

# Measuring the Validity of Clustering Validation Datasets

Hyeon Jeon, Michaël Aupetit, DongHwa Shin, Aeri Cho, Seokhyeon Park, and Jinwook Seo

**Abstract**—Clustering techniques are often validated using benchmark datasets where class labels are used as ground-truth clusters. However, depending on the datasets, class labels may not align with the actual data clusters, and such misalignment hampers accurate validation. Therefore, it is essential to evaluate and compare datasets regarding their cluster-label matching (CLM), i.e., how well their class labels match actual clusters. Internal validation measures (IVMs), like Silhouette, can compare CLM over different labeling of the *same dataset*, but are not designed to do so *across different datasets*. We thus introduce *Adjusted IVMs* as fast and reliable methods to evaluate and compare CLM across datasets. We establish four axioms that require validation measures to be independent of data properties not related to cluster structure (e.g., dimensionality, dataset size). Then, we develop standardized protocols to convert any IVM to satisfy these axioms, and use these protocols to adjust six widely used IVMs. Quantitative experiments (1) verify the necessity and effectiveness of our protocols and (2) show that adjusted IVMs outperform the competitors, including standard IVMs, in accurately evaluating CLM both *within* and *across datasets*. We also show that the datasets can be filtered or improved using our method to form more reliable benchmarks for clustering validation.

**Index Terms**—Clustering, Clustering Validation, Internal Clustering Validation, External Clustering Validation, Clustering Benchmark

## 1 INTRODUCTION

Cluster analysis [1] is an essential exploratory task for data scientists and practitioners in various application domains [2]–[7]. It commonly relies on unsupervised clustering techniques, that is, machine learning algorithms that partition data into subsets called groups or clusters. These algorithms maximize between-cluster separation and within-cluster compactness based on a given distance function [8].

Clustering validation measures [9], [10] or quality measures [11] are used to evaluate clustering results. They are categorized as *internal* measures and *external* measures [8]. Internal validation measures (IVM) (e.g., Silhouette score [12]), also known as relative measures [1], give high scores to partitions in which data points with high or low similarities to each other are assigned to the same or different clusters, respectively. External validation measures (EVM) [13], [14] (e.g., adjusted mutual information [15]) quantify how well a clustering result matches an externally given ground truth partition.

Using the classes of a labeled dataset as a ground truth partition is a typical approach to conduct external validation [9]. This approach promotes a clustering technique that precisely distinguishes labeled classes as separated clusters. The underlying assumption is that the classes of the dataset align well with the clusters [9], [16], [17] (Figure 1A). We name it the *Cluster-Label Matching (CLM)* assumption.

If the CLM of a labeled dataset is accurate, EVMs work as intended. EVM scores become low only when the clustering technique fails to capture the clusters represented by the

classes (Figure 1G). That is, external validation is reliable only when the CLM assumption is valid.

However, datasets can have poor CLM due to labels split across clusters (Figure 1B magenta-colored points). For instance, images of buses and cars both assigned to the `Vehicle` class, may form distinct clusters in the image space. Conversely, data sets can also have poor CLM when they have multiple labels overlap within a single cluster (Figure 1B magenta- and yellow-colored points). This happens, for example, when images of visually similar categories, such as leopards and cheetahs, cause `Leopard` and `Cheetah` classes to overlap in the image space.

With such poor CLM datasets, EVMs become unreliable, producing low scores regardless of a clustering technique’s capacity to capture clusters. In fact, a suboptimal clustering technique may receive a low EVM score because its incorrect cluster partition is unlikely to align well with the already inaccurate class label partition (Figure 1H). Conversely, even a well-performing clustering technique, despite correctly capturing the “natural” clusters, may also receive a low EVM score because these accurately identified clusters do not align well with class labels in bad-CLM datasets (Figure 1E). In essence, poor CLM undermines EVMs, making it difficult to differentiate between effective and ineffective clustering techniques.

It is thus crucial to evaluate CLM—the intrinsic validity of the ground truth labeled dataset—to compare clustering techniques on a reliable basis. Not checking CLM before executing external validation casts doubt on the results obtained. Moreover, such validation might lead to an erroneous conclusion when ranking and comparing clustering techniques (Section 7.1). Still, external validation is often conducted without considering the CLM of benchmark datasets [18]–[21]. Our goal, therefore, is to evaluate the CLM of labeled datasets to distinguish credible benchmark

Hyeon Jeon, Aeri Cho, Seokhyeon Park, and Jinwook Seo are with Seoul National University. e-mail: {hj, archo, shpark}@hclil.snu.ac.kr, jseo@snu.ac.kr  
 Michaël Aupetit is with Qatar Computing Research Institute, Hamad Bin Khalifa University. e-mail: maupetit@hbku.edu.qa  
 DongHwa Shin is with Kwangwoon University. e-mail: dhshin@kw.ac.kr  
 Jinwook Seo is a corresponding author.

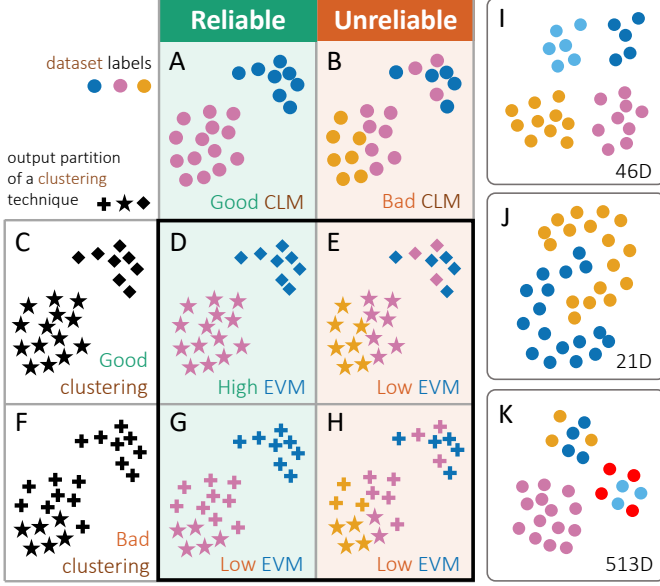


Fig. 1. The illustration of how the degree of Cluster-Label Matching (CLM) affects the reliability of external validation. An external validation measure (EVM) evaluates how well the clustering results (markers' shape) match the ground truth partition, typically given by class labels (markers' hue). The CLM is *good* (A) if the partition formed by the *class labels* (fill) matches well the *clusters* formed by the point distribution (encoded by position). If CLM is *good* (A), the EVM between the dataset labels (A) and clustering results (C, F) gives a *reliable* evaluation (D, G) of the *clustering technique*: *high/low EVM* (D/G) aligns well with *good/bad clustering* (C/F). But if CLM is *bad* (B), the EVM is always *low* (E, H) and *unreliable* to evaluate whether the *clustering* is good (C) or bad (F). We aim to evaluate and compare the CLM *across datasets* (I,J,K) having diverse characteristics (e.g., dimension, size, data, and class distributions) to inform external validation by distinguishing valid benchmark datasets.

datasets for clustering validation. We moreover aim to inform the community to use such datasets to conduct more reliable external validation.

Yet, evaluating and comparing the CLM of datasets is a challenging problem. IVMs can be natural candidates for quantifying CLM since well-clustered class partitions tend to naturally receive higher IVM scores. However, they are designed to compare different partitions of a *single* dataset (Figure 1C *vs.* Figure 1F), and not to compare partitions *across different* datasets (Figure 1IJK). This limitation arises because IVMs depend not only on clustering quality but also on the dataset's size and dimensionality, and the distributions of data and class labels. Consequently, comparing IVM scores of different datasets is not reliable, making them improper for measuring and comparing CLM *across* different datasets (Section 6.2; Table 2). This underscores the need for a more suitable measure of CLM to validate clustering benchmark datasets.

To address this gap, we design *adjusted* internal validation measures (IVM<sub>AS</sub>) to assess and compare CLM *across* datasets. Here are our key contributions:

- We first establish four *across-dataset* axioms in Section 3, that form the theoretical basis for designing IVM<sub>AS</sub>, complementing Ackerman and Ben-David's *within-dataset* axioms [11] already satisfied by standard IVMs.
- We propose adjustment protocols in Section 4 for trans-

forming an IVM into an IVM<sub>A</sub> that satisfies each across-dataset axiom.

- In Section 5, we use our protocols to adjust six of the most widely used IVMs [8] (e.g., Silhouette coefficient [12] and Calinski-Harabasz index [22]). The resulting IVM<sub>AS</sub> are practical ways to quantify and compare CLM both within and across datasets.
- In Section 6, we conduct an ablation study to verify the validity and necessity of our new axioms and adjustment protocols. We also verify the effectiveness of IVM<sub>AS</sub> in accurately evaluating and comparing CLM, both within and between datasets. Finally, a runtime analysis demonstrates the advantage of IVM<sub>AS</sub> in terms of scalability.
- Finally, we present two applications of these IVM<sub>AS</sub> in Section 7 to demonstrate their benefits for the practitioners. We first validate the importance of evaluating the CLM over benchmark datasets by showing the instability of external validation when this step is overlooked. By doing so, we identify the top-CLM benchmark datasets that practitioners can use with higher confidence for evaluating clustering techniques. Second, we show that IVM<sub>AS</sub> can also be used to correct bad-CLM benchmark datasets by searching for data subspaces that maximize IVM<sub>AS</sub>.

## 2 BACKGROUNDS AND RELATED WORK

One common way to distinguish good and bad clustering techniques is to use EVMs. EVMs quantify how much the resulting clustering matches with a ground truth partition of a given dataset. For example, adjusted mutual information [15] measures the agreement of two assignments (clustering and ground truth partitions) in terms of information gain corrected for chance effects.

The classes in labeled datasets have been used extensively as ground truth partitions for EVM [9]. However, despite the potential risk of violating the CLM (Section 1; Figure 1), no principled procedure has yet been proposed to evaluate the reliability of such a ground truth. Our research aims to fill this gap by proposing measures that can evaluate and compare CLM across datasets. A similar endeavor has been engaged in the supervised learning community to quantify datasets' difficulty for classification tasks [23].

A natural approach is to use classification scores as a proxy for CLM [24], [25]. This approach is based on the assumption that the classes of a labeled dataset with good classification scores would provide reliable ground truth for EVM. Still, a classifier can hardly distinguish between two "adjacent" classes forming a single cluster (Figure 1B orange and magenta points in the bottom left cluster, good class separation but bad CLM) and two "separated" classes forming distant clusters (Figure 1A blue and magenta clusters, good CLM). It also cannot distinguish different within-class structures, such as a class forming a single cluster (Figure 1A blue class, good CLM) and one made of several distant clusters (Figure 1B magenta class, bad CLM). In addition, classifiers require expensive training time (Section 6.4).

A more direct approach is to examine how well the clustering techniques capture class labels, as well-separated classes will be easily captured by the techniques (Figure 1D). However, this approach is also computationally expensive

(Section 6.4). Moreover, it is not based on principled axioms independent of any clustering technique, so it is likely to be biased with respect to a certain type of cluster. Still, we can approximate ground truth CLM by running multiple and diverse clustering techniques [26] and aggregating their EVM scores. For lack of a better option, we use this ensemble approach to obtain an approximate ground truth in our experiments to validate our axiom-based solution.

In contrast, IVMs are inexpensive to compute (Section 6.4). They also examine cluster structure in more detail, relying on two criteria, namely *compactness* (i.e., the pairwise proximity of data points within a cluster) and *separability* (i.e., the degree to which clusters lie apart from one another) [8], [27]. For example, in Figure 1, an IVM would give a higher score to clustering partition C than to F. Moreover, following the axiomatization of clustering by Kleinberg [28], Ackerman and Ben-David [11] proposed four *within-dataset* axioms that give a common ground to IVMs: scale invariance, consistency, richness, and isomorphism invariance (Section 3.1). These axioms set the general requirements that a function should satisfy to work properly as an IVM.

Nevertheless, IVMs were originally designed to compare and rank different partitions of the *same* dataset (Figure 1A-H). Therefore, IVMs are not only affected by cluster structure but also dependent on the characteristics of the datasets, such as the number of points, classes, and dimensions (Figure 1I-K), which means that they are cannot properly compare CLM *across* different datasets. Here, we propose four additional axioms that IVMs should satisfy to compare CLM across datasets and derive new *adjusted* IVMs satisfying the axioms (i.e.,  $IVM_{AS}$ ).  $IVM_{AS}$  play the role of a proxy of CLM, evaluating the intrinsic validity of a benchmark dataset for external clustering validation.

### 3 NEW AXIOMS FOR ADJUSTED IVM

We propose new *across-dataset* axioms that adjusted IVMs (i.e.,  $IVM_{AS}$ ) should satisfy to properly evaluate and compare CLM across datasets, complementing the within-dataset ones of Ackerman and Ben-David [11].

#### 3.1 Within-Dataset Axioms

Ackerman and Ben-David (A&B) introduced *within-dataset* axioms [11] that specify the requirements for IVM to properly evaluate clustering partitions (Appendix A): **W1: Scale Invariance** requires measures to be invariant to distance scaling; **W2: Consistency** is satisfied by a measure that increases when within-cluster compactness or between-cluster separability increases; **W3: Richness** requires measures possible to give any fixed cluster partition the best score over the domain by only modifying the distance function; and **W4: Isomorphism Invariance** ensures that an IVM does not depend on the external identity of points (e.g., class labels).

#### 3.2 Across-Dataset Axioms

Within-dataset axioms do not consider the case of comparing scores across datasets; rather, they assume that the dataset is invariant. We propose four additional *across-dataset* axioms that a function should satisfy to fairly compare cluster partitions across datasets.

**Notations.** We begin by defining four fundamental building blocks of our axioms, using notation identical to that of A&B:

- A **finite domain set**  $X \subset \mathcal{D}$  of dimension  $\Delta_X$ , where  $\mathcal{D}$  denotes the data space.
- A **clustering partition** of  $X$  as  $C = \{C_1, C_2, \dots, C_{|C|}\}$ , where  $\forall i \neq j, C_i \cap C_j = \emptyset$  and  $\bigcup_{i=1}^{|C|} C_i = X$ .
- A **distance function**  $\delta : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ , satisfying  $\delta(x, y) \geq 0$ ,  $\delta(x, y) = \delta(y, x)$  and  $\delta(x, y) = 0$  if  $x = y \forall x, y \in \mathcal{D}$ . We do not require the triangle inequality.
- A **measure**  $f$  as a function that takes  $C, X, \delta$  as input and returns a real number. Higher  $f$  implies a better clustering.

We extend these notations with:

- $\bar{X}'$  the **centroid** of  $X'$ , where  $X' \subset X$ .
- $\underline{W}_\alpha$  a **random subsample** of the set  $W$  ( $\underline{W}_\alpha \stackrel{D}{=} W$ ) such that  $|\underline{W}_\alpha|/|W| = \alpha$ , and the corresponding clustering partition is noted  $\underline{C}_\alpha = \{C_{i_\alpha}\}_{i=1, \dots, |C|}$ .

**Goals and factors at play.** A&B's within-dataset axioms are based on the assumption that the measures that satisfy these axioms properly evaluate the quality of a clustering partition. However, the axioms do not consider that the measures  $f$  could operate on varying  $C$ ,  $\delta$ , and  $X$ . For example, isomorphism invariance (W4) assumes fixed  $X$  and  $\delta$ ; consistency (W2) and richness (W3) define how functions  $f$  should react to the change of  $\delta$ , but do not consider how  $\delta$  changes in real terms, affected by various aspects of  $X$  (e.g., dimensionality); scale invariance (W1) considers such variations, but only in terms of the global scaling. Thus, the satisfaction of A&B's axioms is a way to ensure IVMs focus on measuring clustering quality within a single dataset but not across datasets.

In contrast,  $IVM_{AS}$  shall operate on varying  $C$ ,  $\delta$ , and  $X$ . Thus, several aspects of the varying datasets now come into play, and their influence on  $IVM_A$  shall be minimized. The sample size  $|X|$  is one of them (**Axiom A1**), and the dimension  $\Delta_X$  of the data is another one (**Axiom A2**). Moreover, what matters is the *matching* between natural clusters and data labels more than the number of clusters or labels; therefore, we shall reduce the influence of the number of labels  $|C|$  (**Axiom A3**). Lastly, we need to align  $IVM_A$  to a comparable range of values (**Axiom A4**) across datasets, in essence capturing all remaining hard-to-control factors unrelated to clustering quality. We now explain the new axioms in detail:

#### Axiom A1: Data-Cardinality Invariance

Invariance of the sample size  $|X|$  is ensured if subsampling all clusters in the same proportion does not affect the  $IVM_A$  score. This leads to the first axiom:

**A1 – Data-Cardinality Invariance.** A measure  $f$  satisfies data-cardinality invariance if  $\forall X, \forall \delta$  and  $\forall C$  over  $(X, \delta)$ ,  $f(C, X, \delta) = f(\underline{C}_\alpha, X_\alpha, \delta)$  with  $X_\alpha = \bigcup_{i=1}^{|C|} \underline{C}_{i_\alpha} \forall \alpha \in ]0, 1]$ .

#### Axiom A2: Shift Invariance

We shall consider that data dimension  $\Delta_X$  varies across datasets. An important aspect of the dimension called the concentration of distance phenomenon, which is related to

the curse of dimensionality, affects the distance measures involved in  $\text{IVM}_A$ . As the dimension grows, the variance of the pairwise distance for any data tends to be constant, while its mean value increases with the dimension [29]–[31]. Therefore, in high-dimensional spaces,  $\delta$  will act as a constant function for any data  $X$ , and thus an  $\text{IVM}_A$   $f$  will generate similar scores for all datasets. To mitigate this phenomenon, and as a way to reduce the influence of the dimension, we require that the measure  $f$  be shift invariant [31], [32] so that the shift of the distances (i.e., growth of the mean) can be canceled out.

**A2 – Shift Invariance.** A measure  $f$  satisfies the shift invariance if  $\forall X, \forall \delta$ , and  $\forall C$  over  $(X, \delta)$ ,  $f(C, X, \delta) = f(C, X, \delta + \beta)$   $\forall \beta > 0$ , where  $\delta + \beta$  is a distance function satisfying  $(\delta + \beta)(x, y) = \delta(x, y) + \beta, \forall x, y \in X$ .

#### Axiom A3: Class-Cardinality Invariance

The number of classes should not affect an  $\text{IVM}_A$ ; for example, two well-clustered classes should get an  $\text{IVM}_A$  score similar to 10 well-clustered classes. A&B proposed that the minimum, maximum, and average class-pairwise aggregations of IVMs form yet other valid IVMs. We follow this principle as an axiom for  $\text{IVM}_A$ .

**A3 – Class-Cardinality Invariance.** A measure  $f$  satisfies class-cardinality invariance if  $\forall X, \forall \delta$  and  $\forall C$  over  $(X, \delta)$ ,  $f(C, X, \delta) = \text{agg}_{S \subseteq C, |S|=2} f'(S, X, \delta)$  where function  $\text{agg}_S \in \{\text{avg}_S, \text{min}_S, \text{max}_S\}$  and  $f'$  is an IVM.

#### Axiom A4: Range Invariance

Lastly, we need to ensure that an  $\text{IVM}_A$  takes a common range of values across datasets. In detail, we want their minimum and maximum values to correspond to the datasets with the worst and the best CLM, respectively, and that these extrema are aligned across datasets (we set them arbitrarily to 0 and 1), as follows:

**A4 – Range Invariance.** A measure  $f$  satisfies range invariance if  $\forall X, \forall \delta$ , and  $\forall C$  over  $(X, \delta)$ ,  $\min_C f(C, X, \delta) = 0$  and  $\max_C f(C, X, \delta) = 1$ .

## 4 GENERALIZATION PROTOCOLS

We introduce four technical protocols (T1-T4), designed to generate  $\text{IVM}_A$ s that satisfy the corresponding axioms A1-A4, respectively.

#### T1: Approaching Data-Cardinality Invariance (A1)

We cannot guarantee the invariance of a measure for any subsampling of the data (e.g., very small sample size). However, we can obtain robustness to random subsampling if we use consistent estimators of population statistics [33] as building blocks of the measure. For example, we can use the mean, the median, or the standard deviation of the points within a class or the whole dataset or quantities derived from them, such as the average distance between all points of two classes.

#### T2: Achieving Shift Invariance (A2)

**T2-a,b: Exponential protocol.** Considering a vector of distances  $u = (u_1 \dots u_n)$ , we can define a shift-invariant function by using a ratio of exponential functions:

$$g_j(u) = \frac{e^{u_j}}{\sum_k e^{u_k}}. \quad (1)$$

We observe that  $\forall \beta \in \mathbb{R}$ ,

$$g_j(u + \beta) = \frac{e^{u_j + \beta}}{\sum_k e^{u_k + \beta}} = \frac{e^{u_j}}{\sum_k e^{u_k}} \frac{e^\beta}{e^\beta} = g_j(u), \quad (2)$$

hence  $g_j$  is shift invariant. Thus, the measure  $f$  is shift invariant if it consists of ratios of the exponential distances (T2-a). Note that this protocol is at the core of the  $t$ -SNE loss function [31]. If a building block is a sum or average of distances, the exponential should be applied to the average of distances rather than individuals (T2-b), as the shift occurs to the average distances [30].

**T2-c: Equalizing shifting.** The exponential protocol can be safely applied only if the measure incorporates the distance between data points within  $X$  (Type-1 distance). We do not know, in general, how the shift of type-1 distances affects the distances between data points and their centroid (Type-2), nor do we know how the shift affects the distance between two centroids (Type-3), even though they are common building blocks in IVM [8]. Fortunately, if  $\delta$  is the square of Euclidean distances (i.e.,  $\delta = d^2$  where  $d(x, y)$  denotes the Euclidean distance between points  $x$  and  $y$ ), we can prove that the shift of type-1 distances by  $\beta$  results in the shift of type-2 distances by  $\beta/2$ , and in no shift of type-3 distances, which is stated by the following theorems (proof in Appendix D.1).

**Theorem 1 (Type-2 Shift).**  $\forall X' \subset X, \forall \beta > 0$ , and for any Euclidean distance functions  $d_L$  and  $d_H$  satisfying  $d_H^2 = d_L^2 + \beta$ ,  $\sum_{x \in X'} d_H^2(x, c) = \sum_{x \in X'} d_L^2(x, c) + \beta/2$ , where  $c = \bar{X}'$ .

**Theorem 2 (Type-3 Shift).**  $\forall X', X'' \subset X, \forall \beta > 0$ , and for any Euclidean distance functions  $d_L$  and  $d_H$  satisfying  $d_H^2 = d_L^2 + \beta$ ,  $d_H^2(c', c'') = d_L^2(c', c'')$ , where  $c' = \bar{X}'$  and  $c'' = \bar{X}''$ .

Therefore, if an IVM consists of different types of distances, we should use  $\delta = d^2$  and apply the exponential protocol with the same type of distances for both its numerator and denominator (T2-c).

**T2-d: Recovering Scale Invariance.** After applying the exponential protocol,  $g_j$  is no more scale-invariant:

$$\forall \lambda \in \mathbb{R}, g_j(\lambda u) = \frac{e^{\lambda u_j}}{\sum_k e^{\lambda u_k}} \neq g_j(u), \quad (3)$$

and so it will not satisfy axiom W1. We can recover scale-invariance by normalizing each distance  $u_i$  by a term that scales with all of the  $u_k$  together, such as their standard deviation,  $\sigma(u)$ . Now,

$$g_j(\lambda u / \sigma(\lambda u)) = g_j(\lambda u / \lambda \sigma(u)) = g_j(u / \sigma(u)), \quad (4)$$

is both shift and scale invariant (T2-d).

### T3: Achieving Class-Cardinality Invariance (A3)

Class-cardinality invariance can be achieved by following the definition of Axiom A3; that is, by defining the global measure as the aggregation of class-pairwise local measures,  $f_c(C, X, d) = \text{agg}_{S \subseteq C, |S|=2} f(S, X, \delta)$ , where  $\text{agg} = \{\text{avg}, \text{min}, \text{max}\}$ .

### T4: Achieving Range Invariance (A4)

**T4-a,b: Scaling.** A common approach to get a unit range for  $f$  is to use min-max scaling  $f_u = (f - f_{\min}) / (f_{\max} - f_{\min})$ . However, determining the minimum and maximum values of  $f$  for any data  $X$  is nontrivial. Theoretical extrema are usually computed for edge cases far from realistic  $X$  and  $C$ . Wu et al. [14] proposed estimating the worst score over a given dataset  $X$  by the expectation  $\hat{f}_{\min} = E_{\pi}(f(C^{\pi}, X, \delta))$  of  $f$  computed over random partitions  $C^{\pi}$  of  $(X, \delta)$  preserving class proportions  $|C_i^{\pi}| = |C_i| \forall i$  (**T4-a**), which are arguably the worst possible clustering partitions of  $X$ . In contrast, it is hard to estimate the maximum achievable score over  $X$ , as this is the very objective of clustering techniques. If the theoretical maximum  $f_{\max}$  is known and finite, we use it by default; otherwise, if  $f_{\max} \rightarrow +\infty$  then the scaled measure  $f_u \rightarrow 0, \forall f$ . We propose to use a logistic function  $f' = 1 / (1 + e^{-k \cdot f})$  (**T4-b**) before applying the normalization so  $f'_{\max} = 1$  and  $f'_{\min} = \hat{f}'_{\min}$ .

**T4-c: Calibrating logistic growth rate  $k$ .** We can arbitrarily make a logistic function to pull or push all scores toward the minimum or maximum value by tuning the growth rate  $k$ . We thus propose calibrating  $k$  with datasets with ground truth CLM scores. Assume a set of labeled datasets  $\mathcal{X} = \{X^1, \dots, X^n\}$  with class labels  $\mathcal{C} = \{C^1, \dots, C^n\}$  and the corresponding ground truth CLM scores  $\mathcal{S} = \{s^1, \dots, s^n\}$  where  $\min \mathcal{S} = 0$  (worst) and  $\max \mathcal{S} = 1$  (best). Here, we can optimize  $k$  to make  $\mathcal{S}' = \{s'^1, \dots, s'^n\}$  best matches with  $\mathcal{S}$ , where  $s'^i = f(C^i, X^i, \delta)$ . In practice, we use Bayesian optimization [34] targeting the  $R^2$  score.

We propose using human-driven separability scores as a proxy for CLM, building upon available human-labeled datasets acquired from a user study [35] and used in several works on visual perception of cluster patterns [36], [37], [38]. Each dataset consists of a pair of Gaussian clusters (classes) with diverse hyperparameters (e.g., covariance, position), graphically represented as a monochrome scatterplot. The perceived separability score of each pair of clusters was obtained by aggregating the judgments of 34 participants of whether they could see one or more than one cluster in these plots. The separability score of each dataset is defined as the proportion of participants who detected more than one cluster. We used these datasets  $\mathcal{X}$  and separability scores  $\mathcal{S}$  for calibration because they are not biased by a certain clustering technique or validation measure; they are based on human perception following a recent research trend in clustering [36], [39], and the probabilistic scores naturally range from 0 to 1. However, as the scores are not uniformly distributed, we bin them and weigh each dataset in proportion to the inverse size of the bin they belong to (see Appendix B for details).

## 5 ADJUSTING IVM INTO IVM<sub>A</sub>

We use the proposed protocols (Section 4) to adjust six baseline IVMs: Calinski-Harabasz index ( $CH$ ) [22], Dunn index ( $DI$ ) [40], I index ( $II$ ) [41], Xie-Beni index ( $XB$ ) [42], Davies-Bouldin index ( $DB$ ) [43], and Silhouette coefficient ( $SC$ ) [12], into IVM<sub>A</sub>s that satisfy both within- and across-dataset axioms. We pick the IVMs from the list of the survey done by Liu et al. [8]. We select every IVM except the ones optimized based on the elbow rule (e.g., modified Hubert  $\Gamma$  statistic [44]) and those that require several clustering results (e.g., S\_Dbw index [45]). Our choice covers the most widely used IVMs that have clear variety in the way of examining cluster structure.

Here, we explain the adjustment of  $CH$ . We selected  $CH$  because it does not satisfy any of the across-dataset axioms, so that we can demonstrate the application of all protocols (Section 4). The adjusted  $CH$  ( $CH_A$ ) also turned out to be the best IVM<sub>A</sub> in our evaluations (Section 6).

### 5.1 Adjusting the Calinski-Harabasz Index

$CH$  [22] is defined as:

$$CH(C, X, d^2) = \frac{|X| - |C|}{|C| - 1} \cdot \frac{\sum_{i=1}^{|C|} |C_i| d^2(c_i, c)}{\sum_{i=1}^{|C|} \sum_{x \in C_i} d^2(x, c_i)}, \quad (5)$$

where  $c_i = \overline{C_i}$  and  $c = \overline{X}$ . A higher value implies a better CLM. The denominator and numerator measure compactness and separability, respectively. The adjustment procedure is as follows:

**Applying T1 (Data-cardinality invariance).** Both the denominator and numerator of  $CH$  are already robust estimators of population statistics (T1). However, as the term  $(|X| - |C|)$  makes the score grow proportional to the size of the datasets, we remove the term to eliminate the influence of data-cardinality, resulting in:

$$CH_1(C, X, d^2) = \frac{\sum_{i=1}^{|C|} |C_i| d^2(c_i, c)}{(|C| - 1) \sum_{i=1}^{|C|} \sum_{x \in C_i} d^2(x, c_i)}. \quad (6)$$

**Applying T2 (Shift invariance).**  $CH_1$ 's numerator and denominator consists of type-3 and type-2 distances, respectively. To equalize the shift before applying exponential (T2-c), we add the sum of the squared distances of the data points to their centroid as a factor to the numerator, which does not affect separability or compactness. This leads to:

$$CH_2(C, X, d^2) = \frac{\sum_{x \in X} d^2(x, c)}{|C|} \cdot \frac{\sum_{i=1}^{|C|} |C_i| d^2(c_i, c)}{\sum_{i=1}^{|C|} \sum_{x \in C_i} d^2(x, c_i)} \cdot \frac{|C|}{|C| - 1}. \quad (7)$$

As the left term is a fraction of the sum of type-2 distances, we get shift invariance by dividing both the numerator and the denominator by  $|X|$  (i.e., the sum becomes an average; T2-b), then by applying the exponential normalized by the standard deviation  $\sigma_{d^2}$  of the square distances of the data points to their centroid (T2-a); i.e.,  $\sigma_{d^2} = \text{std}(\{d^2(x, c) | x \in X\})$ . The right term does not need an exponential protocol as type-3 distances do not shift as the dimension grows.

We still divide the term with  $|X|$  and  $\sigma_{d^2}$  to ensure data-cardinality and scale invariance, respectively. This leads to:

$$CH_3(C, X, d^2) = \frac{e^{\sum_{x \in X} \frac{d^2(x, c)}{\sigma_{d^2} \cdot |X|}}}{e^{\sum_{i=1}^{|C|} \frac{\sum_{x \in C_i} d^2(x, c_i)}{\sigma_{d^2} \cdot |X|}}} \times \frac{\sum_{i=1}^{|C|} |C_i| d^2(c_i, c)}{\sigma_{d^2} \cdot |X| \cdot (|C| - 1)}. \quad (8)$$

**Applying T4 (Range invariance).** We apply min-max scaling to make the measure range invariant.

As  $\max(CH_3) \rightarrow +\infty$ , we transform it through a logistic function (T4-b), resulting in:

$$CH_4 = 1/(1 + e^{-k \cdot CH_3}), \quad \therefore CH_{4\max} \rightarrow 1. \quad (9)$$

We then estimate the worst score  $CH_{4\min}$  (T4-a) as the average  $CH_4$  score computed over  $T$  Monte-Carlo simulations with random clustering partitions  $C^\pi$ :

$$CH_{4\min} = \frac{1}{T} \sum_{t=1}^T CH_4(C^{\pi_t}, X, d^2), \quad (10)$$

We then get:

$$CH_5 = (CH_4 - CH_{4\min}) / (CH_{4\max} - CH_{4\min}), \quad (11)$$

where we set the logistic growth rate  $k$  by calibrating the  $CH_5$  scores ( $S'$ ) with human-judgment scores ( $S$ ) (T4-c).

**Applying T3 (Class-cardinality invariance).** Lastly, we make our measure satisfy class-cardinality invariance (Axiom A3) by averaging class-pairwise scores (T3), which finally determines the adjusted Calinski-Harabasz index:

$$CH_A(C, X, d^2) = \frac{1}{\binom{|C|}{2}} \sum_{S \subseteq C, |S|=2} CH_5(S, X, d^2). \quad (12)$$

Unlike  $CH$ , which misses all across-dataset axioms,  $CH_A$  satisfies all of them (Refer to Appendix D for the proofs).

**Removing Monte-Carlo simulations.** We can reduce the computing time of  $CH_A$  by removing Monte-Carlo Simulations for estimating  $CH_{4\min}$ . Indeed, as randomly permuting class labels make all  $C_i \in C$  satisfy  $C_i \stackrel{D}{=} X$ , we can assume  $c \simeq c_i \forall c_i$ . Therefore,  $CH_3(C^\pi, X, d^2) \simeq 0$  as it contains  $d^2(c_i, c) \simeq 0$  in the second term, which leads to:

$$\begin{aligned} CH_{4\min} &= E_\pi(CH_4(C^\pi, X, d^2)) \\ &= E_\pi(1/2) = 1/2. \end{aligned} \quad (13)$$

This approximation also makes  $CH_A$  deterministic.

**Computational complexity.** Table 1 presents the time complexities of the IVMs. Since the complexity of  $CH_A$  is linear with respect to all parameters, where the only additional parameter compared to  $CH$  is  $|C|$ , the measure scales efficiently to datasets with large sizes and high dimensionality (Section 6.4).

TABLE 1

Time complexity of  $CH$ ,  $CH_A$  and their variants generated by applying our protocols (Section 5). The simulation refers to  $T$  runs of Monte-Carlo experiments used to compute  $CH_{4\min}$  Equation 10). IVMs in the first two rows do not depend on Monte-Carlo simulations.

IVMs	w/ simulation	w/o simulation
$CH, CH_1, CH_2, CH_3, CH_4$		$O( X \Delta_X)$
$CH_{4\max}$		$O(1)$
$CH_{4\min}$	$O(T X \Delta_X)$	$O( X \Delta_X)$
$CH_5$	$O(T X \Delta_X)$	$O( X \Delta_X)$
$CH_A$	$O(T X \Delta_X C )$	$O( X \Delta_X C )$

## 5.2 Adjusting the Remaining IVMs

The adjustment processes of the remaining IVMs are not significantly different from those of  $CH$ . As  $DI$  misses all across-dataset axioms, it goes through all protocols like  $CH$ .  $II$ ,  $XB$ , and  $DB$  require the shift (T2), range (T4), and class-cardinality (T3) invariance protocols as they are already data-cardinality invariant (A1). After passing through these protocols,  $II$  and  $XB$  become identical (i.e.,  $II_A = XB_A$ ). For  $SC$ , only shift and class-cardinality invariance protocols are required, as data-cardinality and range invariance (A4) are already satisfied. We thus obtain five adjusted IVMs:  $CH_A$ ,  $DI_A$ ,  $\{II, XB\}_A$ ,  $DB_A$ , and  $SC_A$ . Please refer to Appendix C for detailed adjustment processes.

## 6 EVALUATION

We conduct four experiments to evaluate our protocols (Section 4) and  $IVM_A$ s. The first experiment evaluates the effectiveness of each individual protocol by an ablation study (Section 6.1). The second and third experiments investigate how well  $IVM_A$  correlates with the ground truth CLM across (Section 6.2) and within (Section 6.3) datasets, comparing their performance to competitors like standard IVMs and supervised classifiers (Section 2). Finally, we investigate the runtime of  $IVM_A$ s and competitors in quantifying the CLM of datasets (Section 6.4).

### 6.1 Ablation Study on the Protocols

**Objectives and design.** We want to evaluate our protocols in making  $IVM_A$ s to satisfy the new axioms. For datasets, we consider the ones that we used to fit the logistic growth rate  $k$  (Section 4, T4-c), which consists of two bivariate Gaussian clusters (as classes) with various levels of CLM [35], to which we add noisy dimensions. For each  $IVM_A$   $Z_A$ , we consider its variants  $Z_v$  made by switching *on* and *off* the protocols ( $Z_v \in \{Z_P | P \in 2^{\mathcal{T}_Z}\}$  where  $\mathcal{T}_Z$  is the set of protocols switched for  $Z$ ). In the case of  $CH$  and  $DI$ , we make eight variants by switching data-cardinality (T1; D), shift invariance (T2; S), and range invariance (T4; R) protocols ( $\mathcal{T}_{CH} = \mathcal{T}_{DI} = \{T1, T2, T4\} = \{D, S, R\}$ ), and for  $II$ ,  $XB$ ,  $DB$ , we switch T2 and T4 ( $\mathcal{T}_{II} = \mathcal{T}_{XB} = \mathcal{T}_{DB} = \{T2, T4\} = \{S, R\}$ ), resulting in four variants. For  $SC$ , only T2 is turned *on* and *off*, which leads to two variants ( $\mathcal{T}_{SC} = \{T2\} = \{S\}$ ). The effect of the class-cardinality invariance protocol (T3) is not evaluated because the ground truth synthetic datasets [35] contain only two classes. We control the cardinality (A1) and dimension (A2) of the datasets to evaluate how sensitive the variants are to variations of these



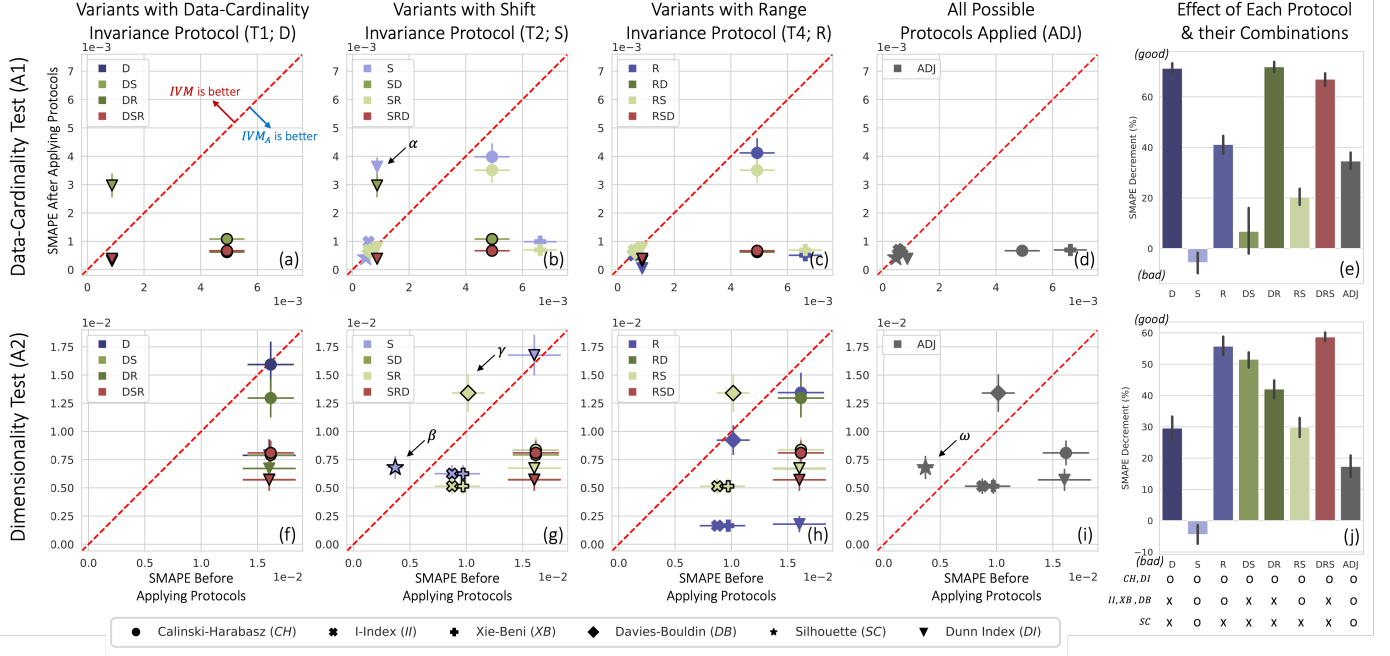


Fig. 2. Results of the ablation study (Section 6.1). Top: data-cardinality test, bottom: dimensionality test. Scatterplots (a-d, f-i) show how the SMAPE varies between IVMs and their variants in average, where each dot corresponds to a single IVM variant  $Z_v$ . The shape and the color of the dots correspond to  $Z$  and  $v$ , respectively, and  $x$  and  $y$  coordinates represent the error made by  $Z$  (before applying the protocols) and  $Z_v$  (after applying the protocols), respectively. If a dot is located in the lower-right half of the scatterplot, it means that the protocol-based variant  $Z_v$  produces less error than  $Z$ . Note that  $II$  and  $XB$  are horizontally aligned as their adjusted versions are identical. The first three columns show the effect of each protocol and its variants, and the fourth column shows the variant in which all possible protocols are applied ( $Z_v = Z_A$ ). Bar plots (e, i) show the error reduction rate made by the protocols and their combinations averaged over all the IVMs; the higher the bar, the better. Notice that ADJ shows lower error reduction compared to other combinations because not all IVMs benefit from the entire set of protocols. The table below the bar plots indicates what combinations of protocols are applied (O) to each IVM or not (X).

conditions (the lower, the better). We do not control class-cardinality (A3) as the number of classes (which is 2) is imposed by the available data. Range invariance (A4) is not controlled either, as it is imposed by the min-max protocol (T4) and is not a characteristic of the datasets.

**Datasets.** We use 1,000 base datasets  $\{X_1 \dots X_{1000}\}$ , each consisting of  $|X| = 1,000$  points sampled from two Gaussian clusters ( $|C| = 2$ ) within the 2D space and augmented with 98 noisy dimensions ( $\Delta = 100$ ). We control the eight independent parameters (ip) of the Gaussians: two covariance matrices (3 ip each), class proportions (1 ip), and the distance between Gaussian means (1 ip), following previous studies [35], [36]. New to this work, we add Gaussian noise along the supplementary dimensions, to each cluster-generated data, with mean 0 and variance equal to the minimum span of that cluster's covariance. We generate any dataset  $X_{i,t}$  by specifying a triplet  $(X_i, N_t, \Delta_t)$  with  $X_i$  a base dataset,  $N_t$  the number of data randomly sampled from  $X_i$  preserving cluster proportions, and  $\Delta_t$  its dimension where the first two dimensions always correspond to the 2D cluster space. **Sensitivity to data-cardinality (A1):** (Figure 2 top) For each of the 1,000 base data  $X_i$ , we generated 11 datasets  $X_{i,t} = (X_i, N_t, \Delta_t)_{i \in [1 \dots 1000], t \in [0 \dots 10]}$ , with the controlled data cardinality set to  $N_t = 50t + 500$  and  $\Delta_t$  drawn uniformly at random from  $[2, \dots, 100]$ . **Sensitivity to dimensionality (A2):** (Figure 2 bottom) For each of the 1000 base data  $X_i$ , we generated 11 datasets  $X_{i,t} = (X_i, N_t, \Delta_t)_{i \in [1 \dots 1000], t \in [0 \dots 10]}$ , with  $N_t$  drawn uniformly at random from  $[500, \dots, 1000]$  and the controlled dimension

set to  $\Delta_0 = 2$  or  $\Delta_t = 10t, \forall t > 0$ .

**Measurements.** We quantify the extent to which each variant of a score changes due to the feature alteration of base datasets. Formally, for each variant  $Z_v$ , we evaluate the matching between a pair  $(a, b)$  of values of the controlled factor  $t$  (e.g.  $(\Delta_a, \Delta_b) = (10, 30)$ ) across all the 1,000 base datasets using:

$$S_{k \in \{1, \dots, 1000\}}(w_k(Z_v(C_k, X_{k,a}, d) - M), w_k(Z_v(C_k, X_{k,b}, d) - M)). \quad (14)$$

where  $S$  is the symmetric mean absolute percentage error (SMAPE) [46], which is defined as

$$S_{k \in K}(F_k, G_k) = \frac{\sum_{k \in K} |F_k - G_k|}{\sum_{k \in K} (|F_k| + |G_k|)} \quad (15)$$

(0 best, 1 worst), and  $M$  is the overall minimum empirical score across all base datasets:

$$M = \min \left( \min_{k \in [1 \dots 1000]} Z_v(C_k, X_{k,a}, d), \min_{k \in [1 \dots 1000]} Z_v(C_k, X_{k,b}, d) \right). \quad (16)$$

Note that  $C_k$  is the class label of  $X_k$  and  $w_k$  is the weight of  $X_k$  to mitigate the distribution imbalance, which is also used to adjust the growth rate in the range invariance protocol (T4-c; Appendix B). We adapt SMAPE to compare measures with different ranges equally and align scores to 0 by subtracting  $M$ , as SMAPE can be over-forecasted by shifting  $F_k$  and  $G_k$  to the positive side [47].

**Results and discussions.** Figure 2 shows that the IVMs without applying our technical protocols fail to produce consistent scores across dimensions and data cardinality, whereas they succeed when satisfying the axioms. The results demonstrate the superiority of our adjusted IVMs and the adjustment process.

All three protocols substantially contribute to achieving the axioms. The scatterplots (a-d, f-i) depict how SMAPE along the pairs of the controlled factor (data-cardinality, dimensionality) is changed between IVMs and their variants in average, while the bar plots (e, j) show the average error reduction made by the protocols and their combinations. The bar graphs indicate that the data-cardinality invariance protocol (T1; D) alone works as expected; it reduces the error by about 70% in the data-cardinality test (bar D in e) and about 30% in the dimensionality test (bar D in j). The range invariance protocol (T4; R) also plays an important role in making measures less sensitive to varying factors, reducing the errors by about 40% and 55% alone for the data-cardinality (bar R in e) and dimensionality (bar R in j) tests, respectively. Unexpectedly, the shift invariance protocol (T2; S) alone does not reduce error; in both tests, there exists a case in which the error increases after applying T2 (b,  $\alpha$ ; g,  $\beta\gamma$ ). However, the error generally reduces when we apply the shift invariance protocol together with data-cardinality or range invariance protocols (DS, RS, DRS bars in e, j). We interpret that the shift invariance protocol alone does not make apparent enhancement as the exponential protocol (T4-a,b) amplifies errors that are not yet reduced by either data-cardinality or range invariance protocols; applying other protocols reduces such errors and thus reveals the benefit of the shift invariance protocol.

We find that interplays exist not only for the shift invariance protocol and generally provide positive effects. In the dimensionality test, for example, both the shift invariance protocol and the data-cardinality invariance protocol perform better when they are combined (DS) than alone (D, S). The combination of all protocols (DRS) (red dots in e-d and f-i; DRS bar in e, j) also consistently shows good performance in making measures less sensitive to the change of dimensionality and data-cardinality. Meanwhile, applying all *possible* protocols (IVM<sub>A</sub>; denoted as ADJ) does not show better performance than most of the other combinations. This is possibly because some IVMs do not fully benefit from the interplay of all protocols (table on the bottom right). For example, for *SC*, only the shift invariance protocol, which shows a negative effect without other protocols, is applied. *SC*<sub>A</sub> thus fails to reduce errors, compared to *SC* (i,  $\omega$ ). Still, applying all possible protocols generally makes IVM less sensitive to cardinality and dimension (d, i), which confirms the importance of our overall adjustment procedures (i.e., protocols) and their underlying axioms. Exploring and interpreting the interplays between the protocols and specific IVMs in more detail would be interesting for future work.

## 6.2 Across-Dataset Rank Correlation Analysis

**Objectives and design.** Five IVM<sub>A</sub> (*CH*<sub>A</sub>, *DI*<sub>A</sub>,  $\{II, XB\}_A$ , *DB*<sub>A</sub>, and *SC*<sub>A</sub>) are assessed against competitors (IVMs and classifiers) for estimating the CLM ranking of labeled datasets. We approximate a ground truth (GT) CLM ranking

of labeled datasets using multiple clustering techniques. We then compare the rankings made by all competitors to the GT using Spearman’s rank correlation.

**Datasets.** We collect 96 publicly available benchmark labeled datasets from various sources (e.g., UCI ML repository [48] and Kaggle [49]), with diverse numbers of data points, class labels, cluster patterns (presumably), and dimensionality (refer to Appendix G).

**Approximating the GT CLM.** For the lack of definite GT clusters in multidimensional real datasets, we use the maximum EVM score achieved by nine various clustering techniques (see below) on a labeled dataset as an approximation of the GT CLM score for that dataset. These GT scores are used to obtain the GT-ranking of all data sets. This scheme relies on the fact that high EVM implies good CLM (Section 1; Figure 1 A and D). We use Bayesian optimization [34] to find the best hyperparameter setting for each clustering technique. We obtain the GT-ranking based on the following four EVMs: adjusted rand index (*arand*) [50], adjusted mutual information (*ami*) [15], v-measure (*vm*) [51], and normalized mutual information (*nmi*) [52] with geometric mean. We select these measures because they are “normalized” or “adjusted” so that their scores can be compared across datasets [14], and also widely used in literature [21], [53]–[55]. For clustering techniques, we use HDBSCAN [56], DBSCAN [57], *K*-Means [58], [59], *K*-Medoids [60], *X*-Means [61], Birch [62], and single, average, and complete variants of Agglomerative clustering [63] (Appendix F).

**Competitors.** We compare IVM<sub>AS</sub>, IVMs, and classifiers, which are natural competitors in measuring CLM (Section 2), to the GT-ranking. For classifiers, we use Support Vector Machine (SVM), *k*-Nearest Neighbors (*k*NN), Multilayer Perceptron (MLP), Naive Bayesian Networks (NB), Random Forest (RF), Logistic Regression (LR), Linear Discriminant Analysis (LDA), following Rodríguez et al. [25]. We also use XGBoost (XGB), an advanced classifier based on tree boosting [64]. We use XGBoost as it adapts well regardless of the datasets’ format [64], [65], thus being suitable to all the 96 datasets composed of tabular, image, and text datasets. XGBoost also outperforms recent deep-learning-based models in classifying tabular datasets [66], a preponderant type among our datasets. Finally, we test the ensemble of classifiers. We measure the classification score of a given labeled dataset using five-fold cross-validation and Bayesian optimization [34] to ensure the fairness of the evaluation. The accuracy in predicting class labels is averaged over the five validation sets to get a proxy of the CLM score for that dataset. For the ensemble, we get the proxy as the highest accuracy score among the eight classifiers for each dataset independently [25].

**Results and discussions.** Table 2 shows that for every EVM, IVM<sub>AS</sub> outperform the competitors; first (\*\*), second (\*\*), and third (\*) places are all part of the IVM<sub>A</sub> category. IVM<sub>AS</sub> achieve about 17% (*DB*) to 81% (*DI*) of performance improvement, compared with their corresponding IVMs (average: 48%), and have *strong* (light-red cells) or *very strong* (red cells) correlation with GT-ranking according to Prion et al.’s criteria [67]. These results show that the adjustment procedure (T1-T4) relying on the new axioms



TABLE 2

The results of across-dataset rank correlation analysis (Section 6.2). The numbers are the rank correlations between the approximated ground truth CLM ranking based on nine clustering techniques and the estimated CLM ranking obtained by  $IVM_A$ ,  $IVMs$ , and classifiers.

		Ground truth ranking made by EVMs			
		ami	arand	vm	nmi
Classifiers	NB	.4126	.5276	.3157	.3130
	MLP	.4405	.5386	.3600	.3761
	LR	.4456	.5382	.3666	.3873
	XGB	.4543	.5247	.3373	.3377
	kNN	.4876	.5810	.3974	.4094
	RF	.4893	.5741	.3991	.3889
	LDA	.4999	.5726	.3945	.3606
	SVM	.5427	.6235	.4625	.4827
	Ensemble	.5536	.6162	.4486	.4531
IVM	$CH$	.5923	.6222	.4487	.3810
	$DI$	.4026	.3534	.5366	.5979
	$II$	.5668	.5957	.6086	.6454
	$XB$	.6201	.7019	.4934	.4446
	$DB$	.7091	.7513	.5719	.5015
	$SC$	.5648	.6800	.4549	.4208
$IVM_A$	$CH_A$	** .8714	** .8472	*** .8300	*** .7836
	$DI_A$	.7293	.7177	.7504	.7427
	$\{II, XB\}_A$	* .8463	* .8442	* .8060	** .7818
	$DB_A$	.8315	.8111	.7856	.7436
	$SC_A$	*** .8955	*** .8769	** .8217	* .7733

- (1) Every result was validated to be statistically significant ( $p < .001$ ) by Spearman's rank correlation test.
- (2) \*\*\* / \*\* / \*: 1st- / 2nd- / 3rd-highest scores for each EVM
- (3) ■ / ■ : very strong ( $> 0.8$ ) / strong ( $> 0.6$ ) correlation [67]

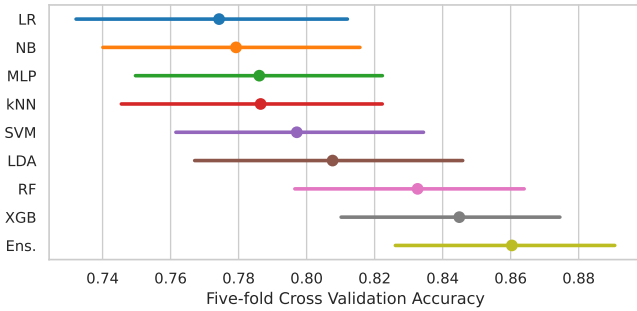


Fig. 3. The cross-validation accuracy of classifiers in classifying classes in 96 labeled datasets. The classifiers are ordered based on the mean accuracy. Error bars indicate 95% confidence interval.

(A1-A4), is beneficial to all  $IVMs$ , systematically increasing their correlation with the GT-ranking. Hence,  $IVM_A$ s are the most suitable measure to compare and rank datasets based on their CLM. Within the  $IVM_A$ s,  $CH_A$  and  $SC_A$  show the best performances, with a slight advantage for  $CH_A$  (first place for both  $vm$  and  $nmi$ , and runner-up for both  $ami$  and  $arand$ ).

In contrast, as expected, supervised classifiers fall behind the  $IVMs$  and  $IVM_A$ s, indicating that they should not be relied upon for predicting CLMs. A notable finding is the most advanced model, XGB, shows relatively poor performance in estimating CLM compared to classical models such as SVM, kNN, and LDA; even an ensemble of classifiers falls behind SVM in terms of  $arand$ ,  $vm$ , and  $nmi$  (Table 2). This is because XGB and ensemble classifiers effectively discriminate classes regardless of whether they are well-separated

TABLE 3

The results of within-dataset rank correlation analysis (Section 6.3). We compared the pairs of rankings obtained by  $IVM$  and  $IVM_A$ , with the ground truth noisy labels ranking (NR) on the 96 collected datasets.

$IVM$	$IVM$ vs. $IVM_A$	NR vs. $IVM$	NR vs. $IVM_A$
$CH$	.848 ± .263	.876 ± .296	.879 ± .277
$DI$	.253 ± .429	.451 ± .335	.381 ± .727
$II$	.825 ± .296	.857 ± .284	.881 ± .307
$XB$	.820 ± .308	.832 ± .341	.881 ± .307
$DB$	.855 ± .257	.884 ± .268	.876 ± .294
$SC$	.515 ± .574	.530 ± .601	.878 ± .322

■ / ■ : very strong ( $> 0.8$ ) / strong ( $> 0.6$ ) correlation [67]

by a large margin or not in the data space (Figure 3 gray and yellow points), leading them to classify most datasets as having similarly good CLM. This finding indicates that improving classification accuracy does not necessarily help achieve better CLM measurement, further emphasizing the significance of our contribution.

### 6.3 Within-Dataset Rank Correlation Analysis

**Objectives and design.** We want to evaluate the  $IVM_A$ 's ability to evaluate and compare CLM within a dataset, which is the original purpose of an  $IVM$ . For this purpose, we generate several noisy label variants of each dataset and compare how the scores achieved by  $IVMs$  and their adjusted counterparts are correlated with the ground truth noisy label ranking (NR). Assume a set of datasets  $\mathcal{X} = \{X_i | i = 1, \dots, n\}$  and their corresponding labels  $\{C_j | j = 1, \dots, n\}$ . For each dataset  $(X_k, C_k)$ , we run the following process. First, we generate 11 noisy label variants of each dataset  $\{C_{k,l} | l = 0, \dots, 10\}$  by randomly shuffling  $l \cdot 10\%$  of their labels. The  $l$ -th noisy label dataset is authoritatively ranked at the  $(11 - l)$ -th place of the NR (i.e., the larger the proportion  $l$  of shuffled labels is, the lower the expected CLM). Then, for each  $IVM Z \in \{CH, DI, II, XB, DB, SC\}$  and its corresponding  $IVM_A$  (i.e.,  $Z_A$ ), we compute the CLM ranking of these noisy label datasets based on  $Z(C_{k,l}, X_k, \delta)$  and  $Z_A(C_{k,l}, X_k, \delta)$ , respectively. We examine how the ranking generated by  $IVMs$  and  $IVM_A$ s are similar to NR using Spearman's rank correlation. We also check the rank correlation between the rankings from  $IVMs$  and  $IVM_A$ s.

**Datasets.** For  $\mathcal{X}$ , we use the 96 labeled datasets from Section 6.2.

**Results and discussions.** As shown in Table 3, every  $IVM_A$  has a very strong rank correlation ( $> 0.8$ ) with both NR and  $IVM$  for every case except for  $DI_A$ . The  $IVM_A$ s showed equal ( $CH, DI, II, XB, DB$ ) or better ( $SC$ ) performance in estimating the CLM within a dataset. We also see that the discrepancy between  $IVM$  and  $IVM_A$  rankings follows the one between  $IVM$  and GT noisy labels ranking. Such results verify the effectiveness of our protocols and  $IVM_A$ s in precisely measuring CLM within a dataset.

### 6.4 Runtime Analysis

**Objectives and design.** We compare the runtime of the approaches explored in Section 6.2 to estimate the CLM of 96 labeled datasets. For classifiers and the clustering ensemble,

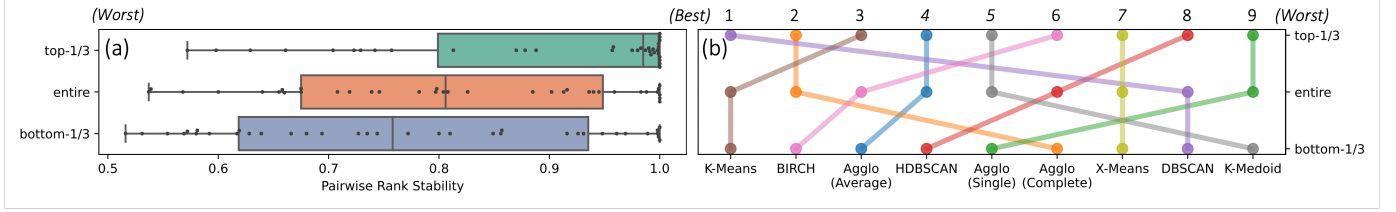


Fig. 4. (a) Distribution of pairwise rank stability for bottom-1/3 (blue;  $P^-$ ), entire (orange;  $P^*$ ), and top-1/3 (green;  $P^+$ ) subsets of 96 labeled datasets based on the CLM score computed by  $CH_A$ . (b) Rankings of clustering techniques for each set; rankings are not stable and can change dramatically if we use low-quality datasets ((a) blue and orange bars). All rankings are based on  $ami$  averaged over each subset. Using the top-ranked datasets leads to more stable and reliable rankings ((a) green bar).

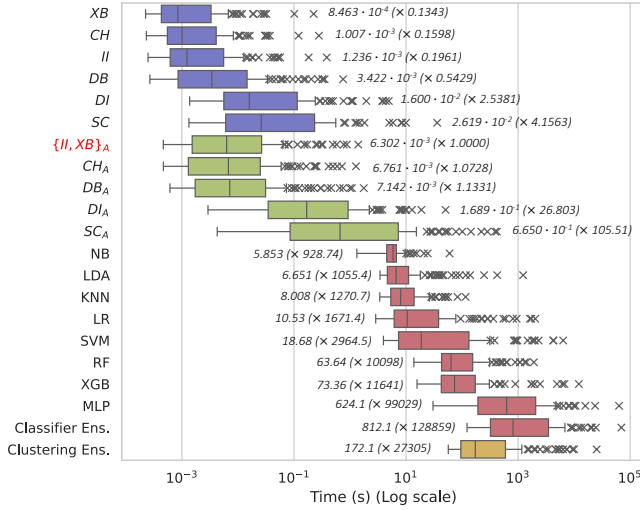


Fig. 5. The runtime of the IVM (blue),  $IVM_A$  (green), classifiers (red), and the clustering ensemble (Ens.; yellow) in computing the CLM of 96 datasets. The numbers next to each box depict the median runtime of the corresponding measure (left) and the relative time compared to  $\{II, XB\}_A$ , the fastest  $IVM_A$  (e.g.,  $CH_A$  is 1.07 times slower than  $\{II, XB\}_A$ ).

we measure the time of the entire optimization (Section 6.2). See Appendix F for experimental settings, including the apparatus we use.

**Results and discussion.** As a result (Figure 5),  $IVM_{AS}$  are up to one order of magnitude slower than  $XB$ , the fastest IVM. However, they are up to *four orders of magnitude* ( $\times 10,000$ ) faster than the competitors like clustering ensemble used to estimate ground truth CLM in Section 6.2. This verifies that most  $IVM_{AS}$ , among which is  $CH_A$ , show an excellent tradeoff between accuracy and speed. Despite  $SC_A$  being as accurate as  $CH_A$  (Section 6.2), it is *two orders of magnitude* slower ( $\times 0.01$ ), making  $CH_A$  the best  $IVM_A$  to use in practice.

## 7 APPLICATIONS

We present two applications of  $IVM_{AS}$ . First, we show that evaluating the CLM of benchmark datasets beforehand and using only those with the highest CLM scores enhances the stability and robustness of external validation and ranking of clustering techniques (Section 7.1). We also show that  $IVM_{AS}$  can be leveraged to improve the CLM of benchmark datasets (Section 7.2).

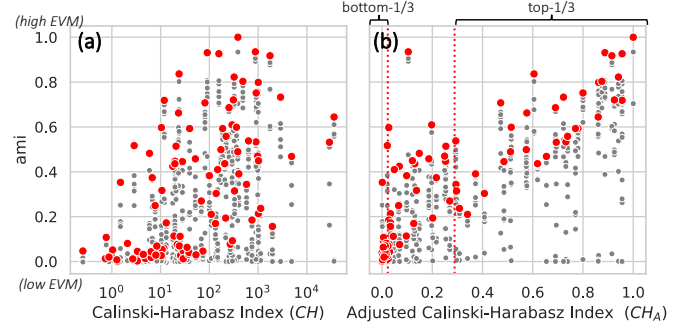


Fig. 6. All (gray points) and best (red points)  $ami$  scores of GT clustering techniques for the 96 benchmark datasets ranked by  $CH$  (left) and  $CH_A$  (right). The correlation with ground truth is clearly better for the adjusted index (Table 2). The top-1/3 datasets ( $P^+$ ) in terms of  $CH_A$  (right) are the most reliable datasets to use to evaluate and compare clustering techniques using EVMs (Section 7.1).

### 7.1 Ranking Benchmark Datasets for Reliable EVM

**Objectives and design.** We want to demonstrate the importance of evaluating the CLM of benchmark datasets prior to conducting the external validation. Here, we consider the entire set of 96 labeled datasets ( $P^*$ ), and also the top-1/3 ( $P^+$ ) and bottom-1/3 ( $P^-$ ) datasets (Figure 6) based on their  $CH_A$  ranking. We consider simulating the situation whereby a data scientist would arbitrarily choose 10 benchmark datasets ( $B$ ) among the datasets at hand for the task  $T$  of ranking clustering techniques according to  $EVM_B$ , the average EVM over  $B$ . For each  $P \in \{P^+, P^*, P^-\}$ , we simulate 100 times picking  $B$  at random among  $P$ . For each  $P$ , we measure the pairwise rank stability  $P_B(A, B) = \max(1 - p, p)$  of clustering techniques  $A$  and  $B$  over  $B$  by counting the proportion  $p$  of cases  $ami_B(A) > ami_B(B)$ .

**Hypothesis.** We expect that conducting  $T$  on any subset of good-CLM datasets will provide similar rankings (Figure 1A), where pairwise ranks remain stable ( $\forall(A, B), P_B(A, B) \approx 1$ ), whereas conducting  $T$  using bad-CLM datasets will lead to arbitrary and unstable rankings ( $\forall(A, B), P_B(A, B)$  spreads over  $[0.5, 1]$ ) (Figure 1BEH).

**Results and discussion.** As depicted in Figure 4a, picking datasets from  $P^+$  provides more stable rankings of clustering techniques, compared with  $P^-$  and  $P^*$ , which validates our hypothesis. Moreover, we find that the rankings of techniques made by  $ami_{P^+}$ ,  $ami_{P^*}$ , and  $ami_{P^-}$  are drastically different (Figure 4b; e.g., DBSCAN is in first place with  $P^+$ , but in eighth place with  $P^*$  or  $P^-$ ). Still, some datasets within  $P^-$  (e.g., Spambase and Hepatitis [48]) were used

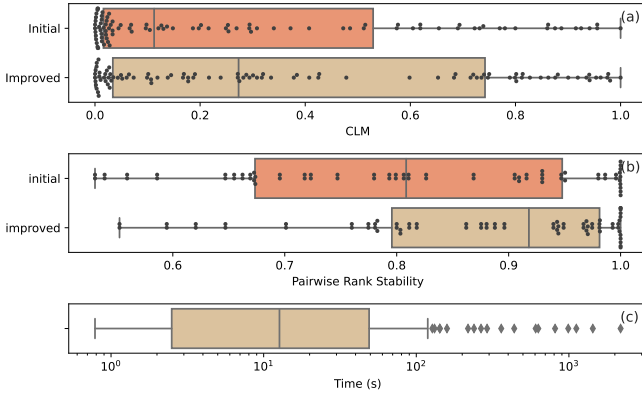


Fig. 7. (a) A comparison of the CLM of 96 labeled datasets before (initial) and after (improved) applying our algorithm that uses  $CH_A$  to enhance CLM (Section 7.2). (b) A comparison of rank stability of EVM using the initial and improved dataset. (c) The runtime of the algorithm. The results confirm that  $IVM_{AS}$  can enhance the reliability of EVM benchmark datasets within a reasonable time, lower than a minute for most datasets.

for external validation in previous studies [18], [19] without CLM evaluation, casting doubt on their conclusions and showing that this issue shall gain more attention in the clustering community.

## 7.2 Improving the CLM of Benchmark Datasets

**Objectives and design.** While datasets with good CLM lead to more reliable EVM (Section 7.1), there is a limited number of such datasets. We thus propose to improve the CLM of existing datasets to enhance diversity and robustness in executing EVM. As a proof-of-concept, we show that  $IVM_{AS}$  can be used to improve the CLM of labeled datasets by implementing a feature selection algorithm that finds a subspace of a given dataset that maximizes  $IVM_A$  scores. Formally, given  $X$ ,  $C$ ,  $\delta$ , and  $IVM_A$   $f$ , the algorithm finds the binary weight vector  $w^* \in \{0, 1\}^{|X|}$  that satisfies:

$$w^* = \arg \max_{w \in \{0, 1\}^{|X|}} f(C, w \cdot X, \delta), \quad (17)$$

where  $\cdot$  denotes the column-wise weight product:

$$w \cdot X = (w^T * \mathbf{1}) \odot X, \quad (18)$$

where  $\odot$  denotes the Hadamard product and  $*$  represents a standard matrix product. The algorithm returns the column-filtered data  $X^* = w^* \cdot X$  as output. We generate 1,000 random weight vectors  $w$  and pick the one that maximizes  $f$ , while using  $CH_A$  as  $f$ . We run the algorithm for each of the 96 labeled datasets ( $P^*$ ) and compare the CLM of the original datasets to their optimally-filtered counterparts. We also repeat the rank stability experiment (Section 7.1) using the improved datasets. We evaluate how the improved set differs from the original set in terms of rank stability. Finally, we record the runtime of the algorithm to check whether the improvement is achieved in a reasonable time.

**Results and discussions.** As illustrated in Figure 7a, there is a substantial overall increase in the CLM of datasets, which confirms the effectiveness of  $IVM_{AS}$  in improving CLM. The improved dataset also outperformed the original datasets in terms of rank stability (Figure 7b), which shows that

the improved dataset substantially enhances the reliability of EVM. We also find that the algorithm completes in less than 20 seconds on average (Figure 7c), and takes less than an hour even for the largest dataset. In summary, utilizing  $IVM_A$  allows for readily generating reliable benchmark datasets for EVM within a practical time frame.

## 8 DISCUSSIONS

### 8.1 Benefits of Adjusted IVMs

Utilizing an  $IVM_A$  to estimate the CLM of a dataset drastically reduces the runtime by *four orders of magnitude* compared to the clustering ensemble (Section 6.4); this translates to a one-year CLM computation being shortened to just 53 minutes. Given this computational efficiency,  $IVM_{AS}$  can be used to improve the CLM of labeled datasets by selecting their dimensions, even with a naive and costly random search (Section 7.2). Such improvement enables applications that require real-time measurement of CLM. For example, it is now possible to estimate the CLM quality on the fly for streaming data without resorting to distributed computing or approximate solutions. We can also compute the CLM of very large datasets, e.g., image or text data in their transformer [68] space during the training phase, to evaluate the quality of this representation at each iteration or in each layer [69], yet a task impracticable with clustering ensembles.

### 8.2 Consistency, Soundness, and Completeness of the Axioms

We discuss the consistency, soundness, and completeness of the extended axioms (within- and across-dataset axioms), following A&B [11]. We remind the definition of these three characteristics: **(1) Consistency:** the set of axioms is consistent if at least one ideal object (i.e.,  $IVM_A$ ) satisfies all of them, which means that there is no contradiction between the axioms. We use the same terminology with W2 (consistency of IVM), but with a different meaning (consistency of axioms). **(2) Soundness:** the set of axioms is sound if every object existing in the target group (i.e., every  $IVM_A$ ) satisfies all the axioms. **(3) Completeness:** the set of axioms is complete if any object not included in the target group (i.e., non- $IVM_A$ ) fails at least one axiom. We discuss them for  $IVM_{AS}$ :

**Consistency.** The existence of five  $IVM_{AS}$  that satisfy all within- and across-dataset axioms (Appendix D) validates their consistency.

**Soundness.** Soundness cannot be proven as we lack a clear definition that differentiates  $IVM_{AS}$  and non- $IVM_{AS}$ . However, if we define  $IVM_A$  as the set of measures that can be adjusted from any IVM, the axioms are sound as our technical protocols guarantee that all adjusted measures satisfy or approach the axioms by design. We suggest both defining a valid  $IVM_A$  as a “function that satisfies all within-dataset and across-dataset axioms” and using the axioms as guidelines to adjust other IVMs or to design new  $IVM_A$ , thereby preserving the soundness of all the axioms.

**Completeness.** Our extended axioms are more complete than the within-dataset axioms in terms of defining  $IVM_A$ ,

given that standard IVMs do not meet the new across-dataset axioms. The across-dataset rank correlation analysis (Section 6.2) empirically quantifies and validates this increased completeness. However, it remains uncertain if our axioms cover every aspect that can vary across datasets. There might be a function that fulfills all axioms but is not a valid IVM<sub>A</sub>. Searching for such a function and making the axioms more complete will be an interesting avenue for future work.

## 9 CONCLUSION

In this research, we provide a grounded way to evaluate the validity of labeled datasets used as benchmarks for external clustering validation. We propose doing so by measuring their level of cluster-label matching (CLM). We propose new across-dataset axioms and technical protocols to generate measures that satisfy the axioms. We use these protocols to design five adjusted internal validation measures (IVM<sub>AS</sub>), generalizing standard IVMs, to estimate the CLM of benchmark labeled datasets and rank them accordingly. A series of experiments confirm IVM<sub>AS</sub>' accuracy, scalability, and practical effectiveness in supporting reliable external clustering validation. As the primary practical outcome of this work, the 96 datasets ranked by CLM estimated by the proposed IVM<sub>AS</sub> measures are available at [github.com/hj-n/labeled-datasets](https://github.com/hj-n/labeled-datasets) for use by practitioners to generate more trustworthy comparative evaluations of clustering techniques.

As future work, we would like to explore other uses of IVM<sub>A</sub> and design new axioms to build better clustering benchmarks. For example, designing IVM<sub>AS</sub> that consider non-globular and hierarchical clusters would be an interesting path to explore. Developing better optimization techniques for maximizing the CLM of benchmark datasets is another path worthy of further study. Finally, given the high scalability of IVM<sub>AS</sub>, we envision using these CLM measures to compare cluster structures of labeled data across different latent spaces, like in the layers of foundational models for exploring concept representations during pre-training or fine-tuning.

The datasets and code are available at [github.com/hj-n/labeled-datasets](https://github.com/hj-n/labeled-datasets) and [github.com/hj-n/clm](https://github.com/hj-n/clm), respectively.

## ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2023R1A2C200520911) and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]. Hyeon Jeon is in part supported by Google Ph.D. Fellowship.

## REFERENCES

- [1] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
- [2] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007. doi: 10.1016/j.cosrev.2007.05.001.
- [3] S. L'Yi, B. Ko, D. Shin, Y.-J. Cho, J. Lee, B. Kim, and J. Seo, "Xclusim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data," *BMC bioinformatics*, vol. 16, no. 11, pp. 1–15, 2015. doi: 10.1186/1471-2105-16-S11-S5.
- [4] H. Jeon, H. Lee, Y.-H. Kuo, T. Yang, D. Archambault, S. Ko, T. Fujiwara, K.-L. Ma, and J. Seo, "Unveiling high-dimensional backstage: A survey for reliable visual analytics with dimensionality reduction," 2025. doi: 10.48550/arXiv.2501.10168.
- [5] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [6] S. Bonakala, M. Aupetit, H. Bensmail, and F. El-Mellouhi, "A human-in-the-loop approach for visual clustering of overlapping materials science data," *Digital Discovery*, vol. 3, pp. 502–513, 2024. doi: 10.1039/D3DD00179B. [Online]. Available: <http://dx.doi.org/10.1039/D3DD00179B>
- [7] K. Darwish, P. Stefanov, M. Aupetit, and P. Nakov, "Unsupervised user stance detection on twitter," in *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, M. D. Choudhury, R. Chunara, A. Culotta, and B. F. Welles, Eds. AAAI Press, 2020, pp. 141–152. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/7286>
- [8] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 911–916. doi: 10.1109/ICDM.2010.35.
- [9] I. Färber, S. Günemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek, "On using class-labels in evaluation of clusterings," in *MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD*, 2010, p. 1.
- [10] B. A. Hassan, N. B. Tayfor, A. A. Hassan, A. M. Ahmed, T. A. Rashid, and N. N. Abdalla, "From a-to-z review of clustering validation indices," *Neurocomputing*, vol. 601, p. 128198, 2024. doi: <https://doi.org/10.1016/j.neucom.2024.128198>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S092523122400969X>
- [11] M. Ackerman and S. Ben-David, "Measures of clustering quality: A working set of axioms for clustering," in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21. Curran Associates, Inc., 2008.
- [12] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. doi: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- [13] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013. doi: 10.1109/TPAMI.2013.57.
- [14] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09, 2009, p. 877–886. doi: 10.1145/1557019.1557115.
- [15] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.
- [16] M. Aupetit, "Sanity check for class-coloring-based evaluation of dimension reduction techniques," in *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, ser. BELIV '14, 2014, p. 134–141. doi: 10.1145/2669557.2669578.
- [17] H. Jeon, Y.-H. Kuo, M. Aupetit, K.-L. Ma, and J. Seo, "Classes are not clusters: Improving label-based evaluation of dimensionality reduction," pp. 781–791, 2024. doi: 10.1109/TVCG.2023.3327187.
- [18] M. K. Khan, S. Sarker, S. M. Ahmed, and M. H. A. Khan, "K-cosine-means clustering algorithm," in *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, 2021, pp. 1–4. doi: 10.1109/ICECIT54077.2021.9641480.
- [19] N. Monath, M. Zaheer, D. Silva, A. McCallum, and A. Ahmed, "Gradient-based hierarchical clustering using continuous representations of trees in hyperbolic space," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19, New York, NY, USA, 2019, p. 714–722. doi: 10.1145/3292500.3330997.
- [20] N. Masuyama, Y. Nojima, C. K. Loo, and H. Ishibuchi, "Multi-label classification via adaptive resonance theory-based clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2022. doi: 10.1109/TPAMI.2022.3230414.



- [21] H. Liu, Z. Tao, and Y. Fu, "Partition level constrained clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2469–2483, 2018. doi: 10.1109/TPAMI.2017.2763945.
- [22] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974. doi: 10.1080/03610927408827101.
- [23] K. Ethayarajh, Y. Choi, and S. Swayamdipta, "Understanding dataset difficulty with V-usable information," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 17–23 Jul 2022, pp. 5988–6008.
- [24] O. Abul, A. Lo, R. Alhaji, F. Polat, and K. Barker, "Cluster validity analysis using subsampling," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, vol. 2, 2003, pp. 1435–1440 vol.2. doi: 10.1109/ICSMC.2003.1244614.
- [25] J. Rodríguez, M. A. Medina-Pérez, A. E. Gutierrez-Rodríguez, R. Monroy, and H. Terashima-Marín, "Cluster validation using an ensemble of supervised classifiers," *Knowledge-Based Systems*, vol. 145, pp. 134–144, 2018. doi: <https://doi.org/10.1016/j.knsys.2018.01.010>.
- [26] S. Vega-Pons and J. Ruiz-Shulcloper, "A survey of clustering ensemble algorithms," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 03, pp. 337–372, 2011. doi: 10.1142/S0218001411008683.
- [27] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to data mining, addison," ed: Boston, MA USA: Wesley Longman, Publishing Co., Inc, 2005.
- [28] J. Kleinberg, "An impossibility theorem for clustering," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15. MIT Press, 2002.
- [29] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *International conference on database theory*. Springer, 1999, pp. 217–235. doi: 10.1007/3-540-49257-7\_15.
- [30] D. Francois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873–886, 2007. doi: 10.1109/TKDE.2007.1037.
- [31] J. A. Lee and M. Verleysen, "Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants," *Procedia Computer Science*, vol. 4, pp. 538–547, 2011. doi: 10.1016/j.procs.2011.04.056., proceedings of the International Conference on Computational Science, ICCS 2011.
- [32] J. Lee and M. Verleysen, "Two key properties of dimensionality reduction methods," in *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2014, pp. 163–170. doi: 10.1109/CIDM.2014.7008663.
- [33] A. W. v. d. Vaart, *Asymptotic Statistics*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.
- [34] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [35] M. M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail, "Clustme: A visual quality measure for ranking monochrome scatterplots based on cluster patterns," *Computer Graphics Forum*, vol. 38, no. 3, pp. 225–236, 2019. doi: <https://doi.org/10.1111/cgf.13684>.
- [36] M. Aupetit, M. Sedlmair, M. M. Abbas, A. Baggag, and H. Bensmail, "Toward perception-based evaluation of clustering techniques for visual analytics," in *30th IEEE Visualization Conference, IEEE VIS 2019 - Short Papers*. IEEE, 2019, pp. 141–145. doi: 10.1109/VISUAL.2019.8933620.
- [37] M. M. Abbas, E. Ullah, A. Baggag, H. Bensmail, M. Sedlmair, and M. Aupetit, "Clustml: A measure of cluster pattern complexity in scatterplots learnt from human-labeled groupings," *Information Visualization*, vol. 23, no. 2, pp. 105–122, 2024. doi: 10.1177/14738716231220536. [Online]. Available: <https://doi.org/10.1177/14738716231220536>
- [38] H. Jeon, G. J. Quadri, H. Lee, P. Rosen, D. A. Szafir, and J. Seo, "Clams: A cluster ambiguity measure for estimating perceptual variability in visual clustering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 1, pp. 770–780, 2024. doi: 10.1109/TVCG.2023.3327201.
- [39] G. Blasilli, D. Kerrigan, E. Bertini, and G. Santucci, "Towards a visual perception-based analysis of clustering quality metrics," in *2024 IEEE Visualization in Data Science (VDS)*, 2024, pp. 15–24. doi: 10.1109/VDS63897.2024.00007.
- [40] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974. doi: 10.1080/01969727408546059.
- [41] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002. doi: 10.1109/TPAMI.2002.1114856.
- [42] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 8, pp. 841–847, 1991. doi: 10.1109/34.85677.
- [43] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.
- [44] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985. doi: 10.1007/BF01908075.
- [45] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: finding the optimal partitioning of a data set," in *Proceedings 2001 IEEE International Conference on Data Mining*, 2001, pp. 187–194. doi: 10.1109/ICDM.2001.989517.
- [46] C. Tofallis, "A better measure of relative prediction accuracy for model selection and model estimation," *Journal of the Operational Research Society*, vol. 66, no. 8, pp. 1352–1362, 2015. doi: 10.1057/jors.2014.103.
- [47] D. Putz, M. Gumhalter, and H. Auer, "A novel approach to multi-horizon wind power forecasting based on deep neural architecture," *Renewable Energy*, vol. 178, pp. 494–505, 2021. doi: <https://doi.org/10.1016/j.renene.2021.06.099>.
- [48] A. Asuncion and D. Newman, "Uci machine learning repository."
- [49] "Kaggle," <https://www.kaggle.com>, 2010.
- [50] J. M. Santos and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *Artificial Neural Networks – ICANN 2009*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 175–184. doi: 10.1007/978-3-642-04277-5\_18.
- [51] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007, pp. 410–420.
- [52] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.
- [53] C. Xiong, D. M. Johnson, and J. J. Corso, "Active clustering with model-based uncertainty reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 5–17, 2017. doi: 10.1109/TPAMI.2016.2539965.
- [54] Y. Zhang and Y.-m. Cheung, "Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3560–3576, 2022. doi: 10.1109/TPAMI.2021.3056510.
- [55] S. Chakraborty and S. Das, "Detecting meaningful clusters from high-dimensional data: A strongly consistent sparse center-based clustering approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2894–2908, 2022. doi: 10.1109/TPAMI.2020.3047489.
- [56] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172. doi: 978-3-642-37456-2\_14.
- [57] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Db-scan revisited, revisited: Why and how you should (still) use dbscan," *ACM Trans. Database Syst.*, vol. 42, no. 3, jul 2017. doi: 10.1145/3068335.
- [58] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979. doi: 10.2307/2346830.
- [59] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003. doi: [https://doi.org/10.1016/S0031-3203\(02\)00060-2](https://doi.org/10.1016/S0031-3203(02)00060-2), biometrics.

- [60] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3336–3341, 2009. doi: <https://doi.org/10.1016/j.eswa.2008.01.039>.
- [61] D. Pelleg, A. W. Moore *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters." in *Icml*, vol. 1, 2000, pp. 727–734.
- [62] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: An efficient data clustering method for very large databases," in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '96. New York, NY, USA: Association for Computing Machinery, 1996, p. 103–114. doi: 10.1145/233269.233324.
- [63] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011. doi: 10.48550/arXiv.1109.2378.
- [64] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. doi: 10.1145/2939672.2939785.
- [65] M. Bohacek and M. Bravansky, "When XGBoost outperforms GPT-4 on text classification: A case study," in *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 51–60. doi: 10.18653/v1/2024.trustnlp-1.5.
- [66] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022, pp. 507–520.
- [67] S. Prion and K. A. Haerling, "Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient," *Clinical Simulation in Nursing*, vol. 10, no. 10, pp. 535–536, 2014. doi: 10.1016/j.ecns.2014.07.005.
- [68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [69] F. Dalvi, A. R. Khan, F. Alam, N. Durrani, J. Xu, and H. Sajjad, "Discovering latent concepts learned in BERT," in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=POTMtpYI1xH>