Comparing high dimensional partitions with the Coclustering Adjusted Rand Index

Valérie Robert, Yann Vasseur

Comparing high dimensional partitions with the Coclustering Adjusted Rand Index.

Valerie Robert
Laboratoire de Mathématiques, UMR 8628, Bâtiment 425,
Université Paris Saclay, F-91405, Orsay, France
and
Yann Vasseur
Laboratoire de Mathématiques, UMR 8628, Bâtiment 425,
Université Paris Saclay, F-91405, Orsay, France
*

May 31, 2017

Abstract

The popular Adjusted Rand Index (ARI) is extended to the task of simultaneous clustering of the rows and columns of a given matrix. This new index called Coclustering Adjusted Rand Index (CARI) remains convenient and competitive facing other indices. Indeed, partitions with high numbers of clusters can be considered and it does not require any convention when the numbers of clusters in partitions are different. Experiments on simulated partitions are presented and the performance of this index to measure the agreement between two pairs of partitions is assessed. Comparison with other indices is discussed.

Keywords: Coclustering, Adjusted Rand Index, Agreement, Partition

*The authors are grateful to Gilles Celeux and Christine Keribin for initiating this work, to Vincent Brault and Gerard Govaert for useful discussions and valuable suggestions.

1 Introduction

With the advent of large datasets in statistics, coclustering arouses a genuine interest for last years in many fields of applications (text mining (Dhillon et al., 2003), genomics (Jagalur et al., 2007), recommendation systems (Shan and Banerjee, 2008; Wyse et al., 2017), pharmacoepidemiology (Robert et al., 2015), and so on ...). Initiated by Hartigan (1975), this useful technique aims at reducing the data matrix in a simpler one with the same structure (Govaert and Nadif, 2013). Indeed, taking profit of the two-dimensional nature of the issue, it enables to provide a simultaneous partition of two sets A (rows, observations, individuals) and B (columns, variables, attributes). To assess the performances of coclustering, partitions obtained by the procedure need to be evaluated. Objective criteria are therefore required to measure how close are these partitions to a reference. On the one hand, Charrad et al. (2010) suggest a first solution and artificially extend several standard indices from clustering (Dunn index, Baker and Hubert index, Davies and Bouldin index, Calinsky and Harabsz index, Silhouette criterion, Hubert and Levin index, Krzanowski and Lai index and the differential method). Wyse et al. (2017) also extend in the same way, another index relied on the normalized mutual information measure introduced in Vinh et al. (2010). However, proceeding in such a manner by just defining a linear combination between the index for row partitions and the index for column partitions, the coclustering structure is not preserved. On the other hand, Lomet (2012) proposes a distance dedicated to coclustering. Nevertheless, the computation of this index is dependent on the number of partition permutations and this property makes it time-consuming so that numbers of clusters can barely exceed nine in each direction. Moreover, no convention is given when the number of clusters of compared partitions is different. The aim of the present paper is to go further and to adapt the very popular and consensual *Adjusted Rand Index* (ARI)

developed by Hubert and Arabie (1985) from a coclustering point of view. To challenge other indices and tackle the problem of high dimensional partitions with numbers of clusters possibly different, this new index takes into account the coclustering structure while its computation remains time saving. The paper is organized as follows. In the next section, the *Adjusted Rand Index* (ARI) on which our index is based on, is presented. In Section 3, the *Coclustering Adjusted Rand Index* (CARI) is detailed and its properties ensuring its efficiency are demonstrated. In Section 4, this new index is exemplified on some partitions. Section 5 is devoted to numerical experiments to illustrate the behaviour of the index and a comparison with other coclustering indices. Finally a conclusion section ends this paper.

2 Statistical framework

In order to assess clustering results, objective criteria are required. For this purpose, distances of agreement between two partitions are developed. We will present the popular measure on which we base our new criterion.

2.1 Notation

Let two partitions be $\boldsymbol{z} = (z_1, \ldots, z_H)$ and $\boldsymbol{z'} = (z'_1, \ldots, z'_{H'})$ on a set $A = \{O_1, \ldots, O_I\}$, with Card(A)=I. $\boldsymbol{z}$ denotes for example an external reference and $\boldsymbol{z'}$ a clustering result.

2.2 The Rand Index and the Adjusted Rand Index

The *Rand Index* (RI) developed by Rand (1971), is a measure of the similarity between two data clusterings $\boldsymbol{z}$ and $\boldsymbol{z'}$, and is calculated as follows:

$$\frac{a+d}{a+b+c+d} = \frac{a+d}{\binom{I}{2}},$$

where,

- a denotes the number of pairs of elements that are placed in the same cluster in $\boldsymbol{z}$ and in the same cluster in $\boldsymbol{z'}$,
- b denotes the number of pairs of elements in the same cluster in $\boldsymbol{z}$ but not in the same cluster in $\boldsymbol{z'}$,
- c denotes the number of pairs of elements in the same cluster in $\boldsymbol{z'}$ but not in the same cluster in $\boldsymbol{z}$,
- d denotes the number of pairs of elements in different clusters in both partitions. The values a and d can be interpreted as agreements, and b and c as disagreements.

To compute all these values, a contingency table can be introduced. Let $\boldsymbol{n^{zz'}} = (n^{zz'}_{h,h'})_{H\times H'}$ be the matrix where $n^{zz'}_{h,h'}$ denotes the number of elements of the set A which belong both the cluster z_h and the cluster $z'_{h'}$. The row and column margins $n^{zz'}_{h,\bullet}$ and $n^{zz'}_{\bullet,h'}$ denote respectively the number of elements in the cluster z_h and $z'_{h'}$. We have the following correspondence (Santos and Embrechts, 2009):

- $a = \sum_{h}^{H}\sum_{h'}^{H'} \binom{n^{zz'}_{h,h'}}{2} = \frac{\sum_{h}^{H}\sum_{h'}^{H'} \left(n^{zz'}_{h,h'}\right)^2 - I}{2}$,
- $b = \sum_{h}^{H} \binom{n^{zz'}_{h,\bullet}}{2} - a = \frac{\sum_{h}^{H}(n^{zz'}_{h,\bullet})^2 - \sum_{h}^{H}\sum_{h'}^{H'} \left(n^{zz'}_{h,h'}\right)^2}{2}$,
- $c = \sum_{h'}^{H'} \binom{n^{zz'}_{\bullet,h'}}{2} - a = \frac{\sum_{h'}^{H'}(n^{zz'}_{\bullet,h'})^2 - \sum_{h}^{H}\sum_{h'}^{H'} \left(n^{zz'}_{h,h'}\right)^2}{2}$,
- $d = \binom{I}{2} - a - b - c = \sum_{h}^{H} \binom{n^{zz'}_{h,\bullet}}{2} - \sum_{h'}^{H'} \binom{n^{zz'}_{\bullet,h'}}{2} + a = \frac{\sum_{h}^{H}\sum_{h'}^{H'} \left(n^{zz'}_{h,h'}\right)^2 + I^2 - \sum_{h}^{H} \left(n^{zz'}_{h,\bullet}\right)^2 - \sum_{h'}^{H'} \left(n^{zz'}_{\bullet,h'}\right)^2}{2}$.

This symmetric index lies between 0 and 1 and takes the value 1 when the two partitions agree perfectly up to a permutation. Thus, by comparing pairs of elements, this index does not need to review all the permutations of studied partitions and its computation is efficient.

Although, the expected value of the *Rand Index* for two random partitions does not take a constant value and its taken values are concentrated in a small interval close to 1 (Meilă (2007)). The *Adjusted Rand Index* (ARI) proposed by Hubert and Arabie (1985) enables to overcome such drawbacks. This corrected version assumes the generalized hypergeometric distribution as the model of randomness, that is to say partitions are chosen randomly such that the number of elements in the clusters are fixed. The general form of this index which is the normalized difference between the *Rand Index* and its expected value under the generalized hypergeometric distribution assumption, is as follows:

$$\text{ARI} = \frac{\text{Index-Expected Index}}{\text{MaxIndex-Expected Index}}. \tag{1}$$

This index is bounded by 1, and takes this value when the two partitions are equal up to a permutation. It can also take negative values, which corresponds to a less agreement than expected by chance.

From Equation (1), Hubert and Arabie (1985) show the ARI can be written in this way:

$$\begin{aligned} \text{ARI}(\boldsymbol{z}, \boldsymbol{z'}) &= \frac{\sum_{h,h'} \binom{n^{zz'}_{h,h'}}{2} - \sum_{h} \binom{n^{zz'}_{h,\bullet}}{2} \sum_{h'} \binom{n^{zz'}_{\bullet,h'}}{2} / \binom{I}{2}}{\frac{1}{2}\left[\sum_{h} \binom{n^{zz'}_{h,\bullet}}{2} + \sum_{h'} \binom{n^{zz'}_{\bullet,h'}}{2} \right] - \left[\sum_{h} \binom{n^{zz'}_{h,\bullet}}{2} \sum_{h'} \binom{n^{zz'}_{\bullet,h'}}{2} \right] / \binom{I}{2}} \\ &= \frac{2(ad - bc)}{b^2 + c^2 + 2ad + (a+d)(b+c)}. \end{aligned} \tag{2}$$

Like the RI, the ARI is symmetric, that is to say $\text{ARI}(\boldsymbol{z}, \boldsymbol{z'})$=$\text{ARI}(\boldsymbol{z'}, \boldsymbol{z})$. Indeed, when the $\text{ARI}(\boldsymbol{z'}, \boldsymbol{z})$ is considered, the associated contingency table is $t(\boldsymbol{n^{zz'}})$, where t denotes the tranpose of a matrix. Besides, in Equation (2) defining the ARI, the margins of the

contingency table work in a symmetric way. That is why, while considering $\boldsymbol{n}^{\boldsymbol{zz'}}$ or its tranpose matrix $t(\boldsymbol{n}^{\boldsymbol{zz'}})$, the ARI remains unchanged. This remark would be particularly interesting in the next section, when the new index we develop is studied.

3 The Coclustering Adjusted Index

We extend the *Adjusted Rand Index* from a coclustering point of view to compare two coclustering partitions which define blocks, and not clusters anymore.

3.1 Notation

Let two partitions be $\boldsymbol{z} = (z_1, \ldots, z_h, \ldots z_H)$ and $\boldsymbol{z'} = (z'_1, \ldots, z'_{h'}, \ldots, z'_{H'})$ on a set A and let two partitions be $\boldsymbol{w} = (w_1, \ldots, w_\ell, \ldots, w_L)$ and $\boldsymbol{w'} = (w'_1, \ldots, w_{\ell'}, \ldots, w'_{L'})$ on a set B. $(\boldsymbol{z}, \boldsymbol{w})$ and $(\boldsymbol{z'}, \boldsymbol{w'})$ are two coclustering partitions on the set $A \times B$ where an observation is denoted by $x_{ij}, i = 1, \ldots, I; j = 1, \ldots, J$, with $\text{Card}(A \times B) = I \times J$. Notice that $\boldsymbol{z}$ and $\boldsymbol{z'}$ are called row partitions. Similarly, $\boldsymbol{w}$ and $\boldsymbol{w'}$ are called column partitions.

3.2 The Coclustering Adjusted Rand Index

Definition 3.1. *The contingency table* $\boldsymbol{n}^{\boldsymbol{zwz'w'}} = (n_{p,q}^{zwz'w'})_{(H\times L)\times(H'\times L')}$ *is defined such as* $n_{p,q}^{zwz'w'}$ *denotes the number of observations of the set* $A \times B$ *which belongs to the block* p *(related to a pair* (h, ℓ) *defined by* $(\boldsymbol{z}, \boldsymbol{w})$*) and the block* q *(related to a pair* (h', ℓ') *defined by* $(\boldsymbol{z'}, \boldsymbol{w'})$*).*

The contingency table can be seen as a block matrix which consists of $H \times H'$ blocks of size $L \times L'$ (see Table 1).

$$\begin{pmatrix}
n_{1,1}^{zwz'w'} & n_{1,2}^{zwz'w'} & \dots & n_{1,L'}^{zwz'w'} & \dots & n_{1,(H'\text{-}1)L+1}^{zwz'w'} & \dots & \dots & n_{1,H'L'}^{zwz'w'} \\
n_{2,1}^{zwz'w'} & n_{2,2}^{zwz'w'} & \dots & n_{2,L'}^{zwz'w'} & \dots & n_{2,(H'\text{-}1)L+1}^{zwz'w'} & \dots & \dots & n_{2,H'L'}^{zwz'w'} \\
\vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\
n_{L,1}^{zwz'w'} & n_{L,2}^{zwz'w'} & \dots & n_{L,L'}^{zwz'w'} & \dots & n_{L,(H'\text{-}1)L+1}^{zwz'w'} & \dots & \dots & n_{L,H'L'}^{zwz'w'} \\
\vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & & \vdots \\
\vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & & \vdots \\
n_{(H\text{-}1)L+1,1}^{zwz'w'} & n_{(H\text{-}1)L+1,2}^{zwz'w'} & \dots & n_{(H\text{-}1)L+1,L'}^{zwz'w'} & \dots & n_{(H\text{-}1)L+1,(H'-1)L+1}^{zwz'w'} & \dots & \dots & n_{(H\text{-}1)L+1,H'L'}^{zwz'w'} \\
\vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\
n_{HL,1}^{zwz'w'} & n_{HL,2}^{zwz'w'} & \dots & n_{HL,L'}^{zwz'w'} & \dots & n_{HL,(H'\text{-}1)L+1}^{zwz'w'} & \dots & \dots & n_{HL,H'L'}^{zwz'w'}
\end{pmatrix}$$

Table 1: Contingency table to compare two pairs of coclustering partitions.

Notice that a bijection can be defined between the index p of the rows of the contingency table, and the block (h, ℓ) defined by $(\boldsymbol{z}, \boldsymbol{w})$.

An analogous correspondence is defined for the index q and the block (h', ℓ') defined by $(\boldsymbol{z'}, \boldsymbol{w'})$. Thus the notation $(h_p\ \ell_p)$ and $(h'_q\ \ell'_q)$ could be used. We will see afterwards, this trick enables us to describe $\boldsymbol{n}^{\boldsymbol{zwz'w'}}$ in such a convenient way.

Definition 3.2. *Let $\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{z'}, \boldsymbol{w'}$ and $\boldsymbol{n}^{\boldsymbol{zwz'w'}}$ specified as in Definition 3.1. The Coclustering Adjusted Rand Index (CARI) is defined as follows:*

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

Like the ARI, this index is symmetric and takes the value 1 when the couples of partitions agree perfectly up to a permutation. But unlike the index proposed by Lomet (2012)

with which we will compare in Section 5, no convention is needed when the number of clusters is different in partitions. Moreover, it does not rely on the permutations of partitions and can therefore be easily computed even if the number of row clusters or column clusters exceeds nine. Though, the nave complexity to compute $\boldsymbol{n}^{\boldsymbol{zwz'w'}}$ is still substancial.

Fortunately, we manage to exhib a link between $\boldsymbol{n}^{\boldsymbol{zwz'w'}}$, $\boldsymbol{n}^{\boldsymbol{zz'}}$ and $\boldsymbol{n}^{\boldsymbol{ww'}}$ which makes the computation of the CARI much faster and competitive in a high dimensional setting:

Theorem 3.3. *Let $\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{z'}, \boldsymbol{w'}, \boldsymbol{n}^{\boldsymbol{zwzw'}}, \boldsymbol{n}^{\boldsymbol{zz'}}$ and $\boldsymbol{n}^{\boldsymbol{ww'}}$ be defined as in Definition 3.1. Then we have the following relation,*

$$\boldsymbol{n}^{\boldsymbol{zwzw'}} = \boldsymbol{n}^{\boldsymbol{zz'}} \otimes \boldsymbol{n}^{\boldsymbol{ww'}}, \tag{3}$$

where $\otimes$ denotes the Kronecker product between two matrices.

The proof of this theorem is postponed to Appendix A.1.

Thanks to this property, the contingency table $\boldsymbol{n}^{\boldsymbol{zwzw'}}$ can be computed more efficiently and its complexity is now $\mathcal{O}(HH'+LL'+HH'LL')$. Moreover, even if the Kronecker product is not commutative, it behaves well with both the transpose operator and the margins, and the initial properties of CARI are kept:

Corollary 1. *1. $\forall (p,q) \in (H \times L) \times (H' \times L')$, we have the relations between the margins,*

$$n_{\bullet,q}^{zwzw'} = n_{\bullet h'_q}^{zz'} \otimes n_{\bullet,\ell'_q}^{ww'} \text{ and } n_{p,\bullet}^{zwzw'} = n_{h_p,\bullet}^{zz'} \otimes n_{\ell_p,\bullet}^{ww'}.$$

2. The CARI associated with the contingency table $\boldsymbol{n}^{\boldsymbol{zwzw'}}$ defined as in Equation (3) remains symmetric, that is to say,

$$CARI((\boldsymbol{z}, \boldsymbol{w}), (\boldsymbol{z'}, \boldsymbol{w'})) = CARI((\boldsymbol{z'}, \boldsymbol{w'}), (\boldsymbol{z}, \boldsymbol{w})).$$

The proof of this corollary is postponed to Appendix A.2.

In the further sections, the contingency table $\boldsymbol{n}^{\boldsymbol{zwzw'}}$ is now defined by Equation (3).

4 Examples

4.1 Pairs of equal partitions up to a permutation.

Let consider the following couples of partitions $(\boldsymbol{z}, \boldsymbol{w}) = ((1,1,3,2),(1,2,1,4,3,))$ and $(\boldsymbol{z'}, \boldsymbol{w'}) = ((2,2,1,3),(2,1,2,3,4))$ which are equal up to a permutation. The contingency table (see Table 2) associated with CARI$((\boldsymbol{z}, \boldsymbol{w}),(\boldsymbol{z'}, \boldsymbol{w'}))$ has a size of $(3 \times 4, 3 \times 4)$.

Thus, the CARI$((\boldsymbol{z}, \boldsymbol{w}),(\boldsymbol{z'}, \boldsymbol{w'}))$ behaves well and is equal to $\frac{11-121/190}{1/2\times 22-121/190} = 1$.

Block	$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$	$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$
$(1,1)$	0	0	0	0	0	4	0	0	0	0	0	0
$(1,2)$	0	0	0	0	2	0	0	0	0	0	0	0
$(1,3)$	0	0	0	0	0	0	0	2	0	0	0	0
$(1,4)$	0	0	0	0	0	0	2	0	0	0	0	0
$(2,1)$	0	0	0	0	0	0	0	0	0	2	0	0
$(2,2)$	0	0	0	0	0	0	0	0	1	0	0	0
$(2,3)$	0	0	0	0	0	0	0	0	0	0	0	1
$(2,4)$	0	0	0	0	0	0	0	0	0	0	1	0
$(3,1)$	0	2	0	0	0	0	0	0	0	0	0	0
$(3,2)$	1	0	0	0	0	0	0	0	0	0	0	0
$(3,3)$	0	0	0	1	0	0	0	0	0	0	0	0
$(3,4)$	0	0	1	0	0	0	0	0	0	0	0	0

Table 2: Initial contingency table $\boldsymbol{n}^{\boldsymbol{zwz'w'}}$ (see Definition 3.1).

4.2 Pairs of partitions with a different number of clusters

Let us now consider the following partitions $(\boldsymbol{z}, \boldsymbol{w}) = ((1,2,2,2,1),(1,1,2,1,1,2))$ and $(\boldsymbol{z'}, \boldsymbol{w'}) = ((1,1,2,1,1),(1,1,2,1,3,2))$. Remark that partitions $\boldsymbol{w}$ and $\boldsymbol{w'}$ do not have the same number of clusters. The initial contingency tables related to ARI$(\boldsymbol{z}, \boldsymbol{z'})$, ARI$(\boldsymbol{w}, \boldsymbol{w'})$ and CARI$((\boldsymbol{z}, \boldsymbol{w}), (\boldsymbol{z'}, \boldsymbol{w'}))$ are described in Tables 3 et 4. We observe as announced that

$$\boldsymbol{n}^{\boldsymbol{zwz'w'}} = \boldsymbol{n}^{\boldsymbol{zz'}} \otimes \boldsymbol{n}^{\boldsymbol{ww'}}.$$

The values of the ARIs and the CARI are available in Table 5. We notably observe that the ARI's value for rows is negative.

Cluster	1	2	Margin
1	2	0	2
2	2	1	3
Margin	4	1	5

Cluster	1	2	3	Margin
1	3	0	1	4
2	0	2	0	2
Margin	3	2	1	6

Table 3: Contingency tables $\boldsymbol{n}^{\boldsymbol{zz'}}$(at left) and $\boldsymbol{n}^{\boldsymbol{ww'}}$(at right).

Block	$(1,1)$	$(1,2)$	$(1,3)$	$(2,1)$	$(2,2)$	$(2,3)$	Margin
$(1,1)$	6	0	2	0	0	0	8
$(1,2)$	0	4	0	0	0	0	4
$(2,1)$	6	0	2	3	0	1	12
$(2,2)$	0	4	0	0	2	0	6
Margin	12	8	4	3	2	1	30

Table 4: Initial contingency table $\boldsymbol{n}^{\boldsymbol{zz'ww'}}$ (see Definition 3.1).

	ARI($\boldsymbol{z}, \boldsymbol{z'}$)	ARI($\boldsymbol{w}, \boldsymbol{w'}$)	CARI($(\boldsymbol{z}, \boldsymbol{w}), (\boldsymbol{z'}, \boldsymbol{w'})$)
Value	-0.1538	0.5872	0.2501

Table 5: Comparison of the values of ARI($\boldsymbol{z}, \boldsymbol{z'}$), ARI($\boldsymbol{w}, \boldsymbol{w'}$) and CARI(($\boldsymbol{z}, \boldsymbol{w}$),($\boldsymbol{z'}, \boldsymbol{w'}$)).

5 Comparison between different coclustering indices

We will present the indices that we consider in the further simulation study. The notations refer to Section 3.1.

5.1 Other coclustering indices

5.1.1 Classification error

The classification distance presented in Lomet (2012) studies the misclassification rate of the observations in the blocks:

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

where $\mathfrak{S}(\{1,..H\})$ denotes the set of permutations on the set $\{1, \ldots, H\}$.

The classification error (CE) is then defined when the cost function measures the difference between the pairs of reference $(\boldsymbol{z}^*, \boldsymbol{w}^*)$ partitions and an estimation $(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{w}})$:

$$\text{CE}((\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{w}}), (\boldsymbol{z}^*, \boldsymbol{w}^*)) \quad = \quad \text{dist}_{(I,H)\times(J,L)}((\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{w}}), (\boldsymbol{z}^*, \boldsymbol{w}^*)).$$

The classification error is between 0 and 1. Thus, the observation x_{ij} is not in the block (h, ℓ) if the row i is not in the cluster h *or* if the column j is not in the cluster ℓ. When a column is improperly classified, all the cells of this column are penalized, and the classification error is increased by $\frac{1}{J}$.

Furthermore, the distance related to the row partitions can be also defined as follows:

$$\text{dist}_{I,H}(\boldsymbol{z}, \boldsymbol{z'}) = 1 - \max_{\sigma \in \mathfrak{S}(\{1,..H\})} \frac{1}{I} \sum_{i,h} z_{ih} z'_{i\sigma(h)}.$$

When the partitions do not include the same number of clusters, a suitable convention we can propose, is to consider H as the maximal number of clusters and the created additional clusters are assumed to be empty. Besides, the computation of this distance when H is higher than nine, remains difficult as the order of the set $\mathfrak{S}(\{1,..H\})$ is $H!$.

In a symmetric way, the distance related to the column partitions is denoted by $\text{dist}_{J,L}$.

Lomet (2012) shows that the classification error could be expressed in terms of the distance related to the row partitions and the distance related to the column partitions:

$$\begin{aligned} \text{dist}_{(I,H)\times(J,L)}((\boldsymbol{z}, \boldsymbol{w}), (\boldsymbol{z'}, \boldsymbol{w'})) &= \text{dist}_{I,H}(\boldsymbol{z}, \boldsymbol{z'}) + \text{dist}_{J,L}(\boldsymbol{w}, \boldsymbol{w'}) \\ &- \text{dist}_{I,H}(\boldsymbol{z}, \boldsymbol{z'}) \times \text{dist}_{J,L}(\boldsymbol{w}, \boldsymbol{w'}). \end{aligned}$$

5.1.2 Extended Generalized Mutual Information

The generalized mutual information introduced by Vinh et al. (2010), is extended by Wyse et al. (2017) to compare two coclustering partitions. Originally, the generalized mutual information between two partitions $\boldsymbol{z} = (z_1, \ldots, z_H)$ and $\boldsymbol{z'} = (z'_1, \ldots, z'_{H'})$ on a same set $A = \{O_1, \ldots, O_I\}$ is as follows:

$$\text{MI}(\boldsymbol{z}, \boldsymbol{z'}) = \sum_{h,h'} P_{h,h'} \log\left(\frac{P_{h,h'}}{P_h P_{h'}}\right),$$

$$\text{where,} \quad P_{h,h'} = \frac{1}{I} \sum_{i,i'}^{I} \mathbb{1}_{\{z_i = h, z'_{i'} = h'\}}, \quad P_h = \frac{1}{I} \sum_{i}^{I} \mathbb{1}_{\{z_i = h\}} \quad \text{and} \quad P_{h'} = \frac{1}{I} \sum_{i'}^{I} \mathbb{1}_{\{z'_{i'} = h'\}}.$$

When the two partitions do not present the same number of clusters, the quantity is normalized as follows:

$$\frac{\text{MI}(\boldsymbol{z}, \boldsymbol{z'})}{\max(\mathcal{H}(\boldsymbol{z}), \mathcal{H}(\boldsymbol{z'}))},$$

$$\text{where,} \quad \mathcal{H}(\boldsymbol{z}) = -\sum_{h} P_h \log P_h, \text{ and } \mathcal{H}(z') = -\sum_{h'} P_{h'} \log P_{h'}.$$

Thus, the proposed measure to compare two coclustering partitions ($\boldsymbol{z} = (z_1, \ldots, z_H)$, $\boldsymbol{w} = (w_1, \ldots, w_L)$) and ($\boldsymbol{z'} = (z'_1, \ldots, z'_{H'})$, $\boldsymbol{w'} = (w'_1, \ldots, w'_{L'})$) on a set $A \times B$ is based on a linear combination of the generalized mutual information of $\boldsymbol{z}$ and $\boldsymbol{z'}$, and the the generalized mutual information of $\boldsymbol{w}$ and $\boldsymbol{w'}$:

$$\text{MI}((\boldsymbol{z}, \boldsymbol{w}), (\boldsymbol{z'}, \boldsymbol{w'})) = \text{MI}(\boldsymbol{z}, \boldsymbol{z'}) + \text{MI}(\boldsymbol{w}, \boldsymbol{w'}).$$

The maximal value of this index is equal to 2 when the partitions perfectly match up to a permutation and is equal to 0 when the correspondence between them is extremely weak. Remark that, by extending this index in this way, the coclustering structure of the problem is not preserved and this major drawback will be tackled in the next section.

5.2 Simulation study

To compare the CARI with the other indices, we first propose to test their computation complexity as a function of the number of observations or clusters. Then, we assess their performance to measure how close are two coclustering partitions from a coclustering point of view. Finally, we investigate if there exists any simple link between the indices. The `R` code to reproduce the simulation experiments is provided in the *supplementary material*.

To achieve these objectives, we propose a simulation methodology to generate a set of coclustering partitions more or less close to the considered initial ones. Remark that a simulation approach already exists in the task of clustering in one dimension (Fowlkes and Mallows, 1983; Saporta and Youness, 2002; Youness and Saporta, 2004), but another point of view is developed here. Our procedure can now be described as follows:

Fix the sizes (I, J) and the number of clusters (H, L) of the coclustering partitions that would be studied. Consider the initial coclustering partitions $(\boldsymbol{z^{(0)}}, \boldsymbol{w^{(0)}})$ in the *balanced*

or *unbalanced* case, that is to say, where the number of individuals in each cluster is the same or not. For $i = 1, ..., N$ iterations:

1) Choose a coordinate of $\boldsymbol{z}^{(\boldsymbol{i-1})}$ at random and allocate to it, a new label chosen randomly between 1 and H. The new vector is named $\boldsymbol{z}^{(\boldsymbol{i})}$.

2) Reproduce the step 1) with the vector $\boldsymbol{w}^{(\boldsymbol{i-1})}$. The new vector is named $\boldsymbol{w}^{(\boldsymbol{i})}$.

3) Compute the different indices between $(\boldsymbol{z}^{(\boldsymbol{0})}, \boldsymbol{w}^{(\boldsymbol{0})})$ and $(\boldsymbol{z}^{(\boldsymbol{i})}, \boldsymbol{w}^{(\boldsymbol{i})})$.

Thus, at each iteration i, the coclustering partitions $(\boldsymbol{z}^{(\boldsymbol{i-1})}, \boldsymbol{w}^{(\boldsymbol{i-1})})$ and $(\boldsymbol{z}^{(\boldsymbol{i})}, \boldsymbol{w}^{(\boldsymbol{i})})$ can differ from only one coordinate in each vector. Gradually, the procedure produces a set of coclustering partitions more and more discordant with the initial coclustering partitions $(\boldsymbol{z}^{(\boldsymbol{0})}, \boldsymbol{w}^{(\boldsymbol{0})})$. The support of the studied indices from high values to small values, can therefore be well explored if the number of iterations N is high enough.

5.2.1 Time comparison

The complexity of the three indices related to the number of observations and the number of clusters is assessed. For this purpose, the procedure is run with $N = 10\ 000$ iterations considering two situations $(I, J) = (315, 315)$ observations and $(I, J) = (630, 630)$ observations when the number of clusters varies as follows, $(H, L) \in \{(5, 5), (7, 7), (9, 9)\}$, in the *balanced case*. The results are presented in Figures 1 and 2. We observe that the elapsed time computation in log scale of the MI is the smallest and seems not to be sensitive to the number of clusters or observations. This optimistic result is notably explained by the fact that the MI ignores the coclustering structure in a pair of partitions. The CARI which takes into account this structure, also behaves well whatever the number of clusters or observations. More precisely, the elapsed time computation of the CARI on a run of the

procedure with $N = 2000$, $(I, J) = (2000, 2000)$ and $(H, L) = (20, 20)$, is three seconds on average, which is reasonable in a high dimensional coclustering context. On the contrary, the time computation of the CE significantly increases with the number of clusters, which illustrates its dependence on the factorial of this quantity.

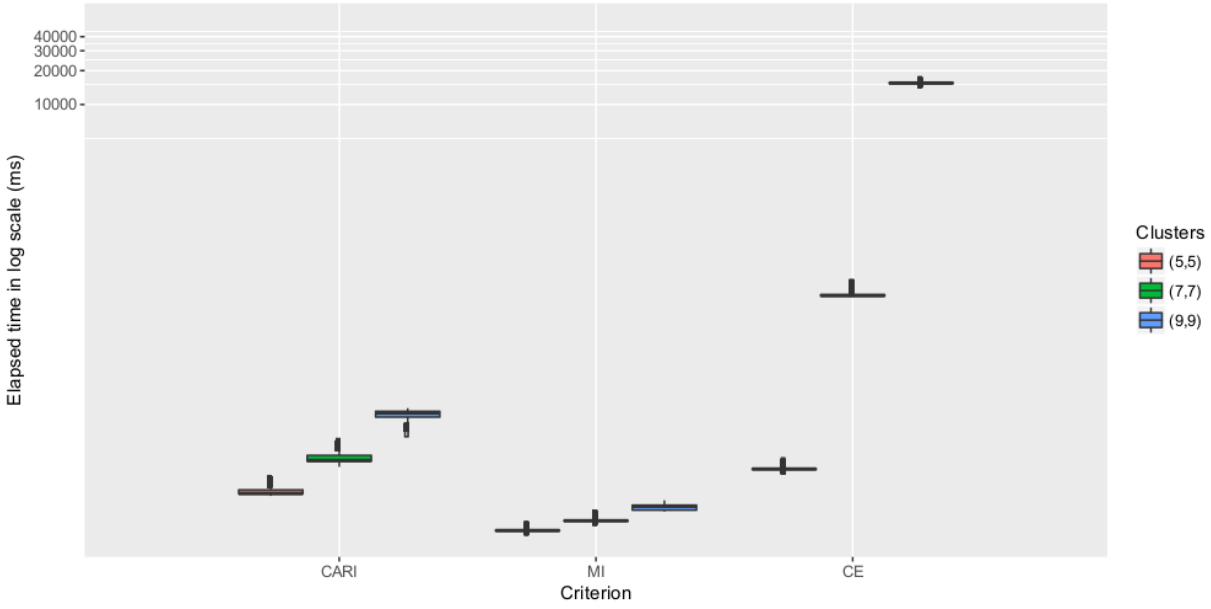

Figure 1: Boxplot of the elapsed time computation in log scale in milliseconds, of the CARI, the MI and the CE, for $N = 10\ 000$ iterations of the procedure, with $(I, J) = (315, 315)$ observations and for different numbers of clusters, $(H, L) \in \{(5, 5), (7, 7), (9, 9)\}$.

5.2.2 Behaviour comparison

The first comparison between the three indices is performed by running the procedure with $N = 10\ 000$ iterations, $(H, L) = (5, 5)$ and the following sample sizes $(I, J) = (50, 50)$, $(I, J) = (500, 500)$ and $(I, J) = (1000, 1000)$. The results are presented in Figure 3 in the

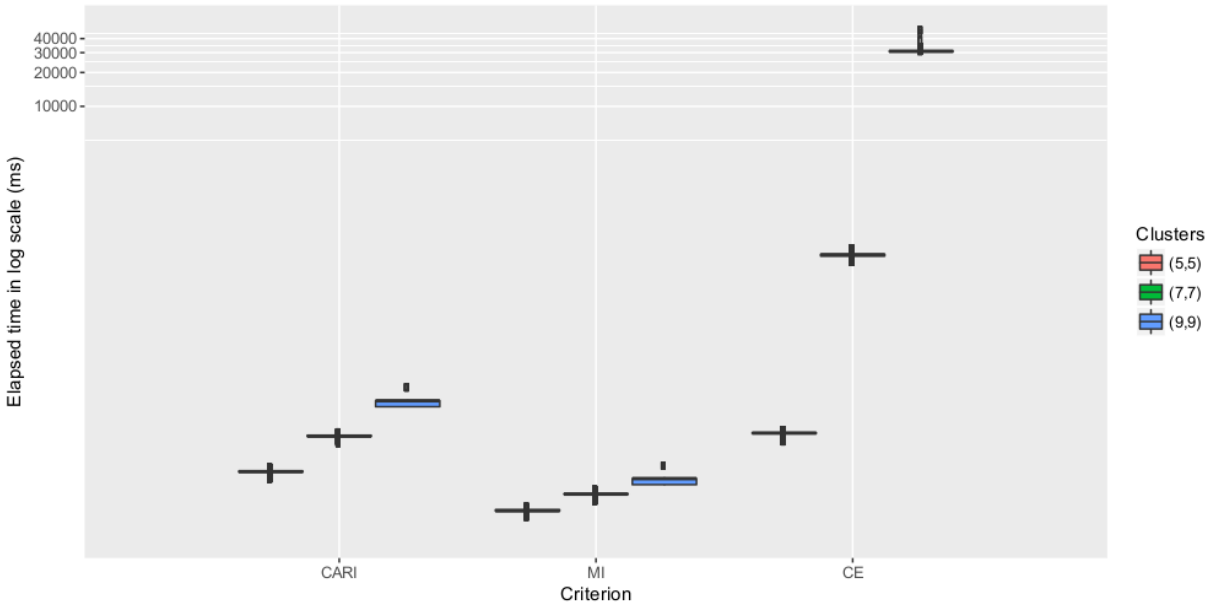

Figure 2: Boxplot of the elapsed time computation in log scale in milliseconds, of the CARI, the MI and the CE, for $N = 10\ 000$ iterations of the procedure, with $(I, J) = (630, 630)$ observations and for different numbers of clusters, $(H, L) \in \{(5, 5), (7, 7), (9, 9)\}$.

balanced case and in Figure 4 in the unbalanced case. In the unbalanced case, the number of observations in each cluster of the initial coclustering partitions is defined in Table 6.

Remark that we considerer the quantity $1-$CE which is more convenient to compare with the CARI. Indeed, a perfect matching between partitions is now corresponding to the value 1 for both indices. First of all, the experiment enables to scan all the supports of the indices, except for the CARI where negative values are not reached. To our knowledge, this phenomenon rather appeared when the agreement of the considered coclustering partitions are very weak and the number of observations is very small (less than the considered case here, (50,50)).

		cluster number				
		1	2	3	4	5
(I,J)	(50,50)	4	7	10	13	16
	(500,500)	20	35	100	165	180
	(1000,1000)	30	70	200	300	400

Table 6: Repartition of the observations in $(z^{(0)}, w^{(0)})$ for the *unbalanced* case.

Then, we observe in Figures 3 and 4, that the behaviour of the three indices are different enough as all the curves are far from the line bisector, and no simple link, like a linear one for example, can be exhibited. We also notice that in general, the CARI tends to be more demanding and penalizing than the other indices.

Moreover, we notice that when the number of observations is small, the fact that an observation is missclasified, impacts more the values of the three indices as the blue circles are more widely spaced, where the values of the indices are high in Figures 3 and 4.

In the unbalanced case (see Figure 4), the compared behaviour between the CARI and the quantity $1-\text{CE}$ seems to be globally the same whatever the number of observations. Conversely, we remark a changement in the compared behaviour between the CARI and the MI. Indeed, when the number of observations is high and the compared colustering partitions differed from few observations (corresponding to the part of red square curves with the highest values for the CARI and the MI in Figure 4, at left), the MI and the CARI behave in the same way, whereas the CARI is more demanding when the compared colustering partitions are very discordant.

The second comparison consists of observing how each criterion behaves when the compared pairs of coclustering partitions have the same row partition or the same column

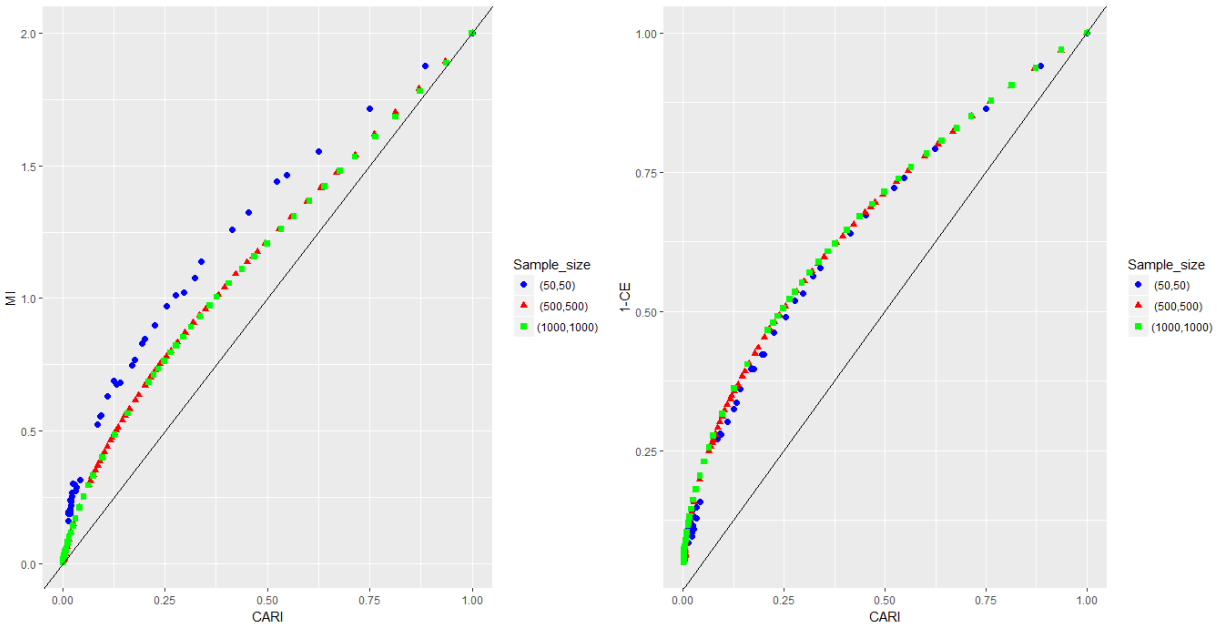

Figure 3: Comparison of the values of the CARI (on the horizontal axis) versus the values of the MI (at the left, on the vertical axis), and versus the values of the 1-CE (at right, on the vertical axis) in the *balanced case*, on a run of the procedure with $N = 10\ 000$, for different sample sizes $(I, J) = (50, 50)$ (blue circle), $(I, J) = (500, 500)$ (red triangle), $(I, J) = (1000, 1000)$ (green square).

partition. That is why we use again the procedure, presented in Section 5.2 and we complete the step 3) of the procedure for each iteration $i = 1...N$, as follows:

3) Compute the indices between $(\boldsymbol{z^{(0)}}, \boldsymbol{w^{(0)}})$ and $(\boldsymbol{z^{(i)}}, \boldsymbol{w^{(i)}})$, between $(\boldsymbol{z^{(0)}}, \boldsymbol{w^{(0)}})$ and $(\boldsymbol{z^{(0)}}, \boldsymbol{w^{(i)}})$ and between $(\boldsymbol{z^{(0)}}, \boldsymbol{w^{(0)}})$ and $(\boldsymbol{z^{(i)}}, \boldsymbol{w^{(0)}})$.

The results shown in Figures 5 and 6, respectively illustrate the comparison of the CARI versus the MI and versus the quantity $1-$CE on a run of the procedure with $N = 10\ 000$,

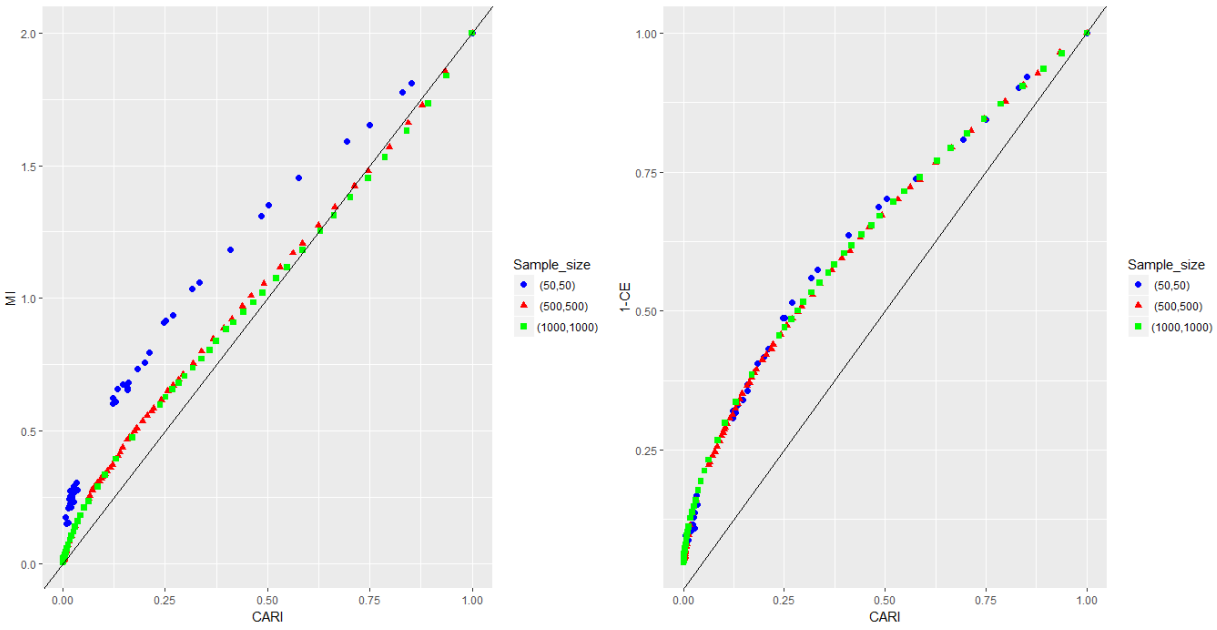

Figure 4: Comparison of the values of the CARI (on the horizontal axis) versus the values of the MI (at the left, on the vertical axis), and versus the values of the 1-CE (at right, on the vertical axis) in the *unbalanced case*, on a run of the procedure with $N = 10\ 000$, for different sample sizes $(I, J) = (50, 50)$ (blue circle), $(I, J) = (500, 500)$ (red triangle), $(I, J) = (1000, 1000)$ (green square).

$(H, L) = (7, 5)$, $(I, J) = (630, 630)$ in the *balanced* case. Each index is computed at each iteration i for the following pairs of coclustering partitions: $(\boldsymbol{z}^{(\mathbf{0})}, \boldsymbol{w}^{(\mathbf{0})})$ and $(\boldsymbol{z}^{(\boldsymbol{i})}, \boldsymbol{w}^{(\boldsymbol{i})})$ (red triangle), between $(\boldsymbol{z}^{(\mathbf{0})}, \boldsymbol{w}^{(\mathbf{0})})$ and $(\boldsymbol{z}^{(\mathbf{0})}, \boldsymbol{w}^{(\boldsymbol{i})})$(blue circle), between $(\boldsymbol{z}^{(\mathbf{0})}, \boldsymbol{w}^{(\mathbf{0})})$ and $(\boldsymbol{z}^{(\boldsymbol{i})}, \boldsymbol{w}^{(\mathbf{0})})$ (green square). In Figure 5 representing the comparison of the CARI versus the MI, we notice that the curves defined by blue circles and green squares are above the curve defined by red triangles. We therefore infer that the CARI is more penalizing than the MI when the compared pairs of coclustering partitions have the same row partition or the

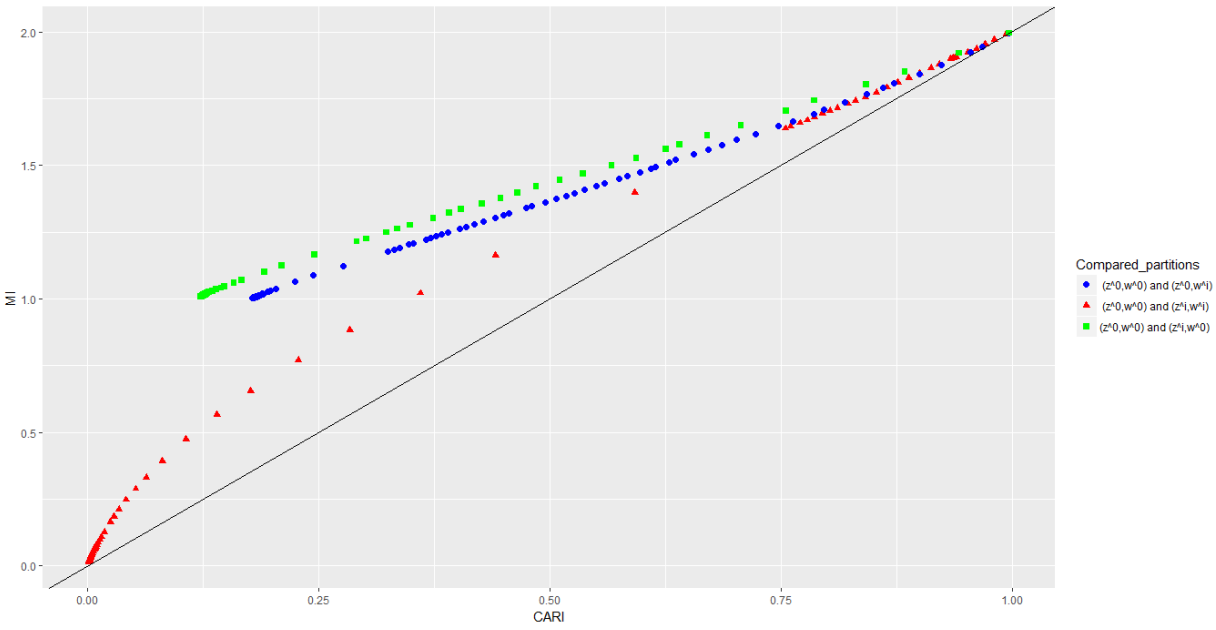

Figure 5: Comparison of the CARI's values (on the horizontal axis) versus the MI's values (on the vertical axis) on a run of the procedure with $N = 10\ 000$, $(H, L) = (7, 5)$, $(I, J) = (630, 630)$ in the *balanced case*. Each index is computed at each iteration i between $(\boldsymbol{z}^{(\boldsymbol{0})}, \boldsymbol{w}^{(\boldsymbol{0})})$ and $(\boldsymbol{z}^{(\boldsymbol{i})}, \boldsymbol{w}^{(\boldsymbol{i})})$ (red triangle), between $(\boldsymbol{z}^{(\boldsymbol{0})}, \boldsymbol{w}^{(\boldsymbol{0})})$ and $(\boldsymbol{z}^{(\boldsymbol{0})}, \boldsymbol{w}^{(\boldsymbol{i})})$(blue circle), between $(\boldsymbol{z}^{(\boldsymbol{0})}, \boldsymbol{w}^{(\boldsymbol{0})})$ and $(\boldsymbol{z}^{(\boldsymbol{i})}, \boldsymbol{w}^{(\boldsymbol{0})})$ (green square).

same column partition. Besides, in this case, when one partition is fixed (curves defined by blue circles and green squares in Figure 5), we observe that the MI, whose maximal value is 2, always remains above 1 even when the partitions $\boldsymbol{w}$ and $\boldsymbol{w'}$ or $\boldsymbol{z}$ and $\boldsymbol{z'}$ are very discordant. From the coclustering point of view, this type of configuration should be very penalised, which does the CARI, but does not the MI due to its construction as a linear combination of a row distance and column distance. Indeed, the CARI takes into account in its construction, the linkage between row partition and column partition, whereas the

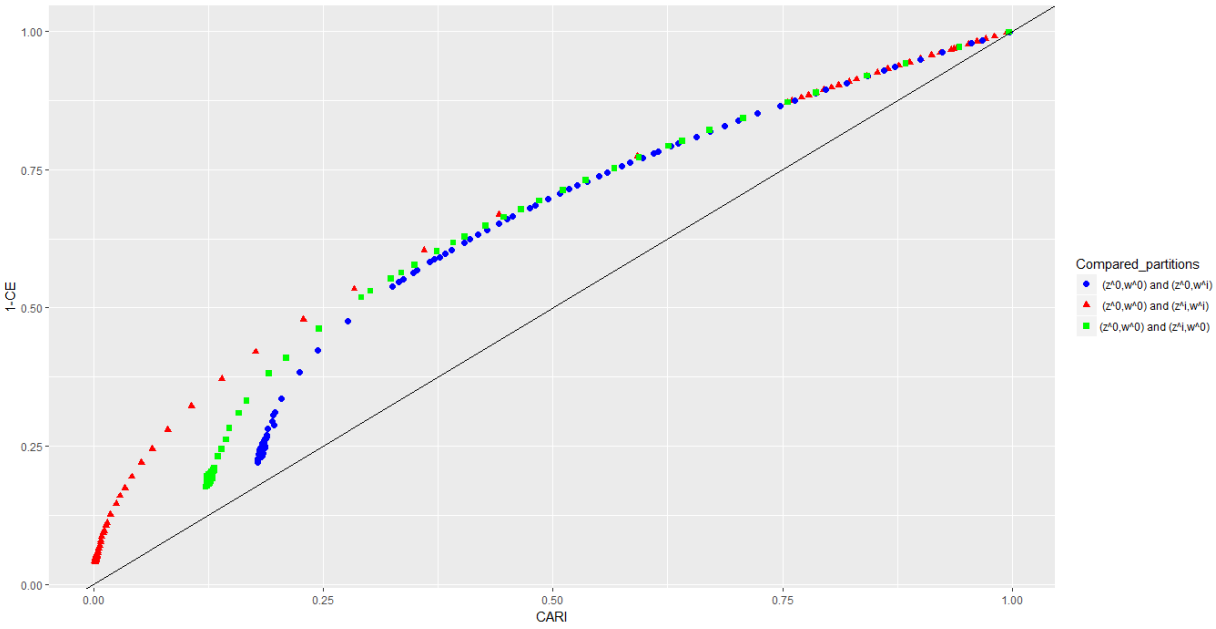

Figure 6: Comparison of the CARI's values (on the horizontal axis) versus the values of the quantity 1−CE (on the vertical axis) on a run of the procedure with $N = 10\ 000$, $(H, L) = (7, 5)$, $(I, J) = (630, 630)$ in the *balanced case*. Each index is computed at each iteration i between $(\boldsymbol{z}^{(\mathbf{0})}, \boldsymbol{w}^{(\mathbf{0})})$ and $(\boldsymbol{z}^{(\boldsymbol{i})}, \boldsymbol{w}^{(\boldsymbol{i})})$ (red triangle), between $(\boldsymbol{z}^{(\mathbf{0})}, \boldsymbol{w}^{(\mathbf{0})})$ and $(\boldsymbol{z}^{(\mathbf{0})}, \boldsymbol{w}^{(\boldsymbol{i})})$(blue circle), between $(\boldsymbol{z}^{(\mathbf{0})}, \boldsymbol{w}^{(\mathbf{0})})$ and $(\boldsymbol{z}^{(\boldsymbol{i})}, \boldsymbol{w}^{(\mathbf{0})})$ (green square).

MI deals with row partition and column partition in a separated way.

6 Conclusion

In this article, we introduced a new coclustering index named *Coclustering Adjusted Rand Index* (CARI) and based on the very popular ARI. We prove that, like the classification error proposed by Lomet (2012) but unlike the criterion developed by Wyse et al. (2017), the CARI measures the agreement between two pairs of partitions from a coclustering point of view. In addition, we show that the CARI could be computed in an efficient way, whatever the number of clusters or observations, thanks to a simple trick. These good charasteristics makes the CARI convenient and useful in a high dimensional setting, which is a highly topical issue nowadays.

A Proofs

A.1 Proof of Theorem 3.3

Theorem 3.3. *Let* $\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{z'}, \boldsymbol{w'}, \boldsymbol{n}^{\boldsymbol{zwzw'}}, \boldsymbol{n}^{\boldsymbol{zz'}}$ *and* $\boldsymbol{n}^{\boldsymbol{ww'}}$ *be defined as in Definition 3.1. Then we have the following relation,*

$$\boldsymbol{n}^{\boldsymbol{zwzw'}} = \boldsymbol{n}^{\boldsymbol{zz'}} \otimes \boldsymbol{n}^{\boldsymbol{ww'}},$$

where $\otimes$ *denotes the Kronecker product between two matrices.*

Proof. Let recall the definition of the Kronecker product. Let $\boldsymbol{A} = (a_{i,j})$ be a matrix of size $H \times H'$ and $\boldsymbol{B}$ be a matrix of size $L \times L'$. The Kronecker product is the matrix $\boldsymbol{A} \otimes \boldsymbol{B}$

of size $H \times L$ by $H' \times L'$, defined by successive blocks of size $L \times L'$. The block of the index i, j is equal to $a_{i,j} \times \boldsymbol{B}$:

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{1,1}\boldsymbol{B} & \dots & a_{1,H'}\boldsymbol{B} \\ \dots & \dots & \dots \\ a_{H,1}\boldsymbol{B} & \dots & a_{H,H'}\boldsymbol{B} \end{pmatrix}.$$

We started by remarking a common trick used in computer science. Indeed, for all $p \in \{1, \dots, HL\}$, the associated pair (h, ℓ) denoting a block of $(\boldsymbol{z}, \boldsymbol{w})$, is respectively the quotient plus 1 and the remainder plus 1 of the Euclidean division of $(p-1)$ by L. In other words, we have:

$$(p-1) = (h-1) \times L + (\ell - 1).$$

We can easily deduce that there is a bijection between each index p and the pairs (h, ℓ). In the same way, the assertion is valid for q and the pairs (h', ℓ').

The next proposition is the last step before proving the final result:

Proposition 1. *For all pairs of indices p and q associated respectively with blocks (h, ℓ) and (h', ℓ'),*

$$n_{p,q}^{zwz'w'} = n_{h,h'}^{zz'} n_{\ell,\ell'}^{ww'}.$$

Proof. We notice that the observation x_{ij} is in the block (h, ℓ) if and only if the row i is in the cluster h and the column j is in the cluster ℓ. Thanks to this remark, we can easily see that an observation x_{ij} belongs to the block (h, ℓ) and the block (h', ℓ') if and only if the row i belongs at the same time to the cluster h and the cluster h', and the column j belongs at the same time to the cluster ℓ and the cluster ℓ'.

□

With the previous results, we finally have:

$$
\begin{aligned}
\boldsymbol{n}^{\boldsymbol{zz'ww'}} &= \begin{pmatrix}
n_{1,1}^{zwz'w'} & n_{1,2}^{zwz'w'} & \cdots & n_{1,L'}^{zwz'w'} & n_{1,L'+1}^{zwz'w'} & \cdots & n_{1,H'L'}^{zwz'w'} \\
n_{2,1}^{zwz'w'} & n_{2,2}^{zwz'w'} & \cdots & n_{2,L'}^{zwz'w'} & n_{2,L'+1}^{zwz'w'} & \cdots & n_{2,H'L'}^{zwz'w'} \\
\vdots & \vdots & \ddots & & & & \vdots \\
n_{L,1}^{zwz'w'} & n_{L,2}^{zwz'w'} & \cdots & n_{L,L'}^{zwz'w'} & n_{L,L'+1}^{zwz'w'} & \cdots & n_{L,H'L'}^{zwz'w'} \\
n_{L+1,1}^{zwz'w'} & n_{L+1,2}^{zwz'w'} & \cdots & n_{L+1,L'}^{zwz'w'} & n_{L+1,L'+1}^{zwz'w'} & \cdots & n_{L+1,H'L'}^{zwz'w'} \\
\vdots & \vdots & & & & \ddots & \vdots \\
n_{HL,1}^{zwz'w'} & n_{HL,2}^{zwz'w'} & \cdots & n_{HL,L'}^{zwz'w'} & n_{HL,L'+1}^{zwz'w'} & \cdots & n_{HL,H'L'}^{zwz'w'}
\end{pmatrix} \\
&= \begin{pmatrix}
n_{1,1}^{zz'}n_{1,1}^{ww'} & n_{1,1}^{zz'}n_{1,2}^{ww'} & \cdots & n_{1,1}^{zz'}n_{1,H'}^{ww'} & n_{1,2}^{zz'}n_{1,1}^{ww'} & \cdots & n_{1,L'}^{zz'}n_{1,H'}^{ww'} \\
n_{1,1}^{z'}n_{2,1}^{ww'} & n_{1,1}^{zz'}n_{2,2}^{ww'} & \cdots & n_{1,1}^{zz'}n_{2,H'}^{ww'} & n_{1,2}^{zz'}n_{1,1}^{ww'} & \cdots & n_{1,L'}^{zz'}n_{2,H'}^{ww'} \\
\vdots & \vdots & \ddots & & & & \vdots \\
n_{1,1}^{zz'}n_{H,1}^{ww'} & n_{1,1}^{zz'}n_{H,2}^{ww'} & \cdots & n_{1,1}^{zz'}n_{H,H'}^{ww'} & n_{1,2}^{zz'}n_{1,1}^{ww'} & \cdots & n_{1,L'}^{zz'}n_{H,H'}^{ww'} \\
n_{2,1}^{zz'}n_{1,1}^{ww'} & n_{2,1}^{zz'}n_{1,2}^{ww'} & \cdots & n_{2,1}^{z,z'}n_{1,H'}^{ww'} & n_{2,2}^{zz'}n_{1,1}^{ww'} & \cdots & n_{2,L'}^{zz'}n_{1,H'}^{ww'} \\
\vdots & \vdots & & & & \ddots & \vdots \\
n_{L,1}^{zz'}n_{H,1}^{ww'} & n_{L,1}^{zz'}n_{H,2}^{ww'} & \cdots & n_{L,1}^{zz'}n_{HH'}^{ww'} & n_{L,2}^{zz'}n_{H,1}^{ww'} & \cdots & n_{L,L'}^{zz'}n_{H,H'}^{ww'}
\end{pmatrix} \\
&= \begin{pmatrix}
n_{1,1}^{zz'}\boldsymbol{n}^{\boldsymbol{ww'}} & n_{1,2}^{zz'}\boldsymbol{n}^{\boldsymbol{ww'}} & \cdots & n_{1,L'}^{zz'}\boldsymbol{n}^{\boldsymbol{ww'}} \\
n_{2,1}^{zz'}\boldsymbol{n}^{\boldsymbol{ww'}} & n_{2,2}^{zz'}\boldsymbol{n}^{\boldsymbol{ww'}} & \cdots & n_{2,L'}^{zz'}\boldsymbol{n}^{\boldsymbol{ww'}} \\
\vdots & & \ddots & \vdots \\
n_{L,1}^{zz'}\boldsymbol{n}^{\boldsymbol{ww'}} & n_{L,2}^{zz'}\boldsymbol{n}^{\boldsymbol{ww'}} & \cdots & n_{L,L'}^{zz'}\boldsymbol{n}^{\boldsymbol{ww'}}
\end{pmatrix} \\
&= \boldsymbol{n}^{\boldsymbol{zz'}} \otimes \boldsymbol{n}^{\boldsymbol{ww'}}.
\end{aligned}
$$

□

A.2 Proof of Corollary 3.4

Corollary 3.4.

1. *$\forall (p,q) \in (H \times L) \times (H' \times L')$, we have the following relations between the margins,*

$$n_{\bullet,q}^{zwzw'} = n_{\bullet h'_q}^{zz'} \otimes n_{\bullet,\ell'_q}^{ww'} \text{ and } n_{p,\bullet}^{zwzw'} = n_{h_p,\bullet}^{zz'} \otimes n_{\ell_p,\bullet}^{ww'}.$$

2. *The CARI associated with the contingency table $\boldsymbol{n^{zwzw'}}$ defined as in Equation (3) remains symmetric, that is to say,*

$$CARI((\boldsymbol{z},\boldsymbol{w}),(\boldsymbol{z'},\boldsymbol{w'})) = CARI((\boldsymbol{z'},\boldsymbol{w'}),(\boldsymbol{z},\boldsymbol{w})).$$

Proof.

1. This assertion forms part of the known properties of the Kronecker product.

2. The proof of this result is the direct consequence of the following Lemma:

   Lemma A.1. *Let $\boldsymbol{z},\boldsymbol{w},\boldsymbol{z'},\boldsymbol{w'},\boldsymbol{n^{zz'}}$, and $\boldsymbol{n^{ww'}}$ be defined as in Definition 3.1 and $\boldsymbol{n^{zwzw'}}$ be defined according to Theorem 3.3. Then we have,*

$$\boldsymbol{n^{z'w'zw}} = t(\boldsymbol{n^{zwz'w'}}),$$

   where t denotes the tranpose of a matrix.

   Proof. Thanks to the property of the Kronecker product with the transpose, we have,

$$\begin{aligned} \boldsymbol{n^{z'w'zw}} &= \boldsymbol{n^{z'z}} \otimes \boldsymbol{n^{w'w}} \\ &= t(\boldsymbol{n^{zz'}}) \otimes t(\boldsymbol{n^{ww'}}) \end{aligned}$$

$$
\begin{aligned}
&= t\left(\boldsymbol{n}^{zz'} \otimes \boldsymbol{n}^{ww'}\right) \\
&= t(\boldsymbol{n}^{zwz'w'}).
\end{aligned}
$$

□

□

References

Aubert, J., T. Ha, and T. MaryHuard (2014). Modele à blocs latents pour l'analyse de données métagénomiques. In $46^{ème}$ *journées de Statistiques de la SFdS*, Rennes.

Charrad, M., Y. Lechevallier, G. Saporta, and M. Ben Ahmed (2010). Détermination du nombre de classes dans les méthodes de bipartitionnement. In *17ème Rencontres de la Société Francophone de Classification*, Saint-Denis de la Réunion, pp. 119–122.

Dhillon, I. S., S. Mallela, and D. S. Modha (2003). Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 89–98. ACM.

Fowlkes, E. B. and C. L. Mallows (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78*(383), 553–569.

Govaert, G. and M. Nadif (2013). *Co-Clustering*. ISTE Ltd and John Wiley & Sons, Inc.

Hartigan, J. A. (1975). *Clustering Algorithms* (99th ed.). John Wiley & Sons.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.

Jagalur, M., C. Pal, E. Learned-Miller, R. T. Zoeller, and D. Kulp (2007). Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. *BMC Bioinformatics, 8*(10), S5.

Lomet, A. (2012). *Sélection de modèle pour la classification croisée de données continues.* Thèse, Université de Technologie de Compiègne.

Meilă, M. (2007). Comparing clusterings - an information based distance. *Journal of Multivariate Analysis, 98*(5), 873–895.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850.

Robert, V., G. Celeux, and C. Keribin (2015). Un modèle statistique pour la pharmacovigilance. In *47èmes Journées de Statistique de la SFdS*.

Santos, J. and M. Embrechts (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. *Artificial neural networks*, 175–184.

Saporta, G. and G. Youness (2002). Comparing two partitions: Some proposals and experiments. In *Compstat*, pp. 243–248. Springer.

Shan, H. and A. Banerjee (2008). Bayesian co-clustering. In *Eighth IEEE International Conference on Data Mining*, pp. 530–539.

Vinh, N. X., J. Epps, and J. Bailey (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research, 11*(Oct), 2837–2854.

Wyse, J., N. Friel, and P. Latouche (2017). Inferring structure in bipartite networks using the latent blockmodel and exact ICL. *Network Science, 5*(1), 45–69.

Youness, G. and G. Saporta (2004). Une méthodologie pour la comparaison de partitions. *Revue de statistique appliquée, 52*(1), 97–120.