Journal of Machine Learning Research 18 (2017) 1-28 Submitted 1/17; Revised 6/17; Published 7/17

The Impact of Random Models on Clustering Similarity

Alexander J. Gates AJGATES@INDIANA.EDU

Yong-Yeol Ahn YYAHN@INDIANA.EDU

Department of Informatics and Program in Cognitive Science
Indiana University
919 East 10th Street
Bloomington, IN 47408, USA

Editor: Sebastian Nowozin

Abstract

Clustering is a central approach for unsupervised learning. After clustering is applied, the most fundamental analysis is to quantitatively compare clusterings. Such comparisons are crucial for the evaluation of clustering methods as well as other tasks such as consensus clustering. It is often argued that, in order to establish a baseline, clustering similarity should be assessed in the context of a random ensemble of clusterings. The prevailing assumption for the random clustering ensemble is the permutation model in which the number and sizes of clusters are fixed. However, this assumption does not necessarily hold in practice; for example, multiple runs of K-means clustering returns clusterings with a fixed number of clusters, while the cluster size distribution varies greatly. Here, we derive corrected variants of two clustering similarity measures (the Rand index and Mutual Information) in the context of two random clustering ensembles in which the number and sizes of clusters vary. In addition, we study the impact of one-sided comparisons in the scenario with a reference clustering. The consequences of different random models are illustrated using synthetic examples, handwriting recognition, and gene expression data. We demonstrate that the choice of random model can have a drastic impact on the ranking of similar clustering pairs, and the evaluation of a clustering method with respect to a random baseline; thus, the choice of random clustering model should be carefully justified.

Keywords: clustering comparison, clustering evaluation, adjustment for chance, Rand index, normalized mutual information

1. Introduction

Clustering is one of the most fundamental techniques of unsupervised learning and one of the most common ways to analyze data. Naturally, numerous methods have been developed and studied (Jain, 2010). To interpret clustering results, it is crucial to compare them to each other. For instance, the evaluation of a clustering method is usually carried out by comparing the method's results with a planted reference clustering, assuming that the more similar the method's solution is to the reference clustering, the better the method. This is particularly common in the field of complex networks in which clustering similarity measures are used to justify the performance of community detection methods (Danon et al., 2005; Lancichinetti and Fortunato, 2009). As quantitative comparison is a fundamental operation, it plays a key role in many other tasks. For instance, comparisons of clusterings can facilitate taxonomies for clustering solutions, can be used as a criteria for parameter estimation, and form the basis of consensus clustering methods (Meila, 2005; Vinh et al., 2009; Yeung et al., 2001).

Among the many clustering comparison methods (see, e.g. Meila, 2005; Pfitzner et al., 2009), two of the most prominent measures are the Rand index (Rand, 1971) and the Normalized Mutual Information (NMI, Danon et al., 2005). In both cases, the similarity score exists in the range $[0, 1]$, where 1 corresponds to identical clusterings and 0 implies maximally dissimilar clusterings. However, in practice, both measures do not efficiently use the full range of values in between 0 and 1, with many comparisons concentrating near the extreme values (Vinh et al., 2009; Hubert and Arabie, 1985). This makes it difficult to directly interpret the results of a comparison.

Thus, it is often argued that clustering similarity should be assessed in the context of a random ensemble of clusterings (Vinh et al., 2009; Hubert and Arabie, 1985; DuBien and Warde, 1981; DuBien et al., 2004; Albatineh et al., 2006; Romano et al., 2014; Zhang, 2015; Romano et al., 2016) and rescaled (see Equation 1). Such a correction for chance establishes a baseline by using the expected similarity of all pair-wise comparisons between clusterings specified by a random model; the resulting similarity values have a new interpretation that facilitates comparisons within a set of clusterings. Specifically, once corrected for chance, a similarity value of 1 still corresponds to identical clusterings, but a value of 0 now corresponds to the expected value amongst random clusterings. Positive values of corrected similarity better reflect an intuitive comparison of clusterings (Hubert and Arabie, 1985; Steinley et al., 2016). The correction may also introduce negative values when two clusterings are less similar than expected by chance.

The correction procedure requires two choices: *a model for random clusterings* and *how clusterings are drawn from the random model*. However, even the existence of these choices is usually ignored or relegated to the status of technical trivialities. Here, we demonstrate that these choices may dramatically affect results, and therefore the choice of a particular model for random clusterings should be justified based on the understanding of the clustering scenario. A poor choice of the random model may “not be random enough” and encode crucial features of the clusterings in all of the random clusterings, providing a poor baseline. At the same time, a random model may be “too random” in which crucial features are lost in a sea of random clusterings that are not representative of the particular problem. Characterizing random models is an important topic of research across statistical physics, network science, and combinatorial mathematics (Sethna, 2006; Goldenberg et al., 2010; Mansour, 2012). Yet, despite the importance of random model selection, almost no study that uses clustering comparison provides a justification for their choice of random model.

By far, the most common approach to correct clustering similarity for chance assumes that both clusterings are uniformly and independently sampled from the *permutation model* (M_{perm}). In the permutation model, the number and size of clusters within a clustering are fixed, and all random clusterings are generated by shuffling the elements between the fixed clusters. However, the premises of the permutation model are frequently violated; in many clustering scenarios, either the number of clusters, the size distribution of those clusters, or both vary drastically (Hubert and Arabie, 1985; Wallace, 1983). For example, K-means clustering, probably the most common technique, fixes the number of clusters but not the sizes of those clusters (Jain, 2010). Later, we explore a real example in which K-means produces clusterings with large variations in the clusterings' cluster size sequences. This suggests that comparing K-means clusterings based on M_{perm} is misleading.

Furthermore, even the assumption that both clusterings were randomly drawn from the same random model (a two-sided comparison) is often problematic. For example, when comparing against a given reference clustering, it is more reasonable to find the expected similarity of the reference clustering with all of the random clusterings from the random model. This *one-sided* comparison accounts for the fixed structure of the reference clustering which is always present in the comparisons, providing a more meaningful baseline.

Here, we present a general framework to adjust measures of clustering similarity for chance by considering a broader class of random clustering models and one-sided comparisons. Specifically, we consider two other random models for clusterings: a uniform distribution over the ensemble of all clusterings of N elements with the same number of clusters (M_{num}), and a uniform distribution over the ensemble of all clusterings of N elements (M_{all}). The resulting expectations for the Rand index under all three random models are summarized in Table 1 and for Mutual Information in Table 2, with the full derivations given in Section 4 and Section 5 respectively. The adjusted similarity measures used throughout this work rescale the Rand and MI measures by these expectations according to Equation 1. We

Adjusted Rand Index

$$\mathrm{ARI}_{\mathrm{model}}(\mathcal{A},\mathcal{B})=\frac{\mathrm{RI}(\mathcal{A},\mathcal{B})-\mathbb{E}_{\mathrm{model}}[\mathrm{RI}(\mathcal{A},\mathcal{B})]}{1.0-\mathbb{E}_{\mathrm{model}}[\mathrm{RI}(\mathcal{A},\mathcal{B})]}$$

Permutation Model

Two-sided = One-sided

$$\mathbb{E}_{\mathrm{perm}}[\mathrm{RI}(\mathcal{A},\mathcal{B})]=\frac{\sum_i\binom{a_i}{2}}{\binom{N}{2}}\frac{\sum_j\binom{b_j}{2}}{\binom{N}{2}}+\left(1-\frac{\sum_i\binom{a_i}{2}}{\binom{N}{2}}\right)\left(1-\frac{\sum_j\binom{b_j}{2}}{\binom{N}{2}}\right)$$

Fixed Number of Clusters

Two-sided

$$\mathbb{E}_{\mathrm{num}}[\mathrm{RI}(\mathcal{A},\mathcal{B})]=\frac{S(N-1,K_{\mathcal{A}})}{S(N,K_{\mathcal{A}})}\frac{S(N-1,K_{\mathcal{B}})}{S(N,K_{\mathcal{B}})}+\left(1-\frac{S(N-1,K_{\mathcal{A}})}{S(N,K_{\mathcal{A}})}\right)\left(1-\frac{S(N-1,K_{\mathcal{B}})}{S(N,K_{\mathcal{B}})}\right)$$

One-sided

$$\mathbb{E}^{1}_{\mathrm{num}}[\mathrm{RI}(\mathcal{A},\mathcal{G})]=\left(\frac{S(N-1,K_{\mathcal{A}})}{S(N,K_{\mathcal{A}})}\frac{\sum_j\binom{g_j}{2}}{\binom{N}{2}}\right)+\left(1-\frac{S(N-1,K_{\mathcal{A}})}{S(N,K_{\mathcal{A}})}\right)\left(1-\frac{\sum_j\binom{g_j}{2}}{\binom{N}{2}}\right)$$

All Clusterings

Two-sided

$$\mathbb{E}_{\mathrm{all}}[\mathrm{RI}(\mathcal{A},\mathcal{B})]=\left(\frac{B_{N-1}}{B_N}\right)^2+\left(1-\frac{B_{N-1}}{B_N}\right)^2$$

One-sided

$$\mathbb{E}^{1}_{\mathrm{all}}[\mathrm{RI}(\mathcal{A},\mathcal{G})]=\frac{B_{N-1}}{B_N}\frac{\sum_j\binom{g_j}{2}}{\binom{N}{2}}+\left(1-\frac{B_{N-1}}{B_N}\right)\left(1-\frac{\sum_j\binom{g_j}{2}}{\binom{N}{2}}\right)$$

Table 1: The expected Rand index between two random clusterings $\mathcal{A}$ and $\mathcal{B}$ of N elements, or random clustering $\mathcal{A}$ and reference clustering $\mathcal{G}$ under different random models. Details and derivations are given in Section 4.

Adjusted Mutual Information

$$\mathrm{AMI}_{\mathrm{model}}(\mathcal{A},\mathcal{B})=\frac{\mathrm{MI}(\mathcal{A},\mathcal{B})-\mathbb{E}_{\mathrm{model}}[\mathrm{MI}(\mathcal{A},\mathcal{B})]}{\max_{\mathrm{model}}[\mathrm{MI}(\mathcal{A},\mathcal{B})]-\mathbb{E}_{\mathrm{model}}[\mathrm{MI}(\mathcal{A},\mathcal{B})]}$$

Permutation Model

Two-sided = One-sided

$$\mathbb{E}_{\mathrm{perm}}[\mathrm{MI}(\mathcal{A},\mathcal{B})]=\mathbb{E}_{\mathrm{perm}}[H(\mathcal{A})]+\mathbb{E}_{\mathrm{perm}}[H(\mathcal{B})]-\mathbb{E}_{\mathrm{perm}}[H(\mathcal{A},\mathcal{B})]$$

$$\mathbb{E}_{\mathrm{perm}}[H(\mathcal{A})] \quad =-\sum_{i=1}^{K_{\mathcal{A}}}\frac{a_i}{N}\log\frac{a_i}{N}$$

$$\mathbb{E}_{\mathrm{perm}}[H(\mathcal{A},\mathcal{B})] =\sum_{i=1}^{K_{\mathcal{A}}}\sum_{j=1}^{K_{\mathcal{B}}}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{b_j}{n}\binom{N-b_j}{a_i-n}}{\binom{N}{a_i}}$$

Upper Bound

$$\max_{\mathrm{perm}}[\mathrm{MI}(\mathcal{A},\mathcal{B})]=\min\{\log H(\mathcal{A}),H(\mathcal{B})\}\ \mathrm{OR}\ \sqrt{H(\mathcal{A})H(\mathcal{B})}\ \mathrm{OR}\ \frac{1}{2}(H(\mathcal{A})+H(\mathcal{B}))\ \mathrm{OR}\ \max\{H(\mathcal{A}),H(\mathcal{B})\}$$

Fixed Number of Clusters

Two-sided

$$\mathbb{E}_{\mathrm{num}}[\mathrm{MI}(\mathcal{A},\mathcal{B})]=\mathbb{E}_{\mathrm{num}}[H(\mathcal{A})]+\mathbb{E}_{\mathrm{num}}[H(\mathcal{B})]-\mathbb{E}_{\mathrm{num}}[H(\mathcal{A},\mathcal{B})]$$

$$\mathbb{E}_{\mathrm{num}}[H(\mathcal{A})] \quad =-\sum_{k=1}\binom{N}{k}\frac{S(N-k,K_{\mathcal{A}}-1)}{S(N,K_{\mathcal{A}})}\frac{k}{N}\log\left(\frac{k}{N}\right)$$

$$\mathbb{E}_{\mathrm{num}}[H(\mathcal{A},\mathcal{B})] =-\sum_{k=1}\binom{N}{k}\frac{S(N-k,K_{\mathcal{A}}-1)}{S(N,K_{\mathcal{A}})}\sum_{m=1}\binom{N}{m}\frac{S(N-m,K_{\mathcal{B}}-1)}{S(N,K_{\mathcal{B}})}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{m}{n}\binom{N-m}{k-n}}{\binom{N}{k}}$$

One-sided

$$\mathbb{E}^{1}_{\mathrm{num}}[\mathrm{MI}(\mathcal{A},\mathcal{G})]=\mathbb{E}_{\mathrm{num}}[H(\mathcal{A})]+\mathbb{E}_{\mathrm{perm}}[H(\mathcal{G})]-\mathbb{E}^{1}_{\mathrm{num}}[H(\mathcal{A},\mathcal{G})]$$

$$\mathbb{E}^{1}_{\mathrm{num}}[H(\mathcal{A},\mathcal{G})] =-\sum_{k=1}\binom{N}{k}\frac{S(N-k,K_{\mathcal{A}}-1)}{S(N,K_{\mathcal{A}})}\sum_{j=1}^{K_{\mathcal{G}}}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{g_j}{n}\binom{N-g_j}{k-n}}{\binom{N}{k}}$$

Upper Bound

$$\max_{\mathrm{num}}[\mathrm{MI}(\mathcal{A},\mathcal{B})]=\min\{\log K_{\mathcal{A}},\log K_{\mathcal{B}}\}\ \mathrm{OR}\ \sqrt{\log K_{\mathcal{A}}\log K_{\mathcal{B}}}\ \mathrm{OR}\ \frac{1}{2}\log K_{\mathcal{A}}K_{\mathcal{B}}\ \mathrm{OR}\ \max\{\log K_{\mathcal{A}},\log K_{\mathcal{B}}\}$$

All Clusterings

Two-sided

$$\mathbb{E}_{\mathrm{all}}[\mathrm{MI}(\mathcal{A},\mathcal{B})]=\mathbb{E}_{\mathrm{all}}[H(\mathcal{A})]+\mathbb{E}_{\mathrm{all}}[H(\mathcal{B})]-\mathbb{E}_{\mathrm{all}}[H(\mathcal{A},\mathcal{B})]$$

$$\mathbb{E}_{\mathrm{all}}[H(\mathcal{A})] \quad =-\sum_{k=1}\binom{N}{k}\frac{B_{N-k}}{B_N}\frac{k}{N}\log\left(\frac{k}{N}\right)$$

$$\mathbb{E}_{\mathrm{all}}[H(\mathcal{A},\mathcal{B})] =-\sum_{k=1}\binom{N}{k}\frac{B_{N-k}}{B_N}\sum_{m=1}\binom{N}{m}\frac{B_{N-m}}{B_N}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{m}{n}\binom{N-m}{k-n}}{\binom{N}{k}}$$

One-sided

$$\mathbb{E}^{1}_{\mathrm{all}}[\mathrm{MI}(\mathcal{A},\mathcal{G})]=\mathbb{E}_{\mathrm{all}}[H(\mathcal{A})]+\mathbb{E}_{\mathrm{perm}}[H(\mathcal{G})]-\mathbb{E}^{1}_{\mathrm{all}}[H(\mathcal{A},\mathcal{G})]$$

$$\mathbb{E}^{1}_{\mathrm{all}}[H(\mathcal{A},\mathcal{B})] =-\sum_{k=1}\binom{N}{k}\frac{B_{N-k}}{B_N}\sum_{j=1}^{K_{\mathcal{G}}}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{g_j}{n}\binom{N-g_j}{k-n}}{\binom{N}{k}}$$

Upper Bound

$$\max_{\mathrm{all}}[\mathrm{MI}(\mathcal{A},\mathcal{B})]=\log N$$

Table 2: The expected Mutual Information between two random clusterings $\mathcal{A}$ and $\mathcal{B}$ of N elements, or random clustering $\mathcal{A}$ and reference clustering $\mathcal{G}$ under different random models. Details and derivations are given in Section 5.

also introduce one-sided variants of the adjusted Rand index and adjusted Mutual Information when using the M_{num} or M_{all} random models (for M_{perm}, the one-sided similarity is equivalent to the two-sided case).

The impact of our framework is illustrated in the case of two common tasks for adjusted clustering similarity measures: 1) ranking the similarity between pairs of clusterings (or finding the most similar clustering pair), and 2) evaluating the performance of a clustering method with respect to a random baseline. In Section 6, these tasks are demonstrated in the context of several examples: a synthetic clustering example, K-means clustering of a handwritten digits data set (MNIST), and an evaluation of hierarchical clustering applied to gene expression data. Our results demonstrate that both the choice of random model for clusterings and the choice of one-sided comparisons can affect results significantly. Therefore, we argue that clustering comparisons should be accompanied by a proper justification for the random model.

2. Clusterings

We first explicitly introduce a clustering of elements. Given a set of N distinct elements $V = \{v_1, \ldots, v_N\}$ (i.e. data points or vertices), a clustering is a partition of V into a set $\mathcal{C} = \{C_1, \ldots, C_{K_\mathcal{C}}\}$ of $K_\mathcal{C}$ non-empty disjoint subsets of V, the clusters, C_k, such that

1. $\forall C_i, C_j$ if $i \neq j$, then $C_i \cap C_j = \varnothing$
2. $\bigcup_{k=1}^{K_\mathcal{C}} C_k = V$.

Each clustering specifies a sequence of cluster sizes, namely, letting $c_i = |C_i|$ be the size of the i-th cluster, then the sequence of cluster sizes is $[c_1, c_2, \ldots, c_{K_\mathcal{C}}]$.

Throughout this paper, we focus on the similarity of two clusterings over the same set of N labeled elements, $\mathcal{A} = \{A_1, \ldots, A_{K_\mathcal{A}}\}$ (with $K_\mathcal{A}$ clusters of sizes a_i) and $\mathcal{B} = \{B_1, \ldots, B_{K_\mathcal{B}}\}$ (with $K_\mathcal{B}$ clusters of sizes b_j).

3. Correction for Chance

Given a clustering similarity measure s and a random model for clusterings: model, the expected clustering similarity $\mathbb{E}_{\text{model}}[s]$ of pair-wise comparisons within the random ensemble defined by the model corrects s for chance as follows (Hubert and Arabie, 1985)

$$\frac{s - \mathbb{E}_{\text{model}}[s]}{s_{\max} - \mathbb{E}_{model}[s]}. \tag{1}$$

The denominator rescales the adjusted similarity by the maximum similarity of pair-wise comparisons within the ensemble $s_{\max}$ so identical clusterings always have a similarity of 1.0. For some clustering similarity measures, the value of $s_{\max}$ is independent of the random model used; for example, the Rand index is always bounded above by 1.0. However, in the case of mutual information, the value of $s_{\max}$ depends on the random model used.

4. Rand Index

The Rand index (Rand, 1971) compares the number of element pairs which are either co-assigned to the same cluster, or assigned to different clusters in both clusterings, to the

$\mathcal{A}/\mathcal{B}$	B_1	B_2	$\ldots$	$B_{K_\mathcal{B}}$	Sums
A_1	n_{11}	n_{12}	$\ldots$	$n_{1K_\mathcal{B}}$	a_1
A_2	n_{21}	n_{22}	$\ldots$	$n_{2K_\mathcal{B}}$	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$A_{K_\mathcal{A}}$	$n_{K_\mathcal{A}1}$	$n_{K_\mathcal{A}2}$	$\ldots$	$n_{K_\mathcal{A}K_\mathcal{B}}$	$a_{K_\mathcal{A}}$
Sums	b_1	b_2	$\ldots$	$b_{K_\mathcal{B}}$	$\sum_{ij} n_{ij} = N$

Table 3: The contingency table $\mathcal{T}$ for two clusterings $\mathcal{A} = \{A_1, \ldots, A_{K_\mathcal{A}}\}$ and $\mathcal{B} = \{B_1, \ldots, B_{K_\mathcal{B}}\}$ of N elements, where $n_{ij} = |A_i \cap B_j|$ are the number of elements that are in both cluster $A_i \in \mathcal{A}$ and cluster $B_j \in \mathcal{B}$.

total number of element pairs. The most common formulation of the Rand index focuses on the following four sets of the $\binom{N}{2}$ element pairs: N_{11} the number of element pairs which are grouped in the same cluster in both clusterings, N_{10} the number of element pairs which are grouped in the same cluster by $\mathcal{A}$ but in different clusters by $\mathcal{B}$, N_{01} the number of element pairs which are grouped in the same cluster by $\mathcal{B}$ but in different clusters by $\mathcal{A}$, and N_{00} the number of element pairs which are grouped in different clusters by both $\mathcal{A}$ and $\mathcal{B}$. Intuitively, N_{11} and N_{00} are indicators of the agreement between the two clusterings, while N_{10} and N_{01} reflect the disagreement between the clusterings.

The aforementioned pair counts are identified from the contingency table $\mathcal{T}$ between two clusterings, shown in Table 3, by the following set of equations

$$\begin{aligned}
N_{11} &= \sum_{k,m=1}^{K_\mathcal{A},K_\mathcal{B}} \binom{n_{km}}{2} = \frac{1}{2}\left(\sum_{k,m=1}^{K_\mathcal{A},K_\mathcal{B}} n_{km}^2 - N\right) \\
N_{10} &= \sum_{k=1}^{K_\mathcal{A}} \binom{a_k}{2} - N_{11} = \frac{1}{2}\left(\sum_{k=1}^{K_\mathcal{A}} a_k^2 - \sum_{k,m=1}^{K_\mathcal{A},K_\mathcal{B}} n_{km}^2\right) \\
N_{01} &= \sum_{m=1}^{K_\mathcal{B}} \binom{b_m}{2} - N_{11} = \frac{1}{2}\left(\sum_{m=1}^{K_\mathcal{B}} b_m^2 - \sum_{k,m=1}^{K_\mathcal{A},K_\mathcal{B}} n_{km}^2\right) \\
N_{00} &= \binom{N}{2} - N_{11} - N_{10} - N_{01}
\end{aligned} \tag{2}$$

The Rand index between clusterings $\mathcal{A}$ and $\mathcal{B}$, $\mathrm{RI}(\mathcal{A}, \mathcal{B})$ is then given by the function

$$\begin{aligned}
\mathrm{RI}(\mathcal{A}, \mathcal{B}) &= \frac{N_{11} + N_{00}}{\binom{N}{2}} \\
&= \frac{2\sum_{k,m=1}^{K_\mathcal{A},K_\mathcal{B}} \binom{n_{km}}{2} - \sum_{k=1}^{K_\mathcal{A}} \binom{a_k}{2} - \sum_{m=1}^{K_\mathcal{B}} \binom{b_m}{2} + \binom{N}{2}}{\binom{N}{2}}.
\end{aligned} \tag{3}$$

It lies between 0 and 1, where 1 indicates the clusterings are identical and 0 occurs for clusters which do not share a single pair of elements (this only happens when one clustering is the full set of elements and the other clustering groups each element into its own cluster). As the number of clustered elements increases, the measure becomes dominated by the number of pairs which were classified into different clusters (N_{00}), resulting in decreased sensitivity to co-occurring element pairs (Fowlkes and Mallows, 1983).

Another formulation of the Rand index, used in our later derivations, focuses on a binary representation of the element pairs. Specifically, consider the vector $U_{\mathcal{A}} = [u_1, \ldots, u_{\binom{N}{2}}]$ with binary entries $u_\alpha \in \{-1, 1\}$ corresponding to all possible element pairs. Using α to index over all element pairs by $\alpha = \binom{N}{2} - \binom{N-i+1}{2} + j - i$, for $i < j \leq N$, then $u_\alpha = 1$ if elements v_i and v_j are in the same cluster in $\mathcal{A}$ and $u_\alpha = -1$ if elements v_i and v_j are in different clusters in $\mathcal{A}$. There are $Q_1^{\mathcal{A}}$ 1s in $U_{\mathcal{A}}$ and $Q_{-1}^{\mathcal{A}}$ -1s in $U_{\mathcal{A}}$ with

$$Q_1^{\mathcal{A}} = \sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2}, \quad Q_{-1}^{\mathcal{A}} = \binom{N}{2} - \sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2}. \tag{4}$$

The Rand index is found from the vectors $U_{\mathcal{A}}$ and $U_{\mathcal{B}}$, for clusterings $\mathcal{A}$ and $\mathcal{B}$ respectively, as the number of 1s in their product vector, $U_{\mathcal{A}} \odot U_{\mathcal{B}}$, using element-wise multiplication and normalized by the total size of the vectors, $\binom{N}{2}$.

4.1 Expected Rand Index, Permutation Model ($M_{\mathbf{perm}}$)

The expectation of the Rand index with respect to the permutation model follows from drawing the entries in Table 3 from the generalized hypergeometric distribution. Utilizing the previous notation with $Q_1^{\mathcal{A}} = \sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2}$, the expectation $\mathbb{E}_{\text{perm}}[RI(\mathcal{A}, \mathcal{B})]$ of the Rand index with respect to the permutation model for the cluster size sequences of clusterings $\mathcal{A}$ and $\mathcal{B}$ is given by

$$\mathbb{E}_{\text{perm}}[\text{RI}(\mathcal{A}, \mathcal{B})] = \frac{2Q_1^{\mathcal{A}}Q_1^{\mathcal{B}} - \binom{N}{2}\left(Q_1^{\mathcal{A}} + Q_1^{\mathcal{B}}\right) + \binom{N}{2}^2}{\binom{N}{2}^2} \tag{5}$$

(see Fowlkes and Mallows, 1983, Hubert and Arabie, 1985, or Albatineh and Niewiadomska-Bugaj, 2011 for the full derivation).

The commonly used adjusted Rand index (ARI) of Hubert and Arabie (1985) uses M_{perm} to calculate the expectation of the Rand index, $\mathbb{E}_{\text{perm}}[\text{RI}(\mathcal{A}, \mathcal{B})]$, as found in Equation 5. This expectation is then used in Equation 1, along with the fact that the maximum value of the Rand index is $\max_{\text{perm}}[\text{RI}] = 1.0$, to give

$$\text{ARI}_{\text{perm}}(\mathcal{A}, \mathcal{B}) = \frac{\binom{N}{2} \sum_{k,m=1}^{K_{\mathcal{A}} K_{\mathcal{B}}} \binom{n_{km}}{2} - \sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2} \sum_{m=1}^{K_{\mathcal{B}}} \binom{b_m}{2}}{\frac{1}{2}\binom{N}{2} \left[\sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2} + \sum_{m=1}^{K_{\mathcal{B}}} \binom{b_m}{2}\right] - \sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2} \sum_{m=1}^{K_{\mathcal{B}}} \binom{b_m}{2}}. \tag{6}$$

4.2 Expected Rand Index, Fixed Number of Clusters

We follow DuBien and Warde (1981) to calculate the Rand index between two clusterings under the assumptions that both clusterings were independently and uniformly drawn from the ensemble of clusterings with a fixed number of clusters (M_{num}). Recall that the Rand index between two clusterings $\mathcal{A}$ and $\mathcal{B}$ is given by the number of 1s in the element-wise product of the binary representations vectors $U_{\mathcal{A}}$ and $U_{\mathcal{B}}$. The expected Rand index under any random model is then the expected number of 1s in this product vector, normalized by the total size of the vector

$$\begin{aligned}\mathbb{E}[RI(\mathcal{A},\mathcal{B})] &= \mathbb{E}\left[\frac{1}{\binom{N}{2}}\sum_{\alpha=1}^{\binom{N}{2}} \mathbf{1}_{u_\alpha^{\mathcal{A}}\cdot u_\alpha^{\mathcal{B}}=1}\right] \qquad (7)\\ &= \frac{1}{\binom{N}{2}}\sum_{\alpha=1}^{\binom{N}{2}} \mathbb{E}\left[\mathbf{1}_{u_\alpha^{\mathcal{A}}\cdot u_\alpha^{\mathcal{B}}=1}\right]\\ &= \frac{1}{\binom{N}{2}}\sum_{\alpha=1}^{\binom{N}{2}} P(u_\alpha^{\mathcal{A}}\cdot u_\alpha^{\mathcal{B}}=1)\end{aligned}$$

The product $u_\alpha^{\mathcal{A}} \cdot u_\alpha^{\mathcal{B}}$ equals 1 when either $u_\alpha^{\mathcal{A}} = 1$ and $u_\alpha^{\mathcal{B}} = 1$, or $u_\alpha^{\mathcal{A}} = -1$ and $u_\alpha^{\mathcal{B}} = -1$. Since we assumed both clusterings were independent, this gives

$$P(u_\alpha^{\mathcal{A}}\cdot u_\alpha^{\mathcal{B}}=1) = P(u_\alpha^{\mathcal{A}}=1)P(u_\alpha^{\mathcal{B}}=1) + P(u_\alpha^{\mathcal{A}}=-1)P(u_\alpha^{\mathcal{B}}=-1) \qquad (8)$$

where $P(u_\alpha^{\mathcal{A}} = 1)$ is the probability that the two elements v_i and v_j are in the same cluster in clustering $\mathcal{A}$, where the element pair is indexed by $\alpha = \binom{N}{2} - \frac{(N-i)(N-i+1)}{2} + j - i$ with $i < j \leq N$. Likewise, $P(u_\alpha^{\mathcal{A}} = -1)$ is the probability that the two elements v_i and v_j are in different clusters.

Under the assumption of M_{num}, there is a uniform probability of selecting a clustering from the $S(N, K_{\mathcal{A}})$ clusterings of N elements into $K_{\mathcal{A}}$ clusters; we define, $P_{\text{num}}(u_\alpha^{\mathcal{A}} = 1)$ as the proportion of these clusterings with elements v_i and v_j in the same cluster. To find this proportion, notice that we can ensure v_i is in the same cluster as v_j by first partitioning all elements besides v_i into $K_{\mathcal{A}}$ clusters; then, we can add v_i to the same cluster as v_j. Since there are $S(N-1, K_{\mathcal{A}})$ such clusterings without element v_i, this gives

$$P_{\text{num}}(u_\alpha^{\mathcal{A}}=1) = \frac{S(N-1,K_{\mathcal{A}})}{S(N,K_{\mathcal{A}})} \qquad (9)$$

$$P_{\text{num}}(u_\alpha^{\mathcal{A}}=-1) = 1 - \frac{S(N-1,K_{\mathcal{A}})}{S(N,K_{\mathcal{A}})}. \qquad (10)$$

Finally, the expected Rand index between two clusterings $\mathcal{A}$ and $\mathcal{B}$ with $K_{\mathcal{A}}$ and $K_{\mathcal{B}}$ clusters assuming M_{num} is given by

$$\begin{aligned}\mathbb{E}_{\text{num}}[\text{RI}(\mathcal{A},\mathcal{B})] = {}& \frac{S(N-1,K_{\mathcal{A}})}{S(N,K_{A})}\frac{S(N-1,K_{\mathcal{B}})}{S(N,K_{\mathcal{B}})}\\ &+\left(1-\frac{S(N-1,K_{\mathcal{A}})}{S(N,K_{\mathcal{A}})}\right)\left(1-\frac{S(N-1,K_{\mathcal{B}})}{S(N,K_{\mathcal{B}})}\right). \qquad (11)\end{aligned}$$

When N is large, we can approximate the Stirling numbers of the second kind for a fixed K by $S(N, K) \approx \frac{K^N}{K!}$. This can be inserted into equation (11) to give the following approximation for the mean of the Rand index assuming M_{num}

$$\mathbb{E}_{\text{num}}[\text{RI}(\mathcal{A},\mathcal{B})] \approx \frac{1}{K_{\mathcal{A}}K_{\mathcal{B}}} + \left(1-\frac{1}{K_{\mathcal{A}}}\right)\left(1-\frac{1}{K_{\mathcal{B}}}\right). \qquad (12)$$

Interestingly, this suggests that the Rand index goes to 1 at a rate inversely related to the smaller number of clusters $\mathcal{O}(\max\{K_{\mathcal{A}}^{-1}, K_{\mathcal{B}}^{-1}\})$.

4.3 Expected Rand Index, All Clusterings M_{all}

The average of the Rand index between two clusterings under the assumption that the clusterings were drawn with uniform probability from the set of all clusterings directly follows from the random model with a fixed number of clusters previously discussed. Namely, because Bell numbers are related to Stirling numbers of the second kind by $B_N = \sum_{k=1}^{N} S(N,k)$, a similar reasoning as followed for equation (9) gives

$$\begin{align} P_{\text{all}}(u_\alpha = 1) &= \sum_{k=1}^{N} \frac{S(N,k)}{B_N} P_{\text{num}}(u_\alpha^k = 1) \tag{13}\\ &= \frac{1}{B_N} \sum_{k=1}^{N} S(N,k) \frac{S(N-1,k)}{S(N,k)} \notag\\ &= \frac{B_{N-1}}{B_N}. \tag{14} \end{align}$$

Using this probability for the expectation in equation (7) gives the expected Rand index under the assumption that both clusterings were uniformly drawn from the set of all clusterings of N elements

$$\mathbb{E}_{\text{all}}[\text{RI}(\mathcal{A},\mathcal{B})] = \left(\frac{B_{N-1}}{B_N}\right)^2 + \left(1 - \frac{B_{N-1}}{B_N}\right)^2. \tag{15}$$

When N is large, we can approximate the ratio of successive Bell numbers by $\frac{B_{N+1}}{B_N} \approx \frac{N}{\log N}$. Using this approximation in equation (15) gives the following approximation for the mean of the Rand index in M_{all}

$$\mathbb{E}_{\text{all}}[\text{RI}(\mathcal{A},\mathcal{B})] \approx \left(\frac{\log N}{N}\right)^2 + \left(1 - \frac{\log N}{N}\right)^2. \tag{16}$$

Interestingly, this suggests that the expected Rand index between two random clusterings goes to 1 at a rate $\mathcal{O}\left(\frac{\log(N)}{N}\right)$, inversely proportional to the number of elements.

4.4 One-Sided Rand

Consider a reference clustering $\mathcal{G}$ that has the cluster size sequence $[g_1, \ldots, g_{K_{\mathcal{G}}}]$. The binary pair vector representation of $\mathcal{G}$ has $Q_1^{\mathcal{G}}$ 1s and $Q_{-1}^{\mathcal{G}} = \binom{N}{2} - Q_1^{\mathcal{G}}$, -1s. The one-sided expectation of the Rand index under the assumption that clustering $\mathcal{A}$ was randomly drawn from either the M_{num} or M_{all} random models follows from treating the two clusterings independently as in equation (8). Since the cluster sequence for the reference clustering is fixed, the probability that a random entry in the binary pair vector is 1 is given by the fraction of 1s in the vector

$$
\begin{aligned}
P_{\text{num}}(u^{\mathcal{G}}_{\alpha} = 1) &= \frac{1}{\binom{N}{2}} Q^{\mathcal{G}}_1 \\
&= \frac{1}{\binom{N}{2}} \sum_{i=1}^{K_{\mathcal{G}}} \binom{g_i}{2}.
\end{aligned} \tag{17}
$$

The one-sided expectation of the Rand index under the assumption that clustering $\mathcal{A}$ was randomly drawn from the set of all clusterings with a fixed number of clusters M^1_{num} is

$$
\mathbb{E}^1_{\text{num}}[\text{RI}(\mathcal{A}, \mathcal{G})] = \left(\frac{S(N-1, K_{\mathcal{A}})}{S(N, K_{\mathcal{A}})} \frac{Q^{\mathcal{G}}_1}{\binom{N}{2}} \right) + \left(1 - \frac{S(N-1, K_{\mathcal{A}})}{S(N, K_{\mathcal{A}})} \right) \left(1 - \frac{Q^{\mathcal{G}}_1}{\binom{N}{2}} \right). \tag{18}
$$

The one-sided expectation of the Rand index with the assumption that the random clustering $\mathcal{A}$ is drawn from the ensemble of all partitions M^1_{all} is

$$
\mathbb{E}^1_{\text{all}}[\text{RI}(\mathcal{A}, \mathcal{G})] = \frac{B_{N-1}}{B_N} \frac{Q^{\mathcal{G}}_1}{\binom{N}{2}} + \left(1 - \frac{B_{N-1}}{B_N} \right) \left(1 - \frac{Q^{\mathcal{G}}_1}{\binom{N}{2}} \right). \tag{19}
$$

5. Mutual Information

Another prominent family of clustering similarity measures is based on the Shannon information between probabilistic representations of each clustering. These probability distributions are also calculated from the contingency table $\mathcal{T}$, Table 3. The partition entropy H of a clustering $\mathcal{A}$ is given by

$$
H(\mathcal{A}) = -\sum_{k=1}^{K_{\mathcal{A}}} \frac{a_k}{N} \log \frac{a_k}{N}. \tag{20}
$$

Using this entropy, the mutual information $\text{MI}(\mathcal{A}, \mathcal{B})$ between two clusterings $\mathcal{A}$ and $\mathcal{B}$ is given by

$$
\begin{aligned}
\text{MI}(\mathcal{A}, \mathcal{B}) &= H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}, \mathcal{B}) \\
&= \sum_{k,m=1}^{K_{\mathcal{A}}, K_{\mathcal{B}}} \frac{n_{km}}{N} \log \frac{n_{km} N}{a_k b_m}.
\end{aligned} \tag{21}
$$

The mutual information can be interpreted as an inverse measure of independence between the clusterings, or a measure of the amount of information each clustering has about the other. As it can vary in the range $[0, \min\{H(\mathcal{A}), H(\mathcal{B})\}]$, to facilitate comparisons, it is desirable to normalize it to the range $[0, 1]$. There are at least six proposals in the literature for this upper bound, each with different advantages and drawbacks

$$
\begin{aligned}
\min\{H(\mathcal{A}), H(\mathcal{B})\} \leq \sqrt{H(\mathcal{A}) H(\mathcal{B})} \leq \frac{H(\mathcal{A}) + H(\mathcal{B})}{2} \\
\leq \max\{H(\mathcal{A}), H(\mathcal{B})\} \leq \max\{\log K_{\mathcal{A}}, \log K_{\mathcal{B}}\} \leq \log N.
\end{aligned} \tag{22}
$$

The resulting measures are all known as normalized mutual information (NMI). This measure has been said to exhibit more desirable properties than the Rand index; for example, it is dependent on the relative proportions of the cluster sizes in each clustering rather than the number of elements. However, due to its dependence on the number of clusters in each clustering, it is known to favor comparisons between clusterings with more clusters regardless of any other shared clustering features (White and Liu, 1994; Vinh et al., 2010; Amelio and Pizzuti, 2015).

5.1 Expected Mutual Information, Permutation Model ($M_{\mathbf{perm}}$)

The mutual information between two clusterings has also previously been studied under the assumption that both clusterings were randomly generated from the permutation model (Vinh et al., 2009; Romano et al., 2014; Vinh et al., 2010). Expanding the definition of the mutual information gives

$$\begin{aligned}\mathbb{E}_{\text{perm}}[\text{MI}(\mathcal{A},\mathcal{B})] &= \mathbb{E}_{\text{perm}}[H(\mathcal{A})] + \mathbb{E}_{\text{perm}}[H(\mathcal{B})] - \mathbb{E}_{\text{perm}}[H(\mathcal{A},\mathcal{B})] \\ &= H(\mathcal{A}) + H(\mathcal{B}) - \mathbb{E}_{\text{perm}}[H(\mathcal{A},\mathcal{B})]\end{aligned} \tag{23}$$

where the second line follows from the fact that all cluster sizes (and hence the entropy) are the same for every clustering in M_{perm}.

The expectation of the joint entropy with respect to M_{perm} for the cluster size distributions of clusterings $\mathcal{A}$ and $\mathcal{B}$ is the average over all possible contingency tables $\mathcal{T}$ with entries n

$$\mathbb{E}_{\text{perm}}[H(\mathcal{A},\mathcal{B})] = -\sum_{\mathcal{T}} p(\mathcal{T}|\mathcal{A},\mathcal{B}) \sum_{k=1}^{K_{\mathcal{A}}} \sum_{m=1}^{K_{\mathcal{B}}} \frac{n}{N} \log\left(\frac{n}{N}\right). \tag{24}$$

Rearranging the summations, and recalling that the entries of the contingency tables are hyper-geometrically distributed such that the probability of each entry

$$p(n) = \frac{\binom{b_m}{n}\binom{N-b_m}{a_k-n}}{\binom{N}{a_k}} \tag{25}$$

is only dependent on the row sum a_k and column sum b_m, gives

$$\mathbb{E}_{\text{perm}}[H(\mathcal{A},\mathcal{B})] = -\sum_{k=1}^{K_{\mathcal{A}}} \sum_{m=1}^{K_{\mathcal{B}}} \sum_{n} \frac{n}{N} \log\left(\frac{n}{N}\right) \frac{\binom{b_m}{n}\binom{N-b_m}{a_k-n}}{\binom{N}{a_k}}. \tag{26}$$

According to the hyper-geometric distribution, the summation over table entries n_{km} occurs between the lower bound: $\max\{0, a_k + b_m - N\}$ and the upper bound: $\min\{a_k, b_m\}$. Combining this expression with the individual entropies $H(\mathcal{A})$ and $H(\mathcal{B})$ gives (Vinh et al., 2009)

$$\mathbb{E}_{\text{perm}}[\text{MI}(\mathcal{A},\mathcal{B})] = \sum_{k=1}^{K_{\mathcal{A}}} \sum_{m=1}^{K_{\mathcal{B}}} \sum_{n} \frac{n}{N} \log\left(\frac{Nn}{a_k b_m}\right) \frac{\binom{b_m}{n}\binom{N-b_m}{a_k-n}}{\binom{N}{a_k}}. \tag{27}$$

As shown in Romano et al. (2014), the computational complexity of calculating the expected mutual information assuming the permutation model is of order $\mathcal{O}(\max\{K_{\mathcal{A}}N, K_{\mathcal{B}}N\})$.

The adjusted mutual information (AMI) of Vinh et al. (2009) uses M_{perm} to correct the MI for chance according to equation (1) and selecting an upper bound max[MI] from equation (22) to give

$$\text{AMI}(\mathcal{A},\mathcal{B}) = \frac{\text{MI}(\mathcal{A},\mathcal{B}) - \mathbb{E}_{\text{perm}}[\text{MI}(\mathcal{A},\mathcal{B})]}{\max_{\text{perm}}[\text{MI}] - \mathbb{E}_{\text{perm}}[\text{MI}(\mathcal{A},\mathcal{B})]}. \tag{28}$$

5.2 Expected Mutual Information, Fixed Number of Clusters (M_{num})

Next, we consider the Mutual Information between two clusterings under the assumptions that both clusterings were independently and uniformly drawn from the ensemble of clusterings with a fixed number of clusters (M_{num}). In this case, the expected mutual information is dependent on both the average partition entropy and the joint partition entropy

$$\mathbb{E}_{\text{num}}[\text{MI}(\mathcal{A},\mathcal{B})] = \mathbb{E}_{\text{num}}[H(\mathcal{A})] + \mathbb{E}_{\text{num}}[H(\mathcal{B})] - \mathbb{E}_{\text{num}}[H(\mathcal{A},\mathcal{B})]. \tag{29}$$

This expectation can be found by considering the average partition entropy and joint partition entropy separately. Recall that in the permutation model $\mathbb{E}_{\text{perm}}[H(\mathcal{A})] = H(\mathcal{A})$ since the cluster sizes remain unchanged; however, the same does not hold in M_{num}. Denoting a random clustering with $K_{\mathcal{A}}$ clusters as $\pi_{K_{\mathcal{A}}}$, and using the notion $\sum_{\sigma_i \in \pi_{K_{\mathcal{A}}}}$ to indicate the summation over all clusters in the clustering $\pi_{K_{\mathcal{A}}}$, where the cardinality of the cluster is $|\sigma_i| = a$, then the expected partition entropy of a random clustering in M_{num} is

$$\mathbb{E}_{\text{num}}[H(\mathcal{A})] = -\sum_{\pi_{K_{\mathcal{A}}}} p_{\text{num}}(\pi_{K_{\mathcal{A}}}) \sum_{\sigma_i \in \pi_{K_{\mathcal{A}}}} \frac{a}{N} \log\left(\frac{a}{N}\right). \tag{30}$$

Note that this expression only depends on the size $a = |\sigma_i|$ for $\sigma_i \in \pi_{K_{\mathcal{A}}}$ of the clusters in the clustering. This means the expected entropy of a random clustering can be rewritten in terms of the expected contribution to the entropy from a random cluster of size a. A counting argument gives the number of clusters of size a which appear in all of the clusterings in the random ensemble. First, choose a of the N elements to form the cluster. Each clustering in M_{num} must have $K_{\mathcal{A}}$ clusters, so the remaining $N-a$ elements have to be arranged into $K_{\mathcal{A}}-1$ other clusters. There are $S(N-a, K_{\mathcal{A}}-1)$ ways to partition these remaining elements. This gives $\binom{N}{a} S(N-a, K_{\mathcal{A}}-1)$ clusters of size a (Chern et al., 2014). The expected number of clusters $n_{\text{num}}(a)$ in a random clustering drawn from M_{num} is then $\binom{N}{a}\frac{S(N-a,K_{\mathcal{A}}-1)}{S(N,K_{\mathcal{A}})}$. Therefore, the expected clustering entropy in M_{num} is:

$$\mathbb{E}_{\text{num}}[H(\mathcal{A})] = -\sum_{a=1}^{N-(K_{\mathcal{A}}-1)} \binom{N}{a} \frac{S(N-a, K_{\mathcal{A}}-1)}{S(N, K_{\mathcal{A}})} \frac{a}{N} \log\left(\frac{a}{N}\right). \tag{31}$$

where the summation is over all possible cluster sizes $[1, N-(K_{\mathcal{A}}-1)]$ encountered when partitioning N elements into $K_{\mathcal{A}}$ clusters.

Similarly, the expected joint entropy of two random clusterings drawn independently from M_{num} is given by the expected number of clusters of size a from a clustering with

$K_{\mathcal{A}}$ clusters, the expected number of clusters of size b from a clustering with $K_{\mathcal{B}}$ clusters, and then considering the probability of overlap $p(n_{km}) = \frac{\binom{b}{n_{km}}\binom{N-b}{a-n_{km}}}{\binom{N}{a}}$ from the resulting random contingency table

$$\begin{aligned}\mathbb{E}_{\text{num}}[H(\mathcal{A},\mathcal{B})] &= -\sum_{\pi_{K_{\mathcal{A}}}} p_{\text{num}}(\pi_{K_{\mathcal{A}}}) \sum_{\pi_{K_{\mathcal{B}}}} p_{\text{num}}(\pi_{K_{\mathcal{B}}}) \\ &\times \sum_{k=1}^{K_{\mathcal{A}}} \sum_{m=1}^{K_{\mathcal{B}}} \sum_{n} \frac{n}{N} \log\left(\frac{Nn}{a_k b_m}\right) \frac{\binom{b_m}{n}\binom{N-b_m}{a_k-n}}{\binom{N}{a_k}} \\ &= -\sum_{a=1}^{N-(K_{\mathcal{A}}-1)} \sum_{b=1}^{N-(K_{\mathcal{B}}-1)} \sum_{n} \left[\binom{N}{a}\frac{S(N-a,K_{\mathcal{A}}-1)}{S(N,K_{\mathcal{A}})}\right. \\ &\times \left.\binom{N}{b}\frac{S(N-b,K_{\mathcal{B}}-1)}{S(N,K_{\mathcal{B}})}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{b}{n}\binom{N-b}{a-n}}{\binom{N}{a}}\right].\end{aligned} \tag{32}$$

Note that, when using equation (1) to adjust the mutual information for chance under the assumption of M_{num}, the maximum value for the measure over the entire ensemble of random clusterings has to be used. When considering clusterings with a fixed number of clusters, we know that $H(\mathcal{A}) \leq \log K_{\mathcal{A}}$. This means that the choices for $\max_{\text{num}}[\text{MI}(\mathcal{A},\mathcal{B})]$ are

$$\min\{\log K_{\mathcal{A}}, \log K_{\mathcal{B}}\} \leq \sqrt{\log K_{\mathcal{A}} \log K_{\mathcal{B}}} \leq \frac{1}{2}\log K_{\mathcal{A}}K_{\mathcal{B}} \leq \max\{\log K_{\mathcal{A}}, \log K_{\mathcal{B}}\}. \tag{33}$$

As is apparent from the summations in equation (32), the computational complexity of exactly calculating the expected mutual information assuming M_{num} is of order $\mathcal{O}(N^3)$.

5.3 Expected Mutual Information, All Clusterings M_{all}

The expected Mutual Information between two clusterings under the assumption that both clusterings were independently and uniformly drawn from the set of all possible clusterings of N elements, M_{all}, has a similar derivation as the previous case of M_{num}. Both the expectations for the entropy of a single clustering and the joint entropy of the two clusterings need to be considered separately and can be rewritten in terms of the contributions from individual clusters of a given size. In M_{all}, the number of clusters of size a is again found by choosing a of the N elements for the cluster and then partitioning the remaining $N-a$ elements; there are now B_{N-a} possible ways to cluster the remaining elements (Chern et al., 2014). This gives

$$\mathbb{E}_{\text{all}}[H(\mathcal{A})] = -\sum_{a=1}^{N}\binom{N}{a}\frac{B_{N-a}}{B_N}\frac{a}{N}\log\left(\frac{a}{N}\right). \tag{34}$$

The expected joint entropy for two clusterings is then

$$\begin{aligned}
\mathbb{E}_{\text{all}}[H(\mathcal{A},\mathcal{B})] &= -\sum_{\pi_a} p(\pi_a) \sum_{\pi_b} p(\pi_b) \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{n} \frac{n}{N}\log\left(\frac{Nn}{a_i b_j}\right)\frac{\binom{b_j}{n}\binom{N-b_j}{a_i-n}}{\binom{N}{a_i}} \\
&= -\sum_{a=1}^{N}\binom{N}{a}\frac{B_{N-a}}{B_N}\sum_{b=1}^{N}\binom{N}{b}\frac{B_{N-b}}{B_N}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{b}{n}\binom{N-b}{a-n}}{\binom{N}{a}} \\
&= -2\sum_{a=1}^{N}\sum_{b=1}^{a-1}\binom{N}{a}\frac{B_{N-a}}{B_N}\binom{N}{b}\frac{B_{N-b}}{B_N}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{b}{n}\binom{N-b}{a-n}}{\binom{N}{a}} \qquad (35)\\
&\quad -\sum_{a=1}^{N}\left(\binom{N}{a}\frac{B_{N-a}}{B_N}\right)^2\sum_{n}\frac{n_{ij}}{N}\log\left(\frac{n}{N}\right)\frac{\binom{a}{n}\binom{N-a}{a-n}}{\binom{N}{a}},
\end{aligned}$$

with the last simplification resulting from the symmetry of the hyper-geometric term with respect to a and b.

As in the previous case, the maximum bound of the measure must be consider over the entire ensemble of clusterings. Again, we consider the bound $H(\mathcal{A}) \leq \log N$. This reduces to only one choice for $\max_{\text{all}}[\text{MI}(\mathcal{A},\mathcal{B})] = \log N$.

5.4 One-Sided Mutual Information

As was the case for the one-sided expected Rand index, the one-sided expectation of mutual information follows from the fact that the cluster sequence for the reference clustering is fixed. This results in the following one-sided expected joint entropy when the random clustering $\mathcal{A}$ is drawn from the M_{num} model

$$\mathbb{E}^1_{\text{num}}[H(\mathcal{A},\mathcal{G})] = -\sum_{a=1}^{N}\binom{N}{a}\frac{S(N-a,K_{\mathcal{A}}-1)}{S(N,K_{\mathcal{A}})}\sum_{b=1}^{K_{\mathcal{G}}}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{g_b}{n}\binom{N-g_b}{a-n}}{\binom{N}{a}}. \qquad (36)$$

The corresponding one-sided expected MI assuming M^1_{num} is

$$\begin{aligned}
\mathbb{E}^1_{\text{num}}[\text{MI}(\mathcal{A},\mathcal{G})] &= -\sum_{a=1}^{K_{\mathcal{A}}}\binom{N}{a}\frac{S(N-a,K_{\mathcal{A}}-1)}{S(N,K_{\mathcal{A}})}\frac{a}{N}\log\left(\frac{a}{N}\right) - \sum_{b=1}^{K_{\mathcal{G}}}\frac{g_b}{N}\log\left(\frac{g_b}{N}\right) \qquad (37)\\
&\quad +\sum_{a=1}^{N}\binom{N}{a}\frac{S(N-a,K_{\mathcal{A}}-1)}{S(N,K_{\mathcal{A}})}\sum_{b=1}^{K_{\mathcal{G}}}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{g_b}{n}\binom{N-g_b}{a-n}}{\binom{N}{a}}.
\end{aligned}$$

The one-sided expected joint entropy when the random clustering $\mathcal{A}$ is drawn from the M^1_{all} model is

$$\mathbb{E}^1_{\text{all}}[H(\mathcal{A},\mathcal{G})] = -\sum_{a=1}^{N}\binom{N}{a}\frac{B_{N-a}}{B_N}\sum_{b=1}^{K_{\mathcal{G}}}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{g_b}{n}\binom{N-g_b}{a-n}}{\binom{N}{a}}, \qquad (38)$$

and the one-sided expectation of the MI when the random clustering $\mathcal{A}$ is drawn from the M^1_{all} model is

$$\begin{aligned}\mathbb{E}^1_{\text{all}}[\text{MI}(\mathcal{A},\mathcal{G})] = &-\sum_{a=1}^{K_{\mathcal{A}}} \binom{N}{a}\frac{B_{N-a}}{B_N}\frac{a}{N}\log\left(\frac{a}{N}\right) - \sum_{b=1}^{K_{\mathcal{G}}}\frac{g_b}{N}\log\left(\frac{g_b}{N}\right) \\ &+\sum_{a=1}^{N}\binom{N}{a}\frac{B_{N-a}}{B_N}\sum_{b=1}^{K_{\mathcal{G}}}\sum_{n}\frac{n}{N}\log\left(\frac{n}{N}\right)\frac{\binom{g_b}{n}\binom{N-g_b}{a-n}}{\binom{N}{a}}.\end{aligned} \tag{39}$$

Again, the maximum bound must be chosen with respect to the measure maximum over the clusterings present in the random model.

6. Results

The choice of random model for clusterings and the choice of one-sided comparisons can significantly affect results of clustering comparisons. We first illustrate that the ranking of similar clustering pairs (or, equivalently, finding the most similar clustering pair) depends on the choices of random models in a hypothetical example (Section 6.1) and K-means clustering of a handwritten digits data set (Section 6.2). One of the primary reasons such strong discrepancies occur is that the cluster size sequences are fixed within samples from M_{perm}. This means that adjusted comparisons using M_{perm} are unable to differentiate random clusterings with drastically different cluster size sequences, as we illustrate through our third example in Section 6.3. Second, we demonstrate that the interpretation of adjusted clustering similarity measures with respect to a random baseline also depends on the random model through an evaluation of hierarchical clustering applied to gene expression data in Section 6.4. Crucially, all of these examples illustrate that conclusions based on corrected similarity measures can change depending on the random model for clusterings.

6.1 Clustering Similarity Ranking

Our first example demonstrates how rankings assigned by the similarity score can change depending on the assumed random model. Consider the four hypothetical clusterings of 20 elements presented in Figure 1a. Clustering $\mathcal{W}$ contains four equally sized clusters; clustering $\mathcal{X}$ is generated by shifting the membership of one element from $\mathcal{W}$; clustering $\mathcal{Y}$ groups the elements into 10 equally sized clusters; and clustering $\mathcal{Z}$ groups the elements into 10 heterogeneous clusters. The similarity (from the most similar at the top to the least similar at the bottom) of all 6 clustering pairs is ranked using the Rand index and each of its three adjusted variants in Figure 1b. Note that the adjusted Rand index can be negative. The unadjusted Rand index ranking serves as a reference to illustrate how the random models change rankings.

As one would expect, all four Rand measures identify clusterings $\mathcal{W}$ and $\mathcal{X}$ as the most similar (Figure 1b). However, the ranking of the other five comparisons varies widely as a result of the underlying random models. These changes can be understood by tracking comparisons to clustering $\mathcal{Z}$. The low Rand index for the three comparisons with clustering $\mathcal{Z}$ (≈ 0.76) reflects the fact that clustering $\mathcal{Z}$ has a drastically different number of clusters or cluster size sequence from the other three clusterings. The permutation model retains these

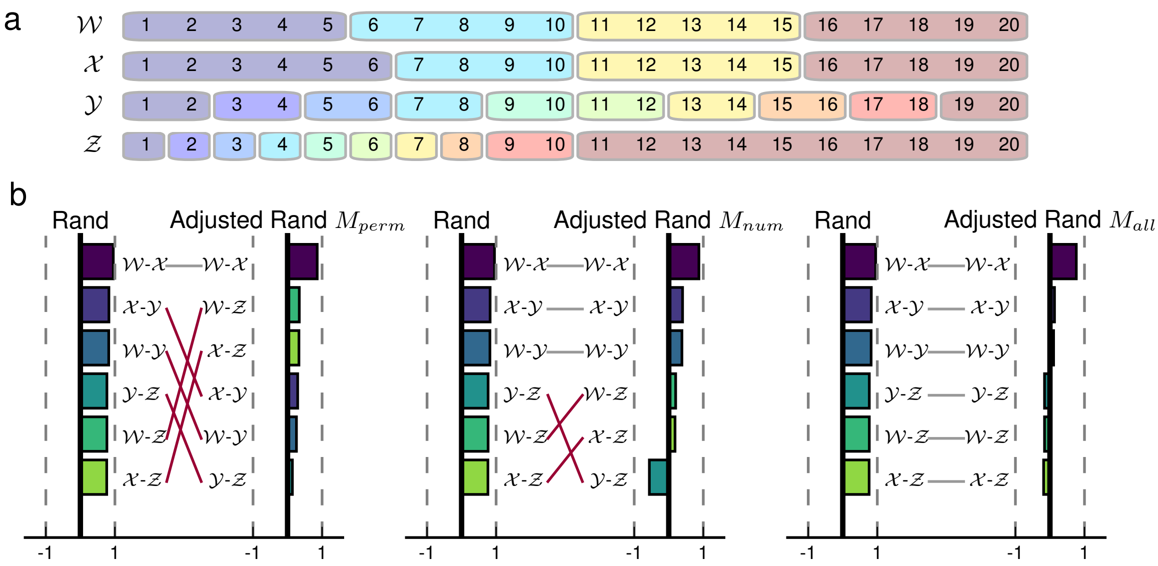


Figure 1: The choice of random model for the Rand index has a significant impact on the rankings of clustering similarity. **a**, Four clusterings of $N = 20$ elements; $\mathcal{W}$ and $\mathcal{X}$ each contain four clusters and differ by the assignment of one element (6), $\mathcal{Y}$ and $\mathcal{Z}$ each contain ten clusters. **b**, Rankings for the similarity of clustering pairs using the Rand index, the Adjusted Rand index assuming M_{perm}, the Adjusted Rand index assuming M_{num}, and the Adjusted Rand index assuming M_{all}. Rankings which change as a function of random model are highlighted in dark red.

differences in all random clusterings; the resulting adjusted index thus treats comparisons between clustering $\mathcal{Z}$ and either $\mathcal{W}$ or $\mathcal{X}$ more favorably than those to clustering $\mathcal{Y}$. On the other hand, the cluster size sequence for clustering $\mathcal{Z}$ is relatively rare in both M_{num} and M_{all}. Since clusterings $\mathcal{Y}$ and $\mathcal{Z}$ have the same number of clusters, the differences in their adjusted scores using M_{num} are a consequence of their cluster size sequences. Finally, all four clusterings are over 20 elements—the only factor that specifies the expected Rand index assuming M_{all}—so they are all adjusted by the same amount when M_{all} is used. Note that in our example, clustering $\mathcal{Z}$ has a negative adjusted Rand score using the M_{all} when compared with all three other clusterings, thus it is less similar to the other three clusterings than one would have expected from comparing two completely random clusterings.

This example illustrates an important property of the M_{all} model. Namely, the ranking provided by the Rand index remains unchanged whenever the M_{all} model is used for adjustment because all clusterings have the same number of elements. However, the corrected baseline now provides a strong interpretation for negative scores: Two randomly selected clusterings are expected to be more similar. This is an important consideration for the evaluation of clustering methods; if the derived clustering is no more similar than would be expected when comparing completely random clusterings, the solution is likely not a meaningful representation of the data.

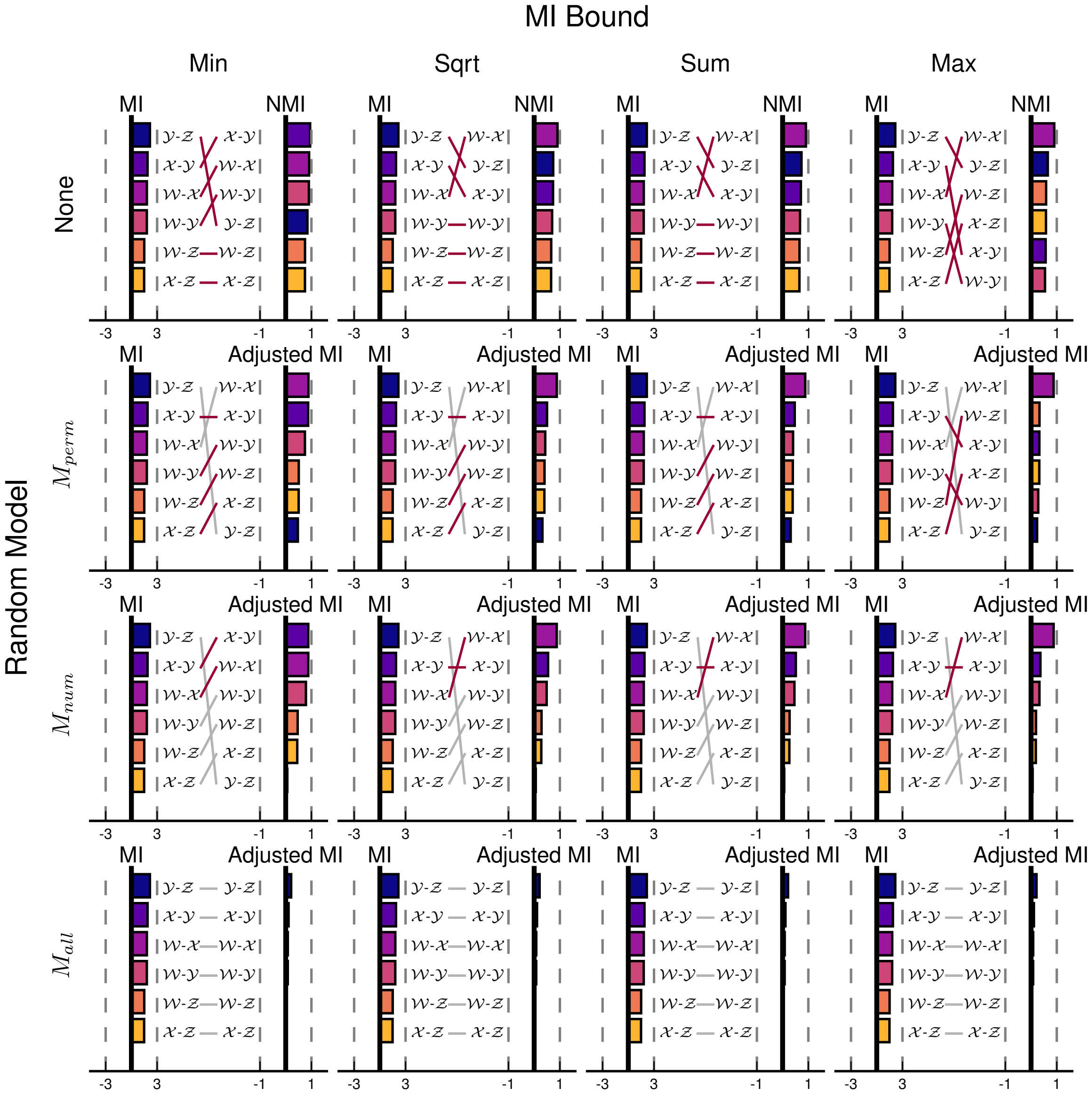


Figure 2: Both the random model and maximum bound have a significant impact on the rankings of clustering similarity using Mutual Information (MI). Rankings for the similarity of clustering pairs from Figure 1a using (vertically) the raw MI, the Adjusted MI assuming M_{perm}, the Adjusted MI assuming M_{num}, and the Adjusted MI assuming M_{all}. MI similarity also depends on the choice of maximum bound (horizontal) as a function of the two clusterings' model entropies; minimum (Min), square-root (Sqrt), average (Sum), and maximum (Max). The MI measures which are normalized but not adjusted by a random model are all members of the common family of normalized MI (NMI). For a given random model, similarity rankings which change as a function of maximum bound are highlighted in dark red.

We then turn our attention to clustering similarity measured by mutual information (MI). Rankings using the adjusted MI depend on two dimensions of variation: the random model and the maximum bound for the measure. This variation is illustrated in Figure 2 using the same 6 comparisons between pairs of clusterings from Figure 1a. We consider four cases for the MI maximum bound: Min, Sqrt, Sum, and Max, corresponding to the minimum of the two model partition entropies, the geometric mean of the two model partition entropies, the average of the two model partition entropies, and the maximum of the two model partition entropies, respectively (see Appendix 5 for details). For the permutation model, the model partition entropies are calculated from the cluster size sequences, while the model partition entropies in M_{num} are bounded by the logarithm of the number of clusters and the model partition entropies in M_{all} are bounded by the logarithm of the number of elements. As a point of reference, all rankings are illustrated in comparison to the raw mutual information score, unnormalized and without a random model adjustment (None). All adjustments of the mutual information without a random model (None, first row) are members of the commonly used family of Normalized Mutual Information (NMI) measures (Danon et al., 2005).

The rankings in Figure 2 demonstrate that both the random model and the maximum bound affect the relative similarity between clusterings when adjusting MI. Firstly, the only random model whose adjustments are independent of MI's maximum bound is M_{all}. This occurs because every choice of the maximum bound reduces to $\log N$ (the entropy of the clustering that places each element into its own cluster). In the other three random model scenarios, the maximum bound depends on the clusterings under comparison. Secondly, MI is highly dependent on the number of clusters in each of the clusterings: When either no normalization and random model adjustment are used, or the M_{all} model is used, MI ranks the similarity of clusterings $\mathcal{Y}$ and $\mathcal{Z}$ above that of $\mathcal{W}$ and $\mathcal{X}$ because of the greater number of clusters in the former case. This bias is mitigated to varying extents by the NMI variations; while NMI using the Sqrt, Sum, and Max normalization terms all produce the intuitive ranking of $\mathcal{W}$ and $\mathcal{X}$ as the most similar pair, NMI using Min for normalization still succumbs to the larger number of clusters in $\mathcal{Y}$, and ranks $\mathcal{X}$ and $\mathcal{Y}$ as the most similar clustering pair. The adjustments provided by both the M_{perm} and M_{num} random models control for the number of clusters; this reduces the impact of the number of clusters when the cluster sizes are regular, but the bias re-occurs when there is a large imbalance between the cluster sizes.

6.2 Appropriate Random Model for Comparing K-means Clusterings

Clustering similarity measures are commonly used to evaluate the results of clustering methods in relation to a known reference clustering. Since the number of clusters can vary between instances, appropriately corrected similarity measures are necessary. However, as we have already seen, the choice of similarity measure and its chance corrected variants can affect the results of the comparisons and suggest drastically different interpretations for the effectiveness of the method.

We demonstrate the importance of the random ensemble assumption through a comparison of the clusterings uncovered by 400 runs of K-means on a collection of hand-written digits (Alimoglu and Alpaydin, 1996, see Appendix B.1 for details). The K-means cluster-

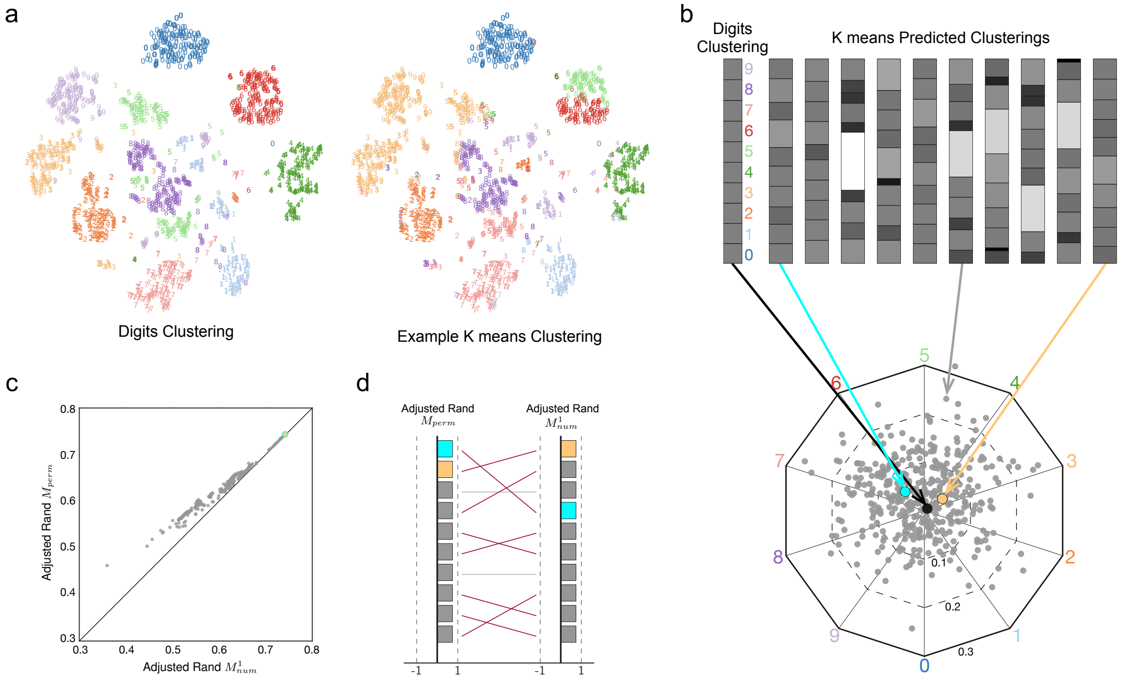


Figure 3: The impact of random model choice on the evaluation of K-means clustering. **a**, The digits data set contains $1,797$ points in 64 dimensions (projected to 2 dimensions using t-SNE dimensionality reduction for visualization, Van der Maaten and Hinton, 2008) with a ground truth clustering corresponding to the digit, and an example K-means clustering. **b**, The original cluster size sequence (top left) and 10 cluster size sequences uncovered by K-means clustering with random initialization (top right). Intensity represents cluster sizes that are smaller (darker) or larger (lighter) than the ground truth clusters. (bottom) The cluster size sequence for 400 clusterings uncovered by K-means clustering with random initialization using Barycentric coordinates. The actual clustering size sequence (black) and the most similar clustering determined by the Adjusted Rand index assuming M_{perm} (light blue) and M^1_{num} (light orange). **c**, The similarity between the actual digits clusterings and each of the 400 K-means clusterings as measured by the Adjusted Rand index assuming M_{perm} (y-axis) and M^1_{num} (x-axis). **d**, The ranking of the most similar 10 K-means clusterings as determined by the Adjusted Rand index assuming M_{perm} (left) and M^1_{num} (right).

ing method groups elements so as to minimize the average (Euclidean) distance from the cluster centroid. In most scenarios, it uncovers clusterings with a pre-specified number of clusters (K). For our example, the digits naturally fall into 10 disjoint clusters, shown in Figure 3a, with relative cluster sizes given on the left of Figure 3b. Interestingly, almost all 400 clusterings produced by K-means have a different cluster size sequence (Figure 3b, bottom) and the cluster sizes vary over a wide range (Figure 3b, top). This suggests that both the specific assignment of elements to clusters and the size sequence of the clusters are major factors differentiating the K-means clusterings. Both sources of variation need to be captured by the random model in order to have a meaningful baseline.

Since the number of clusters does not change between runs, but the size sequence of those clusters changes considerably, it is more appropriate to assess similarity within the context of random clusterings with a fixed number of clusters rather than those given by the permutation model. Furthermore, since all of the comparisons are made against the same reference clustering, a one-sided similarity metric better captures the comparison scenario. In Figure 3c, the similarity of the reference clustering compared to each of the 400 uncovered clusterings is shown using the Adjusted Rand index assuming M_{perm} and the Adjusted Rand index assuming M^1_{num}. While the measures are strongly correlated (the black line indicates perfect agreement), the Adjusted Rand index assuming M_{perm} is consistently biased towards higher similarity. Most importantly, the bulk of the uncovered clusterings change their relative ranking when considered in the context of M^1_{num} compared to M_{perm} as demonstrated by the rankings of the top 10 most similar clusterings in Figure 3d.

6.3 Random Models and Inhomogeneous Cluster Sizes

For both the Rand index and MI, the permutation model is invariant to differences in the cluster size sequence. This invariance is explicitly demonstrated in our next example by the difference between the adjusted similarity measures assuming M_{perm} and M_{num}. To generate an increasing disparity in cluster sizes, we use a preferential attachment model of element assignment. At each step of the algorithm, a random element is uniformly chosen for reassignment to a new cluster based on the current sizes of those clusters. A move is rejected if it results in an empty cluster.

In Figure 4, we compare a clustering of $1,000$ elements grouped into 50 equally sized clusters and a randomized variant of the same clustering throughout 10^6-steps of our preferential attachment algorithm using the Adjusted Rand index (Figure 4a) and Adjusted MI (Figure 4b). Cluster size inhomogeneity is measured by the entropy of the clustering size sequence; equally sized clusters have the maximum entropy ($\log_2 50 \approx 5.64$), while greater inhomogeneity in cluster sizes decreases the entropy of the cluster size sequence. In both cases, the comparisons assuming M_{perm} are invariant to the inhomogeneity of the cluster size sequences. On the other hand, comparisons assuming M_{num} reflect the changes in the cluster size sequence.

6.4 Performing at Random in Tumor Gene Expression Clustering

Finally, recall that adjusted clustering similarity measures have the added interpretation with respect to a random baseline. Such random baselines play an important role when

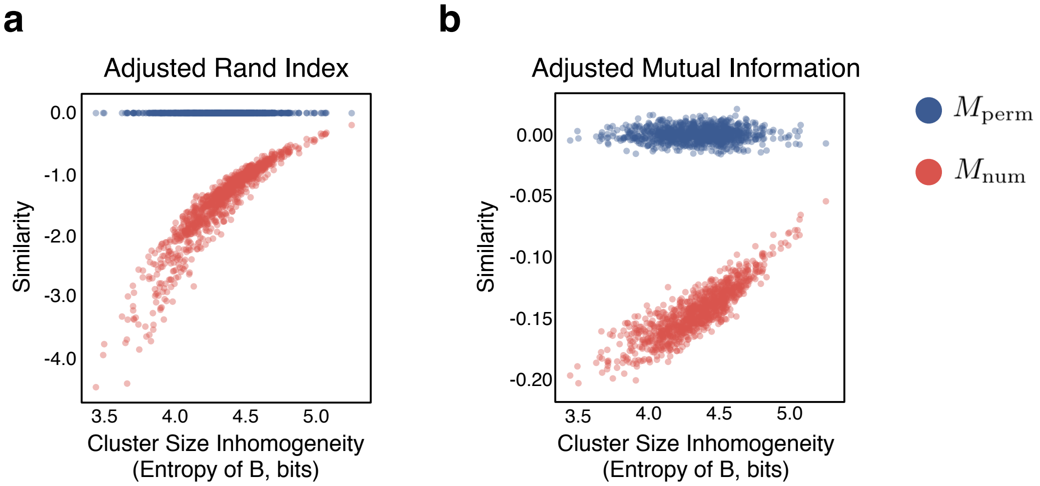
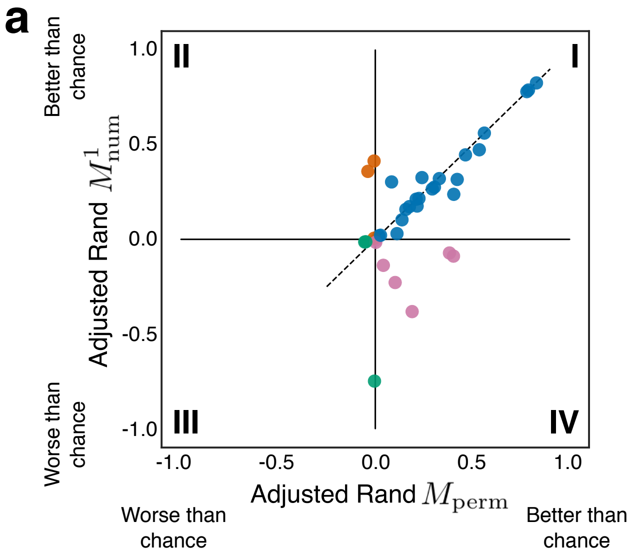


Figure 4: The invariance to inhomogeneous cluster size sequences when assuming M_{perm}. A clustering with 50 equal-sized clusters is compared to a second clustering B generated by a preferential attachment model. The cluster size sequence inhomogenity for clustering B is measured by the cluster size sequence entropy (low entropy is indicative of large cluster size inhomogenity). The similarity is calculated using the adjusted similarity assuming the permutation model M_{perm} and M_{num} for **a**, the Adjusted Rand index, and **b**, the Adjusted Mutual Information. In both cases, the similarity assuming M_{perm} is relatively constant (near 0), while the similarity assuming M_{num} increases with increasing entropy.

evaluating methods in unsupervised learning and classification. In our case, the adjusted similarity measures answer the question: *Is the result of our clustering method more similar to the desired clustering than if we selected a random clustering?* The adjusted similarity measure quantifies an answer to this question: positive scores indicate performance above random, while negative scores indicate a random clustering is more similar.

The interpretation of the adjusted similarity as a random baseline is highly dependent on the assumption of the random model. Critically, if the random model does not reflect the actual ensemble in which the clustering method is searching, the baseline does not accurately reflect the scenario in question. Thus, methods are incorrectly assessed as performing better than randomly generating a clustering.

We illustrate the dependence of adjusted similarity baseline on the choice of random model using a gene expression data set. Specifically, we use a collection of 35 cancer gene expression studies assembled in de Souto et al. (2008). The studies in the collection aim to differentiate the gene expression in cancerous cell tissue samples from those in healthy controls. Each study contains anywhere from 22 to 248 data points (individual tissue samples) for which between 85 and $4,553$ features (individual gene expression) were measured after removing the uninformative and missing genes. For details on the individual studies and filtering methodologies, see de Souto et al. (2008) and references therein.



b

vs. M_{perm}

	M^1_{num}	M_{num}	M_{all}	M^1_{all}
I	22	20	1	17
II	3	3	0	5
III	3	3	6	1
IV	7	9	28	12

Figure 5: The impact of random model choice on the evaluation of gene expression clustering with respect to the random baseline. The results of agglomerative hierarchical clustering identified from the gene expression in tissue samples from cancerous and healthy cells in 35 studies. **a**, The uncovered clusterings are compared to the reference clustering using the Adjusted Rand index assuming the permutation model M_{perm} (x-axis), and the one-sided Adjusted Rand index assuming a fixed-number of clusters M^1_{num} (y-axis). The dashed grey line indicates numerical agreement between the similarity measures. There are four possibilities when using two measures to assess similarity with respect to the random baseline: both random models conclude better than chance (blue, quadrant I), both random models conclude worse than chance (green, quadrant III), M_{perm} concludes better than chance but M^1_{num} concludes worse than chance (pink, quadrant IV), and visa-versa (orange, quadrant II). **b**, The assumed random model affects the classification of clustering comparisons with respect to the random baseline in all four random models considered here ($M^1_{\text{num}}, M_{\text{num}}, M_{\text{all}}$ and M^1_{all}) vs. M_{perm}.

Clusterings are identified via agglomerative hierarchical clustering using correlation to compute the average linkage between data points, a common clustering methodology in biology. While many other methods could be used (and indeed, were compared in de Souto et al., 2008), we use hierarchical clustering as a representative example to illustrate the consequences of the random model. Since hierarchical clustering produces a clustering with the user specified number of clusters, its similarity should be adjusted using the one-sided Adjusted Rand index assuming M^1_{num}, where the reference clustering is specified for each study individually.

Figure 5 shows the similarity between the derived clustering and the reference clustering for each of the 35 studies. The Adjusted Rand index assuming M_{perm} is shown on the x-axis; positive scores (blue and pink points) denote the method performed better than the random baseline, while negative scores (orange and green points) denote the method performed worse than the random baseline. When the Adjusted Rand index assuming M^1_{num} is used (y-axis), a different classification of method performance with respect to the random baseline is found. Of particular note are the seven studies for which the method performed better than chance according to M_{perm}, yet, M^1_{num} concludes the method actually performed worse than chance (pink points). *In this case, a random clustering drawn from the model with a fixed number of clusters would actually perform better than agglomerative hierarchical clustering, yet the practitioner using the permutation model would incorrectly conclude the method was performing better than chance.* This discrepancy occurs even when the values of the Adjusted Rand index assuming M_{perm} are relatively high (> 0.4). Similarly interesting are the three studies in which the method performed worse than chance according to M_{perm}, yet, M^1_{num} concludes the method actually performed better than chance (orange points).

7. Discussion

Given the prevalence of clustering methods for analyzing data, clustering comparison is a fundamental problem that is pertinent to numerous areas of science. In particular, the correction of clustering similarity for chance serves to establish a baseline that facilitates comparisons between different clustering solutions. Expanding previous studies on the selection of an appropriate model for random clusterings (Meila, 2005; Vinh et al., 2009; Romano et al., 2016), our work provides an extensive summary of random models and clearly demonstrates the strong impact of the random model on the interpretation of clustering results.

Our results underpin the importance of selecting the appropriate random model for a given context. To that end, we offer the following guidelines:

1. Consider what is fixed by the clustering method: do all clusterings have a user specified number of clusters (use M_{num}), or is the cluster size sequence fixed (use M_{perm})?

2. Is the comparison against a reference clustering (use a one-sided comparison), or are you comparing two derived clusterings (then use a two-sided comparison)?

The specific comparisons studied here are not meant to establish the superiority of a particular clustering identification technique or a specific random clustering model, rather, they illustrate the importance of the *choice* of the random model. Crucially, conclusions based on corrected similarity measures can change depending on the random model for clusterings. Therefore, previous studies which did promote methods based on evidence from corrected similarity measures should be re-evaluated in the context of the appropriate random model for clusterings (Yeung et al., 2001; de Souto et al., 2008; Yeung and Ruzzo, 2001; Thalamuthu et al., 2006; McNicholas and Murphy, 2010).

Throughout this work, we assumed a uniform probability of selecting a partition given a constraint on the types of partitions in the ensemble. However, other probability distributions could be used which better model the clusterings encountered in practice. For example, instead of using a uniform distribution over the number of clusters, one could consider an inferred distribution for the number of clusters actually uncovered by a given method (e.g. affinity propagation). This is particularly relevant when considering M_{all}, an extreme case for random partitions. Additionally, given that many systems exhibit clusterings with a heavy-tailed cluster size sequence, clusterings with such skewed cluster size distributions could be favored. Changes to the prior probabilities would likely change the expectations of the clustering similarity measures.

The behavior of the Rand index and Mutual Information in the context of the random clustering models discussed here further reveals problems with both measures. Specifically, the expected similarity of random clusterings increases as the number of elements grows. Intuition would suggest the opposite; the similarity of two randomly selected clusterings should decrease as the number of elements increases because it is harder to match the element memberships to clusters between two random clusterings. Instead, both MI and Rand are dominated by the fact that the expected number of clusters and cluster size distribution are converging with increasing N (Mansour, 2012). Our analysis also illustrates the dependency on the normalization term for MI, which, combined with a previously established bias on the number of clusters, suggests more care should be taken when interpreting the results of MI clustering comparisons.

In conclusion, our framework for the correction of clustering similarity for chance allows for more conscious comparisons between clusterings. The practitioner should always provide justification for their choice of random clustering model and treatment of one-sided comparisons.

Acknowledgments

We thank Ian Wood, Santosh Manicka, James Bagrow, Sune Lehmann, Aaron Clauset, Randall D. Beer, and Luis M. Rocha for helpful discussions. We also thank two anonymous reviewers for their wonderful insights and suggestions to improve the paper. A package to computer the adjusted similarity measures is available on the author's github:
`https://github.com/ajgates42/clusim`

Appendix A. Stirling and Bell Numbers

The Stirling number of the second kind $S(n, k)$ gives the number of ways to partition a set of n elements into k clusters, where

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n. \tag{40}$$

There are several recurrence relations which also give $S(n, k)$, one of the most useful is the relation

$$\begin{gathered} S(0,0) = 1 \qquad S(n,0) = S(0,n) = 0 \\ S(n+1,k) = kS(n,k) + S(n,k-1). \end{gathered} \tag{41}$$

As $n \to \infty$, an asymptotic approximation to the Stirling numbers of the second kind for a fixed k is given by $S(n,k) \approx \frac{k^n}{k!}$.

The Bell number B_n is the total number of clusterings over a set with n elements. It is related the Stirling numbers of the second kind by the summation over k for a fixed n, $B_n = \sum_{k=0}^{n} S(n,k)$. There is also a useful recurrence relation for Bell numbers: $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$. As $n \to \infty$, an asymptotic approximation to the ratio of the n-th and $(n+1)$-th Bell numbers is $\frac{B_n}{B_{n+1}} \approx \frac{\log n}{n}$. See Mansour (2012) for an extended discussion of both the Stirling numbers of the second kind and the Bell numbers.

In practice, calculating the Bell numbers and Stirling numbers of the second kind from their recurrence relations can be computationally expensive. However, many efficient approximations and implementations are available (Temme, 1993; Mansour, 2012). Here, we make use of the mpmath arbitrary precision library for Python developed by Johansson et al. (2013). This library takes advantage of Dobiǹski's Formula to approximate the Bell numbers (Dobiński, 1877; Chen and Yeh, 1994).

Appendix B. Application Data Sets

B.1 Digits Data Set

The digits data set is bundled with the sci-kit learn source code and consists of 1,797 images of 8×8 gray level pixels of handwritten digits. The reference clustering contains 10 clusters corresponding to the true digit. The data set was originally assembled in Alimoglu and Alpaydin (1996). To provide a visualization, the data was projected to 2-d using the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction method (Van der Maaten and Hinton, 2008) initialized from the pca decomposition.

B.2 Gene Expression Data Set

The data was assembled in de Souto et al. (2008) and is freely available from `http://bioinformatics.rutgers.edu/Publications/deSouto2008c/index.html`. The studies represent two prominent methods for determining gene expression in cell tissue samples from cancer tumors or healthy controls, Affymetrix microarrays and cDNA microarrays, which, respectively, measure the number of RNA copies found in the cell and the ratio of the number of copies vs a control sample. Each study contains anywhere from 22 to 248 data points (individual tissue samples) for which between 85 and 4,553 features (individual gene expression) were measured after removing the uninformative and missing genes. Please see de Souto et al. (2008) for details of this selection process.

References

Ahmed N. Albatineh and Magdalena Niewiadomska-Bugaj. Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Advances in Data Analysis and Classification*, 5(3):179–200, 2011.

Ahmed N. Albatineh, Magdalena Niewiadomska-Bugaj, and Daniel Mihalko. On similarity indices and correction for chance agreement. *Journal of Classification*, 23(2):301–313, 2006.

Fevzi Alimoglu and Ethem Alpaydin. Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition. In *Proceedings of the Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium (TAINN 96*, 1996.

Alessia Amelio and Clara Pizzuti. Is normalized mutual information a fair measure for comparing community detection methods? In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1584–1585. ACM, 2015.

Beifang Chen and Yeong-Nan N. Yeh. Some explanations of dobinski's formula. *Studies in Applied Mathematics*, 92(3):191–199, 1994.

Bobbie Chern, Persi Diaconis, Daniel M. Kane, and Robert C. Rhoades. Closed expressions for averages of set partition statistics. *Research in the Mathematical Sciences*, 1(1):1–32, 2014.

Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005 (09):P09008, 2005.

Marcilio CP de Souto, Ivan G. Costa, Daniel SA de Araujo, Teresa B. Ludermir, and Alexander Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9(1):1, 2008.

Donald G. Dobiński. Summirung der reihe $\sum n^m/n!$ für m= 1, 2, 3, 4, 5,…. *Arch. der Mat. und Physik*, 61:333–336, 1877.

Janice L. DuBien and William D. Warde. Some distributional results concerning a comparative statistic used in cluster analysis. In *ASA Proceedings of the Social Statistics Section*, pages 309–313, 1981.

Janice L. DuBien, William D. Warde, and Seong S. Chae. Moments of Rand's C statistic in cluster analysis. *Statistics & Probability Letters*, 69(3):243–252, 2004.

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

Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2 2010.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2 (1):193–218, December 1985.

Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31 (8):651–666, 2010.

Fredrik Johansson et al. *mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 0.18)*, December 2013. `http://mpmath.org/`.

Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5):056117, August 2009.

Toufik Mansour. *Combinatorics of Set Partitions.* CRC Press, 2012.

Paul D. McNicholas and Thomas Brendan Murphy. Model-based clustering of microarray expression data via latent gaussian mixture models. *Bioinformatics*, 26(21):2705–2712, 2010.

Marina Meila. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 577–584, New York, NY, USA, 2005. ACM.

Darius Pfitzner, Richard Leibbrandt, and David Powers. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, 19 (3):361–394, 2009.

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

Simone Romano, James Bailey, Vinh Nguyen, and Karin Verspoor. Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1143–1151, 2014.

Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. Adjusting for chance clustering comparison measures. *Journal of Machine Learning Research*, 17:1–32, 2016.

James Sethna. *Statistical mechanics: entropy, order parameters, and complexity*, volume 14. Oxford University Press, 2006.

D. Steinley, M. J. Brusco, and L. Hubert. The variance of the adjusted Rand index. *Psychological Methods*, 2016.

Nico M. Temme. Asymptotic estimates of stirling numbers. *Studies in Applied Mathematics*, 89(3):233–243, 1993.

Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng, and George C. Tseng. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405–2412, 2006.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML-09)*, pages 1073–1080, 2009.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

David L. Wallace. A method for comparing two hierarchical clusterings: comment. *Journal of the American Statistical Association*, 78(383):569–576, 1983.

Allan P. White and Wei Zhong Liu. Technical note: bias in information-based measures in decision tree induction. *Machine Learning*, 15(3):321–329, 1994.

Ka Yee Yeung and Walter L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

Ka Yee Yeung, Chris Fraley, Alejandro Murua, Adrian E. Raftery, and Walter L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

Pan Zhang. Evaluating accuracy of community detection using the relative normalized mutual information. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(11): P11006, 2015.