Comparing Partitions through the Matching Error

Mathias Bourel

Joint work with Badih Ghattas (Aix-Marseille) and Meliza González

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

August 27, 2019

Introduction

Most clustering approaches result in a partition of the data set and often a partition of the space where the data lie.

Validation:

- **Internal validation:** measure the clustering output internal coherency (only other the information of the data, and can be used to obtain optimal number of groups)
- **External validation:** Comparison of different methods of clustering (in particular, the clustering output with the true classification, i.e over a supervised data set).

This talk focuses other the latter case, which could be extended to the framework of comparing two partitions of a set.

Several indices may be used to compare partitions coming from a same data set, among which the Rand index (Rand (1971)), the Adjusted Rand Index (Hubert and Arabie (1985)), the Jaccard Index (Hultsch (2004)), etc. Most of these existing indices lack real mathematical analysis, and almost no information exists about their distribution.

1. Related Works
   - Measures based on counting pairs
   - Measurements based on set overlaps
   - Measures based on counting pairs

2. The Matching Error

3. Hypothesis Test

Plan

External Validation

External measures use indices We can divide comparison indices in three big groups (Wagner and Wagner (2007)):

1. Measures based on counting pairs
2. Measurements based on set overlaps
3. Measures based on mutual information

We denote by $\mathcal{L} = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^p$ a sample of n independent realizations of a multivariate random variable $X = (X_1, \ldots, X_p)$. Clustering seeks to form disjoint subgroups of observations such that individuals within the same cluster are similar to each other and relatively different from those of the other clusters. Let $\mathscr{C}$ be a partition of $\mathcal{L}$ obtained by a cluster analysis, that is $\mathscr{C}$ is a collection of disjoint subsets $\{C_1, \ldots, C_J\}$ such that their union is $\mathcal{L}$. The set of all possible partitions of $\mathcal{L}$ is denoted $\mathcal{P}(\mathcal{L})$. Let $\mathscr{C}' = \{C'_1, \ldots, C'_L\} \in \mathcal{P}(\mathcal{L})$ be a second partition of $\mathcal{L}$. The number of clusters of partitions $\mathscr{C}$ and $\mathscr{C}'$ (J and L respectively) may be different.

Measures based on counting pairs

A natural way to compare partitions is by counting pairs of observations belonging to a same cluster in both partitions.

- The confusion matrix:

$\mathscr{C}/\mathscr{C}'$	C_1'	C_2	$\ldots$	C_L'	Suma
C_1	n_{11}	n_{12}	$\ldots$	n_{1L}	a_1
C_2	n_{21}	n_{22}	$\ldots$	n_{2L}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
C_J	n_{J1}	n_{J2}	$\ldots$	n_{JL}	a_J
Suma	b_1	b_2	$\ldots$	b_L	$\sum_{i,j} n_{ij} = n$

  is such that $n_{ij} = \left|C_i \cap C_j'\right|$, $1 \leq i \leq J$, $1 \leq j \leq L$. We will suppose that $J \leq L$.

- The set of all (unordered) pairs of $\mathcal{L}$ is the disjoint union of the following sets:
  - A={pairs of observations that are in the same cluster in $\mathscr{C}$ and $\mathscr{C}'$}
  - B={pairs of observations that are in different clusters in $\mathscr{C}$ and $\mathscr{C}'$}
  - C={pairs of observations that are in the same cluster in $\mathscr{C}$ but in different clusters in $\mathscr{C}'$}
  - D={pairs of observations that are in the same cluster in $\mathscr{C}'$ but in different clusters in $\mathscr{C}$}

  Sets A, B, C and D are disjoint and if $a = |A|, b = |B|, c = |C|$ and $d = |D|$, where $|\cdot|$ stands for the cardinal) we have $a + b + c + d = \frac{n(n-1)}{2}$.

Measures based on counting pairs - Rand index

A very common and most used index based on counting pairs is the *Rand index* (Rand (1971)) defined by:

$$R\left(\mathscr{C},\mathscr{C}'\right)=\frac{a+b}{a+b+c+d}=\frac{2(a+b)}{n(n-1)}=\frac{\binom{n}{2}+2\sum_{i,j}\binom{n_{ij}}{2}-\left[\sum_{i}\binom{a_i}{2}+\sum_{j}\binom{b_j}{2}\right]}{\binom{n}{2}}$$

- It counts the proportion of pairs classified in a same way by the two clusterings.
- It is equal to zero when there exist no pairs of observations classified in the same way by both clustering, and it is equal to one when the two partitions are identical.
- Expectation of the index for two independent random partitions is not contant.
- As number of cluster grows, the Rand statistic approaches its upper limit of unity.
- Theoretical distribution of the indice is derived in Idrissi (2000) with restrictive conditions: if the "Rand index"[1] compares two independent partitions, with J equiprobable classes, the asymptotic distribution is Normal with expectation $\mathbb{E}(R')=1-\frac{2}{J}+\frac{2}{J^2}$ and variance $\mathbb{V}(R')=\frac{1}{n^2}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{J}+\frac{2}{J^2}\right)\left(\frac{2}{J}-\frac{2}{J^2}\right)$.
  This result is not valid for J small, especially $J=2$ and is only approximately valid for n large.

[1] Actually it is a more generalized version of Rand index, considering pairs (x_i, x_j), (x_j, x_i) and (x_1, x_i)

Measures based on counting pairs - Adjusted Rand index

Because the expected value of the Rand index of two random partitions is not constant, Hubert and Arabie (1985) proposed an adjustment based on the hypothesis that the clusterings are generated randomly subject to a fixed number of groups and fixed cluster size. The *Adjusted Rand* index is a normalized version of the Rand index and is defined as:

$$R_{adj}\left(\mathscr{C},\mathscr{C}'\right)=\frac{R\left(\mathscr{C},\mathscr{C}'\right)-\mathbb{E}(R\left(\mathscr{C},\mathscr{C}'\right))}{1-\mathbb{E}(R\left(\mathscr{C},\mathscr{C}'\right))}$$

which is equivalent to:

$$R_{adj}\left(\mathscr{C},\mathscr{C}'\right)=\frac{a-((a+d)(a+c)/(a+b+c+d))}{\frac{(a+d)+(a+c)}{2}-\frac{(a+d)(a+c)}{a+b+c+d}}$$

which has expected value zero and maximum value 1

However this index could be negative.

Measures based on counting pairs - Jaccard Index

Jaccard index (Hultsch (2004)) measures the similarity between two partitions. It is very similar to the Rand index, but it dismisses the pairs of elements that are in different clusters in the compared partitions. It is defined as

$$J(\mathscr{C}, \mathscr{C}') = \frac{a}{a + c + d}$$

Measurements based on set overlaps

Meilă and Heckerman (2001) introduced an index called *the classification error* inspired from the misclassification error used in supervised learning. Consider that one of the two compared clusterings ($\mathscr{C}$ for instance) corresponds to the true labels of each observation and the other clustering ($\mathscr{C}'$) to the predicted ones. The supervised classification error may be computed for all the possible permutations of the predicted labels (in $\mathscr{C}'$), and the maximum error over all the permutations may be taken. Thus the classification error for comparing both partitions may be written as

$$CE(\mathscr{C}, \mathscr{C}') = 1 - \frac{1}{n} \max_{\sigma} \sum_{i=1}^{J} n_{i\sigma(i)} \tag{1}$$

where σ is an injective mapping of $\{1, \ldots, J\}$ into $\{1, \ldots, L\}$ (Meilă (2005))

Measures based on mutual information

The entropy of a partition $\mathscr{C}$ is defined by $H(\mathscr{C}) = -\sum_{i=1}^{J} p(i) \log_2 p(i)$ where $p(i) = |C_i| / n$ is the estimate of the probability that an element is in cluster $C_i \in \mathscr{C}$.

The *mutual information* can be used to measure the independence of two partitions $\mathscr{C}$ and $\mathscr{C}'$ and is given by:

$$I(\mathscr{C}, \mathscr{C}') = \sum_{i=1}^{J} \sum_{j=1}^{L} p(i,j) \log_2 \frac{p(i,j)}{p(i)p(j)}$$

where $p(i,j)$ is the estimate of the probability that an element belongs to cluster C_i of $\mathscr{C}$ and C'_j of $\mathscr{C}'$.

- Mutual information is a metric over the space of all clusterings, but its value is not bounded which makes it difficult to interpret.
- As $I(\mathscr{C}, \mathscr{C}') \leq \min(H(\mathscr{C}), H(\mathscr{C}'))$, other bounded indices have been proposed such as *Normalized Mutual Information* (Strehl and Ghosh (2002), Fred and Jain (2003)) where $I(\mathscr{C}, \mathscr{C}')$ is divided either by the arithmetic or the geometric mean of the clustering entropies.
- Meila (Meilă (2003)) has also proposed an index based on Mutual information called *Variation of Information*.

$$VI(\mathscr{C}, \mathscr{C}') = H(\mathscr{C}) + H(\mathscr{C}') - 2I(\mathscr{C}, \mathscr{C}')$$

Plan

The Matching Error

Let us consider a data set $\mathcal{L} = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^p$ and $\mathscr{C}$ and $\mathscr{C}' \in \mathcal{P}(\mathcal{L})$ two partitions of $\mathcal{L}$. The labels of each observation in the first partition are denoted $\{y_1, \ldots, y_n\}$ and those of the second partition $\{\hat{y}_1, \ldots, \hat{y}_n\}$, so $y_i \in \{1, \ldots, J\}$ and $\hat{y}_i \in \{1, \ldots, L\}\, \forall i = 1, \ldots, n$.

For simplicity, we assume that the two compared partitions have the same number of clusters, that is $J = L$. If S_J is the set of permutations of $\{1, \ldots, J\}$, we define the *Matching Error* (ME) as:

$$\tau = ME(\mathscr{C}, \mathscr{C}') = \min_{\sigma \in S_J} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}} \tag{2}$$

Observe that $ME(\mathscr{C}, \mathscr{C}')$ is another formulation of $CE(\mathscr{C}, \mathscr{C}')$ used in the works of Meilă and we are going to analyze its distribution.

Example

- $\mathscr{C} = \{y_1 = 1, y_2 = 1, y_3 = 2, y_4 = 2, y_5 = 2, y_6 = 3\} = \{C_1 = (x_1, x_2), C_2 = (x_3, x_4, x_5), C_3 = (x_6)\}$
- $\mathscr{C}' = \{\hat{y}_1 = 3, \hat{y}_2 = 3, \hat{y}_3 = 1, \hat{y}_4 = 1, \hat{y}_5 = 1, \hat{y}_6 = 2\} = \{C_1 = (x_3, x_4, x_5), C_2 = (x_6), C_3 = (x_1, x_2)\}$
- $\mathscr{C}'' = \{\check{y}_1 = 3, \check{y}_2 = 1, \check{y}_3 = 1, \check{y}_4 = 1, \check{y}_5 = 2, \check{y}_6 = 2\} = \{C_1 = (x_2, x_3, x_4), C_2 = (x_5, x_6), C_3 = (x_1)\}$

Permutations of S_3 are:

$\sigma_1 = id$	$\sigma_2 = (23)$	$\sigma_3 = (12)$	$\sigma_4 = (123)$	$\sigma_5 = (132)$	$\sigma_6 = (13)$
$1 \longrightarrow 1$	$1 \longrightarrow 1$	$1 \longrightarrow 2$	$1 \longrightarrow 2$	$1 \longrightarrow 3$	$1 \longrightarrow 3$
$2 \longrightarrow 2$	$2 \longrightarrow 3$	$2 \longrightarrow 1$	$2 \longrightarrow 3$	$2 \longrightarrow 1$	$2 \longrightarrow 2$
$3 \longrightarrow 3$	$3 \longrightarrow 2$	$3 \longrightarrow 3$	$3 \longrightarrow 1$	$3 \longrightarrow 2$	$3 \longrightarrow 1$

Then:

- $\mathscr{C}$ con $\mathscr{C}'$ tenemos que $\tau_{\sigma_1} = \frac{1}{6} \sum_{i=1}^{6} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}} = 1$; $\tau_{\sigma_2} = \frac{1}{2}$; $\tau_{\sigma_3} = \frac{5}{6}$; $\tau_{\sigma_4} = \frac{4}{6}$; $\tau_{\sigma_5} = 1$; $\tau_{\sigma_6} = 0$ and $MCE(\mathscr{C}, \mathscr{C}') = \min\{\tau_{\sigma_1}, \tau_{\sigma_2}, \tau_{\sigma_3}, \tau_{\sigma_4}, \tau_{\sigma_5}, \tau_{\sigma_6}\} = 0$
- $\mathscr{C}$ con $\mathscr{C}''$, $\tau_{\sigma_1} = \frac{4}{6}$; $\tau_{\sigma_2} = \frac{4}{6}$; $\tau_{\sigma_3} = \frac{4}{6}$; $\tau_{\sigma_4} = \frac{4}{6}$; $\tau_{\sigma_5} = \frac{6}{6}$; $\tau_{\sigma_6} = \frac{2}{6}$ and $MCE(\mathscr{C}, \mathscr{C}'') = \frac{1}{3}$

The Matching Error

Let denote $\tau_\sigma = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}}$. To derive the distribution of the *ME* we need to know the distribution of the different classification errors τ_σ where $\sigma \in S_J$.
For y_i and $\hat{y}_i$ two independent realizations of a discrete random variable Y taking values in $\{1, \ldots, J\}$, let:

- $p_j = \mathbb{P}[y_i = j]$, $\hat{p}_j = \mathbb{P}[\hat{y}_i = j]$, and
- $\theta = \mathbb{P}[y_i \neq \sigma(\hat{y}_i)] = 1 - \mathbb{P}[y_i = \sigma(\hat{y}_i)] = 1 - \sum_{j=1}^{J} \mathbb{P}[y_i = j, \sigma(\hat{y}_i) = j] = 1 - \sum_{j=1}^{J} p_j \hat{p}_j$

Assuming that Y is uniform on $\{1, \ldots, J\}$, the random variable $n\tau_\sigma$ is binomial with parameters n and θ, and:

$$\mathbb{E}(\tau_\sigma) = 1 - \frac{1}{J}; \quad \text{and} \quad \mathbb{V}(\tau_\sigma) = \frac{1}{n}\frac{1}{J}\left(1 - \frac{1}{J}\right)$$

and for large values of n: $n\tau_\sigma \sim \mathcal{N}\left(n\theta, n\theta(1-\theta)\right)$.

Proposición 1

1. *As* $\sum_{j=1}^{J!} n\tau_{\sigma_j} = n(J-1)!(J-1)$, *the random variables* $n\tau_{\sigma_1}, n\tau_{\sigma_2}, \ldots, n\tau_{\sigma_{J!}}$ *are not independent.*
2. *The ME is bounded:*

$$0 \leq \tau \leq \frac{J-1}{J}$$

The Matching Error: correlations of $n\tau_{\sigma_l}$ and $n\tau_{\sigma_k}$.

Let σ_l and σ_k be two permutations of S_J. We say that σ_l and σ_k share a point $j \in \{1, \ldots, J\}$ if $\sigma_l(j) = \sigma_k(j)$. Note that two permutations of S_J can share at most $J-2$ points.

Proposición 2

If σ_j and σ_l share s points, then

$$\mathbb{COR}(n\tau_{\sigma_l}, n\tau_{\sigma_k}) = \frac{s-1}{J-1}$$

where $s = 0, \ldots, J-2$

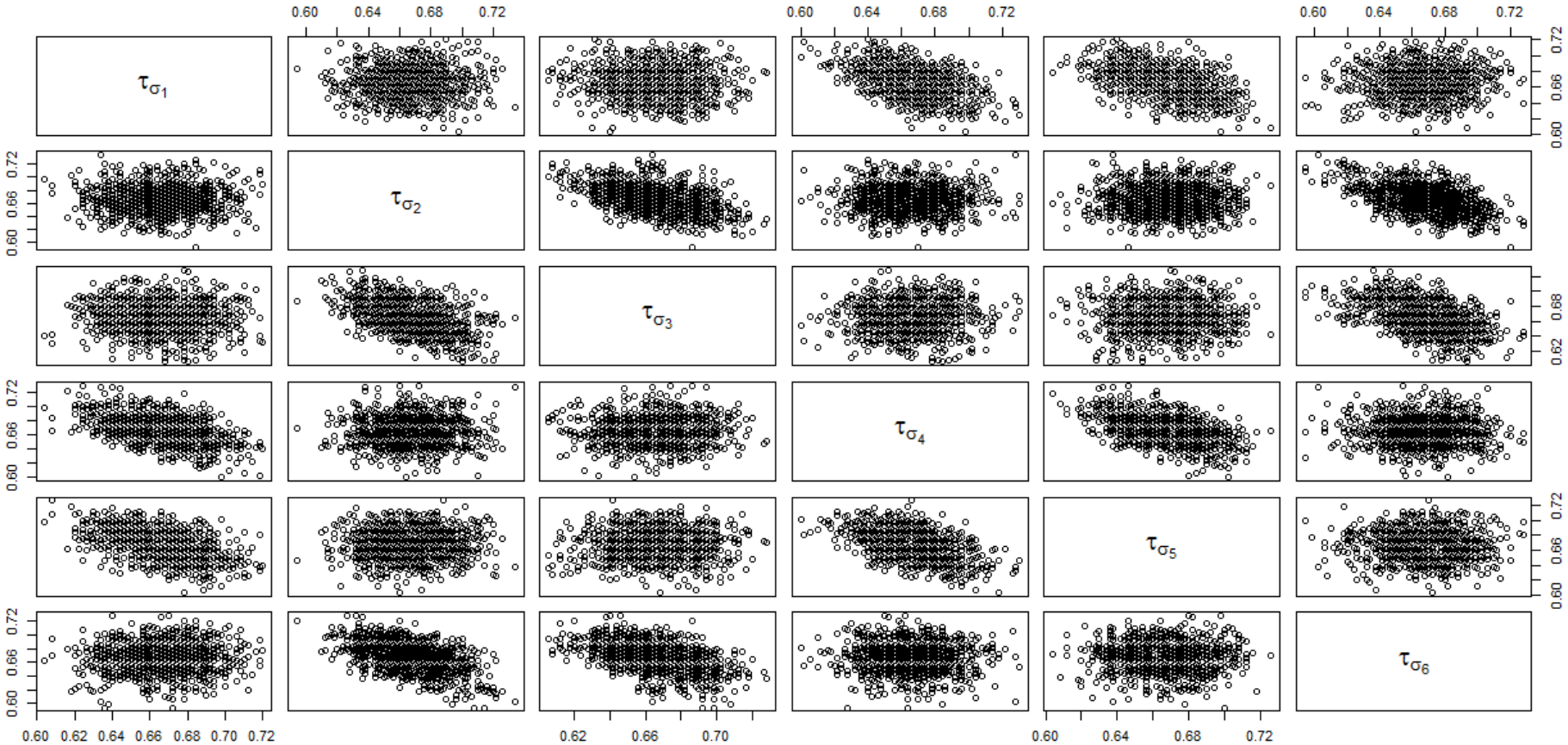
Example with $J = 3$: $\sigma_1 = Id, \sigma_2 = (23), \sigma_3 = (12), \sigma_4 = (123), \sigma_5 = (132), \sigma_6 = (13)$ with $n = 500$ and over $N = 1000$ observations for each τ_{σ_i}.

	$n\tau_{\sigma_1}$	$n\tau_{\sigma_2}$	$n\tau_{\sigma_3}$	$n\tau_{\sigma_4}$	$n\tau_{\sigma_5}$	$n\tau_{\sigma_6}$
$n\tau_{\sigma_1}$	1.00	-0.01	-0.06	-0.50	-0.50	0.06
$n\tau_{\sigma_2}$	-0.01	1.00	-0.47	-0.01	0.02	-0.53
$n\tau_{\sigma_3}$	-0.06	-0.47	1.00	0.04	0.02	-0.50
$n\tau_{\sigma_4}$	-0.50	-0.01	0.04	1.00	-0.50	-0.03
$n\tau_{\sigma_5}$	-0.50	0.02	0.02	-0.50	1.00	-0.03
$n\tau_{\sigma_6}$	0.06	-0.53	-0.50	-0.03	-0.03	1.00

Observe that: $\frac{\frac{-n}{J^2}}{n\left(1-\frac{1}{J}\right)\frac{1}{J}} = \frac{\frac{-500}{9}}{500 \times 2/9} = -0,5$

The Matching Error: correlations between $n\tau_{\sigma_l}$ and $n\tau_{\sigma_k}$.

Example with $J = 3$: $\sigma_1 = Id, \sigma_2 = (23), \sigma_3 = (12), \sigma 4 = (123), \sigma_5 = (132), \sigma_6 = (13)$



The Matching Error: distribution function for the case $J = 2$

If $J = 2$, the range of $n\tau$ are natural numbers in $\left[0, \frac{n}{2}\right]$ and the permutations involved are two: $\sigma_1 = Id$ and $\sigma_2 = (12)$. From the observations above we have:

- $n\tau_{\sigma_1} + n\tau_{\sigma_2} = n$; $\quad n\tau_{\sigma_1}, n\tau_{\sigma_2} \sim \mathcal{B}_{n,1/2}$ where $\mathcal{B}_{n,1/2}$ denotes the binomial distribution with parameters n and θ,
- $\mathbb{E}(n\tau_{\sigma_i}) = \frac{n}{2}$ $\quad$ and $\quad$ $\mathbb{V}(n\tau_{\sigma_i}) = \frac{n}{4}, \forall i = 1, 2$

As $n\tau_{\sigma_1} + n\tau_{\sigma_2} = n$, if $n\tau_{\sigma_1} = x$ and $n\tau_{\sigma_2} = y$, it is easy to establish that:

$$\mathbb{P}(n\tau_{\sigma_1} = x, n\tau_{\sigma_2} = y) = \begin{cases} 0 & \text{if } x + y \neq n \\ \binom{n}{x}\left(\frac{1}{2}\right)^n & \text{if } x + y = n \end{cases}$$

and the joint probability table for $n\tau_{\sigma_1}$ and $n\tau_{\sigma_2}$ has the following shape:

$n\tau_{\sigma_1}/n\tau_{\sigma_2}$	0	1	2	$\cdots$	$n-2$	$n-1$	n
0	0	0	0	0	0	0	$\binom{n}{0}\left(\frac{1}{2}\right)^n$
1	0	0	0	0	0	$\binom{n}{1}\left(\frac{1}{2}\right)^n$	0
2	0	0	0	0	$\binom{n}{2}\left(\frac{1}{2}\right)^n$	0	0
$\vdots$							
	0	0	0	$\cdots$	0	0	0
$n-2$	0	0	$\binom{n}{n-2}\left(\frac{1}{2}\right)^n$	0	0	0	0
$n-1$	0	$\binom{n}{n-1}\left(\frac{1}{2}\right)^n$	0	0	0	0	0
n	$\binom{n}{n}\left(\frac{1}{2}\right)^n$	0	0	0	0	0	0

Proposición 3

The distribution function $F_{n\tau}(z)$ *of* $n\tau = \min\{n\tau_{\sigma_1}, n\tau_{\sigma_2}\}$ *is:*

$$\mathbb{P}(n\tau \leq z) = \begin{cases} \sum_{i=0}^{z} 2\binom{n}{i}\left(\frac{1}{2}\right)^n = 2\mathcal{B}_{n,1/2}(z) & \text{if } z+1 \leq \frac{n}{2} \\ 1 & \text{if } z > \frac{n}{2} \end{cases}$$

where $B_{n,1/2}$ *is the binomial distribution function with parameters* $(n, \frac{1}{2})$.

Proposición 4

$$\mathbb{E}(n\tau) = \begin{cases} \frac{n}{2} - \left(\frac{1}{2}\right)^n \binom{n}{n/2}\frac{n}{2} & \text{if } n \text{ is even} \\ \frac{n}{2} - \left(\frac{1}{2}\right)^n \binom{n-1}{\frac{n-1}{2}} n & \text{if } n \text{ is odd} \end{cases}$$

Hence the distribution of ME is explicit for $J = 2$.

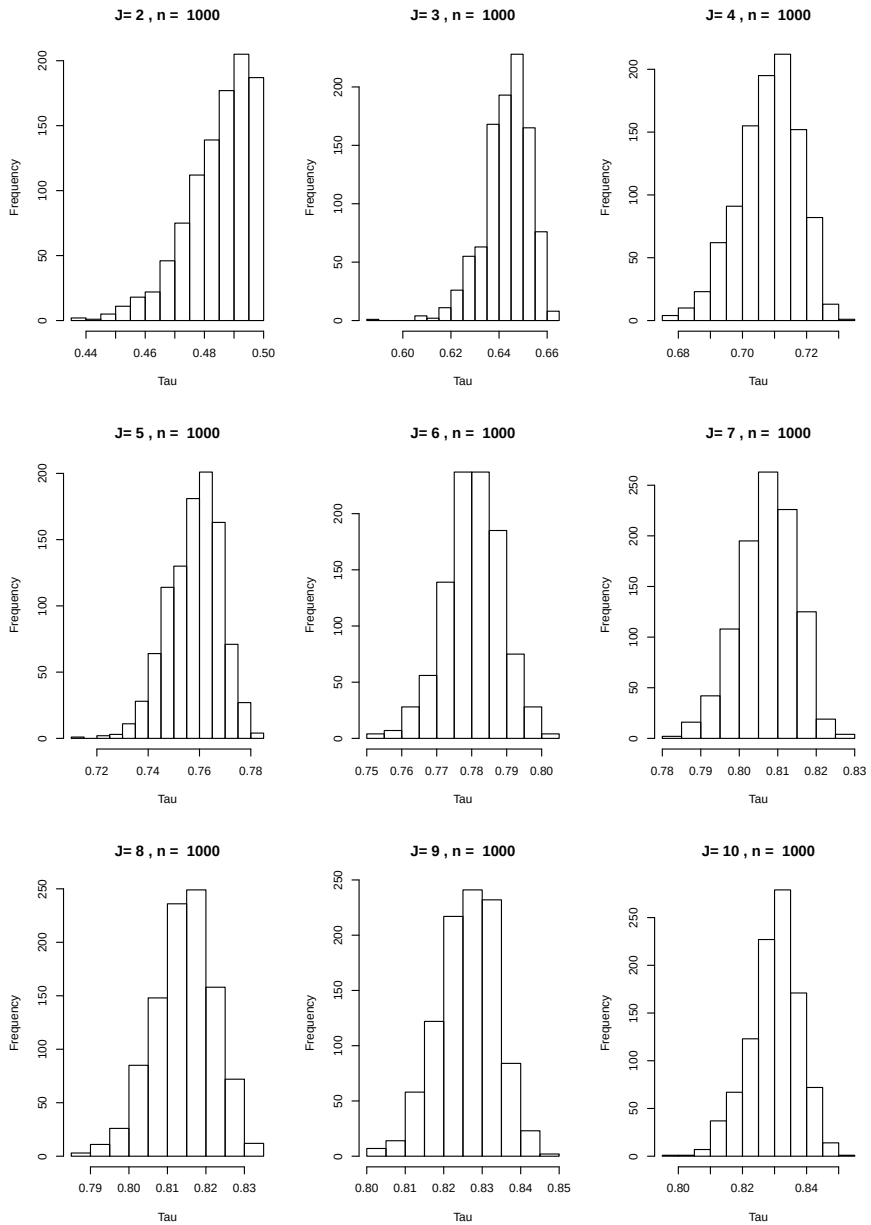
Empirical Distribution of ME

We study the empirical distribution of ME between two partitions $\mathscr{C}$ and $\mathscr{C}'$.

- The results are considered and compared in various experimental conditions: different number of groups ($J \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$), different number of observations ($n \in \{50, 100, 200, 300, 400, 500, 1000\}$), independent partitions, different degrees of dependence between partitions, balanced and imbalanced clusters.
- For each configuration of these parameters, we directly generate the partition vectors, that is, each observation's label for both partitions. These could be the result of two clustering analysis. For each value of J and n, we generate $N = 1000$ independent partition pairs and the index $\tau = ME(\mathscr{C}, \mathscr{C}')$ is computed. This produces 1000 values of the index for each configuration.
- To simulate dependent partitions, we start with two equal vectors Y and $\hat{Y}$ and modify at random a proportion $\gamma \in \{0.1, 0.4, 0.6, 0.9\}$ of labels of $\hat{Y}$.

Distribution of ME

In Figure below we show the distribution of τ for different values of J and for $n = 1000$.



Correlation between ME and Rand and Jaccard indices

We check the correlations between the ME, and both the Rand and Jaccard indices. Tables below give the obtained results for the four scenarios, for $\gamma = 0.4$.

Figure: Correlation between $ME(\mathscr{C}, \mathscr{C}')$ and $R(\mathscr{C}, \mathscr{C}')$ over $N = 1000$ repetitions of the indices, where $\mathscr{C}$ and $\mathscr{C}'$ are partitions with $J = 2, \ldots, 10$ groups and $n = 1000$ observations in the four scenarios explained above.

		J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10	mean
Dependence	Unbalanced groups	-1	-0,93	-0,89	-0,83	-0,81	-0,79	-0,73	-0,74	-0,70	-0,83
	Balanced groups	-1	-1	-1	-1	-0,99	-0,99	-0,99	-0,99	-0,98	-0,99
Independence	Unbalanced groups	-0,94	-0,28	-0,43	-0,21	-0,21	-0,12	-0,14	-0,13	-0,05	-0,28
	Balanced groups	-0,93	-0,86	-0,83	-0,81	-0,8	-0,79	-0,78	-0,78	-0,77	-0,82

Figure: Correlation between $ME(\mathscr{C}, \mathscr{C}')$ and Jaccard$(\mathscr{C}, \mathscr{C}')$ over $N = 1000$ repetitions of the indices, where $\mathscr{C}$ and $\mathscr{C}'$ are partitions with $J = 2, \ldots, 10$ groups and $n = 1000$ observations in the four scenarios explained above.

		J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10	mean
Dependence	Unbalanced groups	-0,99	-0,95	-0,91	-0,87	-0,85	-0,82	-0,78	-0,77	-0,75	-0,86
	Balanced groups	-1	-1	-1	-1	-1	-1	-1	-1	-1	-0,99
Independence	Unbalanced groups	-0,58	-0,45	-0,6	-0,52	-0,56	-0,59	-0,57	-0,63	-0,54	-0,56
	Balanced groups	-0,93	-0,86	-0,83	-0,81	-0,8	-0,79	-0,78	-0,78	-0,77	-0,82

Plan

Hypothesis Test

Our main purpose when analysing the distribution of the ME index is to design a hypothesis test to decide whether two partitions are statistically independent.

- The properties proved above were derived under some assumptions and may be used to compare partitions at least for $J = 2$ groups equally distributed.
- We present this test and analyse its performance on simulated data.
  Given two partitions $\mathscr{C}$ and $\mathscr{C}'$, the test proposal is:
  (H_0): Partitions $\mathscr{C}$ and $\mathscr{C}'$ are independent.
  (H_1): Partitions $\mathscr{C}$ and $\mathscr{C}'$ are not independent.
  1. The test statistic is τ and, as we know its distribution for $J = 2$ under H_0 it is straightforward to derive decision.
  2. For $J > 2$, we propose the following reasoning.
     1. For two partitions $\mathscr{C}$ and $\mathscr{C}'$ we compute $\tau_0 = ME(\mathscr{C}, \mathscr{C}')$ and then take B "perturbations" of $\mathscr{C}'$ denoted $\mathscr{C}'_1, \mathscr{C}'_2, \ldots, \mathscr{C}'_B$, changing at random a proportion π of its labels.
     2. To estimate the p-value $\mathbb{P}(\tau \leq \tau_0)$ of the independence hypothesis (null hypothesis) we take the proportion of values of $\tau_b = ME(\mathscr{C}, \mathscr{C}'_b)$ which are less than τ_0.

We consider two independent balanced partitions of $n = 1000$ observations with $J = 15$ groups and fix $B = 1000$. In this case $\tau_0 = 0.881$ and the estimated $p-$value is 0.464, which is coherent with not rejecting the hypothesis of independence of the two partitions.

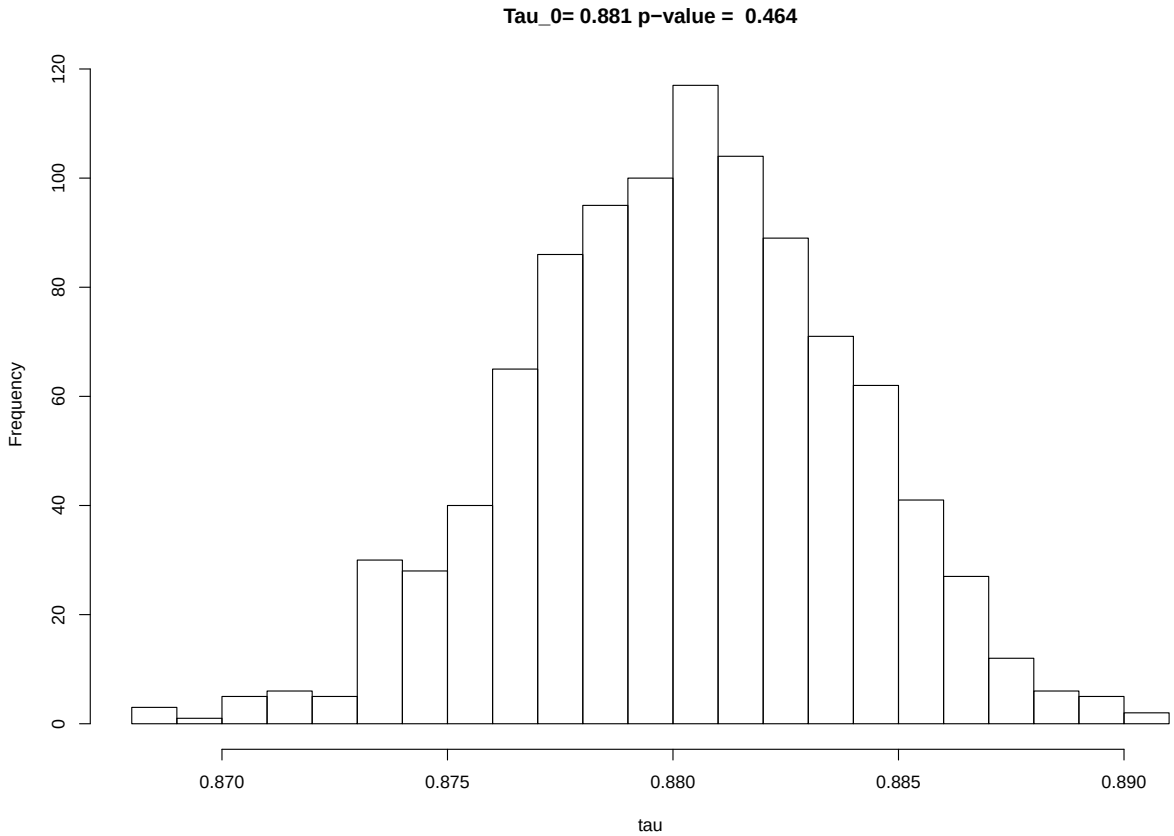


Figure: Histogram of the $B = 1000$ values of τ obtained by changing 20% of the labels randomly with two independent partitions.

Performance of the hypothesis test

To evaluate the performance of the test, we calculate the error averaging over $N = 1000$ simulations.
We take $B = 1000$ and proportion π equals 0.2.
At level $\alpha = 0.1$, we estimate the number of times we make type I and type II error using the empirical values of the p-value.

- As it was expected, taking proportion of values of p-value less than 0.1, the estimation of doing a type I error is small, less than 8%, when we compare two independent partitions.

Table: At level 0.1, proportion of times a type I error is made obtained by averaging $N = 1000$ comparisons of independent partitions with balanced classes.

	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	1.1	1.6	3.3	3.1	4.3	6.1	7.9	7.1
n=100	0.3	1.2	1.6	3.1	3.6	3.1	4.0	4.5
n=200	0.2	0.8	1.2	1.5	1.7	2.5	2.8	3.0
n=300	0.4	0.7	1.1	1.3	1.6	1.1	2.4	2.3
n=400	0.2	0.4	0.6	1.3	1.8	1.3	1.4	1.7
n=500	0.0	0.2	1.1	0.9	1.2	1.7	1.4	2.8
n=1000	0.0	0.1	0.8	0.9	0.8	1.4	1.4	2.2

We consider two dependent partitions ($J = 15$) where the second is constructed from the first by changing 20% of the labels randomly. In this case $\tau_0 = 0.186$ and the estimated p-value equals 0 so we reject the hypothesis of independence.

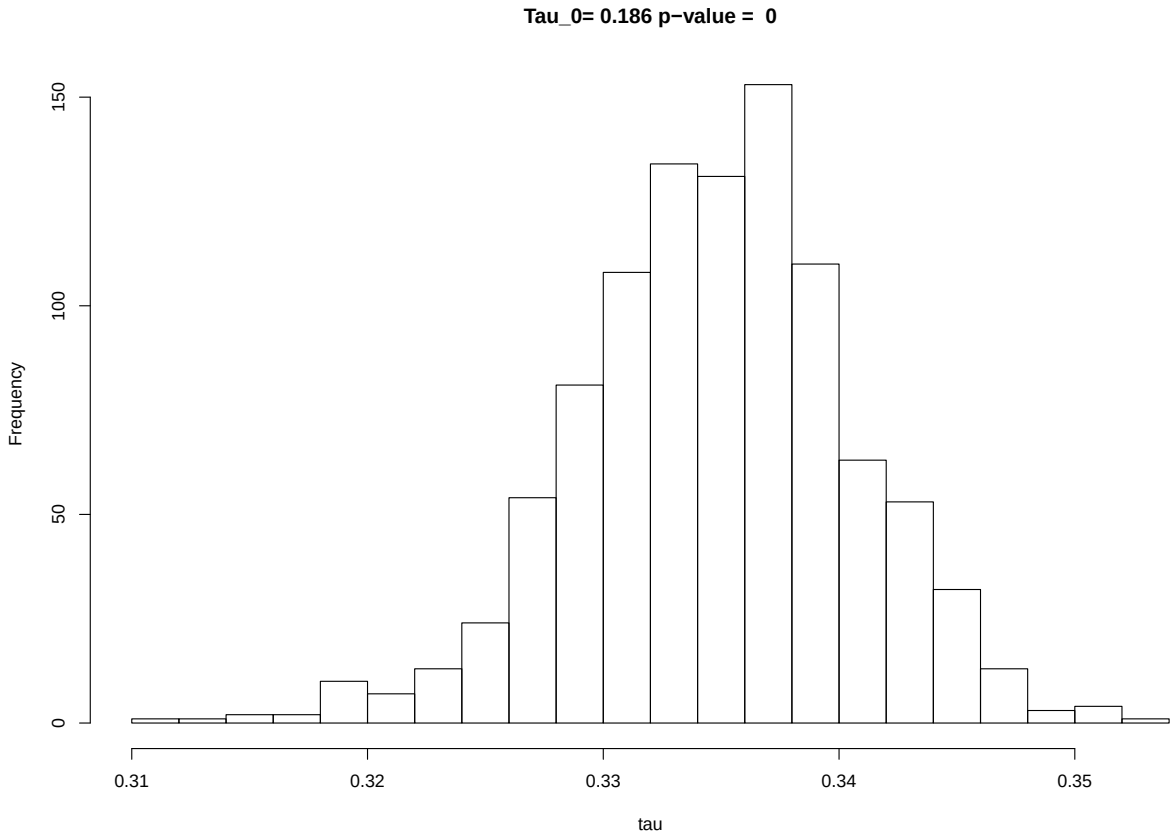


Figure: Histogram of the $B = 1000$ values of τ with two dependent partitions.

Performance of the hypothesis test

On the other hand, with dependent partitions, to estimate the proportion of type II errors over the N simulations, we take the proportion of p-values larger than 0.1. Partitions with balanced clusters and different degrees of dependence are simulated where dependency strength γ varies in $\{0.1, 0.4, 0.6, 0.9\}$.

- This estimation equals zero when $\gamma = 0.1$ and $\gamma = 0.4$ accordingly with a high dependence of the partitions.
- For $\gamma = 0.6$ it is null except for $n = 50$: for $J = 3$ it equals 17.8%, and it is less than 3% only for $4 \leq J \leq 10$. When $\gamma = 0.9$, which is a scenario very close to independence, it has high values for $n = 50, 100$ an 200 but decreases to 0 when n and J grow.

Table: At level 0.1, proportion of times a type II error is made obtained averaging $N = 1000$ comparisons of dependent partitions with $\gamma = 0.9$, with balanced classes.

	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	98.3	97.4	96.0	95.3	94.4	93.3	92.2	91.9
n=100	97.7	96.4	95.1	93.6	91.1	91.3	90.4	91.8
n=200	94.3	89.0	87.7	83.0	83.1	82.5	79.2	82.8
n=300	89.5	78.9	72.7	65.9	67.5	63.9	63.2	62.7
n=400	81.8	63.0	53.4	46.1	43.1	40.8	40.5	38.1
n=500	73.1	46.5	35.0	27.9	21.5	23.4	20.3	17.3
n=1000	22.2	3.7	1.6	0.2	0.2	0.0	0.0	0.0

Conclusions

- We have suggested an hypothesis test for comparing two partitions, useful for comparing the results of two clustering approaches over a same dataset.
- Our test is based on the mismatch error inspired from the misclassification error in supervised learning. We have analyzed the properties and the distribution of this index in several conditions and compared it to other common indices.
- A closed form of the statistic distribution under the null hypothesis was given for two clusters under mild conditions.
- For more than two clusters, the simulations show that the test is quite robust and reliable in various experimental conditions, but the statistic distribution under the null hypothesis is still unavailable.

Bibliografía

Fred, A.L., Jain, A.K., 2003. Robust data clustering., in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE. pp. 128–136.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of classification 2, 193–218.

Hultsch, L., 2004. Untersuchung zur besiedlung einer sprengfläche im pockautal durch die tiergruppen heteroptera (wanzen) und auchenorrhyncha(zikaden). Studienarbeit TU Bergakademie Freiberg, Studiengang Geoökologie .

Idrissi, A.N., 2000. Contribution à l'unification de critères d'association pour variables qualitatives. Ph.D. thesis. Université Paris 6, France.

Meilă, M., 2003. Comparing clusterings by the variation of information, in: Schölkopf, B., Warmuth, M.K. (Eds.), Learning Theory and Kernel Machines, Springer Berlin Heidelberg. pp. 173–187.

Meilă, M., 2005. Comparing clusterings: an axiomatic view, in: Proceedings of the 22nd international conference on Machine learning, ACM. pp. 577–584.

Meilă, M., Heckerman, D., 2001. An experimental comparison of model-based clustering methods. Machine learning 42, 9–29.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66. URL: `http://gen.lib.rus.ec/scimag/index.php?s=10.2307/2284239`, doi:`10.2307/2284239`.

Strehl, A., Ghosh, J., 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research 3, 583–617.

Wagner, S., Wagner, D., 2007. Comparing clusterings: an overview. Universität Karlsruhe, Fakultät für Informatik Karlsruhe.

Youness, G., Saporta, G., 2004. Some measures of agreement between close partitions. Student 51, 1–12.