Check for updates

the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2021), 74, 203–231
© 2020 The British Psychological Society
www.wileyonlinelibrary.com

A close-up comparison of the misclassification error distance and the adjusted Rand index for external clustering evaluation

José E. Chacón
Departamento de Matemáticas, Universidad de Extremadura, Badajoz, Spain

The misclassification error distance and the adjusted Rand index are two of the most common criteria used to evaluate the performance of clustering algorithms. This paper provides an in-depth comparison of the two criteria, with the aim of better understand exactly what they measure, their properties and their differences. Starting from their population origins, the investigation includes many data analysis examples and the study of particular cases in great detail. An exhaustive simulation study provides insight into the criteria distributions and reveals some previous misconceptions.

1. Introduction

The adjusted Rand index (ARI), introduced in Hubert and Arabie (1985), is one of the most commonly used measures of performance for clustering evaluation. Indeed, it was the recommended choice in the seminal paper of Milligan and Cooper (1986), where five criteria were examined with regard to the task of comparison of hierarchical clustering algorithms across different hierarchy levels. Their recommendation is based on the fact that, for data in the null case (i.e., for a synthetic sample with randomly assigned class labels, showing no significant cluster structure), the ARI was the only index that produced a flat response curve across hierarchy levels, with mean values close to zero, hence indicating that the agreement between the randomly assigned labels and the algorithm solution was due to chance.

Another popular measure for clustering validation, not included in Milligan and Cooper's study, is the misclassification error distance (MED). Its first appearance in the literature dates back at least to Régnier (1965), where it was introduced as a distance between partitions of a finite set, and it was called the transfer distance. It is also referred to as the partition distance (Gusfield, 2002) or maximum matching distance (Rossi, 2015). Steinley (2004) supports Milligan and Cooper's recommendation by inspecting the performance of both the ARI and the MED in an exhaustive simulation study. On the other hand, Meilă (2016) suggests that the MED 'comes closest to satisfying everyone' in terms of its properties and ease of interpretation, while Denœud and Guénoche (2006) state that the MED is much appropriate for small sample sizes from their study of all the clusterings at a close number of transfers from a given one, and von Luxburg (2010) considers the MED as 'the most convenient choice from a theoretical point of view'.

*Correspondence should be addressed to José E. Chacón, Departamento de Matemáticas, Universidad de Extremadura, E-06006 Badajoz, Spain. (email: jechacon@unex.es).

DOI:10.1111/bmsp.12212

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

It must be stressed that both criteria are commonly categorized as 'external', in the sense that they are used to measure the performance of a data-based clustering algorithm against a true cluster structure, known in advance in a simulation scenario or after a data inspection by an expert, which is taken as the ideal clustering solution but is external to the clustering methodology itself. Internal criteria (such as those based on cohesion, entropy, cluster separation, etc.) are also frequently used, but they are not the focus of this paper; see Hennig (2019) for a thorough review of internal cluster validation indexes.

The Rand index belongs to the class of similarity measures based on inspecting pairs of points. The ARI is a variant of the Rand index obtained after a correction for chance. The effect of this correction on such pair-counting measures was explored in detail in Albatineh *et al*. (2006) and Warrens (2008a). The need for this correction is motivated, in part, by the features of the Rand index shown in Fowlkes and Mallows (1983), who computed its expected value and variance, and noted that its possible range of values is quite narrow.

Some other properties of the ARI were shown in Steinley (2003, 2004), where it was observed that the ARI appears to be practically invariant to changes in the number of clusters, objects, and relative cluster size. Maximization of the ARI was considered in Brusco and Steinley (2008) and Steinley, Hendrickson and Brusco (2015) (see also Messatfa, 1992) and its variance was obtained in Steinley, Brusco and Hubert (2016), whereas the issue of its expected value has recently been re-examined in Steinley and Brusco (2018). Further comparisons to other measures can be found in Pfitzner, Leibbrandt and Powers (2009), Vinh, Epps and Bailey (2010) or Albatineh and Niewiadomska-Bugaj (2011).

The MED was studied in Meilă (2005, 2007, 2016), where it was shown that it is a true metric satisfying convex additivity. It gives 'meaningful and interpretable results' when comparing partitions that are not very dissimilar (the most common situation in practical applications), but it suffers some resolution loss as the clusterings become more different. Its maximum value was investigated in Charon *et al.,* (2006), Charon, Denœud and Hudry (2007) and Denœud (2008). In any case, there exist very few studies that closely examine the behaviour of the Rand index, the ARI and the MED together, with notable representatives being Steinley (2004), Denœud and Guénoche (2006) and Meilă (2016).

This paper aims to provide further comparisons between these important measures, the MED and the ARI, at several levels. Indeed, many other external criteria could be considered as well, and they are also reviewed in the aforementioned comparative studies, but here the discussion is restricted to the former two because they are usually recognized as the main criteria used in practice. The close-up inspection that is provided here examines a wide range of features, previously little explored, if at all. Section 2 first glances through their population origins (i.e., their counterparts in the case where the true underlying data distribution is fully known) and then elaborates on their traditional, and more common, data-based versions. The comparison of these empirical analogues is the subject of Sections 3 (theoretically) and 4 (through simulation). The theoretical study comprises their computation, some illustrations by means of simple examples in order to better understand precisely what they measure, and an analysis of their extreme values in relation to the case of independent clusterings. The simulation scenarios investigate the distributions of the criteria in the null case and how they evolve as two clusterings drift apart from perfect agreement. Finally, Section 5 discusses the new findings and their implications.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2. Population and empirical distances between clusterings

2.1. The population version of cluster analysis

Cluster analysis is mostly posed as a sample problem, and perhaps that is one of the reasons why many authors have called attention to the lack of theoretical results for clustering (von Luxburg & Ben-David, 2005; Milligan, 1996), as opposed to regression or classification, where the population background is much more clearly established.

Traditionally, the aim of clustering techniques is to provide a partitioning of a data set into groups. For that purpose, it suffices to have an algorithm which is appropriate for the data set at hand. However, from a statistical perspective, such a given data set is not simply a set of points in the space, but a sample from some probability distribution P. Hence, the aim of clustering methodologies cannot be reduced to partitioning only the data set at hand; they must provide a mechanism to assign group labels to any point in the space, or, at least, to all the points in the sample space, since they could have been equally drawn as sample points. Such a view of clustering is shared by many authors, including Györfi *et al.* (2002, p. 245), Ben-David, von Luxburg and Pál (2006), Klemelä (2009, p. 196), Chacón (2015) or Wasserman (2018, Section 2.3).

Hence, the object that clustering algorithms should produce is not just a partition of the data set, but a whole-space partition. This means that if Ω denotes the sample space, a whole-space clustering is a class of sets $\mathscr{C} = \{C_1, \ldots, C_r\}$ such that $C_i \cap C_j = \emptyset$ for all $i \neq j$ and $C_1 \cup \cdots \cup C_r = \Omega$. Indeed, most existing clustering methodologies are able to produce this type of object; this is the case, for instance, for K-means clustering, modal clustering or mixture model clustering (see Chacón, 2015). Obviously, any partition of Ω induces a partition of the observed data set as well. To avoid confusion, these are referred to as a whole-space clustering and a clustering of the data, respectively. Also note that both objects can have a population version (the partition that would be made if the true underlying distribution were fully known) and a data-based version (the partition that would be made after observing the data).

That having been made clear, to evaluate the performance of clustering methods from a statistical point of view it is necessary to employ a distance between whole-space clusterings. While there exist many notions of distance between partitions of a finite set (Day, 1981; Meilă, 2016) notions of distance between whole-space clusterings do not abound in the literature. Two of them are described next.

First, since the parts of a clustering (i.e., the clusters) are sets, it seems natural for distances between clusterings to be built upon a notion of discrepancy between sets. One way to express the discrepancy between two sets C and D is by quantifying the content of their symmetric difference $C \Delta D$. This difference is defined as the elements that C and D do not have in common; that is, $C \Delta D = (C \cup D) \setminus (C \cap D)$. Then, taking into account the distinctive features of a partition, this natural distance between sets can be extended to define a distance between two clusterings $\mathscr{C} = \{C_1, \ldots, C_r\}$ and $\mathscr{D} = \{D_1, \ldots, D_s\}$, by adding up the contributions of the regions that their most similar clusters do not have in common. Specifically, Chacón (2015) defined the distance in measure between $\mathscr{C}$ and $\mathscr{D}$ as

$$d_{\mathrm{M}}(\mathscr{C}, \mathscr{D}) = \frac{1}{2} \min_{\sigma \in \mathscr{P}_s} \sum_{i=1}^{s} P\left(C_i \Delta D_{\sigma(i)}\right), \tag{1}$$

where $\mathscr{P}_s$ is the set of all permutations of s elements and, without loss of generality, it is assumed that $r \leq s$ so that $\mathscr{C}$ would be enlarged by adding $s - r$ empty sets $C_{r+1} = \cdots = C_s = \emptyset$ if necessary. More intuitively, $d_{\mathrm{M}}(\mathscr{C}, \mathscr{D})$ represents the minimum probability mass that needs to be moved (or relabelled) to transform $\mathscr{C}$ into $\mathscr{D}$, or vice versa.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The above $d_{\mathrm{M}}(\mathscr{C},\mathscr{D})$ is a clustering distance, in the sense of Ben-David, von Luxburg and Pál (2006, Definition 3). Nevertheless, these authors considered a different distance between whole-space clusterings, $d_{\mathrm{H}}(\mathscr{C},\mathscr{D})$, which they called the Hamming distance. This second distance is more closely related to the Rand index (as detailed below), since it is defined as the probability that two independent random observations (drawn from P) belong to the same cluster with respect to one of the clusterings and to different clusters with respect to the other clustering. Hence, it can be shown that an explicit expression for this Hamming distance is.

$$d_H(\mathscr{C},\mathscr{D}) = \sum_{i=1}^{r} P^2(C_i) + \sum_{j=1}^{s} P^2\left(D_j\right) - 2\sum_{i=1}^{r}\sum_{j=1}^{s} P^2\left(C_i \cap D_j\right). \tag{2}$$

The dependence of this measure on squared probabilities may appear somewhat unnatural, but it is a consequence of the fact that it is based on comparing the cluster labels of pairs of points.

2.2. *Comparing two clusterings of the data*

In a simulation setting, where the true underlying distribution P is fully known, it is possible to compute the ideal population clustering, that is, the whole-space partition that would be made on the basis of this knowledge of P (this ideal partition varies from one methodology to another, depending on the type of cluster sought). Hence, it is natural to evaluate the performance of a clustering technique by means of the distance from the data-based clustering produced to its population counterpart. Since both are clusterings of the whole space, any of the previously mentioned distances between whole-space clusterings can be employed.

Of course, things are different when dealing with real data. Suppose that we have observed n data points $\mathscr{X} = \{x_1,\ldots,x_n\}$. Even if the usual methods are able to produce whole-space clusterings just with the information provided by $\mathscr{X}$, the fact that a clustering distance depends on P (Ben-David, von Luxburg & Pál, 2006), which is unknown for real data sets, implies that to compute the clustering distance in practice it is necessary to replace P by the empirical distribution P_n, which assigns probability mass $1/n$ to each data point. This means that only the labels of the data points are used in the comparison between the two clusterings, so that a distance between whole-space clusterings becomes in fact a distance between two clusterings of the data.

When this reasoning is applied to the two distances in the previous subsection, it results in two well-known distances between partitions of a finite set. To see this, given two partitions $\mathscr{C} = \{C_1,\ldots,C_r\}$ and $\mathscr{D} = \{D_1,\ldots,D_s\}$ of $\mathscr{X}$, with $r \le s$, denote by n_{ij}, $n_{i+} = \sum_{j=1}^{s} n_{ij}$ and $\mathrm{n}_{+j} = \sum_{i=1}^{r} n_{ij}$ the cardinalities of $C_i \cap D_j$, C_i and D_j, respectively. The $(r \times s)$ matrix $\mathbf{N} = \left(n_{ij}\right)$ is known as the confusion matrix (or contingency table), and the vectors $(n_{1+},\ldots,n_{r+})$ and $(n_{+1},\ldots,n_{+s})$ constitute its rowwise and columnwise margins, respectively. Then, taking into account that $P\left(C_i \Delta D_j\right) = P(C_i) + P\left(D_j\right) - 2P\left(C_i \cap D_j\right)$, it follows that the empirical version of the distance in measure (1) is

$$d_{\mathrm{M}}(\mathscr{C},\mathscr{D}) = \frac{1}{2}\min_{\sigma\in\mathscr{P}_s} n^{-1}\sum_{i=1}^{s}\left\{n_{i+} + n_{+\sigma(i)} - 2n_{i,\sigma(i)}\right\} = 1 - n^{-1}\max_{\sigma\in\mathscr{P}_s}\sum_{i=1}^{r} n_{i,\sigma(i)},$$

which coincides with the definition of the misclassification error distance (see Meilă, 2005), so that it will henceforth be denoted by $\mathrm{MED}(\mathscr{C},\mathscr{D}) = \underset{\mathrm{M}}{d}(\mathscr{C},\mathscr{D})$ (or simply MED, if it

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is obvious which clusterings are being compared). The MED inherits from its population version a clear interpretation as the minimum proportion of data points that would need to be relabelled so that $\mathscr{C}$ and $\mathscr{D}$ coincided, and that is why it is also known as transfer distance (Régnier, 1965).

On the other hand, the empirical equivalent of the Hamming distance (2) is.

$$d_H(\mathscr{C},\mathscr{D}) = n^{-2}\left\{\sum_{i=1}^{r} n_{i+}^2 + \sum_{j=1}^{s} n_{+j}^2 - 2\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}^2\right\}, \tag{3}$$

which is also known as the equivalence mismatch coefficient (Mirkin, 1996; Mirkin & Chernyi, 1970, p. 241) or as the *n*-invariant Mirkin metric (Meilă, 2016). Being a sample equivalent of (2), $d_H(\mathscr{C},\mathscr{D})$ equals the proportion of pairs $\{(x_k,x_l) : k,l = 1,\ldots,n\}$ that belong to the same cluster in one of the clusterings and to different clusters in the other clustering. Note that, somewhat artificially, this empirical version of the Hamming distance takes into account data pairs of type (x_k, x_k) as well.

In statistical terms, if X, Y are independent random variables with distribution P and we denote by I_A the indicator function of a set A, the squared probability $P^2(C_i) = E[I_{C_i}(X)I_{C_i}(Y)]$ appearing in (2) is estimated in (3) by the observed value of the V-statistic $n^{-2}\sum_{k,l=1}^{n} I_{C_i}(x_k)I_{C_i}(x_l) = n_{i+}^2/n^2 = P_n^2(C_i)$. However, U-statistics theory (Lee, 1990) shows that a better estimate of $P^2(C_i)$ is $\binom{n}{2}^{-1}\sum_{1\le k<l\le n} I_{C_i}(x_k)I_{C_i}(x_l) = \binom{n}{2}^{-1}\binom{n_{i+}}{2}$. Reasoning similarly for the other terms in (2) and making these changes everywhere in (3) yields the definition of the Rand distance.

$$\mathrm{RD}(\mathscr{C},\mathscr{D}) = \binom{n}{2}^{-1}\left\{\sum_{i=1}^{r}\binom{n_{i+}}{2} + \sum_{j=1}^{s}\binom{n_{+j}}{2} - 2\sum_{i=1}^{r}\sum_{j=1}^{s}\binom{n_{ij}}{2}\right\} \tag{4}$$

which equals the proportion of unordered data pairs $\mathscr{U} = \{\{x_k,x_l\} : k,l = 1,\ldots,n, k\neq l\}$ that belong to the same cluster in one of the clusterings and to different clusters in the other clustering (see Filkov & Skiena, 2004). This distance was called the *symmetrical difference distance* in Denœud and Guénoche (2006), and it is also considered in Azizyan *et al*. (2015), under the name *pairwise clustering loss*. In any case, it is not hard to check that

$$\mathrm{RD}(\mathscr{C},\mathscr{D}) = \frac{n}{n-1} d_H(\mathscr{C},\mathscr{D}),$$

so in fact there is little difference between these two empirical versions of d_{H}.

Instead of measuring the dissimilarity between clusterings using a distance, clustering comparisons can be based on indices that quantify the agreement between them, with values close to 1 indicating greater similarity. In this sense, the Rand index (Rand, 1971) is defined as $\mathrm{RI}(\mathscr{C},\mathscr{D}) = 1 - \mathrm{RD}(\mathscr{C},\mathscr{D})$. An important feature of the RI is that it also has a clear interpretation as the proportion of unordered data pairs that belong either to the same cluster or to different clusters in both clusterings. However, Fowlkes and Mallows (1983) noted that, when comparing two clusterings with $r = s$, the range of possible values of the RI is quite narrow and its expected value E(RI) quickly approaches 1 as $r = s \to n$. This expectation is meant with respect to a random choice of the entries of the confusion

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

matrix, while keeping its margins fixed, intended to reproduce a null scenario corresponding to independent clusterings. To amend this problem, Hubert and Arabie (1985) proposed to correct the RI for chance, so that it yields an expected value of zero in such a null scenario, and introduced the adjusted Rand index $\mathrm{ARI}(\mathscr{C},\mathscr{D}) = \{\mathrm{RI}(\mathscr{C},\mathscr{D}) - E(\mathrm{RI})\}/\{1 - E(\mathrm{RI})\}$. Milligan and Cooper (1986) showed that, in addition, this correction also results in a much wider range of possible values for the ARI compared to the RI. A recent in-depth study of the properties of the RI can be found in Warrens and van der Hoef (2020).

The most notable loss resulting from this correction is the interpretation; for example, it is not easy to discern what an ARI value of .78 means, or if an ARI of .82 for two clusterings denotes a higher agreement between them than an ARI of .73 for a different pair of clusterings, since the baseline $E(\mathrm{RI})$ could be different. In a series of papers (later collected into a single volume), Goodman and Kruskal (1979) emphasized, for association measures in cross-classifications, the importance of having a clear operational interpretation. Wallace (1983) raised some doubts with respect to the choice of the null scenario in the computation of $E(\mathrm{RI})$ and, more recently, Gates and Ahn (2017) showed that the use of different null models for index adjustment can lead to disparate conclusions.

Despite these drawbacks, the ARI is one of the most popular and most widely employed indicators for clustering comparison, in close competition with the MED. Hence, one of the main contributions of this paper is to provide a detailed examination of both of them, by means of simple examples, to help understand their behaviour and their differences. Additional references providing in-depth investigation of these criteria include Warrens (2008b), Steinley, Brusco and Hubert (2016) and Steinley and Brusco (2018) for the ARI, and Charon *et al*. (2006), Charon, Denœud and Hudry (2007) and Denœud (2008) for the MED.

Here, since the MED is a distance and the ARI is an index, to facilitate their comparison the ARI will first be transformed into a semi-metric (since it is not guaranteed to satisfy the triangle inequality), which will be called the adjusted Rand distance and is defined as

$$\mathrm{ARD}(\mathscr{C},\mathscr{D}) = 1 - \mathrm{ARI}(\mathscr{C},\mathscr{D}) = \mathrm{RD}(\mathscr{C},\mathscr{D})/E(\mathrm{RD}),$$

that is, as the Rand distance normalized by its expected value under the null model. Thus, the ARD has unit expected value under the null model.

3. Detailed comparison of the MED and the ARD

In the following, several aspects of the MED and the ARD will be compared in detail. In Section 3.1 explicit computation of the two criteria is addressed. Then, in Section 3.2, the differences between the two are illustrated through several specific examples. In Section 3.3 an exhaustive study of the simplest case of a 2×2 confusion matrix is provided, with emphasis on exploring the most dissimilar situation between two clusterings. Finally, in Section 3.4, some of the lessons learned from the 2×2 case are generalized for two clusterings of arbitrary size.

3.1. Computation

One undeniable advantage of the ARD over the MED is its simpler definition, which readily translates into a much simpler computation.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Let us write $x_k \sim_{\mathscr{C}} x_l$ if the data points x_k and x_l belong to the same cluster in $\mathscr{C}$ (and $x_k \nsim_{\mathscr{C}} x_l$ otherwise), and consider the cardinalities of the sets of (unordered) data pairs that cover all the possibilities of belonging either to the same or to different clusters in $\mathscr{C}$ and $\mathscr{D}$, denoted by

$$a = |\{\{x_k, x_l\} \in \mathscr{U} : x_k \sim_{\mathscr{C}} x_l, x_k \sim_{\mathscr{D}} x_l\}|,$$

$$b = |\{\{x_k, x_l\} \in \mathscr{U} : x_k \sim_{\mathscr{C}} x_l, x_k \nsim_{\mathscr{D}} x_l\}|,$$

$$c = |\{\{x_k, x_l\} \in \mathscr{U} : x_k \nsim_{\mathscr{C}} x_l, x_k \sim_{\mathscr{D}} x_l\}|,$$

$$d = |\{\{x_k, x_l\} \in \mathscr{U} : x_k \nsim_{\mathscr{C}} x_l, x_k \nsim_{\mathscr{D}} x_l\}|.$$

Then it is clear that $\mathrm{RD}(\mathscr{C}, \mathscr{D}) = \binom{n}{2}^{-1}(b+c)$. Moreover, Steinley (2004) provided the very simple formula $E(\mathrm{RI}) = \binom{n}{2}^{-2}\{(a+b)(a+c)+(c+d)(b+d)\}$, which entails that $E(\mathrm{RD}) = \binom{n}{2}^{-2}\{(a+b)(b+d)+(a+c)(c+d)\}$, so that

$$\mathrm{ARD}(\mathscr{C}, \mathscr{D}) = \binom{n}{2}(b+c)/\{(a+b)(b+d)+(a+c)(c+d)\}. \tag{5}$$

This is very easy to implement, taking into account that $a = \sum_{i=1}^{r}\sum_{j=1}^{s}\binom{n_{ij}}{2}$, $b = \sum_{j=1}^{s}\binom{n_{+j}}{2} - a$, $c = \sum_{i=1}^{r}\binom{n_{i+}}{2} - a$ and $d = \binom{n}{2} - a - b - c$ can be immediately computed from the confusion matrix (Jain & Dubes, 1988, Section 4.4.1).

In contrast, computation of the MED requires solving a discrete minimization problem over $\max\{r!, s!\}$ possible inputs, so its implementation is not that simple, which surely hinders its usage. To fully describe the problem, assume that $r \leq s$ and define $n_{i+} = n_{ij} = 0$ for all $i = r+1, \ldots, s$ (if any). Writing $m_{ij} = n_{i+} + n_{+j} - 2n_{ij}$ for $i, j = 1, \ldots, s$, then computation of the MED involves finding $\min_{\sigma \in \mathscr{P}_s}\sum_{i=1}^{s} m_{i,\sigma(i)}$, where $\mathscr{P}_s$ denotes the set of all possible permutations of s elements. Despite its apparent complexity, this is a form of the well-known assignment problem, and very efficient algorithms exist to find its solution (see Burkard, Dell'Amico & Martello, 2009). Appendix A offers a simple implementation using the popular R language (R Core Team, 2019).

3.2. Examples

To help understand what the ARD and the MED represent and how they are computed in practice it is useful to start with some simple real data examples.

The first example uses the famous iris data set (Anderson, 1935), including four measurements on $n = 150$ flowers of three species of iris: *Iris setosa*, *I. versicolor* and *I.*

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. Confusion matrix for normal mixture clustering against the true cluster labels for the iris data set

	Data-based labels		
True labels	1	2	3
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	1	49

virginica. Clustering these data, e.g., using a normal mixture model (Fraley & Raftery, 2002) with $G = 3$ components, results in the confusion matrix given in Table 1.

Thus, $\mathrm{MED} = (2+1)/150 = 0.02$ since only three data points would need to be relabelled for the two partitions to coincide. On the other hand, there are $(2+1) \times (48+49) = 291$ data pairs that belong to the same cluster in one of the partitions and to different clusters in the other, and that accounts for a proportion of $\mathrm{RD} = 291/\binom{150}{2} = 0.026$ of the total number of possible data pairs. Finally, using Equation (5) the adjusted Rand distance for those two partitions is ARD = 0.059.

This is a very simple example because the two partitions have the same (small) number of clusters, and besides, they are quite similar. Nevertheless, it is helpful to perceive the differences between the MED, the RD and the ARD. Here, perhaps the MED is the easiest criterion to compute and interpret, since it only involves counting misplaced *individual data points*. Obtaining the RD from the confusing matrix (by eye) is a bit more complex, since it involves counting *data pairs*. And the corrected version, ARD, lacks the interpretability of the former two, but it still yields a very small number, indicating that the two partitions have a high degree of agreement.

Our second example concerns the DLBCL data set, introduced in Aghaeepour *et al*. (2013). It contains the records of the CD3, CD5 and CD19 antibodies in a set of $n = 8183$ cells of a patient with diffuse large B-cell lymphoma (DLBCL), along with the true cluster labels in five groups ($A, \ldots E$) found manually by an expert. In Chacón (2019), this data set was analysed using several component merging techniques for mixture model clustering, in particular through the so-called `modclust` and `entmerge` methods. The former suggested the existence of three clusters, while the latter correctly identified five clusters; both confusion matrices are given in Table 2.

Regarding the confusion matrix for the `modclust` labels, again it is not hard to compute the MED: the group matching leading to a higher degree of agreement would be

Table 2. Confusion matrices for the clusterings obtained by `modclust` and `entmerge` for the DLBCL data set, as compared to the true cluster labels

	`modclust` labels			`entmerge` labels				
True labels	1	2	3	1	2	3	4	5
A	47	197	7	16	7	0	14	214
B	0	1,408	153	0	146	929	417	69
C	0	278	1,216	0	1,191	81	63	159
D	0	62	0	0	0	0	0	62
E	4,813	2	0	4,809	0	0	1	5

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

B with 2, *C* with 3 and *E* with 1, whereas the remaining 47 + 197 + 7 + 153 + 278 + 62 + 2 = 276 data points would need to be relabelled to make the two partitions coincide, thus yielding MED = 746/8183 = 0.091. Similarly, for the `entmerge` labels it can be checked that MED = 1040/8183 = 0.127, so that the `modclust` clustering is closer to the true expert labels regarding the MED, despite showing a smaller number of clusters. The reason is that, despite the fact that the `entmerge` method returned the true number of clusters, its assignments to clusters 4 and 5 were so poor (especially the splitting of cluster *B* into two significant groups in clusters 3 and 4) that a large number of relabellings is needed to make this partition equal to the true one. In contrast, the ARD for these two confusion matrices can be computed to be ARD = 0.112 and ARD = 0.097 for the `modclust` and `entmerge` partitions, respectively. As noted before, this does not yield such an intelligible comparison regarding the relative closeness of the two data-based partitions to the true clustering, because the baseline *E*(RD) is different for the two contingency tables. Nevertheless, it must be noted that the unadjusted distances (RD = 0.055 and RD = 0.047, respectively) also suggest that, in terms of data-pair disagreements, the `entmerge` clustering seems to be slightly closer to the expert partition than the `modclust` one.

The two previous examples illustrate the common scenario in real data analysis, where data-based partitions are not too dissimilar from the true clustering. To finish this subsection, a synthetic example concerning quite distant partitions is examined. The confusion matrix shown in Table 3 corresponds to the two assignments of $n = 13$ objects into $r = s = 5$ clusters in Steinley (2003, Table 2).

To appreciate how distant these two clusterings are, it is worth noting that ARD = 1.164, greater than 1, meaning that the disagreement between the two is higher than the average that would be obtained if the labels were randomly assigned (following the null model). The number of data pairs that are in the same group in one clustering and in different groups in the other can be computed to be 22, out of the total of $\binom{13}{2} = 78$ possible data pairs, which leads to RD = 22/78 = 0.282 And, by considering any permutation of the columns of the confusion matrix that preserves all its diagonal entries as 1, adding up the off-diagonal figures leads to MED = 8/13 = 0.615. This example further illustrates how counting 'discordant' data pairs seems to be less intuitive than counting 'discordant' individual data points. But also, it shows that the permutation for which the MED is attained may not be unique: for instance, rearranging the columns of the confusion matrix according to the permutation (45123) yields the same MED value, as already noted in Steinley (2003, 2004). In any case, it is easy to check that the values of the RD and the ARD also remain the same under that permutation. Such a phenomenon is

Table 3. Confusion matrix for Steinley's example

	Clustering $\mathscr{D}$				
Clustering *C*	D_1	D_2	D_3	D_4	D_5
C_1	1	0	1	1	0
C_2	0	1	0	0	1
C_3	1	0	1	0	1
C_4	0	1	0	1	0
C_5	1	0	1	0	1

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

expected to occur for the comparison of very dissimilar clusterings; for example, in the extreme case where the confusion matrix $\mathbf{N}$ has all its entries equal to 1 (representing independent label assignments), any permutation of its column leads to the same MED, RD and ARD values.

3.3. Two clusters in each clustering

In order to gain a deeper understanding of the behaviour of the MED and the ARD the next step is to analyse in detail the simplest scenarios. Arguably, the simplest comparison between two clusterings arises when either $r = 1$ or $s = 1$, but that could be considered a degenerate case, since in fact one of the partitions would show no clusters. So the next simplest case is $r = s = 2$; we will focus our attention on this case first, and then we will generalize some of our findings to the case of arbitrary r and s.

Independently of the criterion employed to compare clusterings, any researcher would probably agree that having a diagonal confusion matrix is synonymous with a perfect agreement between the two partitions. But that is also the case if the confusion matrix is anti-diagonal, which means, for $r = s = 2$, that

$$\mathbf{N} = \begin{pmatrix} 0 & n_{12} \\ n_{21} & 0 \end{pmatrix}.$$

This clearly illustrates a key difference between classification and clustering: since classification is a supervised learning problem, the training data are already equipped with labels with precise meanings, and hence an anti-diagonal confusion matrix must be interpreted as the result of a totally incorrect classification; in contrast, a clustering algorithm labels the groups as it finds them and, hence, the coding designation is not important (group 1 might just as well have been called group 2, and vice versa) so that an anti-diagonal confusion matrix also represents perfect agreement, since the groups discovered are exactly the same, only differing in their (arbitrary) names. Mathematically, this means that distances between clusterings must be invariant with respect to permutations of the cluster labels (Meilă, 2012).

It is precisely the way to measure deviations from the diagonal or anti-diagonal situation that gives rise to the different distances between clusterings. For the case $r = s = 2$, let us consider the 2×2 confusion matrix $\mathbf{N} = (n_{ij})$, and denote by $d_1 = n_{11} + n_{22}$ and $d_2 = n_{12} + n_{21}$ the total sum of its diagonal and anti-diagonal entries, respectively. In this case, it is not hard to show that the MED and the RD can be simply expressed as.

$$\mathrm{MED} = n^{-1} \min\{d_1, d_2\},$$

$$\mathrm{RD} = \binom{n}{2}^{-1} d_1 d_2.$$

To graphically appreciate the differences between the MED and the RD, and noting that $d_2 = n - d_1$, Figure 1 shows the possible values of these criteria for $n = 20$, as a function of d_1. The linear and quadratic appearances of the MED and the RD, respectively, are

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

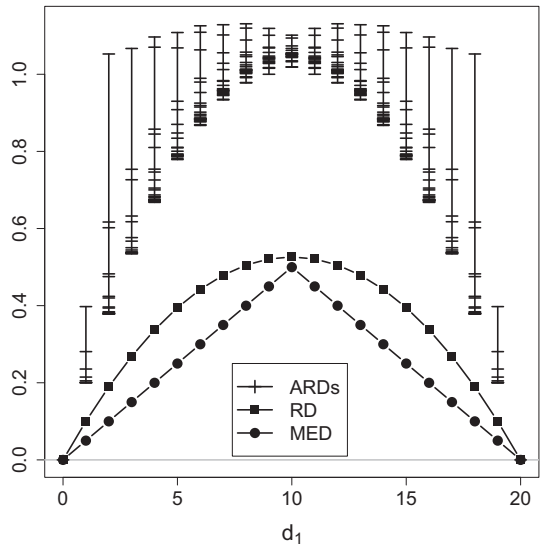

Figure 1. MED (solid circles), RD (solid squares) and possible ARD values (ticks) as a function of d_1 for $n = 20$.

explained by the fact that they can be equivalently expressed as $\text{MED} = \frac{1}{2} - \left|d_1/n - \frac{1}{2}\right|$ and $\text{RD} = \binom{n}{2}^{-1} d_1(n - d_1)$.

On the other hand, it is not possible to express the ARD as a function of d_1 and d_2 only. For a given value of d_1, there exist configurations of the confusion matrix that result in different ARD values. Figure 1 also shows all these possible ARD values for each given d_1 (marked with a tick over the whole possible range, which is indicated by a vertical line). This reveals a somehow erratic behaviour of the ARD in some cases, and inspecting such cases more closely allows us to clarify how the ARD works. For instance, for $n = 20$ consider the confusion matrices

$$\mathbf{N}_1 = \begin{pmatrix} 16 & 2 \\ 2 & 0 \end{pmatrix}, \quad \mathbf{N}_2 = \begin{pmatrix} 11 & 0 \\ 4 & 5 \end{pmatrix}.$$

The two matrices have $d_1 = 16$, so that MED = 0.2 and RD = .337 for both $\mathbf{N}_1$ and $\mathbf{N}_2$. However, ARD = 1.097 for $\mathbf{N}_1$, whereas ARD = 0.668 for $\mathbf{N}_2$. In the first configuration, in both clusterings there is a big cluster with 18 elements and a relatively small one with only two elements; both clusterings agree on most of the elements in the big cluster, but show no agreement at all on the small cluster, since none of the data points has been simultaneously assigned to the small cluster in both clusterings. In the second configuration, the first clustering presents two quite balanced clusters, say $\mathscr{C} = \{C_1, C_2\}$, of sizes 11 and 9 respectively, while the second clustering has clusters of sizes 15 and 5, which can be obtained from $\mathscr{C}$ by transferring four elements from C_2 to C_1. The ARD seems to penalize the first configuration much more severely than the second one. The effect of pronounced cluster size imbalance (as in $\mathbf{N}_1$) on similarity indices was recently studied in van der Hoef and Warrens (2019) and Warrens and van der Hoef (2019).

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.3.1. Worst-case scenario

The previous formulas for the MED and the RD in terms of d_1 and d_2 are also useful to analyse the worst-case scenario: the situation in which two given clusterings are as dissimilar as possible. If n is even, then the maximum possible MED is 1/2 and is attained for $d_1 = d_2 = n/2$. Thus, it is worth remarking that even for the two most dissimilar possible clusterings the MED is not going to be higher than 0.5 for the case of $r = s = 2$. This could make a case against the use of the MED, since one would expect this distance to attain a maximum of 1 when comparing the most dissimilar clusterings. However, a moment of reflection reveals that this maximum of 0.5 makes perfect sense in the context of clustering comparison, due to the aforementioned feature that any cluster label permutation should not affect distances between clusterings: having a proportion of label disagreements greater than 0.5 would mean that exchanging the labels would yield a proportion smaller than 0.5. Nevertheless, it is helpful to keep the value of the maximum possible distance in mind at the time of judging how far two clusterings are: a MED of 0.4 always has the same interpretation, but in relative terms it represents a worse result if the maximum possible MED is 0.5 than if it is 0.95. Hence, this suggests the introduction of a normalized MED, defined as NMED = MED/maxMED, to record how large the MED is with respect to its maximum possible value (given fixed values of r, s and n). This should not replace the unnormalized MED, since they offer different information, but they should be given together. In the previous example, having MED = 0.4, NMED = 0.8 versus MED = 0.4, NMED = 0.42 indicates that the former situation is closer to the case of totally dissimilar clusterings than the latter. Notice that this is a very different adjustment from the usual one, since it is not based on the expected value of the index under some null model; indeed, it does not rely on any choice of a null model.

The difficulty of such a normalization is that it is necessary to analyse which is the worst-case scenario for each index. Continuing with the 2×2 table, it is not hard to check that maxMED = (n-1)/(2n) if n is odd, which is attained for both d_1 = (n − 1)/2 and d_1 = (n + 1)/2. Therefore, $\mathrm{NMED} = 2n^{-1}\min\{d_1, d_2\}$ for even n and $\mathrm{NMED} = 2(n-1)^{-1}\min\{d_1, d_2\}$ for odd n. Regarding the RD, its maximum is attained at the same value of d_1 as for the MED, resulting in $\mathrm{maxRD} = n/\{2(n-1)\}$ for even n and $\mathrm{maxRD} = (n+1)/(2n)$ for odd n, so that it approaches 1/2 as n increases. Hence, the normalized RD, defined as $\mathrm{NRD} = \mathrm{RD}/\mathrm{maxRD}$, can be explicitly written as $\mathrm{NRD} = 4n^{-2}d_1d_2$ for even n and $\mathrm{NRD} = 4\{(n-1)(n+1)\}^{-1}d_1d_2$ for odd n. For the ARD, it would have been expected that its maximum were attained among the possible configurations with $d_1 = n/2$ for even n (or $d_1 = (n-1)/2$ for odd n), but Figure 1 shows that this does not happen in general. For instance, for $n = 20$ the maximum ARD is attained for a configuration with $d_1 = 12$; more precisely, for $\mathbf{N} = \begin{pmatrix} 12 & 4 \\ 4 & 0 \end{pmatrix}$, which gives $\mathrm{maxARD} = 95/84 \cong 1.131$. It is somewhat counterintuitive that the maximum value of the ARD is not attained for $\mathbf{N} = \begin{pmatrix} 5 & 5 \\ 5 & 5 \end{pmatrix}$, which represents the situation where the labels of the first clustering are perfectly independent of the labels in the second clustering.

In fact, it would be interesting to study what are the possible maximum and minimum values of the ARD for a given d_1. Since d_1 and $d_2 = n - d_1$ are fixed, the numerator in the definition of the ARD is constant, so this problem is equivalent to finding the minimum and maximum values of $E(\mathrm{RD})$ for a given d_1. After examining a large number of cases with

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$n \leq 20$, it appears (although it was not possible to find a simple proof) that for $n \geq 7$ and a given $d_1 \geq d_2$, the maximum ARD is attained for

$$\mathbf{N} = \begin{pmatrix} d_1 & d_2/2 \\ d_2/2 & 0 \end{pmatrix} \text{ or } \mathbf{N} = \begin{pmatrix} d_1 & (d_2-1)/2 \\ (d_2+1)/2 & 0 \end{pmatrix}, \tag{6}$$

provided $d_2 \geq 2$ is even or odd, respectively, and that for $d_1 = n-1, d_2 = 1$ the confusion matrix configuration that maximizes the ARD is $\begin{pmatrix} n-2 & 0 \\ 1 & 1 \end{pmatrix}$. Using the form for even d_2, the resulting maximum ARD for a given d_1 can be expressed as

$$\alpha_n(d_1) = \frac{4n(n-1)d_1}{(d_1+n)\{d_1^2 + n(n-2)\}}$$

for $d_1 \geq d_2 \geq 2$. Maximizing $\alpha_n(d_1)$ with respect to d_1 yields maxARD, but it is not clear how to obtain an explicit expression for such a maximum.

Moreover, as noted by an anonymous reviewer, since both marginals are fixed when computing the expected value of the RD, a further interesting open problem (perhaps more closely connected with the ARD definition) would be that of finding the maximum ARD value given specific marginals. This is indeed cast in Hubert and Arabie (1985, p. 199) as 'a very difficult problem of combinatorial optimization', but some preliminary numerical work seems to indicate that, given n, n_{1+} and n_{+1}, the 2×2 confusion matrix that attains the maximum ARD with such marginals is the one whose n_{11} entry is the closest to $(n_{1+} + n_{+1})/2 - n/4$ in its feasible region (this will be further investigated in a separate paper). Nevertheless, note that Brusco and Steinley (2008) and Steinley, Hendrickson and Brusco (2015) proposed a binary integer program and a heuristic algorithm, respectively, for maximizing the ARI, which could be reversed to numerically maximize the ARD.

3.3.2. Close clusterings

Similarly, this in-depth examination of the 2×2 case is also useful for understanding how these measures of dissimilarity between two clusterings evolve when such clusterings are very close. All these distances obviously return a zero value if $n_{12} = n_{21} = 0$, but the question that will be addressed here is how they behave as $n_{12} \to 0$ and $n_{21} \to 0$ before reaching their null limit. More precisely, the aim is to provide a linear approximation of the MED and the RD for small values of n_{12} and n_{21}.

Such an approximation is very easy to find for the MED, since as both $n_{12}, n_{21} \to 0$ it is clear that $\min\{d_1, d_2\} = n_{12} + n_{21}$, so that $\text{MED} = (n_{12} + n_{21})/n$ for small values of n_{12} and n_{21} (this is an equality rather than an approximation). On the other hand, it is possible to write $\text{RD} = \binom{n}{2}^{-1} \{n(n_{12} + n_{21}) - (n_{12} + n_{21})^2\}$, so that a Taylor expansion gives $\text{RD} \simeq 2(n_{12} + n_{21})/(n-1)$ as $n_{12}, n_{21} \to 0$. This means that, for small values of n_{12} and n_{21}, the RD will be roughly twice the MED.

For instance, for $\mathbf{N} = \begin{pmatrix} 55 & 6 \\ 4 & 35 \end{pmatrix}$ we have MED = .1 and RD = .182, while the approximation formula for the RD gives $2(6 + 4) = (100 - 1) = 0.202$.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.4. Arbitrary number of clusters

The case $r = s = 2$ is surely the easiest one to analyse in detail, and its analysis results in a deeper understanding of how the MED, RD and ARD behave. Here, such an analysis is extended for the comparison of two clusterings with an arbitrary number of clusters.

One of the findings for $r = s = 2$ is that the MED attains its maximum when the clustering labels are perfectly independent. In general, this refers to the situation where n is a multiple of rs and the $(r \times s)$ confusion matrix $\mathbf{N}$ has all its entries equal to n/(rs). In that case, note that $\sum_{i=1}^{r} n_{i,\sigma(i)} = n/s$ for any $\sigma \in \mathscr{P}_s$ and, therefore, $\mathrm{med} = 1 - 1/s$. But, assuming $r \leq s \leq n$, Charon *et al*. (2006, Lemma 1) showed that an upper bound for the MED is $1 - \lceil n/s \rceil / n$ (with $\lceil \cdot \rceil$ denoting the ceiling function), which generalizes the bounds obtained for odd or even n in the previous subsection for $s = 2$. Hence, once again the maximum MED is attained for the case of perfectly independent clustering labels. Thus, the corresponding normalized MED is defined in general as $\mathrm{NMED} = \frac{n}{n - \lceil n/q \rceil} \mathrm{MED}$, where $q = \max\{r, s\}$. The effect of this normalization is noticeable for small values of q, but becomes negligible as q increases.

In contrast, the story for the RD with arbitrary $r \leq s$ is different from the case $r = s = 2$. An exhaustive enumeration of all the possible confusion matrix configurations for small values of r, s and n (namely, $r + s \leq 8$ and $n \leq 40$) suggests that, given $r \leq s$ and $n \geq 2(r-1) + s$, the maximum value of the RD is always attained for a matrix of the form

$$\mathbf{N} = \begin{pmatrix} q_1 & q_2 & \cdots & q_s \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}, \tag{7}$$

with $q_1, \ldots, q_s \in \mathbb{N}$ and $q_1 \geq \ldots \geq q_s$ (the formal proof of this fact remains an open question). This does not mean that the maximizing matrix is necessarily unique; in fact, as noted before, for $r = s = 2$ and $n = 20$ the maximum RD is attained for any confusion matrix with $d_1 = 10$, for instance for $\mathbf{N} = \begin{pmatrix} 10 & 9 \\ 1 & 0 \end{pmatrix}$ which has the form (7), and also for $\mathbf{N} = \begin{pmatrix} 5 & 5 \\ 5 & 5 \end{pmatrix}$, which represents the perfectly independent situation. In addition, assuming that the conjectured form of the maximizer is correct, it is shown in Appendix B that maxRD is attained by taking $q_1 = k + r - 1$, the next $\ell \geq 0$ coordinates $q_2 = \cdots = q_{\ell+1} = k + 1$ and the remaining $s - \ell - 1 \geq 0$ coordinates $q_{\ell+2} = \cdots = q_s = k$, where $k = \lfloor \{n - 2(r-1)\}/s \rfloor \in \mathbb{N}$ (with $\lfloor \cdot \rfloor$ denoting the floor function) and $\ell = n - 2(r-1) - ks \in \{0, 1, \ldots, s-1\}$; that is, k and ℓ are the quotient and the remainder of the (Euclidean) division of $n - 2(r-1)$ by s, respectively. With such a choice, it follows that

$$n(n-1)\,\mathrm{maxRD} = (n - r + 1)^2 + (r-1)(2r-3) - sk^2 - l(2k+1). \tag{8}$$

This allows us to explicitly define the normalization $\mathrm{NRD} = \mathrm{RD}/\mathrm{maxRD}$.

Further, when n is a multiple of rs and the two clusterings are perfectly independent it is easy to check that $n(n-1)\mathrm{RD} = n^2(r + s - 2)/(rs)$. Since $\mathrm{maxRD} \tilde{} 1 - 1/s$ as $n \to \infty$, it follows that the maximum RD is not attained for perfectly independent clusterings for big enough n if $s > 2$. Moreover, in practice this seems to be the case for all n, as shown in Figure 2. This figure represents the normalized RD for the case of two perfectly

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

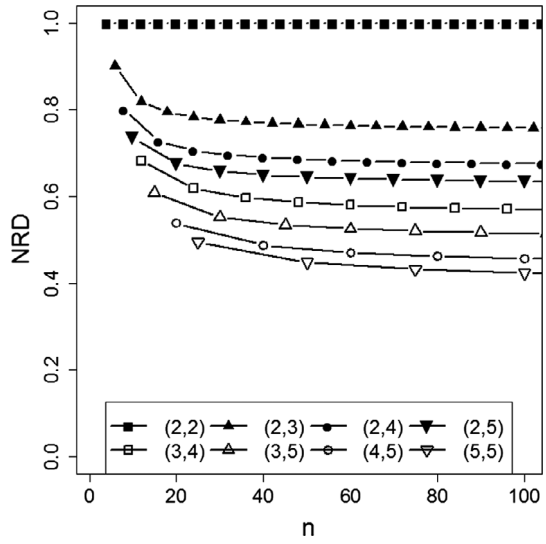

Figure 2. Normalized RD for perfectly independent clusterings with r and s clusters, as a function of the sample size n, for several combinations of (r, s) as indicated in the legend.

independent clusterings for several combinations of r and s. Only for the pair $(r,s) = (2,2)$ does the RD for perfectly independent clusterings match its maximum possible value. For any other combination, having two totally unrelated clusterings does not yield the maximum possible RD; indeed, this phenomenon becomes more and more severe as $r+s$ increases and, for instance, for $(r,s) = (5,5)$ and $n = 100$ the confusion matrix with all its entries equal to 4 results in a RD that is only 42% of the maximum achievable RD, attained for a matrix of the form (7) with $q_1 = 22$, $q_2 = q_3 = 19$ and $q_4 = q_5 = 18$ Finally, it is worth noting that $(r+s-2)/(rs)$ decreases as r and/or s increases, so that the RD for the case of perfectly independent clusterings becomes quite small when both r and s are large and, hence, the RD does not seem useful to detect this important instance of unrelated clusterings. Fowlkes and Mallows (1983, p. 555) already noted this phenomenon, upon examining the expected value and variance of the Rand index under the null model.

For the ARD, it was not possible to provide an explicit formula for its maximum for a given n and $r = s = 2$, and the problem is of course more intricate for arbitrary r and s. Nevertheless, it seems clear that the maximum ARD is not attained for the case of independent clustering labels, in general. Instead, the inspection of all possible confusion matrix configurations for small values of r, s and sufficiently large n (namely, $r+s \leq 8$ and $20 \leq n \leq 40$) seems to suggest that the maximum value of the ARD is always attained for a matrix of the form

$$\mathbf{N} = \begin{pmatrix} p_1 & q_2 & \cdots & q_s \\ p_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ p_r & 0 & \cdots & 0 \end{pmatrix}, \tag{9}$$

with $p_1, \ldots, p_r, q_2, \ldots, q_s \in \mathbb{N}$ and $p_1 \geq p_2 \geq \cdots \geq p_r$, $p_1 \geq q_2 \geq \cdots \geq q_s$ (furthermore, with $(p_2, \ldots, p_r) = (q_2, \ldots, q_s)$ if $r = s$). Indeed, confusion matrices with ARD greater than the

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

value corresponding to the perfectly independent case can be constructed by following the guidelines described above for $r = s = 2$. For instance, if $n = 24$, $r = 2$, $s = 3$, then the confusion matrices

$$\mathbf{N}_1 = \begin{pmatrix} 4 & 4 & 4 \\ 4 & 4 & 4 \end{pmatrix} \text{ and } \mathbf{N}_2 = \begin{pmatrix} 15 & 4 & 1 \\ 4 & 0 & 0 \end{pmatrix}$$

lead to ARDs of 1.062 and 1.143, respectively, and for $n = 27$, $r = 3$, $s = 3$, the confusion matrices

$$\mathbf{N}_3 = \begin{pmatrix} 3 & 3 & 3 \\ 3 & 3 & 3 \\ 3 & 3 & 3 \end{pmatrix} \text{ and } \mathbf{N}_4 = \begin{pmatrix} 15 & 5 & 1 \\ 5 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

yield ARDs of 1.083 and 1.157, respectively. Moreover, when n is a multiple of rs and the two clusterings are perfectly independent, it is easy to show that $\text{ARD} = (n-1)/\{n-(2rs-r-s)/(r+s-2)\}$, which approaches 1 (from above) as n increases.

Even if for a given, large enough n, the maximum ARD seems to be attained for a confusion matrix like (9), an anonymous reviewer noted that such might not be the case for small sample sizes. For instance, for $n = 9$ and $r = s = 3$ the confusion matrix with all entries equal to unity has ARD = 1.333 and the maximum ARD among the class of matrices (9) is ARD = 1.257, attained for $p_1 = 5$ and $p_2 = p_3 = q_2 = q_3 = 1$. In fact, for $r \leq s$ and $n = r + s - 1$, the matrix of the form (9) with $p_1 = \ldots = p_r = q_2 = \ldots = q_s = 1$ has

$$\text{ARD} = \frac{(r+s-1)(r+s-2)\{r(r-1)+s(s-1)\}}{r^4+2r^3(s-2)-r^2(4s-5)+2r(s-1)(s^2-s+1)+s(s-2)(s-1)^2},$$

which represents the maximum achievable ARD for given $r \leq s$ and arbitrary n, as shown in Chacón and Rastrojo (2020).

4. Numerical experiments

In this section the distributions of the MED, RD, ARD and the normalized versions NMED and NRD will be compared in different simulated scenarios.

As noted in van Mechelen *et al*. (2018), benchmarking studies for cluster analysis do not abound. Nevertheless, the task of comparing different external criteria via simulation was addressed in the seminal paper by Milligan and Cooper (1986) and also more recently in Steinley (2004), Denœud and Guénoche (2006) and Steinley and Brusco (2018).

Broadly speaking, these studies handle two possible scenarios. The first explores the performance of the criteria in the null case, that is, when the agreement between the compared clusterings is only due to chance. The second is concerned with how the criteria of interest behave as the two compared clusterings drift apart, starting from perfect similarity. Both scenarios are considered separately in the following subsections.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.1. The null case

As noted above, the null case scenario covers the situation where the clustering agreements are solely due to chance. However, as remarked in Gates and Ahn (2017), different choices for the model for random clusterings can be made, and a careful model selection is needed to provide a baseline that is neither based on a model that is 'not random enough' nor on a model that is 'too random'.

Gates and Ahn (2017) considered three models for random clusterings, with increasing level of randomness, starting with the permutation model (where the number of clusters and their sizes are fixed), followed by the model where only the number of clusters is fixed, and finally the model encompassing all possible clusterings, with arbitrary number of clusters and cluster sizes. As a compromise for intermediate randomness level, in this section the null case refers to the situation where random labels are drawn uniformly after fixing the number of clusters.

Hence, the distribution of the criteria considered in the null case is explored by computing their values on a large enough number B of random clustering pairs of n objects, obtained by independently drawing two uniform samples of size n from $\{1,\ldots,r\}$ and $\{1,\ldots,s\}$, respectively. The number of synthetic replicates was set to $B = 10{,}000$, in order to obtain a precise approximation of the distributions; the number of clusters was considered equal ($r = s$), in common with some of the aforementioned previous studies,

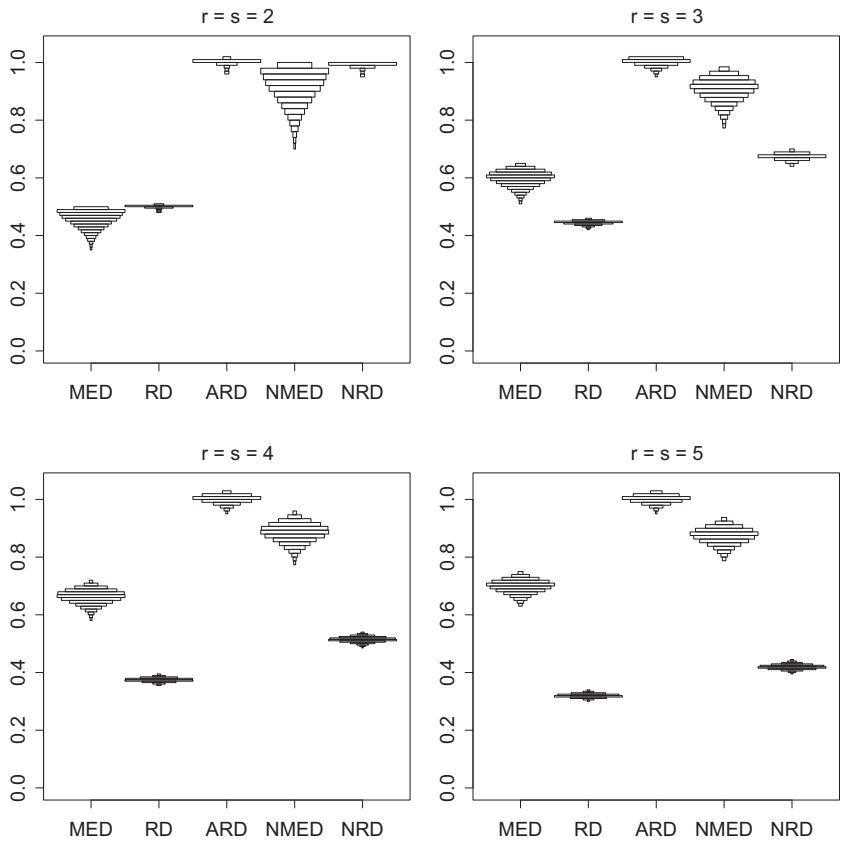

Figure 3. Distribution of the criteria in the null case for sample size $n = 100$.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

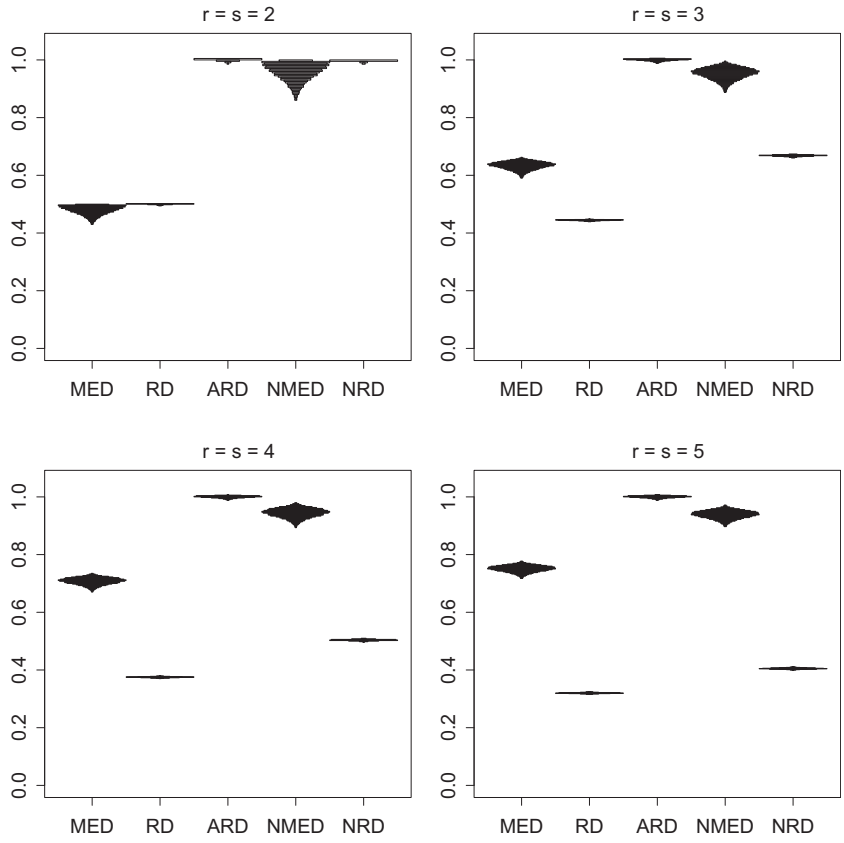

Figure 4. Distribution of the criteria in the null case for sample size $n = 400$.

and ranging in $\{2, 3, 4, 5\}$ and samples sizes $n = 100$ and $n = 400$ were used to investigate the effect of an increasing number of data points. The distributions of the criteria studied are depicted in Figures 3 and 4 for $n = 100$ and $n = 400$, respectively, by means of side-by-side vertical histograms (with the bars mirrored with respect to the vertical axis), whose bars have been rescaled so that each of them has maximum bar length equal to 1, to aid visualization.

Despite being corrected for chance according to the permutation model (which is not exactly the null model in this study), the distribution of the ARD seems to be centred at 1 in all cases, so this type of adjustment makes it possible to compare its behaviour for the different configurations. Its variability is the second lowest among the criteria compared, and it seems not to change with the number of clusters but quickly decreases with the sample size, as also noted in Steinley, Brusco and Hubert (2016). This suggests that the ARD may give rise to a powerful tool for detecting clustering independence.

The RD is the least variable criterion of those considered here. This is not surprising in view of Figure 1, since its quadratic nature entails a least pronounced descent around its null-case value than the MED, for instance (see also Warrens & van der Hoef, 2020). Besides, as remarked in the previous section, under this null scenario the RD only achieves its maximum value for the case $r = s = 2$, which yields a tightly concentrated distribution of the NRD with a maximum of 1 in that case. However, as shown in Figure 2, the

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

maximum possible value of the RD becomes rather larger than its value for independent clusterings as the number of clusters increases, and this explains why even the distributions of the normalized RD are far from 1. In other words, confusion matrices corresponding to randomly generated clusterings are usually far from something like (7). It might be possible to obtain NRD distributions much closer to 1 if the random clusters were generated to produce confusion matrices only slightly deviating from (7), but that does not seem to be an appropriate null model.

The MED is notably more variable than the ARD and RD, with standard deviations about 1.6–2.1 times greater than those of the ARD, and 3.5–4.2 times greater than those of the RD, for $n = 100$ (3.1–4.2 and 6.7–8.5, respectively, for $n = 400$). Its variability, though, appears to decrease slightly as the number of clusters grows. Its approximated distribution shows an upper bound that agrees with the results in the previous section (e.g., for $n = 100$ a maximum value of 0.5, 0.66, 0.75 and 0.8 for $r = s = 2, 3, 4, 5$, respectively), yielding location features that naturally change with the number of clusters, and hence making it inappropriate to aggregate its results across the different simulation configurations. This upper bound also implies that NMED certainly attains a maximum value of 1 for this null scenario of random clusterings. However, it must be pointed out that the probability of attaining such a maximum value seems to decrease with the number of clusters.

Indeed, in some cases it is possible even to give an exact expression for such a probability. For $r = s = 2$ and even n, for instance, it corresponds to $P(d_1 = n/2)$, where d_1 is the sum of the two diagonal terms in the confusion matrix. In the null scenario, d_1 is a random variable following a binomial distribution, with n as the number of trials and probability of success $p = 1/2$ (the probability that two uniform and independent choices from {1,2} are the same). Hence, $P(d_1 = n/2) = \binom{n}{n/2}/2^n$. More generally, here the random variable $n \cdot \mathrm{MED}$ follows a folded binomial distribution (Gart, 1970).

4.2. Diverging clusterings

The second simulation scenario concerns studying the evolution of the criteria compared as two clusterings move away from each other, starting from a situation of perfect agreement.

Interestingly, in most of the existing simulation studies (see, for instance, Denœud & Guénoche, 2006; Steinley, 2004), the process of 'moving away from each other' is quantified by measuring the proportion of data points that are differently clustered from the initial stage of perfect agreement. Steinley (2004) called this proportion the 'degree of overlap' (DO), and more recently Steinley and Brusco (2018, Section 3.2.2) referred to this measure of deviation from the perfect agreement as the misclassification rate. In this section, we investigate how the other clustering distances evolve as compared to the MED (see Figure 3 in Steinley, 2004, or Figure 1 in Denœud & Guénoche, 2006).

As an aside, it should be noted that what Steinley (2004) and Steinley and Brusco (2018) called the DO is not exactly the same as the MED. The simulation set-up in those papers concerns a diagonal confusion matrix as the starting point (hence, a situation of perfect agreement) that is progressively perturbed by randomly taking a proportion of objects from the diagonal and placing them in off-diagonal cells. This proportion of off-diagonal objects is what is called the DO, and in the aforementioned studies it is allowed to vary in $.05, .10, \ldots, .95$. However, this is not the same as the misclassification rate: while the

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

DO and the MED usually coincide for low DO values, when the DO is too high the resulting clusterings may become closer with respect to the MED, instead of further away. For example, consider the confusion matrices.

$$\mathbf{N}_1 = \begin{pmatrix} 8 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix}, \quad \mathbf{N}_2 = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}, \quad \mathbf{N}_3 = \begin{pmatrix} 1 & 2 & 5 \\ 3 & 1 & 2 \\ 2 & 4 & 0 \end{pmatrix}.$$

For all of them, $n = 20$. Matrix $\mathbf{N}_2$ stems from $\mathbf{N}_1$ after removing a total of 13 objects from the diagonal, so that DO = 13/20=.65 for $\mathbf{N}_2$; it can be checked that MED = 0.65 for $\mathbf{N}_2$ as well. Five additional objects are removed from the diagonal when going from $\mathbf{N}_2$ to $\mathbf{N}_3$, representing a total DO = 18/20=.9 with respect to $\mathbf{N}_1$, but MED = 0.4 for $\mathbf{N}_3$, lower than for $\mathbf{N}_2$. Of course, this is due to the fact that maxMED = 0.65 for $n = 20$ and $r = s = 3$, so that it does not seem appropriate to consider DO values greater than .65 in this case.

In Denœud and Guénoche (2006) several agreement indices were compared as a function of an increasing MED. A given starting partition is recursively perturbed by randomly selecting one element and a new class label for it. This procedure aims to randomly and equiprobably generate partitions at a precise MED of the given one. However, there the number of clusters is not fixed and, hence, their study has a higher degree of uncertainty.

Nevertheless, for the aim of examining the evolution of the different distances as a function of the MED, the most exhaustive procedure is surely that based on computing the measures involved for all the possible confusion matrix configurations, that is, for all the matrices in $\mathcal{N}(r,s,n) = \{\mathbf{N} \in \mathcal{M}_{r\times s} : n_{i+} > 0 \text{ for all } i,\ n_{+j} > 0 \text{ for all } j, \text{ and } \sum_{i,j} n_{ij} = n\}$. Indeed, this is precisely what Figure 1 represents for $r = s = 2$ and $n = 20$. But this could be accomplished in that case because the cardinality $|\mathcal{N}(2,2,20)| = 1691$ was reasonably small.

For $r = s = 3$ and $n = 20$ the class of all possible confusion matrices is considerably larger, namely $|\mathcal{N}(3,3,20)| = 2,806,281$, but still not prohibitive, so its exhaustive enumeration is still feasible. Therefore, the ARD, MED and RD were obtained for each of these possible confusion matrix configurations, yielding a large amount of interesting information. Figure 5 shows boxplots for the conditional distributions of the RD (left) and the ARD (right), given the MED, along with the (mean) regression curve. These plots contain the same information as Figure 1, but a first notable difference is that now the RD corresponding to a given MED is no longer a single value, as it was for $r = s = 2$; instead, for $r = s = 3$ all the possible confusion matrices with the same MED result in a wide range of different RD values.

Most of the conditional distributions given the MED are fairly symmetric, but it is worth remarking on some interesting features that arise, especially for very low or very high MED values. For instance, Figure 6 focuses on the distribution of the RD given the particular values of MED = 0.1 (left) and MED = 0.6 (right), which clearly show a high degree of skewness. The conditional distribution of the ARD also shows some peculiarities: its maximum value (ARD = 1.205) is attained for a confusion matrix with MED = 0.5, but its conditional mean attains its maximum at MED = 0.65 (the maximum possible MED value). The outlier in the conditional distribution of the ARD given MED = 0.2 is particularly striking, with a value of 1.108, attained at the confusion matrix

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

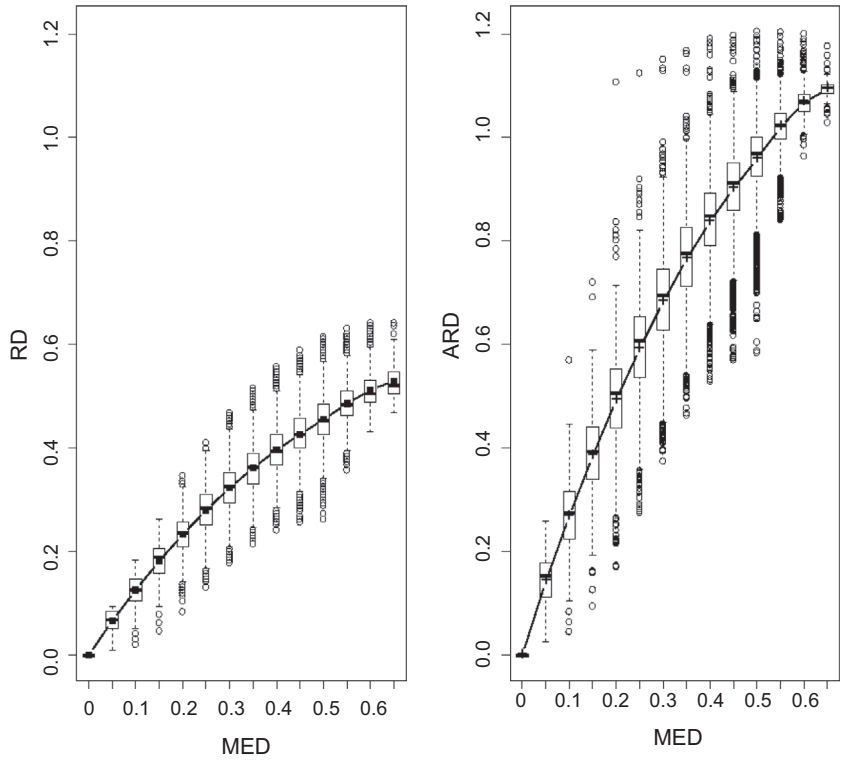

Figure 5. RD (left) and ARD (right) versus MED for $r = s = 3$ and $n = 20$. The boxplots represent the conditional distributions for a given value of MED, and the curves depict the conditional means.

$$\mathbf{N} = \begin{pmatrix} 16 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

despite being only four data point transfers from the perfect agreement situation.

This extensive enumeration study is also useful for examining the individual distributions of each criterion. For example, Figure 7 shows the distribution of all the MED values for $r = s = 3$ and $n = 20$, revealing a very different scenario from that suggested in Steinley (2004, Figure 1). In Steinley's simulation study, the distribution of the MED appeared to be somewhat uniform, which strongly contrasts with the distribution shape shown in Figure 7 from exhaustive enumeration. The reason is that in Steinley's paper the distribution of the MED is investigated by aggregation of all the multiple simulation conditions. And, as noted before, these simulation conditions involved uniformly varying the DO from .05 to .95. It was already noted above that the DO is not exactly the same as the MED, but they are closely related, so forcing a fixed given number of simulations for every DO level naturally results in a (nearly) uniform distribution for the MED. However, the exhaustive examination of all the possible confusion matrix configurations in Figure 7 shows that the MED distribution is quite different from the uniform one.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

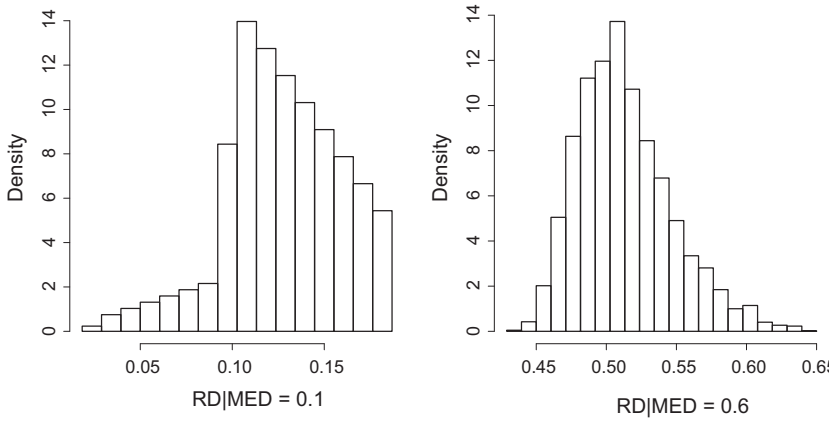

Figure 6. Conditional distribution of the RD given MED = 0:1 (left) and MED = 0:6 (right) for $r = s = 3$ and $n = 20$

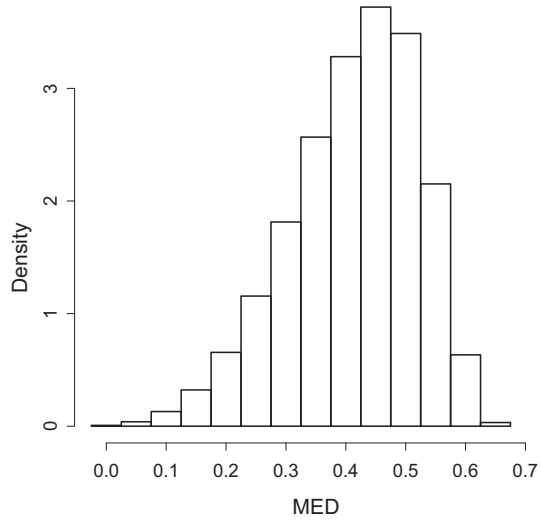

Figure 7. Distribution of the possible MED values for $r = s = 3$ and $n = 20$.

For higher values of r, s or n, it is not possible to enumerate all the confusion matrices in $\mathscr{N}(r,s,n)$, since its cardinality becomes exorbitant. An alternative way to approximate the criteria distributions, for these higher values of r, s and n, would be to randomly sample a large number of matrices from $\mathscr{N}(r,s,n)$ (in an equiprobable way), and then compute their MEDs, RDs and ARDs in order to obtain an equivalent approximation of Figure 5. This suggestion is also not without problems, for two reasons: first, it is not straightforward to uniformly sample from $\mathscr{N}(r,s,n)$ (see Appendix C for a valid procedure); and second, the fact that some of the possible MED values occur only for a small number of confusion matrices makes it difficult to procure an accurate approximation of the conditional distributions given such MED values. For example, from the exhaustive enumeration of $\mathscr{N}(3,3,20)$ it follows that the probability of obtaining a confusion matrix with MED = 0.65 by uniform sampling is approximately 1.6×10^{-3}, so a

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

large simulation size would be required in order to approximate the distribution of the other distances given MED = 0.65.

In any case, following the procedure suggested in Appendix C, a random sample of size 10^7 was drawn from $\mathcal{N}(5,5,80)$, and the values of the MED, RD and ARD for these confusion matrices were recorded. It must be remarked that the cardinality of $\mathcal{N}(5,5,80)$ is approximately 2.309×10^{23}, which makes the exhaustive enumeration approach impossible. From that sample, it is possible to approximate the conditional means of the RD and ARD given the MED (Figure 8, left) and to provide an approximate analogue of Figure 7 for $n = 80$ and $r = s = 5$ (Figure 8, right). Notice also that, even if the possible MED values are $\{i/80 : i = 0, 1, \ldots, 64\}$ (because maxMED = 0.8 for $r = s = 5$ and $n = 80$), the range of MED values for which 10 or more observations were obtained in this particular sample reduced to $\{i/80 : i = 15, 16, \ldots, 60\}$ (i.e., the others had sample frequencies smaller than 10^{-6}), and that is why Figure 8 has some missing parts.

5. Discussion

When comparing two partitions of a finite data set, surely the confusion matrix is the object that yields the most complete information. However, when the number of cells is large, the confusion matrix provides too much detail and it becomes necessary to resort to summary statistics to extract useful information. The MED, the RD and the ARD are examples of such summary statistics, each of them offering a different synopsis.

The first two are empirical versions of distances between whole-space clusterings and, intuitively, correspond to computing the proportion of 'differently placed' individual data points (MED) or data pairs (RD) in the clusterings compared. Considering data pairs instead of individual data points appears rather less intuitive. In fact, as pointed out in

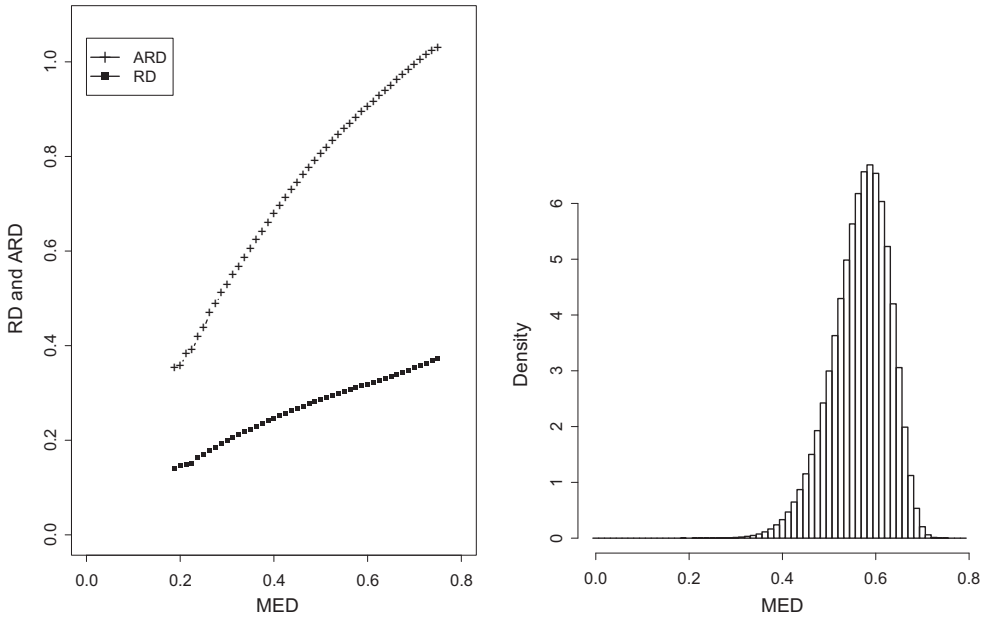

Figure 8. Approximated conditional means of the RD and the ARD given the MED (left) and distribution of possible MED values (right), for $r = s = 5$ and $n = 80$.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hubert and Arabie (1985, Section 4), one could equally consider using data triplets or, more generally, data k-tuples. Comparisons, however, become more and more intricate as k increases. It is in this sense that the choice of $k = 1$ (i.e., the MED) represents the simplest option.

But it is not just a matter of simplicity. As long as $\min\{r, s\} > 2$, the RD also has a more serious drawback: the case of completely unrelated clusterings does not correspond to the most dissimilar clustering pair, according to the RD, and this phenomenon becomes more and more severe as the clustering sizes increase, as shown in Figure 2 (see also Warrens & van der Hoef, 2020). This unfortunate feature is not shared by the MED, which does point out unrelated clusterings as an instance of extreme dissimilarity, and furthermore has a maximum value that quickly approaches 1 as $\max\{r, s\}$ increases.

A possible correction for the aforementioned flaw is to consider the relative size of the RD with respect to the average RD value when the two clusterings are generated at random; this is what the ARD provides. It exhibits the natural advantage of creating a criterion that is always centred at 1 for the null case, but on the other hand introduces a distorting element that further complicates the interpretation: now the ARD represents the relative size of the proportion of differently treated data pairs in the compared clusterings with respect to a baseline, taken as the average value of that proportion when the cluster labels are assigned at random while maintaining the number of clusters and cluster sizes fixed. This also implies that, if the baseline changes (as usually happens when examining two different confusion matrices), then the relative comparison of the two scenarios by means of ARD scores becomes unclear.

An alternative remedy, also aimed at examining the relative size of a criterion, but this time against the worst possible case, is to normalize such a criterion with respect to its maximum value. This is a different kind of adjustment, which does not produce a criterion that is centred at 1 for a null model (in fact, it does not rely on a specific null model), but it ensures that all the resulting values lie in $[0, 1]$ instead. When applied to the MED and the RD it results in the new NMED and NRD criteria, which are not advised to be used alone, but jointly with their unnormalized counterparts, since the latter retain the most straightforward interpretation. In addition, not achieving its maximum for unrelated clusterings also hinders this approach for the RD, as shown in Figures 3 and 4, since it entails that the distribution of the NRD can be far from 1 under the null model. In contrast, the NMED distribution is indeed close to its upper bound of 1 in the null case, more so for higher sample sizes, although it must be pointed out that it seems more and more unlikely to reach this upper bound as the number of clusters increase.

In any case, it seems clear that the distributions of all these criteria (the MED, RD and ARD) deserve further study, since their investigation through exhaustive enumeration or uniform sampling from the set of all possible confusion matrices has revealed some previous misconceptions and unexpected features.

Acknowledgements

The author is grateful to the Associate Editor, Professor Michael Brusco, and to two anonymous reviewers for their constructive remarks. The author acknowledges the support of the Spanish Ministerio de Economa y Competitividad grant MTM2016-78751-P and the Junta de Extremadura grant GR18016.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

References

Azizyan, M., Chen, Y.-C., Singh, A., & Wasserman, L. (2015). Risk bounds for mode clustering. Preprint, arXiv:1505.00482.

Aghaeepour, N., Finak, G., The FlowCAP Consortium, The DREAM Consortium, Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., & Scheuermann, R. H. (2013). Critical assessment of automated flow cytometry analysis techniques. *Nature Methods*, *10*, 228–38.

Albatineh, A. N., & Niewiadomska-Bugaj, M. (2011). MCS: A method for finding the number of clusters. *Journal of Classification*, *28*, 184–209.

Albatineh, A. N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, *23*, 301–13.

Anderson, E. (1935). The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, *59*, 2–5.

Ben-David, S., von Luxburg, U., & Pál, D. (2006). A sober look at clustering stability. In G. Lugosi & H.-U. Simon (Eds.), *Proceedings of the 19th Annual Conference on Learning Theory (COLT)* (pp. 5–19). Berlin: Springer.

Brusco, M. J., & Steinley, D. (2008). A binary integer program to maximize the agreement between partitions. *Journal of Classification*, *25*, 185–93.

Burkard, R., Dell'Amico, M., & Martello, S. (2009). *Assignment problems*. Philadelphia: SIAM.

Chacón, J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science*, *30*, 518–32.

Chacón, J. E. (2019). Mixture model modal clustering. *Advances in Data Analysis and Classification*, *13*, 379–404.

Chacón, J. E., & Rastrojo, A. I. (2020). Minimum adjusted Rand index for two clusterings of a given size. Preprint, arXiv:2002.03677.

Charon, I., Denœud, L., Guénoche, A., & Hudry, O. (2006). Maximum transfer distance between partitions. *Journal of Classification*, *23*, 103–21.

Charon, I., Denœud, L., & Hudry, O. (2007). Maximum de la distance de transfert à une partition donnée. *Mathématiques et Sciences Humaines*, *179*, 45–83.

Day, W. H. E. (1981). The complexity of computing metric distances between partitions. *Mathematical Social Sciences*, *1*, 269–87.

Denœud, L. (2008). Transfer distance between partitions. *Advances in Data Analysis and Classification*, *2*, 279–94.

Denœud, L., & Guénoche, A. (2006). Comparison of distance indices between partitions. In V. Batagelj, H.-H. Bock, A. Ferligoj & A. Žiberna (Eds.), *Data Science and Classification* (pp. 21–8). Berlin: Springer-Verlag.

Filkov, V., & Skiena, S. (2004). Integrating microarray data by consensus clustering. *International Journal on Artificial Intelligence Tools*, *13*, 863–80.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*, 553–69.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611–31.

Gart, J. J. (1970). A locally most powerful test for the symmetric folded binomial distribution. *Biometrics*, *26*, 129–38.

Gates, A. J., & Ahn, Y.-Y. (2017). The impact of random models on clustering similarity. *Journal of Machine Learning Research*, *18*, 3049–76.

Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classifications*. New York: Springer-Verlag.

Gusfield, D. (2002). Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, *82*, 159–64.

Györfi, L., Kohler, M., Krzyzak, M., & Walk, M. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer-Verlag.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hennig, C. (2019). Cluster validation by measurement of clustering characteristics relevant to the user. In C. H. Skiadas & J. R. Bozeman (Eds.), *Data analysis and applications 1: Clustering and regression, modeling-estimating, forecasting and data mining* (pp. 1–24). London: ISTE.

Hornik, K. (2005). A CLUE for CLUster Ensembles. *Journal of Statistical Software*, *14*(12).

Hornik, K. (2018). *clue: Cluster ensembles*. R package version 0.3-55.

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

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

Klemelä, J. (2009). *Smoothing of multivariate data: Density estimation and visualization*. Hoboken, NJ: John Wiley & Sons.

Lee, A. J. (1990). *U-statistics: Theory and practice*. New York: Marcel Dekker.

Meilă, M. (2005). Comparing clusterings – an axiomatic view. In S. Wrobel & L. De Raedt (Eds.), *Proceedings of the International Machine Learning Conference (ICML)*. New York: ACM Press.

Meilă, M. (2007). Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, *98*, 873–95.

Meilă, M. (2012). Local equivalences of distances between clusterings – a geometric perspective. *Machine Learning*, *86*, 369–89.

Meilă, M. (2016). Criteria for comparing clusterings. In C. Hennig, M. Meilă, F. Murtagh & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 619–35). Boca Raton, FL: CRC Press.

Messatfa, H. (1992). An algorithm to maximize the agreement between partitions. *Journal of Classification*, *9*, 5–15.

Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. J. Hubert & G. De Soete (Eds.), *Clustering and classification* (pp. 341–75). River Edge, NJ: World Scientific.

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, *21*, 441–58.

Mirkin, B. G. (1996). *Mathematical classification and clustering*. Dordrecht: Kluwer Academic.

Mirkin, B. G., & Cherny, L. B. (1970). On a distance measure between partitions of a finite set. *Automation and Remote Control*, *31*, 786–92.

Nijenhuis, A., & Wilf, H. S. (1978). *Combinatorial algorithms. For computers and calculators* (2nd ed.). New York: Academic Press.

Pfitzner, D., Leibbrandt, R., & Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, *19*, 361–94.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Auatria: R Foundation for Statistical Computing.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–50.

Régnier, S. (1965). Quelques aspects mathématiques des problèmes de classification automatique. *I.C.C. Bulletin*, *4*, 175–91. Reprinted (1983) in Mathématiques et Sciences Humaines, 82, 31–44.

Rossi, G. (2015). Hamming distance between partitions, clustering comparison and information. In *Proceedings of the International Conference on Pure Mathematics, Applied Mathematics and Computational Methods (PMAMCM2015)* (pp. 101–7).

Steinley, D. (2003). Local optima in -means clustering: What you don't know may hurt you. *Psychological Methods*, *8*, 294–304.

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods*, *9*, 386–96.

Steinley, D., & Brusco, M. J. (2018). A note on the expected value of the Rand index. *British Journal of Mathematical and Statistical Psychology*, *71*, 287–99.

Steinley, D., Brusco, M. J., & Hubert, L. (2016). The variance of the adjusted Rand index. *Psychological Methods*, *21*, 261–72.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Steinley, D., Hendrickson, G., & Brusco, M. J. (2015). A note on maximizing the agreement between partitions: A stepwise optimal algorithm and some properties. *Journal of Classification*, *32*, 114–26.

van der Hoef, H., & Warrens, M. J. (2019). Understanding information theoretic measures for comparing clusterings. *Behaviormetrika*, *46*, 353–70.

van Mechelen, I., Boulesteix, A.-L., Dangl, R., Dean, N., Guyon, I., Hennig, C., Leisch, F., & Steinley, D. (2018). Benchmarking in cluster analysis: A white paper. Preprint, arXiv:1809.10496v2.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 2837–54.

von Luxburg, U. (2010). Clustering stability: An overview. *Foundations and Trends in Machine Learning*, *2*, 235–74.

von Luxburg, U., & Ben-David, S. (2005). Towards a statistical theory for clustering. In *PASCAL Workshop on Statistics and Optimization of Clustering*.

Wallace, D. L. (1983). A method for comparing two hierarchical clusterings: Comment. *Journal of the American Statistical Association*, *78*, 569–76.

Warrens, M. J. (2008a). On similarity coefficients for tables and correction for chance. *Psychometrika*, *73*, 487–502.

Warrens, M. J. (2008b). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification*, *25*, 177–83.

Warrens, M. J., & van der Hoef, H. (2019). Understanding partition comparison indices based on counting object pairs. Preprint, arXiv:1901.01777.

Warrens, M. J., & van der Hoef, H. (2020). Understanding the rand index. In T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto & M. Vichi (Eds.), *Advanced researches in classification and data science*. Tokyo: Springer.

Wasserman, L. (2018). Topological data analysis. *Annual Review of Statistics and Its Application*, *5*, 501–32.

Appendix A:

R function for misclassification error distance computation

Recall from Section 3.1 that, given a confusion matrix $\mathbf{N} = (n_{ij}) \in \mathcal{N}(r,s,n)$ with $r \leq s$, the main computational problem is to find

$$\min_{\sigma \in \mathscr{P}_s} \sum_{i=1}^{s} m_{i,\sigma(i)}, \tag{10}$$

where $\mathscr{P}_s$ denotes the set of all possible permutations of s elements. Here, $n_{i+} = n_{ij} = 0$ for all $i = r+1,\ldots,s$ (if any) and $m_{ij} = n_{i+} + n_{+j} - 2n_{ij}$ for $i,j = 1,\ldots,s$. Fortunately, (10) is a linear sum assignment problem, whose solution can be efficiently found through the function solve_LSAP included in the R library clue (Hornik, 2005, 2018). So once that library is loaded, with the command library(clue), a function to compute the MED from two equal-size vectors containing the cluster labels with respect to each clustering can be obtained through the following simple code:

```
med <- function(labels1, labels2){
   n <- length(labels1)
   N <- table(labels1, labels2)
   r <- nrow(N)
   s <- ncol(N)
   if (r>s) N <- t(N); r <- nrow(N); s <- ncol(N)
   if (r<s) N <- rbind(N, matrix(0, nrow = s-r, ncol = s))
   M <- matrix(rowSums(N), nrow = s, ncol = s) +
```

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

```
    matrix(colSums(N), nrow = s, ncol = s, byrow = TRUE) - 2 * N
  optimal.permutation <- solve_LSAP(M)
  result <- sum(M[cbind(seq_along(optimal.permutation),
    optimal.permutation)]) / (2 * n)
  return(result)
}
```

Appendix B:

The maximum Rand distance

Assuming as true the conjecture that there is always a maximizer of the RD of the form (7), here it is shown that the maximum value of the RD satisfies (8). First notice that, for a confusion matrix of the form (7), the function $M(q_1,\ldots,q_s)=n(n-1)\mathrm{RD}=n^2 d_H$ can be written explicitly as

$$M(q_1,\ldots,q_s)=2(r-1)q_1-\sum_{j=1}^{s}q_j^2+c, \tag{11}$$

where $c=(n-r+1)^2+(r-1)(r-2)$. Then the aim is to maximize $M(q_1,\ldots,q_s)$ with the constraint that the total number of data points is n, which for the matrix (7) yields $\sum q_j+r-1=n$. The method of Lagrange multipliers yields the maximizer over real-valued choices of $q_1,\ldots,q_s$ as $q_1^*=\{n-2(r-1)\}/s+r-1$, $q_2^*=\cdots=q_s^*=\{n-2(r-1)\}/s$, but recall that the aim is to find the maximizer for non-negative integer values of $q_1,\ldots,q_s$.

As in Section 3.4, write $n-2(r-1)=ks+\ell$, with $k\in\mathbb{N}$ and $\ell\in\{0,1,\ldots,s-1\}$. If $\{n-2(r-1)\}/s\in\mathbb{N}$ (corresponding to the case $\ell=0$), then the real-valued maximizer is also integer-valued and leads to the maximizer and maximum value given in (7) and (8) for $\ell=0$.

If $\ell>0$ then the real-valued maximizer is rational, with all the coordinates $q_1^*,\ldots,q_s^*$ having the same fractional part ℓ/s. Due to the total sum constraint, to find the integer-valued maximizer of (11) these fractional reminders need to be redistributed into the coordinates $q_1,\ldots,q_s$, to make them integer, while at the same time trying to decrease the value of $M(q_1,\ldots,q_s)$ as less as possible with respect to the real-valued maximizer. To achieve this, first note that (11) is a concave function with the same curvature in every direction, so the least decrease with integer coordinates with respect to the real-valued maximum corresponds to rounding up to the least greater integer as few coordinates of $q_1^*,\ldots,q_s^*$ as possible. Having s fractional reminders of size ℓ/s, that entails that the integer-valued maximizer is found by rounding up exactly ℓ coordinates to the least greater integer (and rounding down the remaining $s-\ell$ coordinates). Finally, since $q_2,\ldots,q_s$ play a symmetric role in (11), the only two cases to study comprise either rounding up q_1^* and $\ell-1$ of the remaining coordinates, or rounding up ℓ coordinates among $q_2^*,\ldots,q_s^*$ (say, the first ℓ of them). The first of these cases yields $q_1=k+r$, $q_2=\cdots=q_\ell=k+1$, $q_{\ell+1}=\ldots=q_s=k$, while the second entails $q_1=k+r-1$, $q_2=\cdots=q_{\ell+1}=k+1$, $q_{\ell+2}=\ldots=q_s=k$. It is easy to check that these two choices achieve the same value for $M(q_1,\ldots,q_s)$, and the second one agrees with the form posited in (7) and (8), which is thus valid for arbitrary ℓ.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Appendix C:

Uniform sampling from $\mathcal{N}(r,s,n)$

A composition of a positive integer n into k parts is a representation $n=\sum_{i-1}^{k}p_i$ in which $p_1,\ldots,p_k$ are non-negative integers and the order of the summands matters. There are $J(n,k)=\binom{n+k-1}{n}$ possible compositions of n into k parts, and it is easy to draw a composition at random (uniformly) without necessarily generating the set of all of them (see Nijenhuis & Wilf, 1978, Chapters 5 and 6).

The entries of any confusion matrix $\mathbf{N}\in\mathcal{N}(r,s,n)$ constitute a composition of n into rs parts. The additional conditions $n_{i+}>0$ and $n_{+j}>0$ for all i,j, imposed in the definition of $\mathcal{N}(r,s,n)$ to ensure that the sizes of the associated compared clusterings match r and s, respectively, can be checked after arranging each drawn composition of n into rs parts by columns (say) into an $r\times s$ matrix, so uniform sampling from $\mathcal{N}(r,s,n)$ is guaranteed by rejection sampling.

For the simulations in Section 4.2 a sample of size 10^7 was drawn from $N(5,5,80)$ using the previous approach. The recorded rejection rate during the process was very low, approximately 0.47%, so the sampling algorithm is very efficient. Moreover, that rejection rate can also be interpreted as an estimate of the proportion of compositions of $n=80$ into $r\cdot s=25$ parts that cannot be converted (by columns) into a confusion matrix of $\mathcal{N}(5,5,80)$ and, since $J(80,25)\cong 2.319\times 10^{23}$, that yields an estimate of 2.309×10^{23} for the cardinality of $\mathcal{N}(5,5,80)$.

20448317, 2021, 2, Downloaded from https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bmsp.12212 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License