Comparing Subspace Clusterings[*]
Technical Report UW-CSE-2004-10-01

Anne Patrikainen[†] Marina Meilă[‡]

Abstract

We present the first attempt at designing a framework for comparing subspace clusterings. We propose several subspace clustering distance measures, including generalizations of well-known distance measures for ordinary clusterings. We describe a set of desirable properties for any measure for comparing subspace clusterings, as well as a systematic comparison of our proposed measures in terms of these properties. We validate the usefulness of our subspace clustering distance measures by comparing clusterings produced by the algorithms FastDOC, HARP, PROCLUS, ORCLUS, and SSPC.

keywords: subspace clustering, projected clustering, biclustering, clustering comparison, distance measure, cluster validation

1 Introduction

Clustering algorithms often use a preprocessing step of feature selection in order to overcome the curse of dimensionality in high-dimensional spaces. However, finding a single set of features to describe the whole data set is often not sufficient. Subspace clustering methods associate a distinct set of features with different groups of data points. Subspace clustering methods have rapidly gained popularity, and new algorithms are being developed at an increasing rate.

A subspace cluster can be defined as a pair (subset of data points, subspace). The data points of the cluster are similar in the associated subspace. A subspace clustering is a collection of subspace clusters.[1] The first subspace clustering algorithm CLIQUE [1] was published in 1998, being soon followed by a vast number of related methods [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. The algorithms have been applied succesfully for instance to clustering gene expression data: it is often the case that a group of genes behaves similarly only in a subset of experiments (i.e. in a subspace) [4, 6, 8, 12, 13, 14, 15]. Reviews on some of the existing subspace clustering algorithms can be found in [21, 22].

Surprisingly, despite the multitude of subspace clustering algorithms, there are no suitable methods for comparing their outputs. Assume that we have a data set X for which the true subspace clustering $\mathcal{T}$ is known. We have two algorithms, A and A' which have produced subspace clusterings $\mathcal{S}$ and $\mathcal{S}'$. We would like to know which of these clusterings is closer to the true clustering. To the best of our knowledge, nobody has proposed a method for comparing two subspace clusterings in a way that takes into account the data point groups and the subspaces simultaneously. In this paper, we will present a framework for comparing subspace clusterings by generalizing well-known distance measures for ordinary clusterings.

In the existing literature, authors most commonly compare only the grouping of data points into clusters, ignoring the similarity or dissimilarity of the associated subspaces [3, 11, 6, 21, 12]. Sometimes the subspaces are compared and the data point groups ignored [7, 22]. At best, the data points and the subspaces are compared separately [14, 2]. All these approaches fail to compare subspace clusterings in a fair manner. Many papers present extensive experiments comparing the performance of several algorithms on various data sets, hoping to prove the superiority of a given algorithm. Unfortunately, in the absence of a reliable method to compare the clustering results, these efforts are wasted.

Evaluating a clustering result by comparing it to a true clustering is often referred to as *external cluster validation* [23]. External cluster validation is a convenient tool for investigating the behavior of a clustering algorithm in controlled circumstances with synthetic data sets. However, in real clustering applications, the true clustering is not available. In these situations, we can

[*]This work was done with partial support from NSF grant No. 0313339.

[†]HIIT Basic Research Unit, Laboratory of Computer and Information Science, Helsinki University of Technology. annep@james.hut.fi. The work has been done while the author was visiting University of Washington.

[‡]Department of Statistics, University of Washington. mmp@stat.washington.edu.

[1]Other names that have been used for the same or a closely related task are projected clustering, projective clustering, biclustering, co-clustering, two-way clustering, multiway clustering, simultaneous clustering, direct clustering, block clustering, and clustering on subsets of attributes.

evaluate the quality of a clustering with *internal cluster validation* techniques. There are two main approaches to internal cluster validation: the point configuration based approach and the stability based approach [23]. The point configuration based approach (methods such as the Davies-Bouldin index or the Dunn index) is not directly applicable to subspace clusterings. In a stability based approach, a clustering is considered stable and therefore good if the clustering results for different subsets of the data are similar [24]. To measure stability, we have to compare clusterings with each other, and therefore a distance measure for subspace clusterings is needed also here. Hence, having a subspace clustering comparison distance is a basic requirement for both external and internal cluster validation.

There are various subtypes of subspace clusters and subspace clusterings. Perhaps the most common type of a subspace cluster is an *axis-aligned subspace cluster*, in which the subspace is spanned by a subset of the attributes. In this case, an equivalent representation for the cluster is a pair (subset of data points, subset of attributes). A more general type of subspace cluster is a *non-axis-aligned subspace cluster*, in which the subspaces can be arbitrarily oriented. We present distance measures for both types of subspace clusterings and also address the closely related topics of *co-clusterings* and *attribute weighted clusterings*.

This paper is organized as follows. In Section 2, we present an overview of the most popular methods for comparing ordinary clusterings. To keep the presentation simple, we start by comparing axis-aligned subspace clusterings in Section 3. In Section 3.2, we describe a set of important properties of a subspace clustering comparison distance. In Section 3.3, we introduce four candidate distance measures (CE, Rand index, VI, and RNIA), investigate their properties in detail, and compare them. We will extend our analysis to the case of non-axis-aligned subspace clusterings in Section 4 and to attribute weighted clusterings and co-clusterings in Section 5. In Section 6, we apply our distance measures to compare subspace clusterings produced by the algorithms FastDOC, HARP, PROCLUS, ORCLUS, and SSPC on synthetic data sets.

2 Comparing Ordinary Clusterings

A *clustering*[2] $\mathcal{C}$ is a partitioning of the set of m data points into disjoint clusters $C_1, C_2, \ldots, C_K$ of sizes $m_1, m_2, \ldots, m_K$, where $\sum_i m_i = m$. Virtually all criteria for comparing clusterings are based on the so-called *confusion matrix*. Assume that we have two clusterings $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ and $\mathcal{C}' = \{C'_1, C'_2, \ldots, C'_{K'}\}$. The confusion matrix $M = (m_{ij})$ is a $K \times K'$ matrix whose ijth element is the number of points in the intersection of clusters C_i and C'_j, i.e. $m_{ij} = |C_i \cap C'_j|$.

An intuitive way to compare clusterings is to calculate the *clustering error*. It is the proportion of points which are clustered differently in $\mathcal{C}$ and $\mathcal{C}'$ after an optimal matching of clusters. In other words, it is the scaled sum of the non-diagonal elements of the confusion matrix, minimized over all possible permutations of rows and columns. In practise, we do not need to try out all possible permutations; clustering error can be computed efficiently by the Hungarian method [25]. Clustering error is a metric.

An important class of criteria for comparing clusterings is based on counting the pairs of points on which two clusterings agree/disagree. Each pair of data points falls in one of the four categories labeled as N_{11}, N_{10}, N_{01}, and N_{00}. The category N_{11} contains the point pairs that are in the same cluster in both $\mathcal{C}$ and $\mathcal{C}'$. The category N_{10} contains the point pairs that are in the same cluster in $\mathcal{C}$ but not in $\mathcal{C}'$. The definitions of N_{01} and N_{00} are similar. All four counts can be obtained from the confusion matrix; for instance, $2N_{11} = \sum_{i,j} m_{ij}^2 - m$.

The best-known clustering distances based on point pair counts are the Wallace indices, the Fowlkes-Mallows index, the Rand index, the Jacard index, and the Mirkin metric [26]. For instance, the Rand index is the proportion of point pairs on which the two clusterings agree, given as

$$\text{Rand}(\mathcal{C}, \mathcal{C}') = \frac{N_{11} + N_{00}}{N}, \tag{2.1}$$

where $N = N_{11} + N_{10} + N_{01} + N_{00}$, the total number of point pairs. Out of the methods mentioned, the quantity (1 - Rand index) and the Mirkin measure are metrics.

Variation of information (VI) is a recently proposed clustering criterion based on information theoretic concepts [26]. It measures the amount of information that we gain and lose when going from clustering $\mathcal{C}$ to another clustering $\mathcal{C}'$. It is defined as $VI(\mathcal{C}, \mathcal{C}') = \text{H}(\mathcal{C}|\mathcal{C}') + \text{H}(\mathcal{C}'|\mathcal{C})$, where $\text{H}(\mathcal{C}|\mathcal{C}')$ is the conditional entropy of $\mathcal{C}$ given $\mathcal{C}'$. An equivalent way of writing the VI distance is

$$\text{VI}(\mathcal{C}, \mathcal{C}') = \frac{1}{m} \sum_{i=1}^{K} \sum_{j=1}^{K'} m_{ij} \log \frac{m_i m'_j}{m_{ij}^2}. \tag{2.2}$$

The VI distance is also a metric.

[2] In this paper we only discuss hard clusterings.

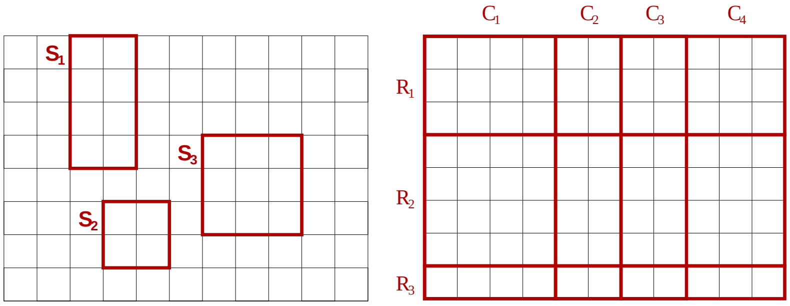

Figure 1: (Left) An example of an axis-aligned subspace clustering $\mathcal{S} = \{S_1, S_2, S_3\}$. The clusters are submatrices of the data matrix. In general, visualization of a subspace clustering is more difficult: It is often not possible to find a permutation of the rows and the columns such that the visual structure of the submatrices is preserved. (Right) An example of a co-clustering $\mathcal{S} = (\{R_1, R_2, R_3\}, \{C_1, C_2, C_3, C_4\})$.

3 Comparing Axis-Aligned Subspace Clusterings.

For the sake of simplicity, we will start by considering only axis-aligned subspace clusterings, the simplest and most popular type of subspace clusterings. We will extend our analysis to non-axis-aligned subspace clusterings in Section 4 and discuss other related types of clusterings in Section 5.

3.1 Definitions. Let us start by introducing our notation. Our data matrix $X = (x_{ij})$ has N_R rows $\{r_1, r_2, \ldots, r_{N_R}\}$ and N_C columns $\{c_1, c_2, \ldots, c_{N_C}\}$. The rows can be interpreted as data points and the columns as attributes.

An *axis-aligned subspace cluster* S is a pair (R, C), where $R \subset \{r_1, r_2, \ldots, r_{N_R}\}$ is a subset of the rows and $C \subset \{c_1, c_2, \ldots, c_{N_C}\}$ is a subset of the columns. An axis-aligned subspace cluster can be visualized as a submatrix of the data matrix X (see Fig. 2 for an example). We do not consider a single element of the data matrix as a cluster. A *axis-aligned subspace clustering* $\mathcal{S}$ is a collection of K subspace clusters $\{S_1, S_2, \ldots, S_K\}$. Examples of axis-aligned subspace clustering algorithms include CLIQUE [1], ENCLUS [27], MAFIA [28], PROCLUS [2], DOC [11], SUBCLU [5], SAMBA [12], and the algorithm by Cheng&Church [14]. Axis-aligned subspace clustering is a generalization of feature selection; instead of having a single set of features (attributes) for the whole data, we have a distinct feature set for each cluster.

In order to construct a clustering distance measure for subspace clusterings, we need to define the intersection of subspace clusters and clusterings. To this end, we define the *support* of a cluster S_k as the set of matrix elements in it, given as $\operatorname{supp} S_k = \{x_{ij} | r_i \in R_k \wedge c_j \in C_k\}$. The support of a clustering $\mathcal{S}$ is $\operatorname{supp} \mathcal{S} = \bigcup_k \operatorname{supp} S_k$. We denote the union of two clusterings $\mathcal{S}$ and $\mathcal{S}'$ by $U = \operatorname{supp} \mathcal{S} \cup \operatorname{supp} \mathcal{S}'$ and the intersection by $I = \operatorname{supp} \mathcal{S} \cap \operatorname{supp} \mathcal{S}'$.

To simplify the analysis, we will start by considering only *disjoint subspace clusterings*, i.e. clusterings in which the clusters are disjoint in the sense that they do not share matrix elements: $\operatorname{supp} S_k \cap \operatorname{supp} S_l = \emptyset \quad \forall k, l \in \{1, \ldots, K\}, \quad k \neq l$. It is important to note that any given data point or an attribute may still belong to multiple clusters. A large number of subspace clustering algorithms are guaranteed to produce disjoint clusterings; exceptions include algorithms of the type of CLIQUE [1] and SUBCLU [5]. We will discuss the extension of the analysis to non-disjoint subspace clusterings in Section 3.4.

3.2 Properties of a Distance Measure. Comparing subspace clusterings differs greatly from comparing ordinary clusterings, since subspace clusterings are not partitions of the data matrix. We will therefore start by introducing a set of properties for describing a subspace clustering comparison distance $d(\mathcal{S}, \mathcal{S}')$. We will use these properties to characterize and compare different distance measure candidates.

Metric. A good distance measure would satisfy the three metric axioms: positivity, symmetricity, and triangle inequality.

Positive. $d(\mathcal{S}, \mathcal{S}') \geq 0$ and $d(\mathcal{S}, \mathcal{S}') = 0 \Leftrightarrow \mathcal{S} = \mathcal{S}' \quad \forall \mathcal{S}, \mathcal{S}'$.

Symmetric. $d(\mathcal{S}, \mathcal{S}') = d(\mathcal{S}', \mathcal{S}) \quad \forall \mathcal{S}, \mathcal{S}'$.

Triangle inequality. $d(\mathcal{S}, \mathcal{S}') \leq d(\mathcal{S}, \mathcal{S}'') + d(\mathcal{S}'', \mathcal{S}') \quad \forall \mathcal{S}, \mathcal{S}', \mathcal{S}''$.

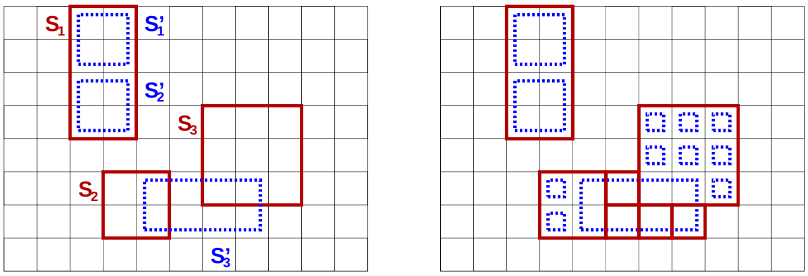

Figure 2: (Left) Two axis-aligned subspace clusterings which we wish to compare. The solid rectangles depict the clustering $\mathcal{S} = \{S_1, S_2, S_3\}$, and the dashed rectangles the clustering $\mathcal{S}' = \{S'_1, S'_2, S'_3\}$. For these clusterings, $\mathrm{CE}(\mathcal{S}, \mathcal{S}') = 19/25$ and $\mathrm{RNIA}(\mathcal{S}, \mathcal{S}') = 13/25$. (Right) To create a partition of the matrix elements, we have filled the non-intersecting areas with singleton clusters $S_4, \ldots, S_7$ and $S'_4, \ldots, S'_{12}$. This allows us to compute $\mathrm{VI}(\mathcal{S}, \mathcal{S}') = 1.68$ and $1 - \mathrm{Rand}(\mathcal{S}, \mathcal{S}') = 82/300$.

Label permutation invariant. Permuting the cluster labels of $\mathcal{S}$ and/or $\mathcal{S}'$ should not affect $d(\mathcal{S}, \mathcal{S}')$.

Penalty for non-intersecting area. If the non-intersecting area $U - I$ of the two clusterings grows while everything else remains unchanged, $d(\mathcal{S}, \mathcal{S}')$ should increase.

Background independent. The distance $d(\mathcal{S}, \mathcal{S}')$ should not be affected by the size of the data matrix X, only the size of the union U of the two clusterings. For instance, increasing the size of X by adding noise rows and columns should not move the clusterings $\mathcal{S}$ and $\mathcal{S}'$ closer to each other.

Scale invariant. Consider scaling a subspace clustering $\mathcal{S}$ by a constant c, or in other words, scaling all areas by c. Denote the scaled clustering by $c\mathcal{S}$. For a good distance measure d it would hold that $d(c\mathcal{S}, c\mathcal{S}') = d(\mathcal{S}, \mathcal{S}')$.

Copy invariant. Consider introducing a disjoint copy of a clustering pair $(\mathcal{S}, \mathcal{S}')$ in a large data matrix. The distance between these 'double clusterings' might or might not be equal to the original distance between the 'single clusterings'.

Requires partitioning. Computing $d(\mathcal{S}, \mathcal{S}')$ might require that we are able to define a partition for the data matrix elements in the union of the two clusterings U. To create a partition, we might have to fill the non-intersecting area $U - I$ with extra singleton clusters (see Fig. 2 for an example).

Multiple cluster coverage penalty. Consider a case in which $\mathcal{S} = \{S_1\}$ consists only of a single cluster and $\mathcal{S}' = \{S'_1, \ldots, S'_K\}$ consists of K disjoint clusters of equal size such that $\operatorname{supp} \mathcal{S} = \operatorname{supp} \mathcal{S}'$. For an example of the case $K = 2$, see Fig. 2. We might want the distance measure to penalize for covering a single cluster by multiple clusters by being greater than zero in this case.

Generalizable. If we have a distance measure for axis-aligned subspace clusterings, it might be straightforward to generalize it to non-axis-aligned subspace clusterings and other related types of clusterings, such as attribute weighted clusterings.

Handles ordinary clusterings. We would like a subspace clustering distance measure $d(\mathcal{S}, \mathcal{S}')$ to be able to handle the special case in which $\mathcal{S}$ and $\mathcal{S}'$ are ordinary clusterings (partitions of the same element set).

Easy to compute. A good distance measure would be easy to implement and efficient to compute.

Intuitive and understandable. A good distance measure would be intuitive and understandable for the practitioner.

3.3 Distance Measures for Subspace Clusterings. We present a framework for comparing subspace clusterings by generalizing well-known distance measures for ordinary clusterings. We consider clustering error (CE), Rand index (as a well-known representative of the point pair counting based methods), and variation of information (VI). In addition, we will introduce a new distance measure, relative non-intersecting area (RNIA). We will define and briefly discuss each of these four distance measures below; a comprehensive comparison of their properties is presented in Table 1.

Since we are comparing subspace clusterings, we consider the set of the data matrix elements $\{x_{ij}\}$ as our base element set, instead of the set of data points (rows).

3.3.1 Clustering Error.

Consider subspace clusterings $\mathcal{S} = (S_i)$ and $\mathcal{S}' = (S'_j)$ of K and K' clusters, respectively. Since the subspace clusterings are not partitions of the data matrix elements, we cannot form a confusion matrix M as in Section 2. Instead, let us define the *cluster intersection matrix* $T = (t_{ij})$ as a $K \times K'$ matrix in which t_{ij} is the number of data matrix elements shared by the clusters S_i and S'_j. More formally, $t_{ij} = |\operatorname{supp} S_i \cap \operatorname{supp} S'_j|$.

Let us find a permutation of the cluster labels such that the sum of the diagonal elements of T is maximized. Denote this maximized sum by D_{max}. Now, we define the clustering error (CE) for subspace clusterings as

$$\mathrm{CE}(\mathcal{S}, \mathcal{S}') = \frac{|U| - D_{max}}{|U|}. \tag{3.3}$$

In the case of ordinary clusterings (partitions of the rows of the data matrix), the clustering error defined here reduces to the clustering error of Section 2.

For the two clusterings of Fig. 2, the cluster intersection matrix T is

	S'_1	S'_2	S'_3
S_1	4	4	0
S_2	0	0	2
S_3	0	0	2

We also have $|U| = 25$, $D_{max} = 6$, and thus $\mathrm{CE}(\mathcal{S}, \mathcal{S}') = 19/25$.

3.3.2 Rand Index.

The Rand index for ordinary clusterings is based on counting data point pairs, but the Rand index for subspace clusterings is based on counting matrix element pairs. Recall that we need the quantities N_{11} (the number of matrix element pairs in the same cluster in both $\mathcal{S}$ and $\mathcal{S}'$), N_{00} (the number of matrix element pairs in a different cluster in both $\mathcal{S}$ and $\mathcal{S}'$) and N (the total number of matrix element pairs) for calculating the value of the Rand index according to Eq. 2.1.

Since we want our distance measure to be background independent, we will only count matrix element pairs in U, the union of the two clusterings. Therefore $N = |U|(|U| - 1)/2$. However, to compute the values of N_{01} and N_{10}, $\mathcal{S}$ and $\mathcal{S}'$ have to be partitions of U. To this end, we will make $\mathcal{S}$ a partition by filling the non-intersecting area $U \backslash \operatorname{supp} \mathcal{S}$ with extra clusters. We will similarly convert $\mathcal{S}'$ into a partition of U.

If we considered this non-intersecting area area as a single big extra cluster, we would end up having zero distances for many disjoint clusterings[3]. The only way around this difficulty seems to be to fill the non-intersecting area with extra singleton clusters, as illustrated in Fig. 2. After this filling procedure for both clusterings, we are able to compute the values of N_{11} and N_{00} as usual, for instance with the help of the confusion matrix.

When we discuss the Rand index, we actually refer to $1 - \mathrm{Rand}(\mathcal{S}, \mathcal{S}') = (N_{01} + N_{10})/N$, since this is a proper distance measure: It assumes zero values for indentical clusterings and positive values for non-identical clusterings.

If we transform the two clusterings of Fig. 2 into partitions by adding singleton clusters, the resulting confusion matrix M is

	S'_1	S'_2	S'_3	S'_4	S'_5	S'_6	$\ldots$	S'_{12}	
S_1	4	4	0	0	0	0	$\ldots$	0	8
S_2	0	0	2	1	1	0	$\ldots$	0	4
S_3	0	0	2	0	0	1	$\ldots$	1	9
S_4	0	0	1						1
S_5	0	0	1			0			1
S_6	0	0	1						1
S_7	0	0	1						1
	4	4	8	1	1	1	$\ldots$	1	25

Based on this confusion matrix, we get $N = 300$, $N_{11} = 14$, $N_{00} = 204$, $N_{01} = 26$, $N_{10} = 56$, and $1 - \mathrm{Rand}(\mathcal{S}, \mathcal{S}') = 82/300$.

3.3.3 Variation of Information.

Like the Rand index, the Variation of Information (VI) also requires the clusterings to be partitions. For reasons similar to the case of the Rand index, we fill the non-intersecting areas of the clusterings with extra singleton clusters. After this step, we can easily compute the confusion matrix and thereby calculate the VI distance according to Eq. 2.2. For the two clusterings of Fig. 2, $\mathrm{VI}(\mathcal{S}, \mathcal{S}') = 1.68$.

3.3.4 Relative Non-Intersecting Area.

If we want to compare a subspace clustering $\mathcal{S}$ to a true clustering $\mathcal{S}'$, a simple approach would be to calculate the *precision*, *recall*, and *F-measure*, used widely in the information retrieval literature to measure the success of the retrieval task [29]. Retrieval is similar to subspace clustering in that it aims to extract a subset of the data that is alike in some respect, while the rest of the data is not assumed to be grouped in any way. Hence, a subspace clustering is like the unsupervised retrieval of several disjoint groups.

[3] For instance, consider two single-cluster clusterings $\mathcal{S} = \{S_1\}$ and $\mathcal{S}' = \{S'_1\}$.

A big drawback of these measures is that they are not symmetric. A symmetric alternative is the *relative non-intersecting area (RNIA)* of the two clusterings:

$$\text{RNIA}(\mathcal{S}, \mathcal{S}') = \frac{|U| - |I|}{|U|}. \tag{3.4}$$

For the two clusterings of Fig. 2, $\text{RNIA}(\mathcal{S}, \mathcal{S}') = 19/25$.

3.3.5 Comparing Distance Measures. Table 1 is a summary of the various properties of CE, Rand index, VI, and RNIA. Rand index and VI have some serious drawbacks: They do not satisfy all metric axioms, and they fail to penalize for the non-intersecting area in certain special cases[4]. On the other hand, CE and RNIA behave very well. The major difference between the two is that RNIA does not penalize for multiple clusters covering a single cluster. In fact, since CE requires one-to-one matching between clusters, but RNIA rewards for all overlaps, we have the following proposition.

PROPOSITION 3.1. *$CE(\mathcal{S}, \mathcal{S}') \geq RNIA(\mathcal{S}, \mathcal{S}') \quad \forall\, \mathcal{S}, \mathcal{S}'$.*

Another difference is that the subspace clustering version of CE reduces to the ordinary CE in the case of ordinary clusterings. However, RNIA cannot be used with ordinary clusterings, since it will always return zero distance for two partitions of the same data set.

Due to their desirable properties, we will focus on CE and RNIA in the rest of this article.

3.4 Comparing Non-Disjoint Clusterings. In a non-disjoint subspace clustering $\mathcal{S}$, some of the clusters share matrix elements. Comparing this kind of clusterings can be handled by duplicating certain matrix elements in a way such that the result clusterings become disjoint. The previously introduced methods can be then applied. Essentially, the element duplication procedure corresponds to re-defining the union size of two clusterings $\mathcal{S}$ and $\mathcal{S}'$ as

$$|U| = \sum_{i,j} \max(n_{ij}^{\mathcal{S}}, n_{ij}^{\mathcal{S}'}) \tag{3.5}$$

and the intersection size as

$$|I| = \sum_{i,j} \min(n_{ij}^{\mathcal{S}}, n_{ij}^{\mathcal{S}'}). \tag{3.6}$$

Here, $n_{ij}^{\mathcal{S}}$ stands for the number of clusters of clustering $\mathcal{S}$ that contain the data matrix element x_{ij}.

[4]Consider clusterings $\mathcal{S}$ and $\mathcal{S}'$. Add a few clusters of size 2 to $\mathcal{S}$, such that these new clusters are not included in $\operatorname{supp} \mathcal{S}'$. After the addition, $d(\mathcal{S}, \mathcal{S}')$ might decrease, contradicting intuition.

Plugging in these definitions for $|U|$ and $|I|$, RNIA can be computed straightforwardly using Eq. 3.4. As for CE, the cluster intersection matrix can be formed and its diagonal sum maximized as usual. After this, Eq. 3.3 can be used together with the above definition of $|U|$. The previously introduced properties of RNIA and CE are preserved in the element duplication procedure.

4 Comparing Non-Axis-Aligned Subspace Clusterings.

A *non-axis-aligned subspace cluster* S is a pair (R, W), where $R \subset \{r_1, r_2, \ldots, r_{N_R}\}$ is a subset of the data points and W is a collection of vectors $\{w_1, w_2, \ldots, w_D\}$, $w_i \in \mathbb{R}^{N_C}$. The vectors in W form a basis for an arbitrary subspace of the original N_C dimensional data space. We will use W also to denote this subspace. Naturally, an axis-aligned subspace cluster is a special case of a non-axis-aligned subspace cluster.

A *non-axis-aligned subspace clustering* $\mathcal{S}$ is a collection of K non-axis aligned subspace clusters $\{S_1, S_2, \ldots, S_K\}$. ORCLUS [3] and KSM [16] produce these clusterings; 4C [9] is also closely related. Non-axis-aligned subspace clustering is a generalization of feature extraction; instead of defining a single set of features for the whole data, we have a distinct set of features for each cluster.

Analogously to the axis-aligned case, we require that the clusters of a non-axis-aligned subspace clustering are disjoint. By this we mean that if two clusters share data points, the associated subspaces must be orthogonal. An extension to the case of non-disjoint clusterings seems possible but complicated, and we leave it for future studies.

We can compare non-axis-aligned subspace clusterings just as we compared axis-aligned ones if we have a definition for the union and intersection of two clusters. These are introduced next, by means of the principal angles between two subspaces.

4.1 Principal Angles. Let us consider two subspaces of $\mathbb{R}^n$, $\mathcal{F}$ and $\mathcal{G}$, such that $p = \dim \mathcal{F} \geq \dim \mathcal{G} = q \geq 1$. The q *principal angles* $\theta_1, \theta_2, \ldots, \theta_q \in [0, \pi/2]$ can be used to measure the similarity of the subspaces. The angles can be defined sequentially for $k = 1, 2, \ldots, q$ by

$$\cos(\theta_k) := \max_{w \in \mathcal{F}} \max_{v \in \mathcal{G}} w^T v =: w_k^T v_k \tag{4.7}$$

subject to

$$\begin{aligned} &||w|| = ||v|| = 1, \quad w^T w_i = 0, \quad v^T v_i = 0, \\ &\quad i = 1, 2, \ldots, k-1. \end{aligned} \tag{4.8}$$

The vectors $w_1, w_2, \ldots, w_q$ and $v_1, v_2, \ldots, v_q$ are referred to as the *principal vectors*.

	CE	RNIA	VI	1-Rand
Positive	✓	✓	✓	✓
Symmetric	✓	✓	✓	✓
Triangle inequality	✓	✓	–	–
Label permutation invariant	✓	✓	✓	✓
Penalty for non-intersecting area	✓	✓	–	–
Background independent	✓	✓	✓	✓
Scale invariant	✓	✓	–	–
Copy invariant	✓	✓	✓	–
Lower bound	0	0	0	0
Upper bound	1	1	$\log(U)$	1
Requires partitioning	–	–	✓	✓
Multiple cluster coverage penalty	$\frac{K-1}{K}U$	0	$\log(K)$	$\frac{U(K-1)}{K(U-1)}$
Generalizable	✓	✓	✓	–
Handles ordinary clusterings	✓	–	✓	✓
Easy to compute	✓	✓	✓	✓
Intuitive and understandable	✓	✓	✓	✓

Table 1: Subspace clustering comparison properties of clustering error (CE), relative non-intersecting area (RNIA), variation of information (VI), and 1-Rand. The proofs of the bold-face properties can be found in the Appendix; the rest of the proofs are straightforward.

Singular value decomposition (SVD) is a convenient way to compute the principal angles. Let the matrices $Q_{\mathcal{F}} \in \mathbb{R}^{n\times p}$ and $Q_{\mathcal{G}} \in \mathbb{R}^{n\times q}$ contain orthonormal bases for the subspaces $\mathcal{F}$ and $\mathcal{G}$, respectively. The SVD of $Q_{\mathcal{F}}^T Q_{\mathcal{G}}$ is

$$Y^T Q_{\mathcal{F}}^T Q_{\mathcal{G}} Z = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_q), \tag{4.9}$$

where $Y \in \mathbb{R}^{p\times p}$ and $Z \in \mathbb{R}^{q\times q}$ are orthogonal matrices. The ith singular value σ_i is just the cosine of the ith principal angle, i.e. $\sigma_i = \cos(\theta_i)$. [30]

4.2 Intersection of Non-Axis-Aligned Subspace Clusters. In order to use CE, RNIA, or other distance measures with non-axis-aligned subspace clusterings, we need to define the size of a non-axis-aligned subspace cluster and the union and intersection of two such clusters. In case of axis-aligned subspace clusters, our base elements were the matrix elements, or the pairs (data point, attribute). Analogously, the base element here is (data point l, basis vector w). We count the number of these base elements in a cluster $S_i = (R_i, W_i)$, whose size is naturally $m_i = |R_i| \cdot \dim(W_i)$. Continuing with the analogy, we define the intersection of two clusters $S_i = (R_i, W_i) \in \mathcal{S}$ and $S'_j = (R'_j, V_j) \in \mathcal{S}'$ as $m_{ij} = |R_i \cap R'_j| \sum_{k=1}^{q} \sigma_k^2$, where $p = \dim(W_i)$, $p' = \dim(V_j)$, $q = \min(p, p')$, and $\{\sigma_k\}$ are the principal angles between W_i and V_j.

This seemingly arbitrary definition has a twofold motivation. First, it is geometrically consistent (see Theorem 4.1) and in the axis-aligned case it coincides with the previously defined m_{ij}. The second, probabilistic, motivation comes from viewing m_{ij}/m_i as the probability of seeing label j in $\mathcal{S}'$ if we randomly pick a point from cluster $S_i \in \mathcal{S}$. This view is used in the original definition of the VI distance and is implicit also in the other distance measures we have considered [26].

For any vector w denote by $\Pi_V\, w$ the projection of w on subspace V. Note that $||\Pi_V\, w||_2^2 = \cos^2(w, V)||w||_2^2$. We imagine the following sampling process: (1) pick uniformly a point $l \in R_i$, and if $l \notin R'_j$, stop with 0 successes; (2) else, pick a random orthonormal basis $\mathcal{B}_{W_i}$ in W_i; (3) for each $w \in \mathcal{B}_{W_i}$, "map w probabilistically to V_j" by counting a success with probability $\cos^2(w, V_j)$ and a failure with probability $\sin^2(w, V_j)$. It can be shown that the expected number of successes in m_i trials is equal to m_{ij} as defined above.

Before we can start computing clustering distances, we have make sure that the size of the intersection of a subspace cluster S_i with the clusters of the other clustering does not exceed the size of S_i. The following theorem shows this; due to space limitations, the proof is presented in the Appendix.

Theorem 4.1. *Assume that we have a p-dimensional subspace $\mathcal{F}$ and k orthogonal subspaces $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_k$ of dimensionalities $q_1, q_2, \ldots, q_k$. We compute the principal angles $\theta^1_{\mathcal{F},\mathcal{G}_i}, \ldots, \theta^{a_i}_{\mathcal{F},\mathcal{G}_i}$ between all subspace pairs*

$\mathcal{F}$, $\mathcal{G}_i$; *here* $a_i = \min(p, q_i)$. *It always holds that* $\sum_{i=1}^{k} \sum_{j=1}^{a_i} \cos^2(\theta^j_{\mathcal{F},\mathcal{G}_i}) \leq p$. *Equality is attained iff* $\mathcal{F}$ *admits an orthogonal decomposition over the subpaces* $\{\mathcal{G}_i\}_i$.

The size of the intersection $|I|$ of two clusterings is naturally the sum of the intersections of the cluster pairs: $|I| = \sum_{ij} m_{ij}$. The union size $|U|$ of two clusterings can be computed as $|U| = \sum_i m_i + \sum_j m_j - \sum_{ij} m_{ij}$. The cluster intersection matrix can be constructed as usual. After these definitions, we can simply use Eqs. 3.3 and 3.4 to compute CE and RNIA, respectively.

4.3 Example. Assume that we have two non-axis-aligned subspace clusterings, $S = \{S_1, S_2\}$ and $S' = \{S'_1, S'_2\}$, defined as follows.

$$\begin{array}{rcl} S_1 &=& (\{1,2,3\}, [1,1,0,0]^T/\sqrt{2}) \\ S_2 &=& (\{5,6,7\}, \{[0,0,1,0]^T, \\ && [0,1,0,1]^T/\sqrt{2}, [1,-2,0,2]^T/3\}) \\ S'_1 &=& (\{2,3,4,5\}, \{[2,0,0,1]^T/\sqrt{(5)}, [0,1,0,0]^T\}) \\ S'_2 &=& (\{5,6,7\}, \{[0,0,1,0]^T, [-1,0,0,2]^T/\sqrt{5}\}) \end{array}$$

We immediately observe that the cluster sizes are $m_1 = 3 \cdot 1 = 3$, $m_2 = 3 \cdot 3 = 9$, $m'_1 = 4 \cdot 2 = 8$, and $m'_2 = 3 \cdot 2 = 6$.

We store the orthonormal basis vectors in the matrices Q_{S_1}, Q_{S_2}, $Q_{S'_1}$, and $Q_{S'_2}$. We need these matrices for calculating the principal angles between the subspaces. For instance, the principal angle $\theta^1_{S_1,S'_1}$ is the non-zero singular value of the matrix $Q^T_{S_1} Q_{S'_1}$. The principal angles are $\theta^1_{S_1,S'_1} = 0.32$ $(18.44°)$, $\theta^1_{S_1,S'_2} = 1.25$ $(71.57°)$, $\theta^1_{S_2,S'_1} = 0$ $(0°)$, $\theta^2_{S_2,S'_1} = 0.89$ $(50.77°)$, $\theta^1_{S_2,S'_2} = 0$ $(0°)$, and $\theta^2_{S_2,S'_2} = 0.69$ $(39.23°)$.

These numbers allow us to compute the intersections between the cluster pairs. For instance, for S_1 and S'_1 we get the intersection size $|I|_{S_1,S'_1} = 2 \cdot \cos(0.32) = 1.90$. Similarly, $|I|_{S_1,S'_2} = 0$, $|I|_{S_2,S'_1} = 1.63$, and $|I|_{S_2,S'_2} = 5.31$. It follows that the union area of the two clusterings is $|U| = \sum_i m_i + \sum_j m'_j - \sum_{i,j} |I|_{S_i,S'_j} = 26 - 8.84 = 17.16$.

Finally, $\text{RNIA}(\mathcal{S}, \mathcal{S}') = (17.16 - 8.84)/17.16 = 0.48$ and $\text{CE}(\mathcal{S}, \mathcal{S}') = (17.16 - 7.21)/17.16 = 0.58$.

5 Comparing Other Types of Clusterings.

5.1 Attribute Weighted Clusterings. An *attribute weighted cluster* S is a pair $(R, \mathbf{b})$, where $R \subset \{r_1, r_2, \ldots, r_{N_R}\}$ is a subset of the data points and $\mathbf{b}$ is a vector $[b_1, b_2, \ldots, b_{N_C}]^T$, $b_i \geq 0$, $\sum_{i=1}^{N_C} b_i = 1$. The vector $\mathbf{b}$ defines an importance weight for each column (attribute), used for calculating the within-cluster distances. For instance, if the rows $r_i, r_j \in R$, we could calculate the weighted L_2 distance of the corresponding data vectors as $\left(\sum_{k=1}^{N_C} b_k (x_{ik} - x_{jk})^2\right)^{1/2}$.

An *attribute weighted clustering* $\mathcal{S}$ is a collection of K attribute weighted clusters $\{S_1, S_2, \ldots, S_K\}$. The algorithms COSA [4] and LAC [6] produce this type of clusterings. Attribute weighted clustering is a generalization of attribute (feature) weighting; instead of defining a single weight vector for the whole data set, we define a distinct weight vector for each cluster.

We can easily utilize CE and RNIA to compare attribute weighted clusterings if we first define the sizes, unions and intersections for this kind of clusterings. We define the size of the cluster $(R, \mathbf{b})$ as $|R|$, the number of data points in the cluster. To define the intersection of two clusters, we first need the intersection of two attribute weight vectors:

$$\mathbf{b_1} \cap \mathbf{b_2} = 1 - \frac{1}{2} \sum_{i=1}^{N_C} |b_{1i} - b_{2i}|. \tag{5.10}$$

Note that this could be defined in various ways; The present definition corresponds to the *variation distance* commonly used with discrete probability distributions [31].

It always holds that $0 \leq \mathbf{b_1} \cap \mathbf{b_2} \leq 1$. Now the intersection of two clusters $(R_1, \mathbf{b}_1)$ and $(R_2, \mathbf{b}_2)$ is given as $|R_1 \cap R_2| \cdot |\mathbf{b_1} \cap \mathbf{b_2}|$. To bound the sum of the intersection sizes, we again require that if two clusters of a clustering share data points, the inner product of the associated attribute weight vectors has to be zero. In practice this requirement is not necessary, since the existing attribute weighted clustering algorithms partition the set of data points, and no data point sharing therefore occurs [4, 6].

5.2 Co-Clusterings. A *co-clustering* $\mathcal{S} = (\mathcal{R}, \mathcal{C})$ is a simultaneous partitioning of the rows and the columns of the data matrix; $\mathcal{R} = \{R_1, R_2, \ldots, R_L\}$ denotes the collection of row clusters and $\mathcal{C} = \{C_1, C_2, \ldots, C_M\}$ the collection of column clusters. A co-clustering is just a special case of a subspace clustering: The whole data matrix is partitioned into subspace clusters S_{ij}, $i \in \{1, 2, \ldots, L\}$, $j \in \{1, 2, \ldots, M\}$. A co-clustering can be visualized as a grid-like partitioning of the data matrix. Co-clustering algorithms are presented for instance in [32, 8, 17].

Since co-clusterings are always partitionings of the data matrix elements, we can straightforwardly use any of the ordinary clustering distance measures of Section 2. Also, it is possible to derive relationships for the distances between two clusterings and their corresponding row and column clusterings. For instance, the following two propositions hold. The proofs are presented in the Appendix.

Proposition 5.1. $VI(\mathcal{S}, \mathcal{S}') = VI(\mathcal{R}, \mathcal{R}') + VI(\mathcal{C}, \mathcal{C}')$.

Proposition 5.2. $CE(\mathcal{S}, \mathcal{S}') = CE(\mathcal{R}, \mathcal{R}') + CE(\mathcal{C}, \mathcal{C}') - CE(\mathcal{R}, \mathcal{R}')CE(\mathcal{C}, \mathcal{C}')$.

6 Experimental Results

We now examine the performance of our distance measures on clusterings produced by a variety of subspace clustering algorithms on synthetic data sets. In our first experiment, detailed in Section 6.1, we show how the use of the subspace clustering distance measures can give more information than traditional row-based or column-based comparison approaches. In Section 6.2, we describe an experiment illustrative of how subspace clustering distance measures can be used for studying the consistency of various clusterings produced by a single algorithm on a given data set. These results indicate that our distance measures could also be useful for internal cluster validation.

6.1 Comparing Subspace Clusterings, Row Clusterings, and Column Clusterings.

Algorithms: We compare the performance of four algorithms, PROCLUS [2], FASTDOC [11], HARP [15], and ORCLUS [3] on synthetic data sets. The first three algorithms are axis-aligned subspace clustering algorithms, and ORCLUS is a non-axis-aligned one. We compare clusterings produced by these algorithms using our extended CE and RNIA distance measures; however, we only present the results for CE, since the RNIA results are very similar.

Data Sets: We use clustering results from earlier work [15] that compared the algorithms across 8 synthetic data sets. Each data set has 500 data points, 20 attributes, and 5 axis-aligned subspace clusters that form a partitioning of all data points. The number of data points in each cluster varies from 15% to 25% of the total number of points. The 8 data sets differ in the dimensionality of the subspace clusters, which varies from 4 in the first data set to 18 in the last one. In an attribute relevant to a subspace, the standard deviation of the within-cluster data is between 3% and 5% of the global standard deviation on that attribute. For each data set, we have several clustering results for each algorithm with various parameter values (except for HARP, which is deterministic and does not require input parameters). Some of these results are visualized in Fig. 3 (see figure for details).

Results: Fig. 4 shows a comparison between the algorithms on all data sets using three different distance measures; CE distance for subspace clusterings, CE distance using row information only, and CE distance using column information only[5]. Note that the column clusterings are non-disjoint, so we need to apply the element duplication procedure from Section 3.4 here.

The figure clearly brings out differences between the subspace clustering distances, the row clustering distances, and the column clustering distances, indicating that it is indeed worthwhile to pay attention to the choice of the clustering type to compare. For instance, according to the row clustering results, HARP performs well for all data sets and always gives the best result. This is somewhat misleading, since the other two distance measures reveal that HARP's choice of subspaces leaves room for improvement. Thus, irrespective of whether we wish to compare algorithms to each other or analyze the performance of a given algorithm across data sets, the row, column, or subspace based distance measures give different information.[6]

6.2 Clustering Clusterings. We conducted another experiment on a new data set consisting of 1000 rows, 100 columns, and 5 subspace clusters, generated in a way similar to the previous data sets. The goal here is to compare the performance of PROCLUS and SSPC [20] over subsamples of this data set with various parameter settings. Our non-disjoint axis-aligned subspace clusters are 10-dimensional and each cluster has approximately 200 rows. We created 5 subsamples of this data set by removing 10% of the rows and 10% of the columns. These 5 subsamples were then clustered by PROCLUS and SSPC. PROCLUS was run with 9 different parameter values for each subsample, resulting in 45 clusterings, and SSPC was run with 10 different parameter values per subsample, resulting in 50 clusterings.

We computed the subspace clustering CE distance between all 95 clusterings; the distance matrix is shown in Figure 5. The SSPC results are clustered both by parameter value (horizontal and vertical strips) and by subsample (dark diagonals). They form two subclusters, one of which is closer to the true clustering. The PROCLUS results are much farther from each other; mainly the results 52–56 corresponding to the first parameter value are relatively close to each other and to the true clustering. Using a subspace clustering distance measure has therefore provided us information on the stability of the two algorithms.

[5] Since we do not yet have a method for handling non-disjoint non-axis-aligned clusterings, the column distance measure for ORCLUS is not shown.

[6] We have chosen to plot only the best clustering results of each algorithm, since the dependence on parameter values is strong, and it does not make sense in this case to plot the means with error bars.

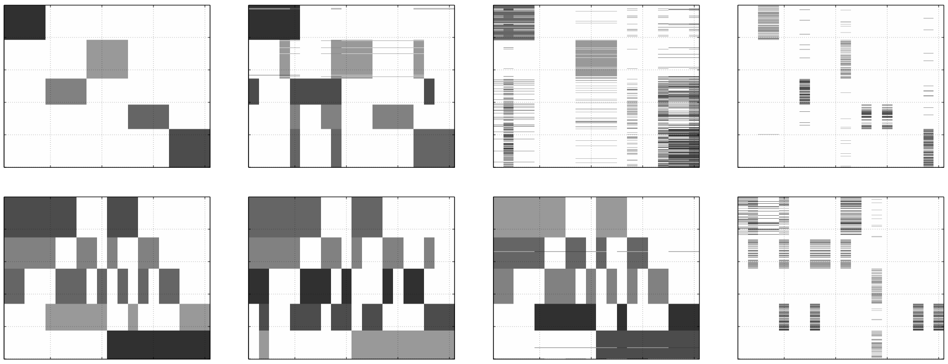

Figure 3: The columns represent the true clustering and the best clustering result from HARP, PROCLUS, and FASTDOC, respectively. Each subspace cluster is represented by a different shade of gray (the colors do not imply correspondence between clusters of different clusterings), and the unclustered background is white. The two rows are for data set 1 (5 clusters of dimensionality 4) and data set 4 (5 clusters of dimensionality 10), respectively. As can be seen, the clusterings produced by the various algorithms have significant differences, especially in the choice of subspaces. The results of ORCLUS cannot be visualized here, since ORCLUS produces non-axis-aligned subspaces.

7 Conclusions

In this paper, we have addressed the problem of comparing subspace clusterings. Subspace clusterings differ from ordinary clusterings in that they are *partial clusterings*: they are not partitions of a base element set. In addition, clusters of a subspace clustering might be *non-disjoint*: portions of the same base element might belong to several clusters. Comparing subspace clusterings is therefore a more general problem than comparing ordinary clusterings. In this paper, we have developed methods for comparing subspace clusterings; we are not aware of any existing work in this field.

We have introduced a set of properties to describe a distance measure for subspace clusterings. We have presented four distance measure candidates for subspace clusterings, namely CE, Rand, VI, and RNIA. While the first three are generalizations of well-known distance measures for ordinary clusterings, RNIA is a newly proposed measure. We have shown how to apply these distance measures to axis-aligned subspace clusterings, non-axis-aligned subspace clusterings, attribute weighted clusterings, and co-clusterings. After a careful theoretical analysis of the properties of these candidate measures, we have concluded that CE and RNIA are the best choices. The CE and RNIA measures differ in philosophy since the former is aimed at comparing groupings of data matrix elements, and the latter attempts to measure retrieval accuracy; however with our choice of algorithms and datasets the two measures showed no significant empirical difference.

A distance measure for subspace clusterings is essential in both external and internal cluster validation. Our experiments have supported this claim: We have shown comparisons between a true subspace clustering and clusterings produced by the algorithms FastDOC, HARP, PROCLUS, and ORCLUS. The results show clear qualitative differences between comparing subspace clusterings and comparing only the corresponding row clusterings or only the corresponding column clusterings. Further, we have briefly shown how to apply subspace clustering distances to clustering clusterings, useful for instance in stability experiments. A detailed study of internal cluster validation for subspace clusterings is reserved for future work.

There are several avenues to be explored in the future. We have not shown how to handle non-disjoint non-axis aligned subspace clusterings. Weighting the rows and the columns of the data matrix is another potentially useful feature. We have only discussed hard clusterings; probabilistic subspace clusterings would require separate analysis[7]. Finally, depending on the definition of the closeness of the data points in the subspace (for instance, distance-based [2] or pattern-based [13]), the rows and the columns of the data matrix may or may not be symmetric, and a successful comparison method should take this into account.

[7] However, we are only aware of one algorithms producing probabilistic subspace clusterings [7].

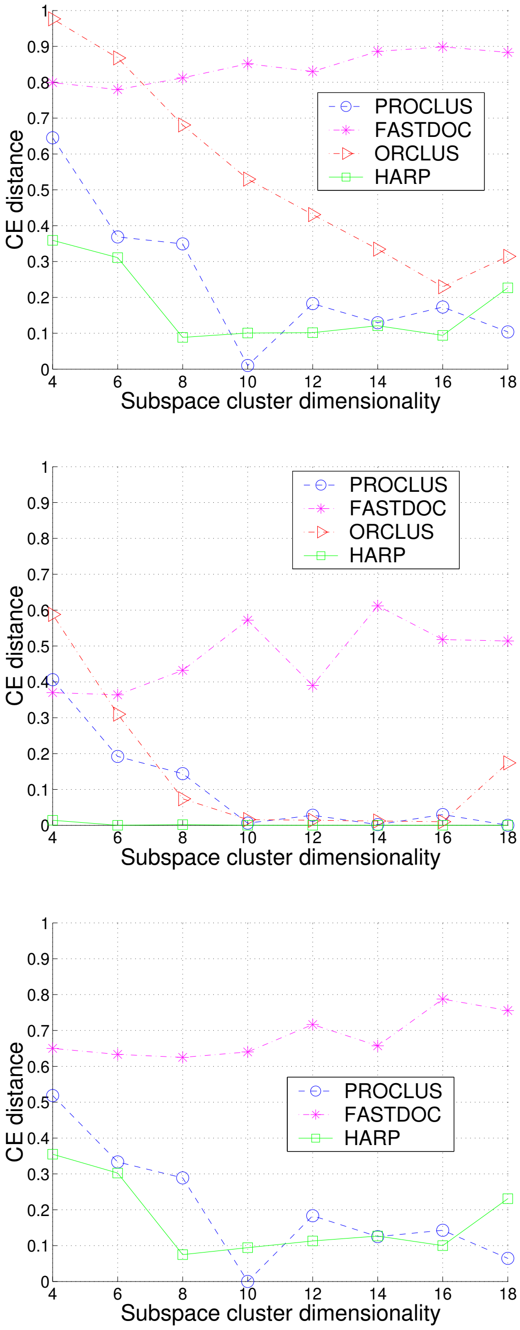

Figure 4: CE distance between PROCLUS output and the true clustering, FastDOC output and the true clustering, etc., for the eight different data sets (cluster dimensionalities $4, 6, \ldots, 18$). Only the best clustering results of each algorithm are shown. (Top) CE distances for subspace clusterings. (Middle) CE distances for the corresponding row clusterings. (Bottom) CE distances for the corresponding column clusterings.

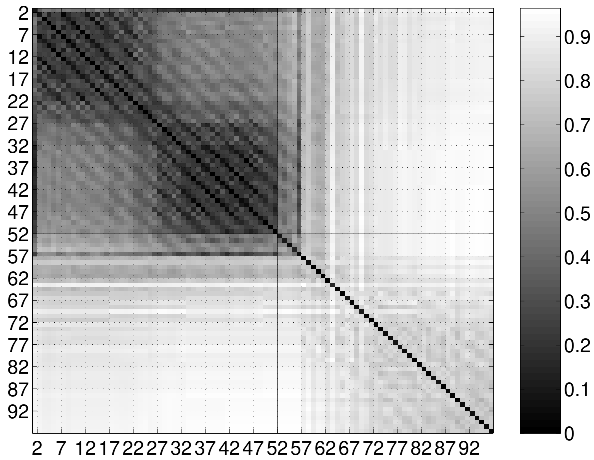

Figure 5: CE subspace clustering distance matrix for 50 clusterings by SSPC and 45 clusterings by PROCLUS. 1: True clustering. 2–51: SSPC clusterings. 52–96: PROCLUS clusterings. 2–6: SSPC clusterings with parameter value 1 for the five subsamples; 7–11: SSPC clusterings with parameter value 2 for the five subsamples; etc.

8 Acknowledgements

We are grateful to Kevin Yip from the University of Hong Kong for providing us the subspace clustering results we used in our experiments. We also wish to thank Profs. Heikki Mannila and Anna Karlin for providing Anne Patrikainen an opportunity to visit University of Washington.

References

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *ACM SIGMOD*, pages 94–105, 1998.

[2] C. Aggarwal, C. Procopiuc, J. Wolf, P. Yu, and J. Park. A framework for finding projected clusters in high dimensional spaces. In *ACM SIGMOD*, 1999.

[3] C. C. Aggarwal and P. S. Yu. Finding generalized projected clusters in high dimensional spaces. In *ACM SIGMOD*, pages 70–81, 2000.

[4] J. H. Friedman and J. J. Meulman. Clustering objects on subsets of attributes. *J. R. Statist. Soc. B*, 66:1–25, 2004.

[5] K. Kailing, H. Kriegel, and P. Kröger. Density-connected subspace clustering for high-dimensional data. In *SIAM Data Mining*, 2004.

[6] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma. Subspace clustering of high dimensional data. In *SIAM Data Mining*, 2004.

[7] A. Patrikainen and H. Mannila. Subspace clustering of high dimensional binary data — a probabilistic approach. In *SIAM Data Mining, Workshop on Clustering High Dimensional Data*, 2004.
[8] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra. Minimum sum-squared residue co-clustering of gene expression data. In *SIAM Data Mining*, 2004.
[9] C. Böhm, K. Kailing, P. Kröger, and A. Zimek. Computing clusters of correlation connected objects. In *ACM SIGMOD*, 2004.
[10] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *ACM SIGKDD*, 2003.
[11] C. M. Procopiuc, M. T. Jones, P. K. Agarwal, and T. M. Murali. A Monte carlo algorithm for fast projective clustering. In *ACM SIGMOD*, 2002.
[12] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18, 2002.
[13] J. Yang, W. Wang, H. Wang, and P. S. Yu. delta-cluster: Capturing subspace correlation in a large data set. In *ICDE*, 2002.
[14] Yizong Cheng and George M. Church. Biclustering of expression data. In *ISMB*, 2000.
[15] K. Y. Yip, D. W. Cheung, and M. K. Ng. HARP: A practical projected clustering algorithm. IEEE TKDE (To appear).
[16] P. K. Agarwal and N. H. Mustafa. k-means projective clustering. In *PODS*, 2004.
[17] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *KDD*, 2004.
[18] A. A. Melkman and E. Shaham. Sleeved coclustering. In *KDD*, 2004.
[19] J. Liu, W. Wang, and J. Yang. A framework for ontology-driven subspace clustering. In *KDD*, 2004.
[20] K. Y. Yip, D. W. Cheung, and M. K. Ng. On discovery of extremely low-dimensional clusters using semi-supervised projected clustering. In *ICDE*, 2005.
[21] Kevin Yip, Michael Ng, and David Cheung. A review on projected clustering algorithms. *International Journal of Applied Mathematics*, 13:35–47, 2003.
[22] L. Parsons, E. Haque, and H. Liu. Evaluating subspace clustering algorithms. In *SIAM Data Mining, Workshop on Clustering High Dimensional Data*, 2004.
[23] V. Roth, T. Lange, M. Braun, and J. Buhmann. A resampling approach to cluster validation. In *COMPSTAT*, 2002.
[24] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16:1299–1323, 2004.
[25] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization, Algorithms and Complexity*. Prentice-Hall, 1982.
[26] M. Meila. Comparing clusterings by the variation of information. In *COLT*, 2003.
[27] C. H. Cheng, A. W. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In *Knowledge Discovery and Data Mining*, pages 84–93, 1999.
[28] H. Nagesh, S. Goil, and A. Choudhary. Mafia: Efficient and scalable subspace clustering for very large data sets, 1999.
[29] C. Van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.
[30] M. E. Argentati. Principal angles between subspaces. `http://www-math.cudenver.edu/~aknyazev/teaching/rico/talk_defense.pdf`.
[31] D. Aldous and J. Fill. Reversible Markov Chains and random walks on graphs. `http://www.stat.berkeley.edu/users/aldous/RWG/book.html`.
[32] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *KDD*, 2003.

9 Appendix 1: Proofs for Table 1

9.1 Preliminaries. In this section, we review our notation and derive new formulas for VI and the Rand index. For CE and RNIA, we use the familiar expressions given in Eqs. 3.3 and 3.4, respectively.

9.1.1 Cluster Sizes and Intersections. Consider two subspace clusterings $\mathcal{S} = (S_1, \ldots, S_K)$ and $\mathcal{S}' = (S'_1, \ldots, S'_{K'})$. The total number of matrix elements in the union of the two clusters is $m = |U|$. The cluster sizes are $m_1, \ldots, m_K$ and $m'_1, \ldots, m'_{K'}$. We write m_{ij} for the cluster intersection $|S_i \cap S'_j|$.

9.1.2 VI. To compute the VI distance for two subspace clusterings, we have used Eq. 2.2 after filling the non-intersecting area of these clusterings with singleton clusters. We now show how to derive a simplified version of the VI distance for subspace clusterings.

Consider two axis-aligned subspace clusterings $\mathcal{S} = \{S_1, S_2, \ldots, S_K\}$ and $\mathcal{S}' = \{S'_1, S'_2, \ldots, S'_{K'}\}$. We fill the non-intersecting areas of these clusterings with singleton clusters. The confusion matrix of the resulting clusterings is shown in Fig. 6.

In the confusion matrix, $m_{0i} = m_i - \sum_{j=1}^{K'} m_{ij}$, i.e. the number of matrix elements in cluster C_i which are not covered by any cluster of $\mathcal{S}'$. We have written C_{0i} as a shorthand for all resulting unit clusters. Similarly, $m'_{0j} = m_j - \sum_{i=1}^{K} m_{ij}$.

We calculate the VI score of Eq.2.2 for this confusion matrix as follows.

$$\begin{aligned}
&\mathrm{VI}(\mathcal{S}, \mathcal{S}') \\
&= \tfrac{1}{|U|} \Big[\textstyle\sum_{i=1}^{K} \sum_{j=1}^{K'} m_{ij} \log \tfrac{m_i m'_j}{m_{ij}^2} \\
&+ \textstyle\sum_{i=1}^{K} m_{i0} \log m_i + \sum_{j=1}^{K'} m'_{j0} \log m'_j \Big] \\
&= \tfrac{1}{|U|} \Big[\textstyle\sum_{i=1}^{K} \sum_{j=1}^{K'} m_{ij} \log \tfrac{m_i m'_j}{m_{ij}^2} \\
&+ \textstyle\sum_{i=1}^{K} m_i \log m_i + \sum_{j=1}^{K'} m'_j \log m'_j \\
&- \textstyle\sum_{i=1}^{K} \sum_{j=1}^{K'} m_{ij} \log m_i \\
&- \textstyle\sum_{j=1}^{K'} \sum_{i=1}^{K} m'_{ij} \log m'_i \Big].
\end{aligned}$$

Our expression does not contain any terms related to the unit clusters. We simplify it further, obtaining

$$\begin{aligned}
&\mathrm{VI}(\mathcal{S}, \mathcal{S}') \\
&= \tfrac{1}{|U|} \Big(\textstyle\sum_{i=1}^{K} m_i \log m_i + \sum_{j=1}^{K'} m'_j \log m'_j \\
&-2 \textstyle\sum_{i=1}^{K} \sum_{j=1}^{K'} m_{ij} \log m_{ij} \Big).
\end{aligned} \tag{9.11}$$

Note that all we need to compute the VI distance are the cluster size vectors $\mathbf{m} = (m_i)$ and $\mathbf{m}' = (m'_i)$ together with the *cluster intersection matrix* $T = (m_{ij})$. An essential difference between T and the confusion matrix is that the cluster intersection matrix contains the intersections of the original clusters only, whereas the confusion matrix contains also the unit clusters. It is also important to note that $\mathbf{m}$ and $\mathbf{m}'$ are not marginals of the cluster intersection matrix. Only in the special case of fully intersecting clusters we have $m_i = \sum_{j=1}^{K'} m_{ij}$ and $m'_j = \sum_{i=1}^{K} m_{ij}$. In this case, we recover the original VI.

9.1.3 Rand. Using the confusion matrix, the quantity (1-Rand) can be written as

(9.12)

$$\begin{aligned}
&1 - \mathrm{Rand}(\mathcal{S}, \mathcal{S}') \\
&= \tfrac{N_{10}+N_{01}}{N} \\
&= \tfrac{\sum_{i=1}^{L'} \sum_{j=1}^{L} \sum_{k=j+1}^{L} m_{ij} m_{ik} + \sum_{i=1}^{L} \sum_{j=1}^{L'} \sum_{k=j+1}^{L'} m_{ji} m_{ki}}{1/2m(m-1)}.
\end{aligned}$$

Another way to write (1-Rand) is

(9.13)

$$1 - \mathrm{Rand}(\mathcal{S}, \mathcal{S}') = \tfrac{\sum_{i=1}^{L} m_i^2 + \sum_{j=1}^{L'} m_j'^2 - 2\sum_{i=1}^{L} \sum_{j=1}^{L'} m_{ij}^2}{m(m-1)}.$$

In these two equations, $L \geq K$ and $L' \geq K'$ stand for the number of clusters in $\mathcal{S}$ and $\mathcal{S}'$ after filling the non-intersecting area with singleton clusters.

9.2 Triangle Inequality

THEOREM 9.1. *(Triangle inequality for CE.) $CE(\mathcal{A}, \mathcal{B}) \leq CE(\mathcal{A}, \mathcal{C}) + CE(\mathcal{B}, \mathcal{C})$ for any subspace clusterings $\mathcal{A}, \mathcal{B}, \mathcal{C}$.*

PROPOSITION 9.1. *If $\mathcal{A}$ and $\mathcal{B}$ are arbitrary and $\mathcal{C}' \subseteq (\mathcal{A} \cup \mathcal{B})$, then $CE(\mathcal{A}, \mathcal{B}) \leq CE(\mathcal{A}, \mathcal{C}') + CE(\mathcal{B}, \mathcal{C}')$.*

Proof. (Proposition 9.1) For simplicity, we adopt a shorthand notation and write $\mathcal{A}$ instead of $\operatorname{supp} \mathcal{A}$, $\mathcal{A} \backslash \mathcal{B}$ instead of $\operatorname{supp} \mathcal{A} \setminus \operatorname{supp} \mathcal{B}$, etc. Also, we write $\mathcal{A}_{\mathcal{A} \cap \mathcal{B}}$ for the part of the clustering $\mathcal{A}$ in $\mathcal{A} \cap \mathcal{B}$, $\mathcal{A}_{\mathcal{A} \backslash \mathcal{B}}$ for the part of the clustering $\mathcal{A}$ in $\mathcal{A} \backslash \mathcal{B}$, etc. Let us write $\mathrm{H}(u, v)$ for the Hamming distance between two cluster label vectors u and v, or in other words, the total number of differences these two vectors. Then the subspace clustering CE distance becomes

$$\mathrm{CE}(\mathcal{A}, \mathcal{B}) = \frac{\min \mathrm{H}(\mathcal{A}_{\mathcal{A} \cap \mathcal{B}}, \mathcal{B}_{\mathcal{A} \cap \mathcal{B}}) + |\mathcal{A} \backslash \mathcal{B}| + |\mathcal{B} \backslash \mathcal{A}|}{|\mathcal{A} \cup \mathcal{B}|},$$

where the minimum is taken over all permutations of the cluster labels. Let us consider three subspace clusterings $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}'$ such that $\mathcal{A}$ and $\mathcal{B}$ are arbitrary and $\mathcal{C}' \subseteq (\mathcal{A} \cup \mathcal{B})$. We consider $\mathcal{C}'$ in three disjoint parts: $\mathcal{C}' = \mathcal{C}'_{\mathcal{A} \cap \mathcal{B}} \cup \mathcal{C}'_{\mathcal{A} \backslash \mathcal{B}} \cup \mathcal{C}'_{\mathcal{B} \backslash \mathcal{A}}$. Let us fix the cluster labels of $\mathcal{C}'$ and choose the permutation of labels in $\mathcal{A}$ to minimize $\mathrm{H}(\mathcal{A}_{\mathcal{A} \cap \mathcal{C}'}, \mathcal{C}'_{\mathcal{A} \cap \mathcal{C}'})$ and the permutation of labels in $\mathcal{B}$ to minimize $\mathrm{H}(\mathcal{B}_{\mathcal{B} \cap \mathcal{C}'}, \mathcal{C}'_{\mathcal{B} \cap \mathcal{C}'})$. Using these

$$
\begin{array}{r|cccccccc|c}
 & S'_1 & S'_2 & \dots & S'_{K'} & \overbrace{S_{01}}^{m_{01}} & \overbrace{S_{02}}^{m_{02}} & \dots & \overbrace{S_{0k}}^{m_{0K}} & \\
\hline
S_1 & m_{11} & m_{12} & \dots & m_{1K'} & 1 & 0 & \dots & 0 & m_1 \\
S_2 & m_{21} & m_{22} & \dots & m_{2K'} & 0 & 1 & \dots & 0 & m_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
S_K & m_{K1} & m_{K2} & \dots & m_{KK'} & 0 & 0 & \dots & 1 & m_{K'} \\
m'_{01}\left\{ S'_{01}\right. & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 \\
m'_{02}\left\{ S'_{02}\right. & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & 1 \\
m'_{0K'}\left\{ S'_{0K'}\right. & 0 & 0 & \dots & 1 & 0 & 0 & 0 & 0 & 1 \\
\hline
 & m'_1 & m'_2 & \dots & m'_K & 1 & 1 & 1 & 1 & |U|
\end{array}
$$

Figure 6: Confusion matrix for two subspace clusterings $\mathcal{S} = \{S_1, S_2, \dots, S_K\}$ and $\mathcal{S}' = \{S'_1, S'_2, \dots, S'_{K'}\}$ with the non-itersecting area filled with singleton clusters.

labels, we have

$$
\begin{array}{ll}
 & \mathrm{CE}(\mathcal{A},\mathcal{C}') + \mathrm{CE}(\mathcal{B},\mathcal{C}') - \mathrm{CE}(\mathcal{A},\mathcal{B}) \\
\geq & \frac{\mathrm{H}(\mathcal{A}_{C'_{\mathcal{A}\cap\mathcal{B}}},\mathcal{C}'_{A\cap B})+\mathrm{H}(\mathcal{A}_{C'_{\mathcal{A}\setminus\mathcal{B}}},\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}})}{|\mathcal{A}\cup\mathcal{C}'|} \\
+ & \frac{|\mathcal{A}|-|\mathcal{C}'_{\mathcal{A}\cap\mathcal{B}}|-|\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}}|+|\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}}|}{|\mathcal{A}\cup\mathcal{C}'|} \\
+ & \frac{\mathrm{H}(\mathcal{B}_{C'_{\mathcal{A}\cap\mathcal{B}}},\mathcal{C}'_{\mathcal{A}\cap\mathcal{B}})+\mathrm{H}(\mathcal{B}_{C'_{\mathcal{B}\setminus\mathcal{A}}},\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}})}{|\mathcal{B}\cup\mathcal{C}'|} \\
+ & \frac{|\mathcal{B}|-|\mathcal{C}'_{\mathcal{A}\cap\mathcal{B}}|-|\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}}|+|\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}}|}{|\mathcal{B}\cup\mathcal{C}'|} \\
- & \frac{\mathrm{H}(\mathcal{A}_{\mathcal{A}\cap\mathcal{B}},\mathcal{B}_{\mathcal{A}\cap\mathcal{B}})+|\mathcal{A}\setminus\mathcal{B}|+|\mathcal{B}\setminus\mathcal{A}|}{|\mathcal{A}\cup\mathcal{B}|} \\
\geq & \frac{\mathrm{H}(\mathcal{A}_{C'_{\mathcal{A}\cap\mathcal{B}}},\mathcal{C}'_{A\cap B})}{|\mathcal{A}\cup\mathcal{C}'|} + \frac{\mathrm{H}(\mathcal{B}_{C'_{\mathcal{A}\cap\mathcal{B}}},\mathcal{C}'_{A\cap B})}{|\mathcal{B}\cup\mathcal{C}'|} \\
- & \frac{\mathrm{H}(\mathcal{A}_{C'_{\mathcal{A}\cap\mathcal{B}}},\mathcal{B}_{C'_{A\cap B}})}{|\mathcal{A}\cup\mathcal{B}|} \\
+ & \frac{\mathrm{H}(\mathcal{A}_{C'_{\mathcal{A}\setminus\mathcal{B}}},\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}})}{|\mathcal{A}\cup\mathcal{C}'|} + \frac{\mathrm{H}(\mathcal{B}_{C'_{\mathcal{B}\setminus\mathcal{A}}},\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}})}{|\mathcal{B}\cup\mathcal{C}'|} \\
+ & \frac{(|\mathcal{A}\cap\mathcal{B}|-|\mathcal{C}'_{\mathcal{A}\cap\mathcal{B}}|)+(|\mathcal{A}\setminus\mathcal{B}|-|\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}}|)+|\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}}|}{|\mathcal{A}\cup\mathcal{C}'|} \\
+ & \frac{(|\mathcal{A}\cap\mathcal{B}|-|\mathcal{C}'_{\mathcal{A}\cap\mathcal{B}}|)+(|\mathcal{B}\setminus\mathcal{A}|-|\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}}|)+|\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}}|}{|\mathcal{B}\cup\mathcal{C}'|} \\
- & \frac{|\mathcal{A}\setminus\mathcal{B}|+|\mathcal{B}\setminus\mathcal{A}|+(|\mathcal{A}\cap\mathcal{B}|-|C'_{\mathcal{A}\cap\mathcal{B}}|)}{|\mathcal{A}\cup\mathcal{B}|} \\
\geq & \frac{(|\mathcal{A}\setminus B|-|\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}}|)+|\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}}|}{|\mathcal{A}\cup\mathcal{C}'|} + \frac{(|\mathcal{B}\setminus\mathcal{A}|-|\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}}|)+|\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}}|}{|\mathcal{B}\cup\mathcal{C}'|} \\
- & \frac{|\mathcal{A}\setminus\mathcal{B}|+|\mathcal{B}\setminus\mathcal{A}|}{|\mathcal{A}\cup\mathcal{B}|} \\
\geq & \frac{|\mathcal{A}\setminus B|-|\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}}|+|\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}}|+|\mathcal{B}\setminus\mathcal{A}|-|\mathcal{C}'_{\mathcal{B}\setminus\mathcal{A}}|+|\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}}|-|\mathcal{A}\setminus\mathcal{B}|+|\mathcal{B}\setminus\mathcal{A}|}{|\mathcal{A}\cup\mathcal{B}|} \\
= & 0.
\end{array}
$$

We get the above inequalities by noticing that $\mathrm{H}(\mathcal{A}_{\mathcal{A}\cap\mathcal{B}},\mathcal{B}_{\mathcal{A}\cap\mathcal{B}}) \leq \mathrm{H}(\mathcal{A}_{C'_{\mathcal{A}\cap\mathcal{B}}},\mathcal{B}_{C'_{\mathcal{A}\cap\mathcal{B}}}) + |\mathcal{A}\cap\mathcal{B}| - |\mathcal{C}'_{\mathcal{A}\cap\mathcal{B}}|$, $|\mathcal{A}\cup\mathcal{B}| \geq |\mathcal{A}\cup\mathcal{C}'|$, $|\mathcal{A}\cup\mathcal{B}| \geq |\mathcal{B}\cup\mathcal{C}'|$, $\mathrm{H}(\cdot,\cdot) \geq 0$, $|\mathcal{A}\cap\mathcal{B}| - |\mathcal{C}'_{\mathcal{A}\cap\mathcal{B}}| \geq 0$, $|\mathcal{A}\setminus\mathcal{B}| - |\mathcal{C}'_{\mathcal{A}\setminus\mathcal{B}}| \geq 0$, etc.

Proposition 9.2. *If $\mathcal{A}$ and $\mathcal{B}$ are arbitrary and $\mathcal{C} = \mathcal{C}' \cup \mathcal{C}''$, where $\mathcal{C}' \subset (\mathcal{A}\cup\mathcal{B})$ and $\mathcal{C}'' \cap (\mathcal{A}\cup\mathcal{B}) = \emptyset$, then $CE(\mathcal{A},\mathcal{C}) \geq CE(\mathcal{A},\mathcal{C}')$ and $CE(\mathcal{B},\mathcal{C}) \geq CE(\mathcal{B},\mathcal{C}')$.*

Proof. (Proposition 9.2)

$$
\begin{array}{ll}
 & \mathrm{CE}(\mathcal{A},\mathcal{C}) \\
= & \frac{\mathrm{H}(\mathcal{A}_{\mathcal{A}\cap\mathcal{C}'},\mathcal{C}'_{\mathcal{A}\cap\mathcal{C}'})+|\mathcal{A}\setminus\mathcal{C}'|+|\mathcal{C}'\setminus\mathcal{A}|+|\mathcal{C}''|}{|\mathcal{A}\cap\mathcal{C}'|+|\mathcal{A}\setminus\mathcal{C}'|+|\mathcal{C}'\setminus\mathcal{A}|+|\mathcal{C}''|} \\
\geq & \frac{\mathrm{H}(\mathcal{A}_{\mathcal{A}\cap\mathcal{C}'},\mathcal{C}'_{\mathcal{A}\cap\mathcal{C}'})+|\mathcal{A}\setminus\mathcal{C}'|+|\mathcal{C}'\setminus\mathcal{A}|}{|\mathcal{A}\cap\mathcal{C}'|+|\mathcal{A}\setminus\mathcal{C}'|+|\mathcal{C}'\setminus\mathcal{A}|} \\
= & \mathrm{CE}(\mathcal{A},\mathcal{C}').
\end{array}
$$

The case of $\mathcal{B}$ and $\mathcal{C}$ can be proven analogously.

Proof. (Theorem 9.1) Let us choose arbitrary subspace clusterings $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, for which $\mathcal{C} = \mathcal{C}' \cup \mathcal{C}''$ such that $\mathcal{C}' \subseteq (\mathcal{A}\cup\mathcal{B})$ and $\mathcal{C}'' \cap (\mathcal{A}\cup\mathcal{B}) = \emptyset$. By Propositions 9.1 and 9.2, $\mathrm{CE}(\mathcal{A},\mathcal{C}) + \mathrm{CE}(\mathcal{B},\mathcal{C}) \geq \mathrm{CE}(\mathcal{A},\mathcal{C}') + \mathrm{CE}(\mathcal{B},\mathcal{C}') \geq \mathrm{CE}(\mathcal{A},\mathcal{B})$.

Theorem 9.2. *(Triangle inequality for RNIA.) $RNIA(\mathcal{A},\mathcal{B}) \leq RNIA(\mathcal{A},\mathcal{C}) + RNIA(\mathcal{B},\mathcal{C})$ for any subspace clusterings $\mathcal{A},\mathcal{B},\mathcal{C}$.*

Proof. (Theorem 9.2) Let us pick subspace clusterings $\mathcal{A},\mathcal{B},\mathcal{C}$ for a data matrix $X = (x_{ij})$. Let us write n_A for the number of the elements of X that are clustered only by $\mathcal{A}$, n_{AB} for the number of the elements of X that are clustered by $\mathcal{A}$ and $\mathcal{B}$, and n_{ABC} for the number of the elements of X that are clustered by all three clusterings. We define n_B, n_C, n_{AC}, and n_{BC} similarly. Now, we can write

$$
\begin{array}{ll}
 & \mathrm{RNIA}(\mathcal{A},\mathcal{C}) + \mathrm{RNIA}(\mathcal{B},\mathcal{C}) - \mathrm{RNIA}(\mathcal{A},\mathcal{B}) \\
= & \frac{n_A+n_C+n_{AB}+n_{BC}}{n_A+n_C+n_{AB}+n_{BC}+n_{AC}+n_{ABC}} \\
+ & \frac{n_B+n_C+n_{AC}+n_{AB}}{n_B+n_C+n_{AC}+n_{AB}+n_{BC}+n_{ABC}} \\
- & \frac{n_A+n_B+n_{AC}+n_{BC}}{n_A+n_B+n_{AC}+n_{BC}+n_{AB}+n_{ABC}}
\end{array}
$$

Once the expression is fully expanded, all negative terms disappear. Since the expression is always greater than zero, the triangle inequality holds.

THEOREM 9.3. *(Triangle inequality for VI.) VI does not satisfy the triangle inequality in the case of subspace clusterings.*

Proof. We show by counterexample that VI does not satisfy the triangle inequality. Consider three subspace clusterings $\mathcal{A} = (A_1) = (\{1,2\},\{1,2,3,4\})$, $\mathcal{B} = (B_1) = (\{2,3\},\{1,2,3,4\})$, and $\mathcal{C} = (C_1, C_2) = (\{1,2\},\{1,2,3,4\},\{1,2\},\{6\})$. We have $\mathrm{VI}(\mathcal{A},\mathcal{B}) = 8/3 \log 2 \approx 1.84$, $\mathrm{VI}(\mathcal{A},\mathcal{C}) = 1/5 \log 2 \approx 0.14$, and $\mathrm{VI}(\mathcal{B},\mathcal{C}) = 17/7 \log 2 \approx 1.68$. Thus $\mathrm{VI}(\mathcal{A},\mathcal{C}) + \mathrm{VI}(\mathcal{B},\mathcal{C}) < \mathrm{VI}(\mathcal{A},\mathcal{B})$, showing that the triangle inequality does not hold. It does not hold for these three clusterings even for VI distances scaled by the logarithm of the union area.

THEOREM 9.4. *(Triangle inequality for Rand.) The Rand index does not satisfy the triangle inequality in the case of subspace clusterings.*

Proof. We show by counterexample that (1-Rand) does not satisfy the triangle inequality. Consider the example clusterings from the proof of the Theorem 9.3. For these clusterings, we have $1 - \mathrm{Rand}(\mathcal{A},\mathcal{B}) = (22 + 22)/66 \approx 0.67$, $1 - \mathrm{Rand}(\mathcal{A},\mathcal{C}) = 1/45 \approx 0.02$, and $1 - \mathrm{Rand}(\mathcal{B},\mathcal{C}) = 45/91 \approx 0.49$. The triangle inequality does not hold, since $1 - \mathrm{Rand}(\mathcal{A},\mathcal{C}) + 1 - \mathrm{Rand}(\mathcal{B},\mathcal{C}) < 1 - \mathrm{Rand}(\mathcal{A},\mathcal{B})$.

9.3 Penalty for Non-Intersecting Area. Consider adding $k \geq 1$ units of non-intersecting area to two subspace clusterings $\mathcal{A}$ and $\mathcal{B}$ and denote the resulting clusterings by $\mathcal{A}^U$ and $\mathcal{B}^U$ (note that one of these might actually equal to the original clustering). Are our distance measures able to penalize for this added non-intersecting area?

THEOREM 9.5. *(Penalty for Non-Intersecting Area with CE.) $CE(\mathcal{A}^U,\mathcal{B}^U) \geq CE(\mathcal{A},\mathcal{B})$ for all subspace clusterings $\mathcal{A}$, $\mathcal{B}$.*

Proof.

$$\begin{aligned}\mathrm{CE}(\mathcal{A}^U,\mathcal{B}^U) &= \tfrac{(|U|+k)-D_{max}}{(|U|+k)}\\ &\geq \tfrac{|U|-D_{max}}{|U|}\\ &= \mathrm{CE}(\mathcal{A},\mathcal{B}).\end{aligned}$$

THEOREM 9.6. *(Penalty for Non-Intersecting Area with RNIA.) $RNIA(\mathcal{A}^U,\mathcal{B}^U) \geq RNIA(\mathcal{A},\mathcal{B})$ for all subspace clusterings $\mathcal{A}$, $\mathcal{B}$.*

Proof.

$$\begin{aligned}\mathrm{RNIA}(\mathcal{A}^U,\mathcal{B}^U) &= \tfrac{(|U|+k)-|I|}{(|U|+k)}\\ &\geq \tfrac{|U|-|I|}{|U|}\\ &= \mathrm{RNIA}(\mathcal{A},\mathcal{B}).\end{aligned}$$

THEOREM 9.7. *(Penalty for Non-Intersecting Area with VI.) VI does not always penalize for non-intersecting area.*

Proof. We show by counterexample that VI does not always penalize for the non-intersecting area. Consider the clusterings in the proof of Theorem 9.3. We have the same intersection size but larger non-intersecting area for clusterings $\mathcal{B}$ and $\mathcal{C}$ than for clusterings $\mathcal{B}$ and $\mathcal{A}$, so $\mathcal{B}$ and $\mathcal{C}$ should be farther apart. Despite this, $\mathrm{VI}(\mathcal{B},\mathcal{C}) < \mathrm{VI}(\mathcal{A},\mathcal{B})$. This is true also if we scale the VI distances by the logarithms of the appropriate union areas.

THEOREM 9.8. *(Penalty for Non-Intersecting Area with Rand.) The Rand index does not always penalize for non-intersecting area.*

Proof. Consider the clusterings from the proof of Theorem 9.4. With similar reasoning as in the proof of Theorem 9.7, the fact that $1 - \mathrm{Rand}(\mathcal{B},\mathcal{C}) < 1 - \mathrm{Rand}(\mathcal{A},\mathcal{B})$ shows that (1-Rand) fails to penalize for the growing non-intersecting area.

9.4 Scale Invariant. Consider scaling all areas by a constant $c \geq 1$.

THEOREM 9.9. *(Scale Invariance for CE.) $CE(c\mathcal{A}, c\mathcal{B}) = CE(\mathcal{A},\mathcal{B})$ for all subspace clusterings $\mathcal{A}$, $\mathcal{B}$.*

Proof.

$$\begin{aligned}\mathrm{CE}(c\mathcal{A},c\mathcal{B}) &= \tfrac{c|U|-cD_{max}}{c|U|}\\ &= \tfrac{|U|-D_{max}}{|U|}\\ &= \mathrm{CE}(\mathcal{A},\mathcal{B}).\end{aligned}$$

THEOREM 9.10. *(Scale Invariance for RNIA.) $RNIA(c\mathcal{A}, c\mathcal{B}) = RNIA(\mathcal{A},\mathcal{B})$ for all subspace clusterings $\mathcal{A}$, $\mathcal{B}$.*

Proof.

$$\begin{aligned}\mathrm{RNIA}(c\mathcal{A},c\mathcal{B}) &= \tfrac{c|U|-c|I|}{c|U|}\\ &= \tfrac{|U|-|I|}{|U|}\\ &= \mathrm{RNIA}(\mathcal{A},\mathcal{B}).\end{aligned}$$

THEOREM 9.11. *(Scale Invariance for VI.) VI is not scale invariant in the case of subspace clusterings.*

Proof.

$$\begin{aligned}&\mathrm{VI}(c\mathcal{A},c\mathcal{B})\\ &= \tfrac{1}{c|U|}\Big[\textstyle\sum_i (cm_i)\log(cm_i) + \sum_j (cm'_j)\log(cm'_j)\\ &-2\textstyle\sum_i\sum_j (cm_{ij})\log(cm_{ij})\Big]\\ &= \mathrm{VI}(\mathcal{A},\mathcal{B}) + \log c\Big[\textstyle\sum_i m_i + \sum_j m'_j - 2\sum_i\sum_j m_{ij}\Big]\\ &\geq \mathrm{VI}(\mathcal{A},\mathcal{B}),\end{aligned}$$

since for subspace clusterings, $\sum_i m_i \geq \sum_i \sum_j m_{ij}$ and $\sum_j m'_j \geq \sum_i \sum_j m_{ij}$. The equality (and thus scale invariance) only holds for ordinary clusterings, for which $m_i = \sum_j m_{ij}$ and $m'_j = \sum_i m_{ij} \quad \forall\, i, j$.

THEOREM 9.12. *(Scale Invariance for Rand.) The Rand index is not scale invariant in the case of subspace clusterings.*

Proof.

$$\begin{aligned}
&1 - \text{Rand}(c\mathcal{A}, c\mathcal{B}) \\
&= \tfrac{\sum_{i=1}^{L'} \sum_{j=1}^{L} \sum_{k=j+1}^{L} (cm_{ij})(cm_{ik})}{1/2(cm)((cm)-1)} \\
&= \tfrac{\sum_{i=1}^{L} \sum_{j=1}^{L'} \sum_{k=j+1}^{L'} (cm_{ji})(cm_{ki})}{1/2(cm)((cm)-1)} \\
&= \tfrac{\sum_{i=1}^{L'} \sum_{j=1}^{L} \sum_{k=j+1}^{L} m_{ij}m_{ik} + \sum_{i=1}^{L} \sum_{j=1}^{L'} \sum_{k=j+1}^{L'} m_{ji}m_{ki}}{1/2m(m-1/c)} \\
&\leq 1 - \text{Rand}(\mathcal{A}, \mathcal{B}).
\end{aligned}$$

The equality holds only for $c = 1$, i.e. when there is no scaling.

9.5 Copy Invariant. Consider introducing two disjoint copies of the same clustering $\mathcal{S}$ in a large data matrix. Denote the new 'double clustering' by $\mathcal{S}^D$.

THEOREM 9.13. *(Copy Invariance for CE.) CE$(\mathcal{A}^D, \mathcal{B}^D)$ = CE$(\mathcal{A}, \mathcal{B})$ for all subspace clusterings $\mathcal{A}$, $\mathcal{B}$.*

Proof.

$$\begin{aligned}
\text{CE}(\mathcal{A}^D, \mathcal{B}^D) &= \tfrac{2|U| - 2D_{max}}{2|U|} \\
&= \tfrac{|U| - D_{max}}{|U|} \\
&= \text{CE}(\mathcal{A}, \mathcal{B}).
\end{aligned}$$

THEOREM 9.14. *(Copy Invariance for RNIA.) RNIA$(\mathcal{A}^D, \mathcal{B}^D)$ = RNIA$(\mathcal{A}, \mathcal{B})$ for all subspace clusterings $\mathcal{A}$, $\mathcal{B}$.*

Proof.

$$\begin{aligned}
\text{RNIA}(\mathcal{A}^D, \mathcal{B}^D) &= \tfrac{2|U| - 2|I|}{2|U|} \\
&= \tfrac{|U| - |I|}{|U|} \\
&= \text{RNIA}(\mathcal{A}, \mathcal{B}).
\end{aligned}$$

THEOREM 9.15. *(Copy Invariance for VI.) VI$(\mathcal{A}^D, \mathcal{B}^D)$ = VI$(\mathcal{A}, \mathcal{B})$ for all subspace clusterings $\mathcal{A}$, $\mathcal{B}$.*

Proof.

$$\begin{aligned}
&\text{VI}(\mathcal{A}^D, \mathcal{B}^D) \\
&= \tfrac{1}{2|U|} \Big[2 \textstyle\sum_i m_i \log m_i + 2 \sum_j m'_j \log m'_j \\
&-2 \cdot 2 \textstyle\sum_i \sum_j m_{ij} \log m_{ij} \Big] \\
&= \text{VI}(\mathcal{A}, \mathcal{B}).
\end{aligned}$$

THEOREM 9.16. *(Copy Invariance for Rand.) The Rand index is not copy invariant.*

Proof. Note that (1-Rand) can also be written as

$$\begin{aligned}
&1 - \text{Rand}(\mathcal{A}, \mathcal{B}) \\
&= \tfrac{1/2(\sum_{i=1}^{L'} \sum_{j=1}^{L} \sum_{k=1}^{L} m_{ij}m_{ik} - \sum_{i=1}^{L} \sum_{j=1}^{L'} m_{ij}^2)}{1/2m(m-1)} \\
&+ \tfrac{1/2(\sum_{i=1}^{L} \sum_{j=1}^{L'} \sum_{k=1}^{L'} m_{ji}m_{ki} - \sum_{i=1}^{L} \sum_{j=1}^{L'} m_{ij}^2)}{1/2m(m-1)}
\end{aligned}$$

Using this, we have

$$\begin{aligned}
&1 - \text{Rand}(\mathcal{A}^D, \mathcal{B}^D) \\
&= \tfrac{1/2(2\sum_{i=1}^{L'} 2\sum_{j=1}^{L} 2\sum_{k=1}^{L} m_{ij}m_{ik} - 2\sum_{i=1}^{L} 2\sum_{j=1}^{L'} m_{ij}^2)}{1/2(2m)((2m)-1)} \\
&+ \tfrac{1/2(2\sum_{i=1}^{L} 2\sum_{j=1}^{L'} 2\sum_{k=1}^{L'} m_{ji}m_{ki} - 2\sum_{i=1}^{L} 2\sum_{j=1}^{L'} m_{ij}^2)}{1/2(2m)((2m)-1)} \\
&= \tfrac{1/2(2\sum_{i=1}^{L'} \sum_{j=1}^{L} \sum_{k=1}^{L} m_{ij}m_{ik} - \sum_{i=1}^{L} \sum_{j=1}^{L'} m_{ij}^2)}{1/2m(m-1/2)} \\
&+ \tfrac{1/2(2\sum_{i=1}^{L} \sum_{j=1}^{L'} \sum_{k=1}^{L'} m_{ji}m_{ki} - \sum_{i=1}^{L} \sum_{j=1}^{L'} m_{ij}^2)}{1/2m(m-1/2)} \\
&\neq 1 - \text{Rand}(\mathcal{A}, \mathcal{B})
\end{aligned}$$

in the general case.

9.6 Multiple Cluster Coverage Penalty. Consider two subspace clusterings $\mathcal{A} = (A_1)$ and $\mathcal{B} = (B_1, \ldots, B_K)$ such that the clusters B_i are disjoint, of equal size $|A_1|/K = |U|/K$, and fully cover supp A_1.

THEOREM 9.17. *(Multiple Cluster Coverage Penalty for CE.) CE$(\mathcal{A}, \mathcal{B}) = \frac{K-1}{K}|U|$.*

Proof. We allow only one of the clusters B_i to be matched with A_1, so $K-1$ clusters are left unmatched, and the clustering error becomes CE$(\mathcal{A}, \mathcal{B}) = \frac{K-1}{K}|U|$.

THEOREM 9.18. *(Multiple Cluster Coverage Penalty for RNIA.) RNIA$(\mathcal{A}, \mathcal{B}) = 0$.*

Proof. In this case $I = U$ and therefore RNIA$(\mathcal{A}, \mathcal{B}) = 0$.

THEOREM 9.19. *(Multiple Cluster Coverage Penalty for VI.) VI$(\mathcal{A}, \mathcal{B}) = \log K$.*

Proof. In this case, $M = (m_{1j})$ for $j = 1, \ldots, K$, where $m_{1j} = |U|/K$. Also, $m_1 = |U|$ and $m'_j = |U|/K$. We thus get

$$\begin{aligned}
&\text{VI}(\mathcal{A}, \mathcal{B}) \\
&= \tfrac{1}{|U|} [|U| \log |U| + K(|U|/K) \log(|U|/K) \\
&-2K(|U|/K) \log(|U|/K)] \\
&= \log K.
\end{aligned}$$

THEOREM 9.20. *(Multiple Cluster Coverage Penalty for Rand.) $1 -$ Rand$(\mathcal{A}, \mathcal{B}) = \frac{|U|(K-1)}{K(|U|-1)}$.*

Proof. In this case, $N = |U|(|U| - 1)/2$, $N_{11} = 1/2(|U|/K)(|U|/K - 1)K$, and $N_{00} = 0$, so (1-Rand) becomes

$$\begin{aligned} & 1 - \text{Rand}(\mathcal{A}, \mathcal{B}) \\ & = 1 - \frac{N_{00}+N_{11}}{N} \\ & = 1 - \frac{1/2(|U|/K)(|U|/K-1)K}{|U|(|U|-1)/2} \\ & = \frac{|U|(K-1)}{K(|U|-1)}. \end{aligned}$$

9.7 Generalizable. If we have cluster sizes, unions, and intersections, but do not have a confusion matrix (like in the case of non-axis-aligned subspace clusterings or attribute weighted clusterings), are we able to apply our distance measures?

THEOREM 9.21. *(Generalizability for CE) CE is generalizable.*

Proof. Computing $\text{CE}(\mathcal{A}, \mathcal{B})$ according to Eq. 3.3 only requires cluster intersection sizes $|A_i \cap B_j| = m_{ij}$ and the union size $|U|$, so it follows that CE is generalizable.

THEOREM 9.22. *(Generalizability for RNIA) RNIA is generalizable.*

Proof. Computing $\text{RNIA}(\mathcal{A}, \mathcal{B})$ according to Eq. 3.4 only requires the union size $|U|$ and the intersection size $|I|$ of the two clusterings, so RNIA is generalizable.

THEOREM 9.23. *(Generalizability for VI) VI is generalizable.*

Proof. Computing $\text{VI}(\mathcal{A}, \mathcal{B})$ according to Eq. 9.11 only requires cluster intersection sizes $|A_i \cap B_j| = m_{ij}$ and the union size $|U|$, so VI is generalizable.

THEOREM 9.24. *(Generalizability for Rand) The Rand index is not generalizable.*

Proof. To compute (1-Rand), we need the confusion matrix of the two clusterings, so Rand is not generalizable.

10 Appendix 2: Proof for Theorem 4.1

PROPOSITION 10.1. *If (u_j, v_j) is a principal pair, then $u_j = \sigma_j^{-1}\Pi_{\mathcal{F}} v_j$ where $\Pi_{\mathcal{F}} x$ represents the projection of vector x on the subspace $\mathcal{F}$.*

Proof. (Proposition 10.1) True for $j = 1$ by the definition of the projection operation. For $j > 1$, we have by the definition of (u_j, v_j) that $u_j = \sigma_j^{-1}\Pi_{(u_1,\ldots u_{j-1})^{\perp_{\mathcal{F}}}} v_j$ where $(u_1, \ldots u_{j-1})^{\perp_{\mathcal{F}}}$ represents the orthogonal complement of $(u_1, \ldots u_{j-1})$ in $\mathcal{F}$. By the elementary "3 perpendicular theorem" we also have that $v_j \perp u_{j'}$ for any $j' < j$. Therefore, $\Pi_{(u_1,\ldots u_{j-1})} v_j = 0$ and $\Pi_{\mathcal{F}} v_j = \sigma_j u_j$.

PROPOSITION 10.2. *If U is a $n \times a$ matrix of rank at most p, then $||U||_F^2 \leq p||U||_2^2$, where $|| \; ||_F$ represents the Frobenius norm.*

Proof. (Proposition 10.2) $||U||_F^2 = \operatorname{tr} U^T U \stackrel{1}{=} \sum_{j=1}^{p} \lambda_i(U^T U) \leq p\lambda_{max}(U^T U) = p||U||_2^2$. Equality $\stackrel{1}{=}$ holds because at most p eigenvalues of $U^T U$ are nonzero.

Proof. (Theorem 4.1) Let $\sigma_{ij} = \cos(\theta^j_{\mathcal{F},\mathcal{G}_i})$ and let (u_{ij}, v_{ij}), $u_{ij} \in \mathcal{F}$, $v_{ij} \in \mathcal{G}_i$, $i = 1, \ldots K$, $j = 1, \ldots a_i$ be the principal vectors. Note that $\{v_{ij}\}_{ij}$ form an orthonormal system and denote by V the $n \times a$ matrix $V = [v_{11}\, v_{12} \ldots v_{Ka_K}]$. Define $\tilde{u}_{ij} = \Pi_{\mathcal{F}} v_{ij}$; $\tilde{u}_{ij}$ is a vector of length σ_{ij} with the same direction as u_{ij} by Proposition 10.1. Form the matrix U having $\tilde{u}_{ij}$, $i = 1, \ldots K$, $j = 1, \ldots a_i$ as columns. Then $||U||_F^2 = \operatorname{tr} U^T U = \sum_{ij} ||\tilde{u}_{ij}||^2 = \sum_{ij} \sigma_{ij}^2$.

It remains to show that $||U||_F^2 \leq p$. But, in matrix notation, $U = HV$ where H is the symmetric, idempotent matrix representing the projection onto $\mathcal{F}$. It is easy to verify that $||V||_2 = 1$ and $||H||_2 = 1$. Therefore, $||U||_2 \leq ||H||_2||V||_2 = 1$ and by virtue of Proposition 10.2 we obtain $||U||_F^2 \leq p$.

11 Appendix 3: Proofs for Section 5.2

Consider two co-clusterings $\mathcal{S} = (S_{ij})$, $\mathcal{S}' = (S'_{ij})$ together with the corresponding row clusterings $\mathcal{R} = (R_i)$, $\mathcal{R}' = (R'_i)$ and the column clusterings $\mathcal{C} = (C_i)$, $\mathcal{C}' = (C'_i)$. For the row clusterings, we define the cluster sizes as $r_i = |R_i|$, $r'_i = |R'_i|$, and the cluster intersection sizes as $r_{ij} = |R_i \cap R'_j|$. Similarly, for the column clusterings, we have $c_i = |C_i|$, $c'_i = |C'_i|$, and $c_{ij} = |C_i \cap C'_j|$. For the co-clusterings, the sizes of the clusters are defined as $m_{ij} = |S_{ij}| = r_i c_j$, $m'_{ij} = |S'_{ij}| = r'_i c'_j$, and the intersection of two co-clusters is $m_{ijkl} = |S_{ij} \cap S'_{kl}| = r_{ik} c_{kl}$. Recall that N_R and N_C stand for the number of data matrix rows and columns, respectively. Also recall that D_{max} is the sum of the diagonal elements of the co-clustering confusion matrix after an optimal permutation of the co-cluster labels. Let us write D^R_{max} for the corresponding sum for the row clustering confusion matrix and D^C_{max} for the column clustering confusion matrix.

THEOREM 11.1. *(CE and Co-Clusterings) $CE(\mathcal{S}, \mathcal{S}') = CE(\mathcal{R}, \mathcal{R}') + CE(\mathcal{C}, \mathcal{C}') - CE(\mathcal{R}, \mathcal{R}')CE(\mathcal{C}, \mathcal{C}')$ for any co-clusterings $\mathcal{S} = (\mathcal{R}, \mathcal{C})$, $\mathcal{S}' = (\mathcal{R}', \mathcal{C}')$.*

Proof. Fix the permutation of the co-cluster labels that minimizes $\text{CE}(\mathcal{S}, \mathcal{S}')$. The same permutation of the labels also minimizes $\mathrm{CE}(\mathcal{R},\mathcal{R}')$ and $\mathrm{CE}(\mathcal{C},\mathcal{C}')$.

$$
\begin{aligned}
&\mathrm{CE}(\mathcal{S},\mathcal{S}')\\
&= \tfrac{N_R N_C - D_{max}}{N_R N_C}\\
&= \tfrac{N_R N_C - \sum_i \sum_j m_{ijij}}{N_R N_C}\\
&= \tfrac{N_R N_C - \sum_i \sum_j r_{ii} c_{jj}}{N_R N_C}\\
&= \tfrac{N_R N_C - \sum_i r_{ii} \sum_j c_{jj}}{N_R N_C}\\
&= \tfrac{N_R N_C - D^R_{max} D^C_{max}}{N_R N_C}\\
&= \tfrac{N_R - D^R_{max}}{N_R} + \tfrac{N_C - D^C_{max}}{N_C} + \tfrac{(N_R - D^R_{max})(N_C - D^C_{max})}{N_R N_C}\\
&= \mathrm{CE}(\mathcal{R},\mathcal{R}') + \mathrm{CE}(\mathcal{C},\mathcal{C}') - \mathrm{CE}(\mathcal{R},\mathcal{R}')\mathrm{CE}(\mathcal{C},\mathcal{C}').
\end{aligned}
$$

THEOREM 11.2. *(RNIA and Co-Clusterings) RNIA*$(\mathcal{S},\mathcal{S}') = 0$ *for any co-clusterings* $\mathcal{S} = (\mathcal{R},\mathcal{C})$, $\mathcal{S}' = (\mathcal{R}',\mathcal{C}')$.

Proof. For co-clusterings it always holds that $I = U$.

THEOREM 11.3. *(VI and Co-Clusterings) VI*$(\mathcal{S},\mathcal{S}') =$ *VI*$(\mathcal{R},\mathcal{R}') +$ *VI*$(\mathcal{C},\mathcal{C}')$ *for any co-clusterings* $\mathcal{S} = (\mathcal{R},\mathcal{C})$, $\mathcal{S}' = (\mathcal{R}',\mathcal{C}')$.

Proof.

$$
\begin{aligned}
&\mathrm{VI}(\mathcal{S},\mathcal{S}')\\
&= \tfrac{1}{N_R N_C} \sum_i \sum_j \sum_k \sum_l m_{ijkl} \log \tfrac{m_{ij} m'_{kl}}{m^2_{ijkl}}\\
&= \tfrac{1}{N_R N_C} \sum_i \sum_j \sum_k \sum_l r_{ik} c_{jl} \log \tfrac{r_i c_j r'_k c'_l}{r^2_{ik} c'^2_{jl}}\\
&= \tfrac{1}{N_R N_C} \sum_i \sum_k r_{ik} \log \tfrac{r_i r'_k}{r^2_{jk}} \sum_j \sum_l c_{jl}\\
&+ \tfrac{1}{N_R N_C} \sum_j \sum_l c_{jl} \log \tfrac{c_j c'_l}{c^2_{jl}} \sum_i \sum_k r_{ik}\\
&= \tfrac{1}{N_R} \sum_i \sum_k r_{ik} \log \tfrac{r_i r'_k}{r^2_{ik}} + \tfrac{1}{N_C} \sum_j \sum_l c_{jl} \log \tfrac{c_j c'_l}{c^2_{jl}}\\
&= \mathrm{VI}(\mathcal{R},\mathcal{R}') + \mathrm{VI}(\mathcal{C},\mathcal{C}').
\end{aligned}
$$

We get this by noticing that $\sum_j \sum_l c_{jl} = N_C$ and that $\sum_i \sum_k r_{ik} = N_R$.

It does not seem possible to derive a simple relationship between the Rand index for the co-clusterings and the Rand indices for the corresponding row/column clustering. However, we are able to observe the following connection.

Based on Eq. 9.13, we have the following expression for the row clusterings:

$$1 - \mathrm{Rand}(\mathcal{R},\mathcal{R}') = \tfrac{\sum_i r_i^2 + \sum_k r'^2_k - 2\sum_i \sum_k r^2_{ik}}{N_R(N_R - 1)}.$$

For the column clusterings, we have

$$1 - \mathrm{Rand}(\mathcal{C},\mathcal{C}') = \tfrac{\sum_j c_j^2 + \sum_l c'^2_l - 2\sum_j \sum_l c^2_{jl}}{N_C(N_C - 1)},$$

and for the co-clusterings,

$$
\begin{aligned}
&1 - \mathrm{Rand}(\mathcal{S},\mathcal{S}')\\
&= \tfrac{\sum_i \sum_j m^2_{ij} + \sum_k \sum_l m^2_{kl} - 2\sum_i \sum_j \sum_k \sum_l m^2_{ijkl}}{m(m-1)}\\
&= \tfrac{\sum_i \sum_j r_i^2 c_j^2 + \sum_k \sum_l r'^2_k c'^2_l - 2\sum_i \sum_j \sum_k \sum_l r^2_{ik} r^2_{jl}}{N_R N_C (N_R N_C - 1)}\\
&= \tfrac{\sum_i r_i^2 \sum_j c_j^2 + \sum_k r'^2_k \sum_l c'^2_l - 2\sum_i \sum_k r^2_{ik} \sum_j \sum_l c^2_{jl}}{N_R N_C (N_R N_C - 1)}.
\end{aligned}
$$