package *clusterCrit* for R, the bibliographic reference and the date of the article where they were originally defined.

1.2.1 The Ball-Hall index

The mean dispersion of a cluster is the mean of the squared distances of the points of the cluster with respect to their barycenter. The Ball-Hall index is the mean, through all the clusters, of their mean dispersion:

$$\boxed{\mathcal{C} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in I_k} ||M_i^{\{k\}} - G^{\{k\}}||^2} \tag{30}$$

In the particular case where all the clusters have the same size N/K, this sum reduces to $\frac{1}{N}WGSS$.

Index	*Name in R*	*Ref.*	*Date*
Ball-Hall	Ball_Hall	[2]	1965
Banfeld-Raftery	Banfeld_Raftery	[3]	1974
C index	C_index	[15]	1976
Calinski-Harabasz	Calinski_Harabasz	[5]	1974
Davies-Bouldin	Davies_Bouldin	[6]	1979
$\|T\|/\|W\|$	Det_Ratio	[24]	1971
Dunn	Dunn	[7]	1974
Dunn generalized	GDImn	[4]	1998
Gamma	Gamma	[1]	1975
$G+$	G_plus	[23]	1974
$k^2\|W\|$	Ksq_DetW	[16]	1975
$\log(\|T\|/\|W\|)$	Log_Det_Ratio	[24]	1971
$\log(BGSS/WGSS)$	Log_SS_Ratio	[14]	1975
McClain-Rao	McClain_Rao	[17]	2001
PBM	PBM	[19]	2004
Point biserial	Point_biserial	[18]	1981
Ratkowsky-Lance	Ratkowsky_Lance	[21]	1978
Ray-Turi	Ray_Turi	[22]	1999
Scott-Symons	Scott_Symons	[24]	1971
SD	SD_Scat	[13]	2001
SD	SD_Dis	[13]	2001
S_Dbw	S_Dbw	[12]	2001
Silhouette	Silhouette	[20]	1987
$\mathrm{Tr}(W)$	Trace_W	[8]	1965
$\mathrm{Tr}(W^{-1}B)$	Trace_WiB	[10]	1967
Wemmert-Gançarski	Wemmert_Gancarski		
Xie-Beni	Xie_Beni	[25]	1991

Table 1: Index names in the package *clusterCrit* for R and bibliographic references.

1.2.2 The Banfeld-Raftery index

This index is the weighted sum of the logarithms of the traces of the variance-covariance matrix of each cluster.

The index can be written like this:

$$\boxed{\mathcal{C} = \sum_{k=1}^{K} n_k \log\left(\frac{\mathrm{Tr}(WG^{\{k\}})}{n_k}\right)} \tag{31}$$

The quantity $\mathrm{Tr}(WG^{\{k\}})/n_k$ can be interpreted, after equation (15), as the mean of the squared distances between the points in cluster C_k and their barycenter $G^{\{k\}}$. If a cluster contains a single point, this trace is equal to 0 and the logarithm is undefined.

1.2.3 The C index

Let us consider the distances between the pairs of points inside each cluster. The numbers N_W and N_T have been defined in section 1.1.4. One computes the following three quantities:

- S_W is the sum of the N_W distances between all the pairs of points inside each cluster ;
- S_{min} is the sum of the N_W smallest distances between all the pairs of points in the entire data set. There are N_T such pairs (see section 1.1.4): one takes the sum of the N_W smallest values ;
- S_{max} is the sum of the N_W largest distances between all the pairs of points in the entire data set. There are N_T such pairs: one takes the sum of the N_W largest values.

The C index is defined like this:

$$\boxed{\mathcal{C} = \frac{S_W - S_{min}}{S_{max} - S_{min}}} \tag{32}$$

If one considers the N_T distances between pairs of points as a sequence of values sorted in increasing order, the $\mathcal{C}$ index uses the N_W smallest values and the N_W largest values in order to compute the sums S_{min} and S_{max}: the sum S involves the N_W distances in this sequence which correspond to pairs present in some cluster (that is to say pairs whose two points are in a same cluster). No more than $3N_W$ distances are effectively retained in the calculation of this index.

1.2.4 The Calinski-Harabasz index

Using the notations of equations (16) and (23), the Calinski-Harabasz index is defined like this:

$$\boxed{\mathcal{C} = \frac{BGSS/(K-1)}{WGSS/(N-K)} = \frac{N-K}{K-1}\frac{BGSS}{WGSS}} \tag{33}$$

1.2.5 The Davies-Bouldin index

Let us denote by δ_k the mean distance of the points belonging to cluster C_k to their barycenter $G^{\{k\}}$:

$$\delta_k = \frac{1}{n_k} \sum_{i \in I_k} ||M_i^{\{k\}} - G^{\{k\}}|| \tag{34}$$

Let us also denote by

$$\Delta_{kk'} = d(G^{\{k\}}, G^{\{k'\}}) = ||G^{\{k'\}} - G^{\{k\}}||$$

the distance between the barycenters $G^{\{k\}}$ and $G^{\{k'\}}$ of clusters C_k and $C_{k'}$.

One computes, for each cluster k, the maximum M_k of the quotients $\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}}$ for all indices $k' \neq k$. The Davies-Bouldin index is the mean value, among all the clusters, of the quantities M_k:

$$\boxed{\mathcal{C} = \frac{1}{K} \sum_{k=1}^{K} M_k = \frac{1}{K} \sum_{k=1}^{K} \max_{k' \neq k} \left(\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}} \right)} \tag{35}$$

1.2.6 The Det_Ratio index

The Det_Ratio index is defined like this:

$$\boxed{\mathcal{C} = \frac{\det(T)}{\det(WG)}} \tag{36}$$

T designates the total scatter matrix defined in section 1.1.1. This is the sum of matrices BG and WG defined in equations (14) and (20).

1.2.7 The Dunn index

Let us denote by d_{min} the minimal distance between points of different clusters and d_{max} the largest within-cluster distance.

The distance between clusters C_k and $C_{k'}$ is measured by the distance between their closest points:

$$d_{kk'} = \min_{\substack{i \in I_k \\ j \in I_{k'}}} ||M_i^{\{k\}} - M_j^{\{k'\}}|| \tag{37}$$

and d_{min} is the smallest of these distances $d_{kk'}$:

$$d_{min} = \min_{k \neq k'} d_{kk'} \tag{38}$$

For each cluster C_k, let us denote by D_k the largest distance separating two distinct points in the cluster (sometimes called the diameter of the cluster):

$$D_k = \max_{\substack{i,j \in I_k \\ i \neq j}} ||M_i^{\{k\}} - M_j^{\{k\}}||. \tag{39}$$

Then d_{max} is the largest of these distances D_k:

$$d_{max} = \max_{1 \leq k \leq K} D_k \tag{40}$$

The Dunn index is defined as the quotient of d_{min} and d_{max}:

$$\boxed{\mathcal{C} = \frac{d_{min}}{d_{max}}} \tag{41}$$

1.2.8 The Baker-Hubert Gamma index

The Gamma index of Baker-Hubert is an adaptation, in the context of clustering, of the index Γ of correlation between two vectors of data A and B with the same size.

Generally, for two indices i and j such that $a_i < a_j$, one says that the two vectors are *concordant* if $b_i < b_j$, in other words, if the values classify in the same order in both vectors. One calculates the number s^+ of concordant pairs $\{i, j\}$ and the number s^- of discordant pairs. Note that the inequalities are strict, meaning that ties are dropped.

In this context, the Γ index is classically defined like this (see [11]):

$$\boxed{\mathcal{C} = \Gamma = \frac{s^+ - s^-}{s^+ + s^-}} \tag{42}$$

Its value is between -1 and 1.

In the context of a partition, the first vector A is chosen to be the set of distances d_{ij} between pairs of points $\{M_i, M_j\}$ (with $i < j$). The second vector B is a binary vector: in this vector, the coordinate corresponding to a pair $\{M_i, M_j\}$ has value 0 if the two points lie in the same cluster and 1 otherwise. These two vectors have length $N_T = N(N-1)/2$.

The number s^+ represents the number of times a distance between two points which belong to the same cluster (that is to say a pair for which the value of vector B is 0) is strictly smaller than the distance between two points not belonging to the same cluster (that is to say a pair for which the value of vector B is 1). The number s^- represents the number of times the opposite situation occurs, that is to say that a distance between two points lying in the same cluster (value 0 in B) is strictly greater than a distance between two points not belonging to the same cluster (value 1 in B). The cases where there is equality (ties or *ex-aequos*) are not taken into account. As defined in section 1.1.4, there are N_B between-cluster distances and, for each of them, one compares with the N_W within-cluster distances: one finally performs $N_B \times N_W$ comparisons.

One can write the numbers s^+ and s^- in the following form:

$$s^+ = \sum_{(r,s)\in I_B} \sum_{(u,v)\in I_W} \mathbf{1}_{\{d_{uv} < d_{rs}\}} \tag{43}$$

$$s^- = \sum_{(r,s)\in I_B} \sum_{(u,v)\in I_W} \mathbf{1}_{\{d_{uv} > d_{rs}\}} \tag{44}$$

Their difference is:

$$s^+ - s^- = \sum_{(r,s)\in I_B} \sum_{(u,v)\in I_W} \operatorname{sgn}(d_{rs} - d_{uv}) \tag{45}$$

1.2.9 The GDI index

The GDI indices are generalisations of the Dunn index seen in section 1.2.7 (GDI is the abbreviation of *Generalized Dunn's Indices*). They use different quantities in order to evaluate the between-clusters and within-groups distances.

Let us denote by the letter δ a measure of the between-cluster distance and by Δ a measure of the within-cluster distance (which is also called the diameter of the cluster). The GDI index, relatively to these distances, is defined like this:

$$\boxed{\mathcal{C} = \frac{\min_{k \neq k'} \delta(C_k, C_{k'})}{\max_k \Delta(C_k)}} \tag{46}$$

with $1 \leq k \leq K$ and $1 \leq k' \leq K$.

Six different definitions of δ (denoted as δ_1 through δ_6) and three definitions of Δ (denoted as Δ_1 through Δ_3) have been suggested. This leads to 18 different indices denoted as $\mathcal{C}_{uv}$: here u is an integer designating the between-clusters distance ($1 \leq u \leq 6$) and v an integer designating the within-groups distance ($1 \leq v \leq 3$).

The definitions of the within-cluster distances Δ are:

$$\Delta_1(C_k) = \max_{\substack{i,j \in I_k \\ i \neq j}} d(M_i, M_j) \tag{47}$$

$$\Delta_2(C_k) = \frac{1}{n_k(n_k-1)} \sum_{\substack{i,j \in I_k \\ i \neq j}} d(M_i, M_j) \tag{48}$$

$$\Delta_3(C_k) = \frac{2}{n_k} \sum_{i \in I_k} d(M_i, G^{\{k\}}) \tag{49}$$

Here d is the euclidean distance. The facteur 2 in the definition of Δ_3 allows us to interpret the value as a diameter rather than a radius.

The definitions of the between-cluster distances δ are:

$$\delta_1(C_k, C_{k'}) = \min_{\substack{i \in I_k \\ j \in I_{k'}}} d(M_i, M_j) \tag{50}$$

$$\delta_2(C_k, C_{k'}) = \max_{\substack{i \in I_k \\ j \in I_{k'}}} d(M_i, M_j) \tag{51}$$

$$\delta_3(C_k, C_{k'}) = \frac{1}{n_k n_{k'}} \sum_{\substack{i \in I_k \\ j \in I_{k'}}} d(M_i, M_j) \tag{52}$$

$$\delta_4(C_k, C_{k'}) = d(G^{\{k\}}, G^{\{k'\}}) \tag{53}$$

$$\delta_5(C_k, C_{k'}) = \frac{1}{n_k + n_{k'}} \left(\sum_{i \in I_k} d(M_i, G^{\{k\}}) + \sum_{j \in I_{k'}} d(M_j, G^{\{k'\}}) \right) \tag{54}$$

$$\delta_6(C_k, C_{k'}) = \max \left\{ \sup_{i \in I_k} \inf_{j \in I_{k'}} d(M_i, M_j), \sup_{j \in I_{k'}} \inf_{i \in I_k} d(M_i, M_j) \right\} \tag{55}$$

The first four distances (δ_1 to δ_4) occur in ascendant clustering algorithms and are called *single linkage, complete linkage, average linkage, centroid linkage* respectively. The measure δ_5 is the weighted mean (with weights n_k and $n_{k'}$) of the mean distances between the points in clusters C_k and $C_{k'}$ and their respective barycenter. The measure δ_6 is the Hausdorff distance D_H.

1.2.10 The G_plus index

Using the same notations as for the Baker-Hubert Γ index seen in section 1.2.8, the $G+$ index is defined like this:

$$\boxed{\mathcal{C} = \frac{s^-}{N_T(N_T-1)/2} = \frac{2s^-}{N_T(N_T-1)}} \tag{56}$$

This is the proportion of discordant pairs among all the pairs of distinct points.

1.2.11 The Ksq_DetW index

The Ksq_DetW index (also denoted as $k^2\,|W|$) is defined like this:

$$\boxed{\mathcal{C} = K^2 \det(WG)} \tag{57}$$

where WG is defined as in equation (14).

1.2.12 The Log_Det_Ratio index

The Log_Det_Ratio index is defined like this:

$$\boxed{\mathcal{C} = N\,\log\left(\frac{\det(T)}{\det(WG)}\right)} \tag{58}$$

where T is the scatter matrix defined in section 1.1.1 and WG is defined by equation (14). This is a logarithmic variant of the Det_Ratio index seen in section .

1.2.13 The Log_SS_Ratio index

The Log_SS_Ratio index is defined like this:

$$\boxed{\mathcal{C} = \log\left(\frac{BGSS}{WGSS}\right)} \tag{59}$$

where $BGSS$ and $WGSS$ are defined by equations (23) and (16) respectively: they are the traces of the BG and WG matrices respectively.

1.2.14 The McClain-Rao index

As for the C index seen in section 1.2.3, let us denote by S_W the sum of the within-cluster distances:

$$S_W = \sum_{(i,j)\in I_W} d(M_i, M_j) = \sum_{k=1}^{K} \sum_{\substack{i,j\in I_k \\ i<j}} d(M_i, M_j) \tag{60}$$

Recall that the total number of distances between pairs of points belonging to a same cluster is N_W.

Let us denote by S_B the sum of the between-cluster distances:

$$S_B = \sum_{(i,j)\in I_B} d(M_i, M_j) = \sum_{k<k'} \sum_{\substack{i\in I_k,\, j\in I_{k'} \\ i<j}} d(M_i, M_j) \tag{61}$$

The total number of distances between pairs of points which do not belong to the same cluster is $N_B = N(N-1)/2 - N_W$.

The McClain-Rao index is defined as the quotient between the mean within-cluster and between-cluster distances:

$$\boxed{\mathcal{C} = \frac{S_W/N_W}{S_B/N_B} = \frac{N_B}{N_W}\frac{S_W}{S_B}} \tag{62}$$

1.2.15 The PBM index

The PBM index (acronym constituted of the initals of the names of its authors, Pakhira, Bandyopadhyay and Maulik) is calculated using the distances between the points and their barycenters and the distances between the barycenters themselves.

Let us denote by D_B the largest distance between two cluster barycenters:

$$D_B = \max_{k<k'} d\big(G^{\{k\}}, G^{\{k'\}}\big) \tag{63}$$

On the other hand, let us denote by E_W the sum of the distances of the points of each cluster to their barycenter and E_T the sum of the distances of all the points to the barycenter G of the entire data set:

$$E_W = \sum_{k=1}^{K}\sum_{i\in I_k} d(M_i, G^{\{k\}}) \tag{64}$$

$$E_T = \sum_{i=1}^{N} d(M_i, G) \tag{65}$$

The PBM index is defined like this:

$$\boxed{\mathcal{C} = \left(\frac{1}{K} \times \frac{E_T}{E_W} \times D_B\right)^2} \tag{66}$$

E_T is a constant which does not depend on the partition, nor on the number of clusters.

1.2.16 The Point-Biserial index

Generally speaking, in statistics, the *point-biserial* coefficient is a correlation measure between a continuous variable A and a binary variable B (i-e a variable whose values are 0 or 1). A and B are sets with the same length n.

The values of A are dispatched into two groups A_0 and A_1 depending on the corresponding value in B being 0 or 1.

Let us denote by M_{A_0} and M_{A_1} the means in A_0 and A_1, and n_{A_0} and n_{A_1} the number of elements in each group. The point-biserial correlation coefficient is defined as the quantity:

$$r_{pb}(A,B) = \frac{M_{A_1} - M_{A_0}}{s_n}\sqrt{\frac{n_{A_0} n_{A_1}}{n^2}} \tag{67}$$

where s_n is the standard deviation of A.

In the context of a comparison between different clusterings, the term s_n may be omitted because it does not depend on the partitions but only on the set of data.

As in the case of the Γ index seen in section 1.2.8, one adapts this definition by choosing A to be the set of the N_T distances between pairs of points M_i and M_j. The corresponding value in B is 1 if the two points lie in the same cluster and 0 otherwise:

$$A_{ij} = d(M_i, M_j) \tag{68}$$

$$B_{ij} = \begin{cases} 1 & \text{if } (i,j) \in I_W \\ 0 & \text{otherwise} \end{cases} \tag{69}$$

M_{A_1} is the mean of all the within-cluster distances and M_{A_0} is the mean of all the between-cluster distances.

Using the notations introduced in section 1.2.14, the definition of the point-biserial index is :

$$\boxed{\mathcal{C} = s_n \times r_{pb}(A,B) = (S_W/N_W - S_B/N_B)\frac{\sqrt{N_W N_B}}{N_T}} \tag{70}$$

1.2.17 The Ratkowsky-Lance index

One computes the mean $\bar{R}$ of the quotients between BGSS and TSS for each dimension of the data, that is to say for each column of the matrix A.

Let us denote

$$BGSS_j = \sum_{k=1}^{K} n_k (\mu_j^{\{k\}} - \mu_j)^2 = b_{jj} \tag{71}$$

$$TSS_j = N\mathrm{Var}(V_j) = \sum_{i=1}^{N} (a_{ij} - \mu_j)^2 \tag{72}$$

Then

$$\bar{c}^2 = \bar{R} = \frac{1}{p}\sum_{j=1}^{p} \frac{BGSS_j}{TSS_j} \tag{73}$$

$BGSS_j$ is in fact the j-th diagonal term of the matrix BG defined by equation (20).

The Ratkowsky_Lance index ($\bar{c}/\sqrt{K}$) is defined like this:

$$\boxed{\mathcal{C} = \sqrt{\frac{\bar{R}}{K}} = \frac{\bar{c}}{\sqrt{K}}} \tag{74}$$

1.2.18 The Ray-Turi index

The Ray-Turi index is defined as a quotient:

- the numerator is the mean of the squared distances of all the points with respect to the barycenter of the cluster they belong to:

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i \in I_k} ||M_i^{\{k\}} - G^{\{k\}}||^2 = \frac{1}{N}\sum_{k=1}^{K} WGSS^{\{k\}} = \frac{1}{N} WGSS$$

- the denominator is the minimum of the squared distances $\Delta_{kk'}$ between all the cluster barycenters:

$$\min_{k<k'} \Delta_{kk'}^2 = \min_{k<k'} d(G^{\{k\}}, G^{\{k'\}})^2 = \min_{k<k'} ||G^{\{k\}} - G^{\{k'\}}||^2 \tag{75}$$

So the Ray-Turi index can be written like this:

$$\boxed{\mathcal{C} = \frac{1}{N} \frac{WGSS}{\min\limits_{k<k'} \Delta_{kk'}^2}} \tag{76}$$

1.2.19 The Scott-Symons index

This index is the weighted sum of the logarithms of the determinants of the variance-covariance matrix of each cluster.

It can be written like this:

$$\boxed{\mathcal{C} = \sum_{k=1}^{K} n_k \log\det\left(\frac{WG^{\{k\}}}{n_k}\right)} \tag{77}$$

The determinants of the matrices $WG^{\{k\}}$ are greater than or equal to 0 because these matrices are positive semi-definite. If one of them is equal to 0, the index is undefined.

1.2.20 The SD index

One defines two quantities $\mathcal{S}$ and $\mathcal{D}$ called respectively the *average scattering for clusters* and the *total separation between clusters*.

The average scattering for the clusters, noted $\mathcal{S}$, is defined as follows. Let us consider the vector of variances for each variable in the data set. It is a vector $\mathcal{V}$ of size p defined by:

$$\mathcal{V} = (\mathrm{Var}(V_1), \ldots, \mathrm{Var}(V_p)) \tag{78}$$

Similarly, one defines variance vectors $\mathcal{V}^{\{k\}}$ for each cluster C_k:

$$\mathcal{V}^{\{k\}} = (\mathrm{Var}(V_1^{\{k\}}), \ldots, \mathrm{Var}(V_p^{\{k\}})). \tag{79}$$

The quantity $\mathcal{S}$ is the mean of the norms of the vectors $\mathcal{V}^{\{k\}}$ divided by the norm of vector $\mathcal{V}$:

$$\mathcal{S} = \frac{\dfrac{1}{K}\displaystyle\sum_{k=1}^{K} ||\mathcal{V}^{\{k\}}||}{||\mathcal{V}||}. \tag{80}$$

On the other hand, the total separation between clusters, noted $\mathcal{D}$, is defined as follows. Let us denote by D_{max} and D_{min} respectively the largest and the smallest distance between the barycenters of the clusters:

$$D_{max} = \max_{k \neq k'} ||G^{\{k\}} - G^{\{k'\}}|| \tag{81}$$

$$D_{min} = \min_{k \neq k'} ||G^{\{k\}} - G^{\{k'\}}|| \tag{82}$$

Let us denote

$$\mathcal{D} = \frac{D_{max}}{D_{min}} \sum_{k=1}^{K} \frac{1}{\displaystyle\sum_{\substack{k'=1 \\ k' \neq k}}^{K} ||G^{\{k\}} - G^{\{k'\}}||} \tag{83}$$

The SD index is finally defined like this:

$$\boxed{\mathcal{C} = \alpha\mathcal{S} + \mathcal{D}} \tag{84}$$

where α is a weight equal to the value of $\mathcal{D}$ obtained for the partition with the greatest number of clusters. In order to compare several partitions of the data, one must first calculate the value of $\mathcal{D}$ corresponding to the greatest number of clusters in order to find the value of the coefficient α and then calculate the other indices based on this coefficient.

1.2.21 The S_Dbw index

This index relies on the notion of density of points belonging to two clusters. One first defines a limit value σ equal to the square root of the sum of the norms of the variance vectors $\mathcal{V}^{\{k\}}$ (introduced in section 1.2.20) divided by the number of clusters:

$$\sigma = \frac{1}{K}\sqrt{\sum_{k=1}^{K} ||\mathcal{V}^{\{k\}}||} \tag{85}$$

The density $\gamma_{kk'}$ for 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. Geometrically, this amounts to considering the ball with radius σ centered at the given point and counting the number of points of $C_k \cup C_{k'}$ located in this ball.

For each pair of clusters, let us evaluate the densities for the barycenters $G^{\{k\}}$ and $G^{\{k'\}}$ of the clusters and for their midpoint $H_{kk'}$. One forms the quotient $R_{kk'}$ between the density at the midpoint and the largest density at the two barycenters:

$$R_{kk'} = \frac{\gamma_{kk'}(H_{kk'})}{\max\left(\gamma_{kk'}(G^{\{k\}}),\, \gamma_{kk'}(G^{\{k'\}})\right)} \tag{86}$$

On the other hand, one defines a between-cluster density $\mathcal{G}$ as the mean of the quotients $R_{kk'}$:

$$\mathcal{G} = \frac{2}{K(K-1)} \sum_{k<k'} R_{kk'} \tag{87}$$

The S-Dbw index is defined as the sum of the mean dispersion in the clusters $\mathcal{S}$ (defined in section 1.2.20) and of the between-cluster density $\mathcal{G}$:

$$\boxed{\mathcal{C} = \mathcal{S} + \mathcal{G}} \tag{88}$$

1.2.22 The Silhouette index

Let us consider, for each point M_i, its mean distance to each cluster. One defines the within-cluster mean distance $a(i)$ as the mean distance of point M_i to the other points of the cluster it belongs to: if $M_i \in C_k$, we thus have

$$a(i) = \frac{1}{n_k - 1} \sum_{\substack{i' \in I_k \\ i' \neq i}} d(M_i, M_{i'}) \tag{89}$$

On the other hand, let us evaluate the mean distance $\mathfrak{d}(M_i, C_{k'})$ of M_i to the points of each of the other clusters $C_{k'}$:

$$\mathfrak{d}(M_i, C_{k'}) = \frac{1}{n_{k'}} \sum_{i' \in I_{k'}} d(M_i, M_{i'}) \tag{90}$$

Let us also denote by $b(i)$ the smallest of these mean distances:

$$b(i) = \min_{k' \neq k} \mathfrak{d}(M_i, C_{k'}) \tag{91}$$

The value k' which realizes this minimum indicates the best choice for reaffecting, if necessary, the point M_i to another cluster than the one it currently belongs to.

For each point M_i, one then forms the quotient

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \tag{92}$$

which is called the silhouette width of the point. It is a quantity between -1 and 1: a value near 1 indicates that the point M_i is affected to the right cluster whereas a value near -1 indicates that the point should be affected to another cluster.

The mean of the silhouette widths for a given cluster C_k is called the cluster mean silhouette and is denoted as $\mathfrak{s}_k$:

$$\mathfrak{s}_k = \frac{1}{n_k} \sum_{i \in I_k} s(i) \tag{93}$$

Finally, the global silhouette index is the mean of the mean silhouettes through all the clusters:

$$\boxed{\mathcal{C} = \frac{1}{K} \sum_{k=1}^{K} \mathfrak{s}_k} \tag{94}$$

1.2.23 The Tau index

Using the same notations as for the Gamma index in section 1.2.8, the τ index of Kendall between two vectors of data of length N_T is classically defined in statistics as the quantity:

$$\tau = \frac{s^+ - s^-}{\dfrac{N_T(N_T - 1)}{2}} \tag{95}$$

The numbers s^+ and s^- do not count ties, so if a between-cluster distance and a within-cluster distance are equal, they do not enter in the numerator. In order to take ties into account, one modifies the denominator and defines the corrected index τ_c like this:

$$\tau_c = \frac{s^+ - s^-}{\sqrt{(\nu_0 - \nu_1)(\nu_0 - \nu_2)}} \tag{96}$$

with

$$\nu_0 = \frac{N_T(N_T-1)}{2} \tag{97}$$

$$\nu_1 = \sum_i \frac{t_i(t_i-1)}{2} \tag{98}$$

$$\nu_2 = \sum_j \frac{u_j(u_j-1)}{2} \tag{99}$$

where t_i is the number of values in the i-th group of ties for the vector A and u_j is the number of values in the j-th group of ties for the vector B. Here the vector B is constituted only of values 0 and 1 (corresponding to the between-cluster and within-cluster pairs respectively) and we thus have:

$$\nu_2 = N_B(N_B-1)/2 + N_W(N_W-1)/2 \tag{100}$$

An easy calculation shows that $\nu_0 - \nu_2 = N_B N_W$.

If one makes the reasonable hypothesis that the vector A contains few identical values, one can estimate that ν_2 is negligible with respect to ν_0. This justifies the following definition of the Tau index of clustering:

$$\boxed{\mathcal{C} = \frac{s^+ - s^-}{\sqrt{N_B N_W \left(\frac{N_T(N_T-1)}{2}\right)}}} \tag{101}$$

1.2.24 The Trace_W index

The Trace_W index is defined like this:

$$\boxed{\mathcal{C} = \mathrm{Tr}(WG) = WGSS} \tag{102}$$

where WG and $WGSS$ are defined by equations (14) and (16) respectively.

1.2.25 The Trace_WiB index

The Trace_WiB (or Trace_$W^{-1}B$) index is defined like this:

$$\boxed{\mathcal{C} = \mathrm{Tr}(WG^{-1}\,.\,BG)} \tag{103}$$

where WG and BG are defined by equations (14) and (20) respectively.

1.2.26 The Wemmert-Gançarski index

The Wemmert-Gançarski index is built using quotients of distances between the points and the barycenters of all the clusters.

For a point M belonging to cluster C_k, one forms the quotient $R(M)$ between the distance of this point to the barycenter of the cluster it belongs to and the smallest distance of this point to the barycenters of all the other clusters:

$$R(M) = \frac{||M - G^{\{k\}}||}{\min\limits_{k' \neq k} ||M - G^{\{k'\}}||} \tag{104}$$

One then takes the mean of these quotients in each cluster. If this mean is greater than 1, it is ignored, otherwise one takes its complement to 1. Precisely, let us define:

$$J_k = \max\left\{0, 1 - \frac{1}{n_k}\sum_{i \in I_k} R(M_i)\right\} \tag{105}$$

The Wemmert-Gançarski index is defined as the weighted mean, for all the clusters, of the quantities J_k like this:

$$\boxed{\mathcal{C} = \frac{1}{N}\sum_{k=1}^{K} n_k J_k} \tag{106}$$

This expression can be rewritten as follows:

$$\mathcal{C} = \frac{1}{N}\sum_{k=1}^{K} \max\left\{0, n_k - \sum_{i \in I_k} R(M_i)\right\} \tag{107}$$

1.2.27 The Xie-Beni index

The Xie-Beni index is an index of fuzzy clustering, but it is also applicable to crisp clustering.

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.

The mean quadratic error, in the case of a crisp clustering, is simply the quantity $\frac{1}{N}WGSS$, in other words the mean of the squared distances of all the points with respect to the barycenter of the cluster they belong to.

Using the same notation as in section 1.2.9, one has

$$\delta_1(C_k, C_{k'}) = \min_{\substack{i \in I_k \\ j \in I_{k'}}} d(M_i, M_j) \tag{108}$$

and the Xie-Beni index can be written like this:

$$\boxed{\mathcal{C} = \frac{1}{N}\frac{WGSS}{\min\limits_{k<k'} \delta_1(C_k, C_{k'})^2}} \tag{109}$$

1.3 Choice of the best partition

In order to find the best partition of the data, one usually executes a clustering algorithm with different values of the expected number of clusters K: let us say that $K_m \leq K \leq K_M$. The clustering algorithm which is applied could be an ascending hierarchical clustering (AHC) or the k-means algorithm or any other technique. One then computes a quality index Q_K for each value of K and selects the partition which led to the "best" value for Q_K. This section explains what is considered the "best" value for the different quality indices.

Table 2 summarizes, for each index, which rule must be applied in order to determine the best index value. For instance, in the case of the Calinski-Harabasz index, if the quality index has been computed for different partitions of the data, the best partition is the one corresponding to the greatest value of the index.

Index	*Rule*
Ball_Hall	*max diff*
Banfeld_Raftery	*min*
C_index	*min*
Calinski_Harabasz	*max*
Davies_Bouldin	*min*
Det_Ratio	*min diff*
Dunn	*max*
GDI	*max*
Gamma	*max*
G_plus	*min*
Ksq_DetW	*max diff*
Log_Det_Ratio	*min diff*
Log_SS_Ratio	*min diff*
McClain_Rao	*min*
PBM	*max*
Point_biserial	*max*
Ratkowsky_Lance	*max*
Ray_Turi	*min*
Scott_Symons	*min*
SD	*min*
S_Dbw	*min*
Silhouette	*max*
Tau	*max*
Trace_W	*max diff*
Trace_WiB	*max diff*
Wemmert_Gancarski	*max*
Xie_Beni	*min*

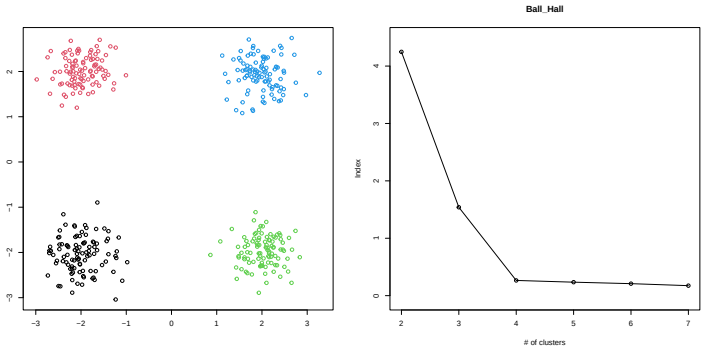
Table 2: Method to determine the best partition.

The decision rules called *max* and *min* in table 2 mean that one should select respectively the greatest or the smallest index value.

The decision rule called *max diff* means that the best value for K is the one corresponding to the greatest difference between two successive slopes. On a diagram representing the index values against the number of selected clusters, this corresponds to an elbow. More precisely, let us denote $V_i = Q_{i+1} - Q_i$ the slope between to successive points of the diagram. Then K is defined by:

$$K = \arg \max_{K_m < i \leq K_M} (V_i - V_{i-1}) \tag{110}$$

This is better explained on the following graphic. The figure on the right displays the values of the Hall_Ball index corresponding to different clusterings of the data represented by the figure on the left. The index has been computed with tentative partitions made of 2 to 7 clusters. The figure exhibits an elbow for the four-clusters partition and indeed the data clearly belong to four distinct groups.



2 External comparison indices

The external indices of comparison are indices designed to measure the similitude between two partitions. They take into account only the distribution of the points in the different clusters and do not allow to measure the quality of this distribution.

2.1 Notation

All the suggested indices rely on a confusion matrix representing the count of pairs of points depending on whether they are considered as belonging to the same cluster or not according to partition P_1 or to partition P_2. There are thus four possibilities:

- the two points belong to the same cluster, according to both P_1 and P_2
- the two points belong to the same cluster according to P_1 but not to P_2
- the two points belong to the same cluster acording to P_2 but not to P_1
- the two points do not belong to the same cluster, according to both P_1 and P_2.

Let us denote by yy, yn, ny, nn (y means *yes*, and n means *no*) the number of points belonging to these four categories respectively. N_T being the total number of pairs of points, one has:

$$N_T = \frac{N(N-1)}{2} = yy + yn + ny + nn. \tag{111}$$

2.2 Precision and recall coefficients

If partition P_1 is used as a reference, one defines the *precision coefficient* as the proportion of points rightly grouped together in P_2, that is to say which are also grouped together according to the reference partition P_1. Among the $yy + ny$ points grouped together according to P_2, yy are rightly grouped. One thus has:

$$\mathcal{P} = \frac{yy}{yy + ny}. \tag{112}$$

Similarly, one defines the *recall coefficient* as the proportion of points grouped together in P_1 which are also grouped together in partition P_2. This is the proportion of points which are supposed to be grouped together according to the reference partition P_1 and which are effectively marked as such by partition P_2. Among the $yy+yn$ points grouped together in P_1, yy are also grouped together in P_2. One thus has:

$$\mathcal{R} = \frac{yy}{yy + yn} \tag{113}$$

In terms of conditional probabilities, one can write

$$\mathcal{P} = P(gp_1|gp_2) \quad \text{and} \quad \mathcal{R} = P(gp_2|gp_1) \tag{114}$$

where the events gp_1 and gp_2 mean that two points are grouped together in P_1 and in P_2 respectively.

The $\mathcal{F}$-measure is the harmonic mean of the precision and recall coefficients:

$$\mathcal{F} = \frac{2}{\frac{1}{\mathcal{P}} + \frac{1}{\mathcal{R}}} = \frac{2\mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}} = \frac{2yy}{2yy + yn + ny} \tag{115}$$

There is also a weighted version of this measure, called the $\mathcal{F}_\alpha$-measure, defined like this:

$$\mathcal{F}_\alpha = \frac{(1+\alpha)\mathcal{P} \times \mathcal{R}}{\alpha\mathcal{P} + \mathcal{R}} \text{ with } \alpha > 0 \tag{116}$$

2.3 Indicator variables

Let us associate to each partition P_a ($a = 1, 2$) the binary random variable X_a defined on the set of indices i and j such that $i < j$ as follows: its value is 1 if the points M_i and M_j are classified in the same cluster than in partition P_a and 0 otherwise. The variable X_a works as an indicator variable.

There are N_T pairs of points and one is interested only in the indices i and j such that $i < j$. Let us consider the mean and the standard deviation of X_a:

$$\mu_{X_a} = \frac{1}{N_T} \sum_{i<j} X_a(i,j) \tag{117}$$

$$\sigma^2_{X_a} = \frac{1}{N_T} \sum_{i<j} X_a(i,j)^2 - \mu^2_{X_a} \tag{118}$$

The following formulas establish a link between these random variables and the

concordant and discordant count variables:

$$yy + yn = \sum_{i<j} X_1(i,j) \tag{119}$$

$$yy + ny = \sum_{i<j} X_2(i,j) \tag{120}$$

$$yy = \sum_{i<j} X_1(i,j)X_2(i,j) \tag{121}$$

From this we get:

$$\mu_{X_1} = \frac{yy+yn}{N_T} \qquad \sigma^2_{X_1} = \frac{yy+yn}{N_T} - \left(\frac{yy+yn}{N_T}\right)^2$$
$$\mu_{X_2} = \frac{yy+ny}{N_T} \qquad \sigma^2_{X_2} = \frac{yy+ny}{N_T} - \left(\frac{yy+ny}{N_T}\right)^2$$

2.4 External indices definition

The following sections give the definition of several (more or less) widely used external indices.

2.4.1 The Czekanowski-Dice index

The Czekanowski-Dice index (aka the Ochiai index) is defined like this:

$$\boxed{\mathcal{C} = \frac{2yy}{2yy + yn + ny}} \tag{122}$$

This index is the harmonic mean of the precision and recall coefficients, that is to say it is identical to the $\mathcal{F}$-measure defined in section 2.2:

$$\mathcal{C} = 2\frac{\mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{123}$$

2.4.2 The Folkes-Mallows index

The Folkes-Mallows index is defined like this:

$$\boxed{\mathcal{C} = \frac{yy}{\sqrt{(yy+yn) \times (yy+ny)}}} \tag{124}$$

This index is the geometric mean of the precision and recall coefficients:

$$\mathcal{C} = \sqrt{\mathcal{P}\mathcal{R}} \tag{125}$$

2.4.3 The Hubert $\hat{\Gamma}$ index

The index of Hubert $\hat{\Gamma}$ is the correlation coefficient of the indicator variables introduced in section 2.3. It is defined like this:

$$\boxed{\mathcal{C} = \mathrm{Corr}(X_1, X_2) = \frac{\sum_{i<j} \big(X_1(i,j) - \mu_{X_1}\big)\big(X_2(i,j) - \mu_{X_2}\big)}{N_T\, \sigma_{X_1} \sigma_{X_2}}} \tag{126}$$

Comparing with equation (136), the index of Hubert $\hat{\Gamma}$ appears as a standardized variant (centered and reduced) of the Russel-Rao index defined in section 2.4.10. Its value is between -1 and 1.

Using the relations of section 2.3, one may write the $\hat{\Gamma}$ index as follows:

$$\mathcal{C} = \frac{N_T \times yy - (yy + yn)(yy + ny)}{\sqrt{(yy + yn)(yy + ny)(nn + yn)(nn + ny)}} \tag{127}$$

2.4.4 The Jaccard index

The Jaccard index is defined like this:

$$\boxed{\mathcal{C} = \frac{yy}{(yy + yn + ny)}} \tag{128}$$

2.4.5 The Kulczynski index

The Kulczynski index is defined like this:

$$\boxed{\mathcal{C} = \frac{1}{2}\left(\frac{yy}{yy + ny} + \frac{yy}{yy + yn}\right)} \tag{129}$$

This index is the arithmetic mean of the precision and recall coefficients:

$$\mathcal{C} = \frac{1}{2}(\mathcal{P} + \mathcal{R}) \tag{130}$$

2.4.6 The McNemar index

The McNemar index is defined like this:

$$\boxed{\mathcal{C} = \frac{yn - ny}{\sqrt{yn + ny}}} \tag{131}$$

Under the null hypothesis H_0 that the discordances between the partitions P_1 and P_2 are random, the index $\mathcal{C}$ follows approximatively a normal distribution. It is an adaptation of the non-parametric test of McNemar for the comparison of frequencies between two paired samples: the statistic of McNemar's test (called the χ^2 distance) is the square of the index

$$\mathcal{C}^2 = \frac{(yn - ny)^2}{yn + ny}$$

and follows, under the null hypothesis of marginal homogeneity of the contingency table, a χ^2 distribution with 1 degree of freedom.

2.4.7 The Phi index

The Phi index is a classical measure of the correlation between two dichotomic variables. It is defined like this:

$$\boxed{\mathcal{C} = \frac{yy \times nn - yn \times ny}{(yy + yn)(yy + ny)(yn + nn)(ny + nn)}} \tag{132}$$

2.4.8 The Rand index

The Rand index is defined like this:

$$\boxed{\mathcal{C} = \frac{yy + nn}{N_T}} \tag{133}$$

2.4.9 The Rogers-Tanimoto index

The Rogers-Tanimoto index is defined like this:

$$\boxed{\mathcal{C} = \frac{yy + nn}{yy + nn + 2(yn + ny)}} \tag{134}$$

2.4.10 The Russel-Rao index

The Russel-Rao index measures the proportion of concordances between the two partitions. It is defined like this:

$$\boxed{\mathcal{C} = \frac{yy}{N_T}} \tag{135}$$

Using the notations introduced in section 2.3, this index can be written:

$$\mathcal{C} = \frac{1}{N_T} \sum_{i<j} X_1(i,j) X_2(i,j) \tag{136}$$

2.4.11 The Sokal-Sneath indices

There are two versions of the Sokal-Sneath index. They are defined respectively like this:

$$\boxed{\begin{array}{rcl} \mathcal{C}_1 & = & \dfrac{yy}{yy + 2(yn + ny)} \\ \mathcal{C}_2 & = & \dfrac{yy + nn}{yy + nn + \frac{1}{2}(yn + ny)} \end{array}} \tag{137}$$

3 Usage of the *clusterCrit* package

The *clusterCrit* package for R provides an implementation of all the indices described in the preceding sections. The core of the package is written in Fortran and is optimized in order to avoid duplicate calculations.

It can be installed from the R console with the following instruction:

```
install.packages(clusterCrit)
```

Once it is installed, it can be loaded in an R session with the following instruction:

```
load(clusterCrit)
```

3.1 Available commands

The *clusterCrit* package defines several functions which let you compute internal quality indices or external comparison indices. The partitions are specified as an integer vector giving the index of the cluster each observation belongs to. The possible values are integers between 1 and K, where K is the number of clusters.

The *intCriteria* function calculates one or several internal quality indices. Its syntax is:

```
intCriteria(traj, part, crit)
```

The *traj* argument is the matrix of observations (aka as trajectories). The *part* argument is the partition vector. The *crit* argument is a list containing the names of the indices to compute. One can use the keyword `"all"` in order to compute all the available indices. See the *getCriteriaNames* function to see the names of the currently available indices. All the names are case insensitive and can be abbreviated as long as the abbreviation remains unambiguous.

The *extCriteria* function calculates one or several external indices (including the precision and recall coefficients). Its syntax is:

```
extCriteria(part1, part2, crit)
```

The *part1* and *part2* arguments are the partition vectors. The meaning of the *crit* argument is the same as for the *intCriteria* function.

Given a vector of several clustering quality index values computed with a given criterion, the function *bestCriterion* returns the index of the one which must be considered as the best in the sense of the specified criterion. Its syntax is:

```
bestCriterion(x, crit)
```

The *x* argument is a numeric vector of quality index values. The *crit* argument is the name of the criterion: it is case insensitive and can be abbreviated.

Typically, a set of data is clusterized several times (using different algorithms or specifying a different number of clusters) and a clustering index is calculated each time: the *bestCriterion* function tells which value is considered the best. For instance, if one uses the Calinski_Harabasz index, the best value is the largest one.

The *concordance* function calculates the concordance matrix between two partitions of the same data. Its syntax is:

```
concordance(part1, part2)
```

The arguments are the partition vectors. The function returns a 2×2 matrix of the form:

$$\begin{pmatrix} yy & yn \\ ny & nn \end{pmatrix}$$

These are the number of pairs classified as belonging or not belonging to the same cluster with respect to both partitions. Since there are $N(N-1)/2$ pairs of distinct points, one has:

$$yy + yn + ny + nn = N(N-1)/2$$

The *getCriteriaNames* function is a convenience function which returns the names of the currently implemented indices. Its syntax is:

```
getCriteriaNames(isInternal)
```

where the argument *isInternal* is a logical value: if TRUE it returns the names of the internal indices, otherwise it returns the names of the external ones.

3.2 Examples of use

First load the package:

```
> library(clusterCrit)
```

Let us create some artificial data:

```
> x <- rbind(matrix(rnorm(100, mean = 0, sd = 0.5), ncol = 2),
+            matrix(rnorm(100, mean = 1, sd = 0.5), ncol = 2),
+            matrix(rnorm(100, mean = 2, sd = 0.5), ncol = 2))
```

Now perform the *kmeans* algorithm in order to get a partition with 3 clusters (the *kmeans* function is provided by R in the *stats* package and is available by default):

```
> cl <- kmeans(x, 3)
```

Let us get the names of the internal indices:

```
> getCriteriaNames(TRUE)
```

```
 [1] "Ball_Hall"         "Banfeld_Raftery"   "C_index"
 [4] "Calinski_Harabasz" "Davies_Bouldin"    "Det_Ratio"
 [7] "Dunn"              "Gamma"             "G_plus"
[10] "GDI11"             "GDI12"             "GDI13"
[13] "GDI21"             "GDI22"             "GDI23"
[16] "GDI31"             "GDI32"             "GDI33"
[19] "GDI41"             "GDI42"             "GDI43"
[22] "GDI51"             "GDI52"             "GDI53"
[25] "Ksq_DetW"          "Log_Det_Ratio"     "Log_SS_Ratio"
[28] "McClain_Rao"       "PBM"               "Point_Biserial"
[31] "Ray_Turi"          "Ratkowsky_Lance"   "Scott_Symons"
[34] "SD_Scat"           "SD_Dis"            "S_Dbw"
[37] "Silhouette"        "Tau"               "Trace_W"
[40] "Trace_WiB"         "Wemmert_Gancarski" "Xie_Beni"
```

Let us compute all the internal indices and display one of them:

```
> intIdx <- intCriteria(x,cl$cluster,"all")
> length(intIdx)

[1] 42

> intIdx[["trace_w"]]

[1] 62.88847
```

It is possible to compute only a few indices:

```
> intCriteria(x,cl$cluster,c("C_index","Calinski_Harabasz","Dunn"))
```

```
$c_index
[1] 0.07132942

$calinski_harabasz
[1] 238.3786

$dunn
[1] 0.06448147
```

The names are case insensitive and can be abbreviated:

```
> intCriteria(x,cl$cluster,c("det","cal","dav"))

$det_ratio
[1] 9.306902

$calinski_harabasz
[1] 238.3786

$davies_bouldin
[1] 0.8213292
```

Here is now an example of the external criteria. Let us generate two artificial partitions:

```
> part1<-sample(1:3,150,replace=TRUE)
> part2<-sample(1:5,150,replace=TRUE)
```

Let us get the names of the external indices:

```
> getCriteriaNames(FALSE)

 [1] "Czekanowski_Dice" "Folkes_Mallows"   "Hubert"           "Jaccard"
 [5] "Kulczynski"       "McNemar"          "Phi"              "Precision"
 [9] "Rand"             "Recall"           "Rogers_Tanimoto"  "Russel_Rao"
[13] "Sokal_Sneath1"    "Sokal_Sneath2"
```

Let us compute all the external indices and retrieve one of them:

```
> extIdx <- extCriteria(part1,part2,"all")
> length(extIdx)

[1] 14

> extIdx[["jaccard"]]

[1] 0.1411675
```

Let us compute only some of them:

```
> extCriteria(part1,part2,c("Rand","Folkes"))

$rand
[1] 0.5971364

$folkes_mallows
[1] 0.2551908
```

The names are case insensitive and can be abbreviated:

```
> extCriteria(part1,part2,c("ra","fo"))

$rand
[1] 0.5971364

$folkes_mallows
[1] 0.2551908
```

3.3 Benchmark

The *clusterCrit* package is written in Fortran which makes the calculations quite fast. Nevertheless some indices are more demanding and require more computations than the others. The following timings have been evaluated using the *rbenchmark* package: the various indices have been computed separately 100 times on a set of 400 points partitionned in four groups. The results are not interesting *per se* but rather to compare the amount of computations required by the different indices.

The following table summarizes the timings for the internal indices (they are expressed in seconds for 100 replications, so they must all be divided by 100):

all	3.095
Ball_Hall	0.944
Banfeld_Raftery	0.946
C_index	2.898
Calinski_Harabasz	0.930
Davies_Bouldin	0.926
Det_Ratio	0.930
Dunn	0.969
Gamma	2.188
G_plus	2.170
GDI11	0.985
GDI12	0.971
GDI13	0.957
GDI21	0.966
GDI22	0.961
GDI23	0.953
GDI31	0.959
GDI32	0.957
GDI33	0.948
GDI41	0.936
GDI42	0.933
GDI43	0.923
GDI51	0.934
GDI52	0.934
GDI53	0.921
Ksq_DetW	0.930
Log_Det_Ratio	0.930
Log_SS_Ratio	0.923
McClain_Rao	0.958
PBM	0.928
Point_Biserial	0.959
Ray_Turi	0.923
Ratkowsky_Lance	0.923
Scott_Symons	0.965
SD_Scat	0.930
SD_Dis	0.923
S_Dbw	0.924
Silhouette	0.992
Tau	2.174
Trace_W	0.945
Trace_WiB	0.952
Wemmert_Gancarski	0.960
Xie_Beni	0.978

We observe that the C index is the most time consuming. The gamma, g_plus and tau indices also need intensive calculations because the concordance and discordance counts concern a huge quantity of pairs of points. All the other indices yield more or less the same values.

Using the keyword `"all"` in the *intCriteria* function is quite efficient because the

code is optimized to avoid duplicate calculations and to reuse values already computed for other indices. The timing result for calculating all the indices simultaneously 100 times is 3.095.

On the contrary, benchmarking the external indices does not exhibit any noticeable difference. They all take more or less the same time and are very fast. Here are the results for 100 replications of the *extCriteria* function applied to two partitions containing 150 items:

all	0.010
Czekanowski_Dice	0.010
Folkes_Mallows	0.010
Hubert	0.011
Jaccard	0.010
Kulczynski	0.011
McNemar	0.010
Phi	0.010
Precision	0.010
Rand	0.010
Recall	0.011
Rogers_Tanimoto	0.010
Russel_Rao	0.011
Sokal_Sneath1	0.010
Sokal_Sneath2	0.009

References

[1] F. B. Baker and L. J. Hubert. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70:31–38, 1975.

[2] G. H. Ball and D. J. Hall. Isodata: A novel method of data analysis and pattern classification. *Menlo Park: Stanford Research Institute. (NTIS No. AD 699616)*, 1965.

[3] J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[4] J. C. Bezdek and N. R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and CyberneticsÑPART B: CYBERNETICS*, 28, no. 3:301–315, 1998.

[5] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3, no. 1:1–27, 1974.

[6] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, no. 2:224–227, 1979.

[7] J. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4:95–104, 1974.

[8] A. W. F. Edwards and L. Cavalli-Sforza. A method for cluster analysis. *Biometrika*, 56:362–375, 1965.

[9] B.S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. Arnold, London, 2001.

[10] H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62:1159–1178, 1967.

[11] L. Goodman and W. Kruskal. Measures of associations for cross-validations. *J. Am. Stat. Assoc.*, 49:732–764, 1954.

[12] M. Halkidi and M. Vazirgiannis. Clustering validity assessment: Finding the optimal partitioning of a data set. *Proceedings IEEE International Conference on Data Mining*, pages 187–194, 2001.

[13] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, 2001.

[14] J. A. Hartigan. Clustering algorithms. *New York: Wiley*, 1975.

[15] L. Hubert and J. Schultz. Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychologie*, 29:190–241, 1976.

[16] F. H. B. Marriot. Practical problems in a method of cluster analysis. *Biometrics*, 27:456–460, 1975.

[17] J. O. McClain and V. R. Rao. Clustisz: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research*, 12:456–460, 1975.

[18] G. W. Milligan. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46, no. 2:187–199, 1981.

[19] Bandyopadhyay S. Pakhira M. K. and Maulik U. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37:487–501, 2004.

[20] Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[21] D. A. Ratkowsky and G. N. Lance. A criterion for determining the number of groups in a classification. *Australian Computer Journal*, 10:115–117, 1978.

[22] S. Ray and Rose H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. *in Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, pages 137–143, 1999.

[23] F. J. Rohlf. Methods of comparing classifications. *Annual Review of Ecology and Systematics*, 5:101–113, 1974.

[24] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

[25] X.L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):841–846, 1991.